IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF DELAWARE

TWINSTRAND BIOSCIENCES, INC., & UNIVERSITY OF WASHINGTON, Plaintiffs and Counterclaim Defendants, v. GUARDANT HEALTH, INC., Defendant and Counterclaim Plaintiff.	C.A. No. 21-1126-LPS **JURY TRIAL DEMANDED**

DEFENDANT GUARDANT HEALTH, INC.'S AMENDED ANSWER, AFFIRMATIVE DEFENSES, AND COUNTERCLAIMS TO PLAINTIFFS' COMPLAINT

Defendant Guardant Health, Inc. ("Guardant") hereby answers the Complaint filed by TwinStrand Biosciences, Inc., ("TwinStrand") and University of Washington ("UW," collectively, "Plaintiffs") on August 3, 2021, and submits its affirmative answers and counterclaims as follows.

Except as expressly admitted, Guardant denies each and every allegation of Plaintiffs' Complaint. Guardant's specific responses to the numbered allegations are set forth below and correspond to the numbered allegations in Plaintiffs' Complaint. The opening paragraph of Plaintiffs' Complaint does not appear to contain any allegations requiring a response. To the extent the opening paragraph contains allegations directed to Guardant, Guardant admits that Plaintiffs' pleading purports to be a complaint for patent infringement, but denies that Guardant has infringed any valid and enforceable patent claim or that Plaintiffs are entitled to any relief.

For ease of reference, Guardant has included the headings from Plaintiffs' Compliant. The inclusion of these headings is not an admission by Guardant and to the extent that any heading or

other non-numbered statement in Plaintiffs' Complaint contains or implies any allegations, Guardant denies each and every allegation therein.

NATURE OF THE ACTION

1. Guardant admits that the Complaint alleges infringement of United States Patent Nos. 10,287,631 ("the '631 patent"); 10,689,699 ("the '699 patent"); 10,752,951 ("the '951 patent"); and 10,760,127 ("the '127 patent") (collectively, "the Asserted Patents") under the patent laws of the United States, 35 U.S.C. § 1, *et seq.* Guardant denies that Plaintiffs are entitled to any relief, including any relief under the patent laws of the United States. Except as expressly admitted herein, Guardant denies each and every allegation contained in Paragraph 1 of the Complaint.

THE PARTIES

2. Guardant is without knowledge or information sufficient to form a belief as to the truth of the allegations of this paragraph, and therefore denies them.

3. Guardant is without knowledge or information sufficient to form a belief as to the truth of the allegations of this paragraph, and therefore denies them.

4. Guardant admits that Guardant is a corporation organized and existing under the laws of the State of Delaware with its principal place of business at 505 Penobscot Dr., Redwood City, CA 94063.

JURISDICTION AND VENUE

5. Guardant admits that Plaintiffs purport to assert claims for patent infringement but denies that Guardant has infringed any valid and enforceable patent claim or that Plaintiffs are entitled to any relief. For purposes of this action, Guardant admits that this Court has jurisdiction over the subject matter of this action under 28 U.S.C. §§ 1331, 1338(a), 2201, and 2202 to the

extent that Plaintiffs are able to demonstrate a proper claim for patent infringement. The remaining allegations of this paragraph are legal conclusions, and, on that basis, Guardant denies them.

6. Guardant admits that it is a Delaware corporation. For purposes of this action, Guardant does not contest personal jurisdiction. The remaining allegations of this paragraph are legal conclusions, and, on that basis, Guardant denies them.

7. For purposes of this action, Guardant does not contest personal jurisdiction, but it denies committing, directly or indirectly, any acts of infringement. The remaining allegations of this paragraph are legal conclusions, and, on that basis, Guardant denies them.

8. Guardant admits that Guardant is a Delaware corporation. For purposes of this action, Guardant does not contest that venue is proper but denies that venue is convenient. The remaining allegations of this paragraph are legal conclusions, and, on that basis, Guardant denies them.

THE PATENTS-IN-SUIT

9. Guardant admits that a document that purports to be a copy of the ’631 patent is attached as Exhibit A to the Complaint. On its face, Exhibit A appears to be entitled “Methods of Lowering the Error Rate of Massively Parallel DNA Sequencing Using Duplex Consensus Sequencing” and appears to have been issued on May 14, 2019. Guardant denies the remaining allegations of this paragraph.

10. Guardant admits that a document that purports to be a copy of the ’699 patent is attached as Exhibit B to the Complaint. On its face, Exhibit B appears to be entitled “Methods of Lowering the Error Rate of Massively Parallel DNA Sequencing Using Duplex Consensus Sequencing” and appears to have been issued on June 23, 2020. Guardant denies the remaining allegations of this paragraph.

11. Guardant admits that a document that purports to be a copy of the '951 patent is attached as Exhibit C to the Complaint. On its face, Exhibit C appears to be entitled "Methods of Lowering the Error Rate of Massively Parallel DNA Sequencing Using Duplex Consensus Sequencing" and appears to have been issued on August 25, 2020. Guardant denies the remaining allegations of this paragraph.

12. Guardant admits that a document that purports to be a copy of the '127 patent is attached as Exhibit D to the Complaint. On its face, Exhibit D appears to be entitled "Methods of Lowering the Error Rate of Massively Parallel DNA Sequencing Using Duplex Consensus Sequencing" and appears to have been issued on September 1, 2020. Guardant denies the remaining allegations of this paragraph.

BACKGROUND

13. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however, Guardant states it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

14. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however, Guardant states it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

A. The Need for High-Accuracy, High-Throughput Sequencing Methods

15. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however, Guardant states that detecting genetic mutations is important in the diagnosis of cancers and other

illnesses. As to the remaining allegations, Guardant states that it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

16. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however, Guardant states that a sequencing method known as Sanger sequencing was used in some manner as part of the Human Genome Project. As to the remaining allegations, Guardant states that it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

17. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however, Guardant states that, in general, some forms of Next Generation Sequencing ("NGS") have the ability to sequence DNA molecules at a much higher throughput compared to older sequencing methods. As to the remaining allegations, Guardant states that it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

18. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however, Guardant states that it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

B. The Inventions

19. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however, Guardant states that it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

20. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however, Guardant states that it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

21. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however, Guardant states that it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

22. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however, Guardant states that it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

23. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however, Guardant states that it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

24. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however, Guardant states that it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

25. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however,

Guardant states that it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

26. The allegations made in this paragraph do not appear to be directed at Guardant, and therefore no response is required. To the extent any response may be required, however, Guardant states that it lacks knowledge or information sufficient to form a belief about the truth of the allegations and thus denies the same.

C. Guardant's [Alleged] Willful Infringement of the Asserted Patents

1. Guardant's [Alleged] infringing products and services

27. Guardant admits that it began selling certain products and services under the moniker "Digital Sequencing Technology" within the United States in 2014. Guardant admits that a document that purports to be a copy of the Guardant IPO Prospectus is attached as Exhibit F to the Complaint. Except as expressly admitted, Guardant denies the allegations of this paragraph.

28. Guardant admits that it sells diagnostic kits with the names Guardant360 lab developed test ("LDT"), Guardant360 CDx ("CDx"), GuardantOMNI ("Omni"), Guardant Reveal ("Reveal"), Guardant360 Response, and Guardant360 TissueNext (collectively, the "Guardant Kits").[1] Guardant admits that it has customers outside of the United States. Except as expressly admitted, Guardant denies the allegations of this paragraph.

29. Guardant admits that in 2017, Dr. Rick Lanman was Guardant's Chief Medical Officer. As to the remaining allegations of this paragraph, Guardant is without knowledge or information sufficient at this time to form a belief as to the truth of the allegations of this paragraph and, on that basis, denies them.

[1] Guardant's LUNAR-2 is currently still in clinical trial which is recruiting through the end of 2021 (Evaluation of the ctDNA LUNAR test in an Average Patient Screening Episode or "ECLIPSE" trial). *See* ClinicalTraial.gov Identified: NCT04136002.Data from this trial will support submission to the FDA.

30. Guardant admits that it provides certain data services to its customers and that it uses the terms Guardant Connect and GuardantINFORM (collectively, "the Guardant Services") to describe some of its services. Except as expressly admitted, Guardant denies the allegations of this paragraph.

2. Guardant [allegedly] knew or should have known of the Asserted Patents and that its conduct amounted to infringement

31. Guardant admits that it attempted to license certain technology from UW. Except as expressly admitted, Guardant denies the allegations of this paragraph.

32. Guardant admits that during prosecution of its patents, it has cited certain patents and publications to the United States Patent & Trademark Office ("PTO"), including certain Asserted Patents and published patent applications belonging to Plaintiffs. Guardant further admits that during prosecution of its patents, the PTO has referenced certain Asserted Patents belonging to Plaintiffs. Except as expressly admitted, Guardant denies the allegations of this paragraph.

33. Guardant admits that the Petition for *Inter Partes* Review in *Foundation Medicine, Inc. v. Guardant Health, Inc.*, PTAB IPR2019-00130, alleged that U.S. Patent No. 9,752,188 is purportedly prior art to Guardant's U.S. Patent 9,598,731. Guardant also admits that the Petition for *Inter Partes* Review in *Foundation Medicine, Inc. v. Guardant Health, Inc.*, PTAB IPR2019-00653, alleged that U.S. Patent No. 9,752,188 is purportedly prior art to Guardant's U.S. Patent 9,843,822. Guardant further admits that the Petition in PTAB IPR2019-00653 includes U.S. Provisional Application 61/613,413 as EX1012. Except as expressly admitted, Guardant denies the allegations of this paragraph.

34. Guardant admits that PCT application 2013/032665 was cited in EPO proceedings regarding Guardant's European Patent Nos. 3087204, 3378952, and 2893040. Except as expressly admitted, Guardant denies the allegations of this paragraph.

35. Guardant denies the allegations of this paragraph.

COUNT I

([Alleged] Direct Infringement of U.S. Patent No. 10,287,631)

36. Guardant incorporates its responses to Paragraphs 1-35 above, as if fully set forth herein.

37. Guardant admits that this paragraph recites claim 1 of the '631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

38. Guardant denies the allegations of this paragraph.

39. Guardant admits that this paragraph recites a portion of the preamble from claim 1 of the '631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

40. Guardant admits that this paragraph recites a portion of claim 1 of the '631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

41. Guardant admits that this paragraph recites a portion of claim 1 of the '631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

42. Guardant admits that this paragraph recites a portion of claim 1 of the '631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

43. Guardant admits that this paragraph recites a portion of claim 1 of the ’631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

44. Guardant admits that this paragraph recites a portion of claim 1 of the ’631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

45. Guardant admits that this paragraph recites a portion of claim 1 of the ’631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

46. Guardant admits that this paragraph recites a portion of claim 1 of the ’631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

47. Guardant admits that this paragraph recites a portion of claim 1 of the ’631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

48. Guardant admits that this paragraph recites a portion of claim 1 of the ’631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

49. Guardant admits that this paragraph recites a portion of claim 1 of the ’631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

50. Guardant admits that this paragraph recites a portion of claim 1 of the ’631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

51. Guardant admits that this paragraph recites a portion of claim 1 of the ’631 Patent as set forth in Exhibit A. Except as expressly admitted, Guardant denies the allegations of this paragraph.

52. Guardant denies the allegations of this paragraph.

53. Guardant denies the allegations of this paragraph.

54. Guardant denies the allegations of this paragraph.

55. Guardant denies the allegations of this paragraph.

56. Guardant denies the allegations of this paragraph.

COUNT II

([Alleged] Infringement of U.S. Patent No. 10,689,699)

57. Guardant incorporates its responses to Paragraphs 1-56 above, as if fully set forth herein.

58. Guardant admits that this paragraph recites claim 1 of the ’699 Patent as set forth in Exhibit B. Except as expressly admitted, Guardant denies the allegations of this paragraph.

59. Guardant denies the allegations of this paragraph.

60. Guardant admits that this paragraph recites a portion of claim 1 of the ’699 Patent as set forth in Exhibit B. Except as expressly admitted, Guardant denies the allegations of this paragraph.

61. Guardant admits that this paragraph recites a portion of claim 1 of the ’699 Patent as set forth in Exhibit B. Except as expressly admitted, Guardant denies the allegations of this paragraph.

62. Guardant admits that this paragraph recites a portion of claim 1 of the ’699 Patent as set forth in Exhibit B. Except as expressly admitted, Guardant denies the allegations of this paragraph.

63. Guardant admits that this paragraph recites a portion of claim 1 of the ’699 Patent as set forth in Exhibit B. Except as expressly admitted, Guardant denies the allegations of this paragraph.

64. Guardant admits that this paragraph recites a portion of claim 1 of the ’699 Patent as set forth in Exhibit B. Except as expressly admitted, Guardant denies the allegations of this paragraph.

65. Guardant admits that this paragraph recites a portion of claim 1 of the ’699 Patent as set forth in Exhibit B. Except as expressly admitted, Guardant denies the allegations of this paragraph.

66. Guardant admits that this paragraph recites a portion of claim 1 of the ’699 Patent as set forth in Exhibit B. Except as expressly admitted, Guardant denies the allegations of this paragraph.

67. Guardant admits that this paragraph recites a portion of claim 1 of the ’699 Patent as set forth in Exhibit B. Except as expressly admitted, Guardant denies the allegations of this paragraph.

68. Guardant admits that this paragraph recites a portion of claim 1 of the ’699 Patent as set forth in Exhibit B. Except as expressly admitted, Guardant denies the allegations of this paragraph.

69. Guardant admits that this paragraph recites a portion of claim 1 of the ’699 Patent as set forth in Exhibit B. Except as expressly admitted, Guardant denies the allegations of this paragraph.

70. Guardant admits that this paragraph recites a portion of claim 1 of the ’699 Patent as set forth in Exhibit B. Except as expressly admitted, Guardant denies the allegations of this paragraph.

71. Guardant denies the allegations of this paragraph.

72. Guardant denies the allegations of this paragraph.

73. Guardant denies the allegations of this paragraph.

74. Guardant denies the allegations of this paragraph.

75. Guardant denies the allegations of this paragraph.

COUNT III

([Alleged] Infringement of U.S. Patent No. 10,752,951)

76. Guardant incorporates its responses to Paragraphs 1-75 above, as if fully set forth herein.

77. Guardant admits that this paragraph recites claim 1 of the ’951 Patent as set forth in Exhibit C. Except as expressly admitted, Guardant denies the allegations of this paragraph.

78. Guardant denies the allegations of this paragraph.

79. Guardant admits that this paragraph recites a portion of claim 1 of the ’951 Patent as set forth in Exhibit C. Except as expressly admitted, Guardant denies the allegations of this paragraph.

80. Guardant admits that this paragraph recites a portion of claim 1 of the ’951 Patent as set forth in Exhibit C. Except as expressly admitted, Guardant denies the allegations of this paragraph.

81. Guardant admits that this paragraph recites a portion of claim 1 of the ’951 Patent as set forth in Exhibit C. Except as expressly admitted, Guardant denies the allegations of this paragraph.

82. Guardant admits that this paragraph recites a portion of claim 1 of the ’951 Patent as set forth in Exhibit C. Except as expressly admitted, Guardant denies the allegations of this paragraph.

83. Guardant admits that this paragraph recites a portion of claim 1 of the ’951 Patent as set forth in Exhibit C. Except as expressly admitted, Guardant denies the allegations of this paragraph.

84. Guardant admits that this paragraph recites a portion of claim 1 of the ’951 Patent as set forth in Exhibit C. Except as expressly admitted, Guardant denies the allegations of this paragraph.

85. Guardant admits that this paragraph recites a portion of claim 1 of the ’951 Patent as set forth in Exhibit C. Except as expressly admitted, Guardant denies the allegations of this paragraph.

86. Guardant admits that this paragraph recites a portion of claim 1 of the '951 Patent as set forth in Exhibit C. Except as expressly admitted, Guardant denies the allegations of this paragraph.

87. Guardant admits that this paragraph recites a portion of claim 1 of the '951 Patent as set forth in Exhibit C. Except as expressly admitted, Guardant denies the allegations of this paragraph.

88. Guardant admits that this paragraph recites a portion of claim 1 of the '951 Patent as set forth in Exhibit C. Except as expressly admitted, Guardant denies the allegations of this paragraph.

89. Guardant admits that this paragraph recites a portion of claim 1 of the '951 Patent as set forth in Exhibit C. Except as expressly admitted, Guardant denies the allegations of this paragraph.

90. Guardant denies the allegations of this paragraph.

91. Guardant denies the allegations of this paragraph.

92. Guardant denies the allegations of this paragraph.

93. Guardant denies the allegations of this paragraph.

94. Guardant denies the allegations of this paragraph.

COUNT IV

([Alleged] Infringement of U.S. Patent No. 10,760,127)

95. Guardant incorporates its responses to Paragraphs 1-94 above, as if fully set forth herein.

96. Guardant admits that this paragraph recites claim 1 of the '127 Patent as set forth in Exhibit D. Except as expressly admitted, Guardant denies the allegations of this paragraph.

97. Guardant denies the allegations of this paragraph.

98. Guardant admits that this paragraph recites a portion of claim 1 of the '127 Patent as set forth in Exhibit D. Except as expressly admitted, Guardant denies the allegations of this paragraph.

99. Guardant admits that this paragraph recites a portion of claim 1 of the '127 Patent as set forth in Exhibit D. Except as expressly admitted, Guardant denies the allegations of this paragraph.

100. Guardant admits that this paragraph recites a portion of claim 1 of the '127 Patent as set forth in Exhibit D. Except as expressly admitted, Guardant denies the allegations of this paragraph.

101. Guardant admits that this paragraph recites a portion of claim 1 of the '127 Patent as set forth in Exhibit D. Except as expressly admitted, Guardant denies the allegations of this paragraph.

102. Guardant admits that this paragraph recites a portion of claim 1 of the '127 Patent as set forth in Exhibit D. Except as expressly admitted, Guardant denies the allegations of this paragraph.

103. Guardant admits that this paragraph recites a portion of claim 1 of the '127 Patent as set forth in Exhibit D. Except as expressly admitted, Guardant denies the allegations of this paragraph.

104. Guardant admits that this paragraph recites a portion of claim 1 of the '127 Patent as set forth in Exhibit D. Except as expressly admitted, Guardant denies the allegations of this paragraph.

105. Guardant denies the allegations of this paragraph.

106. Guardant denies the allegations of this paragraph.

107. Guardant denies the allegations of this paragraph.

108. Guardant denies the allegations of this paragraph.

109. Guardant denies the allegations of this paragraph.

JURY DEMAND

110. This paragraph sets forth Plaintiffs' request for a jury trial to which no response is required.

PRAYER FOR RELIEF

These paragraphs set forth the statement of relief requested by Plaintiffs, to which no response is required. To the extent a response is required, Guardant denies that Plaintiffs are entitled to any relief whatsoever, either legal or equitable, from Guardant or this Court. Further, Guardant denies that this is an exceptional case such that an award of attorneys' fees to Plaintiffs or an award of enhanced damages.

GENERAL DENIAL AND NON-WAIVER

Each and every allegation in the Complaint not specifically admitted is denied, and the failure to deny a specific allegation, or assert a specific defense, shall not be deemed to be an admission of an allegation or a waiver of a defense.

AFFIRMATIVE DEFENSES

Without assuming any burden of proof that it would not otherwise bear, Guardant asserts the following separate and additional defenses, all of which are pled in the alternative, and none of which constitutes an admission that Guardant is in any way liable to Plaintiffs, that Plaintiffs have been or will be injured or damaged in any way, or that Plaintiffs are entitled to any relief whatsoever. In addition to the affirmative defenses described below, subject to its responses

above, Guardant specifically reserves all rights to allege additional affirmative defenses that become known through the course of discovery.

FIRST AFFIRMATIVE DEFENSE
(Failure to State a Claim)

Plaintiffs' claims are barred wholly or in part because Plaintiffs have failed to state any claim upon which relief can be granted with respect to Guardant with sufficient specificity and/or factual support to put Guardant on notice as to the claim being made and therefore failing to state a claim on which relief can be granted.

Plaintiffs fail to state a claim for willful infringement for failing to identify any pre-suit factual basis that Guardant had actual knowledge of or was willfully blind to any of the Asserted Patents or Guardant's alleged infringement, and for failing to plausibly allege facts sufficient to show that this is an egregious case of culpable behavior warranting enhanced damages.

SECOND AFFIRMATIVE DEFENSE
(No Infringement)

Guardant has not, does not, and will not directly or indirectly infringe any valid and enforceable claim of the '631 patent, the '699 patent, the '951 patent, and the '127 patent in any manner under 35 U.S.C. §§ 271(a), (b), (c), (f), and/or (g) literally or under the doctrine of equivalents, willfully or otherwise. Guardant has not performed any act and is not proposing to perform any act in violation of any rights validly belonging to Plaintiffs.

Certain claims of the '631 patent, the '699 patent, the '951 patent, and the '127 patent recite methods and require, for direct infringement, Guardant to perform every recited step of every asserted claim according to the limitations described in each asserted claim. Neither Guardant itself, nor any Guardant employee, performs all of the required steps of the asserted method claims while using any accused product.

THIRD AFFIRMATIVE DEFENSE
(Invalidity)

One or more claims of the '631 patent, the '699 patent, the '951 patent, and the '127 patent are invalid for failure to comply with one or more of the conditions of patentability set forth in Title 35 of the United States Code including, but not limited to, Sections 101, 102, 103 and/or 112.

FOURTH AFFIRMATIVE DEFENSE
(Inventorship/Derivation)

One or more claims of the '631 patent, the '699 patent, the '951 patent, and the '127 patent are invalid for failure to name the correct inventors pursuant to 35 U.S.C. § 116 and/or the claims were derived from the inventions of others.

FIFTH AFFIRMATIVE DEFENSE
(Standing)

One or both of the Plaintiffs lack standing to assert one or more of the Asserted Patents.

SIXTH AFFIRMATIVE DEFENSE
(Prosecution History Estoppel)

One or more claims of the '631 patent, the '699 patent, the '951 patent, and the '127 patent are and were limited by amendment, the prior art and/or by the statements made during prosecution before the PTO such that Plaintiffs are now estopped and/or otherwise precluded from maintaining that such claims of the '631 patent, the '699 patent, the '951 patent, and the '127 patent are of sufficient scope to cover the products accused of infringement in this case either literally or under the doctrine of equivalents.

SEVENTH AFFIRMATIVE DEFENSE
(Waiver, Estoppel)

Plaintiffs' Complaint is barred by the doctrines of waiver and/or estoppel due to Plaintiffs' actions before, during, and after prosecution of the Asserted Patents, and Plaintiffs' conduct before filing this action and in filing of this action.

EIGHTH AFFIRMATIVE DEFENSE
(Unclean Hands)

Plaintiffs' Complaint is barred by the doctrine of unclean hands as a result of the acts, conduct, and omissions of Plaintiffs or others that are attributable to Plaintiffs before, during, and after prosecution of the Asserted Patent and Plaintiffs' conduct before filing this action and in filing of this action.

NINTH AFFIRMATIVE DEFENSE
(Notice, Damages and Cost)

Plaintiffs' claims for damages are statutorily limited under 35 U.S.C. § 286 and 287.

Plaintiffs are barred from recovering costs in connection with this action under 35 U.S.C. § 288.

Plaintiffs' claims for relief are limited by the doctrines of full compensation, exhaustion, first sale and Plaintiffs are not entitled to double recovery.

TENTH AFFIRMATIVE DEFENSE
(No Right to Injunctive Relief)

Plaintiffs are not entitled to preliminary or permanent injunctive relief, *inter alia*, because any injury to Plaintiffs is not immediate nor irreparable, and Plaintiffs have an adequate monetary remedy for any claim that they can prove, subject to any applicable defense and/or set-off.

ELEVENTH AFFIRMATIVE DEFENSE
(No Pre-Suit Damages, Failure to Mark)

Plaintiffs' claims for pre-suit damages are limited by any noncompliance with the marking and actual notice requirements set forth in 35 U.S.C. § 287(a).

TWELFTH AFFIRMATIVE DEFENSE
(Failure to Comply With Patentability Requirements of 35 U.S.C. § 135)

One or more claims of the '699 patent are invalid for failure to comply with one or more of the conditions of patentability set forth in 35 U.S.C. § 135.

RESERVATION OF AFFIRMATIVE DEFENSES

As Guardant's investigation is ongoing and discovery has not yet been completed, Guardant is without complete information regarding the existence or non-existence of other facts or acts that would constitute a defense to the purported causes of action in Plaintiffs' Complaint. Accordingly, Guardant reserves the right to assert any other defenses that discovery may reveal.

COUNTERCLAIMS

Guardant submits the following counterclaims against TwinStrand Biosciences, Inc. and University of Washington:

PARTIES

111. Counterclaim-Plaintiff Guardant is a corporation organized and existing under the laws of the State of Delaware and having its principal place of business at 505 Penobscot Dr., Redwood City, CA, 94063.

112. As alleged in Plaintiffs' Complaint, TwinStrand Biosciences, Inc. ("TwinStrand") is a corporation organized and existing under the laws of the State of Delaware and having its principal place of business at 3131 Elliott Ave., Suite 750, Seattle, WA, 98121.

113. As alleged in Plaintiffs' Complaint, University of Washington ("UW") is a public institution of higher education and an agency of the State of Washington and its principal place of business is in Seattle, Washington.

114. Guardant's Counterclaims arise under the patent laws of the United States, Title 35 of the United States Code. This Court has subject matter jurisdiction over Guardant's Counterclaims under 28 U.S.C. §§ 1331 and 1338.

115. Venue in this case is proper in this District because TwinStrand is a Delaware corporation, and by virtue of TwinStrand's filing of the Complaint in this District, which gave rise to these Counterclaims. Plaintiffs contend in their Complaint that venue is proper in this District.

FACTUAL BACKGROUND

116. Guardant is one of the leading precision oncology companies dedicated to helping conquer cancer with data obtained through its proprietary blood tests. Guardant was founded in 2012 by Helmy Eltoukhy, PhD and AmirAli Talasaz, PhD, pioneers in DNA sequencing and

cancer diagnostics. Since its inception, Guardant has focused its expertise on the development of liquid biopsy cancer assays, including Guardant360® CDx—the first FDA-approved liquid biopsy test. To date, over 150,000 patients have benefitted from Guardant's tests. Guardant's expansive portfolio of IP includes nearly 500 patents and applications worldwide.

117. A liquid biopsy is the sampling and analysis of non-solid biological tissue, as opposed to a conventional biopsy which involves the removal of tissue for examination and is often conducted surgically or with a needle. Guardant was one of the first companies to commercialize a comprehensive liquid biopsy assay to identify genomic biomarkers, which are biological molecules found in blood or tissues that may be indicative of a disease such as cancer using "cell-free DNA," or "cfDNA," taken from simple, non-invasive blood draws. Guardant's liquid biopsy technology enables patients, including those who are ineligible for traditional tissue biopsies, to obtain detailed genomic information about their cancer. This allows physicians to select the best treatment, while also helping to detect cancer early.

118. For example, the Guardant360® CDx test is a liquid biopsy assay that provides clinically actionable sequence information from cfDNA molecules obtained from a routine blood draw taken from cancer patients. The assay sequences a panel of genes commonly mutated in cancer and detects genetic aberrations such as single nucleotide variants, indels (insertions or deletions of nucleotides), gene fusions, and copy number variants. Guardant360® CDx provides highly sensitive and accurate detection of genetic aberrations using innovative laboratory preparation techniques for cfDNA in concert with bioinformatic algorithms specific for single nucleotide variants, indels, gene fusions, and copy number variants.

119. In addition to the FDA-approved Guardant360® CDx test, Guardant currently offers six other tests: the Guardant360® laboratory developed test (LDT), Guardant360

Response™, Guardant360 TissueNext™, GuardantOMNI®, Guardant Reveal™, and LUNAR-2[2] tests. These tests span the cancer care continuum, including early cancer screening, treatment selection, and residual disease and recurrence monitoring.

120. Guardant's liquid biopsy technology has several additional advantages when compared to traditional tissue biopsies. Traditional tumor-based genotyping assays are limited in the number of genes interrogated, require invasive biopsies, and often take upwards of 15 days before results are generated. Guardant's liquid biopsies are less painful and do not require hospital services. This technology is also less expensive and generates results in a shorter period of time. Further, since cfDNA samples are heterogenous samples, they allow for the detection of mutations that may be missed by a tissue biopsy sample.

121. As alleged in Plaintiffs' Complaint, TwinStrand markets and sells kits and services under its TwinStrand Duplex Sequencing™ technology platform. TwinStrand's "platform includes access to a secure cloud-based bioinformatics pipeline." (https://TwinStrandbio.com/technology/).

122. As alleged in Plaintiffs' Complaint, Guardant is informed and believes and thereon alleges that TwinStrand exclusively licenses the Asserted TwinStrand Patents from UW.

123. As alleged in Plaintiffs' Complaint, Guardant is informed and believes and thereon alleges that UW is the owner of the Asserted TwinStrand Patents.

124. TwinStrand was founded in 2015. Since its inception, TwinStrand has targeted Guardant and its innovative technologies with threats of patent infringement and more. Over the course of its interactions, TwinStrand has continuously exaggerated the scope of its patent portfolio and thwarted Guardant's business opportunities with other licensors. TwinStrand has

[2] Guardant's LUNAR-2 is currently in clinical trials.

also shamelessly attacked Guardant's patent portfolio by copying Guardant's own patent claims into TwinStrand patent applications—misrepresenting TwinStrand employees and founders as the true inventors of the claimed inventions—and licensing the same to Guardant's competitors in an attempt to usurp Guardant's patent rights and undermine Guardant's success in helping to establish the market for liquid biopsies. While doing so, TwinStrand has refused to acknowledge that its claims of infringement lack merit or, alternatively, consider reasonable royalty terms to resolve the ongoing dispute.

125. Further, TwinStrand's egregious behavior of copying patent claims of other companies into its own patent applications is not just limited to Guardant. TwinStrand has taken the same approach with other competitors—attempting to secure additional patents to which they are not entitled. Not only has TwinStrand copied Guardant and others' patent claims, it has also failed to dutifully inform the PTO of its claim copying in violation of 35 U.S.C. §§ 101, 115, and/or 135(b)(1).

126. In early 2015, and well into the development of its own proprietary liquid biopsy technology, Guardant became aware of patent applications filed by UW concerning a different method of performing sequence analysis. These applications later issued as U.S. Patent Nos. 9,752,188 (the "'188 Patent") and 10,287,631 (the "'631 Patent") (collectively the "UW Licensed Patents"). Neither of the UW Licensed Patents include claims directed to liquid biopsy, and none of the claims recite "cfDNA" or equivalents thereof. TwinStrand did not pursue claims to cfDNA until 2019. Although Guardant had developed its own, proprietary technology for analysis of cfDNA, it was interested in exploring a potential collaboration and/or licensing arrangement with UW concerning the alternative methodology.

127. While these discussions and negotiations were ongoing, the named inventors of the TwinStrand Patents formed TwinStrand and immediately entered into negotiations with UW to acquire an exclusive license to the very patent applications that Guardant was discussing. Ultimately, UW gave TwinStrand the exclusive rights to the patent applications and UW informed Guardant that any further discussions regarding the patent applications should be had with TwinStrand.

128. As UW instructed, Guardant commenced discussions with TwinStrand, but almost immediately those discussions grew aggressive and unproductive. In contrast to the negotiations that had taken place with UW, Guardant found itself being accused of practicing the methodology set forth in the patent applications and TwinStrand sought to extract a large royalty-based payment on Guardant products that did not use the subject methodology.

129. Through a series of long, drawn-out negotiations and discussions, TwinStrand attempted to coax Guardant into disclosing its own proprietary technology under the auspices of negotiating a license. Despite numerous meetings and discussions, TwinStrand maintained its assertion that Guardant infringed the UW Licensed Patents. However, Guardant and TwinStrand were unable to reach a licensing agreement because TwinStrand's demands for royalties proved untenable.

THE GUARDANT PATENTS-IN-SUIT

130. On October 13, 2020, the United States Patent and Trademark Office lawfully issued U.S. Patent No. 10,801,063 (the "'063 patent"), entitled "Methods and systems for detecting genetic variants." A true and correct copy of the patent is attached hereto as Exhibit A.

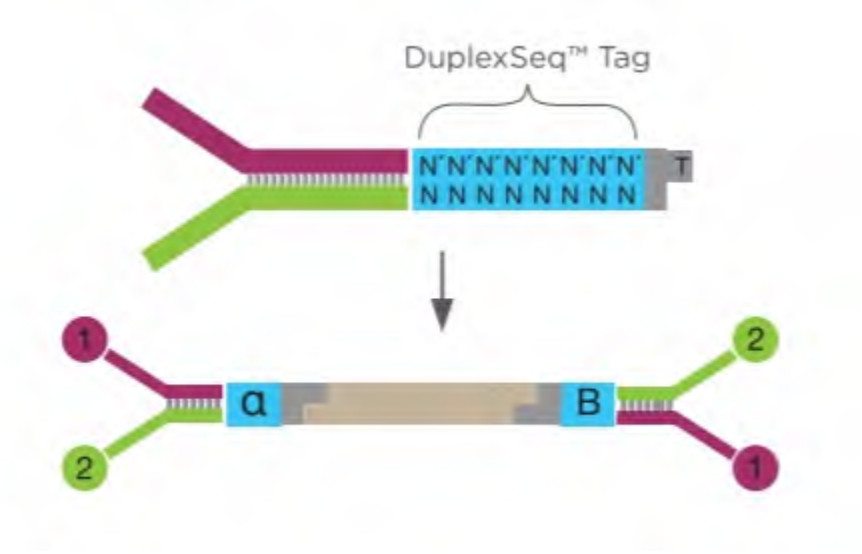
131. On January 12, 2021, the United States Patent and Trademark Office lawfully issued U.S. Patent No. 10,889,858 (the "'858 patent"), entitled "Methods and systems for detecting genetic variants." A true and correct copy of the patent is attached hereto as Exhibit B.

132. On September 14, 2021, the United States Patent and Trademark Office lawfully issued U.S. Patent No. 11,118,221 (the "'221 patent"), entitled "Methods and systems for detecting genetic variants." A true and correct copy of the patent is attached hereto as Exhibit C.

133. On October 19, 2021, the United States Patent and Trademark Office lawfully issued U.S. Patent No. 11,149,306 (the "'306 patent"), entitled "Methods and systems for detecting genetic variants." A true and correct copy of the patent is attached hereto as Exhibit D.

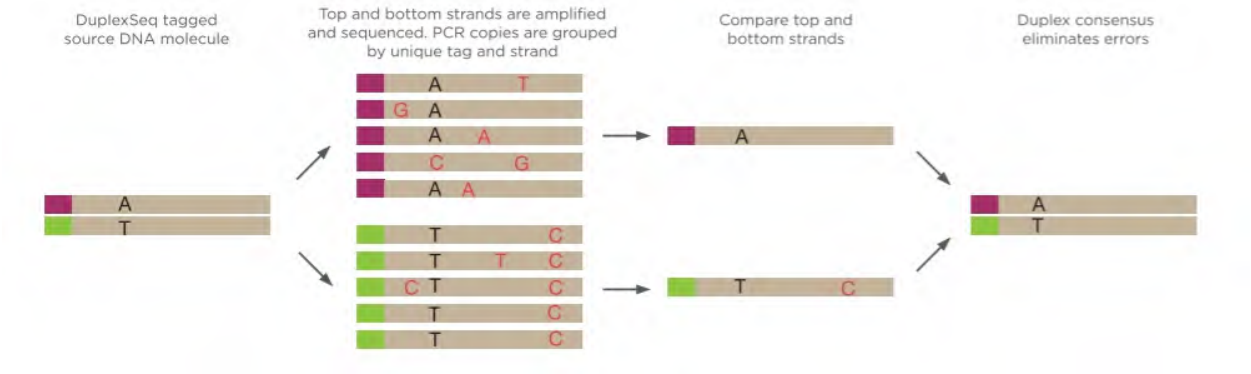
TWINSTRAND'S INFRINGEMENT OF THE GUARDANT PATENTS-IN-SUIT

134. TwinStrand manufactures and sells kits that utilize TwinStrand Duplex Sequencing™. (https://TwinStrandbio.com/technology/). According to public information, TwinStrand Duplex Sequencing separately and uniquely labels complementary strands of individual DNA molecules:



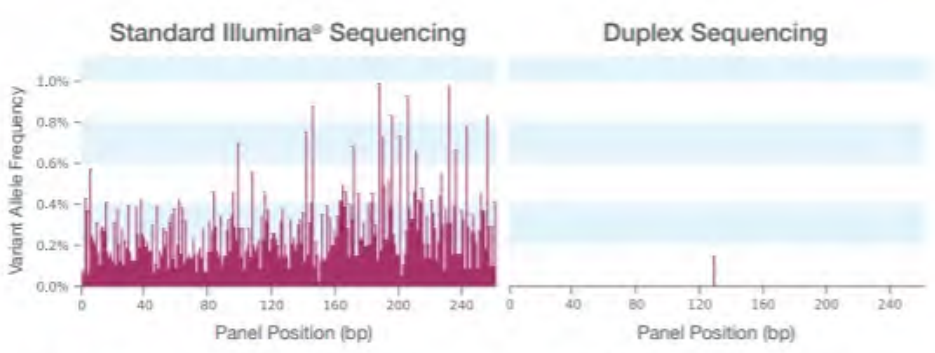
(https://TwinStrandbio.com/wp-content/uploads/TwinStrand-Technology-Brochure-10JUL2020OL-WEB.pdf)

135. TwinStrand's tagging method allows one strand of each molecule to be differentiated from the other strand of that molecule, which allows for the comparison of complementary strands of the original target molecule so that true mutations, which occur on both strands, can be distinguished from sequencing errors, which occur on only one strand:



(https://TwinStrandbio.com/wp-content/uploads/TwinStrand-Technology-Brochure-10JUL2020 OL-WEB.pdf)

136. According to public information, TwinStrand Duplex Sequencing™ removes the noise found in standard NGS sequencing, where every position in the sequenced gene appears mutated in 0.1 to 1% of the molecules sequenced, allowing for the identification of low-frequency true mutations:

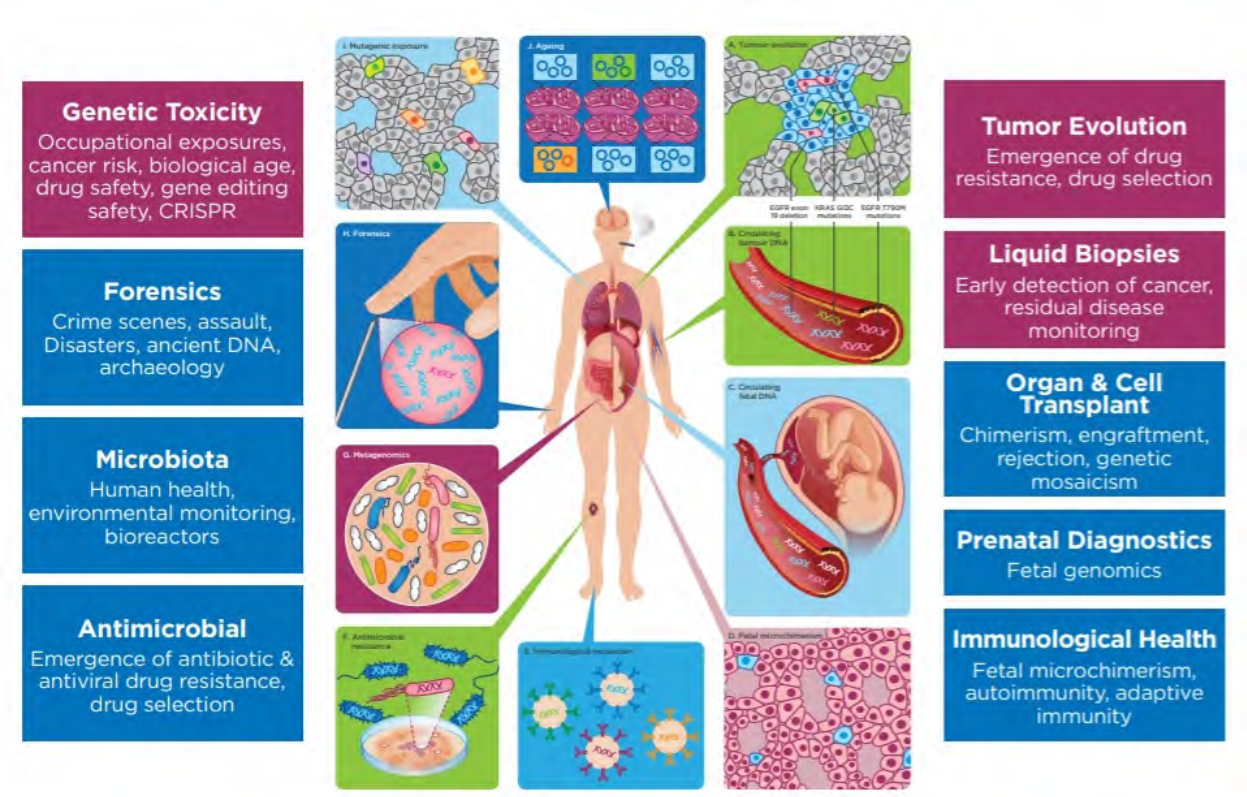


The same gene sequenced by standard Illumina® Sequencing (left) and with Duplex Sequencing (right). With conventional methods, every position in the gene appears mutated in 0.1-1% of molecules sequenced. Duplex Sequencing eliminates the background noise, revealing the previously hidden true mutation.

(https://TwinStrandbio.com/wp-content/uploads/TwinStrand-Technology-Brochure-10JUL2020OL-WEB.pdf)

137. TwinStrand also offers a bioinformatics solution "[p]urpose-built for DuplexSeq data" which "[p]rocesses raw read files to produce duplex error-corrected sequences, variant calls, and assay quality metrics." (https://TwinStrandbio.com/technology/).

138. Based on TwinStrand's promotional materials, TwinStrand Duplex Sequencing™ may be used in liquid biopsies:



(https://TwinStrandbio.com/wp-content/uploads/TwinStrand-Technology-Brochure-10JUL2020OL-WEB.pdf)

139. By performing TwinStrand Duplex Sequencing™ and utilizing its bioinformatics solution, and by instructing its customers to do the same, TwinStrand has infringed at least claim 15 of the '063 patent, claim 1 of the '858 patent, claim 1 of the '212 patent, and claim 17 of the '306 patent. *See, e.g.*, Complaint at ¶26 ("TwinStrand provides services for nucleic acid analysis using the patented Duplex Sequencing methods and provides customers with Duplex Sequencing kits. These kits include a DNA library prep kit containing the reagents, adapters, and other components necessary to practice its Duplex Sequencing process. TwinStrand also provides access to bioinformatics software to process raw sequence read files and produce error-corrected sequences according to the patented processes.").

140. In addition to infringing Guardant's patents, TwinStrand also has engaged in a predatory practice of copying Guardant's and other third parties' intellectual property and claiming it as its own. Specifically, TwinStrand has copied patent claims from Guardant's and other third parties' patents and patent applications into its own patent applications without permission and in violation of 35 U.S.C. §§ 101, 115, and/or 135(b)(1).

141. For example, TwinStrand substantially copied patent claims from Guardant's U.S. Patent No. 9,834,822 ("'822 Patent") into the application that issued as TwinStrand's U.S. Patent No. 10,689,699, which it is now asserting against Guardant.

142. Guardant's '822 Patent titled "Systems and methods to detect rare mutations and copy number variation" issued on December 5, 2017 and names AmirAli Talasaz as inventor of the claimed subject matter.

143. TwinStrand did not notify the Examiner at any point during the prosecution of the '699 patent that the claims were copied.

144. TwinStrand copied the claims of Guardant's '822 patent more than one year after the '822 patent issued, thereby violating 35 U.S.C. § 135(b)(1).

145. TwinStrand also copied patent claims from Guardant Patent Application WO2015/100427 into TwinStrand's Application No. 16/514,931 (which issued as U.S. Patent No. 10,752,951). The '951 Patent is now being asserted against Guardant.

146. The named inventors of the subject matter of Guardant's Patent Application WO2015/100427 are Helmy Eltoukhy and AmirAli Talasaz.

147. TwinStrand did not notify the Examiner at any point during the prosecution of the '951 Patent that the claims were copied.

148. TwinStrand has also copied claims from Illumina Inc. and Grail patents and applications.

COUNT I
(Direct and Indirect Infringement of U.S. Patent No. 10,801,063)

149. Guardant re-alleges and incorporated by reference Paragraphs 1-148 above, as if fully set forth herein.

150. The ’063 patent is directed to a method of determining genetic variants. Claim 15 of the ’063 patent recites:

> 15. A method for classifying unique sequencing reads generated from sequencing reads derived from double-stranded cell-free deoxyribonucleic acid (cfDNA) molecules from a bodily fluid sample of a human subject, the method comprising:
>
> (a) tagging a population of double-stranded cfDNA molecules from the bodily fluid sample with more than a 10× molar excess of adapters comprising molecular barcodes, relative to the double-stranded cfDNA molecules in the population, to generate tagged parent polynucleotides,
>
> wherein at least 20% of the cfDNA molecules are ligated with the adapters comprising the molecular barcodes at both ends of a molecule of the double-stranded cfDNA molecules;
>
> (b) amplifying a plurality of the tagged parent polynucleotides to produce progeny polynucleotides;
>
> (c) sequencing a plurality of the progeny polynucleotides to produce a set of sequencing reads;
>
> (d) mapping a plurality of sequencing reads from the set of sequencing reads to a reference sequence;

(e) determining unique sequencing reads from the set of mapped sequencing reads based at least on the molecular barcode sequences, wherein a unique sequencing read from among the unique sequencing reads is representative of a tagged parent polynucleotide from among the tagged parent polynucleotides; and

(f) classifying one or more of the unique sequencing reads as either (1) paired sequences generated from sequencing reads representing a Watson strand and a Crick strand of a tagged parent polynucleotide or (2) unpaired sequences generated from sequencing reads representing only one of either a Watson strand or a Crick strand of a tagged parent polynucleotide.

151. TwinStrand has willfully infringed and continues to infringe at least claim 15 of the '063 patent, literally or under the doctrine of equivalents, by performing the methods of the '063 patent in the United States. It is expected that discovery likely will reveal additional evidentiary support that TwinStrand performs the above limitations of the '063 patent and Guardant reserves the right to supplement the evidence cited below as additional evidentiary support is obtained.

152. TwinStrand practices the preamble of claim 15 of the '063 patent, which provides "[a] method for classifying unique sequencing reads generated from sequencing reads derived from double-stranded cell-free deoxyribonucleic acid (cfDNA) molecules from a bodily fluid sample of a human subject." For example, TwinStrand's Technology Brochure describes the sequencing methods used by TwinStrand to reveal mutations in double-stranded DNA, including cfDNA. *See* "TwinStrand Duplex Sequencing™" ("TS Brochure") at 1-2 (Exhibit E). Applications described in the TS Brochure include "liquid biopsies" for early detection of cancer and residual disease monitoring. *Id.* at 2.

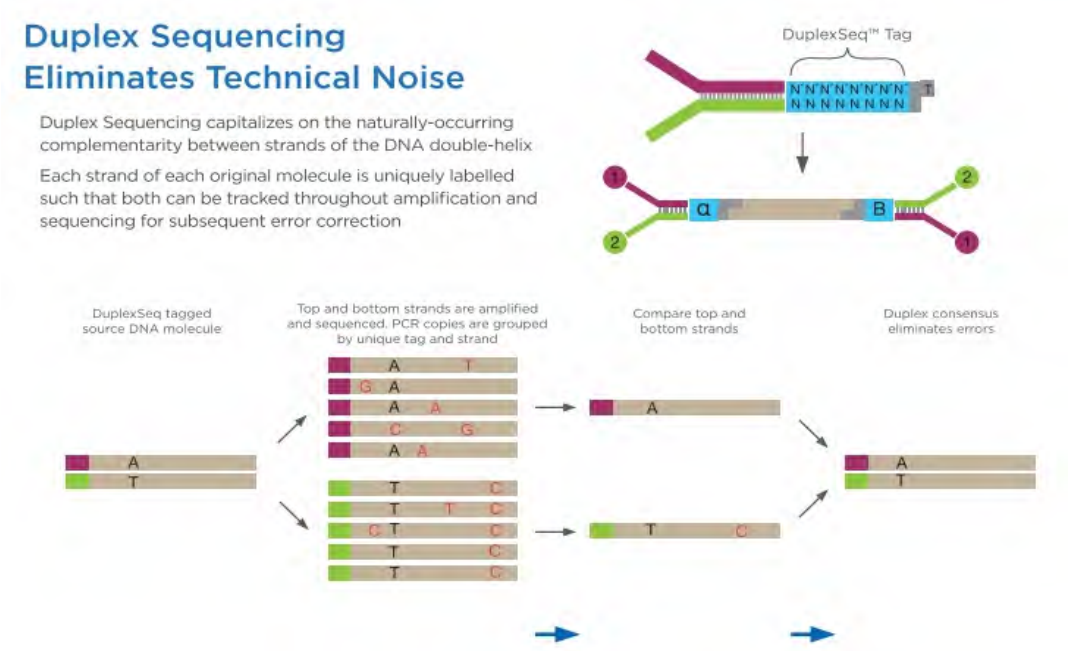
153. Dr. Jesse Salk, CEO and Founder of TwinStrand, has made several public statements regarding TwinStrand's sequencing methods including that "TwinStrand has developed [] a high-accuracy NGS technique called duplex sequencing, which uses both strands of a DNA molecule in the sequencing process to check for errors." *See* "Sequencing Stars Urge Exploration of New Dimensions," Genetic Engineering and Biotechnology News (April 1, 2020) ("GEB April 2020") at 5 (Exhibit F); "Liquid Biopsies Generate Ample Insights from Sparse Clues;" Genetic Engineering and Biotechnology News (August 2, 2020) ("GEB August 2020") at 2 (Exhibit G). Dr. Salk has also stated that TwinStrand is "offering duplex sequencing kits and bioinformatics analysis" to consumers. *See* GEB August 2020 at 3.

154. TwinStrand has also posted several publications on its website describing the methods used by TwinStrand. For example, TwinStrand describes an ultra-sensitive sequencing method for cancer detection in cfDNA. *See* Salk, et al., "Ultra-Sensitive TP53 Sequencing for Cancer Detection Reveals Progressive Clonal Selection in Normal Tissues over a Century of Human Lifespan," Cell Reports, 28(1): 132-144 (July 2, 2019) ("Salk 2019") (Exhibit H). This article states "[t]o explore the abundance of *TP53* mutations in liquid biopsies of clinical interest, we analyzed plasma-derived cell-free DNA (cfDNA) and peritoneal fluid from the 46-year old woman" and "[t]he fact that we identified one pathogenic mutation commonly found in cancers within cfDNA from a healthy woman (Figure S10; Table 5) raises important concerns over specificity in cancer-screening studies based on mutation detection in plasma." *Id.* at 141, FIGS. S10(A), S10(B) and S6.

155. Further, in its Complaint, TwinStrand itself acknowledges that it practices the methods of the Asserted Patents, which include claims directed to methods for classifying unique

sequencing reads generated from sequencing reads derived from double-stranded polynucleotide molecules. *See* e.g., Compl. at ¶¶ 13, 18, 20, 22; '951 patent at claims 1, 11, and 18.

156. TwinStrand also practices the "tagging" step of claim 15 of the '063 patent. This element recites "(a) tagging a population of double-stranded cfDNA molecules from the bodily fluid sample with more than a 10× molar excess of adapters comprising molecular barcodes, relative to the double-stranded cfDNA molecules in the population, to generate tagged parent polynucleotides, wherein at least 20% of the cfDNA molecules are ligated with the adapters comprising the molecular barcodes at both ends of a molecule of the double-stranded cfDNA molecules." For example, the TS Brochure shows tagging of DNA molecules:



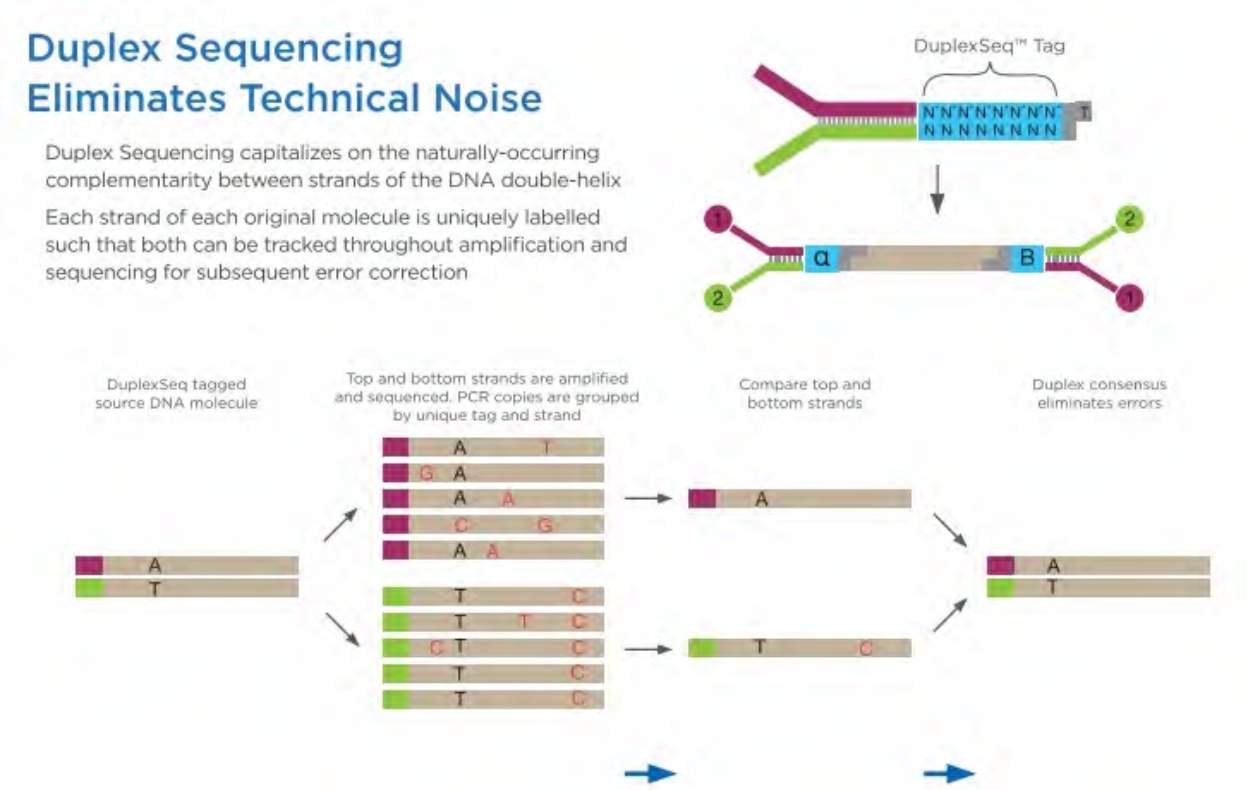
(TS Brochure at 1).

157. Further, publications posted on TwinStrand's website show TwinStrand uses tagging of a population of double-stranded cfDNA molecules from the bodily fluid sample with more than a 10× molar excess of adapters comprising molecular barcodes relative to the double-stranded cfDNA molecules in the population wherein at least 20% of the cfDNA molecules are

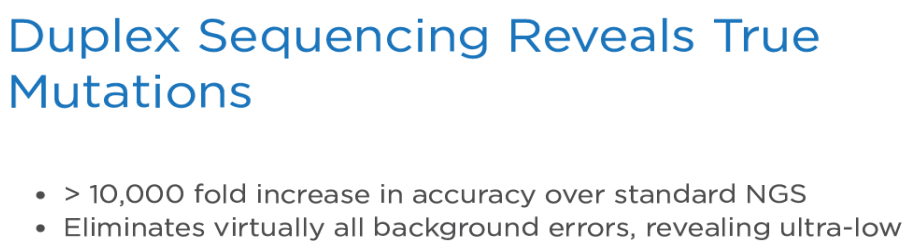
ligated with adaptors. *See* TS Brochure at 1; Salk 2019 at e2. For example, Kennedy, et. al. describes methods used by TwinStrand, including the disclosure that during tagging a "20-fold molar excess of adapter relative to DNA" is used. *See* Kennedy, et. al., "Detecting ultralow-frequency mutations by Duplex Sequencing." *Nature Protocols* 9(11): 2586-2606 (October 9, 2014) ("Kennedy 2014") at Abstract, FIG. 1, 2586, 2591, 2597 (Exhibit I). Kennedy further instructs in its methods during ligation to "calculate the amount of DS adaptors…that will be needed for a 20:1 molar excess relative to the total DNA concentration in the final ligation reaction." *Id.* at 2597; *see also* Compl. at ¶¶ 20, 22; '951 patent at claim 1; 16:38-46; FIG. 1.

158. TwinStrand also practices the "amplifying" step of claim 15 of the '063 patent. This element recites "(b) amplifying a plurality of the tagged parent polynucleotides to produce progeny polynucleotides." For example, TwinStrand advertises that "each strand of original molecule is uniquely labeled such that both ends can be tracked throughout amplification and sequencing for subsequent error detection." TS Brochure at 1; *see also* TwinStrand's website at 7 (Exhibit J); Kennedy 2014 at Abstract ("Individually labeled strands are then PCR-amplified, creating sequence 'families' that share a common tag sequence derived from the two original complementary strands."); Salk 2019 at e2 ("After initial amplification, 120 bp biotinylated oligonucleotide probes (Integrated DNA Technologies, Coralville, Iowa) were used to capture the coding region of TP53"); Compl. at ¶ 20; '951 patent at FIG. 1. Amplification includes PCR amplification.

159. TwinStrand also practices the "sequencing" step of claim 15 of the '063 patent. This element recites "(c) sequencing a plurality of the progeny polynucleotides to produce a set of sequencing reads." For example, TwinStrand's method is a sequencing method that produces a set of sequencing reads:



TS Brochure at 1.



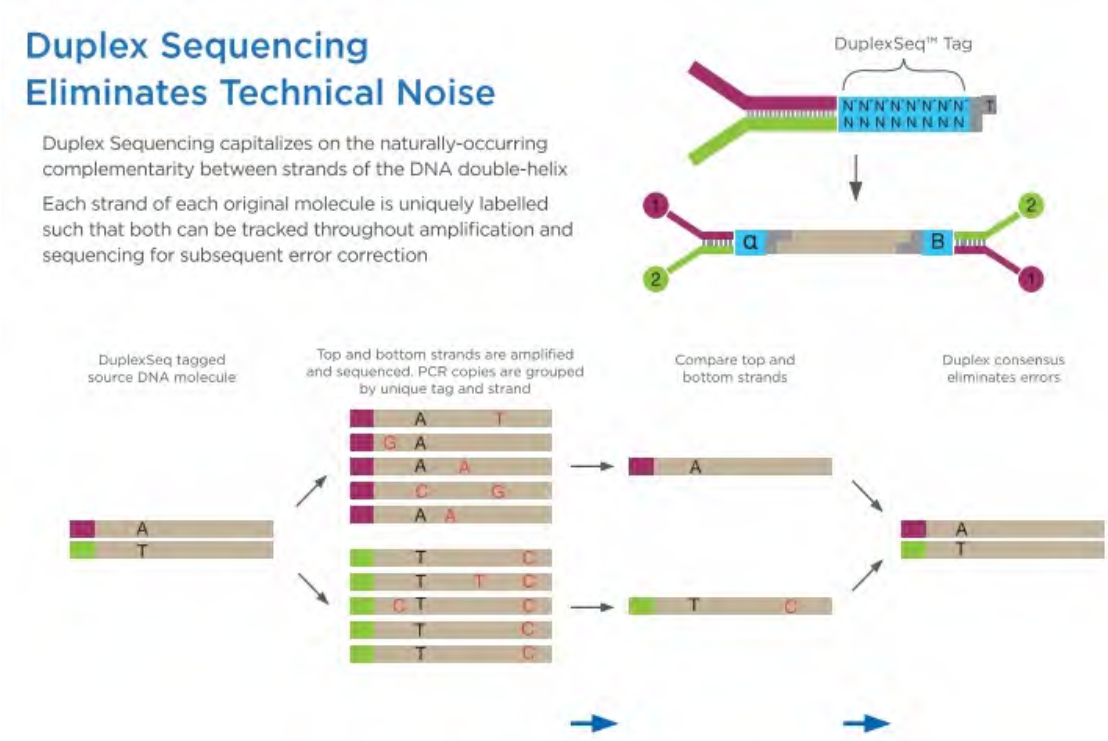
TwinStrand Website at 6; *see also id.* at 4, 13; Kennedy 2014 at FIG. 1(C), 2587; Compl. at 26 ("TwinStrand provides services for nucleic acid analysis using the patented Duplex Sequencing methods and provides customers with Duplex Sequencing kits. These kits include a DNA library prep kit containing the reagents, adapters, and other components necessary to practice its Duplex Sequencing process. TwinStrand also provides access to bioinformatics software to process raw

sequence read files and produce error-corrected sequences according to the patented processes"); '951 patent at claim 1; FIG. 1.

160. TwinStrand also practices the "mapping" step of claim 15 of the '063 patent. This element recites "(d) mapping a plurality of sequencing reads from the set of sequencing reads to a reference sequence." For example, Kennedy 2014 describes "a comprehensive computational pipeline that we use to process and analyze our data (**Fig. 2).** The computational workflow for DS uses a number of standard software packages to process the sequencing data; these include the Burrows-Wheeler Aligner (BWA)37, SAMtools38, Picard and the Genome Analysis Toolkit (GATK)39,40, as well as several custom Python scripts. The computational workflow is broken down into three major steps: (i) Tag parsing and initial alignment; (ii) SSCS assembly; and (iii) DCS assembly. The latest version of the DS software package can be downloaded from https://github.com/loeblab/Duplex-Sequencing." Kennedy 2014 at 2589. "The reads are then aligned to the reference genome using BWA37. After alignment, reads sharing the same tag sequence and genomic coordinates are identified and grouped to form 'tag families' with a python script called 'ConsensusMaker.py'. *Id.* at 2589, 2601; *see also* '951 patent at claim 1 and 13; 28:43-45 ("Initial nucleotides represent raw reads that contain the expected fixed adapter sequence following 12 degenerate nucleotides and map to the reference genome.").

161. TwinStrand also practices the "determining unique sequencing reads" step of claim 15 of the '063 patent. This element recites "determining unique sequencing reads from the set of mapped sequencing reads based at least on the molecular barcode sequences, wherein a unique sequencing read from among the unique sequencing reads is representative of a tagged parent polynucleotide from among the tagged parent polynucleotides." For example, TwinStrand's

method produces duplex consensus sequences resulting from the comparison of sequencing reads from tagged parent polynucleotides:



TS Brochure at 1; *see also* Kennedy 2014 at Abstract ("Individually labeled strands are then PCR-amplified, creating sequence 'families' that share a common tag sequence derived from the two original complementary strands. Mutations are scored only if the variant is present in the PCR families arising from both of the two DNA strands. Here we provide a detailed protocol for efficient DS adapter synthesis, library preparation and target enrichment, as well as an overview of the data analysis workflow."), FIG. 1; Salk 2019 at e2 ("Sequencing reads were aligned to hg19 then reads sharing a common molecular tag in both strand orientations were grouped and assembled into an error-corrected Duplex Consensus Sequence.").

162. TwinStrand also practices the "classifying unique sequencing reads" step of claim 15 of the '063 patent. This element recites "classifying one or more of the unique sequencing reads as either (1) paired sequences generated from sequencing reads representing a Watson strand

and a Crick strand of a tagged parent polynucleotide or (2) unpaired sequences generated from sequencing reads representing only one of either a Watson strand or a Crick strand of a tagged parent polynucleotide." For example, TwinStrand's method involves the step of tagging both the Watson and Crick strand of the parent polynucleotides. *See* ¶¶ 156-157 above.

163. Further, in its Complaint, TwinStrand admits that it practices the claims of the Asserted Patents, including claims 1 and 27 of the '951 patent. Claim 1 of the '951 patent recites "(d) sorting sequence reads into paired sequence reads and unpaired sequence reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from an original parent polynucleotide molecule in said set, and (ii) each unpaired read represents a first tagged strand having no second differently tagged complementary strand derived from an original parent polynucleotide molecule represented among said sequence reads in said set of sequence reads; and (e) quantifying at least two of (i) said paired sequence reads, (ii) said unpaired sequence reads, (iii) read depth of said paired sequence reads and (iv) read depth of said unpaired sequence reads." Claim 27 recites "The method of claim 1, further comprising quantifying a total number of original double-stranded polynucleotide molecules or genome-equivalents in the sample based at least in part on said quantification of said paired sequence reads and said unpaired sequence reads."

164. By making, selling, and offering to sell kits that utilize TwinStrand Duplex Sequencing™ with the express purpose of encouraging, aiding and/or causing such customers to practice at least claim 15 of the '063 patent, TwinStrand willfully indirectly infringes at least claim 15 of the '063 patent through inducement and contributory infringement under 35 U.S.C. § 271 *et seq.* TwinStrand's indirect infringement has and continues to be willful.

165. Upon information and belief, TwinStrand has been aware of the existence of the '063 patent since at least October 13, 2020 and/or has willfully blinded itself as to the infringing nature of its activity since the issuance of the '063 patent.

166. By its actions, TwinStrand's infringement of the '063 patent has irreparably harmed Guardant. Unless TwinStrand's infringing acts are enjoined by this Court, Guardant will continue to suffer additional irreparable injury. Guardant has no adequate remedy at law.

167. By its actions, TwinStrand's infringement of the '063 patent has damaged and continues to damage Guardant in an amount yet to be determined of at least a reasonable royalty and/or lost profits that Guardant would have made but for TwinStrand's infringing acts.

COUNT II
(Direct and Indirect Infringement of U.S. Patent No. 10,889,858)

168. Guardant re-alleges and incorporated by reference Paragraphs 1-167 above, as if fully set forth herein.

169. The '858 patent is directed to a method of determining genetic variants. Claim 1 of the patent recites:

> 1. A method for analyzing sequencing reads of double-stranded cell-free deoxyribonucleic acid (cfDNA) molecules from a sample of a subject, comprising:
>
> (a) tagging a plurality of double-stranded cfDNA molecules from a population of double-stranded cfDNA molecules from the sample with a set of library adaptors comprising a plurality of molecular barcodes to generate tagged parent polynucleotides,
>
> wherein the tagging comprises ligating a plurality of library adaptors from the set of library adaptors to the plurality of double-stranded cfDNA molecules from the population using more than a 10× molar excess of

library adaptors as compared to the double-stranded cfDNA molecules of the population;

wherein the tagging produces at least 20% of the double-stranded cfDNA molecules of the populations having library adaptors ligated to both ends of a molecule of the double-stranded cfDNA molecules;

(b) amplifying a plurality of the tagged parent polynucleotides to produce progeny polynucleotides;

(c) sequencing a plurality of the progeny polynucleotides to produce a set of sequencing reads; and

(d) determining, based at least on sequence information from the molecular barcodes, individual double-stranded cfDNA molecules from among the tagged parent polynucleotides for which either (1) both a Watson strand and a Crick strand of the individual double-stranded cfDNA molecule are detected or (2) only one of a Watson strand or a Crick strand of the individual double-stranded cfDNA molecule is detected from a plurality of sequencing reads from the set of sequencing reads.

170. TwinStrand has willfully infringed and continues to infringe at least claim 1 of the ’858 patent, literally or under the doctrine of equivalents, by performing the methods of the ’858 patent in the United States. It is expected that discovery likely will reveal additional evidentiary support that TwinStrand performs the above limitations of the ’858 patent and Guardant reserves the right to supplement the evidence cited below as additional evidentiary support is obtained.

171. TwinStrand practices the preamble of claim 1 of the ’858 patent, which provides “[a] method for analyzing sequencing reads of double-stranded cell-free deoxyribonucleic acid

(cfDNA) molecules from a sample of a subject." For example, TwinStrand's Technology Brochure describes the sequencing methods used by TwinStrand to reveal mutations in double-stranded DNA, including cfDNA. *See* TS Brochure at 1-2.

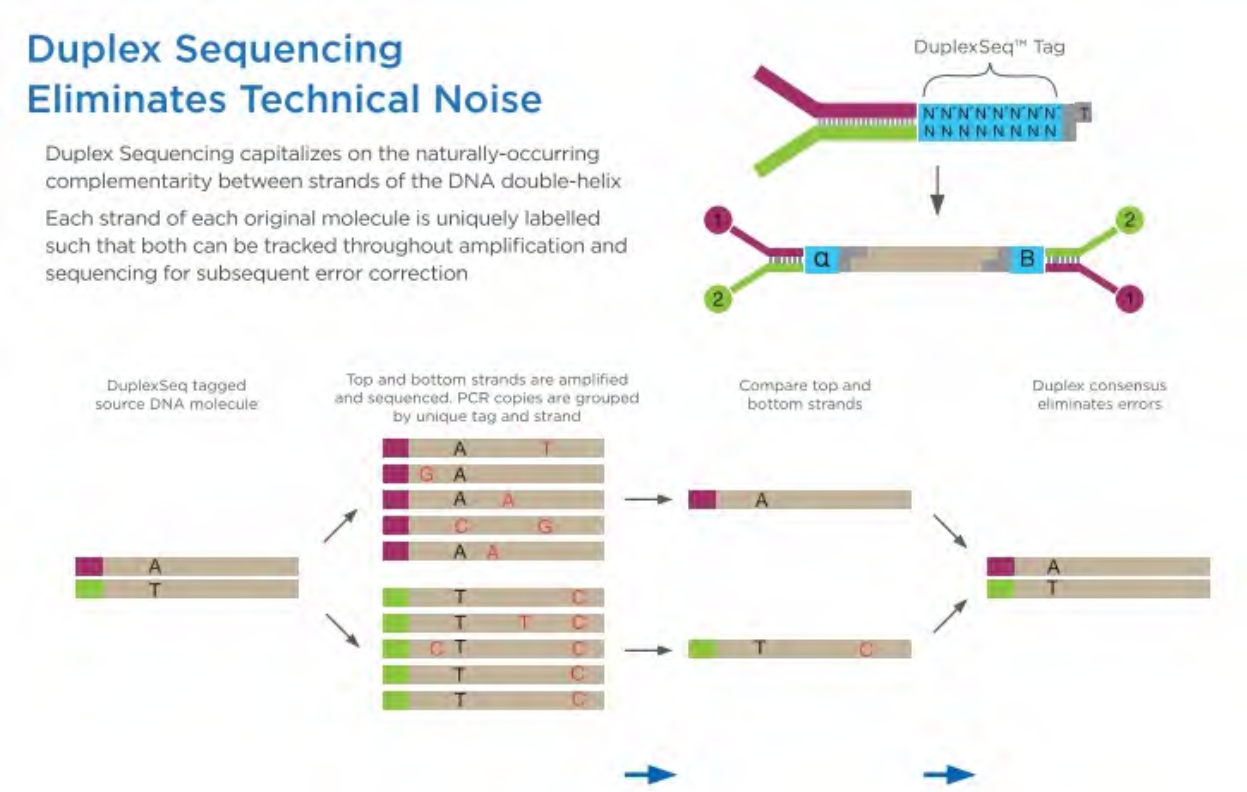
172. Dr. Jesse Salk, CEO and Founder of TwinStrand, has made several public statements regarding TwinStrand's sequencing methods including that "TwinStrand has developed [] a high-accuracy NGS technique called duplex sequencing, which uses both strands of a DNA molecule in the sequencing process to check for errors." *See* GEB April 2020 at 5 (Exhibit F); GEB August 2020" at 2 (Exhibit G). Dr. Salk has also stated that TwinStrand is "offering duplex sequencing kits and bioinformatics analysis" to consumers. *See* GEB August 2020 at 3.

173. TwinStrand has also posted several publications on its website describing the methods used by TwinStrand. For example, TwinStrand describes an ultra-sensitive sequencing method for cancer detection in cfDNA. *See* Salk 2019 (Exhibit H). This article states "[t]o explore the abundance of *TP53* mutations in liquid biopsies of clinical interest, we analyzed plasma-derived cell-free DNA (cfDNA) and peritoneal fluid from the 46-year old woman" and "[t]he fact that we identified one pathogenic mutation commonly found in cancers within cfDNA from a healthy woman (Figure S10; Table 5) raises important concerns over specificity in cancer-screening studies based on mutation detection in plasma." *Id.* at 141.

174. Further, in its Complaint, TwinStrand itself acknowledges that it practices the methods of the Asserted Patents, which include claims directed to methods for classifying unique sequencing reads generated from sequencing reads derived from double-stranded polynucleotide molecules. *See* e.g., Compl. at ¶¶ 20, 22; '951 patent at claims 1, 11, and 18.

175. TwinStrand also practices the "tagging" step of claim 1 of the '858 patent. This element recites "(a) tagging a plurality of double-stranded cfDNA molecules from a population of

double-stranded cfDNA molecules from the sample with a set of library adaptors comprising a plurality of molecular barcodes to generate tagged parent polynucleotides." For example:



TS Brochure at 1.

176. This "tagging" step includes tagging cfDNA. For example, "[t]o explore the abundance of TP53 mutations in liquid biopsies of clinical interest, we analyzed plasma-derived cell-free DNA (cfDNA) and peritoneal fluid from the 46-year old woman. TP53 mutations were identified in both, with cancer-like features similar to what was observed for solid tissue biopsies, UL, and leukocytes (Figure S10)." Salk 2019 at 141; *see also id.* at e2. Kennedy further instructs in its methods during ligation to "calculate the amount of DS adaptors…that will be needed for a 20:1 molar excess relative to the total DNA concentration in the final ligation reaction." Kennedy 2014 at 2597; Abstract, FIG. 1, 2586, 2591, 2597; *see also* Compl. at ¶¶ 20, 22; '951 patent at claims 1 and 21-25; 16:38-46 ("The step of tagging of both strands of individual duplex DNA may

be accomplished by ligation of degenerate or semi-degenerate complementary DNA sequences; as the complementary nature of the two strands of such a tag sequence allows the two molecules to be grouped together for error correction."); FIG. 1.

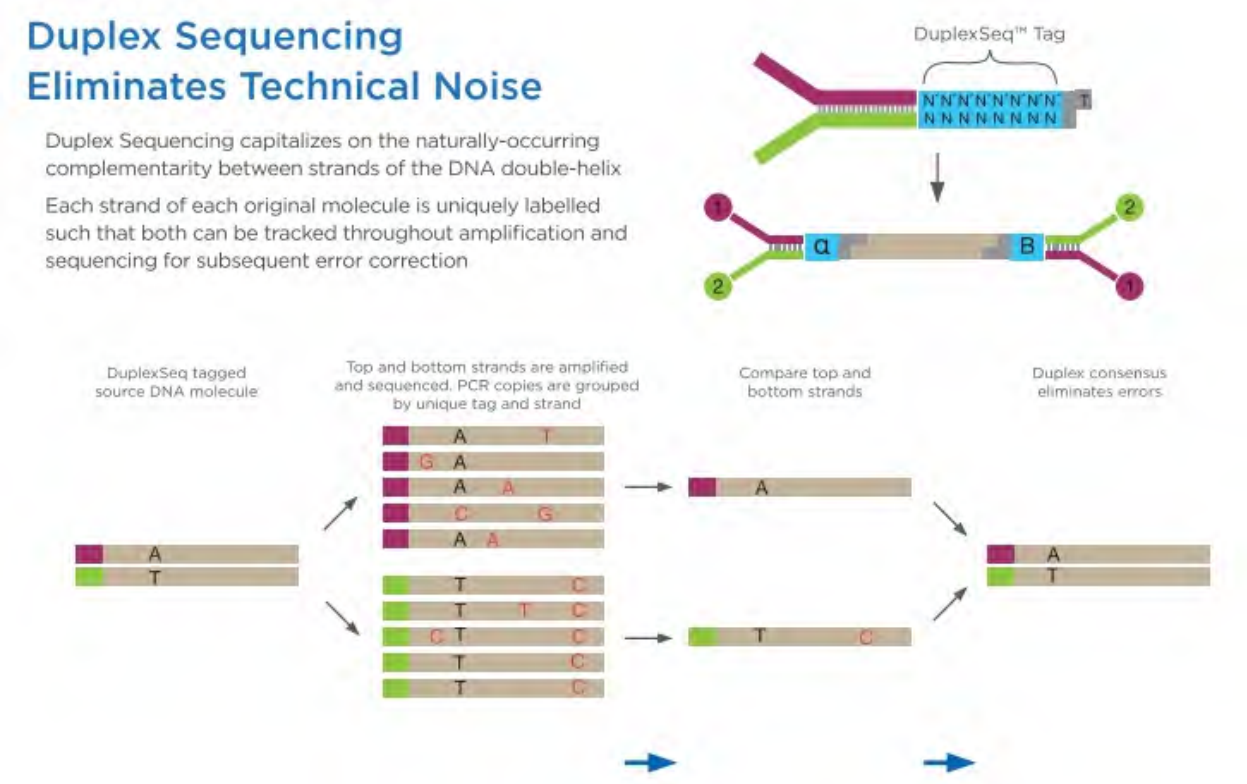
177. TwinStrand also practices the "tagging" step of claim 1 of the '858 patent "wherein the tagging comprises ligating a plurality of library adaptors from the set of library adaptors to the plurality of double-stranded cfDNA molecules from the population using more than a 10 times molar excess of library adaptors as compared to the double-stranded cfDNA molecules of the population." For example, TwinStrand uses ligating a plurality of library adaptors from the set of library adaptors to the plurality of double-stranded cfDNA molecules from the population using more than a 10 times molar excess of library adaptors as compared to the double-stranded cfDNA molecules of the population. *See* Salk 2019 at e2; Kennedy 2014 at Abstract, FIG. 1, 2591 ("To minimize the presence of adapter dimers, we determine the molar concentration of DNA molecules in the preligated DNA library on an Agilent TapeStation 2200 or Bioanalyzer 2100 and then use 20-fold molar excess of adapter relative to DNA."); 2597 ("By using the amount of sample DNA determined in Step 28, calculate the amount of DS adaptors…that will be needed for a 20:1 molar excess relative to the total DNA concentration in the final ligation reaction."); *see also* Compl. at ¶¶ 20, 22; '951 patent at claims 1 and 21-25; 16:38-46; FIG. 1.

178. TwinStrand also practices the "tagging" step of claim 1 of the '858 patent "wherein the tagging produces at least 20% of the double-stranded cfDNA molecules of the populations having library adaptors ligated to both ends of a molecule of the double-stranded cfDNA molecules." *See* Kennedy 2014 at 2587, 2590 (end repair and dA-tailing); FIG.1.

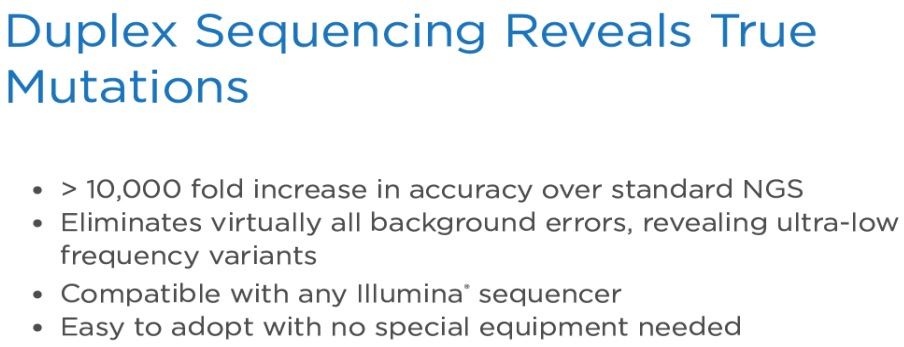
179. TwinStrand also practices the "amplifying" step of claim 1 of the '858 patent. This element recites "(b) amplifying a plurality of the tagged parent polynucleotides to produce progeny

polynucleotides." For example, TwinStrand advertises that "each strand of original molecule is uniquely labeled such that both ends can be tracked throughout amplification and sequencing for subsequent error detection." TS Brochure at 1; *see also* TwinStrand's website at 7; Kennedy 2014 at Abstract ("Individually labeled strands are then PCR-amplified, creating sequence 'families' that share a common tag sequence derived from the two original complementary strands."); Salk 2019 at e2 ("After initial amplification, 120 bp biotinylated oligonucleotide probes (Integrated DNA Technologies, Coralville, Iowa) were used to capture the coding region of TP53"); Compl. at ¶ 20; '951 patent at FIG. 1. Amplification includes PCR amplification.

180. TwinStrand also practices the "sequencing" step of claim 1 of the '858 patent. This element recites "(c) sequencing a plurality of the progeny polynucleotides to produce a set of sequencing reads." For example, TwinStrand's method is a sequencing method that produces a set of sequencing reads:

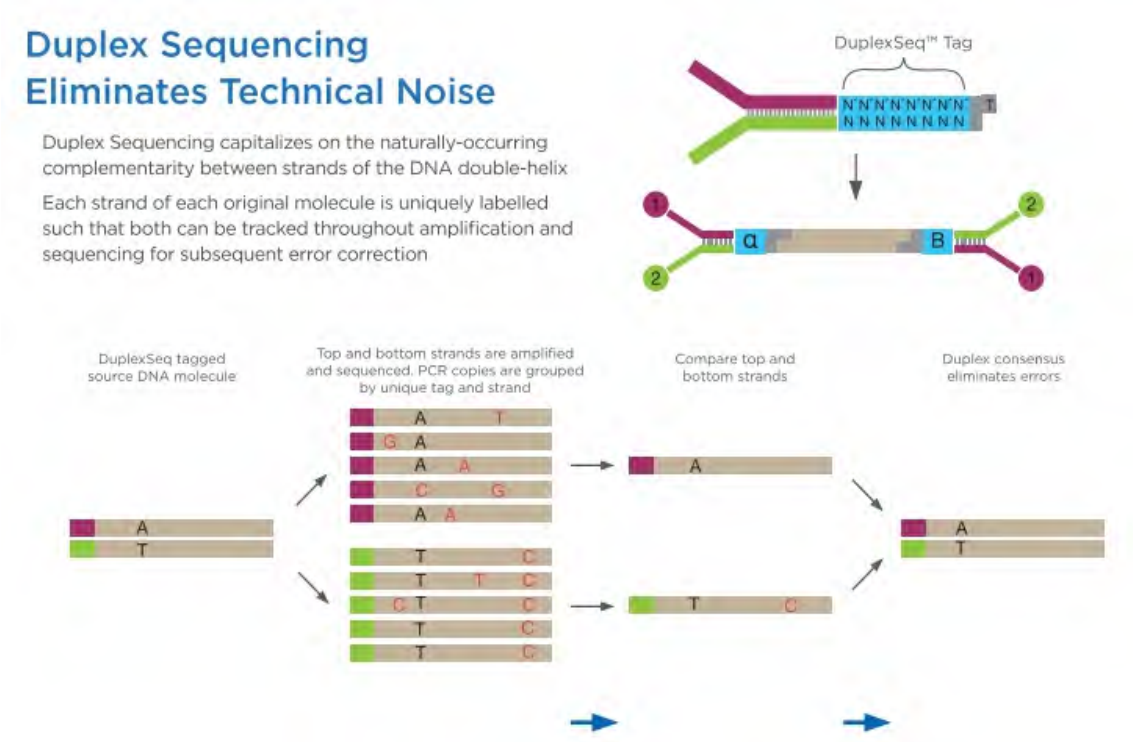


TS Brochure at 1.



TwinStrand Website at 6; *see also id.* at 4, 13; Kennedy 2014 at FIG. 1(C), 2587; Compl. at 26 ("TwinStrand provides services for nucleic acid analysis using the patented Duplex Sequencing methods and provides customers with Duplex Sequencing kits. These kits include a DNA library prep kit containing the reagents, adapters, and other components necessary to practice its Duplex Sequencing process. TwinStrand also provides access to bioinformatics software to process raw sequence read files and produce error-corrected sequences according to the patented processes"); '951 patent at claim 1; FIG. 1.

181. TwinStrand also practices the "determining individual double-stranded cfDNA molecules from among the tagged parent polynucleotides" step of claim 1 of the '858 patent. This element recites "(d) determining, based at least on sequence information from the molecular barcodes, individual double-stranded cfDNA molecules from among the tagged parent polynucleotides for which either (1) both a Watson strand and a Crick strand of the individual double-stranded cfDNA molecule are detected or (2) only one of a Watson strand or a Crick strand of the individual double-stranded cfDNA molecule is detected from a plurality of sequencing reads from the set of sequencing reads." For example, TwinStrand's method produces duplex consensus sequences resulting from the comparison of sequencing reads from tagged parent polynucleotides:



TS Brochure at 1; *see also* Kennedy 2014 at Abstract ("Individually labeled strands are then PCR-amplified, creating sequence 'families' that share a common tag sequence derived from the two original complementary strands. Mutations are scored only if the variant is present in the PCR families arising from both of the two DNA strands. Here we provide a detailed protocol for efficient DS adapter synthesis, library preparation and target enrichment, as well as an overview of the data analysis workflow."), FIG. 1; Salk 2019 at e2 ("Sequencing reads were aligned to hg19 then reads sharing a common molecular tag in both strand orientations were grouped and assembled into an error-corrected Duplex Consensus Sequence.").

182. Further, in its Complaint, TwinStrand admits that it practices the claims of the Asserted Patents, including claims 1 and 27 of the '951 patent. Claim 1 of the '951 patent recites "(d) sorting sequence reads into paired sequence reads and unpaired sequence reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from an original parent polynucleotide molecule

in said set, and (ii) each unpaired read represents a first tagged strand having no second differently tagged complementary strand derived from an original parent polynucleotide molecule represented among said sequence reads in said set of sequence reads; and (e) quantifying at least two of (i) said paired sequence reads, (ii) said unpaired sequence reads, (iii) read depth of said paired sequence reads and (iv) read depth of said unpaired sequence reads." Claim 27 recites "The method of claim 1, further comprising quantifying a total number of original double-stranded polynucleotide molecules or genome-equivalents in the sample based at least in part on said quantification of said paired sequence reads and said unpaired sequence reads."

183. By making, selling, and offering to sell kits that utilize TwinStrand Duplex Sequencing™ with the express purpose of encouraging, aiding and/or causing such customers to practice at least claim 1 of the '858 patent, TwinStrand willfully indirectly infringes at least claim 1 of the '858 patent through inducement and contributory infringement under 35 U.S.C. § 271 *et seq.* TwinStrand's indirect infringement has and continues to be willful.

184. Upon information and belief, TwinStrand has been aware of the existence of the '858 patent since at least January 12, 2021 and/or has willfully blinded itself as to the infringing nature of its activity since the issuance of the '858 patent.

185. By its actions, TwinStrand's infringement of the '858 patent has irreparably harmed Guardant. Unless TwinStrand's infringing acts are enjoined by this Court, Guardant will continue to suffer additional irreparable injury. Guardant has no adequate remedy at law.

186. By its actions, TwinStrand's infringement of the '858 patent has damaged and continues to damage Guardant in an amount yet to be determined of at least a reasonable royalty and/or lost profits that Guardant would have made but for TwinStrand's infringing acts.

COUNT III
(Direct and Indirect Infringement of U.S. Patent No. 11,118,221)

187. Guardant re-alleges and incorporated by reference Paragraphs 1-186 above, as if fully set forth herein.

188. The ’221 patent is directed to a method of determining genetic variants. Claim 1 of the patent recites:

1. A method, comprising:

(a) providing a population of cell-free deoxyribonucleic acid (cfDNA) molecules having first and second complementary strands;

(b) tagging a plurality of the cfDNA molecules of the population with a set of duplex tags comprising molecular barcodes from a set of molecular barcodes to produce tagged parent polynucleotides, wherein duplex tags from the set of duplex tags are attached at both ends of a molecule of the plurality of the cfDNA molecules;

(c) amplifying a plurality of the tagged parent polynucleotides to produce amplified progeny polynucleotides;

(d) sequencing at least a subset of the amplified progeny polynucleotides to produce a set of sequence reads; and

(e) reducing or tracking redundancy in the set of sequence reads using at least sequence information from the molecular barcodes to generate a plurality of consensus sequences representative of original cfDNA molecules from among the tagged parent polynucleotides, wherein the plurality of consensus sequences is generated from (i) paired reads corresponding to sequence reads generated from a first tagged strand and a second tagged complementary strand derived from a cfDNA molecule from among the tagged parent polynucleotides, and (ii) unpaired

reads corresponding to sequence reads generated from a first tagged strand having no second tagged complementary strand derived from a cfDNA molecule from among the tagged parent polynucleotides.

189. TwinStrand has willfully infringed and continues to infringe at least claim 1 of the '221 patent, literally or under the doctrine of equivalents, by performing the methods of the '221 patent in the United States. It is expected that discovery likely will reveal additional evidentiary support that TwinStrand performs the above limitations of the '221 patent and Guardant reserves the right to supplement the evidence cited below as additional evidentiary support is obtained.

190. TwinStrand practices the preamble of claim 1 of the '221 patent, which provides "a method." For example, TwinStrand's Technology Brochure describes the methods used by TwinStrand to reveal true mutations in double-stranded DNA, including cfDNA. TS Brochure at 1-2.

191. Dr. Jesse Salk, CEO and Founder of TwinStrand, has made several public statements regarding TwinStrand's sequencing methods including that "TwinStrand has developed [] a high-accuracy NGS technique called duplex sequencing, which uses both strands of a DNA molecule in the sequencing process to check for errors." *See* GEB April 2020" at 5 (Exhibit F); GEB August 2020 at 2 (Exhibit G). Dr. Salk has also stated that TwinStrand is "offering duplex sequencing kits and bioinformatics analysis" to consumers. *See* GEB August 2020 at 3.

192. TwinStrand has also posted several publications on its website describing the methods used by TwinStrand. For example, TwinStrand describes an ultra-sensitive sequencing method for cancer detection in cfDNA. *See* Salk 2019 (Exhibit H). This article states "[t]o explore the abundance of *TP53* mutations in liquid biopsies of clinical interest, we analyzed plasma-derived cell-free DNA (cfDNA) and peritoneal fluid from the 46-year old woman" and "[t]he fact

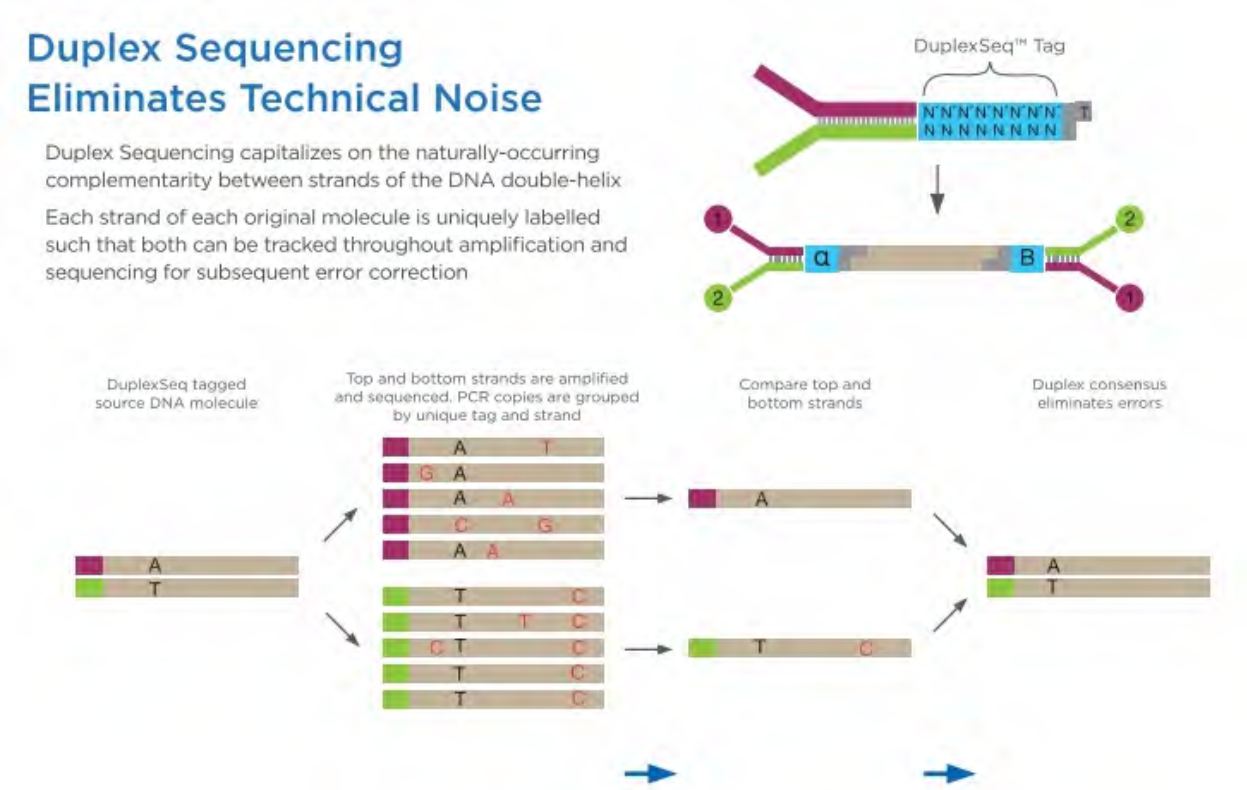
that we identified one pathogenic mutation commonly found in cancers within cfDNA from a healthy woman (Figure S10; Table 5) raises important concerns over specificity in cancer-screening studies based on mutation detection in plasma." *Id.* at 141.

193. Further, in its Complaint, TwinStrand itself acknowledges that it practices the methods of the Asserted Patents, which include claims directed to methods for classifying unique sequencing reads generated from sequencing reads derived from double-stranded polynucleotide molecules. *See* e.g., Compl. at ¶¶ 20, 22; '951 patent at claim 1.

194. TwinStrand also practices the "providing a cfDNA molecule" step of claim 1 of the '221 patent. This element recites "(a) providing a population of cell-free deoxyribonucleic acid (cfDNA) molecules having first and second complementary strands." For example, TwinStrand describes an ultra-sensitive sequencing method for cancer detection in cfDNA. *See* Salk 2019 at 141.

195. Further, in its Complaint, TwinStrand itself acknowledges that it practices the methods of the Asserted Patents, which include claims directed to a method comprising "(a) providing a sample comprising a set of double-stranded polynucleotide molecules, each double-stranded polynucleotide molecule including first and second complementary strands." '951 patent at claim 1.

196. TwinStrand also practices the "tagging" step of claim 1 of the '221 patent. This element recites "(b) tagging a plurality of the cfDNA molecules of the population with a set of duplex tags comprising molecular barcodes from a set of molecular barcodes to produce tagged parent polynucleotides, wherein duplex tags from the set of duplex tags are attached at both ends of a molecule of the plurality of the cfDNA molecules." For example:



TS Brochure at 1.

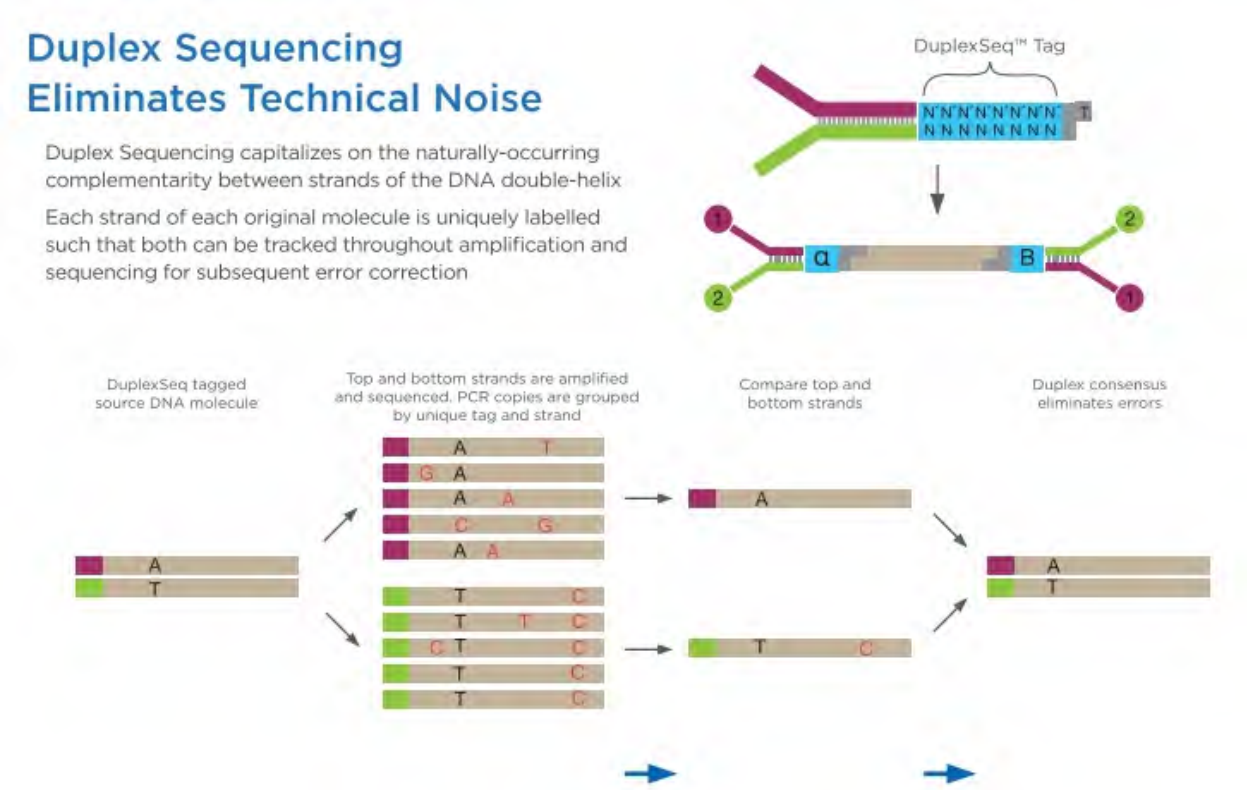
197. This "tagging" step includes tagging cfDNA. For example, "[t]o explore the abundance of TP53 mutations in liquid biopsies of clinical interest, we analyzed plasma-derived cell-free DNA (cfDNA) and peritoneal fluid from the 46-year old woman. T P53 mutations were identified in both, with cancer-like features similar to what was observed for solid tissue biopsies, UL, and leukocytes (Figure S10)." Salk 2019 at 141; *see also id.* at e2.

198. Additional publications posted on TwinStrand's website also show TwinStrand uses tagging of a population of double-stranded DNA molecules. *See* Kennedy 2014 at Abstract, FIG. 1 ("(b) Ligation of the adaptors to the sample DNA results in a unique 12-nt tag sequence on both ends of the molecule,"); 2586, 2591, 2597; *see also* Compl. at ¶¶ 20, 22; '951 patent at claims 1 and 21-25; 16:38-46 ("The step of tagging of both strands of individual duplex DNA may be accomplished by ligation of degenerate or semi-degenerate complementary DNA sequences; as

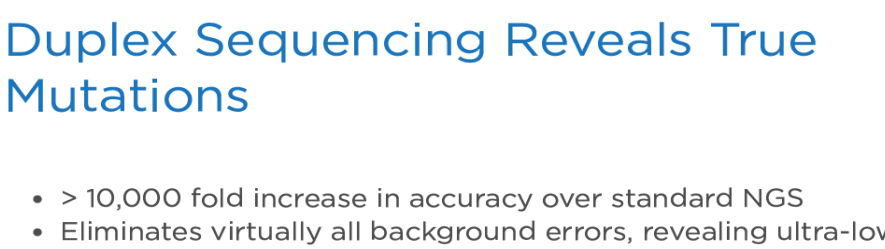
the complementary nature of the two strands of such a tag sequence allows the two molecules to be grouped together for error correction."); FIG. 1.

199. TwinStrand also practices the "amplifying" step of claim 1 of the '221 patent. This element recites "amplifying a plurality of the tagged parent polynucleotides to produce amplified progeny polynucleotides." For example, TwinStrand advertises that "each strand of original molecule is uniquely labeled such that both ends can be tracked throughout amplification and sequencing for subsequent error detection." TS Brochure at 1; *see also* TwinStrand's website at 7; Kennedy 2014 at Abstract ("Individually labeled strands are then PCR-amplified, creating sequence 'families' that share a common tag sequence derived from the two original complementary strands."); Salk 2019 at e2 ("After initial amplification, 120 bp biotinylated oligonucleotide probes (Integrated DNA Technologies, Coralville, Iowa) were used to capture the coding region of TP53"); Compl. at ¶ 20; '951 patent at FIG. 1. Amplification includes PCR amplification.

200. TwinStrand also practices the "sequencing" step of claim 1 of the '221 patent. This element recites "(d) sequencing at least a subset of the amplified progeny polynucleotides to produce a set of sequence reads." For example, TwinStrand's method is a sequencing method that produces a set of sequencing reads:



TS Brochure at 1.



TwinStrand Website at 6; *see also id.* at 4, 13; Kennedy 2014 at FIG. 1(C), 2587; Compl. at 26 ("TwinStrand provides services for nucleic acid analysis using the patented Duplex Sequencing methods and provides customers with Duplex Sequencing kits. These kits include a DNA library prep kit containing the reagents, adapters, and other components necessary to practice its Duplex Sequencing process. TwinStrand also provides access to bioinformatics software to process raw

sequence read files and produce error-corrected sequences according to the patented processes"); '951 patent at claim 1; FIG. 1.

201. TwinStrand also practices the "reducing or tracking redundancy in the set of sequencing reads" step of claim 1 of the '221 patent. This element recites "(e) reducing or tracking redundancy in the set of sequence reads using at least sequence information from the molecular barcodes to generate a plurality of consensus sequences representative of original cfDNA molecules from among the tagged parent polynucleotides, wherein the plurality of consensus sequences is generated from (i) paired reads corresponding to sequence reads generated from a first tagged strand and a second tagged complementary strand derived from a cfDNA molecule from among the tagged parent polynucleotides, and (ii) unpaired reads corresponding to sequence reads generated from a first tagged strand having no second tagged complementary strand derived from a cfDNA molecule from among the tagged parent polynucleotides."

202. For example, in its Complaint TwinStrand admits that it practices the methods of the Asserted Patents, including claim 1 of the '951 patent, which recites "(d) sorting sequence reads into paired sequence reads and unpaired sequence reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from an original parent polynucleotide molecule in said set, and (ii) each unpaired read represents a first tagged strand having no second differently tagged complementary strand derived from an original parent polynucleotide molecule represented among said sequence reads in said set of sequence reads; and (e) quantifying at least two of (i) said paired sequence reads, (ii) said unpaired sequence reads, (iii) read depth of said paired sequence reads and (iv) read depth of said unpaired sequence reads." *See also* claims 2-4 and 27-28.

203. By making, selling, and offering to sell kits that utilize TwinStrand Duplex Sequencing™ with the express purpose of encouraging, aiding and/or causing such customers to practice at least claim 1 of the '221 patent, TwinStrand indirectly infringes at least claim 1 of the '221 patent through inducement and contributory infringement under 35 U.S.C. § 271 *et seq*. TwinStrand's indirect infringement has and continues to be willful.

204. Upon information and belief, TwinStrand has been aware of the existence of the '221 patent since at least September 14, 2021 and/or has willfully blinded itself as to the infringing nature of its activity since the issuance of the '221 patent.

205. By its actions, TwinStrand's infringement of the '221 patent has irreparably harmed Guardant. Unless TwinStrand's infringing acts are enjoined by this Court, Guardant will continue to suffer additional irreparable injury. Guardant has no adequate remedy at law.

206. By its actions, TwinStrand's infringement of the '221 patent has damaged and continues to damage Guardant in an amount yet to be determined of at least a reasonable royalty and/or lost profits that Guardant would have made but for TwinStrand's infringing acts.

COUNT IV
(Direct and Indirect Infringement of U.S. Patent No. 11,149,306)

207. Guardant re-alleges and incorporated by reference Paragraphs 1-206 above, as if fully set forth herein.

208. The '306 patent is directed to a method of determining genetic variants. Claim 17 of the patent recites:

> 17. A method, comprising:
>
> (a) tagging a population of double-stranded cell-free deoxyribonucleic acid (cfDNA) molecules obtained or derived from a sample of a subject with a set of tags comprising molecular barcodes to produce tagged parent polynucleotides;

(b) amplifying a plurality of the tagged parent polynucleotides to produce amplified progeny polynucleotides;

(c) sequencing at least a subset of the amplified progeny polynucleotides to produce a set of sequence reads; and

(d) sorting a plurality of sequence reads from the set of sequence reads into (i) families comprising paired reads corresponding to sequence reads generated from a first tagged strand and a second tagged complementary strand derived from double-stranded cfDNA molecules from among the tagged parent polynucleotides, and (ii) families comprising unpaired reads corresponding to sequence reads generated from a first tagged strand having no second tagged complementary strand derived from double-stranded cfDNA molecules from among the tagged parent polynucleotides.

209. TwinStrand has willfully infringed and continues to infringe at least claim 17 of the ’306 patent, literally or under the doctrine of equivalents, by performing the methods of the ’306 patent in the United States. It is expected that discovery likely will reveal additional evidentiary support that TwinStrand performs the above limitations of the ’306 patent and Guardant reserves the right to supplement the evidence cited below as additional evidentiary support is obtained.

210. TwinStrand practices the preamble of claim 17 of the ’306 patent, which provides “a method.” For example, TwinStrand’s Technology Brochure describes the methods used by TwinStrand to reveal true mutations in double-stranded DNA, including cfDNA. *See* TS Brochure at 1-2.

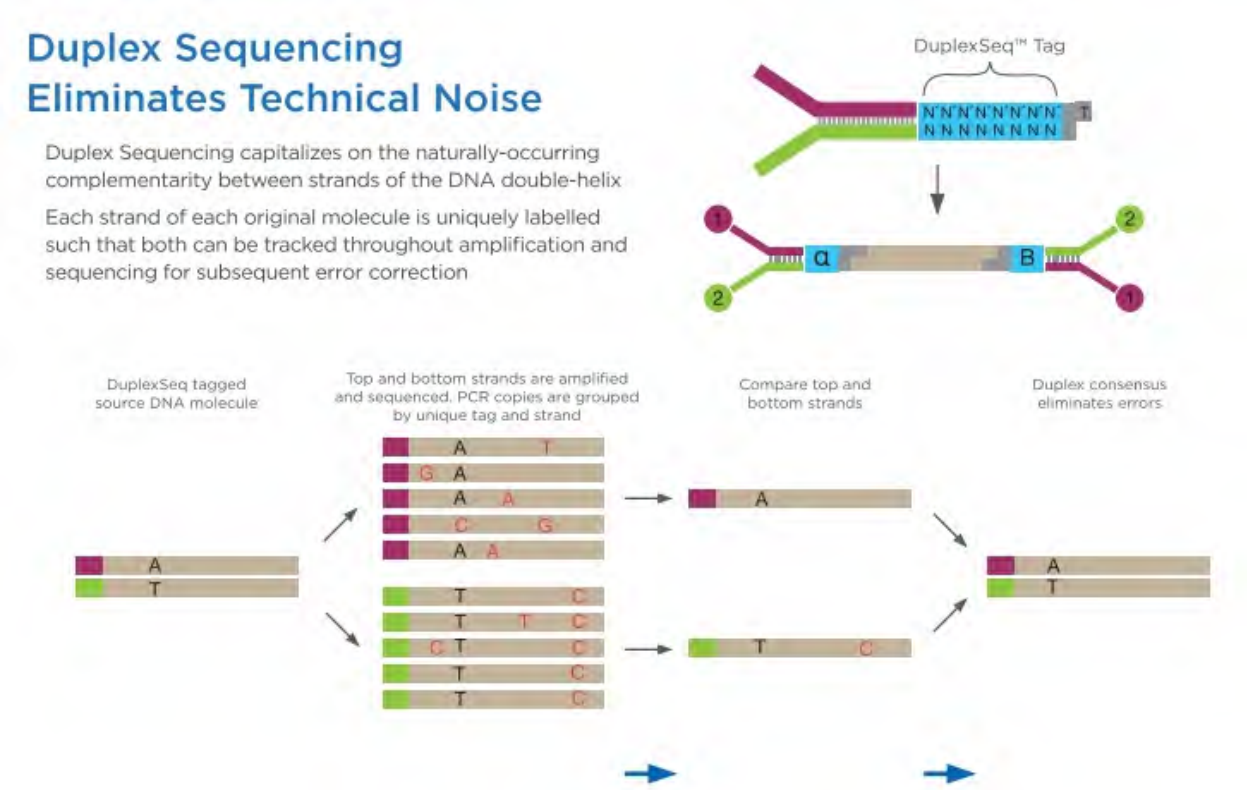
211. Dr. Jesse Salk, CEO and Founder of TwinStrand, has made several public statements regarding TwinStrand’s sequencing methods including that “TwinStrand has developed

[] a high-accuracy NGS technique called duplex sequencing, which uses both strands of a DNA molecule in the sequencing process to check for errors." *See* GEB April 2020" at 5 (Exhibit F); GEB August 2020 at 2 (Exhibit G). Dr. Salk has also stated that TwinStrand is "offering duplex sequencing kits and bioinformatics analysis" to consumers. *See* GEB August 2020 at 3.

212. TwinStrand has also posted several publications on its website describing the methods used by TwinStrand. For example, TwinStrand describes an ultra-sensitive sequencing method for cancer detection in cfDNA. *See* Salk 2019 (Exhibit H). This article states "[t]o explore the abundance of TP53 mutations in liquid biopsies of clinical interest, we analyzed plasma-derived cell-free DNA (cfDNA) and peritoneal fluid from the 46-year old woman" and "[t]he fact that we identified one pathogenic mutation commonly found in cancers within cfDNA from a healthy woman (Figure S10; Table 5) raises important concerns over specificity in cancer-screening studies based on mutation detection in plasma." *Id*. at 141.

213. Further, in its Complaint, TwinStrand itself acknowledges that it practices the methods of the Asserted Patents, which include claims directed to methods for classifying unique sequencing reads generated from sequencing reads derived from double-stranded polynucleotide molecules. *See* e.g., Compl. at ¶¶ 20, 22; '951 patent at claim 1.

214. TwinStrand also practices the "tagging" step of claim 17 of the '306 patent. This element recites "(a) tagging a population of double-stranded cell-free deoxyribonucleic acid (cfDNA) molecules obtained or derived from a sample of a subject with a set of tags comprising molecular barcodes to produce tagged parent polynucleotides." For example:



TS Brochure at 1.

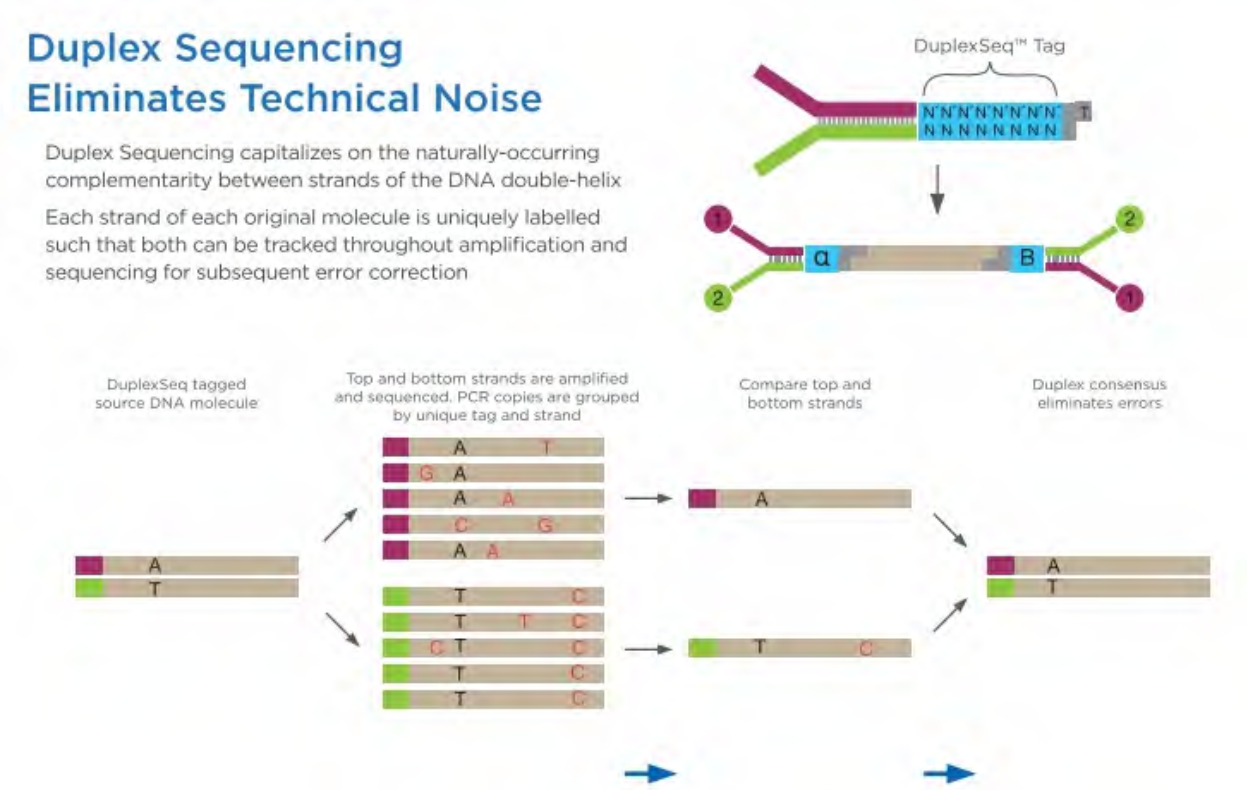
215. This "tagging" step includes tagging cfDNA. For example, "[t]o explore the abundance of TP53 mutations in liquid biopsies of clinical interest, we analyzed plasma-derived cell-free DNA (cfDNA) and peritoneal fluid from the 46-year old woman. TP53 mutations were identified in both, with cancer-like features similar to what was observed for solid tissue biopsies, UL, and leukocytes (Figure S10)." Salk 2019 at 141; *see also id.* at e2.

216. Additional publications posted on TwinStrand's website also show TwinStrand uses tagging of a population of double-stranded DNA molecules. *See* Kennedy 2014 at Abstract, FIG. 1 ("(b) Ligation of the adaptors to the sample DNA results in a unique 12-nt tag sequence on both ends of the molecule,"); 2586, 2590-91, 2597; *see also* Compl. at ¶¶ 20, 22; '951 patent at claims 1 and 21-25; 16:38-46 ("The step of tagging of both strands of individual duplex DNA may be accomplished by ligation of degenerate or semi-degenerate complementary DNA sequences; as

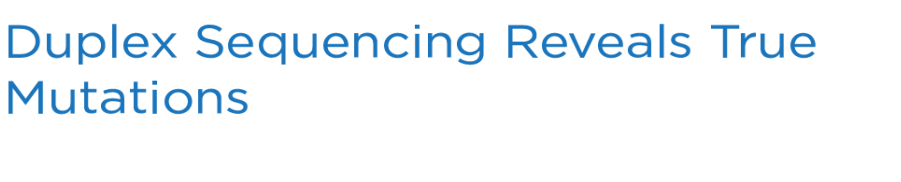
the complementary nature of the two strands of such a tag sequence allows the two molecules to be grouped together for error correction."); FIG. 1.

217. TwinStrand also practices the "amplifying" step of claim 17 of the '306 patent. This element recites "(b) amplifying a plurality of the tagged parent polynucleotides to produce amplified progeny nucleotides." For example, TwinStrand advertises that "each strand of original molecule is uniquely labeled such that both ends can be tracked throughout amplification and sequencing for subsequent error detection." TS Brochure at 1; *see also* TwinStrand's website at 7; Kennedy 2014 at Abstract ("Individually labeled strands are then PCR-amplified, creating sequence 'families' that share a common tag sequence derived from the two original complementary strands."); Salk 2019 at e2 ("After initial amplification, 120 bp biotinylated oligonucleotide probes (Integrated DNA Technologies, Coralville, Iowa) were used to capture the coding region of TP53"); Compl. at ¶ 20; '951 patent at FIG. 1. Amplification includes PCR amplification.

218. TwinStrand also practices the "sequencing" step of claim 17 of the '306 patent. This element recites "(c) sequencing at least a subset of the amplified progeny polynucleotides to produce a set of sequence reads." For example, TwinStrand's method is a sequencing method that produces a set of sequencing reads:



TS Brochure at 1.



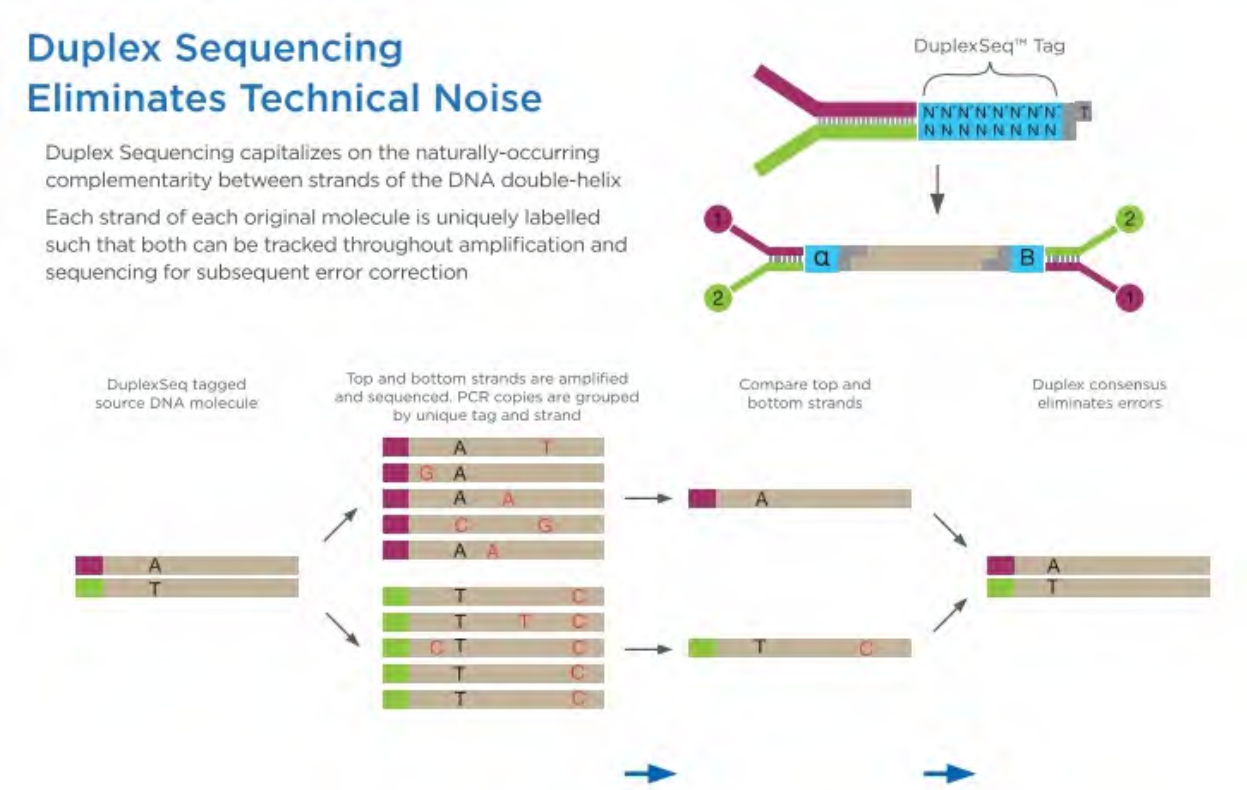
TwinStrand Website at 6; *see also id.* at 4, 13; Kennedy 2014 at FIG. 1(C), 2587; Compl. at 26 ("TwinStrand provides services for nucleic acid analysis using the patented Duplex Sequencing methods and provides customers with Duplex Sequencing kits. These kits include a DNA library prep kit containing the reagents, adapters, and other components necessary to practice its Duplex Sequencing process. TwinStrand also provides access to bioinformatics software to process raw

sequence read files and produce error-corrected sequences according to the patented processes"); '951 patent at claim 1; FIG. 1.

219. TwinStrand also practices the "sorting" step of claim 17 of the '306 patent. This element recites "(d) sorting a plurality of sequence reads from the set of sequence reads into (i) families comprising paired reads corresponding to sequence reads generated from a first tagged strand and a second tagged complementary strand derived from double-stranded cfDNA molecules from among the tagged parent polynucleotides, and (ii) families comprising unpaired reads corresponding to sequence reads generated from a first tagged strand having no second tagged complementary strand derived from double-stranded cfDNA molecules from among the tagged parent polynucleotides."

220. For example, in its Complaint, TwinStrand admits that it practices the claims of the Asserted Patents, including claim 1 of the '951 patent. Claim 1 of the '951 patent recites "(d) sorting sequence reads into paired sequence reads and unpaired sequence reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from an original parent polynucleotide molecule in said set, and (ii) each unpaired read represents a first tagged strand having no second differently tagged complementary strand derived from an original parent polynucleotide molecule represented among said sequence reads in said set of sequence reads; and (e) quantifying at least two of (i) said paired sequence reads, (ii) said unpaired sequence reads, (iii) read depth of said paired sequence reads and (iv) read depth of said unpaired sequence reads."

221. Additional publications on TwinStrand's website also show that TwinStrand practices the "sorting" step of claim 17 of the '306 patent. For example:



TS Brochure at 1; *see also* Kennedy 2014 at FIG. 1. 2587.

222. By making, selling, and offering to sell kits that utilize TwinStrand Duplex Sequencing™ with the express purpose of encouraging, aiding and/or causing such customers to practice at least claim 17 of the '306 patent, TwinStrand indirectly infringes at least claim 1 of the '306 patent through inducement and contributory infringement under 35 U.S.C. § 271 *et seq*. TwinStrand's indirect infringement has and continues to be willful.

223. Upon information and belief, TwinStrand has been aware of the existence of the '306 patent since at least October 19, 2021 and/or has willfully blinded itself as to the infringing nature of its activity since the issuance of the '306 patent.

224. By its actions, TwinStrand's infringement of the '306 patent has irreparably harmed Guardant. Unless TwinStrand's infringing acts are enjoined by this Court, Guardant will continue to suffer additional irreparable injury. Guardant has no adequate remedy at law.

225. By its actions, TwinStrand's infringement of the '306 patent has damaged and continues to damage Guardant in an amount yet to be determined of at least a reasonable royalty and/or lost profits that Guardant would have made but for TwinStrand's infringing acts.

COUNT V
(Declaration of Unpatentability of the '699 Patent Under 35 U.S.C. §§ 101 *et seq*)

226. Guardant re-alleges and incorporates by reference Paragraphs 1-225 above, as if fully set forth herein.

227. The claims of the '699 Patent are unpatentable and invalid for failure to comply with 35 U.S.C. §§ 101, 115, and 135(b).

228. The '699 Patent fails to name the actual inventor of the claimed subject matter.

229. Chapter 2157 of the Manual of Patent Examining Procedure states: "Although the AIA eliminated pre-AIA 35 U.S.C. § 102(f), the patent laws still require naming of the actual inventors of the claimed subject matter . . . In the rare situation where it [is] clear that the application does not name the correct inventorship and the applicant has not filed a request to correct inventorship under 37 C.F.R. § 1.48, Office personnel should reject the claims under 35 U.S.C. § 101 and 35 U.S.C. § 115."

230. 35 U.S.C. § 101 states: "whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefore, subject to the conditions and requirements of this title."

231. 35 U.S.C. § 115(a) states: "an applicant for a patent that is filed under 111(a) . . . shall include, or be amended to include, the name of the inventor for any invention claimed in the application."

232. 35 U.S.C. § 135(b) states that a "claim which is the same as, or for the same or substantially the same subject matter as, a claim of an issued patent may not be made in any

application unless such a claim is made prior to one year from the date on which the patent was granted."

233. The '699 Patent fails to meet the statutory requirements of 35 U.S.C. §§ 101 and 115, because, to the extent it claims inventive subject matter, such invention was not made by the named inventors.

234. The '699 Patent fails to meet the statutory requirements of 35 USC § 135(b)(1), because the claims of the '699 Patent were copied from Guardant's '822 Patent more than one year after the issue date of the '822 Patent.

235. There exists as actual controversy between Guardant and TwinStrand regarding the patentability of the claims of the '699 Patent given Guardant's allegation of infringement, and a judicial declaration of unpatentability and/or invalidity is necessary and appropriate at this time.

COUNT VI
(Declaration of Unpatentability of the '951 Patent Under 35 U.S.C. §§ 101 *et seq*)

236. Guardant re-alleges and incorporates by reference Paragraphs 1-235 above, as if fully set forth herein.

237. The claims of the '951 Patent are unpatentable and invalid for failure to comply with 35 U.S.C. §§ 101 and 115.

238. The '951 Patent fails to name the actual inventor of the claimed subject matter.

239. Chapter 2157 of the Manual of Patent Examining Procedure states: "Although the AIA eliminated pre-AIA 35 U.S.C. § 102(f), the patent laws still require naming of the actual inventors of the claimed subject matter . . . In the rare situation where it [is] clear that the application does not name the correct inventorship and the applicant has not filed a request to

correct inventorship under 37 C.F.R. § 1.48, Office personnel should reject the claims under 35 U.S.C. § 101 and 35 U.S.C. § 115."

240. 35 U.S.C. § 101 states: "whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefore, subject to the conditions and requirements of this title."

241. 35 U.S.C. § 115(a) states: "an applicant for a patent that is filed under 111(a) . . . shall include, or be amended to include, the name of the inventor for any invention claimed in the application."

242. The '951 Patent fails to meet the statutory requirements of 35 U.S.C. §§ 101 and 115, because, to the extent it claims inventive subject matter, such invention was not made by the named inventors.

243. There exists as actual controversy between Guardant and TwinStrand regarding the patentability of the claims of the '951 Patent given Guardant's allegation of infringement, and a judicial declaration of unpatentability and/or invalidity is necessary and appropriate at this time.

COUNT VII
(Declaration of Unenforceability of the '699 Patent For Inequitable Conduct)

244. Guardant re-alleges and incorporates by reference Paragraphs 1-243 above, as if fully set forth herein.

245. During prosecution of the '699 patent, TwinStrand made misrepresentations and omissions material to patentability and did so with the specific intent to mislead or deceive the PTO.

246. The TwinStrand applicants deliberately and improperly copied claims from the '822 Patent more than one year after issuance of the '822 Patent in violation of 35 U.S.C. § 135(b).

Under pre-AIA 35 U.S.C. § 135(b)(1), "[a] claim which is the same as, or for the same or substantially the same subject matter as, a claim of an issued patent may not be made in any application unless such a claim is made prior to one year from the date on which the patent was granted." By doing so outside of the one-year timeframe, TwinStrand attempted to evade an interference proceeding.

247. Not only did the TwinStrand applicants violate the timing provisions of § 135(b)(1), but they also never informed the Patent Examiner that the claims were copied.

248. On May 13, 2019, Lara J. Dueppen, a patent attorney, filed U.S. Patent Application No. 16/411,066 ("'066 Application"), which issued as TwinStrand's U.S. Patent No. 10,689,699, which it is now asserting against Guardant. The '066 Application names Jesse Salk, Lawrence Loeb, and Michael Schmitt as inventors. All three inventors submitted declarations pursuant to 37 C.F.R. § 1.76. The '066 Application claims priority to Provisional Application No. 61/613,413, filed on March 20, 2012.

249. The '066 Application presented 36 claims, which were reduced to 20 claims in a preliminary amendment filed on the same day. Claims 37-66 were first added in a May 24, 2019 amendment, and subsequently amended on October 29, 2019, and February 4, 2020.

250. Claims 37-39 and 41-43 of the '066 Application were exact copies of claims 1-3, 6-7, and 10 from Guardant's '822 Patent, which issued on December 5, 2017. The subsequent amendment on October 29, 2019 and February 4, 2020 did not materially alter the copied claims.

251. The inventors and the assignee of the '822 Patent are unrelated to Jesse Salk, Lawrence Loeb, and Michael Schmitt.

252. Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen cited the '822 Patent in their June 18, 2019 Information Disclosure Statement. They were thus aware of the '822 Patent and its claims.

253. Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen are experienced before the PTO. Lara J. Dueppen was admitted to the State Bar of California in December 2008 and has been registered as a patent attorney since August 2009. Jesse Salk, Lawrence Loeb, and Michael Schmitt are each named as inventors on multiple patents and applications for which they have submitted oaths and declarations.

254. Patent applicants and attorneys are subject to the duties of good faith, candor, and disclosure, among others, including a duty to disclose to the PTO all information known to that individual to be material to patentability of the claimed invention. 37 C.F.R. § 1.56.

255. 7 C.F.R. § 1.607(c) requires that "[w]hen an applicant [for patent] presents a claim which corresponds exactly or substantially to a claim of a patent, the applicant shall identify the patent and the number of the patent claim. . . ."

256. MPEP § 2001.06(d) (8th ed., rev. 8, July 2010), as it existed at the time the '066 Application was filed, states that "[w]here claims are copied or substantially copied from a patent, 37 C.F.R. 1.607(c) requires applicant shall, at the time he or she presents the claim(s), identify the patent and the numbers of the patent claims..."

257. Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen did not notify the PTO that claims 37-39 and 41-43 of the '066 Application were copied from the '822 Patent.

258. 37 C.F.R. § 10.23(c)(7) states that "[k]nowingly withholding from the [PTO] information identifying a patent or patent application of another from which one or more claims have been copied" constitutes disreputable or gross misconduct.

259. MPEP § 2001.06(d) (8th ed., rev. 8, July 2010), as it existed at the time the '066 Application was filed, states that "the information required by 37 CFR 1.607(c) as to the source of copied claims is material information under 37 CFR 1.56. . . ."

260. Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen's failure to notify the PTO that claims 37-39 and 41-43 of the '066 Application were copied from the '822 Patent and to identify which claims of the '822 Patent were copied, constitutes affirmative egregious misconduct.

261. Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen's failure to notify the PTO of the copied claims was material to the patentability of at least claims 1-5 of the '699 Patent which correspond to originally filed claims 37-39 and 42-43 in the '066 Application, because failure to identify such information to the PTO is per se material to patentability.

262. In addition, applicants' failure to notify the PTO of the copied claims was material to the patentability of at least claims 1-5 of the '699 Patent because identifying the copied claims would have triggered an interference, a much more thorough and in-depth examination as to whether the claims of the '066 Application were entitled to be issued in light of the existence of identical claims in a previously-issued U.S. Patent.

263. On information and belief, the '699 Patent would not have issued if Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen had notified the PTO of the copied claims.

264. On information and belief, Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen affirmatively chose not to notify the PTO that they copied the claims of the '066 Application from the '822 Patent with the intent to mislead the PTO. Alternatively, such intent to mislead the PTO is the only reasonable inference from the totality of the circumstances alleged in this counterclaim.

265. This is consistent with Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen's practice of copying claims into their own applications without disclosing that the claims are copied, which has occurred in at least the following instances:

- U.S. Application No. 16/118,290, filed August 30, 2018, included at least one copied claim from U.S. Application No. 2016/0319345, assigned to Illumina Inc. U.S. Application No. 2016/0319345 was cited in an IDS during prosecution of U.S. Application No. 16/118,290.
- U.S. Application No. 16/120,072 (issued as U.S. Patent No. 10,385,393), filed August 31, 2018, included at least one copied claim from U.S. Publication No. 2016/0319345, assigned to Illumina Inc. U.S. Publication No. 2016/0319345 was cited in an IDS during prosecution of U.S. Application No. 16/120,072.
- U.S. Application No. 16/120,091 (issued as U.S. Patent No. 10,370,713), filed August 31, 2018, included at least one copied claim from U.S. Patent No. 9,920,366, assigned to Guardant. U.S. Patent No. 9,920,366 was cited in an IDS after issuance of U.S. Application No. 16/120,091.
- U.S. Application No. 16/411,045 (issued as U.S. Patent No. 10,604,804), filed May 13, 2019, included at least one copied claim from U.S. Patent No. 9,598,731, assigned to Guardant. U.S. Patent No. 9,598,731 was cited in an IDS during prosecution of U.S. Application No. 16/411,045.
- U.S. Application No. 16/514,931, filed July 17, 2019, included at least one copied claim from WO2015/100427A1, assigned to Guardant.

266. In at least one instance of claim copying, Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen did notify the PTO of the copied claims.

267. In a December 14, 2020 response to a restriction requirement in U.S. Application No. 16/118,305, Lara J. Dueppen submitted the following statement "Claim Copying. Claims 39 and 57 (newly introduced in a Preliminary Amendment filed on December 10, 2018 in the subject application) were substantially copied from claim 2 of U.S. Patent Appl. Publ. No. 2018/0291438. Applicant hereby specifically directs the Office's attention to claims 39 and 57 that substantially correspond to claim 2 of U.S. Patent. Appl. Publ. No.2018/0291438."

268. After notifying the PTO of the claim copying, no patent issued from U.S. Application No. 16/118,305.

269. Therefore, Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen knew they had a duty to disclose claim copying to the PTO. Their failure to notify the PTO of the copied claims in the application leading to the '699 Patent constitutes inequitable conduct that renders the '699 Patent unenforceable.

COUNT VIII
(Declaration of Unenforceability of the '951 Patent For Inequitable Conduct)

270. Guardant re-alleges and incorporates by reference Paragraphs 1-269 above, as if fully set forth herein.

271. During prosecution of the '951 patent, TwinStrand made misrepresentations and omissions material to patentability and did so with the specific intent to mislead or deceive the PTO.

272. The TwinStrand applicants deliberately and improperly copied claims from Guardant's Patent Application WO2015/100427 ("'427 Application") more than one year after the publication of the '427 Application in violation of 35 U.S.C. § 135(b). Under pre-AIA 35 U.S.C. § 135(b)(2), "[a] claim which is the same as, or for the same or substantially the same subject matter as, a claim of an application published under section 122(b) may be made in an application

filed after the application is published only if the claim is made before 1 year after the date on which the application is published." By doing so outside of the one-year timeframe, TwinStrand attempted to evade an interference proceeding.

273. Not only did the TwinStrand applicants violate the timing provisions of § 135(b)(2), but they also never informed the Patent Examiner that the claims were copied.

274. On July 17, 2019, Lara J. Dueppen, a patent attorney, filed U.S. Patent Application No. 16/514,931 ("'931 Application"), which issued as TwinStrand's U.S. Patent No. 10,752,951, which it is now asserting against Guardant. The '931 Application names Jesse Salk, Lawrence Loeb, and Michael Schmitt as inventors. All three inventors submitted declarations pursuant to 37 C.F.R. § 1.76. The '931 Application presented 29 claims. The '931 Application claims priority to Provisional Application No. 61/613,413, filed on March 20, 2012.

275. Claims 1-2 of the '931 Application were almost exact copies of claims 77-78 from Guardant's '427 Application, which was published on July 7, 2015.

276. The inventors and the assignee of the '427 Application are unrelated to Jesse Salk, Lawrence Loeb, and Michael Schmitt.

277. Jesse Salk, Lawrence Loeb, and Michael Schmitt cited the '427 Application at least in their own application, WO2017/100441, published on June 15, 2017. Thus Jesse Salk, Lawrence Loeb, and Michael Schmitt were aware of the '427 Application and its claims two years prior to filing the '931 Application.

278. Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen are experienced before the PTO. Lara J. Dueppen was admitted to the State Bar of California in December 2008 and has been registered as a patent attorney since August 2009. Jesse Salk, Lawrence Loeb, and

Michael Schmitt are each named as inventors on multiple patents and applications for which they have submitted oaths and declarations.

279. Patent applicants and attorneys are subject to the duties of good faith, candor, and disclosure, among others, including a duty to disclose to the United States Patent and Trademark Office ("PTO") all information known to that individual to be material to patentability.

280. 7 C.F.R. § 1.607(c) requires that "[w]hen an applicant [for patent] presents a claim which corresponds exactly or substantially to a claim of a patent, the applicant shall identify the patent and the number of the patent claim. . . ."

281. MPEP § 2001.06(d) (8th ed., rev. 8, July 2010), both in its current incarnation and as it existed at the time the '931 Application was filed, states that "[w]here claims are copied or substantially copied from a patent, 37 C.F.R. 1.607(c) requires applicant shall, at the time he or she presents the claim(s), identify the patent and the numbers of the patent claims..."

282. Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen did not notify the PTO that claims 1-2 of the '931 Application were copied from the '427 Application.

283. 37 C.F.R. § 10.23(c)(7) states that "[k]nowingly withholding from the [PTO] information identifying a patent or patent application of another from which one or more claims have been copied" constitutes disreputable or gross misconduct.

284. MPEP § 2001.06(d) (8th ed., rev. 8, July 2010), both in its current incarnation and as it existed at the time the '931 Application was filed, states that "the information required by 37 CFR 1.607(c) as to the source of copied claims is material information under 37 CFR 1.56. . . ."

285. Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen's failure to notify the PTO that claims 37-39 and 41-43 of the '931 Application were copied from the '427

Application and to identify which claims of the '427 Application were copied, constitutes affirmative egregious misconduct.

286. Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen's failure to notify the PTO of the copied claims was material to the patentability of at least claims 1-2 of the '951 Patent which correspond to originally filed claims 77-78 in the '427 Application, because failure to identify such information to the PTO is per se material to patentability.

287. In addition, Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen's failure to notify the PTO of the copied claims was material to the patentability of at least claims 1-2 of the '951 Patent because identifying the copied claims would have triggered an interference, a much more thorough and in-depth examination as to whether the claims of the '931 Application were entitled to be issued in light of the existence of identical claims in a previously-issued U.S. Patent.

288. On information and belief, the '951 Patent would not have issued if the Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen had notified the PTO of the copied claims.

289. On information and belief, Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen affirmatively chose not to notify the PTO that they copied the claims of the '931 Application from '427 Application with the intent to mislead the PTO. Alternatively, such intent to mislead the PTO is the only reasonable inference from the totality of the circumstances alleged in this counterclaim.

290. This is consistent with Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen's practice of copying claims into their own applications without disclosing that the claims are copied, which has occurred in at least the following instances:

- U.S. Application No. 16/118,290, filed August 30, 2018, included at least one copied claim from U.S. Application No. 2016/0319345, assigned to Illumina Inc. U.S. Application No. 2016/0319345 was cited in an IDS during prosecution of U.S. Application No. 16/118,290.
- U.S. Application No. 16/120,072 (issued as U.S. Patent No. 10,385,393), filed August 31, 2018, included at least one copied claim from U.S. Publication No. 2016/0319345, assigned to Illumina Inc. U.S. Publication No. 2016/0319345 was cited in an IDS during prosecution of U.S. Application No. 16/120,072.
- U.S. Application No. 16/120,091 (issued as U.S. Patent No. 10,370,713), filed August 31, 2018, included at least one copied claim from U.S. Patent No. 9,920,366, assigned to Guardant. U.S. Patent No. 9,920,366 was cited in an IDS after issuance of U.S. Application No. 16/120,091.
- U.S. Application No. 16/411,045 (issued as U.S. Patent No. 10,604,804), filed May 13, 2019, included at least one copied claim from U.S. Patent No. 9,598,731, assigned to Guardant. U.S. Patent No. 9,598,731 was cited in an IDS during prosecution of U.S. Application No. 16/411,045.
- U.S. Application No. 16/411,066 (issued as U.S. Patent No. 10,689,699), filed May 13, 2019, included at least one copied claim from U.S. Patent No. 9,834,822, assigned to Guardant. U.S. Patent No. 9,834,822 was cited in an IDS during prosecution of U.S. Application No. 16/411,066.

291. In at least one instance of claim copying, Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen did notify the PTO of the copied claims.

292. In a December 14, 2020 response to a restriction requirement in U.S. Application No. 16/118,305, Lara J. Dueppen submitted the following statement "Claim Copying. Claims 39

and 57 (newly introduced in a Preliminary Amendment filed on December 10, 2018 in the subject application) were substantially copied from claim 2 of U.S. Patent Appl. Publ. No. 2018/0291438. Applicant hereby specifically directs the Office's attention to claims 39 and 57 that substantially correspond to claim 2 of U.S. Patent. Appl. Publ. No.2018/0291438."

293. After notifying the PTO of the claim copying, no patent issued from U.S. Application No. 16/118,305.

294. Therefore, Jesse Salk, Lawrence Loeb, Michael Schmitt, and Lara J. Dueppen knew they had a duty to disclose claim copying to the PTO and their failure to notify the PTO of the copied claims in the application leading to the '951 Patent constitutes inequitable conduct that renders the '951 Patent unenforceable.

JURY DEMAND

Pursuant to Rule 38(b) of the Federal Rules of Civil Procedure, Guardant demands a trial by jury of all issues presented by Plaintiffs' Complaint and Guardant's counterclaims that are so triable.

PRAYER FOR RELIEF

WHEREFORE, Guardant prays for judgment as follows:

1. A judgment dismissing Plaintiff's Complaint against Guardant with prejudice;

2. A judgement that Guardant has not infringed, contributed to the infringement of, or induced others to infringe, either directly or indirectly, any valid claims of '631 Patent, the '699 Patent, the '951 Patent, and the '127 Patent;

3. A judgment that the '631 Patent, the '699 Patent, the '951 Patent, and the '127 Patent be found invalid;

4. A judgment that the claims of the '699 Patent and '951 Patent be found unpatentable and/or unenforceable;

5. A judgment that the ’063 Patent, the ’858 Patent, the ’221 Patent, and the ’306 Patent be found valid and infringed by TwinStrand;

6. A judgment that TwinStrand’s infringement of the ’063 Patent, the ’858 Patent, the ’221 Patent, and the ’306 Patent has been willful;

7. An award of damages adequate to compensate Guardant for the patent infringements that have occurred, together with pre-judgment interest and costs;

8. An accounting for acts of infringement not presented at trial and/or up to the judgment and an award by the Court of additional damage for any such acts of infringement;

9. A permanent injunction against TwinStrand, its affiliates, subsidiaries, officers directors, agents, and employees from further infringement, or alternatively, an award an ongoing royalty for TwinStrand’s post-verdict infringement, payable on each product or service offered by TwinStrand that is found to infringe one or more of the patents asserted herein, and on all future products and services that are not substantially different from those found to infringe;

10. An award of all other damages permitted by 35 U.S.C. § 284, including increased damages of three times the amount of compensatory damages found;

11. A judgment that the Court find this case exceptional under 35 U.S.C. § 285 and order an award to Guardant of its fees and expenses of litigation; and

12. Such other and further relief as this Court may deem just and proper.

Dated: January 11, 2022

Respectfully submitted,

DLA PIPER LLP (US)

/s/ Brian A. Biggs

Brian A. Biggs (DE Bar No. 5591)
Erin E. Larson (DE Bar No. 6616)
1201 North Market Street, Suite 2100
Wilmington, DE 19801
Telephone: (302) 468-5700
Facsimile: (302) 394-2341
brian.biggs@us.dlapiper.com
erin.larson@us.dlapiper.com

Attorneys for Defendant and Counterclaim Plaintiff Guardant Health, Inc.

OF COUNSEL:

Mark Fowler *(*admitted *pro hac vice*)
Susan Krumplitsch *(*admitted *pro hac vice*)
Monica De Lazzari *(*admitted *pro hac vice*)
DLA PIPER LLP (US)
2000 University Avenue
East Palo Alto, California 94303-2214
Telephone: 650-833-2048
Facsimile: 650-687-1166
mark.fowler@us.dlapiper.com
susan.krumplitsch@us.dlapiper.com
monica.delazzari@us.dlapiper.com

Michael A. Sitzman *(*admitted *pro hac vice*)
DLA PIPER LLP (US)
555 Mission Street
Suite 2400
San Francisco, CA 94105-2933
Telephone: 415-615-6175
Facsimile: 415-659-7411
michael.sitzman@us.dlapiper.com

Ellen Scordino *(*admitted *pro hac vice*)
Kristin Beale *(*admitted *pro hac vice*)
DLA PIPER LLP (US)
33 Arch Street
26th Floor
Boston, MA 02110-1447
Telephone: 617-406-6085
ellen.scordino@us.dlapiper.com
kristin.beale@us.dlapiper.com

EXHIBIT A

US010801063B2

(12) **United States Patent**
Eltoukhy et al.

(10) **Patent No.: US 10,801,063 B2**
(45) **Date of Patent: *Oct. 13, 2020**

(54) **METHODS AND SYSTEMS FOR DETECTING GENETIC VARIANTS**

(71) Applicant: **GUARDANT HEALTH, INC.**, Redwood City, CA (US)

(72) Inventors: **Helmy Eltoukhy**, Atherton, CA (US); **AmirAli Talasaz**, Atherton, CA (US); **Stefanie Ann Ward Mortimer**, Morgan Hill, CA (US)

(73) Assignee: **GUARDANT HEALTH, INC.**, Redwood City, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **16/601,168**

(22) Filed: **Oct. 14, 2019**

(65) **Prior Publication Data**

US 2020/0123602 A1 Apr. 23, 2020

Related U.S. Application Data

(63) Continuation of application No. 15/892,178, filed on Feb. 8, 2018, which is a continuation of application No. 14/861,989, filed on Sep. 22, 2015, now Pat. No. 9,920,366, which is a continuation of application No. PCT/US2014/072383, filed on Dec. 24, 2014.

(60) Provisional application No. 61/948,509, filed on Mar. 5, 2014, provisional application No. 61/921,456, filed on Dec. 28, 2013.

(51) **Int. Cl.**

C12P 19/34	(2006.01)
C12Q 1/6869	(2018.01)
C12Q 1/6886	(2018.01)
G16B 15/00	(2019.01)

(52) **U.S. Cl.**

CPC ***C12Q 1/6869*** (2013.01); ***C12Q 1/6886*** (2013.01); *C12Q 2535/122* (2013.01); *C12Q 2600/158* (2013.01); *G16B 15/00* (2019.02)

(58) **Field of Classification Search**

None

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,725,536	A	2/1988	Fritsch et al.
4,942,124	A	7/1990	Church
5,124,246	A	6/1992	Urdea et al.
5,149,625	A	9/1992	Church et al.
5,200,314	A	4/1993	Urdea
5,424,186	A	6/1995	Fodor et al.
5,424,413	A	6/1995	Hogan et al.
5,445,934	A	8/1995	Fodor et al.
5,604,097	A	2/1997	Brenner
5,635,352	A	6/1997	Urdea et al.
5,635,400	A	6/1997	Brenner
5,648,245	A	7/1997	Fire et al.
5,654,413	A	8/1997	Brenner
5,656,731	A	8/1997	Urdea
5,658,737	A	8/1997	Nelson et al.
5,714,330	A	2/1998	Brenner et al.
5,744,305	A	4/1998	Fodor et al.
5,759,778	A	6/1998	Li et al.
5,763,175	A	6/1998	Brenner
5,800,992	A	9/1998	Fodor et al.
5,846,719	A	12/1998	Brenner et al.
5,854,033	A	12/1998	Lizardi
5,871,928	A	2/1999	Fodor et al.
5,925,525	A	7/1999	Fodor et al.
5,935,793	A	8/1999	Wong
5,952,170	A	9/1999	Stroun et al.
5,968,740	A	10/1999	Fodor et al.
5,981,176	A	11/1999	Wallace
5,981,179	A	11/1999	Lorinez et al.
6,013,445	A	1/2000	Albrecht et al.
6,020,124	A	2/2000	Sorenson
6,040,138	A	3/2000	Lockhart et al.
6,046,005	A	4/2000	Ju et al.
6,060,596	A	5/2000	Lerner et al.
6,117,631	A	9/2000	Nilsen
6,124,092	A	9/2000	O'Neill et al.
6,138,077	A	10/2000	Brenner
6,140,489	A	10/2000	Brenner
6,172,214	B1	1/2001	Brenner
6,197,506	B1	3/2001	Fodor et al.
6,235,475	B1	5/2001	Brenner et al.
6,242,186	B1	6/2001	Salonen
6,268,152	B1	7/2001	Fodor et al.
6,284,460	B1	9/2001	Fodor et al.
6,300,077	B1	10/2001	Shuber et al.
6,309,822	B1	10/2001	Fodor et al.
6,309,823	B1	10/2001	Cronin et al.
6,326,148	B1	12/2001	Pauletti et al.

(Continued)

FOREIGN PATENT DOCUMENTS

CN	102933721	A	2/2013
EP	0799897	B1	6/2006

(Continued)

OTHER PUBLICATIONS

Daser, et al. "Interrogation of genomes by molecular copy-number counting (MCC)." Nature Methods, 3(6): 447-453 (2006).

(Continued)

Primary Examiner — Kenneth R Horlick
(74) *Attorney, Agent, or Firm* — Timothy Hott

(57) **ABSTRACT**

Disclosed herein in are methods and systems for determining genetic variants (e.g., copy number variation) in a polynucleotide sample. A method for determining copy number variations includes tagging double-stranded polynucleotides with duplex tags, sequencing polynucleotides from the sample and estimating total number of polynucleotides mapping to selected genetic loci. The estimate of total number of polynucleotides can involve estimating the number of double-stranded polynucleotides in the original sample for which no sequence reads are generated. This number can be generated using the number of polynucleotides for which reads for both complementary strands are detected and reads for which only one of the two complementary strands is detected.

28 Claims, 11 Drawing Sheets

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,355,431 B1 3/2002 Chee et al.
6,355,432 B1 3/2002 Fodor et al.
6,395,491 B1 5/2002 Fodor et al.
6,406,848 B1 6/2002 Bridgham et al.
6,440,667 B1 8/2002 Fodor et al.
6,440,706 B1 8/2002 Vogelstein et al.
6,451,536 B1 9/2002 Fodor et al.
6,458,530 B1 10/2002 Morris et al.
6,468,744 B1 10/2002 Cronin et al.
6,489,114 B2 12/2002 Laayoun et al.
6,492,121 B2 12/2002 Kurane et al.
6,498,012 B2 12/2002 Laken
6,503,718 B2 1/2003 Shuber et al.
6,512,105 B1 1/2003 Hogan et al.
6,514,699 B1 2/2003 O'Neill et al.
6,544,739 B1 4/2003 Fodor et al.
6,551,784 B2 4/2003 Fodor et al.
6,576,424 B2 6/2003 Fodor et al.
6,582,908 B2 6/2003 Fodor et al.
6,586,177 B1 7/2003 Shuber
6,600,996 B2 7/2003 Webster et al.
6,629,040 B1 9/2003 Goodlett et al.
6,653,077 B1 11/2003 Brenner
6,753,147 B2 6/2004 Vogelstein et al.
6,849,403 B1 2/2005 Shuber
6,849,404 B2 2/2005 Park et al.
6,852,488 B2 2/2005 Fodor et al.
6,858,412 B2 2/2005 Willis et al.
6,964,846 B1 11/2005 Shuber
7,163,789 B2 1/2007 Chen et al.
7,208,275 B2 4/2007 Gocke et al.
7,406,385 B2 7/2008 Sorenson
7,410,764 B2 8/2008 Gocke et al.
7,424,368 B2 9/2008 Huang et al.
7,424,371 B2 9/2008 Kamentsky
7,537,897 B2 5/2009 Brenner et al.
7,700,286 B2 4/2010 Stroun et al.
7,704,687 B2 4/2010 Wang et al.
7,727,720 B2 6/2010 Dhallan
7,803,929 B2 9/2010 Melkonyan et al.
7,811,757 B2 10/2010 Shuber
7,822,555 B2 10/2010 Huang et al.
7,824,889 B2 11/2010 Vogelstein et al.
7,838,647 B2 11/2010 Hahn et al.
7,915,015 B2 3/2011 Vogelstein et al.
7,935,487 B2 5/2011 Gocke et al.
7,937,225 B2 5/2011 Mishra et al.
7,957,913 B2 6/2011 Chinitz et al.
7,972,817 B2 7/2011 Kopreski
7,981,612 B2 7/2011 Shuber et al.
8,094,312 B2 1/2012 Ulmer
8,168,385 B2 5/2012 Brenner
8,195,415 B2 6/2012 Fan et al.
8,209,130 B1 6/2012 Kennedy et al.
8,216,789 B2 7/2012 Disis et al.
8,236,532 B2 8/2012 Ronaghi et al.
8,361,726 B2 1/2013 Gocke et al.
8,383,338 B2 2/2013 Kitzman et al.
8,383,345 B2 2/2013 Shendure et al.
8,455,193 B2 6/2013 Travers et al.
8,470,996 B2 6/2013 Brenner
8,481,292 B2 7/2013 Casbon et al.
8,603,749 B2 12/2013 Gillevet
8,685,678 B2 4/2014 Casbon et al.
8,697,408 B2 4/2014 Kucera et al.
8,704,165 B2 4/2014 Huang
8,715,967 B2 5/2014 Casbon et al.
8,722,368 B2 5/2014 Casbon et al.
8,728,766 B2 5/2014 Casbon et al.
8,741,606 B2 6/2014 Casbon et al.
8,775,092 B2 7/2014 Colwell et al.
8,835,358 B2 9/2014 Fodor et al.
8,865,410 B2 10/2014 Shendure et al.
9,018,365 B2 4/2015 Brenner
9,085,798 B2 7/2015 Chee
9,260,753 B2 2/2016 Xie et al.
9,340,830 B2 5/2016 Lipson et al.
9,376,719 B2 6/2016 Eijk et al.
9,404,156 B2 8/2016 Hicks et al.
9,476,095 B2 10/2016 Vogelstein et al.
9,598,731 B2 3/2017 Talasaz
9,752,188 B2 9/2017 Schmitt et al.
9,834,822 B2 12/2017 Talasaz
9,840,743 B2 12/2017 Talasaz
9,920,366 B2 3/2018 Eltoukhy et al.
10,287,631 B2 5/2019 Salk et al.
10,370,713 B2 8/2019 Salk et al.
10,385,393 B2 8/2019 Salk et al.
10,388,403 B2 8/2019 Rava et al.
10,604,804 B2 3/2020 Salk et al.
2001/0053519 A1 12/2001 Fodor et al.
2002/0072058 A1 6/2002 Voelker et al.
2003/0003490 A1 1/2003 Fan et al.
2003/0049616 A1 3/2003 Brenner et al.
2003/0104436 A1 6/2003 Morris et al.
2003/0152490 A1 8/2003 Trulson et al.
2003/0165978 A1 9/2003 Firth et al.
2003/0186251 A1 10/2003 Dunn et al.
2003/0207300 A1 11/2003 Matray et al.
2004/0096892 A1 5/2004 Wang et al.
2004/0146901 A1 7/2004 Morris et al.
2004/0157243 A1 8/2004 Huang et al.
2004/0209299 A1 10/2004 Pinter et al.
2004/0259118 A1 12/2004 Macevicz
2005/0164241 A1 7/2005 Hahn et al.
2005/0221314 A1 10/2005 Berlin et al.
2005/0250147 A1 11/2005 Macevicz
2006/0035258 A1 2/2006 Tadakamalla et al.
2006/0046258 A1 3/2006 Lapidus et al.
2006/0073506 A1 4/2006 Christians et al.
2006/0211030 A1 9/2006 Brenner
2007/0020640 A1 1/2007 McCloskey et al.
2007/0065823 A1 3/2007 Dressman et al.
2007/0111233 A1 5/2007 Bianchi et al.
2007/0128724 A1 6/2007 Miles et al.
2007/0172839 A1 7/2007 Smith et al.
2007/0172873 A1 7/2007 Brenner et al.
2008/0014146 A1 1/2008 Hoff et al.
2008/0124721 A1 5/2008 Fuchs et al.
2008/0161420 A1 7/2008 Shuber
2008/0261204 A1 10/2008 Lexow
2008/0293055 A1 11/2008 Freeman et al.
2009/0029377 A1 1/2009 Lo et al.
2009/0087847 A1 4/2009 Lo et al.
2009/0098547 A1 4/2009 Ghosh
2009/0105959 A1 4/2009 Braverman et al.
2009/0162836 A1 6/2009 Widschwendter
2009/0298075 A1 12/2009 Travers et al.
2009/0298709 A1 12/2009 Ma
2009/0317818 A1 12/2009 Ehrich et al.
2010/0041048 A1 2/2010 Diehl et al.
2010/0062494 A1 3/2010 Church et al.
2010/0069250 A1 3/2010 White et al.
2010/0112575 A1 5/2010 Fan et al.
2010/0143932 A1 6/2010 Lapidus
2010/0166744 A1 7/2010 Wong
2010/0196898 A1 8/2010 Sugarbaker et al.
2010/0264331 A1* 10/2010 Sacko G01N 21/6428 250/459.1
2010/0323348 A1 12/2010 Hamady et al.
2010/0330571 A1 12/2010 Robins et al.
2011/0014607 A1 1/2011 Jirtle et al.
2011/0160078 A1 6/2011 Fodor et al.
2011/0160290 A1 6/2011 Tewari
2011/0171640 A1 7/2011 Bhatt et al.
2011/0177512 A1 7/2011 Shuber
2011/0183330 A1 7/2011 Lo et al.
2011/0201507 A1 8/2011 Rava et al.
2011/0230358 A1 9/2011 Rava
2011/0230360 A1 9/2011 Stephan et al.
2011/0245482 A1 10/2011 Hahn et al.
2011/0264376 A1 10/2011 Chinitz et al.
2011/0275084 A1 11/2011 Byron et al.
2011/0281736 A1 11/2011 Drmanac et al.

(56) **References Cited**

U.S. PATENT DOCUMENTS

2011/0319272 A1 12/2011 Fan et al.
2012/0003637 A1 1/2012 Lo et al.
2012/0034685 A1 2/2012 Sparks et al.
2012/0046877 A1 2/2012 Hyland et al.
2012/0053073 A1 3/2012 Kassis
2012/0059594 A1 3/2012 Hatchwell et al.
2012/0065081 A1 3/2012 Chee
2012/0095697 A1 4/2012 Halpern et al.
2012/0100548 A1 4/2012 Rava et al.
2012/0164630 A1 6/2012 Porreca et al.
2012/0165202 A1 6/2012 Porreca et al.
2012/0191367 A1 7/2012 Stuelpnagel et al.
2012/0208706 A1 8/2012 Downing et al.
2012/0214163 A1 8/2012 Sugarbaker et al.
2012/0214678 A1 8/2012 Rava et al.
2012/0220478 A1 8/2012 Shaffer
2012/0231479 A1 9/2012 Puskas et al.
2012/0238464 A1 9/2012 Koi et al.
2012/0270212 A1 10/2012 Rabinowitz et al.
2012/0316074 A1 12/2012 Saxonov
2012/0322681 A1 12/2012 Kung et al.
2013/0005585 A1 1/2013 Anderson et al.
2013/0017958 A1 1/2013 Benz et al.
2013/0022977 A1 1/2013 Lapidus et al.
2013/0029852 A1 1/2013 Rava et al.
2013/0034546 A1 2/2013 Rava et al.
2013/0040824 A1 2/2013 Lo et al.
2013/0053256 A1 2/2013 Hubbell
2013/0060483 A1 3/2013 Struble et al.
2013/0078626 A1 3/2013 Wasserstrom et al.
2013/0085681 A1 4/2013 Deciu et al.
2013/0102485 A1 4/2013 Lee
2013/0102487 A1 4/2013 Cos et al.
2013/0116127 A1 5/2013 Schuetz et al.
2013/0116130 A1 5/2013 Fu et al.
2013/0122499 A1 5/2013 Morris et al.
2013/0130923 A1 5/2013 Ehrich et al.
2013/0137588 A1 5/2013 Shendure et al.
2013/0143747 A1 6/2013 Gutin et al.
2013/0210643 A1 8/2013 Casbon et al.
2013/0210645 A1 8/2013 Volgelstein et al.
2013/0224743 A1 8/2013 Casbon et al.
2013/0237431 A1 9/2013 Lo et al.
2013/0237458 A1 9/2013 Casbon et al.
2013/0260381 A1 10/2013 Ramakrishnan
2013/0267424 A1 10/2013 Casbon et al.
2013/0288244 A1 10/2013 Deciu et al.
2013/0338933 A1 12/2013 Deciu et al.
2014/0011694 A1 1/2014 Couronne
2014/0065609 A1 3/2014 Hicks et al.
2014/0065630 A1 3/2014 Bubnoff et al.
2014/0066317 A1 3/2014 Talasaz
2014/0100121 A1 4/2014 Lo et al.
2014/0242588 A1 8/2014 Boom et al.
2014/0274731 A1 9/2014 Raymond et al.
2014/0274740 A1 9/2014 Srinivasan et al.
2014/0296094 A1 10/2014 Domanus
2014/0303008 A1 10/2014 Schutz et al.
2014/0336943 A1 11/2014 Pellini et al.
2014/0350130 A1 11/2014 Sanborn et al.
2014/0357500 A1 12/2014 Vigneault et al.
2015/0004158 A1 1/2015 Shipp et al.
2015/0024950 A1 1/2015 Bielas et al.
2015/0044687 A1 2/2015 Schmitt et al.
2015/0050647 A1 2/2015 Luo et al.
2015/0051085 A1 2/2015 Vogelstein et al.
2015/0065358 A1 3/2015 Comstock et al.
2015/0080266 A1 3/2015 Volkmuth et al.
2015/0087535 A1 3/2015 Patel
2015/0167069 A1 6/2015 Schutz et al.
2015/0197786 A1 7/2015 Osborne et al.
2015/0275289 A1 10/2015 Otwinowski et al.
2015/0299812 A1 10/2015 Talasaz
2015/0329917 A1 11/2015 Shuber
2015/0344970 A1 12/2015 Vogelstein et al.
2015/0366866 A1 12/2015 Ali et al.
2015/0368708 A1 12/2015 Talasaz
2016/0002739 A1 1/2016 Schütz et al.
2016/0002741 A1 1/2016 Kitano et al.
2016/0024576 A1 1/2016 Chee
2016/0026758 A1 1/2016 Jabara et al.
2016/0032396 A1 2/2016 Diehn et al.
2016/0040229 A1 2/2016 Talasaz et al.
2016/0046986 A1 2/2016 Eltoukhy et al.
2016/0053301 A1 2/2016 Raymond et al.
2016/0060691 A1 3/2016 Giresi et al.
2016/0071432 A1 3/2016 Kurowski et al.
2016/0115553 A1 4/2016 Stephan et al.
2016/0130649 A1 5/2016 Xie et al.
2016/0251704 A1 9/2016 Talasaz et al.
2016/0319345 A1 11/2016 Gnerre et al.
2016/0333417 A1 11/2016 Talasaz
2016/0376647 A1 12/2016 Travers et al.
2017/0051347 A1 2/2017 Vogelstein et al.
2017/0073774 A1 3/2017 Lo et al.
2017/0145516 A1 5/2017 Kopetz et al.
2017/0159120 A1 6/2017 Eijk et al.
2017/0211143 A1 7/2017 Shendure et al.
2017/0218459 A1 8/2017 Talasaz et al.
2017/0218460 A1 8/2017 Talasaz
2018/0023125 A1 1/2018 Talasaz et al.
2019/0271040 A1 9/2019 Salk et al.
2019/0292597 A1 9/2019 Salk et al.
2019/0338358 A1 11/2019 Salk et al.
2019/0352714 A1 11/2019 Salk et al.

FOREIGN PATENT DOCUMENTS

EP 1647600 A2 6/2006
EP 2110442 A1 10/2009
EP 3070177 A1 9/2016
EP 3178941 A1 6/2017
WO 1997007241 A1 2/1997
WO 1997010365 A1 3/1997
WO 1999028505 A1 6/1999
WO 2000058516 A2 10/2000
WO 2002056014 A2 7/2002
WO 2005080604 A2 9/2005
WO 2005111242 A2 11/2005
WO 2006102264 A1 9/2006
WO 2007037678 A2 4/2007
WO 2008070144 A2 6/2008
WO 2008154317 A1 12/2008
WO 2009152928 A2 12/2009
WO 2011060240 A1 5/2011
WO 2011087760 A2 7/2011
WO 2011091046 A1 7/2011
WO 2011103236 A2 8/2011
WO 2011140510 A2 11/2011
WO 2011155833 A2 12/2011
WO 2012012693 A2 1/2012
WO 2012014877 A1 2/2012
WO 2012019200 A2 2/2012
WO 2012028746 A1 3/2012
WO 2012038839 A2 3/2012
WO 2012042374 A2 4/2012
WO 2012048341 A1 4/2012
WO 2012054873 A2 4/2012
WO 2012066451 A1 5/2012
WO 2012071621 A1 6/2012
WO 2012088348 A2 6/2012
WO 2012106559 A1 8/2012
WO 2012129363 A2 9/2012
WO 2012142213 A2 10/2012
WO 2012142611 A2 10/2012
WO 2012103031 A3 1/2013
WO 2013019075 A2 2/2013
WO 2013033721 A1 3/2013
WO 2013106737 A1 7/2013
WO 2013123442 A1 8/2013
WO 2013130512 A2 9/2013
WO 2013130674 A1 9/2013
WO 2013138510 A1 9/2013
WO 2013142213 A1 9/2013

(56) **References Cited**

FOREIGN PATENT DOCUMENTS

WO	2013142389	A1	9/2013
WO	2013148496	A1	10/2013
WO	2013159035	A2	10/2013
WO	2013173394	A2	11/2013
WO	2013181170	A1	12/2013
WO	2013188471	A2	12/2013
WO	2013190441	A2	12/2013
WO	2014004726	A1	1/2014
WO	2014014497	A1	1/2014
WO	2014015319	A1	1/2014
WO	2014039556	A1	3/2014
WO	2014093330	A1	6/2014
WO	2014145078	A1	9/2014
WO	2014149134	A2	9/2014
WO	2014151117	A1	9/2014
WO	2014152990	A1	9/2014
WO	2015159293	A2	10/2015
WO	2016015058	A2	1/2016
WO	2016040901	A1	3/2016
WO	2017100441	A1	6/2017
WO	2017181146	A1	10/2017

OTHER PUBLICATIONS

De Saizieu, et al. "Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays." Nature Biotechnology, 16: 45-48 (1998).

Diehl et al. Circulating mutant DNA to assess tumor dynamics. Nat Med 14(9):985-990 (2008).

Diehl, et al. Detection and quantification of mutations in the plasma of patients with colorectal tumors. Proc Natl Acad Sci US A. Nov. 8, 2005;102(45):16368-73. Epub Oct. 28, 2005.

Diehl, F. et al., "Analysis of Mutations in DNA Isolated From Plasma and Stool of Colorectal Cancer Patients," Gastroenterology (2008) 135(2):489-498.

Ding, L. et al. "Clonal evolution in relapsed acute myeloid leukemia revealed by whole-genome sequencing" Nature (2012) 481(7382):506-510.

Ding, L. et al., "Analysis of next-generation genomic data in cancer: accomplishments and challenges," Hum. Mol. Genet. 2010 19(R2), R188-R196.

Duncan, D.L. et al. "Next-Generation Sequencing in the Clinical Laboratory" Diagnostic Molecular Pathology: A Guide to Applied Molecular Testing 25-33 (Coleman and Tsongalis eds., 1st ed. 2016).

Ehrich, et al. Noninvasive detection of fetal trisomy 21 by sequencing of DNA in maternal blood: a study in a clinical setting. Am J Obstet Gynecol. Mar. 2011;204(3):205.e1-11. doi: 10.1016/j.ajog.2010.12.060. Epub Feb. 18, 2011.

Eisenmann, et al. 5q-myelodysplastic syndromes: chromosome 5q genes direct a tumor-suppression network sensing actin dynamics. Oncogene. Oct. 1, 2009;28(39):3429-41. doi: 10.1038/onc.2009.207. Epub Jul. 13, 2009.

Elshire, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. May 4, 2011;6(5):e19379. doi: 10.1371/journal.pone.0019379.

European search report and search opinion dated Mar. 29, 2016 for EP Application No. 13834427.0.

Extended European search report and opinion dated May 4, 2017 for EP Application No. 14874157.2.

Extended European search report and opinion dated Nov. 21, 2016 for EP Application No. 14771159.

Fan, et al. "Microfluidic digital PCR enables rapid prenatal diagnosis of fetal aneuploidy." Am Obstet Gynecol. 2009; 200:541e1-543.e7.

Fan, et al. "Parallel Genotyping of Human SNPs Using Generic High-density Oligonucleotide Tag Arrays." Genome Research, 10: 853-860 (2000).

Fan, et al. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci U S A. Oct. 21, 2008;105(42):16266-71. Epub Oct. 6, 2008.

Fan, et al. Non-invasive prenatal measurement of the fetal genome. Nature. Jul. 19, 2012;487(7407):320-4. doi: 10.1038/nature11251.

Fleischhacker, M. et al. "Circulating nucleic acids (CNAs) and cancer—A survey" Biochimica et Biophysica Acta (2007) 1775:181-232.

Fonatsch, C. The role of chromosome 21 in hematology and oncology. Genes Chromosomes Cancer. Jun. 2010;49(6)497-508. doi: 10.1002/gcc.20764.

Fong, S.L. et al. "Comparison of 7 Methods for Extracting Cell-Free DNA from Serum Samples of Colorectal Cancer Patients" Clinical Chem (2009) 55(3):587-598.

Forshew, T. et al. "Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA" Sci Transl Med (2012) 4(136) ra68.

Forster, et al. From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software. Nucleic Acids Res. Jan. 7, 2013;41(1):e16. doi: 10.1093/nar/gks836. Epub Sep. 10, 2012.

Fournie, et al. Plasma DNA as a marker of cancerous cell death. Investigations in patients suffering from lung cancer and in nude mice bearing human tumours. Cancer Lett. May 8, 1995;91(2):221-7.

Freeman, et al. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. Oct. 2009;19(10):1817-24. doi: 10.1101/gr.092924.109. Epub Jun. 18, 2009.

Fu, et al. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc Natl Acad Sci U S A. May 31, 2011;108(22):9026-31. Epub May 11, 2011.

Fujiwara, et al. Identification of epigenetic aberrant promoter methylation in serum DNA is useful for early detection of lung cancer Clin Cancer Res. Feb. 1, 2005;11(3):1219-25.

Fujiwara, et al. Plasma DNA microsatellites as tumor-specific markers and indicators of tumor progression in melanoma patients. Cancer Res. Apr. 1, 1999;59(7):1567-71.

Genome Analyzer IIx Systems Specification Sheet (2009).

Gerry, et al. Universal DNA microarray method for multiplex detection of low abundance point mutations. Journal of Molecular Biology, 292(2): 251-262 (1999).

Gillespie. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry, 81(25): 2340-2361 (1977).

Gordian, et al. Serum free circulating DNA is a useful biomarker to distinguish benign versus malignant prostate disease. Cancer Epidemiol Biomarkers Prev. Aug. 2010;19(8):1984-91. doi: 10.1158/1055-9965.EPI-10-0287. Epub Jul. 20, 2010.

Gordon, D.J. et al., "Causes and consequences of aneuploidy in cancer," Nat. Rev. Genet. 2012 13(3), 189-203.

Gormally, et al. Amount of DNA in plasma and cancer risk: a prospective study. Int J Cancer. Sep. 20, 2004;111(5):746-9.

Grant, et al. SNP genotyping on a genome-wide amplified DOP-PCR template. Nucleic Acids Res. Nov. 15, 2002;30(22):e25.

Gregory et al. "Targeted Single Molecule Mutation Detection with Massively Parallel Sequencing" Nucleic Acids Research (2015) 44(3):2-11.

Grutzmann, et al. Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. PLoS One. 2008;3(11):e3759. doi: 10.1371/journal.pone.0003759. Epub Nov. 19, 2008.

Guardant Health vs. *FMI* 1st Amended Answer to Second Amended Complaint dated May 6, 2019 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Contentions dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit A-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit A-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit A-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit A-8, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

(56) **References Cited**

OTHER PUBLICATIONS

Guardant Health vs. *FMI* Invalidity Exhibit B-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit B-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit B-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit B-7, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit C-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit C-11, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit C-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Pinkel, et al. Comparative Genomic Hybridization. Annual Review of Genomics and Human Genetics, 6: 331-354 (2005).
Pleasance, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. Jan. 14, 2010;463(7278)184-90. doi: 10.1038/nature08629. Epub Dec. 16, 2009.
Pollack, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet. Sep. 1999;23(1):41-6.
Qiu, et al. DNA sequence-based "bar codes" for tracking the origins of expressed sequence tags from a maize cDNA library constructed using multiple mRNA sources. Plant Physiol. Oct. 2003;133(2):475-81.
Quail, et al. A large genome center's improvements to the Illumina sequencing system. Nat Methods. Dec. 2008; 5(12):1005-10. doi: 10.1038/nmeth.1270.
Quinlan, A.R. et al., "Pyrobayes: an improved base caller for SNP discovery in pyrosequences," Nat. Methods 2008 5(2), 179-181.
Rafi, et al. "Cell-free fetal DNA and non-invasive prenatal diagnosis," Br. J. Gen. Pract. May 1, 2009; 59(562):e146-e148.
Rand, et al. Headloop suppression PCR and its application to selective amplification of methylated DNA sequences. Nucleic Acids Res. Aug. 9, 2005;33(14):e127.
Redon, R. et al., "Global variation in copy number in the human genome," Nature 2006 444(7118), 444-454.
Report and Recommendation in Civil Action No. 17-1623-LPS-CJB between Guardant Health, Inc. and Foundation Medicine and between Guardant Health, Inc. and Personal Genome Diagnostics, Inc., dated Oct. 11, 2019.
Rizzo, J.M. et al. "Key Principles and Clinical Applications of 'Next Generation' DNA Sequencing," Cancer Prev. Res., (2012) 5, 887-900.
Ryan, et al. A prospective study of circulating mutant KRAS2 in the serum of patients with colorectal neoplasia: strong prognostic indicator in postoperative follow up. Gut. Jan. 2003;52(1):101-8.
Rygaard, et al. Abnormalities in structure and expression of the retinoblastoma gene in small cell lung cancer cell lines and xenografts in nude mice. Cancer Res. Sep. 1, 1990;50(17):5312-7.
Sanger, F. et al. "DNA sequencing with chain-terminating inhibitors" PNAS (1977) 74(12):5463-5467.
Sausen, M. et al. "Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma", Nature Genetics 2013, 45(1), 12-17.
Schmitt et al. Supplemental Information http://www.pnas.org/content/suppl/2012/08/01/1208715109. DCSupplemental.
Schmitt, et al. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A. Sep. 4, 2012;109(36)14508-13. doi: 10.1073/pnas.1208715109. Epub Aug. 1, 2012.
Schwarzenbach, et al. A critical evaluation of loss of heterozygosity detected in tumor tissues, blood serum and bone marrow plasma from patients with breast cancer. Breast Cancer Res. 2007;9(5):R66.
Schwarzenbach, et al. Cell-free tumor DNA in blood plasma as a marker for circulating tumor cells in prostate cancer. Clin Cancer Res. Feb. 1, 2009;15(3):1032-8. doi: 10.1158/1078-0432.CCR-08-1910.
Schwarzenback, H. et al. "Cell-free nucleic acids as biomarkers in cancer patients" Nature Reviews Cancer (2011) 11:426-437.
Schweiger et al., "Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis," PLoS One 2009, 4(5), e5548.
Sebat, et al. Large-Scale Copy Number Polymorphism in the Human Genome. Science, 305: 525-528 (2004).
Sehnert, A.J. et al. "Optimal Detection of Fetal Chromosomal Abnormalities by Massively Parallel DNA Sequencing of Cell-Free Fetal DNA from Maternal Blood" Clin Chem (2011) 57(7):1042-1049.
Shaw et al. Genomic analysis of circulating cell-free DNA infers breast cancer dormancy. Genome Research 22 (2):220-231 (Feb. 2012).
Shaw, et al. Microsatellite alterations plasma DNA of primary breast cancer patients. Clin Cancer Res. Mar. 2000;6(3)1119-24.
Shendure, J. et al. "Next-generation DNA sequencing," Nat. Biotechnol. 2008 26(10), 1135-1145.
Shinozaki, et al. Utility of circulating B-RAF DNA mutation in serum for monitoring melanoma patients receiving biochemotherapy. Clin Cancer Res. Apr. 1, 2007;13(7)2068-74.
Shiroguchi, et al. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci U S A. Jan. 24, 2012;109(4):1347-52. doi: 10.1073/pnas.1118018109. Epub Jan. 9, 2012.
Shoemaker, et al. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. Nature Genetics, 14: 450-456 (1996).
Simpson, et al. Copy number variant detection in inbred strains from short read sequence data. Bioinformatics. Feb. 15, 2010;26(4):565-7. doi: 10.1093/bioinformatics/btp693. Epub Dec. 18, 2009.
Smith, et al. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Research, 38(13): e142 (2010).
Smith, T.F. et al. "Comparison of Biosequences" Adv App Math (1981) 2:482-489.
Sorenson, et al. Soluble normal and mutated DNA sequences from single-copy genes in human blood. Cancer Epidemiol Biomarkers Prev. Jan.-Feb. 1994;3(1):67-71.
Sparks, et al. Selective analysis of cell-free DNA in maternal blood for evaluation of fetal trisomy. Prenat Diagn. Jan. 2012;32(1):3-9. doi: 10.1002/pd.2922. Epub Jan. 6, 2012.
Stein, et al. "The case for cloud computing in genome informatics", Genome Biol. 2010; 11 (5):207. Epub May 5, 2010.
Steinman. Free DNA in serum and plasma from normal adults. J Clin Invest. Aug. 1975;56(2):512-5.
Stitziel, et al. Computational and statistical approaches to analyzing variants identified byexome sequencing. Genome Biol. Sep. 14, 2011;12(9):227. doi: 10.1186/gb-2011-12-9-227.
Stroun, M, et al, "About the possible orgin and mechanism of circulating DNA apoptosis and active DNA release", Clin Chim Acta, vol. 313, No. 1-2, pp. 139-142, (2001).
Stumm, M. et al. "Noninvasive prenatal detection of chromosomal aneuploidies using different next generation sequencing strategies and algorithms" Prenatal Diagnosis (2012) 32:569-577.
Taback, et al. Detection of tumor-specific genetic alterations in bone marrow from early-stage breast cancer patients. Cancer Res. Apr. 15, 2003;63(8):1884-7.
Tan, et al. Genome-wide comparison of DNA hydroxymethylation in mouse embryonic stem cells and neural progenitor cells by a new comparative hMeDIP-seq method. Nucleic Acids Res. Apr. 2013;41(7):e84. doi: 10.1093/nar/gkt091. Epub Feb. 13, 2013.
Taudien, et al. Haplotyping and copy number estimation of the highly polymorphic human beta-defensin locus on 8p23 by 454 amplicon sequencing. BMC Genomics. Apr. 19, 2010;11:252. doi: 10.1186/1471-2164-11-252.
Teer, J.K. et al. "Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing" Genome Res (2010) 20(10):1420-1431.
Tie, J. et al., "Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer," Sci. Transl. Med. 2016, 8(346):346ra92.

(56) **References Cited**

OTHER PUBLICATIONS

Tomaz, et al. Differential methylation as a cause of allele dropout at the imprinted GNAS locus. Genet Test Mol Biomarkers. Aug. 2010;14(4):455-60. doi: 10.1089/gtmb.2010.0029.

Tomlinson, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet. Aug. 2007;39(8):984-8. Epub Jul. 8, 2007.

Tsai, et al. Discovery of rare mutations in populations: TILLING by sequencing. Plant Physiol. Jul. 2011;156(3):1257-68. doi: 10.1104/pp.110.169748. Epub Apr. 29, 2011.

U.S. Appl. No. 14/102,285, filed Dec. 10, 2013.

U.S. Appl. No. 61/613,413 ("Schmitt '413 provisional") filed Mar. 20, 2012.

U.S. Appl. No. 61/625,319 ("Schmitt '319 provisional") filed Mar. 20, 2012.

Koyanagi, et al. Association of circulating tumor cells with serum tumor-related methylated DNA in peripheral blood of melanoma patients. Cancer Res. Jun. 15, 2006;66(12):6111-7.

Krimmel, J.D. et al., "Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues," Proc. Natl. Acad. Sci. USA 2016, 113(21), 6005-6010.

Kukita, Y. et al. "High-fidelity target sequencing of individual molecules identified using barcode sequences: de novo detection and absolute quantitation of mutations in plasma cell-free DNA from cancer patients" DNA Research (2015) 22(4)269-277.

Lam, et al. Plasma DNA as a prognostic marker for stroke patients with negative neuroimaging within the first 24 h of symptom onset. Resuscitation. Jan. 2006;68(1):71-8. Epub Dec. 1, 2005.

Lanman, et al., Analytical and Clinical Validation of a Digital Sequencing Panel for Quantitative, Highly Accurate Evaluation of Cell-Free Circulating Tumor DNA PLoS One, Oct. 2015, 10(10), e0140712. doi:10.1371/journal.pone.0140712.

Larson, et al. A single molecule view of of gene expression. Trends Cell Biol. Nov. 2009;19(11):630-7. Epub Oct. 8, 2009.

Leary, et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Sci Transl Med. Nov. 28, 2012;4(162):162ra154. doi: 10.1126/scitranslmed.3004742.

Leary, et al. Development of personalized tumor biomarkers using massively parallel sequencing. Sci Transl Med. Feb. 24, 2010;2(20):20ra14. doi: 10.1126/scitranslmed.3000702.

Lecomte, et al. Detection of free-circulating tumor-associated DNA in plasma of colorectal cancer patients and its association with prognosis. Int J Cancer Aug. 10, 2002;100(5):542-8.

Ledergerber, C. et al., "Base-calling for next-generation sequencing platforms," Brief Bioinform. 2011 12(5), 489-497.

Leon, et al. Free DNA in the serum of cancer patients and the effect of therapy. Cancer Res. Mar. 1977;37(3):646-50.

Leung et al. Plasma Epstein-Barr viral deoxyribonucleic acid quantitation complements tumor-node-metastasis staging prognostication in nasopharyngeal carcinoma. J Clin Oncol. Dec. 1, 2006;24(34):5414-8.

Li, et al. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. Nov. 2008;18(11):1851-8. doi: 10.1101/gr.078212.108. Epub Aug. 19, 2008.

Li et al. Structure-independent and quantitative ligation of single-stranged DNA. Anal Biochem. Feb. 15, 2006;349(2):242-6. Epub Nov. 18, 2005.

Li, H. et al., "Fast and accurate short read alignment with Burrows-Wheeler transform," Bioinformatics 2009, 25(14), 1754-1760.

Liang, K-C. et al., "Bayesian basecalling for DNA sequence analysis using hidden Markov models," IEE/ACM Trans. Comput. Biol. Bioinform. 2007 4(3), 430-440.

Liao. G.J.W. et al., "Targeted massively parallel sequencing of maternal plasma DNA permits efficient and unbiased detection of fetal alleles," Clin. Chem. 2011, 57(1), 92-101.

Lizardi, et al. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. Nat Genet. Jul. 1998;19(3):225-32.

Lo, et al. Quantitative Analysis of Cell-free Epstein-Barr Virus DNA in Plasma of Patients with Nasopharyngeal Carcinoma, Cancer Research, 59(6):1188-1191 (Mar. 1999).

Lo, Y.M.D et al. "Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis," Am. J. Hum. Genet. 1998, 62(4), 768-775.

Lo, Y.M.D. et al. "Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus," Sci Transl Med (2010) 2(61):1-13.

Lockhart, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology, 14: 1675-1680 (1996).

Lodes, et al. Novel Chimera-Free, High Efficiency Library Preparation for NGS Platforms.Poster. Presented at the Plant & Animal Genome XX Conference. San Diego, California. Jan. 14-18, 2012. 1 page.

Lucito, et al. Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. Genome Res. Nov. 2000;10(11):1726-36.

Lucito, et al. Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation. Genome Research (2003), 13: 2291-2305.

Lunter, G. et al., "Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads," Genome Res. 2011, 21(6):936-939.

Maamar, et al. Noise in Gene Expression Determines Cell Fate in Bacillus subtilis. Science, 317: 526-529 (2007).

Mahmud, et al. Fast MCMC sampling for hidden Markov Models to determine copy number variations. BMC Bioinformatics. Nov. 2, 2011;12:428. doi: 10.1186/1471-2105-12-428.

Makrigiorgos, et al., A PCR-Based amplification method retaining quantative difference between two complex genomes. Nature Biotech, vol. 20, No. 9, pp. 936-939 (Sep. 2002).

Mamanova, L. et al., "Target-enrichment strategies for next-generation sequencing," Nat. Methods 2010, 7(2), 111-118.

Mandel, et al. Les acides nucleiques du plasma sanguin chez l'homme. C R Seances Soc Biol Fil. Feb. 1948;142(3-4):241-3.

Marsit, et al. Epigenetic profiling reveals etiologically distinct patterns of DNA methylation in head and neck squamous cell carcinoma. Carcinogenesis. Mar. 2009;30(3):416-22. doi: 10.1093/carcin/bgp006. Epub Jan. 6, 2009.

McCloskey, et al. Encoding PCR products with batch-stamps and barcodes. Biochem Genet. Dec. 2007;45(11-12):761-7. Epub Oct. 23, 2007.

McKernon, K.J. et al., "Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding," Genome Res. 2009, 19(9), 1527-1541.

Medvedev, et al. Detecting copy number variation with mated short reads. Genome Res. Nov. 2010;20(11):1613-22. doi: 10.1101/gr.106344.110. Epub Aug. 30, 2010.

Mei, et al. Identification of recurrent regions of Copy-Number Variants across multiple individuals. BMC Bioinformatics. Mar. 22, 2010;11:147. doi: 10.1186/1471-2105-11-147.

Meldrum, et al. Next generation sequencing for cancer diagnostics: a practical perspective. Clin Biochem Rev. Nov. 2011;32(4):177-95.

Mertes, F. et al. "Targeted enrichment of genomic DNA regions for next-generation sequencing" Brief Functional Genomics (2011) 10(6):374-386.

Metzker, M.L. "Sequencing technologies—the next generation" Nature Reviews Genetics (2010) 11:31-46.

Meyer, M. et al. "Illumina sequencing library preparation for highly multiplexed target capture and sequencing," Cold Spring Harb. Protoc. 2010, (6), prot5448.

Meyerson, M. et al. "Advances in understanding cancer genomes through second-generation sequencing" Nature Reviews Genetics (2010) 11:685-696.

Milbury, et al. "Enabling Clinical Cancer Genomics for Rare Mutations: COLD-PCR Magnifies Mutations Prior to Targeted Amplicon Re-Sequencing". Clin Chem. Mar. 2012;58(3):580-9. doi: 10.1373/clinchem.2011.176198. Epub Dec. 21, 2011.

(56) **References Cited**

OTHER PUBLICATIONS

Miner, et al. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. Nucleic Acids Res. Sep. 30, 2004;32(17):e135.
Mitchell, et al. Circulating microRNAS as stable blood-based markers for cancer detection. Proc Natl Acad Sci U S 4. Jul. 29, 2008;105(30):10513-8. Epub Jul. 28, 2008.
Mitelman, F. et al. "The impact of translocations and gene fusions on cancer causation" Nature Rev. Cancer (2007) 7(4)233-245.
Mohan, et al., Changes in colorectal carcinoma genomes under anti-EGFR therapy identified by whole-genome plasma DNA sequencing.,PLoS Genet,doi: 10.1371/journal.pgen.1004271. eCollection 2014.,Mar. 27, 2014,10(3), e1004271.
Mori, et al. Predictive utility of circulating methylated DNA in serum of melanoma patients receiving biochemotherapy. J Clin Oncol. Dec. 20, 2005;23(36):9351-8.
Morrissy, A.S. et al. "Next-generation tag sequencing for cancer gene expression profiling" Genome Research (2009) 19(10): 1825-1835.
Mortazavi, et al. Mapping and quantifying mammalian transcriptomes by RNA—Seq. Nat. Methods. 2008; 5, 621-628.
Narayan, et al. Ultrasensitive measurement of hotspot mutations in tumor DNA in blood using error-suppressed multiplexed deep sequencing. Cancer Res. Jul. 15, 2012;72(14):3492-8. doi: 10.1158/0008-5472.CAN-11-4037. Epub May 10, 2012.
UCSC Genome Bioinformatics. About the UCSC Genome Bioinformatics Site. http://genome.ucsc.edu/index.html. Accessed on May 26, 2015. 2 pgs.
Umetani et al. Prediction of Breast Tumor Progression by Integrity of Free Circulating DNA in Serum. Journal of Clinical Oncology 24(26):4270-4276 (Sep. 10, 2006).
U.S. Appl. No. 61/384,001, filed Sep. 17, 2010.
U.S. Appl. No. 61/432,119, filed Jan. 12, 2011.
Utting, et al. Microsatellite analysis of free tumor DNA in urine, serum, and plasma of patients: a minimally invasive method for the detection of bladder cancer. Clin Cancer Res. Jan. 2002;8(1):35-40.
Van Houten, et al. Mutated p53 as a molecular marker for the diagnosis of head and neck cancer. J Pathol. Dec. 2002;198(4):476-86.
Van Orsouw, N. et al. "Complexity Reduction of Polymorphic Sequences (CRoPSTM): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes" PLoS One (2007) 11(e1172):1-10.
Vasyukhin, V. et al., "K-Ras Point Mutations in the Blood Plasma DNA of Patients with Colorectal Tumors" Challenges of Modern Medicine, 141-150 (Verna and Shamoo eds, 1994).
Velculescu, et al. Characterization of the Yeast Transcriptome. Cell, 88: 243-251 (1997).
Velculescu, et al. Serial Analysis of Gene Expression. Science, 270: 484-487 (1995).
Vogelstein, et al. Digital PCR. Proc. Natl. Acad. Sci., 96(16): 9236-9241(1999).
Wagle, N. et al., "High-throughput Detection of actionable Genomic alterations in clinical tumor samples by targeted, Massively Parallel sequencing," Cancer Discov. 2012, (2)1:82-93.
Walker, et al. Isothermal in vitro amplification of DNA by a restriction enzyme/DNA polymerase system. Proc Natl Acad Sci U S A. Jan. 1, 1992;89(1):392-6.
Walsh, et al. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc Natl Acad Sci U S A. Jul. 13, 2010;107(28):12629-33. doi: 10.1073/pnas.1007983107. Epub Jun. 28, 2010.
Wang, et al. iCLIP predicts the dual splicing effects of TIA-RNA interactions, Oct. 2010, PLoS Biol, 8(10):e1000530.
Wang, et al. Molecular detection of APC, K- ras, and p53 mutations in the serum of colorectal cancer patients as circulating biomarkers. World J Surg. Jul. 2004;28(7):721-6. Epub Jun. 8, 2004.
Wang, et al. RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics, 10: 57-63 (2009).
Wang, T.L. et al. "Digital Karyotyping" PNAS (2002) 99(25):16156-16161.
Wang, Y. et al., "Detection of somatic mutations and HPV in the saliva and plasma of patients with head and neck squamous cell carcinomas," Sci. Transl. Med. 2015, 7(293):293ra104.
Wang, Y. et al., "Detection of tumor-derived DNA in cerebrospinal fluid of patients with primary tumors of the brain and spinal cord," Proc. Natl. Acad. Sci. USA 2015, 112(31), 9704-9709.
Wang, Y. et al., "Diagnostic potential of tumor DNA from ovarian cyst fluid," eLife 2016, 5:e15175.
Weber, et al. A real-time polymerase chain reaction assay for quantification of allele ratios and correction of amplification bias. Anal Biochem. Sep. 15, 2003;320(2):252-8.
Wheeler, D.A. et al., "The complete genome of an individual by massively parallel DNA sequencing," Nature 2008, 452(7189), 872-876.
Williford, A. et al., "Gene Fusion," eLS 2013, 1-8.
Wittes, et al. Searching for Evidence of Altered Gene Expression: a Comment on Statistical Analysis of Microarray Data. Journal of the National Cancer Institute, 91(5): 400-401 (1999).
Wodicka, et al. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nature Biotechnology, 15: 1359-1367 (1997).
Wojdacs, et al. Primer design versus PCR bias in methylation independent PCR amplifications. Epigenetics. May 16, 2009;4(4):231-4. Epub May 14, 2009.
Wood, et al. The genomic landscapes of human breast and colorectal cancers. Science. Nov. 16, 2007;318(5853):1108-13. Epub Oct. 11, 2007.
Wood, et al. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucleic Acids Res. Aug. 2010;38(14): a151. doi: 10.1093/nar/gkq510. Epub Jun. 4, 2010.
Xi, et al. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc Natl Acad Sci U S A. Nov. 15, 2011;108(46):E1128-36. doi: 10.1073/pnas.1110574108. Epub Nov. 7, 2011.
Yandell, et al. A probabilistic disease-gene finder for personal genomes. Genome Res. Sep. 2011;21(9):1529-42. doi: 10.1101/gr.123158.111. Epub Jun. 23, 2011.
Yang, et al., EGFR gene copy number as a predictive biomarker for the treatment of metastatic colorectal cancer with anti-EGFR monoclonal antibodies: a meta-analysis.,J Hematol Oncol, Aug. 16, 2012,5:52,1-9.
Yang. Simple binary segmentation frameworks for identifying variation in DNA copy number. BMC Bioinformatics. Oct. 30, 2012;13:277. doi: 10.1186/1471-2105-13-277.
Ye, et al. Fluorescent microsphere-based readout technology for multiplexed human single nucleotide polymorphism analysis and bacterial identification. Human Mutation, 17(4): 305-316 (2001).
Yoon, et al. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. Sep. 2009;19(9):1586-92. doi: 10.1101/gr.092981.109. Epub Aug. 5, 2009.
Zhang et al. "Comprehensive One-Step Molecular Analysis of Mitochondrial Genome by Massively Parallel Sequencing" Clinical Chem (2012) 58(9):1322-1331.
Zhang et al. The impact of next-generation sequencing on genomics. J Genet Genomics. Mar. 20, 2011;38(3):95-109. doi: 10.1016/j.jgg.2011.02.003. Epub Mar. 15, 2011.
Zhao, et al. Homozygous Deletions and Chromosome Amplifications in Human Lung Carcinomas Revealed by Single Nucleotide Polymorphism Array Analysis. Cancer Research, 65: 5561-5570 (2005).
Zhou, et al. Counting alleles reveals a connection between chromosome 18q loss and vascularinvasion. Nature Biotechnology, 19: 78-81 (2001).
Guardant Health vs. *FMI* Invalidity Exhibit C-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit C-4, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit C-5, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

(56) **References Cited**

OTHER PUBLICATIONS

Guardant Health vs. *FMI* Invalidity Exhibit D-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit D-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit D-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit D-5, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Second Amended Complaint dated Mar. 6, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Suppl Invalidity References dated Feb. 15, 2019 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Supplemental Invalidity Contentions dated Mar. 29, 2019 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *PGDx* Amended Answer to Second Amended Complaint dated Apr. 30, 2019 (C.A. No. 17-av-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart A, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart B, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart C, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart D, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart E, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart F, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart G, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart H, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart I, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart J, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart K, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart L, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Contentions dated Jun. 25, 2018 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Second Amended Complaint dated Mar. 23, 2018 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Second Updated Invalidity Contentions dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Updated Invalidity Contentions dated Mar. 29, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Gundry, et al. "Direct mutation analysis by high-throughput sequencing: from gremlin to low-abundant, somatic variants" Mutat Res Jan. 3, 2012; 729(1-2):1-15. doi: 10.1016/mrfmmm.2011.10.001. Pub Oct. 12, 2011.
Gundry, et al. Direct, genome-wide assessment of DNA mutations in single cells. Nucleic Acids Res. Mar. 2012;40(5):2032-40. doi: 10.1093/nar/gkr949. Epub Nov. 15, 2011.
Hacia, et al. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nature Genetics, 22: 164-167 (1999).
Hafner, et al., RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA eDNA libraries, RNA Sep. 1, 2011, 17(9):1697-1712.
Hall, A., Short-Read DNA Sequence Alignmnet with Custom Designed FPGA-based Hardware, Thesis 2010.
Hamady, et al. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methods. Mar. 2008;5(3):235-7. doi: 10.1038/nmeth.1184. Epub Feb. 10, 2008.
Hensel, et al. Simultaneous identification of bacterial virulence genes by negative selection. Science. Jul. 21, 1995;269(5222):400-3.
Hiatt, et al. Parallel, tag-directed assembly of locally derived short sequence reads. Nat Methods. Feb. 2010;7(2)119-22. doi: 10.1038/nmeth.1416. Epub Jan. 17, 2010.
Hiatt, et al. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. Genome Res. May 2013;23(5):843-54. doi: 10.1101/gr.147686.112. Epub Feb. 4, 2013.
Hibi, et al. Molecular detection of genetic alterations in the serum of colorectal cancer patients. Cancer Res. Apr. 1, 1998;58(7):1405-7.
HiSeq 2000 User Guide (2010).
Holdenrieder, et al. Circulating nucleosomes and cytokeratin 19-fragments in patients with colorectal cancer luring chemotherapy. Anticancer Res. May-Jun. 2005;25(3A):1795-801.
Holies, et al. A stochastic approach to count RNA molecules using DNA sequencing methods. 2003, Lecture Notes in Computer Science, vol. 2812, pp. 55-62.
Hollas, et al. A stochastic approach to count RNA molecules using DNA sequencing methods. Lecture Notes in Computer Science, 2812: 55-62 (2003).
Hollstein, M. et al. "p53 mutations in human cancers" Science (1991) 253(5015):49-53.
Hoque, et al. Detection of aberrant methylation of four genes in plasma DNA for the detection of breast cancer. J Clin Oncol. Sep. 10, 2006;24(26):4262-9. Epub Aug. 14, 2006.
Howe, et al. Retinoblastoma growth suppressor and a 300-kDa protein appear to regulate cellular DNA synthesis. Proc Natl Acad Sci U S A. Aug. 1990;87(15):5883-7.
Hug, et al. Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. J Theor Biol. Apr. 21, 2003;221(4):615-24.
Hyland, et al. The normal and tumor spectrum of copy number variation: Copy number alterations correlate with changes in gene expression in tumor transcriptome. Nov. 15, 2009. 1 page. Retreived from:http://tools.thermofisher.com/content/sfs/posters/cms_073633.pdf.
Iafrate, et al. Detection of large-scale variation in the human genome. Nat Genet. Sep. 2004;36(9):949-51. Epub Aug. 1, 2004.
Ikeguchi, et al. Detection of loss of heterozygosityat microsatellite loci in esophageal squamous-cell carcinoma. Oncology. 1999;56(2)164-8.
Illumina "Multiplexed Sequencing with Illumina Genome Analyzer System" https://www.illumina.com/Documents/products/datasheets/datasheet_sequencing_multiplex.pdf (Dec. 2008).
Illumina Preparing Samples for Paired-End Sequencing (2008).
Ingolia, et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. Apr. 10, 2009;324(5924)118-23. Epub Feb. 12, 2009.
Instructions for Norit Rapid DNA Ligation Kit (Nov. 6, 2004).
International search report and written opinion dated Apr. 3, 2015 for PCT/US2014/072383.
International search report and written opinion dated May 7, 2012 for PCT/IB2011/003160.
International search report and written opinion dated Jun. 6, 2012 for PCT/US2011/065291.
International search report and written opinion dated Jul. 15, 2015 for PCT/US2015/030639.
International search report and written opinion dated Sep. 5, 2014 for PCT/US2014/000048.
International search report and written opinion dated Sep. 6, 2017 for PCT Application No. US-201727809.
International search report and written opinion dated Nov. 18, 2013 for PCT/US2013/058061.
Invitrogen Instructions for T4 DNA Ligase (May 5, 2002).
IPR2018-00130, Petition for Inter Partes Review of U.S. Pat. No. 9,598,731, dated Nov. 7, 2018.
IPR2019-00130, Decision Denying Inter Partes Review U.S. Pat. No. 9,598,731, issued Jun. 5, 2019.
IPR2019-00130, Preliminary Response to Petition for Inter Partes Review U.S. Pat. No. 9,598,731, filed Mar. 6, 2019.
IPR2019-00634, Decision Instituting Inter Partes Review U.S. Pat. No. 9,840,743, issued Aug. 19, 2019.

(56) **References Cited**

OTHER PUBLICATIONS

IPR2019-00634, Petition for Inter Partes Review of U.S. Pat. No. 9,840,743, dated Feb. 1, 2019.
IPR2019-00636 & 637, Exhibit 3001 Memo for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Jul. 1, 2019.
IPR2019-00636 & 637, Order—Conduct of the Proceedings for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Jul. 22, 2019.
IPR2019-00636 and IPR2019-00637, Decision Denying Inter Partes Review U.S. Pat. No. 9,902,992, issued Aug. 20, 2019.
IPR2019-00636, Petition for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Feb. 1, 2019.
IPR2019-00637, Petition for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Feb. 1, 2019.
IPR2019-00652, Decision Granting Inter Partes Review U.S. Pat. No. 9,834,822, issued Aug. 19, 2019.
IPR2019-00652, Petition for Inter Partes Review of U.S. Pat. No. 9,834,822, dated Feb. 1, 2019.
IPR2019-00653, Decision Denying Inter Partes Review U.S. Pat. No.9,834,822, issued Aug. 19, 2019.
IPR2019-00653, Petition for Inter Partes Review of U.S. Pat. No. 9,834,822, dated Feb. 1, 2019.
Jabara, C. Capturing the cloud: High throughput sequencing of multiple individual genomes from a retroviral population. Biology Lunch Bunch Series, Training Initiatives in Biomedical & Biological Sciences of the University of North Carolina at Chapel Hill, Apr. 23, 2010.
Jabara, et al. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. (Paper # 665), The 18th Annual Conference on Retroviruses and Opportunistic Infections, Boston, Massachusetts, Mar. 2011.
Jabara, et al. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl Acad Sci U S A. Dec. 13, 2011;108(50):20166-71. doi: 10.1073/pnas.1110064108. Epub Nov. 30, 2011.
Jahr, et al. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. Cancer Res. Feb. 15, 2001;61(4):1659-65.
Jeronimo, et al. Quantitative GSTP1 hypermethylation in bodily fluids of patients with prostate cancer. Urology. Dec. 2002;60(6):1131-5.
Jiang, H. et al., "SeqMap: mapping massive amount of oligonucleotides to the genome," Bioinformatics 2008, 24(20):2395-2396.
Kanagawa. Bias and artifacts in multitemplate polymerase chain reactions (PCR), 2003, Journal of Bioscience and Bioengineering, vol. 96, No. 4, p. 317-323.
Kao,W-C. et al., "BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing," Genome Res. 2009 19(10), 1884-1895.
Karow, J. "Hopkins Team Develops Method to Improve Rare Mutation Detection for Early Cancer Dx" Genome Web (2011) 3 pages.
Kennedy, S.R. et al., "Detecting ultralow-frequency mutations by Duplex Sequencing" Nature Protocols (2014) 9(11):2586-2606.
Kennedy, S.R. et al., "Ultra-Sensitive Sequencing Reveals an Age-Related Increase in Somatic Mitochondrial Mutations That are Inconsistent with Oxidative Damage" PLOS Genetics (2013) 9:e1003794.
Kennedy, S.R. et al., "Detecting ultralow-frequency mutations by Duplex Sequencing," Nat. Protoc. 2014, 9(11), 2586-2606.
Kimura, et al. EGFR mutation status in tumour-derived DNA from pleural effusion fluid is a practical basis for predicting the response to gefitinib. Br J Cancer. Nov. 20, 2006;95(10):1390-5. Epub Oct. 24, 2006.
Kinde, et al. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. Jun. 7, 2011;108(23):9530-5. doi: 10.1073/pnas.1105422108. Epub May 17, 2011.
Kinde, et al. Supplemental Information, Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. Jun. 7, 2011;108(23):1-10.
Kinde, I. et al., "Evaluation of DNA from the Papanicolaou Test to Detect Ovarian and Endometrial Cancers," Sci. Transl. Med. 2013, 5(167):167ra4.
Kinde, I. et al., "FAST-SeqS: a simple and efficient method for the detection of aneuploidy by massively parallel sequencing," PLoS One 2012, 7(7), e41162.
Kircher, M. et al., "Improved base calling for the Illumina Genome Analyzer using machine learning strategies," Genome Biol. 2009, 10(8), R83.
Kirsch, S. et al. "Sequence error storms and the landscape of mutations in cancer," Proc. Natl. Acad. Sci. USA 2012, 109(36), 14289-14290.
Kivioja, T., et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. Nov. 20, 2011;9(1):72-4. doi: 10.1038/nmeth.1778.
Koboldt, D.C.et al., "The next-generation sequencing revolution and its impact on genomics," Cell 2013, 155(1), 27-38.
Koboldt, et al. Massively parallel sequencing approaches for characterization of structural variation. Aug. 12, 2011 Methods Mol Biol. 2012;838:369-84. doi: 10.1007/978-1-61779-507-7_18.
Kolble, et al. Microsatellite alterations in serum DNA of patients with colorectal cancer. Lab Invest. Sep. 1999;79(9):1145-50.
Konig, et al. iCLIP reveals the function of hnRNAP particles in splicing at individual nucleotide resolution, Jul. 2010, Nature Structural & Molecular Biology, 17(7):909-916.
Kopreski, et al. Detection of mutant K-ras DNA in plasma or serum of patients with colorectal cancer. Br J Cancer. 1997;76(10):1293-9.
Korbel, J.O. et al., "Paired-end mapping reveals extensive structural variation in the human genome," Science 2007, 318(5849), 420-426.
Nawroz, et al. Microsatellite alterations in serum DNA of head and neck cancer patients. Nat Med. Sep. 1996;2(9):1035-7.
Newman, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. May 2014;20(5):548-54. doi: 10.1038/nm.3519. Epub Apr. 6, 2014.
Ng, S.B., et al., "Targeted capture and massively parallel sequencing of 12 human exomes," Nature 2009, 461(7261), 272-276.
Nielsen, R. et al. "Genotype and SNP calling from next-generation sequencing data" Nature Reviews Genetics (2011) 12(6):443-451.
Nord, et al. Accurate and exact CNV identification from targeted high-throughput sequence data. BMC Genomics. Apr. 12, 2011;12:184.
Notice of Allowance dated Jan. 31, 2017 for U.S. Appl. No. 14/712,754.
Notice of allowance dated Mar. 21, 2014 for U.S. Appl. No. 12/969,581.
Notice of allowance dated Jun. 15, 2017 for U.S. Appl. No. 15/076,565.
Notice of allowance dated Jun. 19, 2014 for U.S. Appl. No. 12/969,581.
Notice of allowance dated Aug. 1, 2017 for U.S. Appl. No. 15/492,659.
Notice of allowance dated Aug. 4, 2017 for U.S. Appl. No. 15/467,570.
Notice of allowance dated Aug. 22, 2014 for U.S. Appl. No. 12/969,581.
Notice of allowance dated Aug. 29, 2017 for U.S. Appl. No. 15/492,659.
Notice of allowance dated Sep. 11, 2017 for U.S. Appl. No. 15/076,565.
Notice of allowance dated Oct. 3, 2017 for U.S. Appl. No. 15/076,565.
Notice of allowance dated Oct. 25, 2017 for U.S. Appl. No. 14/861,989.
Notice of allowance dated Dec. 28, 2017 for U.S. Appl. No. 14/861,989.
Office Action dated Feb. 9, 2017 for U.S. Appl. No. 15/076,565.
Office action dated May 13, 2019 for U.S. Appl. No. 15/669,779.
Office action dated May 20, 2016 for U.S. Appl. No. 14/855,301.
Office action dated May 31, 2016 for U.S. Appl. No. 14/712,754.
Office action dated Jun. 1, 2017 for U.S. Appl. No. 15/467,570.
Office action dated Jun. 3, 2016 for U.S. Appl. No. 14/861,989.
Office action dated Jun. 12, 2017 for U.S. Appl. No. 15/492,659.
Office action dated Jul. 18, 2017 for U.S. Appl. No. 14/861,989.
Office Action dated Jul. 30, 2019 for U.S. Appl. No. 16/283,635.
Office action dated Aug. 6, 2019 for U.S. Appl. No. 14/855,301.
Office action dated Oct. 3, 2013 for U.S. Appl. No. 12/969,581.

(56) **References Cited**

OTHER PUBLICATIONS

Office Action dated Oct. 12, 2016 for U.S. Appl. No. 14/712,754.
Office action dated Oct. 20, 2017 for U.S. Appl. No. 14/425,189.
Office Action dated Nov. 21, 2016 for U.S. Appl. No. 14/855,301.
Office action dated Dec. 4, 2015 for U.S. Appl. No. 14/712,754.
Office action dated Dec. 7, 2017 for U.S. Appl. No. 14/855,301.
Office Action dated Dec. 9, 2016 for U.S. Appl. No. 14/861,989.
Office action dated Dec. 21, 2018 for U.S. Appl. No. 15/978,848.
Ogino, et al. Quantification of PCR bias caused by a single nucleotide polymorphism in SMN gene dosage analysis. J Mol Diagn. Nov. 2002;4(4):185-90.
Opposition Brief dated Aug. 7, 2018 for Japanese Opposition No. 2018-700659 to JP Patent 6275145.
Opposition Decision dated Jan. 15, 2019 for Japanese Opposition No. 2018-700659 to JP Patent 6275145.
Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Foundation Medicine Inc.
Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Grunecker Patent-und Rechtsanwalte PartG mbB.
Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Personal Genome Diagnostics, Inc.
Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Strawman Limited.
Opposition Form and Statement to EP3087204 filed Nov. 14, 2018.
Pacific Biosciences. Template Preparation and Sequencing Guide. Publication date: Oct. 14, 2014. Pacific Biosciences website http://www.pacificbiosciences.com/support/pubmap/documentation.html.
Pan, et al. Loss of heterozygosity patterns provide fingerprints for genetic heterogeneity in multistep cancer progression of tobacco smoke-induced non-small cell lung cancer. Cancer Res. Mar. 1, 2005;65(5):1664-9.
Park, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet. May 2010;42(5):400-5. doi: 10.1038/ng.555. Epub Apr. 4, 2010.
Parkinson, N.J. et al., "Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA," Genome Res. 2012, 22(1), 125-133.
Pel, J. et al. "Duplex proximity sequencing (pro-seq): A method to improve DNA sequencing accuracy without the cost of molecular barcoding redundancy" bioRxiv (2017) https://doi.org/10.1101/163444.
Perakis, S. et al. "Advances in Circulating Tumor DNA Analysis" Adv Clin Chem (2017) pp. 1-81.
Pihlak, et al. Rapid genome sequencing with short universal tiling probes. Nature Biotechnology, 26: 676-684 (2008).
"Blood Plasma" Oxford Dictionary of Biochemistry and Molecular Biology 81 (2d ed. 2006).
"LB-246112—Detection of circulating tumor DNA in early stage cancers," Available at https://www.genomeweb.com/cancer/ngs-error-correction-method-described-aacr-excitement-grows-early-cancer-detection, Accessed on Apr. 26, 2017.
"Cohesive End," Oxford Dictionary of Biochemistry and Molecular Biology 132 (2d ed. 2006).
"Control," Oxford Dictionary of Biochemistry and Molecular Biology (2d ed. 2006), 143.
Ajay, S.S. et al., "Accurate and comprehensive sequencing of personal genomes," Genome Res. 2011, 21(9), 1498-1505.
Albert, T.J. et al., "Direct selection of human genomic loci by microarray hybridization," Nat. Methods 2007, 4(11), 903-905.
Alkan, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. Oct. 2009;41(10):1061-7. doi: 10.1038/ngA37. Epub Aug. 30, 2009.
Andersson, et al. Bayesian detection of periodic mRNA time profiles without use of training examples. BMC Bioinformatics. Feb. 9, 2006;7:63.
Angeloni, D. Molecular analysis of deletions in human chromosome 3p21 and the role of resident cancer genes in disease. Brief Funct Genomic Proteomic. Mar. 2007;6(1):19-39. Epub May 24, 2007.
Ansorge "Next-generation DNA sequencing techniques." New Biotechnology, 25(4):195-203 (2009).
Appel, Maryke. Part II: It's all about conversion. Kapa Biosystems. Accessed Mar. 29, 2017. Printed Apr. 11, 2017. 5 pages. URL:https://www.kapabiosystems.com/ngs/part-ii-conversion/.
Arneson, N. et al., "Whole-Genome Amplification by Adaptor-Ligation PCR of Randomly Sheared Genomic DNA (PRSG)," CSH Protocols, 2008, 3(1).
Ashford, Monica. NGS Error Correction Method Described at AACR as Excitement Grows for Early Cancer Detection. Genomeweb.com. Apr. 5, 2017. 3 pages. URL:https://www.genomeweb.com/cancer/ngs-error-correction-method-described-aacr-excitement-grows-early-cancer-detection.
Atanur, et al. The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. Genome Res. Jun. 2010;20(6):791-803. doi: 10.1101/gr.103499.109. Epub Apr. 29, 2010.
Audic, et al. "The Significance of Digital Gene Expression Profiles." Genome Research, 7: 986-995 (1997).
Barzon, et al. Evaluation of circulating thyroid-specific transcripts as markers of thyroid cancer relapse. Int J Cancer. Jul. 20, 2004;110(6):914-20.
Bendich, et al. Circulating DNA as a possible factor in oncogenesis. Science. Apr. 16, 1965;148(3668):374-6.
Bentley, D.R. et al., "Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry," Nature, 2008, 456(7218):53-59.
Bettegowda et al., "Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies," Sci. Transl. Med. 2014, 6(224):224ra24.
Bianchi, D.W. et al. "Genome-Wide Fetal Aneuploidy Detection by Maternal Plasma DNA Sequencing," Obstetrics & Gynecology, (2012) 119(5), 1-12.
Bonaldo, et al. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. Sep. 1996;6(9):791-806.
Bowcock, et al. Exclusion of the retinoblastoma gene and chromosome 13q as the site of a primary lesion for human breast cancer. Am J Hum Genet. Jan. 1990;46(1):12-7.
Braha, et al. Simultaneous stochastic sensing of divalent metal ions. Nature Biotechnology, 18: 1005-1007 (2000).
Bremnes, et al. Circulating tumour-derived DNA and RNA markers in blood: a tool for early detection, diagnostics, and follow-up? Lung Cancer. Jul. 2005;49(1):1-12.
Brenner, et al. "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays" Nature Biotechnology, 18: 630-634 (2000).
Brenner, et al. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. Proc Natl Acad Sci U S A. Feb. 15, 2000;97(4):1665-70.
Bryzgunova,O. et al., "A reliable method to concentrate circulating DNA," Analytical Biochem., 2010, 408:354-356.
Campbell, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. Jun. 2008;40(6):722-9. doi: 10.1038/ng.128. Epub Apr. 27, 2008.
Caramazza, et al. Chromosome 1 abnormalities in myeloid malignancies: a literature survey and karyotype-phenotype associations. Eur J Haematol. Mar. 2010;84(3):191-200. doi: 10.1111/j.1600-0609.2009.01392.x. Epub Nov. 30, 2009.
Carr, et al. Inferring relative proportions of DNA variants from sequencing electropherograms. Bioinformatics. Dec. 15, 2009;25(24):3244-50. doi: 10.1093/bioinformatics/btp583. Epub Oct. 9, 2009.
Casava v1.8.2 User Guide http://gensoft.pasteur.fr/docs/casava/1.8.2/CASAVA_1_8_2_UG_15011196C.pdf (Oct. 2011).
Casbon, et al. A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res. Jul. 2011;39(12):e81. doi: 10.1093/nar/gkr217. Epub Apr. 13, 2011.
Castle, et al. DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. BMC Genomics. Apr. 16, 2016;11:244. doi: 10.1186/1471-2164-11-244.

(56) **References Cited**

OTHER PUBLICATIONS

Chang, et al. Detection of allelic imbalance in ascitic supernatant by digital single nucleotide polymorphism analysis. Clin Cancer Res. Aug. 2002;8(8):2580-5.

Chee, et al. "Accessing genetic information with high-density DNA arrays." Science, 274: 610-614 (1996).

Chee. "Enzymatic multiplex DNA sequencing." Nucleic Acids Research, 19(12): 3301-3305 (1991).

Chen, et al. Microsatellite alterations in plasma DNA of small cell lung cancer patients. Nat Med. Sep. 1996;2(9):1033-5.

Chin, et al. A SNP in a let-7 microRNA complementary site in the KRAS 3' untranslated region increases non-small cell lung cancer risk. Cancer Res. Oct. 15, 2008;68(20):8535-40. doi: 10.1158/0008-5472.CAN-08-2129.

Chiu, et al. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. BMJ. Jan. 11, 2011;342:c7401. doi: 10.1136/bmj.c7401.

Chiu, et al. Quantitative analysis of circulating mitochondrial DNA in plasma. Clin Chem. May 2003;49(5):719-26.

Chiu, R.W.K. et al. "Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma" PNAS (2008) 105(51):20458-20463.

Church, et al. "Multiplex DNA sequencing." Science, 240: 185-188 (1988).

Cook, et al. Methylated DNA labels for marking objects. Biotechnol Lett. Jan. 2003;25(1):89-94.

Costello, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. Apr. 1, 2013;41(6):e67. doi: 10.1093/nar/gks1443. Epub Jan. 8, 2013.

Coulet, et al. Detection of plasma tumor DNA in head and neck squamous cell carcinoma by microsatellite typing and p53 mutation analysis. Cancer Res. Feb. 1, 2000;60(3):707-11.

Cox, J. "Bar coding objects with DNA." Analyst. May 2001;126(5):545-7.

Cox. Bar coding objects with DNA. Analyst. May 2001;126(5):545-7.

Craig, et al. Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods. Oct. 2008;5(10):887-93. doi: 10.1038/nmeth.1251. Epub Sep. 14, 2006.

Daines, et al. High-throughput multiplex sequencing to discover copy number variants in *Drosophila*. Genetics. Aug. 2009;182(4):935-41. doi: 10.1534/genetics.109.103218. Epub Jun. 15, 2009.

D'Antoni, et al. Rapid quantitative analysis using a single molecule counting approach. Anal Biochem. May 1, 2006;352(1):97-109. Epub Feb. 10, 2006.

Clark, T.A. et al. "Analytical Validation of a Hybrid Capture Based Next-Generation Sequencing Clinical Assay for Genomic Profiling of Cell-Free Circulating Tumor DNA," J. Mol. Diagnostics (2018) 20(5):686-702.

Paweletz, C.P. et al. "Bias-corrected targeted next-generation sequencing for rapid, multiplexed detection of actionable alterations in cell-free DNA from advanced lung cancer patients" Clin Canc Res (2016) 22(4):915-922.

Phallen, J. et al. "Direct detection of early-stage cancers using circulating tumor DNA" Sci Trans Med (2017) vol. 9, Issue 403, eaan2415DOI: 10.1126/scitranslmed.aan2415.

Shiroguchi, et al. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci U S A. 2012 Supplemental Information (8 pages).

Kamps-Hughes, N. et al. "ERASE-Seq: Leveraging replicate measurements to enhance ultralow frequency variant detection in NGS data" PLOS One (2018).

Lennon, N.J. et al. "Technological considerations for genome-guided diagnosis and management of cancer" Gen Med (2016) 8:112.

Mishra, S. et al. "Different Facets of Copy Number Changes: Permanent, Transient, and Adaptive" Mol Cell Biol (2016) 36(7):1050-1063.

Moench, S. Genomic Profiling Using Guardant 360 Cell-Free DNA-Based Assay vs Tumor-Based Genotyping Assays in Advanced NSCLC, Cancer Therapy Advisor (Feb. 28, 2019), https://www.cancertherapyadvisor.com/home/news/conferencecoverage/american-association-for-cancer-research-aacr/aacr-2019/genomic-profiling-using-guardant-360-cell-free-dna-based assay-vs-tumor-based-genotyping-assays-in-advanced-nsclc/ (lastaccessed Nov. 30, 2019).

Newman, A. et al. "Integrated digital error suppression for improved detection of circulating tumor DNA" Nature Biotech (2016) 34(5):547-555.

Odegaard, J.I. et al. "Validation of a Plasma-Based Comprehensive Cancer Genotyping Assay Utilizing Orthogonal Tissue- and Plasma-Based Methodologies" Clin Canc Res (2018) 24(15):3539-3549.

Ou, Shi et al. "Liquid Biopsy to Identify Actionable Genomic Alterations" Am Soc Clin Onc (2018) 978.

Sathirapongsasuti, J.F. et al. "Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV" BioInformatics (2011) 27(19):2648-2654.

Trapnell, C. et al. "How to map billions of short reads onto genomes" Nature Biotech (2009) 27(5):455-457.

Van Loo, P. et al. "Allele-specific copy number analysis of tumors" PNAS (2010) 107(39):16910-16915.

Wang, T.T. et al. "High efficiency error suppression for accurate detection of low-frequency variants" NAR (2019) 47(15):e87.

IPR2019-00634, Final Written Decision of U.S. Pat. No. 9,840,743, dated Aug. 18, 2020.

IPR2019-00652, Final Written Decision of U.S. Pat. No. 9,834,822, dated Aug. 18, 2020.

* cited by examiner

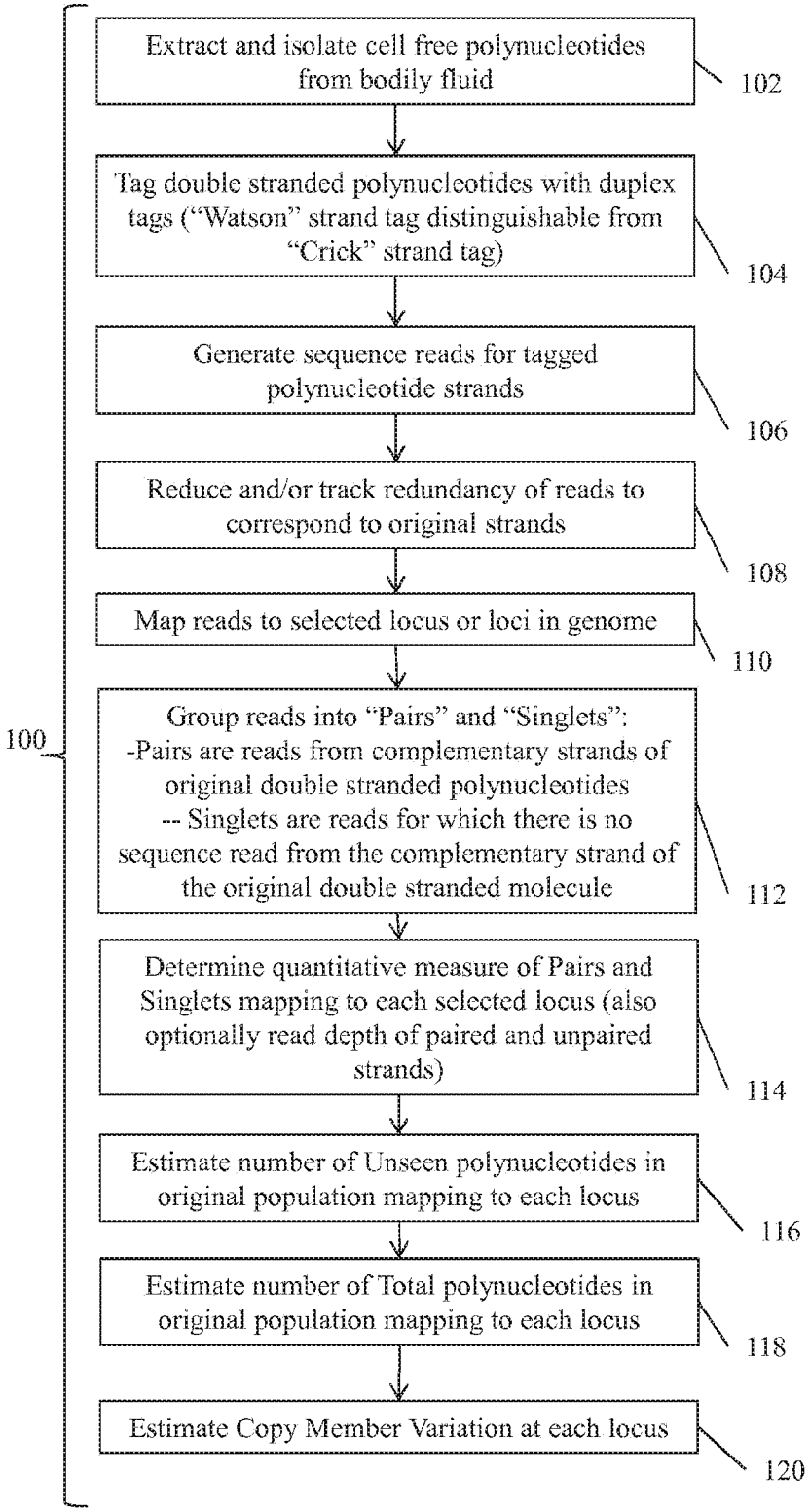

Extract and isolate cell free polynucleotides from bodily fluid
102
Tag double stranded polynucleotides with duplex tags ("Watson" strand tag distinguishable from "Crick" strand tag)
104
Generate sequence reads for tagged polynucleotide strands
106
Reduce and/or track redundancy of reads to correspond to original strands
108
Map reads to selected locus or loci in genome
110
100
Group reads into "Pairs" and "Singlets":
-Pairs are reads from complementary strands of original double stranded polynucleotides
-- Singlets are reads for which there is no sequence read from the complementary strand of the original double stranded molecule
112
Determine quantitative measure of Pairs and Singlets mapping to each selected locus (also optionally read depth of paired and unpaired strands)
114
Estimate number of Unseen polynucleotides in original population mapping to each locus
116
Estimate number of Total polynucleotides in original population mapping to each locus
118
Estimate Copy Member Variation at each locus
120

Fig. 1

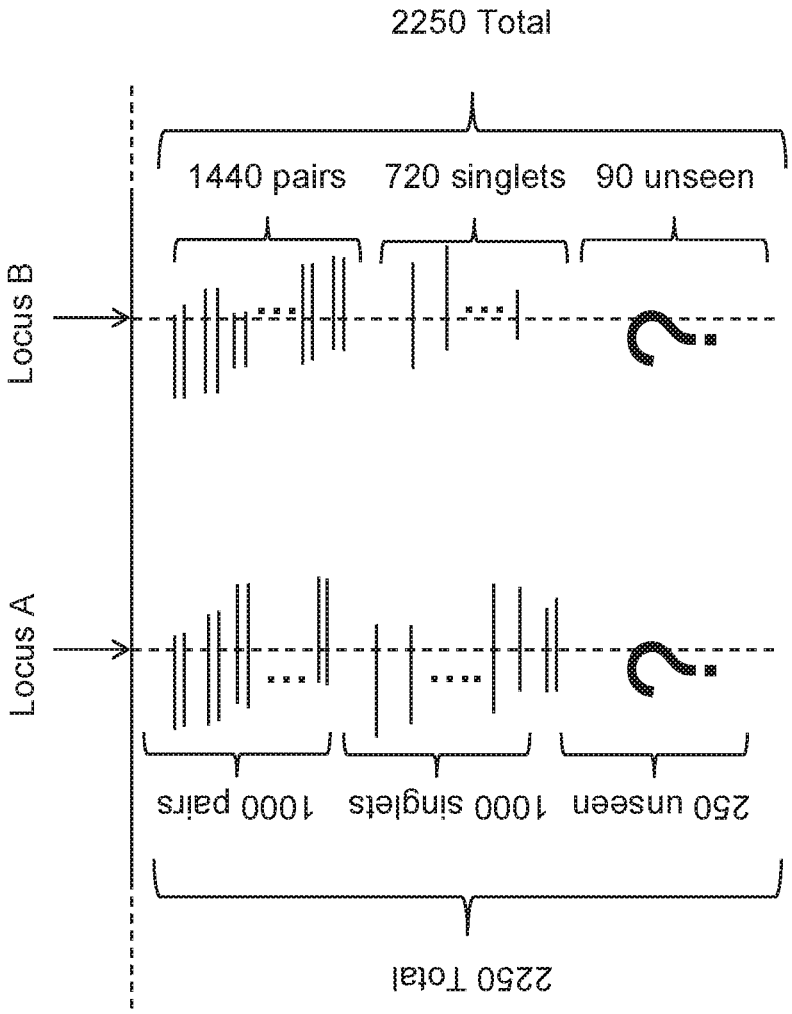

2250 Total
1440 pairs
720 singlets
90 unseen
Locus B
Locus A
1000 pairs
1000 singlets
250 unseen
2250 Total

Fig. 2

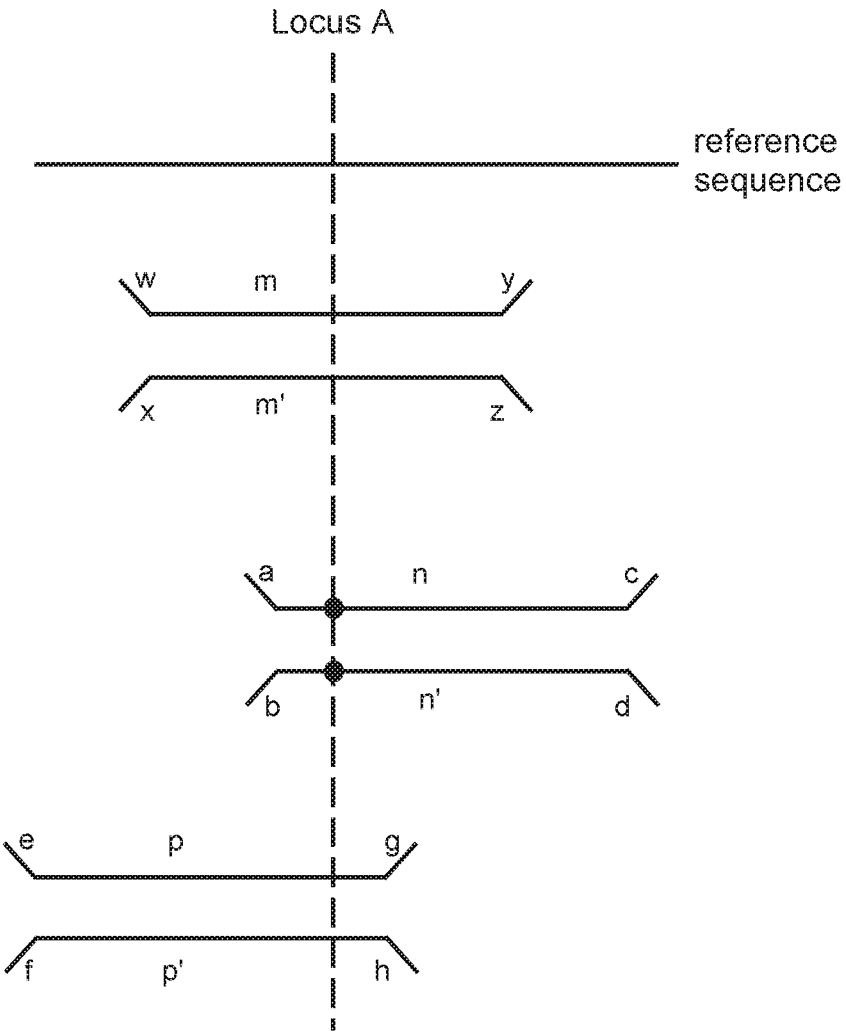


Fig. 3

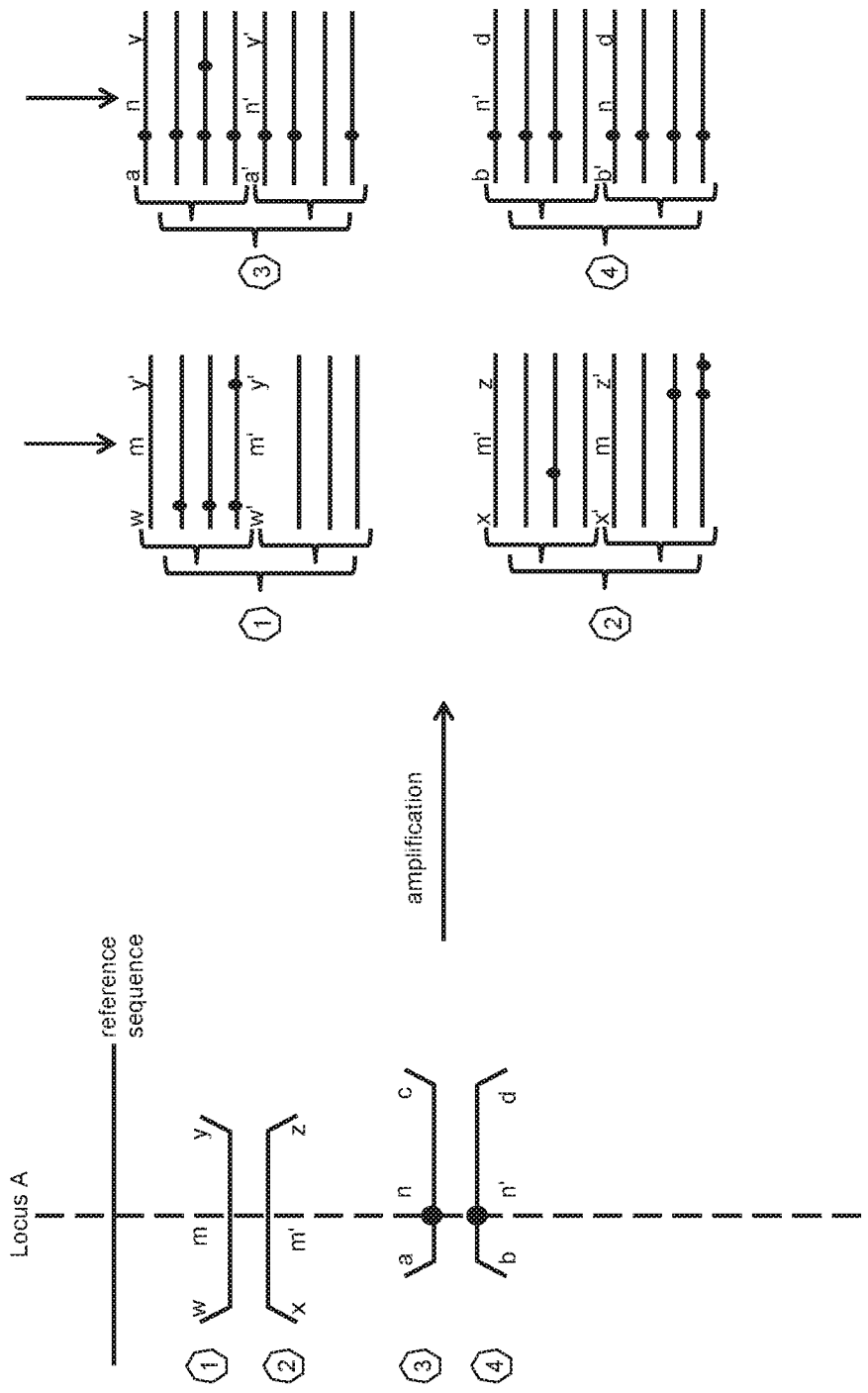

Locus A
reference sequence
amplification
w
m
y
x
m'
z
a
n
c
b
n'
d
1
2
3
4

Fig. 4A

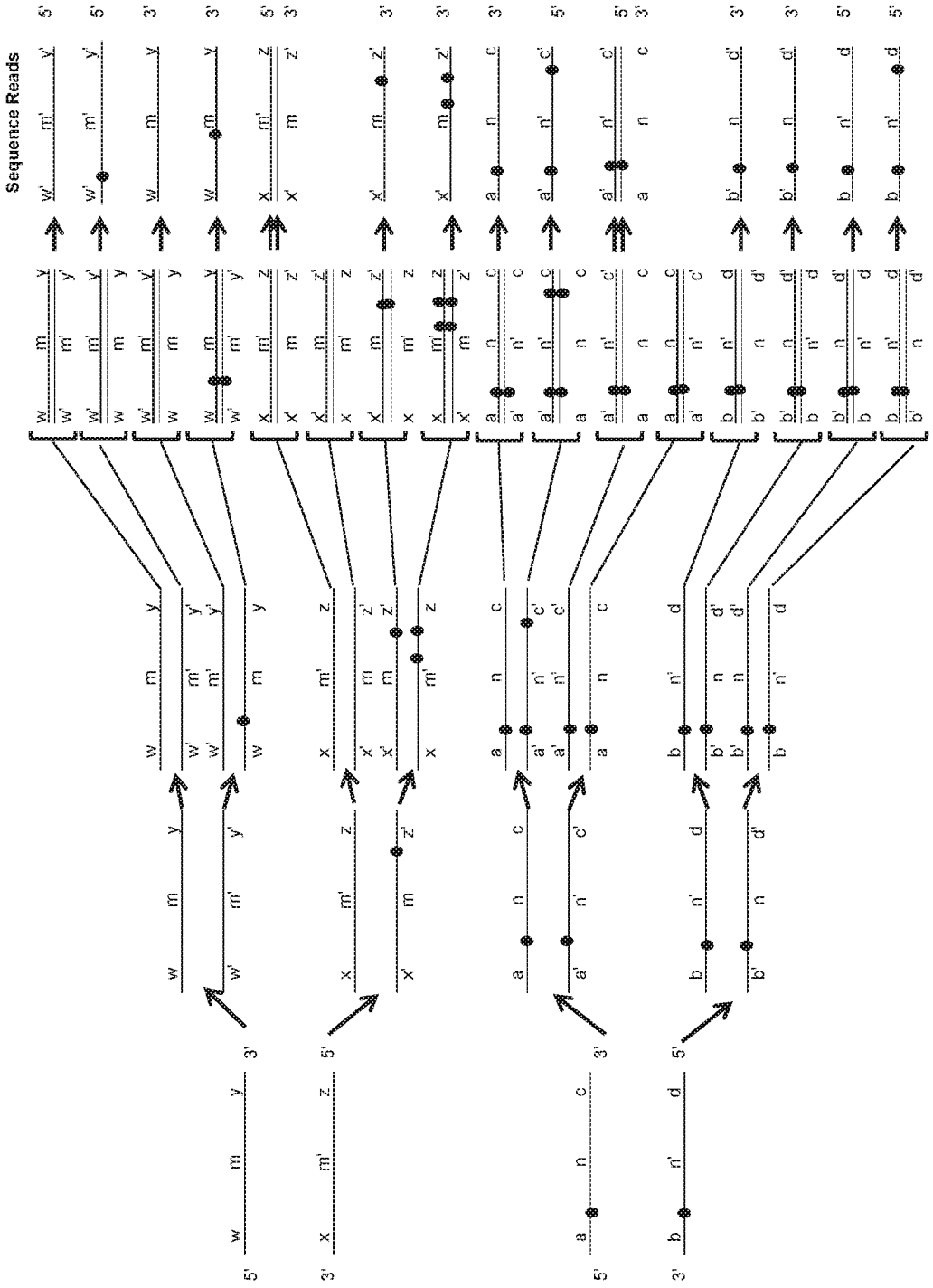


Fig. 4B

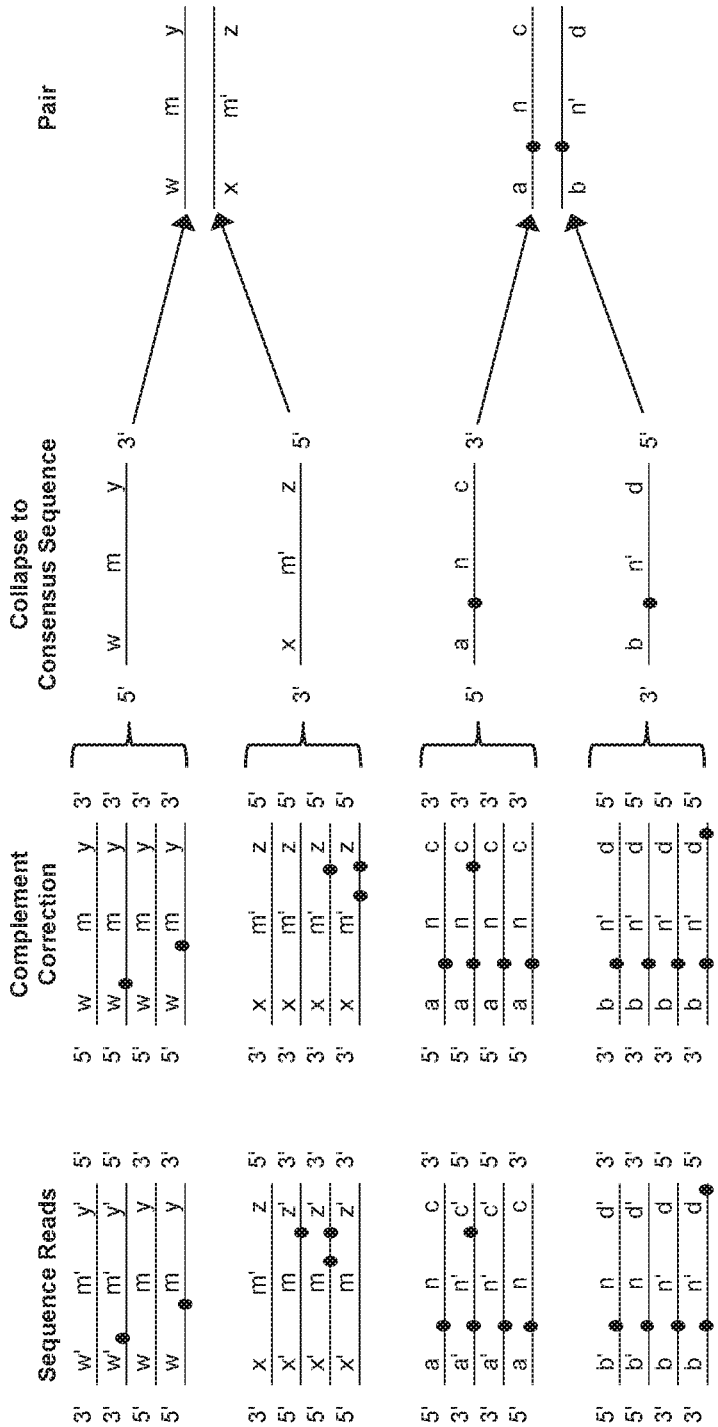

Sequence Reads
Complement Correction
Collapse to Consensus Sequence
Pair

Fig. 4C

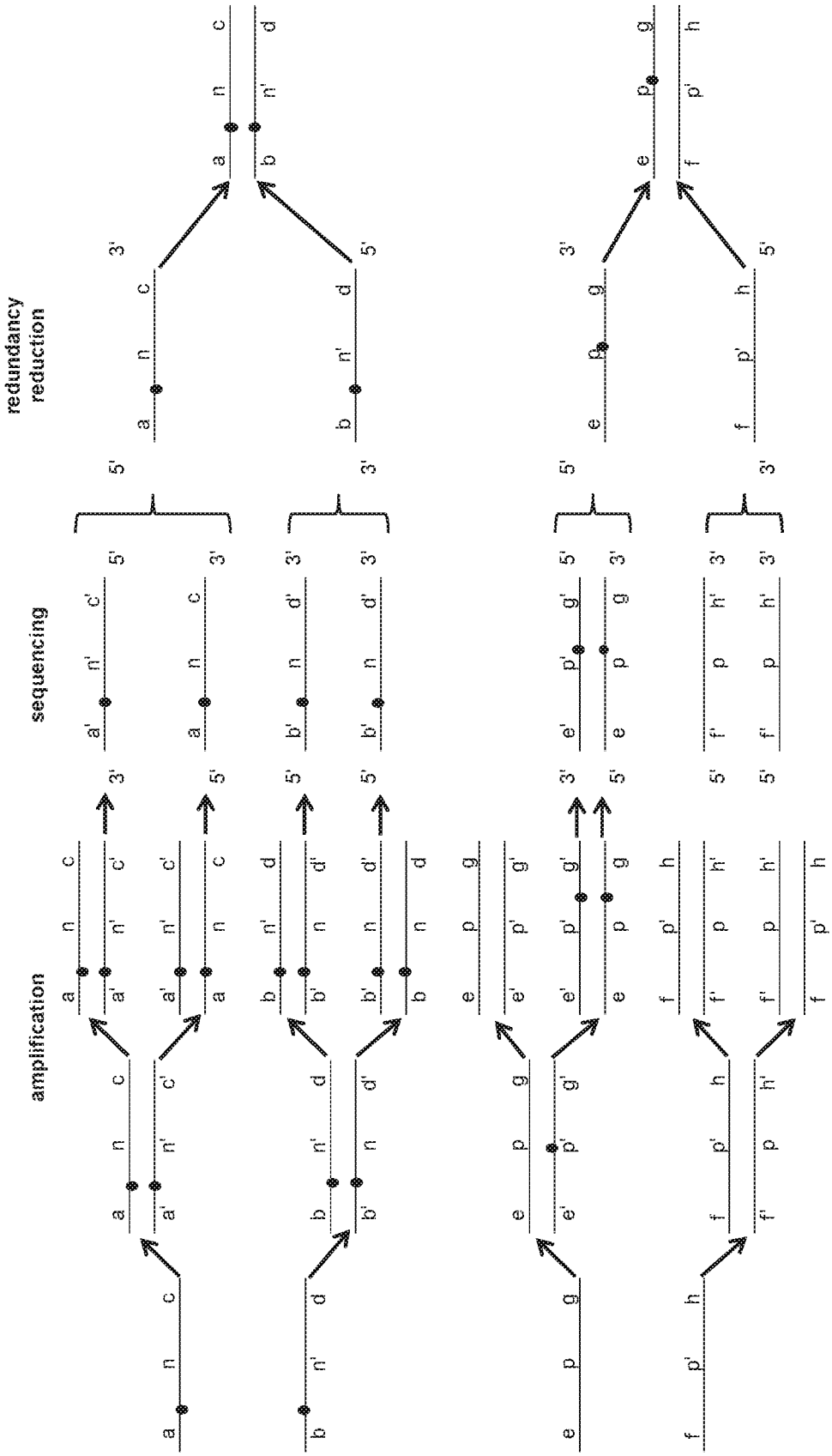


Fig. 5

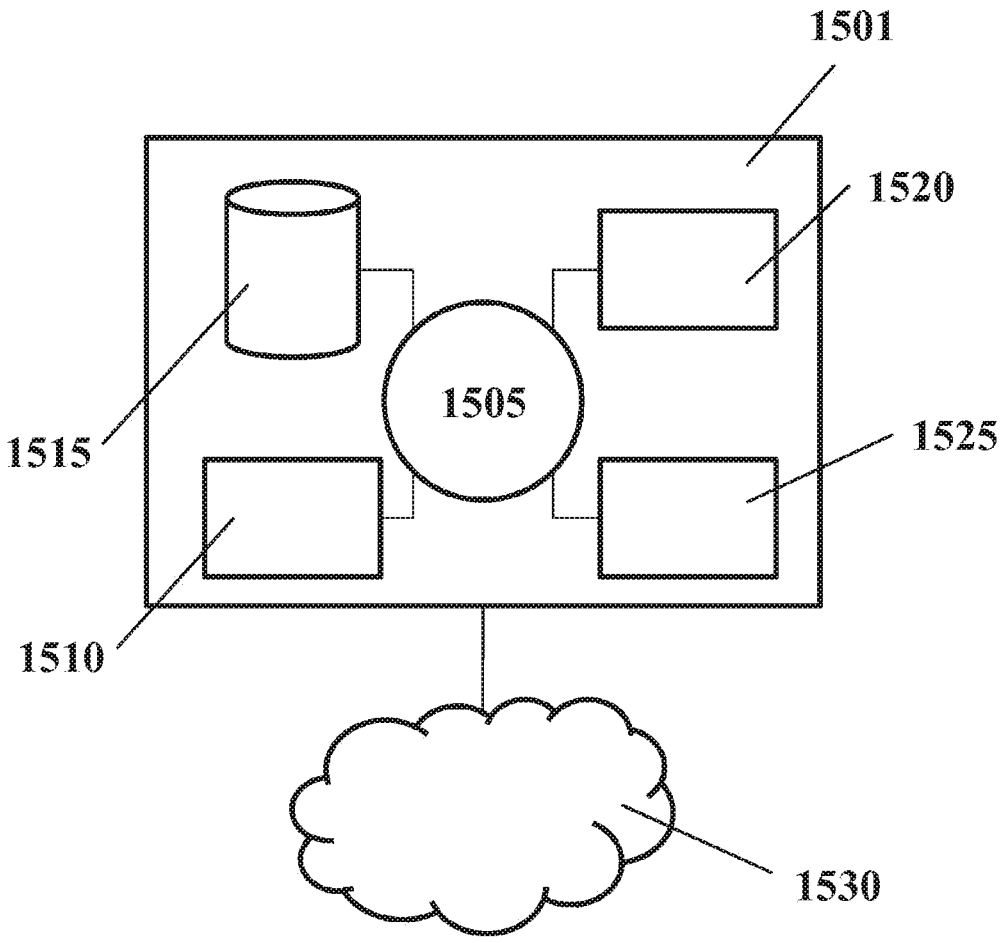


Fig. 6

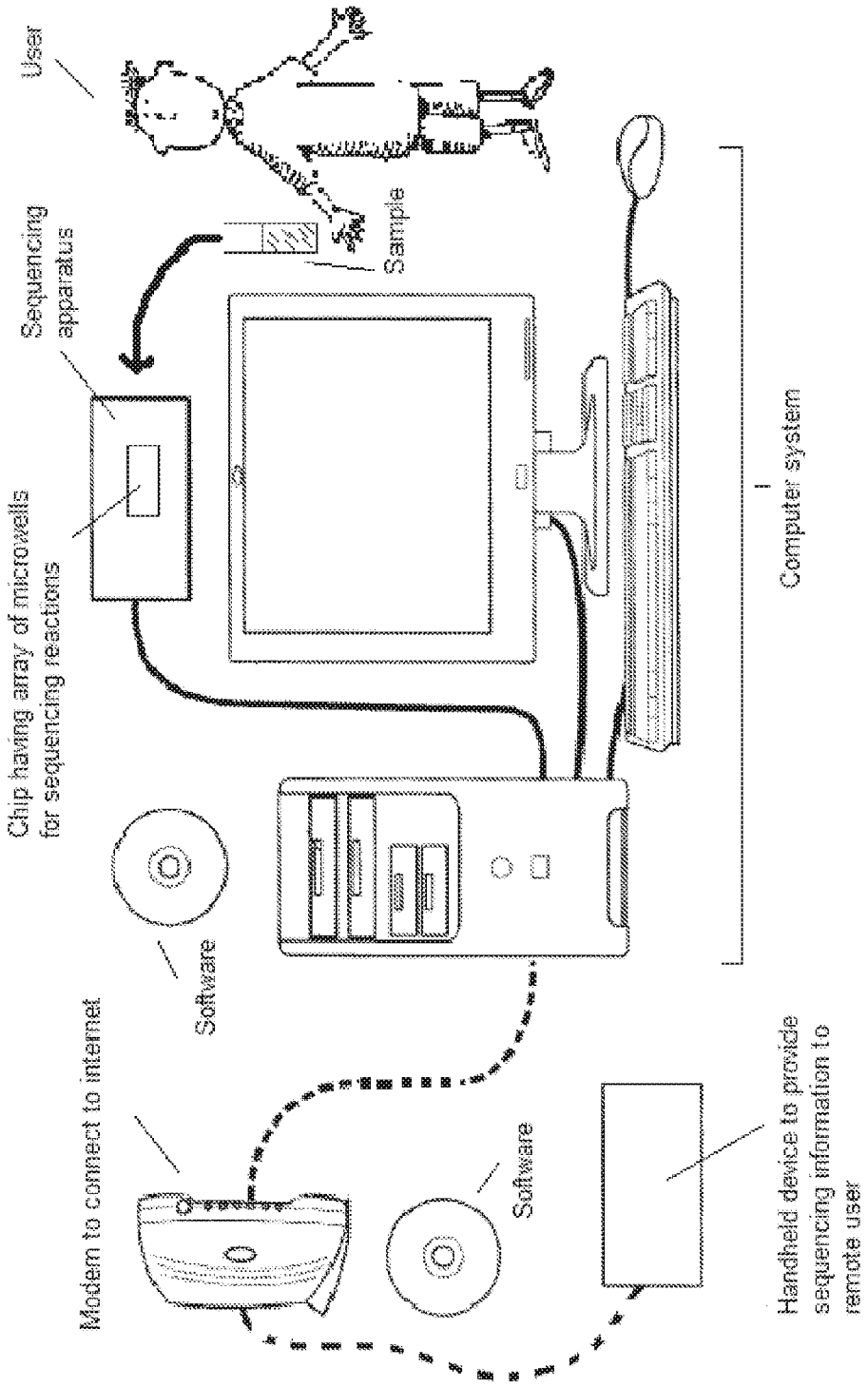

User
Sample
Sequencing apparatus
Chip having array of microwells for sequencing reactions
Computer system
Software
Modem to connect to internet
Software
Handheld device to provide sequencing information to remote user

Fig. 7

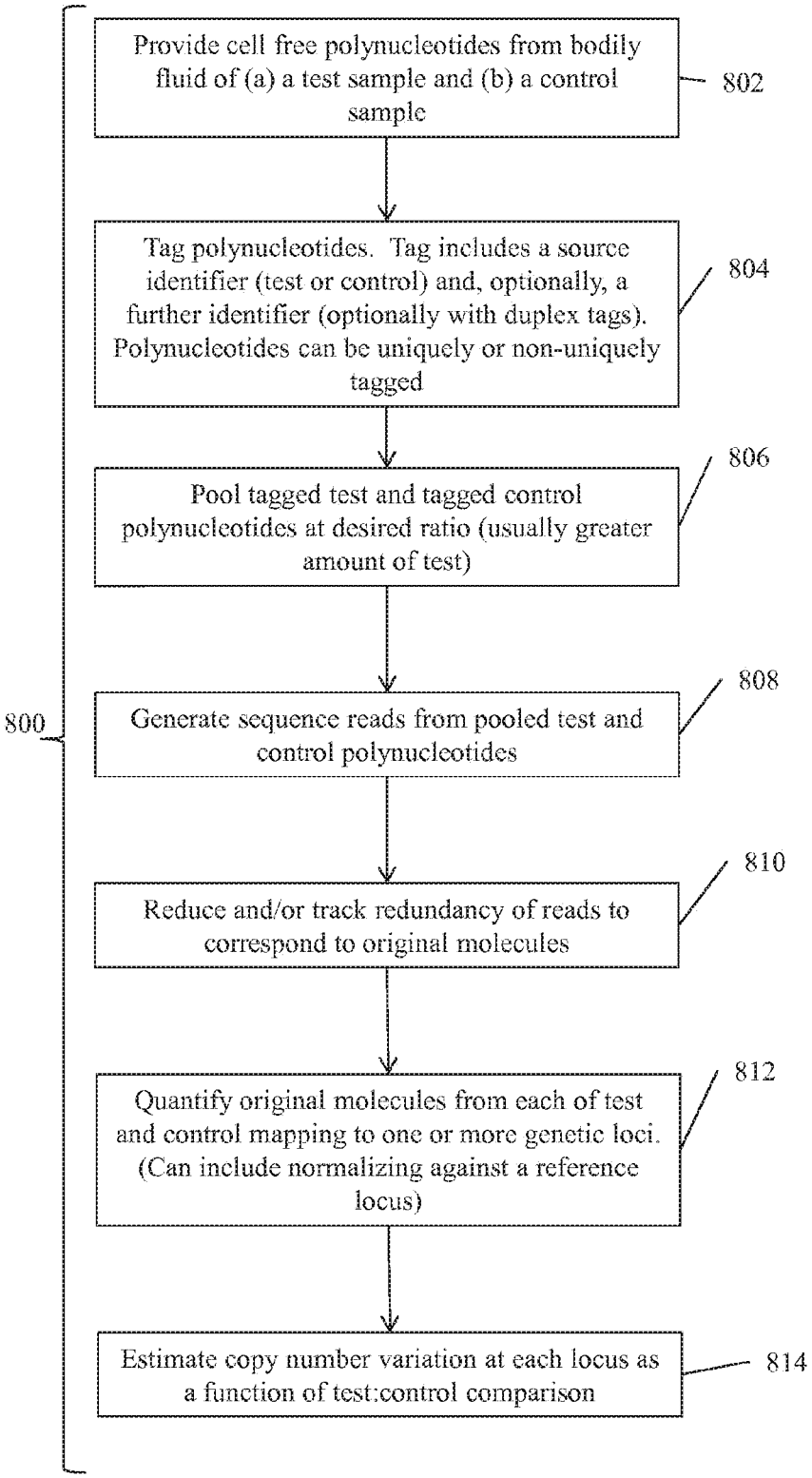

Provide cell free polynucleotides from bodily fluid of (a) a test sample and (b) a control sample
802
Tag polynucleotides. Tag includes a source identifier (test or control) and, optionally, a further identifier (optionally with duplex tags). Polynucleotides can be uniquely or non-uniquely tagged
804
Pool tagged test and tagged control polynucleotides at desired ratio (usually greater amount of test)
806
Generate sequence reads from pooled test and control polynucleotides
808
800
Reduce and/or track redundancy of reads to correspond to original molecules
810
Quantify original molecules from each of test and control mapping to one or more genetic loci. (Can include normalizing against a reference locus)
812
Estimate copy number variation at each locus as a function of test:control comparison
814

Fig. 8

Fig. 9A

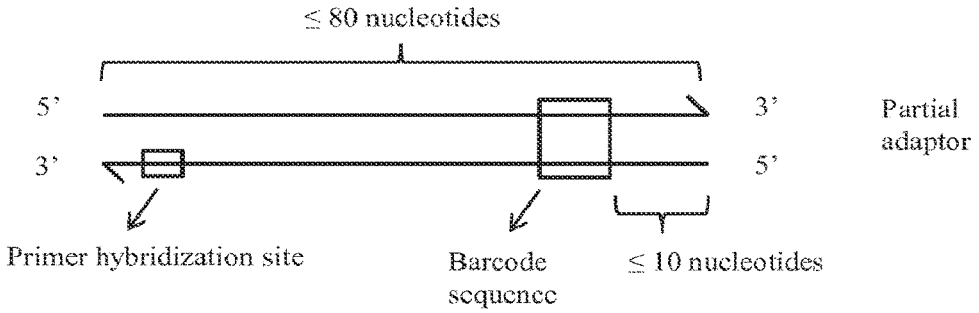


Fig. 9B

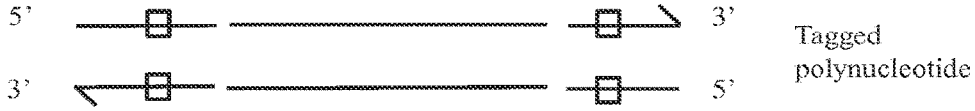
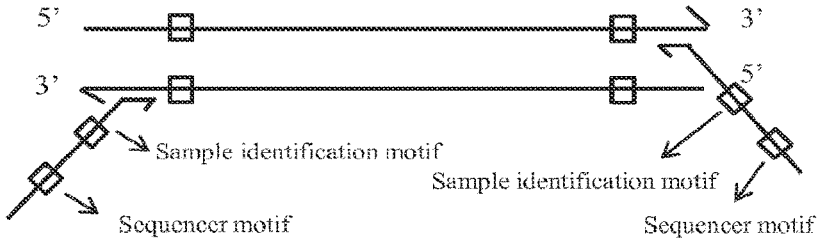


Fig. 9C



METHODS AND SYSTEMS FOR DETECTING GENETIC VARIANTS

CROSS-REFERENCE

This application is a continuation of U.S. application Ser. No. 15/892,178, filed Feb. 8, 2018, which is a continuation of U.S. application Ser. No. 14/861,989, filed Sep. 22, 2015 (now U.S. Pat. No. 9,920,366), which is a continuation application of International Application No. PCT/US2014/072383, filed Dec. 24, 2014, which application claims the benefit under 35 U.S.C. § 119(e) of U.S. Provisional Application No. 61/921,456, filed Dec. 28, 2013, and U.S. Provisional Application No. 61/948,509, filed Mar. 5, 2014, each of which is entirely incorporated herein by reference.

BACKGROUND

The detection and quantification of polynucleotides is important for molecular biology and medical applications, such as diagnostics. Genetic testing is particularly useful for a number of diagnostic methods. For example, disorders that are caused by rare genetic alterations (e.g., sequence variants) or changes in epigenetic markers, such as cancer and partial or complete aneuploidy, may be detected or more accurately characterized with DNA sequence information.

Early detection and monitoring of genetic diseases, such as cancer, is often useful and needed in the successful treatment or management of the disease. One approach may include the monitoring of a sample derived from cell-free nucleic acids, a population of polynucleotides that can be found in different types of bodily fluids. In some cases, disease may be characterized or detected based on detection of genetic aberrations, such as copy number variation and/or sequence variation of one or more nucleic acid sequences, or the development of other certain rare genetic alterations. Cell-free DNA (cfDNA) may contain genetic aberrations associated with a particular disease. With improvements in sequencing and techniques to manipulate nucleic acids, there is a need in the art for improved methods and systems for using cell-free DNA to detect and monitor disease.

In particular, many methods have been developed for accurate copy number variation estimation, especially for heterogeneous genomic samples, such as tumor-derived gDNA or for cfDNA for many applications (e.g., prenatal, transplant, immune, metagenomics or cancer diagnostics). Most of these methods include sample preparation whereby the original nucleic acids are converted into a sequenceable library, followed by massively parallel sequencing, and finally bioinformatics to estimate copy number variation at one or more loci.

SUMMARY

Although many of these methods are able to reduce or combat the errors introduced by the sample preparation and sequencing processes for all molecules that are converted and sequenced, these methods are not able to infer the counts of molecules that were converted but not sequenced. Since this count of converted by unsequenced molecules can be highly variable from genomic region to region, these counts can dramatically and adversely affect the sensitivity that can be achieved.

To address this issue, input double-stranded deoxyribonucleic acid (DNA) can be converted by a process that tags both halves of the individual double-stranded molecule, in some cases differently. This can be performed using a variety of techniques, including ligation of hairpin, bubble, or forked adapters or other adaptors having double-stranded and single stranded segments (the unhybridized portion of a bubble, forked or hairpin adapter are deemed single-stranded herein). If tagged correctly, each original Watson and Crick (i.e., strand) side of the input double-stranded DNA molecule can be differently tagged and identified by the sequencer and subsequent bioinformatics. For all molecules in a particular region, counts of molecules where both Watson and Crick sides were recovered ("Pairs") versus those where only one half was recovered ("Singlets") can be recorded. The number of unseen molecules can be estimated based on the number of Pairs and Singlets detected.

An aspect of the present disclosure provides a method for detecting and/or quantifying rare deoxyribonucleic acid (DNA) in a heterogeneous population of original DNA fragments, comprising tagging the original DNA fragments in a single reaction using a library of a plurality of different tags such that greater than 30% of the fragments are tagged at both ends, wherein each of the tags comprises a molecular barcode. The single reaction can be in a single reaction vessel. Greater than 50% of the fragments can be tagged at both ends. The plurality of different tags can be no more than any of 100, 500, 1000, 10,000 or 100,000 different tags.

Another aspect provides a set of library adaptors that can be used to tag the molecules of interest (e.g., by ligation, hybridization, etc.). The set of library adaptors can comprise plurality of polynucleotide molecules with molecular barcodes, wherein the plurality of polynucleotide molecules are less than or equal to 80 nucleotide bases in length, wherein the molecular barcodes are at least 4 nucleotide bases in length, and wherein (a) the molecular barcodes are different from one another and have an edit distance of at least 1 between one another; (b) the molecular barcodes are located at least one nucleotide base away from a terminal end of their respective polynucleotide molecules; (c) optionally, at least one terminal base is identical in all of the polynucleotide molecules; and (d) none of the polynucleotide molecules contains a complete sequencer motif.

In some embodiments, the library adaptors (or adapters) are identical to one another but for the molecular barcodes. In some embodiments, each of the plurality of library adaptors comprises at least one double-stranded portion and at least one single-stranded portion (e.g., a non-complementary portion or an overhang). In some embodiments, the double-stranded portion has a molecular barcode selected from a collection of different molecular barcodes. In some embodiments, the given molecular barcode is a randomer. In some embodiments, each of the library adaptors further comprises a strand-identification barcode on the at least one single-stranded portion. In some embodiments, the strand-identification barcode includes at least 4 nucleotide bases. In some embodiments, the single-stranded portion has a partial sequencer motif. In some embodiments, the library adaptors do not include a complete sequencer motif.

In some embodiments, none of the library adaptors contains a sequence for hybridizing to a flow cell or forming a hairpin for sequencing.

In some embodiments, all of the library adaptors have a terminal end with nucleotide(s) that are the same. In some embodiments, the identical terminal nucleotide(s) are over two or more nucleotide bases in length.

In some embodiments, each of the library adapters is Y-shaped, bubble shaped or hairpin shaped. In some embodiments, none of the library adapters contains a sample identification motif. In some embodiments, each of the library adapters comprises a sequence that is selectively

hybridizable to a universal primer. In some embodiments, each of the library adapters comprises a molecular barcode that is at least 5, 6, 7, 8, 9 and 10 nucleotide bases in length. In some embodiments, each of the library adapters is from 10 nucleotide bases to 80 in length, or 30 to 70 nucleotide bases in length, or 40 to 60 nucleotide bases in length. In some embodiments, at least 1, 2, 3, or 4 terminal bases are identical in all of the library adaptors. In some embodiments, at least 4 terminal bases are identical in all of the library adaptors.

In some embodiments, the edit distance of the molecular barcodes of the library adapters is a Hamming distance. In some embodiments, the edit distance is at least 1, 2, 3, 4 or 5. In some embodiments, the edit distance is with respect to individual bases of the plurality of polynucleotide molecules. In some embodiments, the molecular barcodes are located at least 10 nucleotide base away from a terminal end of an adapter. In some embodiments, the plurality of library adapters includes at least 2, 4, 6, 8, 10, 20, 30, 40 or 50 different molecular barcodes, or from 2-100, 4-80, 6-60 or 8-40 different molecular barcodes. In any of the embodiments herein, there are more polynucleotides (e.g., cfDNA fragments) to be tagged than there are different molecular barcodes such that the tagging is not unique.

In some embodiments, the terminal end of an adaptor is configured for ligation (e.g., to a target nucleic acid molecule). In some embodiments, the terminal end of an adaptor is a blunt end.

In some embodiments, the adaptors are purified and isolated. In some embodiments, the library comprises one or more non-naturally occurring bases.

In some embodiments, the polynucleotide molecules comprise a primer sequence positioned 5' with respect to the molecular barcodes.

In some embodiments, the set of library adaptors consists essentially of the plurality of polynucleotide molecules.

In another aspect, a method comprises (a) tagging a collection of polynucleotides with a plurality of polynucleotide molecules from a library of adaptors to create a collection of tagged polynucleotides; and (b) amplifying the collection of tagged polynucleotides in the presence of sequencing adaptors, wherein the sequencing adaptors have primers with nucleotide sequences that are selectively hybridizable to complementary sequences in the plurality of polynucleotide molecules. The library of adaptors may be as described above or elsewhere herein. In some embodiments, each of the sequencer adaptors further comprises an index tag, which can be a sample identification motif.

Another aspect, provides a method for detecting and/or quantifying rare DNA in a heterogeneous population of original DNA fragments, wherein the rare DNA has a concentration that is less than 1%, the method comprising (a) tagging the original DNA fragments in a single reaction such that greater than 30% of the original DNA fragments are tagged at both ends with library adaptors that comprise molecular barcodes, thereby providing tagged DNA fragments; (b) performing high-fidelity amplification on the tagged DNA fragments; (c) optionally, selectively enriching a subset of the tagged DNA fragments; (d) sequencing one or both strands of the tagged, amplified and optionally selectively enriched DNA fragments to obtain sequence reads comprising nucleotide sequences of the molecular barcodes and at least a portion of the original DNA fragments; (e) from the sequence reads, determining consensus reads that are representative of single-strands of the original DNA fragments; and (f) quantifying the consensus reads to detect and/or quantify the rare DNA at a specificity that is greater than 99.9%.

In some embodiments, (e) comprises comparing sequence reads having the same or similar molecular barcodes and the same or similar end of fragment sequences. In some embodiments, the comparing further comprises performing a phylogentic analysis on the sequence reads having the same or similar molecular barcodes. In some embodiments, the molecular barcodes include a barcode having an edit distance of up to 3. In some embodiments, the end of fragment sequence includes fragment sequences having an edit distance of up to 3.

In some embodiments, the method further comprises sorting sequence reads into paired reads and unpaired reads, and quantifying a number of paired reads and unpaired reads that map to each of one or more genetic loci.

In some embodiments, the tagging occurs by having an excess amount of library adaptors as compared to original DNA fragments. In some embodiments, n the excess is at least a 5-fold excess. In some embodiments, the tagging comprises using a ligase. In some embodiments, the tagging comprises attachment to blunt ends.

In some embodiments, the method further comprises binning the sequence reads according to the molecular barcodes and sequence information from at least one end of each of the original DNA fragments to create bins of single stranded reads. In some embodiments, the method further comprises, in each bin, determining a sequence of a given original DNA fragment among the original DNA fragments by analyzing sequence reads. In some embodiments, the method further comprises detecting and/or quantifying the rare DNA by comparing a number of times each base occurs at each position of a genome represented by the tagged, amplified, and optionally enriched DNA fragments.

In some embodiments, the library adaptors do not contain complete sequencer motifs. In some embodiments, the method further comprises selectively enriching a subset of the tagged DNA fragments. In some embodiments, the method further comprises, after enriching, amplifying the enriched tagged DNA fragments in the presence of sequencing adaptors comprising primers. In some embodiments, (a) provides tagged DNA fragments having from 2 to 1000 different combinations of molecular barcodes.

In some embodiments, the DNA fragments are tagged with polynucleotide molecules from a library of adaptors as described above or elsewhere herein.

In another aspect, a method for processing and/or analyzing a nucleic acid sample of a subject comprises (a) exposing polynucleotide fragments from the nucleic acid sample to a set of library adaptors to generate tagged polynucleotide fragments; and (b) subjecting the tagged polynucleotide fragments to nucleic acid amplification reactions under conditions that yield amplified polynucleotide fragments as amplification products of the tagged polynucleotide fragments. The set of library adaptors comprises a plurality of polynucleotide molecules with molecular barcodes, wherein the plurality of polynucleotide molecules are less than or equal to 80 nucleotide bases in length, wherein the molecular barcodes are at least 4 nucleotide bases in length, and wherein (1) the molecular barcodes are different from one another and have an edit distance of at least 1 between one another; (2) the molecular barcodes are located at least one nucleotide base away from a terminal end of their respective polynucleotide molecules; (3) optionally, at least one terminal base is identical in all of the polynucle-

otide molecules; and (4) none of the polynucleotide molecules contains a complete sequencer motif.

In some embodiments, the method further comprises determining nucleotide sequences of the amplified tagged polynucleotide fragments. In some embodiments, the nucleotide sequences of the amplified tagged polynucleotide fragments are determined without polymerase chain reaction (PCR). In some embodiments, the method further comprises analyzing the nucleotide sequences with a programmed computer processor to identify one or more genetic variants in the nucleotide sample of the subject. In some embodiments, the one or more genetic variants are selected from the group consisting of base change(s), insertion(s), repeat(s), deletion(s), copy number variation(s) and transversion(s). In some embodiments, the one or more genetic variants include one or more tumor associated genetic alterations.

In some embodiments, the subject has or is suspected of having a disease. In some embodiments, the disease is cancer. In some embodiments, the method further comprises collecting the nucleic acid sample from the subject. In some embodiments, the nucleic acid sample is collected from a location selected from the group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, cerebral spinal fluid and tears of the subject. In some embodiments, the nucleic acid sample is a cell-free nucleic acid sample. In some embodiments, the nucleic acid sample is collected from no more than 100 nanograms (ng) of double-stranded polynucleotide molecules of the subject.

In some embodiments, the polynucleotide fragments comprise double-stranded polynucleotide molecules. In some embodiments, in (a), the plurality of polynucleotide molecules couple to the polynucleotide fragments via blunt end ligation, sticky end ligation, molecular inversion probes, PCR, ligation-based PCR, multiplex PCR, single stranded ligation, and single stranded circularization. In some embodiments, exposing the polynucleotide fragments of the nucleic acid sample to the plurality of polynucleotide molecules yields the tagged polynucleotide fragments with a conversion efficiency of at least 10%. In some embodiments, any of at least 5%, 6%, 7%, 8%, 9%, 10%, 20%, or 25% of the tagged polynucleotide fragments share a common polynucleotide molecule or sequence. In some embodiments, the method further comprises generating the polynucleotide fragments from the nucleic acid sample.

In some embodiments, the subjecting comprises amplifying the tagged polynucleotide fragments from sequences corresponding to genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

In another aspect, a method comprises (a) generating a plurality of sequence reads from a plurality of polynucleotide molecules, wherein the plurality of polynucleotide molecules cover genomic loci of a target genome, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1; (b) grouping with a computer processor the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides; (c) for each of the families, merging sequence reads to generate a consensus sequence; (d) calling the consensus sequence at a given genomic locus among the genomic loci; and (e) detecting at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls, and total number of alterations among the calls.

In some embodiments, each family comprises sequence reads from only one of the template polynucleotides. In some embodiments, the given genomic locus comprises at least one nucleic acid base. In some embodiments, the given genomic locus comprises a plurality of nucleic acid bases. In some embodiments, the calling comprises calling at least one nucleic acid base at the given genomic locus. In some embodiments, the calling comprises calling a plurality of nucleic acid bases at the given genomic locus. In some embodiments, the calling comprises any one of phylogenetic analysis, voting, weighing, assigning a probability to each read at the locus in a family and calling the base with the highest probability.

In some embodiments, the method further comprises performing (d)-(e) at an additional genomic locus among the genomic loci. In some embodiments, the method further comprises determining a variation in copy number at one of the given genomic locus and additional genomic locus based on counts at the given genomic locus and additional genomic locus.

In some embodiments, the grouping comprises classifying the plurality of sequence reads into families by identifying (i) different molecular barcodes coupled to the plurality of polynucleotide molecules and (ii) similarities between the plurality of sequence reads, wherein each family includes a plurality of nucleic acid sequences that are associated with a different combination of molecular barcodes and similar or identical sequence reads. Different molecular barcodes have different sequences.

In some embodiments, the consensus sequence is generated by evaluating a quantitative measure or a statistical significance level for each of the sequence reads. In some embodiments, the quantitative measure comprises use of a binomial distribution, exponential distribution, beta distribution, or empirical distribution. In some embodiments, the method further comprises mapping the consensus sequence to the target genome. In some embodiments, the plurality of genes includes at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50 or all of the plurality of genes selected from the group.

Another aspect of the present disclosure provides a method, comprising (a) providing template polynucleotide molecules and a set of library adaptors in a single reaction vessel, wherein the library adaptors are polynucleotide molecules that have different molecular barcodes (e.g., from 2 to 1,000 different molecular barcodes), and wherein none of the library adaptors contains a complete sequencer motif; (b) in the single reaction vessel, coupling the library adaptors to the template polynucleotide molecules at an efficiency of at least 10%, thereby tagging each template polynucleotide

with a tagging combination that is among a plurality of different tagging combinations (e.g., 4 to 1,000,000 different tagging combinations), to produce tagged polynucleotide molecules; (c) subjecting the tagged polynucleotide molecules to an amplification reaction under conditions that yield amplified polynucleotide molecules as amplification products of the tagged polynucleotide molecules; and (d) sequencing the amplified polynucleotide molecules.

In some embodiments, the template polynucleotide molecules are blunt ended or sticky-ended. In some embodiments, the library adaptors are identical but for the molecular barcodes. In some embodiments, each of the library adaptors has a double stranded portion and at least one single-stranded portion. In some embodiments, the double-stranded portion has a molecular barcode among the molecular barcodes. In some embodiments, each of the library adaptors further comprises a strand-identification barcode on the at least one single-stranded portion. In some embodiments, the single-stranded portion has a partial sequencer motif. In some embodiments, the library adaptors have a sequence of terminal nucleotides that are the same. In some embodiments, the template polynucleotide molecules are double-stranded. In some embodiments, the library adaptors couple to both ends of the template polynucleotide molecules.

In some embodiments, subjecting the tagged polynucleotide molecules to the amplification reaction comprises non-specifically amplifying the tagged polynucleotide molecules.

In some embodiments, the amplification reaction comprises use of a priming site to amplify each of the tagged polynucleotide molecules. In some embodiments, the priming site is a primer. In some embodiments, the primer is a universal primer. In some embodiments, the priming site is a nick.

In some embodiments, the method further comprises, prior to (e), (i) separating polynucleotide molecules comprising one or more given sequences from the amplified polynucleotide molecules, to produce enriched polynucleotide molecules; and (ii) amplifying the enriched polynucleotide molecules with sequencing adaptors.

In some embodiments, the efficiency is at least 30%, 40%, or 50%. In some embodiments, the method further comprises identifying genetic variants upon sequencing the amplified polynucleotide molecules. In some embodiments, the sequencing comprises (i) subjecting the amplified polynucleotide molecules to an additional amplification reaction under conditions that yield additional amplified polynucleotide molecules as amplification products of the amplified polynucleotide molecules, and (ii) sequencing the additional amplified polynucleotide molecules. In some embodiments, the additional amplification is performed in the presence of sequencing adaptors.

In some embodiments, (b) and (c) are performed without aliquoting the tagged polynucleotide molecules. In some embodiments, the tagging is non-unique tagging.

Another aspect, provides a system for analyzing a target nucleic acid molecule of a subject, comprising a communication interface that receives nucleic acid sequence reads for a plurality of polynucleotide molecules that cover genomic loci of a target genome; computer memory that stores the nucleic acid sequence reads for the plurality of polynucleotide molecules received by the communication interface; and a computer processor operatively coupled to the communication interface and the memory and programmed to (i) group the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides, (ii) for each of the families, merge sequence reads to generate a consensus sequence, (iii) call the consensus sequence at a given genomic locus among the genomic loci, and (iv) detect at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls; and total number of alterations among the calls, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

In another aspect, a set of oligonucleotide molecules that selectively hybridize to at least 5 genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

In some embodiments, the oligonucleotide molecules are from 10-200 bases in length. In some embodiments, the oligonucleotide molecules selectively hybridize to exon regions of the at least 5 genes. In some embodiments, the oligonucleotide molecules selectively hybridize to at least 30 exons in the at least 5 genes. In some embodiments, multiple oligonucleotide molecules selectively hybridize to each of the at least 30 exons. In some embodiments, the oligonucleotide molecules that hybridize to each exon have sequences that overlap with at least 1 other oligonucleotide molecule.

In another aspect, a kit comprises a first container containing a plurality of library adaptors each having a different molecular barcode; and a second container containing a plurality of sequencing adaptors, each sequencing adaptor comprising at least a portion of a sequencer motif and optionally a sample barcode. The library adaptors can be as described above or elsewhere herein.

In some embodiments, the sequencing adaptor comprises the sample barcode. In some embodiments, the library adaptors are blunt ended and Y-shaped, and are less than or equal to 80 nucleic acid bases in length. In some embodiments, the sequencing adaptor is up to 70 bases from end to end.

In another aspect, a method for detecting sequence variants in a cell free DNA sample, comprising detecting rare DNA at a concentration less than 1% with a specificity that is greater than 99.9%.

In another aspect, a method comprises detecting genetic variants in a sample comprising DNA with a detection limit of at least 1% and specificity greater than 99.9%. In some embodiments, the method further comprises converting cDNA (e.g. cfDNA) into adaptor tagged DNA with a con-

version efficiency of at least 30%, 40%, or 50% and reducing sequencing noise (or distortion) by eliminating false positive sequence reads.

Another aspect provides a method, comprising (a) providing a sample comprising a set of double-stranded polynucleotide molecules, each double-stranded polynucleotide molecule including first and second complementary strands; (b) tagging the double-stranded polynucleotide molecules with a set of duplex tags, wherein each duplex tag differently tags the first and second complementary strands of a double-stranded polynucleotide molecule in the set; (c) sequencing at least some of the tagged strands to produce a set of sequence reads; (d) reducing and/or tracking redundancy in the set of sequence reads; (e) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; (f) determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci; and (g) estimating with a programmed computer processor a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

In some embodiments, the method further comprises (h) detecting copy number variation in the sample by determining a normalized total quantitative measure determined in step (g) at each of the one or more genetic loci and determining copy number variation based on the normalized measure. In some embodiments, the sample comprises double-stranded polynucleotide molecules sourced substantially from cell-free nucleic acids. In some embodiments, the duplex tags are not sequencing adaptors.

In some embodiments, reducing redundancy in the set of sequence reads comprises collapsing sequence reads produced from amplified products of an original polynucleotide molecule in the sample back to the original polynucleotide molecule. In some embodiments, the method further comprises determining a consensus sequence for the original polynucleotide molecule. In some embodiments, the method further comprises identifying polynucleotide molecules at one or more genetic loci comprising a sequence variant. In some embodiments, the method further comprises determining a quantitative measure of paired reads that map to a locus, wherein both strands of the pair comprise a sequence variant. In some embodiments, the method further comprises determining a quantitative measure of paired molecules in which only one member of the pair bears a sequence variant and/or determining a quantitative measure of unpaired molecules bearing a sequence variant. In some embodiments, the sequence variant is selected from the group consisting of a single nucleotide variant, an indel, a transversion, a translocation, an inversion, a deletion, a chromosomal structure alteration, a gene fusion, a chromosome fusion, a gene truncation, a gene amplification, a gene duplication and a chromosomal lesion.

Another aspect provides a system comprising a computer readable medium comprising machine-executable code that, upon execution by a computer processor, implements a method comprising (a) receiving into memory a set of sequence reads of polynucleotides tagged with duplex tags; (b) reducing and/or tracking redundancy in the set of sequence reads; (c) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; (d) determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci; and (e) estimating a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

Another aspect provides a method, comprising (a) providing a sample comprising a set of double-stranded polynucleotide molecules, each double-stranded polynucleotide molecule including first and second complementary strands; (b) tagging the double-stranded polynucleotide molecules with a set of duplex tags, wherein each duplex tag differently tags the first and second complementary strands of a double-stranded polynucleotide molecule in the set; (c) sequencing at least some of the tagged strands to produce a set of sequence reads; (d) reducing and/or tracking redundancy in the set of sequence reads; (e) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; and (f) determining quantitative measures of at least two of (i) the paired reads, (ii) the unpaired reads that map to each of one or more genetic loci, (iii) read depth of the paired reads and (iv) read depth of unpaired reads.

In some embodiments, (f) comprises determining quantitative measures of at least three of (i)-(iv). In some embodiments, (f) comprises determining quantitative measures of all of (i)-(iv). In some embodiments, the method further comprises (g) estimating with a programmed computer processor a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads and their read depths mapping to each locus.

In another aspect, a method comprises (a) tagging control parent polynucleotides with a first tag set to produce tagged control parent polynucleotides, wherein the first tag set comprises a plurality of tags, wherein each tag in the first tag set comprises a same control tag and an identifying tag, and wherein the tag set comprises a plurality of different identifying tags; (b) tagging test parent polynucleotides with a second tag set to produce tagged test parent polynucleotides, wherein the second tag set comprises a plurality of tags, wherein each tag in the second tag set comprises a same test tag that is distinguishable from the control tag and an identifying tag, and wherein the second tag set comprises a plurality of different identifying tags; (c) mixing tagged control parent polynucleotides with tagged test parent polynucleotides to form a pool; (d) amplifying tagged parent polynucleotides in the pool to form a pool of amplified, tagged polynucleotides; (e) sequencing amplified, tagged

polynucleotides in the amplified pool to produce a plurality of sequence reads; (f) grouping sequence reads into families, each family comprising sequence reads generated from a same parent polynucleotide, which grouping is optionally based on information from an identifying tag and from start/end sequences of the parent polynucleotides, and, optionally, determining a consensus sequence for each of a plurality of parent polynucleotides from the plurality of sequence reads in a group; (g) classifying each family or consensus sequence as a control parent polynucleotide or as a test parent polynucleotide based on having a test tag or a control tag; (h) determining a quantitative measure of control parent polynucleotides and control test polynucleotides mapping to each of at least two genetic loci; and (i) determining copy number variation in the test parent polynucleotides at at least one locus based on relative quantity of test parent polynucleotides and control parent polynucleotides mapping to the at least one locus.

In another aspect, a method comprises (a) generating a plurality of sequence reads from a plurality of template polynucleotides, each polynucleotide mapped to a genomic locus; (b) grouping the sequence reads into families, each family comprising sequence reads generated from one of the template polynucleotides; (c) calling a base (or sequence) at the genomic locus for each of the families; (d) detecting at the genomic locus any of genomic alterations among the calls, frequency of a genetic alteration among the calls, total number of calls and total number of alterations among the calls.

In some embodiments, calling comprises any of phylogenetic analysis, voting, weighing, assigning a probability to each read at the locus in a family, and calling the base with the highest probability. In some embodiments, the method is performed at two loci, comprising determining CNV at one of the loci based on counts at each of the loci.

Another aspect provides a method for determining a quantitative measure indicative of a number of individual double-stranded DNA fragments in a sample comprising (a) determining a quantitative measure of individual DNA molecules for which both strands are detected; (b) determining a quantitative measure of individual DNA molecules for which only one of the DNA strands are detected; (c) inferring from (a) and (b) above a quantitative measure of individual DNA molecules for which neither strand was detected; and (d) using (a)-(c) determining the quantitative measure indicative of a number of individual double-stranded DNA fragments in the sample.

In some embodiments, the method further comprises detecting copy number variation in the sample by determining a normalized quantitative measure determined in step (d) at each of one or more genetic loci and determining copy number variation based on the normalized measure. In some embodiments, the sample comprises double-stranded polynucleotide molecules sourced substantially from cell-free nucleic acids.

In some embodiments, determining the quantitative measure of individual DNA molecules comprises tagging the DNA molecules with a set of duplex tags, wherein each duplex tag differently tags complementary strands of a double-stranded DNA molecule in the sample to provide tagged strands. In some embodiments, the method further comprises sequencing at least some of the tagged strands to produce a set of sequence reads. In some embodiments, the method further comprises sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads. In some embodiments, the method further comprises determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci to determine a quantitative measure of total double-stranded DNA molecules in the sample that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

In another aspect, a method for reducing distortion in a sequencing assay, comprises (a) tagging control parent polynucleotides with a first tag set to produce tagged control parent polynucleotides; (b) tagging test parent polynucleotides with a second tag set to produce tagged test parent polynucleotides; (c) mixing tagged control parent polynucleotides with tagged test parent polynucleotides to form a pool; (d) determining quantities of tagged control parent polynucleotides and tagged test parent polynucleotides; and (e) using the quantities of tagged control parent polynucleotides to reduce distortion in the quantities of tagged test parent polynucleotides.

In some embodiments, the first tag set comprises a plurality of tags, wherein each tag in the first tag set comprises a same control tag and an identifying tag, and wherein the first tag set comprises a plurality of different identifying tags. In some embodiments, the second tag set comprises a plurality of tags, wherein each tag in the second tag set comprises a same test tag and an identifying tag, wherein the test tag is distinguishable from the control tag, and wherein the second tag set comprises a plurality of different identifying tags. In some embodiments, (d) comprises amplifying tagged parent polynucleotides in the pool to form a pool of amplified, tagged polynucleotides, and sequencing amplified, tagged polynucleotides in the amplified pool to produce a plurality of sequence reads. In some embodiments, the method further comprises grouping sequence reads into families, each family comprising sequence reads generated from a same parent polynucleotide, which grouping is optionally based on information from an identifying tag and from start/end sequences of the parent polynucleotides, and, optionally, determining a consensus sequence for each of a plurality of parent polynucleotides from the plurality of sequence reads in a group.

In some embodiments, (d) comprises determining copy number variation in the test parent polynucleotides at greater than or equal to one locus based on relative quantity of test parent polynucleotides and control parent polynucleotides mapping to the locus.

Another aspect provides a method comprising (a) ligating adaptors to double-stranded DNA polynucleotides, wherein ligating is performed in a single reaction vessel, and wherein the adaptors comprise molecular barcodes, to produce a tagged library comprising an insert from the double-stranded DNA polynucleotides, and having between 4 and 1 million different tags; (b) generating a plurality of sequence reads for each of the double-stranded DNA polynucleotides in the tagged library; (c) grouping sequence reads into families, each family comprising sequence reads generated from a single DNA polynucleotide among the double-stranded DNA polynucleotides, based on information in a tag and information at an end of the insert; and (d) calling bases at each position in the double-stranded DNA molecule based on bases at the position in members of a family. In some

embodiments, (b) comprises amplifying each of the double-stranded DNA polynucleotide molecules in the tagged library to generate amplification products, and sequencing the amplification products. In some embodiments, the method further comprises sequencing the double-stranded DNA polynucleotide molecules a plurality of times. In some embodiments, (b) comprises sequencing the entire insert. In some embodiments, (c) further comprises collapsing sequence reads in each family to generate a consensus sequence. In some embodiments, (d) comprises calling a plurality of sequential bases from at least a subset of the sequence reads to identify single nucleotide variations (SNV) in the double-stranded DNA molecule.

Another aspect provides a method of detecting disease cell heterogeneity from a sample comprising polynucleotides from somatic cells and disease cells. The method comprises quantifying polynucleotides in the sample bearing a nucleotide sequence variant at each of a plurality of genetic loci; determining copy number variation (CNV) at each of the plurality of genetic loci, wherein the CNV indicates a genetic dose of a locus in the disease cell polynucleotides; determining with a programmed computer processor a relative measure of quantity of polynucleotides bearing a sequence variant at a locus per the genetic dose at the locus for each of a plurality of the loci; and comparing the relative measures at each of the plurality of loci, wherein different relative measures is indicative of tumor heterogeneity.

In another aspect, a method comprises subjecting a subject to one or more pulsed therapy cycles, each pulsed therapy cycle comprising (a) a first period during which a drug is administered at a first amount; and (b) a second period during which the drug is administered at a second, reduced amount, wherein (i) the first period is characterized by a tumor burden detected above a first clinical level; and (ii) the second period is characterized by a tumor burden detected below a second clinical level.

Additional aspects and advantages of the present disclosure will become readily apparent to those skilled in this art from the following detailed description, wherein only illustrative embodiments of the present disclosure are shown and described. As will be realized, the present disclosure is capable of other and different embodiments, and its several details are capable of modifications in various obvious respects, all without departing from the disclosure. Accordingly, the drawings and description are to be regarded as illustrative in nature, and not as restrictive.

INCORPORATION BY REFERENCE

All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings (also "figure" and "FIG." herein), of which:

FIG. **1** is a flowchart representation of a method of the present disclosure for determining copy number variation (CNV);

FIG. **2** depicts mapping of pairs and singlets to Locus A and Locus B in a genome;

FIG. **3** shows a reference sequence encoding a genetic Locus A;

FIGS. **4**A-C shows amplification, sequencing, redundancy reduction and pairing of complementary molecules;

FIG. **5** shows increased confidence in detecting sequence variants by pairing reads from Watson and Crick strands;

FIG. **6** shows a computer system that is programmed or otherwise configured to implement various methods of the present disclosure;

FIG. **7** is schematic representation of a system for analyzing a sample comprising nucleic acids from a user, including a sequencer; bioinformatic software and internet connection for report analysis by, for example, a hand held device or a desk top computer;

FIG. **8** is a flowchart representation of a method of this invention for determining CNV using pooled test and control pools; and

FIGS. **9**A-**9**C schematically illustrate a method for tagging a polynucleotide molecule with a library adaptor and subsequently a sequencing adaptor.

DETAILED DESCRIPTION

While various embodiments of the invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions may occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed.

The term "genetic variant," as used herein, generally refers to an alteration, variant or polymorphism in a nucleic acid sample or genome of a subject. Such alteration, variant or polymorphism can be with respect to a reference genome, which may be a reference genome of the subject or other individual. Single nucleotide polymorphisms (SNPs) are a form of polymorphisms. In some examples, one or more polymorphisms comprise one or more single nucleotide variations (SNVs), insertions, deletions, repeats, small insertions, small deletions, small repeats, structural variant junctions, variable length tandem repeats, and/or flanking sequences. Copy number variants (CNVs), transversions and other rearrangements are also forms of genetic variation. A genomic alternation may be a base change, insertion, deletion, repeat, copy number variation, or transversion.

The term "polynucleotide," as used herein, generally refers to a molecule comprising one or more nucleic acid subunits. A polynucleotide can include one or more subunits selected from adenosine (A), cytosine (C), guanine (G), thymine (T) and uracil (U), or variants thereof. A nucleotide can include A, C, G, T or U, or variants thereof. A nucleotide can include any subunit that can be incorporated into a growing nucleic acid strand. Such subunit can be an A, C, G, T, or U, or any other subunit that is specific to one or more complementary A, C, G, T or U, or complementary to a purine (i.e., A or G, or variant thereof) or a pyrimidine (i.e., C, T or U, or variant thereof). A subunit can enable individual nucleic acid bases or groups of bases (e.g., AA, TA, AT, GC, CG, CT, TC, GT, TG, AC, CA, or uracil-counterparts thereof) to be resolved. In some examples, a polynucleotide is deoxyribonucleic acid (DNA) or ribonucleic

acid (RNA), or derivatives thereof. A polynucleotide can be single-stranded or double stranded.

The term "subject," as used herein, generally refers to an animal, such as a mammalian species (e.g., human) or avian (e.g., bird) species, or other organism, such as a plant. More specifically, the subject can be a vertebrate, a mammal, a mouse, a primate, a simian or a human. Animals include, but are not limited to, farm animals, sport animals, and pets. A subject can be a healthy individual, an individual that has or is suspected of having a disease or a pre-disposition to the disease, or an individual that is in need of therapy or suspected of needing therapy. A subject can be a patient.

The term "genome" generally refers to an entirety of an organism's hereditary information. A genome can be encoded either in DNA or in RNA. A genome can comprise coding regions that code for proteins as well as non-coding regions. A genome can include the sequence of all chromosomes together in an organism. For example, the human genome has a total of 46 chromosomes. The sequence of all of these together constitutes a human genome.

The terms "adaptor(s)", "adapter(s)" and "tag(s)" are used synonymously throughout this specification. An adaptor or tag can be coupled to a polynucleotide sequence to be "tagged" by any approach including ligation, hybridization, or other approaches.

The term "library adaptor" or "library adapter" as used herein, generally refers to a molecule (e.g., polynucleotide) whose identity (e.g., sequence) can be used to differentiate polynucleotides in a biological sample (also "sample" herein).

The term "sequencing adaptor," as used herein, generally refers to a molecule (e.g., polynucleotide) that is adapted to permit a sequencing instrument to sequence a target polynucleotide, such as by interacting with the target polynucleotide to enable sequencing. The sequencing adaptor permits the target polynucleotide to be sequenced by the sequencing instrument. In an example, the sequencing adaptor comprises a nucleotide sequence that hybridizes or binds to a capture polynucleotide attached to a solid support of a sequencing system, such as a flow cell. In another example, the sequencing adaptor comprises a nucleotide sequence that hybridizes or binds to a polynucleotide to generate a hairpin loop, which permits the target polynucleotide to be sequenced by a sequencing system. The sequencing adaptor can include a sequencer motif, which can be a nucleotide sequence that is complementary to a flow cell sequence of other molecule (e.g., polynucleotide) and usable by the sequencing system to sequence the target polynucleotide. The sequencer motif can also include a primer sequence for use in sequencing, such as sequencing by synthesis. The sequencer motif can include the sequence(s) needed to couple a library adaptor to a sequencing system and sequence the target polynucleotide.

As used herein the terms "at least", "at most" or "about", when preceding a series, refers to each member of the series, unless otherwise identified.

The term "about" and its grammatical equivalents in relation to a reference numerical value can include a range of values up to plus or minus 10% from that value. For example, the amount "about 10" can include amounts from 9 to 11. In other embodiments, the term "about" in relation to a reference numerical value can include a range of values plus or minus 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1% from that value.

The term "at least" and its grammatical equivalents in relation to a reference numerical value can include the reference numerical value and greater than that value. For example, the amount "at least 10" can include the value 10 and any numerical value above 10, such as 11, 100, and 1,000.

The term "at most" and its grammatical equivalents in relation to a reference numerical value can include the reference numerical value and less than that value. For example, the amount "at most 10" can include the value 10 and any numerical value under 10, such as 9, 8, 5, 1, 0.5, and 0.1.

1. Methods for Processing and/or Analyzing a Nucleic Acid Sample

An aspect of the present disclosure provides methods for determining a genomic alternation in a nucleic acid sample of a subject. FIG. **1** shows a method of determining copy number variation (CNV). The method can be implemented to determine other genomic alternations, such as SNVs.

A. Polynucleotide Isolation

Methods disclosed herein can comprise isolating one or more polynucleotides. A polynucleotide can comprise any type of nucleic acid, for example, a sequence of genomic nucleic acid, or an artificial sequence (e.g., a sequence not found in genomic nucleic acid). For example, an artificial sequence can contain non-natural nucleotides. Also, a polynucleotide can comprise both genomic nucleic acid and an artificial sequence, in any portion. For example, a polynucleotide can comprise 1 to 99% of genomic nucleic acid and 99% to 1% of artificial sequence, where the total adds up to 100%. Thus, fractions of percentages are also contemplated. For example, a ratio of 99.1% to 0.9% is contemplated.

A polynucleotide can comprise any type of nucleic acids, such as DNA and/or RNA. For example, if a polynucleotide is DNA, it can be genomic DNA, complementary DNA (cDNA), or any other deoxyribonucleic acid. A polynucleotide can also be cell-free DNA (cfDNA). For example, the polynucleotide can be circulating DNA. The circulating DNA can comprise circulating tumor DNA (ctDNA). A polynucleotide can be double-stranded or single-stranded. Alternatively, a polynucleotide can comprise a combination of a double-stranded portion and a single-stranded portion.

Polynucleotides do not have to be cell-free. In some cases, the polynucleotides can be isolated from a sample. For example, in step (**102**) (FIG. **1**), double-stranded polynucleotides are isolated from a sample. A sample can be any biological sample isolated from a subject. For example, a sample can comprise, without limitation, bodily fluid, whole blood, platelets, serum, plasma, stool, red blood cells, white blood cells or leucocytes, endothelial cells, tissue biopsies, synovial fluid, lymphatic fluid, ascites fluid, interstitial or extracellular fluid, the fluid in spaces between cells, including gingival crevicular fluid, bone marrow, cerebrospinal fluid, saliva, mucous, sputum, semen, sweat, urine, or any other bodily fluids. A bodily fluid can include saliva, blood, or serum. For example, a polynucleotide can be cell-free DNA isolated from a bodily fluid, e.g., blood or serum. A sample can also be a tumor sample, which can be obtained from a subject by various approaches, including, but not limited to, venipuncture, excretion, ejaculation, massage, biopsy, needle aspirate, lavage, scraping, surgical incision, or intervention or other approaches.

A sample can comprise various amount of nucleic acid that contains genome equivalents. For example, a sample of about 30 ng DNA can contain about 10,000 (10^4) haploid human genome equivalents and, in the case of cfDNA, about 200 billion (2×10^{11}) individual polynucleotide molecules. Similarly, a sample of about 100 ng of DNA can contain about 30,000 haploid human genome equivalents and, in the case of cfDNA, about 600 billion individual molecules.

A sample can comprise nucleic acids from different sources. For example, a sample can comprise germline DNA or somatic DNA. A sample can comprise nucleic acids carrying mutations. For example, a sample can comprise DNA carrying germline mutations and/orsomatic mutations. A sample can also comprise DNA carrying cancer-associated mutations (e.g., cancer-associated somatic mutations).

B. Tagging

Polynucleotides disclosed herein can be tagged. For example, in step (**104**) (FIG. **1**) the double-stranded polynucleotides are tagged with duplex tags, tags that differently label the complementary strands (i.e., the "Watson" and "Crick" strands) of a double-stranded molecule. In one embodiment the duplex tags are polynucleotides having complementary and non-complementary portions.

Tags can be any types of molecules attached to a polynucleotide, including, but not limited to, nucleic acids, chemical compounds, florescent probes, or radioactive probes. Tags can also be oligonucleotides (e.g., DNA or RNA). Tags can comprise known sequences, unknown sequences, or both. A tag can comprise random sequences, pre-determined sequences, or both. A tag can be double-stranded or single-stranded. A double-stranded tag can be a duplex tag. A double-stranded tag can comprise two complementary strands. Alternatively, a double-stranded tag can comprise a hybridized portion and a non-hybridized portion. The double-stranded tag can be Y-shaped, e.g., the hybridized portion is at one end of the tag and the non-hybridized portion is at the opposite end of the tag. One such example are the "Y adapters" used in Illumina sequencing. Other examples include hairpin shaped adapters or bubble shaped adapters. Bubble shaped adapters have non-complementary sequences flanked on both sides by complementary sequences.

Tagging disclosed herein can be performed using any method. A polynucleotide can be tagged with an adaptor by hybridization. For example, the adaptor can have a nucleotide sequence that is complementary to at least a portion of a sequence of the polynucleotide. As an alternative, a polynucleotide can be tagged with an adaptor by ligation.

For example, tagging can comprise using one or more enzymes. The enzyme can be a ligase. The ligase can be a DNA ligase. For example, the DNA ligase can be a T4 DNA ligase, *E. coli* DNA ligase, and/or mammalian ligase. The mammalian ligase can be DNA ligase I, DNA ligase III, or DNA ligase IV. The ligase can also be a thermostable ligase. Tags can be ligated to a blunt-end of a polynucleotide (blunt-end ligation). Alternatively, tags can be ligated to a sticky end of a polynucleotide (sticky-end ligation). Efficiency of ligation can be increased by optimizing various conditions. Efficiency of ligation can be increased by optimizing the reaction time of ligation. For example, the reaction time of ligation can be less than 12 hours, e.g., less than 1, less than 2, less than 3, less than 4, less than 5, less than 6, less than 7, less than 8, less than 9, less than 10, less than 11, less than 12, less than 13, less than 14, less than 15, less than 16, less than 17, less than 18, less than 19, or less than 20 hours. In a particular example, reaction time of ligation is less than 20 hours. Efficiency of ligation can be increased by optimizing the ligase concentration in the reaction. For example, the ligase concentration can be at least 10, at least 50, at least 100, at least 150, at least 200, at least 250, at least 300, at least 400, at least 500, or at least 600 unit/microliter. Efficiency can also be optimized by adding or varying the concentration of an enzyme suitable for ligation, enzyme cofactors or other additives, and/or optimizing a temperature of a solution having the enzyme. Efficiency can also be optimized by varying the addition order of various components of the reaction. The end of tag sequence can comprise dinucleotide to increase ligation efficiency. When the tag comprises a non-complementary portion (e.g., Y-shaped adaptor), the sequence on the complementary portion of the tag adaptor can comprise one or more selected sequences that promote ligation efficiency. Preferably such sequences are located at the terminal end of the tag. Such sequences can comprise 1, 2, 3, 4, 5, or 6 terminal bases. Reaction solution with high viscosity (e.g., a low Reynolds number) can also be used to increase ligation efficiency. For example, solution can have a Reynolds number less than 3000, less than 2000, less than 1000, less than 900, less than 800, less than 700, less than 600, less than 500, less than 400, less than 300, less than 200, less than 100, less than 50, less than 25, or less than 10. It is also contemplated that roughly unified distribution of fragments (e.g., tight standard deviation) can be used to increase ligation efficiency. For example, the variation in fragment sizes can vary by less than 20%, less than 15%, less than 10%, less than 5%, or less than 1%. Tagging can also comprise primer extension, for example, by polymerase chain reaction (PCR). Tagging can also comprise any of ligation-based PCR, multiplex PCR, single strand ligation, or single strand circularization.

In some instances, the tags herein comprise molecular barcodes. Such molecular barcodes can be used to differentiate polynucleotides in a sample. Preferably molecular barcodes are different from one another. For example, molecular barcodes can have a difference between them that can be characterized by a predetermined edit distance or a Hamming distance. In some instances, the molecular barcodes herein have a minimum edit distance of 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10. To further improve efficiency of conversion (e.g., tagging) of untagged molecular to tagged molecules, one preferably utilizes short tags. For example, in some embodiments, a library adapter tag can be up to 65, 60, 55, 50, 45, 40, or 35 nucleotide bases in length. A collection of such short library barcodes preferably includes a number of different molecular barcodes, e.g., at least 2, 4, 6, 8, 10, 12, 14, 16, 18 or 20 different barcodes with a minimum edit distance of 1, 2, 3 or more.

Thus, a collection of molecules can include one or more tags. In some instances, some molecules in a collection can include an identifying tag ("identifier") such as a molecular barcode that is not shared by any other molecule in the collection. For example, in some instances of a collection of molecules, at least 50%, at least 51%, at least 52%, at least 53%, at least 54%, at least 55%, at least 56%, at least 57%, at least 58%, at least 59%, at least 60%, at least 61%, at least 62%, at least 63%, at least 64%, at least 65%, at least 66%, at least 67%, at least 68%, at least 69%, at least 70%, at least 71%, at least 72%, at least 73%, at least 74%, at least 75%, at least 76%, at least 77%, at least 78%, at least 79%, at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% of the molecules in the collection can include an identifier or molecular barcode that is not shared by any other molecule in the collection. As used herein, a collection of molecules is considered to be "uniquely tagged" if each of at least 95% of the molecules in the collection bears an identifier that is not shared by any other molecule in the collection ("unique tag" or "unique identifier"). A collection of molecules is considered to be "non-uniquely tagged" if each of at least 1%, at least 5%, at

least 10%, at least 15%, at least 20%, at least 25%, at least 30%, at least 35%, at least 40%, at least 45%, or at least or about 50% of the molecules in the collection bears an identifying tag or molecular barcode that is shared by at least one other molecule in the collection ("non-unique tag" or "non-unique identifier"). Accordingly, in a non-uniquely tagged population no more than 1% of the molecules are uniquely tagged. For example, in a non-uniquely tagged population, no more than 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, or 50% of the molecules can be uniquely tagged.

A number of different tags can be used based on the estimated number of molecules in a sample. In some tagging methods, the number of different tags can be at least the same as the estimated number of molecules in the sample. In other tagging methods, the number of different tags can be at least two, three, four, five, six, seven, eight, nine, ten, one hundred or one thousand times as many as the estimated number of molecules in the sample. In unique tagging, at least two times (or more) as many different tags can be used as the estimated number of molecules in the sample.

The molecules in the sample may be non-uniquely tagged. In such instances a fewer number of tags or molecular barcodes is used then the number of molecules in the sample to be tagged. For example, no more than 100, 50, 40, 30, 20 or 10 unique tags or molecular barcodes are used to tag a complex sample such as a cell free DNA sample with many more different fragments.

The polynucleotide to be tagged can be fragmented, such as either naturally or using other approaches, such as, for example, shearing. The polynucleotides can be fragmented by certain methods, including but not limited to, mechanical shearing, passing the sample through a syringe, sonication, heat treatment (e.g., for 30 minutes at 90° C.), and/or nuclease treatment (e.g., using DNase, RNase, endonuclease, exonuclease, and/or restriction enzyme).

The polynucleotides fragments (prior to tagging) can comprise sequences of any length. For example, polynucleotide fragments (prior to tagging) can comprise at least 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, 275, 280, 285, 290, 295, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000 or more nucleotides in length. The polynucleotide fragment are preferably about the average length of cell-free DNA. For example, the polynucleotide fragments can comprise about 160 bases in length. The polynucleotide fragment can also be fragmented from a larger fragment into smaller fragments about 160 bases in length.

Polynucleotides tagged can comprise sequences associated with cancer. The cancer-associated sequences can comprise single nucleotide variation (SNV), copy number variation (CNV), insertions, deletions, and/or rearrangements.

The polynucleotides can comprise sequences associated with cancer, such as acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), adrenocortical carcinoma, Kaposi Sarcoma, anal cancer, basal cell carcinoma, bile duct cancer, bladder cancer, bone cancer, osteosarcoma, malignant fibrous histiocytoma, brain stem glioma, brain cancer, craniopharyngioma, ependymoblastoma, ependymoma, medulloblastoma, medulloepithelioma, pineal parenchymal tumor, breast cancer, bronchial tumor, Burkitt lymphoma, Non-Hodgkin lymphoma, carcinoid tumor, cervical cancer, chordoma, chronic lymphocytic leukemia (CLL), chronic myelogenous leukemia (CML), colon cancer, colorectal cancer, cutaneous T-cell lymphoma, ductal carcinoma in situ, endometrial cancer, esophageal cancer, Ewing Sarcoma, eye cancer, intraocular melanoma, retinoblastoma, fibrous histiocytoma, gallbladder cancer, gastric cancer, glioma, hairy cell leukemia, head and neck cancer, heart cancer, hepatocellular (liver) cancer, Hodgkin lymphoma, hypopharyngeal cancer, kidney cancer, laryngeal cancer, lip cancer, oral cavity cancer, lung cancer, non-small cell carcinoma, small cell carcinoma, melanoma, mouth cancer, myelodysplastic syndromes, multiple myeloma, medulloblastoma, nasal cavity cancer, paranasal sinus cancer, neuroblastoma, nasopharyngeal cancer, oral cancer, oropharyngeal cancer, osteosarcoma, ovarian cancer, pancreatic cancer, papillomatosis, paraganglioma, parathyroid cancer, penile cancer, pharyngeal cancer, pituitary tumor, plasma cell neoplasm, prostate cancer, rectal cancer, renal cell cancer, rhabdomyosarcoma, salivary gland cancer, Sezary syndrome, skin cancer, nonmelanoma, small intestine cancer, soft tissue sarcoma, squamous cell carcinoma, testicular cancer, throat cancer, thymoma, thyroid cancer, urethral cancer, uterine cancer, uterine sarcoma, vaginal cancer, vulvar cancer, Waldenstrom macroglobulinemia, and/or Wilms Tumor.

A haploid human genome equivalent has about 3 picograms of DNA. A sample of about 1 microgram of DNA contains about 300,000 haploid human genome equivalents. Improvements in sequencing can be achieved as long as at least some of the duplicate or cognate polynucleotides bear unique identifiers with respect to each other, that is, bear different tags. However, in certain embodiments, the number of tags used is selected so that there is at least a 95% chance that all duplicate molecules starting at any one position bear unique identifiers. For example, in a sample comprising about 10,000 haploid human genome equivalents of fragmented genomic DNA, e.g., cfDNA, z is expected to be between 2 and 8. Such a population can be tagged with between about 10 and 100 different identifiers, for example, about 2 identifiers, about 4 identifiers, about 9 identifiers, about 16 identifiers, about 25 identifiers, about 36 different identifiers, about 49 different identifiers, about 64 different identifiers, about 81 different identifiers, or about 100 different identifiers.

Nucleic acid barcodes having identifiable sequences including molecular barcodes, can be used for tagging. For example, a plurality of DNA barcodes can comprise various numbers of sequences of nucleotides. A plurality of DNA barcodes having 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more identifiable sequences of nucleotides can be used. When attached to only one end of a polynucleotide, the plurality of DNA barcodes can produce 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more different identifiers. Alternatively, when attached to both ends of a polynucleotide, the plurality DNA barcodes can produce 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 324, 361, 400 or more different identifiers (which is the ^2 of when the DNA barcode is attached to only 1 end of a polynucleotide). In one example, a plurality of DNA barcodes having 6, 7, 8, 9 or 10 identifiable sequences of nucleotides can be used. When attached to both ends of a polynucleotide, they produce 36, 49, 64, 81 or 100 possible different identifiers, respectively. In a particular example, the plurality of DNA barcodes can comprise 8 identifiable sequences of nucleotides. When attached to only one end of a polynucleotide, the plurality of DNA barcodes can produce 8 different identifiers. Alternatively, when attached to both ends of a polynucleotide, the

plurality of DNA barcodes can produce 64 different identifiers. Samples tagged in such a way can be those with a range of about 10 ng to any of about 100 ng, about 1 about 10 μg of fragmented polynucleotides, e.g., genomic DNA, e.g., cfDNA.

A polynucleotide can be uniquely identified in various ways. A polynucleotide can be uniquely identified by a unique DNA barcode. For example, any two polynucleotides in a sample are attached two different DNA barcodes. Alternatively, a polynucleotide can be uniquely identified by the combination of a DNA barcode and one or more endogenous sequences of the polynucleotide. For example, any two polynucleotides in a sample can be attached the same DNA barcode, but the two polynucleotides can still be identified by different endogenous sequences. The endogenous sequence can be on an end of a polynucleotide. For example, the endogenous sequence can be adjacent (e.g., base in between) to the attached DNA barcode. In some instances the endogenous sequence can be at least 2, 4, 6, 8, 10, 20, 30, 40, 50, 60, 70, 80, 90, or 100 bases in length. Preferably, the endogenous sequence is a terminal sequence of the fragment/polynucleotides to be analyzed. The endogenous sequence may be the length of the sequence. For example, a plurality of DNA barcodes comprising 8 different DNA barcodes can be attached to both ends of each polynucleotide in a sample. Each polynucleotide in the sample can be identified by the combination of the DNA barcodes and about 10 base pair endogenous sequence on an end of the polynucleotide. Without being bound by theory, the endogenous sequence of a polynucleotide can also be the entire polynucleotide sequence.

Also disclosed herein are compositions of tagged polynucleotides. The tagged polynucleotide can be single-stranded. Alternatively, the tagged polynucleotide can be double-stranded (e.g., duplex-tagged polynucleotides). Accordingly, this invention also provides compositions of duplex-tagged polynucleotides. The polynucleotides can comprise any types of nucleic acids (DNA and/or RNA). The polynucleotides comprise any types of DNA disclosed herein. For example, the polynucleotides can comprise DNA, e.g., fragmented DNA or cfDNA. A set of polynucleotides in the composition that map to a mappable base position in a genome can be non-uniquely tagged, that is, the number of different identifiers can be at least 2 and fewer than the number of polynucleotides that map to the mappable base position. The number of different identifiers can also be at least 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 and fewer than the number of polynucleotides that map to the mappable base position.

In some instances, as a composition goes from about 1 ng to about 10 μg or higher, a larger set of different molecular barcodes can be used. For example, between 5 and 100 different library adaptors can be used to tag polynucleotides in a cfDNA sample.

The systems and methods disclosed herein may be used in applications that involve the assignment of molecular barcodes. The molecular barcodes can be assigned to any types of polynucleotides disclosed in this invention. For example, the molecular barcodes can be assigned to cell-free polynucleotides (e.g., cfDNAs). Often, an identifier disclosed herein can be a barcode oligonucleotide that is used to tag the polynucleotide. The barcode identifier may be a nucleic acid oligonucleotide (e.g., a DNA oligonucleotide). The barcode identifier can be single-stranded. Alternatively, the barcode identifier can be double-stranded. The barcode identifier can be attached to polynucleotides using any method disclosed herein. For example, the barcode identifier can be attached to the polynucleotide by ligation using an enzyme. The barcode identifier can also be incorporated into the polynucleotide through PCR. In other cases, the reaction may comprise addition of a metal isotope, either directly to the analyte or by a probe labeled with the isotope. Generally, assignment of unique or non-unique identifiers or molecular barcodes in reactions of this disclosure may follow methods and systems described by, for example, U.S. patent applications 2001/0053519, 2003/0152490, 2011/0160078 and U.S. Pat. No. 6,582,908, each of which is entirely incorporated herein by reference.

Identifiers or molecular barcodes used herein may be completely endogenous whereby circular ligation of individual fragments may be performed followed by random shearing or targeted amplification. In this case, the combination of a new start and stop point of the molecule and the original intramolecular ligation point can form a specific identifier.

Identifiers or molecular barcodes used herein can comprise any types of oligonucleotides. In some cases, identifiers may be predetermined, random, or semi-random sequence oligonucleotides. Identifiers can be barcodes. For example, a plurality of barcodes may be used such that barcodes are not necessarily unique to one another in the plurality. Alternatively, a plurality of barcodes may be used such that each barcode is unique to any other barcode in the plurality. The barcodes can comprise specific sequences (e.g., predetermined sequences) that can be individually tracked. Further, barcodes may be attached (e.g., by ligation) to individual molecules such that the combination of the barcode and the sequence it may be ligated to creates a specific sequence that may be individually tracked. As described herein, detection of barcodes in combination with sequence data of beginning (start) and/or end (stop) portions of sequence reads can allow assignment of a unique identity to a particular molecule. The length or number of base pairs of an individual sequence read may also be used to assign a unique identity to such a molecule. As described herein, fragments from a single strand of nucleic acid having been assigned a unique identity, may thereby permit subsequent identification of fragments from the parent strand. In this way the polynucleotides in the sample can be uniquely or substantially uniquely tagged. A duplex tag can include a degenerate or semi-degenerate nucleotide sequence, e.g., a random degenerate sequence. The nucleotide sequence can comprise any number of nucleotides. For example, the nucleotide sequence can comprise 1 (if using a non-natural nucleotide), 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more nucleotides. In a particular example, the sequence can comprise 7 nucleotides. In another example, the sequence can comprise 8 nucleotides. The sequence can also comprise 9 nucleotides. The sequence can comprise 10 nucleotides.

A barcode can comprise contiguous or non-contiguous sequences. A barcode that comprises at least 1, 2, 3, 4, 5 or more nucleotides is a contiguous sequence or non-contiguous sequence. if the 4 nucleotides are uninterrupted by any other nucleotide. For example, if a barcode comprises the sequence TTGC, a barcode is contiguous if the barcode is TTGC. On the other hand, a barcode is non-contiguous if the barcode is TTXGC, where X is a nucleic acid base.

An identifier or molecular barcode can have an n-mer sequence which may be 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45,

46, 47, 48, 49, 50 or more nucleotides in length. A tag herein can comprise any range of nucleotides in length. For example, the sequence can be between 2 to 100, 10 to 90, 20 to 80, 30 to 70, 40 to 60, or about 50 nucleotides in length.

The tag can comprise a double-stranded fixed reference sequence downstream of the identifier or molecular barcode. Alternatively, the tag can comprise a double-stranded fixed reference sequence upstream or downstream of the identifier or molecular barcode. Each strand of a double-stranded fixed reference sequence can be, for example, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 nucleotides in length.

C. Adaptors

A library of polynucleotide molecules can be synthesized for use in sequencing. For example, a library of polynucleotides comprising a plurality of polynucleotide molecules that are each less than or equal to 100, 90, 80, 70, 60, 50, 45, 40, or 35 nucleic acid (or nucleotide) bases in length can be made. A plurality of polynucleotide molecules can be each less than or equal to 35 nucleic acid bases in length. A plurality of polynucleotide molecules can be each less than or equal to 30 nucleic acid bases in length. A plurality of polynucleotide molecules can also be less than or equal to 250, 200, 150, 100, or 50 nucleic acid bases. Additionally, the plurality of polynucleotide molecules can also be less than or equal to 100, 99, 98, 97, 96, 95, 94, 93, 92, 91, 90, 89, 88, 87, 86, 85, 84, 83, 82, 81, 80, 79, 78, 77, 76, 75, 74, 73, 72, 71, 70, 69, 68, 67, 66, 65, 64, 63, 62, 61, 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 nucleic acid bases.

A library of polynucleotides comprising a plurality of polynucleotide molecules can also have distinct (with respect to each other) molecular barcode sequences (or molecular barcodes) with respect to at least 4 nucleic acid bases. A molecular barcode (also "barcode" or "identifier" herein) sequence is a nucleotide sequence that distinguishes one polynucleotide from another. In other embodiments, the polynucleotide molecules can also have different barcode sequences with respect to 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more nucleic acid bases.

A library of polynucleotides comprising a plurality of polynucleotide molecules can also have a plurality of different barcode sequences. For example, a plurality of polynucleotide molecules can have at least 4 different molecular barcode sequences. In some cases, the plurality of polynucleotide molecules has from 2-100, 4-50, 4-30, 4-20, or 4-10 different molecular barcode sequences. The plurality of polynucleotides molecules can also have other ranges of different barcode sequences such as, 1-4, 2-5, 3-6, 4-7, 5-8, 6-9, 7-10, 8-11, 9-12, 10-13, 11-14, 12-15, 13-16, 14-17, 15-18, 16-19, 17-20, 18-21, 19-22, 20-23, 21-24, or 22-25 different barcode sequences. In other cases, a plurality of polynucleotide molecules can have at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or 100 more different barcode sequences. In a particular example, the plurality library adapters comprise at least 8 different sequences.

The location of the different barcode sequences can vary within the plurality of polynucleotides. For example, the different barcode sequences can be within 20, 15, 10, 9, 8, 7, 6, 5, 4, 3, or 2 nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules. In an example, a plurality of polynucleotide molecules has distinct barcode sequences that are within 10 nucleic acid bases from the terminal end. In another example, a plurality of polynucleotide molecules has distinct barcode sequences that are within 5 or 1 nucleic acid bases from the terminal end. In other instances, the distinct barcode sequences can be at the terminal end of a respective one of the plurality of polynucleotide molecules. Other variations include that the distinct molecular barcode sequences can be within 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, or more nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules.

The terminal end of the plurality of polynucleotide molecules can be adapted for ligation to a target nucleic acid molecule. For example, the terminal end can be a blunt end. In some other cases, the terminal end is adapted for hybridization to a complementary sequence of a target nucleic acid molecule.

A library of polynucleotides comprising a plurality of polynucleotide molecules can also have an edit distance of at least 1. In some cases, the edit distance is with respect to individual bases of the plurality of polynucleotide molecules. In other cases, the plurality of polynucleotide molecules can have an edit distance of at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more. The edit distance can be a Hamming distance.

In some cases, the plurality of polynucleotides does not contain sequencing adaptors. A sequence adaptor can be a polynucleotide that comprises a sequence that hybridizes to one or more sequencing adaptors or primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, e.g., a flow cell sequence. The term "flow cell sequence" and its grammatical equivalents as used herein, refers to a sequence that permits hybridization to a substrate, for example, by way of a primer attached to the substrate. The substrate can be bead or a planar surface. In some embodiments, a flow cell sequence can allow a polynucleotide to attach to a flow cell or surface (e.g., surface of a bead, for example, an Illumina flow cell.

When a plurality of polynucleotide molecules does not contain sequencing adaptors or primers, each polynucleotide molecule of the plurality does not contain a nucleic acid sequence or other moiety that is adapted to permit sequencing of a target nucleic acid molecule with a given sequencing approach, such as Illumina, SOLiD, Pacific Biosciences, GeneReader, Oxford Nanopore, Complete Genomics, Gnu-

Bio, Ion Torrent, Oxford Nanopore or Genia. In some examples, when a plurality of polynucleotide molecules does not contain sequencing adaptors or primers, the plurality of polynucleotide molecules does not contain flow cell sequences. For example, the plurality of polynucleotide molecules cannot bind to flow cells, such as used in Illumina flow cell sequencers. However, these flow cell sequences, if desired, can be added to the plurality of polynucleotide molecules by methods such as PCR amplification or ligation. At this point, Illumina flow cell sequencers can be used. Alternatively, when the plurality of polynucleotide molecules does not contain sequencing adaptors or primers, the plurality of polynucleotide molecules does not contain hairpin shaped adaptors or adaptors for generating hairpin loops in a target nucleic acid molecule, such as Pacific Bioscience SMRTbell™ adaptors. However, these hairpin shaped adaptors, if desired, can be added to the plurality of polynucleotide molecules by methods such as PCR amplification or ligation. The plurality of polynucleotide molecules can be circular or linear.

A plurality of polynucleotide molecules can be double stranded. In some cases, the plurality of polynucleotide molecules can be single stranded, or can comprise hybridized and non-hybridized regions. A plurality of polynucleotide molecules can be non-naturally occurring polynucleotide molecules.

Adaptors can be polynucleotide molecules. The polynucleotide molecules can be Y-shaped, bubble-shaped or hairpin-shaped. A hairpin adaptor may contain a restriction site(s) or a Uracil containing base. Adaptors can comprise a complementary portion and a non-complementary portion. The non-complementary portion can have an edit distance (e.g., Hamming distance). For example, the edit distance can be at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, or at least 30. The complementary portion of the adaptor can comprise sequences that are selected to enable and/or promote ligation to a polynucleotide, e.g., a sequence to enable and/or promote ligation to a polynucleotide at a high yield.

A plurality of polynucleotide molecules as disclosed herein can be purified. In some cases, a plurality of polynucleotide molecules as disclosed herein can be isolated polynucleotide molecules. In other cases, a plurality of polynucleotide molecules as disclosed herein can be purified and isolated polynucleotide molecules.

In certain aspects, each of the plurality of polynucleotide molecules is Y-shaped or hairpin-shaped. Each of the plurality of polynucleotide molecules can comprise a different barcode. The different barcode can be a randomer in the complementary portion (e.g., double stranded portion) of the Y-shaped or hairpin-shaped adaptor. Alternatively, the different barcode can be in one strand of the non-complementary portion (e.g., one of the Y-shaped arms). As discussed above, the different barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, e.g., 7 bases. The barcode can be contiguous or non-contiguous sequences, as described above. The plurality of polynucleotide molecules is from 10 nucleic acid bases to 35 nucleic acid bases (or any length as described above) in length. Further, the plurality of polynucleotide molecules can comprise an edit distance (as described above), that is a Hamming distance. A plurality of polynucleotide molecules can have distinct barcode sequences that are within 10 nucleic acid bases from the terminal end.

In another aspect, a plurality of polynucleotide molecules can be sequencing adaptors. A sequencing adaptor can comprise a sequence hybridizing to one or more sequencing primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, e.g., a flow cell sequence. For example, a sequencing adaptor can be a flow cell adaptor. The sequencing adaptors can be attached to one or both ends of a polynucleotide fragment. In another example, a sequencing adaptor can be hairpin shaped. For example, the hairpin shaped adaptor can comprise a complementary double-stranded portion and a loop portion, where the double-stranded portion can be attached (e.g., ligated) to a double-stranded polynucleotide. Hairpin shaped sequencing adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. A sequencing adaptor can be up to 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, or more bases from end to end. For example, a sequencing adaptor can be up to 70 bases from end to end. The sequencing adaptor can comprise 20-30, 20-40, 30-50, 30-60, 40-60, 40-70, 50-60, 50-70, bases from end to end. In a particular example, the sequencing adaptor can comprise 20-30 bases from end to end. In another example, the sequencing adaptor can comprise 50-60 bases from end to end. A sequencing adaptor can comprise one or more barcodes. For example, a sequencing adaptor can comprise a sample barcode. The sample barcode can comprise a pre-determined sequence. The sample barcodes can be used to identify the source of the polynucleotides. The sample barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, e.g., at least 8 bases. The barcode can be contiguous or non-contiguous sequences, as described above.

The plurality of polynucleotide molecules as described herein can be used as adaptors. Adaptors can comprise one or more identifiers. An adaptor can comprise an identifier with a random sequence. Alternatively, an adaptor can comprise an identifier with pre-determined sequences. Some adaptors can comprise an identifier with a random sequence and another identifier with a pre-determined sequence. The adaptors comprising identifiers can be double-stranded or single-stranded adaptors. The adaptors comprising identifiers can be Y-shaped adaptors. A Y-shaped adaptor can comprise one or more identifiers with a random sequence. The one or more identifiers can be on the hybrid portion and/or non-hybridized portion of the Y-shaped adaptor. A Y-shaped adaptor can comprise one or more identifiers with a pre-determined sequence. The one or more identifiers with pre-determined sequence can be on the hybridized portion and/or non-hybridized portion of the Y-shaped adaptor. A Y-shaped adaptor can comprise one or more identifiers with a random sequence and one or more identifiers with a pre-determined sequence. For example, the one or more identifiers with a random sequence can be on the hybridized portion of the Y-shaped adaptor and/or the non-hybridized portion of the Y-shaped adaptor. The one or more identifiers with a pre-determined sequence can be on the hybridized portion of the Y-shaped adaptor and/or the non-hybridized

portion of the Y-shaped adaptor. In a particular example, a Y-shaped adaptor can comprise an identifier with a random sequence on its hybridized portion and an identifier with a pre-determined sequence on its non-hybridized portion. The identifiers can be in any length disclosed herein. For example, a Y-shaped adaptor can comprise an identifier with a random sequence of 7 nucleotides on its hybridized portion and an identifier with a pre-determined sequence of 8 nucleotides on its non-hybridized portion.

An adaptor can include a double-stranded portion with a molecular barcode and at least one or two single-stranded portion. For example, the adaptor can be Y-shaped and include a double-stranded portion and two single-stranded portions. The single-stranded portions can include sequences that are not complementary to one another.

The adaptor can include a terminal end that has a sequence that is selected to permit the adaptor to be efficiently (e.g., at an efficiency of at least about 20%, 30%, 40%, 50%) ligated or otherwise coupled to a polynucleotide. In some examples, terminal nucleotides in a double-stranded portion of an adaptor are selected from a combination of purines and pyrimidines to provide for efficient ligation.

In some examples, a set of library adaptors comprises a plurality of polynucleotide molecules (library adaptors) with molecular barcodes. The library adaptors are less than or equal to 80, 70, 60, 50, 45, or 40 nucleotide bases in length. The molecular barcodes can be at least 4 nucleotide bases in length, but may be from 4 to 20 nucleotide bases in length. The molecular barcodes can be different from one another and have an edit distance of at least 1, 2, 3, 4, or 5 between one another. The molecular barcodes are located at least 1, 2, 3, 4, 5, 10, or 20 nucleotide bases away from a terminal end of their respective library adaptors. In some cases, the at least one terminal base is identical in all of the library adaptors.

The library adaptors can be identical but for the molecular barcodes. For example, the library adaptors can have identical sequences but differ only with respect to nucleotide sequences of the molecular barcodes.

Each of the library adaptors can have a double stranded portion and at least one single-stranded portion. By "single stranded portion" is meant an area of non-complementarity or an overhang. In some cases, each of the library adaptors has a double-stranded portion and two single-stranded portions. The double-stranded portion can have a molecular barcode. In some cases, the molecular barcode is a randomer. Each of the library adaptors can further include a strand-identification barcode on a single-stranded portion. The strand-identification barcode can include at least 4 nucleotide bases, in some cases from 4 to 20 nucleotide bases.

In some examples, each of the library adaptors has a double-stranded portion with a molecular barcode and two single-stranded portions. The single-stranded portions may not hybridize to one another. The single-stranded portions may not be completely complementary to one another.

The library adaptors can have a sequence of terminal nucleotides in a double-stranded portion that are the same. The sequence of terminal nucleotides can be at least 2, 3, 4, 5 or 6 nucleotide bases in length. For example, one strand of a double-stranded portion of the library adaptor can have the sequence ACTT, TCGC, or TACC at the terminal end, while the other strand can have a complementary sequence. In some cases, such a sequence is selected to optimize the efficiency at which the library adaptors ligate to target polynucleotides. Such sequences can be selected to optimize a binding interaction between the ends of the library adaptors and the target polynucleotides.

In some cases, none of the library adaptors contains a sample identification motif (or sample molecular barcode). Such sample identification motif can be provided via sequencing adaptors. A sample identification motif can include a sequencer of at least 4, 5, 6, 7, 8, 9, 10, 20, 30, or 40 nucleotide bases that permits the identification of polynucleotide molecules from a given sample from polynucleotide molecules from other samples. For example, this can permit polynucleotide molecules from two subjects to be sequenced in the same pool and sequence reads for the subjects subsequently identified.

A sequencer motif includes nucleotide sequence(s) needed to couple a library adaptor to a sequencing system and sequence a target polynucleotide coupled to the library adaptor. The sequencer motif can include a sequence that is complementary to a flow cell sequence and a sequence (sequencing initiation sequence) that is selectively hybridizable to a primer (or priming sequence) for use in sequencing. For example, such sequencing initiation sequence can be complementary to a primer that is employed for use in sequence by synthesis (e.g., Illumina). Such primer can be included in a sequencing adaptor. A sequencing initiation sequence can be a primer hybridization site.

In some cases, none of the library adaptors contains a complete sequencer motif. The library adaptors can contain partial or no sequencer motifs. In some cases, the library adaptors include a sequencing initiation sequence. The library adaptors can include a sequencing initiation sequence but no flow cell sequence. The sequence initiation sequence can be complementary to a primer for sequencing. The primer can be a sequence specific primer or a universal primer. Such sequencing initiation sequences may be situated on single-stranded portions of the library adaptors. As an alternative, such sequencing initiation sequences may be priming sites (e.g., kinks or nicks) to permit a polymerase to couple to the library adaptors during sequencing.

In some cases, partial or complete sequencer motifs are provided by sequencing adaptors. A sequencing adaptor can include a sample molecular barcode and a sequencer motif. The sequencing adaptors can be provided in a set that is separate from the library adaptors. The sequencing adaptors in a given set can be identical—i.e., they contain the same sample barcode and sequencer motif.

Sequencing adaptors can include sample identification motifs and sequencer motifs. Sequencer motifs can include primers that are complementary to a sequencing initiation sequence. In some cases, sequencer motifs also include flow cell sequences or other sequences that permit a polynucleotide to a configured or arranged in a manner that permits the polynucleotide to be sequenced by a sequencer.

Library adaptors and sequencing adaptors can each be partial adaptors, that is, containing part but not all of the sequences necessary to enable sequencing by a sequencing platform. Together they provide complete adaptors. For example, library adaptors can include partial or no sequencer motifs, but such sequencer motifs are provided by sequencing adaptors.

FIGS. **9**A-**9**C schematically illustrate a method for tagging a target polynucleotide molecule with library adaptors. FIG. **9**A shows a library adaptor as a partial adaptor containing a primer hybridization site on one of the strands and a molecular barcode towards another end. The primer hybridization site can be a sequencing initiation sequence for subsequent sequencing. The library adaptor is less than or equal to 80 nucleotide bases in length. In FIG. **9**B, the

library adaptors are ligated at both ends of the target polynucleotide molecule to provide a tagged target polynucleotide molecule. The tagged target polynucleotide molecule may be subjected to nucleic acid amplification to generate copies of the target. Next, in FIG. **9**C, sequencing adaptors containing sequencer motifs are provided and hybridized to the tagged target polynucleotide molecule. The sequencing adaptors contain sample identification motifs. The sequencing adaptors can contain sequences to permit sequencing of the tagged target with a given sequencer.

D. Sequencing

Tagged polynucleotides can be sequenced to generate sequence reads (e.g., as shown in step (**106**), FIG. **1**). For example, a tagged duplex polynucleotide can be sequenced. Sequence reads can be generated from only one strand of a tagged duplex polynucleotide. Alternatively, both strands of a tagged duplex polynucleotide can generate sequence reads. The two strands of the tagged duplex polynucleotide can comprise the same tags. Alternatively, the two strands of the tagged duplex polynucleotide can comprise different tags. When the two strands of the tagged duplex polynucleotide are differently tagged, sequence reads generated from one strand (e.g., a Watson strand) can be distinguished from sequence reads generated from the other strands (e.g., a Crick strand). Sequencing can involve generating multiple sequence reads for each molecule. This occurs, for example, as a result the amplification of individual polynucleotide strands during the sequencing process, e.g., by PCR.

Methods disclosed herein can comprise amplifying of polynucleotides. Polynucleotides amplification can result in the incorporation of nucleotides into a nucleic acid molecule or primer thereby forming a new nucleic acid molecule complementary to a template nucleic acid. The newly formed polynucleotide molecule and its template can be used as templates to synthesize additional polynucleotides. The polynucleotides being amplified can be any nucleic acids, for example, deoxyribonucleic acids, including genomic DNAs, cDNAs (complementary DNA), cfDNAs, and circulating tumor DNAs (ctDNAs). The polynucleotides being amplified can also be RNAs. As used herein, one amplification reaction may comprise many rounds of DNA replication. DNA amplification reactions can include, for example, polymerase chain reaction (PCR). One PCR reaction may comprise 2-100 "cycles" of denaturation, annealing, and synthesis of a DNA molecule. For example, 2-7, 5-10, 6-11, 7-12, 8-13, 9-14, 10-15, 11-16, 12-17, 13-18, 14-19, or 15-20 cycles can be performed during the amplification step. The condition of the PCR can be optimized based on the GC content of the sequences, including the primers.

Nucleic acid amplification techniques can be used with the assays described herein. Some amplification techniques are the PCR methodologies which can include, but are not limited to, solution PCR and in situ PCR. For example, amplification may comprise PCR-based amplification. Alternatively, amplification may comprise non PCR-based amplification. Amplification of the template nucleic acid may comprise use of one or more polymerases. For example, the polymerase may be a DNA polymerase or an RNA polymerase. In some cases, high fidelity amplification is performed such as with the use of high fidelity polymerase (e.g., Phusion® High-Fidelity DNA Polymerase) or PCR protocols. In some cases, the polymerase may be a high fidelity polymerase. For example, the polymerase may be KAPA HiFi DNA polymerase. The polymerase may also be Phusion DNA polymerase. The polymerase may be used under reaction conditions that reduce or minimize amplification biases, e.g., due to fragment length, GC content, etc.

Amplification of a single strand of a polynucleotide by PCR will generate copies both of that strand and its complement. During sequencing, both the strand and its complement will generate sequence reads. However, sequence reads generated from the complement of, for example, the Watson strand, can be identified as such because they bear the complement of the portion of the duplex tag that tagged the original Watson strand. In contrast, a sequence read generated from a Crick strand or its amplification product will bear the portion of the duplex tag that tagged the original Crick strand. In this way, a sequence read generated from an amplified product of a complement of the Watson strand can be distinguished from a complement sequence read generated from an amplification product of the Crick strand of the original molecule.

All amplified polynucleotides can be submitted to a sequencing device for sequencing. Alternatively, a sampling, or subset, of all of the amplified polynucleotides is submitted to a sequencing device for sequencing. With respect to any original double-stranded polynucleotide there can be three results with respect to sequencing. First, sequence reads can be generated from both complementary strands of the original molecule (that is, from both the Watson strand and from the Crick strand). Second, sequence reads can be generated from only one of the two complementary strands (that is, either from the Watson strand or from the Crick strand, but not both). Third, no sequence read may be generated from either of the two complementary strands. Consequently, counting unique sequence reads mapping to a genetic locus will underestimate the number of double-stranded polynucleotides in the original sample mapping to the locus. Described herein are methods of estimating the unseen and uncounted polynucleotides.

The sequencing method can be massively parallel sequencing, that is, simultaneously (or in rapid succession) sequencing any of at least 100, 1000, 10,000, 100,000, 1 million, 10 million, 100 million, or 1 billion polynucleotide molecules. Sequencing methods may include, but are not limited to: high-throughput sequencing, pyrosequencing, sequencing-by-synthesis, single-molecule sequencing, nanopore sequencing, semiconductor sequencing, sequencing-by-ligation, sequencing-by-hybridization, RNA-Seq (Illumina), Digital Gene Expression (Helicos), Next generation sequencing, Single Molecule Sequencing by Synthesis (SMSS)(Helicos), massively-parallel sequencing, Clonal Single Molecule Array (Solexa), shotgun sequencing, Maxam-Gilbert or Sanger sequencing, primer walking, sequencing using PacBio, SOLiD, Ion Torrent, or Nanopore platforms and any other sequencing methods known in the art.

For example, duplex-tagged polynucleotides can be amplified, by for example PCR (see e.g., FIG. **4**A duplex-tagged polynucleotides are referred to as mm' and nn'). In FIG. **4**A, the strand of the duplex polynucleotide including sequence m bears sequence tags w and y, while the strand of the duplex polynucleotide including sequence m' bears sequence tags x and z. Similarly, the strand of the duplex polynucleotide including sequence n bears sequence tags a and c, while the strand of the duplex polynucleotide including sequence n' bears sequence tags b and d. During amplification, each strand produces itself and its complementary sequence. However, for example, an amplification progeny of original strand m that includes the complementary sequence, m', is distinguishable from an amplification progeny of original strand m' because the progeny from original

strand m will have the sequence 5'-y'm'w'-3' and the progeny of the original m' strand one strand will have the sequence 5'-zm'x-3'. FIG. **4**B shows amplification in more detail. During amplification, errors can be introduced into the amplification progeny, represented by dots. The application progeny are sampled for sequencing, so that not all strands produce sequence reads, resulting in the sequence reads indicated. Because sequence reads can come from either of a strand or its complement, both sequences and complement sequences will be included in the set of sequence reads. It should be noted that it is possible that a polynucleotide would bear the same tag on each end. Thus, for a tag "a", and polynucleotide "m", a first strand could be tagged a-m-a', and the complement could be tagged a-m'-a.

E. Determining Consensus Sequence Reads

Methods disclosed herein can comprise determining consensus sequence reads in sequence reads (e.g., as shown in step (**108**), FIG. **1**), such as by reducing or tracking redundancy. Sequencing of amplified polynucleotides can produce reads of the several amplification products from the same original polynucleotide, referred to as "redundant reads". By identifying redundant reads, unique molecules in the original sample can be determined. If the molecules in a sample are uniquely tagged, then reads generated from amplification of a single unique original molecule can be identified based on their distinct barcode. Ignoring barcodes, reads from unique original molecules can be determined based on sequences at the beginning and end of a read, optionally in combination with the length of the read. In certain cases, however, a sample may be expected to have a plurality of original molecules having the same start stop sequences and the same length. Without barcoding, these molecules are difficult to distinguish from one another. However, if a collection of polynucleotides is non-uniquely tagged (that is, an original molecule shares the same identifier with at least one other original molecule), combining information from a barcode with start/stop sequence and/or polynucleotide length significantly increases the probability that any sequence read can be traced back to an original polynucleotide. This is because, in part, even without unique tagging, it is unlikely that any two original polynucleotides having the same start/stop sequence and length also will be tagged with the same identifier.

F. Collapsing

Collapsing allows for reduction in noise (i.e., background) that is generated at each step of the process. Methods disclosed herein can comprise collapsing, e.g., generating a consensus sequence by comparing multiple sequence reads. For example, sequence reads generated from a single original polynucleotide can be used to generate a consensus sequence of that original polynucleotide. Iterative rounds of amplification can introduce errors into progeny polynucleotides. Also, sequencing typically may not be performed with perfect fidelity so sequencing errors are introduced at this stage as well. However, comparison of sequence reads of molecules derived from a single original molecule, including those that have sequence variants, can be analyzed so as to determine the original, or "consensus" sequence. This can be done phylogenetically. Consensus sequences can be generated from families of sequence reads by any of a variety of methods. Such methods include, for example, linear or non-linear methods of building consensus sequences (such as voting (e.g., biased voting), averaging, statistical, maximum a posteriori or maximum likelihood detection, dynamic programming, Bayesian, hidden Markov or support vector machine methods, etc.) derived from digital communication theory, information theory, or bioinformatics. For example, if all or most of the sequence reads tracking back to an original molecule bear the same sequence variant, that variant probably existed in the original molecule. On the other hand, if a sequence variant exists in a subset of redundant sequence reads, that variant may have been introduced during amplification/sequencing and represents an artifact not existing in the original. Furthermore, if only sequence reads derived from the Watson or Crick strand of an original polynucleotide contain the variant, the variant may have been introduced through single-sided DNA damage, first-cycle PCR error or through contaminating polynucleotides that were amplified from a different sample.

After fragments are amplified and the sequences of amplified fragments are read and aligned, the fragments are subjected to base calling, e.g., determining for each locus the most likely nucleotide. However, variations in the number of amplified fragments and unseen amplified fragments (e.g., those without being read their sequences; reasons could be too many such as amplification errors, sequencing reading errors, too long, too short, being chopped, etc.) may introduce errors in base calling. If there are too many unseen amplified fragments with respect to the seen amplified fragments (amplified fragments actually being read), the reliability of base calling may be diminished.

Therefore, disclosed herein is a method to correct for the number of unseen fragments in base calling. For example, when base calling for locus A (an arbitrary locus), it is first assumed that there are N amplified fragments. The sequence readouts can come from two types of fragments: double-strand fragments and single-strand fragments. Therefore, we assign N1, N2, and N3 as the numbers of double-strands, single-strands, and unseen fragments, respectively. Thus, N=N1+N2+N3 (N1 and N2 are known from the sequence readouts, and N and N3 are unknown). If the formula is solved for N (or N3), then N3 (or N) will be inferred.

Probability is used to estimate N. For example, we assign "p" to be the probability of having detected (or having read) a nucleotide of locus A in a sequence readout of a single-strand.

For sequence readouts from double-strands, the nucleotide call from a double-strand amplified fragment has a probability of p*p=p^2, seeing all N1 double-strands has the following equation: N1=N*(p^2).

For sequence readouts from a single-strand. Assuming that one of the 2 strands is seen, and the other is unseen, the probability of seeing one strand is "p", but the probability of missing the other strand is (1−p). Furthermore, by not distinguishing the single strand sourcing from 5-primer and sourcing from 3-primer, there is a factor of 2. Therefore, the nucleotide call from a single-strand amplified fragment has a probability 2×p×(1−p). Thus, seeing all N2 single-strands has the following equation: N2=N×2×p×(1−p).

"p" is also unknown. To solve p, the ratio of N1 to N2 is used to solve for "p":

$$R = \frac{N1}{N2} = \frac{Np^2}{2Np(1-p)} = \frac{p^2}{2p(1-p)} = \frac{p}{2(1-p)}$$

Once "p" is found, N can be found. After N is found, can be found N3=N−N1−N2.

Besides the ratio of paired versus unpaired strands (which is a measure post-collapsing), there is useful information in the pre-collapsing read depth at each locus. This information

can be used to further improve the call for total molecule count and/or increase confidence of calling variants.

For example, FIG. **4**C demonstrates sequence reads corrected for complementary sequences. Sequences generated from an original Watson strand or an original Crick strand can be differentiated on the basis of their duplex tags. Sequences generated from the same original strand can be grouped. Examination of the sequences can allow one to infer the sequence of the original strand (the "consensus sequence"). In this case, for example, the sequence variant in the nn' molecule is included in the consensus sequence because it included in every sequence read while other variants are seen to be stray errors. After collapsing sequences, original polynucleotide pairs can be identified based on their complementary sequences and duplex tags.

FIG. **5** demonstrates increased confidence in detecting sequence variants by pairing reads from Watson and Crick strands. Sequence nn' can include a sequence variant indicated by a dot. In some cases, sequence pp' does not include a sequence variant. Amplification, sequencing, redundancy reduction and pairing can result in both Watson and Crick strands of the same original molecule including the sequence variant. In contrast, as a result of errors introduced during amplification and sampling during sequencing, the consensus sequence of the Watson strand p can contain a sequence variant, while the consensus sequence of the Crick strand p' does not. It is less likely that amplification and sequencing will introduce the same variant into both strands (nn' sequence) of a duplex than onto one strand (pp' sequence). Therefore, the variant in the pp' sequence is more likely to be an artifact, and the variant in the nn' sequence is more likely to exist in the original molecule.

Methods disclosed herein can be used to correct errors resulted from experiments, e.g., PCR, amplification, and/or sequencing. For example, such a method can comprises attaching one or more double stranded adaptors to both ends of a double stranded polynucleotide, thereby providing a tagged double stranded polynucleotide; amplifying the double stranded tagged polynucleotide; sequencing both strands of the tagged polynucleotide; comparing the sequence of one strand with its complement to determine any errors introduced during sequencing; and correcting errors in the sequence based on (d). The adaptors used in this method can be any adaptors disclosed herein, e.g., Y-shaped adaptors. The adaptor can comprise any barcodes (e.g., distinct barcodes) disclosed herein.

G. Mapping

Sequence reads or consensus sequences can be mapped to one or more selected genetic loci (e.g., as shown step (**110**), FIG. **1**). A genetic locus can be, for example, a specific nucleotide position in the genome, a sequence of nucleotides (for example, an open reading frame), a fragment of a chromosome, a whole chromosome, or an entire genome. A genetic locus can be a polymorphic locus. Polymorphic locus can be a locus at which sequence variation exists in the population and/or exists in a subject and/or a sample. A polymorphic locus can be generated by two or more distinct sequences coexisting at the same location of the genome. The distinct sequences can differ from one another by one or more nucleotide substitutions, a deletion/insertion, and/or a duplication of any number of nucleotides, generally a relatively small number of nucleotides, such as less than 50, 45, 40, 35, 30, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1 nucleotide(s), among others. A polymorphic locus can be created by a single nucleotide position that varies within the population, e.g. a single nucleotide variation (SNV) or a single nucleotide polymorphism (SNP).

A reference genome for mapping can include the genome of any species of interest. Human genome sequences useful as references can include the hg19 assembly or any previous or available hg assembly. Such sequences can be interrogated using the genome browser available at genome.ucsc.edu/index.html. Other species genomes include, for example PanTro2 (chimp) and mm9 (mouse).

In methods disclosed herein, collapsing can be performed before or after mapping. In some aspects, collapsing can be performed before mapping. For example, sequence reads can be grouped into families based on their tags and one or more endogenous sequences, without regard to where the reads map in the genome. Then, the members of a family can be collapsed into a consensus sequence. The consensus sequence can be generated using any collapsing method disclosed herein. Then the consensus sequence can be mapped to locations in the genome. Reads mapped to a locus can be quantified (e.g., counted). Percentage of reads carrying a mutation at a locus can also be determined. Alternatively, collapsing can be performed after mapping. For example, all reads can first be mapped to the genome. Then the reads can be grouped into families based on their tags and one or more endogenous sequences. Since the reads have been mapped to the genome, consensus bases can be determined for each family at each locus. In other aspects, consensus sequence can be generated for one strand of a DNA molecule (e.g., for a Watson strand or a Crick strand). Mapping can be performed before or after the consensus sequence for one strand of the DNA molecule is determined. Numbers of Doublets and Singlets can be determined. These numbers can be used to calculate unseen molecules. For example, the unseen molecules can be calculated using the following equation: $N=D+S+U$; $D=Np(2)$, $S=N2pq$, where $p=1-q$, where p is the probability of seeing; q is the probability of missing a strand.

H. Grouping

Methods disclosed herein can also comprise grouping sequence reads. Sequence reads can be grouped based on various types of sequences, e.g., sequences of an oligonucleotide tag (e.g., a barcode), sequence of a polynucleotide fragments, or any combinations. For example, as shown in step (**112**) (FIG. **1**), sequence reads can be grouped as follows: Sequence reads generated from a "Watson" strand and those generated from a "Crick" strand of a double-stranded polynucleotide in the sample are identifiable based on the duplex tags that they bear. In this way, a sequence read or consensus sequence from a Watson strand of a duplex polynucleotide can be paired with a sequence read or consensus sequence from its complementary Crick strand. Paired sequence reads are referred to as a "Pair".

Sequence reads for which no sequence read corresponding to a complementary strand can be found among the sequence reads are termed "Singlets".

Double-stranded polynucleotides for which a sequence read for neither of the two complementary strands has been generated are referred to as "Unseen" molecules.

I. Quantifying

Methods disclosed herein also comprise quantifying sequence reads. For example, as shown in step (**114**) (FIG. **1**), Pairs and Singlets mapping to a selected genetic locus, or to each of a plurality of selected genetic loci, are quantified, e.g., counted.

The quantifying can comprise estimating number of polynucleotides in the sample (e.g., Pairs polynucleotides, Sin-

glets polynucleotides, or Unseen polynucleotides. For example, as shown in step (**116**) (FIG. **1**), the number of double-stranded polynucleotides in the sample for which no sequence reads were generated ("Unseen" polynucleotides) is estimated. The probability that a double strand polynucleotide generates no sequence reads can be determined based on the relative number of Pairs and Singlets at any locus. Using this probability, the number of Unseen polynucleotide can be estimated.

In step (**118**) an estimate for the total number of double-stranded polynucleotides in a sample mapping to a selected locus is the sum of the number of Pairs, the number of Singlets and the number of Unseen molecules mapping to the locus.

The number of Unseen original molecules in a sample can be estimated based on the relative number of Pairs and Singlets (FIG. **2**). Referring to FIG. **2**, as an example, counts for a particular genomic locus, Locus A, are recorded, where 1000 molecules are paired and 1000 molecules are unpaired. Assuming a uniform probability, p, for an individual Watson or Crick strand to make it through the process subsequent to conversion, one can calculate the proportion of molecules that fail to make it through the process (Unseen) as follows: Let R=ratio of paired to unpaired molecules=1, so $R=1=p^2/(2p(1-p))$. This implies that p=2/3 and that the quantity of lost molecules is equal to $(1-p)^2=1/9$. Thus in this example, approximately 11% of converted molecules are lost and never detected. Consider another genomic locus, Locus B, in the same sample where 1440 molecules are paired and 720 are unpaired. Using the same method, we can infer the number of molecules that are lost, is only 4%. Comparing the two areas, it may be assumed that Locus A had 2000 unique molecules as compared to 2160 molecules in Locus B—a difference of almost 8%. However, by correctly adding in the lost molecules in each region, we infer there are 2000/(8/9)=2250 molecules in Locus A and 2160/0.96=2250 molecules in Locus B. Hence, the counts in both regions are actually equal. This correction and thus much higher sensitivity can be achievable by converting the original double-stranded nucleic acid molecules and bioinformatically keeping track of all those that are paired and unpaired at the end of the process. Similarly, the same procedure can be used to infer true copy number variations in regions that appear to have similar counts of observed unique molecules. By taking the number of unseen molecules into consideration in the two or more regions, the copy number variation becomes apparent.

In addition to using binomial distribution, other methods of estimating numbers of unseen molecules include exponential, beta, gamma or empirical distributions based on the redundancy of sequence reads observed. In the latter case, the distribution of read counts for paired and unpaired molecules can be derived from such redundancy to infer the underlying distribution of original polynucleotide molecules at a particular locus. This can often lead to a better estimation of the number of unseen molecules.

J. CNV Detection

Methods disclosed herein also comprise detecting CNV. For example, as shown in step (**120**) (FIG. **1**), once the total number of polynucleotides mapping to a locus is determined, this number can be used in standard methods of determining CNV at the locus. A quantitative measure can be normalized against a standard. The standard can be an amount of any polynucleotides. In one method, a quantitative measure at a test locus can be standardized against a quantitative measure of polynucleotides mapping to a control locus in the genome, such as gene of known copy number. Quantitative measures can be compared against the amount of nucleic acid in any sample disclosed herein. For example, in another method, the quantitative measure can be compared against the amount of nucleic acid in the original sample. For example, if the original sample contained 10,000 haploid gene equivalents, the quantitative measure can be compared against an expected measure for diploidy. In another method, the quantitative measure can be normalized against a measure from a control sample, and normalized measures at different loci can be compared.

In some cases, in which copy number variation analysis is desired, sequence data may be: 1) aligned with a reference genome; 2) filtered and mapped; 3) partitioned into windows or bins of sequence; 4) coverage reads counted for each window; 5) coverage reads can then be normalized using a stochastic or statistical modeling algorithm; 6) and an output file can be generated reflecting discrete copy number states at various positions in the genome. In other cases, in which rare mutation analysis is desired, sequence data may be 1) aligned with a reference genome; 2) filtered and mapped; 3) frequency of variant bases calculated based on coverage reads for that specific base; 4) variant base frequency normalized using a stochastic, statistical or probabilistic modeling algorithm; 5) and an output file can be generated reflecting mutation states at various positions in the genome.

After the sequence read coverage ratios have been determined, a stochastic modeling algorithm can be optionally applied to convert the normalized ratios for each window region into discrete copy number states. In some cases, this algorithm may comprise a Hidden Markov Model. In other cases, the stochastic model may comprise dynamic programming, support vector machine, Bayesian modeling, probabilistic modeling, trellis decoding, Viterbi decoding, expectation maximization, Kalman filtering methodologies, or neural networks.

Methods disclosed herein can comprise detecting SNVs, CNVs, insertions, deletions, and/or rearrangements at a specific region in a genome. The specific genomic region can comprise a sequence in a gene, such as ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, or NTRK1.

In some cases, the method uses a panel which comprises exons of one or more genes. The panel can comprise introns of one or more genes as well. The panel can also comprise exons and introns of one or more genes. The one or more genes can be those disclosed above. The panel can comprise about 80,000 bases which cover a panel of genes. The panel can comprise about 1000, 2000, 3000, 4000, 5000, 10000, 15000, 20000, 25000, 30000, 35000, 40000, 45000, 50000, 55000, 60000, 65000, 70000, 75000, 80000, 85000, 90000, 95000, 100000, 105000, 110000, 115000, 120000, 125000, or more bases.

In some aspects, copy number of a gene can be reflected in the frequency of a genetic form of the gene in a sample. For example, in a healthy individual, no copy number variation is reflected in a variant in a gene in one chromosome (e.g., heterozygosity) being detected in about 50% of detected molecules in a sample. Also, in a healthy indi-

vidual, duplication of a gene bearing a variant can be reflected in the variant being detected in about 66% of detected molecules in a sample. Accordingly, if the tumor burden in a DNA sample is 10%, the frequency of a somatic mutation in a gene in one chromosome of cancer cells, without CNV, can be about 5%. The converse can be true in the case of aneuploidy.

The methods disclosed herein can be used to determine whether a sequence variant is more likely present in the germ line level or resulted from a somatic cell mutation, e.g., in a cancer cell. For example, a sequence variant in a gene detected at levels arguably consistent with heterozygosity in the germ line is more likely the product of a somatic mutation if CNV is also detected in that gene. In some cases, to the extent we expect that a gene duplication in the germ line bears a variant consistent with genetic dose (e.g., 66% for trisomy at a locus), detection gene amplification with a sequence variant dose that deviates significantly from this expected amount indicates that the CNV is more likely present as a result of somatic cell mutation.

The methods disclosed herein can also be used to infer tumor heterogeneity in a situation in which sequence variants in two genes are detected at different frequencies. For example, tumor heterogeneity can be inferred when two genes are detected at different frequencies but their copy numbers are relatively equal. Alternatively, tumor homogeneity can be inferred when the difference in frequency between two sequence variants is consistent with difference in copy number for the two genes. Thus, for example, if an EGFR variant is detected at 11% and a KRAS variant is detected at 5%, and no CNV is detected at these genes, the difference in frequency likely reflects tumor heterogeneity (e.g., all tumor cells carry an EGFR mutant and half the tumor cells also carry a KRAS mutant). Alternatively, if the EGFR gene carrying the mutant is detected at 2-times normal copy number, one interpretation is a homogenous population of tumor cells, each cell carrying a mutant in the EGFR and KRAS genes, but in which the KRAS gene is duplicated.

In response to chemotherapy, a dominant tumor form can eventually give way through Darwinian selection to cancer cells carrying mutants that render the cancer unresponsive to the therapy regimen. Appearance of these resistance mutants can be delayed through methods of this invention. In one embodiment of this method, a subject is subjected to one or more pulsed therapy cycles, each pulsed therapy cycle comprising a first period during which a drug is administered at a first amount and a second cycle during which the drug is administered at a second, reduced amount. The first period can be characterized by a tumor burden detected above a first clinical level. The second period can be characterized by a tumor burden detected below a second clinical level. First and second clinical levels can be different in different pulsed therapy cycles. For example, the first clinical level can be lower in succeeding cycles. A plurality of cycles can include at least 2, 3, 4, 5, 6, 7, 8 or more cycles. For example, the BRAF mutant V600E may be detected in polynucleotides of a disease cell at an amount indicating a tumor burden of 5% in cfDNA. Chemotherapy can commence with dabrafenib. Subsequent testing can show that the amount of the BRAF mutant in the cfDNA falls below 0.5% or to undetectable levels. At this point, dabrafenib therapy can stop or be significantly curtailed. Further subsequent testing may find that DNA bearing the BRAF mutation has risen to 2.5% of polynucleotides in cfDNA. At this point, dabrafenib therapy can be re-started, e.g., at the same level as the initial treatment. Subsequent testing may find that DNA bearing the BRAF mutation has decreased to 0.5% of polynucleotides in cfDNA. Again, dabrafenib therapy can be stopped or reduced. The cycle can be repeated a number of times.

A therapeutic intervention can also be changed upon detection of the rise of a mutant form resistant to an original drug. For example, cancers with the EGFR mutation L858R respond to therapy with erlotinib. However, cancers with the EGFR mutation T790M are resistant to erlotinib. However, they are responsive to ruxolitinib. A method of this invention involves monitoring changes in tumor profile and changing a therapeutic intervention when a genetic variant associated with drug resistance rises to a predetermined clinical level.

Methods disclosed in this invention can comprise a method of detecting disease cell heterogeneity from a sample comprising polynucleotides from somatic cells and disease cells, the method comprising: a) quantifying polynucleotides in the sample bearing a sequence variant at each of a plurality of genetic loci; b) determining CNV at each of the plurality of genetic loci; different relative amounts of disease molecules at a locus, wherein the CNV indicates a genetic dose of a locus in the disease cell polynucleotides; c) determining a relative measure of quantity of polynucleotides bearing a sequence variant at a locus per genetic dose at the locus for each of a plurality of the loci; and d) comparing the relative measures at each of the plurality of loci, wherein different relative measures indicates tumor heterogeneity. In the methods disclosed herein, the genetic dose can be determined on a total molecule basis. For example, if there are 1× total molecules at a first locus, and 1.2× molecules mapped to a second locus, then the genetic dose is 1.2. Variants at this locus can be divided by 1.2. In some aspects, the method disclosed herein can be used to detect any disease cell heterogeneity, e.g., tumor cell heterogeneity. The methods can be used to detect disease cell heterogeneity from a sample comprising any types of polynucleotides, e.g., cfDNA, genomic DNA, cDNA, or ctDNA. In the methods, the quantifying can comprise, for example, determining the number or relative amount of the polynucleotides. Determining CNV can comprise mapping and normalizing different relative amounts of total molecules to a locus.

In another aspect, in response to chemotherapy, a dominant tumor form can eventually give way through Darwinian selection to cancer cells carrying mutants that render the cancer unresponsive to the therapy regimen. Appearance of these resistance mutants can be delayed through methods disclosed throughout. The methods disclosed herein can comprise a method comprising: a) subjecting a subject to one or more pulsed therapy cycles, each pulsed therapy cycle comprising (i) a first period during which a drug is administered at a first amount and (ii) a second period during which the drug is administered at a second, reduced amount; wherein (A) the first period is characterized by a tumor burden detected above a first clinical level; and (B) the second period is characterized by a tumor burden detected below a second clinical level.

K. Sequence Variant Detection

Systems and methods disclosed herein can be used to detect sequence variants, e.g., SNVs. For example, a sequence variant can be detected from consensus sequences from multiple sequence reads, for example, from at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29,

at least 30, at least 31, at least 32, at least 33, at least 34, at least 35, at least 36, at least 37, at least 38, at least 39, at least 40, at least 41, at least 42, at least 43, at least 44, at least 45, at least 46, at least 47, at least 48, at least 49, at least 50, at least 51, at least 52, at least 53, at least 54, at least 55, at least 56, at least 57, at least 58, at least 59, at least 60, at least 61, at least 62, at least 63, at least 64, at least 65, at least 66, at least 67, at least 68, at least 69, at least 70, at least 71, at least 72, at least 73, at least 74, at least 75, at least 76, at least 77, at least 78, at least 79, at least 80, at least 81, at least 82, at least 83, at least 84, at least 85, at least 86, at least 87, at least 88, at least 89, at least 90, at least 91, at least 92, at least 93, at least 94, at least 95, at least 96, at least 97, at least 98, at least 99, at least 100, at least 200, at least 300, at least 400, at least 500, at least 600, at least 700, at least 800, at least 900, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, at least 8000, at least 9000, at least 10000 or more sequence reads. A consensus sequence can be from sequence reads of a single strand polynucleotide. A consensus sequence can also be from sequence reads of one strand of a double-stranded polynucleotide (e.g., pairing reads). In an exemplary method, pairing reads allows one to identify with increased confidence the existence of a sequence variant in a molecule. For example, if both strands of a Pair include the same variant, one can be reasonably sure that the variant existed in the original molecule, as the chance that the same variant is introduced into both strands during amplification/sequencing is rare. In contrast, if only one strand of a Pair includes the sequence variant, this is more likely to be an artifact. Similarly, the confidence that a Singlet bearing a sequence variant existed in the original molecule is less than the confidence if the variant exists in a Duplex, as there is higher probability that the variant can be introduced once than twice during amplification/sequencing.

Other methods of copy number variation detection and the sequence variant detection are described in PCT/US2013/058061, which is entirely incorporated herein by reference.

Sequence reads can be collapsed to generate a consensus sequence, which can be mapped to a reference sequence to identify genetic variants, such as CNV or SNV. As an alternative, the sequence reads are mapped prior to or even without mapping. In such a case, the sequence reads can be individually mapped to the reference to identify a CNV or SNV.

FIG. **3** shows a reference sequence encoding a genetic Locus A. The polynucleotides in FIG. **3** may be Y-shaped or have other shapes, such as hairpin.

In some cases, an SNV or multiple-nucleotide variant (MNV) can be determined across multiple sequence reads at a given locus (e.g., nucleotide base) by aligning sequence reads that correspond to that locus. Next, a plurality of sequential nucleotide bases from at least a subset of the sequence reads are mapped to the reference to a SNV or MNV in a polynucleotide molecule or portion thereof that corresponds to the reads. The plurality of sequential nucleotide bases can span an actual, inferred or suspected location of the SNV or MNV. The plurality of sequential nucleotide bases can span at least 3, 4, 5, 6, 7, 8, 9, or 10 nucleotide bases.

L. Detecting/Quantifying Nucleic Acids

The methods described throughout can be used to tag nucleic acids fragments, such as deoxyribonucleic acid (DNA), at extremely high efficiency. This efficient tagging allows a person to efficiently and accurately detect rare DNA in heterogenous populations of original DNA fragments (such as in cfDNA). A rare polynucleotide (e.g., rare DNA) can be a polynucleotide that comprises a genetic variant occurring in a population of polynucleotides at a frequency of less than 10%, 5%, 4%, 3%, 2%, 1%, or 0.1%. A rare DNA can be a polynucleotide with a detectable property at a concentration less than 50%, 25%, 10%, 5%, 1%, or 0.1%

Tagging can occur in a single reaction. In some cases, two or more reactions can be performed and pooled together. Tagging each original DNA fragments in a single reaction can result in tagging such that greater than 50% (e.g., 60%, 70%, 80%, 90%, 95%, or 99%) of the original DNA fragments are tagged at both ends with tags that comprise molecular barcodes, thereby providing tagged DNA fragments. Tagging can also result in greater than 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% of the original DNA fragments tagged at both ends with tags that comprise molecular barcodes. Tagging can also result in 100% of the original DNA fragments tagged at both ends with tags that comprise molecular barcodes. Tagging can also result in single end tagging.

Tagging can also occur by using an excess amount of tags as compared to the original DNA fragments. For example, the excess can be at least 5-fold excess. In other cases, the excess can be at least 1.25, 1.5, 1.75, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100 or more fold excess. Tagging can comprise attachment to blunt ends or sticky ends. Tagging can also be performed by hybridization PCR. Tagging can also be performed in low reaction volumes, such as 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or 100 pico- and/or microliters.

The method can also include performing high fidelity amplification on the tagged DNA fragments. Any high fidelity DNA polymerases can be used. For example, the polymerase may be KAPA HiFi DNA polymerase or Phusion DNA polymerase.

Further, the method can comprise selectively enriching a subset of the tagged DNA fragments. For example, selective enrichment can be performed by hybridization or amplification techniques. The selective enrichment can be performed using a solid support (e.g., beads). The solid support (e.g., beads) can comprise probes (e.g., oligonucleotides specifically hybridizing to certain sequences. For example, the probes can hybridize with certain genomic regions, e.g., genes. In some cases, the genomic regions, e.g., genes, can be regions associated with diseases, e.g., cancer. After enrichment, the selected fragmented can be attached any sequencing adaptor disclosed in this invention. For example, a sequence adaptor can comprise a flow cell sequence, a sample barcode, or both. In another example, a sequence adaptor can be a hairpin shaped adaptor and/or comprises a sample barcode. Further, the resulting fragments can be amplified and sequenced. In some cases, the adaptor does not comprise a sequencing primer region.

The method can include sequencing one or both strands of the DNA fragments. In one case, both strands of the DNA fragment are independently sequenced. The tagged, amplified, and/or selectively enriched DNA fragments are sequenced to obtain sequence reads that comprise sequence

information of the molecular barcodes and at least a portion of the original DNA fragments.

The method can include reducing or tracking redundancy (as described above) in the sequence reads to determine consensus reads that are representative of single-strands of the original DNA fragments. For example, to reduce or track redundancy, the method can include comparing sequence reads having the same or similar molecular barcodes and the same or similar end of fragment sequences. The method can comprise performing a phylogentic analysis on the sequence reads having the same or similar molecular barcodes. The molecular barcodes can have a barcode with varying edit distances (including any edit distances as described throughout), for example, an edit distance of up to 3. The end of the fragment sequences can include fragment sequences having an edit distance with varying distances (including any edit distances as described throughout), for example, an edit distance of up to 3.

The method can comprise binning the sequence reads according to the molecular barcodes and sequence information. For example, binning the sequence reads according to the molecular barcodes and sequence information can be performed from at least one end of each of the original DNA fragments to create bins of single stranded reads. The method can further comprise in each bin, determining a sequence of a given original DNA fragment among the original DNA fragments by analyzing sequence reads.

In some cases, sequence reads in each bin can be collapsed to a consensus sequence and subsequently mapped to a genome. As an alternative, sequence reads can be mapped to a genome prior to binning and subsequently collapsed to a consensus sequence.

The method can also comprise sorting sequence reads into paired reads and unpaired reads. After sorting, the number of paired reads and unpaired reads that map to each of one or more genetic loci can be quantified.

The method can include quantifying the consensus reads to detect and/or quantify the rare DNA, which are described throughout. The method can comprise detecting and/or quantifying the rare DNA by comparing a number of times each base occurs at each position of a genome represented by the tagged, amplified, and/or enriched DNA fragments.

The method can comprise tagging the original DNA fragments in a single reaction using a library of tags. The library can include at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 50, at least 100, at least 500, at least 1000, at least 5000, at least 10000, or any number of tags as disclosed throughout. For example, the library of tags can include at least 8 tags. The library of tags can include 8 tags (which can generate 64 different possible combinations). The method can be conducted such that a high percentage of fragments, e.g., greater than 50% (or any percentages as described throughout) are tagged at both ends, wherein each of the tags comprises a molecular barcode.

M. Processing and/or Analyzing Nucleic Acids

The methods described throughout can be used for processing and/or analyzing a nucleic acid sample of a subject. The method can comprising exposing polynucleotide fragments of the nucleic acid sample to a plurality of polynucleotide molecules to yield tagged polynucleotide fragments. The plurality of polynucleotide molecules that can be used are described throughout the application.

For example, the plurality of polynucleotide molecules can be each less than or equal to 40 nucleic acid bases in length and have distinct barcode sequences with respect to at least 4 nucleic acid bases and an edit distance of at least 1, wherein each of the distinct barcode sequences is within 20 nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules, and wherein the plurality of polynucleotide molecules are not sequencing adaptors.

The tagged polynucleotide fragments can be subjected to nucleic acid amplification reactions under conditions that yield amplified polynucleotide fragments as amplification products of the tagged polynucleotide fragments. After amplification, the nucleotide sequence of the amplified tagged polynucleotide fragments is determined. In some cases, the nucleotide sequences of the amplified tagged polynucleotide fragments are determined without the use of polymerase chain reaction (PCR).

The method can comprise analyzing the nucleotide sequences with a programmed computer processor to identify one or more genetic variants in the nucleotide sample of the subject. Any genetic alterations can be identified, including but not limited to, base change(s), insertion(s), repeat(s), deletion(s), copy number variation(s), epigenetic modification(s), nucleosome binding site(s), copy number change(s) due to origin(s) of replication, and transversion(s). Other genetic alterations can include, but are not limited to, one or more tumor associated genetic alterations.

The subject of the methods can be suspected of having a disease. For example, the subject can be suspected of having cancer. The method can comprise collecting a nucleic acid sample from a subject. The nucleic acid sample can be collected from blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, cerebral spinal fluid, skin, hair, sweat, and/or tears. The nucleic acid sample can be a cell-free nucleic acid sample. In some cases, the nucleic acid sample is collected from no more than 100 nanograms (ng) of double-stranded polynucleotide molecules of the subject.

The polynucleotide fragments can comprise double-stranded polynucleotide molecules. In some cases, the plurality of polynucleotide molecules are coupled to the polynucleotide fragments via blunt end ligation, sticky end ligation, molecular inversion probes, polymerase chain reaction (PCR), ligation-based PCR, multiplex PCR, single strand ligation, or single strand circularization.

The method as described herein results in high efficiency tagging of nucleic acids. For example, exposing the polynucleotide fragments of the nucleic acid sample to the plurality of polynucleotide molecules yields the tagged polynucleotide fragments with a conversion efficiency of at least 30%, e.g., of at least 50% (e.g., 60%, 70%, 80%, 90%, 95%, or 99%). Conversion efficiency of at least 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% can be achieved.

The method can result in a tagged polynucleotide fragment that share common polynucleotide molecules. For example, any of at least 5%, 6%, 7%, 8%, 9%, 10%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% or 100% of the tagged polynucleotide fragments share a common polynucleotide molecule. The method can comprise generating the polynucleotide fragments from the nucleic acid sample.

In some cases, the subjecting of the method comprises amplifying the tagged polynucleotide fragments in the pres-

ence primers corresponding to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. Additionally, any combination of these genes can be amplified. For example, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, or all 54 of these genes can be amplified.

The methods described herein can comprise generating a plurality of sequence reads from a plurality of polynucleotide molecules. The plurality of polynucleotide molecules can cover genomic loci of a target genome. For example, the genomic loci can correspond to a plurality of genes as listed above. Further, the genomic loci can be any combination of these genes. Any given genomic locus can comprise at least two nucleic acid bases. Any given genomic locus can also comprise a plurality of nucleic acid bases, for example, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, or more nucleic acid bases.

The method can comprise grouping with a computer processor the plurality of sequence reads into families. Each of the family can comprises sequence reads from one of the template polynucleotides. Each family can comprise sequence reads from only one of the template polynucleotides. For each of the family, the sequence reads can be merged to generate a consensus sequence. The grouping can comprise classifying the plurality of sequence reads into families by identifying (i) distinct molecular barcodes coupled to the plurality of polynucleotide molecules and (ii) similarities between the plurality of sequence reads, wherein each family includes a plurality of nucleic acid sequences that are associated with a distinct combination of molecular barcodes and similar or identical sequence reads.

Once merged, a consensus sequence can be called at a given genomic locus among the genomic loci. At any given genomic loci, any of the following can be determined: i) genetic variants among the calls; ii) frequency of a genetic alteration among the calls; iii) total number of calls; and iv) total number of alterations among the calls. The calling can comprise calling at least one nucleic acid base at the given genomic locus. The calling can also comprise calling a plurality of nucleic acid bases at the given genomic locus. In some cases, the calling can comprise phylogenetic analysis, voting (e.g., biased voting), weighing, assigning a probability to each read at the locus in a family, or calling the base with the highest probability. The consensus sequence can be generated by evaluating a quantitative measure or a statistical significance level for each of the sequence reads. If a quantitative measure is performed, the method can comprise use of a binomial distribution, exponential distribution, beta distribution, or empirical distribution. However, frequency of the base at the particular location can also be used for calling, for example, if 51% or more of the reads is a "A" at the location, then the base may be called an "A" at that particular location. The method can further comprise mapping a consensus sequence to a target genome.

The method can further comprising performing consensus calling at an additional genomic locus among the genomic loci. The method can comprise determining a variation in copy number at one of the given genomic locus and additional genomic locus based on counts at the given genomic locus and additional genomic locus.

The methods described herein can comprise providing template polynucleotide molecules and a library of adaptor polynucleotide molecules in a reaction vessel. The adaptor polynucleotide molecules can have from 2 to 1,000 different barcode sequences and in some cases are not sequencing adaptors. Other variations of adaptor polynucleotide molecules are described throughout, which can also be used in the methods.

The polynucleotide molecules of the adaptors can have the same sample tag. The adaptor polynucleotide molecules can be coupled to both ends of the template polynucleotide molecules. The method can comprise coupling the adaptor polynucleotide molecules to the template polynucleotide molecules at an efficiency of at least 30%, e.g., of at least 50% (e.g., 60%, 70%, 80%, 90%, 95%, or 99%), thereby tagging each template polynucleotide with a tagging combination that is among 4 to 1,000,000 different tagging combinations, to produce tagged polynucleotide molecules. In some cases, the reaction can occur in a single reaction vessel. Coupling efficiency can also be at least 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99%. Tagging can be non-unique tagging.

The tagged polynucleotide molecules can then be subject to an amplification reaction under conditions that will yield amplified polynucleotide molecules as amplification products of the tagged polynucleotide molecules. The template polynucleotide molecules can be double-stranded. Further, the template polynucleotide molecules can be blunt ended. In some cases, the amplification reaction comprises non-specifically amplifying the tagged polynucleotide molecules. The amplification reaction can also comprises using a priming site to amplify each of the tagged polynucleotide molecules. The priming site can be a primer, e.g., a universal primer. The priming site can also be a nick.

The method can also comprise sequencing the amplified polynucleotide molecules. The sequencing can comprise (i) subjecting the amplified polynucleotide molecules to an additional amplification reaction under conditions that yield additional amplified polynucleotide molecules as amplification products of the amplified polynucleotide molecules, and/or (ii) sequencing the additional amplified polynucleotide molecules. The additional amplification can be performed in the presence of primers comprising flow cells sequences, which will produce polynucleotide molecules that are capable of binding to a flow cell. The additional amplification can also be performed in the presence of primers comprising sequences for hairpin shaped adaptors. The hairpin shaped adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. The method can further comprise identifying genetic variants upon sequencing the amplified polynucleotide molecules.

The method can further comprising separating polynucleotide molecules comprising one or more given sequences from the amplified polynucleotide molecules, to produce

enriched polynucleotide molecules. The method can also comprise amplifying the enriched polynucleotide molecules with primers comprising the flow cell sequences. This amplification with primers comprising flow cell sequences will produce polynucleotide molecules that are capable of binding to a flow cell. The amplification can also be performed in the presence of primers comprising sequences for hairpin shaped adaptors. The hairpin shaped adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times.

Flow cell sequences or hairpin shaped adaptors can be added by non-amplification methods such as through ligation of such sequences. Other techniques such as hybridization methods can be used, e.g., nucleotide overhangs.

The method can be performed without aliquoting the tagged polynucleotide molecules. For example, once the tagged polynucleotide molecule is made, the amplification and sequencing can occur in the same tube without any further preparation.

The methods described herein can be useful in detecting single nucleotide variations (SNV), copy number variations (CNV), insertions, deletions, and/or rearrangements. In some cases, the SNVs, CNVs, insertions, deletions, and/or rearrangements, can be associated with disease, for example, cancer.

N. Monitoring a Patient's Status

Methods disclosed herein can also be used to monitor a patient's disease status. The disease of a subject can be monitored over time to determine a progression of the disease (e.g., regression). Markers indicative of the disease can be monitored in a biological sample of the subject, such as a cell-free DNA sample.

For example, monitoring a subject's cancer status can comprise (a) determining an amount of one or more SNVs or copy numbers of a plurality of genes (e.g., in an exon), (b) repeating such determination at different points in time, and (c) determining if there is a difference in the number of SNVs, level of SNVs, number or level of genomic rearrangements, or copy numbers between (a) and (b). The genes can be selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The genes can be selected from any 5, 10, 15, 20, 30, 40, 50, or all of the genes in this group.

O. Sensitivity and Specificity

Methods disclosed herein can be used to detect cancer polynucleotides in a sample, and cancer in a subject, with high measures of agreement, e.g., high sensitivity and/or specificity. For example, such methods can detect cancer polynucleotides (e.g., rare DNA) in a sample at a concentration that is less than 5%, 1%, 0.5%, 0.1%, 0.05%, or 0.01%, at a specificity of at least 99%, 99.9%, 99.99%, 99.999%, 99.9999%, or 99.99999%. Such polynucleotides may be indicative of cancer or other disease. Further, such methods can detect cancer polynucleotides in a sample with a positive predictive value of at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.9%, 99.99%, 99.999%, or 99.9999%.

Subjects identified as positive in a test that are in reality positive are referred as true positives (TP). Subjects identified as positive in a test that are in reality negative are referred as false positives (FP). Subjects identified as negative in a test that are in reality negative are referred as true negatives (TN). Subjects identified as negative in a test that are in reality positive are referred as false negatives (FN). Sensitivity is the percentage of actual positives identified in a test as positive. This includes, for example, instances in which one should have found a cancer genetic variant and did. (Sensitivity=TP/(TP+FN).) Specificity is the percentage of actual negatives identified in a test as negative. This includes, for example, instances in which one should have found no cancer genetic variant and did not. Specificity can be calculated using the following equation: Specificity=TN/(TN+FP). Positive predictive value (PPV) can be measured by the percentage of subjects who test positive that are true positives. PPV can be calculated using the following equation: PPV=TP/(TP+FP). Positive predictive value can be increased by increasing sensitivity (e.g., chance of an actual positive being detected) and/or specificity (e.g., chance of not mistaking an actual negative for a positive).

Low conversion rates of polynucleotides into adaptor-tagged polynucleotides can compromise sensitivity as it decreases the chance of converting, and therefore detecting, rare polynucleotide targets. Noise in a test can compromise specificity as it increases the number of false positives detected in a test. Both low conversion rate and noise compromise positive predictive value as they decrease the percentage of true positives and increase the percentage of false positives.

The methods disclosed herein can achieve high levels of agreement, e.g., sensitivity and specificity, leading to high positive predictive values. Methods of increasing sensitivity include high efficiency conversion of polynucleotides into adaptor-tagged polynucleotides in a sample. Methods of increasing specificity include reducing sequencing errors, for example, by molecular tracking.

Methods of the present disclosure can be used to detect genetic variation in non-uniquely tagged initial starting genetic material (e.g., rare DNA) at a concentration that is less than 5%, 1%, 0.5%, 0.1%, 0.05%, or 0.01%, at a specificity of at least 99%, 99.9%, 99.99%, 99.999%, 99.9999%, or 99.99999%. In some aspects, the methods can further comprise converting polynucleotides in the initial starting material at an efficiency of at least at least 10%, at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, or at least 90%. Sequence reads of tagged polynucleotides can be subsequently tracked to generate consensus sequences for polynucleotides with an error rate of no more than 2%, 1%, 0.1%, or 0.01%.

2. Pooling Methods

Disclosed herein are methods of detecting copy number variation and/or sequence variants at one or more genetic loci in a test sample. One embodiment is shown in FIG. **8**. Typically, detecting copy number variation involves determining a quantitative measure (e.g., an absolute or relative number) of polynucleotides mapping to a genetic locus of interest in a genome of a test sample, and comparing that number to a quantitative measure of polynucleotides mapping to that locus in a control sample. In certain methods, the quantitative measure is determined by comparing the number of molecules in the test sample that map to a locus of interest with a number of molecules in the test sample

mapping to a reference sequence, e.g., a sequence expected to be present at wild type ploidy number. In some examples, the reference sequence is HG19, build 37, or build 38. The comparison could involve, for example, determining a ratio. Then, this measure is compared with a similar measure determined in a control sample. So, for example, if a test sample has a ratio of 1.5:1 for locus of interest versus reference locus, and a control sample has a ratio of 1:1 for the same loci, one may conclude that the test sample exhibits polyploidy at the locus of interest.

When the test sample and the control sample are analyzed separately, the work flow can introduce distortions between final numbers in the control and test samples.

In one method disclosed herein (e.g., flow chart **800**), polynucleotides are provided from a test and a control sample (**802**). Polynucleotides in a test sample and those in a control sample are tagged with tags that identify the polynucleotides as originating from the test or control sample (a source tag). (**804**.) The tag can be, for example, a polynucleotide sequence or barcode that unambiguously identifies the source.

The polynucleotides in each of the control and test samples also can be tagged with identifier tags that will be carried by all amplification progeny of a polynucleotide. Information from start and end sequences of a polynucleotide and identifier tags can identify sequence reads from polynucleotides amplified from an original parent molecule. Each molecule can be uniquely tagged compared with other molecules in the sample. Alternatively, each molecule need not be uniquely tagged compared with other molecules in the sample. That is, the number of different identifier sequences can be fewer than that the number of molecules in sample. By combining identifier information with start/stop sequence information, the probability of confusing two molecules having the same start/stop sequence is significantly diminished.

Number of different identifiers used to tag a nucleic acid (e.g., cfDNA) can dependent on the number of different haploid genome equivalents. Different identifiers can be used to tag at least 2, least 10, least 100, least 200, least 300, least 400, least 500, least 600, least 700, least 800, least 900, least 1,000, least 2,000, least 3,000, least 4,000, least 5,000, least 6,000, least 7,000, least 8,000, least 9,000, least 10,000 or more different haploid genome equivalents. Accordingly, the number of different identifiers used to tag a nucleic acid sample, e.g., cell-free DNA from 500 to 10,000 different haploid genome equivalents and be between any of 1, 2, 3, 4 and 5 and no more than 100, 90, 80, 70, 60, 50, 40 or 30. For example, the number of different identifier used to tag a nucleic acid sample from 500 to 10,000 different haploid genome equivalents can be 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 or less.

Polynucleotides can be tagged by ligation of adaptors comprising the tags or identifiers before amplification. Ligation can be performed using an enzyme, e.g., a ligase. For example, tagging can be performed using a DNA ligase. The DNA ligase can be a T4 DNA ligase, *E. coli* DNA ligase, and/or mammalian ligase. The mammalian ligase can be DNA ligase I, DNA ligase III, or DNA ligase IV. The ligase may also be a thermostable ligase. Tags can be ligated to a blunt-end of a polynucleotide (blunt-end ligation). Alternatively, tags can be ligated to a sticky end of a polynucleotide (sticky-end ligation). The polynucleotides can be tagged by blunt end ligation using adaptors (e.g., adaptors having forked ends). High efficiency of ligation can be achieved using high excess of adaptors (e.g., more than 1.5×, more than 2×, more than 3×, more than 4×, more than 5×, more than 6×, more than 7×, more than 8×, more than 9×, more than 10×, more than 11×, more than 12×, more than 13×, more than 14×, more than 15×, more than 20×, more than 25×, more than 30×, more than 35×, more than 40×, more than 45×, more than 50×, more than 55×, more than 60×, more than 65×, more than 70×, more than 75×, more than 80×, more than 85×, more than 90×, more than 95×, or more than 100).

Once tagged with tags that identify source of the polynucleotides, polynucleotides from different sources (e.g., different samples) can be pooled. After pooling, polynucleotides from different sources (e.g., different samples) can be distinguished by any measurement using the tags, including any process of quantitative measurement. For example, as shown in (**806**) (FIG. **8**), polynucleotides from the control sample and the test sample can be pooled. The pooled molecules can be subject to the sequencing (**808**) and bioinformatic work flow. Both will be subject to the same variations in the process and, therefore, any differential bias is reduced. Because molecules originating from control and test samples are differently tagged, they can be distinguished in any process of quantitative measurement.

The relative amount of control and test sample pooled can be varied. The amount of control sample can be same as the amount of test sample. The amount of control sample can also be larger than the amount of test sample. Alternatively, the amount of control sample can be smaller than the amount of test sample. The smaller the relative amount of one sample to the total, the fewer identifying tags needed in the original tagging process. A number can be selected to reduce to acceptable levels the probability that two parent molecules having the same start/end sequences will bear the same identifying tag. This probability can be less than 10%, less than 1%, less than 0.1% or less than 0.01%. The probability can be less than 25%, 24%, 23%, 22%, 21%, 20%, 19%, 18%, 17%, 16%, 15%, 14%, 13%, 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1%.

Methods disclosed herein can also comprise grouping sequence reads. For example, bioinformatic workflow can include grouping sequence reads produced from progeny of a single parent molecule, as shown in (**810**) (FIG. **8**). This can involve any of the redundancy reduction methods described herein. Molecules sourced from test and control samples can be differentiated based on source tags they carry (**812**). Molecules mapping to a target locus are quantified for both test-sourced and control-sourced molecules (**812**). This can include the normalization methods discussed herein, e.g., in which numbers at a target locus are normalized against numbers at a reference locus.

Normalized (or raw) quantities at a target locus from test and control samples are compared to determine presence of copy number variation (**814**).

3. Computer Control Systems

The present disclosure provides computer control systems that are programmed to implement methods of the disclosure. FIG. **6** shows a computer system **1501** that is programmed or otherwise configured to implement the methods of the present disclosure. The computer system **1501** can regulate various aspects sample preparation, sequencing and/or analysis. In some examples, the computer system **1501** is configured to perform sample preparation and sample analysis, including nucleic acid sequencing. The

computer system **1501** can be an electronic device of a user or a computer system that is remotely located with respect to the electronic device. The electronic device can be a mobile electronic device.

The computer system **1501** includes a central processing unit (CPU, also "processor" and "computer processor" herein) **1505**, which can be a single core or multi core processor, or a plurality of processors for parallel processing. The computer system **1501** also includes memory or memory location **1510** (e.g., random-access memory, read-only memory, flash memory), electronic storage unit **1515** (e.g., hard disk), communication interface **1520** (e.g., network adapter) for communicating with one or more other systems, and peripheral devices **1525**, such as cache, other memory, data storage and/or electronic display adapters. The memory **1510**, storage unit **1515**, interface **1520** and peripheral devices **1525** are in communication with the CPU **1505** through a communication bus (solid lines), such as a motherboard. The storage unit **1515** can be a data storage unit (or data repository) for storing data. The computer system **1501** can be operatively coupled to a computer network ("network") **1530** with the aid of the communication interface **1520**. The network **1530** can be the Internet, an internet and/or extranet, or an intranet and/or extranet that is in communication with the Internet. The network **1530** in some cases is a telecommunication and/or data network. The network **1530** can include one or more computer servers, which can enable distributed computing, such as cloud computing. The network **1530**, in some cases with the aid of the computer system **1501**, can implement a peer-to-peer network, which may enable devices coupled to the computer system **1501** to behave as a client or a server.

The CPU **1505** can execute a sequence of machine-readable instructions, which can be embodied in a program or software. The instructions may be stored in a memory location, such as the memory **1510**. The instructions can be directed to the CPU **1505**, which can subsequently program or otherwise configure the CPU **1505** to implement methods of the present disclosure. Examples of operations performed by the CPU **1505** can include fetch, decode, execute, and writeback.

The CPU **1505** can be part of a circuit, such as an integrated circuit. One or more other components of the system **1501** can be included in the circuit. In some cases, the circuit is an application specific integrated circuit (ASIC).

The storage unit **1515** can store files, such as drivers, libraries and saved programs. The storage unit **1515** can store user data, e.g., user preferences and user programs. The computer system **1501** in some cases can include one or more additional data storage units that are external to the computer system **1501**, such as located on a remote server that is in communication with the computer system **1501** through an intranet or the Internet.

The computer system **1501** can communicate with one or more remote computer systems through the network **1530**. For instance, the computer system **1501** can communicate with a remote computer system of a user (e.g., an operator). Examples of remote computer systems include personal computers (e.g., portable PC), slate or tablet PC's (e.g., Apple® iPad, Samsung® Galaxy Tab), telephones, Smart phones (e.g., Apple® iPhone, Android-enabled device, Blackberry®), or personal digital assistants. The user can access the computer system **1501** via the network **1530**.

Methods as described herein can be implemented by way of machine (e.g., computer processor) executable code stored on an electronic storage location of the computer system **1501**, such as, for example, on the memory **1510** or electronic storage unit **1515**. The machine executable or machine readable code can be provided in the form of software. During use, the code can be executed by the processor **1505**. In some cases, the code can be retrieved from the storage unit **1515** and stored on the memory **1510** for ready access by the processor **1505**. In some situations, the electronic storage unit **1515** can be precluded, and machine-executable instructions are stored on memory **1510**.

The code can be pre-compiled and configured for use with a machine have a processor adapted to execute the code, or can be compiled during runtime. The code can be supplied in a programming language that can be selected to enable the code to execute in a pre-compiled or as-compiled fashion.

Aspects of the systems and methods provided herein, such as the computer system **1501**, can be embodied in programming. Various aspects of the technology may be thought of as "products" or "articles of manufacture" typically in the form of machine (or processor) executable code and/or associated data that is carried on or embodied in a type of machine readable medium. Machine-executable code can be stored on an electronic storage unit, such memory (e.g., read-only memory, random-access memory, flash memory) or a hard disk. "Storage" type media can include any or all of the tangible memory of the computers, processors or the like, or associated modules thereof, such as various semiconductor memories, tape drives, disk drives and the like, which may provide non-transitory storage at any time for the software programming. All or portions of the software may at times be communicated through the Internet or various other telecommunication networks. Such communications, for example, may enable loading of the software from one computer or processor into another, for example, from a management server or host computer into the computer platform of an application server. Thus, another type of media that may bear the software elements includes optical, electrical and electromagnetic waves, such as used across physical interfaces between local devices, through wired and optical landline networks and over various air-links. The physical elements that carry such waves, such as wired or wireless links, optical links or the like, also may be considered as media bearing the software. As used herein, unless restricted to non-transitory, tangible "storage" media, terms such as computer or machine "readable medium" refer to any medium that participates in providing instructions to a processor for execution.

Hence, a machine readable medium, such as computer-executable code, may take many forms, including but not limited to, a tangible storage medium, a carrier wave medium or physical transmission medium. Non-volatile storage media include, for example, optical or magnetic disks, such as any of the storage devices in any computer(s) or the like, such as may be used to implement the databases, etc. shown in the drawings. Volatile storage media include dynamic memory, such as main memory of such a computer platform. Tangible transmission media include coaxial cables; copper wire and fiber optics, including the wires that comprise a bus within a computer system. Carrier-wave transmission media may take the form of electric or electromagnetic signals, or acoustic or light waves such as those generated during radio frequency (RF) and infrared (IR) data communications. Common forms of computer-readable media therefore include for example: a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD or DVD-ROM, any other optical medium, punch cards paper tape, any other physical storage

medium with patterns of holes, a RAM, a ROM, a PROM and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave transporting data or instructions, cables or links transporting such a carrier wave, or any other medium from which a computer may read programming code and/or data. Many of these forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to a processor for execution.

The computer system **1501** can include or be in communication with an electronic display **1535** that comprises a user interface (UI) **1540**. The UI can allow a user to set various conditions for the methods described herein, for example, PCR or sequencing conditions. Examples of UI's include, without limitation, a graphical user interface (GUI) and web-based user interface.

Methods and systems of the present disclosure can be implemented by way of one or more algorithms. An algorithm can be implemented by way of software upon execution by the central processing unit **1505**. The algorithm can, for example, process the reads to generate a consequence sequence.

FIG. **7** schematically illustrates another system for analyzing a sample comprising nucleic acids from a subject. The system includes a sequencer, bioinformatic software and internet connection for report analysis by, for example, a hand held device or a desktop computer

Disclosed herein is a system for analyzing a target nucleic acid molecule of a subject, comprising: a communication interface that receives nucleic acid sequence reads for a plurality of polynucleotide molecules that cover genomic loci of a target genome; computer memory that stores the nucleic acid sequence reads for the plurality of polynucleotide molecules received by the communication interface; and a computer processor operatively coupled to the communication interface and the memory and programmed to (i) group the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides, (ii) for each of the families, merge sequence reads to generate a consensus sequence, (iii) call the consensus sequence at a given genomic locus among the genomic loci, and (iv) detect at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls; and total number of alterations among the calls, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The different variations of each component of the system are described throughout the disclosure within the methods and compositions. These individual components and variations thereof, are also applicable in this system.

4. Kits

Kits comprising the compositions as described herein. The kits can be useful in performing the methods as described herein. Disclosed herein is a kit comprising a plurality of oligonucleotide probes that selectively hybridize to least 5, 6, 7, 8, 9, 10, 20, 30, 40 or all genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The number genes to which the oligonucleotide probes can selectively hybridize can vary. For example, the number of genes can comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, or 54. The kit can include a container that includes the plurality of oligonucleotide probes and instructions for performing any of the methods described herein.

The oligonucleotide probes can selectively hybridize to exon regions of the genes, e.g., of the at least 5 genes. In some cases, the oligonucleotide probes can selectively hybridize to at least 30 exons of the genes, e.g., of the at least 5 genes. In some cases, the multiple probes can selectively hybridize to each of the at least 30 exons. The probes that hybridize to each exon can have sequences that overlap with at least 1 other probe. In some embodiments, the oligoprobes can selectively hybridize to non-coding regions of genes disclosed herein, for example, intronic regions of the genes. The oligoprobes can also selectively hybridize to regions of genes comprising both exonic and intronic regions of the genes disclosed herein.

Any number of exons can be targeted by the oligonucleotide probes. For example, at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, 275, 280, 285, 290, 295, 300, 400, 500, 600, 700, 800, 900, 1,000, or more, exons can be targeted.

The kit can comprise at least 4, 5, 6, 7, or 8 different library adaptors having distinct molecular barcodes and identical sample barcodes. The library adaptors may not be sequencing adaptors. For example, the library adaptors do not include flow cell sequences or sequences that permit the formation of hairpin loops for sequencing. The different variations and combinations of molecular barcodes and sample barcodes are described throughout, and are applicable to the kit. Further, in some cases, the adaptors are not sequencing adaptors. Additionally, the adaptors provided with the kit can also comprise sequencing adaptors. A sequencing adaptor can comprise a sequence hybridizing to one or more sequencing primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, e.g., a flow cell sequence. For example, a sequencing adaptor can be a flow cell adaptor. The sequencing adaptors can be attached to one or both ends of a polynucleotide fragment. In some cases, the kit can comprise at least 8 different library adaptors having distinct molecular barcodes and identical sample barcodes. The library adaptors may not be sequencing adaptors. The kit can further include a sequencing adaptor having a first sequence that selectively hybridizes to the library adaptors and a second sequence that selectively hybridizes to a flow cell sequence. In another example, a sequencing adaptor can be hairpin shaped. For

example, the hairpin shaped adaptor can comprise a complementary double stranded portion and a loop portion, where the double stranded portion can be attached (e.g., ligated) to a double-stranded polynucleotide. Hairpin shaped sequencing adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. A sequencing adaptor can be up to 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, or more bases from end to end. The sequencing adaptor can comprise 20-30, 20-40, 30-50, 30-60, 40-60, 40-70, 50-60, 50-70, bases from end to end. In a particular example, the sequencing adaptor can comprise 20-30 bases from end to end. In another example, the sequencing adaptor can comprise 50-60 bases from end to end. A sequencing adaptor can comprise one or more barcodes. For example, a sequencing adaptor can comprise a sample barcode. The sample barcode can comprise a pre-determined sequence. The sample barcodes can be used to identify the source of the polynucleotides. The sample barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, e.g., at least 8 bases. The barcode can be contiguous or non-contiguous sequences, as described above.

The library adaptors can be blunt ended and Y-shaped and can be less than or equal to 40 nucleic acid bases in length. Other variations of the can be found throughout and are applicable to the kit.

EXAMPLES

Example 1. Methods for Copy Number Variation Detection

Blood Collection

10-30 mL Blood samples are collected at room temperature. The samples are centrifuged to remove cells. Plasma is collected after centrifugation.

cfDNA Extraction

The sample is subjected to proteinase K digestion. DNA is precipitated with isopropanol. DNA is captured on a DNA purification column (e.g., a QIAamp DNA Blood Mini Kit) and eluted in 100 μl solution. DNAs below 500 bp are selected with Ampure SPRI magnetic bead capture (PEG/salt). The resulting production is suspended in 30 μl H_2O. Size distribution is checked (major peak=166 nucleotides; minor peak=330 nucleotides) and quantified. 5 ng of extracted DNA contain approximately 1700 haploid genome equivalents ("HGE"). The general correlation between the amount of DNA and HGE is as follow: 3 pg DNA=1 HGE; 3 ng DNA=1K HGE; 3 μg DNA=1M HGE; 10 pg DNA=3 HE; 10 ng DNA=3K HGE; 10 μg DNA=3M HGE.

"Single Molecule" Library Prep

High-efficiency DNA tagging (>80%) is performed by blunt-end repair and ligation with 8 different octomers (i.e., 64 combinations) with overloaded hairpin adaptors. 2.5 ng DNA (i.e. approximately 800 HGE) is used as the starting material. Each hairpin adaptor comprises a random sequence on its non-complementary portion. Both ends of each DNA fragment are attached with hairpin adaptors. Each tagged fragment can be identified by the random sequence on the hairpin adaptors and a 10 p endogenous sequence on the fragment.

Tagged DNA is amplified by 10 cycles of PCR to produce about 1-7 μg DNAs that contain approximately 500 copies of each of the 800 HGE in the starting material.

Buffer optimization, polymerase optimization and cycle reduction may be performed to optimize the PCR reactions. Amplification bias, e.g., non-specific bias, GC bias, and/or size bias are also reduced by optimization. Noise(s) (e.g., polymerase-introduced errors) are reduced by using high-fidelity polymerases.

The Library may be prepared using Verniata or Sequenom methods.

Sequences may be enriched as follow: DNAs with regions of interest (ROI) are captured using biotin-labeled bead with probe to ROIs. The ROIs are amplified with 12 cycles of PCR to generate a 2000 times amplification. The resulting DNA is then denatured and diluted to 8 pM and loaded into an Illumina sequencer.

Massively Parallel Sequencing

0.1 to 1% of the sample (approximately 100pg) are used for sequencing.

Digital Bioinformatics

Sequence reads are grouped into families, with about 10 sequence reads in each family. Families are collapsed into consensus sequences by voting (e.g., biased voting) each position in a family. A base is called for consensus sequence if 8 or 9 members agree. A base is not called for consensus sequence if no more than 60% of the members agree.

The resulting consensus sequences are mapped to a reference genome. Each base in a consensus sequence is covered by about 3000 different families. A quality score for each sequence is calculated and sequences are filtered based on the their quality scores.

Sequence variation is detected by counting distribution of bases at each locus. If 98% of the reads have the same base (homozygous) and 2% have a different base, the locus is likely to have a sequence variant, presumably from cancer DNA.

CNV is detected by counting the total number of sequences (bases) mapping to a locus and comparing with a control locus. To increase CNV detection, CNV analysis is performed specific regions, including regions on ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, or NTRK1 genes.

Example 2. Method for Correcting Base Calling by Determining the Total Number Unseen Molecules in a Sample

After fragments are amplified and the sequences of amplified fragments are read and aligned, the fragments are subjected to base calling. Variations in the number of amplified fragments and unseen amplified fragments can introduce errors in base calling. These variations are corrected by calculating the number of unseen amplified fragments.

When base calling for locus A (an arbitrary locus), it is first assumed that there are N amplified fragments. The sequence readouts can come from two types of fragments: double-strand fragments and single-strand fragments. The following is a theoretical example of calculating the total number of unseen molecules in a sample.

N is the total number of molecules in the sample.

Assuming 1000 is the number of duplexes detected.

Assuming 500 is the number of single-stranded molecule detected.

P is the probability of seeing a strand.

Q is the probability of not detecting a strand.

Since Q=1−P.

1000=NP(2).

500=N2PQ.

1000/P(2)=N.

500÷2 PQ=N.

1000/P(2)=500+2PQ.

1000*2 PQ=500 P(2).

2000 PQ=500 P(2).

2000 Q=500 P.

2000 (1−P)=500P

2000-2000 P=500P.

2000=500P+2000 P.

2000=2500 P.

2000±2500=P.

0.8=P.

1000/P(2)=N.

1000÷0.64=N.

1562=N.

Number of unseen fragments=62.

Example 3. Identification of Genetic Variants in Cancer-Associated Somatic Variants in a Patient

An assay is used to analyze a panel of genes to identify genetic variants in cancer-associated somatic variants with high sensitivity.

Cell-free DNA is extracted from plasma of a patient and amplified by PCR. Genetic variants are analyzed by massively parallel sequencing of the amplified target genes. For one set of genes, all exons are sequenced as such sequencing coverage had shown to have clinically utility (Table 1). For another set of genes, sequencing coverage included those exons with a previously reported somatic mutation (Table 2). The minimum detectable mutant allele (limit of detection) is dependent on the patient's sample cell-free DNA concentration, which varied from less than 10 to over 1,000 genomic equivalents per mL of peripheral blood. Amplification may not be detected in samples with lower amounts of cell-free DNA and/or low-level gene copy amplification. Certain sample or variant characteristics resulted in reduced analytic sensitivity, such as low sample quality or improper collection.

The percentage of genetic variants found in cell-free DNA circulating in blood is related to the unique tumor biology of this patient. Factors that affected the amount/percentages of detected genetic variants in circulating cell-free DNA in blood include tumor growth, turn-over, size, heterogeneity, vascularization, disease progression or treatment. Table 3 annotates the percentage, or allele frequency, of altered circulating cell-free DNA (% cfDNA) detected in this patient. Some of the detected genetic variants are listed in descending order by % cfDNA.

Genetic variants are detected in the circulating cell-free DNA isolated from this patient's blood specimen. These genetic variants are cancer-associated somatic variants, some of which have been associated with either increased or reduced clinical response to specific treatment. "Minor Alterations" are defined as those alterations detected at less than 10% the allele frequency of "Major Alterations". The detected allele frequencies of these alterations (Table 3) and associated treatments for this patient are annotated.

All genes listed in Tables 1 and 2 are analyzed as part of the Guardant360™ test. Amplification is not detected for ERBB2, EGFR, or MET in the circulating cell-free DNA isolated from this patient's blood specimen.

Patient test results comprising the genetic variants are listed in Table 4.

TABLE 1

Genes in which all exons are sequenced
GENES IN WHICH ALL EXONS ARE SEQUENCED

ALK	<0.1%	APC	<0.1%
AR	<0.1%	BRAF	<0.1%
CDKN2A	<0.1%	EGFR	<0.1%
ERBB2	<0.1%	FBXW7	<0.1%
KRAS	<0.1%	MET	<0.1%
MYC	<0.1%	NOTCH1	<0.1%
NRAS	<0.1%	PIK3CA	<0.1%
PTEN	<0.1%	PROC	<0.1%
RB1	<0.1%	TP53	<0.1%

LOD: Limit of Detection. The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected.

TABLE 2

Genes in which exons with a previously reported somatic mutation are sequenced
GENES IN WHICH EXONS WITH A PREVIOUSLY REPORTED SOMATIC MUTATION ARE SEQUENCED

ABL1	<0.1%	AKT1	<0.1%
ATM	<0.1%	CDH1	<0.1%
CSF1R	<0.1%	CTNNB1	<0.1%
ERBB4	<0.1%	EZH2	<0.1%
FGFR1	<0.1%	FGFR2	<0.1%
FGFR3	<0.1%	FLT3	<0.1%
GNA11	<0.1%	GNAQ	<0.1%
GNAS	<0.1%	HNF1A	<0.1%
HRAS	<0.1%	IDH1	<0.1%
IDH2	<0.1%	JAK2	<0.1%
JAK3	<0.1%	KDR	<0.1%
KIT	<0.1%	MLH1	<0.1%
MPL	<0.1%	NPM1	<0.1%
PDGFRA	<0.1%	PTPN11	<0.1%
RET	<0.1%	SMAD4	<0.1%
SMARCB1	<0.1%	SMO	<0.1%
SRC	<0.1%	STK11	<0.1%
TERT	<0.1%	VHL	<0.1%

LOD: Limit of Detection. The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected.

TABLE 3

Allele frequency of altered circulating cell-free DNA detected in this patient

Gene	cfDNA with alterations (%)	cfDNA without alterations (%)
BRAF V600E	8.9	91.1
NRAS Q61K	6.2	93.8
JAK V617F	1.5	98.6

TABLE 4

Genomic alterations detected in selected genes
Detected: 51 Genomic Alterations

Gene	Chromosome	Position	Mutation (nt)	Mutation (AA)	Percentage	Cosmic ID	DBSNP ID
KRAS	12	25368462	C > T		100.0%		rs4362222
ALK	2	29416572	T > C	I1461V	100.0%		rs1670283
ALK	2	29444095	C > T		100.0%		rs1569156
ALK	2	29543663	T > C	Q500Q	100.0%		rs2293564
ALK	2	29940529	A > T	P234P	100.0%		rs2246745
APC	5	112176756	T > A	V1822D	100.0%		rs459552
CDKN2A	9	21968199	C > G		100.0%	COSM14251	rs11515
FGFR3	4	1807894	G > A	T651T	100.0%		rs7688609
NOTCH1	9	139410424	A > G		100.0%		rs3125006
PDGFRA	4	55141055	A > G	P567P	100.0%		rs1873778
HRAS	11	534242	A > G	H27H	100.0%	COSM249860	rs12628
EGFR	7	55214348	C > T	N158N	99.9%	COSM42978	rs2072454
TP53	17	7579472	G > C	P72R	99.8%		rs1042522
APC	5	112162854	T > C	Y486Y	55.0%		rs2229992
APC	5	112177171	G > A	P1960P	53.8%		rs465899
EGFR	7	55266417	T > C	T903T	53.6%		rs1140475
APC	5	112176325	G > A	G1678G	53.2%		rs42427
APC	5	112176559	T > G	S1756S	53.0%		rs866006
EGFR	7	55229255	G > A	R521K	53.0%		
MET	7	116397572	A > G	Q648Q	52.7%		
APC	5	112175770	G > A	T1493T	52.7%		rs41115
EGFR	7	55249063	G > A	Q787Q	52.6%		rs1050171
NOTCH1	9	139411714	T > C		52.4%		rs11145767
EGFR	7	55238874	T > A	T629T	52.0%		rs2227984
ERBB2	17	37879588	A > G	I655V	51.6%		rs1136201
NOTCH1	9	139397707	G > A	D1698D	51.3%	COSM33747	rs10521
ALK	2	30143499	G > C	L9L	51.0%		rs4358080
APC	5	112164561	G > A	A545A	51.0%		rs351771
FLT3	13	28610183	A > G		50.8%		rs2491231
NOTCH1	9	139418260	A > G	N104N	50.5%		rs4489420
ALK	2	29444076	G > T		50.4%		rs1534545
PIK3CA	3	178917005	A > G		50.3%		rs3729674
NOTCH1	9	139412197	G > A		50.2%		rs9411208
ALK	2	29455267	A > G	G845G	50.0%	COSM148825	rs2256740
KIT	4	55593464	A > C	M541L	49.9%	COSM28026	
NOTCH1	9	139391636	G > A	D2185D	48.9%		rs2229974
PDGFRA	4	55152040	C > T	V824V	48.9%	COSM22413	rs2228230
ALK	2	29416481	T > C	K1491R	48.9%	COSM1130802	rs1881420
ALK	2	29445458	G > T	G1125G	48.6%		rs3795850
NOTCH1	9	139410177	T > C		48.5%		rs3124603
RET	10	43613843	G > T	L769L	48.2%		rs1800861
EGFR	7	55214443	G > A		48.0%		rs7801956
ALK	2	29416366	G > C	D1529E	47.2%		rs1881421
EGFR	7	55238087	C > T		45.5%		rs10258429
RET	10	43615633	C > G	S904S	44.8%		rs1800863
BRAF	7	140453136	A > T	V600E	8.9%	COSM476	
NRAS	1	115256530	G > T	Q61K	6.2%	COSM580	rs121913254
JAK2	9	5073770	G > T	V617F	1.5%	COSM12600	rs77375493

Example 4. Determining Patient-Specific Limits of Detection for Genes Analyzed by Guardant360™ assays.

Using the method of Example 3, Genetic alterations in cell-free DNA of a patient are detected. The sequence reads of these genes include exon and/or intron sequences.

Limits of detection of the test are shown in Table 5. The limits of detection values are dependent on cell-free DNA concentration and sequencing coverage for each gene.

TABLE 5

Limits of Detection of selected genes in a patient using Guardant

Complete Exon and Partial Intron Coverage

APC	0.1%	AR *	0.2%	ARID1A	
BRAF *	0.1%	BRCA1		BRCA2	
CCND1 *		CCND2 *		CCNE1 *	

TABLE 5-continued

Limits of Detection of selected genes in a patient using Guardant

CDK4 *		CDK6 *		CDKN2A	0.1%
CDKN2B		EGFR *	<0.1%	ERBB2 *	0.1%
FGFR1 *	<0.1%	FGFR2 *	0.1%	HRAS	0.1%
KIT *	0.1%	KRAS *	0.1%	MET *	0.1%
MYC *	0.1%	NF1		NRAS	0.1%
PDGFRA *	0.1%	PIK3CA *	0.1%	PTEN	0.1%
RAF1 *		TP53	0.1%		

Exons Covered with Reported Somatic Mutations

AKT1	0.1%	ALK	<0.1%	ARAF	
ATM	0.1%	CDH1	0.1%	CTNNB1	0.1%
ESR1		EZH2	0.1%	FBXW7	0.1%
FGFR3	0.1%	GATA3		GNA11	0.1%
GNAQ	0.1%	GNAS	0.1%	HNF1A	0.1%
IDH1	0.1%	IDH2	0.1%	JAK2	0.1%
JAK3	0.1%	MAP2K1		MAP2K2	
MLH1	0.1%	MPL	0.2%	NFE2L2	
NOTCH1	0.1%	NPM1	0.1%	PTPN11	0.1%

TABLE 5-continued

Limits of Detection of selected genes in a patient using Guardant					
RET	0.1%	RHEB		RHOA	
RIT1		ROS1		SMAD4	0.1%
SMO	0.1%	SRC	<0.1%	STK11	0.2%
TERT	0.1%	VHL	0.2%		
Fusions					
ALK	<0.1%	RET	0.1%	ROS1	
NTRK1					

LOD: Limit of Detection. The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected. * indicates CNV genes.

Example 5. Correcting Sequence Errors Comparing Watson and Crick Sequences

Double-stranded cell-free DNA is isolated from the plasma of a patient. The cell-free DNA fragments are tagged using 16 different bubble-containing adaptors, each of which comprises a distinctive barcode. The bubble-containing adaptors are attached to both ends of each cell-free DNA fragment by ligation. After ligation, each of the cell-free DNA fragment can be distinctly identified by the sequence of the distinct barcodes and two 20 bp endogenous sequences at each end of the cell-free DNA fragment.

The tagged cell-free DNA fragments are amplified by PCR. The amplified fragments are enriched using beads comprising oligonucleotide probes that specifically bind to a group of cancer-associated genes. Therefore, cell-free DNA fragments from the group of cancer-associated genes are selectively enriched.

Sequencing adaptors, each of which comprises a sequencing primer binding site, a sample barcode, and a cell-flow sequence, are attached to the enriched DNA molecules. The resulting molecules are amplified by PCR.

Both strands of the amplified fragments are sequenced. Because each bubble-containing adaptor comprises a non-complementary portion (e.g., the bubble), the sequence of the one strand of the bubble-containing adaptor is different from the sequence of the other strand (complement). Therefore, the sequence reads of amplicons derived from the Watson strand of an original cell-free DNA can be distinguished from amplicons from the Crick strand of the original cell-free DNA by the attached bubble-containing adaptor sequences.

The sequence reads from a strand of an original cell-free DNA fragment are compared to the sequence reads from the other strand of the original cell-free DNA fragment. If a variant occurs in only the sequence reads from one strand, but not other strand, of the original cell-free DNA fragment, this variant will be identified as an error (e.g., resulted from PCR and/or amplification), rather than a true genetic variant.

The sequence reads are grouped into families. Errors in the sequence reads are corrected. The consensus sequence of each family is generated by collapsing.

While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. It is not intended that the invention be limited by the specific examples provided within the specification. While the invention has been described with reference to the aforementioned specification, the descriptions and illustrations of the embodiments herein are not meant to be construed in a limiting sense. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. Furthermore, it shall be understood that all aspects of the invention are not limited to the specific depictions, configurations or relative proportions set forth herein which depend upon a variety of conditions and variables. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is therefore contemplated that the invention shall also cover any such alternatives, modifications, variations or equivalents. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

What is claimed is:

1. A method for classifying consensus sequences generated from sequencing reads derived from double-stranded cell-free deoxyribonucleic acid (cfDNA) molecules from a sample of a human subject, the method comprising:

(a) non-uniquely tagging a population of double-stranded cfDNA molecules from the sample with more than a 10× molar excess of adapters comprising molecular barcodes, relative to the double-stranded cfDNA molecules in the population, to generate non-uniquely tagged parent polynucleotides,

wherein the double-stranded cfDNA molecules that map to a mappable base position of a reference sequence are tagged with a number of different molecular barcodes ranging from at least 2 to fewer than a number of double-stranded cfDNA molecules that map to the mappable base position, and

wherein at least 20% of the double-stranded cfDNA molecules are non-uniquely tagged with the adapters comprising the molecular barcodes at both ends of a molecule of the double-stranded cfDNA molecules;

(b) amplifying a plurality of the non-uniquely tagged parent polynucleotides to produce progeny polynucleotides;

(c) enriching a plurality of the progeny polynucleotides for target regions of interest to generate enriched progeny polynucleotides;

(d) sequencing a plurality of the enriched progeny polynucleotides to produce a set of sequencing reads;

(e) mapping a plurality of sequencing reads from the set of sequencing reads to the reference sequence;

(f) grouping a plurality of the mapped sequencing reads into families of mapped sequencing reads based at least on (i) sequence information from the molecular barcodes and (ii) a beginning base position and an ending base position of the mapped sequencing reads;

(g) generating a consensus sequence for each family from among one or more of the families to produce a set of consensus sequences; and

(h) classifying one or more consensus sequences from among the set of consensus sequences as (1) paired consensus sequences generated from sequencing reads representing a Watson strand and a Crick strand of a non-uniquely tagged parent polynucleotide or (2) unpaired consensus sequences generated from sequencing reads representing only one of either a Watson strand or a Crick strand of a non-uniquely tagged parent polynucleotide.

2. The method of claim **1**, wherein the population of double-stranded cfDNA molecules comprises 1 nanogram (ng) to 100 ng of double-stranded cfDNA molecules.

3. The method of claim **1**, wherein the sample is blood, plasma, or serum.

4. The method of claim 1, wherein the adapters comprising the molecular barcodes are non-uniquely tagged to the double-stranded cfDNA molecules by blunt-end ligation or sticky-end ligation.

5. The method of claim 1, wherein at least 40% of the double-stranded cfDNA molecules are non-uniquely tagged with the adapters comprising the molecular barcodes at both ends of a molecule of the double-stranded cfDNA molecules.

6. The method of claim 1, wherein more than a 90× molar excess of the adapters relative to the double-stranded cfDNA molecules in the population is used to generate the non-uniquely tagged parent polynucleotides.

7. The method of claim 1, wherein the molecular barcodes are from a set of molecular barcodes having 2 to 1,000 different molecular barcode sequences.

8. The method of claim 1, wherein the molecular barcodes are from a set of molecular barcodes having 5 to 100 different molecular barcode sequences that have a length of 5 to 20 nucleotides.

9. The method of claim 1, wherein the target regions of interest associated with cancer comprise genetic sequences of a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

10. The method of claim 1, further comprising amplifying a plurality of the enriched progeny polynucleotides prior to sequencing.

11. The method of claim 1, further comprising:

(i) calculating a first quantitative measure of paired consensus sequences that map to a locus of the reference sequence, and

(ii) calculating a second quantitative measure of unpaired consensus sequences that map to the locus of the reference sequence.

12. The method of claim 11, further comprising:

(iii) calculating a third quantitative measure of double-stranded cfDNA molecules that map to the locus of the reference sequence for which neither complementary strand of the double-stranded cfDNA molecules is detected in the set of consensus sequences, wherein the third quantitative measure is calculated based at least in part on the first quantitative measure and the second quantitative measure.

13. The method of claim 12, further comprising:

(iv) calculating a fourth quantitative measure of a number of non-uniquely tagged double-stranded cfDNA molecules, wherein the number of non-uniquely tagged double-stranded cfDNA molecules comprises unseen double-stranded cfDNA molecules, and wherein the number of non-uniquely tagged double-stranded cfDNA molecules is calculated based at least in part on the first quantitative measure, the second quantitative measure, and the third quantitative measure.

14. The method of claim 13, comprising calculating the first quantitative measure, the second quantitative measure, the third quantitative measure, and the fourth quantitative measure with a programmed computer processor.

15. A method for classifying unique sequencing reads generated from sequencing reads derived from double-stranded cell-free deoxyribonucleic acid (cfDNA) molecules from a bodily fluid sample of a human subject, the method comprising:

(a) tagging a population of double-stranded cfDNA molecules from the bodily fluid sample with more than a 10× molar excess of adapters comprising molecular barcodes, relative to the double-stranded cfDNA molecules in the population, to generate tagged parent polynucleotides,

wherein at least 20% of the cfDNA molecules are ligated with the adapters comprising the molecular barcodes at both ends of a molecule of the double-stranded cfDNA molecules;

(b) amplifying a plurality of the tagged parent polynucleotides to produce progeny polynucleotides;

(c) sequencing a plurality of the progeny polynucleotides to produce a set of sequencing reads;

(d) mapping a plurality of sequencing reads from the set of sequencing reads to a reference sequence;

(e) determining unique sequencing reads from the set of mapped sequencing reads based at least on the molecular barcode sequences, wherein a unique sequencing read from among the unique sequencing reads is representative of a tagged parent polynucleotide from among the tagged parent polynucleotides; and

(f) classifying one or more of the unique sequencing reads as either (1) paired sequences generated from sequencing reads representing a Watson strand and a Crick strand of a tagged parent polynucleotide or (2) unpaired sequences generated from sequencing reads representing only one of either a Watson strand or a Crick strand of a tagged parent polynucleotide.

16. The method of claim 15, further comprising selectively enriching a plurality of the progeny polynucleotides for target regions of.

17. The method of claim 16, wherein determining the unique sequencing reads comprises grouping a plurality of the mapped sequencing reads into families, wherein a family of the families comprises mapped sequencing reads of progeny polynucleotides amplified from a same tagged parent polynucleotide from among the tagged parent polynucleotides.

18. The method of claim 15, further comprising:

(i) calculating a first quantitative measure of paired sequences that map to a locus of the reference sequence, and

(ii) calculating a second quantitative measure of unpaired sequences that map to the locus of the reference sequence.

19. The method of claim 18, further comprising:

(iii) calculating a third quantitative measure of double-stranded cfDNA molecules that map to the locus of the reference sequence for which neither complementary strand of the double-stranded cfDNA molecules is detected in the set of unique sequence reads, wherein the third quantitative measure is calculated based at least in part on the first quantitative measure and the second quantitative measure.

20. The method of claim 19, further comprising:

(iv) calculating a fourth quantitative measure of a number of non-uniquely tagged double-stranded cfDNA molecules, wherein the number of non-uniquely tagged double-stranded cfDNA molecules comprises unseen double-stranded cfDNA molecules, and wherein the number of non-uniquely tagged double-stranded

cfDNA molecules is determined based at least in part on the first quantitative measure, the second quantitative measure, and the third quantitative measure.

21. The method of claim **20**, comprising calculating the first quantitative measure, the second quantitative measure, the third quantitative measure, and the fourth quantitative measure with a programmed computer processor.

22. The method of claim **15**, wherein the molecular barcodes are from a set of molecular barcodes having 2 to 10,000 different molecular barcode sequences.

23. The method of claim **15**, wherein the molecular barcodes are pre-determined sequences.

24. The method of claim **15**, wherein at least 40% of the cfDNA molecules are tagged with the adapters comprising the molecular barcodes at both ends of a molecule of the double-stranded cfDNA molecules.

25. The method of claim **15**, wherein more than a 90× molar excess of the adapters relative to the double-stranded cfDNA molecules in the population is used to generate the tagged parent polynucleotides.

26. The method of claim **15**, further comprising quantifying a number of the unique sequencing reads determined from the set of mapped sequencing reads.

27. The method of claim **15**, wherein determining the unique sequencing reads is further based on (1) a start base position of a given sequencing read from among the set of mapped sequencing reads at which the given sequencing read starts aligning to the reference sequence, and (2) a stop base position of the given sequencing read at which the given sequencing read stops aligning to the reference sequence.

28. The method of claim **1**, further comprising determining a quantity of the set of consensus sequences.

* * * * *

EXHIBIT B

US010889858B2

(12) **United States Patent**
Talasaz et al.

(10) **Patent No.: US 10,889,858 B2**
(45) **Date of Patent: *Jan. 12, 2021**

(54) **METHODS AND SYSTEMS FOR DETECTING GENETIC VARIANTS**

(71) Applicant: **GUARDANT HEALTH, INC.**, Redwood City, CA (US)

(72) Inventors: **AmirAli Talasaz**, Atherton, CA (US); **Stefanie Ann Ward Mortimer**, Morgan Hill, CA (US)

(73) Assignee: **GUARDANT HEALTH, INC.**, Redwood City, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **16/714,579**

(22) Filed: **Dec. 13, 2019**

(65) **Prior Publication Data**

US 2020/0131568 A1 Apr. 30, 2020

Related U.S. Application Data

(63) Continuation of application No. 16/672,267, filed on Jan. 7, 2020, which is a continuation of application No. 16/601,168, filed on Oct. 14, 2019, which is a continuation of application No. 15/892,178, filed on Feb. 8, 2018, which is a continuation of application No. 14/861,989, filed on Sep. 22, 2015, now Pat. No. 9,920,366, which is a continuation of application No. PCT/US2014/072383, filed on Dec. 24, 2014.

(60) Provisional application No. 61/948,509, filed on Mar. 5, 2014, provisional application No. 61/921,456, filed on Dec. 28, 2013.

(51) **Int. Cl.**

C12P 19/34	(2006.01)
C12Q 1/6869	(2018.01)
C12Q 1/6886	(2018.01)
G16B 15/00	(2019.01)

(52) **U.S. Cl.**

CPC ***C12Q 1/6869*** (2013.01); ***C12Q 1/6886*** (2013.01); *C12Q 2535/122* (2013.01); *C12Q 2600/158* (2013.01); *G16B 15/00* (2019.02)

(58) **Field of Classification Search**

None

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,725,536	A	2/1988	Fritsch et al.
4,942,124	A	7/1990	Church
5,124,246	A	6/1992	Urdea et al.
5,149,625	A	9/1992	Church et al.
5,200,314	A	4/1993	Urdea
5,424,186	A	6/1995	Fodor et al.
5,424,413	A	6/1995	Hogan et al.
5,445,934	A	8/1995	Fodor et al.
5,604,097	A	2/1997	Brenner
5,635,352	A	6/1997	Urdea et al.
5,635,400	A	6/1997	Brenner
5,648,245	A	7/1997	Fire et al.
5,654,413	A	8/1997	Brenner
5,656,731	A	8/1997	Urdea
5,658,737	A	8/1997	Nelson et al.
5,714,330	A	2/1998	Brenner et al.
5,744,305	A	4/1998	Fodor et al.
5,759,778	A	6/1998	Li et al.
5,763,175	A	6/1998	Brenner
5,800,992	A	9/1998	Fodor et al.
5,846,719	A	12/1998	Brenner et al.
5,854,033	A	12/1998	Lizardi
5,871,928	A	2/1999	Fodor et al.
5,925,525	A	7/1999	Fodor et al.
5,935,793	A	8/1999	Wong
5,952,170	A	9/1999	Stroun et al.
5,968,740	A	10/1999	Fodor et al.
5,981,176	A	11/1999	Wallace
5,981,179	A	11/1999	Lorinez et al.
6,013,445	A	1/2000	Albrecht et al.

(Continued)

FOREIGN PATENT DOCUMENTS

CN	102933721	A	2/2013
EP	0799897	B1	6/2006

(Continued)

OTHER PUBLICATIONS

Ingolia, et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. Apr. 10, 2009;324(5924)118-23. Epub Feb. 12, 2009.

Instructions for Norit Rapid DNA Ligation Kit (Nov. 6, 2004).

International search report and written opinion dated Apr. 3, 2015 for PCT/US2014/072383.

International search report and written opinion dated May 7, 2012 for PCT/IB2011/003160.

International search report and written opinion dated Jun. 6, 2012 for PCT/US2011/065291.

(Continued)

Primary Examiner — Kenneth R Horlick

(74) *Attorney, Agent, or Firm* — Timothy A. Hott

(57) **ABSTRACT**

Disclosed herein in are methods and systems for determining genetic variants (e.g., copy number variation) in a polynucleotide sample. A method for determining copy number variations includes tagging double-stranded polynucleotides with duplex tags, sequencing polynucleotides from the sample and estimating total number of polynucleotides mapping to selected genetic loci. The estimate of total number of polynucleotides can involve estimating the number of double-stranded polynucleotides in the original sample for which no sequence reads are generated. This number can be generated using the number of polynucleotides for which reads for both complementary strands are detected and reads for which only one of the two complementary strands is detected.

29 Claims, 11 Drawing Sheets

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,020,124 A 2/2000 Sorenson
6,040,138 A 3/2000 Lockhart et al.
6,046,005 A 4/2000 Ju et al.
6,060,596 A 5/2000 Lemer et al.
6,117,631 A 9/2000 Nilsen
6,124,092 A 9/2000 O'Neill et al.
6,138,077 A 10/2000 Brenner
6,140,489 A 10/2000 Brenner
6,172,214 B1 1/2001 Brenner
6,197,506 B1 3/2001 Fodor et al.
6,235,475 B1 5/2001 Brenner et al.
6,242,186 B1 6/2001 Salonen
6,268,152 B1 7/2001 Fodor et al.
6,284,460 B1 9/2001 Fodor et al.
6,300,077 B1 10/2001 Shuber et al.
6,309,822 B1 10/2001 Fodor et al.
6,309,823 B1 10/2001 Cronin et al.
6,326,148 B1 12/2001 Pauletti et al.
6,355,431 B1 3/2002 Chee et al.
6,355,432 B1 3/2002 Fodor et al.
6,395,491 B1 5/2002 Fodor et al.
6,406,848 B1 6/2002 Bridgham et al.
6,440,667 B1 8/2002 Fodor et al.
6,440,706 B1 8/2002 Vogelstein et al.
6,451,536 B1 9/2002 Fodor et al.
6,458,530 B1 10/2002 Morris et al.
6,468,744 B1 10/2002 Cronin et al.
6,489,114 B2 12/2002 Laayoun et al.
6,492,121 B2 12/2002 Kurane et al.
6,498,012 B2 12/2002 Laken
6,503,718 B2 1/2003 Shuber et al.
6,512,105 B1 1/2003 Hogan et al.
6,514,699 B1 2/2003 O'Neill et al.
6,544,739 B1 4/2003 Fodor et al.
6,551,784 B2 4/2003 Fodor et al.
6,576,424 B2 6/2003 Fodor et al.
6,582,908 B2 6/2003 Fodor et al.
6,586,177 B1 7/2003 Shuber
6,600,996 B2 7/2003 Webster et al.
6,629,040 B1 9/2003 Goodlett et al.
6,653,077 B1 11/2003 Brenner
6,753,147 B2 6/2004 Vogelstein et al.
6,849,403 B1 2/2005 Shuber
6,849,404 B2 2/2005 Park et al.
6,852,488 B2 2/2005 Fodor et al.
6,858,412 B2 2/2005 Willis et al.
6,964,846 B1 11/2005 Shuber
7,163,789 B2 1/2007 Chen et al.
7,208,275 B2 4/2007 Gocke et al.
7,406,385 B2 7/2008 Sorenson
7,410,764 B2 8/2008 Gocke et al.
7,424,368 B2 9/2008 Huang et al.
7,424,371 B2 9/2008 Kamentsky
7,537,897 B2 5/2009 Brenner et al.
7,700,286 B2 4/2010 Stroun et al.
7,704,687 B2 4/2010 Wang et al.
7,727,720 B2 6/2010 Dhallan
7,803,929 B2 9/2010 Melkonyan et al.
7,811,757 B2 10/2010 Shuber
7,822,555 B2 10/2010 Huang et al.
7,824,889 B2 11/2010 Vogelstein et al.
7,838,647 B2 11/2010 Hahn et al.
7,915,015 B2 3/2011 Vogelstein et al.
7,935,487 B2 5/2011 Gocke et al.
7,937,225 B2 5/2011 Mishra et al.
7,957,913 B2 6/2011 Chinitz et al.
7,972,817 B2 7/2011 Kopreski
7,981,612 B2 7/2011 Shuber et al.
8,094,312 B2 1/2012 Ulmer
8,168,385 B2 5/2012 Brenner
8,195,415 B2 6/2012 Fan et al.
8,209,130 B1 6/2012 Kennedy et al.
8,216,789 B2 7/2012 Disis et al.
8,236,532 B2 8/2012 Ronaghi et al.
8,361,726 B2 1/2013 Gocke et al.
8,383,338 B2 2/2013 Kitzman et al.
8,383,345 B2 2/2013 Shendure et al.
8,455,193 B2 6/2013 Travers et al.
8,470,996 B2 6/2013 Brenner
8,481,292 B2 7/2013 Casbon et al.
8,603,749 B2 12/2013 Gillevet
8,685,678 B2 4/2014 Casbon et al.
8,697,408 B2 4/2014 Kucera et al.
8,704,165 B2 4/2014 Huang
8,715,967 B2 5/2014 Casbon et al.
8,722,368 B2 5/2014 Casbon et al.
8,728,766 B2 5/2014 Casbon et al.
8,741,606 B2 6/2014 Casbon et al.
8,775,092 B2 7/2014 Colwell et al.
8,835,358 B2 9/2014 Fodor et al.
8,865,410 B2 10/2014 Shendure et al.
9,018,365 B2 4/2015 Brenner
9,085,798 B2 7/2015 Chee
9,260,753 B2 2/2016 Xie et al.
9,340,830 B2 5/2016 Lipson et al.
9,376,719 B2 6/2016 Eijk et al.
9,404,156 B2 8/2016 Hicks et al.
9,476,095 B2 10/2016 Vogelstein et al.
9,598,731 B2 3/2017 Talasaz
9,752,188 B2 9/2017 Schmitt et al.
9,834,822 B2 12/2017 Talasaz
9,840,743 B2 12/2017 Talasaz
9,920,366 B2 3/2018 Eltoukhy et al.
10,287,631 B2 5/2019 Salk et al.
10,370,713 B2 8/2019 Salk et al.
10,385,393 B2 8/2019 Salk et al.
10,388,403 B2 8/2019 Rava et al.
10,604,804 B2 3/2020 Salk et al.
2001/0053519 A1 12/2001 Fodor et al.
2002/0072058 A1 6/2002 Voelker et al.
2003/0003490 A1 1/2003 Fan et al.
2003/0049616 A1 3/2003 Brenner et al.
2003/0104436 A1 6/2003 Morris et al.
2003/0152490 A1 8/2003 Trulson et al.
2003/0165978 A1 9/2003 Firth et al.
2003/0186251 A1 10/2003 Dunn et al.
2003/0207300 A1 11/2003 Matray et al.
2004/0096892 A1 5/2004 Wang et al.
2004/0146901 A1 7/2004 Morris et al.
2004/0157243 A1 8/2004 Huang et al.
2004/0209299 A1 10/2004 Pinter et al.
2004/0259118 A1 12/2004 Macevicz
2005/0164241 A1 7/2005 Hahn et al.
2005/0221314 A1 10/2005 Berlin et al.
2005/0250147 A1 11/2005 Macevicz
2006/0035258 A1 2/2006 Tadakamalla et al.
2006/0046258 A1 3/2006 Lapidus et al.
2006/0073506 A1 4/2006 Christians et al.
2006/0211030 A1 9/2006 Brenner
2007/0020640 A1 1/2007 McCloskey et al.
2007/0065823 A1 3/2007 Dressman et al.
2007/0111233 A1 5/2007 Bianchi et al.
2007/0128724 A1 6/2007 Miles et al.
2007/0172839 A1 7/2007 Smith et al.
2007/0172873 A1 7/2007 Brenner et al.
2008/0014146 A1 1/2008 Hoff et al.
2008/0124721 A1 5/2008 Fuchs et al.
2008/0161420 A1 7/2008 Shuber
2008/0261204 A1 10/2008 Lexow
2008/0293055 A1 11/2008 Freeman et al.
2009/0029377 A1 1/2009 Lo et al.
2009/0087847 A1 4/2009 Lo et al.
2009/0098547 A1 4/2009 Ghosh
2009/0105959 A1 4/2009 Braverman et al.
2009/0162836 A1 6/2009 Widschwendter
2009/0298075 A1 12/2009 Travers et al.
2009/0298709 A1 12/2009 Ma
2009/0317818 A1 12/2009 Ehrich et al.
2010/0041048 A1 2/2010 Diehl et al.
2010/0062494 A1 3/2010 Church et al.
2010/0069250 A1 3/2010 White et al.
2010/0112575 A1 5/2010 Fan et al.
2010/0143932 A1 6/2010 Lapidus
2010/0166744 A1 7/2010 Wong

(56) **References Cited**

U.S. PATENT DOCUMENTS

2010/0196898	A1	8/2010	Sugarbaker et al.	
2010/0264331	A1	10/2010	Sacko et al.	
2010/0323348	A1	12/2010	Hamady et al.	
2010/0330571	A1	12/2010	Robins et al.	
2011/0014607	A1	1/2011	Jirtle et al.	
2011/0160078	A1	6/2011	Fodor et al.	
2011/0160290	A1	6/2011	Tewari	
2011/0171640	A1	7/2011	Bhatt et al.	
2011/0177512	A1	7/2011	Shuber	
2011/0183330	A1	7/2011	Lo et al.	
2011/0201507	A1	8/2011	Rava et al.	
2011/0230358	A1	9/2011	Rava	
2011/0230360	A1	9/2011	Stephan et al.	
2011/0245482	A1	10/2011	Hahn et al.	
2011/0264376	A1	10/2011	Chinitz et al.	
2011/0275084	A1	11/2011	Byron et al.	
2011/0281736	A1	11/2011	Drmanac et al.	
2011/0319272	A1	12/2011	Fan et al.	
2012/0003637	A1	1/2012	Lo et al.	
2012/0034685	A1	2/2012	Sparks et al.	
2012/0046877	A1	2/2012	Hyland et al.	
2012/0053073	A1	3/2012	Kassis	
2012/0059594	A1	3/2012	Hatchwell et al.	
2012/0065081	A1	3/2012	Chee	
2012/0095697	A1	4/2012	Halpern et al.	
2012/0100548	A1	4/2012	Rava et al.	
2012/0164630	A1	6/2012	Porreca et al.	
2012/0165202	A1	6/2012	Porreca et al.	
2012/0191367	A1	7/2012	Stuelpnagel et al.	
2012/0208706	A1	8/2012	Downing et al.	
2012/0214163	A1	8/2012	Sugarbaker et al.	
2012/0214678	A1	8/2012	Rava et al.	
2012/0220478	A1	8/2012	Shaffer	
2012/0231479	A1	9/2012	Puskas et al.	
2012/0238464	A1	9/2012	Koi et al.	
2012/0270212	A1	10/2012	Rabinowitz et al.	
2012/0316074	A1	12/2012	Saxonov	
2012/0322681	A1	12/2012	Kung et al.	
2013/0005585	A1	1/2013	Anderson et al.	
2013/0017958	A1	1/2013	Benz et al.	
2013/0022977	A1	1/2013	Lapidus et al.	
2013/0029852	A1	1/2013	Rava et al.	
2013/0034546	A1	2/2013	Rava et al.	
2013/0040824	A1	2/2013	Lo et al.	
2013/0053256	A1	2/2013	Hubbell	
2013/0060483	A1	3/2013	Struble et al.	
2013/0078626	A1	3/2013	Wasserstrom et al.	
2013/0085681	A1	4/2013	Deciu et al.	
2013/0096011	A1	4/2013	Rava et al.	
2013/0102485	A1	4/2013	Lee	
2013/0102487	A1	4/2013	Cos et al.	
2013/0116127	A1	5/2013	Schuetz et al.	
2013/0116130	A1	5/2013	Fu et al.	
2013/0122499	A1	5/2013	Morris et al.	
2013/0130923	A1	5/2013	Ehrich et al.	
2013/0137588	A1	5/2013	Shendure et al.	
2013/0143747	A1	6/2013	Gutin et al.	
2013/0210643	A1	8/2013	Casbon et al.	
2013/0210645	A1	8/2013	Volgelstein et al.	
2013/0224743	A1	8/2013	Casbon et al.	
2013/0237431	A1	9/2013	Lo et al.	
2013/0237458	A1	9/2013	Casbon et al.	
2013/0260381	A1	10/2013	Ramakrishnan	
2013/0267424	A1	10/2013	Casbon et al.	
2013/0288244	A1*	10/2013	Deciu	G16B 20/00 435/6.11
2013/0338933	A1	12/2013	Deciu et al.	
2014/0011694	A1	1/2014	Couronne	
2014/0065609	A1	3/2014	Hicks et al.	
2014/0065630	A1	3/2014	Bubnoff et al.	
2014/0066317	A1	3/2014	Talasaz	
2014/0100121	A1	4/2014	Lo et al.	
2014/0242588	A1	8/2014	Boom et al.	
2014/0274731	A1	9/2014	Raymond et al.	
2014/0274740	A1	9/2014	Srinivasan et al.	
2014/0296094	A1	10/2014	Domanus	
2014/0303008	A1	10/2014	Schutz et al.	
2014/0336943	A1	11/2014	Pellini et al.	
2014/0350130	A1	11/2014	Sanborn et al.	
2014/0357500	A1	12/2014	Vigneault et al.	
2015/0004158	A1	1/2015	Shipp et al.	
2015/0024950	A1	1/2015	Bielas et al.	
2015/0044687	A1	2/2015	Schmitt et al.	
2015/0050647	A1	2/2015	Luo et al.	
2015/0051085	A1	2/2015	Vogelstein et al.	
2015/0065358	A1	3/2015	Comstock et al.	
2015/0080266	A1	3/2015	Volkmuth et al.	
2015/0087535	A1	3/2015	Patel	
2015/0167069	A1	6/2015	Schutz et al.	
2015/0197786	A1	7/2015	Osborne et al.	
2015/0275289	A1	10/2015	Otwinowski et al.	
2015/0299812	A1	10/2015	Talasaz	
2015/0329917	A1	11/2015	Shuber	
2015/0344970	A1	12/2015	Vogelstein et al.	
2015/0366866	A1	12/2015	Ali et al.	
2015/0368708	A1	12/2015	Talasaz	
2016/0002739	A1	1/2016	Schütz et al.	
2016/0002741	A1	1/2016	Kitano et al.	
2016/0024576	A1	1/2016	Chee	
2016/0026758	A1	1/2016	Jabara et al.	
2016/0032396	A1	2/2016	Diehn et al.	
2016/0040229	A1	2/2016	Talasaz et al.	
2016/0046986	A1	2/2016	Eltoukhy et al.	
2016/0053301	A1	2/2016	Raymond et al.	
2016/0060691	A1	3/2016	Giresi et al.	
2016/0071432	A1	3/2016	Kurowski et al.	
2016/0115553	A1	4/2016	Stephan et al.	
2016/0130649	A1	5/2016	Xie et al.	
2016/0251704	A1	9/2016	Talasaz et al.	
2016/0319345	A1	11/2016	Gnerre et al.	
2016/0333417	A1	11/2016	Talasaz	
2016/0376647	A1	12/2016	Travers et al.	
2017/0051347	A1	2/2017	Vogelstein et al.	
2017/0073774	A1	3/2017	Lo et al.	
2017/0145516	A1	5/2017	Kopetz et al.	
2017/0159120	A1	6/2017	Eijk et al.	
2017/0211143	A1	7/2017	Shendure et al.	
2017/0218459	A1	8/2017	Talasaz et al.	
2017/0218460	A1	8/2017	Talasaz	
2018/0023125	A1	1/2018	Talasaz et al.	
2019/0271040	A1	9/2019	Salk et al.	
2019/0292597	A1	9/2019	Salk et al.	
2019/0338358	A1	11/2019	Salk et al.	
2019/0352714	A1	11/2019	Salk et al.	

FOREIGN PATENT DOCUMENTS

EP	1647600	A2	6/2006
EP	2110442	A1	10/2009
EP	3070177	A1	9/2016
EP	3178941	A1	6/2017
WO	1997007241	A1	2/1997
WO	1997010365	A1	3/1997
WO	1999028505	A1	6/1999
WO	2000058516	A2	10/2000
WO	2002056014	A2	7/2002
WO	2005080604	A2	9/2005
WO	2005111242	A2	11/2005
WO	2006102264	A1	9/2006
WO	2007037678	A2	4/2007
WO	2008070144	A2	6/2008
WO	2008154317	A1	12/2008
WO	2009152928	A2	12/2009
WO	2011060240	A1	5/2011
WO	2011087760	A2	7/2011
WO	2011091046	A1	7/2011
WO	2011103236	A2	8/2011
WO	2011140510	A2	11/2011
WO	2011155833	A2	12/2011
WO	2012012693	A2	1/2012
WO	2012014877	A1	2/2012
WO	2012019200	A2	2/2012
WO	2012028746	A1	3/2012
WO	2012038839	A2	3/2012

(56) **References Cited**

FOREIGN PATENT DOCUMENTS

WO	2012042374	A2	4/2012
WO	2012048341	A1	4/2012
WO	2012054873	A2	4/2012
WO	2012066451	A1	5/2012
WO	2012071621	A1	6/2012
WO	2012088348	A2	6/2012
WO	2012106559	A1	8/2012
WO	2012129363	A2	9/2012
WO	2012142213	A2	10/2012
WO	2012142611	A2	10/2012
WO	2012148477	A1	11/2012
WO	2012103031	A3	1/2013
WO	2013019075	A2	2/2013
WO	2013033721	A1	3/2013
WO	2013106737	A1	7/2013
WO	2013123442	A1	8/2013
WO	2013130512	A2	9/2013
WO	2013130674	A1	9/2013
WO	2013138510	A1	9/2013
WO	2013142213	A1	9/2013
WO	2013142389	A1	9/2013
WO	2013148496	A1	10/2013
WO	2013159035	A2	10/2013
WO	2013173394	A2	11/2013
WO	2013181170	A1	12/2013
WO	2013188471	A2	12/2013
WO	2013190441	A2	12/2013
WO	2014004726	A1	1/2014
WO	2014014497	A1	1/2014
WO	2014015319	A1	1/2014
WO	2014039556	A1	3/2014
WO	2014093330	A1	6/2014
WO	2014145078	A1	9/2014
WO	2014149134	A2	9/2014
WO	2014151117	A1	9/2014
WO	2014152990	A1	9/2014
WO	2015159293	A2	10/2015
WO	2016015058	A2	1/2016
WO	2016040901	A1	3/2016
WO	2017100441	A1	6/2017
WO	2017181146	A1	10/2017

OTHER PUBLICATIONS

International search report and written opinion dated Jul. 15, 2015 for PCT/US2015/030639.

International search report and written opinion dated Sep. 5, 2014 for PCT/US2014/000048.

International search report and written opinion dated Sep. 6, 2017 for PCT Application No. US-201727809.

International search report and written opinion dated Nov. 18, 2013 for PCT/US2013/058061.

Invitrogen Instructions for T4 DNA Ligase (May 5, 2002).

IPR2018-00130, Petition for Inter Partes Review of U.S. Pat. No. 9,598,731, dated Nov. 7, 2018.

IPR2019-00130, Decision Denying Inter Partes Review U.S. Pat. No. 9,598,731, issued Jun. 5, 2019.

IPR2019-00130, Preliminary Response to Petition for Inter Partes Review U.S. Pat. No. 9,598,731, filed Mar. 6, 2019.

IPR2019-00634, Decision Instituting Inter Partes Review U.S. Pat. No. 9,840,743, issued Aug. 19, 2019.

IPR2019-00634, Petition for Inter Partes Review of U.S. Pat. No. 9,840,743, dated Feb. 1, 2019.

IPR2019-00636 & 637, Exhibit 3001 Memo for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Jul. 1, 2019.

IPR2019-00636 & 637, Order—Conduct of the Proceedings for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Jul. 22, 2019.

IPR2019-00636 and IPR2019-00637, Decision Denying Inter Partes Review U.S. Pat. No. 9,902,992, issued Aug. 20, 2019.

IPR2019-00636, Petition for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Feb. 1, 2019.

IPR2019-00637, Petition for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Feb. 1, 2019.

IPR2019-00652, Decision Granting Inter Partes Review U.S. Pat. No. 9,834,822, issued Aug. 19, 2019.

IPR2019-00652, Petition for Inter Partes Review of U.S. Pat. No. 9,834,822, dated Feb. 1, 2019.

IPR2019-00653, Decision Denying Inter Partes Review U.S. Pat. No. 9,834,822, issued Aug. 19, 2019.

IPR2019-00653, Petition for Inter Partes Review of U.S. Pat. No. 9,834,822, dated Feb. 1, 2019.

Jabara, C. Capturing the cloud: High throughput sequencing of multiple individual genomes from a retroviral population. Biology Lunch Bunch Series, Training Initiatives in Biomedical & Biological Sciences of the University of North Carolina at Chapel Hill, Apr. 23, 2010.

Jabara, et al. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. (Paper # 665), The 18th Annual Conference on Retroviruses and Opportunistic Infections, Boston, Massachusetts, Mar. 2011.

Jabara, et al. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl Acad Sci U S A. Dec. 13, 2011;108(50):20166-71. doi: 10.1073/pnas.1110064108. Epub Nov. 30, 2011.

Jahr, et al. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. Cancer Res. Feb. 15, 2001;61(4):1659-65.

Jeronimo, et al. Quantitative GSTP1 hypermethylation in bodily fluids of patients with prostate cancer. Urology. Dec. 2002;60(6):1131-5.

Jiang, H. et al., "SeqMap: mapping massive amount of oligonucleotides to the genome," Bioinformatics 2008, 24(20):2395-2396.

Kanagawa. Bias and artifacts in multitemplate polymerase chain reactions (PCR), 2003, Journal of Bioscience and Bioengineering, vol. 96, No. 4, p. 317-323.

Kao,W-C. et al., "BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing," Genome Res. 2009 19(10), 1884-1895.

Karow, J. "Hopkins Team Develops Method to Improve Rare Mutation Detection for Early Cancer Dx" Genome Web (2011) 3 pages.

Kennedy, S.R. et al., "Detecting ultralow-frequency mutations by Duplex Sequencing" Nature Protocols (2014) 9(11):2586-2606.

Kennedy, S.R. et al., "Ultra-Sensitive Sequencing Reveals an Age-Related Increase in Somatic Mitochondrial Mutations That Are Inconsistent with Oxidative Damage" PLoS Genetics (2013) 9:e1003794.

Kennedy, S.R. et al., "Detecting ultralow-frequency mutations by Duplex Sequencing," Nat. Protoc. 2014, 9(11), 2586-2606.

Kimura, et al. EGFR mutation status in tumour-derived DNA from pleural effusion fluid is a practical basis for predicting the response to gefitinib. Br J Cancer Nov. 20, 2006;95(10):1390-5. Epub Oct. 24, 2006.

Kinde, et al. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. Jun. 7, 2011;108(23):9530-5. doi: 10.1073/pnas.1105422108. Epub May 17, 2011.

Kinde, et al. Supplemental Information, Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. Jun. 7, 2011;108(23):1-10.

Kinde, I. et al., "Evaluation of DNA from the Papanicolaou Test to Detect Ovarian and Endometrial Cancers," Sci. Transl. Med. 2013, 5(167):167ra4.

Kinde, I. et al., "FAST-SeqS: a simple and efficient method for the detection of aneuploidy by massively parallel sequencing," PLoS One 2012, 7(7), e41162.

Kircher, M. et al., "Improved base calling for the Illumina Genome Analyzer using machine learning strategies," Genome Biol. 2009, 10(8), R83.

Kirsch, S. et al. "Sequence error storms and the landscape of mutations in cancer," Proc. Natl. Acad. Sci. USA 2012, 109(36), 14289-14290.

Kivioja, T., et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. Nov. 20, 2011;9(1):72-4. doi: 10.1038/nmeth.1778.

(56) **References Cited**

OTHER PUBLICATIONS

Koboldt, D.C.et al., "The next-generation sequencing revolution and its impact on genomics," Cell 2013, 155(1), 27-38.
Koboldt, et al. Massively parallel sequencing approaches for characterization of structural variation. Aug. 12, 2011. Methods Mol Biol. 2012;838:369-84. doi: 10.1007/978-1-61779-507-7_18.
Kolble, et al. Microsatellite alterations in serum DNA of patients with colorectal cancer. Lab Invest. Sep. 1999;79(9):1145-50.
Konig, et al. iCLIP reveals the function of hnRNAP particles in splicing at individual nucleotide resolution, Jul. 2010, Nature Structural & Molecular Biology, 17(7):909-916.
Kopreski, et al. Detection of mutant K-ras DNA in plasma or serum of patients with colorectal cancer. Br J Cancer. 1997;76(10):1293-9.
Korbel, J.O. et al., "Paired-end mapping reveals extensive structural variation in the human genome," Science 2007, 318(5849), 420-426.
"Blood Plasma" Oxford Dictionary of Biochemistry and Molecular Biology 81 (2d ed. 2006).
"LB-246/12—Detection of circulating tumor DNA in early stage cancers," Available at https://www.genomeweb.com/cancer/ngs-error-correction-method-described-aacr-excitement-grows-early-cancer-detection, Accessed on Apr. 26, 2017.
"Cohesive End," Oxford Dictionary of Biochemistry and Molecular Biology 132 (2d ed. 2006).
"Control," Oxford Dictionary of Biochemistry and Molecular Biology (2d ed. 2006), 143.
Ajay, S.S. et al., "Accurate and comprehensive sequencing of personal genomes," Genome Res. 2011, 21(9), 1498-1505.
Albert, T.J. et al., "Direct selection of human genomic loci by microarray hybridization," Nat. Methods 2007, 4(11), 903-905.
Alkan, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. Oct. 2009;41(10):1061-7. doi: 10.1038/ng.437. Epub Aug. 30, 2009.
Andersson, et al. Bayesian detection of periodic mRNA time profiles without use of training examples. BMC Bioinformatics. Feb. 9, 2006;7:63.
Angeloni, D. Molecular analysis of deletions in human chromosome 3p21 and the role of resident cancer genes in disease. Brief Funct Genomic Proteomic. Mar. 2007;6(1):19-39. Epub May 24, 2007.
Ansorge "Next-generation DNA sequencing techniques." New Biotechnology, 25(4):195-203 (2009).
Appel, Maryke. Part II: It's all about conversion. Kapa Biosystems. Accessed Mar. 29, 2017. Printed Apr. 11, 2017. 5 pages. URL:https://www.kapabiosystems.com/ngs/part-ii-conversion/.
Arneson, N. et al., "Whole-Genome Amplification by Adaptor-Ligation PCR of Randomly Sheared Genomic DNA (PRSG)," CSH Protocols, 2008, 3(1).
Ashford, Monica. NGS Error Correction Method Described at AACR as Excitement Grows for Early Cancer Detection. Genomeweb.com. Apr. 5, 2017. 3 pages. URL:https://www.genomeweb.com/cancerings-error-correction-method-described-aacr-excitement-grows-early-cancer-detection.
Atanur, et al. The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. Genome Res. Jun. 2010;20(6):791-803. doi: 10.1101/gr.103499.109. Epub Apr. 29, 2010.
Audic, et al. "The Significance of Digital Gene Expression Profiles." Genome Research, 7: 986-995 (1997).
Barzon, et al. Evaluation of circulating thyroid-specific transcripts as markers of thyroid cancer relapse. Int J Cancer. Jul. 20, 2004;110(6):914-20.
Bendich, et al. Circulating DNA as a possible factor in oncogenesis. Science. Apr. 16, 1965;148(3668):374-6.
Bentley, D.R. et al., "Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry," Nature, 2008, 456(7218):53-59.
Bettegowda et al., "Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies," Sci. Transl. Med. 2014, 6(224):224ra24.
Bianchi, D.W. et al. "Genome-Wide Fetal Aneuploidy Detection by Maternal Plasma DNA Sequencing," Obstetrics & Gynecology, (2012) 119(5), 1-12.
Bonaldo, et al. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. Sep. 1996;6(9):791-806.
Bowcock, et al. Exclusion of the retinoblastoma gene and chromosome 13q as the site of a primary lesion for human breast cancer. Am J Hum Genet. Jan. 1990;46(1):12-7.
Braha, et al. Simultaneous stochastic sensing of divalent metal ions. Nature Biotechnology, 18: 1005-1007 (2000).
Bremnes, et al. Circulating tumour-derived DNA and RNA markers in blood: a tool for early detection, diagnostics, and follow-up? Lung Cancer. Jul. 2005;49(1):1-12.
Brenner, et al. "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays." Nature Biotechnology, 18: 630-634 (2000).
Brenner, et al. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. Proc Natl Acad Sci U S A. Feb. 15, 2000;97(4):1665-70.
Bryzgunova,O. et al., "A reliable method to concentrate circulating DNA," Analytical Biochem., 2010, 408:354-356.
Campbell, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. Jun. 2008;40(6):722-9. doi: 10.1038/ng.128. Epub Apr. 27, 2008.
Caramazza, et al. Chromosome 1 abnormalities in myeloid malignancies: a literature survey and karyotype-phenotype associations. Eur J Haematol. Mar. 2010;84(3):191-200. doi: 10.1111/j.1600-0609.2009.01392.x. Epub Nov. 30, 2009.
Carr, et al. Inferring relative proportions of DNA variants from sequencing electropherograms. Bioinformatics. Dec. 15, 2009;25(24):3244-50. doi: 10.1093/bioinformatics/btp583. Epub Oct. 9, 2009.
CASAVA v1.8.2 User Guide http://gensoft.pasteur.fr/docs/casava/1.8.2/CASAVA_1_8_2_UG_15011196C.pdf (Oct. 2011).
Casbon, et al. A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res. Jul. 2011;39(12):e81. doi: 10.1093/nar/gkr217. Epub Apr. 13, 2011.
Castle, et al. DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. BMC Genomics. Apr. 16, 2010;11:244. doi: 10.1186/1471-2164-11-244.
Chang, et al. Detection of allelic imbalance in ascitic supernatant by digital single nucleotide polymorphism analysis. Clin Cancer Res. Aug. 2002;8(8):2580-5.
Chee, et al. "Accessing genetic information with high-density DNA arrays." Science, 274: 610-614 (1996).
Chee. "Enzymatic multiplex DNA sequencing." Nucleic Acids Research, 19(12): 3301-3305 (1991).
Chen, et al. Microsatellite alterations in plasma DNA of small cell lung cancer patients. Nat Med. Sep. 1996;2(9):1033-5.
Chin, et al. A SNP in a let-7 microRNA complementary site in the KRAS 3' untranslated region increases non-small cell lung cancer risk. Cancer Res. Oct. 15, 2008;68(20):8535-40. doi: 10.1158/0008-5472.CAN-08-2129.
Chiu, et al. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. BMJ. Jan. 11, 2011;342:c7401. doi: 10.1136/bmj.c7401.
Chiu, et al. Quantitative analysis of circulating mitochondrial DNA in plasma. Clin Chem. May 2003;49(5):719-26.
Chiu, R.W.K. et al. "Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma" PNAS (2008) 105(51):20458-20463.
Church, et al. "Multiplex DNA sequencing." Science, 240: 185-188 (1988).
Cook, et al. Methylated DNA labels for marking objects. Biotechnol Lett. Jan. 2003;25(1):89-94.
Costello, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. Apr. 1, 2013;41(6):e67. doi: 10.1093/nar/gks1443. Epub Jan. 8, 2013.

(56) **References Cited**

OTHER PUBLICATIONS

Coulet, et al. Detection of plasma tumor DNA in head and neck squamous cell carcinoma by microsatellite typing and p53 mutation analysis. Cancer Res. Feb. 1, 2000;60(3):707-11.

Cox, J. "Bar coding objects with DNA." Analyst. May 2001;126(5):545-7.

Cox. Bar coding objects with DNA. Analyst. May 2001;126(5):545-7.

Craig, et al. Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods. Oct. 2008;5(10):887-93. doi: 10.1038/nmeth.1251. Epub Sep. 14, 2008.

Daines, et al. High-throughput multiplex sequencing to discover copy number variants in *Drosophila*. Genetics. Aug. 2009;182(4):935-41. doi: 10.1534/genetics.109.103218. Epub Jun. 15, 2009.

D'Antoni, et al. Rapid quantitative analysis using a single molecule counting approach. Anal Biochem. May 1, 2006;352(1):97-109. Epub Feb. 10, 2006.

Koyanagi, et al. Association of circulating tumor cells with serum tumor-related methylated DNA in peripheral blood of melanoma patients. Cancer Res. Jun. 15, 2006;66(12):6111-7.

Krimmel, J.D. et al., "Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 nutations in noncancerous tissues," Proc. Natl. Acad. Sci. USA 2016, 113(21), 6005-6010.

Kukita, Y. et al. "High-fidelity target sequencing of individual molecules identified using barcode sequences: de novo detection and absolute quantitation of mutations in plasma cell-free DNA from cancer patients" DNA Research (2015) 22(4):269-277.

Lam, et al. Plasma DNA as a prognostic marker for stroke patients with negative neuroimaging within the first 24 h of symptom onset. Resuscitation. Jan. 2006;68(1):71-8. Epub Dec. 1, 2005.

Lanman, et al., Analytical and Clinical Validation of a Digital Sequencing Panel for Quantitative, Highly Accurate Evaluation of Cell-Free Circulating Tumor DNA PLoS One, Oct. 2015, 10(10), e0140712. doi:10.1371/journal.oone.0140712.

Larson, et al. A single molecule view of of gene expression. Trends Cell Biol. Nov. 2009;19(11):630-7. Epub Oct. 8, 2009.

Leary, et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Sci Transl Med. Nov. 28, 2012;4(162):162ra154. doi: 10.1126/scitranslmed.3004742.

Leary, et al. Development of personalized tumor biomarkers using massively parallel sequencing. Sci Transl Med. Feb. 24, 2010;2(20):20ra14. doi: 10.1126/scitranslmed.3000702.

Lecomte, et al. Detection of free-circulating tumor-associated DNA in plasma of colorectal cancer patients and its association with prognosis. Int J Cancer Aug. 10, 2002;100(5):542-8.

Ledergerber, C. et al., "Base-calling for next-generation sequencing platforms," Brief Bioinform. 2011 12(5), 489-497.

Leon, et al. Free DNA in the serum of cancer patients and the effect of therapy. Cancer Res. Mar. 1977;37(3):646-50.

Leung, et al. Plasma Epstein-Barr viral deoxyribonucleic acid quantitation complements tumor-node-metastasis staging prognostication in nasopharyngeal carcinoma. J Clin Oncol. Dec. 1, 2006;24(34):5414-8.

Li, et al. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. Nov. 2008;18(11):1851-8. doi: 10.1101/gr.078212.108. Epub Aug. 19, 2008.

Li, et al. Structure-independent and quantitative ligation of single-stranged DNA. Anal Biochem. Feb. 15, 2006;349(2):242-6. Epub Nov. 18, 2005.

Li, H. et al "Fast and accurate short read alignment with Burrows—Wheeler transform," Bioinformatics 2009, 25(14), 1754-1760.

Liang, K-C. et al., "Bayesian basecalling for DNA sequence analysis using hidden Markov models," IEE/ACM Trans. Comput. Biol. Bioinform. 2007 4(3), 430-440.

Liao, G.J.W et al., "Targeted massively parallel sequencing of maternal plasma DNA permits efficient and unbiased detection of fetal alleles," Clin. Chem. 2011, 57(1), 92-101.

Lizardi, et al. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. Nat Genet. Jul. 1998;19(3):225-32.

Lo, et al. Quantitative Analysis of Cell-free Epstein-Barr Virus DNA in Plasma of Patients with Nasopharyngeal Carcinoma, Cancer Research, 59(6):1188-1191 (Mar. 1999).

Lo, Y M D et al. "Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis," Am. J. Hum. Genet. 1998, 62(4), 768-775.

Lo, Y.M.D. et al. "Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus," Sci Transl Med (2010) 2(61):1-13.

Lockhart, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology, 14: 1675-1680 (1996).

Lodes, et al. Novel Chimera-Free, High Efficiency Library Preparation for NGS Plafforms.Poster. Presented at the Plant & Animal Genome XX Conference. San Diego, California. Jan. 14-18, 2012. 1 page.

Lucito, et al. Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. Genome Res. Nov. 2000;10(11):1726-36.

Lucito, et al. Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation. Genome Research (2003), 13: 2291-2305.

Lunter, G. et al., "Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads," Genome Res. 2011, 21(6):936-939.

Maamar, et al. Noise in Gene Expression Determines Cell Fate in Bacillus subtilis. Science, 317: 526-529 (2007).

Mahmud, et al. Fast MCMC sampling for hidden Markov Models to determine copy number variations. BMC Bioinformatics. Nov. 2, 2011;12:428. doi: 10.1186/1471-2105-12-428.

Makrigiorgos, et al., A PCR-Based amplification method retaining quantative difference between two complex genomes. Nature Biotech, vol. 20, No. 9, pp. 936-939 (Sep. 2002).

Mamanova, L. et al., "Target-enrichment strategies for next-generation sequencing," Nat. Methods 2010, 7(2), 111-118.

Mandel, et al. Les acides nucleiques du plasma sanguin chez l'homme. C R Seances Soc Biol Fil. Feb. 1948;142(3-4):241-3.

Marsit, et al. Epigenetic profiling reveals etiologically distinct patterns of DNA methylation in head and neck squamous cell carcinoma. Carcinogenesis. Mar. 2009;30(3):416-22. doi: 10.1093/carcin/bgp006. Epub Jan. 6, 2009.

McCloskey, et al. Encoding PCR products with batch-stamps and barcodes. Biochem Genet. Dec. 2007;45(11-12):761-7. Epub Oct. 23, 2007.

McKernon, K.J. et al., "Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding," Genome Res. 2009, 19(9), 1527-1541.

Medvedev, et al. Detecting copy number variation with mated short reads. Genome Res. Nov. 2010;20(11):1613-22. doi 10.1101/gr.106344.110. Epub Aug. 30, 2010.

Mei, et al. Identification of recurrent regions of Copy-Number Variants across multiple individuals. BMC Bioinformatics. Mar. 22, 2010;11:147. doi: 10.1186/1471-2105-11-147.

Meldrum et al. Next generation sequencing for cancer diagnostics: a practical perspective. Clin Biochem Rev. Nov. 2011;32(4):177-95.

Mertes, F. et al. "Targeted enrichment of genomic DNA regions for next-generation sequencing" Brief Functional Genomics (2011) 10(6):374-386.

Metzker, M.L. "Sequencing technologies—the next generation" Nature Reviews Genetics (2010) 11:31-46.

Meyer, M. et al. "Illumina sequencing library preparation for highly multiplexed target capture and sequencing," Cold Spring Harb. Protoc. 2010, (6), prot5448.

Meyerson, M. et al. "Advances in understanding cancer genomes through second-generation sequencing" Nature Reviews Genetics (2010) 11:685-696.

Milbury, et al. "Enabling Clinical Cancer Genomics for Rare Mutations: COLD-PCR Magnifies Mutations Prior to Targeted

(56) **References Cited**

OTHER PUBLICATIONS

Amplicon Re-Sequencing". Clin Chem. Mar. 2012;58(3):580-9. doi: 10.1373/clinchem.2011.176198. Epub Dec. 21, 2011.
Miner, et al. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. Nucleic Acids Res. Sep. 30, 2004;32(17):e135.
Mitchell, et al. Circulating microRNAS as stable blood-based markers for cancer detection. Proc Natl Acad Sci U S A. Jul. 29, 2008;105(30):10513-8. Epub Jul. 28, 2008.
Mitelman, F. et al. "The impact of translocations and gene fusions on cancer causation" Nature Rev. Cancer (2007) 7(4)233-245.
Mohan, et al.,Changes in colorectal carcinoma genomes under anti-EGFR therapy identified by whole-genome plasma DNA sequencing.,PLoS Genet,doi: 10.1371/journal.pgen.1004271. eCollection 2014.,Mar. 27, 2014,10(3), e1004271.
Mori, et al. Predictive utility of circulating methylated DNA in serum of melanoma patients receiving biochemotherapy. J Clin Oncol. Dec. 20, 2005;23(36):9351-8.
Morrissy, A.S. et al. "Next-generation tag sequencing for cancer gene expression profiling" Genome Research (2009) 19(10): 1825-1835.
Mortazavi, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. 2008; 5, 621-628.
Narayan, et al. Ultrasensitive measurement of hotspot mutations in tumor DNA in blood using error-suppressed multiplexed deep sequencing. Cancer Res. Jul. 15, 2012;72(14):3492-8. doi: 10.1158/0008-5472.CAN-11-4037. Epub May 10, 2012.
Nawroz, et al. Microsatellite alterations in serum DNA of head and neck cancer patients. Nat Med. Sep. 1996;2(9):1035-7.
Newman, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. May 2014;20(5):548-54. doi: 10.1038/nm.3519. Epub Apr. 6, 2014.
Ng, S.B., et al., "Targeted capture and massively parallel sequencing of 12 human exomes," Nature 2009, 461(7261), 272-276.
Nielsen, R. et al. "Genotype and SNP calling from next-generation sequencing data" Nature Reviews Genetics (2011)12(6):443-451.
Nord, et al. Accurate and exact CNV identification from targeted high-throughput sequence data. BMC Genomics. Apr. 12, 2011;12:184.
Notice of Allowance dated Jan. 31, 2017 for U.S. Appl. No. 14/712,754.
Notice of allowance dated Mar. 21, 2014 for U.S. Appl. No. 12/969,581.
Notice of allowance dated Jun. 15, 2017 for U.S. Appl. No. 15/076,565.
Notice of allowance dated Jun. 19, 2014 for U.S. Appl. No. 12/969,581.
Notice of allowance dated Aug. 1, 2017 for U.S. Appl. No. 15/492,659.
Notice of allowance dated Aug. 4, 2017 for U.S. Appl. No. 15/467,570.
Notice of allowance dated Aug. 22, 2014 for U.S. Appl. No. 12/969,581.
Notice of allowance dated Aug. 29, 2017 for U.S. Appl. No. 15/492,659.
Notice of allowance dated Sep. 11, 2017 for U.S. Appl. No. 15/076,565.
Notice of allowance dated Oct. 3, 2017 for U.S. Appl. No. 15/076,565.
Notice of allowance dated Oct. 25, 2017 for U.S. Appl. No. 14/861,989.
Notice of allowance dated Dec. 28, 2017 for U.S. Appl. No. 14/861,989.
Office Action dated Feb. 9, 2017 for U.S. Appl. No. 15/076,565.
Office action dated May 13, 2019 for U.S. Appl. No. 15/669,779.
Office action dated May 20, 2016 for U.S. Appl. No. 14/855,301.
Office action dated May 31, 2016 for U.S. Appl. No. 14/712,754.
Office action dated Jun. 1, 2017 for U.S. Appl. No. 15/467,570.
Office action dated Jun. 3, 2016 for U.S. Appl. No. 14/861,989.
Office action dated Jun. 12, 2017 for U.S. Appl. No. 15/492,659.
Office action dated Jul. 18, 2017 for U.S. Appl. No. 14/861,989.
Office Action dated Jul. 30, 2019 for U.S. Appl. No. 16/283,635.
Office action dated Aug. 6, 2019 for U.S. Appl. No. 14/855,301.
Office action dated Oct. 3, 2013 for U.S. Appl. No. 12/969,581.
Office Action dated Oct. 12, 2016 for U.S. Appl. No. 14/712,754.
Office action dated Oct. 20, 2017 for U.S. Appl. No. 14/425,189.
Office Action dated Nov. 21, 2016 for U.S. Appl. No. 14/855,301.
Office action dated Dec. 4, 2015 for U.S. Appl. No. 14/712,754.
Office action dated Dec. 7, 2017 for U.S. Appl. No. 14/855,301.
Office Action dated Dec. 9, 2016 for U.S. Appl. No. 14/861,989.
Office action dated Dec. 21, 2018 for U.S. Appl. No. 15/978,848.
Ogino, et al. Quantification of PCR bias caused by a single nucleotide polymorphism in SMN gene dosage analysis. J Mol Diagn. Nov. 2002;4(4):185-90.
Opposition Brief dated Aug. 7, 2018 for Japanese Opposition No. 2018-700659 to JP Patent 6275145.
Opposition Decision dated Jan. 15, 2019 for Japanese Opposition No. 2018-700659 to JP Patent 6275145.
Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Foundation Medicine Inc.
Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Grunecker Patent-und Rechtsanwalte PartG mbB.
Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Personal Genome Diagnostics, Inc.
Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Strawman Limited.
Opposition Form and Statement to EP3087204 filed Nov. 14, 2018.
Pacific Biosciences. Template Preparation and Sequencing Guide. Publication date: Oct. 14, 2014. Pacific Biosciences website http://www.pacificbiosciences.com/support/pubmap/documentation.html.
Pan, et al. Loss of heterozygosity patterns provide fingerprints for genetic heterogeneity in multistep cancer progression of tobacco smoke-induced non-small cell lung cancer. Cancer Res. Mar. 1, 2005;65(5):1664-9.
Park, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet. May 2010;42(5):400-5. doi: 10.1038/ng.555. Epub Apr. 4, 2010.
Parkinson, N.J. et al., "Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA," Genome Res. 2012, 22(1), 125-133.
Pel, J. et al. "Duplex proximity sequencing (pro-seq): A method to improve DNA sequencing accuracy without the cost of molecular barcoding redundancy" bioRxiv (2017) https://doi.org/10.1101/163444.
Perakis, S. et al. "Advances in Circulating Tumor DNA Analysis" Adv Clin Chem (2017) pp. 1-81.
Pihlak, et al. Rapid genome sequencing with short universal tiling probes. Nature Biotechnology, 26: 676-684 (2008).
Pinkel, et al. Comparative Genomic Hybridization. Annual Review of Genomics and Human Genetics, 6: 331-354 (2005).
Pleasance, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. Jan. 14, 2010;463(7278)184-90. doi: 10.1038/nature08629. Epub Dec. 16, 2009.
Pollack, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet. Sep. 1999;23(1):41-6.
Qiu et al. DNA sequence-based "bar codes" for tracking the origins of expressed sequence tags from a maize cDNA library constructed using multiple mRNA sources. Plant Physiol. Oct. 2003;133(2):475-81.
Quail, et al. A large genome center's improvements to the Illumina sequencing system. Nat Methods. Dec. 2008;5(12):1005-10. doi: 10.1038/nmeth.1270.
Quinlan, A.R. et al., "Pyrobayes: an improved base caller for SNP discovery in pyrosequences," Nat. Methods 2008 5(2), 179-181.
Rafi, I. et al. "Cell-free fetal DNA and non-invasive prenatal diagnosis," Br. J. Gen. Pract. May 1, 2009; 59(562):e146-e148.
Rand, et al. Headloop suppression PCR and its application to selective amplification of methylated DNA sequences. Nucleic Acids Res. Aug. 9, 2005;33(14):e127.
Redon, R. et al., "Global variation in copy number in the human genome," Nature 2006 444(7118), 444-454.
Report and Recommendation in Civil Action No. 17-1623-LPS-CJB between Guardant Health, Inc. and Foundation Medicine and between Guardant Health, Inc. and Personal Genome Diagnostics, Inc., dated Oct. 11, 2019.

(56) **References Cited**

OTHER PUBLICATIONS

Rizzo, J.M. et al. "Key Principles and Clinical Applications of 'Next Generation' DNA Sequencing," Cancer Prev. Res., (2012) 5, 887-900.

Ryan, et al. A prospective study of circulating mutant KRAS2 in the serum of patients with colorectal neoplasia: strong prognostic indicator in postoperative follow up. Gut. Jan. 2003;52(1):101-8.

Rygaard, et al. Abnormalities in structure and expression of the retinoblastoma gene in small cell lung cancer cell lines and xenografts in nude mice. Cancer Res. Sep. 1, 1990;50(17):5312-7.

Sanger, F. et al. "DNA sequencing with chain-terminating inhibitors" PNAS (1977) 74(12):5463-5467.

Sausen, M. et al. "Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma", Nature Genetics 2013, 45(1), 12-17.

Schmitt et al. Supplemental Information http://www.pnas.org/content/suppl/2012/08/01/1208715109. DCSupplemental.

Schmitt, et al. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A. Sep. 4, 2012;109(36):14508-13. doi: 10.1073/pnas.1208715109. Epub Aug. 1, 2012.

Schwarzenbach, et al. A critical evaluation of loss of heterozygosity detected in tumor tissues, blood serum and bone marrow plasma from patients with breast cancer. Breast Cancer Res. 2007;9(5):R66.

Schwarzenbach, et al. Cell-free tumor DNA in blood plasma as a marker for circulating tumor cells in prostate cancer. Clin Cancer Res. Feb. 1, 2009;15(3):1032-8. doi: 10.1158/1078-0432.CCR-08-1910.

Schwarzenback, H. et al. "Cell-free nucleic acids as biomarkers in cancer patients" Nature Reviews Cancer (2011)11:426-437.

Schweiger et al., "Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis," PLoS One 2009, 4(5), e5548.

Sebat, et al. Large-Scale Copy Number Polymorphism in the Human Genome. Science, 305: 525-528 (2004).

Sehnert, A.J. et al. "Optimal Detection of Fetal Chromosomal Abnormalities by Massively Parallel DNA Sequencing of Cell-Free Fetal DNA from Maternal Blood" Clin Chem (2011) 57(7):1042-1049.

Shaw et al. Genomic analysis of circulating cell-free DNA infers breast cancer dormancy. Genome Research 22(2):220-231 (Feb. 2012).

Shaw, et al. Microsatellite alterations plasma DNA of primary breast cancer patients. Clin Cancer Res. Mar. 2000;6(3):1119-24.

Shendure, J. et al. "Next-generation DNA sequencing," Nat. Biotechnol. 2008 26(10), 1135-1145.

Shinozaki, et al. Utility of circulating B-RAF DNA mutation in serum for monitoring melanoma patients receiving Biochemotherapy. Clin Cancer Res. Apr. 1, 2007;13(7)2068-74.

Shiroguchi, et al. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci U S A. Jan. 24, 2012;109(4):1347-52. doi: 10.1073/onas.1118018109. Epub Jan. 9, 2012.

Shoemaker, et al. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. Nature Genetics, 14: 450-456 (1996).

Simpson, et al. Copy number variant detection in inbred strains from short read sequence data. Bioinformatics. Feb. 15, 2010;26(4):565-7. doi: 10.1093/bioinformatics/btp693. Epub Dec. 18, 2009.

Smith, et al. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Research, 38(13): e142 (2010).

Smith, T.F. et al. "Comparison of Biosequences" Adv App Math (1981) 2:482-489.

Sorenson, et al. Soluble normal and mutated DNA sequences from single-copy genes in human blood. Cancer Epidemiol Biomarkers Prev. Jan.-Feb. 1994;3(1):67-71.

Sparks, et al. Selective analysis of cell-free DNA in maternal blood for evaluation of fetal trisomy. Prenat Diagn. Jan. 2012;32(1):3-9. doi: 10.1002/pd.2922. Epub Jan. 6, 2012.

Stein, et al. "The case for cloud computing in genome informatics", Genome Biol. 2010; 11 (5):207. Epub May 5, 2010.

Steinman. Free DNA in serum and plasma from normal adults. J Clin Invest. Aug. 1975;56(2):512-5.

Stitziel, et al. Computational and statistical approaches to analyzing variants identified byexome sequencing. Genome Biol. Sep. 14, 2011;12(9):227. doi: 10.1186/gb-2011-12-9-227.

Stroun, M, et al, "About the possible orgin and mechanism of circulating DNA apoptosis and active DNA release", Clin Chim Acta, vol. 313, No. 1-2, pp. 139-142, (2001).

Stumm, M. et al. "Noninvasive prenatal detection of chromosomal aneuploidies using different next generation sequencing strategies and algorithms" Prenatal Diagnosis (2012) 32:569-577.

Taback, et al. Detection of tumor-specific genetic alterations in bone marrow from early-stage breast cancer patients. Cancer Res. Apr. 15, 2003;63(8):1884-7.

Tan, et al. Genome-wide comparison of DNA hydroxymethylation in mouse embryonic stem cells and neural progenitor cells by a new comparative hMeDIP-seq method. Nucleic Acids Res. Apr. 2013;41(7):e84. doi: 10.1093/nar/gkt091. Epub Feb. 13, 2013.

Taudien, et al. Haplotyping and copy number estimation of the highly polymorphic human beta-defensin locus on 8p23 by 454 amplicon sequencing. BMC Genomics. Apr. 19, 2010;11:252. doi: 10.118611471-2164-11-252.

Teer, J.K. et al. "Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing" Genome Res (2010) 20(10):1420-1431.

Tie, J. et al., "Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer,"Sci. Transl. Med. 2016, 8(346):346ra92.

Tomaz, et al. Differential methylation as a cause of allele dropout at the imprinted GNAS locus. Genet Test Mol Biomarkers. Aug. 2010;14(4):455-60. doi: 10.1089/gtmb.2010.0029.

Tomlinson, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet. Aug. 2007;39(8):984-8. Epub Jul. 8, 2007.

Tsai, et al. Discovery of rare mutations in populations: Tilling by sequencing. Plant Physiol. Jul. 2011;156(3):1257-68. doi: 10.1104/pp. 110.169748. Epub Apr. 29, 2011.

U.S. Appl. No. 14/102,285, filed Dec. 10, 2013.

U.S. Appl. No. 61/613,413 ("Schmitt '413 provisional") (filed Mar. 20, 2012).

U.S. Appl. No. 61/625,319 ("Schmitt '319 provisional") (filed Apr. 17, 2012).

Guardant Health vs. *FMI* Invalidity Exhibit C-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit C-4, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit C-5, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit D-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit D-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit D-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit D-5, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Second Amended Complaint dated Mar. 6, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Suppl Invalidity References dated Feb. 15, 2019 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Supplemental Invalidity Contentions dated Mar. 29, 2019 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *PGDx* Amended Answer to Second Amended Complaint dated Apr. 30, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart A, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart B, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart C, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

(56) **References Cited**

OTHER PUBLICATIONS

Guardant Health vs. *PGDx* Invalidity Claim Chart D, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart E, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart F, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart G, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart H, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart I, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart J, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart K, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart L, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Contentions dated Jun. 25, 2018 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Second Amended Complaint dated Mar. 23, 2018 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Second Updated Invalidity Contentions dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Updated Invalidity Contentions dated Mar. 29, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Gundry, et al. "Direct mutation analysis by high-throughput sequencing: from gremlin to low-abundant, somatic variants" Mutat Res Jan. 3, 2012; 729(1-2):1-15. doi: 10.1016/mrfmmm.2011.10.001. Pub Oct. 12, 2011.
Gundry, et al. Direct, genome-wide assessment of DNA mutations in single cells. Nucleic Acids Res. Mar. 2012;40(5):2032-40. doi: 10.1093/nar/gkr949. Epub Nov. 15, 2011.
Hacia, et al. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nature Genetics, 22: 164-167 (1999).
Hafner, et al. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA eDNA libraries, RNA Sep. 1, 2011 17(9):1697-1712.
Hall, A., Short-Read DNA Sequence Alignmnet with Custom Designed FPGA-based Hardware, Thesis 2010.
Hamady, et al. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methods. Mar. 2008;5(3):235-7. doi: 10.1038/nmeth.1184. Epub Feb. 10, 2008.
Hensel, et al. Simultaneous identification of bacterial virulence genes by negative selection. Science. Jul. 21, 1995;269(5222):400-3.
Hiatt, et al. Parallel, tag-directed assembly of locally derived short sequence reads. Nat Methods. Feb. 2010;7(2)119-22. doi: 10.1038/nmeth.1416. Epub Jan. 17, 2010.
Hiatt, et al. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. Genome Res. May 2013;23(5):843-54. doi: 10.1101/gr.147686.112. Epub Feb. 4, 2013.
Hibi, et al. Molecular detection of genetic alterations in the serum of colorectal cancer patients. Cancer Res. Apr. 1, 1998;58(7):1405-7.
HiSeq 2000 User Guide (2010).
Holdenrieder, et al. Circulating nucleosomes and cytokeratin 19-fragments in patients with colorectal cancer during chemotherapy. Anticancer Res. May-Jun. 2005;25(3A):1795-801.
Holies, et al. A stochastic approach to count RNA molecules using DNA sequencing methods. 2003, Lecture Notes in Computer Science, vol. 2812, pp. 55-62.
Hollas, et al. A stochastic approach to count RNA molecules using DNA sequencing methods. Lecture Notes in Computer Science, 2812: 55-62 (2003).
Hollstein, M. et al. "p53 mutations in human cancers" Science (1991) 253(5015):49-53.
Hoque, et al. Detection of aberrant methylation of four genes in plasma DNA for the detection of breast cancer. J Clin Oncol. Sep. 10, 2006;24(26):4262-9. Epub Aug. 14, 2006.
Howe, et al. Retinoblastoma growth suppressor and a 300-kDa protein appear to regulate cellular DNA synthesis. Proc Natl Acad Sci U S A. Aug. 1990;87(15):5883-7.
Hug, et al. Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. J Theor Biol. Apr. 21, 2003;221(4):615-24.
Hyland, et al. The normal and tumor spectrum of copy number variation: Copy number alterations correlate with changes in gene expression in tumor transcriptome. Nov. 15, 2009. 1 page. Retreived from:http://tools.thermofisher.com/content/sfs/posters/cms_073633.pdf.
Iafrate, et al. Detection of large-scale variation in the human genome. Nat Genet. Sep. 2004;36(9):949-51. Epub Aug. 1, 2004.
Ikeguchi, et al. Detection of loss of heterozygosityat microsatellite loci in esophageal squamous-cell carcinoma. Oncology. 1999;56(2)164-8.
Illumina "Multiplexed Sequencing with Illumina Genome Analyzer System" https://www.illumina.com/Documents/products/datasheets/datasheet_sequencing_multiplex.pdf (Dec. 2008).
Illumina Preparing Samples for Paired-End Sequencing (2008).
Daser, et al. "Interrogation of genomes by molecular copy-number counting (MCC)." Nature Methods, 3(6): 447-453 (2006).
De Saizieu, et al. "Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays." Nature Biotechnology, 16: 45-48 (1998).
Diehl et al. Circulating mutant DNA to assess tumor dynamics. Nat Med 14(9):985-990 (2008).
Diehl, et al. Detection and quantification of mutations in the plasma of patients with colorectal tumors. Proc Natl Acad Sci US A. Nov. 8, 2005;102(45):16368-73. Epub Oct. 28, 2005.
Diehl, F. et al., "Analysis of Mutations in DNA Isolated From Plasma and Stool of Colorectal Cancer Patients," Gastroenterology (2008) 135(2):489-498.
Ding, L. et al. "Clonal evolution in relapsed acute myeloid leukemia revealed by whole-genome sequencing" Nature (2012) 481(7382):506-510.
Ding, L. et al., "Analysis of next-generation genomic data in cancer: accomplishments and challenges," Hum. Mol. Genet. 2010 19(R2), R188-R196.
Duncan, D.L. et al. "Next-Generation Sequencing in the Clinical Laboratory" Diagnostic Molecular Pathology: A Guide to Applied Molecular Testing 25-33 (Coleman and Tsongalis eds., 1st ed. 2016).
Ehrich, et al. Noninvasive detection of fetal trisomy 21 by sequencing of DNA in maternal blood: a study in a clinical setting. Am J Obstet Gynecol. Mar. 2011;204(3):205.e1-11. doi: 10.1016/j.ajog.2010.12.060. Epub Feb. 18, 2011.
Eisenmann, et al. 5q-myelodysplastic syndromes: chromosome 5q genes direct a tumor-suppression network sensing actin dynamics. Oncogene. Oct. 1, 2009;28(39):3429-41. doi: 10.1038/onc.2009.207. Epub Jul. 13, 2009.
Elshire, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. May 4, 2011;6(5):e19379. doi: 10.1371/journal.pone.0019379.
European search report and search opinion dated Mar. 29, 2016 for EP Application No. 13834427.0.
Extended European search report and opinion dated May 4, 2017 for EP Application No. 14874157.2.
Extended European search report and opinion dated Nov. 21, 2016 for EP Application No. 14771159.
Fan, et al. "Microfluidic digital PCR enables rapid prenatal diagnosis of fetal aneuploidy." Am Obstet Gynecol. 2009; 200:543.e1-543.e7.
Fan, et al. "Parallel Genotyping of Human SNPs Using Generic High-density Oligonucleotide Tag Arrays." Genome Research, 10: 853-860 (2000).
Fan, et al. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci U S A. Oct. 21, 2008;105(42):16266-71. Epub Oct. 6, 2008.

(56) **References Cited**

OTHER PUBLICATIONS

Fan, et al. Non-invasive prenatal measurement of the fetal genome. Nature. Jul. 19, 2012;487(7407):320-4. doi: 10.1038/nature 11251.
Fleischhacker, M. et al. "Circulating nucleic acids (CNAs) and cancer—A survey" Biochimica et Biophysica Acta (2007) 1775:181-232.
Fonatsch, C. The role of chromosome 21 in hematology and oncology. Genes Chromosomes Cancer. Jun. 2010;49(6):497-508. doi: 10.1002/gcc.20764.
Fong, S.L. et al. "Comparison of 7 Methods for Extracting Cell-Free DNA from Serum Samples of Colorectal Cancer Patients" Clinical Chem (2009) 55(3):587-598.
Forshew, T. et al. "Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA" Sci Transl Med (2012) 4(136) ra68.
Forster, et al. From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software. Nucleic Acids Res. Jan. 7, 2013;41(1):e16. doi: 10.1093/nar/gks836. Epub Sep. 10, 2012.
Fournie, et al. Plasma DNA as a marker of cancerous cell death. Investigations in patients suffering from lung cancer and in nude mice bearing human tumours. Cancer Lett. May 8, 1995;91(2):221-7.
Freeman, et al. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. Oct. 2009;19(10):1817-24. doi: 10.1101/gr.092924.109. Epub Jun. 18, 2009.
Fu, et al. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc Natl Acad Sci U S A. May 31, 2011;108(22):9026-31. Epub May 11, 2011.
Fujiwara, et al. Identification of epigenetic aberrant promoter methylation in serum DNA is useful for early detection of lung cancer Clin Cancer Res. Feb. 1, 2005;11(3):1219-25.
Fujiwara, et al. Plasma DNA microsatellites as tumor-specific markers and indicators of tumor progression in melanoma patients. Cancer Res. Apr. 1, 1999;59(7):1567-71.
Genome Analyzer IIx Systems Specification Sheet (2009).
Gerry, et al. Universal DNA microarray method for multiplex detection of low abundance point mutations. Journal of Molecular Biology, 292(2): 251-262 (1999).
Gillespie. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry, 81(25): 2340-2361 (1977).
Gordian, et al. Serum free circulating DNA is a useful biomarker to distinguish benign versus malignant prostate disease. Cancer Epidemiol Biomarkers Prev. Aug. 2010;19(8):1984-91. doi: 10.1158/1055-9965.EPI-10-0287. Epub Jul. 20, 2010.
Gordon, D.J. et al., "Causes and consequences of aneuploidy in cancer," Nat. Rev. Genet. 2012 13(3), 189-203.
Gormally, et al. Amount of DNA in plasma and cancer risk: a prospective study. Int J Cancer. Sep. 20, 2004;111(5):746-9.
Grant, et al. SNP genotyping on a genome-wide amplified DOP-PCR template. Nucleic Acids Res. Nov. 15, 2002;30(22):e25.
Gregory et al. "Targeted Single Molecule Mutation Detection with Massively Parallel Sequencing" Nucleic Acids Research (2015) 44(3):2-11.
Grutzmann, et al. Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. PLoS One. 2008;3(11):e3759. doi: 10.1371/journal.pone.0003759. Epub Nov. 19, 2008.
Guardant Health vs. *FMI* 1st Amended Answer to Second Amended Complaint dated May 6, 2019 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Contentions dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit A-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit A-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit A-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit A-8, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit B-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit B-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit B-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit B-7, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit C-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit C-11, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit C-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
UCSC Genome Bioinformatics. About the UCSC Genome Bioinformatics Site. http://genome.ucsc.edu/index.html. Accessed on May 26, 2015. 2 pgs.
Umetani et al. Prediction of Breast Tumor Progression by Integrity of Free Circulating DNA in Serum. Journal of Clinical Oncology 24(26):4270-4276 (Sep. 10, 2006).
U.S. Appl. No. 61/384,001, filed Sep. 17, 2010.
U.S. Appl. No. 61/432,119, filed Jan. 12, 2011.
Utting, et al. Microsatellite analysis of free tumor DNA in urine, serum, and plasma of patients: a minimally invasive method for the detection of bladder cancer. Clin Cancer Res. Jan. 2002;8(1):35-40.
Van Houten, et al. Mutated p53 as a molecular marker for the diagnosis of head and neck cancer. J Pathol. Dec. 2002;198(4):476-86.
Van Orsouw, N. et al. "Complexity Reduction of Polymorphic Sequences (CRoPSTM): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes" PLoS One (2007) 11(e1172):1-10.
Vasyukhin, V. et al., "K-Ras Point Mutations in the Blood Plasma DNA of Patients with Colorectal Tumors" Challenges of Modern Medicine, 141-150 (Verna and Shamoo eds, 1994).
Velculescu, et al. Characterization of the Yeast Transcriptome. Cell, 88: 243-251 (1997).
Velculescu, et al. Serial Analysis of Gene Expression. Science, 270: 484-487 (1995).
Vogelstein, et al. Digital PCR. Proc. Natl. Acad. Sci., 96(16): 9236-9241(1999).
Wagle, N. et al., "High-throughput Detection of actionable Genomic alterations in clinical tumor samples by targeted, Massively Parallel sequencing," Cancer Discov. 2012, (2)1:82-93.
Walker, et al. Isothermal in vitro amplification of DNA by a restriction enzyme/DNA polymerase system. Proc Natl Acad Sci U S A. Jan. 1, 1992;89(1):392-6.
Walsh, et al. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc Natl Acad Sci U S A. Jul. 13, 2010;107(28):12629-33. doi: 10.1073/pnas.1007983107. Epub Jun. 28, 2010.
Wang, et al. iCLIP predicts the dual splicing effects of TIA-RNA interactions, Oct. 2010, PLoS Biol, 8(10):e1000530.
Wang, et al. Molecular detection of APC, K-ras, and p53 mutations in the serum of colorectal cancer patients as circulating biomarkers. World J Surg. Jul. 2004;28(7):721-6. Epub Jun. 8, 2004.
Wang, et al. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10: 57-63 (2009).
Wang, T.L. et al. "Digital Karyotyping" PNAS (2002) 99(25):16156-16161.
Wang, Y. et al., "Detection of somatic mutations and HPV in the saliva and plasma of patients with head and neck squamous cell carcinomas," Sci. Transl. Med. 2015, 7(293):293ra104.
Wang, Y. et al., "Detection of tumor-derived DNA in cerebrospinal fluid of patients with primary tumors of the brain and spinal cord," Proc. Natl. Acad. Sci. USA 2015, 112(31), 9704-9709.
Wang, Y. et al., "Diagnostic potential of tumor DNA from ovarian cyst fluid," eLife 2016, 5:e15175.
Weber, et al. A real-time polymerase chain reaction assay for quantification of allele ratios and correction of amplification bias. Anal Biochem. Sep. 15, 2003;320(2):252-8.

(56) **References Cited**

OTHER PUBLICATIONS

Wheeler, D.A. et al., "The complete genome of an individual by massively parallel DNA sequencing," Nature 2008, 452(7189),872-876.
Williford, A. et al., "Gene Fusion," eLS 2013, 1-8.
Wittes, et al. Searching for Evidence of Altered Gene Expression: a Comment on Statistical Analysis of Microarray Data. Journal of the National Cancer Institute, 91(5): 400-401 (1999).
Wodicka, et al. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nature Biotechnology, 15: 1359-1367 (1997).
Wojdacs, et al. Primer design versus PCR bias in methylation independent PCR amplifications. Epigenetics. May 16, 2009;4(4):231-4. Epub May 14, 2009.
Wood, et al. The genomic landscapes of human breast and colorectal cancers. Science. Nov. 16, 2007;318 (5853)1108-13. Epub Oct. 11, 2007.
Wood, et al. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucleic Acids Res. Aug. 2010;38(14): e151. doi: 10.1093/nar/gkq510. Epub Jun. 4, 2010.
Xi, et al. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc Natl Acad Sci U S A. Nov. 15, 2011;108(46):E1128-36. doi: 10.1073/pnas.1110574108. Epub Nov. 7, 2011.
Yandell, et al. A probabilistic disease-gene finder for personal genomes. Genome Res. Sep. 2011;21(9):1529-42. doi: 10.1101/gr.123158.111. Epub Jun. 23, 2011.
Yang, et al., EGFR gene copy number as a predictive biomarker for the treatment of metastatic colorectal cancer with anti-EGFR monoclonal antibodies: a meta-analysis.,J Hematol Oncol, Aug. 16, 2012,5:52,1-9.
Yang. Simple binary segmentation frameworks for identifying variation in DNA copy number. BMC Bioinformatics. Oct. 30, 2012;13:277. doi: 10.1186/1471-2105-13-277.
Ye, et al. Fluorescent microsphere-based readout technology for multiplexed human single nucleotide polymorphism analysis and bacterial identification. Human Mutation, 17(4): 305-316 (2001).
Yoon, et al. Sensitive and accurate detection of copy numbers variants using read depth of coverage. Genome Res. Sep. 2009;19(9):1586-92. doi: 10.1101/gr.092981.109. Epub Aug. 5, 2009.
Zhang, et al. "Comprehensive One-Step Molecular Analysis of Mitochondrial Genome by Massively Parallel Sequencing" Clinical Chem (2012) 58(9):1322-1331.
Zhang et al. The impact of next-generation sequencing on genomics. J Genet Genomics. Mar. 20, 2011;38(3):95-109. doi:10.1016/.jgg.2011.02.003. Epub Mar. 15, 2011.
Zhao, et al. Homozygous Deletions and Chromosome Amplifications in Human Lung Carcinomas Revealed by Single Nucleotide Polymorphism Array Analysis. Cancer Research, 65: 5561-5570 (2005).
Zhou, et al. Counting alleles reveals a connection between chromosome 18q loss and vascularinvasion. Nature Biotechnology, 19: 78-81 (2001).
Clark, T.A. et al. "Analytical Validation of a Hybrid Capture Based Next-Generation Sequencing Clinical Assay for Genomic Profiling of Cell-Free Circulating Tumor DNA," J. Mol. Diagnostics (2018) 20(5):686-702.
Kamps-Hughes, N. et al. "ERASE-Seq: Leveraging replicate measurements to enhance ultralow frequency variant detection in NGS data"PLoS One (2018).
Lennon, N.J. et al. "Technological considerations for genome-guided diagnosis and management of cancer" Gen Med (2016) 8:112.
Mishra, S. et al. "Different Facets of Copy Number Changes: Permanent, Transient, and Adaptive" Mol Cell Biol (2016) 36(7):1050-1063.
Moench, S. Genomic Profiling Using Guardant 360 Cell-Free DNA-Based Assay vs Tumor-Based Genotyping Assays in Advanced NSCLC, Cancer Therapy Advisor (Feb. 28, 2019), https://www.cancertherapyadvisor.com/home/news/conferencecoverage/american-association-for-cancer-research-aacr/aacr-2019/genomic-profiling-using-guardant-360-cell-free-dna-based assay-vs-tumor-based-genotyping-assays-in-advanced-nsclc/ (lastaccessed Nov. 30, 2019).
Newman, A. et al. "Integrated digital error suppression for improved detection of circulating tumor DNA" Nature Biotech (2016) 34(5):547-555.
Odegaard, J.I. et al. "Validation of a Plasma-Based Comprehensive Cancer Genotyping Assay Utilizing Orthogonal Tissue- and Plasma-Based Methodologies" Clin Canc Res (2018) 24(15):3539-3549.
Ou, Shi et al. "Liquid Biopsy to Identify Actionable Genomic Alterations" Am Soc Clin Onc (2018) 978.
Paweletz, C.P. et al. "Bias-corrected targeted next-generation sequencing for rapid, multiplexed detection of actionable alterations in cell-free DNA from advanced lung cancer patients" Clin Canc Res (2016) 22(4):915-922.
Phallen, J. et al. "Direct detection of early-stage cancers using circulating tumor DNA" Sci Trans Med (2017) vol. 9, Issue 403, eaan2415DOI: 10.1126/scitranslmed.aan2415.
Sathirapongsasuti, J.F. et al. "Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV" BioInformatics (2011) 27(19):2648-2654.
Shiroguchi, et al. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci U S A. 2012 Supplemental Information (8 pages).
Trapnell, C. et al. "How to map billions of short reads onto genomes" Nature Biotech (2009) 27(5):455-457.
Van Loo, P. et al. "Allele-specific copy number analysis of tumors" PNAS (2010) 107(39):16910-16915.
Wang, T.T. et al. "High efficiency error suppression for accurate detection of low-frequency variants" NAR (2019) 47(15):e87.
Guardant Health, Inc. Response to Notices of Opposition in EP2893040 filed May 29, 2020.
IPR2019-00634, Final Written Decision of U.S. Pat. No. 9,840,743, dated Aug. 18, 2020.
IPR2019-00652, Final Written Decision of U.S. Pat. No. 9,834,822, dated Aug. 18, 2020.
EPO Preliminary Opinion, dated Aug. 25, 2020, in EP2893040.
Lo, Dennis Interview with Professor Dennis Lo, Qiagen News (2002).
Opponents Reply dated Aug. 20, 2020 to Proprietor's Observation in EP2893040.
Opposition Form and Statement to EP3470533 filed Aug. 6, 2020 by Foundation Medicine Inc.
U.S. Appl. No. 61/625,623, filed Apr. 17, 2012.
Jiang et al. "Basics in Bioinformations Lecture Notes of the Graduate Summer School on Bioinformatics of China" Springer Heidelburg New York Dordrecht London (2013).
Opposition Form and Statement to EP3378952 filed Nov. 4, 2020 by Dirk Buhler.
Opposition Form and Statement to EP3378952 filed Nov. 5, 2020 by Foundation Medicine Inc.
Opposition Form and Statement to EP3378952 filed Nov. 5, 2020 by Grunecker.

* cited by examiner

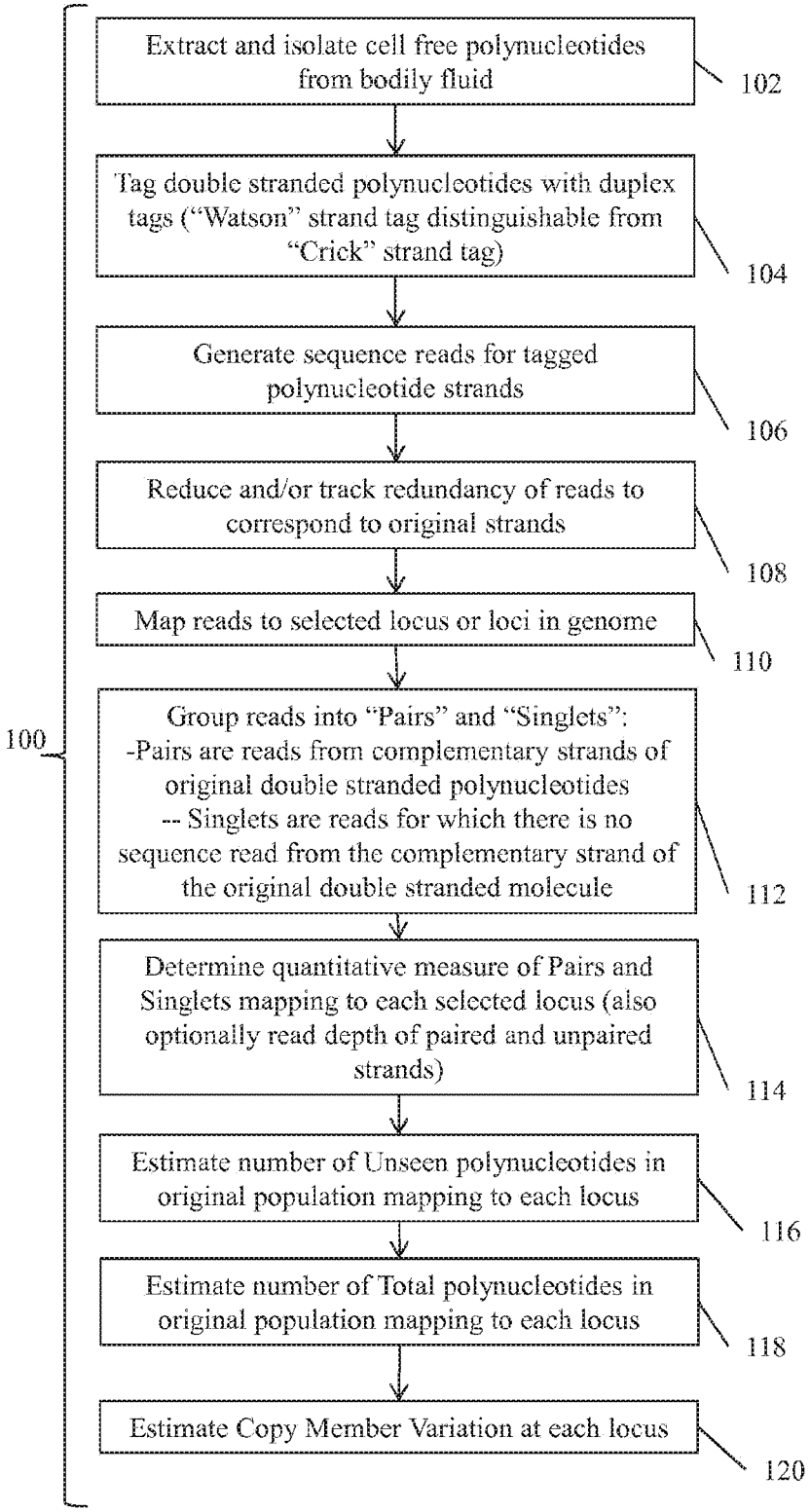

Extract and isolate cell free polynucleotides from bodily fluid
102
Tag double stranded polynucleotides with duplex tags ("Watson" strand tag distinguishable from "Crick" strand tag)
104
Generate sequence reads for tagged polynucleotide strands
106
Reduce and/or track redundancy of reads to correspond to original strands
108
Map reads to selected locus or loci in genome
110
100
Group reads into "Pairs" and "Singlets":
-Pairs are reads from complementary strands of original double stranded polynucleotides
-- Singlets are reads for which there is no sequence read from the complementary strand of the original double stranded molecule
112
Determine quantitative measure of Pairs and Singlets mapping to each selected locus (also optionally read depth of paired and unpaired strands)
114
Estimate number of Unseen polynucleotides in original population mapping to each locus
116
Estimate number of Total polynucleotides in original population mapping to each locus
118
Estimate Copy Member Variation at each locus
120

Fig. 1

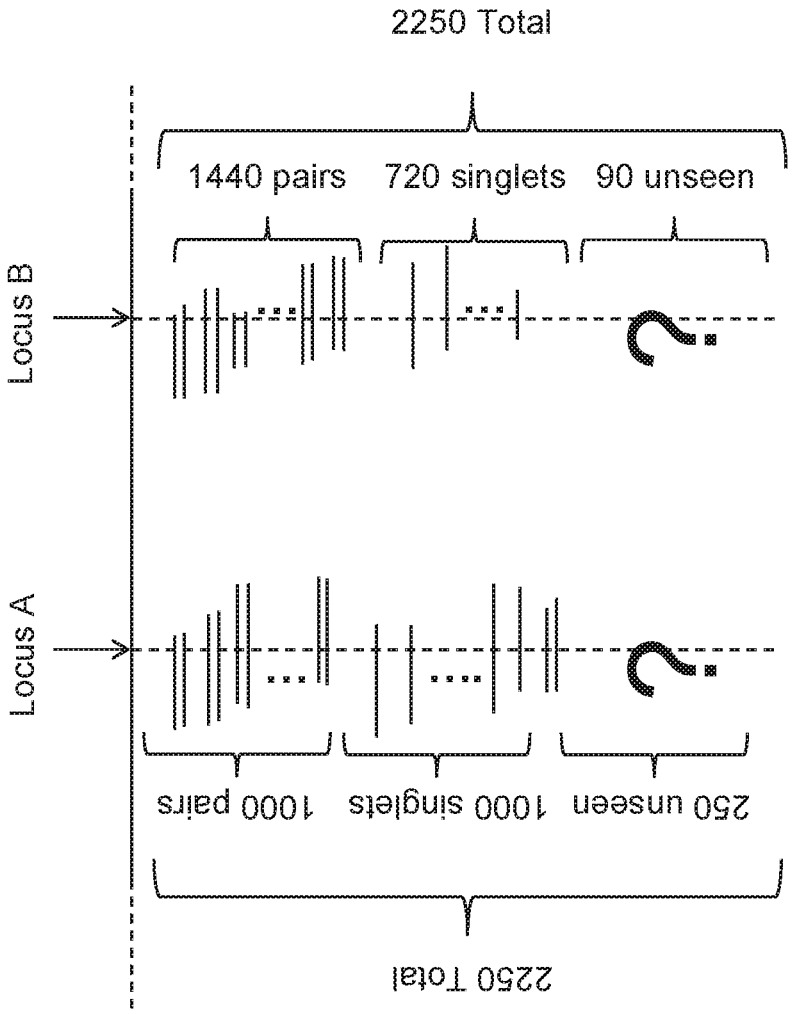

2250 Total
1440 pairs
720 singlets
90 unseen
Locus B
Locus A
1000 pairs
1000 singlets
250 unseen
2250 Total

Fig. 2

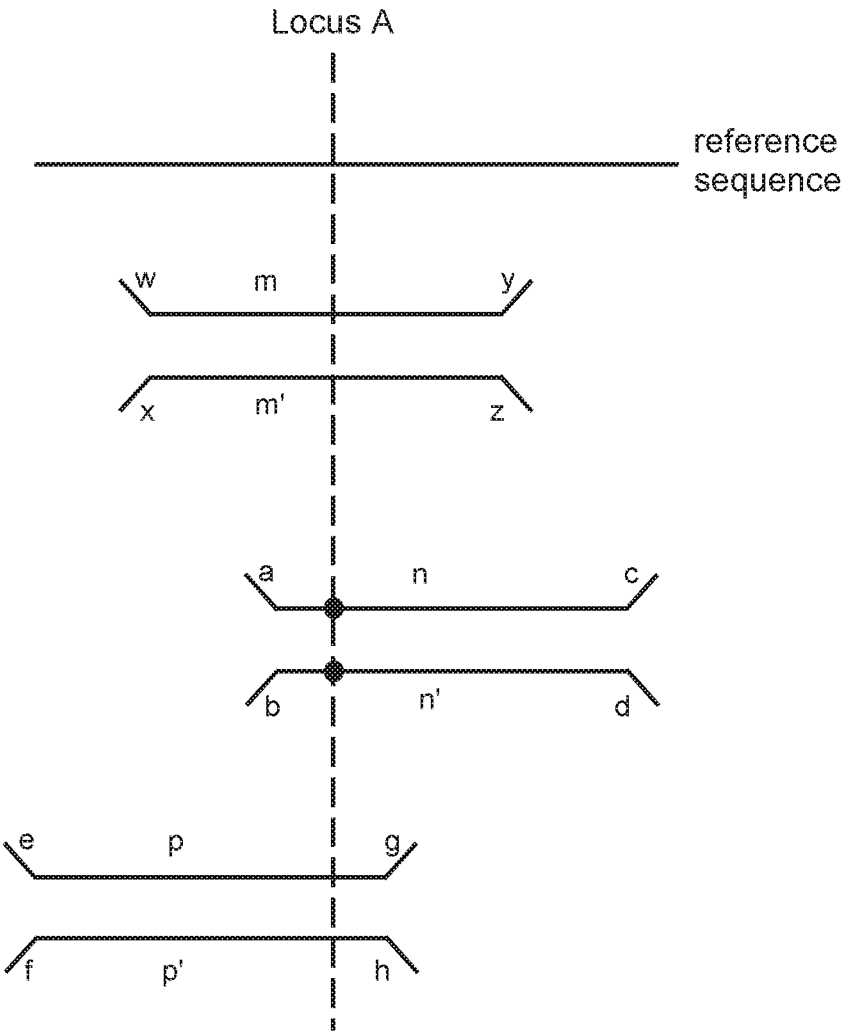


Fig. 3

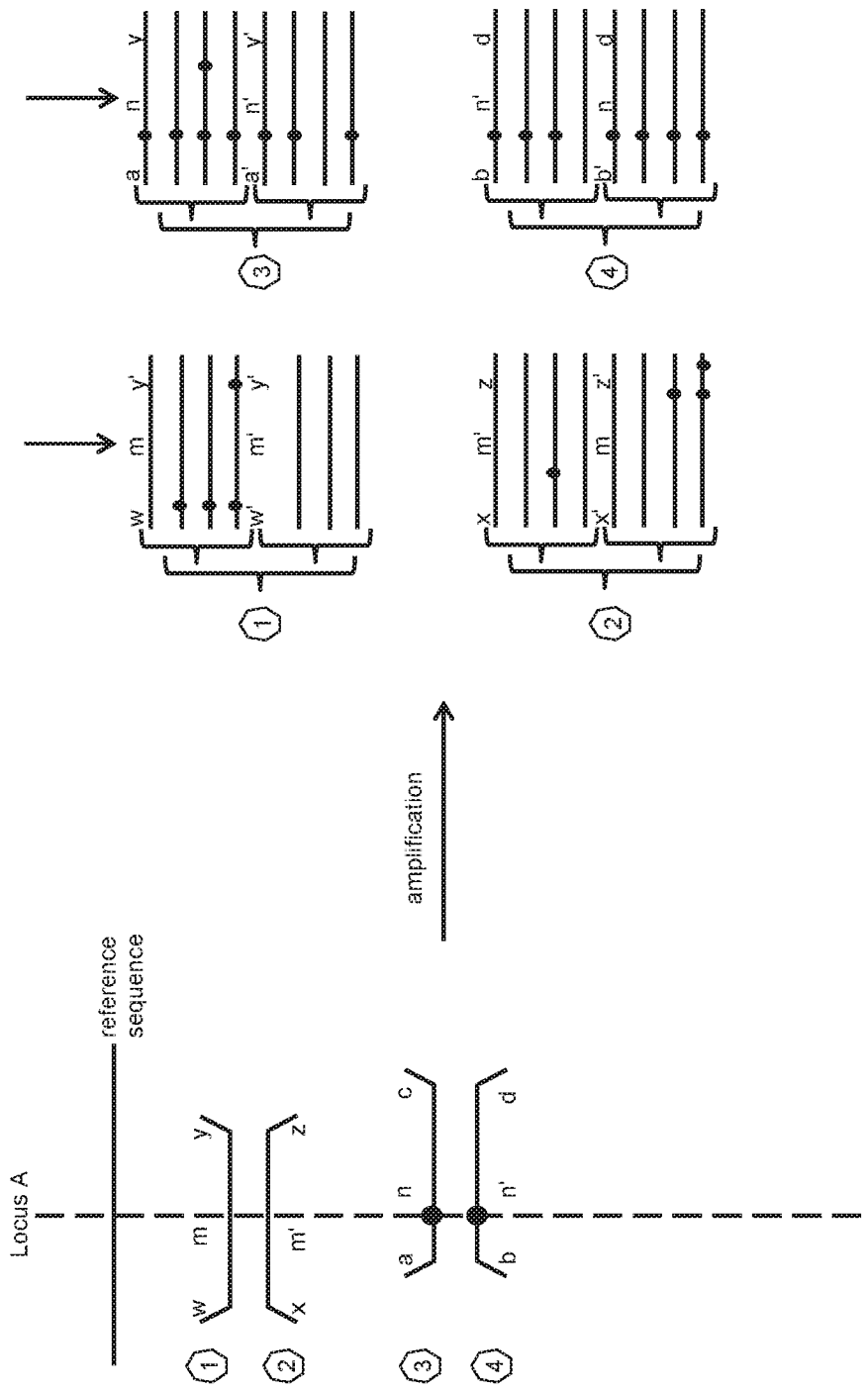

Locus A
reference sequence
amplification
w
m
y
x
m'
z
a
n
c
b
n'
d
1
2
3
4

Fig. 4A

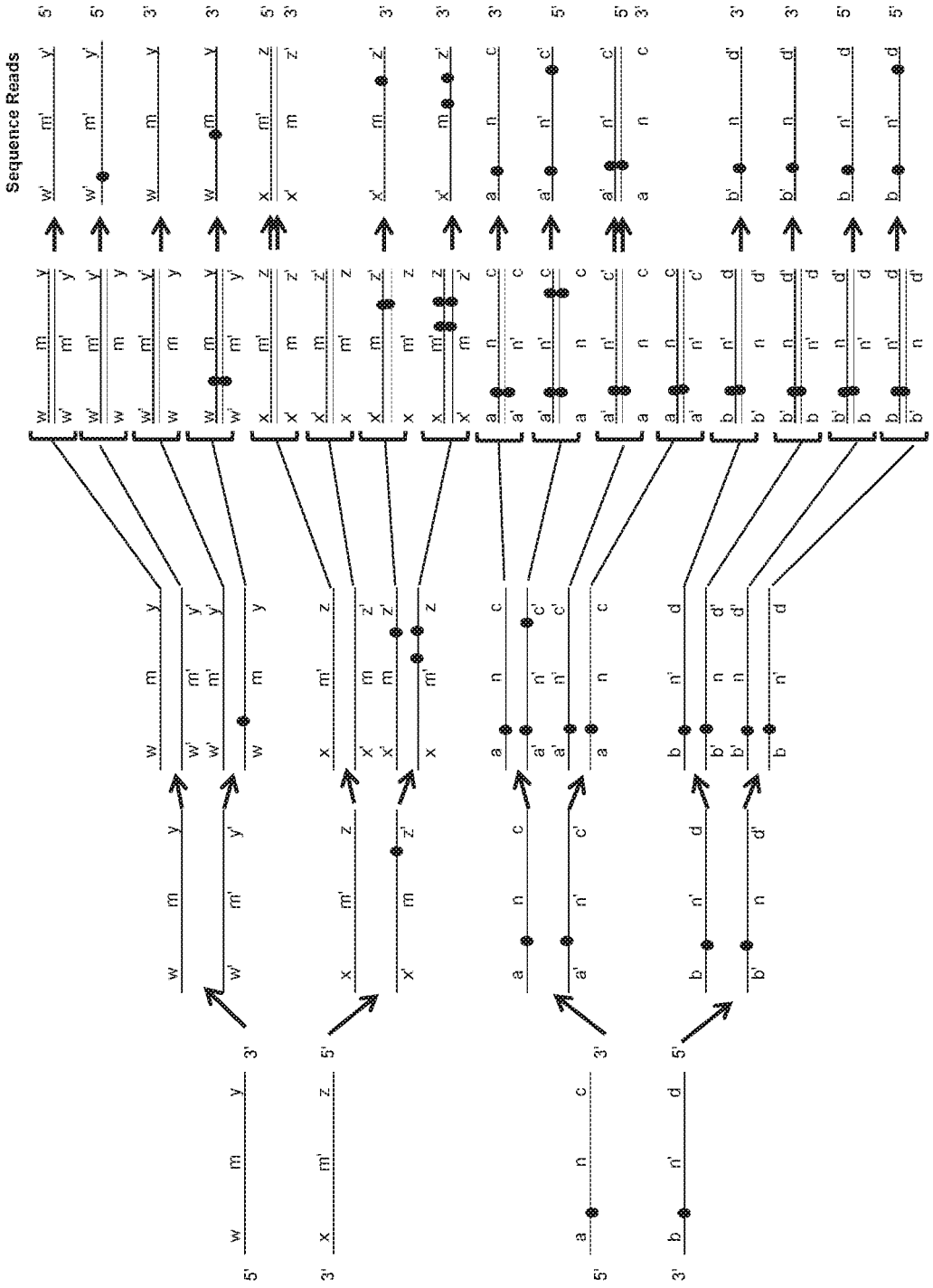


Fig. 4B

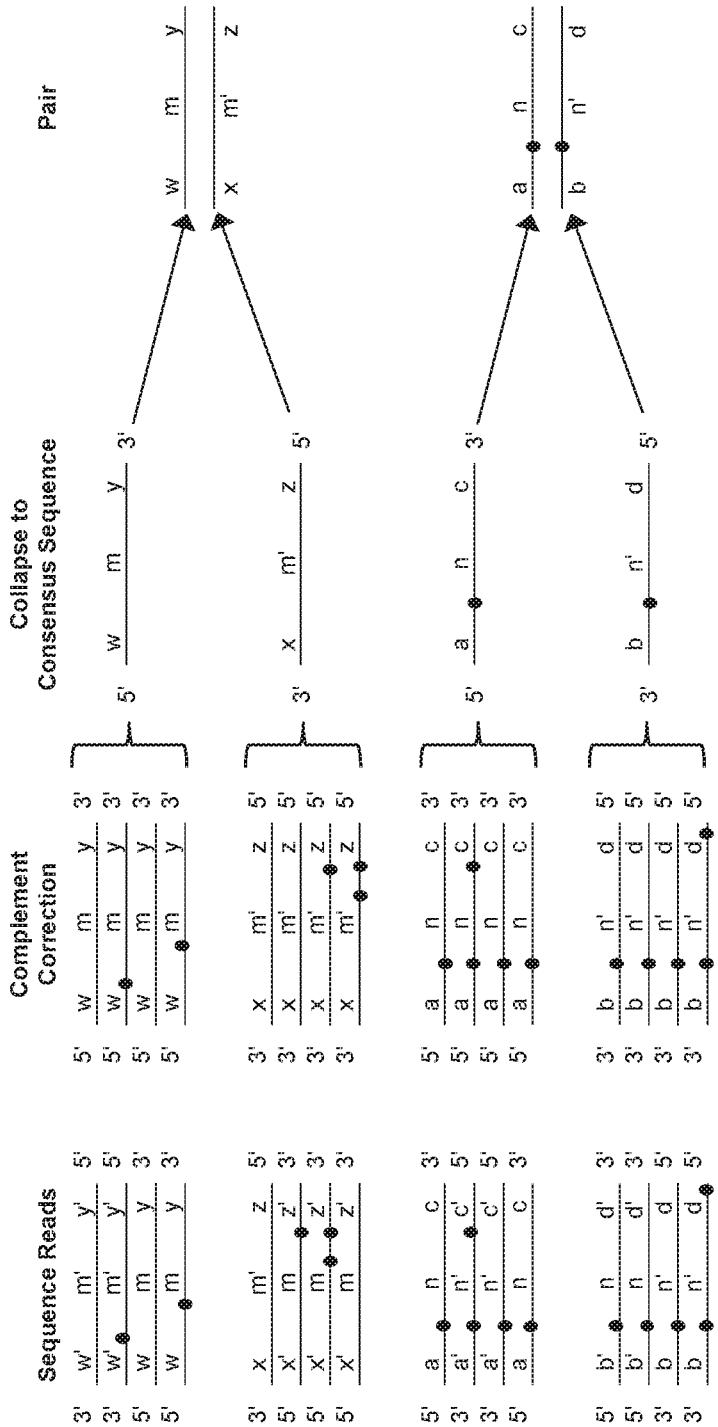

Sequence Reads
Complement Correction
Collapse to Consensus Sequence
Pair

Fig. 4C

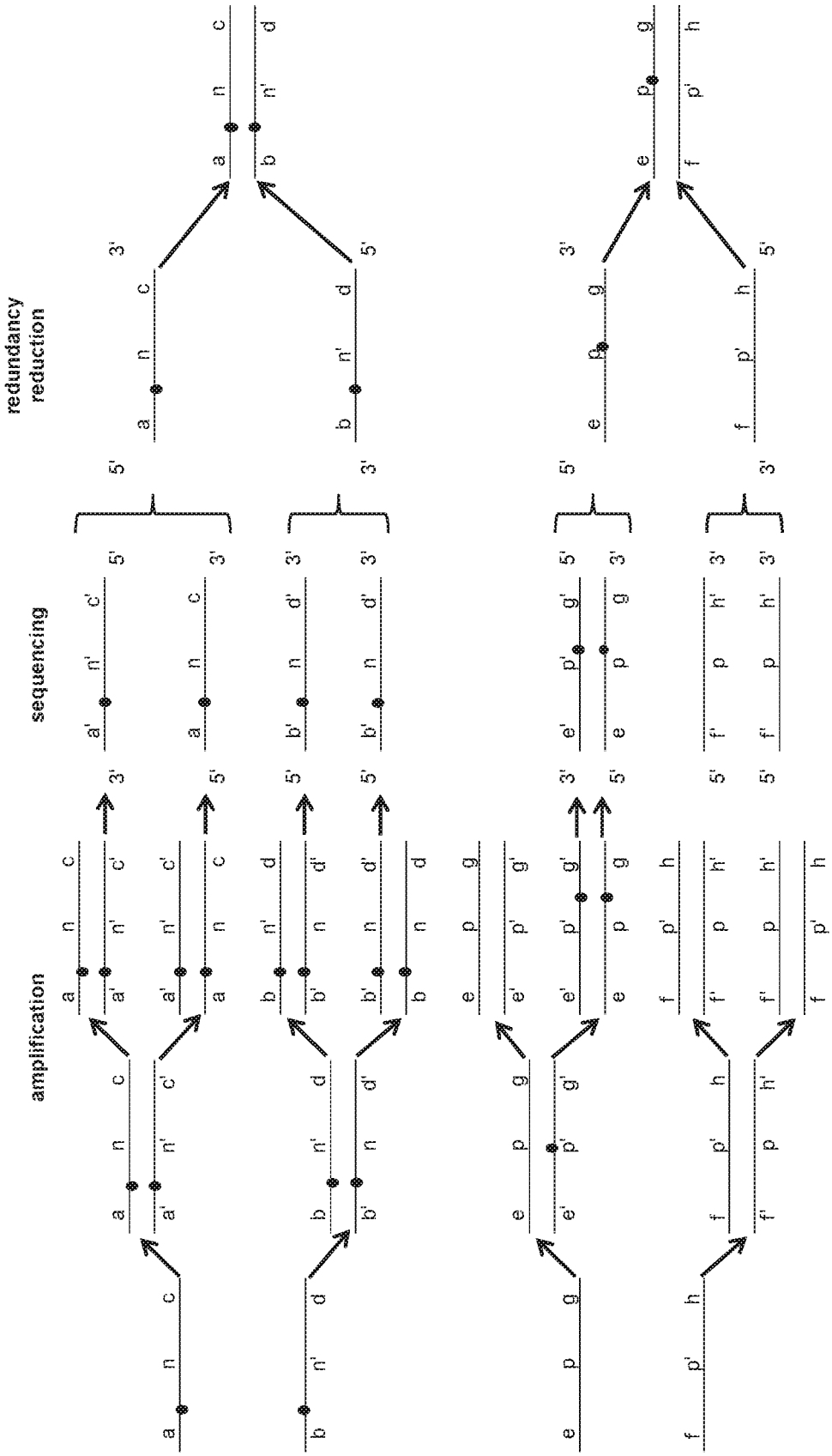


Fig. 5

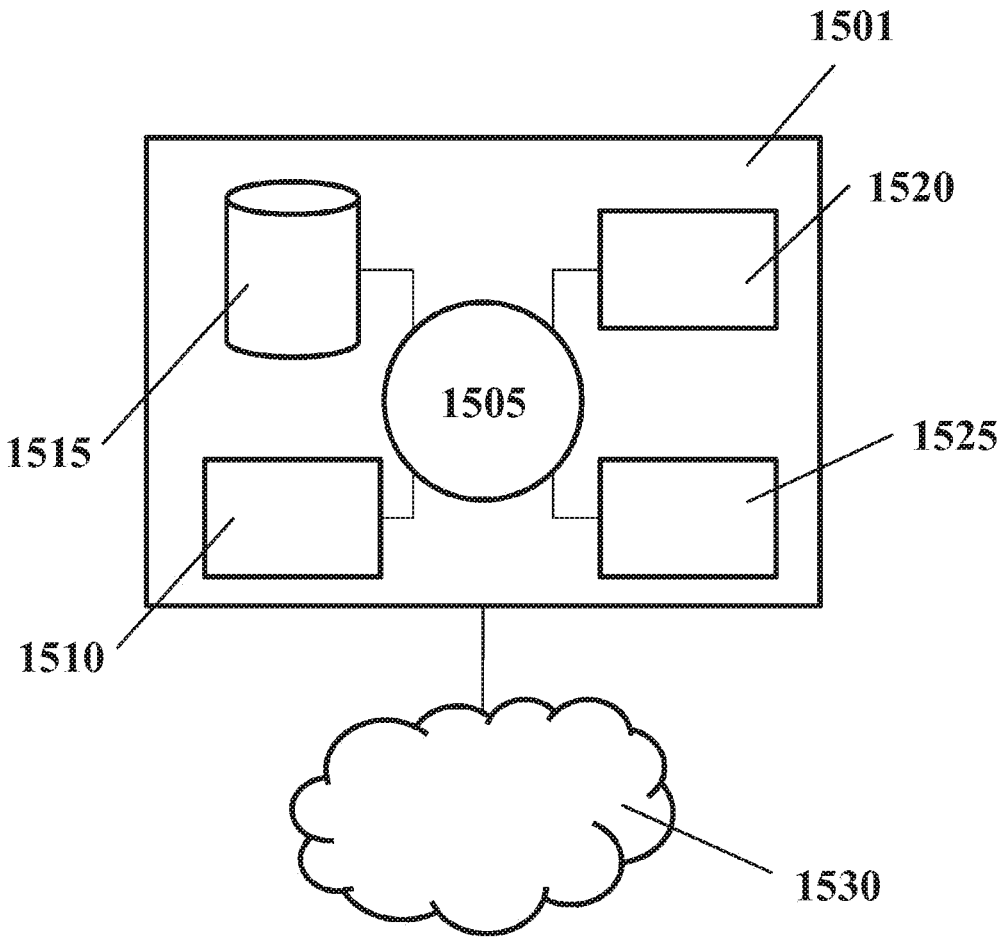


Fig. 6

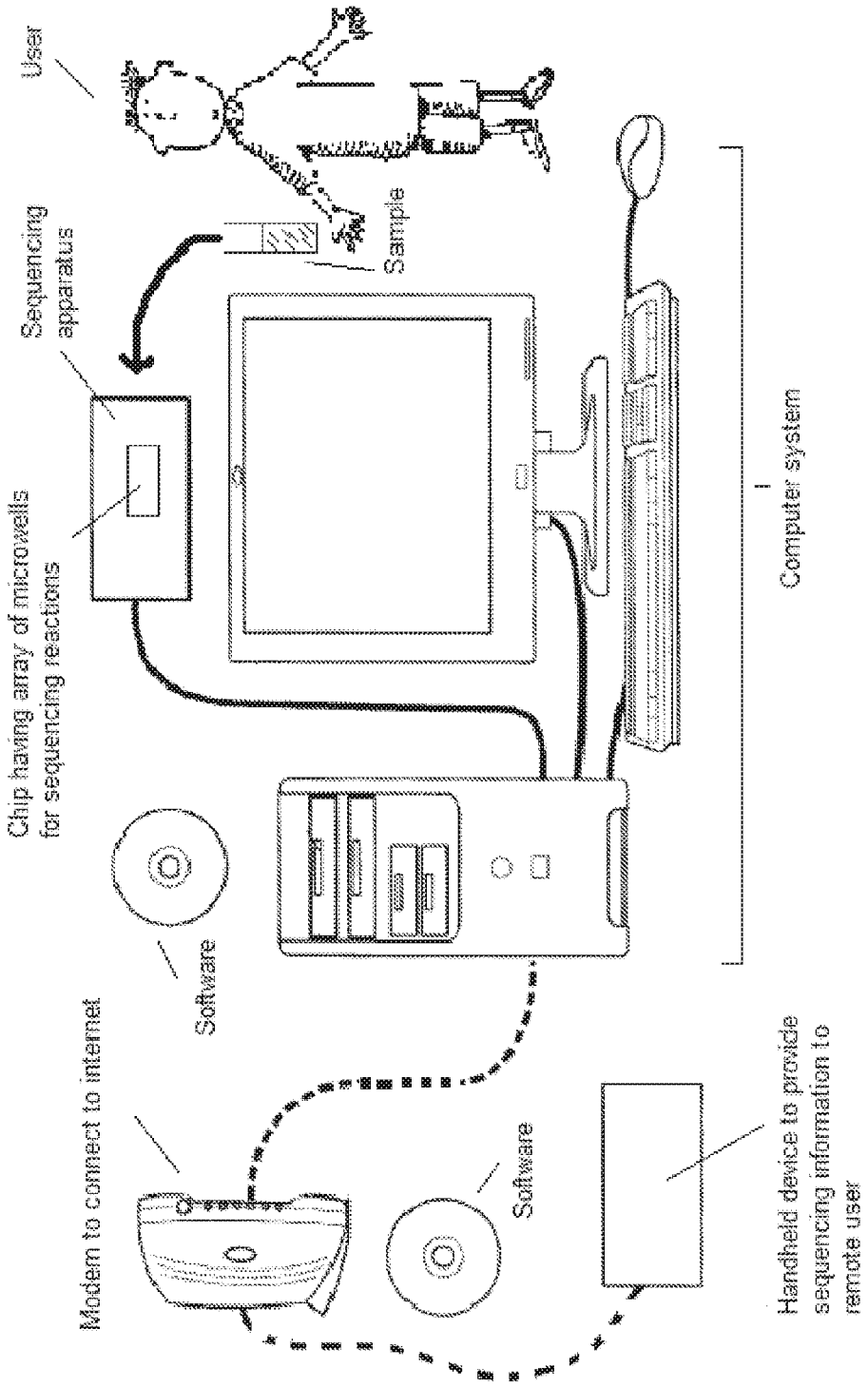

User
Sample
Sequencing apparatus
Chip having array of microwells for sequencing reactions
Computer system
Software
Modem to connect to internet
Software
Handheld device to provide sequencing information to remote user

Fig. 7

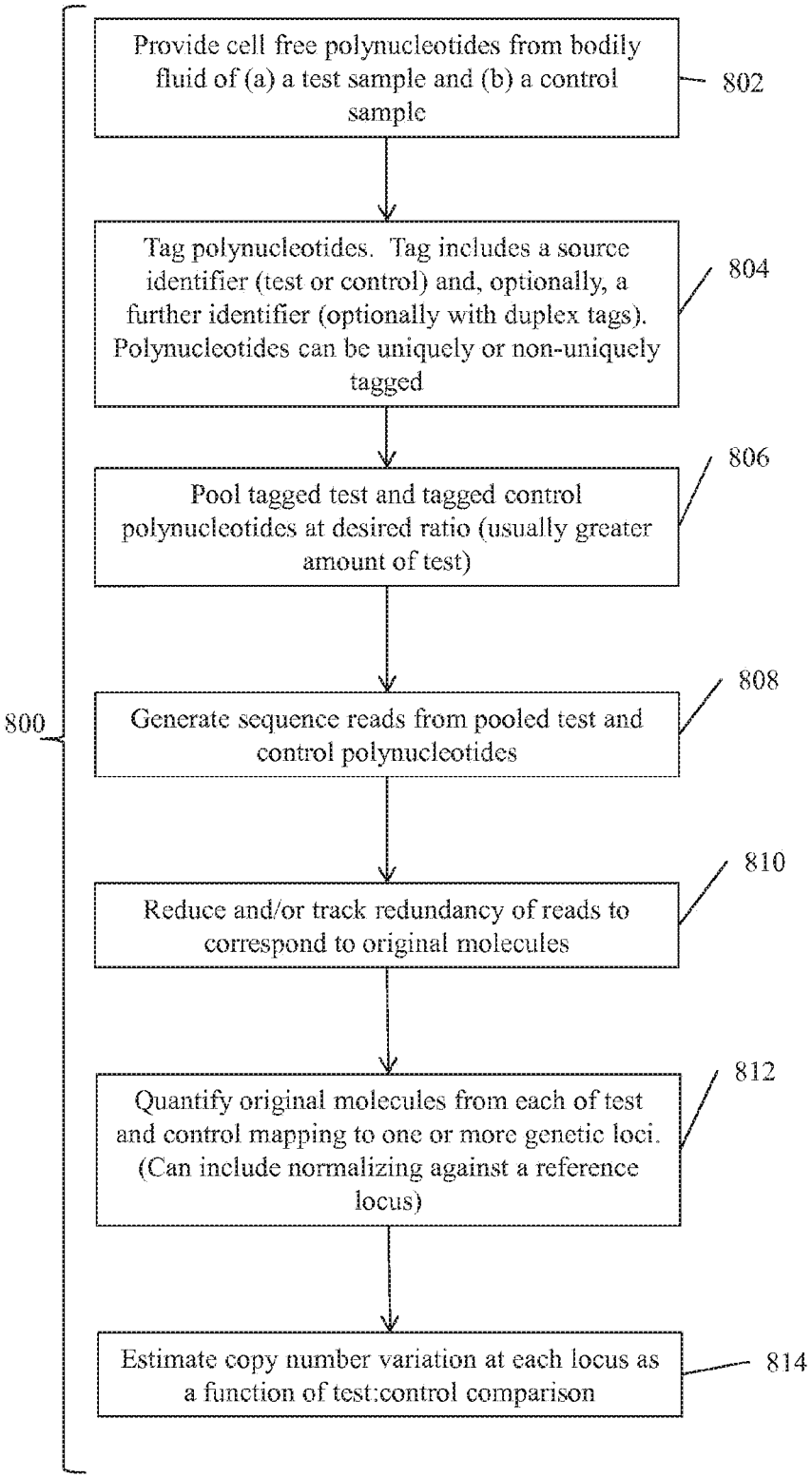

Provide cell free polynucleotides from bodily fluid of (a) a test sample and (b) a control sample
802
Tag polynucleotides. Tag includes a source identifier (test or control) and, optionally, a further identifier (optionally with duplex tags). Polynucleotides can be uniquely or non-uniquely tagged
804
Pool tagged test and tagged control polynucleotides at desired ratio (usually greater amount of test)
806
Generate sequence reads from pooled test and control polynucleotides
808
800
Reduce and/or track redundancy of reads to correspond to original molecules
810
Quantify original molecules from each of test and control mapping to one or more genetic loci. (Can include normalizing against a reference locus)
812
Estimate copy number variation at each locus as a function of test:control comparison
814

Fig. 8

Fig. 9A

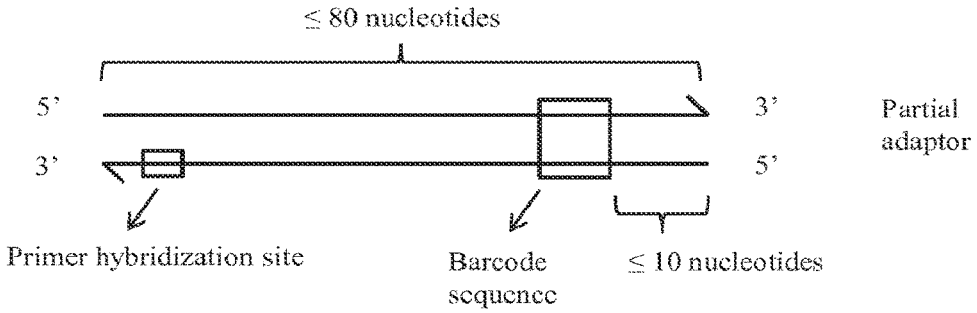


Fig. 9B

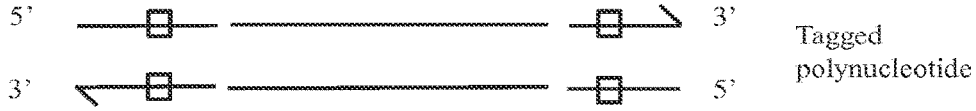
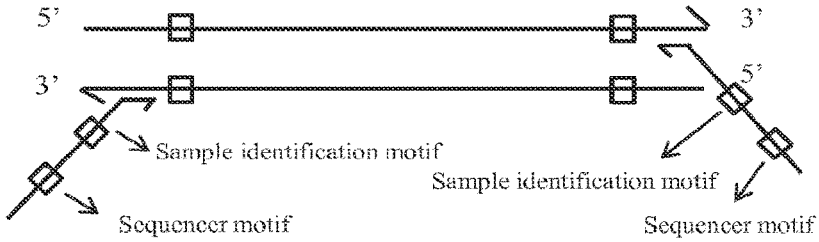


Fig. 9C



METHODS AND SYSTEMS FOR DETECTING GENETIC VARIANTS

CROSS-REFERENCE

This application is a continuation of U.S. application Ser. No. 16/672,267, filed Nov. 1, 2019, which is a continuation of U.S. application Ser. No. 16/601,168, filed Oct. 14, 2019, which is a continuation of U.S. application Ser. No. 15/892,178, filed Feb. 8, 2018, which is a continuation of U.S. application Ser. No. 14/861,989, filed Sep. 22, 2015 (now U.S. Pat. No. 9,920,366, issued Mar. 20, 2018), which is a continuation application of International Application No. PCT/US2014/072383, filed Dec. 24, 2014, which application claims the benefit under 35 U.S.C. § 119(e) of U.S. Provisional Application No. 61/921,456, filed Dec. 28, 2013, and U.S. Provisional Application No. 61/948,509, filed Mar. 5, 2014, each of which is entirely incorporated herein by reference.

BACKGROUND

The detection and quantification of polynucleotides is important for molecular biology and medical applications, such as diagnostics. Genetic testing is particularly useful for a number of diagnostic methods. For example, disorders that are caused by rare genetic alterations (e.g., sequence variants) or changes in epigenetic markers, such as cancer and partial or complete aneuploidy, may be detected or more accurately characterized with DNA sequence information.

Early detection and monitoring of genetic diseases, such as cancer, is often useful and needed in the successful treatment or management of the disease. One approach may include the monitoring of a sample derived from cell-free nucleic acids, a population of polynucleotides that can be found in different types of bodily fluids. In some cases, disease may be characterized or detected based on detection of genetic aberrations, such as copy number variation and/or sequence variation of one or more nucleic acid sequences, or the development of other certain rare genetic alterations. Cell-free DNA (cfDNA) may contain genetic aberrations associated with a particular disease. With improvements in sequencing and techniques to manipulate nucleic acids, there is a need in the art for improved methods and systems for using cell-free DNA to detect and monitor disease.

In particular, many methods have been developed for accurate copy number variation estimation, especially for heterogeneous genomic samples, such as tumor-derived gDNA or for cfDNA for many applications (e.g., prenatal, transplant, immune, metagenomics or cancer diagnostics). Most of these methods include sample preparation whereby the original nucleic acids are converted into a sequenceable library, followed by massively parallel sequencing, and finally bioinformatics to estimate copy number variation at one or more loci.

SUMMARY

Although many of these methods are able to reduce or combat the errors introduced by the sample preparation and sequencing processes for all molecules that are converted and sequenced, these methods are not able to infer the counts of molecules that were converted but not sequenced. Since this count of converted by unsequenced molecules can be highly variable from genomic region to region, these counts can dramatically and adversely affect the sensitivity that can be achieved.

To address this issue, input double-stranded deoxyribonucleic acid (DNA) can be converted by a process that tags both halves of the individual double-stranded molecule, in some cases differently. This can be performed using a variety of techniques, including ligation of hairpin, bubble, or forked adapters or other adaptors having double-stranded and single stranded segments (the unhybridized portion of a bubble, forked or hairpin adapter are deemed single-stranded herein). If tagged correctly, each original Watson and Crick (i.e., strand) side of the input double-stranded DNA molecule can be differently tagged and identified by the sequencer and subsequent bioinformatics. For all molecules in a particular region, counts of molecules where both Watson and Crick sides were recovered ("Pairs") versus those where only one half was recovered ("Singlets") can be recorded. The number of unseen molecules can be estimated based on the number of Pairs and Singlets detected.

An aspect of the present disclosure provides a method for detecting and/or quantifying rare deoxyribonucleic acid (DNA) in a heterogeneous population of original DNA fragments, comprising tagging the original DNA fragments in a single reaction using a library of a plurality of different tags such that greater than 30% of the fragments are tagged at both ends, wherein each of the tags comprises a molecular barcode. The single reaction can be in a single reaction vessel. Greater than 50% of the fragments can be tagged at both ends. The plurality of different tags can be no more than any of 100, 500, 1000, 10,000 or 100,000 different tags.

Another aspect provides a set of library adaptors that can be used to tag the molecules of interest (e.g., by ligation, hybridization, etc.). The set of library adaptors can comprise plurality of polynucleotide molecules with molecular barcodes, wherein the plurality of polynucleotide molecules are less than or equal to 80 nucleotide bases in length, wherein the molecular barcodes are at least 4 nucleotide bases in length, and wherein (a) the molecular barcodes are different from one another and have an edit distance of at least 1 between one another; (b) the molecular barcodes are located at least one nucleotide base away from a terminal end of their respective polynucleotide molecules; (c) optionally, at least one terminal base is identical in all of the polynucleotide molecules; and (d) none of the polynucleotide molecules contains a complete sequencer motif.

In some embodiments, the library adaptors (or adapters) are identical to one another but for the molecular barcodes. In some embodiments, each of the plurality of library adaptors comprises at least one double-stranded portion and at least one single-stranded portion (e.g., a non-complementary portion or an overhang). In some embodiments, the double-stranded portion has a molecular barcode selected from a collection of different molecular barcodes. In some embodiments, the given molecular barcode is a randomer. In some embodiments, each of the library adaptors further comprises a strand-identification barcode on the at least one single-stranded portion. In some embodiments, the strand-identification barcode includes at least 4 nucleotide bases. In some embodiments, the single-stranded portion has a partial sequencer motif. In some embodiments, the library adaptors do not include a complete sequencer motif.

In some embodiments, none of the library adaptors contains a sequence for hybridizing to a flow cell or forming a hairpin for sequencing.

In some embodiments, all of the library adaptors have a terminal end with nucleotide(s) that are the same. In some embodiments, the identical terminal nucleotide(s) are over two or more nucleotide bases in length.

In some embodiments, each of the library adapters is Y-shaped, bubble shaped or hairpin shaped. In some embodiments, none of the library adapters contains a sample identification motif. In some embodiments, each of the library adapters comprises a sequence that is selectively hybridizable to a universal primer. In some embodiments, each of the library adapters comprises a molecular barcode that is at least 5, 6, 7, 8, 9 and 10 nucleotide bases in length. In some embodiments, each of the library adapters is from 10 nucleotide bases to 80 in length, or 30 to 70 nucleotide bases in length, or 40 to 60 nucleotide bases in length. In some embodiments, at least 1, 2, 3, or 4 terminal bases are identical in all of the library adaptors. In some embodiments, at least 4 terminal bases are identical in all of the library adaptors.

In some embodiments, the edit distance of the molecular barcodes of the library adapters is a Hamming distance. In some embodiments, the edit distance is at least 1, 2, 3, 4 or 5. In some embodiments, the edit distance is with respect to individual bases of the plurality of polynucleotide molecules. In some embodiments, the molecular barcodes are located at least 10 nucleotide base away from a terminal end of an adapter. In some embodiments, the plurality of library adapters includes at least 2, 4, 6, 8, 10, 20, 30, 40 or 50 different molecular barcodes, or from 2-100, 4-80, 6-60 or 8-40 different molecular barcodes. In any of the embodiments herein, there are more polynucleotides (e.g., cfDNA fragments) to be tagged than there are different molecular barcodes such that the tagging is not unique.

In some embodiments, the terminal end of an adaptor is configured for ligation (e.g., to a target nucleic acid molecule). In some embodiments, the terminal end of an adaptor is a blunt end.

In some embodiments, the adaptors are purified and isolated. In some embodiments, the library comprises one or more non-naturally occurring bases.

In some embodiments, the polynucleotide molecules comprise a primer sequence positioned 5' with respect to the molecular barcodes.

In some embodiments, the set of library adaptors consists essentially of the plurality of polynucleotide molecules.

In another aspect, a method comprises (a) tagging a collection of polynucleotides with a plurality of polynucleotide molecules from a library of adaptors to create a collection of tagged polynucleotides; and (b) amplifying the collection of tagged polynucleotides in the presence of sequencing adaptors, wherein the sequencing adaptors have primers with nucleotide sequences that are selectively hybridizable to complementary sequences in the plurality of polynucleotide molecules. The library of adaptors may be as described above or elsewhere herein. In some embodiments, each of the sequencer adaptors further comprises an index tag, which can be a sample identification motif.

Another aspect, provides a method for detecting and/or quantifying rare DNA in a heterogeneous population of original DNA fragments, wherein the rare DNA has a concentration that is less than 1%, the method comprising (a) tagging the original DNA fragments in a single reaction such that greater than 30% of the original DNA fragments are tagged at both ends with library adaptors that comprise molecular barcodes, thereby providing tagged DNA fragments; (b) performing high-fidelity amplification on the tagged DNA fragments; (c) optionally, selectively enriching a subset of the tagged DNA fragments; (d) sequencing one or both strands of the tagged, amplified and optionally selectively enriched DNA fragments to obtain sequence reads comprising nucleotide sequences of the molecular barcodes and at least a portion of the original DNA fragments; (e) from the sequence reads, determining consensus reads that are representative of single-strands of the original DNA fragments; and (f) quantifying the consensus reads to detect and/or quantify the rare DNA at a specificity that is greater than 99.9%.

In some embodiments, (e) comprises comparing sequence reads having the same or similar molecular barcodes and the same or similar end of fragment sequences. In some embodiments, the comparing further comprises performing a phylogentic analysis on the sequence reads having the same or similar molecular barcodes. In some embodiments, the molecular barcodes include a barcode having an edit distance of up to 3. In some embodiments, the end of fragment sequence includes fragment sequences having an edit distance of up to 3.

In some embodiments, the method further comprises sorting sequence reads into paired reads and unpaired reads, and quantifying a number of paired reads and unpaired reads that map to each of one or more genetic loci.

In some embodiments, the tagging occurs by having an excess amount of library adaptors as compared to original DNA fragments. In some embodiments, n the excess is at least a 5-fold excess. In some embodiments, the tagging comprises using a ligase. In some embodiments, the tagging comprises attachment to blunt ends.

In some embodiments, the method further comprises binning the sequence reads according to the molecular barcodes and sequence information from at least one end of each of the original DNA fragments to create bins of single stranded reads. In some embodiments, the method further comprises, in each bin, determining a sequence of a given original DNA fragment among the original DNA fragments by analyzing sequence reads. In some embodiments, the method further comprises detecting and/or quantifying the rare DNA by comparing a number of times each base occurs at each position of a genome represented by the tagged, amplified, and optionally enriched DNA fragments.

In some embodiments, the library adaptors do not contain complete sequencer motifs. In some embodiments, the method further comprises selectively enriching a subset of the tagged DNA fragments. In some embodiments, the method further comprises, after enriching, amplifying the enriched tagged DNA fragments in the presence of sequencing adaptors comprising primers. In some embodiments, (a) provides tagged DNA fragments having from 2 to 1000 different combinations of molecular barcodes.

In some embodiments, the DNA fragments are tagged with polynucleotide molecules from a library of adaptors as described above or elsewhere herein.

In another aspect, a method for processing and/or analyzing a nucleic acid sample of a subject comprises (a) exposing polynucleotide fragments from the nucleic acid sample to a set of library adaptors to generate tagged polynucleotide fragments; and (b) subjecting the tagged polynucleotide fragments to nucleic acid amplification reactions under conditions that yield amplified polynucleotide fragments as amplification products of the tagged polynucleotide fragments. The set of library adaptors comprises a plurality of polynucleotide molecules with molecular barcodes, wherein the plurality of polynucleotide molecules are less than or equal to 80 nucleotide bases in length, wherein the molecular barcodes are at least 4 nucleotide bases in length, and wherein (1) the molecular barcodes are different from one another and have an edit distance of at least 1 between one another; (2) the molecular barcodes are located at least one nucleotide base away from a terminal end of

their respective polynucleotide molecules; (3) optionally, at least one terminal base is identical in all of the polynucleotide molecules; and (4) none of the polynucleotide molecules contains a complete sequencer motif.

In some embodiments, the method further comprises determining nucleotide sequences of the amplified tagged polynucleotide fragments. In some embodiments, the nucleotide sequences of the amplified tagged polynucleotide fragments are determined without polymerase chain reaction (PCR). In some embodiments, the method further comprises analyzing the nucleotide sequences with a programmed computer processor to identify one or more genetic variants in the nucleotide sample of the subject. In some embodiments, the one or more genetic variants are selected from the group consisting of base change(s), insertion(s), repeat(s), deletion(s), copy number variation(s) and transversion(s). In some embodiments, the one or more genetic variants include one or more tumor associated genetic alterations.

In some embodiments, the subject has or is suspected of having a disease. In some embodiments, the disease is cancer. In some embodiments, the method further comprises collecting the nucleic acid sample from the subject. In some embodiments, the nucleic acid sample is collected from a location selected from the group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, cerebral spinal fluid and tears of the subject. In some embodiments, the nucleic acid sample is a cell-free nucleic acid sample. In some embodiments, the nucleic acid sample is collected from no more than 100 nanograms (ng) of double-stranded polynucleotide molecules of the subject.

In some embodiments, the polynucleotide fragments comprise double-stranded polynucleotide molecules. In some embodiments, in (a), the plurality of polynucleotide molecules couple to the polynucleotide fragments via blunt end ligation, sticky end ligation, molecular inversion probes, PCR, ligation-based PCR, multiplex PCR, single stranded ligation, and single stranded circularization. In some embodiments, exposing the polynucleotide fragments of the nucleic acid sample to the plurality of polynucleotide molecules yields the tagged polynucleotide fragments with a conversion efficiency of at least 10%. In some embodiments, any of at least 5%, 6%, 7%, 8%, 9%, 10%, 20%, or 25% of the tagged polynucleotide fragments share a common polynucleotide molecule or sequence. In some embodiments, the method further comprises generating the polynucleotide fragments from the nucleic acid sample.

In some embodiments, the subjecting comprises amplifying the tagged polynucleotide fragments from sequences corresponding to genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

In another aspect, a method comprises (a) generating a plurality of sequence reads from a plurality of polynucleotide molecules, wherein the plurality of polynucleotide molecules cover genomic loci of a target genome, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1; (b) grouping with a computer processor the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides; (c) for each of the families, merging sequence reads to generate a consensus sequence; (d) calling the consensus sequence at a given genomic locus among the genomic loci; and (e) detecting at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls, and total number of alterations among the calls.

In some embodiments, each family comprises sequence reads from only one of the template polynucleotides. In some embodiments, the given genomic locus comprises at least one nucleic acid base. In some embodiments, the given genomic locus comprises a plurality of nucleic acid bases. In some embodiments, the calling comprises calling at least one nucleic acid base at the given genomic locus. In some embodiments, the calling comprises calling a plurality of nucleic acid bases at the given genomic locus. In some embodiments, the calling comprises any one of phylogenetic analysis, voting, weighing, assigning a probability to each read at the locus in a family and calling the base with the highest probability.

In some embodiments, the method further comprises performing (d)-(e) at an additional genomic locus among the genomic loci. In some embodiments, the method further comprises determining a variation in copy number at one of the given genomic locus and additional genomic locus based on counts at the given genomic locus and additional genomic locus.

In some embodiments, the grouping comprises classifying the plurality of sequence reads into families by identifying (i) different molecular barcodes coupled to the plurality of polynucleotide molecules and (ii) similarities between the plurality of sequence reads, wherein each family includes a plurality of nucleic acid sequences that are associated with a different combination of molecular barcodes and similar or identical sequence reads. Different molecular barcodes have different sequences.

In some embodiments, the consensus sequence is generated by evaluating a quantitative measure or a statistical significance level for each of the sequence reads. In some embodiments, the quantitative measure comprises use of a binomial distribution, exponential distribution, beta distribution, or empirical distribution. In some embodiments, the method further comprises mapping the consensus sequence to the target genome. In some embodiments, the plurality of genes includes at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50 or all of the plurality of genes selected from the group.

Another aspect of the present disclosure provides a method, comprising (a) providing template polynucleotide molecules and a set of library adaptors in a single reaction vessel, wherein the library adaptors are polynucleotide molecules that have different molecular barcodes (e.g., from 2 to 1,000 different molecular barcodes), and wherein none of the library adaptors contains a complete sequencer motif; (b) in the single reaction vessel, coupling the library adaptors to

the template polynucleotide molecules at an efficiency of at least 10%, thereby tagging each template polynucleotide with a tagging combination that is among a plurality of different tagging combinations (e.g., 4 to 1,000,000 different tagging combinations), to produce tagged polynucleotide molecules; (c) subjecting the tagged polynucleotide molecules to an amplification reaction under conditions that yield amplified polynucleotide molecules as amplification products of the tagged polynucleotide molecules; and (d) sequencing the amplified polynucleotide molecules.

In some embodiments, the template polynucleotide molecules are blunt ended or sticky-ended. In some embodiments, the library adaptors are identical but for the molecular barcodes. In some embodiments, each of the library adaptors has a double stranded portion and at least one single-stranded portion. In some embodiments, the double-stranded portion has a molecular barcode among the molecular barcodes. In some embodiments, each of the library adaptors further comprises a strand-identification barcode on the at least one single-stranded portion. In some embodiments, the single-stranded portion has a partial sequencer motif. In some embodiments, the library adaptors have a sequence of terminal nucleotides that are the same. In some embodiments, the template polynucleotide molecules are double-stranded. In some embodiments, the library adaptors couple to both ends of the template polynucleotide molecules.

In some embodiments, subjecting the tagged polynucleotide molecules to the amplification reaction comprises non-specifically amplifying the tagged polynucleotide molecules.

In some embodiments, the amplification reaction comprises use of a priming site to amplify each of the tagged polynucleotide molecules. In some embodiments, the priming site is a primer. In some embodiments, the primer is a universal primer. In some embodiments, the priming site is a nick.

In some embodiments, the method further comprises, prior to (e), (i) separating polynucleotide molecules comprising one or more given sequences from the amplified polynucleotide molecules, to produce enriched polynucleotide molecules; and (ii) amplifying the enriched polynucleotide molecules with sequencing adaptors.

In some embodiments, the efficiency is at least 30%, 40%, or 50%. In some embodiments, the method further comprises identifying genetic variants upon sequencing the amplified polynucleotide molecules. In some embodiments, the sequencing comprises (i) subjecting the amplified polynucleotide molecules to an additional amplification reaction under conditions that yield additional amplified polynucleotide molecules as amplification products of the amplified polynucleotide molecules, and (ii) sequencing the additional amplified polynucleotide molecules. In some embodiments, the additional amplification is performed in the presence of sequencing adaptors.

In some embodiments, (b) and (c) are performed without aliquoting the tagged polynucleotide molecules. In some embodiments, the tagging is non-unique tagging.

Another aspect, provides a system for analyzing a target nucleic acid molecule of a subject, comprising a communication interface that receives nucleic acid sequence reads for a plurality of polynucleotide molecules that cover genomic loci of a target genome; computer memory that stores the nucleic acid sequence reads for the plurality of polynucleotide molecules received by the communication interface; and a computer processor operatively coupled to the communication interface and the memory and programmed to (i) group the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides, (ii) for each of the families, merge sequence reads to generate a consensus sequence, (iii) call the consensus sequence at a given genomic locus among the genomic loci, and (iv) detect at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls; and total number of alterations among the calls, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

In another aspect, a set of oligonucleotide molecules that selectively hybridize to at least 5 genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

In some embodiments, the oligonucleotide molecules are from 10-200 bases in length. In some embodiments, the oligonucleotide molecules selectively hybridize to exon regions of the at least 5 genes. In some embodiments, the oligonucleotide molecules selectively hybridize to at least 30 exons in the at least 5 genes. In some embodiments, multiple oligonucleotide molecules selectively hybridize to each of the at least 30 exons. In some embodiments, the oligonucleotide molecules that hybridize to each exon have sequences that overlap with at least 1 other oligonucleotide molecule.

In another aspect, a kit comprises a first container containing a plurality of library adaptors each having a different molecular barcode; and a second container containing a plurality of sequencing adaptors, each sequencing adaptor comprising at least a portion of a sequencer motif and optionally a sample barcode. The library adaptors can be as described above or elsewhere herein.

In some embodiments, the sequencing adaptor comprises the sample barcode. In some embodiments, the library adaptors are blunt ended and Y-shaped, and are less than or equal to 80 nucleic acid bases in length. In some embodiments, the sequencing adaptor is up to 70 bases from end to end.

In another aspect, a method for detecting sequence variants in a cell free DNA sample, comprising detecting rare DNA at a concentration less than 1% with a specificity that is greater than 99.9%.

In another aspect, a method comprises detecting genetic variants in a sample comprising DNA with a detection limit of at least 1% and specificity greater than 99.9%. In some

embodiments, the method further comprises converting cDNA (e.g. cfDNA) into adaptor tagged DNA with a conversion efficiency of at least 30%, 40%, or 50% and reducing sequencing noise (or distortion) by eliminating false positive sequence reads.

Another aspect provides a method, comprising (a) providing a sample comprising a set of double-stranded polynucleotide molecules, each double-stranded polynucleotide molecule including first and second complementary strands; (b) tagging the double-stranded polynucleotide molecules with a set of duplex tags, wherein each duplex tag differently tags the first and second complementary strands of a double-stranded polynucleotide molecule in the set; (c) sequencing at least some of the tagged strands to produce a set of sequence reads; (d) reducing and/or tracking redundancy in the set of sequence reads; (e) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; (f) determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci; and (g) estimating with a programmed computer processor a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

In some embodiments, the method further comprises (h) detecting copy number variation in the sample by determining a normalized total quantitative measure determined in step (g) at each of the one or more genetic loci and determining copy number variation based on the normalized measure. In some embodiments, the sample comprises double-stranded polynucleotide molecules sourced substantially from cell-free nucleic acids. In some embodiments, the duplex tags are not sequencing adaptors.

In some embodiments, reducing redundancy in the set of sequence reads comprises collapsing sequence reads produced from amplified products of an original polynucleotide molecule in the sample back to the original polynucleotide molecule. In some embodiments, the method further comprises determining a consensus sequence for the original polynucleotide molecule. In some embodiments, the method further comprises identifying polynucleotide molecules at one or more genetic loci comprising a sequence variant. In some embodiments, the method further comprises determining a quantitative measure of paired reads that map to a locus, wherein both strands of the pair comprise a sequence variant. In some embodiments, the method further comprises determining a quantitative measure of paired molecules in which only one member of the pair bears a sequence variant and/or determining a quantitative measure of unpaired molecules bearing a sequence variant. In some embodiments, the sequence variant is selected from the group consisting of a single nucleotide variant, an indel, a transversion, a translocation, an inversion, a deletion, a chromosomal structure alteration, a gene fusion, a chromosome fusion, a gene truncation, a gene amplification, a gene duplication and a chromosomal lesion.

Another aspect provides a system comprising a computer readable medium comprising machine-executable code that, upon execution by a computer processor, implements a method comprising (a) receiving into memory a set of sequence reads of polynucleotides tagged with duplex tags; (b) reducing and/or tracking redundancy in the set of sequence reads; (c) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; (d) determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci; and (e) estimating a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

Another aspect provides a method, comprising (a) providing a sample comprising a set of double-stranded polynucleotide molecules, each double-stranded polynucleotide molecule including first and second complementary strands; (b) tagging the double-stranded polynucleotide molecules with a set of duplex tags, wherein each duplex tag differently tags the first and second complementary strands of a double-stranded polynucleotide molecule in the set; (c) sequencing at least some of the tagged strands to produce a set of sequence reads; (d) reducing and/or tracking redundancy in the set of sequence reads; (e) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; and (f) determining quantitative measures of at least two of (i) the paired reads, (ii) the unpaired reads that map to each of one or more genetic loci, (iii) read depth of the paired reads and (iv) read depth of unpaired reads.

In some embodiments, (f) comprises determining quantitative measures of at least three of (i)-(iv). In some embodiments, (f) comprises determining quantitative measures of all of (i)-(iv). In some embodiments, the method further comprises (g) estimating with a programmed computer processor a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads and their read depths mapping to each locus.

In another aspect, a method comprises (a) tagging control parent polynucleotides with a first tag set to produce tagged control parent polynucleotides, wherein the first tag set comprises a plurality of tags, wherein each tag in the first tag set comprises a same control tag and an identifying tag, and wherein the tag set comprises a plurality of different identifying tags; (b) tagging test parent polynucleotides with a second tag set to produce tagged test parent polynucleotides, wherein the second tag set comprises a plurality of tags, wherein each tag in the second tag set comprises a same test tag that is distinguishable from the control tag and an identifying tag, and wherein the second tag set comprises a plurality of different identifying tags; (c) mixing tagged control parent polynucleotides with tagged test parent polynucleotides to form a pool; (d) amplifying tagged parent

polynucleotides in the pool to form a pool of amplified, tagged polynucleotides; (e) sequencing amplified, tagged polynucleotides in the amplified pool to produce a plurality of sequence reads; (f) grouping sequence reads into families, each family comprising sequence reads generated from a same parent polynucleotide, which grouping is optionally based on information from an identifying tag and from start/end sequences of the parent polynucleotides, and, optionally, determining a consensus sequence for each of a plurality of parent polynucleotides from the plurality of sequence reads in a group; (g) classifying each family or consensus sequence as a control parent polynucleotide or as a test parent polynucleotide based on having a test tag or a control tag; (h) determining a quantitative measure of control parent polynucleotides and control test polynucleotides mapping to each of at least two genetic loci; and (i) determining copy number variation in the test parent polynucleotides at at least one locus based on relative quantity of test parent polynucleotides and control parent polynucleotides mapping to the at least one locus.

In another aspect, a method comprises (a) generating a plurality of sequence reads from a plurality of template polynucleotides, each polynucleotide mapped to a genomic locus; (b) grouping the sequence reads into families, each family comprising sequence reads generated from one of the template polynucleotides; (c) calling a base (or sequence) at the genomic locus for each of the families; (d) detecting at the genomic locus any of genomic alterations among the calls, frequency of a genetic alteration among the calls, total number of calls and total number of alterations among the calls.

In some embodiments, calling comprises any of phylogenetic analysis, voting, weighing, assigning a probability to each read at the locus in a family, and calling the base with the highest probability. In some embodiments, the method is performed at two loci, comprising determining CNV at one of the loci based on counts at each of the loci.

Another aspect provides a method for determining a quantitative measure indicative of a number of individual double-stranded DNA fragments in a sample comprising (a) determining a quantitative measure of individual DNA molecules for which both strands are detected; (b) determining a quantitative measure of individual DNA molecules for which only one of the DNA strands are detected; (c) inferring from (a) and (b) above a quantitative measure of individual DNA molecules for which neither strand was detected; and (d) using (a)-(c) determining the quantitative measure indicative of a number of individual double-stranded DNA fragments in the sample.

In some embodiments, the method further comprises detecting copy number variation in the sample by determining a normalized quantitative measure determined in step (d) at each of one or more genetic loci and determining copy number variation based on the normalized measure. In some embodiments, the sample comprises double-stranded polynucleotide molecules sourced substantially from cell-free nucleic acids.

In some embodiments, determining the quantitative measure of individual DNA molecules comprises tagging the DNA molecules with a set of duplex tags, wherein each duplex tag differently tags complementary strands of a double-stranded DNA molecule in the sample to provide tagged strands. In some embodiments, the method further comprises sequencing at least some of the tagged strands to produce a set of sequence reads. In some embodiments, the method further comprises sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads. In some embodiments, the method further comprises determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci to determine a quantitative measure of total double-stranded DNA molecules in the sample that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

In another aspect, a method for reducing distortion in a sequencing assay, comprises (a) tagging control parent polynucleotides with a first tag set to produce tagged control parent polynucleotides; (b) tagging test parent polynucleotides with a second tag set to produce tagged test parent polynucleotides; (c) mixing tagged control parent polynucleotides with tagged test parent polynucleotides to form a pool; (d) determining quantities of tagged control parent polynucleotides and tagged test parent polynucleotides; and (e) using the quantities of tagged control parent polynucleotides to reduce distortion in the quantities of tagged test parent polynucleotides.

In some embodiments, the first tag set comprises a plurality of tags, wherein each tag in the first tag set comprises a same control tag and an identifying tag, and wherein the first tag set comprises a plurality of different identifying tags. In some embodiments, the second tag set comprises a plurality of tags, wherein each tag in the second tag set comprises a same test tag and an identifying tag, wherein the test tag is distinguishable from the control tag, and wherein the second tag set comprises a plurality of different identifying tags. In some embodiments, (d) comprises amplifying tagged parent polynucleotides in the pool to form a pool of amplified, tagged polynucleotides, and sequencing amplified, tagged polynucleotides in the amplified pool to produce a plurality of sequence reads. In some embodiments, the method further comprises grouping sequence reads into families, each family comprising sequence reads generated from a same parent polynucleotide, which grouping is optionally based on information from an identifying tag and from start/end sequences of the parent polynucleotides, and, optionally, determining a consensus sequence for each of a plurality of parent polynucleotides from the plurality of sequence reads in a group.

In some embodiments, (d) comprises determining copy number variation in the test parent polynucleotides at greater than or equal to one locus based on relative quantity of test parent polynucleotides and control parent polynucleotides mapping to the locus.

Another aspect provides a method comprising (a) ligating adaptors to double-stranded DNA polynucleotides, wherein ligating is performed in a single reaction vessel, and wherein the adaptors comprise molecular barcodes, to produce a tagged library comprising an insert from the double-stranded DNA polynucleotides, and having between 4 and 1 million different tags; (b) generating a plurality of sequence reads for each of the double-stranded DNA polynucleotides in the tagged library; (c) grouping sequence reads into families, each family comprising sequence reads generated from a single DNA polynucleotide among the double-stranded DNA polynucleotides, based on information in a tag and information at an end of the insert; and (d) calling bases at

each position in the double-stranded DNA molecule based on bases at the position in members of a family. In some embodiments, (b) comprises amplifying each of the double-stranded DNA polynucleotide molecules in the tagged library to generate amplification products, and sequencing the amplification products. In some embodiments, the method further comprises sequencing the double-stranded DNA polynucleotide molecules a plurality of times. In some embodiments, (b) comprises sequencing the entire insert. In some embodiments, (c) further comprises collapsing sequence reads in each family to generate a consensus sequence. In some embodiments, (d) comprises calling a plurality of sequential bases from at least a subset of the sequence reads to identify single nucleotide variations (SNV) in the double-stranded DNA molecule.

Another aspect provides a method of detecting disease cell heterogeneity from a sample comprising polynucleotides from somatic cells and disease cells. The method comprises quantifying polynucleotides in the sample bearing a nucleotide sequence variant at each of a plurality of genetic loci; determining copy number variation (CNV) at each of the plurality of genetic loci, wherein the CNV indicates a genetic dose of a locus in the disease cell polynucleotides; determining with a programmed computer processor a relative measure of quantity of polynucleotides bearing a sequence variant at a locus per the genetic dose at the locus for each of a plurality of the loci; and comparing the relative measures at each of the plurality of loci, wherein different relative measures is indicative of tumor heterogeneity.

In another aspect, a method comprises subjecting a subject to one or more pulsed therapy cycles, each pulsed therapy cycle comprising (a) a first period during which a drug is administered at a first amount; and (b) a second period during which the drug is administered at a second, reduced amount, wherein (i) the first period is characterized by a tumor burden detected above a first clinical level; and (ii) the second period is characterized by a tumor burden detected below a second clinical level.

Additional aspects and advantages of the present disclosure will become readily apparent to those skilled in this art from the following detailed description, wherein only illustrative embodiments of the present disclosure are shown and described. As will be realized, the present disclosure is capable of other and different embodiments, and its several details are capable of modifications in various obvious respects, all without departing from the disclosure. Accordingly, the drawings and description are to be regarded as illustrative in nature, and not as restrictive.

INCORPORATION BY REFERENCE

All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings (also "figure" and "FIG." herein), of which:

FIG. **1** is a flowchart representation of a method of the present disclosure for determining copy number variation (CNV);

FIG. **2** depicts mapping of pairs and singlets to Locus A and Locus B in a genome;

FIG. **3** shows a reference sequence encoding a genetic Locus A;

FIGS. **4**A-C shows amplification, sequencing, redundancy reduction and pairing of complementary molecules;

FIG. **5** shows increased confidence in detecting sequence variants by pairing reads from Watson and Crick strands;

FIG. **6** shows a computer system that is programmed or otherwise configured to implement various methods of the present disclosure;

FIG. **7** is schematic representation of a system for analyzing a sample comprising nucleic acids from a user, including a sequencer; bioinformatic software and internet connection for report analysis by, for example, a hand held device or a desk top computer;

FIG. **8** is a flowchart representation of a method of this invention for determining CNV using pooled test and control pools; and

FIGS. **9**A-**9**C schematically illustrate a method for tagging a polynucleotide molecule with a library adaptor and subsequently a sequencing adaptor.

DETAILED DESCRIPTION

While various embodiments of the invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions may occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed.

The term "genetic variant," as used herein, generally refers to an alteration, variant or polymorphism in a nucleic acid sample or genome of a subject. Such alteration, variant or polymorphism can be with respect to a reference genome, which may be a reference genome of the subject or other individual. Single nucleotide polymorphisms (SNPs) are a form of polymorphisms. In some examples, one or more polymorphisms comprise one or more single nucleotide variations (SNVs), insertions, deletions, repeats, small insertions, small deletions, small repeats, structural variant junctions, variable length tandem repeats, and/or flanking sequences. Copy number variants (CNVs), transversions and other rearrangements are also forms of genetic variation. A genomic alternation may be a base change, insertion, deletion, repeat, copy number variation, or transversion.

The term "polynucleotide," as used herein, generally refers to a molecule comprising one or more nucleic acid subunits. A polynucleotide can include one or more subunits selected from adenosine (A), cytosine (C), guanine (G), thymine (T) and uracil (U), or variants thereof. A nucleotide can include A, C, G, T or U, or variants thereof. A nucleotide can include any subunit that can be incorporated into a growing nucleic acid strand. Such subunit can be an A, C, G, T, or U, or any other subunit that is specific to one or more complementary A, C, G, T or U, or complementary to a purine (i.e., A or G, or variant thereof) or a pyrimidine (i.e., C, T or U, or variant thereof). A subunit can enable individual nucleic acid bases or groups of bases (e.g., AA, TA, AT, GC, CG, CT, TC, GT, TG, AC, CA, or uracil-counterparts thereof) to be resolved. In some examples, a polynucleotide is deoxyribonucleic acid (DNA) or ribonucleic

acid (RNA), or derivatives thereof. A polynucleotide can be single-stranded or double stranded.

The term "subject," as used herein, generally refers to an animal, such as a mammalian species (e.g., human) or avian (e.g., bird) species, or other organism, such as a plant. More specifically, the subject can be a vertebrate, a mammal, a mouse, a primate, a simian or a human. Animals include, but are not limited to, farm animals, sport animals, and pets. A subject can be a healthy individual, an individual that has or is suspected of having a disease or a pre-disposition to the disease, or an individual that is in need of therapy or suspected of needing therapy. A subject can be a patient.

The term "genome" generally refers to an entirety of an organism's hereditary information. A genome can be encoded either in DNA or in RNA. A genome can comprise coding regions that code for proteins as well as non-coding regions. A genome can include the sequence of all chromosomes together in an organism. For example, the human genome has a total of 46 chromosomes. The sequence of all of these together constitutes a human genome.

The terms "adaptor(s)", "adapter(s)" and "tag(s)" are used synonymously throughout this specification. An adaptor or tag can be coupled to a polynucleotide sequence to be "tagged" by any approach including ligation, hybridization, or other approaches.

The term "library adaptor" or "library adapter" as used herein, generally refers to a molecule (e.g., polynucleotide) whose identity (e.g., sequence) can be used to differentiate polynucleotides in a biological sample (also "sample" herein).

The term "sequencing adaptor," as used herein, generally refers to a molecule (e.g., polynucleotide) that is adapted to permit a sequencing instrument to sequence a target polynucleotide, such as by interacting with the target polynucleotide to enable sequencing. The sequencing adaptor permits the target polynucleotide to be sequenced by the sequencing instrument. In an example, the sequencing adaptor comprises a nucleotide sequence that hybridizes or binds to a capture polynucleotide attached to a solid support of a sequencing system, such as a flow cell. In another example, the sequencing adaptor comprises a nucleotide sequence that hybridizes or binds to a polynucleotide to generate a hairpin loop, which permits the target polynucleotide to be sequenced by a sequencing system. The sequencing adaptor can include a sequencer motif, which can be a nucleotide sequence that is complementary to a flow cell sequence of other molecule (e.g., polynucleotide) and usable by the sequencing system to sequence the target polynucleotide. The sequencer motif can also include a primer sequence for use in sequencing, such as sequencing by synthesis. The sequencer motif can include the sequence(s) needed to couple a library adaptor to a sequencing system and sequence the target polynucleotide.

As used herein the terms "at least", "at most" or "about", when preceding a series, refers to each member of the series, unless otherwise identified.

The term "about" and its grammatical equivalents in relation to a reference numerical value can include a range of values up to plus or minus 10% from that value. For example, the amount "about 10" can include amounts from 9 to 11. In other embodiments, the term "about" in relation to a reference numerical value can include a range of values plus or minus 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1% from that value.

The term "at least" and its grammatical equivalents in relation to a reference numerical value can include the reference numerical value and greater than that value. For example, the amount "at least 10" can include the value 10 and any numerical value above 10, such as 11, 100, and 1,000.

The term "at most" and its grammatical equivalents in relation to a reference numerical value can include the reference numerical value and less than that value. For example, the amount "at most 10" can include the value 10 and any numerical value under 10, such as 9, 8, 5, 1, 0.5, and 0.1.

1. Methods for Processing and/or Analyzing a Nucleic Acid Sample

An aspect of the present disclosure provides methods for determining a genomic alternation in a nucleic acid sample of a subject. FIG. **1** shows a method of determining copy number variation (CNV). The method can be implemented to determine other genomic alternations, such as SNVs.

A. Polynucleotide Isolation

Methods disclosed herein can comprise isolating one or more polynucleotides. A polynucleotide can comprise any type of nucleic acid, for example, a sequence of genomic nucleic acid, or an artificial sequence (e.g., a sequence not found in genomic nucleic acid). For example, an artificial sequence can contain non-natural nucleotides. Also, a polynucleotide can comprise both genomic nucleic acid and an artificial sequence, in any portion. For example, a polynucleotide can comprise 1 to 99% of genomic nucleic acid and 99% to 1% of artificial sequence, where the total adds up to 100%. Thus, fractions of percentages are also contemplated. For example, a ratio of 99.1% to 0.9% is contemplated.

A polynucleotide can comprise any type of nucleic acids, such as DNA and/or RNA. For example, if a polynucleotide is DNA, it can be genomic DNA, complementary DNA (cDNA), or any other deoxyribonucleic acid. A polynucleotide can also be cell-free DNA (cfDNA). For example, the polynucleotide can be circulating DNA. The circulating DNA can comprise circulating tumor DNA (ctDNA). A polynucleotide can be double-stranded or single-stranded. Alternatively, a polynucleotide can comprise a combination of a double-stranded portion and a single-stranded portion.

Polynucleotides do not have to be cell-free. In some cases, the polynucleotides can be isolated from a sample. For example, in step (**102**) (FIG. **1**), double-stranded polynucleotides are isolated from a sample. A sample can be any biological sample isolated from a subject. For example, a sample can comprise, without limitation, bodily fluid, whole blood, platelets, serum, plasma, stool, red blood cells, white blood cells or leucocytes, endothelial cells, tissue biopsies, synovial fluid, lymphatic fluid, ascites fluid, interstitial or extracellular fluid, the fluid in spaces between cells, including gingival crevicular fluid, bone marrow, cerebrospinal fluid, saliva, mucous, sputum, semen, sweat, urine, or any other bodily fluids. A bodily fluid can include saliva, blood, or serum. For example, a polynucleotide can be cell-free DNA isolated from a bodily fluid, e.g., blood or serum. A sample can also be a tumor sample, which can be obtained from a subject by various approaches, including, but not limited to, venipuncture, excretion, ejaculation, massage, biopsy, needle aspirate, lavage, scraping, surgical incision, or intervention or other approaches.

A sample can comprise various amount of nucleic acid that contains genome equivalents. For example, a sample of about 30 ng DNA can contain about 10,000 (10^4) haploid human genome equivalents and, in the case of cfDNA, about 200 billion (2×10^{11}) individual polynucleotide molecules. Similarly, a sample of about 100 ng of DNA can contain about 30,000 haploid human genome equivalents and, in the case of cfDNA, about 600 billion individual molecules.

A sample can comprise nucleic acids from different sources. For example, a sample can comprise germline DNA or somatic DNA. A sample can comprise nucleic acids carrying mutations. For example, a sample can comprise DNA carrying germline mutations and/orsomatic mutations. A sample can also comprise DNA carrying cancer-associated mutations (e.g., cancer-associated somatic mutations).

B. Tagging

Polynucleotides disclosed herein can be tagged. For example, in step (**104**) (FIG. **1**) the double-stranded polynucleotides are tagged with duplex tags, tags that differently label the complementary strands (i.e., the "Watson" and "Crick" strands) of a double-stranded molecule. In one embodiment the duplex tags are polynucleotides having complementary and non-complementary portions.

Tags can be any types of molecules attached to a polynucleotide, including, but not limited to, nucleic acids, chemical compounds, fluorescent probes, or radioactive probes. Tags can also be oligonucleotides (e.g., DNA or RNA). Tags can comprise known sequences, unknown sequences, or both. A tag can comprise random sequences, pre-determined sequences, or both. A tag can be double-stranded or single-stranded. A double-stranded tag can be a duplex tag. A double-stranded tag can comprise two complementary strands. Alternatively, a double-stranded tag can comprise a hybridized portion and a non-hybridized portion. The double-stranded tag can be Y-shaped, e.g., the hybridized portion is at one end of the tag and the non-hybridized portion is at the opposite end of the tag. One such example are the "Y adapters" used in Illumina sequencing. Other examples include hairpin shaped adapters or bubble shaped adapters. Bubble shaped adapters have non-complementary sequences flanked on both sides by complementary sequences.

Tagging disclosed herein can be performed using any method. A polynucleotide can be tagged with an adaptor by hybridization. For example, the adaptor can have a nucleotide sequence that is complementary to at least a portion of a sequence of the polynucleotide. As an alternative, a polynucleotide can be tagged with an adaptor by ligation.

For example, tagging can comprise using one or more enzymes. The enzyme can be a ligase. The ligase can be a DNA ligase. For example, the DNA ligase can be a T4 DNA ligase, *E. coli* DNA ligase, and/or mammalian ligase. The mammalian ligase can be DNA ligase I, DNA ligase III, or DNA ligase IV. The ligase can also be a thermostable ligase. Tags can be ligated to a blunt-end of a polynucleotide (blunt-end ligation). Alternatively, tags can be ligated to a sticky end of a polynucleotide (sticky-end ligation). Efficiency of ligation can be increased by optimizing various conditions. Efficiency of ligation can be increased by optimizing the reaction time of ligation. For example, the reaction time of ligation can be less than 12 hours, e.g., less than 1, less than 2, less than 3, less than 4, less than 5, less than 6, less than 7, less than 8, less than 9, less than 10, less than 11, less than 12, less than 13, less than 14, less than 15, less than 16, less than 17, less than 18, less than 19, or less than 20 hours. In a particular example, reaction time of ligation is less than 20 hours. Efficiency of ligation can be increased by optimizing the ligase concentration in the reaction. For example, the ligase concentration can be at least 10, at least 50, at least 100, at least 150, at least 200, at least 250, at least 300, at least 400, at least 500, or at least 600 unit/microliter. Efficiency can also be optimized by adding or varying the concentration of an enzyme suitable for ligation, enzyme cofactors or other additives, and/or optimizing a temperature of a solution having the enzyme. Efficiency can also be optimized by varying the addition order of various components of the reaction. The end of tag sequence can comprise dinucleotide to increase ligation efficiency. When the tag comprises a non-complementary portion (e.g., Y-shaped adaptor), the sequence on the complementary portion of the tag adaptor can comprise one or more selected sequences that promote ligation efficiency. Preferably such sequences are located at the terminal end of the tag. Such sequences can comprise 1, 2, 3, 4, 5, or 6 terminal bases. Reaction solution with high viscosity (e.g., a low Reynolds number) can also be used to increase ligation efficiency. For example, solution can have a Reynolds number less than 3000, less than 2000, less than 1000, less than 900, less than 800, less than 700, less than 600, less than 500, less than 400, less than 300, less than 200, less than 100, less than 50, less than 25, or less than 10. It is also contemplated that roughly unified distribution of fragments (e.g., tight standard deviation) can be used to increase ligation efficiency. For example, the variation in fragment sizes can vary by less than 20%, less than 15%, less than 10%, less than 5%, or less than 1%. Tagging can also comprise primer extension, for example, by polymerase chain reaction (PCR). Tagging can also comprise any of ligation-based PCR, multiplex PCR, single strand ligation, or single strand circularization.

In some instances, the tags herein comprise molecular barcodes. Such molecular barcodes can be used to differentiate polynucleotides in a sample. Preferably molecular barcodes are different from one another. For example, molecular barcodes can have a difference between them that can be characterized by a predetermined edit distance or a Hamming distance. In some instances, the molecular barcodes herein have a minimum edit distance of 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10. To further improve efficiency of conversion (e.g., tagging) of untagged molecular to tagged molecules, one preferably utilizes short tags. For example, in some embodiments, a library adapter tag can be up to 65, 60, 55, 50, 45, 40, or 35 nucleotide bases in length. A collection of such short library barcodes preferably includes a number of different molecular barcodes, e.g., at least 2, 4, 6, 8, 10, 12, 14, 16, 18 or 20 different barcodes with a minimum edit distance of 1, 2, 3 or more.

Thus, a collection of molecules can include one or more tags. In some instances, some molecules in a collection can include an identifying tag ("identifier") such as a molecular barcode that is not shared by any other molecule in the collection. For example, in some instances of a collection of molecules, at least 50%, at least 51%, at least 52%, at least 53%, at least 54%, at least 55%, at least 56%, at least 57%, at least 58%, at least 59%, at least 60%, at least 61%, at least 62%, at least 63%, at least 64%, at least 65%, at least 66%, at least 67%, at least 68%, at least 69%, at least 70%, at least 71%, at least 72%, at least 73%, at least 74%, at least 75%, at least 76%, at least 77%, at least 78%, at least 79%, at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% of the molecules in the collection can include an identifier or molecular barcode that is not shared by any other molecule in the collection. As used herein, a collection of molecules is considered to be "uniquely tagged" if each of at least 95% of the molecules in the collection bears an identifier that is not shared by any other molecule in the collection ("unique tag" or "unique identifier"). A collection of molecules is considered to be "non-uniquely tagged" if each of at least 1%, at least 5%, at

least 10%, at least 15%, at least 20%, at least 25%, at least 30%, at least 35%, at least 40%, at least 45%, or at least or about 50% of the molecules in the collection bears an identifying tag or molecular barcode that is shared by at least one other molecule in the collection ("non-unique tag" or "non-unique identifier"). Accordingly, in a non-uniquely tagged population no more than 1% of the molecules are uniquely tagged. For example, in a non-uniquely tagged population, no more than 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, or 50% of the molecules can be uniquely tagged.

A number of different tags can be used based on the estimated number of molecules in a sample. In some tagging methods, the number of different tags can be at least the same as the estimated number of molecules in the sample. In other tagging methods, the number of different tags can be at least two, three, four, five, six, seven, eight, nine, ten, one hundred or one thousand times as many as the estimated number of molecules in the sample. In unique tagging, at least two times (or more) as many different tags can be used as the estimated number of molecules in the sample.

The molecules in the sample may be non-uniquely tagged. In such instances a fewer number of tags or molecular barcodes is used then the number of molecules in the sample to be tagged. For example, no more than 100, 50, 40, 30, 20 or 10 unique tags or molecular barcodes are used to tag a complex sample such as a cell free DNA sample with many more different fragments.

The polynucleotide to be tagged can be fragmented, such as either naturally or using other approaches, such as, for example, shearing. The polynucleotides can be fragmented by certain methods, including but not limited to, mechanical shearing, passing the sample through a syringe, sonication, heat treatment (e.g., for 30 minutes at 90° C.), and/or nuclease treatment (e.g., using DNase, RNase, endonuclease, exonuclease, and/or restriction enzyme).

The polynucleotides fragments (prior to tagging) can comprise sequences of any length. For example, polynucleotide fragments (prior to tagging) can comprise at least 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, 275, 280, 285, 290, 295, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000 or more nucleotides in length. The polynucleotide fragment are preferably about the average length of cell-free DNA. For example, the polynucleotide fragments can comprise about 160 bases in length. The polynucleotide fragment can also be fragmented from a larger fragment into smaller fragments about 160 bases in length.

Polynucleotides tagged can comprise sequences associated with cancer. The cancer-associated sequences can comprise single nucleotide variation (SNV), copy number variation (CNV), insertions, deletions, and/or rearrangements.

The polynucleotides can comprise sequences associated with cancer, such as acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), adrenocortical carcinoma, Kaposi Sarcoma, anal cancer, basal cell carcinoma, bile duct cancer, bladder cancer, bone cancer, osteosarcoma, malignant fibrous histiocytoma, brain stem glioma, brain cancer, craniopharyngioma, ependymoblastoma, ependymoma, medulloblastoma, medulloepithelioma, pineal parenchymal tumor, breast cancer, bronchial tumor, Burkitt lymphoma, Non-Hodgkin lymphoma, carcinoid tumor, cervical cancer, chordoma, chronic lymphocytic leukemia (CLL), chronic myelogenous leukemia (CML), colon cancer, colorectal cancer, cutaneous T-cell lymphoma, ductal carcinoma in situ, endometrial cancer, esophageal cancer, Ewing Sarcoma, eye cancer, intraocular melanoma, retinoblastoma, fibrous histiocytoma, gallbladder cancer, gastric cancer, glioma, hairy cell leukemia, head and neck cancer, heart cancer, hepatocellular (liver) cancer, Hodgkin lymphoma, hypopharyngeal cancer, kidney cancer, laryngeal cancer, lip cancer, oral cavity cancer, lung cancer, non-small cell carcinoma, small cell carcinoma, melanoma, mouth cancer, myelodysplastic syndromes, multiple myeloma, medulloblastoma, nasal cavity cancer, paranasal sinus cancer, neuroblastoma, nasopharyngeal cancer, oral cancer, oropharyngeal cancer, osteosarcoma, ovarian cancer, pancreatic cancer, papillomatosis, paraganglioma, parathyroid cancer, penile cancer, pharyngeal cancer, pituitary tumor, plasma cell neoplasm, prostate cancer, rectal cancer, renal cell cancer, rhabdomyosarcoma, salivary gland cancer, Sezary syndrome, skin cancer, nonmelanoma, small intestine cancer, soft tissue sarcoma, squamous cell carcinoma, testicular cancer, throat cancer, thymoma, thyroid cancer, urethral cancer, uterine cancer, uterine sarcoma, vaginal cancer, vulvar cancer, Waldenstrom macroglobulinemia, and/or Wilms Tumor.

A haploid human genome equivalent has about 3 picograms of DNA. A sample of about 1 microgram of DNA contains about 300,000 haploid human genome equivalents. Improvements in sequencing can be achieved as long as at least some of the duplicate or cognate polynucleotides bear unique identifiers with respect to each other, that is, bear different tags. However, in certain embodiments, the number of tags used is selected so that there is at least a 95% chance that all duplicate molecules starting at any one position bear unique identifiers. For example, in a sample comprising about 10,000 haploid human genome equivalents of fragmented genomic DNA, e.g., cfDNA, z is expected to be between 2 and 8. Such a population can be tagged with between about 10 and 100 different identifiers, for example, about 2 identifiers, about 4 identifiers, about 9 identifiers, about 16 identifiers, about 25 identifiers, about 36 different identifiers, about 49 different identifiers, about 64 different identifiers, about 81 different identifiers, or about 100 different identifiers.

Nucleic acid barcodes having identifiable sequences including molecular barcodes, can be used for tagging. For example, a plurality of DNA barcodes can comprise various numbers of sequences of nucleotides. A plurality of DNA barcodes having 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more identifiable sequences of nucleotides can be used. When attached to only one end of a polynucleotide, the plurality of DNA barcodes can produce 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more different identifiers. Alternatively, when attached to both ends of a polynucleotide, the plurality DNA barcodes can produce 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 324, 361, 400 or more different identifiers (which is the ^2 of when the DNA barcode is attached to only 1 end of a polynucleotide). In one example, a plurality of DNA barcodes having 6, 7, 8, 9 or 10 identifiable sequences of nucleotides can be used. When attached to both ends of a polynucleotide, they produce 36, 49, 64, 81 or 100 possible different identifiers, respectively. In a particular example, the plurality of DNA barcodes can comprise 8 identifiable sequences of nucleotides. When attached to only one end of a polynucleotide, the plurality of DNA barcodes can produce 8 different identifiers. Alternatively, when attached to both ends of a polynucleotide, the

plurality of DNA barcodes can produce 64 different identifiers. Samples tagged in such a way can be those with a range of about 10 ng to any of about 100 ng, about 1 μg, about 10 μg of fragmented polynucleotides, e.g., genomic DNA, e.g., cfDNA.

A polynucleotide can be uniquely identified in various ways. A polynucleotide can be uniquely identified by a unique DNA barcode. For example, any two polynucleotides in a sample are attached two different DNA barcodes. Alternatively, a polynucleotide can be uniquely identified by the combination of a DNA barcode and one or more endogenous sequences of the polynucleotide. For example, any two polynucleotides in a sample can be attached the same DNA barcode, but the two polynucleotides can still be identified by different endogenous sequences. The endogenous sequence can be on an end of a polynucleotide. For example, the endogenous sequence can be adjacent (e.g., base in between) to the attached DNA barcode. In some instances the endogenous sequence can be at least 2, 4, 6, 8, 10, 20, 30, 40, 50, 60, 70, 80, 90, or 100 bases in length. Preferably, the endogenous sequence is a terminal sequence of the fragment/polynucleotides to be analyzed. The endogenous sequence may be the length of the sequence. For example, a plurality of DNA barcodes comprising 8 different DNA barcodes can be attached to both ends of each polynucleotide in a sample. Each polynucleotide in the sample can be identified by the combination of the DNA barcodes and about 10 base pair endogenous sequence on an end of the polynucleotide. Without being bound by theory, the endogenous sequence of a polynucleotide can also be the entire polynucleotide sequence.

Also disclosed herein are compositions of tagged polynucleotides. The tagged polynucleotide can be single-stranded. Alternatively, the tagged polynucleotide can be double-stranded (e.g., duplex-tagged polynucleotides). Accordingly, this invention also provides compositions of duplex-tagged polynucleotides. The polynucleotides can comprise any types of nucleic acids (DNA and/or RNA). The polynucleotides comprise any types of DNA disclosed herein. For example, the polynucleotides can comprise DNA, e.g., fragmented DNA or cfDNA. A set of polynucleotides in the composition that map to a mappable base position in a genome can be non-uniquely tagged, that is, the number of different identifiers can be at least 2 and fewer than the number of polynucleotides that map to the mappable base position. The number of different identifiers can also be at least 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 and fewer than the number of polynucleotides that map to the mappable base position.

In some instances, as a composition goes from about 1 ng to about 10 μg or higher, a larger set of different molecular barcodes can be used. For example, between 5 and 100 different library adaptors can be used to tag polynucleotides in a cfDNA sample.

The systems and methods disclosed herein may be used in applications that involve the assignment of molecular barcodes. The molecular barcodes can be assigned to any types of polynucleotides disclosed in this invention. For example, the molecular barcodes can be assigned to cell-free polynucleotides (e.g., cfDNAs). Often, an identifier disclosed herein can be a barcode oligonucleotide that is used to tag the polynucleotide. The barcode identifier may be a nucleic acid oligonucleotide (e.g., a DNA oligonucleotide). The barcode identifier can be single-stranded. Alternatively, the barcode identifier can be double-stranded. The barcode identifier can be attached to polynucleotides using any method disclosed herein. For example, the barcode identifier can be attached to the polynucleotide by ligation using an enzyme. The barcode identifier can also be incorporated into the polynucleotide through PCR. In other cases, the reaction may comprise addition of a metal isotope, either directly to the analyte or by a probe labeled with the isotope. Generally, assignment of unique or non-unique identifiers or molecular barcodes in reactions of this disclosure may follow methods and systems described by, for example, U.S. patent applications 2001/0053519, 2003/0152490, 2011/0160078 and U.S. Pat. No. 6,582,908, each of which is entirely incorporated herein by reference.

Identifiers or molecular barcodes used herein may be completely endogenous whereby circular ligation of individual fragments may be performed followed by random shearing or targeted amplification. In this case, the combination of a new start and stop point of the molecule and the original intramolecular ligation point can form a specific identifier.

Identifiers or molecular barcodes used herein can comprise any types of oligonucleotides. In some cases, identifiers may be predetermined, random, or semi-random sequence oligonucleotides. Identifiers can be barcodes. For example, a plurality of barcodes may be used such that barcodes are not necessarily unique to one another in the plurality. Alternatively, a plurality of barcodes may be used such that each barcode is unique to any other barcode in the plurality. The barcodes can comprise specific sequences (e.g., predetermined sequences) that can be individually tracked. Further, barcodes may be attached (e.g., by ligation) to individual molecules such that the combination of the barcode and the sequence it may be ligated to creates a specific sequence that may be individually tracked. As described herein, detection of barcodes in combination with sequence data of beginning (start) and/or end (stop) portions of sequence reads can allow assignment of a unique identity to a particular molecule. The length or number of base pairs of an individual sequence read may also be used to assign a unique identity to such a molecule. As described herein, fragments from a single strand of nucleic acid having been assigned a unique identity, may thereby permit subsequent identification of fragments from the parent strand. In this way the polynucleotides in the sample can be uniquely or substantially uniquely tagged. A duplex tag can include a degenerate or semi-degenerate nucleotide sequence, e.g., a random degenerate sequence. The nucleotide sequence can comprise any number of nucleotides. For example, the nucleotide sequence can comprise 1 (if using a non-natural nucleotide), 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more nucleotides. In a particular example, the sequence can comprise 7 nucleotides. In another example, the sequence can comprise 8 nucleotides. The sequence can also comprise 9 nucleotides. The sequence can comprise 10 nucleotides.

A barcode can comprise contiguous or non-contiguous sequences. A barcode that comprises at least 1, 2, 3, 4, 5 or more nucleotides is a contiguous sequence or non-contiguous sequence. if the 4 nucleotides are uninterrupted by any other nucleotide. For example, if a barcode comprises the sequence TTGC, a barcode is contiguous if the barcode is TTGC. On the other hand, a barcode is non-contiguous if the barcode is TTXGC, where X is a nucleic acid base.

An identifier or molecular barcode can have an n-mer sequence which may be 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45,

46, 47, 48, 49, 50 or more nucleotides in length. A tag herein can comprise any range of nucleotides in length. For example, the sequence can be between 2 to 100, 10 to 90, 20 to 80, 30 to 70, 40 to 60, or about 50 nucleotides in length.

The tag can comprise a double-stranded fixed reference sequence downstream of the identifier or molecular barcode. Alternatively, the tag can comprise a double-stranded fixed reference sequence upstream or downstream of the identifier or molecular barcode. Each strand of a double-stranded fixed reference sequence can be, for example, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 nucleotides in length.

C. Adaptors

A library of polynucleotide molecules can be synthesized for use in sequencing. For example, a library of polynucleotides comprising a plurality of polynucleotide molecules that are each less than or equal to 100, 90, 80, 70, 60, 50, 45, 40, or 35 nucleic acid (or nucleotide) bases in length can be made. A plurality of polynucleotide molecules can be each less than or equal to 35 nucleic acid bases in length. A plurality of polynucleotide molecules can be each less than or equal to 30 nucleic acid bases in length. A plurality of polynucleotide molecules can also be less than or equal to 250, 200, 150, 100, or 50 nucleic acid bases. Additionally, the plurality of polynucleotide molecules can also be less than or equal to 100, 99, 98, 97, 96, 95, 94, 93, 92, 91, 90, 89, 88, 87, 86, 85, 84, 83, 82, 81, 80, 79, 78, 77, 76, 75, 74, 73, 72, 71, 70, 69, 68, 67, 66, 65, 64, 63, 62, 61, 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 nucleic acid bases.

A library of polynucleotides comprising a plurality of polynucleotide molecules can also have distinct (with respect to each other) molecular barcode sequences (or molecular barcodes) with respect to at least 4 nucleic acid bases. A molecular barcode (also "barcode" or "identifier" herein) sequence is a nucleotide sequence that distinguishes one polynucleotide from another. In other embodiments, the polynucleotide molecules can also have different barcode sequences with respect to 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more nucleic acid bases.

A library of polynucleotides comprising a plurality of polynucleotide molecules can also have a plurality of different barcode sequences. For example, a plurality of polynucleotide molecules can have at least 4 different molecular barcode sequences. In some cases, the plurality of polynucleotide molecules has from 2-100, 4-50, 4-30, 4-20, or 4-10 different molecular barcode sequences. The plurality of polynucleotides molecules can also have other ranges of different barcode sequences such as, 1-4, 2-5, 3-6, 4-7, 5-8, 6-9, 7-10, 8-11, 9-12, 10-13, 11-14, 12-15, 13-16, 14-17, 15-18, 16-19, 17-20, 18-21, 19-22, 20-23, 21-24, or 22-25 different barcode sequences. In other cases, a plurality of polynucleotide molecules can have at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or 100 more different barcode sequences. In a particular example, the plurality library adapters comprise at least 8 different sequences.

The location of the different barcode sequences can vary within the plurality of polynucleotides. For example, the different barcode sequences can be within 20, 15, 10, 9, 8, 7, 6, 5, 4, 3, or 2 nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules. In an example, a plurality of polynucleotide molecules has distinct barcode sequences that are within 10 nucleic acid bases from the terminal end. In another example, a plurality of polynucleotide molecules has distinct barcode sequences that are within 5 or 1 nucleic acid bases from the terminal end. In other instances, the distinct barcode sequences can be at the terminal end of a respective one of the plurality of polynucleotide molecules. Other variations include that the distinct molecular barcode sequences can be within 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, or more nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules.

The terminal end of the plurality of polynucleotide molecules can be adapted for ligation to a target nucleic acid molecule. For example, the terminal end can be a blunt end. In some other cases, the terminal end is adapted for hybridization to a complementary sequence of a target nucleic acid molecule.

A library of polynucleotides comprising a plurality of polynucleotide molecules can also have an edit distance of at least 1. In some cases, the edit distance is with respect to individual bases of the plurality of polynucleotide molecules. In other cases, the plurality of polynucleotide molecules can have an edit distance of at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more. The edit distance can be a Hamming distance.

In some cases, the plurality of polynucleotides does not contain sequencing adaptors. A sequence adaptor can be a polynucleotide that comprises a sequence that hybridizes to one or more sequencing adaptors or primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, e.g., a flow cell sequence. The term "flow cell sequence" and its grammatical equivalents as used herein, refers to a sequence that permits hybridization to a substrate, for example, by way of a primer attached to the substrate. The substrate can be bead or a planar surface. In some embodiments, a flow cell sequence can allow a polynucleotide to attach to a flow cell or surface (e.g., surface of a bead, for example, an Illumina flow cell.

When a plurality of polynucleotide molecules does not contain sequencing adaptors or primers, each polynucleotide molecule of the plurality does not contain a nucleic acid sequence or other moiety that is adapted to permit sequencing of a target nucleic acid molecule with a given sequencing approach, such as Illumina, SOLiD, Pacific Biosciences, GeneReader, Oxford Nanopore, Complete Genomics, Gnu-

Bio, Ion Torrent, Oxford Nanopore or Genia. In some examples, when a plurality of polynucleotide molecules does not contain sequencing adaptors or primers, the plurality of polynucleotide molecules does not contain flow cell sequences. For example, the plurality of polynucleotide molecules cannot bind to flow cells, such as used in Illumina flow cell sequencers. However, these flow cell sequences, if desired, can be added to the plurality of polynucleotide molecules by methods such as PCR amplification or ligation. At this point, Illumina flow cell sequencers can be used. Alternatively, when the plurality of polynucleotide molecules does not contain sequencing adaptors or primers, the plurality of polynucleotide molecules does not contain hairpin shaped adaptors or adaptors for generating hairpin loops in a target nucleic acid molecule, such as Pacific Bioscience SMRTbell™ adaptors. However, these hairpin shaped adaptors, if desired, can be added to the plurality of polynucleotide molecules by methods such as PCR amplification or ligation. The plurality of polynucleotide molecules can be circular or linear.

A plurality of polynucleotide molecules can be double stranded. In some cases, the plurality of polynucleotide molecules can be single stranded, or can comprise hybridized and non-hybridized regions. A plurality of polynucleotide molecules can be non-naturally occurring polynucleotide molecules.

Adaptors can be polynucleotide molecules. The polynucleotide molecules can be Y-shaped, bubble-shaped or hairpin-shaped. A hairpin adaptor may contain a restriction site(s) or a Uracil containing base. Adaptors can comprise a complementary portion and a non-complementary portion. The non-complementary portion can have an edit distance (e.g., Hamming distance). For example, the edit distance can be at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, or at least 30. The complementary portion of the adaptor can comprise sequences that are selected to enable and/or promote ligation to a polynucleotide, e.g., a sequence to enable and/or promote ligation to a polynucleotide at a high yield.

A plurality of polynucleotide molecules as disclosed herein can be purified. In some cases, a plurality of polynucleotide molecules as disclosed herein can be isolated polynucleotide molecules. In other cases, a plurality of polynucleotide molecules as disclosed herein can be purified and isolated polynucleotide molecules.

In certain aspects, each of the plurality of polynucleotide molecules is Y-shaped or hairpin-shaped. Each of the plurality of polynucleotide molecules can comprise a different barcode. The different barcode can be a randomer in the complementary portion (e.g., double stranded portion) of the Y-shaped or hairpin-shaped adaptor. Alternatively, the different barcode can be in one strand of the non-complementary portion (e.g., one of the Y-shaped arms). As discussed above, the different barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, e.g., 7 bases. The barcode can be contiguous or non-contiguous sequences, as described above. The plurality of polynucleotide molecules is from 10 nucleic acid bases to 35 nucleic acid bases (or any length as described above) in length. Further, the plurality of polynucleotide molecules can comprise an edit distance (as described above), that is a Hamming distance. A plurality of polynucleotide molecules can have distinct barcode sequences that are within 10 nucleic acid bases from the terminal end.

In another aspect, a plurality of polynucleotide molecules can be sequencing adaptors. A sequencing adaptor can comprise a sequence hybridizing to one or more sequencing primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, e.g., a flow cell sequence. For example, a sequencing adaptor can be a flow cell adaptor. The sequencing adaptors can be attached to one or both ends of a polynucleotide fragment. In another example, a sequencing adaptor can be hairpin shaped. For example, the hairpin shaped adaptor can comprise a complementary double-stranded portion and a loop portion, where the double-stranded portion can be attached (e.g., ligated) to a double-stranded polynucleotide. Hairpin shaped sequencing adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. A sequencing adaptor can be up to 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, or more bases from end to end. For example, a sequencing adaptor can be up to 70 bases from end to end. The sequencing adaptor can comprise 20-30, 20-40, 30-50, 30-60, 40-60, 40-70, 50-60, 50-70, bases from end to end. In a particular example, the sequencing adaptor can comprise 20-30 bases from end to end. In another example, the sequencing adaptor can comprise 50-60 bases from end to end. A sequencing adaptor can comprise one or more barcodes. For example, a sequencing adaptor can comprise a sample barcode. The sample barcode can comprise a pre-determined sequence. The sample barcodes can be used to identify the source of the polynucleotides. The sample barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, e.g., at least 8 bases. The barcode can be contiguous or non-contiguous sequences, as described above.

The plurality of polynucleotide molecules as described herein can be used as adaptors. Adaptors can comprise one or more identifiers. An adaptor can comprise an identifier with a random sequence. Alternatively, an adaptor can comprise an identifier with pre-determined sequences. Some adaptors can comprise an identifier with a random sequence and another identifier with a pre-determined sequence. The adaptors comprising identifiers can be double-stranded or single-stranded adaptors. The adaptors comprising identifiers can be Y-shaped adaptors. A Y-shaped adaptor can comprise one or more identifiers with a random sequence. The one or more identifiers can be on the hybrid portion and/or non-hybridized portion of the Y-shaped adaptor. A Y-shaped adaptor can comprise one or more identifiers with a pre-determined sequence. The one or more identifiers with pre-determined sequence can be on the hybridized portion and/or non-hybridized portion of the Y-shaped adaptor. A Y-shaped adaptor can comprise one or more identifiers with a random sequence and one or more identifiers with a pre-determined sequence. For example, the one or more identifiers with a random sequence can be on the hybridized portion of the Y-shaped adaptor and/or the non-hybridized portion of the Y-shaped adaptor. The one or more identifiers with a pre-determined sequence can be on the hybridized portion of the Y-shaped adaptor and/or the non-hybridized

portion of the Y-shaped adaptor. In a particular example, a Y-shaped adaptor can comprise an identifier with a random sequence on its hybridized portion and an identifier with a pre-determined sequence on its non-hybridized portion. The identifiers can be in any length disclosed herein. For example, a Y-shaped adaptor can comprise an identifier with a random sequence of 7 nucleotides on its hybridized portion and an identifier with a pre-determined sequence of 8 nucleotides on its non-hybridized portion.

An adaptor can include a double-stranded portion with a molecular barcode and at least one or two single-stranded portion. For example, the adaptor can be Y-shaped and include a double-stranded portion and two single-stranded portions. The single-stranded portions can include sequences that are not complementary to one another.

The adaptor can include a terminal end that has a sequence that is selected to permit the adaptor to be efficiently (e.g., at an efficiency of at least about 20%, 30%, 40%, 50%) ligated or otherwise coupled to a polynucleotide. In some examples, terminal nucleotides in a double-stranded portion of an adaptor are selected from a combination of purines and pyrimidines to provide for efficient ligation.

In some examples, a set of library adaptors comprises a plurality of polynucleotide molecules (library adaptors) with molecular barcodes. The library adaptors are less than or equal to 80, 70, 60, 50, 45, or 40 nucleotide bases in length. The molecular barcodes can be at least 4 nucleotide bases in length, but may be from 4 to 20 nucleotide bases in length. The molecular barcodes can be different from one another and have an edit distance of at least 1, 2, 3, 4, or 5 between one another. The molecular barcodes are located at least 1, 2, 3, 4, 5, 10, or 20 nucleotide bases away from a terminal end of their respective library adaptors. In some cases, the at least one terminal base is identical in all of the library adaptors.

The library adaptors can be identical but for the molecular barcodes. For example, the library adaptors can have identical sequences but differ only with respect to nucleotide sequences of the molecular barcodes.

Each of the library adaptors can have a double stranded portion and at least one single-stranded portion. By "single stranded portion" is meant an area of non-complementarity or an overhang. In some cases, each of the library adaptors has a double-stranded portion and two single-stranded portions. The double-stranded portion can have a molecular barcode. In some cases, the molecular barcode is a randomer. Each of the library adaptors can further include a strand-identification barcode on a single-stranded portion. The strand-identification barcode can include at least 4 nucleotide bases, in some cases from 4 to 20 nucleotide bases.

In some examples, each of the library adaptors has a double-stranded portion with a molecular barcode and two single-stranded portions. The single-stranded portions may not hybridize to one another. The single-stranded portions may not be completely complementary to one another.

The library adaptors can have a sequence of terminal nucleotides in a double-stranded portion that are the same. The sequence of terminal nucleotides can be at least 2, 3, 4, 5 or 6 nucleotide bases in length. For example, one strand of a double-stranded portion of the library adaptor can have the sequence ACTT, TCGC, or TACC at the terminal end, while the other strand can have a complementary sequence. In some cases, such a sequence is selected to optimize the efficiency at which the library adaptors ligate to target polynucleotides. Such sequences can be selected to optimize a binding interaction between the ends of the library adaptors and the target polynucleotides.

In some cases, none of the library adaptors contains a sample identification motif (or sample molecular barcode). Such sample identification motif can be provided via sequencing adaptors. A sample identification motif can include a sequencer of at least 4, 5, 6, 7, 8, 9, 10, 20, 30, or 40 nucleotide bases that permits the identification of polynucleotide molecules from a given sample from polynucleotide molecules from other samples. For example, this can permit polynucleotide molecules from two subjects to be sequenced in the same pool and sequence reads for the subjects subsequently identified.

A sequencer motif includes nucleotide sequence(s) needed to couple a library adaptor to a sequencing system and sequence a target polynucleotide coupled to the library adaptor. The sequencer motif can include a sequence that is complementary to a flow cell sequence and a sequence (sequencing initiation sequence) that is selectively hybridizable to a primer (or priming sequence) for use in sequencing. For example, such sequencing initiation sequence can be complementary to a primer that is employed for use in sequence by synthesis (e.g., Illumina). Such primer can be included in a sequencing adaptor. A sequencing initiation sequence can be a primer hybridization site.

In some cases, none of the library adaptors contains a complete sequencer motif. The library adaptors can contain partial or no sequencer motifs. In some cases, the library adaptors include a sequencing initiation sequence. The library adaptors can include a sequencing initiation sequence but no flow cell sequence. The sequence initiation sequence can be complementary to a primer for sequencing. The primer can be a sequence specific primer or a universal primer. Such sequencing initiation sequences may be situated on single-stranded portions of the library adaptors. As an alternative, such sequencing initiation sequences may be priming sites (e.g., kinks or nicks) to permit a polymerase to couple to the library adaptors during sequencing.

In some cases, partial or complete sequencer motifs are provided by sequencing adaptors. A sequencing adaptor can include a sample molecular barcode and a sequencer motif. The sequencing adaptors can be provided in a set that is separate from the library adaptors. The sequencing adaptors in a given set can be identical—i.e., they contain the same sample barcode and sequencer motif.

Sequencing adaptors can include sample identification motifs and sequencer motifs. Sequencer motifs can include primers that are complementary to a sequencing initiation sequence. In some cases, sequencer motifs also include flow cell sequences or other sequences that permit a polynucleotide to a configured or arranged in a manner that permits the polynucleotide to be sequenced by a sequencer.

Library adaptors and sequencing adaptors can each be partial adaptors, that is, containing part but not all of the sequences necessary to enable sequencing by a sequencing platform. Together they provide complete adaptors. For example, library adaptors can include partial or no sequencer motifs, but such sequencer motifs are provided by sequencing adaptors.

FIGS. **9**A-**9**C schematically illustrate a method for tagging a target polynucleotide molecule with library adaptors. FIG. **9**A shows a library adaptor as a partial adaptor containing a primer hybridization site on one of the strands and a molecular barcode towards another end. The primer hybridization site can be a sequencing initiation sequence for subsequent sequencing. The library adaptor is less than or equal to 80 nucleotide bases in length. In FIG. **9**B, the

library adaptors are ligated at both ends of the target polynucleotide molecule to provide a tagged target polynucleotide molecule. The tagged target polynucleotide molecule may be subjected to nucleic acid amplification to generate copies of the target. Next, in FIG. **9**C, sequencing adaptors containing sequencer motifs are provided and hybridized to the tagged target polynucleotide molecule. The sequencing adaptors contain sample identification motifs. The sequencing adaptors can contain sequences to permit sequencing of the tagged target with a given sequencer.

D. Sequencing

Tagged polynucleotides can be sequenced to generate sequence reads (e.g., as shown in step (**106**), FIG. **1**). For example, a tagged duplex polynucleotide can be sequenced. Sequence reads can be generated from only one strand of a tagged duplex polynucleotide. Alternatively, both strands of a tagged duplex polynucleotide can generate sequence reads. The two strands of the tagged duplex polynucleotide can comprise the same tags. Alternatively, the two strands of the tagged duplex polynucleotide can comprise different tags. When the two strands of the tagged duplex polynucleotide are differently tagged, sequence reads generated from one strand (e.g., a Watson strand) can be distinguished from sequence reads generated from the other strands (e.g., a Crick strand). Sequencing can involve generating multiple sequence reads for each molecule. This occurs, for example, as a result the amplification of individual polynucleotide strands during the sequencing process, e.g., by PCR.

Methods disclosed herein can comprise amplifying of polynucleotides. Polynucleotides amplification can result in the incorporation of nucleotides into a nucleic acid molecule or primer thereby forming a new nucleic acid molecule complementary to a template nucleic acid. The newly formed polynucleotide molecule and its template can be used as templates to synthesize additional polynucleotides. The polynucleotides being amplified can be any nucleic acids, for example, deoxyribonucleic acids, including genomic DNAs, cDNAs (complementary DNA), cfDNAs, and circulating tumor DNAs (ctDNAs). The polynucleotides being amplified can also be RNAs. As used herein, one amplification reaction may comprise many rounds of DNA replication. DNA amplification reactions can include, for example, polymerase chain reaction (PCR). One PCR reaction may comprise 2-100 "cycles" of denaturation, annealing, and synthesis of a DNA molecule. For example, 2-7, 5-10, 6-11, 7-12, 8-13, 9-14, 10-15, 11-16, 12-17, 13-18, 14-19, or 15-20 cycles can be performed during the amplification step. The condition of the PCR can be optimized based on the GC content of the sequences, including the primers.

Nucleic acid amplification techniques can be used with the assays described herein. Some amplification techniques are the PCR methodologies which can include, but are not limited to, solution PCR and in situ PCR. For example, amplification may comprise PCR-based amplification. Alternatively, amplification may comprise non PCR-based amplification. Amplification of the template nucleic acid may comprise use of one or more polymerases. For example, the polymerase may be a DNA polymerase or an RNA polymerase. In some cases, high fidelity amplification is performed such as with the use of high fidelity polymerase (e.g., Phusion® High-Fidelity DNA Polymerase) or PCR protocols. In some cases, the polymerase may be a high fidelity polymerase. For example, the polymerase may be KAPA HiFi DNA polymerase. The polymerase may also be Phusion DNA polymerase. The polymerase may be used under reaction conditions that reduce or minimize amplification biases, e.g., due to fragment length, GC content, etc.

Amplification of a single strand of a polynucleotide by PCR will generate copies both of that strand and its complement. During sequencing, both the strand and its complement will generate sequence reads. However, sequence reads generated from the complement of, for example, the Watson strand, can be identified as such because they bear the complement of the portion of the duplex tag that tagged the original Watson strand. In contrast, a sequence read generated from a Crick strand or its amplification product will bear the portion of the duplex tag that tagged the original Crick strand. In this way, a sequence read generated from an amplified product of a complement of the Watson strand can be distinguished from a complement sequence read generated from an amplification product of the Crick strand of the original molecule.

All amplified polynucleotides can be submitted to a sequencing device for sequencing. Alternatively, a sampling, or subset, of all of the amplified polynucleotides is submitted to a sequencing device for sequencing. With respect to any original double-stranded polynucleotide there can be three results with respect to sequencing. First, sequence reads can be generated from both complementary strands of the original molecule (that is, from both the Watson strand and from the Crick strand). Second, sequence reads can be generated from only one of the two complementary strands (that is, either from the Watson strand or from the Crick strand, but not both). Third, no sequence read may be generated from either of the two complementary strands. Consequently, counting unique sequence reads mapping to a genetic locus will underestimate the number of double-stranded polynucleotides in the original sample mapping to the locus. Described herein are methods of estimating the unseen and uncounted polynucleotides.

The sequencing method can be massively parallel sequencing, that is, simultaneously (or in rapid succession) sequencing any of at least 100, 1000, 10,000, 100,000, 1 million, 10 million, 100 million, or 1 billion polynucleotide molecules. Sequencing methods may include, but are not limited to: high-throughput sequencing, pyrosequencing, sequencing-by-synthesis, single-molecule sequencing, nanopore sequencing, semiconductor sequencing, sequencing-by-ligation, sequencing-by-hybridization, RNA-Seq (Illumina), Digital Gene Expression (Helicos), Next generation sequencing, Single Molecule Sequencing by Synthesis (SMSS) (Helicos), massively-parallel sequencing, Clonal Single Molecule Array (Solexa), shotgun sequencing, Maxam-Gilbert or Sanger sequencing, primer walking, sequencing using PacBio, SOLiD, Ion Torrent, or Nanopore platforms and any other sequencing methods known in the art.

For example, duplex-tagged polynucleotides can be amplified, by for example PCR (see e.g., FIG. **4**A duplex-tagged polynucleotides are referred to as mm' and nn'). In FIG. **4**A, the strand of the duplex polynucleotide including sequence m bears sequence tags w and y, while the strand of the duplex polynucleotide including sequence m' bears sequence tags x and z. Similarly, the strand of the duplex polynucleotide including sequence n bears sequence tags a and c, while the strand of the duplex polynucleotide including sequence n' bears sequence tags b and d. During amplification, each strand produces itself and its complementary sequence. However, for example, an amplification progeny of original strand m that includes the complementary sequence, m', is distinguishable from an amplification progeny of original strand m' because the progeny from original

strand m will have the sequence 5'-y'm'w'-3' and the progeny of the original m' strand one strand will have the sequence 5'-zm'x-3'. FIG. **4**B shows amplification in more detail. During amplification, errors can be introduced into the amplification progeny, represented by dots. The application progeny are sampled for sequencing, so that not all strands produce sequence reads, resulting in the sequence reads indicated. Because sequence reads can come from either of a strand or its complement, both sequences and complement sequences will be included in the set of sequence reads. It should be noted that it is possible that a polynucleotide would bear the same tag on each end. Thus, for a tag "a", and polynucleotide "m", a first strand could be tagged a-m-a', and the complement could be tagged a-m'-a.

E. Determining Consensus Sequence Reads

Methods disclosed herein can comprise determining consensus sequence reads in sequence reads (e.g., as shown in step (**108**), FIG. **1**), such as by reducing or tracking redundancy. Sequencing of amplified polynucleotides can produce reads of the several amplification products from the same original polynucleotide, referred to as "redundant reads". By identifying redundant reads, unique molecules in the original sample can be determined. If the molecules in a sample are uniquely tagged, then reads generated from amplification of a single unique original molecule can be identified based on their distinct barcode. Ignoring barcodes, reads from unique original molecules can be determined based on sequences at the beginning and end of a read, optionally in combination with the length of the read. In certain cases, however, a sample may be expected to have a plurality of original molecules having the same start stop sequences and the same length. Without barcoding, these molecules are difficult to distinguish from one another. However, if a collection of polynucleotides is non-uniquely tagged (that is, an original molecule shares the same identifier with at least one other original molecule), combining information from a barcode with start/stop sequence and/or polynucleotide length significantly increases the probability that any sequence read can be traced back to an original polynucleotide. This is because, in part, even without unique tagging, it is unlikely that any two original polynucleotides having the same start/stop sequence and length also will be tagged with the same identifier.

F. Collapsing

Collapsing allows for reduction in noise (i.e., background) that is generated at each step of the process. Methods disclosed herein can comprise collapsing, e.g., generating a consensus sequence by comparing multiple sequence reads. For example, sequence reads generated from a single original polynucleotide can be used to generate a consensus sequence of that original polynucleotide. Iterative rounds of amplification can introduce errors into progeny polynucleotides. Also, sequencing typically may not be performed with perfect fidelity so sequencing errors are introduced at this stage as well. However, comparison of sequence reads of molecules derived from a single original molecule, including those that have sequence variants, can be analyzed so as to determine the original, or "consensus" sequence. This can be done phylogenetically. Consensus sequences can be generated from families of sequence reads by any of a variety of methods. Such methods include, for example, linear or non-linear methods of building consensus sequences (such as voting (e.g., biased voting), averaging, statistical, maximum a posteriori or maximum likelihood detection, dynamic programming, Bayesian, hidden Markov or support vector machine methods, etc.) derived from digital communication theory, information theory, or bioinformatics. For example, if all or most of the sequence reads tracking back to an original molecule bear the same sequence variant, that variant probably existed in the original molecule. On the other hand, if a sequence variant exists in a subset of redundant sequence reads, that variant may have been introduced during amplification/sequencing and represents an artifact not existing in the original. Furthermore, if only sequence reads derived from the Watson or Crick strand of an original polynucleotide contain the variant, the variant may have been introduced through single-sided DNA damage, first-cycle PCR error or through contaminating polynucleotides that were amplified from a different sample.

After fragments are amplified and the sequences of amplified fragments are read and aligned, the fragments are subjected to base calling, e.g., determining for each locus the most likely nucleotide. However, variations in the number of amplified fragments and unseen amplified fragments (e.g., those without being read their sequences; reasons could be too many such as amplification errors, sequencing reading errors, too long, too short, being chopped, etc.) may introduce errors in base calling. If there are too many unseen amplified fragments with respect to the seen amplified fragments (amplified fragments actually being read), the reliability of base calling may be diminished.

Therefore, disclosed herein is a method to correct for the number of unseen fragments in base calling. For example, when base calling for locus A (an arbitrary locus), it is first assumed that there are N amplified fragments. The sequence readouts can come from two types of fragments: double-strand fragments and single-strand fragments. Therefore, we assign N1, N2, and N3 as the numbers of double-strands, single-strands, and unseen fragments, respectively. Thus, N=N1+N2+N3 (N1 and N2 are known from the sequence readouts, and N and N3 are unknown). If the formula is solved for N (or N3), then N3 (or N) will be inferred.

Probability is used to estimate N. For example, we assign "p" to be the probability of having detected (or having read) a nucleotide of locus A in a sequence readout of a single-strand.

For sequence readouts from double-strands, the nucleotide call from a double-strand amplified fragment has a probability of p*p=pˆ2, seeing all N1 double-strands has the following equation: N1=N*(pˆ2).

For sequence readouts from a single-strand. Assuming that one of the 2 strands is seen, and the other is unseen, the probability of seeing one strand is "p", but the probability of missing the other strand is (1−p). Furthermore, by not distinguishing the single strand sourcing from 5-primer and sourcing from 3-primer, there is a factor of 2. Therefore, the nucleotide call from a single-strand amplified fragment has a probability 2×p×(1−p). Thus, seeing all N2 single-strands has the following equation: N2=N×2×p×(1−p).

"p" is also unknown. To solve p, the ratio of N1 to N2 is used to solve for "p":

$$R=N1/N2=Np^2/2Np(1-p)=p^2/2p(1-p)=p/2(1-p)$$

Once "p" is found, N can be found. After N is found, can be found N3=N−N1−N2.

Besides the ratio of paired versus unpaired strands (which is a measure post-collapsing), there is useful information in the pre-collapsing read depth at each locus. This information can be used to further improve the call for total molecule count and/or increase confidence of calling variants.

For example, FIG. **4**C demonstrates sequence reads corrected for complementary sequences. Sequences generated from an original Watson strand or an original Crick strand

can be differentiated on the basis of their duplex tags. Sequences generated from the same original strand can be grouped. Examination of the sequences can allow one to infer the sequence of the original strand (the "consensus sequence"). In this case, for example, the sequence variant in the nn' molecule is included in the consensus sequence because it included in every sequence read while other variants are seen to be stray errors. After collapsing sequences, original polynucleotide pairs can be identified based on their complementary sequences and duplex tags.

FIG. **5** demonstrates increased confidence in detecting sequence variants by pairing reads from Watson and Crick strands. Sequence nn' can include a sequence variant indicated by a dot. In some cases, sequence pp' does not include a sequence variant. Amplification, sequencing, redundancy reduction and pairing can result in both Watson and Crick strands of the same original molecule including the sequence variant. In contrast, as a result of errors introduced during amplification and sampling during sequencing, the consensus sequence of the Watson strand p can contain a sequence variant, while the consensus sequence of the Crick strand p' does not. It is less likely that amplification and sequencing will introduce the same variant into both strands (nn' sequence) of a duplex than onto one strand (pp' sequence). Therefore, the variant in the pp' sequence is more likely to be an artifact, and the variant in the nn' sequence is more likely to exist in the original molecule.

Methods disclosed herein can be used to correct errors resulted from experiments, e.g., PCR, amplification, and/or sequencing. For example, such a method can comprises attaching one or more double stranded adaptors to both ends of a double stranded polynucleotide, thereby providing a tagged double stranded polynucleotide; amplifying the double stranded tagged polynucleotide; sequencing both strands of the tagged polynucleotide; comparing the sequence of one strand with its complement to determine any errors introduced during sequencing; and correcting errors in the sequence based on (d). The adaptors used in this method can be any adaptors disclosed herein, e.g., Y-shaped adaptors. The adaptor can comprise any barcodes (e.g., distinct barcodes) disclosed herein.

G. Mapping

Sequence reads or consensus sequences can be mapped to one or more selected genetic loci (e.g., as shown step (**110**), FIG. **1**). A genetic locus can be, for example, a specific nucleotide position in the genome, a sequence of nucleotides (for example, an open reading frame), a fragment of a chromosome, a whole chromosome, or an entire genome. A genetic locus can be a polymorphic locus. Polymorphic locus can be a locus at which sequence variation exists in the population and/or exists in a subject and/or a sample. A polymorphic locus can be generated by two or more distinct sequences coexisting at the same location of the genome. The distinct sequences can differ from one another by one or more nucleotide substitutions, a deletion/insertion, and/or a duplication of any number of nucleotides, generally a relatively small number of nucleotides, such as less than 50, 45, 40, 35, 30, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1 nucleotide(s), among others. A polymorphic locus can be created by a single nucleotide position that varies within the population, e.g. a single nucleotide variation (SNV) or a single nucleotide polymorphism (SNP).

A reference genome for mapping can include the genome of any species of interest. Human genome sequences useful as references can include the hg19 assembly or any previous or available hg assembly. Such sequences can be interrogated using the genome browser available at genome.ucsc.edu/index.html. Other species genomes include, for example PanTro2 (chimp) and mm9 (mouse).

In methods disclosed herein, collapsing can be performed before or after mapping. In some aspects, collapsing can be performed before mapping. For example, sequence reads can be grouped into families based on their tags and one or more endogenous sequences, without regard to where the reads map in the genome. Then, the members of a family can be collapsed into a consensus sequence. The consensus sequence can be generated using any collapsing method disclosed herein. Then the consensus sequence can be mapped to locations in the genome. Reads mapped to a locus can be quantified (e.g., counted). Percentage of reads carrying a mutation at a locus can also be determined. Alternatively, collapsing can be performed after mapping. For example, all reads can first be mapped to the genome. Then the reads can be grouped into families based on their tags and one or more endogenous sequences. Since the reads have been mapped to the genome, consensus bases can be determined for each family at each locus. In other aspects, consensus sequence can be generated for one strand of a DNA molecule (e.g., for a Watson strand or a Crick strand). Mapping can be performed before or after the consensus sequence for one strand of the DNA molecule is determined. Numbers of Doublets and Singlets can be determined. These numbers can be used to calculate unseen molecules. For example, the unseen molecules can be calculated using the following equation: $N=D+S+U$; $D=Np(2)$, $S=N2pq$, where $p=1-q$, where p is the probability of seeing; q is the probability of missing a strand.

H. Grouping

Methods disclosed herein can also comprise grouping sequence reads. Sequence reads can be grouped based on various types of sequences, e.g., sequences of an oligonucleotide tag (e.g., a barcode), sequence of a polynucleotide fragments, or any combinations. For example, as shown in step (**112**) (FIG. **1**), sequence reads can be grouped as follows: Sequence reads generated from a "Watson" strand and those generated from a "Crick" strand of a double-stranded polynucleotide in the sample are identifiable based on the duplex tags that they bear. In this way, a sequence read or consensus sequence from a Watson strand of a duplex polynucleotide can be paired with a sequence read or consensus sequence from its complementary Crick strand. Paired sequence reads are referred to as a "Pair".

Sequence reads for which no sequence read corresponding to a complementary strand can be found among the sequence reads are termed "Singlets".

Double-stranded polynucleotides for which a sequence read for neither of the two complementary strands has been generated are referred to as "Unseen" molecules.

I. Quantifying

Methods disclosed herein also comprise quantifying sequence reads. For example, as shown in step (**114**) (FIG. **1**), Pairs and Singlets mapping to a selected genetic locus, or to each of a plurality of selected genetic loci, are quantified, e.g., counted.

The quantifying can comprise estimating number of polynucleotides in the sample (e.g., Pairs polynucleotides, Singlets polynucleotides, or Unseen polynucleotides. For example, as shown in step (**116**) (FIG. **1**), the number of double-stranded polynucleotides in the sample for which no sequence reads were generated ("Unseen" polynucleotides) is estimated. The probability that a double strand polynucleotide generates no sequence reads can be determined based

on the relative number of Pairs and Singlets at any locus. Using this probability, the number of Unseen polynucleotide can be estimated.

In step (**118**) an estimate for the total number of double-stranded polynucleotides in a sample mapping to a selected locus is the sum of the number of Pairs, the number of Singlets and the number of Unseen molecules mapping to the locus.

The number of Unseen original molecules in a sample can be estimated based on the relative number of Pairs and Singlets (FIG. **2**). Referring to FIG. **2**, as an example, counts for a particular genomic locus, Locus A, are recorded, where 1000 molecules are paired and 1000 molecules are unpaired. Assuming a uniform probability, p, for an individual Watson or Crick strand to make it through the process subsequent to conversion, one can calculate the proportion of molecules that fail to make it through the process (Unseen) as follows: Let R=ratio of paired to unpaired molecules=1, so $R=1=p^2/(2p(1-p))$. This implies that p=2/3 and that the quantity of lost molecules is equal to $(1-p)^2=1/9$. Thus in this example, approximately 11% of converted molecules are lost and never detected. Consider another genomic locus, Locus B, in the same sample where 1440 molecules are paired and 720 are unpaired. Using the same method, we can infer the number of molecules that are lost, is only 4%. Comparing the two areas, it may be assumed that Locus A had 2000 unique molecules as compared to 2160 molecules in Locus B—a difference of almost 8%. However, by correctly adding in the lost molecules in each region, we infer there are 2000/(8/9)=2250 molecules in Locus A and 2160/0.96=2250 molecules in Locus B. Hence, the counts in both regions are actually equal. This correction and thus much higher sensitivity can be achievable by converting the original double-stranded nucleic acid molecules and bioinformatically keeping track of all those that are paired and unpaired at the end of the process. Similarly, the same procedure can be used to infer true copy number variations in regions that appear to have similar counts of observed unique molecules. By taking the number of unseen molecules into consideration in the two or more regions, the copy number variation becomes apparent.

In addition to using binomial distribution, other methods of estimating numbers of unseen molecules include exponential, beta, gamma or empirical distributions based on the redundancy of sequence reads observed. In the latter case, the distribution of read counts for paired and unpaired molecules can be derived from such redundancy to infer the underlying distribution of original polynucleotide molecules at a particular locus. This can often lead to a better estimation of the number of unseen molecules.

J. CNV Detection

Methods disclosed herein also comprise detecting CNV. For example, as shown in step (**120**) (FIG. **1**), once the total number of polynucleotides mapping to a locus is determined, this number can be used in standard methods of determining CNV at the locus. A quantitative measure can be normalized against a standard. The standard can be an amount of any polynucleotides. In one method, a quantitative measure at a test locus can be standardized against a quantitative measure of polynucleotides mapping to a control locus in the genome, such as gene of known copy number. Quantitative measures can be compared against the amount of nucleic acid in any sample disclosed herein. For example, in another method, the quantitative measure can be compared against the amount of nucleic acid in the original sample. For example, if the original sample contained 10,000 haploid gene equivalents, the quantitative measure can be compared against an expected measure for diploidy. In another method, the quantitative measure can be normalized against a measure from a control sample, and normalized measures at different loci can be compared.

In some cases, in which copy number variation analysis is desired, sequence data may be: 1) aligned with a reference genome; 2) filtered and mapped; 3) partitioned into windows or bins of sequence; 4) coverage reads counted for each window; 5) coverage reads can then be normalized using a stochastic or statistical modeling algorithm; 6) and an output file can be generated reflecting discrete copy number states at various positions in the genome. In other cases, in which rare mutation analysis is desired, sequence data may be 1) aligned with a reference genome; 2) filtered and mapped; 3) frequency of variant bases calculated based on coverage reads for that specific base; 4) variant base frequency normalized using a stochastic, statistical or probabilistic modeling algorithm; 5) and an output file can be generated reflecting mutation states at various positions in the genome.

After the sequence read coverage ratios have been determined, a stochastic modeling algorithm can be optionally applied to convert the normalized ratios for each window region into discrete copy number states. In some cases, this algorithm may comprise a Hidden Markov Model. In other cases, the stochastic model may comprise dynamic programming, support vector machine, Bayesian modeling, probabilistic modeling, trellis decoding, Viterbi decoding, expectation maximization, Kalman filtering methodologies, or neural networks.

Methods disclosed herein can comprise detecting SNVs, CNVs, insertions, deletions, and/or rearrangements at a specific region in a genome. The specific genomic region can comprise a sequence in a gene, such as ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDHL IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, or NTRK1.

In some cases, the method uses a panel which comprises exons of one or more genes. The panel can comprise introns of one or more genes as well. The panel can also comprise exons and introns of one or more genes. The one or more genes can be those disclosed above. The panel can comprise about 80,000 bases which cover a panel of genes. The panel can comprise about 1000, 2000, 3000, 4000, 5000, 10000, 15000, 20000, 25000, 30000, 35000, 40000, 45000, 50000, 55000, 60000, 65000, 70000, 75000, 80000, 85000, 90000, 95000, 100000, 105000, 110000, 115000, 120000, 125000, or more bases.

In some aspects, copy number of a gene can be reflected in the frequency of a genetic form of the gene in a sample. For example, in a healthy individual, no copy number variation is reflected in a variant in a gene in one chromosome (e.g., heterozygosity) being detected in about 50% of detected molecules in a sample. Also, in a healthy individual, duplication of a gene bearing a variant can be reflected in the variant being detected in about 66% of detected molecules in a sample. Accordingly, if the tumor burden in a DNA sample is 10%, the frequency of a somatic

mutation in a gene in one chromosome of cancer cells, without CNV, can be about 5%. The converse can be true in the case of aneuploidy.

The methods disclosed herein can be used to determine whether a sequence variant is more likely present in the germ line level or resulted from a somatic cell mutation, e.g., in a cancer cell. For example, a sequence variant in a gene detected at levels arguably consistent with heterozygosity in the germ line is more likely the product of a somatic mutation if CNV is also detected in that gene. In some cases, to the extent we expect that a gene duplication in the germ line bears a variant consistent with genetic dose (e.g., 66% for trisomy at a locus), detection gene amplification with a sequence variant dose that deviates significantly from this expected amount indicates that the CNV is more likely present as a result of somatic cell mutation.

The methods disclosed herein can also be used to infer tumor heterogeneity in a situation in which sequence variants in two genes are detected at different frequencies. For example, tumor heterogeneity can be inferred when two genes are detected at different frequencies but their copy numbers are relatively equal. Alternatively, tumor homogeneity can be inferred when the difference in frequency between two sequence variants is consistent with difference in copy number for the two genes. Thus, for example, if an EGFR variant is detected at 11% and a KRAS variant is detected at 5%, and no CNV is detected at these genes, the difference in frequency likely reflects tumor heterogeneity (e.g., all tumor cells carry an EGFR mutant and half the tumor cells also carry a KRAS mutant). Alternatively, if the EGFR gene carrying the mutant is detected at 2-times normal copy number, one interpretation is a homogenous population of tumor cells, each cell carrying a mutant in the EGFR and KRAS genes, but in which the KRAS gene is duplicated.

In response to chemotherapy, a dominant tumor form can eventually give way through Darwinian selection to cancer cells carrying mutants that render the cancer unresponsive to the therapy regimen. Appearance of these resistance mutants can be delayed through methods of this invention. In one embodiment of this method, a subject is subjected to one or more pulsed therapy cycles, each pulsed therapy cycle comprising a first period during which a drug is administered at a first amount and a second cycle during which the drug is administered at a second, reduced amount. The first period can be characterized by a tumor burden detected above a first clinical level. The second period can be characterized by a tumor burden detected below a second clinical level. First and second clinical levels can be different in different pulsed therapy cycles. For example, the first clinical level can be lower in succeeding cycles. A plurality of cycles can include at least 2, 3, 4, 5, 6, 7, 8 or more cycles. For example, the BRAF mutant V600E may be detected in polynucleotides of a disease cell at an amount indicating a tumor burden of 5% in cfDNA. Chemotherapy can commence with dabrafenib. Subsequent testing can show that the amount of the BRAF mutant in the cfDNA falls below 0.5% or to undetectable levels. At this point, dabrafenib therapy can stop or be significantly curtailed. Further subsequent testing may find that DNA bearing the BRAF mutation has risen to 2.5% of polynucleotides in cfDNA. At this point, dabrafenib therapy can be re-started, e.g., at the same level as the initial treatment. Subsequent testing may find that DNA bearing the BRAF mutation has decreased to 0.5% of polynucleotides in cfDNA. Again, dabrafenib therapy can be stopped or reduced. The cycle can be repeated a number of times.

A therapeutic intervention can also be changed upon detection of the rise of a mutant form resistant to an original drug. For example, cancers with the EGFR mutation L858R respond to therapy with erlotinib. However, cancers with the EGFR mutation T790M are resistant to erlotinib. However, they are responsive to ruxolitinib. A method of this invention involves monitoring changes in tumor profile and changing a therapeutic intervention when a genetic variant associated with drug resistance rises to a predetermined clinical level.

Methods disclosed in this invention can comprise a method of detecting disease cell heterogeneity from a sample comprising polynucleotides from somatic cells and disease cells, the method comprising: a) quantifying polynucleotides in the sample bearing a sequence variant at each of a plurality of genetic loci; b) determining CNV at each of the plurality of genetic loci; different relative amounts of disease molecules at a locus, wherein the CNV indicates a genetic dose of a locus in the disease cell polynucleotides; c) determining a relative measure of quantity of polynucleotides bearing a sequence variant at a locus per genetic dose at the locus for each of a plurality of the loci; and d) comparing the relative measures at each of the plurality of loci, wherein different relative measures indicates tumor heterogeneity. In the methods disclosed herein, the genetic dose can be determined on a total molecule basis. For example, if there are 1× total molecules at a first locus, and 1.2× molecules mapped to a second locus, then the genetic dose is 1.2. Variants at this locus can be divided by 1.2. In some aspects, the method disclosed herein can be used to detect any disease cell heterogeneity, e.g., tumor cell heterogeneity. The methods can be used to detect disease cell heterogeneity from a sample comprising any types of polynucleotides, e.g., cfDNA, genomic DNA, cDNA, or ctDNA. In the methods, the quantifying can comprise, for example, determining the number or relative amount of the polynucleotides. Determining CNV can comprise mapping and normalizing different relative amounts of total molecules to a locus.

In another aspect, in response to chemotherapy, a dominant tumor form can eventually give way through Darwinian selection to cancer cells carrying mutants that render the cancer unresponsive to the therapy regimen. Appearance of these resistance mutants can be delayed through methods disclosed throughout. The methods disclosed herein can comprise a method comprising: a) subjecting a subject to one or more pulsed therapy cycles, each pulsed therapy cycle comprising (i) a first period during which a drug is administered at a first amount and (ii) a second period during which the drug is administered at a second, reduced amount; wherein (A) the first period is characterized by a tumor burden detected above a first clinical level; and (B) the second period is characterized by a tumor burden detected below a second clinical level.

K. Sequence Variant Detection

Systems and methods disclosed herein can be used to detect sequence variants, e.g., SNVs. For example, a sequence variant can be detected from consensus sequences from multiple sequence reads, for example, from at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, at least 30, at least 31, at least 32, at least 33, at least 34, at least 35, at least 36, at least 37, at least 38, at least 39, at least 40, at least 41, at least 42, at least 43, at least 44, at least 45, at least 46, at least 47, at least 48, at least 49, at least 50, at

least 51, at least 52, at least 53, at least 54, at least 55, at least 56, at least 57, at least 58, at least 59, at least 60, at least 61, at least 62, at least 63, at least 64, at least 65, at least 66, at least 67, at least 68, at least 69, at least 70, at least 71, at least 72, at least 73, at least 74, at least 75, at least 76, at least 77, at least 78, at least 79, at least 80, at least 81, at least 82, at least 83, at least 84, at least 85, at least 86, at least 87, at least 88, at least 89, at least 90, at least 91, at least 92, at least 93, at least 94, at least 95, at least 96, at least 97, at least 98, at least 99, at least 100, at least 200, at least 300, at least 400, at least 500, at least 600, at least 700, at least 800, at least 900, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, at least 8000, at least 9000, at least 10000 or more sequence reads. A consensus sequence can be from sequence reads of a single strand polynucleotide. A consensus sequence can also be from sequence reads of one strand of a double-stranded polynucleotide (e.g., pairing reads). In an exemplary method, pairing reads allows one to identify with increased confidence the existence of a sequence variant in a molecule. For example, if both strands of a Pair include the same variant, one can be reasonably sure that the variant existed in the original molecule, as the chance that the same variant is introduced into both strands during amplification/sequencing is rare. In contrast, if only one strand of a Pair includes the sequence variant, this is more likely to be an artifact. Similarly, the confidence that a Singlet bearing a sequence variant existed in the original molecule is less than the confidence if the variant exists in a Duplex, as there is higher probability that the variant can be introduced once than twice during amplification/sequencing.

Other methods of copy number variation detection and the sequence variant detection are described in PCT/US2013/058061, which is entirely incorporated herein by reference.

Sequence reads can be collapsed to generate a consensus sequence, which can be mapped to a reference sequence to identify genetic variants, such as CNV or SNV. As an alternative, the sequence reads are mapped prior to or even without mapping. In such a case, the sequence reads can be individually mapped to the reference to identify a CNV or SNV.

FIG. **3** shows a reference sequence encoding a genetic Locus A. The polynucleotides in FIG. **3** may be Y-shaped or have other shapes, such as hairpin.

In some cases, an SNV or multiple-nucleotide variant (MNV) can be determined across multiple sequence reads at a given locus (e.g., nucleotide base) by aligning sequence reads that correspond to that locus. Next, a plurality of sequential nucleotide bases from at least a subset of the sequence reads are mapped to the reference to a SNV or MNV in a polynucleotide molecule or portion thereof that corresponds to the reads. The plurality of sequential nucleotide bases can span an actual, inferred or suspected location of the SNV or MNV. The plurality of sequential nucleotide bases can span at least 3, 4, 5, 6, 7, 8, 9, or 10 nucleotide bases.

L. Detecting/Quantifying Nucleic Acids

The methods described throughout can be used to tag nucleic acids fragments, such as deoxyribonucleic acid (DNA), at extremely high efficiency. This efficient tagging allows a person to efficiently and accurately detect rare DNA in heterogenous populations of original DNA fragments (such as in cfDNA). A rare polynucleotide (e.g., rare DNA) can be a polynucleotide that comprises a genetic variant occurring in a population of polynucleotides at a frequency of less than 10%, 5%, 4%, 3%, 2%, 1%, or 0.1%. A rare DNA can be a polynucleotide with a detectable property at a concentration less than 50%, 25%, 10%, 5%, 1%, or 0.1%

Tagging can occur in a single reaction. In some cases, two or more reactions can be performed and pooled together. Tagging each original DNA fragments in a single reaction can result in tagging such that greater than 50% (e.g., 60%, 70%, 80%, 90%, 95%, or 99%) of the original DNA fragments are tagged at both ends with tags that comprise molecular barcodes, thereby providing tagged DNA fragments. Tagging can also result in greater than 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% of the original DNA fragments tagged at both ends with tags that comprise molecular barcodes. Tagging can also result in 100% of the original DNA fragments tagged at both ends with tags that comprise molecular barcodes. Tagging can also result in single end tagging.

Tagging can also occur by using an excess amount of tags as compared to the original DNA fragments. For example, the excess can be at least 5-fold excess. In other cases, the excess can be at least 1.25, 1.5, 1.75, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100 or more fold excess. Tagging can comprise attachment to blunt ends or sticky ends. Tagging can also be performed by hybridization PCR. Tagging can also be performed in low reaction volumes, such as 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or 100 pico- and/or microliters.

The method can also include performing high fidelity amplification on the tagged DNA fragments. Any high fidelity DNA polymerases can be used. For example, the polymerase may be KAPA HiFi DNA polymerase or Phusion DNA polymerase.

Further, the method can comprise selectively enriching a subset of the tagged DNA fragments. For example, selective enrichment can be performed by hybridization or amplification techniques. The selective enrichment can be performed using a solid support (e.g., beads). The solid support (e.g., beads) can comprise probes (e.g., oligonucleotides specifically hybridizing to certain sequences. For example, the probes can hybridize with certain genomic regions, e.g., genes. In some cases, the genomic regions, e.g., genes, can be regions associated with diseases, e.g., cancer. After enrichment, the selected fragmented can be attached any sequencing adaptor disclosed in this invention. For example, a sequence adaptor can comprise a flow cell sequence, a sample barcode, or both. In another example, a sequence adaptor can be a hairpin shaped adaptor and/or comprises a sample barcode. Further, the resulting fragments can be amplified and sequenced. In some cases, the adaptor does not comprise a sequencing primer region.

The method can include sequencing one or both strands of the DNA fragments. In one case, both strands of the DNA fragment are independently sequenced. The tagged, amplified, and/or selectively enriched DNA fragments are sequenced to obtain sequence reads that comprise sequence information of the molecular barcodes and at least a portion of the original DNA fragments.

The method can include reducing or tracking redundancy (as described above) in the sequence reads to determine consensus reads that are representative of single-strands of the original DNA fragments. For example, to reduce or track redundancy, the method can include comparing sequence reads having the same or similar molecular barcodes and the same or similar end of fragment sequences. The method can comprise performing a phylogentic analysis on the sequence reads having the same or similar molecular barcodes. The molecular barcodes can have a barcode with varying edit distances (including any edit distances as described throughout), for example, an edit distance of up to 3. The end of the fragment sequences can include fragment sequences having an edit distance with varying distances (including any edit distances as described throughout), for example, an edit distance of up to 3.

The method can comprise binning the sequence reads according to the molecular barcodes and sequence information. For example, binning the sequence reads according to the molecular barcodes and sequence information can be performed from at least one end of each of the original DNA fragments to create bins of single stranded reads. The method can further comprise in each bin, determining a sequence of a given original DNA fragment among the original DNA fragments by analyzing sequence reads.

In some cases, sequence reads in each bin can be collapsed to a consensus sequence and subsequently mapped to a genome. As an alternative, sequence reads can be mapped to a genome prior to binning and subsequently collapsed to a consensus sequence.

The method can also comprise sorting sequence reads into paired reads and unpaired reads. After sorting, the number of paired reads and unpaired reads that map to each of one or more genetic loci can be quantified.

The method can include quantifying the consensus reads to detect and/or quantify the rare DNA, which are described throughout. The method can comprise detecting and/or quantifying the rare DNA by comparing a number of times each base occurs at each position of a genome represented by the tagged, amplified, and/or enriched DNA fragments.

The method can comprise tagging the original DNA fragments in a single reaction using a library of tags. The library can include at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 50, at least 100, at least 500, at least 1000, at least 5000, at least 10000, or any number of tags as disclosed throughout. For example, the library of tags can include at least 8 tags. The library of tags can include 8 tags (which can generate 64 different possible combinations). The method can be conducted such that a high percentage of fragments, e.g., greater than 50% (or any percentages as described throughout) are tagged at both ends, wherein each of the tags comprises a molecular barcode.

M. Processing and/or Analyzing Nucleic Acids

The methods described throughout can be used for processing and/or analyzing a nucleic acid sample of a subject. The method can comprising exposing polynucleotide fragments of the nucleic acid sample to a plurality of polynucleotide molecules to yield tagged polynucleotide fragments. The plurality of polynucleotide molecules that can be used are described throughout the application.

For example, the plurality of polynucleotide molecules can be each less than or equal to 40 nucleic acid bases in length and have distinct barcode sequences with respect to at least 4 nucleic acid bases and an edit distance of at least 1, wherein each of the distinct barcode sequences is within 20 nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules, and wherein the plurality of polynucleotide molecules are not sequencing adaptors.

The tagged polynucleotide fragments can be subjected to nucleic acid amplification reactions under conditions that yield amplified polynucleotide fragments as amplification products of the tagged polynucleotide fragments. After amplification, the nucleotide sequence of the amplified tagged polynucleotide fragments is determined. In some cases, the nucleotide sequences of the amplified tagged polynucleotide fragments are determined without the use of polymerase chain reaction (PCR).

The method can comprise analyzing the nucleotide sequences with a programmed computer processor to identify one or more genetic variants in the nucleotide sample of the subject. Any genetic alterations can be identified, including but not limited to, base change(s), insertion(s), repeat(s), deletion(s), copy number variation(s), epigenetic modification(s), nucleosome binding site(s), copy number change(s) due to origin(s) of replication, and transversion(s). Other genetic alterations can include, but are not limited to, one or more tumor associated genetic alterations.

The subject of the methods can be suspected of having a disease. For example, the subject can be suspected of having cancer. The method can comprise collecting a nucleic acid sample from a subject. The nucleic acid sample can be collected from blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, cerebral spinal fluid, skin, hair, sweat, and/or tears. The nucleic acid sample can be a cell-free nucleic acid sample. In some cases, the nucleic acid sample is collected from no more than 100 nanograms (ng) of double-stranded polynucleotide molecules of the subject.

The polynucleotide fragments can comprise double-stranded polynucleotide molecules. In some cases, the plurality of polynucleotide molecules are coupled to the polynucleotide fragments via blunt end ligation, sticky end ligation, molecular inversion probes, polymerase chain reaction (PCR), ligation-based PCR, multiplex PCR, single strand ligation, or single strand circularization.

The method as described herein results in high efficiency tagging of nucleic acids. For example, exposing the polynucleotide fragments of the nucleic acid sample to the plurality of polynucleotide molecules yields the tagged polynucleotide fragments with a conversion efficiency of at least 30%, e.g., of at least 50% (e.g., 60%, 70%, 80%, 90%, 95%, or 99%). Conversion efficiency of at least 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% can be achieved.

The method can result in a tagged polynucleotide fragment that share common polynucleotide molecules. For example, any of at least 5%, 6%, 7%, 8%, 9%, 10%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% or 100% of the tagged polynucleotide fragments share a common polynucleotide molecule. The method can comprise generating the polynucleotide fragments from the nucleic acid sample.

In some cases, the subjecting of the method comprises amplifying the tagged polynucleotide fragments in the presence primers corresponding to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A,

EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. Additionally, any combination of these genes can be amplified. For example, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, or all 54 of these genes can be amplified.

The methods described herein can comprise generating a plurality of sequence reads from a plurality of polynucleotide molecules. The plurality of polynucleotide molecules can cover genomic loci of a target genome. For example, the genomic loci can correspond to a plurality of genes as listed above. Further, the genomic loci can be any combination of these genes. Any given genomic locus can comprise at least two nucleic acid bases. Any given genomic locus can also comprise a plurality of nucleic acid bases, for example, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, or more nucleic acid bases.

The method can comprise grouping with a computer processor the plurality of sequence reads into families. Each of the family can comprises sequence reads from one of the template polynucleotides. Each family can comprise sequence reads from only one of the template polynucleotides. For each of the family, the sequence reads can be merged to generate a consensus sequence. The grouping can comprise classifying the plurality of sequence reads into families by identifying (i) distinct molecular barcodes coupled to the plurality of polynucleotide molecules and (ii) similarities between the plurality of sequence reads, wherein each family includes a plurality of nucleic acid sequences that are associated with a distinct combination of molecular barcodes and similar or identical sequence reads.

Once merged, a consensus sequence can be called at a given genomic locus among the genomic loci. At any given genomic loci, any of the following can be determined: i) genetic variants among the calls; ii) frequency of a genetic alteration among the calls; iii) total number of calls; and iv) total number of alterations among the calls. The calling can comprise calling at least one nucleic acid base at the given genomic locus. The calling can also comprise calling a plurality of nucleic acid bases at the given genomic locus. In some cases, the calling can comprise phylogenetic analysis, voting (e.g., biased voting), weighing, assigning a probability to each read at the locus in a family, or calling the base with the highest probability. The consensus sequence can be generated by evaluating a quantitative measure or a statistical significance level for each of the sequence reads. If a quantitative measure is performed, the method can comprise use of a binomial distribution, exponential distribution, beta distribution, or empirical distribution. However, frequency of the base at the particular location can also be used for calling, for example, if 51% or more of the reads is a "A" at the location, then the base may be called an "A" at that particular location. The method can further comprise mapping a consensus sequence to a target genome.

The method can further comprising performing consensus calling at an additional genomic locus among the genomic loci. The method can comprise determining a variation in copy number at one of the given genomic locus and additional genomic locus based on counts at the given genomic locus and additional genomic locus.

The methods described herein can comprise providing template polynucleotide molecules and a library of adaptor polynucleotide molecules in a reaction vessel. The adaptor polynucleotide molecules can have from 2 to 1,000 different barcode sequences and in some cases are not sequencing adaptors. Other variations of adaptor polynucleotide molecules are described throughout, which can also be used in the methods.

The polynucleotide molecules of the adaptors can have the same sample tag. The adaptor polynucleotide molecules can be coupled to both ends of the template polynucleotide molecules. The method can comprise coupling the adaptor polynucleotide molecules to the template polynucleotide molecules at an efficiency of at least 30%, e.g., of at least 50% (e.g., 60%, 70%, 80%, 90%, 95%, or 99%), thereby tagging each template polynucleotide with a tagging combination that is among 4 to 1,000,000 different tagging combinations, to produce tagged polynucleotide molecules. In some cases, the reaction can occur in a single reaction vessel. Coupling efficiency can also be at least 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99%. Tagging can be non-unique tagging.

The tagged polynucleotide molecules can then be subject to an amplification reaction under conditions that will yield amplified polynucleotide molecules as amplification products of the tagged polynucleotide molecules. The template polynucleotide molecules can be double-stranded. Further, the template polynucleotide molecules can be blunt ended. In some cases, the amplification reaction comprises non-specifically amplifying the tagged polynucleotide molecules. The amplification reaction can also comprises using a priming site to amplify each of the tagged polynucleotide molecules. The priming site can be a primer, e.g., a universal primer. The priming site can also be a nick.

The method can also comprise sequencing the amplified polynucleotide molecules. The sequencing can comprise (i) subjecting the amplified polynucleotide molecules to an additional amplification reaction under conditions that yield additional amplified polynucleotide molecules as amplification products of the amplified polynucleotide molecules, and/or (ii) sequencing the additional amplified polynucleotide molecules. The additional amplification can be performed in the presence of primers comprising flow cells sequences, which will produce polynucleotide molecules that are capable of binding to a flow cell. The additional amplification can also be performed in the presence of primers comprising sequences for hairpin shaped adaptors. The hairpin shaped adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. The method can further comprise identifying genetic variants upon sequencing the amplified polynucleotide molecules.

The method can further comprising separating polynucleotide molecules comprising one or more given sequences from the amplified polynucleotide molecules, to produce enriched polynucleotide molecules. The method can also comprise amplifying the enriched polynucleotide molecules

with primers comprising the flow cell sequences. This amplification with primers comprising flow cell sequences will produce polynucleotide molecules that are capable of binding to a flow cell. The amplification can also be performed in the presence of primers comprising sequences for hairpin shaped adaptors. The hairpin shaped adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times.

Flow cell sequences or hairpin shaped adaptors can be added by non-amplification methods such as through ligation of such sequences. Other techniques such as hybridization methods can be used, e.g., nucleotide overhangs.

The method can be performed without aliquoting the tagged polynucleotide molecules. For example, once the tagged polynucleotide molecule is made, the amplification and sequencing can occur in the same tube without any further preparation.

The methods described herein can be useful in detecting single nucleotide variations (SNV), copy number variations (CNV), insertions, deletions, and/or rearrangements. In some cases, the SNVs, CNVs, insertions, deletions, and/or rearrangements, can be associated with disease, for example, cancer.

N. Monitoring a Patient's Status

Methods disclosed herein can also be used to monitor a patient's disease status. The disease of a subject can be monitored over time to determine a progression of the disease (e.g., regression). Markers indicative of the disease can be monitored in a biological sample of the subject, such as a cell-free DNA sample.

For example, monitoring a subject's cancer status can comprise (a) determining an amount of one or more SNVs or copy numbers of a plurality of genes (e.g., in an exon), (b) repeating such determination at different points in time, and (c) determining if there is a difference in the number of SNVs, level of SNVs, number or level of genomic rearrangements, or copy numbers between (a) and (b). The genes can be selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The genes can be selected from any 5, 10, 15, 20, 30, 40, 50, or all of the genes in this group.

O. Sensitivity and Specificity

Methods disclosed herein can be used to detect cancer polynucleotides in a sample, and cancer in a subject, with high measures of agreement, e.g., high sensitivity and/or specificity. For example, such methods can detect cancer polynucleotides (e.g., rare DNA) in a sample at a concentration that is less than 5%, 1%, 0.5%, 0.1%, 0.05%, or 0.01%, at a specificity of at least 99%, 99.9%, 99.99%, 99.999%, 99.9999%, or 99.99999%. Such polynucleotides may be indicative of cancer or other disease. Further, such methods can detect cancer polynucleotides in a sample with a positive predictive value of at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.9%, 99.99%, 99.999%, or 99.9999%.

Subjects identified as positive in a test that are in reality positive are referred as true positives (TP). Subjects identified as positive in a test that are in reality negative are referred as false positives (FP). Subjects identified as negative in a test that are in reality negative are referred as true negatives (TN). Subjects identified as negative in a test that are in reality positive are referred as false negatives (FN). Sensitivity is the percentage of actual positives identified in a test as positive. This includes, for example, instances in which one should have found a cancer genetic variant and did. (Sensitivity=TP/(TP+FN).) Specificity is the percentage of actual negatives identified in a test as negative. This includes, for example, instances in which one should have found no cancer genetic variant and did not. Specificity can be calculated using the following equation: Specificity=TN/(TN+FP). Positive predictive value (PPV) can be measured by the percentage of subjects who test positive that are true positives. PPV can be calculated using the following equation: PPV=TP/(TP+FP). Positive predictive value can be increased by increasing sensitivity (e.g., chance of an actual positive being detected) and/or specificity (e.g., chance of not mistaking an actual negative for a positive).

Low conversion rates of polynucleotides into adaptor-tagged polynucleotides can compromise sensitivity as it decreases the chance of converting, and therefore detecting, rare polynucleotide targets. Noise in a test can compromise specificity as it increases the number of false positives detected in a test. Both low conversion rate and noise compromise positive predictive value as they decrease the percentage of true positives and increase the percentage of false positives.

The methods disclosed herein can achieve high levels of agreement, e.g., sensitivity and specificity, leading to high positive predictive values. Methods of increasing sensitivity include high efficiency conversion of polynucleotides into adaptor-tagged polynucleotides in a sample. Methods of increasing specificity include reducing sequencing errors, for example, by molecular tracking.

Methods of the present disclosure can be used to detect genetic variation in non-uniquely tagged initial starting genetic material (e.g., rare DNA) at a concentration that is less than 5%, 1%, 0.5%, 0.1%, 0.05%, or 0.01%, at a specificity of at least 99%, 99.9%, 99.99%, 99.999%, 99.9999%, or 99.99999%. In some aspects, the methods can further comprise converting polynucleotides in the initial starting material at an efficiency of at least at least 10%, at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, or at least 90%. Sequence reads of tagged polynucleotides can be subsequently tracked to generate consensus sequences for polynucleotides with an error rate of no more than 2%, 1%, 0.1%, or 0.01%.

2. Pooling Methods

Disclosed herein are methods of detecting copy number variation and/or sequence variants at one or more genetic loci in a test sample. One embodiment is shown in FIG. **8**. Typically, detecting copy number variation involves determining a quantitative measure (e.g., an absolute or relative number) of polynucleotides mapping to a genetic locus of interest in a genome of a test sample, and comparing that number to a quantitative measure of polynucleotides mapping to that locus in a control sample. In certain methods, the quantitative measure is determined by comparing the number of molecules in the test sample that map to a locus of interest with a number of molecules in the test sample mapping to a reference sequence, e.g., a sequence expected to be present at wild type ploidy number. In some examples, the reference sequence is HG19, build 37, or build 38. The

comparison could involve, for example, determining a ratio. Then, this measure is compared with a similar measure determined in a control sample. So, for example, if a test sample has a ratio of 1.5:1 for locus of interest versus reference locus, and a control sample has a ratio of 1:1 for the same loci, one may conclude that the test sample exhibits polyploidy at the locus of interest.

When the test sample and the control sample are analyzed separately, the work flow can introduce distortions between final numbers in the control and test samples.

In one method disclosed herein (e.g., flow chart **800**), polynucleotides are provided from a test and a control sample (**802**). Polynucleotides in a test sample and those in a control sample are tagged with tags that identify the polynucleotides as originating from the test or control sample (a source tag). (**804**.) The tag can be, for example, a polynucleotide sequence or barcode that unambiguously identifies the source.

The polynucleotides in each of the control and test samples also can be tagged with identifier tags that will be carried by all amplification progeny of a polynucleotide. Information from start and end sequences of a polynucleotide and identifier tags can identify sequence reads from polynucleotides amplified from an original parent molecule. Each molecule can be uniquely tagged compared with other molecules in the sample. Alternatively, each molecule need not be uniquely tagged compared with other molecules in the sample. That is, the number of different identifier sequences can be fewer than that the number of molecules in sample. By combining identifier information with start/stop sequence information, the probability of confusing two molecules having the same start/stop sequence is significantly diminished.

Number of different identifiers used to tag a nucleic acid (e.g., cfDNA) can dependent on the number of different haploid genome equivalents. Different identifiers can be used to tag at least 2, least 10, least 100, least 200, least 300, least 400, least 500, least 600, least 700, least 800, least 900, least 1,000, least 2,000, least 3,000, least 4,000, least 5,000, least 6,000, least 7,000, least 8,000, least 9,000, least 10,000 or more different haploid genome equivalents. Accordingly, the number of different identifiers used to tag a nucleic acid sample, e.g., cell-free DNA from 500 to 10,000 different haploid genome equivalents and be between any of 1, 2, 3, 4 and 5 and no more than 100, 90, 80, 70, 60, 50, 40 or 30. For example, the number of different identifier used to tag a nucleic acid sample from 500 to 10,000 different haploid genome equivalents can be 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 or less.

Polynucleotides can be tagged by ligation of adaptors comprising the tags or identifiers before amplification. Ligation can be performed using an enzyme, e.g., a ligase. For example, tagging can be performed using a DNA ligase. The DNA ligase can be a T4 DNA ligase, *E. coli* DNA ligase, and/or mammalian ligase. The mammalian ligase can be DNA ligase I, DNA ligase III, or DNA ligase IV. The ligase may also be a thermostable ligase. Tags can be ligated to a blunt-end of a polynucleotide (blunt-end ligation). Alternatively, tags can be ligated to a sticky end of a polynucleotide (sticky-end ligation). The polynucleotides can be tagged by blunt end ligation using adaptors (e.g., adaptors having forked ends). High efficiency of ligation can be achieved using high excess of adaptors (e.g., more than 1.5×, more than 2×, more than 3×, more than 4×, more than 5×, more than 6×, more than 7×, more than 8×, more than 9×, more than 10×, more than 11×, more than 12×, more than 13×, more than 14×, more than 15×, more than 20×, more than 25×, more than 30×, more than 35×, more than 40×, more than 45×, more than 50×, more than 55×, more than 60×, more than 65×, more than 70×, more than 75×, more than 80×, more than 85×, more than 90×, more than 95×, or more than 100).

Once tagged with tags that identify source of the polynucleotides, polynucleotides from different sources (e.g., different samples) can be pooled. After pooling, polynucleotides from different sources (e.g., different samples) can be distinguished by any measurement using the tags, including any process of quantitative measurement. For example, as shown in (**806**) (FIG. **8**), polynucleotides from the control sample and the test sample can be pooled. The pooled molecules can be subject to the sequencing (**808**) and bioinformatic work flow. Both will be subject to the same variations in the process and, therefore, any differential bias is reduced. Because molecules originating from control and test samples are differently tagged, they can be distinguished in any process of quantitative measurement.

The relative amount of control and test sample pooled can be varied. The amount of control sample can be same as the amount of test sample. The amount of control sample can also be larger than the amount of test sample. Alternatively, the amount of control sample can be smaller than the amount of test sample. The smaller the relative amount of one sample to the total, the fewer identifying tags needed in the original tagging process. A number can be selected to reduce to acceptable levels the probability that two parent molecules having the same start/end sequences will bear the same identifying tag. This probability can be less than 10%, less than 1%, less than 0.1% or less than 0.01%. The probability can be less than 25%, 24%, 23%, 22%, 21%, 20%, 19%, 18%, 17%, 16%, 15%, 14%, 13%, 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1%.

Methods disclosed herein can also comprise grouping sequence reads. For example, bioinformatic workflow can include grouping sequence reads produced from progeny of a single parent molecule, as shown in (**810**) (FIG. **8**). This can involve any of the redundancy reduction methods described herein. Molecules sourced from test and control samples can be differentiated based on source tags they carry (**812**). Molecules mapping to a target locus are quantified for both test-sourced and control-sourced molecules (**812**). This can include the normalization methods discussed herein, e.g., in which numbers at a target locus are normalized against numbers at a reference locus.

Normalized (or raw) quantities at a target locus from test and control samples are compared to determine presence of copy number variation (**814**).

3. Computer Control Systems

The present disclosure provides computer control systems that are programmed to implement methods of the disclosure. FIG. **6** shows a computer system **1501** that is programmed or otherwise configured to implement the methods of the present disclosure. The computer system **1501** can regulate various aspects sample preparation, sequencing and/or analysis. In some examples, the computer system **1501** is configured to perform sample preparation and sample analysis, including nucleic acid sequencing. The computer system **1501** can be an electronic device of a user

or a computer system that is remotely located with respect to the electronic device. The electronic device can be a mobile electronic device.

The computer system **1501** includes a central processing unit (CPU, also "processor" and "computer processor" herein) **1505**, which can be a single core or multi core processor, or a plurality of processors for parallel processing. The computer system **1501** also includes memory or memory location **1510** (e.g., random-access memory, read-only memory, flash memory), electronic storage unit **1515** (e.g., hard disk), communication interface **1520** (e.g., network adapter) for communicating with one or more other systems, and peripheral devices **1525**, such as cache, other memory, data storage and/or electronic display adapters. The memory **1510**, storage unit **1515**, interface **1520** and peripheral devices **1525** are in communication with the CPU **1505** through a communication bus (solid lines), such as a motherboard. The storage unit **1515** can be a data storage unit (or data repository) for storing data. The computer system **1501** can be operatively coupled to a computer network ("network") **1530** with the aid of the communication interface **1520**. The network **1530** can be the Internet, an internet and/or extranet, or an intranet and/or extranet that is in communication with the Internet. The network **1530** in some cases is a telecommunication and/or data network. The network **1530** can include one or more computer servers, which can enable distributed computing, such as cloud computing. The network **1530**, in some cases with the aid of the computer system **1501**, can implement a peer-to-peer network, which may enable devices coupled to the computer system **1501** to behave as a client or a server.

The CPU **1505** can execute a sequence of machine-readable instructions, which can be embodied in a program or software. The instructions may be stored in a memory location, such as the memory **1510**. The instructions can be directed to the CPU **1505**, which can subsequently program or otherwise configure the CPU **1505** to implement methods of the present disclosure. Examples of operations performed by the CPU **1505** can include fetch, decode, execute, and writeback.

The CPU **1505** can be part of a circuit, such as an integrated circuit. One or more other components of the system **1501** can be included in the circuit. In some cases, the circuit is an application specific integrated circuit (ASIC).

The storage unit **1515** can store files, such as drivers, libraries and saved programs. The storage unit **1515** can store user data, e.g., user preferences and user programs. The computer system **1501** in some cases can include one or more additional data storage units that are external to the computer system **1501**, such as located on a remote server that is in communication with the computer system **1501** through an intranet or the Internet.

The computer system **1501** can communicate with one or more remote computer systems through the network **1530**. For instance, the computer system **1501** can communicate with a remote computer system of a user (e.g., an operator). Examples of remote computer systems include personal computers (e.g., portable PC), slate or tablet PC's (e.g., Apple® iPad, Samsung® Galaxy Tab), telephones, Smart phones (e.g., Apple® iPhone, Android-enabled device, Blackberry®), or personal digital assistants. The user can access the computer system **1501** via the network **1530**.

Methods as described herein can be implemented by way of machine (e.g., computer processor) executable code stored on an electronic storage location of the computer system **1501**, such as, for example, on the memory **1510** or electronic storage unit **1515**. The machine executable or machine readable code can be provided in the form of software. During use, the code can be executed by the processor **1505**. In some cases, the code can be retrieved from the storage unit **1515** and stored on the memory **1510** for ready access by the processor **1505**. In some situations, the electronic storage unit **1515** can be precluded, and machine-executable instructions are stored on memory **1510**.

The code can be pre-compiled and configured for use with a machine have a processor adapted to execute the code, or can be compiled during runtime. The code can be supplied in a programming language that can be selected to enable the code to execute in a pre-compiled or as-compiled fashion.

Aspects of the systems and methods provided herein, such as the computer system **1501**, can be embodied in programming. Various aspects of the technology may be thought of as "products" or "articles of manufacture" typically in the form of machine (or processor) executable code and/or associated data that is carried on or embodied in a type of machine readable medium. Machine-executable code can be stored on an electronic storage unit, such memory (e.g., read-only memory, random-access memory, flash memory) or a hard disk. "Storage" type media can include any or all of the tangible memory of the computers, processors or the like, or associated modules thereof, such as various semiconductor memories, tape drives, disk drives and the like, which may provide non-transitory storage at any time for the software programming. All or portions of the software may at times be communicated through the Internet or various other telecommunication networks. Such communications, for example, may enable loading of the software from one computer or processor into another, for example, from a management server or host computer into the computer platform of an application server. Thus, another type of media that may bear the software elements includes optical, electrical and electromagnetic waves, such as used across physical interfaces between local devices, through wired and optical landline networks and over various air-links. The physical elements that carry such waves, such as wired or wireless links, optical links or the like, also may be considered as media bearing the software. As used herein, unless restricted to non-transitory, tangible "storage" media, terms such as computer or machine "readable medium" refer to any medium that participates in providing instructions to a processor for execution.

Hence, a machine readable medium, such as computer-executable code, may take many forms, including but not limited to, a tangible storage medium, a carrier wave medium or physical transmission medium. Non-volatile storage media include, for example, optical or magnetic disks, such as any of the storage devices in any computer(s) or the like, such as may be used to implement the databases, etc. shown in the drawings. Volatile storage media include dynamic memory, such as main memory of such a computer platform. Tangible transmission media include coaxial cables; copper wire and fiber optics, including the wires that comprise a bus within a computer system. Carrier-wave transmission media may take the form of electric or electromagnetic signals, or acoustic or light waves such as those generated during radio frequency (RF) and infrared (IR) data communications. Common forms of computer-readable media therefore include for example: a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD or DVD-ROM, any other optical medium, punch cards paper tape, any other physical storage medium with patterns of holes, a RAM, a ROM, a PROM

and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave transporting data or instructions, cables or links transporting such a carrier wave, or any other medium from which a computer may read programming code and/or data. Many of these forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to a processor for execution.

The computer system **1501** can include or be in communication with an electronic display **1535** that comprises a user interface (UI) **1540**. The UI can allow a user to set various conditions for the methods described herein, for example, PCR or sequencing conditions. Examples of UI's include, without limitation, a graphical user interface (GUI) and web-based user interface.

Methods and systems of the present disclosure can be implemented by way of one or more algorithms. An algorithm can be implemented by way of software upon execution by the central processing unit **1505**. The algorithm can, for example, process the reads to generate a consequence sequence.

FIG. **7** schematically illustrates another system for analyzing a sample comprising nucleic acids from a subject. The system includes a sequencer, bioinformatic software and internet connection for report analysis by, for example, a hand held device or a desktop computer

Disclosed herein is a system for analyzing a target nucleic acid molecule of a subject, comprising: a communication interface that receives nucleic acid sequence reads for a plurality of polynucleotide molecules that cover genomic loci of a target genome; computer memory that stores the nucleic acid sequence reads for the plurality of polynucleotide molecules received by the communication interface; and a computer processor operatively coupled to the communication interface and the memory and programmed to (i) group the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides, (ii) for each of the families, merge sequence reads to generate a consensus sequence, (iii) call the consensus sequence at a given genomic locus among the genomic loci, and (iv) detect at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls; and total number of alterations among the calls, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The different variations of each component of the system are described throughout the disclosure within the methods and compositions. These individual components and variations thereof, are also applicable in this system.

4. Kits

Kits comprising the compositions as described herein. The kits can be useful in performing the methods as described herein. Disclosed herein is a kit comprising a plurality of oligonucleotide probes that selectively hybridize to least 5, 6, 7, 8, 9, 10, 20, 30, 40 or all genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The number genes to which the oligonucleotide probes can selectively hybridize can vary. For example, the number of genes can comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, or 54. The kit can include a container that includes the plurality of oligonucleotide probes and instructions for performing any of the methods described herein.

The oligonucleotide probes can selectively hybridize to exon regions of the genes, e.g., of the at least 5 genes. In some cases, the oligonucleotide probes can selectively hybridize to at least 30 exons of the genes, e.g., of the at least 5 genes. In some cases, the multiple probes can selectively hybridize to each of the at least 30 exons. The probes that hybridize to each exon can have sequences that overlap with at least 1 other probe. In some embodiments, the oligoprobes can selectively hybridize to non-coding regions of genes disclosed herein, for example, intronic regions of the genes. The oligoprobes can also selectively hybridize to regions of genes comprising both exonic and intronic regions of the genes disclosed herein.

Any number of exons can be targeted by the oligonucleotide probes. For example, at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, 275, 280, 285, 290, 295, 300, 400, 500, 600, 700, 800, 900, 1,000, or more, exons can be targeted.

The kit can comprise at least 4, 5, 6, 7, or 8 different library adaptors having distinct molecular barcodes and identical sample barcodes. The library adaptors may not be sequencing adaptors. For example, the library adaptors do not include flow cell sequences or sequences that permit the formation of hairpin loops for sequencing. The different variations and combinations of molecular barcodes and sample barcodes are described throughout, and are applicable to the kit. Further, in some cases, the adaptors are not sequencing adaptors. Additionally, the adaptors provided with the kit can also comprise sequencing adaptors. A sequencing adaptor can comprise a sequence hybridizing to one or more sequencing primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, e.g., a flow cell sequence. For example, a sequencing adaptor can be a flow cell adaptor. The sequencing adaptors can be attached to one or both ends of a polynucleotide fragment. In some cases, the kit can comprise at least 8 different library adaptors having distinct molecular barcodes and identical sample barcodes. The library adaptors may not be sequencing adaptors. The kit can further include a sequencing adaptor having a first sequence that selectively hybridizes to the library adaptors and a second sequence that selectively hybridizes to a flow cell sequence. In another example, a sequencing adaptor can be hairpin shaped. For example, the hairpin shaped adaptor can comprise a comple-

mentary double stranded portion and a loop portion, where the double stranded portion can be attached (e.g., ligated) to a double-stranded polynucleotide. Hairpin shaped sequencing adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. A sequencing adaptor can be up to 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, or more bases from end to end. The sequencing adaptor can comprise 20-30, 20-40, 30-50, 30-60, 40-60, 40-70, 50-60, 50-70, bases from end to end. In a particular example, the sequencing adaptor can comprise 20-30 bases from end to end. In another example, the sequencing adaptor can comprise 50-60 bases from end to end. A sequencing adaptor can comprise one or more barcodes. For example, a sequencing adaptor can comprise a sample barcode. The sample barcode can comprise a pre-determined sequence. The sample barcodes can be used to identify the source of the polynucleotides. The sample barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, e.g., at least 8 bases. The barcode can be contiguous or non-contiguous sequences, as described above.

The library adaptors can be blunt ended and Y-shaped and can be less than or equal to 40 nucleic acid bases in length. Other variations of the can be found throughout and are applicable to the kit.

EXAMPLES

Example 1

Methods for Copy Number Variation Detection

Blood Collection

10-30 mL Blood samples are collected at room temperature. The samples are centrifuged to remove cells. Plasma is collected after centrifugation.

cfDNA Extraction

The sample is subjected to proteinase K digestion. DNA is precipitated with isopropanol. DNA is captured on a DNA purification column (e.g., a QlAamp DNA Blood Mini Kit) and eluted in 100 μsolution. DNAs below 500 bp are selected with Ampure SPRI magnetic bead capture (PEG/salt). The resulting production is suspended in 30 μl H_2O. Size distribution is checked (major peak=166 nucleotides; minor peak=330 nucleotides) and quantified. 5 ng of extracted DNA contain approximately 1700 haploid genome equivalents ("HGE"). The general correlation between the amount of DNA and HGE is as follow: 3 pg DNA=1 HGE; 3 ng DNA=1K HGE; 3 μg DNA=1M HGE; 10 pg DNA=3 HE; 10 ng DNA=3K HGE; 10 μg DNA=3M HGE.

"Single Molecule" Library Prep

High-efficiency DNA tagging (>80%) is performed by blunt-end repair and ligation with 8 different octomers (i.e., 64 combinations) with overloaded hairpin adaptors. 2.5 ng DNA (i.e. approximately 800 HGE) is used as the starting material. Each hairpin adaptor comprises a random sequence on its non-complementary portion. Both ends of each DNA fragment are attached with hairpin adaptors. Each tagged fragment can be identified by the random sequence on the hairpin adaptors and a 10 p endogenous sequence on the fragment.

Tagged DNA is amplified by 10 cycles of PCR to produce about 1-7 μg DNAs that contain approximately 500 copies of each of the 800 HGE in the starting material.

Buffer optimization, polymerase optimization and cycle reduction may be performed to optimize the PCR reactions. Amplification bias, e.g., non-specific bias, GC bias, and/or size bias are also reduced by optimization. Noise(s) (e.g., polymerase-introduced errors) are reduced by using high-fidelity polymerases.

The Library may be prepared using Verniata or Sequenom methods.

Sequences may be enriched as follow: DNAs with regions of interest (ROI) are captured using biotin-labeled bead with probe to ROIs. The ROIs are amplified with 12 cycles of PCR to generate a 2000 times amplification. The resulting DNA is then denatured and diluted to 8 pM and loaded into an Illumina sequencer.

Massively Parallel Sequencing

0.1 to 1% of the sample (approximately 100 pg) are used for sequencing.

Digital Bioinformatics

Sequence reads are grouped into families, with about 10 sequence reads in each family. Families are collapsed into consensus sequences by voting (e.g., biased voting) each position in a family. A base is called for consensus sequence if 8 or 9 members agree. A base is not called for consensus sequence if no more than 60% of the members agree.

The resulting consensus sequences are mapped to a reference genome. Each base in a consensus sequence is covered by about 3000 different families. A quality score for each sequence is calculated and sequences are filtered based on the their quality scores.

Sequence variation is detected by counting distribution of bases at each locus. If 98% of the reads have the same base (homozygous) and 2% have a different base, the locus is likely to have a sequence variant, presumably from cancer DNA.

CNV is detected by counting the total number of sequences (bases) mapping to a locus and comparing with a control locus. To increase CNV detection, CNV analysis is performed specific regions, including regions on ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET,SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, or NTRK1 genes.

Example 2

Method for Correcting Base Calling by Determining the Total Number Unseen Molecules in a Sample

After fragments are amplified and the sequences of amplified fragments are read and aligned, the fragments are subjected to base calling. Variations in the number of amplified fragments and unseen amplified fragments can

introduce errors in base calling. These variations are corrected by calculating the number of unseen amplified fragments.

When base calling for locus A (an arbitrary locus), it is first assumed that there are N amplified fragments. The sequence readouts can come from two types of fragments: double-strand fragments and single-strand fragments. The following is a theoretical example of calculating the total number of unseen molecules in a sample.

N is the total number of molecules in the sample.

Assuming 1000 is the number of duplexes detected.

Assuming 500 is the number of single-stranded molecule detected.

P is the probability of seeing a strand.

Q is the probability of not detecting a strand.

Since $Q=1-P$.

$1000=NP(2)$.

$500=N2PQ$.

$1000/P(2)=N$.

$500\div 2PQ=N$.

$1000/P(2)=500\div 2PQ$.

$1000*2PQ=500P(2)$.

$2000PQ=500P(2)$.

$2000Q=500P$.

$2000(1-P)=500P$

$2000-2000P=500P$.

$2000=500P+2000P$.

$2000=2500P$.

$2000\div 2500=P$.

$0.8=P$.

$1000/P(2)=N$.

$1000\div 0.64=N$.

$1562=N$.

Number of unseen fragments=62.

Example 3

Identification of Genetic Variants in Cancer-Associated Somatic Variants in a Patient

An assay is used to analyze a panel of genes to identify genetic variants in cancer-associated somatic variants with high sensitivity.

Cell-free DNA is extracted from plasma of a patient and amplified by PCR. Genetic variants are analyzed by massively parallel sequencing of the amplified target genes. For one set of genes, all exons are sequenced as such sequencing coverage had shown to have clinically utility (Table 1). For another set of genes, sequencing coverage included those exons with a previously reported somatic mutation (Table 2). The minimum detectable mutant allele (limit of detection) is dependent on the patient's sample cell-free DNA concentration, which varied from less than 10 to over 1,000 genomic equivalents per mL of peripheral blood. Amplification may not be detected in samples with lower amounts of cell-free DNA and/or low-level gene copy amplification. Certain sample or variant characteristics resulted in reduced analytic sensitivity, such as low sample quality or improper collection.

The percentage of genetic variants found in cell-free DNA circulating in blood is related to the unique tumor biology of this patient. Factors that affected the amount/percentages of detected genetic variants in circulating cell-free DNA in blood include tumor growth, turn-over, size, heterogeneity, vascularization, disease progression or treatment. Table 3 annotates the percentage, or allele frequency, of altered circulating cell-free DNA (% cfDNA) detected in this patient. Some of the detected genetic variants are listed in descending order by % cfDNA.

Genetic variants are detected in the circulating cell-free DNA isolated from this patient's blood specimen. These genetic variants are cancer-associated somatic variants, some of which have been associated with either increased or reduced clinical response to specific treatment. "Minor Alterations" are defined as those alterations detected at less than 10% the allele frequency of "Major Alterations". The detected allele frequencies of these alterations (Table 3) and associated treatments for this patient are annotated.

All genes listed in Tables 1 and 2 are analyzed as part of the Guardant360™ test. Amplification is not detected for ERBB2, EGFR, or MET in the circulating cell-free DNA isolated from this patient's blood specimen.

Patient test results comprising the genetic variants are listed in Table 4.

TABLE 1

Genes in which all exons are sequenced
GENES IN WHICH ALL EXONS ARE SEQUENCED

ALK	<0.1%	APC	<0.1%
AR	<0.1%	BRAF	<0.1%
CDKN2A	<0.1%	EGFR	<0.1%
ERBB2	<0.1%	FBXW7	<0.1%
KRAS	<0.1%	MET	<0.1%
MYC	<0.1%	NOTCH1	<0.1%
NRAS	<0.1%	PIK3CA	<0.1%
PTEN	<0.1%	PROC	<0.1%
RB1	<0.1%	TP53	<0.1%

LOD: Limit of Detection. The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected.

TABLE 2

Genes in which exons with a previously reported somatic mutation are sequenced
GENES IN WHICH EXONS WITH A PREVIOUSLY REPORTED SOMATIC MUTATION ARE SEQUENCED

ABL1	<0.1%	AKT1	<0.1%
ATM	<0.1%	CDH1	<0.1%
CSF1R	<0.1%	CTNNB1	<0.1%
ERBB4	<0.1%	EZH2	<0.1%
FGFR1	<0.1%	FGFR2	<0.1%
FGFR3	<0.1%	FLT3	<0.1%
GNA11	<0.1%	GNAQ	<0.1%
GNAS	<0.1%	HNF1A	<0.1%
HRAS	<0.1%	IDH1	<0.1%
IDH2	<0.1%	JAK2	<0.1%
JAK3	<0.1%	KDR	<0.1%
KIT	<0.1%	MLH1	<0.1%
MPL	<0.1%	NPM1	<0.1%
PDGFRA	<0.1%	PTPN11	<0.1%

TABLE 2-continued

Genes in which exons with a previously reported somatic mutation are sequenced GENES IN WHICH EXONS WITH A PREVIOUSLY REPORTED SOMATIC MUTATION ARE SEQUENCED			
RET	<0.1%	SMAD4	<0.1%
SMARCB1	<0.1%	SMO	<0.1%
SRC	<0.1%	STK11	<0.1%
TERT	<0.1%	VHL	<0.1%

LOD: Limit of Detection. The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected.

TABLE 3

Allele frequency of altered circulating cell-free DNA detected in this patient		
Gene	cfDNA with alterations (%)	cfDNA without alterations (%)
BRAE V600E	8.9	91.1
NRAS Q61K	6.2	93.8
JAK V61F	1.5	98.6

TABLE 4

Genomic alterations detected in selected genes
Detected: 51 Genomic Alterations

Gene	Chromosome	Position	Mutation (nt)	Mutation (AA)	Percentage	Cosmic ID	DBSNP ID
KRAS	12	25368462	C > T		100.0%		rs4362222
ALK	2	29416572	T > C	I1461V	100.0%		rs1670283
ALK	2	29444095	C > T		100.0%		rs1569156
ALK	2	29543663	T > C	Q500Q	100.0%		rs2293564
ALK	2	29940529	A > T	P234P	100.0%		rs2246745
APC	5	112176756	T > A	V1822D	100.0%		rs459552
CDKN2A	9	21968199	C > G		100.0%	COSM14251	rs11515
FGFR3	4	1807894	G > A	T651T	100.0%		rs7688609
NOTCH1	9	139410424	A > G		100.0%		rs3125006
PDGFRA	4	55141055	A > G	P567P	100.0%		rs1873778
HRAS	11	534242	A > G	H27H	100.0%	COSM249860	rs12628
EGFR	7	5214348	C > T	N158N	99.9%	COSM42978	rs2072454
TP53	17	7579472	G > C	P72R	99.8%		rs1042522
APC	5	112162854	T > C	Y486Y	55.0%		rs2229992
APC	5	112177171	G > A	P1960P	53.8%		rs465899
EGFR	7	55266417	T > C	T903T	53.6%		rs1140475
APC	5	112176325	G > A	G1678G	53.2%		rs42427
APC	5	112176559	T > G	S1756S	53.0%		rs866006
EGFR	7	55229255	G > A	R521K	53.0%		
MET	7	116397572	A > G	Q648Q	52.7%		
APC	5	112175770	G > A	T1493T	52.7%		rs41115
EGFR	7	55249063	G > A	Q787Q	52.6%		rs1050171
NOTCH1	9	139411714	T > C		52.4%		rs11145767
EGFR	7	55238874	T > A	T629T	52.0%		rs2227984
ERBB2	17	37879588	A > G	I655V	51.6%		rs1136201
NOTCH1	9	139397707	G > A	D1698D	51.3%	COSM33747	rs10521
ALK	2	30143499	G > C	L9L	51.0%		rs4358080
APC	5	112164561	G > A	A545A	51.0%		rs351771
FLT3	13	28610183	A > G		50.8%		rs2491231
NOTCH1	9	139418260	A > G	N104N	50.5%		rs4489420
ALK	2	29444076	G > T		50.4%		rs1534545
PIK3CA	3	178917005	A > G		50.3%		rs3729674
NOTCH1	9	139412197	G > A		50.2%		rs9411208
ALK	2	29455267	A > G	G845G	50.0%	COSM148825	rs2256740
KIT	4	55593464	A > C	M541L	49.9%	COSM28026	
NOTCH1	9	139391636	G > A	D2185D	48.9%		rs2229974
PDGFRA	4	55152040	C > T	V824V	48.9%	COSM22413	rs2228230
ALK	2	29416481	T > C	K1491R	48.9%	COSM1130802	rs1881420
ALK	2	29445458	G > T	G1125G	48.6%		rs3795850
NOTCH1	9	139410177	T > C		48.5%		rs3124603
RET	10	43613843	G > T	L769L	48.2%		rs1800861
EGFR	7	55214443	G > A		48.0%		rs7801956
ALK	2	29416366	G > C	D1529E	47.2%		rs1881421
EGFR	7	55238087	C > T		45.5%		rs10258429
RET	10	43615633	C > G	S904S	44.8%		rs1800863
BRAF	7	140453136	A > T	V600E	8.9%	COSM476	
NRAS	1	115256530	G > T	Q61K	6.2%	COSM580	rs121913254
JAK2	9	5073770	G > T	V617F	1.5%	COSM12600	rs77375493

Example 4

Determining Patient-Specific Limits of Detection for Genes Analyzed by Guardant360™ Assays

Using the method of Example 3, Genetic alterations in cell-free DNA of a patient are detected. The sequence reads of these genes include exon and/or intron sequences.

Limits of detection of the test are shown in Table 5. The limits of detection values are dependent on cell-free DNA concentration and sequencing coverage for each gene.

TABLE 5

Limits of Detection of selected genes in a patient using Guardant

Complete Exon and Partial Intron Coverage					
APC	0.1%	AR *	0.2%	ARID1A	
BRAF *	0.1%	BRCA1		BRCA2	
CCND1 *		CCND2 *		CCNE1 *	
CDK4 *		CDK6 *		CDKN2A	0.1%
CDKN2B		EGFR *	<0.1%	ERBB2 *	0.1%
FGFR1 *	<0.1%	FGFR2 *	0.1%	HRAS	0.1%
KIT *	0.1%	KRAS *	0.1%	MET *	0.1%
MYC *	0.1%	NF1		NRAS	0.1%
PDGFRA *	0.1%	PIK3CA *	0.1%	PTEN	0.1%
RAF1 *		TP53	0.1%		
Exons Covered with Reported Somatic Mutations					
AKT1	0.1%	ALK	<0.1%	ARAF	
ATM	0.1%	CDH1	0.1%	CTNNB1	0.1%
ESR1		EZH2	0.1%	FBXW7	0.1%
FGFR3	0.1%	GATA3		GNA11	0.1%
GNAQ	0.1%	GNAS	0.1%	HNF1A	0.1%
IDH1	0.1%	IDH2	0.1%	JAK2	0.1%
JAK3	0.1%	MAP2K1		MAP2K2	
MLH1	0.1%	MPL	0.2%	NFE2L2	
NOTCH1	0.1%	NPM1	0.1%	PTPN11	0.1%
RET	0.1%	RHEB		RHOA	
RIT1		ROS1		SMAD4	0.1%
SMO	0.1%	SRC	<0.1%	STK11	0.2%
TERT	0.1%	VHL	0.2%		
Fusions					
ALK	<0.1%	RET	0.1%	ROS1	
NTRK1					

LOD: Limit of Detection. The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected.
* indicates CNV genes.

Example 5

Correcting Sequence Errors Comparing Watson and Crick Sequences

Double-stranded cell-free DNA is isolated from the plasma of a patient. The cell-free DNA fragments are tagged using 16 different bubble-containing adaptors, each of which comprises a distinctive barcode. The bubble-containing adaptors are attached to both ends of each cell-free DNA fragment by ligation. After ligation, each of the cell-free DNA fragment can be distinctly identified by the sequence of the distinct barcodes and two 20 bp endogenous sequences at each end of the cell-free DNA fragment.

The tagged cell-free DNA fragments are amplified by PCR. The amplified fragments are enriched using beads comprising oligonucleotide probes that specifically bind to a group of cancer-associated genes. Therefore, cell-free DNA fragments from the group of cancer-associated genes are selectively enriched.

Sequencing adaptors, each of which comprises a sequencing primer binding site, a sample barcode, and a cell-flow sequence, are attached to the enriched DNA molecules. The resulting molecules are amplified by PCR.

Both strands of the amplified fragments are sequenced. Because each bubble-containing adaptor comprises a non-complementary portion (e.g., the bubble), the sequence of the one strand of the bubble-containing adaptor is different from the sequence of the other strand (complement). Therefore, the sequence reads of amplicons derived from the Watson strand of an original cell-free DNA can be distinguished from amplicons from the Crick strand of the original cell-free DNA by the attached bubble-containing adaptor sequences.

The sequence reads from a strand of an original cell-free DNA fragment are compared to the sequence reads from the other strand of the original cell-free DNA fragment. If a variant occurs in only the sequence reads from one strand, but not other strand, of the original cell-free DNA fragment, this variant will be identified as an error (e.g., resulted from PCR and/or amplification), rather than a true genetic variant.

The sequence reads are grouped into families. Errors in the sequence reads are corrected. The consensus sequence of each family is generated by collapsing.

While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. It is not intended that the invention be limited by the specific examples provided within the specification. While the invention has been described with reference to the aforementioned specification, the descriptions and illustrations of the embodiments herein are not meant to be construed in a limiting sense. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. Furthermore, it shall be understood that all aspects of the invention are not limited to the specific depictions, configurations or relative proportions set forth herein which depend upon a variety of conditions and variables. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is therefore contemplated that the invention shall also cover any such alternatives, modifications, variations or equivalents. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

What is claimed is:

1. A method for analyzing sequencing reads of double-stranded cell-free deoxyribonucleic acid (cfDNA) molecules from a sample of a subject, comprising:

(a) tagging a plurality of double-stranded cfDNA molecules from a population of double-stranded cfDNA molecules from the sample with a set of library adaptors comprising a plurality of molecular barcodes to generate tagged parent polynucleotides,

wherein the tagging comprises ligating a plurality of library adaptors from the set of library adaptors to the plurality of double-stranded cfDNA molecules from the population using more than a 10× molar excess of library adaptors as compared to the double-stranded cfDNA molecules of the population;

wherein the tagging produces at least 20% of the double-stranded cfDNA molecules of the populations having library adaptors ligated to both ends of a molecule of the double-stranded cfDNA molecules;

(b) amplifying a plurality of the tagged parent polynucleotides to produce progeny polynucleotides;

(c) sequencing a plurality of the progeny polynucleotides to produce a set of sequencing reads; and

(d) determining, based at least on sequence information from the molecular barcodes, individual double-stranded cfDNA molecules from among the tagged parent polynucleotides for which either (1) both a Watson strand and a Crick strand of the individual double-stranded cfDNA molecule are detected or (2) only one of a Watson strand or a Crick strand of the individual double-stranded cfDNA molecule is detected from a plurality of sequencing reads from the set of sequencing reads.

2. The method of claim **1**, wherein the sample is selected from the group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, and tears.

3. The method of claim **1**, wherein the plurality of library adaptors is ligated to the plurality of double-stranded cfDNA molecules by blunt-end ligation or sticky-end ligation.

4. The method of claim **1**, wherein the ligating comprises using more than a 100×molar excess of library adaptors as compared to the double-stranded cfDNA molecules.

5. The method of claim **1**, wherein the tagging produces at least 40% of the double-stranded cfDNA molecules of the population having library adaptors ligated to both ends of a molecule of the double-stranded cfDNA molecules.

6. The method of claim **1**, wherein the plurality of molecular barcodes of the plurality of library adaptors has between 2 and 100,000 different molecular barcode sequences.

7. The method of claim **1**, wherein the plurality of molecular barcodes of the plurality of library adaptors has between 5 and 100 different molecular barcode sequences.

8. The method of claim **7**, wherein molecular barcodes of the plurality of molecular barcodes have a length of between 5 and 20 nucleotides.

9. The method of claim **1**, wherein a plurality of the plurality of molecular barcodes have different sequences from one another and have an edit distance of at least 1 nucleotide base between one another.

10. The method of claim **1**, wherein library adaptors of the plurality of library adaptors are Y-shaped adaptors.

11. The method of claim **1**, further comprising selectively enriching a plurality of the progeny polynucleotides for genomic regions of interest.

12. The method of claim **11**, wherein the genomic regions of interest comprises exon sequences.

13. The method of claim **11**, wherein the genomic regions of interest comprise sequences from one or more genes selected from the group consisting of:

ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

14. The method of claim **11**, wherein selectively enriching the plurality of the progeny polynucleotides for the genomic regions of interest comprises using a set of probes that hybridize to at least a portion of the genomic regions of interest.

15. The method of claim **11**, further comprising amplifying a plurality of the enriched progeny polynucleotides.

16. A method for analyzing double-stranded cell-free deoxyribonucleic acid (cfDNA) molecules from a sample of a subject, comprising:

(a) tagging a plurality of double-stranded cfDNA molecules from a population of double-stranded cfDNA molecules from the sample with a set of library adaptors comprising a plurality of molecular barcodes to generate tagged parent polynucleotides,

wherein the tagging comprises ligating a plurality of library adaptors from the set of library adaptors to the plurality of double-stranded cfDNA molecules from the population using more than a 10× molar excess of library adaptors as compared to the double-stranded cfDNA molecules of the population;

wherein the tagging produces at least 20% of the double-stranded cfDNA molecules of the population having library adaptors ligated to both ends of a molecule of the double-stranded cfDNA molecules;

(b) amplifying a plurality of the tagged parent polynucleotides to progeny polynucleotides;

(c) determining nucleotide sequences of a plurality of the progeny polynucleotides; and

(d) analyzing a plurality of the nucleotide sequences with a programmed computer processor, the analyzing comprising:

mapping a plurality of the nucleotide sequences to a reference sequence to produce mapped sequences,

grouping a plurality of the mapped sequences into families based on a combination of sequence information from the molecular barcodes and start and stop positions of the mapped sequences, wherein a family of the families is representative of an individual double-stranded cfDNA molecule from among the tagged parent polynucleotides, and

identifying a plurality of the families as having nucleotide sequences representing either (1) both a Watson strand and a Crick strand of an individual double-stranded cfDNA molecule from among the tagged parent polynucleotides or (2) only one of a Watson strand or a Crick strand of an individual double-stranded cfDNA molecule from among the tagged parent polynucleotides.

17. The method of claim **16**, wherein the sample is selected from the group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, and tears.

18. The method of claim **16**, wherein the plurality of library adaptors is ligated to the plurality of double-stranded cfDNA molecules by blunt-end ligation or sticky-end ligation.

19. The method of claim **16**, wherein the ligating comprises using more than a 100× molar excess of library adaptors as compared to the double-stranded cfDNA molecules.

20. The method of claim **16**, wherein the tagging produces at least 40% of the double-stranded cfDNA molecules of the population having library adaptors ligated to both ends of a molecule of the double-stranded cfDNA molecules.

21. The method of claim **16**, further comprising selectively enriching a plurality of the progeny polynucleotides for genomic regions of interest prior to determining the nucleotide sequences, wherein the genomic regions of interest comprise sequences from one or more genes selected from the group consisting of:

ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN,

RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

22. The method of claim **16**, wherein double-stranded cfDNA molecules of the plurality of double-stranded cfDNA molecules are non-uniquely tagged, wherein the double-stranded cfDNA molecules that map to a mappable base position of a reference sequence are tagged with a number of different molecular barcodes ranging from at least 2 to fewer than a number of the double-stranded cfDNA molecules that map to the mappable base position.

23. The method of claim **16**, wherein nucleotide sequences determined from a Watson strand of an individual double-stranded cfDNA molecule from among the tagged parent polynucleotides are distinguished from nucleotide sequences generated from a Crick strand of the individual double-stranded cfDNA molecule from among the tagged parent polynucleotides based on sequence information from the molecular barcodes and/or sequence information from endogenous nucleotide sequences of the progeny polynucleotides.

24. The method of claim **1**, wherein the tagging comprises non-uniquely tagging the plurality of double-stranded cfDNA molecules, wherein double-stranded cfDNA molecules of the plurality of double-stranded cfDNA molecules that map to a mappable base position of a reference sequence are tagged with a number of different molecular barcodes ranging from at least 2 to fewer than a number of the double-stranded cfDNA molecules that map to the mappable base position.

25. The method of claim **1**, wherein sequencing reads produced from a Watson strand of a double-stranded cfDNA molecule from among the tagged parent polynucleotides are distinguished from sequencing reads generated from a Crick strand of the double-stranded cfDNA molecule from among the tagged parent polynucleotides based on sequence information from the molecular barcodes and/or sequence information from endogenous nucleotide sequences of the progeny polynucleotides.

26. The method of claim **1**, further comprising, prior to (d), mapping a plurality of sequencing reads from the set of sequencing reads to a reference sequence to produce mapped sequencing reads.

27. The method of claim **26**, further comprising grouping a plurality of the mapped sequencing reads into families based on a combination of sequence information from the molecular barcodes and start and stop positions of the mapped sequencing reads, wherein a family of the families is representative of an individual double-stranded cfDNA molecule from among the tagged parent polynucleotides.

28. The method of claim **1**, further comprising:

(e) for a plurality of genetic loci of a reference sequence, estimating a number of tagged parent polynucleotides that map to the genetic locus of the reference sequence based at least on quantitative measures of the individual double-stranded cfDNA molecules determined in (d) that map to the genetic locus.

29. The method of claim **28**, wherein estimating the number of tagged parent polynucleotides is further based on inferring, from at least the quantitative measures of the individual double-stranded cfDNA molecules determined in (d), a quantitative measure of individual double-stranded cfDNA molecules that map to the genetic locus for which neither a Watson strand nor a Crick strand is detected.

* * * * *

EXHIBIT C

US011118221B2

(12) **United States Patent**
Talasaz et al.

(10) **Patent No.: US 11,118,221 B2**
(45) **Date of Patent: *Sep. 14, 2021**

(54) **METHODS AND SYSTEMS FOR DETECTING GENETIC VARIANTS**

(71) Applicant: **GUARDANT HEALTH, INC.**, Redwood City, CA (US)

(72) Inventors: **AmirAli Talasaz**, Atherton, CA (US); **Stefanie Ann Ward Mortimer**, Morgan Hill, CA (US); **Helmy Eltoukhy**, Atherton, CA (US)

(73) Assignee: **GUARDANT HEALTH, INC.**, Redwood City, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **16/672,267**

(22) Filed: **Jan. 7, 2020**

(65) **Prior Publication Data**

US 2020/0115746 A1 Apr. 16, 2020

Related U.S. Application Data

(63) Continuation of application No. 16/601,168, filed on Oct. 14, 2019, now Pat. No. 10,801,063, which is a continuation of application No. 15/892,178, filed on Feb. 8, 2018, which is a continuation of application No. 14/861,989, filed on Sep. 22, 2015, now Pat. No. 9,920,366, which is a continuation of application No. PCT/US2014/072383, filed on Dec. 24, 2014.

(60) Provisional application No. 61/948,509, filed on Mar. 5, 2014, provisional application No. 61/921,456, filed on Dec. 28, 2013.

(51) **Int. Cl.**

C12P 19/34	(2006.01)
C12Q 1/6869	(2018.01)
C12Q 1/6886	(2018.01)
G16B 15/00	(2019.01)

(52) **U.S. Cl.**

CPC ***C12Q 1/6869*** (2013.01); ***C12Q 1/6886*** (2013.01); *C12Q 2535/122* (2013.01); *C12Q 2600/158* (2013.01); *G16B 15/00* (2019.02)

(58) **Field of Classification Search**

CPC C12Q 1/6869

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,725,536	A	2/1988	Fritsch et al.
4,942,124	A	7/1990	Church
5,124,246	A	6/1992	Urdea et al.
5,149,625	A	9/1992	Church et al.
5,200,314	A	4/1993	Urdea
5,424,186	A	6/1995	Fodor et al.
5,424,413	A	6/1995	Hogan et al.
5,445,934	A	8/1995	Fodor et al.
5,604,097	A	2/1997	Brenner
5,635,352	A	6/1997	Urdea et al.
5,635,400	A	6/1997	Brenner
5,648,245	A	7/1997	Fire et al.
5,654,413	A	8/1997	Brenner
5,656,731	A	8/1997	Urdea
5,658,737	A	8/1997	Nelson et al.
5,714,330	A	2/1998	Brenner et al.
5,744,305	A	4/1998	Fodor et al.
5,759,778	A	6/1998	Li et al.
5,763,175	A	6/1998	Brenner
5,800,992	A	9/1998	Fodor et al.
5,846,719	A	12/1998	Brenner et al.
5,854,033	A	12/1998	Lizardi
5,871,928	A	2/1999	Fodor et al.
5,925,525	A	7/1999	Fodor et al.
5,935,793	A	8/1999	Wong
5,952,170	A	9/1999	Stroun et al.
5,968,740	A	10/1999	Fodor et al.
5,981,176	A	11/1999	Wallace
5,981,179	A	11/1999	Lorinez et al.
6,013,445	A	1/2000	Albrecht et al.
6,020,124	A	2/2000	Sorenson

(Continued)

FOREIGN PATENT DOCUMENTS

CN	102933721 A	2/2013
EP	0799897 B1	6/2006

(Continued)

OTHER PUBLICATIONS

Nawroz, et al. Microsatellite alterations in serum DNA of head and neck cancer patients. Nat Med. Sep. 1996;2(9):1035-7.

Newman, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. May 2014;20(5):548-54. doi: 10.1038/nm.3519. Epub Apr. 6, 2014.

Ng, S.B., et al., "Targeted capture and massively parallel sequencing of 12 human exomes," Nature 2009, 461(7261), 272-276.

Nielsen, R. et al. "Genotype and SNP calling from next-generation sequencing data" Nature Reviews Genetics (2011) 12(6):443-451.

(Continued)

Primary Examiner — Kenneth R Horlick

(74) *Attorney, Agent, or Firm* — Timothy A. Hott

(57) **ABSTRACT**

Disclosed herein in are methods and systems for determining genetic variants (e.g., copy number variation) in a polynucleotide sample. A method for determining copy number variations includes tagging double-stranded polynucleotides with duplex tags, sequencing polynucleotides from the sample and estimating total number of polynucleotides mapping to selected genetic loci. The estimate of total number of polynucleotides can involve estimating the number of double-stranded polynucleotides in the original sample for which no sequence reads are generated. This number can be generated using the number of polynucleotides for which reads for both complementary strands are detected and reads for which only one of the two complementary strands is detected.

30 Claims, 11 Drawing Sheets

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,040,138 A 3/2000 Lockhart et al.
6,046,005 A 4/2000 Ju et al.
6,060,596 A 5/2000 Lerner et al.
6,117,631 A 9/2000 Nilsen
6,124,092 A 9/2000 O'Neill et al.
6,138,077 A 10/2000 Brenner
6,140,489 A 10/2000 Brenner
6,172,214 B1 1/2001 Brenner
6,197,506 B1 3/2001 Fodor et al.
6,235,475 B1 5/2001 Brenner et al.
6,242,186 B1 6/2001 Salonen
6,268,152 B1 7/2001 Fodor et al.
6,284,460 B1 9/2001 Fodor et al.
6,300,077 B1 10/2001 Shuber et al.
6,309,822 B1 10/2001 Fodor et al.
6,309,823 B1 10/2001 Cronin et al.
6,326,148 B1 12/2001 Pauletti et al.
6,355,431 B1 3/2002 Chee et al.
6,355,432 B1 3/2002 Fodor et al.
6,395,491 B1 5/2002 Fodor et al.
6,406,848 B1 6/2002 Bridgham et al.
6,440,667 B1 8/2002 Fodor et al.
6,440,706 B1 8/2002 Vogelstein et al.
6,451,536 B1 9/2002 Fodor et al.
6,458,530 B1 10/2002 Morris et al.
6,468,744 B1 10/2002 Cronin et al.
6,489,114 B2 12/2002 Laayoun et al.
6,492,121 B2 12/2002 Kurane et al.
6,498,012 B2 12/2002 Laken
6,503,718 B2 1/2003 Shuber et al.
6,512,105 B1 1/2003 Hogan et al.
6,514,699 B1 2/2003 O'Neill et al.
6,544,739 B1 4/2003 Fodor et al.
6,551,784 B2 4/2003 Fodor et al.
6,576,424 B2 6/2003 Fodor et al.
6,582,908 B2 6/2003 Fodor et al.
6,586,177 B1 7/2003 Shuber
6,600,996 B2 7/2003 Webster et al.
6,629,040 B1 9/2003 Goodlett et al.
6,653,077 B1 11/2003 Brenner
6,753,147 B2 6/2004 Vogelstein et al.
6,849,403 B1 2/2005 Shuber
6,849,404 B2 2/2005 Park et al.
6,852,488 B2 2/2005 Fodor et al.
6,858,412 B2 2/2005 Willis et al.
6,964,846 B1 11/2005 Shuber
7,163,789 B2 1/2007 Chen et al.
7,208,275 B2 4/2007 Gocke et al.
7,406,385 B2 7/2008 Sorenson
7,410,764 B2 8/2008 Gocke et al.
7,424,368 B2 9/2008 Huang et al.
7,424,371 B2 9/2008 Kamentsky
7,537,897 B2 5/2009 Brenner et al.
7,700,286 B2 4/2010 Stroun et al.
7,704,687 B2 4/2010 Wang et al.
7,727,720 B2 6/2010 Dhallan
7,803,929 B2 9/2010 Melkonyan et al.
7,811,757 B2 10/2010 Shuber
7,822,555 B2 10/2010 Huang et al.
7,824,889 B2 11/2010 Vogelstein et al.
7,838,647 B2 11/2010 Hahn et al.
7,915,015 B2 3/2011 Vogelstein et al.
7,935,487 B2 5/2011 Gocke et al.
7,937,225 B2 5/2011 Mishra et al.
7,957,913 B2 6/2011 Chinitz et al.
7,972,817 B2 7/2011 Kopreski
7,981,612 B2 7/2011 Shuber et al.
8,094,312 B2 1/2012 Ulmer
8,168,385 B2 5/2012 Brenner
8,195,415 B2 6/2012 Fan et al.
8,209,130 B1 6/2012 Kennedy et al.
8,216,789 B2 7/2012 Disis et al.
8,236,532 B2 8/2012 Ronaghi et al.
8,361,726 B2 1/2013 Gocke et al.
8,383,338 B2 2/2013 Kitzman et al.
8,383,345 B2 2/2013 Shendure et al.
8,455,193 B2 6/2013 Travers et al.
8,470,996 B2 6/2013 Brenner
8,481,292 B2 7/2013 Casbon et al.
8,603,749 B2 12/2013 Gillevet
8,685,678 B2 4/2014 Casbon et al.
8,697,408 B2 4/2014 Kucera et al.
8,704,165 B2 4/2014 Huang
8,715,967 B2 5/2014 Casbon et al.
8,722,368 B2 5/2014 Casbon et al.
8,728,766 B2 5/2014 Casbon et al.
8,741,606 B2 6/2014 Casbon et al.
8,775,092 B2 7/2014 Colwell et al.
8,835,358 B2 9/2014 Fodor et al.
8,865,410 B2 10/2014 Shendure et al.
9,018,365 B2 4/2015 Brenner
9,085,798 B2 7/2015 Chee
9,260,753 B2 2/2016 Xie et al.
9,340,830 B2 5/2016 Lipson et al.
9,376,719 B2 6/2016 Eijk et al.
9,404,156 B2 8/2016 Hicks et al.
9,476,095 B2 10/2016 Vogelstein et al.
9,598,731 B2 3/2017 Talasaz
9,752,188 B2 9/2017 Schmitt et al.
9,834,822 B2 12/2017 Talasaz
9,840,743 B2 12/2017 Talasaz
9,920,366 B2 3/2018 Eltoukhy et al.
10,287,631 B2 5/2019 Salk et al.
10,370,713 B2 8/2019 Salk et al.
10,385,393 B2 8/2019 Salk et al.
10,388,403 B2 8/2019 Rava et al.
10,604,804 B2 3/2020 Salk et al.
10,752,951 B2 8/2020 Salk et al.
2001/0053519 A1 12/2001 Fodor et al.
2002/0072058 A1 6/2002 Voelker et al.
2003/0003490 A1 1/2003 Fan et al.
2003/0049616 A1 3/2003 Brenner et al.
2003/0104436 A1 6/2003 Morris et al.
2003/0152490 A1 8/2003 Trulson et al.
2003/0165978 A1 9/2003 Firth et al.
2003/0186251 A1 10/2003 Dunn et al.
2003/0207300 A1 11/2003 Matray et al.
2004/0096892 A1 5/2004 Wang et al.
2004/0146901 A1 7/2004 Morris et al.
2004/0157243 A1 8/2004 Huang et al.
2004/0209299 A1 10/2004 Pinter et al.
2004/0259118 A1 12/2004 Macevicz
2005/0164241 A1 7/2005 Hahn et al.
2005/0221314 A1 10/2005 Berlin et al.
2005/0250147 A1 11/2005 Macevicz
2006/0035258 A1 2/2006 Tadakamalla et al.
2006/0046258 A1 3/2006 Lapidus et al.
2006/0073506 A1 4/2006 Christians et al.
2006/0211030 A1 9/2006 Brenner
2007/0020640 A1 1/2007 McCloskey et al.
2007/0065823 A1 3/2007 Dressman et al.
2007/0111233 A1 5/2007 Bianchi et al.
2007/0128724 A1 6/2007 Miles et al.
2007/0172839 A1 7/2007 Smith et al.
2007/0172873 A1 7/2007 Brenner et al.
2008/0014146 A1 1/2008 Hoff et al.
2008/0124721 A1 5/2008 Fuchs et al.
2008/0161420 A1 7/2008 Shuber
2008/0261204 A1 10/2008 Lexow
2008/0293055 A1 11/2008 Freeman et al.
2009/0029377 A1 1/2009 Lo et al.
2009/0087847 A1 4/2009 Lo et al.
2009/0098547 A1 4/2009 Ghosh
2009/0105959 A1 4/2009 Braverman et al.
2009/0162836 A1 6/2009 Widschwendter
2009/0298075 A1 12/2009 Travers et al.
2009/0298709 A1 12/2009 Ma
2009/0317818 A1 12/2009 Ehrich et al.
2010/0041048 A1 2/2010 Diehl et al.
2010/0062494 A1 3/2010 Church et al.
2010/0069250 A1 3/2010 White et al.
2010/0112575 A1 5/2010 Fan et al.
2010/0143932 A1 6/2010 Lapidus
2010/0166744 A1 7/2010 Wong

(56) **References Cited**

U.S. PATENT DOCUMENTS

2010/0196898	A1	8/2010	Sugarbaker et al.	
2010/0264331	A1*	10/2010	Sacko	G01N 21/6428 250/459.1
2010/0323348	A1	12/2010	Hamady et al.	
2010/0330571	A1	12/2010	Robins et al.	
2011/0014607	A1	1/2011	Jirtle et al.	
2011/0160078	A1	6/2011	Fodor et al.	
2011/0160290	A1	6/2011	Tewari	
2011/0171640	A1	7/2011	Bhatt et al.	
2011/0177512	A1	7/2011	Shuber	
2011/0183330	A1	7/2011	Lo et al.	
2011/0201507	A1	8/2011	Rava et al.	
2011/0230358	A1	9/2011	Rava	
2011/0230360	A1	9/2011	Stephan et al.	
2011/0245482	A1	10/2011	Hahn et al.	
2011/0264376	A1	10/2011	Chinitz et al.	
2011/0275084	A1	11/2011	Byron et al.	
2011/0281736	A1	11/2011	Drmanac et al.	
2011/0319272	A1	12/2011	Fan et al.	
2012/0003637	A1	1/2012	Lo et al.	
2012/0034685	A1	2/2012	Sparks et al.	
2012/0046877	A1	2/2012	Hyland et al.	
2012/0053073	A1	3/2012	Kassis	
2012/0059594	A1	3/2012	Hatchwell et al.	
2012/0065081	A1	3/2012	Chee	
2012/0095697	A1	4/2012	Halpern et al.	
2012/0100548	A1	4/2012	Rava et al.	
2012/0164630	A1	6/2012	Porreca et al.	
2012/0165202	A1	6/2012	Porreca et al.	
2012/0191367	A1	7/2012	Stuelpnagel et al.	
2012/0208706	A1	8/2012	Downing et al.	
2012/0214163	A1	8/2012	Sugarbaker et al.	
2012/0214678	A1	8/2012	Rava et al.	
2012/0220478	A1	8/2012	Shaffer	
2012/0231479	A1	9/2012	Puskas et al.	
2012/0238464	A1	9/2012	Koi et al.	
2012/0270212	A1	10/2012	Rabinowitz et al.	
2012/0316074	A1	12/2012	Saxonov	
2012/0322681	A1	12/2012	Kung et al.	
2013/0005585	A1	1/2013	Anderson et al.	
2013/0017958	A1	1/2013	Benz et al.	
2013/0022977	A1	1/2013	Lapidus et al.	
2013/0029852	A1	1/2013	Rava et al.	
2013/0034546	A1	2/2013	Rava et al.	
2013/0040824	A1	2/2013	Lo et al.	
2013/0053256	A1	2/2013	Hubbell	
2013/0060483	A1	3/2013	Struble et al.	
2013/0078626	A1	3/2013	Wasserstrom et al.	
2013/0085681	A1	4/2013	Deciu et al.	
2013/0096011	A1	4/2013	Rava et al.	
2013/0102485	A1	4/2013	Lee	
2013/0102487	A1	4/2013	Cos et al.	
2013/0116127	A1	5/2013	Schuetz et al.	
2013/0116130	A1	5/2013	Fu et al.	
2013/0122499	A1	5/2013	Morris et al.	
2013/0130923	A1	5/2013	Ehrich et al.	
2013/0137588	A1	5/2013	Shendure et al.	
2013/0143747	A1	6/2013	Gutin et al.	
2013/0210643	A1	8/2013	Casbon et al.	
2013/0210645	A1	8/2013	Volgelstein et al.	
2013/0224743	A1	8/2013	Casbon et al.	
2013/0237431	A1	9/2013	Lo et al.	
2013/0237458	A1	9/2013	Casbon et al.	
2013/0260381	A1	10/2013	Ramakrishnan	
2013/0267424	A1	10/2013	Casbon et al.	
2013/0288244	A1*	10/2013	Deciu	G16B 20/00 435/6.11
2013/0338933	A1	12/2013	Deciu et al.	
2014/0011694	A1	1/2014	Couronne	
2014/0065609	A1	3/2014	Hicks et al.	
2014/0065630	A1	3/2014	Bubnoff et al.	
2014/0066317	A1	3/2014	Talasaz	
2014/0100121	A1	4/2014	Lo et al.	
2014/0242588	A1	8/2014	Boom et al.	
2014/0274731	A1	9/2014	Raymond et al.	
2014/0274740	A1	9/2014	Srinivasan et al.	
2014/0296081	A1	10/2014	Diehn et al.	
2014/0296094	A1	10/2014	Domanus	
2014/0303008	A1	10/2014	Schutz et al.	
2014/0336943	A1	11/2014	Pellini et al.	
2014/0350130	A1	11/2014	Sanborn et al.	
2014/0357500	A1	12/2014	Vigneault et al.	
2015/0004158	A1	1/2015	Shipp et al.	
2015/0024950	A1	1/2015	Bielas et al.	
2015/0044687	A1	2/2015	Schmitt et al.	
2015/0050647	A1	2/2015	Luo et al.	
2015/0051085	A1	2/2015	Vogelstein et al.	
2015/0065358	A1	3/2015	Comstock et al.	
2015/0080266	A1	3/2015	Volkmuth et al.	
2015/0087535	A1	3/2015	Patel	
2015/0167069	A1	6/2015	Schutz et al.	
2015/0197786	A1	7/2015	Osborne et al.	
2015/0275289	A1	10/2015	Otwinowski et al.	
2015/0299812	A1	10/2015	Talasaz	
2015/0329917	A1	11/2015	Shuber	
2015/0344970	A1	12/2015	Vogelstein et al.	
2015/0366866	A1	12/2015	Ali et al.	
2015/0368708	A1	12/2015	Talasaz	
2016/0002739	A1	1/2016	Schütz et al.	
2016/0002741	A1	1/2016	Kitano et al.	
2016/0024576	A1	1/2016	Chee	
2016/0026758	A1	1/2016	Jabara et al.	
2016/0032396	A1	2/2016	Diehn et al.	
2016/0040229	A1	2/2016	Talasaz et al.	
2016/0046986	A1	2/2016	Eltoukhy et al.	
2016/0053301	A1	2/2016	Raymond et al.	
2016/0060691	A1	3/2016	Giresi et al.	
2016/0071432	A1	3/2016	Kurowski et al.	
2016/0115553	A1	4/2016	Stephan et al.	
2016/0130649	A1	5/2016	Xie et al.	
2016/0251704	A1	9/2016	Talasaz et al.	
2016/0319345	A1	11/2016	Gnerre et al.	
2016/0333417	A1	11/2016	Talasaz	
2016/0376647	A1	12/2016	Travers et al.	
2017/0051347	A1	2/2017	Vogelstein et al.	
2017/0073774	A1	3/2017	Lo et al.	
2017/0145516	A1	5/2017	Kopetz et al.	
2017/0159120	A1	6/2017	Eijk et al.	
2017/0211143	A1	7/2017	Shendure et al.	
2017/0218459	A1	8/2017	Talasaz et al.	
2017/0218460	A1	8/2017	Talasaz	
2018/0023125	A1	1/2018	Talasaz et al.	
2019/0271040	A1	9/2019	Salk et al.	
2019/0292597	A1	9/2019	Salk et al.	
2019/0338358	A1	11/2019	Salk et al.	
2019/0352714	A1	11/2019	Salk et al.	
2021/0222243	A1	7/2021	Bielas	

FOREIGN PATENT DOCUMENTS

EP	1647600	A2	6/2006
EP	2110442	A1	10/2009
EP	3070177	A1	9/2016
EP	3178941	A1	6/2017
WO	1997007241	A1	2/1997
WO	1997010365	A1	3/1997
WO	1999028505	A1	6/1999
WO	2000058516	A2	10/2000
WO	2002056014	A2	7/2002
WO	2005080604	A2	9/2005
WO	2005111242	A2	11/2005
WO	2006102264	A1	9/2006
WO	2007037678	A2	4/2007
WO	2008070144	A2	6/2008
WO	2008154317	A1	12/2008
WO	2009152928	A2	12/2009
WO	2011060240	A1	5/2011
WO	2011087760	A2	7/2011
WO	2011091046	A1	7/2011
WO	2011103236	A2	8/2011
WO	2011140510	A2	11/2011
WO	2011155833	A2	12/2011
WO	2012012693	A2	1/2012
WO	2012014877	A1	2/2012

(56) **References Cited**

FOREIGN PATENT DOCUMENTS

WO	2012019200	A2	2/2012
WO	2012028746	A1	3/2012
WO	2012038839	A2	3/2012
WO	2012042374	A2	4/2012
WO	2012048341	A1	4/2012
WO	2012054873	A2	4/2012
WO	2012066451	A1	5/2012
WO	2012071621	A1	6/2012
WO	2012088348	A2	6/2012
WO	2012099832	A2	7/2012
WO	2012106559	A1	8/2012
WO	2012129363	A2	9/2012
WO	2012142213	A2	10/2012
WO	2012142611	A2	10/2012
WO	2012148477	A1	11/2012
WO	2012103031	A3	1/2013
WO	2013019075	A2	2/2013
WO	2013033721	A1	3/2013
WO	2013106737	A1	7/2013
WO	2013123442	A1	8/2013
WO	2013130512	A2	9/2013
WO	2013130674	A1	9/2013
WO	2013138510	A1	9/2013
WO	2013142213	A1	9/2013
WO	2013142389	A1	9/2013
WO	2013148496	A1	10/2013
WO	2013159035	A2	10/2013
WO	2013173394	A2	11/2013
WO	2013181170	A1	12/2013
WO	2013188471	A2	12/2013
WO	2013190441	A2	12/2013
WO	2014004726	A1	1/2014
WO	2014014497	A1	1/2014
WO	2014015319	A1	1/2014
WO	2014039556	A1	3/2014
WO	2014093330	A1	6/2014
WO	2014145078	A1	9/2014
WO	2014149134	A2	9/2014
WO	2014151117	A1	9/2014
WO	2014152990	A1	9/2014
WO	2015159293	A2	10/2015
WO	2016015058	A2	1/2016
WO	2016040901	A1	3/2016
WO	2017100441	A1	6/2017
WO	2017181146	A1	10/2017

OTHER PUBLICATIONS

Nord, et al. Accurate and exact CNV identification from targeted high-throughput sequence data. BMC Genomics. Apr. 12, 2011;12:184.

Notice of Allowance dated Jan. 31, 2017 for U.S. Appl. No. 14/712,754.

Notice of allowance dated Mar. 21, 2014 for U.S. Appl. No. 12/969,581.

Notice of allowance dated Jun. 15, 2017 for U.S. Appl. No. 15/076,565.

Notice of allowance dated Jun. 19, 2014 for U.S. Appl. No. 12/969,581.

Notice of allowance dated Aug. 1, 2017 for U.S. Appl. No. 15/492,659.

Notice of allowance dated Aug. 4, 2017 for U.S. Appl. No. 15/467,570.

Notice of allowance dated Aug. 22, 2014 for U.S. Appl. No. 12/969,581.

Notice of allowance dated Aug. 29, 2017 for U.S. Appl. No. 15/492,659.

Notice of allowance dated Sep. 11, 2017 for U.S. Appl. No. 15/076,565.

Notice of allowance dated Oct. 3, 2017 for U.S. Appl. No. 15/076,565.

Notice of allowance dated Oct. 25, 2017 for U.S. Appl. No. 14/861,989.

Notice of allowance dated Dec. 28, 2017 for U.S. Appl. No. 14/861,989.

Office Action dated Feb. 9, 2017 for U.S. Appl. No. 15/076,565.

Office action dated May 13, 2019 for U.S. Appl. No. 15/669,779.

Office action dated May 20, 2016 for U.S. Appl. No. 14/855,301.

Office action dated May 31, 2016 for U.S. Appl. No. 14/712,754.

Office action dated Jun. 1, 2017 for U.S. Appl. No. 15/467,570.

Office action dated Jun. 3, 2016 for U.S. Appl. No. 14/861,989.

Office action dated Jun. 12, 2017 for U.S. Appl. No. 15/492,659.

Office action dated Jul. 18, 2017 for U.S. Appl. No. 14/861,989.

Office Action dated Jul. 30, 2019 for U.S. Appl. No. 16/283,635.

Office action dated Aug. 6, 2019 for U.S. Appl. No. 14/855,301.

Office action dated Oct. 3, 2013 for U.S. Appl. No. 12/969,581.

Office Action dated Oct. 12, 2016 for U.S. Appl. No. 14/712,754.

Office action dated Oct. 20, 2017 for U.S. Appl. No. 14/425,189.

Office Action dated Nov. 21, 2016 for U.S. Appl. No. 14/855,301.

Office action dated Dec. 4, 2015 for U.S. Appl. No. 14/712,754.

Office action dated Dec. 7, 2017 for U.S. Appl. No. 14/855,301.

Office Action dated Dec. 9, 2016 for U.S. Appl. No. 14/861,989.

Office action dated Dec. 21, 2018 for U.S. Appl. No. 15/978,848.

Ogino, et al. Quantification of PCR bias caused by a single nucleotide polymorphism in SMN gene dosage analysis. J Mol Diagn. Nov. 2002;4(4):185-90.

Opposition Brief dated Aug. 7, 2018 for Japanese Opposition No. 2018-700659 to JP Patent 6275145.

Opposition Decision dated Jan. 15, 2019 for Japanese Opposition No. 2018-700659 to JP Patent 6275145.

Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Foundation Medicine Inc.

Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Grunecker Patent-und Rechtsanwalte PartG mbB.

Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Personal Genome Diagnostics, Inc.

Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Strawman Limited.

Opposition Form and Statement to EP3087204 filed Nov. 14, 2018.

Pacific Biosciences. Template Preparation and Sequencing Guide. Publication date: Oct. 14, 2014. Pacific Biosciences website http://www.pacificbiosciences.com/support/pubmap/documentation.html.

Pan, et al. Loss of heterozygosity patterns provide fingerprints for genetic heterogeneity in multistep cancer progression of tobacco smoke-induced non-small cell lung cancer. Cancer Res. Mar. 1, 2005;65(5):1664-9.

Park, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet. May 2010;42(5):400-5. doi: 10.1038/ng.555. Epub Apr. 4, 2010.

Parkinson, N.J. et al., "Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA," Genome Res. 2012, 22(1), 125-133.

Pel, J. et al. "Duplex proximity sequencing (pro-seq): A method to improve DNA sequencing accuracy without the cost of molecular barcoding redundancy" bioRxiv (2017) https://doi.org/10.1101/163444.

Perakis, S. et al. "Advances in Circulating Tumor DNA Analysis" Adv Clin Chem (2017) pp. 1-81.

Pihlak, et al. Rapid genome sequencing with short universal tiling probes. Nature Biotechnology, 26: 676-684 (2008).

Guardant Health vs. *FMI* Invalidity Exhibit C-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit C-4, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit C-5, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit D-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit D-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit D-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit D-5, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Second Amended Complaint dated Mar. 6, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Suppl Invalidity References dated Feb. 15, 2019 (C.A. No. 17-cv-1616-LPS-CJB).

(56) **References Cited**

OTHER PUBLICATIONS

Guardant Health vs. *FMI* Supplemental Invalidity Contentions dated Mar. 29, 2019 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *PGDx* Amended Answer to Second Amended Complaint dated Apr. 30, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart A, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart B, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart C, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart D, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart E, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart F, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart G, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart H, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart I, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart J, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart K, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Claim Chart L, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Invalidity Contentions dated Jun. 25, 2018 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Second Amended Complaint dated Mar. 23, 2018 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Second Updated Invalidity Contentions dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Guardant Health vs. *PGDx* Updated Invalidity Contentions dated Mar. 29, 2019 (C.A. No. 17-cv-1623-LPS-CJB).
Gundry, et al. "Direct mutation analysis by high-throughput sequencing: from gremlin to low-abundant, somatic variants" Mutat Res Jan. 3, 2012; 729(1-2):1-15. doi: 10.1016/mrfmmm.2011.10.001. Pub Oct. 12, 2011.
Gundry, et al. Direct, genome-wide assessment of DNA mutations in single cells. Nucleic Acids Res. Mar. 2012;40(5):2032-40. doi: 10.1093/nar/gkr949. Epub Nov. 15, 2011.
Hacia, et al. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nature Genetics, 22: 164-167 (1999).
Hafner, et al., RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA eDNA libraries, RNA Sep. 1, 2011, 17(9):1697-1712.
Hall, A., Short-Read DNA Sequence Alignmnet with Custom Designed FPGA-based Hardware, Thesis 2010.
Hamady, et al. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methods. Mar. 2008;5(3):235-7. doi: 10.1038/nmeth.1184. Epub Feb. 10, 2008.
Hensel, et al. Simultaneous identification of bacterial virulence genes by negative selection. Science. Jul. 21, 1995;269(5222):400-3.
Hiatt, et al. Parallel, tag-directed assembly of locally derived short sequence reads. Nat Methods. Feb. 2010;7(2):119-22. doi: 10.1038/nmeth.1416. Epub Jan. 17, 2010.
Hiatt, et al. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. Genome Res. May 2013;23(5):843-54. doi: 10.1101/gr.147686.112. Epub Feb. 4, 2013.
Hibi, et al. Molecular detection of genetic alterations in the serum of colorectal cancer patients. Cancer Res. Apr. 1, 1998;58(7):1405-7.
HiSeq 2000 User Guide (2010).
Holdenrieder, et al. Circulating nucleosomes and cytokeratin 19-fragments in patients with colorectal cancer during chemotherapy. Anticancer Res. May-Jun. 2005;25(3A):1795-801.
Holies, et al. A stochastic approach to count RNA molecules using DNA sequencing methods. 2003, Lecture Notes in Computer Science, vol. 2812, pp. 55-62.
Hollas, et al. A stochastic approach to count RNA molecules using DNA sequencing methods. Lecture Notes in Computer Science, 2812: 55-62 (2003).
Hollstein, M. et al. "p53 mutations in human cancers" Science (1991) 253(5015):49-53.
Hoque, et al. Detection of aberrant methylation of four genes in plasma DNA for the detection of breast cancer. J Clin Oncol. Sep. 10, 2006;24(26):4262-9. Epub Aug. 14, 2006.
Howe, et al. Retinoblastoma growth suppressor and a 300-kDa protein appear to regulate cellular DNA synthesis. Proc Natl Acad Sci U S A. Aug. 1990;87(15):5883-7.
Hug, et al. Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. J Theor Biol. Apr. 21, 2003;221(4):615-24.
Hyland, et al. The normal and tumor spectrum of copy number variation: Copy number alterations correlate with changes in gene expression in tumor transcriptome. Nov. 15, 2009. 1 page. Retreived from:http://tools.thermofisher.com/content/sfs/posters/cms_073633.pdf.
Iafrate, et al. Detection of large-scale variation in the human genome. Nat Genet. Sep. 2004;36(9):949-51. Epub Aug. 1, 2004.
Ikeguchi, et al. Detection of loss of heterozygosityat microsatellite loci in esophageal squamous-cell carcinoma. Oncology. 1999;56(2):164-8.
Illumina "Multiplexed Sequencing with Illumina Genome Analyzer System" https://www.illumina.com/Documents/products/datasheets/datasheet_sequencing_multiplex.pdf (Dec. 2008).
Illumina Preparing Samples for Paired-End Sequencing (2008).
Koyanagi, et al. Association of circulating tumor cells with serum tumor-related methylated DNA in peripheral blood of melanoma patients. Cancer Res. Jun. 15, 2006;66(12):6111-7.
Krimmel, J.D. et al., "Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues," Proc. Natl. Acad. Sci. USA 2016, 113(21), 6005-6010.
Kukita, Y. et al. "High-fidelity target sequencing of individual molecules identified using barcode sequences: de novo detection and absolute quantitation of mutations in plasma cell-free DNA from cancer patients" DNA Research (2015) 22(4):269-277.
Lam, et al. Plasma DNA as a prognostic marker for stroke patients with negative neuroimaging within the first 24 h of symptom onset. Resuscitation. Jan. 2006;68(1):71-8. Epub Dec. 1, 2005.
Lanman, et al., Analytical and Clinical Validation of a Digital Sequencing Panel for Quantitative, Highly Accurate Evaluation of Cell-Free Circulating Tumor DNA PLoS One, Oct. 2015, 10(10), e0140712. doi:10.1371/journal.pone.0140712.
Larson, et al. A single molecule view of of gene expression. Trends Cell Biol. Nov. 2009;19(11):630-7. Epub Oct. 8, 2009.
Leary, et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Sci Transl Med. Nov. 28, 2012;4(162):162ra154. doi: 10.1126/scitranslmed.3004742.
Leary, et al. Development of personalized tumor biomarkers using massively parallel sequencing. Sci Transl Med. Feb. 24, 2010;2(20):20ra14. doi: 10.1126/scitranslmed.3000702.
Lecomte, et al. Detection of free-circulating tumor-associated DNA in plasma of colorectal cancer patients and its association with prognosis. Int J Cancer. Aug. 10, 2002;100(5):542-8.
Ledergerber, C. et al., "Base-calling for next-generation sequencing platforms," Brief Bioinform. 2011 12(5), 489-497.
Leon, et al. Free DNA in the serum of cancer patients and the effect of therapy. Cancer Res. Mar. 1977;37(3):646-50.
Leung, et al. Plasma Epstein-Barr viral deoxyribonucleic acid quantitation complements tumor-node-metastasis staging prognostication in nasopharyngeal carcinoma. J Clin Oncol. Dec. 1, 2006;24(34):5414-8.

(56) **References Cited**

OTHER PUBLICATIONS

Li, et al. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. Nov. 2008;18(11):1851-8. doi: 10.1101/gr.078212.108. Epub Aug. 19, 2008.

Li, et al. Structure-independent and quantitative ligation of single-stranged DNA. Anal Biochem. Feb. 15, 2006;349(2):242-6. Epub Nov. 18, 2005.

Li, H. et al., "Fast and accurate short read alignment with Burrows-Wheeler transform," Bioinformatics 2009, 25(14), 1754-1760.

Liang, K-C. et al., "Bayesian basecalling for DNA sequence analysis using hidden Markov models," IEE/ACM Trans. Comput. Biol. Bioinform. 2007 4(3), 430-440.

Liao, G.J.W. et al., "Targeted massively parallel sequencing of maternal plasma DNA permits efficient and unbiased detection of fetal alleles," Clin. Chem. 2011, 57(1), 92-101.

Lizardi, et al. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. Nat Genet. Jul. 1998;19(3):225-32.

Lo, et al. Quantitative Analysis of Cell-free Epstein-Barr Virus DNA in Plasma of Patients with Nasopharyngeal Carcinoma, Cancer Research, 59(6):1188-1191 (Mar. 1999).

Lo, Y.M.D et al., "Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis," Am. J. Hum. Genet. 1998, 62(4), 768-775.

Lo, Y.M.D. et al. "Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus," Sci Transl Med (2010) 2(61):1-13.

Lockhart, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology, 14: 1675-1680 (1996).

Lodes, et al. Novel Chimera-Free, High Efficiency Library Preparation for NGS Platforms.Poster. Presented at the Plant & Animal Genome XX Conference. San Diego, California. Jan. 14-18, 2012. 1 page.

Lucito, et al. Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. Genome Res. Nov. 2000;10(11):1726-36.

Lucito, et al. Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation. Genome Research (2003), 13: 2291-2305.

Lunter, G. et al., "Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads," Genome Res. 2011, 21(6):936-939.

Maamar, et al. Noise in Gene Expression Determines Cell Fate in Bacillus subtilis. Science, 317: 526-529 (2007).

Mahmud, et al. Fast MCMC sampling for hidden Markov Models to determine copy number variations. BMC Bioinformatics. Nov. 2, 2011;12:428. doi: 10.1186/1471-2105-12-428.

Makrigiorgos, et al., A PCR-Based amplification method retaining quantative difference between two complex genomes. Nature Biotech, vol. 20, No. 9, pp. 936-939 (Sep. 2002).

Mamanova, L. et al., "Target-enrichment strategies for next-generation sequencing," Nat. Methods 2010, 7(2), 111-118.

Mandel, et al. Les acides nucleiques du plasma sanguin chez l'homme. C R Seances Soc Biol Fil. Feb. 1948;142(3-4):241-3.

Marsit, et al. Epigenetic profiling reveals etiologically distinct patterns of DNA methylation in head and neck squamous cell carcinoma. Carcinogenesis. Mar. 2009;30(3):416-22. doi: 10.1093/carcin/bgp006. Epub Jan. 6, 2009.

McCloskey, et al. Encoding PCR products with batch-stamps and barcodes. Biochem Genet. Dec. 2007;45(11-12):761-7. Epub Oct. 23, 2007.

McKernon, K.J. et al., "Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding," Genome Res. 2009, 19(9), 1527-1541.

Medvedev, et al. Detecting copy number variation with mated short reads. Genome Res. Nov. 2010;20(11):1613-22. doi: 10.1101/gr.106344.110. Epub Aug. 30, 2010.

Mei, et al. Identification of recurrent regions of Copy-Number Variants across multiple individuals. BMC Bioinformatics. Mar. 22, 2010;11:147. doi: 10.1186/1471-2105-11-147.

Meldrum, et al. Next generation sequencing for cancer diagnostics: a practical perspective. Clin Biochem Rev. Nov. 2011;32(4):177-95.

Mertes, F. et al. "Targeted enrichment of genomic DNA regions for next-generation sequencing" Brief Functional Genomics (2011) 10(6):374-386.

Metzker, M.L. "Sequencing technologies—the next generation" Nature Reviews Genetics (2010) 11:31-46.

Meyer, M. et al. "Illumina sequencing library preparation for highly multiplexed target capture and sequencing," Cold Spring Harb. Protoc. 2010, (6), prot5448.

Meyerson, M. et al. "Advances in understanding cancer genomes through second-generation sequencing" Nature Reviews Genetics (2010) 11:685-696.

Milbury, et al. "Enabling Clinical Cancer Genomics for Rare Mutations: COLD-PCR Magnifies Mutations Prior to Targeted Amplicon Re-Sequencing". Clin Chem. Mar. 2012;58(3):580-9. doi: 10.1373/clinchem.2011.176198. Epub Dec. 21, 2011.

Miner, et al. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. Nucleic Acids Res. Sep. 30, 2004;32(17):e135.

Mitchell, et al. Circulating microRNAS as stable blood-based markers for cancer detection. Proc Natl Acad Sci U S A. Jul. 29, 2008;105(30):10513-8. Epub Jul. 28, 2008.

Mitelman, F. et al. "The impact of translocations and gene fusions on cancer causation" Nature Rev. Cancer (2007) 7(4):233-245.

Mohan, et al.,Changes in colorectal carcinoma genomes under anti-EGFR therapy identified by whole-genome plasma DNA sequencing.,PLoS Genet,doi: 10.1371/journal.pgen.1004271. eCollection 2014.,Mar. 27, 2014,10(3), e1004271.

Mori, et al. Predictive utility of circulating methylated DNA in serum of melanoma patients receiving biochemotherapy. J Clin Oncol. Dec. 20, 2005;23(36):9351-8.

Morrissy, A.S. et al. "Next-generation tag sequencing for cancer gene expression profiling" Genome Research (2009) 19(10): 1825-1835.

Mortazavi, et al. Mapping and quantifying mammalian transcriptomes by RNA—Seq. Nat. Methods. 2008; 5, 621-628.

Narayan, et al. Ultrasensitive measurement of hotspot mutations in tumor DNA in blood using error-suppressed multiplexed deep sequencing. Cancer Res. Jul. 15, 2012;72(14):3492-8. doi: 10.1158/0008-5472.CAN-11-4037. Epub May 10, 2012.

UCSC Genome Bioinformatics. About the UCSC Genome Bioinformatics Site. http://genome.ucsc.edu/index.html. Accessed on May 26, 2015. 2 pgs.

Umetani et al. Prediction of Breast Tumor Progression by Integrity of Free Circulating DNA in Serum. Journal of Clinical Oncology 24(26):4270-4276 (Sep. 10, 2006).

U.S. Appl. No. 61/384,001, filed Sep. 17, 2010.

U.S. Appl. No. 61/432,119, filed Jan. 12, 2011.

Utting, et al. Microsatellite analysis of free tumor DNA in urine, serum, and plasma of patients: a minimally invasive method for the detection of bladder cancer. Clin Cancer Res. Jan. 2002;8(1):35-40.

Van Houten, et al. Mutated p53 as a molecular marker for the diagnosis of head and neck cancer. J Pathol. Dec. 2002;198(4):476-86.

Van Orsouw, N. et al. "Complexity Reduction of Polymorphic Sequences (CRoPSTM): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes" PLoS One (2007) 11(e1172):1-10.

Vasyukhin, V. et al., "K-Ras Point Mutations in the Blood Plasma DNA of Patients with Colorectal Tumors" Challenges of Modern Medicine, 141-150 (Verna and Shamoo eds, 1994).

Velculescu, et al. Characterization of the Yeast Transcriptome. Cell, 88: 243-251 (1997).

Velculescu, et al. Serial Analysis of Gene Expression. Science, 270: 484-487 (1995).

Vogelstein, et al. Digital PCR. Proc. Natl. Acad. Sci., 96(16): 9236-9241(1999).

(56) **References Cited**

OTHER PUBLICATIONS

Wagle, N. et al., "High-throughput Detection of actionable Genomic alterations in clinical tumor samples by targeted, Massively Parallel sequencing," Cancer Discov. 2012, (2)1:82-93.

Walker, et al. Isothermal in vitro amplification of DNA by a restriction enzyme/DNA polymerase system. Proc Natl Acad Sci U S A. Jan. 1, 1992;89(1):392-6.

Walsh, et al. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc Natl Acad Sci U S A. Jul. 13, 2010;107(28):12629-33. doi: 10.1073/pnas.1007983107. Epub Jun. 28, 2010.

Wang, et al. iCLIP predicts the dual splicing effects of TIA-RNA interactions, Oct. 2010, PLoS Biol, 8(10):e1000530.

Wang, et al. Molecular detection of APC, K-ras, and p53 mutations in the serum of colorectal cancer patients as circulating biomarkers. World J Surg. Jul. 2004;28(7):721-6. Epub Jun. 8, 2004.

Wang, et al. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10: 57-63 (2009).

Wang, T.L. et al. "Digital Karyotyping" PNAS (2002) 99(25):16156-16161.

Wang, Y. et al., "Detection of somatic mutations and HPV in the saliva and plasma of patients with head and neck squamous cell carcinomas," Sci. Transl. Med. 2015, 7(293):293ra104.

Wang, Y. et al., "Detection of tumor-derived DNA in cerebrospinal fluid of patients with primary tumors of the brain and spinal cord," Proc. Natl. Acad. Sci. USA 2015, 112(31), 9704-9709.

Wang, Y. et al., "Diagnostic potential of tumor DNA from ovarian cyst fluid," eLife 2016, 5:e15175.

Weber, et al. A real-time polymerase chain reaction assay for quantification of allele ratios and correction of amplification bias. Anal Biochem. Sep. 15, 2003;320(2):252-8.

Wheeler, D.A. et al., "The complete genome of an individual by massively parallel DNA sequencing," Nature 2008, 452(7189), 872-876.

Williford, A. et al., "Gene Fusion," eLS 2013, 1-8.

Wittes, et al. Searching for Evidence of Altered Gene Expression: a Comment on Statistical Analysis of Microarray Data. Journal of the National Cancer Institute, 91(5): 400-401 (1999).

Wodicka, et al. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nature Biotechnology, 15: 1359-1367 (1997).

Wojdacs, et al. Primer design versus PCR bias in methylation independent PCR amplifications. Epigenetics. May 16, 2009;4(4):231-4. Epub May 14, 2009.

Wood, et al. The genomic landscapes of human breast and colorectal cancers. Science. Nov. 16, 2007;318(5853):1108-13. Epub Oct. 11, 2007.

Wood, et al. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucleic Acids Res. Aug. 2010;38(14):e151. doi: 10.1093/nar/gkq510. Epub Jun. 4, 2010.

Xi, et al. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc Natl Acad Sci U S A. Nov. 15, 2011;108(46):E1128-36. doi: 10.1073/pnas.1110574108. Epub Nov. 7, 2011.

Yandell, et al. A probabilistic disease-gene finder for personal genomes. Genome Res. Sep. 2011;21(9):1529-42. doi: 10.1101/gr.123158.111. Epub Jun. 23, 2011.

Yang, et al., EGFR gene copy number as a predictive biomarker for the treatment of metastatic colorectal cancer with anti-EGFR monoclonal antibodies: a meta-analysis.,J Hematol Oncol, Aug. 16, 2012,5:52,1-9.

Yang. Simple binary segmentation frameworks for identifying variation in DNA copy number. BMC Bioinformatics. Oct. 30, 2012;13:277. doi: 10.1186/1471-2105-13-277.

Ye, et al. Fluorescent microsphere-based readout technology for multiplexed human single nucleotide polymorphism analysis and bacterial identification. Human Mutation, 17(4): 305-316 (2001).

Yoon, et al. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. Sep. 2009;19(9):1586-92. doi: 10.1101/gr.092981.109. Epub Aug. 5, 2009.

Zhang et al. "Comprehensive One-Step Molecular Analysis of Mitochondrial Genome by Massively Parallel Sequencing" Clinical Chem (2012) 58(9):1322-1331.

Zhang, et al. The impact of next-generation sequencing on genomics. J Genet Genomics. Mar. 20, 2011;38(3):95-109. doi: 10.1016/j.jgg.2011.02.003. Epub Mar. 15, 2011.

Zhao et al. Homozygous Deletions and Chromosome Amplifications in Human Lung Carcinomas Revealed by Single Nucleotide Polymorphism Array Analysis. Cancer Research, 65: 5561-5570 (2005).

Zhou, et al. Counting alleles reveals a connection between chromosome 18q loss and vascularinvasion. Nature Biotechnology, 19: 78-81 (2001).

Pinkel, et al. Comparative Genomic Hybridization. Annual Review of Genomics and Human Genetics, 6: 331-354 (2005).

Pleasance, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. Jan. 14, 2010;463(7278):184-90. doi: 10.1038/nature08629. Epub Dec. 16, 2009.

Pollack, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet. Sep. 1999;23(1):41-6.

Qiu, et al. DNA sequence-based "bar codes" for tracking the origins of expressed sequence tags from a maize cDNA library constructed using multiple mRNA sources. Plant Physiol. Oct. 2003;133(2):475-81.

Quail, et al. A large genome center's improvements to the Illumina sequencing system. Nat Methods. Dec. 2008;5(12):1005-10. doi: 10.1038/nmeth.1270.

Quinlan, A.R. et al., "Pyrobayes: an improved base caller for SNP discovery in pyrosequences," Nat. Methods 2008 5(2), 179-181.

Rafi, I. et al. "Cell-free fetal DNA and non-invasive prenatal diagnosis," Br. J. Gen. Pract. May 1, 2009; 59(562):e146-e148.

Rand, et al. Headloop suppression PCR and its application to selective amplification of methylated DNA sequences. Nucleic Acids Res. Aug. 9, 2005;33(14):e127.

Redon, R. et al., "Global variation in copy number in the human genome," Nature 2006 444(7118), 444-454.

Report and Recommendation in Civil Action No. 17-1623-LPS-CJB between *Guardant Health, Inc.* and *Foundation Medicine* and between *Guardant Health, Inc.* and *Personal Genome Diagnostics, Inc.*, dated Oct. 11, 2019.

Rizzo, J.M. et al. "Key Principles and Clinical Applications of 'Next Generation' DNA Sequencing," Cancer Prev. Res., (2012) 5, 887-900.

Ryan, et al. A prospective study of circulating mutant KRAS2 in the serum of patients with colorectal neoplasia: strong prognostic indicator in postoperative follow up. Gut. Jan. 2003;52(1):101-8.

Rygaard, et al. Abnormalities in structure and expression of the retinoblastoma gene in small cell lung cancer cell lines and xenografts in nude mice. Cancer Res. Sep. 1, 1990;50(17):5312-7.

Sanger, F. et al. "DNA sequencing with chain-terminating inhibitors" PNAS (1977) 74(12):5463-5467.

Sausen, M. et al. "Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma", Nature Genetics 2013, 45(1), 12-17.

Schmitt et al. Supplemental Information http://www.pnas.org/content/suppl/2012/08/01/1208715109. DCSupplemental.

Schmitt, et al. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A. Sep. 4, 2012;109(36):14508-13. doi: 10.1073/pnas.1208715109. Epub Aug. 1, 2012.

Schwarzenbach, et al. A critical evaluation of loss of heterozygosity detected in tumor tissues, blood serum and bone marrow plasma from patients with breast cancer. Breast Cancer Res. 2007;9(5):R66.

Schwarzenbach, et al. Cell-free tumor DNA in blood plasma as a marker for circulating tumor cells in prostate cancer. Clin Cancer Res. Feb. 1, 2009;15(3):1032-8. doi: 10.1158/1078-0432.CCR-08-1910.

Schwarzenback, H. et al. "Cell-free nucleic acids as biomarkers in cancer patients" Nature Reviews Cancer (2011) 11:426-437.

(56) **References Cited**

OTHER PUBLICATIONS

Schweiger et al., "Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis," PLoS One 2009, 4(5), e5548.
Sebat, et al. Large-Scale Copy Number Polymorphism in the Human Genome. Science, 305: 525-528 (2004).
Sehnert, A.J. et al. "Optimal Detection of Fetal Chromosomal Abnormalities by Massively Parallel DNA Sequencing of Cell-Free Fetal DNA from Maternal Blood" Clin Chem (2011) 57(7):1042-1049.
Shaw et al. Genomic analysis of circulating cell-free DNA infers breast cancer dormancy. Genome Research 22(2):220-231 (Feb. 2012).
Shaw, et al. Microsatellite alterations plasma DNA of primary breast cancer patients. Clin Cancer Res. Mar. 2000;6(3):1119-24.
Shendure, J. et al. "Next-generation DNA sequencing," Nat. Biotechnol. 2008 26(10), 1135-1145.
Shinozaki, et al. Utility of circulating B-RAF DNA mutation in serum for monitoring melanoma patients receiving biochemotherapy. Clin Cancer Res. Apr. 1, 2007;13(7):2068-74.
Shiroguchi, et al. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci U S A. Jan. 24, 2012;109(4):1347-52. doi: 10.1073/pnas.1118018109. Epub Jan. 9, 2012.
Shoemaker, et al. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. Nature Genetics, 14: 450-456 (1996).
Simpson, et al. Copy number variant detection in inbred strains from short read sequence data. Bioinformatics. Feb. 15, 2010;26(4):565-7. doi: 10.1093/bioinformatics/btp693. Epub Dec. 18, 2009.
Smith, et al. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Research, 38(13): e142 (2010).
Smith, T.F. et al. "Comparison of Biosequences" Adv App Math (1981) 2:482-489.
Sorenson, et al. Soluble normal and mutated DNA sequences from single-copy genes in human blood. Cancer Epidemiol Biomarkers Prev. Jan.-Feb. 1994;3(1):67-71.
Sparks, et al. Selective analysis of cell-free DNA in maternal blood for evaluation of fetal trisomy. Prenat Diagn. Jan. 2012;32(1):3-9. doi: 10.1002/pd.2922. Epub Jan. 6, 2012.
Stein, et al. "The case for cloud computing in genome informatics", Genome Biol. 2010; 11 (5):207. Epub May 5, 2010.
Steinman. Free DNA in serum and plasma from normal adults. J Clin Invest. Aug. 1975;56(2):512-5.
Stitziel, et al. Computational and statistical approaches to analyzing variants identified byexome sequencing. Genome Biol. Sep. 14, 2011;12(9):227. doi: 10.1186/gb-2011-12-9-227.
Stroun, M, et al, "About the possible orgin and mechanism of circulating DNA apoptosis and active DNA release", Clin Chim Acta, vol. 313, No. 1-2, pp. 139-142, (2001).
Stumm, M. et al. "Noninvasive prenatal detection of chromosomal aneuploidies using different next generation sequencing strategies and algorithms" Prenatal Diagnosis (2012) 32:569-577.
Taback, et al. Detection of tumor-specific genetic alterations in bone marrow from early-stage breast cancer patients. Cancer Res. Apr. 15, 2003;63(8):1884-7.
Tan, et al. Genome-wide comparison of DNA hydroxymethylation in mouse embryonic stem cells and neural progenitor cells by a new comparative hMeDIP-seq method. Nucleic Acids Res. Apr. 2013;41(7):e84. doi: 10.1093/nar/gkt091. Epub Feb. 13, 2013.
Taudien, et al. Haplotyping and copy number estimation of the highly polymorphic human beta-defensin locus on 8p23 by 454 amplicon sequencing. BMC Genomics. Apr. 19, 2010;11:252. doi: 10.1186/1471-2164-11-252.
Teer, J.K. et al. "Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing" Genome Res (2010) 20(10):1420-1431.
Tie, J. et al., "Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer," Sci. Transl. Med. 2016, 8(346):346ra92.
Tomaz, et al. Differential methylation as a cause of allele dropout at the imprinted GNAS locus. Genet Test Mol Biomarkers. Aug. 2010;14(4):455-60. doi: 10.1089/gtmb.2010.0029.
Tomlinson, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet. Aug. 2007;39(8):984-8. Epub Jul. 8, 2007.
Tsai, et al. Discovery of rare mutations in populations: TILLING by sequencing. Plant Physiol. Jul. 2011;156(3):1257-68. doi: 10.1104/pp. 110.169748. Epub Apr. 29, 2011.
U.S. Appl. No. 14/102,285, filed Dec. 10, 2013.
U.S. Appl. No. 61/613,413 ("Schmitt '413 provisional") (Mar. 20, 2012).
U.S. Appl. No. 61/625,319 ("Schmitt '319 provisional") (Apr. 17, 2012).
Daser, et al. "Interrogation of genomes by molecular copy-number counting (MCC)." Nature Methods, 3(6): 447-453 (2006).
De Saizieu, et al. "Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays." Nature Biotechnology, 16: 45-48 (1998).
Diehl et al. Circulating mutant DNA to assess tumor dynamics. Nat Med 14(9):985-990 (2008).
Diehl, et al. Detection and quantification of mutations in the plasma of patients with colorectal tumors. Proc Natl Acad Sci US A. Nov. 8, 2005;102(45):16368-73. Epub Oct. 28, 2005.
Diehl, F. et al., "Analysis of Mutations in DNA Isolated From Plasma and Stool of Colorectal Cancer Patients," Gastroenterology (2008) 135(2):489-498.
Ding, L. et al. "Clonal evolution in relapsed acute myeloid leukemia revealed by whole-genome sequencing" Nature (2012) 481(7382):506-510.
Ding, L. et al., "Analysis of next-generation genomic data in cancer: accomplishments and challenges," Hum. Mol. Genet. 2010 19(R2), R188-R196.
Duncan, D.L. et al. "Next-Generation Sequencing in the Clinical Laboratory" Diagnostic Molecular Pathology: A Guide to Applied Molecular Testing 25-33 (Coleman and Tsongalis eds., 1st ed. 2016).
Ehrich, et al. Noninvasive detection of fetal trisomy 21 by sequencing of DNA in maternal blood: a study in a clinical setting. Am J Obstet Gynecol. Mar. 2011;204(3):205.e1-11. doi: 10.1016/j.ajog.2010.12.060. Epub Feb. 18, 2011.
Eisenmann, et al. 5q-myelodysplastic syndromes: chromosome 5q genes direct a tumor-suppression network sensing actin dynamics. Oncogene. Oct. 1, 2009;28(39):3429-41. doi: 10.1038/onc.2009.207. Epub Jul. 13, 2009.
Elshire, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. May 4, 2011;6(5):e19379. doi: 10.1371/journal.pone.0019379.
European search report and search opinion dated Mar. 29, 2016 for EP Application No. 13834427.0.
Extended European search report and opinion dated May 4, 2017 for EP Application No. 14874157.2.
Extended European search report and opinion dated Nov. 21, 2016 for EP Application No. 14771159.
Fan, et al. "Microfluidic digital PCR enables rapid prenatal diagnosis of fetal aneuploidy." Am Obstet Gynecol. 2009;200:543.e1-543.e7.
Fan, et al. "Parallel Genotyping of Human SNPs Using Generic High-density Oligonucleotide Tag Arrays." Genome Research, 10: 853-860 (2000).
Fan, et al. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci U S A. Oct. 21, 2008;105(42):16266-71. Epub Oct. 6, 2008.
Fan, et al. Non-invasive prenatal measurement of the fetal genome. Nature. Jul. 19, 2012;487(7407):320-4. doi: 10.1038/nature 11251.
Fleischhacker, M. et al. "Circulating nucleic acids (CNAs) and cancer—A survey" Biochimica et Biophysica Acta (2007) 1775:181-232.

(56) **References Cited**

OTHER PUBLICATIONS

Fonatsch, C. The role of chromosome 21 in hematology and oncology. Genes Chromosomes Cancer. Jun. 2010;49(6):497-508. doi: 10.1002/gcc.20764.

Fong, S.L. et al. "Comparison of 7 Methods for Extracting Cell-Free DNA from Serum Samples of Colorectal Cancer Patients" Clinical Chem (2009) 55(3):587-598.

Forshew, T. et al. "Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA" Sci Transl Med (2012) 4(136) ra68.

Forster, et al. From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software. Nucleic Acids Res. Jan. 7, 2013;41(1):e16. doi: 10.1093/nar/gks836. Epub Sep. 10, 2012.

Fournie, et al. Plasma DNA as a marker of cancerous cell death. Investigations in patients suffering from lung cancer and in nude mice bearing human tumours. Cancer Lett. May 8, 1995;91(2):221-7.

Freeman, et al. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. Oct. 2009;19(10):1817-24. doi: 10.1101/gr.092924.109. Epub Jun. 18, 2009.

Fu, et al. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc Natl Acad Sci U S A. May 31, 2011;108(22):9026-31. Epub May 11, 2011.

Fujiwara, et al. Identification of epigenetic aberrant promoter methylation in serum DNA is useful for early detection of lung cancer Clin Cancer Res. Feb. 1, 2005;11(3):1219-25.

Fujiwara, et al. Plasma DNA microsatellites as tumor-specific markers and indicators of tumor progression in melanoma patients. Cancer Res. Apr. 1, 1999;59(7):1567-71.

Genome Analyzer IIx Systems Specification Sheet (2009).

Gerry, et al. Universal DNA microarray method for multiplex detection of low abundance point mutations. Journal of Molecular Biology, 292(2): 251-262 (1999).

Gillespie. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry, 81(25): 2340-2361 (1977).

Gordian, et al. Serum free circulating DNA is a useful biomarker to distinguish benign versus malignant prostate disease. Cancer Epidemiol Biomarkers Prev. Aug. 2010;19(8):1984-91. doi: 10.1158/1055-9965.EPI-10-0287. Epub Jul. 20, 2010.

Gordon, D.J. et al., "Causes and consequences of aneuploidy in cancer," Nat. Rev. Genet. 2012 13(3), 189-203.

Gormally, et al. Amount of DNA in plasma and cancer risk: a prospective study. Int J Cancer. Sep. 20, 2004;111(5):746-9.

Grant, et al. SNP genotyping on a genome-wide amplified DOP-PCR template. Nucleic Acids Res. Nov. 15, 2002;30(22):e25.

Gregory et al. "Targeted Single Molecule Mutation Detection with Massively Parallel Sequencing" Nucleic Acids Research (2015) 44(3):2-11.

Grutzmann, et al. Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. PLoS One. 2008;3(11):e3759. doi: 10.1371/journal.pone.0003759. Epub Nov. 19, 2008.

Guardant Health vs. *FMI* 1st Amended Answer to Second Amended Complaint dated May 6, 2019 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Contentions dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit A-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit A-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit A-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit A-8, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit B-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit B-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit B-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit B-7, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit C-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit C-11, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit C-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

"Blood Plasma" Oxford Dictionary of Biochemistry and Molecular Biology 81 (2d ed. 2006).

"LB-246/12—Detection of circulating tumor DNA in early stage cancers," Available at https://www.genomeweb.com/cancer/ngs-error-correction-method-described-aacr-excitement-grows-early-cancer-detection, Accessed on Apr. 26, 2017.

"Cohesive End," Oxford Dictionary of Biochemistry and Molecular Biology 132 (2d ed. 2006).

"Control," Oxford Dictionary of Biochemistry and Molecular Biology (2d ed. 2006), 143.

Ajay, S.S. et al., "Accurate and comprehensive sequencing of personal genomes," Genome Res. 2011, 21(9), 1498-1505.

Albert, T.J. et al., "Direct selection of human genomic loci by microarray hybridization," Nat. Methods 2007, 4(11), 903-905.

Alkan, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. Oct. 2009;41(10):1061-7. doi: 10.1038/ng.437. Epub Aug. 30, 2009.

Andersson, et al. Bayesian detection of periodic mRNA time profiles without use of training examples. BMC Bioinformatics. Feb. 9, 2006;7:63.

Angeloni, D. Molecular analysis of deletions in human chromosome 3p21 and the role of resident cancer genes in disease. Brief Funct Genomic Proteomic. Mar. 2007;6(1):19-39. Epub May 24, 2007.

Ansorge "Next-generation DNA sequencing techniques." New Biotechnology, 25(4):195-203 (2009).

Appel, Maryke. Part II: It's all about conversion. Kapa Biosystems. Accessed Mar. 29, 2017. Printed Apr. 11, 2017. 5 pages. URL:https://www.kapabiosystems.com/ngs/part-ii-conversion/.

Arneson, N. et al., "Whole-Genome Amplification by Adaptor-Ligation PCR of Randomly Sheared Genomic DNA (PRSG)," CSH Protocols, 2008, 3(1).

Ashford, Monica. NGS Error Correction Method Described at AACR as Excitement Grows for Early Cancer Detection. Genomeweb.com. Apr. 5, 2017. 3 pages. URL:https://www.genomeweb.com/cancer/ngs-error-correction-method-described-aacr-excitement-grows-early-cancer-detection.

Atanur, et al. The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. Genome Res. Jun. 2010;20(6):791-803. doi: 10.1101/gr.103499.109. Epub Apr. 29, 2010.

Audic, et al. "The Significance of Digital Gene Expression Profiles." Genome Research, 7: 986-995 (1997).

Barzon, et al. Evaluation of circulating thyroid-specific transcripts as markers of thyroid cancer relapse. Int J Cancer. Jul. 20, 2004;110(6):914-20.

Bendich, et al. Circulating DNA as a possible factor in oncogenesis. Science. Apr. 16, 1965;148(3668):374-6.

Bentley, D.R. et al., "Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry," Nature, 2008, 456(7218):53-59.

Bettegowda et al., "Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies," Sci. Transl. Med. 2014, 6(224):224ra24.

Bianchi, D.W. et al. "Genome-Wide Fetal Aneuploidy Detection by Maternal Plasma DNA Sequencing," Obstetrics & Gynecology, (2012) 119(5), 1-12.

Bonaldo, et al. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. Sep. 1996;6(9):791-806.

Bowcock, et al. Exclusion of the retinoblastoma gene and chromosome 13q as the site of a primary lesion for human breast cancer. Am J Hum Genet. Jan. 1990;46(1):12-7.

Braha, et al. Simultaneous stochastic sensing of divalent metal ions. Nature Biotechnology, 18: 1005-1007 (2000).

(56) **References Cited**

OTHER PUBLICATIONS

Bremnes, et al. Circulating tumour-derived DNA and RNA markers in blood: a tool for early detection, diagnostics, and follow-up? Lung Cancer. Jul. 2005;49(1):1-12.

Brenner, et al. "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays." Nature Biotechnology, 18: 630-634 (2000).

Brenner, et al. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. Proc Natl Acad Sci U S A. Feb. 15, 2000;97(4):1665-70.

Bryzgunova,O. et al., "A reliable method to concentrate circulating DNA," Analytical Biochem., 2010, 408:354-356.

Campbell, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. Jun. 2008;40(6):722-9. doi: 10.1038/ng.128. Epub Apr. 27, 2008.

Caramazza, et al. Chromosome 1 abnormalities in myeloid malignancies: a literature survey and karyotype-phenotype associations. Eur J Haematol. Mar. 2010;84(3):191-200. doi: 10.1111/j.1600-0609.2009.01392.x. Epub Nov. 30, 2009.

Carr, et al. Inferring relative proportions of DNA variants from sequencing electropherograms. Bioinformatics. Dec. 15, 2009;25(24):3244-50. doi: 10.1093/bioinformatics/btp583. Epub Oct. 9, 2009.

Casava v1.8.2 User Guide http://gensoft.pasteur.fr/docs/casava/1.8.2/CASAVA_1_8_2_UG_15011196C.pdf (Oct. 2011).

Casbon, et al. A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res. Jul. 2011;39(12):e81. doi: 10.1093/nar/gkr217. Epub Apr. 13, 2011.

Castle, et al. DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing. BMC Genomics. Apr. 16, 2010;11:244. doi: 10.1186/1471-2164-11-244.

Chang, et al. Detection of allelic imbalance in ascitic supernatant by digital single nucleotide polymorphism analysis. Clin Cancer Res. Aug. 2002;8(8):2580-5.

Chee, et al. "Accessing genetic information with high-density DNA arrays." Science, 274: 610-614 (1996).

Chee. "Enzymatic multiplex DNA sequencing." Nucleic Acids Research, 19(12): 3301-3305 (1991).

Chen, et al. Microsatellite alterations in plasma DNA of small cell lung cancer patients. Nat Med. Sep. 1996;2(9):1033-5.

Chin, et al. A SNP in a let-7 microRNA complementary site in the KRAS 3' untranslated region increases non-small cell lung cancer risk. Cancer Res. Oct. 15, 2008;68(20):8535-40. doi: 10.1158/0008-5472.CAN-08-2129.

Chiu, et al. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. BMJ. Jan. 11, 2011;342:c7401. doi: 10.1136/bmj.c7401.

Chiu, et al. Quantitative analysis of circulating mitochondrial DNA in plasma. Clin Chem. May 2003;49(5):719-26.

Chiu, R.W.K. et al. "Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma" PNAS (2008) 105(51):20458-20463.

Church, et al. "Multiplex DNA sequencing." Science, 240: 185-188 (1988).

Cook, et al. Methylated DNA labels for marking objects. Biotechnol Lett. Jan. 2003;25(1):89-94.

Costello, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. Apr. 1, 2013;41(6):e67. doi: 10.1093/nar/gks1443. Epub Jan. 8, 2013.

Coulet, et al. Detection of plasma tumor DNA in head and neck squamous cell carcinoma by microsatellite typing and p53 mutation analysis. Cancer Res. Feb. 1, 2000;60(3):707-11.

Cox, J. "Bar coding objects with DNA." Analyst. May 2001;126(5):545-7.

Cox. Bar coding objects with DNA. Analyst. May 2001;126(5):545-7.

Craig, et al. Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods. Oct. 2008;5(10):887-93. doi: 10.1038/nmeth.1251. Epub Sep. 14, 2008.

Daines, et al. High-throughput multiplex sequencing to discover copy number variants in *Drosophila*. Genetics. Aug. 2009;182(4):935-41. doi: 10.1534/genetics.109.103218. Epub Jun. 15, 2009.

D'Antoni, et al. Rapid quantitative analysis using a single molecule counting approach. Anal Biochem. May 1, 2006;352(1):97-109. Epub Feb. 10, 2006.

Ingolia, et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. Apr. 10, 2009;324(5924):218-23. Epub Feb. 12, 2009.

Instructions for Norit Rapid DNA Ligation Kit (Nov. 6, 2004).

International search report and written opinion dated Apr. 3, 2015 for PCT/US2014/072383.

International search report and written opinion dated May 7, 2012 for PCT/IB2011/003160.

International search report and written opinion dated Jun. 6, 2012 for PCT/US2011/065291.

International search report and written opinion dated Jul. 15, 2015 for PCT/US2015/030639.

International search report and written opinion dated Sep. 5, 2014 for PCT/US2014/000048.

International search report and written opinion dated Sep. 6, 2017 for PCT Application No. US-201727809.

International search report and written opinion dated Nov. 18, 2013 for PCT/US2013/058061.

Invitrogen Instructions for T4 DNA Ligase (May 5, 2002).

IPR2018-00130, Petition for Inter Partes Review of U.S. Pat. No. 9,598,731, dated Nov. 7, 2018.

IPR2019-00130, Decision Denying Inter Partes Review U.S. Pat. No. 9,598,731, issued Jun. 5, 2019.

IPR2019-00130, Preliminary Response to Petition for Inter Partes Review U.S. Pat. No. 9,598,731, filed Mar. 6, 2019.

IPR2019-00634, Decision Instituting Inter Partes Review U.S. Pat. No. 9,840,743, issued Aug. 19, 2019.

IPR2019-00634, Petition for Inter Partes Review of U.S. Pat. No. 9,840,743, dated Feb. 1, 2019.

IPR2019-00636 & 637, Exhibit 3001 Memo for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Jul. 1, 2019.

IPR2019-00636 & 637, Order—Conduct of the Proceedings for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Jul. 22, 2019.

IPR2019-00636 and IPR2019-00637, Decision Denying Inter Partes Review U.S. Pat. No. 9,902,992, issued Aug. 20, 2019.

IPR2019-00636, Petition for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Feb. 1, 2019.

IPR2019-00637, Petition for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Feb. 1, 2019.

IPR2019-00652, Decision Granting Inter Partes Review U.S. Pat. No. 9,834,822, issued Aug. 19, 2019.

IPR2019-00652, Petition for Inter Partes Review of U.S. Pat. No. 9,834,822, dated Feb. 1, 2019.

IPR2019-00653, Decision Denying Inter Partes Review U.S. Pat. No. 9,834,822, issued Aug. 19, 2019.

IPR2019-00653, Petition for Inter Partes Review of U.S. Pat. No. 9,834,822, dated Feb. 1, 2019.

Jabara, C. Capturing the cloud: High throughput sequencing of multiple individual genomes from a retroviral population. Biology Lunch Bunch Series, Training Initiatives in Biomedical & Biological Sciences of the University of North Carolina at Chapel Hill, Apr. 23, 2010.

Jabara, et al. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. (Paper # 665), The 18th Annual Conference on Retroviruses and Opportunistic Infections, Boston, Massachusetts, Mar. 2011.

Jabara, et al. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl Acad Sci U S A. Dec. 13, 2011;108(50):20166-71. doi: 10.1073/pnas.1110064108. Epub Nov. 30, 2011.

Jahr, et al. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. Cancer Res. Feb. 15, 2001;61(4):1659-65.

(56) **References Cited**

OTHER PUBLICATIONS

Jeronimo, et al. Quantitative GSTP1 hypermethylation in bodily fluids of patients with prostate cancer. Urology. Dec. 2002;60(6):1131-5.

Jiang, H. et al., "SeqMap: mapping massive amount of oligonucleotides to the genome," Bioinformatics 2008, 24(20):2395-2396.

Kanagawa. Bias and artifacts in multitemplate polymerase chain reactions (PCR), 2003, Journal of Bioscience and Bioengineering, vol. 96, No. 4, p. 317-323.

Kao,W-C. et al., "BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing," Genome Res. 2009 19(10), 1884-1895.

Karow, J. "Hopkins Team Develops Method to Improve Rare Mutation Detection for Early Cancer Dx" Genome Web (2011) 3 pages.

Kennedy, S.R. et al., "Detecting ultralow-frequency mutations by Duplex Sequencing" Nature Protocols (2014) 9(11):2586-2606.

Kennedy, S.R. et al., "Ultra-Sensitive Sequencing Reveals an Age-Related Increase in Somatic Mitochondrial Mutations That Are Inconsistent with Oxidative Damage" PLOS Genetics (2013) 9:e1003794.

Kennedy, S.R. et al., "Detecting ultralow-frequency mutations by Duplex Sequencing," Nat. Protoc. 2014, 9(11), 2586-2606.

Kimura, et al. EGFR mutation status in tumour-derived DNA from pleural effusion fluid is a practical basis for predicting the response to gefitinib. Br J Cancer Nov. 20, 2006;95(10):1390-5. Epub Oct. 24, 2006.

Kinde, et al. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. Jun. 7, 2011;108(23):9530-5. doi: 10.1073/pnas.1105422108. Epub May 17, 2011.

Kinde, et al. Supplemental Information, Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. Jun. 7, 2011;108(23):1-10.

Kinde, I. et al., "Evaluation of DNA from the Papanicolaou Test to Detect Ovarian and Endometrial Cancers," Sci. Transl. Med. 2013, 5(167):167ra4.

Kinde, I. et al., "Fast-SeqS: a simple and efficient method for the detection of aneuploidy by massively parallel sequencing," PLoS One 2012, 7(7), e41162.

Kircher, M. et al., "Improved base calling for the Illumina Genome Analyzer using machine learning strategies," Genome Biol. 2009, 10(8), R83.

Kirsch, S. et al. "Sequence error storms and the landscape of mutations in cancer," Proc. Natl. Acad. Sci. USA 2012, 109(36), 14289-14290.

Kivioja, T., et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. Nov. 20, 2011;9(1):72-4. doi: 10.1038/nmeth.1778.

Koboldt, D.C.et al., "The next-generation sequencing revolution and its impact on genomics," Cell 2013, 155(1), 27-38.

Koboldt, et al. Massively parallel sequencing approaches for characterization of structural variation. Aug. 12, 2011. Methods Mol Biol. 2012;838:369-84. doi: 10.1007/978-1-61779-507-7_18.

Kolble, et al. Microsatellite alterations in serum DNA of patients with colorectal cancer. Lab Invest. Sep. 1999;79(9):1145-50.

Konig, et al. iCLIP reveals the function of hnRNAP particles in splicing at individual nucleotide resolution, Jul. 2010, Nature Structural & Molecular Biology, 17(7):909-916.

Kopreski, et al. Detection of mutant K-ras DNA in plasma or serum of patients with colorectal cancer. Br J Cancer. 1997;76(10):1293-9.

Korbel, J.O. et al., "Paired-end mapping reveals extensive structural variation in the human genome," Science 2007, 318(5849), 420-426.

Clark, T.A. et al. "Analytical Validation of a Hybrid Capture Based Next-Generation Sequencing Clinical Assay for Genomic Profiling of Cell-Free Circulating Tumor DNA," J. Mol. Diagnostics (2018) 20(5):686-702.

Paweletz, C.P. et al. "Bias-corrected targeted next-generation sequencing for rapid, multiplexed detection of actionable alterations in cell-free DNA from advanced lung cancer patients" Clin Canc Res (2016) 22(4):915-922.

Phallen, J. et al. "Direct detection of early-stage cancers using circulating tumor DNA" Sci Trans Med (2017) vol. 9, Issue 403, eaan2415DOI: 10.1126/scitranslmed.aan2415.

Shiroguchi, et al. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci U S A. 2012 Supplemental Information (8 pages).

Kamps-Hughes, N. et al. "ERASE-Seq: Leveraging replicate measurements to enhance ultralow frequency variant detection in NGS data" PLOS One (2018).

Lennon, N.J. et al. "Technological considerations for genome-guided diagnosis and management of cancer" Gen Med (2016) 8:112.

Mishra, S. et al. "Different Facets of Copy Number Changes: Permanent, Transient, and Adaptive" Mol Cell Biol (2016) 36(7):1050-1063.

Moensch, S. Genomic Profiling Using Guardant 360 Cell-Free DNA-Based Assay vs Tumor-Based Genotyping Assays in Advanced NSCLC, Cancer Therapy Advisor (Feb. 28, 2019), https://www.cancertherapyadvisor.com/home/news/conferencecoverage/american-association-for-cancer-research-aacr/aacr-2019/genomic-profiling-using-guardant-360-cell-free-dna-based assay-vs-tumor-based-genotyping-assays-in-advanced-nsclc/ (lastaccessed Nov. 30, 2019).

Newman, A. et al. "Integrated digital error suppression for improved detection of circulating tumor DNA" Nature Biotech (2016) 34(5):547-555.

Odegaard, J.I. et al. "Validation of a Plasma-Based Comprehensive Cancer Genotyping Assay Utilizing Orthogonal Tissue- and Plasma-Based Methodologies" Clin Canc Res (2018) 24(15):3539-3549.

Ou, Shi et al. "Liquid Biopsy to Identify Actionable Genomic Alterations" Am Soc Clin Onc (2018) 978.

Sathirapongsasuti, J.F. et al. "Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV" BioInformatics (2011) 27(19):2648-2654.

Trapnell, C. et al. "How to map billions of short reads onto genomes" Nature Biotech (2009) 27(5):455-457.

Van Loo, P. et al. "Allele-specific copy number analysis of tumors" PNAS (2010) 107(39):16910-16915.

Wang, T.T. et al. "High efficiency error suppression for accurate detection of low-frequency variants" NAR (2019) 47(15): e87.

EPO Preliminary Opinion, dated Aug. 25, 2020, in EP2893040.

Lo, Dennis Interview with Professor Dennis Lo, Qiagen News (2002).

Opponents Reply dated Aug. 20, 2020 to Proprietor's Observation in EP2893040.

Opposition Form and Statement to EP3470533 filed Aug. 6, 2020 by Foundation Medicine Inc.

U.S. Appl. No. 61/625,623, filed Apr. 17, 2012.

Guardant Health, Inc. Response to Notices of Opposition in EP2893040 filed May 29, 2020.

IPR2019-00634, Final Written Decision of U.S. Pat. No. 9,840,743, dated Aug. 18, 2020.

IPR2019-00652, Final Written Decision of U.S. Pat. No. 9,834,822, dated Aug. 18, 2020.

Jiang et al. "Basics in Bioinformatics Lecture Notes of the Graduate Summer School on Bioinformatics of China" Springer Heidelberg New York Dordrecht London (2013).

Opposition Form and Statement to EP3378952 filed Nov. 4, 2020 by Dirk Buhler.

Opposition Form and Statement to EP3378952 filed Nov. 5, 2020 by Foundation Medicine Inc.

Opposition Form and Statement to EP3378952 filed Nov. 5, 2020 by Grunecker.

Office Action dated Dec. 21, 2020 for U.S. Appl. No. 16/945,124.

Crowley, E. et al. "Liquid biopsy: monitoring cancer-genetics in the blood" Nat Rev Clin Oncology (2013) 8:472-478.

Extended European search report and opinion dated Dec. 7, 2020 for EP Application No. 20183626.9.

(56) **References Cited**

OTHER PUBLICATIONS

Greaves, L.C. et al. "Quantification of mitochondrial DNA mutation load" Aging Cell (2009) 8(5): 566-572.
Guardant Health, Inc. Response to Notice of Opposition in EP3378952 dated Mar. 29, 2021.
Guardant Health, Inc. v. *FMI* Defendant's Answer and Counter Claims, filed Jan. 14, 2021 (C.A. No. 20--1580(LPS)).
Guardant Health, Inc. v. *FMI* Defendant's Answering Brief in Opposition to Plaintiff Guardant Health, Inc.'sMotion for Preliminary Injunction, filed Mar. 5, 2021 (C.A. No. 20-1580 (LPS)).
Office action dated Jun. 23, 2021 for U.S. Appl. No. 16/913,965.
Opposition to EP3524694, dated Apr. 15, 2021 by Foundation Medicine, Inc.
Opposition to EP324694, dated Apr. 15, 2021 by Maiwlad International.

* cited by examiner

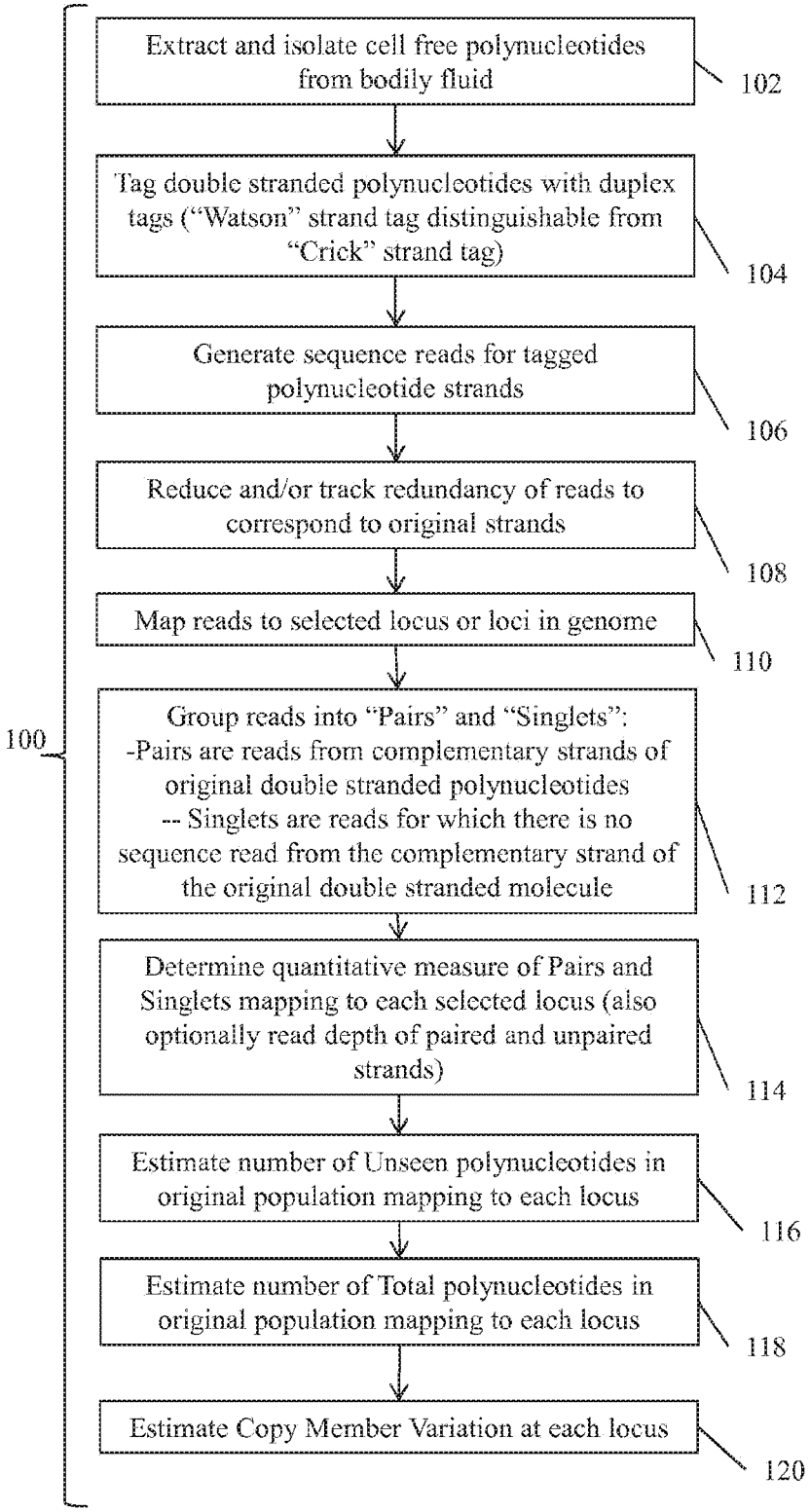

Extract and isolate cell free polynucleotides from bodily fluid
102
Tag double stranded polynucleotides with duplex tags ("Watson" strand tag distinguishable from "Crick" strand tag)
104
Generate sequence reads for tagged polynucleotide strands
106
Reduce and/or track redundancy of reads to correspond to original strands
108
Map reads to selected locus or loci in genome
110
100
Group reads into "Pairs" and "Singlets":
-Pairs are reads from complementary strands of original double stranded polynucleotides
-- Singlets are reads for which there is no sequence read from the complementary strand of the original double stranded molecule
112
Determine quantitative measure of Pairs and Singlets mapping to each selected locus (also optionally read depth of paired and unpaired strands)
114
Estimate number of Unseen polynucleotides in original population mapping to each locus
116
Estimate number of Total polynucleotides in original population mapping to each locus
118
Estimate Copy Member Variation at each locus
120

Fig. 1

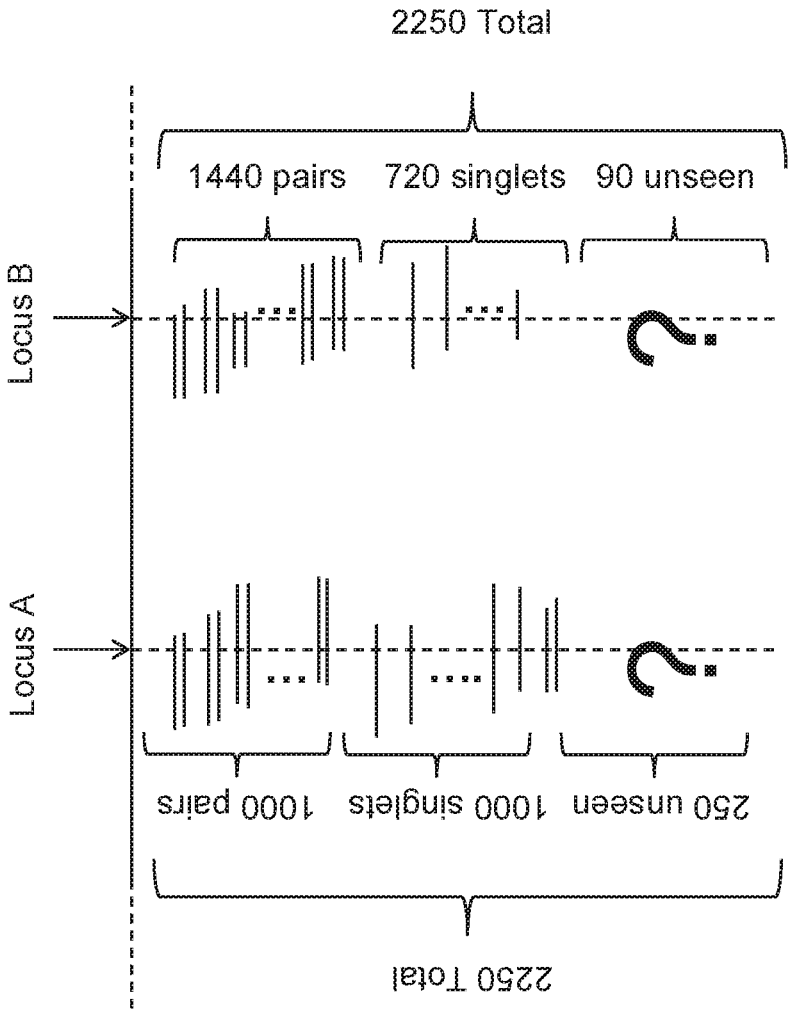

2250 Total
1440 pairs
720 singlets
90 unseen
Locus B
Locus A
1000 pairs
1000 singlets
250 unseen
2250 Total

Fig. 2

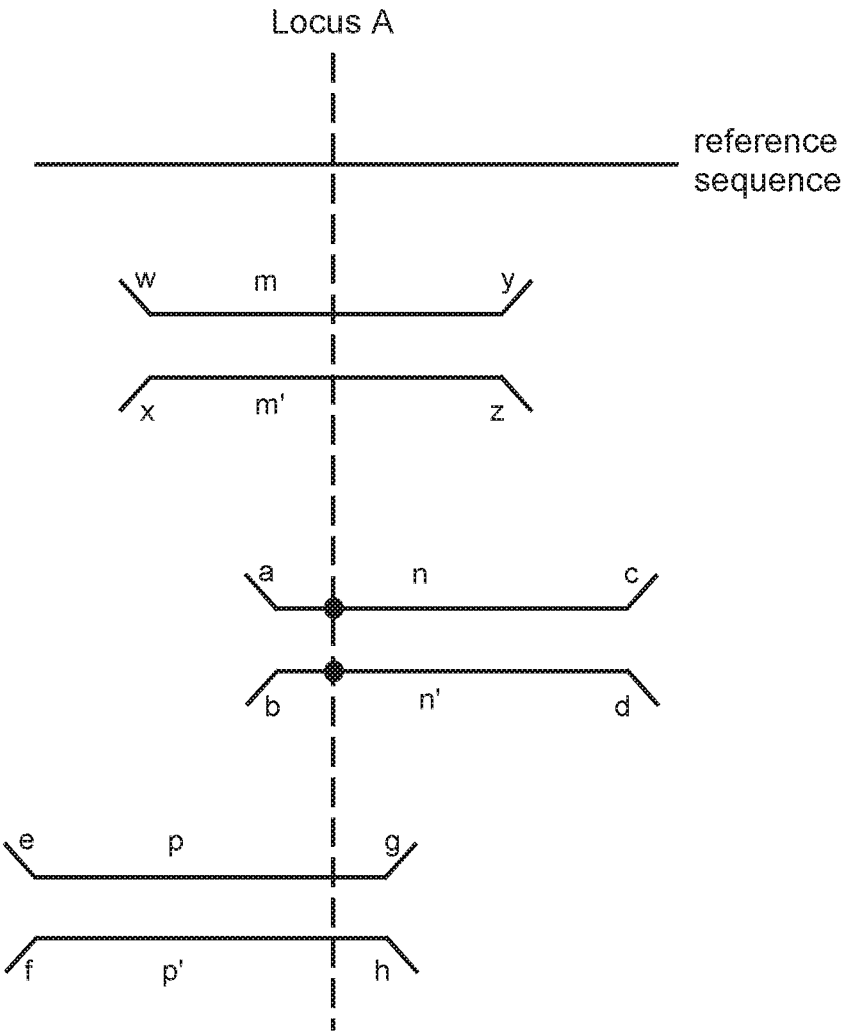


Fig. 3

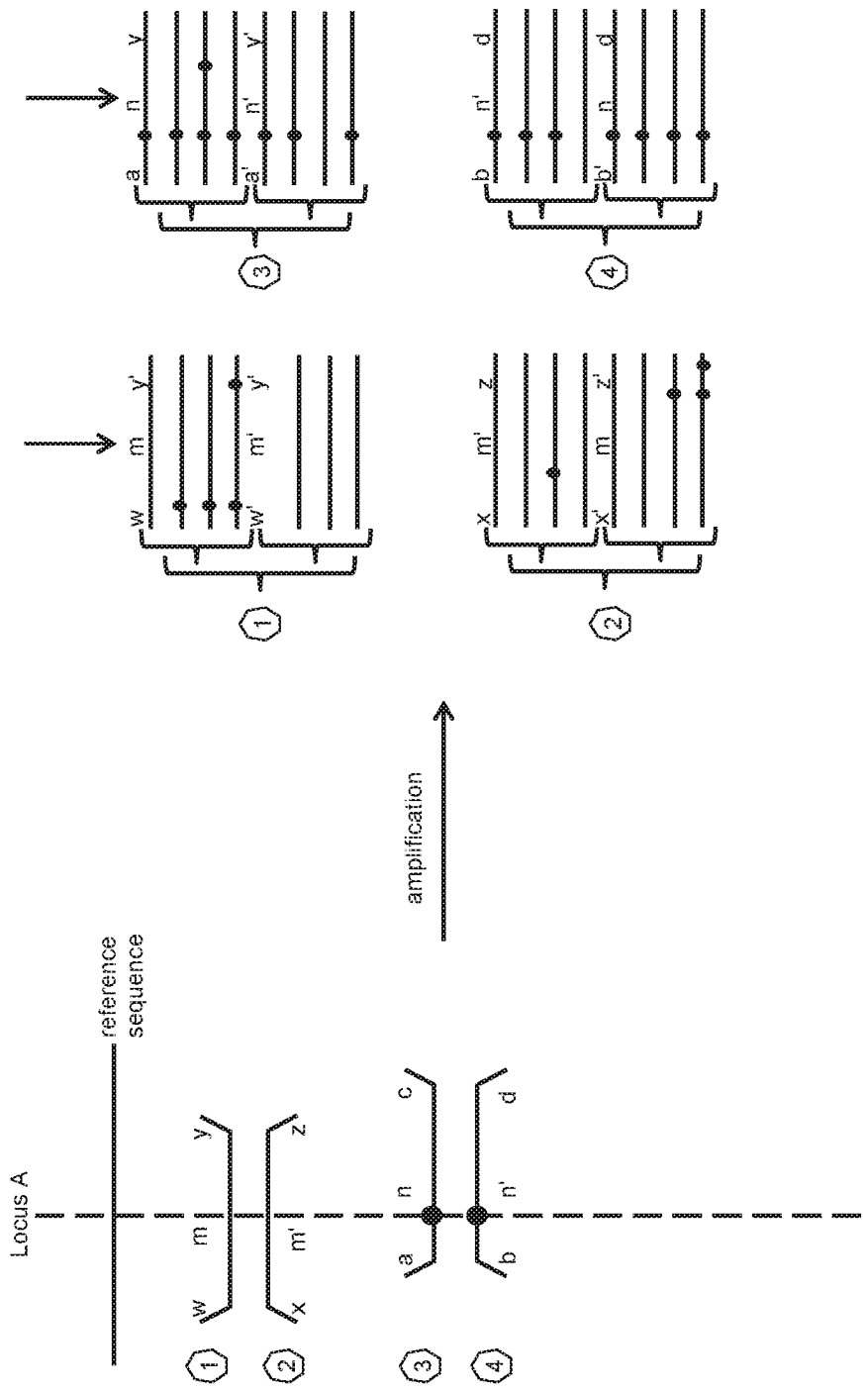

Locus A
reference sequence
amplification
w
m
y
x
m'
z
a
n
c
b
n'
d
1
2
3
4

Fig. 4A

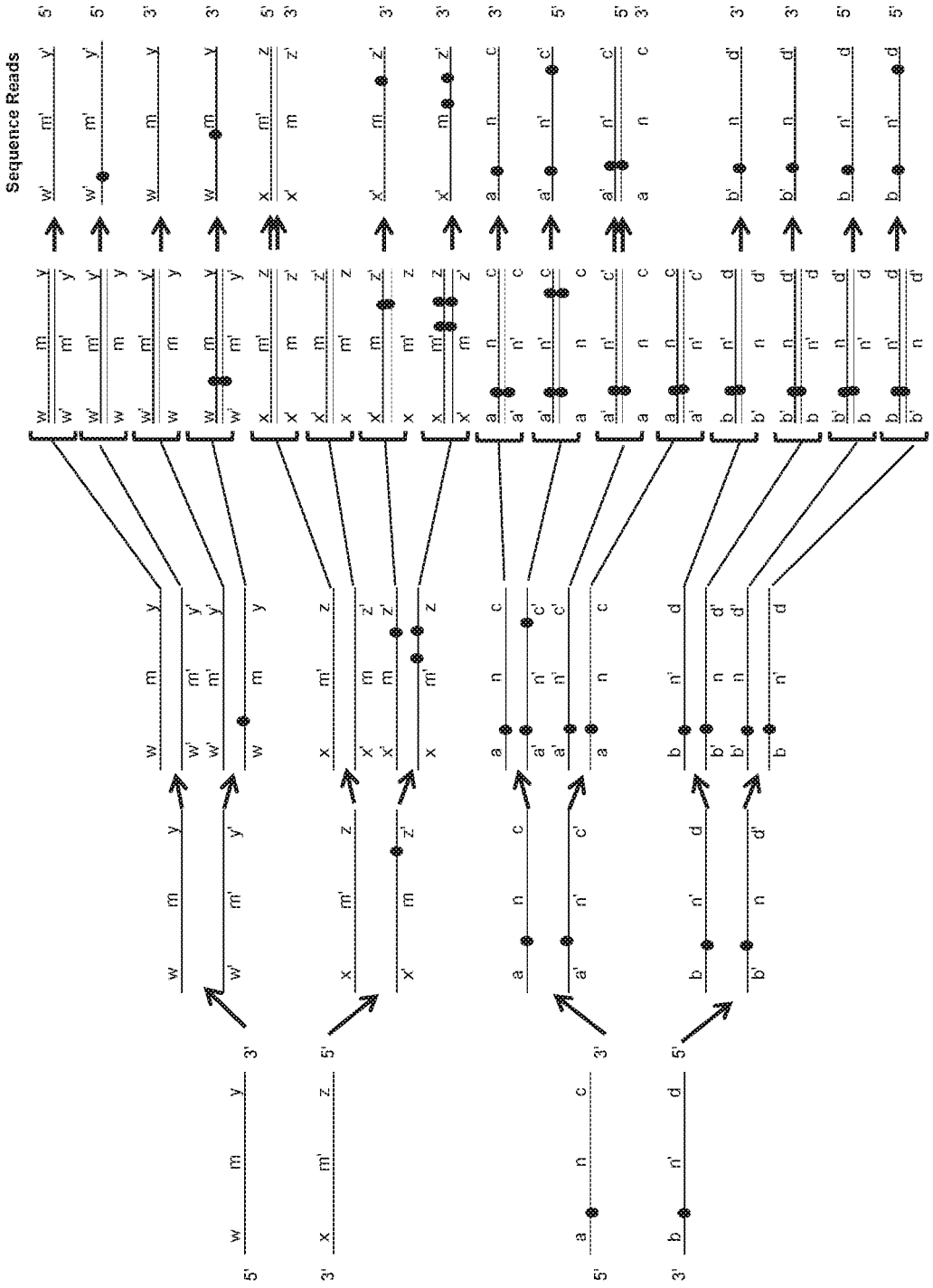


Fig. 4B

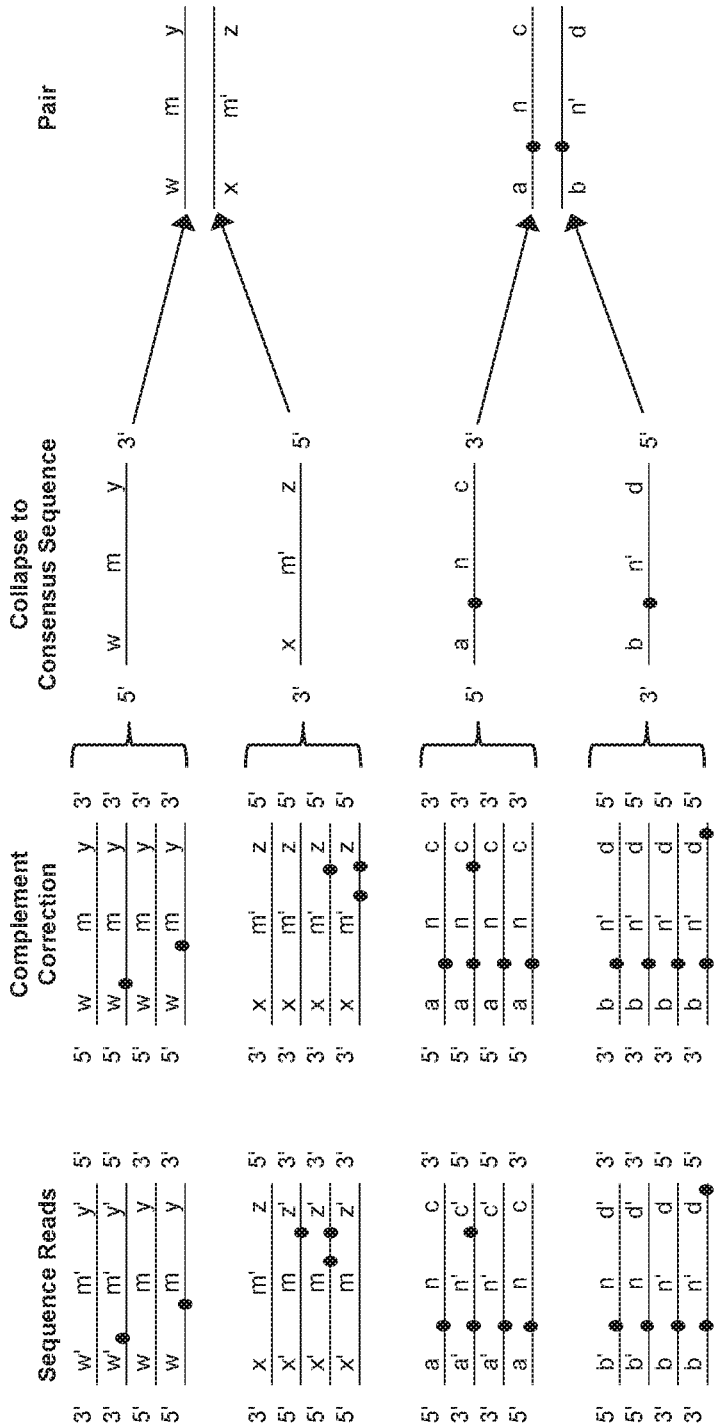

Sequence Reads
Complement Correction
Collapse to Consensus Sequence
Pair

Fig. 4C

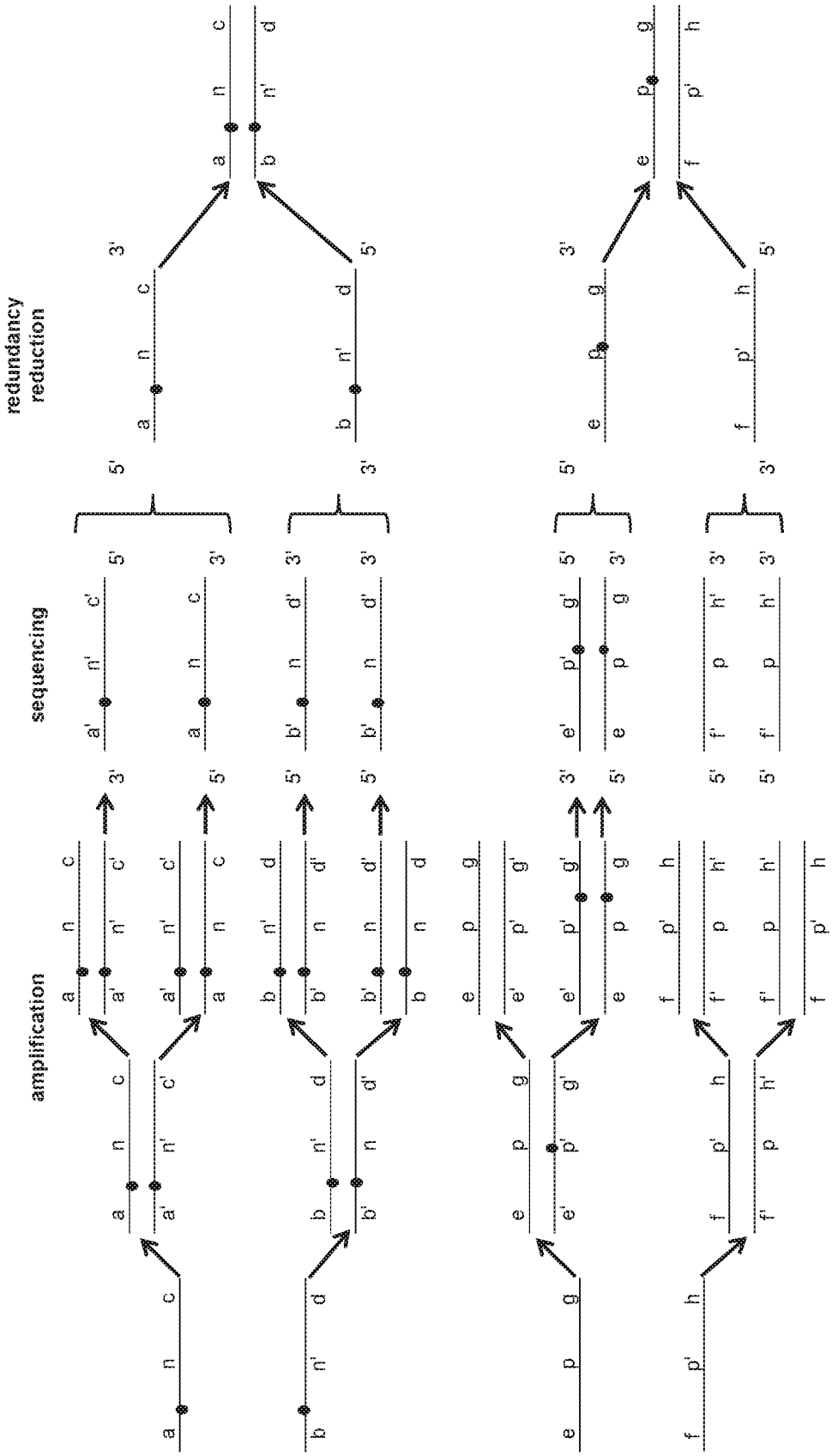


Fig. 5

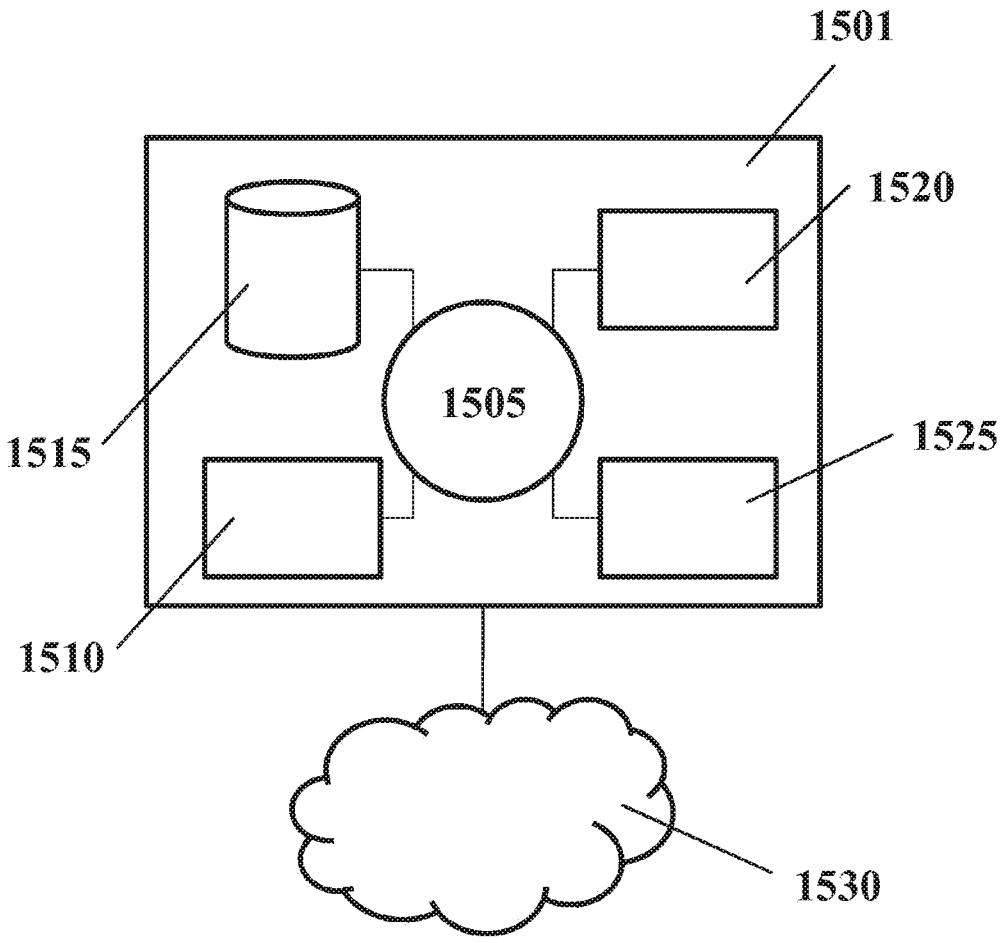


Fig. 6

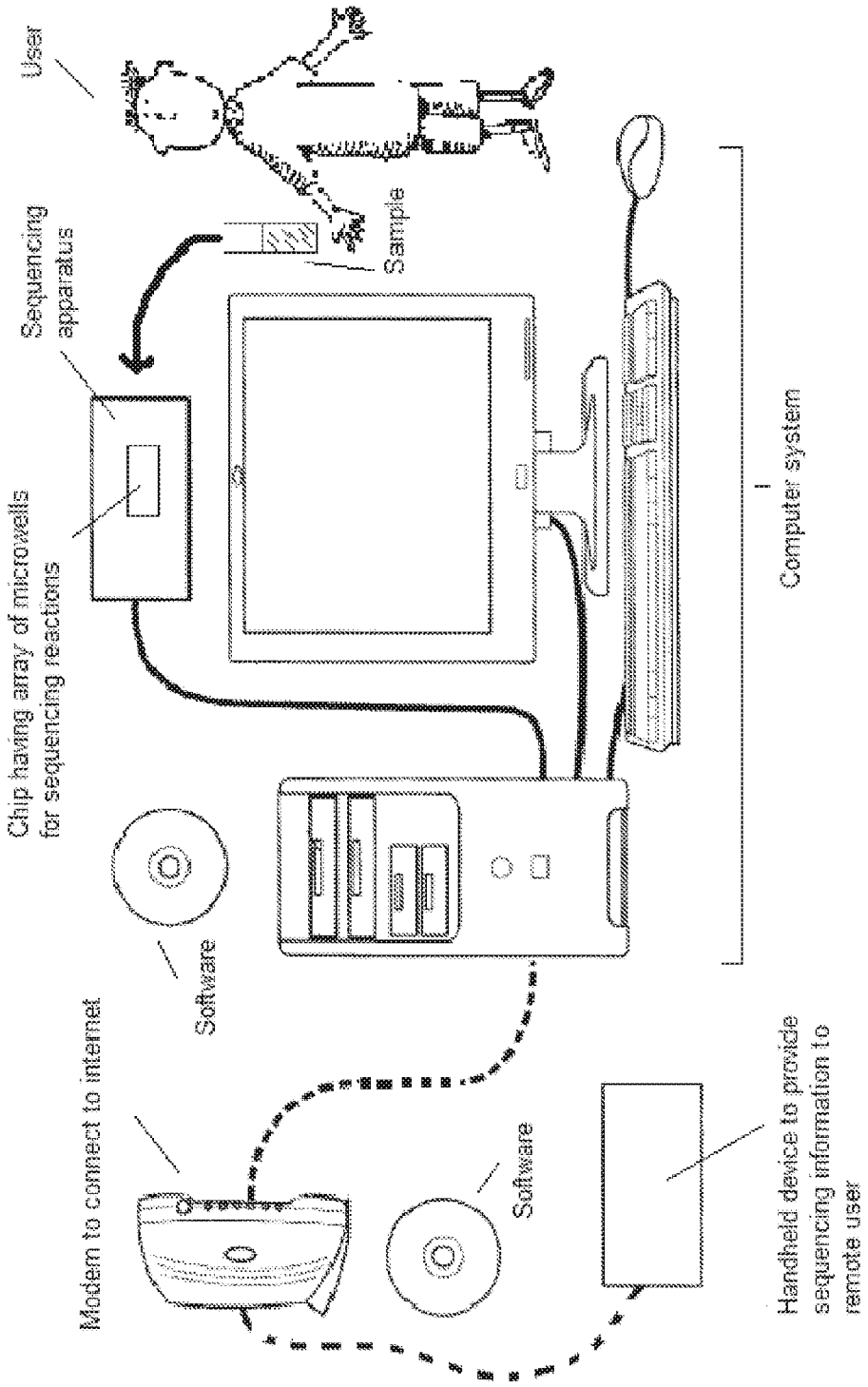

User
Sample
Sequencing apparatus
Chip having array of microwells for sequencing reactions
Computer system
Software
Modem to connect to internet
Software
Handheld device to provide sequencing information to remote user

Fig. 7

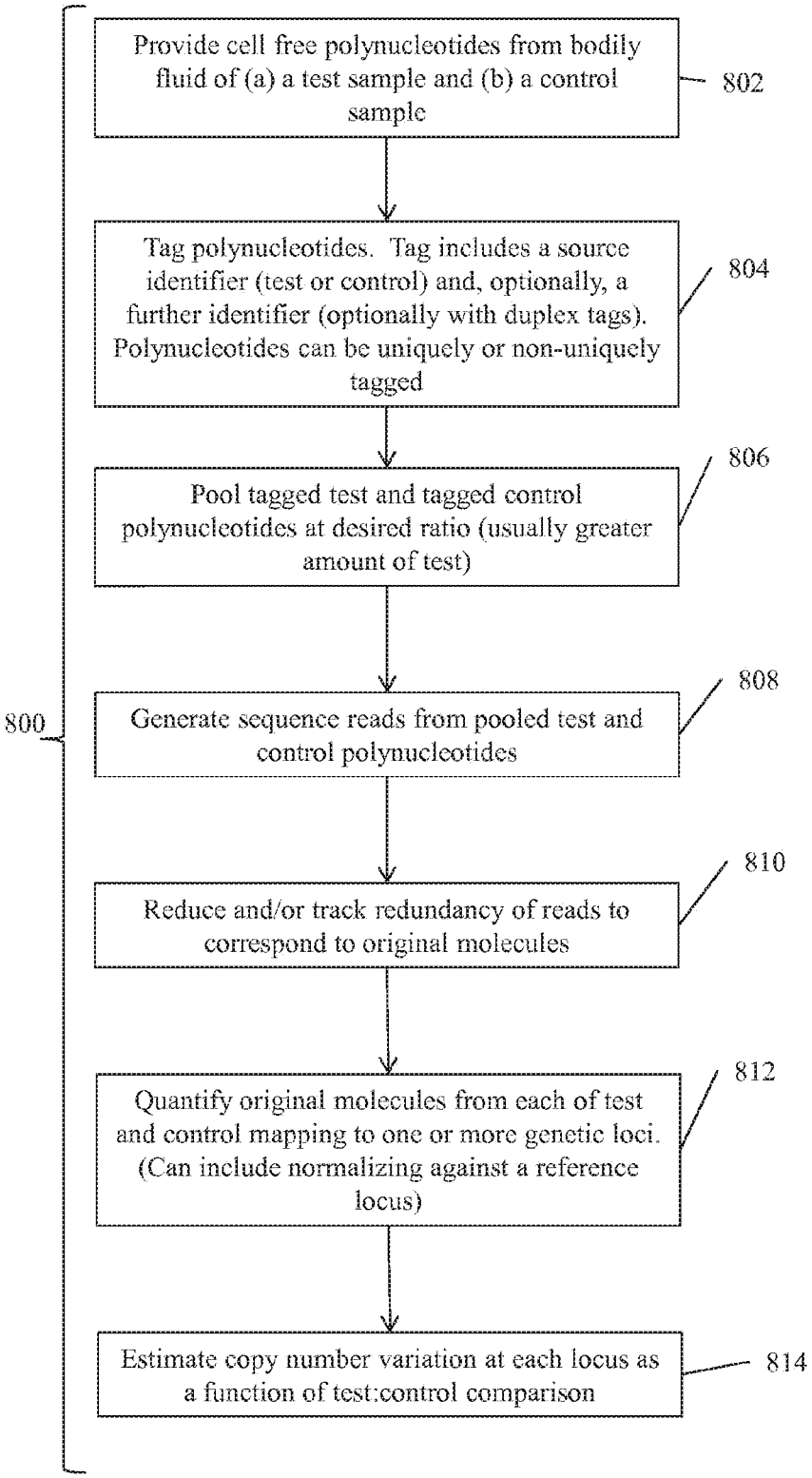

Provide cell free polynucleotides from bodily fluid of (a) a test sample and (b) a control sample
802
Tag polynucleotides. Tag includes a source identifier (test or control) and, optionally, a further identifier (optionally with duplex tags). Polynucleotides can be uniquely or non-uniquely tagged
804
Pool tagged test and tagged control polynucleotides at desired ratio (usually greater amount of test)
806
Generate sequence reads from pooled test and control polynucleotides
808
800
Reduce and/or track redundancy of reads to correspond to original molecules
810
Quantify original molecules from each of test and control mapping to one or more genetic loci. (Can include normalizing against a reference locus)
812
Estimate copy number variation at each locus as a function of test:control comparison
814

Fig. 8

Fig. 9A

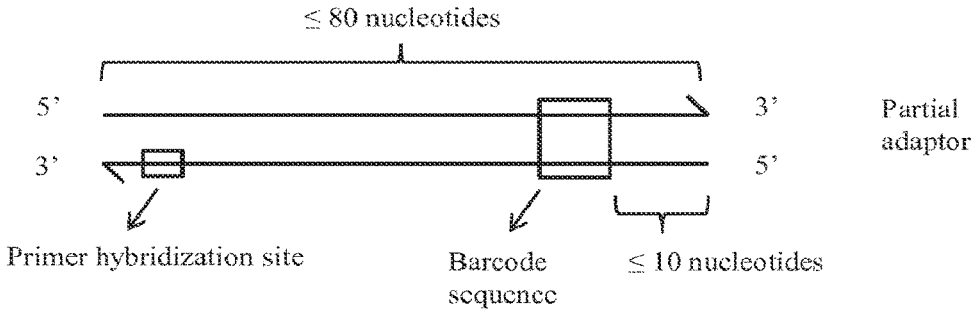


Fig. 9B

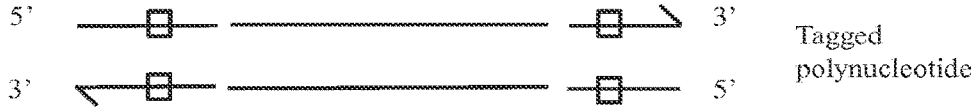
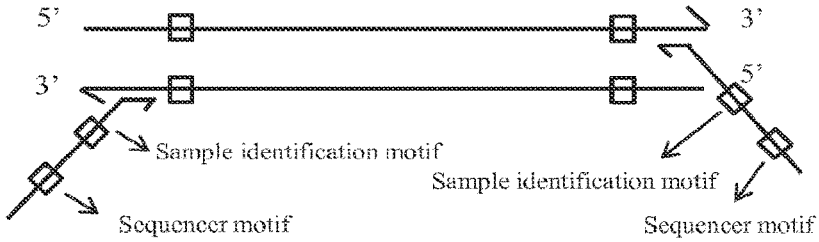


Fig. 9C



METHODS AND SYSTEMS FOR DETECTING GENETIC VARIANTS

CROSS-REFERENCE

This application is a continuation of U.S. application Ser. No. 16/601,168, filed Oct. 14, 2019 (now U.S. Pat. No. 10,801,063, issued Oct. 13, 2020), which is a continuation of U.S. application Ser. No. 15/892,178, filed Feb. 8, 2018 (now U.S. Pat. No. 10,883,139, issued Jan. 5, 2021), which is a continuation of U.S. application Ser. No. 14/861,989, filed Sep. 22, 2015 (now U.S. Pat. No. 9,920,366, issued Mar. 20, 2018), which is a continuation application of International Application No. PCT/US2014/072383, filed Dec. 24, 2014, which application claims the benefit under 35 U.S.C. § 119(e) of U.S. Provisional Application No. 61/921,456, filed Dec. 28, 2013, and U.S. Provisional Application No. 61/948,509, filed Mar. 5, 2014, each of which is entirely incorporated herein by reference.

BACKGROUND

The detection and quantification of polynucleotides is important for molecular biology and medical applications, such as diagnostics. Genetic testing is particularly useful for a number of diagnostic methods. For example, disorders that are caused by rare genetic alterations (e.g., sequence variants) or changes in epigenetic markers, such as cancer and partial or complete aneuploidy, may be detected or more accurately characterized with DNA sequence information.

Early detection and monitoring of genetic diseases, such as cancer, is often useful and needed in the successful treatment or management of the disease. One approach may include the monitoring of a sample derived from cell-free nucleic acids, a population of polynucleotides that can be found in different types of bodily fluids. In some cases, disease may be characterized or detected based on detection of genetic aberrations, such as copy number variation and/or sequence variation of one or more nucleic acid sequences, or the development of other certain rare genetic alterations. Cell-free DNA (cfDNA) may contain genetic aberrations associated with a particular disease. With improvements in sequencing and techniques to manipulate nucleic acids, there is a need in the art for improved methods and systems for using cell-free DNA to detect and monitor disease.

In particular, many methods have been developed for accurate copy number variation estimation, especially for heterogeneous genomic samples, such as tumor-derived gDNA or for cfDNA for many applications (e.g., prenatal, transplant, immune, metagenomics or cancer diagnostics). Most of these methods include sample preparation whereby the original nucleic acids are converted into a sequenceable library, followed by massively parallel sequencing, and finally bioinformatics to estimate copy number variation at one or more loci.

SUMMARY

Although many of these methods are able to reduce or combat the errors introduced by the sample preparation and sequencing processes for all molecules that are converted and sequenced, these methods are not able to infer the counts of molecules that were converted but not sequenced. Since this count of converted by unsequenced molecules can be highly variable from genomic region to region, these counts can dramatically and adversely affect the sensitivity that can be achieved.

To address this issue, input double-stranded deoxyribonucleic acid (DNA) can be converted by a process that tags both halves of the individual double-stranded molecule, in some cases differently. This can be performed using a variety of techniques, including ligation of hairpin, bubble, or forked adapters or other adaptors having double-stranded and single stranded segments (the unhybridized portion of a bubble, forked or hairpin adapter are deemed single-stranded herein). If tagged correctly, each original Watson and Crick (i.e., strand) side of the input double-stranded DNA molecule can be differently tagged and identified by the sequencer and subsequent bioinformatics. For all molecules in a particular region, counts of molecules where both Watson and Crick sides were recovered ("Pairs") versus those where only one half was recovered ("Singlets") can be recorded. The number of unseen molecules can be estimated based on the number of Pairs and Singlets detected.

An aspect of the present disclosure provides a method for detecting and/or quantifying rare deoxyribonucleic acid (DNA) in a heterogeneous population of original DNA fragments, comprising tagging the original DNA fragments in a single reaction using a library of a plurality of different tags such that greater than 30% of the fragments are tagged at both ends, wherein each of the tags comprises a molecular barcode. The single reaction can be in a single reaction vessel. Greater than 50% of the fragments can be tagged at both ends. The plurality of different tags can be no more than any of 100, 500, 1000, 10,000 or 100,000 different tags.

Another aspect provides a set of library adaptors that can be used to tag the molecules of interest (e.g., by ligation, hybridization, etc.). The set of library adaptors can comprise plurality of polynucleotide molecules with molecular barcodes, wherein the plurality of polynucleotide molecules are less than or equal to 80 nucleotide bases in length, wherein the molecular barcodes are at least 4 nucleotide bases in length, and wherein (a) the molecular barcodes are different from one another and have an edit distance of at least 1 between one another; (b) the molecular barcodes are located at least one nucleotide base away from a terminal end of their respective polynucleotide molecules; (c) optionally, at least one terminal base is identical in all of the polynucleotide molecules; and (d) none of the polynucleotide molecules contains a complete sequencer motif.

In some embodiments, the library adaptors (or adapters) are identical to one another but for the molecular barcodes. In some embodiments, each of the plurality of library adaptors comprises at least one double-stranded portion and at least one single-stranded portion (e.g., a non-complementary portion or an overhang). In some embodiments, the double-stranded portion has a molecular barcode selected from a collection of different molecular barcodes. In some embodiments, the given molecular barcode is a randomer. In some embodiments, each of the library adaptors further comprises a strand-identification barcode on the at least one single-stranded portion. In some embodiments, the strand-identification barcode includes at least 4 nucleotide bases. In some embodiments, the single-stranded portion has a partial sequencer motif. In some embodiments, the library adaptors do not include a complete sequencer motif.

In some embodiments, none of the library adaptors contains a sequence for hybridizing to a flow cell or forming a hairpin for sequencing.

In some embodiments, all of the library adaptors have a terminal end with nucleotide(s) that are the same. In some embodiments, the identical terminal nucleotide(s) are over two or more nucleotide bases in length.

In some embodiments, each of the library adapters is Y-shaped, bubble shaped or hairpin shaped. In some embodiments, none of the library adapters contains a sample identification motif. In some embodiments, each of the library adapters comprises a sequence that is selectively hybridizable to a universal primer. In some embodiments, each of the library adapters comprises a molecular barcode that is at least 5, 6, 7, 8, 9 and 10 nucleotide bases in length. In some embodiments, each of the library adapters is from 10 nucleotide bases to 80 in length, or 30 to 70 nucleotide bases in length, or 40 to 60 nucleotide bases in length. In some embodiments, at least 1, 2, 3, or 4 terminal bases are identical in all of the library adaptors. In some embodiments, at least 4 terminal bases are identical in all of the library adaptors.

In some embodiments, the edit distance of the molecular barcodes of the library adapters is a Hamming distance. In some embodiments, the edit distance is at least 1, 2, 3, 4 or 5. In some embodiments, the edit distance is with respect to individual bases of the plurality of polynucleotide molecules. In some embodiments, the molecular barcodes are located at least 10 nucleotide base away from a terminal end of an adapter. In some embodiments, the plurality of library adapters includes at least 2, 4, 6, 8, 10, 20, 30, 40 or 50 different molecular barcodes, or from 2-100, 4-80, 6-60 or 8-40 different molecular barcodes. In any of the embodiments herein, there are more polynucleotides (e.g., cfDNA fragments) to be tagged than there are different molecular barcodes such that the tagging is not unique.

In some embodiments, the terminal end of an adaptor is configured for ligation (e.g., to a target nucleic acid molecule). In some embodiments, the terminal end of an adaptor is a blunt end.

In some embodiments, the adaptors are purified and isolated. In some embodiments, the library comprises one or more non-naturally occurring bases.

In some embodiments, the polynucleotide molecules comprise a primer sequence positioned 5' with respect to the molecular barcodes.

In some embodiments, the set of library adaptors consists essentially of the plurality of polynucleotide molecules.

In another aspect, a method comprises (a) tagging a collection of polynucleotides with a plurality of polynucleotide molecules from a library of adaptors to create a collection of tagged polynucleotides; and (b) amplifying the collection of tagged polynucleotides in the presence of sequencing adaptors, wherein the sequencing adaptors have primers with nucleotide sequences that are selectively hybridizable to complementary sequences in the plurality of polynucleotide molecules. The library of adaptors may be as described above or elsewhere herein. In some embodiments, each of the sequencer adaptors further comprises an index tag, which can be a sample identification motif.

Another aspect, provides a method for detecting and/or quantifying rare DNA in a heterogeneous population of original DNA fragments, wherein the rare DNA has a concentration that is less than 1%, the method comprising (a) tagging the original DNA fragments in a single reaction such that greater than 30% of the original DNA fragments are tagged at both ends with library adaptors that comprise molecular barcodes, thereby providing tagged DNA fragments; (b) performing high-fidelity amplification on the tagged DNA fragments; (c) optionally, selectively enriching a subset of the tagged DNA fragments; (d) sequencing one or both strands of the tagged, amplified and optionally selectively enriched DNA fragments to obtain sequence reads comprising nucleotide sequences of the molecular barcodes and at least a portion of the original DNA fragments; (e) from the sequence reads, determining consensus reads that are representative of single-strands of the original DNA fragments; and (f) quantifying the consensus reads to detect and/or quantify the rare DNA at a specificity that is greater than 99.9%.

In some embodiments, (e) comprises comparing sequence reads having the same or similar molecular barcodes and the same or similar end of fragment sequences. In some embodiments, the comparing further comprises performing a phylogentic analysis on the sequence reads having the same or similar molecular barcodes. In some embodiments, the molecular barcodes include a barcode having an edit distance of up to 3. In some embodiments, the end of fragment sequence includes fragment sequences having an edit distance of up to 3.

In some embodiments, the method further comprises sorting sequence reads into paired reads and unpaired reads, and quantifying a number of paired reads and unpaired reads that map to each of one or more genetic loci.

In some embodiments, the tagging occurs by having an excess amount of library adaptors as compared to original DNA fragments. In some embodiments, n the excess is at least a 5-fold excess. In some embodiments, the tagging comprises using a ligase. In some embodiments, the tagging comprises attachment to blunt ends.

In some embodiments, the method further comprises binning the sequence reads according to the molecular barcodes and sequence information from at least one end of each of the original DNA fragments to create bins of single stranded reads. In some embodiments, the method further comprises, in each bin, determining a sequence of a given original DNA fragment among the original DNA fragments by analyzing sequence reads. In some embodiments, the method further comprises detecting and/or quantifying the rare DNA by comparing a number of times each base occurs at each position of a genome represented by the tagged, amplified, and optionally enriched DNA fragments.

In some embodiments, the library adaptors do not contain complete sequencer motifs. In some embodiments, the method further comprises selectively enriching a subset of the tagged DNA fragments. In some embodiments, the method further comprises, after enriching, amplifying the enriched tagged DNA fragments in the presence of sequencing adaptors comprising primers. In some embodiments, (a) provides tagged DNA fragments having from 2 to 1000 different combinations of molecular barcodes.

In some embodiments, the DNA fragments are tagged with polynucleotide molecules from a library of adaptors as described above or elsewhere herein.

In another aspect, a method for processing and/or analyzing a nucleic acid sample of a subject comprises (a) exposing polynucleotide fragments from the nucleic acid sample to a set of library adaptors to generate tagged polynucleotide fragments; and (b) subjecting the tagged polynucleotide fragments to nucleic acid amplification reactions under conditions that yield amplified polynucleotide fragments as amplification products of the tagged polynucleotide fragments. The set of library adaptors comprises a plurality of polynucleotide molecules with molecular barcodes, wherein the plurality of polynucleotide molecules are less than or equal to 80 nucleotide bases in length, wherein the molecular barcodes are at least 4 nucleotide bases in length, and wherein (1) the molecular barcodes are different from one another and have an edit distance of at least 1 between one another; (2) the molecular barcodes are located at least one nucleotide base away from a terminal end of

their respective polynucleotide molecules; (3) optionally, at least one terminal base is identical in all of the polynucleotide molecules; and (4) none of the polynucleotide molecules contains a complete sequencer motif.

In some embodiments, the method further comprises determining nucleotide sequences of the amplified tagged polynucleotide fragments. In some embodiments, the nucleotide sequences of the amplified tagged polynucleotide fragments are determined without polymerase chain reaction (PCR). In some embodiments, the method further comprises analyzing the nucleotide sequences with a programmed computer processor to identify one or more genetic variants in the nucleotide sample of the subject. In some embodiments, the one or more genetic variants are selected from the group consisting of base change(s), insertion(s), repeat(s), deletion(s), copy number variation(s) and transversion(s). In some embodiments, the one or more genetic variants include one or more tumor associated genetic alterations.

In some embodiments, the subject has or is suspected of having a disease. In some embodiments, the disease is cancer. In some embodiments, the method further comprises collecting the nucleic acid sample from the subject. In some embodiments, the nucleic acid sample is collected from a location selected from the group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, cerebral spinal fluid and tears of the subject. In some embodiments, the nucleic acid sample is a cell-free nucleic acid sample. In some embodiments, the nucleic acid sample is collected from no more than 100 nanograms (ng) of double-stranded polynucleotide molecules of the subject.

In some embodiments, the polynucleotide fragments comprise double-stranded polynucleotide molecules. In some embodiments, in (a), the plurality of polynucleotide molecules couple to the polynucleotide fragments via blunt end ligation, sticky end ligation, molecular inversion probes, PCR, ligation-based PCR, multiplex PCR, single stranded ligation, and single stranded circularization. In some embodiments, exposing the polynucleotide fragments of the nucleic acid sample to the plurality of polynucleotide molecules yields the tagged polynucleotide fragments with a conversion efficiency of at least 10%. In some embodiments, any of at least 5%, 6%, 7%, 8%, 9%, 10%, 20%, or 25% of the tagged polynucleotide fragments share a common polynucleotide molecule or sequence. In some embodiments, the method further comprises generating the polynucleotide fragments from the nucleic acid sample.

In some embodiments, the subjecting comprises amplifying the tagged polynucleotide fragments from sequences corresponding to genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

In another aspect, a method comprises (a) generating a plurality of sequence reads from a plurality of polynucleotide molecules, wherein the plurality of polynucleotide molecules cover genomic loci of a target genome, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1; (b) grouping with a computer processor the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides; (c) for each of the families, merging sequence reads to generate a consensus sequence; (d) calling the consensus sequence at a given genomic locus among the genomic loci; and (e) detecting at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls, and total number of alterations among the calls.

In some embodiments, each family comprises sequence reads from only one of the template polynucleotides. In some embodiments, the given genomic locus comprises at least one nucleic acid base. In some embodiments, the given genomic locus comprises a plurality of nucleic acid bases. In some embodiments, the calling comprises calling at least one nucleic acid base at the given genomic locus. In some embodiments, the calling comprises calling a plurality of nucleic acid bases at the given genomic locus. In some embodiments, the calling comprises any one of phylogenetic analysis, voting, weighing, assigning a probability to each read at the locus in a family and calling the base with the highest probability.

In some embodiments, the method further comprises performing (d)-(e) at an additional genomic locus among the genomic loci. In some embodiments, the method further comprises determining a variation in copy number at one of the given genomic locus and additional genomic locus based on counts at the given genomic locus and additional genomic locus.

In some embodiments, the grouping comprises classifying the plurality of sequence reads into families by identifying (i) different molecular barcodes coupled to the plurality of polynucleotide molecules and (ii) similarities between the plurality of sequence reads, wherein each family includes a plurality of nucleic acid sequences that are associated with a different combination of molecular barcodes and similar or identical sequence reads. Different molecular barcodes have different sequences.

In some embodiments, the consensus sequence is generated by evaluating a quantitative measure or a statistical significance level for each of the sequence reads. In some embodiments, the quantitative measure comprises use of a binomial distribution, exponential distribution, beta distribution, or empirical distribution. In some embodiments, the method further comprises mapping the consensus sequence to the target genome. In some embodiments, the plurality of genes includes at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50 or all of the plurality of genes selected from the group.

Another aspect of the present disclosure provides a method, comprising (a) providing template polynucleotide molecules and a set of library adaptors in a single reaction vessel, wherein the library adaptors are polynucleotide molecules that have different molecular barcodes (e.g., from 2 to 1,000 different molecular barcodes), and wherein none of the library adaptors contains a complete sequencer motif; (b) in the single reaction vessel, coupling the library adaptors to

the template polynucleotide molecules at an efficiency of at least 10%, thereby tagging each template polynucleotide with a tagging combination that is among a plurality of different tagging combinations (e.g., 4 to 1,000,000 different tagging combinations), to produce tagged polynucleotide molecules; (c) subjecting the tagged polynucleotide molecules to an amplification reaction under conditions that yield amplified polynucleotide molecules as amplification products of the tagged polynucleotide molecules; and (d) sequencing the amplified polynucleotide molecules.

In some embodiments, the template polynucleotide molecules are blunt ended or sticky-ended. In some embodiments, the library adaptors are identical but for the molecular barcodes. In some embodiments, each of the library adaptors has a double stranded portion and at least one single-stranded portion. In some embodiments, the double-stranded portion has a molecular barcode among the molecular barcodes. In some embodiments, each of the library adaptors further comprises a strand-identification barcode on the at least one single-stranded portion. In some embodiments, the single-stranded portion has a partial sequencer motif. In some embodiments, the library adaptors have a sequence of terminal nucleotides that are the same. In some embodiments, the template polynucleotide molecules are double-stranded. In some embodiments, the library adaptors couple to both ends of the template polynucleotide molecules.

In some embodiments, subjecting the tagged polynucleotide molecules to the amplification reaction comprises non-specifically amplifying the tagged polynucleotide molecules.

In some embodiments, the amplification reaction comprises use of a priming site to amplify each of the tagged polynucleotide molecules. In some embodiments, the priming site is a primer. In some embodiments, the primer is a universal primer. In some embodiments, the priming site is a nick.

In some embodiments, the method further comprises, prior to (e), (i) separating polynucleotide molecules comprising one or more given sequences from the amplified polynucleotide molecules, to produce enriched polynucleotide molecules; and (ii) amplifying the enriched polynucleotide molecules with sequencing adaptors.

In some embodiments, the efficiency is at least 30%, 40%, or 50%. In some embodiments, the method further comprises identifying genetic variants upon sequencing the amplified polynucleotide molecules. In some embodiments, the sequencing comprises (i) subjecting the amplified polynucleotide molecules to an additional amplification reaction under conditions that yield additional amplified polynucleotide molecules as amplification products of the amplified polynucleotide molecules, and (ii) sequencing the additional amplified polynucleotide molecules. In some embodiments, the additional amplification is performed in the presence of sequencing adaptors.

In some embodiments, (b) and (c) are performed without aliquoting the tagged polynucleotide molecules. In some embodiments, the tagging is non-unique tagging.

Another aspect, provides a system for analyzing a target nucleic acid molecule of a subject, comprising a communication interface that receives nucleic acid sequence reads for a plurality of polynucleotide molecules that cover genomic loci of a target genome; computer memory that stores the nucleic acid sequence reads for the plurality of polynucleotide molecules received by the communication interface; and a computer processor operatively coupled to the communication interface and the memory and programmed to (i) group the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides, (ii) for each of the families, merge sequence reads to generate a consensus sequence, (iii) call the consensus sequence at a given genomic locus among the genomic loci, and (iv) detect at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls; and total number of alterations among the calls, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

In another aspect, a set of oligonucleotide molecules that selectively hybridize to at least 5 genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

In some embodiments, the oligonucleotide molecules are from 10-200 bases in length. In some embodiments, the oligonucleotide molecules selectively hybridize to exon regions of the at least 5 genes. In some embodiments, the oligonucleotide molecules selectively hybridize to at least 30 exons in the at least 5 genes. In some embodiments, multiple oligonucleotide molecules selectively hybridize to each of the at least 30 exons. In some embodiments, the oligonucleotide molecules that hybridize to each exon have sequences that overlap with at least 1 other oligonucleotide molecule.

In another aspect, a kit comprises a first container containing a plurality of library adaptors each having a different molecular barcode; and a second container containing a plurality of sequencing adaptors, each sequencing adaptor comprising at least a portion of a sequencer motif and optionally a sample barcode. The library adaptors can be as described above or elsewhere herein.

In some embodiments, the sequencing adaptor comprises the sample barcode. In some embodiments, the library adaptors are blunt ended and Y-shaped, and are less than or equal to 80 nucleic acid bases in length. In some embodiments, the sequencing adaptor is up to 70 bases from end to end.

In another aspect, a method for detecting sequence variants in a cell free DNA sample, comprising detecting rare DNA at a concentration less than 1% with a specificity that is greater than 99.9%.

In another aspect, a method comprises detecting genetic variants in a sample comprising DNA with a detection limit of at least 1% and specificity greater than 99.9%. In some

embodiments, the method further comprises converting cDNA (e.g. cfDNA) into adaptor tagged DNA with a conversion efficiency of at least 30%, 40%, or 50% and reducing sequencing noise (or distortion) by eliminating false positive sequence reads.

Another aspect provides a method, comprising (a) providing a sample comprising a set of double-stranded polynucleotide molecules, each double-stranded polynucleotide molecule including first and second complementary strands; (b) tagging the double-stranded polynucleotide molecules with a set of duplex tags, wherein each duplex tag differently tags the first and second complementary strands of a double-stranded polynucleotide molecule in the set; (c) sequencing at least some of the tagged strands to produce a set of sequence reads; (d) reducing and/or tracking redundancy in the set of sequence reads; (e) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; (f) determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci; and (g) estimating with a programmed computer processor a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

In some embodiments, the method further comprises (h) detecting copy number variation in the sample by determining a normalized total quantitative measure determined in step (g) at each of the one or more genetic loci and determining copy number variation based on the normalized measure. In some embodiments, the sample comprises double-stranded polynucleotide molecules sourced substantially from cell-free nucleic acids. In some embodiments, the duplex tags are not sequencing adaptors.

In some embodiments, reducing redundancy in the set of sequence reads comprises collapsing sequence reads produced from amplified products of an original polynucleotide molecule in the sample back to the original polynucleotide molecule. In some embodiments, the method further comprises determining a consensus sequence for the original polynucleotide molecule. In some embodiments, the method further comprises identifying polynucleotide molecules at one or more genetic loci comprising a sequence variant. In some embodiments, the method further comprises determining a quantitative measure of paired reads that map to a locus, wherein both strands of the pair comprise a sequence variant. In some embodiments, the method further comprises determining a quantitative measure of paired molecules in which only one member of the pair bears a sequence variant and/or determining a quantitative measure of unpaired molecules bearing a sequence variant. In some embodiments, the sequence variant is selected from the group consisting of a single nucleotide variant, an indel, a transversion, a translocation, an inversion, a deletion, a chromosomal structure alteration, a gene fusion, a chromosome fusion, a gene truncation, a gene amplification, a gene duplication and a chromosomal lesion.

Another aspect provides a system comprising a computer readable medium comprising machine-executable code that, upon execution by a computer processor, implements a method comprising (a) receiving into memory a set of sequence reads of polynucleotides tagged with duplex tags; (b) reducing and/or tracking redundancy in the set of sequence reads; (c) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; (d) determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci; and (e) estimating a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

Another aspect provides a method, comprising (a) providing a sample comprising a set of double-stranded polynucleotide molecules, each double-stranded polynucleotide molecule including first and second complementary strands; (b) tagging the double-stranded polynucleotide molecules with a set of duplex tags, wherein each duplex tag differently tags the first and second complementary strands of a double-stranded polynucleotide molecule in the set; (c) sequencing at least some of the tagged strands to produce a set of sequence reads; (d) reducing and/or tracking redundancy in the set of sequence reads; (e) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; and (f) determining quantitative measures of at least two of (i) the paired reads, (ii) the unpaired reads that map to each of one or more genetic loci, (iii) read depth of the paired reads and (iv) read depth of unpaired reads.

In some embodiments, (f) comprises determining quantitative measures of at least three of (i)-(iv). In some embodiments, (f) comprises determining quantitative measures of all of (i)-(iv). In some embodiments, the method further comprises (g) estimating with a programmed computer processor a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads and their read depths mapping to each locus.

In another aspect, a method comprises (a) tagging control parent polynucleotides with a first tag set to produce tagged control parent polynucleotides, wherein the first tag set comprises a plurality of tags, wherein each tag in the first tag set comprises a same control tag and an identifying tag, and wherein the tag set comprises a plurality of different identifying tags; (b) tagging test parent polynucleotides with a second tag set to produce tagged test parent polynucleotides, wherein the second tag set comprises a plurality of tags, wherein each tag in the second tag set comprises a same test tag that is distinguishable from the control tag and an identifying tag, and wherein the second tag set comprises a plurality of different identifying tags; (c) mixing tagged control parent polynucleotides with tagged test parent polynucleotides to form a pool; (d) amplifying tagged parent

polynucleotides in the pool to form a pool of amplified, tagged polynucleotides; (e) sequencing amplified, tagged polynucleotides in the amplified pool to produce a plurality of sequence reads; (f) grouping sequence reads into families, each family comprising sequence reads generated from a same parent polynucleotide, which grouping is optionally based on information from an identifying tag and from start/end sequences of the parent polynucleotides, and, optionally, determining a consensus sequence for each of a plurality of parent polynucleotides from the plurality of sequence reads in a group; (g) classifying each family or consensus sequence as a control parent polynucleotide or as a test parent polynucleotide based on having a test tag or a control tag; (h) determining a quantitative measure of control parent polynucleotides and control test polynucleotides mapping to each of at least two genetic loci; and (i) determining copy number variation in the test parent polynucleotides at at least one locus based on relative quantity of test parent polynucleotides and control parent polynucleotides mapping to the at least one locus.

In another aspect, a method comprises (a) generating a plurality of sequence reads from a plurality of template polynucleotides, each polynucleotide mapped to a genomic locus; (b) grouping the sequence reads into families, each family comprising sequence reads generated from one of the template polynucleotides; (c) calling a base (or sequence) at the genomic locus for each of the families; (d) detecting at the genomic locus any of genomic alterations among the calls, frequency of a genetic alteration among the calls, total number of calls and total number of alterations among the calls.

In some embodiments, calling comprises any of phylogenetic analysis, voting, weighing, assigning a probability to each read at the locus in a family, and calling the base with the highest probability. In some embodiments, the method is performed at two loci, comprising determining CNV at one of the loci based on counts at each of the loci.

Another aspect provides a method for determining a quantitative measure indicative of a number of individual double-stranded DNA fragments in a sample comprising (a) determining a quantitative measure of individual DNA molecules for which both strands are detected; (b) determining a quantitative measure of individual DNA molecules for which only one of the DNA strands are detected; (c) inferring from (a) and (b) above a quantitative measure of individual DNA molecules for which neither strand was detected; and (d) using (a)-(c) determining the quantitative measure indicative of a number of individual double-stranded DNA fragments in the sample.

In some embodiments, the method further comprises detecting copy number variation in the sample by determining a normalized quantitative measure determined in step (d) at each of one or more genetic loci and determining copy number variation based on the normalized measure. In some embodiments, the sample comprises double-stranded polynucleotide molecules sourced substantially from cell-free nucleic acids.

In some embodiments, determining the quantitative measure of individual DNA molecules comprises tagging the DNA molecules with a set of duplex tags, wherein each duplex tag differently tags complementary strands of a double-stranded DNA molecule in the sample to provide tagged strands. In some embodiments, the method further comprises sequencing at least some of the tagged strands to produce a set of sequence reads. In some embodiments, the method further comprises sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads. In some embodiments, the method further comprises determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci to determine a quantitative measure of total double-stranded DNA molecules in the sample that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

In another aspect, a method for reducing distortion in a sequencing assay, comprises (a) tagging control parent polynucleotides with a first tag set to produce tagged control parent polynucleotides; (b) tagging test parent polynucleotides with a second tag set to produce tagged test parent polynucleotides; (c) mixing tagged control parent polynucleotides with tagged test parent polynucleotides to form a pool; (d) determining quantities of tagged control parent polynucleotides and tagged test parent polynucleotides; and (e) using the quantities of tagged control parent polynucleotides to reduce distortion in the quantities of tagged test parent polynucleotides.

In some embodiments, the first tag set comprises a plurality of tags, wherein each tag in the first tag set comprises a same control tag and an identifying tag, and wherein the first tag set comprises a plurality of different identifying tags. In some embodiments, the second tag set comprises a plurality of tags, wherein each tag in the second tag set comprises a same test tag and an identifying tag, wherein the test tag is distinguishable from the control tag, and wherein the second tag set comprises a plurality of different identifying tags. In some embodiments, (d) comprises amplifying tagged parent polynucleotides in the pool to form a pool of amplified, tagged polynucleotides, and sequencing amplified, tagged polynucleotides in the amplified pool to produce a plurality of sequence reads. In some embodiments, the method further comprises grouping sequence reads into families, each family comprising sequence reads generated from a same parent polynucleotide, which grouping is optionally based on information from an identifying tag and from start/end sequences of the parent polynucleotides, and, optionally, determining a consensus sequence for each of a plurality of parent polynucleotides from the plurality of sequence reads in a group.

In some embodiments, (d) comprises determining copy number variation in the test parent polynucleotides at greater than or equal to one locus based on relative quantity of test parent polynucleotides and control parent polynucleotides mapping to the locus.

Another aspect provides a method comprising (a) ligating adaptors to double-stranded DNA polynucleotides, wherein ligating is performed in a single reaction vessel, and wherein the adaptors comprise molecular barcodes, to produce a tagged library comprising an insert from the double-stranded DNA polynucleotides, and having between 4 and 1 million different tags; (b) generating a plurality of sequence reads for each of the double-stranded DNA polynucleotides in the tagged library; (c) grouping sequence reads into families, each family comprising sequence reads generated from a single DNA polynucleotide among the double-stranded DNA polynucleotides, based on information in a tag and information at an end of the insert; and (d) calling bases at

each position in the double-stranded DNA molecule based on bases at the position in members of a family. In some embodiments, (b) comprises amplifying each of the double-stranded DNA polynucleotide molecules in the tagged library to generate amplification products, and sequencing the amplification products. In some embodiments, the method further comprises sequencing the double-stranded DNA polynucleotide molecules a plurality of times. In some embodiments, (b) comprises sequencing the entire insert. In some embodiments, (c) further comprises collapsing sequence reads in each family to generate a consensus sequence. In some embodiments, (d) comprises calling a plurality of sequential bases from at least a subset of the sequence reads to identify single nucleotide variations (SNV) in the double-stranded DNA molecule.

Another aspect provides a method of detecting disease cell heterogeneity from a sample comprising polynucleotides from somatic cells and disease cells. The method comprises quantifying polynucleotides in the sample bearing a nucleotide sequence variant at each of a plurality of genetic loci; determining copy number variation (CNV) at each of the plurality of genetic loci, wherein the CNV indicates a genetic dose of a locus in the disease cell polynucleotides; determining with a programmed computer processor a relative measure of quantity of polynucleotides bearing a sequence variant at a locus per the genetic dose at the locus for each of a plurality of the loci; and comparing the relative measures at each of the plurality of loci, wherein different relative measures is indicative of tumor heterogeneity.

In another aspect, a method comprises subjecting a subject to one or more pulsed therapy cycles, each pulsed therapy cycle comprising (a) a first period during which a drug is administered at a first amount; and (b) a second period during which the drug is administered at a second, reduced amount, wherein (i) the first period is characterized by a tumor burden detected above a first clinical level; and (ii) the second period is characterized by a tumor burden detected below a second clinical level.

Additional aspects and advantages of the present disclosure will become readily apparent to those skilled in this art from the following detailed description, wherein only illustrative embodiments of the present disclosure are shown and described. As will be realized, the present disclosure is capable of other and different embodiments, and its several details are capable of modifications in various obvious respects, all without departing from the disclosure. Accordingly, the drawings and description are to be regarded as illustrative in nature, and not as restrictive.

INCORPORATION BY REFERENCE

All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings (also "figure" and "FIG." herein), of which:

FIG. **1** is a flowchart representation of a method of the present disclosure for determining copy number variation (CNV);

FIG. **2** depicts mapping of pairs and singlets to Locus A and Locus B in a genome;

FIG. **3** shows a reference sequence encoding a genetic Locus A;

FIGS. **4**A-C shows amplification, sequencing, redundancy reduction and pairing of complementary molecules;

FIG. **5** shows increased confidence in detecting sequence variants by pairing reads from Watson and Crick strands;

FIG. **6** shows a computer system that is programmed or otherwise configured to implement various methods of the present disclosure;

FIG. **7** is schematic representation of a system for analyzing a sample comprising nucleic acids from a user, including a sequencer; bioinformatic software and internet connection for report analysis by, for example, a hand held device or a desk top computer;

FIG. **8** is a flowchart representation of a method of this invention for determining CNV using pooled test and control pools; and

FIGS. **9**A-**9**C schematically illustrate a method for tagging a polynucleotide molecule with a library adaptor and subsequently a sequencing adaptor.

DETAILED DESCRIPTION

While various embodiments of the invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions may occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed.

The term "genetic variant," as used herein, generally refers to an alteration, variant or polymorphism in a nucleic acid sample or genome of a subject. Such alteration, variant or polymorphism can be with respect to a reference genome, which may be a reference genome of the subject or other individual. Single nucleotide polymorphisms (SNPs) are a form of polymorphisms. In some examples, one or more polymorphisms comprise one or more single nucleotide variations (SNVs), insertions, deletions, repeats, small insertions, small deletions, small repeats, structural variant junctions, variable length tandem repeats, and/or flanking sequences,. Copy number variants (CNVs), transversions and other rearrangements are also forms of genetic variation. A genomic alternation may be a base change, insertion, deletion, repeat, copy number variation, or transversion.

The term "polynucleotide," as used herein, generally refers to a molecule comprising one or more nucleic acid subunits. A polynucleotide can include one or more subunits selected from adenosine (A), cytosine (C), guanine (G), thymine (T) and uracil (U), or variants thereof. A nucleotide can include A, C, G, T or U, or variants thereof. A nucleotide can include any subunit that can be incorporated into a growing nucleic acid strand. Such subunit can be an A, C, G, T, or U, or any other subunit that is specific to one or more complementary A, C, G, T or U, or complementary to a purine (i.e., A or G, or variant thereof) or a pyrimidine (i.e., C, T or U, or variant thereof). A subunit can enable individual nucleic acid bases or groups of bases (e.g., AA, TA, AT, GC, CG, CT, TC, GT, TG, AC, CA, or uracil-counterparts thereof) to be resolved. In some examples, a polynucleotide is deoxyribonucleic acid (DNA) or ribonucleic

acid (RNA), or derivatives thereof. A polynucleotide can be single-stranded or double stranded.

The term "subject," as used herein, generally refers to an animal, such as a mammalian species (e.g., human) or avian (e.g., bird) species, or other organism, such as a plant. More specifically, the subject can be a vertebrate, a mammal, a mouse, a primate, a simian or a human. Animals include, but are not limited to, farm animals, sport animals, and pets. A subject can be a healthy individual, an individual that has or is suspected of having a disease or a pre-disposition to the disease, or an individual that is in need of therapy or suspected of needing therapy. A subject can be a patient.

The term "genome" generally refers to an entirety of an organism's hereditary information. A genome can be encoded either in DNA or in RNA. A genome can comprise coding regions that code for proteins as well as non-coding regions. A genome can include the sequence of all chromosomes together in an organism. For example, the human genome has a total of 46 chromosomes. The sequence of all of these together constitutes a human genome.

The terms "adaptor(s)", "adapter(s)" and "tag(s)" are used synonymously throughout this specification. An adaptor or tag can be coupled to a polynucleotide sequence to be "tagged" by any approach including ligation, hybridization, or other approaches.

The term "library adaptor" or "library adapter" as used herein, generally refers to a molecule (e.g., polynucleotide) whose identity (e.g., sequence) can be used to differentiate polynucleotides in a biological sample (also "sample" herein).

The term "sequencing adaptor," as used herein, generally refers to a molecule (e.g., polynucleotide) that is adapted to permit a sequencing instrument to sequence a target polynucleotide, such as by interacting with the target polynucleotide to enable sequencing. The sequencing adaptor permits the target polynucleotide to be sequenced by the sequencing instrument. In an example, the sequencing adaptor comprises a nucleotide sequence that hybridizes or binds to a capture polynucleotide attached to a solid support of a sequencing system, such as a flow cell. In another example, the sequencing adaptor comprises a nucleotide sequence that hybridizes or binds to a polynucleotide to generate a hairpin loop, which permits the target polynucleotide to be sequenced by a sequencing system. The sequencing adaptor can include a sequencer motif, which can be a nucleotide sequence that is complementary to a flow cell sequence of other molecule (e.g., polynucleotide) and usable by the sequencing system to sequence the target polynucleotide. The sequencer motif can also include a primer sequence for use in sequencing, such as sequencing by synthesis. The sequencer motif can include the sequence(s) needed to couple a library adaptor to a sequencing system and sequence the target polynucleotide.

As used herein the terms "at least", "at most" or "about", when preceding a series, refers to each member of the series, unless otherwise identified.

The term "about" and its grammatical equivalents in relation to a reference numerical value can include a range of values up to plus or minus 10% from that value. For example, the amount "about 10" can include amounts from 9 to 11. In other embodiments, the term "about" in relation to a reference numerical value can include a range of values plus or minus 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1% from that value.

The term "at least" and its grammatical equivalents in relation to a reference numerical value can include the reference numerical value and greater than that value. For example, the amount "at least 10" can include the value 10 and any numerical value above 10, such as 11, 100, and 1,000.

The term "at most" and its grammatical equivalents in relation to a reference numerical value can include the reference numerical value and less than that value. For example, the amount "at most 10" can include the value 10 and any numerical value under 10, such as 9, 8, 5, 1, 0.5, and 0.1.

1. Methods for Processing and/or Analyzing a Nucleic Acid Sample

An aspect of the present disclosure provides methods for determining a genomic alternation in a nucleic acid sample of a subject. FIG. **1** shows a method of determining copy number variation (CNV). The method can be implemented to determine other genomic alternations, such as SNVs.

A. Polynucleotide Isolation

Methods disclosed herein can comprise isolating one or more polynucleotides. A polynucleotide can comprise any type of nucleic acid, for example, a sequence of genomic nucleic acid, or an artificial sequence (e.g., a sequence not found in genomic nucleic acid). For example, an artificial sequence can contain non-natural nucleotides. Also, a polynucleotide can comprise both genomic nucleic acid and an artificial sequence, in any portion. For example, a polynucleotide can comprise 1 to 99% of genomic nucleic acid and 99% to 1% of artificial sequence, where the total adds up to 100%. Thus, fractions of percentages are also contemplated. For example, a ratio of 99.1% to 0.9% is contemplated.

A polynucleotide can comprise any type of nucleic acids, such as DNA and/or RNA. For example, if a polynucleotide is DNA, it can be genomic DNA, complementary DNA (cDNA), or any other deoxyribonucleic acid. A polynucleotide can also be cell-free DNA (cfDNA). For example, the polynucleotide can be circulating DNA. The circulating DNA can comprise circulating tumor DNA (ctDNA). A polynucleotide can be double-stranded or single-stranded. Alternatively, a polynucleotide can comprise a combination of a double-stranded portion and a single-stranded portion.

Polynucleotides do not have to be cell-free. In some cases, the polynucleotides can be isolated from a sample. For example, in step (**102**) (FIG. **1**), double-stranded polynucleotides are isolated from a sample. A sample can be any biological sample isolated from a subject. For example, a sample can comprise, without limitation, bodily fluid, whole blood, platelets, serum, plasma, stool, red blood cells, white blood cells or leucocytes, endothelial cells, tissue biopsies, synovial fluid, lymphatic fluid, ascites fluid, interstitial or extracellular fluid, the fluid in spaces between cells, including gingival crevicular fluid, bone marrow, cerebrospinal fluid, saliva, mucous, sputum, semen, sweat, urine, or any other bodily fluids. A bodily fluid can include saliva, blood, or serum. For example, a polynucleotide can be cell-free DNA isolated from a bodily fluid, e.g., blood or serum. A sample can also be a tumor sample, which can be obtained from a subject by various approaches, including, but not limited to, venipuncture, excretion, ejaculation, massage, biopsy, needle aspirate, lavage, scraping, surgical incision, or intervention or other approaches.

A sample can comprise various amount of nucleic acid that contains genome equivalents. For example, a sample of about 30 ng DNA can contain about 10,000 (10^4) haploid human genome equivalents and, in the case of cfDNA, about

200 billion (2×10^{11}) individual polynucleotide molecules. Similarly, a sample of about 100 ng of DNA can contain about 30,000 haploid human genome equivalents and, in the case of cfDNA, about 600 billion individual molecules.

A sample can comprise nucleic acids from different sources. For example, a sample can comprise germline DNA or somatic DNA. A sample can comprise nucleic acids carrying mutations. For example, a sample can comprise DNA carrying germline mutations and/orsomatic mutations. A sample can also comprise DNA carrying cancer-associated mutations (e.g., cancer-associated somatic mutations).

B. Tagging

Polynucleotides disclosed herein can be tagged. For example, in step (**104**) (FIG. **1**) the double-stranded polynucleotides are tagged with duplex tags, tags that differently label the complementary strands (i.e., the "Watson" and "Crick" strands) of a double-stranded molecule. In one embodiment the duplex tags are polynucleotides having complementary and non-complementary portions.

Tags can be any types of molecules attached to a polynucleotide, including, but not limited to, nucleic acids, chemical compounds, florescent probes, or radioactive probes. Tags can also be oligonucleotides (e.g., DNA or RNA). Tags can comprise known sequences, unknown sequences, or both. A tag can comprise random sequences, pre-determined sequences, or both. A tag can be double-stranded or single-stranded. A double-stranded tag can be a duplex tag. A double-stranded tag can comprise two complementary strands. Alternatively, a double-stranded tag can comprise a hybridized portion and a non-hybridized portion. The double-stranded tag can be Y-shaped, e.g., the hybridized portion is at one end of the tag and the non-hybridized portion is at the opposite end of the tag. One such example are the "Y adapters" used in Illumina sequencing. Other examples include hairpin shaped adapters or bubble shaped adapters. Bubble shaped adapters have non-complementary sequences flanked on both sides by complementary sequences.

Tagging disclosed herein can be performed using any method. A polynucleotide can be tagged with an adaptor by hybridization. For example, the adaptor can have a nucleotide sequence that is complementary to at least a portion of a sequence of the polynucleotide. As an alternative, a polynucleotide can be tagged with an adaptor by ligation.

For example, tagging can comprise using one or more enzymes. The enzyme can be a ligase. The ligase can be a DNA ligase. For example, the DNA ligase can be a T4 DNA ligase, *E. coli* DNA ligase, and/or mammalian ligase. The mammalian ligase can be DNA ligase I, DNA ligase III, or DNA ligase IV. The ligase can also be a thermostable ligase. Tags can be ligated to a blunt-end of a polynucleotide (blunt-end ligation). Alternatively, tags can be ligated to a sticky end of a polynucleotide (sticky-end ligation). Efficiency of ligation can be increased by optimizing various conditions. Efficiency of ligation can be increased by optimizing the reaction time of ligation. For example, the reaction time of ligation can be less than 12 hours, e.g., less than 1, less than 2, less than 3, less than 4, less than 5, less than 6, less than 7, less than 8, less than 9, less than 10, less than 11, less than 12, less than 13, less than 14, less than 15, less than 16, less than 17, less than 18, less than 19, or less than 20 hours. In a particular example, reaction time of ligation is less than 20 hours. Efficiency of ligation can be increased by optimizing the ligase concentration in the reaction. For example, the ligase concentration can be at least 10, at least 50, at least 100, at least 150, at least 200, at least 250, at least 300, at least 400, at least 500, or at least 600 unit/microliter. Efficiency can also be optimized by adding or varying the concentration of an enzyme suitable for ligation, enzyme cofactors or other additives, and/or optimizing a temperature of a solution having the enzyme. Efficiency can also be optimized by varying the addition order of various components of the reaction. The end of tag sequence can comprise dinucleotide to increase ligation efficiency. When the tag comprises a non-complementary portion (e.g., Y-shaped adaptor), the sequence on the complementary portion of the tag adaptor can comprise one or more selected sequences that promote ligation efficiency. Preferably such sequences are located at the terminal end of the tag. Such sequences can comprise 1, 2, 3, 4, 5, or 6 terminal bases. Reaction solution with high viscosity (e.g., a low Reynolds number) can also be used to increase ligation efficiency. For example, solution can have a Reynolds number less than 3000, less than 2000, less than 1000, less than 900, less than 800, less than 700, less than 600, less than 500, less than 400, less than 300, less than 200, less than 100, less than 50, less than 25, or less than 10. It is also contemplated that roughly unified distribution of fragments (e.g., tight standard deviation) can be used to increase ligation efficiency. For example, the variation in fragment sizes can vary by less than 20%, less than 15%, less than 10%, less than 5%, or less than 1%. Tagging can also comprise primer extension, for example, by polymerase chain reaction (PCR). Tagging can also comprise any of ligation-based PCR, multiplex PCR, single strand ligation, or single strand circularization.

In some instances, the tags herein comprise molecular barcodes. Such molecular barcodes can be used to differentiate polynucleotides in a sample. Preferably molecular barcodes are different from one another. For example, molecular barcodes can have a difference between them that can be characterized by a predetermined edit distance or a Hamming distance. In some instances, the molecular barcodes herein have a minimum edit distance of 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10. To further improve efficiency of conversion (e.g., tagging) of untagged molecular to tagged molecules, one preferably utilizes short tags. For example, in some embodiments, a library adapter tag can be up to 65, 60, 55, 50, 45, 40, or 35 nucleotide bases in length. A collection of such short library barcodes preferably includes a number of different molecular barcodes, e.g., at least 2, 4, 6, 8, 10, 12, 14, 16, 18 or 20 different barcodes with a minimum edit distance of 1, 2, 3 or more.

Thus, a collection of molecules can include one or more tags. In some instances, some molecules in a collection can include an identifying tag ("identifier") such as a molecular barcode that is not shared by any other molecule in the collection. For example, in some instances of a collection of molecules, at least 50%, at least 51%, at least 52%, at least 53%, at least 54%, at least 55%, at least 56%, at least 57%, at least 58%, at least 59%, at least 60%, at least 61%, at least 62%, at least 63%, at least 64%, at least 65%, at least 66%, at least 67%, at least 68%, at least 69%, at least 70%, at least 71%, at least 72%, at least 73%, at least 74%, at least 75%, at least 76%, at least 77%, at least 78%, at least 79%, at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% of the molecules in the collection can include an identifier or molecular barcode that is not shared by any other molecule in the collection. As used

herein, a collection of molecules is considered to be "uniquely tagged" if each of at least 95% of the molecules in the collection bears an identifier that is not shared by any other molecule in the collection ("unique tag" or "unique identifier"). A collection of molecules is considered to be "non-uniquely tagged" if each of at least 1%, at least 5%, at least 10%, at least 15%, at least 20%, at least 25%, at least 30%, at least 35%, at least 40%, at least 45%, or at least or about 50% of the molecules in the collection bears an identifying tag or molecular barcode that is shared by at least one other molecule in the collection ("non-unique tag" or "non-unique identifier"). Accordingly, in a non-uniquely tagged population no more than 1% of the molecules are uniquely tagged. For example, in a non-uniquely tagged population, no more than 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, or 50% of the molecules can be uniquely tagged.

A number of different tags can be used based on the estimated number of molecules in a sample. In some tagging methods, the number of different tags can be at least the same as the estimated number of molecules in the sample. In other tagging methods, the number of different tags can be at least two, three, four, five, six, seven, eight, nine, ten, one hundred or one thousand times as many as the estimated number of molecules in the sample. In unique tagging, at least two times (or more) as many different tags can be used as the estimated number of molecules in the sample.

The molecules in the sample may be non-uniquely tagged. In such instances a fewer number of tags or molecular barcodes is used then the number of molecules in the sample to be tagged. For example, no more than 100, 50, 40, 30, 20 or 10 unique tags or molecular barcodes are used to tag a complex sample such as a cell free DNA sample with many more different fragments.

The polynucleotide to be tagged can be fragmented, such as either naturally or using other approaches, such as, for example, shearing. The polynucleotides can be fragmented by certain methods, including but not limited to, mechanical shearing, passing the sample through a syringe, sonication, heat treatment (e.g., for 30 minutes at 90° C.), and/or nuclease treatment (e.g., using DNase, RNase, endonuclease, exonuclease, and/or restriction enzyme).

The polynucleotides fragments (prior to tagging) can comprise sequences of any length. For example, polynucleotide fragments (prior to tagging) can comprise at least 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, 275, 280, 285, 290, 295, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000 or more nucleotides in length. The polynucleotide fragment are preferably about the average length of cell-free DNA. For example, the polynucleotide fragments can comprise about 160 bases in length. The polynucleotide fragment can also be fragmented from a larger fragment into smaller fragments about 160 bases in length.

Polynucleotides tagged can comprise sequences associated with cancer. The cancer-associated sequences can comprise single nucleotide variation (SNV), copy number variation (CNV), insertions, deletions, and/or rearrangements.

The polynucleotides can comprise sequences associated with cancer, such as acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), adrenocortical carcinoma, Kaposi Sarcoma, anal cancer, basal cell carcinoma, bile duct cancer, bladder cancer, bone cancer, osteosarcoma, malignant fibrous histiocytoma, brain stem glioma, brain cancer, craniopharyngioma, ependymoblastoma, ependymoma, medulloblastoma, medulloeptithelioma, pineal parenchymal tumor, breast cancer, bronchial tumor, Burkitt lymphoma, Non-Hodgkin lymphoma, carcinoid tumor, cervical cancer, chordoma, chronic lymphocytic leukemia (CLL), chronic myelogenous leukemia (CML), colon cancer, colorectal cancer, cutaneous T-cell lymphoma, ductal carcinoma in situ, endometrial cancer, esophageal cancer, Ewing Sarcoma, eye cancer, intraocular melanoma, retinoblastoma, fibrous histiocytoma, gallbladder cancer, gastric cancer, glioma, hairy cell leukemia, head and neck cancer, heart cancer, hepatocellular (liver) cancer, Hodgkin lymphoma, hypopharyngeal cancer, kidney cancer, laryngeal cancer, lip cancer, oral cavity cancer, lung cancer, non-small cell carcinoma, small cell carcinoma, melanoma, mouth cancer, myelodysplastic syndromes, multiple myeloma, medulloblastoma, nasal cavity cancer, paranasal sinus cancer, neuroblastoma, nasopharyngeal cancer, oral cancer, oropharyngeal cancer, osteosarcoma, ovarian cancer, pancreatic cancer, papillomatosis, paraganglioma, parathyroid cancer, penile cancer, pharyngeal cancer, pituitary tumor, plasma cell neoplasm, prostate cancer, rectal cancer, renal cell cancer, rhabdomyosarcoma, salivary gland cancer, Sezary syndrome, skin cancer, nonmelanoma, small intestine cancer, soft tissue sarcoma, squamous cell carcinoma, testicular cancer, throat cancer, thymoma, thyroid cancer, urethral cancer, uterine cancer, uterine sarcoma, vaginal cancer, vulvar cancer, Waldenstrom macroglobulinemia, and/or Wilms Tumor.

A haploid human genome equivalent has about 3 picograms of DNA. A sample of about 1 microgram of DNA contains about 300,000 haploid human genome equivalents. Improvements in sequencing can be achieved as long as at least some of the duplicate or cognate polynucleotides bear unique identifiers with respect to each other, that is, bear different tags. However, in certain embodiments, the number of tags used is selected so that there is at least a 95% chance that all duplicate molecules starting at any one position bear unique identifiers. For example, in a sample comprising about 10,000 haploid human genome equivalents of fragmented genomic DNA, e.g., cfDNA, z is expected to be between 2 and 8. Such a population can be tagged with between about 10 and 100 different identifiers, for example, about 2 identifiers, about 4 identifiers, about 9 identifiers, about 16 identifiers, about 25 identifiers, about 36 different identifiers, about 49 different identifiers, about 64 different identifiers, about 81 different identifiers, or about 100 different identifiers.

Nucleic acid barcodes having identifiable sequences including molecular barcodes, can be used for tagging. For example, a plurality of DNA barcodes can comprise various numbers of sequences of nucleotides. A plurality of DNA barcodes having 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more identifiable sequences of nucleotides can be used. When attached to only one end of a polynucleotide, the plurality of DNA barcodes can produce 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more different identifiers. Alternatively, when attached to both ends of a polynucleotide, the plurality DNA barcodes can produce 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 324, 361, 400 or more different identifiers (which is the ^2 of when the DNA barcode is attached to only 1 end of a polynucleotide). In one example, a plurality of DNA barcodes having 6, 7, 8, 9 or 10 identifiable sequences of nucleotides can be used. When attached to both ends of a polynucleotide, they produce 36,

49, 64, 81 or 100 possible different identifiers, respectively. In a particular example, the plurality of DNA barcodes can comprise 8 identifiable sequences of nucleotides. When attached to only one end of a polynucleotide, the plurality of DNA barcodes can produce 8 different identifiers. Alternatively, when attached to both ends of a polynucleotide, the plurality of DNA barcodes can produce 64 different identifiers. Samples tagged in such a way can be those with a range of about 10 ng to any of about 100 ng, about 1 μg, about 10 μg of fragmented polynucleotides, e.g., genomic DNA, e.g., cfDNA.

A polynucleotide can be uniquely identified in various ways. A polynucleotide can be uniquely identified by a unique DNA barcode. For example, any two polynucleotides in a sample are attached two different DNA barcodes. Alternatively, a polynucleotide can be uniquely identified by the combination of a DNA barcode and one or more endogenous sequences of the polynucleotide. For example, any two polynucleotides in a sample can be attached the same DNA barcode, but the two polynucleotides can still be identified by different endogenous sequences. The endogenous sequence can be on an end of a polynucleotide. For example, the endogenous sequence can be adjacent (e.g., base in between) to the attached DNA barcode. In some instances the endogenous sequence can be at least 2, 4, 6, 8, 10, 20, 30, 40, 50, 60, 70, 80, 90, or 100 bases in length. Preferably, the endogenous sequence is a terminal sequence of the fragment/polynucleotides to be analyzed. The endogenous sequence may be the length of the sequence. For example, a plurality of DNA barcodes comprising 8 different DNA barcodes can be attached to both ends of each polynucleotide in a sample. Each polynucleotide in the sample can be identified by the combination of the DNA barcodes and about 10 base pair endogenous sequence on an end of the polynucleotide. Without being bound by theory, the endogenous sequence of a polynucleotide can also be the entire polynucleotide sequence.

Also disclosed herein are compositions of tagged polynucleotides. The tagged polynucleotide can be single-stranded. Alternatively, the tagged polynucleotide can be double-stranded (e.g., duplex-tagged polynucleotides). Accordingly, this invention also provides compositions of duplex-tagged polynucleotides. The polynucleotides can comprise any types of nucleic acids (DNA and/or RNA). polynucleotides comprise any types of DNA disclosed herein. For example, the polynucleotides can comprise DNA, e.g., fragmented DNA or cfDNA. A set of polynucleotides in the composition that map to a mappable base position in a genome can be non-uniquely tagged, that is, the number of different identifiers can be at least 2 and fewer than the number of polynucleotides that map to the mappable base position. The number of different identifiers can also be at least 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 and fewer than the number of polynucleotides that map to the mappable base position.

In some instances, as a composition goes from about 1 ng to about 10 μg or higher, a larger set of different molecular barcodes can be used. For example, between 5 and 100 different library adaptors can be used to tag polynucleotides in a cfDNA sample.

The systems and methods disclosed herein may be used in applications that involve the assignment of molecular barcodes. The molecular barcodes can be assigned to any types of polynucleotides disclosed in this invention. For example, the molecular barcodes can be assigned to cell-free polynucleotides (e.g., cfDNAs). Often, an identifier disclosed herein can be a barcode oligonucleotide that is used to tag the polynucleotide. The barcode identifier may be a nucleic acid oligonucleotide (e.g., a DNA oligonucleotide). The barcode identifier can be single-stranded. Alternatively, the barcode identifier can be double-stranded. The barcode identifier can be attached to polynucleotides using any method disclosed herein. For example, the barcode identifier can be attached to the polynucleotide by ligation using an enzyme. The barcode identifier can also be incorporated into the polynucleotide through PCR. In other cases, the reaction may comprise addition of a metal isotope, either directly to the analyte or by a probe labeled with the isotope. Generally, assignment of unique or non-unique identifiers or molecular barcodes in reactions of this disclosure may follow methods and systems described by, for example, U.S. patent applications 2001/0053519, 2003/0152490, 2011/0160078 and U.S. Pat. No. 6,582,908, each of which is entirely incorporated herein by reference.

Identifiers or molecular barcodes used herein may be completely endogenous whereby circular ligation of individual fragments may be performed followed by random shearing or targeted amplification. In this case, the combination of a new start and stop point of the molecule and the original intramolecular ligation point can form a specific identifier.

Identifiers or molecular barcodes used herein can comprise any types of oligonucleotides. In some cases, identifiers may be predetermined, random, or semi-random sequence oligonucleotides. Identifiers can be barcodes. For example, a plurality of barcodes may be used such that barcodes are not necessarily unique to one another in the plurality. Alternatively, a plurality of barcodes may be used such that each barcode is unique to any other barcode in the plurality. The barcodes can comprise specific sequences (e.g., predetermined sequences) that can be individually tracked. Further, barcodes may be attached (e.g., by ligation) to individual molecules such that the combination of the barcode and the sequence it may be ligated to creates a specific sequence that may be individually tracked. As described herein, detection of barcodes in combination with sequence data of beginning (start) and/or end (stop) portions of sequence reads can allow assignment of a unique identity to a particular molecule. The length or number of base pairs of an individual sequence read may also be used to assign a unique identity to such a molecule. As described herein, fragments from a single strand of nucleic acid having been assigned a unique identity, may thereby permit subsequent identification of fragments from the parent strand. In this way the polynucleotides in the sample can be uniquely or substantially uniquely tagged. A duplex tag can include a degenerate or semi-degenerate nucleotide sequence, e.g., a random degenerate sequence. The nucleotide sequence can comprise any number of nucleotides. For example, the nucleotide sequence can comprise 1 (if using a non-natural nucleotide), 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more nucleotides. In a particular example, the sequence can comprise 7 nucleotides. In another example, the sequence can comprise 8 nucleotides. The sequence can also comprise 9 nucleotides. The sequence can comprise 10 nucleotides.

A barcode can comprise contiguous or non-contiguous sequences. A barcode that comprises at least 1, 2, 3, 4, 5 or more nucleotides is a contiguous sequence or non-contiguous sequence. if the 4 nucleotides are uninterrupted by any other nucleotide. For example, if a barcode comprises the sequence TTGC, a barcode is contiguous if the barcode is

TTGC. On the other hand, a barcode is non-contiguous if the barcode is TTXGC, where X is a nucleic acid base.

An identifier or molecular barcode can have an n-mer sequence which may be 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more nucleotides in length. A tag herein can comprise any range of nucleotides in length. For example, the sequence can be between 2 to 100, 10 to 90, 20 to 80, 30 to 70, 40 to 60, or about 50 nucleotides in length.

The tag can comprise a double-stranded fixed reference sequence downstream of the identifier or molecular barcode. Alternatively, the tag can comprise a double-stranded fixed reference sequence upstream or downstream of the identifier or molecular barcode. Each strand of a double-stranded fixed reference sequence can be, for example, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 nucleotides in length.

C. Adaptors

A library of polynucleotide molecules can be synthesized for use in sequencing. For example, a library of polynucleotides comprising a plurality of polynucleotide molecules that are each less than or equal to 100, 90, 80, 70, 60, 50, 45, 40, or 35 nucleic acid (or nucleotide) bases in length can be made. A plurality of polynucleotide molecules can be each less than or equal to 35 nucleic acid bases in length. A plurality of polynucleotide molecules can be each less than or equal to 30 nucleic acid bases in length. A plurality of polynucleotide molecules can also be less than or equal to 250, 200, 150, 100, or 50 nucleic acid bases. Additionally, the plurality of polynucleotide molecules can also be less than or equal to 100, 99, 98, 97, 96, 95, 94, 93, 92, 91, 90, 89, 88, 87, 86, 85, 84, 83, 82, 81, 80, 79, 78, 77, 76, 75, 74, 73, 72, 71, 70, 69, 68, 67, 66, 65, 64, 63, 62, 61, 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 nucleic acid bases.

A library of polynucleotides comprising a plurality of polynucleotide molecules can also have distinct (with respect to each other) molecular barcode sequences (or molecular barcodes) with respect to at least 4 nucleic acid bases. A molecular barcode (also "barcode" or "identifier" herein) sequence is a nucleotide sequence that distinguishes one polynucleotide from another. In other embodiments, the polynucleotide molecules can also have different barcode sequences with respect to 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more nucleic acid bases.

A library of polynucleotides comprising a plurality of polynucleotide molecules can also have a plurality of different barcode sequences. For example, a plurality of polynucleotide molecules can have at least 4 different molecular barcode sequences. In some cases, the plurality of polynucleotide molecules has from 2-100, 4-50, 4-30, 4-20, or 4-10 different molecular barcode sequences. The plurality of polynucleotides molecules can also have other ranges of different barcode sequences such as, 1-4, 2-5, 3-6, 4-7, 5-8, 6-9, 7-10, 8-11, 9-12, 10-13, 11-14, 12-15, 13-16, 14-17, 15-18, 16-19, 17-20, 18-21, 19-22, 20-23, 21-24, or 22-25 different barcode sequences. In other cases, a plurality of polynucleotide molecules can have at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or 100 more different barcode sequences. In a particular example, the plurality library adapters comprise at least 8 different sequences.

The location of the different barcode sequences can vary within the plurality of polynucleotides. For example, the different barcode sequences can be within 20, 15, 10, 9, 8, 7, 6, 5, 4, 3, or 2 nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules. In an example, a plurality of polynucleotide molecules has distinct barcode sequences that are within 10 nucleic acid bases from the terminal end. In another example, a plurality of polynucleotide molecules has distinct barcode sequences that are within 5 or 1 nucleic acid bases from the terminal end. In other instances, the distinct barcode sequences can be at the terminal end of a respective one of the plurality of polynucleotide molecules. Other variations include that the distinct molecular barcode sequences can be within 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, or more nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules.

The terminal end of the plurality of polynucleotide molecules can be adapted for ligation to a target nucleic acid molecule. For example, the terminal end can be a blunt end. In some other cases, the terminal end is adapted for hybridization to a complementary sequence of a target nucleic acid molecule.

A library of polynucleotides comprising a plurality of polynucleotide molecules can also have an edit distance of at least 1. In some cases, the edit distance is with respect to individual bases of the plurality of polynucleotide molecules. In other cases, the plurality of polynucleotide molecules can have an edit distance of at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more. The edit distance can be a Hamming distance.

In some cases, the plurality of polynucleotides does not contain sequencing adaptors. A sequence adaptor can be a polynucleotide that comprises a sequence that hybridizes to one or more sequencing adaptors or primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, e.g., a flow cell sequence. The term "flow cell sequence" and its grammatical equivalents as used herein, refers to a sequence that permits hybridization to a substrate, for example, by way of a primer attached to the substrate. The substrate can be bead or a planar surface. In some embodiments, a flow cell sequence can allow a polynucle-

otide to attach to a flow cell or surface (e.g., surface of a bead, for example, an Illumina flow cell.

When a plurality of polynucleotide molecules does not contain sequencing adaptors or primers, each polynucleotide molecule of the plurality does not contain a nucleic acid sequence or other moiety that is adapted to permit sequencing of a target nucleic acid molecule with a given sequencing approach, such as Illumina, SOLiD, Pacific Biosciences, GeneReader, Oxford Nanopore, Complete Genomics, GnuBio, Ion Torrent, Oxford Nanopore or Genia. In some examples, when a plurality of polynucleotide molecules does not contain sequencing adaptors or primers, the plurality of polynucleotide molecules does not contain flow cell sequences. For example, the plurality of polynucleotide molecules cannot bind to flow cells, such as used in Illumina flow cell sequencers. However, these flow cell sequences, if desired, can be added to the plurality of polynucleotide molecules by methods such as PCR amplification or ligation. At this point, Illumina flow cell sequencers can be used. Alternatively, when the plurality of polynucleotide molecules does not contain sequencing adaptors or primers, the plurality of polynucleotide molecules does not contain hairpin shaped adaptors or adaptors for generating hairpin loops in a target nucleic acid molecule, such as Pacific Bioscience SMARTbell™ adaptors. However, these hairpin shaped adaptors, if desired, can be added to the plurality of polynucleotide molecules by methods such as PCR amplification or ligation. The plurality of polynucleotide molecules can be circular or linear.

A plurality of polynucleotide molecules can be double stranded. In some cases, the plurality of polynucleotide molecules can be single stranded, or can comprise hybridized and non-hybridized regions. A plurality of polynucleotide molecules can be non-naturally occurring polynucleotide molecules.

Adaptors can be polynucleotide molecules. The polynucleotide molecules can be Y-shaped, bubble-shaped or hairpin-shaped. A hairpin adaptor may contain a restriction site(s) or a Uracil containing base. Adaptors can comprise a complementary portion and a non-complementary portion. The non-complementary portion can have an edit distance (e.g., Hamming distance). For example, the edit distance can be at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, or at least 30. The complementary portion of the adaptor can comprise sequences that are selected to enable and/or promote ligation to a polynucleotide, e.g., a sequence to enable and/or promote ligation to a polynucleotide at a high yield.

A plurality of polynucleotide molecules as disclosed herein can be purified. In some cases, a plurality of polynucleotide molecules as disclosed herein can be isolated polynucleotide molecules. In other cases, a plurality of polynucleotide molecules as disclosed herein can be purified and isolated polynucleotide molecules.

In certain aspects, each of the plurality of polynucleotide molecules is Y-shaped or hairpin-shaped. Each of the plurality of polynucleotide molecules can comprise a different barcode. The different barcode can be a randomer in the complementary portion (e.g., double stranded portion) of the Y-shaped or hairpin-shaped adaptor. Alternatively, the different barcode can be in one strand of the non-complementary portion (e.g., one of the Y-shaped arms). As discussed above, the different barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, e.g., 7 bases. The barcode can be contiguous or non-contiguous sequences, as described above. The plurality of polynucleotide molecules is from 10 nucleic acid bases to 35 nucleic acid bases (or any length as described above) in length. Further, the plurality of polynucleotide molecules can comprise an edit distance (as described above), that is a Hamming distance. A plurality of polynucleotide molecules can have distinct barcode sequences that are within 10 nucleic acid bases from the terminal end.

In another aspect, a plurality of polynucleotide molecules can be sequencing adaptors. A sequencing adaptor can comprise a sequence hybridizing to one or more sequencing primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, e.g., a flow cell sequence. For example, a sequencing adaptor can be a flow cell adaptor. The sequencing adaptors can be attached to one or both ends of a polynucleotide fragment. In another example, a sequencing adaptor can be hairpin shaped. For example, the hairpin shaped adaptor can comprise a complementary double-stranded portion and a loop portion, where the double-stranded portion can be attached (e.g., ligated) to a double-stranded polynucleotide. Hairpin shaped sequencing adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. A sequencing adaptor can be up to 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, or more bases from end to end. For example, a sequencing adaptor can be up to 70 bases from end to end. The sequencing adaptor can comprise 20-30, 20-40, 30-50, 30-60, 40-60, 40-70, 50-60, 50-70, bases from end to end. In a particular example, the sequencing adaptor can comprise 20-30 bases from end to end. In another example, the sequencing adaptor can comprise 50-60 bases from end to end. A sequencing adaptor can comprise one or more barcodes. For example, a sequencing adaptor can comprise a sample barcode. The sample barcode can comprise a pre-determined sequence. The sample barcodes can be used to identify the source of the polynucleotides. The sample barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, e.g., at least 8 bases. The barcode can be contiguous or non-contiguous sequences, as described above.

The plurality of polynucleotide molecules as described herein can be used as adaptors. Adaptors can comprise one or more identifiers. An adaptor can comprise an identifier with a random sequence. Alternatively, an adaptor can comprise an identifier with pre-determined sequences. Some adaptors can comprise an identifier with a random sequence and another identifier with a pre-determined sequence. The adaptors comprising identifiers can be double-stranded or single-stranded adaptors. The adaptors comprising identifiers can be Y-shaped adaptors. A Y-shaped adaptor can comprise one or more identifiers with a random sequence. The one or more identifiers can be on the hybrid portion and/or non-hybridized portion of the Y-shaped adaptor. A Y-shaped adaptor can comprise one or more identifiers with a pre-determined sequence. The one or more identifiers with pre-determined sequence can be on the hybridized portion

and/or non-hybridized portion of the Y-shaped adaptor. A Y-shaped adaptor can comprise one or more identifiers with a random sequence and one or more identifiers with a pre-determined sequence. For example, the one or more identifiers with a random sequence can be on the hybridized portion of the Y-shaped adaptor and/or the non hybridized portion of the Y-shaped adaptor. The one or more identifiers with a pre-determined sequence can be on the hybridized portion of the Y-shaped adaptor and/or the non-hybridized portion of the Y-shaped adaptor. In a particular example, a Y-shaped adaptor can comprise an identifier with a random sequence on its hybridized portion and an identifier with a pre-determined sequence on its non- hybridized portion. The identifiers can be in any length disclosed herein. For example, a Y-shaped adaptor can comprise an identifier with a random sequence of 7 nucleotides on its hybridized portion and an identifier with a pre-determined sequence of 8 nucleotides on its non- hybridized portion.

An adaptor can include a double-stranded portion with a molecular barcode and at least one or two single-stranded portion. For example, the adaptor can be Y-shaped and include a double-stranded portion and two single-stranded portions. The single-stranded portions can include sequences that are not complementary to one another.

The adaptor can include a terminal end that has a sequence that is selected to permit the adaptor to be efficiently (e.g., at an efficiency of at least about 20%, 30%, 40%, 50%) ligated or otherwise coupled to a polynucleotide. In some examples, terminal nucleotides in a double-stranded portion of an adaptor are selected from a combination of purines and pyrimidines to provide for efficient ligation.

In some examples, a set of library adaptors comprises a plurality of polynucleotide molecules (library adaptors) with molecular barcodes. The library adaptors are less than or equal to 80, 70, 60, 50, 45, or 40 nucleotide bases in length. The molecular barcodes can be at least 4 nucleotide bases in length, but may be from 4 to 20 nucleotide bases in length. The molecular barcodes can be different from one another and have an edit distance of at least 1, 2, 3, 4, or 5 between one another. The molecular barcodes are located at least 1, 2, 3, 4, 5, 10, or 20 nucleotide bases away from a terminal end of their respective library adaptors. In some cases, the at least one terminal base is identical in all of the library adaptors.

The library adaptors can be identical but for the molecular barcodes. For example, the library adaptors can have identical sequences but differ only with respect to nucleotide sequences of the molecular barcodes.

Each of the library adaptors can have a double stranded portion and at least one single-stranded portion. By "single stranded portion" is meant an area of non-complementarity or an overhang. In some cases, each of the library adaptors has a double-stranded portion and two single-stranded portions. The double-stranded portion can have a molecular barcode. In some cases, the molecular barcode is a randomer. Each of the library adaptors can further include a strand-identification barcode on a single-stranded portion. The strand-identification barcode can include at least 4 nucleotide bases, in some cases from 4 to 20 nucleotide bases.

In some examples, each of the library adaptors has a double-stranded portion with a molecular barcode and two single-stranded portions. The single-stranded portions may not hybridize to one another. The single-stranded portions may not be completely complementary to one another.

The library adaptors can have a sequence of terminal nucleotides in a double-stranded portion that are the same. The sequence of terminal nucleotides can be at least 2, 3, 4, 5 or 6 nucleotide bases in length. For example, one strand of a double-stranded portion of the library adaptor can have the sequence ACTT, TCGC, or TACC at the terminal end, while the other strand can have a complementary sequence. In some cases, such a sequence is selected to optimize the efficiency at which the library adaptors ligate to target polynucleotides. Such sequences can be selected to optimize a binding interaction between the ends of the library adaptors and the target polynucleotides.

In some cases, none of the library adaptors contains a sample identification motif (or sample molecular barcode). Such sample identification motif can be provided via sequencing adaptors. A sample identification motif can include a sequencer of at least 4, 5, 6, 7, 8, 9, 10, 20, 30, or 40 nucleotide bases that permits the identification of polynucleotide molecules from a given sample from polynucleotide molecules from other samples. For example, this can permit polynucleotide molecules from two subjects to be sequenced in the same pool and sequence reads for the subjects subsequently identified.

A sequencer motif includes nucleotide sequence(s) needed to couple a library adaptor to a sequencing system and sequence a target polynucleotide coupled to the library adaptor. The sequencer motif can include a sequence that is complementary to a flow cell sequence and a sequence (sequencing initiation sequence) that is selectively hybridizable to a primer (or priming sequence) for use in sequencing. For example, such sequencing initiation sequence can be complementary to a primer that is employed for use in sequence by synthesis (e.g., Illumina). Such primer can be included in a sequencing adaptor. A sequencing initiation sequence can be a primer hybridization site.

In some cases, none of the library adaptors contains a complete sequencer motif. The library adaptors can contain partial or no sequencer motifs. In some cases, the library adaptors include a sequencing initiation sequence. The library adaptors can include a sequencing initiation sequence but no flow cell sequence. The sequence initiation sequence can be complementary to a primer for sequencing. The primer can be a sequence specific primer or a universal primer. Such sequencing initiation sequences may be situated on single-stranded portions of the library adaptors. As an alternative, such sequencing initiation sequences may be priming sites (e.g., kinks or nicks) to permit a polymerase to couple to the library adaptors during sequencing.

In some cases, partial or complete sequencer motifs are provided by sequencing adaptors. A sequencing adaptor can include a sample molecular barcode and a sequencer motif. The sequencing adaptors can be provided in a set that is separate from the library adaptors. The sequencing adaptors in a given set can be identical—i.e., they contain the same sample barcode and sequencer motif.

Sequencing adaptors can include sample identification motifs and sequencer motifs. Sequencer motifs can include primers that are complementary to a sequencing initiation sequence. In some cases, sequencer motifs also include flow cell sequences or other sequences that permit a polynucleotide to a configured or arranged in a manner that permits the polynucleotide to be sequenced by a sequencer.

Library adaptors and sequencing adaptors can each be partial adaptors, that is, containing part but not all of the sequences necessary to enable sequencing by a sequencing platform. Together they provide complete adaptors. For example, library adaptors can include partial or no sequencer motifs, but such sequencer motifs are provided by sequencing adaptors.

FIGS. **9**A-**9**C schematically illustrate a method for tagging a target polynucleotide molecule with library adaptors. FIG. **9**A shows a library adaptor as a partial adaptor containing a primer hybridization site on one of the strands and a molecular barcode towards another end. The primer hybridization site can be a sequencing initiation sequence for subsequent sequencing. The library adaptor is less than or equal to 80 nucleotide bases in length. In FIG. **9**B, the library adaptors are ligated at both ends of the target polynucleotide molecule to provide a tagged target polynucleotide molecule. The tagged target polynucleotide molecule may be subjected to nucleic acid amplification to generate copies of the target. Next, in FIG. **9**C, sequencing adaptors containing sequencer motifs are provided and hybridized to the tagged target polynucleotide molecule. The sequencing adaptors contain sample identification motifs. The sequencing adaptors can contain sequences to permit sequencing of the tagged target with a given sequencer.

D. Sequencing

Tagged polynucleotides can be sequenced to generate sequence reads (e.g., as shown in step (**106**), FIG. **1**). For example, a tagged duplex polynucleotide can be sequenced. Sequence reads can be generated from only one strand of a tagged duplex polynucleotide. Alternatively, both strands of a tagged duplex polynucleotide can generate sequence reads. The two strands of the tagged duplex polynucleotide can comprise the same tags. Alternatively, the two strands of the tagged duplex polynucleotide can comprise different tags. When the two strands of the tagged duplex polynucleotide are differently tagged, sequence reads generated from one strand (e.g., a Watson strand) can be distinguished from sequence reads generated from the other strands (e.g., a Crick strand). Sequencing can involve generating multiple sequence reads for each molecule. This occurs, for example, as a result the amplification of individual polynucleotide strands during the sequencing process, e.g., by PCR.

Methods disclosed herein can comprise amplifying of polynucleotides. Polynucleotides amplification can result in the incorporation of nucleotides into a nucleic acid molecule or primer thereby forming a new nucleic acid molecule complementary to a template nucleic acid. The newly formed polynucleotide molecule and its template can be used as templates to synthesize additional polynucleotides. The polynucleotides being amplified can be any nucleic acids, for example, deoxyribonucleic acids, including genomic DNAs, cDNAs (complementary DNA), cfDNAs, and circulating tumor DNAs (ctDNAs). The polynucleotides being amplified can also be RNAs. As used herein, one amplification reaction may comprise many rounds of DNA replication. DNA amplification reactions can include, for example, polymerase chain reaction (PCR). One PCR reaction may comprise 2-100 "cycles" of denaturation, annealing, and synthesis of a DNA molecule. For example, 2-7, 5-10, 6-11, 7-12, 8-13, 9-14, 10-15, 11-16, 12-17, 13-18, 14-19, or 15-20 cycles can be performed during the amplification step. The condition of the PCR can be optimized based on the GC content of the sequences, including the primers.

Nucleic acid amplification techniques can be used with the assays described herein. Some amplification techniques are the PCR methodologies which can include, but are not limited to, solution PCR and in situ PCR. For example, amplification may comprise PCR-based amplification. Alternatively, amplification may comprise non PCR-based amplification. Amplification of the template nucleic acid may comprise use of one or more polymerases. For example, the polymerase may be a DNA polymerase or an RNA polymerase. In some cases, high fidelity amplification is performed such as with the use of high fidelity polymerase (e.g., Phusion® High-Fidelity DNA Polymerase) or PCR protocols. In some cases, the polymerase may be a high fidelity polymerase. For example, the polymerase may be KAPA HiFi DNA polymerase. The polymerase may also be Phusion DNA polymerase. The polymerase may be used under reaction conditions that reduce or minimize amplification biases, e.g., due to fragment length, GC content, etc.

Amplification of a single strand of a polynucleotide by PCR will generate copies both of that strand and its complement. During sequencing, both the strand and its complement will generate sequence reads. However, sequence reads generated from the complement of, for example, the Watson strand, can be identified as such because they bear the complement of the portion of the duplex tag that tagged the original Watson strand. In contrast, a sequence read generated from a Crick strand or its amplification product will bear the portion of the duplex tag that tagged the original Crick strand. In this way, a sequence read generated from an amplified product of a complement of the Watson strand can be distinguished from a complement sequence read generated from an amplification product of the Crick strand of the original molecule.

All amplified polynucleotides can be submitted to a sequencing device for sequencing. Alternatively, a sampling, or subset, of all of the amplified polynucleotides is submitted to a sequencing device for sequencing. With respect to any original double-stranded polynucleotide there can be three results with respect to sequencing. First, sequence reads can be generated from both complementary strands of the original molecule (that is, from both the Watson strand and from the Crick strand). Second, sequence reads can be generated from only one of the two complementary strands (that is, either from the Watson strand or from the Crick strand, but not both). Third, no sequence read may be generated from either of the two complementary strands. Consequently, counting unique sequence reads mapping to a genetic locus will underestimate the number of double-stranded polynucleotides in the original sample mapping to the locus. Described herein are methods of estimating the unseen and uncounted polynucleotides.

The sequencing method can be massively parallel sequencing, that is, simultaneously (or in rapid succession) sequencing any of at least 100, 1000, 10,000, 100,000, 1 million, 10 million, 100 million, or 1 billion polynucleotide molecules. Sequencing methods may include, but are not limited to: high-throughput sequencing, pyrosequencing, sequencing-by-synthesis, single-molecule sequencing, nanopore sequencing, semiconductor sequencing, sequencing-by-ligation, sequencing-by-hybridization, RNA-Seq (Illumina), Digital Gene Expression (Helicos), Next generation sequencing, Single Molecule Sequencing by Synthesis (SMSS) (Helicos), massively-parallel sequencing, Clonal Single Molecule Array (Solexa), shotgun sequencing, Maxam-Gilbert or Sanger sequencing, primer walking, sequencing using PacBio, SOLiD, Ion Torrent, or Nanopore platforms and any other sequencing methods known in the art.

For example, duplex-tagged polynucleotides can be amplified, by for example PCR (see e.g., FIG. **4**A duplex-tagged polynucleotides are referred to as mm' and nn'). In FIG. **4**A, the strand of the duplex polynucleotide including sequence m bears sequence tags w and y, while the strand of the duplex polynucleotide including sequence m' bears

sequence tags x and z. Similarly, the strand of the duplex polynucleotide including sequence n bears sequence tags a and c, while the strand of the duplex polynucleotide including sequence n' bears sequence tags b and d. During amplification, each strand produces itself and its complementary sequence. However, for example, an amplification progeny of original strand m that includes the complementary sequence, m', is distinguishable from an amplification progeny of original strand m' because the progeny from original strand m will have the sequence 5'-y'm'w'-3' and the progeny of the original m' strand one strand will have the sequence 5'-zm'x-3'. FIG. **4**B shows amplification in more detail. During amplification, errors can be introduced into the amplification progeny, represented by dots. The application progeny are sampled for sequencing, so that not all strands produce sequence reads, resulting in the sequence reads indicated. Because sequence reads can come from either of a strand or its complement, both sequences and complement sequences will be included in the set of sequence reads. It should be noted that it is possible that a polynucleotide would bear the same tag on each end. Thus, for a tag "a", and polynucleotide "m", a first strand could be tagged a-m-a', and the complement could be tagged a-m'-a.

E. Determining Consensus Sequence Reads

Methods disclosed herein can comprise determining consensus sequence reads in sequence reads (e.g., as shown in step (**108**), FIG. **1**), such as by reducing or tracking redundancy. Sequencing of amplified polynucleotides can produce reads of the several amplification products from the same original polynucleotide, referred to as "redundant reads". By identifying redundant reads, unique molecules in the original sample can be determined. If the molecules in a sample are uniquely tagged, then reads generated from amplification of a single unique original molecule can be identified based on their distinct barcode. Ignoring barcodes, reads from unique original molecules can be determined based on sequences at the beginning and end of a read, optionally in combination with the length of the read. In certain cases, however, a sample may be expected to have a plurality of original molecules having the same start stop sequences and the same length. Without barcoding, these molecules are difficult to distinguish from one another. However, if a collection of polynucleotides is non-uniquely tagged (that is, an original molecule shares the same identifier with at least one other original molecule), combining information from a barcode with start/stop sequence and/or polynucleotide length significantly increases the probability that any sequence read can be traced back to an original polynucleotide. This is because, in part, even without unique tagging, it is unlikely that any two original polynucleotides having the same start/stop sequence and length also will be tagged with the same identifier.

F. Collapsing

Collapsing allows for reduction in noise (i.e., background) that is generated at each step of the process. Methods disclosed herein can comprise collapsing, e.g., generating a consensus sequence by comparing multiple sequence reads. For example, sequence reads generated from a single original polynucleotide can be used to generate a consensus sequence of that original polynucleotide. Iterative rounds of amplification can introduce errors into progeny polynucleotides. Also, sequencing typically may not be performed with perfect fidelity so sequencing errors are introduced at this stage as well. However, comparison of sequence reads of molecules derived from a single original molecule, including those that have sequence variants, can be analyzed so as to determine the original, or "consensus" sequence. This can be done phylogenetically. Consensus sequences can be generated from families of sequence reads by any of a variety of methods. Such methods include, for example, linear or non-linear methods of building consensus sequences (such as voting (e.g., biased voting), averaging, statistical, maximum a posteriori or maximum likelihood detection, dynamic programming, Bayesian, hidden Markov or support vector machine methods, etc.) derived from digital communication theory, information theory, or bioinformatics. For example, if all or most of the sequence reads tracking back to an original molecule bear the same sequence variant, that variant probably existed in the original molecule. On the other hand, if a sequence variant exists in a subset of redundant sequence reads, that variant may have been introduced during amplification/sequencing and represents an artifact not existing in the original. Furthermore, if only sequence reads derived from the Watson or Crick strand of an original polynucleotide contain the variant, the variant may have been introduced through single-sided DNA damage, first-cycle PCR error or through contaminating polynucleotides that were amplified from a different sample.

After fragments are amplified and the sequences of amplified fragments are read and aligned, the fragments are subjected to base calling, e.g., determining for each locus the most likely nucleotide. However, variations in the number of amplified fragments and unseen amplified fragments (e.g., those without being read their sequences; reasons could be too many such as amplification errors, sequencing reading errors, too long, too short, being chopped, etc.) may introduce errors in base calling. If there are too many unseen amplified fragments with respect to the seen amplified fragments (amplified fragments actually being read), the reliability of base calling may be diminished.

Therefore, disclosed herein is a method to correct for the number of unseen fragments in base calling. For example, when base calling for locus A (an arbitrary locus), it is first assumed that there are N amplified fragments. The sequence readouts can come from two types of fragments: double-strand fragments and single-strand fragments. Therefore, we assign N1, N2, and N3 as the numbers of double-strands, single-strands, and unseen fragments, respectively. Thus, N=N1+N2+N3 (N1 and N2 are known from the sequence readouts, and N and N3 are unknown). If the formula is solved for N (or N3), then N3 (or N) will be inferred.

Probability is used to estimate N. For example, we assign "p" to be the probability of having detected (or having read) a nucleotide of locus A in a sequence readout of a single-strand.

For sequence readouts from double-strands, the nucleotide call from a double-strand amplified fragment has a probability of p * p=p^2, seeing all N1 double-strands has the following equation: N1=N *(p^2).

For sequence readouts from a single-strand. Assuming that one of the 2 strands is seen, and the other is unseen, the probability of seeing one strand is "p", but the probability of missing the other strand is (1−p). Furthermore, by not distinguishing the single strand sourcing from 5-primer and sourcing from 3-primer, there is a factor of 2. Therefore, the nucleotide call from a single-strand amplified fragment has a probability 2×p×(1−p). Thus, seeing all N2 single-strands has the following equation: N2=N×2×p×(1−p).

"p" is also unknown. To solve p, the ratio of N1 to N2 is used to solve for "p":

$$R = \frac{N1}{N2} = \frac{Np^2}{2Np(1-p)} = \frac{p^2}{2p(1-p)} = \frac{p}{2(1-p)}$$

Once "p" is found, N can be found. After N is found, can be found N3=N−N1−N2.

Besides the ratio of paired versus unpaired strands (which is a measure post-collapsing), there is useful information in the pre-collapsing read depth at each locus. This information can be used to further improve the call for total molecule count and/or increase confidence of calling variants.

For example, FIG. **4**C demonstrates sequence reads corrected for complementary sequences. Sequences generated from an original Watson strand or an original Crick strand can be differentiated on the basis of their duplex tags. Sequences generated from the same original strand can be grouped. Examination of the sequences can allow one to infer the sequence of the original strand (the "consensus sequence"). In this case, for example, the sequence variant in the nn' molecule is included in the consensus sequence because it included in every sequence read while other variants are seen to be stray errors. After collapsing sequences, original polynucleotide pairs can be identified based on their complementary sequences and duplex tags.

FIG. **5** demonstrates increased confidence in detecting sequence variants by pairing reads from Watson and Crick strands. Sequence nn' can include a sequence variant indicated by a dot. In some cases, sequence pp' does not include a sequence variant. Amplification, sequencing, redundancy reduction and pairing can result in both Watson and Crick strands of the same original molecule including the sequence variant. In contrast, as a result of errors introduced during amplification and sampling during sequencing, the consensus sequence of the Watson strand p can contain a sequence variant, while the consensus sequence of the Crick strand p' does not. It is less likely that amplification and sequencing will introduce the same variant into both strands (nn' sequence) of a duplex than onto one strand (pp' sequence). Therefore, the variant in the pp' sequence is more likely to be an artifact, and the variant in the nn' sequence is more likely to exist in the original molecule.

Methods disclosed herein can be used to correct errors resulted from experiments, e.g., PCR, amplification, and/or sequencing. For example, such a method can comprises attaching one or more double stranded adaptors to both ends of a double stranded polynucleotide, thereby providing a tagged double stranded polynucleotide; amplifying the double stranded tagged polynucleotide; sequencing both strands of the tagged polynucleotide; comparing the sequence of one strand with its complement to determine any errors introduced during sequencing; and correcting errors in the sequence based on (d). The adaptors used in this method can be any adaptors disclosed herein, e.g., Y-shaped adaptors. The adaptor can comprise any barcodes (e.g., distinct barcodes) disclosed herein.

G. Mapping

Sequence reads or consensus sequences can be mapped to one or more selected genetic loci (e.g., as shown step (**110**), FIG. **1**). A genetic locus can be, for example, a specific nucleotide position in the genome, a sequence of nucleotides (for example, an open reading frame), a fragment of a chromosome, a whole chromosome, or an entire genome. A genetic locus can be a polymorphic locus. Polymorphic locus can be a locus at which sequence variation exists in the population and/or exists in a subject and/or a sample. A polymorphic locus can be generated by two or more distinct sequences coexisting at the same location of the genome. The distinct sequences can differ from one another by one or more nucleotide substitutions, a deletion/insertion, and/or a duplication of any number of nucleotides, generally a relatively small number of nucleotides, such as less than 50, 45, 40, 35, 30, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1 nucleotide(s), among others. A polymorphic locus can be created by a single nucleotide position that varies within the population, e.g. a single nucleotide variation (SNV) or a single nucleotide polymorphism (SNP).

A reference genome for mapping can include the genome of any species of interest. Human genome sequences useful as references can include the hg19 assembly or any previous or available hg assembly. Such sequences can be interrogated using the genome browser available at genome.ucsc.edu/index.html. Other species genomes include, for example PanTro2 (chimp) and mm9 (mouse).

In methods disclosed herein, collapsing can be performed before or after mapping. In some aspects, collapsing can be performed before mapping. For example, sequence reads can be grouped into families based on their tags and one or more endogenous sequences, without regard to where the reads map in the genome. Then, the members of a family can be collapsed into a consensus sequence. The consensus sequence can be generated using any collapsing method disclosed herein. Then the consensus sequence can be mapped to locations in the genome. Reads mapped to a locus can be quantified (e.g., counted). Percentage of reads carrying a mutation at a locus can also be determined. Alternatively, collapsing can be performed after mapping. For example, all reads can first be mapped to the genome. Then the reads can be grouped into families based on their tags and one or more endogenous sequences. Since the reads have been mapped to the genome, consensus bases can be determined for each family at each locus. In other aspects, consensus sequence can be generated for one strand of a DNA molecule (e.g., for a Watson strand or a Crick strand). Mapping can be performed before or after the consensus sequence for one strand of the DNA molecule is determined. Numbers of Doublets and Singlets can be determined. These numbers can be used to calculate unseen molecules. For example, the unseen molecules can be calculated using the following equation: N=D+S+U; D=Np(2), S=N2pq, where p=1−q, where p is the probability of seeing; q is the probability of missing a strand.

H. Grouping

Methods disclosed herein can also comprise grouping sequence reads. Sequence reads can be grouped based on various types of sequences, e.g., sequences of an oligonucleotide tag (e.g., a barcode), sequence of a polynucleotide fragments, or any combinations. For example, as shown in step (**112**) (FIG. **1**), sequence reads can be grouped as follows: Sequence reads generated from a "Watson" strand and those generated from a "Crick" strand of a double-stranded polynucleotide in the sample are identifiable based on the duplex tags that they bear. In this way, a sequence read or consensus sequence from a Watson strand of a duplex polynucleotide can be paired with a sequence read or

consensus sequence from its complementary Crick strand. Paired sequence reads are referred to as a "Pair".

Sequence reads for which no sequence read corresponding to a complementary strand can be found among the sequence reads are termed "Singlets".

Double-stranded polynucleotides for which a sequence read for neither of the two complementary strands has been generated are referred to as "Unseen" molecules.

I. Quantifying

Methods disclosed herein also comprise quantifying sequence reads. For example, as shown in step (**114**) (FIG. **1**), Pairs and Singlets mapping to a selected genetic locus, or to each of a plurality of selected genetic loci, are quantified, e.g., counted.

The quantifying can comprise estimating number of polynucleotides in the sample (e.g., Pairs polynucleotides, Singlets polynucleotides, or Unseen polynucleotides. For example, as shown in step (**116**) (FIG. **1**), the number of double-stranded polynucleotides in the sample for which no sequence reads were generated ("Unseen" polynucleotides) is estimated. The probability that a double strand polynucleotide generates no sequence reads can be determined based on the relative number of Pairs and Singlets at any locus. Using this probability, the number of Unseen polynucleotide can be estimated.

In step (**118**) an estimate for the total number of double-stranded polynucleotides in a sample mapping to a selected locus is the sum of the number of Pairs, the number of Singlets and the number of Unseen molecules mapping to the locus.

The number of Unseen original molecules in a sample can be estimated based on the relative number of Pairs and Singlets (FIG. **2**). Referring to FIG. **2**, as an example, counts for a particular genomic locus, Locus A, are recorded, where 1000 molecules are paired and 1000 molecules are unpaired. Assuming a uniform probability, p, for an individual Watson or Crick strand to make it through the process subsequent to conversion, one can calculate the proportion of molecules that fail to make it through the process (Unseen) as follows: Let R=ratio of paired to unpaired molecules =1, so $R=1=p^2/(2p(1-p))$. This implies that p=2/3 and that the quantity of lost molecules is equal to $(1-p)^2=1/9$. Thus in this example, approximately 11% of converted molecules are lost and never detected. Consider another genomic locus, Locus B, in the same sample where 1440 molecules are paired and 720 are unpaired. Using the same method, we can infer the number of molecules that are lost, is only 4%. Comparing the two areas, it may be assumed that Locus A had 2000 unique molecules as compared to 2160 molecules in Locus B—a difference of almost 8%. However, by correctly adding in the lost molecules in each region, we infer there are 2000/(8/9)=2250 molecules in Locus A and 2160/0.96=2250 molecules in Locus B. Hence, the counts in both regions are actually equal. This correction and thus much higher sensitivity can be achievable by converting the original double-stranded nucleic acid molecules and bioinformatically keeping track of all those that are paired and unpaired at the end of the process. Similarly, the same procedure can be used to infer true copy number variations in regions that appear to have similar counts of observed unique molecules. By taking the number of unseen molecules into consideration in the two or more regions, the copy number variation becomes apparent.

In addition to using binomial distribution, other methods of estimating numbers of unseen molecules include exponential, beta, gamma or empirical distributions based on the redundancy of sequence reads observed. In the latter case, the distribution of read counts for paired and unpaired molecules can be derived from such redundancy to infer the underlying distribution of original polynucleotide molecules at a particular locus. This can often lead to a better estimation of the number of unseen molecules.

J. CNV Detection

Methods disclosed herein also comprise detecting CNV. For example, as shown in step (**120**) (FIG. **1**), once the total number of polynucleotides mapping to a locus is determined, this number can be used in standard methods of determining CNV at the locus. A quantitative measure can be normalized against a standard. The standard can be an amount of any polynucleotides. In one method, a quantitative measure at a test locus can be standardized against a quantitative measure of polynucleotides mapping to a control locus in the genome, such as gene of known copy number. Quantitative measures can be compared against the amount of nucleic acid in any sample disclosed herein. For example, in another method, the quantitative measure can be compared against the amount of nucleic acid in the original sample. For example, if the original sample contained 10,000 haploid gene equivalents, the quantitative measure can be compared against an expected measure for diploidy. In another method, the quantitative measure can be normalized against a measure from a control sample, and normalized measures at different loci can be compared.

In some cases, in which copy number variation analysis is desired, sequence data may be: 1) aligned with a reference genome; 2) filtered and mapped; 3) partitioned into windows or bins of sequence; 4) coverage reads counted for each window; 5) coverage reads can then be normalized using a stochastic or statistical modeling algorithm; 6) and an output file can be generated reflecting discrete copy number states at various positions in the genome. In other cases, in which rare mutation analysis is desired, sequence data may be 1) aligned with a reference genome; 2) filtered and mapped; 3) frequency of variant bases calculated based on coverage reads for that specific base; 4) variant base frequency normalized using a stochastic, statistical or probabilistic modeling algorithm; 5) and an output file can be generated reflecting mutation states at various positions in the genome.

After the sequence read coverage ratios have been determined, a stochastic modeling algorithm can be optionally applied to convert the normalized ratios for each window region into discrete copy number states. In some cases, this algorithm may comprise a Hidden Markov Model. In other cases, the stochastic model may comprise dynamic programming, support vector machine, Bayesian modeling, probabilistic modeling, trellis decoding, Viterbi decoding, expectation maximization, Kalman filtering methodologies, or neural networks.

Methods disclosed herein can comprise detecting SNVs, CNVs, insertions, deletions, and/or rearrangements at a specific region in a genome. The specific genomic region can comprise a sequence in a gene, such as ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1 IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1,

CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, or NTRK1.

In some cases, the method uses a panel which comprises exons of one or more genes. The panel can comprise introns of one or more genes as well. The panel can also comprise exons and introns of one or more genes. The one or more genes can be those disclosed above. The panel can comprise about 80,000 bases which cover a panel of genes. The panel can comprise about 1000, 2000, 3000, 4000, 5000, 10000, 15000, 20000, 25000, 30000, 35000, 40000, 45000, 50000, 55000, 60000, 65000, 70000, 75000, 80000, 85000, 90000, 95000, 100000, 105000, 110000, 115000, 120000, 125000, or more bases.

In some aspects, copy number of a gene can be reflected in the frequency of a genetic form of the gene in a sample. For example, in a healthy individual, no copy number variation is reflected in a variant in a gene in one chromosome (e.g., heterozygosity) being detected in about 50% of detected molecules in a sample. Also, in a healthy individual, duplication of a gene bearing a variant can be reflected in the variant being detected in about 66% of detected molecules in a sample. Accordingly, if the tumor burden in a DNA sample is 10%, the frequency of a somatic mutation in a gene in one chromosome of cancer cells, without CNV, can be about 5%. The converse can be true in the case of aneuploidy.

The methods disclosed herein can be used to determine whether a sequence variant is more likely present in the germ line level or resulted from a somatic cell mutation, e.g., in a cancer cell. For example, a sequence variant in a gene detected at levels arguably consistent with heterozygosity in the germ line is more likely the product of a somatic mutation if CNV is also detected in that gene. In some cases, to the extent we expect that a gene duplication in the germ line bears a variant consistent with genetic dose (e.g., 66% for trisomy at a locus), detection gene amplification with a sequence variant dose that deviates significantly from this expected amount indicates that the CNV is more likely present as a result of somatic cell mutation.

The methods disclosed herein can also be used to infer tumor heterogeneity in a situation in which sequence variants in two genes are detected at different frequencies. For example, tumor heterogeneity can be inferred when two genes are detected at different frequencies but their copy numbers are relatively equal. Alternatively, tumor homogeneity can be inferred when the difference in frequency between two sequence variants is consistent with difference in copy number for the two genes. Thus, for example, if an EGFR variant is detected at 11% and a KRAS variant is detected at 5%, and no CNV is detected at these genes, the difference in frequency likely reflects tumor heterogeneity (e.g., all tumor cells carry an EGFR mutant and half the tumor cells also carry a KRAS mutant). Alternatively, if the EGFR gene carrying the mutant is detected at 2-times normal copy number, one interpretation is a homogenous population of tumor cells, each cell carrying a mutant in the EGFR and KRAS genes, but in which the KRAS gene is duplicated.

In response to chemotherapy, a dominant tumor form can eventually give way through Darwinian selection to cancer cells carrying mutants that render the cancer unresponsive to the therapy regimen. Appearance of these resistance mutants can be delayed through methods of this invention. In one embodiment of this method, a subject is subjected to one or more pulsed therapy cycles, each pulsed therapy cycle comprising a first period during which a drug is administered at a first amount and a second cycle during which the drug is administered at a second, reduced amount. The first period can be characterized by a tumor burden detected above a first clinical level. The second period can be characterized by a tumor burden detected below a second clinical level. First and second clinical levels can be different in different pulsed therapy cycles. For example, the first clinical level can be lower in succeeding cycles. A plurality of cycles can include at least 2, 3, 4, 5, 6, 7, 8 or more cycles. For example, the BRAF mutant V600E may be detected in polynucleotides of a disease cell at an amount indicating a tumor burden of 5% in cfDNA. Chemotherapy can commence with dabrafenib. Subsequent testing can show that the amount of the BRAF mutant in the cfDNA falls below 0.5% or to undetectable levels. At this point, dabrafenib therapy can stop or be significantly curtailed. Further subsequent testing may find that DNA bearing the BRAF mutation has risen to 2.5% of polynucleotides in cfDNA. At this point, dabrafenib therapy can be re-started, e.g., at the same level as the initial treatment. Subsequent testing may find that DNA bearing the BRAF mutation has decreased to 0.5% of polynucleotides in cfDNA. Again, dabrafenib therapy can be stopped or reduced. The cycle can be repeated a number of times.

A therapeutic intervention can also be changed upon detection of the rise of a mutant form resistant to an original drug. For example, cancers with the EGFR mutation L858R respond to therapy with erlotinib. However, cancers with the EGFR mutation T790M are resistant to erlotinib. However, they are responsive to ruxolitinib. A method of this invention involves monitoring changes in tumor profile and changing a therapeutic intervention when a genetic variant associated with drug resistance rises to a predetermined clinical level.

Methods disclosed in this invention can comprise a method of detecting disease cell heterogeneity from a sample comprising polynucleotides from somatic cells and disease cells, the method comprising: a) quantifying polynucleotides in the sample bearing a sequence variant at each of a plurality of genetic loci; b) determining CNV at each of the plurality of genetic loci; different relative amounts of disease molecules at a locus, wherein the CNV indicates a genetic dose of a locus in the disease cell polynucleotides; c) determining a relative measure of quantity of polynucleotides bearing a sequence variant at a locus per genetic dose at the locus for each of a plurality of the loci; and d) comparing the relative measures at each of the plurality of loci, wherein different relative measures indicates tumor heterogeneity. In the methods disclosed herein, the genetic dose can be determined on a total molecule basis. For example, if there are 1× total molecules at a first locus, and 1.2× molecules mapped to a second locus, then the genetic dose is 1.2. Variants at this locus can be divided by 1.2. In some aspects, the method disclosed herein can be used to detect any disease cell heterogeneity, e.g., tumor cell heterogeneity. The methods can be used to detect disease cell heterogeneity from a sample comprising any types of polynucleotides, e.g., cfDNA, genomic DNA, cDNA, or ctDNA. In the methods, the quantifying can comprise, for example, determining the number or relative amount of the polynucleotides. Determining CNV can comprise mapping and normalizing different relative amounts of total molecules to a locus.

In another aspect, in response to chemotherapy, a dominant tumor form can eventually give way through Darwinian selection to cancer cells carrying mutants that render the cancer unresponsive to the therapy regimen. Appearance of these resistance mutants can be delayed through methods

disclosed throughout. The methods disclosed herein can comprise a method comprising: a) subjecting a subject to one or more pulsed therapy cycles, each pulsed therapy cycle comprising (i) a first period during which a drug is administered at a first amount and (ii) a second period during which the drug is administered at a second, reduced amount; wherein (A) the first period is characterized by a tumor burden detected above a first clinical level; and (B) the second period is characterized by a tumor burden detected below a second clinical level.

K. Sequence Variant Detection

Systems and methods disclosed herein can be used to detect sequence variants, e.g., SNVs. For example, a sequence variant can be detected from consensus sequences from multiple sequence reads, for example, from at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, at least 30, at least 31, at least 32, at least 33, at least 34, at least 35, at least 36, at least 37, at least 38, at least 39, at least 40, at least 41, at least 42, at least 43, at least 44, at least 45, at least 46, at least 47, at least 48, at least 49, at least 50, at least 51, at least 52, at least 53, at least 54, at least 55, at least 56, at least 57, at least 58, at least 59, at least 60, at least 61, at least 62, at least 63, at least 64, at least 65, at least 66, at least 67, at least 68, at least 69, at least 70, at least 71, at least 72, at least 73, at least 74, at least 75, at least 76, at least 77, at least 78, at least 79, at least 80, at least 81, at least 82, at least 83, at least 84, at least 85, at least 86, at least 87, at least 88, at least 89, at least 90, at least 91, at least 92, at least 93, at least 94, at least 95, at least 96, at least 97, at least 98, at least 99, at least 100, at least 200, at least 300, at least 400, at least 500, at least 600, at least 700, at least 800, at least 900, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, at least 8000, at least 9000, at least 10000 or more sequence reads. A consensus sequence can be from sequence reads of a single strand polynucleotide. A consensus sequence can also be from sequence reads of one strand of a double-stranded polynucleotide (e.g., pairing reads). In an exemplary method, pairing reads allows one to identify with increased confidence the existence of a sequence variant in a molecule. For example, if both strands of a Pair include the same variant, one can be reasonably sure that the variant existed in the original molecule, as the chance that the same variant is introduced into both strands during amplification/sequencing is rare. In contrast, if only one strand of a Pair includes the sequence variant, this is more likely to be an artifact. Similarly, the confidence that a Singlet bearing a sequence variant existed in the original molecule is less than the confidence if the variant exists in a Duplex, as there is higher probability that the variant can be introduced once than twice during amplification/sequencing.

Other methods of copy number variation detection and the sequence variant detection are described in PCT/US2013/058061, which is entirely incorporated herein by reference.

Sequence reads can be collapsed to generate a consensus sequence, which can be mapped to a reference sequence to identify genetic variants, such as CNV or SNV. As an alternative, the sequence reads are mapped prior to or even without mapping. In such a case, the sequence reads can be individually mapped to the reference to identify a CNV or SNV.

FIG. **3** shows a reference sequence encoding a genetic Locus A. The polynucleotides in FIG. **3** may be Y-shaped or have other shapes, such as hairpin.

In some cases, an SNV or multiple-nucleotide variant (MNV) can be determined across multiple sequence reads at a given locus (e.g., nucleotide base) by aligning sequence reads that correspond to that locus. Next, a plurality of sequential nucleotide bases from at least a subset of the sequence reads are mapped to the reference to a SNV or MNV in a polynucleotide molecule or portion thereof that corresponds to the reads. The plurality of sequential nucleotide bases can span an actual, inferred or suspected location of the SNV or MNV. The plurality of sequential nucleotide bases can span at least 3, 4, 5, 6, 7, 8, 9, or 10 nucleotide bases.

L. Detecting/Quantifying Nucleic Acids

The methods described throughout can be used to tag nucleic acids fragments, such as deoxyribonucleic acid (DNA), at extremely high efficiency. This efficient tagging allows a person to efficiently and accurately detect rare DNA in heterogenous populations of original DNA fragments (such as in cfDNA). A rare polynucleotide (e.g., rare DNA) can be a polynucleotide that comprises a genetic variant occurring in a population of polynucleotides at a frequency of less than 10%, 5%, 4%, 3%, 2%, 1%, or 0.1%. A rare DNA can be a polynucleotide with a detectable property at a concentration less than 50%, 25%, 10%, 5%, 1%, or 0.1%

Tagging can occur in a single reaction. In some cases, two or more reactions can be performed and pooled together. Tagging each original DNA fragments in a single reaction can result in tagging such that greater than 50% (e.g., 60%, 70%, 80%, 90%, 95%, or 99%) of the original DNA fragments are tagged at both ends with tags that comprise molecular barcodes, thereby providing tagged DNA fragments. Tagging can also result in greater than 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% of the original DNA fragments tagged at both ends with tags that comprise molecular barcodes. Tagging can also result in 100% of the original DNA fragments tagged at both ends with tags that comprise molecular barcodes. Tagging can also result in single end tagging.

Tagging can also occur by using an excess amount of tags as compared to the original DNA fragments. For example, the excess can be at least 5-fold excess. In other cases, the excess can be at least 1.25, 1.5, 1.75, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100 or more fold excess. Tagging can comprise attachment to blunt ends or sticky ends. Tagging can also be performed by hybridization PCR. Tagging can also be performed in low reaction volumes, such as 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or 100 pico- and/or microliters.

The method can also include performing high fidelity amplification on the tagged DNA fragments. Any high

fidelity DNA polymerases can be used. For example, the polymerase may be KAPA HiFi DNA polymerase or Phusion DNA polymerase.

Further, the method can comprise selectively enriching a subset of the tagged DNA fragments. For example, selective enrichment can be performed by hybridization or amplification techniques. The selective enrichment can be performed using a solid support (e.g., beads). The solid support (e.g., beads) can comprise probes (e.g., oligonucleotides specifically hybridizing to certain sequences. For example, the probes can hybridize with certain genomic regions, e.g., genes. In some cases, the genomic regions, e.g., genes, can be regions associated with diseases, e.g., cancer. After enrichment, the selected fragmented can be attached any sequencing adaptor disclosed in this invention. For example, a sequence adaptor can comprise a flow cell sequence, a sample barcode, or both. In another example, a sequence adaptor can be a hairpin shaped adaptor and/or comprises a sample barcode. Further, the resulting fragments can be amplified and sequenced. In some cases, the adaptor does not comprise a sequencing primer region.

The method can include sequencing one or both strands of the DNA fragments. In one case, both strands of the DNA fragment are independently sequenced. The tagged, amplified, and/or selectively enriched DNA fragments are sequenced to obtain sequence reads that comprise sequence information of the molecular barcodes and at least a portion of the original DNA fragments.

The method can include reducing or tracking redundancy (as described above) in the sequence reads to determine consensus reads that are representative of single-strands of the original DNA fragments. For example, to reduce or track redundancy, the method can include comparing sequence reads having the same or similar molecular barcodes and the same or similar end of fragment sequences. The method can comprise performing a phylogentic analysis on the sequence reads having the same or similar molecular barcodes. The molecular barcodes can have a barcode with varying edit distances (including any edit distances as described throughout), for example, an edit distance of up to 3. The end of the fragment sequences can include fragment sequences having an edit distance with varying distances (including any edit distances as described throughout), for example, an edit distance of up to 3.

The method can comprise binning the sequence reads according to the molecular barcodes and sequence information. For example, binning the sequence reads according to the molecular barcodes and sequence information can be performed from at least one end of each of the original DNA fragments to create bins of single stranded reads. The method can further comprise in each bin, determining a sequence of a given original DNA fragment among the original DNA fragments by analyzing sequence reads.

In some cases, sequence reads in each bin can be collapsed to a consensus sequence and subsequently mapped to a genome. As an alternative, sequence reads can be mapped to a genome prior to binning and subsequently collapsed to a consensus sequence.

The method can also comprise sorting sequence reads into paired reads and unpaired reads. After sorting, the number of paired reads and unpaired reads that map to each of one or more genetic loci can be quantified.

The method can include quantifying the consensus reads to detect and/or quantify the rare DNA, which are described throughout. The method can comprise detecting and/or quantifying the rare DNA by comparing a number of times each base occurs at each position of a genome represented by the tagged, amplified, and/or enriched DNA fragments.

The method can comprise tagging the original DNA fragments in a single reaction using a library of tags. The library can include at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 50, at least 100, at least 500, at least 1000, at least 5000, at least 10000, or any number of tags as disclosed throughout. For example, the library of tags can include at least 8 tags. The library of tags can include 8 tags (which can generate 64 different possible combinations). The method can be conducted such that a high percentage of fragments, e.g., greater than 50% (or any percentages as described throughout) are tagged at both ends, wherein each of the tags comprises a molecular barcode.

M. Processing and/or Analyzing Nucleic Acids

The methods described throughout can be used for processing and/or analyzing a nucleic acid sample of a subject. The method can comprising exposing polynucleotide fragments of the nucleic acid sample to a plurality of polynucleotide molecules to yield tagged polynucleotide fragments. The plurality of polynucleotide molecules that can be used are described throughout the application.

For example, the plurality of polynucleotide molecules can be each less than or equal to 40 nucleic acid bases in length and have distinct barcode sequences with respect to at least 4 nucleic acid bases and an edit distance of at least 1, wherein each of the distinct barcode sequences is within 20 nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules, and wherein the plurality of polynucleotide molecules are not sequencing adaptors.

The tagged polynucleotide fragments can be subjected to nucleic acid amplification reactions under conditions that yield amplified polynucleotide fragments as amplification products of the tagged polynucleotide fragments. After amplification, the nucleotide sequence of the amplified tagged polynucleotide fragments is determined. In some cases, the nucleotide sequences of the amplified tagged polynucleotide fragments are determined without the use of polymerase chain reaction (PCR).

The method can comprise analyzing the nucleotide sequences with a programmed computer processor to identify one or more genetic variants in the nucleotide sample of the subject. Any genetic alterations can be identified, including but not limited to, base change(s), insertion(s), repeat(s), deletion(s), copy number variation(s) , epigenetic modification(s), nucleosome binding site(s), copy number change(s) due to origin(s) of replication, and transversion(s). Other genetic alterations can include, but are not limited to, one or more tumor associated genetic alterations.

The subject of the methods can be suspected of having a disease. For example, the subject can be suspected of having cancer. The method can comprise collecting a nucleic acid sample from a subject. The nucleic acid sample can be collected from blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, cerebral spinal fluid, skin, hair, sweat, and/or tears. The nucleic acid sample can be a cell-free nucleic acid sample. In some cases, the nucleic acid sample is collected from no more than 100 nanograms (ng) of double-stranded polynucleotide molecules of the subject.

The polynucleotide fragments can comprise double-stranded polynucleotide molecules. In some cases, the plu-

rality of polynucleotide molecules are coupled to the polynucleotide fragments via blunt end ligation, sticky end ligation, molecular inversion probes, polymerase chain reaction (PCR), ligation-based PCR, multiplex PCR, single strand ligation, or single strand circularization.

The method as described herein results in high efficiency tagging of nucleic acids. For example, exposing the polynucleotide fragments of the nucleic acid sample to the plurality of polynucleotide molecules yields the tagged polynucleotide fragments with a conversion efficiency of at least 30%, e.g., of at least 50% (e.g., 60%, 70%, 80%, 90%, 95%, or 99%). Conversion efficiency of at least 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% can be achieved.

The method can result in a tagged polynucleotide fragment that share common polynucleotide molecules. For example, any of at least 5%, 6%, 7%, 8%, 9%, 10%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% or 100% of the tagged polynucleotide fragments share a common polynucleotide molecule. The method can comprise generating the polynucleotide fragments from the nucleic acid sample.

In some cases, the subjecting of the method comprises amplifying the tagged polynucleotide fragments in the presence primers corresponding to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. Additionally, any combination of these genes can be amplified. For example, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, or all 54 of these genes can be amplified.

The methods described herein can comprise generating a plurality of sequence reads from a plurality of polynucleotide molecules. The plurality of polynucleotide molecules can cover genomic loci of a target genome. For example, the genomic loci can correspond to a plurality of genes as listed above. Further, the genomic loci can be any combination of these genes. Any given genomic locus can comprise at least two nucleic acid bases. Any given genomic locus can also comprise a plurality of nucleic acid bases, for example, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, or more nucleic acid bases.

The method can comprise grouping with a computer processor the plurality of sequence reads into families. Each of the family can comprises sequence reads from one of the template polynucleotides. Each family can comprise sequence reads from only one of the template polynucleotides. For each of the family, the sequence reads can be merged to generate a consensus sequence. The grouping can comprise classifying the plurality of sequence reads into families by identifying (i) distinct molecular barcodes coupled to the plurality of polynucleotide molecules and (ii) similarities between the plurality of sequence reads, wherein each family includes a plurality of nucleic acid sequences that are associated with a distinct combination of molecular barcodes and similar or identical sequence reads.

Once merged, a consensus sequence can be called at a given genomic locus among the genomic loci. At any given genomic loci, any of the following can be determined: i) genetic variants among the calls; ii) frequency of a genetic alteration among the calls; iii) total number of calls; and iv) total number of alterations among the calls. The calling can comprise calling at least one nucleic acid base at the given genomic locus. The calling can also comprise calling a plurality of nucleic acid bases at the given genomic locus. In some cases, the calling can comprise phylogenetic analysis, voting (e.g., biased voting), weighing, assigning a probability to each read at the locus in a family, or calling the base with the highest probability. The consensus sequence can be generated by evaluating a quantitative measure or a statistical significance level for each of the sequence reads. If a quantitative measure is performed, the method can comprise use of a binomial distribution, exponential distribution, beta distribution, or empirical distribution. However, frequency of the base at the particular location can also be used for calling, for example, if 51% or more of the reads is a "A" at the location, then the base may be called an "A" at that particular location. The method can further comprise mapping a consensus sequence to a target genome.

The method can further comprising performing consensus calling at an additional genomic locus among the genomic loci. The method can comprise determining a variation in copy number at one of the given genomic locus and additional genomic locus based on counts at the given genomic locus and additional genomic locus.

The methods described herein can comprise providing template polynucleotide molecules and a library of adaptor polynucleotide molecules in a reaction vessel. The adaptor polynucleotide molecules can have from 2 to 1,000 different barcode sequences and in some cases are not sequencing adaptors. Other variations of adaptor polynucleotide molecules are described throughout, which can also be used in the methods.

The polynucleotide molecules of the adaptors can have the same sample tag. The adaptor polynucleotide molecules can be coupled to both ends of the template polynucleotide molecules. The method can comprise coupling the adaptor polynucleotide molecules to the template polynucleotide molecules at an efficiency of at least 30%, e.g., of at least 50% (e.g., 60%, 70%, 80%, 90%, 95%, or 99%), thereby tagging each template polynucleotide with a tagging combination that is among 4 to 1,000,000 different tagging combinations, to produce tagged polynucleotide molecules. In some cases, the reaction can occur in a single reaction vessel. Coupling efficiency can also be at least 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99%. Tagging can be non-unique tagging.

The tagged polynucleotide molecules can then be subject to an amplification reaction under conditions that will yield amplified polynucleotide molecules as amplification products of the tagged polynucleotide molecules. The template polynucleotide molecules can be double-stranded. Further,

the template polynucleotide molecules can be blunt ended. In some cases, the amplification reaction comprises non-specifically amplifying the tagged polynucleotide molecules. The amplification reaction can also comprises using a priming site to amplify each of the tagged polynucleotide molecules. The priming site can be a primer, e.g., a universal primer. The priming site can also be a nick.

The method can also comprise sequencing the amplified polynucleotide molecules. The sequencing can comprise (i) subjecting the amplified polynucleotide molecules to an additional amplification reaction under conditions that yield additional amplified polynucleotide molecules as amplification products of the amplified polynucleotide molecules, and/or (ii) sequencing the additional amplified polynucleotide molecules. The additional amplification can be performed in the presence of primers comprising flow cells sequences, which will produce polynucleotide molecules that are capable of binding to a flow cell. The additional amplification can also be performed in the presence of primers comprising sequences for hairpin shaped adaptors. The hairpin shaped adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. The method can further comprise identifying genetic variants upon sequencing the amplified polynucleotide molecules.

The method can further comprising separating polynucleotide molecules comprising one or more given sequences from the amplified polynucleotide molecules, to produce enriched polynucleotide molecules. The method can also comprise amplifying the enriched polynucleotide molecules with primers comprising the flow cell sequences. This amplification with primers comprising flow cell sequences will produce polynucleotide molecules that are capable of binding to a flow cell. The amplification can also be performed in the presence of primers comprising sequences for hairpin shaped adaptors. The hairpin shaped adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times.

Flow cell sequences or hairpin shaped adaptors can be added by non-amplification methods such as through ligation of such sequences. Other techniques such as hybridization methods can be used, e.g., nucleotide overhangs.

The method can be performed without aliquoting the tagged polynucleotide molecules. For example, once the tagged polynucleotide molecule is made, the amplification and sequencing can occur in the same tube without any further preparation.

The methods described herein can be useful in detecting single nucleotide variations (SNV), copy number variations (CNV), insertions, deletions, and/or rearrangements. In some cases, the SNVs, CNVs, insertions, deletions, and/or rearrangements, can be associated with disease, for example, cancer.

N. Monitoring a Patient's Status

Methods disclosed herein can also be used to monitor a patient's disease status. The disease of a subject can be monitored over time to determine a progression of the disease (e.g., regression). Markers indicative of the disease can be monitored in a biological sample of the subject, such as a cell-free DNA sample.

For example, monitoring a subject's cancer status can comprise (a) determining an amount of one or more SNVs or copy numbers of a plurality of genes (e.g., in an exon), (b) repeating such determination at different points in time, and (c) determining if there is a difference in the number of SNVs, level of SNVs, number or level of genomic rearrangements, or copy numbers between (a) and (b). The genes can be selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The genes can be selected from any 5, 10, 15, 20, 30, 40, 50, or all of the genes in this group.

O. Sensitivity and Specificity

Methods disclosed herein can be used to detect cancer polynucleotides in a sample, and cancer in a subject, with high measures of agreement, e.g., high sensitivity and/or specificity. For example, such methods can detect cancer polynucleotides (e.g., rare DNA) in a sample at a concentration that is less than 5%, 1%, 0.5%, 0.1%, 0.05%, or 0.01%, at a specificity of at least 99%, 99.9%, 99.99%, 99.999%, 99.9999%, or 99.99999%. Such polynucleotides may be indicative of cancer or other disease. Further, such methods can detect cancer polynucleotides in a sample with a positive predictive value of at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.9%, 99.99%, 99.999%, or 99.9999%.

Subjects identified as positive in a test that are in reality positive are referred as true positives (TP). Subjects identified as positive in a test that are in reality negative are referred as false positives (FP). Subjects identified as negative in a test that are in reality negative are referred as true negatives (TN). Subjects identified as negative in a test that are in reality positive are referred as false negatives (FN). Sensitivity is the percentage of actual positives identified in a test as positive. This includes, for example, instances in which one should have found a cancer genetic variant and did. (Sensitivity=TP/(TP+FN).) Specificity is the percentage of actual negatives identified in a test as negative. This includes, for example, instances in which one should have found no cancer genetic variant and did not. Specificity can be calculated using the following equation: Specificity=TN/(TN+FP). Positive predictive value (PPV) can be measured by the percentage of subjects who test positive that are true positives. PPV can be calculated using the following equation: PPV=TP/(TP+FP). Positive predictive value can be increased by increasing sensitivity (e.g., chance of an actual positive being detected) and/or specificity (e.g., chance of not mistaking an actual negative for a positive).

Low conversion rates of polynucleotides into adaptor-tagged polynucleotides can compromise sensitivity as it decreases the chance of converting, and therefore detecting, rare polynucleotide targets. Noise in a test can compromise specificity as it increases the number of false positives detected in a test. Both low conversion rate and noise compromise positive predictive value as they decrease the percentage of true positives and increase the percentage of false positives.

The methods disclosed herein can achieve high levels of agreement, e.g., sensitivity and specificity, leading to high positive predictive values. Methods of increasing sensitivity

include high efficiency conversion of polynucleotides into adaptor-tagged polynucleotides in a sample. Methods of increasing specificity include reducing sequencing errors, for example, by molecular tracking.

Methods of the present disclosure can be used to detect genetic variation in non-uniquely tagged initial starting genetic material (e.g., rare DNA) at a concentration that is less than 5%, 1%, 0.5%, 0.1%, 0.05%, or 0.01%, at a specificity of at least 99%, 99.9%, 99.99%, 99.999%, 99.9999%, or 99.99999%. In some aspects, the methods can further comprise converting polynucleotides in the initial starting material at an efficiency of at least at least 10%, at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, or at least 90%. Sequence reads of tagged polynucleotides can be subsequently tracked to generate consensus sequences for polynucleotides with an error rate of no more than 2%, 1%, 0.1%, or 0.01%.

2. Pooling Methods

Disclosed herein are methods of detecting copy number variation and/or sequence variants at one or more genetic loci in a test sample. One embodiment is shown in FIG. **8**. Typically, detecting copy number variation involves determining a quantitative measure (e.g., an absolute or relative number) of polynucleotides mapping to a genetic locus of interest in a genome of a test sample, and comparing that number to a quantitative measure of polynucleotides mapping to that locus in a control sample. In certain methods, the quantitative measure is determined by comparing the number of molecules in the test sample that map to a locus of interest with a number of molecules in the test sample mapping to a reference sequence, e.g., a sequence expected to be present at wild type ploidy number. In some examples, the reference sequence is HG19, build 37, or build 38. The comparison could involve, for example, determining a ratio. Then, this measure is compared with a similar measure determined in a control sample. So, for example, if a test sample has a ratio of 1.5:1 for locus of interest versus reference locus, and a control sample has a ratio of 1:1 for the same loci, one may conclude that the test sample exhibits polyploidy at the locus of interest.

When the test sample and the control sample are analyzed separately, the work flow can introduce distortions between final numbers in the control and test samples.

In one method disclosed herein (e.g., flow chart **800**), polynucleotides are provided from a test and a control sample (**802**). Polynucleotides in a test sample and those in a control sample are tagged with tags that identify the polynucleotides as originating from the test or control sample (a source tag). (**804**.) The tag can be, for example, a polynucleotide sequence or barcode that unambiguously identifies the source.

The polynucleotides in each of the control and test samples also can be tagged with identifier tags that will be carried by all amplification progeny of a polynucleotide. Information from start and end sequences of a polynucleotide and identifier tags can identify sequence reads from polynucleotides amplified from an original parent molecule. Each molecule can be uniquely tagged compared with other molecules in the sample. Alternatively, each molecule need not be uniquely tagged compared with other molecules in the sample. That is, the number of different identifier sequences can be fewer than that the number of molecules in sample. By combining identifier information with start/stop sequence information, the probability of confusing two molecules having the same start/stop sequence is significantly diminished.

Number of different identifiers used to tag a nucleic acid (e.g., cfDNA) can dependent on the number of different haploid genome equivalents. Different identifiers can be used to tag at least 2, least 10, least 100, least 200, least 300, least 400, least 500, least 600, least 700, least 800, least 900, least 1,000, least 2,000, least 3,000, least 4,000, least 5,000, least 6,000, least 7,000, least 8,000, least 9,000, least 10,000 or more different haploid genome equivalents. Accordingly, the number of different identifiers used to tag a nucleic acid sample, e.g., cell-free DNA from 500 to 10,000 different haploid genome equivalents and be between any of 1, 2, 3, 4 and 5 and no more than 100, 90, 80, 70, 60, 50, 40 or 30. For example, the number of different identifier used to tag a nucleic acid sample from 500 to 10,000 different haploid genome equivalents can be 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 or less.

Polynucleotides can be tagged by ligation of adaptors comprising the tags or identifiers before amplification. Ligation can be performed using an enzyme, e.g., a ligase. For example, tagging can be performed using a DNA ligase. The DNA ligase can be a T4 DNA ligase, *E. coli* DNA ligase, and/or mammalian ligase. The mammalian ligase can be DNA ligase I, DNA ligase III, or DNA ligase IV. The ligase may also be a thermostable ligase. Tags can be ligated to a blunt-end of a polynucleotide (blunt-end ligation). Alternatively, tags can be ligated to a sticky end of a polynucleotide (sticky-end ligation). The polynucleotides can be tagged by blunt end ligation using adaptors (e.g., adaptors having forked ends). High efficiency of ligation can be achieved using high excess of adaptors (e.g., more than 1.5×, more than 2×, more than 3×, more than 4×, more than 5×, more than 6×, more than 7×, more than 8×, more than 9×, more than 10×, more than 11×, more than 12×, more than 13×, more than 14×, more than 15×, more than 20×, more than 25×, more than 30×, more than 35×, more than 40×, more than 45×, more than 50×, more than 55×, more than 60×, more than 65×, more than 70×, more than 75×, more than 80×, more than 85×, more than 90×, more than 95×, or more than 100).

Once tagged with tags that identify source of the polynucleotides, polynucleotides from different sources (e.g., different samples) can be pooled. After pooling, polynucleotides from different sources (e.g., different samples) can be distinguished by any measurement using the tags, including any process of quantitative measurement. For example, as shown in (**806**) (FIG. **8**), polynucleotides from the control sample and the test sample can be pooled. The pooled molecules can be subject to the sequencing (**808**) and bioinformatic work flow. Both will be subject to the same variations in the process and, therefore, any differential bias is reduced. Because molecules originating from control and test samples are differently tagged, they can be distinguished in any process of quantitative measurement.

The relative amount of control and test sample pooled can be varied. The amount of control sample can be same as the amount of test sample. The amount of control sample can also be larger than the amount of test sample. Alternatively, the amount of control sample can be smaller than the amount of test sample. The smaller the relative amount of one

sample to the total, the fewer identifying tags needed in the original tagging process. A number can be selected to reduce to acceptable levels the probability that two parent molecules having the same start/end sequences will bear the same identifying tag. This probability can be less than 10%, less than 1%, less than 0.1% or less than 0.01%. The probability can be less than 25%, 24%, 23%, 22%, 21%, 20%, 19%, 18%, 17%, 16%, 15%, 14%, 13%, 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1%.

Methods disclosed herein can also comprise grouping sequence reads. For example, bioinformatic workflow can include grouping sequence reads produced from progeny of a single parent molecule, as shown in (**810**) (FIG. **8**). This can involve any of the redundancy reduction methods described herein. Molecules sourced from test and control samples can be differentiated based on source tags they carry (**812**). Molecules mapping to a target locus are quantified for both test-sourced and control-sourced molecules (**812**). This can include the normalization methods discussed herein, e.g., in which numbers at a target locus are normalized against numbers at a reference locus.

Normalized (or raw) quantities at a target locus from test and control samples are compared to determine presence of copy number variation (**814**).

3. Computer Control Systems

The present disclosure provides computer control systems that are programmed to implement methods of the disclosure. FIG. **6** shows a computer system **1501** that is programmed or otherwise configured to implement the methods of the present disclosure. The computer system **1501** can regulate various aspects sample preparation, sequencing and/or analysis. In some examples, the computer system **1501** is configured to perform sample preparation and sample analysis, including nucleic acid sequencing. The computer system **1501** can be an electronic device of a user or a computer system that is remotely located with respect to the electronic device. The electronic device can be a mobile electronic device.

The computer system **1501** includes a central processing unit (CPU, also "processor" and "computer processor" herein) **1505**, which can be a single core or multi core processor, or a plurality of processors for parallel processing. The computer system **1501** also includes memory or memory location **1510** (e.g., random-access memory, read-only memory, flash memory), electronic storage unit **1515** (e.g., hard disk), communication interface **1520** (e.g., network adapter) for communicating with one or more other systems, and peripheral devices **1525**, such as cache, other memory, data storage and/or electronic display adapters. The memory **1510**, storage unit **1515**, interface **1520** and peripheral devices **1525** are in communication with the CPU **1505** through a communication bus (solid lines), such as a motherboard. The storage unit **1515** can be a data storage unit (or data repository) for storing data. The computer system **1501** can be operatively coupled to a computer network ("network") **1530** with the aid of the communication interface **1520**. The network **1530** can be the Internet, an internet and/or extranet, or an intranet and/or extranet that is in communication with the Internet. The network **1530** in some cases is a telecommunication and/or data network. The network **1530** can include one or more computer servers, which can enable distributed computing, such as cloud computing. The network **1530**, in some cases with the aid of the computer system **1501**, can implement a peer-to-peer network, which may enable devices coupled to the computer system **1501** to behave as a client or a server.

The CPU **1505** can execute a sequence of machine-readable instructions, which can be embodied in a program or software. The instructions may be stored in a memory location, such as the memory **1510**. The instructions can be directed to the CPU **1505**, which can subsequently program or otherwise configure the CPU **1505** to implement methods of the present disclosure. Examples of operations performed by the CPU **1505** can include fetch, decode, execute, and writeback.

The CPU **1505** can be part of a circuit, such as an integrated circuit. One or more other components of the system **1501** can be included in the circuit. In some cases, the circuit is an application specific integrated circuit (ASIC).

The storage unit **1515** can store files, such as drivers, libraries and saved programs. The storage unit **1515** can store user data, e.g., user preferences and user programs. The computer system **1501** in some cases can include one or more additional data storage units that are external to the computer system **1501**, such as located on a remote server that is in communication with the computer system **1501** through an intranet or the Internet.

The computer system **1501** can communicate with one or more remote computer systems through the network **1530**. For instance, the computer system **1501** can communicate with a remote computer system of a user (e.g., an operator). Examples of remote computer systems include personal computers (e.g., portable PC), slate or tablet PC's (e.g., Apple® iPad, Samsung® Galaxy Tab), telephones, Smart phones (e.g., Apple® iPhone, Android-enabled device, Blackberry®), or personal digital assistants. The user can access the computer system **1501** via the network **1530**.

Methods as described herein can be implemented by way of machine (e.g., computer processor) executable code stored on an electronic storage location of the computer system **1501**, such as, for example, on the memory **1510** or electronic storage unit **1515**. The machine executable or machine readable code can be provided in the form of software. During use, the code can be executed by the processor **1505**. In some cases, the code can be retrieved from the storage unit **1515** and stored on the memory **1510** for ready access by the processor **1505**. In some situations, the electronic storage unit **1515** can be precluded, and machine-executable instructions are stored on memory **1510**.

The code can be pre-compiled and configured for use with a machine have a processer adapted to execute the code, or can be compiled during runtime. The code can be supplied in a programming language that can be selected to enable the code to execute in a pre-compiled or as-compiled fashion.

Aspects of the systems and methods provided herein, such as the computer system **1501**, can be embodied in programming. Various aspects of the technology may be thought of as "products" or "articles of manufacture" typically in the form of machine (or processor) executable code and/or associated data that is carried on or embodied in a type of machine readable medium. Machine-executable code can be stored on an electronic storage unit, such memory (e.g., read-only memory, random-access memory, flash memory) or a hard disk. "Storage" type media can include any or all of the tangible memory of the computers, processors or the like, or associated modules thereof, such as various semiconductor memories, tape drives, disk drives and the like, which may provide non-transitory storage at any time for the software programming. All or portions of the software may

at times be communicated through the Internet or various other telecommunication networks. Such communications, for example, may enable loading of the software from one computer or processor into another, for example, from a management server or host computer into the computer platform of an application server. Thus, another type of media that may bear the software elements includes optical, electrical and electromagnetic waves, such as used across physical interfaces between local devices, through wired and optical landline networks and over various air-links. The physical elements that carry such waves, such as wired or wireless links, optical links or the like, also may be considered as media bearing the software. As used herein, unless restricted to non-transitory, tangible "storage" media, terms such as computer or machine "readable medium" refer to any medium that participates in providing instructions to a processor for execution.

Hence, a machine readable medium, such as computer-executable code, may take many forms, including but not limited to, a tangible storage medium, a carrier wave medium or physical transmission medium. Non-volatile storage media include, for example, optical or magnetic disks, such as any of the storage devices in any computer(s) or the like, such as may be used to implement the databases, etc. shown in the drawings. Volatile storage media include dynamic memory, such as main memory of such a computer platform. Tangible transmission media include coaxial cables; copper wire and fiber optics, including the wires that comprise a bus within a computer system. Carrier-wave transmission media may take the form of electric or electromagnetic signals, or acoustic or light waves such as those generated during radio frequency (RF) and infrared (IR) data communications. Common forms of computer-readable media therefore include for example: a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD or DVD-ROM, any other optical medium, punch cards paper tape, any other physical storage medium with patterns of holes, a RAM, a ROM, a PROM and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave transporting data or instructions, cables or links transporting such a carrier wave, or any other medium from which a computer may read programming code and/or data. Many of these forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to a processor for execution.

The computer system **1501** can include or be in communication with an electronic display **1535** that comprises a user interface (UI) **1540**. The UI can allow a user to set various conditions for the methods described herein, for example, PCR or sequencing conditions. Examples of UI's include, without limitation, a graphical user interface (GUI) and web-based user interface.

Methods and systems of the present disclosure can be implemented by way of one or more algorithms. An algorithm can be implemented by way of software upon execution by the central processing unit **1505**. The algorithm can, for example, process the reads to generate a consequence sequence.

FIG. **7** schematically illustrates another system for analyzing a sample comprising nucleic acids from a subject. The system includes a sequencer, bioinformatic software and internet connection for report analysis by, for example, a hand held device or a desktop computer

Disclosed herein is a system for analyzing a target nucleic acid molecule of a subject, comprising: a communication interface that receives nucleic acid sequence reads for a plurality of polynucleotide molecules that cover genomic loci of a target genome; computer memory that stores the nucleic acid sequence reads for the plurality of polynucleotide molecules received by the communication interface; and a computer processor operatively coupled to the communication interface and the memory and programmed to (i) group the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides, (ii) for each of the families, merge sequence reads to generate a consensus sequence, (iii) call the consensus sequence at a given genomic locus among the genomic loci, and (iv) detect at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls; and total number of alterations among the calls, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The different variations of each component of the system are described throughout the disclosure within the methods and compositions. These individual components and variations thereof, are also applicable in this system.

4. Kits

Kits comprising the compositions as described herein. The kits can be useful in performing the methods as described herein. Disclosed herein is a kit comprising a plurality of oligonucleotide probes that selectively hybridize to least 5, 6, 7, 8, 9, 10, 20, 30, 40 or all genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The number genes to which the oligonucleotide probes can selectively hybridize can vary. For example, the number of genes can comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, or 54. The kit can include a container that includes the plurality of oligonucleotide probes and instructions for performing any of the methods described herein.

The oligonucleotide probes can selectively hybridize to exon regions of the genes, e.g., of the at least 5 genes. In some cases, the oligonucleotide probes can selectively hybridize to at least 30 exons of the genes, e.g., of the at least 5 genes. In some cases, the multiple probes can selectively hybridize to each of the at least 30 exons. The probes that hybridize to each exon can have sequences that overlap with at least 1 other probe. In some embodiments, the oligoprobes

can selectively hybridize to non-coding regions of genes disclosed herein, for example, intronic regions of the genes. The oligoprobes can also selectively hybridize to regions of genes comprising both exonic and intronic regions of the genes disclosed herein.

Any number of exons can be targeted by the oligonucleotide probes. For example, at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, 275, 280, 285, 290, 295, 300, 400, 500, 600, 700, 800, 900, 1,000, or more, exons can be targeted.

The kit can comprise at least 4, 5, 6, 7, or 8 different library adaptors having distinct molecular barcodes and identical sample barcodes. The library adaptors may not be sequencing adaptors. For example, the library adaptors do not include flow cell sequences or sequences that permit the formation of hairpin loops for sequencing. The different variations and combinations of molecular barcodes and sample barcodes are described throughout, and are applicable to the kit. Further, in some cases, the adaptors are not sequencing adaptors. Additionally, the adaptors provided with the kit can also comprise sequencing adaptors. A sequencing adaptor can comprise a sequence hybridizing to one or more sequencing primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, e.g., a flow cell sequence. For example, a sequencing adaptor can be a flow cell adaptor. The sequencing adaptors can be attached to one or both ends of a polynucleotide fragment. In some cases, the kit can comprise at least 8 different library adaptors having distinct molecular barcodes and identical sample barcodes. The library adaptors may not be sequencing adaptors. The kit can further include a sequencing adaptor having a first sequence that selectively hybridizes to the library adaptors and a second sequence that selectively hybridizes to a flow cell sequence. In another example, a sequencing adaptor can be hairpin shaped. For example, the hairpin shaped adaptor can comprise a complementary double stranded portion and a loop portion, where the double stranded portion can be attached (e.g., ligated) to a double-stranded polynucleotide. Hairpin shaped sequencing adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. A sequencing adaptor can be up to 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, or more bases from end to end. The sequencing adaptor can comprise 20-30, 20-40, 30-50, 30-60, 40-60, 40-70, 50-60, 50-70, bases from end to end. In a particular example, the sequencing adaptor can comprise 20-30 bases from end to end. In another example, the sequencing adaptor can comprise 50-60 bases from end to end. A sequencing adaptor can comprise one or more barcodes. For example, a sequencing adaptor can comprise a sample barcode. The sample barcode can comprise a pre-determined sequence. The sample barcodes can be used to identify the source of the polynucleotides. The sample barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, e.g., at least 8 bases. The barcode can be contiguous or non-contiguous sequences, as described above.

The library adaptors can be blunt ended and Y-shaped and can be less than or equal to 40 nucleic acid bases in length. Other variations of the can be found throughout and are applicable to the kit.

EXAMPLES

Example 1. Methods for Copy Number Variation Detection

Blood Collection

10-30 mL Blood samples are collected at room temperature. The samples are centrifuged to remove cells. Plasma is collected after centrifugation.

cfDNA Extraction

The sample is subjected to proteinase K digestion. DNA is precipitated with isopropanol. DNA is captured on a DNA purification column (e.g., a QIAamp DNA Blood Mini Kit) and eluted in 100 µl solution. DNAs below 500 bp are selected with Ampure SPRI magnetic bead capture (PEG/salt). The resulting production is suspended in 30 µl H_2O. Size distribution is checked (major peak=166 nucleotides; minor peak=330 nucleotides) and quantified. 5 ng of extracted DNA contain approximately 1700 haploid genome equivalents ("HGE"). The general correlation between the amount of DNA and HGE is as follow: 3 pg DNA=1 HGE; 3 ng DNA=1K HGE; 3 µg DNA=1M HGE; 10 pg DNA=3 HE; 10 ng DNA=3K HGE; 10 µg DNA=3M HGE.

"Single Molecule" Library Prep

High-efficiency DNA tagging (>80%) is performed by blunt-end repair and ligation with 8 different octomers (i.e., 64 combinations) with overloaded hairpin adaptors. 2.5 ng DNA (i.e. approximately 800 HGE) is used as the starting material. Each hairpin adaptor comprises a random sequence on its non-complementary portion. Both ends of each DNA fragment are attached with hairpin adaptors. Each tagged fragment can be identified by the random sequence on the hairpin adaptors and a 10 p endogenous sequence on the fragment.

Tagged DNA is amplified by 10 cycles of PCR to produce about 1-7 µg DNAs that contain approximately 500 copies of each of the 800 HGE in the starting material.

Buffer optimization, polymerase optimization and cycle reduction may be performed to optimize the PCR reactions. Amplification bias, e.g., non-specific bias, GC bias, and/or size bias are also reduced by optimization. Noise(s) (e.g., polymerase-introduced errors) are reduced by using high-fidelity polymerases.

The Library may be prepared using Verniata or Sequenom methods.

Sequences may be enriched as follow: DNAs with regions of interest (ROI) are captured using biotin-labeled bead with probe to ROIs. The ROIs are amplified with 12 cycles of PCR to generate a 2000 times amplification. The resulting DNA is then denatured and diluted to 8 pM and loaded into an Illumina sequencer.

Massively Parallel Sequencing

0.1 to 1% of the sample (approximately 100 pg) are used for sequencing. BDigital bioinformatics

Sequence reads are grouped into families, with about 10 sequence reads in each family. Families are collapsed into consensus sequences by voting (e.g., biased voting) each position in a family. A base is called for consensus sequence if 8 or 9 members agree. A base is not called for consensus sequence if no more than 60% of the members agree.

The resulting consensus sequences are mapped to a reference genome. Each base in a consensus sequence is covered by about 3000 different families. A quality score for each sequence is calculated and sequences are filtered based on the their quality scores.

Sequence variation is detected by counting distribution of bases at each locus. If 98% of the reads have the same base (homozygous) and 2% have a different base, the locus is likely to have a sequence variant, presumably from cancer DNA.

CNV is detected by counting the total number of sequences (bases) mapping to a locus and comparing with a control locus. To increase CNV detection, CNV analysis is performed specific regions, including regions on ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, or NTRK1 genes.

Example 2. Method for Correcting Base Calling by Determining the Total Number Unseen Molecules in a Sample

After fragments are amplified and the sequences of amplified fragments are read and aligned, the fragments are subjected to base calling. Variations in the number of amplified fragments and unseen amplified fragments can introduce errors in base calling. These variations are corrected by calculating the number of unseen amplified fragments.

When base calling for locus A (an arbitrary locus), it is first assumed that there are N amplified fragments. The sequence readouts can come from two types of fragments: double-strand fragments and single-strand fragments. The following is a theoretical example of calculating the total number of unseen molecules in a sample.

N is the total number of molecules in the sample.

Assuming 1000 is the number of duplexes detected.

Assuming 500 is the number of single-stranded molecule detected.

P is the probability of seeing a strand.

Q is the probability of not detecting a strand.

Since Q=1−P.

1000=NP(2).

500=N2PQ.

1000/P(2)=N.

500÷2 PQ=N.

1000/P(2)=500÷2PQ.

1000*2 PQ=500 P(2).

2000 PQ=500 P(2).

2000 Q=500 P.

2000 (1−P)=500P

2000−2000 P=500P.

2000=500P+2000 P.

2000=2500 P.

2000÷2500=P.

0.8=P.

1000/P(2)=N.

1000÷0.64=N.

1562=N.

Number of unseen fragments=62.

Example 3. Identification of Genetic Variants in Cancer-Associated Somatic Variants in a Patient

An assay is used to analyze a panel of genes to identify genetic variants in cancer-associated somatic variants with high sensitivity.

Cell-free DNA is extracted from plasma of a patient and amplified by PCR. Genetic variants are analyzed by massively parallel sequencing of the amplified target genes. For one set of genes, all exons are sequenced as such sequencing coverage had shown to have clinically utility (Table 1). For another set of genes, sequencing coverage included those exons with a previously reported somatic mutation (Table 2). The minimum detectable mutant allele (limit of detection) is dependent on the patient's sample cell-free DNA concentration, which varied from less than 10 to over 1,000 genomic equivalents per mL of peripheral blood. Amplification may not be detected in samples with lower amounts of cell-free DNA and/or low-level gene copy amplification. Certain sample or variant characteristics resulted in reduced analytic sensitivity, such as low sample quality or improper collection.

The percentage of genetic variants found in cell-free DNA circulating in blood is related to the unique tumor biology of this patient. Factors that affected the amount/percentages of detected genetic variants in circulating cell-free DNA in blood include tumor growth, turn-over, size, heterogeneity, vascularization, disease progression or treatment. Table 3 annotates the percentage, or allele frequency, of altered circulating cell-free DNA (% cfDNA) detected in this patient. Some of the detected genetic variants are listed in descending order by % cfDNA.

Genetic variants are detected in the circulating cell-free DNA isolated from this patient's blood specimen. These genetic variants are cancer-associated somatic variants, some of which have been associated with either increased or reduced clinical response to specific treatment. "Minor Alterations" are defined as those alterations detected at less than 10% the allele frequency of "Major Alterations". The detected allele frequencies of these alterations (Table 3) and associated treatments for this patient are annotated.

All genes listed in Tables 1 and 2 are analyzed as part of the Guardant360™ test. Amplification is not detected for ERBB2, EGFR, or MET in the circulating cell-free DNA isolated from this patient's blood specimen.

Patient test results comprising the genetic variants are listed in Table 4.

TABLE 1

Genes in which all exons are sequenced

GENES IN WHICH ALL EXONS ARE SEQUENCED			
ALK	<0.1%	APC	<0.1%
AR	<0.1%	BRAF	<0.1%

TABLE 1-continued

Genes in which all exons are sequenced
GENES IN WHICH ALL EXONS ARE SEQUENCED

CDKN2A	<0.1%	EGFR	<0.1%
ERBB2	<0.1%	FBXW7	<0.1%
KRAS	<0.1%	MET	<0.1%
MYC	<0.1%	NOTCH1	<0.1%
NRAS	<0.1%	PIK3CA	<0.1%
PTEN	<0.1%	PROC	<0.1%
RB1	<0.1%	TP53	<0.1%

LOD: Limit of Detection.
The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected.

TABLE 2

Genes in which exons with a previously reported somatic mutation are sequenced
GENES IN WHICH EXONS WITH A PREVIOUSLY REPORTED SOMATIC MUTATION ARE SEQUENCED

ABL1	<0.1%	AKT1	<0.1%
ATM	<0.1%	CDH1	<0.1%
CSF1R	<0.1%	CTNNB1	<0.1%
ERBB4	<0.1%	EZH2	<0.1%
FGFR1	<0.1%	FGFR2	<0.1%
FGFR3	<0.1%	FLT3	<0.1%
GNA11	<0.1%	GNAQ	<0.1%
GNAS	<0.1%	HNF1A	<0.1%
HRAS	<0.1%	IDH1	<0.1%

TABLE 2-continued

Genes in which exons with a previously reported somatic mutation are sequenced
GENES IN WHICH EXONS WITH A PREVIOUSLY REPORTED SOMATIC MUTATION ARE SEQUENCED

IDH2	<0.1%	JAK2	<0.1%
JAK3	<0.1%	KDR	<0.1%
KIT	<0.1%	MLH1	<0.1%
MPL	<0.1%	NPM1	<0.1%
PDGFRA	<0.1%	PTPN11	<0.1%
RET	<0.1%	SMAD4	<0.1%
SMARCB1	<0.1%	SMO	<0.1%
SRC	<0.1%	STK11	<0.1%
TERT	<0.1%	VHL	<0.1%

LOD: Limit of Detection.
The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected.

TABLE 3

Allele frequency of altered circulating cell-free DNA detected in this patient

Gene	cfDNA with alterations (%)	cfDNA without alterations (%)
BRAF V600E	8.9	91.1
NRAS Q61K	6.2	93.8
JAK V617F	1.5	98.6

TABLE 4

Genomic alterations detected in selected genes
Detected: 51 Genomic Alterations

Gene	Chromosome	Position	Mutation (nt)	Mutation (AA)	Percentage	Cosmic ID	DBSNP ID
KRAS	12	25368462	C > T		100.0%		rs4362222
ALK	2	29416572	T > C	I1461V	100.0%		rs1670283
ALK	2	29444095	C > T		100.0%		rs1569156
ALK	2	29543663	T > C	Q500Q	100.0%		rs2293564
ALK	2	29940529	A > T	P234P	100.0%		rs2246745
APC	5	112176756	T > A	V1822D	100.0%		rs459552
CDKN2A	9	21968199	C > G		100.0%	COSM14251	rs11515
FGFR3	4	1807894	G > A	T651T	100.0%		rs7688609
NOTCH1	9	139410424	A > G		100.0%		rs3125006
PDGFRA	4	55141055	A > G	P567P	100.0%		rs1873778
HRAS	11	534242	A > G	H27H	100.0%	COSM249860	rs12628
EGFR	7	55214348	C > T	N158N	99.9%	COSM42978	rs2072454
TP53	17	7579472	G > C	P72R	99.8%		rs1042522
APC	5	112162854	T > C	Y486Y	55.0%		rs2229992
APC	5	112177171	G > A	P1960P	53.8%		rs465899
EGFR	7	55266417	T > C	T903T	53.6%		rs1140475
APC	5	112176325	G > A	G1678G	53.2%		rs42427
APC	5	112176559	T > G	S1756S	53.0%		rs866006
EGFR	7	55229255	G > A	R521K	53.0%		
MET	7	116397572	A > G	Q648Q	52.7%		
APC	5	112175770	G > A	T1493T	52.7%		rs41115
EGFR	7	55249063	G > A	Q787Q	52.6%		rs1050171
NOTCH1	9	139411714	T > C		52.4%		rs11145767
EGFR	7	55238874	T > A	T629T	52.0%		rs2227984
ERBB2	17	37879588	A > G	I655V	51.6%		rs1136201
NOTCH1	9	139397707	G > A	D1698D	51.3%	COSM33747	rs10521
ALK	2	30143499	G > C	L9L	51.0%		rs4358080
APC	5	112164561	G > A	A545A	51.0%		rs351771
FLT3	13	28610183	A > G		50.8%		rs2491231
NOTCH1	9	139418260	A > G	N104N	50.5%		rs4489420
ALK	2	29444076	G > T		50.4%		rs1534545
PIK3CA	3	178917005	A > G		50.3%		rs3729674
NOTCH1	9	139412197	G > A		50.2%		rs9411208
ALK	2	29455267	A > G	G845G	50.0%	COSM148825	rs2256740
KIT	4	55593464	A > C	M541L	49.9%	COSM28026	
NOTCH1	9	139391636	G > A	D2185D	48.9%		rs2229974
PDGFRA	4	55152040	C > T	V824V	48.9%	COSM22413	rs2228230
ALK	2	29416481	T > C	K1491R	48.9%	COSM1130802	rs1881420
ALK	2	29445458	G > T	G1125G	48.6%		rs3795850

TABLE 4-continued

Genomic alterations detected in selected genes Detected: 51 Genomic Alterations							
Gene	Chromosome	Position	Mutation (nt)	Mutation (AA)	Percentage	Cosmic ID	DBSNP ID
NOTCH1	9	139410177	T > C		48.5%		rs3124603
RET	10	43613843	G > T	L769L	48.2%		rs1800861
EGFR	7	55214443	G > A		48.0%		rs7801956
ALK	2	29416366	G > C	D1529E	47.2%		rs1881421
EGFR	7	55238087	C > T		45.5%		rs10258429
RET	10	43615633	C > G	S904S	44.8%		rs1800863
BRAF	7	140453136	A > T	V600E	8.9%	COSM476	
NRAS	1	115256530	G > T	Q61K	6.2%	COSM580	rs121913254
JAK2	9	5073770	G > T	V617F	1.5%	COSM12600	rs77375493

Example 4. Determining Patient-Specific Limits of Detection for Genes Analyzed by Guardant360™ Assays

Using the method of Example 3, Genetic alterations in cell-free DNA of a patient are detected. The sequence reads of these genes include exon and/or intron sequences.

Limits of detection of the test are shown in Table 5. The limits of detection values are dependent on cell-free DNA concentration and sequencing coverage for each gene.

TABLE 5

Limits of Detection of selected genes in a patient using Guardant					
Complete Exon and Partial Intron Coverage					
APC	0.1%	AR*	0.2%	ARID1A	
BRAF*	0.1%	BRCA1		BRCA2	
CCND1*		CCND2*		CCNE1*	
CDK4*		CDK6*		CDKN2A	0.1%
CDKN2B		EGFR*	<0.1%	ERBB2*	0.1%
FGFR1*	<0.1%	FGFR2*	0.1%	HRAS	0.1%
KIT*	0.1%	KRAS*	0.1%	MET*	0.1%
MYC*	0.1%	NF1		NRAS	0.1%
PDGFRA*	0.1%	PIK3CA*	0.1%	PTEN	0.1%
RAF1*		TP53	0.1%		
Exons Covered with Reported Somatic Mutations					
AKT1	0.1%	ALK	<0.1%	ARAF	
ATM	0.1%	CDH1	0.1%	CTNNB1	0.1%
ESR1		EZH2	0.1%	FBXW7	0.1%
FGFR3	0.1%	GATA3		GNA11	0.1%
GNAQ	0.1%	GNAS	0.1%	HNF1A	0.1%
IDH1	0.1%	IDH2	0.1%	JAK2	0.1%
JAK3	0.1%	MAP2K1		MAP2K2	
MLH1	0.1%	MPL	0.2%	NFE2L2	
NOTCH1	0.1%	NPM1	0.1%	PTPN11	0.1%
RET	0.1%	RHEB		RHOA	
RIT1		ROS1		SMAD4	0.1%
SMO	0.1%	SRC	<0.1%	STK11	0.2%
TERT	0.1%	VHL	0.2%		
Fusions					
ALK	<0.1%	RET	0.1%	ROS1	
NTRK1					

LOD: Limit of Detection.

The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected.

*indicates CNV genes.

Example 5. Correcting Sequence Errors Comparing Watson and Crick Sequences

Double-stranded cell-free DNA is isolated from the plasma of a patient. The cell-free DNA fragments are tagged using 16 different bubble-containing adaptors, each of which comprises a distinctive barcode. The bubble-containing adaptors are attached to both ends of each cell-free DNA fragment by ligation. After ligation, each of the cell-free DNA fragment can be distinctly identified by the sequence of the distinct barcodes and two 20 bp endogenous sequences at each end of the cell-free DNA fragment.

The tagged cell-free DNA fragments are amplified by PCR. The amplified fragments are enriched using beads comprising oligonucleotide probes that specifically bind to a group of cancer-associated genes. Therefore, cell-free DNA fragments from the group of cancer-associated genes are selectively enriched.

Sequencing adaptors, each of which comprises a sequencing primer binding site, a sample barcode, and a cell-flow sequence, are attached to the enriched DNA molecules. The resulting molecules are amplified by PCR.

Both strands of the amplified fragments are sequenced. Because each bubble-containing adaptor comprises a non-complementary portion (e.g., the bubble), the sequence of the one strand of the bubble-containing adaptor is different from the sequence of the other strand (complement). Therefore, the sequence reads of amplicons derived from the Watson strand of an original cell-free DNA can be distinguished from amplicons from the Crick strand of the original cell-free DNA by the attached bubble-containing adaptor sequences.

The sequence reads from a strand of an original cell-free DNA fragment are compared to the sequence reads from the other strand of the original cell-free DNA fragment. If a variant occurs in only the sequence reads from one strand, but not other strand, of the original cell-free DNA fragment, this variant will be identified as an error (e.g., resulted from PCR and/or amplification), rather than a true genetic variant.

The sequence reads are grouped into families. Errors in the sequence reads are corrected. The consensus sequence of each family is generated by collapsing.

While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. It is not intended that the invention be limited by the specific examples provided within the specification. While the invention has been described with reference to the aforementioned specification, the descriptions and illustrations of the embodiments herein are not meant to be construed in a limiting sense. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. Furthermore, it shall be understood that all aspects of the invention are not limited to the specific depictions, configurations or relative proportions set forth herein which depend upon a variety of conditions and variables. It should be understood that various alternatives to the embodiments of

the invention described herein may be employed in practicing the invention. It is therefore contemplated that the invention shall also cover any such alternatives, modifications, variations or equivalents. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

What is claimed is:

1. A method, comprising:

(a) providing a population of cell-free deoxyribonucleic acid (cfDNA) molecules having first and second complementary strands;

(b) tagging a plurality of the cfDNA molecules of the population with a set of duplex tags comprising molecular barcodes from a set of molecular barcodes to produce tagged parent polynucleotides, wherein duplex tags from the set of duplex tags are attached at both ends of a molecule of the plurality of the cfDNA molecules;

(c) amplifying a plurality of the tagged parent polynucleotides to produce amplified progeny polynucleotides;

(d) sequencing at least a subset of the amplified progeny polynucleotides to produce a set of sequence reads; and

(e) reducing or tracking redundancy in the set of sequence reads using at least sequence information from the molecular barcodes to generate a plurality of consensus sequences representative of original cfDNA molecules from among the tagged parent polynucleotides, wherein the plurality of consensus sequences is generated from (i) paired reads corresponding to sequence reads generated from a first tagged strand and a second tagged complementary strand derived from a cfDNA molecule from among the tagged parent polynucleotides, and (ii) unpaired reads corresponding to sequence reads generated from a first tagged strand having no second tagged complementary strand derived from a cfDNA molecule from among the tagged parent polynucleotides.

2. The method of claim **1**, wherein the population of cfDNA molecules is obtained or derived from a sample from a subject having cancer.

3. The method of claim **1**, wherein the plurality of cfDNA molecules comprises between 1 nanogram (ng) and 100 ng of cfDNA molecules.

4. The method of claim **1**, wherein the tagging comprises ligating the duplex tags to the plurality of the cfDNA molecules using more than a 10× excess of duplex tags as compared to the population of cfDNA molecules.

5. The method of claim **4**, wherein at least 20% of the cfDNA molecules of the population are tagged with the duplex tags.

6. The method of claim **1**, wherein the tagging comprises non-uniquely tagging the plurality of the cfDNA molecules with the set of duplex tags comprising molecular barcodes from the set of molecular barcodes, wherein the cfDNA molecules that map to a mappable base position of a reference sequence are tagged with a number of different molecular barcodes ranging from at least 2 to fewer than a number of the cfDNA molecules that map to the mappable base position.

7. The method of claim **1**, wherein the molecular barcodes of the set of molecular barcodes have pre-determined sequences.

8. The method of claim **1**, wherein the molecular barcodes of the set of molecular barcodes have between 5 and 10,000 different molecular barcode sequences and have a length of between 5 and 20 base pairs.

9. The method of claim **1**, further comprising enriching a plurality of the amplified progeny polynucleotides for target regions of interest prior to the sequencing to produce enriched progeny polynucleotides.

10. The method of claim **9**, wherein the target regions of interest comprise genetic sequences of a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

11. The method of claim **9**, further comprising amplifying a plurality of the enriched progeny polynucleotides prior to the sequencing.

12. The method of claim **1**, wherein the duplex tags of the set of duplex tags are part of sequencing adapters.

13. The method of claim **12**, wherein the sequencing adapters are Y-shaped adapters.

14. The method of claim **1**, wherein the reducing or tracking redundancy in the set of sequence reads comprises mapping a plurality of the set of sequence reads to a reference sequence.

15. The method of claim **1**, further comprising:

(f) determining quantitative measures of at least two of (i) the paired reads, (ii) the unpaired reads, (iii) read depth of the paired reads, and (iv) read depth of the unpaired reads at one or more loci of a reference sequence.

16. The method of claim **15**, further comprising:

(g) estimating with a programmed computer processor a quantitative measure of tagged parent polynucleotides based at least in part on the quantitative measures of the at least two of (i) the paired reads, (ii) the unpaired reads, (iii) the read depth of the paired reads, and (iv) the read depth of the unpaired reads at the one or more loci.

17. The method of claim **16**, wherein (f) comprises determining quantitative measures of the paired reads and the unpaired reads, and wherein in (g), the quantitative measure of the tagged parent polynucleotides is determined based at least in part on the quantitative measures of the paired reads and the unpaired reads.

18. A method, comprising:

(a) providing a population of double-stranded cell-free deoxyribonucleic acid (cfDNA) molecules having first and second complementary strands;

(b) non-uniquely tagging a plurality of the double-stranded cfDNA molecules of the population with a set of duplex tags comprising molecular barcodes from a set of molecular barcodes to produce non-uniquely tagged parent polynucleotides,

wherein the double-stranded cfDNA molecules that map to a mappable base position of a reference sequence are tagged with a number of different molecular barcodes ranging from at least 2 to fewer than a number of the double-stranded cfDNA molecules that map to the mappable base position;

(c) amplifying a plurality of the non-uniquely tagged parent polynucleotides to produce amplified progeny polynucleotides;

(d) sequencing at least a subset of the amplified progeny polynucleotides to produce a set of sequence reads;

(e) reducing or tracking redundancy in the set of sequence reads using at least sequence information from the molecular barcodes;

(f) sorting the set of sequence reads into paired reads and unpaired reads, wherein (i) a paired read corresponds to sequence reads generated from a first tagged strand and a second tagged complementary strand derived from a double-stranded cfDNA molecule from among the non-uniquely tagged parent polynucleotides, and (ii) an unpaired read corresponds to sequence reads generated from a first tagged strand having no second tagged complementary strand derived from a double-stranded cfDNA molecule from among the non-uniquely tagged parent polynucleotides; and

(g) determining, at one or more loci of a reference sequence, quantitative measures of at least two of (i) the paired reads, (ii) the unpaired reads, (iii) read depth of the paired reads, and (iv) read depth of the unpaired reads.

19. The method of claim **18**, wherein the population of double-stranded cfDNA molecules is obtained or derived from a sample from a subject having cancer.

20. The method of claim **18**, wherein the plurality of double-stranded cfDNA molecules comprises between 1 nanogram (ng) and 100 ng of double-stranded cfDNA molecules.

21. The method of claim **18**, wherein the non-uniquely tagging comprises ligating the duplex tags to the plurality of the double-stranded cfDNA molecules.

22. The method of claim **18**, wherein the molecular barcodes of the set of molecular barcodes have between 2 and 10,000 different molecular barcode sequences.

23. The method of claim **18**, wherein the molecular barcodes of the set of molecular barcodes have between 5 and 10,000 different molecular barcode sequences and have a length of between 5 and 20 base pairs in length.

24. The method of claim **18**, further comprising enriching a plurality of the amplified progeny polynucleotides for target regions of interest prior to the sequencing to produce enriched progeny polynucleotides.

25. The method of claim **24**, wherein the target regions of interest comprise genetic sequences of a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

26. The method of claim **24**, further comprising amplifying a plurality of the enriched progeny polynucleotides prior to the sequencing.

27. The method of claim **18**, wherein reducing or tracking the redundancy in the set of sequence reads comprises collapsing a plurality of the set of sequence reads to generate consensus sequences representative of original double-stranded cfDNA molecules from among the non-uniquely tagged parent polynucleotides.

28. The method of claim **27**, further comprising mapping a plurality of the set of sequence reads or the consensus sequences to a reference sequence.

29. The method of claim **18**, further comprising:

(h) estimating with a programmed computer processor a quantitative measure of non-uniquely tagged parent polynucleotides based at least in part on the quantitative measures of the at least two of (i) the paired reads, (ii) the unpaired reads, (iii) the read depth of the paired reads, and (iv) the read depth of the unpaired reads at the one or more loci.

30. The method of claim **29**, wherein (g) comprises determining quantitative measures of the paired reads and the unpaired reads, and wherein in (h), the quantitative measure of the non-uniquely tagged parent polynucleotides is determined based at least in part on the quantitative measures of the paired reads and the unpaired reads.

* * * * *

EXHIBIT D

US011149306B2

(12) **United States Patent**
Talasaz et al.

(10) **Patent No.: US 11,149,306 B2**
(45) **Date of Patent: *Oct. 19, 2021**

(54) **METHODS AND SYSTEMS FOR DETECTING GENETIC VARIANTS**

(71) Applicant: **GUARDANT HEALTH, INC.**, Redwood City, CA (US)

(72) Inventors: **AmirAli Talasaz**, Atherton, CA (US); **Helmy Eltoukhy**, Atherton, CA (US); **Stefanie Ann Ward Mortimer**, Morgan Hill, CA (US)

(73) Assignee: **GUARDANT HEALTH, INC.**, Redwood City, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **16/945,124**

(22) Filed: **Jul. 31, 2020**

(65) **Prior Publication Data**

US 2020/0362405 A1 Nov. 19, 2020

Related U.S. Application Data

(63) Continuation of application No. 16/601,168, filed on Oct. 14, 2019, now Pat. No. 10,801,063, which is a continuation of application No. 15/892,178, filed on Feb. 8, 2018, now Pat. No. 10,883,139, which is a continuation of application No. 14/861,989, filed on Sep. 22, 2015, now Pat. No. 9,920,366, which is a continuation of application No. PCT/US2014/072383, filed on Dec. 24, 2014.

(60) Provisional application No. 61/948,509, filed on Mar. 5, 2014, provisional application No. 61/921,456, filed on Dec. 28, 2013.

(51) **Int. Cl.**

C12P 19/34	(2006.01)
C12Q 1/6869	(2018.01)
C12Q 1/6886	(2018.01)
G16B 15/00	(2019.01)

(52) **U.S. Cl.**

CPC ***C12Q 1/6869*** (2013.01); ***C12Q 1/6886*** (2013.01); *C12Q 2535/122* (2013.01); *C12Q 2600/158* (2013.01); *G16B 15/00* (2019.02)

(58) **Field of Classification Search**

CPC C12Q 1/6869

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,725,536	A	2/1988	Fritsch et al.
4,942,124	A	7/1990	Church
5,124,246	A	6/1992	Urdea et al.
5,149,625	A	9/1992	Church et al.
5,200,314	A	4/1993	Urdea
5,424,186	A	6/1995	Fodor et al.
5,424,413	A	6/1995	Hogan et al.
5,445,934	A	8/1995	Fodor et al.
5,604,097	A	2/1997	Brenner
5,635,352	A	6/1997	Urdea et al.
5,635,400	A	6/1997	Brenner
5,648,245	A	7/1997	Fire et al.
5,654,413	A	8/1997	Brenner
5,656,731	A	8/1997	Urdea
5,658,737	A	8/1997	Nelson et al.
5,714,330	A	2/1998	Brenner et al.
5,744,305	A	4/1998	Fodor et al.
5,759,778	A	6/1998	Li et al.
5,763,175	A	6/1998	Brenner
5,800,992	A	9/1998	Fodor et al.
5,846,719	A	12/1998	Brenner et al.
5,854,033	A	12/1998	Lizardi
5,871,928	A	2/1999	Fodor et al.
5,925,525	A	7/1999	Fodor et al.
5,935,793	A	8/1999	Wong
5,952,170	A	9/1999	Stroun et al.
5,968,740	A	10/1999	Fodor et al.
5,981,176	A	11/1999	Wallace
5,981,179	A	11/1999	Lorinez et al.
6,013,445	A	1/2000	Albrecht et al.
6,020,124	A	2/2000	Sorenson
6,040,138	A	3/2000	Lockhart et al.

(Continued)

FOREIGN PATENT DOCUMENTS

CN	102933721	A	2/2013
EP	0799897	B1	6/2006

(Continued)

OTHER PUBLICATIONS

Guardant Health vs. *FMI* Invalidity Exhibit C-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit C-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit C-4, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit C-5, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit D-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).
Guardant Health vs. *FMI* Invalidity Exhibit D-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

(Continued)

Primary Examiner — Kenneth R Horlick
(74) *Attorney, Agent, or Firm* — Timothy A. Hott

(57) **ABSTRACT**

Disclosed herein in are methods and systems for determining genetic variants (e.g., copy number variation) in a polynucleotide sample. A method for determining copy number variations includes tagging double-stranded polynucleotides with duplex tags, sequencing polynucleotides from the sample and estimating total number of polynucleotides mapping to selected genetic loci. The estimate of total number of polynucleotides can involve estimating the number of double-stranded polynucleotides in the original sample for which no sequence reads are generated. This number can be generated using the number of polynucleotides for which reads for both complementary strands are detected and reads for which only one of the two complementary strands is detected.

29 Claims, 11 Drawing Sheets

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,046,005 A 4/2000 Ju et al.
6,060,596 A 5/2000 Lerner et al.
6,117,631 A 9/2000 Nilsen
6,124,092 A 9/2000 O'Neill et al.
6,138,077 A 10/2000 Brenner
6,140,489 A 10/2000 Brenner
6,172,214 B1 1/2001 Brenner
6,197,506 B1 3/2001 Fodor et al.
6,235,475 B1 5/2001 Brenner et al.
6,242,186 B1 6/2001 Salonen
6,268,152 B1 7/2001 Fodor et al.
6,284,460 B1 9/2001 Fodor et al.
6,300,077 B1 10/2001 Shuber et al.
6,309,822 B1 10/2001 Fodor et al.
6,309,823 B1 10/2001 Cronin et al.
6,326,148 B1 12/2001 Pauletti et al.
6,355,431 B1 3/2002 Chee et al.
6,355,432 B1 3/2002 Fodor et al.
6,395,491 B1 5/2002 Fodor et al.
6,406,848 B1 6/2002 Bridgham et al.
6,440,667 B1 8/2002 Fodor et al.
6,440,706 B1 8/2002 Vogelstein et al.
6,451,536 B1 9/2002 Fodor et al.
6,458,530 B1 10/2002 Morris et al.
6,468,744 B1 10/2002 Cronin et al.
6,489,114 B2 12/2002 Laayoun et al.
6,492,121 B2 12/2002 Kurane et al.
6,498,012 B2 12/2002 Laken
6,503,718 B2 1/2003 Shuber et al.
6,512,105 B1 1/2003 Hogan et al.
6,514,699 B1 2/2003 O'Neill et al.
6,544,739 B1 4/2003 Fodor et al.
6,551,784 B2 4/2003 Fodor et al.
6,576,424 B2 6/2003 Fodor et al.
6,582,908 B2 6/2003 Fodor et al.
6,586,177 B1 7/2003 Shuber
6,600,996 B2 7/2003 Webster et al.
6,629,040 B1 9/2003 Goodlett et al.
6,653,077 B1 11/2003 Brenner
6,753,147 B2 6/2004 Vogelstein et al.
6,849,403 B1 2/2005 Shuber
6,849,404 B2 2/2005 Park et al.
6,852,488 B2 2/2005 Fodor et al.
6,858,412 B2 2/2005 Willis et al.
6,964,846 B1 11/2005 Shuber
7,163,789 B2 1/2007 Chen et al.
7,208,275 B2 4/2007 Gocke et al.
7,406,385 B2 7/2008 Sorenson
7,410,764 B2 8/2008 Gocke et al.
7,424,368 B2 9/2008 Huang et al.
7,424,371 B2 9/2008 Kamentsky
7,537,897 B2 5/2009 Brenner et al.
7,700,286 B2 4/2010 Stroun et al.
7,704,687 B2 4/2010 Wang et al.
7,727,720 B2 6/2010 Dhallan
7,803,929 B2 9/2010 Melkonyan et al.
7,811,757 B2 10/2010 Shuber
7,822,555 B2 10/2010 Huang et al.
7,824,889 B2 11/2010 Vogelstein et al.
7,838,647 B2 11/2010 Hahn et al.
7,915,015 B2 3/2011 Vogelstein et al.
7,935,487 B2 5/2011 Gocke et al.
7,937,225 B2 5/2011 Mishra et al.
7,957,913 B2 6/2011 Chinitz et al.
7,972,817 B2 7/2011 Kopreski
7,981,612 B2 7/2011 Shuber et al.
8,094,312 B2 1/2012 Ulmer
8,168,385 B2 5/2012 Brenner
8,195,415 B2 6/2012 Fan et al.
8,209,130 B1 6/2012 Kennedy et al.
8,216,789 B2 7/2012 Disis et al.
8,236,532 B2 8/2012 Ronaghi et al.
8,361,726 B2 1/2013 Gocke et al.
8,383,338 B2 2/2013 Kitzman et al.
8,383,345 B2 2/2013 Shendure et al.
8,455,193 B2 6/2013 Travers et al.
8,470,996 B2 6/2013 Brenner
8,481,292 B2 7/2013 Casbon et al.
8,603,749 B2 12/2013 Gillevet
8,685,678 B2 4/2014 Casbon et al.
8,697,408 B2 4/2014 Kucera et al.
8,704,165 B2 4/2014 Huang
8,715,967 B2 5/2014 Casbon et al.
8,722,368 B2 5/2014 Casbon et al.
8,728,766 B2 5/2014 Casbon et al.
8,741,606 B2 6/2014 Casbon et al.
8,775,092 B2 7/2014 Colwell et al.
8,835,358 B2 9/2014 Fodor et al.
8,865,410 B2 10/2014 Shendure et al.
9,018,365 B2 4/2015 Brenner
9,085,798 B2 7/2015 Chee
9,260,753 B2 2/2016 Xie et al.
9,340,830 B2 5/2016 Lipson et al.
9,376,719 B2 6/2016 Eijk et al.
9,404,156 B2 8/2016 Hicks et al.
9,476,095 B2 10/2016 Vogelstein et al.
9,598,731 B2 3/2017 Talasaz
9,752,188 B2 9/2017 Schmitt et al.
9,834,822 B2 12/2017 Talasaz
9,840,743 B2 12/2017 Talasaz
9,920,366 B2 3/2018 Eltoukhy et al.
10,287,631 B2 5/2019 Salk et al.
10,370,713 B2 8/2019 Salk et al.
10,385,393 B2 8/2019 Salk et al.
10,388,403 B2 8/2019 Rava et al.
10,604,804 B2 3/2020 Salk et al.
10,752,951 B2 8/2020 Salk et al.
2001/0053519 A1 12/2001 Fodor et al.
2002/0072058 A1 6/2002 Voelker et al.
2003/0003490 A1 1/2003 Fan et al.
2003/0049616 A1 3/2003 Brenner et al.
2003/0104436 A1 6/2003 Morris et al.
2003/0152490 A1 8/2003 Trulson et al.
2003/0165978 A1 9/2003 Firth et al.
2003/0186251 A1 10/2003 Dunn et al.
2003/0207300 A1 11/2003 Matray et al.
2004/0096892 A1 5/2004 Wang et al.
2004/0146901 A1 7/2004 Morris et al.
2004/0157243 A1 8/2004 Huang et al.
2004/0209299 A1 10/2004 Pinter et al.
2004/0259118 A1 12/2004 Macevicz
2005/0164241 A1 7/2005 Hahn et al.
2005/0221314 A1 10/2005 Berlin et al.
2005/0250147 A1 11/2005 Macevicz
2006/0035258 A1 2/2006 Tadakamalla et al.
2006/0046258 A1 3/2006 Lapidus et al.
2006/0073506 A1 4/2006 Christians et al.
2006/0211030 A1 9/2006 Brenner
2007/0020640 A1 1/2007 McCloskey et al.
2007/0065823 A1 3/2007 Dressman et al.
2007/0111233 A1 5/2007 Bianchi et al.
2007/0128724 A1 6/2007 Miles et al.
2007/0172839 A1 7/2007 Smith et al.
2007/0172873 A1 7/2007 Brenner et al.
2008/0014146 A1 1/2008 Hoff et al.
2008/0124721 A1 5/2008 Fuchs et al.
2008/0161420 A1 7/2008 Shuber
2008/0261204 A1 10/2008 Lexow
2008/0293055 A1 11/2008 Freeman et al.
2009/0029377 A1 1/2009 Lo et al.
2009/0087847 A1 4/2009 Lo et al.
2009/0098547 A1 4/2009 Ghosh
2009/0105959 A1 4/2009 Braverman et al.
2009/0162836 A1 6/2009 Widschwendter
2009/0298075 A1 12/2009 Travers et al.
2009/0298709 A1 12/2009 Ma
2009/0317818 A1 12/2009 Ehrich et al.
2010/0041048 A1 2/2010 Diehl et al.
2010/0062494 A1 3/2010 Church et al.
2010/0069250 A1 3/2010 White et al.
2010/0112575 A1 5/2010 Fan et al.
2010/0143932 A1 6/2010 Lapidus
2010/0166744 A1 7/2010 Wong
2010/0196898 A1 8/2010 Sugarbaker et al.

(56) **References Cited**

U.S. PATENT DOCUMENTS

2010/0264331 A1 10/2010 Sacko et al.
2010/0323348 A1 12/2010 Hamady et al.
2010/0330571 A1 12/2010 Robins et al.
2011/0014607 A1 1/2011 Jirtle et al.
2011/0160078 A1 6/2011 Fodor et al.
2011/0160290 A1 6/2011 Tewari
2011/0171640 A1 7/2011 Bhatt et al.
2011/0177512 A1 7/2011 Shuber
2011/0183330 A1 7/2011 Lo et al.
2011/0201507 A1 8/2011 Rava et al.
2011/0230358 A1 9/2011 Rava
2011/0230360 A1 9/2011 Stephan et al.
2011/0245482 A1 10/2011 Hahn et al.
2011/0264376 A1 10/2011 Chinitz et al.
2011/0275084 A1 11/2011 Byron et al.
2011/0281736 A1 11/2011 Drmanac et al.
2011/0319272 A1 12/2011 Fan et al.
2012/0003637 A1 1/2012 Lo et al.
2012/0034685 A1 2/2012 Sparks et al.
2012/0046877 A1 2/2012 Hyland et al.
2012/0053073 A1 3/2012 Kassis
2012/0059594 A1 3/2012 Hatchwell et al.
2012/0065081 A1 3/2012 Chee
2012/0095697 A1 4/2012 Halpern et al.
2012/0100548 A1 4/2012 Rava et al.
2012/0164630 A1 6/2012 Porreca et al.
2012/0165202 A1 6/2012 Porreca et al.
2012/0191367 A1 7/2012 Stuelpnagel et al.
2012/0208706 A1 8/2012 Downing et al.
2012/0214163 A1 8/2012 Sugarbaker et al.
2012/0214678 A1 8/2012 Rava et al.
2012/0220478 A1 8/2012 Shaffer
2012/0231479 A1 9/2012 Puskas et al.
2012/0238464 A1 9/2012 Koi et al.
2012/0270212 A1 10/2012 Rabinowitz et al.
2012/0316074 A1 12/2012 Saxonov
2012/0322681 A1 12/2012 Kung et al.
2013/0005585 A1 1/2013 Anderson et al.
2013/0017958 A1 1/2013 Benz et al.
2013/0022977 A1 1/2013 Lapidus et al.
2013/0029852 A1 1/2013 Rava et al.
2013/0034546 A1 2/2013 Rava et al.
2013/0040824 A1 2/2013 Lo et al.
2013/0053256 A1 2/2013 Hubbell
2013/0060483 A1 3/2013 Struble et al.
2013/0078626 A1 3/2013 Wasserstrom et al.
2013/0085681 A1 4/2013 Deciu et al.
2013/0096011 A1 4/2013 Rava et al.
2013/0102485 A1 4/2013 Lee
2013/0102487 A1 4/2013 Cos et al.
2013/0116127 A1 5/2013 Schuetz et al.
2013/0116130 A1 5/2013 Fu et al.
2013/0122499 A1 5/2013 Morris et al.
2013/0130923 A1 5/2013 Ehrich et al.
2013/0137588 A1 5/2013 Shendure et al.
2013/0143747 A1 6/2013 Gutin et al.
2013/0210643 A1 8/2013 Casbon et al.
2013/0210645 A1 8/2013 Volgelstein et al.
2013/0224743 A1 8/2013 Casbon et al.
2013/0237431 A1 9/2013 Lo et al.
2013/0237458 A1 9/2013 Casbon et al.
2013/0260381 A1 10/2013 Ramakrishnan
2013/0267424 A1 10/2013 Casbon et al.
2013/0288244 A1 10/2013 Deciu et al.
2013/0338933 A1 12/2013 Deciu et al.
2014/0011694 A1 1/2014 Couronne
2014/0065609 A1 3/2014 Hicks et al.
2014/0065630 A1 3/2014 Bubnoff et al.
2014/0066317 A1 3/2014 Talasaz
2014/0100121 A1 4/2014 Lo et al.
2014/0242588 A1 8/2014 Boom et al.
2014/0274731 A1 9/2014 Raymond et al.
2014/0274740 A1 9/2014 Srinivasan et al.
2014/0296081 A1 10/2014 Diehn et al.
2014/0296094 A1 10/2014 Domanus
2014/0303008 A1 10/2014 Schutz et al.
2014/0336943 A1 11/2014 Pellini et al.
2014/0350130 A1 11/2014 Sanborn et al.
2014/0357500 A1 12/2014 Vigneault et al.
2015/0004158 A1 1/2015 Shipp et al.
2015/0024950 A1 1/2015 Bielas et al.
2015/0044687 A1 2/2015 Schmitt et al.
2015/0050647 A1 2/2015 Luo et al.
2015/0051085 A1 2/2015 Vogelstein et al.
2015/0065358 A1 3/2015 Comstock et al.
2015/0080266 A1 3/2015 Volkmuth et al.
2015/0087535 A1 3/2015 Patel
2015/0167069 A1 6/2015 Schutz et al.
2015/0197786 A1 7/2015 Osborne et al.
2015/0275289 A1 10/2015 Otwinowski et al.
2015/0299812 A1 10/2015 Talasaz
2015/0329917 A1 11/2015 Shuber
2015/0344970 A1 12/2015 Vogelstein et al.
2015/0366866 A1 12/2015 Ali et al.
2015/0368708 A1 12/2015 Talasaz
2016/0002739 A1 1/2016 Schütz et al.
2016/0002741 A1 1/2016 Kitano et al.
2016/0024576 A1 1/2016 Chee
2016/0026758 A1 1/2016 Jabara et al.
2016/0032396 A1 2/2016 Diehn et al.
2016/0040229 A1 2/2016 Talasaz et al.
2016/0046986 A1 2/2016 Eltoukhy et al.
2016/0053301 A1 2/2016 Raymond et al.
2016/0060691 A1 3/2016 Giresi et al.
2016/0071432 A1 3/2016 Kurowski et al.
2016/0115553 A1 4/2016 Stephan et al.
2016/0130649 A1 5/2016 Xie et al.
2016/0251704 A1 9/2016 Talasaz et al.
2016/0319345 A1 11/2016 Gnerre et al.
2016/0333417 A1 11/2016 Talasaz
2016/0376647 A1 12/2016 Travers et al.
2017/0051347 A1 2/2017 Vogelstein et al.
2017/0073774 A1 3/2017 Lo et al.
2017/0145516 A1 5/2017 Kopetz et al.
2017/0159120 A1 6/2017 Eijk et al.
2017/0211143 A1 7/2017 Shendure et al.
2017/0218459 A1 8/2017 Talasaz et al.
2017/0218460 A1 8/2017 Talasaz
2018/0023125 A1 1/2018 Talasaz et al.
2019/0271040 A1 9/2019 Salk et al.
2019/0292597 A1 9/2019 Salk et al.
2019/0338358 A1 11/2019 Salk et al.
2019/0352714 A1 11/2019 Salk et al.
2021/0222243 A1 7/2021 Bielas

FOREIGN PATENT DOCUMENTS

EP 1647600 A2 6/2006
EP 2110442 A1 10/2009
EP 3070177 A1 9/2016
EP 3178941 A1 6/2017
WO 1997007241 A1 2/1997
WO 1997010365 A1 3/1997
WO 1999028505 A1 6/1999
WO 2000058516 A2 10/2000
WO 2002056014 A2 7/2002
WO 2005080604 A2 9/2005
WO 2005111242 A2 11/2005
WO 2006102264 A1 9/2006
WO 2007037678 A2 4/2007
WO 2008070144 A2 6/2008
WO 2008154317 A1 12/2008
WO 2009152928 A2 12/2009
WO 2011060240 A1 5/2011
WO 2011087760 A2 7/2011
WO 2011091046 A1 7/2011
WO 2011103236 A2 8/2011
WO 2011140510 A2 11/2011
WO 2011155833 A2 12/2011
WO 2012012693 A2 1/2012
WO 2012014877 A1 2/2012
WO 2012019200 A2 2/2012
WO 2012028746 A1 3/2012
WO 2012038839 A2 3/2012

(56) **References Cited**

FOREIGN PATENT DOCUMENTS

WO	2012042374	A2	4/2012
WO	2012048341	A1	4/2012
WO	2012054873	A2	4/2012
WO	2012066451	A1	5/2012
WO	2012071621	A1	6/2012
WO	2012088348	A2	6/2012
WO	2012099832	A2	7/2012
WO	2012106559	A1	8/2012
WO	2012129363	A2	9/2012
WO	2012142213	A2	10/2012
WO	2012142611	A2	10/2012
WO	2012148477	A1	11/2012
WO	2012103031	A3	1/2013
WO	2013019075	A2	2/2013
WO	2013033721	A1	3/2013
WO	2013106737	A1	7/2013
WO	2013123442	A1	8/2013
WO	2013130512	A2	9/2013
WO	2013130674	A1	9/2013
WO	2013138510	A1	9/2013
WO	2013142213	A1	9/2013
WO	2013142389	A1	9/2013
WO	2013148496	A1	10/2013
WO	2013159035	A2	10/2013
WO	2013173394	A2	11/2013
WO	2013181170	A1	12/2013
WO	2013188471	A2	12/2013
WO	2013190441	A2	12/2013
WO	2014004726	A1	1/2014
WO	2014014497	A1	1/2014
WO	2014015319	A1	1/2014
WO	2014039556	A1	3/2014
WO	2014093330	A1	6/2014
WO	2014145078	A1	9/2014
WO	2014149134	A2	9/2014
WO	2014151117	A1	9/2014
WO	2014152990	A1	9/2014
WO	2015159293	A2	10/2015
WO	2016015058	A2	1/2016
WO	2016040901	A1	3/2016
WO	2017100441	A1	6/2017
WO	2017181146	A1	10/2017

OTHER PUBLICATIONS

Guardant Health vs. *FMI* Invalidity Exhibit D-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit D-5, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Second Amended Complaint dated Mar. 6, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Suppl Invalidity References dated Feb. 15, 2019 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Supplemental Invalidity Contentions dated Mar. 29, 2019 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *PGDx* Amended Answer to Second Amended Complaint dated Apr. 30, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart A, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart B, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart C, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart D, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart E, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart F, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart G, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart H, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart I, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart J, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart K, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Claim Chart L, dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Invalidity Contentions dated Jun. 25, 2018 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Second Amended Complaint dated Mar. 23, 2018 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Second Updated Invalidity Contentions dated May 13, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Guardant Health vs. *PGDx* Updated Invalidity Contentions dated Mar. 29, 2019 (C.A. No. 17-cv-1623-LPS-CJB).

Gundry, et al. "Direct mutation analysis by high-throughput sequencing: from gremlin to low-abundant, somatic variants" Mutat Res Jan. 3, 2012; 729(1-2):1-15. doi: 10.1016/mrfmmm.2011.10.001. Pub Oct. 12, 2011.

Gundry, et al. Direct, genome-wide assessment of DNA mutations in single cells. Nucleic Acids Res. Mar. 2012;40 (5):2032-40. doi: 10.1093/nar/gkr949. Epub Nov. 15, 2011.

Hacia, et al. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nature Genetics, 22: 164-167 (1999).

Hafner, et al., RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA eDNA libraries, RNA Sep. 1, 2011, 17(9):1697-1712.

Hall, A., Short-Read DNA Sequence Alignmnet with Custom Designed FPGA-based Hardware, Thesis 2010.

Hamady, et al. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methods. Mar. 2008;5(3):235-7. doi: 10.1038/nmeth.1184. Epub Feb. 10, 2008.

Hensel, et al. Simultaneous identification of bacterial virulence genes by negative selection. Science. Jul. 21, 1995;269(5222):400-3.

Hiatt, et al. Parallel, tag-directed assembly of locally derived short sequence reads. Nat Methods. Feb. 2010;7(2):119-22. doi: 10.1038/nmeth.1416. Epub Jan. 17, 2010.

Hiatt, et al. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. Genome Res May 2013;23(5):843-54. doi: 10.1101/gr.147686.112. Epub Feb. 4, 2013.

Hibi, et al. Molecular detection of genetic alterations in the serum of colorectal cancer patients. Cancer Res. Apr. 1, 1998;58(7):1405-7.

HiSeq 2000 User Guide (2010).

Holdenrieder, et al. Circulating nucleosomes and cytokeratin 19-fragments in patients with colorectal cancer during chemotherapy. Anticancer Res. May-Jun. 2005;25(3A):1795-801.

Holies, et al. A stochastic approach to count RNA molecules using DNA sequencing methods. 2003, Lecture Notes in Computer Science, vol. 2812, pp. 55-62.

Hollas, et al. A stochastic approach to count RNA molecules using DNA sequencing methods. Lecture Notes in Computer Science, 2812: 55-62 (2003).

Hollstein, M. et al. "p53 mutations in human cancers" Science (1991) 253(5015):49-53.

Hoque, et al. Detection of aberrant methylation of four genes in plasma DNA for the detection of breast cancer. J Clin Oncol. Sep. 10, 2006;24(26):4262-9. Epub Aug. 14, 2006.

Howe, et al. Retinoblastoma growth suppressor and a 300-kDa protein appear to regulate cellular DNA synthesis. Proc Natl Acad Sci U S A. Aug. 1990;87(15):5883-7.

Hug, et al. Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. J Theor Biol. Apr. 21, 2003;221(4):615-24.

Hyland, et al. The normal and tumor spectrum of copy number variation: Copy number alterations correlate with changes in gene

(56) **References Cited**

OTHER PUBLICATIONS

expression in tumor transcriptome. Nov. 15, 2009. 1 page. Retrieved from:http://tools.thermofisher.com/content/sfs/posters/cms_073633.pdf.

Iafrate, et al. Detection of large-scale variation in the human genome. Nat Genet. Sep. 2004;36(9):949-51. Epub Aug. 1, 2004.

Ikeguchi, et al. Detection of loss of heterozygosityat microsatellite loci in esophageal squamous-cell carcinoma. Oncology. 1999;56(2):164-8.

Illumina "Multiplexed Sequencing with Illumina Genome Analyzer System" https://www.illumina.com/Documents/products/datasheets/datasheet_sequencing_multiplex.pdf (Dec. 2008).

Stumm, M. et al. "Noninvasive prenatal detection of chromosomal aneuploidies using different next generation sequencing strategies and algorithms" Prenatal Diagnosis (2012) 32:569-577.

Taback, et al. Detection of tumor-specific genetic alterations in bone marrow from early-stage breast cancer patients. Cancer Res. Apr. 15, 2003;63(8):1884-7.

Tan, et al. Genome-wide comparison of DNA hydroxymethylation in mouse embryonic stem cells and neural progenitor cells by a new comparative hMeDIP-seq method. Nucleic Acids Res. Apr. 2013;41(7):e84. doi: 10.1093/nar/gkt091. Epub Feb. 13, 2013.

Taudien, et al. Haplotyping and copy number estimation of the highly polymorphic human beta-defensin locus on 8p23 by 454 amplicon sequencing. BMC Genomics. Apr. 19, 2010;11:252. doi: 10.1186/1471-2164-11-252.

Teer, J.K. et al. "Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing" Genome Res (2010) 20(10):1420-1431.

Tie, J. et al., "Circulating tumor DNA analysis detects minimal residual disease and predicts recurrence in patients with stage II colon cancer,"Sci. Transl. Med. 2016, 8(346):346ra92.

Tomaz, et al. Differential methylation as a cause of allele dropout at the imprinted GNAS locus. Genet Test Mol Biomarkers. Aug. 2010;14(4):455-60. doi: 10.1089/gtmb.2010.0029.

Tomlinson, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet. Aug. 2007;39(8):984-8. Epub Jul. 8, 2007.

Trapnell, C. et al. "How to map billions of short reads onto genomes" Nature Biotech (2009) 27(5):455-457.

Tsai, et al. Discovery of rare mutations in populations: Tilling by sequencing. Plant Physiol. Jul. 2011;156(3):1257-68. doi: 10.1104/pp.110.169748. Epub Apr. 29, 2011.

U.S. Appl. No. 14/102,285, filed Dec. 10, 2013.

U.S. Appl. No. 61/613,413 ("Schmitt '413 provisional"), filed Mar. 20, 2012.

U.S. Appl. No. 61/625,319 ("Schmitt '319 provisional"), filed Apr. 17, 2012.

UCSC Genome Bioinformatics. About the UCSC Genome Bioinformatics Site. http://genome.ucsc.edu/index.html. Accessed on May 26, 2015. 2 pgs.

Umeiani et al. Prediction of Breast Tumor Progression by Integrity of Free Circulating DNA in Serum. Journal of Clinical Oncology 24(26):4270-4276 (Sep. 10, 2006).

U.S. Appl. No. 61/384,001, filed Sep. 17, 2010.

U.S. Appl. No. 61/432,119, filed Jan. 12, 2011.

Utting, et al. Microsatellite analysis of free tumor DNA in urine, serum, and plasma of patients: a minimally invasive method for the detection of bladder cancer. Clin Cancer Res. Jan. 2002;8(1):35-40.

Van Houten, et al. Mutated p53 as a molecular marker for the diagnosis of head and neck cancer. J Pathol. Dec. 2002;198(4):476-86.

Van Loo, P. et al. "Allele-specific copy number analysis of tumors" PNAS (2010) 107(39):16910-16915.

Van Orsouw, N. et al. "Complexity Reduction of Polymorphic Sequences (CRoPSTM): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes" PLoS One (2007) 11(e1172):1-10.

Vasyukhin, V. et al., "K-Ras Point Mutations in the Blood Plasma DNA of Patients with Colorectal Tumors" Challenges of Modern Medicine, 141-150 (Verna and Shamoo eds, 1994).

Velculescu, et al. Characterization of the Yeast Transcriptome. Cell, 88: 243-251 (1997).

Velculescu, et al. Serial Analysis of Gene Expression. Science, 270: 484-487 (1995).

Vogelstein, et al. Digital PCR. Proc. Natl. Acad. Sci., 96(16): 9236-9241(1999).

Wagle, N. et al., "High-throughput Detection of actionable Genomic alterations in clinical tumor samples by targeted, Massively Parallel sequencing," Cancer Discov. 2012, (2)1:82-93.

Walker, et al. Isothermal in vitro amplification of DNA by a restriction enzyme/DNA polymerase system. Proc Natl Acad Sci U S A. Jan. 1, 1992;89(1):392-6.

Walsh, et al. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc Natl Acad Sci U S A. Jul. 13, 2010;107(28):12629-33. doi: 10.1073/pnas.1007983107. Epub Jun. 28, 2010.

Wang, et al. iCLIP predicts the dual splicing effects of TIA-RNA interactions, Oct. 2010, PLoS Biol, 8(10):e1000530.

Wang, et al. Molecular detection of APC, K- ras, and p53 mutations in the serum of colorectal cancer patients as circulating biomarkers. World J Surg. Jul. 2004;28(7):721-6. Epub Jun. 8, 2004.

Wang, et al. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10: 57-63 (2009).

Wang, T.L. et al. "Digital Karyotyping" PNAS (2002) 99(25):16156-16161.

Wang, T.T. et al. "High efficiency error suppression for accurate detection of low-frequency variants" NAR (2019) 47(15).

Wang, Y. et al., "Detection of somatic mutations and HPV in the saliva and plasma of patients with head and neck squamous cell carcinomas," Sci. Transl. Med. 2015, 7(293):293ra104.

Wang, Y. et al., "Detection of tumor-derived DNA in cerebrospinal fluid of patients with primary tumors of the brain and spinal cord," Proc. Natl. Acad. Sci. USA 2015, 112(31), 9704-9709.

Wang, Y. et al., "Diagnostic potential of tumor DNA from ovarian cyst fluid," eLife 2016, 5:e15175.

Weber, et al. A real-time polymerase chain reaction assay for quantification of allele ratios and correction of amplification bias. Anal Biochem. Sep. 15, 2003;320(2):252-8.

Wheeler, D.A. et al., "The complete genome of an individual by massively parallel DNA sequencing," Nature 2008, 452(7189), 872-876.

Williford, A. et al., "Gene Fusion," eLS 2013, 1-8.

Wittes, et al. Searching for Evidence of Altered Gene Expression: a Comment on Statistical Analysis of Microarray Data. Journal of the National Cancer Institute, 91(5): 400-401 (1999).

Wodicka, et al. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. Nature Biotechnology, 15:1359-1367 (1997).

Wojdacs, et al. Primer design versus PCR bias in methylation independent PCR amplifications. Epigenetics. May 16, 2009;4(4):231-4. Epub May 14, 2009.

Wood, et al. The genomic landscapes of human breast and colorectal cancers. Science. Nov. 16, 2007;318(5853):1108-13. Epub Oct. 11, 2007.

Wood, et al. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucleic Acids Res. Aug. 2010;38(14):e151. doi: 10.1093/nar/gkq510. Epub Jun. 4, 2010.

Xi, et al. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. Proc Natl Acad Sci U S A. Nov. 15, 2011;108(46):E1128-36. doi: 10.1073/pnas.1110574108. Epub Nov. 7, 2011.

Yandell, et al. A probabilistic disease-gene finder for personal genomes. Genome Res. Sep. 2011;21(9):1529-42. doi: 10.1101/gr.123158.111 Epub Jun. 23, 2011.

Yang, et al., EGFR gene copy number as a predictive biomarker for the treatment of metastatic colorectal cancer with anti-EGFR monoclonal antibodies: a meta-analysis.,J Hematol Oncol, Aug. 16, 2012,5:52,1-9.

(56) **References Cited**

OTHER PUBLICATIONS

Yang. Simple binary segmentation frameworks for identifying variation in DNA copy number. BMC Bioinformatics. Oct. 30, 2020;13:277. doi: 10.1186/1471-2105-13-277.

Ye, et al. Fluorescent microsphere-based readout technology for multiplexed human single nucleotide polymorphism analysis and bacterial identification. Human Mutation, 17(4): 305-316 (2001).

Yoon, et al. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. Sep. 2009;19(9):1586-92. doi: 10.1101/gr.092981.109. Epub Aug. 5, 2009.

EPO Preliminary Opinion, dated Aug. 25, 2020, in EP2893040.

Guardant Health, Inc. Response to Notices of Opposition in EP2893040 filed May 29, 2020.

IPR2019-00634, Final Written Decision of U.S. Pat. No. 9,840,743, dated Aug. 18, 2020.

IPR2019-00652, Final Written Decision of U.S. Pat. No. 9,834,822, dated Aug. 18, 2020.

Lo, Dennis Interview with Professor Dennis Lo, Qiagen News (2002).

Opponents Reply dated Aug. 20, 2020 to Proprietor's Observation in EP2893040.

Opposition Form and Statement to EP3470533 filed Aug. 6, 2020 by Foundation Medicine Inc.

U.S. Appl. No. 61/625,623, filed Apr. 17, 2012.

Illumina Preparing Samples for Paired-End Sequencing (2008).

Ingolia, et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. Apr. 10, 2009;324(5924):218-23. Epub Feb. 12, 2009.

Instructions for Norit Rapid DNA Ligation Kit (Nov. 6, 2004).

International search report and written opinion dated Apr. 3, 2015 for PCT/US2014/072383.

International search report and written opinion dated May 7, 2012 for PCT/IB2011/003160.

International search report and written opinion dated Jun. 6, 2012 for PCT/US2011/065291.

International search report and written opinion dated Jul. 15, 2015 for PCT/US2015/030639.

International search report and written opinion dated Sep. 5, 2014 for PCT/US2014/000048.

International search report and written opinion dated Sep. 6, 2017 for PCT Application No. US-201727809.

International search report and written opinion dated Nov. 18, 2013 for PCT/US2013/058061.

Invitrogen Instructions for T4 DNA Ligase (May 5, 2002).

IPR2018-00130, Petition for Inter Partes Review of U.S. Pat. No. 9,598,731, dated Nov. 7, 2018.

IPR2019-00130, Decision Denying Inter Partes Review U.S. Pat. No. 9,598,731, issued Jun. 5, 2019.

IPR2019-00130, Preliminary Response to Petition for Inter Partes Review U.S. Pat. No. 9,598,731, filed Mar. 6, 2019.

IPR2019-00634, Decision Instituting Inter Partes Review U.S. Pat. No. 9,840,743, issued Aug. 19, 2019.

IPR2019-00634, Petition for Inter Partes Review of U.S. Pat. No. 9,840,743, dated Feb. 1, 2019.

IPR2019-00636 & 637, Exhibit 3001 Memo for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Jul. 1, 2019.

IPR2019-00636 & 637, Order—Conduct of the Proceedings for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Jul. 22, 2019.

IPR2019-00636 and IPR2019-00637, Decision Denying Inter Partes Review U.S. Pat. No. 9,902,992, issued Aug. 20, 2019.

IPR2019-00636, Petition for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Feb. 1, 2019.

IPR2019-00637, Petition for Inter Partes Review of U.S. Pat. No. 9,902,992, dated Feb. 1, 2019.

IPR2019-00652, Decision Granting Inter Partes Review U.S. Pat. No. 9,834,822, issued Aug. 19, 2019.

IPR2019-00652, Petition for Inter Partes Review of U.S. Pat. No. 9,834,822, dated Feb. 1, 2019.

IPR2019-00653, Decision Denying Inter Partes Review U.S. Pat. No. 9,834,822, issued Aug. 19, 2019.

IPR2019-00653, Petition for Inter Partes Review of U.S. Pat. No. 9,834,822, dated Feb. 1, 2019.

Jabara, C. Capturing the cloud: High throughput sequencing of multiple individual genomes from a retroviral population. Biology Lunch Bunch Series, Training Initiatives in Biomedical & Biological Sciences of the University of North Carolina at Chapel Hill, Apr. 23, 2010.

Jabara, et al. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. (Paper # 665), The 18th Annual Conference on Retroviruses and Opportunistic Infections, Boston, Massachusetts, Mar. 2011.

Jabara, et al. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl Acad Sci U S A. Dec. 13, 2011;108(50):20166-71. doi: 10.1073/pnas.1110064108. Epub Nov. 30, 2011.

Jahr, et al. DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. Cancer Res. Feb. 15, 2001;61(4):1659-65.

Jeronimo, et al. Quantitative GSTP1 hypermethylation in bodily fluids of patients with prostate cancer. Urology. Dec. 2002;60(6):1131-5.

Jiang, H. et al., "SeqMap: mapping massive amount of oligonucleotides to the genome," Bioinformatics 2008, 24 (20):2395-2396.

Kamps-Hughes, N. et al. "ERASE-Seq: Leveraging replicate measurements to enhance ultralow frequency variant detection in NGS data"PLOS One (2018).

Kanagawa. Bias and artifacts in multitemplate polymerase chain reactions (PCR), 2003, Journal of Bioscience and Bioengineering, vol. 96, No. 4, p. 317-323.

Kao,W-C et al., "BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing," Genome Res. 2009 19(10), 1884-1895.

Karow, J. "Hopkins Team Develops Method to Improve Rare Mutation Detection for Early Cancer Dx" Genome Web (2011) 3 pages.

Kennedy, S.R. et al., "Detecting ultralow-frequency mutations by Duplex Sequencing" Nature Protocols (2014) 9(11):2586-2606.

Kennedy, S.R. et al., "Ultra-Sensitive Sequencing Reveals an Age-Related Increase in Somatic Mitochondrial Mutations That Are Inconsistent with Oxidative Damage" PLOS Genetics (2013)9:e1003794.

Kennedy, S.R. et al., "Detecting ultralow-frequency mutations by Duplex Sequencing," Nat. Protoc. 2014, 9(11), 2586-2606.

Kimura, et al. EGFR mutation status in tumour-derived DNA from pleural effusion fluid is a practical basis for predicting the response to gefitinib. Br J Cancer. Nov. 20, 2006;95(10):1390-5. Epub Oct. 24, 2006.

Kinde, et al. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. Jun. 7, 2011;108(23):9530-5. doi: 10.1073/pnas.1105422108. Epub May 17, 2011.

Kinde, et al. Supplemental Information, Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. Jun. 7, 2011;108(23):1-10.

Kinde, I. et al., "Evaluation of DNA from the Papanicolaou Test to Detect Ovarian and Endometrial Cancers," Sci. Transl. Med. 2013, 5(167):167ra4.

Kinde, I et al., "Fast-SeqS: a simple and efficient method for the detection of aneuploidy by massively parallel sequencing," PLoS One 2012, 7(7), e41162.

Kircher, M. et al., "Improved base calling for the Illumina Genome Analyzer using machine learning strategies," Genome Biol. 2009, 10(8), R83.

Kirsch, S. et al. "Sequence error storms and the landscape of mutations in cancer," Proc. Natl. Acad. Sci. USA 2012, 109(36), 14289-14290.

Kivioja, T., et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. Nov. 20, 2011;9(1):72-4 doi: 10.1038/nmeth.1778.

Koboldt, D.C.et al., "The next-generation sequencing revolution and its impact on genomics," Cell 2013, 155(1), 27-38.

(56) **References Cited**

OTHER PUBLICATIONS

Koboldt, et al. Massively parallel sequencing approaches for characterization of structural variation. Aug. 12, 2011. Methods Mol Biol. 2012;838:369-84. doi: 10.1007/978-1-61779-507-7_18.
Kolble, et al. Microsatellite alterations in serum DNA of patients with colorectal cancer. Lab Invest. Sep. 1999;79(9):1145-50.
Konig, et al. iCLIP reveals the function of hnRNAP particles in splicing at individual nucleotide resolution, Jul. 2010, Mature Structural & Molecular Biology, 17(7):909-916.
Mohan, et al.,Changes in colorectal carcinoma genomes under anti-EGFR therapy identified by whole-genome plasma DNA sequencing.,PLoS Genet,doi: 10.1371/journal.pgen.1004271. eCollection 2014.,Mar. 27, 2014, 10(3), e1004271.
Miori, et al. Predictive utility of circulating methylated DNA in serum of melanoma patients receiving biochemotherapy. J Clin Oncol. Dec. 20, 2005;23(36):9351-8.
Morrissy, A.S. et al. "Next-generation tag sequencing for cancer gene expression profiling" Genome Research (2009)19(10): 1825-1835.
Mortazavi, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. 2008; 5, 621-628.
Narayan, et al. Ultrasensitive measurement of hotspot mutations in tumor DNA in blood using error-suppressed multiplexed deep sequencing. Cancer Res. Jul. 15, 2012;72(14):3492-8. doi: 10.1158/0008-5472.CAN-11-4037. Epub May 10, 2012.
Nawroz, et al. Microsatellite alterations in serum DNA of head and neck cancer patients. Nat Med. Sep. 1996;2 (9):1035-7.
Newman, A. et al. "Integrated digital error suppression for improved detection of circulating tumor DNA" Nature Biotech (2016) 34(5):547-555.
Newman, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. May 2014;20(5):548-54. doi: 10.1038/nm 3519. Epub Apr. 6, 2014.
Ng, S B., et al., "Targeted capture and massively parallel sequencing of 12 human exomes," Nature 2009,461(7261), 272-276.
Nielsen, R. et al. "Genotype and SNP calling from next-generation sequencing data" Nature Reviews Genetics (2011) 12(6):443-451.
Nord, et al. Accurate and exact CNV identification from targeted high-throughput sequence data. BMC Genomics. Apr. 12, 2011;12:184.
Notice of Allowance dated Jan. 31, 2017 for U.S. Appl. No. 14/712,754.
Notice of allowance dated Mar. 21, 2014 for U.S. Appl. No. 12/969,581.
Notice of allowance dated Jun. 15, 2017 for U.S. Appl. No. 15/076,565.
Notice of allowance dated Jun. 19, 2014 for U.S. Appl. No. 12/969,581.
Notice of allowance dated Aug. 1, 2017 for U.S. Appl. No. 15/492,659.
Notice of allowance dated Aug. 4, 2017 for U.S. Appl. No. 15/467,570.
Notice of allowance dated Aug. 22, 2014 for U.S. Appl. No. 12/969,581.
Notice of allowance dated Aug. 29, 2017 for U.S. Appl. No. 15/492,659.
Notice of allowance dated Sep. 11, 2017 for U.S. Appl. No. 15/076,565.
Notice of allowance dated Oct. 3, 2017 for U.S. Appl. No. 15/076,565.
Notice of allowance dated Oct. 25, 2017 for U.S. Appl. No. 14/861,989.
Notice of allowance dated Dec. 28, 2017 for U.S. Appl. No. 14/861,989.
Odegaard, J.I et al. "Validation of a Plasma-Based Comprehensive Cancer Genotyping Assay Utilizing Orthogonal Tissue- and Plasma-Based Methodologies" Clin Canc Res (2018) 24(15):3539-3549.
Office Action dated Feb. 9, 2017 for U.S. Appl. No. 15/076,565.
Office action dated May 13, 2019 for U.S. Appl. No. 15/669,779.
Office action dated May 20, 2016 for U.S. Appl. No. 14/855,301.
Office action dated May 31, 2016 for U.S. Appl. No. 14/712,754.
Office action dated Jun. 1, 2017 for U.S. Appl. No. 15/467,570.
Office action dated Jun. 3, 2016 for U.S. Appl. No. 14/861,989.
Office action dated Jun. 12, 2017 for U.S. Appl. No. 15/492,659.
Office action dated Jul. 18, 2017 for U.S. Appl. No. 14/861,989.
Office Action dated Jul. 30, 2019 for U.S. Appl. No. 16/283,635.
Office action dated Aug. 6, 2019 for U.S. Appl. No. 14/855,301.
Office action dated Oct. 3, 2013 for U.S. Appl. No. 12/969,581.
Office Action dated Oct. 12, 2016 for U.S. Appl. No. 14/712,754.
Office action dated Oct. 20, 2017 for U.S. Appl. No. 14/425,189.
Office Action dated Nov. 21, 2016 for U.S. Appl. No. 14/855,301.
Office action dated Dec. 4, 2015 for U.S. Appl. No. 14/712,754.
Office action dated Dec. 7, 2017 for U.S. Appl. No. 14/855,301.
Office Action dated Dec. 9, 2016 for U.S. Appl. No. 14/861,989.
Office action dated Dec. 21, 2018 for U.S. Appl. No. 15/978,848.
Ogino, et al. Quantification of PGR bias caused by a single nucleotide polymorphism in SMN gene dosage analysis. J Mol Diagn. Nov. 2002;4(4):185-90.
Opposition Brief dated Aug. 7, 2018 for Japanese Opposition No. 2018-700659 to JP Patent 6275145.
Opposition Decision dated Jan. 15, 2019 for Japanese Opposition No. 2018-700659 to JP Patent 6275145.
Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Foundation Medicine Inc.
Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Grunecker Patent-und Rechtsanwalte PartG mbB.
Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Personal Genome Diagnostics, Inc.
Opposition Form and Statement to EP2893040 filed Oct. 2, 2019 by Strawman Limited.
Opposition Form and Statement to EP3087204 filed Nov. 14, 2018.
D'Antoni, et al. Rapid quantitative analysis using a single molecule counting approach. Anal Biochem. May 1, 2006;352(1):97-109. Epub Feb. 10, 2006.
Daser, et al. "Interrogation of genomes by molecular copy-number counting (MCC)." Nature Methods, 3(6): 447-453 (2006).
De Saizieu, et al. "Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays." Nature Biotechnology, 16: 45-48 (1998).
Diehl et al. Circulating mutant DNA to assess tumor dynamics. Nat Med 14(9):985-990 (2008).
Diehl, et al. Detection and quantification of mutations in the plasma of patients with colorectal tumors Proc Natl Acad Sci US A. Nov. 8, 2005; 102(45): 16368-73. Epub Oct. 28, 2005.
Diehl, F. et al., "Analysis of Mutations in DNA Isolated From Plasma and Stool of Colorectal Cancer Patients," Gastroenterology (2008) 135(2):489-498.
Ding, L. et al. "Clonal evolution in relapsed acute myeloid leukemia revealed by whole-genome sequencing" Nature (2012)481(7382):506-510.
Ding, L. et al., "Analysis of next-generation genomic data in cancer: accomplishments and challenges," Hum. Mol. Genet. 2010 19(R2), R188-R196.
Duncan, D.L. et al. "Next-Generation Sequencing in the Clinical Laboratory" Diagnostic Molecular Pathology: A Guide to Applied Molecular Testing 25-33 (Coleman and Tsongalis eds., 1st ed. 2016).
Ehrich, et al. Noninvasive detection of fetal trisomy 21 by sequencing of DNA in maternal blood: a study in a clinical setting Am J Obstet Gynecol. Mar. 2011;204(3):205.e1-11. doi: 10.1016/j.ajog.2010.12.060. Epub Feb. 18, 2011.
Eisenmann, et al. 5q- myelodysplastic syndromes: chromosome 5q genes direct a tumor-suppression network sensing actin dynamics. Oncogene. Oct. 1, 2009;28(39):3429-41. doi: 10.1038/onc.2009.207. Epub Jul. 13, 2009.
Elshire, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. May 4, 2011;6(5):e19379. doi: 10.1371/journal.pone.0019379.
European search report and search opinion dated Mar. 29, 2016 for EP Application No. 13834427.0.
Extended European search report and opinion dated May 4, 2017 for EP Application No. 14874157.2.
Extended European search report and opinion dated Nov. 21, 2016 for EP Application No. 14771159.
Fan, et al. "Microfluidic digital PCR enables rapid prenatal diagnosis of fetal aneuploidy." Am Obstet Gynecol. 2009; 200:543.e1-543.e7.

(56) **References Cited**

OTHER PUBLICATIONS

Fan, et al. "Parallel Genotyping of Human SNPs Using Generic High-density Oligonucleotide Tag Arrays." Genome Research, 10: 853-860 (2000).

Fan, et al. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci U S A. Oct. 21, 2008;105(42):16266-71. Epub Oct. 6, 2008.

Fan, et al. Non-invasive prenatal measurement of the fetal genome. Nature. Jul. 19, 2012;487(7407):320-4. doi:10.1038/nature 11251.

Fleischhacker, M. et al. "Circulating nucleic acids (CNAs) and cancer—A survey" Biochimica et Biophysica Acta (2007) 1775:181-232.

Fonatsch, C. The role of chromosome 21 in hematology and oncology. Genes Chromosomes Cancer. Jun. 2010;49(6):497-508. doi: 10.1002/gcc.20764.

Fong, S.L. et al. "Comparison of 7 Methods for Extracting Cell-Free DNA from Serum Samples of Colorectal Cancer Patients" Clinical Chem (2009) 55(3):587-598.

Forshew, T. et al. "Noninvasive Identification and Monitoring of Cancer Mutations by Targeted Deep Sequencing of Plasma DNA" Sci Transl Med (2012) 4(136) ra68.

Forster, et al. From next-generation sequencing alignments to accurate comparison and validation of single-nucleotide variants: the pibase software. Nucleic Acids Res. Jan. 7, 2013;41(1):e16. doi: 10.1093/nar/gks836. Epub Sep. 10, 2012.

Fournie, et al. Plasma DNA as a marker of cancerous cell death. Investigations in patients suffering from lung cancer and in nude mice bearing human tumours. Cancer Lett. May 8, 1995;91(2):221-7.

Freeman, et al. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. Oct. 2009;19(10):1817-24. doi: 10.1101/gr.092924.109. Epub Jun. 18, 2009.

Fu, et al. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc Natl Acad Sci U S A. May 31, 2011;108(22):9026-31. Epub May 11, 2011.

Fujiwara, et al. Identification of epigenetic aberrant promoter methylation in serum DNA is useful for early detection of lung cancer. Clin Cancer Res. Feb. 1, 2005;11(3):1219-25.

Fujiwara, et al. Plasma DNA microsatellites as tumor-specific markers and indicators of tumor progression in melanoma patients Cancer Res. Apr. 1, 1999;59(7):1567-71.

Genome Analyzer IIx Systems Specification Sheet (2009).

Gerry, et al. Universal DNA microarray method for multiplex detection of low abundance point mutations. Journal of Molecular Biology, 292(2): 251-262 (1999).

Gillespie. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry, 81(25): 2340-2361 (1977).

Gordian, et al. Serum free circulating DNA is a useful biomarker to distinguish benign versus malignant prostate disease. Cancer Epidemiol Biomarkers Prev. Aug. 2010;19(8):1984-91. doi: 10.1158/1055-9965.EPI-10-0287. Epub Jul. 20, 2010.

Gordon, D.J. et al., "Causes and consequences of aneuploidy in cancer," Nat. Rev. Genet. 2012 13(3), 189-203.

Gormally, et al. Amount of DNA in plasma and cancer risk: a prospective study. Int J Cancer. Sep. 20, 2004;111 (5):746-9.

Grant, et al. SNP genotyping on a genome-wide amplified DOP-PCR template. Nucleic Acids Res. Nov. 15, 2002;30(22):e25.

Gregory et al. "Targeted Single Molecule Mutation Detection with Massively Parallel Sequencing" Nucleic Acids Research (2015) 44(3):2-11.

Grutzmann, et al. Sensitive detection of colorectal cancer in peripheral blood by septin 9 DNA methylation assay. PLoS One. 2008;3(11):e3759. doi: 10.1371/journal.pone.0003759. Epub Nov. 19, 2008.

Guardant Health vs. *FMI* 1st Amended Answer to Second Amended Complaint dated May 6, 2019 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Contentions dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit A-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit A-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit A-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit A-8, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit B-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit B-2, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit B-3, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit B-7, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit C-1, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Guardant Health vs. *FMI* Invalidity Exhibit C-11, dated Jun. 25, 2018 (C.A. No. 17-cv-1616-LPS-CJB).

Kopreski, et al. Detection of mutant K-ras DNA in plasma or serum of patients with colorectal cancer. Br J Cancer. 1997;76(10):1293-9.

Korbel, J.O. et al., "Paired-end mapping reveals extensive structural variation in the human genome," Science 2007, 318(5849), 420-426.

Koyanagi, et al. Association of circulating tumor cells with serum tumor-related methylated DNA in peripheral blood of melanoma patients. Cancer Res. Jun. 15, 2006;66(12):6111-7.

Krimmel, J.D et al., "Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues," Proc. Natl. Acad. Sci. USA 2016, 113(21), 6005-6010.

Kukita, Y. et al. "High-fidelity target sequencing of individual molecules identified using barcode sequences: de novo detection and absolute quantitation of mutations in plasma cell-free DNA from cancer patients" DNA Research (2015) 22(4):269-277.

Lam, et al. Plasma DNA as a prognostic marker for stroke patients with negative neuroimaging within the first 24 h of symptom onset. Resuscitation. Jan. 2006;68(1):71-8. Epub Dec. 1, 2005.

Lanman, et al., Analytical and Clinical Validation of a Digital Sequencing Panel for Quantitative, Highly Accurate Evaluation of Cell-Free Circulating Tumor DNA PLoS One, Oct. 2015, 10(10), e0140712. doi:10.1371/journal.pone.0140712.

Larson, et al. A single molecule viewof of gene expression. Trends Cell Biol. Nov. 2009;19(11):630-7. Epub Oct. 8, 2009.

Leary, et al. Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Sci Transl Med. Nov. 28, 2012;4(162):162ra154. doi: 10.1126/scitranslmed. 3004742.

Leary, et al. Development of personalized tumor biomarkers using massively parallel sequencing. Sci Transl Med. Feb. 24, 2010;2(20):20ra14. doi: 10.1126/scitranslmed.3000702.

Lecomte, et al. Detection of free-circulating tumor-associated DNA in plasma of colorectal cancer patients and its association with prognosis. Int J Cancer. Aug. 10, 2002;100(5):542-8.

Ledergerber, C. et al., "Base-calling for next-generation sequencing platforms," Brief Bioinform. 2011 12(5), 489-497.

Lennon, N.J. et al. "Technological considerations for genome-guided diagnosis and management of cancer" Gen Med (2016)8:112.

Leon, et al. Free DNA in the serum of cancer patients and the effect of therapy. Cancer Res. Mar. 1977;37(3):646-50.

Leung, et al. Plasma Epstein-Barr viral deoxyribonucleic acid quantitation complements tumor-node-metastasis staging prognostication in nasopharyngeal carcinoma. J Clin Oncol. Dec. 1, 2006;24(34):5414-8.

Li, et al. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. Nov. 2008;18(11):1851-8. doi: 10.1101/gr.078212.108 Epub Aug. 19, 2008.

Li, et al. Structure-independent and quantitative ligation of single-stranged DNA. Anal Biochem. Feb. 15, 2006;349(2):242-6. Epub Nov. 18, 2005.

Li, H. et al., "Fast and accurate short read alignment with Burrows-Wheeler transform," Bioinformatics 2009, 25(14), 1754-1760.

(56) **References Cited**

OTHER PUBLICATIONS

Liang, K-C. et al., "Bayesian basecalling for DNA sequence analysis using hidden Markov models," IEE/ACM Trans. Comput. Biol. Bioinform. 2007 4(3), 430-440.

Liao, G.J.W. et al., "Targeted massively parallel sequencing of maternal plasma DNA permits efficient and unbiased detection of fetal alleles," Clin. Chem. 2011, 57(1), 92-101.

Lizardi, et al. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. Nat Genet. Jul. 1998;19(3):225-32.

Lo, et al. Quantitative Analysis of Cell-free Epstein-Barr Virus DNA in Plasma of Patients with Nasopharyngeal Carcinoma, Cancer Research, 59(6):1188-1191 (Mar. 1999).

Lo, Y.M.D et al., "Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis," Am. J. Hum. Genet. 1998, 62(4), 768-775.

Lo, Y.M.D. et al. "Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus," Sci Transl Med (2010) 2(61): 1-13.

Lockhart, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology, 14: 1675-1680 (1996).

Lodes, et al. Novel Chimera-Free, High Efficiency Library Preparation for NGS Platforms.Poster. Presented at the Plant & Animal Genome XX Conference. San Diego, California. Jan. 14-18, 2012. 1 page.

Lucito, et al. Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. Genome Res. Nov. 2000;10(11):1726-36.

Lucito, et al. Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation. Genome Research (2003), 13: 2291-2305.

Lunter, G. et al., "Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads," Genome Res. 2011, 21(6):936-939.

Maamar, et al. Noise in Gene Expression Determines Cell Fate in Bacillus subtilis. Science, 317: 526-529 (2007).

Mahmud, et al. Fast MCMC sampling for hidden Markov Models to determine copy number variations. BMC Bioinformatics. Nov. 2, 2011;12:428. doi: 10.1186/1471-2105-12-428.

Makrigiorgos, et al., A PCR-Based amplification method retaining quantative difference between two complex genomes. Nature Biotech, vol. 20, No. 9, pp. 936-939 (Sep. 2002).

Mamanova, L. et al., "Target-enrichment strategies for next-generation sequencing," Nat. Methods 2010, 7(2), 111-118.

Mandel, et al. Les acides nucleiques du plasma sanguin chez I'homme. C R Seances Soc Biol Fil. Feb. 1948; 142 (3-4):241-3.

Marsit, et al. Epigenetic profiling reveals etiologically distinct patterns of DNA methylation in head and neck squamous cell carcinoma. Carcinogenesis. Mar. 2009;30(3):416-22. doi: 10.1093/carcin/bgp006. Epub Jan. 6, 2009.

McCloskey, et al. Encoding PCR products with batch-stamps and barcodes. Biochem Genet. Dec. 2007;45 (11-12):761-7. Epub Oct. 23, 2007.

McKernon, K.J. et al., "Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding," Genome Res. 2009,19(9), 1527-1541.

Medvedev, et al. Detecting copy number variation with mated short reads. Genome Res. Nov. 2010;20(11):1613-22. doi: 10.1101/gr.106344.110 Epub Aug. 30, 2010.

Mei, et al. Identification of recurrent regions of Copy-Number Variants across multiple individuals. BMC Bioinformatics. Mar. 22, 2010;11:147. doi: 10.1186/1471-2105-11-147.

Meldrum, et al. Next generation sequencing for cancer diagnostics: a practical perspective. Clin Biochem Rev. Nov. 2011;32(4):177-95.

Mertes, F. et al. "Targeted enrichment of genomic DNA regions for next-generation sequencing" Brief Functional Genomics (2011) 10(6):374-386.

Metzker, M.L. "Sequencing technologies—the next generation" Nature Reviews Genetics (2010) 11:31-46.

Meyer, M. et al. "Illumina sequencing library preparation for highly multiplexed target capture and sequencing," Cold Spring Harb. Protoc. 2010, (6), prot5448.

Meyerson, M. et al. "Advances in understanding cancer genomes through second-generation sequencing" Nature Reviews Genetics (2010) 11:685-696.

Milbury, et al. "Enabling Clinical Cancer Genomics for Rare Mutations: COLD-PCR Magnifies Mutations Prior to Targeted Amplicon Re-Sequencing". Clin Chem. Mar. 2012;58(3):580-9. doi: 10.1373/clinchem.2011.176198. Epub Dec. 21, 2011.

Miner, et al. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. Nucleic Acids Res. Sep. 30, 2004;32(17):e135.

Mishra, S. et al. "Different Facets of Copy Number Changes: Permanent, Transient, and Adaptive" Mol Cell Biol (2016) 36(7):1050-1063.

Mitchell, et al. Circulating microRNAS as stable blood-based markers for cancer detection. Proc Natl Acad Sci U S A. Jul. 29, 2008;105(30):10513-8. Epub Jul. 28, 2008.

Mitelman, F. et al. "The impact of translocations and gene fusions on cancer causation" Nature Rev. Cancer (2007) 7(4):233-245.

Moench, S. Genomic Profiling Using Guardant 360 Cell-Free DNA-Based Assay vs Tumor-Based Genotyping Assays in Advanced NSCLC, Cancer Therapy Advisor (Feb. 28, 2019), https://www.cancertherapyadvisor.com/home/news/conferencecoverage/american-association-for-cancer-research-aacr/aacr-2019/genomic-profiling-using-guardant-360-cell-free-dna-based assay-vs-tumor-based-genotyping-assays-in-advanced-nsclc/ (lastaccessed Nov. 30, 2019).

Ou, Shi et al. "Liquid Biopsy to Identify Actionable Genomic Alterations" Am Soc Clin One (2018) 978.

Pacific Biosciences. Template Preparation and Sequencing Guide. Publication date: Oct. 14, 2014. Pacific Biosciences website http://www.pacificbiosciences.com/support/pubmap/documentation.html.

Pan, et al. Loss of heterozygosity patterns provide fingerprints for genetic heterogeneity in multistep cancer progression of tobacco smoke-induced non-small cell lung cancer. Cancer Res. Mar. 1, 2005;65(5):1664-9.

Park, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet. May 2010;42(5):400-5. doi: 10.1038/ng.555. Epub Apr. 4, 2010.

Parkinson, N.J. et al., "Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA," Genome Res. 2012, 22(1), 125-133.

Paweleiz, C.P. et al. "Bias-corrected targeted next-generation sequencing for rapid, multiplexed detection of actionable alterations in cell-free DNA from advanced lung cancer patients" Clin Canc Res (2016) 22(4):915-922.

Pel, J. et al. "Duplex proximity sequencing (pro-seq): A method to improve DNA sequencing accuracy without the cost of molecular barcoding redundancy" bioRxiv (2017) https://doi.org/10.1101/163444.

Perakis, S. et al. "Advances in Circulating Tumor DNA Analysis" Adv Clin Chem (2017) pp. 1-81.

Phallen, J. et al. "Direct detection of early-stage cancers using circulating tumor DNA" Sci Trans Med (2017) vol. 9, Issue 403, eaan2415DOI: 10.1126/scitranslmed.aan2415.

Pihlak, et al. Rapid genome sequencing with short universal tiling probes. Nature Biotechnology, 26: 676-684 (2008).

Pinkel, et al. Comparative Genomic Hybridization. Annual Review of Genomics and Human Genetics, 6: 331-354 (2005).

Pleasance, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature. Jan. 14, 2010;463(7278):184-90. doi: 10 1038/nature08629. Epub Dec. 16, 2009.

Pollack, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet. Sep. 1999;23(1):41-6.

Qiu, et al. DNA sequence-based "bar codes" for tracking the origins of expressed sequence tags from a maize cDNA library constructed using multiple mRNA sources. Plant Physiol Oct. 2003;133(2):475-81.

(56) **References Cited**

OTHER PUBLICATIONS

Quail, et al. A large genome center's improvements to the Illumina sequencing system. Nat Methods. Dec. 2008;5(12):1005-10. doi: 10.1038/nmeth.1270.

Quinlan, A.R. et al., "Pyrobayes: an improved base caller for SNP discovery in pyrosequences," Nat. Methods 2008 5(2), 179-181.

Rafi, I. et al. "Cell-free fetal DNA and non-invasive prenatal diagnosis," Br. J. Gen. Pract. May 1, 2009; 59(562)e146-e148.

Rand, et al. Headloop suppression PCR and its application to selective amplification of methylated DNA sequences. Nucleic Acids Res. Aug. 9, 2005;33(14):e127.

Redon, R. et al., "Global variation in copy number in the human genome," Nature 2006 444(7118), 444-454.

Report and Recommendation in Civil Action No. 17-1623-LPS-CJB between Guardant Health, Inc. and Foundation Medicine and between Guardant Health, Inc. and Personal Genome Diagnostics, Inc., dated Oct. 11, 2019.

Rizzo, J.M. et al. "Key Principles and Clinical Applications of 'Next Generation' DNA Sequencing," Cancer Prev. Res., (2012)5, 887-900.

Ryan, et al. A prospective study of circulating mutant KRAS2 in the serum of patients with colorectal neoplasia: strong prognostic indicator in postoperative follow up. Gut. Jan. 2003;52(1):101-8.

Rygaard, et al. Abnormalities in structure and expression of the retinoblastoma gene in small cell lung cancer cell ines and xenografts in nude mice Cancer Res. Sep. 1, 1990;50(17):5312-7.

Sanger, F. et al. "DNA sequencing with chain-terminating inhibitors" PNAS (1977) 74(12):5463-5467.

Sathirapongsasuti, J.F. et al. "Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV" BioInformatics (2011) 27(19):2648-2654.

Sausen, M. et al. "Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma", Nature Genetics 2013, 45(1), 12-17.

Schmitt et al. Supplemental Information http://www.pnas.org/content/suppl/2012/08/01/1208715109. DCSupplemental.

Schmitt, et al. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A. Sep. 4, 2012;109(36):14508-13. doi: 10.1073/pnas. 1208715109 Epub Aug. 1, 2012.

Schwarzenbach, et al. A critical evaluation of loss of heterozygosity detected in tumor tissues, blood serum and bone marrow plasma from patients with breast cancer. Breast Cancer Res. 2007;9(5):R66.

Schwarzenbach, et al. Cell-free tumor DNA in blood plasma as a marker for circulating tumor cells in prostate cancer. Clin Cancer Res. Feb. 1, 2009;15(3):1032-8. doi: 10.1158/1078-0432.CCR-08-1910.

Schwarzenback, H. et al. "Cell-free nucleic acids as biomarkers in cancer patients" Nature Reviews Cancer (2011) 11:426-437.

Schweiger et al., "Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis," PLoS One 2009,4(5), e5548.

Sebat, et al. Large-Scale Copy number Polymorphism in the Human Genome. Science, 305: 525-528 (2004).

Sehnert, A.J. et al. "Optimal Detection of Fetal Chromosomal Abnormalities by Massively Parallel DNA Sequencing of Cell-Free Fetal DNA from Maternal Blood" Clin Chem (2011) 57(7):1042-1049.

Shaw et al. Genomic analysis of circulating cell-free DNA infers breast cancer dormancy. Genome Research 22 (2):220-231 (Feb. 2012).

Shaw, et al. Microsatellite alterations plasma DNA of primary breast cancer patients. Clin Cancer Res. Mar. 2000;6(3):1119-24.

Shendure, J. et al. "Next-generation DNA sequencing," Nat. Biotechnol. 2008 26(10), 1135-1145.

Shinozaki, et al. Utility of circulating B-RAF DNA mutation in serum for monitoring melanoma patients receiving biochemotherapy. Clin Cancer Res. Apr. 1, 2007;13(7):2068-74.

Shiroguchi, et al. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci U S A. Jan. 24, 2012;109(4):1347-52. doi: 10.1073/pnas.1118018109. Epub Jan. 9, 2012.

Shiroguchi, et al. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes Proc Natl Acad Sci U S A 2012 Supplemental Information (8 pages).

Shoemaker, et al. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. Nature Genetics, 14:450-456 (1996).

Simpson, et al. Copy number variant detection in inbred strains from short read sequence data. Bioinformatics. Feb. 15, 2010;26(4):565-7. doi: 10.1093/bioinformatics/btp693. Epub Dec. 18, 2009.

Smith, et al. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic Acids Research, 38(13): e142 (2010).

Smith, T.F. et al. "Comparison of Biosequences" Adv App Math (1981) 2:482-489.

Sorenson, et al. Soluble normal and mutated DNA sequences from single-copy genes in human blood. Cancer Epidemiol Biomarkers Prev. Jan.-Feb. 1994;3(1):67-71.

Sparks, et al. Selective analysis of cell-free DNA in maternal blood for evaluation of fetal trisomy. Prenat Diagn. Jan. 2012;32(1):3-9. doi: 10.1002/pd.2922. Epub Jan. 6, 2012.

Stein, et al. "The case for cloud computing in genome informatics", Genome Biol. 2010; 11 (5):207. Epub May 5, 2010.

Steinman. Free DNA in serum and plasma from normal adults. J Clin Invest. Aug. 1975;56(2):512-5.

Stitziel, et al. Computational and statistical approaches to analyzing variants identified byexome sequencing. Genome Biol. Sep. 14, 2011;12(9):227. doi: 10.1186/gb-2011-12-9-227.

Stroun, M, et al., "About the possible orgin and mechanism of circulating DNA apoptosis and active DNA release", Clin Chim Acta, vol. 313, No. 1-2, pp. 139-142, (2001).

"Blood Plasma" Oxford Dictionary of Biochemistry and Molecular Biology 81 (2d ed. 2006).

"LB-246/12—Detection of circulating tumor DNA in early stage cancers," Available at https://www.genomeweb.com/cancer/ngs-error-correction-method-described-aacr-excitement-grows-early-cancer-detection. Accessed on Apr. 26, 2017.

"Cohesive End," Oxford Dictionary of Biochemistry and Molecular Biology 132 (2d ed. 2006).

"Control," Oxford Dictionary of Biochemistry and Molecular Biology (2d ed. 2006), 143.

Ajay, S.S. et al., "Accurate and comprehensive sequencing of personal genomes," Genome Res. 2011, 21(9), 1498-1505.

Albert, T.J. et al., "Direct selection of human genomic loci by microarray hybridization," Nat. Methods 2007,4(11), 903-905.

Alkan, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. Oct. 2009;41(10):1061-7. doi: 10.1038/ng.437. Epub Aug. 30, 2009.

Andersson, et al. Bayesian detection of periodic mRNA time profiles without use of training examples. BMC Bioinformatics. Feb. 9, 2006;7:63.

Angeloni, D. Molecular analysis of deletions in human chromosome 3p21 and the role of resident cancer genes in disease Brief Funct Genomic Proteomic Mar. 2007;6(1): 19-39 Epub May 24, 2007.

Ansorge "Next-generation DNA sequencing techniques." New Biotechnology, 25(4):195-203 (2009).

Appel, Maryke. Part II: It's all about conversion. Kapa Biosystems. Accessed Mar. 29, 2017. Printed Apr. 11, 2017. 5 pages. URL:https://www.kapabiosystems.com/ngs/part-ii-conversion/.

Arneson, N et al., "Whole-Genome Amplification by Adaptor-Ligation PCR of Randomly Sheared Genomic DNA (PRSG)," CSH Protocols, 2008, 3(1).

Ashford, Monica. NGS Error Correction Method Described at AACR as Excitement Grows For Early Cancer Detection. Genomeweb.com. Apr. 5, 2017. 3 pages. URL:https://www.genomeweb.com/cancer/ngs-error-correction-method-described-aacr-excitement-grows-early-cancer-detection.

(56) **References Cited**

OTHER PUBLICATIONS

Atanur, et al. The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. Genome Res. Jun. 2010;20(6):791-803. doi: 10.1101/gr.103499.109. Epub Apr. 29, 2010.

Audic, et al. "The Significance of Digital Gene Expression Profiles." Genome Research, 7: 986-995 (1997).

Barzon, et al. Evaluation of circulating thyroid-specific transcripts as markers of thyroid cancer relapse. Int J Cancer. Jul. 20, 2004;110(6):914-20.

Bendich, et al. Circulating DNA as a possible factor in oncogenesis. Science. Apr. 16, 1965;148(3668):374-6.

Bentley, D.R. et al., "Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry," Nature, 2008, 456(7218):53-59.

Bettegowda et al., "Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies," Sci. Transl. Med. 2014, 6(224):224ra24.

Bianchi, D.W. et al. "Genome-Wide Fetal Aneuploidy Detection by Maternal Plasma DNA Sequencing," Obstetrics & Synecology, (2012) 119(5), 1-12.

Bonaldo, et al. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res. Sep. 1996;6(9):791-806.

Bowcock, et al. Exclusion of the retinoblastoma gene and chromosome 13q as the site of a primary lesion for human breast cancer. Am J Hum Genet. Jan. 1990;46(1):12-7.

Braha, et al. Simultaneous stochastic sensing of divalent metal ions. Nature Biotechnology, 18:1005-1007 (2000).

Bremnes, et al. Circulating tumour-derived DNA and RNA markers in blood: a tool for early detection, diagnostics, and follow-up? Lung Cancer. Jul. 2005;49(1):1-12.

Brenner, et al. "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays." Nature Biotechnology, 18: 630-634 (2000).

Brenner, et al. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially axpressed cDNAs. Proc Natl Acad Sci U S A. Feb. 15, 2000;97(4):1665-70.

Bryzgunova,O. et al., "A reliable method to concentrate circulating DNA," Analytical Biochem., 2010, 408:354-356.

Campbell, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing Nat Genet. Jun. 2008;40(6):722-9. doi: 10.1038/ng.128. Epub Apr. 27, 2008.

Caramazza, et al. Chromosome 1 abnormalities in myeloid malignancies: a literature survey and karyotype-phenotype associations. Eur J Haematol. Mar. 2010;84(3):191-200. doi: 10.1111/j.1600-0609.2009.01392.x. Epub Nov. 30, 2009.

Carr, et al. Inferring relative proportions of DNA variants from sequencing electropherograms. Bioinformatics. Dec. 15, 2009;25(24):3244-50. doi: 10.1093/bioinformatics/btp583. Epub Oct. 9, 2009.

Casava v1.8.2 User Guide http://gensoft.pasteur.fr/docs/casava/18.2/CASAVA_1_8_2_UG_15011196C.pdf (Oct. 2011).

Casbon, et al. A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res. Jul. 2011;39(12):e81. doi: 10.1093/nar/gkr217. Epub Apr. 13, 2011.

Castle, et al. DNA copy number including telomeres and mitochondria, assayed using next-generation sequencing. BMC Genomics. Apr. 16, 2010;11:244. doi: 10 1186/1471-2164-11-244.

Chang, et al. Detection of allelic imbalance in ascitic supernatant by digital single nucleotide polymorphism analysis. Clin Cancer Res. Aug. 2002;8(8):2580-5.

Chee, et al. "Accessing genetic information with high-density DNA arrays." Science, 274: 610-614 (1996).

Chee. "Enzymatic multiplex DNA sequencing." Nucleic Acids Research, 19(12): 3301-3305 (1991).

Chen, et al. Microsatellite alterations in plasma DNA of small cell lung cancer patients. Nat Med. Sep. 1996;2(9):1033-5.

Chin, et al. A SNP in a let-7 microRNA complementary site in the KRAS 3' untranslated region increases non-small cell lung cancer risk Cancer Res. Oct. 15, 2008;68(20):8535-40. doi: 10.1158/0008-5472.CAN-08-2129.

Chiu, et al. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study BMJ Jan. 11, 2011;342:c7401 doi: 10.1136/bmj.c7401.

Chiu, et al. Quantitative analysis of circulating mitochondrial DNA in plasma. Clin Chem. May 2003;49(5):719-26.

Chiu, R.W.K. et al. "Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma" PNAS (2008) 105(51):20458-20463.

Church, et al. "Multiplex DNA sequencing." Science, 240:185-188 (1988).

Clark, T.A. et al. "Analytical Validation of a Hybrid Capture Based Next-Generation Sequencing Clinical Assay for Genomic Profiling of Cell-Free Circulating Tumor DNA," J. Mol. Diagnostics (2018) 20(5):686-702.

Cook, et al. Methylated DNA labels for marking objects. Biotechnol Lett. Jan. 2003;25(1):89-94.

Costello, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res. Apr. 1, 2013;41(6):e67. doi: 10.1093/nar/gks1443. Epub Jan. 8, 2013.

Coulet, et al. Detection of plasma tumor DNA in head and neck squamous cell carcinoma by microsatellite typing and p53 mutation analysis. Cancer Res Feb. 1, 2000;60(3):707-11.

Cox, J. "Bar coding objects with DNA." Analyst. May 2001;126(5):545-7.

Cox. Bar coding objects with DNA. Analyst. May 2001;126(5):545-7.

Craig, et al. Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods. Oct. 2008;5(10):887-93. doi: 10.1038/nmeth.1251. Epub Sep. 14, 2008.

Daines, et al. High-throughput multiplex sequencing to discover copy number variants in *Drosophila*. Genetics. Aug. 2009;182(4):935-41. doi: 10.1534/genetics.109.103218. Epub Jun. 15, 2009.

Zhang et al. "Comprehensive One-Step Molecular Analysis of Mitochondrial Genome by Massively Parallel Sequencing" Clinical Chem (2012) 58(9):1322-1331.

Zhang, et al. The impact of next-generation sequencing on genomics. J Genet Genomics. Mar. 20, 2011;38(3):95-109. doi: 10.1016/j.jgg.2011.02.003. Epub Mar. 15, 2011.

Zhao, et al. Homozygous Deletions and Chromosome Amplifications in Human Lung Carcinomas Revealed by Single Nucleotide Polymorphism Array Analysis. Cancer Research, 65: 5561-5570 (2005).

Zhou, et al. Counting alleles reveals a connection between chromosome 18q loss and vascularinvasion. Nature Biotechnology, 19: 78-81 (2001).

Jiang et al. "Basics in Bioinformatics Lecture Notes of the Graduate Summer School on Bioinformatics of China" Springer Heidelberg New York Dordrecht London (2013).

Opposition Form and Statement to EP3378952 filed Nov. 4, 2020 by Dirk Buhler.

Opposition Form and Statement to EP3378952 filed Nov. 5, 2020 by Foundation Medicine Inc.

Opposition Form and Statement to EP3378952 filed Nov. 5, 2020 by Grunecker.

Crowley, E. et al. "Liquid biopsy: monitoring cancer-genetics in the blood" Nat Rev Clin Oncology (2013) 8:472-478.

Greaves, L.C. et al. "Quantification of mitochondrial DNA mutation load" Aging Cell (2009) 8(5): 566-572.

Guardant Health, Inc. Response to Notice of Opposition in EP3378952 filed Mar. 29, 2021.

Guardant Health, Inc. v. *FMI* Defendant's Answer and Counter Claims, filed Jan. 14, 2021 (C.A. No. 20-1580 (LPS)).

Guardant Health, Inc. v. *FMI* Defendant's Answering Brief in Opposition to Plaintiff Guardant Health, Inc's Motion for Preliminary Injunction, filed Mar. 5, 2021 (C.A. No. 20-1580 (LPS)).

Opposition to EP3524694, dated Apr. 15, 2021 by Foundation Medicine, Inc.

(56) **References Cited**

OTHER PUBLICATIONS

Opposition to EP3524694, filed Apr. 16, 2021 by Maiwald International.
Extended European search report and opinion dated Dec. 7, 2020 for EP Application No. 20183626.9.
Office action dated Jun. 23, 2021 for U.S. Appl. No. 16/913,965.

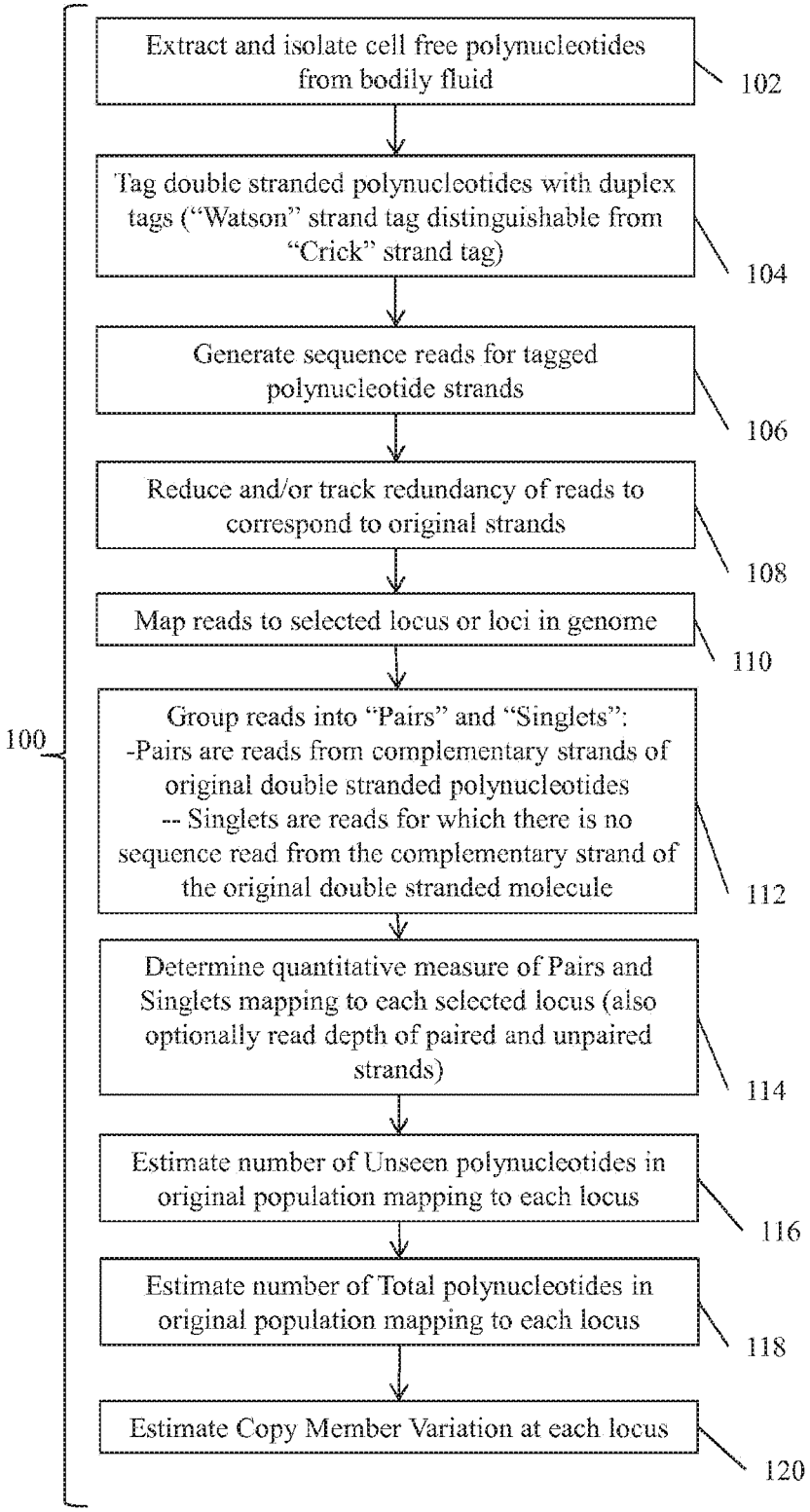

Extract and isolate cell free polynucleotides from bodily fluid
102
Tag double stranded polynucleotides with duplex tags ("Watson" strand tag distinguishable from "Crick" strand tag)
104
Generate sequence reads for tagged polynucleotide strands
106
Reduce and/or track redundancy of reads to correspond to original strands
108
Map reads to selected locus or loci in genome
110
100
Group reads into "Pairs" and "Singlets":
-Pairs are reads from complementary strands of original double stranded polynucleotides
-- Singlets are reads for which there is no sequence read from the complementary strand of the original double stranded molecule
112
Determine quantitative measure of Pairs and Singlets mapping to each selected locus (also optionally read depth of paired and unpaired strands)
114
Estimate number of Unseen polynucleotides in original population mapping to each locus
116
Estimate number of Total polynucleotides in original population mapping to each locus
118
Estimate Copy Member Variation at each locus
120

Fig. 1

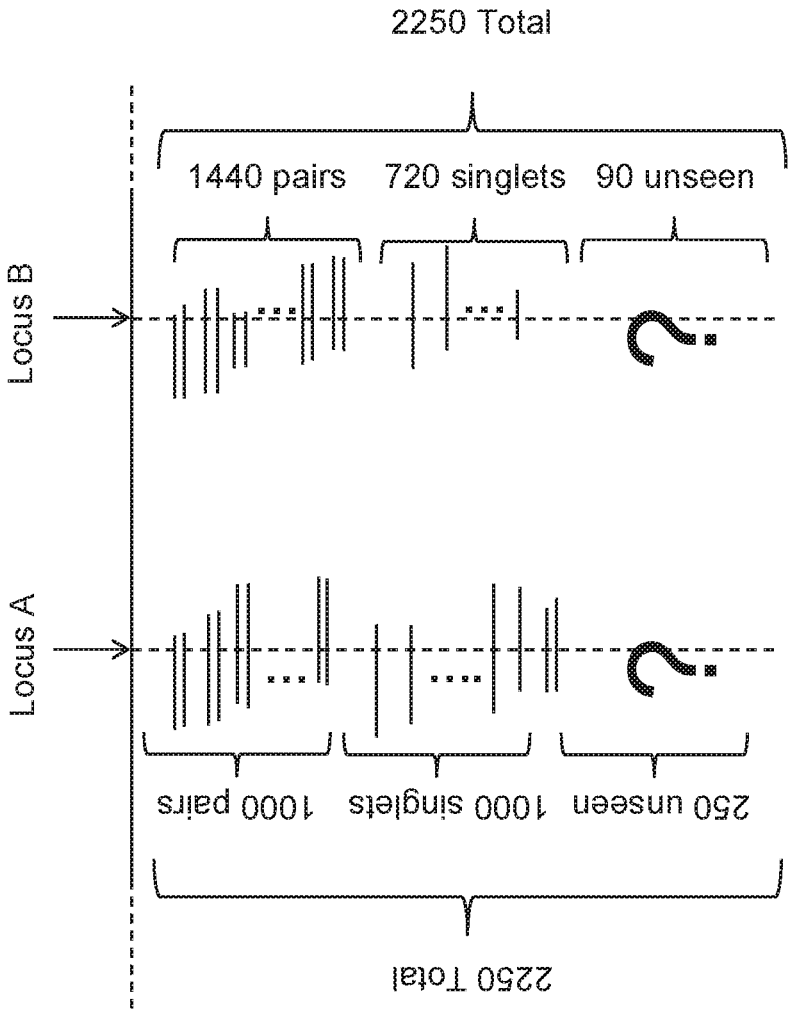

2250 Total
1440 pairs
720 singlets
90 unseen
Locus B
Locus A
1000 pairs
1000 singlets
250 unseen
2250 Total

Fig. 2

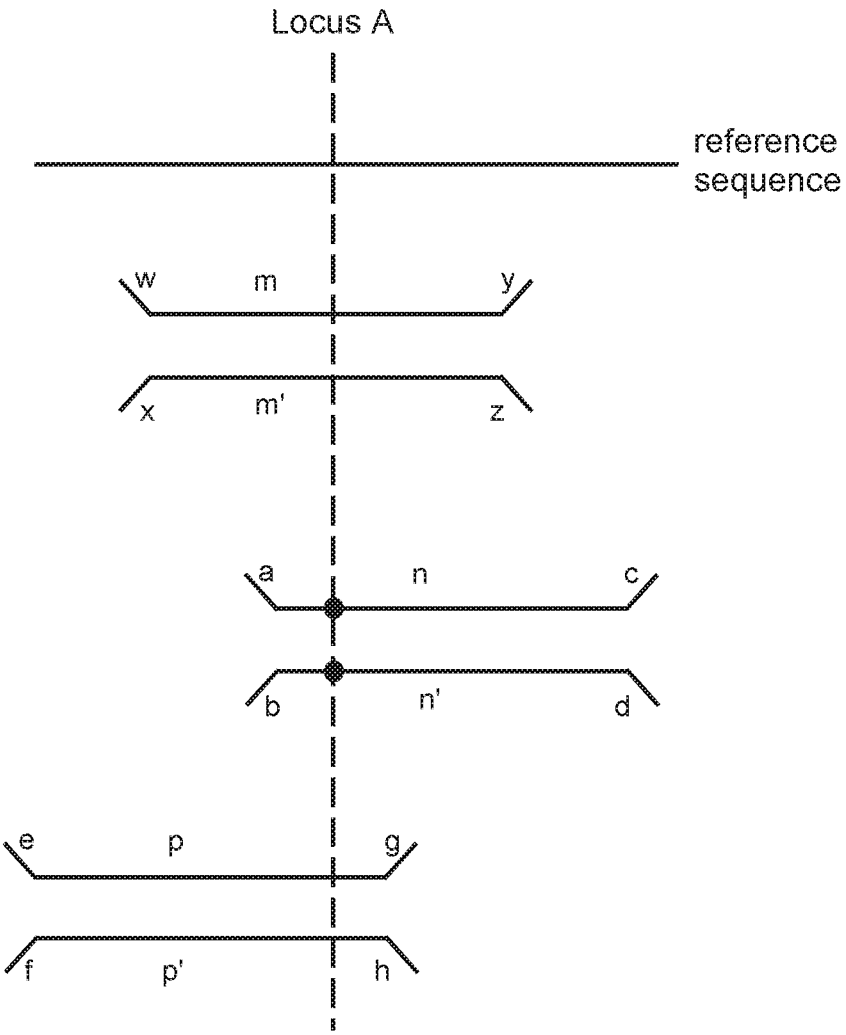


Fig. 3

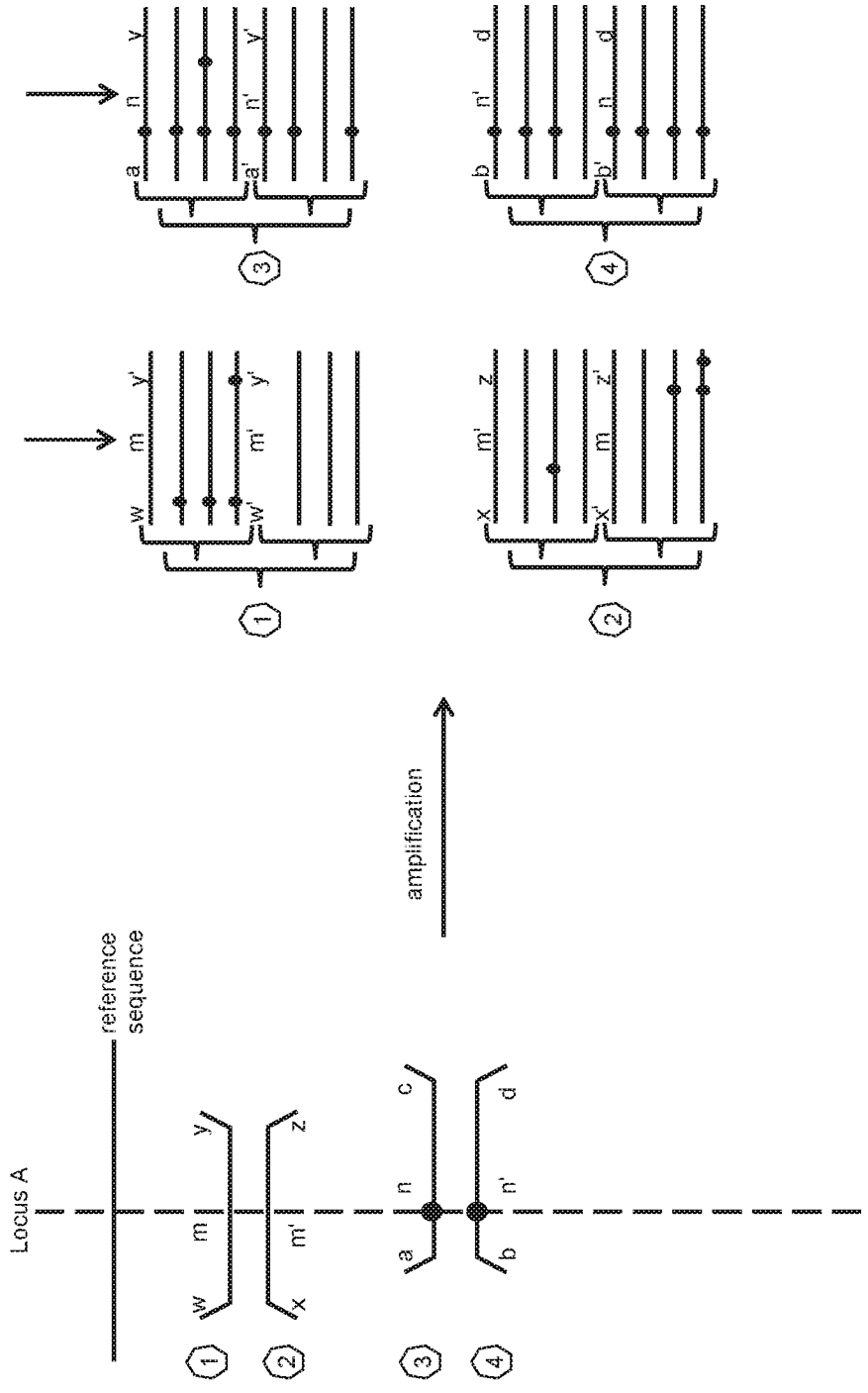

Locus A
reference sequence
amplification
w
m
y
x
m'
z
a
n
c
b
n'
d
1
2
3
4

Fig. 4A

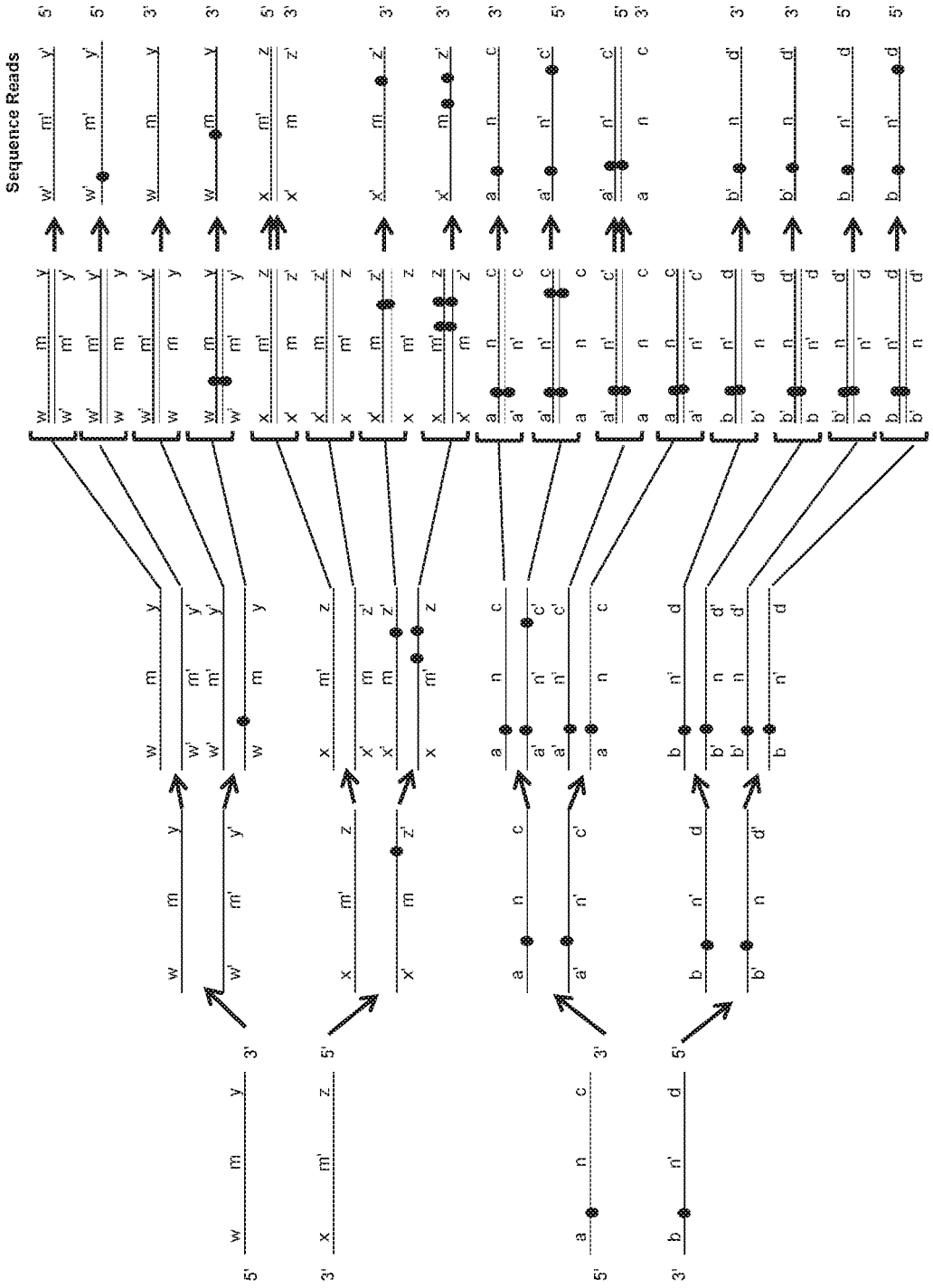


Fig. 4B

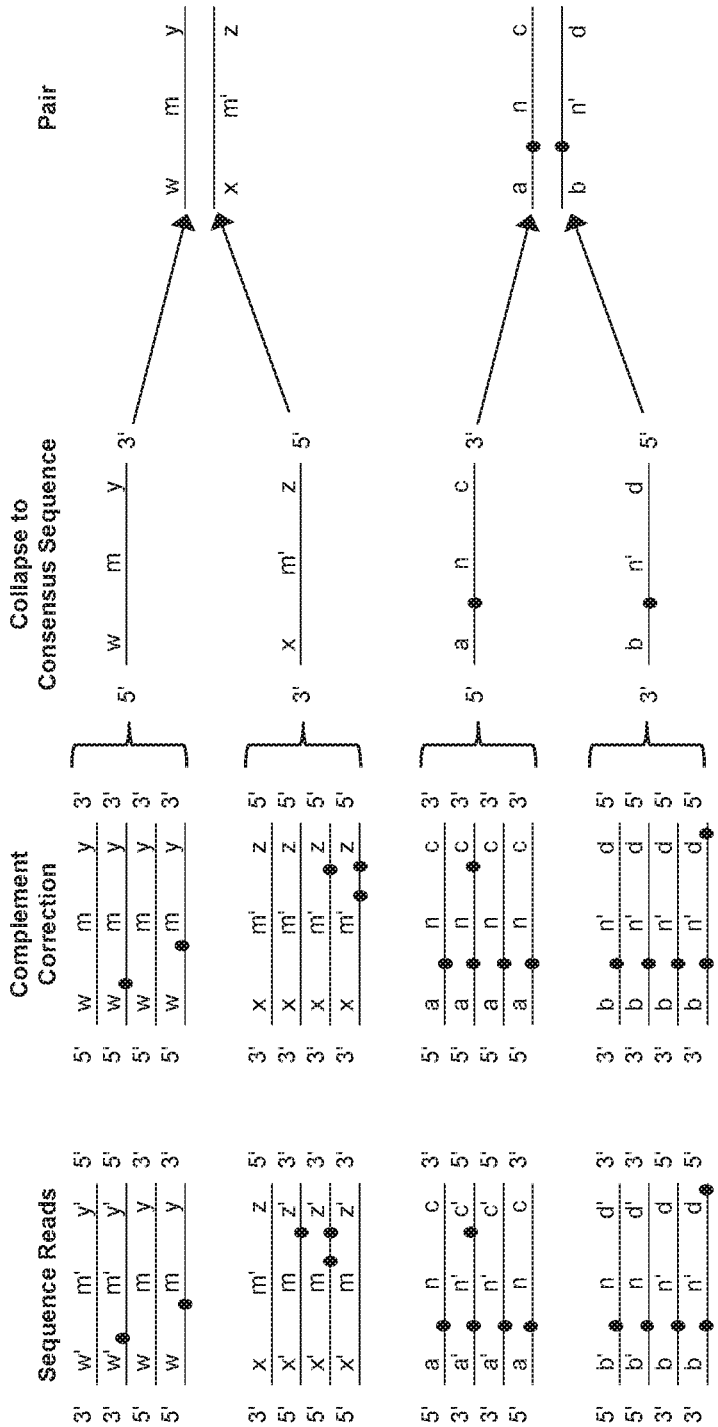

Sequence Reads
Complement Correction
Collapse to Consensus Sequence
Pair

Fig. 4C

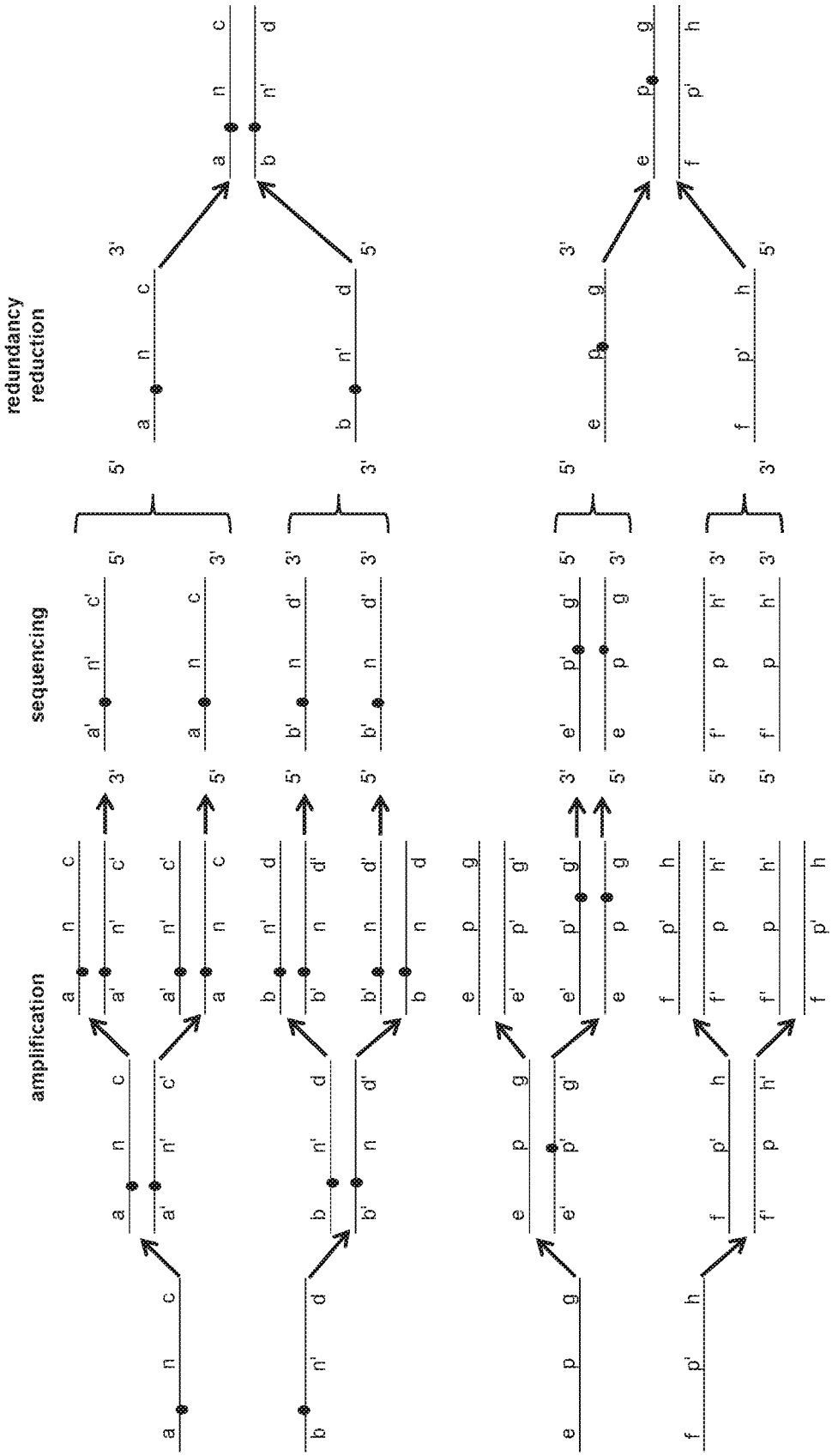


Fig. 5

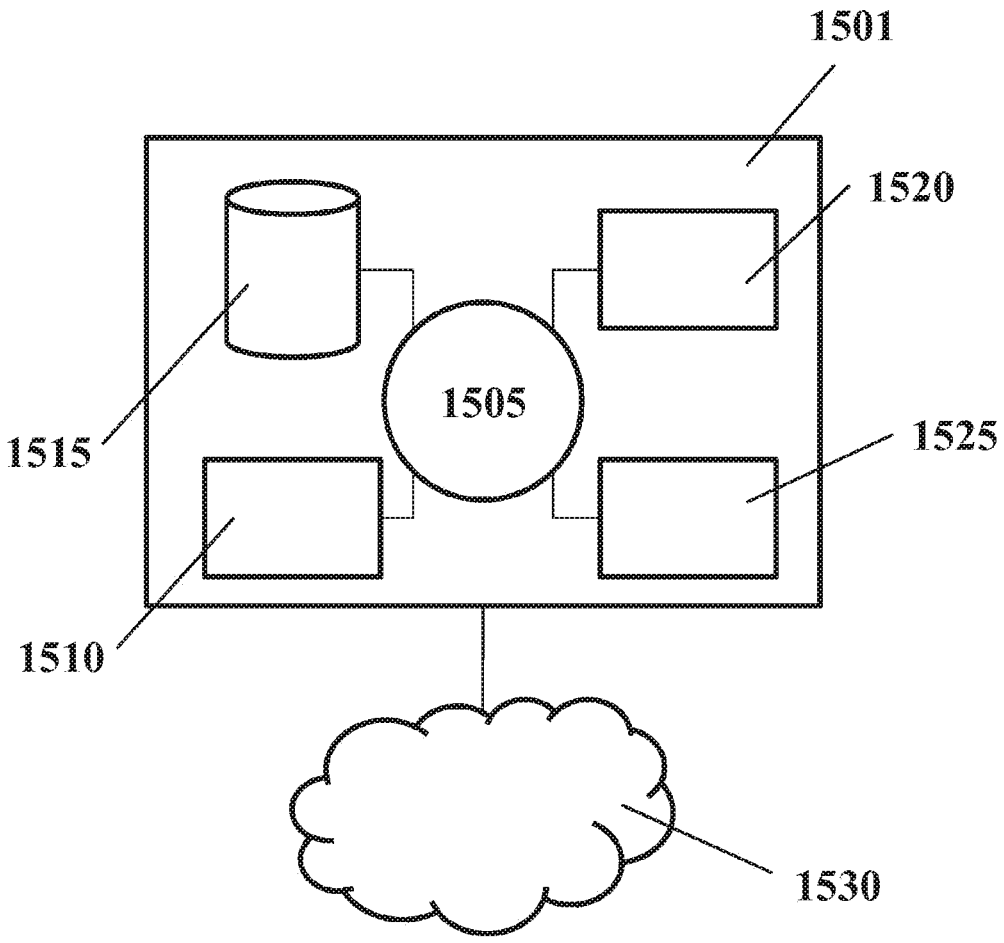


Fig. 6

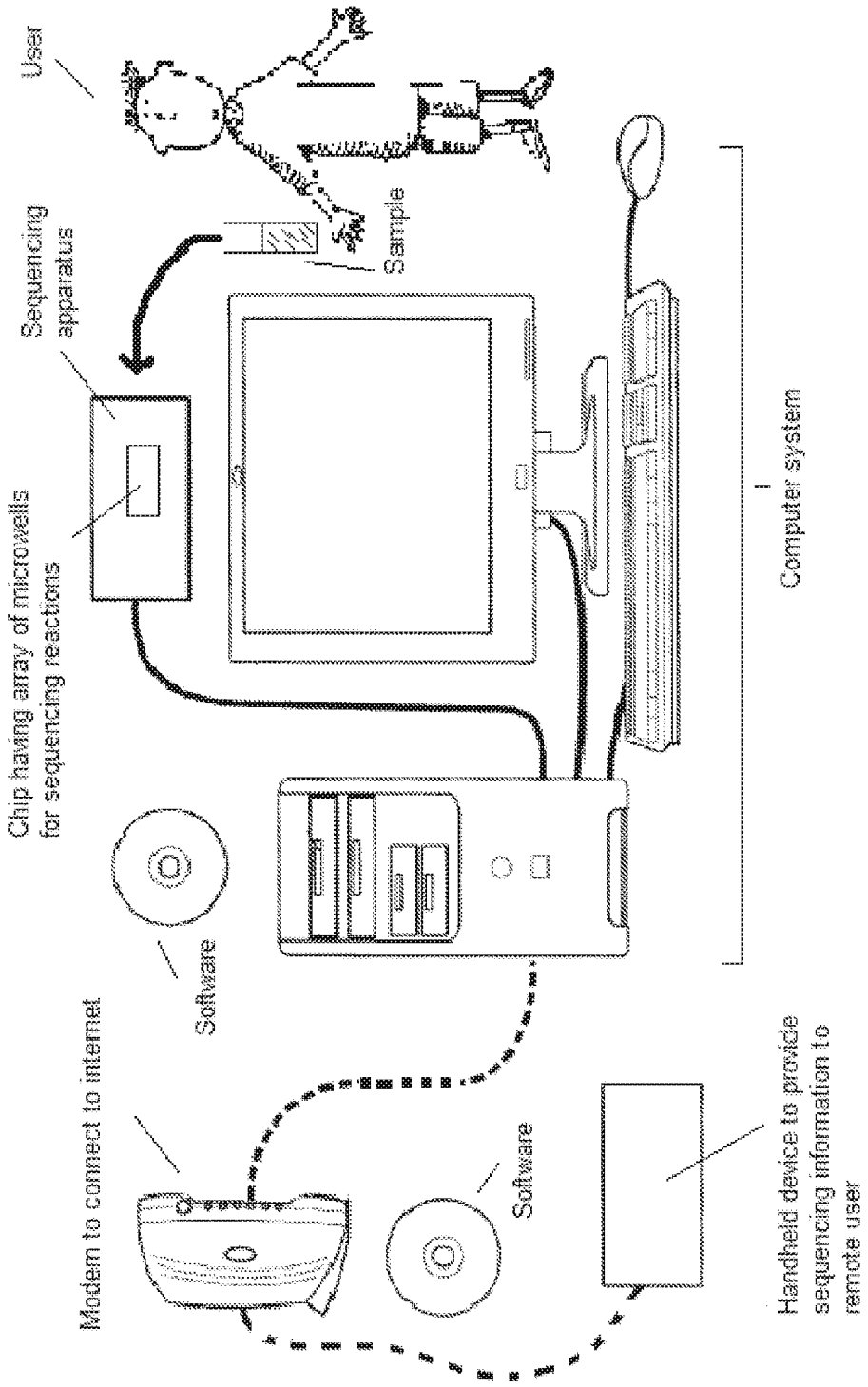

User
Sample
Sequencing apparatus
Chip having array of microwells for sequencing reactions
Computer system
Software
Modem to connect to internet
Software
Handheld device to provide sequencing information to remote user

Fig. 7

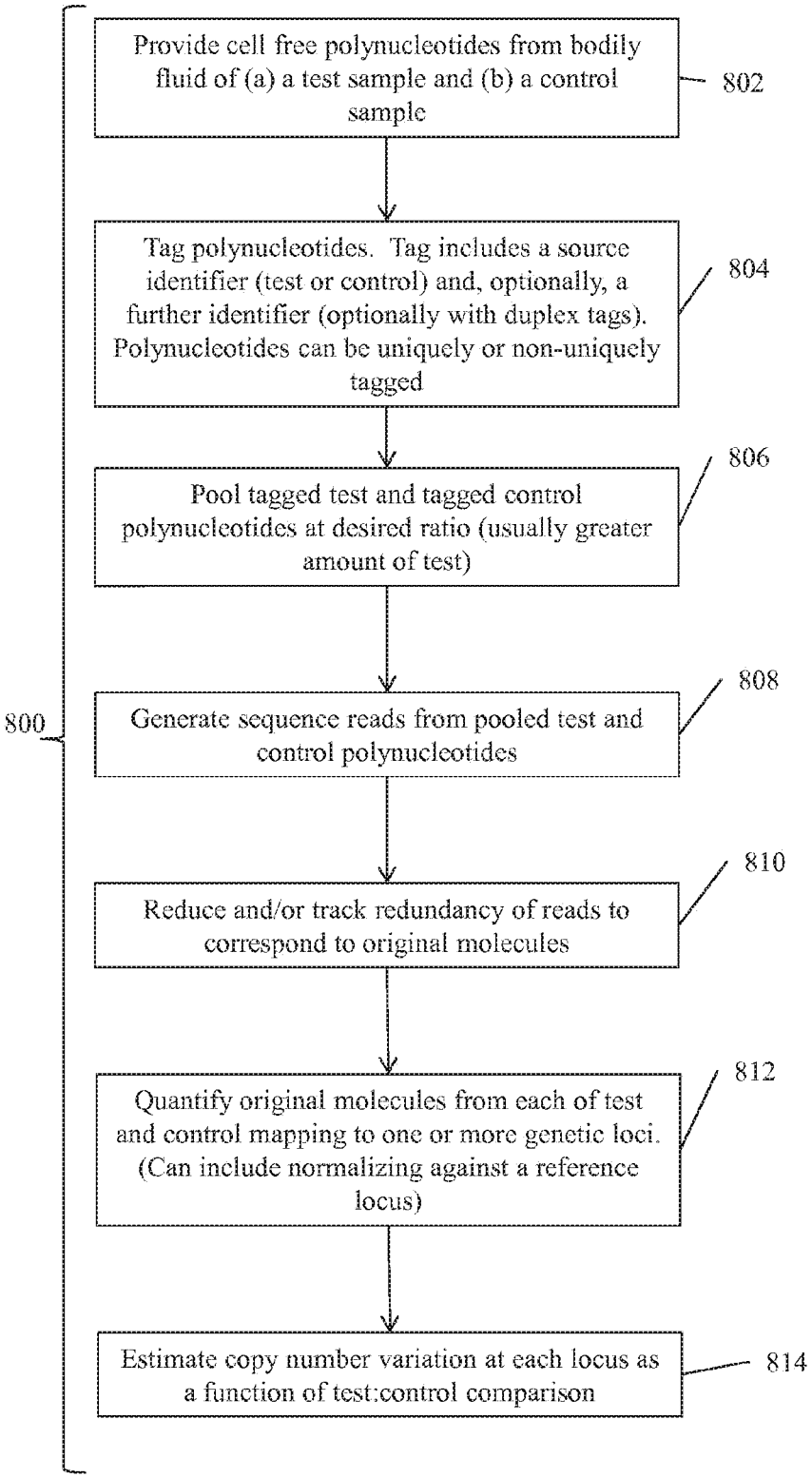

Provide cell free polynucleotides from bodily fluid of (a) a test sample and (b) a control sample
802
Tag polynucleotides. Tag includes a source identifier (test or control) and, optionally, a further identifier (optionally with duplex tags). Polynucleotides can be uniquely or non-uniquely tagged
804
Pool tagged test and tagged control polynucleotides at desired ratio (usually greater amount of test)
806
Generate sequence reads from pooled test and control polynucleotides
808
800
Reduce and/or track redundancy of reads to correspond to original molecules
810
Quantify original molecules from each of test and control mapping to one or more genetic loci. (Can include normalizing against a reference locus)
812
Estimate copy number variation at each locus as a function of test:control comparison
814

Fig. 8

Fig. 9A

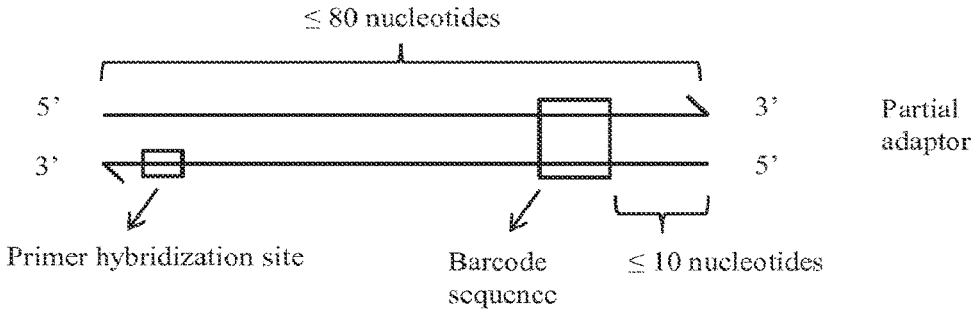


Fig. 9B

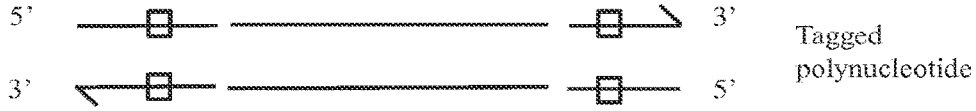
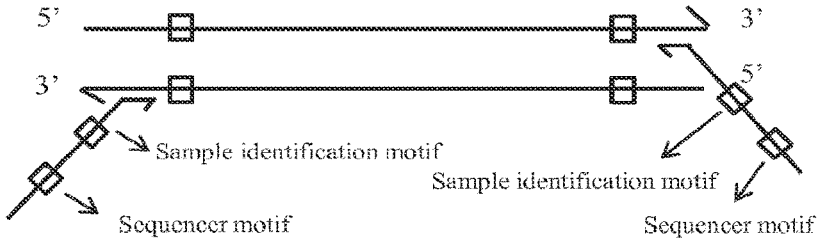


Fig. 9C



METHODS AND SYSTEMS FOR DETECTING GENETIC VARIANTS

CROSS-REFERENCE

This application is a continuation of U.S. application Ser. No. 16/601,168, filed Oct. 14, 2019 (now U.S. Pat. No. 10,801,063, issued Oct. 13, 2020), which is a continuation of U.S. application Ser. No. 15/892,178, filed Feb. 8, 2018 (now U.S. Pat. No. 10,883,139, issued Jan. 5, 2021), which is a continuation of U.S. application Ser. No. 14/861,989, filed Sep. 22, 2015 (now U.S. Pat. No. 9,920,366, issued Mar. 20, 2018), which is a continuation application of International Application No. PCT/US2014/072383, filed Dec. 24, 2014, which application claims the benefit under 35 U.S.C. § 119(e) of U.S. Provisional Application No. 61/921,456, filed Dec. 28, 2013, and U.S. Provisional Application No. 61/948,509, filed Mar. 5, 2014, each of which is entirely incorporated herein by reference.

BACKGROUND

The detection and quantification of polynucleotides is important for molecular biology and medical applications, such as diagnostics. Genetic testing is particularly useful for a number of diagnostic methods. For example, disorders that are caused by rare genetic alterations (e.g., sequence variants) or changes in epigenetic markers, such as cancer and partial or complete aneuploidy, may be detected or more accurately characterized with DNA sequence information.

Early detection and monitoring of genetic diseases, such as cancer, is often useful and needed in the successful treatment or management of the disease. One approach may include the monitoring of a sample derived from cell-free nucleic acids, a population of polynucleotides that can be found in different types of bodily fluids. In some cases, disease may be characterized or detected based on detection of genetic aberrations, such as copy number variation and/or sequence variation of one or more nucleic acid sequences, or the development of other certain rare genetic alterations. Cell-free DNA (cfDNA) may contain genetic aberrations associated with a particular disease. With improvements in sequencing and techniques to manipulate nucleic acids, there is a need in the art for improved methods and systems for using cell-free DNA to detect and monitor disease.

In particular, many methods have been developed for accurate copy number variation estimation, especially for heterogeneous genomic samples, such as tumor-derived gDNA or for cfDNA for many applications (e.g., prenatal, transplant, immune, metagenomics or cancer diagnostics). Most of these methods include sample preparation whereby the original nucleic acids are converted into a sequenceable library, followed by massively parallel sequencing, and finally bioinformatics to estimate copy number variation at one or more loci.

SUMMARY

Although many of these methods are able to reduce or combat the errors introduced by the sample preparation and sequencing processes for all molecules that are converted and sequenced, these methods are not able to infer the counts of molecules that were converted but not sequenced. Since this count of converted by unsequenced molecules can be highly variable from genomic region to region, these counts can dramatically and adversely affect the sensitivity that can be achieved.

To address this issue, input double-stranded deoxyribonucleic acid (DNA) can be converted by a process that tags both halves of the individual double-stranded molecule, in some cases differently. This can be performed using a variety of techniques, including ligation of hairpin, bubble, or forked adapters or other adaptors having double-stranded and single stranded segments (the unhybridized portion of a bubble, forked or hairpin adapter are deemed single-stranded herein). If tagged correctly, each original Watson and Crick (i.e., strand) side of the input double-stranded DNA molecule can be differently tagged and identified by the sequencer and subsequent bioinformatics. For all molecules in a particular region, counts of molecules where both Watson and Crick sides were recovered ("Pairs") versus those where only one half was recovered ("Singlets") can be recorded. The number of unseen molecules can be estimated based on the number of Pairs and Singlets detected.

An aspect of the present disclosure provides a method for detecting and/or quantifying rare deoxyribonucleic acid (DNA) in a heterogeneous population of original DNA fragments, comprising tagging the original DNA fragments in a single reaction using a library of a plurality of different tags such that greater than 30% of the fragments are tagged at both ends, wherein each of the tags comprises a molecular barcode. The single reaction can be in a single reaction vessel. Greater than 50% of the fragments can be tagged at both ends. The plurality of different tags can be no more than any of 100, 500, 1000, 10,000 or 100,000 different tags.

Another aspect provides a set of library adaptors that can be used to tag the molecules of interest (e.g., by ligation, hybridization, etc.). The set of library adaptors can comprise plurality of polynucleotide molecules with molecular barcodes, wherein the plurality of polynucleotide molecules are less than or equal to 80 nucleotide bases in length, wherein the molecular barcodes are at least 4 nucleotide bases in length, and wherein (a) the molecular barcodes are different from one another and have an edit distance of at least 1 between one another; (b) the molecular barcodes are located at least one nucleotide base away from a terminal end of their respective polynucleotide molecules; (c) optionally, at least one terminal base is identical in all of the polynucleotide molecules; and (d) none of the polynucleotide molecules contains a complete sequencer motif.

In some embodiments, the library adaptors (or adapters) are identical to one another but for the molecular barcodes. In some embodiments, each of the plurality of library adaptors comprises at least one double-stranded portion and at least one single-stranded portion (e.g., a non-complementary portion or an overhang). In some embodiments, the double-stranded portion has a molecular barcode selected from a collection of different molecular barcodes. In some embodiments, the given molecular barcode is a randomer. In some embodiments, each of the library adaptors further comprises a strand-identification barcode on the at least one single-stranded portion. In some embodiments, the strand-identification barcode includes at least 4 nucleotide bases. In some embodiments, the single-stranded portion has a partial sequencer motif. In some embodiments, the library adaptors do not include a complete sequencer motif.

In some embodiments, none of the library adaptors contains a sequence for hybridizing to a flow cell or forming a hairpin for sequencing.

In some embodiments, all of the library adaptors have a terminal end with nucleotide(s) that are the same. In some embodiments, the identical terminal nucleotide(s) are over two or more nucleotide bases in length.

In some embodiments, each of the library adapters is Y-shaped, bubble shaped or hairpin shaped. In some embodiments, none of the library adapters contains a sample identification motif. In some embodiments, each of the library adapters comprises a sequence that is selectively hybridizable to a universal primer. In some embodiments, each of the library adapters comprises a molecular barcode that is at least 5, 6, 7, 8, 9 and 10 nucleotide bases in length. In some embodiments, each of the library adapters is from 10 nucleotide bases to 80 in length, or 30 to 70 nucleotide bases in length, or 40 to 60 nucleotide bases in length. In some embodiments, at least 1, 2, 3, or 4 terminal bases are identical in all of the library adaptors. In some embodiments, at least 4 terminal bases are identical in all of the library adaptors.

In some embodiments, the edit distance of the molecular barcodes of the library adapters is a Hamming distance. In some embodiments, the edit distance is at least 1, 2, 3, 4 or 5. In some embodiments, the edit distance is with respect to individual bases of the plurality of polynucleotide molecules. In some embodiments, the molecular barcodes are located at least 10 nucleotide base away from a terminal end of an adapter. In some embodiments, the plurality of library adapters includes at least 2, 4, 6, 8, 10, 20, 30, 40 or 50 different molecular barcodes, or from 2-100, 4-80, 6-60 or 8-40 different molecular barcodes. In any of the embodiments herein, there are more polynucleotides (e.g., cfDNA fragments) to be tagged than there are different molecular barcodes such that the tagging is not unique.

In some embodiments, the terminal end of an adaptor is configured for ligation (e.g., to a target nucleic acid molecule). In some embodiments, the terminal end of an adaptor is a blunt end.

In some embodiments, the adaptors are purified and isolated. In some embodiments, the library comprises one or more non-naturally occurring bases.

In some embodiments, the polynucleotide molecules comprise a primer sequence positioned 5' with respect to the molecular barcodes.

In some embodiments, the set of library adaptors consists essentially of the plurality of polynucleotide molecules.

In another aspect, a method comprises (a) tagging a collection of polynucleotides with a plurality of polynucleotide molecules from a library of adaptors to create a collection of tagged polynucleotides; and (b) amplifying the collection of tagged polynucleotides in the presence of sequencing adaptors, wherein the sequencing adaptors have primers with nucleotide sequences that are selectively hybridizable to complementary sequences in the plurality of polynucleotide molecules. The library of adaptors may be as described above or elsewhere herein. In some embodiments, each of the sequencer adaptors further comprises an index tag, which can be a sample identification motif.

Another aspect, provides a method for detecting and/or quantifying rare DNA in a heterogeneous population of original DNA fragments, wherein the rare DNA has a concentration that is less than 1%, the method comprising (a) tagging the original DNA fragments in a single reaction such that greater than 30% of the original DNA fragments are tagged at both ends with library adaptors that comprise molecular barcodes, thereby providing tagged DNA fragments; (b) performing high-fidelity amplification on the tagged DNA fragments; (c) optionally, selectively enriching a subset of the tagged DNA fragments; (d) sequencing one or both strands of the tagged, amplified and optionally selectively enriched DNA fragments to obtain sequence reads comprising nucleotide sequences of the molecular barcodes and at least a portion of the original DNA fragments; (e) from the sequence reads, determining consensus reads that are representative of single-strands of the original DNA fragments; and (f) quantifying the consensus reads to detect and/or quantify the rare DNA at a specificity that is greater than 99.9%.

In some embodiments, (e) comprises comparing sequence reads having the same or similar molecular barcodes and the same or similar end of fragment sequences. In some embodiments, the comparing further comprises performing a phylogentic analysis on the sequence reads having the same or similar molecular barcodes. In some embodiments, the molecular barcodes include a barcode having an edit distance of up to 3. In some embodiments, the end of fragment sequence includes fragment sequences having an edit distance of up to 3.

In some embodiments, the method further comprises sorting sequence reads into paired reads and unpaired reads, and quantifying a number of paired reads and unpaired reads that map to each of one or more genetic loci.

In some embodiments, the tagging occurs by having an excess amount of library adaptors as compared to original DNA fragments. In some embodiments, n the excess is at least a 5-fold excess. In some embodiments, the tagging comprises using a ligase. In some embodiments, the tagging comprises attachment to blunt ends.

In some embodiments, the method further comprises binning the sequence reads according to the molecular barcodes and sequence information from at least one end of each of the original DNA fragments to create bins of single stranded reads. In some embodiments, the method further comprises, in each bin, determining a sequence of a given original DNA fragment among the original DNA fragments by analyzing sequence reads. In some embodiments, the method further comprises detecting and/or quantifying the rare DNA by comparing a number of times each base occurs at each position of a genome represented by the tagged, amplified, and optionally enriched DNA fragments.

In some embodiments, the library adaptors do not contain complete sequencer motifs. In some embodiments, the method further comprises selectively enriching a subset of the tagged DNA fragments. In some embodiments, the method further comprises, after enriching, amplifying the enriched tagged DNA fragments in the presence of sequencing adaptors comprising primers. In some embodiments, (a) provides tagged DNA fragments having from 2 to 1000 different combinations of molecular barcodes.

In some embodiments, the DNA fragments are tagged with polynucleotide molecules from a library of adaptors as described above or elsewhere herein.

In another aspect, a method for processing and/or analyzing a nucleic acid sample of a subject comprises (a) exposing polynucleotide fragments from the nucleic acid sample to a set of library adaptors to generate tagged polynucleotide fragments; and (b) subjecting the tagged polynucleotide fragments to nucleic acid amplification reactions under conditions that yield amplified polynucleotide fragments as amplification products of the tagged polynucleotide fragments. The set of library adaptors comprises a plurality of polynucleotide molecules with molecular barcodes, wherein the plurality of polynucleotide molecules are less than or equal to 80 nucleotide bases in length, wherein the molecular barcodes are at least 4 nucleotide bases in length, and wherein (1) the molecular barcodes are different from one another and have an edit distance of at least 1 between one another; (2) the molecular barcodes are located at least one nucleotide base away from a terminal end of

their respective polynucleotide molecules; (3) optionally, at least one terminal base is identical in all of the polynucleotide molecules; and (4) none of the polynucleotide molecules contains a complete sequencer motif.

In some embodiments, the method further comprises determining nucleotide sequences of the amplified tagged polynucleotide fragments. In some embodiments, the nucleotide sequences of the amplified tagged polynucleotide fragments are determined without polymerase chain reaction (PCR). In some embodiments, the method further comprises analyzing the nucleotide sequences with a programmed computer processor to identify one or more genetic variants in the nucleotide sample of the subject. In some embodiments, the one or more genetic variants are selected from the group consisting of base change(s), insertion(s), repeat(s), deletion(s), copy number variation(s) and transversion(s). In some embodiments, the one or more genetic variants include one or more tumor associated genetic alterations.

In some embodiments, the subject has or is suspected of having a disease. In some embodiments, the disease is cancer. In some embodiments, the method further comprises collecting the nucleic acid sample from the subject. In some embodiments, the nucleic acid sample is collected from a location selected from the group consisting of blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, cerebral spinal fluid and tears of the subject. In some embodiments, the nucleic acid sample is a cell-free nucleic acid sample. In some embodiments, the nucleic acid sample is collected from no more than 100 nanograms (ng) of double-stranded polynucleotide molecules of the subject.

In some embodiments, the polynucleotide fragments comprise double-stranded polynucleotide molecules. In some embodiments, in (a), the plurality of polynucleotide molecules couple to the polynucleotide fragments via blunt end ligation, sticky end ligation, molecular inversion probes, PCR, ligation-based PCR, multiplex PCR, single stranded ligation, and single stranded circularization. In some embodiments, exposing the polynucleotide fragments of the nucleic acid sample to the plurality of polynucleotide molecules yields the tagged polynucleotide fragments with a conversion efficiency of at least 10%. In some embodiments, any of at least 5%, 6%, 7%, 8%, 9%, 10%, 20%, or 25% of the tagged polynucleotide fragments share a common polynucleotide molecule or sequence. In some embodiments, the method further comprises generating the polynucleotide fragments from the nucleic acid sample.

In some embodiments, the subjecting comprises amplifying the tagged polynucleotide fragments from sequences corresponding to genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, REHB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

In another aspect, a method comprises (a) generating a plurality of sequence reads from a plurality of polynucleotide molecules, wherein the plurality of polynucleotide molecules cover genomic loci of a target genome, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1; (b) grouping with a computer processor the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides; (c) for each of the families, merging sequence reads to generate a consensus sequence; (d) calling the consensus sequence at a given genomic locus among the genomic loci; and (e) detecting at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls, and total number of alterations among the calls.

In some embodiments, each family comprises sequence reads from only one of the template polynucleotides. In some embodiments, the given genomic locus comprises at least one nucleic acid base. In some embodiments, the given genomic locus comprises a plurality of nucleic acid bases. In some embodiments, the calling comprises calling at least one nucleic acid base at the given genomic locus. In some embodiments, the calling comprises calling a plurality of nucleic acid bases at the given genomic locus. In some embodiments, the calling comprises any one of phylogenetic analysis, voting, weighing, assigning a probability to each read at the locus in a family and calling the base with the highest probability.

In some embodiments, the method further comprises performing (d)-(e) at an additional genomic locus among the genomic loci. In some embodiments, the method further comprises determining a variation in copy number at one of the given genomic locus and additional genomic locus based on counts at the given genomic locus and additional genomic locus.

In some embodiments, the grouping comprises classifying the plurality of sequence reads into families by identifying (i) different molecular barcodes coupled to the plurality of polynucleotide molecules and (ii) similarities between the plurality of sequence reads, wherein each family includes a plurality of nucleic acid sequences that are associated with a different combination of molecular barcodes and similar or identical sequence reads. Different molecular barcodes have different sequences.

In some embodiments, the consensus sequence is generated by evaluating a quantitative measure or a statistical significance level for each of the sequence reads. In some embodiments, the quantitative measure comprises use of a binomial distribution, exponential distribution, beta distribution, or empirical distribution. In some embodiments, the method further comprises mapping the consensus sequence to the target genome. In some embodiments, the plurality of genes includes at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50 or all of the plurality of genes selected from the group.

Another aspect of the present disclosure provides a method, comprising (a) providing template polynucleotide molecules and a set of library adaptors in a single reaction vessel, wherein the library adaptors are polynucleotide molecules that have different molecular barcodes (e.g., from 2 to 1,000 different molecular barcodes), and wherein none of the library adaptors contains a complete sequencer motif; (b) in the single reaction vessel, coupling the library adaptors to

the template polynucleotide molecules at an efficiency of at least 10%, thereby tagging each template polynucleotide with a tagging combination that is among a plurality of different tagging combinations (e.g., 4 to 1,000,000 different tagging combinations), to produce tagged polynucleotide molecules; (c) subjecting the tagged polynucleotide molecules to an amplification reaction under conditions that yield amplified polynucleotide molecules as amplification products of the tagged polynucleotide molecules; and (d) sequencing the amplified polynucleotide molecules.

In some embodiments, the template polynucleotide molecules are blunt ended or sticky-ended. In some embodiments, the library adaptors are identical but for the molecular barcodes. In some embodiments, each of the library adaptors has a double stranded portion and at least one single-stranded portion. In some embodiments, the double-stranded portion has a molecular barcode among the molecular barcodes. In some embodiments, each of the library adaptors further comprises a strand-identification barcode on the at least one single-stranded portion. In some embodiments, the single-stranded portion has a partial sequencer motif. In some embodiments, the library adaptors have a sequence of terminal nucleotides that are the same. In some embodiments, the template polynucleotide molecules are double-stranded. In some embodiments, the library adaptors couple to both ends of the template polynucleotide molecules.

In some embodiments, subjecting the tagged polynucleotide molecules to the amplification reaction comprises non-specifically amplifying the tagged polynucleotide molecules.

In some embodiments, the amplification reaction comprises use of a priming site to amplify each of the tagged polynucleotide molecules. In some embodiments, the priming site is a primer. In some embodiments, the primer is a universal primer. In some embodiments, the priming site is a nick.

In some embodiments, the method further comprises, prior to (e), (i) separating polynucleotide molecules comprising one or more given sequences from the amplified polynucleotide molecules, to produce enriched polynucleotide molecules; and (ii) amplifying the enriched polynucleotide molecules with sequencing adaptors.

In some embodiments, the efficiency is at least 30%, 40%, or 50%. In some embodiments, the method further comprises identifying genetic variants upon sequencing the amplified polynucleotide molecules. In some embodiments, the sequencing comprises (i) subjecting the amplified polynucleotide molecules to an additional amplification reaction under conditions that yield additional amplified polynucleotide molecules as amplification products of the amplified polynucleotide molecules, and (ii) sequencing the additional amplified polynucleotide molecules. In some embodiments, the additional amplification is performed in the presence of sequencing adaptors.

In some embodiments, (b) and (c) are performed without aliquoting the tagged polynucleotide molecules. In some embodiments, the tagging is non-unique tagging.

Another aspect, provides a system for analyzing a target nucleic acid molecule of a subject, comprising a communication interface that receives nucleic acid sequence reads for a plurality of polynucleotide molecules that cover genomic loci of a target genome; computer memory that stores the nucleic acid sequence reads for the plurality of polynucleotide molecules received by the communication interface; and a computer processor operatively coupled to the communication interface and the memory and programmed to (i) group the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides, (ii) for each of the families, merge sequence reads to generate a consensus sequence, (iii) call the consensus sequence at a given genomic locus among the genomic loci, and (iv) detect at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls; and total number of alterations among the calls, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

In another aspect, a set of oligonucleotide molecules that selectively hybridize to at least 5 genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

In some embodiments, the oligonucleotide molecules are from 10-200 bases in length. In some embodiments, the oligonucleotide molecules selectively hybridize to exon regions of the at least 5 genes. In some embodiments, the oligonucleotide molecules selectively hybridize to at least 30 exons in the at least 5 genes. In some embodiments, multiple oligonucleotide molecules selectively hybridize to each of the at least 30 exons. In some embodiments, the oligonucleotide molecules that hybridize to each exon have sequences that overlap with at least 1 other oligonucleotide molecule.

In another aspect, a kit comprises a first container containing a plurality of library adaptors each having a different molecular barcode; and a second container containing a plurality of sequencing adaptors, each sequencing adaptor comprising at least a portion of a sequencer motif and optionally a sample barcode. The library adaptors can be as described above or elsewhere herein.

In some embodiments, the sequencing adaptor comprises the sample barcode. In some embodiments, the library adaptors are blunt ended and Y-shaped, and are less than or equal to 80 nucleic acid bases in length. In some embodiments, the sequencing adaptor is up to 70 bases from end to end.

In another aspect, a method for detecting sequence variants in a cell free DNA sample, comprising detecting rare DNA at a concentration less than 1% with a specificity that is greater than 99.9%.

In another aspect, a method comprises detecting genetic variants in a sample comprising DNA with a detection limit of at least 1% and specificity greater than 99.9%. In some

embodiments, the method further comprises converting cDNA (e.g. cfDNA) into adaptor tagged DNA with a conversion efficiency of at least 30%, 40%, or 50% and reducing sequencing noise (or distortion) by eliminating false positive sequence reads.

Another aspect provides a method, comprising (a) providing a sample comprising a set of double-stranded polynucleotide molecules, each double-stranded polynucleotide molecule including first and second complementary strands; (b) tagging the double-stranded polynucleotide molecules with a set of duplex tags, wherein each duplex tag differently tags the first and second complementary strands of a double-stranded polynucleotide molecule in the set; (c) sequencing at least some of the tagged strands to produce a set of sequence reads; (d) reducing and/or tracking redundancy in the set of sequence reads; (e) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; (f) determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci; and (g) estimating with a programmed computer processor a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

In some embodiments, the method further comprises (h) detecting copy number variation in the sample by determining a normalized total quantitative measure determined in step (g) at each of the one or more genetic loci and determining copy number variation based on the normalized measure. In some embodiments, the sample comprises double-stranded polynucleotide molecules sourced substantially from cell-free nucleic acids. In some embodiments, the duplex tags are not sequencing adaptors.

In some embodiments, reducing redundancy in the set of sequence reads comprises collapsing sequence reads produced from amplified products of an original polynucleotide molecule in the sample back to the original polynucleotide molecule. In some embodiments, the method further comprises determining a consensus sequence for the original polynucleotide molecule. In some embodiments, the method further comprises identifying polynucleotide molecules at one or more genetic loci comprising a sequence variant. In some embodiments, the method further comprises determining a quantitative measure of paired reads that map to a locus, wherein both strands of the pair comprise a sequence variant. In some embodiments, the method further comprises determining a quantitative measure of paired molecules in which only one member of the pair bears a sequence variant and/or determining a quantitative measure of unpaired molecules bearing a sequence variant. In some embodiments, the sequence variant is selected from the group consisting of a single nucleotide variant, an indel, a transversion, a translocation, an inversion, a deletion, a chromosomal structure alteration, a gene fusion, a chromosome fusion, a gene truncation, a gene amplification, a gene duplication and a chromosomal lesion.

Another aspect provides a system comprising a computer readable medium comprising machine-executable code that, upon execution by a computer processor, implements a method comprising (a) receiving into memory a set of sequence reads of polynucleotides tagged with duplex tags; (b) reducing and/or tracking redundancy in the set of sequence reads; (c) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; (d) determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci; and (e) estimating a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

Another aspect provides a method, comprising (a) providing a sample comprising a set of double-stranded polynucleotide molecules, each double-stranded polynucleotide molecule including first and second complementary strands; (b) tagging the double-stranded polynucleotide molecules with a set of duplex tags, wherein each duplex tag differently tags the first and second complementary strands of a double-stranded polynucleotide molecule in the set; (c) sequencing at least some of the tagged strands to produce a set of sequence reads; (d) reducing and/or tracking redundancy in the set of sequence reads; (e) sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads; and (f) determining quantitative measures of at least two of (i) the paired reads, (ii) the unpaired reads that map to each of one or more genetic loci, (iii) read depth of the paired reads and (iv) read depth of unpaired reads.

In some embodiments, (f) comprises determining quantitative measures of at least three of (i)-(iv). In some embodiments, (f) comprises determining quantitative measures of all of (i)-(iv). In some embodiments, the method further comprises (g) estimating with a programmed computer processor a quantitative measure of total double-stranded polynucleotide molecules in the set that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads and their read depths mapping to each locus.

In another aspect, a method comprises (a) tagging control parent polynucleotides with a first tag set to produce tagged control parent polynucleotides, wherein the first tag set comprises a plurality of tags, wherein each tag in the first tag set comprises a same control tag and an identifying tag, and wherein the tag set comprises a plurality of different identifying tags; (b) tagging test parent polynucleotides with a second tag set to produce tagged test parent polynucleotides, wherein the second tag set comprises a plurality of tags, wherein each tag in the second tag set comprises a same test tag that is distinguishable from the control tag and an identifying tag, and wherein the second tag set comprises a plurality of different identifying tags; (c) mixing tagged control parent polynucleotides with tagged test parent polynucleotides to form a pool; (d) amplifying tagged parent

polynucleotides in the pool to form a pool of amplified, tagged polynucleotides; (e) sequencing amplified, tagged polynucleotides in the amplified pool to produce a plurality of sequence reads; (f) grouping sequence reads into families, each family comprising sequence reads generated from a same parent polynucleotide, which grouping is optionally based on information from an identifying tag and from start/end sequences of the parent polynucleotides, and, optionally, determining a consensus sequence for each of a plurality of parent polynucleotides from the plurality of sequence reads in a group; (g) classifying each family or consensus sequence as a control parent polynucleotide or as a test parent polynucleotide based on having a test tag or a control tag; (h) determining a quantitative measure of control parent polynucleotides and control test polynucleotides mapping to each of at least two genetic loci; and (i) determining copy number variation in the test parent polynucleotides at at least one locus based on relative quantity of test parent polynucleotides and control parent polynucleotides mapping to the at least one locus.

In another aspect, a method comprises (a) generating a plurality of sequence reads from a plurality of template polynucleotides, each polynucleotide mapped to a genomic locus; (b) grouping the sequence reads into families, each family comprising sequence reads generated from one of the template polynucleotides; (c) calling a base (or sequence) at the genomic locus for each of the families; (d) detecting at the genomic locus any of genomic alterations among the calls, frequency of a genetic alteration among the calls, total number of calls and total number of alterations among the calls.

In some embodiments, calling comprises any of phylogenetic analysis, voting, weighing, assigning a probability to each read at the locus in a family, and calling the base with the highest probability. In some embodiments, the method is performed at two loci, comprising determining CNV at one of the loci based on counts at each of the loci.

Another aspect provides a method for determining a quantitative measure indicative of a number of individual double-stranded DNA fragments in a sample comprising (a) determining a quantitative measure of individual DNA molecules for which both strands are detected; (b) determining a quantitative measure of individual DNA molecules for which only one of the DNA strands are detected; (c) inferring from (a) and (b) above a quantitative measure of individual DNA molecules for which neither strand was detected; and (d) using (a)-(c) determining the quantitative measure indicative of a number of individual double-stranded DNA fragments in the sample.

In some embodiments, the method further comprises detecting copy number variation in the sample by determining a normalized quantitative measure determined in step (d) at each of one or more genetic loci and determining copy number variation based on the normalized measure. In some embodiments, the sample comprises double-stranded polynucleotide molecules sourced substantially from cell-free nucleic acids.

In some embodiments, determining the quantitative measure of individual DNA molecules comprises tagging the DNA molecules with a set of duplex tags, wherein each duplex tag differently tags complementary strands of a double-stranded DNA molecule in the sample to provide tagged strands. In some embodiments, the method further comprises sequencing at least some of the tagged strands to produce a set of sequence reads. In some embodiments, the method further comprises sorting sequence reads into paired reads and unpaired reads, wherein (i) each paired read corresponds to sequence reads generated from a first tagged strand and a second differently tagged complementary strand derived from a double-stranded polynucleotide molecule in the set, and (ii) each unpaired read represents a first tagged strand having no second differently tag complementary strand derived from a double-stranded polynucleotide molecule represented among the sequence reads in the set of sequence reads. In some embodiments, the method further comprises determining quantitative measures of (i) the paired reads and (ii) the unpaired reads that map to each of one or more genetic loci to determine a quantitative measure of total double-stranded DNA molecules in the sample that map to each of the one or more genetic loci based on the quantitative measure of paired reads and unpaired reads mapping to each locus.

In another aspect, a method for reducing distortion in a sequencing assay, comprises (a) tagging control parent polynucleotides with a first tag set to produce tagged control parent polynucleotides; (b) tagging test parent polynucleotides with a second tag set to produce tagged test parent polynucleotides; (c) mixing tagged control parent polynucleotides with tagged test parent polynucleotides to form a pool; (d) determining quantities of tagged control parent polynucleotides and tagged test parent polynucleotides; and (e) using the quantities of tagged control parent polynucleotides to reduce distortion in the quantities of tagged test parent polynucleotides.

In some embodiments, the first tag set comprises a plurality of tags, wherein each tag in the first tag set comprises a same control tag and an identifying tag, and wherein the first tag set comprises a plurality of different identifying tags. In some embodiments, the second tag set comprises a plurality of tags, wherein each tag in the second tag set comprises a same test tag and an identifying tag, wherein the test tag is distinguishable from the control tag, and wherein the second tag set comprises a plurality of different identifying tags. In some embodiments, (d) comprises amplifying tagged parent polynucleotides in the pool to form a pool of amplified, tagged polynucleotides, and sequencing amplified, tagged polynucleotides in the amplified pool to produce a plurality of sequence reads. In some embodiments, the method further comprises grouping sequence reads into families, each family comprising sequence reads generated from a same parent polynucleotide, which grouping is optionally based on information from an identifying tag and from start/end sequences of the parent polynucleotides, and, optionally, determining a consensus sequence for each of a plurality of parent polynucleotides from the plurality of sequence reads in a group.

In some embodiments, (d) comprises determining copy number variation in the test parent polynucleotides at greater than or equal to one locus based on relative quantity of test parent polynucleotides and control parent polynucleotides mapping to the locus.

Another aspect provides a method comprising (a) ligating adaptors to double-stranded DNA polynucleotides, wherein ligating is performed in a single reaction vessel, and wherein the adaptors comprise molecular barcodes, to produce a tagged library comprising an insert from the double-stranded DNA polynucleotides, and having between 4 and 1 million different tags; (b) generating a plurality of sequence reads for each of the double-stranded DNA polynucleotides in the tagged library; (c) grouping sequence reads into families, each family comprising sequence reads generated from a single DNA polynucleotide among the double-stranded DNA polynucleotides, based on information in a tag and information at an end of the insert; and (d) calling bases at

each position in the double-stranded DNA molecule based on bases at the position in members of a family. In some embodiments, (b) comprises amplifying each of the double-stranded DNA polynucleotide molecules in the tagged library to generate amplification products, and sequencing the amplification products. In some embodiments, the method further comprises sequencing the double-stranded DNA polynucleotide molecules a plurality of times. In some embodiments, (b) comprises sequencing the entire insert. In some embodiments, (c) further comprises collapsing sequence reads in each family to generate a consensus sequence. In some embodiments, (d) comprises calling a plurality of sequential bases from at least a subset of the sequence reads to identify single nucleotide variations (SNV) in the double-stranded DNA molecule.

Another aspect provides a method of detecting disease cell heterogeneity from a sample comprising polynucleotides from somatic cells and disease cells. The method comprises quantifying polynucleotides in the sample bearing a nucleotide sequence variant at each of a plurality of genetic loci; determining copy number variation (CNV) at each of the plurality of genetic loci, wherein the CNV indicates a genetic dose of a locus in the disease cell polynucleotides; determining with a programmed computer processor a relative measure of quantity of polynucleotides bearing a sequence variant at a locus per the genetic dose at the locus for each of a plurality of the loci; and comparing the relative measures at each of the plurality of loci, wherein different relative measures is indicative of tumor heterogeneity.

In another aspect, a method comprises subjecting a subject to one or more pulsed therapy cycles, each pulsed therapy cycle comprising (a) a first period during which a drug is administered at a first amount; and (b) a second period during which the drug is administered at a second, reduced amount, wherein (i) the first period is characterized by a tumor burden detected above a first clinical level; and (ii) the second period is characterized by a tumor burden detected below a second clinical level.

Additional aspects and advantages of the present disclosure will become readily apparent to those skilled in this art from the following detailed description, wherein only illustrative embodiments of the present disclosure are shown and described. As will be realized, the present disclosure is capable of other and different embodiments, and its several details are capable of modifications in various obvious respects, all without departing from the disclosure. Accordingly, the drawings and description are to be regarded as illustrative in nature, and not as restrictive.

INCORPORATION BY REFERENCE

All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings (also "figure" and "FIG." herein), of which:

FIG. **1** is a flowchart representation of a method of the present disclosure for determining copy number variation (CNV);

FIG. **2** depicts mapping of pairs and singlets to Locus A and Locus B in a genome;

FIG. **3** shows a reference sequence encoding a genetic Locus A;

FIGS. **4**A-C shows amplification, sequencing, redundancy reduction and pairing of complementary molecules;

FIG. **5** shows increased confidence in detecting sequence variants by pairing reads from Watson and Crick strands;

FIG. **6** shows a computer system that is programmed or otherwise configured to implement various methods of the present disclosure;

FIG. **7** is schematic representation of a system for analyzing a sample comprising nucleic acids from a user, including a sequencer; bioinformatic software and internet connection for report analysis by, for example, a hand held device or a desk top computer;

FIG. **8** is a flowchart representation of a method of this invention for determining CNV using pooled test and control pools; and

FIGS. **9**A-**9**C schematically illustrate a method for tagging a polynucleotide molecule with a library adaptor and subsequently a sequencing adaptor.

DETAILED DESCRIPTION

While various embodiments of the invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions may occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed.

The term "genetic variant," as used herein, generally refers to an alteration, variant or polymorphism in a nucleic acid sample or genome of a subject. Such alteration, variant or polymorphism can be with respect to a reference genome, which may be a reference genome of the subject or other individual. Single nucleotide polymorphisms (SNPs) are a form of polymorphisms. In some examples, one or more polymorphisms comprise one or more single nucleotide variations (SNVs), insertions, deletions, repeats, small insertions, small deletions, small repeats, structural variant junctions, variable length tandem repeats, and/or flanking sequences. Copy number variants (CNVs), transversions and other rearrangements are also forms of genetic variation. A genomic alternation may be a base change, insertion, deletion, repeat, copy number variation, or transversion.

The term "polynucleotide," as used herein, generally refers to a molecule comprising one or more nucleic acid subunits. A polynucleotide can include one or more subunits selected from adenosine (A), cytosine (C), guanine (G), thymine (T) and uracil (U), or variants thereof. A nucleotide can include A, C, G, T or U, or variants thereof. A nucleotide can include any subunit that can be incorporated into a growing nucleic acid strand. Such subunit can be an A, C, G, T, or U, or any other subunit that is specific to one or more complementary A, C, G, T or U, or complementary to a purine (i.e., A or G, or variant thereof) or a pyrimidine (i.e., C, T or U, or variant thereof). A subunit can enable individual nucleic acid bases or groups of bases (e.g., AA, TA, AT, GC, CG, CT, TC, GT, TG, AC, CA, or uracil-counterparts thereof) to be resolved. In some examples, a polynucleotide is deoxyribonucleic acid (DNA) or ribonucleic

acid (RNA), or derivatives thereof. A polynucleotide can be single-stranded or double stranded.

The term "subject," as used herein, generally refers to an animal, such as a mammalian species (e.g., human) or avian (e.g., bird) species, or other organism, such as a plant. More specifically, the subject can be a vertebrate, a mammal, a mouse, a primate, a simian or a human. Animals include, but are not limited to, farm animals, sport animals, and pets. A subject can be a healthy individual, an individual that has or is suspected of having a disease or a pre-disposition to the disease, or an individual that is in need of therapy or suspected of needing therapy. A subject can be a patient.

The term "genome" generally refers to an entirety of an organism's hereditary information. A genome can be encoded either in DNA or in RNA. A genome can comprise coding regions that code for proteins as well as non-coding regions. A genome can include the sequence of all chromosomes together in an organism. For example, the human genome has a total of 46 chromosomes. The sequence of all of these together constitutes a human genome.

The terms "adaptor(s)", "adapter(s)" and "tag(s)" are used synonymously throughout this specification. An adaptor or tag can be coupled to a polynucleotide sequence to be "tagged" by any approach including ligation, hybridization, or other approaches.

The term "library adaptor" or "library adapter" as used herein, generally refers to a molecule (e.g., polynucleotide) whose identity (e.g., sequence) can be used to differentiate polynucleotides in a biological sample (also "sample" herein).

The term "sequencing adaptor," as used herein, generally refers to a molecule (e.g., polynucleotide) that is adapted to permit a sequencing instrument to sequence a target polynucleotide, such as by interacting with the target polynucleotide to enable sequencing. The sequencing adaptor permits the target polynucleotide to be sequenced by the sequencing instrument. In an example, the sequencing adaptor comprises a nucleotide sequence that hybridizes or binds to a capture polynucleotide attached to a solid support of a sequencing system, such as a flow cell. In another example, the sequencing adaptor comprises a nucleotide sequence that hybridizes or binds to a polynucleotide to generate a hairpin loop, which permits the target polynucleotide to be sequenced by a sequencing system. The sequencing adaptor can include a sequencer motif, which can be a nucleotide sequence that is complementary to a flow cell sequence of other molecule (e.g., polynucleotide) and usable by the sequencing system to sequence the target polynucleotide. The sequencer motif can also include a primer sequence for use in sequencing, such as sequencing by synthesis. The sequencer motif can include the sequence(s) needed to couple a library adaptor to a sequencing system and sequence the target polynucleotide.

As used herein the terms "at least", "at most" or "about", when preceding a series, refers to each member of the series, unless otherwise identified.

The term "about" and its grammatical equivalents in relation to a reference numerical value can include a range of values up to plus or minus 10% from that value. For example, the amount "about 10" can include amounts from 9 to 11. In other embodiments, the term "about" in relation to a reference numerical value can include a range of values plus or minus 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1% from that value.

The term "at least" and its grammatical equivalents in relation to a reference numerical value can include the reference numerical value and greater than that value. For example, the amount "at least 10" can include the value 10 and any numerical value above 10, such as 11, 100, and 1,000.

The term "at most" and its grammatical equivalents in relation to a reference numerical value can include the reference numerical value and less than that value. For example, the amount "at most 10" can include the value 10 and any numerical value under 10, such as 9, 8, 5, 1, 0.5, and 0.1.

1. Methods for Processing and/or Analyzing a Nucleic Acid Sample

An aspect of the present disclosure provides methods for determining a genomic alternation in a nucleic acid sample of a subject. FIG. **1** shows a method of determining copy number variation (CNV). The method can be implemented to determine other genomic alternations, such as SNVs.

A. Polynucleotide Isolation

Methods disclosed herein can comprise isolating one or more polynucleotides. A polynucleotide can comprise any type of nucleic acid, for example, a sequence of genomic nucleic acid, or an artificial sequence (e.g., a sequence not found in genomic nucleic acid). For example, an artificial sequence can contain non-natural nucleotides. Also, a polynucleotide can comprise both genomic nucleic acid and an artificial sequence, in any portion. For example, a polynucleotide can comprise 1 to 99% of genomic nucleic acid and 99% to 1% of artificial sequence, where the total adds up to 100%. Thus, fractions of percentages are also contemplated. For example, a ratio of 99.1% to 0.9% is contemplated.

A polynucleotide can comprise any type of nucleic acids, such as DNA and/or RNA. For example, if a polynucleotide is DNA, it can be genomic DNA, complementary DNA (cDNA), or any other deoxyribonucleic acid. A polynucleotide can also be cell-free DNA (cfDNA). For example, the polynucleotide can be circulating DNA. The circulating DNA can comprise circulating tumor DNA (ctDNA). A polynucleotide can be double-stranded or single-stranded. Alternatively, a polynucleotide can comprise a combination of a double-stranded portion and a single-stranded portion.

Polynucleotides do not have to be cell-free. In some cases, the polynucleotides can be isolated from a sample. For example, in step (**102**) (FIG. **1**), double-stranded polynucleotides are isolated from a sample. A sample can be any biological sample isolated from a subject. For example, a sample can comprise, without limitation, bodily fluid, whole blood, platelets, serum, plasma, stool, red blood cells, white blood cells or leucocytes, endothelial cells, tissue biopsies, synovial fluid, lymphatic fluid, ascites fluid, interstitial or extracellular fluid, the fluid in spaces between cells, including gingival crevicular fluid, bone marrow, cerebrospinal fluid, saliva, mucous, sputum, semen, sweat, urine, or any other bodily fluids. A bodily fluid can include saliva, blood, or serum. For example, a polynucleotide can be cell-free DNA isolated from a bodily fluid, e.g., blood or serum. A sample can also be a tumor sample, which can be obtained from a subject by various approaches, including, but not limited to, venipuncture, excretion, ejaculation, massage, biopsy, needle aspirate, lavage, scraping, surgical incision, or intervention or other approaches.

A sample can comprise various amount of nucleic acid that contains genome equivalents. For example, a sample of about 30 ng DNA can contain about 10,000 (10^4) haploid human genome equivalents and, in the case of cfDNA, about 200 billion (2×10^{11}) individual polynucleotide molecules. Similarly, a sample of about 100 ng of DNA can contain about 30,000 haploid human genome equivalents and, in the case of cfDNA, about 600 billion individual molecules.

A sample can comprise nucleic acids from different sources. For example, a sample can comprise germline DNA or somatic DNA. A sample can comprise nucleic acids carrying mutations. For example, a sample can comprise DNA carrying germline mutations and/orsomatic mutations. A sample can also comprise DNA carrying cancer-associated mutations (e.g., cancer-associated somatic mutations).

B. Tagging

Polynucleotides disclosed herein can be tagged. For example, in step (**104**) (FIG. **1**) the double-stranded polynucleotides are tagged with duplex tags, tags that differently label the complementary strands (i.e., the "Watson" and "Crick" strands) of a double-stranded molecule. In one embodiment the duplex tags are polynucleotides having complementary and non-complementary portions.

Tags can be any types of molecules attached to a polynucleotide, including, but not limited to, nucleic acids, chemical compounds, florescent probes, or radioactive probes. Tags can also be oligonucleotides (e.g., DNA or RNA). Tags can comprise known sequences, unknown sequences, or both. A tag can comprise random sequences, pre-determined sequences, or both. A tag can be double-stranded or single-stranded. A double-stranded tag can be a duplex tag. A double-stranded tag can comprise two complementary strands. Alternatively, a double-stranded tag can comprise a hybridized portion and a non-hybridized portion. The double-stranded tag can be Y-shaped, e.g., the hybridized portion is at one end of the tag and the non-hybridized portion is at the opposite end of the tag. One such example are the "Y adapters" used in Illumina sequencing. Other examples include hairpin shaped adapters or bubble shaped adapters. Bubble shaped adapters have non-complementary sequences flanked on both sides by complementary sequences.

Tagging disclosed herein can be performed using any method. A polynucleotide can be tagged with an adaptor by hybridization. For example, the adaptor can have a nucleotide sequence that is complementary to at least a portion of a sequence of the polynucleotide. As an alternative, a polynucleotide can be tagged with an adaptor by ligation.

For example, tagging can comprise using one or more enzymes. The enzyme can be a ligase. The ligase can be a DNA ligase. For example, the DNA ligase can be a T4 DNA ligase, *E. coli* DNA ligase, and/or mammalian ligase. The mammalian ligase can be DNA ligase I, DNA ligase III, or DNA ligase IV. The ligase can also be a thermostable ligase. Tags can be ligated to a blunt-end of a polynucleotide (blunt-end ligation). Alternatively, tags can be ligated to a sticky end of a polynucleotide (sticky-end ligation). Efficiency of ligation can be increased by optimizing various conditions. Efficiency of ligation can be increased by optimizing the reaction time of ligation. For example, the reaction time of ligation can be less than 12 hours, e.g., less than 1, less than 2, less than 3, less than 4, less than 5, less than 6, less than 7, less than 8, less than 9, less than 10, less than 11, less than 12, less than 13, less than 14, less than 15, less than 16, less than 17, less than 18, less than 19, or less than 20 hours. In a particular example, reaction time of ligation is less than 20 hours. Efficiency of ligation can be increased by optimizing the ligase concentration in the reaction. For example, the ligase concentration can be at least 10, at least 50, at least 100, at least 150, at least 200, at least 250, at least 300, at least 400, at least 500, or at least 600 unit/microliter. Efficiency can also be optimized by adding or varying the concentration of an enzyme suitable for ligation, enzyme cofactors or other additives, and/or optimizing a temperature of a solution having the enzyme. Efficiency can also be optimized by varying the addition order of various components of the reaction. The end of tag sequence can comprise dinucleotide to increase ligation efficiency. When the tag comprises a non-complementary portion (e.g., Y-shaped adaptor), the sequence on the complementary portion of the tag adaptor can comprise one or more selected sequences that promote ligation efficiency. Preferably such sequences are located at the terminal end of the tag. Such sequences can comprise 1, 2, 3, 4, 5, or 6 terminal bases. Reaction solution with high viscosity (e.g., a low Reynolds number) can also be used to increase ligation efficiency. For example, solution can have a Reynolds number less than 3000, less than 2000, less than 1000, less than 900, less than 800, less than 700, less than 600, less than 500, less than 400, less than 300, less than 200, less than 100, less than 50, less than 25, or less than 10. It is also contemplated that roughly unified distribution of fragments (e.g., tight standard deviation) can be used to increase ligation efficiency. For example, the variation in fragment sizes can vary by less than 20%, less than 15%, less than 10%, less than 5%, or less than 1%. Tagging can also comprise primer extension, for example, by polymerase chain reaction (PCR). Tagging can also comprise any of ligation-based PCR, multiplex PCR, single strand ligation, or single strand circularization.

In some instances, the tags herein comprise molecular barcodes. Such molecular barcodes can be used to differentiate polynucleotides in a sample. Preferably molecular barcodes are different from one another. For example, molecular barcodes can have a difference between them that can be characterized by a predetermined edit distance or a Hamming distance. In some instances, the molecular barcodes herein have a minimum edit distance of 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10. To further improve efficiency of conversion (e.g., tagging) of untagged molecular to tagged molecules, one preferably utilizes short tags. For example, in some embodiments, a library adapter tag can be up to 65, 60, 55, 50, 45, 40, or 35 nucleotide bases in length. A collection of such short library barcodes preferably includes a number of different molecular barcodes, e.g., at least 2, 4, 6, 8, 10, 12, 14, 16, 18 or 20 different barcodes with a minimum edit distance of 1, 2, 3 or more.

Thus, a collection of molecules can include one or more tags. In some instances, some molecules in a collection can include an identifying tag ("identifier") such as a molecular barcode that is not shared by any other molecule in the collection. For example, in some instances of a collection of molecules, at least 50%, at least 51%, at least 52%, at least 53%, at least 54%, at least 55%, at least 56%, at least 57%, at least 58%, at least 59%, at least 60%, at least 61%, at least 62%, at least 63%, at least 64%, at least 65%, at least 66%, at least 67%, at least 68%, at least 69%, at least 70%, at least 71%, at least 72%, at least 73%, at least 74%, at least 75%, at least 76%, at least 77%, at least 78%, at least 79%, at least 80%, at least 81%, at least 82%, at least 83%, at least 84%, at least 85%, at least 86%, at least 87%, at least 88%, at least 89%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or 100% of the molecules in the collection an include an identifier or molecular barcode that is not shared by any other molecule in the collection. As used herein, a collection of molecules is considered to be "uniquely tagged" if each of at least 95% of the molecules in the collection bears an identifier that is not shared by any other molecule in the collection ("unique tag" or "unique identifier"). A collection of molecules is considered to be "non-uniquely tagged" if each of at least 1%, at least 5%, at

least 10%, at least 15%, at least 20%, at least 25%, at least 30%, at least 35%, at least 40%, at least 45%, or at least or about 50% of the molecules in the collection bears an identifying tag or molecular barcode that is shared by at least one other molecule in the collection ("non-unique tag" or "non-unique identifier"). Accordingly, in a non-uniquely tagged population no more than 1% of the molecules are uniquely tagged. For example, in a non-uniquely tagged population, no more than 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, or 50% of the molecules can be uniquely tagged.

A number of different tags can be used based on the estimated number of molecules in a sample. In some tagging methods, the number of different tags can be at least the same as the estimated number of molecules in the sample. In other tagging methods, the number of different tags can be at least two, three, four, five, six, seven, eight, nine, ten, one hundred or one thousand times as many as the estimated number of molecules in the sample. In unique tagging, at least two times (or more) as many different tags can be used as the estimated number of molecules in the sample.

The molecules in the sample may be non-uniquely tagged. In such instances a fewer number of tags or molecular barcodes is used then the number of molecules in the sample to be tagged. For example, no more than 100, 50, 40, 30, 20 or 10 unique tags or molecular barcodes are used to tag a complex sample such as a cell free DNA sample with many more different fragments.

The polynucleotide to be tagged can be fragmented, such as either naturally or using other approaches, such as, for example, shearing. The polynucleotides can be fragmented by certain methods, including but not limited to, mechanical shearing, passing the sample through a syringe, sonication, heat treatment (e.g., for 30 minutes at 90° C.), and/or nuclease treatment (e.g., using DNase, RNase, endonuclease, exonuclease, and/or restriction enzyme).

The polynucleotides fragments (prior to tagging) can comprise sequences of any length. For example, polynucleotide fragments (prior to tagging) can comprise at least 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, 275, 280, 285, 290, 295, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000 or more nucleotides in length. The polynucleotide fragment are preferably about the average length of cell-free DNA. For example, the polynucleotide fragments can comprise about 160 bases in length. The polynucleotide fragment can also be fragmented from a larger fragment into smaller fragments about 160 bases in length.

Polynucleotides tagged can comprise sequences associated with cancer. The cancer-associated sequences can comprise single nucleotide variation (SNV), copy number variation (CNV), insertions, deletions, and/or rearrangements.

The polynucleotides can comprise sequences associated with cancer, such as acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), adrenocortical carcinoma, Kaposi Sarcoma, anal cancer, basal cell carcinoma, bile duct cancer, bladder cancer, bone cancer, osteosarcoma, malignant fibrous histiocytoma, brain stem glioma, brain cancer, craniopharyngioma, ependymoblastoma, ependymoma, medulloblastoma, medulloepithelioma, pineal parenchymal tumor, breast cancer, bronchial tumor, Burkitt lymphoma, Non-Hodgkin lymphoma, carcinoid tumor, cervical cancer, chordoma, chronic lymphocytic leukemia (CLL), chronic myelogenous leukemia (CML), colon cancer, colorectal cancer, cutaneous T-cell lymphoma, ductal carcinoma in situ, endometrial cancer, esophageal cancer, Ewing Sarcoma, eye cancer, intraocular melanoma, retinoblastoma, fibrous histiocytoma, gallbladder cancer, gastric cancer, glioma, hairy cell leukemia, head and neck cancer, heart cancer, hepatocellular (liver) cancer, Hodgkin lymphoma, hypopharyngeal cancer, kidney cancer, laryngeal cancer, lip cancer, oral cavity cancer, lung cancer, non-small cell carcinoma, small cell carcinoma, melanoma, mouth cancer, myelodysplastic syndromes, multiple myeloma, medulloblastoma, nasal cavity cancer, paranasal sinus cancer, neuroblastoma, nasopharyngeal cancer, oral cancer, oropharyngeal cancer, osteosarcoma, ovarian cancer, pancreatic cancer, papillomatosis, paraganglioma, parathyroid cancer, penile cancer, pharyngeal cancer, pituitary tumor, plasma cell neoplasm, prostate cancer, rectal cancer, renal cell cancer, rhabdomyosarcoma, salivary gland cancer, Sezary syndrome, skin cancer, nonmelanoma, small intestine cancer, soft tissue sarcoma, squamous cell carcinoma, testicular cancer, throat cancer, thymoma, thyroid cancer, urethral cancer, uterine cancer, uterine sarcoma, vaginal cancer, vulvar cancer, Waldenstrom macroglobulinemia, and/or Wilms Tumor.

In certain embodiments, a population of polynucleotides in a sample of fragmented genomic DNA is tagged with n different unique identifiers, wherein n is at least 2 and no more than 100,000*z, wherein z is a measure of central tendency (e.g., mean, median, mode) of an expected number of duplicate molecules having the same start and stop positions. In certain embodiments, n is at least any of 2*z, 3*z, 4*z, 5*z, 6*z, 7*z, 8*z, 9*z, 10*z, 11*z, 12*z, 13*z, 14*z, 15*z, 16*z, 17*z, 18*z, 19*z, or 20*z (e.g., lower limit). In other embodiments, n is no greater than 100,000*z, 10,000*z, 1000*z or 100*z (e.g., upper limit). Thus, n can range between any combination of these lower and upper limits. In certain embodiments, n is between 5*z and 15*z, between 8*z and 12*z, or about 10*z. A haploid human genome equivalent has about 3 picograms of DNA. A sample of about 1 microgram of DNA contains about 300,000 haploid human genome equivalents. Improvements in sequencing can be achieved as long as at least some of the duplicate or cognate polynucleotides bear unique identifiers with respect to each other, that is bear different tags. However, in certain embodiments, the number of tags used is selected so that there is at least a 95% chance that all duplicate molecules starting at any one position bear unique identifiers. For example, in a sample comprising about 10,000 haploid human genome equivalents of fragmented genomic DNA, e.g., cfDNA, z is expected to be between 2 and 8. Such a population can be tagged with between about 10 and 100 different identifiers, for example, about 2 identifiers, about 4 identifiers, about 9 identifiers, about 16 identifiers, about 25 identifiers, about 36 different identifiers, about 49 different identifiers, about 64 different identifiers, about 81 different identifiers, about 100 different identifiers.

Nucleic acid barcodes having identifiable sequences including molecular barcodes, can be used for tagging. For example, a plurality of DNA barcodes can comprise various numbers of sequences of nucleotides. A plurality of DNA barcodes having 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more identifiable sequences of nucleotides can be used. When attached to only one end of a polynucleotide, the plurality of DNA barcodes can produce 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 or more different identifiers. Alternatively, when attached to both ends of a polynucleotide, the plurality

DNA barcodes can produce 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 324, 361, 400 or more different identifiers (which is the ^2 of when the DNA barcode is attached to only 1 end of a polynucleotide). In one example, a plurality of DNA barcodes having 6, 7, 8, 9 or 10 identifiable sequences of nucleotides can be used. When attached to both ends of a polynucleotide, they produce 36, 49, 64, 81 or 100 possible different identifiers, respectively. In a particular example, the plurality of DNA barcodes can comprise 8 identifiable sequences of nucleotides. When attached to only one end of a polynucleotide, the plurality of DNA barcodes can produce 8 different identifiers. Alternatively, when attached to both ends of a polynucleotide, the plurality of DNA barcodes can produce 64 different identifiers. Samples tagged in such a way can be those with a range of about 10 ng to any of about 100 ng, about 1 μg, about 10 μg of fragmented polynucleotides, e.g., genomic DNA, e.g., cfDNA.

A polynucleotide can be uniquely identified in various ways. A polynucleotide can be uniquely identified by a unique DNA barcode. For example, any two polynucleotides in a sample are attached two different DNA barcodes. Alternatively, a polynucleotide can be uniquely identified by the combination of a DNA barcode and one or more endogenous sequences of the polynucleotide. For example, any two polynucleotides in a sample can be attached the same DNA barcode, but the two polynucleotides can still be identified by different endogenous sequences. The endogenous sequence can be on an end of a polynucleotide. For example, the endogenous sequence can be adjacent (e.g., base in between) to the attached DNA barcode. In some instances the endogenous sequence can be at least 2, 4, 6, 8, 10, 20, 30, 40, 50, 60, 70, 80, 90, or 100 bases in length. Preferably, the endogenous sequence is a terminal sequence of the fragment/polynucleotides to be analyzed. The endogenous sequence may be the length of the sequence. For example, a plurality of DNA barcodes comprising 8 different DNA barcodes can be attached to both ends of each polynucleotide in a sample. Each polynucleotide in the sample can be identified by the combination of the DNA barcodes and about 10 base pair endogenous sequence on an end of the polynucleotide. Without being bound by theory, the endogenous sequence of a polynucleotide can also be the entire polynucleotide sequence.

Also disclosed herein are compositions of tagged polynucleotides. The tagged polynucleotide can be single-stranded. Alternatively, the tagged polynucleotide can be double-stranded (e.g., duplex-tagged polynucleotides). Accordingly, this invention also provides compositions of duplex-tagged polynucleotides. The polynucleotides can comprise any types of nucleic acids (DNA and/or RNA). The polynucleotides comprise any types of DNA disclosed herein. For example, the polynucleotides can comprise DNA, e.g., fragmented DNA or cfDNA. A set of polynucleotides in the composition that map to a mappable base position in a genome can be non-uniquely tagged, that is, the number of different identifiers can be at least 2 and fewer than the number of polynucleotides that map to the mappable base position. The number of different identifiers can also be at least 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 and fewer than the number of polynucleotides that map to the mappable base position.

In some instances, as a composition goes from about 1 ng to about 10 μg or higher, a larger set of different molecular barcodes can be used. For example, between 5 and 100 different library adaptors can be used to tag polynucleotides in a cfDNA sample.

The systems and methods disclosed herein may be used in applications that involve the assignment of molecular barcodes. The molecular barcodes can be assigned to any types of polynucleotides disclosed in this invention. For example, the molecular barcodes can be assigned to cell-free polynucleotides (e.g., cfDNAs). Often, an identifier disclosed herein can be a barcode oligonucleotide that is used to tag the polynucleotide. The barcode identifier may be a nucleic acid oligonucleotide (e.g., a DNA oligonucleotide). The barcode identifier can be single-stranded. Alternatively, the barcode identifier can be double-stranded. The barcode identifier can be attached to polynucleotides using any method disclosed herein. For example, the barcode identifier can be attached to the polynucleotide by ligation using an enzyme. The barcode identifier can also be incorporated into the polynucleotide through PCR. In other cases, the reaction may comprise addition of a metal isotope, either directly to the analyte or by a probe labeled with the isotope. Generally, assignment of unique or non-unique identifiers or molecular barcodes in reactions of this disclosure may follow methods and systems described by, for example, U.S. patent applications 2001/0053519, 2003/0152490, 2011/0160078 and U.S. Pat. No. 6,582,908, each of which is entirely incorporated herein by reference.

Identifiers or molecular barcodes used herein may be completely endogenous whereby circular ligation of individual fragments may be performed followed by random shearing or targeted amplification. In this case, the combination of a new start and stop point of the molecule and the original intramolecular ligation point can form a specific identifier.

Identifiers or molecular barcodes used herein can comprise any types of oligonucleotides. In some cases, identifiers may be predetermined, random, or semi-random sequence oligonucleotides. Identifiers can be barcodes. For example, a plurality of barcodes may be used such that barcodes are not necessarily unique to one another in the plurality. Alternatively, a plurality of barcodes may be used such that each barcode is unique to any other barcode in the plurality. The barcodes can comprise specific sequences (e.g., predetermined sequences) that can be individually tracked. Further, barcodes may be attached (e.g., by ligation) to individual molecules such that the combination of the barcode and the sequence it may be ligated to creates a specific sequence that may be individually tracked. As described herein, detection of barcodes in combination with sequence data of beginning (start) and/or end (stop) portions of sequence reads can allow assignment of a unique identity to a particular molecule. The length or number of base pairs of an individual sequence read may also be used to assign a unique identity to such a molecule. As described herein, fragments from a single strand of nucleic acid having been assigned a unique identity, may thereby permit subsequent identification of fragments from the parent strand. In this way the polynucleotides in the sample can be uniquely or substantially uniquely tagged. A duplex tag can include a degenerate or semi-degenerate nucleotide sequence, e.g., a random degenerate sequence. The nucleotide sequence can comprise any number of nucleotides. For example, the nucleotide sequence can comprise 1 (if using a non-natural nucleotide), 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more nucleotides. In a particular example, the sequence can comprise 7 nucleotides. In another example,

the sequence can comprise 8 nucleotides. The sequence can also comprise 9 nucleotides. The sequence can comprise 10 nucleotides.

A barcode can comprise contiguous or non-contiguous sequences. A barcode that comprises at least 1, 2, 3, 4, 5 or more nucleotides is a contiguous sequence or non-contiguous sequence. if the 4 nucleotides are uninterrupted by any other nucleotide. For example, if a barcode comprises the sequence TTGC, a barcode is contiguous if the barcode is TTGC. On the other hand, a barcode is non-contiguous if the barcode is TTXGC, where X is a nucleic acid base.

An identifier or molecular barcode can have an n-mer sequence which may be 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more nucleotides in length. A tag herein can comprise any range of nucleotides in length. For example, the sequence can be between 2 to 100, 10 to 90, 20 to 80, 30 to 70, 40 to 60, or about 50 nucleotides in length.

The tag can comprise a double-stranded fixed reference sequence downstream of the identifier or molecular barcode. Alternatively, the tag can comprise a double-stranded fixed reference sequence upstream or downstream of the identifier or molecular barcode. Each strand of a double-stranded fixed reference sequence can be, for example, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 nucleotides in length.

C. Adaptors

A library of polynucleotide molecules can be synthesized for use in sequencing. For example, a library of polynucleotides comprising a plurality of polynucleotide molecules that are each less than or equal to 100, 90, 80, 70, 60, 50, 45, 40, or 35 nucleic acid (or nucleotide) bases in length can be made. A plurality of polynucleotide molecules can be each less than or equal to 35 nucleic acid bases in length. A plurality of polynucleotide molecules can be each less than or equal to 30 nucleic acid bases in length. A plurality of polynucleotide molecules can also be less than or equal to 250, 200, 150, 100, or 50 nucleic acid bases. Additionally, the plurality of polynucleotide molecules can also be less than or equal to 100, 99, 98, 97, 96, 95, 94, 93, 92, 91, 90, 89, 88, 87, 86, 85, 84, 83, 82, 81, 80, 79, 78, 77, 76, 75, 74, 73, 72, 71, 70, 69, 68, 67, 66, 65, 64, 63, 62, 61, 60, 59, 58, 57, 56, 55, 54, 53, 52, 51, 50, 49, 48, 47, 46, 45, 44, 43, 42, 41, 40, 39, 38, 37, 36, 35, 34, 33, 32, 31, 30, 29, 28, 27, 26, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, or 10 nucleic acid bases.

A library of polynucleotides comprising a plurality of polynucleotide molecules can also have distinct (with respect to each other) molecular barcode sequences (or molecular barcodes) with respect to at least 4 nucleic acid bases. A molecular barcode (also "barcode" or "identifier" herein) sequence is a nucleotide sequence that distinguishes one polynucleotide from another. In other embodiments, the polynucleotide molecules can also have different barcode sequences with respect to 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more nucleic acid bases.

A library of polynucleotides comprising a plurality of polynucleotide molecules can also have a plurality of different barcode sequences. For example, a plurality of polynucleotide molecules can have at least 4 different molecular barcode sequences. In some cases, the plurality of polynucleotide molecules has from 2-100, 4-50, 4-30, 4-20, or 4-10 different molecular barcode sequences. The plurality of polynucleotides molecules can also have other ranges of different barcode sequences such as, 1-4, 2-5, 3-6, 4-7, 5-8, 6-9, 7-10, 8-11, 9-12, 10-13, 11-14, 12-15, 13-16, 14-17, 15-18, 16-19, 17-20, 18-21, 19-22, 20-23, 21-24, or 22-25 different barcode sequences. In other cases, a plurality of polynucleotide molecules can have at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or 100 more different barcode sequences. In a particular example, the plurality library adapters comprise at least 8 different sequences.

The location of the different barcode sequences can vary within the plurality of polynucleotides. For example, the different barcode sequences can be within 20, 15, 10, 9, 8, 7, 6, 5, 4, 3, or 2 nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules. In an example, a plurality of polynucleotide molecules has distinct barcode sequences that are within 10 nucleic acid bases from the terminal end. In another example, a plurality of polynucleotide molecules has distinct barcode sequences that are within 5 or 1 nucleic acid bases from the terminal end. In other instances, the distinct barcode sequences can be at the terminal end of a respective one of the plurality of polynucleotide molecules. Other variations include that the distinct molecular barcode sequences can be within 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, or 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, 200, or more nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules.

The terminal end of the plurality of polynucleotide molecules can be adapted for ligation to a target nucleic acid molecule. For example, the terminal end can be a blunt end. In some other cases, the terminal end is adapted for hybridization to a complementary sequence of a target nucleic acid molecule.

A library of polynucleotides comprising a plurality of polynucleotide molecules can also have an edit distance of at least 1. In some cases, the edit distance is with respect to individual bases of the plurality of polynucleotide molecules. In other cases, the plurality of polynucleotide molecules can have an edit distance of at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50 or more. The edit distance can be a Hamming distance.

In some cases, the plurality of polynucleotides does not contain sequencing adaptors. A sequence adaptor can be a polynucleotide that comprises a sequence that hybridizes to one or more sequencing adaptors or primers. A sequencing adaptor can further comprise a sequence hybridizing to a

solid support, e.g., a flow cell sequence. The term "flow cell sequence" and its grammatical equivalents as used herein, refers to a sequence that permits hybridization to a substrate, for example, by way of a primer attached to the substrate. The substrate can be bead or a planar surface. In some embodiments, a flow cell sequence can allow a polynucleotide to attach to a flow cell or surface (e.g., surface of a bead, for example, an Illumina flow cell.

When a plurality of polynucleotide molecules does not contain sequencing adaptors or primers, each polynucleotide molecule of the plurality does not contain a nucleic acid sequence or other moiety that is adapted to permit sequencing of a target nucleic acid molecule with a given sequencing approach, such as Illumina, SOLiD, Pacific Biosciences, GeneReader, Oxford Nanopore, Complete Genomics, GnuBio, Ion Torrent, Oxford Nanopore or Genia. In some examples, when a plurality of polynucleotide molecules does not contain sequencing adaptors or primers, the plurality of polynucleotide molecules does not contain flow cell sequences. For example, the plurality of polynucleotide molecules cannot bind to flow cells, such as used in Illumina flow cell sequencers. However, these flow cell sequences, if desired, can be added to the plurality of polynucleotide molecules by methods such as PCR amplification or ligation. At this point, Illumina flow cell sequencers can be used. Alternatively, when the plurality of polynucleotide molecules does not contain sequencing adaptors or primers, the plurality of polynucleotide molecules does not contain hairpin shaped adaptors or adaptors for generating hairpin loops in a target nucleic acid molecule, such as Pacific Bioscience SMRTbell™ adaptors. However, these hairpin shaped adaptors, if desired, can be added to the plurality of polynucleotide molecules by methods such as PCR amplification or ligation. The plurality of polynucleotide molecules can be circular or linear.

A plurality of polynucleotide molecules can be double stranded. In some cases, the plurality of polynucleotide molecules can be single stranded, or can comprise hybridized and non-hybridized regions. A plurality of polynucleotide molecules can be non-naturally occurring polynucleotide molecules.

Adaptors can be polynucleotide molecules. The polynucleotide molecules can be Y-shaped, bubble-shaped or hairpin-shaped. A hairpin adaptor may contain a restriction site(s) or a Uracil containing base. Adaptors can comprise a complementary portion and a non-complementary portion. The non-complementary portion can have an edit distance (e.g., Hamming distance). For example, the edit distance can be at least 1, at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, or at least 30. The complementary portion of the adaptor can comprise sequences that are selected to enable and/or promote ligation to a polynucleotide, e.g., a sequence to enable and/or promote ligation to a polynucleotide at a high yield.

A plurality of polynucleotide molecules as disclosed herein can be purified. In some cases, a plurality of polynucleotide molecules as disclosed herein can be isolated polynucleotide molecules. In other cases, a plurality of polynucleotide molecules as disclosed herein can be purified and isolated polynucleotide molecules.

In certain aspects, each of the plurality of polynucleotide molecules is Y-shaped or hairpin-shaped. Each of the plurality of polynucleotide molecules can comprise a different barcode. The different barcode can be a randomer in the complementary portion (e.g., double stranded portion) of the Y-shaped or hairpin-shaped adaptor. Alternatively, the different barcode can be in one strand of the non-complementary portion (e.g., one of the Y-shaped arms). As discussed above, the different barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, e.g., 7 bases. The barcode can be contiguous or non-contiguous sequences, as described above. The plurality of polynucleotide molecules is from 10 nucleic acid bases to 35 nucleic acid bases (or any length as described above) in length. Further, the plurality of polynucleotide molecules can comprise an edit distance (as described above), that is a Hamming distance. A plurality of polynucleotide molecules can have distinct barcode sequences that are within 10 nucleic acid bases from the terminal end.

In another aspect, a plurality of polynucleotide molecules can be sequencing adaptors. A sequencing adaptor can comprise a sequence hybridizing to one or more sequencing primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, e.g., a flow cell sequence. For example, a sequencing adaptor can be a flow cell adaptor. The sequencing adaptors can be attached to one or both ends of a polynucleotide fragment. In another example, a sequencing adaptor can be hairpin shaped. For example, the hairpin shaped adaptor can comprise a complementary double-stranded portion and a loop portion, where the double-stranded portion can be attached (e.g., ligated) to a double-stranded polynucleotide. Hairpin shaped sequencing adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. A sequencing adaptor can be up to 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, or more bases from end to end. For example, a sequencing adaptor can be up to 70 bases from end to end. The sequencing adaptor can comprise 20-30, 20-40, 30-50, 30-60, 40-60, 40-70, 50-60, 50-70, bases from end to end. In a particular example, the sequencing adaptor can comprise 20-30 bases from end to end. In another example, the sequencing adaptor can comprise 50-60 bases from end to end. A sequencing adaptor can comprise one or more barcodes. For example, a sequencing adaptor can comprise a sample barcode. The sample barcode can comprise a pre-determined sequence. The sample barcodes can be used to identify the source of the polynucleotides. The sample barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, e.g., at least 8 bases. The barcode can be contiguous or non-contiguous sequences, as described above.

The plurality of polynucleotide molecules as described herein can be used as adaptors. Adaptors can comprise one or more identifiers. An adaptor can comprise an identifier with a random sequence. Alternatively, an adaptor can comprise an identifier with pre-determined sequences. Some adaptors can comprise an identifier with a random sequence and another identifier with a pre-determined sequence. The adaptors comprising identifiers can be double-stranded or single-stranded adaptors. The adaptors comprising identifiers can be Y-shaped adaptors. A Y-shaped adaptor can

comprise one or more identifiers with a random sequence. The one or more identifiers can be on the hybrid portion and/or non-hybridized portion of the Y-shaped adaptor. A Y-shaped adaptor can comprise one or more identifiers with a pre-determined sequence. The one or more identifiers with pre-determined sequence can be on the hybridized portion and/or non-hybridized portion of the Y-shaped adaptor. A Y-shaped adaptor can comprise one or more identifiers with a random sequence and one or more identifiers with a pre-determined sequence. For example, the one or more identifiers with a random sequence can be on the hybridized portion of the Y-shaped adaptor and/or the non-hybridized portion of the Y-shaped adaptor. The one or more identifiers with a pre-determined sequence can be on the hybridized portion of the Y-shaped adaptor and/or the non-hybridized portion of the Y-shaped adaptor. In a particular example, a Y-shaped adaptor can comprise an identifier with a random sequence on its hybridized portion and an identifier with a pre-determined sequence on its non-hybridized portion. The identifiers can be in any length disclosed herein. For example, a Y-shaped adaptor can comprise an identifier with a random sequence of 7 nucleotides on its hybridized portion and an identifier with a pre-determined sequence of 8 nucleotides on its non-hybridized portion.

An adaptor can include a double-stranded portion with a molecular barcode and at least one or two single-stranded portion. For example, the adaptor can be Y-shaped and include a double-stranded portion and two single-stranded portions. The single-stranded portions can include sequences that are not complementary to one another.

The adaptor can include a terminal end that has a sequence that is selected to permit the adaptor to be efficiently (e.g., at an efficiency of at least about 20%, 30%, 40%, 50%) ligated or otherwise coupled to a polynucleotide. In some examples, terminal nucleotides in a double-stranded portion of an adaptor are selected from a combination of purines and pyrimidines to provide for efficient ligation.

In some examples, a set of library adaptors comprises a plurality of polynucleotide molecules (library adaptors) with molecular barcodes. The library adaptors are less than or equal to 80, 70, 60, 50, 45, or 40 nucleotide bases in length. The molecular barcodes can be at least 4 nucleotide bases in length, but may be from 4 to 20 nucleotide bases in length. The molecular barcodes can be different from one another and have an edit distance of at least 1, 2, 3, 4, or 5 between one another. The molecular barcodes are located at least 1, 2, 3, 4, 5, 10, or 20 nucleotide bases away from a terminal end of their respective library adaptors. In some cases, the at least one terminal base is identical in all of the library adaptors.

The library adaptors can be identical but for the molecular barcodes. For example, the library adaptors can have identical sequences but differ only with respect to nucleotide sequences of the molecular barcodes.

Each of the library adaptors can have a double stranded portion and at least one single-stranded portion. By "single stranded portion" is meant an area of non-complementarity or an overhang. In some cases, each of the library adaptors has a double-stranded portion and two single-stranded portions. The double-stranded portion can have a molecular barcode. In some cases, the molecular barcode is a randomer. Each of the library adaptors can further include a strand-identification barcode on a single-stranded portion. The strand-identification barcode can include at least 4 nucleotide bases, in some cases from 4 to 20 nucleotide bases.

In some examples, each of the library adaptors has a double-stranded portion with a molecular barcode and two single-stranded portions. The single-stranded portions may not hybridize to one another. The single-stranded portions may not be completely complementary to one another.

The library adaptors can have a sequence of terminal nucleotides in a double-stranded portion that are the same. The sequence of terminal nucleotides can be at least 2, 3, 4, 5 or 6 nucleotide bases in length. For example, one strand of a double-stranded portion of the library adaptor can have the sequence ACTT, TCGC, or TACC at the terminal end, while the other strand can have a complementary sequence. In some cases, such a sequence is selected to optimize the efficiency at which the library adaptors ligate to target polynucleotides. Such sequences can be selected to optimize a binding interaction between the ends of the library adaptors and the target polynucleotides.

In some cases, none of the library adaptors contains a sample identification motif (or sample molecular barcode). Such sample identification motif can be provided via sequencing adaptors. A sample identification motif can include a sequencer of at least 4, 5, 6, 7, 8, 9, 10, 20, 30, or 40 nucleotide bases that permits the identification of polynucleotide molecules from a given sample from polynucleotide molecules from other samples. For example, this can permit polynucleotide molecules from two subjects to be sequenced in the same pool and sequence reads for the subjects subsequently identified.

A sequencer motif includes nucleotide sequence(s) needed to couple a library adaptor to a sequencing system and sequence a target polynucleotide coupled to the library adaptor. The sequencer motif can include a sequence that is complementary to a flow cell sequence and a sequence (sequencing initiation sequence) that is selectively hybridizable to a primer (or priming sequence) for use in sequencing. For example, such sequencing initiation sequence can be complementary to a primer that is employed for use in sequence by synthesis (e.g., Illumina). Such primer can be included in a sequencing adaptor. A sequencing initiation sequence can be a primer hybridization site.

In some cases, none of the library adaptors contains a complete sequencer motif. The library adaptors can contain partial or no sequencer motifs. In some cases, the library adaptors include a sequencing initiation sequence. The library adaptors can include a sequencing initiation sequence but no flow cell sequence. The sequence initiation sequence can be complementary to a primer for sequencing. The primer can be a sequence specific primer or a universal primer. Such sequencing initiation sequences may be situated on single-stranded portions of the library adaptors. As an alternative, such sequencing initiation sequences may be priming sites (e.g., kinks or nicks) to permit a polymerase to couple to the library adaptors during sequencing.

In some cases, partial or complete sequencer motifs are provided by sequencing adaptors. A sequencing adaptor can include a sample molecular barcode and a sequencer motif. The sequencing adaptors can be provided in a set that is separate from the library adaptors. The sequencing adaptors in a given set can be identical—i.e., they contain the same sample barcode and sequencer motif.

Sequencing adaptors can include sample identification motifs and sequencer motifs. Sequencer motifs can include primers that are complementary to a sequencing initiation sequence. In some cases, sequencer motifs also include flow cell sequences or other sequences that permit a polynucleotide to a configured or arranged in a manner that permits the polynucleotide to be sequenced by a sequencer.

Library adaptors and sequencing adaptors can each be partial adaptors, that is, containing part but not all of the sequences necessary to enable sequencing by a sequencing platform. Together they provide complete adaptors. For example, library adaptors can include partial or no sequencer motifs, but such sequencer motifs are provided by sequencing adaptors.

FIGS. **9**A-**9**C schematically illustrate a method for tagging a target polynucleotide molecule with library adaptors. FIG. **9**A shows a library adaptor as a partial adaptor containing a primer hybridization site on one of the strands and a molecular barcode towards another end. The primer hybridization site can be a sequencing initiation sequence for subsequent sequencing. The library adaptor is less than or equal to 80 nucleotide bases in length. In FIG. **9**B, the library adaptors are ligated at both ends of the target polynucleotide molecule to provide a tagged target polynucleotide molecule. The tagged target polynucleotide molecule may be subjected to nucleic acid amplification to generate copies of the target. Next, in FIG. **9**C, sequencing adaptors containing sequencer motifs are provided and hybridized to the tagged target polynucleotide molecule. The sequencing adaptors contain sample identification motifs. The sequencing adaptors can contain sequences to permit sequencing of the tagged target with a given sequencer.

D. Sequencing

Tagged polynucleotides can be sequenced to generate sequence reads (e.g., as shown in step (**106**), FIG. **1**). For example, a tagged duplex polynucleotide can be sequenced. Sequence reads can be generated from only one strand of a tagged duplex polynucleotide. Alternatively, both strands of a tagged duplex polynucleotide can generate sequence reads. The two strands of the tagged duplex polynucleotide can comprise the same tags. Alternatively, the two strands of the tagged duplex polynucleotide can comprise different tags. When the two strands of the tagged duplex polynucleotide are differently tagged, sequence reads generated from one strand (e.g., a Watson strand) can be distinguished from sequence reads generated from the other strands (e.g., a Crick strand). Sequencing can involve generating multiple sequence reads for each molecule. This occurs, for example, as a result the amplification of individual polynucleotide strands during the sequencing process, e.g., by PCR.

Methods disclosed herein can comprise amplifying of polynucleotides. Polynucleotides amplification can result in the incorporation of nucleotides into a nucleic acid molecule or primer thereby forming a new nucleic acid molecule complementary to a template nucleic acid. The newly formed polynucleotide molecule and its template can be used as templates to synthesize additional polynucleotides. The polynucleotides being amplified can be any nucleic acids, for example, deoxyribonucleic acids, including genomic DNAs, cDNAs (complementary DNA), cfDNAs, and circulating tumor DNAs (ctDNAs). The polynucleotides being amplified can also be RNAs. As used herein, one amplification reaction may comprise many rounds of DNA replication. DNA amplification reactions can include, for example, polymerase chain reaction (PCR). One PCR reaction may comprise 2-100 "cycles" of denaturation, annealing, and synthesis of a DNA molecule. For example, 2-7, 5-10, 6-11, 7-12, 8-13, 9-14, 10-15, 11-16, 12-17, 13-18, 14-19, or 15-20 cycles can be performed during the amplification step. The condition of the PCR can be optimized based on the GC content of the sequences, including the primers.

Nucleic acid amplification techniques can be used with the assays described herein. Some amplification techniques are the PCR methodologies which can include, but are not limited to, solution PCR and in situ PCR. For example, amplification may comprise PCR-based amplification. Alternatively, amplification may comprise non PCR-based amplification. Amplification of the template nucleic acid may comprise use of one or more polymerases. For example, the polymerase may be a DNA polymerase or an RNA polymerase. In some cases, high fidelity amplification is performed such as with the use of high fidelity polymerase (e.g., Phusion® High-Fidelity DNA Polymerase) or PCR protocols. In some cases, the polymerase may be a high fidelity polymerase. For example, the polymerase may be KAPA HiFi DNA polymerase. The polymerase may also be Phusion DNA polymerase. The polymerase may be used under reaction conditions that reduce or minimize amplification biases, e.g., due to fragment length, GC content, etc.

Amplification of a single strand of a polynucleotide by PCR will generate copies both of that strand and its complement. During sequencing, both the strand and its complement will generate sequence reads. However, sequence reads generated from the complement of, for example, the Watson strand, can be identified as such because they bear the complement of the portion of the duplex tag that tagged the original Watson strand. In contrast, a sequence read generated from a Crick strand or its amplification product will bear the portion of the duplex tag that tagged the original Crick strand. In this way, a sequence read generated from an amplified product of a complement of the Watson strand can be distinguished from a complement sequence read generated from an amplification product of the Crick strand of the original molecule.

All amplified polynucleotides can be submitted to a sequencing device for sequencing. Alternatively, a sampling, or subset, of all of the amplified polynucleotides is submitted to a sequencing device for sequencing. With respect to any original double-stranded polynucleotide there can be three results with respect to sequencing. First, sequence reads can be generated from both complementary strands of the original molecule (that is, from both the Watson strand and from the Crick strand). Second, sequence reads can be generated from only one of the two complementary strands (that is, either from the Watson strand or from the Crick strand, but not both). Third, no sequence read may be generated from either of the two complementary strands. Consequently, counting unique sequence reads mapping to a genetic locus will underestimate the number of double-stranded polynucleotides in the original sample mapping to the locus. Described herein are methods of estimating the unseen and uncounted polynucleotides.

The sequencing method can be massively parallel sequencing, that is, simultaneously (or in rapid succession) sequencing any of at least 100, 1000, 10,000, 100,000, 1 million, 10 million, 100 million, or 1 billion polynucleotide molecules. Sequencing methods may include, but are not limited to: high-throughput sequencing, pyrosequencing, sequencing-by-synthesis, single-molecule sequencing, nanopore sequencing, semiconductor sequencing, sequencing-by-ligation, sequencing-by-hybridization, RNA-Seq (Illumina), Digital Gene Expression (Helicos), Next generation sequencing, Single Molecule Sequencing by Synthesis (SMSS) (Helicos), massively-parallel sequencing, Clonal Single Molecule Array (Solexa), shotgun sequencing, Maxam-Gilbert or Sanger sequencing, primer walking, sequencing using PacBio, SOLiD, Ion Torrent, or Nanopore platforms and any other sequencing methods known in the art.

For example, duplex-tagged polynucleotides can be amplified, by for example PCR (see e.g., FIG. **4**A duplex-tagged polynucleotides are referred to as mm' and nn'). In FIG. **4**A, the strand of the duplex polynucleotide including sequence m bears sequence tags w and y, while the strand of the duplex polynucleotide including sequence m' bears sequence tags x and z. Similarly, the strand of the duplex polynucleotide including sequence n bears sequence tags a and c, while the strand of the duplex polynucleotide including sequence n' bears sequence tags b and d. During amplification, each strand produces itself and its complementary sequence. However, for example, an amplification progeny of original strand m that includes the complementary sequence, m', is distinguishable from an amplification progeny of original strand m' because the progeny from original strand m will have the sequence 5'-y'm'w'-3' and the progeny of the original m' strand one strand will have the sequence 5'-zm'x-3'. FIG. **4**B shows amplification in more detail. During amplification, errors can be introduced into the amplification progeny, represented by dots. The application progeny are sampled for sequencing, so that not all strands produce sequence reads, resulting in the sequence reads indicated. Because sequence reads can come from either of a strand or its complement, both sequences and complement sequences will be included in the set of sequence reads. It should be noted that it is possible that a polynucleotide would bear the same tag on each end. Thus, for a tag "a", and polynucleotide "m", a first strand could be tagged a-m-a', and the complement could be tagged a-m'-a.

E. Determining Consensus Sequence Reads

Methods disclosed herein can comprise determining consensus sequence reads in sequence reads (e.g., as shown in step (**108**), FIG. **1**), such as by reducing or tracking redundancy. Sequencing of amplified polynucleotides can produce reads of the several amplification products from the same original polynucleotide, referred to as "redundant reads". By identifying redundant reads, unique molecules in the original sample can be determined. If the molecules in a sample are uniquely tagged, then reads generated from amplification of a single unique original molecule can be identified based on their distinct barcode. Ignoring barcodes, reads from unique original molecules can be determined based on sequences at the beginning and end of a read, optionally in combination with the length of the read. In certain cases, however, a sample may be expected to have a plurality of original molecules having the same start stop sequences and the same length. Without barcoding, these molecules are difficult to distinguish from one another. However, if a collection of polynucleotides is non-uniquely tagged (that is, an original molecule shares the same identifier with at least one other original molecule), combining information from a barcode with start/stop sequence and/or polynucleotide length significantly increases the probability that any sequence read can be traced back to an original polynucleotide. This is because, in part, even without unique tagging, it is unlikely that any two original polynucleotides having the same start/stop sequence and length also will be tagged with the same identifier.

F. Collapsing

Collapsing allows for reduction in noise (i.e., background) that is generated at each step of the process. Methods disclosed herein can comprise collapsing, e.g., generating a consensus sequence by comparing multiple sequence reads. For example, sequence reads generated from a single original polynucleotide can be used to generate a consensus sequence of that original polynucleotide. Iterative rounds of amplification can introduce errors into progeny polynucleotides. Also, sequencing typically may not be performed with perfect fidelity so sequencing errors are introduced at this stage as well. However, comparison of sequence reads of molecules derived from a single original molecule, including those that have sequence variants, can be analyzed so as to determine the original, or "consensus" sequence. This can be done phylogenetically. Consensus sequences can be generated from families of sequence reads by any of a variety of methods. Such methods include, for example, linear or non-linear methods of building consensus sequences (such as voting (e.g., biased voting), averaging, statistical, maximum a posteriori or maximum likelihood detection, dynamic programming, Bayesian, hidden Markov or support vector machine methods, etc.) derived from digital communication theory, information theory, or bioinformatics. For example, if all or most of the sequence reads tracking back to an original molecule bear the same sequence variant, that variant probably existed in the original molecule. On the other hand, if a sequence variant exists in a subset of redundant sequence reads, that variant may have been introduced during amplification/sequencing and represents an artifact not existing in the original. Furthermore, if only sequence reads derived from the Watson or Crick strand of an original polynucleotide contain the variant, the variant may have been introduced through single-sided DNA damage, first-cycle PCR error or through contaminating polynucleotides that were amplified from a different sample.

After fragments are amplified and the sequences of amplified fragments are read and aligned, the fragments are subjected to base calling, e.g., determining for each locus the most likely nucleotide. However, variations in the number of amplified fragments and unseen amplified fragments (e.g., those without being read their sequences; reasons could be too many such as amplification errors, sequencing reading errors, too long, too short, being chopped, etc.) may introduce errors in base calling. If there are too many unseen amplified fragments with respect to the seen amplified fragments (amplified fragments actually being read), the reliability of base calling may be diminished.

Therefore, disclosed herein is a method to correct for the number of unseen fragments in base calling. For example, when base calling for locus A (an arbitrary locus), it is first assumed that there are N amplified fragments. The sequence readouts can come from two types of fragments: double-strand fragments and single-strand fragments. Therefore, we assign N1, N2, and N3 as the numbers of double-strands, single-strands, and unseen fragments, respectively. Thus, N=N1+N2+N3 (N1 and N2 are known from the sequence readouts, and N and N3 are unknown). If the formula is solved for N (or N3), then N3 (or N) will be inferred.

Probability is used to estimate N. For example, we assign "p" to be the probability of having detected (or having read) a nucleotide of locus A in a sequence readout of a single-strand.

For sequence readouts from double-strands, the nucleotide call from a double-strand amplified fragment has a probability of p*p=p^2, seeing all N1 double-strands has the following equation: N1=N*(p^2).

For sequence readouts from a single-strand. Assuming that one of the 2 strands is seen, and the other is unseen, the probability of seeing one strand is "p", but the probability of missing the other strand is (1−p). Furthermore, by not distinguishing the single strand sourcing from 5-primer and sourcing from 3-primer, there is a factor of 2. Therefore, the nucleotide call from a single-strand amplified fragment has

a probability 2×p×(1−p). Thus, seeing all N2 single-strands has the following equation: N2=N×2×p×(1−p).

"p" is also unknown. To solve p, the ratio of N1 to N2 is used to solve for "p":

$$R = \frac{N1}{N2} = \frac{Np^2}{2Np(1-p)} = \frac{p^2}{2p(1-p)} = \frac{p}{2(1-p)}$$

Once "p" is found, N can be found. After N is found, can be found N3=N−N1−N2.

Besides the ratio of paired versus unpaired strands (which is a measure post-collapsing), there is useful information in the pre-collapsing read depth at each locus. This information can be used to further improve the call for total molecule count and/or increase confidence of calling variants.

For example, FIG. **4**C demonstrates sequence reads corrected for complementary sequences. Sequences generated from an original Watson strand or an original Crick strand can be differentiated on the basis of their duplex tags. Sequences generated from the same original strand can be grouped. Examination of the sequences can allow one to infer the sequence of the original strand (the "consensus sequence"). In this case, for example, the sequence variant in the nn' molecule is included in the consensus sequence because it included in every sequence read while other variants are seen to be stray errors. After collapsing sequences, original polynucleotide pairs can be identified based on their complementary sequences and duplex tags.

FIG. **5** demonstrates increased confidence in detecting sequence variants by pairing reads from Watson and Crick strands. Sequence nn' can include a sequence variant indicated by a dot. In some cases, sequence pp' does not include a sequence variant. Amplification, sequencing, redundancy reduction and pairing can result in both Watson and Crick strands of the same original molecule including the sequence variant. In contrast, as a result of errors introduced during amplification and sampling during sequencing, the consensus sequence of the Watson strand p can contain a sequence variant, while the consensus sequence of the Crick strand p' does not. It is less likely that amplification and sequencing will introduce the same variant into both strands (nn' sequence) of a duplex than onto one strand (pp' sequence). Therefore, the variant in the pp' sequence is more likely to be an artifact, and the variant in the nn' sequence is more likely to exist in the original molecule.

Methods disclosed herein can be used to correct errors resulted from experiments, e.g., PCR, amplification, and/or sequencing. For example, such a method can comprises attaching one or more double stranded adaptors to both ends of a double stranded polynucleotide, thereby providing a tagged double stranded polynucleotide; amplifying the double stranded tagged polynucleotide; sequencing both strands of the tagged polynucleotide; comparing the sequence of one strand with its complement to determine any errors introduced during sequencing; and correcting errors in the sequence based on (d). The adaptors used in this method can be any adaptors disclosed herein, e.g., Y-shaped adaptors. The adaptor can comprise any barcodes (e.g., distinct barcodes) disclosed herein.

G. Mapping

Sequence reads or consensus sequences can be mapped to one or more selected genetic loci (e.g., as shown step (**110**), FIG. **1**). A genetic locus can be, for example, a specific nucleotide position in the genome, a sequence of nucleotides (for example, an open reading frame), a fragment of a chromosome, a whole chromosome, or an entire genome. A genetic locus can be a polymorphic locus. Polymorphic locus can be a locus at which sequence variation exists in the population and/or exists in a subject and/or a sample. A polymorphic locus can be generated by two or more distinct sequences coexisting at the same location of the genome. The distinct sequences can differ from one another by one or more nucleotide substitutions, a deletion/insertion, and/or a duplication of any number of nucleotides, generally a relatively small number of nucleotides, such as less than 50, 45, 40, 35, 30, 25, 24, 23, 22, 21, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1 nucleotide(s), among others. A polymorphic locus can be created by a single nucleotide position that varies within the population, e.g. a single nucleotide variation (SNV) or a single nucleotide polymorphism (SNP).

A reference genome for mapping can include the genome of any species of interest. Human genome sequences useful as references can include the hg19 assembly or any previous or available hg assembly. Such sequences can be interrogated using the genome browser available at genome.ucsc.edu/index.html. Other species genomes include, for example PanTro2 (chimp) and mm9 (mouse).

In methods disclosed herein, collapsing can be performed before or after mapping. In some aspects, collapsing can be performed before mapping. For example, sequence reads can be grouped into families based on their tags and one or more endogenous sequences, without regard to where the reads map in the genome. Then, the members of a family can be collapsed into a consensus sequence. The consensus sequence can be generated using any collapsing method disclosed herein. Then the consensus sequence can be mapped to locations in the genome. Reads mapped to a locus can be quantified (e.g., counted). Percentage of reads carrying a mutation at a locus can also be determined. Alternatively, collapsing can be performed after mapping. For example, all reads can first be mapped to the genome. Then the reads can be grouped into families based on their tags and one or more endogenous sequences. Since the reads have been mapped to the genome, consensus bases can be determined for each family at each locus. In other aspects, consensus sequence can be generated for one strand of a DNA molecule (e.g., for a Watson strand or a Crick strand). Mapping can be performed before or after the consensus sequence for one strand of the DNA molecule is determined. Numbers of Doublets and Singlets can be determined. These numbers can be used to calculate unseen molecules. For example, the unseen molecules can be calculated using the following equation: N=D+S+U; D=Np(2), S=N2pq, where p=1−q, where p is the probability of seeing; q is the probability of missing a strand.

H. Grouping

Methods disclosed herein can also comprise grouping sequence reads. Sequence reads can be grouped based on various types of sequences, e.g., sequences of an oligonucleotide tag (e.g., a barcode), sequence of a polynucleotide fragments, or any combinations. For example, as shown in step (**112**) (FIG. **1**), sequence reads can be grouped as follows: Sequence reads generated from a "Watson" strand and those generated from a "Crick" strand of a double-stranded polynucleotide in the sample are identifiable based on the duplex tags that they bear. In this way, a sequence read or consensus sequence from a Watson strand of a duplex polynucleotide can be paired with a sequence read or consensus sequence from its complementary Crick strand. Paired sequence reads are referred to as a "Pair".

Sequence reads for which no sequence read corresponding to a complementary strand can be found among the sequence reads are termed "Singlets".

Double-stranded polynucleotides for which a sequence read for neither of the two complementary strands has been generated are referred to as "Unseen" molecules.

I. Quantifying

Methods disclosed herein also comprise quantifying sequence reads. For example, as shown in step (**114**) (FIG. **1**), Pairs and Singlets mapping to a selected genetic locus, or to each of a plurality of selected genetic loci, are quantified, e.g., counted.

The quantifying can comprise estimating number of polynucleotides in the sample (e.g., Pairs polynucleotides, Singlets polynucleotides, or Unseen polynucleotides. For example, as shown in step (**116**) (FIG. **1**), the number of double-stranded polynucleotides in the sample for which no sequence reads were generated ("Unseen" polynucleotides) is estimated. The probability that a double strand polynucleotide generates no sequence reads can be determined based on the relative number of Pairs and Singlets at any locus. Using this probability, the number of Unseen polynucleotide can be estimated.

In step (**118**) an estimate for the total number of double-stranded polynucleotides in a sample mapping to a selected locus is the sum of the number of Pairs, the number of Singlets and the number of Unseen molecules mapping to the locus.

The number of Unseen original molecules in a sample can be estimated based on the relative number of Pairs and Singlets (FIG. **2**). Referring to FIG. **2**, as an example, counts for a particular genomic locus, Locus A, are recorded, where 1000 molecules are paired and 1000 molecules are unpaired. Assuming a uniform probability, p, for an individual Watson or Crick strand to make it through the process subsequent to conversion, one can calculate the proportion of molecules that fail to make it through the process (Unseen) as follows: Let R=ratio of paired to unpaired molecules=1, so $R=1=p^2/(2p(1-p))$. This implies that p=2/3 and that the quantity of lost molecules is equal to $(1-p)^2=1/9$. Thus in this example, approximately 11% of converted molecules are lost and never detected. Consider another genomic locus, Locus B, in the same sample where 1440 molecules are paired and 720 are unpaired. Using the same method, we can infer the number of molecules that are lost, is only 4%. Comparing the two areas, it may be assumed that Locus A had 2000 unique molecules as compared to 2160 molecules in Locus B—a difference of almost 8%. However, by correctly adding in the lost molecules in each region, we infer there are 2000/(8/9)=2250 molecules in Locus A and 2160/0.96=2250 molecules in Locus B. Hence, the counts in both regions are actually equal. This correction and thus much higher sensitivity can be achievable by converting the original double-stranded nucleic acid molecules and bioinformatically keeping track of all those that are paired and unpaired at the end of the process. Similarly, the same procedure can be used to infer true copy number variations in regions that appear to have similar counts of observed unique molecules. By taking the number of unseen molecules into consideration in the two or more regions, the copy number variation becomes apparent.

In addition to using binomial distribution, other methods of estimating numbers of unseen molecules include exponential, beta, gamma or empirical distributions based on the redundancy of sequence reads observed. In the latter case, the distribution of read counts for paired and unpaired molecules can be derived from such redundancy to infer the underlying distribution of original polynucleotide molecules at a particular locus. This can often lead to a better estimation of the number of unseen molecules.

J. CNV Detection

Methods disclosed herein also comprise detecting CNV. For example, as shown in step (**120**) (FIG. **1**), once the total number of polynucleotides mapping to a locus is determined, this number can be used in standard methods of determining CNV at the locus. A quantitative measure can be normalized against a standard. The standard can be an amount of any polynucleotides. In one method, a quantitative measure at a test locus can be standardized against a quantitative measure of polynucleotides mapping to a control locus in the genome, such as gene of known copy number. Quantitative measures can be compared against the amount of nucleic acid in any sample disclosed herein. For example, in another method, the quantitative measure can be compared against the amount of nucleic acid in the original sample. For example, if the original sample contained 10,000 haploid gene equivalents, the quantitative measure can be compared against an expected measure for diploidy. In another method, the quantitative measure can be normalized against a measure from a control sample, and normalized measures at different loci can be compared.

In some cases, in which copy number variation analysis is desired, sequence data may be: 1) aligned with a reference genome; 2) filtered and mapped; 3) partitioned into windows or bins of sequence; 4) coverage reads counted for each window; 5) coverage reads can then be normalized using a stochastic or statistical modeling algorithm; 6) and an output file can be generated reflecting discrete copy number states at various positions in the genome. In other cases, in which rare mutation analysis is desired, sequence data may be 1) aligned with a reference genome; 2) filtered and mapped; 3) frequency of variant bases calculated based on coverage reads for that specific base; 4) variant base frequency normalized using a stochastic, statistical or probabilistic modeling algorithm; 5) and an output file can be generated reflecting mutation states at various positions in the genome.

After the sequence read coverage ratios have been determined, a stochastic modeling algorithm can be optionally applied to convert the normalized ratios for each window region into discrete copy number states. In some cases, this algorithm may comprise a Hidden Markov Model. In other cases, the stochastic model may comprise dynamic programming, support vector machine, Bayesian modeling, probabilistic modeling, trellis decoding, Viterbi decoding, expectation maximization, Kalman filtering methodologies, or neural networks.

Methods disclosed herein can comprise detecting SNVs, CNVs, insertions, deletions, and/or rearrangements at a specific region in a genome. The specific genomic region can comprise a sequence in a gene, such as ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDHL IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, or NTRK1.

In some cases, the method uses a panel which comprises exons of one or more genes. The panel can comprise introns of one or more genes as well. The panel can also comprise

exons and introns of one or more genes. The one or more genes can be those disclosed above. The panel can comprise about 80,000 bases which cover a panel of genes. The panel can comprise about 1000, 2000, 3000, 4000, 5000, 10000, 15000, 20000, 25000, 30000, 35000, 40000, 45000, 50000, 55000, 60000, 65000, 70000, 75000, 80000, 85000, 90000, 95000, 100000, 105000, 110000, 115000, 120000, 125000, or more bases.

In some aspects, copy number of a gene can be reflected in the frequency of a genetic form of the gene in a sample. For example, in a healthy individual, no copy number variation is reflected in a variant in a gene in one chromosome (e.g., heterozygosity) being detected in about 50% of detected molecules in a sample. Also, in a healthy individual, duplication of a gene bearing a variant can be reflected in the variant being detected in about 66% of detected molecules in a sample. Accordingly, if the tumor burden in a DNA sample is 10%, the frequency of a somatic mutation in a gene in one chromosome of cancer cells, without CNV, can be about 5%. The converse can be true in the case of aneuploidy.

The methods disclosed herein can be used to determine whether a sequence variant is more likely present in the germ line level or resulted from a somatic cell mutation, e.g., in a cancer cell. For example, a sequence variant in a gene detected at levels arguably consistent with heterozygosity in the germ line is more likely the product of a somatic mutation if CNV is also detected in that gene. In some cases, to the extent we expect that a gene duplication in the germ line bears a variant consistent with genetic dose (e.g., 66% for trisomy at a locus), detection gene amplification with a sequence variant dose that deviates significantly from this expected amount indicates that the CNV is more likely present as a result of somatic cell mutation.

The methods disclosed herein can also be used to infer tumor heterogeneity in a situation in which sequence variants in two genes are detected at different frequencies. For example, tumor heterogeneity can be inferred when two genes are detected at different frequencies but their copy numbers are relatively equal. Alternatively, tumor homogeneity can be inferred when the difference in frequency between two sequence variants is consistent with difference in copy number for the two genes. Thus, for example, if an EGFR variant is detected at 11% and a KRAS variant is detected at 5%, and no CNV is detected at these genes, the difference in frequency likely reflects tumor heterogeneity (e.g., all tumor cells carry an EGFR mutant and half the tumor cells also carry a KRAS mutant). Alternatively, if the EGFR gene carrying the mutant is detected at 2-times normal copy number, one interpretation is a homogenous population of tumor cells, each cell carrying a mutant in the EGFR and KRAS genes, but in which the KRAS gene is duplicated.

In response to chemotherapy, a dominant tumor form can eventually give way through Darwinian selection to cancer cells carrying mutants that render the cancer unresponsive to the therapy regimen. Appearance of these resistance mutants can be delayed through methods of this invention. In one embodiment of this method, a subject is subjected to one or more pulsed therapy cycles, each pulsed therapy cycle comprising a first period during which a drug is administered at a first amount and a second cycle during which the drug is administered at a second, reduced amount. The first period can be characterized by a tumor burden detected above a first clinical level. The second period can be characterized by a tumor burden detected below a second clinical level. First and second clinical levels can be different in different pulsed therapy cycles. For example, the first clinical level can be lower in succeeding cycles. A plurality of cycles can include at least 2, 3, 4, 5, 6, 7, 8 or more cycles. For example, the BRAF mutant V600E may be detected in polynucleotides of a disease cell at an amount indicating a tumor burden of 5% in cfDNA. Chemotherapy can commence with dabrafenib. Subsequent testing can show that the amount of the BRAF mutant in the cfDNA falls below 0.5% or to undetectable levels. At this point, dabrafenib therapy can stop or be significantly curtailed. Further subsequent testing may find that DNA bearing the BRAF mutation has risen to 2.5% of polynucleotides in cfDNA. At this point, dabrafenib therapy can be re-started, e.g., at the same level as the initial treatment. Subsequent testing may find that DNA bearing the BRAF mutation has decreased to 0.5% of polynucleotides in cfDNA. Again, dabrafenib therapy can be stopped or reduced. The cycle can be repeated a number of times.

A therapeutic intervention can also be changed upon detection of the rise of a mutant form resistant to an original drug. For example, cancers with the EGFR mutation L858R respond to therapy with erlotinib. However, cancers with the EGFR mutation T790M are resistant to erlotinib. However, they are responsive to ruxolitinib. A method of this invention involves monitoring changes in tumor profile and changing a therapeutic intervention when a genetic variant associated with drug resistance rises to a predetermined clinical level.

Methods disclosed in this invention can comprise a method of detecting disease cell heterogeneity from a sample comprising polynucleotides from somatic cells and disease cells, the method comprising: a) quantifying polynucleotides in the sample bearing a sequence variant at each of a plurality of genetic loci; b) determining CNV at each of the plurality of genetic loci; different relative amounts of disease molecules at a locus, wherein the CNV indicates a genetic dose of a locus in the disease cell polynucleotides; c) determining a relative measure of quantity of polynucleotides bearing a sequence variant at a locus per genetic dose at the locus for each of a plurality of the loci; and d) comparing the relative measures at each of the plurality of loci, wherein different relative measures indicates tumor heterogeneity. In the methods disclosed herein, the genetic dose can be determined on a total molecule basis. For example, if there are 1× total molecules at a first locus, and 1.2× molecules mapped to a second locus, then the genetic dose is 1.2. Variants at this locus can be divided by 1.2. In some aspects, the method disclosed herein can be used to detect any disease cell heterogeneity, e.g., tumor cell heterogeneity. The methods can be used to detect disease cell heterogeneity from a sample comprising any types of polynucleotides, e.g., cfDNA, genomic DNA, cDNA, or ctDNA. In the methods, the quantifying can comprise, for example, determining the number or relative amount of the polynucleotides. Determining CNV can comprise mapping and normalizing different relative amounts of total molecules to a locus.

In another aspect, in response to chemotherapy, a dominant tumor form can eventually give way through Darwinian selection to cancer cells carrying mutants that render the cancer unresponsive to the therapy regimen. Appearance of these resistance mutants can be delayed through methods disclosed throughout. The methods disclosed herein can comprise a method comprising: a) subjecting a subject to one or more pulsed therapy cycles, each pulsed therapy cycle comprising (i) a first period during which a drug is administered at a first amount and (ii) a second period during which the drug is administered at a second, reduced amount;

wherein (A) the first period is characterized by a tumor burden detected above a first clinical level; and (B) the second period is characterized by a tumor burden detected below a second clinical level.

K. Sequence Variant Detection

Systems and methods disclosed herein can be used to detect sequence variants, e.g., SNVs. For example, a sequence variant can be detected from consensus sequences from multiple sequence reads, for example, from at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, at least 30, at least 31, at least 32, at least 33, at least 34, at least 35, at least 36, at least 37, at least 38, at least 39, at least 40, at least 41, at least 42, at least 43, at least 44, at least 45, at least 46, at least 47, at least 48, at least 49, at least 50, at least 51, at least 52, at least 53, at least 54, at least 55, at least 56, at least 57, at least 58, at least 59, at least 60, at least 61, at least 62, at least 63, at least 64, at least 65, at least 66, at least 67, at least 68, at least 69, at least 70, at least 71, at least 72, at least 73, at least 74, at least 75, at least 76, at least 77, at least 78, at least 79, at least 80, at least 81, at least 82, at least 83, at least 84, at least 85, at least 86, at least 87, at least 88, at least 89, at least 90, at least 91, at least 92, at least 93, at least 94, at least 95, at least 96, at least 97, at least 98, at least 99, at least 100, at least 200, at least 300, at least 400, at least 500, at least 600, at least 700, at least 800, at least 900, at least 1000, at least 2000, at least 3000, at least 4000, at least 5000, at least 6000, at least 7000, at least 8000, at least 9000, at least 10000 or more sequence reads. A consensus sequence can be from sequence reads of a single strand polynucleotide. A consensus sequence can also be from sequence reads of one strand of a double-stranded polynucleotide (e.g., pairing reads). In an exemplary method, pairing reads allows one to identify with increased confidence the existence of a sequence variant in a molecule. For example, if both strands of a Pair include the same variant, one can be reasonably sure that the variant existed in the original molecule, as the chance that the same variant is introduced into both strands during amplification/sequencing is rare. In contrast, if only one strand of a Pair includes the sequence variant, this is more likely to be an artifact. Similarly, the confidence that a Singlet bearing a sequence variant existed in the original molecule is less than the confidence if the variant exists in a Duplex, as there is higher probability that the variant can be introduced once than twice during amplification/sequencing.

Other methods of copy number variation detection and the sequence variant detection are described in PCT/US2013/058061, which is entirely incorporated herein by reference.

Sequence reads can be collapsed to generate a consensus sequence, which can be mapped to a reference sequence to identify genetic variants, such as CNV or SNV. As an alternative, the sequence reads are mapped prior to or even without mapping. In such a case, the sequence reads can be individually mapped to the reference to identify a CNV or SNV.

FIG. **3** shows a reference sequence encoding a genetic Locus A. The polynucleotides in FIG. **3** may be Y-shaped or have other shapes, such as hairpin.

In some cases, an SNV or multiple-nucleotide variant (MNV) can be determined across multiple sequence reads at a given locus (e.g., nucleotide base) by aligning sequence reads that correspond to that locus. Next, a plurality of sequential nucleotide bases from at least a subset of the sequence reads are mapped to the reference to a SNV or MNV in a polynucleotide molecule or portion thereof that corresponds to the reads. The plurality of sequential nucleotide bases can span an actual, inferred or suspected location of the SNV or MNV. The plurality of sequential nucleotide bases can span at least 3, 4, 5, 6, 7, 8, 9, or 10 nucleotide bases.

L. Detecting/Quantifying Nucleic Acids

The methods described throughout can be used to tag nucleic acids fragments, such as deoxyribonucleic acid (DNA), at extremely high efficiency. This efficient tagging allows a person to efficiently and accurately detect rare DNA in heterogenous populations of original DNA fragments (such as in cfDNA). A rare polynucleotide (e.g., rare DNA) can be a polynucleotide that comprises a genetic variant occurring in a population of polynucleotides at a frequency of less than 10%, 5%, 4%, 3%, 2%, 1%, or 0.1%. A rare DNA can be a polynucleotide with a detectable property at a concentration less than 50%, 25%, 10%, 5%, 1%, or 0.1%

Tagging can occur in a single reaction. In some cases, two or more reactions can be performed and pooled together. Tagging each original DNA fragments in a single reaction can result in tagging such that greater than 50% (e.g., 60%, 70%, 80%, 90%, 95%, or 99%) of the original DNA fragments are tagged at both ends with tags that comprise molecular barcodes, thereby providing tagged DNA fragments. Tagging can also result in greater than 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% of the original DNA fragments tagged at both ends with tags that comprise molecular barcodes. Tagging can also result in 100% of the original DNA fragments tagged at both ends with tags that comprise molecular barcodes. Tagging can also result in single end tagging.

Tagging can also occur by using an excess amount of tags as compared to the original DNA fragments. For example, the excess can be at least 5-fold excess. In other cases, the excess can be at least 1.25, 1.5, 1.75, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100 or more fold excess. Tagging can comprise attachment to blunt ends or sticky ends. Tagging can also be performed by hybridization PCR. Tagging can also be performed in low reaction volumes, such as 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, or 100 pico- and/or microliters.

The method can also include performing high fidelity amplification on the tagged DNA fragments. Any high fidelity DNA polymerases can be used. For example, the polymerase may be KAPA HiFi DNA polymerase or Phusion DNA polymerase.

Further, the method can comprise selectively enriching a subset of the tagged DNA fragments. For example, selective enrichment can be performed by hybridization or amplification techniques. The selective enrichment can be performed using a solid support (e.g., beads). The solid support (e.g., beads) can comprise probes (e.g., oligonucleotides specifically hybridizing to certain sequences. For example, the probes can hybridize with certain genomic regions, e.g., genes. In some cases, the genomic regions, e.g., genes, can

be regions associated with diseases, e.g., cancer. After enrichment, the selected fragmented can be attached any sequencing adaptor disclosed in this invention. For example, a sequence adaptor can comprise a flow cell sequence, a sample barcode, or both. In another example, a sequence adaptor can be a hairpin shaped adaptor and/or comprises a sample barcode. Further, the resulting fragments can be amplified and sequenced. In some cases, the adaptor does not comprise a sequencing primer region.

The method can include sequencing one or both strands of the DNA fragments. In one case, both strands of the DNA fragment are independently sequenced. The tagged, amplified, and/or selectively enriched DNA fragments are sequenced to obtain sequence reads that comprise sequence information of the molecular barcodes and at least a portion of the original DNA fragments.

The method can include reducing or tracking redundancy (as described above) in the sequence reads to determine consensus reads that are representative of single-strands of the original DNA fragments. For example, to reduce or track redundancy, the method can include comparing sequence reads having the same or similar molecular barcodes and the same or similar end of fragment sequences. The method can comprise performing a phylogentic analysis on the sequence reads having the same or similar molecular barcodes. The molecular barcodes can have a barcode with varying edit distances (including any edit distances as described throughout), for example, an edit distance of up to 3. The end of the fragment sequences can include fragment sequences having an edit distance with varying distances (including any edit distances as described throughout), for example, an edit distance of up to 3.

The method can comprise binning the sequence reads according to the molecular barcodes and sequence information. For example, binning the sequence reads according to the molecular barcodes and sequence information can be performed from at least one end of each of the original DNA fragments to create bins of single stranded reads. The method can further comprise in each bin, determining a sequence of a given original DNA fragment among the original DNA fragments by analyzing sequence reads.

In some cases, sequence reads in each bin can be collapsed to a consensus sequence and subsequently mapped to a genome. As an alternative, sequence reads can be mapped to a genome prior to binning and subsequently collapsed to a consensus sequence.

The method can also comprise sorting sequence reads into paired reads and unpaired reads. After sorting, the number of paired reads and unpaired reads that map to each of one or more genetic loci can be quantified.

The method can include quantifying the consensus reads to detect and/or quantify the rare DNA, which are described throughout. The method can comprise detecting and/or quantifying the rare DNA by comparing a number of times each base occurs at each position of a genome represented by the tagged, amplified, and/or enriched DNA fragments.

The method can comprise tagging the original DNA fragments in a single reaction using a library of tags. The library can include at least 2, at least 3, at least 4, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 50, at least 100, at least 500, at least 1000, at least 5000, at least 10000, or any number of tags as disclosed throughout. For example, the library of tags can include at least 8 tags. The library of tags can include 8 tags (which can generate 64 different possible combinations). The method can be conducted such that a high percentage of fragments, e.g., greater than 50% (or any percentages as described throughout) are tagged at both ends, wherein each of the tags comprises a molecular barcode.

M. Processing and/or Analyzing Nucleic Acids

The methods described throughout can be used for processing and/or analyzing a nucleic acid sample of a subject. The method can comprising exposing polynucleotide fragments of the nucleic acid sample to a plurality of polynucleotide molecules to yield tagged polynucleotide fragments. The plurality of polynucleotide molecules that can be used are described throughout the application.

For example, the plurality of polynucleotide molecules can be each less than or equal to 40 nucleic acid bases in length and have distinct barcode sequences with respect to at least 4 nucleic acid bases and an edit distance of at least 1, wherein each of the distinct barcode sequences is within 20 nucleic acid bases from a terminal end of a respective one of the plurality of polynucleotide molecules, and wherein the plurality of polynucleotide molecules are not sequencing adaptors.

The tagged polynucleotide fragments can be subjected to nucleic acid amplification reactions under conditions that yield amplified polynucleotide fragments as amplification products of the tagged polynucleotide fragments. After amplification, the nucleotide sequence of the amplified tagged polynucleotide fragments is determined. In some cases, the nucleotide sequences of the amplified tagged polynucleotide fragments are determined without the use of polymerase chain reaction (PCR).

The method can comprise analyzing the nucleotide sequences with a programmed computer processor to identify one or more genetic variants in the nucleotide sample of the subject. Any genetic alterations can be identified, including but not limited to, base change(s), insertion(s), repeat(s), deletion(s), copy number variation(s), epigenetic modification(s), nucleosome binding site(s), copy number change(s) due to origin(s) of replication, and transversion(s). Other genetic alterations can include, but are not limited to, one or more tumor associated genetic alterations.

The subject of the methods can be suspected of having a disease. For example, the subject can be suspected of having cancer. The method can comprise collecting a nucleic acid sample from a subject. The nucleic acid sample can be collected from blood, plasma, serum, urine, saliva, mucosal excretions, sputum, stool, cerebral spinal fluid, skin, hair, sweat, and/or tears. The nucleic acid sample can be a cell-free nucleic acid sample. In some cases, the nucleic acid sample is collected from no more than 100 nanograms (ng) of double-stranded polynucleotide molecules of the subject.

The polynucleotide fragments can comprise double-stranded polynucleotide molecules. In some cases, the plurality of polynucleotide molecules are coupled to the polynucleotide fragments via blunt end ligation, sticky end ligation, molecular inversion probes, polymerase chain reaction (PCR), ligation-based PCR, multiplex PCR, single strand ligation, or single strand circularization.

The method as described herein results in high efficiency tagging of nucleic acids. For example, exposing the polynucleotide fragments of the nucleic acid sample to the plurality of polynucleotide molecules yields the tagged polynucleotide fragments with a conversion efficiency of at least 30%, e.g., of at least 50% (e.g., 60%, 70%, 80%, 90%, 95%, or 99%). Conversion efficiency of at least 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%,

78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99% can be achieved.

The method can result in a tagged polynucleotide fragment that share common polynucleotide molecules. For example, any of at least 5%, 6%, 7%, 8%, 9%, 10%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, 99% or 100% of the tagged polynucleotide fragments share a common polynucleotide molecule. The method can comprise generating the polynucleotide fragments from the nucleic acid sample.

In some cases, the subjecting of the method comprises amplifying the tagged polynucleotide fragments in the presence primers corresponding to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. Additionally, any combination of these genes can be amplified. For example, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, or all 54 of these genes can be amplified.

The methods described herein can comprise generating a plurality of sequence reads from a plurality of polynucleotide molecules. The plurality of polynucleotide molecules can cover genomic loci of a target genome. For example, the genomic loci can correspond to a plurality of genes as listed above. Further, the genomic loci can be any combination of these genes. Any given genomic locus can comprise at least two nucleic acid bases. Any given genomic locus can also comprise a plurality of nucleic acid bases, for example, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, or more nucleic acid bases.

The method can comprise grouping with a computer processor the plurality of sequence reads into families. Each of the family can comprises sequence reads from one of the template polynucleotides. Each family can comprise sequence reads from only one of the template polynucleotides. For each of the family, the sequence reads can be merged to generate a consensus sequence. The grouping can comprise classifying the plurality of sequence reads into families by identifying (i) distinct molecular barcodes coupled to the plurality of polynucleotide molecules and (ii) similarities between the plurality of sequence reads, wherein each family includes a plurality of nucleic acid sequences that are associated with a distinct combination of molecular barcodes and similar or identical sequence reads.

Once merged, a consensus sequence can be called at a given genomic locus among the genomic loci. At any given genomic loci, any of the following can be determined: i) genetic variants among the calls; ii) frequency of a genetic alteration among the calls; iii) total number of calls; and iv) total number of alterations among the calls. The calling can comprise calling at least one nucleic acid base at the given genomic locus. The calling can also comprise calling a plurality of nucleic acid bases at the given genomic locus. In some cases, the calling can comprise phylogenetic analysis, voting (e.g., biased voting), weighing, assigning a probability to each read at the locus in a family, or calling the base with the highest probability. The consensus sequence can be generated by evaluating a quantitative measure or a statistical significance level for each of the sequence reads. If a quantitative measure is performed, the method can comprise use of a binomial distribution, exponential distribution, beta distribution, or empirical distribution. However, frequency of the base at the particular location can also be used for calling, for example, if 51% or more of the reads is a "A" at the location, then the base may be called an "A" at that particular location. The method can further comprise mapping a consensus sequence to a target genome.

The method can further comprising performing consensus calling at an additional genomic locus among the genomic loci. The method can comprise determining a variation in copy number at one of the given genomic locus and additional genomic locus based on counts at the given genomic locus and additional genomic locus.

The methods described herein can comprise providing template polynucleotide molecules and a library of adaptor polynucleotide molecules in a reaction vessel. The adaptor polynucleotide molecules can have from 2 to 1,000 different barcode sequences and in some cases are not sequencing adaptors. Other variations of adaptor polynucleotide molecules are described throughout, which can also be used in the methods.

The polynucleotide molecules of the adaptors can have the same sample tag. The adaptor polynucleotide molecules can be coupled to both ends of the template polynucleotide molecules. The method can comprise coupling the adaptor polynucleotide molecules to the template polynucleotide molecules at an efficiency of at least 30%, e.g., of at least 50% (e.g., 60%, 70%, 80%, 90%, 95%, or 99%), thereby tagging each template polynucleotide with a tagging combination that is among 4 to 1,000,000 different tagging combinations, to produce tagged polynucleotide molecules. In some cases, the reaction can occur in a single reaction vessel. Coupling efficiency can also be at least 30%, 35%, 40%, 45%, 50%, 51%, 52%, 53%, 54%, 55%, 56%, 57%, 58%, 59%, 60%, 61%, 62%, 63%, 64%, 65%, 66%, 67%, 68%, 69%, 70%, 71%, 72%, 73%, 74%, 75%, 76%, 77%, 78%, 79%, 80%, 81%, 82%, 83%, 84%, 85%, 86%, 87%, 88%, 89%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, or 99%. Tagging can be non-unique tagging.

The tagged polynucleotide molecules can then be subject to an amplification reaction under conditions that will yield amplified polynucleotide molecules as amplification products of the tagged polynucleotide molecules. The template polynucleotide molecules can be double-stranded. Further, the template polynucleotide molecules can be blunt ended. In some cases, the amplification reaction comprises non-specifically amplifying the tagged polynucleotide molecules. The amplification reaction can also comprises using a priming site to amplify each of the tagged polynucleotide molecules. The priming site can be a primer, e.g., a universal primer. The priming site can also be a nick.

The method can also comprise sequencing the amplified polynucleotide molecules. The sequencing can comprise (i) subjecting the amplified polynucleotide molecules to an additional amplification reaction under conditions that yield additional amplified polynucleotide molecules as amplification products of the amplified polynucleotide molecules, and/or (ii) sequencing the additional amplified polynucleotide molecules. The additional amplification can be per-

formed in the presence of primers comprising flow cells sequences, which will produce polynucleotide molecules that are capable of binding to a flow cell. The additional amplification can also be performed in the presence of primers comprising sequences for hairpin shaped adaptors. The hairpin shaped adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. The method can further comprise identifying genetic variants upon sequencing the amplified polynucleotide molecules.

The method can further comprising separating polynucleotide molecules comprising one or more given sequences from the amplified polynucleotide molecules, to produce enriched polynucleotide molecules. The method can also comprise amplifying the enriched polynucleotide molecules with primers comprising the flow cell sequences. This amplification with primers comprising flow cell sequences will produce polynucleotide molecules that are capable of binding to a flow cell. The amplification can also be performed in the presence of primers comprising sequences for hairpin shaped adaptors. The hairpin shaped adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times.

Flow cell sequences or hairpin shaped adaptors can be added by non-amplification methods such as through ligation of such sequences. Other techniques such as hybridization methods can be used, e.g., nucleotide overhangs.

The method can be performed without aliquoting the tagged polynucleotide molecules. For example, once the tagged polynucleotide molecule is made, the amplification and sequencing can occur in the same tube without any further preparation.

The methods described herein can be useful in detecting single nucleotide variations (SNV), copy number variations (CNV), insertions, deletions, and/or rearrangements. In some cases, the SNVs, CNVs, insertions, deletions, and/or rearrangements, can be associated with disease, for example, cancer.

N. Monitoring a Patient's Status

Methods disclosed herein can also be used to monitor a patient's disease status. The disease of a subject can be monitored over time to determine a progression of the disease (e.g., regression). Markers indicative of the disease can be monitored in a biological sample of the subject, such as a cell-free DNA sample.

For example, monitoring a subject's cancer status can comprise (a) determining an amount of one or more SNVs or copy numbers of a plurality of genes (e.g., in an exon), (b) repeating such determination at different points in time, and (c) determining if there is a difference in the number of SNVs, level of SNVs, number or level of genomic rearrangements, or copy numbers between (a) and (b). The genes can be selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The genes can be selected from any 5, 10, 15, 20, 30, 40, 50, or all of the genes in this group.

O. Sensitivity and Specificity

Methods disclosed herein can be used to detect cancer polynucleotides in a sample, and cancer in a subject, with high measures of agreement, e.g., high sensitivity and/or specificity. For example, such methods can detect cancer polynucleotides (e.g., rare DNA) in a sample at a concentration that is less than 5%, 1%, 0.5%, 0.1%, 0.05%, or 0.01%, at a specificity of at least 99%, 99.9%, 99.99%, 99.999%, 99.9999%, or 99.99999%. Such polynucleotides may be indicative of cancer or other disease. Further, such methods can detect cancer polynucleotides in a sample with a positive predictive value of at least 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%, 99%, 99.9%, 99.99%, 99.999%, or 99.9999%.

Subjects identified as positive in a test that are in reality positive are referred as true positives (TP). Subjects identified as positive in a test that are in reality negative are referred as false positives (FP). Subjects identified as negative in a test that are in reality negative are referred as true negatives (TN). Subjects identified as negative in a test that are in reality positive are referred as false negatives (FN). Sensitivity is the percentage of actual positives identified in a test as positive. This includes, for example, instances in which one should have found a cancer genetic variant and did. (Sensitivity=TP/(TP+FN).) Specificity is the percentage of actual negatives identified in a test as negative. This includes, for example, instances in which one should have found no cancer genetic variant and did not. Specificity can be calculated using the following equation: Specificity=TN/(TN+FP). Positive predictive value (PPV) can be measured by the percentage of subjects who test positive that are true positives. PPV can be calculated using the following equation: PPV=TP/(TP+FP). Positive predictive value can be increased by increasing sensitivity (e.g., chance of an actual positive being detected) and/or specificity (e.g., chance of not mistaking an actual negative for a positive).

Low conversion rates of polynucleotides into adaptor-tagged polynucleotides can compromise sensitivity as it decreases the chance of converting, and therefore detecting, rare polynucleotide targets. Noise in a test can compromise specificity as it increases the number of false positives detected in a test. Both low conversion rate and noise compromise positive predictive value as they decrease the percentage of true positives and increase the percentage of false positives.

The methods disclosed herein can achieve high levels of agreement, e.g., sensitivity and specificity, leading to high positive predictive values. Methods of increasing sensitivity include high efficiency conversion of polynucleotides into adaptor-tagged polynucleotides in a sample. Methods of increasing specificity include reducing sequencing errors, for example, by molecular tracking.

Methods of the present disclosure can be used to detect genetic variation in non-uniquely tagged initial starting genetic material (e.g., rare DNA) at a concentration that is less than 5%, 1%, 0.5%, 0.1%, 0.05%, or 0.01%, at a specificity of at least 99%, 99.9%, 99.99%, 99.999%, 99.9999%, or 99.99999%. In some aspects, the methods can further comprise converting polynucleotides in the initial starting material at an efficiency of at least at least 10%, at least 20%, at least 30%, at least 40%, at least 50%, at least 60%, at least 70%, at least 80%, or at least 90%. Sequence reads of tagged polynucleotides can be subsequently tracked to generate consensus sequences for polynucleotides with an error rate of no more than 2%, 1%, 0.1%, or 0.01%.

2. Pooling Methods

Disclosed herein are methods of detecting copy number variation and/or sequence variants at one or more genetic loci in a test sample. One embodiment is shown in FIG. **8**. Typically, detecting copy number variation involves determining a quantitative measure (e.g., an absolute or relative number) of polynucleotides mapping to a genetic locus of interest in a genome of a test sample, and comparing that number to a quantitative measure of polynucleotides mapping to that locus in a control sample. In certain methods, the quantitative measure is determined by comparing the number of molecules in the test sample that map to a locus of interest with a number of molecules in the test sample mapping to a reference sequence, e.g., a sequence expected to be present at wild type ploidy number. In some examples, the reference sequence is HG19, build 37, or build 38. The comparison could involve, for example, determining a ratio. Then, this measure is compared with a similar measure determined in a control sample. So, for example, if a test sample has a ratio of 1.5:1 for locus of interest versus reference locus, and a control sample has a ratio of 1:1 for the same loci, one may conclude that the test sample exhibits polyploidy at the locus of interest.

When the test sample and the control sample are analyzed separately, the work flow can introduce distortions between final numbers in the control and test samples.

In one method disclosed herein (e.g., flow chart **800**), polynucleotides are provided from a test and a control sample (**802**). Polynucleotides in a test sample and those in a control sample are tagged with tags that identify the polynucleotides as originating from the test or control sample (a source tag). (**804**.) The tag can be, for example, a polynucleotide sequence or barcode that unambiguously identifies the source.

The polynucleotides in each of the control and test samples also can be tagged with identifier tags that will be carried by all amplification progeny of a polynucleotide. Information from start and end sequences of a polynucleotide and identifier tags can identify sequence reads from polynucleotides amplified from an original parent molecule. Each molecule can be uniquely tagged compared with other molecules in the sample. Alternatively, each molecule need not be uniquely tagged compared with other molecules in the sample. That is, the number of different identifier sequences can be fewer than that the number of molecules in sample. By combining identifier information with start/stop sequence information, the probability of confusing two molecules having the same start/stop sequence is significantly diminished.

Number of different identifiers used to tag a nucleic acid (e.g., cfDNA) can dependent on the number of different haploid genome equivalents. Different identifiers can be used to tag at least 2, least 10, least 100, least 200, least 300, least 400, least 500, least 600, least 700, least 800, least 900, least 1,000, least 2,000, least 3,000, least 4,000, least 5,000, least 6,000, least 7,000, least 8,000, least 9,000, least 10,000 or more different haploid genome equivalents. Accordingly, the number of different identifiers used to tag a nucleic acid sample, e.g., cell-free DNA from 500 to 10,000 different haploid genome equivalents and be between any of 1, 2, 3, 4 and 5 and no more than 100, 90, 80, 70, 60, 50, 40 or 30. For example, the number of different identifier used to tag a nucleic acid sample from 500 to 10,000 different haploid genome equivalents can be 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 or less.

Polynucleotides can be tagged by ligation of adaptors comprising the tags or identifiers before amplification. Ligation can be performed using an enzyme, e.g., a ligase. For example, tagging can be performed using a DNA ligase. The DNA ligase can be a T4 DNA ligase, *E. coli* DNA ligase, and/or mammalian ligase. The mammalian ligase can be DNA ligase I, DNA ligase III, or DNA ligase IV. The ligase may also be a thermostable ligase. Tags can be ligated to a blunt-end of a polynucleotide (blunt-end ligation). Alternatively, tags can be ligated to a sticky end of a polynucleotide (sticky-end ligation). The polynucleotides can be tagged by blunt end ligation using adaptors (e.g., adaptors having forked ends). High efficiency of ligation can be achieved using high excess of adaptors (e.g., more than 1.5×, more than 2×, more than 3×, more than 4×, more than 5×, more than 6×, more than 7×, more than 8×, more than 9×, more than 10×, more than 11×, more than 12×, more than 13×, more than 14×, more than 15×, more than 20×, more than 25×, more than 30×, more than 35×, more than 40×, more than 45×, more than 50×, more than 55×, more than 60×, more than 65×, more than 70×, more than 75×, more than 80×, more than 85×, more than 90×, more than 95×, or more than 100).

Once tagged with tags that identify source of the polynucleotides, polynucleotides from different sources (e.g., different samples) can be pooled. After pooling, polynucleotides from different sources (e.g., different samples) can be distinguished by any measurement using the tags, including any process of quantitative measurement. For example, as shown in (**806**) (FIG. **8**), polynucleotides from the control sample and the test sample can be pooled. The pooled molecules can be subject to the sequencing (**808**) and bioinformatic work flow. Both will be subject to the same variations in the process and, therefore, any differential bias is reduced. Because molecules originating from control and test samples are differently tagged, they can be distinguished in any process of quantitative measurement.

The relative amount of control and test sample pooled can be varied. The amount of control sample can be same as the amount of test sample. The amount of control sample can also be larger than the amount of test sample. Alternatively, the amount of control sample can be smaller than the amount of test sample. The smaller the relative amount of one sample to the total, the fewer identifying tags needed in the original tagging process. A number can be selected to reduce to acceptable levels the probability that two parent molecules having the same start/end sequences will bear the same identifying tag. This probability can be less than 10%, less than 1%, less than 0.1% or less than 0.01%. The probability can be less than 25%, 24%, 23%, 22%, 21%, 20%, 19%, 18%, 17%, 16%, 15%, 14%, 13%, 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, or 1%.

Methods disclosed herein can also comprise grouping sequence reads. For example, bioinformatic workflow can include grouping sequence reads produced from progeny of a single parent molecule, as shown in (**810**) (FIG. **8**). This can involve any of the redundancy reduction methods described herein. Molecules sourced from test and control samples can be differentiated based on source tags they carry (**812**). Molecules mapping to a target locus are quantified for both test-sourced and control-sourced molecules (**812**). This can include the normalization methods discussed herein, e.g., in which numbers at a target locus are normalized against numbers at a reference locus.

Normalized (or raw) quantities at a target locus from test and control samples are compared to determine presence of copy number variation (**814**).

3. Computer Control Systems

The present disclosure provides computer control systems that are programmed to implement methods of the disclosure. FIG. **6** shows a computer system **1501** that is programmed or otherwise configured to implement the methods of the present disclosure. The computer system **1501** can regulate various aspects sample preparation, sequencing and/or analysis. In some examples, the computer system **1501** is configured to perform sample preparation and sample analysis, including nucleic acid sequencing. The computer system **1501** can be an electronic device of a user or a computer system that is remotely located with respect to the electronic device. The electronic device can be a mobile electronic device.

The computer system **1501** includes a central processing unit (CPU, also "processor" and "computer processor" herein) **1505**, which can be a single core or multi core processor, or a plurality of processors for parallel processing. The computer system **1501** also includes memory or memory location **1510** (e.g., random-access memory, read-only memory, flash memory), electronic storage unit **1515** (e.g., hard disk), communication interface **1520** (e.g., network adapter) for communicating with one or more other systems, and peripheral devices **1525**, such as cache, other memory, data storage and/or electronic display adapters. The memory **1510**, storage unit **1515**, interface **1520** and peripheral devices **1525** are in communication with the CPU **1505** through a communication bus (solid lines), such as a motherboard. The storage unit **1515** can be a data storage unit (or data repository) for storing data. The computer system **1501** can be operatively coupled to a computer network ("network") **1530** with the aid of the communication interface **1520**. The network **1530** can be the Internet, an internet and/or extranet, or an intranet and/or extranet that is in communication with the Internet. The network **1530** in some cases is a telecommunication and/or data network. The network **1530** can include one or more computer servers, which can enable distributed computing, such as cloud computing. The network **1530**, in some cases with the aid of the computer system **1501**, can implement a peer-to-peer network, which may enable devices coupled to the computer system **1501** to behave as a client or a server.

The CPU **1505** can execute a sequence of machine-readable instructions, which can be embodied in a program or software. The instructions may be stored in a memory location, such as the memory **1510**. The instructions can be directed to the CPU **1505**, which can subsequently program or otherwise configure the CPU **1505** to implement methods of the present disclosure. Examples of operations performed by the CPU **1505** can include fetch, decode, execute, and writeback.

The CPU **1505** can be part of a circuit, such as an integrated circuit. One or more other components of the system **1501** can be included in the circuit. In some cases, the circuit is an application specific integrated circuit (ASIC).

The storage unit **1515** can store files, such as drivers, libraries and saved programs. The storage unit **1515** can store user data, e.g., user preferences and user programs. The computer system **1501** in some cases can include one or more additional data storage units that are external to the computer system **1501**, such as located on a remote server that is in communication with the computer system **1501** through an intranet or the Internet.

The computer system **1501** can communicate with one or more remote computer systems through the network **1530**. For instance, the computer system **1501** can communicate with a remote computer system of a user (e.g., an operator). Examples of remote computer systems include personal computers (e.g., portable PC), slate or tablet PC's (e.g., Apple® iPad, Samsung® Galaxy Tab), telephones, Smart phones (e.g., Apple® iPhone, Android-enabled device, Blackberry®), or personal digital assistants. The user can access the computer system **1501** via the network **1530**.

Methods as described herein can be implemented by way of machine (e.g., computer processor) executable code stored on an electronic storage location of the computer system **1501**, such as, for example, on the memory **1510** or electronic storage unit **1515**. The machine executable or machine readable code can be provided in the form of software. During use, the code can be executed by the processor **1505**. In some cases, the code can be retrieved from the storage unit **1515** and stored on the memory **1510** for ready access by the processor **1505**. In some situations, the electronic storage unit **1515** can be precluded, and machine-executable instructions are stored on memory **1510**.

The code can be pre-compiled and configured for use with a machine have a processer adapted to execute the code, or can be compiled during runtime. The code can be supplied in a programming language that can be selected to enable the code to execute in a pre-compiled or as-compiled fashion.

Aspects of the systems and methods provided herein, such as the computer system **1501**, can be embodied in programming. Various aspects of the technology may be thought of as "products" or "articles of manufacture" typically in the form of machine (or processor) executable code and/or associated data that is carried on or embodied in a type of machine readable medium. Machine-executable code can be stored on an electronic storage unit, such memory (e.g., read-only memory, random-access memory, flash memory) or a hard disk. "Storage" type media can include any or all of the tangible memory of the computers, processors or the like, or associated modules thereof, such as various semiconductor memories, tape drives, disk drives and the like, which may provide non-transitory storage at any time for the software programming. All or portions of the software may at times be communicated through the Internet or various other telecommunication networks. Such communications, for example, may enable loading of the software from one computer or processor into another, for example, from a management server or host computer into the computer platform of an application server. Thus, another type of media that may bear the software elements includes optical, electrical and electromagnetic waves, such as used across physical interfaces between local devices, through wired and optical landline networks and over various air-links. The physical elements that carry such waves, such as wired or wireless links, optical links or the like, also may be considered as media bearing the software. As used herein, unless restricted to non-transitory, tangible "storage" media, terms such as computer or machine "readable medium" refer to any medium that participates in providing instructions to a processor for execution.

Hence, a machine readable medium, such as computer-executable code, may take many forms, including but not limited to, a tangible storage medium, a carrier wave medium or physical transmission medium. Non-volatile storage media include, for example, optical or magnetic disks, such as any of the storage devices in any computer(s) or the like, such as may be used to implement the databases,

etc. shown in the drawings. Volatile storage media include dynamic memory, such as main memory of such a computer platform. Tangible transmission media include coaxial cables; copper wire and fiber optics, including the wires that comprise a bus within a computer system. Carrier-wave transmission media may take the form of electric or electromagnetic signals, or acoustic or light waves such as those generated during radio frequency (RF) and infrared (IR) data communications. Common forms of computer-readable media therefore include for example: a floppy disk, a flexible disk, hard disk, magnetic tape, any other magnetic medium, a CD-ROM, DVD or DVD-ROM, any other optical medium, punch cards paper tape, any other physical storage medium with patterns of holes, a RAM, a ROM, a PROM and EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave transporting data or instructions, cables or links transporting such a carrier wave, or any other medium from which a computer may read programming code and/or data. Many of these forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to a processor for execution.

The computer system **1501** can include or be in communication with an electronic display **1535** that comprises a user interface (UI) **1540**. The UI can allow a user to set various conditions for the methods described herein, for example, PCR or sequencing conditions. Examples of UI's include, without limitation, a graphical user interface (GUI) and web-based user interface.

Methods and systems of the present disclosure can be implemented by way of one or more algorithms. An algorithm can be implemented by way of software upon execution by the central processing unit **1505**. The algorithm can, for example, process the reads to generate a consequence sequence.

FIG. **7** schematically illustrates another system for analyzing a sample comprising nucleic acids from a subject. The system includes a sequencer, bioinformatic software and internet connection for report analysis by, for example, a hand held device or a desktop computer

Disclosed herein is a system for analyzing a target nucleic acid molecule of a subject, comprising: a communication interface that receives nucleic acid sequence reads for a plurality of polynucleotide molecules that cover genomic loci of a target genome; computer memory that stores the nucleic acid sequence reads for the plurality of polynucleotide molecules received by the communication interface; and a computer processor operatively coupled to the communication interface and the memory and programmed to (i) group the plurality of sequence reads into families, wherein each family comprises sequence reads from one of the template polynucleotides, (ii) for each of the families, merge sequence reads to generate a consensus sequence, (iii) call the consensus sequence at a given genomic locus among the genomic loci, and (iv) detect at the given genomic locus any of genetic variants among the calls, frequency of a genetic alteration among the calls, total number of calls; and total number of alterations among the calls, wherein the genomic loci correspond to a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The different variations of each component of the system are described throughout the disclosure within the methods and compositions. These individual components and variations thereof, are also applicable in this system.

4. Kits

Kits comprising the compositions as described herein. The kits can be useful in performing the methods as described herein. Disclosed herein is a kit comprising a plurality of oligonucleotide probes that selectively hybridize to least 5, 6, 7, 8, 9, 10, 20, 30, 40 or all genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1. The number genes to which the oligonucleotide probes can selectively hybridize can vary. For example, the number of genes can comprise 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, or 54. The kit can include a container that includes the plurality of oligonucleotide probes and instructions for performing any of the methods described herein.

The oligonucleotide probes can selectively hybridize to exon regions of the genes, e.g., of the at least 5 genes. In some cases, the oligonucleotide probes can selectively hybridize to at least 30 exons of the genes, e.g., of the at least 5 genes. In some cases, the multiple probes can selectively hybridize to each of the at least 30 exons. The probes that hybridize to each exon can have sequences that overlap with at least 1 other probe. In some embodiments, the oligoprobes can selectively hybridize to non-coding regions of genes disclosed herein, for example, intronic regions of the genes. The oligoprobes can also selectively hybridize to regions of genes comprising both exonic and intronic regions of the genes disclosed herein.

Any number of exons can be targeted by the oligonucleotide probes. For example, at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140, 145, 150, 155, 160, 165, 170, 175, 180, 185, 190, 195, 200, 205, 210, 215, 220, 225, 230, 235, 240, 245, 250, 255, 260, 265, 270, 275, 280, 285, 290, 295, 300, 400, 500, 600, 700, 800, 900, 1,000, or more, exons can be targeted.

The kit can comprise at least 4, 5, 6, 7, or 8 different library adaptors having distinct molecular barcodes and identical sample barcodes. The library adaptors may not be sequencing adaptors. For example, the library adaptors do not include flow cell sequences or sequences that permit the formation of hairpin loops for sequencing. The different variations and combinations of molecular barcodes and sample barcodes are described throughout, and are applicable to the kit. Further, in some cases, the adaptors are not sequencing adaptors. Additionally, the adaptors provided with the kit can also comprise sequencing adaptors. A sequencing adaptor can comprise a sequence hybridizing to

one or more sequencing primers. A sequencing adaptor can further comprise a sequence hybridizing to a solid support, e.g., a flow cell sequence. For example, a sequencing adaptor can be a flow cell adaptor. The sequencing adaptors can be attached to one or both ends of a polynucleotide fragment. In some cases, the kit can comprise at least 8 different library adaptors having distinct molecular barcodes and identical sample barcodes. The library adaptors may not be sequencing adaptors. The kit can further include a sequencing adaptor having a first sequence that selectively hybridizes to the library adaptors and a second sequence that selectively hybridizes to a flow cell sequence. In another example, a sequencing adaptor can be hairpin shaped. For example, the hairpin shaped adaptor can comprise a complementary double stranded portion and a loop portion, where the double stranded portion can be attached (e.g., ligated) to a double-stranded polynucleotide. Hairpin shaped sequencing adaptors can be attached to both ends of a polynucleotide fragment to generate a circular molecule, which can be sequenced multiple times. A sequencing adaptor can be up to 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, or more bases from end to end. The sequencing adaptor can comprise 20-30, 20-40, 30-50, 30-60, 40-60, 40-70, 50-60, 50-70, bases from end to end. In a particular example, the sequencing adaptor can comprise 20-30 bases from end to end. In another example, the sequencing adaptor can comprise 50-60 bases from end to end. A sequencing adaptor can comprise one or more barcodes. For example, a sequencing adaptor can comprise a sample barcode. The sample barcode can comprise a pre-determined sequence. The sample barcodes can be used to identify the source of the polynucleotides. The sample barcode can be at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, or more (or any length as described throughout) nucleic acid bases, e.g., at least 8 bases. The barcode can be contiguous or non-contiguous sequences, as described above.

The library adaptors can be blunt ended and Y-shaped and can be less than or equal to 40 nucleic acid bases in length. Other variations of the can be found throughout and are applicable to the kit.

EXAMPLES

Example 1

Methods for Copy Number Variation Detection

Blood Collection

10-30 mL Blood samples are collected at room temperature. The samples are centrifuged to remove cells. Plasma is collected after centrifugation.

cfDNA Extraction

The sample is subjected to proteinase K digestion. DNA is precipitated with isopropanol. DNA is captured on a DNA purification column (e.g., a QIAamp DNA Blood Mini Kit) and eluted in 100 μl solution. DNAs below 500 bp are selected with Ampure SPRI magnetic bead capture (PEG/salt). The resulting production is suspended in 30 μl H_2O. Size distribution is checked (major peak=166 nucleotides; minor peak=330 nucleotides) and quantified. 5 ng of extracted DNA contain approximately 1700 haploid genome equivalents ("HGE"). The general correlation between the amount of DNA and HGE is as follow: 3 pg DNA=1 HGE; 3 ng DNA=1K HGE; 3 μg DNA=1M HGE; 10 pg DNA=3 HE; 10 ng DNA=3K HGE; 10 μg DNA=3M HGE.

"Single Molecule" Library Prep

High-efficiency DNA tagging (>80%) is performed by blunt-end repair and ligation with 8 different octomers (i.e., 64 combinations) with overloaded hairpin adaptors. 2.5 ng DNA (i.e. approximately 800 HGE) is used as the starting material. Each hairpin adaptor comprises a random sequence on its non-complementary portion. Both ends of each DNA fragment are attached with hairpin adaptors. Each tagged fragment can be identified by the random sequence on the hairpin adaptors and a 10 p endogenous sequence on the fragment.

Tagged DNA is amplified by 10 cycles of PCR to produce about 1-7 μg DNAs that contain approximately 500 copies of each of the 800 HGE in the starting material.

Buffer optimization, polymerase optimization and cycle reduction may be performed to optimize the PCR reactions. Amplification bias, e.g., non-specific bias, GC bias, and/or size bias are also reduced by optimization. Noise(s) (e.g., polymerase-introduced errors) are reduced by using high-fidelity polymerases.

The Library may be prepared using Verniata or Sequenom methods.

Sequences may be enriched as follow: DNAs with regions of interest (ROI) are captured using biotin-labeled bead with probe to ROIs. The ROIs are amplified with 12 cycles of PCR to generate a 2000 times amplification. The resulting DNA is then denatured and diluted to 8 pM and loaded into an Illumina sequencer.

Massively Parallel Sequencing

0.1 to 1% of the sample (approximately 100 pg) are used for sequencing.

Digital Bioinformatics

Sequence reads are grouped into families, with about 10 sequence reads in each family. Families are collapsed into consensus sequences by voting (e.g., biased voting) each position in a family. A base is called for consensus sequence if 8 or 9 members agree. A base is not called for consensus sequence if no more than 60% of the members agree.

The resulting consensus sequences are mapped to a reference genome. Each base in a consensus sequence is covered by about 3000 different families. A quality score for each sequence is calculated and sequences are filtered based on the their quality scores.

Sequence variation is detected by counting distribution of bases at each locus. If 98% of the reads have the same base (homozygous) and 2% have a different base, the locus is likely to have a sequence variant, presumably from cancer DNA.

CNV is detected by counting the total number of sequences (bases) mapping to a locus and comparing with a control locus. To increase CNV detection, CNV analysis is performed specific regions, including regions on ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, or NTRK1 genes.

Example 2

Method for Correcting Base Calling by Determining the Total Number Unseen Molecules in a Sample

After fragments are amplified and the sequences of amplified fragments are read and aligned, the fragments are subjected to base calling. Variations in the number of amplified fragments and unseen amplified fragments can introduce errors in base calling. These variations are corrected by calculating the number of unseen amplified fragments.

When base calling for locus A (an arbitrary locus), it is first assumed that there are N amplified fragments. The sequence readouts can come from two types of fragments: double-strand fragments and single-strand fragments. The following is a theoretical example of calculating the total number of unseen molecules in a sample.

N is the total number of molecules in the sample.

Assuming 1000 is the number of duplexes detected.

Assuming 500 is the number of single-stranded molecule detected.

P is the probability of seeing a strand.

Q is the probability of not detecting a strand.

Since $Q=1-P$.

$1000=NP(2)$.

$500=N2PQ$.

$1000/P(2)=N$.

$500\div 2\ PQ=N$.

$1000/P(2)=500\div 2PQ$.

$1000*2\ PQ=500\ P(2)$.

$2000\ PQ=500\ P(2)$.

$2000\ Q=500\ P$.

$2000\ (1-P)=500P$

$2000-2000\ P=500P$.

$2000=500P+2000\ P$.

$2000=2500\ P$.

$2000\div 2500=P$.

$0.8=P$.

$1000/P(2)=N$.

$1000\div 0.64=N$.

$1562=N$.

Number of unseen fragments=62.

Example 3

Identification of Genetic Variants in Cancer-Associated Somatic Variants in a Patient

An assay is used to analyze a panel of genes to identify genetic variants in cancer-associated somatic variants with high sensitivity.

Cell-free DNA is extracted from plasma of a patient and amplified by PCR. Genetic variants are analyzed by massively parallel sequencing of the amplified target genes. For one set of genes, all exons are sequenced as such sequencing coverage had shown to have clinically utility (Table 1). For another set of genes, sequencing coverage included those exons with a previously reported somatic mutation (Table 2). The minimum detectable mutant allele (limit of detection) is dependent on the patient's sample cell-free DNA concentration, which varied from less than 10 to over 1,000 genomic equivalents per mL of peripheral blood. Amplification may not be detected in samples with lower amounts of cell-free DNA and/or low-level gene copy amplification. Certain sample or variant characteristics resulted in reduced analytic sensitivity, such as low sample quality or improper collection.

The percentage of genetic variants found in cell-free DNA circulating in blood is related to the unique tumor biology of this patient. Factors that affected the amount/percentages of detected genetic variants in circulating cell-free DNA in blood include tumor growth, turn-over, size, heterogeneity, vascularization, disease progression or treatment. Table 3 annotates the percentage, or allele frequency, of altered circulating cell-free DNA (% cfDNA) detected in this patient. Some of the detected genetic variants are listed in descending order by % cfDNA.

Genetic variants are detected in the circulating cell-free DNA isolated from this patient's blood specimen. These genetic variants are cancer-associated somatic variants, some of which have been associated with either increased or reduced clinical response to specific treatment. "Minor Alterations" are defined as those alterations detected at less than 10% the allele frequency of "Major Alterations". The detected allele frequencies of these alterations (Table 3) and associated treatments for this patient are annotated.

All genes listed in Tables 1 and 2 are analyzed as part of the Guardant360™ test. Amplification is not detected for ERBB2, EGFR, or MET in the circulating cell-free DNA isolated from this patient's blood specimen.

Patient test results comprising the genetic variants are listed in Table 4.

TABLE 1

Genes in which all exons are sequenced

GENES IN WHICH ALL EXONS ARE SEQUENCED			
ALK	<0.1%	APC	<0.1%
AR	<0.1%	BRAF	<0.1%
CDKN2A	<0.1%	EGFR	<0.1%
ERBB2	<0.1%	FBXW7	<0.1%
KRAS	<0.1%	MET	<0.1%
MYC	<0.1%	NOTCH1	<0.1%
NRAS	<0.1%	PIK3CA	<0.1%
PTEN	<0.1%	PROC	<0.1%
RB1	<0.1%	TP53	<0.1%

LOD: Limit of Detection. The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected.

TABLE 2

Genes in which exons with a previously reported somatic mutation are sequenced
GENES IN WHICH EXONS WITH A PREVIOUSLY REPORTED SOMATIC MUTATION ARE SEQUENCED

ABL1	<0.1%	AKT1	<0.1%
ATM	<0.1%	CDH1	<0.1%
CSF1R	<0.1%	CTNNB1	<0.1%
ERBB4	<0.1%	EZH2	<0.1%
FGFR1	<0.1%	FGFR2	<0.1%
FGFR3	<0.1%	FLT3	<0.1%
GNA11	<0.1%	GNAQ	<0.1%
GNAS	<0.1%	HNFlA	<0.1%
HRAS	<0.1%	IDH1	<0.1%
IDH2	<0.1%	JAK2	<0.1%
JAK3	<0.1%	KDR	<0.1%
KIT	<0.1%	MLH1	<0.1%
MPL	<0.1%	NPM1	<0.1%
PDGFRA	<0.1%	PTPN11	<0.1%
RET	<0.1%	SMAD4	<0.1%
SMARCB1	<0.1%	SMO	<0.1%

TABLE 2-continued

Genes in which exons with a previously reported somatic mutation are sequenced
GENES IN WHICH EXONS WITH A PREVIOUSLY REPORTED SOMATIC MUTATION ARE SEQUENCED

SRC	<0.1%	STK11	<0.1%
TERT	<0.1%	VHL	<0.1%

LOD: Limit of Detection. The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected.

TABLE 3

Allele frequency of altered circulating cell-free DNA detected in this patient

Gene	cfDNA with alterations (%)	cfDNA without alterations (%)
BRAF V600E	8.9	91.1
NRAS Q61K	6.2	93.8
JAK V617F	1.5	98.6

TABLE 4

Genomic alterations detected in selected genes
Detected: 51 Genomic Alterations

Gene	Chromosome	Position	Mutation (nt)	Mutation (AA)	Percentage	Cosmic ID	DBSNP ID
KRAS	12	25368462	C > T		100.0%		rs4362222
ALK	2	29416572	T > C	I1461V	100.0%		rs1670283
ALK	2	29444095	C > T		100.0%		rs1569156
ALK	2	29543663	T > C	Q500Q	100.0%		rs2293564
ALK	2	29940529	A > T	P234P	100.0%		rs2246745
APC	5	112176756	T > A	V1822D	100.0%		rs459552
CDKN2A	9	21968199	C > G		100.0%	COSM14251	rs11515
FGFR3	4	1807894	G > A	T651T	100.0%		rs7688609
NOTCH1	9	139410424	A > G		100.0%		rs3125006
PDGFRA	4	55141055	A > G	P567P	100.0%		rs1873778
HRAS	11	534242	A > G	H27H	100.0%	COSM249860	rs12628
EGFR	7	55214348	C > T	N158N	99.9%	COSM42978	rs2072454
TP53	17	7579472	G > C	P72R	99.8%		rs1042522
APC	5	112162854	T > C	Y486Y	55.0%		rs2229992
APC	5	112177171	G > A	P1960P	53.8%		rs465899
EGFR	7	55266417	T > C	T903T	53.6%		rs1140475
APC	5	112176325	G > A	G1678G	53.2%		rs42427
APC	5	112176559	T > G	S1756S	53.0%		rs866006
EGFR	7	55229255	G > A	R521K	53.0%		
MET	7	116397572	A > G	Q648Q	52.7%		
APC	5	112175770	G > A	T1493T	52.7%		rs41115
EGFR	7	55249063	G > A	Q787Q	52.6%		rs1050171
NOTCH1	9	139411714	T > C		52.4%		rs11145767
EGFR	7	55238874	T > A	T629T	52.0%		rs2227984
ERBB2	17	37879588	A > G	I655V	51.6%		rs1136201
NOTCH1	9	139397707	G > A	D1698D	51.3%	COSM33747	rs10521
ALK	2	30143499	G > C	L9L	51.0%		rs4358080
APC	5	112164561	G > A	A545A	51.0%		rs351771
FLT3	13	28610183	A > G		50.8%		rs2491231
NOTCH1	9	139418260	A > G	N104N	50.5%		rs4489420
ALK	2	29444076	G > T		50.4%		rs1534545
PIK3CA	3	178917005	A > G		50.3%		rs3729674
NOTCH1	9	139412197	G > A		50.2%		rs9411208
ALK	2	29455267	A > G	G845G	50.0%	COSM148825	rs2256740
KIT	4	55593464	A > C	M541L	49.9%	COSM28026	
NOTCH1	9	139391636	G > A	D2185D	48.9%		rs2229974
PDGFRA	4	55152040	C > T	V824V	48.9%	COSM22413	rs2228230
ALK	2	29416481	T > C	K1491R	48.9%	COSM1130802	rs1881420
ALK	2	29445458	G > T	G1125G	48.6%		rs3795850
NOTCH1	9	139410177	T > C		48.5%		rs3124603
RET	10	43613843	G > T	L769L	48.2%		rs1800861
EGFR	7	55214443	G > A		48.0%		rs7801956
ALK	2	29416366	G > C	D1529E	47.2%		rs1881421

TABLE 4-continued

Genomic alterations detected in selected genes Detected: 51 Genomic Alterations							
Gene	Chromosome	Position	Mutation (nt)	Mutation (AA)	Percentage	Cosmic ID	DBSNP ID
EGFR	7	55238087	C > T		45.5%		rs10258429
RET	10	43615633	C > G	S904S	44.8%		rs1800863
BRAF	7	140453136	A > T	V600E	8.9%	COSM476	
NRAS	1	115256530	G > T	Q61K	6.2%	COSM580	rs121913254
JAK2	9	5073770	G > T	V617F	1.5%	COSM12600	rs77375493

Example 4

Determining Patient-Specific Limits of Detection for Genes Analyzed by Guardant360™ Assays

Using the method of Example 3, Genetic alterations in cell-free DNA of a patient are detected. The sequence reads of these genes include exon and/or intron sequences.

Limits of detection of the test are shown in Table 5. The limits of detection values are dependent on cell-free DNA concentration and sequencing coverage for each gene.

TABLE 5

Limits of Detection of selected genes in a patient using Guardant					
Complete Exon and Partial Intron Coverage					
APC	0.1%	AR *	0.2%	ARID1A	
BRAF *	0.1%	BRCA1		BRCA2	
CCND1 *		CCND2 *		CCNE1 *	
CDK4 *		CDK6 *		CDKN2A	0.1%
CDKN2B		EGFR *	<0.1%	ERBB2 *	0.1%
FGFR1 *	<0.1%	FGFR2 *	0.1%	HRAS	0.1%
KIT *	0.1%	KRAS *	0.1%	MET *	0.1%
MYC *	0.1%	NF1		NRAS	0.1%
PDGFRA *	0.1%	PIK3CA *	0.1%	PTEN	0.1%
RAF1 *		TP53	0.1%		
Exons Covered with Reported Somatic Mutations					
AKT1	0.1%	ALK	<0.1%	ARAF	
ATM	0.1%	CDH1	0.1%	CTNNB1	0.1%
ESR1		EZH2	0.1%	FBXW7	0.1%
FGFR3	0.1%	GATA3		GNA11	0.1%
GNAQ	0.1%	GNAS	0.1%	HNF1A	0.1%
IDH1	0.1%	IDH2	0.1%	JAK2	0.1%
JAK3	0.1%	MAP2K1		MAP2K2	
MLH1	0.1%	MPL	0.2%	NFE2L2	
NOTCH1	0.1%	NPM1	0.1%	PTPN11	0.1%
RET	0.1%	RHEB		RHOA	
RIT1		ROS1		SMAD4	0.1%
SMO	0.1%	SRC	<0.1%	STK11	0.2%
TERT	0.1%	VHL	0.2%		
Fusions					
ALK	<0.1%	RET	0.1%	ROS1	
NTRK1					

LOD: Limit of Detection. The minimum detectable mutant allele frequency for this specimen in which 80% of somatic variants is detected.
* indicates CNV genes.

Example 5

Correcting Sequence Errors Comparing Watson and Crick Sequences

Double-stranded cell-free DNA is isolated from the plasma of a patient. The cell-free DNA fragments are tagged using 16 different bubble-containing adaptors, each of which comprises a distinctive barcode. The bubble-containing adaptors are attached to both ends of each cell-free DNA fragment by ligation. After ligation, each of the cell-free DNA fragment can be distinctly identified by the sequence of the distinct barcodes and two 20 bp endogenous sequences at each end of the cell-free DNA fragment.

The tagged cell-free DNA fragments are amplified by PCR. The amplified fragments are enriched using beads comprising oligonucleotide probes that specifically bind to a group of cancer-associated genes. Therefore, cell-free DNA fragments from the group of cancer-associated genes are selectively enriched.

Sequencing adaptors, each of which comprises a sequencing primer binding site, a sample barcode, and a cell-flow sequence, are attached to the enriched DNA molecules. The resulting molecules are amplified by PCR.

Both strands of the amplified fragments are sequenced. Because each bubble-containing adaptor comprises a non-complementary portion (e.g., the bubble), the sequence of the one strand of the bubble-containing adaptor is different from the sequence of the other strand (complement). Therefore, the sequence reads of amplicons derived from the Watson strand of an original cell-free DNA can be distinguished from amplicons from the Crick strand of the original cell-free DNA by the attached bubble-containing adaptor sequences.

The sequence reads from a strand of an original cell-free DNA fragment are compared to the sequence reads from the other strand of the original cell-free DNA fragment. If a variant occurs in only the sequence reads from one strand, but not other strand, of the original cell-free DNA fragment, this variant will be identified as an error (e.g., resulted from PCR and/or amplification), rather than a true genetic variant.

The sequence reads are grouped into families. Errors in the sequence reads are corrected. The consensus sequence of each family is generated by collapsing.

While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. It is not intended that the invention be limited by the specific examples provided within the specification. While the invention has been described with reference to the aforementioned specification, the descriptions and illustrations of the embodiments herein are not meant to be construed in a limiting sense. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. Furthermore, it shall be understood that all aspects of the invention are not limited to the specific depictions, configurations or relative proportions set forth herein which depend upon a variety of conditions and variables. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is therefore contemplated that the invention shall also cover any such alternatives, modifica-

tions, variations or equivalents. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

What is claimed is:

1. A method, comprising:

(a) providing a population of cell-free deoxyribonucleic acid (cfDNA) molecules having first and second complementary strands;

(b) tagging a plurality of the cfDNA molecules in the population with duplex tags comprising molecular barcodes to produce tagged parent polynucleotides, wherein the duplex tags are attached to both ends of a molecule of the plurality of the cfDNA molecules,

wherein the plurality of the cfDNA molecules are tagged with n different combinations of molecular barcodes, wherein n is at least 2 and no more than 100,000*z, wherein z is a mean of an expected number of duplicate molecules in the population of cfDNA molecules that map to identical start and stop positions on a reference sequence;

(c) amplifying a plurality of the tagged parent polynucleotides to produce amplified progeny polynucleotides;

(d) sequencing at least a subset of the amplified progeny polynucleotides to produce a set of sequence reads; and

(e) reducing or tracking redundancy of a plurality of sequence reads from the set of sequence reads using at least sequencing information from the molecular barcodes of the duplex tags to determine distinct cfDNA molecules from among the tagged parent polynucleotides, wherein the distinct cfDNA molecules are determined based on (i) paired reads corresponding to sequence reads generated from a first tagged strand and a second tagged complementary strand derived from cfDNA molecules from among the tagged parent polynucleotides, or (ii) unpaired reads corresponding to sequence reads generated from a first tagged strand having no second tagged complementary strand derived from cfDNA molecules from among the tagged parent polynucleotides, wherein reducing or tracking the redundancy of the plurality of sequence reads comprises mapping at least a subset of the plurality of sequence reads to the reference sequence.

2. The method of claim **1**, wherein the population of cfDNA molecules is obtained or derived from a subject having cancer.

3. The method of claim **1**, wherein the population of cfDNA molecules comprises 1 nanogram (ng) to 100 ng of cfDNA molecules.

4. The method of claim **1**, wherein the molecular barcodes are ligated to the plurality of the cfDNA molecules using more than a 10× excess of duplex tags as compared to the population of cfDNA molecules, wherein at least 20% of the cfDNA molecules from the population are tagged with the duplex tags.

5. The method of claim **1**, wherein the molecular barcodes are ligated to the plurality of the cfDNA molecules using more than an 80× excess of duplex tags as compared to the population of cfDNA molecules.

6. The method of claim **5**, wherein at least 40% of the cfDNA molecules from the population are tagged with the duplex tags.

7. The method of claim **1**, wherein z is between 2 and 8.

8. The method of claim **1**, wherein the molecular barcodes have a length of 5 to 20 base pairs.

9. The method of claim **1**, further comprising, prior to the sequencing, enriching at least a subset of the amplified progeny polynucleotides for target regions of interest to produce enriched progeny polynucleotides.

10. The method of claim **9**, wherein the target regions of interest comprise genetic sequences of a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF 1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

11. The method of claim **10**, further comprising, prior to the sequencing, amplifying a plurality of the enriched progeny polynucleotides.

12. The method of claim **1**, wherein the duplex tags are part of sequencing adapters.

13. The method of claim **1**, wherein reducing or tracking the redundancy of the plurality of sequence reads comprises grouping the paired reads or the unpaired reads into families based at least in part on (i) the molecular barcodes associated with the paired reads or the unpaired reads and (ii) sequence information at the start and stop positions of the paired reads or the unpaired reads.

14. The method of claim **1**, further comprising:

(f) determining quantitative measures of (i) paired reads or (ii) unpaired reads that map to a genomic locus of the reference sequence.

15. The method of claim **14**, further comprising:

(g) estimating with a programmed computer processor a quantitative measure of tagged parent polynucleotides that map to the genomic locus based on the quantitative measures of the paired reads and the unpaired reads.

16. The method of claim **15**, further comprising:

(h) detecting copy number variation in the population of cfDNA molecules by determining a normalized quantitative measure determined in (g) at each of one or more genomic loci.

17. A method, comprising:

(a) tagging a population of double-stranded cell-free deoxyribonucleic acid (cfDNA) molecules obtained or derived from a sample of a subject with a set of tags comprising molecular barcodes to produce tagged parent polynucleotides;

(b) amplifying a plurality of the tagged parent polynucleotides to produce amplified progeny polynucleotides;

(c) sequencing at least a subset of the amplified progeny polynucleotides to produce a set of sequence reads; and

(d) sorting a plurality of sequence reads from the set of sequence reads into (i) families comprising paired reads corresponding to sequence reads generated from a first tagged strand and a second tagged complementary strand derived from double-stranded cfDNA molecules from among the tagged parent polynucleotides, and (ii) families comprising unpaired reads corresponding to sequence reads generated from a first tagged strand having no second tagged complementary strand derived from double-stranded cfDNA molecules from among the tagged parent polynucleotides.

18. The method of claim **17**, wherein the sample is blood, plasma, or serum.

19. The method of claim **17**, wherein the population of double-stranded cfDNA molecules comprises 1 nanogram (ng) to 100 ng of double-stranded cfDNA molecules.

20. The method of claim **17**, wherein the tagging comprises ligating the molecular barcodes to double-stranded cfDNA molecules.

21. The method of claim **17**, wherein the set of tags comprises 2 to 10,000 different molecular barcode sequences.

22. The method of claim **17**, further comprising, prior to the sequencing, enriching at least a subset of the amplified progeny polynucleotides for target regions of interest to produce enriched progeny polynucleotides.

23. The method of claim **22**, wherein the target regions of interest comprise genetic sequences of a plurality of genes selected from the group consisting of ALK, APC, BRAF, CDKN2A, EGFR, ERBB2, FBXW7, KRAS, MYC, NOTCH1, NRAS, PIK3CA, PTEN, RB1, TP53, MET, AR, ABL1, AKT1, ATM, CDH1, CSF1R, CTNNB1, ERBB4, EZH2, FGFR1, FGFR2, FGFR3, FLT3, GNA11, GNAQ, GNAS, HNF1A, HRAS, IDH1, IDH2, JAK2, JAK3, KDR, KIT, MLH1, MPL, NPM1, PDGFRA, PROC, PTPN11, RET, SMAD4, SMARCB1, SMO, SRC, STK11, VHL, TERT, CCND1, CDK4, CDKN2B, RAF1, BRCA1, CCND2, CDK6, NF1, TP53, ARID1A, BRCA2, CCNE1, ESR1, RIT1, GATA3, MAP2K1, RHEB, ROS1, ARAF, MAP2K2, NFE2L2, RHOA, and NTRK1.

24. The method of claim **17**, wherein (d) further comprises reducing or tracking redundancy of a plurality of sequence reads from the set of sequencing reads, wherein the reducing or tracking comprises mapping at least a subset of the plurality of sequence reads to a reference sequence, and the reducing or tracking is based on (i) the molecular barcodes associated the paired reads or the unpaired reads and (ii) sequence information at the start and stop positions of the paired reads or the unpaired reads.

25. The method of claim **24**, wherein reducing or tracking the redundancy of the plurality of sequence reads comprises determining a base call at one or more genomic loci for a plurality of sequence reads that map to the one or more genomic loci on the reference sequence.

26. The method of claim **24**, wherein reducing or tracking the redundancy of the plurality of sequence reads comprises collapsing the plurality of sequence reads to produce consensus sequences representative of a sequence of the original double-stranded cfDNA molecules from among the tagged parent polynucleotides.

27. The method of claim **24**, further comprising determining quantitative measures of at least two of (i) paired reads that map to one or more genomic loci, (ii) unpaired reads that map to one or more genomic loci, (iii) a read depth of the paired reads, and (iv) a read depth of the unpaired reads.

28. The method of claim **27**, further comprising estimating with a programmed computer processor a quantitative measure of tagged parent polynucleotides that map to a genomic locus of the reference sequence based on the quantitative measures of at least two of (i) paired reads, (ii) unpaired reads, (iii) a read depth of the paired reads, and (iv) a read depth of the unpaired reads at the genomic locus.

29. The method of claim **1**, wherein the distinct cfDNA molecules in (e) are determined based on (i) the paired reads and (ii) the unpaired reads.

* * * * *

EXHIBIT E

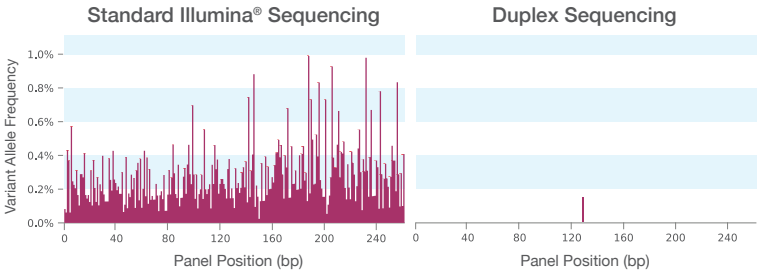
TwinStrand Duplex Sequencing™

TWINSTRAND BIOSCIENCES

Duplex Sequencing is an ultra-high accuracy sequencing method that overcomes the limitations of Next-Generation Sequencing by independently tracking both strands of individual DNA molecules. The paired sequences are compared to eliminate errors.

Duplex Sequencing Reveals True Mutations

- > 10,000-fold increase in accuracy over standard NGS
- Eliminates virtually all background errors, revealing ultra-low frequency variants
- Compatible with any Illumina® sequencer
- Easy to adopt with no special equipment needed

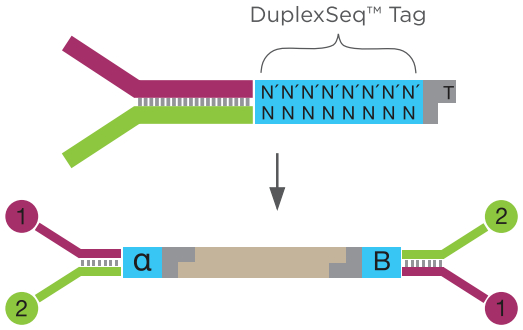


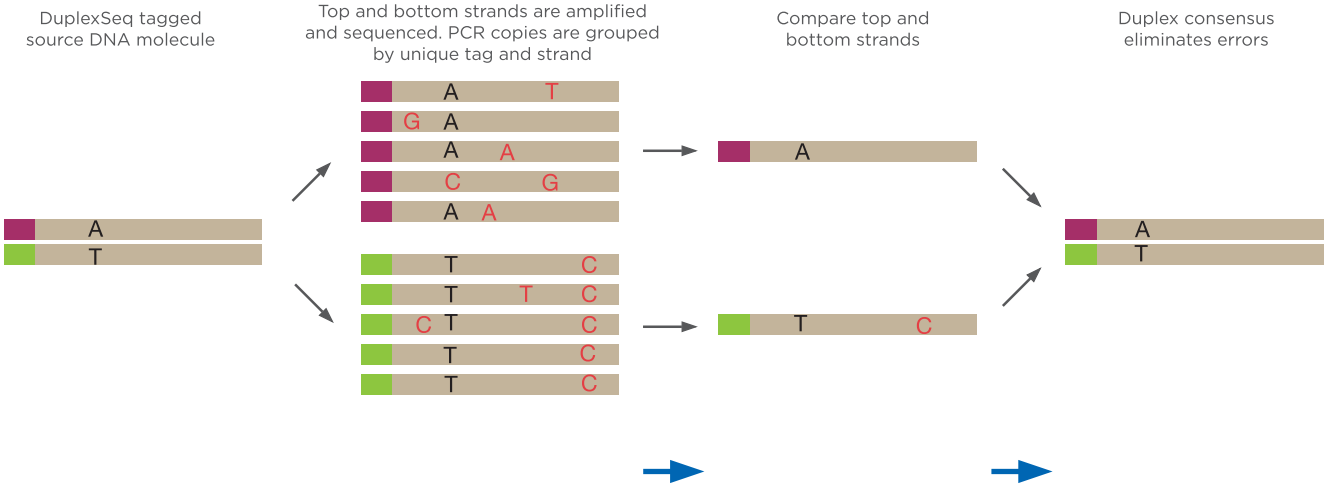
The same gene sequenced by standard Illumina® Sequencing (left) and with Duplex Sequencing (right). With conventional methods, every position in the gene appears mutated in 0.1-1% of molecules sequenced. Duplex Sequencing eliminates the background noise, revealing the previously hidden true mutation.

Duplex Sequencing Eliminates Technical Noise

Duplex Sequencing capitalizes on the naturally-occurring complementarity between strands of the DNA double-helix

Each strand of each original molecule is uniquely labelled such that both can be tracked throughout amplification and sequencing for subsequent error correction





TwinStrand Duplex Sequencing™

TWINSTRAND BIOSCIENCES

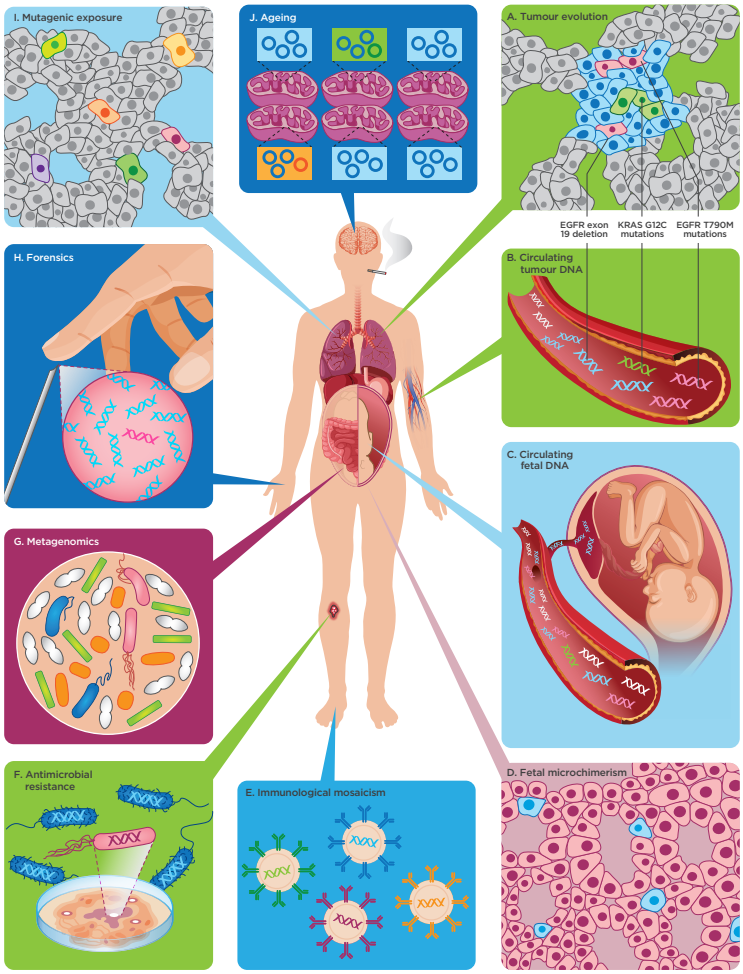
Duplex Sequencing Solves Genetic Needle-In-A-Haystack Challenges

Genetic Toxicity
Occupational exposures, cancer risk, biological age, drug safety, gene editing safety, CRISPR

Forensics
Crime scenes, assault, Disasters, ancient DNA, archaeology

Microbiota
Human health, environmental monitoring, bioreactors

Antimicrobial
Emergence of antibiotic & antiviral drug resistance, drug selection



Tumor Evolution
Emergence of drug resistance, drug selection

Liquid Biopsies
Early detection of cancer, residual disease monitoring

Organ & Cell Transplant
Chimerism, engraftment, rejection, genetic mosaicism

Prenatal Diagnostics
Fetal genomics

Immunological Health
Fetal microchimerism, autoimmunity, adaptive immunity

Duplex Sequencing is a fundamentally enabling platform technology for rare variant detection with many and diverse applications (above). A need for ultra-high accuracy in DNA sequencing is the common thread tying these applications together. TwinStrand has purpose-built solutions for mutagenesis and AML MRD available now with more applications currently in development.

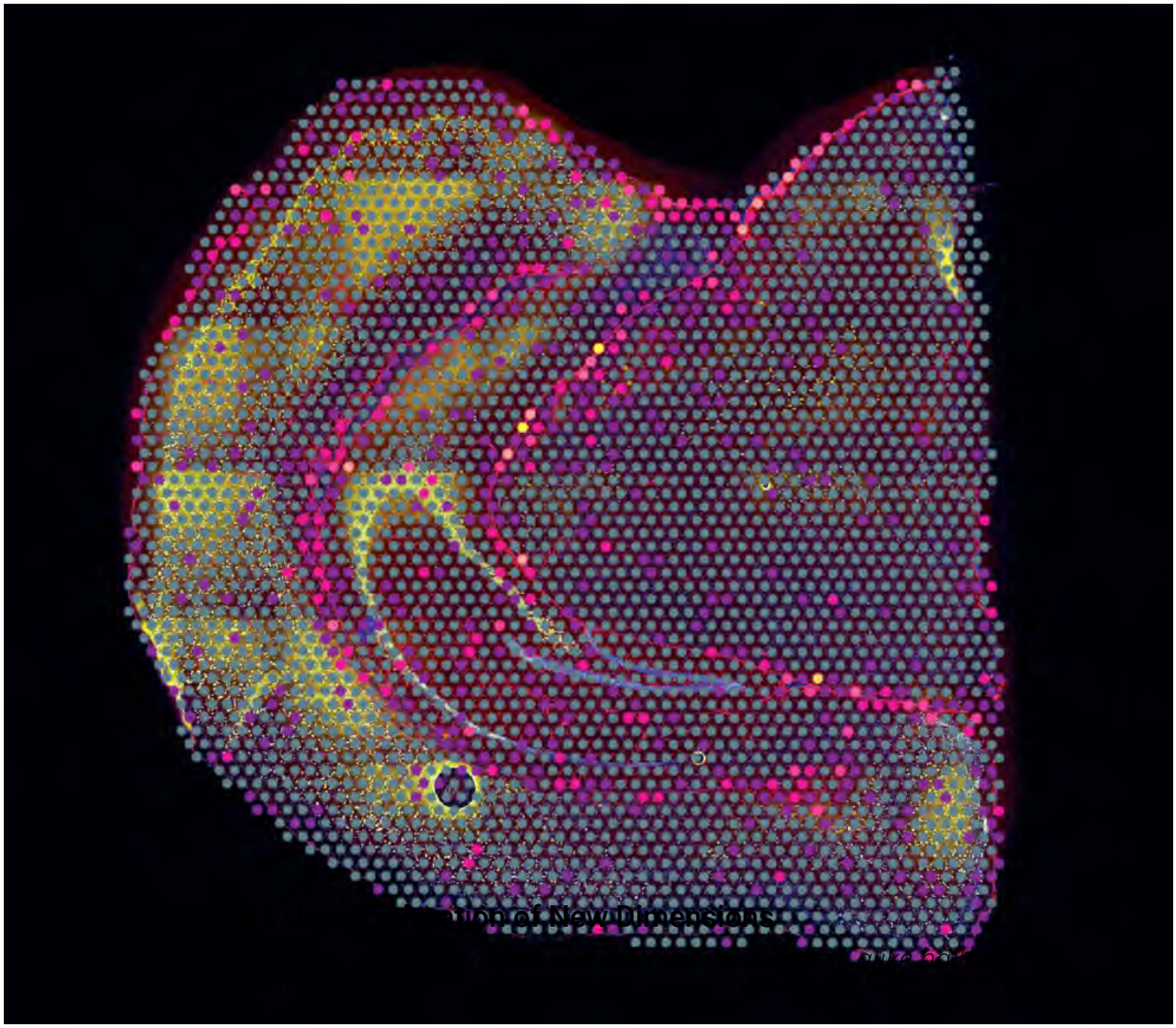
MKT0007 v1.0

For Research Use Only. Not for use in diagnostic procedures.

© 2020 TwinStrand Biosciences, Inc. All rights reserved. All trademarks are the property of TwinStrand Biosciences, Inc. or their respective owners.

To learn more visit **twinstrandbio.com**

EXHIBIT F



(https://www.genengnews.com/wp-content/uploads/2020/03/GenomeMap_AGBT_10xGenomics-scaled-e1585670045996.jpg)

By Julianna LeMieux, PhD (https://www.genengnews.com/author/juliannalemieux/) April | 2020

At Left. Combining whole transcriptome analysis with morphological context, spatial transcriptomics allows researchers to uncover the "where" for every "what," according to 10X Genomics. Here, a section of mouse brain, visualized using the company's spatial transcriptomics platform, the Visiu Gene Expression Solution, shows immunofluorescence staining of glial cells (GFAP: magenta). Overlaid spots represent GFAP mRNA selected from whole transcriptome data from the Visium Solution. [10x Genomics]

A casual observer might have failed to see past the merriment—the sun and sand, the food and drink, the singing and dancing. But make no mistake, despite occurring in a resort atmosphere, the most recent Advances in Genome Biology and Technology (AGBT) meeting was mostly business.

The meeting, which took place last February in Marco Island, FL, marked a significant anniversary (officially the 20th, but technically the 21st). The robust conference program featured presentations by the pioneers of genomics, luminaries such as Richard Wilson, PhD, Francis Collins, MD, PhD, Barbara Wold, PhD, Leroy Hood, MD, PhD, George Church, PhD, Rick Myers, PhD, and Marco Marra, PhD, while showcasing the current work of up-and- coming researchers.



Barbara Wold, PhD, professor at the California Institute of Technology and a pioneer in the field of genomics, has led many development efforts, including the one that produced the large-scale chromatin immunoprecipitation assay known as ChIPSeq. [AGBT]

The sharing of insights was lively and intense. One of the genomics pioneers, Wold, emphasized another kind of sharing, the sharing of data. A culture of data sharing, she insisted, represents a hard-won achievement, and it contributes, she continued, to genomic science as much as inspiration, grit, and funding.

Focusing largely on technology, AGBT meetings have often served as the place to introduce big developments. For example, over the years, AGBT meetings have highlighted the launches of next-generation sequencing stars such as Complete Genomics, Pacific Biosciences, Ion Torrent, and Oxford Nanopore. This year's meeting did not disappoint. From the much anticipated "$100 genome" to the launch of multiple spatial transcriptomics platforms, conversation from the pools to the beach cabanas was buzzing with news of a field that shows no signs of slowing down.



Francis S. Collins, MD, PhD, director of the National Institutes of Health, is a physician-geneticist noted for his landmark discoveries of disease genes and his leadership of the international Human Genome Project. At the annual Advances in Genome Biology and Technology meeting, his plenary session included this quote: "Two decades of breathtaking progress in genomics—but the best is yet to come." [AGBT]

The AGBT meeting, then, was mainly about the work of genomic science's movers and shakers, notwithstanding the spectacle, one evening, of scientists moving and shaking to an oceanfront performance of "Love Shack" by the B-52s. Given the venue, the song's refrain, the "bang bang bang on the door," seemed less about the urge to party, and more about the determination to propel genomic science forward.

Exploring the transcriptome-space continuum

At this year's AGBT meeting, the biggest stir was caused by spatial transcriptomics, a technique to visualize gene expression in the context of a tissue. The technique allows researchers to perform transcriptomics while retaining the positional information for the cells. That is, the cells are allowed to occupy intact tissue sections instead of being obliged to contribute to homogenized biopsies.

Spatial transcriptomics is "at the nexus of pathology, molecular biology, computational biology, cell biology, and genomics" noted Alex Swarbrick, PhD, principal research fellow at the Garvan Institute in Sydney, Australia. He told *GEN* that spatial transcriptomics could be used to answer broad research questions as well as "directed" translational questions. Regardless of the application, Swarbrick asserted, spatial transcriptomics could almost always "bring a unique perspective."

Last year, 10x Genomics took a deep dive into this field when it acquired Spatial Transcriptomics, a Swedish company that developed the technology that supports 10x Genomics' spatial transcriptomics platform, the Visium Spatial Gene Expression Solution. Two other spatial transcriptomics platforms were announced at the AGBT meeting. The first, from NanoString, is the GeoMX digital spatial profiler, and the second, from Readcoor, is the RC2 multi-omic spatial solution.

The launch of 10x Genomics' Chromium, about four years ago, expanded the utility of single-cell genomics by moving the technology into the hands of researchers. The company is looking to the Visium to do the same with spatial transcriptomics. Visium is basically a slide and reagent kit that is

readily adoptable within existing lab infrastructure.



Kit Fuhrman, PhD, the GeoMx Digital Spatial Profiler product manager at Nanostring, delivers a presentation at the company's suite at the annual meeting of Advances in Genome Biology and Technology. Here, he explains the nuts and bolts of the GeoMX platform. From dawn to dusk, the Nanostring team answered scientists' questions and explained how researchers are currently using the platform in their research. [AGBT]

By contrast, Nanostring's GeoMX is big, robust, and powerful—and expensive (roughly $300k). It provides an incredibly dense and complex array of information with an easy-to-use interface.

Both protein and RNA expression can be obtained on slides that present fresh tissue or formalin-fixed, paraffin-embedded (FFPE) tissue. The researcher can evaluate two distinct areas, for example, comparing the tissue of a tumor with the stroma or microenvironment surrounding it. Alternatively, the researcher can pinpoint an area, enclosing just a few cells within a circle.

The antibodies that bind the tissue have photo-cleavable oligonucleotide tags (or barcodes) that are attached by a UV light–sensitive linker. When the UV light hits the sample, the tags released from the bound antibodies, and the oligos are moved onto a microtiter plate where they can be analyzed by sequencing. The barcode design with sequencing readout "will open up the imagination of every scientist," assert Brad Gray, president and CEO of Nanostring.

Readcoor was founded by Richard Terry, who is now the company's CEO and CTO. Previously, he had spent several years in George Church's lab at Harvard Medical School. Then he developed Readcoor's technology while working at the Wyss Institute.

Although Readcoor's launch was signposted all over the conference center, information about the company's RC2 platform was hard to find. Terry told *GEN* that Readcoor is "different than other companies" and has "a new approach." The RC2 is powered by proprietary FISSEQ (fluorescent in situ sequencing) technology, but other details have yet to be revealed. However, the company has divulged that the RC2 has been designed to provide higher resolution than other spatial transcriptomics platforms—down to a single cell.

The proliferation of platforms, which is bound to continue, suggests that spatial transcriptomics has arrived. Spatial analysis of tissue is "already revealing new biology across many research disciplines like cancer biology, neuroscience, and infectious disease," notes Gray. He adds that in the next 5 to 10 years, we will be able to create visual constructs of how cells are organized in organs, and that these constructs will provide new insights into how organs function and respond to stimuli.

Chasing the $100 Genome



At the AGBT conference, the last presentation—a coveted speaking slot—promised to be a blockbuster. Its title: "First $100 genome sequencing enabled by new extreme throughput DNBSEQ platform." Its speaker: Radoje (Rade) Drmanac, PhD, CSO of MGI (a subsidiary of China's genomics powerhouse BGI) and a co-founder of Complete Genomics.

Drmanac presented details of two MGI products: the DNBSeq Tx sequencing platform and the CoolMPS sequencing chemistry. He also announced that MGI's sequencing platforms will be made available in the

United States for the first time in April 2020.

Radoje (Rade) Drmanac, PhD, CSO of MGI

The DNBSeq Tx can boast of several improvements over its predecessors. These improvements, Drmanac asserts are "not about going from bad to good." Rather they are "about going from excellent to extremely excellent." MGI reports on average one error in 170 kilobases of sequenced DNA. According to Drmanac, users are usually happy with about one error in one thousand.

With a surface that is five times larger and a doubling of the number of spots that can capture DNA, the DNBSeq Tx can sequence up to 700 genomes on a single run—a capacity that is an order of magnitude larger than any sequencer on the market today. Also notable is the new fluidics system designed to handle the big arrays in the DNBSeq Tx. Rather than a flow cell, where the solutions wash over the samples, a robot dips a large panel through a series of solutions, conserving reagents and cutting cost.

CoolMPS (MPS stands for massively parallel sequencing) is the first nucleobase-specific antibody-based sequencing chemistry. Drmanac told *GEN* that MGI wanted to avoid the DNA "scars" that can accumulate with traditional sequencing methods that use dye-labeled reversibly terminated nucleotides (RTs) and affect the accuracy of subsequent reads.

"MGI has taken a really out-of-the-box approach to drive the per genome cost down for a large project—eliminating fluidics and instead dipping huge slides in reagents" noted Keith Robison, PhD, a leading genomics blogger and authority on next-generation sequencing platforms. The new platform, he says, is "unlikely to influence the pattern of sequencer design," but it's "a clever way to minimize the waste of expensive reagents."

"It's not unthinkable that a $10 genome may be a useful genome," Drmanac suggested. There is "no theoretical limitation." He ended his talk on an optimistic note: "At future AGBTs, we'll hopefully present even crazier innovations and breakthroughs."

Searching for rare variants is becoming more common

"Launching at AGBT is a big deal for us because this is where the technical leaders in the field come," noted Jesse Salk, MD, PhD, CEO of TwinStrand Biosciences. "We want to put our nickel down [physician-speak for 'stand our ground'] and say that our product is ready to go and ready to be widely used."



Jesse Salk, MD, PhD, CEO of TwinStrand Biosciences, presented duplex sequencing technology data to attendees of the annual meeting of Advances in Genome Biology and Technology. The information generated—mutation frequencies, simple mutation spectra, and trinucleotide mutation spectra—can be used together to detect,

TwinStrand has developed duplex sequencing technology to improve sequencing accuracy, making it possible to find low-frequency variants at a level of detection that has not been achievable before. At AGBT, Salk spoke about two of the primary applications where their product could be useful, the first of which is cancer—more specifically, the detection of minimal residual disease in acute myeloid leukemia (AML) patients. The hope is to be able to assess how well treatment works or if cancer remains after treatment. This information could spare patients additional, unnecessary treatments and allow drug companies to run smaller, safer, faster clinical trials.

quantify, and characterize induced DNA mutagenesis. [AGBT]

Why did the company start in this area? Salk, who maintains a part-time position as an oncologist at a University of Washington–affiliated VA hospital, explained to *GEN*, "This is the area where I live." He added that he wants to focus on something that would bring an immediate benefit by addressing this particular unmet need, which is one that he sees on a daily basis. He said, "It's not just nice, it's necessary."

Besides cancer applications, TwinStrand's duplex sequencing could be used in almost any application where detection of rare variants is useful. The technology, Salk suggested, could detect the impact of environmental factors on DNA mutagenesis and recognize patterns of mutations that result from smoking, aflatoxin, aristolochic acid, urethane, etc. This work could be important in areas of research such as forensics, NASA's studies on cosmic radiation, or CRISPR-mediated effects of off-target mutagenesis.

Salk noted that interest has been expressed for applications that he would never have known were important to people. He said that "every few weeks, something new comes up," and that he has a hard time not getting distracted because there are so many new, cool things the company could do.

PacBio embraces being single

Last year at AGBT, Jonas Korlach, PhD, the CSO of Pacific Biosciences (PacBio) could not speak about much of anything. At that time, PacBio was stuck in the middle of a prolonged process (beginning in November 2018) of merging with Illumina. But, with the deal abandoned earlier this year, *GEN* asked Korlach, "What is next for PacBio?"

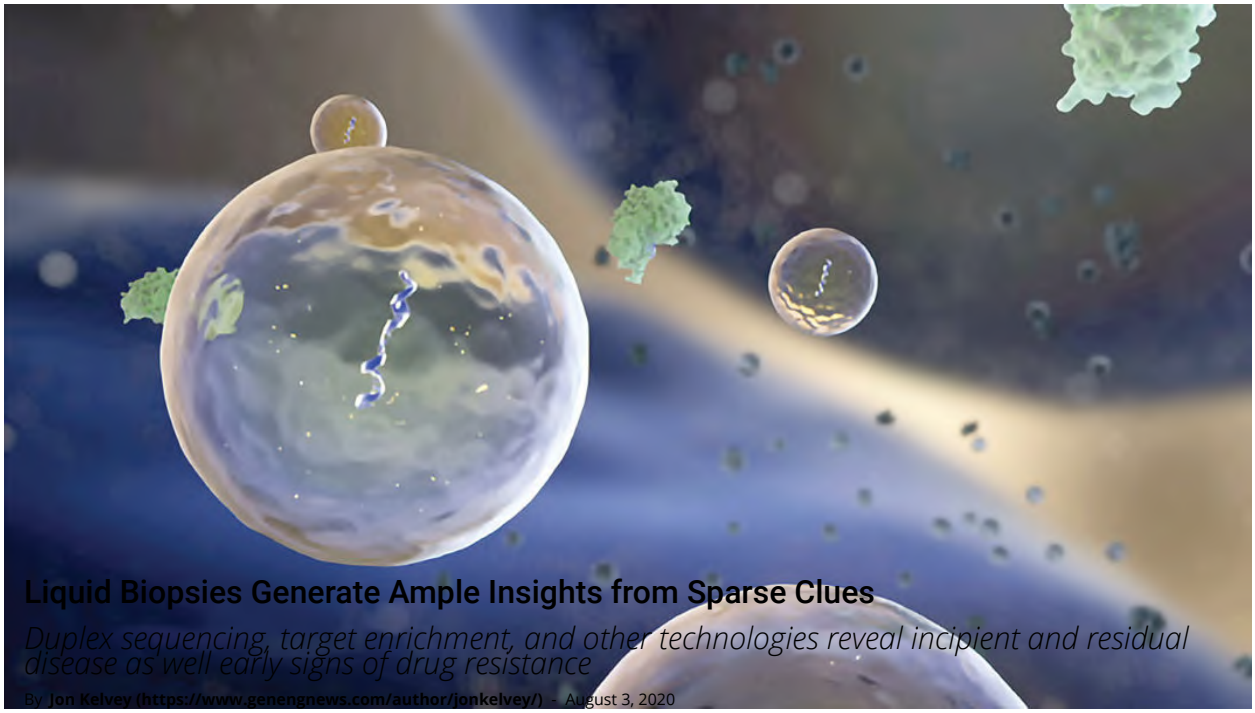
Korlach's answer had a strong, singular message: PacBio is firing on all cylinders. Its HiFi sequencing technology—which offers reads that are long and accurate—is better than ever.

"We feel really good about where we are going," declared Korlach. "A year ago, it wasn't clear what the future would be." Korlach explained that people stayed on the task at hand—to launch the Sequel II—which turned out to be PacBio's most successful launch ever. Now the gurus of long-read sequencing have a place to stand and can turn toward the future.

When asked about the continuous lowering cost of sequencing, Korlach told *GEN* that "there are limitations to the cheap genomes." A cheap genome, he asserted, isn't really a genome. It's an "SNP table" or a series of reads that is mapped to the reference genome. Those are useful, he conceded, but he insisted that when it comes to wanting a high-quality human reference assembly, long stretches are important. This is where he sees PacBio's HiFi continuous long reads playing a key part. They are not as long as those some other long-read technologies can produce, notably nanopore sequencing, but they are more accurate. "To do good functional biology," Korlach concluded, "you need a good genome."

We use cookies to give you a better experience on genengnews.com. By continuing to use our site, you are agreeing to the use of cookies as set in our privacy policy. (https://privacy.liebertpub.com)
Got it!

EXHIBIT G



(https://www.genengnews.com/wp-content/uploads/2020/07/LiquidBiopsies_NIH_ExtracellularVesicles.jpg)

According to Captis Diagnostics, extracellular vesicles (EVs) have several advantages over other liquid biopsies such as circulating tumor DNA (ctDNA) and circulating tumor cells (CTCs). EVs in blood contain circulating biomarkers such as DNA, RNA, lipids, and proteins for potentially comprehensive biomarker analyses. (This image, from the NIH, shows extracellular RNA encapsulated in an EV.)

Like many other professional gatherings in the bioscience world, this year's Liquid Biopsy Summit was canceled due to the COVID-19 (https://www.genengnews.com/coronavirus) pandemic. It was to have been the fifth such event organized by the Cambridge Health Institute, which had arranged for liquid biopsy's leaders to present their latest advances June 15–17 in Seattle.

The cancellation of any bioscience event is unwelcome news, but the loss of this year's Liquid Biopsy Summit may be felt especially keenly because it comes just when the liquid biopsy field is burgeoning. According to Grand View Research, the field attained a value of $24 million in 2016 and is projected to grow to more than $2 billion by 2030.

In partial compensation for the loss, *GEN* presents this article. It is a virtual summit of sorts. It highlights what would have been discussed at the real summit by several key presenters. The emphasis here is on emerging technologies that are helping the liquid biopsy field overcome its biggest challenges: weak signals, subtly variable fragments, and vexing interpretive puzzles.

Tiny differences in tiny samples

"If you [take a liquid biopsy], you are sampling a liquid to detect something," says Jesse Salk, MD, PhD, the CEO of TwinStrand Biosciences. "A blood draw is a liquid biopsy. Taking a urine sample—that's a liquid biopsy. Taking a sample of cerebrospinal fluid—that's a liquid biopsy. You could say a complete blood count is a liquid biopsy."

But when it comes to a liquid biopsy designed to detect signs of a particular cancer, or a rare mutation in a tumor, the signal-to-noise ratio can cause real problems for existing next-generation sequencing (NGS) techniques, according to Salk, who notes that limitations in both optics and biochemistry typically lead to background error rates of 0.1–1%"That's not a problem at all," he says, "if you're looking at the genetic difference between person A and person B while focusing, for example, on the gene that codes for green eyes versus brown eyes."

The background error rate, however, becomes a big problem when you're looking at tiny differences in a tiny sample. "Suppose you're trying to detect one leukemia cell carrying one leukemia-defining mutation mixed in with 100,000 normal cells," Salk suggests. "If your background error rate is 1%, you can't do that."

What TwinStrand has developed is a high-accuracy NGS technique called duplex sequencing, which uses both strands of a DNA molecule in the sequencing process to check for errors. Duplex sequencing, Salk maintains, can reduce the error rate to about one in 10 million. "We can," he continues, "detect incredibly low-frequency variants and mutations." What he had planned for the summit was a presentation on two applications of duplex sequencing: the detection of residual disease in patients being treated for blood cancers, and the detection of emerging drug resistance in those patients.

In patients with acute myeloid leukemia, the front-line treatment, Salk says, is chemotherapy. But determining whether the cancer is truly gone or whether there are residual cancer cells is difficult with existing technologies, and giving additional, cytotoxic treatment carries its own risks for patients who might not actually carry any residual disease.

"Historically, we have done bone marrow biopsies because it gives us a higher concentration of cells, but we know they are pretty unpleasant," Salk acknowledges. "We would love to be able to take a blood sample and look for one in a million cancer cells left." He asserts that if duplex sequencing were performed instead, it would be possible to detect leukemia-related mutations at levels of one in a million. Doing so, he argues, could improve the development of new therapies and better inform treatment decisions. For example, clinicians could identify people who should start or continue taking anticancer therapies. "Clinicians could," Salk emphasizes, "prevent people from getting toxicities from unnecessary treatments when their cancer is already cured."

Duplex sequencing could also help clinicians detect cancer drug resistance as soon as it emerges because the genetic changes that characterize drug resistance in a tumor are understood. At present, Salk says, these changes seldom inform assessments of drug resistance. Instead, the typical clinical sign that a tumor has become resistant to a chemotherapy drug is that a patient relapses.

"What if," Salk asks "you could see that happening really, really early—like months before somebody apparently relapsed. What if you saw their leukemia count come back then?"

With duplex sequencing, he says, clinicians could see resistance occurring at "low, low frequency." He adds that in clinical practices where there are multiple generations of drugs already approved, "you just switch somebody to a different one to prevent their relapse."

TwinStrand Biosciences is currently offering duplex sequencing kits and bioinformatics analysis, but the company is also collaborating with other companies, discusses Salk, in some cases helping the other companies identify patients most suitable to clinical trials.

"You really want to find the subset of patients who are likely to benefit and not the ones who are already cured or not at risk of relapse," he advises. "They just slow down your trial. They flood out your signal." Working with the right patients, he points out, can "really help these trials move along quickly so these drugs can get approved and get to patients."

One means for many ends

At Claret Bioscience, research has focused on capturing genetic information that is typically lost during analysis of the cell-free DNA (cfDNA) in a liquid biopsy, relates Varsha Rao, PhD, the company's director of clinical research and development.

"When DNA is fragmented, it has different types of ends," she says. "There can be blunt ends, 5′ ends, 5′ overhangs, and 3′ overhangs, depending on the nuclear activity. This information is usually lost because you have to do end polishing. You fill in 5′ overhangs or you cleave 3′ overhangs, and then you basically generate blunt-ended molecules."

Rao was headed to the Liquid Biopsy Summit to present on Claret's Single Reaction Single-Stranded LibrarY (SRSLY) technology, which uses a single-step phosphorylation/ligation reaction to produce complex libraries from small samples of DNA. This approach avoids end polishing and preserves the native ends of DNA fragments.

"We believe that DNA fragmentation is important in trying to understand the biology of any disease," she declares. "Our aim is to generate technologies that will retain the DNA ends."

Claret's technology also captures many shorter fragments of DNA, providing more information than other techniques, asserts Rao, who adds that there are a lot of fragments between 30 and 120 base pairs that are lost in conventional methods. "Our technology addresses the shortcomings in current approaches that suggest you have to ignore your short fragments, and that you have to do end repair and lose information about DNA fragmentation points."

Claret is currently offering SRSLY library preparation kits in 12- or 96-reaction formats. According to Rao, the company is just getting started. "We just recently came out of stealth mode," she explains. "We've been selling the kits since November and gaining a lot of attention."

Magnetic enrichment of short fragments

When Apostle entered the liquid biopsy space, the company resolved to develop technology that could capture information that eluded existing technologies. Soon, the company began exploring what it characterizes as a unique solution to a common problem.

Apostle's CEO, David Ge, MD, PhD, notes that any liquid biopsy technology needs to detect a small signal (generated by a diseased tissue) within a huge amount of noise (generated by healthy tissues). In addition to this basic requirement, Ge points to more advanced challenges: "When nucleic acids come from diseased organs or tissues, they are for the most part shorter—

or sometimes even longer—than the nucleic acids from normal tissues. In general, the small pieces from diseased tissues are more diverse or varied compared with those from normal tissues."

The cfDNA molecules of interest are not the sections of genomic DNA that are so large that one might use conventional technology to sequence them. Instead of focusing on these sections, which are hundreds to thousands of base pairs long, Apostle distinguishes between cfDNA fragments that are much shorter. These cfDNA fragments often range from 165 to 175 base pairs if they come from healthy tissues, and they are typically about 20 base pairs shorter if they come from diseased organs.

The very subtle difference between short and even shorter fragments could be very helpful—if it could be measured efficiently in a clinical setting. "What we tried to do," Ge says, "is invent a new way to isolate those circulating, small fragments of genetic materials."

At present, cfDNA enrichment technology usually involves spin columns or magnetic nanoparticles. Seeing an opportunity to use magnetic nanoparticles more effectively, Apostle developed its MiniEnrich technology. "We invented a new way to bind and enrich the functional groups to interact with the nucleic acid fragments, on top of a new magnetic bead core," Ge details. The result, he says, is a technology that can recover DNA fragments as small as 20 base pairs. He adds that MiniEnrich technology provides the resolution necessary to utilize the length of fragments as a means of detecting disease.

Ge is willing to entertain different options for MiniEnrich's introduction to the market. The technology could be developed into a standalone product, or it could be integrated into another company's existing product line. Either way, he hopes that Apostle will bring a finished product to the market before 2021.

A peek inside extracellular vesicles

If the Liquid Biopsy Summit hadn't been canceled, Captis Diagnostics CEO HongZhang He, PhD, would have discussed his company's focus on a very different medium for conducting liquid biopsy. Rather than detect or analyze non-encasulated cfDNA in the blood, Captis harnesses information contained in nanoscale extracellular vesicles (EVs) from tumor tissues.

"EVs have several unique advantages over tissue biopsies and other liquid biopsies such as circulating tumor DNA (ctDNA) and circulating tumor cells (CTCs)," He tells *GEN*. "EVs in blood contain circulating biomarkers such as DNA, RNA, lipids, and proteins for potentially comprehensive biomarker analyses."

Among the more interesting biomarkers found in liquid biopsies are the molecular fragments shed by tumors. Circulating tumor DNA (ctDNA), for example, is being subjected to increasingly sophisticated analyses that can tune into vanishingly weak signals while filtering out enormous amounts of background noise. [Jill George/NIH]

Captis has developed a lipid nanoprobe that can label EVs for magnetic enrichment and isolate EVs from plasma within 15 minutes. According to He, the lipid nanoprobe "does not require expensive and bulky equipment, which is suitable for miniaturization, automation, and high-throughput instrumentation."

Captis published a proof-of-concept study of this technology in 2017, in *Nature*, but He said that the company has more work to do before its approach is ready for the clinic. "In the next phase," He notes, "we will validate the analytical sensitivity/specificity, the clinical sensitivity/specificity, and the clinical utility of the developed assay in large clinical trials evaluating patients with advanced cancer."

Making sequencing a snap

NGS is essential to liquid biopsy approaches, but according to Avida Biomed's head of clinical research, Grace Zhao, PhD, the technology has several drawbacks—it's complicated, time consuming, and often quite expensive, especially when it is used to search for cancer-specific mutations and biomarkers through methylation analysis. "Nowadays, if you need whole genome methylation analysis, you might pay $3000 for cfDNA analysis," she says. "Even if you need smaller coverage, you might pay up to $1000. It's still too expensive if, say, you want to do early screening."

So, Avida's focus as a company, according to Zhao, has been to make NGS as simple and inexpensive as possible. "We named our technology Point-n-Seq™ because we want to make NGS as easy to use as a point-and-shoot camera," she points out.

Point-n-Seq, as Zhao would have discussed at the summit, is a platform for combined methylation sequencing and mutation analysis of cfDNA signals of cancer in blood plasma. It may be used to assess all disease stages, but it may be most valuable as a means of enabling early detection, which requires high sensitivity. Typically, only minuscule signals are emitted by early-stage tumors.

"Current bisulfite-based technologies have a very low recovery rate. If you put in, say, 10 nanograms of human DNA, which is around 3000 copies of the genome, you may get 100–200 copies back," she explains. "If your recovery is so low, you've lost a lot of signal. You've lost your sensitivity."

Point-n-Seq, by contrast, can handle nanogram-level inputs and perform a combined methylation and genetic alteration analysis without splitting a DNA sample. And by focusing on small panels sufficient for one cancer type, Avida can keep the cost of sequencing low, the workflow efficient, and the fidelity of the signal high.

"If you're looking at a huge panel, your background noise is very high," Zhao points out. "The target capture step in Point-n-Seq for targeted methylation sequencing is before C-to-T conversion and before any amplification, which enables highly specific target enrichment even in very focused panels. If we do a very focused panel, it is cheap and fast, and the background noise can be very low."

The result is an NGS tool that has a turnaround time of 1–2 days instead of 5–10 days, Zhao asserts, and at 1/10th the usual sequencing cost.

"Our goal is to make our platform technology readily available to researchers and clinicians doing cutting-edge work, and we're open to partnering with collaborators," she indicates. "If a big pharmaceutical company looks for an assay to screen for patients—say, it wants to stratify the patients at the early relapse stage—we believe that our assay can provide such a tool."

Challenges ahead

The emerging technologies in the liquid biopsy field, like all emerging technologies, present new opportunities and challenges. As Salk notes, liquid biopsy technologies that enhance sensitivity not only probe deeper and detect more fine variations in genetic samples, they also generate data that can be hard to interpret.

"Mutations accumulate at low frequency during life that look completely like they come from cancer, but they are actually just a part of aging," he says. "One has to be very careful now that we are able to detect these, to not automatically assume that this means someone has cancer. With great sensitivity comes great responsibility."

And then there is the more immediate and universal challenge posed by the COVID-19 pandemic, which is complicating operations in every field. The bioscience field—as the cancellation of the Liquid Biopsy Summit demonstrates—is no exception.

"The COVID-19 situation has had a very lasting impact on our operations," says Ge, who adds that the impact goes beyond work-from-home arrangements. "We were working to get to the next stage in clinical studies with a larger sample size," he explains. "That has been paused as

well, as every hospital is fighting COVID-19 at the moment."

Still, Ge believes that the challenges presented by COVID-19, like the liquid biopsy field's technological challenges, will shake out with time. "I personally believe it will be over in no time, within several months," he says. "It will be over, and we will get back on track very soon."

Smarter Immuno-Oncology Trials Using Tumor-Informed ctDNA Analysis

By Alexey Aleshin

Immune checkpoint inhibitors are being increasingly investigated through clinical trials for a range of cancers. However, only a minority of patients benefit from these interventions, and current biomarkers do not reliably predict treatment response.

A prospective Phase II clinical trial, described in *Nature Cancer*, assessed circulating tumor DNA (ctDNA) as a biomarker of response to immune checkpoint inhibitors in patients with advanced solid tumors being treated with pembrolizumab. It demonstrated that sensitive ctDNA assessment, using a tumor-informed assay—namely, the Signatera assay from Natera—can predict response to pembrolizumab in a histology-agnostic manner.

Signatera provided a readout on the trial in as short a time as six weeks, while complementing standard radiologic imaging to differentiate pseudoprogression from true progression, potentially enabling patients who are deriving clinical benefit to continue therapy, while sparing others from unnecessary toxicities and costs.

We use cookies to give you a better experience on genengnews.com. By continuing to use our site, you are agreeing to the use of cookies as set in our privacy policy. (https://privacy.liebertpub.com)

Got it!

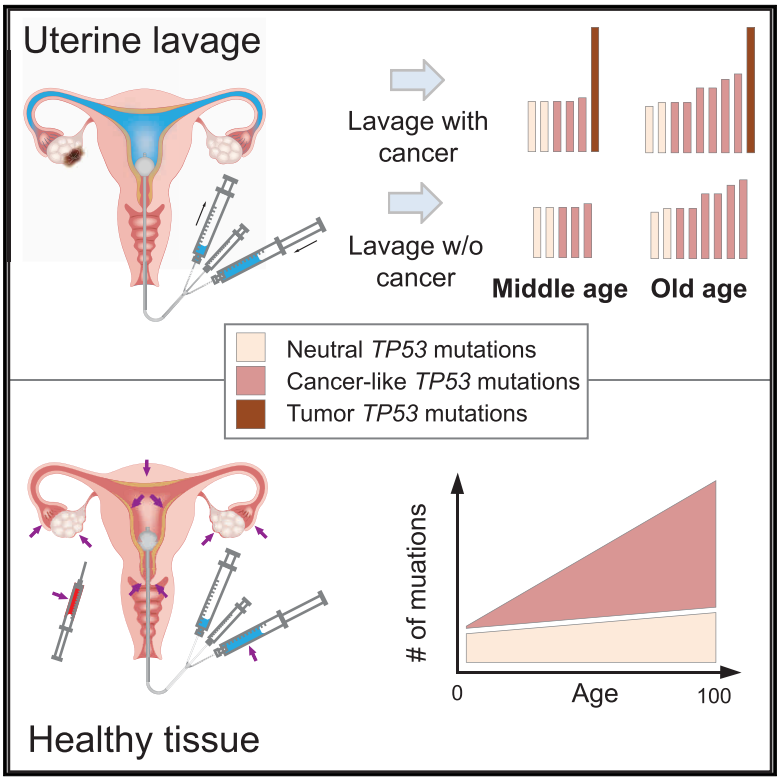
EXHIBIT H

Article

Cell Reports

Ultra-Sensitive *TP53* Sequencing for Cancer Detection Reveals Progressive Clonal Selection in Normal Tissue over a Century of Human Lifespan

Graphical Abstract



Authors

Jesse J. Salk, Kaitlyn Loubet-Senear, Elisabeth Maritschnegg, ..., Robert Zeillinger, Paul Speiser, Rosa Ana Risques

Correspondence

rrisques@uw.edu

In Brief

Salk et al. demonstrate that ultra-sensitive DNA sequencing to identify *TP53* mutations among cells shed into uterine fluid shows potential for minimally invasive ovarian cancer detection. Yet they also reveal ubiquitous age-related accumulations of cancer-like *TP53* mutations in the normal tissues of healthy women. This highlights an important challenge of using tumor driver mutations for cancer screening.

Highlights

- Ovarian cancer can be detected by ultra-accurate sequencing of uterine lavage DNA
- However, low-frequency *TP53* mutations also exist in normal tissue of healthy women
- *TP53* mutations are increasingly selected for with age, revealing somatic evolution
- Age-associated, cancer-like mutations challenge specificity for cancer detection

Salk et al., 2019, Cell Reports *28*, 132–144
July 2, 2019 © 2019 The Authors.
https://doi.org/10.1016/j.celrep.2019.05.109

CellPress

OPEN ACCESS
CellPress
Cell Reports
Article

Ultra-Sensitive *TP53* Sequencing for Cancer Detection Reveals Progressive Clonal Selection in Normal Tissue over a Century of Human Lifespan

Jesse J. Salk,[1,2] Kaitlyn Loubet-Senear,[3,12] Elisabeth Maritschnegg,[4,13] Charles C. Valentine,[2] Lindsey N. Williams,[2] Jacob E. Higgins,[2] Reinhard Horvat,[5] Adriaan Vanderstichele,[6] Daniela Nachmanson,[3,14] Kathryn T. Baker,[3,15] Mary J. Emond,[7] Emily Loter,[8] Maria Tretiakova,[3] Thierry Soussi,[9,10,11] Lawrence A. Loeb,[3] Robert Zeillinger,[4] Paul Speiser,[4] and Rosa Ana Risques[3,16,*]

[1]Department of Medicine, Division of Medical Oncology, University of Washington, Seattle, WA 98195, USA
[2]TwinStrand Biosciences, Seattle, WA 98121, USA
[3]Department of Pathology, University of Washington, Seattle, WA 98195, USA
[4]Molecular Oncology Group, Department of Obstetrics and Gynecology, Comprehensive Cancer Center-Gynecologic Cancer Unit, Medical University of Vienna, Vienna, Austria
[5]Department of Pathology, Medical University of Vienna, Vienna, Austria
[6]Department of Gynecologic Oncology, Leuven Cancer Institute, University Hospitals Leuven, Katholieke Universiteit, Leuven, Belgium
[7]Department of Statistics, University of Washington, Seattle, WA 98195, USA
[8]Department of Pathology, Seattle Children's Hospital, Seattle, WA 98105, USA
[9]Sorbonne Université, UPMC Université Paris 06, 75005 Paris, France
[10]Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden
[11]INSERM, U1138, Centre de Recherche des Cordeliers, Paris, France
[12]Present address: Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA
[13]Present address: VIB-KU Leuven Center for Brain and Disease Research, Leuven, Belgium
[14]Present address: Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA, USA
[15]Present address: Knight Cancer Institute, Oregon Health and Science University, Portland, OR, USA
[16]Lead Contact
*Correspondence: rrisques@uw.edu
https://doi.org/10.1016/j.celrep.2019.05.109

SUMMARY

High-accuracy next-generation DNA sequencing promises a paradigm shift in early cancer detection by enabling the identification of mutant cancer molecules in minimally invasive body fluid samples. We demonstrate 80% sensitivity for ovarian cancer detection using ultra-accurate Duplex Sequencing to identify *TP53* mutations in uterine lavage. However, in addition to tumor DNA, we also detect low-frequency *TP53* mutations in nearly all lavages from women with and without cancer. These mutations increase with age and share the selection traits of clonal *TP53* mutations commonly found in human tumors. We show that low-frequency *TP53* mutations exist in multiple healthy tissues, from newborn to centenarian, and progressively increase in abundance and pathogenicity with older age across tissue types. Our results illustrate that subclonal cancer evolutionary processes are a ubiquitous part of normal human aging, and great care must be taken to distinguish tumor-derived from age-associated mutations in high-sensitivity clinical cancer diagnostics.

INTRODUCTION

Worldwide, >250,000 new cases of ovarian cancer are diagnosed each year, and two-thirds of these women die from the disease (Bray et al., 2018). This high mortality is largely due to the high frequency of metastasis before diagnosis and a lack of effective screening and early detection methods. More than 60% of cases are diagnosed at an advanced stage, when the 5-year survival rate is only 29% (Siegel et al., 2017). In contrast, survival for women with localized disease is 92%, indicating that early ovarian cancer detection could vastly decrease mortality, yet diagnosis at this stage is rare. The most used approach for ovarian cancer screening involves a combination of serum protein CA-125 level and transvaginal ultrasound, but this has not demonstrated survival benefit, and it may result in harm due to false-positives, such as unnecessary surgeries in cancer-free women (Henderson et al., 2018). Thus, the US Preventive Services Task Force recommends against its use (Grossman et al., 2018). Better tools for early ovarian cancer detection remains an urgent and unmet clinical need (Drescher and Anderson, 2018).

In recent years, it has been demonstrated that cancers can be non-invasively detected in "liquid biopsies," that is, blood or other body fluids in which cancers shed cells or DNA (Diaz and Bardelli, 2014). Proof-of-principle for this approach in ovarian cancer detection was initially accomplished via the identification

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

OPEN ACCESS
CellPress

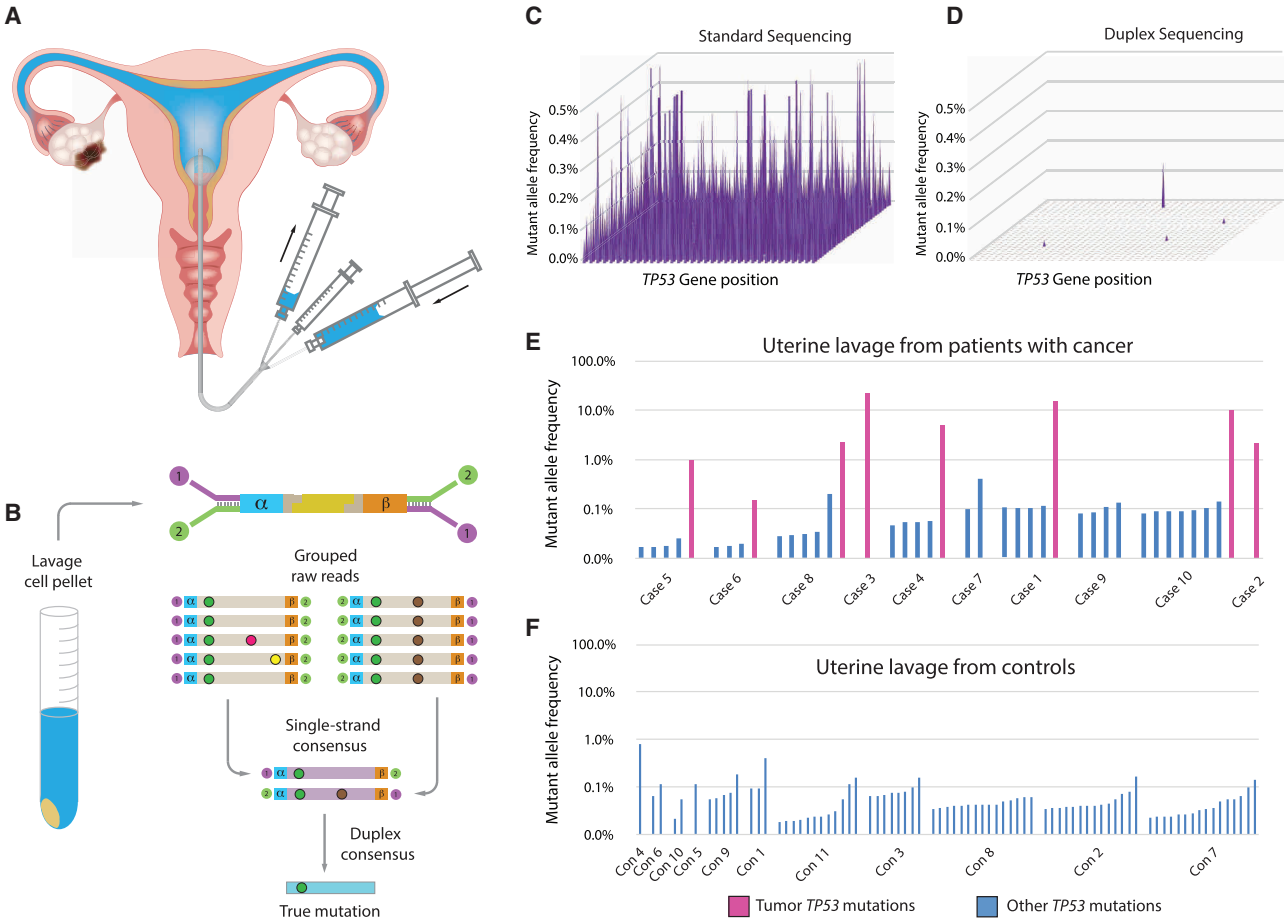


Figure 1. Detection of Ovarian Cancer Using UL and DS

(A) UL is carried out by passing a small catheter through the cervix, followed by concurrent flushing and aspiration with 10 mL of saline.

(B) After cell isolation by centrifugation, DNA is extracted, fragmented, and ligated with DS adapters, which include degenerate molecular tags (α and β). Following amplification, hybrid capture, and sequencing, reads sharing the same tags are grouped and mutations are scored only if present in both strands of each original DNA molecule.

(C and D) Each spot on the 2D surface represents one of the 1,179 coding positions in *TP53*. The y axis indicates mutant allele frequency (MAF). By standard NGS, all positions show false mutations (C). DS of the same sample (case 6 in E) eliminates errors and reveals only true mutations (D).

(E and F) *TP53* mutations identified by DS in UL from women with ovarian cancer (E) and women who are cancer-free (controls; F). Fuchsia bars, matching tumor mutation; blue bars, biological background mutations. Mutations are sorted by ascending MAF within each patient and patients are sorted by age.

of tumor-derived mutations in DNA extracted from routine Papanicolaou (Pap) tests (Kinde et al., 2013). Although the sensitivity for ovarian cancer detection was only 41%, these findings supported the exciting possibility that ovarian cancer could be detected by the genetic identification of cancer cells disseminated into the gynecological tract. A follow-up study recently reported that improved sensitivity, up to 63%, could be obtained by combining mutation detection in Pap tests and plasma. In addition, sampling with an intrauterine brush also improved sensitivity, probably due to increased tumor cell recovery by more proximal collection to the anatomical site of tumors (Wang et al., 2018).

An alternate means for tumor cell collection, developed by members of our team, consists of trans-cervical lavage of the uterine cavity (Figure 1A; Maritschnegg et al., 2015). This method improves the efficiency of collection by rinsing the uterus and fallopian tubes, the latter being the site of origin of most serous ovarian cancers (Labidi-Galy et al., 2017). This lavage technique demonstrated 80% sensitivity for ovarian cancer detection (Maritschnegg et al., 2015). The challenge, however, was that cancer-derived mutations, particularly those from early-stage tumors, were often present in a very small fraction of the total lavage DNA. To detect these mutations, digital droplet PCR (ddPCR) was required, which is a sensitive method but not practical for prospective screening because the tumor mutation needs to be known *a priori*.

Next-generation DNA sequencing (NGS) is a widely used, variant-agnostic form of mutation detection, but it has a background error rate of up to ~1%, which precludes confident identification of lower-frequency mutations (Salk et al., 2018). Of the mutations that make up the 80% sensitivity achieved in our previous study, conventional NGS missed 25% (Maritschnegg

et al., 2015). Currently, the most accurate NGS method is Duplex Sequencing (DS) (Salk et al., 2018), which uses double-stranded molecular barcodes for error correction and decreases the error rate of sequencing from 10^{-3} to $<10^{-7}$ (Kennedy et al., 2014; Schmitt et al., 2012). We previously demonstrated that DS can detect ovarian cancer-derived mutations in DNA extracted from peritoneal fluid at frequencies as low as 1 tumor mutation per 25,000 normal genomes (Krimmel et al., 2016). This extreme sensitivity for mutation detection also led to the discovery of prevalent yet very-low-frequency (<0.01%) *TP53* mutations in both the peritoneal fluid and peripheral blood from healthy women. These "biological background" mutations resembled *TP53* mutations found in cancers, but appeared to result from the normal aging process. This observation was among the first of an emerging body of literature that has identified age-related, cancer-associated mutations within non-cancerous tissue (Risques and Kennedy, 2018).

In the present study, we combine the most sensitive reported sampling technique for ovarian cancer detection, uterine lavage (UL), with the highest-accuracy sequencing technology available, DS. High-grade serous ovarian carcinoma (HGSOC) is both the most common and the most deadly histological type, accounting for 70%–80% of ovarian cancer deaths (Bowtell et al., 2015). More than 98% of HGSOCs carry mutations in *TP53*, which makes this gene an ideal, cost-effective target for sequencing (Ahmed et al., 2010; The Cancer Genome Atlas Research Network, 2011; Vang et al., 2016). In addition, *TP53* mutations are one of the earliest genetic events in HGSOC formation, as demonstrated by their presence in early serous epithelial proliferations found in the fallopian tubes (often termed p53 signatures), as well as serous tubal intraepithelial carcinomas (Chien et al., 2015; Kuhn et al., 2012; Kurman and Shih, 2010; Soong et al., 2019).

The primary goal of this study is to demonstrate the technical feasibility of using DS to deeply sequence *TP53* from UL as a potential test for ovarian cancer detection. We capitalize on the accuracy of DS to identify true-positive cancer-derived mutations and to uniquely detect low-frequency, age-related mutations that may affect diagnostic performance. To better understand the extent and nature of these biological background mutations, we perform a detailed characterization of somatic *TP53* mutations in multiple gynecologic tissues from women without ovarian cancer of ages spanning a century of human lifetime.

RESULTS

DS Detects Ovarian Cancer Mutations in ULs with High Sensitivity

We used DS to examine the coding region of *TP53* in DNA extracted from the lavage cell pellet (Figures 1A and 1B) of 10 women with ovarian cancer and 11 controls under blinded conditions. Most of the lavages from women with ovarian cancer were included in the original study that reported 80% sensitivity for ovarian cancer detection with prior knowledge of tumor mutation (Maritschnegg et al., 2015; Table S1). We sought to determine whether similarly high sensitivity was possible without prior knowledge of the tumor mutation by using DS. DS uses special adapters with double-stranded molecular barcodes, which allow the identification of sequencing reads that are derived from both strands of each starting DNA molecule. Mutations are only scored if they are present in the majority of reads from both DNA strands, effectively eliminating sequencing and PCR artifacts (Figure 1B). The estimated error rate of DS is <1 in 10 million (Schmitt et al., 2012), which allows for extreme sensitivity and specificity of mutation detection when carrying out high-depth sequencing. To illustrate the superior accuracy of DS compared with standard NGS, an example of a UL sample (case 6) processed by both methods is shown in Figures 1C and 1D. Standard NGS entails alignment and variant calling from Illumina sequencing reads. Whereas every nucleotide position in the gene artifactually appears mutated in 0.1%–0.5% of molecules with standard sequencing, DS eliminates these tens of thousands of erroneous mutations to reveal the known tumor mutation at a mutant allele frequency (MAF) of 0.15%, a value that is very close to the frequency previously determined by ddPCR (0.12%; case 6 in Table S2; Maritschnegg et al., 2015).

Among the 10 lavages from women with ovarian cancer, we identified the expected tumor mutation (fuchsia bars, Figure 1E) in 8, matching the post hoc 80% sensitivity of the previous study. In the subset of these lavages that had been analyzed by conventional NGS and/or ddPCR, we confirmed tumor mutations at similar allele frequencies in most cases (Table S2). In addition to the tumor mutations, in nearly all of the lavages from women with and without ovarian cancer, we identified very-low-frequency *TP53* mutations (blue bars, Figures 1E and 1F). To confirm that these mutations were not due to technical errors, two of the mutations identified in controls (lavages con2 and con7) were assessed by ddPCR (Table S2). This orthogonal assay demonstrated that these mutations, present at a comparable frequency of <0.1% by both assays, were authentic. As a further demonstration of the high sensitivity, accuracy, and precision of DS, we carried out a mixing experiment whereby cell lines bearing point mutations in *TP53* and two other genes were spiked into a normal DNA sample at ratios from 1/100 to 1/100,000. Three replicates of this mixture were prepared and sequenced on different days at DS depths of up to 400,000×. All of the variants were identified in each replicate at expected frequencies (R^2 of 0.95–0.98), demonstrating excellent reproducibility and accuracy (Figure S1).

Although *TP53* background mutations were common, their MAF was always <1% (Figures 1E and 1F), which could be used as a threshold to optimally identify patients with ovarian cancers from patients who are cancer-free. In this, albeit small, pilot study, a 1% threshold yielded a sensitivity of 70% and a specificity of 100%. Furthermore, the maximum mutation frequency was significantly higher in cases than controls ($p = 0.0005$ for test of no difference in log maximum frequency between groups and no change after adjustment for age), and in the lavages in which the tumor mutation was identified, its frequency was at least 10-fold above the highest background mutation in that individual.

TP53 Mutations in UL Increase with Age

To better understand the basis of *TP53* background mutations, we examined the association of their abundance with

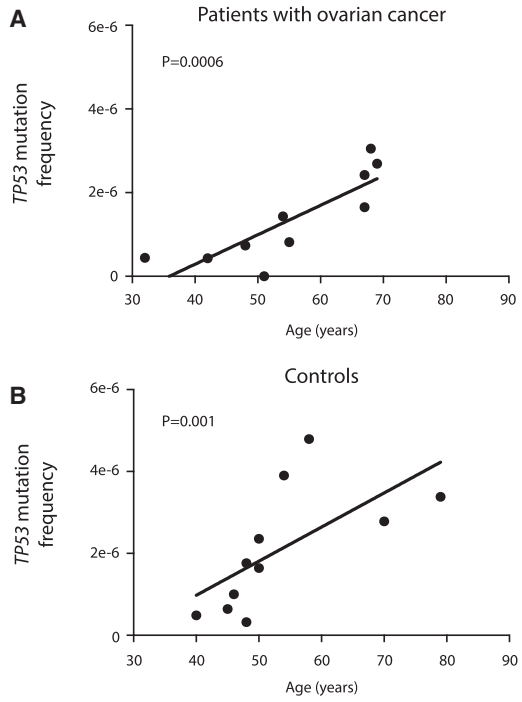


Figure 2. The Frequency of *TP53* Mutations in UL Increases with Age

Frequency is calculated as the total number of unique *TP53* mutations identified in each sample (including exons and flanking intronic regions) divided by the total number of DS nucleotides sequenced.

(A) UL samples from patients with ovarian cancer; n = 10, r = 0.89, p = 0.0006 by Spearman's correlation test.

(B) UL samples from control patients without cancer; n = 11, r = 0.83, p = 0.001 by Spearman's correlation test.

age. When patients were ordered by ascending age (Figures 1E and 1F), it appeared that older patients carried more mutations. However, the number of mutations found depends on the total number of nucleotides sequenced (Figure S2), which was variable across samples and tended to be higher in controls due to increased sequencing depths (Table S1). To compensate for this variation, for each sample we calculated the total *TP53* mutation frequency by dividing the number of mutations identified in UL (including exons and flanking intronic sequences) by the total number of DS bases sequenced. For patients with cancer, we excluded the tumor mutation from this calculation to fairly reflect only *TP53* background mutations. For patients with ovarian cancer, as well as cancer-free control patients, the *TP53* mutation frequency significantly increased with age (Figure 2; p = 0.0006 for ovarian cancer, p = 0.001 for controls, Spearman's correlation test). This trend was identical to prior observations of *TP53* background mutation frequency in peritoneal fluid and peripheral blood (Krimmel et al., 2016).

TP53 Mutations in UL Are Not Random, but Rather Are Positively Selected

The *TP53* gene is a tumor suppressor, the genetic disruption of which facilitates cell proliferation in tumors, even when only one allele is mutated (Leroy et al., 2014). An age-associated increase in ultra-low-frequency *TP53* background mutations could result from random, age-related mutagenic processes or, alternatively, from mutagenesis coupled with clonal selection of pathogenic variants. To distinguish between these possibilities, we performed a detailed analysis of traits of selection among the 112 age-associated *TP53* background mutations collectively identified among all 21 patients (tumor mutations and intronic and UTR mutations excluded) (Figure 3; Table S3).

First, we calculated the proportion of non-synonymous *TP53* mutations in ULs from cases and controls and compared it with the expected proportion when considering all of the possible mutations in the *TP53* gene (n = 3,546). The percentage of non-synonymous mutations in both UL controls (90.5%) and cases (93.5%) was significantly higher than expected under no selection (76.6%) with p = 0.0035 and p = 0.031, respectively (Figure 3A, exact binomial test). The excess of non-synonymous mutations was not driven by a subset of outlier samples, but rather was uniformly observed across nearly all of the lavage samples (Figure S3A).

Second, we examined the metrics of selection related to the genic location of mutations. Background *TP53* mutations were not randomly distributed along the gene but clustered in certain regions of biological significance. Nearly 25% of *TP53* lavage background mutations occurred in the context of CpG dinucleotides, which is remarkable given the fact that these dinucleotides make up <5% of the coding region of *TP53* (Figure 3B, $p = 2.2 \times 10^{-10}$ for controls and $p = 1.4 \times 10^{-5}$ for ovarian cancer mutations, by the exact binomial test). Mutations were also enriched in exons 5–8, which encode the DNA-binding domain of the protein (Figure 3C, $p = 2.4 \times 10^{-6}$ for controls and $p = 4.5 \times 10^{-4}$ for ovarian cancer mutations, by exact binomial test). The most significant enrichment, however, was observed in *TP53* cancer-associated hotspot codons, which are the codons most recurrently mutated in cancer. We considered the nine most abundantly mutated codons in the Universal Mutation Database (UMD; April 2017 version) (Leroy et al., 2014). These codons encompass only 2.3% of the coding region of *TP53*, yet >25% of lavage background mutations clustered within these 27 bp (Figure 3D, $p = 5.1 \times 10^{-17}$ for controls and $p = 2.4 \times 10^{-9}$ for ovarian cancer mutations, by exact binomial test), and among these, all were non-synonymous. The biases for each characteristic were not driven by outliers but were distributed evenly across samples in both groups (Figure S3B–S3D). Cases and controls were not significantly different for any of these traits (using Fisher's exact test or using generalized estimating equations [GEEs]) to account for the correlation between patients with and without adjustment for sequencing depth).

To assess the impact of these mutations on *TP53* protein function, we took advantage of Seshat, a recently developed online tool for *TP53* analysis that provides comprehensive mutational information, including prediction of impact on protein activity and pathogenicity (Tikkanen et al., 2018). We queried all background *TP53* mutations identified in the 21 lavages and color-coded them according to 5 binned categories of protein activity and predicted pathogenicity. Nearly all of the samples carried at least one *TP53* mutation that inactivated the protein totally or

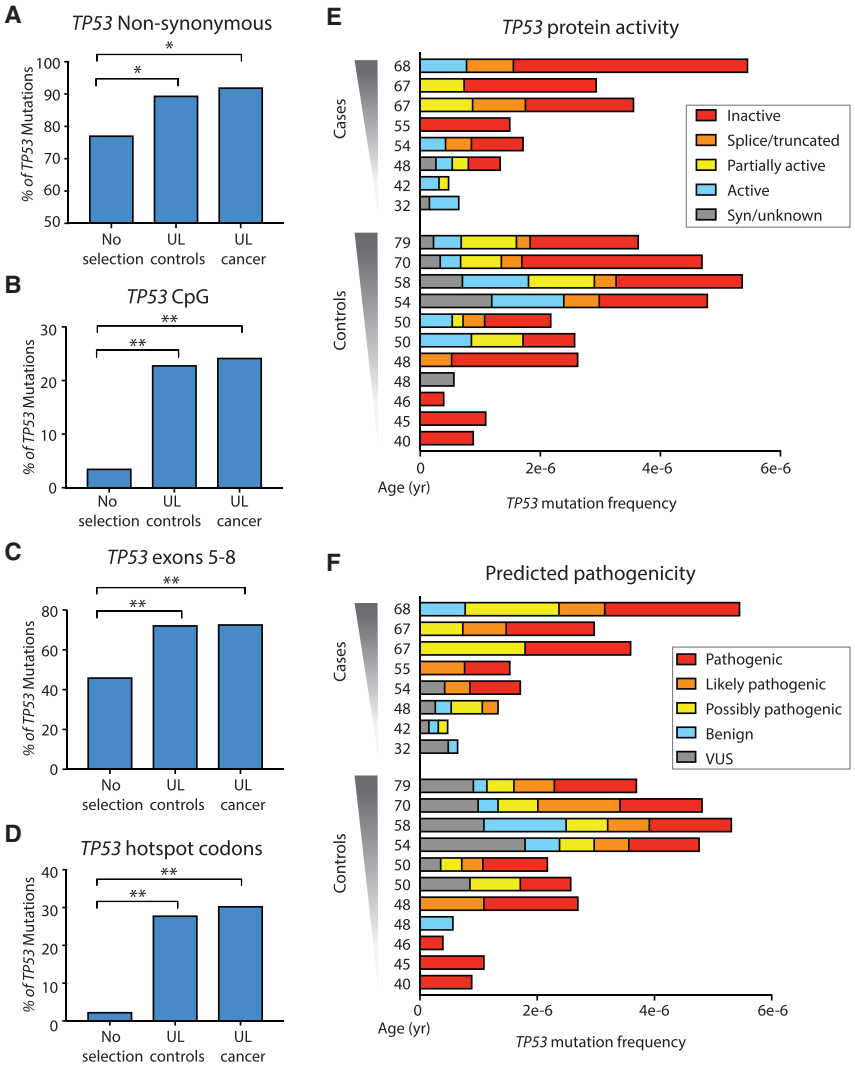


Figure 3. Evidence of Positive Selection in *TP53* Background Mutations from ULs

(A–D) Percentage of non-synonymous *TP53* mutations (A), and percentage of *TP53* mutations localized in CpG dinucleotides (B), in exons 5–8 (C), and in hotspot codons (D). For (A–D), *TP53* mutations identified in UL from controls and cancer significantly exceed expected values without selection. *$p < 0.05$, **$p < 0.0001$ by binomial exact test, n = 79 for UL controls, and n = 33 for UL cancer. (E and F) Protein activity (E) and predicted pathogenicity (F) color-coded as five groups from Seshat data. Patients are sorted by ascending age. For each patient, *TP53* mutation frequency is calculated as the number of mutations in the coding region divided by the total number of DS nucleotides sequenced in that region. Two cancer patients with unusually low sequencing depth and no identified *TP53* mutations are not shown. Nearly all cases and controls carry mutations that have an impact on protein activity and predicted pathogenicity.

partially (Figure 3E) and/or was predicted to be pathogenic (Figure 3F). Cases and controls were not significantly different when comparing the proportion of mutations that inactivated protein activity (categories of "inactive," "splice/truncated," and "partially inactive" in Figure 3E) or the proportion of mutations with predicted pathogenicity (categories of "pathogenic," "likely pathogenic," and "possibly pathogenic" in Figure 3F) (Fisher's exact test p values are 0.49 and 0.99, respectively). The unambiguous signature across six distinct metrics of positive selection (Figures 3A–3F) within the ultra-low-frequency *TP53* mutations observed in all lavages, regardless of cancer status, indicate that these mutations expanded under strong selective pressure and are not the result of technical errors.

TP53 Mutations in UL Resemble Mutations in Cancer

We next compared the features of the selection of *TP53* mutations identified in lavages to *TP53* mutations from cancers. For this analysis, we used all of the cancer mutations present in the UMD cancer database (April 2017, n = 71,051). We determined the percentage of these mutations that reside at CpG sites, cancer hotspots, and exons 5–8, as well as the percentage of mutations that affect protein activity (first 3 categories in Figure 3E) or are predicted to be pathogenic (first 3 categories in Figure 3F). For each trait, we compared the distribution of mutations in the theoretical absence of selection, in the 21 ULs, and in the cancer database (Figure 4A). Remarkably, for all of the traits, *TP53* background mutations from ULs far more strongly resembled *TP53* mutations in the cancer database than the pattern expected by random chance. This was true for *TP53* mutations found in lavages from cases and controls (Figure S4), as expected due to the fact that mutations in both groups were not statistically different for these traits (Figure 3). We also used a feature of Seshat that categorizes *TP53* mutations according to their frequency in the UMD database. Nearly all of the UL samples harbored *TP53* mutations listed as "frequent" or "very frequent" in the database (Figure 4B). Again, cases and controls did not significantly differ in the proportion of mutations that were common in the cancer database (neither when comparing "very frequent" and "frequent" versus the rest nor when comparing all groups separately; Fisher's exact test, p = 0.84 and p = 0.27, respectively).

To further characterize background *TP53* mutations in UL in comparison to those in cancers, we compared mutation type, spectrum, and gene location. Non-cancer-derived mutations in ULs from women with and without cancer were predominantly missense, similar to mutations in the database (Figure 4C), and displayed a mutational spectrum enriched in G > A and C > T transitions, comparable to cancer mutations (Figure 4D). The distribution of low-frequency *TP53* background mutations from only 21 women along the length of the gene is a mirror image of the distribution of *TP53* mutations from >71,000 different

OPEN ACCESS
CellPress

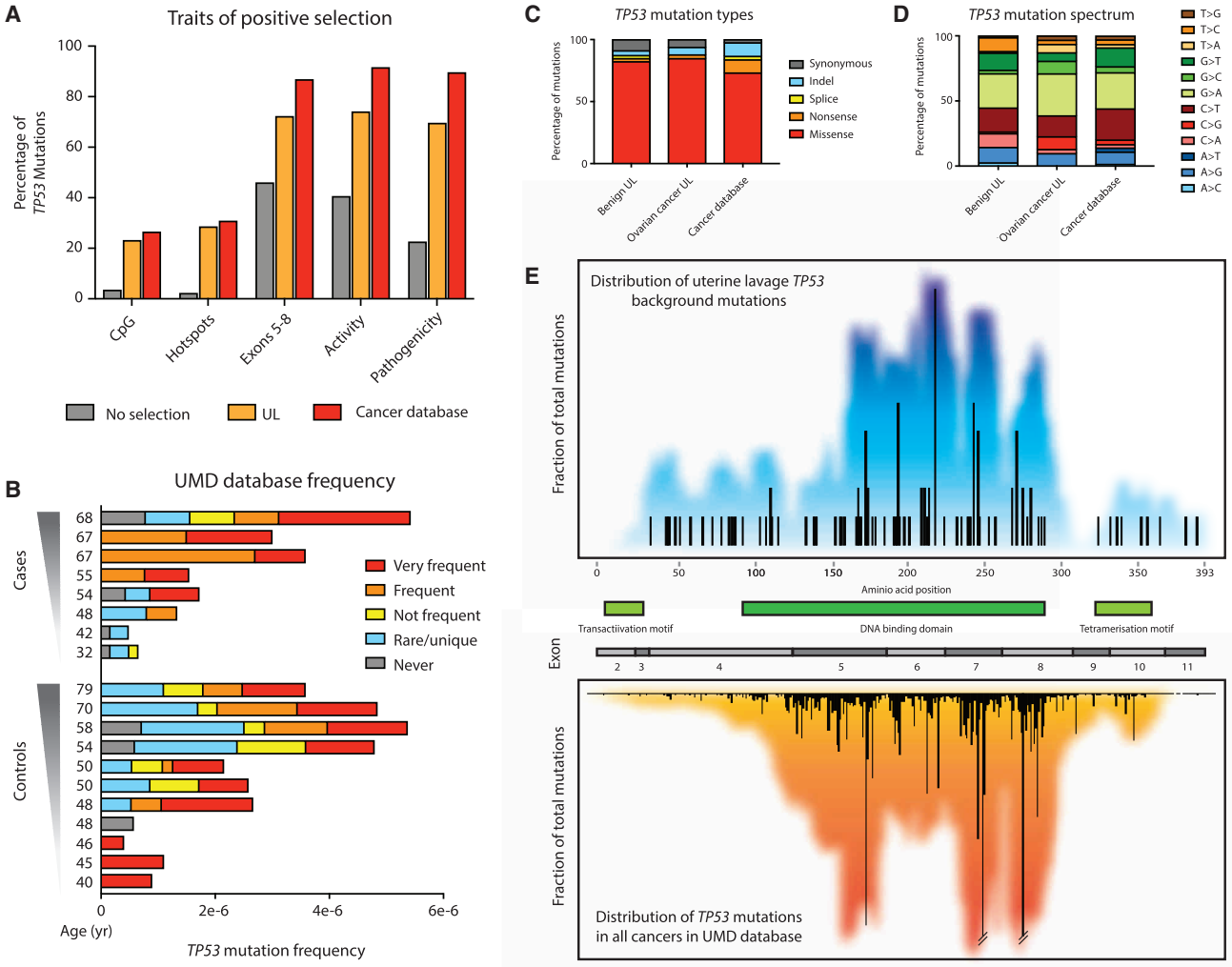


Figure 4. *TP53* Mutations in UL Are Very Similar to *TP53* Mutations Found in Human Cancers

(A) Traits of positive selection are compared between all of the possible mutations in the *TP53* coding region (no selection; n = 3,546), *TP53* background mutations found in UL (n = 112), and *TP53* mutations reported in the UMD cancer database (n = 71,051).

(B) UL mutations in cases and controls color-coded by their abundance in the UMD *TP53* database. For each sample, *TP53* mutation frequency is calculated as the number of mutations in the coding region divided by the total number of DS nucleotides sequenced in that region. Most samples harbor *TP53* mutations that are common in the database.

(C and D) *TP53* mutation type (C) and *TP53* mutation spectrum (D) are compared between mutations identified in UL from controls (n = 79), UL from cases (n = 33), and the UMD database (n = 71,051).

(E) *TP53* mutation distribution map for UL (top) and the UMD cancer database (bottom). Bars quantify the frequency of mutations at each codon. Colored background reflects a 20-bp sliding window average of mutation density around each position. Gene exons and protein domains are indicated in the center section.

tumors included in the database (Figure 4E). Thus, the somatic *TP53* mutations recovered from cells sloughed into the uterine cavity from normal healthy women are not random, but appear to emerge from an evolutionary process of mutation, selection, and clonal expansion akin to what takes place in tumors, but within normal tissue.

TP53 Mutations Are Common in Healthy Tissues from Middle-Aged Women

These striking results prompted us to consider what the tissue origin of the mutation-bearing cells in the ULs may be. To address this question, we sequenced *TP53* from DNA obtained from pre-operative UL and peripheral blood, as well as multiple gynecological tissues collected postoperatively following total hysterectomy and bilateral salpingo-oophorectomy for symptomatic fibroids (benign leiomyomas) from two middle-aged women (Figure 5A; Table S4). DNA was extracted and processed for DS as before, except that samples were sequenced to a higher average depth (Table S4). We identified *TP53* mutations in all of the samples from a 56-year-old woman and in all but 2 samples from a 46-year-old woman (Figure 5B). When we compared the mutation frequency across all of the samples,

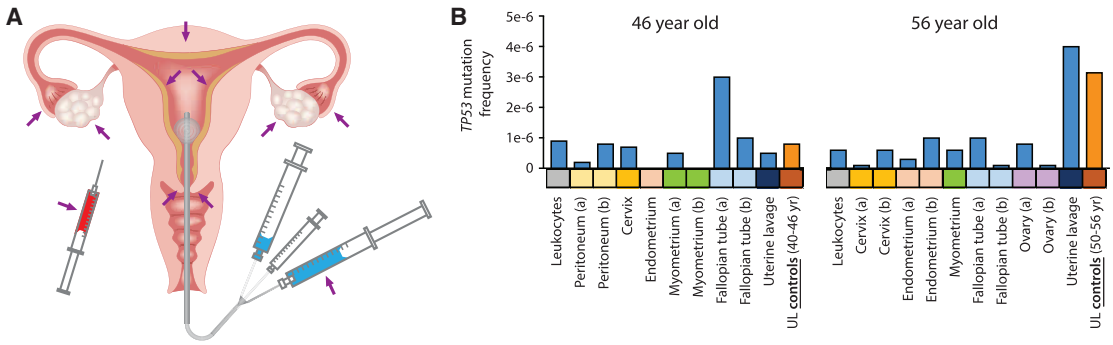


Figure 5. *TP53* Mutations in Normal Tissues and UL from Two Middle-Aged Women

(A) Normal tissues collected included leukocytes, peritoneum, cervix, endometrium, myometrium, fallopian tube, ovary, and UL.

(B) *TP53* mutation frequency for each sample calculated as the number of *TP53* mutations in the coding region divided by the total number of DS nucleotides sequenced. (a) and (b) indicate two spatially separated samples from the same tissue. The blue bars correspond to the samples in this study; the orange bars correspond to the mean values for ULs from control women in the first study.

several interesting observations emerged. First, the UL from the 56-year-old had a mutation frequency that appeared disproportionally high, both when compared to that of most other tissues and when compared with the lavage of the 46-year-old. However, when compared to the mean values of ULs from control women of similar ages (50- to 56- and 40- to 46-year-olds) from the first part of the study, the frequencies by age were quite similar (Figure 5B).

Moreover, the distribution of mutations according to each trait of positive selection (type, frequency in the cancer database, predicted activity and pathogenicity, exon clustering, CpG clustering, and enrichment for cancer hotspots) was comparable to the lavages previously analyzed (Figure S5). Both lines of reasoning support the conclusion that the elevated frequency of mutations in the UL of the 56-year-old woman is not artifactual and confirm the previously observed age effect.

For other tissues, however, we did not observe an obvious increase in *TP53* mutation frequency between 46 and 56 years of age. There was substantial variability in the mutation content of different tissues and of different biopsies of the same tissue, which reflects either a stochastic effect or the imprecision of macrodissection for obtaining exactly comparable tissue samples (e.g., the depth of endometrium harvested, how distal the tubal fimbriae were cut). No single tissue stood out as obviously more mutation prone than another, nor could any tissue be identified as a dominant source of the mutations found in lavages. When we checked for mutations shared between tissues, we did not find any common mutation between UL and any of the rest of the tissues analyzed from the 49-year-old woman and only 1 mutation in the 56-year-old woman that was shared between UL and leukocytes (Figure S6).

TP53 Mutations Increase in Number and Cancer-like Features during Normal Human Aging

With the hope of observing a stronger aging mutational signal, we looked to tissue samples from greater extremes of age. While the procurement of such material was challenging, we obtained several gynecologic tissues at autopsy from a neonate who died from a congenital vascular malformation and from a 101-year-old female who died of natural causes (Table S4). Together with the middle-aged samples, this unique specimen collection represents the full breadth of a century of human lifespan.

Although the tissue types available were not fully identical across all four subjects, the pattern of *TP53* mutations, nevertheless, yielded unique insights. To help more intuitively visualize this multiparametric data, in Figure 6 we annotated all of the mutations found among the different tissues of the four subjects as color-coded boxes for each feature of selection: red for "cancer-like," blue for non-cancer-like. The number of columns of colored boxes per sample reflects the total number of mutations identified. When viewed in this format, it is apparent that *TP53* mutations are not only more abundant with age but are also more "cancer-like." Mutations found in older tissues are disproportionately observed in cancers and are predicted to inactivate the protein or be otherwise pathogenic. In contrast, mutations found in the newborn are rarely found in cancer, tend to preserve the protein activity, and are not predicted to be pathogenic.

Different tissues and different biopsies within the same tissue showed substantial variability in both the number of mutations and their cancer-like features. In aggregate, fallopian tube epithelium appeared to be a "hot" tissue, with a high number of mutations and a high percentage of cancer-like mutations. However, in the 56-year-old, one fallopian tube sample harbored only a single synonymous mutation, which is consistent with the notion of "hot" and "cold" zones within a tissue. This was similarly seen in the two distinct endometrial biopsies of the centenarian.

In addition to a larger number of clones, with aging we would predict an increase in the size of clones, as would be reflected by a higher MAF of each variant found. However, in this study, some samples were sequenced at a lower depth, which may lead to outlier (low event count) biases in the calculation of MAFs, thus precluding a fair comparison between samples (Figure S7). Despite this caveat, 2 large clones were clearly seen within the peripheral blood leukocytes of the 101-year-old (Figure S7; Table S5; c.659A > G MAF: 1.2×10^{-2}, and c.455C > T MAF: 4.5×10^{-3}). The exact *TP53* mutation that defined each of these

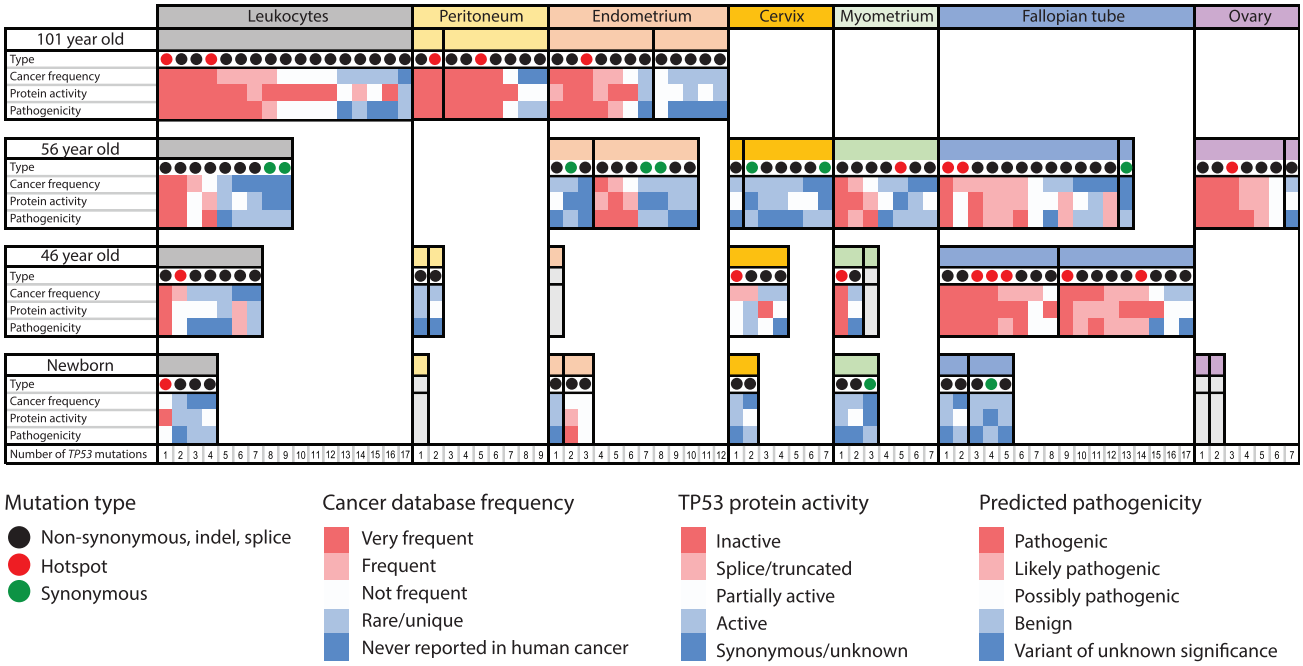


Figure 6. Characterization of *TP53* Mutations in Normal Tissues over a Century of Human Lifespan

TP53 mutations identified by DS in leukocytes and gynecological tissues are indicated as columns within each tissue. Each mutation is characterized by four parameters: type (synonymous, non-synonymous, and hotspots), frequency in cancer, protein activity, and predicted pathogenicity. The last three parameters are color-coded, with red indicating "cancer-like" mutations and blue indicating benign mutations. Vertical black lines separate different biopsies from the same tissue. Within each biopsy, mutations are ordered left to right by decreasing cancer frequency. The biopsies that were sequenced but no mutations were identified are shown in gray. Note that the female newborn in the study had no blood collected. The leukocytes included here correspond to a similarly aged newborn male.

clones was also detected at lower frequencies in peritoneal and endometrial samples from the same subject, revealing an apparent contribution of leukocyte DNA to those tissue samples (Figure S6).

In fact, this cross-tissue mutation sharing was common in the 101-year-old woman, suggesting that aged leukocytes may indeed harbor relatively large clones that recurrently contribute to the mutations found when sequencing other biopsies. Mutation sharing was less prevalent in middle-aged women. While very-low-frequency mutations are often hard to replicate due to the low precision of the measurement resulting purely from sampling statistics (not technical accuracy), it is important to keep in mind that certain mutations may also be recurrently identified simply because they are hotspots, and thus a common origin cannot automatically be assumed. For example, the hotspot mutation c.659A > G was identified in the large blood clone of the 101-year-old woman and in a myometrium sample and a fallopian tube biopsy of the 46-year-old woman (Figure S6). The processing of these particular samples was done on different days, making a cross-contamination explanation improbable.

As already considered, an important limitation of this study was the different depth of sequencing achieved across samples, due to the inherent variability in library preparations and differences in DNA availability. Because numerically more mutations will be identified in samples with more sequencing (Figure S8A), it is essential to compare samples based on their mutation frequency, which is a sequencing-normalized value calculated as the number of mutations divided by the number of total DS error-corrected nucleotides sequenced (Figure S8B). *TP53* mutation frequency tended to be higher at older ages in the three tissue types shared by the neonate and the centenarian (leukocytes, peritoneum, and endometrium), although there was substantial variability across samples.

TP53 Mutations in Newborn Tissue Are Random, yet Become Positively Selected over a Lifetime

As further illustration of the increase of cancer-like mutations with aging, we divided all *TP53* mutations into two binary categories: "common in cancer" and "not common in cancer," with the former being defined by those classified as "frequent" or "very frequent" in the UMD cancer database. When plotted by age, the progressive enrichment for cancer-like mutations was easily apparent, especially in certain tissues such as fallopian tubes and leukocytes (Figure 7A).

We then examined the five traits of selection previously calculated for the UL study, but for *TP53* mutations found in the newborn, middle-aged, and centenarian tissues (Figures 7B and S9). For mutations found in newborn tissue, all five traits yielded values consistent with random processes (e.g., absence of selection), yet in middle-aged and even more so in centenarian tissue, values reflected selection to an extent that neared that

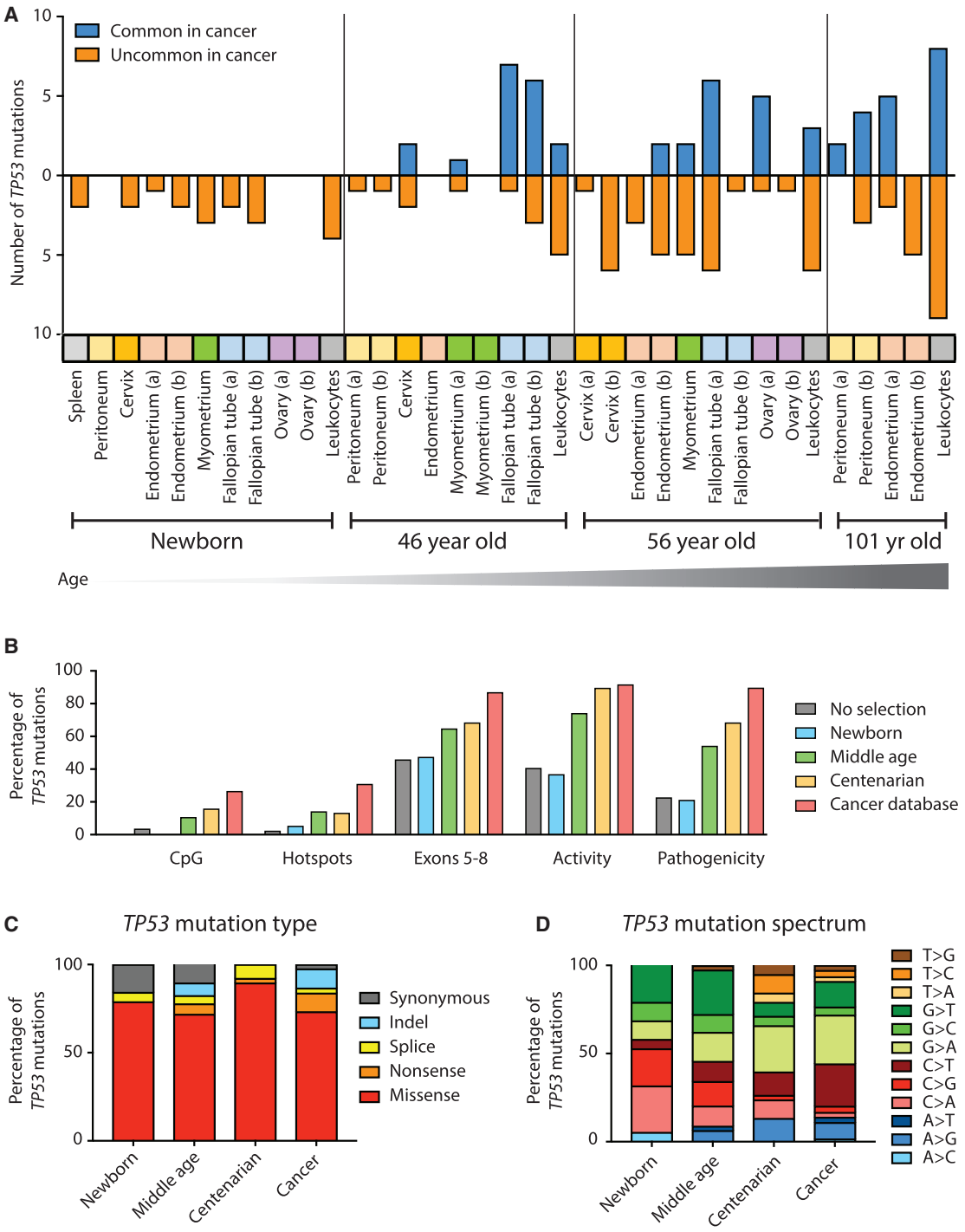


Figure 7. Cancer-Associated *TP53* Mutations Are Positively Selected during Normal Aging

(A) Across a variety of human tissues, *TP53* mutations accumulate with age and are progressively enriched for mutations commonly found in cancers. Tissues are color-coded. (a) and (b) indicate two biopsies from the same tissue.

(B) Traits of positive selection are compared between all of the possible mutations in the *TP53* coding region (n = 3,546); *TP53* mutations found in newborns (n = 19), middle-aged individuals (n = 85), and centenarian individuals (n = 38); and *TP53* mutations reported in the UMD cancer database (n = 71,051).

(C and D) Distribution of *TP53* mutation type (C) and mutation spectrum (D) for newborn, middle-aged, and centenarian mutations compared to the UMD cancer database.

seen in mutations from tumors in the UMD database. Analysis of mutation type (Figure 7C) revealed a decrease in synonymous mutations with age (in fact, no synonymous mutations were identified in centenarian tissue; Table S5).

Lastly, regarding the mutation spectrum, we noted an interesting preponderance of C and G mutations in newborn tissue, which progressively shifted toward an increased representation of A and T mutations in centenarian tissue, which is more similar to the pattern in cancers. The significance of this shift is unknown as it could represent both biases in the nucleotide composition of the gene at selectable hotspots and differential age-associated mutagenic processes (Alexandrov et al., 2015), which disproportionately contribute to the clonal mutation burden of tumors because tumors mostly arise in the elderly.

TP53 Mutations in cfDNA and Peritoneal Fluid Follow the Same Patterns as Solid Tissue

To explore the abundance of *TP53* mutations in liquid biopsies of clinical interest, we analyzed plasma-derived cell-free DNA (cfDNA) and peritoneal fluid from the 46-year-old woman. *TP53* mutations were identified in both, with cancer-like features similar to what was observed for solid tissue biopsies, UL, and leukocytes (Figure S10). None of the mutations identified in cfDNA or peritoneal fluid overlapped with mutations detected in other samples from the same woman (Figure S6). Peritoneal fluid is routinely collected for disease staging during gynecologic surgery, and we previously demonstrated that it carries *TP53* background mutations with cancer-like features (Krimmel et al., 2016), in agreement with our findings here. cfDNA had not been analyzed previously by DS. The fact that we identified one pathogenic mutation commonly found in cancers within cfDNA from a healthy woman (Figure S10; Table S5) raises important concerns over specificity in cancer-screening studies based on mutation detection in plasma.

DISCUSSION

We have demonstrated that DS can detect *TP53* ovarian cancer mutations in UL, providing proof-of-principle for an innovative approach with the potential for ovarian cancer detection. Our combined approach of UL plus DS improves upon past mutation-based screening efforts through the use of a collection method that recovers cancer cells very close to the anatomical site of the tumor and an ultra-accurate DNA sequencing technology that can resolve exceptionally low-frequency mutations. We were able to achieve remarkable sensitivity and specificity without prior knowledge of the tumor mutation and using a MAF threshold for differentiating cancers from controls. The study, however, was small and patients had advanced ovarian cancer, important limitations that will need to be addressed with much larger prospective trials. These trials will be facilitated by the fact that UL is minimally invasive and can be practically integrated into routine gynecologic primary care (Maritschnegg et al., 2018).

However, the most profound finding of this work is not the biomarker performance itself, but the incidentally found mutational patterns that reflect a somatic evolutionary process that appears operative throughout much of human life in normal tissues. Specifically, we identified widespread low-frequency *TP53* mutations that were heavily enriched for pathogenic variants. This enrichment reflects a process of natural selection that favors the survival and proliferation of cells with mutations that are identical to those observed in cancer, but as part of routine aging. The unambiguous selection signature is supported by multiple orthogonal metrics and cannot be explained by technical errors; both the biological and diagnostic implications are substantial.

One of the main reasons that cancer biomarkers fail to reach the clinic is their inability to achieve the extremely high specificity that is required for screening (Diamandis, 2012). This is critical for cancers with low incidence and that require an invasive procedure to follow up positive screening tests, such as the case with ovarian cancer (Drescher and Anderson, 2018). Harms due to false-positives and a lack of proven reduction in mortality are the reasons for the recent recommendation against the use of CA-125 and transvaginal sonography for screening asymptomatic women (Grossman et al., 2018). In recent years, mutation-based cancer screening from plasma or other body fluids has emerged as a promising method to detect cancer based on the supposition that cancer-associated mutations found in liquid biopsies are a specific indication of cancer somewhere in the body (Aravanis et al., 2017). Here, we demonstrate that cancer-associated mutations can be found in most normal tissues and, therefore, they are not cancer specific.

The detection of cancer-associated mutations in normal tissues is not entirely new (Risques and Kennedy, 2018). In 2014, 3 groups reported acute myeloid leukemia mutations found as minority subclones in the blood of ~10% of healthy elderly individuals—a phenomena dubbed clonal hematopoiesis of indeterminate potential (CHIP) (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). One year later, Martincorena et al. (2015) observed hundreds of tiny clones carrying cancer-associated driver mutations on sun-exposed eyelids, a finding recently replicated in normal aged esophagus (Martincorena et al., 2018). Cancer-associated mutations have been similarly reported in abnormal but non-cancerous tissues, including endometriosis (Anglesio et al., 2017; Suda et al., 2018) and benign dermal nevi (Shain et al., 2015). The use of laser capture microdissection in recent studies has revealed that as many as 1% of normal colorectal crypts of middle-aged individuals (Lee-Six et al., 2018) and >50% of normal endometrial glands of middle-aged women (Suda et al., 2018) carry mutations in cancer driver genes. A broader body of work related to quantifying the accumulation of neutral mutations in aging normal tissues (Abyzov et al., 2017; Welch et al., 2012; Yadav et al., 2016) has afforded important insights into the mechanisms of age-associated mutagenic processes.

The relation between mutations and cancer has been known for decades, yet the delay in appreciating their presence outside cancer can be largely attributed to technical limitations. The advent of NGS enabled surveying wide swaths of the genome and detection of mutations present clonally or as modest-size subclones. In the above studies, standard whole exome or multi-gene NGS was able to identify driver mutations because of unique scenarios in which clones were either relatively large

(CHIP in a subset of very elderly individuals) (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014) or spatially coherent and comprising a sizeable percentage of cells when very small biopsies were taken (Lee-Six et al., 2018; Martincorena et al., 2015, 2018; Suda et al., 2018). With higher-accuracy NGS techniques able to resolve lower-frequency subclones, later studies have found that CHIP mutations are abundant in middle-aged adults (Acuna-Hidalgo et al., 2017; Young et al., 2016). Using ultra-accurate DS, we found extremely low-frequency cancer-associated *TP53* mutations in both the blood and peritoneal fluid of women without cancer, and, in both sample types, the abundance of mutations increased with age (Krimmel et al., 2016). A subsequent study that used UL for endometrial cancer detection found pathogenic mutations in cancer driver genes in lavages of cases as well as controls (Nair et al., 2016), in agreement with the data reported here.

In addition to UL, we examined 10 different tissue or sample types from a unique cohort of individuals spanning more than a century of human lifespan and assessed the pattern of mutations found using multiple different metrics of selection. A significant finding was that not only do *TP53* mutations increase in abundance with age but also the relative representation of random mutations versus cancer-associated mutations transitioned from almost entirely the former to almost entirely the latter from birth to the end of life. Moreover, the extent of mutation frequency varied considerably by sample type and tissue. Our research complements recent efforts to characterize the accumulation of random, unselected somatic mutations with aging (Blokzijl et al., 2016; Hoang et al., 2016) and suggests that these findings are likely only the tip of the iceberg. Further work is important to expand beyond the main limitations of this study, which include the analysis of mostly gynecological tissues from only four individuals at a coarse spacing along the aging continuum; imperfect matching of sample types for each subject due to the inherent challenges of tissue acquisition at the extremes of age; and focus on a single gene, albeit the one most commonly mutated in cancer.

The implications of our findings are important as a cautionary message for mutation-based cancer biomarkers. At the same time as we have shown that highly sensitive NGS methods are essential for maximal mutation detection, we have also illustrated a substantial specificity challenge related to biology, not technology, the extent of which has been under-appreciated. This is not limited to one or a few tissues; rather, it seems to be ubiquitous among the epithelial, mesenchymal, and hematopoietic cell lineages that we investigated. Moreover, cancer-associated mutations were found in liquid biopsies, including UL and cfDNA. This suggests that ongoing large-scale efforts to develop universal liquid biopsy cancer screening tests via deep sequencing of cfDNA need to be approached with great caution (Alix-Panabières and Pantel, 2016; Aravanis et al., 2017).

Despite biological background mutations and with the caveat of this being a quite small study from a biomarker perspective, our approach worked remarkably well as a minimally invasive cancer test. We identified 80% of tumor mutations, and 70% of those were above the 1% MAF threshold we used to distinguish cases from controls. The only tumor mutation missed below this threshold corresponded to a 42-year-old woman, one of the youngest in the study. Younger women tended to have fewer background mutations and mutations with lower MAFs, which suggest that, moving forward, sensitivity could probably be increased by using age-adjusted thresholds. In addition, specificity could likely be improved by uniform lavage collection at the luteal phase in premenopausal women, sequencing of peripheral blood to identify and exclude CHIP clones that may be present in lavage, and longitudinal assessment of mutations to identify MAF increases over time. Of note, we have demonstrated detection of intermediate- and late-stage cancers, but the most critical targets for screening are early-stage cancers because they are the most curable. In that regard, monitoring of high-risk populations, such as *BRCA1* and *BRCA2* carriers, may be the highest-impact near-term clinical implementation.

The sensitivity improvements lent by new sequencing technologies are forcing a far more nuanced genetic definition of what distinguishes a cancer cell from simply an old cell. Our results show that CHIP clones are merely one relatively easy to detect manifestation of a far broader phenomena that appears to extend to most, if not all, tissues in the body. From a biomarker perspective, the fact that those who are at greatest risk of cancer and for whom cancer screening holds the most benefit (older adults) are also the population with the most cancer-like age-associated background mutations is particularly inconvenient. Ongoing improvements will be needed to find ways to maximize specificity through careful MAF threshold calibration and combination with orthogonal biomarkers. Further investigation into the significance of biological background mutations from the perspective of human aging and biology is similarly warranted. While the notion that our somatic genomes are steadily evolving toward neoplasia with each passing decade may viewed as disheartening, an alternative perspective is that despite this, most people do not develop overt cancer in their lifetime. This serves as a reminder of just how much remains unknown about the body's many complex mechanisms of tumor suppression, a toolkit that we can perhaps augment with future technologies for cancer prevention.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAIL
  - Uterine Lavage
  - Normal Tissue
- METHOD DETAILS
  - Digital Droplet Polymerase Chain Reaction
  - Duplex Sequencing
  - Characterization of *TP53* mutations using Seshat and the UMD *TP53* database
  - *TP53* cancer database mutational analysis
  - *TP53* mutations without selection
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY
  - Code availability
  - Data availability

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2019.05.109.

ACKNOWLEDGMENTS

This work was supported in part by NIH grant R01CA181308 to R.A.R.; Mary Kay Foundation grant 045-15 to R.A.R.; Rivkin Center for Ovarian Cancer grant 567612 to R.A.R.; T32CA009515 to J.J.S.; R44CA221426 to J.J.S. and L.N.W.; CA077852 and CA193649 to L.A.L.; and Radiumhemmets Forskningsfonder 174261 to T.S. We thank members of the Loeb, Risques, Kennedy, and Swisher labs at the University of Washington for helpful discussions, as well as participating members of the LUSTIC and LUDOC clinical trials. Most important, we deeply thank the patients and families who volunteered to provide clinical samples, without which this research would not have been possible.

AUTHOR CONTRIBUTIONS

J..J.S., R.Z., P.S., E.M., and R.A.R. designed the study; E.M., E.L., R.H., and A.V. procured the samples; J.J.S., K.L.-S., E.M., J.E.H., M.T., and L.N.W. processed the samples; J.J.S., C.C.V., D.N., K.T.B., T.S., and R.A.R. contributed to the data analysis and visualization; M.J.E. and R.A.R. performed the statistical analyses; L.A.L., R.Z., and P.S. contributed expertise and invaluable critical discussion; J.J.S. and R.A.R. wrote the article.

DECLARATION OF INTERESTS

J.J.S. and L.A.L. are founders and equity holders at TwinStrand Biosciences. J.J.S., C.C.V., L.N.W., and J.E.H. are employees and equity holders at TwinStrand Biosciences. P.S. is a founder and equity holder in Ovartec. R.Z. is a founder and equity holder in Oncolab GmbH. R.A.R. shares equity in NanoString Technologies and is the principal investigator in an NIH Small Business Innovation Research (SBIR) subcontract research agreement with TwinStrand Biosciences. Commercial uses of Duplex Sequencing are protected by multiple patents held or licensed by the University of Washington and TwinStrand Biosciences. Commercial uses of uterine lavage for cancer screening and diagnosis are protected by multiple patents licensed or held by the Medical University of Vienna and Ovartec.

Received: December 14, 2018
Revised: April 6, 2019
Accepted: May 29, 2019
Published: July 2, 2019

REFERENCES

Abyzov, A., Tomasini, L., Zhou, B., Vasmatzis, N., Coppola, G., Amenduni, M., Pattni, R., Wilson, M., Gerstein, M., Weissman, S., et al. (2017). One thousand somatic SNVs per skin fibroblast cell set baseline of mosaic mutational load with patterns that suggest proliferative origin. Genome Res. *27*, 512–523.

Acuna-Hidalgo, R., Sengul, H., Steehouwer, M., van de Vorst, M., Vermeulen, S.H., Kiemeney, L.A.L.M., Veltman, J.A., Gilissen, C., and Hoischen, A. (2017). Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. Am. J. Hum. Genet. *101*, 50–64.

Ahmed, A.A., Etemadmoghadam, D., Temple, J., Lynch, A.G., Riad, M., Sharma, R., Stewart, C., Fereday, S., Caldas, C., Defazio, A., et al. (2010). Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. J. Pathol. *221*, 49–56.

Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S., and Stratton, M.R. (2015). Clock-like mutational processes in human somatic cells. Nat. Genet. *47*, 1402–1407.

Alix-Panabières, C., and Pantel, K. (2016). Clinical Applications of Circulating Tumor Cells and Circulating Tumor DNA as Liquid Biopsy. Cancer Discov. *6*, 479–491.

Anglesio, M.S., Papadopoulos, N., Ayhan, A., Nazeran, T.M., Noë, M., Horlings, H.M., Lum, A., Jones, S., Senz, J., Seckin, T., et al. (2017). Cancer-Associated Mutations in Endometriosis without Cancer. N. Engl. J. Med. *376*, 1835–1848.

Aravanis, A.M., Lee, M., and Klausner, R.D. (2017). Next-Generation Sequencing of Circulating Tumor DNA for Early Cancer Detection. Cell *168*, 571–574.

Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P., et al. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. Nature *538*, 260–264.

Bowtell, D.D., Böhm, S., Ahmed, A.A., Aspuria, P.J., Bast, R.C., Jr., Beral, V., Berek, J.S., Birrer, M.J., Blagden, S., Bookman, M.A., et al. (2015). Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. Nat. Rev. Cancer *15*, 668–679.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. *68*, 394–424.

Chien, J., Sicotte, H., Fan, J.B., Humphray, S., Cunningham, J.M., Kalli, K.R., Oberg, A.L., Hart, S.N., Li, Y., Davila, J.I., et al. (2015). TP53 mutations, tetraploidy and homologous recombination repair defects in early stage high-grade serous ovarian cancer. Nucleic Acids Res. *43*, 6945–6958.

Diamandis, E.P. (2012). The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem? BMC Med. *10*, 87.

Diaz, L.A., Jr., and Bardelli, A. (2014). Liquid biopsies: genotyping circulating tumor DNA. J. Clin. Oncol. *32*, 579–586.

Drescher, C.W., and Anderson, G.L. (2018). The Yet Unrealized Promise of Ovarian Cancer Screening. JAMA Oncol. *4*, 456–457.

Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. *371*, 2477–2487.

Grossman, D.C., Curry, S.J., Owens, D.K., Barry, M.J., Davidson, K.W., Doubeni, C.A., Epling, J.W., Jr., Kemper, A.R., Krist, A.H., Kurth, A.E., et al.; US Preventive Services Task Force (2018). Screening for Ovarian Cancer: US Preventive Services Task Force Recommendation Statement. JAMA *319*, 588–594.

Henderson, J.T., Webber, E.M., and Sawaya, G.F. (2018). Screening for Ovarian Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. JAMA *319*, 595–606.

Hoang, M.L., Kinde, I., Tomasetti, C., McMahon, K.W., Rosenquist, T.A., Grollman, A.P., Kinzler, K.W., Vogelstein, B., and Papadopoulos, N. (2016). Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. Proc. Natl. Acad. Sci. USA *113*, 9846–9851.

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. *371*, 2488–2498.

Kato, S., Han, S.Y., Liu, W., Otsuka, K., Shibata, H., Kanamaru, R., and Ishioka, C. (2003). Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. Proc. Natl. Acad. Sci. USA *100*, 8424–8429.

Kennedy, S.R., Schmitt, M.W., Fox, E.J., Kohrn, B.F., Salk, J.J., Ahn, E.H., Prindle, M.J., Kuong, K.J., Shen, J.C., Risques, R.A., and Loeb, L.A. (2014). Detecting ultralow-frequency mutations by Duplex Sequencing. Nat. Protoc. *9*, 2586–2606.

Kinde, I., Bettegowda, C., Wang, Y., Wu, J., Agrawal, N., Shih, IeM., Kurman, R., Dao, F., Levine, D.A., Giuntoli, R., et al. (2013). Evaluation of DNA from the

Papanicolaou test to detect ovarian and endometrial cancers. Sci. Transl. Med. *5*, 167ra4.

Krimmel, J.D., Schmitt, M.W., Harrell, M.I., Agnew, K.J., Kennedy, S.R., Emond, M.J., Loeb, L.A., Swisher, E.M., and Risques, R.A. (2016). Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. Proc. Natl. Acad. Sci. USA *113*, 6005–6010.

Kuhn, E., Kurman, R.J., Vang, R., Sehdev, A.S., Han, G., Soslow, R., Wang, T.L., and Shih, IeM. (2012). TP53 mutations in serous tubal intraepithelial carcinoma and concurrent pelvic high-grade serous carcinoma–evidence supporting the clonal relationship of the two lesions. J. Pathol. *226*, 421–426.

Kurman, R.J., and Shih, IeM. (2010). The origin and pathogenesis of epithelial ovarian cancer: a proposed unifying theory. Am. J. Surg. Pathol. *34*, 433–443.

Labidi-Galy, S.I., Papp, E., Hallberg, D., Niknafs, N., Adleff, V., Noe, M., Bhattacharya, R., Novak, M., Jones, S., Phallen, J., et al. (2017). High grade serous ovarian carcinomas originate in the fallopian tube. Nat. Commun. *8*, 1093.

Lee-Six, H., Ellis, P., Osborne, R.J., Sanders, M.A., Moore, L., Georgakopoulos, N., Torrente, F., Noorani, A., Goddard, M., Robinson, P., et al. (2018). The landscape of somatic mutation in normal colorectal epithelial cells. bioRxiv. https://doi.org/10.1101/416800.

Leroy, B., Anderson, M., and Soussi, T. (2014). TP53 mutations in human cancer: database reassessment and prospects for the next decade. Hum. Mutat. *35*, 672–688.

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum. Mutat. *37*, 235–241.

Maritschnegg, E., Wang, Y., Pecha, N., Horvat, R., Van Nieuwenhuysen, E., Vergote, I., Heitz, F., Sehouli, J., Kinde, I., Diaz, L.A., Jr., et al. (2015). Lavage of the Uterine Cavity for Molecular Detection of Müllerian Duct Carcinomas: A Proof-of-Concept Study. J. Clin. Oncol. *33*, 4293–4300.

Maritschnegg, E., Heitz, F., Pecha, N., Bouda, J., Trillsch, F., Grimm, C., Vanderstichele, A., Agreiter, C., Harter, P., Obermayr, E., et al. (2018). Uterine and Tubal Lavage for Earlier Cancer Detection Using an Innovative Catheter: A Feasibility and Safety Study. Int. J. Gynecol. Cancer *28*, 1692–1698.

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D.C., Fullam, A., Alexandrov, L.B., Tubio, J.M., et al. (2015). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science *348*, 880–886.

Martincorena, I., Fowler, J.C., Wabik, A., Lawson, A.R.J., Abascal, F., Hall, M.W.J., Cagan, A., Murai, K., Mahbubani, K., Stratton, M.R., et al. (2018). Somatic mutant clones colonize the human esophagus with age. Science *362*, 911–917.

Nair, N., Camacho-Vanegas, O., Rykunov, D., Dashkoff, M., Camacho, S.C., Schumacher, C.A., Irish, J.C., Harkins, T.T., Freeman, E., Garcia, I., et al. (2016). Genomic Analysis of Uterine Lavage Fluid Detects Early Endometrial Cancers and Reveals a Prevalent Landscape of Driver Mutations in Women without Histopathologic Evidence of Cancer: A Prospective Cross-Sectional Study. PLoS Med. *13*, e1002206.

Risques, R.A., and Kennedy, S.R. (2018). Aging and the rise of somatic cancer-associated mutations in normal tissues. PLoS Genet. *14*, e1007108.

Salk, J.J., Schmitt, M.W., and Loeb, L.A. (2018). Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat. Rev. Genet. *19*, 269–285.

Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B., and Loeb, L.A. (2012). Detection of ultra-rare mutations by next-generation sequencing. Proc. Natl. Acad. Sci. USA *109*, 14508–14513.

Schmitt, M.W., Fox, E.J., Prindle, M.J., Reid-Bayliss, K.S., True, L.D., Radich, J.P., and Loeb, L.A. (2015). Sequencing small genomic targets with high efficiency and extreme accuracy. Nat. Methods *12*, 423–425.

Shain, A.H., Yeh, I., Kovalyshyn, I., Sriharan, A., Talevich, E., Gagnon, A., Dummer, R., North, J., Pincus, L., Ruben, B., et al. (2015). The Genetic Evolution of Melanoma from Precursor Lesions. N. Engl. J. Med. *373*, 1926–1936.

Siegel, R.L., Miller, K.D., and Jemal, A. (2017). Cancer Statistics, 2017. CA Cancer J. Clin. *67*, 7–30.

Soong, T.R., Howitt, B.E., Horowitz, N., Nucci, M.R., and Crum, C.P. (2019). The fallopian tube, "precursor escape" and narrowing the knowledge gap to the origins of high-grade serous carcinoma. Gynecol. Oncol. *152*, 426–433.

Suda, K., Nakaoka, H., Yoshihara, K., Ishiguro, T., Tamura, R., Mori, Y., Yamawaki, K., Adachi, S., Takahashi, T., Kase, H., et al. (2018). Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. Cell Rep. *24*, 1777–1789.

The Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609–615.

Tikkanen, T., Leroy, B., Fournier, J.L., Risques, R.A., Malcikova, J., and Soussi, T. (2018). Seshat: A Web service for accurate annotation, validation, and analysis of TP53 variants generated by conventional and next-generation sequencing. Hum. Mutat. *39*, 925–933.

Vang, R., Levine, D.A., Soslow, R.A., Zaloudek, C., Shih, IeM., and Kurman, R.J. (2016). Molecular Alterations of TP53 are a Defining Feature of Ovarian High-Grade Serous Carcinoma: A Rereview of Cases Lacking TP53 Mutations in The Cancer Genome Atlas Ovarian Study. Int. J. Gynecol. Pathol. *35*, 48–55.

Wang, Y., Li, L., Douville, C., Cohen, J.D., Yen, T.T., Kinde, I., Sundfelt, K., Kjær, S.K., Hruban, R.H., Shih, I.M., et al. (2018). Evaluation of liquid from the Papanicolaou test and other liquid biopsies for the detection of endometrial and ovarian cancers. Sci. Transl. Med. *10*, eaap8793.

Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., et al. (2012). The origin and evolution of mutations in acute myeloid leukemia. Cell *150*, 264–278.

Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, M.C., McMichael, J.F., Schmidt, H.K., Yellapantula, V., Miller, C.A., et al. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat. Med. *20*, 1472–1478.

Yadav, V.K., DeGregori, J., and De, S. (2016). The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection. Nucleic Acids Res. *44*, 2075–2084.

Young, A.L., Challen, G.A., Birmann, B.M., and Druley, T.E. (2016). Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. Nat. Commun. *7*, 12484.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Uterine lavage	Medical University of Vienna (Austria), Charles University Pilsen (Czech Republic) and University Hospitals Leuven (Belgium).	N/A
Newborn tissue	Autopsies from Seattle Childrens' Hospital	N/A
Middle age tissue	Hysterectomies from Medical University of Vienna (Austria)	N/A
Centenarian tissue	Autopsy from Tissue for Research Inc.	N/A
Critical Commercial Assays		
Custom TaqMan SNP Genotyping Assays for ddPCR	Life Technologies	No.4331349
Droplet Digital PCR system	BioRad	QX200
Massively parallel DNA sequencer	Illumina	HiSeq 2500
Massively parallel DNA sequencer	Illumina	NextSeq 550
KAPA HyperPrep library Kit	Roche Sequencing	KK8505
Deposited Data		
TP53 Duplex Sequencing data	This article	PRJNA503496
Oligonucleotides		
Primers for ddPCR	Maritschnegg et al., 2015	N/A
Duplex Sequencing primers and adapters	Kennedy et al., 2014	N/A
Software and Algorithms		
Duplex Sequencing Analysis	Kennedy et al., 2014	https://github.com/loeblab/Duplex-Sequencing
Seshat	Tikkanen et al. 2018	https://p53.fr/tp53-database/seshat
IGV	Broad Institute	https://software.broadinstitute.org/software/igv/
Other		
UMD TP53 database	*TP53* website	https://p53.fr/tp53-database

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Rosana Risques (rrisques@uw.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAIL

We performed two complementary studies. The objective of the uterine lavage study was to determine the ability of DS to detect ovarian cancer through deep sequencing of *TP53* mutations in uterine lavage. This study included 10 patients with high grade serous ovarian cancer (cases) and 11 with benign gynecological masses (controls). The objective of the normal tissue study was to characterize somatic *TP53* mutations that accumulate during aging. This study included tissue from two newborn subjects (one newborn male only provided blood), two middle age women (ages 46 and 56 year-old) and one centenarian woman (101 year-old). Clinico-pathological information for all subjects is listed in Table S1.

Uterine Lavage

In the first study, we analyzed uterine lavages collected by a trans-cervical catheter from women undergoing procedures for suspected gynecological malignancies (Table S1). None of the women had endometriosis. Lavages were collected immediately

pre-operatively as previously described (Maritschnegg et al., 2015). Lavage samples were centrifuged at 300x g for 10 minutes at room temperature and DNA was isolated from the cell pellet (QIAamp MinElute Kit, QIAGEN, Hilden, Germany). Patients were recruited in three institutions: Medical University of Vienna (Austria), Charles University Pilsen (Czech Republic) and University Hospitals Leuven (Belgium). Sample procurement was performed in accordance with the institutional review boards of the Medical University of Vienna (EK#1148/2011 and EK#1766/2013), the Catholic University Leuven (B322201214864/S54406) and the Medical Faculty Hospital Pilsen (No 502/2013).

Normal Tissue

In the second study, multiple gynecological tissues were collected per Table S4. Not all sample types were available for all subjects. Newborn and centenarian tissue was collected at autopsy, while tissue from middle age women was collected following hysterectomy. Peripheral blood was unavailable from the female newborn, so we sequenced peripheral blood from a similar aged neonatal male. No other tissue was collected from the newborn male other than blood. The two newborn autopsies were performed at Seattle Childrens' Hospital and tissues were collected under research IRB #52304. The diagnoses were vein of Galen malformation (newborn female) and bronchopulmonary dysplasia (newborn male). The two middle age women had uterine leiomyoma and were operated on at the Medical University of Vienna. Samples were collected with informed consent and according to approved IRB EK# 1152/2014. Uterine lavage was collected with the same procedure as in the first study. For the 46 year-old woman, cfDNA was collected preoperatively and peritoneal lavage collected intraoperatively. For the centenarian woman, tissue was obtained via rapid autopsy from Tissue for Research Inc. and processed at the University of Washington under IRB waiver 2016-52304. All samples were collected using sterile new instruments between biopsies and frozen over liquid nitrogen immediately after collection and stored at −80°C until DNA extraction. To confirm normal histology, tissue biopsies immediately adjacent to the biopsy used for DNA extraction were embedded in OCT, sectioned, stained with H&E and reviewed by a pathologist (M.T.). In all samples, morphologic examination revealed only normal tissue, without inflammation, necrosis, hyperplasia or neoplasia.

METHOD DETAILS

Digital Droplet Polymerase Chain Reaction

Lavage DNA from 5 ovarian cancer cases and 2 controls was analyzed by ddPCR (Table S2). In ovarian cancer lavage, ddPCR amplified the tumor mutation whereas in benign lavage, the assay targeted two mutations previously identified by DS at frequencies below 0.1%. ddPCR was performed with the QX100 Droplet Digital PCR system (Bio-Rad Laboratories, Hercules, CA) using custom TaqMan SNP Genotyping Assays (Life Technologies, Carlsbard, CA) designed using Primer Express 3.0 software (ThermoFisher). 10-20ng of DNA were used in each reaction and samples were analyzed at least in duplicates. A positive control and a wild-type control were included in every run.

Duplex Sequencing

Duplex Sequencing was performed as previously described with minor modifications (Kennedy et al., 2014). Briefly for most samples, DNA was sonicated, end-repaired, A-tailed, and ligated with DS adapters using the KAPA HyperPrep library kit (Roche Sequencing, Pleasanton, CA). DNA from the two sets of normal tissues obtained from the middle-aged hysterectomy specimens were prepared with a prototype Duplex Sequencing kit (TwinStrand Biosciences, Seattle, WA). After initial amplification, 120 bp biotinylated oligonucleotide probes (Integrated DNA Technologies, Coralville, Iowa) were used to capture the coding region of *TP53*. Two successive rounds of captures were performed to ensure sufficient target enrichment, as previously described (Schmitt et al., 2015). Indexed libraries were pooled and sequenced on an Illumina HiSeq2500 or NextSeq500. Sequencing reads were aligned to hg19 then reads sharing a common molecular tag in both distinct strand orientations were grouped and assembled into an error-corrected Duplex Consensus Sequence as previously described (Kennedy et al., 2014).

The total number of Duplex nucleotides sequenced for each uterine lavage and tissue sample is listed in Tables S1 and S4. In aggregate, we sequenced 587,169,708 unique nucleotides, 319,576,913 of which corresponded to coding nucleotides. We targeted a median DS depth of ~1000x. Three tissue biopsies were excluded because of insufficient depth. Because Duplex reads correspond to original DNA molecules, DS depth reflects the total number of haploid genomes sequenced. For each sample, *TP53* mutation frequency was calculated as the number of identified mutations divided by the total number of Duplex nucleotides sequenced. For each individual mutation, mutant allele frequency (MAF) was calculated as number of mutated Duplex bases divided by the total DS depth at a given nucleotide position. Mutations identified as SNPs in the 1000 genome database were excluded from mutation analysis. All mutations were manually reviewed with the Integrative Genome Viewer (IGV).

Characterization of *TP53* mutations using Seshat and the UMD *TP53* database

The final list of mutations from all samples in the study (uterine lavage study n = 166, normal tissue study n = 264) was converted into a Variant Call Format (VCF) file and submitted to Seshat (https://p53.fr/TP53-database/seshat), a web service that performs *TP53* mutation annotation using data derived from the UMD *TP53* database (Tikkanen et al., 2018). This database is the most updated and comprehensive repository of *TP53* variants. From the Seshat output, the following variables were extracted: cDNA variant, HG19 Variant, Variant Classification, Frequency, Activity, Pathogenicity, Exon, Codon, CpG, Mutational Event and Variant Comment. These

variables were used to annotate the DS pipeline-generated mutational calls in Table S3 (uterine lavage) and Table S5 (normal tissue). The human genomic reference hg19 (GRCH37) was used for data reporting. Mutations occurring in the coding region and adjacent splice sites were selected for mutational analysis (uterine lavage n = 112, normal tissue n = 180).

Mutations were annotated based on type (missense, nonsense, splice, indel, synonymous), mutation spectrum (each of the 12 possible nucleotide substitutions), localization to CpG dinucleotides, localization in exons 5 to 8 (encoding the protein's DNA binding domain), localization to mutational hotspot (9 most common mutated codons in the UMD *TP53* database: 175, 179, 213, 220, 245, 248, 249, 273, 282), frequency of the mutation in the cancer database, functional activity, and predicted pathogenicity. Functional activity was assessed by a transcriptional activity chart assay for 3,000 variants performed by Kato et al. (2003). Pathogenicity was based on multiple predictive algorithms included from dbNSFP (Liu et al., 2016) as well as functional activity (Tikkanen et al., 2018). For the last 3 variables (frequency in cancer database, activity, and pathogenicity), mutations were aggregated into 5 categories and 2 categories. Categories of frequency in cancer database included very frequent, frequent, not frequent, rare/unique, and never identified in human cancer. The first 2 categories were considered "common in cancer" and the last 3 categories "not common in cancer." Categories of activity included inactive, splice/truncated, partially active, active, and synonymous/unknown. The first 3 categories were considered "impaired activity" and the last 2 categories "active/likely active." Categories of pathogenicity included pathogenic, likely pathogenic, possibly pathogenic, benign, and variants of uncertain significance. The first 3 categories were considered "expected pathogenic" and the last 2 categories "unlikely pathogenic."

TP53 cancer database mutational analysis

From the UMD *TP53* database (April 2017 version), we selected the set of 71,051 mutations reported within human tumors of all types (mutations from cell lines, normal, and premalignant tissue were excluded). Then we determined the distribution of mutations in the following categories: CpG, hotspots, exons 5-8, activity, pathogenicity, mutation type, and mutation spectrum. These values were used as a comparator for *TP53* mutations identified in uterine lavage and normal tissues.

TP53 mutations without selection

To assess the distribution of *TP53* mutations in the absence of selection, we generated a list of all possible mutations in the gene coding region (n = 3,546) *in silico*. Then we submitted this list to Seshat to determine the distribution of mutations in the same categories as above. The values obtained represent the distribution of all possible *TP53* mutations in the absence of selection and were used as a comparator for *TP53* mutations identified in uterine lavage and normal tissue.

QUANTIFICATION AND STATISTICAL ANALYSIS

Correlations were tested with Spearman's rank test due to high variability in the outcomes (non-normality). Comparisons by groups were performed with Fisher's exact tests (for comparison between case and control groups) or with the exact binomial test for comparison of one group with a hypothesized null probability of a success (mutation). Generalized estimating equations (GEE) also was investigated to take into account possible correlation within patients (lack of independence among observations/mutations with a patient) for these comparisons, but the results from the exact tests that assume independence were more conservative and thus, are the ones reported. Further, taking into account the sequencing depth either as an adjustment in binomial models or an offset in Poisson count models did not affect the significance of results. Sensitivity and specificity were calculated as simple proportions. Adjustment for age was performed by fitting log (maximum frequency) in a linear model. All tests were two-sided at an alpha level (type 1 error rate) of 0.05. Statistical analyses were performed with SPSS and R.

DATA AND CODE AVAILABILITY

Code availability

Software for DS data analysis is available at https://github.com/loeblab/Duplex-Sequencing

Data availability

The accession number for the sequencing data reported in this paper is Sequence Read Archive BioProject ID: PRJNA503496.

Cell Reports, Volume *28*

Supplemental Information

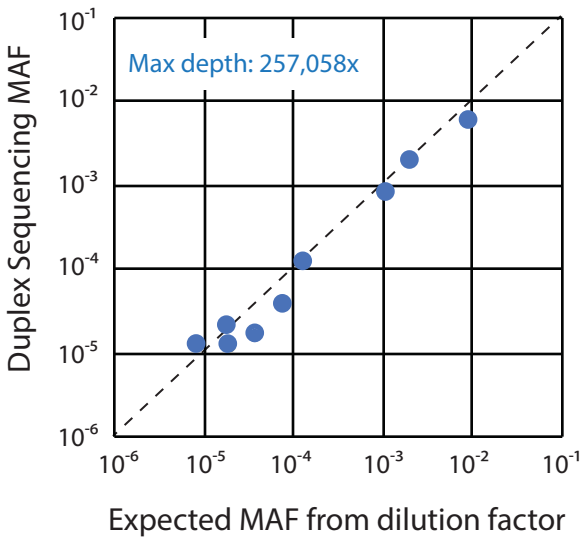
Ultra-Sensitive *TP53* Sequencing for Cancer Detection Reveals Progressive Clonal Selection in Normal Tissue over a Century of Human Lifespan

Jesse J. Salk, Kaitlyn Loubet-Senear, Elisabeth Maritschnegg, Charles C. Valentine, Lindsey N. Williams, Jacob E. Higgins, Reinhard Horvat, Adriaan Vanderstichele, Daniela Nachmanson, Kathryn T. Baker, Mary J. Emond, Emily Loter, Maria Tretiakova, Thierry Soussi, Lawrence A. Loeb, Robert Zeillinger, Paul Speiser, and Rosa Ana Risques

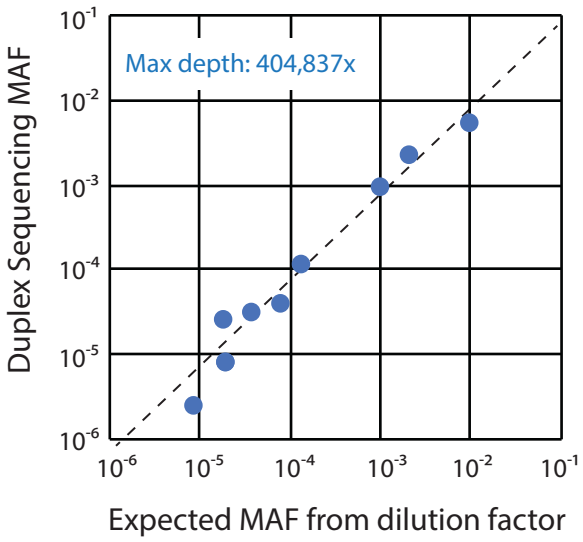
SUPPLEMENTAL FIGURES

Figure S1. Duplex Sequencing spike-in test of reproducibility and accuracy
Figure S2. Association between number of independent *TP53* mutations detected and total number of Duplex nucleotides sequenced
Figure S3. *TP53* mutation frequency and characteristics by age for individual patient lavages in case-control study
Figure S4. Comparison of traits of positive selection between *TP53* mutations in the UMD cancer database and uterine lavages.
Figure S5. *TP53* mutation frequency and characteristics by age including uterine lavages from the two middle age women in the normal tissue study
Figure S6. Analysis of mutations shared across multiple tissue samples within the same individual
Figure S7. Mutant allele frequency as a function of Duplex Sequencing depth
Figure S8. *TP53* mutation frequency by tissue type
Figure S9. *TP53* mutation characteristics by age for individual tissue samples
Figure S10. *TP53* mutation characteristics within non-invasively collected body fluids from a 46 year old woman

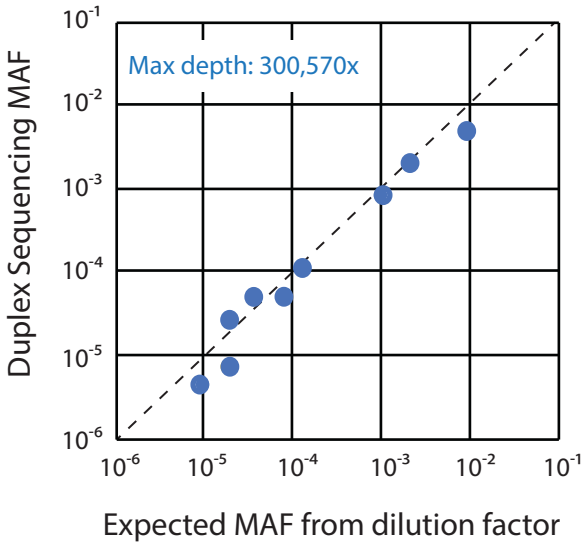
A.



B.



C.



D.

Gene	Mutation	Dilution
TP53	Q136fs	1:100
TP53	R249S	1:500
TP53	R248Q	1:1,000
NRAS	Q61R	1:5,000
TP53	G245S	1:10,000
NRAS	Q61L	1:25,000
TP53	R175H	1:50,000
KRAS	G12D (het.)	1:50,000
KRAS	G12V	1:100,000

Figure S1. Duplex Sequencing spike-in test of reproducibility and accuracy [related to Fig. 1B-E]. (A-C) DNAs from 9 cell lines, each carrying one or more unique single nucleotide mutation in *TP53* or a *RAS* gene, were serially diluted into DNA obtained from the peripheral blood of a healthy 26 year old male donor at levels ranging from 1:100-1:100,000. This single mixture was divided into three portions, with each prepared into Duplex libraries on three different days and sequenced on three independent runs. The total Duplex depth achieved approximately 1-million-fold (i.e. a million independent genome equivalents). All mutations were detected in all runs, with R^2 values of measured vs. expected MAFs respectively being 0.98, 0.96 and 0.95. Larger MAF variations among replicates at higher dilutions reflects the greater impact of stochastic Poisson sampling with lower frequency mutations. (D) Table of each spike-in mutation and the dilution factor used to generate the mixture. Expected MAF values for each data point in A-C were uniformly adjusted from the dilution factor listed in D by known copy number variation/zygosity at the loci of interest, as determined by Duplex Sequencing of pure DNA from each cell line.

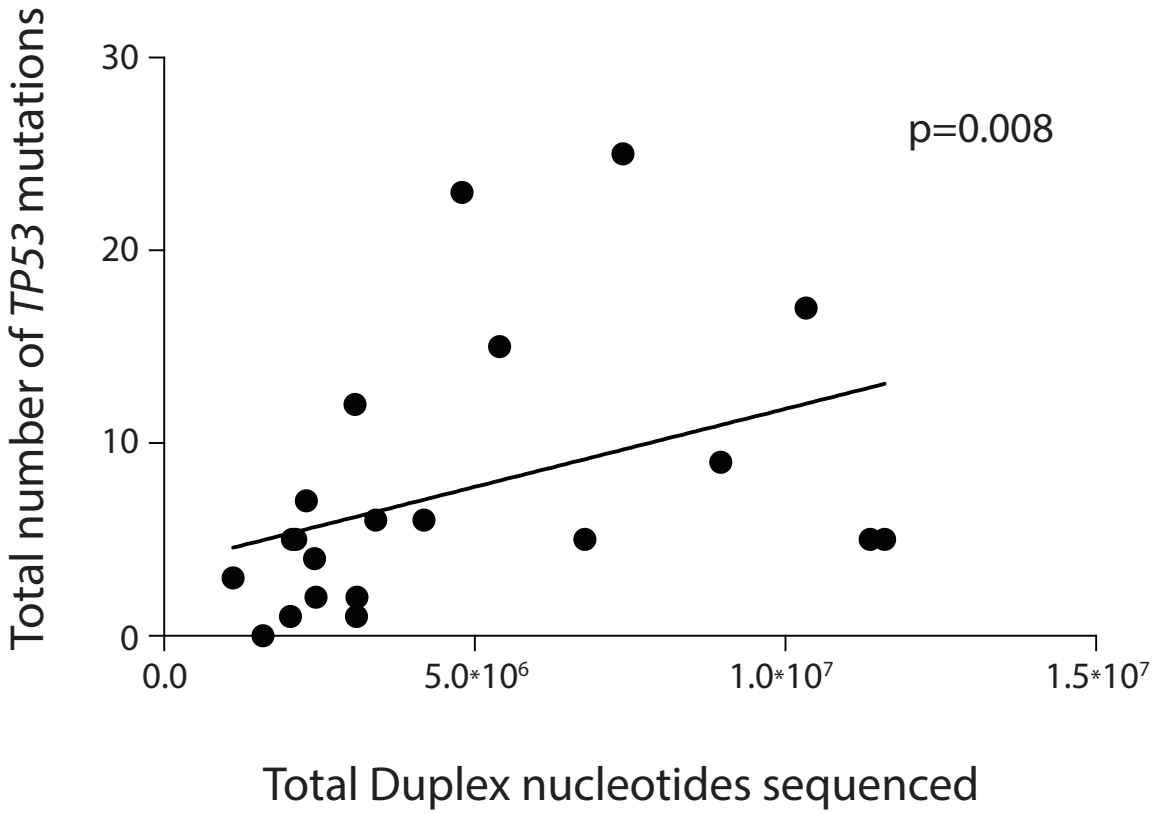


Figure S2. Association between number of independent *TP53* mutations detected and total number of Duplex nucleotides sequenced [related to Fig. 1E-F]. The total number of *TP53* mutations found (including exons and flanking intronic regions) in the 21 uterine lavages of the study was plotted against the total number of Duplex nucleotides sequenced in each sample. More *TP53* mutations were identified in samples with more nucleotides sequenced (p=0.008 by Spearman's correlation test).

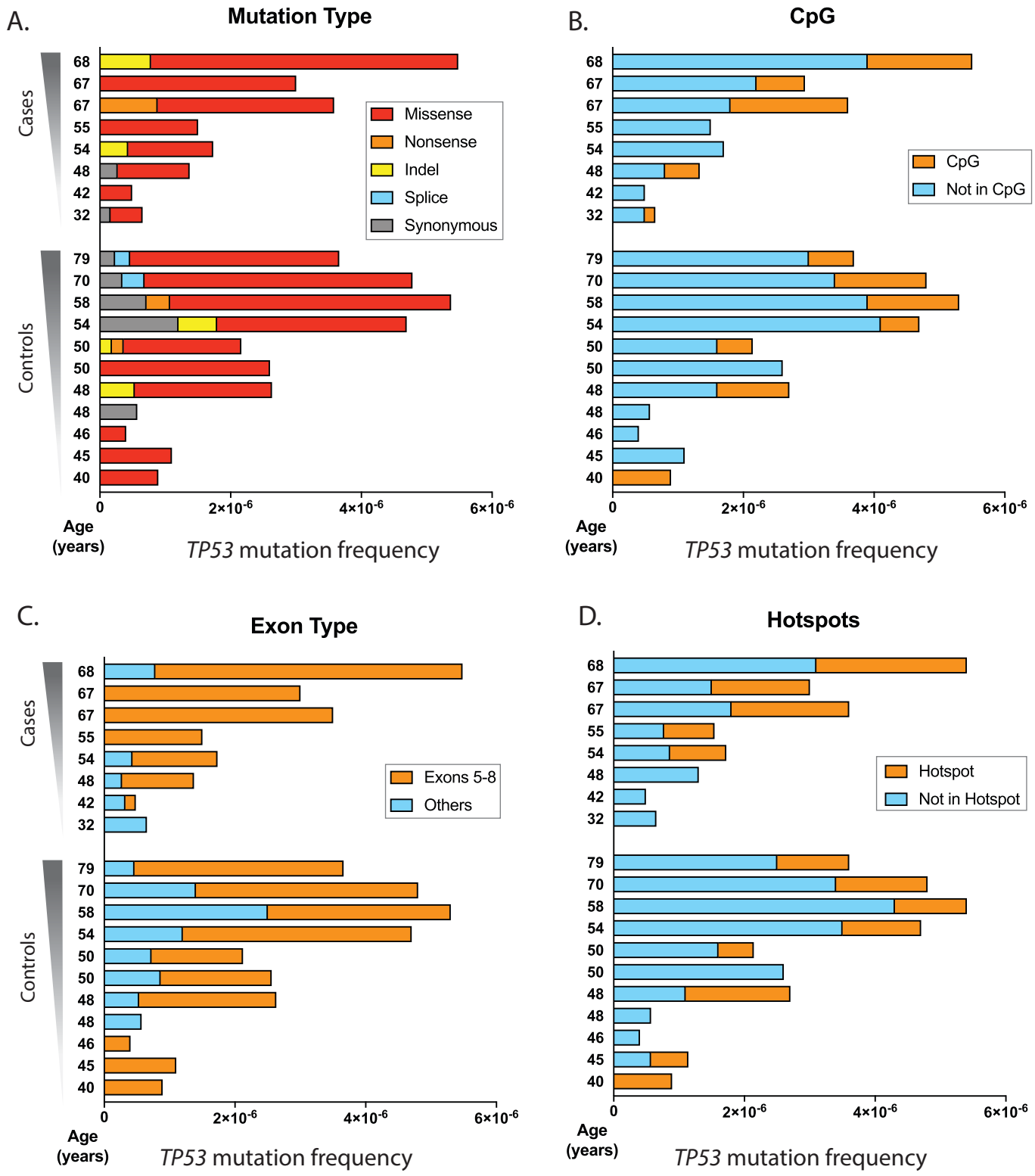


Figure S3. *TP53* mutation frequency and characteristics by age for individual patient lavages in case-control study [related to Fig. 3A-D]. Data is parsed by (A) mutation type, (B) CpG dinucleotide site, (C) exon type and (D) cancer-associated hotspots. Patients are divided in cases (women with ovarian cancer) and controls (women without ovarian cancer) and ordered by age within each group. For each patient, *TP53* mutation frequency was calculated as the number of *TP53* mutations identified in the coding region divided by the total number of Duplex nucleotides sequenced in that region. For each trait, the fraction of mutations corresponding to each of the categories of analysis is indicated by color.

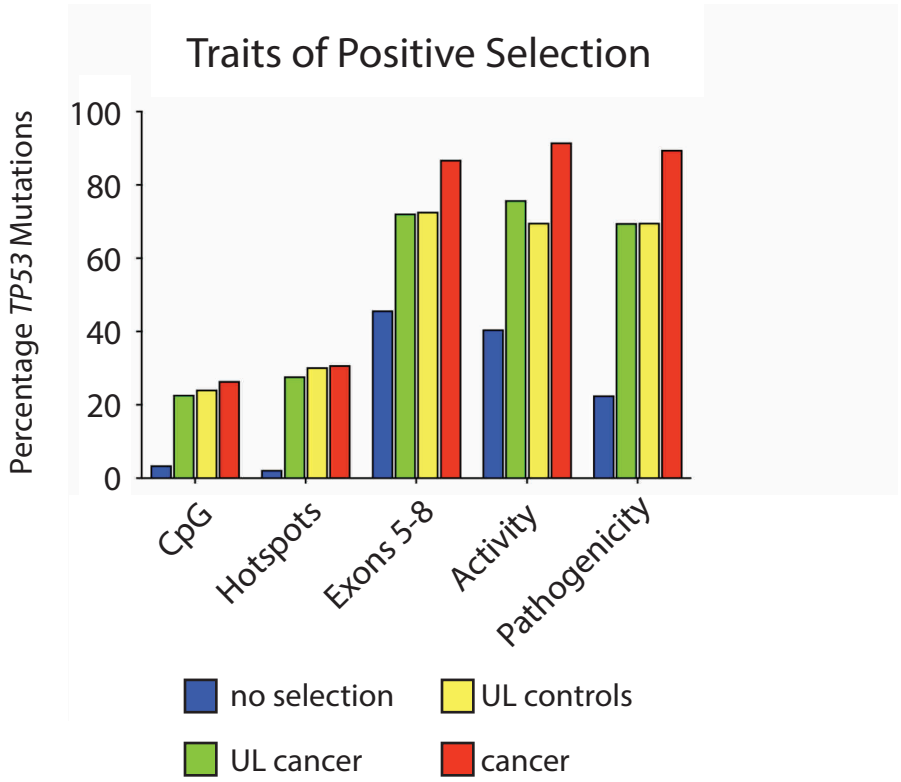


Figure S4. Comparison of traits of positive selection between *TP53* mutations in the UMD cancer database and uterine lavages [related to Fig. 4A]. For each trait, the percentage of observed *TP53* mutations is color-coded for each group. The 'no selection' group includes all possible mutations in the *TP53* coding region (n=3,546). *TP53* background mutations found in uterine lavage from women without ovarian cancer (controls, n= 79) and uterine lavage from women with ovarian cancer (cases, n=33) show similar distribution of mutational traits, which more closely resemble mutations found in cancers (n=71,051) than mutations expected in the absence of selection. UL: Uterine lavage.

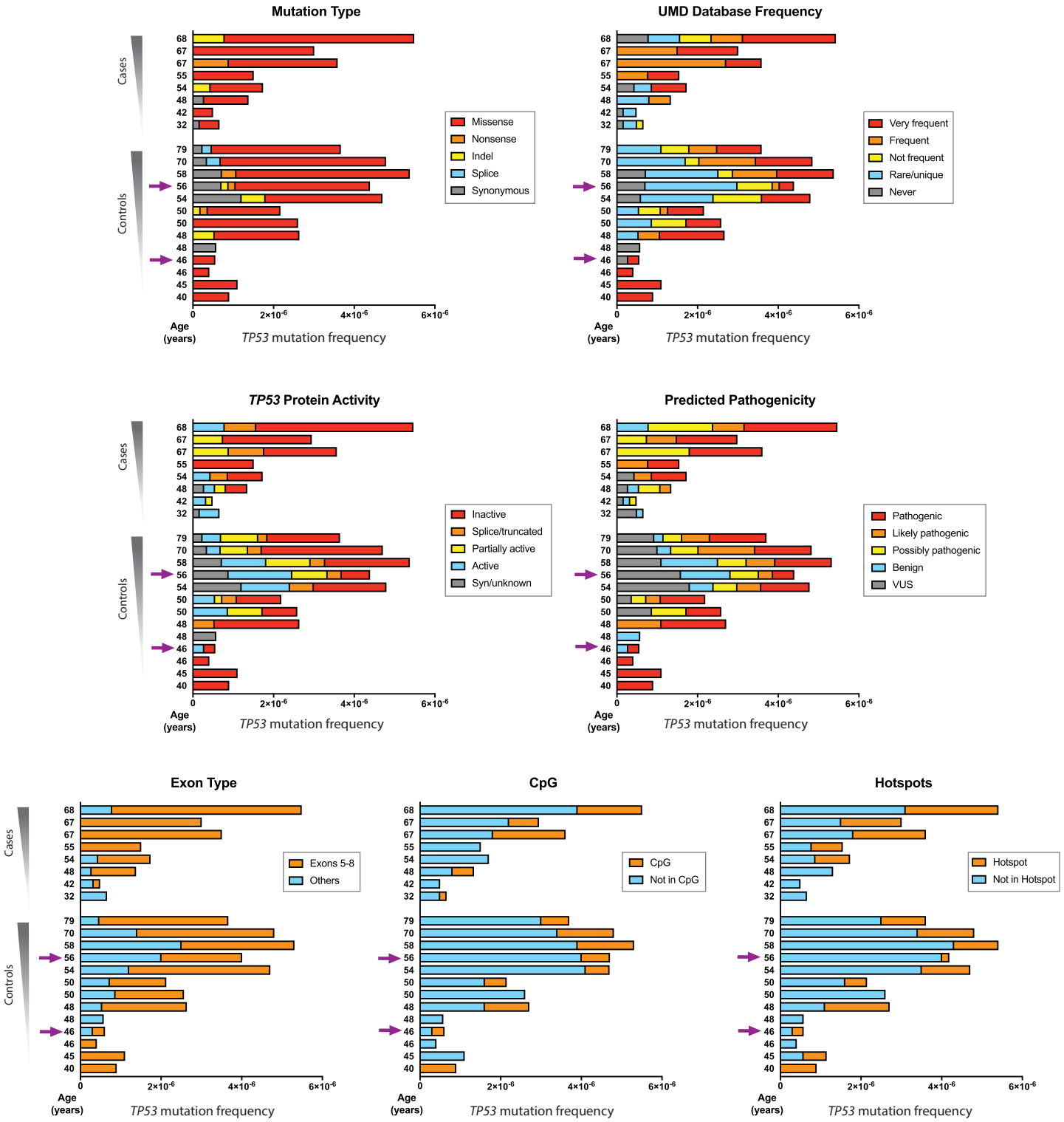


Figure S5. *TP53* mutation frequency and characteristics by age including uterine lavages from the two middle age women in the normal tissue study [related to Figs. 3E-F and 4B]. The two new lavages correspond to a 46 year old woman and a 56 year old woman and are indicated by arrows. *TP53* mutation type, frequency in cancer database, activity, pathogenicity, exon 5-8 location, CpG location, and hotspot location are indicated by color, with warm colors indicating 'cancer-like' features. The *TP53* mutation frequency and the distribution of mutational cancer-like traits in the two new lavages are very similar to the data obtained for women of comparable age in the first part of the study.

TISSUE	MUTATIONS FOUND IN DIFFERENT TISSUES OF SAME INDIVIDUAL					
101 yo	**c.659A>G**	**c.596G>A**	**c.517G>A**	**c.455C>T**	**c.389T>G**	**c.149T>C**
A001-Leukocytes	85/6971	2/11543		44/9587	1/10797	1/11512
A001-Peritoneum (a)	1/4501			1/4939		
A001-Peritoneum (b)	1/5344	3/8248	1/7342	2/6633		
A001-Endometrium (a)	1/5712		1/7803		1/7961	
A001-Endometrium (b)						1/9988

56 yo	**c.151G>T**
A004-Leukocytes	1/13560
A004-Cervix (a)	
A004-Cervix (b)	
A004-Endometrium (a)	
A004-Endometrium (b)	
A004-Myometrium	
A004-FT (a)	
A004-FT (b)	
A004-Ovary (a)	
A004-Ovary (b)	
A004-Uterine lavage	1/4843

46 yo	**c.659A>G**	**c.524G>T**
A006-Leukocytes		1/6665
A006-Peritoneum (a)		
A006-Peritoneum (b)		
A006-Cervix		1/4349
A006-Endometrium		
A006-Myometrium (a)	16/2694	
A006-Myometrium (b)		
A006-FT (a)	1/1918	
A006-FT (b)		
A006-Uterine lavage		
A006-Peritoneal fluid		
A006-ctDNA		

	MAF>1%
	MAF>0.1%

Figure S6. Analysis of mutations shared across multiple tissue samples within the same individual [related to Figs 5B and 6]. For each individual, all analyzed samples are listed and color coded by tissue type. Mutations identified in more than one biopsy are indicated in columns with ratios provided for the biopsies in which the mutation was identified. The ratios indicate the number of Duplex reads with the given mutation divided by the depth of sequencing in that position. Mutations found at MAF>1% and MAF>0.1% are indicated. It should be noted that the first mutation listed for the 101 year old woman and the 46 year old woman is the same and corresponds to codon 220, one of the most common hotspots in *TP53*. FT: fallopian tube.

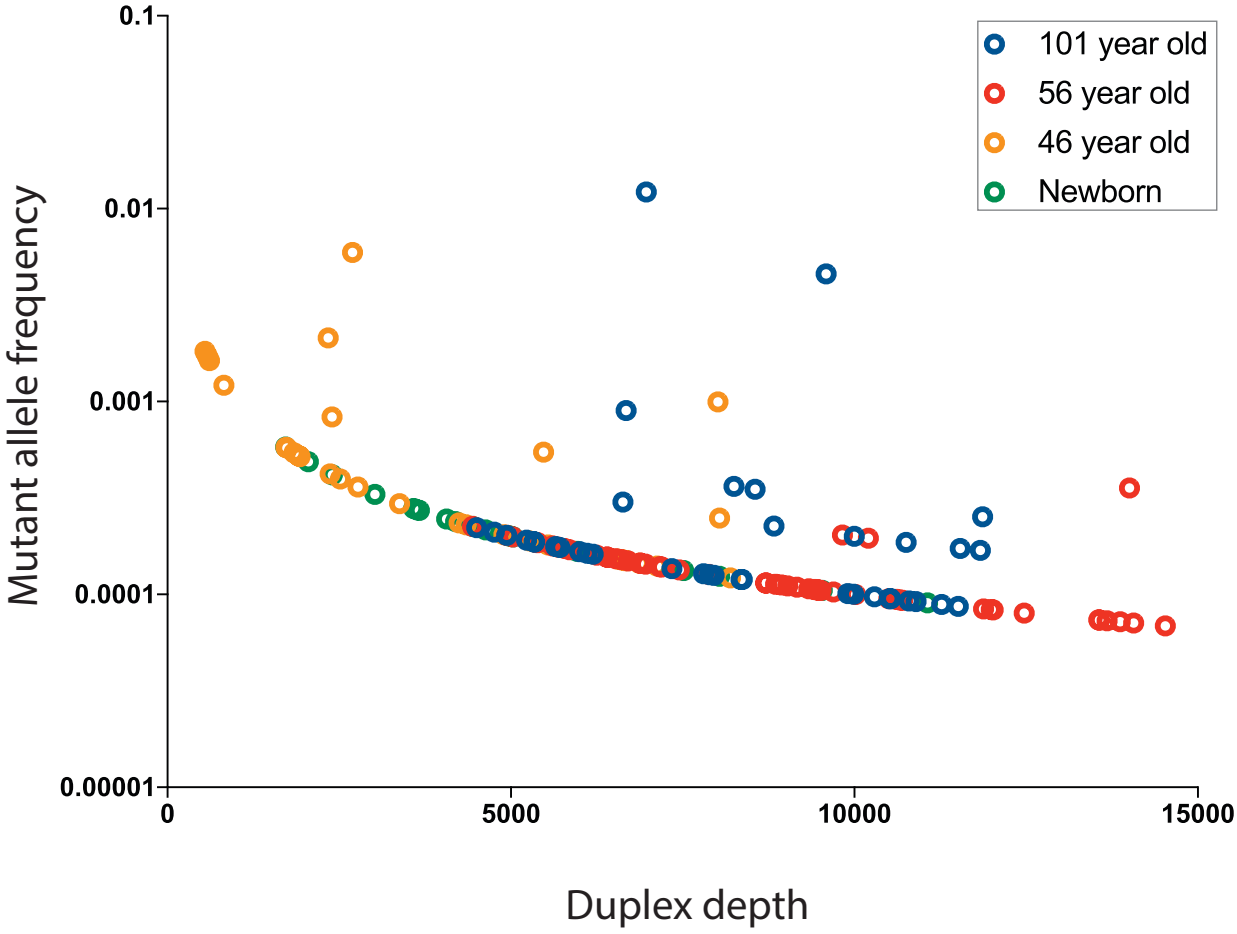


Figure S7. Mutant allele frequency as a function of Duplex Sequencing depth [related to Fig. 5B]. Each dot corresponds to a *TP53* mutation identified in normal tissue. Mutations are color coded by subject. MAF is calculated as the number of times a mutation was observed divided by the depth of sequencing at the given position. Because MAF is inversely associated with depth, mutations identified in biopsies sequenced at a lower depth (mostly from the 46 year old woman and newborn) present with higher MAF.

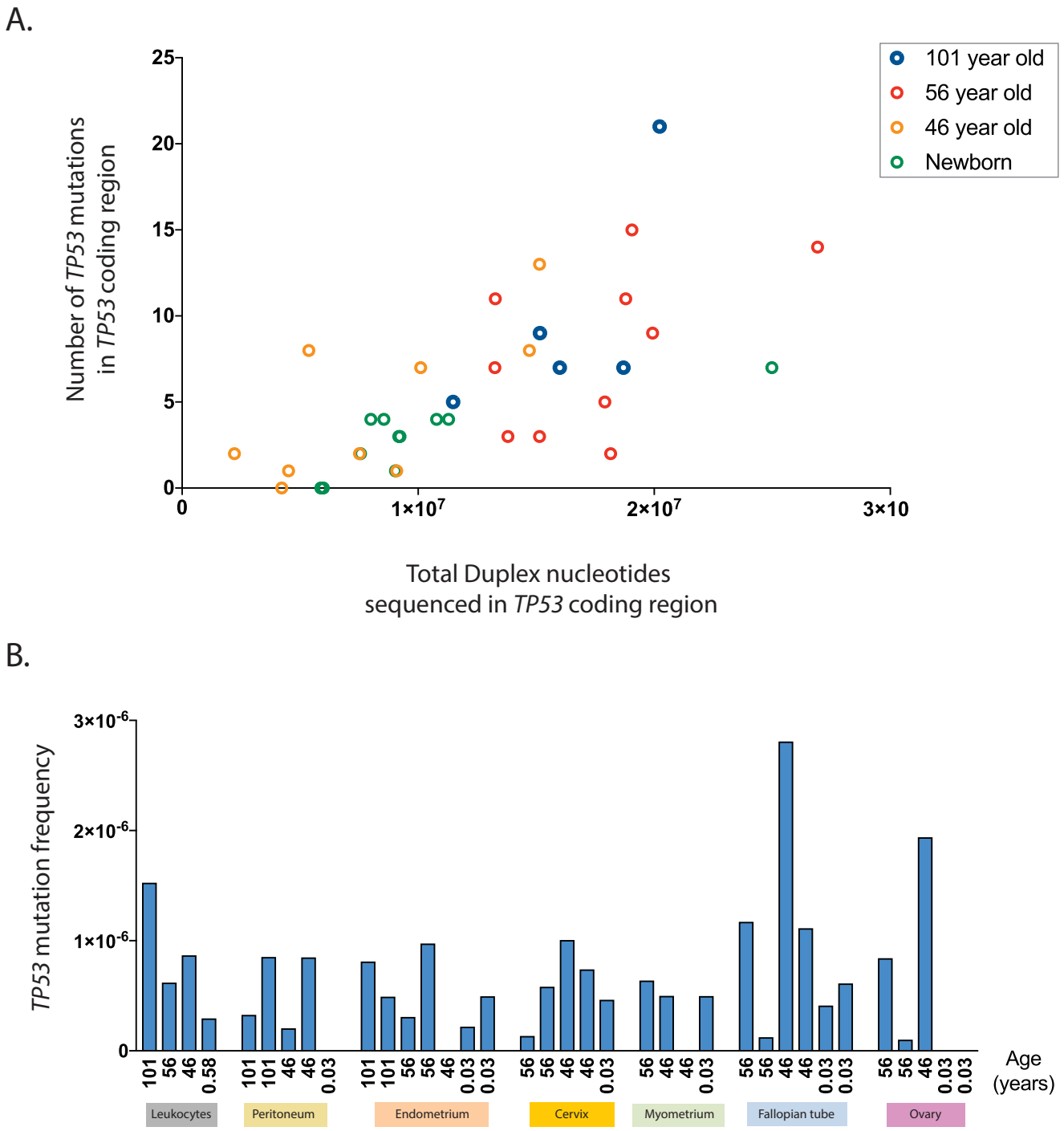


Figure S8. *TP53* mutation frequency by tissue type [related to Figs. 5B, 6, 7A]. (A) Association between number of *TP53* mutations and total number of Duplex nucleotides sequenced in the *TP53* coding region. Dots correspond to samples and are color coded by individual of origin. (B) For each sample, *TP53* mutation frequency was calculated as the number of *TP53* mutations identified in the coding region divided by the total number of Duplex nucleotides sequenced in that region. Subject age is indicated in the X-axis.

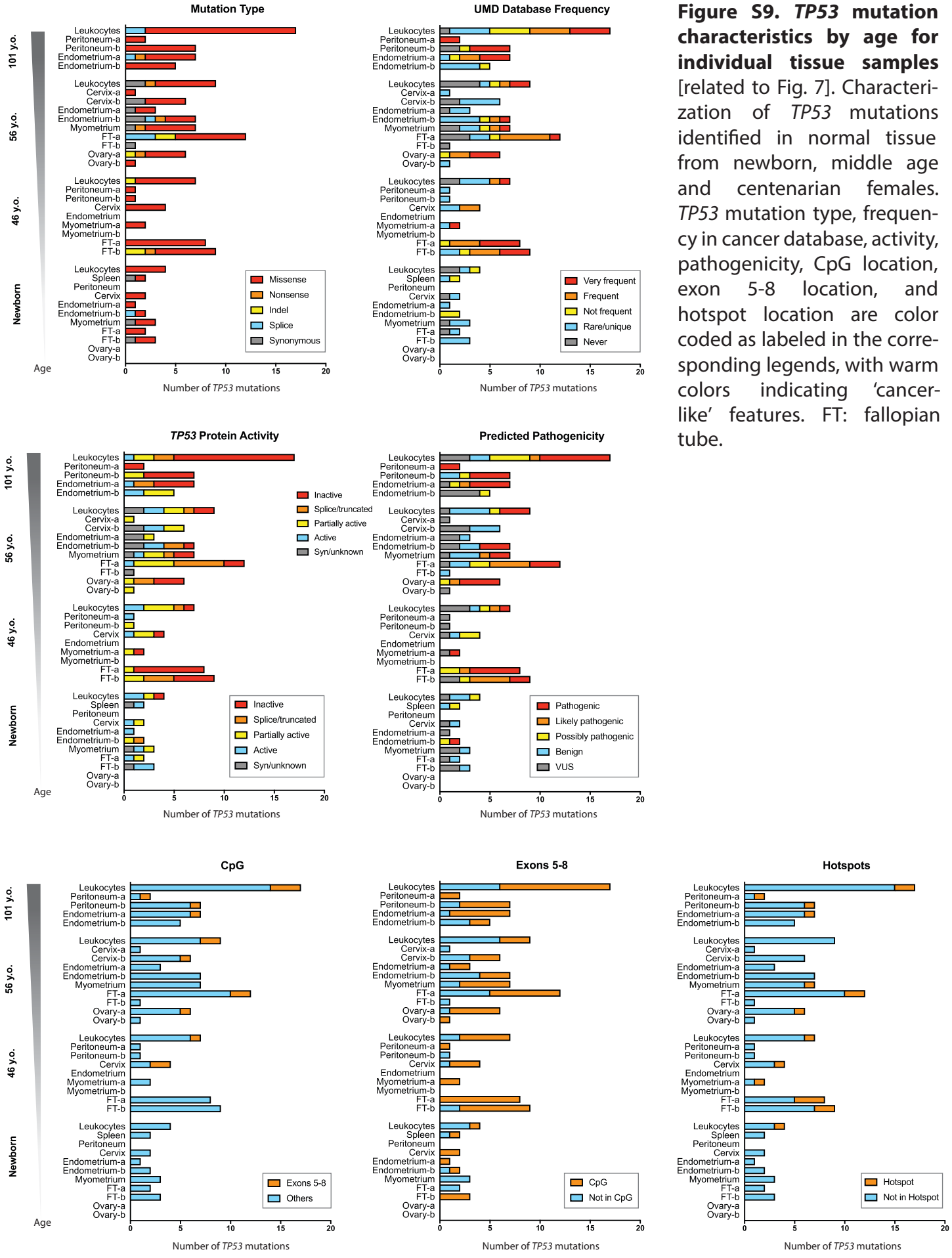


Figure S9. *TP53* mutation characteristics by age for individual tissue samples [related to Fig. 7]. Characterization of *TP53* mutations identified in normal tissue from newborn, middle age and centenarian females. *TP53* mutation type, frequency in cancer database, activity, pathogenicity, CpG location, exon 5-8 location, and hotspot location are color coded as labeled in the corresponding legends, with warm colors indicating 'cancer-like' features. FT: fallopian tube.

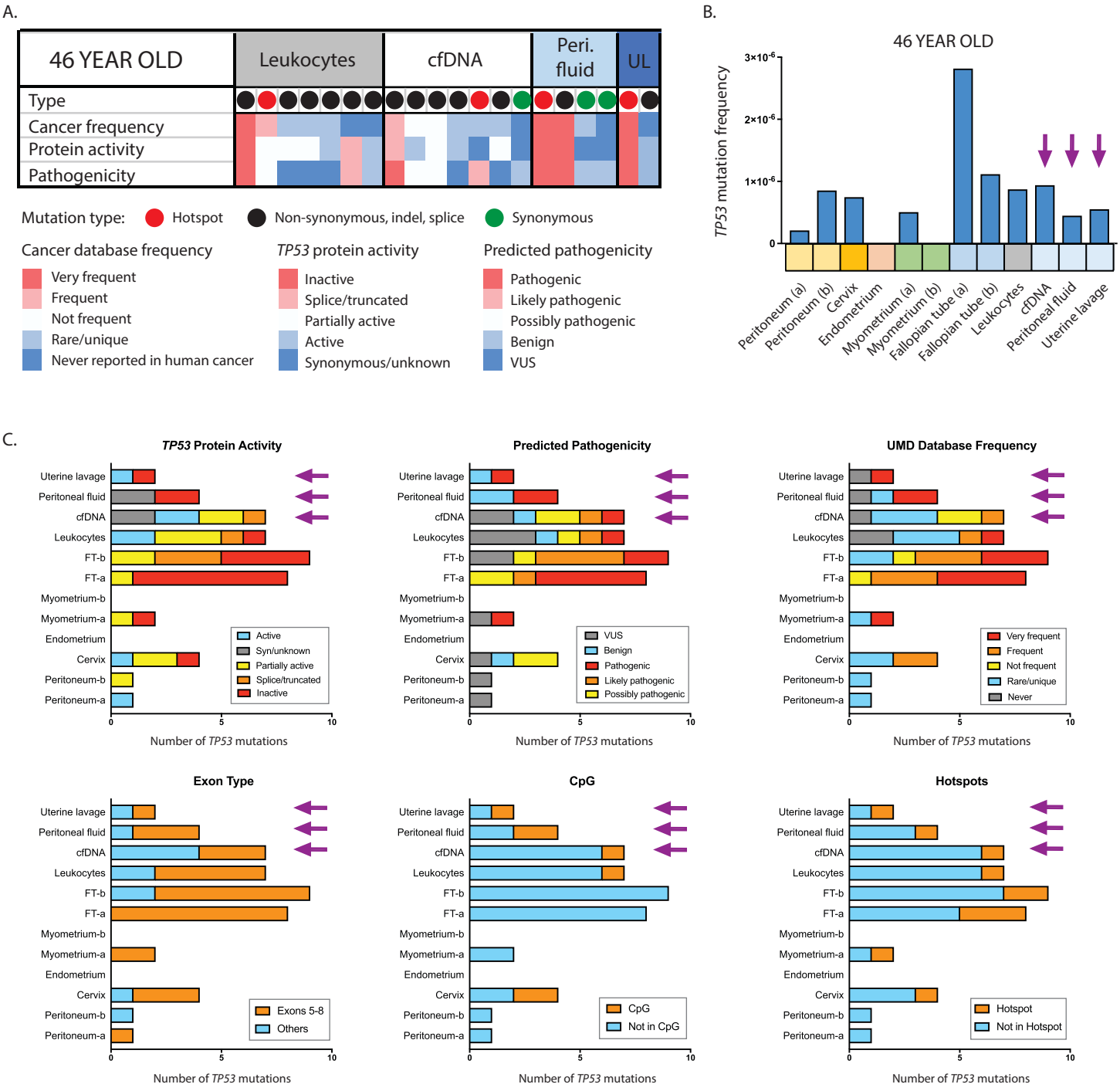


Figure S10. *TP53* mutation characteristics within non-invasively collected body fluids from a 46 year-old woman [related to Figs. 5B and 6]. (A) Heatmap indicating mutation type, frequency in cancer database, impact on protein activity, and predicted pathogenicity for each of the *TP53* mutations identified in leukocytes, cfDNA, peritoneal fluid, and uterine lavage. Categories and color coding are the same as for Figure 6. Each column represents a mutation. (B) Comparison of *TP53* mutation frequency in all tissues collected from the 46 year old woman, including liquid biopsies. (C) Comparison of *TP53* mutational features in all tissues collected from the 46 year old woman, including liquid biopsies (indicated by arrows). FT: fallopian tube.

EXHIBIT I

PROTOCOL

Detecting ultralow-frequency mutations by Duplex Sequencing

Scott R Kennedy[1], Michael W Schmitt[2], Edward J Fox[1], Brendan F Kohrn[3], Jesse J Salk[2], Eun Hyun Ahn[1], Marc J Prindle[1], Kawai J Kuong[1], Jiang-Cheng Shen[1], Rosa-Ana Risques[1] & Lawrence A Loeb[1,4]

[1]Department of Pathology, University of Washington, Seattle, USA. [2]Department of Medicine, University of Washington, Seattle, USA. [3]Department of Biology, Portland State University, Portland, Oregon, USA. [4]Department of Biochemistry, University of Washington, Seattle, USA. Correspondence should be addressed to L.A.L. (laloeb@uw.edu).

Published online 9 October 2014; corrected online 22 October 2014 (details online); doi:10.1038/nprot.2014.170

Duplex Sequencing (DS) is a next-generation sequencing methodology capable of detecting a single mutation among >1 × 10^7 wild-type nucleotides, thereby enabling the study of heterogeneous populations and very-low-frequency genetic alterations. DS can be applied to any double-stranded DNA sample, but it is ideal for small genomic regions of <1 Mb in size. The method relies on the ligation of sequencing adapters harboring random yet complementary double-stranded nucleotide sequences to the sample DNA of interest. Individually labeled strands are then PCR-amplified, creating sequence 'families' that share a common tag sequence derived from the two original complementary strands. Mutations are scored only if the variant is present in the PCR families arising from both of the two DNA strands. Here we provide a detailed protocol for efficient DS adapter synthesis, library preparation and target enrichment, as well as an overview of the data analysis workflow. The protocol typically takes 1–3 d.

© 2014 Nature America, Inc. All rights reserved.

INTRODUCTION

Next-generation DNA sequencing (NGS) defines the modern genomic era. This powerful technology has revolutionized traditional genetics and has made feasible the emerging field of personalized medicine. It is now routine to sequence billions of nucleotides and to identify inherited clonal mutations. However, all NGS approaches have a relatively high error rate: on the order of one erroneous base call per 100–1,000 sequenced nucleotides (**Table 1**; ref. 1). Although this error rate is acceptable for studying inherited mutations, it greatly limits the analysis of subclonal mutations, which are defined as mutations that are present in only a fraction of cells within a population.

There is growing need for technologies capable of resolving subclonal mutations. For example, genetic heterogeneity has long been proposed to be an intrinsic driver of cancer initiation and progression[2]. Recent tumor genome sequencing studies suggest that human cancers exhibit extreme levels of genetic heterogeneity[3–6]. Subclonal mutations are probably a major factor in cancer relapse and in rapid emergence of chemotherapy resistance[7–9]. However, the study of cancer subclones requires the confident detection of mutations that are present in <1% of cells—a level of resolution that cannot be obtained by conventional sequencing approaches. Similarly, the genetic diversity found within microbial populations underlies their ability to adapt to changing environments, including development of drug resistance[10–13], but this genetic diversity is difficult to directly assess owing to the high background error rate of conventional NGS sequencing. Other fields with a similar need for robust low-frequency mutation detection include forensics[14], paleogenomics[15,16], evolution[17] and toxicology[18], as high-accuracy sequencing would allow one to assess the potential mutagenicity of new chemical compounds without the need for a genetic selection system to identify mutant genes.

The concept of DS

To overcome the high error rate of next-generation sequencing and thereby facilitate the study of subclonal and random mutations, we recently developed a highly sensitive sequencing methodology termed Duplex Sequencing (DS). DS yields unprecedented accuracy in sequencing of double-stranded DNA, with a >10,000-fold improvement compared with conventional NGS, and it has the unique ability to detect a single mutation among >10^7 sequenced bases[19]. DS takes advantage of the inherent complementarity of double-stranded DNA by using degenerate molecular tags[20–26] to label each fragmented DNA molecule with its own unique DNA sequence. By tagging duplex DNA with adapters containing random yet complementary double-stranded nucleotide sequences, it becomes feasible to trace every sequence read back to one of the two strands of the original double-stranded DNA molecule (**Fig. 1a**). After adapter ligation, the individually labeled strands are PCR-amplified to create sequence families that share the same tag sequences derived from each of the two single parental strands (**Fig. 1b**). After sequencing, members of each tag family are grouped and a consensus sequence is established for each of the two strands to form 'single-strand consensus sequences' (SSCSs; **Fig. 1c**). The two complementary consensus sequences derived from the two strands of an individual DNA duplex are then compared with each other, and the base identity at each position is

TABLE 1 | NGS platforms and their associated errors[1].

Platform	Primary error type	Background (%)
Pacific Biosciences	G/C deletions	16
Life Ion Torrent	Short deletions, homopolymers	1
ABI SOLiD	A-T bias	0.2
Illumina MiSeq	Single nucleotide	0.1
Illumina HiSeq	Single nucleotide	0.1

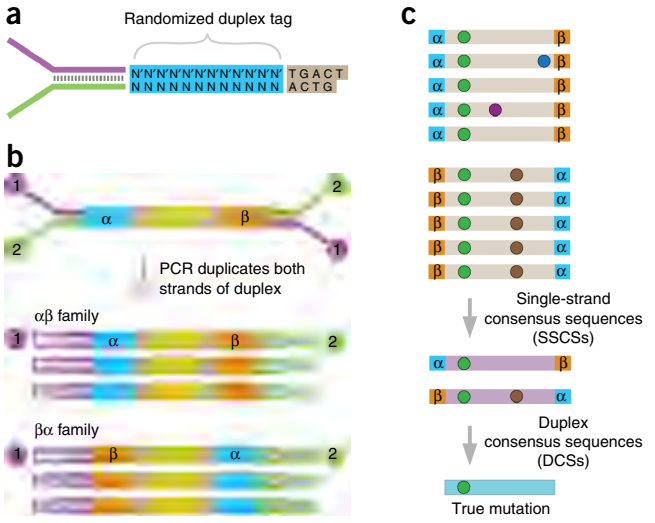


Figure 1 | Overview of Duplex Sequencing. (**a**) Schematic of a Duplex Sequencing adapter, showing the random double-stranded tag and the invariant spacer sequence. (**b**) Ligation of the adapters to the sample DNA results in a unique 12-nt tag sequence on both ends of the molecule. PCR amplification of each strand of a DNA duplex results in two distinct, but related, PCR products. (**c**) Reads sharing unique α and β tag sequences are grouped together into tag families of form αβ or βα, and an SSCS is created for each tag family. Mutations are of three different types: sequencing mistakes (blue or purple dots); first-round PCR errors (brown dots); true mutations (green dots). Formation of the SSCS removes the first type of error, but not first-round PCR errors. Comparing SSCSs from the paired families with tags αβ and βα generates a DCS, which eliminates these first-round PCR errors. True mutations are scored if and only if they are present at the same position in both strands of the DNA. Figure is adapted from ref. 33, © 2013 Kennedy *et al.*

retained only if the two strands match perfectly at that position, yielding a 'duplex consensus sequence' (DCS; **Fig. 1c**). Mutations introduced during PCR by DNA polymerase misincorporations or arising from DNA damage will appear in only one of the two DNA strands and thus are not counted as real mutations.

Alternatives to DS

As the need for accurate sequencing methods has increased, three main strategies besides DS have emerged: (i) single-cell sequencing[4,27–29], (ii) single-stranded molecular barcoding[20,21,23] and (iii) circle sequencing (CirSeq)[30,31]. Although each approach has unique strengths, all three of these methods involve sequencing DNA derived from a single strand of a double-stranded molecule. Misincorporation events by DNA polymerase occurring during the first round of amplification will inherently be propagated to the daughter molecules, and they are likely to be erroneously scored as mutations. In the case of single-cell sequencing, the use of random primer sequences in conjunction with a strand-displacing DNA polymerase results in random priming from the newly synthesized DNA and the generation of 'copies of copies', thus propagating any initial misincorporation events to all of the reads. Because all the reads are derived from a single cell, the propagated error would be incorrectly called a genetic variant. A similar process can occur in CirSeq and single-stranded barcoding, whereby a misinsertion event occurring during the first round of synthesis can be propagated to subsequently synthesized daughter molecules. When sequenced, these subpopulations of related molecules will all contain a copy of the original misincorporation event, and they will be erroneously scored as a mutation. However, because the same mutation is unlikely to occur in unrelated molecules, these artifactual variants would give the appearance of a subclonal mutation. DNA polymerases typically used in library construction make misinsertions at a frequency between 10^{-4} and 10^{-6}, which can lead to thousands of false positives on a typical sequencing run. Damaged or degraded DNA is particularly sensitive to this form of error because of the prevalence of DNA adducts that cause erroneous base pairings during polymerization. Removal of DNA damage with the addition of glycosylases or *in vitro* repair kits has been shown to reduce the number of false mutations in these samples[31,32]. However, not all mutagenic lesions are recognized by these enzymes, nor is the fidelity of repair perfect, thus limiting their utility in error correction.

Advantages of DS

The primary advantage of DS over other approaches is its superior accuracy. In our original publication on DS[19], we demonstrated the direct measurement of the mutation frequency for the M13 bacteriophage as 2.5×10^{-6} mutations per base pair. We have subsequently measured mutation frequencies as low as 5×10^{-8} in human nuclear DNA (E.J.F. and L.A.L., unpublished results). In addition, we recently applied DS to measuring the mutation frequency of human mtDNA in brain tissue from young and old individuals[33]. We showed that the frequency of mtDNA point mutations is 10–100-fold lower than that previously reported, and the frequency increases approximately fivefold over an 80-year lifespan. Surprisingly, we observed no significant age-associated increase in the mutations most commonly associated with oxidative damage to mtDNA. This finding is inconsistent with free-radical theories of aging and suggests that previous studies indicating that reactive oxygen species induced mutations increase with age may be the result of PCR artifacts.

Similarly to our results with mtDNA, we have found that DS can help mitigate artifactual mutations derived from chemically damaged DNA. Numerous studies using NGS to sequence DNA from formalin-fixed tissues have reported a high number of false mutations[34–36]. Because complementary artifacts from damage events are unlikely to occur at the same corresponding position on paired DNA strands, the use of sequencing information from both strands to correct these errors makes DS extremely resistant to damage-induced misincorporations. By using DS, we have sequenced paired formalin-fixed and unfixed tissue samples, and we found only a twofold change in the mutation frequency, indicating that DS is useful for removing mutational artifacts even in extremely damaged and degraded DNA (M.J.P., E.J.F. and L.A.L., unpublished results).

Limitations of DS

DS is a uniquely sensitive method that is capable of detecting and quantifying mutations that occur at extremely low frequencies; however, this ability comes at a cost. Owing to the method's reliance on sequencing multiple PCR duplicates of both DNA strands, DS requires much larger sequencing capacity than conventional NGS to produce a given depth of sequencing. At present, the use

© 2014 Nature America, Inc. All rights reserved.

Box 1 | Synthesis of DS adapters ● TIMING 2 d

The adapter synthesis protocol generates a sufficient amount of adapters for several hundred samples.

1. Anneal the two oligonucleotides (MWS51 and MWS55 in **Table 3**) by combining 100 µl of each 100 µM oligonucleotide in a 0.2-ml PCR tube; heat the tube to 95 °C for 5 min in a thermocycler with a heated lid. Turn off the machine and leave it for 1 h. Remove and save 1 µl for quality control; label it as 'annealed adapters'.

▲ CRITICAL STEP It is extremely important that the oligonucleotides be allowed to cool slowly. Make sure that the thermocycler does not automatically cool after the heating cycle.

2. Extend the annealed adapters by mixing the components in the table below by gently pipetting up and down, and then splitting into two 0.2-ml PCR tubes and incubating for 1 h at 37 °C.

Reagent	Volume (µl)	Final concentration
DNA from step 1	199	Variable
10× NEB buffer #2	27.9	1×
10 mM dNTP mix	27.9	1 mM
ddH_2O	11.6	N/A
Klenow exo$^-$ (5 U/µl)	11.6	0.5 U/µl

3. Ethanol-precipitate the DNA and resuspend it with 200 µl of ddH_2O. We recommend saving 1 µl of the resuspended adapters for quality control purposes; label as 'extended adapters'.

▲ CRITICAL STEP It is essential that the two strands never melt. If they do, the complementarity of the tags will be lost, and DS will not work. The strands could melt if the DNA is overdried after precipitation.

? TROUBLESHOOTING

4. Cleave the extended adapters by mixing the components in the table below by carefully pipetting up and down and then dividing into four 0.2-ml PCR tubes. Incubate the tubes for 16 h at 37 °C in a thermocycler with a heated lid. After 16 h, remove and save 1 µl for quality control purposes; label as 'cut adapters'.

Reagent	Volume (µl)	Final concentration
DNA from step 3	200	Variable
10× NEB CutSmart buffer	50	1×
ddH_2O	235	N/A
HpyCH4III (5 U/µl)	15	0.5 U/µl

5. Add 50 µl of 3 M sodium acetate (pH 5.2) and mix. The final volume should be 550 µl.

6. Divide the solution into aliquots in 1.5-ml tubes, 275 µl per tube (~2 tubes total), and add 675 µl (i.e., 2.5 volumes) of room-temperature 100% ethanol to each tube; mix the tubes by inversion.

? TROUBLESHOOTING

7. Centrifuge the mixture for 30 min at 4 °C at >10,000*g*.

8. Remove the supernatant, add 1 ml of 75% (vol/vol) ethanol to each tube and mix by inversion. Immediately centrifuge the tubes for 30 min at 4 °C at >10,000*g*.

9. Remove the supernatant. Dry the tube for 10 min inverted on a paper towel, and then for 5 min upright.

▲ CRITICAL STEP Dry the DNA pellet for the exact times specified. Overdrying the DNA could lead to strand melting.

10. Resuspend the pellet in each tube with 100 µl TE_{low}. After resuspension, pool all the tubes together for a final volume of 200 µl. Remove and save 1 µl for quality control; label as 'final adapters'.

11. Divide the adapters into 50-µl aliquots, and place them at –80 °C for long-term storage. The final concentration will be ~50 µM.

12. (Optional) Radiolabel the sample from each step and run it on an 8 M urea, 14% (wt/vol) acrylamide gel (**Fig. 2a**). Although this step is optional, it is strongly recommended. See **Box 2** for a detailed protocol.

! CAUTION γ-^{32}P-ATP is radioactive. Take appropriate steps to avoid exposure to radioactive compounds.

? TROUBLESHOOTING

of current NGS technology with DS to sequence large genomes or genome targets >1–2 Mbp in size to a high depth of coverage is prohibitively expensive; however, as the cost of sequencing continues to fall, we anticipate that this will become increasingly practical.

Overview of the procedure

Sequencing library construction for DS is similar to the standard Illumina library preparation protocol. The protocol follows the basic standard steps of DNA shearing by sonication, size selection, end repair, 3′ dA-tailing, adapter ligation, PCR amplification and, optionally, targeted DNA capture. We have made several important updates and optimizations to the protocol since its initial publication, which have substantially increased its reliability and reproducibility. Specifically, we have re-designed the sequencing adapters to allow for more efficient dA-tailing of the sample DNA. As part of this new design, the adapters require a different synthesis method that we have included in **Boxes 1** and **2**. An additional change in the protocol involves the amount of DNA used during the PCR step. In the original publication, we specified that ~40 attomoles of DNA was optimal. However, we have since determined that the PCR input amount depends on a number of factors, including the number of reads devoted to a sample, the

© 2014 Nature America, Inc. All rights reserved.

Box 2 | Optional quality control steps for adapter synthesis using radiolabeled adapters and PAGE ● TIMING 1 d

The following steps provide a method by which to evaluate the synthesis of the adapters. We have found that the use of radiolabeling with polyacrylamide electrophoresis provides the most consistent and sensitive way to evaluate the success of the adapter synthesis steps. Agarose electrophoresis could also be used; however, the fluorescent dyes used require double-stranded DNA. Therefore, the evaluation of the extension step (step 3) will be limited. Technical solutions, such as the Agilent TapeStation 2200 or Bioanalyzer 2100, do not have enough resolution to detect the 9-bp size change that occurs during enzymatic restriction (step 4) and should be avoided.

1. Dilute each DNA sample from steps 1, 3, 4 and 5 from **Box 1** in 9 µl of ddH_2O.
2. For each DNA sample, mix the following components and incubate the mixture at 37 °C for 30 min.

! CAUTION γ-^{32}P-ATP is radioactive. Take appropriate steps to avoid exposure.

Reagent	Volume (µl)	Final concentration
DNA from step 1	1	Variable
10× NEB polynucleotide kinase (PNK) buffer	1	1×
γ-^{32}P-ATP (6,000 Ci/mmol)	0.25	0.4 µM
ddH_2O	16.25	N/A
T4 PNK (10 U/µl)	0.5	0.5 U/µl

3. Add 20 µl of the denaturing gel-loading buffer containing sequencing gel dye mix for a total volume of 40 µl.
4. Boil the DNA:buffer mixture for 1 min.
5. Mix 100 ml of the 8 M urea, 14% (wt/vol) polyacrylamide gel mixture with 130 µl of ammonium persulfate and 75 µl of tetramethylethylenediamine (TEMED). Pour it rapidly into the sequencing gel plate setup, as described by the manufacturer.
6. Run 5 µl of the final DNA:buffer mixture on the 8 M urea, 14% (wt/vol) polyacrylamide gel at 60 W for 1.5–2 h or until the bromophenol blue band reaches three-fourths of the way to the bottom of the gel.

▲ CRITICAL STEP The small fragment that is removed by ethanol precipitation migrates near the bottom of the bromophenol blue marker. Do not run the gel for too long to avoid loss of the fragment. See **Figure 3** for an example gel.

use of targeted DNA capture, genome and target size, and gene and target copy number. We formally outline the relationship between these parameters and provide the optimal DNA input amounts to maximize the final amount of data. Finally, we have included the option to use targeted DNA capture, which greatly expands the utility of DS beyond small, highly pure genomes, such as mtDNA and other small plasmids that were presented in our original publication.

In addition to the updated protocol, we also include a comprehensive computational pipeline that we use to process and analyze our data (**Fig. 2**). The computational workflow for DS uses a number of standard software packages to process the sequencing data; these include the Burrows-Wheeler Aligner (BWA)[37], SAMtools[38], Picard and the Genome Analysis Toolkit (GATK)[39,40], as well as several custom Python scripts. The computational workflow is broken down into three major steps: (i) Tag parsing and initial alignment; (ii) SSCS assembly; and (iii) DCS assembly. The latest version of the DS software package can be downloaded from https://github.com/loeblab/Duplex-Sequencing.

Each read obtained from a DS run consists of a 12-nt tag sequence, followed by an invariant 5-bp sequence corresponding to the ligation site. First, the invariant 5-bp sequence is computationally removed from each read, and the 12-nt tag present on each of the two paired-end reads is combined to a single 24-nt tag that is stored in the read header. Sequences with ambiguous nucleotides or homopolymers greater than nine bases within the tag are discarded. These steps are all performed by the custom python script called 'tag_to_header.py' (**Supplementary Fig. 1**). The reads are then aligned to the reference genome using BWA[37]. After alignment, reads sharing the same tag sequence and genomic coordinates are identified and grouped to form 'tag families' with a python script called 'ConsensusMaker.py'. By default, the script requires three members to result in a tag family. The family members are then compared at each sequence position, and the identity of a position is kept only when at least 70% of the members have the same sequence at that position. Positions that cannot form a consensus are replaced by an 'N' and are considered undefined. The resulting data are referred to as SSCSs (**Supplementary Fig. 2**). We have experimented with requiring more than three members per family or >70% sequence agreement, but we have found that this reduces data yield without any appreciable change in the method's accuracy (**Supplementary Fig. 3**).

Next, the two related SSCS reads corresponding to the two initial DNA strands are grouped together and compared by a script called 'DuplexMaker.py'. The 24-nt tag associated with each sequence read consists of two 12-nt sequences, and the tags corresponding to a pair of SSCS reads can be grouped by virtue of being transposed relative to one another. Specifically, if the two 12-nt sub-tag sequences are designated α and β, then a sequence with a tag $\alpha\beta$ in read 1 is compared with the sequence having the tag $\beta\alpha$ in read 2. The paired strand SSCS reads are then compared at each position, with only matching bases being kept to produce a DCS. Non-matching bases are considered undefined, and they are replaced by 'N'. Reads containing a high proportion of Ns (>30%) are filtered out during this step (**Supplementary Fig. 4**). The processed DCS data are then re-aligned to the reference genome.

Experimental design

In the following sections, we will highlight important considerations that should be taken into account for each stage in the protocol, as well as provide details about the changes and improvements we have made since the initial publication[19].

© 2014 Nature America, Inc. All rights reserved.

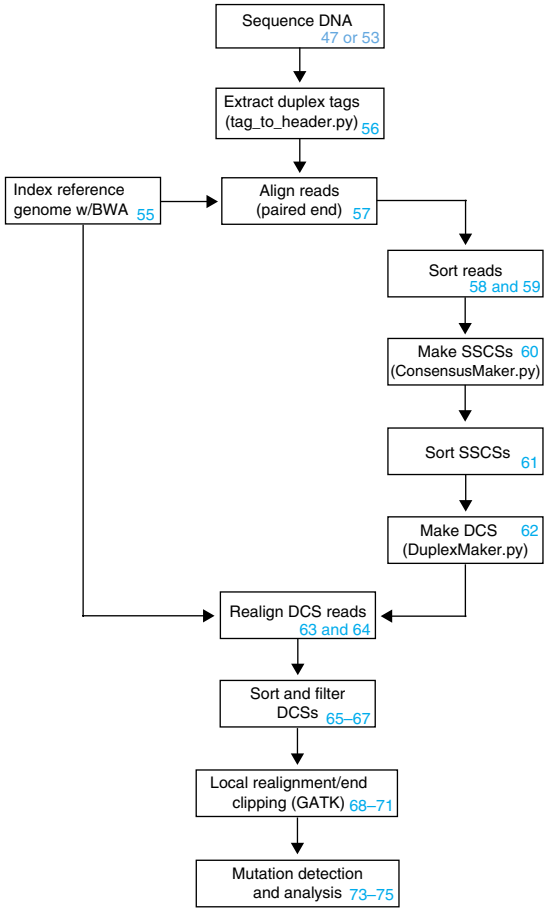


Figure 2 | Schematic of the basic computational workflow for Duplex Sequencing. The blue numbers correspond to the steps in the PROCEDURE.

Sample requirements. The amount of fragmented DNA actually used in the PCR amplification step is typically in the low-attomole range, so the theoretical amount of starting DNA needed for DS is extremely low. However, when feasible, it is most convenient to start with excess amounts of DNA to allow for easy quantification of the DNA throughout the initial sample preparation steps and to account for expected sample loss at the enzymatic and purification steps. Typically, 1–3 µg of DNA is required for targeted DNA capture and 100–300 ng of DNA for smaller genomes, such as purified plasmid or mtDNA, that do not require targeted capture. When performing DNA isolation, it is imperative to avoid any manipulations that could result in melting of the two DNA strands, such as heating above 80 °C, the use of chaotropic agents such as urea or overdrying the DNA during ethanol precipitations.

DNA fragmentation. We physically shear the DNA by sonication using a Covaris acoustic ultrasonicator. The shearing time is varied depending on the size of the genome, with shorter times for small genomes (such as mtDNA or viral DNA) and longer times for nuclear DNA. Specific shearing parameters are given in Equipment Setup below. Other methods such as nebulization or enzymatic digestion should both be compatible with DS, but would probably require optimization. 'Tagmentation' methods, which involve simultaneous fragmentation of the sample DNA and ligation of sequencing adapters by a transposase (e.g., Illumina's Nextera DNA sample prep kit), are currently incompatible with DS because they make use of an invariant transposon sequence that is incompatible with molecular barcodes. Because the DNA strands are not marked with molecular barcodes, it is impossible to identify unique DNA molecules and perform the comparison between complementary strands.

Size selection. Many protocols for next-generation sequencing library preparation use PAGE, with excision and purification of the desired fragment range from the gel. However, gel-based size selection can result in melting of the two DNA strands, precluding its use with DS. In addition, the use of gels increases the introduction of DNA damage during UV transillumination. To mitigate these problems, we perform size selection with Ampure XP beads, which has the additional benefits of higher recovery and greater speed.

End repair and dA-tailing. Similar to the standard Illumina library construction protocol, the sheared DNA is subjected to end repair and 3′-end dA-tailing. Notably, in our initial description of DS[19], we used A-tailed adapters and enzymatically added a 3′ dT overhang on the fragmented DNA sample. However, since then, we have determined enzymatic T-tailing to be considerably less efficient than A-tailing (**Supplementary Fig. 5**), which results in a substantially reduced number of final reads. We have re-engineered the adapters to have a 3′ dT overhang (**Fig. 1a**, see 'Adapter synthesis' section and **Boxes 1** and **2**), which simultaneously increases the final data yield (**Supplementary Table 1**) and makes the protocol more consistent with conventional NGS library preparations.

Adapter synthesis. DS adapters are constructed by annealing two oligonucleotides, one of which contains a 12-nt single-stranded randomized tag sequence. A DNA polymerase is used to copy the degenerate tag sequence, thereby converting it to a double-stranded form (**Fig. 3**). The extended product contains



Figure 3 | Quality control of the sequencing adapters at each step of synthesis. (**a**) Representative 14% (wt/vol) polyacrylamide gel for each step of the adapter synthesis. Lane numbers correspond to the following steps in the synthesis protocol described in **Box 1**. Lane 1: step 1 (annealed adapters); lane 2: step 3 (extended adapters); lane 3: step 4 (cut adapters); and lane 4: step 10 (final adapters). Band sizes are as follows: a = 58 nt; b = 56 nt; c = 83 nt; d = 75 nt; e = 48 nt; and f = 9 nt. (**b**) Schematic of the adapters at each step on the synthesis. Lane and gel band designations correspond to the designations in **a**.

© 2014 Nature America, Inc. All rights reserved.

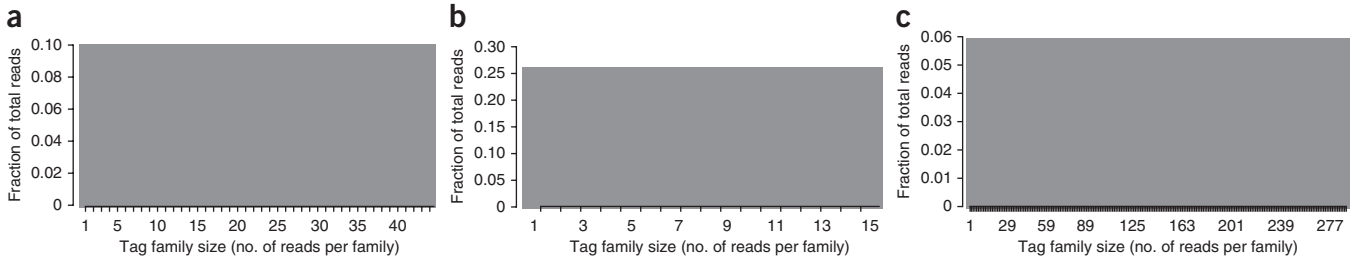


Figure 4 | Representative tag family size distributions. (**a**) Optimal family size distribution. (**b**) Tag family size distribution that is too small because of too much PCR input. (**c**) Tag family size distribution that is too large because of too little PCR input. We have found that a family size that is centered around approximately six members maximizes the final number of DCS reads. Samples that exhibit a small peak family size can be sequenced again and the raw sequencing data from the two sequencing runs can be combined and analyzed. Importantly, further sequencing of a sample with a large peak family size will not increase the final depth of coverage.

an HpyCH4III restriction site downstream of the tag sequence. Cleavage of this site results in a 3′ dT overhang on the end of the final adapter (**Figs. 1a** and **3b**), which can be ligated to the 3′ dA overhang on the DNA fragment library to be sequenced. Twelve random nucleotides per adapter (24 nt per final ligated molecule) significantly exceed the degeneracy needed to ensure unique labeling of every molecule in a library. However, a tag length of <12 is incompatible with the Illumina sequencer because of technical limitations of the platform's 'phasing' requirement and should be avoided.

Adapter ligation. The DNA:adapter ratio is a crucial variable, as too few adapters lead to inefficient ligation, and the use of excess adapters can result in adapter dimers, which, owing to their small size, preferentially amplify during PCR. The presence of these adapter dimers interferes with DNA sequencing. To minimize the presence of adapter dimers, we determine the molar concentration of DNA molecules in the preligated DNA library on an Agilent TapeStation 2200 or Bioanalyzer 2100 and then use 20-fold molar excess of adapter relative to DNA. Unligated adapters and adapter dimers are then removed by purification with Ampure XP beads (**Supplementary Fig. 6**).

Setting tag family size by PCR amplification. The sequencing library is PCR-amplified for the purpose of creating multiple copies of each strand of a double-stranded DNA molecule. The number of DNA fragments used in the PCR, along with the fraction of a sequencing lane dedicated to a particular sample, are the primary adjustable variables that dictate the number of sequencing reads that share the same tag sequence (i.e., tag family size), which strongly influences the final number of DCSs formed. If there are too few reads sharing the same tag sequence (i.e., small tag family size), a consensus sequence cannot be calculated; conversely, too many reads having the same tag sequence (i.e., large tag family size) wastes sequencing capacity without appreciably improving data yield. Because the number of reads varies between different tag families and occur in a distribution (**Fig. 4**), we use 'peak family size' as our preferred metric to refer to tag family sizes generated under a given set of conditions. This distribution occurs during PCR amplification, and it is the result of different amplification efficiencies of the DNA molecules present in the library. By plotting the proportion of reads belonging to tag families of the same size (e.g., tag families can have the same number of reads as different tags) as a function of tag family size (the 'PE_reads.tagstats' file generated in Step 60 of the PROCEDURE section provides these data), we typically observe a distribution of tag family sizes with a solitary peak at a tag family size of one, which is probably the result of sequencing errors in the tag region, and a broader distribution centered at the peak family size (see **Fig. 4a** as an example). We formally define peak family size as the tag family size >1 containing the highest proportion of reads. On the basis of the analysis of samples with different values for the peak family size, we have determined that a peak family size of six maximizes the efficiency of DS: that is, requiring the smallest number of raw reads to produce a single DCS (**Fig. 5a**). Although

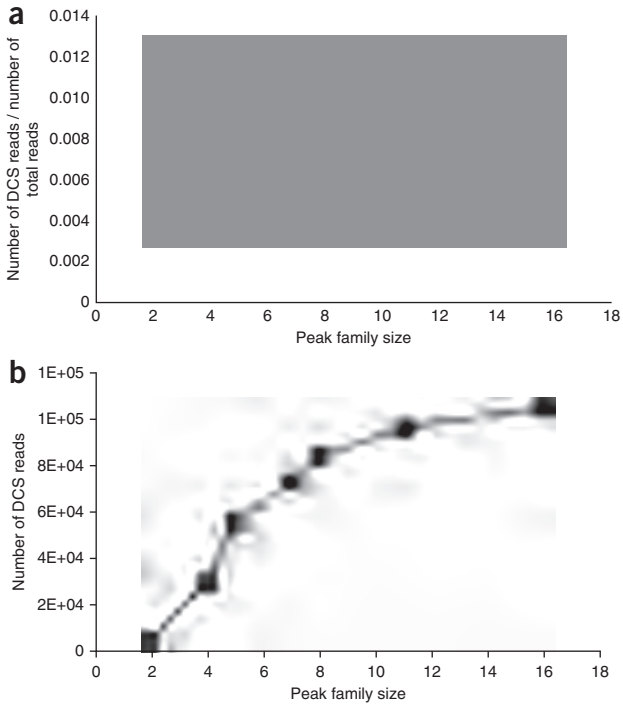


Figure 5 | Optimal peak family size. Replicates of the same sample at different lane fractions. (**a**) Plot compares peak family size to the number of final DCSs that are formed for every read originally dedicated to a sample. The maximum efficiency of DCS formations occurs at a peak family size of six and corresponds to ~40 raw reads being required to form one DCS read. (**b**) The total number of DCSs increases until a peak family size of ~16 is reached. A peak family size >16 does not result in an increase in the final number of DCSs.

© 2014 Nature America, Inc. All rights reserved.

a peak family size of six is optimal, it should be noted that data yield continues to increase, albeit at a decreasing overall efficiency, for peak family sizes ranging between six and 16 members (**Fig. 5b**). As the peak family size approaches 16 reads per tag family, nearly all additional reads are redundant and consume sequencing capacity without yielding further DCS data. Therefore, when deciding on DNA input amounts for PCR, we aim for a peak family size ranging between six and 12 members. This range allows for some tolerance in peak family size because of pipetting or measurement error, while maximizing the final number of DCSs that we obtain.

To achieve greater sequencing depth for a given target, or to achieve the same depth for a larger genome or genome target, more DCSs are needed. To increase the number of reads obtained in DS, it is necessary to proportionally increase both the input amounts of DNA for PCR and the dedicated lane fraction to maintain a peak family size of six (**Table 2**). Nonproportional changes to PCR input and lane fraction lead to suboptimal peak family sizes and substantial decreases in DS efficiency (**Figs. 4b**,**c** and **5a**). The choice of lane fraction (i.e., the number of reads devoted to a particular sample) will be influenced by the size of the targeted region and the desired depth of coverage. In particular, for a given lane fraction, the final sequencing depth of a sample will decrease as the size of the genome target increases. The following formula can be used to estimate the number of reads to devote to a sample:

$$N = \frac{40DG}{R}$$

Where N is the number of paired-end reads devoted to a sample (i.e., lane fraction), D is the desired average depth of coverage, G is the genome or genome target size in base pairs and R is the postanalysis read length in bases (76 in our analysis). The value of 40 is an empirically determined pseudo-constant that approximately corresponds to the average number of unprocessed read-pairs needed to form a single DCS. This value roughly corresponds to the product of the optimal peak family size (i.e., six) and the number of SSCSs typically needed to form a single DCS (we typically obtain SSCS:DCS ratios between four and ten, with an average of around six). However, because not all raw reads are of high quality, several extra reads are typically present that fail to form a DCS, and this tends to raise the number of raw reads needed to form a single DCS to ~40. Notably, this formula provides a rough estimate of the depth that will be obtained provided that the peak family size and SSCS:DCS ratio are optimal. After determining the per-sample lane fraction, this value can be referred to in **Table 2** to determine the amount of target DNA that should be used in the PCR to obtain the appropriate peak family size of six.

For samples that do not make use of targeted capture, such as purified mtDNA or plasmid DNA, we have found that a PCR input of 40 amol of DNA produces an optimal peak family size of six when devoting eight million paired-end reads to a sample (i.e., ~5% sequencing capacity of a HiSeq2500 lane). For the 16.5-kb human mtDNA, this typically results in a depth of 500–1,000. Given the linear relationship between PCR input and lane fraction needed to keep the peak family size constant, 200 attomoles of input mtDNA for 25% of a HiSeq lane results in a depth of 2,500–5,000.

When performing targeted capture, a PCR input amount of 4 amol of target DNA is used for ~8 million paired-end reads (i.e., ~5% sequencing capacity of a HiSeq2500 lane). However, in contrast to a noncapture sample, this value refers specifically to the amount of *target* DNA being captured in the sample, which is generally much less than the amount of total DNA. The amount of total DNA used in the PCR is dependent on its relative copy number and the size of the genomic target, and it can be estimated by the following formula:

$$\text{Total DNA} = \frac{mG}{NT}$$

Where m is the desired amount of *target* DNA in attomoles of fragments (**Table 2**), G is the size of total genome in base pairs, T is the size of the target DNA in base pairs and N is the number of copies of the target per copy of the nontarget portion of the genome. For single-copy genes, N will have a value of 1 and can be ignored. However, in the case of multicopy targets such as mtDNA or ribosomal genes, N can be on the order of hundreds or thousands of copies per nontarget genome and can markedly influence the amount of input DNA. The value of N can be reasonably estimated by performing copy number analysis with quantitative PCR (qPCR) using primers that are specific to the target and nontarget DNA. Similar to the nontarget DNA inputs, the target DNA input amounts should also be proportionally increased with increasing lane fraction (**Table 2**).

The number of PCR cycles is an important variable. Additional cycles beyond the exponential phase can result in the appearance of nonspecific, higher-molecular-weight products (**Fig. 6a**,**b**). To avoid this complication, we carefully monitor the number of PCR cycles. For samples requiring <10 fmol of input DNA, such as for mtDNA, we monitor the reaction by qPCR and remove the reactions one or two cycles before the SYBR Green signal plateaus, which typically occurs between cycles 15 and 20. Notably, the

TABLE 2 | DNA input amounts for PCR.

Reads per sample	Attomoles of DNA (with capture)	Attomoles of DNA (with no capture)
8×10^6	4	40
16×10^6	8	80
24×10^6	12	120
32×10^6	16	160
40×10^6	20	200
48×10^6	24	240
56×10^6	28	280
64×10^6	32	320
72×10^6	36	360
80×10^6	40	400

© 2014 Nature America, Inc. All rights reserved.

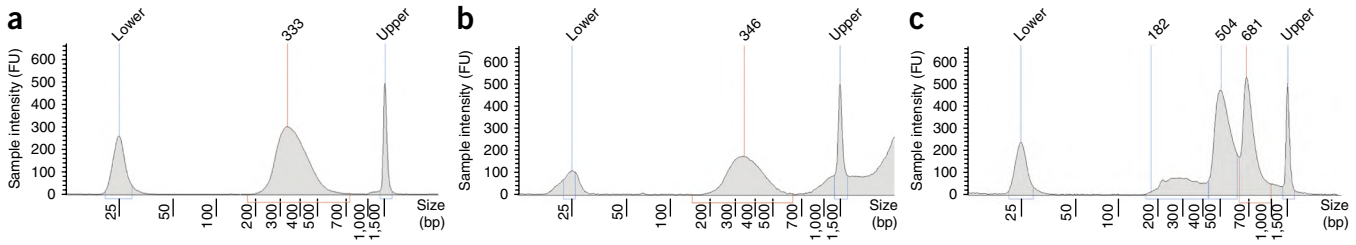


Figure 6 | Example Agilent TapeStation 2200 electropherograms. (**a**) Electropherogram of an optimal post-PCR sample at Step 47. (**b**) Electropherogram from Step 47 showing higher molecular species resulting from too many PCR cycles. See Experimental design for details on determining the number of PCR cycles. (**c**) Electropherogram of the postligation at Step 36; note that the double peaks are normal. The peaks can vary in size and intensity without affecting the final results.

specific plateau point will be dependent on the amount of DNA used in the PCR and should be carefully monitored.

If a sample requires >10 fmol of DNA, such as when performing targeted capture of nuclear DNA, we have found that the SYBR Green signal can be unreliable. As a benchmark for determining the number of cycles that should be used during PCR, we have found that samples requiring ~600 fmol of total input DNA should be stopped between cycles six and eight. Importantly, for every twofold increase or decrease in the amount of DNA used in the PCR, the reactions should be stopped one cycle earlier or later, respectively. When first setting up DS for a new experimental system or genomic target, we recommend testing the number of cycles needed to maximize yield while avoiding higher-molecular-weight products by pausing the thermocycler every two cycles and carefully removing a small aliquot from the reaction. The aliquots can then be quantified on an Agilent TapeStation 2200 or Bioanalyzer 2100 to determine the optimal cycle number (**Fig. 6a**,**b**).

Targeted DNA capture. Targeted capture of the sequences of interest is often desirable. We have used the Agilent SureSelectXT target enrichment system. We have designed and used several custom-designed targeted capture sets against both nuclear and mtDNA from a number of organisms, including humans, mice and flies, with good success. Hybridization-based enrichment methods offered by other manufacturers would probably work as well. However, we have not evaluated these technologies, and they are likely to require some optimization. Of note, PCR-based enrichment methods are not compatible with DS because the targeted duplex DNA is melted into its single-stranded components during PCR, and the complementarity of the original DNA is lost.

Capture is readily performed on large targets, or on high-copy-number targets such as mtDNA. However, capture is inefficient when targeting small, low-copy-number regions. The reason is that capture typically results in a 10,000-fold enrichment of the target DNA. For example, the mitochondrial genome, owing to its high copy number, composes ~0.2% of the total mass of DNA in a human cell. Therefore, a 10,000-fold enrichment will result in nearly 100% purity of the targeted DNA. However, a 20-kb locus in genomic DNA, which is present in a single copy per haploid genome, composes only 0.0007% of the total DNA. Thus, a 10,000-fold enrichment will result in only 7% of the final DNA being on-target. However, we have found that repeating the targeted capture steps typically results in >85% of reads mapping to the target sequence.

Quality metrics. The overall efficiency of DS is calculated by dividing the number of DCS reads by the number of raw reads, and in our hands it can range from 1 to 10%. We have found that two primary factors determine overall data yield: the peak family size and the SSCS:DCS ratio. The peak family size is determined by the amount of DNA used for the PCR and the fraction of a lane dedicated to a sample, and it should optimally be adjusted to six, as discussed previously (see the 'Setting tag family size by PCR amplification' section; **Fig. 5a**). Should a sample exhibit a peak family size that is smaller than what is desired, then an optimal peak family size can be obtained by resequencing the same sample using the information in **Supplementary Table 2** and combining the data with the previous run(s). Importantly, the resequenced sample must be a technical replicate; performing the protocol again on the same biological sample will not work. The SSCS:DCS ratio is the other key factor. A ratio of two is the theoretically ideal and would indicate that every SSCS can find its partner and form a DCS. We typically obtain SSCS:DCS ratios between four and ten, with an average of ~6. An excessively high ratio (i.e., >14–15) can occur for several reasons. If peak family sizes are too small, then families with at least three members might not form from both DNA strands. This issue is readily overcome by increasing the family size, as noted above, although it does not appreciably improve above peak sizes of 5–6 (**Fig. 5b**). If a poor ratio is seen despite adequate family sizes, this suggests that only one of the two DNA strands was successfully ligated and amplified during PCR. In such cases, the sample library will need to be remade in order to obtain more data, with attention to optimizing ligation conditions and library tailing steps.

Finally, mutations should theoretically be distributed throughout all possible read positions; however, we observe an increased frequency of mutations at the 5′ and 3′ends of reads. These artifacts are likely due to errors that occur during the end-repair process and alignment. These artifacts can be mitigated by performing local re-alignment with GATK, followed by clipping the first and last five bases of each read[39].

© 2014 Nature America, Inc. All rights reserved.

MATERIALS

REAGENTS

- DNA of interest ▲ CRITICAL It must be purified under non-strand-denaturing conditions.
- Agencourt AMPure XP beads (Beckman Coulter, cat. no. A63880) Alternatively, Agencourt RNAclean XP (Beckman Coulter, cat. no. A63987) may be used
- Klenow fragment (3′→5′exo⁻; New England Biolabs (NEB) cat. no. M0212L, includes NEB buffer #2)
- 10× TBE electrophoresis buffer (Sigma-Aldrich, cat. no. T4415)
- TE_{low} (Affymetrix, cat. no. 75793)
- Nuclease-free water (Life Technologies, cat. no. AM9937)
- T4 polynucleotide kinase (NEB, cat. no. M0201; includes reaction buffer)
- HpyCH4III (NEB, cat. no. R0618; includes CutSmart buffer)
- dNTPs, 2.5 mM each (Promega, cat. no. U1511)
- Ethanol (100%, 200 proof; Sigma-Aldrich, cat. no. E7023) ! CAUTION Ethanol is flammable. Keep it away from flames.
- Sodium acetate (Sigma-Aldrich, cat. no. S2889)
- γ-^{32}P-ATP (6,000 Ci/mmol; MP Biomedicals, cat. no. 0135001U01) ! CAUTION γ-^{32}P-ATP is radioactive. Take appropriate care when handling it.
- Urea (Sigma-Aldrich, cat. no. U6504)
- Acrylamide (Supplied as a 19:1 30% (wt/vol) acrylamide/bis-acrylamide solution; Bio-Rad, cat. no. 161-0154) ! CAUTION Acrylamide is a known neurotoxin; handle it with care.
- Denaturing gel loading buffer (Life Technologies, cat. no. AM8546G)
- DS adapter oligonucleotides (IDT; see **Table 3** for sequences) ▲ CRITICAL Oligonucleotides must be PAGE purified.
- Sequencing adapter PCR oligonucleotide primers (IDT; see **Table 3** for sequences
- NEBNext end-repair module (NEB, cat. no. E6050L; includes reaction buffer)
- NEBNext dA-tailing module (NEB, cat. no. E6053L; includes reaction buffer)
- Ultrapure T4 ligase (600,000 U/ml; Enzymatics, cat. no. L6030-HC-L; supplied with 10× Ultrapure ligation buffer)
- KAPA HiFi HotStart DNA polymerase (1 U/ml; KAPA Biosystems, cat. no. KK2502; supplied with 5× fidelity buffer and 10 mM dNTP mix)
- SYBR Green (10,000×; Life Technologies, cat. no. S-7563)
- Agilent DNA high-sensitivity D1K TapeStation kit (Agilent, cat. no. 5067-5363 and 5067-5364). Alternatively, an Agilent high-sensitivity DNA for the Bioanalyzer 2100 may be used (Agilent, cat. no. 5067-4626)
- Agilent SureSelectXT target enrichment Set (protocol version 1.6; Agilent) ▲ CRITICAL The use of target enrichment is optional for this protocol. The target enrichment section uses Agilent SureSelectXT. The use of other targeted capture products will probably work; however, they have not been evaluated and their compatibility is unknown.
- SureSelectXT reagent kit, HSQ, 16 (Agilent, cat. no. G9611A) This item is optional and is needed only if targeted capture is performed

EQUIPMENT

- Eight-well PCR strip tubes (0.2 ml; BioExpress, cat. no. T-3035-1)
- DynaMag-96 side magnetic plate separator (Life Technologies, cat. no. 12331D)
- Vacuum centrifuge
- Microcentrifuge tubes (1.7 ml, Eppendorf)
- qPCR thermocycler (Bio-Rad)
- qPCR-compatible eight-well PCR strip tubes (0.2 ml; Bio-Rad, cat. no. TLS0851)
- qPCR-compatible eight-well PCR strip caps (Bio-Rad, cat. no. TCS0803)
- Microcentrifuge (Eppendorf, cat. no. 5424 000.410)
- Heating block (37 °C)
- Heating stir plate
- Covaris Sonicator S220 (Covaris)
- Sonication tubes (Covaris, cat. no. 520045)
- Agilent TapeStation 2200 (Agilent, cat. no. G2964AA) or Agilent Bioanalyzer 2100 (Agilent, cat. no. G2938C)
- Illumina HiSeq2500 and associated equipment (Illumina): NextSeq500, HiSeq1000 and HiSeq2000 are all compatible with this protocol. The MiSeq uses a substantially reduced number of clusters relative to the other Illumina platforms and, although not incompatible with this protocol, is not recommended. Non-Illumina sequencing platforms have not been evaluated and would require re-designing of the adapters
- Pipette tips
- BWA software package (http://bio-bwa.sourceforge.net/): the protocol is known to work with versions ≤0.6.2. The software distributed with this protocol has been designed for use with BWA; different aligners may not be compatible
- SAMtools software package (http://samtools.sourceforge.net/): the protocol is known to work with versions ≤0.1.18. The software distributed with this protocol has been designed for use with SAMtools; the use of other software to manipulate sam/bam files may not be compatible
- Python software package (http://python.org/): the scripts used in this protocol have been tested with v2.7.x. They are not currently compatible with Python 3.x
- BioPython software package (http://biopython.org/): the scripts used in this protocol are compatible with v1.62.
- Pysam software package (http://code.google.com/p/pysam/): the scripts used in this protocol are compatible with v0.7.5
- Genome Analysis Toolkit (GATK) software package (http://www.broadinstitute.org/gatk/): the scripts used in this protocol are known to be compatible with v2.4-9
- Picard software package (http://picard.sourceforge.net/): the scripts used in this protocol are known to be compatible with v1.107
- Distance software package: the scripts used in this protocol are compatible with v0.1.3.

TABLE 3 | Sequences of the adapter and PCR oligonucleotides.

Name	Sequence	Purpose
MWS51	5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′	Duplex sequencing adapter oligos. Annealed adapters prepared as described in **Box 1** and used in PROCEDURE Step 30
MWS55	5′-TCTTCTACAGTCANNNNNNNNNNNNNAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3′	
MWS13	5′-AATGATACGGCGACCACCGAG-3′	PCR primers used at PROCEDURE Steps 38 and 51 to generate amplified tag families, with (Steps 38B and 51) or without (Step 38A) targeted capture
MWS20	5′-GTGACTGGAGTTCAGACGTGTGC-3′	
MWS21	5′-CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTGC-3′	

Xs represent the sample indexing sequence. We recommend using the Illumina TruSeq index sequences available on the Illumina website.

© 2014 Nature America, Inc. All rights reserved.

REAGENT SETUP

Urea 8 M, polyacrylamide gel mixture 14% (wt/vol) Add 480 g of urea to 466 ml of acrylamide and 100 ml of 10× TBE. Allow the urea to dissolve by slowly stirring with mild heat using a stir plate. After the urea is dissolved, bring the volume up to 1 liter with ddH_2O. Store it at 4 °C for up to 1–2 months.

Sodium acetate solution, 3 M Add 24.6 g of sodium acetate and dissolve it in 90 ml of ddH_2O. Adjust the pH to 5.2 with 1 M HCl, and then add ddH_2O to adjust the final volume to 100 ml. Sterilize the solution by autoclaving. This solution can be stored indefinitely at room temperature (20–25 °C).

Ethanol solution, 75% (vol/vol) Add 25 ml of ddH_2O to 75 ml of 100% ethanol. This solution can be stored at room temperature for up to 1 month.

SYBR Green, 12.5× Dilute 1 µl of 10,000× SYBR Green in 799 µl of ddH_2O, and divide the solution into 50-µl aliquots; store the aliquots at −20 °C for up to 6 months.

Sequencing adapter oligonucleotides Dissolve the oligos in TE or ddH_2O in separate microcentrifuge tubes to a final concentration of 100 µM each. Immediately freeze the oligonucleotides at −20 °C until further use. Oligonucleotides can be stored at this temperature for 6–12 months.

Sequencing adapter PCR primers Dissolve each respective PCR primer in TE or ddH_2O in separate microcentrifuge tubes to a final concentration of 20 µM each. Immediately freeze the oligonucleotides at −20 °C until further use. Primers can be stored at this temperature for 6–12 months.

AMPure XP magnetic beads The beads are normally stored at 4 °C. Allow the beads to warm to room temperature before use. ▲ CRITICAL The beads will not function properly if they are not at room temperature.

EQUIPMENT SETUP

Covaris S220 Sonicator The following settings are used to shear nuclear DNA to ~300 bp (range 100–500 bp): duty cycle, 10%; intensity, 5; cycles/burst, 200; and time, 20 s × 6. For small genomes, such as mtDNA, viral or plasmid DNA, the following settings should be used: duty cycle, 10%; intensity, 5; cycles/burst, 100; and time, 20 s × 3. ▲ CRITICAL The water-bath temperature should be 4 °C, and the water bath should be degassed for 30 min before shearing.

© 2014 Nature America, Inc. All rights reserved.

PROCEDURE

Sonication of DNA ● TIMING 1 h

▲ CRITICAL The following procedure is presented for a single sample; however, the protocol can be scaled up for an arbitrary number of DNA samples.

1| For each library, dilute the DNA into a final volume of 130 µl of TE_{low}. For samples that do not use targeted capture, we typically start with 200–500 ng of DNA. For targeted capture, we typically start with 1–3 µg of DNA.

2| Transfer the DNA to a Covaris sonication tube and shear DNA using the settings outlined in Equipment Setup.
▲ CRITICAL STEP Ensure that there are no air bubbles in the bottom of the tube after loading the sample. In the case of air bubbles, gently tap or shake to bring the solution to the bottom of the well. In addition, the water bath temperature should be 4 °C, and the water bath should be degassed for 30 min before shearing.

3| Vortex the room-temperature AMPure XP bead mixture to resuspend any magnetic particles that may have settled.
▲ CRITICAL STEP The beads must be at room temperature before use. A cold bead mixture significantly reduces yield.

4| Split the 130 µl of sonicated DNA into 2 × 65-µl aliquots in 0.2-ml PCR tubes.

5| Add 130 µl of AMPure XP beads to each PCR tube (2:1 bead:sample ratio) and mix well by vortexing or by pipetting up and down. The 2:1 ratio is used to ensure the maximal retention of all DNA.

6| Incubate the mixture for 5 min at room temperature.

7| Place PCR tubes onto the DynaMag-96 side magnetic plate separator (or equivalent) for at least 2 min or until all the beads are out of solution. Visually confirm that the beads have moved to the side of the tubes and that the solution is clear.

8| Aspirate the supernatant from each tube and discard it. Keep the tubes on the magnetic plate separator.

9| Carefully dispense 200 µl of room-temperature 75% (vol/vol) ethanol to each PCR tube and incubate the tubes for 30 s at room temperature.

10| Aspirate out the ethanol and discard it. Repeat Steps 9 and 10 once more for a total of two washes.

11| Allow the beads to dry for 5 min at room temperature until there are no visible traces of ethanol.
▲ CRITICAL STEP Overdrying or underdrying the beads reduces yield.

12| Remove the PCR tubes from the magnetic plate separator and add 40 µl of ddH_2O to only one of the two tubes of each sample and resuspend.

13| Carefully transfer the resuspended bead mixture to the second sample PCR tube and resuspend the beads. Let the bead mixture sit for at least 2 min.

14| Place the tube containing the combined sample back on the magnetic plate separator for 2 min or until the supernatant is clear. Transfer the supernatant to a new 0.2-ml PCR tube. Discard the beads.
■ **PAUSE POINT** Samples can be stored at –20 °C for several weeks.

End repair of sonicated DNA and size selection ● TIMING 1 h
15| Combine the components in the table below in a 0.2-ml PCR tube and mix them carefully by pipetting up and down. Incubate the mixture in a thermocycler for 30 min at 20 °C.

Reagent	Volume (μl)	Final concentration
DNA from Step 14	40	Variable
10× NEBNext end-repair buffer	5	1×
NEBNext end-repair enzyme Mix	5	N/A

16| Add 35 μl of AMPure XP beads (0.7:1 bead:sample ratio) and incubate the tube at room temperature for 5 min. This selects against large (>500 bp), poorly sheared DNA fragments.

17| Place the tube containing the end-repaired DNA onto the magnetic plate separator for at least 2 min or until all the beads are out of solution, and then transfer the supernatant (should be 85 μl) to a new 0.2-ml PCR tube. Discard the beads.

18| Add 55 μl of AMPure XP beads to the sample and incubate the tube at room temperature for 5 min. Place the PCR tube onto the magnetic plate separator for at least 2 min or until all the beads are out of solution, and discard the 140-μl supernatant. Keep the beads.

19| Repeat the double 75% (vol/vol) ethanol wash outlined in Steps 9–11.

20| Remove the PCR tube from the magnet and add 42 μl of ddH_2O to each sample and gently pipette up and down until the beads are resuspended. Let the bead mixture sit for at least 2 min.

21| Place the PCR tube back onto the magnetic plate separator for at least 2 min or until all the beads are out of solution, and then transfer the supernatant to a new 0.2-ml PCR tube.
■ **PAUSE POINT** Samples can be stored at –20 °C for up to several weeks.

3′ A-tailing of blunt-ended DNA library ● TIMING 1 h
22| Combine the components in the table below in a 0.2-ml PCR tube and mix them carefully by pipetting up and down. Incubate the mixture in a thermocycler for 30 min at 37 °C.

Reagent	Volume (μl)	Final concentration
DNA from Step 21	42	Variable
10× NEBNext dA-tailing buffer	5	1×
NEBNext Klenow^{exo-}	3	

23| Add 60 μl of AMPure XP beads (1.2:1 bead:sample ratio) and mix. The bead ratio selects against DNA fragments <150 bp in size. Incubate the tube for 5 min at room temperature.

24| Place A-tailed DNA sample onto the magnetic plate separator for at least 2 min or until all the beads are out of solution, and then remove and discard the supernatant.

© 2014 Nature America, Inc. All rights reserved.

25| Repeat the double 75% (vol/vol) ethanol wash outlined in Steps 9–11.

26| Remove the tube from the magnetic plate separator and resuspend the beads in 37 µl of ddH_2O. Let the bead mixture sit for at least 2 min.

27| Place the sample back onto the magnetic plate separator for at least 2 min or until all the beads are out of solution, and then transfer the 37 µl of eluent to a new 0.2-ml PCR tube.
■ **PAUSE POINT** Samples can be stored at –20 °C for several weeks.

28| Remove 2 µl from the A-tailed library and quantify the amount of DNA using the high-sensitivity kits for either the Agilent TapeStation 2200 or Agilent Bioanalyzer 2100, according to the manufacturer's instructions.
? TROUBLESHOOTING

Ligation of DS adapters to sample DNA ● TIMING 1 h
29| By using the amount of sample DNA determined in Step 28, calculate the amount of DS adapters (synthesized as described in **Boxes 1** and **2**) that will be needed for a 20:1 molar excess relative to the total DNA concentration in the final ligation reaction. If necessary, dilute the adapters in ddH_2O to an easily pipetteable volume.

30| Mix the following components, in order, in a 0.2-ml PCR tube, and mix them carefully by pipetting up and down. As noted in Step 29, the volumes of adapters and water that are used will need to be adjusted on the basis of the DNA concentration of your sample, but the total volume should be 60 µl.

Reagent	Volume (µl)	Final concentration
DNA from Step 27	35	Variable
10× Ultrapure ligation buffer	6	1×
Duplex Sequencing adapters (50 µM)	Variable	Variable
ddH_2O	Variable	N/A
Ultrapure T4 ligase (600 U/µl)	6	60 U/µl
Total volume	60	

31| Incubate the tube for 15 min at 25 °C in a thermocycler (no heated lid).

32| Add 60 µl of AMPure beads (1.2× bead:sample ratio) and mix. This ratio selects against DNA fragments <150 bp in size. Incubate the tube for 5 min at room temperature.

33| Place the PCR tube containing the ligated sample onto the magnetic plate separator for at least 2 min or until all the beads are out of solution. Remove and discard the supernatant.

34| Repeat the double 75% (vol/vol) ethanol washes outlined in Steps 9–11, with the exception that the beads be fully resuspended in the ethanol by vortexing before being placed on the magnetic separator. Dry the beads for 5 min at 37 °C in a heat block with the tube cap opened.
▲ **CRITICAL STEP** Resuspending the magnetic beads during the ethanol washes is essential for removing adapter dimers that form during ligation. Failure to do so will adversely affect the downstream PCR steps. Removal of any residual ethanol by drying at 37 °C is also important, as ethanol inhibits PCR.

35| Remove samples from the magnetic separator and add 30 µl of TE_{low} to each sample. After 2 min, place the samples back on the magnetic separator for 2 min or until the supernatant is clear. Transfer the 30 µl of supernatant to a new 0.2-ml PCR tube.
■ **PAUSE POINT** Samples can be stored at –20 °C for several weeks.

© 2014 Nature America, Inc. All rights reserved.

PCR amplification ● **TIMING 1 h**

36| Quantify the molar amount of adapter-ligated DNA library from Step 35 on an Agilent TapeStation 2200 or Bioanalyzer 2100. The ligation reaction can result in multiple peaks; we quantify all DNA seen between 200 and 900 bp (**Fig. 6c**).
? TROUBLESHOOTING

37| Determine the number of attomoles of input DNA required for PCR on the basis of target size and lane fraction. For noncapture experiments, we use 40 amol of input for every 8 million paired-end reads. For capture of mtDNA, we use ~2 fmol of total DNA for the same number of reads. For nuclear DNA capture (20 kb target size), we use ~600 fmol of total DNA per 8 million reads. The formula for estimating the total DNA input is

$$\text{Total DNA} = \frac{mG}{NT}$$

where m is the desired amount of target DNA in attomoles of fragments, G is the size of total genome in base pairs, T is the size of the target DNA in base pairs and N is the number of copies of the target per copy of the nontarget portion of the genome. (Refer to the INTRODUCTION and **Table 2** for further details.)

38| Generate amplified tag families by qPCR. To generate tag families for small homogeneous genomes such as mtDNA or plasmids that do not require targeted capture, follow option A. To generate duplicate families of genomic targets that require targeted capture, follow option B.

(A) Generation of duplicate families without targeted capture

(i) Mix the components in the table below in a qPCR-compatible PCR tube, and mix carefully by pipetting up and down. The volumes of DNA and water that are used will need to be adjusted on the basis of the DNA concentration and the desired input amount of the sample, but the total volume should be 50 µl. See **Table 3** for MWS13 and MWS21 primer sequences.

Reagent	Volume (µl)	Final concentration
ddH_2O	Variable	N/A
5× KAPA fidelity buffer	10	1×
10 mM dNTP mix	1.5	0.3 mM
20 µM Primer MWS13	1	0.4 µM
20 µM Primer MWS21	1	0.4 µM
12.5× SYBR Green	1	0.25×
DNA (From Step 37)	Variable	Variable
KAPA HiFi DNA polymerase (1 U/µl)	1	0.02 U/µl
Total volume	50	

(ii) Incubate the mixture in a qPCR thermocycler as follows:

Cycle	Denature	Anneal	Extend
1	95 °C for 4 min	—	—
2–variable	98 °C for 20 s	60 °C for 20 s	72 °C for 15 s

▲ **CRITICAL STEP** The number of PCR cycles is a crucial variable; a difference of a single cycle can result in nonspecific, higher-molecular-weight products appearing (**Fig. 6b**). See 'PCR Amplification' in the INTRODUCTION for details on determining the number of PCR cycles.

© 2014 Nature America, Inc. All rights reserved.

(B) Generation of duplicate families for targeted DNA capture

(i) Dilute the DNA from Step 37 to a final volume of 30 µl with ddH_2O (see 'PCR Amplification' in the INTRODUCTION and **Table 2**). We have found that splitting the input DNA into three equal aliquots and performing a separate PCR on each sample aliquot provides enough PCR product for targeted capture (each PCR will contain one-third of the total input DNA).

(ii) For each PCR, mix the components in the tables below in a qPCR-compatible PCR tube and mix carefully by pipetting up and down (see **Table 3** for MWS13 and MWS20 primer sequences):

Reagent	Volume (µl)	Final concentration
ddH_2O	24.5	N/A
5× KAPA fidelity buffer	10	1×
10 mM dNTP mix	1.5	0.3 mM
20 µM primer MWS13	1	0.4 µM
20 µM primer MWS20	1	0.4 µM
12.5× SYBR Green	1	0.25×
DNA (from Step 38B(i))	10	Variable
KAPA HiFi DNA polymerase (1 U/µl)	1	0.02 U/µl

▲ **CRITICAL STEP** It is important that the nonindexed primer MWS20 be used and not MWS21.

(iii) Incubate the mixture in a qPCR thermocycler as follows:

Cycle	Denature	Anneal	Extend
1	95 °C for 4 min	—	—
2–Variable	98 °C for 20 s	60 °C for 20 s	72 °C for 15 s

▲ **CRITICAL STEP** The number of PCR cycles is a crucial variable; a difference of a single cycle can result in nonspecific, higher-molecular-weight products appearing (**Fig. 6b**). See 'PCR Amplification' in the INTRODUCTION for details on determining the number of PCR cycles.

39| Transfer and pool all the PCRs from a single sample into a single 1.7-ml microcentrifuge tube.

40| Add 1.0 volumes of AMPure XP beads to the pooled PCRs (typically 50 µl of beads for every 50 µl of PCR). The bead ratio selects against DNA fragments <200 bp in size. Incubate the mixture for 5 min at room temperature.

41| Place the PCR sample onto a magnetic separator for at least 2 min or until all the beads are out of solution, and then remove and discard the supernatant.

42| Remove the beads from the magnetic plate and resuspend them in 200 µl of room-temperature 75% (vol/vol) ethanol by vortexing.

▲ **CRITICAL STEP** Failure to completely resuspend the beads during this step can lead to the retention of nonspecific PCR products (**Supplementary Fig. 6**).

43| Place the microcentrifuge tube onto a magnetic separator for at least 2 min or until all the beads are out of solution, and then remove and discard the supernatant.

44| Repeat Steps 42 and 43. Let the beads dry for 5 min.

45| Remove the microcentrifuge tube from the magnetic plate separator and resuspend the beads in 20 µl of ddH_2O. Let the bead mixture sit for at least 2 min.

© 2014 Nature America, Inc. All rights reserved.

46| Place the sample onto the magnetic separator for at least 2 min or until all the beads are out of solution, and then transfer the 20 µl of eluent to a new 0.2-ml PCR tube.
■ **PAUSE POINT** Samples can be stored at –20 °C for several weeks.

47| Quantify the final bead-purified PCR products between 200 and 900 bp using an Agilent TapeStation 2200 or Bioanalyzer 2100. If targeted capture is being performed, continue to Step 48; otherwise, the sample can be submitted for sequencing at this point.
? TROUBLESHOOTING

(Optional) Targeted DNA capture ● TIMING 24 h
48| Remove 200 ng of purified PCR product from Step 47 and lyophilize it completely using a vacuum centrifuge. Resuspend it in 1.7 µl of ddH_2O.
? TROUBLESHOOTING

49| Perform targeted DNA capture according to the Agilent SureSelectXT instructions (version 1.6). Alternatively, we routinely perform the targeted capture at 0.25× volume reduction, relative to those recommended by the manufacturer. The lower volume capture greatly reduces cost, with no adverse effects noted. The 0.25× capture follows the standard Agilent SureSelectXT protocol with all reagent volumes reduced by three-quarters. The final capture DNA should be eluted in 30 µl of ddH_2O.

50| PCR-amplify the 30 µl of captured sample by mixing the components in the table below in a qPCR-compatible PCR tube and mix carefully by pipetting up and down (see **Table 3** for MWS13 and MWS21 primer sequences):

Reagent	Volume (µl)	Final concentration
ddH_2O	5	N/A
5× KAPA fidelity buffer	10	1×
10 mM dNTP mix	1	0.2 mM
20 µM primer MWS13	1	0.4 µM
20 µM primer MWS21	1	0.4 µM
12.5× SYBR Green	1	0.25×
Captured DNA (from Step 49)	30	Variable
KAPA HiFi DNA polymerase (1 U/µl)	1	0.02 U/µl

▲ **CRITICAL STEP** For DNA prepared using target capture, it is essential to use all the DNA from the targeted capture to maintain sample diversity.

51| Incubate the mixture in a qPCR thermocycler as follows:

Cycle	Denature	Anneal	Extend
1	95 °C for 4 min	—	—
2–30	98 °C for 20 s	60 °C for 20 s	72 °C for 15 s

▲ **CRITICAL STEP** The number of PCR cycles is a crucial variable; a difference of a single cycle can result in nonspecific, higher-molecular-weight products appearing. We monitor the reaction by the SYBR Green signal, and we remove the tube one or two cycles before the signal plateaus. The SYBR Green signal typically begins to plateau between cycles 15 and 22.

52| Purify each completed PCR by repeating Steps 40–44.
■ **PAUSE POINT** Samples can be stored at –20 °C for several weeks.

53| Quantify the final bead-purified PCR products between 200 and 900 bp using an Agilent TapeStation 2200 or Bioanalyzer 2100 and submit for sequencing.
? TROUBLESHOOTING

© 2014 Nature America, Inc. All rights reserved.

Bioinformatic processing: prepare a reference genome for use with BWA and SAMtools ● TIMING variable

54| First, download and decompress the genome of interest. For the human nuclear and mitochondrial genomes, we use v37 of the human reference genome from the 1,000 Genomes Project. This reference genome can be obtained at ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/human_g1k_v37.fasta.gz. Alternatively, the hg19 reference genome from the University of California, Santa Cruz (UCSC) Genome Browser can be used.

55| The reference genome must be indexed before being used with BWA. Follow the instructions on the website (http://bio-bwa.sourceforge.net/bwa.shtml) using the recommended options relevant to the genome of interest.
▲ **CRITICAL STEP** The bioinformatics steps outlined in this protocol are specific to BWA. We have not used other sequencing aligners, and their compatibility with the downstream bioinformatics steps is not known.

Bioinformatic processing: parsing and filtering duplex tags ● TIMING variable

56| Concatenate the 12-nt tag sequences from the paired reads and evaluate for tag quality using 'tag_to_header.py' using the following command:

```
python tag_to_header.py --infile1 read_1.fq --infile2
read_2.fq --outfile1 read_1.fq.smi --outfile2
read_2.fq.smi --barcode_length 12 --spacer_length 5
```

This command results in two fastq files (one for each read) that are ready to be aligned. A flowchart schematic can be found in **Supplementary Figure 1**.
▲ **CRITICAL STEP** If samples are multiplexed within the same sequencing lane, then these samples should be sorted into separate files based on their index sequence before tag parsing.

57| Align each read to the reference genome:

```
bwa aln reference_genome.fasta read_1.fq.smi > read_1.aln
```

and

```
bwa aln reference_genome.fasta read_2.fq.smi > read_2.aln
```

58| Make a single paired-end .sam file:

```
bwa sampe -s reference_genome.fasta read_1.aln
read_2.aln read_1.fq.smi read_2.fq.smi > PE_reads.sam
```

59| Convert to .bam format and sort by position:

```
samtools view -Sbu PE_reads.sam | samtools sort - PE_reads.sort
```

Bioinformatics processing: making SSCSs ● TIMING variable

60| Run the Python program 'ConsensusMaker.py' to collapse PCR duplicates into SSCS:

```
python ConsensusMaker.py --infile PE_reads.sort.bam --
tagfile PE_reads.tagcounts --outfile SSCS.bam --min 3 --
max 1000 --cutoff 0.7 --Ncutoff 0.3 --readlength 84 --
read_type dpm --filt osn
```

A detailed flowchart schematic of the program can be found in **Supplementary Figure 2**.

61| Sort the SSCS reads:

```
samtools view -bu SSCS.bam | samtools sort - SSCS.sort
```

© 2014 Nature America, Inc. All rights reserved.

Bioinformatics processing: making DCSs ● TIMING variable

62| Construct DCSs from SSCSs using 'DuplexMaker.py'.

```
python DuplexMaker.py --infile SSCS.sort.bam --outfile
DCS_data --Ncutoff 0.3 --read_length 84
```

A detailed flowchart schematic of the program can be found in **Supplementary Figure 3**.

63| Align each DCS .fastq to the reference genome:

```
bwa aln reference_genome.fasta DCS_data_read_1.fq >
DCS_data_read_1.aln
```

and

```
bwa aln reference_genome.fasta DCS_data_read_2.fq >
DCS_data_read_2.aln
```

64| Make a paired-end .sam file for the DCS data:

```
bwa sampe -s reference_genome.fasta DCS_data_read_1.aln
DCS_data_read_2.aln DCS_data_read_1.fq
DCS_data_read_2.fq > DCS_PE.aln.sam
```

65| Convert to .bam format and sort by position:

```
samtools view -Sbu DCS_PE.aln.sam | samtools sort -
DCS_PE.aln.sort
```

Bioinformatics processing: prepare files for analysis ● TIMING variable

66| Index the final sorted DCS .bam file:

```
samtools index DCS_PE.aln.sort.bam
```

67| Filter out unmapped reads from the final DCS .bam file:

```
samtools view -F 4 -b DCS_PE.aln.sort.bam >
DCS_PE.filt.bam
```

68| Add readgroups field to the header of the final DCS .bam file with Picard to allow for compatibility with the GATK using Picard tools. An example command is provided below:

```
java -jar -Xmx2g AddOrReplaceReadGroups.jar
INPUT=DCS.filt.bam OUTPUT=DCS_PE.filt.readgroups.bam
RGLB=UW RGPL=Illumina RGPU=ATATAT RGSM=default
```

69| Index the final sorted DCS .bam file:

```
samtools index DCS_PE.filt.readgroups.bam
```

70| Perform local re-alignment of the reads using GATK. First identify the genome targets for local re-alignment. An example command is as follows:

```
java -Xmx2g -jar GenomeAnalysisTK.jar -T
RealignerTargetCreator -R reference_genome.fasta -I
DCS_PE.filt.readgroups.bam -o
DCS_PE.filt.readgroups.intervals
```

© 2014 Nature America, Inc. All rights reserved.

This command is followed by the actual local re-alignment. An example command is as follows:

```
java -Xmx2g -jar GenomeAnalysisTK.jar -T IndelRealigner
-R reference_genome.fasta -I DCS_PE.filt.readgroups.bam
-targetIntervals DCS_PE.filt.readgroups.intervals -o
DCS_PE.filt.readgroups.realign.bam
```

▲ CRITICAL STEP Alignment increases the occurrence of false mutations that occur at the 3′ and 5′ ends of the reads, so this step is highly recommended. In addition, local re-alignment must be done after the last alignment step.

71| Perform end-trimming of DCS reads. An example command that trims five bases from both the 3′ and 5′ ends of each read is provided:

```
java -Xmx2g -jar GenomeAnalysisTK.jar -T ClipReads -I
DCS_PE_reads.filt.readgroups.realign.bam -o DCS-
final.bam -R reference_genome.fasta --cyclesToTrim "1-
5,80-84" --clipRepresentation SOFTCLIP_BASES
```

▲ CRITICAL STEP Failure to clip the ends of reads increases the occurrence of false positives that occur during the enzymatic steps used during sequencing library preparation.

Bioinformatics analysis: quality metrics and basic mutation analysis ● TIMING variable

▲ CRITICAL The processed DCS .bam files can be used in any downstream analysis that is desired. We present the basic data analysis protocol that we currently use.

72| Plot the tag family size distribution (see **Fig. 4** as an example). The information for this plot is found in the 'PE_reads.tagstats' file created in Step 60. Plot the family size (column 1) on the *x* axis and the fraction of total reads (column 3) on the *y* axis.

? TROUBLESHOOTING

73| Make a pileup file from the final DCS reads using the following example command:

```
samtools mpileup -B -A -d 500000 -f
reference_genome.fasta DCS-final.bam > DCS-
final.bam.pileup
```

74| Count the number of unique mutations present in the final DCS sequences and calculate their frequencies:

```
cat DCS-final.bam.pileup | python CountMuts.py -d 100 -c 0 -C
0.01 -u > DCS-final.bam.pileup.countmuts
```

The optional argument -u invokes unique mutation counting; if it is left out, all mutations are counted, instead of only counting each type of mutation at each position once. The argument -d specifies that a minimum depth of 100 is required to score a position, and -C indicates that only mutations present at a clonality of <1% are counted. The output file lists the number and frequency of mutations with Wilson confidence intervals for each type of mutation, as well as total mutation frequency.

75| Locate the genomic position of each mutation:

```
python mut-position.py -i DCS-final.bam.pileup -o DCS-
final.bam.pileup.mutpos -d 100 -C .01
```

This program outputs a tab-delimited file containing the following information for each genomic position in the reference genome: chromosome, position, wild-type base identity, depth, total number of mutations observed, number of nonreference Ts, number of nonreference Cs, number of nonreference Gs, number of nonreference As, number of nonreference insertions, and number of nonreference deletions.

© 2014 Nature America, Inc. All rights reserved.

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 4**.

TABLE 4 | Troubleshooting table.

Step	Problem	Possible reason	Solution
Box 1, steps 3 and 6	Sample becomes cloudy upon addition of ethanol	Precipitation of enzymes and buffer salts	Precipitate does not affect downstream steps or library preparation. Continue with the protocol
Box 1, Step 12	Adapters are not properly synthesized (bands in lane 4 of **Fig. 3a** do not match)	Adapters are not fully extended (**Fig. 3a**, lane 2 does not match)	Replace reagents, re-anneal oligonucleotides and perform the extension reaction again
		Adapters are incompletely cut (**Fig. 2a**, lane 3 does not match)	Add an additional 2 µl of HpyCH4III and incubate overnight at 37 °C. If necessary, replace enzyme and start the synthesis over again
		Restriction fragment is still present after final ethanol precipitation (**Fig. 3a**, band f still present in lane 4)	Perform ethanol precipitation again using the protocol outlined in step 5 of **Box 1**
28, 36	No DNA is detectable during quantification	Ethanol concentration used during the washes is too low	Remake the 75% (vol/vol) ethanol and start over
		Not enough DNA 1s used during library preparation	Repeat the library preparation with a larger amount of input DNA
47, 53	Dimerized adapter or nonspecific PCR products are present after cleaning with AMPure XP beads	Adapter dimers are not efficiently removed	Bring the volume up to 50 µl with ddH_2O and repeat the bead clean-up (Steps 40–44; **Supplementary Fig. 6b**)
	No DNA is detectable during quantitation or large-molecular-weight products are present	DNA concatemerization owing to too many PCR cycles	Run a titration of the PCR cycle number. Quantify the unpurified PCR on a TapeStation or Bioanalyzer to confirm that the expected product is present (see **Fig. 6a** for an example). Rerun the PCR steps as outlined in the PROCEDURE and remove the samples at the determined cycle
		Ethanol concentration used during the washes is too low	Remake the 75% (vol/vol) ethanol and repeat PCR
48	Not enough DNA is present to perform targeted capture	Not enough DNA is produced during PCR amplification	Perform additional PCRs under the same conditions to obtain additional DNA, pool all the reactions, and purify the DNA (Steps 38B–47)
72	Not enough depth is obtained and the peak family size is too small	Too much DNA is used during the initial PCR or too small a lane fraction was used	Sequence the sample again (do not re-do the PCR) and pool the raw sequencing reads prior to data analysis. See **Supplementary Table 2** to determine the lane fraction needed to obtain a peak family size of six
	Not enough depth is obtained and the peak family size is too large	Too little DNA is used during the initial PCR	Depth is directly proportional to the number of DCS reads, which is directly proportional to the number of DNA molecules used as input in the PCR. If the number of molecules was too low, the depth will be too low. Further resequencing will only increase the family size, not the depth. Divide your peak family size by six, which is the optimal peak family size. Then increase the DNA amount in the PCR proportionally. For example, if the obtained peak family size is 24, 24/6 = 4, 4× more DNA is needed in the PCR

© 2014 Nature America, Inc. All rights reserved.

● **TIMING**

Steps 1–14, sonication of DNA: 1 h
Steps 15–21, end repair of sonicated DNA and size selection: 1 h
Steps 22–28, 3′ dA-tailing of end-repaired DNA: 1 h
Steps 29–35, ligation of DS adapters to sample DNA: 1 h
Steps 36–47, PCR amplification: 1 h
Steps 48–53 (optional), targeted DNA capture: 24 h
Steps 54–75, bioinformatic processing: variable and dependent on computational resources
Box 1, synthesis of DS adapters: 2 d
Box 2, optional quality control steps for adapter synthesis using radiolabeled adapters and PAGE: 1 d

ANTICIPATED RESULTS

We have applied DS to a number of biological systems, including the M13mp2 phagemid[19], human mtDNA from frozen brain tissue[33] and human nuclear DNA from both frozen and formalin-fixed samples (M.J.P., E.J.F. and L.A.L.; unpublished results). We consistently observe a >50,000-fold reduction in the apparent mutation frequency in these samples, relative to conventional NGS methods. We have found that SSCS formation is capable of removing almost all sequencer-derived false positives, which account for >99% of all artifacts. However, this step is unable to remove artifacts arising from first-round PCR errors that have been propagated to all tag family members. The mutational spectrum of the SSCS typically shows a high frequency of G→T mutations, which is the result of fixation of oxidative damage occurring during the DNA purification and processing steps[19]. Formation of the DCS reads consistently removes these artifacts, as well as those from other damage events, and it results in a further 90–99+% reduction in mutational artifacts.

© 2014 Nature America, Inc. All rights reserved.

npg

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS We thank D. Crispin and A. Lawrence for testing the protocol as written and for helping to clarify several important steps of the protocol. We also thank A. Herr for helpful comments and discussion. Support for this research was provided by the following US National Institutes of Health grants from the National Institute on Aging (NIA) and the National Cancer Institute (NCI): NIA P01-AG001751, NCI P01-CA77852, NCI R01-CA160674 and NCI R01-CA102029 to L.A.L. S.R.K. was further supported by the Genetic Approaches to Aging Training grant (NIA T32-AG000057).

AUTHOR CONTRIBUTIONS S.R.K., M.W.S., J.J.S. and E.J.F. developed the original protocol. S.R.K., M.W.S. and B.F.K. developed the data analysis software. S.R.K., M.W.S., E.J.F., M.J.P., E.H.A., J.-C.S., K.J.K. and R.-A.R. optimized the protocol. S.R.K., M.W.S., J.J.S., B.F.K. and L.A.L. wrote the paper.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Glenn, T.C. Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* **11**, 759–769 (2011).
2. Loeb, L.A., Springgate, C.F. & Battula, N. Errors in DNA replication as a basis of malignant change. *Cancer Res.* **34**, 2311–2321 (1974).
3. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
4. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
5. Bolli, N. *et al.* Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat. Commun.* **5**, 2997 (2014).
6. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
7. Kreso, A. *et al.* Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science* **339**, 543–548 (2013).
8. Johnson, B.E. *et al.* Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189–193 (2014).
9. Schmitt, M.W., Prindle, M.J. & Loeb, L.A. Implications of genetic heterogeneity in cancer. *Ann. NY Acad. Sci.* **1267**, 110–116 (2012).
10. Liu, Z. *et al.* Patterns of diversifying selection in the phytotoxin-like scr74 gene family of *Phytophthora infestans*. *Mol. Biol. Evol.* **22**, 659–672 (2005).
11. Srivatsan, A. *et al.* High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. *PLoS Genet.* **4**, e1000139 (2008).
12. Holt, K.E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in *Salmonella typhi*. *Nat. Genet.* **40**, 987–993 (2008).
13. Loh, E., Salk, J.J. & Loeb, L.A. Optimization of DNA polymerase mutation rates during bacterial evolution. *Proc. Natl. Acad. Sci. USA* **107**, 1154–1159 (2010).
14. Budowle, B. & van Daal, A. Extracting evidence from forensic DNA analyses: future molecular biology directions. *Biotechniques* **46**, 339–340, 342–350 (2009).
15. Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A. & Paabo, S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* **29**, 4793–4799 (2001).
16. Knapp, M. & Hofreiter, M. Next-generation sequencing of ancient DNA: requirements, strategies and perspectives. *Genes* **1**, 227–243 (2010).
17. Carlson, C.A. *et al.* Decoding cell lineage from acquired mutations using arbitrary deep sequencing. *Nat. Methods* **9**, 78–80 (2012).
18. Besaratinia, A. *et al.* A high-throughput next-generation sequencing-based method for detecting the mutational fingerprint of carcinogens. *Nucleic Acids Res.* **40**, e116 (2012).
19. Schmitt, M.W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl. Acad. Sci. USA* **109**, 14508–14513 (2012).
20. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. USA* **108**, 9530–9535 (2011).
21. Hiatt, J.B., Patwardhan, R.P., Turner, E.H., Lee, C. & Shendure, J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat. Methods* **7**, 119–122 (2010).
22. McCloskey, M.L., Stoger, R., Hansen, R.S. & Laird, C.D. Encoding PCR products with batch-stamps and barcodes. *Biochem. Genet.* **45**, 761–767 (2007).
23. Jabara, C.B., Jones, C.D., Roach, J., Anderson, J.A. & Swanstrom, R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a primer ID. *Proc. Natl. Acad. Sci. USA* **108**, 20166–20171 (2011).
24. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).

25. Casbon, J.A., Osborne, R.J., Brenner, S. & Lichtenstein, C.P. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res.* **39**, e81 (2011).
26. Shiroguchi, K., Jia, T.Z., Sims, P.A. & Xie, X.S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl. Acad. Sci. USA* **109**, 1347–1352 (2012).
27. Zong, C., Lu, S., Chapman, A.R. & Xie, X.S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).
28. Wang, J., Fan, H.C., Behr, B. & Quake, S.R. Genome-wide single-cell analysis of recombination activity and *de novo* mutation rates in human sperm. *Cell* **150**, 402–412 (2012).
29. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single-nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
30. Acevedo, A., Brodsky, L. & Andino, R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**, 686–690 (2014).
31. Lou, D.I. *et al.* High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl. Acad. Sci. USA* **110**, 19872–19877 (2013).
32. Diegoli, T.M., Farr, M., Cromartie, C., Coble, M.D. & Bille, T.W. An optimized protocol for forensic application of the PreCR repair mix to multiplex STR amplification of UV-damaged DNA. *Forensic Sci. Int. Genet.* **6**, 498–503 (2012).
33. Kennedy, S.R., Salk, J.J., Schmitt, M.W. & Loeb, L.A. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. *PLoS Genet.* **9**, e1003794 (2013).
34. Yost, S.E. *et al.* Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic Acids Res.* **40**, e107 (2012).
35. Kerick, M. *et al.* Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med. Genomics* **4**, 68 (2011).
36. Spencer, D.H. *et al.* Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. *J. Mol. Diagn.* **15**, 623–633 (2013).
37. Li, H. & Durbin, R. Fast and accurate short-read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
38. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
39. McKenna, A. *et al.* The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
40. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

© 2014 Nature America, Inc. All rights reserved.

Corrigendum: Detecting ultralow-frequency mutations by Duplex Sequencing

Scott R Kennedy, Michael W Schmitt, Edward J Fox, Brendan F Kohrn, Jesse J Salk, Eun Hyun Ahn, Marc J Prindle, Kawai J Kuong, Jiang-Cheng Shen, Rosa-Ana Risques & Lawrence A Loeb

Nat. Protoc. **9**, 2586–2606 (2014); doi:10.1038/nprot.2014.170; corrected online 22 October 2014

In the version of this article initially published online, the sequence for the MWS21 oligonucleotide was incorrectly listed in Table 3 as 5′-CCAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTGC-3′. The second nucleotide in the sequence should be A, not C, and the correct sequence is 5′-CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGTTCAGACGTGTGC-3′. The error has been corrected in the PDF and HTML versions of this article.

EXHIBIT J

TWINSTRAND BIOSCIENCES

What was hidden is now **revealed.**

TwinStrand Duplex Sequencing™ Technology

LEARN MORE

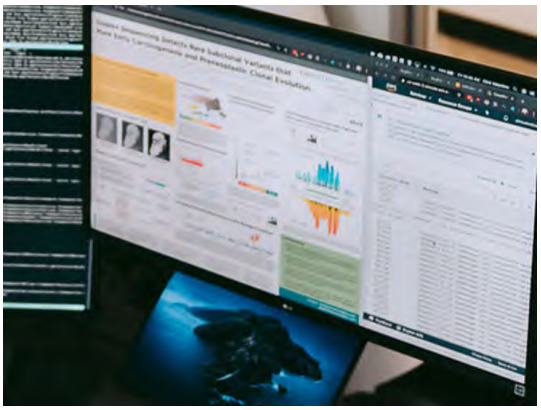
News



TwinStrand Biosciences Licenses Duplex Sequencing Technology to Foundation Medicine – 11/18/2020

TWINSTRAND BIOSCIENCES

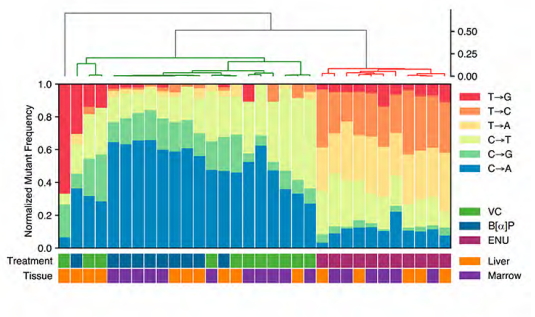
Virtual Events



American Society of Hematology (ASH 2020) December 5-8, 2020
Poster Presentation

SEE ALL EVENTS

Science



Direct Quantification of *In Vivo* Mutagenesis and

TWINSTRAND BIOSCIENCES

It's a simple but revolutionary idea—overcome the limitations of standard sequencing by independently tracking *both* strands of individual DNA molecules and comparing the results to eliminate errors.

Using a combination of proprietary biochemistry and cloud-based informatics, Duplex Sequencing greatly increases the resolution of next generation sequencing by reducing error rates from about *1-in-100 to 1-in-10 million*.

Now even the rarest variants can be detected with unparalleled sensitivity and unprecedented specificity.

A complete solution

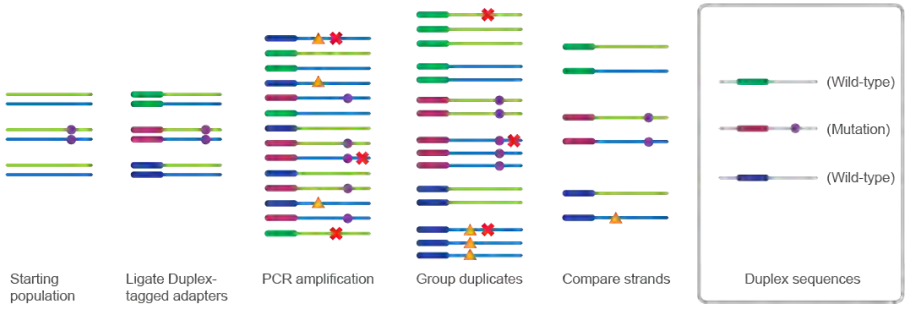
Overcoming obscurity in *5 Days*
TwinStrand Duplex Sequencing technology is an end-to-end solution, from pre-sequencing biochemistry to post-sequencing bioinformatics.



(click to expand)

Superior error correction

Using standard sequencing systems and an easy-to-adopt workflow, errors are reduced to below one artifactual mutation per ten million nucleotides.



(click to expand)

LEARN MORE

Imagine the possibilities

Duplex Sequencing can inform critical decisions in clinical medicine, public health, and other fields of science—on a faster time scale, where actions are most impactful.

Cancers can be detected and treated sooner. Drug resistance can be identified earlier. Residual diseases can be confirmed quickly. Mutagenetic chemicals can be recognized faster.

TWINSTRAND BIOSCIENCES

Put Duplex Sequencing to work for you.

To get started, click “contact us” below to tell us about your current work and to learn how our Duplex Sequencing technology will revolutionize your research.

CONTACT US

TwinStrand Biosciences, Inc.
3131 Elliott Ave, Suite 750 • Seattle, WA 98121 • USA
+1 (877) 202-TWIN

For Research Use Only. Not for use in diagnostic procedures.

©2020 TwinStrand Biosciences, Inc. All rights reserved. All trademarks are the property of TwinStrand Biosciences, Inc. or their respective owners. Legal Notices

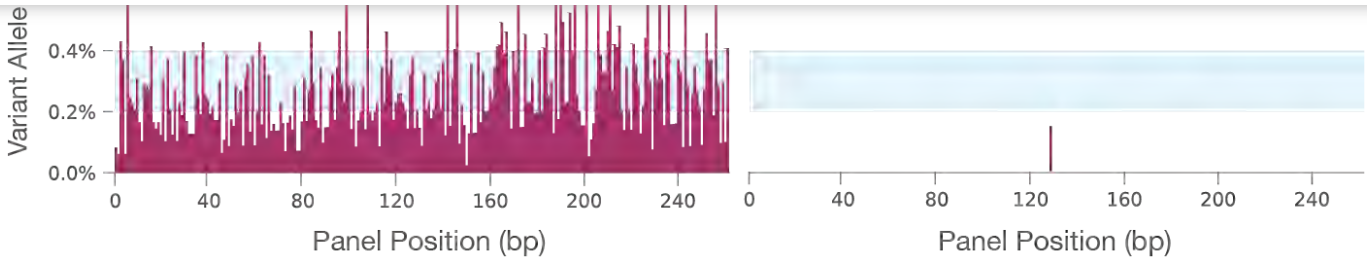
TWINSTRAND BIOSCIENCES

Duplex Sequencing Technology

Duplex Sequencing is an ultra-high accuracy sequencing method that overcomes the limitations of Next-Generation Sequencing by independently tracking both strands of individual DNA molecules. The paired sequences are compared to eliminate errors.

Duplex Sequencing Reveals True Mutations

- > 10,000 fold increase in accuracy over standard NGS
- Eliminates virtually all background errors, revealing ultra-low frequency variants
- Compatible with any Illumina® sequencer
- Easy to adopt with no special equipment needed

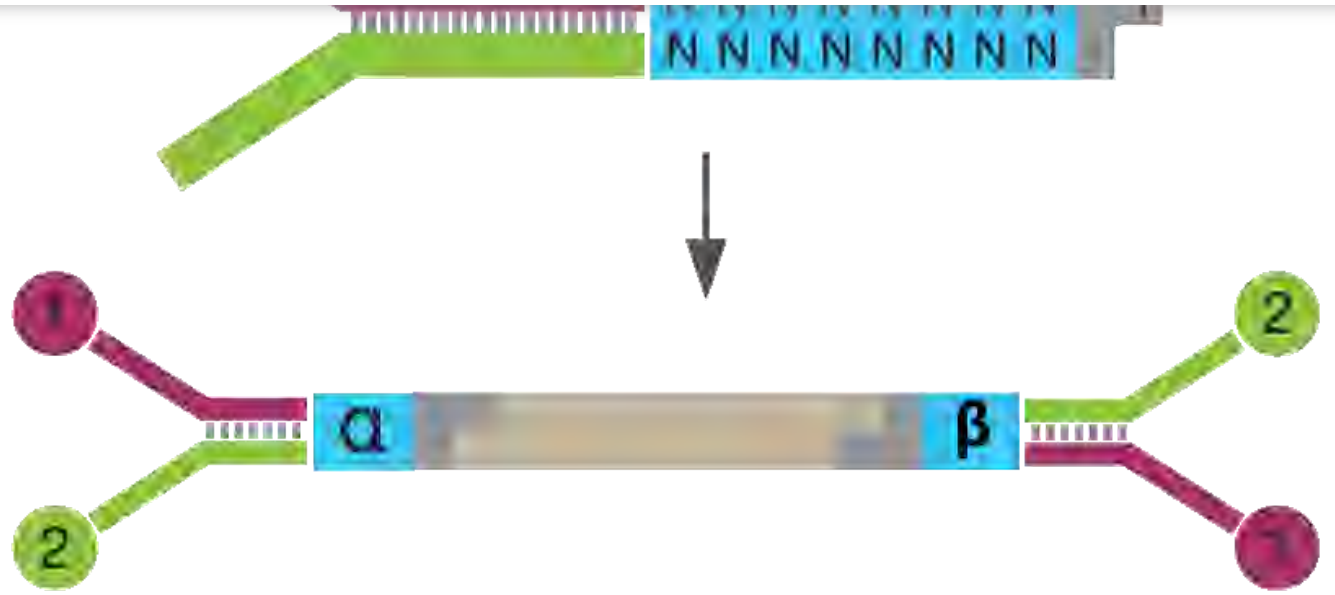


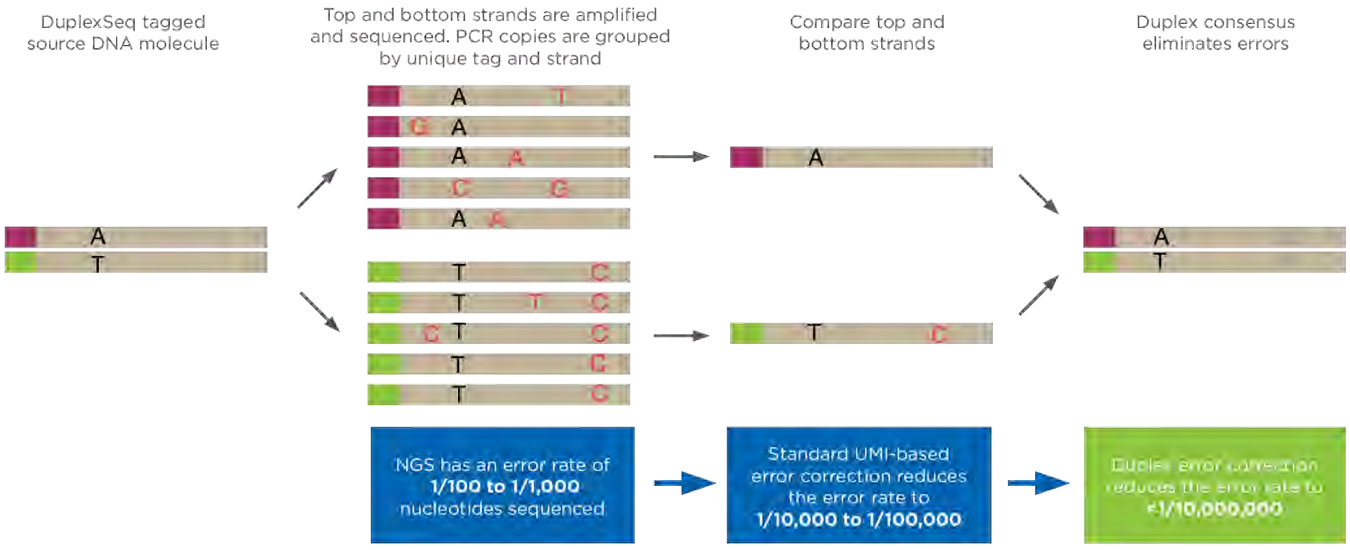
The same gene sequenced by standard Illumina® sequencing (left) and with Duplex Sequencing (right). With conventional methods, every position in the gene appears mutated in 0.1 – 1% of molecules sequenced. Duplex Sequencing eliminates the background noise, revealing the previously hidden true mutation.

Duplex Sequencing Eliminates Technical Noise

- Duplex Sequencing capitalizes on the naturally-occurring complementarity between strands of the DNA double-helix
- Each strand of each original molecule is uniquely labelled such that both can be tracked throughout amplification and sequencing for subsequent error correction

TWINSTRAND™ BIOSCIENCES





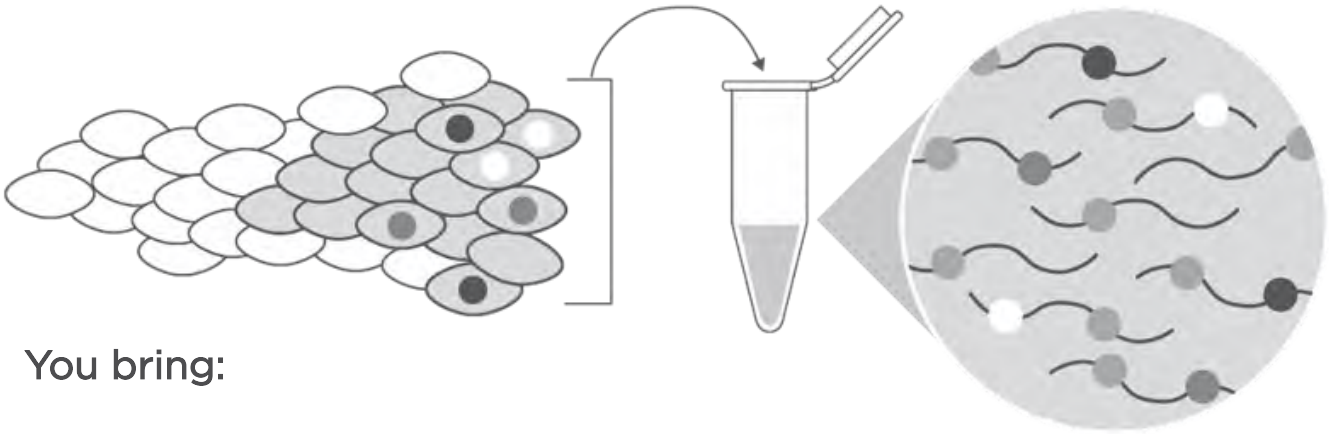
Our Solution

based workflow is familiar to and easily integrated into existing workflows.

Our platform includes access to a secure cloud-based bioinformatics pipeline. The pipeline is expertly tuned and optimized for accurate, easy-to-use, primary and secondary analysis.

TwinStrand Duplex Sequencing

TWINSTRAND™
BIOSCIENCES



You bring:

- DNA sample
- Water and ethanol

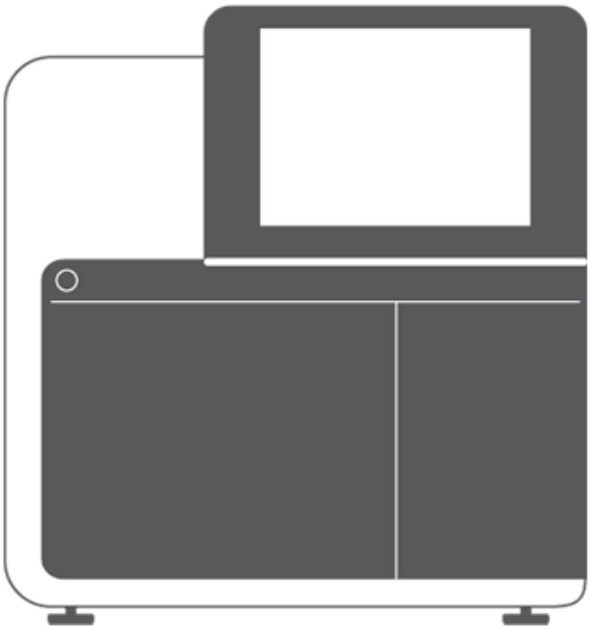




We pro library prep

- Duple optimized library p
- DuplexSeq™ adapters
- Dual unique indexes
- PCR reagents
- Hybrid capture reagents
- Magnetic beads
- TwinStrand DNA standards

Multi-box configuration facilitates best practices, separating pre- and post- PCR protocol steps.



You bring:

- Illumina® sequencer



We provide bioinformatics solution:

- Purpose-built for DuplexSeq data
- Easy-to-use, secure cloud-based software hosted on DNAnexus®
- Processes raw read files to produce duplex error-corrected sequences, variant calls, and assay quality metrics
- Industry standard formats compatible with existing annotation

Inputs: demultiplexed FASTQ files from any Illumina® platform

Outputs: error corrected BAM, VCF, and CSV files

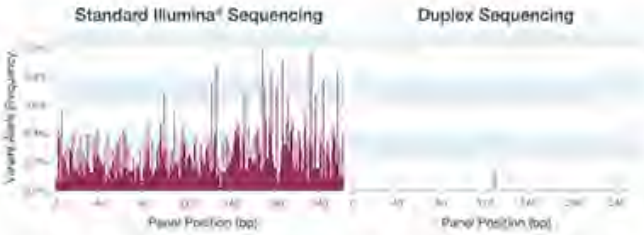

TWINSTRAND
BIOSCIENCES

TWINSTRAND™ BIOSCIENCES

Duplex Sequencing is an ultra-high accuracy sequencing method that overcomes the limitations of Next-Generation Sequencing by independently tracking both strands of individual DNA molecules. The paired sequences are compared to eliminate errors.

Duplex Sequencing Reveals True Mutations

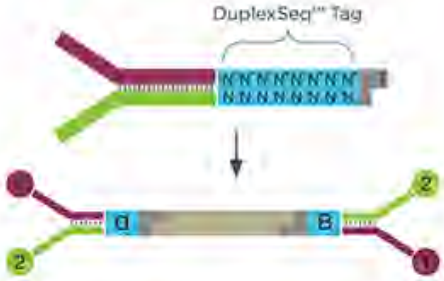
- > 10,000-fold increase in accuracy over standard NGS
- Eliminates virtually all background errors, revealing ultra-low frequency variants
- Compatible with any Illumina® sequencer
- Easy to adopt with no special equipment needed

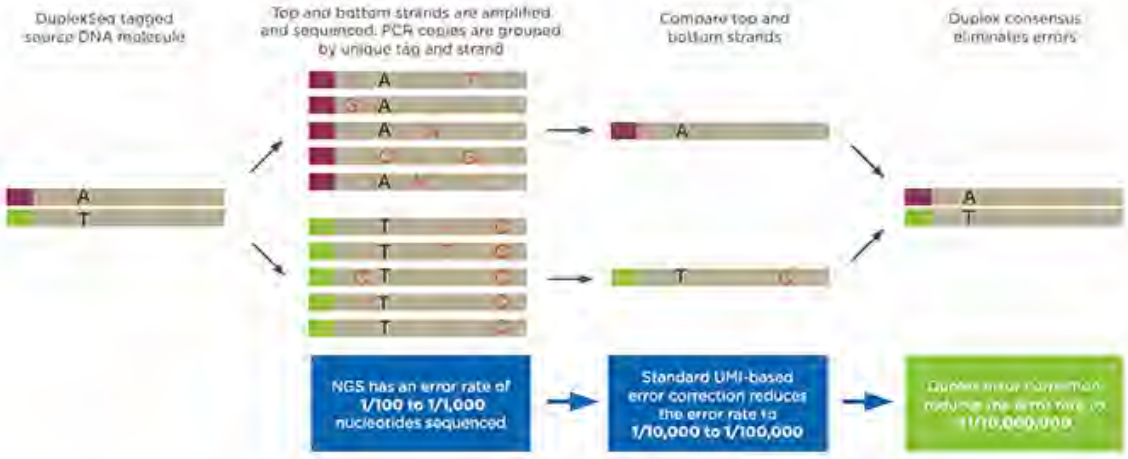


The same gene sequenced by standard Illumina® Sequencing (left) and with Duplex Sequencing (right). With conventional methods, every position in the gene appears mutated in 0.1-1% of molecules sequenced. Duplex Sequencing eliminates the background noise, revealing the previously hidden true mutation.

Duplex Sequencing Eliminates Technical Noise

- Duplex Sequencing capitalizes on the naturally-occurring complementarity between strands of the DNA double-helix
- Each strand of each original molecule is uniquely labelled such that both can be tracked throughout amplification and sequencing for subsequent error correction





TWINSTRAND BIOSCIENCES

AML MRD

MUTAGENESIS

COLLABORATION

PUBLICATIONS

Want to learn more or try TwinStrand Duplex Sequencing Technology?

CONTACT US

TwinStrand Biosciences, Inc.
3131 Elliott Ave, Suite 750 • Seattle, WA 98121 • USA
+1 (877) 202-TWIN

For Research Use Only. Not for use in diagnostic procedures.

©2020 TwinStrand Biosciences, Inc. All rights reserved. All trademarks are the property of TwinStrand Biosciences, Inc. or their respective owners. Legal Notices


TWINSTRAND
BIOSCIENCES

TWINSTRAND™ BIOSCIENCES

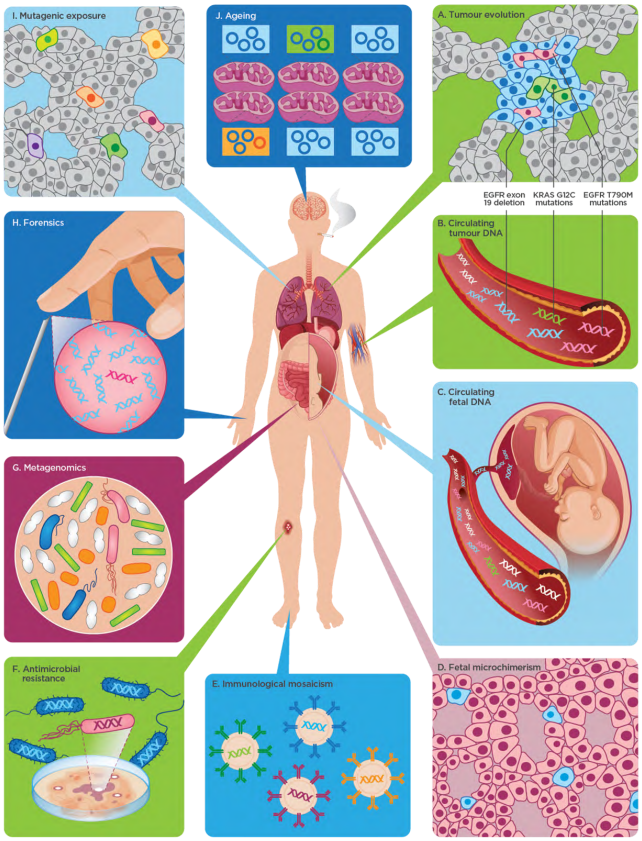
Duplex Sequencing solves genetic needle-in-a-haystack challenges

Genetic Toxicity
Occupational exposures, cancer risk, biological age, drug safety, gene editing safety, CRISPR

Forensics
Crime scenes, assault, Disasters, ancient DNA, archaeology

Microbiota
Human health, environmental monitoring, bioreactors

Antimicrobial
Emergence of antibiotic & antiviral drug resistance, drug selection



Tumor Evolution
Emergence of drug resistance, drug selection

Liquid Biopsies
Early detection of cancer, residual disease monitoring

Organ & Cell Transplant
Chimerism, engraftment, rejection, genetic mosaicism

Prenatal Diagnostics
Fetal genomics

Immunological Health
Fetal microchimerism, autoimmunity, adaptive immunity

for rare variant detection with many diverse applications. TwinStrand has purpose-built solutions for mutagenesis and AML MRD available now with more applications currently in development.

Explore Applications and Publications

AML MRD

MUTAGENESIS

COLLABORATION

PUBLICATIONS

TwinStrand Biosciences, Inc.
3131 Elliott Ave, Suite 750 • Seattle, WA 98121 • USA
+1 (877) 202-TWIN

For Research Use Only. Not for use in diagnostic procedures.

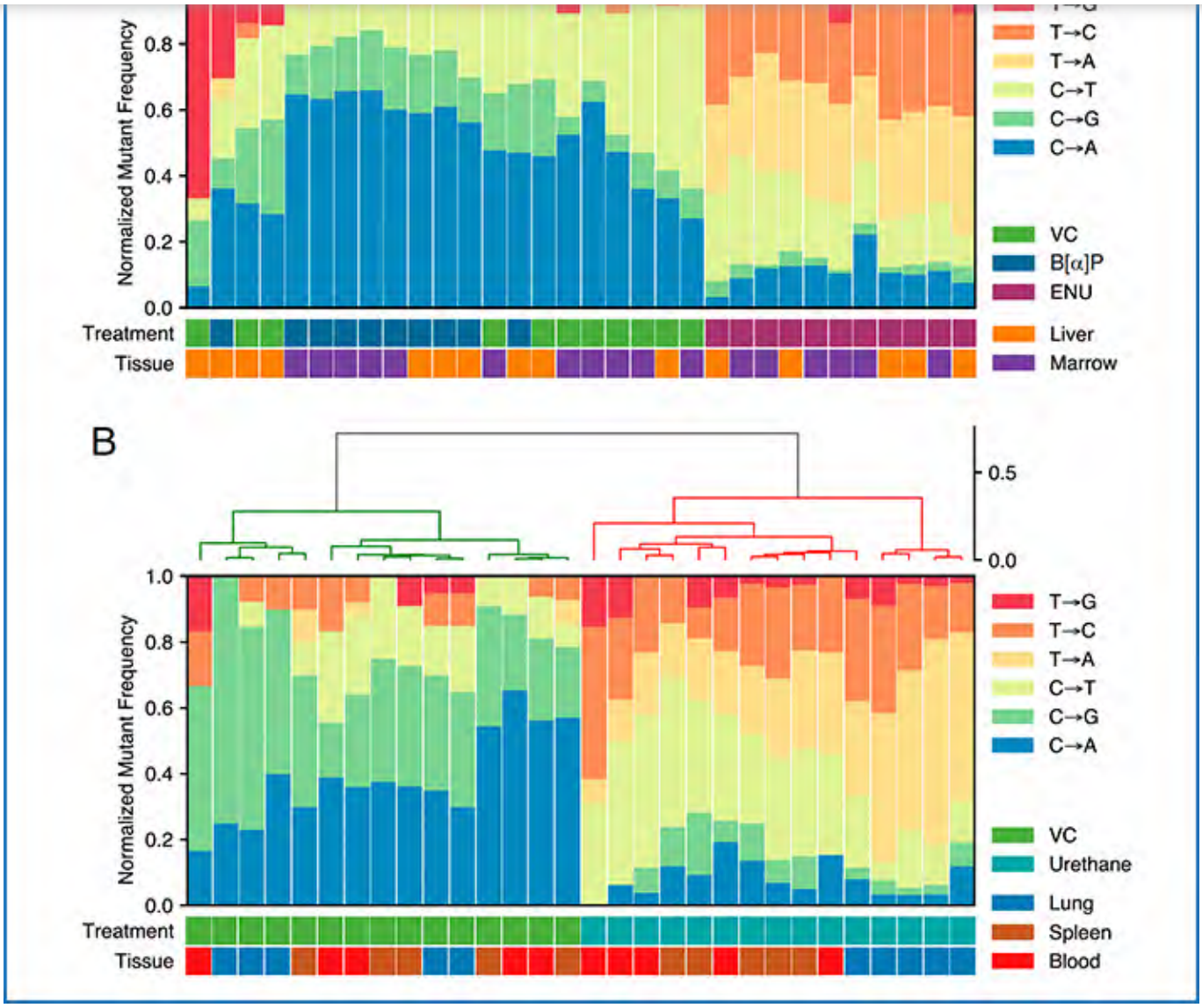
©2020 TwinStrand Biosciences, Inc. All rights reserved. All trademarks are the property of TwinStrand Biosciences, Inc. or their respective owners. Legal Notices


TWINSTRAND
BIOSCIENCES

Publications

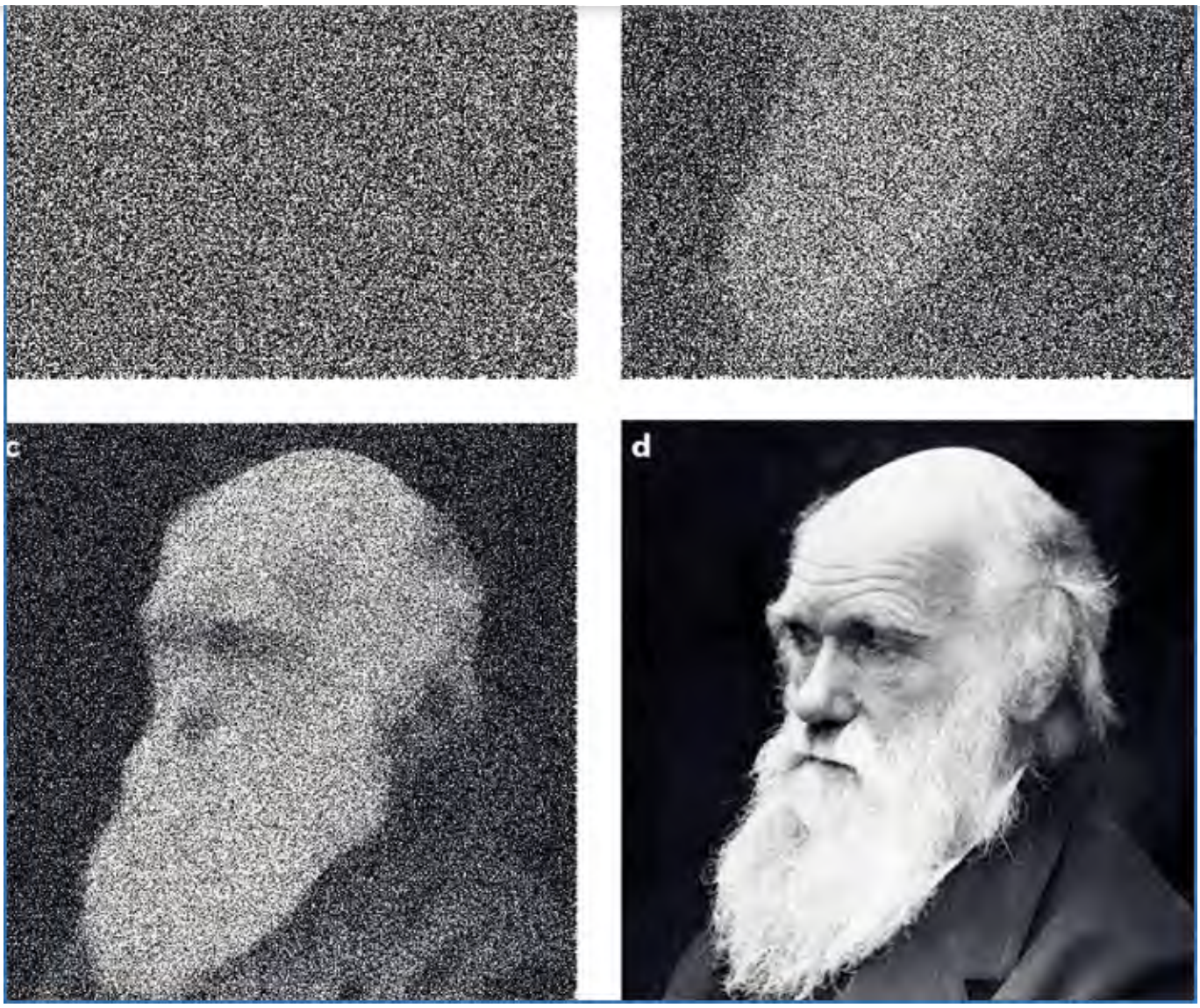
Featured Publications

TWINSTRAND BIOSCIENCES



Direct Quantification of *In Vivo* Mutagenesis and Carcinogenesis Using Duplex Sequencing

TWINSTRAND™ BIOSCIENCES



TWINSTRAND BIOSCIENCES

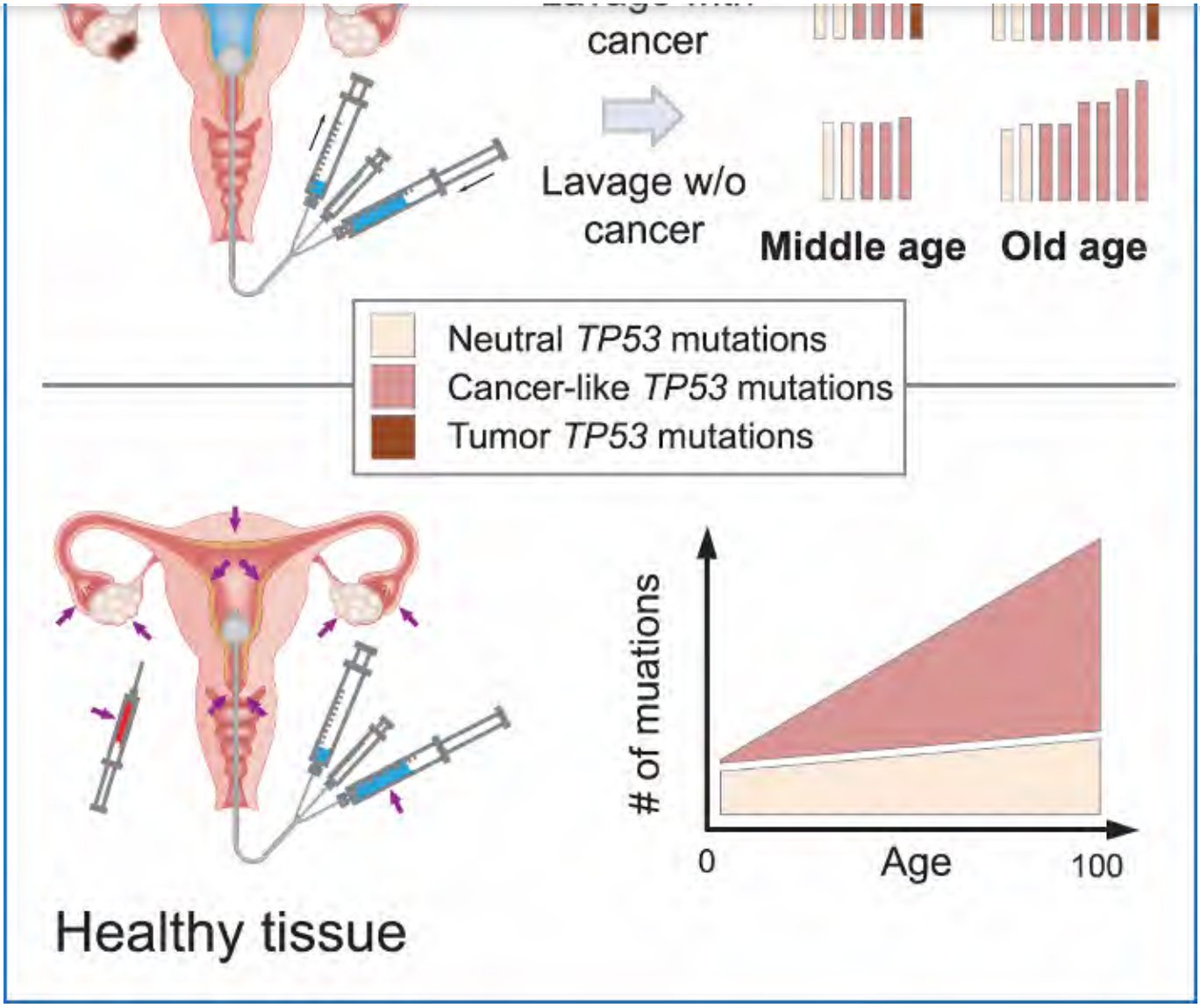
Enhancing the Accuracy of Next-Generation Sequencing for Detecting Rare and Subclonal Mutations

TWINSTRAND BIOSCIENCES



Next-Generation Genotoxicology: Using Modern Sequencing Technologies to Assess Somatic Mutagenesis and Cancer Risk

TWINSTRAND BIOSCIENCES



TWINSTRAND BIOSCIENCES

PNAS. December 2020

Charles C. Valentine III, Robert R. Young, Mark R. Fielden et al.

Direct Quantification of *In Vivo* Mutagenesis and Carcinogenesis Using Duplex Sequencing

pnas.org/content/117/52/33414

Blood Cancer Journal. 2020 Nicholas J. Short and Hagop Kantarjian et al.

Ultra-accurate Duplex Sequencing for the assessment of pretreatment ABL1 kinase domain mutations in Ph+ ALL

DOI: 10.1038/s41408-020-0329-y PubMed: 32457305

PNAS. 2019 Lawrence A. Loeb and Brendan F. Kohrn et al.

Extensive Subclonal Mutational Diversity in Human Colorectal Cancer and its Significance

DOI: 10.1016/j.ygyno.2019.11.124 PubMed: 31806761

PNAS. 2019 Jiang-Cheng Shen and Ashwini S. Kamath-Loeb et al.

TWINSTRAND BIOSCIENCES

Gynecologic Oncology. 2019

Jeffrey D. Krimmel-Morrison and Talayeh S. Ghezelayagh et al.

Characterization of *TP53* Mutations in Pap Test DNA of Women with and without Serous Ovarian Carcinoma

DOI: 10.1016/j.ygyno.2019.11.124 PubMed: 31839337

Environmental and Molecular Mutagenesis. 2019

Jesse J. Salk and Scott R. Kennedy

Next-Generation Genotoxicology: Using Modern Sequencing Technologies to Assess Somatic Mutagenesis and Cancer Risk

10.1002/em.22342 PubMed: 31595553

Nature Communication. 2019 Simonetta Andreazza and Colby L. Samstag et al.

Mitochondrially-Targeted *APOBEC1* is a Potent mtDNA Mutator Affecting Mitochondrial Function and Organismal Fitness in *Drosophila*

DOI: 10.1038/s41467-019-10857-y PubMed: 31337756

Cell Reports. 2019 Jesse J. Salk andKaitlyn Loubet-Senear et al.

doi.org/10.1016 PubMed: 31269435

eLife. 2019 Jianing Xu and Ed Reznik et al.

Abnormal Oxidative Metabolism in a Quiet Genomic Background Underlies Clear Cell Papillary Renal Cell Carcinoma

DOI: 10.7554/eLife.38986 PubMed: 30924768

Nature. 2018 Gannie Tzoneva and Chelsea L. Dieck et al.

Clonal Evolution Mechanisms in *NT5C2* Mutant-Relapsed Acute Lymphoblastic Leukaemia

DOI: 10.1038/nature25186 PubMed: 29342136

PLoS Genetics. 2018 Colby L. Samstag and Jake G. Hoekstra et al.

Deleterious Mitochondrial DNA Point Mutations are Overrepresented in Drosophila Expressing a Proofreading-Defective DNA Polymerase γ

DOI: 10.1371/journal.pgen.1007805 PubMed: 30452458

Genome Research. 2018 Daniela Nachmanson and Shenyi Lian et al.

Targeted Genome Fragmentation with CRISPR/Cas9 Enables Fast and Efficient Enrichment of Small Genomic Regions and

TWINSTRAND BIOSCIENCES

Clinical Cancer Research. 2018 Michael W. Schmitt and Justin R. Pritchard et al.

Single-Molecule Sequencing Reveals Patterns of Preexisting Drug Resistance That Suggest Treatment Strategies in Philadelphia-Positive Leukemias

DOI: 10.1158/1078-0432.CCR-18-0167 PubMed: 30042204

Nature Review Genetics. 2018

Jesse J. Salk, Michael W. Schmitt, and Lawrence A. Loeb

Enhancing the Accuracy of Next-Generation Sequencing for Detecting Rare and Subclonal Mutations

10.1038/nrg.2017.117 PubMed: 29576615

PNAS. 2017 Supawadee Chawanthayatham and Charles C. Valentine III et al.

Mutational Spectra of Aflatoxin B1 *in vivo* Establish Biomarkers of Exposure for Human Hepatocellular Carcinoma

DOI: 10.1073/pnas.1700759114 PubMed: 28351974

PNAS. 2016 Kate S. Reid-Bayliss and Sarah T. Arron et al.

Why Cockayne Syndrome Patients Do Not Get Cancer Despite Their DNA Repair Deficiency

TWINSTRAND BIOSCIENCES

Mitochondrial DNA Mutations Increase in Early Stage Alzheimer Disease and Are Inconsistent with Oxidative Damage

DOI: 10.1002/ana.24709 PubMed: 27315116

eLife. 2016 Maulik R. Patel and Ganesh K. Miriyala et al.

A Mitochondrial DNA Hypomorph of Cytochrome Oxidase Specifically Impairs Male Fertility in Drosophila Melanogaster

DOI: 10.7554/eLife.16923 PubMed: 27481326

PNAS. 2016 Jeffrey D. Krimmel and Michael W. Schmitt et al.

Ultra-Deep Sequencing Detects Ovarian Cancer Cells in Peritoneal Fluid and Reveals Somatic *TP53* Mutations in Noncancerous Tissues

DOI: 10.1073/pnas.1601311113 PubMed: 27152024

Nature Review Clinical Oncology. 2015

Michael W. Schmitt and Lawrence A. Loeb et al.

The Influence of Subclonal Resistance Mutations on Targeted Cancer Therapy

DOI: 10.1038/nrclinonc.2015.175 PubMed: 26483300

TWINSTRAND BIOSCIENCES

DOI: 10.1371/journal.pone.0136216 PubMed: 26305705

Neuron. 2015 Alicia M. Pickrell and Chiu-Hui Huang et al.

Endogenous Parkin Preserves Dopaminergic Substantia Nigral Neurons Following Mitochondrial DNA Mutagenic Stress

DOI: 10.1016/j.neuron.2015.06.034 PubMed: 26182419

Nature Methods. 2015 Michael W. Schmitt and Edward J. Fox et al.

Sequencing Small Genomic Targets with High Efficiency

DOI: 10.1038/nmeth.3351 PubMed: 25849638

Nature Protocols. 2014 Scott R. Kennedy and Michael W. Schmitt et al.

Detecting Ultralow-Frequency Mutations by Duplex Sequencing

DOI: 10.1038/nprot.2014.170 PubMed: 25299156

PLoS Genetics. 2013 Scott R. Kennedy and Jesse J. Salk et al.

Ultra-Sensitive Sequencing Reveals an Age-Related Increase in Somatic Mitochondrial Mutations That Are Inconsistent with Oxidative Damage

TWINSTRAND BIOSCIENCES

Detection of Ultra-Rare Mutations by Next-Generation Sequencing

PNAS.1208715109 PubMed: 22853953

TwinStrand Biosciences, Inc.
3131 Elliott Ave, Suite 750 • Seattle, WA 98121 • USA
+1 (877) 202-TWIN

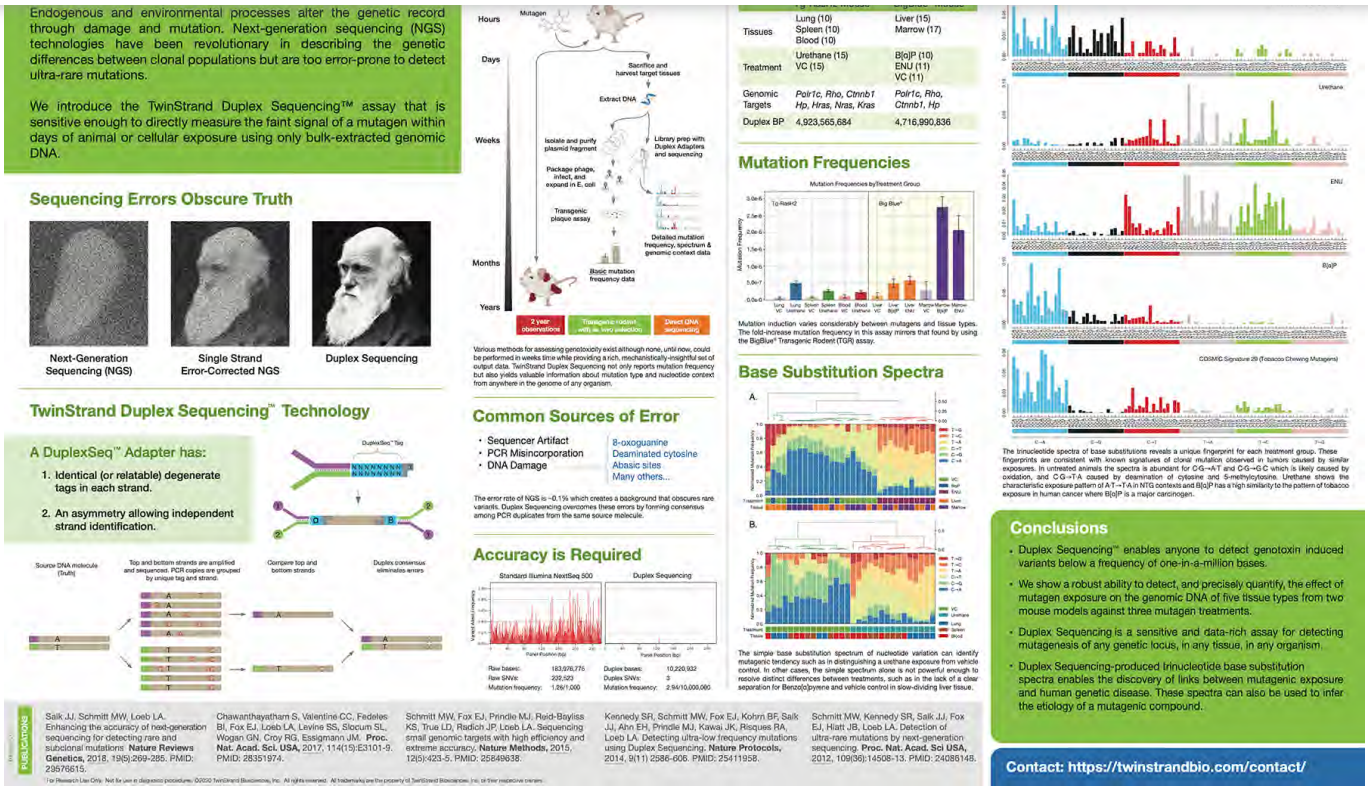
For Research Use Only. Not for use in diagnostic procedures.

©2020 TwinStrand Biosciences, Inc. All rights reserved. All trademarks are the property of TwinStrand Biosciences, Inc. or their respective owners. Legal Notices

Posters

At TwinStrand we put science first in everything we do. We routinely present at academic conferences to share our own discoveries as well as those made with our many collaborators across the scientific spectrum.

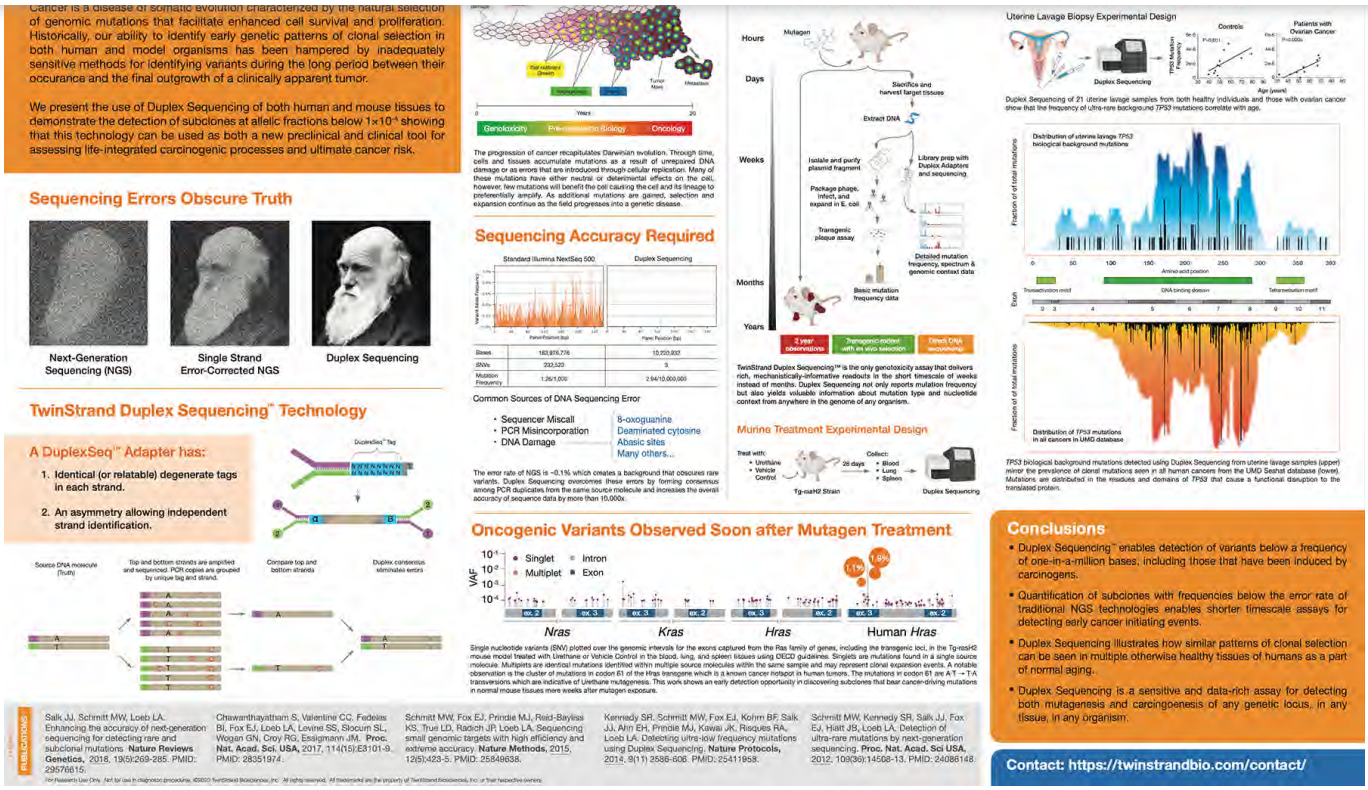
TWINSTRAND™ BIOSCIENCES



POSTER: Detection of Ultra-rare Mutations *in vivo* Establish biomarkers of Endogenous and Environmental Exposure

DOWNLOAD HIGH-RESOLUTION PDF

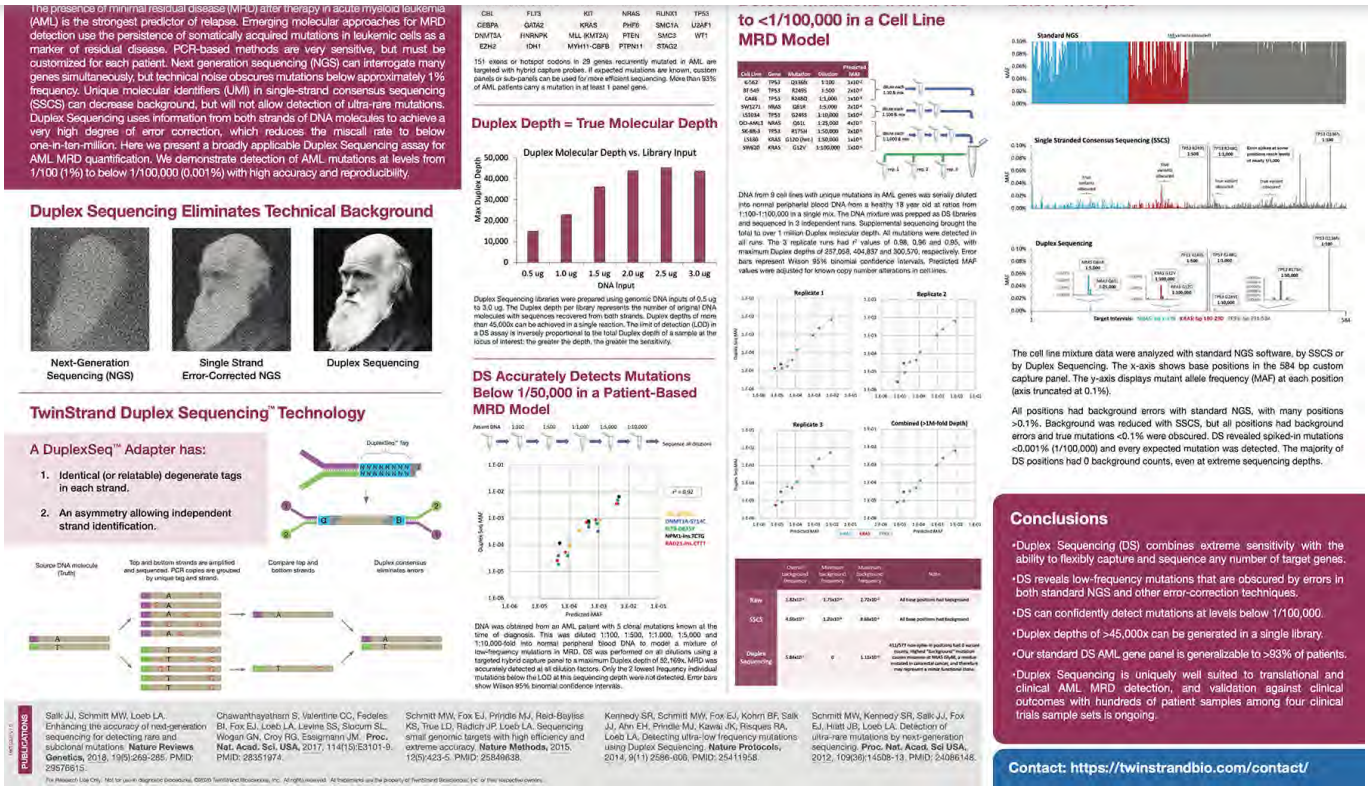
TWINSTRAND™ BIOSCIENCES



POSTER: Duplex Sequencing Detects Rare Subclonal Variants that Mark Early Carcinogenesis and Preneoplastic Clonal Evolution

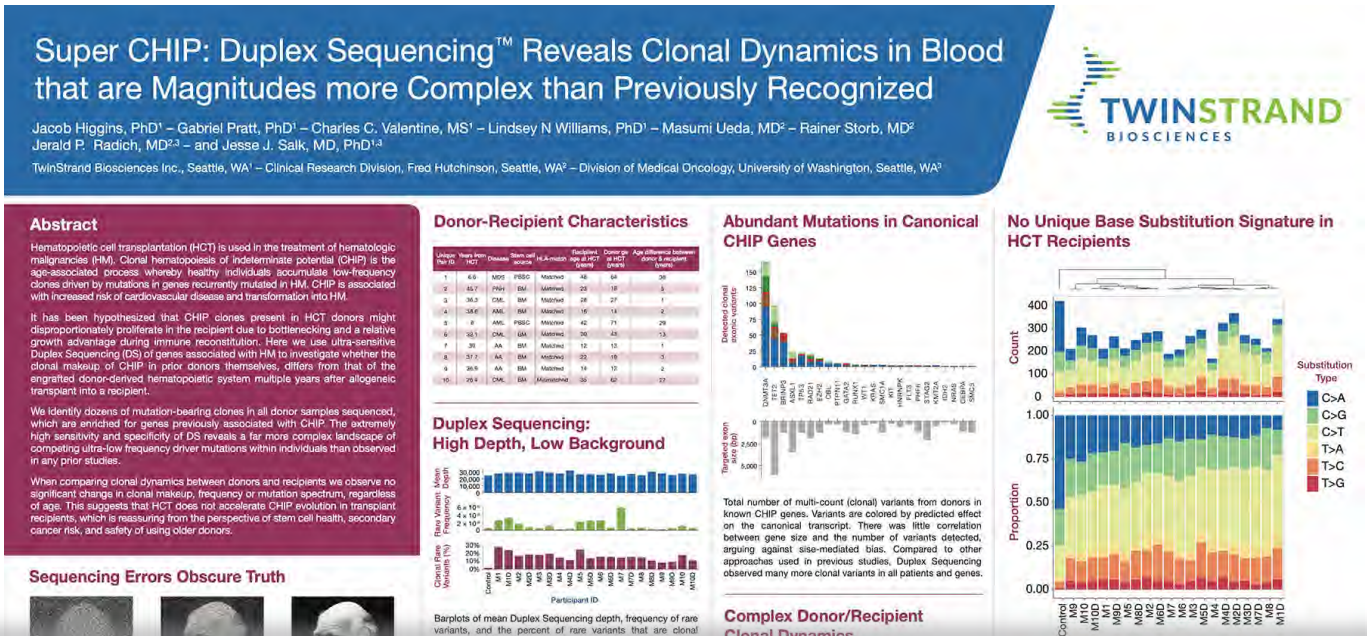
DOWNLOAD HIGH-RESOLUTION PDF

TWINSTRAND BIOSCIENCES

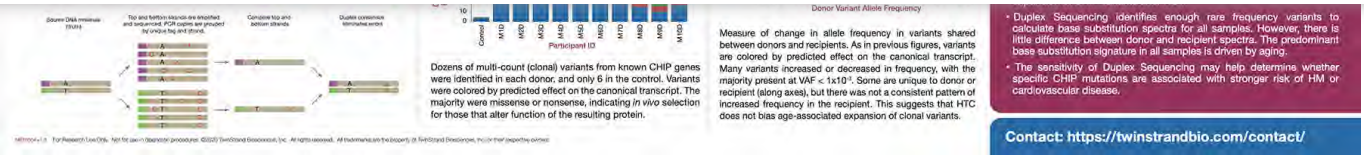


POSTER: Duplex Sequencing for MRD Detection in Acute Myeloid Leukemia

DOWNLOAD HIGH-RESOLUTION PDF

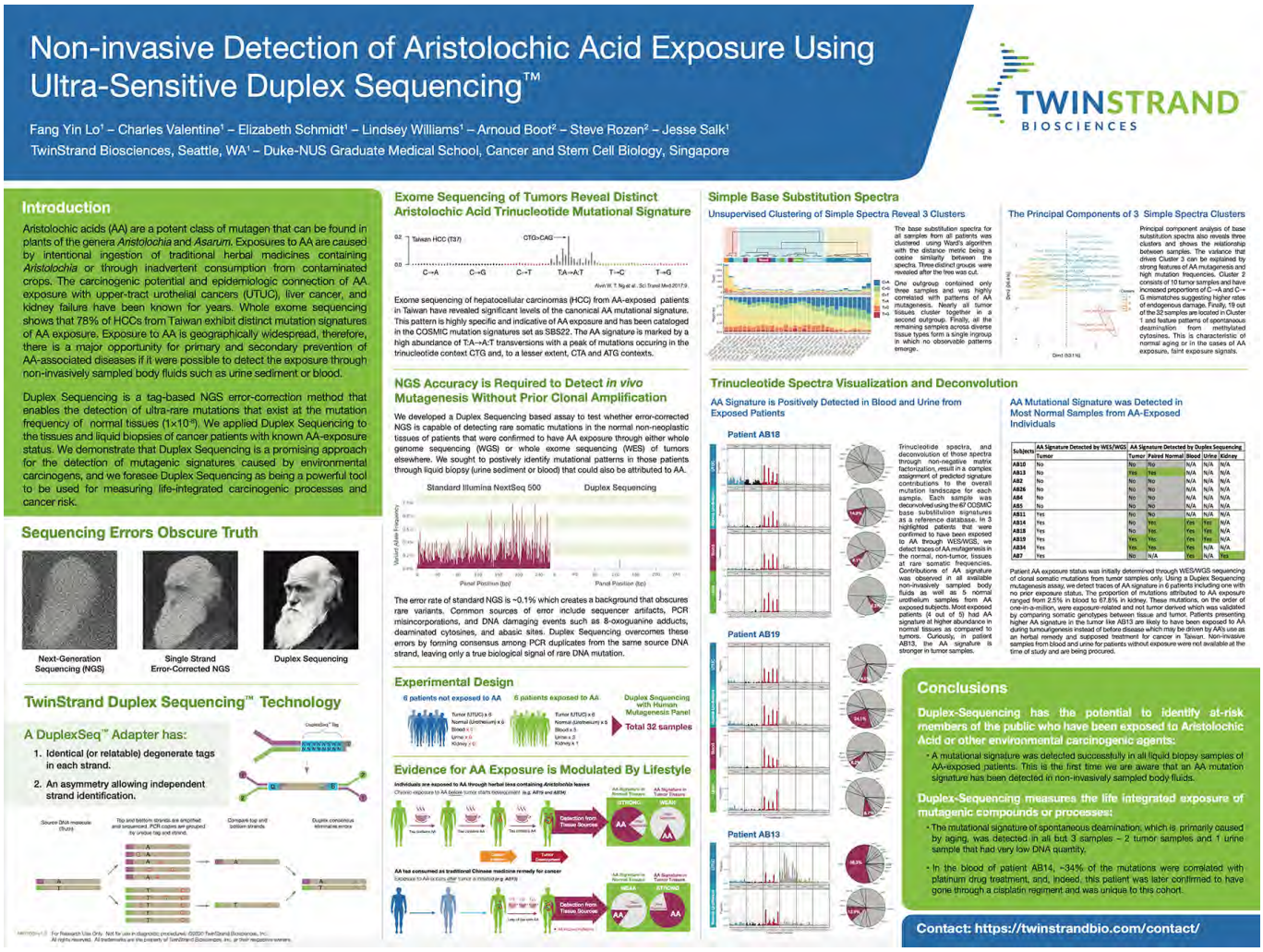


TWINSTRAND BIOSCIENCES



POSTER: Super CHIP: Duplex Sequencing Reveals Clonal Dynamics in Blood that are Magnitudes more Complex than Previously Recognized

DOWNLOAD HIGH-RESOLUTION PDF



POSTER: Non-invasive Detection of Aristolochic Acid Exposure Using Ultra-Sensitive Duplex Sequencing

TWINSTRAND BIOSCIENCES

Redefining "Gold Standard": Ultra-Sensitive Characterization of Commercial DNA Standards with Duplex Sequencing™

Jacob Higgins, PhD[1] – Gabriel Pratt, PhD[1] – Charles C. Valentine, MS[1] – Lindsey N Williams, PhD[1] – and Jesse J. Salk, MD, PhD[1,2]
TwinStrand Biosciences, Seattle, WA[1] – Division of Medical Oncology, University of Washington, Seattle, WA[2]

Abstract

DNA standards are an essential resource in diagnostic and research labs. However, as genomic technologies become more sensitive, the metrics defining what a "gold standard" entails have not been carefully re-evaluated. Cell culture exposes DNA to oxidative damage and clonal selection, and synthetic DNA has a mutation frequency inherently higher than biologically-derived genomic DNA.

Here we apply ultra-accurate Duplex Sequencing to characterize the background mutation frequency in two commercial mutation standards, in control genomic DNA from a healthy young donor and in custom mutation mixes using that young donor DNA as a base. We observe that synthetic (highest) and cell line (intermediate) DNA-based mixes exhibit significantly elevated background mutation frequency relative to carefully selected genomic DNA or mutation mixes where genomic DNA is the dilutant into which patient or cell-line derived mutant molecules are spiked.

When employing high-sensitivity genomic technologies with the ability to resolve mutations at levels below one-in-one-million, rigorous attention to the techniques and substrates used for manufacturing mutation standards are critical.

Sequencing Errors Obscure Truth

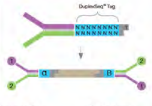


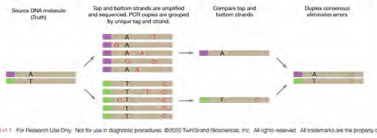
Next-Generation Sequencing (NGS) | Single Strand Error-Corrected NGS | Duplex Sequencing

TwinStrand Duplex Sequencing™ Technology

A DuplexSeq™ Adapter has:

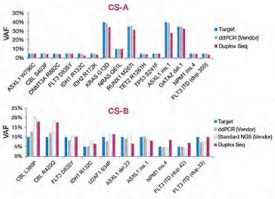
1. Identical (or relatable) degenerate tags in each strand.
2. An asymmetry allowing independent strand identification.





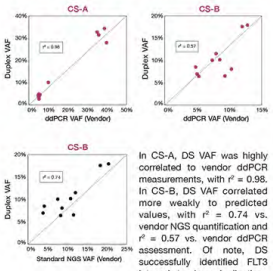
MKT0006 v1.1 For Research Use Only. Not for use in diagnostic procedures. ©2020 TwinStrand Biosciences, Inc. All rights reserved. All trademarks are the property of TwinStrand Biosciences, Inc. or their respective owners.

Duplex Sequencing of Commercially Available DNA Mutation Standards



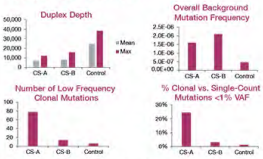
Duplex Sequencing (DS) libraries from commercial standard A (CS-A) and commercial standard B (CS-B) were captured with probes targeting 29 genes recurrently mutated in acute myeloid leukemia (AML). CS-A is a mix of cell line DNA with known mutations, and CS-B is a mix of synthetic mutant DNA fragments spiked into DNA from a single cell line. Fifteen mutations at predicted variant allele frequencies (VAF) from 5-40% were captured in CS-A, and 10 mutations at predicted VAF of 5-15% in CS-B. All mutations were identified by DS.

Duplex Sequencing Correlation with ddPCR and Standard NGS



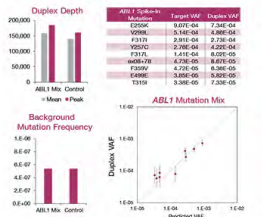
In CS-A, DS VAF was highly correlated to vendor ddPCR measurements, with $r^2 = 0.98$. In CS-B, DS VAF correlated more weakly to predicted values, with $r^2 = 0.74$ vs. vendor NGS quantification and $r^2 = 0.57$ vs. vendor ddPCR assessment. Of note, DS successfully identified FLT3 internal tandem duplications (ITD) from 33-300 bp, as well as other indels up to 23 bp.

Background Mutations in Commercial Standards vs. Negative Control



CS-A and CS-B were sequenced to mean Duplex depth of 7,012x and 8,178x, respectively. Negative control DNA from a healthy 18-year old never-smoker was sequenced to mean Duplex depth of 24,751x. Background mutations were compared in CS-A, CS-B and the control. Despite being sequenced to approximately 3-fold greater Duplex depth, the control DNA had fewer low-frequency clonal mutations (mutations with < 1% VAF and ≥ 2 alternate allele counts) than CS-A or CS-B, in terms of mutant sites and proportion of total background mutations. The control also had lower overall background frequency ($4.7x10^{-7}$) than either CS-A or CS-B.

Duplex Sequencing of Custom In-House *ABL1* Mutation Mix

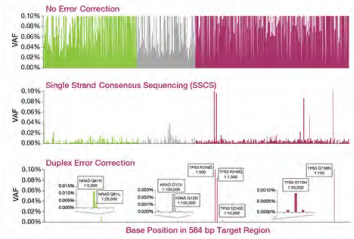


A custom *ABL1* mutation mix was prepared by mixing *ABL1*-mutant DNA into control DNA from a healthy 18-year old donor. Nine mutations were spiked in at target frequencies from 1/1,000 to 1/30,000. The *ABL1* mutation mix was Duplex Sequenced to a mean Duplex depth of 158,823x, and the control DNA was sequenced to a mean depth of 139,819x. All mutations were detected at close to expected frequencies ($r^2 = 0.93$.). Error bars represent 95% confidence intervals. Overall background mutation frequencies were $5.4x10^{-7}$ in the mutation mix and the negative control. The negative control had zero alternate allele counts for any of the *ABL1* spike-in mutations.

Duplex Sequencing of Custom In-House AML Mutation Mix

Gene	Mutation	Predicted VAF
TP53	Q136fs	1x10⁻³
TP53	R248S	2x10⁻³
TP53	R248Q	1x10⁻³
NRAS	Q61R	2x10⁻³
TP53	G245S	1x10⁻³
NRAS	Q61L	4x10⁻³
TP53	R175H	2x10⁻³
KRAS	G12D	1x10⁻³
KRAS	G12V	1x10⁻³

DNA from 9 cell lines, each harboring a unique mutation in *KRAS*, *NRAS* and/or *TP53*, was mixed into negative control DNA such that expected frequencies in the DNA mix ranged from 1/100 to 1/100,000. Replicate libraries were Duplex Sequenced, and the data were combined for over 1 million total Duplex depth. Sequencing data were analyzed with no error correction, with single strand error correction only, or with Duplex error correction. Plots below represent alternate allele signal at each position in the 584 bp target region.



All spiked-in mutations were identified above background signal in the Duplex data, down to VAFs of 0.001%. Mutations below approximately 0.1% are obscured in the SSCS data, and all mutations are obscured in the uncorrected data. Every position in the uncorrected or SSCS data has background signal, whereas the vast majority of Duplex base positions have zero background counts. The overall background frequency of non-spiked in variants with Duplex Sequencing is $5.8x10^{-7}$, which reflects a biological, rather than technical, background.

Conclusions

- Duplex Sequencing of cancer driver genes in apheresis DNA from a healthy 18-year old control reveals a background mutation frequency below 1-in-2-million, which is biological, not technical.
- Cell lines have a higher background mutation frequency and a greater proportion of clonal (multi-count) background mutations than DNA from a healthy young donor, even when sequencing the donor to 3x greater depth.
- For mutation mixes, use of DNA from a young donor as the base results in a lower background mutation than with cell line DNA as the base dilutant.
- Duplex Sequencing readily detects single-nucleotide variants (SNVs) as well as large insertions and deletions at ultra low frequencies.
- Given sufficient sequencing depth, Duplex Sequencing detects mutations at frequencies from 1/30,000 to 1/100,000 with high sensitivity and specificity.

Contact: https://twinstrandbio.com/contact/

POSTER: Redefining "Gold Standard" Ultra-Sensitive Characterization of Commercial DNA Standards with Duplex Sequencing

DOWNLOAD HIGH-RESOLUTION PDF

TwinStrand Biosciences, Inc.
3131 Elliott Ave, Suite 750 • Seattle, WA 98121 • USA

TWINSTRAND BIOSCIENCES

TWINSTRAND™ BIOSCIENCES

Company

We are a vibrant young company committed to applying a powerful new genomic technology across disciplines to improve human health and accelerate scientific discovery. Our mission is to develop and deliver the unprecedented accuracy of Duplex Sequencing™ Technology to applications in clinical medicine and life sciences where it can do the greatest good for the greatest number of people. Our approach is multifaceted and encompasses internal development projects, external collaborations and empowering of partner organizations. Our operation is strongly rooted in academic principles of responsible citizenship within the broader scientific community, yet executed with the nimbleness of a startup, such that research can be rapidly translated into products that benefit people.

TWINSTRAND™ BIOSCIENCES

EVENTS

NEWS

TwinStrand Biosciences, Inc.
3131 Elliott Ave, Suite 750 • Seattle, WA 98121 • USA
+1 (877) 202-TWIN

For Research Use Only. Not for use in diagnostic procedures.

©2020 TwinStrand Biosciences, Inc. All rights reserved. All trademarks are the property of TwinStrand Biosciences, Inc. or their respective owners. Legal Notices

Our Team

TWINSTRAND™ BIOSCIENCES



Jesse Salk MD PhD
Chief Executive Officer, Chief Scientific Officer, Board Member



Elena Cant
Chief Operating Officer

TWINSTRAND BIOSCIENCES



Jeff Grainger JD
Chief Legal Officer

TWINSTRAND™ BIOSCIENCES



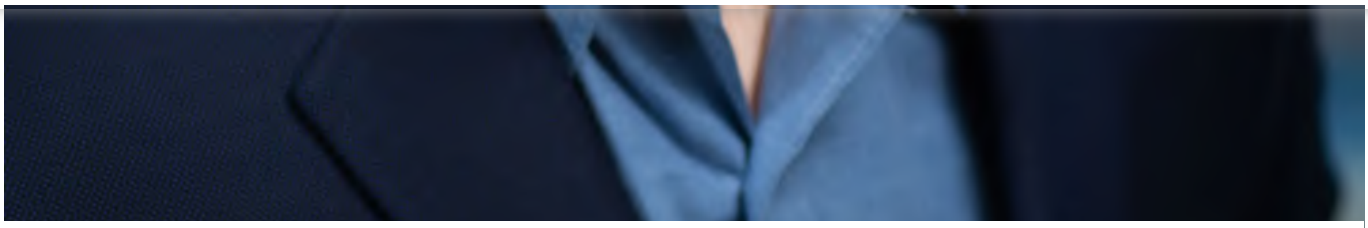
Kellie Bickel PhD
Vice President of Intellectual Property

TWINSTRAND™ BIOSCIENCES



TWINSTRAND BIOSCIENCES



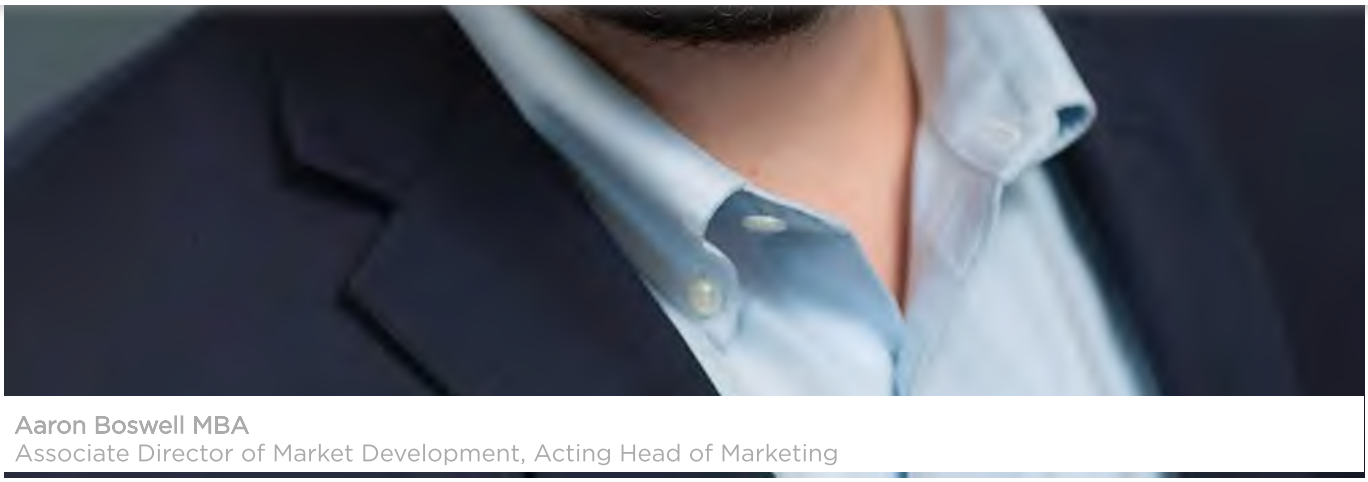


Clint Valentine MS
Director of Computational Biology



Brett Fritz MS
Head of Global Sales and Commercial Development





Aaron Boswell MBA
Associate Director of Market Development, Acting Head of Marketing

Board of Directors

TWINSTRAND™ BIOSCIENCES



Chad Waite
Chairman of the Board

TWINSTRAND™
BIOSCIENCES



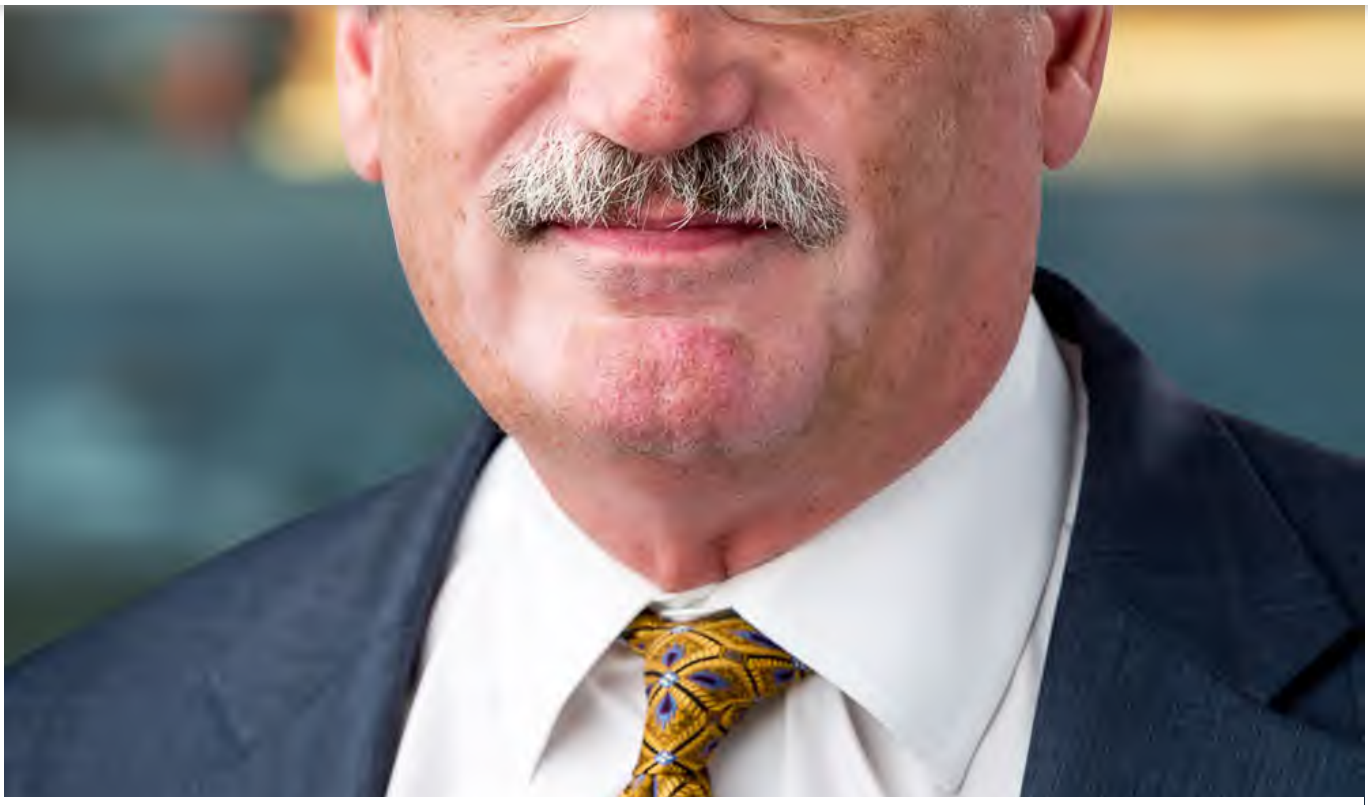
Steve Dow
Board Member



TWINSTRAND™ BIOSCIENCES



TWINSTRAND™ BIOSCIENCES



D. Gary Gilliland MD PhD
Board Member



TWINSTRAND BIOSCIENCES



Jesse Salk MD PhD
Chief Executive Officer, Chief Scientific Officer, Board Member

Scientific Advisors

TWINSTRAND™ BIOSCIENCES



Lawrence A. Loeb MD PhD
Co-founder, Scientific Advisor



TWINSTRAND™
BIOSCIENCES



Jerald Radich MD
Scientific Advisor

TwinStrand Biosciences, Inc.
3131 Elliott Ave, Suite 750 • Seattle, WA 98121 • USA
+1 (877) 202-TWIN


TWINSTRAND
BIOSCIENCES

TWINSTRAND™ BIOSCIENCES

In the News



PRNewswire - November 18, 2020

technology.

READ MORE

PRNewswire - January 6, 2021

TwinStrand Biosciences Appoints Dr. Gary Gilliland to Board of Directors

READ MORE

PRNewswire - December 15, 2020

TwinStrand Biosciences Announces Peer-Reviewed Publication of Study Results Highlighting Ability of Duplex Sequencing Technology to Rapidly Detect the Mutagenic Activity of Chemicals

READ MORE

PRNewswire - November 18, 2020

TwinStrand Biosciences Licenses Duplex Sequencing Technology to Foundation Medicine

READ MORE

Puget Sound Business Journal - October 1, 2020

GEN - August 1, 2020

Liquid Biopsies Generate Ample Insights from Sparse Clues

READ MORE

Gene Online - July 20, 2020

The Profound Impact of Duplex Sequencing: An Interview with Jesse Salk, CEO, TwinStrand Biosciences

READ MORE

Bio-IT World - July 15, 2020

In The Midst Of A Pandemic, Duplex Sequencing Comes Of Age

READ MORE

Timmerman Report - June 15, 2020

Timmerman Report- DNA Sequencing Startups Aim to Reduce Cost, Increase Meaning in the Data

READ MORE

GEN - April 1, 2020

Sequencing's Stars Urge Exploration of New Dimensions

GEN - March 1, 2020

Genomic Medicine Identifies Reconcilable Differences

READ MORE

DeciBio - February 27, 2020

AGBT 2020: DeciBio Highlights – Spatial Profiling Reloaded & MGI Unleashed 2

READ MORE

PRNewswire - February 25, 2020

TwinStrand Biosciences Launches Duplex Sequencing Technology

READ MORE

GeekWire - January 7, 2020

TwinStrand raises $16M from Madrona and others for next-gen DNA sequencing, promising early cancer detection and other breakthroughs

READ MORE

LinkedIn - January 7, 2020

Leveraging big data to improve genome sequencing accuracy by 10,000x

TwinStrand Biosciences, Inc.
3131 Elliott Ave, Suite 750 • Seattle, WA 98121 • USA
+1 (877) 202-TWIN

For Research Use Only. Not for use in diagnostic procedures.

©2020 TwinStrand Biosciences, Inc. All rights reserved. All trademarks are the property of TwinStrand Biosciences, Inc. or their respective owners. Legal Notices