



(12)发明专利申请

(10)申请公布号 CN 106536756 A

(43)申请公布日 2017.03.22

(21)申请号 201580034149.4

(22)申请日 2015.06.26

(30)优先权数据

62/017,808 2014.06.26 US

62/072,214 2014.10.29 US

(85)PCT国际申请进入国家阶段日
2016.12.23

(86)PCT国际申请的申请数据
PCT/US2015/038141 2015.06.26

(87)PCT国际申请的公布数据
W02015/200869 EN 2015.12.30

(71)申请人 10X基因组学有限公司
地址 美国加利福尼亚州

(72)发明人 迈克尔·史诺-莱文
米尔娜·雅罗什

克里斯多佛·辛德森 凯文·奈斯
瑟奇·萨克森诺夫
本杰明·辛德森
格雷斯·X·Y·郑
帕特里克·马科斯
约翰·史图尔普纳格

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262

代理人 郑霞

(51)Int.Cl.

C12Q 1/68(2006.01)

C12N 15/10(2006.01)

C40B 40/06(2006.01)

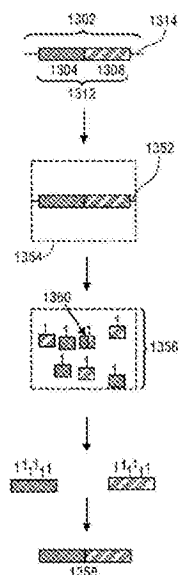
权利要求书10页 说明书42页 附图23页

(54)发明名称

核酸序列的分析

(57)摘要

本公开涉及用于单倍型定相和拷贝数变异分析的方法、组合物以及系统。本公开内包括用于将包含条形码的珠粒与样品组合于多个单独的分区中的方法和系统,以及对条形码化样品进行处理、测序和分析的方法。



1. 一种用于识别核酸中的一种或多种变异的方法,其包括:
 - (a) 提供所述核酸的第一片段,其中所述第一片段具有大于10千碱基(kb)的长度;
 - (b) 对所述第一片段的多个第二片段进行测序,以提供多个片段序列,所述多个片段序列共有共同的条形码序列;
 - (c) 根据所述共同条形码序列的存在将所述多个片段序列归属于所述第一片段;
 - (d) 使用所述多个片段序列测定所述第一片段的核酸序列,其中以低于1%的误差率测定所述核酸序列;以及
 - (e) 识别(d)中所测定的所述第一片段的所述核酸序列中的所述一种或多种变异,由此识别所述核酸内的所述一种或多种变异。
2. 如权利要求1所述的方法,其中所述第一片段在多个离散分区之中的离散分区中。
3. 如权利要求2所述的方法,其中所述离散分区为乳液中的微滴。
4. 如权利要求1所述的方法,其中所述识别包括由所述第一片段的所述核酸序列识别所述核酸中的定相变体。
5. 如权利要求1所述的方法,其中所述识别包括由所述第一片段的所述核酸序列识别所述核酸中的一种或多种结构变异。
6. 如权利要求1所述的方法,其中所述第一片段具有大于15kb的长度。
7. 如权利要求1所述的方法,其中所述第一片段具有大于20kb的长度。
8. 如权利要求1所述的方法,其中所述测定包括将所述多个片段序列映射至参考序列。
9. 如权利要求1所述的方法,其中所述测定包括组装具有所述共同条形码序列的所述多个片段序列。
10. 如权利要求1所述的方法,其进一步包括提供所述核酸的多个第一片段,所述多个第一片段的长度为至少10kb,并且其中所述识别包括测定所述多个第一片段中的每一者的核酸序列以及由所述多个第一片段中的每一者的所述核酸序列识别所述核酸中的所述一种或多种变异。
11. 如权利要求10所述的方法,其进一步包括基于所述多个第一片段的两个或更多个核酸序列的重叠核酸序列将所述两个或更多个核酸序列连接成推测重叠群,其中最大推测重叠群长度为至少10kb。
12. 如权利要求11所述的方法,其中所述最大推测重叠群长度为至少20kb。
13. 如权利要求12所述的方法,其中所述最大推测重叠群长度为至少40kb。
14. 如权利要求13所述的方法,其中所述最大推测重叠群长度为至少50kb。
15. 如权利要求14所述的方法,其中所述最大推测重叠群长度为至少100kb。
16. 如权利要求15所述的方法,其中所述最大推测重叠群长度为至少200kb。
17. 如权利要求16所述的方法,其中所述最大推测重叠群长度为至少500kb。
18. 如权利要求17所述的方法,其中所述最大推测重叠群长度为至少750kb。
19. 如权利要求18所述的方法,其中所述最大推测重叠群长度为至少1兆碱基(Mb)。
20. 如权利要求19所述的方法,其中所述最大推测重叠群长度为至少1.75Mb。
21. 如权利要求20所述的方法,其中所述最大推测重叠群长度为至少2.5Mb。
22. 如权利要求10所述的方法,其进一步包括基于所述多个第一片段的两个或更多个核酸序列内的重叠定相变体将所述多个第一片段的所述两个或更多个核酸序列连接成相

位区块,其中最大相位区块长度为至少10kb。

23. 如权利要求22所述的方法,其中所述最大相位区块长度为至少20kb。

24. 如权利要求23所述的方法,其中所述最大相位区块长度为至少40kb。

25. 如权利要求24所述的方法,其中所述最大相位区块长度为至少50kb。

26. 如权利要求25所述的方法,其中所述最大相位区块长度为至少100kb。

27. 如权利要求26所述的方法,其中所述最大相位区块长度为至少200kb。

28. 如权利要求27所述的方法,其中所述最大相位区块长度为至少500kb。

29. 如权利要求28所述的方法,其中所述最大相位区块长度为至少750kb。

30. 如权利要求29所述的方法,其中所述最大相位区块长度为至少1Mb。

31. 如权利要求30所述的方法,其中所述最大相位区块长度为至少1.75Mb。

32. 如权利要求31所述的方法,其中所述最大相位区块长度为至少2.5Mb。

33. 如权利要求10所述的方法,其进一步包括基于所述多个第一片段的两个或更多个核酸序列的重叠核酸序列将所述两个或更多个核酸序列连接成推测重叠群,由此形成推测重叠群的群体,其中所述推测重叠群的群体的N50为至少10kb。

34. 如权利要求33所述的方法,其中所述推测重叠群的群体的所述N50为至少20kb。

35. 如权利要求34所述的方法,其中所述推测重叠群的群体的所述N50为至少40kb。

36. 如权利要求35所述的方法,其中所述推测重叠群的群体的所述N50为至少50kb。

37. 如权利要求36所述的方法,其中所述推测重叠群的群体的所述N50为至少100kb。

38. 如权利要求37所述的方法,其中所述推测重叠群的群体的所述N50为至少200kb。

39. 如权利要求38所述的方法,其中所述推测重叠群的群体的所述N50为至少500kb。

40. 如权利要求39所述的方法,其中所述推测重叠群的群体的所述N50为至少750kb。

41. 如权利要求40所述的方法,其中所述推测重叠群的群体的所述N50为至少1Mb。

42. 如权利要求41所述的方法,其中所述推测重叠群的群体的所述N50为至少1.75Mb。

43. 如权利要求42所述的方法,其中所述推测重叠群的群体的所述N50为至少2.5Mb。

44. 如权利要求10所述的方法,其进一步包括基于所述多个第一片段的两个或更多个核酸序列内的重叠定相变体将所述多个第一片段的所述两个或更多个核酸序列连接成相位区块,由此形成相位区块的群体,其中所述相位区块的群体的N50为至少10kb。

45. 如权利要求44所述的方法,其中所述相位区块的群体的所述N50为至少20kb。

46. 如权利要求45所述的方法,其中所述相位区块的群体的所述N50为至少40kb。

47. 如权利要求46所述的方法,其中所述相位区块的群体的所述N50为至少50kb。

48. 如权利要求47所述的方法,其中所述相位区块的群体的所述N50为至少100kb。

49. 如权利要求48所述的方法,其中所述相位区块的群体的所述N50为至少200kb。

50. 如权利要求49所述的方法,其中所述相位区块的群体的所述N50为至少500kb。

51. 如权利要求50所述的方法,其中所述相位区块的群体的所述N50为至少750kb。

52. 如权利要求51所述的方法,其中所述相位区块的群体的所述N50为至少1Mb。

53. 如权利要求52所述的方法,其中所述相位区块的群体的所述N50为至少1.75Mb。

54. 如权利要求53所述的方法,其中所述相位区块的群体的所述N50为至少2.5Mb。

55. 一种确定核酸结构变异的存在的方法,其包括:

(a) 提供所述核酸的多个第一片段分子,其中所述多个第一片段分子中的给定第一片

段分子包含所述结构变异；

(b) 对所述多个第一片段分子中的每一者的多个第二片段分子进行测序，以提供多个片段序列，其中所述多个片段序列中对应于给定第一片段分子的每一者共有共同的条形码序列；以及

(c) 通过以下方式确定所述结构变异的存在：(i) 将所述多个片段序列映射至参考序列，(ii) 识别共有所述共同条形码序列的所述多个片段序列，并且 (iii) 基于在比所述给定第一片段分子的长度相隔更远的位置存在升高量的映射至所述参考序列的共有所述共同条形码序列的所述多个片段序列来识别所述结构变异，所述升高量是相对于缺少所述结构变异的序列。

56. 如权利要求55所述的方法，其中所述升高量相对于源于所述核酸中具有所述结构变异的区域的所述第一片段分子的总数为1%或更多。

57. 如权利要求56所述的方法，其中所述升高量相对于源于所述核酸中具有所述结构变异的区域的所述第一片段分子的总数为2%或更多。

58. 如权利要求55所述的方法，其中所述位置相隔至少约100个碱基。

59. 如权利要求58所述的方法，其中所述位置相隔至少约500个碱基。

60. 如权利要求59所述的方法，其中所述位置相隔至少约1千碱基 (kb)。

61. 如权利要求60所述的方法，其中所述位置相隔至少约10kb。

62. 如权利要求55所述的方法，其进一步包括通过由所述多个片段序列形成所述给定第一片段分子的组合物来识别所述结构变异，其中基于所述共同条形码序列的存在选择所述多个片段序列作为所述组合物的输入物。

63. 如权利要求62所述的方法，其中所述组合物通过由所述多个片段序列产生共有序列来形成。

64. 如权利要求55所述的方法，其中所述结构变异包括易位。

65. 一种表征变体核酸序列的方法，其包括：

(a) 将变体核酸片段化以提供具有大于10千碱基 (kb) 的长度的多个第一片段；

(b) 将所述多个第一片段分离至离散分区中；

(c) 由各个第一片段各自的分区内的各个第一片段形成多个第二片段，所述多个第二片段具有与其连接的条形码序列，给定分区内的所述条形码序列为共同条形码序列；

(d) 对所述多个第二片段和与其连接的所述条形码序列进行测序，以提供多个第二片段序列；

(e) 至少部分基于所述共同条形码序列的存在将所述第二片段序列归属于原始第一片段，以提供所述第二片段序列的第一片段序列环境；以及

(f) 由所述第一片段序列环境识别所述变体核酸的变体部分，由此表征所述变体核酸序列。

66. 如权利要求65所述的方法，其中所述归属包括至少部分基于所述共同条形码序列的存在由所述多个第二片段序列组装所述多个第一片段中的单个片段的序列的至少一部分。

67. 如权利要求65所述的方法，其中所述归属包括至少部分基于所述共同条形码序列将所述多个第二片段序列映射至所述多个第一片段中的单个第一片段。

68. 如权利要求65所述的方法,其进一步包括基于所述多个第一片段中的两者或更多者之间的重叠序列将所述多个第一片段中的所述两者或更多者连接成推测重叠群。

69. 如权利要求65所述的方法,其中所述识别包括由所述第一片段序列环境识别一个或多个定相变体。

70. 如权利要求65所述的方法,其进一步包括基于所述多个第一片段中的两者或更多者之间的重叠定相变体将所述多个第一片段中的所述两者或更多者连接成相位区块。

71. 如权利要求70所述的方法,其中所述识别包括由所述第一片段序列环境识别一种或多种结构变异。

72. 如权利要求71所述的方法,其中所述一种或多种结构变异独立地选自插入、缺失、易位、反转录转座子、倒位以及重复。

73. 如权利要求72所述的方法,其中所述结构变异包括插入或易位,并且所述第一片段序列环境指示所述插入或易位的存在。

74. 一种识别核酸的序列中的变体的方法,其包括:

获得所述核酸的多个单个片段分子的核酸序列,所述多个单个片段分子的所述核酸序列各自具有至少1千碱基(kb)的长度;

将所述多个单个片段分子中的一者或多者的序列连接成一个或多个推测重叠群;以及
由所述一个或多个推测重叠群识别一个或多个变体。

75. 如权利要求74所述的方法,其中所述获得包括获得长度大于10kb的多个片段分子的所述核酸序列。

76. 如权利要求75所述的方法,其中所述获得包括:

提供所述多个单个片段分子中的各个单个片段分子的多个条形码化片段,给定单个片段分子的所述条形码化片段具有共同条形码;

对所述多个单个片段分子的所述多个条形码化片段进行测序,所述测序提供低于1%的测序误差率;以及

由所述多个条形码化片段和其相关条形码的序列测定所述多个单个片段分子的序列。

77. 如权利要求76所述的方法,其中所述连接包括识别两个或更多个单个片段分子之间的一个或多个重叠序列以将所述两个或更多个单个片段分子连接成一个或多个推测重叠群。

78. 如权利要求77所述的方法,其中所述连接包括识别两个或更多个单个片段分子之间的一个或多个共同变体以将所述两个或更多个单个片段分子连接成所述一个或多个推测重叠群。

79. 如权利要求78所述的方法,其中所述一个或多个共同变体为定相变体,并且所述一个或多个推测重叠群包含至少100kb的最大相位区块长度。

80. 如权利要求74所述的方法,其中在所述识别中所识别的所述一个或多个变体包含结构变异。

81. 如权利要求80所述的方法,其中所述结构变异选自插入、缺失、易位、反转录转座子、倒位以及重复。

82. 一种表征核酸的方法,其包括:

获得具有至少10千碱基(kb)的长度的多个片段分子的核酸序列;

识别所述多个片段分子的所述核酸序列中的一个或多个定相变体位置；

基于所述第一和第二片段分子内的一个或多个共同定相变体位置的存在将至少一个第一片段分子的所述核酸序列连接至至少一个第二片段分子，以提供具有至少10kb的最大相位区块长度的相位区块；以及

由具有至少10kb的所述最大相位区块长度的所述相位区块识别一个或多个定相变体。

83. 如权利要求82所述的方法，其进一步包括由所述相位区块识别一个或多个额外定相变体。

84. 如权利要求82所述的方法，其中所述多个片段分子在离散分区中。

85. 如权利要求84所述的方法，其中所述离散分区为乳液中的微滴。

86. 如权利要求82所述的方法，其中所述多个片段分子的长度为至少50kb。

87. 如权利要求86所述的方法，其中所述多个片段分子的长度为至少100kb。

88. 如权利要求82所述的方法，其中所述最大相位区块长度为至少50kb。

89. 如权利要求88所述的方法，其中所述最大相位区块长度为至少100kb。

90. 如权利要求89所述的方法，其中所述最大相位区块长度为至少1Mb。

91. 如权利要求90所述的方法，其中所述最大相位区块长度为至少2Mb。

92. 如权利要求91所述的方法，其中所述最大相位区块长度为至少2.5Mb。

93. 一种方法，其包括：

a) 将第一核酸分配至第一分区中，其中所述第一核酸包含源于生物体的第一染色体的靶序列；

b) 将第二核酸分配至第二分区中，其中所述第二核酸包含源于所述生物体的第二染色体的所述靶序列；

c) 在所述第一分区中，将第一条形码序列连接至所述第一核酸的片段或所述第一核酸的诸多个部分的拷贝以提供第一条形码化片段；

d) 在所述第二分区，将第二条形码序列连接至所述第二核酸的片段或所述第二核酸的诸多个部分的拷贝以提供第二条形码化片段，所述第二条形码序列不同于所述第一条形码序列；

e) 测定所述第一和第二条形码化片段的所述核酸序列，并且组装所述第一和第二核酸的核酸序列；以及

f) 比较所述第一和第二核酸的所述核酸序列以分别将所述第一和第二核酸表征为源于第一和第二染色体。

94. 如权利要求93所述的方法，其中将包含所述第一条形码序列的寡核苷酸与所述第一核酸共分配，并且将包含所述第二条形码序列的寡核苷酸与所述第二核酸共分配。

95. 如权利要求94所述的方法，其中将包含所述第一条形码序列的所述寡核苷酸可释放地连接至第一珠粒，并且将包含所述第二条形码序列的所述寡核苷酸可释放地连接至第二珠粒，并且所述共分配包括分别将所述第一和第二珠粒共分配至所述第一和第二分区中。

96. 如权利要求93所述的方法，其中所述第一和第二分区包括乳液中的微滴。

97. 如权利要求93所述的方法，其中所述第一染色体为父体染色体，并且所述第二染色体为母体染色体。

98. 如权利要求93所述的方法,其中所述第一染色体和所述第二染色体为同源染色体。
99. 如权利要求93所述的方法,其中所述第一核酸和所述第二核酸包含一种或多种变异。
100. 一种方法,其包括:
- a) 将第一核酸分配至第一分区中,其中所述第一核酸包含源于生物体的第一染色体的靶序列;
 - b) 将第二核酸分配至第二分区中,其中所述第二核酸包含源于所述生物体的第二染色体的所述靶序列;
 - c) 在所述第一分区中,将第一条形码序列连接至所述第一核酸的片段或所述第一核酸的诸多个部分的拷贝以提供第一条形码化片段;
 - d) 在所述第二分区,将第二条形码序列连接至所述第二核酸的片段或所述第二核酸的诸多个部分的拷贝以提供第二条形码化片段,所述第二条形码序列不同于所述第一条形码序列;
 - e) 测定所述第一和第二条形码化片段的所述核酸序列,并且组装所述第一和第二核酸的核酸序列;以及
 - f) 比较所述第一和第二核酸的所述核酸序列以识别所述第一和第二核酸的所述核酸序列之间的任何变异。
101. 如权利要求100所述的方法,其中将包含所述第一条形码序列的寡核苷酸与所述第一核酸共分配,并且将包含所述第二条形码序列的寡核苷酸与所述第二核酸共分配。
102. 如权利要求101所述的方法,其中将包含所述第一条形码序列的所述寡核苷酸可释放地连接至第一珠粒,并且将包含所述第二条形码序列的所述寡核苷酸可释放地连接至第二珠粒,并且所述共分配包括分别将所述第一和第二珠粒共分配至所述第一和第二分区中。
103. 如权利要求100所述的方法,其中所述第一和第二分区包括乳液中的微滴。
104. 如权利要求100所述的方法,其中所述第一染色体为父体染色体,并且所述第二染色体为母体染色体。
105. 如权利要求100所述的方法,其中所述第一染色体和所述第二染色体为同源染色体。
106. 如权利要求100所述的方法,其中所述第一核酸和所述第二核酸包含一种或多种变异。
107. 如权利要求93或权利要求100所述的方法,其中所述第一和第二染色体源于胎儿。
108. 如权利要求93或权利要求100所述的方法,其中所述第一和第二核酸是从取自孕妇的样品获得。
109. 如权利要求93或权利要求100所述的方法,其中所述第一染色体为染色体21、18或13。
110. 如权利要求93或权利要求100所述的方法,其中所述第二染色体为染色体21、18或13。
111. 如权利要求93或权利要求100所述的方法,其进一步包括测定所述第一或第二染色体的相对量。

112. 如权利要求93或权利要求100所述的方法,其进一步包括测定所述第一或第二染色体相对于参考染色体的量。

113. 如权利要求93或权利要求100所述的方法,其中所述第一染色体或第二染色体或两者的拷贝数增加。

114. 如权利要求113所述的方法,其中所述拷贝数增加是癌症或非整倍性的结果。

115. 如权利要求93或权利要求100所述的方法,其中所述第一染色体或第二染色体或两者的拷贝数减少。

116. 如权利要求115所述的方法,其中所述拷贝数减少是癌症或非整倍性的结果。

117. 一种用于表征胎儿核酸序列的方法,其包括:

(a) 通过以下方式测定母体核酸序列,其中所述母体核酸源于胎儿的怀孕母亲:

(i) 将母体核酸片段化以提供多个第一母体片段;

(ii) 将所述多个第一母体片段分离至母体分区中;

(iii) 由第一母体片段各自的母体分区内的所述第一母体片段中的每一者形成多个第二母体片段,所述多个第二母体片段具有与其连接的第一条形码序列,其中在所述母体分区中的给定母体分区内,所述第二母体片段包含与其连接的第一共同条形码序列;

(iv) 对所述多个第二母体片段进行测序以提供多个母体片段序列;

(v) 至少部分基于所述第一共同条形码序列的存在将所述母体片段序列归属于原始第一母体片段以测定所述母体核酸序列;

(b) 通过以下方式测定父体核酸序列,其中所述父体核酸源于所述胎儿的父亲:

(i) 将父体核酸片段化以提供多个第一父体片段;

(ii) 将所述多个第一父体片段分离至父体离散分区中;

(iii) 从各个第一父体片段各自的分区内的各个第一父体片段形成多个第二父体片段,所述多个第二父体片段具有与其连接的第二条形码序列,其中在给定父体分区内,所述第二父体片段包含与其连接的第二共同条形码序列;

(iv) 对所述多个第二父体片段和与其连接的所述第二条形码序列进行测序,以提供多个父体片段序列;

(v) 至少部分基于所述第二共同条形码序列的存在将所述父体片段序列归属于原始第一父体片段以测定所述父体核酸序列;

(c) 从所述怀孕母亲获得胎儿核酸并且使用所述母体核酸序列和所述父体核酸序列测定所述胎儿核酸的序列和/或所述胎儿核酸的所述序列的一种或多种基因变异。

118. 如权利要求117所述的方法,其中使用所述父体片段序列和所述母体片段序列各者来将序列连接成一个或多个推测重叠群。

119. 如权利要求118所述的方法,其中使用所述推测重叠群来构建母体和父体相位区块。

120. 如权利要求119所述的方法,其中将所述胎儿核酸的所述序列与所述母体和父体相位区块相比较来构建胎儿相位区块。

121. 如权利要求120所述的方法,其中组装所述父体片段序列以产生所述多个第一父体片段的序列的至少一部分,由此测定所述父体核酸序列,并且其中组装所述母体片段序列以产生所述多个第一母体片段的序列的至少一部分,由此测定所述母体核酸序列。

122. 如权利要求117所述的方法,其中所述测定所述父体核酸序列包括将所述父体片段序列映射至父体参考序列,并且其中所述测定所述母体核酸序列包括将所述母体片段序列映射至母体参考序列。

123. 如权利要求117所述的方法,其中以至少99%的准确度测定所述胎儿核酸的所述序列。

124. 如权利要求117所述的方法,其中以至少99%的准确度测定所述胎儿核酸的所述序列的所述一种或多种基因变异。

125. 如权利要求117所述的方法,其中所述一种或多种基因变异选自结构变异和单核苷酸多态性(SNP)。

126. 如权利要求117所述的方法,其中所述一种或多种基因变异为选自拷贝数变异、插入、缺失、易位、反转录转座子、倒位、重排、重复扩增以及重复的结构变异。

127. 如权利要求117所述的方法,其进一步包括,在(c)中,使用针对所述母体核酸序列和所述父体核酸序列所测定的一种或多种基因变异测定所述胎儿核酸的所述序列的所述一种或多种基因变异。

128. 如权利要求117所述的方法,其进一步包括,在(c)中,测定所述胎儿核酸的一个或多个从头突变。

129. 如权利要求117所述的方法,其进一步包括,在(c)期间或之后,测定与所述胎儿核酸相关的非整倍性。

130. 如权利要求117所述的方法,其进一步包括,在(a)中在(v)期间或之后,对所述母体核酸序列进行单倍型分析以提供单倍型解析型母体核酸序列,以及在(b)中在(v)期间或之后,对所述父体核酸序列进行单倍型分析以提供单倍型解析型父体核酸序列。

131. 如权利要求130所述的方法,其进一步包括,在(c)中,使用所述单倍型解析型母体核酸序列和所述单倍型解析型父体核酸序列测定所述胎儿核酸的所述序列和/或所述一种或多种基因变异。

132. 如权利要求117所述的方法,其中所述母体核酸和所述父体核酸中的一者或多者为基因组脱氧核糖核酸(DNA)。

133. 如权利要求117所述的方法,其中在(c)中,所述胎儿核酸包括无细胞核酸。

134. 如权利要求117所述的方法,其进一步包括,在(a)中,以至少99%的准确度测定所述母体核酸序列。

135. 如权利要求117所述的方法,其进一步包括,在(b)中,以至少99%的准确度测定所述父体核酸序列。

136. 如权利要求117所述的方法,其中所述母体核酸序列和/或所述父体核酸序列具有大于10千碱基(kb)的长度。

137. 如权利要求117所述的方法,其中所述母体和父体分区包括乳液中的微滴。

138. 如权利要求117所述的方法,其中在(a)中,在所述给定母体分区中提供可释放地连接至第一粒子的所述第一条形码序列。

139. 如权利要求138所述的方法,其中在(b)中,在所述给定父体分区中提供可释放地连接至第二粒子的所述第二条形码序列。

140. 一种用于表征样品核酸的方法,其包括:

- (a) 从受试者获得生物样品,所述生物样品包括无细胞样品核酸;
- (b) 在微滴中,将条形码序列连接至所述无细胞样品核酸的片段或所述样品核酸的诸多个部分的拷贝,以提供条形码化样品片段;
- (c) 测定所述条形码化样品片段的核酸序列并且基于所述条形码化样品片段的所述核酸序列提供样品核酸序列;
- (d) 使用经过编程的计算机处理器来产生所述样品核酸序列与参考核酸序列的比较,所述参考核酸序列具有大于10千碱基(kb)的长度和至少99%的准确度;以及
- (e) 使用所述比较来识别所述样品核酸序列中的一种或多种基因变异,由此将所述样品核酸与疾病相关联。

141. 如权利要求140所述的方法,其中所述样品核酸序列中的所述一种或多种基因变异选自结构变异和单核苷酸多态性(SNP)。

142. 如权利要求140所述的方法,其中所述样品核酸序列的所述一种或多种基因变异为选自拷贝数变异、插入、缺失、反转录转座子、易位、倒位、重排、重复扩增以及重复的结构变异。

143. 如权利要求140所述的方法,其中在(c)中,以至少99%的准确度提供所述样品核酸序列。

144. 如权利要求140所述的方法,其中在(b)中,在所述微滴中提供可释放地连接至粒子的所述条形码序列,并且其中(b)进一步包括在所述连接所述条形码序列之前使所述条形码序列从所述粒子释放至所述微滴中。

145. 如权利要求144所述的方法,其中在(b)中,提供所述条形码序列作为可释放地连接至所述粒子的引物序列的一部分,其中所述引物序列还包括随机N-mer序列,并且其中(b)进一步包括在所述连接所述条形码序列之前使所述引物序列从所述粒子释放至所述微滴中。

146. 如权利要求145所述的方法,其进一步包括,在(b)中,在扩增反应中使用所述引物将所述条形码序列连接至所述无细胞样品核酸的所述片段或所述无细胞样品核酸的诸多个部分的所述拷贝。

147. 如权利要求140所述的方法,其进一步包括(i)在额外微滴中,将额外条形码序列连接至参考核酸的片段或所述参考核酸的诸多个部分的拷贝以提供条形码化参考片段;以及(ii)测定所述条形码化参考片段的核酸序列并且基于所述条形码化参考片段的所述核酸序列测定所述参考核酸序列。

148. 如权利要求147所述的方法,其中所述测定所述参考核酸序列包括组装所述条形码化参考片段的所述核酸序列。

149. 如权利要求147所述的方法,其进一步包括在所述额外微滴中提供可释放地连接至粒子的所述额外条形码序列,以及在所述连接所述额外条形码序列之前使所述额外条形码序列从所述粒子释放至所述额外分区中。

150. 如权利要求149所述的方法,其进一步包括提供所述额外条形码序列作为可释放地连接至所述粒子的引物序列的一部分,其中所述引物序列还包括随机N-mer序列,以及在所述连接所述额外条形码序列之前使所述引物从所述粒子释放至所述额外微滴中。

151. 如权利要求150所述的方法,其进一步包括在扩增反应中使用所述引物将所述额

外条形码序列连接至所述参考核酸的所述片段或所述参考核酸的诸多个部分的所述拷贝。

152. 如权利要求147所述的方法, 其进一步包括测定所述参考核酸序列中的一种或多种基因变异。

153. 如权利要求152所述的方法, 其中所述参考核酸序列中的所述一种或多种基因变异选自结构变异和单核苷酸多态性 (SNP)。

154. 如权利要求152所述的方法, 其中所述参考核酸序列中的所述一种或多种基因变异为选自拷贝数变异、插入、缺失、反转录转座子、易位、倒位、重排、重复扩增以及重复的结构变异。

155. 如权利要求147所述的方法, 其中所述参考核酸包括生殖系核酸序列。

156. 如权利要求147所述的方法, 其中所述参考核酸包括癌症核酸序列。

157. 如权利要求140所述的方法, 其中所述样品核酸序列具有大于10kb的长度。

158. 如权利要求147所述的方法, 其中所述参考核酸源于指示疾病状态不存在的基因组。

159. 如权利要求147所述的方法, 其中所述参考核酸源于指示疾病状态的基因组。

160. 如权利要求158或159中任一项所述的方法, 其中所述疾病状态包括癌症。

161. 如权利要求158或159中任一项所述的方法, 其中所述疾病状态包括非整倍性。

162. 如权利要求140所述的方法, 其中所述无细胞样品核酸包括肿瘤核酸。

163. 如权利要求162所述的方法, 其中所述肿瘤核酸包括循环肿瘤核酸。

核酸序列的分析

[0001] 交叉引用

[0002] 本申请要求2014年6月26日提交的美国临时专利申请号62/017,808以及2014年10月29日提交的美国临时专利申请号62/072,214的优先权,这些申请中的每一者出于所有目的以全文引用的方式并入本文中。

[0003] 背景

[0004] 对特定人基因组的基本理解可能不只需要简单识别存在或不存在某些基因变异,诸如突变。确定某些基因变异是出现在同一染色体还是不同染色体上(也称为定相(phasing))也是重要的。关于基因变异模式(诸如单倍型)的信息也是重要的,关于基因拷贝的数目的信息也重要。

[0005] 术语“单倍型”是指在连续区块(block)中一起遗传的DNA序列变体(等位基因)的集合。一般来说,人基因组含有各基因的两个拷贝-母体拷贝和父体拷贝。对于各自具有两个可能的等位基因的一对基因,例如基因等位基因“A”和“a”,以及基因等位基因“B”和“b”,给定个体的基因组将包括两种单倍型中的一种:“AB/ab”,其中A和B等位基因位于同一染色体上(“顺式”构型);或“Ab/aB”,其中A和B等位基因位于不同染色体上(“反式”构型)。可使用定相方法或分析来确定指定等位基因集合是位于同一染色体还是不同染色体上。在一些情况下,限定单倍型的若干相连等位基因可能与特定疾病表型相关联或相关;在此类情况下,就患者是否将展示疾病来说单倍型(而不是任何一种特定基因变体)可能是最具决定性的因素。

[0006] 基因拷贝数也在一些疾病表型中发挥作用。大多数基因通常以两个拷贝存在,然而,扩增的基因为以超过两个功能拷贝的形式存在的基因。在一些情况下,基因还可经历功能拷贝减少。基因拷贝数的减少或增加可导致异常mRNA和蛋白质表达水平的产生,从而潜在地导致癌性状态或其他病症。癌症和其他基因病症经常与异常(增加或减少)的染色体数目(“非整倍性”)相关联。可使用诸如荧光原位杂交或比较基因组杂交等细胞遗传技术来检测异常基因的存在或染色体拷贝数。本领域中需要改进的检测基因定相信息、单倍型或拷贝数变异的方法。

[0007] 概述

[0008] 本公开提供可适用于在基因材料的表征方面提供显著进步的方法和系统。这些方法和系统可适用于提供使用通常可获得的技术可能大体上较难实现的基因表征,包括例如单倍型定相、识别结构变异(例如缺失、重复、拷贝数变体、插入、倒位、易位、长串联重复(LTR)、短串联重复(STR))以及多种其他适用表征。

[0009] 本公开一个方面提供一种用于识别核酸中的一种或多种变异的方法,其包括:a)提供核酸的第一片段,其中第一片段具有大于10千碱基(kb)的长度;b)对第一片段的多个第二片段进行测序以提供多个片段序列,所述多个片段序列共有共同的条形码序列;c)根据共同条形码序列的存在将多个片段序列归属于第一片段;d)使用多个片段序列测定第一片段的核酸序列,其中以低于1%的误差率测定核酸序列;以及(e)识别(d)中所测定的第一片段的核酸序列中的一种或多种变异,由此识别核酸内的一种或多种变异。

[0010] 在一些情况下,第一片段在多个离散分区之中的离散分区中。在一些情况下,离散分区为乳液中的微滴。在一些情况下,识别包括识别第一片段的核酸序列中的定相变体。在一些情况下,识别包括由第一片段的核酸序列识别核酸中的一种或多种结构变异。在一些情况下,第一片段具有大于15kb的长度。在一些情况下,第一片段具有大于20kb的长度。在一些情况下,测定包括将多个片段序列映射至参考序列。在一些情况下,测定包括组装具有共同条形码序列的多个片段序列。

[0011] 在一些情况下,用于识别一种或多种变异的方法进一步包括提供核酸的多个第一片段,其长度为至少10kb,并且识别包括测定多个第一片段中的每一者的核酸序列以及由多个第一片段中的每一者的核酸序列识别核酸中的一种或多种变异。

[0012] 在一些情况下,用于识别一种或多种变异的方法进一步包括基于多个第一片段的两个或更多个核酸序列的重叠核酸序列将两个或更多个核酸序列连接成推测重叠群(inferred contig),其中最大推测重叠群长度为至少10kb。在一些情况下,最大推测重叠群长度为至少20kb。在一些情况下,最大推测重叠群长度为至少40kb。在一些情况下,最大推测重叠群长度为至少50kb。在一些情况下,最大推测重叠群长度为至少100kb。在一些情况下,最大推测重叠群长度为至少200kb。在一些情况下,最大推测重叠群长度为至少500kb。在一些情况下,最大推测重叠群长度为至少750kb。在一些情况下,最大推测重叠群长度为至少1兆碱基(Mb)。在一些情况下,最大推测重叠群长度为至少1.75Mb。在一些情况下,最大推测重叠群长度为至少2.5Mb。

[0013] 在一些情况下,识别一种或多种变异的方法进一步包括基于多个第一片段的两个或更多个核酸序列内的重叠定相变体将多个第一片段的两个或更多个核酸序列连接成相位区块(phase block),其中最大相位区块长度为至少10kb。在一些情况下,最大相位区块长度为至少20kb。在一些情况下,最大相位区块长度为至少40kb。在一些情况下,最大相位区块长度为至少50kb。在一些情况下,最大相位区块长度为至少100kb。在一些情况下,最大相位区块长度为至少200kb。在一些情况下,最大相位区块长度为至少500kb。在一些情况下,最大相位区块长度为至少750kb。在一些情况下,最大相位区块长度为至少1Mb。在一些情况下,最大相位区块长度为至少1.75Mb。在一些情况下,最大相位区块长度为至少2.5Mb。

[0014] 在一些情况下,用于识别一种或多种变异的方法进一步包括基于多个第一片段的两个或更多个核酸序列的重叠核酸序列将两个或更多个核酸序列连接成推测重叠群,由此形成推测重叠群的群体,其中推测重叠群的群体的N50为至少10kb。在一些情况下,推测重叠群的群体的N50为至少20kb。在一些情况下,推测重叠群的群体的N50为至少40kb。在一些情况下,推测重叠群的群体的N50为至少50kb。在一些情况下,推测重叠群的群体的N50为至少100kb。在一些情况下,推测重叠群的群体的N50为至少200kb。在一些情况下,推测重叠群的群体的N50为至少500kb。在一些情况下,推测重叠群的群体的N50为至少750kb。在一些情况下,推测重叠群的群体的N50为至少1Mb。在一些情况下,推测重叠群的群体的N50为至少1.75Mb。在一些情况下,推测重叠群的群体的N50为至少2.5Mb。

[0015] 在一些情况下,用于识别一种或多种变异的方法进一步包括基于多个第一片段的两个或更多个核酸序列内的重叠定相变体将多个第一片段的两个或更多个核酸序列连接成相位区块,由此形成相位区块的群体,其中相位区块的群体的N50为至少10kb。在一些情况下,相位区块的群体的N50为至少20kb。在一些情况下,相位区块的群体的N50为至少

40kb。在一些情况下,相位区块的群体的N50为至少50kb。在一些情况下,相位区块的群体的N50为至少100kb。在一些情况下,相位区块的群体的N50为至少200kb。在一些情况下,相位区块的群体的N50为至少500kb。在一些情况下,相位区块的群体的N50为至少750kb。在一些情况下,相位区块的群体的N50为至少1Mb。在一些情况下,相位区块的群体的N50为至少1.75Mb。在一些情况下,相位区块的群体的N50为至少2.5Mb。

[0016] 本公开的额外方面提供一种用于确定核酸结构变异的存在的方法。所述方法可包括:(a)提供核酸的多个第一片段分子,其中多个第一片段分子中的给定第一片段分子包含结构变异;(b)对多个第一片段分子中的每一者的多个第二片段分子进行测序以提供多个片段序列,其中多个片段序列中对应于给定第一片段分子的每一者共有共同的条形码序列;以及(c)通过以下方式确定结构变异的存在:(i)将多个片段序列映射至参考序列,(ii)识别共有共同的条形码序列的多个片段序列,并且(iii)基于在比给定第一片段分子的长度相隔更远的位置存在升高量的映射至参考序列的共有共同条形码序列的多个片段序列来识别结构变异,所述升高量是相对于缺少结构变异的序列。

[0017] 在一些情况下,升高量相对于源于核酸中具有结构变异的区域的第一片段分子的总数为1%或更多。在一些情况下,升高量相对于源于核酸中具有结构变异的区域的第一片段分子的总数为2%或更多。在一些情况下,所述位置相隔至少约100个碱基。在一些情况下,所述位置相隔至少约500个碱基。在一些情况下,所述位置相隔至少约1千碱基(kb)。在一些情况下,所述位置相隔至少约10kb。

[0018] 在一些情况下,确定核酸的结构变异的存在的方法进一步包括通过由多个片段序列形成给定第一片段分子的组装物(assembly)来识别结构变异,其中基于共同条形码序列的存在选择多个片段序列作为组装物的输入物(input)。在一些情况下,组装物通过由多个片段序列产生共有序列来形成。在一些情况下,结构变异包括易位。

[0019] 本公开的额外方面提供一种表征变体核酸序列的方法。在一些情况下,所述方法可包括:(a)将变体核酸片段化以提供具有大于10千碱基(kb)的长度的多个第一片段;(b)将多个第一片段分离至离散分区中;(c)从各个第一片段各自的分区内的各个第一片段形成多个第二片段,所述多个第二片段具有与其连接的条形码序列,给定分区内的所述条形码序列为共同条形码序列;(d)对多个第二片段和与其连接的条形码序列进行测序,以提供多个第二片段序列;(e)至少部分基于共同条形码序列的存在将第二片段序列归属于原始第一片段,以提供第二片段序列的第一片段序列环境(context);以及(f)由第一片段序列环境识别变体核酸的变体部分,由此表征变体核酸序列。在一些情况下,归属包括至少部分基于共同条形码序列的存在由多个第二片段序列组装多个第一片段中的单个片段的序列的至少一部分。在一些情况下,归属包括至少部分基于共同条形码序列将多个第二片段序列映射至多个第一片段中的单个第一片段。

[0020] 在一些情况下,表征变体核酸序列的方法进一步包括基于多个第一片段中的两者或更多者之间的重叠序列将多个第一片段中的两者或更多者连接成推测重叠群。在一些情况下,识别包括由第一片段序列环境识别一个或多个定相变体。在一些情况下,表征变体核酸序列的方法进一步包括基于多个第一片段中的两者或更多者之间的重叠定相变体将多个第一片段中的两者或更多者连接成相位区块。在一些情况下,识别包括由第一片段序列环境识别一种或多种结构变异。在一些情况下,一种或多种结构变异独立地选自插入、缺

失、易位、反转录转座子、倒位以及重复。在一些情况下，结构变异包括插入或易位，并且第一片段序列环境指示插入或易位的存在。

[0021] 本公开的额外方面提供一种识别核酸序列中的变体的方法。在一些情况下，所述方法包括：获得核酸的多个单个片段分子的核酸序列，多个单个片段分子的核酸序列各自具有至少1千碱基(kb)的长度；将多个单个片段分子中的一者或多者的序列连接成一个或多个推测重叠群；以及由一个或多个推测重叠群识别一个或多个变体。在一些情况下，获得包括获得长度大于10kb的多个片段分子的核酸序列。在一些情况下，获得包括：提供多个单个片段分子中的各个单个片段分子的多个条形码化片段，给定单个片段分子的条形码化片段具有共同条形码；对多个单个片段分子的多个条形码化片段进行测序，测序提供低于1%的测序误差率；以及由多个条形码化片段和其相关条形码的序列测定多个单个片段分子的序列。

[0022] 在一些情况下，连接包括识别两个或更多个单个片段分子之间的一个或多个重叠序列以将两个或更多个单个片段分子连接成一个或多个推测重叠群。在一些情况下，连接包括识别两个或更多个单个片段分子之间的一个或多个共同变体以将两个或更多个单个片段分子连接成一个或多个推测重叠群。在一些情况下，一个或多个共同变体为定相变体，并且一个或多个推测重叠群包含至少100kb的最大相位区块长度。在一些情况下，所述识别中所识别的一个或多个变体包含结构变异。在一些情况下，结构变异选自插入、缺失、易位、反转录转座子、倒位以及重复。

[0023] 本公开的额外方面提供一种表征核酸的方法。在一些情况下，所述方法包括：获得具有至少10千碱基(kb)的长度的多个片段分子的核酸序列；识别多个片段分子的核酸序列中的一个或多个定相变体位置；基于第一和第二片段分子内的一个或多个共同定相变体位置的存在将至少一个第一片段分子的核酸序列连接至至少一个第二片段分子，以提供具有至少10kb的最大相位区块长度的相位区块；以及由具有至少10kb的最大相位区块长度的相位区块识别一个或多个定相变体。在一些情况下，表征核酸的方法进一步包括由相位区块识别一个或多个额外定相变体。在一些情况下，多个片段分子在离散分区中。在一些情况下，离散分区为乳液中的微滴。在一些情况下，多个片段分子的长度为至少50kb。在一些情况下，多个片段分子的长度为至少100kb。在一些情况下，最大相位区块长度为至少50kb。在一些情况下，最大相位区块长度为至少100kb。在一些情况下，最大相位区块长度为至少1Mb。在一些情况下，最大相位区块长度为至少2Mb。在一些情况下，最大相位区块长度为至少2.5Mb。

[0024] 本公开的额外方面提供一种方法，其包括：(a) 将第一核酸分配至第一分区中，其中第一核酸包含源于生物体的第一染色体的靶序列；(b) 将第二核酸分配至第二分区中，其中第二核酸包含源于生物体的第二染色体的靶序列；(c) 在第一分区中，将第一条形码序列连接至第一核酸的片段或第一核酸的诸多个部分的拷贝以提供第一条形码化片段；(d) 在第二分区中，将第二条形码序列连接至第二核酸的片段或第二核酸的诸多个部分的拷贝以提供第二条形码化片段，第二条形码序列不同于第一条形码序列；(e) 测定第一和第二条形码化片段的核酸序列，并且组装第一和第二核酸的核酸序列；以及(f) 比较第一和第二核酸的核酸序列，以分别将第一和第二核酸表征为源于第一和第二染色体。在一些情况下，将包含第一条形码序列的寡核苷酸与第一核酸共分配，并且将包含第二条形码序列的寡核苷酸

与第二核酸共分配。在一些情况下,将包含第一条形码序列的寡核苷酸可释放地连接至第一珠粒,并且将包含第二条形码序列的寡核苷酸可释放地连接至第二珠粒,并且共分配包括分别将第一和第二珠粒共分配至第一和第二分区中。在一些情况下,第一和第二分区包括乳液中的微滴。在一些情况下,第一染色体为父体染色体,并且第二染色体为母体染色体。在一些情况下,第一染色体和第二染色体为同源染色体。在一些情况下,第一核酸和第二核酸包含一种或多种变异。

[0025] 在一些情况下,第一和第二染色体源于胎儿。在一些情况下,第一和第二核酸从取自孕妇的样品获得。在一些情况下,第一染色体为染色体21、18或13。在一些情况下,第二染色体为染色体21、18或13。在一些情况下,所述方法进一步包括测定第一或第二染色体的相对量。在一些情况下,所述方法进一步包括测定第一或第二染色体相对于参考染色体的量。在一些情况下,第一染色体或第二染色体或两者的拷贝数增加。在一些情况下,拷贝数增加是癌症或非整倍性的结果。在一些情况下,第一染色体或第二染色体或两者的拷贝数减少。在一些情况下,拷贝数减少是癌症或非整倍性的结果。

[0026] 本公开的额外方面提供一种方法,其包括:(a)将第一核酸分配至第一分区中,其中第一核酸包含源于生物体的第一染色体的靶序列;(b)将第二核酸分配至第二分区中,其中第二核酸包含源于生物体的第二染色体的靶序列;(c)在第一分区中,将第一条形码序列连接至第一核酸的片段或第一核酸的诸多个部分的拷贝以提供第一条形码化片段;(d)在第二分区中,将第二条形码序列连接至第二核酸的片段或第二核酸的诸多个部分的拷贝以提供第二条形码化片段,第二条形码序列不同于第一条形码序列;(e)测定第一和第二条形码化片段的核酸序列,并且组装第一和第二核酸的核酸序列;以及(f)比较第一和第二核酸的核酸序列,以识别第一和第二核酸的核酸序列之间的任何变异。在一些情况下,将包含第一条形码序列的寡核苷酸与第一核酸共分配,并且将包含第二条形码序列的寡核苷酸与第二核酸共分配。在一些情况下,将包含第一条形码序列的寡核苷酸可释放地连接至第一珠粒,并且将包含第二条形码序列的寡核苷酸可释放地连接至第二珠粒,并且共分配包括分别将第一和第二珠粒共分配至第一和第二分区中。在一些情况下,第一和第二分区包括乳液中的微滴。在一些情况下,第一染色体为父体染色体,并且第二染色体为母体染色体。在一些情况下,第一染色体和第二染色体为同源染色体。在一些情况下,第一核酸和第二核酸包含一种或多种变异。在一些情况下,第一和第二染色体源于胎儿。在一些情况下,第一和第二核酸从取自孕妇的样品获得。在一些情况下,第一染色体为染色体21、18或13。在一些情况下,第二染色体为染色体21、18或13。在一些情况下,所述方法进一步包括测定第一或第二染色体的相对量。在一些情况下,所述方法进一步包括测定第一或第二染色体相对于参考染色体的量。在一些情况下,第一染色体或第二染色体或两者的拷贝数增加。在一些情况下,拷贝数增加是癌症或非整倍性的结果。在一些情况下,第一染色体或第二染色体或两者的拷贝数减少。在一些情况下,拷贝数减少是癌症或非整倍性的结果。

[0027] 本公开的额外方面提供一种用于表征胎儿核酸序列的方法。在一些情况下,所述方法包括:(a)通过以下方式测定母体核酸序列,其中母体核酸源于胎儿的怀孕母亲:(i)将母体核酸片段化以提供多个第一母体片段;(ii)将多个第一母体片段分离至母体分区中;(iii)从第一母体片段各自的母体分区内的第一母体片段中的每一者形成多个第二母体片段,多个第二母体片段具有与其连接的第一条形码序列,其中在母体分区中的给定母体分

区内,第二母体片段包含与其连接的第一共同条形码序列;(iv)对多个第二母体片段进行测序以提供多个母体片段序列;(v)至少部分基于第一共同条形码序列的存在将母体片段序列归属于原始第一母体片段以测定母体核酸序列;(b)通过以下方式测定父体核酸序列,其中父体核酸源于胎儿的父亲:(i)将父体核酸片段化以提供多个第一父体片段;(ii)将多个第一父体片段分离至父体离散分区中;(iii)从各个第一父体片段各自的分区内的各个第一父体片段形成多个第二父体片段,多个第二父体片段具有与其连接的第二条形码序列,其中在给定父体分区内,第二父体片段包含与其连接的第二共同条形码序列;(iv)对多个第二父体片段和与其连接的第二条形码序列进行测序,以提供多个父体片段序列;(v)至少部分基于第二共同条形码序列的存在将父体片段序列归属于原始第一父体片段以测定父体核酸序列;(c)从怀孕母亲获得胎儿核酸并且使用母体核酸序列和父体核酸序列测定胎儿核酸的序列和/或胎儿核酸的序列的一种或多种基因变异。

[0028] 在一些情况下,使用父体片段序列和母体片段序列各者来将序列连接成一个或多个推测重叠群。在一些情况下,使用推测重叠群来构建母体和父体相位区块。在一些情况下,将胎儿核酸的序列与母体和父体相位区块相比较以构建胎儿相位区块。在一些情况下,组装父体片段序列以产生多个第一父体片段的序列的至少一部分,由此测定父体核酸序列,并且其中组装母体片段序列以产生多个第一母体片段的序列的至少一部分,由此测定母体核酸序列。在一些情况下,测定父体核酸序列包括将父体片段序列映射至父体参考序列,并且其中测定母体核酸序列包括将母体片段序列映射至母体参考序列。

[0029] 在一些情况下,以至少99%的准确度测定胎儿核酸的序列。在一些情况下,以至少99%的准确度测定胎儿核酸的序列的一种或多种基因变异。在一些情况下,一种或多种基因变异选自结构变异和单核苷酸多态性(SNP)。在一些情况下,一种或多种基因变异为选自拷贝数变异、插入、缺失、易位、反转录转座子、倒位、重排、重复扩增以及重复的结构变异。

[0030] 在一些情况下,用于表征胎儿核酸序列的方法进一步包括,在(c)中,使用针对母体核酸序列和父体核酸序列所测定的一种或多种基因变异测定胎儿核酸的序列的一种或多种基因变异。在一些情况下,用于表征胎儿核酸序列的方法进一步包括在(c)中,测定胎儿核酸的一个或多个从头突变。在一些情况下,用于表征胎儿核酸序列的方法进一步包括,在(c)期间或之后,测定与胎儿核酸相关的非整倍性。

[0031] 在一些情况下,用于表征胎儿核酸序列的方法进一步包括,在(a)中在(v)期间或之后,对母体核酸序列进行单倍型分析以提供单倍型解析型母体核酸序列,以及在(b)中在(v)期间或之后,对父体核酸序列进行单倍型分析以提供单倍型解析型父体核酸序列。在一些情况下,用于表征胎儿核酸序列的方法进一步包括在(c)中,使用单倍型解析型母体核酸序列和单倍型解析型父体核酸序列测定胎儿核酸的序列和/或一种或多种基因变异。在一些情况下,母体核酸和父体核酸中的一者或多者为基因组脱氧核糖核酸(DNA)。在一些情况下,在(c)中,胎儿核酸包括无细胞核酸。在一些情况下,用于表征胎儿核酸序列的方法进一步包括,在(a)中,以至少99%的准确度测定母体核酸序列。在一些情况下,表征胎儿核酸序列的方法进一步包括,在(b)中,以至少99%的准确度测定父体核酸序列。

[0032] 在一些情况下,母体核酸序列和/或父体核酸序列具有大于10千碱基(kb)的长度。在一些情况下,母体和父体分区包括乳液中的微滴。在一些情况下,在(a)中,在给定母体分区中提供可释放地连接至第一粒子的第一条形码序列。在一些情况下,在(b)中,在给定父

体分区中提供可释放地连接至第二粒子的第二条形码序列。

[0033] 本公开的额外方面提供一种用于表征样品核酸的方法。在一些情况下,方法包括:(a)从受试者获得生物样品,所述生物样品包括无细胞样品核酸;(b)在微滴中,将条形码序列连接至无细胞样品核酸的片段或样品核酸的诸多个部分的拷贝,以提供条形码化样品片段;(c)测定条形码化样品片段的核酸序列并且基于条形码化样品片段的核酸序列提供样品核酸序列;(d)使用经过编程的计算机处理器来产生样品核酸序列与参考核酸序列的比较,所述参考核酸序列具有大于10千碱基(kb)的长度和至少99%的准确度;以及(e)使用所述比较来识别样品核酸序列中的一种或多种基因变异,由此将样品核酸与疾病相关联。在一些情况下,样品核酸序列中的一种或多种基因变异选自结构变异和单核苷酸多态性(SNP)。在一些情况下,样品核酸序列的一种或多种基因变异为选自拷贝数变异、插入、缺失、反转录转座子、易位、倒位、重排、重复扩增以及重复的结构变异。在一些情况下,在(c)中,以至少99%的准确度提供样品核酸序列。在一些情况下,在(b)中,在微滴中提供可释放地连接至粒子的条形码序列,并且其中(b)进一步包括在连接条形码序列之前使条形码序列从粒子释放至微滴中。在一些情况下,在(b)中,提供条形码序列作为可释放地连接至的粒子的引物序列的一部分,其中引物序列还包括随机N-mer序列,并且其中(b)进一步包括在连接条形码序列之前使引物序列从粒子释放至微滴中。在一些情况下,在(b)中,在扩增反应中使用引物将条形码序列连接至无细胞样品核酸的片段或无细胞样品核酸的诸多个部分的拷贝。

[0034] 在一些情况下,用于表征样品核酸的方法进一步包括:(i)在额外微滴中,将额外条形码序列连接至参考核酸的片段或参考核酸的诸多个部分的拷贝以提供条形码化参考片段;以及(ii)测定条形码化参考片段的核酸序列并且基于条形码化参考片段的核酸序列测定参考核酸序列。在一些情况下,测定参考核酸序列包括组装条形码化参考片段的核酸序列。在一些情况下,用于表征样品核酸的方法进一步包括在额外微滴中提供可释放地连接至粒子的额外条形码序列以及在连接额外条形码序列之前使额外条形码序列从粒子释放至额外分区中。在一些情况下,用于表征样品核酸的方法进一步包括提供额外条形码序列作为可释放地连接至粒子的引物序列的一部分,其中引物序列还包括随机N-mer序列,并且在连接额外条形码序列之前使引物从粒子释放至额外微滴中。在一些情况下,用于表征样品核酸的方法进一步包括在扩增反应中使用引物将额外条形码序列连接至参考核酸的片段或参考核酸的诸多个部分的拷贝。在一些情况下,用于表征样品核酸的方法进一步包括测定参考核酸序列中的一种或多种基因变异。

[0035] 在一些情况下,参考核酸序列中的一种或多种基因变异选自结构变异和单核苷酸多态性(SNP)。在一些情况下,参考核酸序列中的一种或多种基因变异为选自拷贝数变异、插入、缺失、反转录转座子、易位、倒位、重排、重复扩增以及复制的结构变异。在一些情况下,参考核酸包括生殖系核酸序列。在一些情况下,参考核酸包括癌症核酸序列。在一些情况下,样品核酸序列具有大于10kb的长度。在一些情况下,参考核酸源于指示疾病状态不存在的基因组。在一些情况下,参考核酸源于指示疾病状态的基因组。在一些情况下,疾病状态包括癌症。在一些情况下,疾病状态包括非整倍性。在一些情况下,无细胞样品核酸包括肿瘤核酸。在一些情况下,肿瘤核酸包括循环肿瘤核酸。

[0036] 由以下详细描述本公开的额外方面和优点对本领域技术人员来说将变得轻易显

而易见,其中仅示出和描述了本公开的说明性实施方案。如将认识到的,本公开能够实现其他和不同实施方案,并且在各个明显的方面其若干细节能够进行修改,所有这些都不脱离本公开。因此,图式和描述将被视为在本质上是说明性的,而不是限制性的。

[0037] 以引用的方式并入

[0038] 本说明书中提到的所有出版物、专利以及专利申请以全文引用的方式并入本文中,其程度如同每个单个出版物、专利或专利申请被具体地和单个地指示以引用的方式并入一般。

[0039] 附图简述

[0040] 本发明的新颖特征在所附权利要求书中被特别阐述。通过参考以下使用本发明原理阐述说明性实施方案的详细描述以及附图将获得对本发明的特征和优点的更好的理解,在附图中:

[0041] 图1提供使用常规方法与本文所描述的示例性方法和系统来识别和分析定相变体的示意性说明。

[0042] 图2提供使用常规方法与本文所描述的示例性方法和系统来识别和分析结构变异的示意性说明。

[0043] 图3说明使用本文所公开的方法和组合物进行分析来检测拷贝数或单倍型的示例性工作流程。

[0044] 图4提供将核酸样品与珠粒组合并且将核酸和珠粒分配至离散微滴中的示例性方法的示意性说明。

[0045] 图5提供条形码化和扩增染色体核酸片段的示例性方法的示意性说明。

[0046] 图6提供条形码化染色体核酸片段在将序列数据归属于单个染色体中的示例性用途的示意性说明。

[0047] 图7提供定相测序方法的实例的示意性说明。

[0048] 图8提供健康患者(上部图)以及单倍型拷贝数增加(中部图)或单倍型拷贝数减少(下部图)的癌症患者的基因组的示例性子集的示意性说明。

[0049] 图9A-B提供:(a)示出肿瘤DNA的相对贡献的示意性说明,以及(b)通过普通测序方法检测此类拷贝增加和减少的图示。

[0050] 图10提供使用单一变体位置(左侧图)和组合变体位置(右侧图)检测拷贝增加和减少的实例的示意性说明。

[0051] 图11提供所描述的方法和系统用于识别拷贝数增加和减少的潜力的示意性说明。

[0052] 图12说明基于使用本文所描述的方法和组合物对染色体数目和拷贝数变异的测定来进行非整倍性测试的示例性工作流程。

[0053] 图13A-B说明用于识别基因样品中的诸如易位和基因融合等结构变异的方法的示例性概括。

[0054] 图14说明基于使用本文所描述的方法和组合物对拷贝数变异的测定来进行癌症诊断测试的示例性工作流程。

[0055] 图15提供NCI-H2228癌症细胞系的EML-4-ALK结构变异的示意性说明。

[0056] 图16A和16B使用本文所描述的系统提供条形码映射数据,以便识别与阴性对照细胞系(图16B)相比在癌细胞系(图16A)中存在图15中所示的EML-4-ALK变体结构。

[0057] 图17示意性描述如本文所描述分析父体核酸序列的示例性工作流程。

[0058] 图18示意性描绘如本文所描述分析母体核酸序列的示例性工作流程。

[0059] 图19示意性描绘如本文所描述分析胎儿核酸序列的示例性工作流程。

[0060] 图20示意性描绘如本文所描述分析参考核酸序列的示例性工作流程。

[0061] 图21示意性描绘如本文所描述分析样品核酸序列的示例性工作流程。

[0062] 图22示意性描绘示例性计算机控制系统。

[0063] 详细描述

[0064] 虽然本文中已示出和描述了本发明的各个实施方案,但对本领域技术人员来说将显而易见的是此类实施方案仅仅是通过举例而提供。本领域技术人员会想到许多变化、改变以及替换,而不会脱离本发明。应了解,可采用对本文中所描述的本发明实施方案的各种替代方案。

[0065] 如本文中所用,术语“生物体”通常是指连续活系统。生物体的非限制性实例包括动物(例如人、其他类型的哺乳动物、鸟类、爬行动物、昆虫、本文中别处所描述的其他示例性类型的动物)、植物、真菌以及细菌。

[0066] 如本文中所用,术语“重叠群”通常是指给定长度的连续核酸序列。连续序列可源于单个序列读段,包括短读段或长读段序列读段,或来自序列读段的组装物,所述序列读段是比对过的并且基于读段内的重叠序列被组装或基于其他已知连接数据(例如如本文中别处所描述用共同条形码标记)被定义为在片段内相连接。这些重叠序列读段可同样包括短读段,例如小于500个碱基,例如在一些情况下,约100至500个碱基,并且在一些情况下,100至250个碱基,或基于更长的序列读段,例如大于500个碱基、1000个碱基或甚至大于10,000个碱基。

[0067] I. 概括

[0068] 本公开提供适用于在基因材料的表征方面提供显著进步的方法和系统。在一些情况下,所述方法和系统可适用于提供使用通常可获得的技术非常难或甚至不可能实现的基因表征,包括例如单倍型定相、识别结构变异(例如缺失、重复、拷贝数变体、插入、倒位、反转录转座子、易位、LTR、STR)以及多种其他适用表征。

[0069] 一般来说,本文所描述的方法和系统通过提供对长单个核酸分子的测序来完成以上目标,所述测序允许识别和使用长范围变体信息,例如将变异与不同序列区段相联系,包括含有其他变异的序列区段,其在起源序列中隔开例如长于由短读段测序技术所提供的显著距离。然而,这些方法和系统实现这些目标,具有短读段测序技术的极低的并且远低于报告过的长读段长度测序技术(例如单分子测序,诸如SMRT测序和纳米孔测序技术)的测序误差率的优点。

[0070] 一般来说,本文所描述的方法和系统将长核酸分子区段化成更小片段,所述更小片段是使用高通量、更高准确度短读段测序技术可测序的,但此类区段化是以允许源于更小片段的序列信息被归属于更长的单个起源核酸分子的方式进行。通过将序列读段归属于更长的起源核酸分子,可获得关于所述更长核酸序列的显著表征信息,这是单独由短序列读段通常不能获得的。如所提到的,此类表征信息可包括单倍型定相、识别结构变异以及识别拷贝数变异。

[0071] 本文所描述的方法和系统的优点是关于大量一般实例来描述。在第一实例中,使

用本文所描述的方法和系统识别和表征定相序列变体。图1示意性说明了定相变体识别(phased variant calling)的挑战以及通过本文所描述的方法提供的解决方案。如图所示,图I中的核酸102和104代表不同染色体(例如母体和父体遗传染色体)的相同区域的两个单倍体序列。各序列在表征各单倍体序列的不同等位基因处包括一系列变体,例如核酸102上的变体106-114,以及核酸104上的变体116-122。由于其序列读段非常短,所以大多数测序技术不能提供单个变体相对于同一单倍体序列上的其他变体的环境。另外,因为它们依赖于不分离单个分子组分(例如各单倍体序列)的样品制备技术,所以不能识别各种变体的定相,例如变体所源于的单倍体序列。因此,这些短读段技术不能将这些变体解析至其起源分子。图IIa和IIIa示意性说明了使用此方法的困难。简要地说,对图IIa中所示的从两个单倍体序列汇集的片段进行测序,产生大量短序列读段124,并且组装所得序列126(图IIIa中所示)。如图所示,因为没有图IIa中的更短序列读段中的任一者的相对定相环境,所以在组装过程中将不能解析两个不同单倍体序列之间的变体。因此,图IIIa中所示的所得组装物产生单一共有序列组装物126,包括所有变体106-122。

[0072] 相比之下,并且如图1的图IIb中所示,如使用上文所描述的方法,本文所描述的方法和系统将更长核酸102和104分解或区段化成更短的可测序片段,但在那些片段存在下保留将其归属于其起源分子环境的能力。图IIb中对此进行了示意性说明,其中将不同片段根据其起源分子环境进行分组或“区室化(compartmentalize)”。在本公开的上下文中,此分组可通过将片段物理分配至保留分子环境的诸多个组中以及标记那些片段以随后能够阐明环境中的一者或两者来完成。

[0073] 此分组被示意性说明为在分别代表来自核酸102和104的短序列读段的群组128和130之间分派较短序列读段。因为在整个测序过程中起源序列环境被保留,所以可将所述环境用于解析原始分子环境,例如分别定相序列102和104之间的各种变体106-114和116-122。

[0074] 在另一示例性有利应用中,所述方法和系统适用于表征使用短读段序列技术通常不可识别或至少难以识别的结构变体。

[0075] 图2中参考简单易位事件对此进行了示意性说明。如图所示,基因组样品可包括核酸,所述核酸包括易位事件,例如基因元件206从序列202至序列204的易位。此类易位可为多种不同易位类型中的任一种,包括例如不同染色体之间(无论是到相同的还是不同的区域)、同一染色体的不同区域之间的易位。

[0076] 再次,如在上文的图1中所说明的实例中,常规测序是通过将图I中的序列202和204分解成小片段并且如图IIa中所示从那些片段产生短序列读段208而开始。因为这些序列片段208为相对短的,所以在组装过程期间易位序列206的环境(即起源于相同或不同序列上的变体位置时)容易丢失。此外,由于其短读段长度,经常预测在使用参考序列时序列组装几乎根据定义就不会反映结构变异。因此,将不变地无视易位序列206的适当位置而组装短序列读段208,并且实际上将如图IIIa中所示组装非变体序列210和212。

[0077] 相比之下,使用本文所描述的方法和系统,为源于序列202和204的短序列读段提供区室化,在图IIb中被示出为群组214和216,其保留更小序列片段的原始分子分组,从而允许其组装为图IIIb中所示的序列218和220,从而允许归属回到起源序列202和204,并且识别易位变异,例如如图IIIb中所说明的恰当序列组装物218和220中的易位序列区段

206a。

[0078] 如上文所提到,本文所描述的方法和系统提供更长核酸的短序列读段的单个分子环境。如本文中所用,单个分子环境是指特定序列读段以外的序列环境,例如与邻近或接近序列相关,所述邻近或接近序列不包括在序列读段本身内,并且因此,通常将使得它们不会被整个或部分地包括在短序列读段中,例如约150个碱基或对于成对读段来说约300个碱基的读段。在一些方面,所述方法和系统提供短序列读段的长范围序列环境。此类长范围环境包括给定序列读段与彼此距离长于1千碱基(kb)、长于5kb、长于10kb、长于15kb、长于20kb、长于30kb、长于40kb、长于50kb、长于60kb、长于70kb、长于80kb、长于90kb或甚至长于100kb之内或更长的序列读段的关系或联系。通过提供更长范围单个分子环境,本文所描述的方法和系统还提供长得多的推测分子环境。序列环境如本文所描述可包括更低解析环境,例如由将短序列读段映射至单个更长分子或相连分子的重叠群;以及更高解析序列环境,例如由更长单个分子中例如具有单个分子的连续的所测定序列的较大部分的长范围测序,其中此类所测定序列长于1kb、长于5kb、长于10kb、长于15kb、长于20kb、长于30kb、长于40kb、长于50kb、长于60kb、长于70kb、长于80kb、长于90kb或甚至长于100kb。如在序列环境的情况下,将短序列归属于更长核酸(例如均为单个长核酸分子或相连核酸分子的集合或重叠群)可包括将短序列映射至更长核酸区段以提供高水平序列环境,以及由短序列通过这些更长核酸提供组装序列。

[0079] 此外,虽然可利用与长单个分子相关的长范围序列环境,但具有此类长范围序列环境还允许推测甚至更长范围的序列环境。举一个例子,通过提供上文所描述的长范围分子环境,可识别长序列之中来自不同起源分子的重叠变体部分,例如定相变体、易位序列等,从而允许得到那些分子之间的推测联系。此类推测联系或分子环境在本文中被称作“推测重叠群”。在一些情况下,当在定相序列的背景下论述时,推测重叠群通常可代表定相序列,例如在凭借重叠定相变体的情况下,可推测长度大体上大于单个起源分子的定相重叠群。这些定相重叠群在本文中被称作“相位区块”。

[0080] 通过以更长单分子读段开始,可得到比使用短读段测序技术或其他定相测序方法以其他方式将可达到的更长的推测重叠群或相位区块。参见例如已出版的美国专利公布号2013/0157870,该专利公布的全部公开内容以全文引用的方式并入本文中。特定而言,使用本文所描述的方法和系统,可获得N50(重叠群或相位区块长度,对于所述重叠群或相位区块长度,具有所述长度或更长的所有相位区块或重叠群的集合含有所有重叠群或相位区块的长度总和的至少一半,并且对于所述重叠群或相位区块长度,具有所述长度或更短的所有重叠群或相位区块的集合也含有所有重叠群或相位区块的长度总和的至少一半)、众数、平均值或中值为至少约10千碱基(kb)、至少约20kb、至少约50kb的推测重叠群或相位区块长度。在一些方面,推测重叠群或相位区块长度的N50、众数、平均值或中值为至少约100kb、至少约150kb、至少约200kb,并且在一些情况下为至少约250kb、至少约300kb、至少约350kb、至少约400kb,并且在一些情况下达到至少约500kb、至少约750kb、至少约1Mb、至少约1.75Mb、至少约2.5Mb或更大。在其他情况下,可获得至少或超过20kb、40kb、50kb、100kb、200kb、300kb、400kb、500kb、750kb、1兆碱基(Mb)、1.75Mb、2Mb或2.5Mb的最大推测重叠群或相位区块长度。在其他情况下,推测重叠群或相位区块长度可为至少约20kb、至少约40kb、至少约50kb、至少约100kb、至少约200kb,并且在一些情况下为至少约500kb、至少约750kb、

至少约1Mb,并且在一些情况下为至少约1.75Mb、至少约2.5Mb或更长。

[0081] 在一个方面,本文所描述的方法和系统提供样品核酸或其片段区室化、沉积或分配至离散区室或分区(本文中可互换地称为分区)中,其中各分区保持其自己的内容物与其他分区的内容物隔开。可事先、随后或同时将独特的标识(例如条形码)递送至容纳被区室化或分配的样品核酸的分区,以允许随后将特征(例如核酸序列信息)归属于特定区室内所包括的样品核酸,并且特别是归属于可能最初沉积至分区中的相对长的连续样品核酸区段。

[0082] 可对样品核酸进行分配使得核酸以相对长的连续核酸分子片段或区段存在于分区中。这些片段可代表所要分析的全部样品核酸的许多重叠片段,例如整个染色体、外显子组或其他大基因组片段。这些样品核酸可包括全基因组、单个染色体、外显子组、扩增子或所关注的多种不同核酸中的任一种。在一些情况下,样品核酸的这些片段可长于100个碱基、长于500个碱基、长于1kb、长于5kb、长于10kb、长于15kb、长于20kb、长于30kb、长于40kb、长于50kb、长于60kb、长于70kb、长于80kb、长于90kb或甚至长于100kb,这允许上文所描述的更长范围的分子环境。

[0083] 还可以使得给定分区包括起始样品核酸的两个重叠片段的概率极低的水平对样品核酸进行分配。这可通过在分配过程期间以低输入量和/或浓度提供样品核酸来完成。因此,在一些情况下,给定分区可包括起始样品核酸的许多长但不重叠的片段。然后将不同分区中的样品核酸与独特的标识相关联,其中对于任何给定分区,其中所含的核酸具有相同的独特标识,但其中不同分区可包括不同的独特标识。另外,因为分配将样品组分分派至极小体积的分区或微滴中,应了解为实现如上文所阐述的分派,不需要如在更高体积方法中(例如在管或多孔板的孔中)可能会要求的对样品进行实质性稀释。此外,因为本文所描述的系统采用如此高水平的条形码多样性,可如上文所提供将多样的条形码分派在更高数目的基因组当量之中。特定而言,先前所描述的多孔板方法(参见例如美国专利公布号2013/0079231和2013/0157870,这些美国专利公布的全部公开内容以全文引用的方式并入本文中)仅可使用一百至数百个不同的条形码序列来操作,并且采用其样品的有限稀释过程以便能够将条形码归属于不同的细胞/核酸。因此,它们通常使用远少于100个细胞来操作,这将会提供大约1:10并且当然远高于1:100的基因组:(条形码类型)比率。另一方面,本文所描述的系统由于高水平的条形码多样性(例如超过10,000、100,000、500,000(等)种多样的条形码类型)而可在大约1:50或更小、1:100或更小、1:1000或更小或甚至更小比率的基因组:(条形码型)比率下操作,同时还允许在仍提供提高很多的每一基因组条形码多样性的同时加载更高数目的基因组(例如大约每次分析大于100个基因组、每次分析大于500个基因组、每次分析1000个基因组或甚至更多)。

[0084] 经常,在分配之前将样品与可释放地连接至珠粒的寡核苷酸标签的集合组合。寡核苷酸可包含至少第一和第二区域。第一区域可为条形码区域,其在给定分区内的寡核苷酸之间可大体上为相同的条形码序列,但在不同分区之间可能并且在大多数情况下为不同的条形码序列。第二区域可为N-mer(例如随机N-mer或被设计成靶向特定序列的N-mer),其可用于引导分区内的样品内的核酸。在一些情况下,在N-mer被设计成靶向特定序列的情况下,其可被设计成靶向特定染色体(例如染色体1、13、18或21)或染色体区域,例如外显子组或其他靶向区域。在一些情况下,N-mer可被设计成靶向特定基因或基因区域,诸如与疾病

或病症(例如癌症)相关的基因或区域。在分区内,可使用第二N-mer进行扩增反应以在沿核酸长度的不同位置引导核酸样品。由于扩增,各分区可含有核酸的扩增产物,所述核酸的扩增产物连接至同一或近乎同一的条形码,并且在各分区中可呈现核酸的重叠的更小片段。条形码可充当标记物,所述标记物表示核酸的集合起源于同一分区,并且因此可能也起源于核酸的同一链。在扩增之后,可汇集核酸,使用测序算法进行测序和比对。因为更短序列读段可凭借其相关条形码序列进行比对并且被归属于样品核酸的单一长片段,所以可将所述序列上所有所识别的变体归属于单一起源片段和单一起源染色体。此外,通过比对多个长片段上的多个共定位变体,可进一步表征所述染色体贡献。因此,然后可得出关于特定基因变体定相的结论。此类信息可适用于识别单倍型,所述单倍型通常为位于同一核酸链上或不同核酸链上的指定基因变体集合。拷贝数变异也可以此方式进行识别。

[0085] 所描述的方法和系统提供优于当前核酸测序技术和其相关样品制备方法的显著优点。因为对生物样品(例如血液、细胞或组织样品)进行整体处理来从细胞整体提取基因材料,并且将其转化成被配置成特定用于给定测序技术的测序文库,所以通过对基因组DNA进行测序通常不可获得单倍型定相和拷贝数变异数据。由于此整体样品处理方法,测序数据通常提供非定相基因型,以此方式不可能确定基因信息是存在于同一染色体还是不同染色体上。

[0086] 除了不能将基因特征归属于特定染色体,此类整体样品制备和测序方法还倾向于主要识别和表征样品中的多数组分,并且未被设计成识别和表征少数组分,例如由一个染色体或由一个或数个细胞所贡献的基因材料或构成所提取样品中的全部DNA的较小百分比的在血流中循环的片段化肿瘤细胞DNA分子。所描述的方法和系统还提供检测存在于更大样品中的较小群体的显著优点。因此,它们可适用于评估样品中的拷贝数变异,因为经常仅一小部分的临床样品含有具有拷贝数变异的组织。举例来说,如果样品为来自孕妇的血液样品,那么仅一小部分的样品会含有循环无细胞胎儿DNA。

[0087] 本文所公开的条形码化技术的使用赋予所述技术提供给定基因标记物集合的单个分子环境的能力,即将给定基因标记物集合(与单一标记物不同)归属于单个样品核酸分子,并且通过变体协调性组装来在多个样品核酸分子之中提供更宽或甚至更长范围的推测单个分子环境;和/或归属于特定染色体的能力。这些基因标记物可包括特定基因座,例如变体,诸如SNP;或者它们可包括短序列。此外,条形码化的使用赋予以下额外优点:促成区分从样品提取例如用于检测和表征血流中的循环肿瘤DNA的全部核酸群体中的少数组分与多数组分以及减少或消除任何扩增期间的扩增偏向的能力。另外,以微流体模式实现赋予以极小样品体积和低输入量的DNA工作的能力,以及快速处理大量样品分区(例如微滴)以促进基因组范围标记的能力。

[0088] 如先前所描述,本文所描述的方法和系统的优点在于它们可通过使用广泛可用的短读段测序技术来实现结果。此类技术具有以下优点:可轻易获得并且在研究界内分布范围很广,并且具有充分表征和高度有效的方案和试剂系统。这些短读段测序技术包括可从例如Illumina, Inc. (例如GXII、NextSeq、MiSeq、HiSeq、X10)、Thermo-Fisher的Ion Torrent分公司(例如Ion Proton和Ion PGM)获得的那些、焦磷酸测序法以及其他技术。

[0089] 特别有利的是,本文所描述的方法和系统利用这些短读段测序技术并且这样做具有相关低误差率。特定而言,如上文所描述,本文所描述的方法和系统实现单个分子读段长

度或环境,但具有单个测序读段,从而排除mate pair延伸,其短于1000bp、短于500bp、短于300bp、短于200bp、短于150bp或甚至更短;并且对于此类单个分子读段长度来说,具有低于5%、低于1%、低于0.5%、低于0.1%、低于0.05%、低于0.01%、低于0.005%或甚至低于0.001%的测序误差率。

[0090] II. 工作流程概括

[0091] 在一个示例性方面,本公开中所描述的方法和系统使得单个样品(例如核酸)沉积或分配至离散分区中,其中各分区维持其自身的内容物与其他分区的内容物分离。如本文中所示,分区是指器皿或容器,其可包括多种不同形式,例如孔、管、微孔或纳米孔、通孔等。然而,在一些方面,分区可在流体流内流动。这些容器可包含例如具有包围内部流体中心或核心的外部屏障的微胶囊或微囊泡,或其可为能够夹带和/或保留基质内的材料的多孔基质。然而,在一些方面,这些分区可包含非水性连续相(例如油相)内的水性流体的微滴。多种不同的容器描述于例如2013年8月13日提交的美国专利申请号13/966,150中。同样地,用于形成非水性或油性连续相中的稳定微滴的乳液系统详细描述于例如美国专利公布号2010/0105112中,所述专利公布的全部公开内容以全文引用的方式并入本文中。在某些情况下,微流体通道网络可适合用于产生如本文所描述的分区。此类微流体装置的实例包括详细描述于2014年4月10日提交的美国临时专利申请号61/977,804中的那些,所述临时专利申请的全部公开内容出于所有目的以全文引用的方式并入本文中。在分配单个细胞时还可采用替代机制,包括多孔膜,细胞的水性混合物穿过所述多孔膜被挤压至非水性流体中。此类系统通常可自例如Nanomi, Inc.获得。

[0092] 在乳液中的液滴的情况下,将样品材料(例如核酸)分配至离散分区中通常可通过以下方式完成:使水性的含有样品的流流至接头中,也使分配流体(例如氟化油)的非水性流流至所述接头中,使得在流动的流分配流体内形成水性微滴,其中此类微滴包括样品材料。如下所述,分区(例如微滴)还可包括共分配的条形码寡核苷酸。可通过控制系统的各种不同参数来调节任何特定分区内的样品材料的相对量,所述各种不同参数包括例如水性流中的样品浓度、水性流和/或非水性流的流速等。本文所描述的分区的特征常常为具有极小的体积。举例来说,在基于微滴的分区的情况下,微滴可具有小于1000皮升(pL)、小于900pL、小于800pL、小于700pL、小于600pL、小于500pL、小于400pL、小于300pL、小于200pL、小于100pL、小于50pL、小于20pL、小于10pL或甚至小于1pL的总体积。在与珠粒共分配的情况下,应了解,分区内的样品流体体积可为上文所描述的体积的小于90%、小于80%、小于70%、小于60%、小于50%、小于40%、小于30%、小于20%或甚至为上文所描述的体积的小于10%。在一些情况下,在与极小量的起始试剂(例如输入核酸)进行反应时,使用低反应体积分区可为有利的。用于在低输入核酸情况下分析样品的方法和系统提供于2014年6月26日提交的美国临时专利申请号62/017,580中,该临时专利申请的全部公开内容以全文引用的方式并入本文中。

[0093] 一旦将样品引入其各自的分区中,根据本文所描述的方法和系统,分区内的样品核酸通常具备独特的标识,使得在表征那些核酸之后,可将其归因为源于其各自的起源。因此,可将样品核酸与独特的标识(例如条形码序列)共分配。在一些方面,以包含可连接至那些样品的核酸条形码序列的寡核苷酸的形式来提供独特的标识。对寡核苷酸进行分配,使得在给定分区中的寡核苷酸之间,其中所含的核酸条形码序列相同,但在不同分区之间,寡

核苷酸可具有不同的条形码序列。在一些方面,仅一个核酸条形码序列可与给定分区相关联,不过在一些情况下,可存在两个或更多个不同的条形码序列。

[0094] 核酸条形码序列可在寡核苷酸的序列内包括6至约20个或更多个核苷酸。这些核苷酸可为完全连续的,即呈单段相邻核苷酸的形式,或者它们可被分隔至由一个或多个核苷酸隔开的两个或更多个单独的子序列中。在一些情况下,隔开的子序列的长度可为约4至约16个核苷酸。

[0095] 共分配的寡核苷酸还可包含适用于处理共分配的核酸的其他功能序列。这些序列包括例如靶向型或随机/通用型扩增引物序列,其用于扩增分区内的单个细胞的基因组DNA,同时连接相关条形码序列、测序引物、杂交或探测序列,例如用于识别序列的存在或用于向下拉动条形码化核酸;或许多其他潜在功能序列中的任一种。再次,寡核苷酸和相关条形码以及其他功能序列连同样品材料的共分配描述于例如2014年2月7日提交的美国临时专利申请号61/940,318和2014年5月9日提交的美国临时专利申请号61/991,018以及2014年6月26日提交的美国专利申请号14/316,383以及2014年2月7日提交的美国专利申请号14/175,935中,这些专利申请的全部公开内容以全文引用的方式并入本文中。

[0096] 简要地说,在一种示例性方法中,提供珠粒,所述珠粒各自可包括大量上文所描述的可释放地连接至珠粒的寡核苷酸,其中连接至特定珠粒的所有寡核苷酸可包括相同的核酸条形码序列,但其中在所用的珠粒群体中可呈现大量多样的条形码序列。在一些情况下,珠粒群体可提供多样的条形码序列文库,其可包括至少1000个不同的条形码序列、至少10,000个不同的条形码序列、至少100,000个不同的条形码序列或在一些情况下,至少1,000,000个不同的条形码序列。另外,各珠粒可具备所连接的大量寡核苷酸分子。特定而言,单个珠粒上包括条形码序列的寡核苷酸分子的数目可为至少约10,000个寡核苷酸、至少100,000个寡核苷酸分子、至少1,000,000个寡核苷酸分子、至少100,000,000个寡核苷酸分子,且在一些情况下为至少十亿个寡核苷酸分子。

[0097] 在对珠粒施加特定刺激后,寡核苷酸可从珠粒释放。在一些情况下,刺激可为光刺激,例如通过可释放寡核苷酸的光不稳定性键的裂解。在一些情况下,可使用热刺激,其中珠粒环境的温度升高可能会导致键的裂解或寡核苷酸从珠粒的其他释放。在一些情况下,可使用化学刺激,从而裂解寡核苷酸与珠粒的键,或以其他方式可使得寡核苷酸从珠粒释放。

[0098] 根据本文所描述的方法和系统,可将包括连接的寡核苷酸的珠粒与单个样品共分配,使得单个分区内含有单一珠粒和单一样品。在一些情况下,在需要单珠粒分区的情况下,可控制流体的相对流速,使得所述分区平均每个分区含有少于一个珠粒,以确保被占用的那些分区主要是单一占用的。同样地,可能希望控制流速以使得更高百分比的分区被占用,从而例如仅允许较小百分比的未占用分区。在一些方面,控制流量和通道结构,以确保所需数目的单一占用分区、低于某一水平的未占用分区以及低于某一水平的多重占用分区。

[0099] 图3说明用于条形码化样品核酸并且随后测序诸如用于拷贝数变异或单倍型分析的示例性方法。首先,可从来源获得包含核酸的样品300,并且还获得条形码化珠粒的集合310。可使珠粒键联至含有一个或多个条形码序列以及引物(诸如随机N-mer或其他引物)的寡核苷酸。在一些情况下,条形码序列可从条形码化珠粒释放,例如通过在条形码与珠粒之

间的键的裂解或通过下面珠粒的降解来释放条形码,或两种途径的组合。举例来说,在一些方面,条形码化珠粒可由诸如还原剂等试剂降解或溶解以释放条形码序列。在此实例中,将较低量的包含核酸的样品305、条形码化珠粒315以及(在一些情况下)其他试剂(例如还原剂)320组合并且进行分配。举例来说,此类分配可涉及将组分引入微滴产生系统,诸如微流体装置325。在微流体装置325的辅助下,可形成油包水乳液330,其中所述乳液含有水性微滴,所述水性微滴含有样品核酸305、还原剂320以及条形码化珠粒315。还原剂可溶解或降解条形码化珠粒,由此使具有条形码和随机N-mer的寡核苷酸从微滴内的珠粒释放335。随机N-mer可然后引导样品核酸的不同区域,从而在扩增之后产生样品的扩增拷贝,其中将各拷贝用条形码序列标记340。在一些情况下,各微滴含有寡核苷酸的集合,所述寡核苷酸的集合含有相同的条形码序列和不同的随机N-mer序列。随后,将乳液破坏345,并且可经由例如扩增方法350(例如PCR)来添加额外序列(例如辅助特定测序方法的序列、额外条形码等)。然后可进行测序355,并且应用算法来解释测序数据360。测序算法通常能够例如对条形码进行分析以比对测序读段和/或识别特定序列读段所属的样品。

[0100] 如上文所提到,虽然单一珠粒占用可能为所需的,但是应了解可能经常存在多重占用分区或未占用分区。图4中示意性说明了用于对样品和包含条形码寡核苷酸的珠粒进行共分配的微流体通道结构的实例。如图所示,以在通道接头412处流体连通的形式提供通道区段402、404、406、408以及410。使包含单个样品414的水性流通过通道区段402流至通道接头412。如本文别处所描述,可在分配过程之前将这些样品悬浮于水性流体内。

[0101] 同时,使包含携带条形码的珠粒的水性流416通过通道区段404流至通道接头412。将非水性分配流体从侧通道406和408中的每一者引入通道接头412中,并且使组合流流至出口通道410中。在通道接头412内,将来自通道区段402和404的两条组合水性流组合在一起,并且分配至微滴418中,所述微滴包括共分配的样品414和珠粒416。如前面提到的,通过控制在通道接头412处组合的各个流体的流动特征,以及控制通道接头的几何结构,可优化组合和分配以实现珠粒、样品或两者在所产生的分区418内的所需占用水平。

[0102] 如应了解,可将许多其他试剂与样品和珠粒一起共分配,包括例如化学刺激物;核酸延伸、转录和/或扩增试剂,诸如聚合酶、反转录酶、三磷酸核苷或NTP类似物、引物序列和额外辅因子(诸如用于此类反应中的二价金属离子)、连接反应试剂(诸如连接酶和连接序列);染料、标签或其他标记试剂。

[0103] 一旦共分配,即可使用设置于珠粒上的寡核苷酸来对所分配的样品进行条形码化和扩增。在对样品进行扩增和条形码化时使用这些条形码寡核苷酸的示例性方法详细描述于2014年2月7日提交的美国专利申请号61/940,318和2014年5月9日提交的美国专利申请号61/991,018以及2014年6月26日提交的美国专利申请号14/316,383中,这些专利申请的全部公开内容以全文引用的方式并入本文中。简要地说,在一个方面,寡核苷酸存在于与样品共分配的珠粒上并且与样品一起从其珠粒释放至分区中。寡核苷酸可(连同条形码序列一起)在其5'端包括引物序列。此引物序列可为意在随机引导样品的许多不同区域的随机寡核苷酸序列或其可为以引导样品的特定靶向区域的上游为目标的特定引物序列。

[0104] 一旦被释放,寡核苷酸的引物部分即可与样品的互补区域退火。也与样品和珠粒共分配的延伸反应试剂(例如DNA聚合酶、三磷酸核苷、辅因子(例如 Mg^{2+} 或 Mn^{2+} 等))然后使用样品作为模板来延伸引物序列,以产生与引物退火的模板链的互补片段,其中互补片段

包括寡核苷酸和其相关条形码序列。多个引物与样品的不同部分的退火和延伸可产生样品的重叠互补片段的大型汇集物,所述重叠互补片段各自具有其自己的指示其在其中形成的分区的条形码序列。在一些情况下,这些互补片段本身可用作模板,所述模板由存在于分区中的寡核苷酸引导以产生互补序列的互补序列,其又包括条形码序列。在一些情况下,此复制过程被配置为使得当第一互补序列重复时,其产生位于或靠近其末端的两个互补序列,以允许形成发夹结构或部分发夹结构,从而降低所述分子成为产生其他重复拷贝的基础的能力。图5中示出了对此情况的一个实例的示意性说明。

[0105] 如该图所示,将包括条形码序列的寡核苷酸与样品核酸504一起共分配于例如乳液中的微滴502中。如本文别处所提到,如图A中所示,寡核苷酸508可提供于与样品核酸504共分配的珠粒506上,所述寡核苷酸可从珠粒506释放。寡核苷酸508除一个或多个功能序列(例如序列510、514以及516)之外还包括条形码序列512。举例来说,寡核苷酸508被示出为包含条形码序列512以及可充当给定测序系统的连接或固定序列的序列510,例如用于在Illumina HiSeq或Miseq系统的流动细胞中进行连接的P5序列。如图所示,寡核苷酸还包括引物序列516,其可包括用于引导样品核酸504的诸多个部分的复制的随机或靶向型N-mer。寡核苷酸508内还包括序列514,其可提供测序引导区,诸如“读段1”或R1引导区,所述引导区用于通过测序系统中的合成反应来引导聚合酶介导的模板定向测序。在一些情况下,条形码序列512、固定序列510以及R1序列514对于连接至给定珠粒的所有寡核苷酸来说可为共同的。引物序列516可能因随机N-mer引物而不同,或者在某些靶向应用中对于给定珠粒上的寡核苷酸来说可为共同的。

[0106] 基于存在引物序列516,寡核苷酸能够如图B中所示引导样品核酸,这允许使用也与珠粒506和样品核酸504共分配的聚合酶和其他延伸试剂来延伸寡核苷酸508和508a。如图C中所示,在对于随机N-mer引物来说将与样品核酸504的多个不同区域退火的寡核苷酸延伸之后;形成核酸的多个重叠互补序列或片段,例如片段518和520。虽然包括与样品核酸的诸多个部分互补的序列部分,例如序列522和524,但是这些构建体在本文中通常被称为包含样品核酸504中具有连接的条形码序列的片段。如应了解,如上文所描述的模板序列的复制部分在本文中经常被称为所述模板序列的“片段”。然而,尽管如此,术语“片段”涵盖起源核酸序列(例如模板或样品核酸)的一部分的任何表示,包括通过提供模板序列的诸多个部分的其他机制(诸如给定序列分子的实际片段化,例如通过酶促、化学或机械片段化)所形成的那些。然而,在一些方面,模板或样品核酸序列的片段可表示基础序列的复制部分或其互补序列。

[0107] 然后可例如通过序列分析对条形码化核酸片段进行表征,或可如在如图D中所示的过程中将其进一步扩增。举例来说,也从珠粒306释放的额外寡核苷酸(例如寡核苷酸508b)可引导片段518和520。特定而言,再次,基于随机N-mer引物516b存在于寡核苷酸508b中(这在一些情况下可不同于给定分区中的其他随机N-mer,例如引物序列516),寡核苷酸与片段518退火,并且延伸以形成片段518中包括序列528的至少一部分的互补序列526,其包含样品核酸序列的一部分的重复。寡核苷酸508b继续延伸直到它已通过片段518的寡核苷酸部分508复制。如本文别处所提到,并且如图D中所说明,寡核苷酸可被配置成提示通过聚合酶进行的复制在所需点停止,例如在通过寡核苷酸508的包括在片段518内的序列516和514复制之后停止。如本文所描述,这可通过不同方法来实现,包括例如并入不能由所用

的聚合酶处理的不同核苷酸和/或核苷酸类似物。举例来说,这可包括在序列区域512内纳入含尿嘧啶的核苷酸来防止非尿嘧啶耐受型聚合酶使所述区域的复制停止。结果,形成片段526,其在一个末端包括全长寡核苷酸508b,包括条形码序列512、连接序列510、R1引物区514以及随机N-mer序列516b。在序列的另一个末端可包括第一寡核苷酸508的随机N-mer的互补序列516',以及整个或一部分的R1序列的互补序列以(以序列514'示出)。R1序列514和其互补序列514'然后能够杂交在一起以形成部分发夹结构528。如应了解,因为不同寡核苷酸之间的随机N-mer不同,这些序列和其互补序列预期不会参与发夹形成,例如序列516'(其为随机N-mer 516的互补序列)预期不会与随机N-mer序列516b互补。对于其他应用来说不会是这种情况,例如靶向型引物,其中在给定分区内寡核苷酸之间的N-mer将为共同的。

[0108] 通过形成这些部分发夹结构,允许从进一步的复制中去除样品序列的第一级重复,从而例如防止拷贝的重复拷贝。部分发夹结构还提供适用于所形成的片段(例如片段526)的后续处理的结构。

[0109] 可然后如本文所描述从多个不同分区汇集所有片段以便在高通量测序仪上进行测序。因为各片段是关于其起源分区而被编码,所以基于条形码的存在可将所述片段的序列归属回到其起源。图6中对这进行了示意性说明。如一个实例中所示,将起源于第一来源600(例如单个染色体、核酸链等)的核酸604和源于不同染色体602或核酸链的核酸606各自如上文所描述与其自己的条形码寡核苷酸集合一起分配。

[0110] 在各分区内,各核酸604和606然后被处理,以单独地提供第一片段的重叠的第二片段集合,例如第二片段集合608和610。此处理还提供第二片段,其中来源于特定第一片段的第二片段中的每一者的条形码序列是相同的。如图所示,第二片段集合608的条形码序列由“1”表示,而片段集合610的条形码序列由“2”表示。可使用多样的条形码文库来区别地条形码化大量不同片段集合。然而,没有必要用不同的条形码序列来条形码化来自不同第一片段的每一个第二片段集合。在一些情况下,可同时处理多个不同的第一片段以包括相同的条形码序列。本文在别处详细描述了多样的条形码文库。

[0111] 然后可汇集例如来自片段集合608和610的条形码化片段,以便使用例如通过可从Illumina或Thermo Fisher, Inc.的Ion Torrent分公司获得的合成技术获得的序列进行测序。一旦经过测序,即可将序列读段612至少部分基于所包括的条形码并且在一些情况下部分基于其片段的序列而归属于其各自的片段集合,例如如聚集读段614和616中所示。然后组装被归属于各片段集合的序列读段以提供各样品片段的组装序列,例如序列618和620,所述组装序列又可被进一步归属回到其各自的原始染色体(600和602)。用于组装基因组序列的方法和系统描述于例如2014年6月26日提交的美国临时专利申请号62/017,589中,该临时专利申请的全部公开内容以全文引用的方式并入本文中。在一些实例中,通过从头组装和/或基于参考序列的组装(例如映射至参考序列)来组装基因组序列。

[0112] III. 将方法和系统应用于定相和拷贝数分析

[0113] 在本文所描述的系统和方法的一个方面,将序列读段归属于更长起源分子的能力用于测定关于序列的相位信息。在一个实例中,比较与显示两个或更多个特定基因变体序列(例如等位基因、基因标记物)的序列相关的条形码以确定基因标记物的集合是否位于样品中的同一染色体或不同染色体上。可使用此类定相信息来确定样品中某些靶染色体或基因的相对拷贝数。所描述的方法和系统(symptom)的优点在于可使用多个位置、基因座、变

体等来识别它们所起源于的单个染色体或核酸链以测定定相和拷贝数信息。经常,使用沿染色体的多个位置(例如大于2、3、4、5、6、7、8、9、10、20、30、40、50、100、500、1000、5000、10000、50000、100000或500000个)来测定本文所描述的定相、单倍型以及拷贝数变异信息。

[0114] 举例来说,如上文所提到,尽管利用可提供相对更短的序列读段的测序技术,本文所描述的方法和系统凭借上文所描述的分配和归属方面可适用于从单个核酸片段(例如单个核酸分子)提供有效长序列读段。因为这些长序列读段可被归属于单一起始片段或分子,所以序列中的变体位置可同样被归属于单一分子,并且通过外推法归属于单一染色体。另外,可采用任何给定片段上的多个位置作为邻近片段的比对特征来提供可被推测为起源于同一染色体的比对序列。举例来说,可对第一片段进行测序,并且凭借上文所描述的归属方法和系统,可将存在于所述序列上的变体全部归属于单一染色体。然后可将共有被确定为仅存在于一个染色体上的多个这些变体的第二片段假定为源于同一染色体,并且因此与第一片段比对,以形成两个片段的定相比对。重复此过程允许识别长范围相位信息。可从已知参考物(例如HapMap)或从例如显示以其他方式同一的序列区段上的不同变体的测序数据的集合获得单一染色体上的变体的识别。

[0115] 图7提供示例性定相测序方法的示意性说明。如图所示,可将起源核酸702(诸如染色体、染色体片段、外显子组或其他大的单核酸分子)片段化成多个大片段704、706、708。起源核酸702可包括特定核酸分子(例如染色体)所特有的许多序列变体(A、B、C、D、E、F以及G)。根据本文所描述的方法,可将起源核酸片段化成多个大的重叠片段704、706以及708,其包括相关序列变体的子集。然后可如本文所描述将各片段分配,进一步片段化成子片段,并且条形码化,以提供更大片段的多个重叠的条形码化子片段,其中给定更大片段的子片段带有相同的条形码序列。举例来说,与条形码序列“1”和条形码序列“2”相关的子片段分别显示于分区710和712中。然后可汇集条形码化的子片段,测序,并且组装测序过的子片段以提供长片段序列714、716以及717。长片段序列714、716以及717中的一者或多者可包括多个变体。然后可基于序列714、716以及717的重叠定相变体信息进一步组装长片段序列以提供定相序列718,从所述定相序列可确定定相位置。

[0116] 一旦确定了定相位置,即可进一步以多种方式来探究信息。举例来说,可利用评估某些病症的基因风险时对定相变体的了解,识别父体和母体特征,识别非整倍性,或识别单倍型分析信息。

[0117] 在本文所公开的系统和方法的一些方面,使用同时检测两种或更多种定相基因标记物来进行拷贝数变异分析以提高拷贝数计数的准确度。与在原初方法下仅基于多个基因座上单倍型之间的计数读段的变异相比,利用定相信息可增加信号的相对强度。另外,利用定相信息允许位置特异性偏向的正规化,从而大体上进一步增强信号。拷贝数变异(CNV)准确度可取决于众多因素,包括测序长度、CNV长度、拷贝数目等)。本文所提供的方法和系统可以至少70%、80%、85%、90%、91%、92%、93%、94%、95%、99%、99.1%、99.2%、99.3%、99.4%、99.5%、99.6%、99.7%、99.8%、99.9%、99.95%、99.99%、99.995%或99.999%的准确度测定CNV。在一些情况下,本文所提供的方法和系统以低于10%、9%、8%、7%、6%、5%、4%、3%、2%、1%、0.1%、0.05%、0.01%、0.005%、0.001%、0.0005%、0.0001%、0.00005%、0.00001%或0.000005%的误差率测定CNV。类似地,本文所提供的方法和系统可以至少70%、80%、85%、90%、91%、92%、93%、94%、95%、99%、99.1%、

99.2%、99.3%、99.4%、99.5%、99.6%、99.7%、99.8%、99.9%、99.95%、99.99%、99.995%或99.999%的准确度检测两个或更多个基因变体的定相/单倍型信息。在一些情况下,本文所提供的方法和系统以低于10%、9%、8%、7%、6%、5%、4%、3%、2%、1%、0.1%、0.05%、0.01%、0.005%、0.001%、0.0005%、0.0001%、0.00005%、0.00001%或0.000005%的误差率测定定相或单倍型信息。本公开还提供去除基因座特异性偏向的方法,其中基因座特异性变异减少了至少2倍、3倍、4倍、5倍、10倍、20倍、30倍、40倍、50倍、60倍、70倍、80倍、90倍、100倍、200倍、500倍、1000倍、5000倍或10000倍。可使用本文所提供的方法和系统来检测拷贝数变异,诸如在拷贝数变化反映染色体数目或染色体的诸多个部分的变化变化的情况下。在一些情况下,可使用本文所提供的方法和系统来检测存在于同一染色体上的基因的拷贝数变异。

[0118] 图8(上部图)为说明健康患者的基因组的子集的示意图。此患者具有在所示基因座处的杂合基因型和位于单独染色体链上的两个单独的单倍型(1和2) 805、810。患者的天然存在的变异(诸如SNP或缺失)以圆圈描绘。图8还描绘癌症患者的基因组815。某些癌症与单倍型拷贝数增加相关。中部图描绘单倍型2增加810。如图8中的底部图中所示显示单倍型2减少820的底部图中所示,癌症也可能与单倍型数目减少相关。常用测序技术不能准确确定此单倍型拷贝减少或增加。如图9a中所示,这是部分归因于以下事实:患者血液中的肿瘤贡献的DNA 910仅为全部DNA中的一小部分,全部DNA中多数是由正常组织贡献的DNA905。肿瘤DNA的此低浓度使得在一般测序技术下对拷贝数的检测不精确,参见图9b。难以检测无拷贝变异920在平均深度D 935处的预期计数的峰与拷贝减少925(940)和拷贝增加930(945)的峰差异。对于任何给定的单个标记物,在重复测试中拷贝数分析的结果的分布可以近似于泊松分布的方式公布于正确答案周围,其中分布的宽度取决于分析中的随机误差的各种来源。因为对于给定样品来说,拷贝数的变化可为样品中相对小的部分,所以监测单一基因标记物时的宽的概率分布会掩盖正确的结果。此困难是归因于以下事实:如图10(左侧图)中所示,一般测序技术一次仅关注单倍型的一个单一变体位置。使用此类技术,在代表拷贝减少1025、正常拷贝1020以及拷贝增加1030的峰之间可存在显著重叠。本文所公开的技术允许检测全部(或部分)单倍型、增加解析率以及改善对拷贝增加和减少的检测,图10(右侧图)。图11中示意性示出此改善,其中一般检测1100产生展开的重叠峰,而本文中的技术1110允许更精细的峰和对拷贝增加或减少的改善的解析。使用同时监测两种或更多种定相基因标记物,特别是已知共定位于单一染色体上并且可能因此最有可能总是以同步化非随机方式以更大或更小数目出现的标记物,具有使预期结果分布的宽度变窄并且同时提高计数准确度的作用。

[0119] 除了在检测和诊断癌症方面的优点,本文所提供的方法和系统还提供更准确和灵敏的检测胎儿非整倍性的方法。

[0120] 胎儿非整倍性为胎儿染色体数目失常。非整倍性通常引起显著的物理和神经损伤。举例来说,X染色体数目减少引起特纳氏综合征(Turner's syndrome)。21号染色体的拷贝数增加引起唐氏综合征(Down Syndrome)。侵入性测试(诸如羊膜穿刺术或绒毛膜绒毛取样(CVS))可导致流产风险,并且此处使用较无创的测试母体血液的方法。

[0121] 本文所描述的方法可适用于无创检测胎儿非整倍性。图12中示出了示例性方法。对处于携带具有非整倍体基因组的胎儿的风险中的孕妇进行测试1200。收集含有胎儿基因

材料的母体血液样品1205。然后从血液样品提取基因材料(例如无细胞核酸)1210。还可获得条形码化珠粒的集合1215。可使珠粒键联至含有一个或多个条形码序列以及引物(诸如随机N-mer或其他引物)的寡核苷酸。在一些情况下,条形码序列可从条形码化珠粒释放,例如通过在条形码与珠粒之间的键的裂解或通过下面珠粒的降解来释放条形码,或两种途径的组合。举例来说,在一些方面,条形码化珠粒可由诸如还原剂等试剂降解或溶解以释放条形码序列。在此实例中,将样品1210、条形码化珠粒1220以及(在一些情况下)其他试剂(例如还原剂)组合在一起,并且进行分配。举例来说,此类分配可涉及将组分引导至微滴产生系统,诸如微流体装置1225。在微流体装置1225的辅助下,可形成油包水乳液1230,其中所述乳液含有水性微滴,所述水性微滴含有样品核酸1210、条形码化珠粒1215以及(在一些情况下)还原剂。还原剂可溶解或降解条形码化珠粒,由此使具有条形码和随机N-mer的寡核苷酸从微滴内的珠粒释放1235。随机N-mer可然后引导样品核酸的不同区域,从而在扩增之后产生样品的扩增拷贝,其中将各拷贝用条形码序列标记1240。在一些情况下,各微滴含有寡核苷酸的集合,所述寡核苷酸的集合含有相同的条形码序列和不同的随机N-mer序列。在其他实施方案中,单个微滴包含独特的条形码序列;或者在一些情况下,整个微滴群体中的某一比例具有独特的序列。随后,将乳液破坏1245,并且可经由例如扩增方法(例如PCR)来添加额外序列(例如辅助特定测序方法的序列、额外条形码等)。然后可经由任何适合类型的测序平台(例如Illumina、Ion Torrent、Pacific Biosciences SMRT、Roche 454测序、SOLiD测序等)进行测序1250,并且应用算法来解释测序数据1255。测序算法通常能够例如对条形码进行分析以比对测序读段和/或识别特定序列读段所属的样品。可基于所连接的独特条形码条形码将比对序列进一步归属于其各自的基因起源(例如,染色体)。然后将染色体拷贝的数目与正常二倍体染色体相比较1260。患者被告知不同染色体的任何拷贝数失常和相关风险/疾病1265。

[0122] 例如确定基因变体为相连的或位于不同染色体上的定相可为多种应用提供适用信息。举例来说,定相适用于确定基因组中是否存在与疾病相关的某些易位。此类易位的检测还可允许鉴别诊断和改良的治疗。确定基因组中哪些等位基因是相连的可适用于考虑基因是如何遗传的。

[0123] 知道染色体对中的各个单个染色体的等位基因的模式、单倍型经常可能是有用的。举例来说,存在于一个染色体上的失活突变的两个拷贝可具有有限的效果,但如果分布在两个染色体之间就可具有显著效果,例如在任一染色体提供活性基因产物的情况下。例如在疾病风险增加或对某些药物的反应缺少的情况下可表现这些效果。

[0124] IV. 应用方法和系统来识别/表征结构变异

[0125] 在其他应用中,本文所描述的方法和系统高度适用于获得长范围分子序列信息以识别和表征广泛范围的不同基因结构变异。如上文所提到,这些变异包括多种不同的变体事件,包括插入、缺失、重复、反转录转座子、易位、倒位、短和长串联重复等。这些结构变异受到大量科学关注,因为它们被认为与一系列多样的基因疾病相关。

[0126] 尽管这些变异受到关注,几乎没有识别和表征这些结构变异的有效且高效的方法。这部分是因为这些变异的特征不是存在异常序列区段,而是实际上涉及将被视为是正常序列区段的异常序列环境,或简单地丢失序列信息。由于其相对短的读段长度,大多数测序技术不能提供其所产生的序列读段的显著环境,以及尤其例如其读段长度以外的长范围

序列环境,并且因此损失在组装过程中对这些变异的识别。识别这些变异时的困难因这些技术的整体方法而进一步复杂化,其中将许多分子(例如多个染色体)组合以产生共有序列,所述共有序列可包括包括与不包括变异的基因组材料。

[0127] 然而,在当前所描述的方法和系统的情形下,可利用短读段测序技术来得到长范围序列信息,所述长范围序列信息可归属于单个起源核酸分子,并且因此保留整个或部分地含于那些单个分子中的变体区域的长范围序列环境。

[0128] 如上文所描述,本文所描述的方法和系统能够提供长范围序列信息,所述长范围序列信息可归属于单个起源核酸分子,并且此外,在处理此长范围序列信息时,通过这些更长序列信息的比较和重叠推测甚至更长范围的序列环境。此类长范围序列信息和/或推测的序列环境允许识别和表征使用可获得的技术不能轻易识别的许多结构变异。

[0129] 虽然以简化的方式在图2中进行了说明,但图13A和13B提供更详细的使用本文所描述的方法和系统识别某些类型的结构变异的示例性方法。如图所示,生物体的基因组或生物体的组织一般会包括图13A中所说明的第一基因型,其中将包括第一基因1304的第一基因区域1302与包括第二基因1208的第二基因区域1306分离。此分离可反映基因之间的一系列距离,包括例如同一外显子中的不同区域、同一染色体上的不同外显子、不同染色体等。然而,如图13B中所示,示出了一个基因型,所述基因型反映易位事件已发生,其中基因1308被插入基因区域1304中,使得其在基因1304与1308之间形成如变体序列1314中的基因融合1312的基因融合。

[0130] 当前用于检测大基因组结构变体(诸如大的倒位或易位)的方法依赖于读段对,所述读段对跨越变体的断点(例如其中易位部分融合在一起的基因组基因座)。为确保在测序实验期间观察到此类读段对,可能需要非常深入的测序。在靶向测序(诸如外显子组测序)中,使用当前测序技术检测结构变体几乎是不可能的,除非断点在靶向的区域内(例如在外显子中),这是非常不可能的。

[0131] 然而,本文所描述的条形码方法和系统所提供的信息可极大地提高检测结构变体的能力。直观上,在断点左侧和右侧的基因座,可倾向于位于基因组DNA的共同片段上,并且因此保持在单一分区内,并且因此用共同或共有条形码序列条形码化。由于剪切的随机性质,这种条形码共有随着序列距离断点更远而减少。使用统计方法,可确定两个基因组基因座之间的条形码重叠是否显著大于偶然将预料的情况。此类重叠暗示断点的存在。重要地,条形码信息补充由传统测序提供的信息(诸如来自跨越断点的读段的信息),如果此类信息是可获得的。

[0132] 在本文所描述的方法的情况下,如上文所描述,将来自生物体的基因组材料(包括相关基因区域)片段化,使得其包括相对长的片段。这是相对于图13A中的非易位基因型来说明。如图所示,形成分别包括基因区域1302和1306的两个长单个第一分子片段1316和1318。分别将这些片段单独分配至分区1320和1322中,并且将第一片段中的每一者分别在分区内片段化成许多第二片段1324和1326,此片段化过程将独特标识标签或条形码序列连接至第二片段,所述独特标识标签或条形码序列对给定分区内的所有第二片段来说是共同的。对于分区1320和1322中的每一者,标签或条形码分别由“1”或“2”表示。因此,完全分开的基因1304和1308可产生第二片段的区别分配和区别条形码化群组。

[0133] 一旦条形码化,然后即可汇集第二片段并且进行核酸测序过程,所述核酸测序过

程可提供第二片段的序列以及所述片段的条形码序列。基于特定条形码(例如1或2)的存在,然后如由条形码归属于各序列所示可将第二片段序列归属于某一起源序列,例如基因1304或1308。在一些情况下,条形码化的第二片段序列关于单独的起源第一片段序列的映射可足够明确确定未发生易位。然而,在一些情况下,可组装第二片段序列以提供全部或部分起源第一片段序列的例如如由组装序列1330和1332所示的组装序列。

[0134] 与图13A中所示的非易位基因型实例相比之下,图13B示出了相同方法应用于含有易位的基因型的示意性说明。如图所示,由变体序列1314产生第一长核酸片段1352,并且包括至少一部分的易位变体,例如基因融合1312。然后将第一片段1352分配至离散分区1354中。在分区1354内,将第一片段1352进一步片段成第二片段1356,所述第二片段又包括独特条形码,所述独特条形码对于分区1354内的所有第二片段1356来说是相同的(以条形码“1”示出)。如上所述,汇集第二片段并且测序提供第二片段以及其相关条形码的基础序列。然后可将这些条形码化序列归属于其各自的基因序列。然而,如图所示,两种基因均可反映包括相同条形码序列的经归属的第二片段序列,表明其起源于相同分区,并且潜在地起源于相同起源分子,从而指示基因融合。这可通过提供许多重叠的第一片段来进一步验证,所述许多重叠的第一片段也包括基因融合的至少诸多个部分,但在不同分区中用不同条形码加以处理。

[0135] 在一些情况下,存在归属于最初分离的基因中的每一者的多个不同条形码序列(和其基础片段序列)可指示存在基因融合或其他易位事件。在一些情况下,将至少2个条形码、至少3个不同条形码、至少4个不同条形码、至少5个不同条形码、至少10个不同条形码、至少20个不同条形码或更多归属于基于参考序列将被预期已分离的两个基因区域可提供易位事件已将那些区域放置于彼此近端、附近或以其他方式使彼此整合的指示。在一些情况下,被分配的片段的尺寸可指示可识别变体联系的灵敏性。特定而言,在给定微滴中的片段的长度为10kb的情况下,将预期在所述10kb尺寸范围内的所述联系将为可检测的。

[0136] 同样地,在变体与野生型结构均在相同的10kb片段内的情况下,将预期所述变体的识别会更困难,因为通过共同或共有条形码两者均将显示联系。因此,可使用片段尺寸选择来调节所检测的相连序列的相对接近度,无论是野生型还是变体。然而,一般来说,在本文中可通过识别变体基因组中那些不相连的序列区段之间的联系而轻易识别产生一般由超过100个碱基、超过500个碱基、超过1kb、10kb、超过20kb、超过30kb、超过40kb、超过50kb、超过60kb、超过70kb、超过80kb、超过90kb、超过100kb、超过200kb或甚至更大距离的隔开的近端序列的结构变体,所述联系是由共有或共同的条形码和/或如所提到的由跨越断点的序列数据指示。当那些相连序列在基因组序列内隔开小于50kb、小于40kb、小于30kb、小于20kb、小于10kb、小于5kb、小于4kb、小于3kb、小于2kb、小于1kb、小于500个碱基、小于200个碱基或甚至更小距离时,此类联系通常是可识别的。

[0137] 在一些情况下,产生位于彼此近端或相连的两个序列(其中它们通常会隔开例如超过10kb、超过20kb、超过30kb、超过40kb或超过50kb或更多)的结构变异可通过占可映射条形码化序列的总数的百分比来识别,所述可映射条形码化序列包括条形码,所述条形码对于两个序列部分来说为共同的。

[0138] 如应了解,在一些情况下,本文所描述的方法可确保在某一序列距离内的序列(无论是野生型还是变体序列)将被包括在单一分区内,例如作为单一核酸片段。举例来说,在

共同或重叠条形码序列为映射至两个序列的条形码的总数的大于1%的情况下,其可用于识别两个序列区段之间并且特别是一般将不相连的两个序列区段之间的联系,例如结构变异。在一些情况下,共有或共同条形码可占可映射至两个一般分开的序列的全部条形码的超过2%、超过3%、超过4%、超过5%、超过6%、超过7%、超过8%并且在一些情况下超过9%或甚至超过10%,以识别构成基因组内的结构变异的联系。在一些情况下,可以统计显著大于已知不具有结构变异的对照基因组的比例或数目检测到共有或共同条形码。另外,在第二序列片段跨越变体序列遇到“正常”序列的点或“断点”(例如如在第二片段1358中)的情况下,可使用此信息作为基因融合的额外证据。

[0139] 再次,如上所述,可通过组装第二片段序列以产生基因融合1312的组装序列(以组装序列1360示出)来进一步阐明基因融合1312的结构。

[0140] 此外,虽然条形码序列的存在允许将短序列组装成更长起源片段的序列,但这些更长片段也允许从由不同的重叠起源长片段组装的重叠长片段推测更长范围序列信息。此所得组装允许基因融合1312的更长范围序列水平识别和表征。

[0141] 在一些情况下,上文所描述的方法适用于识别反转录转座子的存在。可通过剪接信使RNA(mRNA)的转录继之以反转录以及插入基因组中的新位置来形成反转录转座子。因此,这些结构变体缺少内含子并且经常为染色体间的,但以其他方式具有多样的特征。当反转录转座子引入基因的功能拷贝时,它们被称为反转录基因(retrogene),所述反转录基因在人和果蝇(*Drosophila*)基因组中已有报告。在其他情况下,反转录拷贝可含有整个转录物、特定转录物同种型或不完整转录物。另外,替代转录起始位点和启动子序列有时位于转录物内,所以反转录转座子有时在基因组的再插入区域内引入启动子序列,这会驱使下游序列的表达。

[0142] 不像串联重复,反转录转座子远离亲代基因插入外显子或内含子内。当在基因附近插入时,反转录转座子可利用就近调控序列来进行表达。在基因附近插入还可使接收基因失活或形成新的嵌合体转录物。反转录转座子介导的嵌合基因转录物在人样品的RNA-Seq数据中已有报告。

[0143] 尽管反转录转座子具重要意义,但其检测可能被限于定向方法,所述定向方法依赖于来自mate pair文库的成对读段支持、全基因组测序(WGS)中的外显子-外显子接合发现或反转录转座子嵌合体的RNA-Seq识别。所有这些方法均可能具有使分析复杂化的假阳性。

[0144] 可使用本文所描述的系统和方法从全基因组文库识别反转录转座子,并且可使用上文所论述的条形码映射来定位其插入位点。举例来说,Ceph NA12878基因组具有SKA3-DDX10嵌合反转录转座子。SKA3无内含子转录物被插入DDX10的外显子10与11之间。此外,还可使用本文所描述的方法检测NA12878中的CBX3-C15ORF17反转录转座子。CBX3的同种型2被插入C15ORF17的外显子2与3之间。已在20%的来自HapMap计划的欧洲RNA-Seq样品中观测到此嵌合转录物(D.R.Schrider等*PLoS Genetics* 2013)。

[0145] 还可使用本文所描述的方法和系统在所制备的全外显子组文库中检测反转录转座子。虽然在外显子组靶向的情况下反转录转座子容易富集,但可能难以或不可能区分易位事件和反转录转座子,因为内含子在捕获过程中被去除。然而,使用本文所描述的系统和方法,通过为疑似反转录转座子引入内含子诱饵可在全外显子组测序(WES)文库中识别反

转录转座子(也参见2014年10月29日提交的美国临时专利申请号62/072,164,该临时专利申请出于所有目的以全文引用的方式并入本文中)。缺少内含子信号可指示反转录转座子结构变体,而内含子信号可指示易位。

[0146] 如应了解,在识别和表征上文所描述的变异时使用更长范围序列环境的能力同样适用于通过将条形码映射至变异内和/或跨越变异的区域来识别其他结构变异的范围,包括插入、缺失、反转录转座子、倒位等。

[0147] V.由拷贝数变异引起的疾病和病症

[0148] 本发明方法和系统提供高度准确和灵敏的用于诊断和/或检测广泛范围的疾病和病症的方法。与拷贝数变异相关的疾病可包括例如迪乔治/腭心面综合征(DiGeorge/velocardiofacial syndrome)(22q11.2缺失)、普拉德-威利综合征(Prader-Willi syndrome)(15q11-q13缺失)、威廉-博伊伦综合征(Williams-Beuren syndrome)(7q11.23缺失)、米勒-狄克综合征(Miller-Dieker syndrome)(MDLS)(17p13.3微缺失)、史密斯-马吉利斯综合征(Smith-Magenis syndrome)(SMS)(17p11.2微缺失)、神经纤维瘤病1型(NF1)(17q11.2微缺失)、费伦-麦克德米德综合征(Phelan-McErmid Syndrome)(22q13缺失)、雷特综合征(Rett syndrome)(染色体Xq28上的MECP2的功能缺失突变)、梅茨巴赫病(Merzbacher disease)(PLP1的CNV)、脊髓性肌萎缩(SMA)(染色体5q13上端粒SMN1的纯合性不存在)、波托茨基-鲁普斯基综合征(Potocki-Lupski Syndrome)(PTLS,染色体17p.11.2重复)。PMP22基因的额外拷贝可与沙-马-图神经病变IA型(Charcot-Marie-Tooth neuropathy type IA,CMT1A)和遗传性压力易感性神经病变(hereditary neuropathy with liability to pressure palsies,HNPP)相关。所述疾病可为描述于Lupski J.(2007)Nature Genetics 39:S43-S47中的疾病。

[0149] 本文所提供的方法和系统还可以准确检测或诊断广泛范围的胎儿非整倍性。经常,本文所提供的方法包括分析从孕妇取得的样品(例如血液样品)以评估样品内的胎儿核酸。胎儿非整倍性可包括例如13三体(帕韬氏综合征(Patau syndrome))、18三体(爱德华兹综合征(Edwards syndrome))、21三体(唐氏综合征)、柯林菲特氏症(Klinefelter Syndrome)(XXY)、一个或多个染色体的单体性(X染色体单体性,特纳氏综合征)、X三体性、一个或多个染色体的三体性、一个或多个染色体的四体性或五体性(例如XXXX、XXYY、XXXY、XXYY、XXXXX、XXXXY、XXYY、XXXXY以及XXYY)、三倍性(每个染色体有三个,例如人中的69个染色体)、四倍性(每个染色体有四个,例如人中的92个染色体)以及多倍性。在一些实施方案中,非整倍性可为区段非整倍性。区段非整倍性可包括例如1p36重复、dup(17)(p11.2p11.2)综合征、唐氏综合征、佩利措伊斯-梅茨巴赫病(Pelizaeus-Merzbacher disease)、dup(22)(q11.2q11.2)综合征以及猫眼综合征。在一些情况下,异常基因型(例如胎儿基因型)是归因于性染色体或常染色体的一个或多个缺失,此可导致诸如以下病症:猫叫综合征(Cri-du-chat syndrome)、沃夫-贺许宏氏症(Wolf-Hirschhorn)、威廉-博伊伦综合征、沙-马-图病(Charcot-Marie-Tooth disease)、遗传性压力易感性神经病变、史密斯-马吉利斯综合征、神经纤维瘤病、阿拉吉耶综合征(Alagille syndrome)、腭心面综合征(Velocardiofacial syndrome)、迪乔治综合征(DiGeorges syndrome)、类固醇硫酸酯酶缺乏症、卡曼氏综合征(Kallmann syndrome)、小眼球线性皮肤缺损、肾上腺发育不良、甘油激酶缺乏症、佩利措伊斯-梅茨巴赫病、Y上睾丸决定因子、无精症(因子a)、无精症(因子b)、无精

症(因子c)或1p36缺失。在一些实施方案中,染色体数目减少导致XO综合征。

[0150] 过度基因组DNA拷贝数变异也与李-佛美尼癌症倾向综合征(Li-Fraumeni cancer predisposition syndrome)相关(Shlien等(2008)PNAS105:11264-9)。CNV与畸形综合征相关,包括CHARGE(眼部缺损、心脏异常、后鼻孔闭锁、发育迟缓、生殖器以及耳部异常)、彼得斯-普拉斯综合征(Peters-Plus)、皮特-霍普金斯综合征(Pitt-Hopkins)以及血小板减少-桡骨缺失综合征(thrombocytopenia-absent radius syndrome)(参见例如Ropers HH(2007)Am J of Hum Genetics 81:199-207)。拷贝数变异与癌症之间的关系描述于例如Shlien A.和Malkin D.(2009)Genome Med.1(6):62中。拷贝数变异与例如自闭症、精神分裂症以及特发性学习障碍相关。参见例如Sebat J.等(2007)Science 316:445-9;Pinto J.等。

[0151] 如本文所描述,本文所提供的方法和系统还适用于检测与不同类型的癌症相关的CNV。举例来说,可使用所述方法和系统来检测EGFR拷贝数,在非小细胞肺癌中其可为增加的。

[0152] 还可使用本文所提供的方法和系统来测定受试者对特定疾病或病症的易感性水平,包括病原体感染易感性(例如病毒、细菌、微生物、真菌等)。举例来说,鉴于相对高水平的CCL3L1与更低HIV感染易感性相关,可使用所述方法通过分析CCL3L1的拷贝数来测定受试者对HIV感染的易感性(Gonzalez E.等(2005)Science 307:1434-1440)。在另一实例中,可使用所述方法来测定受试者对系统性红斑狼疮的易感性。在此类情况下,举例来说,可使用所述方法来检测FCGR3B(CD16细胞表面免疫球蛋白受体)的拷贝数,因为此分子的低拷贝数与增加的系统性红斑狼疮易感性相关(Aitman T.J.等(2006)Nature 439:851-855)。还可使用本文所提供的方法和系统来检测与其他疾病或病症相关的CNV,诸如与自闭症、精神分裂症以及特发性学习障碍相关的CNV(Kinght等,(1999)TheLancet 354(9191):1676-81)。类似地,可使用所述方法和系统来检测常染色体显性小耳症,其与染色体4p16处的拷贝数可变区的五个串联拷贝有关(Balikova I.(2008)Am J.Hum Genet.82:181-187)。

[0153] VI.疾病和病症的检测、诊断以及治疗

[0154] 本文所提供的方法和系统还可协助检测、诊断以及治疗疾病或病症。在一些情况下,一种方法包括使用本文所描述的系统或方法检测疾病或病症,并且基于对疾病的检测进一步为受试者提供治疗。举例来说,如果检测到癌症,那么可通过外科手术、通过施用被设计成治疗此类癌症的药物、通过提供激素治疗以及/或者通过施用辐射或更一般化的化学治疗来治疗受试者。

[0155] 经常,所述方法和系统还允许鉴别诊断并且可进一步包括用靶向治疗来治疗患者。一般来说,可通过以下方式实现疾病或病症的鉴别诊断(或其不存在):测定和表征从怀疑患有所述疾病或病症的受试者获得的样品核酸的序列,并且通过将其与指示病症或疾病状态存在(或不存在)的参考核酸的序列和/或序列表征相比较将样品核酸进一步表征为指示病症或疾病状态(或其不存在)。

[0156] 参考核酸序列可源于指示疾病或病症状态不存在的基因组(例如生殖系核酸)或可源于指示疾病或病症状态的基因组(例如癌症核酸、指示非整倍性等的核酸)。另外,可在一个或多个方面表征参考核酸序列(例如具有长于1kb、长于5kb、长于10kb、长于15kb、长于20kb、长于30kb、长于40kb、长于50kb、长于60kb、长于70kb、长于80kb、长于90kb或甚至长于

100kb的长度),其中非限制性实例包括确定特定序列的存在(或不存在)、确定特定单倍型的存在(或不存在)、确定一种或多种基因变异(例如结构变异(例如拷贝数变异、插入、缺失、易位、倒位、反转录转座子、重排、重复扩增、重复等)、单核苷酸多态性(SNP)等)的存在(或不存在)以及其组合。另外,可使用任何合适类型和数目的参考序列的序列特征来表征样品核酸的序列。举例来说,可使用参考核酸序列的一种或多种基因变异(或其缺乏)或结构变异(或其缺乏)作为用于将参考核酸识别为指示病症或疾病状态存在(或不存在)的序列标签。基于对所利用的参考核酸序列的表征,可以类似方式表征样品核酸序列并且基于其是否展示与参考核酸序列类似的性质将其进一步表征/识别为源于(或不源于)指示病症或疾病的核酸。在一些情况下,可在经过编程的计算机处理器的辅助下完成样品核酸序列和/或参考核酸序列的表征以及其比较。在一些情况下,此类经过编程的计算机处理器可被包括于计算机控制系统中,诸如本文中别处所描述的示例性计算机控制系统中。

[0157] 可从任何合适的来源获得样品核酸,包括本文中别处所描述的样品来源和生物样品来源。在一些情况下,样品核酸可包括无细胞核酸。在一些情况下,样品核酸可包括肿瘤核酸(例如,肿瘤DNA)。在一些情况下,样品核酸可包括循环肿瘤核酸(例如,循环肿瘤DNA(ctDNA))。循环肿瘤核酸可源于循环肿瘤细胞(CTC)和/或可从例如受试者的血液、血浆、其他体液或组织获得。

[0158] 图20-21说明用于在疾病检测和诊断的背景下表征样品核酸的示例性方法。图20展现用于诸如以类似于图6中所示的方式由更短条形码化片段测定参考核酸(例如生殖系核酸(例如生殖系基因组DNA)、与特定病症或疾病状态相关的核酸)的长范围序列环境的示例性方法。就图20来说,可获得参考核酸2000,并且可获得条形码化珠粒的集合2010。可使珠粒键联至含有一个或多个条形码序列以及引物(诸如随机N-mer或其他引物)的寡核苷酸。在一些情况下,条形码序列可从条形码化珠粒释放,例如通过在条形码与珠粒之间的键的裂解或通过下面珠粒的降解来释放条形码,或两种途径的组合。举例来说,在一些方面,条形码化珠粒可由诸如还原剂等试剂降解或溶解以释放条形码序列。在此实例中,将参考核酸2005、条形码化珠粒、2015以及(在一些情况下)其他试剂(例如还原剂)2020组合并且进行分配。在一些情况下,可在分配之前将参考核酸2000片段化,并且将所得片段中的至少一些如2005进行分配以便进行条形码化。举例来说,此类分配可涉及将组分引入微滴产生系统,诸如微流体装置2025。在微流体装置2025的辅助下,可形成油包水乳液2030,其中所述乳液含有水性微滴,所述水性微滴含有参考核酸2005、还原剂2020以及条形码化珠粒2015。还原剂可溶解或降解条形码化珠粒,由此使具有条形码和随机N-mer的寡核苷酸从微滴内的珠粒释放2035。随机N-mer可然后引导参考核酸的不同区域,从而在扩增之后产生参考核酸的扩增拷贝,其中将各拷贝用条形码序列标记2040。在一些情况下,可通过类似于本文中别处所描述并且示意性描绘于图5中的方法来实现扩增2040。在一些情况下,各微滴含有寡核苷酸的集合,所述寡核苷酸的集合含有相同的条形码序列和不同的随机N-mer序列。随后,将乳液破坏2045,并且可经由例如扩增方法2050(例如PCR)来添加额外序列(例如辅助特定测序方法的序列、额外条形码等)。然后可进行测序2055,并且应用算法来解释测序数据2060。在一些情况下,测序数据的解释2060可包括提供参考核酸的至少一部分的序列。在一些情况下,获得参考核酸的长范围序列环境并且进行表征,诸如在参考核酸源于疾病状态的情况下(例如如本文中别处所描述的一个或多个单倍型的测定、一种或多种结构变

异(例如拷贝数变异、插入、缺失、易位、倒位、重排、重复扩增、重复、反转录转座子、基因融合等)的测定、一个或多个SNP等的分辨等)。在一些情况下,可针对从来源获得的各种参考核酸和所产生的推测重叠群来识别变体以提供更长范围序列环境,诸如本文中别处关于图7所描述。

[0159] 图21展现由如图20中所示所获得的参考2060表征来表征样品核酸序列的实例。可如本文中别处所描述诸如经由示意性描绘于图6中的方法由更短条形码化片段的测序获得样品核酸的长范围序列环境。如图21中所示,可从怀疑患有病症或疾病(例如癌症)的受试者获得核酸样品(例如包含循环肿瘤核酸的样品)2100,并且还可获得条形码化珠粒2110。可使珠粒键联至含有一个或多个条形码序列以及引物(诸如随机N-mer或其他引物)的寡核苷酸。在一些情况下,条形码序列可从条形码化珠粒释放,例如通过在条形码与珠粒之间的键的裂解或通过下面珠粒的降解来释放条形码,或两种途径的组合。举例来说,在一些方面,条形码化珠粒可由诸如还原剂等试剂降解或溶解以释放条形码序列。在此实例中,将样品核酸2105、条形码化珠粒2115以及(在一些情况下)其他试剂(例如还原剂)2120组合并且进行分配。在一些情况下,在分配之前将胎儿样品2100片段化并且对所得片段中的至少一些如2105进行分配以便条形码化。举例来说,此类分配可涉及将组分引入微滴产生系统(诸如微流体装置)2125。在微流体装置2125的辅助下,可形成油包水乳液2130,其中所述乳液含有水性微滴,所述水性微滴含有样品核酸2105、还原剂2120以及条形码化珠粒2115。还原剂可溶解或降解条形码化珠粒,由此使具有条形码和随机N-mer的寡核苷酸从微滴内的珠粒释放2135。随机N-mer可然后引导样品核酸的不同区域,从而在扩增之后产生样品核酸的扩增拷贝,其中将各拷贝用条形码序列标记2140。在一些情况下,可通过类似于本文中别处所描述并且示意性描绘于图5中的方法来实现扩增2140。在一些情况下,各微滴含有寡核苷酸的集合,所述寡核苷酸的集合含有相同的条形码序列和不同的随机N-mer序列。随后,将乳液破坏2145,并且可经由例如扩增方法2150(例如PCR)来添加额外序列(例如辅助特定测序方法的序列、额外条形码等)。然后可进行测序2155并且应用算法来解释测序数据2160。在一些情况下,测序数据的解释2160可包括提供样品核酸的序列。在一些情况下,获得核酸样品的长范围序列环境。可使用对参考核酸序列的表征2060来表征样品核酸序列2160(例如测定如本文中别处所描述的一个或多个单倍型、测定一种或多种结构变异(例如拷贝数变异、插入、缺失、易位、倒位、重排、重复扩增、重复、反转录转座子、基因融合等)。基于样品核酸序列及其表征与参考核酸的序列和表征的比较,可进行关于病症或疾病状态的存在(或不存在)的鉴别诊断2170。

[0160] 如可理解,参考核酸和样品核酸的分析可作为单独分配分析来完成或可作为单一分配分析的一部分来完成。举例来说,可将样品和参考核酸添加至同一装置,并且根据图20和21在微滴中产生条形码化的样品和参考片段,其中乳液包含用于两种类型的核酸的微滴。然后可将乳液破坏,并且汇集微滴的内容物,进一步处理(例如经由PCR批量添加额外序列)并且如本文中别处所描述进行测序。可经由条形码序列将来自条形码化片段的单个测序读段归属至其各自的样品序列。可基于对参考核酸序列的表征来表征从样品核酸获得的序列。

[0161] 利用本文的方法和系统可提高测定核酸的长范围序列环境(包括如本文所描述的参考和样品核酸序列的长范围序列环境)的准确度。本文所提供的方法和系统可以至少

70%、80%、85%、90%、91%、92%、93%、94%、95%、99%、99.1%、99.2%、99.3%、99.4%、99.5%、99.6%、99.7%、99.8%、99.9%、99.95%、99.99%、99.995%、或99.999%的准确度测定参考和/或样品核酸的长范围序列环境。在一些情况下,本文所提供的方法和系统可以低于10%、9%、8%、7%、6%、5%、4%、3%、2%、1%、0.1%、0.05%、0.01%、0.005%、0.001%、0.0005%、0.0001%、0.00005%、0.00001%、或0.000005%的误差率测定参考和/或样品核酸的长范围序列环境。

[0162] 另外,本文的方法和系统还可在一个或多个方面(例如序列的测定、一种或多种基因变异的测定、单倍型的测定等)提高表征参考核酸序列和/或样品核酸序列时的准确度。因此,本文所提供的方法和系统可以至少70%、80%、85%、90%、91%、92%、93%、94%、95%、99%、99.1%、99.2%、99.3%、99.4%、99.5%、99.6%、99.7%、99.8%、99.9%、99.95%、99.99%、99.995%、或99.999%的准确度在一个或多个方面表征参考核酸序列和/或样品核酸序列。在一些情况下,本文所提供的方法和系统可以低于10%、9%、8%、7%、6%、5%、4%、3%、2%、1%、0.1%、0.05%、0.01%、0.005%、0.001%、0.0005%、0.0001%、0.00005%、0.00001%、或0.000005%的误差率在一个或多个方面表征参考核酸序列和/或样品核酸序列。

[0163] 另外,如上文所论述,测定参考核酸的长范围序列环境 and 对其进行表征时的准确度提高可使得对样品核酸进行测序和表征以及随后用于鉴别诊断病症或疾病时的准确度提高。因此,可以低于10%、9%、8%、7%、6%、5%、4%、3%、2%、1%、0.1%、0.05%、0.01%、0.005%、0.001%、0.0005%、0.0001%、0.00005%、0.00001%、或0.000005%的误差率由参考核酸序列的分析提供样品核酸序列(包括长范围序列环境)。在一些情况下,可通过与参考核酸的序列和/或序列表征相比较使用样品核酸序列以至少70%、80%、85%、90%、91%、92%、93%、94%、95%、99%、99.1%、99.2%、99.3%、99.4%、99.5%、99.6%、99.7%、99.8%、99.9%、99.95%、99.99%、99.995%、或99.999%的准确度来鉴别诊断病症或疾病(或其不存在)。在一些情况下,可通过与参考核酸的序列和/或序列表征相比较使用样品核酸序列以低于10%、9%、8%、7%、6%、5%、4%、3%、2%、1%、0.1%、0.05%、0.01%、0.005%、0.001%、0.0005%、0.0001%、0.00005%、0.00001%、或0.000005%的误差率来鉴别诊断病症或疾病(或其不存在)。

[0164] 在一个实例中,可使用所述方法和系统来检测肺癌患者的拷贝数变异以确定肺癌是否是EGFR基因变异相关的非小细胞肺癌。在此类诊断之后,可细化患者的治疗方案以与鉴别诊断相关联。靶向治疗或分子靶向治疗是癌症医学治疗(药物治疗)的主要形式之一,其他是激素治疗和细胞毒性化学治疗。靶向治疗通过干扰为致癌作用和肿瘤生长所需的特定靶向分子,而不是通过简单地干扰所有快速分裂的细胞(例如在传统化学治疗情况下)来阻碍癌细胞生长。

[0165] 图14示出了鉴别诊断非小细胞肺癌的示例性方法。对慢性咳嗽、体重减轻以及呼吸急促患者进行关于肺癌的测试1400。从患者抽取血液1405,并且从血液得到样品(例如循环肿瘤细胞、无细胞DNA、循环核酸(例如循环肿瘤核酸)等)1410。还可获得条形码化珠粒的集合1415。可使珠粒键联至含有一个或多个条形码序列以及引物(诸如随机N-mer或其他引物)的寡核苷酸。在一些情况下,条形码序列可从条形码化珠粒释放,例如通过在条形码与珠粒之间的键的裂解或通过下面珠粒的降解来释放条形码,或两种途径的组合。举例来说,

在一些方面,条形码化珠粒可由诸如还原剂等试剂降解或溶解以释放条形码序列。在此实例中,将样品1410、条形码化珠粒1420以及(在一些情况下)其他试剂(例如还原剂)组合并且进行分配。举例来说,此类分配可涉及将组分引入微滴产生系统,诸如微流体装置1425。在微流体装置1425的辅助下,可形成油包水乳液1430,其中所述乳液含有水性微滴,所述水性微滴含有样品核酸1410、条形码化珠粒1415以及(在一些情况下)还原剂。还原剂可溶解或降解条形码化珠粒,由此使具有条形码和随机N-mer的寡核苷酸从微滴内的珠粒释放1435。随机N-mer可然后引导样品核酸的不同区域,从而在扩增之后产生样品的扩增拷贝,其中将各拷贝用条形码序列标记1440。在一些情况下,各微滴含有寡核苷酸的集合,所述寡核苷酸的集合含有相同的条形码序列和不同的随机N-mer序列。随后,将乳液破坏1445,并且可经由例如扩增方法(例如PCR)来添加额外序列(例如辅助特定测序方法的序列、额外条形码等)。然后可进行测序1450,并且应用算法来解释测序数据1455。测序算法通常能够例如对条形码进行分析以比对测序读段和/或识别特定序列读段所属的样品。

[0166] 然后将分析过的序列与已知基因组参考序列相比较以确定不同基因的CNV 1460。如果DNA中的EGFR拷贝数高于正常,那么可将患者鉴别诊断为患有非小细胞肺癌(NSCLC)而不是小细胞肺癌1465。非小细胞肺癌的CTC还具有其他拷贝数变异,所述其他拷贝数变异可使其与小细胞肺癌进一步区分开。视癌症的阶段而定,嘱咐进行手术、化学治疗或放射治疗1470。在一些情况下,为被诊断为患有NSLC的患者施用针对此类癌症的药物,诸如ALK抑制剂(例如克唑替尼(Crizotinib))。在EGFR的变异的一些情况下,为患者施用西妥昔单抗(cetuximab)、帕尼单抗(panitumumab)、拉帕替尼(lapatinib)以及/或者卡培他滨(capecitabine)。在不同的情况下,目标可为不同基因,诸如ERBB2,并且治疗包括曲妥珠单抗(trastuzumab)(赫赛汀(Herceptin))。(2010) Nature 466:368-72; Cook E. H. 和 Scherer S. W. (2008) Nature 455:919-923。

[0167] 靶向治疗的主要类别为小分子、小分子药物缀合物以及单克隆抗体。小分子可包括酪氨酸激酶抑制剂,诸如伊马替尼(Imatinib)甲磺酸盐(格列卫(Gleevec),也被称为STI-571)(其被批准用于慢性粒细胞性白血病、胃肠道基质肿瘤以及一些其他类型的癌症);吉非替尼(Gefitinib)(易瑞沙(Iressa),也被称为ZD1839)(其靶向表皮生长因子受体(EGFR)酪氨酸激酶并且在美国被批准用于非小细胞肺癌);厄洛替尼(Erlotinib)(以特罗凯(Tarceva)形式出售);硼替佐米(Bortezomib)(万珂(Velcade))(其为诱导细胞凋亡的蛋白酶体抑制剂药物,其通过干扰蛋白质而使得癌细胞经历细胞死亡);他莫昔芬(tamoxifen);JAK抑制剂(例如托法替尼(tofacitinib)、ALK抑制剂(例如克唑替尼);Bcl-2抑制剂(例如在临床试验中的奥巴克拉(obatoclax)、ABT-263以及棉酚(Gossypol));PARP抑制剂(例如依尼帕尼(Iniparib)、在临床试验中的奥拉帕尼(Olaparib));PI3K抑制剂(例如在III期试验中的哌立福辛(perifosine))。阿帕替尼(其为选择性VEGF受体2抑制剂);AN-152, (AEZS-108)与[D-Lys(6)]-LHRH键联的多柔比星(doxorubicin);Braf抑制剂(维罗非尼(vemurafenib)、达拉菲尼(dabrafenib)、LGX818)(用于治疗具有BRAF V600E突变的转移性黑色素瘤);MEK抑制剂(曲美替尼(trametinib), MEK162);CDK抑制剂,例如PD-0332991、在临床试验中的LEE011;Hsp90抑制剂;以及沙利霉素(Salinomycin)。

[0168] 其他治疗包括小分子药物缀合物,诸如Vintafolide,其为由靶向叶酸受体的小分子组成的小分子药物缀合物。

[0169] 单克隆抗体为另一类型的治疗,其可作为本文所提供的方法的一部分进行施用。还可施用单克隆药物缀合物。示例性单克隆抗体包括:利妥昔单抗(Rituximab)(以MabThera或Rituxan出售)(其靶向存在于B细胞上的CD20,并且靶向非霍奇金淋巴瘤(non Hodgkin lymphoma));曲妥珠单抗(赫赛汀)(其靶向在一些类型的乳房癌中表达的Her2/neu(也称为ErbB2)受体);西妥昔单抗(以(Erbitux出售)以及帕尼单抗贝伐单抗(Bevacizumab)(以Avastin形式出售)(其靶向VEGF配位体)。

[0170] VII. 由亲代核酸表征胎儿核酸

[0171] 如本文中别处所提到,还可使用本文所描述的方法和系统来表征受试者的血液或血浆内的循环核酸。此类分析包括分析循环肿瘤DNA,以便用于识别患者中的潜在疾病状态;或怀孕女性的血液或血浆内的循环胎儿DNA,以例如在没有通过羊膜穿刺术或其他侵入性程序进行直接取样的情况下按无创的方式表征胎儿DNA。

[0172] 在一些情况下,可至少部分基于亲代核酸序列的分析使用所述方法来表征胎儿核酸序列,例如循环胎儿DNA。举例来说,可使用本文所描述的方法和系统由更短条形码化片段测定父体与母体核酸的长范围序列环境(例如具有长于1kb、长于5kb、长于10kb、长于15kb、长于20kb、长于30kb、长于40kb、长于50kb、长于60kb、长于70kb、长于80kb、长于90kb或甚至长于100kb的长度)。可使用长范围序列环境来测定一个或多个单倍型和一种或多种基因变异,包括父体与母体核酸序列中的单核苷酸多态性(SNP)、结构变异(例如拷贝数变异、插入、缺失、易位、倒位、重排、重复扩增、反转录转座子、重复、基因融合等)。另外,可使用父体和母体核酸的长范围序列环境以及任何所测定的SNP、单倍型和/或结构变异信息来表征从怀孕母亲获得的胎儿核酸(例如循环胎儿核酸,诸如无细胞胎儿核酸)的序列。在一些情况下,经由与母体和父体序列和表征相比较来表征胎儿核酸可在经过编程的计算机处理器的辅助下完成。在一些情况下,此类经过编程的计算机处理器可被包括于计算机控制系统中,诸如本文中别处所描述的示例性计算机控制系统中。

[0173] 举例来说,可使用亲代和/或母体核酸的序列和/或长范围序列环境作为用于表征胎儿核酸(包括胎儿核酸序列)的参考。事实上,由本文所描述的方法和系统获得的长范围序列环境可提供父体和母体核酸的改善的长范围序列环境信息,由此可表征胎儿核酸序列。在一些情况下,由亲代核酸作为参考来表征胎儿核酸序列可包括测定胎儿核酸的至少一部分的序列,以及/或者识别胎儿核酸序列的一个或多个SNP,测定胎儿核酸序列的一个或多个从头突变,测定胎儿核酸序列的一个或多个单倍型,以及/或者测定和表征胎儿核酸序列中的一种或多种结构变异等。

[0174] 图17-19说明经由对更短条形码化片段进行测序由针对父体和母体核酸所获得的更长范围序列环境表征胎儿核酸的示例性方法。图17展现可用于诸如以类似于图6中所示的方法由更短条形码化片段测定父体核酸样品(例如父体基因组DNA)的更长范围序列环境的示例性方法。就图17来说,可从胎儿的父亲获得包含父体核酸的样品1700,并且还可获得条形码化珠粒集合1710。可使珠粒键联至含有一个或多个条形码序列以及引物(诸如随机N-mer或其他引物)的寡核苷酸。在一些情况下,条形码序列可从条形码化珠粒释放,例如通过在条形码与珠粒之间的键的裂解或通过下面珠粒的降解来释放条形码,或两种途径的组合。举例来说,在一些方面,条形码化珠粒可由诸如还原剂等试剂降解或溶解以释放条形码序列。在此实例中,将包含核酸的父体样品1705、条形码化珠粒1715以及(在一些情况下)其

他试剂(例如还原剂) 1720组合并且进行分配。在一些情况下,在分配之前将父体样品1700片段化,并且对所得片段中的至少一些如1705进行分配以便进行条形码化。举例来说,此类分配可涉及将组分引入微滴产生系统,诸如微流体装置1725。在微流体装置1725的辅助下,可形成油包水乳液1730,其中所述乳液含有水性微滴,所述水性微滴含有父体样品核酸1705、还原剂1720以及条形码化珠粒1715。还原剂可溶解或降解条形码化珠粒,由此使具有条形码和随机N-mer的寡核苷酸从微滴内的珠粒释放1735。随机N-mer可然后引导父体样品核酸的不同区域,从而在扩增之后产生父体样品的扩增拷贝,其中将各拷贝用条形码序列标记1740。在一些情况下,可通过类似于本文中别处所描述并且示意性描绘于图5中的方法来实现扩增1740。在一些情况下,各微滴含有寡核苷酸的集合,所述寡核苷酸的集合含有相同的条形码序列和不同的随机N-mer序列。随后,将乳液破坏1745,并且可经由例如扩增方法1750(例如PCR)来添加额外序列(例如辅助特定测序方法的序列、额外条形码等)。然后可进行测序1755,并且应用算法来解释测序数据1760。在一些情况下,举例来说,测序数据的解释1760可包括提供父体核酸的至少一部分的序列。在一些情况下,可获得父体核酸样品的长范围序列环境并且进行表征(例如测定如本文中别处所描述的一个或多个单倍型、测定一种或多种结构变异(例如拷贝数变异、插入、缺失、易位、倒位、重排、重复扩增、重复、反转录转座子、基因融合等)、识别一个或多个SNP、测定一种或多种其他基因变异等)。在一些情况下,可针对各种父体核酸和所产生的推测重叠群来识别变体以提供更长范围序列环境,诸如本文中别处关于图7所描述。

[0175] 图18展现可用于诸如以类似于图6中所示的方法由更短条形码化片段测定母体核酸样品(例如母体基因组DNA)的长范围序列环境的示例性方法。就图18来说,可从胎儿的怀孕母亲获得包含母体核酸的样品1800,并且还可获得条形码化珠粒1810。可使珠粒键联至含有一个或多个条形码序列以及引物(诸如随机N-mer或其他引物)的寡核苷酸。在一些情况下,条形码序列可从条形码化珠粒释放,例如通过在条形码与珠粒之间的键的裂解或通过下面珠粒的降解来释放条形码,或两种途径的组合。举例来说,在一些方面,条形码化珠粒可由诸如还原剂等试剂降解或溶解以释放条形码序列。在此实例中,将包含核酸的母体样品1805、条形码化珠粒1815以及(在一些情况下)其他试剂(例如还原剂) 1820组合并且进行分配。在一些情况下,在分配之前将母体样品1800片段化并且对所得片段中的至少一些如1805进行分配以便进行条形码化。举例来说,此类分配可涉及将组分引入微滴产生系统,诸如微流体装置1825。在微流体装置1825的辅助下,可形成油包水乳液1830,其中所述乳液含有水性微滴,所述水性微滴含有母体样品核酸1805、还原剂1820以及条形码化珠粒1815。还原剂可溶解或降解条形码化珠粒,由此使具有条形码和随机N-mer的寡核苷酸从微滴内的珠粒释放1835。随机N-mer可然后引导母体样品核酸的不同区域,从而在扩增之后产生母体样品的扩增拷贝,其中将各拷贝用条形码序列标记1840。在一些情况下,可通过类似于本文中别处所描述并且示意性展示于图5中的方法来实现扩增1840。在一些情况下,各微滴含有寡核苷酸的集合,所述寡核苷酸的集合含有相同的条形码序列和不同的随机N-mer序列。随后,将乳液破坏1845,并且可经由例如扩增方法1850(例如PCR)来添加额外序列(例如辅助特定测序方法的序列、额外条形码等)。然后可进行测序1855,并且应用算法来解释测序数据1860。在一些情况下,举例来说,测序数据的解释1860可包括提供母体核酸的至少一部分的序列。在一些情况下,可获得母体核酸样品的长范围序列环境并且进行表征(例如测定

如本文中别处所描述的一个或多个单倍型、测定一种或多种结构变异(例如拷贝数变异、插入、缺失、易位、倒位、重排、重复扩增、重复、反转录转座子、基因融合等)、识别一个或多个 SNP、测定一个或多个其他基因变异等。在一些情况下,可针对从样品获得的各种母体核酸和所产生的推测重叠群来识别变体以提供更长范围序列环境,诸如本文中别处关于图7所描述。

[0176] 图19展现由分别如图17和图18中所示所获得父体1760和母体1860表征来表征胎儿样品序列的实例。如图19中所示,可从怀孕母亲获得胎儿核酸样品1900。可如本文中别处所描述诸如经由示意性描绘于图6中的方法由更短条形码化片段的测序获得胎儿核酸的长范围序列环境。在一些情况下,胎儿核酸样品可为循环胎儿DNA和/或无细胞DNA,其可例如从怀孕母亲的血液、血浆、其他身体流体或组织获得。还可获得条形码化珠粒的集合1910。可使珠粒键联至含有一个或多个条形码序列以及引物(诸如随机N-mer或其他引物)的寡核苷酸。在一些情况下,条形码序列可从条形码化珠粒释放,例如通过在条形码与珠粒之间的键的裂解或通过下面珠粒的降解来释放条形码,或两种途径的组合。举例来说,在一些方面,条形码化珠粒可由诸如还原剂等试剂降解或溶解以释放条形码序列。在此实例中,将包含核酸的胎儿样品1905、条形码化珠粒1915以及(在一些情况下)其他试剂(例如还原剂)1920组合并且进行分配1905。在一些情况下,在分配之前将胎儿样品1900片段化并且对所得片段中的至少一些如1905进行分配以便进行条形码化。举例来说,此类分配可涉及将组分引入微滴产生系统,诸如微流体装置1925。在微流体装置1925的辅助下,可形成油包水乳液1930,其中所述乳液含有水性微滴,所述水性微滴含有母体样品核酸1905、还原剂1920以及条形码化珠粒1915。还原剂可溶解或降解条形码化珠粒,由此使具有条形码和随机N-mer的寡核苷酸从微滴内的珠粒释放1935。随机N-mer可然后引导胎儿样品核酸的不同区域,从而在扩增之后产生胎儿样品的扩增拷贝,其中将各拷贝用条形码序列标记1940。在一些情况下,可通过类似于本文中别处所描述并且示意性描绘于图5中的方法来实现扩增1940。在一些情况下,各微滴含有寡核苷酸的集合,所述寡核苷酸的集合含有相同的条形码序列和不同的随机N-mer序列。随后,将乳液破坏1945,并且可经由例如扩增方法1950(例如PCR)来添加额外序列(例如辅助特定测序方法的序列、额外条形码等)。然后可进行测序1955,并且应用算法来解释测序数据1960。一般来说,可从测序过的更短条形码化片段获得胎儿核酸样品的更长范围序列环境。在一些情况下,举例来说,测序数据的解释1960可包括提供胎儿核酸的至少一部分的序列。可使用父体1760和母体1860样品的长范围序列环境和/或表征来表征胎儿核酸序列1960(例如测定如本文中别处所描述的一个或多个单倍型、测定一种或多种结构变异(例如拷贝数变异、插入、缺失、易位、倒位、重排、重复扩增、重复、反转录转座子、基因融合等)、测定一个或多个从头突变、识别一个或多个SNP等)。在一些情况下,可通过将胎儿核酸序列与母体和父体相位区块相比较来测定胎儿核酸的相位区块。

[0177] 如可理解,父体核酸、母体核酸以及/或者胎儿核酸的分析可作为单独分配分析的一部分来完成或可作为一个或多个组合分配分析的一部分来完成。举例来说,可将父体、母体以及胎儿核酸添加至同一装置,并且根据图17-19在微滴中产生条形码化母体、父体以及胎儿片段,其中乳液包含用于三种类型的核酸的微滴。然后可将乳液破坏,并且汇集微滴的内容物,进一步处理(例如经由PCR批量添加额外序列)并且如本文中别处所描述进行测序。可经由条形码序列将来自条形码化片段的单个测序读段归属至其各自的样品序列。

[0178] 在一些情况下,可由使用本文所描述的方法和系统获得的长范围父体和母体序列环境和表征测定胎儿核酸的序列(包括胎儿基因组的序列)和/或胎儿核酸序列中的基因变异。举例来说,可使用父体和母体基因组的基因组测序以及循环胎儿核酸的测序来测定相应的胎儿基因组序列。由亲代基因组和无细胞胎儿核酸的序列分析测定基因组胎儿核酸的序列的实例可见于Kitzman等(2012年6月6日)Sci Transl.Med.4(137):137ra76中,该文献以引用的方式全部并入本文中。胎儿基因组的测定可适用于胎儿的基因病症(包括例如胎儿非整倍性)的产前确定和诊断。如本文中别处所论述,本文所提供的方法和系统可适用于解析核酸序列中的单倍型。可分别测定父体和母体样品核酸序列的单倍型解析型父体和母体序列,其可辅助更准确测定胎儿基因组的序列和/或对其进行表征。

[0179] 利用本文的方法和系统可提高测定核酸的长范围序列环境(包括亲代核酸序列(例如母体核酸序列、父体核酸序列)的长范围序列环境)时的准确度。本文所提供的方法和系统可以至少70%、80%、85%、90%、91%、92%、93%、94%、95%、99%、99.1%、99.2%、99.3%99.4%、99.5%、99.6%、99.7%、99.8%、99.9%、99.95%、99.99%、99.995%或99.999%的准确度测定亲代核酸的长范围序列环境。在一些情况下,本文所提供的方法和系统可以低于10%、9%、8%、7%、6%、5%、4%、3%、2%、1%、0.1%、0.05%、0.01%、0.005%、0.001%、0.0005%、0.0001%、0.00005%、0.00001%或0.000005%的误差率测定亲代核酸的长范围序列环境。另外,本文的方法和系统还可可在一个或多个方面(例如序列的测定、一种或多种基因变异的测定、一个或多个结构变体的测定、单倍型的测定等)提高表征父体核酸序列时的准确度。因此,本文所提供的方法和系统可以至少70%、80%、85%、90%、91%、92%、93%、94%、95%、99%、99.1%、99.2%、99.3%99.4%、99.5%、99.6%、99.7%、99.8%、99.9%、99.95%、99.99%、99.995%或99.999%准确度在一个或多个方面表征父体核酸序列。在一些情况下,本文所提供的方法和系统可以低于10%、9%、8%、7%、6%、5%、4%、3%、2%、1%、0.1%、0.05%、0.01%、0.005%、0.001%、0.0005%、0.0001%、0.00005%、0.00001%或0.000005%的误差率在一个或多个方面表征亲代核酸序列。

[0180] 另外,如上文所论述,测定亲代核酸的长范围序列环境和对其进行表征时的准确度提高可使得对胎儿核酸进行测序和表征时的准确度提高。因此,在一些情况下,可由亲代核酸序列的分析以至少70%、80%、85%、90%、91%、92%、93%、94%、95%、99%、99.1%、99.2%、99.3%99.4%、99.5%、99.6%、99.7%、99.8%、99.9%、99.95%、99.99%、99.995%或99.999%的准确度提供胎儿核酸序列(包括长范围序列环境)。在一些情况下,可由亲代核酸序列的分析以低于10%、9%、8%、7%、6%、5%、4%、3%、2%、1%、0.1%、0.05%、0.01%、0.005%、0.001%、0.0005%、0.0001%、0.00005%、0.00001%或0.000005%的误差率提供胎儿核酸序列(包括长范围序列环境)。在一些情况下,可以至少70%、80%、85%、90%、91%、92%、93%、94%、95%、99%、99.1%、99.2%、99.3%99.4%、99.5%、99.6%、99.7%、99.8%、99.9%、99.95%、99.99%、99.995%或99.999%的准确度在一个或多个方面经由如本文所描述的亲代核酸序列的分析(例如序列的测定、一种或多种基因变异的测定、一种或多种结构变异的测定、单倍型的测定等)表征胎儿核酸序列。在一些情况下,可以低于10%、9%、8%、7%、6%、5%、4%、3%、2%、1%、0.1%、0.05%、0.01%、0.005%、0.001%、0.0005%、0.0001%、0.00005%、0.00001%或0.000005%的误

差率在一个或多个方面经由如本文所描述的亲代核酸序列的分析(例如序列的测定、一种或多种基因变异的测定、单倍型的测定、一种或多种结构变异的测定等)表征胎儿核酸序列。

[0181] VIII. 样品

[0182] 可以从患者获得样品来开始疾病或病症的检测。如本文中所用,术语“样品”通常是指生物样品。生物样品的实例包括核酸分子、氨基酸、多肽、蛋白质、碳水化合物、脂肪或病毒。在一个实例中,生物样品为包括一个或多个核酸分子的核酸样品。示例性样品可包括聚核苷酸、核酸、寡核苷酸、无细胞核酸(例如无细胞DNA(cfDNA))、循环无细胞核酸、循环肿瘤核酸(例如循环肿瘤DNA(ctDNA))、循环肿瘤细胞(CTC)核酸、核酸片段、核苷酸、DNA、RNA、肽聚核苷酸、互补DNA(cDNA)、双链DNA(dsDNA)、单链DNA(ssDNA)、质粒DNA、粘粒DNA、染色体DNA、基因组DNA(gDNA)、病毒DNA、细菌DNA、mtDNA(线粒体DNA)、核糖体RNA、无细胞DNA、无细胞胎儿DNA(cffDNA)、mRNA、rRNA、tRNA、nRNA、siRNA、snRNA、snoRNA、scaRNA、微RNA、dsRNA、病毒RNA等。总之,所用的样品可视特定处理需要而变化。

[0183] 包含核酸的任何物质均可作为样品的来源。物质可为流体,例如生物流体。流体物质可包括但不限于血液、脐带血、唾液、尿液、汗液、血清、精液、阴道流体、胃部和消化流体、脊髓流体、胎盘流体、空腔流体、眼部流体、血清、乳房乳汁、淋巴流体或其组合。物质可为实体组织,例如生物组织。物质可包含正常的健康组织、疾病组织或健康与疾病组织的混合物。在一些情况下,物质可包含肿瘤。肿瘤可为良性的(非癌症)或恶性的(癌症)。肿瘤的非限制性实例可包括:纤维肉瘤、粘液肉瘤、脂肪肉瘤、软骨肉瘤、骨源性肉瘤、脊索瘤、血管肉瘤、内皮肉瘤、淋巴管肉瘤、淋巴管内皮肉瘤、滑膜瘤、间皮瘤、尤因氏肿瘤(Ewing's),平滑肌肉瘤、横纹肌肉瘤、胃肠系统癌瘤、结肠癌瘤、胰腺癌、乳房癌、泌尿生殖系统癌瘤、卵巢癌、前列腺癌、鳞状细胞癌瘤、基底细胞癌瘤、腺癌瘤、汗腺癌瘤、皮脂腺癌瘤、乳头状癌瘤、乳头状腺癌瘤、囊腺癌瘤、髓样癌瘤、支气管癌瘤、肾细胞癌瘤、肝细胞癌瘤、胆管癌瘤、绒毛膜癌瘤、精原细胞瘤、胚胎性癌瘤、维尔姆斯氏肿瘤(Wilms'tumor)、子宫颈癌、内分泌系统癌瘤、睾丸肿瘤、肺癌瘤、小细胞肺癌瘤、非小细胞肺癌瘤、膀胱癌瘤、上皮癌瘤、神经胶质瘤、星形细胞瘤、髓母细胞瘤、颅咽管瘤、室管膜瘤、松果体瘤、成血管细胞瘤、听神经瘤、少突神经胶质瘤、脑膜瘤、黑色素瘤、成神经细胞瘤、成视网膜细胞瘤或其组合。物质可与各种类型的器官相关。器官的非限制性实例可包括脑、肝、肺、肾、前列腺、卵巢、脾、淋巴结(包括扁桃体)、甲状腺、胰腺、心脏、骨骼肌、肠、喉、食管、胃或其组合。在一些情况下,物质包含多种细胞,包括但不限于:真核细胞、原核细胞、真菌细胞、心脏细胞、肺细胞、肾细胞、肝细胞、胰腺细胞、生殖细胞、干细胞、诱导多能干细胞、胃肠道细胞、血细胞、癌细胞、细菌细胞、从人微生物组样品分离的细菌细胞等。在一些情况下,物质可包含细胞的内容物,诸如单一细胞的内容物或多个细胞的内容物。用于分析单个细胞的方法和系统提供于2014年6月26日提交的例如美国临时专利申请号62/017,558中,该临时专利申请的全部公开内容以全文引用的方式并入本文中。

[0184] 可从各种受试者获得样品。受试者可为活受试者或死受试者。受试者的实例可包括但不限于人、哺乳动物、非人哺乳动物、啮齿动物、两栖动物、爬行动物、犬、猫科动物、牛、马、山羊、绵羊、母鸡、禽类(avine)、小鼠、兔、昆虫、蛞蝓、微生物、细菌、寄生虫或鱼。在一些情况下,受试者可为患有疾病或病症、被怀疑患有疾病或病症或处于发展疾病或病症的风

险之中的患者。在一些情况下,受试者可为孕妇。在某一情况下,受试者可为正常的健康孕妇。在一些情况下,受试者可为处于怀有具有某种先天缺陷的胎儿的风险之中的孕妇。

[0185] 可通过各种方法从受试者获得样品。举例来说,可通过以下方式从受试者获得样品:到达循环系统(例如经由注射器或其他设备静脉内或动脉内地),收集分泌的生物样品(例如唾液、痰尿液、粪便等),手术(例如活检)获得生物样品(例如手术中样品、手术后样品等),擦拭(例如颊拭子、口咽拭子),或移液。

[0186] CNV可与治疗功效相关。举例来说,增加的HER2基因拷贝数可增加晚期非小细胞肺癌中对吉非替尼治疗的反应。参见Cappuzzo F.等(2005) J.Clin.Oncol. 23:5007-5018。高EGFR基因拷贝数可预示对拉帕替尼和卡培他滨的敏感性增加。参见Fabi等(2010) J.Clin.Oncol. 28:15s(2010年ASCO年会)。高EGFR基因拷贝数与对西妥昔单抗和帕尼单抗的敏感性增加相关。

[0187] 拷贝数变异可与癌症患者对某些治疗剂的抗性相关。举例来说,胸苷酸合成酶的扩增可导致转移性结肠直肠癌患者对5-氟尿嘧啶治疗的抗性。参见Wang等(2002) PNAS USA, 第99卷,第16156-61页。

[0188] IX. 计算机控制系统

[0189] 本公开提供计算机系统,其经过编程或以其他方式配置成实现本文所提供的方法,诸如如本文所描述用于核酸测序和基因变异测定、存储参考核酸序列、进行序列分析以及/或者比较样品和参考核酸序列的方法。图22中示出了此类计算机系统的实例。如图22中所示,计算机系统2201包括中央处理单元(CPU,本文中也为“处理器”和“计算机处理器”)2205,其可为单核或多核处理器,或用于平行处理的多个处理器。计算机系统2201还包括存储器或存储位置2210(例如随机存取存储器、只读存储器、闪存存储器)、电子存储单元2215(例如硬盘)、用于与一个或多个其他系统通信的通信接口2220(例如网络适配器)以及外围装置2225,诸如缓存、其他存储器、数据存储和/或电子显示适配器。存储器2210、存储单元2215、接口2220以及外围装置2225通过通信总线(实线)(诸如母板)与CPU 2205通信。存储单元2215可为用于存储数据的数据存储单元(或数据存储库)。计算机系统2201可在通信接口2220的辅助下可操作地耦合至计算机网络(“网络”)2230。网络2230可为因特网、因特网以及/或者外联网,或者与因特网通信的内联网和/或外联网。网络2230在一些情况下为电信和/或数据网络。网络2230可包括一个或多个计算机服务器,所述一个或多个计算机服务器可实现分布式计算,诸如云计算。网络2230在一些情况下在计算机系统2201的辅助下可实现对等网络,所述对等网络可使得耦合至计算机系统2201的装置能够起客户端或服务器的作用。

[0190] CPU 2205可执行机器可读指令的序列,所述机器可读指令的序列可在程序或软件中实现。可将指令存储在存储位置(诸如存储器2210)中。由CPU 2205进行的操作的实例可包括取指令、解码、执行以及写回。

[0191] 存储单元2215可存储文件,诸如驱动器、文库以及保存的程序。存储单元2215可存储用户数据,例如用户偏好和用户程序。计算机系统2201在一些情况下可包括一个或多个额外数据存储单元,所述一个或多个额外数据存储单元在计算机系统2201的外部,诸如位于通过内联网或因特网与计算机系统2201通信的远程服务器上。

[0192] 计算机系统2201可通过网络2230与一个或多个远程计算机系统通信。举例来说,

计算机系统2201可与用户(例如操作者)的远程计算机系统通信。远程计算机系统的实例包括个人计算机(例如便携式PC)、板型或平板PC(例如Apple® iPad、Samsung® Galaxy Tab)、电话、智能手机(例如Apple® iPhone、Android可实现装置、Blackberry®)或个人数字助理。用户可经由网络2230访问计算机系统2201。

[0193] 可通过存储于计算机系统2201的电子存储位置上(诸如在存储器2210或电子存储单元2215上)的机器(例如计算机处理器)可执行的代码来实现如本文所描述的方法。可以软件的形式提供机器可执行或机器可读代码。在使用期间,可由处理器2205执行代码。在一些情况下,可从存储单元2215检索代码并且存储于存储器2210上,以备由处理器2205存取。在一些情况下,可排除电子存储单元2215,并且将机器可执行指令存储于存储器2210上。

[0194] 代码可被预编译并且被配置成与具有适合执行代码的处理器一起使用,或在运行期间被编译。可在编程语言中提供代码,可对所述编程语言加以选择以使得代码能够以预编译或当时编译(as-compiled)方式执行。

[0195] 可在编程中实现本文所提供的诸如计算机系统2201的系统和方法的多个方面。技术的各个方面可被认为是典型地呈在一类机器可读介质上执行或在一种类型的机器可读介质中实现的机器(或处理器)可执行代码和/或相关数据形式的“产品”或“制品”。机器可执行代码可存储于电子存储单元,诸如存储器(例如只读存储器、随机存取存储器、闪存存储器)或硬盘中。“存储”类型介质可包括计算机、处理器等的任何或所有的有形存储器,或其相关模块,诸如各种半导体存储器、磁带驱动器、磁盘驱动器等,其可在软件编程的任何时间提供非暂时存储。软件的全部或部分有时可通过因特网或各种其他电信网络进行通信。此类通信例如可实现软件从一个计算机或处理器加载至另一者中,例如从管理服务器或主机计算机至应用程序服务器的计算机平台中。因此,可承载软件元素的另一类型的介质包括诸如通过有线和光学陆上线路网络并经各种空中链路跨越本地装置之间的物理接口使用的光波、电波以及电磁波。携带此类波的物理元素(诸如有线或无线链路、光学链路等)也可被视为承载软件的介质。如本文中所用,除非限于非暂时有形“存储”介质,否则诸如计算机或机器“可读介质”等术语是指参与提供指令至处理器以执行的任何介质。

[0196] 因此,机器可读介质(诸如计算机可执行代码)可采取许多形式,包括但不限于有形存储介质、载波介质或物理传输介质。非易失性存储介质包括例如光盘或磁盘,诸如图式所示的任何计算机中诸如可用于实现数据库等的任何存储装置等。易失性存储介质包括动态存储器,诸如此类计算机平台的主存储器。有形传输介质包括同轴线缆;铜线以及光纤,包括包含计算机系统内的总线的线。载波传输介质可采取电或电磁信号或声波或光波形式,诸如在射频(RF)和红外线(IR)数据通信期间产生的那些。因此,计算机可读介质的常见形式包括例如:软磁盘、软盘、硬盘、磁带、任何其他磁性介质、CD-ROM、DVD或DVD-ROM、任何其他光学介质、打孔卡纸带、具有孔模式的任何其他物理存储介质、RAM、ROM、PROM和EPROM、FLASH-EPROM、任何其他存储芯片或盒、传送数据或指令的载波、传送此类载波的线缆或链路或计算机可从中读取编程代码和/或数据的任何其他介质。这些形式的计算机可读介质中许多可参与将一个或多个指令的一个或多个序列运送至处理器以执行。

[0197] 计算机系统2201可包括电子显示器2235或与其通信,所述电子显示器包括用于提供例如耦合至计算机系统2201的核酸测序仪器的输出或读出的用户界面(UI)。此类读出可

包括核酸测序读出,诸如包含给定核酸样品的核酸碱基的序列。还可利用此类读出使用UI来展示分析结果。UI的实例包括但不限于图形用户界面(GUI)和基于网络的用户界面。电子显示器2235可为计算机监视器,或电容或电阻式触摸屏。

实施例

[0198] 实施例1:定相变体的识别

[0199] 使用Blue Pippin DNA尺寸选择系统对来自NA12878人细胞系的基因组DNA进行基于尺寸的片段分离以回收长度为约10kb的片段。然后使用微流体分配系统将经过尺寸选择的样品核酸与条形码珠粒共分配于氟化油连续相内的水性微滴中(参见例如2014年4月10日提交并且出于所有目的以全文引用的方式并入本文中的美国临时专利申请号61/977,804),其中水性微滴还包括dNTP、热稳定性DNA聚合酶和用于在微滴内进行扩增的其他试剂以及用于使条形码寡核苷酸从珠粒释放的化学活化剂。对1ng的总输入DNA和2ng的总输入DNA重复此操作。获得作为呈现超过700,000种不同条形码序列的条形码多样性储备文库的子集的条形码珠粒。含有条形码的寡核苷酸包括额外序列组分并且具有以下一般结构:

[0200] 珠粒-P5-BC-R1-N-mer

[0201] 其中P5和R1分别指Illumina连接序列和读段1引物序列,BC表示寡核苷酸的条形码部分,并且N-mer表示用于引导模板核酸的随机10碱基N-mer引导序列。参见例如2014年6月26日提交的美国专利申请号14/316,383,该专利申请的全部公开内容出于所有目的以全文引用的方式并入本文中。

[0202] 在珠粒溶解之后,对微滴进行热循环以允许条形码寡核苷酸针对各微滴内的样品核酸的模板进行的引物延伸。这产生样品核酸的拷贝片段,除了上文所阐述的其他所包括的序列,所述拷贝片段还包括代表起源分区的条形码序列。

[0203] 在对拷贝片段进行条形码标记之后,将包括扩增拷贝片段的微滴的乳液破坏,并且通过额外扩增将额外的测序仪所需组分(例如用于Illumina测序仪的读段2引物序列和P7连接序列)添加至拷贝片段,所述额外扩增将这些序列连接至拷贝片段的另一端。

[0204] 然后在Illumina HiSeq系统上以10X覆盖、20X覆盖以及30X覆盖对测序文库进行测序,并且然后分析所得序列读段和其相关条形码序列。然后将共有共同条形码的邻近映射序列组装成更大重叠群,并且识别单核苷酸多态性并且将其基于其相关条形码和序列映射与单个起始分子相关联,以识别定相SNP。然后基于重叠定相SNP将包括重叠定相SNP的序列组装成定相序列数据的相位区块或推测重叠群。将所得数据与供比较的细胞系的已知单倍型图谱相比较。

[0205] 在至少一种方法中,将一系列杂合变体中的各等位基因指派到两种单倍型中的一者至两者。定义了对数似然函数 $\log P(\text{条形码化读段} | \text{定相指派, 变体})$,其返回所观测到的读段和条形码数据(给定一组变体)的对数似然性以及杂合变体的定相指派(phasing assignment)。对数似然函数的形式源于关于条形码化序列读段数据的两个主要观察:(1)来自一个条形码的读段覆盖单倍体基因组的一小部分,所以一个条形码含有基因组的给定区域中的两种单倍型的读段的概率较小。相反地,基因组局部区域中的一个条形码的读段极有可能来自单一单倍型;(2)所观测到的碱基不同于其所源于的单倍型中的真实碱基的概率是通过由测序仪指派的所观测到的碱基的Phred QV来描述。

[0206] 然后报告了对于给定的条形码化读段和变体集合,使对数似然函数最大化的定相配置。然后通过结构化搜索程序找到了最大似然性评分单倍型配置。首先,使用集束搜索来找到相邻变体(例如约50个变体)的较小区块的最佳定相配置。其次,在区块接合处上以扫描的形式测定诸多个区块的相对定相。此时,找到了总体接近最优的定相配置并且用作进一步优化起始点。然后将单个变体的单倍型指派倒位,以找到对定相局部改善,交换的配置之间的对数似然性的差异提供对定相指派的置信度的评估。最后,将定相配置分解成相位区块,所述相位区块具有高内部校正概率。然后,通过将最佳配置的对数似然性与其中当前SNP右侧的所有SNP均使其单倍型指派倒位的配置相比较来测试是否在各SNP处破坏相位区块。

[0207] 下表提供针对NA 12878基因组获得的定相度量。很明显,从短读段序列数据获得极长相位区块,从而正确识别显著百分比的定相SNP,并且短切换误差或长切换误差极低。

[0208]		10X 覆盖	20X 覆盖	30X 覆盖	30X 集束搜索
	N50 相位区块	193kb	385kb	428kb	489kb
	最长相位区块	2121kb	2514kb	2514kb	3027kb
[0209]	长切换误差	0.0053	0.0021	0.0018	0.0015
	短切换误差	0.004	0.0017	0.0014	0.0012
	定相的 SNP	83%	94%	95%	95.2%

[0210] 其他实验对许多额外样品的SNP进行了定相,包括NA12878trio (NA12878、NA12882以及NA12877)、古吉拉特人(Gujarati) (NA20847)、墨西哥人(NA19662)以及非洲人(NA19701)细胞系样品。实现约1MB的N50相位区块长度,并且定相的SNP大于95%,并且切换误差小于0.3%。相同样品的全外显子组测序(例如其中条形码后进行靶向型向下拉动)显示约90%的基因SNP定相,并且切换误差再次小于0.3%。

[0211] 实施例2:EML-4/ALK基因倒位/易位的识别

[0212] 使用本文所描述的方法和过程来检测所表征的癌细胞系的结构变异。特定而言,NCI-H2228肺癌细胞系已知在其基因组内具有EML4-ALK融合易位。图15中说明了与野生型相比的变异结构。如上部图中所示,在变体结构中,EML-4基因(虽然在同一染色体上)与ALK基因相对隔开或远离,实际上是易位的并且融合至ALK基因(参见例如Choi等, Identification of Novel Isoforms of the EML4-LK Transforming Gene in Non-Small Cell Lung Cancer, J.Cancer Res., 68:4971 (2008年7月))。在易位的同时,EML4基因还是倒位的。图II中进一步说明了易位,与野生型结构相比,其中易位引起EML-4的外显子1-6(以黑框示出)与ALK的外显子20-29(以白框示出)的融合,以及与EML-4的外显子1-19融合的ALK的外显子7-23的融合。

[0213] 为了识别此变异,使用BluePippin®系统(Sage Sciences, Inc.)对来自NCI-H2228细胞系的基因组DNA进行尺寸分离以选择长度为约10kb的片段。

[0214] 然后如上文关于实施例1所描述,将经过尺寸选择的样品核酸与条形码珠粒共分配,扩增并且处理成测序文库,除了在条形码化之后并且在测序之前使用Agilent SureSelect外显子组捕获试剂盒对DNA进行杂交捕获。然后在Illumina HiSeq系统上对测序文库进行测序达到约80X覆盖,并且然后分析所得序列读段和其相关条形码序列。与野生型相比,显然基因组中跨越易位事件的部分之间共有的条形码的数目更高,说明不存在于野生型中的融合组分之间的结构接近性。特定而言,并且如图16A中所示,融合结构显示12个条形码在EML-4外显子1-6与ALK外显子20-29之间以及20个条形码在EML-4外显子7-23与ALK外显子1-19之间的条形码重叠,这类似于杂合细胞系的野生型构建体的重叠条形码。

[0215] 相比之下,如图16B中所示,使用非变体细胞系(NA12878)进行的阴性对照运行仅大体上显示野生型相较于变体构建体的条形码重叠,并且序列覆盖为约140X,并且使用3ng的起始DNA。

[0216] 特定而言,虽然对各个序列区段展示较大数目的全部映射条形码,但通过与展现极高数目的共同或重叠条形码的野生型结构相比较仅在融合结构中看到极小百分比(例如全部映射条形码的少于0.5%)的重叠条形码。因此,跨融合或易位断点的共同映射条形码提供识别那些易位事件的强大基础。

[0217] 还采用了用于SV检测的算法,其首先搜索具有显著条形码相交/重叠的所有基因组基因座对,以高效稀疏矩阵相乘对此搜索进行编码。然后利用合并读段对、分离读段以及条形码数据的概率模型对此第一阶段的候选物进行过滤。NA12878和NA20847上的SV识别使得识别多个大规模缺失和倒位,并且相对于邻近相位区块对其进行定相,从而在上文所描述的核trio中在诸多种遗传模式下显示定相一致性。

[0218] 实施例3:经由CNV筛检来检测增加的狼疮易感性

[0219] 测试患者对狼疮的易感性。从患者抽取血液。使用本文所叙述的技术对无细胞DNA样品进行测序。然后将序列与已知基因组参考序列相比较以测定不同基因的CNV。FCGR3B(CD16细胞表面免疫球蛋白受体)的拷贝数低指示对系统性红斑狼疮的易感性增加。患者被告知任何拷贝数失常和相关风险/疾病。

[0220] 实施例4:经由CNV筛选来检测增加的成神经细胞瘤倾向

[0221] 测试患者的成神经细胞瘤倾向。从患者抽取血液。使用本文所叙述的技术对无细胞DNA样品进行测序。然后将序列与已知基因组参考序列相比较以测定不同基因的CNV。在1q21.1处的CNV指示增加的成神经细胞瘤倾向。患者被告知任何拷贝数失常和相关风险/疾病。

[0222] 实施例5:经由CNV筛选来鉴别诊断肺癌

[0223] 对慢性咳嗽、体重减轻以及呼吸急促患者进行关于肺癌的测试。从患者抽取血液。使用本文所叙述的技术对循环肿瘤细胞(CTC)或无细胞DNA样品进行测序。然后将CTC序列与已知基因组参考序列相比较以测定不同基因的CNV。如果DNA中的EGFR拷贝数高于正常,那么可将患者鉴别诊断为患有非小细胞肺癌(NSCLC)而不是小细胞肺癌。非小细胞肺癌的CTC还具有其他拷贝数变异,所述其他拷贝数变异可使其与小细胞肺癌进一步区分开。视癌症的阶段而定,嘱咐进行手术、化学治疗或放射治疗。

[0224] 小细胞肺癌很多时候比非小细胞肺癌瘤更快速并且广泛地转移(并且因此被区别地划分阶段)。NSCLC通常对化学治疗和/或放射不太敏感,因此,如果被诊断处于早期,手术

为所选治疗,经常使用涉及顺铂的辅助(辅佐)性化学治疗。非小细胞肺癌(NSCLC)患者还可使用靶向治疗,例如ALK抑制剂,诸如克唑替尼。靶向治疗通过干扰为致癌作用和肿瘤生长所需的特定靶向分子,而不是通过简单地干扰所有快速分裂的细胞(例如在传统化学治疗情况下)来阻碍癌细胞生长。

[0225] 实施例6:经由定相来鉴别诊断胎儿非整倍性

[0226] 胎儿非整倍性为染色体数目失常。非整倍性通常引起显著的物理和神经损伤。X染色体数目减少引起特纳氏综合征。21号染色体的拷贝数增加引起唐氏综合征。侵入性测试(诸如羊膜穿刺术或绒毛膜绒毛取样(CVS))可导致流产风险,并且此处使用较无创的测试母体血液的方法。

[0227] 对具有唐氏综合征或特纳氏综合征家族史的怀孕患者进行测试。收集含有胎儿基因材料的母体血液样品。然后如本文所描述将来自不同染色体的核酸与条形码化标签分子一起分离至不同分区中。然后对样品进行测序,并且将各染色体拷贝的数目与正常二倍体染色体上的序列相比较。患者被告知不同染色体的任何拷贝数失常和相关风险/疾病。

[0228] 实施例7:经由定相来检测染色体易位以鉴别诊断伯基特氏淋巴瘤(Burkitt's Lymphoma)

[0229] 伯基特氏淋巴瘤的特征为染色体中的t(8;14)易位。对总体上被诊断为具有淋巴瘤的患者进行关于伯基特氏淋巴瘤的测试。从淋巴结收集肿瘤活检标本。如本文所描述将来自不同染色体的核酸与条形码化标签分子一起分离至不同分区中。然后对样品进行测序,并且与对照DNA样品相比较以检测染色体易位。如果患者被诊断为具有伯基特氏淋巴瘤,那么可能需要比在其他类型的淋巴瘤的情况下更强的化学治疗方案,包括CHOP或R-CHOP方案。CHOP由以下组成:环磷酰胺,一种烷基化剂,其通过与DNA结合并且使得形成交联来损害DNA;羟基佐柔比星(Hydroxydaunorubicin)(也称为多柔比星或阿霉素(Adriamycin)),一种嵌入剂,其通过将自身插入DNA碱基之间来损害DNA;安可平(Oncovin)(长春新碱(vincristine)),其通过与微管蛋白结合来阻止细胞复制;泼尼松(Prednisone)或泼尼松龙(prednisolone),其为皮质类固醇。还可将此方案与单克隆抗体利妥昔单抗组合,因为伯基特氏淋巴瘤是来源于B细胞;此组合被称为R-CHOP。

[0230] 实施例8:通过与亲代基因组比较来定相源于无细胞DNA的胎儿基因组序列

[0231] 收集来自怀孕患者的包含母体DNA的样品和来自胎儿父亲的包含父体DNA的样品。如本文所描述将来自各样品的核酸与分子条形码化标签一起分离至不同分区中。然后对样品进行测序,并且使用所述序列来产生所分配的母体和父体片段中的每一者的推测重叠群。使用推测重叠群来构建母体和父体染色体中的每一者的诸多个部分的单倍型区块。

[0232] 收集含有胎儿基因材料的母体血液样品。对无细胞DNA进行测序以产生母体循环DNA与胎儿循环DNA的序列。将读段与上面所产生的父体和母体相位区块相比较。一些相位区块在减数分裂期间已经历重组。识别与父体相位区块而不与母体相位区块匹配的胎儿材料。在一些情况下,胎儿材料与整个父体相位区块匹配,并且确定胎儿具有父体遗传染色体中的所述父体相位区块。在其他情况下,胎儿材料与一个相位区块的一部分匹配,并且然后与第二相位区块匹配,其中这两个相位区块均位于父体基因组中的同源染色体区域上。确定在此区域发生减数分裂重组事件,确定最有可能的重组点,并且产生作为两个父体相位区块的组合的新颖胎儿相位区块。

[0233] 将循环DNA的序列与母体相位区块相比较。使用母体相位区块中的杂合性位点来确定源于母体的胎儿染色体的最有可能的相位。使用循环DNA序列来测定母体基因组的杂合位点处的拷贝数。特定母体相位区块的拷贝数升高指示胎儿中源于母体的染色体含有升高的相位区块的序列。在一些情况下,类似于在父体情况下所描述,首先同源区域的一个相位区块将出现升高,并且然后同一区域的另一相位区块的一部分将出现升高,表明减数分裂重组已发生。在这些情况下,确定最有可能的重组区域,并且由两个母体相位区块构建新的胎儿相位区块。

[0234] 虽然本文中已示出和描述了本发明的优选实施方案,但对本领域技术人员来说将显而易见的是此类实施方案仅仅是通过举例而提供。本发明不旨在受说明书内所提供的特定实施例限制。虽然已参照上述说明书描述了本发明,但本文中对实施方案的描述和说明并不意在以限制意义来解释。本领域技术人员将会想到许多变化、改变以及替换,而不会脱离本发明。此外,应了解,本发明的所有方面不限于本文所阐述的特定描述、配置或相对比例,其取决于多种条件和变量。应了解,在实践本发明时可采用本文中所描述的本发明实施方案的各种替代方案。因此可以预期的是,本发明还应涵盖任何此类替代、修改、变化或等效物。以下权利要求旨在限定本发明的范围并且从而涵盖这些权利要求和其等效物的范围内的方法和结构。

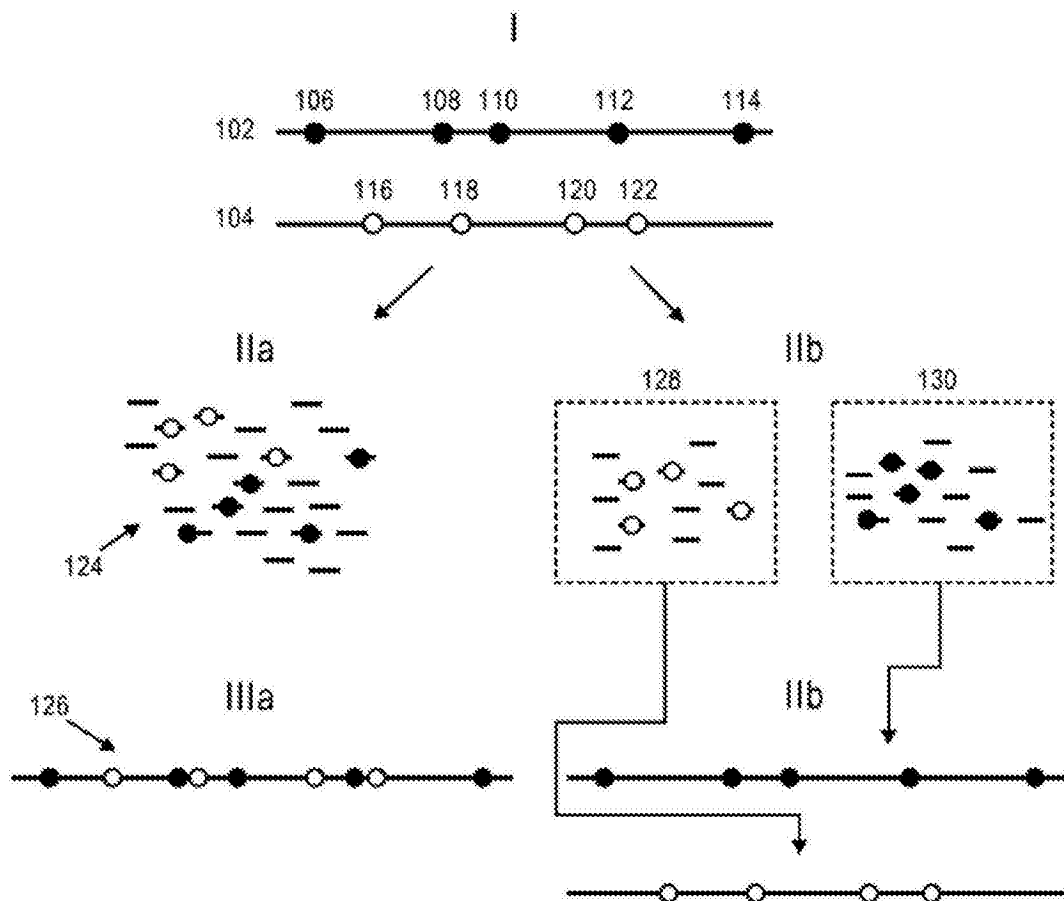


图1

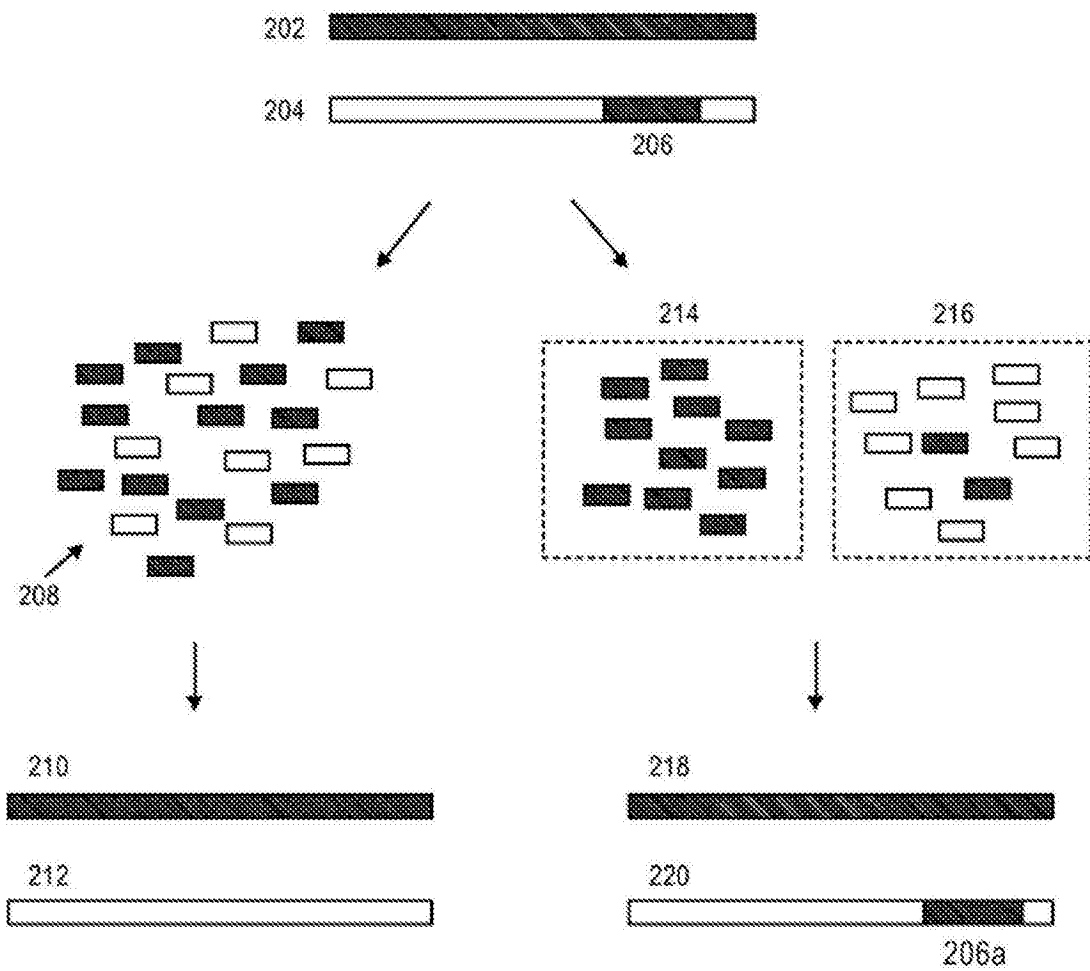


图2

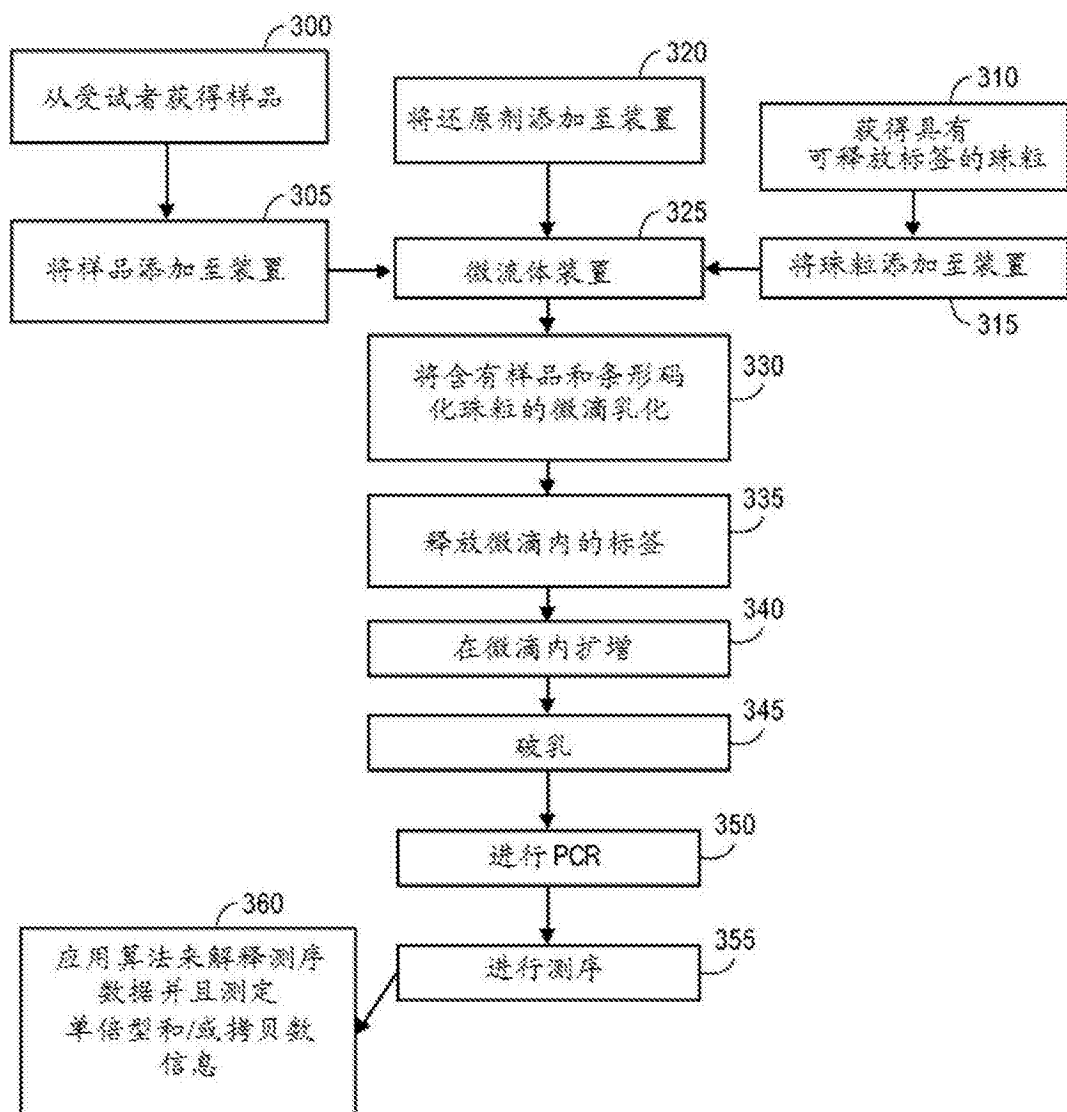


图3

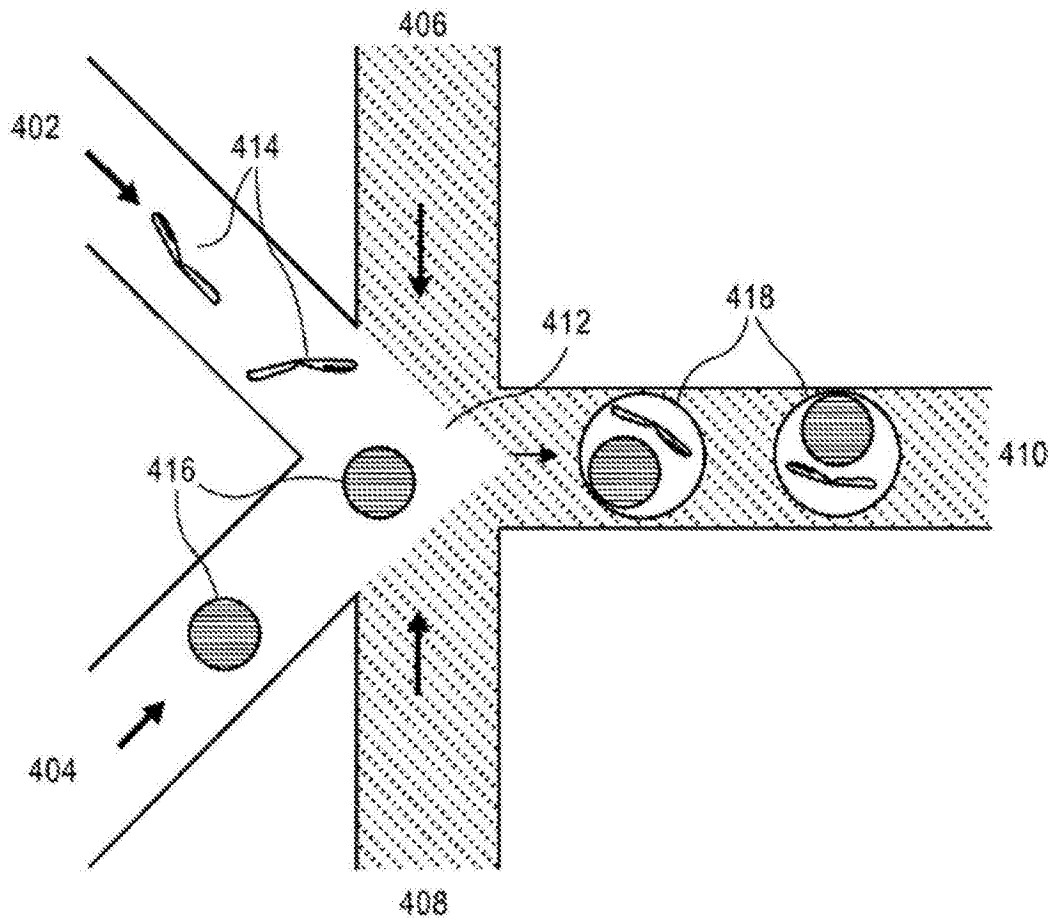


图4

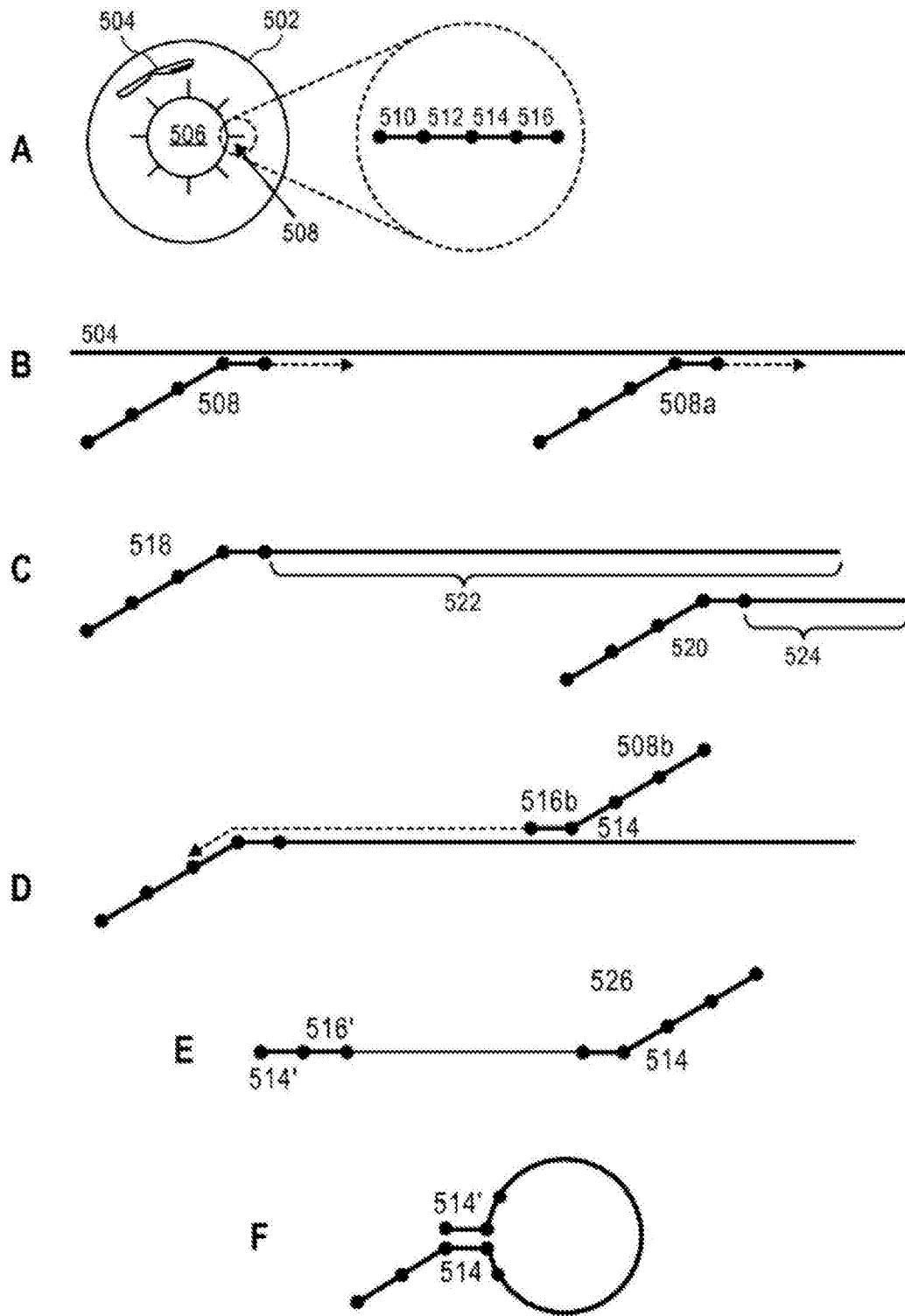


图5

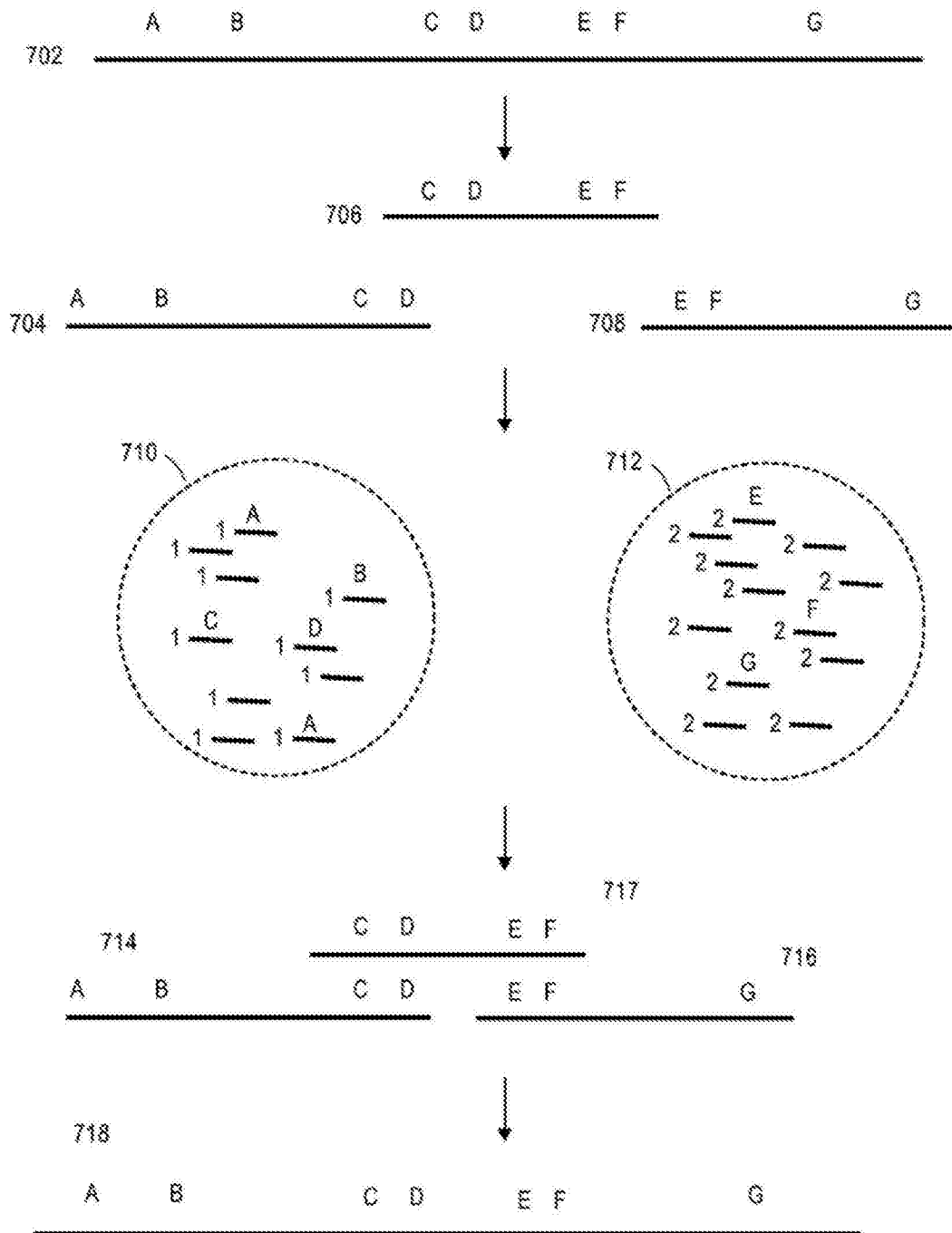


图7

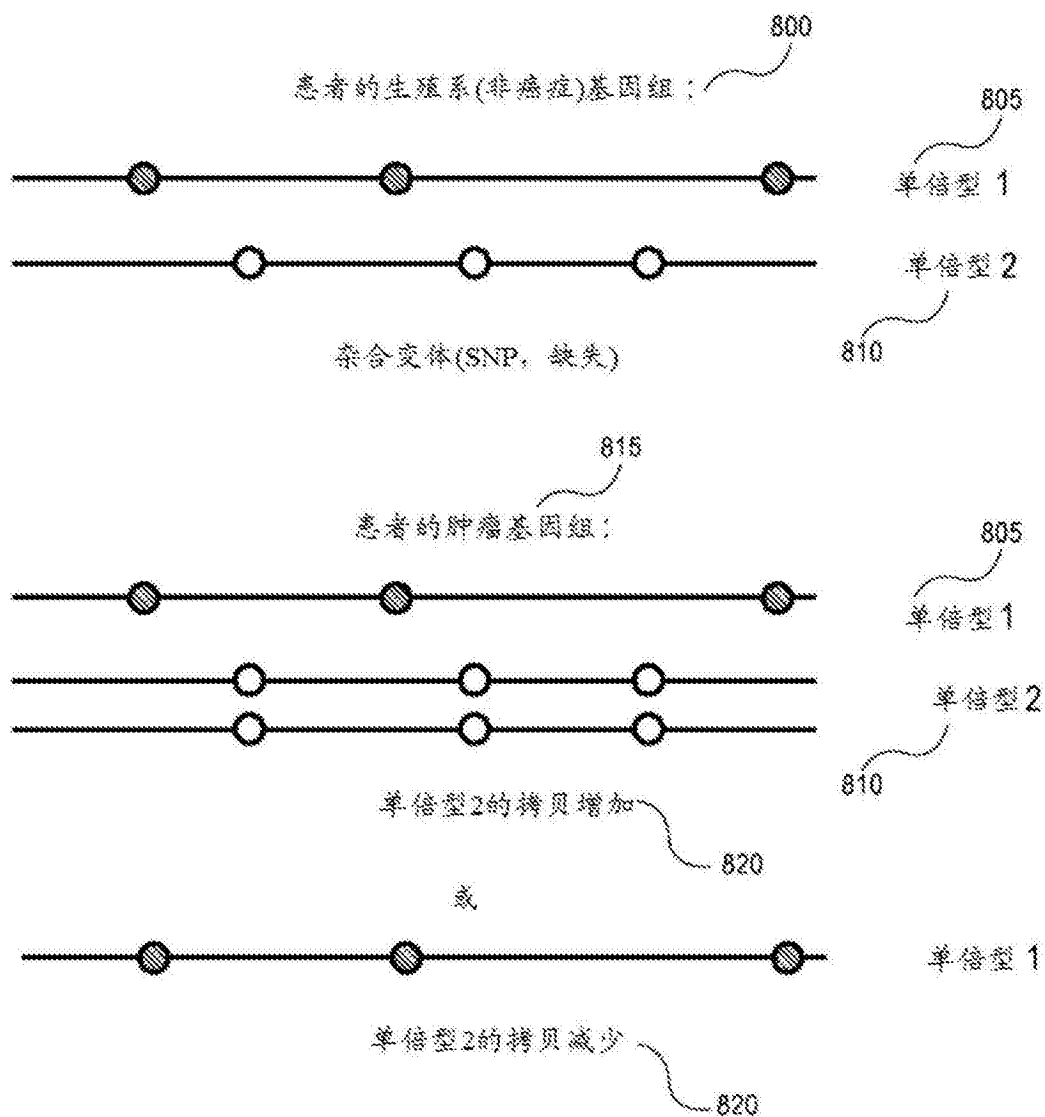


图8

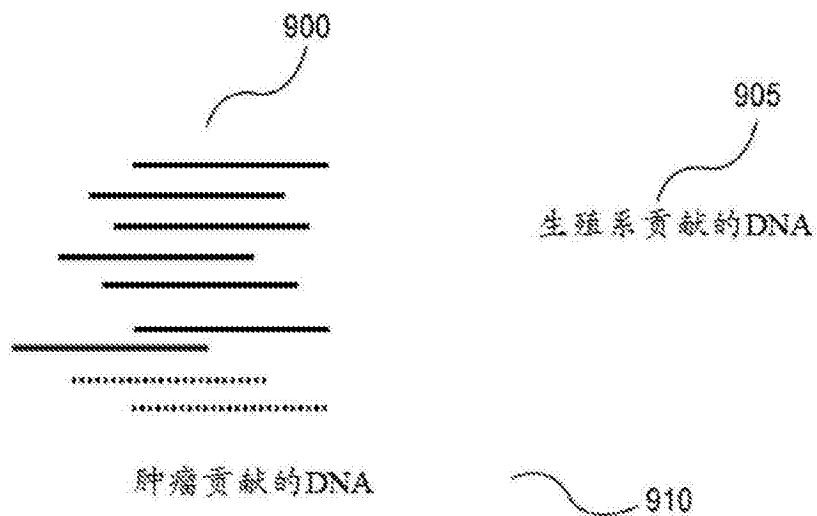


图9A

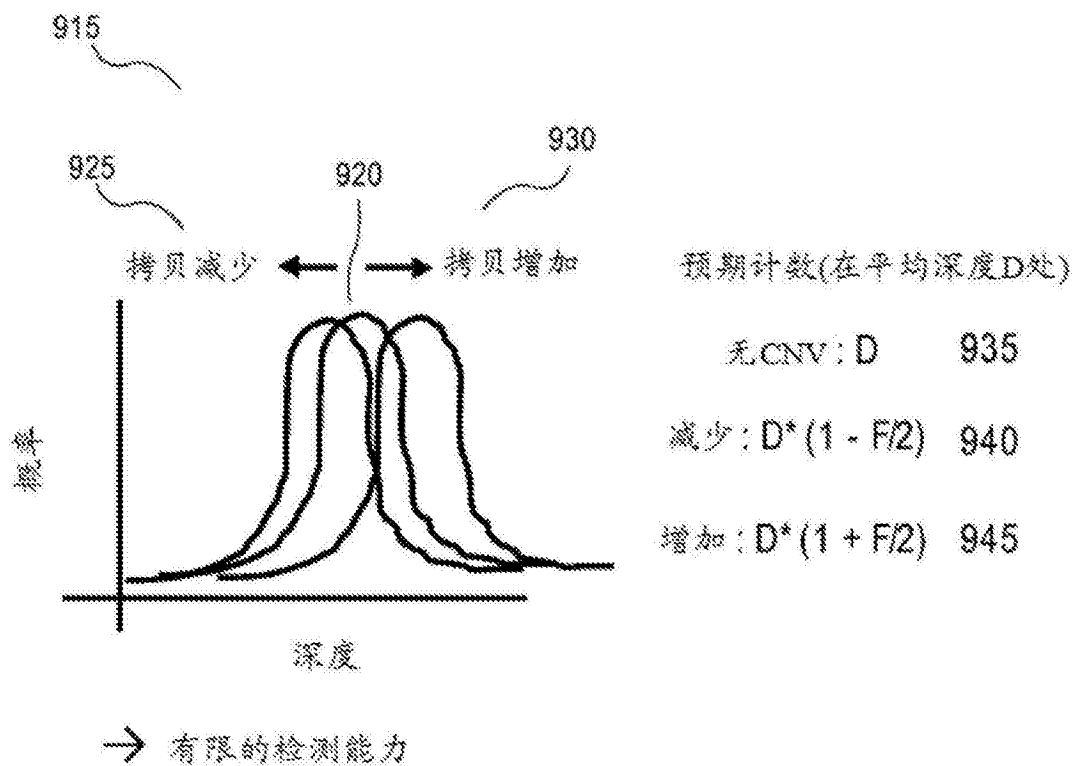


图9B

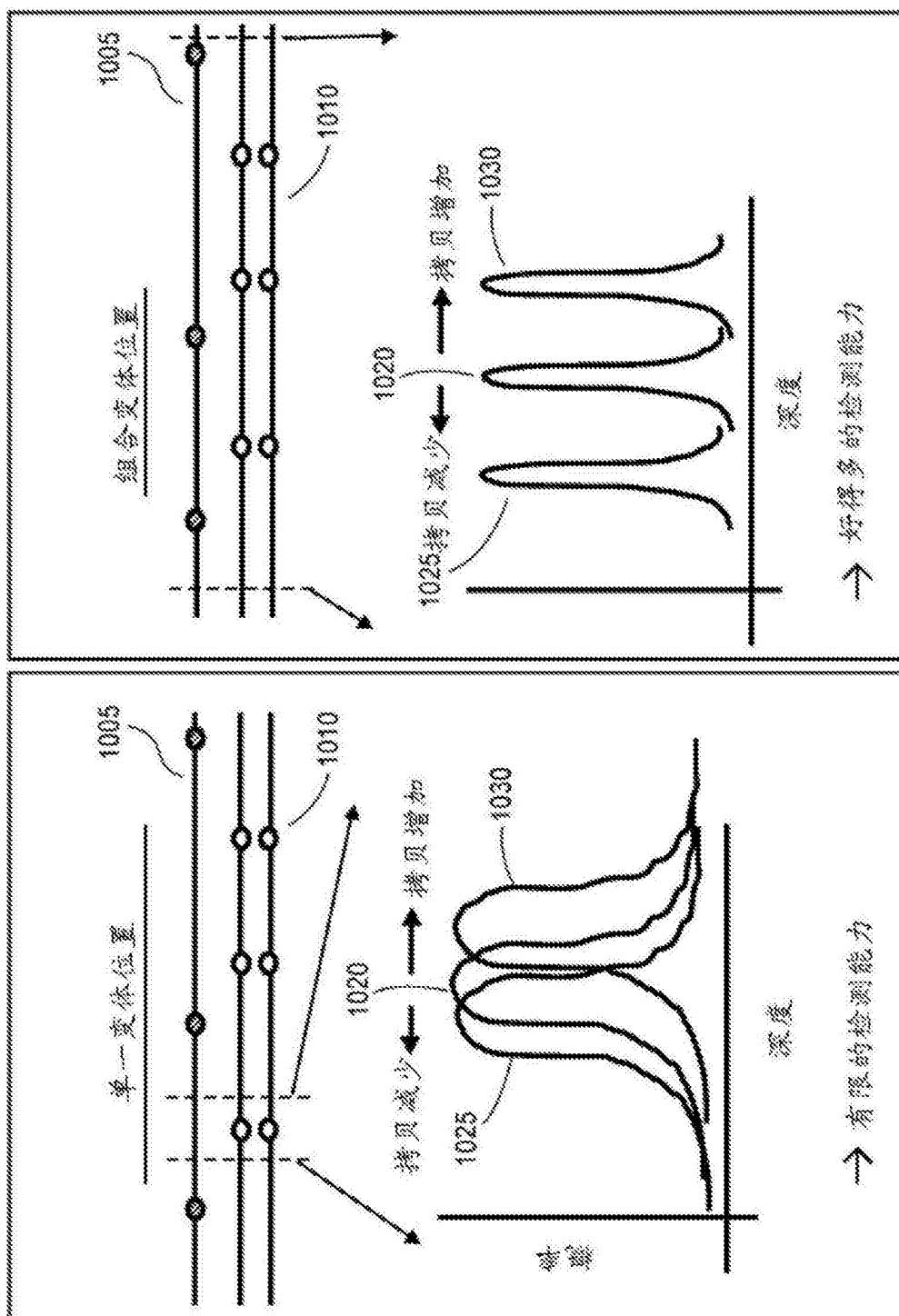


图 10

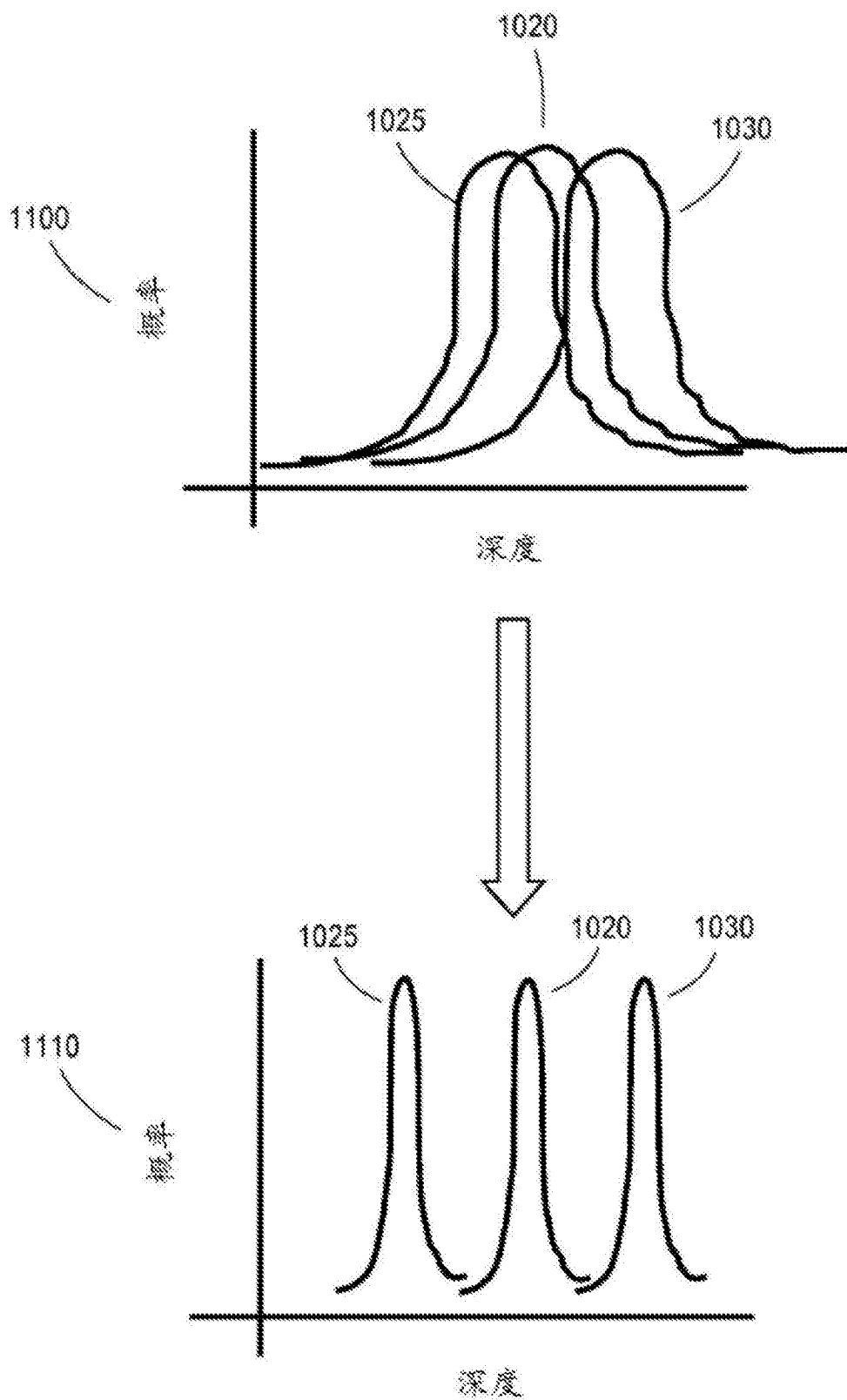


图11

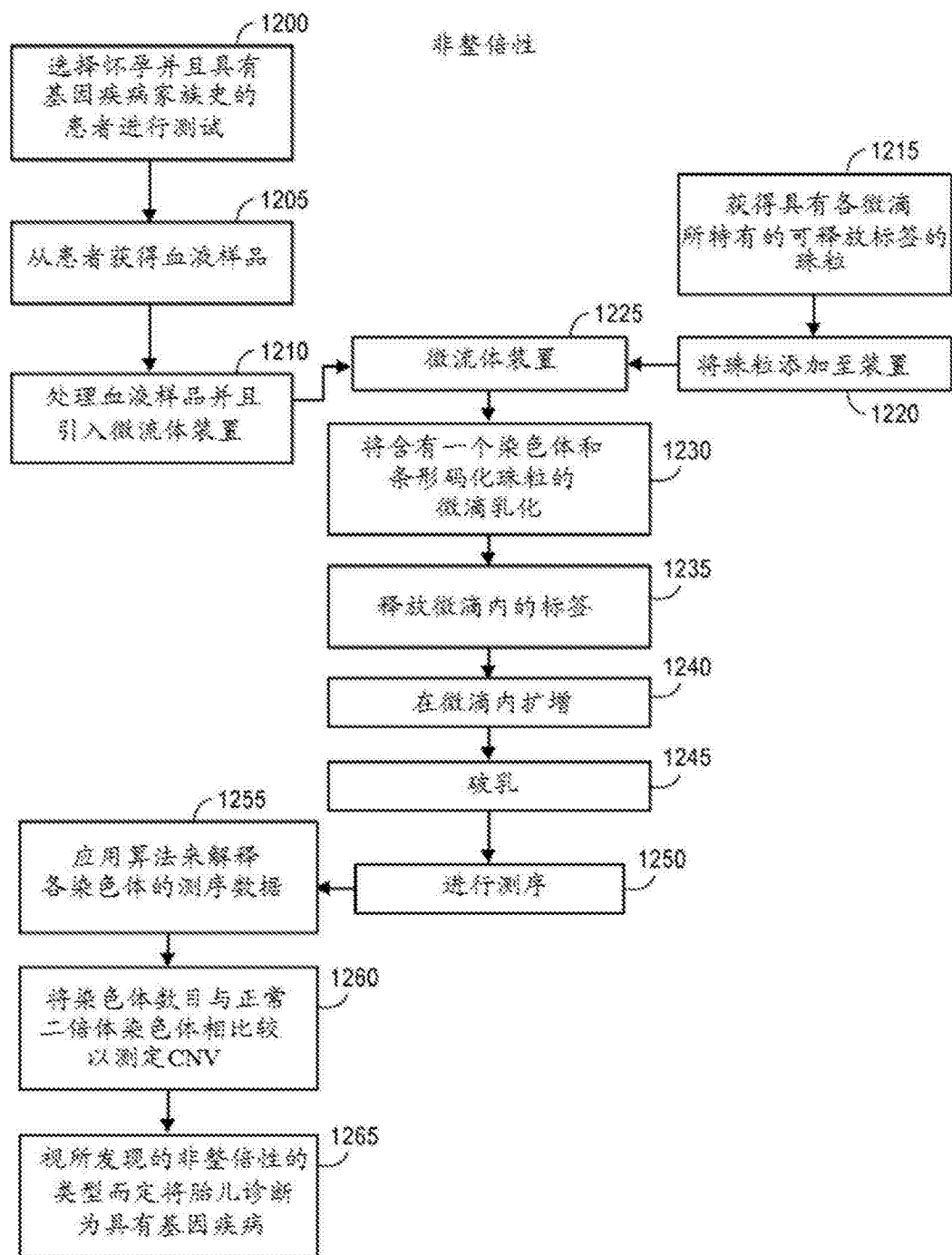


图12

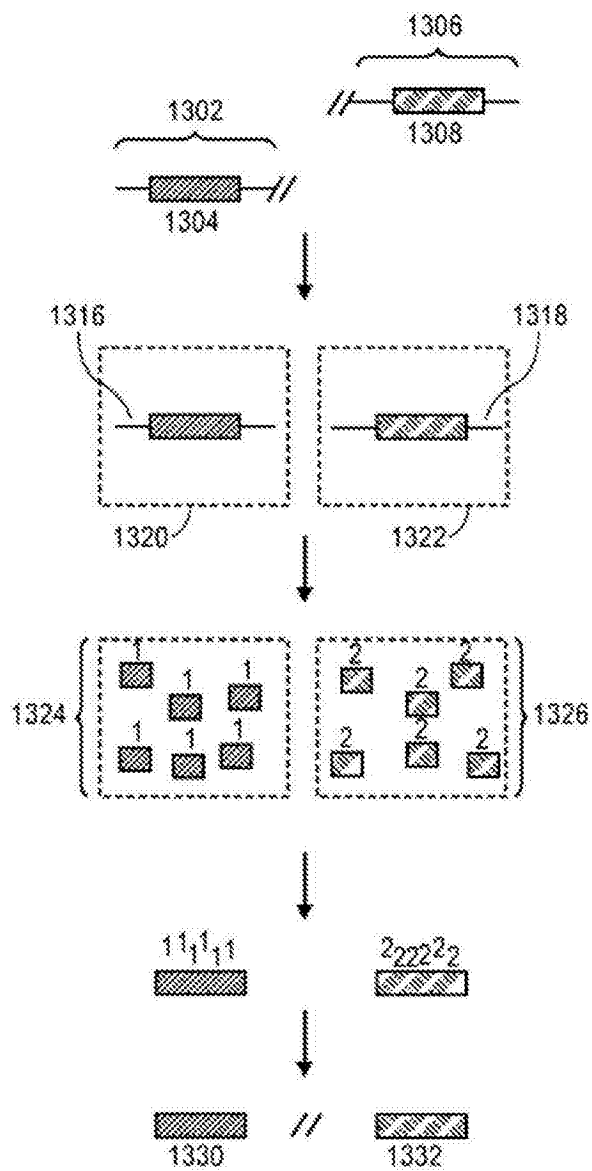


图13A

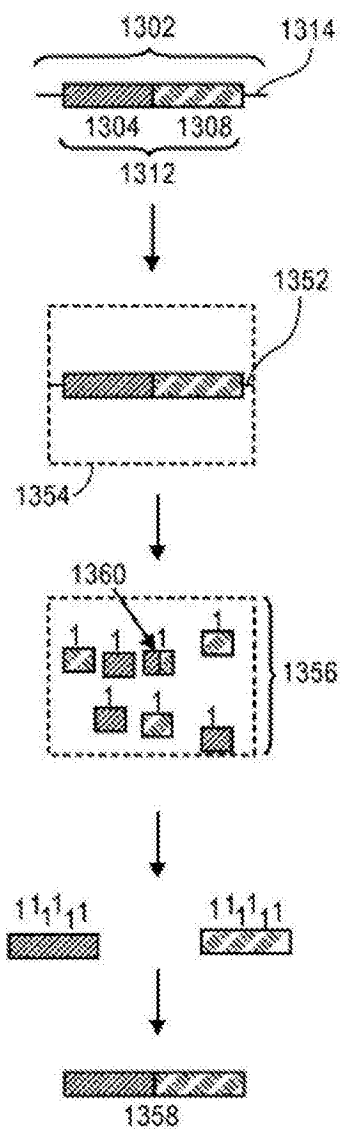


图13B

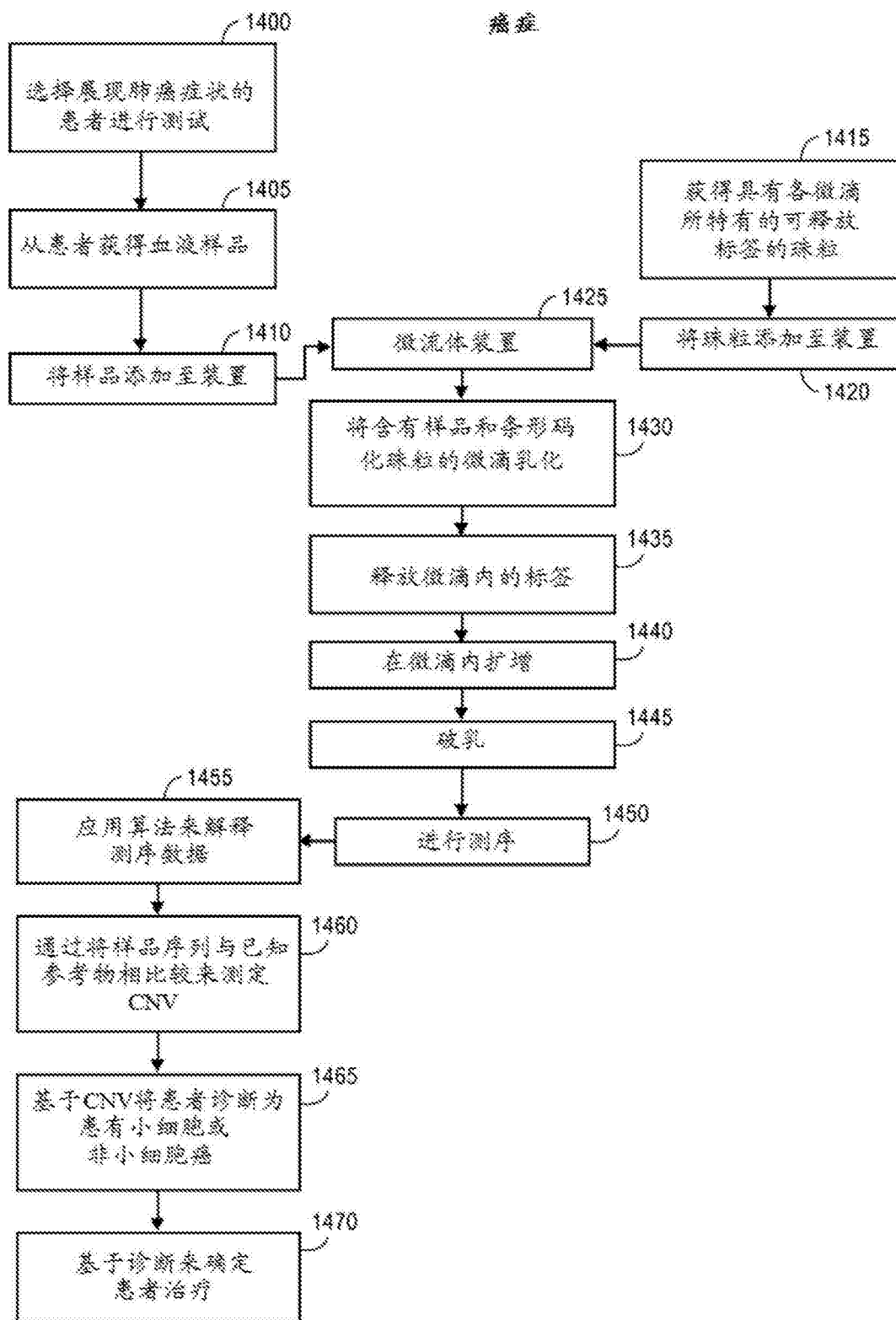


图14

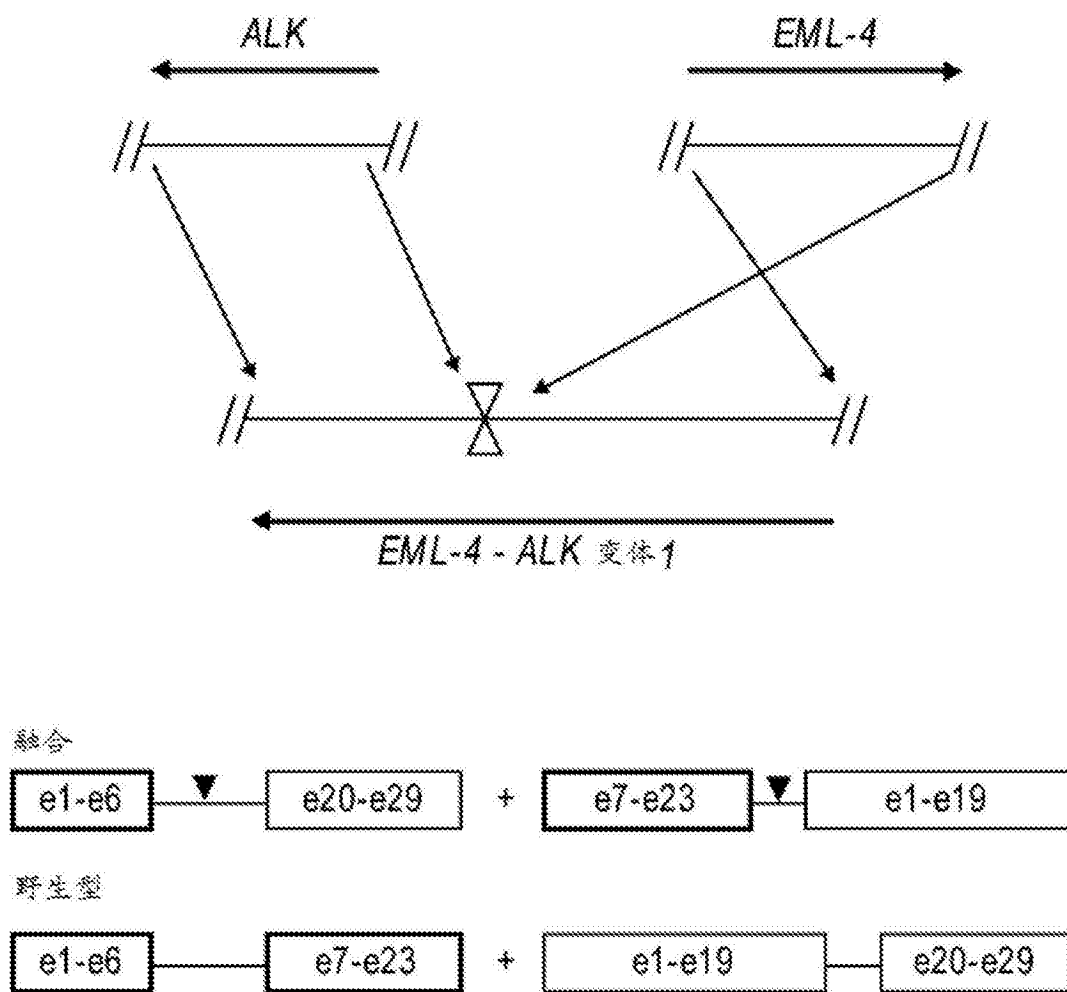


图15

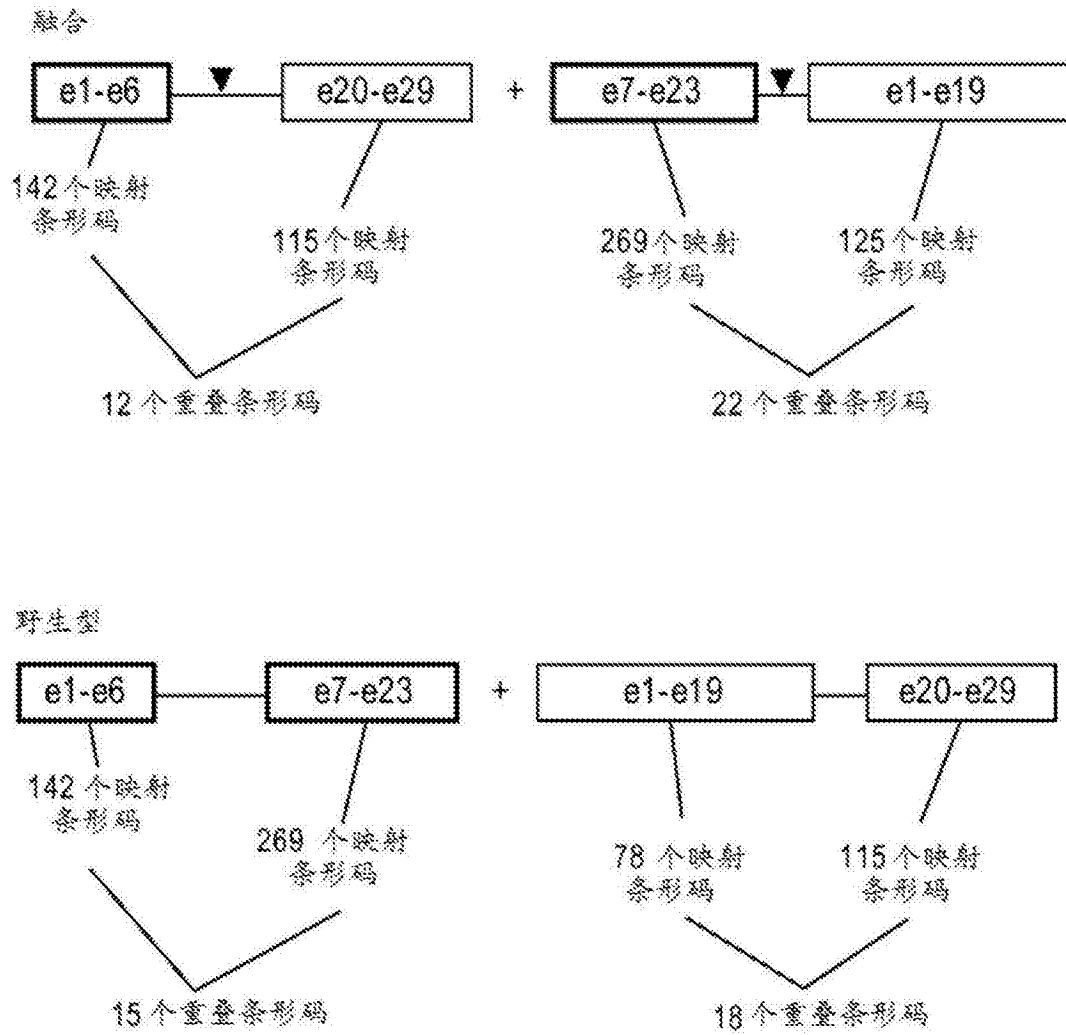


图16A

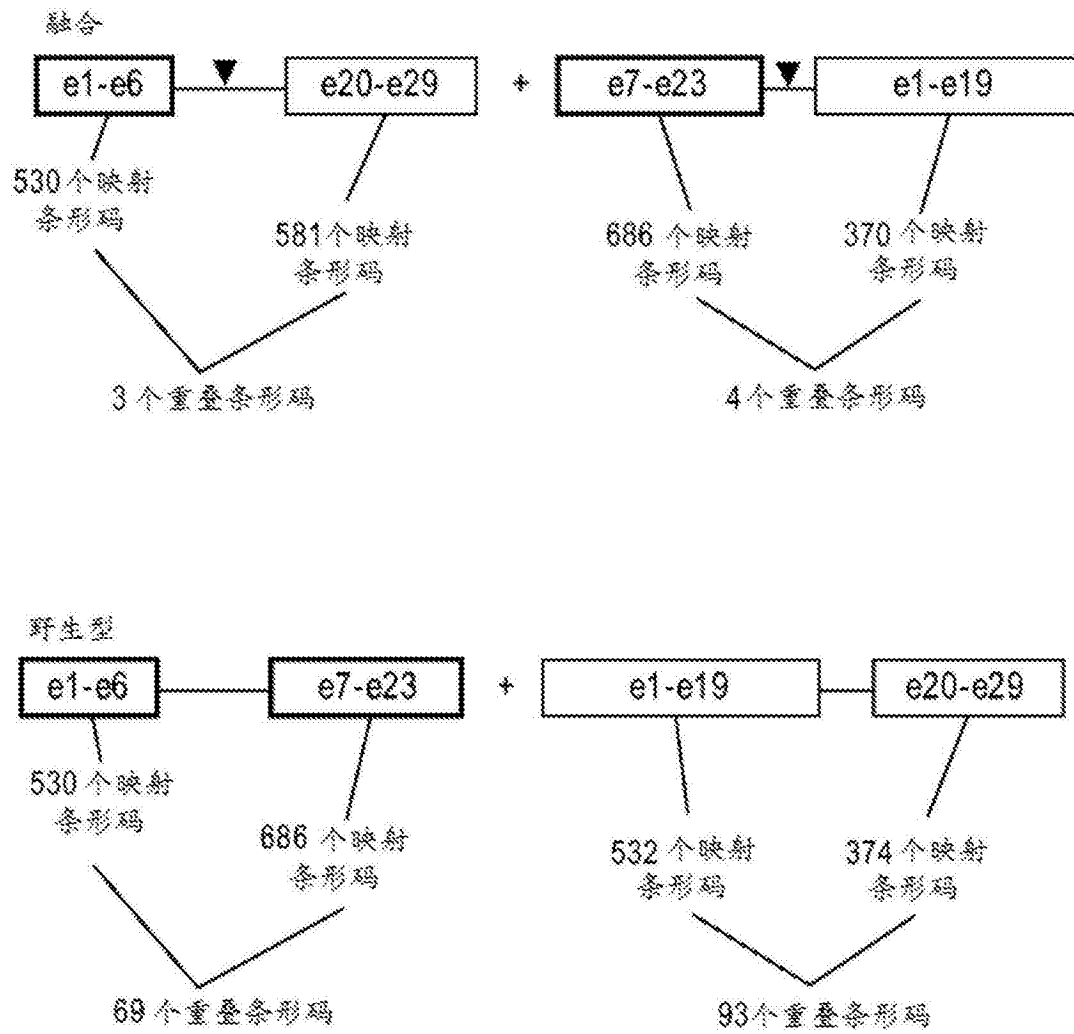


图16B

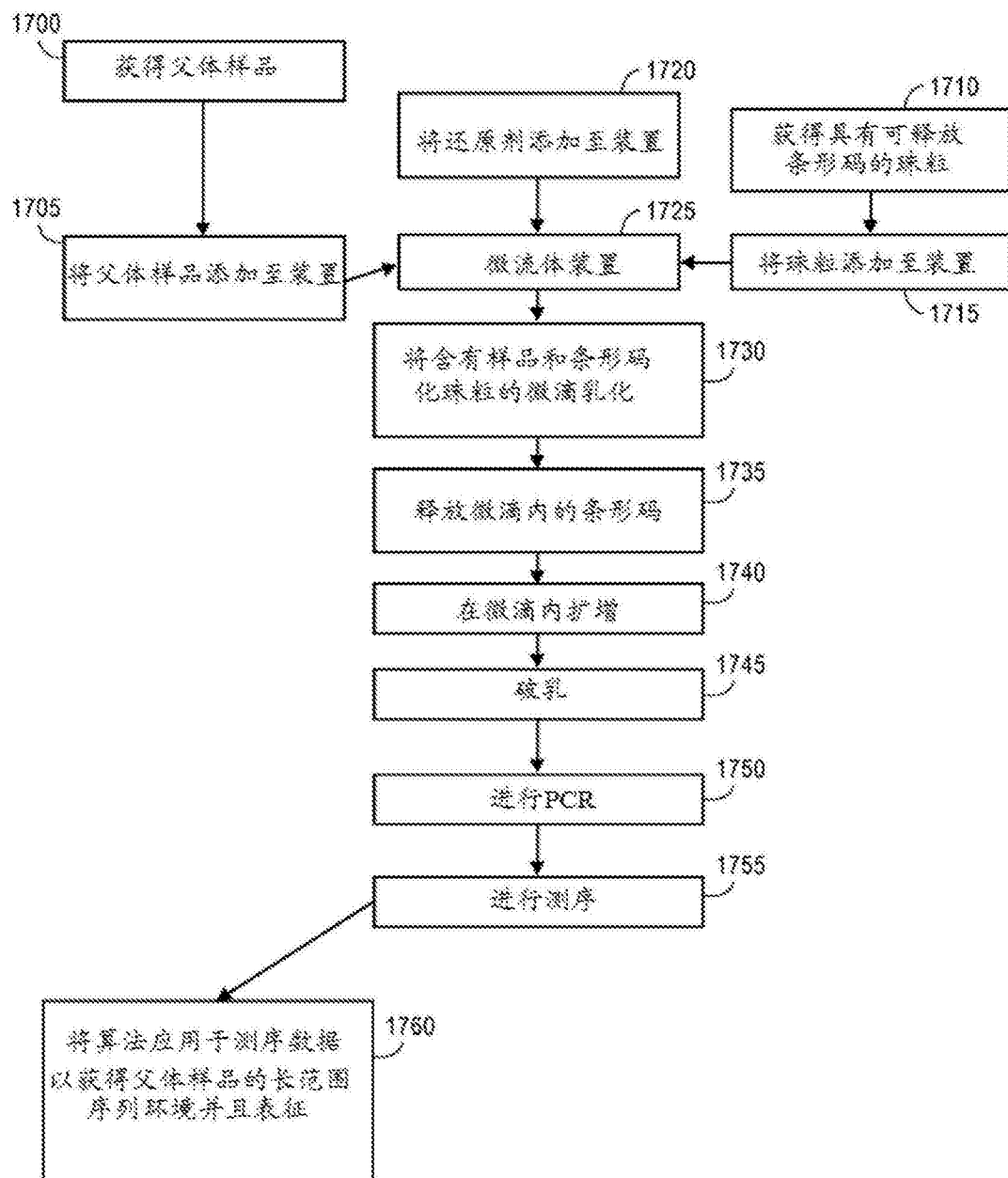


图17

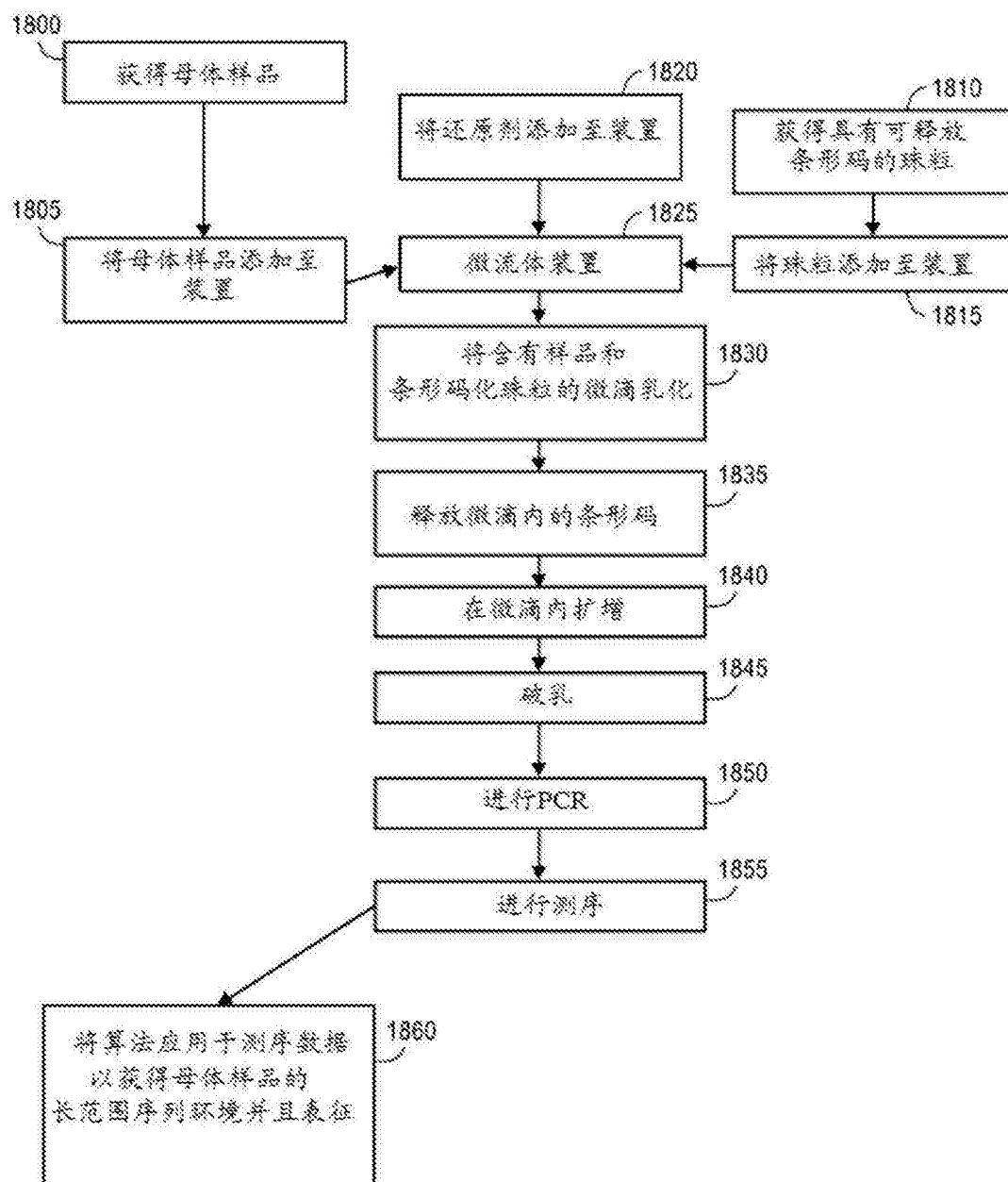


图18

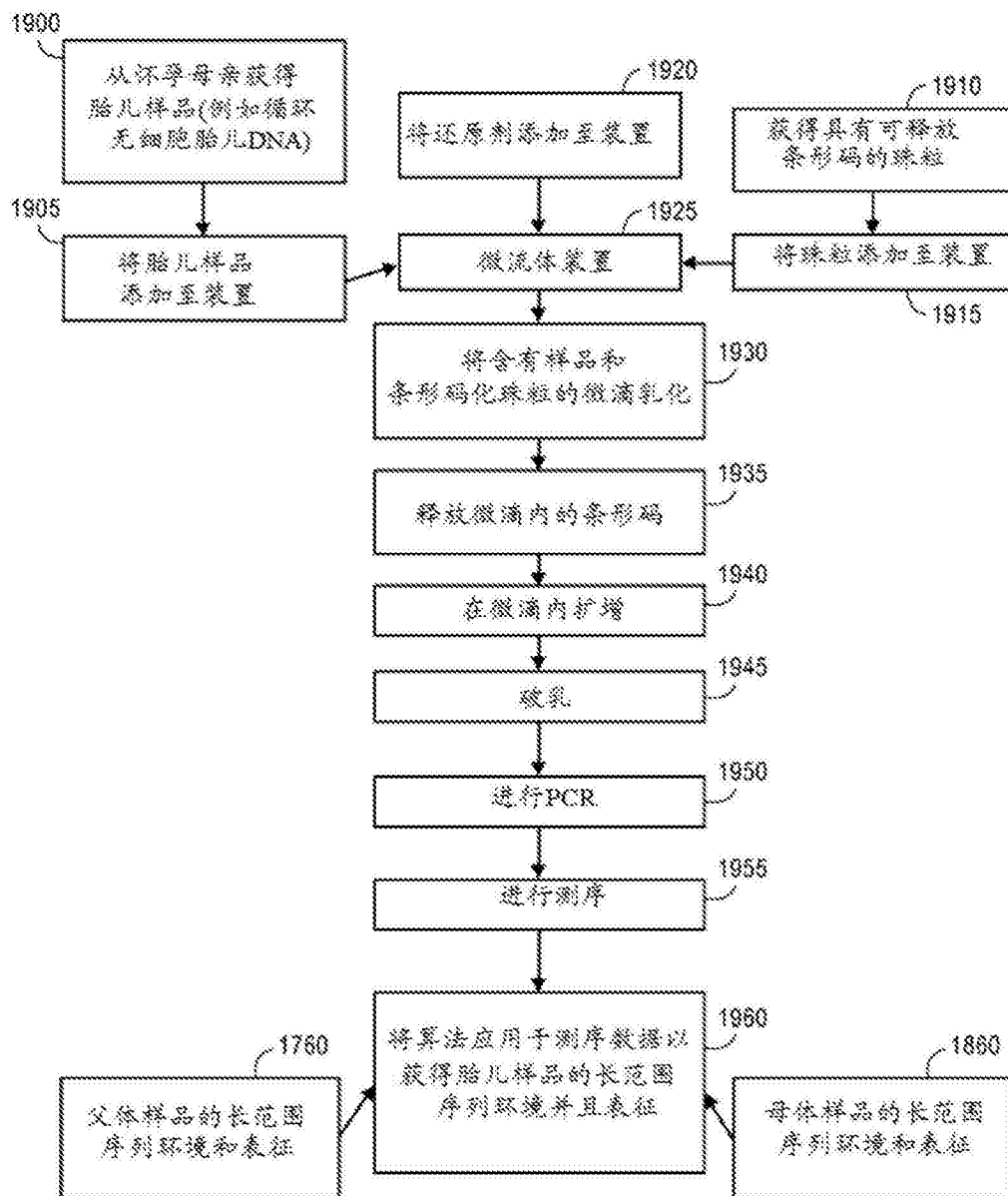


图19

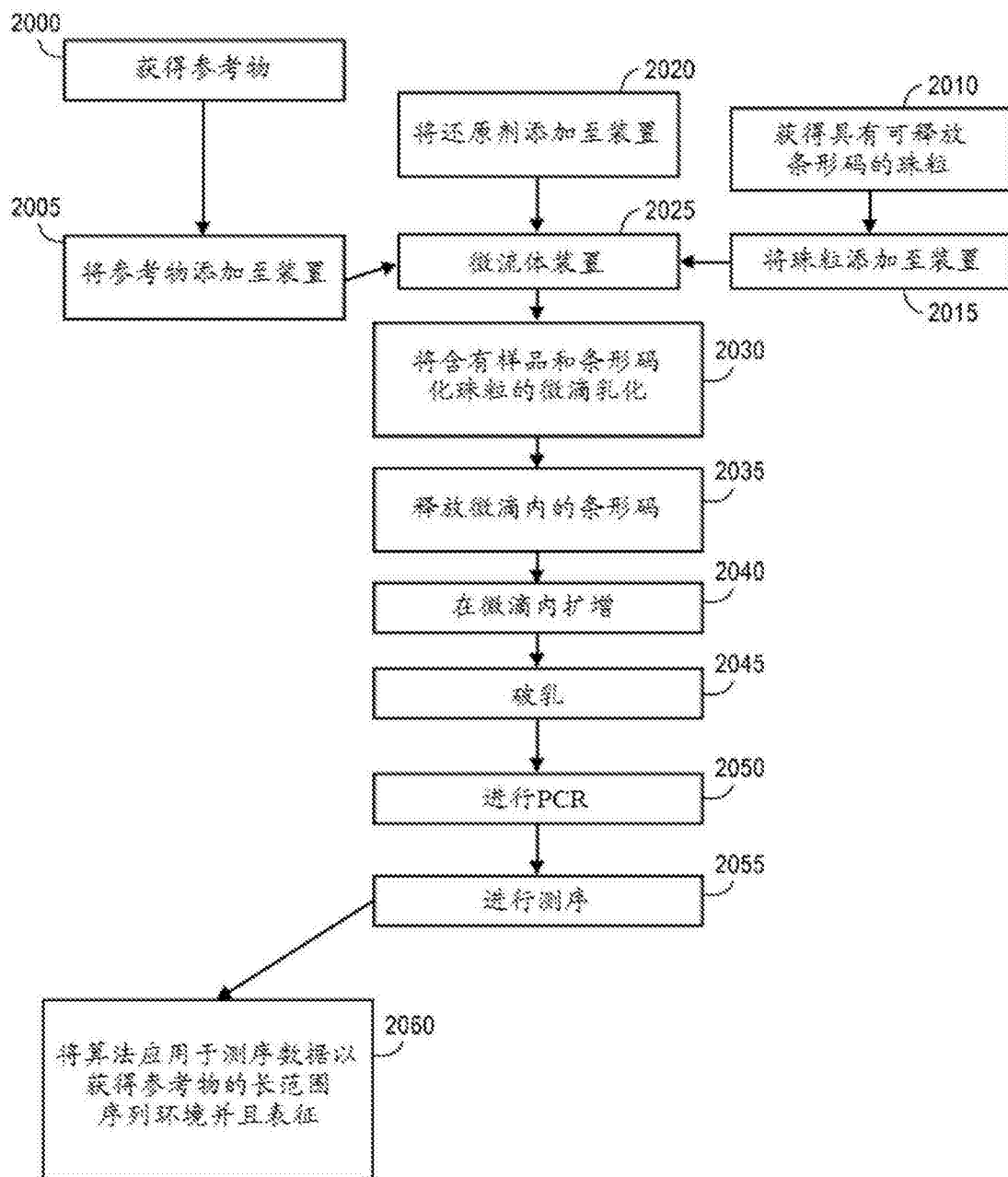


图20

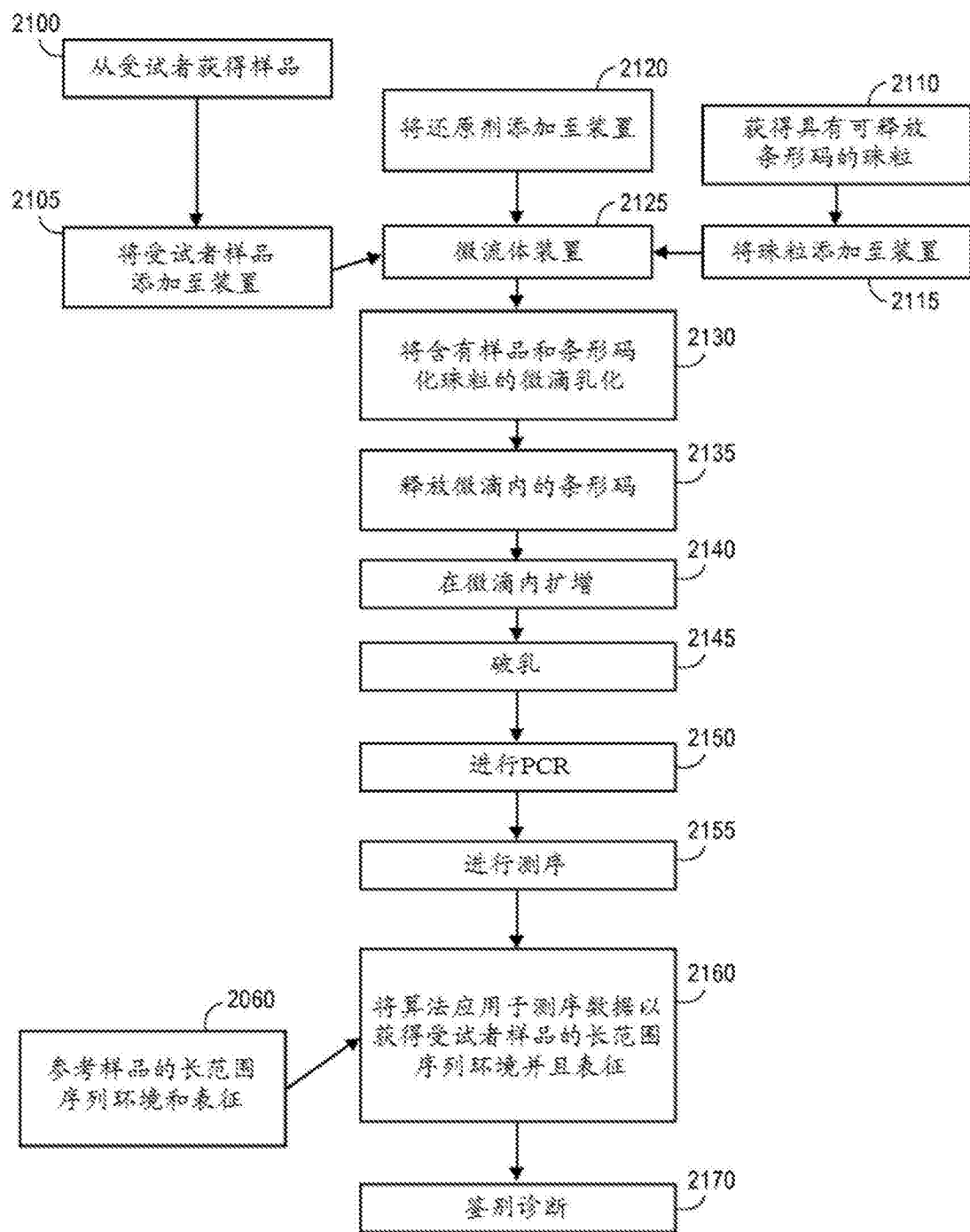


图21

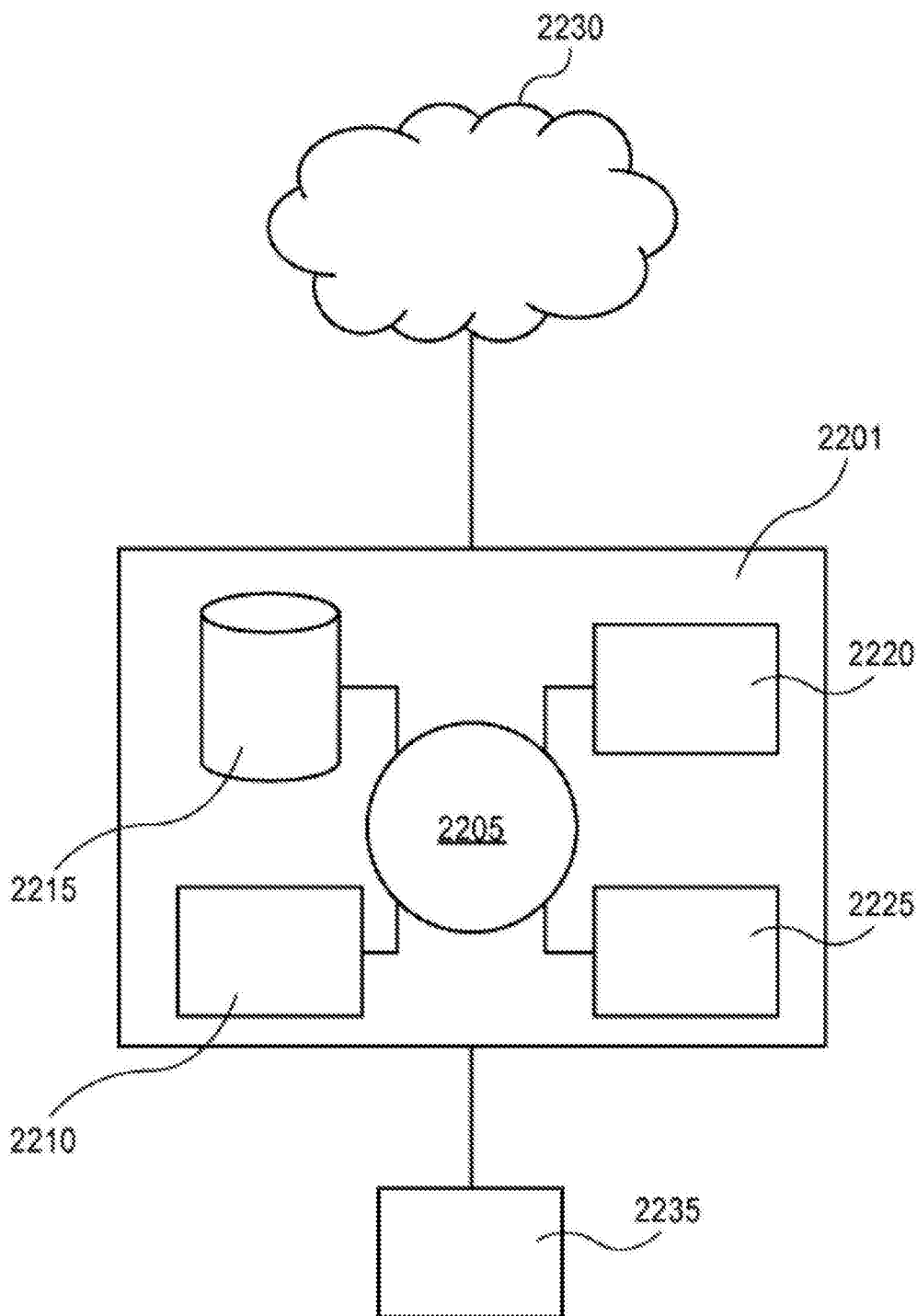


图22