(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105830077 B
(45)授权公告日 2019.07.09

(21)申请号 201480070158.4
(22)申请日 2014.10.21
(65)同一申请的已公布的文献号
申请公布号 CN 105830077 A
(43)申请公布日 2016.08.03
(30)优先权数据
61/893,830 2013.10.21 US
(85)PCT国际申请进入国家阶段日
2016.06.21
(86)PCT国际申请的申请数据
PCT/US2014/061635 2014.10.21
(87)PCT国际申请的公布数据
WO2015/061359 EN 2015.04.30
(73)专利权人 维里纳塔健康公司
地址 美国加利福尼亚州
(72)发明人 达里娅·I·丘多瓦
戴安娜·阿布杜伊瓦
里查德·P·拉瓦
(74)专利代理机构 北京北翔知识产权代理有限公司 11285
代理人 张广育 孙占华
(51)Int.Cl.
G16B 25/10(2019.01)
G16B 30/10(2019.01)
(56)对比文件
WO 2013109981 A1,2013.07.25,
CN 102791881 A,2012.11.21,
CN 101849236 A,2010.09.29,
CN 103003447 A,2013.03.27,
CN 102985561 A,2013.03.20,
Alberto Magi etc..Read count approach for DNA copy number variants detection.《bioinformatics》.2012,第28卷(第4期),
审查员 朱哲

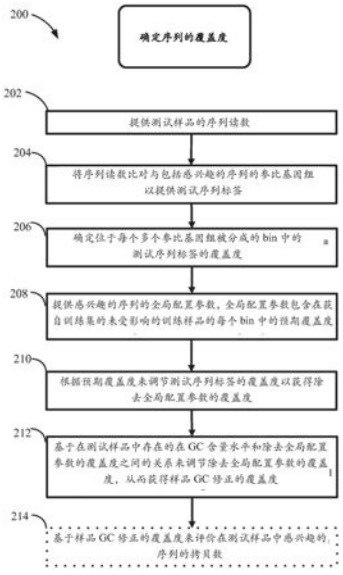
权利要求书6页 说明书66页 附图19页

(54)发明名称

用于在确定拷贝数变异中改善检测的灵敏度的方法

(57)摘要

本发明披露了用于确定已知或疑似与各种各样的医学状况相关的拷贝数变异(CNV)的方法。在一些实施方式中,提供了利用包含母体和胎儿无细胞DNA的母体样品来确定胎儿的拷贝数变异(CNV)的方法。在一些实施方式中,提供了用于确定已知或疑似与各种各样的医学状况相关的CNV的方法。本文披露的一些实施方式提供了通过除去样品中GC含量偏差来改善序列数据分析的灵敏度和/或特异性的方法。在一些实施方式中,样品中GC含量偏差的除去是基于针对通用于不受影响的训练样品的系统性变化修正的序列数据。还披露了用于感兴趣的序列的CNV的评价的系统和计算机程序产品。



CN 105830077 B

1.一种用包括一个或多个处理器和系统存储器的计算机系统来实施的用于评价测试样品中的感兴趣的核酸序列的拷贝数的方法,所述方法包括:

(a)在所述计算机系统中提供通过核酸序列测定仪由所述测试样品获得的序列读数,所述测试样品包含来自一个或多个基因组的核酸分子;

(b)通过所述计算机系统,比对所述测试样品与包含感兴趣的核酸序列的参比基因组的序列读数,从而提供测试序列标签;

(c)通过所述计算机系统,确定位于多个bin中的所述测试序列标签的多个覆盖度,其中所述参比基因组被分成所述多个bin,并且其中所述覆盖度表示bin中序列标签的丰度;

(d)通过所述计算机系统,提供所述感兴趣的核酸序列的全局配置参数,其中所述全局配置参数包含所述多个bin中的多个预期覆盖度,并且其中所述多个预期覆盖度获自未受影响的训练样品的训练集,所述训练样品包含以与所述测试样品基本相同的方式进行测序和比对的核酸分子,所述多个预期覆盖度呈现在bin之间的变化;

(e)通过所述计算机系统,利用每个bin中的预期覆盖度来分别调节每个bin中的所述测试序列标签的覆盖度,从而获得所述感兴趣的核酸序列的全局配置参数修正的覆盖度;

(f)通过所述计算机系统,并基于GC含量水平与所述全局配置参数修正的覆盖度之间的关系,来调节全局配置参数修正的覆盖度,从而获得所述感兴趣的核酸序列的样品-GC-修正的覆盖度,以及

(g)通过所述计算机系统,并基于所述样品-GC-修正的覆盖度,来评价所述测试样品中所述感兴趣的核酸序列的拷贝数,其中所述样品-GC-修正的覆盖度针对用于确定所述感兴趣的核酸序列的拷贝数,改善信号水平和/或降低噪声水平。

2.根据权利要求1所述的方法,进一步包括,在提供序列读数之前,利用测序仪对来自所述测试样品的核酸进行测序,从而产生所述序列读数。

3.根据权利要求2所述的方法,进一步包括,在对所述核酸进行测序之前,使标志物核酸与所述测试样品结合。

4.根据权利要求3所述的方法,其中,所述标志物核酸选自由天然存在的脱氧核糖核酸、天然存在的核糖核酸、肽核酸(PNA)、吗啉代核酸、锁核酸、二醇核酸、苏糖核酸,和它们的任意组合组成的组。

5.根据权利要求1所述的方法,其中,所述序列读数获自孕妇的无细胞DNA和由所述孕妇携带的胎儿的无细胞DNA的序列。

6.根据权利要求1所述的方法,进一步包括施加序列掩码,所述序列掩码排除在掩蔽bin中的考虑覆盖度。

7.根据权利要求6所述的方法,其中,所述序列掩码通过包括以下步骤的方法获得:

在所述计算机系统中提供训练集,所述训练集包含来自多个未受影响的训练样品的序列读数;

通过所述计算机系统,比对所述训练集与所述参比基因组的序列读数,从而提供所述训练样品的训练序列标签;

通过所述计算机系统,将所述参比基因组分成多个bin;

通过所述计算机系统,针对每个训练样品确定每个bin中训练序列标签的覆盖度;以及

通过所述计算机系统,产生包含未掩蔽和掩蔽bin的序列掩码,其中每个掩蔽bin具有

超过掩蔽阈值的分布指数，所述分布指数与所述训练样品的覆盖度的分布有关。

8.根据权利要求7所述的方法，进一步包括，在产生序列掩码之前，根据每个bin中的预期覆盖度来调节所述训练序列标签的覆盖度，从而获得所述bin中所述训练序列标签的全局配置参数修正的覆盖度，所述全局配置参数修正的覆盖度然后用来产生序列掩码。

9.根据权利要求7所述的方法，其中，所述分布指数在数学上与所述训练样品的覆盖度的方差有关。

10.根据权利要求9所述的方法，其中，所述分布指数是变异系数。

11.根据权利要求6所述的方法，其中，所述感兴趣的核酸序列上的掩蔽bin具有第一掩蔽阈值并且归一化序列上的掩蔽bin具有第二掩蔽阈值。

12.根据权利要求11所述的方法，其中，所述第一掩蔽阈值和所述第二掩蔽阈值的组合提供序列掩码，所述序列掩码导致未受影响的样品中的包括所述感兴趣的序列的区域上比利用其他阈值获得的序列掩码更低的覆盖度变异。

13.根据权利要求6所述的方法，其中，所述序列掩码包含由跨整个所述bin内的训练样品的映射质量得分的分布所定义的掩蔽bin和未掩蔽bin，所述映射质量得分来源于多个未受影响的训练样品与所述参比基因组的比对序列读数。

14.根据权利要求1-13中任一项所述的方法，其中，在操作(g)中评价所述测试样品中所述感兴趣的核酸序列的拷贝数包括利用归一化序列的覆盖度信息，针对所述测试样品，来计算所述感兴趣的核酸序列的序列剂量。

15.根据权利要求14所述的方法，其中，计算所述序列剂量包括所述感兴趣的核酸序列中的所述测试序列标签的样品-GC-修正的覆盖度除以归一化序列中的所述测试序列标签的样品-GC-修正的覆盖度。

16.根据权利要求15所述的方法，其中，所述归一化序列包含一个或多个鲁棒常染色体序列或它们的片段。

17.根据权利要求1-13中任一项所述的方法，其中，在操作(g)中评价所述测试样品中所述感兴趣的核酸序列的拷贝数包括利用归一化序列的覆盖度信息，针对所述测试样品，计算所述感兴趣的核酸序列的归一化染色体值或归一化片段值。

18.根据权利要求1-13中任一项所述的方法，其中，所述测试样品包含来自两个不同基因组的核酸的混合物。

19.根据权利要求18所述的方法，其中，所述核酸包含无细胞DNA分子。

20.根据权利要求1-13中任一项所述的方法，其中，所述测试样品包含胎儿和母体无细胞核酸。

21.根据权利要求1-13中任一项所述的方法，其中，所述测试样品包含来自两个或更多个胎儿的胎儿无细胞核酸。

22.根据权利要求1-13中任一项所述的方法，其中，所述测试样品包含来自相同受试者的癌细胞和未受影响的细胞的核酸。

23.根据权利要求1-13中任一项所述的方法，其中，所述评价所述测试样品中所述感兴趣的核酸序列的拷贝数包括确定完全或部分胎儿非整倍性的存在或不存在。

24.根据权利要求1-13中任一项所述的方法，进一步包括，考虑到拷贝数变异的评价，在操作(f)之后，除去样品-GC-修正的覆盖度的离群bin。

25.根据权利要求24所述的方法，其中，所述离群bin包含其中位数样品-GC-修正的覆盖度离所有所述bin的中位数是大于约1中位数绝对偏差的bin。

26.根据权利要求1-13中任一项所述的方法，其中，每个bin中的预期覆盖度包含训练样品的覆盖度的中位数或平均值，并且其中在操作(e)中调节所述测试序列标签的覆盖度包括每个bin的所述测试序列标签的覆盖度除以来自所述相应bin的训练样品的覆盖度的中位数或平均值。

27.根据权利要求1-13中任一项所述的方法，其中，在操作(e)中调节所述测试序列标签的覆盖度包括：(i)在一个或多个鲁棒染色体或区域中的多个bin中获得在所述测试序列标签的覆盖度和所述预期覆盖度之间的关系，以及(ii)将所述数学关系应用于所述感兴趣的序列中的bin，以获得所述全局配置参数修正的覆盖度。

28.根据权利要求27所述的方法，其中

通过线性回归来获得(i)中的关系：

y_a＝截距+斜率*gwp_a

其中y_a是在一个或多个鲁棒染色体或区域中所述测试样品的bin a的覆盖度，并且gwp_a是针对未受影响的训练样品，bin a的全局配置参数；以及

在(ii)中获得所述全局配置参数修正的覆盖度包括如下获得所述全局配置参数修正的覆盖度z_b：

z_b＝y_b/(截距+斜率*gwp_b)-1

其中y_b是在所述感兴趣的序列中所述测试样品的bin b的观测覆盖度，并且gwp_b是针对未受影响的训练样品的bin b的全局配置参数。

29.根据权利要求1-13中任一项所述的方法，其中，来自(e)的所述测试序列标签的全局配置参数修正的覆盖度包含所述感兴趣的核酸序列中bin的全局配置参数修正的覆盖度和归一化序列中bin的全局配置参数修正的覆盖度。

30.根据权利要求1-13中任一项所述的方法，其中，在操作(f)中调节所述全局配置参数修正的覆盖度包括：

将所述参比基因组中的bin分组为多个GC组，每个GC组包含多个bin，其中所述多个bin含有测试序列标签并具有类似的GC含量；

针对多个鲁棒常染色体的每个GC组，确定所述全局配置参数修正的覆盖度的预期值；以及

对于每个GC组，基于相同GC组的确定的预期值，调节所述测试序列标签的全局配置参数修正的覆盖度，从而获得所述感兴趣的核酸序列上的测试序列标签的样品-GC-修正的覆盖度。

31.根据权利要求30所述的方法，其中，所述全局配置参数修正的覆盖度的预期值是多个鲁棒常染色体的GC组的覆盖度的平均值或中位数。

32.根据权利要求30所述的方法，其中，调节所述测试序列标签的全局配置参数修正的覆盖度包括从所述全局配置参数修正的覆盖度减去所述预期值。

33.根据权利要求1-13中任一项所述的方法，其中，在操作(f)中的所述调节所述全局配置参数修正的覆盖度包括：

将线性或非线性数学函数拟合为来自多个鲁棒常染色体的数据点，其中每个数据点使

覆盖度值与GC含量值相关；

基于每个bin的覆盖度的预期值，调节每个bin中的测试序列标签的全局配置参数修正的覆盖度，所述全局配置参数修正的覆盖度等于考虑中的bin的GC含量值处所述数学函数的覆盖度值。

34.根据权利要求33所述的方法，其中，调节所述测试序列标签的全局配置参数修正的覆盖度包括从所述全局配置参数修正的覆盖度减去所述预期值。

35.根据权利要求30所述的方法，其中，所述鲁棒常染色体包含除所述感兴趣的染色体之外的所有常染色体。

36.根据权利要求30所述的方法，其中，所述鲁棒常染色体包含除chr X、Y、13、18、和21之外的所有常染色体。

37.根据权利要求30所述的方法，其中，所述鲁棒常染色体包含除那些确定自偏离自正常二倍体状态的测试样品的常染色体之外的所有常染色体。

38.根据权利要求1-13中任一项所述的方法，进一步包括从多个未受影响的个体和/或所述测试样品提取无细胞DNA。

39.根据权利要求1-13中任一项所述的方法，其中，所述序列读数包含来自个体的整个基因组中的任意处的约20至50-bp的序列。

40.根据权利要求1-13中任一项所述的方法，其中，(a)的序列读数包含条形码25聚体。

41.根据权利要求1-13中任一项所述的方法，其中，所述测试序列标签和所述训练序列标签的覆盖度是基于未排除的位点计数(NES计数)，其中NES计数是映射到未排除的位点的非冗余序列标签的数目。

42.根据权利要求41所述的方法，其中，NES计数是映射到未排除的位点的唯一对齐的非冗余序列标签的数目。

43.根据权利要求1-13中任一项所述的方法，其中，所述bin尺寸为约1000bp至1,000,000bp。

44.根据权利要求1-13中任一项所述的方法，其中，所述bin尺寸为约100,000bp。

45.根据权利要求1-13中任一项所述的方法，进一步包括通过利用所述测试样品的序列读数的数目进行计算来确定所述bin尺寸。

46.根据权利要求45所述的方法，其中，每个bin中序列标签的数目为至少约1000bp。

47.一种用包括一个或多个处理器和系统存储器的计算机系统来实施的用来产生用于评价感兴趣的核酸序列的拷贝数的序列掩码的方法，所述方法包括：

(a)在所述计算机系统中提供训练集，所述训练集包含来自多个未受影响的训练样品的序列读数；

(b)通过所述计算机系统，比对所述训练集与包含所述感兴趣的核酸序列的参比基因组的序列读数，从而提供用于所述训练样品的训练序列标签；

(c)通过所述计算机系统，将所述参比基因组分成多个bin；

(d)通过所述计算机系统，针对每个未受影响的训练样品，确定针对每个训练样品的每个bin中的训练序列标签的覆盖度，其中所述覆盖度表示在bin中序列标签的丰度；

(e)针对每个bin，对所有训练样品，确定所述训练序列标签的预期覆盖度；

(f)通过所述计算机系统，对于每个训练样品，根据每个bin中的预期覆盖度来调节在

每个bin中的训练序列标签的覆盖度，从而针对每个训练样品获得在所述bin中的训练序列标签的全局配置参数修正的覆盖度；

(g)通过所述计算机系统，基于每个bin中跨整个训练样品的全局配置参数修正的覆盖度的变化，产生序列掩码，所述序列掩码包含跨整个所述参比基因组的未掩蔽和掩蔽bin。

48.根据权利要求47所述的方法，其中，在(e)中针对每个bin确定的预期覆盖度包含训练样品的覆盖度的中位数或平均值。

49.根据权利要求48所述的方法，其中，在操作(f)中调节所述训练序列标签的覆盖度包括从每个bin的所述训练序列标签的每个训练样品的覆盖度减去所述中位数或平均值。

50.根据权利要求48所述的方法，其中，在操作(f)中调节所述训练序列标签的覆盖度包括每个bin的所述训练序列标签的每个训练样品的覆盖度除以所述中位数或平均值。

51.根据权利要求47所述的方法，其中，所述感兴趣的核酸序列上的掩蔽bin具有第一掩蔽阈值并且归一化序列上的掩蔽bin具有第二掩蔽阈值。

52.根据权利要求51所述的方法，其中，所述第一掩蔽阈值和所述第二掩蔽阈值的组合提供序列掩码，所述序列掩码导致在包括在未受影响的样品中的所述感兴趣的序列的区域上比利用其他阈值获得的序列掩码更低的覆盖度变异。

53.根据权利要求47所述的方法，进一步包括，在(f)之后并且在(g)之前，通过所述计算机系统，基于在GC含量水平和在每个训练样品中存在的全局配置参数修正的覆盖度之间的关系，调节每个训练样品的bin的全局配置参数修正的覆盖度，从而对于每个训练样品获得所述训练序列标签的样品-GC-修正的覆盖度。

54.根据权利要求53所述的方法，其中，所述调节每个训练样品的所述全局配置参数修正的覆盖度包括：

将所述参比基因组中的所有bin分组为多个GC组，每个GC组包含具有类似的GC含量的多个bin；

对于多个鲁棒常染色体的每个GC组，确定所述全局配置参数修正的覆盖度的预期值；以及

基于相同GC组的确定的预期值，对于每个GC组，调节所述训练序列标签的全局配置参数修正的覆盖度，从而获得在所述感兴趣的核酸序列上的训练序列标签的样品-GC-修正的覆盖度。

55.根据权利要求54所述的方法，其中，对于多个鲁棒常染色体的GC组，所述全局配置参数修正的覆盖度的预期值是所述覆盖度的平均值或中位数。

56.根据权利要求54所述的方法，其中，调节所述训练序列标签的全局配置参数修正的覆盖度包括从所述全局配置参数修正的覆盖度减去所述预期值。

57.根据权利要求53所述的方法，其中，所述调节每个训练样品的所述全局配置参数修正的覆盖度包括：

将线性或非线性数学函数拟合为来自多个鲁棒常染色体的数据点，其中每个数据点使覆盖度值与GC含量值相关；

基于每个bin的覆盖度的预期值，调节每个bin中所述训练序列标签的全局配置参数修正的覆盖度，所述全局配置参数修正的覆盖度等于在所述bin的GC含量值处所述数学函数的覆盖度值。

58.根据权利要求57所述的方法，其中，调节所述训练序列标签的全局配置参数修正的覆盖度包括从所述全局配置参数修正的覆盖度减去所述预期值。

59.根据权利要求1-13和47-58中任一项所述的方法，其中，所述样品是血液样品、尿液样品、或唾液样品。

60.根据权利要求1-13和47-58中任一项所述的方法，其中，所述样品是血浆样品。

61.根据权利要求1-13和47-58中任一项所述的方法，其中，针对每个所述感兴趣的核酸序列确定的序列标签的数目为至少约10,000。

62.一种用于评价测试样品中的感兴趣的核酸序列的拷贝数的系统，所述系统包括：

用于接收来自所述测试样品的核酸的测序仪，其提供来自所述样品的核酸序列信息；

处理器；以及

根据对所述处理器的执行指令已存储的一个或多个计算机可读存储介质，用于利用如下方法来评价所述测试样品中的拷贝数，所述方法包括：

(a)在所述计算机系统中提供所述测试样品的序列读数；

(b)通过所述计算机系统，比对所述测试样品与包含所述感兴趣的核酸序列的参比基因组的序列读数，从而提供测试序列标签；

(c)通过所述计算机系统，确定位于多个bin中的测试序列标签的多个覆盖度，其中所述参比基因组被分成所述多个bin；

(d)通过所述计算机系统，提供针对所述感兴趣的核酸序列的全局配置参数，其中所述全局配置参数包含所述多个bin中的多个预期覆盖度，并且其中所述多个预期覆盖度获自以与所述测试样品基本相同的方式进行测序和比对的未受影响的训练样品的训练集，所述多个预期覆盖度呈现bin之间的变化；

(e)通过所述计算机系统，根据每个bin中的预期覆盖度分别调节每个bin中的所述测试序列标签的覆盖度，从而获得全局配置参数修正的覆盖度；

(f)通过所述计算机系统，基于在GC含量水平和针对所述测试序列标签的bin的全局配置参数修正的覆盖度之间的关系，调节所述全局配置参数修正的覆盖度，从而获得在所述感兴趣的核酸序列上的测试序列标签的样品-GC-修正的覆盖度；以及

(g)通过所述计算机系统，基于所述样品-GC-修正的覆盖度，评价在所述测试样品中所述感兴趣的核酸序列的拷贝数。

用于在确定拷贝数变异中改善检测的灵敏度的方法

[0001] 与相关申请的参考

[0002] 本申请依据35 U.S.C.§119(e)要求于2013年10月21日提交的题为“METHOD FOR IMPROVING THE SENSITIVITY OF DETECTION IN DETERMINING COPY NUMBER VARIATIONS(用于在确定拷贝数变异中改善检测的灵敏度的方法)”美国临时专利申请号61/893,830的优先权,其全部内容以引用方式结合于本文。

背景技术

[0003] 人类医学研究中的关键的努力之一是产生不良健康后果的遗传异常的发现。在许多情况下,在基因组的多个部分中已确定了特定基因和/或关键的诊断标志物,它们是以异常拷贝数存在的。例如,在产前诊断中,全染色体的额外的或丢失的拷贝是频繁发生的遗传性病变。在癌症中,全染色体或染色体片段的拷贝的缺失或倍增,以及基因组的特定区域的较高水平扩增是常见的情况。

[0004] 通过允许识别出结构性异常的细胞遗传学分辨能力已经提供了关于拷贝数变异(CNV)的大部分信息。用于基因筛查和生物学剂量测定的常规程序已经利用了侵入性程序,例如,羊膜穿刺术、脐静脉穿刺术、或绒膜绒毛取样(CVS),来获得用于核型分析的细胞。认识到需要并不要求细胞培养的更快速的测试方法,已经开发了荧光原位杂交(FISH)、定量荧光PCR(QF-PCR)和阵列-比较基因组杂交(阵列-CGH)作为用于拷贝数变异分析的分子细胞遗传学方法。

[0005] 人类医学研究中的关键的努力之一是产生不良健康后果的遗传异常的发现。在许多情况下,在基因组的多个部分中已确定了特定基因和/或关键的诊断标志物,它们是以异常拷贝数存在的。例如,在产前诊断中,全染色体的额外的或丢失的拷贝是频繁发生的遗传性病变。在癌症中,全染色体或染色体片段的拷贝的缺失或倍增,以及基因组的特定区域的较高水平扩增是常见的情况。

[0006] 通过允许识别出结构性异常的细胞遗传学分辨能力已经提供了关于拷贝数变异(CNV)的大部分信息。用于基因筛查和生物学剂量测定的常规程序已经利用了侵入性程序,例如,羊膜穿刺术、脐静脉穿刺术、或绒膜绒毛取样(CVS),来获得用于核型分析的细胞。认识到需要并不要求细胞培养的更快速的测试方法,已经开发了荧光原位杂交(FISH)、定量荧光PCR(QF-PCR)和阵列-比较基因组杂交(阵列-CGH)作为用于拷贝数变异分析的分子细胞遗传学方法。

[0007] 允许在相对较短的时间内对测序整个基因组进行测序的技术的出现以及循环无细胞的DNA(cell-free DNA,cfDNA)的发现已提供了将来自一个有待比较的染色体的遗传物质以与另一个染色体的遗传物质进行比较的机会,而没有伴随侵入性采样方式的风险,其提供了一种诊断感兴趣的基因序列的各种各样拷贝数变异的工具。

[0008] 在一些应用中,拷贝数变异(CNV)的诊断涉及高度的技术挑战。例如,对于异卵多胎(或多卵性,polyzygotic)妊娠的CNV的非侵入性产前诊断(NIPD)比单胎妊娠更加困难,这是因为胎儿cfDNA的总分数与胎儿的数目并不是成正比变化,这使cfDNA的胎儿分数降低

了胎儿数目的一个数量级，其反过来又会降低分析的的信噪比。另外，基于Y染色体的诊断如性别鉴定受到与Y染色体相关限制的影响。具体地，Y染色体的覆盖度(coverage)低于常染色体的覆盖度，并且在Y染色体上的重复序列使得读数到其正确位置的定位的映射复杂化。此外，一些目前的测序方法方法利用超短读数如25聚体读数和标签，从而提出另一个序列比对挑战，因为25聚体标签短于大多数遍在重复元件的典型尺寸。本文披露的一些实施方式提供了在分析用于评价CNV的序列数据时改善灵敏度和/或特异性的方法。

[0009] 无侵入性产前诊断中现有方法存在局限性，包括源于cfDNA的有限水平的灵敏度不足，以及源于基因组信息的固有特性的技术的测序偏差，构成了对能够提供任何或所有的特异性、灵敏度、和适用性，以在各种各样的临床设置中可靠地诊断拷贝数变化的非侵入性方法的持续需要的基础。本文披露的实施方式满足一些上述需要，并且尤其提供了适用于无侵入性产前诊断的实践的可靠方法。

[0010] 发明概述

[0011] 在一些实施方式中，提供了用于确定任何胎儿非整倍性的拷贝数变异(CNV)，以及已知或疑似相关与各种各样的医学状况的CNV的方法。所述方法包括用于减少与基因组序列的GC波动的噪声和误差有关的机制。能够根据本方法确定的CNV包括1-22、X和Y中的任意一个或多个染色体的三体性和单体性，其他染色体多体性，以及任意一个或多个染色体的片段的缺失和/或复制。

[0012] 另一种实施方式提供了一种用于确定在测试样品中感兴趣的核酸序列(例如，临床相关序列)的拷贝数变异(CNV)的方法。所述方法评价感兴趣的序列而不是完整的染色体或染色体的片段的拷贝数变异。

[0013] 在一些实施方式中，用包括一个或多个处理器和系统存储器的计算机系统来实施所述方法，从而评价在包含一个或多个基因组的核酸的测试样品中感兴趣的核酸序列的拷贝数。所述方法包括：(a)提供通过核酸序列测定仪由测试样品获得的序列读数；(b)比对测试样品与包含感兴趣的核酸序列的参比基因组的序列读数，从而提供测试序列标签；(c)确定位于每个bin中的测试序列标签的覆盖度，其中参比基因组被分成多个bin；(d)提供针对感兴趣的核酸序列的全局配置参数(global profile)，其中全局配置参数包含每个bin中的预期覆盖度，以及其中预期覆盖度获自以与测试样品基本相同的方式测序和比对的未受影响的(例如，二倍体)训练样品的训练集(training set)，预期覆盖度呈现bin之间的变化；(e)利用在每个bin中至少感兴趣的核酸序列的预期覆盖度来调节测试序列标签的覆盖度，从而获得针对感兴趣的核酸序列的全局配置参数修正的覆盖度；(f)基于在GC含量水平和全局配置参数修正的覆盖度之间的关系，调节全局配置参数修正的覆盖度，从而获得针对感兴趣的核酸序列的样品-GC-修正的覆盖度；以及(g)基于样品-GC-修正的覆盖度，评价在测试样品中感兴趣的核酸序列的拷贝数。在一些实施方式中，在文库深度差(library depth difference)的归一化之后，获得在步骤(c)中确定的覆盖度。文库归一化可涉及覆盖度除以映射到鲁棒染色体(稳健染色体，robust chromosome)(预期是如本文所描述的二倍体)的读数的总数。可替换地，文库深度归一化可能涉及覆盖度除以映射到全基因组的读数数目，从而产生序列与标签密度比值。在一些实施方式中，样品本身的测序数据可以用来得到估计具有二倍体覆盖度的基因组区，以及将那些区域文库归一化。与通常在(c)之后进行的其他形式的归一化，如归一化在(f)中获得的全局配置参数修正的覆盖度，分别进行文

库深度归一化。另一种形式的“归一化”产生如下文所描述的“序列剂量”。

[0014] 在一些实施方式中，所述方法进一步涉及，在确定bin的覆盖度的操作(c)之前，施加序列掩码(sequence mask)，其排除掩蔽bin中的考虑覆盖度。在一些实施方式中，序列掩码获自多个未受影响的训练样品的序列读数。通过比对训练集与参比基因组的序列读数来获得序列掩码，从而提供针对训练样品的训练序列标签。所述方法还涉及将参比基因组分成多个bin以及针对每个训练样品确定在每个bin中的训练序列标签的覆盖度。所述方法进一步涉及产生包含未掩蔽(unmasked)和掩蔽(masked)bin的序列掩码。每个掩蔽bin具有超过掩蔽阈值(masking threshold)的分布指数，该分布指数与训练样品的覆盖度的分布有关。在一些实施方式中，用来确定掩蔽和未掩蔽bin的分布指数与训练样品的覆盖度的变化(例如，变异系数)在数学上相关。分布指数作为用于掩蔽bin的标准来事实，这是因为在训练样品中呈现较大可变性或变异的bin具有高分布指数，因而对应用于表征拷贝数而言是不可靠的。

[0015] 在一些实施方式中，在产生或施加序列掩码之前，所述方法首先去除常见于未受影响的训练样品的系统性变化(或全局配置参数)。这可以通过根据在每个bin中的预期覆盖度调节训练序列标签的覆盖度来实现，从而获得在bin中的训练序列标签的全局配置参数修正的覆盖度，其然后用来产生序列掩码。在一些实施方式中，归一化覆盖度的量用来计算掩码。归一化覆盖度的量是感兴趣的核酸序列的覆盖度与归一化序列的覆盖度的比率。在一些实施方式中，在感兴趣的核酸序列上的掩蔽bin具有第一掩蔽阈值以及在归一化序列上的掩蔽bin具有第二掩蔽阈值。在一些实施方式中，第一掩蔽阈值和第二掩蔽阈值的组合提供这样的序列掩码，其导致在未受影响的样品中在包括感兴趣的序列的区域内比利用其他阈值获得的掩码更低的覆盖度变异。覆盖度的变化反映了序列掩码在整个样品和运行上控制变异的能力，因而较低变异会使受影响的和未受影响的样品之间的分离增加。在一些实施方式中，掩蔽阈值导致在验证样品中覆盖度的较小的变异系数和/或在ROC分析中较大的d’值。

[0016] 在一些实施方式中，序列掩码包括在bin内由跨整个训练样品的映射质量得分(mapping quality score)的分布所定义的掩蔽bin和未掩蔽bin。映射质量得分来源于多个未受影响的训练样品与参比基因组的比对序列读数。

[0017] 在一些实施方式中，评价在测试样品中感兴趣的核酸序列的拷贝数包括利用归一化序列的覆盖度信息来计算测试样品的感兴趣的核酸序列的序列剂量。在一些实施方式中，计算序列剂量包括在感兴趣的核酸序列中的测试序列标签的覆盖度(例如，样品-GC-修正的覆盖度)除以在归一化序列中的测试序列标签的覆盖度。其他方法可以用来计算序列剂量，如利用线性回归或稳健线性回归并依据基因组的其他归一化区的归一化覆盖度来对感兴趣的序列的归一化覆盖度建模。

[0018] 在一些实施方式中，归一化序列包含一个或多个鲁棒常染色体序列或它们的片段。在一些实施方式中，鲁棒常染色体包括除感兴趣的染色体之外的所有常染色体。在一些实施方式中，鲁棒常染色体包括除chr X、Y、13、18、和21之外的所有常染色体。在一些实施方式中，鲁棒常染色体包括除那些确定自偏离自正常二倍体状态的样品的常染色体之外的所有常染色体。

[0019] 在一些实施方式中，评价拷贝数进一步包括利用归一化序列的覆盖度信息来计算

测试样品的感兴趣的核酸序列的归一化染色体值或归一化片段值。

[0020] 在一些实施方式中，测试样品包括来自两个不同的基因组的核酸的混合物。在一些实施方式中，测试样品包括cfDNA分子。在一些实施方式中，测试样品包括胎儿和母体无细胞核酸。在一些实施方式中，测试样品包括来自两个或更多个胎儿的胎儿无细胞核酸。在一些实施方式中，测试样品包含来自相同受试者的癌细胞和未受影响的细胞的核酸(细胞基因组DNA和/或cfDNA)。

[0021] 在一些实施方式中，评价在测试样品中感兴趣的核酸序列的拷贝数与确定完全或部分胎儿非整倍性的存在或不存在有关。

[0022] 在一些实施方式中，在获得样品-GC-修正的覆盖度的操作(f)之后，考虑到CNV的评价，所述方法进一步涉及除去样品-GC-修正的覆盖度的离群bin(outlier bins)。在一些实施方式中，离群bin是这样的bin，其中位数样品-GC-修正的覆盖度离在每个染色体中的所有bin的中位数是大于约3中位数绝对偏差(median absolute deviation)。

[0023] 在一些实施方式中，在每个bin中的预期覆盖度是跨整个训练样品的中位数或平均值。在一些实施方式中，在计算全局配置参数作为中位数或均值归一化覆盖度之前，针对GC含量变异，修正在训练样品中的覆盖度。

[0024] 在一些实施方式中，通过(i)在一个或多个鲁棒染色体或区域中的多个bin中获得在测试序列标签的覆盖度和预期覆盖度之间的数学关系，以及(ii)将数学关系应用于在感兴趣的序列中的bin来调节测试序列标签的覆盖度。在一些实施方式中，利用在来自未受影响的训练样品的预期覆盖度值和在鲁棒染色体或基因组的其他鲁棒区(robust region)中的测试样品的覆盖度值之间的线性关系，来修正在测试样品中覆盖度的变化。上述调节导致全局配置参数修正的覆盖度。在一些情况下，上述调节涉及获得在鲁棒染色体或区中针对bin的子集的针对测试样品的覆盖度，具体如下：

[0025] y_a＝截距+斜率*gwp_a

[0026] 其中y_a是在一个或多个强大染色体或区域中针对测试样品的bin的覆盖度，以及gwp_a是针对未受影响的训练样品的bin的全局配置参数。然后上述过程计算针对感兴趣的序列或区的全局配置参数修正的覆盖度z_b，作为：

[0027] z_b＝y_b/(截距+斜率*gwp_b)-1

[0028] 其中y_b是针对测试样品在感兴趣的序列中binb的观测覆盖度(其可以位于鲁棒染色体或区之外)，以及gwp_b是针对未受影响的训练样品的binb的全局配置参数。分母(截距+斜率*gwp_b)是binb的覆盖度，其应在未受影响的测试样品中进行观测。在藏匿(harboring)拷贝数变异的感兴趣的序列的情况下，针对binb的观测覆盖度，因而全局配置参数修正的覆盖度值将显著偏离未受影响的样品的覆盖度。例如，在三体样品的情况下，针对在受影响的染色体上的bin，修正的覆盖度z_b将正比于胎儿分数。通过计算在鲁棒染色体上的截距和斜率，此过程在样品内归一化，然后评价目标染色体(或其他感兴趣的序列)如何偏离适用于在同一样品中的鲁棒染色体的关系(如由斜率和截距所描述的)。

[0029] 在一些实施方式中，来自(e)的全局配置参数修正的测试序列标签的覆盖度包含在感兴趣的核酸序列中bin的全局配置参数修正的覆盖度和在归一化序列中bin的全局配置参数修正的覆盖度。

[0030] 在一些实施方式中，在操作(f)中调节全局配置参数修正的覆盖度包括将在参比

基因组中的bin分组为多个GC组，每个GC组包含多个bin，其中多个bin含有测试序列标签并具有类似的GC含量；确定针对多个鲁棒常染色体的每个GC组的全局配置参数修正的覆盖度的预期值；以及基于相同GC组的确定的预期值，调节针对每个GC组的全局配置参数修正的测试序列标签的覆盖度，从而获得在感兴趣的核酸序列上的测试序列标签的样品-GC-修正的覆盖度。

[0031] 在一些实施方式中，全局配置参数修正的覆盖度的预期值是针对多个鲁棒常染色体的GC组的覆盖度的平均值或中位数。在一些实施方式中，通过从全局配置参数修正的覆盖度减去预期值来调节测试序列标签的全局配置参数修正的覆盖度。

[0032] 在一些实施方式中，在操作(f)中调节全局配置参数修正的覆盖度涉及将线性或非线性数学函数拟合于来自多个鲁棒常染色体的数据点，其中每个数据点使覆盖度值相关与GC含量值。然后，通过在所考虑的bin的GC含量值下等于数学函数的覆盖度值的值，所述方法调节覆盖度。在一些实施方式中，所述方法从全局配置参数修正的覆盖度减去预期值。在其他实施方式中，所述方法将覆盖度量除以预期值。

[0033] 在一些实施方式中，用于评价CNV的方法还涉及从多个不受影响的的个体和/或测试样品提取无细胞DNA。在一些实施方式中，所述方法还涉及利用测序仪来测序来自测试样品的核酸，从而产生测试样品的序列读数。在一些实施方式中，序列读数包含来自在个体的整个基因组中任意处的约20至50-bp的序列。在一些实施方式中，序列读数包括条形码25聚体。

[0034] 在一些实施方式中，测试序列标签和训练序列标签的覆盖度是基于未排除的位点计数(non-excluded site counts，NES计数)，其中NES计数是映射到未排除的位点的非冗余的和/或唯一对齐的序列标签的数目。

[0035] 在一些实施方式中，感兴趣的核酸序列被分成约1000bp至1,000,000bp的bin。在一些实施方式中，bin尺寸是约100,000bp。在一些实施方式中，参照测试样品的序列读数的数目来计算bin尺寸。在一些实施方式中，在每个bin中序列标签的数目为至少约1000bp。

[0036] 本文披露的一些实施方式提供了用来产生用于评价感兴趣的核酸序列的拷贝数的序列掩码的方法。所述方法包括：(a)在计算机系统中提供包含来自多个未受影响的训练样品的序列读数的训练集；(b)比对训练集与包含感兴趣的核酸序列的参比基因组的序列读数，从而提供用于训练样品的训练序列标签；(c)将参比基因组分成多个bin；(d)在对于每个训练样品的每个bin中，针对每个未受影响的训练样品，确定训练序列标签的覆盖度；(e)对所有训练样品，针对每个bin，确定训练序列标签的预期覆盖度；(f)根据在每个bin中的预期覆盖度，在对于每个训练样品的每个bin中调节训练序列标签的覆盖度，从而获得在对于每个训练样品的bin中训练序列标签的全局配置参数修正的覆盖度；以及(g)产生包含整个参比基因组的未掩蔽和掩蔽bin的序列掩码，其中每个掩蔽bin具有超过掩蔽阈值的分布特征，以及提供了分布特征，用于调节在跨整个训练样品的bin中训练序列标签的覆盖度。

[0037] 在一些实施方式中，在(e)中针对每个bin确定的预期覆盖度包括训练样品的覆盖度的中位数或均值。在一些实施方式中，在(f)操作中，调节训练序列标签的覆盖度包括从针对每个bin的训练序列标签的每个训练样品的覆盖度减去中位数或均值。在一些实施方式中，通过针对每个bin的训练序列标签的每个训练样品的覆盖度除以中位数或均值来完

成调节。

[0038] 在一些实施方式中，在感兴趣的核酸序列上的掩蔽bin具有第一掩蔽阈值以及在归一化序列上的掩蔽bin具有第二掩蔽阈值。在一些实施方式中，第一掩蔽阈值和第二掩蔽阈值的组合提供这样的序列掩码，其导致在包括在未受影响的样品中的感兴趣的序列的区域内比利用其他阈值获得的掩码更低的覆盖度变异。

[0039] 在一些实施方式中，用于产生序列掩码的方法进一步涉及，在(f)之后以及在(g)之前，基于在GC含量水平和在每个训练样品中存在的全局配置参数修正的覆盖度之间的关系，调节针对每个训练样品的bin的全局配置参数修正的覆盖度，从而获得对于每个训练样品的训练序列标签的样品-GC-修正的覆盖度。

[0040] 在一些实施方式中，对于每个训练样品，全局配置参数修正的覆盖度的调节涉及：将在参比基因组中的所有bin分组为多个GC组，每个GC组包含具有类似的GC含量的多个bin；对于多个鲁棒常染色体的每个GC组，确定全局配置参数修正的覆盖度的预期值；以及基于相同GC组的确定的预期值，对于每个GC组，调节训练序列标签的全局配置参数修正的覆盖度，从而获得在感兴趣的核酸序列上的训练序列标签的样品-GC-修正的覆盖度。

[0041] 在一些实施方式中，全局配置参数修正的覆盖度的预期值是针对多个鲁棒常染色体的GC组的覆盖度的平均值或中位数。在一些实施方式中，调节训练序列标签的全局配置参数修正的覆盖度涉及从全局配置参数修正的覆盖度减去预期值。

[0042] 在一些实施方式中，对于每个训练样品，调节全局配置参数修正的覆盖度涉及：将线性或非线性数学函数拟合为来自多个鲁棒常染色体的数据点，其中每个数据点使覆盖度值相关与GC含量值。然后，基于对于每个bin的覆盖度的预期值，所述方法调节在每个bin中的训练序列标签的全局配置参数修正的覆盖度，其等于在bin的GC含量值下数学函数的覆盖度值。

[0043] 在一些实施方式中，调节训练序列标签的全局配置参数修正的覆盖度包括从全局配置参数修正的覆盖度减去预期值。在其他实施方式中，覆盖度除以预期值。

[0044] 在一些实施方式中，测试样品可以是母体样品，选自血液、血浆、血清、尿液和唾液样品。在任何一种实施方式中，测试样品可以是血浆样品。母体样品的核酸分子是胎儿和母体无细胞DNA分子的混合物。可以下一代测序(NGS)来进行核酸的测序。在一些实施方式中，测序是利用合成法测序(sequencing-by-synthesis)并借助于可逆染料终止子的大规模平行测序。在其他实施方式中，测序是连接法测序(sequencing-by-ligation)。在其他实施方式中，测序是单分子测序。可选地，在测序之前，进行扩增步骤。

[0045] 另一种实施方式提供了用于在测试样品中确定感兴趣的核酸序列(例如，临床相关序列)的拷贝数变异(CNV)的方法。所述方法评价感兴趣的序列的拷贝数变异，而不是完整的染色体或染色体的片段。

[0046] 在用计算机系统实施的某些实施方式中，针对每一个或多个感兴趣的染色体或感兴趣的染色体片段确定的序列标签的数目为至少约10,000、或至少约100,000。

[0047] 所披露的实施方式还提供了计算机程序产品，该产品包括非临时性计算机可读介质，其上提供程序指令，用于执行列举的操作和本文描述的其他计算操作。

[0048] 一些实施方式提供了用于评价在测试样品中的感兴趣的核酸序列的拷贝数的系统。上述系统包括：测序仪，用于接收来自测试样品的核酸，从而提供来自样品的核酸序列

信息;处理器;以及一个或多个计算机可读存储介质,在其上已存储用于用所述处理器加以执行的指令,以利用本文列举的方法来评价在测试样品中的拷贝数。

[0049] 在一些实施方式中,方法另外包括测序所述测试样品的至少一部分的所述核酸分子以获得关于所述测试样品的所述胎儿和母体核酸分子的所述序列信息。上述测序可能涉及对来自母体测试样品的母体和胎儿核酸的大规模平行测序以产生序列读数。

[0050] 虽然在此这些实例涉及人类并且语言主要针对人类问题,但是本发明的概念也适用于来自任何植物或动物的基因组。依据下文的描述和所附的权利要求,本公开内容的这些和其他的目的和特征将变得更充分显而易见的,或可以通过如下文阐述的公开内容的实践来实现。

[0051] 通过引用并入

[0052] 所有专利、专利申请、和其他出版物,包括在这些参考文献中披露的、在本文中提及的所有序列,以引用方式明确地并入本文。在相关部分中,引用的所有文件的全部内容以引用方式结合于本文。然而,任何文件的引用不应当被解释为承认它是关于本公开内容的现有技术。

附图说明

[0053] 图1是用于在包含核酸的混合物的测试样品中确定拷贝数变异的存在或不存在的方法100的流程图。

[0054] 图2描述了用来确定用于拷贝数的评价的感兴趣的核酸序列的覆盖度的过程的流程图。

[0055] 图3A示出用于在来自测试样品的序列数据中减少噪声的过程的一个实施例的流程图。

[0056] 图3B-3K呈现在图3A中描述的过程的不同阶段获得的数据的分析。

[0057] 图4A示出过程的流程图,上述过程用于产生序列掩码,其用于减少在序列数据中的噪声。

[0058] 图4B表明,MapQ得分与归一化覆盖度量CV具有单调的强相关性。

[0059] 图5是用于处理测试样品并最终作出诊断的分散系统的框图。

[0060] 图6示意性地说明,在处理测试样品中的不同的操作如何可以被分组以由系统的不同元件来处理。

[0061] 图7A和7B示出根据在实施例1a中描述的方法简化方法(图7A)以及在实施例1b中描述的方法方法(图7B)所制备的cfDNA测序文库的电泳图。

[0062] 图8示出来自118个双胞胎妊娠的母体血浆样品的归一化染色体值(NCV)分布。(A)染色体21和18的NCV分布,三个样品被分类为T21受影响的(包括对于T21的嵌合体(mosaic)胎儿)以及一个样品被分类为T18受影响的。(B)Y染色体的NCV分布。将群组分为临床上被分类为雌性/雌性的样品或含有至少一个雄性胎儿的样品(雄性/雌性和雄性/雄性)并利用针对Y染色体的NCV来确定Y染色体的存在。

[0063] 图9示出在NIPT研究中分析的双胞胎样品。在各种研究中使用的双胞胎样品的数目,以评价市售NIPT测试的性能。

[0064] 发明详述

[0065] 披露的实施方式涉及用于评价在包含胎儿和母体无细胞核酸的测试样品中Y染色体的拷贝数的方法、装置、和系统。在一些实施方式中，感兴趣的序列包括基因组片段序列，其范围为，例如，千碱基(kb)至兆碱基(Mb)至整个染色体，其已知或疑似相关与遗传或疾病状况。在一些实施方式中，Y染色体的拷贝数用来确定胎儿性别。在一些实施方式中，根据本方法可以确定的CNV包括性Y染色体的单体性和三体性(例如47，XXY和47，XYY)，性染色体的其他多体性如四体性和五体性(例如XXXXY和XYYYY)，以及任何一个或多个性染色体的片段的缺失和/或复制。感兴趣的序列的其他实例包括相关与众所周知的非整倍体的染色体，例如，三体性XXX、三体性21、以及在疾病如癌症中倍增的染色体的片段，例如，在急性髓细胞样白血病中的部分三体性8。

[0066] 除非另有说明，本文披露的方法和系统的实践涉及常规技术和在分子生物学、微生物学、蛋白纯化、蛋白质工程、蛋白质和DNA测序、以及重组DNA领域中通常使用的仪器，其是在本领域的技术范围内。这样的技术和仪器是本领域技术人员已知的并且描述于众多的教科书和参考著作(参见例如，Sambrook等人，“Molecular Cloning:A Laboratory Manual，”第三版(Cold Spring Harbor)，[2001])；以及Ausubel等人，“Current Protocols in Molecular Biology”[1987])。

[0067] 数值范围包括限定范围的数字。意图是，在整个本说明书中给出的每个最大数值限度包括每个较低数值限度，好像在本文中明确写入这样的较低数值限度。在整个本说明书中给出的每个最小数值限度将包括每个较高数值限度，好像在本文中明确写入这样的较高数值限度。在整个本说明书中给出的每个数值范围将包括在上述较宽数值范围之内的每个较窄数值范围，好像在本文中均明确写入这样的较窄数值范围。

[0068] 本文中提供的标题并不旨在限制本公开内容。

[0069] 除非本文中另有定义，本文中使用的所有技术和科学术语具有和本领域普通技术人员通常理解的相同的含义。包括在本文中包括的术语的各种科学词典是众所周知的并且是本领域技术人员可获得的。虽然类似或等同于本文描述的那些方法和材料的任何方法和材料可用于本文披露的实施方式的实践或测试，但描述了一些方法和材料。

[0070] 通过参照作为整体的说明书来更全面地描述下文马上定义的术语。应当理解的是，本公开内容不限于所描述的特定的方法、方法、和试剂，因为它们可以变化，其取决于本领域技术人员使用它们的上下文。

[0071] 定义

[0072] 如在本文中所使用的，单数术语“一个”、“一种”和“该”包括复数对象(除非上下文另外明确指出)。

[0073] 除非另外指明，对应地，核酸是按5′到3′方向从左到右书写并且氨基酸序列是按氨基到羧基方向从左到右书写。

[0074] 当在本文中在分析核酸样品的CNV的情况下使用时，术语“评价”是指通过三种类型的调用之一来表征染色体或片段非整倍体的状态：“正常的”或“未受影响的”、“受影响的”、和“无调用的”。通常设置用于调用正常的和受影响的阈值。在样品中测量涉及到非整倍体或其他拷贝数变异的参数并将测量值相比于阈值。对于复制型非整倍体，如果染色体或片段剂量(或其他测量值序列含量)是高于受影响的样品的定义的阈值设置，则进行受影响的的调用。对于这样的非整倍体，如果染色体或片段剂量是低于针对正常的样品的阈值

设置，则进行正常的的调用。相比之下，对于缺失型非整倍体，如果染色体或片段剂量是低于针对受影响的样品所定义的阈值，则进行受影响的的调用，以及如果染色体或片段剂量是高于针对正常的样品的阈值设置，则进行正常的的调用。例如，在三体性的存在下，由参数值，例如，低于用户定义的可靠性阈值的测试染色体剂量，来确定“正常的”调用，以及由参数，例如，测试染色体剂量，其是高于用户定义的可靠性阈值，来确定“受影响的”调用。由参数，例如，位于进行“正常的”或“受影响的”调用的阈值之间的测试染色体剂量，来确定“无调用的”结果。术语“无调用的”与“未分类的”互换使用。

[0075] 术语“拷贝数变异”在本文中是指，和在参比样品中存在的核酸序列的拷贝数比较，在测试样品中存在的核酸序列的拷贝的数目的变化。在某些实施方式中，核酸序列是1kb或更大。在一些情况下，核酸序列是全染色体或其显著部分。“拷贝数变异体(variant)”是指通过比较在测试样品中的感兴趣的序列与存在于合格样品中的序列，其中发现的拷贝数差异为1kb或更大的核酸的序列。拷贝数变异体/变异包括缺失(包括微缺失)、插入(包括微插入)、复制、倍增、倒位、易位和复杂的多位点变异体。CNV涵盖染色体性非整倍性和部分非整倍性。

[0076] 术语“非整倍性”在本文中是指由获得或丢失整个染色体、或染色体的一部分而引起的遗传物质的不平衡。

[0077] 术语“染色体非整倍性”和“完全染色体非整倍性”在本文中是指由全染色体的获得或丢失所引起的遗传物质的失衡，并且包括种系非整倍性和嵌合性非整倍性。

[0078] 术语“部分非整倍性”和“部分染色体非整倍性”在本文中是指由部分染色体(例如，部分单体性和部分三体性)的获得或丢失所引起的遗传物质的失衡，并且涵盖来自易位、缺失和插入的失衡。

[0079] 术语“多个/多种”在本文中是用于提及一定数目的核酸分子或序列标签，该数目在使用本发明的方法的测试样品和合格样品中足以识别拷贝数变异(例如染色体剂量)中的显著性差异。在一些实施方案中，对于每一测试样品获得了包括在20和40bp读数之间的至少约3×10^6个序列标签、至少约5×10^6个序列标签、至少约8×10^6个序列标签、至少约10×10^6个序列标签、至少约15×10^6个序列标签、至少约20×10^6个序列标签、至少约30×10^6个序列标签、至少约40×10^6个序列标签、或至少约50×10^6个序列标签。

[0080] 术语“多核苷酸”、“核酸”和“核酸分子”可互换使用并且是指核苷酸的共价连接序列(即，用于RNA的核糖核苷酸和用于DNA的脱氧核糖核苷酸)，其中通过磷酸二酯基团将一个核苷酸的戊糖的3’位连接于下一个核苷酸的戊糖的5’位。核苷酸包括任何形式的核酸的序列，包括但不限于RNA和DNA分子如cfDNA分子。术语“多核苷酸”包括但不限于单链和双链多核苷酸。

[0081] 术语“部分”在本文中用来指在生物样品中胎儿和母体核酸分子的序列信息的量，其总量小于1个人类基因组的序列信息。

[0082] 术语“测试样品”在本文中是指这样的样品，其通常来源于生物液体、细胞、组织、器官、或生物体，其包含核酸或核酸的混合物，其包含要被筛查拷贝数变异的至少一个核酸序列。在某些实施方式中，上述样品包含至少一种核酸序列，其拷贝数被疑似已经历变化。这样的样品包括但不限于痰/口腔液、羊水、血液、血液部分、或细针活检样品(例如，手术活检、细针活检等)、尿、腹腔液、胸膜液等。虽然上述样品经常取自人受试者(例如，患者)，但

上述测定可以用于在来自任何哺乳动物的样品中的拷贝数变异(CNV)，包括但不限于狗、猫、马、山羊、绵羊、牛、猪等。当获自生物源或在用来改变样品的特性的预处理之后，可以直接使用上述样品。例如，这样的预处理可以包括从血液制备血浆，稀释粘性液体等。预处理的方法可能还涉及但不限于过滤、沉淀、稀释、蒸馏、混合、离心、冷冻、冷冻干燥、浓缩、扩增、核酸片段化、干扰成分的灭活、试剂的添加、裂解等。如果针对上述样品，采用这样的预处理方法，则这样的预处理方法通常是如此以致感兴趣的核酸留在测试样品中，有时具有正比于在未经处理的测试样品中的浓度的浓度(例如，即未经受任何这样的预处理方法的样品)。针对本文描述的方法，这样的“处理过的”样品仍然被认为是生物“测试”样品。

[0083] 术语“合格样品”或“未受影响的样品”在本文中是指这样的样品，其包含以已知的拷贝数(在测试样品中的核酸将与其比较)存在的核酸的混合物，并且对于感兴趣的核酸序列，它是正常的样品，即，不是非整倍体。在一些实施方式中，合格样品用作训练集的未受影响的训练样品，以得到序列掩码或序列分布图。在某些实施方式中，合格样品用于确定一个或多个归一化染色体或用于在考虑中的染色体的片段。例如，合格样品可以用于确定染色体21的归一化染色体。在这种情况下，合格样品是这样的样品，其不是三体性21样品。另一个实施例涉及仅利用女性作为用于X染色体的合格样品。合格样品还可以用于其他目的如确定用于调用受影响的样品的阈值，确定用于定义在参比序列上的掩码区的阈值，确定针对基因组的不同区的预期覆盖度量等。

[0084] 术语“训练集”在本文中是指一组训练样品，其可以包含受影响的和/或未受影响的样品并且用来开发用于分析测试样品的模型。在一些实施方式中，上述训练集包括未受影响的样品。在这些实施方式中，利用对于感兴趣的拷贝数变异未受影响的样品的训练集来建立用于确定CNV的阈值。在训练集中的未受影响的样品可以用作合格样品来确定归一化序列，例如，归一化染色体，以及未受影响的样品的染色体剂量用来设定对于感兴趣的每个序列，例如，染色体，的阈值。在一些实施方式中，训练集包括受影响的样品。在训练集中的受影响的样品可以用来确认，受影响的测试样品可以容易地区别与未受影响的样品。

[0085] “训练集”在本文中还用于指感兴趣的群体的统计样品的一组个体，上述个体用来确定适用于群体的感兴趣的一个或多个定量值的数据。统计样品是在感兴趣的群体中的一个子集的个体。上述个体可以是人、动物、组织、细胞、其他生物样品(即，统计样品可以包括多种生物样品)，以及其他个别实体提供用于统计分析的数据点。

[0086] 通常，连同验证集(validation set)一起来使用训练集。在本文中参照在统计样品中的一组个体来使用术语“验证集”，所述个体的数据用来验证或评价利用训练集确定的感兴趣的定量值。在一些实施方式中，例如，训练集提供用于计算参比序列的掩码的数据，验证集提供用来验证或评价掩码的数据。

[0087] “拷贝数的评价”在本文中用来指涉及到序列的拷贝数的基因序列的状态的统计评价。例如，在一些实施方式中，评价包括确定基因序列的存在或不存在。在一些实施方式中，评价包括确定基因序列的部分或完全非整倍体。在其他实施方式中，评价包括基于基因序列的拷贝数来区别两个或更多的样品。在一些实施方式中，评价包括基于基因序列的拷贝数的统计分析，例如，归一化和比较。

[0088] 术语“合格核酸”与“合格序列”互换使用，其是这样的序列，相对于其比较测试序列或测试核酸的量。合格序列是在生物样品中存在的序列，其优选具有已知的表示，即，合

格序列的量是已知的。通常，合格序列是在“合格样品”中存在的序列。“感兴趣的合格序列”是这样的合格序列，在合格样品中其量是已知的，并且是这样的序列，其相关与在具有医学状况的个体中序列表示的差异。

[0089] 术语“感兴趣的序列”或“感兴趣的核酸序列”在本文中是指这样的核酸序列，其相关与在健康与患病个体中序列表示的差异。感兴趣的序列可以是在疾病或遗传病下在染色体上所表示的序列，即，过度表示或表示不足的。感兴趣的序列可以是染色体的一部分，即，染色体片段，或全染色体。例如，感兴趣的序列可以是在非整倍体的条件下过度表示的染色体，或在癌症中表示不足的编码肿瘤抑制子的基因。感兴趣的序列包括在总群体、或受试者细胞的亚群中过度表示或表示不足的序列。“感兴趣的合格序列”是在合格样品中的感兴趣的序列。“感兴趣的测试序列”是在测试样品中的感兴趣的序列。

[0090] 术语“归一化序列”在本文中是指这样的序列，其用来归一化映射到相关与归一化序列的感兴趣的序列的序列标签的数目。在一些实施方式中，归一化序列包含鲁棒染色体。“鲁棒染色体”是一种染色体，其不可能是非整倍体。在涉及人染色体的一些情况下，鲁棒染色体是不同于X染色体、Y染色体、染色体13、染色体18、和染色体21的任何染色体。在一些实施方式中，归一化序列显示在样品中映射到它的序列标签的数目的变异性以及测序运行，其接近对其他用作归一化参数的感兴趣的序列的变异性。归一化序列能够区分受影响的样品与一个或多个未受影响的样品。在一些实施方式中，当相比于其他潜在的归一化序列如其他染色体时，归一化序列最好或有效地区分受影响的样品与一个或多个未受影响的样品。在一些实施方式中，归一化序列的变异性被计算为针对整个样品和测序运行的感兴趣的序列的染色体剂量的变异性。在一些实施方式中，在一组不受影响的的样品中确定归一化序列。

[0091] “归一化染色体”、“归一化分母染色体”、或“归一化染色体序列”是“归一化序列”的实例。“归一化染色体序列”可以由单染色体或一组染色体组成。在一些实施方式中，归一化序列包括两个或更多的鲁棒染色体。在某些实施方式中，鲁棒染色体是不同于染色体X、Y、13、18、和21的所有常染色体。“归一化片段”是“归一化序列”的另一个实例。“归一化片段序列”可以由染色体的单片段组成，或它可以由相同或不同染色体的两个或更多片段组成。在某些实施方式中，归一化序列旨在归一化变异性如过程相关的、染色体间的(运行内)、和测序间的(运行间)变异性。

[0092] 术语“可微性”在本文中是指归一化染色体的特性，其使得能够区分一个或多个未受影响的，即，正常的，样品与一个或多个受影响的，即，非整倍体，样品。显示最大“可微性”的归一化染色体是这样染色体或染色体组，其提供在针对在一组合格样品中的感兴趣的染色体的染色体剂量和针对在一个或多个受影响的样品中的在相应染色体中感兴趣的相同染色体的染色体剂量的分布之间的最大的统计学差异。

[0093] 术语“变异性”在本文中是指归一化染色体的另一特性，其使得能够区分一个或多个未受影响的，即，正常的，样品与一个或多个受影响的，即，非整倍体，样品。归一化染色体的变异性，其是在一组合格样品中加以测量，是指被映射到它的序列标签的数目的变异性，其接近被映射到感兴趣的染色体的序列标签的数目的变异性，对其他作为归一化参数。

[0094] 术语“序列标签密度”在本文中是指被映射到参比基因组序列的序列读数的数目，例如，针对染色体21的序列标签密度是通过测序方法产生的被映射到参比基因组的染色体

21的序列读数的数目。

[0095] 术语“序列标签密度比”在本文中是指被映射到参比基因组的染色体，例如，染色体21，的序列标签的数目与参比基因组染色体的长度的比率。

[0096] 术语“序列剂量”在本文中是指这样的参数，其相关针对感兴趣的序列确定的序列标签的数目和针对归一化序列确定的序列标签的数目。在一些情况下，序列剂量是针对感兴趣的序列的序列标签覆盖度与针对归一化序列的序列标签覆盖度的比率。在一些情况下，序列剂量是指这样的参数，其使感兴趣的序列的序列标签密度相关与归一化序列的序列标签密度。“测试序列剂量”是这样的参数，其使在测试样品中确定的感兴趣的序列，例如，染色体21，的序列标签密度相关与归一化序列，例如，染色体9，的序列标签密度。同样地，“合格序列剂量”是这样的参数，其使感兴趣的序列的序列标签密度相关与在合格样品中确定的归一化序列的序列标签密度。

[0097] 术语“覆盖度”是指映射到限定序列的序列标签的丰度。可以通过序列标签密度(或序列标签的计数)、序列标签密度比、归一化覆盖度量、调节的覆盖度值等来定量地表示覆盖度。

[0098] 术语“覆盖度量”是原始覆盖度的修正并且经常表示在基因组的区如bin中序列标签的相对量(有时被称为计数)。可以通过归一化、调节和/或修正基因组的区的原始覆盖度或计数来获得覆盖度量。例如，可以通过映射到一个区的序列标签计数除以映射到整个基因组的总数序列标签来获得针对上述区的归一化覆盖度量。归一化覆盖度量允许比较整个不同样品的bin的覆盖度，其可以具有不同深度的测序。它不同于序列剂量，因为后者通常是通过除以映射到整个基因组的一个子集的标记计数来获得。上述子集是归一化片段或染色体。覆盖度量，无论是否被归一化，可以针对在基因组上的不同区的全局配置参数变化、G-C分数变化、在鲁棒染色体中的离群等加以修正。

[0099] 术语“下一代测序(NGS)”在本文中是指这样的测序方法，其允许克隆扩增分子和单核酸分子的大规模平行测序。NGS的非限制性实例包括利用可逆染料终止子的合成测序，以及连接测序。

[0100] 术语“参数”在本文中是指表征物理性能的数值。经常地，参数数值上表征定量数据集和/或在定量数据集之间的数值关系。例如，在映射到染色体的序列标签的数目和上述标记对其映射的染色体的长度之间的比率(或比率的函数)是一种参数。

[0101] 术语“阈值”和“合格阈值”在本文中是指任何数字，其用作截止值来表征样品如含有来自疑似具有医学状况的生物体的核酸的测试样品。阈值可以相比于参数值，以确定引起这样的参数值的样品是否提示生物体具有医学状况。在某些实施方式中，合格阈值是利用合格数据集加以计算并作为在生物体中拷贝数变异的诊断的限制，例如，非整倍体。如果获自本文披露的方法的结果超过阈值，由受试者可以被诊断为具有拷贝数变异，例如，三体性21。可以通过分析针对样品的训练集计算的归一化值(例如染色体剂量，NCV或NSV)来确定针对本文描述的方法的合适的阈值。可以利用在包含合格(即，未受影响的)样品和受影响的样品的训练集中的合格(即，未受影响的)样品来确定阈值。在已知具有染色体非整倍体的训练集中的样品(即，受影响的样品)可以用来证实，选择的阈值可以用于区分在测试组中的受影响的与未受影响的样品(见本文的实施例)。阈值的选择取决于用户为进行分类所希望具有的置信水平。在一些实施方式中，用来识别合适的阈值的训练集包含至少10、至

少20、至少30、至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400、至少500、至少600、至少700、至少800、至少900、至少1000、至少2000、至少3000、至少4000、或更多合格样品。可能有利的是，使用较大集的合格样品来改善阈值的诊断效用。

[0102] 术语“bin”是指序列的片段或基因组的片段。在一些实施方式中，bin是彼此邻接并在基因组或染色体内通过位置分开。每个bin可以限定在参比基因组中核苷酸的序列。bin的尺寸可以是1kb、100kb、1Mb等，其取决于特定应用所需要的分析和序列标签密度。除它们在参比序列内的位置之外，bin可以具有其他特性如样品覆盖度和序列结构特性如G-C分数。

[0103] 术语“掩蔽阈值”在本文中用来指这样的量，相对于其来比较基于在序列bin中的序列标签的数目的值，其中具有超过掩蔽阈值的值的bin被掩蔽。在一些实施方式中，掩蔽阈值可以是百分等级、绝对计数、映射质量得分、或其他合适的值。在一些实施方式中，掩蔽阈值可被定义为整个多个未受影响的样品的变异系数的百分等级。在其他实施方式中，掩蔽阈值可被定义为映射质量得分，例如，MapQ得分，其涉及到比对序列读数与参比基因组的可靠性。注意，掩蔽阈值不同于拷贝数变异(CNV)阈值，后者是截止值以表征这样的样品，其含有来自疑似具有涉及到CNV的医学状况的生物体的核酸。在一些实施方式中，相对于在本文中别处描述的归一化染色体值(NCV)或归一化片段值(NSV)来定义CNV阈值。

[0104] 术语“归一化值”在本文中是指这样的数值，其使针对感兴趣的序列(例如染色体或染色体片段)确定的序列标签的数目相关与针对归一化序列(例如归一化染色体或归一化染色体片段)确定的序列标签的数目。例如，“归一化值”可以是如在本文中别处描述的染色体剂量，或它可以是NCV，或它可以是如在本文中别处描述的NSV。

[0105] 术语“读数”是指来自一部分核酸样品的序列读数。通常，虽然不一定，读数表示在样品中相邻碱基对的短序列。可以通过样品部分的碱基对序列(用ATCG)来用符号表示读数。它可以被存储在存储器件中并酌情被处理以确定它是否匹配参比序列或满足其他标准。读数可以直接获自测序仪器或间接获自涉及样品的存储的序列信息。在一些情况下，读数是足够长度(例如，至少约25个bp)的DNA序列，其可以用来确定较大序列或区，例如，其可以被比对以及具体地指定于染色体或基因组区或基因。

[0106] 术语“基因组读取”用来指在个体的整个基因组中任何片段的读数。

[0107] 术语“序列标签”在本文中与术语“映射的序列标签”互换使用以指这样的序列读数，通过比对，其已被具体地指定，即，映射至较大序列，例如，参比基因组。映射的序列标签被独特映射到参比基因组，即，它们被指定于到参比基因组的单个位置。除非另有规定，映射到在参比序列上的相同序列的标记被计数一次。可以作为数据结构或数据的其他集合来提供标记。在某些实施方式中，标记含有读数序列和上述读数的相关信息如序列在基因组中的位置，例如，在染色体上的位置。在某些实施方式中，位置指向正链方向。可以定义标记以在与参比基因组的比对中提供有限量的错配。在一些实施方式中，可以被映射到在参比基因组上的一个以上的位置的标记，即，并不独特映射的标记，可以不包括在分析中。

[0108] 术语“非冗余序列标签”是指并不映射到同一位点的序列标签，其被计数，借以在一些实施方式中确定归一化染色体值(NCV)。有时将多个序列读数比对于在参比基因组上的同样位置，从而产生多余的或重复的序列标签。在一些实施方式中，映射到同样位置的重

复序列标签被省略或计数为一个“非冗余序列标签”，借以确定NCV。在一些实施方式中，比对于未排除的位点的非冗余序列标签被计数以产生用于确定NCV的“非排除位点计数”(NES计数)。

[0109] 术语“位点”是指在参比基因组上的独特位置(即，染色体ID、染色体位置和方向)。在一些实施方式中，位点可以是在序列上残基、序列标签、或片段的位置。

[0110] “排除的位点”是在参比基因组的区中发现的已被排除的位点，借以计数序列标签。在一些实施方式中，排除的位点存在于含有重复序列的染色体的区，例如，着丝粒和端粒，以及是一个以上的染色体共有的染色体的区，例如，存在于Y染色体上的区，其还存在于X染色体上。

[0111] “未排除的位点”(NES)是在参比基因组中未排除的位点，借以计数序列标签。

[0112] “非排除位点计数”(NES计数)是被映射到在参比基因组上的NES的序列标签的数目。在一些实施方式中，NES计数是映射到NES的非冗余序列标签的数目。在一些实施方式中，覆盖度和相关参数如归一化覆盖度量、全局配置参数去除的覆盖度量、和染色体剂量是基于NES计数。在一个实例中，染色体剂量被计算为针对感兴趣的染色体的NES计数的数目与针对归一化染色体的NES计数的数目的比率。

[0113] 归一化染色体值(NCV)使测试样品的覆盖度相关与一组训练/合格样品的覆盖度。在一些实施方式中，NCV是基于染色体剂量。在一些实施方式中，NCV涉及到在测试样品中感兴趣的染色体的染色体剂量和在一组合格样品中相应染色体剂量的平均值之间的差异，其作为并可以被计算为：

[0114] $$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

[0115] 其中$\hat{\mu}_j$和$\hat{\sigma}_j$分别是估计的平均值和标准偏差，其是针对在一组合格样品中的第j个染色体剂量，以及x_{ij}是针对测试样品i所观测到的第j个染色体比率(剂量)。

[0116] 在一些实施方式中，可以通过使在测试样品中的感兴趣的染色体的染色体剂量相关于在用同样的流通池测序的多重样品中的相应染色体剂量的中位数来“在运行中”计算NCV，作为：

[0117] $$NCV_{ij} = \frac{x_{ij} - M_j}{\hat{\sigma}_j}$$

[0118] 其中M_j是在用同样的流通池测序的一组多重样品中针对第j染色体剂量的估计的中位数，$\hat{\sigma}_j$是在用一个或多个流动池测序的一组或多组多重样品中针对第j染色体剂量的标准偏差，以及x_j是针对测试样品i的观测到的第j染色体剂量。在此实施方式中，测试样品i是用同样的流通池测序的多重样品之一，据其确定M_j。

[0119] 例如，对于在测试样品A中的感兴趣的染色体21，其被测序为在一个流动池上的64个多重样品之一，针对在测试样品A中的染色体21的NCV被计算为在样品A中染色体21的剂量减去针对在64个多重样品中确定的染色体21的剂量的中位数，除以针对染色体21对于64个多重样品用流动池1、或另外的流动池例如20，确定的剂量的标准偏差。

[0120] 如在本文中所使用的，术语“比对的”、“比对”、或“比对”是指比较读数或标记与参

比序列并从而确定参比序列是否含有读数序列的过程。如果参比序列含有上述读数，则可以将上述读数映射到参比序列，或在某些实施方式中，映射到在参比序列中的特定位置。在一些情况下，比对简单地告诉，读数是否是特定参比序列的数目(即，在参比序列中读数是否是存在或不存在)。例如，读数与人染色体13的参比序列的比对将告诉，读数是否存在于染色体13的参比序列中。提供此信息的工具可被称为一组成员资格测试仪。在一些情况下，比对另外指示在参比序列中其中读数或标记映射到的位置。例如，如果参比序列是全人类基因组序列，则比对可以表明，读数是存在于染色体13上，并且可以进一步指示，读数是在染色体13的特定链和/或位点上。

[0121] 比对的读数或标记是一个或多个序列，就它们的核酸分子的顺序而言，其被确定为匹配于来自参比基因组的已知序列。可以手工完成比对，虽然通常通过计算机算法来实施，因为在合理的时间期限内将不可能比对读数以实施本文披露的方法。来自比对序列的算法的一个实例是核苷酸数据的有效的局部比对(ELAND)计算机程序，其被分布为Illumina Genomics Analysis流水线的一部分。可替换地，布隆过滤器或类似的集合成员资格测试仪可以用来比对读数与参比基因组。见2011年10月27日提交的美国专利申请号61/552,374，其全部内容以引用方式结合于本文。在比对中序列读数的匹配可以是100％序列匹配或小于100％(非完美匹配))。

[0122] 术语“比对分布图”用来指比对于位置的序列标签的分布，其可以被确定为在感兴趣的参比序列中的碱基对bin。

[0123] 在本文中使用的术语“映射”是指，通过比对，将序列读数具体地指派给较大序列，例如，参比基因组。

[0124] 如在本文中所使用的，术语“参比基因组”或“参比序列”是指任何生物体或病毒的任何特定的已知基因组序列(无论是部分的或完全的)，其可以用来提及来自受试者的确定的序列。例如，在国家生物技术信息中心(ncbi.nlm.nih.gov)具有用于人受试者以及许多其他生物体的参比基因组。“基因组”是指用核酸序列表达的生物体或病毒的完整的遗传信息。

[0125] 在不同的实施方式中，参比序列显著大于待比对于它的读数。例如，它可以是至少约100倍更大，或至少约1000倍更大，或至少约10,000倍更大，或至少约10^5倍更大，或至少约10^6倍更大，或至少约10^7倍更大。

[0126] 在一个实例中，参比序列是全长人类基因组的参比序列。这样的序列可被称为基因组参比序列。在另一个实例中，参比序列限于具体的人染色体如染色体13。在一些实施方式中，参比Y染色体是来自人类基因组版本hg19的Y染色体序列。这样的序列可被称为染色体参比序列。参比序列的其他实例包括其他物种的基因组、以及任何物种的染色体、亚染色体区(如链)等。

[0127] 在不同的实施方式中，参比序列是共有序列或来源于多个个体的其他组合。然而，在某些应用中，参比序列可以取自特定个体。

[0128] 术语“临床相关序列”在本文中是指已知或疑似关联或牵连于遗传或疾病状态的核酸序列。确定临床相关序列的存在或不存在可以用于确定或确认医学状况的诊断，或针对疾病的发展提供预后。

[0129] 当在核酸或核酸的混合物的情况下使用时，术语“衍生的”在本文中是指方式，借

此核酸获自它们从其所产生的源。例如，在一种实施方式中，来源于两个不同的基因组的核酸的混合物是指，核酸，例如，cfDNA，是由细胞通过天然存在的过程如坏死或凋亡所自然释放。在另一种实施方式中，来源于两种不同的基因组的核酸的混合物是指，核酸提取自受试者的两种不同类型的细胞。

[0130] 当在获得具体定量值的情况下使用时，术语“基于”在本文中是指利用另一个量作为输入来计算作为输出的具体定量值。

[0131] 术语“患者样品”在本文中是指这样的生物样品，其获自患者，即，医疗照顾、护理或治疗的接受者。患者样品可以是本文描述的任何样品。在某些实施方式中，通过非侵入性程序来获得患者样品，例如，外周血样品或粪便样品。本文描述的方法不需限于人类。因此，设想各种兽医应用，在这种情况下，患者样品可以是来自非人哺乳动物(例如，猫、猪、马、牛等)的样品。

[0132] 术语“混合样品”在本文中是指含有来源于不同的基因组的核酸的混合物的样品。

[0133] 术语“母体样品”在本文中是指这样的生物样品，其获自妊娠受试者，例如，妇女。

[0134] 术语“生物液体”在本文中是指取自生物源的液体并且包括，例如，血液、血清、血浆、痰、灌洗液、脑脊液、尿、精液、汗液、眼泪、唾液等。如在本文中所使用的，术语“血液”、“血浆”和“血清”明确地涵盖其部分或处理的部分。同样地，在样品取自活检、拭子、涂片等的情况下，“样品”明确地涵盖来源于活检、拭子、涂片等的处理过的部分。

[0135] 术语“母体核酸”和“胎儿核酸”在本文中分别是指妊娠雌性受试者的核酸和由妊娠雌性携带的胎儿的核酸。

[0136] 如在本文中所使用的，术语“对应于”有时是指核酸序列，例如，基因或染色体，其存在于不同受试者的基因组中，并且其在所有基因组中不一定具有相同序列，但用来提供同一性而不是感兴趣的序列，例如，基因或染色体，的遗传信息。

[0137] 如在本文中所使用的，连同所希望的样品一起使用的术语“基本上无细胞的”涵盖所希望的样品的制备，从其除去通常相关与样品的细胞成分。例如，通过除去通常相关与它的血细胞，例如，红细胞，来使血浆样品成为基本上无细胞的。在一些实施方式中，处理基本上无细胞的样品以除去否则将有助于待测试CNV的所希望的遗传物质的细胞。

[0138] 如在本文中所使用的，术语“胎儿分数”是指在包含胎儿和母体核酸的样品中存在的胎儿核酸的分数。胎儿分数经常用来表征在母体血液中的cfDNA。

[0139] 如在本文中所使用的，术语“染色体”是指活细胞的承载遗传的基因载体，其来源于染色质链，其包含DNA和蛋白质成分(尤其是组蛋白)。在本文中采用了传统的国际公认的个体人类基因组染色体编号系统。

[0140] 如在本文中所使用的，术语“多核苷酸长度”是指在序列中或在参比基因组的区中核酸分子(核苷酸)的绝对数目。术语“染色体长度”是指以碱基对给出的染色体的已知长度，例如，提供在人染色体的NCB136/hg18装配中，见在万维网上的|genome|.|ucsc|.|edu/cgi-bin/hgTracks?hgsid＝167155613&chromInfoPage＝。

[0141] 术语“受试者”在本文中是指人受试者以及非人受试者如哺乳动物、无脊椎动物、脊椎动物、真菌、酵母、细菌、和病毒。虽然本文的实施例关注人类以及语言主要涉及人文关怀，但本文披露的概念适用于来自任何植物或动物的基因组，并且可用于兽医领域、动物科学、研究实验室等等。

[0142] 术语“状况”在本文中是指“医学状况”，作为广义的术语，其包括所有疾病和病症，但可以包括[受伤]和正常的健康状况，如妊娠，其可能会影响人的健康、来自医疗帮助的益处，或对于医学治疗有影响。

[0143] 当参照染色体非整倍体使用时，术语“完全的”在本文中是指整个染色体的增益或损失。

[0144] 当参照染色体非整倍体使用时，术语“部分的”在本文中是指染色体的一部分，即，片段，的增益或损失。

[0145] 术语“嵌合(mosaic)”在本文中是指在已发育自单受精卵的一个个体中存在具有不同核型的细胞的两个群体。嵌合性可能源于在发育过程中的突变，其仅被传播到成体细胞的一个子集。

[0146] 术语“非嵌合(non-mosaic)”在本文中是指生物体，例如，人胎儿，其由一种核型的细胞组成。

[0147] 当参照确定染色体剂量使用时，术语“使用染色体”在本文中是指使用针对染色体获得的序列信息，即，针对染色体获得的序列标签的数目。

[0148] 如在本文中所使用的，术语“灵敏度”等于真阳性的数目除以真阳性和假阴性的总和。

[0149] 如在本文中所使用的，术语“特异性”等于真阴性的数目除以真阴性和假阳性的总和。

[0150] 术语“富集”在本文中是指以下过程：扩增包含在部分母体样品中的多态目标核酸，并结合扩增产物与从其除去部分的母体样品的剩余物。例如，母体样品的剩余物可以是原始母体样品。

[0151] 术语“原始母体样品”在本文中是指获自妊娠受试者的非富集的生物样品，例如，妇女，其作为从其除去部分以扩增多态目标核酸的源。“原始样品”可以是获自妊娠受试者的任何样品、以及其经处理的部分，例如，提取自母体血浆样品的纯化的cfDNA样品。

[0152] 如在本文中所使用的，术语“引物”是指分离的寡核苷酸，当放置在诱导延伸产物的合成的条件下时(例如，上述条件包括核苷酸、诱导剂如DNA聚合酶、以及适宜的温度和pH)，其能够作为合成的引发点。引物优选是单链，以获得最大扩增效率，但可以可替换地是双链。如果是双链，则首先处理引物以在用来制备延伸产物之前分离其链。优选地，引物是寡脱氧核苷酸。引物必须足够长以在诱导剂的存在下引发延伸产物的合成。引物的确切长度将取决于许多因素，包括温度、引物源、方法的使用、以及用于引物设计的参数。

[0153] 短语“导致待给予”是指由医疗专业人士(例如，医师)或控制或指导受试者的医疗保健的人采取的操作，其控制和/或允许将所考虑的剂/化合物给予受试者。导致待给予可能涉及诊断和/或适当的治疗或预防方案的确定，和/或为受试者规定特定剂/化合物。这样的规定可以包括，例如，起草处方组成、注解医疗记录等。同样地，“导致待实施”，例如，对于诊断程序，是指由医疗专业人士(例如，医师)或控制或指导受试者的医疗保健的人采取的操作，其控制和/或允许对受试者执行一个或多个诊断方法方法。

[0154] 引言

[0155] 本文披露了方法、装置、和系统，用于确定在测试样品中感兴趣的不同序列的拷贝数和拷贝数变异(CNV)，其中上述测试样品包含来源于两个或更多个不同的基因组的核酸

的混合物，并且其已知或被疑似在一个或多个感兴趣的序列的量上有所不同。通过本文披露的方法和仪器确定的拷贝数变异包括整个染色体的增益或损失，涉及显微镜下可见的非常大的染色体片段的变化，以及DNA片段的亚微观拷贝数变异的丰度，在尺寸方面，其范围为单核苷酸至千碱基(kb)、至兆碱基(Mb)。

[0156] 在一些实施方式中，提供了利用含有母体和胎儿无细胞DNA的母体样品来确定胎儿的拷贝数变异(CNV)的方法。本文披露的一些实施方式提供了通过除去样品中GC含量偏差来改善序列数据分析的灵敏度和/或特异性的方法。在一些实施方式中，样品中GC含量偏差的除去是基于针对通用于不受影响的训练样品的系统性变化修正的序列数据。

[0157] 披露的一些实施方式提供了低噪声和高信号地确定序列覆盖度量的方法，从而提供数据来确定涉及到拷贝数和CNV的各种遗传病症，相对于通过常规方法所获得的序列覆盖度量，其具有改善的灵敏度、选择性、和/或效率。已发现上文描述的方法，在具有相对低分数的来自在考虑中的基因组(例如，胎儿的基因组)的DNA的样品中，特别有效地改善信号。这样的样品的一个实例是来自怀上假性双胞胎、三胎等的个体的母体血液样品，其中所述方法评价在胎儿之一的基因组中的拷贝数变异。

[0158] 所述方法适用于确定任何胎儿非整倍性的CNV，以及已知或疑似相关与各种各样的医学状况的CNV。在涉及人受试者的一些实施方式中，根据本方法可以确定的CNV包括任何一个或多个染色体1-22、X和Y的三体性和单体性、其他染色体多体性、以及任何一个或多个染色体的片段的缺失和/或复制，其可以通过测序测试样品的核酸仅一次加以检测。任何非整倍体可以确定自通过测序测试样品的核酸仅一次所获得的测序信息。

[0159] 在人类基因组中的CNV显著影响人类多样性和对疾病的易感性(Redon等人，Nature 23:444-454[2006]，Shaikh等人，Genome Res19:1682-1690[2009])。CNV已经被知道通过不同的机制有助于遗传疾病，从而导致基因剂量的失衡或基因破坏(在大多数情况下)。除它们直接相关与遗传紊乱之外，已知CNV会调节可能是有害的表型改变。最近，若干研究已报道了，当相比于正常对照时，在杂的疾病如自闭症、ADHD、和精神分裂症中稀有或从头CNV的增加的负担，其强调了稀有或特有CNV的潜在的致病性(Sebat等人，316:445-449[2007]；Walsh等人，Science 320:539-543[2008])。CNV产生于基因组重排，其主要是由于缺失、复制、插入、和不平衡的易位事件。

[0160] 本文描述的方法和仪器可以采用下一代测序技术(NGS)，其是大规模平行测序。在某些实施方式中，在流动池内，以大规模平行方式来测序克隆扩增的DNA模板或单个DNA分子(例如在Volkerding等人.Clin Chem55:641-658[2009]；Metzker M Nature Rev 11:31-46[2010]中所描述的)。除高通量序列信息之外，NGS还提供定量信息，这是因为每个序列读数是表示个体克隆DNA模板或单个DNA分子的可计数“序列标签”。NGS的测序技术包括焦磷酸测序、借助于可逆染料终止子的合成测序、通过寡核苷酸探针连接的测序、和离子半导体测序。在单测序运行中，可以单独测序(即，单重测序)来自个体样品的DNA或可以汇集来自多个样品的DNA并测序为索引基因组分子(即，多重测序)，以产生DNA序列的高达数亿读数。下文描述了根据本方法可以用来获得序列信息的测序技术的实例。

[0161] 利用DNA样品的各种CNV分析涉及将来自测序仪的序列读数比对或映射到参比序列。参比序列可以是全基因组的序列、染色体的序列、亚染色体区的序列等。相比于常染色体，由于参比序列的特性，Y染色体的CNV的诊断涉及高难度的技术挑战，这是因为Y染色体

的覆盖度低于常染色体的覆盖度，并且在Y染色体上的重复序列复杂化读数映射到它们的正确位置。存在通过目前NGS技术可访问的约10Mb的独特的Y序列，但在胎儿诊断中性别检测仍然是具有挑战性的任务，其中在母体样品中胎儿cfDNA的量是至少数量级低于母体DNA的量，从而强调了非特异性映射的问题。

[0162] 另外，目前一些测序方法利用超短读取如25聚体读取和标记。在测序方法的方法中利用的超短测序产生短读数长度，其对于序列比对提出了技术挑战，这是因为近一半的人类基因组被重复序列覆盖，其中它们中的许多已经知道大约几十年。从计算的角度来看，在比对中重复序列产生歧义，其转而甚至在全染色体计数水平下也可以产生偏差和差错。

[0163] 评价CNV

[0164] 用于确定CNV的方法

[0165] 相对于利用通过常规方法获得的序列覆盖度值，利用通过本文披露的方法提供的序列覆盖度值，可以确定涉及到序列、染色体、或染色体片段的拷贝数和CNV的各种遗传病症，并具有改善的灵敏度、选择性、和/或效率。例如，在一些实施方式中，掩蔽的参比序列用于在包含胎儿和母体核酸分子的母体测试样品中确定任何两种或更多种不同的完整胎儿染色体非整倍体的存在或不存在。以下提供的示例性方法比对读数与参比序列(包括参比基因组)。可以对未掩蔽的或掩蔽的参比序列进行比对，从而产生映射到参比序列的序列标签。在一些实施方式中，考虑到仅落在参比序列的未掩蔽的片段上的序列标签来确定拷贝数变异。

[0166] 在一些实施方式中，用于在母体测试样品中确定任何完整的胎儿染色体非整倍体的存在或不存在的方法包括(a)在母体测试样品中获得胎儿和母体核酸的序列信息；(b)利用上文描述的序列信息和方法来确定，针对选自染色体1-22、X和Y的每个感兴趣的染色体，从其衍生的序列标签的数目或序列覆盖度量，以及确定针对一个或多个归一化染色体序列的序列标签的数目；(c)利用针对每个感兴趣的染色体确定的序列标签的数目和针对每个归一化染色体确定的序列标签的数目来计算针对每个感兴趣的染色体的单染色体剂量；以及(d)比较每个染色体剂量与阈值，并从而确定在母体测试样品中任何完整的胎儿染色体非整倍体的存在或不存在。

[0167] 在一些实施方式中，上面描述的步骤(a)可以包括测序测试样品的至少一部分的核酸分子以获得关于测试样品的胎儿和母体核酸分子的所述序列信息。在一些实施方式中，步骤(c)包括将每个感兴趣的染色体的单染色体剂量计算为针对每个感兴趣的染色体确定的序列标签的数目和针对归一化染色体序列确定的序列标签的数目的比率。在一些其他实施方式中，染色体剂量是基于来源于序列标签的数目的处理的序列覆盖度量。在一些实施方式中，仅独特的非冗余序列标签用来计算处理的序列覆盖度量。在一些实施方式中，处理的序列覆盖度量是序列标签密度比，其是通过序列长度加以归一化的序列标签的数目。在一些实施方式中，处理的序列覆盖度量是归一化的序列标签，其是感兴趣的序列的序列标签的数目除以所有或主要部分的基因组。在一些实施方式中，根据感兴趣的序列的全局配置参数来调节处理的序列覆盖度量。在一些实施方式中，根据在针对待测试样品的GC含量和序列覆盖度之间的在样品内相关性来调节处理的序列覆盖度量。在一些实施方式中，处理的序列覆盖度量产生于这些过程的组合，其在本文中别处进一步加以描述。

[0168] 在一些实施方式中，染色体剂量被计算为针对每个感兴趣的染色体的处理的序列

覆盖度量和针对归一化染色体序列的处理的序列覆盖度量的比率。

[0169] 在上述任何一种实施方式中，完全染色体非整倍体选自完全染色体三体性、完全染色体单体性和完全染色体多体性。完全染色体非整倍体选自染色体1-22、X、和Y中的任一种的完全非整倍体。例如，所述不同的完全胎儿染色体非整倍体选自三体性2、三体性8、三体性9、三体性20、三体性21、三体性13、三体性16、三体性18、三体性22、47，XXX、47，XYY、和单体性X。

[0170] 在上述任何一种实施方式中，针对来自不同母体受试者的测试样品重复步骤(a)-(d)，以及所述方法包括确定在每个测试样品中任何两种或更多种不同的完整胎儿染色体非整倍体的存在或不存在。

[0171] 在上述任何一种实施方式中，所述方法可以进一步包括计算归一化染色体值(NCV)，其中NCV使染色体剂量相关与在一组合格样品中相应染色体剂量的平均值，作为：

[0172] $$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

[0173] 其中$\hat{\mu}_j$和$\hat{\sigma}_j$分别是针对在一组合格样品中的第j染色体剂量的估计的平均值和标准偏差，以及x_{ij}是针对测试样品i的观测到的第j染色体剂量。

[0174] 在一些实施方式中，通过使在测试样品中的感兴趣的染色体的染色体剂量相关于在用同样的流通池测序的多重样品中相应染色体剂量的中位数，可以“在运行中”计算NCV，作为：

[0175] $$NCV_{ij} = \frac{x_{ij} - M_j}{\hat{\sigma}_j}$$

[0176] 其中M_j是在用同样的流通池测序的一组多重样品中第j染色体剂量的估计的中位数；$\hat{\sigma}_j$是在用一个或多个流动池测序的一组或多组多重样品中第j染色体剂量的标准偏差；以及x_i是针对测试样品i的所观测到第j染色体剂量。在此实施方式中，测试样品i是用同样的流通池(从其确定M_j)测序的多重样品之一。

[0177] 在一些实施方式中，提供了用于确定在包含胎儿和母体核酸的母体测试样品中不同的部分胎儿染色体非整倍体的存在或不存在的方法。所述方法涉及类似于用于检测完全非整倍体(如上所述)的方法的程序。然而，代替分析完整的染色体，分析染色体的片段。见美国专利申请公开号2013/0029852，其以引用方式结合于本文。

[0178] 图1示出了根据一些实施方式用于确定拷贝数变异的存在的方法。在操作130和135中，确定合格序列标签覆盖度和测试序列标签覆盖度。本公开内容提供了用来确定覆盖度量的过程，相对于常规方法，其提供了改善的灵敏度和选择性。用星号来标示操作130和135并通过重线方框加以强调以表示这些操作有助于相对于现有技术的改善。在一些实施方式中，归一化、调节、修剪、和以其他方式处理序列标签覆盖度量以改善分析的灵敏度和选择性。在本文中别处进一步描述这些过程。

[0179] 从概观的角度看，所述方法在测试样品的CNV的确定中利用了合格训练样品的归一化序列。在一些实施方式中，合格训练样品是未受影响的并具有正常的拷贝数。归一化序列提供了机制来归一化用于运行内和运行间变异性的测量。利用来自获自受试者的一组合格样品的序列信息来确定归一化序列，其中上述受试者已知包含对于任何一个感兴趣的序

列，例如，染色体或其片段，具有正常的拷贝数的细胞。归一化序列的确定概述于在图1中描述的方法的实施方式的步骤110、120、130、145和146。在一些实施方式中，归一化序列用来计算针对测试序列的序列剂量。见步骤150。在一些实施方式中，归一化序列还用来计算阈值，相对于其来比较测试序列的序列剂量。见步骤150。获自归一化序列和测试序列的序列信息用于确定在测试样品中染色体非整倍体的有统计学意义的识别(步骤160)。

[0180] 转向根据一些实施方式用于确定拷贝数变异的存在的方法的细节，图1提供了用于在生物样品中确定感兴趣的序列，例如，染色体或其片段，的CNV的一种实施方式的流程图100。在一些实施方式中，生物样品获自受试者并且包含由不同的基因组贡献的核酸的混合物。不同的基因组可以促成两个个体的样品，例如，不同的基因组是由胎儿和携带胎儿的母体所贡献。另外，不同的基因组可以促成三个或更多个体的样品，例如，不同的基因组是由两个或更多胎儿和携带胎儿的母体所贡献。可替换地，基因组有助于来自相同受试者的非整倍体癌细胞和正常整倍体细胞的样品，例如，来自癌症患者的血浆样品。

[0181] 除了分析患者的测试样品，一个或多个归一化染色体或一个或多个归一化染色体片段被选择用于感兴趣的每种可能的染色体。以异步方式，归一化染色体或片段确定自患者样品的正常测试，其可以发生在临床环境中。换句话说，在测试患者样品之前，确定归一化染色体或片段。存储在归一化染色体或片段和感兴趣的染色体或片段之间的关联，以在测试过程中使用。如下面所解释的，通常在跨越许多样品的测试的一段时间内保持这样的关联。下面的讨论涉及用于针对感兴趣的个体染色体或片段选择归一化染色体或染色体片段的实施方式。

[0182] 获得一组合格样品以识别合格归一化序列以及提供用于在测试样品中确定CNV的有统计学意义的识别的方差值。在步骤110中，多个生物合格样品获自多个受试者，其已知包含针对感兴趣的任何一个序列具有正常拷贝数的细胞。在一种实施方式中，合格样品获自怀上胎儿的母体，利用细胞遗传学方法，其已被证实具有染色体的正常拷贝数。生物合格样品可以是生物液体，例如，血浆，或如下所述的任何合适的样品。在一些实施方式中，合格样品含有核酸分子的混合物，例如，cfDNA分子。在一些实施方式中，合格样品是母体血浆样品，其含有胎儿和母体cfDNA分子的混合物。通过，利用任何已知的测序方法，测序至少一部分的核酸，例如，胎儿和母体核酸，来获得针对归一化染色体和/或其片段的序列信息。优选地，在本文中别处描述的下一代测序(NGS)方法的任何一种用来测序作为单或克隆扩增分子的胎儿和母体核酸。在不同的实施方式中，如下文在测序之前和期间所披露的，处理合格样品。可以利用如在本文中披露的仪器、系统、和试剂盒来处理它们。

[0183] 在步骤120中，测序包含在合格样品中的至少一部分的每个所有的合格核酸以产生数以百万计的序列读数，例如，36bp的读数，其被比对于参比基因组，例如，hg18。在一些实施方式中，序列读数包括约20bp、约25bp、约30bp、约35bp、约40bp、约45bp、约50bp、约55bp、约60bp、约65bp、、约70bp、约75bp、约80bp、约85bp、约90bp、约95bp、约100bp、约110bp、约120bp、约130、约140bp、约150bp、约200bp、约250bp、约300bp、约350bp、约400bp、约450bp、或约500bp。预期，当产生配对末端读数时，技术进步将使得能够产生大于500bp的单端读数，从而使得能够产生大于约1000bp的读数。在一种实施方式中，映射的序列读数包含36bp。在另一种实施方式中，映射的序列读数包含25bp。

[0184] 将序列读数比对于参比基因组，以及被独特映射到参比基因组的读数被称为序列

标签。对于CNV的分析，没有计数落在掩蔽的参比序列的掩蔽片段上的序列标签。

[0185] 在一种实施方式中，至少约3×10^6个合格序列标签、至少约5×10^6个合格序列标签、至少约8×10^6合个格序列标签、至少约10×10^6个合格序列标签、至少约15×10^6个合格序列标签、至少约20×10^6个合格序列标签、至少约30×10^6个合格序列标签、至少约40×10^6个合格序列标签、或至少约50×10^6个合格序列标签(包含20至40bp读数)获自独特映射到参比基因组的读数。

[0186] 在步骤130中，计数获自测序在合格样品中的核酸的所有标记以获得合格序列标签覆盖度。同样地，在操作135中，计数获自测试样品的所有标记以获得测试序列标签覆盖度。本公开内容提供了方法来确定覆盖度量，相对于常规方法，其提供改善的灵敏度和选择性。操作130和135是通过星号加以标示并通过重线方框加以强调以表示这些操作有助于相对于现有技术的改善。在一些实施方式中，归一化、调节、修剪、和以其他方式处理序列标签覆盖度量，以改善分析的灵敏度和选择性。在本文中别处进一步描述这些过程。

[0187] 当在每个合格样品中映射和计数所有合格序列标签时，确定在合格样品中的感兴趣的序列，例如，临床相关序列，的序列标签覆盖度，如确定据其随后确定归一化序列的另外序列的序列标签覆盖度。

[0188] 在一些实施方式中，感兴趣的序列是相关与完全染色体非整倍体的染色体，例如，染色体21，以及合格归一化序列是完整的染色体，其并不相关与染色体非整倍体并且其在序列标签覆盖度方面的变化接近感兴趣的序列(即，染色体)，例如，染色体21，的序列标签覆盖度变化。所选的归一化染色体可以是最好接近感兴趣的序列的序列标签覆盖度的变化的一个或一组。染色体1-22、X、和Y的任何一个或多个可以是感兴趣的序列，并且一个或多个染色体可以确定为在合格样品中对于每个的任何一个染色体1-22、X和Y的归一化序列。归一化染色体可以是个体染色体或它可以是一组染色体(如在本文中别处描述的)。

[0189] 在另一种实施方式中，感兴趣的序列是相关与部分非整倍体的染色体的片段，例如，染色体缺失或插入，或不平衡的染色体易位，以及归一化序列是不相关与部分非整倍体的染色体片段(或片段的组)并且其在序列标签覆盖度方面的变化接近相关与部分非整倍体的染色体片段的序列标签覆盖度变化。所选的归一化染色体片段可以是一种或多种归一化染色体片段，其最好接近感兴趣的序列的序列标签覆盖度的变化。任何一个或多个染色体1-22、X、和Y的任何一个或多个片段可以是感兴趣的序列。

[0190] 在其他实施方式中，感兴趣的序列是相关与部分非整倍体的染色体的片段以及归一化序列是全染色体或染色体。在另外的其他实施方式中，感兴趣的序列是相关与非整倍体的全染色体以及归一化序列是不相关与非整倍体的染色体片段。

[0191] 在合格样品中，单个序列或一组序列被确定为针对感兴趣的任何一个或多个序列的归一化序列，可以选择合格归一化序列以具有在序列标签覆盖度方面的变化，其最好或有效地接近如在合格样品中确定的感兴趣的序列的序列标签覆盖度变化。例如，合格归一化序列是这样的序列，当用来归一化感兴趣的序列时，其产生整个合格样品的最小变异性，即，归一化序列的变异性最靠近在合格样品中确定的感兴趣的序列的变异性。换句话说，合格归一化序列是这样的序列，其选来产生整个合格样品的序列剂量(对于感兴趣的序列)的最小变化。因此，所述方法选择这样的序列，当用作归一化染色体时，对于感兴趣的序列，其预期产生在运行到运行染色体剂量中的最小变异性。

[0192] 在合格样品中针对任何一个或多个感兴趣的序列确定的归一化序列仍然是在数天、数周、数月、并且可能数年内选择用于确定在测试样品中非整倍体的存在或不存在的归一化序列，只要，为产生测序文库所需要的程序以及测序样品随时间基本不变。如上面描述的，用于确定非整倍体的存在的归一化序列被选择用于(可能还有其他原因)在样品中被映射到它的序列标签的数目的变异性，例如，不同样品，以及测序运行，例如，发生在同一天和/或不同天的测序运行，其最好接近对其他用作归一化参数的感兴趣的序列的变异性。这些程序的实质性变化将影响被映射到所有序列的标记的数目，其转而将确定序列的哪一个或哪一组将具有在相同和/或在不同的测序运行中并在同一天或在不同天整个样品的变异性，其最紧密接近感兴趣的序列的变异性，其将要求重新确定归一化序列的组。程序的实质性变化包括用于制备测序文库的实验室方法的变化，其包括涉及到为多重测序而不是单重测序而制备样品的变化，以及测序平台的变化，其包括用于测序的化学作用的变化。

[0193] 在一些实施方式中，选择用来归一化感兴趣的特定序列的归一化序列是这样的序列，其最好区分一个或多个合格样品与一个或多个受影响的样品，这意味着，归一化序列是这样的序列，其具有最大可微性(differentiability)，即，归一化序列的可微性是如此，以致它提供与在受影响的测试样品中感兴趣的序列的最佳区别，从而容易区分受影响的测试样品与其他未受影响的样品。在其他实施方式中，归一化序列是具有最小变异性和最大可微性的组合的序列。

[0194] 可微性的水平可以确定为在合格样品的群体中的序列剂量，例如，染色体剂量或片段剂量，和在如下所述的和示于实施例中的一个或多个测试样品中的染色体剂量之间的统计差异。例如，可微性可以数值表示为t检验值，其表示在合格样品的群体中的染色体剂量和在一个或多个测试样品中的染色体剂量之间的统计差异。同样地，可微性可以基于片段剂量而不是染色体剂量。可替换地，可微性可以数值表示为归一化染色体值(NCV)，其是用于染色体剂量的z得分，只要NCV的分布是正常的。同样地，在其中染色体片段是感兴趣的序列的情况下，片段剂量的可微性可以数值表示为归一化片段值(NSV)，其是用于染色体片段剂量的z得分，只要NSV的分布是正常的。在确定z得分时，可以使用在一组合格样品中的染色体或片段剂量的平均值和标准偏差。可替换地，可以使用在包含合格样品和受影响的样品的训练集中的染色体或片段剂量的平均值和标准偏差。在其他实施方式中，归一化序列是具有最小变异性和最大可微性或小变异性和大可微性的最佳组合的序列。

[0195] 所述方法确定这样的序列，其固有地具有类似特性并且其在样品和测序运行中容易出现类似的变化，以及其可用于确定在测试样品中的序列剂量。

[0196] 序列剂量的确定

[0197] 在一些实施方式中，在所有合格样品中确定针对感兴趣的一个或多个染色体或片段的染色体或片段剂量(如在示于图1的步骤146中所描述的)，以及在步骤145中确定归一化染色体或片段序列。在计算序列剂量之前，提供一些归一化序列。然后根据如下文进一步描述的各种标准来确定一个或多个归一化序列，见步骤145。在一些实施方式中，例如，确定的归一化序列导致，对于整个所有合格样品的感兴趣的序列，序列剂量的最小变异性。

[0198] 在步骤146中，基于计算的合格标签密度，对于感兴趣的序列的合格序列剂量，即，染色体剂量或片段剂量，被确定为用于感兴趣的序列的序列标签覆盖度和用于另外的序列的合格序列标签覆盖度的比率，据此在步骤145中随后确定归一化序列。确定的归一化序列

随后用来确定在测试样品中的序列剂量。

[0199] 在一种实施方式中，在合格样品中的序列剂量是染色体剂量，其计算为在合格样品中针对感兴趣的染色体的序列标签的数目和针对归一化染色体序列的序列标签的数目的比率。归一化染色体序列可以是单染色体、一组染色体、一个染色体的片段、或来自不同染色体的一组片段。因此，在合格样品中对于感兴趣的染色体的染色体剂量被确定为针对感兴趣的染色体的标记的数目和针对下述序列的标记的数目的比率：(i)由单染色体组成的归一化染色体序列，(ii)由两个或两个以上的染色体组成的归一化染色体序列，(iii)由染色体的单片段组成的归一化片段序列，(iv)由来自一个染色体的两个或更多片段组成的归一化片段序列，或(v)由两个或两个以上的染色体的两个或更多片段组成的归一化片段序列。用于根据(i)-(v)来确定针对感兴趣的染色体21的染色体剂量的实施例是如下：针对感兴趣的染色体，例如，染色体21，的染色体剂量被确定为染色体21的序列标签覆盖度和以下序列标签覆盖度之一的比率：(i)每个所有剩余的染色体，即，染色体1-20、染色体22、X染色体、和Y染色体；(ii)两个或更多的剩余染色体的所有可能的组合；(iii)另一染色体的片段，例如，染色体9；(iv)另一个染色体的两个片段，例如，染色体9的两个片段；(v)两个不同的染色体的两个片段，例如，9号染色体的片段和14号染色体的片段。

[0200] 在另一种实施方式中，在合格样品中的序列剂量是片段剂量而不是染色体剂量，上述片段剂量被计算为在合格样品中，对于感兴趣的片段，其不是全染色体，的序列标签的数目和对于归一化片段序列的序列标签的数目的比率。归一化片段序列可以是以上讨论的任何归一化染色体或片段序列。

[0201] 归一化序列的鉴定

[0202] 在步骤145中，针对感兴趣的序列，确定归一化序列。在一些实施方式中，例如，归一化序列是基于计算的序列剂量的序列，例如，其导致对于整个所有合格训练样品的感兴趣的序列的序列剂量的最小变异性。所述方法确定这样的序列，其固有地具有类似特性并且在样品和测序运行中容易出现类似的变化，以用其可用于确定在测试样品中的序列剂量。

[0203] 可以在一组合格样品中确定针对一个或多个感兴趣的序列的归一化序列，以及在合格样品中确定的序列随后用来计算针对在每个测试样品中的一个或多个感兴趣的序列的序列剂量(步骤150)，以确定在每个测试样品中非整倍体的存在或不存在。当使用不同的测序平台时和/或当在待测序的核酸的纯化和/或测序文库的制备中存在差异时，针对感兴趣的染色体或片段确定的归一化序列可能不同。根据本文描述的方法的归一化序列的使用提供了染色体或其片段的拷贝数的变化的特异性的和敏感性的度量，而不论样品制备和/或所使用的测序平台。

[0204] 在一些实施方式中，确定一个以上的归一化序列，即，针对一个感兴趣的序列，可以确定不同的归一化序列，以及针对一个感兴趣的序列可以确定多个序列剂量。例如，当使用14号染色体的序列标签覆盖度时，在针对感兴趣的染色体21的染色体剂量中，变化，例如，变异系数(CV＝标准偏差/平均值)，是最小的。然而，可以确定二、三、四、五、六、七、八或更多的归一化序列，用于确定针对在测试样品中的感兴趣的序列的序列剂量。作为例子，利用7号染色体、9号染色体、11号染色体或12号染色体作为归一化染色体序列，可以确定在任何一个测试样品中针对染色体21的第二剂量，因为这些染色体均具有接近14号染色体的CV

的CV。

[0205] 在一些实施方式中，当单染色体被选择为用于感兴趣的染色体的归一化染色体序列时，归一化染色体序列将是这样的染色体，其导致用于感兴趣的染色体的染色体剂量，其具有整个所有测试样品的最小变异性，例如，合格样品。在一些情况下，最好的归一化染色体可以不具有最小变化，但可以具有合格剂量的分布，其最好区分测试样品或来自合格样品的样品，即，最好的归一化染色体可以不具有最低的变化，但可以具有最大可微性。

[0206] 在一些实施方式中，归一化序列包括一个或多个鲁棒常染色体序列或它们的片段。在一些实施方式中，鲁棒常染色体包括除感兴趣的染色体之外的所有常染色体。在一些实施方式中，鲁棒常染色体包括除chr X、Y、13、18、和21之外的所有常染色体。在一些实施方式中，鲁棒常染色体包括除那些确定自将偏离自正常二倍体状态的样品的常染色体之外的所有常染色体，其可以用于确定相对于正常二倍体基因组具有异常拷贝数的癌症基因组。

[0207] 在测试样品中非整倍体的确定

[0208] 基于在合格样品中归一化序列的鉴定，针对在测试样品中的感兴趣的序列，确定序列剂量，其中上述测试样品包含来源于在一个或多个感兴趣的序列方面有所不同的基因组的核酸的混合物。

[0209] 在步骤115中，测试样品获自疑似或已知携带感兴趣的序列的临床相关CNV的受试者。测试样品可以是生物液体，例如，血浆、或如下所述的任何合适的样品。如所解释的，可以利用非侵入性程序如简单的抽血来获得样品。在一些实施方式中，测试样品含有核酸分子的混合物，例如，cfDNA分子。在一些实施方式中，测试样品是含有胎儿和母体cfDNA分子的混合物的母体血浆样品。

[0210] 在步骤125中，如针对合格样品所描述的，测序在测试样品中的至少一部分的测试核酸，以产生数以百万计的序列读数，例如，36bp的读数。如在步骤120中，产生自测序在测试样品中的核酸的读数被独特地映射或比对于参比基因组以产生标记。如在步骤120中所描述的，至少约3×10^6个合格序列标签、至少约5×10^6个合格序列标签、至少约8×10^6个合格序列标签、至少约10×10^6个合格序列标签、至少约15×10^6个合格序列标签、至少约20×10^6个合格序列标签、至少约30×10^6个合格序列标签、至少约40×10^6个合格序列标签、或至少约50×10^6个合格序列标签(包含20至40bp读数)获自独特映射到参比基因组的读数。在某些实施方式中，以电子格式提供通过测序仪器产生的读数。利用如下文所讨论的计算仪器来完成比对。相对于参比基因组来比较个别读数，其常常是庞大的(数以百万计碱基对)，以识别其中读数唯一地对应于参比基因组的位点。在一些实施方式中，比对程序允许在读数和参比基因组之间的有限的错配。在一些情况下，允许在读数中的1、2、或3个碱基对错配在参比基因组中的相应的碱基对，但依然取得映射。

[0211] 在步骤135中，利用如下所述的计算仪器，获自测序在测试样品中的核酸的全部或大部分的标记被计数以确定测试序列标签覆盖度。在一些实施方式中，使每个读数比对于参比基因组的特定区(染色体或片段，在大多数情况下)，并通过追加位点信息于读数，将读数转换为标记。当此过程展开时，计算仪器可能保持映射到参比基因组的每个区(染色体或片段，在大多数情况下)的标记读数的数目的运行计数。针对感兴趣的每个染色体或片段以及每个相应的归一化染色体或片段存储计数。

[0212] 在某些实施方式中，参比基因组具有一个或多个排除区，其是真正的生物基因组的一部分但并不包括在参比基因组中。并不计数潜在比对于这些排除区的读数。排除区的实例包括长重复序列的区、在X和Y染色体之间具有相似性的区等。利用通过上面描述的掩蔽技术获得的掩蔽的参比序列，仅在参比序列的未掩蔽的片段上的标记被考虑到用于CNV的分析。

[0213] 在一些实施方式中，当多个读数比对于在参比基因组或序列上的同一位点时，所述方法确定是否计数标记一次以上。可能存在这样的情况，其时两个标记具有相同序列并因而比对于在参比序列上的相同位。用来计数标记的方法，在某些情况下，可能从计数排除来自相同的测序样品的相同标记。如果在给定样品中标记的不成比例的数目是相同的，则它提示，在程序中存在强烈的偏差或其他缺陷。因此，依照某些实施方式，上述计数方法不计数来自给定样品的标记，其是相同于来自被先前计数的样品的标记。

[0214] 当忽视来自单个样品的相同标记时，可以设定用于选择的各种标准。在某些实施方式中，被计数的标记的定义的百分比必须是唯一的。如果比此阈值更多的标记不是唯一的，则忽视它们。例如，如果定义的百分比需要至少50％是独特的，则不计数相同标记，直到对于样品，独特标记的百分比超过50％，在其他实施方式中，独特标记的阈值数目是至少约60％。在其他实施方式中，独特标记的阈值百分比是至少约75％、或至少约90％、或至少约95％、或至少约98％、或至少约99％。对于21号染色体，阈值可以设定为90％。如果30M标记被比对于21号染色体，那么它们的至少27M必须是唯一的。如果3M计数的标记不是唯一的以及3000万和第一标记不是唯一的，则它不被计数。可以利用适当的统计分析来选择用来确定何时不计数另外相同标记的特定阈值或其他标准。影响此阈值或其他标准的一个因素是相对于标记可以与其比对的基因组的尺寸，测序样品的相对量。其他因素包括读数的大小和类似的考虑。

[0215] 在一种实施方式中，映射到感兴趣的序列的测试序列标签的数目被归一化到它们被映射到的感兴趣的序列的已知长度，以提供测试序列标签密度比。如针对合格样品所描述的，归一化到感兴趣的序列的已知长度是不需要的，并且可以被包括为用来减少数目的位数的步骤，以简化它，从而便于人解释。因为计数在测试样品中的所有映射的测试序列标签，所以确定针对在测试样品中的感兴趣的序列，例如，临床相关序列，的序列标签覆盖度，如确定针对对应于在合格样品中确定的至少一个归一化序列的另外的序列的序列标签覆盖度。

[0216] 在步骤150中，基于在合格样品中至少一个归一化序列的同一性，在测试样品中确定针对感兴趣的序列的测试序列剂量。在不同的实施方式中，利用感兴趣的序列和相应的归一化序列(如本文所描述的)的序列标签覆盖度来计算确定测试序列剂量。负责这项工作的计算仪器将电子访问在感兴趣的序列和它的相关的归一化序列之间的关联，其可以被存储在数据库、表、图形中，或被包括为在程序指令中的代码。

[0217] 如在本文中别处描述的，至少一个归一化序列可以是单个序列或一组序列。对于在测试样品中的感兴趣的序列的序列剂量是针对在测试样品中的感兴趣的序列确定的序列标签覆盖度和在测试样品中确定的至少一个归一化序列的序列标签覆盖度的比率，其中在测试样品中的归一化序列对应于在合格样品中针对感兴趣的特定序列确定的归一化序列。例如，如果在合格样品中针对21号染色体确定的归一化序列被确定为染色体，例如，14

号染色体，那么对于21号染色体(感兴趣的序列)的测试序列剂量被确定为各自在测试样品中确定的对于21号染色体的序列标签覆盖度和对于14号染色体的序列标签覆盖度的比率。同样地，确定了针对染色体13、18、X、Y、和相关与染色体非整倍体的其他染色体的染色体剂量。用于感兴趣的染色体的归一化序列可以是一个或一组染色体、或一个或一组染色体片段。如先前所描述的，感兴趣的序列可以是部分染色体，例如，染色体片段。因此，染色体片段的剂量可以被确定为针对在测试样品中的片段确定的序列标签覆盖度和在测试样品中的归一化染色体片段的序列标签覆盖度的比率，其中在测试样品中的归一化片段对应于在合格样品中针对感兴趣的特定片段确定的归一化片段(单个或一组片段)。在尺寸方面，染色体片段可以为千碱基(kb)至兆碱基(Mb)(例如，约1kb至10kb、或约10kb至100kb、或约100kb至1Mb)。

[0218] 在步骤155中，阈值来源于标准偏差值，其是针对在多个合格样品中确定的合格序列剂量和针对已知是感兴趣的序列的非整倍体的样品确定的序列剂量所建立。注意，通常以异步方式并借助于患者测试样品的分析来进行这种操作。它可以，例如，同时借助于自合格样品的归一化序列的选择来进行。准确的分类取决于在不同类别，即，非整倍体的类型，的概率分布之间的差异。在一些实施例中，阈值选自非整倍体的每个类型，例如，三体性21，的经验分布。为分类三体性13、三体性18、三体性21、和单体性X非整倍体(如在实施例中所描述的)所建立的可能的阈值，其描述用于通过测序提取自包含胎儿和母体核酸的混合物的母体样品的cfDNA来确定染色体非整倍体的方法的使用。被确定以用来区分受影响的样品的染色体的非整倍体的阈值可以是相同于或可以是不同于用于不同非整倍体的阈值。如在实施例中所示，针对感兴趣的每个染色体的阈值确定自整个样品和测序运行的感兴趣的染色体的剂量变异性。针对感兴趣的任何染色体的染色体剂量变化越小，则针对整个所有未受影响的样品的感兴趣的染色体的剂量的范围越窄，其用来设定用于确定不同非整倍体的阈值。

[0219] 回到相关与分类患者测试样品的工艺流程，在步骤160中，通过比较针对感兴趣的序列的测试序列剂量和建立自合格序列剂量的至少一个阈值来确定在测试样品中感兴趣的序列的拷贝数变异。可以通过用来测量序列标签覆盖度和/或计算片段剂量的同样的计算仪器来进行这种操作。

[0220] 在步骤160中，使针对感兴趣的测试序列的计算的剂量相比于设定为阈值的剂量，其是根据用户自定义“可靠性的阈值”加以选择，以将样品分类为“正常的”、“受影响的”、或“无调用的”。“无调用的”样品是这样的样品，对其不能可靠地作出明确诊断。每种类型的受影响的样品(例如，三体性21、部分三体性21、单体性X)具有它自己的阈值，一个用于调用正常的(未受影响的)样品以及另一个调用受影响的样品(虽然在一些情况下，上述两个阈值重合)。如在本文中别处描述的，在一些情形下，如果在测试样品中核酸的胎儿分数是足够高，则可以将无调用的转换为调用(受影响的或正常的)。可以通过在此过程流程的其他操作中采用的计算仪器来报告测试序列的分类。在一些情况下，分类是以电子格式加以报告并且可以可以被显示、发送电子邮件、发短信等给感兴趣的人。

[0221] 在一些实施方式中，CNV的确定包括计算NCV或NSV，其使染色体或片段剂量相关与在一组合格样品中相应的染色体或片段剂量的平均值(如上面描述的)。然后，可能通过比较NCV/NSV与预定的拷贝数评价阈值来确定CNV。

[0222] 可以选择拷贝数评价阈值以优化假阳性与假阴性的比率。拷贝数评价阈值越高，则越小可能发生假阳性。同样地，阈值越低，则越小可能发生假阴性。因此，在第一理想阈值(高于其，仅真阳性被归类)和第二理想阈值(低于其，仅真阴性被归类)之间存在权衡。

[0223] 设定阈值，其很大程度上取决于针对感兴趣的特定染色体的染色体剂量的变异性，如在一组不受影响的的样品中确定的。变异性取决于许多因素，包括在样品中存在的胎儿cDNA的分数。通过对于整个未受影响的样品的群体的染色体剂量的平均值或中位数和标准偏差来确定变异性(CV)。因此，用于分类非整倍体的阈值使用NCV，并根据：

[0224] $$NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

[0225] (其中$\hat{\mu}_j$和$\hat{\sigma}_j$分别是针对在一组合格样品中的第j染色体剂量的估计的平均值和标准偏差，以及x_{ij}是针对测试样品i.所观测到的第j染色体剂量)

[0226] 借助于相关的胎儿分数，为：

[0227] $$FF_{ij} = 2 \times \left| \frac{NCV_{ij} \times \hat{\sigma}_j}{\hat{\mu}_j} \right| = 2 \times NCV \times CV$$

[0228] 因此，基于针对整个未受影响的样品的群体的感兴趣的染色体的染色体比率的平均值和标准偏差，对于感兴趣的染色体的每个NCV，相关与给定NCV值的预期胎儿分数可以计算自CV。

[0229] 其后，基于在胎儿分数和NCV值之间的关系，可以选择判别边界，高于其，样品被确定为阳性(受影响的)(基于正态分布分位数)。如上面描述的，设定阈值，用于获得在真阳性的检测和假阴性结果的比率之间的最佳权衡。因此，选择设定的阈值以优化假阳性和假阴性。

[0230] 某些实施方式提供了用于在包含胎儿和母体核酸分子的生物样品中提供胎儿染色体非整倍体的产前诊断的方法。进行上述诊断，其基于：从来源于生物测试样品，例如，母体血浆样品，的胎儿和母体核酸分子的混合物的至少一部分获得序列信息；依据测序数据，计算用于感兴趣的一个或多个染色体的归一化染色体剂量和/或用于感兴趣的一个或多个片段的归一化片段剂量；以及在测试样品中确定在分别用于感兴趣的染色体的染色体剂量和/或用于感兴趣的片段的片段剂量和在多个合格(正常)样品中建立的阈值之间的统计学显着差异；以及基于统计差异来提供产前诊断。如在方法的步骤160中描述的，进行正常的或受影响的诊断。在不能有信心地对正常的或受影响的进行诊断的情况下，提供“无调用的”。

[0231] 在一些实施方式中，可以选择两个阈值。选择第一阈值以最小化假阳性率，高于其，样品将被分类为“受影响的”，以及选择第二阈值以最小化假阴性率，低于其，样品将被分类为“未受影响的”。具有高于第二阈值但低于第一阈值的NCV的样品可以被分类为“疑似的非整倍体”或“无调用的”样品，对于其，可以通过独立的方式来证实非整倍体的存在或不存在。在第一和第二阈值之间的区可被称为“无调用的”区。

[0232] 在一些实施方式中，疑似的和无调用的阈值示于表2。如可以看到的，NCV的阈值随不同的染色体而变化。在一些实施方式中，对于如以上所解释的样品，阈值根据FF而变化。在一些实施方式中，在这里应用的阈值技术有助于改善的灵敏度和选择性。

[0233] 表2:包括无调用的范围的疑似的和受影响的NCV阈值

[0234]

	疑似的	受影响的
Chr 13	3.5	4.0
Chr 18	3.5	4.5
Chr 21	3.5	4.0
ChT X(XO,XXX)	4.0	4.0
Chr Y(XX vs XY)	6.0	6.0

[0235] 确定序列覆盖度

[0236] 用于确定序列覆盖度的一般过程

[0237] 披露的一些实施方式提供了用来具有低噪声和高信号地确定序列覆盖度量的方法,从而提供数据来确定涉及到拷贝数和CNV的各种遗传病症,相对于通过常规方法获得的序列覆盖度量,其具有改善的敏度、选择性、和/或效率。在某些实施方式中,对来自测试样品的序列加以处理以获得序列覆盖度量。

[0238] 上述过程利用了可获自其他来源的某些信息。在一些实施方式中,所有的这种信息获自已知是未受影响的(例如,不是非整倍体)样品的训练集。在其他实施方式中,一些或所有的信息获自其他测试样品,当在同一过程中分析多个样品时,其可以被“即时”提供。

[0239] 在某些实施方式中,序列掩码用来降低数据噪声。在一些实施方式中,感兴趣的序列和它的归一化序列均被掩蔽。在一些实施方式中,当考虑感兴趣的不同染色体或片段时,可以采用不同的掩码。例如,当13号染色体是感兴趣的染色体时,可以采用一个掩码(或掩码组)以及当21号染色体是感兴趣的染色体时可以采用不同的掩码(或掩码组)。在某些实施方式中,以bin的分辨率来定义掩码。因此,在一个实例中,掩码分辨率是100kb。在一些实施方式中,不同的掩码可应用于Y染色体。可以以比用于其他感兴趣的染色体更加精细的分辨率(1kb)来提供用于Y染色体的掩蔽的排除区,如在于2013年6月17日提交的美国临时专利申请号61/836,057[代理人卷号ARTEP008P]中所描述的。以确定排除的基因组区的文件的形式来提供掩码。

[0240] 在某些实施方式中,上述过程利用归一化覆盖度的期望值以在感兴趣的序列的分布图中除去bin-到-bin变化,上述变化不提供用于确定测试样品的CNV的信息。上述过程,根据针对整个整个基因组的每个bin、或在参比基因组中的至少鲁棒染色体的bin的归一化覆盖度的期望值,来调节归一化覆盖度量(用于以下的操作317)。期望值可以确定自未受影响的样品的训练集。作为例子,期望值可以是整个训练集样品的中位数值。样品的预期覆盖度值可以被确定为比对于bin的独特的非冗余标记的数目除以比对于在参比基因组的鲁棒染色体中的所有bin的独特的非冗余标记的总数。

[0241] 图2描述了用于确定感兴趣的序列的覆盖度的过程200的流程图,其用来在方框214中评价在测试样品中感兴趣的序列的拷贝数。此过程除去通用于不受影响的训练样品的系统性变化,上述变化会增加在用于CNV评价的分析中的噪声。它还消除测试样品特有的GC偏差,从而增加在数据分析中的信噪比。

[0242] 上述过程开始于提供测试样品的序列读数,如在方框202中所示。在一些实施方式中,通过测序获自孕妇的血液的DNA片段,包括母体和胎儿的cfDNA,来获得序列读数。上述

过程进行以将序列读数比对于包括感兴趣的序列的参比基因组，从而提供测试序列标签。方框204。测试在参比序列上的每个bin中的序列标签计数定义了bin的覆盖度。方框206。在一些实施方式中，比对于一个以上位点的读数被排除。在一些实施方式中，比对于同一位点的多个读数被排除或减少到单读数计数。在一些实施方式中，比对于排除的位点的读数也被排除。因此，在一些实施方式中，仅计数比对于未排除的位点的唯一对齐的、非冗余标记，以提供用于确定每个bin的覆盖度的未排除的位点计数(NES计数)。在一些实施方式中，每个bin的覆盖度除以在同一样品中归一化序列的覆盖度，从而提供归一化覆盖度量。

[0243] 然后过程200提供感兴趣的序列的全局配置参数。全局配置参数包括在获自未受影响的训练样品的训练集的每个bin中的预期覆盖度。方框208。通过调节测试序列标签的归一化覆盖度量并根据预期覆盖度，过程200除去常见于训练样品的变化，以获得全局配置参数修正的覆盖度。方框210。在一些实施方式中，获自在方框208中提供的训练集的预期覆盖度是整个训练样品的中位数。在一些实施方式中。通过从归一化覆盖度减去预期覆盖度，操作2010调节归一化覆盖度量。在其他实施方式中，操作2010使归一化覆盖度量除以每个bin的预期覆盖度，以提供全局配置参数修正的覆盖度。

[0244] 此外，通过进一步调节已被调节的覆盖度量，过程200除去测试样品特有的GC偏差，以除去全局配置参数。如在方框212中所示，基于在GC含量水平和在测试样品中存在的全局配置参数修正的覆盖度之间的关系，上述过程调节全局配置参数修正的覆盖度，从而获得样品-GC-修正的覆盖度。在调节常见于未受影响的训练样品的系统性变化和受试者内GC偏差之后，上述过程提供覆盖度量来具有改善的灵敏度和特异性地评价样品的CNV。

[0245] 用于确定序列覆盖度的示例性过程的细节

[0246] 图3A提供了用于减少在来自测试样品的序列数据中的噪声的过程301的一个实施例。图3B-3J显示了在过程的不同阶段的数据分析。如在图3A中所示，描述的方法首先从一个或多个样品提取cfDNA。见方框303。在本文中别处描述了适宜的提取过程和仪器。在一些实施方式中，在2013年3月15日提交的美国专利申请号61/801,126(以引用方式将其全部内容结合于本文)中描述的过程提取cfDNA。在一些实施方式中，上述仪器一起处理来自多个样品的cfDNA以提供复用文库和序列数据。见在图3A中的方框305和307。在一些实施方式中，上述仪器平行处理来自八个或更多测试样品的cfDNA。如在本文中别处描述的，测序系统可以处理提取的cfDNA以产生编码的(例如，带条形码的)cfDNA片段的文库。测序仪测序cfDNA的文库以产生非常大量的序列读数。每个样品编码允许多路分解在多重样品中的读数。八个或更多样品的每一个可以具有几十万或上百万的读数。在图3A中的另外的操作之前，上述过程可以过滤读数。在一些实施方式中，读数过滤是能够通过在测序仪中实施的软件程序加以操作的质量过滤过程以过滤掉错误的和低质量的读数。例如，通过将由测序反应产生的原始图像数据转换成强度得分、基本调用(base call)、质量得分比对、和另外的格式，Illumina的测序控制软件(Sequencing Control Software，SCS)和序列和变化的一致评价软件程序过滤掉错误的和低质量的读数，以提供用于下游分析的生物学相关信息。

[0247] 在测序仪或其他仪器产生针对样品的读数之后，系统的元件将读数计算上对齐于参比基因组。见方框309。在本文中别处描述对齐。上述对齐产生标记，其含有读数序列，并具有带注释的位置信息，其指定在参比基因组上的独特位置。在某些实施方式中，上述系统进行第一通过对齐，而不考虑重复读数(具有相同序列的两个或更多读数)，以及随后除去

重复读数或将重复读数计数为单个读数，以提供非重复序列标签。在其他实施方式中，上述系统并不消除重复读数。在一些实施方式中，上述过程不考虑对齐于在基因组上的多个位置的读数，以产生唯一对齐的标记。在一些实施方式中，考虑映射到未排除的位点(NES)的唯一对齐的非冗余序列标签，以产生未排除的位点计数(NES计数)，其提供数据来估计覆盖度。

[0248] 如在别处解释的，排除的位点是在已被排除的参比基因组的区中发现的位点，借以计数序列标签。在一些实施方式中，排除的位点存在于染色体的区，其含有重复序列，例如，着丝粒和端粒，以及染色体的区，其是一个以上的染色体共有的，例如，在Y染色体上存在的区，其还存在在X染色体上。未排除的位点(NES)是在参比基因组中未排除的位点，借以计数序列标签。

[0249] 其次，上述系统将比对标记分为在参比基因组上的bin。见方框311。沿着参比基因组的长度，隔开bin。在一些实施方式中，整个参比基因组被分成连续bin，其可以具有定义的同等大小(例如，100kb)。可替换地，bin可以具有动态确定的长度(可能基于每个样品)。测序深度影响最佳bin尺寸选择。动态确定大小的bin可以具有由文库大小确定的它们的大小。例如，bin尺寸可以被确定为是为容纳1000个标记所需要有序列长度(平均而言)。

[0250] 每个bin具有来自在考虑中的样品的若干标记。标记的数目，其反映了比对序列的“覆盖度”，作为起点，用于过滤以及以其他方式清除样品数据，以可靠地确定在样品中的拷贝数变异。图3A示出在方框313至321中的清除操作。

[0251] 在图3A中描述的实施方式中，上述过程施加掩码于参比基因组的bin。见方框313。考虑到一些或所有以下处理操作，上述系统可以排除掩蔽bin中的覆盖度。在许多情况下，在图3A中的任何剩余的操作不考虑来自掩蔽bin的覆盖度值。

[0252] 在各种实施方式中，针对发现从样品到样品表现出高变异性的基因组的区，一个或多个掩码用来清除bin。提供这样的掩码，用于感兴趣的染色体(例如，chr13、18、和21)和其他染色体。如在别处解释的，感兴趣的染色体是在考虑中的可能藏匿拷贝数变异或其他畸变的染色体。

[0253] 在一些实施方式中，使用以下方法，掩码被确定自合格样品的训练集。最初，根据在图3A中的操作315至319，处理和过滤每个训练集样品。然后针对每个bin，指出归一化的和修正的覆盖度量，以及针对每个bin，计算统计数据如标准偏差、中位数绝对偏差、和/或变异系数。可以针对感兴趣的每个染色体，评价各种过滤组合。过滤程序组合提供一个过滤程序，用于感兴趣的染色体的bin，以及不同的过滤程序，用于所有其他染色体的bin。

[0254] 在一些实施方式中，在获得掩码(例如，通过选择针对如上面描述的感兴趣的染色体的截止)之后，重新考虑归一化染色体(或染色体组)的选择。在施加序列掩码之后，可以进行选择归一化染色体的过程(如在本文中别处描述的)。例如，染色体的所有可能的组合被评价为归一化染色体并根据它们区别受影响的和未受影响的样品的能力加以排列。此过程可能(或不可能)发现不同的最佳归一化染色体或染色体组。在其他实施方式中，归一化染色体是那些染色体，其导致在针对整个所有合格样品的感兴趣的序列的序列剂量的最小变异性。如果不同的归一化染色体或染色体组被确定，则上述过程可选地执行bin到过滤程序的上述鉴定。可能地，新的归一化染色体导致不同的截止。

[0255] 在某些实施方式中，不同的掩码应用于Y染色体。适宜的Y染色体掩码的实例描述

于2013年6月17日提交的美国临时专利申请号61/836,057[代理人档案号ARTEP008P],其以引用方式结合于本文。

[0256] 在上述系统计算上掩蔽bin之后,它计算上归一化在未由掩码排除的bin中的覆盖度值。见方框315。在某些实施方式中,上述系统,相对于在参比基因组中的大多数或所有的覆盖度或它们的一部分(例如,在参比基因组的鲁棒染色体中的覆盖度),归一化在每个bin中的测试样品覆盖度值(例如,NES计数/bin)。在一些情况下,通过针对在考虑中的bin的计数除以比对于在参比基因组中的所有鲁棒染色体的所有未排除的位点的总数,上述系统归一化测试样品覆盖度值(每bin)。在一些实施方式中,通过进行线性回归,上述系统归一化测试样品覆盖度值(每bin)。例如,上述系统首先将针对在鲁棒染色体中的bin子集的覆盖度计算为ya=截距+斜率*gwpa,其中ya是针对bina的覆盖度,以及gwpa是针对同样bin的全局配置参数。然后上述系统计算归一化覆盖度zb作为:zb=yb/(截距+斜率*gwpb)-1。

[0257] 如上文中所解释的,鲁棒染色体是一种不可能是非整倍体的染色体。在某些实施方式中,鲁棒染色体是不同于染色体13、18、和21的所有常染色体。在一些实施方式中,鲁棒染色体是不同于被确定为偏离自正常二倍体基因组的染色体的所有常染色体。

[0258] bin的转化计数值或覆盖度被称为用于进一步处理的“归一化覆盖度量”。利用每个样品独有的信息进行归一化。通常,不使用来自训练集的信息。归一化允许在平等的基础上处理来自具有不同的文库大小(因而不同数目的读数和标记)的样品的覆盖度量。一些随后的处理操作使用来源于训练样品的覆盖度量,其可以被测序自大于或小于用于在考虑中的测试样品的文库的文库。在一些实施方式中,没有基于比对于整个参比基因组(或至少鲁棒染色体)的读数数目的归一化,利用来源于训练集的参数进行的处理可能不是可靠的或一般化的。

[0259] 图3B示出对于许多样品整个染色体21、13、和18上的覆盖度。彼此不同地处理一些样品。作为结果,可以看到在任何给定的基因组位置处的宽的样品到样品的变化。归一化除去一些样品到样品的变化。图3C的左图描述整个整个基因组的归一化覆盖度量。

[0260] 在图3A的实施方式中,上述系统消除或减小来自在操作315中产生的归一化覆盖度量的“全局配置参数”。见方框317。这种操作除去归一化覆盖度量的系统性偏差,其产生自基因组的结构、文库产生过程、和测序过程。此外,这种操作旨在修正在任何给定样品中与预期分布图的任何系统线性偏差。

[0261] 在一些实施方式中,全局配置参数除去涉及每个bin的归一化覆盖度量除以每个bin的相应预期值。在其他实施方式中,全局配置参数除去涉及从每个bin的归一化覆盖度量减去每个bin的预期值。预期值可以获自训练集的未受影响的样品(或针对X染色体,未受影响的雌性样品)。未受影响的样品是来自已知不具有针对感兴趣的染色体的非整倍体的个体的样品。在一些实施方式中,全局配置参数除去涉及从每个bin的归一化覆盖度量减去每个bin的预期值(获自训练集)。在一些实施方式中,上述过程使用用于每个bin的归一化覆盖度量的中位数值(如使用训练集所确定的)。换句话说,中位数值是预期值。

[0262] 在一些实施方式中,利用对于样品覆盖度对全局配置参数的依赖的线性修正,来实施全局配置参数除去。如所示,全局配置参数是对于每个bin的预期值,如确定自训练集(例如对于每个bin的中位数值)。这些实施方式可以采用通过相对于针对每个bin获得的全局中位数分布图来拟合测试样品的归一化覆盖度量以获得鲁棒线性模型。在一些实施方式

中，通过相对于全局中位数(或其他期望值)分布图，回归样品的观测到的归一化覆盖度量来获得线性模型。

[0263] 线性模型是基于以下假设：样品覆盖度量具有与全局配置参数值的线性关系，上述线性关系应适用于鲁棒染色体/区和感兴趣的序列。见图3D。在这种情况下，样品归一化覆盖度量对全局配置参数的预期覆盖度量的回归将产生具有斜率和截距的线。在某些实施方式中，上述线的斜率和截距用来依据对于bin的全局配置参数值来计算“预测的”覆盖度量。在一些实施方式中，全局配置参数修正涉及通过对于bin的预测的覆盖度量来建模每个bin的归一化覆盖度量。在一些实施方式中，调节测试序列标签的覆盖度，其是通过：(i)在一个或多个强大染色体或区域中的多个bin中，获得在测试序列标签的覆盖度与预期覆盖度之间的数学关系，以及(ii)将上述数学关系应用于在感兴趣的序列中的bin。在一些实施方式中，利用在鲁棒染色体或基因组的其他稳健区中在来自未受影响的训练样品的预期覆盖度值和测试样品的覆盖度值之间的线性关系来修正在测试样品中覆盖度的变化。上述调节导致全局配置参数修正的覆盖度。在一些情况下，上述调节涉及在鲁棒染色体或区中获得针对bin子集的测试样品的覆盖度，具体如下：

[0264] y_a＝截距+斜率*gwp_a

[0265] 其中y_a是在一个或多个强大染色体或区域中测试样品的bina的覆盖度，以及gwp_a是针对未受影响的训练样品的bina的全局配置参数。然后上述过程计算针对感兴趣的序列或区的全局配置参数修正的覆盖度z_b，作为：

[0266] z_b＝y_b/(截距+斜率*gwp_b)-1

[0267] 其中y_b是在感兴趣的序列(其可以位于鲁棒染色体或区之外)中针对测试样品的bin b的观测覆盖度，以及gwp_b是针对未受影响的训练样品的bin b的全局配置参数。分母(截距+斜率*gwp_b)是bin b的覆盖度，基于估计自基因组的强大区的关系，其被预测在未受影响的测试样品中被观测到。在感兴趣的序列藏匿拷贝数变异的情况下，观测覆盖度，因而bin b的全局配置参数修正的覆盖度值将显著偏离未受影响的样品的覆盖度。例如，在在受影响的染色体上的bin的三体样品的情况下，修正的覆盖度z_b将正比于胎儿分数。通过对鲁棒染色体计算截距和斜率，此过程在样品内归一化，然后评价感兴趣的基因组区如何偏离适用于在同一样品中的鲁棒染色体的关系(如由斜率和截距所描述的)。

[0268] 上述斜率和截距获自如在图3D中所示的线。全局配置参数除去的一个实例描述于图3C。左图示出整个许多样品的归一化覆盖度的量的高bin-到-bin变化。右图示出在如上面描述的全局配置参数除去之后的同样的归一化覆盖度的量。

[0269] 在上述系统在方框317处消除或减小全局配置参数变化之后，它修正样品中GC(鸟嘌呤-胞嘧啶)含量变化。见方框319。每个bin具有它自己的来自GC的分数贡献。上述分数是通过在bin中的G和C核苷酸的数目除以在bin中核苷酸的总数(例如，100,000)来确定。一些bin将具有比其他bin更大的GC分数。如在图3E和3F中所示，不同样品表现出不同的GC偏差。下文将进一步解释这些差异和它们的修正。图3E-图3G示出全局配置参数修正的归一化覆盖度量(每bin)，其是作为GC分数(每个bin)的函数。出人意料的是，不同样品表现出不同的GC依赖。一些样品显示单调递减的依赖(如在图3E中)，而其他样品则呈现逗号形状的依赖(如在图3F和图3G中)。由于这些分布图对于每个样品可以是唯一的，所以对于每个样品单独地并唯一地进行在此步骤中描述的修正。

[0270] 在一些实施方式中，基于如在图3E-图3G描述的GC分数，所述系统在计算上安排bin。然后，利用来自具有类似的GC含量的其他bin的信息，它修正bin的全局配置参数修正的归一化覆盖度量。将这种修正应用于每个未掩蔽bin。

[0271] 在一些方法中，以以下方式，修正每个bin的GC含量。上述系统在计算上选择具有类似于在考虑中的bin的GC分数的bin，然后依据在所选的bin中的信息来确定修正参数。在一些实施方式中，利用相似性的任意定义的截止值来选择那些具有类似的GC分数的bin。在一个实例中，选择所有bin的2％。这些bin是上述2％，其具有最类似于在考虑中的bin的GC含量bin。例如，选择1％的具有稍微更大GC含量的bin和1％的具有稍微更小GC含量的bin。

[0272] 利用所选的bin，上述系统在计算上确定修正参数。在一个实例中，上述修正参数是在所选的bin中归一化覆盖度量(在全局配置参数除去之后)的代表性值。这样的代表性值的实例包括在所选的bin中归一化覆盖度量的中位数或均值。上述系统将用于在考虑中的bin的计算的修正参数应用于用于在考虑中的bin的归一化覆盖度量(在全局配置参数除去之后)。在一些实施方式中，从在考虑中的bin的归一化覆盖度量减去代表性值(例如，中位数值)。在一些实施方式中，仅利用针对鲁棒常染色体(不同于染色体13、18、和21的所有常染色体)的覆盖度量来选择归一化覆盖度量的中位数值(或其他代表性值)。

[0273] 在使用例如100kbbin的一个实例中，每个bin将具有GC分数的唯一值，以及基于它们的GC分数含量将bin分为组。例如，将bin分为50组，其中组边界对应于％GC分布的(0,2,4,6,...,和100)分位数。依据映射到相同GC组(在样品中)的鲁棒常染色体，对bin的每个组，计算中位数归一化覆盖度量，然后从归一化覆盖度量减去中位数值(对于在相同GC组中，整个整个基因组的所有bin)。这使估计自在任何给定样品内的鲁棒染色体的GC修正适用于在同一样品中的潜在受影响的染色体。例如，一起分组在鲁棒染色体上的具有0.338660至0.344720的GC含量的所有bin，对于此组计算中位数并从在此GC范围内的bin的归一化覆盖度减去，可以在基因组(排除染色体13、18、21、和X)上的任意处发现上述bin。在某些实施方式中，从这种GC修正过程排除Y染色体。

[0274] 图3G示出GC修正的应用，其中利用中位数归一化覆盖度量作为修正参数(如刚刚描述的)。左图示出未修正的覆盖度量与GC分数分布图。如图所示，上述分布图具有非线性形状。右图示出修正的覆盖度量。图3H示出对于许多样品在GC分数修正之前(左图)以及在GC分数修正之后(右图)的归一化覆盖度。图3I示出对于许多测试样品在GC分数修正之前(红色)和在GC分数修正之后(绿色)的归一化覆盖度的变异系数(CV)，其中GC修正导致归一化覆盖度的显著更小的变化。

[0275] 所述方法GC修正的相对简单的实施。用来修正GC偏差的替代方法采用样条或其他非线性拟合技术，其可以应用于连续GC空间并且不涉及通过GC含量来分级覆盖度量。适宜的技术的实例包括连续黄土修正和光滑样条修正。对于在考虑中的样品，拟合函数可以来源于bin-bin归一化覆盖度量与GC含量。通过将对于在考虑中的bin的GC含量施加于拟合函数来计算对于每个bin的修正。例如，可以通过减去在考虑中的bin的GC含量下样均被涵盖的预期覆盖度值来调节归一化覆盖度量。可替换地，可以根据样均被涵盖拟合并通过划分预期覆盖度值来实现调节。

[0276] 在操作319中修正GC依赖之后，上述系统在计算上除去在考虑中的样品中的离群bin，见方框321。这种操作可被称为单个样品过滤或修剪。图3J示出，甚至在GC修正之后，覆

盖度仍然具有在小区域内的样品特有的变化。见例如在12号染色体上在位置1.1e8处的覆盖度，其中发生出乎意料的与预期值的高偏差。可能的是，这种偏差产生于在母体基因组中的小拷贝数变异。可替换地，这可能是由于在测序中不相关于拷贝数变异的技术原因。通常，这种操作仅适用于鲁棒染色体。

[0277] 作为一个实例，上述系统在计算上过滤任何bin，其具有离整个在染色体(藏匿在考虑中的用于过滤的bin)中的所有bin的GC修正的归一化覆盖度量的中位数大于3个中位数绝对偏差的GC修正的归一化覆盖度量。在一个实例中，截止值被定义为3个中位数绝对偏差，其被调节以一致与标准偏差，所以实际上截止是离中位数1.4826*中位数绝对偏差。在某些实施方式中，这种操作应用于在样品中的所有染色体，包括鲁棒染色体和可疑具有非整倍体的染色体。

[0278] 在某些实施方式中，进行另外的操作，其可以被表征为质量控制。见方框323。在一些实施方式中，质量控制度量涉及检测是否任何潜在的分母染色体，即“归一化染色体”或“鲁棒染色体”，是非整倍体，或以其他方式不适用于确定是否测试样品具有在感兴趣的序列中的拷贝数变异。当所述方法确定鲁棒染色体是不合适的时，所述方法可以忽视测试样品并使得无调用的。可替换地，这种QC度量的失效可以触发替代组的归一化染色体用于调用。在一个实例中，质量控制方法比较鲁棒染色体的实际的归一化覆盖度值与鲁棒常染色体的期望值。可以通过将多元正规模型拟合于未受影响的训练样品的归一化的分布图，根据数据的似然或贝叶斯准则来选择最好的模型结构(例如，利用AIC准则或可能地贝叶斯信息准则来选择模型)，以及固定用于QC的最佳模型，来获得期望值。可以通过，例如，利用聚类技术，其识别针对在正常样品中的染色体覆盖度具有平均值和标准偏差的概率函数，来获得鲁棒染色体的正规模型。当然，可以使用其他的模型形式。鉴于固定的模型参数，所述方法评价在任何进入的测试样品中观测的归一化覆盖度的似然。可以通过借助于模型来记分每个进入的测试样品来做到这一点，以获得似然并从而确定相对于正常样品集的离群。测试样品的似然与训练样品的似然的偏差可能提示，在归一化染色体或样品处理/分析处理伪像中的异常，其可能导致不正确的样品分类。这种QC度量可以用来在相关与任何一个这些样品伪像的分类中减少差错。图3K，右图，示出在x轴上的染色体数目以及y轴示出归一化染色体覆盖度，其是基于与如上面描述的获得的QC模型的比较。图形显示针对2号染色体具有过度覆盖度的一个样品以及针对20号染色体具有过度覆盖度的其他样品。将利用这里描述的QC度量来消除这些样品，或改用替代组的归一化染色体。图3K的左图示出针对染色体的NCV与似然。

[0279] 在图3A中描述的序列可以用于在基因组中的所有染色体的所有bin。在某些实施方式中，不同的过程应用于Y染色体。为了计算染色体或片段剂量、NCV、和/或NSV，使用来自在剂量、NCV、和/或NSV的表达中使用的染色体或片段的bin的修正的归一化覆盖度量(如在图3A中确定的)。见方框325。在某些实施方式中，平均归一化覆盖度量计算自在感兴趣的染色体中的所有bin，归一化染色体、感兴趣的片段、和/或归一化片段用来计算序列剂量、NCV、和/或NSV(如在本文中别处描述的)。

[0280] 在某些实施方式中，不同地处理Y染色体。它可以掩蔽Y染色体独有的一组bin加以过滤。在一些实施方式中，根据在美国临时专利申请号61/836,057(先前以引用方式结合于本文)的方法来确定Y染色体过滤。在一些实施方式中，上述过滤掩蔽这样的bin，其小于那

些在其他染色体的过滤中的bin。例如,Y染色体掩码可以在1kb水平下过滤,而其他染色体掩码可以在100kb水平下过滤。然而,可以在同样bin尺寸下将Y染色体归一化为其他染色体(例如,100kb)。

[0281] 在某些实施方式中,如上面在图3A的操作315中描述的来归一化过滤的Y染色体。然而,以其他方式,并不进一步修正Y染色体。因此,Y染色体bin并不经受全局配置参数除去。同样地,Y染色体bin并不经受其后进行的GC修正或其他过滤步骤。这是因为,当处理样品时,所述方法并不知道样品是男性或女性。雌性样品不应具有比对于Y参比染色体的读数。

[0282] 产生序列掩码

[0283] 本文披露的一些实施方式采用利用序列掩码来过滤掉(或掩蔽)在感兴趣的序列上的非判别序列读数的策略,在用于CNV评价的覆盖度值方面,相对于通过常规方法计算的值,其导致更高的信号和更低的噪声。可以通过各种技术来确定这样的掩码。在一种实施方式中,利用图4A-4B所示的技术来确定掩码(如下面进一步详细解释的)。

[0284] 在一些实施方式中,利用训练集的已知具有感兴趣的序列的正常拷贝数的代表性样品来确定掩码。可以利用一种技术,其首先归一化训练集样品,然后修正整个一系列序列(例如,分布图)的系统性变化,接着修正GC变异性(如下所述的),来确定掩码。对来自训练集的样品,而不对测试样品,进行归一化和修正。掩码被确定一次,然后应用于许多测试样品。

[0285] 图4A示出用于产生这样的序列掩码的过程400的流程图,其可以应用于一个或多个测试样品以除去在拷贝数的评价中考虑到的感兴趣的序列上的bin。上述过程开始于提供训练集,其包括来自多个未受影响的训练样品的序列读数。方框402。上述过程然后将训练集的序列读数比对于包含感兴趣的序列的参比基因组,从而提供用于训练样品的训练序列标签。方框404。在一些实施方式中,仅映射到未排除的位点的唯一对齐的非冗余标记用于进一步分析。上述过程涉及将参比基因组分成多个bin并针对每个未受影响的训练样品确定在对于每个训练样品的每个bin中训练序列标签的覆盖度。方框406。上述过程还对所有训练样品并针对每个bin确定训练序列标签的预期覆盖度。方框408。在一些实施方式中,每个bin的预期覆盖度是整个训练样品的中位数或均值。预期覆盖度构成全局配置参数。然后对于每个训练样品上述过程调节在每个bin中训练序列标签的覆盖度:通过除去在全局配置参数中的变化,从而对于每个训练样品获得在bin中训练序列标签的全局配置参数修正的覆盖度。上述过程然后产生包含整个参比基因组的未掩蔽和掩蔽bin的序列掩码。每个掩蔽bin具有超过掩蔽阈值的分布特征。上述分布特征提供了在跨整个训练样品的bin中训练序列标签的调节的覆盖度。在一些实施方式中,掩蔽阈值可能涉及到在跨整个训练样品的bin内归一化覆盖度的观察到的变化。可以基于相应度量的经验分布来确定具有整个样品的归一化覆盖度的高变化系数或中位数绝对偏差的bin。在一些替代的实施方式中,掩蔽阈值可能涉及到在跨整个训练样品的bin内归一化覆盖度的观察到的变化。可以基于相应度量的经验分布来掩蔽具有整个样品的归一化覆盖度的高变化系数或中位数绝对偏差的bin。

[0286] 在一些实施方式中,用于确定掩蔽bin的分开的截止,即,掩蔽阈值,是针对感兴趣的染色体和针对所有其他染色体所定义。另外,分开的掩蔽阈值可以分别地针对感兴趣的

每个染色体加以定义，以及单掩蔽阈值可以针对所有非受影响的染色体的集合加以定义。作为例子，基于一定掩蔽阈值的掩码是针对13号染色体所定义以及另一个掩蔽阈值用来定义用于其他染色体的掩码。非受影响的染色体还可以具有它们的依照染色体加以定义的掩蔽阈值。

[0287] 可以针对感兴趣的每个染色体来评价各种掩蔽阈值组合。掩蔽阈值组合提供用于感兴趣的染色体的bin的一种掩码和用于所有其他染色体的bin的不同的掩码。

[0288] 在一种方式中，对于变异系数(CV)的一系列值或样品分布截止的度量被定义为binCV值的经验分布的百分位数(例如，95、96、97、98、99)以及这些截止值应用于所有常染色体(排除感兴趣的染色体)。另外，用于CV的一系列的百分位数截止值是针对经验CV分布所定义并且这些截止值应用于感兴趣的染色体(例如，chr21)。在一些实施方式中，感兴趣的染色体是X染色体以及染色体13、18、和21。当然，可以考虑其他方式，例如，可以针对每个染色体进行单独的优化。一起，待平行优化的范围(例如，一个范围用于在考虑中的感兴趣的染色体以及另一个范围用于所有其他染色体)定义CV截止组合的网格。见图4B。整个上述两个截止(一个用于归一化染色体(或不同于感兴趣的染色体的常染色体)和一个用于感兴趣的染色体)来评价上述系统对训练集的性能以及选择表现最好的组合用于最终配置。对于每个感兴趣的染色体，此组合可以是不同的。在某些实施方式中，对验证集而不是训练集来评价性能，即，交叉验证用来评价性能。

[0289] 在一些实施方式中，被优化以确定截止范围的性能是染色体剂量的变异系数(基于归一化染色体的试探性选择)。上述过程选择截止的组合，利用目前所选的归一化染色体，其最小化感兴趣的染色体的染色体剂量(例如，比率)的CV。在一种方式中，上述过程测试在网格中的截止的每个组合的性能，具体如下：(1)应用截止的组合以定义对于所有染色体的掩码发及应用那些掩码来过滤训练集的标记；(2)通过将图3A的过程应用于过滤的标记，来计算整个未受影响的样品的训练集的归一化覆盖度；(3)通过，例如，求和针对在考虑中的染色体的bin的归一化覆盖度来确定代表性归一化覆盖度/染色体；(4)利用目前的归一化染色体来计算染色体剂量；以及(5)确定染色体剂量的CV。通过将它们应用于分离自训练集的原始部分的一组测试样品，上述过程可以评价所选过滤的性能。即，和述过程将原始训练集分成训练和测试子集。上述训练子集用来定义掩码截止(如上面描述的)。

[0290] 在可替代的实施方式中，代替基于覆盖度的CV来定义掩码，掩码可以由来自跨整个训练样品在bin内的比对结果的映射质量得分的分布来定义。映射质量得分反映了借其读数被映射到参比基因组的独特性。换句话说，映射质量得分量化读数被错比对的可能性。低映射质量得分是相关的低独特性(错比对的高可能性)。独特性引起在读数序列中的一个或多个误差(如由测序仪产生的)。映射质量得分的详细描述提供于Li H,Ruan J,Durbin R.(2008)Mapping short DNA sequencing reads and calling variants using mapping quality scores.Genome Research 18:1851-8，其全部内容以引用方式结合于本文。在一些实现方式中，映射质量得分在本文中被称为MapQ得分。图4B示出，MapQ得分具有与处理的覆盖度的CV的单调的强相关性。例如，具有CV高于0.4的bin几乎完全群集在图4B中的绘图的左侧，其具有低于约4的MapQ得分。因此，掩蔽具有小MapQ的bin可以产生相当类似于由掩蔽具有高CV的bin所定义的掩码。

[0291] 样品和样品处理

[0292] 样品

[0293] 用于确定CNV,例如,染色体非整倍体、部分非整倍体等,的样品可以包括取自任何细胞、组织、或器官的样品,其中针对一个或多个感兴趣的序列的拷贝数变异待被确定。期望地,上述样品含有存在于细胞中的核酸和/或为“无细胞”的核酸(例如,cfDNA)。

[0294] 在一些实施方式中,有利的是,获得无细胞核酸,例如,无细胞DNA(cfDNA)。无细胞核酸,包括无细胞DNA,可以通过本领域中已知的各种方法,获自生物样品,包括但不限于血浆、血清、和尿(参见,例如,Fan等人,Proc Natl Acad Sci 105:16266-16271[2008];Koide等人,Prenatal Diagnosis 25:604-607[2005];Chen等人,Nature Med.2:1033-1035[1996];Lo等人,Lancet 350:485-487[1997];Botezatu等人,Clin Chem.46:1078-1084,2000;和Su等人,J Mol.Diagn.6:101-107[2004])。为了从样品中的细胞分离无细胞DNA,可以使用各种方法,包括但不限于分级分离、离心(例如,密度梯度离心)、DNA的特异性沉淀、或高通量细胞分选和/或其他分离方法。用于cfDNA的手动和自动分离的市售试剂盒是可获得的(Roche Diagnostics,Indianapolis,IN,Qiagen,Valencia,CA,Macherey-Nagel,Duren,DE)。通过测序分析,其可以检测染色体非整倍体和/或各种多态性,包含cfDNA的生物样品已用于测定来确定染色体异常,例如,三体性21,的存在或不存在。

[0295] 在不同的实施方式中,可以在使用之前(例如,在制备测序文库之前),特异性地或非特异性地富集在样品中存在的cfDNA。样品DNA的非特异性富集是指样品的基因组DNA片段的全基因组扩增,其可以用来在制备cfDNA测序文库之前增加样品DNA的水平。非特异性富集可以是在包含一种以上的基因组的样品中存在的两种基因组之一的选择性富集。例如,非特异性富集对在母体样品中的胎儿基因组可以选择性的,其可以通过已知方法来获得以增加在样品中胎儿与母体DNA的相对比例。可替换地,非特异性富集可以是在样品中存在的两种基因组的非选择性扩增。例如,非特异性扩增可以是在包含来自胎儿和母体基因组的DNA的混合物的样品中胎儿和母体DNA的扩增。用于全基因组扩增的方法在本领域中是已知的。简并寡核苷酸引发PCR(DOP)、引物延伸PCR技术(PEP)和多重置换扩增(MDA)是全基因组扩增方法的实例。在一些实施方式中,包含来自不同的基因组的cfDNA的混合物的样品对于在混合物中存在的基因组的cfDNA来说是非富集的。在其他实施方式中,包含来自不同的基因组的cfDNA的混合物的样品对于在样品中存在的基因组的任何一种来说是非特异性发富集的。

[0296] 对其应用本文描述的方法的包含核酸的样品通常包括生物样品(“测试样品”),例如,如上面描述的。在一些实施方式中,通过若干众所周知的方法的任何一种来纯化或分离待针对一个或多个CNV加以筛选的核酸。

[0297] 因此,在某些实施方式中,样品包含或组成自纯化的或分离的多核苷酸,或它可以包括样品如组织样品、生物液体样品、细胞样品等。适宜的生物液体样品包括但不限于血液、血浆、血清、汗液、眼泪、痰、尿、痰、耳流出物(ear flow)、淋巴液、唾液、脑脊液、灌洗液(ravages)、骨髓悬浮液、阴道流出物、经宫颈灌洗液、脑液、腹水、乳液、呼吸道、肠道和生殖泌尿道的分泌物、羊水、乳液、和白细胞分离术样品。在一些实施方式中,样品是通过非侵入性程序容易得到的样品,例如,血液、血浆、血清、汗液、眼泪、痰、尿、痰、耳流、唾液或粪便。在某些实施方式中,样品是外周血样品、或外周血样品的血浆和/或血清部分。在其他实施方式中,生物样品是拭子或涂片、活检样品、或细胞培养物。在另一种实施方式中,样品是两

种或更多种生物样品的混合物，例如，生物样品可以包括生物液体样品、组织样品、和细胞培养样品的两种或更多种。如在本文中所使用的，术语“血液”、“血浆”和“血清”明确地涵盖其部分或处理的部分。同样地，当样品取自活检、拭子、涂片等时，“样品”明确地涵盖来源于活检、拭子、涂片等的处理过的部分。

[0298] 在某些实施方式中，样品可以获自来源，包括但不限于来自不同个体的样品、来自相同或不同个体的不同发育阶段的样品、来自不同患病个体(例如，患有癌症或疑似具有遗传紊乱的个体)的样品、来自正常个体的样品、在个体中在疾病的不同阶段获得的样品、获自对于疾病经受不同治疗的个体的样品、来自经受不同的环境因素的个体的样品、来自具有对病变的素质的个体的样品、来自暴露于传染病病原体(例如，HIV)的个体的样品等。

[0299] 在一种说明性的但非限制性的实施方式中，样品是获自妊娠雌性，例如孕妇，的母体样品。在这种情况下，可以利用本文描述的方法来分析样品以提供在胎儿中潜在染色体异常的产前诊断。母体样品可以是组织样品、生物液体样品、或细胞样品。生物液体包括，作为非限制性实例，血液、血浆、血清、汗液、眼泪、痰、尿、痰、耳流、淋巴样液、唾液、脑脊液、灌洗液、骨髓悬浮液、阴道流出物、经宫颈灌洗液、脑液、腹水、乳液、呼吸道、肠道和生殖泌尿道的分泌物、以及白细胞分离术样品。

[0300] 在另一种说明性的但非限制性的实施方式中，母体样品是两种或更多种生物样品的混合物，例如，生物样品可以包括生物液体样品、组织样品、和细胞培养样品的两种或更多种。在一些实施方式中，样品是通过非侵入性程序容易得到的样品，例如，血液、血浆、血清、汗液、眼泪、痰、尿、乳汁、痰、耳流出物、唾液和粪便。在一些实施方式中，生物样品是外周血样品、和/或其血浆和血清部分。在其他实施方式中，生物样品是拭子或涂片、活检样品、或细胞培养物的样品。如上文所披露的，术语“血液”、“血浆”和“血清”明确地涵盖其部分或处理的部分。同样地，当样品取自活检、拭子、涂片等时，“样品”明确地涵盖来源于活检、拭子、涂片等的处理过的部分。

[0301] 在某些实施方式中，样品还可以获自体外培养组织、细胞、或其他含有多核苷酸的来源。培养的样品可以取自来源，包括但不限于维持在不同培养基和均被涵盖件(例如，pH、压力、或温度)下的培养物(例如，组织或细胞)、维持不同时期的长度的培养物(例如，组织或细胞)、用不同因子或试剂(例如，候选药物、或调节剂)处理的培养物(例如，组织或细胞)、或不同类型的组织和/或细胞的培养物。

[0302] 从生物源分离核酸的方法是众所周知的并且将会有所不同，其取决于来源的特性。根据本文描述的方法的需要，本领域的技术人员可以容易地从来源分离核酸。在一些情况下，可以是有利的是，片段化在核酸样品中的核酸分子。如实现的，片段化可以是无规的，或者它可以是特定的，例如，利用限制性内切核酸酶消化。用于随机片段化的方法在本领域中是众所周知的，并且包括，例如，有限DNA酶消化、碱处理和物理剪切。在一种实施方式中，获得样品核酸，作为cfDNA，其未经受片段化。

[0303] 在其他说明性实施方式中，获得样品核酸，作为基因组DNA，将其片段化成大约300或更多、大约400或更多、或大约500或更多碱基对的片段，并且对其可以容易地应用NGS方法。

[0304] 测序文库的制备

[0305] 在一种实施方式中，本文描述的方法可以利用下一代测序技术(NGS)，其允许在单

测序运行中单独测序多个样品，作为基因组分子(即，单重测序)或作为包含索引基因组分子的汇集样品(例如，多重测序)。这些方法可以产生高达数亿读数的DNA序列。在不同的实施方式中，可以利用，例如，本文描述的下一代测序技术(NGS)，来确定基因组核酸、和/或索引基因组核酸的序列。在不同的实施方式中，可以利用如本文所描述的一个或多个处理器来分析利用NGS获得的大量的序列数据。

[0306] 在不同的实施方式中，上述测序技术的应用不涉及测序文库的制备。

[0307] 然而，在某些实施方式中，本文中设想的测序方法涉及测序文库的制备。在一种说明性方式中，测序文库制备涉及准备好加以测序的衔接子修饰DNA片段(例如，多核苷酸)的随机收集的生产。多核苷酸的测序文库可以制备自DNA或RNA，包括DNA或cDNA的等同物、类似物，例如，DNA或cDNA，其是通过逆转录酶的作用产生自RNA模板的互补的或拷贝DNA。多核苷酸可能来源于双链形式(例如，dsDNA如基因组DNA片段、cDNA、PCR扩增产物等)，或在某些实施方式中，多核苷酸可能来源于单链形式(例如，ssDNA、RNA等)并已转为dsDNA形式。通过说明的方式，在某些实施方式中，单链mRNA分子可被复制到适用于制备测序文库的双链cDNA。对于文库制备的方法来说，初级多核苷酸分子的精确序列一般不是实质性的，并且可以是已知或未知的。在一种实施方式中，多核苷酸分子是DNA分子。更特别地，在某些实施方式中，多核苷酸分子表示生物体的全部遗传补体或基本上生物体的全部遗传补体，以及是基因组DNA分子(例如，细胞DNA、无细胞DNA(cfDNA)等)，其通常包括内含子序列和外显子序列(编码序列)、以及非编码调节序列如启动子和增强子序列。在某些实施方式中，初级多核苷酸分子包括在妊娠受试者的外周血中存在的人类基因组DNA分子，例如，cfDNA分子。

[0308] 通过使用包括特定范围的片段尺寸的多核苷酸来促进用于一些NGS测序平台的测序文库的制备。上述文库的制备通常涉及较大多核苷酸(例如，细胞基因组DNA)的片段化以获得在所期望的尺寸范围内的多核苷酸。

[0309] 可以通过本领域技术人员已知的若干方法的任何一种来实现片段化。例如，可以通过机械方式，包括但不限于雾化、超声处理和水剪切，来实现片段化。然而，机械片段化通常在C-O、P-O和C-C键处切割DNA主链，从而导致平端以及3'-和5'-突出端(具有被打断的C-O、P-O和C-C键)的异质混合物(参见，例如，Alnemri和Liwack，J Biol.Chem 265:17323-17333[1990];Richards和Boyer，J Mol Biol 11:327-240[1965])，其可能需要被修复，因为它们可能缺乏用于随后的酶促反应的必要的5'-磷酸，例如，测序衔接子的连接，其是为制备用于测序的DNA所需要的。

[0310] 相比之下，cfDNA，通常存在为小于约300个碱基对的片段，因而，为了利用cfDNA样品来产生测序文库，片段化通常不是必要的。

[0311] 通常，不管多核苷酸被强制片段化(例如，体外被片段化)或天然存在为片段，它们被转化为具有5'-磷酸和3'-羟基的平端DNA。标准方法，例如，用于利用，例如，如在本文中别处描述的Illumina平台来测序的方法，指导用户末端修复样品DNA，以在dA-尾之前纯化末端修复的产物，以及在文库制备的衔接子连接步骤之前纯化dA-尾产物。

[0312] 本文描述的序列文库制备的方法的各种实施方式不需要进行由标准方法通常要求的一个或多个步骤来获得可以通过NGS加以测序的修饰DNA产物。简略方法(ABB方法)、一步法、和两步法是用于制备测序文库的方法的实例，其可以见于2012年7月20日提交的专利申请13/555,037，其全部内容以引用方式结合于本文。

[0313] 用于跟踪和确认样品完整性的标志物核酸(marker nucleic acid)

[0314] 在不同的实施方式中,可以通过测序样品基因组核酸,例如,cfDNA,的混合物,以及补充,例如,在处理之前,已被引入样品的标志物核酸,来完成样品完整性的确认和样品跟踪。

[0315] 标志物核酸可以与测试样品(例如,生物来源样品)结合并经历如下处理,包括,例如,分馏生物来源样品的一个或多个步骤,例如,从全血样品获得基本上无细胞血浆部分,从分馏的,例如,血浆,或未分馏的生物源样品,例如,组织样品,纯化核酸,以及测序。在一些实施方式中,测序包括制备测序文库。选择结合与源样品的标记分子的序列或序列的组合,以是源样品独有的。在一些实施方式中,在样品中的独特的标记分子均具有相同序列。在其他实施方式中,在样品中的独特的标记分子是多个序列,例如,两种、三种、四种、五种、六种、七种、八种、九种、十种、十五种、二十种、或更多种不同序列的组合。在一种实施方式中,可以利用具有相同序列的多种标志物核酸分子来证实样品的完整性。可替换地,可以利用具有至少两种、至少三种、至少四种、至少五种、至少六种、至少七种、至少八种、至少九种、至少十种、至少11种、至少12种、至少13种、至少14种、至少15种、至少16种、至少17种、至少18种、至少19种、至少20种、至少25种、至少30种、至少35种、至少40种、至少50种、或更多种不同序列的多种标志物核酸分子来证实样品的同一性。多种生物的样品,即,两种或更多种生物样品,的完整性的确认需要,用标志物核酸(其具有待标示的多种测试样品的每一种独有的序列)来标示两种或更多种样品的每一种。例如,可以用具有序列A的标志物核酸来标示第一样品,以及可以用具有序列B的标志物核酸来标示第二样品。可替换地,可以用均具有序列A的标志物核酸分子来标示第一样品,以及可以用序列B和C的混合物来标示第二样品,其中序列A、B和C是具有不同序列的标记分子。

[0316] 可以在样品制备的任何阶段将标志物核酸加入样品,其发生在文库制备(如果要制备文库)和测序之前。在一种实施方式中,标记分子可以结合与未经处理的源样品。例如,可以在用来收集血液样品的收集管中提供标志物核酸。可替换地,可以在抽血之后将标志物核酸加入血液样品。在一种实施方式中,将标志物核酸加入用来收集生物液体样品的容器,例如,将标志物核酸加入用来收集血液样品的血液收集管。在另一种实施方式中,将标志物核酸加入生物液体样品的一部分。例如,将标志物核酸加入血液样品的血浆和/或血清部分,例如,母体血浆样品。在又一种实施方式中,将标记分子加入纯化样品,例如,已纯化自生物样品的核酸样品。例如,将标志物核酸加入经纯化的母体和胎儿cfDNA的样品。同样地,可以在处理样品之前将标志物核酸加入活检样品。在一些实施方式中,标志物核酸可以结合与将标记分子递送进入生物样品的细胞的载体。细胞递送载体包括pH敏感的和阳离子脂质体。

[0317] 在不同的实施方式中,标记分子具有反基因组序列,其是缺失自生物源样品的基因组的序列。在一种示例性实施方式中,用来验证人生物源样品的完整性的标记分子具有缺失自人类基因组的序列。在一种可替换的实施方式中,标记分子具有这样的序列,其缺失自源样品和任何一种或多种其他已知的基因组。例如,用来验证人生物源样品的完整性的标记分子具有这样的序列,其缺失自人类基因组和小鼠基因组。上述替代允许确认包含两种或更多种基因组的测试样品的完整性。例如,可以利用具有缺失自人类基因组和影响细菌的基因组的序列的标记分子来证实获自受病原体例如细菌影响的受试者的人无细胞DNA

样品的完整性。许多病原体，例如，细菌、病毒、酵母、真菌、原生动物等，的基因组的序列是在万维网上在ncbi.nlm.nih.gov/genomes处可公开获得的。在另一种实施方式中，标记分子是这样的核酸，其具有缺失自任何已知基因组的序列。可以在算法上随机产生标记分子的序列。

[0318] 在不同的实施方式中，标记分子可以是天然存在的脱氧核糖核酸(DNA)、核糖核酸或人工核酸类似物(核酸模拟物)，包括肽核酸(PNA)、吗啉代核酸、锁核酸、二醇核酸(glycol nucleic acid)、和苏糖核酸，其区别于天然存在的DNA或RNA(通过分子主链的变化)或DNA模拟物(其并不具有磷酸二酯主链)。脱氧核糖核酸可以来自天然存在的基因组或可以在实验室中通过使用酶或通过固相化学合成来产生。化学方法还可以用来产生未在自然界中发现的DNA模拟物。DNA的衍生物是可获得的衍生物，其中磷酸二酯键已被替换但其中脱氧核糖被保留，包括但不限于具有通过硫代甲缩醛或甲酰胺键形成的主链的DNA模拟物，其已被表明是良好的结构DNA模拟物。其他DNA模拟物包括吗啉代衍生物和肽核酸(PNA)，其含有基于N-(2-氨基乙基)甘氨酸的假肽主链(Ann Rev Biophys Biomol Struct 24:167-183[1995])。PNA是DNA(或核糖核酸[RNA])的非常良好的结构模拟物，以及PNA寡聚体能够与沃森-克里克互补DNA和RNA(或PNA)低聚体形成非常稳定的双螺旋结构，并且通过螺旋侵入，它们还可以结合于在双螺旋DNA中的靶(Mol Biotechnol 26:233-248[2004]。可以用作标记分子的DNA类似物的另一种良好的结构模拟物/类似物是硫代磷酸DNA，其中非桥连氧之一被硫替换。这种修饰降低了内切和外切核酸酶2的作用，包括5’至3’和3’至5’DNA POL 1外切核酸酶、核酸酶S1和P1、核糖核酸酶、血清核酸酶以及蛇毒素磷酸二酯酶。

[0319] 标记分子的长度与样品核酸的长度可以是不同的或模糊的，即，标记分子的长度可以类似于样品基因组分子的长度，或它可以大于或小于样品基因组分子的长度。借助于构成标记分子的核苷酸或核苷酸类似物碱基的数目来测量标记分子的长度。利用本领域中已知的分离方法，长度不同于样品基因组分子的标记分子可以区别于源核酸。例如，可以通过电泳分离，例如，毛细管电泳，来确定标记物和样品核酸分子在长度方面的差异。尺寸差异可以有利于量化和评价标记物和样品核酸的质量。优选地，标志物核酸短于基因组核酸，并具有足够长度以使它们不能被映射到样品的基因组。例如，需要30个碱基人序列以独特地将它映射到人类基因组。因此，在某些实施方式中，在人样品的测序生物测定中使用的标记分子的长度应是至少30bp。

[0320] 主要通过用来确认源样品的完整性的测序技术来确定标记分子的长度的选择。还可以考虑待测序的样品基因组核酸的长度。例如，一些测序技术采用多核苷酸的克隆扩增，其可能需要，待克隆扩增的基因组多核苷酸具有最小长度。例如，利用Illumina GAII序列分析仪的测序包括通过多核苷酸的桥接PCR所进行的体外克隆扩增(还被称为群集扩增)，其中多核苷酸具有110bp的最小长度，对其连接衔接子以提供至少200bp并且小于600bp的可以被克隆扩增和测序的核酸。在一些实施方式中，衔接子-连接的标记分子的长度是约200bp至约600bp、约250bp至550bp、约300bp至500bp、或约350至450。在其他实施方式中，衔接子-连接的标记分子的长度是约200bp。例如，当测序在母体样品中存在的胎儿cfDNA时，标记分子的长度可被选择为类似于胎儿cfDNA分子的长度。因此，在一种实施方式中，在包括大规模平行测序在母体样品中的cfDNA以确定胎儿染色体非整倍体的存在或不存在的测定中使用的标记分子的长度可以是约150bp、约160bp、约170bp、约180bp、约190bp或约

200bp，优选地，标记分子是约170pp。其他测序方式，例如，SOLiD测序、Polony测序和454测序使用乳液PCR来克隆扩增用于测序的DNA分子，以及每种技术指定待扩增的分子的最小和最大长度。待测序为克隆扩增核酸的标记分子的长度可以高达约600bp。在一些实施方式中，待测序的标记分子的长度可以大于600bp。

[0321] 单分子测序技术，其并不采用分子的克隆扩增，并且能够在模板长度的非常广泛的范围内测序核酸，在大多数情况下并不要求，待测序的分子具有任何特定长度。然而，序列产率/单位质量取决于3'端羟基的数目，因而具有用于测序的相对较短模板是比具有长模板更加有效的。如果开始于长于1000nt的核酸，则通常可取的是，将核酸剪切到100至200nt的平均长度，以致更多的序列信息可以产生自同样质量的核酸。因此，标记分子的长度可以为几十碱基至数千碱基。用于单分子测序的标记分子的长度可以高达约25bp、高达约50bp、高达约75bp、高达约100bp、高达约200bp、高达约300bp、高达约400bp、高达约500bp、高达约600bp、高达约700bp、高达约800bp、高达约900bp、高达约1000bp、或更大的长度。

[0322] 还通过待测序的基因组核酸的长度来确定为标记分子选择的长度。例如，作为细胞基因组DNA的基因组片段，cfDNA在人血流中循环。在孕妇的血浆中发现的胎儿cfDNA分子通常短于母体cfDNA分子(Chan等人，Clin Chem 50:8892[2004])。循环胎儿DNA的尺寸分级分离已证实，循环胎儿DNA片段的平均长度是＜300bp，而母体DNA已被估计为约0.5至1Kb(Li等人，Clin Chem，50:1002-1011[2004])。这些发现是一致与Fan等人的那些发现，其利用NGS确定了胎儿cfDNA很少＞340bp(Fan等人，Clin Chem56:1279-1286[2010])。借助于标准的基于二氧化硅的方法分离自尿的DNA由两个部分组成：高分子量DNA，其来源于脱落的细胞；以及低分子量(150-250个碱基对)部分的经肾DNA(Tr-DNA)(Botezatu等人，Clin Chem.46:1078-1084，2000；和Su等人，J Mol.Diagn.6:101-107，2004)。用于从体液分离无细胞核酸的新开发的技术应用于经肾核酸的分离已揭示了在尿中DNA和RNA片段(远短于150个碱基对)的存在(美国专利申请公开号20080139801)。在实施方式中，其中cfDNA是待测序的基因组核酸，选择的标记分子可以高达约cfDNA的长度。例如，在母体DNA样品中使用的待测序为单核酸分子或为克隆扩增核酸的标记分子的长度可以是约100bp至600。在其他实施方式中，样品基因组核酸是较大分子的片段。例如，被测序的样品基因组核酸是片段化细胞DNA。在实施方式中，当测序片段化细胞DNA时，标记分子的长度可以高达DNA片段的长度。在一些实施方式中，标记分子的长度至少是为独特地将序列读数映射到适当的参比基因组所需要的最小长度。在其他实施方式中，标记分子的长度是为排除标记分子被映射到样品参比基因组所需要的最小长度。

[0323] 此外，标记分子可以用来确认这样的样品，其没有通过核酸测序加以测定，以及其可以通过不同于测序的常见的生物技术，例如，实时PCR，加以证实。

[0324] 样品对照(例如，在用于测序和/或分析的过程阳性对照中)

[0325] 在不同的实施方式中，引入样品的标记序列，例如，如上面描述的，可以作为阳性对照来证实测序以及随后的处理和分析的准确性和有效性。

[0326] 因此，提供了成分和方法，其用来提供用于测序在样品中的DNA的处理中阳性对照(IPC)。在某些实施方式中，提供了阳性对照，其用于测序在包含基因组的混合物的样品中的cfDNA。IPC可以用来关联在获自不同组样品的序列信息中的基线位移，例如，在不同时间

在不同的测序运行中测序的样品。因此，例如IPC可以使针对母体测试样品获得的序列信息相关与获自一组合格样品的在不同时间测序的序列信息。

[0327] 同样地，在片段分析的情况下，IPC可以使获自受试者的并针对特定片段的序列信息相关与获自(类似序列的)一组合格样品并在不同时间测序的序列。在某些实施方式中，IPC可以使获自受试者并针对特定癌症相关基因座的序列信息相关与获自一组合格样品(例如，来自已知的扩增/缺失等)的序列信息。

[0328] 此外，IPC可以用作标记物来追踪样品(通过测序过程)。IPC还可以提供定性阳性序列剂量值，例如，NCV，用于感兴趣的染色体的一种或多种非整倍体，例如，三体性21、三体性13、三体性18，以提供适当的解释，以及确保数据的可信性和准确性。在某些实施方式中，可以产生IPC以比较来自雄性和雌性基因组的核酸，进而提供针对在母体样品中的染色体X和Y的剂量，以确定胎儿是否是雄性(男性)的。

[0329] 过程中对照的类型和数目取决于所需测试的类型或特性。例如，对于测试，其需要测序来自包含基因组的混合物的样品的DNA以确定染色体非整倍体是否存在，过程中对照可以包含获自样品(已知包含待测试的同样的染色体非整倍体)的DNA。在一些实施方式中，IPC包括这样的DNA，其来自已知包含感兴趣的染色体的非整倍体的样品。例如，用于确定在母体样品中胎儿三体性，例如，三体性21，的存在或不存在的测试的IPC包含获自具有三体性21的个体的DNA。在一些实施方式中，IPC包含获自具有不同非整倍体的两个或更多个个体的DNA的混合物。例如，对于用来确定三体性13、三体性18、三体性21、和单体性X的存在或不存在的测试，IPC包含获自孕妇(各携带具有待测试的三体性之一的胎儿)的DNA样品的组合。除完全染色体非整倍体之外，还可以产生IPC以提供这样的阳性对照，其用于用来确定部分非整倍体的存在或不存在的测试。

[0330] 可以利用获自两个受试者(一个是非整倍体基因组的贡献者)的细胞基因组DNA的混合物来产生作为用于检测单非整倍体的对照的IPC。例如，可以通过结合来自携带三体染色体的雄性或雌性受试者的基因组DNA与已知不携带三体染色体的雌性受试者的基因组DNA，来产生这样的IPC，其被产生为用于用来确定胎儿三体性，例如，三体性21，的测试的对照。基因组DNA可以提取自两个受试者的细胞，并被剪切以提供约100-400bp、约150-350bp、或约200-300bp的片段，从而模拟在母体样品中的循环cfDNA片段。选择来自携带非整倍体，例如，三体性21，的受试者的片段化DNA的比例，以模拟在母体样品中发现的循环胎儿cfDNA的比例，从而提供这样的IPC，其包含片段化DNA的混合物，其中包含来自携带非整倍体的受试者的约5％、约10％、约15％、约20％、约25％、约30％的DNA。上述IPC可以包含来自各携带不同非整倍体的不同受试者的DNA。例如，IPC可以包含约80％的未受影响的雌性DNA，以及剩余20％可以是来自各携带三体染色体21、三体染色体13、和三体染色体18的三个不同受试者的DNA。制备用于测序的片段化DNA的混合物。片段化DNA的混合物的处理可以包括制备测序文库，其可以利用任何大规模平行方法并以单重或多重方式加以测序。可以存储基因组IPC的储备溶液并用于多个诊断测试。

[0331] 可替换地，可以利用获自已知携带具有已知染色体非整倍体的胎儿的母体的cfDNA来产生IPC。例如，cfDNA可以获自携带具有三体性21的胎儿的孕妇。cfDNA提取自母体样品，并被克隆入细菌载体且在细菌中生长，以提供IPC的持续源。利用限制酶，DNA可以提取自细菌载体。可替换地，可以通过，例如，PCR，来扩增克隆cfDNA。可以处理IPC DNA，用于

在和来自测试样品的cfDNA相同的运行中加以测序，其中将分析测试样品中染色体非整倍体的存在或不存在。

[0332] 虽然上文参照三体性描述了IPC的产生，但是应当理解的是，可以产生IPC来反映其他部分非整倍体，包括，例如，各种片段扩增和/或缺失。因此，例如，在各种癌症已知是相关与特定扩增(例如，乳癌相关与20Q13)的情况下，可以产生IPC，其结合那些已知的扩增。

[0333] 测序方法

[0334] 如上文所指出的，作为用于确定拷贝数变异的程序的一部分，测序制备的样品(例如，测序文库)。可以利用若干测序技术的任何一种。

[0335] 一些测序技术是市售的，如来自Affymetrix Inc.(Sunnyvale,CA)的杂交测序平台和来自454 Life Sciences(Bradford,CT)、Illumina/Solexa(Hayward,CA)和Helicos Biosciences(Cambridge,MA)的合成测序平台、以及来自Applied Biosystems(Foster City,CA)的连接测序平台(如下所述)。除利用Helicos Biosciences的合成测序进行的单分子测序之外，其他单分子测序技术包括但不限于Pacific Biosciences的SMRT™技术、ION TORRENTTM技术、和例如，由Oxford Nanopore Technologies开发的纳米孔测序。

[0336] 虽然自动化桑格方法被看作‘第一代’技术，但桑格测序(包括自动化桑格测序)还可以用于本文描述的方法。另外合适的测序方法包括但不限于核酸成像技术，例如，原子力显微法(AFM)或透射电子显微术(TEM)。下文更详细地描述说明性测序技术。

[0337] 在一种说明性的但非限制性的实施方式中，本文描述的方法包括获得在测试样品中核酸的序列信息，例如，在母体样品中的cfDNA，在被筛查癌症的受试者中的cfDNA或细胞DNA等，其中利用Helicos True单分子测序(tSMS)技术的单分子测序技术(例如，如在Harris T.D.等人，Science320:106-109[2008]中所描述的)。在tSMS技术中，将DNA样品切割成大约100至200个核苷酸的链，以及将多聚腺苷酸序列加入每个DNA链的3’端。通过添加荧光标记腺苷核苷酸来标记每均被涵盖链。然后将DNA链杂交于流动池，上述流动池含有数百万的寡T俘获位点，其被固定于流动池表面。在某些实施方式中，模板可以是在约亿个模板/cm2的密度下。然后将流动池装入仪器，例如，HeliScope™测序仪，以及激光照射流动池的表面，从而揭示了每个模板的位置。CCD照相机可以映射模板在流动池表面上的位置。然后切割和冲走模板荧光标记。通过引入DNA聚合酶和荧光标记核苷酸来开始测序反应。寡T核酸作为引物。以模板引导方式，聚合酶将标记核苷酸并入引物。除去聚合酶和未并入的核苷酸。通过成像流动池表面来辨别已指导荧光标记核苷酸的并入的模板。在成像之后，切割步骤除去荧光标记，然后借助于其他荧光标记核苷酸来重复上述过程，直至达到所期望的读数长度。借助于每个核苷酸添加步骤来收集序列信息。在测序文库的制备中，通过单分子测序技术的全基因组测序排除或通常避免基于PCR的扩增，以及所述方法允许直接测量样品，而不是测量上述样品的拷贝。

[0338] 在另一种说明性的但非限制性的实施方式中，本文描述的方法包括获得在测试样品中核酸的序列信息，例如，在母体测试样品中的cfDNA，在被筛查癌症的受试者中的cfDNA或细胞DNA等，其中利用454测序(Roche)(例如在Margulies,M.等人Nature 437:376-380[2005]中所描述的)。454测序通常涉及两个步骤。在第一步骤中，将DNA剪切成大约300-800个碱基对的片段，以及平端片段。然后将寡核苷酸衔接子连接于片段的末端。衔接子充当用于片段的扩增和测序的引物。利用，例如，衔接子B，其含有5’-生物素标记，可以将片段附着

于DNA俘获珠，例如，链霉亲和素涂层珠。在油水乳液的液滴内PCR扩增附着于珠的片段。结果是在每个珠上多个拷贝的克隆扩增DNA片段。在第二步骤中，在孔(例如，皮升尺寸孔)中俘获珠。对每个DNA片段平行进行焦磷酸测序。一个或多个核苷酸的加成产生由测序仪中的CCD照相机所记录的光信号。信号强度正比于并入的核苷酸的数目。焦磷酸测序利用了焦磷酸(PPi)，其是在核苷酸加成之后被释放。在腺苷5'磷酸硫酸的存在下通过ATP硫酸化酶将PPi转化为ATP。荧光素酶使用ATP来将荧光素转化成氧化萤光素，以及此反应产生被测量和分析的光。

[0339] 在另一种说明性但非限制性的实施方式中，本文描述的方法包括利用SOLiD™技术(Applied Biosystems)来获得在测试样品中核酸的序列信息，例如，在母体测试样品中的cfDNA，在被筛查癌症的受试者中的cfDNA或细胞DNA等。在SOLiD™连接测序中，将基因组DNA剪切成片段，并将衔接子附着于片段的5'和3'端，以产生片段文库。可替换地，可以引入内部衔接子：通过将衔接子连接于片段的5'和3'端，环化片段，消化环化的片段以产生内部衔接子，然后将衔接子附着于得到的片段的5'和3'端，以产生伴侣配对文库。其次，在含有珠、引物、模板、和PCR成分的微反应器中制备克隆珠群体。在PCR之后，使模板变性并富集珠以分离珠与延伸的模板。可以对在所选珠上的模板进行3'修饰，其允许结合于载玻片。可以通过部分随机寡核苷酸与通过特定荧光团确定的中央确定的碱基(或碱基对)的连续杂交和连接来确定序列。在记录颜色之后，切割并除去连接的寡核苷酸，然后重复上述过程。

[0340] 在另一种说明性但非限制性的实施方式中，本文描述的方法包括利用Pacific Biosciences的单分子实时(SMRT™)测序技术来获得在测试样品中核酸的序列信息，例如，在母体测试样品中的cfDNA，在被筛查癌症的受试者中的cfDNA或细胞DNA等。在SMRT测序中，在DNA合成过程中成像染料标记核苷酸的连续并入。将单DNA聚合酶分子附着于获得序列信息的个别零模式波长检测器(ZMW检测器)的底表面，同时将磷酸联核苷酸并入不断增长的引物链。ZMW检测器包括约束结构，该结构使得能够相对于荧光核苷酸(其在ZMW之外迅速扩散(例如，在数微秒内))的背景来观察单核苷酸的并入(通过DNA聚合酶)。通常需要几毫秒来将核苷酸并入不断增长的链。在这段时间期间，荧光标记被激发并产生荧光信号，然后切割掉荧光标记。染料的相应荧光的测量指示哪个碱基被并入。重复上述过程以提供序列。

[0341] 在另一种说明性的但非限制性的实施方式中，本文描述的方法包括利用纳米孔测序(例如，在Soni GV和Meller A.Clin Chem 53:1996-2001[2007]中所描述的)来获得在测试样品中核酸的序列信息，例如，在母体测试样品中的cfDNA，在被筛查癌症的受试者中的cfDNA或细胞DNA等。纳米孔测序DNA分析技术是由若干公司所开发，包括，例如，Oxford Nanopore Technologies(Oxford，United Kingdom)、Sequenom、NABsys等。纳米孔测序是单分子测序技术，借此，当它经过纳米孔时，直接测序DNA的单分子。纳米孔是小孔，直径通常大约为1纳米。纳米孔在导电流体中的浸没和整个它电势(电压)的应用导致轻微的电流，这是由于离子通过纳米孔的传导。流动的电流量是对纳米孔的尺寸和形状敏感的。当DNA分子经过纳米孔时，在DNA分子上的每个核苷酸在不同的程度上阻塞纳米孔，从而在不同的程度上改变通过纳米孔的电流的大小。因此，当DNA分子经过纳米孔时，电流的这种变化会提供DNA序列的读数。

[0342] 在另一种说明性但非限制性的实施方式中，本文描述的方法包括利用化学敏感的

场效应晶体管(chemFET)阵列(例如,如在美国专利申请公开号2009/0026082中所描述的)来获得在测试样品中核酸的序列信息,例如,在母体测试样品中的cfDNA,在被筛查癌症的受试者中的cfDNA或细胞DNA等。在这种技术的一个实施例中,可以将DNA分子放入反应室,并且可以将模板分子杂交于结合于聚合酶的测序引物。通过chemFET,在测序引物的3'端处一个或多个三磷酸到新核酸链的并入可以被辨别为电流的变化。阵列可以具有多个chemFET传感器。在另一个实例中,可以将单核酸附着于珠,并可以在珠上扩增核酸,然后可以将单个珠转移到在chemFET阵列上的单个反应室,其中每个室具有chemFET传感器,接着可以测序核酸。

[0343] 在另一种实施方式中,本方法包括利用Halcyon Molecular技术,其使用透射电子显微术(TEM),来获得在测试样品中核酸的序列信息,例如,在母体测试样品中的cfDNA。被称为单分子位置快速纳米转移(IMPRNT)的方法包括利用用重原子标记选择性地标记的高分子量(150kb或更大)DNA的单原子分辨率透射电子显微镜成像以及以具有一致的碱基-碱基间距的超密集的(3nm链到链)的平行阵列在超薄膜上安排这些分子。电子显微镜用来成像在薄膜上的分子以确定重原子标记的位置以及从DNA析取碱基序列信息。所述方法进一步描述于PCT专利公开WO 2009/046445。所述方法允许在不到十分钟内测序完全人类基因组。

[0344] 在另一种实施方式中,DNA测序技术是Ion Torrent单分子测序,其使半导体技术和简单测序化学成对以直接在半导体芯片上将化学编码信息(A、C、G、T)转换成数字信息(0、1)。在自然界中,当通过聚合酶将核苷酸并入DNA的链时,释放氢离子作为副产物。Ion Torrent使用微机械加工孔的高密度阵列,从而以大规模平行方式来进行此生化过程。各个孔包含不同的DNA分子。在孔下方是离子敏感层以及在其下方是离子传感器。当将核苷酸,例如C,加入DNA模板然后并入DNA的链时,将释放氢离子。来自上述离子的电荷将改变溶液的pH,其可以通过Ion Torrent离子传感器加以检测。测序仪,基本上是世界上最小的固态pH计,调用碱基,从而直接从化学信息转到数字信息。然后Ion个人类基因组机(PGM™)测序仪用一个接一个的核苷酸依次充斥芯片。如果充斥芯片的下一核苷酸不是匹配,则将记录不到电压变化并且将不调用碱基。如果在DNA链上存在两个相同的碱基,则电压将是双倍,并且芯片将记录调用的两个相同的碱基。直接检测允许记录核苷酸并入(在几秒钟内)。

[0345] 在另一种实施方式中,本方法包括利用杂交测序来获得在测试样品中核酸的序列信息,例如,在母体测试样品中的cfDNA。杂交测序包括使多种多核苷酸序列接触多种多核苷酸探针,其中多种多核苷酸探针的每一种可以被可选地束缚于基板。基板可以具有包含已知核苷酸序列的阵列的平整表面。杂交于阵列的模式可以用来确定在样品中存在的多核苷酸序列。在其他实施方式中,将每个探针束缚于珠,例如,磁珠等。于珠的杂交可以被确定并且用来识别在样品内的多种多核苷酸序列。

[0346] 在另一种实施方式中,本方法包括通过数百万的DNA片段的大规模平行测序并利用Illumina合成测序和基于可逆终止子的测序化学(例如,如在Bentley等人,Nature 6:53-59[2009]中所描述的)来获得在测试样品中核酸的序列信息,例如,在母体测试样品中的cfDNA。模板DNA可以是基因组DNA,例如,cfDNA。在一些实施方式中,来自分离细胞的基因组DNA用作模板,并且它被片段化成数百碱基对的长度。在其他实施方式中,cfDNA用作模板,并且不需要片段化,因为cfDNA存在为短片段。例如胎儿cfDNA在血流中循环,作为长度

为大约170个碱基对(bp)的片段(Fan等人,Clin Chem 56:1279-1286[2010]),以及在测序之前不需要DNA的片段化。Illumina测序技术依赖于片段化基因组DNA附着于其上结合寡核苷酸锚形体的平面的光学透明表面。末端修复模板DNA以产生5'-磷酸化平端,以及Klenow片段的聚合酶活性用来将单A碱基加入平整的磷酸化DNA片段的3'端。这种加成可制备DNA片段,用于连接于寡核苷酸衔接子,其在它们的3'端处具有单T碱基的突出端,以增加连接效率。衔接子寡核苷酸互补于流动池锚形物。在限制稀释均被涵盖件下,将衔接子修饰的单链模板DNA加入流动池并通过杂交于锚形物加以固定。延伸和桥接扩增附着的DNA片段以产生具有数以亿计簇的超高密度测序流动池,各自包含相同模板的～1,000个拷贝。在一种实施方式中,在经受群集扩增之前,利用PCR来扩增随机片段化基因组DNA,例如,cfDNA。可替换地,使用无扩增基因组文库制备,并利用单独的群集扩增来富集随机片段化基因组DNA,例如,cfDNA(Kozarewa等人,Nature Methods 6:291-295[2009])。利用鲁棒四色DNA合成测序技术,其采用可逆终止子并借助于可除去的荧光染料,来测序模板。利用激光激发和全内反射光学元件来实现高灵敏度荧光检测。将约20-40bp,例如,36bp,的短序列读数比对于重复掩蔽的参比基因组并利用专门开发的数据分析流水线软件来确定短序列读数到参比基因组的独特映射。还可以使用非重复掩蔽的参比基因组。不论使用重复掩蔽的或非重复掩蔽的参比基因组,仅计数独特映射到参比基因组的读数。在第一读取的完成之后,可以原位再生模板以使得能够从片段的相反端进行第二次读取。因此,可以使用DNA片段的单端或成对端测序。进行在样品中存在的DNA片段的部分测序,以及将包含预定长度,例如,36bp,的读数的序列标签映射到已知参比基因组,并加以计数。在一种实施方式中,参比基因组序列是NCBI36/hg18序列,其可以在万维网上并在genome.ucsc.edu/cgi-bin/hgGateway?org＝Human&db＝hg18&hgsid＝166260105)处获得。可替换地,参比基因组序列是GRCh37/hg19,其可以在万维网上并在genome.ucsc.edu/cgi-bin/hgGateway处获得。公开序列信息的其他来源包括GenBank、dbEST、dbSTS、EMBL(欧洲分子生物学实验室)、和DDBJ(日本的DNA数据库)。若干计算机算法可用于比对序列,包括但不限于BLAST(Altschul等人,1990)、BLITZ(MPsrch)(Sturrock&Collins,1993)、FASTA(Person&Lipman,1988)、BOWTIE(Langmead等人,Genome Biology 10:R25.1-R25.10[2009])、或ELAND(Illumina,Inc.,San Diego,CA,USA)。在一种实施方式中,测序血浆cfDNA分子的克隆扩增拷贝的一端并通过针对Illumina基因组分析仪,其使用核苷酸数据库的有效的大规模比对(ELAND)软件,的生物信息学比对分析,加以处理。

[0347] 在本文描述的方法的一些实施方式中,映射的序列标签包含约20bp、约25bp、约30bp、约35bp、约40bp、约45bp、约50bp、约55bp、约60bp、约65bp、约70bp、约75bp、约80bp、约85bp、约90bp、约95bp、约100bp、约110bp、约120bp、约130、约140bp、约150bp、约200bp、约250bp、约300bp、约350bp、约400bp、约450bp、或约500bp的序列读数。预期,技术进步将使得能够产生大于500bp的单端读数,从而当产生配对末端读数时使得能够获得大于约1000bp的读数。在一种实施方式中,映射的序列标签包含为36bp的序列读数。通过比较标记的序列与参比的序列来实现序列标签的映射,以确定测列的核酸(例如cfDNA)分子的染色体起源,以及不需要特定基因序列信息。可以允许小程度的错配(0-2个错配/序列标签)以说明在参比基因组和在混合样品中的基因组之间可能存在的次要多态性。

[0348] 对于每个样品,通常获得多个序列标签。在一些实施方式中,对于每个样品,至少

约3×10^6个序列标签、至少约5×10^6个序列标签、至少约8×10^6个序列标签、至少约10×10^6个序列标签、至少约15×10^6个序列标签、至少约20×10^6个序列标签、至少约30×10^6个序列标签、至少约40×10^6个序列标签、或至少约50×10^6个序列标签(其包含20至40bp读数，例如，36bp)获自将读数映射到参比基因组。在一种实施方式中，将所有序列读数映射到参比基因组的所有区。在一种实施方式中，计数已映射到参比基因组的所有区，例如，所有染色体，的标记，并在混合DNA样品中确定CNV，即，感兴趣的序列，例如，染色体或其部分，的过高或过低表示。所述方法不需要在两个基因组之间的区分。

[0349] 基于在测序运行内映射到在样品中中的参比基因组的序列标签的数目的变化(染色体间变异性)、和在不同测序运行中映射到参比基因组的序列标签的数目的变化(测序间的变异性)，来预测为正确确定是否CNV，例如，非整倍体，在样品中存在或不存在所需要的准确性。例如，对于映射到富含GC或GC贫乏的参比序列的标记，变化可能是特别明显的。其他变体可能来自对于核酸的提取和纯化、测序文库的制备、和不同测序平台的使用，使用了不同方法。基于归一化序列(归一化染色体序列或归一化片段序列)的知识，本方法使用序列剂量(染色体剂量、或片段剂量)，以本质上说明起源于染色体间的(运行内)、和测序间的(运行间)的累计变异性以及依赖于平台的变异性。染色体剂量是基于归一化染色体序列的知识，其可以组成自单染色体，或两个或两个以上的染色体，其选自染色体1-22、X、和Y。可替换地，归一化染色体序列可以组成自单染色体片段，或一个染色体或两个或两个以上的染色体的两个或更多片段。片段剂量是基于归一化片段序列的知识，其可以组成自任何一个染色体的单片段，或染色体1-22、X、和Y的任何两个或更多的两个或更多片段。

[0350] CNV和产前诊断

[0351] 在母体血液中循环的无细胞胎儿DNA和RNA可以用于越来越多的遗传性疾病的早期非侵入性产前诊断(NIPD)，用于妊娠管理和生殖决策。已经知道在血流中循环的无细胞DNA的存在超过50年。最近，发现在妊娠期间的母体血流中存在少量循环胎儿DNA(Lo等人，Lancet 350:485-487[1997])。认为来源于垂死的胎盘细胞，无细胞胎儿DNA(cfDNA)已经表明，由长度小于200bp的短片段组成(Chan等人，Clin Chem 50:88-92[2004])，其可以早在4周妊娠加以辨别(Illanes等人，Early Human Dev83:563-566[2007])，并且已知在递送数小时内被清除自母体循环(Lo等人，Am J Hum Genet 64:218-224[1999])。除cfDNA之外，在母体血流中还可以辨别无细胞胎儿RNA(cfRNA)的片段，其来自在胎儿或胎盘中被转录的基因。来自母体血液样品的这些胎儿遗传元件的提取和随后的分析提供了用于NIPD的新机会。

[0352] 本方法是不依赖多态性的方法，其用于NIPD并且其不需要胎儿cfDNA区别于母体cfDNA来能够确定胎儿非整倍性。在一些实施方式中，非整倍体是完全染色体三体性或单体性、或部分三体性或单体性。部分非整倍体起因于部分染色体的损失或收益，以及涵盖染色体失衡，其来自不平衡的易位、不平衡的倒置、缺失和插入。到目前为止，相容于生命的最常见的已知的非整倍体是三体性21，即，唐氏综合征(DS)，其起因于部分或全部的21号染色体的存在。很少地，DS可能起因于遗传性或散发性缺陷，借此全部或部分的21号染色体的额外拷贝变成附着于另一染色体(通常14号染色体)以形成单异常染色体。DS相关与智能缺陷、严重的学习困难和死亡率过高，其起因于长期健康问题如心脏病。具有已知临床意义的其他非整倍体包括爱德华综合征(三体性18)和帕陶综合征(三体性13)，在生命的最初几个月

内，其经常是致命的。相关与性染色体的数目的异常也是已知的并且包括单体性X，例如，特纳综合征(XO)、和在女婴中的三X染色体综合征(XXX)以及在男婴中的克兰费尔特综合征(XXY)和XYY综合征，其均相关与各种表型，包括不育性和智力技能的降低。单体性X[45,X]是妊娠早期流产的常见原因，其占约7％的自然流产。基于1-2/10,000的45,X(还称为特纳综合征)的活产率，估计，小于1％的45,X受孕将生存至足月。约30％的特纳综合征患者嵌合有45,X细胞系和46,XX细胞系或含有重排X染色体的一种细胞系(Hook和Warburton 1983)。考虑到高胎儿致死率，在活产婴儿中的表型是相对温和的，并且已假设，患有特纳综合征的所有活产女婴可能携带含有两个性染色体的细胞系。单体性X在雌性中可能发生为45,X或45,X/46XX，以及在雄性中可能发生为45,X/46XY。通常认为在人中的常染色体单体性不相容于生命；然而，有相当多的细胞遗传学报告，其描述了在活产儿童中一个21号染色体的全单体性(Vosranova I等人，Molecular Cytogen.1:13[2008]；Joosten等人，Prenatal Diagn.17:271-5[1997]。本文描述的方法可以用来产前诊断这些和其他染色体异常。

[0353] 根据一些实施方式，本文披露的方法可以确定染色体1-22、X和Y中的任一种的染色体三体性的存在或不存在。可以根据本方法加以检测的染色体三体性的实例包括但不限于三体性21(T21；唐氏综合征)、三体性18(T18；爱德华综合征)、三体性16(T16)、三体性20(T20)、三体性22(T22；猫眼综合征)、三体性15(T15；普拉德-威利综合征)、三体性13(T13；帕陶综合征)、三体性8(T8；Warkany综合征)、三体性9、以及XXY(克兰费尔特综合征)、XYY、或XXX三体性。以非嵌合状态存在的其他常染色体的完全三体性是致命的，但当以嵌合状态存在时，可以相容于生命。应当理解的是，根据本文中提供的教导，可以在胎儿cfDNA中确定各种完全三体性，不论以嵌合或非嵌合状态存在，以及部分三体性。

[0354] 可以通过本方法来确定的部分三体性的非限制性实例包括但不限于部分三体性1q32-44、三体性9p、三体性4嵌合性、三体性17p、部分三体性4q26-qter、部分2p三体性、部分三体性1q、和/或部分三体性6p/单体性6q。

[0355] 本文披露的方法还可以用来确定染色体单体性X、染色体单体性21、和部分单体性如，单体性13、单体性15、单体性16、单体性21、和单体性22，其已知涉及妊娠流产。还可以通过本文描述的方法来确定通常涉及完全非整倍体的染色体的部分单体性。可以根据本方法来确定的缺失综合征的非限制性实例包括由染色体的部分缺失引起的综合征。可以根据本文描述的方法来确定的部分缺失的实例包括但不限于染色体1、4、5、7、11、18、15、13、17、22和10的部分缺失，其在下文中描述。

[0356] 1q21.1缺失综合征或1q21.1(复发性)微缺失是1号染色体的罕见畸变。紧邻缺失综合征，还存在1q21.1重复综合征。虽然在患有缺失综合征的情况下，在特定点上存在一部分的DNA丢失，但在患有重复综合征的情况下，在相同点上存在DNA的类似部分的两个或三个拷贝。文献提到缺失和重复作为1q21.1拷贝数变异(CNV)。1q21.1缺失可能相关与TAR综合征(具有Absent半径的血小板减少)。

[0357] 沃-希综合征(WHS)(OMIN#194190)是相邻基因缺失综合征，其相关与染色体4p16.3的半纯合子缺失。沃-希综合征是先天性畸形综合征，其特征是产前和产后生长不足、不同程度的发育性残疾、特征性颅面特点(鼻子的‘希腊武士头盔’外貌、高额、突出的眉间、距离过宽、高拱形眉、突出的眼睛、内眦赘皮、短人中、具有向下的角的截然不同的嘴、和小颌)、和癫痫症。

[0358] 5号染色体的部分缺失，还被称为5p-或5p减，并且命名为猫叫综合征(Cris du Chat syndrome)(OMIN#123450)，起因于5号染色体(5p15.3-p15.2)的短臂(p臂)的缺失。患有这种病症的婴儿经常具有高亢的叫声，其听上去像猫的叫声。上述疾病的特征是智力残疾和延迟发育、小头尺寸(小头畸形)、低出生体重、和弱肌张力(张力过低)(在婴儿期中)、独特的面部特征以及可能的心脏缺陷。

[0359] 威-布综合征(Williams-Beuren Syndrome)，也被称为染色体7q11.23缺失综合征(OMIN 194050)，是相邻基因缺失综合征，其导致起因于在染色体7q11.23，其含有大约28个基因，上的1.5至1.8Mb的半纯合子缺失的多系统疾病。

[0360] 雅各布森综合征(Jacobsen Syndrome)，也被称为11q缺失疾病，是罕见的先天性疾病，其来自11号染色体(其包括带11q24.1)的末端区的缺失。它可以引起智力残疾、独特的面部外观、和各种各样的身体问题，包括心脏缺陷和出血性疾病。

[0361] 18号染色体的部分单体性，被称为单体性18p，是罕见的染色体异常，其中18号染色体的全部或部分的短臂(p)被删除(单体的)。上述疾病是通常特征是身材矮小、不同程度的智力低下、语音延迟、头骨和面部(颅面)区域的畸形、和/或另外的身体异常。相关的颅面缺陷在范围和严重程度上可能因病例不同而相差很大。

[0362] 起因于15号染色体的拷贝的结构或数目的变化的病症包括安格尔曼综合征(Angelman Syndrome)和普拉德-威利综合征(Prader-Willi Syndrome)，其涉及在15号染色体的相同部分，15q11-q13区，中基因活性的丧失。应当理解的是，在载体亲本中若干易位和微缺失可以是无症状的，但可能在后代中会造成重大遗传疾病。例如，携带15q11-q13微缺失的健康母亲能生出患有安格尔曼综合征，一种严重的神经变性疾病，的孩子。因此，本文描述的方法、仪器和系统可以用来鉴定在胎儿中的这样的部分缺失和其他缺失。

[0363] 部分单体性13q是当丢失13号染色体的一均被涵盖长臂(q)(单体的)时导致的罕见的染色体异常。天生具有部分单体性13q的婴儿可能表现出低出生体重、头部和脸部(颅面区)的畸形、骨骼异常(尤其是手和脚的骨骼异常)、和其他身体异常。智力低下是这种病症的特性。在婴儿期期间在天生患有这种疾病的个体中死亡率是高的。部分单体性13q的几乎所有病例是随机发生而没有明显的理由(散发的)。

[0364] 史密斯-马吉利综合征(Smith-Magenis syndrome)(SMS-OMIM#182290)起因于在17号染色体的一个拷贝上遗传物质的缺失、或丢失。这种众所周知的综合征相关与发育延迟、智力低下、先天性异常如心脏和肾脏缺陷、和神经行为异常如严重的睡眠障碍和自伤行为。在大多数情况下(90％)，史密斯-马盖尼斯综合征(SMS)起因于在17号染色体p11.2中的3.7-Mb中间缺失。

[0365] 22q11.2缺失综合征，还被称为迪格奥尔格综合征(DiGeorge syndrome)，是起因于22号染色体的小部分的缺失的综合征。上述缺失(22q11.2)发生在靠近染色体的中间并在染色体对的长臂上。这种综合征的特征有很大的不同，甚至在同一家族的成员中，并且影响身体的许多部分。特征性体征和症状可以包括出生缺陷如先天性心脏疾病、腭缺陷，最常涉及到关于闭合的神经肌肉问题(腭咽关闭不全)、学习障碍、面部特征的温和差异、和复发性感染。在染色体区22q11.2中的微缺失相关与精神分裂症的20至30倍增加的风险。

[0366] 在10号染色体的短臂上的缺失相关与迪格奥尔格综合征(DiGeorge Syndrome)样表型。10号染色体p的部分单体性是罕见的，但已在显示迪格奥尔格综合征的特征的一部分

患者中观测到。

[0367] 在一种实施方式中，本文描述的方法、装置、和系统用来确定部分单体性，包括但不限于染色体1、4、5、7、11、18、15、13、17、22和10的部分单体性，例如，利用所述方法还可以确定部分单体性1q21.11、部分单体性4p16.3、部分单体性5p15.3-p15.2、部分单体性7q11.23、部分单体性11q24.1、部分单体性18p、15号染色体的部分单体性(15q11-q13)、部分单体性13q、部分单体性17p11.2、22号染色体的部分单体性(22q11.2)、以及部分单体性10p。

[0368] 根据本文描述的方法可以确定的其他部分单体性包括不平衡的易位t(8;11)(p23.2;p15.5)；11q23微缺失；17p11.2缺失；22q13.3缺失；Xp22.3微缺失；10p14缺失；20p微缺失，[del(22)(q11.2q11.23)]，7q11.23和7q36缺失；1p36缺失；2p微缺失；1型神经纤维瘤病(17q11.2微缺失)，Yq缺失；4p16.3微缺失；1p36.2微缺失；11q14缺失；19q13.2微缺失；鲁-泰综合症(16p13.3微缺失)；7p21微缺失；米-迪综合征(17p13.3)；和2q37微缺失。部分缺失可以是部分染色体的小缺失，或它们可以是染色体的微缺失，其中可以发生单基因的缺失。

[0369] 已确定起因于部分的染色体臂的重复的一些重复综合征(参见OMIN[Online Mendelian Inheritance in Man viewed online at ncbi.nlm.nih.gov/omim])。在一种实施方式中，本方法可以用来确定染色体1-22、X和Y中的任一种的片段的复制和/或扩增的存在或不存在。可以根据本方法加以确定的重复综合征的非限制性实例包括染色体8、15、12、和17的部分的重复，其在下文中描述。

[0370] 8p23.1重复综合征是罕见的遗传紊乱，其起因于来自人8号染色体的区的重复。这种重复综合征具有在64,000个出生中有一例的估计流行率并且互相关联与8p23.1缺失综合征。8p23.1重复相关与可变表型，包括语音延迟、发育延迟、轻度畸形、具有突出的额和拱形眉、和先天性心脏疾病(CHD)的一种或多种。

[0371] 15号染色体q重复综合征(Chromosome 15q Duplication Syndrome)(Dup15q)是临床上可识别的综合征，其产生于15号染色体q11-13.1的重复。具有Dup15q的婴通常具有张力过低(不良的肌张力)、生长迟缓；他们可能天生具有裂唇和/或腭或心脏、肾脏或其他器官的畸形；他们显示出一定程度的认知迟缓/残疾(智力低下)、言语和语言延迟、以及感觉处理障碍。

[0372] 帕里斯特基利安综合征(Pallister Killian syndrome)是额外的第12号染色体材料的结果。通常存在细胞的混合物(嵌合性)，一些具有额外的第12号材料，以及一些是正常的(46均被涵盖染色体而没有额外的第12号材料)。患有这种综合征的婴儿具有许多问题，包括重度智力低下、不良的肌张力、“粗糙的”面部特征、和突出的额头。他们倾向于具有非常薄的上唇、较厚的下唇和短鼻。其他健康问题包括癫痫发作、喂养困难、关节僵硬、在成年期白内障、听力丧失、和心脏缺陷。患有帕里斯特基利安综合征的人具有缩短的寿命。

[0373] 患有遗传性疾病，被指定为dup(17)(p11.2p11.2)或dup 17p，的个体携带在17号染色体的短臂上的额外的遗传信息(被称为重复)。17号染色体p11.2的重复构成波托茨基-Lupski综合征(Potocki-Lupski syndrome)(PTLS)的基础，其是是在医学文献中仅报告有几十病例的新认定的遗传性疾病。具有这种重复的患者经常具有低肌张力、喂养困难、和在婴儿期期间不能生长发育、以及还存在有运动和言语方面的延迟发育。患有PTLS的许多个

体在发音和语言处理方面具有困难。此外，患者可以具有这样的行为特征，其类似于在患有自闭症或自闭症-谱群疾病的人中看到的那些行为特征。患有PTLS的个体可以具有心脏缺陷和睡眠呼吸暂停。在包括基因PMP22的17号染色体p12中较大区的重复已知会导致沙-马-图病。

[0374] CNV已经与死产相关联。然而，由于常规细胞遗传学的固有局限性，CNV对死产的贡献被认为是代表性不足的(Harris等人，Prenatal Diagn31:932-944[2011])。如在实施例中所示和在本文中别处描述的，本方法能够确定部分非整倍体的存在，例如，染色体片段的缺失和扩增，并且可以用来鉴定和确定相关与死产的CNV的存在或不存在。

[0375] 用于确定CNV的仪器和系统

[0376] 通常利用各种计算机执行的算法和程序来进行测序数据的分析和从其衍生的诊断。因此，某些实施方式采用这样的过程，其涉及存储在一个或多个计算机系统或其他处理系统中的数据或通过一个或多个计算机系统或其他处理系统传输的数据。本文披露的实施方式还涉及用于执行这些操作的仪器。这种仪器可以是为所需目的而专门构建，或它可以是通用计算机(或一组计算机)，其由计算机程序和/或存储在计算机中的数据结构所选择性地激活或重新配置。在一些实施方式中，一组处理器协作地(例如，经由网络或云计算)和/或平行地执行一些或所有的列举的分析操作。用于执行本文描述的方法的处理器或处理器组可以是各种类型，包括微控制器和微处理器如可编程器件(例如，CPLD和FPGA)以及不可编程器件如门阵列ASIC或通用微处理器。

[0377] 此外，某些实施方式涉及有形的和/或非暂时性计算机可读媒体或计算机程序产品，其包括用于执行各种计算机执行操作的程序指令和/或数据(包括数据结构)。计算机可读媒体的实例包括但不限于半导体存储器件，磁介质如磁盘驱动器、磁带，光学介质如CD、磁光介质，和硬件器件，其被特别配置以存储和执行程序指令，如只读存储器件(ROM)和随机存储器(RAM)。计算机可读媒体可以由最终用户来直接控制或可以由最终用户来间接控制。直接控制的媒体的实例包括位于用户设施处的媒体和/或并不与其他实体共享的媒体。间接控制的媒体的实例包括通过外部网络和/或通过提供共享资源的服务如“云”，用户间接可访问的媒体。程序指令的实例包括机器码，如由编译器产生的，和包含更高级别的代码的文件，其可以利用解释程序由计算机来执行。

[0378] 在不同的实施方式中，在所披露的方法和仪器中采用的数据或信息是以电子格式来提供。这样的数据或信息可以包括来源于核酸样品的读数和标记，比对与参比序列的特定区的标记(例如，其比对于染色体或染色体片段)的计数或密度，参比序列(包括唯一或主要地提供多态性的参比序列)，染色体和片段剂量，调用如非整倍体调用，归一化染色体和片段值，染色体或片段和相应的归一化染色体或片段的对，咨询建议，诊断等。如在本文中所使用的，以电子格式提供的数据或其他信息可用于在机器上的存储以及在机器之间的传输。传统上，数字地提供电子格式的数据并且可以以各种数据结构。列表、数据库等被存储为位和/或字节。可以以电子方式、以光学方式等来体现数据。

[0379] 一种实施方式提供了一种计算机程序产品，用于产生输出，其指示在测试样品中非整倍体，例如，胎儿非整倍性或癌症，的存在或不存在。计算机产品可以包含指令，用于执行用来确定染色体异常的任何一种或多种所述方法。如所解释的，上述计算机产品可以包括非暂时性和/或有形的计算机可读介质，其具有记录在其上的计算机可执行的或可编译

的逻辑(例如,指令),用于使处理器能够确定染色体剂量以及,在一些情况下,确定是否胎儿非整倍性是存在或不存在的。在一个实例中,上述计算机产品包含计算机可读介质,其具有记录在其上的计算机可执行的或可编译的逻辑(例如,指令),用于使处理器能够诊断胎儿非整倍性,包括:接收程序,用于接收来自母体生物样品的至少一部分的核酸分子的测序数据,其中所述测序数据包括计算的染色体和/或片段剂量;计算机辅助逻辑,用于依据所述接收的数据来分析胎儿非整倍性;以及输出程序,用于产生输出,其指示所述胎儿非整倍性的存在、不存在或种类。

[0380] 可以将来自在考虑中的样品的序列信息映射到染色体参比序列以识别针对感兴趣的任何一个或多个的染色体的每一个的若干序列标签以及识别针对感兴趣的所述任何一个或多个的染色体的每一个的归一化片段序列的若干序列标签。在不同的实施方式中,将参比序列存储在数据库中如例如关系或对象数据库。

[0381] 应当理解的是,不实际的是,或在大多数情况下甚至不可能的是,独立的人执行本文披露的方法的计算操作。例如,在没有计算仪器的帮助下,将来自样品的单30bp的读数映射到人染色体的任何之一可能需要几年的努力。当然,上述问题是复杂的,因为可靠的非整倍体调用通常需要将数千(例如,至少约10,000)或甚至数百万的读数映射到一个或多个染色体。

[0382] 可以利用用于在测试样品中感兴趣的基因序列的拷贝数的评价的系统来进行本文披露的方法。上述系统包括:(a)测序仪,用于接收来自测试样品的核酸,从而提供来自样品的核酸序列信息;(b)处理器;以及(c)一个或多个计算机可读存储介质,其具有存储在其上的用于在所述处理器上执行的指令,以进行用于确定任何CNV,例如,染色体或部分非整倍体,的方法。

[0383] 在一些实施方式中,通过计算机可读介质来指示所述方法,其中上述计算机可读介质具有存储在其上的用于实施用来确定任何CNV,例如,染色体或部分非整倍体,的方法的计算机可读指令。因此,一种实施方式提供了计算机程序产品,其包括一个或多个计算机可读非临时性存储介质,其具有存储在其上的计算机可执行指令,当由计算机系统的一个或多个处理器执行时其引起计算机系统实施用来对在包含胎儿和母体无细胞核酸的测试样品中的感兴趣的序列进行拷贝数的评价的方法。所述方法包括:(a)提供测试样品的序列读数;(b)将测试样品的序列读数比对与包含感兴趣的序列的参比基因组,从而提供测试序列标签;(c)确定位于每个bin中的测试序列标签的覆盖度,其中参比基因组被分成多个bin;(d)提供用于感兴趣的序列的全局配置参数,其中全局配置参数包含在每个bin中的预期覆盖度,以及其中预期覆盖度获自训练集的未受影响的训练样品,其以和测试样品基本相同的方式加以测序和比对,预期覆盖度呈现bin之间的变化;(e)根据在每个bin中的预期覆盖度,调节测试序列标签的覆盖度,从而获得在测试序列标签的每个bin中的全局配置参数修正的覆盖度;(f)针对测试序列标签的bin,基于在GC含量水平和全局配置参数修正的覆盖度之间的关系,来调节全局配置参数修正的覆盖度,从而获得在感兴趣的序列上的测试序列标签的样品-GC-修正的覆盖度;以及(g)基于样品-GC-修正的覆盖度来评价在测试样品中感兴趣的序列的拷贝数。在一些实施方式中,归一化在步骤(c)中确定的覆盖度。上述归一化可能涉及将覆盖度除以映射到鲁棒染色体的读数的总数或依据映射到鲁棒染色体的读数的总数来建模覆盖度(有时还被称为文库深度归一化)。

[0384] 在一些实施方式中，上述指令可以进一步包括自动记录有关与所述方法的信息如染色体剂量和在提供母体测试样品的人受试者的患者医疗记录中胎儿染色体非整倍体的存在或不存在。患者医疗记录可以由，例如，实验室、医生办公室、医院、健康维护组织、保险公司、或个人医疗记录万维网站来保留。另外，基于处理器实施的分析的结果，所述方法可以进一步涉及规定、开始、和/或改变从其取得母体测试样品的人受试者的治疗。这可能涉及对取自受试者的另外的样品进行一个或多个另外的测试或分析。

[0385] 还可以利用计算机处理系统，其被适应或配置以实施用于确定任何CNV，例如，染色体或部分非整倍体，的方法，来进行披露的方法。一种实施方式提供了计算机处理系统，其被适应或配置以实施如本文所描述的方法。在一种实施方式中，上述仪器包括测序装置，其被调整或配置用于测序在样品中的至少一部分的核酸分子以获得在本文中别处描述的序列信息的类型。上述仪器还可以包括用于处理样品的部件。这样的部件是在本文中别处描述的。

[0386] 可以将序列或其他数据直接或间接地输入计算机或存储在计算机可读介质上。在一种实施方式中，将计算机系统直接耦合到测序装置，其读取和/或分析来自样品的核酸的序列。通过在计算机系统中的接口来提供来自上述工具的序列或其他信息。可替换地，由系统处理的序列提供自序列存储源如数据库或其他存储库。在处理仪器可获得之后，存储器件或大容量存储器至少暂时地缓冲或存储核酸序列。此外，存储器件可以存储针对各种染色体或基因组等的标记计数。上述存储器还可以存储用于分析呈递序列或映射的数据的各种例程和/或程序。这样的程序/例程可以包括用于执行统计分析等的程序。

[0387] 在一个实例中，用户将样品提供到测序仪器。通过连接于计算机的测序仪器来收集和/或分析数据。在计算机上的软件允许数据收集和/或分析。可以对数据进行存储，显示(通过监视器或其他类似的装置)，和/或发送到另一个位置。可以将计算机连接到互联网，其用来将数据发送到由远程用户(例如，医师、科学家或分析师)使用的手持装置。可以理解的是，在发送之前，可以存储和/或分析数据。在一些实施方式中，收集原始数据并发送到远程用户或仪器，其将分析和/或存储数据。可以通过互联网，但也可以通过卫星或其他连接来进行发送。可替代地，可以将数据存储在计算机可读介质上，并且可以将上述介质运送到最终用户(例如，通过邮件)。远程用户可以处于相同或不同的地理位置，包括但不限于建筑物、城市、州、国家或大陆。

[0388] 在一些实施方式中，所述方法还包括收集关于多个多核苷酸序列的数据(例如，读数、标记和/或参比染色体序列)并将数据发送到计算机或其他计算系统。例如，可以将计算机连接到实验室设备，例如，样品收集仪器、核苷酸扩增仪器、核苷酸测序仪器、或杂交仪器。然后计算机可以收集由实验室装置收集的可适用的数据。可以在任何步骤中将数据存储在计算机上，例如，当实时收集时，在发送之前，在发送期间或连同发送一起，或在发送之后。可以将数据存储在计算机可读介质上，其可以析取自计算机。可以将收集或存储的数据从计算机发送到远程位置，例如，通过局部网络或广域网如互联网。在远程位置处，可以对发送的数据进行各种操作(如下所述)。

[0389] 在本文披露的系统、仪器、和方法中可以存储、发送、分析、和/或操作的电子格式数据中有如下：

[0390] 通过测序在测试样品中的核酸所获得的读数

[0391] 通过比对读数于参比基因组或其他参比序列所获得的标记

[0392] 参比基因组或序列

[0393] 序列标签密度-针对参比基因组或其他参比序列的两个或更多区(通常染色体或染色体片段)的每一个的标记的计数或数目

[0394] 针对感兴趣的特定染色体或染色体片段的归一化染色体或染色体片段的标识

[0395] 获自感兴趣的染色体或片段和相应的归一化染色体或片段的染色体或染色体片段(或其他区)的剂量

[0396] 用于调用染色体剂量作为受影响的、非受影响的、或无调用的的阈值

[0397] 染色体剂量的实际调用

[0398] 诊断(相关与调用的临床状况)

[0399] 来源于调用和/或诊断的用于进一步测试的建议

[0400] 来源于调用和/或诊断的治疗和/或监测计划

[0401] 不同的仪器,在一个或多个位置处,可以获得,存储,发送,分析,和/或操作这些各种类型的数据。处理选项跨越广泛的范围。在范围的一端,在处理测试样品的位置处,例如,医生办公室或其他临床环境,存储和使用全部或很多这种信息。在另一端,在一个位置处获得样品,在不同的位置处处理它并可选地测序,在一个或多个不同的位置处比对读数和进行调用,以及在又一个位置处(其可以是获得样品的位置)准备诊断、建议、和/或计划。

[0402] 在不同的实施方式中,借助于测序仪器来产生读数,然后发送到其中处理它们的远程站点,以产生非整倍体调用。在此远程位置,作为例子,使读数比对于参比序列以产生标记,其被计数并指定于感兴趣的染色体或片段。此外,在远程位置处,利用相关的归一化染色体或片段,将计数转化为剂量。再进一步,在远程位置处,剂量用来产生非整倍体调用。

[0403] 在不同的位置处可以采用的处理操作中有以下几种:

[0404] 样品收集

[0405] 在测序之前的样品处理

[0406] 测序

[0407] 分析序列数据和导出非整倍体调用

[0408] 诊断

[0409] 向患者或健康护理提供者报告诊断和/或调用

[0410] 制定用于进一步治疗、测试、和/或监测的计划

[0411] 执行上述计划

[0412] 咨询服务

[0413] 可以自动化任何一种或多种的这些操作(如在本文中别处描述的)。通常,将计算上进行序列数据的测序和分析以及导出非整倍体调用。可以手动或自动地进行其他操作。

[0414] 可以进行样品收集的位置的实例包括保健医生办公室、诊所、患者家(其中提供样品收集工具或试剂盒)、和移动医疗保健车。可以在样品处理之前进行测序的位置的实例包括保健医生办公室、诊所、患者家(其中提供样品处理仪器或试剂盒)、移动医疗保健车、和非整倍体分析提供者的设备。可以进行测序的位置的实例包括保健医生办公室、诊所、保健医生办公室、诊所、患者家里(其中提供样品测序仪器和/或试剂盒)、移动医疗保健车、和非整倍体分析提供者的设备。其中测序发生有位置可以拥有专用网络连接,用于发送电子格

式的序列数据(通常为读数)。这样的连接可以是有线或无线的并且已经和可以被配置来将数据发送到可以处理和/或在传输到处理站点之前聚集数据的站点。可以由保健组织如健康维护组织(HMO)来保留数据聚集器。

[0415] 可以在任何前述位置处或可替换地在另外远程站点处,其专用于计算和/或分析核酸序列数据的服务,来进行分析和/或导出操作。这样的位置包括例如,群集如通用服务器群、非整倍体分析服务业务的设备等。在一些实施方式中,租用或租借用来进行分析的计算仪器。计算资源可以是处理器的互联网可存取收集的一部分如通俗地被称为云的处理资源。在一些情况下,通过彼此关联或无关联的处理器的并行或大规模并行组来进行计算。可以利用分布式处理如群集计算、网格计算等来完成处理。在这样的实施方式中,计算资源的群集或网格共同形成超级虚拟计算机,该计算机组成自多个处理器或计算机,其一起用来进行本文描述的分析和/或导出。这些技术以及更常规的超级计算机可以用来处理序列数据(如本文所描述的)。每一种是一种形式的并行计算,其依赖于处理器或计算机。在网格计算的情况下,通过网络(专用网络、公共网络、或互联网)并根据常规的网络方法如以太网来连接这些处理器(通常整个计算机)。相比之下,超级计算机具有通过局部高速计算机总线加以连接的许多处理器。

[0416] 在某些实施方式中,在和分析操作相同的位置处产生诊断(例如,胎儿具有唐氏综合征或患者患有特定类型的癌症)。在其他实施方式中,在不同的位置处进行它。在一些实施例中,在取得样品的位置处报告诊断,虽然不必如此。可以产生或报告诊断和/或制定计划的位置的实例包括保健医生办公室、诊所、通过计算机可访问的互联网站点、和手持装置如手机、图形输入板、智能手机等,其具有与网络的有线或无线连接。进行咨询服务的位置的实例包括保健医生办公室、诊所、通过计算机可访问的互联网站点、手持装置等。

[0417] 在一些实施方式中,在第一位置处进行样品收集、样品处理、和测序操作以及在第二位置处进行分析和导出操作。然而,在一些情况下,在一个位置(例如,保健工作者办公室或诊所)处进行样品收集并在不同的位置(其可选地是发生分析和导出的同样位置)处进行样品处理和测序。

[0418] 在不同的实施方式中,可以由启动样品收集、样品处理和/或测序的用户或实体来触发一系列的以上所列操作。在一个或多个这些操作已经开始执行之后,其他操作会自然跟随。例如,测序操作可能引起自动收集读数并发送到处理仪器,其然后,经常自动地并且可能无需进一步的用户干预,进行非整倍体操作的序列分析和导出。在一些实施方式中,这种处理操作的结果然后被自动交付,可能重新格式化为诊断,到系统部件或实体,其处理信息并报告给健康专业人员和/或患者。如所解释的,还可以自动处理这样的信息以产生治疗、测试、和/或监测计划,可能连同咨询信息。因此,启动前期操作可以触发端到端的序列,其中健康专业人员、患者或其他相关方拥有诊断、计划、咨询服务和/或可用于作用于身体状况的其他信息。即使整个系统的部分被物理分离并且可能远离例如样品和序列仪器的位置,这也可以被完成。

[0419] 图5示出用于从测试样品来产生调用或诊断的分散系统的一种实施方式。样品收集位置01用于获得来自患者如妊娠雌性或假定的癌症患者的测试样品。然后将样品提供到处理和测序位置03,此处可以处理和测序测试样品(如上文描述的)。位置03包括用于处理样品的仪器以及用于测序经处理的样品的仪器。测序的结果,如在本文中别处描述的,是读

数的收集，其通常是以电子格式加以提供并提供到网络如互联网，在图5中其是由参考数字05来表示。

[0420] 序列数据被提供到远程位置07，此处进行分析和调用产生。此位置可以包括一个或多个鲁棒计算装置如计算机或处理器。在位置07处的计算资源已完成它们的分析并产生来自接收到的序列信息的调用之后，调用被中继回到网络05。在一些实施方式中，在位置07处不仅产生调用而且还产生相关的诊断。然后将调用和/或诊断发送整个网络并回到样品收集位置01，如在图5中所示。如所解释的，这仅仅只是关于相关与产生调用或诊断的各种操作可以如何在各种位置之间划分的许多变化的一种。一个常见的变通例涉及在单个位置处提供样品收集以及处理和测序。另一种变化涉及在和分析和调用产生相同的位置处提供处理和测序。

[0421] 图6详细说明用于在不同的位置处进行各种操作的选项。在图6中描述的最精细意义上，在分开的位置处进行以下每个操作：样品收集、样品处理、测序、读数比对、调用、诊断、和报告和/或计划制定。

[0422] 在聚集这些操作的一些操作的一种实施方式中，在一个位置处进行样品处理和测序以及在分开的位置处进行读数比对，调用，和诊断。见由参考字符A确定的图6的部分。在另一种实施方式中，其是由图6中的字符B所确定，样品收集、样品处理、和测序均在同样位置处进行。在此实施方式中，读数比对和调用在是第二位置处进行。最后，诊断和报告和/或计划制定是在第三位置处进行。在由图6中的字符C描述的实施方式中，在第一位置处进行样品收集，在第二位置处一起进行样品处理、测序、读数比对、调用、和诊断，以及在第三位置处进行报告和/或计划制定。最后，在图6中标记为D的实施方式中，在第一位置处进行样品收集，在第二位置处进行样品处理、测序、读数比对、和调用，以及在第三位置处进行诊断和报告和/或计划管理。

[0423] 一种实施方式提供了用于在包含胎儿和母体核酸的母体测试样品中确定任何一种或多种不同的完全胎儿染色体非整倍体的存在或不存在的系统，上述系统包括用于接收核酸样品和提供来自样品的胎儿和母体核酸序列信息的测序仪；处理器；以及机器可读存储介质，其包含用于在所述处理器上执行的指令，上述指令包括：

[0424] (a)用于获得在样品中所述胎儿和母体核酸的序列信息的代码；

[0425] (b)这样的代码，其用于利用所述序列信息来针对感兴趣的任何一个或多个的染色体(选自染色体1-22、X、和Y)的每一个计算上鉴定来自胎儿和母体核酸的序列标签的数目，以及针对感兴趣的所述任何一个或多个染色体的每一个确定至少一个归一化染色体序列或归一化染色体片段序列的序列标签的数目；

[0426] (c)这样的代码，其用于利用针对感兴趣的所述任何一个或多个的染色体的每一个所确定的序列标签的所述数目和针对每个归一化染色体序列或归一化染色体片段序列所确定的序列标签的所述数目，来计算对于感兴趣的任何一个或多个染色体的每一个的单染色体剂量；以及

[0427] (d)这样的代码，其用于比较对于感兴趣的染色体的任何一个或多个的每一个的每个单染色体剂量与对于感兴趣的一个或多个染色体的每一个的相应的阈值，并从而确定在样品中任何一个或多个完全不同的胎儿染色体非整倍体的存在或不存在。

[0428] 在一些实施方式中，用于计算对于每个感兴趣的任何一个或多个染色体的单染色

体剂量的代码包括这样的代码，其用于计算感兴趣的染色体的所选一种的染色体剂量为针对所选的感兴趣的染色体确定的序列标签的数目和针对所选的感兴趣的染色体的相应的至少一个归一化染色体序列或归一化染色体片段序列所确定的序列标签的数目的比率。

[0429] 在一些实施方式中，上述系统进一步包括这样的代码，其用于重复计算感兴趣的任何一个或多个染色体的任何一个或多个片段的每个的任何剩余染色体片段的染色体剂量。

[0430] 在一些实施方式中，选自染色体1-22、X、和Y的感兴趣的一个或多个染色体包括选自染色体1-22、X、和Y的至少二十均被涵盖染色体，以及其中指令包括用于确定至少二十种不同的完全胎儿染色体非整倍体的存在或不存在的指令。

[0431] 在一些实施方式中，上述至少一种归一化染色体序列是一组染色体，其选自染色体1-22、X、和Y。在其他实施方式中，上述至少一种归一化染色体序列是选自染色体1-22、X、和Y的单染色体。

[0432] 另一种实施方式提供了用于在包含胎儿和母体核酸的母体测试样品中确定任何一种或多种不同的部分胎儿染色体非整倍体的存在或不存在的系统，上述系统包括：测序仪，其用于接收核酸样品并提供来自样品的胎儿和母体核酸序列信息；处理器；以及机器可读存储介质，其包含用于在所述处理器上执行的指令，上述指令包括：

[0433] (a)代码，用于获得关于在所述样品中所述胎儿和母体核酸的序列信息；

[0434] (b)这样的代码，其用于利用所述序列信息来针对感兴趣的任何一个或多个的染色体(选自染色体1-22、X、和Y)的任何一个或多个片段的每一个计算上鉴定来自胎儿和母体核酸的序列标签的数目，以及针对感兴趣的任何一个或多个染色体的所述任何一个或多个片段的每一个确定至少一个归一化片段序列的序列标签的数目；

[0435] (c)这样的代码，其利用针对感兴趣的任何一个或多个染色体的每一个所述任何一个或多个片段所确定的序列标签的所述数目和针对所述归一化片段序列所确定的序列标签的所述数目，来计算对于感兴趣的任何一个或多个染色体的每一个的所述任何一个或多个片段的单染色体片段剂量；以及

[0436] (d)这样的代码，其用于比较对于感兴趣的任何一个或多个染色体的每个所述任何一个或多个片段的每个所述单染色体剂量与对于感兴趣的任何一个或多个染色体的每个所述任何一个或多个染色体片段的相应的阈值，并从而确定在所述样品中一个或多个不同的部分胎儿染色体非整倍体的存在或不存在。

[0437] 在一些实施方式中，用于计算单染色体片段剂量的代码包括这样的代码，其用于计算染色体片段的所选一种的染色体片段剂量为针对所选的染色体片段确定的序列标签的数目与针对所选的染色体片段的相应的归一化片段序列确定的序列标签的数目的比率。

[0438] 在一些实施方式中，上述系统进一步包括这样的代码，其用于重复计算感兴趣的任何一个或多个染色体的任何一个或多个片段的任何剩余染色体片段的每一个的染色体片段剂量。

[0439] 在一些实施方式中，上述系统进一步包括(i)这样的代码，其用于针对来自不同母体受试者的测试样品重复(a)-(d)；以及(ii)这样的代码，其用于确定在每个所述样品中任何一种或多种不同的部分胎儿染色体非整倍体的存在或不存在。

[0440] 在本文中提供的任何系统的其他实施方式中，上述代码进一步包括这样的代码，

其用于自动记录胎儿染色体非整倍体的存在或不存在，如在针对提供母体测试样品的人受试者的患者医疗记录中确定的(d)，其中利用处理器来进行记录。

[0441] 在本文中提供的任何系统的一些实施方式中，测序仪被配置以进行下一代测序(NGS)。在一些实施方式中，测序仪被配置以进行大规模平行测序，其中利用合成测序并借助于可逆染料终止子。在其他实施方式中，测序仪被配置以进行连接测序。在其他实施方式中，测序仪被配置以进行单分子测序。

[0442] 实验

[0443] 实施例1

[0444] 原始和富集测序文库的制备和测序

[0445] a.测序文库的制备-简化方法(ABB)

[0446] 所有测序文库，即，原始和富集文库，制备自大约2ng纯化cfDNA，其提取自母体血浆。利用NEBNext™ DNA Sample Prep DNA Reagent Set 1(产品号E6000L;New England Biolabs,Ipswich,MA)的试剂进行文库制备，对于Illumina®如下。由于无细胞血浆DNA在自然界中被片段化，所以没有通过雾化或超声处理对血浆DNA样品进行进一步片段化。根据NEBNext®End Repair Module，在20℃下，通过在1.5ml微离心管中，并借助于5μl 10X磷酸化缓冲液、2μl脱氧核苷酸溶液混合物(每种dNTP为10mM)、1μl的1:5稀度的DNA聚合酶I、1μl的T4 DNA聚合酶和1μl T4多核苷酸激酶，其提供在NEBNext™ DNA Sample Prep DNA Reagent Set1中，温育cfDNA 15分钟，将包含在40μl中的大约2ng纯化cfDNA片段的突出端转化为磷酸化平端。然后通过在75℃下温育反应混合物5分钟来热灭活上述酶。将混合物冷却至4℃，以及利用10μl的含有Klenow片段(3'至5'外减)(NEBNext™ DNA Sample Prep DNA Reagent Set 1)的dA拖尾主要混合物，并在37℃下温育15分钟，来完成平端DNA的dA拖尾。其后，通过在75℃下温育反应混合物5分钟来热灭活Klenow片段。在Klenow片段的灭活之后，1μl的1:5稀度的Illumina Genomic Adaptor Oligo Mix(产品号1000521;Illumina Inc.,Hayward,CA)用来将Illumina衔接子(非索引Y衔接子)连接于dA拖尾DNA，其中利用4μl的T4 DNA连接酶，其提供在NEBNext™ DNA Sample Prep DNA Reagent Set 1中，并通过在25℃下温育反应混合物15分钟。将混合物冷却至4℃，以及衔接子连接的cfDNA纯化自未连接的衔接子、衔接子二聚体、和其他试剂，其中利用在Agencourt AMPure XP PCR纯化系统(产品号A63881;Beckman Coulter Genomics,Danvers,MA)中提供的磁珠。进行PCR的十八个循环以选择性地富集衔接子连接的cfDNA(25μl)，其中使用Phusion ®High-Fidelity Master Mix(25μl;Finnzymes,Woburn,MA)和Illumina的PCR引物(各自0.5μM)，其互补于衔接子(产品号1000537和1000537)。根据制造商的说明，使用Illumina Genomic PCR引物(产品号100537和1000538)和在NEBNext™ DNA Sample Prep DNA Reagent Set 1中提供的Phusion HF PCR Master Mix，使衔接子连接的DNA经受PCR(98℃下30秒，98℃下10秒、65℃下30秒、和72℃下30秒的18个循环，在72℃下最后延伸5分钟，然后保持在4℃下)。根据在www.beckmangenomics.com/products/AMPureXPProtocol_000387v001.Ddf.处可获得的制造商的说明，使用Agencourt AMPure XP PCR纯化系统(Agencourt Bioscience Corporation,Beverly,MA)来纯化扩增产物。在40μl的Qiagen EB Buffer中洗脱纯化的扩增产物，然后利用用于2100 Bioanalyzer(Agilent technologies Inc.,Santa Clara,CA)的Agilent DNA 1000 Kit来分析扩增文库的浓度和尺寸分布。

[0447] b.测序文库的制备-全长方法

[0448] 这里描述的全长方法基本上是由Illumina提供的标准方法，并且仅在扩增文库的纯化方面不同于Illumina方法。Illumina方法指示，利用凝胶电泳来纯化扩增文库，而本文描述的方法则是磁珠用于同样的纯化步骤。基本上根据制造商的说明，利用用于Illumina®的NEBNext™ DNA Sample Prep DNA Reagent Set 1(产品号E6000L;New England Biolabs,Ipswich,MA)，大约2ng的提取自母体血浆的纯化cfDNA用来制备原始测序文库。除衔接子连接的产物的最后纯化(其是利用Agencourt磁珠和试剂而不是纯化柱来进行)之外的所有步骤是根据上述方法并伴随NEBNext™试剂(其用于基因组DNA文库，其是利用Illumina®GAII加以测序，的样品制备)来进行。NEBNext™方法基本上遵循由Illumina提供的方法，其是在grcf.jhml.edu/hts/protocols/11257047_ChIP_Sample_Prep.pdf.处可获得的。

[0449] 根据NEBNext®End Repair Module，在热循环仪中的200μl微离心管中，在20℃下。通过用5μl 10X磷酸化缓冲液、2μl脱氧核苷酸溶液混合物(每种dNTP为10mM)、1μl的1:5稀度的DNA聚合酶I、1μl T4 DNA聚合酶和1μl T4多核苷酸激酶，其提供在NEBNext™ DNA Sample Prep DNA Reagent Set 1中，温育40μl的cfDNA 30分钟，将包含在40μl中的大约2ng纯化cfDNA片段的突出端转化为磷酸化平端。将样品冷却至4℃，然后利用在QIAQuick PCR纯化试剂盒(QIAGEN Inc.,Valencia,CA)中提供的QIAQuick柱加以纯化如下。将50μl反应混合物转移到1.5ml微离心管，然后添加250μl的Qiagen Buffer PB。将得到的300μl转移到QIAquick柱，其在微型离心机中在13,000RPM下被离心1分钟。用750μl的Qiagen Buffer PE洗涤上述柱，并再次离心。通过在13,000RPM下另外离心5分钟来除去残余乙醇。通过离心，在39μl的Qiagen Buffer EB中洗脱DNA。利用16μl的含有Klenow片段(3’至5’外减)(NEBNext™ DNA Sample Prep DNA Reagent Set 1)的dA拖尾主要混合物，并根据制造商的NEBNext® dA-Tailing Module在37℃下温育30分钟，来完成34μl平端DNA的dA拖尾。将样品冷却至4℃，然后利用在MinElute PCR纯化试剂盒(QIAGEN Inc.,Valencia,CA)中提供的柱加以纯化如下。将50μl反应混合物转移到1.5ml微离心管，并添加250μl的Qiagen Buffer PB。将300μl转移到MinElute柱，在微型离心机中并在13,000RPM下其被离心1分钟。用750μl Qiagen Buffer PE洗涤上述柱，然后再次离心。通过在下13,000RPM下另外离心5分钟来除去残余乙醇。在15μl Qiagen Buffer EB中通过离心来洗脱DNA。根据NEBNext®Quick Ligation Module，在25℃下，用1μl的1:5稀度的Illumina Genomic Adapter Oligo Mix(产品号1000521)、15μl的2X Quick Ligation Reaction Buffer、和4μl Quick T4 DNA连接酶，来温育10微升DNA洗脱液15分钟。将样品冷却至4℃，然后利用MinElute柱加以纯化如下。将一百五十微升的Qiagen Buffer PE加入30μl反应混合物，并将整个容积转移到MinElute柱，在微型离心机中并在13,000RPM下其被离心1分钟。用750μl Qiagen Buffer PE洗涤上述柱，并再次离心。通过在13,000RPM下另外离心5分钟来除去残余乙醇。在28μl Qiagen Buffer EB中通过离心来洗脱DNA。根据制造商的说明，利用Illumina Genomic PCR引物(产品号100537和1000538)和Phusion HF PCR Master Mix，其提供在NEBNext™ DNA Sample Prep DNA Reagent Set 1中，对二十三分微升的衔接子连接的DNA洗脱液进行18个循环的PCR(98℃下30秒;98℃下10秒、65℃下30秒、和72℃下30秒的18次循环;在72℃下最后延伸5分钟，并保持在4℃下)。根据在www.beckmangenomics.com/products/

AMPureXPProtocol_000387v001.pdf处可获得的制造商的说明，利用Agencourt AMPure XP PCR纯化系统(Agencourt Bioscience Corporation，Beverly，MA)来纯化扩增产物。Agencourt AMPure XP PCR纯化系统除去未并入的dNTP、引物、引物二聚体、盐和其他污染物，并回收大于100bp的扩增子。使纯化扩增产物洗脱自在在40μl的Qiagen EB Buffer中的Agencourt珠，然后利用用于2100Bioanalyzer(Agilent technologies Inc.，Santa Clara，CA)的Agilent DNA 1000Kit来分析文库的尺寸分布。

[0450] c.根据精简(a)和全长(b)方法制备的测序文库的分析

[0451] 通过Bioanalyzer产生的电泳图示于图7A和7B。图7A示出制备自cfDNA，其利用在(a)中描述的全长方法纯化自血浆样品M24228，的文库DNA的电泳图，以及图7B示出制备自cfDNA，其利用在(b)中描述的全长方法纯化自血浆样品M24228，的文库DNA的电泳图。在两个图中，峰1和4分别表示15bp下标志物和1,500上标志物；在峰上方的数字指出文库片段的迁移时间；以及水平线指出整合的设置阈值。在图7A中的电泳图示出187bp的片段的副峰，以及263bp的片段的主峰，而在图7B中的电泳图则仅示出在265bp处的一个峰。峰面积的整合导致对于在图7A中的187bp峰的DNA的计算浓度为0.40ng/μl，在图7A中的263bp峰的DNA的浓度为7.34ng/μl，以及对于在图7B中的265bp峰的DNA的浓度为14.72ng/μl。连接于cfDNA的Illumina衔接子已知是92bp，当从265bp减去时，其指出，cfDNA的峰大小是173bp。可能的是，在187bp处的副峰表示端-端连接的两个引物的片段。当使用精简方法时，从最终文库产物消除线性两引物片段。精简方法还消除小于187bp的其他较小片段。在本实施例中，纯化衔接子连接的cfDNA的浓度是利用全长方法产生的衔接子连接的cfDNA的浓度的两倍。已经注意到，衔接子连接的cfDNA片段的浓度总是大于利用全长方法获得的浓度(数据未示出)。

[0452] 因此，利用精简方法来制备测序文库的一个优点在于，获得的文库始终包含在262-267bp范围内的仅一个主峰，而利用全长方法来制备的文库的质量则会变化，如由不同于表示cfDNA峰的峰的数目和移动性所反映的。非cfDNA产物会占据在流动池上的空间，因而降低测序反应的群集扩增和随后成像的质量，其构成非整倍体状态的整体指配的基础。精简方法表明不影响文库的测序。

[0453] 利用精简方法来制备测序文库的另一个优点在于，平端、d-A拖尾、和衔接子连接的三酶促步骤需要不到一个小时来完成以支持快速非整倍体诊断服务的验证和实施。利用精简方法来制备测序文库的另一个优点在于，平端、d-A拖尾、和衔接子连接的三酶促步骤需要不到一个小时来完成以支持快速非整倍体诊断服务的验证和实施。

[0454] 另一个优点在于，在同样的反应管中进行平端、d-A拖尾、和衔接子连接的三酶促步骤，因而避免多个样品传输，其将潜在地导致材料的损失，以及更重要的是，导致可能的样品混合和样品污染。

[0455] 实施例2

[0456] 在双胞胎妊娠中准确的非整倍性检测

[0457] 引言

[0458] 已经表明，利用全基因组大规模平行测序进行的总无细胞DNA(cfDNA)的非侵入性产前测试(NIPT)是检测胎儿染色体非整倍性的非常准确和可靠的方法。参见，Bianchi DW，Platt LD，Goldberg JD，等人Genome-Wide fetal aneuploidy detection by maternal

plasma DNA sequencing.Obstet Gynecol 2012;119:890-901;Fan HC,Blumenfeld YJ,Chitkara U,Hudgins L,Quake SR.Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood.Proc Natl Acad Sci U S A2008;105:16266-71;Sehnert AJ,Rhecs B,Comstock D,等人Optimal detection of fetal chromosomal abnormalities by massively parallel DNA sequencing of cell-free fetal DNA from maternal blood.Clin Chem2011;57:1042-9。上述即时测试检测来自单母体血液样品的三体性21、18、13和性染色体非整倍性。上述即时测试当前指示用于具有10+周并具有胎儿非整倍性的高风险的单胎妊娠的孕妇。最近,American College of Obstetricians and Gynecologists(ACOG)、International Society for Prenatal Diagnosis(ISPD)、American College of Medical Genetics and Genomics(ACMG)和National Society of Genetic Counselors(NSGC)已建议考虑对具有胎儿非整倍性的高风险的妇女使用NIPT。

[0459] 在美国,双胞胎占大约30分之一的活产并且双胞胎出生的比率是在增加(National Center for Health Statistics Data Brief,No.80,2012年1月)。随着女性年龄的增长,她们更可能每个月经周期释放一个以上的卵,因此,超过30岁的妇女占双胞胎妊娠的增加的约1/3。辅助生殖技术,其中在体外受精期间,经常传输一个以上的胎儿,占双胞胎妊娠的大部分的其余增长。

[0460] 初步证据提示,当相比于单胎妊娠时,在双胞胎妊娠中,在母体循环中存在的胎儿DNA的量增加大约35%,但上述研究没有看来源于每个胎儿的cfDNA的量。Canick JA,Kloza EM,Lambert-Messerlian GM,等人DNA sequencing of maternal plasma to identifyDown sydrome and other trosomies in multiple gestations.Prenat Diagn 2012;32:730-4。研究人员已经证明,虽然在双胞胎妊娠中存在循环胎儿DNA的量的总体增加,但对于每个胎儿的cfDNA的量则降低。Srinivasann A,Bianchi D,Liao W,Sehnert A,Rava R.52:Maternal plasma DNA sequencing:effects of multiple gestation on aneuploidy detection and the relative cell-fere fetal DNA(cffDNA)per fetus.American journal of obstetrics and gynecology 2013;208:S31.Srinivasan A,Bianchi DW,Huang H,Sehnert AJ,Rava RP.Noninvasive detection of fetal subchromosome abnormalitocs via deep sequencing of maternal plasma.American journal of human genetics 2013;92:167-76。因此,需要敏感的方法以确保在双胞胎妊娠中非整倍体的正确分类。

[0461] 最大化NIPT准确分类非整倍体样品的能力的因素是在分析中使用的测序读数的数目的增加,以致统计噪声被最小化以及归一化染色体信号以致运行间变异性的能力被降低。最近,申请人已开发了一种改善的、自动化样品制备工作流程,其增加每个样品可用读数的数目,以及一种改善的分析方法,其会增加非整倍体染色体的特定信号。这些增强特性会改善分类非整倍体受影响的样品的总体准确性。

[0462] 此实施例描述了改善的分类算法对迄今使用的最大的双校验群组的应用。我们证明了,改善的SAFeR(用于胎儿结果的选择算法)算法允许在双胞胎样品中准确的非整倍体检测,其中上述双胞胎样品已知具有减少量的无细胞DNA/胎儿。

[0463] 方法

[0464] 作为两项独立临床研究的一部分，收集样品，其涉及高风险和平均风险母体群体。MatErnal BLood IS Source与准确诊断胎儿非整倍性研究(MELISSA;NCT01122524)被设计用来检测在高危妊娠中的全染色体非整倍体。Bianchi DW,Platt LD,Goldberg JD,等人Genome-wide fetal aneuploidy derection by maternal plasma DNA sequencing.Obstet Gynecol2012;119:890-901。非整倍体风险评价试验的比较(CARE;NCT01663350)被设计用来表明，对于在平均风险母体群体中的三体性21和三体性18，相比于常规产前血清筛查方法，即时测试的优越的特异性(提交公布)。数据集的细节示于表3。通过来自产前侵入性程序的核型或通过新生儿体检来确定临床结果。

[0465] 表3：双胞胎样品的核型和即时分类。利用即时产前测试，针对染色体21、18和13的非整倍体以及针对Y染色体的存在，分析了来自118例双胞胎妊娠的母体样品。将即时数据比较与通过核型分析或新生儿体检所获得的临床结果。

[0466]

研究的数量	胎儿1	胎儿2	即时非整倍性分类	即时Y染色性分类
24	46,XX	46,XX	未受影响的	没有检测到
48	46,XX	46,XY	未受影响的	Y检测的
42	46,XY	46,XY	未受影响的	Y检测的
2	47,XY,+21	46,XY	T21受影响的	Y检测的
1	Mos 47,XY,+21[7]/46,XY[11]	46,XX	T21受影响的	Y检测的
1	47,XY,+18	47,XY,+18	T18受影响的	Y检测的

[0467] 无细胞DNA提取自冷冻血浆样品并用如先前所描述的HiSeq2000测序仪加以测序。Sehnert AJ,Rhees B,Comstock D,等人Optimal detection of fetal chromosomal abnomalities by massively parallel DNA sequencing of cell-free fetal DNA from maternal blood.Clin Chem 2011;57:1042-9.。将大规模并行测序(MPS)序列标签映射到人类基因组参比版本hg19并利用改善的分析工作流程，其最大化信噪比以及改善检测的总灵敏度，针对染色体21、18、13、X和Y，来计算归一化染色体值(NCV)。算法部分包括改善的基因组过滤、通过分子生物学步骤引入的系统性偏差的除去以及改善的归一化和分类方法。进行测序的实验室人员对临床结果是不知情的。

[0468] 结果

[0469] 在本研究中研究了来自具有临床上定义的结果的118例双胞胎妊娠的母体血浆样品(表3)。针对研究中的所有样品，产生了针对染色体21、18和13的非整倍体分类，以及正确确定了来自具有一个或多个非整倍体胎儿的妊娠的四个样品(图8)。这些样品的两个是来自双绒膜双胞胎对，其各自具有一个T21受影响的雄性胎儿和一个非受影响的雄性胎儿(47,XY+21/46,XY)；一个是具有47,XY+18核型的单绒膜双胞胎样品；以及一个样品是双绒膜双胞胎，其中一个双胞胎具有嵌合核型47,XY+T21[7]/46,XY[11]。在本研究中没有临床上定义的未受影响的样品((N＝114)被分类为受影响的(对于非整倍体)。

[0470] 可以通过在cfDNA中Y染色体的存在来确定胎儿的性别。本文披露的测试能够在具有至少一个雄性胎儿的所有样品中阳性鉴定Y染色体的存在(图8)。此外，上述测试还正确确定在具有两个雌性胎儿的样品中Y染色体的不存在。

[0471] 结论

[0472] 目前的研究表明一种改善的分析方法,其能够进行双胞胎样品的最敏感的常染色体非整倍体测试。上述增强的分析方法利用了基因组过滤的改善、系统噪声减小和改善的分类方法。对一组118个双胞胎样品,证实了改善的分析工作流程的效用;在MPS的任何确认中,数量最多的样品用来检测在双胞胎中的常染色体非整倍体以及Y染色体的存在(图9)。图9示出在NIPT研究中分析的双胞胎样品。许多双胞胎样品用于各种研究来评价市售NIPT测试的性能。Canick JA,Kloza EM,Lambert-Messerlian GM,等人DNA sequncing of maternal plasma to identify Domn syndrome and other trisomies in multiple gestations.Prenat Diagn 2012;32:730-4.Lau TK,Jiang F,Chan MK,Zhang H,Lo PSS,Wang W.Non-invasive prenatal screening of fetal Domn syndrome by maternal plasma DNA sequencing in twin pregnancies.Journal of Maternal-Fetal and Neonatal Medicine 2013;26:434-7。改善的分析方法表明可准确地进行,其中通过在群组中,包括针对三体性21为嵌合型的受影响的胎儿,正确地检测所有三体性21和三体性18样品的存在,而没有产生任何假阳性结果。另外,改善的分析方法正确地检测在具有至少一个雄性胎儿的所有双胞胎妊娠中Y染色体的存在,以及并不检测在具有两个雌性胎儿的任何双胞胎妊娠中的Y染色体。

[0473] 敏感方法的一个特性是最小化系统噪声和增加总体信噪比的能力。通过产生比任何其他市售NIPT测定(大约28M测序读数/样品)更多的测序读数/样品以及通过改善分析方法来更好地处理伴随复杂的DNA样品的生化操作的系统噪声,目前的研究完成此任务。改善的分析工作流程最终减小归一化染色体计数分布的宽度,从而允许未受影响的和受影响的群体的更好的分离以及借助于少量的胎儿DNA来准确地确定非整倍体受影响的胎儿的改善的能力。

[0474] 具有非常准确和灵敏的方法来检测在双胞胎妊娠中的非整倍体的能力是重要的,这是因为,虽然在双胞胎妊娠中无细胞胎儿DNA的总量增加,但可归因于每个胎儿的量则减少。因此,可以A)忽视此发现和测试样品,好像它们相当于单胎妊娠,并增加假阴性结果的似然,B)拒绝增加数目的样品,由于不足的DNA,或C)建立更加敏感的方法(表4)。

[0475] 表2:利用市售NIPT测试来处理双胞胎妊娠的策略

[0476]

	选项		结果
A	测试双胞胎妊娠,好像存在的cfDNA等同于单胎妊娠。	⇨	假阴性的增加的似然。
B	利用目前的方法来测试双胞胎妊娠。	⇨	拒绝样品,由于不足的DNA
C	使用对于个体cfDN浓度是更加敏感的改善的方法	⇨	对于双胞胎和低水平单胎的更准确的测试并具有更少的假阴性。

[0477] 对SAFeR™算法的分析改善延伸超过了使得能够在双胞胎妊娠中进行精确的非整倍体分类。未受影响的与受影响的群体的改善的分离还降低了被分类为疑似的非整倍体的样品的总频率。另外,改善的分析工作流程可以应用于单胎妊娠,并在非整倍体检测和性别分类方面具有类似的改善。

[0478] 总之,目前的研究描述了改善的分析方法,对于含有少量的胎儿DNA的样品,其导致非整倍体未受影响的与受影响的样品的更好的分离以及更准确的常染色体非整倍体分

类。通过结合这些改善，产前测试能力已被扩展来测试双胞胎妊娠。

[0479] 可以以其他特定形式来具体实施本公开内容而不偏离它的精神或基本特征。所描述的实施方式在各方面仅被看作说明性的而不是限制性的。因此，公开内容的范围是由所附权利要求而不是由上述描述来限定。在权利要求的含义和等效范围内的所有变化均被涵盖在它们的范围内。

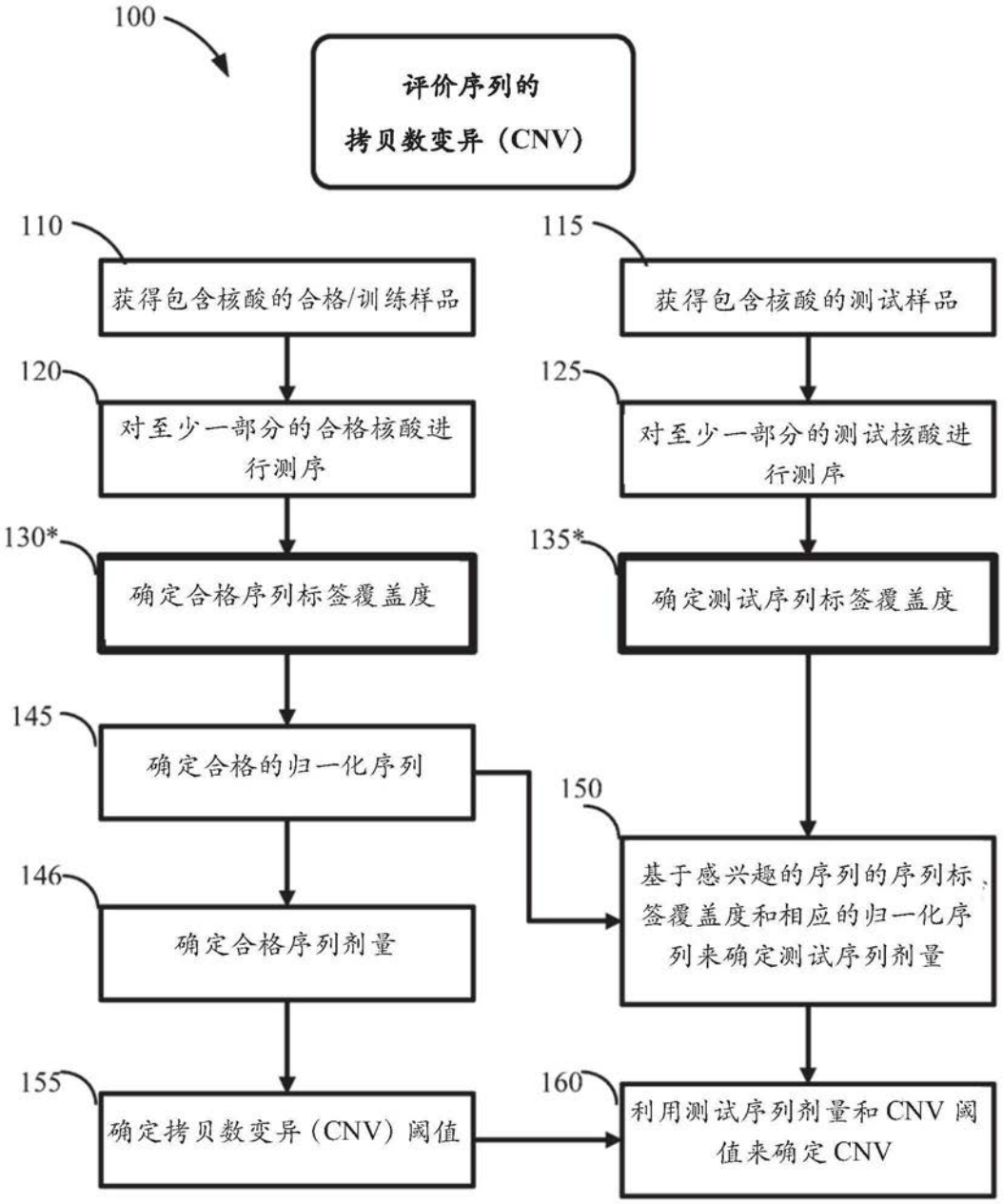

100
评价序列的
拷贝数变异（CNV）
110
获得包含核酸的合格/训练样品
115
获得包含核酸的测试样品
120
对至少一部分的合格核酸进行测序
125
对至少一部分的测试核酸进行测序
130*
确定合格序列标签覆盖度
135*
确定测试序列标签覆盖度
145
确定合格的归一化序列
150
基于感兴趣的序列的序列标签覆盖度和相应的归一化序列来确定测试序列剂量
146
确定合格序列剂量
155
确定拷贝数变异（CNV）阈值
160
利用测试序列剂量和 CNV 阈值来确定 CNV

图1

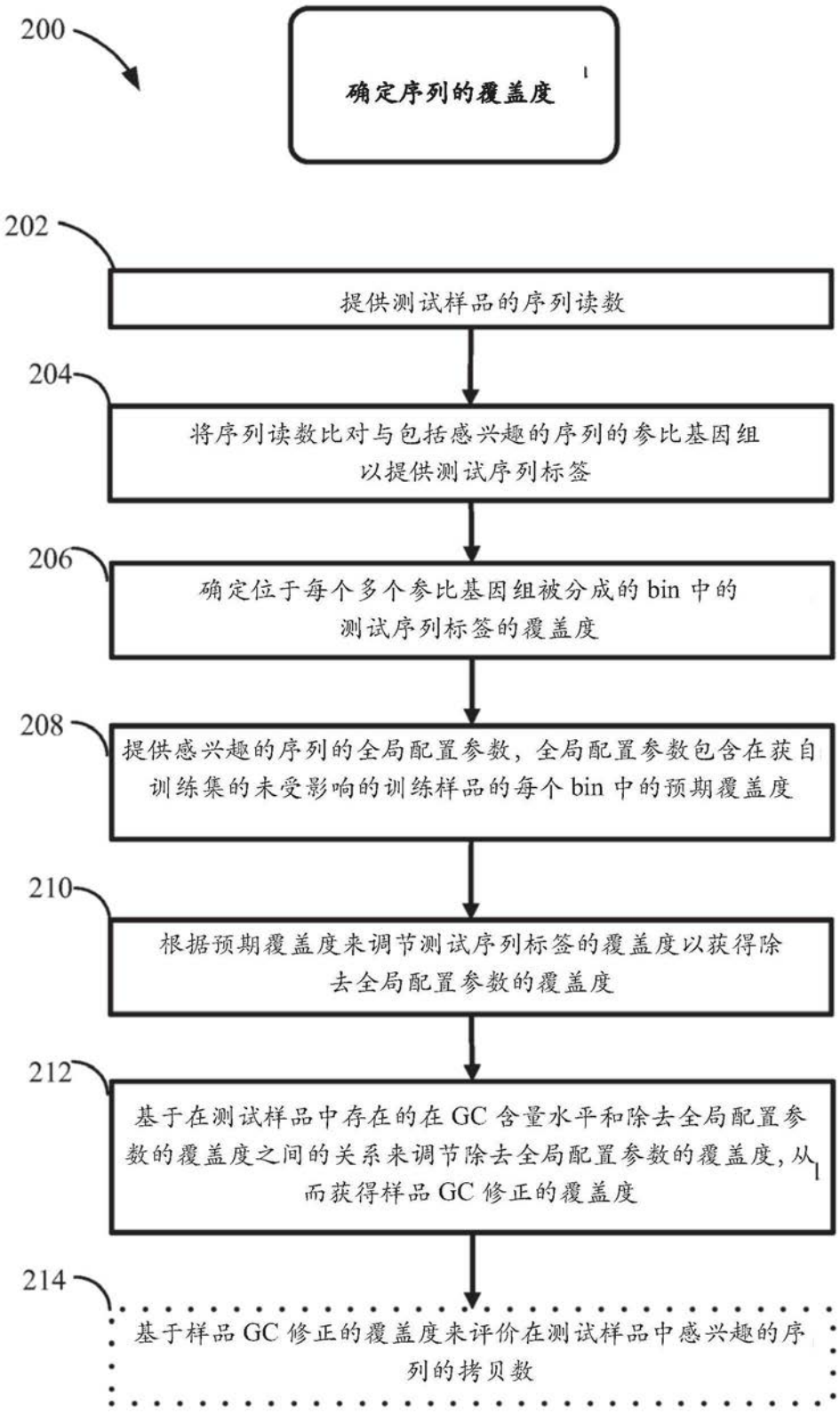

200
确定序列的覆盖度
202
提供测试样品的序列读数
204
将序列读数比对与包括感兴趣的序列的参比基因组以提供测试序列标签
206
确定位于每个多个参比基因组被分成的 bin 中的测试序列标签的覆盖度
208
提供感兴趣的序列的全局配置参数，全局配置参数包含在获自训练集的未受影响的训练样品的每个 bin 中的预期覆盖度
210
根据预期覆盖度来调节测试序列标签的覆盖度以获得除去全局配置参数的覆盖度
212
基于在测试样品中存在的在 GC 含量水平和除去全局配置参数的覆盖度之间的关系来调节除去全局配置参数的覆盖度，从而获得样品 GC 修正的覆盖度
214
基于样品 GC 修正的覆盖度来评价在测试样品中感兴趣的序列的拷贝数

图2

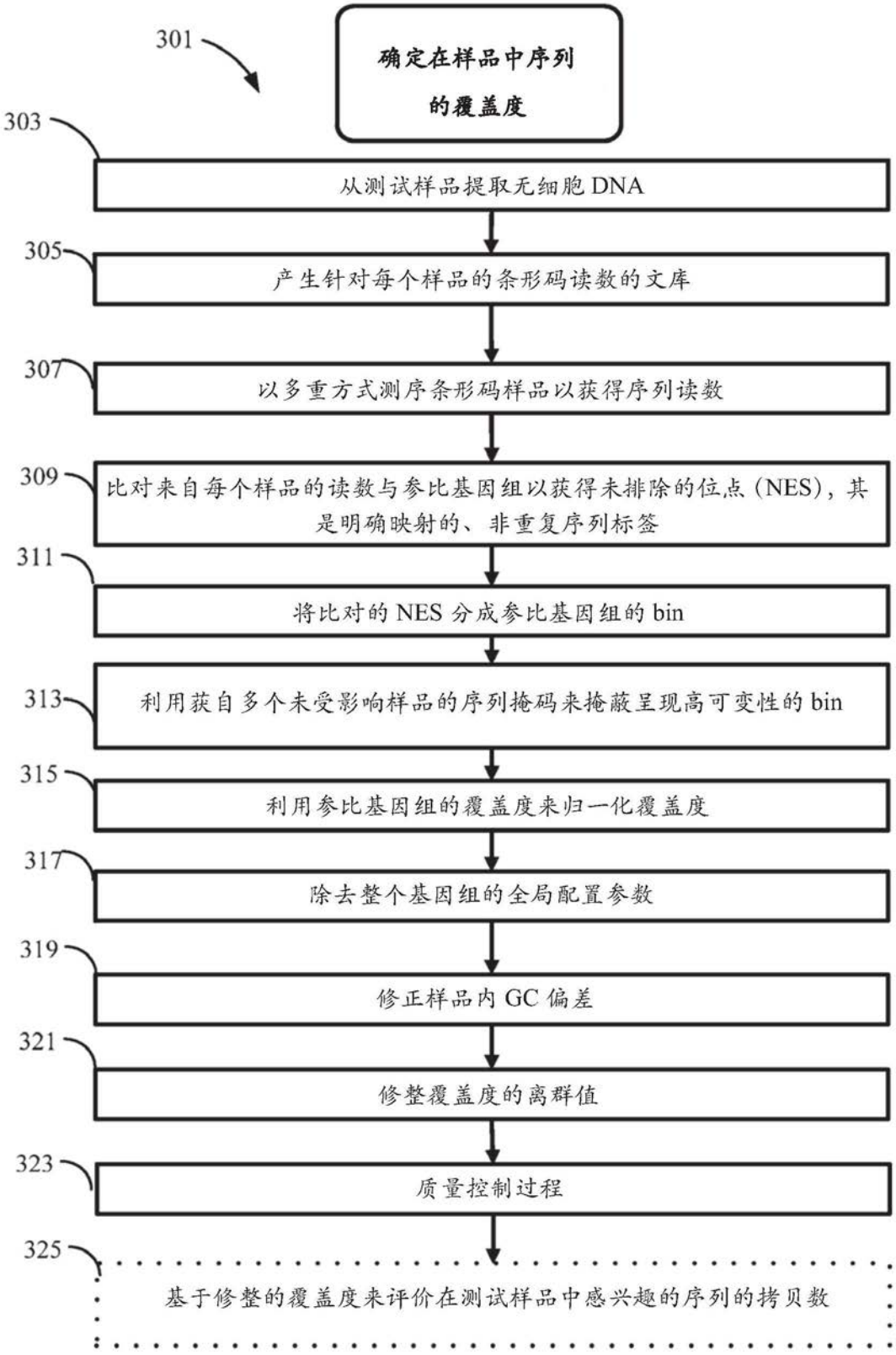

301
确定在样品中序列的覆盖度
303
从测试样品提取无细胞 DNA
305
产生针对每个样品的条形码读数的文库
307
以多重方式测序条形码样品以获得序列读数
309
比对来自每个样品的读数与参比基因组以获得未排除的位点（NES），其是明确映射的、非重复序列标签
311
将比对的 NES 分成参比基因组的 bin
313
利用获自多个未受影响样品的序列掩码来掩蔽呈现高可变性的 bin
315
利用参比基因组的覆盖度来归一化覆盖度
317
除去整个基因组的全局配置参数
319
修正样品内 GC 偏差
321
修整覆盖度的离群值
323
质量控制过程
325
基于修整的覆盖度来评价在测试样品中感兴趣的序列的拷贝数

图3A

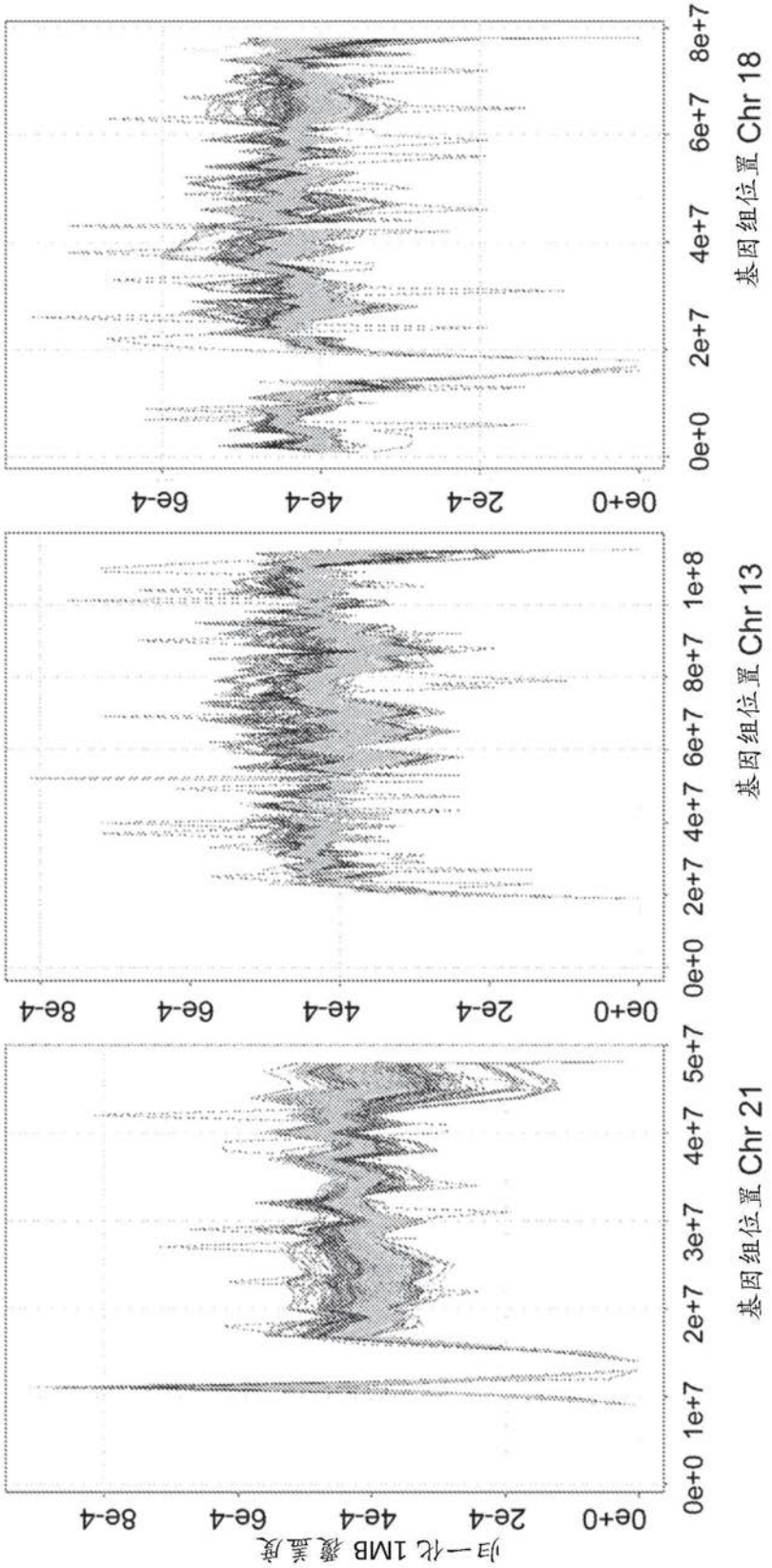

归一化 1MB 覆盖度
0e+0
2e-4
4e-4
6e-4
8e-4
0e+0
1e+7
2e+7
3e+7
4e+7
5e+7
基因组位置 Chr 21
0e+0
2e+7
4e+7
6e+7
8e+7
1e+8
基因组位置 Chr 13
0e+0
2e+7
4e+7
6e+7
8e+7
基因组位置 Chr 18

图3B

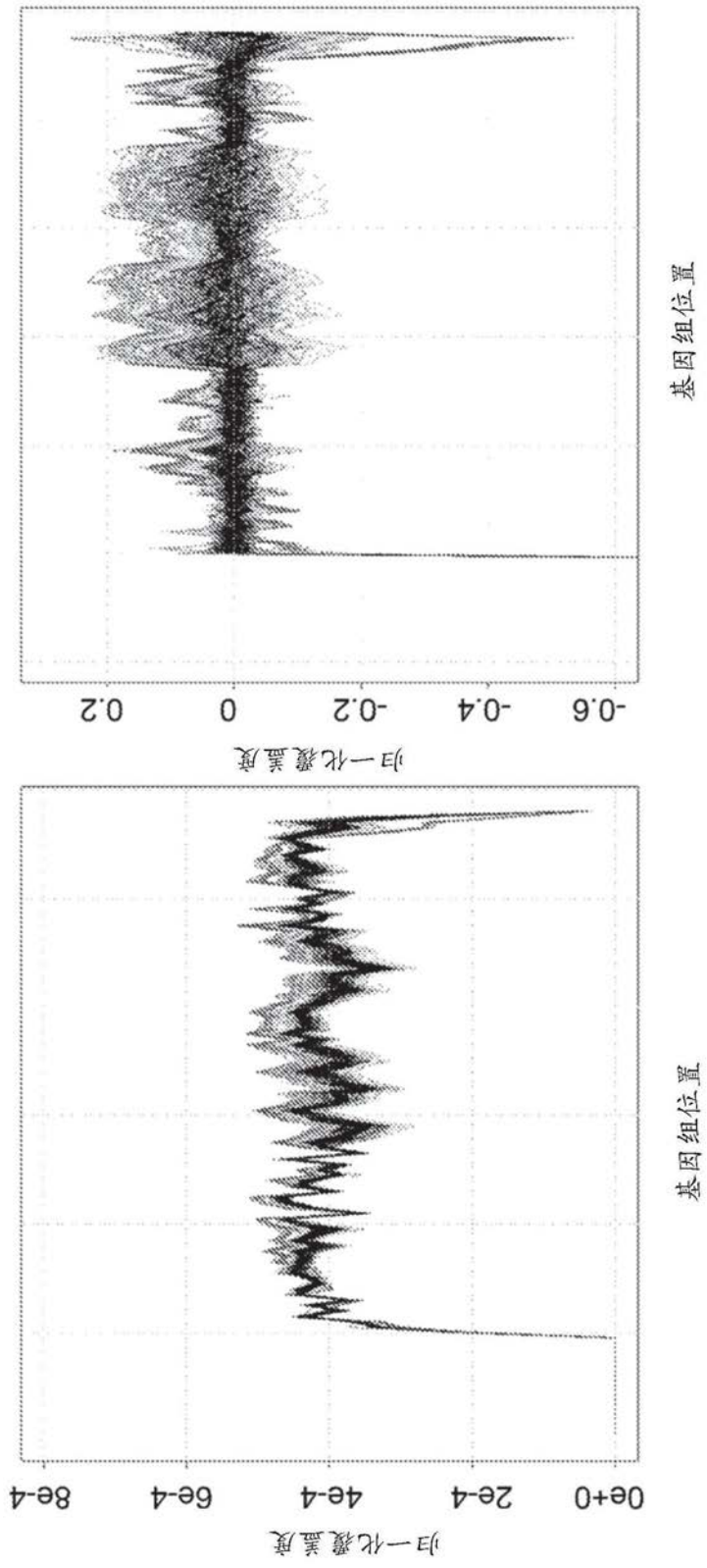


图3C

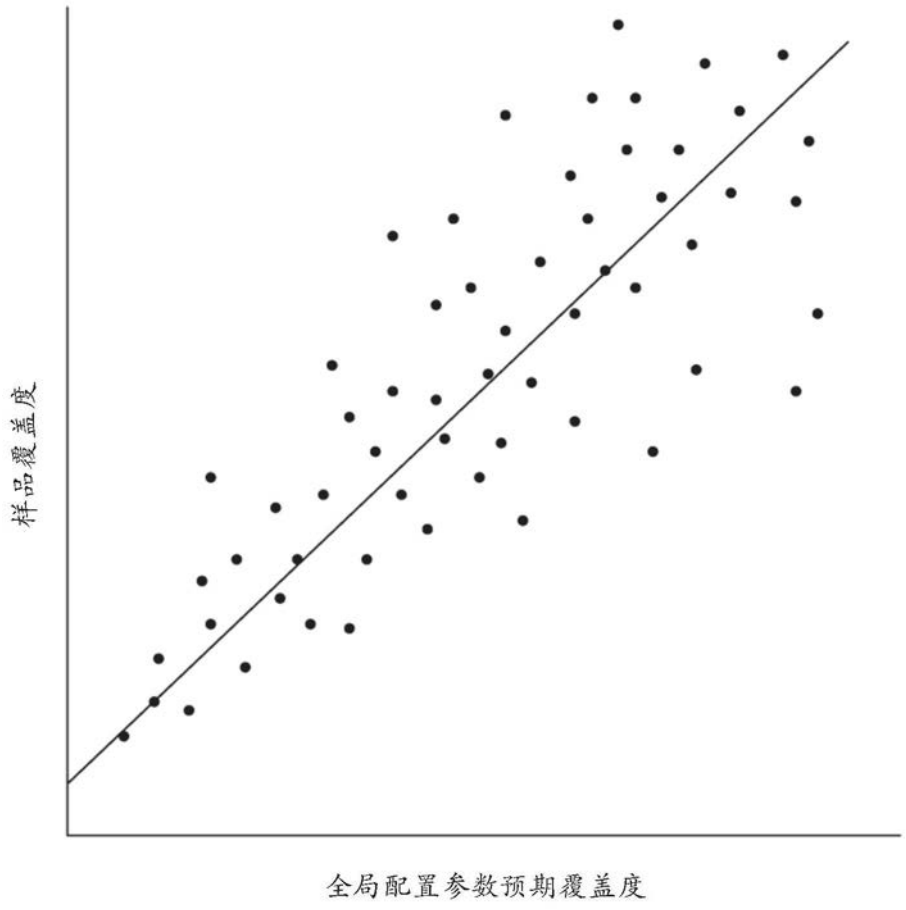


图3D

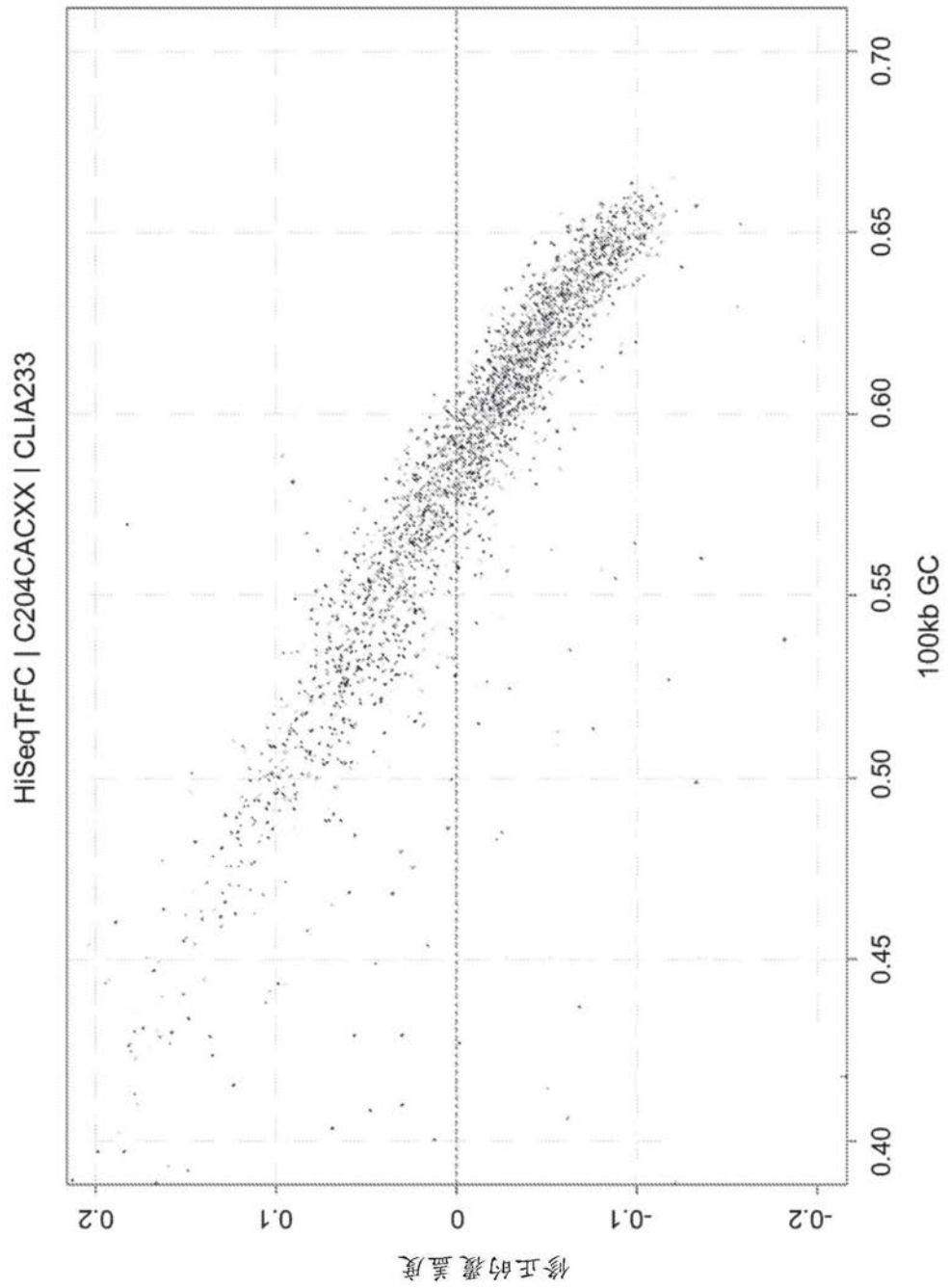

HiSeqTrFC | C204CACXX | CLIA233
修正的覆盖度
0.2
0.1
0
-0.1
-0.2
0.70
0.65
0.60
0.55
0.50
0.45
0.40
100kb GC

图3E

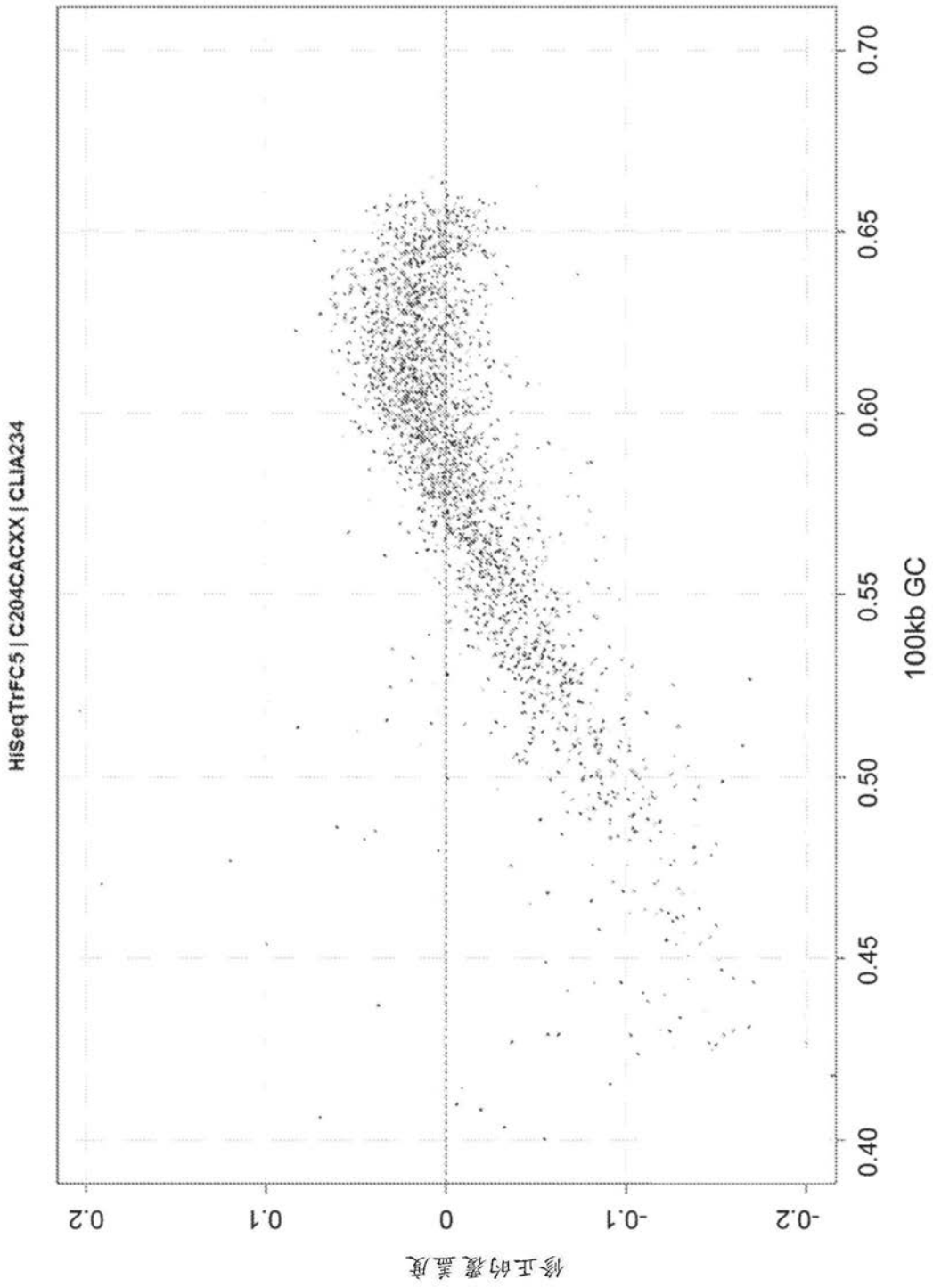

HiSeqTrFC5 | C204CACXX | CLIA234
修正的覆盖度
0.2
0.1
0
-0.1
-0.2
0.70
0.65
0.60
0.55
0.50
0.45
0.40
100kb GC

图3F

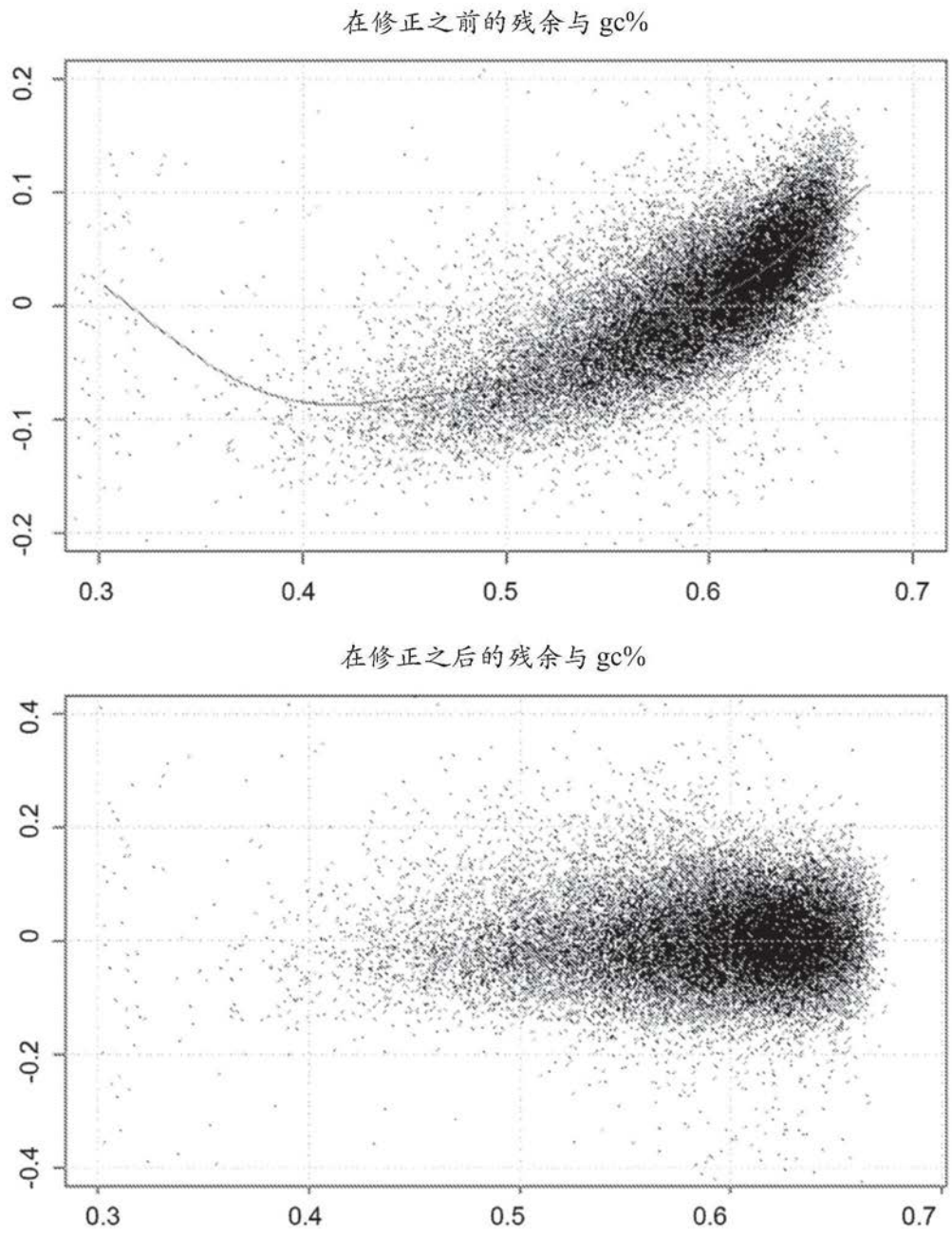

在修正之前的残余与 gc%
0.2
0.1
0
-0.1
-0.2
0.3
0.4
0.5
0.6
0.7
在修正之后的残余与 gc%
0.4
0.2
0
-0.2
-0.4
0.3
0.4
0.5
0.6
0.7

图3G

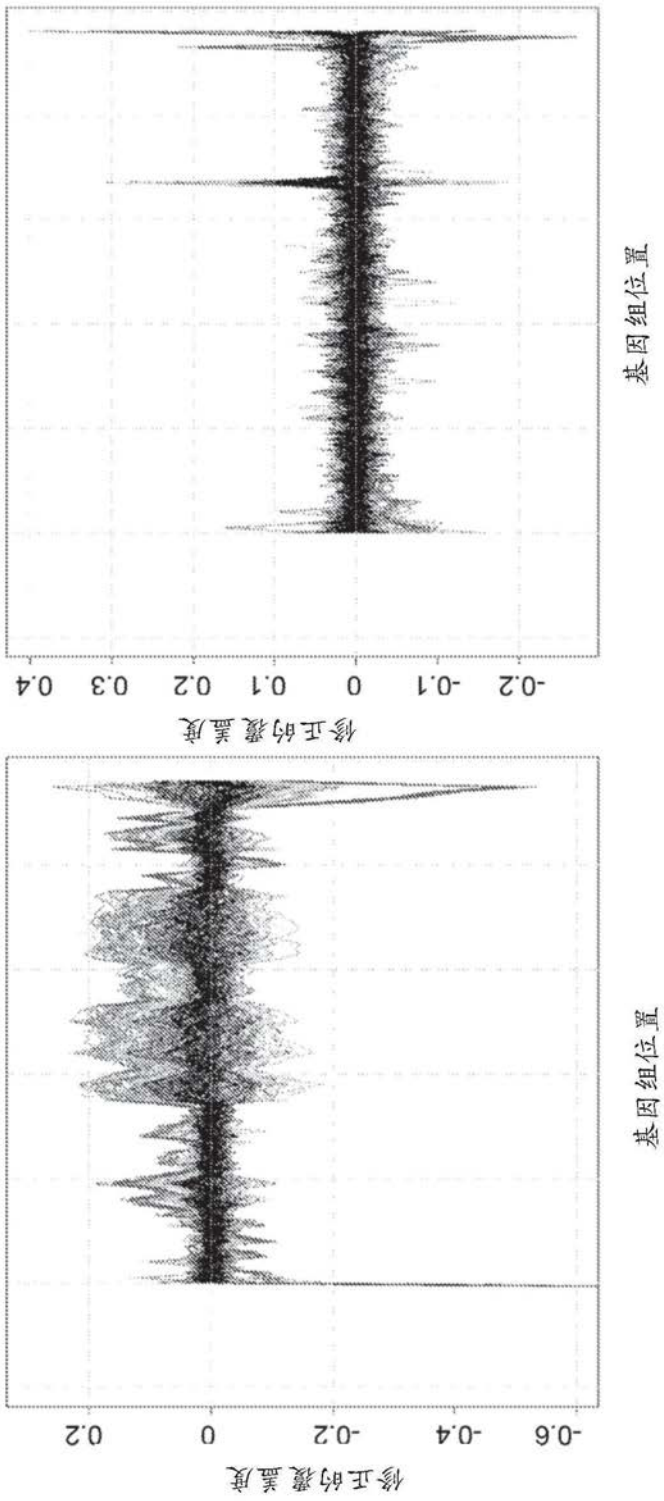


图3H

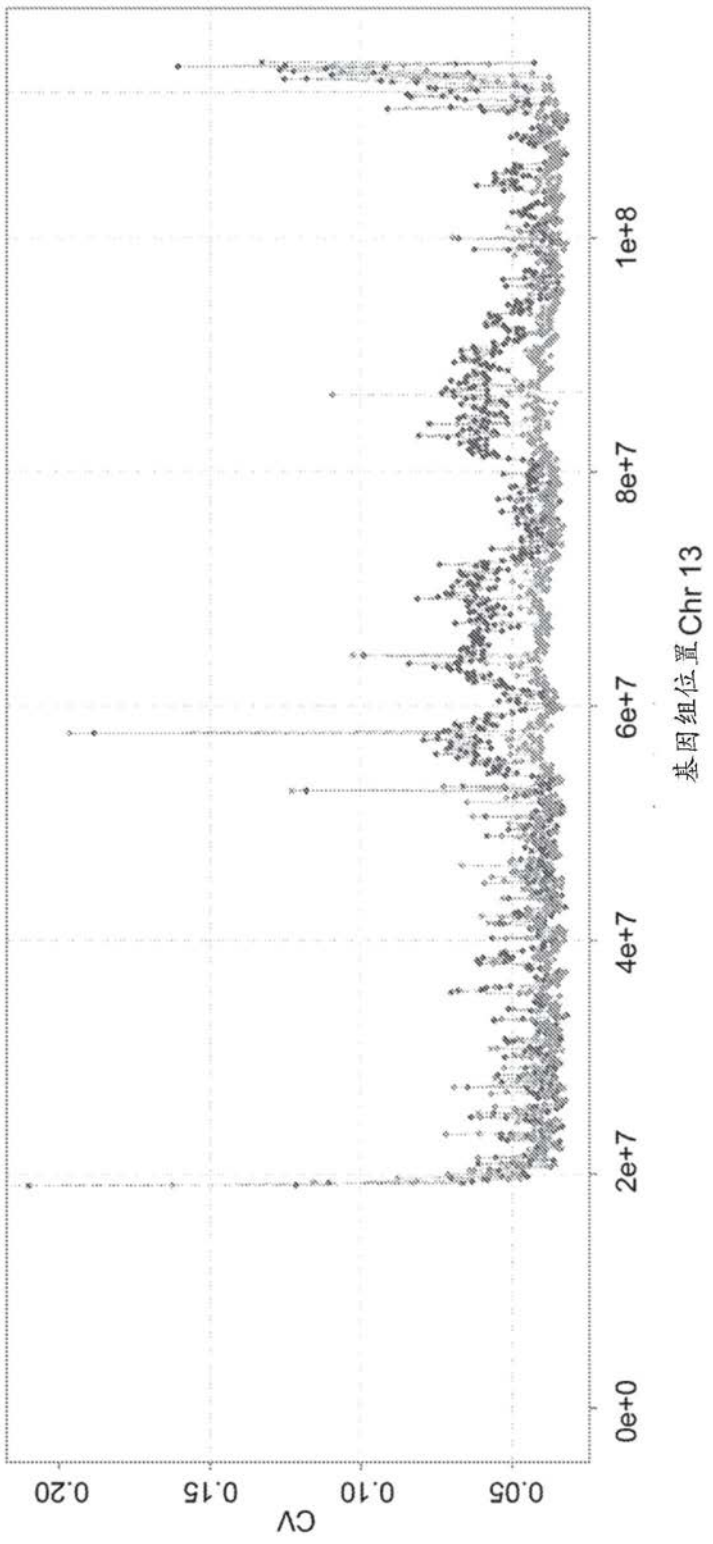


图3I

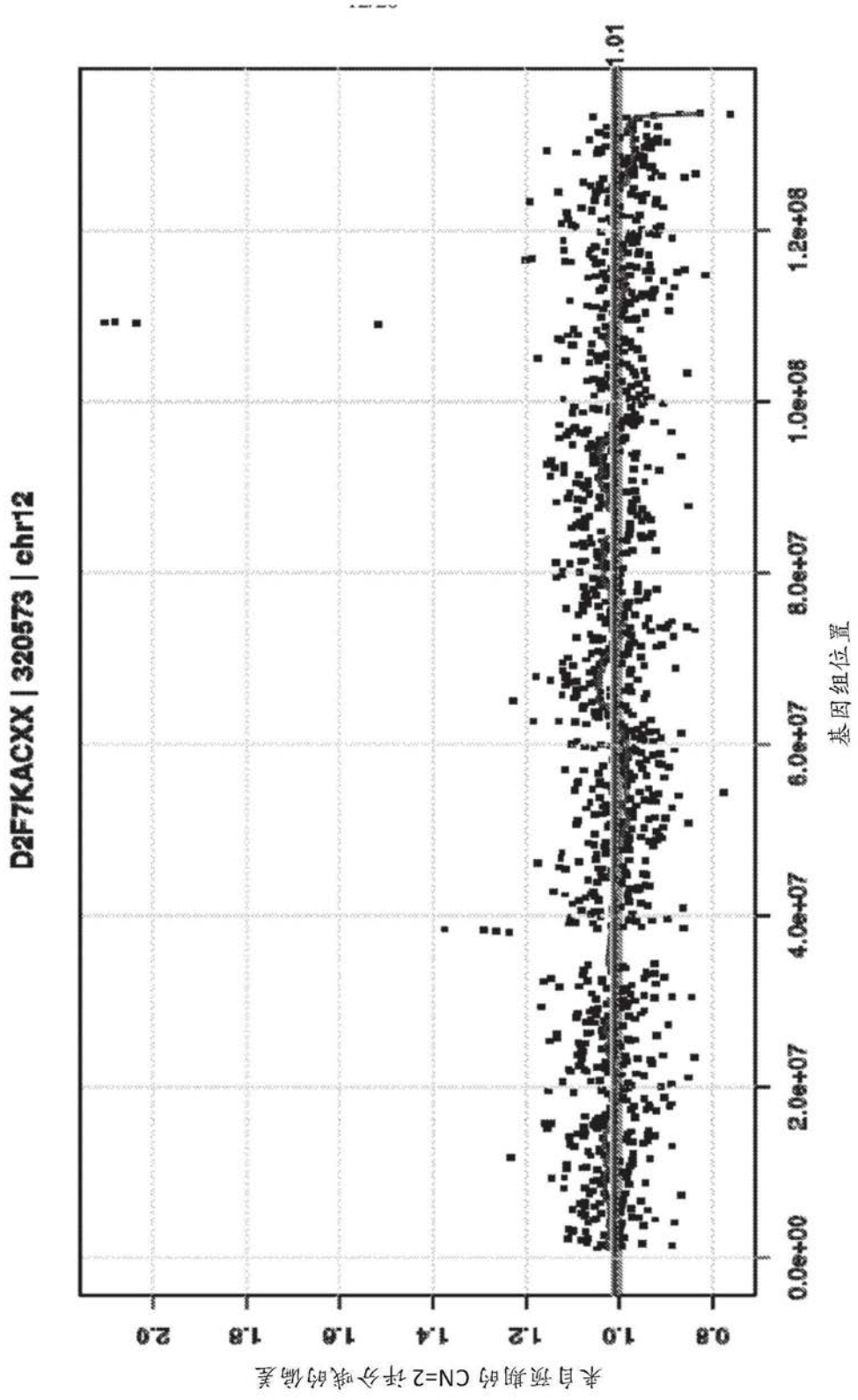

D2F7KACXX | 320573 | chr12
来自预期的 CN=2 评分哦的偏差
基因组位置
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
0.8
1.0
1.2
1.4
1.6
1.8
2.0
1.01

图3J

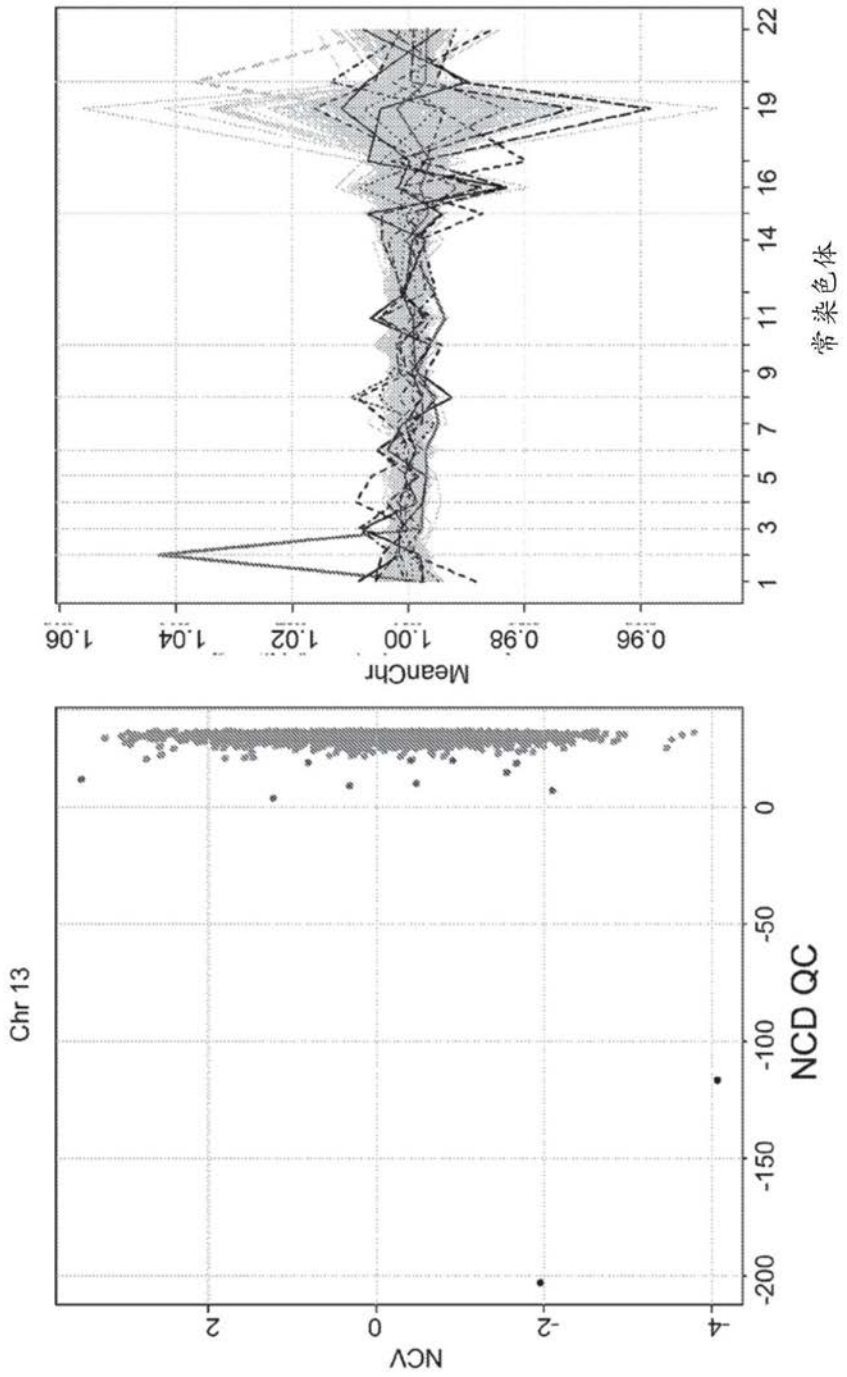

常染色体
1
3
5
7
9
11
14
16
19
22
MeanChr
0.96
0.98
1.00
1.02
1.04
1.06
NCD QC
0
-50
-100
-150
-200
NCV
2
0
-2
-4
Chr 13

图3K

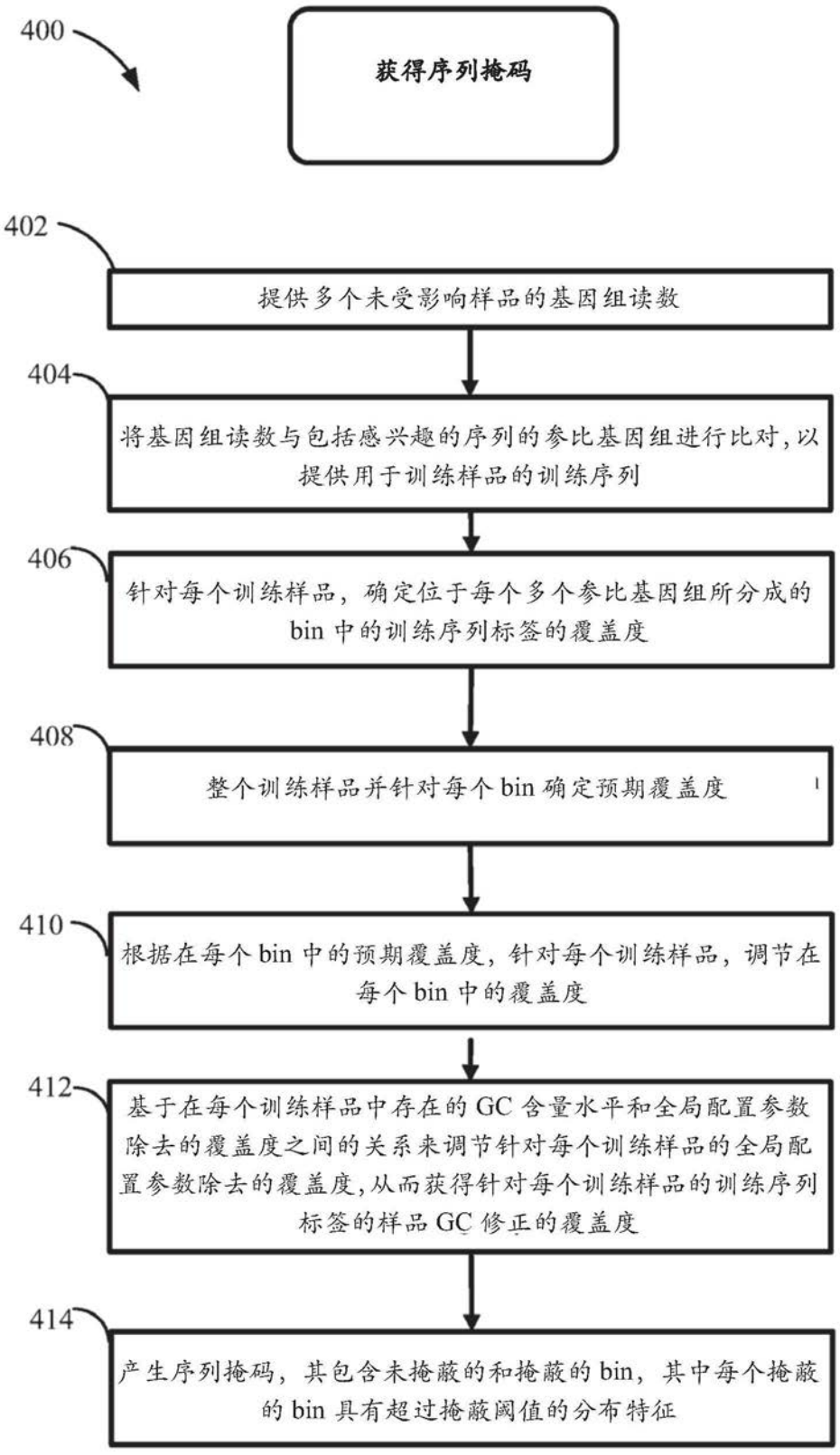

400
获得序列掩码
402
提供多个未受影响样品的基因组读数
404
将基因组读数与包括感兴趣的序列的参比基因组进行比对，以提供用于训练样品的训练序列
406
针对每个训练样品，确定位于每个多个参比基因组所分成的bin中的训练序列标签的覆盖度
408
整个训练样品并针对每个bin确定预期覆盖度
410
根据在每个bin中的预期覆盖度，针对每个训练样品，调节在每个bin中的覆盖度
412
基于在每个训练样品中存在的GC含量水平和全局配置参数除去的覆盖度之间的关系来调节针对每个训练样品的全局配置参数除去的覆盖度，从而获得针对每个训练样品的训练序列标签的样品GC修正的覆盖度
414
产生序列掩码，其包含未掩蔽的和掩蔽的bin，其中每个掩蔽的bin具有超过掩蔽阈值的分布特征

图4A

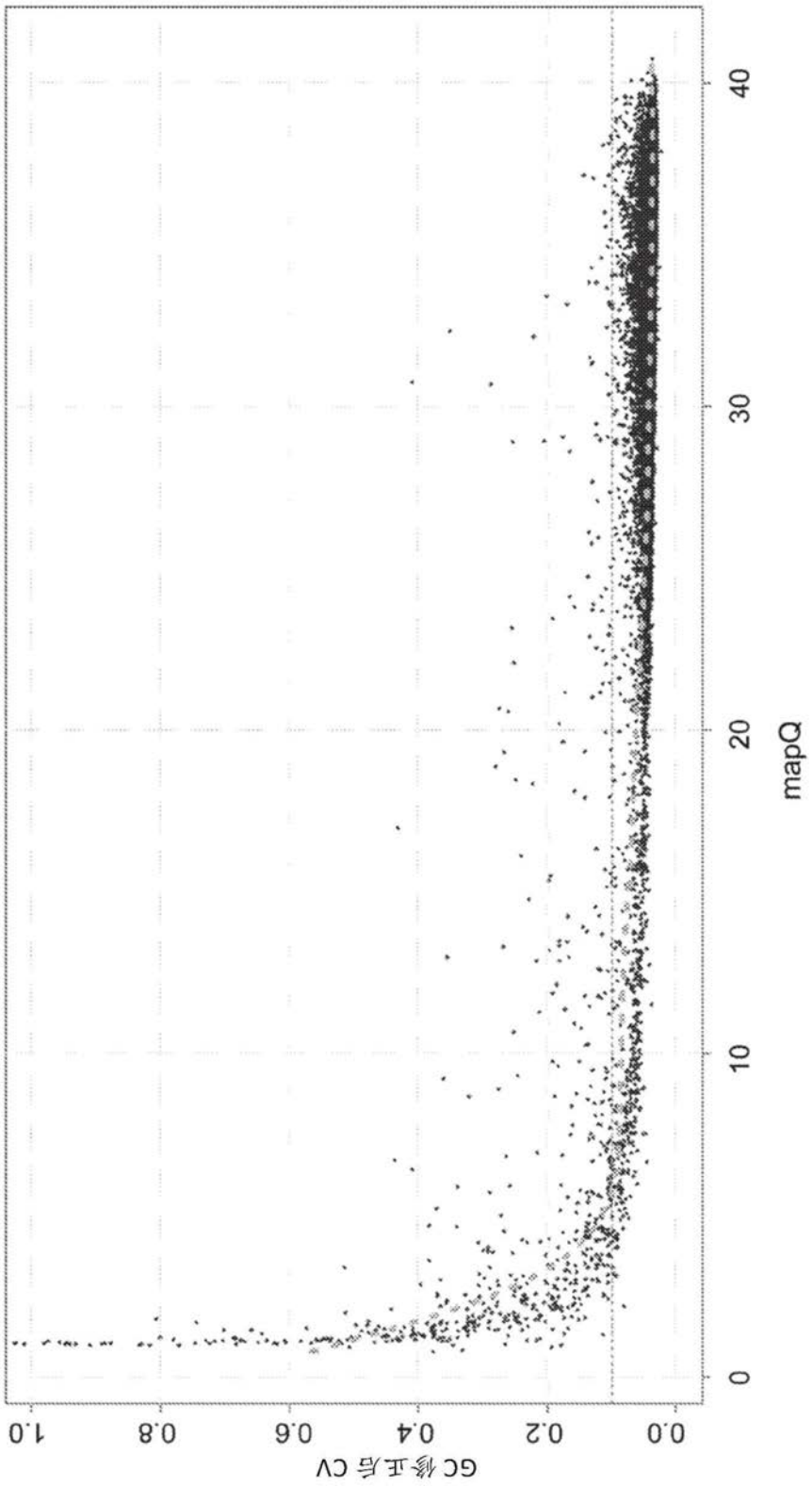


图4B

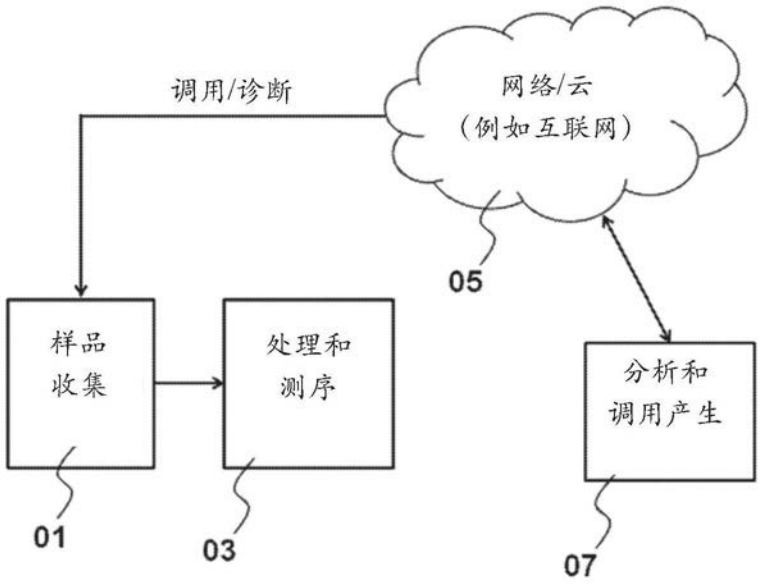

调用/诊断
网络/云
（例如互联网）
05
样品
收集
处理和
测序
分析和
调用产生
01
03
07

图5

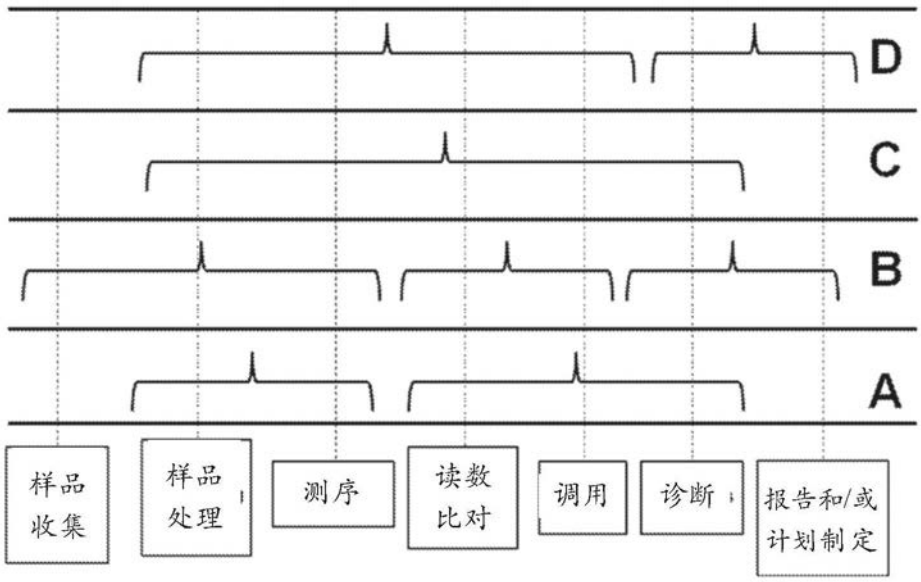

D
C
B
A
样品
收集
样品
处理
测序
读数
比对
调用
诊断
报告和/或
计划制定

图6

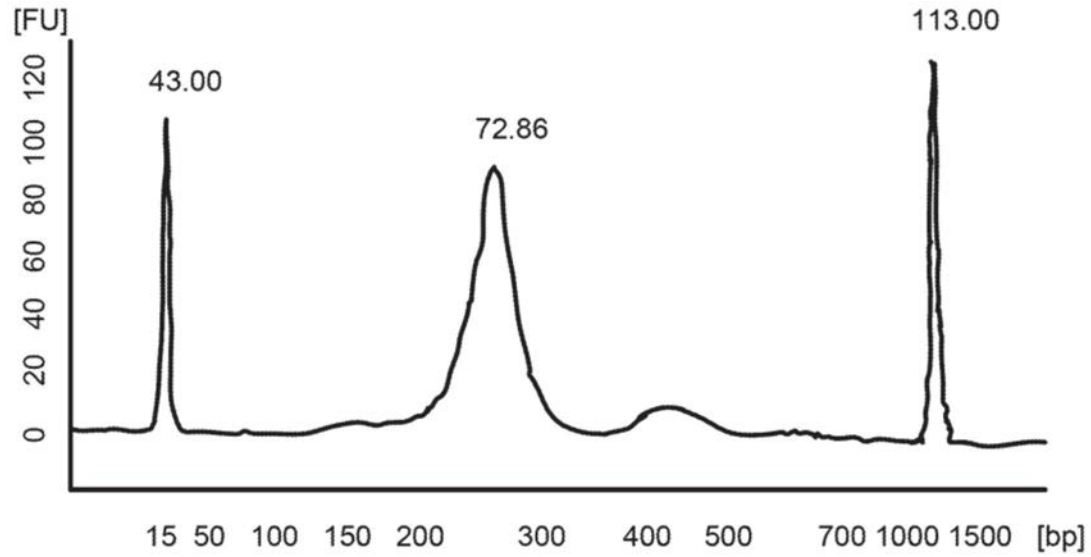

[FU]
120
100
80
60
40
20
0
43.00
72.86
113.00
15 50 100 150 200 300 400 500 700 1000 1500 [bp]

图7A

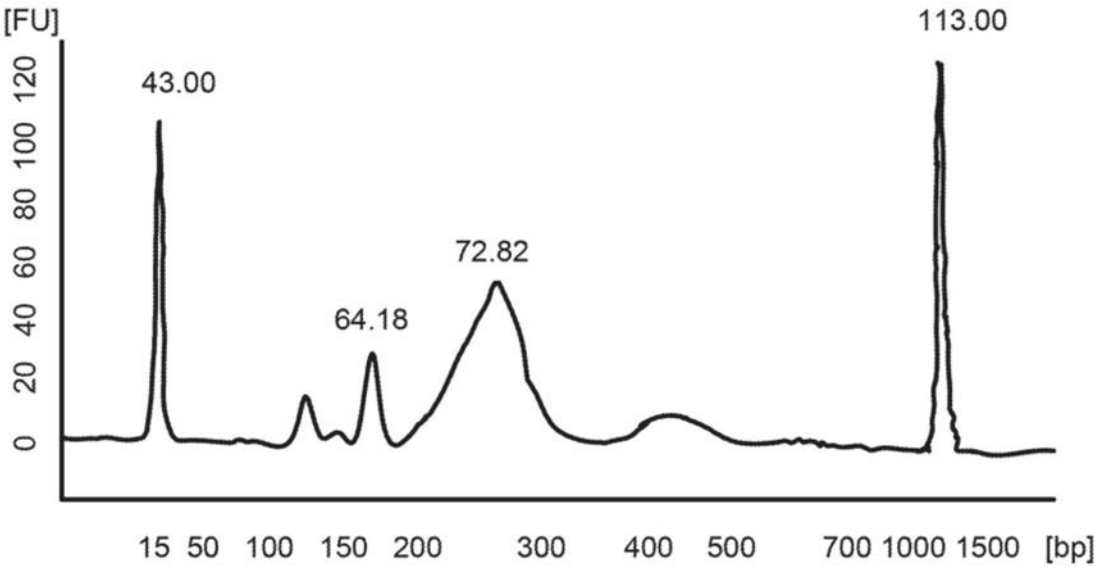

[FU]
120
100
80
60
40
20
0
43.00
64.18
72.82
113.00
15 50 100 150 200 300 400 500 700 1000 1500 [bp]

图7B

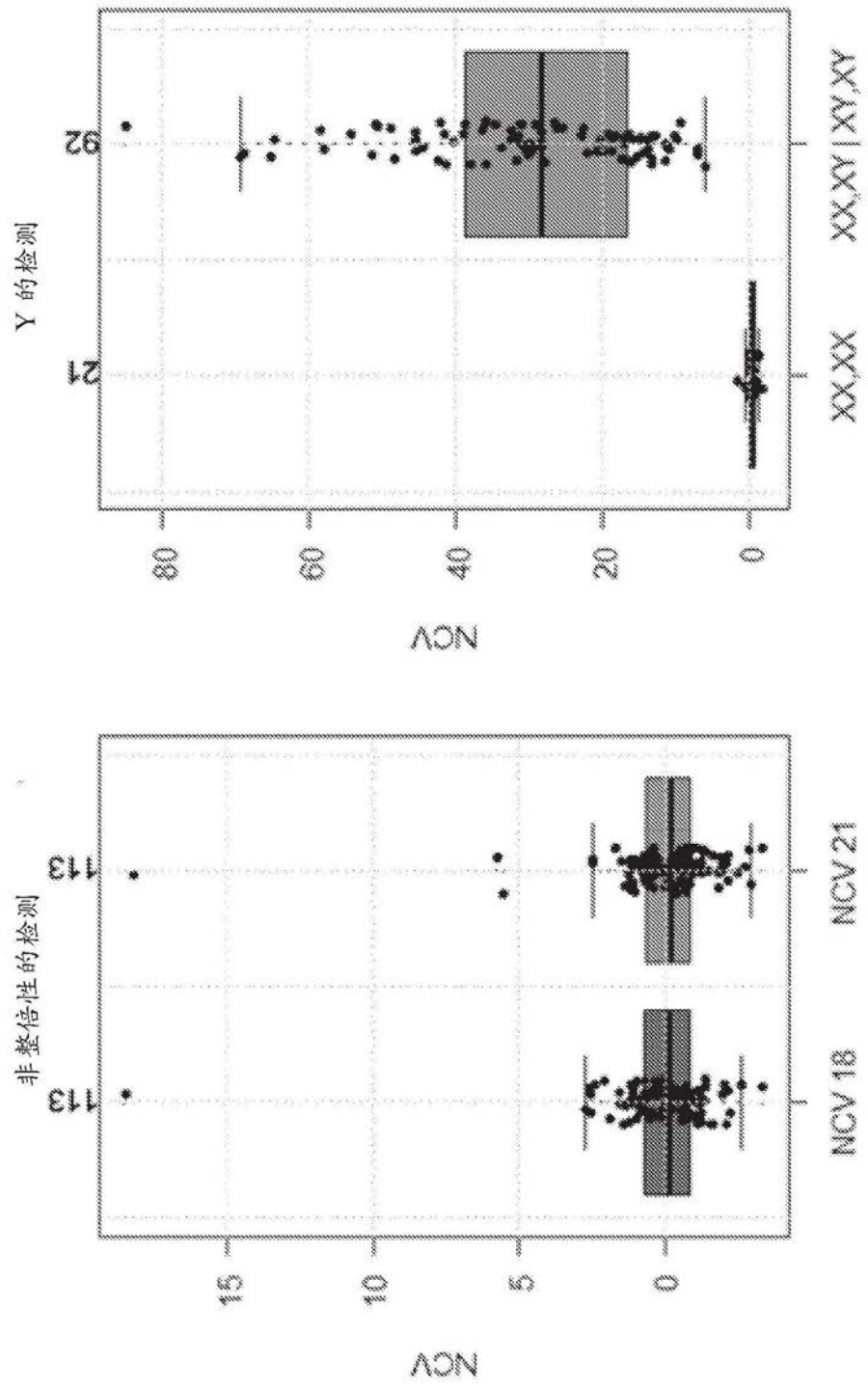
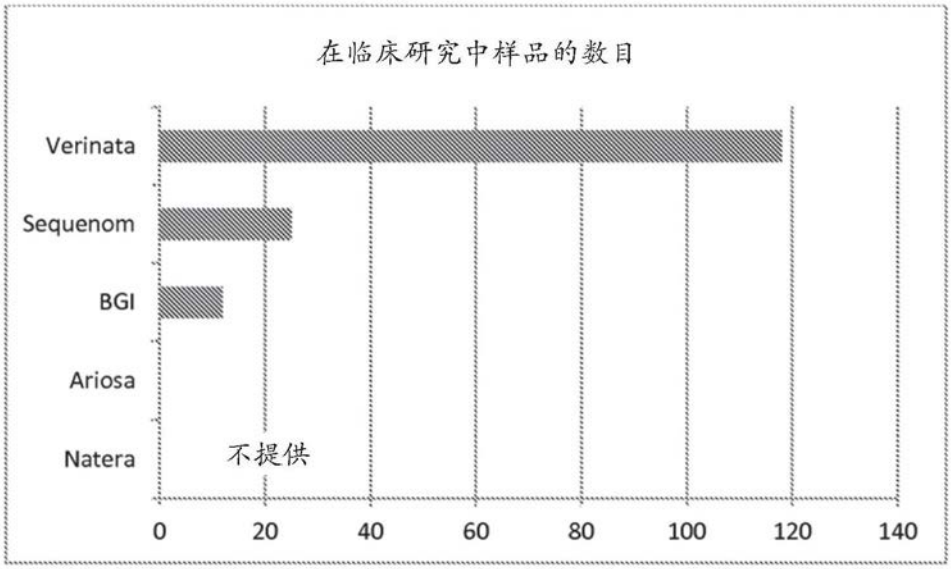

Y 的检测
92
21
XX,XY | XY,XY
XX,XX
NCV
80
60
40
20
0
非整倍性的检测
113
113
NCV 21
NCV 18
15
10
5
0
NCV

图8



Natera	不提供
Ariosa	0
BGI	12
Sequenom	25
Verinata	118

图9