



(10) 授权公告号 CN 112823391 B

(45) 授权公告日 2024. 07. 05

(21) 申请号 202080005447.1

(22) 申请日 2020.06.02

(65) 同一申请的已公布的文献号

申请公布号 CN 112823391 A

(43) 申请公布日 2021.05.18

(30) 优先权数据

62/856,651 2019.06.03 US

(85) PCT国际申请进入国家阶段日

2021.03.30

(86) PCT国际申请的申请数据

PCT/US2020/035787 2020.06.02

(87) PCT国际申请的公布数据

W02020/247411 EN 2020.12.10

(73) 专利权人 ILLUMINA公司

地址 美国加利福尼亚州

(72) 发明人 萨拉·L·金宁斯 科斯明·德丘

巴德里·帕杜卡萨哈斯拉姆

迪米特里·斯库沃佐夫

(74) 专利代理机构 北京英赛嘉华知识产权代理

有限责任公司 11204

专利代理师 王达佐 洪欣

(51) Int.Cl.

G16B 20/10 (2019.01)

G16B 30/10 (2019.01)

G16B 40/20 (2019.01)

G16B 40/30 (2019.01)

G16B 30/00 (2019.01)

(56) 对比文件

CN 108884491 A, 2018.11.23

US 2017316150 A1, 2017.11.02

审查员 柴珍珍

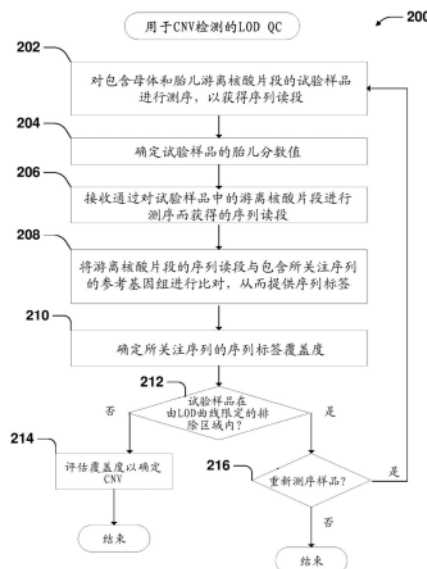
权利要求书8页 说明书55页 附图35页

(54) 发明名称

基于检测限的质量控制度量

(57) 摘要

本发明提供了用于在CNV检测中使用包含源自母体和胎儿的游离核酸片段的试验样品进行样品质量控制的方法和系统。该方法包括确定由至少胎儿分数检测限 (LOD) 曲线限定的排除区域。该胎儿分数LOD曲线随覆盖度值而变化, 并且指示在给定不同覆盖度的情况下实现检测标准所需的胎儿分数的最小值。



1. 一种用于处理包含源自母体和胎儿的游离核酸片段的试验样品的方法,所述方法是使用包括存储器和一个或多个处理器的计算机系统来实现的,所述方法包括:

(a) 确定所述试验样品的胎儿分数的观察值,其中所述试验样品的所述胎儿分数指示所述试验样品中胎源游离核酸片段的相对量;

(b) 由所述计算机系统接收通过对所述试验样品中的所述游离核酸片段进行测序而获得的序列读段;

(c) 由所述计算机系统将所述游离核酸片段的所述序列读段与包含关注序列的参考基因组进行比对,从而提供序列标签;

(d) 由所述计算机系统确定所述参考基因组的至少一部分的所述序列标签的覆盖度;

(e) 基于在 (d) 中确定的所述序列标签覆盖度和在 (a) 中确定的所述胎儿分数的观察值来确定所述试验样品在排除区域内,其中所述排除区域由至少胎儿分数检测限 (LOD) 曲线限定,其中所述胎儿分数LOD曲线随覆盖度值而变化,并且指示在不同覆盖度值的给定覆盖度值的情况下用于实现给定置信水平的胎儿分数的最小观察值,所述给定置信水平为对应的真实胎儿分数大于实现检测标准所需的胎儿分数的最小值;以及

(f) 作为确定所述试验样品在排除区域内的响应,排除所述试验样品用于对所述关注序列的拷贝数变异 (CNV) 作出判定,或对所述试验样品进行重新测序以获得用于对所述关注序列的CNV进行判定的重新测序的序列读段,

其中所述LOD曲线通过以下获得:

(i) 针对多个观察胎儿分数中的每个观察胎儿分数和针对多个覆盖度水平中的每个覆盖度水平,从第一多个训练样品中选择一个或多个选定的训练样品,所述训练样品的真实胎儿分数具有给定置信水平,其中每个训练样品包括源自母体和胎儿的游离核酸片段;

(ii) 使用一个或多个选定的训练样品,获得多个覆盖度水平的真实胎儿分数与观察胎儿分数曲线;

(iii) 使用第二多个训练样品获得真实胎儿分数,所述真实胎儿分数针对多个覆盖度水平至少具有实现检测标准所需的胎儿分数的最小值;

(iv) 使用真实胎儿分数与观察胎儿分数曲线和至少具有实现检测标准所需的胎儿分数的最小值的真实胎儿分数,获得针对具有给定置信度水平所需的多个覆盖度水平的多个最小观察胎儿分数,所述给定置信水平为真实胎儿分数至少具有实现检测标准所需的胎儿分数的最小值;以及

(v) 使用针对多个覆盖度水平的多个最小观察胎儿分数,获得胎儿分数LOD曲线。

2. 根据权利要求1所述的方法,所述方法还包括:

使用所述重新测序的序列读段来重复 (a) - (d); 以及

确定所述试验样品在所述排除区域之外。

3. 根据权利要求1所述的方法,其中基于受CNV影响的受影响训练样品的LOD来获得所述胎儿分数LOD曲线。

4. 根据权利要求3所述的方法,其中所述受影响训练样品包括计算机模拟样品。

5. 根据权利要求3所述的方法,其中所述受影响训练样品包括体外样品。

6. 根据权利要求3所述的方法,其中通过将具有两个或更多个胎儿分数的样品组合来获得所述受影响训练样品。

7. 根据权利要求1所述的方法, 其中所述检测标准是, 在所述胎儿分数的最小观察值和所述给定覆盖度值处, 存在所述给定置信水平, 所述给定置信水平为所述对应的真实胎儿分数大于以一定的检测置信度检测受影响的样品所需的指定LOD。

8. 根据权利要求7所述的方法, 其中所述检测标准是X%置信度, 所述X%置信度为对于所述胎儿分数的最小观察值而言, 所述对应的真实胎儿分数大于以Y%的检测置信度检测受影响的样品所需的指定LOD。

9. 根据权利要求8所述的方法, 其中X%为约50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、99%或99.5%。

10. 根据权利要求8所述的方法, 其中Y%为约50%、55%、60%、65%、70%、75%、80%、85%、90%、95%或99%的检测置信度。

11. 根据权利要求8所述的方法, 其中X%为50%并且Y%为95%。

12. 根据权利要求7所述的方法, 其中所述指定LOD被确定为可检测到Y%的受影响样品的最小观察胎儿分数。

13. 根据权利要求7所述的方法, 其中使用在观察覆盖度处观察胎儿分数的真实胎儿分数的分布来获得针对在所述覆盖度处的所述观察胎儿分数的所述检测标准。

14. 根据权利要求1-13中任一项所述的方法, 其中所述排除区域在所述胎儿分数LOD曲线的下方。

15. 根据权利要求1-13中任一项所述的方法, 其中所述排除区域由所述胎儿分数LOD曲线和覆盖度阈值限定。

16. 根据权利要求1-13中任一项所述的方法, 其中所述排除区域在所述胎儿分数LOD曲线和所述覆盖度阈值两者的下方。

17. 根据权利要求1-13中任一项所述的方法, 其中所述确定所述参考基因组的所述至少一部分的所述序列标签覆盖度包括:

(i) 将所述参考基因组分成多个分组;

(ii) 确定与每个分组进行比对的序列标签数目; 以及

(iii) 使用所述参考基因组的所述至少一部分中的分组中的所述序列标签数目来确定所述序列标签覆盖度。

18. 根据权利要求17所述的方法, 所述方法还包括在(iii)之前, 通过考虑由于除拷贝数变异之外的因素引起的分组间变异, 来调整与所述分组进行比对的所述序列标签数目。

19. 根据权利要求1-13中任一项所述的方法, 其中基于所述游离核酸片段的大小来确定所述试验样品的所述胎儿分数的观察值。

20. 根据权利要求19所述的方法, 其中所述试验样品的所述胎儿分数的观察值通过以下步骤确定:

获得所述游离核酸片段大小的频率分布; 以及

将所述频率分布应用于将胎儿分数与片段大小的频率相关联的模型, 以获得所述胎儿分数的观察值。

21. 根据权利要求17所述的方法, 其中基于所述参考基因组的分组的覆盖度信息来确定所述试验样品的所述胎儿分数的观察值。

22. 根据权利要求21所述的方法, 其中通过以下方式来计算所述胎儿分数的观察值: 将

所述参考基因组的多个分组的覆盖度值应用于将胎儿分数与分组的覆盖度相关联的模型,以获得所述胎儿分数的观察值。

23. 根据权利要求22所述的方法,其中所述参考基因组的所述多个分组具有比其他分组更高的胎儿游离核酸片段分数。

24. 根据权利要求1-13中任一项所述的方法,其中基于性染色体的分组的覆盖度信息来确定所述试验样品的所述胎儿分数的观察值。

25. 一种用于评估试验样品中关注核酸序列的拷贝数的系统,所述系统包括处理器和一个或多个计算机可读存储介质,所述一个或多个计算机可读存储介质具有存储在其上的用于在所述处理器上执行以下操作的指令:

(a) 确定所述试验样品的胎儿分数的观察值,其中所述试验样品的所述胎儿分数指示所述试验样品中胎源游离核酸片段的相对量;

(b) 接收通过对所述试验样品中的所述游离核酸片段进行测序而获得的序列读段;

(c) 将所述游离核酸片段的所述序列读段与包含关注序列的参考基因组进行比对,从而提供序列标签;

(d) 确定所述参考基因组的至少一部分的所述序列标签的覆盖度;以及

(e) 基于在(d)中确定的所述序列标签覆盖度和在(a)中确定的所述胎儿分数的观察值来确定所述试验样品在排除区域内,其中所述排除区域由至少胎儿分数检测限(LOD)曲线限定,其中所述胎儿分数LOD曲线随覆盖度值而变化,并且指示在不同覆盖度值的给定覆盖度值的情况下用于实现给定置信水平的胎儿分数的最小观察值,所述给定置信水平为对应的真实胎儿分数大于实现检测标准所需的胎儿分数的最小值,

其中所述LOD曲线通过以下获得:

(i) 从第一多个训练样品中,针对多个观察胎儿分数中的每个观察胎儿分数和针对多个覆盖度水平中的每个覆盖度水平,选择一个或多个选定的训练样品,所述训练样品的真实胎儿分数具有给定置信水平,其中每个训练样品包括源自母体和胎儿的游离核酸片段;

(ii) 使用一个或多个选定的训练样品,获得多个覆盖度水平的真实胎儿分数与观察胎儿分数曲线;

(iii) 使用第二多个训练样品获得真实胎儿分数,所述真实胎儿分数针对多个覆盖度水平至少具有实现检测标准所需的胎儿分数的最小值;

(iv) 使用真实胎儿分数与观察胎儿分数曲线和至少具有实现检测标准所需的胎儿分数的最小值的真实胎儿分数,获得针对具有给定置信度水平所需的多个覆盖度水平的多个最小观察胎儿分数,所述给定置信水平为真实胎儿分数至少具有实现检测标准所需的胎儿分数的最小值;以及

(v) 使用针对多个覆盖度水平的多个最小观察胎儿分数,获得胎儿分数LOD曲线。

26. 一种存储程序代码的机器可读介质,所述程序代码在由计算机系统的一个或多个处理器执行时,使得所述计算机系统:

(a) 确定试验样品的胎儿分数的观察值,其中所述试验样品的所述胎儿分数指示所述试验样品中胎源游离核酸片段的相对量;

(b) 接收通过对所述试验样品中的所述游离核酸片段进行测序而获得的序列读段;

(c) 将所述游离核酸片段的所述序列读段与包含关注序列的参考基因组进行比对,从

而提供序列标签;

(d) 确定所述参考基因组的至少一部分的所述序列标签的覆盖度;以及

(e) 基于在 (d) 中确定的所述序列标签覆盖度和在 (a) 中确定的所述胎儿分数的观察值来确定所述试验样品在排除区域内,其中所述排除区域由至少胎儿分数检测限 (LOD) 曲线限定,其中所述胎儿分数LOD曲线随覆盖度值而变化,并且指示在不同覆盖度值的给定覆盖度值的情况下用于实现给定置信水平的胎儿分数的最小观察值,所述给定置信水平为对应的真实胎儿分数大于实现检测标准所需的胎儿分数的最小值,

其中所述LOD曲线通过以下获得:

(i) 从第一多个训练样品中,针对多个观察胎儿分数中的每个观察胎儿分数和针对多个覆盖度水平中的每个覆盖度水平,选择一个或多个选定的训练样品,所述训练样品的真实胎儿分数具有给定置信水平,其中每个训练样品包括源自母体和胎儿的游离核酸片段;

(ii) 使用一个或多个选定的训练样品,获得多个覆盖度水平的真实胎儿分数与观察胎儿分数曲线;

(iii) 使用第二多个训练样品获得真实胎儿分数,所述真实胎儿分数针对多个覆盖度水平至少具有实现检测标准所需的胎儿分数的最小值;

(iv) 使用真实胎儿分数与观察胎儿分数曲线和至少具有实现检测标准所需的胎儿分数的最小值的真实胎儿分数,获得针对具有给定置信度水平所需的多个覆盖度水平的多个最小观察胎儿分数,所述给定置信水平为真实胎儿分数至少具有实现检测标准所需的胎儿分数的最小值;以及

(v) 使用针对多个覆盖度水平的多个最小观察胎儿分数,获得胎儿分数LOD曲线。

27. 一种用于评估试验样品中关注核酸序列的拷贝数的系统,所述系统包括一个或多个处理器和一个或多个计算机可读存储介质,所述一个或多个计算机可读存储介质具有存储在其上的用于实施以下的程序代码:

(a) 确定所述试验样品的胎儿分数的观察值,其中所述试验样品的所述胎儿分数指示所述试验样品中胎源游离核酸片段的相对量;

(b) 接收通过对所述试验样品中的所述游离核酸片段进行测序而获得的序列读段;

(c) 将所述游离核酸片段的所述序列读段与包含关注序列的参考基因组进行比对,从而提供序列标签;

(d) 确定所述参考基因组的至少一部分的所述序列标签的覆盖度;

(e) 基于在 (d) 中确定的所述序列标签覆盖度和在 (a) 中确定的所述胎儿分数的观察值来确定所述试验样品在排除区域内,其中所述排除区域由至少胎儿分数检测限 (LOD) 曲线限定,其中所述胎儿分数LOD曲线随覆盖度值而变化,并且指示在不同覆盖度值的给定覆盖度值的情况下用于实现给定置信水平的胎儿分数的最小观察值,所述给定置信水平为对应的真实胎儿分数大于实现检测标准所需的胎儿分数的最小值;以及

(f) 作为确定所述试验样品在排除区域内的响应,排除所述试验样品用于对所述关注序列的拷贝数变异 (CNV) 作出判定,或获得来自对所述试验样品进行重新测序的用于对所述关注序列的CNV进行判定的重新测序的序列读段,

其中所述LOD曲线通过以下获得:

(i) 从第一多个训练样品中,针对多个观察胎儿分数中的每个观察胎儿分数和针对多

个覆盖度水平中的每个覆盖度水平,选择一个或多个选定的训练样品,所述训练样品的真实胎儿分数具有给定置信水平,其中每个训练样品包括源自母体和胎儿的游离核酸片段;

(ii) 使用一个或多个选定的训练样品,获得多个覆盖度水平的真实胎儿分数与观察胎儿分数曲线;

(iii) 使用第二多个训练样品获得真实胎儿分数,所述真实胎儿分数针对多个覆盖度水平至少具有实现检测标准所需的胎儿分数的最小值;

(iv) 使用真实胎儿分数与观察胎儿分数曲线和至少具有实现检测标准所需的胎儿分数的最小值的真实胎儿分数,获得针对具有给定置信度水平所需的多个覆盖度水平的多个最小观察胎儿分数,所述给定置信水平为真实胎儿分数至少具有实现检测标准所需的胎儿分数的最小值;以及

(v) 使用针对多个覆盖度水平的多个最小观察胎儿分数,获得胎儿分数LOD曲线。

28. 根据权利要求27所述的系统,所述程序代码包括用于实施以下的代码:

使用所述重新测序的序列读段来重复(a) - (d);

确定所述试验样品在所述排除区域之外;以及

对所述试验样品做出关于所述关注序列的CNV判定。

29. 根据权利要求27所述的系统,其中基于受CNV影响的受影响训练样品的LOD来获得所述胎儿分数LOD曲线。

30. 根据权利要求29所述的系统,其中所述受影响训练样品包括计算机模拟样品。

31. 根据权利要求29所述的系统,其中所述受影响训练样品包括体外样品。

32. 根据权利要求29所述的系统,其中通过将具有两个或更多个胎儿分数的样品组合来获得所述受影响训练样品。

33. 根据权利要求27所述的系统,其中所述检测标准是,在所述胎儿分数的最小观察值和所述给定覆盖度值处,存在所述给定置信水平,所述给定置信水平为所述对应的真实胎儿分数大于以一定的检测置信度检测受影响的样品所需的指定LOD。

34. 根据权利要求33所述的系统,其中所述检测标准是X%置信度,所述X%置信度为对于所述胎儿分数的最小观察值而言,所述对应的真实胎儿分数大于以Y%的检测置信度检测受影响的样品所需的指定LOD。

35. 根据权利要求34所述的系统,其中X%为约50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、99%或99.5%。

36. 根据权利要求34所述的系统,其中Y%为约50%、55%、60%、65%、70%、75%、80%、85%、90%、95%或99%的检测置信度。

37. 根据权利要求34所述的系统,其中X%为50%并且Y%为95%。

38. 根据权利要求33所述的系统,其中所述指定LOD被确定为可检测到Y%的受影响样品的最小观察胎儿分数。

39. 根据权利要求33所述的系统,其中使用在观察覆盖度处观察胎儿分数的真实胎儿分数的分布来获得针对在所述覆盖度处的所述观察胎儿分数的所述检测标准。

40. 根据权利要求27-39中任一项所述的系统,其中所述排除区域在所述胎儿分数LOD曲线的下方。

41. 根据权利要求27-39中任一项所述的系统,其中所述排除区域由所述胎儿分数LOD

曲线和覆盖度阈值限定。

42. 根据权利要求27-39中任一项所述的系统,其中所述排除区域在所述胎儿分数LOD曲线和所述覆盖度阈值两者的下方。

43. 根据权利要求27-39中任一项所述的系统,其中所述确定所述参考基因组的所述至少一部分的所述序列标签覆盖度包括:

(i) 将所述参考基因组分成多个分组;

(ii) 确定与每个分组进行比对的序列标签数目;以及

(iii) 使用所述参考基因组的所述至少一部分中的分组中的所述序列标签数目来确定所述序列标签覆盖度。

44. 根据权利要求43所述的系统,其中所述确定所述参考基因组的所述部分的所述序列标签覆盖度还包括在(iii)之前,通过考虑由于除拷贝数变异之外的因素引起的分组间变异,来调整与所述分组进行比对的所述序列标签数目。

45. 根据权利要求27-39中任一项所述的系统,其中基于所述游离核酸片段的大小来确定所述试验样品的所述胎儿分数的观察值。

46. 根据权利要求45所述的系统,其中所述试验样品的所述胎儿分数的观察值通过以下步骤确定:

获得所述游离核酸片段大小的频率分布;以及

将所述频率分布应用于将胎儿分数与片段大小的频率相关联的模型,以获得所述胎儿分数的观察值。

47. 根据权利要求43所述的系统,其中基于所述参考基因组的分组的覆盖度信息来确定所述试验样品的所述胎儿分数的观察值。

48. 根据权利要求47所述的系统,其中通过以下方式来计算所述胎儿分数的观察值:将所述参考基因组的多个分组的覆盖度值应用于将胎儿分数与分组的覆盖度相关联的模型,以获得所述胎儿分数的观察值。

49. 根据权利要求48所述的系统,其中所述参考基因组的所述多个分组具有比其他分组更高的胎儿游离核酸片段分数。

50. 根据权利要求27-39中任一项所述的系统,其中基于性染色体的分组的覆盖度信息来确定所述试验样品的所述胎儿分数的观察值。

51. 一种存储程序代码的机器可读介质,所述程序代码在由计算机系统的一个或多个处理器执行时,使得所述计算机系统实施包括以下步骤的方法:

(a) 确定试验样品的胎儿分数的观察值,其中所述试验样品的所述胎儿分数指示所述试验样品中胎源游离核酸片段的相对量;

(b) 由所述计算机系统接收通过对所述试验样品中的所述游离核酸片段进行测序而获得的序列读段;

(c) 由所述计算机系统将所述游离核酸片段的所述序列读段与包含关注序列的参考基因组进行比对,从而提供序列标签;

(d) 由所述计算机系统确定所述参考基因组的至少一部分的所述序列标签的覆盖度;

(e) 基于在(d)中确定的所述序列标签覆盖度和在(a)中确定的所述胎儿分数的观察值来确定所述试验样品在排除区域内,其中所述排除区域由至少胎儿分数检测限(LOD)曲线

限定,其中所述胎儿分数LOD曲线随覆盖度值而变化,并且指示在不同覆盖度值的给定覆盖度值的情况下用于实现给定置信水平的胎儿分数的最小观察值,所述给定置信水平为对应的真实胎儿分数大于实现检测标准所需的胎儿分数的最小值;以及

(f) 作为确定所述试验样品在排除区域内的响应,排除所述试验样品用于对所述关注序列的拷贝数变异 (CNV) 作出判定,或获得来自对所述试验样品进行重新测序的用于对所述关注序列的CNV进行判定的重新测序的序列读段,

其中所述LOD曲线通过以下获得:

(i) 从第一多个训练样品中,针对多个观察胎儿分数中的每个观察胎儿分数和针对多个覆盖度水平中的每个覆盖度水平,选择一个或多个选定的训练样品,所述训练样品的真实胎儿分数具有给定置信水平,其中每个训练样品包括源自母体和胎儿的游离核酸片段;

(ii) 使用一个或多个选定的训练样品,获得多个覆盖度水平的真实胎儿分数与观察胎儿分数曲线;

(iii) 使用第二多个训练样品获得真实胎儿分数,所述真实胎儿分数针对多个覆盖度水平至少具有实现检测标准所需的胎儿分数的最小值;

(iv) 使用真实胎儿分数与观察胎儿分数曲线和至少具有实现检测标准所需的胎儿分数的最小值的真实胎儿分数,获得针对具有给定置信度水平所需的多个覆盖度水平的多个最小观察胎儿分数,所述给定置信水平为真实胎儿分数至少具有实现检测标准所需的胎儿分数的最小值;以及

(v) 使用针对多个覆盖度水平的多个最小观察胎儿分数,获得胎儿分数LOD曲线。

52. 根据权利要求51所述的机器可读介质,其中所述试验样品已被确定对于所述关注序列的CNV判定是阴性的。

53. 根据权利要求51所述的机器可读介质,所述方法还包括以下步骤:

使用所述重新测序的序列读段来重复 (a) - (d);

确定所述试验样品在所述排除区域之外;以及

对所述试验样品做出关于所述关注序列的CNV判定。

54. 根据权利要求51所述的机器可读介质,其中基于受CNV影响的受影响训练样品的LOD来获得所述胎儿分数LOD曲线。

55. 根据权利要求54所述的机器可读介质,其中所述受影响训练样品包括计算机模拟样品。

56. 根据权利要求54所述的机器可读介质,其中所述受影响训练样品包括体外样品。

57. 根据权利要求54所述的机器可读介质,其中通过将具有两个或更多个胎儿分数的样品组合来获得所述受影响训练样品。

58. 根据权利要求51所述的机器可读介质,其中所述检测标准是,在所述胎儿分数的最小观察值和所述给定覆盖度值处,存在所述给定置信水平,所述给定置信水平为所述对应的真实胎儿分数大于以一定的检测置信度检测受影响的样品所需的指定LOD。

59. 根据权利要求58所述的机器可读介质,其中所述检测标准是X%置信度,所述X%置信度为对于所述胎儿分数的最小观察值而言,所述对应的真实胎儿分数大于以Y%的检测置信度检测受影响的样品所需的指定LOD。

60. 根据权利要求59所述的机器可读介质,其中X%为约50%、55%、60%、65%、70%、

75%、80%、85%、90%、95%、99%或99.5%。

61. 根据权利要求59所述的机器可读介质,其中Y%为约50%、55%、60%、65%、70%、75%、80%、85%、90%、95%或99%的检测置信度。

62. 根据权利要求59所述的机器可读介质,其中X%为50%并且Y%为95%。

63. 根据权利要求58所述的机器可读介质,其中所述指定LOD被确定为可检测到Y%的受影响样品的最小观察胎儿分数。

64. 根据权利要求58所述的机器可读介质,其中使用在观察覆盖度处观察胎儿分数的真实胎儿分数的分布来获得针对在所述覆盖度处的所述观察胎儿分数的所述检测标准。

65. 根据权利要求51-64中任一项所述的机器可读介质,其中所述排除区域在所述胎儿分数LOD曲线的下方。

66. 根据权利要求51-64中任一项所述的机器可读介质,其中所述排除区域由所述胎儿分数LOD曲线和覆盖度阈值限定。

67. 根据权利要求51-64中任一项所述的机器可读介质,其中所述排除区域在所述胎儿分数LOD曲线和所述覆盖度阈值两者的下方。

68. 根据权利要求51-64中任一项所述的机器可读介质,其中所述确定所述参考基因组的所述至少一部分的所述序列标签覆盖度包括:

(i) 将所述参考基因组分成多个分组;

(ii) 确定与每个分组进行比对的序列标签数目;以及

(iii) 使用所述参考基因组的所述至少一部分中的分组中的所述序列标签数目来确定所述序列标签覆盖度。

69. 根据权利要求68所述的机器可读介质,所述确定所述参考基因组的所述部分的所述序列标签覆盖度还包括在(iii)之前,通过考虑由于除拷贝数变异之外的因素引起的分组间变异,来调整与所述分组进行比对的所述序列标签数目。

70. 根据权利要求51-64中任一项所述的机器可读介质,其中基于所述游离核酸片段的大小来确定所述试验样品的所述胎儿分数的观察值。

71. 根据权利要求70所述的机器可读介质,其中所述试验样品的所述胎儿分数的观察值通过以下步骤确定:

获得所述游离核酸片段大小的频率分布;以及

将所述频率分布应用于将胎儿分数与片段大小的频率相关联的模型,以获得所述胎儿分数的观察值。

72. 根据权利要求68所述的机器可读介质,其中基于所述参考基因组的分组的覆盖度信息来确定所述试验样品的所述胎儿分数的观察值。

73. 根据权利要求72所述的机器可读介质,其中通过以下方式来计算所述胎儿分数的观察值:将所述参考基因组的多个分组的覆盖度值应用于将胎儿分数与分组的覆盖度相关联的模型,以获得所述胎儿分数的观察值。

74. 根据权利要求73所述的机器可读介质,其中所述参考基因组的所述多个分组具有比其他分组更高的胎儿游离核酸片段分数。

75. 根据权利要求51-64中任一项所述的机器可读介质,其中基于性染色体的分组的覆盖度信息来确定所述试验样品的所述胎儿分数的观察值。

基于检测限的质量控制度量

[0001] 引用说明

[0002] PCT申请表格与本说明书同时提交,作为本申请的一部分。本申请要求同时提交的PCT申请表格中确定的权益或优先权的每项申请以引用的方式全文并入本文并用于所有目的。

背景技术

[0003] 人类医学研究中的重要工作之一是发现会产生不良健康后果的遗传异常。在许多情况下,已经在基因组中拷贝数异常的部分中识别出特定基因和/或关键诊断标记。例如,在产前诊断中,全染色体的额外拷贝或缺失拷贝是经常发生的遗传病变。在癌症中,全染色体或染色体区段的拷贝缺失或倍增以及基因组特定区域的较高水平扩增经常发生。

[0004] 关于拷贝数变异(CNV)的大多数信息已通过允许识别结构异常的细胞遗传学分辨率提供。用于基因筛检和生物剂量测定的常规手术已利用侵入性手术(例如羊膜穿刺术、脐带穿刺术或绒毛取样术(CVS))来获得用于染色体核型分析的细胞。由于认识到无需细胞培养的更快速测试方法的必要性,人们已经开发了荧光原位杂交(FISH)、荧光定量PCR(QF-PCR)和阵列-比较基因组杂交(array-CGH)作为用于分析拷贝数变异的分子细胞遗传学方法。

[0005] 可在相对较短的时间内对全基因组进行测序的技术的出现和循环游离DNA(cfDNA)的发现,提供了将源自一个待比较染色体的遗传物质与源自另一个待比较染色体的遗传物质进行比较而无需考虑与侵入性取样方法相关联的风险的机会,其为诊断所关注基因序列的多种拷贝数变异提供了工具。

[0006] 由于现有非侵入性产前诊断方法存在包括由于cfDNA的水平有限而导致的灵敏度不足和由于基因组信息的固有性质而导致的技术测序偏差的局限性,因此人们对会提供特异性、灵敏度和适用性中的任意一者或全部以在多种临床环境中可靠地诊断拷贝数变异的非侵入性方法存在持续需求。研究已表明,孕妇血浆中胎儿cfDNA片段的平均长度短于母体cfDNA片段。在本文的具体实施中利用母体和胎儿cfDNA之间的这种差异来确定CNV和/或胎儿分数。本文所公开的实施方案满足上述需求中的一些需求。一些实施方案提供了用于控制CNV检测中的样品质量的方法和系统,从而识别具有过低胎儿分数或读段覆盖度的样品以进行进一步处理,诸如重新测序。一些实施方案为非侵入性产前诊断和多种疾病的诊断提供了高分析灵敏度和特异性。

发明内容

[0007] 虽然本文的示例涉及人类并且语言主要针对人类,但本文所述的概念适用于任何植物或动物的基因组。本公开的这些和其他目的以及特征将根据以下描述和所附权利要求变得更加显而易见,或者可通过如下文所述的本公开的实践来了解。

[0008] 本公开的一个方面涉及用于处理试验样品的方法,每个试验样品包含源自母体和胎儿的游离核酸片段。所述方法使用包括存储器和一个或多个处理器的计算机系统来实

现。在一些具体实施中,方法包括:(a)确定试验样品的胎儿分数值,其中该试验样品的该胎儿分数指示该试验样品中胎源游离核酸片段的相对量;(b)由计算机系统接收通过对该试验样品中的所述游离核酸片段进行测序而获得的序列读段;(c)由计算机系统将所述游离核酸片段的所述序列读段与包含所关注序列的参考基因组进行比对,从而提供序列标签;(d)由计算机系统确定该参考基因组的至少一部分的所述序列标签的覆盖度;(e)基于在(d)中确定的所述序列标签覆盖度和在(a)中确定的胎儿分数来确定该试验样品在排除区域内,其中该排除区域由至少胎儿分数检测限(LOD)曲线限定,其中该胎儿分数LOD曲线随覆盖度值而变化,并且指示在给定不同覆盖度的情况下实现检测标准所需的胎儿分数的最小值;以及(f)排除该试验样品用于对该所关注序列的CNV作出判定,或对该试验样品进行重新测序以获得用于对该所关注序列的CNV作出判定的重新测序的序列读段。

[0009] 在一些具体实施中,方法还包括在(f)之前,确定试验样品对于所关注序列的CNV是阴性的。

[0010] 在一些具体实施中,方法还包括:使用重新测序的序列读段来重复(a)-(d);确定试验样品在排除区域之外;以及将试验样品判定为具有所关注序列的CNV或不具有所关注序列的CNV。

[0011] 在一些具体实施中,基于受CNV影响的受影响训练样品的LOD来获得胎儿分数LOD曲线。在一些具体实施中,受影响样品包括计算机模拟样品。在一些具体实施中,受影响样品包括体外样品。在一些具体实施中,通过将具有两个或更多个胎儿分数的样品组合来获得受影响样品。

[0012] 在一些具体实施中,检测标准是对于观察胎儿分数而言基准真值胎儿分数大于指定LOD的期望置信水平。在一些具体实施中,检测标准是对于观察胎儿分数而言基准真值胎儿分数大于LOD Y%的X%置信度。在一些具体实施中,X为约50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、99%或99.5%。在一些具体实施中,Y为约50%、55%、60%、65%、70%、75%、80%、85%、90%、95%或99%的检测置信度。在一些具体实施中,X为50%并且Y为95%。在一些具体实施中,指定LOD被确定为可检测到Y%的受影响样品的最小观察胎儿分数。在一些具体实施中,使用在观察覆盖度处观察胎儿分数的基准真值胎儿分数的分布来获得针对在该观察覆盖度处的该观察胎儿分数的检测标准。

[0013] 在一些具体实施中,排除区域在胎儿分数LOD曲线的下方。

[0014] 在一些具体实施中,排除区域由胎儿分数LOD曲线和覆盖度阈值限定。

[0015] 在一些具体实施中,排除区域在胎儿分数LOD曲线和覆盖度阈值两者的下方。

[0016] 在一些具体实施中,确定参考基因组的部分的序列标签覆盖度包括:(i)将参考基因组分成多个分组;(ii)确定与每个分组进行比对的序列标签数目;以及(iii)使用该参考基因组的部分中的分组中的所述序列标签数目来确定所述序列标签覆盖度。在一些具体实施中,方法还包括在(iii)之前,通过考虑由于除拷贝数变异之外的因素引起的分组间变异,来调整与分组进行比对的序列标签数目。

[0017] 在一些具体实施中,基于游离核酸片段的大小来确定试验样品的胎儿分数值。在一些具体实施中,通过以下方式来确定试验样品的胎儿分数值:获得游离核酸片段大小的频率分布;以及将该频率分布应用于将胎儿分数与片段大小的频率相关联的模型,以获得胎儿分数值。

[0018] 在一些具体实施中,基于参考基因组的分组的覆盖度信息来确定试验样品的胎儿分数值。在一些具体实施中,通过以下方式来计算胎儿分数值:将参考基因组的多个分组的覆盖度值应用于将胎儿分数与分组的覆盖度相关联的模型,以获得胎儿分数值。在一些具体实施中,参考基因组的多个分组具有比其他分组更高的胎儿游离核酸片段分数。

[0019] 在一些具体实施中,基于性染色体的分组的覆盖度信息来确定试验样品的胎儿分数值。

[0020] 本公开的另一方面涉及包括一个或多个处理器和系统存储器的计算机系统。所述一个或多个处理器被配置为执行上述方法中的任意一种方法。

[0021] 本公开的附加方面涉及计算机程序产品,该计算机程序产品包括一个或多个计算机可读非暂态存储介质,所述一个或多个计算机可读非暂态存储介质具有存储在其上的计算机可执行指令,所述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使得该计算机系统实现上述方法中的任意一种方法。

[0022] 引用并入

[0023] 本文提及的所有专利、专利申请和其他出版物,包括在这些参考文献中公开的所有序列,均明确地以引用方式并入本文,其程度如同具体且单独地指出每个单独的出版物、专利或专利申请以引用方式并入本文。所有引用文献的相关部分均全文以引用方式并入本文以用于本文引用的上下文所指示的目的。然而,不可将任何文献的引用理解为是对其作为本公开的现有技术的认可。

附图说明

[0024] 图1示意性地示出了根据覆盖度的确保检测所需的最小胎儿分数。

[0025] 图2示出了用于排除样品的两步覆盖度阈值。

[0026] 图3示出了三个群体或其样品的胎儿分数分布。

[0027] 图4A示出了根据一些具体实施的使用检测限 (LOD) QC方法来确定CNV的工作流程200。

[0028] 图4B示出了用于CNV检测的另一个LOD QC过程。

[0029] 图5示出了例示以统计概念为基础的LOD的经验和假设数据。

[0030] 图6示出了根据胎儿分数的检测概率。

[0031] 图7示出了估计胎儿分数包括导致所述估计胎儿分数偏离真实胎儿分数的误差。

[0032] 图8示出了由于覆盖度和由于胎儿分数值引起的胎儿分数误差。

[0033] 图9模拟了八个水平的观察胎儿分数的真实胎儿分数。

[0034] 图10示出了三个不同水平的误差或覆盖度的真实胎儿分数相对于观察胎儿分数的分布。

[0035] 图11示出了在给定不同误差或覆盖度的情况下2%的观察胎儿分数及其模拟真实胎儿分数分布。

[0036] 图12示出了根据观察胎儿分数的第20百分位(左图)和第25百分位(右图)的真实胎儿分数分布。

[0037] 图13示出了LOD、覆盖度和观察胎儿分数。

[0038] 图14包括可从类似于图13所示的数据中获得的两条胎儿分数LOD曲线。

- [0039] 图15示出了根据一些实施方案的用于确定存在拷贝数变异的方法。
- [0040] 图16A示出了用于评估拷贝数的三途径过程的流程图。
- [0041] 图16B示出了根据本公开的一些具体实施的用于根据覆盖度信息确定胎儿分数的示例性过程800。
- [0042] 图16C示出了根据一些具体实施的用于根据大小分布信息确定胎儿分数的过程。
- [0043] 图16D示出了根据本公开的一些具体实施的用于根据8聚体 (8-mer) 频率信息确定胎儿分数的示例性过程1000。
- [0044] 图17示出了用于从试验样品中产生判定或诊断的分散系统的一个具体实施。
- [0045] 图18示出了用于在不同位置处执行各种操作的选项。
- [0046] 图19示出了根据稀释分数的合成生成的样品的染色体Y覆盖度(左图)和FF级分估计值(右图)。
- [0047] 图20示出了LOD对比覆盖度的线性拟合的结果。
- [0048] 图21示出了预测LOD与观察LOD重叠的结果。
- [0049] 图22示出了NES的概率密度函数。
- [0050] 图23示出了根据观察胎儿分数的NES覆盖度。
- [0051] 图24示出了使用由LOD曲线和读段阈值限定的排除区域的数据排除。
- [0052] 图25示出了现有方法和LOD QC方法第一次运行和第二次运行的通过率和失败率。
- [0053] 图26示出了通过两步阈值法排除并通过胎儿分数LOD曲线方法挽救的数据。
- [0054] 图27示出了通过LOD QC方法挽救的样品。
- [0055] 图28示出了通过LOD QC方法排除的样品。
- [0056] 图29示出了现有方法和LOD QC方法两次运行的通过率和失败率。
- [0057] 图30示出了75%置信度LOD曲线挽救的样品。
- [0058] 图31示出了使用一些具体实施的方法检测T21样品的灵敏度。

具体实施方式

[0059] 定义

[0060] 如本文所用,关于数值的术语“约”是指 $\pm 10\%$ 。

[0061] 术语“由……组成”意指“包括并限于”。

[0062] 术语“基本上由……组成”意指组合物、方法或结构可包括另外的成分、步骤和/或部分,但前提是所述另外的成分、步骤和/或部分不在实质上改变所要求保护的组合物、方法或结构的基本特性和新颖特性。

[0063] 除非另外指明,否则本文所公开的方法和系统的实践包括分子生物学、微生物学、蛋白质纯化、蛋白质工程、蛋白质和DNA测序以及重组DNA领域中常用的常规技术和装置,这些技术和装置在本领域的技术范围内。此类技术和装置是本领域技术人员已知的,并且在许多文本和参考文献中有所描述(参见例如Sambrook等人,“Molecular Cloning: A Laboratory Manual”,第三版(Cold Spring Harbor),[2001年]);和Ausubel等人,“Current Protocols in Molecular Biology”[1987年])。

[0064] 数值范围包括限定该范围的数字。在本说明书通篇中给出的每一最大数值限度旨在包括每一更低数值限度,如同此类更低数值限度在本文中明确地写出一样。在本说明书

通篇中给出的每一最小数值限度将包括每一更高数值限度,如同此类更高数值限度在本文中明确地写出一样。在本说明书通篇中给出的每一数值范围将包括落入此类更宽数值范围内的每一更窄数值范围,如同此类更窄数值范围均在本文中明确写出一样。

[0065] 除非本文中另有定义,否则本文所用的所有技术和科学术语的含义与本发明所属领域的普通技术人员通常理解的含义相同。包括本文所包括的术语的各种科学词典是本领域技术人员熟知的并且是可用的。虽然与本文所述的方法和材料类似或等同的任何方法和材料也可用于本文所公开的实施方案的实践或测试,但本文描述了一些方法和材料。

[0066] 下文紧接着定义的术语通过整体参考本说明书来进行更全面的描述。应当理解,本公开不限于所述的特定方法、方案和试剂,因为这些方法、方案和试剂可根据本领域技术人员使用它们的上下文而变化。如本文所用,除非上下文另有明确指示,否则单数术语“一个”、“一种”和“该”包括复数指代。

[0067] 除非另外指明,否则分别地,以5'至3'的取向从左到右书写核酸,并且以氨基至羧基的取向从左到右书写氨基酸序列。

[0068] 检测限 (LOD) 是可以限定置信度检测的最小水平的信号 (例如,分析物、胎儿分数、指示条件的得分等)。在本申请中,LOD是以限定置信度检测非整倍体/CNV所需的胎儿分数 (或其他分析物) 的最低水平。

[0069] 本文所用的术语“参数”表示其值或其他特性影响相关条件 (诸如拷贝数变异) 的物理特征。在一些情况下,术语参数是指影响数学关系或模型输出的变量,该变量可以是自变量 (即,模型的输入) 或基于一个或多个自变量的中间变量。根据模型的范围,一个模型的输出可成为另一个模型的输入,从而成为另一个模型的参数。

[0070] 术语“片段大小参数”是指与片段或片段集合 (诸如核酸片段) 的大小或长度相关的参数;例如,从体液中获得的cfDNA片段。如本文所用,在以下情况时,参数“偏向片段大小或大小范围”:1) 对片段大小或大小范围的参数进行有利地加权,例如,当与该大小或大小范围的片段相关联时,计数的权重比其他大小或范围的计数的权重更大;或者2) 从对片段大小或大小范围有利地进行加权的值中获得参数,例如,当与该大小或大小范围的片段相关联时,从权重更大的计数中获得的比率。当基因组产生相对于来自另一基因组或同一基因组的另一部分的核酸片段富含或具有更高浓度的大小或大小范围的核酸片段时,片段大小或大小范围可以是该基因组或其一部分的特性。

[0071] 术语“加权”是指使用被认为是“权重”的一个或多个值或函数来修改量 (诸如参数或变量)。在某些实施方案中,将参数或变量乘以权重。在其他实施方案中,参数或变量以指数方式修改。在一些实施方案中,函数可以是线性或非线性函数。适用的非线性函数的示例包括但不限于赫维赛德 (Heaviside) 阶跃函数、箱车函数、阶梯函数或S形函数。对原始参数或变量进行加权可系统性地增大或减小加权变量的值。在各种实施方案中,加权可产生正值、非负值或负值。

[0072] 本文中术语“拷贝数变异”是指与存在于参考样品中的核酸序列的拷贝数相比,存在于试验样品中的核酸序列的拷贝数变异。在某些实施方案中,核酸序列为1kb或更大。在一些情况下,核酸序列是全染色体或其重要部分。“拷贝数变异体”是指通过将试验样品中的所关注核酸序列与所关注核酸序列的预期水平进行比较来发现拷贝数差异的核酸序列。例如,将试验样品中所关注核酸序列的水平与合格样品中存在的核酸序列水平进行比较。

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、倍增和易位。CNV涵盖染色体非整倍体和部分非整倍体。

[0073] 本文中术语“非整倍体”是指由全染色体或染色体的一部分的丢失或获得引起的遗传物质的失衡。

[0074] 本文中术语“染色体非整倍体”和“完整染色体非整倍体”是指由全染色体的丢失或获得引起的遗传物质的失衡,并且包括种系非整倍体和嵌合性非整倍体。

[0075] 本文中术语“部分非整倍体”和“部分染色体非整倍体”是指由染色体的一部分的丢失或获得引起的遗传物质的失衡(例如,部分单倍体性和部分三倍体性),并且涵盖由易位、缺失和插入引起的失衡。

[0076] 术语“多个”是指多于一个元件。例如,本文所用的术语是指使用本文所公开的方法足以识别试验样品和合格样品的拷贝数变异的显著差异的多个核酸分子或序列标签。在一些实施方案中,每个试验样品获得至少约 3×10^6 个约20b至40bp之间的序列标签。在一些实施例方案中,每个试验样品提供至少约 5×10^6 、 8×10^6 、 10×10^6 、 15×10^6 、 20×10^6 、 30×10^6 、 40×10^6 或 50×10^6 个序列标签的数据,每个序列标签的长度在约20bp和40bp之间。

[0077] 术语“配对末端读段”是指来自从核酸片段的每个末端获得一个读段的配对末端测序的读段。配对末端测序可包括将多核苷酸链片段化成称为插入序列的短序列。片段化对于相对较短的多核苷酸(诸如无细胞DNA分子)是可选的或不必要的。

[0078] 术语“多核苷酸”、“核酸”和“核酸分子”可互换使用,并且是指共价连接的核苷酸(即RNA的核糖核苷酸和DNA的脱氧核糖核苷酸)的序列,其中一个核苷酸的戊糖的3'位置通过磷酸二酯基团连接至下一个核苷酸的戊糖的5'位置。核苷酸包括任何形式的核酸的序列,这些核酸包括但不限于RNA和DNA分子(诸如cfDNA分子)。术语“多核苷酸”包括但不限于单链多核苷酸和双链多核苷酸。

[0079] 本文中术语“试验样品”是指通常来源于生物流体、细胞、组织、器官或生物体且包含核酸或核酸混合物的样品,该核酸或核酸混合物包含待筛选拷贝数变异的至少一种核酸序列。在某些实施方案中,样品包含拷贝数疑似已发生变异的至少一条核酸序列。此类样品包括但不限于痰/口腔液、羊水、血液、血液级分或细针活检样品(例如,外科活检、细针活检等)、尿液、腹膜液、胸膜液等。虽然样品通常取自人类受试者(例如,患者),但测定可用于任何哺乳动物的拷贝数变异(CNV),这些哺乳动物包括但不限于狗、猫、马、山羊、绵羊、牛、猪等。样品可按从生物来源中获得的原样直接使用,或者经过预处理以改变样品的性质后使用。例如,此类预处理可包括由血液制备血浆、稀释粘性流体等。预处理的方法还包括但不限于过滤、沉淀、稀释、蒸馏、混合、离心、冷冻、冻干、浓缩、扩增、核酸片段化、干扰组分失活、添加试剂、裂解等。如果对于样品采用此类预处理方法,则此类预处理方法通常使得所关注核酸保留在试验样品中,有时其浓度与未处理的试验样品(例如,即未受到任何此类预处理方法的样品)中的浓度成比例。就本文所述的方法而言,此类“经处理的”或“处理后的”样品仍被视为生物“试验”样品。

[0080] 本文中术语“合格样品”或“未受影响样品”是指包含以(试验样品中的核酸所比较的)已知拷贝数存在的核酸混合物的样品,并且合格样品或未受影响样品是对于所关注核酸序列而言正常,即不是非整倍体的样品。在一些实施方案中,将合格样品用作训练集的未受影响训练样品,以推导序列掩蔽或序列图谱。在某些实施方案中,将合格样品用于识别所

考虑染色体的一条或多条归一化染色体或区段。例如,合格样品可用于识别21号染色体的归一化染色体。在这种情况下,合格样品是非21-三倍体样品的样品。另一个示例包括仅使用女性个体作为X染色体的合格样品。也可将合格样品用于其他目的,诸如确定用于判定受影响样品的阈值、识别用于限定参考序列上的掩蔽区域的阈值、确定基因组不同区域的预期覆盖度量等。

[0081] 本文中术语“训练集”是指可包括受影响和/或未受影响样品并且用来开发用于分析试验样品的模型的训练样品集。在一些实施方案中,训练集包括未受影响样品。在这些实施方案中,使用不受所关注拷贝数变异影响的样品训练集来建立用于确定CNV的阈值。训练集中的未受影响样品可用作合格样品以识别归一化序列(例如,归一化染色体),并且未受影响样品的染色体剂量用于设定用于所关注序列(例如,所关注染色体)中的每一者的阈值。在一些实施方案中,训练集包括受影响样品。训练集中的受影响样品可用于确认受影响试验样品易于与未受影响样本区分开。

[0082] 训练集也是所关注群体中的统计样品,该统计样品不会与生物样品混淆。统计样品通常包括多个个体,这些个体的数据用于确定适用于群体的一个或多个所关注定量值。统计样品是所关注群体中的个体子集。这些个体可以是提供用于统计分析的数据点的人、动物、组织、细胞、其他生物样品(即,统计样品可包括多个生物样品)和其他个体实体。

[0083] 通常,训练集与验证集结合使用。术语“验证集”用于指统计样品中的个体集,这些个体的数据用于验证或评估使用训练集确定的所关注定量值。在一些实施方案中,例如,训练集提供用于计算参考序列的掩蔽的数据,而验证集提供用于评估掩蔽的有效性或效果的数据。

[0084] 本文所用的“拷贝数评估”是指与序列拷贝数相关的基因序列状态的统计评估。例如,在一些实施方案中,评估包括确定是否存在基因序列。在一些实施方案中,评估包括确定基因序列的部分或完整非整倍体。在其他实施方案中,评估包括基于基因序列的拷贝数来区分两个或更多个样品。在一些实施方案中,评估包括基于基因序列的拷贝数来进行统计分析,例如归一化和比较。

[0085] 术语“合格核酸”可与“合格序列”互换使用,“合格序列”是所关注序列或核酸的量所比较的序列。合格序列是优选按已知表示(即,合格序列的量是已知的)存在于生物学样品中的一种序列。一般来讲,合格序列是存在于“合格样品”中的序列。“所关注合格序列”是在合格样品中已知其量的合格序列,并且是与对照受试者和具有医学病症的个体之间的所关注序列的差异相关联的序列。

[0086] 本文中术语“所关注序列”或“所关注核酸序列”是指与健康个体和患病个体之间的序列表示的差异相关联的核酸序列。所关注序列可以是在疾病或遗传病症中错误表示(即过度表示或表示不足)的染色体上的序列。所关注序列可以是染色体的一部分(即染色体区段)或全染色体。例如,所关注序列可以是在非整倍体病症中过度表示的染色体或编码在癌症中表示不足的肿瘤抑制因子的基因。所关注序列包括在受试者的细胞总群或细胞亚群中过度表示或表示不足的序列。“所关注合格序列”是合格样品中的所关注序列。“所关注测试序列”是试验样品中的所关注序列。

[0087] 本文中术语“归一化序列”是指用于将映射到与归一化序列相关联的所关注序列的序列标签数目归一化的序列。在一些实施方案中,归一化序列包含稳健染色体(robust

chromosome)。“稳健染色体”是不可能为非整倍体的染色体。在涉及人类染色体的一些情况下,稳健染色体是除X染色体、Y染色体、13号染色体、18号染色体和21号染色体之外的任何染色体。在一些实施方案中,归一化序列显示在样品和测序运行间映射到其的序列标签数目的变异性,该变异性近似于用作归一化参数的所关注序列的变异性。归一化序列可将受影响样品与一个或多个未受影响样品区分开。在一些具体实施中,当与其他潜在归一化序列(诸如其他染色体)相比时,归一化序列最佳地或有效地将受影响样品与一个或多个未受影响样品区分开。在一些实施方案中,以在样品和测序运行间所关注序列的染色体剂量变异性来计算归一化序列的变异性。在一些实施方案中,在未受影响样品集中识别归一化序列。

[0088] “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”的示例。“归一化染色体序列”可由单条染色体或一组染色体构成。在一些实施方案中,归一化序列包含两条或更多条稳健染色体。在某些实施方案中,稳健染色体均为除X染色体、Y染色体、13号染色体、18号染色体和21号染色体之外的染色体。“归一化区段”是“归一化序列”的另一个示例。“归一化区段序列”可由染色体的单个片段构成,或可由相同或不同染色体的两个或更多个片段构成。在某些实施方案中,归一化序列旨在将变异性诸如过程相关变异性、染色体间(运行内)变异性和测序间(运行间)变异性归一化。

[0089] 本文中术语“可区分性”是指归一化染色体的特性,其使得人们能够将一个或多个未受影响(即正常)样品与一个或多个受影响(即非整倍体)样品区分开。显示最大“可区分性”的归一化染色体是这样的染色体或染色体组,其提供合格样品集中所关注染色体的染色体剂量与一个或多个受影响样品中的对应染色体中相同所关注染色体的染色体剂量的分布之间的最大统计差异。

[0090] 本文中术语“变异性”是指归一化染色体的特性,其使得人们能够将一个或多个未受影响(即正常)样品与一个或多个受影响(即非整倍体)样品区分开。在合格样品集中测量的归一化染色体的变异性是指映射到其的序列标签数目的变异性,该变异性近似于映射到用作归一化参数的所关注染色体的序列标签数目的变异性。

[0091] 本文中术语“序列标签密度”是指映射到参考基因组序列的序列读段的数目,例如21号染色体的序列标签密度是通过测序方法生成的映射到参考基因组的21号染色体的序列读段的数目。

[0092] 本文中术语“序列标签密度比”是指映射到参考基因组的染色体(例如21号染色体)的序列标签数目与参考基因组染色体长度的比率。

[0093] 本文中术语“序列剂量”是指将所识别的所关注序列的序列标签数目或另一个参数与所识别的归一化序列的序列标签数目或另一个参数相关联的参数。在一些情况下,序列剂量是所关注序列的序列标签覆盖度或其他参数与归一化序列的序列标签覆盖度或其他参数的比率。在一些情况下,序列剂量是指将所关注序列的序列标签密度与归一化序列的序列标签密度相关联的参数。“测试序列剂量”是将在试验样品中确定的所关注序列(例如21号染色体)的序列标签密度或其他参数与归一化序列(例如9号染色体)的序列标签密度或其他参数相关联的参数。类似地,“合格序列剂量”是将在合格样品中确定的所关注序列的序列标签密度或其他参数与归一化序列的序列标签密度或其他参数相关的参数。

[0094] 术语“覆盖度”是指映射到限定序列的序列标签的丰度。可通过序列标签密度(或

序列标签的计数)、序列标签密度比、归一化覆盖度量、调整的覆盖度值等定量地表示覆盖度。

[0095] 术语“覆盖度量”是指原始覆盖度的修改,并且通常表示基因组区域诸如分组中序列标记的相对量(有时称为计数)。可通过归一化、调整和/或校正基因组区域的原始覆盖度或计数来获得覆盖度量。例如,可通过将映射到区域的序列标签计数除以映射到全基因组的序列标签总数来获得该区域的归一化覆盖度量。归一化覆盖度量允许在不同样品间比较分组覆盖度,这些样品可具有不同的测序深度。归一化覆盖度量与序列剂量的不同之处在于后者通常是通过除以映射到全基因组的子集的标签计数获得的。该子集是一个或多个归一化区段或染色体。无论是否归一化,覆盖度量都可针对基因组上区域与区域之间的全局图谱变化、G-C分数变化、稳健染色体异常值等加以校正。

[0096] 本文中术语“下一代测序(NGS)”是指允许对克隆扩增分子和单个核酸分子进行大规模平行测序的测序方法。NGS的非限制性示例包括边连接边测序和使用可逆染料终止子的边合成边测序。

[0097] 本文中术语“参数”是指表征系统特性的数值。通常,参数以数值方式表征定量数据集和/或定量数据集之间的数值关系。例如,映射到染色体的序列标签数目和标签所映射的染色体长度之间的比率(或比率的函数)是参数。

[0098] 本文中术语“阈值”和“合格阈值”是指用作表征样品(诸如含有来自疑似具有医学病症的生物体的核酸的试验样品)的截止值的任何数字。可将阈值与参数值进行比较,以确定产生此参数值的样品是否表明生物体具有医学病症。在某些实施方案中,使用合格数据集来计算合格阈值,并将该合格阈值用作生物体中拷贝数变异(例如,非整倍体)的诊断限。如果从本文所公开的方法中获得的结果超过阈值,则受试者可被诊断为具有拷贝数变异,例如,21-三倍体症。可通过分析所计算的样品训练集的归一化值(例如染色体剂量、NCV或NSV)来确定用于本文所述方法的适当阈值。可使用包括合格(即,未受影响)样品和受影响样品两者的训练集中的合格(即,未受影响)样品来确定阈值。训练集中已知具有染色体非整倍体的样品(即,受影响样品)可用于确认所选阈值可用于将试验集中的受影响样品与未受影响样品区分开(参见本文的实施例)。阈值的选择取决于用户希望必须进行分类的置信水平。在一些实施方案中,用于确定适当阈值的训练集包括至少10个、至少20个、至少30个、至少40个、至少50个、至少60个、至少70个、至少80个、至少90个、至少100个、至少200个、至少300个、至少400个、至少500个、至少600个、至少700个、至少800个、至少900个、至少1000个、至少2000个、至少3000个、至少4000个或更多个合格样品。使用更大的合格样品集来改善阈值的诊断效用可能是有利的。

[0099] 术语“分组”是指序列的区段或基因组的区段。在一些实施方案中,分组在基因组或染色体彼此邻接。每个分组可限定参考基因组中的核苷酸序列。根据特定应用和序列标签密度所需的分析,分组的大小可以是1kb、100kb、1Mb等。除了在参考序列内的位置之外,分组还可具有其他特性,诸如样品覆盖度和序列结构特性(诸如G-C分数)。

[0100] 本文所用的术语“掩蔽阈值”是指基于序列分组中序列标签数目的值所比较的量,其中具有超过掩蔽阈值的值的分组被掩蔽。在一些实施方案中,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。在一些实施方案中,掩蔽阈值可被定义为在多个未受影响样品间的变异系数的百分等级。在其他实施方案中,掩蔽阈值可被定义为映射质

量得分(例如MapQ得分),其与将序列读段与参考基因组进行比对的可靠性相关联。需注意,掩蔽阈值不同于拷贝数变异(CNV)阈值,后者是表征包含来自疑似具有与CNV相关的医学病症的生物体的核酸的样品的截止值。在一些实施方案中,相对于本文他处所述的归一化染色体值(NCV)或归一化区段值(NSV)来定义CNV阈值。

[0101] 本文中术语“归一化值”是指将所识别的所关注序列(例如染色体或染色体片段)的序列标签数目与所识别的归一化序列(例如归一化染色体或归一化染色体区段)的序列标签数目相关联的数值。例如,“归一化值”可以是如本文他处所述的染色体剂量,或者其可以是NCV,或者其可以是如本文他处所述的NSV。

[0102] 术语“读段”是指从核酸样品的一部分中获得的序列。通常,尽管不是必须的,读段表示样品中邻接碱基对的短序列。读段可由样品部分的碱基对序列(以A、T、C或G)象征性地表示。读段可存储在存储器设备中并视情况进行处理,以确定该读段是匹配参考序列还是满足其他标准。读段可直接从测序装置获得或间接从所存储的关于样本的序列信息获得。在一些情况下,读段为足够长度(例如,至少约25bp)的DNA序列,其可用于鉴定更大的序列或区域,例如,其可被比对并特异性地分配给染色体或基因组区域或基因。

[0103] 术语“基因组读段”用于指个体的全基因组中任何区段的读段。

[0104] 本文中术语“序列标签”可与术语“映射序列标签”互换使用,是指已通过比对特异性地分配(即,映射)到较大序列(例如,参考基因组)的序列读段。映射序列标签唯一地映射到参考基因组,即,它们被分配到参考基因组的单一位置。除非另外指明,否则映射到参考序列上的相同序列上的标记计数一次。可以数据结构或数据的其他集合体提供标签。在某些实施方案中,标签包含读段序列和与该读段相关的信息,诸如序列在基因组中的位置,例如染色体上的位置。在某些实施方案中,位置专门针对正链取向。与参考基因组的比对中,可限定标签允许有限量的错配。在一些实施方案中,分析中可能不包括可映射到参考基因组上的多于一个位置的标签,即,并不唯一映射的标签。

[0105] 术语“非冗余序列标签”是指并不映射到相同位点的序列标签,在一些实施方案中,为了确定归一化染色体值(NCV),对该相同位点进行计数。有时,将多个序列读段与参考基因组上的相同位置进行比对,从而产生冗余的或重复的序列标签。在一些实施方案中,为了确定NCV,将映射到相同位置的重复序列标签省略或计数为一个“非冗余序列标签”。在一些实施方案中,对与非排除位点进行比对的非冗余序列标签进行计数以产生“非排除位点计数”(NES计数)以用于确定NCV。

[0106] 术语“位点”是指参考基因组上的唯一位置(即,染色体ID、染色体位置和取向)。在一些实施方案中,位点可提供残基、序列标签或区段在序列上的位置。

[0107] “排除位点”是存在于已被排除用于对序列标签进行计数的参考基因组的区域中的位点。在一些实施方案中,排除位点存在于含有重复序列的染色体区域(例如着丝粒和端粒)和多于一种染色体所共有的染色体区域(例如存在于Y染色体上也存在于X染色体上的区域)中。

[0108] “非排除位点”(NSE)是参考基因组中未被排除用于对序列标签进行计数的位点。

[0109] “非排除位点计数”(NES计数)是映射到参考基因组上NSE的序列标签数目。在一些实施方案中,NES计数是映射到NES的非冗余序列标签数目。在一些实施方案中,覆盖度和相关参数(诸如归一化覆盖度量)、全局图谱去除的覆盖度量和染色体剂量基于NES计数。在一

个示例中,以所关注染色体的NES计数与归一化染色体的计数的比率来计算染色体剂量。

[0110] 归一化染色体值 (NCV) 将试验样品的覆盖度与训练/合格样品集的覆盖度相关联。在一些实施方案中,NCV基于染色体剂量。在一些实施方案中,NCV将试验样品中所关注染色体的染色体剂量与合格样品集中对应染色体剂量的平均值之间的差异相关联为,并且可如下计算:

$$[0111] \quad NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

[0112] 其中 $\hat{\mu}_j$ 和 $\hat{\sigma}_j$ 分别是合格样品集中第j个染色体剂量的估计平均值和标准偏差,并且 x_{ij} 是观察的试验样品i的第j个染色体比率(剂量)。

[0113] 在一些实施方案中,可通过将试验样品中所关注染色体的染色体剂量与在相同流通池上测序的多重样品中对应染色体剂量的中值相关联来如下“在线(on the fly)”计算NCV:

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

[0115] 其中 M_j 是在相同流通池上进行测序的多重样品集中第j个染色体剂量的估计中值; $\hat{\sigma}_j$ 是在一个或多个流通池上进行测序的一个或多个多重样集品中第j个染色体剂量的标准偏差,并且 x_{ij} 是观察的试验样品i的第j个染色体剂量。在该实施方案中,试验样品i是在确定 M_j 的相同流通池上进行测序的多重样品中的一个样品。

[0116] 例如,对于在一个流通池上以64个多重样品中的一个样品进行测序的试验样品A中的所关注21号染色体,试验样品A中21号染色体的NCV如下计算:样品A中21号染色体的剂量减去在64个多重样品中确定的21号染色体的剂量的中值,除以确定的在流通池1上的或另外流通池的64个多重样品21号染色体的剂量的标准偏差。

[0117] 如本文所用,术语“比对(aligned、alignment或aligning)”是指将读段或标签与参考序列进行比较,从而确定参考序列是否包含该读段序列的过程。如果参考序列包含该读段,则该读段可映射到参考序列,或者在某些实施方案中,映射到参考序列中的特定位置。在一些情况下,比对简单地告知读段是否为特定参考序列的成员(即,该读段是否存在于该参考序列中)。例如,读段与人类13号染色体的参考序列的比对将告知该读段是否存在于13号染色体的参考序列中。提供该信息的工具可被称为集合成员资格测试仪(set membership tester)。在一些情况下,比对另外指示读段或标签映射到的参考序列中的位置。例如,如果参考序列是人类全基因组序列,则比对可指示读段存在于13号染色体上,并且还可指示该读段存在于13号染色体的特定链和/或位点上。

[0118] 比对的读段或标签是就其核酸分子顺序而言识别为与来自参考基因组的已知序列匹配的一条或多条序列。比对可手动进行,尽管其通常通过计算机算法来实现,因为不可能在实现本文所公开的方法的合理时间段内比对读段。来自比对序列的算法的一个示例是作为Illumina Genomics分析管线的一部分分布的高效核苷酸数据局部比对(ELAND) 计算机程序。另选地,布隆(Bloom) 过滤器或类似的集成员资格测试仪可用于将读段与参考基因组进行比对。参见2011年10月27日提交的美国专利申请61/552,374,其全文以引用方式并入本文。比对中序列读段的匹配可以是100%序列匹配或小于100% (非完全匹配)。

[0119] 本文所用的术语“映射”指通过比对将序列读段特异性地分配给较大的序列,例如参考基因组。

[0120] 如本文所用,术语“参考基因组”或“参考序列”是指可用于参考来自受试者的识别序列的任何生物体或病毒的任何特定已知基因组序列,无论是部分的还是完整的。例如,可在ncbi.nlm.nih.gov的美国国家生物技术信息中心(National Center for Biotechnology Information)找到用于人类受试者以及许多其他生物体的参考基因组。“基因组”是指以核酸序列表达的生物体或病毒的完整遗传信息。

[0121] 在各种实施方案中,参考序列显著大于与其比对的读段。例如,参考序列可为比对读段的至少约100倍大、或至少约1000倍大、或至少约10,000倍大、或至少约 10^5 倍大、或至少约 10^6 倍大、或至少约 10^7 倍大。

[0122] 在一个示例中,参考序列是全长人类基因组的序列。此类序列可称为基因组参考序列。在另一个示例中,参考序列限于特定的人类染色体,诸如13号染色体。在一些实施方案中,参考染色体是来自人类基因组版本hg19的Y染色体序列。此类序列可称为染色体参考序列。参考序列的其他示例包括其他物种的基因组,以及任何物种的染色体、亚染色体区域(诸如链)等。

[0123] 在各种实施方案中,参考序列是来源于多个个体的共有序列或其他组合。然而,在某些应用中,参考序列可取自特定个体。

[0124] 本文中术语“临床相关序列”是指已知或疑似与遗传或疾病病症相关联或相牵连的核酸序列。确定是否存在临床相关序列可用于确定或确认医学病症的诊断,或提供用于疾病发展的预后。

[0125] 当用于核酸或核酸混合物的上下文时,本文中术语“衍生的”是指从其来源中获得核酸的方式。例如,在一个实施方案中,衍生自两种不同基因组的核酸混合物意指核酸(例如cfDNA)由细胞通过自然发生的过程(诸如坏死或细胞凋亡)自然释放。在另一个实施方案中,衍生自两种不同基因组的核酸混合物意指从来自受试者的两种不同类型的细胞中提取核酸。

[0126] 当用于获得特定定量值的上下文时,本文中术语“基于”是指使用另一数量作为输入来计算特定定量值作为输出。

[0127] 本文中术语“患者样品”是指从患者,即医学关注、护理或治疗的接受者中获得的生物样品。患者样品可以是本文所述的任何样品。在某些实施方案中,患者样品通过非侵入性手术获得,例如外周血样品或粪便样品。本文所述的方法不需要限于人类。因此,考虑了各种兽医应用,在这种情况下,患者样品可以是来自非人类哺乳动物(例如,猫科动物、猪、马、牛等)的样品。

[0128] 本文中术语“混合样品”是指含有衍生自不同基因组的核酸混合物的样品。

[0129] 本文中术语“母体样品”是指从妊娠受试者(例如,女性)中获得的生物样品。

[0130] 本文中术语“生物流体”是指取自生物来源的液体,并且包括例如血液、血清、血浆、痰、灌洗液、脑脊液、尿液、精液、汗液、泪液、唾液等。如本文所用,术语“血液”、“血浆”和“血清”明确地涵盖其级分或加工部分。类似地,在样品取自活检、拭子、涂片等的情况中,“样品”明确地涵盖来源于活检、拭子、涂片等的处理级分或部分。

[0131] 本文中术语“母体核酸”和“胎儿核酸”分别指妊娠女性受试者的核酸和该妊娠女

性受试者所怀的胎儿的核酸。

[0132] 如本文所用,术语“对应于”有时是指存在于不同受试者的基因组中的核酸序列(例如基因或染色体)不必在所有基因组中具有相同的序列,而是用于提供所关注序列(例如基因或染色体)的同一性而非遗传信息。

[0133] 如本文所用,术语“胎儿分数”是指存在于包含胎儿和母体核酸的样品中的胎儿核酸的分数。胎儿分数通常用于表征母体血液中的cfDNA。

[0134] 如本文所用,术语“染色体”是指活细胞的携带遗传性的基因载体,其来源于包含DNA和蛋白质组分(尤其是组蛋白)的染色质链。本文采用了常规的国际公认的个体人类基因组染色体编号系统。

[0135] 如本文所用,术语“多核苷酸长度”是指序列中或参考基因组的区域中核苷酸的绝对数目。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,在以存在于万维网上|genome|. |ucsc|. |edu/cgi-bin/hgTracks?hgsid=167155613&chromInfoPage=处的人类染色体的NCBI36/hg18组件提供的长度。

[0136] 本文中术语“受试者”是指人类受试者以及非人受试者,该非人受试者诸如哺乳动物、无脊椎动物、脊椎动物、真菌、酵母、细菌和病毒。虽然本文的示例涉及人类并且语言主要针对人类,但本文所公开的概念适用于来自任何植物或动物的基因组,并且可用于兽医学、动物科学、研究实验室等领域。

[0137] 本文中术语“病症”是指作为广义术语的“医学病症”,其包括所有疾病和紊乱,也可包括可能影响人的健康、受益于医学辅助或对医学治疗有影响的损伤和正常健康情况,诸如妊娠。

[0138] 当用于指染色体非整倍体时,本文中术语“完整”是指全染色体的获得或丢失。

[0139] 当用于指染色体非整倍体时,本文中术语“部分”是指染色体的一部分(即区段)的获得或丢失。

[0140] 本文中术语“嵌合性”是指表示在已从单个受精卵发育成的一个个体中存在具有不同染色体核型的两个细胞群。嵌合可由发育期间的突变引起,这种突变仅增殖成成体细胞的亚群。

[0141] 本文术语“非嵌合性”是指生物体(例如人类胎儿)由一种染色体核型的细胞构成。

[0142] 如本文所用,术语“灵敏度”是指当存在所关注条件时测试结果将为阳性的概率。可用真阳性的数目除以真阳性和假阴性之和来计算灵敏度。

[0143] 如本文所用,术语“特异性”是指当不存在所关注条件时测试结果将为阴性的概率。可用真阴性的数目除以真阴性和假阳性之和来计算特异性。

[0144] 本文中术语“富集”是指扩增母体样品的一部分中包含的多态性靶核酸,并将所扩增的产物与去除该部分的该母体样品的其余部分组合的过程。例如,母体样品的其余部分可为原始母体样品。

[0145] 本文中术语“原始母体样品”是指从作为从中去除一部分以扩增多态性靶核酸的来源的妊娠受试者(例如女性)中获得的未富集生物样品。“原始样品”可以是妊娠受试者中获得的任何样品及其处理级分,例如从母体血浆样品中提取的经纯化cfDNA样品。

[0146] 如本文所用,术语“引物”是指当置于诱导延伸产物合成的条件下(例如,该条件包括核苷酸、诱导剂(诸如DNA聚合酶)以及合适的温度和pH)时能够充当合成起始点的分离寡

核苷酸。引物优选为在扩增中效率最高的单链,但也可以是双链。如果是双链,则在用于制备延伸产物之前首先对引物进行处理以分离其链。优选地,引物是寡脱氧核糖核苷酸。引物必须足够长以在存在诱导剂的情况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、引物来源、方法的用途和用于引物设计的参数。

[0147] 引言及背景

[0148] 人类基因组中的CNV显著影响人类对疾病的多样性和易感性(Redon等人,Nature第23章:第444-454页[2006年],Shaiho等人,Genome Res第19章:第1682-1690页[2009年])。此类疾病包括但不限于癌症、感染性和自身免疫疾病、神经系统疾病、代谢和/或心血管疾病等。

[0149] 已知CNV通过不同的机制导致遗传疾病,在大多数情况下导致基因剂量失衡或基因破坏。已知CNV除了与遗传紊乱直接相关之外,还介导可能有害的表型改变。最近,几项研究报道了与正常对照相比,在复杂疾病诸如自闭症、注意力缺陷多动障碍(ADHD)和精神分裂症中罕见或从头CNV的负担增加,表明了罕见或独特CNV的潜在致病性(Sebat等人,第316章:第445-449页[2007年];Walsh等人,第320章:第539-543页[2008年])。CNV产生于主要由于缺失、复制、插入和不平衡易位事件的基因组重排。

[0150] 研究已表明,胎儿来源的cfDNA片段平均长度短于母体来源的cfDNA片段。基于NGS数据的非侵入性产前检测(Non-invasive prenatal testing,NIPT)已成功实现。目前的方法包括使用短读段(25bp-36bp)对母体样品进行测序,与基因组比对,计算并归一化亚染色体覆盖度,最后评估与正常二倍体基因组相关联的预期归一化覆盖度相比靶染色体的过度表示(13/18/21/X/Y)。因此,传统的NIPT测定和分析依赖于计数或覆盖度来评估胎儿非整倍体的可能性。

[0151] 由于母体血浆样品表示母体和胎儿cfDNA的混合物,因此任何给定NIPT方法的成功取决于其对检测低胎儿分数样品中拷贝数变异的灵敏度。对于基于计数的方法,它们的灵敏度取决于(a)测序深度和(b)数据归一化减小技术差异的能力。本公开提供了用于NIPT和其他应用的分析方法,该方法通过从例如配对端读段中推导片段大小信息,并且在分析管线中使用该信息。提高的分析灵敏度提供了在覆盖度降低(例如,测序深度降低)时应用NIPT方法的能力,这使得该技术能够用于对平均风险妊娠的低成本检测。

[0152] 在一些实施方案中,提供了使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。

[0153] 本公开的方面涉及可识别能够可靠地作出拷贝数变异判定的样品测量值以及不能可靠地作出此类判定的样品测量值的方法和系统。换言之,本发明所公开的实施方案能够区分具有足够信息以进行可信判定的样品测量值与不具有足够信息的样品测量值。

[0154] 在用于区分可靠和不可靠样品测量值的一些现有方法中,将胎儿分数用作确定是否排除样品被判定的度量。此类方法可针对NIPT样品以特定方式运行测序例程。所得样品具有基因组或染色体的非排除区域的覆盖度或读段数目值。此类方法还可确定胎儿分数值。如果所确定的胎儿分数值低于某个阈值,例如3%,则排除样品测量值。在此类图的底部存在样品排除区域,该样品排除区域覆盖有效读段计数相当低和/或所确定的胎儿分数值低的样品测量值。在一些具体实施中,排除区域包括对于特别低的胎儿分数,将即使数目相对较高的有效读段排除在外的步骤。

[0155] 如下文某些附图所示,数据示出了所确定的胎儿分数值可能由于各种原因而不准确的情况的各种示例。该数据示出了胎儿分数误差的某些趋势,诸如胎儿分数误差随判定误差的增加(其可由覆盖度降低引起)而增加,以及胎儿分数误差随真实胎儿分数值的量级减小而增加。由于以胎儿分数作为用于确定要排除哪些样品的模式存在这些问题,因此应使用用于确定要排除哪些样品的另一种技术。

[0156] 胎儿分数影响NIPT中胎儿游离DNA丰度和拷贝数变异的确定。当样品中的胎儿分数降低时,由于随着胎儿分数的降低,信号降低并且相对噪声增加,因此准确确定胎儿游离DNA的相对覆盖度变得更加困难。覆盖度或测序深度也对CNV检测具有类似的影响。由于这两个因素共同影响CNV检测,因此确保检测所需的最低胎儿分数根据覆盖度而改变,具有如图1所示的双曲线形状。

[0157] 现有方法通过排除具有低覆盖度的样品来执行质量控制。例如,一些现有方法根据胎儿分数值来设定两个覆盖度阈值水平。在图2所示的示例中,第一覆盖度阈值设定为小于约5%的胎儿分数的约4.5M非排除位点NES或有效读段(线2002),并且第二覆盖度阈值设定为大于约5%的胎儿分数的约2M(线2006)。线或曲线(2004)连接两个阈值水平。

[0158] 然而,此类方法具有各种局限性。就此类方法对CNV检测的影响而言,此类方法不能充分捕获覆盖度和胎儿分数之间的关系。此外,在这些现有方法中使用的阈值是静态的,相同的阈值应用于不同的样品、群体和平台。如图3所示,不同的样品群体具有不同的胎儿分数分布。图3示出了三个群体或其样品的胎儿分数分布。左图示出了三个群体的概率密度函数。右图示出了三个群体的累积分布。累积分布清楚地示出了0.04的固定胎儿分数阈值排除了来自三个群体的样品的不同部分。

[0159] 本公开提供了CNV检测过程中的质量控制(QC)的方法和系统,诸如图4A、图4B、图15和图16A所示。QC可在进行CNV判定之前或之后执行,以识别胎儿分数水平太低而不能产生可靠结果的样品。可重新运行所识别的样品以获得新的读段。如果重新运行中的测序深度增加,则覆盖度增加,这提高了样品的信号或降低了样品的噪声。当应用于不同样品群体时,本发明所公开的具体实施可考虑样品和群体之间的变异。并且这些具体实施可更有效地识别具有低胎儿分数和/或低读段覆盖度的样品。

[0160] 示例性工作流程

[0161] 除非另外指明,否则本公开中的工作流程和标记的步骤可以与图和示例中所述的顺序不同的顺序来执行。例如,框204所示的确定胎儿分数值的操作可在202、206、208和210中所示的操作之前或之后执行。图4A示出了根据一些具体实施的使用检测限(LOD) QC方法来确定CNV的工作流程200。在该工作流程中,在框212处进行质量检查以检查试验样品是否在由至少胎儿分数LOD曲线限定的排除区域内。在许多具体实施中,排除区域由LOD曲线和覆盖度阈值或胎儿分数阈值限定。该步骤及其下游步骤可应用于下文所述的其他CNV检测过程。该示例性工作流程包括对包含母体游离核酸片段和胎儿游离核酸片段的试验样品进行测序,以获得序列读段。参见框202。可使用各种技术来获得序列读段,包括但不限于下文所述的测序技术。

[0162] 然后过程200继续确定试验样品的胎儿分数值。在一些具体实施中,将从202获得的序列读段用于确定胎儿分数。然而,也可使用来自相同个体的其他核酸。在一些具体实施中,可使用各种技术来确定试验样品的胎儿分数,包括但不限于下文所述的技术。简而言

之,在一些具体实施中,基于游离核酸片段的大小来确定试验样品的胎儿分数值。在一些具体实施中,通过获得游离核酸片段大小的频率分布,以及将该频率分布应用于将胎儿分数与片段大小的频率相关联的模型,以获得胎儿分数值。在一些具体实施中,基于参考基因组的分组的覆盖度信息来确定胎儿分数值,其中将参考基因组分成区段或分组。在一些具体实施中,通过以下方式来计算胎儿分数值:将参考基因组的多个分组的覆盖度值应用于将胎儿分数与分组的覆盖度相关联的模型,以获得胎儿分数值。在一些具体实施中,从训练样品中选择过度表示的胎儿游离核酸片段所比对的分组。与分组进行比对的测试读段用于确定胎儿分数。在一些具体实施中,基于性染色体(诸如从怀有男性胎儿的母体中获得的母体样品中的Y染色体)的分组的覆盖度信息来确定试验样品的胎儿分数值。

[0163] 过程200还包括接收通过对试验样品中的游离核酸片段进行测序而获得的序列读段。参见框206。然后将游离核酸片段的序列读段与包括所关注序列的参考基因组进行比对,从而提供序列标签。参见框208。

[0164] 过程200继续确定所关注序列的序列标签的覆盖度。参见210。确定每个样品的覆盖度。在各种具体实施中,在所有染色体(全参考基因组)上、在染色体的子集上或在亚染色体水平上确定覆盖度。可使用各种技术来确定覆盖度,包括但不限于下文所示的那些技术。在一些具体实施中,通过以下方式确定覆盖度:将参考基因组分成多个分组,确定与每个分组进行比对的序列标签数目,以及使用所关注序列的分组中的所述序列标签数目来确定所述序列标签的覆盖度。在一些具体实施中,方法还包括通过考虑由于拷贝数变异之外的因素引起的分组间变异,来调整与分组进行比对的序列标签数目。下文提供了用于确定覆盖度的方法的更详细描述。

[0165] 过程200中的下一个步骤是QC步骤。QC步骤包括基于208中确定的序列标签的覆盖度和204中确定的胎儿分数来确定试验样品是否在排除区域内。该排除区域由至少胎儿分数LOD曲线限定。该胎儿分数LOD曲线随覆盖度值而变化,并且指示在给定不同覆盖度的情况下实现检测标准所需的胎儿分数的最小值。LOD曲线的示例参见图14。

[0166] 在一些具体实施中,使用受CNV影响的训练样品来获得胎儿分数LOD曲线。受影响样品可包括通过将来自不同个体的序列读段组合而获得的计算机模拟样品。在一些具体实施中,通过将具有两个或更多个胎儿分数的物理样品组合来获得受影响样品,从而提供具有中间胎儿分数值的合成样品。在一些具体实施中,使用关于CNV的受影响样品和未受影响样品来获得胎儿分数LOD曲线。在一些具体实施中,可使用物理或模拟样品来推导LOD曲线。

[0167] 在一些具体实施中,检测标准是在给定观察胎儿分数的情况下,基准真值胎儿分数大于指定LOD的期望置信水平。LOD是可以限定置信度检测的最小水平的信号(分析物、胎儿分数、指示条件的得分等)。在本申请的上下文中,LOD是以限定置信度检测非整倍体/CNV所需的胎儿分数(或其他分析物)的最低水平。需注意,包括两个置信度值:第一置信度是大于指定LOD胎儿分数的基准真值胎儿分数的置信度,并且第二置信度是LOD本身的置信度。在一些具体实施中,检测标准是对于观察胎儿分数而言基准真值胎儿分数大于LOD Y%的X%置信度。基准真值胎儿分数是以实际胎儿分数为基础的推断胎儿分数。在一些具体实施中,X为约50%、55%、60%、65%、70%、75%、80%、85%、90%、95%、99%或99.5%。在一些具体实施中,Y为约50%、55%、60%、65%、70%、75%、80%、85%、90%、95%或99%的检测置信度。在一些具体实施中,X为80%并且Y为95%。换句话讲,检测标准是对于观察胎儿分

数而言基准真值胎儿分数大于LOD95% (或简称为LOD95) 的80%置信度, LOD95是可在95%的时间检测到CNV或非整倍体的最小胎儿分数值。在其他具体实施中, X为50%并且Y为95%。

[0168] 在一些具体实施中, 指定LOD被确定为可检测到Y%的受影响样品的最小观察胎儿分数。在一些具体实施中, 使用在观察覆盖度处观察胎儿分数的真实胎儿分数(或模拟胎儿分数)的分布来获得针对该观察胎儿分数和该观察覆盖度的检测标准。图5示出了例示以统计概念为基础的LOD的经验和假设数据。CNV检测基于指示样品包含CNV的可能性的对数似然比。上图示出了Y轴上的LLR和X轴上的胎儿分数。在每个胎儿分数水平处测量多个样品, 并且可获得平均值和标准偏差。给定样品数据, 人们可推断每个水平的胎儿分数的LLR的群体分布。上图示出了应用截止值(标记为502A)来判定CNV。随着胎儿分数值增加, LLR得分也增加, 进一步远离截止值, 从而允许检测到更多样品。

[0169] 在该示例中, 2.3%是可以95%置信度(单边)检测到非整倍体/CNV的最低胎儿分数。在胎儿分数2.3%处受影响样品的观察LLR在上图中示出为504A, 并且在下图中示出为504B。观察数据的基础群体分布在下图中示出为506。截止LLR在上图中示出为502A, 并且在下图中示出为502B。下图示出了在胎儿分数2.3%处, 5%的基础群体低于LLR截止值, 并且95%的群体高于LLR阈值502B。因此, 群体中95%的样品被检测为具有CNV。

[0170] 如在图5的上图中可见, 随着胎儿分数的增加, LLR和检测概率也增加。图6示出了根据胎儿分数的检测概率。在该示例中, 2.3%胎儿分数的检测概率为95%。因此, 此处胎儿分数LOD95为2.3%。

[0171] 返回图4A, 如果确定试验样品不在由胎儿分数LOD曲线限定的排除区域内, 则该方法继续评估所关注序列的覆盖度以确定CNV。参见从框212到框214的“否”分支。并且在该路径中结束该过程。如果确定试验样品在LOD排除区域内, 并且确定应对样品进行重新测序, 则参见框212的“是”分支和框216的“是”分支。然后对样品进行重新测序, 并且重复从202到212的过程。如果在框216处确定序列不需要进行重新测序, 诸如当其已经进行重新测序时, 则过程结束。

[0172] 如下文中实验数据所示, 所公开的QC方法可有助于提高灵敏度和特异性两者。如果人们希望仅改善灵敏度, 则可首先确定试验样品为阴性还是阳性。只有判定为阴性的样品才能通过如图4B所示的QC过程。

[0173] 图4B示出了用于CNV检测的LOD QC过程, 其类似于图4A中的过程, 不同的是在确定样品是否落入排除区域之前首先执行判定。参见框211。如果判定为阴性, 则该过程之后确定样品是否在由LOD曲线限定的排除区域内。参见框212。如果不是, 则该过程结束。参见框216的“否”分支。如果试验样品在排除区域内(框212的“是”分支)并且确定需要进行重新测序, 则将对样品进行重新测序并重复该过程。如果确定样品不需要进行重新测序, 诸如在其已经进行一次重新测序的情况下, 则过程结束。参见框216的“否”分支。

[0174] LOD和LOD曲线

[0175] 图7示出了估计胎儿分数包括导致所述估计胎儿分数偏离真实胎儿分数的误差。实线示出了观察胎儿分数分布。虚线示出真实胎儿分数分布。观察胎儿分数是根据其他变量推断的, 而不是直接根据胎儿DNA片段和母体DNA片段测量的。灰线是直接根据胎儿Y染色体测量的胎儿分数的分布。由于灰线是直接测量的, 因此它比观察胎儿分数更接近真实胎

几分数。

[0176] 图8示出了由于覆盖度和由于胎儿分数值引起的胎儿分数误差。左图示出了胎儿分数估计值的标准偏差(与误差相关的量度)随覆盖度的增加而减小。右图示出了胎儿分数估计值的标准偏差随真实胎儿分数的增加而减小。

[0177] 图9模拟了八个水平的观察胎儿分数的真实胎儿分数。图9示出了在0%和8%之间以1%为间隔的观察胎儿分数。图9还示出了这些观察胎儿分数的基准真值分布。竖直虚线9002示出了观察胎儿分数0%。实线9004是0%处观察胎儿分数的真实胎儿分数的分布。虚线9006指示8%的观察胎儿分数。实线9008示出了8%的观察胎儿分数的真实胎儿分数的分布。如图所示,当观察胎儿分数较低时,诸如在线9002和0%处,基准真值胎儿分数分布9004偏离观察胎儿分数的距离比在8%处的真实胎儿分数(9008)偏离观察8%胎儿分数(9006)的距离更远。另外,较低胎儿分数分布的尖峰值往往较大。换句话讲,随着观察胎儿分数的增加,真实胎儿分数的分布变得更平坦,并且与观察胎儿分数的偏差更小。

[0178] 图10示出了三个不同水平的误差或覆盖度(注意,覆盖度与误差成反比)的真实胎儿分数相对于观察胎儿分数的分布。图10的左图与图9相同。图10的左图以1%胎儿分数误差来模拟。中间图示出了以1.5%误差模拟的胎儿分数的分布。右图示出了以2%误差模拟的胎儿分数的分布。从图中可以看出,观察胎儿分数和真实胎儿分数之间的差异受覆盖度影响,误差越大,观察胎儿分数和真实胎儿分数之间的差异越大。

[0179] 图11示出了在给定不同误差或覆盖度的情况下2%的观察胎儿分数及其模拟真实胎儿分数分布。观察胎儿分数由线1102指示。以最小误差1%(或最高覆盖度)模拟的真实胎儿分数具有分布1112。分布1114是1.5%误差或中间覆盖度模拟的真实胎儿分数。分布1116是以最高2%误差或最低覆盖度模拟的真实胎儿分数分布。分布1112、1114、1116的第5百分位或95%置信度分别由线1122、1124和1126标记。这些分布表明,随着误差增加或覆盖度减小,真实胎儿分数进一步偏离观察胎儿分数,并且95%置信水平也增加。图11中还示出了分布1116的第20百分位(1128)和第25百分位(1130)。它们分别对应于80%置信水平和75%置信水平。类似地,可确定其他置信水平,诸如50%置信水平。

[0180] 图12示出了根据观察胎儿分数的第20百分位(左图)和第25百分位(右图)的真实胎儿分数分布。随着观察胎儿分数的增加,真实胎儿分数的第20百分位也相应地增加。左图示出了80%置信度数据,即真实胎儿分数分布的第20百分位。右图示出了75%置信度数据。在左图中,第20百分位真实胎儿分数与观察胎儿分数之间的关系示出为不同的线,每条线指示不同的覆盖度水平(和对应的FF误差水平)。左图示出了十个不同的覆盖度水平(1M-10M)。该图示出了三个变量之间的关系:观察胎儿分数、第20百分位(或80%置信度)处的真实胎儿分数和覆盖度。观察例如左图,可以看到2%观察FF(对应于图11中观察到的FF)处的两种模式,(a)真实FF的第20百分位(或80%置信度)高于2%,以及(b)随着覆盖度减小(或误差增加),真实FF的第20百分位(或80%置信度)增加。这两种模式也可见于图11中。

[0181] 在观察胎儿分数的低端,随着覆盖度减小,真实胎儿分数的第20百分位增加。在观察胎儿分数的高端,这种关系发生了逆转。为了计算LOD曲线,首先根据经验确定特定覆盖度的LOD95为某值(为胎儿分数)。该LOD95值也是期望的(例如,20百分位)真实胎儿分数的值。利用该值,人们可根据真实函数对比观察函数(图12的左图中的线)确定观察胎儿分数。

[0182] 例如,根据经验确定覆盖度1百万的LOD95为6.50% FF,这也是1202处的真实胎儿

分数的第20百分位。人们可根据图确定观察胎儿分数为8% (如1204所示)。根据经验确定覆盖度2百万的LOD95为4.59% FF,这也是1206处的真实胎儿分数的第20百分位。人们可根据图确定观察胎儿分数为4.2% (如1208所示)。人们可类似地从图中的其他线获得其他覆盖度的观察胎儿分数。这些点是具有真实胎儿分数超过不同覆盖度的LOD95的80%置信度所需的观察胎儿分数。

[0183] 图13示出了LOD、覆盖度和观察胎儿分数。该表在从左侧起的第一列中示出了有效读段计数,在第三列中示出了LOD95,并且在第四列中示出了达到真实胎儿分数高于LOD95的80%置信度所需的观察胎儿分数。具有75%置信度所需的观察胎儿分数在最右列中示出,其可从图12的右图所示的数据中获得。第二列示出了不同有效读段计数的FF误差,表明随着覆盖度增加,误差减小。

[0184] 图14包括可从类似于图13所示的数据中获得的两条胎儿分数LOD曲线。在一些具体实施中,LOD曲线与覆盖度阈值(诸如图14所示的1M覆盖度阈值)结合使用。

[0185] 在一些具体实施中,如图14所示,排除区域在胎儿分数LOD曲线的下方。从数值上看,排除区域是每个点具有比胎儿分数LOD曲线上的对应点更低的胎儿分数或覆盖度值的区域。在一些具体实施中,排除区域由胎儿分数LOD曲线和覆盖度阈值限定。排除区域在胎儿分数LOD曲线和覆盖度阈值两者的下方。

[0186] 在一些具体实施中,生成LOD曲线的方法可总结如下。使用模拟“样品”获得LOD曲线,其中每个“样品”包括以下项。

- [0187] • 从临床样品中已知的覆盖度值的正态分布中采样的覆盖度值
- [0188] • 根据临床样品中的观察FF分布确定的从已知基础真实FF分布采样的基础真实FF值
- [0189] • 通过将误差(对应于覆盖度值)添加到基础真实FF值来确定的观察FF值

[0190] 模拟了大量此类“样品”。对于每个观察胎儿分数值(以小数增量计),以及在每个覆盖度水平(以增量计1-10M)处,收集该数据集中的所有对应真实FF值,并且选择这些真实FF值的100减去X百分位数。对于每个可能的观察FF值和覆盖度组合,现在存在单个真实FF值,我们可以说在给定该观察FF值的情况下,我们对真实FF至少与所选择的真实FF一样高的置信度为X%。

[0191] 对于每个覆盖度水平,将观察FF相对于所选择的真实FF作图,以得到每个覆盖度水平具有一条曲线的这些图。

[0192] 除此之外,LOD研究诸如实施例1的LOD研究描述了覆盖度和LOD95之间的关系(该关系的推导包括受影响样品)。然后将来自该图的对应于覆盖度水平1-10M(以增量计)的LOD95(Y)值叠加在上述观察FF对比真实FF的图上。这些水平LOD95(Y)线中的每一条线与观察FF对比真实FF图上的对应(按覆盖度)曲线相交的点从x轴上读段,以得到确保真实FF超过LOD95(Y)的X%置信度所需的观察FF。

[0193] 然后相对于获得LOD QC曲线所需的观察FF绘制覆盖度值。

[0194] 评估CNV

[0195] 用于确定CNV的方法

[0196] 使用由本文所公开的方法提供的序列覆盖度值、片段大小参数和/或甲基化水平,人们可以相对于通过常规方法获得的序列覆盖度具有提高的灵敏度、选择性和/或效率来

确定与序列、染色体或染色体区段的拷贝数和CNV相关的各种遗传条件。例如,在一些实施方案中,掩蔽参考序列用于确定在包含胎儿和母体核酸分子的母体试验样品中是否存在任意两种或更多种不同的完整胎儿染色体非整倍体。下文提供的示例性方法将读段与参考序列(包括参考基因组)进行比对。可在未掩蔽或掩蔽参考序列上执行比对,从而产生映射到该参考序列的序列标签。在一些实施方案中,仅将落入参考序列的未掩蔽区段上的序列标签考虑在内以确定拷贝数变异。

[0197] 在一些实施方案中,评估核酸样品的CNV包括通过以下三种类型的判定之一来表征染色体或区段非整倍体的状态:“正常”或“未受影响”、“受影响”和“无判定”。通常设定用于判定正常和受影响的阈值。在样品中测量与非整倍体或其他拷贝数变异相关的参数,并将测量值与阈值进行比较。对于复制型非整倍体,如果染色体或区段剂量(或其他测量值序列含量)高于用于受影响样品的限定阈值集,则作出受影响的判定。对于此类非整倍体,如果染色体或区段剂量低于用于正常样品的阈值设定,则作出正常的判定。相比之下,对于缺失型非整倍体,如果染色体或区段剂量低于用于受影响样品的限定阈值,则判定受影响,并且如果染色体或区段剂量高于针对正常样品的阈值设定,则作出正常的判定。例如,在存在三倍体的情况下,通过低于用户定义的可靠性阈值的参数的值(例如,试验染色体剂量)来确定“正常”判定,并且通过高于用户定义的可靠性阈值的参数(例如,试验染色体剂量)来确定“受影响”判定。通过位于进行“正常”或“受影响”判定的阈值之间的参数(例如,试验染色体剂量)来确定“无判定”结果。术语“无判定”可与“未分类”互换使用。

[0198] 可用于确定CNV的参数包括但不限于覆盖度、片段大小偏倚/加权覆盖度、限定大小范围内片段的分数或比率和片段的甲基化水平。如本文所述,从与参考基因组的区域进行比对的读段的计数中获得覆盖度,并且任选地将该覆盖度归一化以产生序列标签计数。在一些实施方案中,可通过片段大小来对序列标签计数进行加权。

[0199] 在一些实施方案中,片段大小参数偏向基因组中的一个基因组的片段大小特性。片段大小参数是与片段大小相关的参数。在以下情况时,参数偏向片段大小:1)对片段大小参数进行有利地加权,例如,该大小的计数的权重比其他大小的计数的权重更大;或者2)从对片段大小有利地进行加权的值中获得参数,例如,从该大小的权重更大的计数中获得的比率。当基因组具有相对于另一基因组或同一基因组的另一部分富集或更高浓度的核酸时,该大小是该基因组的特性。

[0200] 在一些实施方案中,用于确定在母体试验样品中是否存在任何完整胎儿染色体非整倍体的方法包括(a)获得母体试验样品中胎儿和母体核酸的序列信息;(b)使用上述序列信息和方法来识别选自1-22号染色体、X染色体和Y染色体的所关注染色体中每一条染色体的序列标签数目、序列覆盖度量、片段大小参数或另一个参数,以及识别一条或多条归一化染色体序列的序列标签数目或另一个参数;(c)使用所识别的所述所关注染色体中每一条染色体的所述序列标签数目或所述其他参数,以及所识别的所述归一化染色体中每一条染色体的所述序列标签数目或所述其他参数,来计算所述所关注染色体中每一条序列的单个染色体剂量;以及(d)将每个染色体剂量与阈值进行比较,并由此确定在母体试验样品中是否存在任何完整胎儿染色体非整倍体。

[0201] 在一些实施方案中,上述步骤(a)可包括对试验样品的核酸分子的至少一部分进行测序,以获得该试验样品的胎儿和母体核酸分子的所述序列信息。在一些实施方案中,步

骤(c)包括以所识别的所关注染色体中每一条染色体的序列标签数目或其他参数与所识别的归一化染色体序列的序列标签数目或其他参数的比率,来计算所述所关注染色体中每一条染色体的单个染色体剂量。在一些其他实施方案中,染色体剂量基于来源于序列标签数目的处理序列覆盖度量或另一个参数。在一些实施方案中,仅使用唯一的、非冗余序列标签来计算处理序列覆盖度量或另一个参数。在一些实施方案中,处理序列覆盖度量是序列标签密度比,其是通过序列长度进行标准化的序列标签数目。在一些实施方案中,处理序列覆盖度量或其他参数是归一化序列标签或另一个归一化参数,其是由基因组的全部或大部分的序列标签数目或其他参数来除的所关注序列的序列标签数目或其他参数。在一些实施方案中,根据所关注序列的全局图谱来调整处理序列覆盖度量或其他参数(诸如片段大小参数)。在一些实施方案中,根据试验样品的GC含量与序列覆盖度之间的样品内相关性来调整处理序列覆盖度量或其他参数。在一些实施方案中,处理序列覆盖度量或其他参数由这些过程的组合产生,这些过程在本文他处进一步描述。

[0202] 在一些实施方案中,以所关注染色体中每一条染色体的处理序列覆盖度或其他参数与归一化染色体序列的处理序列覆盖度或其他参数比率来计算染色体剂量。

[0203] 在上述实施方案的任意一个实施方案中,完整染色体非整倍体选自完整染色体三倍体、完整染色体单体和完整染色体多体。完整染色体非整倍体选自1-22号染色体、X染色体和Y染色体中任意一条染色体的完整非整倍体。例如,所述不同的完整胎儿染色体非整倍体选自2-三倍体、8-三倍体、9-三倍体、20-三倍体、21-三倍体、13-三倍体、16-三倍体、18-三倍体、22-三倍体、47,XXX、47,XYY和X单倍体。

[0204] 在上述实施方案的任意一个实施方案中,对来自不同母体受试者的试验样品重复步骤(a)至(d),并且该方法包括确定在试验样品的每个试验样品中是否存在任意两种或更多种不同的完整胎染色体非整倍体。

[0205] 在上述实施方案的任意一个实施方案中,方法还可包括计算归一化染色体值(NCV),其中该NCV将染色体剂量与合格样品集中对应染色体剂量的平均值相关联为:

$$[0206] \quad NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

[0207] 其中 $\hat{\mu}_j$ 和 $\hat{\sigma}_j$ 分别是合格样品集中第j个染色体剂量的估计平均值和标准偏差,并且 $x_{i,j}$ 是观察的试验样品i的第j个染色体剂量。

[0208] 在一些实施方案中,可通过将试验样品中所关注染色体的染色体剂量与在相同流通池上测序的多重样品中对应染色体剂量的中值相关联来如下“在线(on the fly)”计算NCV:

$$[0209] \quad NCV_{ij} = \frac{x_{ij} - M_j}{\hat{\sigma}_j}$$

[0210] 其中 M_j 是在相同流通池上进行测序的多重样品集中第j个染色体剂量的估计中值; $\hat{\sigma}_j$ 是在一个或多个流通池上进行测序的一个或多个多重样集品中第j个染色体剂量的标准偏差,并且 x_i 是观察的试验样品i的第j个染色体剂量。在该实施方案中,试验样品i是在确定 M_j 的相同流通池上进行测序的多重样品之一。

[0211] 在一些实施方案中,提供了用于确定在包含胎儿和母体核酸的母体试验样品中是

否存在不同的部分胎儿染色体非整倍体的方法。该方法包括与上文概述的用于检测完整非整倍体的方法类似的程序。然而,不是分析完整染色体,而是分析染色体的区段。参见美国专利申请公布2013/0029852,其以引用方式并入。

[0212] 图15示出了根据一些实施方案的用于确定存在拷贝数变异的方法。图15所示的过程100使用基于序列标签数目(即序列标签计数)的序列标签覆盖度来确定CNV。然而,类似于上文关于计算NCV的描述,可使用其他变量或参数(诸如大小、大小比率和甲基化水平)来代替覆盖度。在一些具体实施中,将两个或更多个变量组合以确定CNV。此外,可基于标签所来源的片段的大小来对覆盖度和其他参数进行加权。为了便于阅读,在图1所示的过程100中仅提及覆盖度,但应当注意,可使用其他参数(诸如大小、大小比率和甲基化水平、由大小加权的计数等)来代替覆盖度。

[0213] 在操作130和135中,确定合格序列标签覆盖度(或另一个参数的值)和测试序列标签覆盖度(或另一个参数的值)。本公开提供了相对于常规方法提供提高的灵敏度和选择性的用于确定覆盖度量的过程。操作130和135由星号标记并以粗线框加粗以指示这些操作有助于改进现有技术。在一些实施方案中,将序列标签覆盖度量进行归一化、调整、修整和以其他方式处理以提高分析的灵敏度和选择性。这些过程在本文他处进一步描述。

[0214] 从概述上来看,方法利用合格训练样品的归一化序列进行试验样品CNV的确定。在一些实施方案中,合格训练样品不受影响并且具有正常的拷贝数。归一化序列提供了将运行内和运行间变异性的测量值归一化的机制。使用来自合格样品集的序列信息来识别归一化序列,该合格样品集从已知包含具有任意一条所关注序列(如染色体或其区段)的正常拷贝数的细胞的受试者中获得。图1所示方法的实施方案的步骤110、120、130、145和146概述了归一化序列的确定。在一些实施方案中,归一化序列用于计算测试序列的序列剂量。参见步骤150。在一些实施方案中,归一化序列还用于计算测试序列的序列剂量所比较的阈值。参见步骤150。从归一化序列和测试序列中获得的序列信息用于确定试验样品中有统计意义的染色体非整倍体识别(步骤160)。

[0215] 转到根据一些实施方案的用于确定存在拷贝数变异的方法的细节,图15提供了用于确定生物样品中的所关注序列(例如染色体或其区段)的CNV的实施方案的流程图100。在一些实施方案中,生物样品从受试者中获得并且包含由不同基因组贡献的核酸的混合物。可由两个个体将不同的基因组贡献给样品,例如,由胎儿和怀有胎儿的母体贡献不同的基因组。另外,可由三个或更多个个体将不同的基因组贡献给样品,例如,由两个或更多个胎儿和怀有胎儿的母体贡献不同的基因组。另选地,由来自同一受试者的非整倍体癌细胞和正常整倍体细胞(例如,来自癌症患者的血浆样品)将基因组贡献给样品。

[0216] 除了分析患者的试验样品之外,还选择每个可能的所关注染色体的一条或多条归一化染色体或一个或多个归一化染色体区段。归一化染色体或区段的识别与患者样品的正常测试异步,这可在临床环境中进行。换句话讲,在测试患者样品之前识别归一化染色体或区段。将归一化染色体或区段与所关注染色体或区段之间的关联存储起来以供在测试期间使用。如下所解释的,这种关联通常在跨越许多样品的测试的时间段内得以维持。以下讨论涉及用于选择所关注单条染色体或区段的归一化染色体或染色体区段的实施方案。

[0217] 获得合格样品集以识别合格归一化序列并且提供方差值以用于确定试验样品中有统计意义的CNV识别。在步骤110中,从已知包含具有任意一条所关注序列的正常拷贝数

的细胞的多个受试者中获得多个生物合格样品。在一个实施方案中,从怀有已使用细胞遗传学方法确认具有正常拷贝数的染色体的胎儿的母体中获得合格样品。生物合格样品可以是生物流体(例如血浆)或如下所述的任何合适的样品。在一些实施方案中,合格样品含有核酸分子(例如cfDNA分子)的混合物。在一些实施方案中,合格样品是含有胎儿和母体cfDNA分子混合物的母体血浆样品。通过使用任何已知的测序方法对核酸(例如,胎儿核酸和母体核酸)的至少一部分进行测序来获得归一化染色体和/或其区段的序列信息。优选地,本文他处所述的下一代测序(NGS)方法中的任意一种方法用于将胎儿和母体核酸以单个或克隆扩增分子进行测序。在各种实施方案中,在测序之前和测序期间如下所公开的那样处理合格样品。可使用如本文所公开的装置、系统和试剂盒来处理合格样品。

[0218] 在步骤120中,对合格样品中包含的所有合格核酸中的每一个核酸的至少一部分进行测序,以生成与参考基因组(例如hg18)进行比对的数百万个序列读段(例如36bp读段)。在一些实施方案中,序列读段包含约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130bp、约140bp、约150bp、约200bp、约250bp、约300bp、约350bp、约400bp、约450bp或约500bp。预期的是,技术进步将使单末端读段能够大于500bp,从而在生成配对末端读段时,使读段能够大于约1000bp。在一个实施方案中,映射序列读段包含36bp。在另一个实施方案中,映射序列读段包含25bp。

[0219] 将序列读段与参考基因组进行比对,并且将唯一地映射到该参考基因组的读段称为序列标签。不对落入掩蔽参考序列的掩蔽区段上的序列标签进行计数用于CNV分析。

[0220] 在一个实施方案中,从唯一地映射到参考基因组的读段中获得包含20bp至40bp的读段的至少约 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。

[0221] 在步骤130中,对通过对合格样品中的核酸进行测序而获得的所有标签进行计数,以获得合格序列标签覆盖度。类似地,在操作135中,对从试验样品中获得的所有标签进行计数以获得测试序列标签覆盖度。本公开提供了相对于常规方法提供提高的灵敏度和选择性的用于确定覆盖度量的过程。操作130和135由星号标记并以粗线框加粗以指示这些操作有助于改进现有技术。在一些实施方案中,将序列标签覆盖度量进行归一化、调整、修整和以其他方式处理以提高分析的灵敏度和选择性。这些过程在本文他处进一步描述。

[0222] 当在合格样品中的每个合格样品中对所有合格序列标签进行映射和计数时,确定合格样品中所关注序列(例如,临床相关序列)的序列标签覆盖度,以及随后从中识别归一化序列的另外序列的序列标签覆盖度。

[0223] 在一些实施方案中,所关注序列是与完整染色体非整倍体相关联的染色体(例如,21号染色体),并且合格归一化序列是与染色体非整倍体不相关,并且序列标签覆盖度变异近似于所关注序列(即染色体)(例如,21号染色体)的序列标签覆盖度变异的完整染色体。所选择的归一化染色体可以是最近近似于所关注序列的序列标签覆盖度变异的一条染色体或一组染色体。1-22号染色体、X染色体和Y染色体中的任意一条或多条染色体可以是所关注序列,并且可将一条或多条染色体识别为合格样品中任意一条1-22号染色体、X染色体和

Y染色体中的每一条染色体的归一化序列。如本文他处所述,归一化染色体可以是单条染色体或其可为一组染色体。

[0224] 在另一个实施方案中,所关注序列是与部分非整倍体相关联的染色体区段(例如染色体缺失或插入、或失衡的染色体易位),并且归一化序列是与部分非整倍体不相关,并且序列标签覆盖度变异近似于与部分非整倍体相关联的染色体片段的序列标签覆盖度变异的一个染色体区段(或一组染色体区段)。所选择的归一化染色体区段可以是最近似于所关注序列的序列标签覆盖度变异的一个或多个染色体区段。任意一条或多条1-22号染色体、X染色体和Y染色体中的任意一个或多个区段可以是所关注序列。

[0225] 在其他实施方案中,所关注序列是与部分非整倍体相关联的染色体区段,并且归一化序列是一条或多条全染色体。在其他实施方案中,所关注序列是与非整倍体相关联的全染色体,并且归一化序列是与非整倍体不相关的一个或多个染色体区段。

[0226] 无论是否在合格样品中将单条序列或一组序列识别为任意一条或多条所关注序列的归一化序列,都可以选择合格归一化序列以使其具有序列标签覆盖度变异或片段大小参数,该序列标签覆盖度变异或片段大小参数最佳或有效地近似于在合格样品中确定的所关注序列的序列标签覆盖度变异或片段大小参数。例如,合格归一化序列是当用于将所关注序列归一化时在合格样品间产生最小变异性的序列,即,归一化序列的变异性最接近在合格样品中确定的所关注序列的变异性。换言之,合格归一化序列是选择用于在合格样品间产生(所关注序列的)最小序列剂量变异的序列。因此,该方法选择当用作归一化染色体时预期产生所关注序列的运行间染色体剂量的最小变异性的序列。

[0227] 在合格样品中所识别的任意一条或多条所关注序列的归一化序列仍然是用于确定在数天、数周、数月 and 可能数年内试验样品中是否存在非整倍体的所选归一化序列,前提条件是生成测序文库和对样品进行测序所需的程序基本上不随时间推移而改变。如上所述,就在样品(例如,不同样品)和测序运行(例如,在同一天和/或不同天进行的测序运行)间映射至其的序列标签数目或片段大小参数的值的变异性(也可能是其他原因)而选择用于确定存在非整倍体的归一化序列,该变异性最佳近似于用作归一化参数的所关注序列的变异性。这些程序中的实质性改变将影响映射到所有序列的标签数目,这继而将确定在同一天或不同天进行的相同和/或不同测序运行中哪一条或哪一组序列将在样品间具有最接近所关注序列的变异性的变异性,这将需要重新确定归一化序列集。程序中的实质改变包括用于制备测序文库的实验室方案的改变(其包括与制备多重测序而非单重测序的样品相关的改变)和测序平台的改变(其包括用于测序的化学性质的改变)。

[0228] 在一些实施方案中,选择用于将特定所关注序列归一化的归一化序列是将一个或多个合格样品与一个或多个受影响样品最佳区分开的序列,这意味着归一化序列是具有最大可区分性的序列,即,归一化序列的可区分性使得其为受影响试验样品中的所关注序列提供最佳区分,以容易地将受影响试验样品与其他未受影响样品区分开。在其他实施方案中,归一化序列使具有最小变异性和最大可区分性的组合的序列。

[0229] 可将可区分性的水平确定为合格样品群中的序列剂量(例如染色体剂量或区段剂量)与一个或多个试验样品中的染色体剂量之间的统计差异,如下文所述 and 实施例所示。例如,可将可区分性用数字表示为t检验值,其表示合格样品群中的染色体剂量与一个或多个试验样品中的染色体剂量之间的统计差异。类似地,可区分性可基于区段剂量而非染色体

剂量。另选地,可将可区分性用数字表示为归一化染色体值(NCV),其为染色体剂量的 z 得分,只要NCV的分布是正态的。类似地,在染色体区段是所关注序列的情况下,可将区段剂量的可区分性用数字表示为归一化区段值(NSV),其是染色体片段剂量的 z 得分,只要NSV的分布是正态的。在确定 z 得分时,可使用合格样品集中的染色体或区段剂量的平均和标准偏差。另选地,可使用包括合格样品和受影响样品的训练集中的染色体或区段剂量的平均值和标准偏差。在其他实施方案中,归一化序列是具有最小变异性和最大可区分性或小变异性和大可区分性的最佳组合的序列。

[0230] 该方法识别固有相似特性并且易于产生样品和测序运行间相似变异的序列,并且所识别的序列可用于确定试验样品中的序列剂量。

[0231] 序列剂量的确定

[0232] 在一些实施方案中,如图1所示的步骤146中所述,在所有合格样品中确定一条或多条所关注染色体或区段的染色体或区段剂量,并且在步骤145中识别归一化染色体或区段序列。在计算序列剂量之前提供一些归一化序列。然后,如下文进一步所述,根据各种标准来识别一条或多条归一化序列,参见步骤145。在一些实施方案中,例如,所识别的归一化序列在所有合格样品间产生所关注序列的最小序列剂量变异性。

[0233] 在步骤146中,基于计算的合格标签密度,所关注序列的合格序列剂量(即染色体剂量或区段剂量)被确定为该所关注序列的序列标签覆盖度与随后在步骤145中从中识别归一化序列的另外序列的合格序列标签覆盖度的比率。随后使用所识别的归一化序列来确定试验样品中的序列剂量。

[0234] 在一个实施方案中,合格样品中的序列剂量是以合格样品中所关注染色体的序列标签数目或片段大小参数与归一化染色体序列的序列标签数目的比率来计算的染色体剂量。归一化染色体序列可以是单条染色体、一组染色体、一条染色体的区段或来自不同染色体的一组区段。因此,在合格样品中将所关注染色体的染色体剂量确定为所关注染色体的标签数目与(i)由单条染色体构成的归一化染色体序列、(ii)由两条或更多条染色体构成的归一化染色体序列、(iii)由染色体的单个区段构成的归一化区段序列、(iv)由来自一条染色体的两个或更多个区段构成的归一化区段序列或(v)由两条或更多条染色体的两个或更多个片段构成的归一化区段序列的标签数目的比率。用于根据(i)-(v)确定所关注21号染色体的染色体剂量的示例如下:所关注染色体(例如,21号染色体)的染色体剂量被确定为21号染色体的序列标签覆盖度与以下序列标签覆盖度中的一个序列标签覆盖度的比率:(i)所有剩余染色体中的每一条染色体,即1-20号染色体、22号染色体、X染色体和Y染色体;(ii)两条或更多条剩余染色体的所有可能组合;(iii)另一条染色体(例如9号染色体)的区段;(iv)一条其他染色体的两个区段,例如9号染色体的两个区段;(v)两条不同染色体的两个区段,例如9号染色体的区段和14号染色体的区段。

[0235] 在另一个实施方案中,合格样品中的序列剂量是与染色体剂量相对的区段剂量,以合格样品中所关注区段(不是全染色体)的序列标签数目与归一化区段序列的序列标签数目的比率来计算该区段剂量。归一化区段序列可为上文所讨论的归一化染色体或区段序列中的任意一条序列。

[0236] 归一化序列的识别

[0237] 在步骤145中,识别所关注序列的归一化序列。在一些实施方案中,例如,归一化序

列是基于所计算的序列剂量的序列,例如,其在所有合格训练样品间产生所关注序列的最小序列剂量变异性。该方法识别固有相似特性并且易于产生样品和测序运行间相似变异的序列,并且所识别的序列可用于确定试验样品中的序列剂量。

[0238] 可在合格样品集中识别一个或多个所关注序列的归一化序列,并且随后使用在合格样品中所识别的序列来计算试验样品中的每一个试验样品中的一条或多条所关注序列的序列剂量(步骤150),以确定在试验样品中的每个试验样品中是否存在非整倍体。当使用不同的测序平台时和/或当待测序的核酸的纯化和/或测序文库的制备中存在差异时,所识别的所关注染色体或区段的归一化序列可不同。不考虑所使用的样品制备和/或测序平台,根据本文所述的方法使用归一化序列提供了染色体或其区段的拷贝数变异的特异性和灵敏量度。

[0239] 在一些实施方案中,识别多于一条归一化序列,即,可确定一条所关注序列的不同归一化序列,并且可确定一条所关注序列的多个序列剂量。例如,当使用14号染色体的序列标签覆盖度时,所关注21号染色体的染色体剂量变异,例如变异系数($CV = \text{标准偏差} / \text{平均值}$)是最小的。然而,可识别两条、三条、四条、五条、六条、七条、八条或更多条归一化序列以用于确定试验样品中所关注序列的序列剂量。例如,可使用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来确定任意一个试验样品中21号染色体的第二剂量,因为这些染色体均具有接近14号染色体的CV。

[0240] 在一些实施方案中,当选择单条染色体作为所关注染色体的归一化染色体序列时,归一化染色体序列将是产生所关注染色体的染色体剂量的染色体,其在所有试验样品(例如合格样品)间具有最小变异性。在一些情况下,最佳归一化染色体可不具有最小变异,但可具有将一个或多个试验样品与合格样品最佳区分开的合格剂量分布,即最佳归一化染色体可不具有最小变异,但可具有最大可区分性。

[0241] 在一些实施方案中,归一化序列包括一条或多条稳健常染色体序列或其片段。在一些实施方案中,稳健常染色体包括除所关注染色体之外的所有常染色体。在一些实施方案中,稳健常染色体包括除X染色体、Y染色体、13号染色体、18号染色体和21号染色体之外的所有常染色体。在一些实施方案中,稳健常染色体包括除从样品确定的偏离正常二倍体状态的常染色体之外的所有常染色体,这些常染色体可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。

[0242] 试验样品中非整倍体的确定

[0243] 基于识别合格样品中的归一化序列,确定试验样品中所关注序列的序列剂量,该试验样品包含衍生自在一条或多条所关注序列中不同的基因组的核酸混合物。

[0244] 在步骤115中,从疑似或已知携带所关注序列的临床相关CNV的受试者中获得试验样品。试验样品可以是生物流体(例如血浆)或如下所述的任何合适的样品。如所解释的,可使用非侵入性手术(诸如简单的采血)来获得样品。在一些实施方案中,试验样品含有核酸分子(例如cfDNA分子)的混合物。在一些实施方案中,试验样品是含有胎儿和母体cfDNA分子混合物的母体血浆样品。

[0245] 在步骤125中,如针对合格样品所述对试验样品中的测试核酸的至少一部分进行测序,以生成数百万个序列读段,例如36bp读段。在各种实施方案中,将 $2 \times 36\text{bp}$ 配对末端读段用于配对末端测序。如步骤120中所述,将通过对试验样品中的核酸进行测序而生成的读

段唯一地映射到参考基因组或与参考基因组进行比对以产生标签。如步骤120中所述,从唯一地映射到参考基因组的读段中获得包含20bp至40bp的读段的至少约 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。在某些实施方案中,通过测序装置产生的读段以电子格式提供。使用如下所述的计算装置来实现比对。将各个读段与参考基因组进行比较,该参考基因组通常非常庞大(数百万个碱基对)以识别读段唯一地与参考基因组对应的位点。在一些实施方案中,比对程序允许读段与参考基因组之间的有限错配。在一些情况下,允许读段中的1、2或3个碱基对与参考基因组中的对应碱基对错配,但仍然进行映射。

[0246] 在步骤135中,使用如下所述的计算装置对通过对试验样品中的核酸进行测序而获得的标签中的所有或大部分进行计数,以确定测试序列标签覆盖度。在一些实施方案中,将每个读段与参考基因组的特定区域(在大多数情况下为染色体或区段)进行比对,并且通过将位点信息附加到读段来将读段转化为标签。当该过程展开时,计算装置可对映射到参考基因组中的每个区域(在大多数情况下为染色体或区段)的标签/读段的数目持续进行运行计数。将每个所关注染色体或区段以及每个对应的归一化染色体或区段的计数存储起来。

[0247] 在某些实施方案中,参考基因组具有一个或多个排除区域,所述一个或多个排除区域为真实生物基因组的一部分但不包括在参考基因组中。不对可能与这些排除区域进行比对的读段进行计数。排除区域的示例包括长重复序列的区域、X染色体和Y染色体之间的相似性区域等。使用通过上述掩蔽技术获得的掩蔽参考序列,仅考虑将参考序列的未掩蔽片段上的标签用于CNV分析。

[0248] 在一些实施方案中,当将多个读段与参考基因组或序列上的相同位点进行比对时,方法确定是否对标签进行不止一次计数。可能存在两个标签具有相同序列并因此与参考序列上的相同位点进行比对的情况。用于对标签进行计数的方法在某些情况下可从计数中排除来源于相同测序样品的相同标签。如果在给定样品中有不成比例数目的标签相同,则表明在该程序中存在强偏差或其他缺陷。因此,根据某些实施方案,计数方法不对来自给定样品的与来自先前计数的样品中的标签相同的标签进行计数。

[0249] 可设定各种标准以用于选择何时忽略来自单个样品的相同标签。在某些实施方案中,计数的标签的限定百分比必须是唯一的。如果超过该阈值的标签不是唯一的,则忽略它们。例如,如果限定的百分比要求至少50%是唯一的,则直到样品的唯一标签百分比超过50%时才对相同标签进行计数。在其他实施方案中,唯一标签的阈值数目为至少约60%。在其他实施方案中,唯一标签的阈值百分比为至少约75%、或至少约90%、或至少约95%、或至少约98%、或至少约99%。可将用于21号染色体的阈值设定为90%。如果将30M标签与21号染色体进行比对,则它们中的至少27M必须是唯一的。如果3M计数的标签不是唯一的,并且3千万零一个标签不是唯一的,则不进行计数。可使用适当的统计分析对用于确定何时不对其他相同标签进行计数的特定阈值或其他标准进行选择。影响该阈值或其他标准的一个因素是测序样品与标记可比对的基因组大小的相对量。其他因素包括读段大小和类似考虑因素。

[0250] 在一个实施方案中,将映射到所关注序列的测试序列标签数目归一化为它们所映射到的所关注序列的已知长度,以提供测试序列标签密度比。如针对合格样品所述,不需要归一化为所关注序列的已知长度,并且该归一化可作为减少数字中数字位数的步骤包括在内以简化数字以供人类解读。当对试验样品中的所有映射测试序列标签进行计数时,确定试验样品中所关注序列(例如,临床相关序列)的序列标签覆盖度,以及对应于合格样品中识别的至少一条归一化序列的另外序列的序列标签覆盖度。

[0251] 在步骤150中,基于合格样品中至少一条归一化序列的同一性,确定试验样品中所关注序列的测试序列剂量。在各种实施方案中,使用如本文所述的所关注序列和对应归一化序列的序列标签覆盖度来计算确定测试序列剂量。负责该任务的计算装置将以电子方式访问所关注序列和其相关联的归一化序列之间的关联,该关联可存储在数据库、表格、图形中,或作为代码包含在程序指令中。

[0252] 如本文他处所述,至少一条归一化序列可以是单条序列或一组序列。试验样品中所关注序列的序列剂量是在该试验样品中确定的所述所关注序列的序列标签覆盖度与在该试验样品中确定的至少一条归一化序列的序列标签覆盖度的比率,其中该试验样品中的该归一化序列对应于在合格样品中识别的特定所关注序列的归一化序列。例如,如果在合格样品中识别的21号染色体的归一化序列被确定为染色体(例如14号染色体),则21号染色体(所关注序列)的测试序列剂量确定为分别在试验样品中确定的21号染色体的序列标签覆盖度与14号染色体的序列标签覆盖度的比率。类似地,测定与染色体非整倍体相关联的13号染色体、18号染色体、X染色体、Y染色体和其他染色体的染色体剂量。所关注染色体的归一化序列可以是一条或一组染色体,或者一个或一组染色体区段。如前所述,所关注序列可以是染色体的一部分,例如染色体区段。因此,染色体片段的剂量可被确定为在试验样品中确定的区段的序列标签覆盖度与在试验样品中确定的归一化染色体区段的序列标签覆盖度的比率,其中该试验样品中的该归一化区段对应于在合格样品中识别的特定所关注区段的归一化区段(单个或一组区段)。染色体区段的大小可在千碱基(kb)至兆碱基(Mb)的范围内(例如,约1kb至10kb、或约10kb至100kb、或约100kb至1Mb)。

[0253] 在步骤155中,根据建立的多个合格样品中确定的合格序列剂量与针对确定的已知为非整倍体样品的所关注序列的序列剂量的标准偏差值来推导阈值。需注意,该操作通常与患者试验样品的分析异步执行。该操作可例如与从合格样品中选择归一化序列同时进行。准确分类取决于不同类别(即,非整倍体类型)的概率分布之间的差异。在一些示例中,从每种类型的非整倍体(例如,21-三倍体)的经验分布中选择阈值。如实施例中所述,建立用于对13-三倍体、18-三倍体、21-三倍体和X单倍体非整倍体进行分类的可能阈值,这些实施例描述了通过对从包含胎儿和母体核酸混合物的母体样品中提取的cfDNA进行测序来确定染色体非整倍体的方法的用途。确定用于区分受染色体非整倍体影响的样品的阈值可与用于不同非整倍体的阈值相同或不同。如实施例所示,根据所关注染色体在样品和测序运行中的剂量变异性来确定针对每条所关注染色体的阈值。任何所关注染色体的染色体剂量的变异越小,所关注染色体在用于设定用于确定不同非整倍体的阈值的所有未受影响样品中的剂量分布越窄。

[0254] 返回与对患者试验样品进行分类相关联的过程流程,在步骤160中,通过将所关注序列的测试序列剂量与从合格序列剂量建立的至少一个阈值进行比较来确定所关注序列

在试验样品中的拷贝数变异。该操作可由用于测量序列标签覆盖度和/或计算区段剂量的相同计算装置执行。

[0255] 在步骤160中,将所关注测试序列的所计算剂量与设定为根据用户定义的“可靠性阈值”进行选择的阈值的剂量进行比较,以将样品分类为“正常”、“受影响”或“无判定”。“无判定”样品是无法可靠地作出明确诊断的样品。每种类型的受影响样品(例如,21-三倍体、部分21-三倍体、X单倍体)具有其自身的阈值,一种用于判定正常(未受影响)样品,另一种用于判定受影响样品(尽管在一些情况下两个阈值一致)。如本文他处所述,在一些情况下,如果试验样品中核酸的胎儿分数足够高,则可将无判定转换为判定(受影响或正常)。可由在该过程流程的其他操作中采用的计算装置报告测试序列的分类。在一些情况下,分类以电子格式报告,并且可以显示、电子邮件、短信等方式发送给感兴趣的人。

[0256] 在一些实施方案中,CNV的确定包括计算将染色体或区段剂量与如上所述的合格样品集中的对应染色体或区段剂量的平均值关联的NCV或NSV。然后可通过将NCV/NSV与预定拷贝数评估阈值进行比较来确定CNV。

[0257] 可选择拷贝数评估阈值以优化假阳性率和假阴性率。拷贝数评估阈值越高,发生假阳性的可能性越小。类似地,阈值越低,发生假阴性的可能性越小。因此,在第一理想阈值(高于此仅对真阳性进行分类)和第二理想阈值(低于此仅对真阴性进行分类)之间存在权衡。

[0258] 阈值的设定在很大程度上取决于在未受影响样品集中确定的特定所关注染色体的染色体剂量变异性。变异性取决于多种因素,包括存在于样品中的胎儿cDNA的分数。通过染色体剂量在未受影响样品群体间的平均值或中值以及标准偏差来确定变异性(CV)。因此,用于对非整倍体进行分类的阈值根据下式使用NCV:

$$[0259] \quad NCV_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}$$

[0260] (其中 $\hat{\mu}_j$ 和 $\hat{\sigma}_j$ 分别是合格样品集中第j个染色体剂量的估计平均值和标准偏差,并且 x_{ij} 是观察试验样品i的第j个染色体剂量。)

[0261] 其中相关联的胎儿分数如下:

$$[0262] \quad FF_{ij} = 2 \times \left| \frac{NCV_{ij} \times \hat{\sigma}_j}{\hat{\mu}_j} \right| = 2 \times NCV \times CV$$

[0263] 因此,对于所关注染色体的每个NCV,可根据基于所关注染色体在未受影响样品群间的染色体比率的平均值和标准偏差的CV,计算与给定NCV值相关联的预期胎儿分数。

[0264] 随后,可基于胎儿分数与NCV值之间的关系选择决策边界,高于该决策边界基于正态分布分位数确定样品为阳性(受影响)。如上所述,在一些实施方案中,设定用于在真阳性的检测和假阴性率结果之间进行最佳权衡的阈值。即,选择阈值以使真阳性和真阴性之和最大化,或使假阳性和假阴性之和最小化。

[0265] 某些实施方案提供了用于对包含胎儿和母体核酸分子的生物样品中的胎儿染色体非整倍体进行产前诊断的方法。基于以下过程来作出该诊断:从衍生自生物试验样品(例如母体血浆样品)的胎儿和母体核酸分子混合物的至少一部分中获得序列信息,根据测序数据计算一条或多条所关注染色体的归一化染色体剂量和/或一条或多条所关注区段的归

一化区段剂量,以及确定该试验样品中该所关注染色体的该染色体剂量和/或该所关注区段的该区段剂量分别与在多个合格(正常)样品中建立的阈值之间的统计显著差异,并基于该统计差异提供产前诊断。如方法的步骤160中所述,作出正常或受影响的诊断。在无法确信作出正常或受影响的诊断的情况下,提供“无判定”。

[0266] 在一些实施方案中,可选择两个阈值。选择第一阈值以使假阳性率最小化,高于该第一阈值则样品归类为“受影响”,并且选择第二阈值以使假阴性率最小化,低于该第二阈值则样品归类为“未受影响”。具有高于第二阈值但低于第一阈值的NCV的样品可归类为“疑似非整倍体”或“无判定”样品,对于疑似非整倍体或无判定样品,可通过独立方式来确认是否存在非整倍体。第一阈值与第二阈值之间的区域可称为“无判定”区域。

[0267] 在一些实施方案中,疑似和无判定阈值在表1中示出。可以看出,NCV的阈值在不同染色体间变化。在一些实施方案中,如上所述,阈值根据样品的FF变化。在一些实施方案中,此处应用的阈值技术有助于提高灵敏度和选择性。

[0268] 表1.疑似和受影响NCV阈值归类无判定范围

[0269]	疑似	受影响
13号染色体	3.5	4.0
18号染色体	3.5	4.5
21号染色体	3.5	4.0
X染色体(X0,XXX)	4.0	4.0
Y染色体(XX对比XY)	6.0	6.0

[0270] 使用三途径过程、似然比、T统计值和/或胎儿分数确定拷贝数

[0271] 三途径过程

[0272] 图16A示出了用于评估拷贝数的三途径过程的流程图。该图包括工作流程700的三个重叠途径,其中包括对与所有大小的片段相关联的读段覆盖度的途径1(或713A)分析、对与较短片段相关联的读段覆盖度的途径2(或713B)分析和对较短读段相对于所有读段的相对频率的途径3(或713C)分析。

[0273] 在获得读段计数之后,在途径713A中使用来自所有大小的片段的读段来确定覆盖度。在途径713B中使用来自短片段的读段来确定覆盖度。在途径713C中确定来自短片段的读段相对于所有读段的频率。相对频率在本文他处也称为大小比率或大小分数。相对频率是片段大小特性的示例。在一些具体实施中,短片段是短于约150个碱基对的片段。在各种具体实施中,短片段可在约50至150、80至150或110至150个碱基对的大小范围内。在一些具体实施中,第三途径或途径713C是可选的。

[0274] 三条途径713A、713B和713C的数据均进行归一化操作714、716、718、719和722,以去除与所关注序列的拷贝数无关的方差。这些归一化操作在方框723中框出。操作714包括通过将分析量除以参考序列的量的总值来将所关注序列的分析量归一化。该归一化步骤使用从试验样品中获得的值。类似地,操作718和722使用从试验样品中获得的值来将分析量归一化。操作716和719使用从未受影响样品训练集中获得的值。

[0275] 操作716去除从未受影响样品训练集中获得的全局波的方差。操作718来去除个体特定GC方差的方差。

[0276] 操作719使用主成分分析(PCA)方法去除进一步的方差。通过PCA方法去除的方差

归因于与所关注序列的拷贝数无关的因素。每个分组中的分析量(覆盖度、片段大小比率等)为PCA提供了自变量,而未受影响训练集的样品为这些自变量提供了值。训练集的样品均包括具有相同拷贝数的所关注序列的样品,例如两个拷贝的体细胞染色体、一个拷贝的X染色体(当男性样品用作未受影响样品时)或两个拷贝的X染色体(当女性样品用作未受影响样品时)。因此,样品中的方差不是由非整倍体或其他拷贝数方差引起的。训练集的PCA产生与所关注序列的拷贝数无关的主成分。然后可使用这些主成分来去除与所关注序列的拷贝数无关的试验样品中的方差。

[0277] 在某些实施方案中,使用根据所关注序列之外的区域中未受影响样品的数据估计的系数,从试验样品的数据中去除一个或多个主成分的方差。在一些具体实施中,区域表示所有稳健染色体。例如,对训练正常样品的归一化分组覆盖度数据执行PCA,从而提供与可捕获数据的最大方差的维度对应的主成分。如此捕获的方差与所关注序列中的拷贝数变异无关。在已经从训练正常样品中获得主成分之后,将这些主成分应用于测试数据。在来自所关注序列之外的区域的分组间生成以试验样品为响应变量并且以主成分为因变量的线性回归模型。所得的回归系数用于通过减去由所估计的回归系数限定的主分量的线性组合来将所关注区域的分组覆盖度归一化。这从所关注序列中去除了与CNV无关的方差。参见框719。将剩余数据用于下游分析。另外,操作722来去除异常数据点。

[0278] 在框723中进行归一化操作之后,已将所有分组的覆盖度值“归一化”以去除非整倍体或其他拷贝数变异之外的变异来源。在某种意义上,为了检测拷贝数变异,相对于其他分组富集或改变所关注序列的分组。参见框724,该框不是操作,而是表示所得的覆盖度值。大框723中的归一化操作可增加信号和/或减小正在分析的量的噪声。类似地,已将分组的短片段的覆盖度值归一化以去除非整倍体或其他拷贝数变异之外的变异来源,如框728所示,并且已类似地将分组的短片段的相对频率(或大小比率)归一化以去除非整倍体或其他拷贝数变异之外的变异来源,如框732所示。与框724一样,框728和732不是操作,而是表示在处理大框723之后的覆盖度和相对频率值。应当理解,大方框723中的操作可被修改、重排或去除。例如,在一些实施方案中,不执行PCA操作719。在其他实施方案中,不执行校正GC操作718。在其他实施方案中,改变操作的顺序;例如,在校正GC操作718之前执行PCA操作719。

[0279] 在框724所示的归一化和方差去除之后的所有片段的覆盖度用于获得框726中的t统计值。类似地,在框728所示的归一化和方差去除之后的短片段的覆盖度用于获得框730中的t统计值,并且在框732所示的归一化和方差去除之后的短片段的相对频率用于获得框734中的t统计值。

[0280] 将t统计值应用于拷贝数分析可有助于提高分析的准确性。仅使用平均值以及使用平均值和方差两者来区分两个分布并未捕获两个分布之间的方差。T统计值可反映分布的平均值和方差两者。

[0281] 在一些具体实施中,操作726如下计算t统计值:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

[0282]

[0283] 其中 $\bar{x}_1$ 是所关注序列的分组覆盖度, $\bar{x}_2$ 是参考区域/序列的分组覆盖度, s_1 是所关

注序列的覆盖度的标准偏差, s_2 是参考区域的分组的标准偏差, n_1 是所关注序列的分组数目; 并且 n_2 是参考区域的分组的数目。

[0284] 在一些具体实施中, 参考区域包括所有稳健染色体 (例如, 除了最可能包含非整倍体的染色体之外的染色体)。在一些具体实施中, 参考区域包括所关注序列之外的至少一条染色体。在一些模拟中, 参考区域包括不包含所关注序列的稳健染色体。在其他具体实施中, 参考区域包括已确定为训练样品集提供最佳信号检测能力的染色体集 (例如, 选自稳健染色体的染色体子集)。在一些实施方案中, 信号检测能力基于参考区域将包含拷贝数变异的分组与不包含拷贝数变异的分组区分开的能力。在一些实施方案中, 以与标题为“归一化序列的识别”的部分中所述的用于确定“归一化序列”或“归一化染色体”的方式相似的方式来识别参考区域。

[0285] 胎儿分数的确定

[0286] 返回图16A, 可将一个或多个胎儿分数估计值 (框735) 与框726、730和734中的任何 t 统计值组合, 以获得倍体病例的似然估计值。参见框736。在一些具体实施中, 通过图16B中的过程800、图16C中的过程900或图16D的过程1000中的任一过程获得一个或多个胎儿分数。可使用作为图16A中的工作流程700的工作流程来并行地实施这些过程。

[0287] 图16B示出了根据本公开的一些具体实施的用于根据覆盖度信息确定胎儿分数的示例性过程800。过程800开始于从训练集中获得训练样品的覆盖度信息 (例如, 序列剂量值)。参见框802。从已知怀有男性胎儿的孕妇中获得训练集中的每个样品。即, 样品含有男性胎儿的 cfDNA。在一些具体实施中, 操作802可获得以不同于如本文所述的序列剂量的方式进行归一化的序列覆盖度, 或者该操作可获得其他覆盖度值。

[0288] 然后方法800包括计算训练样品的胎儿分数。在一些具体实施中, 可根据序列剂量值计算胎儿分数:

$$[0289] \quad FF_j = -2 \times \frac{Rx_j - \text{median}(Rx_i)}{\text{median}(Rx_i)}$$

[0290] 其中 Rx_j 是男性样品的序列剂量, $\text{median}(Rx_i)$ 是女性样品的序列剂量的中值。在其他具体实施中, 可使用平均值或其他集中趋势量度。在一些具体实施中, 可通过其他方法获得 FF , 方法诸如X染色体和Y染色体的相对频率。参见框804。

[0291] 过程800还包括将参考序列划分成亚序列的多个分组。在一些具体实施中, 参考序列是完整基因组。在一些具体实施中, 分组是100kb分组。在一些具体实施中, 将基因组分成约25,000个分组。然后该过程获得分组的覆盖度。参见框806。在一些具体实施中, 在进行图16A的框723所示的归一化操作之后获得框806中所用的覆盖度。在其他具体实施中, 可使用来自不同大小范围的覆盖度。

[0292] 每个分组与训练集中的样品的覆盖度相关联。因此, 可获得每个分组的样品的覆盖度和样品的胎儿分数之间的相关性。过程800包括获得胎儿分数与所有分组的覆盖度之间的相关性。参见框808。然后该过程选择相关性值高于阈值的分组。参见框810。在一些具体实施中, 选择具有6000个最高相关性值的分组。目的是识别在训练样品中显示覆盖度与胎儿分数之间具有高相关性的分组。然后可使用这些分组来预测试验样品中的胎儿分数。虽然训练样品是男性样品, 但胎儿分数与覆盖度之间的相关性可适用于男性试验样品和女性试验样品。

[0293] 使用所选择的具有高相关性值的分组,该过程获得将胎儿分数与覆盖度相关联的线性模型。参见框812。每个所选择的分组为该线性模型提供自变量。因此,所获得的线性模型还包括每个分组的参数或权重。调整这些分组的权重以将该模型拟合到数据。在获得线性模型之后,过程800包括将试验样品的覆盖度数据应用于该模型以确定试验样品的胎儿分数。参见框814。所应用的试验样品的覆盖度数据用于在胎儿分数和覆盖度之间具有高相关性的分组。

[0294] 图16A示出了用于处理序列读段的工作流程700,其信息可用于获得胎儿分数估计值。在一些具体实施中,框723中的一个或多个归一化操作是可选的。途径1提供覆盖度信息,其可用于图16B所示的过程800的框806中。

[0295] 在一些具体实施中,可将多个胎儿分数估计值组合以提供复合胎儿分数估计值。可使用各种方法来获得胎儿分数估计值。例如,可从覆盖度信息中获得胎儿分数。参见图16B的过程800。在一些具体实施中,还可根据片段的大小分布估计胎儿分数。参见图16C的过程900。在一些具体实施中,还可根据8聚体频率分布估计胎儿分数。参见图16D的过程1000。

[0296] 在包含男性胎儿的cfDNA的试验样品中,还可根据Y染色体和/或X染色体的覆盖度估计胎儿分数。在一些具体实施中,通过使用选自以下项的信息来获得推定男性胎儿的胎儿分数的复合估计值:从分组的覆盖度信息中获得的胎儿分数,从片段大小信息中获得的胎儿分数、从Y染色体覆盖度中获得的胎儿分数、从X染色体中获得的胎儿分数以及它们的任何组合。在一些具体实施中,通过使用Y染色体的覆盖度来获得胎儿的推定性别。可以各种方式将两个或更多个胎儿分数组合以提供胎儿分数的复合估计值。例如,在一些具体实施中,可使用平均值或加权平均值方法,其中加权可基于胎儿分数估计值的统计置信度。

[0297] 在一些具体实施中,通过使用选自以下项的信息来获得推定女性胎儿的胎儿分数的复合估计值:从分组的覆盖度信息中获得的胎儿分数、从片段大小信息中获得的胎儿分数以及它们的任何组合。

[0298] 图16C示出了根据一些具体实施的用于根据大小分布信息确定胎儿分数的过程。过程900开始于从训练集中获得男性训练样品的覆盖度信息(例如,序列剂量值)。参见框902。然后过程900包括使用上文针对框804所述的方法来计算训练样品的胎儿分数。参见框904。

[0299] 过程900继续将大小范围划分成多个分组以提供基于片段大小的分组并确定这些基于片段大小的分组的读段频率。参见框906。在一些具体实施中,在未对框723中所示的因素进行归一化的情况下,获得基于片段大小的分组的频率。在一些具体实施中,在任选地进行图16A的框723所示的归一化操作之后获得基于片段大小的分组的频率。在一些具体实施中,将大小范围分成40个分组。在一些具体实施中,低末端处的分组包括大小小于约55个碱基对的片段。在一些具体实施中,低末端处的分组包括大小在约50至55个碱基对范围内的片段,这排除了读段短于50bp的信息。在一些具体实施中,高末端处的分组包括大小大于约245个碱基对的片段。在一些具体实施中,高末端处的分组包括大小在约245至250个碱基对范围内的片段,这排除了读段长于250bp的信息。

[0300] 过程900继续通过使用训练样品的数据来获得将胎儿分数与基于片段大小的分组的读段频率相关联的线性模型。参见框908。所获得的线性模型包括基于大小的分组的读段

频率的自变量。该模型还包括每个基于大小的分组的参数或权重。调整这些分组的权重以将该模型拟合到数据。在获得线性模型之后,过程900包括将试验样品的读段频率数据应用于该模型以确定试验样品的胎儿分数。参见框910。

[0301] 在一些具体实施中,可使用8聚体频率来计算胎儿分数。图16D示出了根据本公开的一些具体实施的用于根据8聚体频率信息确定胎儿分数的示例性过程1000。过程1000开始于从训练集中获得男性训练样品的覆盖度信息(例如,序列剂量值)。参见框1002。然后过程1000包括使用针对框804所述的方法中的任意一种方法来计算训练样品的胎儿分数。参见框1004。

[0302] 过程1000还包括从每个训练样品的读段中获得8聚体的频率(例如,8个位置处4个核苷酸的所有可能排列)。参见框1006。在一些具体实施中,获得至多或接近65,536个8聚体及它们的频率。在一些具体实施中,在未对框723中所示的因素进行归一化的情况下,获得基于8聚体的频率。在一些具体实施中,在任选地进行归一化操作之后获得8聚体频率。

[0303] 每个8聚体与训练集中的样品的频率相关联。因此,可获得每个8聚体的样品的8聚体频率和样品的胎儿分数之间的相关性。过程1000包括获得胎儿分数与所有8聚体的8聚体频率之间的相关性。参见框1008。然后该过程选择相关性值高于阈值的8聚体。参见框1010。目的是识别在训练样品中显示8聚体频率与胎儿分数之间具有高相关性的8聚体。然后可使用这些分组来预测试验样品中的胎儿分数。虽然训练样品是男性样品,但胎儿分数与8聚体频率之间的相关性可适用于男性试验样品和女性试验样品。

[0304] 使用所选择的具有高相关性值的8聚体,该过程获得将胎儿分数与8聚体频率相关联的线性模型。参见框1012。每个所选择的分组为该线性模型提供自变量。因此,所获得的线性模型还包括每个分组的参数或权重。在获得线性模型之后,过程1000包括将试验样品的8聚体频率数据应用于该模型以确定试验样品的胎儿分数。参见框1014。

[0305] 在一些具体实施中,在计算胎儿分数时,对与基因组的不同部分具有相关性的不同部分的覆盖度(或其他参数)进行加权。此类方法的示例包括在美国专利申请公布US2015/0005176和Kim等人(2015年),Prenatal Diagnosis,第35卷,第1-6页中描述的SeqFF方法,这些文献全文以引用方式并入以用于计算胎儿分数的目的。在一些具体实施中,具有较高游离核酸片段分数的分组权重更大以用于确定胎儿分数。

[0306] 确定似然比

[0307] 返回图16A,在一些具体实施中,过程700包括在操作736中使用基于由操作726提供的所有片段覆盖度的t统计值、由操作726提供的胎儿分数估计值和基于由操作730提供的短片段覆盖度的t统计值来获得最终倍体似然性。这些具体实施使用多变量正态模型来将途径1和途径2的结果组合。在用于评估CNV的一些具体实施中,倍体似然性是非整倍体似然性,其是具有非整倍体假设(例如,三倍体或单倍体)的模型的似然性减去具有整倍体假设的模型的似然性,其中该模型使用基于所有片段覆盖度的t统计值、胎儿分数估计值和基于短片段覆盖度的t统计值作为输入并提供似然性作为输出。

[0308] 在一些具体实施中,倍体似然性表示为似然比。在一些具体实施中,似然比建模为:

$$[0309] \quad LR = \frac{\sum_{ff_{total}} q(ff_{total}) * p_1(T_{short}, T_{all} | ff_{est})}{p_0(T_{short}, T_{all})}$$

[0310] 其中 p_1 表示数据来自表示3拷贝或1拷贝模型的多变量正态分布的似然性, p_0 表示数据来自表示2拷贝模型的多变量正态分布的似然性, T_{short} 、 T_{all} 是根据从短片段和所有片段生成的染色体覆盖度计算的T得分,而 $q(ff_{\text{total}})$ 是考虑到与胎儿分数估计值相关联的误差的胎儿分数的密度分布(根据训练数据估计)。该模型将由短片段生成的覆盖度与由所有片段生成的覆盖度组合,这有助于改善受影响样品和未受影响样品的覆盖度得分之间的间距。在所述的实施方案中,模型还利用胎儿分数,从而进一步改善区分受影响样品与未受影响样品的能力。此处,使用如上所述的基于所有片段覆盖度的t统计值(726)、基于短片段覆盖度的t统计值(730)和由过程800(或框726)、900或1000提供的胎儿分数估计值来计算似然比。在一些具体实施中,该似然比用于分析13号染色体、18号染色体和21号染色体。

[0311] 在一些具体实施中,由操作736获得的倍体似然性仅使用基于由途径3的操作734提供的短片段的相对频率和由操作726、过程800、900或1000提供的胎儿分数估计值获得的t统计值。可根据以下公式来计算似然比:

$$[0312] \quad LR = \frac{\sum_{ff_{\text{total}}} q(ff_{\text{total}}) * p_1(T_{\text{short_freq}} | ff_{\text{est}})}{p_0(T_{\text{short_freq}})}$$

[0313] 其中 p_1 表示数据来自表示3拷贝或1拷贝模型的多变量正态分布的似然性, p_0 表示数据来自表示2拷贝模型的多变量正态分布的似然性, $T_{\text{short-freq}}$ 是根据短片段的相对频率计算的T得分,而 $q(ff_{\text{total}})$ 是考虑到与胎儿分数估计值相关联的误差的胎儿分数的密度分布(根据训练数据估计)。此处,使用如上所述的基于短片段相对频率的t统计值(734)和由过程800(或框726)、900或1000提供的胎儿分数估计值来计算似然比。在一些具体实施中,该似然比用于分析X染色体。

[0314] 在一些具体实施中,使用基于所有片段覆盖度的t统计值(726)、基于短片段覆盖度的t统计值(730)和短片段的相对频率(734)来计算似然比。此外,可将如上所述获得的胎儿分数与t统计值组合来计算似然比。通过将来自三条途径713A、713B和713C中任一条途径的信息组合,可提高倍体评估的区分能力。在一些具体实施中,可使用不同的组合来获得染色体的似然比,组合例如,来自全部三条途径的t统计值、来自第一条和第二条途径的t统计值、胎儿分数和三个t统计值、胎儿分数和一个t统计值等。然后可基于模型性能来选择最佳组合。

[0315] 在上文所述的各种具体实施中,整倍体模型和非整倍体模型将t统计值作为输入值。然而,这些模型当然也可以原始的或以其他方式转换的覆盖度或丰度值作为输入,并且提供似然性作为输出。以其他方式转换的输入值或t统计输入值可有助于提高模型的预测能力,但并不是在所有具体实施中都必须进行这种转换。

[0316] 在用于评估常染色体的一些具体实施中,建模似然比表示已从三倍体或单倍体样品获得的建模数据的似然性相对于已从二倍体样品获得的建模数据的似然性。在一些具体实施中,此类似然比可用于确定常染色体的三倍体或单倍体。

[0317] 在用于评估性染色体的一些具体实施中,评估了X单倍体的似然比和X三倍体的似然比。此外,还评估了X染色体和Y染色体的染色体覆盖度测量值(例如CNV或覆盖度z得分)。在一些具体实施中,使用决策树来评估四个值以确定性染色体的拷贝数。在一些具体实施中,决策树允许确定XX、XY、X、XXY、XXX或XYY的倍体病例。

[0318] 在一些具体实施中,将似然比转换为对数似然比,并且可根据经验来设定用于判

定非整倍体或拷贝数变异的标准或阈值,以获得特定的灵敏度和选择性。例如,当将对数似然比应用于训练集时,可基于模型的灵敏度和选择性将该对数似然比设定为1.5以用于判定13-三倍体或18-三倍体。此外,例如,在一些应用中,可将针对21号染色体的三倍体的判定标准值设定为3。

[0319] 样品和样品处理

[0320] 样品

[0321] 用于确定CNV的样品(例如染色体非整倍体、部分非整倍体等)可包括从一个或多个所关注序列的拷贝数变异待确定的任何细胞、组织或器官中取得的样品。有利地,样品含有存在于细胞中的核酸和/或“游离”的核酸(例如cfDNA)。

[0322] 在一些实施方案中,有利的是获得游离核酸,例如游离DNA(cfDNA)。可通过本领域已知的多种方法从生物样品获得游离核酸(包括游离DNA),这些生物样品包括但不限于血浆、血清和尿液(参见例如Fan等人,Proc Natl Acad Sci第105章:第16266-16271页[2008年];Koide等人,Prenatal Diagnosis第25章:第604-607页[2005年];Chen等人,Nature Med.第2章:第1033-1035页[1996年];Lo等人,Lancet第350章:第485-487页[1997年];Botezatu等人,Clin Chem.第46章:第1078-1084页,2000年;和Su等人,J Mol.Diagn.第6章:第101-107页[2004年])。为了从样品中的细胞中分离游离DNA,可使用各种方法,包括但不限于分馏、离心(例如密度梯度离心)、DNA特异性沉淀或高通量细胞分选和/或其他分离方法。用于手动和自动分离cfDNA的市售试剂盒是可用的(Roche Diagnostics, Indianapolis, IN, Qiagen, Valencia, CA, Macherey-Nagel, Duren, DE)。包含cfDNA的生物样品已用于测定中,以通过可检测染色体非整倍体和/或多种多态性的测序测定来确定是否存在染色体异常(例如,21-三倍体)。

[0323] 在各种实施方案中,存在于样品中的cfDNA可在使用之前(例如,在制备测序文库之前)特异性地或非特异性地富集。样品DNA的非特异性富集是指可用于在制备cfDNA测序文库之前增加样品DNA水平的样品基因组DNA片段的全基因组扩增。非特异性富集可以是存在于包含多于一个基因组的样品中的两个基因组中的一个基因组的富集。例如,非特异性富集可对母体样品中的胎儿基因组具有选择性,其可通过已知方法获得以增加样品中胎儿与母体DNA的相对比例。另选地,非特异性富集可以是存在于样品中的两个基因组的非选择性扩增。例如,非特异性扩增可以是胎儿和母体DNA在包含来自胎儿和母体基因组的DNA混合物的样品中的扩增。用于全基因组扩增的方法是本领域已知的。简并寡核苷酸引物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示例。在一些实施方案中,包含来自不同基因组的cfDNA混合物的样品未富集存在于混合物中的基因组的cfDNA。在其他实施方案中,包含来自不同基因组的cfDNA混合物的样品非特异性富集存在于样品中的基因组中的任意一个基因组。

[0324] 包含本文所述方法所应用的核酸的样品通常包括生物样品(“试验样品”),例如如上所述。在一些实施方案中,通过多种熟知的方法中的任意一种方法纯化或分离待筛选一种或多种CNV的核酸。

[0325] 因此,在某些实施方案中,样品包含经纯化或分离的多核苷酸或由经纯化或分离的多核苷酸组成,或者其可包括样品,诸如组织样品、生物流体样品、细胞样品等。合适的生物流体样品包括但不限于血液、血浆、血清、汗液、泪液、痰、尿液、痰、耳溢液、淋巴液、唾液、

脑脊液、灌洗液、骨髓悬浮液、阴道液、经宫颈灌洗液、脑液、腹水、乳汁、呼吸道、肠道和泌尿生殖道的分泌物、羊水、乳汁和白细胞析离样品。在一些实施方案中,样品是易于通过非侵入性手术获得的样品,例如血液、血浆、血清、汗液、泪液、痰、尿液、痰、耳液、唾液或粪便。在某些实施方案中,样品是外周血样品或外周血样品的血浆和/或血清级分。在其他实施方案中,生物样品是拭子或涂片、活检标本或细胞培养物。在另一实施方案中,样品是两种或更多种生物样品的混合物,例如,生物样品可包括生物流体样品、组织样品和细胞培养样品中的两种或更多种。如本文所用,术语“血液”、“血浆”和“血清”明确地涵盖其级分或加工部分。类似地,在样品取自活检、拭子、涂片等的情况中,“样品”明确地涵盖来源于活检、拭子、涂片等的处理级分或部分。

[0326] 在某些实施方案中,样品可从来源中获得,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患有遗传性疾病的个体)的样品、正常个体、在个体疾病的不同阶段获得的样品、从对疾病进行不同治疗的个体获得的样品、受到不同环境因素影响的个体的样品、对病理易感的个体的样品、从暴露于传染病因子(例如HIV)的个体获得的样品等。

[0327] 在一个示例性但非限制性的实施方案中,样品是从妊娠女性(例如,孕妇)中获得的母体样品。在这种情况下,可使用本文所述的方法分析样品以提供胎儿的潜在染色体异常的产前诊断。母体样品可以是组织样品、生物流体样品或细胞样品。作为非限制性示例,生物流体包括血液、血浆、血清、汗液、泪液、痰、尿液、痰、耳溢液、淋巴液、唾液、脑脊液、灌洗液、骨髓悬浮液、阴道液、经宫颈灌洗液、脑液、腹水、乳汁、呼吸道、肠道和泌尿生殖道的分泌物和白细胞析离样品。

[0328] 在另一个示例性但非限制性的实施方案中,母体样品是两种或更多种生物样品的混合物,例如,生物样品可包括生物流体样品、组织样品和细胞培养样品中的两种或更多种。在一些实施方案中,样品是易于通过非侵入性手术获得的样品,例如血液、血浆、血清、汗液、泪液、痰、尿液、乳汁、痰、耳溢液、唾液和粪便。在一些实施方案中,生物样品是外周血样品和/或其血浆及血清组分。在其他实施方案中,生物样品是拭子或涂片、活检标本或细胞培养物样品。如上所公开,术语“血液”、“血浆”和“血清”明确地涵盖其级分或加工部分。类似地,在样品取自活检、拭子、涂片等的情况中,“样品”明确地涵盖来源于活检、拭子、涂片等的处理级分或部分。

[0329] 在某些实施方案中,也可从体外培养的组织、细胞或其他含多核苷酸的来源中获得样品。培养样品可从来源中获取,包括但不限于在不同培养基和条件(例如pH、压力或温度)中维持的培养物(例如组织或细胞)、维持不同时间的培养物(例如组织或细胞)、用不同的因子或试剂(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。

[0330] 从生物来源分离核酸的方法是熟知的,并且将根据来源的性质而有所不同。本领域技术人员可以容易地从本文所述方法所需的来源分离核酸。在某些情况下,可能有利的是将核酸样品中的核酸分子片段化。片段化可以是随机的,或者它可以是特异性的,如例如使用限制性内切核酸酶消化所实现的。随机片段化的方法是本领域熟知的,并且包括例如限制性DNA酶消化、碱处理和物理剪切。在一个实施方案中,从未经片段化的cfDNA中获得样品核酸。

[0331] 测序文库制备

[0332] 在一个实施方案中,本文所述的方法可利用下一代测序技术(NGS),其允许在单次测序运行中以基因组分子(即,单重测序)或以包含索引基因组分子的合并样品(例如,多重测序)对多个样品进行单独测序。这些方法可生成高达数十亿个DNA序列读段。在各种实施方案中,可使用例如本文所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。在各种实施方案中,可使用如本文所述的一个或多个处理器来执行对使用NGS获得的大量序列数据的分析。

[0333] 在各种实施方案中,此类测序技术的使用不包括测序文库的制备。

[0334] 然而,在某些实施方案中,本文设想的测序方法包括测序文库的制备。在一个示例性方法中,测序文库制备包括随机采集待测序的衔接子修饰的DNA片段(例如,多核苷酸)。可通过逆转录酶的作用从DNA或RNA(包括DNA或cDNA的等同物或类似物,例如由RNA模板产生的互补或拷贝DNA的DNA或cDNA)中制备多核苷酸测序文库。多核苷酸可以双链形式(例如,dsDNA,诸如基因组DNA片段、cDNA、PCR扩增产物等)起始,或者在某些实施方案中,多核苷酸可以单链形式(例如,ssDNA、RNA等)起始并已转化为dsDNA形式。举例来说,在某些实施方案中,单链mRNA分子可拷贝成适用于制备测序文库的双链cDNA。初级多核苷酸分子的精确序列通常对文库制备方法并不重要,并且可以是已知的或未知的。在一个实施方案中,多核苷酸分子是DNA分子。更具体地,在某些实施方案中,多核苷酸分子表示生物体的整个遗传互补序列或生物体的基本上整个遗传互补序列,并且是基因组DNA分子(例如,细胞DNA、游离DNA(cfDNA)等),其通常包括内含子序列和外显子序列(编码序列),以及非编码调控序列诸如启动子和增强子序列。在某些实施方案中,初级多核苷酸分子包括人类基因组DNA分子,例如存在于妊娠受试者的外周血中的cfDNA分子。

[0335] 通过使用包含特定范围的片段大小的多核苷酸来促进一些NGS测序平台的测序文库的制备。此类文库的制备通常包括大的多核苷酸(例如,细胞基因组DNA)的片段化,以获得所需大小范围内的多核苷酸。

[0336] 可通过本领域技术人员已知的多种方法中的任意一种方法来实现片段化。例如,可通过机械方法来实现片段化,机械方法包括但不限于雾化、超声处理和水剪切。然而,机械片段化通常在C-O、P-O和C-C键处切割DNA主链,从而产生具有断裂的C-O、P-O和C-C键的平末端和3'-和5'-突出末端的异质混合物(参见例如Alnemri和Liwack, J Biol.Chem第265章:第17323-17333页[1990年];Richards和Boyer, J Mol Biol第11章:第327-240页[1965年]),该异质混合物可能需要修复,因为它们可能缺少后续酶反应(例如,制备用于测序的DNA所需的测序衔接子的连接)所必需的5'-磷酸。

[0337] 相比之下,cfDNA通常以小于约300个碱基对的片段存在,因此片段化通常不是使用cfDNA样品生成测序文库所必需的。

[0338] 通常,无论多核苷酸是强制片段化的(例如体外片段化的)还是作为片段天然存在的,它们均转化成具有5'-磷酸和3'-羟基的平末端DNA。标准方案,例如,使用例如本文他处所述的Illumina平台进行测序的方案,指示用户对样品DNA进行末端修复,在dA-加尾之前对末端修复的产品进行纯化,以及在文库制备的衔接子连接步骤之前对dA-加尾产品进行纯化。

[0339] 本文所述的测序文库制备方法的各实施方案无需执行标准方案通常要求的一

个或多个步骤来获得可由NGS测序的经修饰的DNA产物。简化方法(ABB方法)、1步方法和2步方法是用于制备测序文库的方法的示例,其可见于2012年7月20日提交的专利申请13/555,037,该专利申请全文以引用方式并入本文。

[0340] 用于跟踪和验证样品完整性的标记核酸

[0341] 在各种实施方案中,可通过例如在处理之前对已引入样品中的样品基因组核酸(例如,cfDNA)和伴随的标记核酸的混合物进行测序来实现样品完整性的验证和样品跟踪。

[0342] 可将标记核酸可与试验样品(例如,生物来源样品)组合并进行包括例如以下步骤中的一个或多个步骤的处理:对生物来源样品进行分馏(例如,从全血样品中获得基本上游离的血浆级分),对来自分馏的(例如,血浆)或未分馏的生物来源样品(例如,组织样品)的核酸进行纯化,以及测序。在一些实施方案中,测序包括制备测序文库。与来源样品组合的标记分子的序列或序列组合选择为对该来源样品而言是唯一的。在一些实施方案中,样品中的唯一标记分子均具有相同的序列。在其他实施方案中,样品中的唯一标记分子是多条序列,例如两条、三条、四条、五条、六条、七条、八条、九条、十条、十五条、二十条或更多条不同序列的组合。

[0343] 在一个实施方案中,可使用具有相同序列的多个标记核酸分子来验证样品的完整性。另选地,可使用多个标记核酸分子来验证样品的同一性,所述标记核酸分子具有至少两条、至少三条、至少四条、至少五条、至少六条、至少七条、至少八条、至少九条、至少十条、至少十一条、至少十二条、至少十三条、至少十四条、至少十五条、至少十六条、至少十七条、至少十八条、至少十九条、至少二十条、至少二十五条、至少三十条、至少三十五条、至少四十条、至少五十条或更多条不同的序列。验证所述多个生物样品(即,两个或更多个生物样品)的完整性要求所述两个或更多个样品中的每一个样品都用标记核酸进行标记,所述标记核酸具有对于被标记的所述多个试验样品中的每一个样品而言是唯一的序列。例如,可用具有序列A的标记核酸来标记第一样品,并且可用具有序列B的标记核酸来标记第二样品。另选地,可用均具有序列A的标记核酸分子来标记第一样品,并且可用序列B和C的混合物来标记第二样品,其中序列A、B和C是具有不同序列的标记分子。

[0344] 可在文库制备(如果要制备文库的话)和测序之前发生的样品制备的任何阶段将标记核酸添加到样品中。在一个实施方案中,可将标记分子与未处理的来源样品组合。例如,可在用于采集血样的采集管中提供标记核酸。另选地,可在抽血后将标记核酸添加到血样中。在一个实施方案中,将标记核酸添加到用于采集生物流体样品的容器中,例如将标记核酸添加到用于采集血样的采血管中。在另一个实施方案中,将标记核酸添加到生物流体样品的级分中。例如,将标记核酸添加到血样(例如母体血浆样品)的血浆和/或浆液组分中。在另一个实施方案中,将标记分子添加到经纯化的样品中,例如已从生物样品中纯化的核酸样品。例如,将标记核酸添加到经纯化的母体和胎儿cfDNA样品中。类似地,可在处理标本之前将标记核酸添加到活检标本中。在一些实施方案中,可将标记核酸与将标记分子递送到生物样品的细胞中的载体组合。细胞递送载体包括pH灵敏型和阳离子脂质体。

[0345] 在各种实施方案中,标记分子具有反基因组序列,这些反基因组序列是不存在于生物来源样品的基因组中的序列。在一个示例性实施方案中,用于验证人类生物来源样品完整性的标记分子具有不存在于人类基因组中的序列。在另选的实施方案中,标记分子具有不存在于来源样品和任意一种或多种其他已知基因组中的序列。例如,用于验证人类生

物来源样品完整性的标记分子具有不存在于人类基因组和小鼠基因组中的序列。替代方案允许验证包含两个或更多个基因组的试验样品的完整性。例如,可使用具有不存在于人类基因组和受影响细菌的基因组两者中的序列的标记分子来验证从受病原体(例如,细菌)影响的受试者中获得的人类游离DNA样品的完整性。可在万维网上ncbi.nlm.nih.gov/genomes处公开获得多种病原体(例如细菌、病毒、酵母菌、真菌、原生动物等)的基因组序列。在另一个实施方案中,标记分子是具有不存在于任何已知基因组中的序列的核酸。可通过算法随机生成标记分子的序列。

[0346] 在各种实施方案中,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),人工核酸类似物包括肽核酸(PNA)、吗啉代核酸、锁核酸、二醇核酸和苏糖核酸,它们与天然存在的DNA或RNA的区别在于不具有磷酸二酯主链的分子或DNA模拟物主链的改变。脱氧核糖核酸可来自天然存在的基因组或可通过使用酶或通过固相化学合成在实验室中生成。化学方法也可用于生成自然界中不存在的DNA模拟物。可用的其中磷酸二酯键已被置换,但脱氧核糖得以保留的DNA的衍生物包括但不限于具有由硫代甲酰缩醛或羧酰胺键形成的主链的DNA模拟物,该DNA模拟物已显示是良好的结构DNA模拟物。其他DNA模拟物包括吗啉代衍生物和肽核酸(PNA),PNA包含基于N-(2-氨基乙基)甘氨酸的假肽主链(Ann Rev Biophys Biomol Struct第24章:第167-183页[1995年])。PNA是极好的DNA(或核糖核酸[RNA])的结构模拟物,并且PNA低聚物能够与沃森-克里克(Watson-Crick)互补DNA和RNA(或PNA)低聚物形成非常稳定的双链体结构,并且它们还可通过螺旋侵入与双链体DNA中的靶标结合(Mol Biotechnol第26章:第233-248页[2004年])。可用作标记分子的DNA类似物的另一种良好的结构模拟物/类似物是硫代磷酸酯DNA,其中非桥联氧中的一个氧被硫置换。此修饰降低了核酸内切酶和核酸外切酶2(包括5'至3'和3'至5'DNA POL 1核酸外切酶)、核酸酶S1和P1、RNA酶、血清核酸酶和蛇毒磷酸二酯酶的作用。

[0347] 标记分子的长度可不同于或类似于样品核酸的长度,即,标记分子的长度可类似于样品基因组分子的长度,或者其可大于或小于样品基因组分子的长度。通过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。可使用本领域已知的分离方法,将长度不同于样品基因组分子的长度的标记分子与来源核酸区分开。例如,可通过电泳分离(例如毛细管电泳)来确定标记和样品核酸分子的长度差异。大小差异可有利于对标记和样品核酸质量的定量和评估。优选地,标记核酸短于基因组核酸,并且具有足够的长度以排除它们映射到样品的基因组。例如,如需要30个碱基的人类序列来唯一地映射到人类基因组的情况。因此,在某些实施方案中,用于对人类样品进行测序生物测定的标记分子的长度应为至少30bp。

[0348] 主要通过用于验证来源样品完整性的测序技术来确定标记分子长度的选择。还可考虑经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,这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,使用Illumina GAII序列分析仪进行测序包括通过桥式PCR(也称为簇扩增)对具有110bp最小长度的多核苷酸进行体外克隆扩增,将衔接子连接到该多核苷酸以提供可进行克隆扩增和测序的至少200bp且小于600bp的核酸。在一些实施方案中,衔接子连接的标记分子的长度在约200bp至约600bp之间、约250bp至550bp之间、约300bp至500bp之间、或约350bp至450bp之间。在其他实施方案中,衔接子连接的标记分子的长度为约200bp。例如,当对存在于母体样品中的胎儿cfDNA进行测

序时,可将标记分子的长度选择为类似于胎儿cfDNA分子的长度。因此,在一个实施方案中,在测定中使用的标记分子的长度可为约150bp、约160bp、170bp、约180bp、约190bp或约200bp,该测定包括对母体样品中的cfDNA进行大规模平行测序以确定是否存在胎儿染色体非整倍体;优选地,标记分子为约170bp。其他测序方法(例如SOLiD测序、Polony测序和454测序)使用乳液PCR来克隆扩增DNA分子进行测序,并且每种技术都规定了待扩增分子的最小和最大长度。待作为克隆扩增核酸进行测序的标记分子的长度可为至多约600bp。在一些实施方案中,待测序的标记分子的长度可大于600bp。

[0349] 单分子测序技术不采用分子的克隆扩增,并且能够在非常宽的模板长度范围内对核酸进行测序,在大多数情况下不需要待测序的分子具有任何特定长度。然而,每单位质量的序列收率取决于3'端羟基基团的数目,因此具有相对较短的测序模板比具有长模板更有效。如果从长于1000nt的核酸开始,则通常建议将核酸剪切至100至200nt的平均长度,使得可从相同质量的核酸产生更多的序列信息。因此,标记分子的长度可在数十个碱基至数千个碱基的范围内。用于单分子测序的标记分子的长度可为至多约25bp、至多约50bp、至多约75bp、至多约100bp、至多约200bp、至多约300bp、至多约400bp、至多约500bp、至多约600bp、至多约700bp、至多约800bp、至多约900bp、至多约1000bp或更长。

[0350] 还通过经测序的基因组核酸的长度来确定所选的标记分子的长度。例如,cfDNA在人类血流中作为细胞基因组DNA的基因组片段循环。孕妇血浆中存在的胎儿cfDNA分子通常比母体cfDNA分子短(Chan等人,Clin Chem第50章:第8892页[2004年])。循环胎儿DNA的大小分馏已证实循环胎儿DNA片段的平均长度为<300bp,而母体DNA据估计为约0.5Kb至1Kb之间(Li等人,Clin Chem,第50章:第1002-1011页[2004年])。这些发现与Fan等人的发现一致,他们使用NGS确定胎儿cfDNA很少为>340bp(Fan等人,Clin Chem第56章:第1279-1286页[2010年])。用标准的基于二氧化硅的方法从尿液中分离的DNA由两种级分组成:源自脱落细胞的高分子量DNA和经肾DNA(Tr-DNA)的低分子量(150-250碱基对)级分(Botezatu等人,Clin Chem.第46章:第1078-1084页,2000年;和Su等人,J Mol.Diagn.第6章:第101-107页,2004年)。将新开发的用于从体液中分离游离核酸的技术应用于分离经肾核酸已表明,尿液中存在的DNA和RNA片段远短于150个碱基对(美国专利申请公布20080139801)。在实施方案中,其中cfDNA是经测序的基因组核酸,所选择的标记分子可为至多约cfDNA的长度。例如,待作为单个核酸分子或克隆扩增核酸进行测序的母体cfDNA样品中使用的标记分子的长度可在约100bp至600bp之间。在其他实施方案中,样品基因组核酸是较大分子的片段。例如,经测序的样品基因组核酸是片段化的细胞DNA。在实施方案中,当对片段化的细胞DNA进行测序时,标记分子的长度可为至多DNA片段的长度。在一些实施方案中,标记分子的长度为至少将序列读段唯一地映射到适当的参考基因组所需的最小长度。在其他实施方案中,标记分子的长度是排除标记分子映射到样品参考基因组所需的最小长度。

[0351] 此外,标记分子可用于验证未通过核酸测序进行测定,并且可通过除测序之外的常见生物技术(例如实时PCR)进行验证的样品。

[0352] 样品对照(例如,用于测序和/或分析的过程中阳性对照)。

[0353] 在各种实施方案中,例如如上所述,引入样品中的标记序列可用作阳性对照来验证测序以及后续加工和分析的准确性和有效性。

[0354] 因此,提供了用于提供对样品中的DNA进行测序的过程中阳性对照(IPC)的组合物

和方法。在某些实施方案中,提供了用于对包含基因组混合物的样品中的cfDNA进行测序的阳性对照。IPC可用于将从不同样品(例如,在不同测序运行上在不同时间进行测序的样品)集中获得的序列信息的基线偏移相关联。因此,例如,IPC可将从母体试验样品中获得的序列信息与从在不同时间进行测序的合格样品集中获得的序列信息相关联。

[0355] 类似地,在区段分析的情况下,IPC可将从受试者中获得的特定区段的序列信息与从在不同时间进行测序的(类似序列的)合格样品集中获得的序列相关联。在某些实施方案中,IPC可将从受试者中获得的特定癌症相关基因座的序列信息与从合格样品集(例如,来自已知的扩增/缺失等)中获得的序列信息相关联。

[0356] 此外,IPC可用作在测序过程中跟踪样品的标记。IPC还可为所关注染色体的一个或多个非整倍体(例如,21-三倍体、13-三倍体、18-三倍体)提供定性阳性序列剂量值(例如,NCV),以提供适当的解释并确保数据的可靠性和准确性。在某些实施方案中,可产生包含来自男性基因组和女性基因组的核酸的IPC,来为母体样品中的X染色体和Y染色体提供剂量。

[0357] 过程中对照的类型和数量取决于所需测试的类型或性质。例如,对于需要对来自包含基因组混合物的样品的DNA进行测序以确定是否存在染色体非整倍体的测试,过程中对照可包含从已知包含正在测试的相同染色体非整倍体的样品中获得的DNA。在一些实施方案中,IPC包含来自已知包含所关注染色体的非整倍体的样品的DNA。例如,用于确定母体样品中是否存在胎儿三倍体(例如21-三倍体)的测试的IPC包含从具有21-三倍体的个体中获得的DNA。在一些实施方案中,IPC包含从具有不同非整倍体的两个或更多个个体中获得的DNA混合物。例如,用于确定是否存在13-三倍体、18-三倍体、21-三倍体和X单倍体的测试的IPC,包含从每个怀有胎儿的孕妇中获得的DNA样品与被测试的三倍体中的一种三倍体的组合。除了完整的染色体非整倍体之外,还可产生IPC来为用于确定是否存在部分非整倍体的测试提供阳性对照。

[0358] 可使用从两个受试者中获得的细胞基因组DNA的混合物来产生用作检测单个非整倍体的对照的IPC,其中一个受试者是非整倍体基因组的贡献者。例如,可通过将来自携带三体染色体的男性或女性受试者的基因组DNA与已知不携带三体染色体的女性受试者的基因组DNA组合来产生IPC,该IPC作为用于确定胎儿三倍体(21-三倍体)的测试的对照。可从两个受试者的细胞中提取基因组DNA,并且将其剪切来提供在约100bp至400bp之间、约150bp至350bp之间、或约200bp至300bp之间的片段,以模拟母体样品中的循环cfDNA片段。选择来自携带非整倍体(例如,21-三倍体)的受试者的片段化DNA的比例,来模拟母体样品中存在的循环胎儿cfDNA的比例,以提供包含片段化DNA的混合物的IPC,该片段化DNA包括约5%、约10%、约15%、约20%、约25%、约30%的来自携带非整倍体的受试者的DNA。IPC可包含来自各自携带不同非整倍体的不同受试者的DNA。例如,IPC可包含约80%的未受影响的女性DNA,并且剩余的20%可以是来自各自携带21-三体染色体、13-三体染色体和18-三体染色体的三个不同受试者的DNA。制备片段化DNA的混合物以进行测序。片段化DNA的混合物的处理可包括制备测序文库,该测序文库可使用任何大规模平行的方法以单重或多重方式进行测序。基因组IPC的储备溶液可储存并用于多个诊断测试中。

[0359] 另选地,可使用从已知怀有具有已知染色体非整倍体的胎儿的母体中获得的cfDNA来产生IPC。例如,可从怀有具有21-三倍体的胎儿的孕妇中获得cfDNA。从母体样品中

提取cfDNA,将其克隆到细菌载体中并在细菌中生长,以提供IPC的持续来源。可使用限制性酶从细菌载体中提取DNA。另选地,可通过例如PCR扩增克隆的cfDNA。可在与来自试验样品的cfDNA相同的运行中处理IPC DNA以进行测序,所述试验样品将分析是否存在染色非整倍体。

[0360] 虽然上文针对三倍体描述了IPC的产生,但应当理解,可产生IPC来反映其他部分非整倍体,包括例如各种区段扩增和/或缺失。因此,例如,在已知与特定扩增相关联的各种癌症(例如,与20Q13相关联的乳腺癌)的情况下,可产生结合这些已知扩增的IPC。

[0361] 测序方法

[0362] 如上所述,将所制备的样品(例如测序文库)作为用于识别拷贝数变异程序的一部分进行测序。可利用多种测序技术中的任意一种技术。

[0363] 一些测序技术是可商购获得的,诸如从Affymetrix公司(Sunnyvale,CA)获得的边杂交边测序平台,从454Life Sciences(Bradford,CT)、Illumina/Solexa(Hayward,CA)和Helicos Biosciences(Cambridge,MA)获得的边合成边测序平台以及从Applied Biosystems(Foster City,CA)获得的边连接边测序平台,如下所述。除了使用Helicos Biosciences的边合成边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 TORRENT™技术,以及例如由Oxford Nanopore Technologies开发的纳米孔测序。

[0364] 虽然自动桑格法(Sanger method)被认为是“第一代”技术,但包括自动桑格测序(Sanger sequencing)的桑格测序也可用于本文所述的方法中。另外的合适测序方法包括但不限于核酸成像技术,例如原子力显微镜(AFM)或透射电子显微镜(TEM)。下文更详细地描述了示例性的测序技术。

[0365] 在一个示例性但非限制性的实施方案中,本文所述的方法包括使用Illumina的边合成边测序和基于可逆终止子的测序化学方法来获得试验样品中核酸的序列信息,例如母体样品中的cfDNA、正在筛查癌症的受试者中的cfDNA或细胞DNA等(例如,如Bentley等人, Nature第6章:第53-59页[2009年]中所述)。模板DNA可以是基因组DNA,例如细胞DNA或cfDNA。在一些实施方案中,将来自分离细胞的基因组DNA用作模板,并且将其片段化成几百个碱基对的长度。在其他实施方案中,将cfDNA用作模板,并且不需要将其片段化,因为cfDNA是作为短片段存在的。例如,胎儿cfDNA在血流中以长度为约170个碱基对(bp)的片段循环(Fan et al., Clin Chem第56章:第1279-1286页[2010年],并且不需要在测序之前将该DNA片段化。Illumina的测序技术依赖于将片段化的基因组DNA连接到锚寡核苷酸所结合的平面的光学透明表面。将模板DNA进行末端修复以产生5'-磷酸化的平末端,并且使用克列诺片段(Klenow fragment)的聚合酶活性将单个A碱基添加到平末端磷酸化的DNA片段的3'末端。此添加制备用于连接至寡核苷酸衔接子的DNA片段,所述寡核苷酸衔接子在其3'末端具有单个T碱基的突出端以提高连接效率。衔接子寡核苷酸与流通池锚寡核苷酸互补(不会与重复扩增分析中的锚/锚定读段混淆)。在有限稀释条件下,将接头修饰的单链模板DNA添加到流通池中,并通过与锚寡核苷酸杂交进行固定。将连接的DNA片段延伸并桥式扩增以产生具有数亿簇的超高密度测序流通池,每个簇包含约1,000个相同模板的拷贝。在一个实施方案中,随机片段化的基因组DNA在其进行簇扩增之前使用PCR进行扩增。另选地,使用无扩增(例如,无PCR)基因组文库制备,并且仅使用簇扩增来富集随机片段化的基因组DNA

(Kozarewa等人, Nature Methods第6章:第291-295页[2009年])。使用稳健的四色DNA边合成边测序技术对模板进行测序,该技术采用具有可去除荧光染料的可逆终止子。使用激光激发和全内反射光学器件来实现高灵敏度荧光检测。将约几十至几百个碱基对的短序列读段与参考基因组进行比对,并且使用专门开发的数据分析管线软件来识别短序列读段与参考基因组的唯一映射。在第一次读取完成之后,模板可原位再生以能够从片段的相反端进行第二次读取。因此,可使用DNA片段的单末端或配对末端测序。

[0366] 本发明的各种实施方案可使用允许配对末端测序的边合成边测序。在一些实施方案中,Illumina的边合成边测序包括聚类片段。进行聚类是其中每个片段分子进行等温扩增的过程。在一些实施方案中,如此处所述的示例,片段具有连接至该片段的两端的不同接头,所述接头允许该片段与流通池泳道表面上的两种不同寡核苷酸杂交。片段还包括或连接至片段两端的两个索引序列,所述索引序列提供标记以在多重测序中识别不同的样品。在一些测序平台中,待测序的片段也称为插入序列。

[0367] 在一些具体实施中,用于在Illumina平台中进行聚类的流通池是具有泳道的载玻片。每个泳道是涂覆有两种类型寡核苷酸的引物基的玻璃通道。通过在表面上的两种类型寡核苷酸中的第一种寡核苷酸来实现杂交。该寡核苷酸与片段一个末端上的第一接头互补。聚合酶产生杂交片段的互补链。双链分子变性,并且原始模板链被洗掉。剩余的链与许多其他剩余的链平行,通过桥式应用进行克隆扩增。

[0368] 在桥式扩增中,一条链折叠,并且该链的第二末端上的第二接头区域与流通池表面上的第二种类型寡核苷酸杂交。聚合酶产生互补链,从而形成双链桥式分子。该双链分子变性,导致两个单链分子通过两种不同的寡核苷酸连接到流通池。然后反复重复该过程,并且同时在数百万个簇中进行,从而导致所有片段的克隆扩增。在桥式扩增之后,反向链被裂解并洗掉,仅留下正向链。封闭3'端以防止不需要的引发。

[0369] 在聚类之后,测序开始于延伸第一测序引物以生成第一次读取。在每次循环中,荧光标记的核苷酸竞争添加到正在增长的链中。基于模板的序列仅掺入一个荧光标记的核苷酸。在添加每个核苷酸后,簇由光源激发,并发出特征性荧光信号。循环次数决定了读段的长度。发射波长和信号强度决定了碱基判定。同时读取给定簇的所有相同的链。以大规模平行的方式对数以亿计的簇进行测序。在完成第一次读取时,将所读取的产物洗掉。

[0370] 在包括两种索引引物的方案的下一步骤中,引入索引1引物并将其与模板上的索引1区域杂交。索引区域提供对片段的识别,这可用于在多重测序过程中解复用样品。生成与第一次读取类似地索引1读取。在完成索引1读取之后,将所读取的产物洗掉,并将链的3'末端去保护。然后模板链折叠并结合至流通池上的第二寡核苷酸。以与索引1相同的方式读取索引2序列。然后在步骤结束时洗掉索引2读段产物。

[0371] 在读取两个索引后,通过使用聚合酶来延伸第二流通池寡核苷酸从而形成双链桥来启动读取2。该双链DNA变性,并且3'端被阻断。原正向链被切割并洗掉,留下反向链。读取2以引入读取2测序引物开始。与读取1一样,重复测序步骤直到实现所需长度。将读取2产物洗掉。该整个过程生成数百万个读段,表示所有片段。基于在样品制备期间引入的唯一索引来分离来自合并样品文库的序列。对于每个样品,对碱基判定的相似延伸的读段进行本地聚类。将正向和反向读段配对,从而产生邻接序列。将这些邻接序列与参考基因组进行比对以进行变异体识别。

[0372] 上述边合成边测序示例包括配对末端读段,其用于本发明所公开的方法的许多实施方案中。配对末端测序包括从片段两个末端进行的2次读取。当一对读段映射到参考序列时,可以确定两个读段之间的碱基对距离,然后可以使用该距离来确定从中获得读段的片段的长度。在一些情况下,跨越两个分组的片段将使其配对末端读取中的一次读取与一个分组进行比对,并且另一次读取与相邻分组进行比对。随着分组变得更长或读段变的更短,这种情况将变得越来越少。可使用各种方法来说明这些片段的分组成员身份。例如,在确定分组的片段大小频率时可以省略这些片段;可对相邻分组中的两者的这些片段进行计数;可将这些片段分配给包含两个分组的更大数量的碱基对的分组;或者可将这些片段分配给具有与每个分组中的碱基对的部分碱基对相关性的权重的两个分组。

[0373] 配对末端读段可使用不同长度的插入序列(即,待测序的不同片段大小)。作为本公开中的默认含义,配对末端读段用于指从各种插入序列长度中获得的读段。在一些情况下,为了区分短插入序列配对末端读段与长插入序列配对末端读段,后者也称为配偶对读段。在包括配偶对读段的一些实施方案中,首先将两个生物素连接衔接子连接到相对长的插入序列(例如,几kb)的两个末端。然后所述生物素连接衔接子连接插入序列的这两个末端以形成环化分子。然后可通过进一步使该环状化分子片段化来获得包含所述生物素连接衔接子的亚片段。然后可通过与上述短插入序列配对末端测序相同的程序对以相反序列顺序包含原始片段的两个末端的亚片段进行测序。使用Illumina平台的配偶对测序的更多细节示于以下URL的在线公布中,该URL以引用方式全文并入本文:res|.|illumina|.|com/documents/products/technotes/technote_nextera_matepair_dat a_processing.关于配对末端测序的附加信息可见于美国专利7601499和美国专利公布2012/0,053,063,这些专利关于配对末端测序方法和装置上的材料以引用方式并入。

[0374] 在DNA片段测序后,将预定长度(例如,100bp)的序列读段映射到已知的参考基因组或与其进行比对。所映射的或比对的读段及它们在参考序列上的对应位置也称为标签。在一个实施方案中,参考基因组序列是NCBI36/hg18序列,其可在万维网上genome dot ucsc dot edu/cgi-bin/hgGateway?org=Human&db=hg18&hgsid=166260105)处获得。另选地,参考基因组序列是GRCh37/hg19,其可在万维网上genome dot ucsc dot edu/cgi-bin/hgGateway处获得。公共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DNA数据库)。多种计算机算法可用于比对序列,包括但不限于BLAST(Altschul等人,1990年)、BLITZ(MPsrch)(Sturrock和Collins,1993年)、FASTA(Person和Lipman,1988年)、BOWTIE(Langmead等人,Genome Biology,10:R25.1-R25.10[2009年])或ELAND(Illumina公司,San Diego,CA,USA)。在一个实施方案中,对血浆cfDNA分子的克隆扩增拷贝的一个末端进行测序,并通过Illumina基因组分析仪的生物信息学比对分析进行处理,该分析仪使用核苷酸数据库(ELAND)软件的高效大规模比对。

[0375] 可使用其他测序方法和系统来获得序列读段。

[0376] 用于确定CNV的装置和系统

[0377] 通常使用各种计算机执行的算法和程序来执行测序数据的分析和从其推导的诊断。因此,某些实施方案采用包括存储在一个或多个计算机系统或其他处理系统中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的过程。本文所公开的实施方案还包括用于执行这些操作的装置。该装置可被专门构造用于所需目的,或者其可以由存储在计算

机中的计算机程序和/或数据结构来选择性地激活或重新配置的通用计算机(或计算机组)。在一些实施方案中,处理器组协同地(例如,经由网络或云计算)和/或并行地执行所述分析操作中的一些或全部操作。用于执行本文所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,微处理器诸如可编程设备(例如,CPLD和FPGA)和不可编程设备(诸如门阵列ASIC或通用微处理器)。

[0378] 此外,某些实施方案涉及有形和/或非暂态计算机可读介质或计算机程序产品,其包括用于执行各种计算机实现的操作的程序指令和/或数据(包括数据结构)。计算机可读介质的示例包括但不限于半导体存储器设备、磁性介质(诸如磁盘驱动器)、磁带、光学介质(诸如CD)、磁光介质以及被专门配置为存储和执行程序指令的硬件设备,诸如只读存储器设备(ROM)和随机存取存储器(RAM)。计算机可读介质可由最终用户直接控制,或者所述介质可由最终用户间接控制。直接控制介质的示例包括位于用户设施处的介质和/或不与其他实体共享的介质。间接控制介质的示例包括用户可经由外部网络和/或经由提供共享资源(诸如“云”)的服务间接访问的介质。程序指令的示例包括诸如由编译器产生的机器代码以及包含可由计算机使用解释器执行的较高级代码的文件。

[0379] 在各种实施方案中,在本发明所公开的方法和装置中采用的数据或信息以电子格式提供。此类数据或信息可包括来源于核酸样品的读段和标签、与参考序列的特定区域进行比对(例如,与染色体或染色体区段进行比对)的此类标签的计数或密度、参考序列(包括仅提供或主要提供多态性的参考序列)、染色体和区段剂量、判定(诸如非整倍体判定)、归一化染色体和区段值、染色体或区段和对应的归一化染色体或区段的配对、咨询建议、诊断等。如本文所用,以电子格式提供的数据或其他信息可用于存储在机器上并在机器之间传输。常规地,电子格式的数据以数字方式提供,并且可作为位和/或字节存储在各种数据结构、列表、数据库等中。数据可以电子方式、光学方式等体现。

[0380] 一个实施方案提供了用于生成输出的计算机程序产品,该输出指示在试验样品中是否存在非整倍体(例如,胎儿非整倍体)或癌症。计算机产品可包含用于执行上述用于确定染色体异常的方法中的任何一种或多种方法的指令。如所解释的,计算机产品可包括非暂态和/或有形计算机可读介质,该非暂态和/或有形计算机可读介质具有记录在其上的用于使处理器能够确定染色体剂量,并且在一些情况下,确定是否存在胎儿非整倍体的计算机可执行或可编译逻辑(例如,指令)。在一个示例中,计算机产品包括计算机可读介质,该计算机可读介质具有记录在其上的用于使处理器能够诊断胎儿非整倍体的计算机可执行或可编译逻辑(例如,指令),该计算机可执行或可编译逻辑包括:用于从来自母体生物样品的核酸分子的至少一部分中接收测序数据,其中所述测序数据包括计算的染色体和/或区段剂量;用于根据所接收的数据分析胎儿非整倍体的计算机辅助逻辑;和用于生成输出的输出程序,该输出指示所述胎儿非整倍体的存在、不存在或种类。

[0381] 可将来自所考虑样品的序列信息映射到染色体参考序列,以识别任意一条或多条所关注染色体中的每一条染色体的序列标签数目,以及识别所述任意一条或多条所关注染色体中的每一条染色体的归一化区段序列的序列标签数目。在各种实施方案中,例如,参考序列存储在数据库诸如关系数据库或对象数据库中。

[0382] 应当理解,对于无辅助的人类而言,执行本文所公开的方法的计算操作是不切实际的,甚至在大多数情况下是不可能的。例如,在没有计算装置辅助的情况下,将来自样品

的单个30bp读段映射到人类染色体中的任意一条染色体可能需要多年的努力。当然,由于可靠的非整倍体判定通常需要将数千(例如,至少约10,000)或甚至数百万个读段映射到一条或多条染色体,因此该问题是复杂的。

[0383] 可使用用于评价试验样品中所关注基因序列的拷贝数的系统来执行本文所公开的方法。该系统包括:(a) 测序仪,该测序仪用于接收来自该试验样品的核酸并提供来自该样品的核酸序列信息;(b) 处理器;和(c) 一个或多个计算机可读介质,所述一个或多个计算机可读介质具有存储在其上的指令,所述指令用于在所述处理器上执行以执行用于识别任何CNV(例如,染色体或部分非整倍体)的方法。

[0384] 在一些实施方案中,由具有存储在其上的计算机可读指令的计算机可读介质来指示方法,所述计算机可读指令用于执行用于识别任何CNV(例如,染色体或部分非整倍体)的方法。因此,一个实施方案提供了计算机程序产品,该计算机程序产品包括一个或多个计算机可读非暂态存储介质,所述一个或多个计算机可读非暂态存储介质具有存储在其上的计算机可执行指令,所述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,使得该计算机系统实现用于评估包含胎儿和母体游离核酸的试验样品中所关注序列的拷贝数的方法。该方法包括:(a) 接收通过对试验样品中的游离核酸片段进行测序而获得的序列读段;(b) 将所述游离核酸片段的所述序列读段与包含所述所关注序列的参考基因组进行比对,从而提供测试序列标签,其中将所述参考基因组分成多个分组;(c) 确定存在于该试验样品中的所述游离核酸片段的大小;(d) 基于从中获得所述标签的游离核酸片段的所述大小来对所述测试序列标签进行加权;(e) 基于(d)的所加权的标签来计算所述分组的覆盖度;以及(f) 根据所计算的覆盖度来识别所述所关注序列中的拷贝数变异。在一些具体实施中,对测试序列标签进行加权包括使覆盖度偏向测试序列标签,所述测试序列标签从具有试验样品中一个基因组的大小或大小范围特性的游离核酸片段中获得。在一些具体实施中,对测试序列标签进行加权包括将值1分配给从大小或大小范围的游离核酸片段中获得的标签,以及将值0分配给其他标签。在一些具体实施中,方法还包括在参考基因组(包括所关注序列)的分组中确定片段大小参数的值,该片段大小参数包括试验样品中片段大小短于或长于阈值的一定量的游离核酸片段。此处,识别所关注序列中的拷贝数变异包括使用片段大小参数的值以及(e)中计算的覆盖度。在一些具体实施中,系统被配置为使用上述各种方法和过程来评估试验样品中的拷贝数。

[0385] 在一些实施方案中,指令还可包括在提供母体试验样品的人类受试者的患者病历中自动记录与方法相关的信息,诸如染色体剂量和胎儿染色体非整倍体的存在与否。可由例如实验室、医师办公室、医院、健康维护组织、保险公司或个人病历网站来维护患者病历。此外,基于处理器实现的分析的结果,方法还可包括开处方、启动和/或改变从中获取母体试验样品的人类受试者的治疗。这可包括对从受试者中获取的另外的样品执行一个或多个另外的测试或分析。

[0386] 本发明所公开的方法还可使用计算机处理系统来执行,该计算机处理系统适于或被配置成执行用于识别任何CNV(例如,染色体或部分非整倍体)的方法。一个实施方案提供了适于或被配置为执行如本文所述的方法的计算机处理系统。在一个实施方案中,装置包括测序设备,该测序设备适于或被配置用于对样品中的核酸分子的至少一部分进行测序,以获得本文他处所述的序列信息的类型。该装置还可包括用于处理样品的部件。此类部件

在本文他处有所描述。

[0387] 序列或数据可直接或间接地输入到计算机中或存储在计算机可读介质上。在一个实施方案中,计算机系统直接联接到读取和/或分析来自样品的核酸序列的测序设备。来自此类工具的序列或其他信息经由计算机系统接口提供。另选地,由系统处理的序列由序列存储源诸如数据库或其他储存库提供。一旦可供处理装置使用,存储器设备或大容量存储装置就会至少暂时地缓冲或存储核酸序列。此外,该存储器设备可存储各种染色体或基因组等的标签计数。该存储器还可存储用于分析呈现序列或映射数据的各种例程和/或程序。此类程序/例程可包括用于执行统计分析的程序等。

[0388] 在一个示例中,用户将样品置于测序装置中。通过连接到计算机的测序装置来采集和/或分析数据。计算机上的软件允许数据采集和/或分析。可将数据存储、显示(经由监视器或其他类似设备)和/或发送到另一个位置。计算机可连接到互联网,该互联网用于将数据传输到由远程用户(例如,医师、科学家或分析员)使用的手持设备。应当理解,可在传输之前存储和/或分析数据。在一些实施方案中,采集原始数据并将其发送到将分析和/或存储该数据的远程用户或装置。传输可经由互联网进行,但也可经由卫星或其他连接进行。另选地,数据可存储在计算机可读介质上,并且该介质可发送到最终用户(例如,经由邮件)。远程用户可位于相同或不同的地理位置,包括但不限于建筑物、城市、州、国家或洲。

[0389] 在一些实施方案中,方法还包括采集关于多条多核苷酸序列的数据(例如读段、标记和/或参考染色体序列)并将该数据发送到计算机或其他计算系统。例如,计算机可连接到实验室装备,例如样品采集装置、核苷酸扩增装置、核苷酸测序装置或杂交装置。然后,计算机可采集由实验室设备收集的适用数据。数据可在任何步骤存储在计算机上,例如,在实时收集时、在发送之前、在发送期间或与发送同时、或在发送之后。数据可存储在可从计算机提取的计算机可读介质上。所采集或存储的数据可例如经由局域网或广域网(诸如互联网)从计算机传输到远程位置。在该远程位置处,可对所传输的数据执行各种操作,如下所述。

[0390] 在本文所公开的系统、装置和方法中,可存储、传输、分析和/或操纵的电子格式数据的类型如下:

[0391] 通过对试验样品中的核酸测序而获得的读段

[0392] 通过将读段与参考基因组或其他参考序列进行比对而获得的标签

[0393] 参考基因组或序列

[0394] 序列标签密度—参考基因组或其他参考序列的两个或更多个区域(通常为染色体或染色体片段)中每一个区域的标签计数或数目

[0395] 归一化染色体或染色体区段对特定所关注染色体或染色体区段的同一性

[0396] 从所关注染色体或区段中获得的染色体或染色体区段(或其他区域)和对应的归一化染色体或区段的剂量

[0397] 将染色体剂量判定为受影响、未受影响或无判定的阈值

[0398] 染色体剂量的实际判定

[0399] 诊断(与判定相关联的临床病症)

[0400] 对来源于判定和/或诊断的进一步测试的建议

[0401] 来源于判定和/或诊断的治疗和/或监测计划

[0402] 可使用不同的装置在一个或多个位置处获得、存储、分析和/或操纵这些各种类型的数据。处理选项范围很广。在最简单的情况下,在处理试验样品的位置处(如医生办公室或其他临床环境)存储和使用所有或大部分的这些信息。在最复杂的情况下,在一个位置处获得样品,在不同位置处对其进行处理并且任选地测序,在一个或多个不同位置处将读段进行比对并且作出判定,并且在又一个位置(其可以是获得样品的位置)处准备诊断、建议和/或计划。

[0403] 在各种实施方案中,用测序装置生成读段,然后将读段传输到远程位点,在该远程位点处对读段进行处理以产生非整倍体判定。在该远程位置处,例如,将读段与参考序列进行比对以产生标签,对所述标签计数并将其指定给所关注染色体或片段。同样在该远程位置处,使用相关联的归一化染色体或片段将计数转换成剂量。还在该远程位置处,使用剂量来生成非整倍体判定。

[0404] 在处理操作中,可在不同位置采用以下操作:

[0405] 样品采集

[0406] 测序前的样品处理

[0407] 测序

[0408] 分析序列数据并推导非整倍体判定

[0409] 诊断

[0410] 向患者或医疗服务人员报告诊断和/或判定

[0411] 开发用于进一步治疗、测试和/或监测的计划

[0412] 执行计划

[0413] 咨询

[0414] 这些操作中的任意一个或多个操作可以是自动化的,如本文他处所述。通常,序列数据的测序和分析以及推导非整倍体判定将通过计算来执行。其他操作可手动或自动执行。

[0415] 可执行样品采集的位置的示例包括医务人员的办公室、诊所、患者的家(其中提供了样品采集工具或试剂盒)和移动医疗车。可执行测序前的样品处理的位置的示例包括医务人员的办公室、诊所、患者的家(其中提供样品处理装置或试剂盒)、移动医疗车和非整倍体分析供应商的设施。可执行测序的位置的示例包括医务人员的办公室、诊所、医务人员的办公室、诊所、患者的家(其中提供样品测序装置和/或试剂盒)、移动医疗车和非整倍体分析供应商的设施。进行测序的位置可设置有用于以电子格式传输序列数据(通常为读段)的专用网络连接。此类连接可以是有线的或无线的,并且具有数据并且可被配置为将数据发送到位点,在该位点处可将数据在传输到处理站点之前处理和/或聚合。数据聚合器可由健康组织诸如健康维护组织(HMO)维护。

[0416] 分析和/或推导操作可在任何前述位置处执行,或者另选地在专用于计算和/或服务于分析核酸序列数据的另外的远程位点处执行。此类位置包括例如簇,诸如通用服务器群、非整倍体分析服务业务的设施等。在一些实施方案中,用于执行分析的计算装置被租用或租借。计算资源可以是可互联网访问的处理器集合的一部分,诸如通称为云的处理资源。在一些情况下,由彼此关联或非关联的并行或大规模并行处理器组来执行计算。可使用分布式处理诸如簇计算、网格计算等来完成该处理。在此类实施方案中,计算资源的簇或网格

共同形成由多个处理器或计算机构成的超级虚拟计算机,所述多个处理器或计算机共同作用以执行本文所述的分析和/或推导。这些技术以及更常规的超级计算机可用于处理如本文所述的序列数据。每项技术都是依赖于处理器或计算机的并行计算的形式。就网络计算而言,这些处理器(通常是全部计算机)通过网络(私人、公共或互联网)通过常规网络协议诸如以太网连接。相比之下,超级计算机具有通过本地高速计算机总线连接的许多处理器。

[0417] 在某些实施方案中,诊断(例如,胎儿患有唐氏综合征(Downs syndrome)或患者患有特定类型的癌症)在与分析操作相同的位置处生成。在其他实施方案中,诊断在不同位置处执行。在一些示例中,报告诊断在获取样品的位置处执行,虽然不一定是这种情况。可生成或报告诊断和/或执行计划开发的位置的示例包括医务人员的办公室、诊所、计算机可访问的互联网站和具有与网络的有线或无线连接的手持设备(诸如手机、平板电脑、智能电话等)。执行咨询的位置的示例包括医务人员的办公室、诊所、计算机可访问的互联网站、手持设备等。

[0418] 在一些实施方案中,样品采集、样品处理和测序操作在第一位置处进行,并且分析和推导操作在第二位置处进行。然而,在一些情况下,在一个位置处(例如,医务人员的办公室或诊所)进行样品采集,并且在不同位置处进行样品处理和测序,所述不同位置任选地为进行分析和推导的相同位置。

[0419] 在多个实施方案中,上文列出的操作的序列可由用户或实体启动样品采集、样品处理和/或测序来引发。在一个或多个这些操作已开始执行之后,其他操作可自然地紧随其后。例如,测序操作可以使读段被自动采集并发送到处理装置,然后该处理装置通常自动地并且可能在没有进一步用户干预的情况下执行序列分析和非整倍体的推导操作。在一些具体实施中,然后将该处理操作的结果(可能会重新格式化为诊断)自动递送到处理将信息报告给健康专业人员 and/或患者的系统部件或实体。如所解释的,此类信息也可进行自动处理以产生治疗、测试和/或监测计划,可能还有咨询信息。因此,启动早期操作可引发端对端顺序,其中为健康专业人员、患者或其他相关方提供可用于根据身体状况采取行动的诊断、计划、咨询和/或其他信息。即使整个系统的各部分在物理上是分开的并且可能远离例如样品和序列装置的位置,也可实现这一点。

[0420] 图17示出了用于从试验样品中产生判定或诊断的分散系统的一个具体实施。样品采集位置01用于从患者(诸如妊娠女性或推定的癌症患者)中获得试验样品。然后将样品提供给处理和测序位置03,在此处可如上所述对试验样品进行处理和测序。位置03包括用于处理样品的装置以及用于对经处理的样品进行测序的装置。如本文他处所述,测序的结果是读段的集合,其通常以电子格式提供并提供给网络诸如互联网,其由图17中的参考标号05指示。

[0421] 将序列数据提供给执行分析和判定生成的远程位置07。该位置可包括一个或多个强大的计算设备,诸如计算机或处理器。在位置07处的计算资源已完成其分析并根据所接收的序列信息生成判定之后,将该判定中继回到网络05。在一些具体实施中,在位置07处不仅生成判定,而且还生成相关联的诊断。然后将该判定和/或诊断通过网络传输并返回至如图17所示的样品采集位置01。如所解释的,这仅仅是关于如何在各种位置之间划分与生成判定或诊断相关联的各种操作的许多变型中的一种变型。一种常见的变异体包括在单个位置处提供样品采集、处理和测序。另一种变型包括在与分析和判定生成相同的位置处提供

处理和测序。

[0422] 图18阐明了用于在不同位置处执行各种操作的选项。在图18中所示的最细化意义上,以下操作中的每一个操作在单独的位置处执行:样品采集、样品处理、测序、读段比对、判定、诊断和报告和/或计划开发。

[0423] 在聚集这些操作中的一些操作的一个实施方案中,样品处理和测序在一个位置进行,而读段比对、判定和诊断在单独的位置进行。参见图18中由参考字符A标识的部分。在图18中由字符B标识的另一个具体实施中,样品采集、样品处理和测序均在相同位置处执行。在该具体实施中,在第二位置处执行读段比对和判定。最后,在第三位置处执行诊断和报告和/或计划开发。在图18中由字符C所示的具体实施中,在第一位置处执行样品采集,在第二位置处一起执行样品处理、测序、读段比对、判定和诊断,并且在第三位置处执行报告和/或计划开发。最后,在图18中标为D的具体实施中,在第一位置处执行样品采集,在第二位置处执行样品处理、测序、读段比对和判定全部,并且在第三位置处执行诊断和报告和/或计划管理。

[0424] 一个实施方案提供了用于确定包含胎儿和母体核酸的试验样品中是否存在非整倍体的系统,该系统包括测序仪,该测序仪用于接收核酸样品并提供来自该样品的胎儿和母体核酸序列信息;一个或多个处理器,所述一个或多个处理器被配置为:(a) 确定该试验样品的胎儿分数值,其中该试验样品的该胎儿分数指示该试验样品中胎源游离核酸片段的相对量;(b) 由计算机系统接收通过该试验样品中的所述游离核酸片段进行测序而获得的序列读段;(c) 由计算机系统将该所述游离核酸片段的所述序列读段与包含所关注序列的参考基因组进行比对,从而提供序列标签;(d) 由计算机系统确定该参考基因组的至少一部分的所述序列标签的覆盖度;以及(e) 基于在(d)中确定的所述序列标签的覆盖度和在(a)中确定的胎儿分数来确定该试验样品在排除区域内,其中该排除区域由至少胎儿分数检测限(LOD)曲线限定,其中该胎儿分数LOD曲线随覆盖度值而变化,并且指示在给定不同覆盖度的情况下实现检测标准所需的胎儿分数的最小值。

[0425] 在本文提供的系统中的任意一种系统的一些实施方案中,测序仪被配置为执行下一代测序(NGS)。在一些实施方案中,测序仪被配置为使用具有可逆染料终止子的边合成边测序来执行大规模并行测序。在其他实施方案中,测序仪被配置为执行边连接边测序。在其他实施方案中,测序仪被配置为执行单分子测序。

[0426] 在本文提供的系统中的任意一种系统的一些实施方案中,一个或多个处理器被编程为执行上述各种方法。

[0427] 本公开的另一方面涉及计算机程序产品,该计算机程序产品包括存储程序代码的非暂态机器可读介质,该程序代码在由计算机系统的一个或多个处理器执行时,使得该计算机系统:(a) 确定试验样品的胎儿分数值,其中该试验样品的胎儿分数指示该试验样品中胎源游离核酸片段的相对量;(b) 由计算机系统接收通过该试验样品中的所述游离核酸片段进行测序而获得的序列读段;(c) 由计算机系统将该所述游离核酸片段的所述序列读段与包含所关注序列的参考基因组进行比对,从而提供序列标签;(d) 由计算机系统确定该参考基因组的至少一部分的所述序列标签的覆盖度;以及(e) 基于在(d)中确定的所述序列标签的覆盖度和在(a)中确定的胎儿分数来确定该试验样品在排除区域内,其中该排除区域由至少胎儿分数检测限(LOD)曲线限定,其中该胎儿分数LOD曲线随覆盖度值而变化,并且

指示在给定不同覆盖度的情况下实现检测标准所需的胎儿分数的最小值。

[0428] 在本文提供的系统的一些实施方案中,计算机程序产品包括存储程序代码的非暂态机器可读介质,该程序代码将由一个或多个处理器执行以执行上述各种方法。

[0429] 实验

[0430] 实施例1

[0431] 该实施例的目的在于阐述如何使用上述方法来获得LOD曲线。

[0432] 研究设计

[0433] 对于该研究,将在VeriSeq NIPT精度研究 (BRIGID-0147VeriSeq NIPT:精度研究方案和BRIGID-0166VeriSeq NIPT—精度研究报告) 中生成的数据中的一部分用于确定VeriSeq NIPT溶液系统的检测限。

[0434] 在VeriSeq NIPT精度研究中,将来自21-三倍体和非妊娠血浆池的合并cfDNA组合,以产生5%胎儿分数21-三倍体受影响池,其通过VeriSeq NIPT测定法进行处理。男性母体池也通过VeriSeq NIPT测定按原样进行处理。

[0435] 表1示出了用于LOD研究的精度研究设计的实验室内精度部分。实验室内精度部分由2台Hamilton仪器、2台NextSeq仪器和三个试剂批次组成,总共在6天内运行12次。

[0436] 表1—精度研究的实验室内精度部分

[0437]

实验室内精度											
位点 3											
Hamilton 3						Hamilton 4					
NextSeq 3			NextSeq 4			NextSeq 4			NextSeq 3		
试剂批次						试剂批次					
1	2	3	1	2	3	1	2	3	1	2	3
运行	运行	运行	运行	运行	运行	运行	运行	运行	运行	运行	运行
10	12	14	16	18	20	11	13	15	17	19	21

[0438] 自动化和测序仪器与一致的操作者配对。将操作者和仪器变型组合。

[0439] 用试剂批次1和2生成的数据用于建立检测限。通过将来自T21的序列数据 (5% FF) 与非妊娠女性的序列数据组合,我们生成了具有不同胎儿分数的T21样品的合成序列数据。针对5个不同水平的测序深度 (2百万、4百万、6百万、8百万和1千万个唯一比对读段覆盖度—参见表2) 重复该过程。在宽范围内的精细网格上选择稀释点,以覆盖所有水平的测序深度 (1.25% -4.5%) 的预期检测限,步长为0.25%。

[0440] 对于每个稀释点和每个覆盖度水平,生成20个重复的合成样品 (更详细的方法描述参见DEV报告-0072,对数似然比得分和检测限预测的数学模型)。

[0441] 使用计算机模拟生成的稀释系列,使用如BRIGID-0150 (VeriSeq NIPT溶液设计验证方案检测限) 中所述的概率单位回归方法针对每个覆盖度水平建立T21的LoD。此外,使用T21的LOD结果以及LOD T21与LOD T13和T18之间的已知关系来确定T13和T18的LOD,如DEV报告-0072中所述。

[0442] 结果

[0443] 通过R&D Clinical Aneuploidy Detection and Analysis Suite (cADAS) v3.2来计算每个重复运行的胎儿分数的定量得分、对数似然比 (LLR) 和估计值。

[0444] 样品过滤和异常值

[0445] 将所有无模板对照 (NTC) 样品排除在分析之外。随后去除这些样品, 得到564个样品, 其中N=48T21池和N-48非妊娠池。将这些样品进一步用于计算机模拟稀释分析。其余样品不直接使用, 而是用作分析的填料。非妊娠样品中的一个样品被标记为NES_FF_QC失败, 该QC失败被忽略, 并且该样品纳入分析, 因为样品是非妊娠的 (其中不具有胎儿DNA), 因此预期在该度量中失败。

[0446] 计算机模拟稀释质量验证

[0447] 由男性血浆样品和女性血浆样品的合并混合物来制备受影响样品, 其中每种样品约占50%。因此, Y染色体DNA存在于样品中并可用于验证计算机模拟稀释过程。图19示出了根据稀释分数的合成生成的样品的染色体Y覆盖度 (左图) 和FF级分估计值 (右图)。两个线性曲线图均示出FF (初始非妊娠FF=0%, 并且受影响FF=5%) 和计算机模拟稀释度之间的协调值。

[0448] 确定T21三倍体测试的检测限

[0449] 根据对指定覆盖度水平的实验观察来确定LOD的程序描述于BRIGID 0150vA中。(VeriSeq NIPT溶液设计验证方案检测限)。

[0450] 每个批次和每个覆盖度水平的T21的测量LOD汇总于表2中。

[0451] 表2检测限

覆盖度	LOD				
	精度研究 (BRIGID-166) 批次			平均值	最大 (最保守)
	批次 1	批次 2	批次 3		
[0452] 2M	4.3	4.1	4.4	4.3	4.4
4M	3.5	3.1	3.1	3.2	3.5
6M	3.1	2.6	2.6	2.8	3.1
8M	2.9	2.2	2.1	2.4	2.9
10M	2.3	2	2.1	2.1	2.3

[0453] 检测限对比覆盖度

[0454] DEV报告-0072中描述的理论模型 (对数似然比得分和检测限预测的数学模型) 预测以Log-Log标度计的LOD对比覆盖度是直线。图20示出了LOD对比覆盖度的线性拟合的结果。我们观察到实验数据与预测行为十分匹配。

[0455] 拟合的LOD对比覆盖度线预测任何覆盖度的检测限。预测LOD与观察LOD叠加的结果在图21中示出。

[0456] 平均情况的检测限

[0457] 样品的总群具有可变的覆盖度。图22示出了使用V2版本的VeriSeq NIPT测定法处理的大量样品 (N=14,400) 中的覆盖度分布 (NES)。以根据覆盖度分布范围内的覆盖度的预期LOD来计算总群的预期LOD。

[0458]
$$\int_{2M}^{20M} LOD(NES) * p(NES) dNES$$
 其中LOD (NES) 是根据覆盖度的LOD, 并且p (NES) 是

图22所示的NES的概率密度函数。T21的所得平均情况LOD示于表3中。

[0459] 表3平均群体覆盖度的检测限

[0460]	度量	批次1	批次2	批次3
	LOD21	2.81 %	2.40 %	2.43 %

[0461] 其他非整倍体T13、T18的检测限

[0462] 如在先前的研究(对数似然比得分和检测限预测的数学模型)中所述,可根据染色体中的至少一条染色体的已知LOD推断不同染色体的非整倍体的检测限。使用来自先前研究的关系,我们可使用先前确定的倍增系数来推断T13三倍体和T18三倍体的LOD,结果示于表4和表5中。

[0463] 表4根据验收标准推断的T13和T18评估的平均群体覆盖度的检测限

[0464]	非整倍体	观察LoD	LoD系数	根据T21推断的LoD
	T21	2.81 %	100 %	2.81 %
	T18	NA	81 %	2.27 %
	T13	NA	72 %	2.02 %

[0465] 表5精度%CV验收标准

[0466]	度量	%FF	通过标准	通过/失败
	LOD13	2.02 %	LOD<4 %	通过
	LOD18	2.27 %	LOD<4 %	通过
	LOD21	2.81 %	LOD<4 %	通过

[0467] 结论

[0468] 已确定VeriSeq NIPT系统对于正常母体二倍体cfDNA背景下的胎儿cfDNA的胎儿分数(三倍体13/18/21)的检测限(LOD)。检测限值汇总于表6中。对于所有测试的非整倍体,发现观察LOD低于测试要求文件中规定的值。基于表6中呈现的数据,所确定的检测限通过了对所有靶向非整倍体的另一项研究的接受标准。

[0469] 表6:非整倍体胎儿分数LOD

[0470]	非整倍体	LOD (%FF)
	T13	2.02 %
	T18	2.27 %
	T21	2.81 %

[0471] 实施例2

[0472] 实施例2示出了标记为上述NES和胎儿分数LOD曲线方法的常规方法的各种性能度量的经验数据。

[0473] 图23示出了根据观察胎儿分数的NES覆盖度。两步覆盖度阈值技术用于排除样品。落入排除区域的各种样品可见于图中。大多数排除的样品具有在0%至20%之间的胎儿分数。并且它们的覆盖度受两个阈值水平的限制。

[0474] 图24示出了使用由LOD曲线和读段阈值限定的排除区域的数据排除。左图的读段阈值为2百万个读段。右图示出了使用1百万个读段的覆盖度阈值的数据。如图所示,在期望排除较少样品的条件下可降低覆盖度阈值以排除较少样品。许多排除的样品具有相对高的读段覆盖度和相对低的观察胎儿分数。相比之下,在图23中,许多排除的样品具有相对高的胎儿分数和相对低的覆盖度。

[0475] 图25示出了现有方法和上述LOD QC方法第一次运行和第二次运行的通过率和失败率。数据显示,这两种方法都具有相似的通过率和失败率。现有方法的最终失败率为0.42%。并且LOD QC方法的最终失败率为0.38%。

[0476] 图26示出了通过两步阈值法排除并通过标记为2604的胎儿分数LOD曲线方法挽救的数据。通过LOD QC方法排除并通过常规方法挽救的样品在区域2602中示出。图26的左图示出了与LOD曲线组合的2百万个读段的覆盖度阈值的数据。右图示出了1百万个读段的较低覆盖度阈值的挽救数据。使用左图2百万个读段的覆盖度阈值,LOD QC方法排除了61个另外的样品。但是当覆盖度阈值降低到1百万个读段时,通过LOD QC方法排除的样品就少了151个。

[0477] 图27示出了通过LOD QC方法挽救的样品。左图示出了在区域2702中通过常规方法排除并且通过LOD QC方法挽救的样品。在重新运行之后,这些样品中的80%落入包含区域2704内。而10%的样品仍留在相同的区域2702中。9.7%的样品落在LOD曲线下方进入到排除区域中。右图示出了通过应用LOD曲线的1百万覆盖度阈值挽救的样品。区域2708中示出了通过常规方法排除的通过1百万个读段阈值挽救的样品。在重新运行样品之后,80%的样品移动到包含区域2710中,13.7%的样品仍留在相同区域2708中,并且6.3%的样品落在LOD曲线下方进入到排除区域中。该图示出了单独的LOD方法或与覆盖度阈值方法组合的LOD方法可挽救原本会被排除的很大百分比的样品。在实施过程中,LOD QC方法有助于避免重新运行被挽救的样品的需要,从而节省了用于CNV检测的时间、成本和资源。

[0478] 图28示出了在区域2802中通过LOD QC方法排除的数据。在重新运行之后,少数样品移动到LOD曲线上方进入到包含区域2708中。更多数据停留在相同2802区域中或进一步移动到LOD曲线2708下方。来自图27和图28的结果表明,LOD QC方法包括实际上满足QC条件的许多样品,并且排除不能满足QC标准的样品,即使在重新运行时也是如此。这说明LOD QC方法可智能且正确地分离阳性和阴性样品。

[0479] 图29示出了现有方法和LOD QC方法两次运行的通过率和失败率。总体而言,对于两种方法,第一次运行和第二次运行中的通过率和失败率都是类似的。

[0480] 图30示出75%置信度LOD曲线挽救了为21-三倍体(T21)假阳性的样品3002。该曲线还挽救了为18-三倍体(T18)假阴性的样品3004。这导致T21假阳性减少100%,并且T18假阴性减少50%。说明LOD QC方法可增加灵敏度和特异性两者。

[0481] 图31示出,对于模拟的T21样品,可以99.7%的灵敏度检测通过LOD QC的样品,而可以88.1%的灵敏度检测通过LOD QC的样品。无论LOD QC状态如何,所有样品的总体灵敏度均为99.7%。

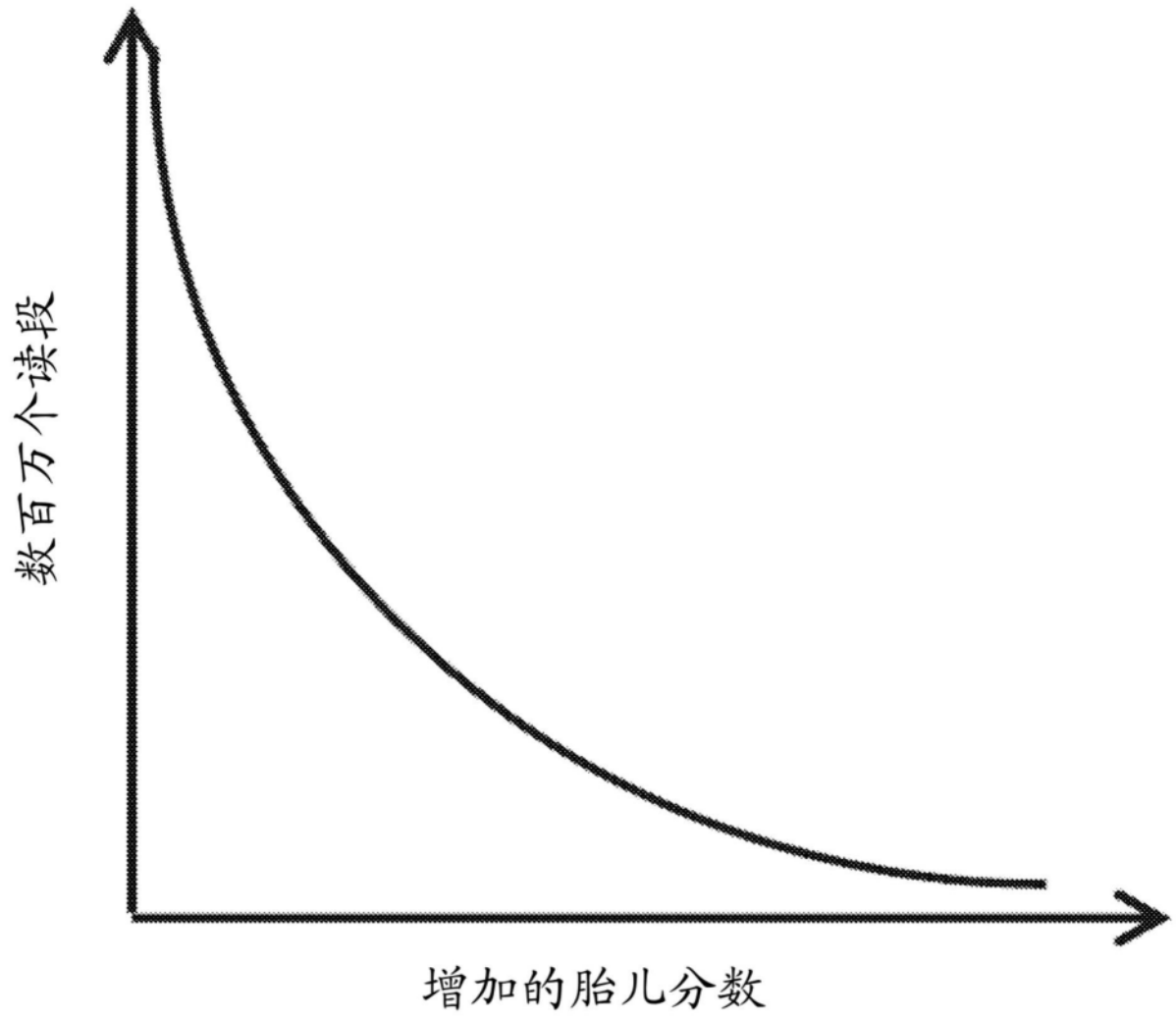


图1

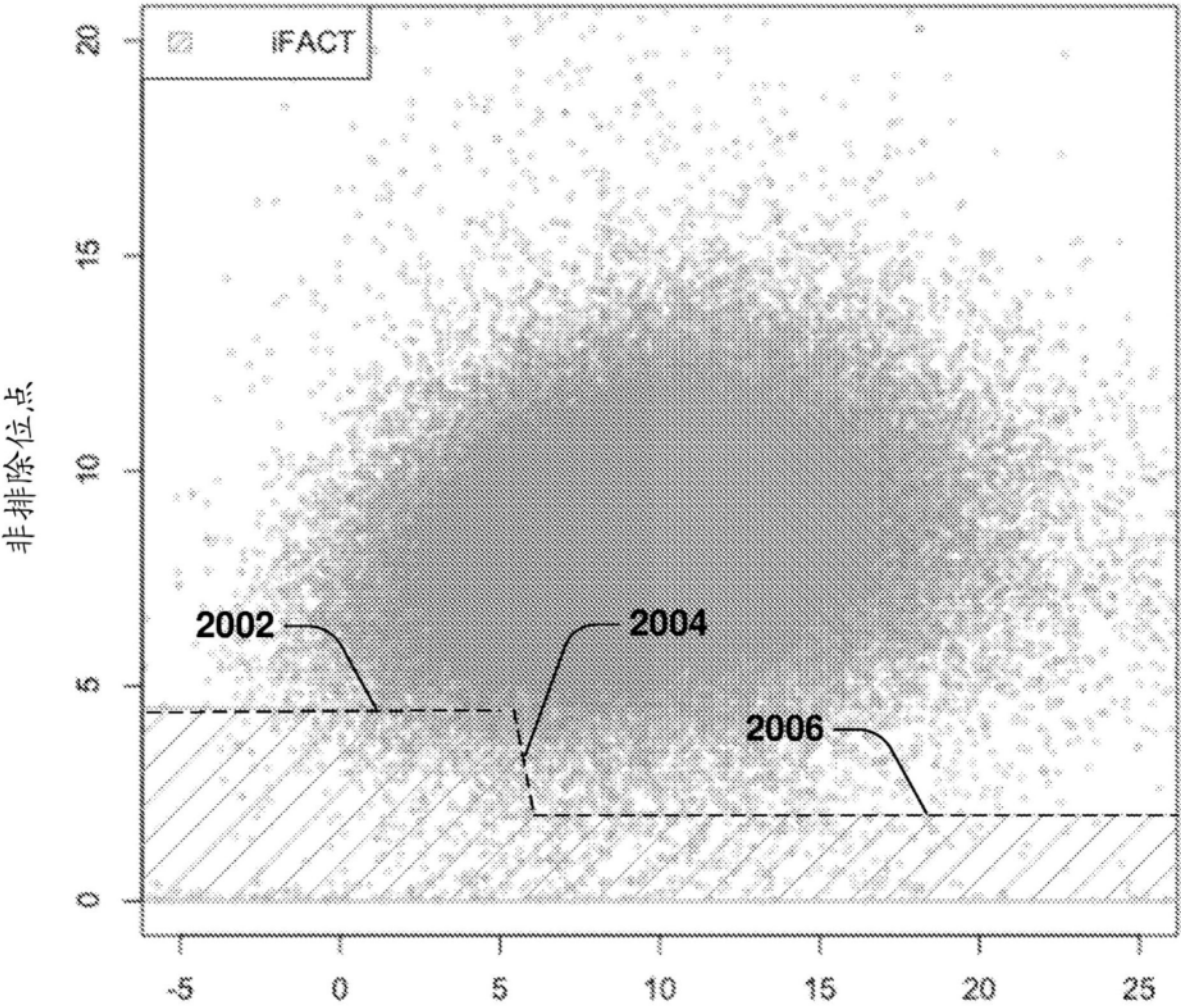


图2

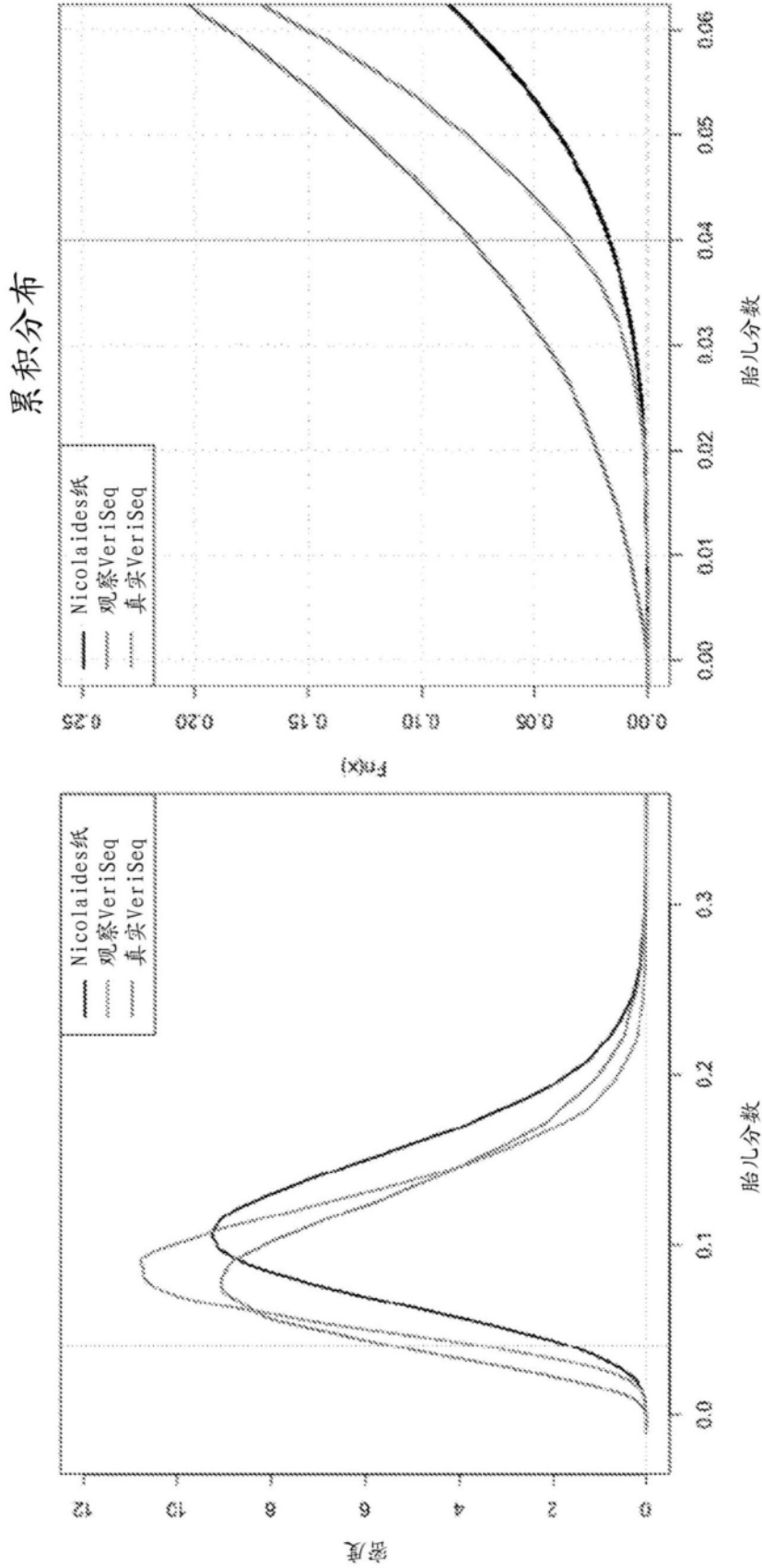


图3

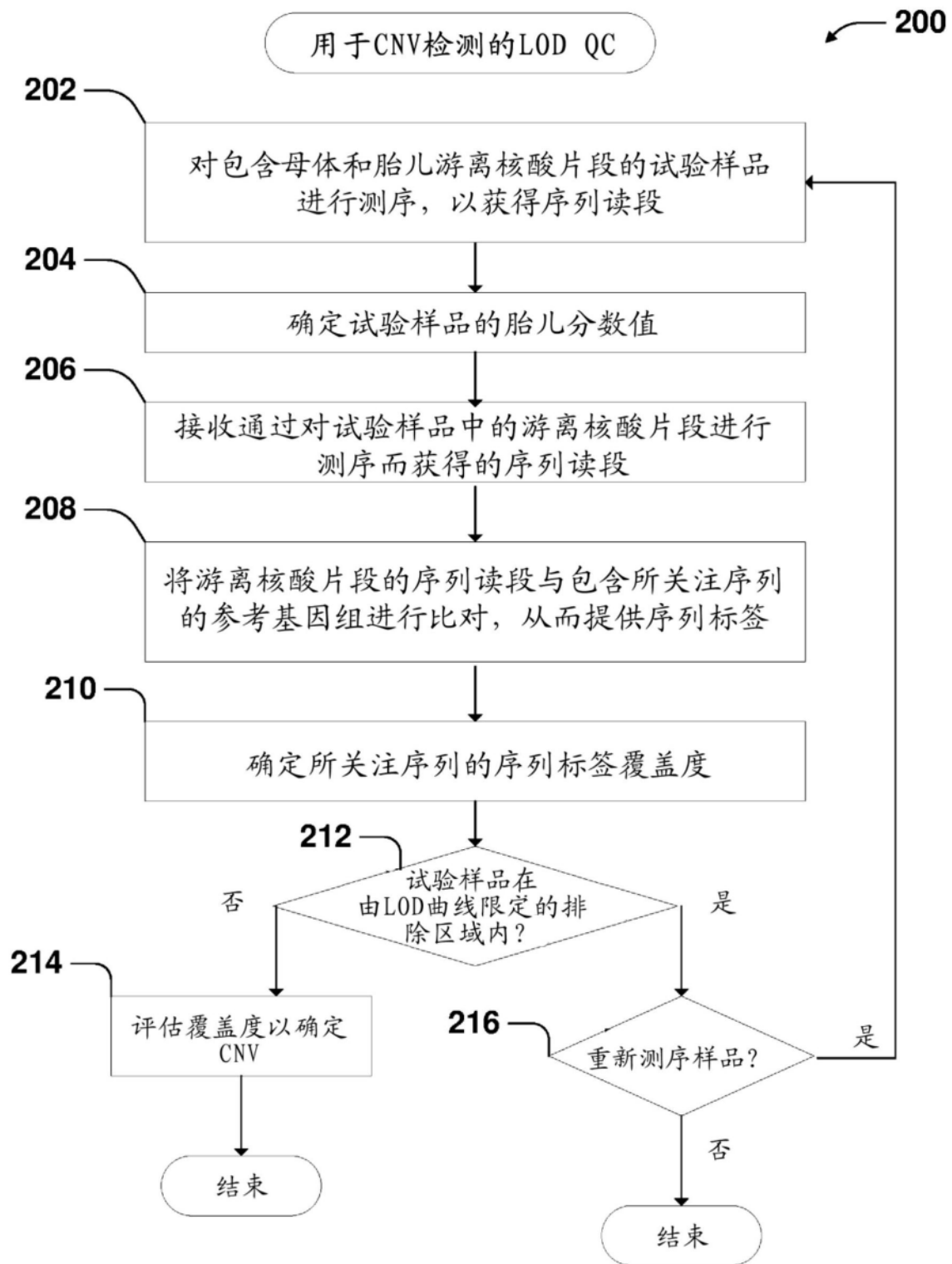


图4A

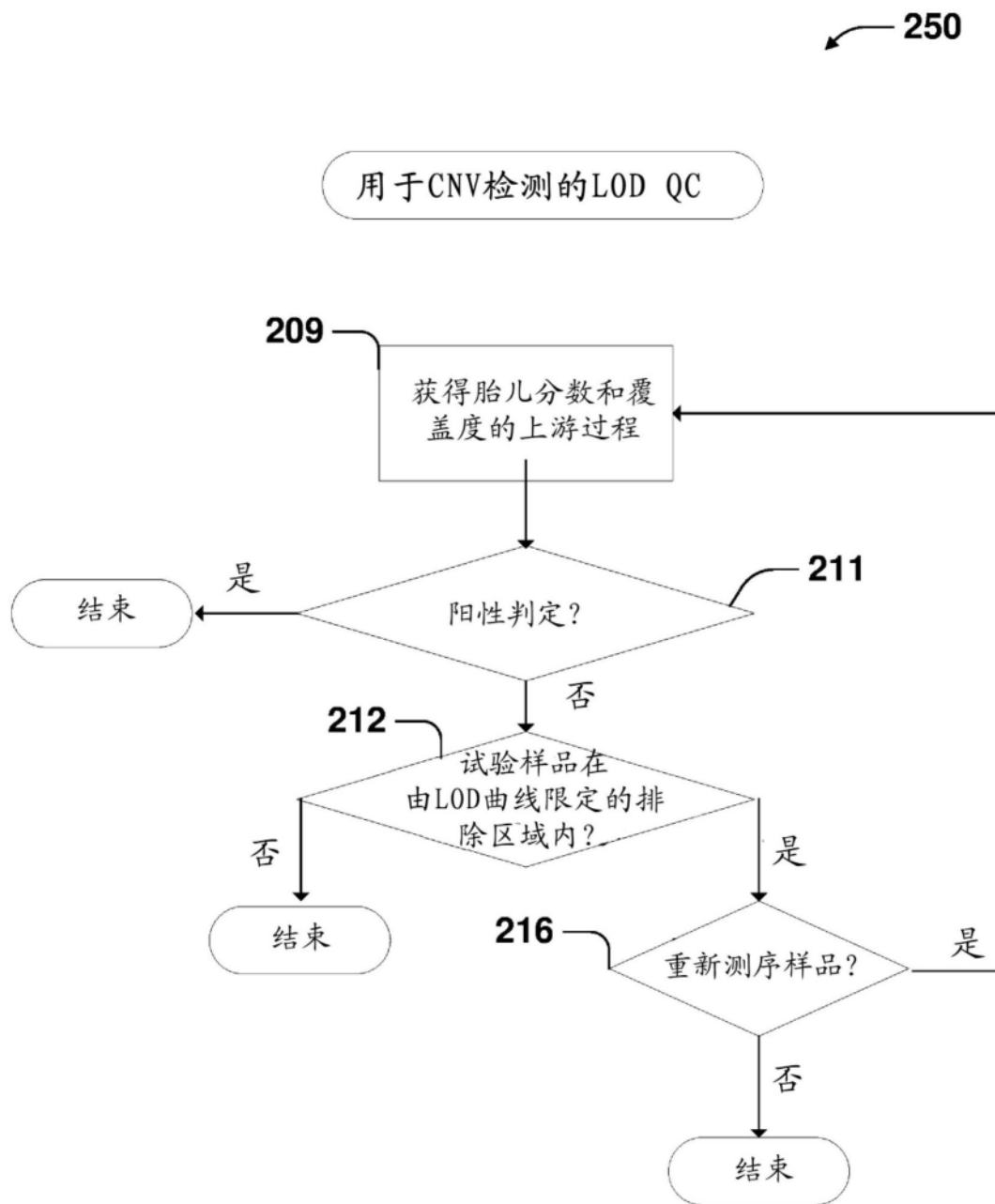


图4B

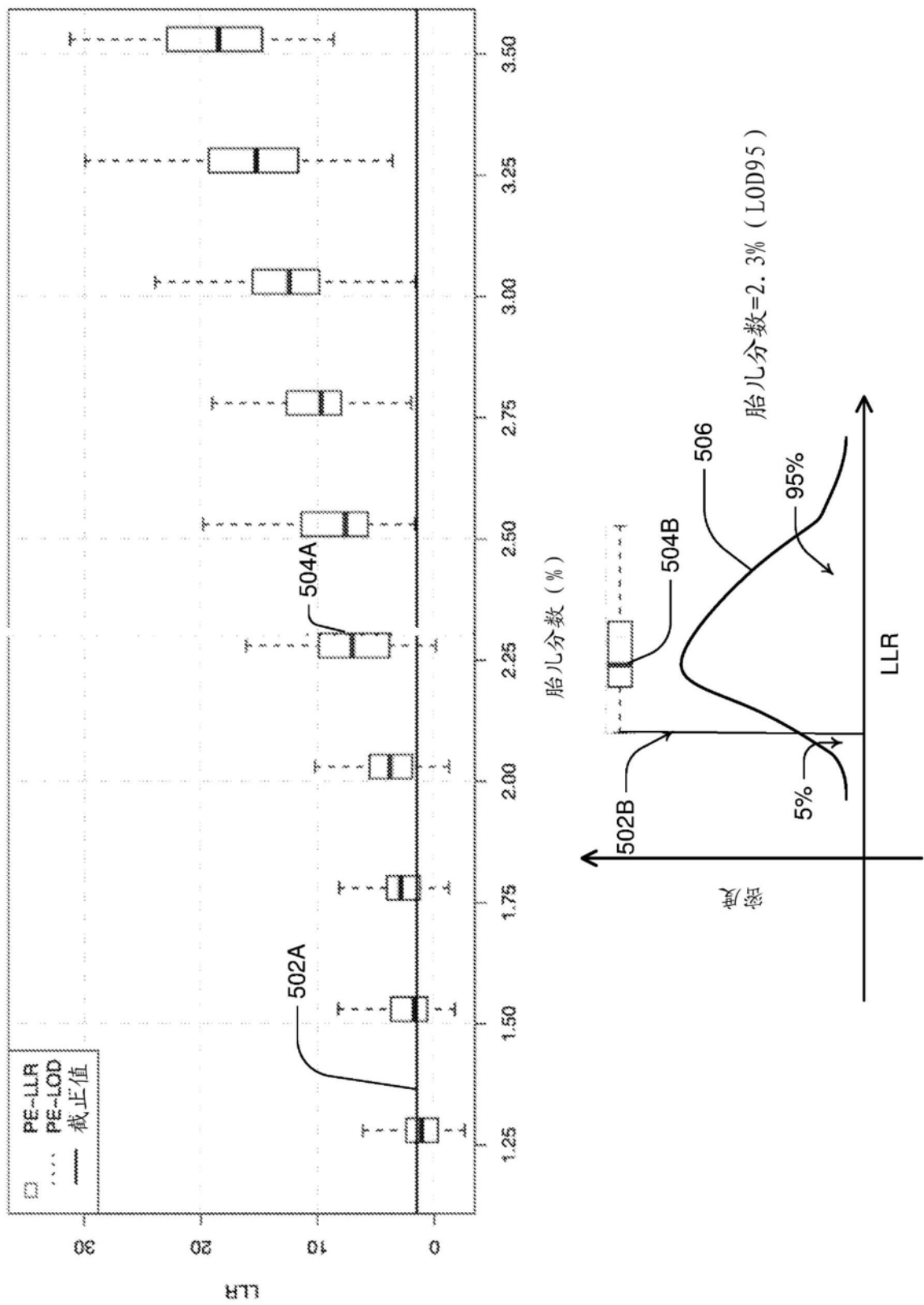


图5

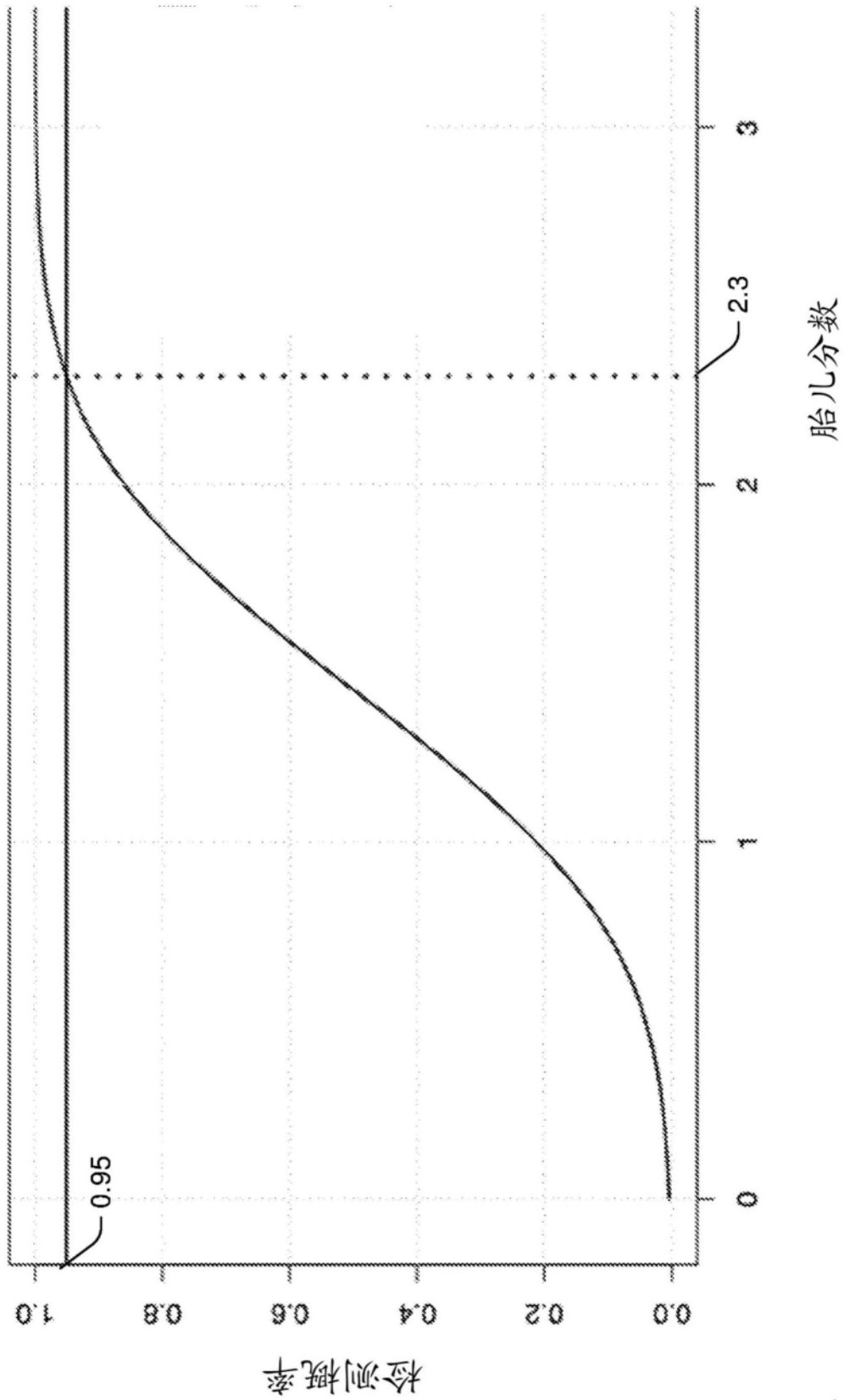


图6

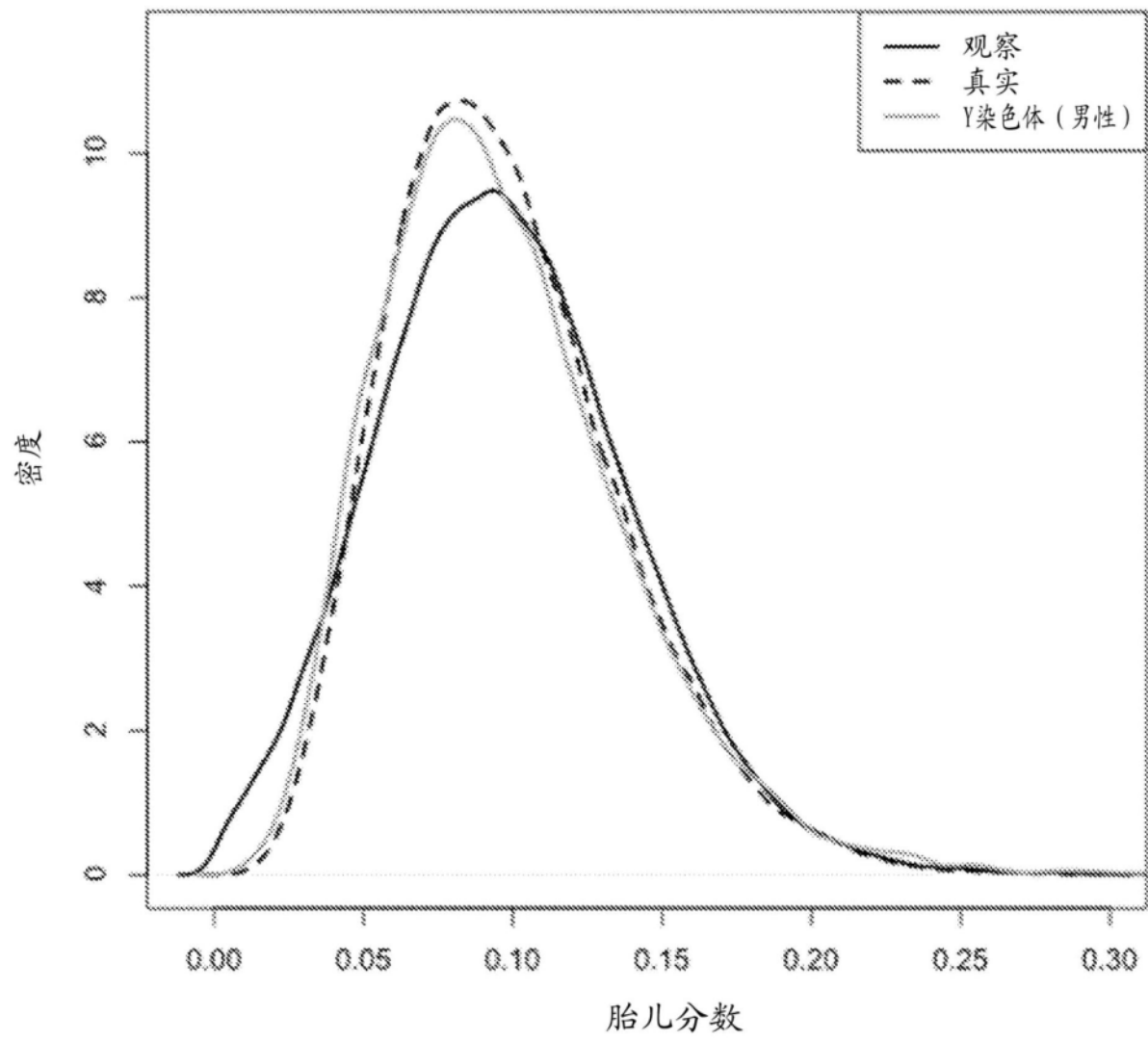


图7

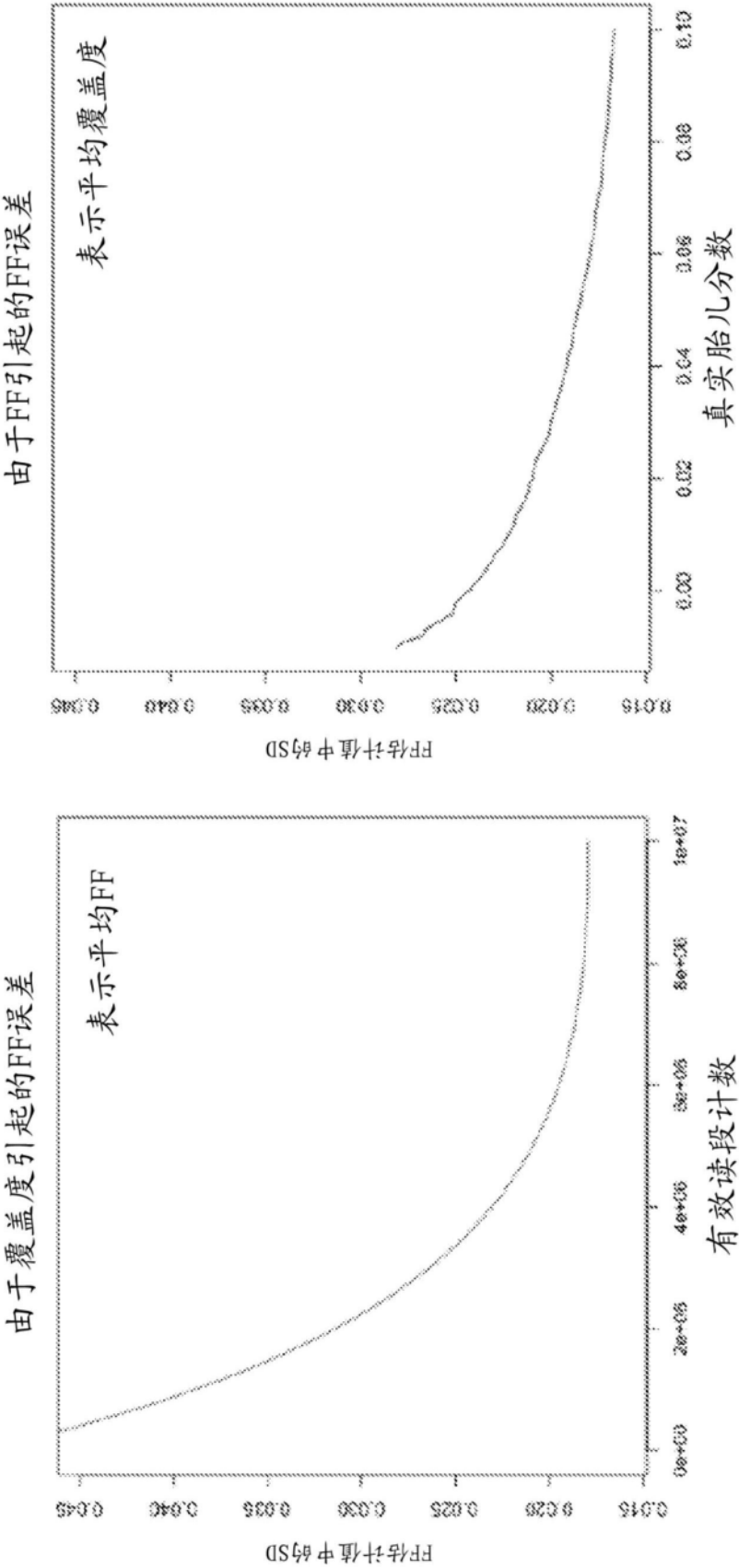


图8

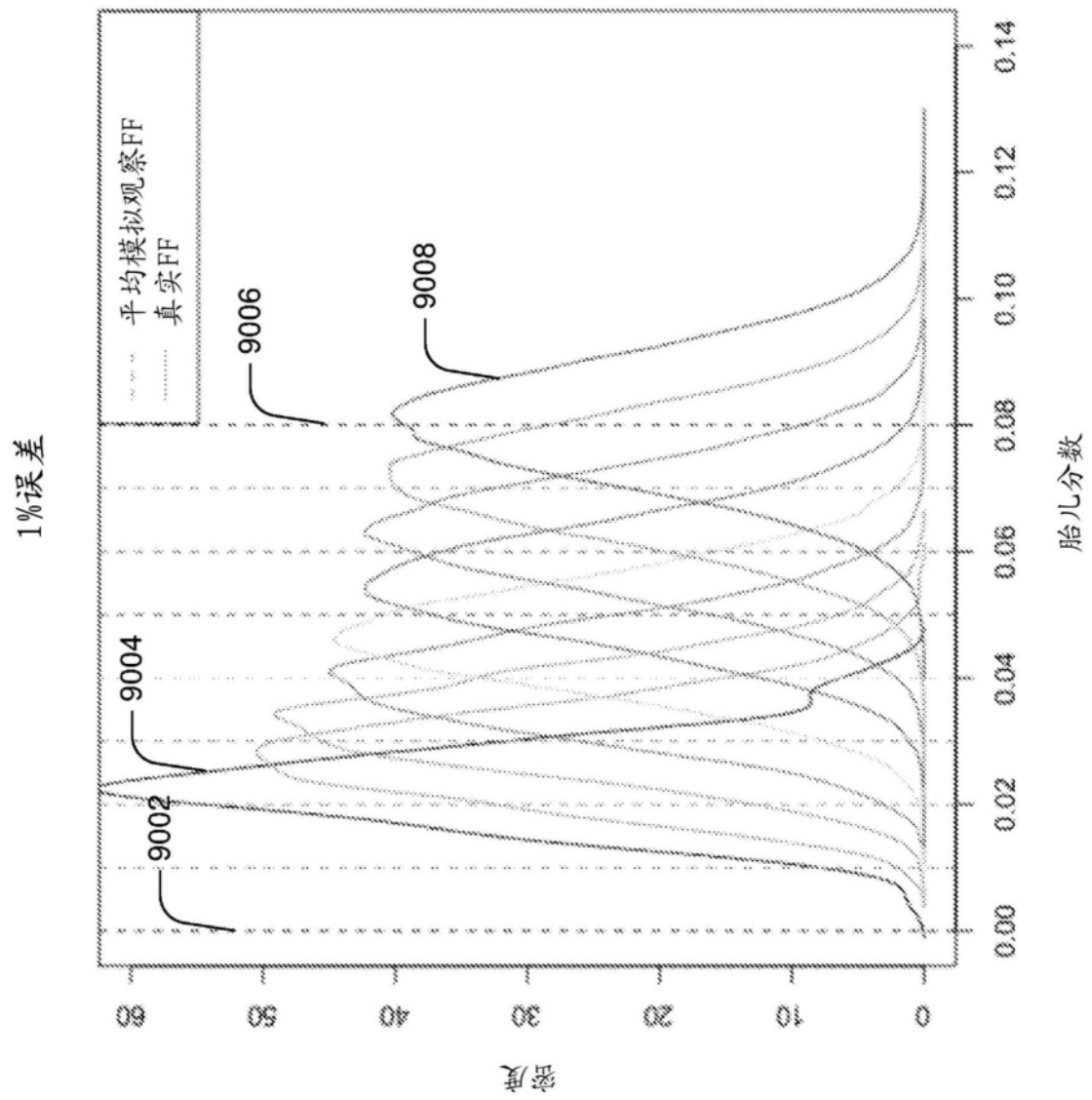
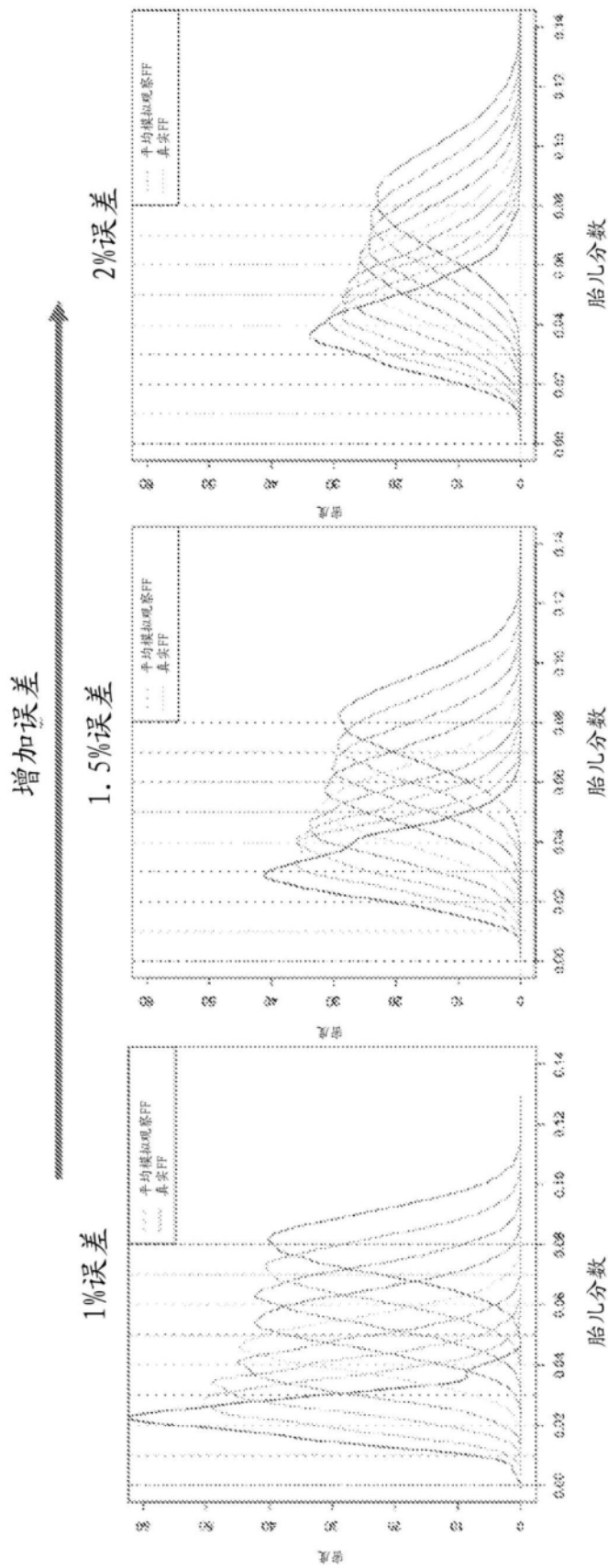


图6



每个图示出了观察FF为0%至8%的样品

图10

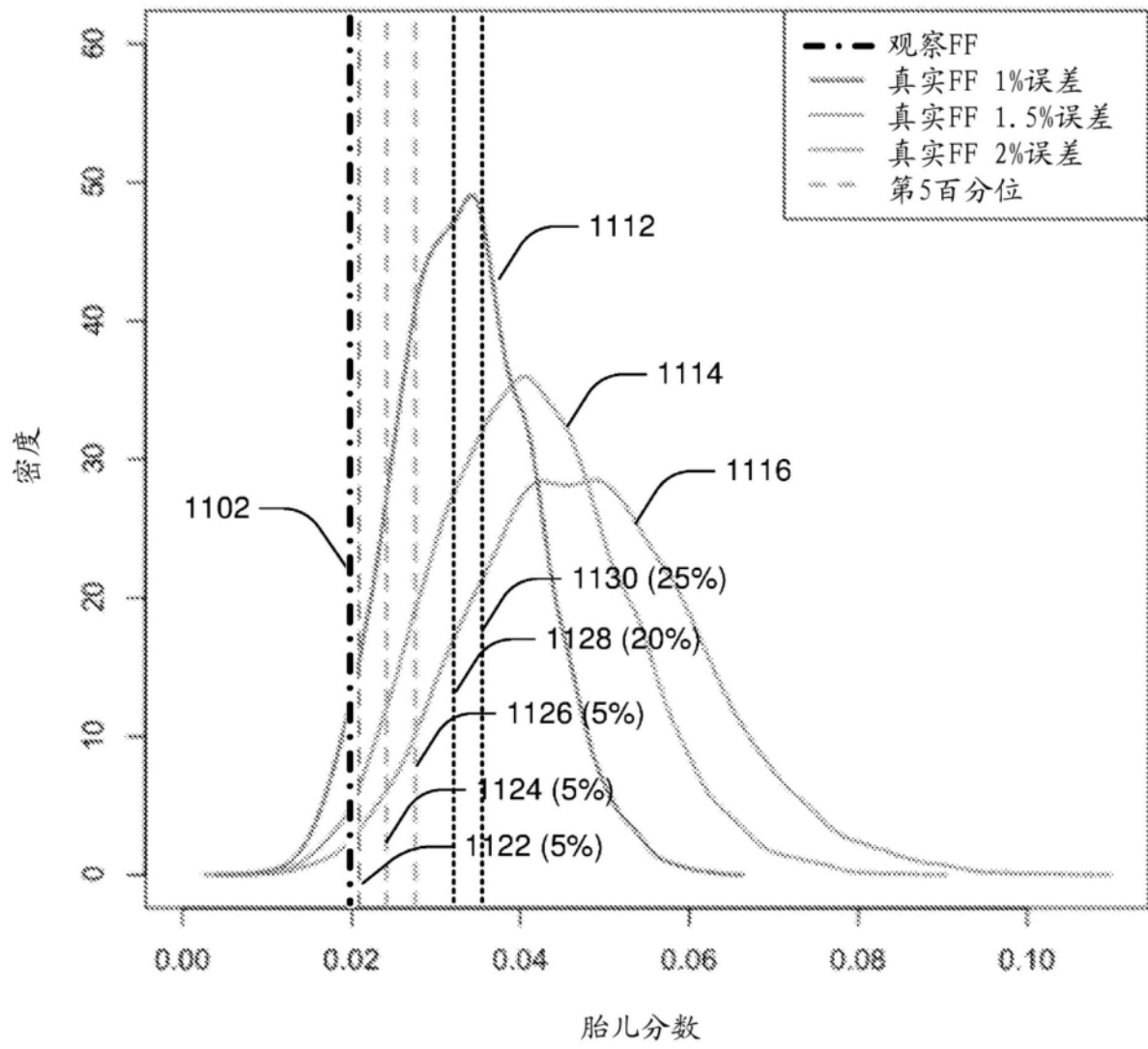


图11

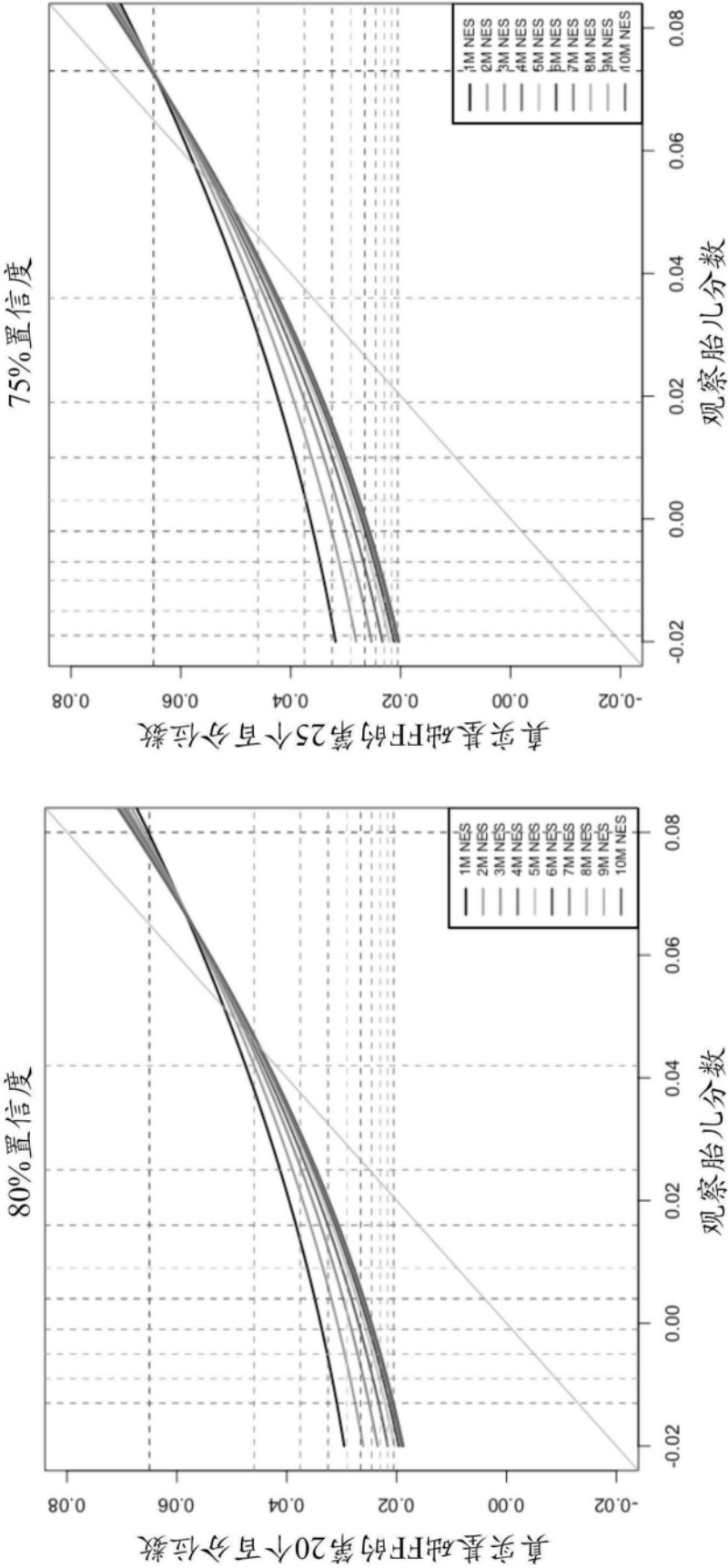


图12

有效读段 计数 (M)	基于ERC 的FF误差 (%)	LOD95 (%)	所需观察FF (%) 80%置 信度	所需观察FF (%) 75%置 信度
1	3.89	6.50	8.0	7.3
2	3.14	4.59	4.2	3.6
3	2.64	3.75	2.5	1.9
4	2.31	3.25	1.6	1.0
5	2.09	2.91	0.9	0.3
6	1.95	2.65	0.4	-0.2
7	1.87	2.46	-0.1	-0.7
8	1.82	2.30	-0.5	-1.0
9	1.80	2.17	-0.9	-1.5
10	1.80	2.05	-1.3	-1.9

图13

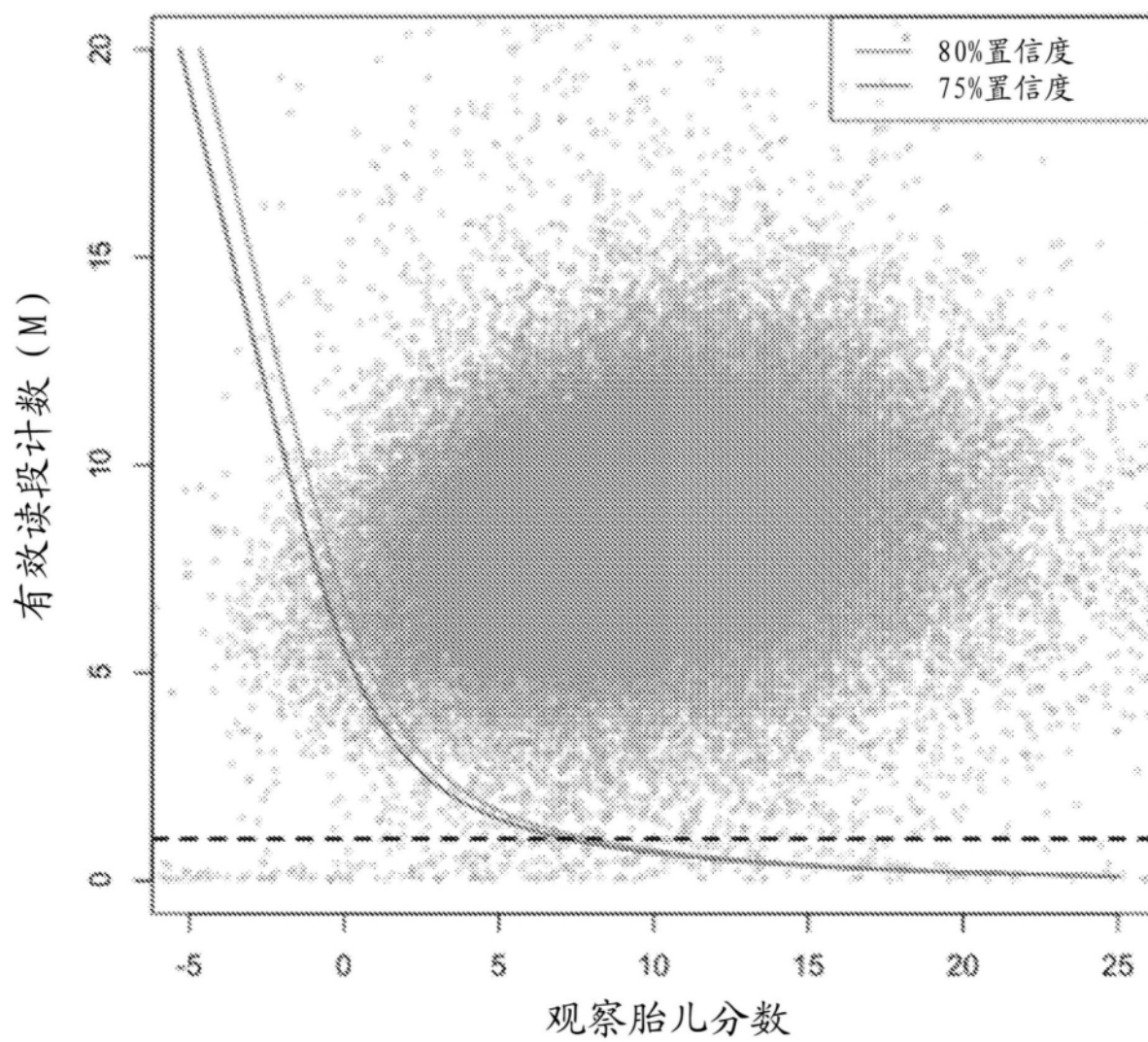


图14

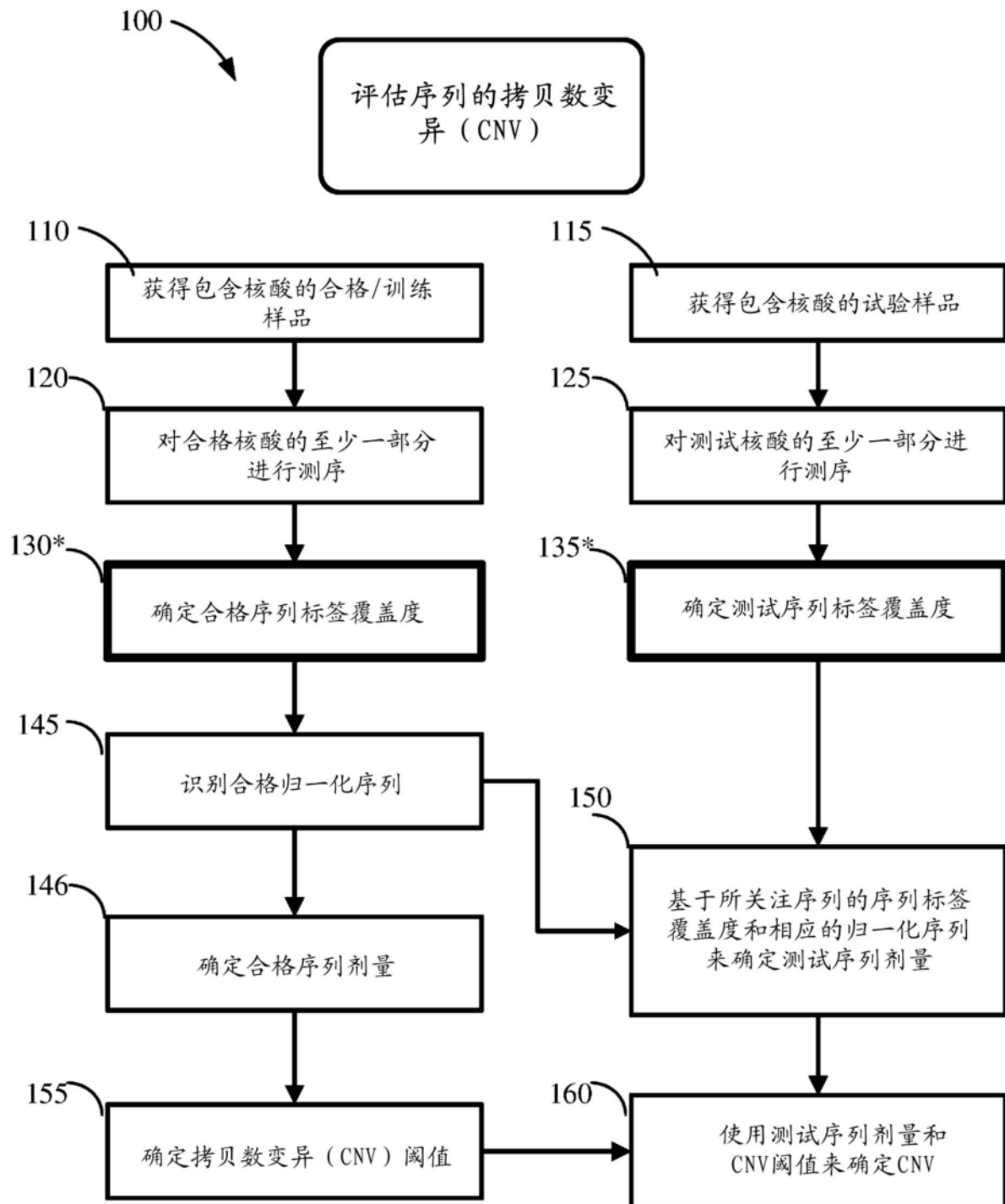


图15

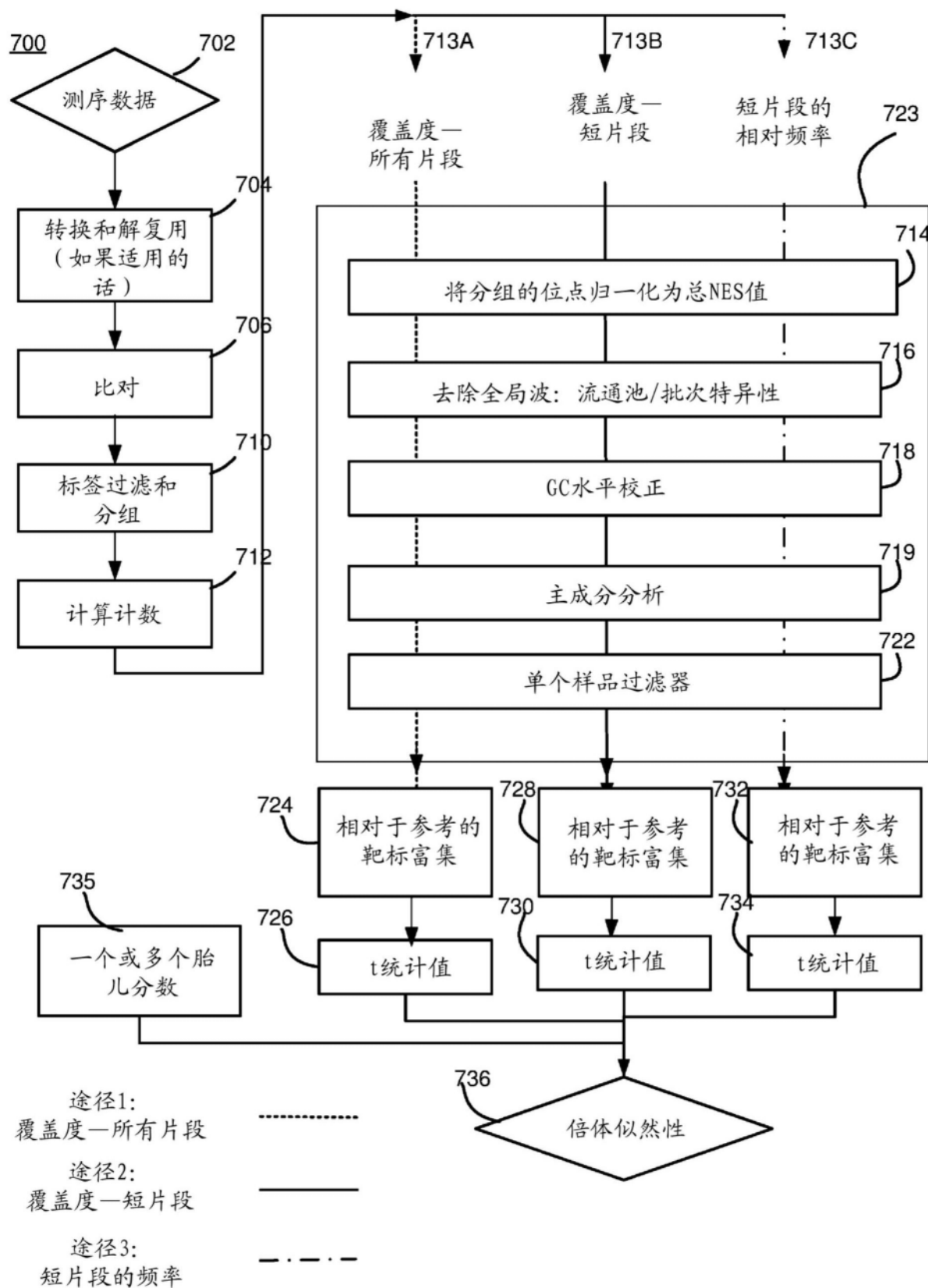


图16A

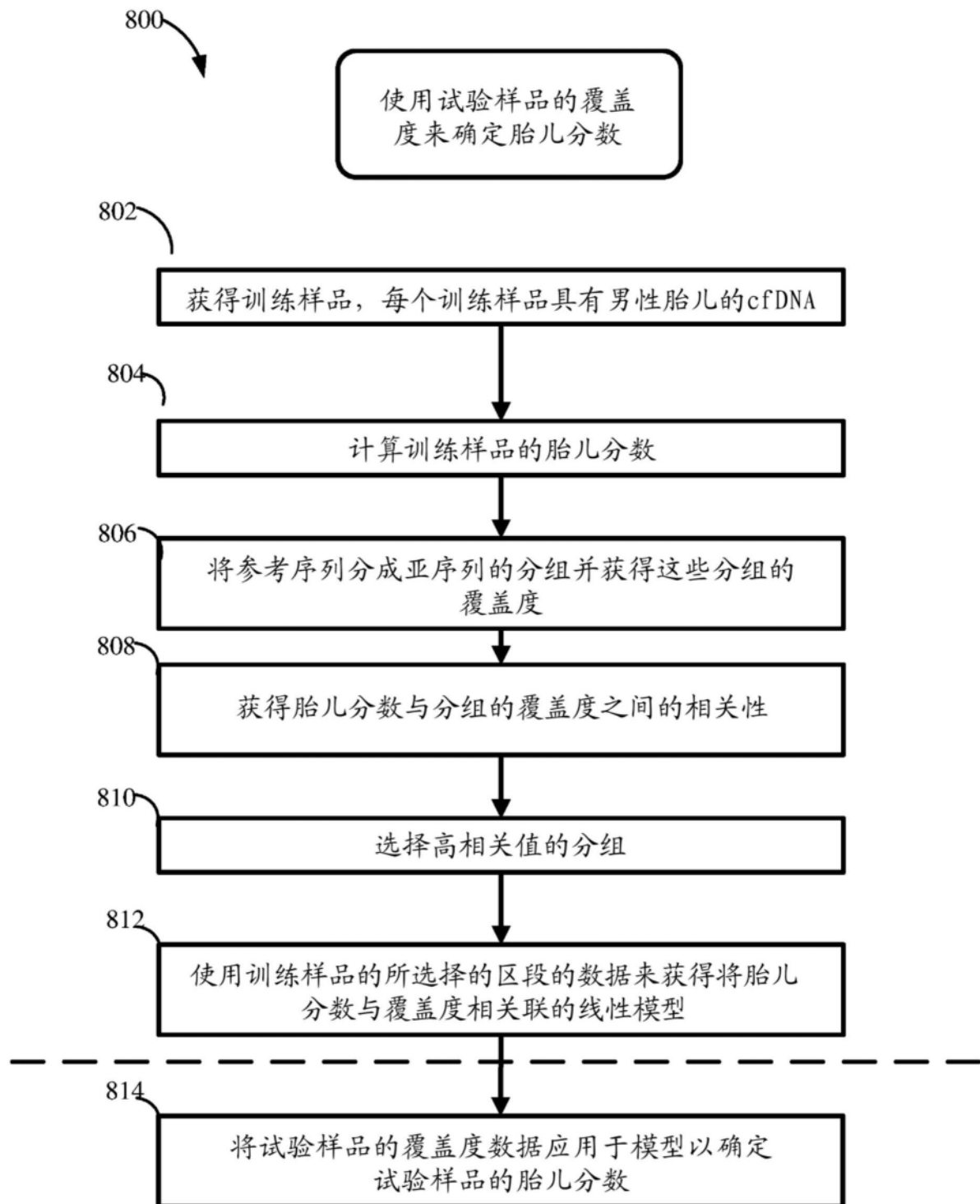


图16B

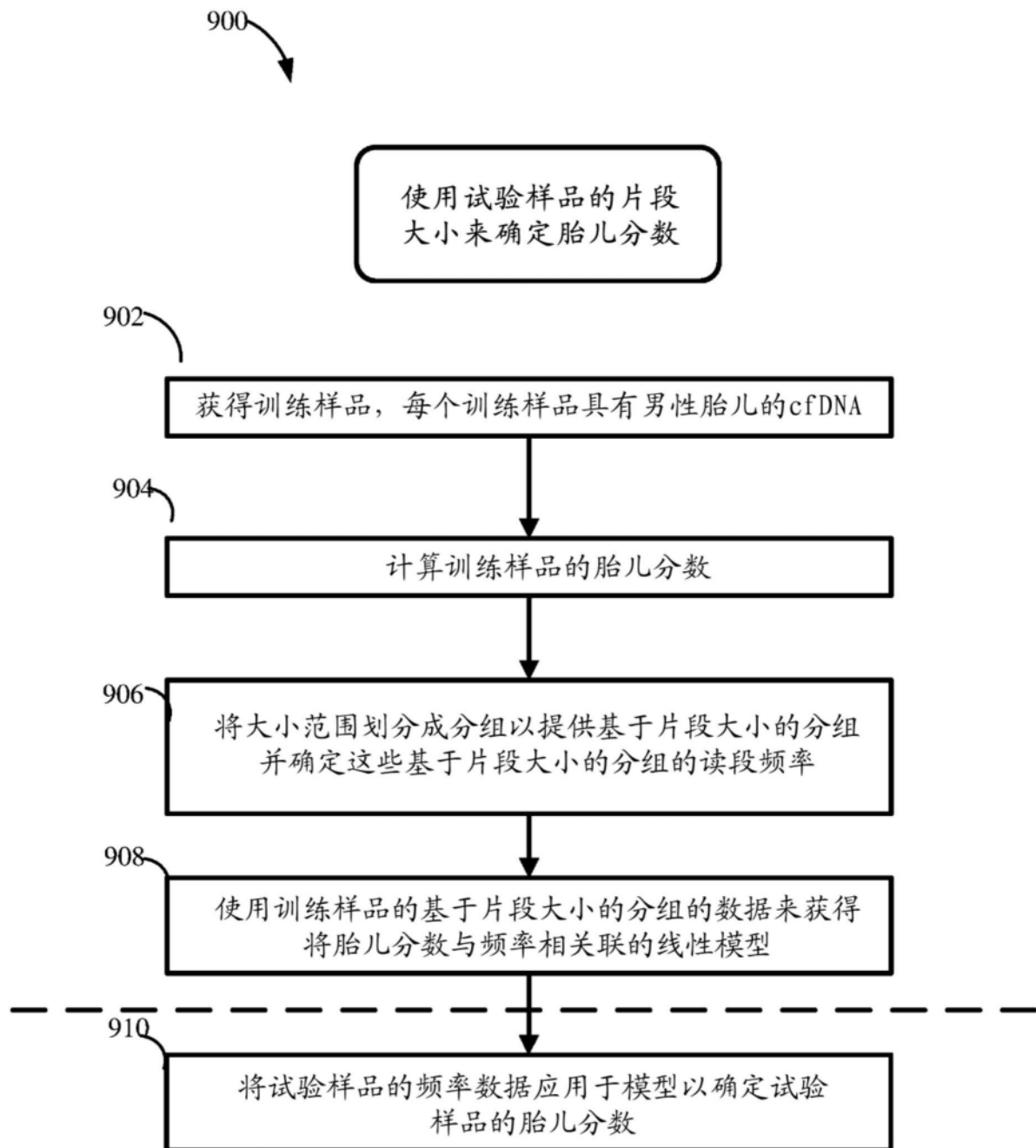


图16C

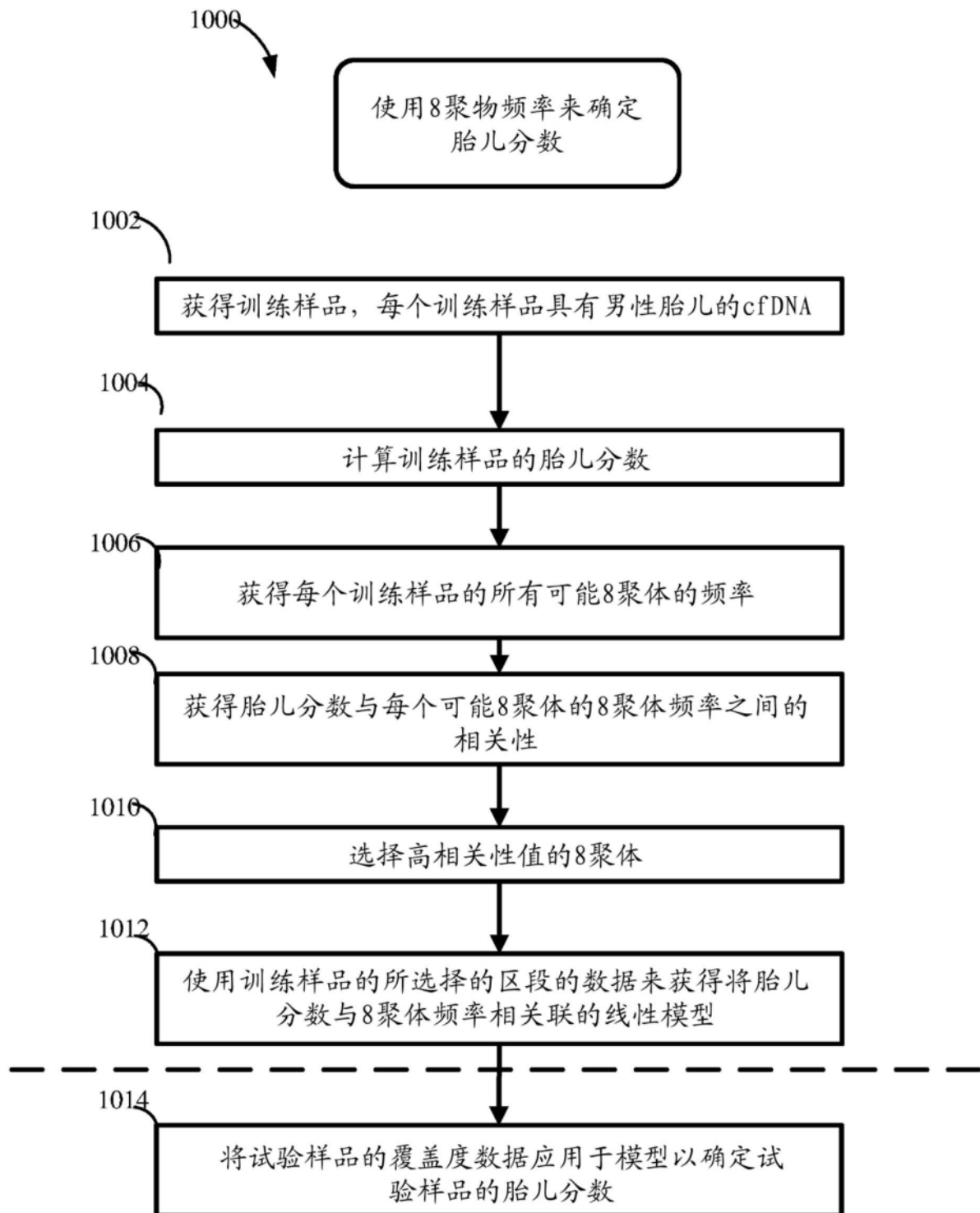


图16D

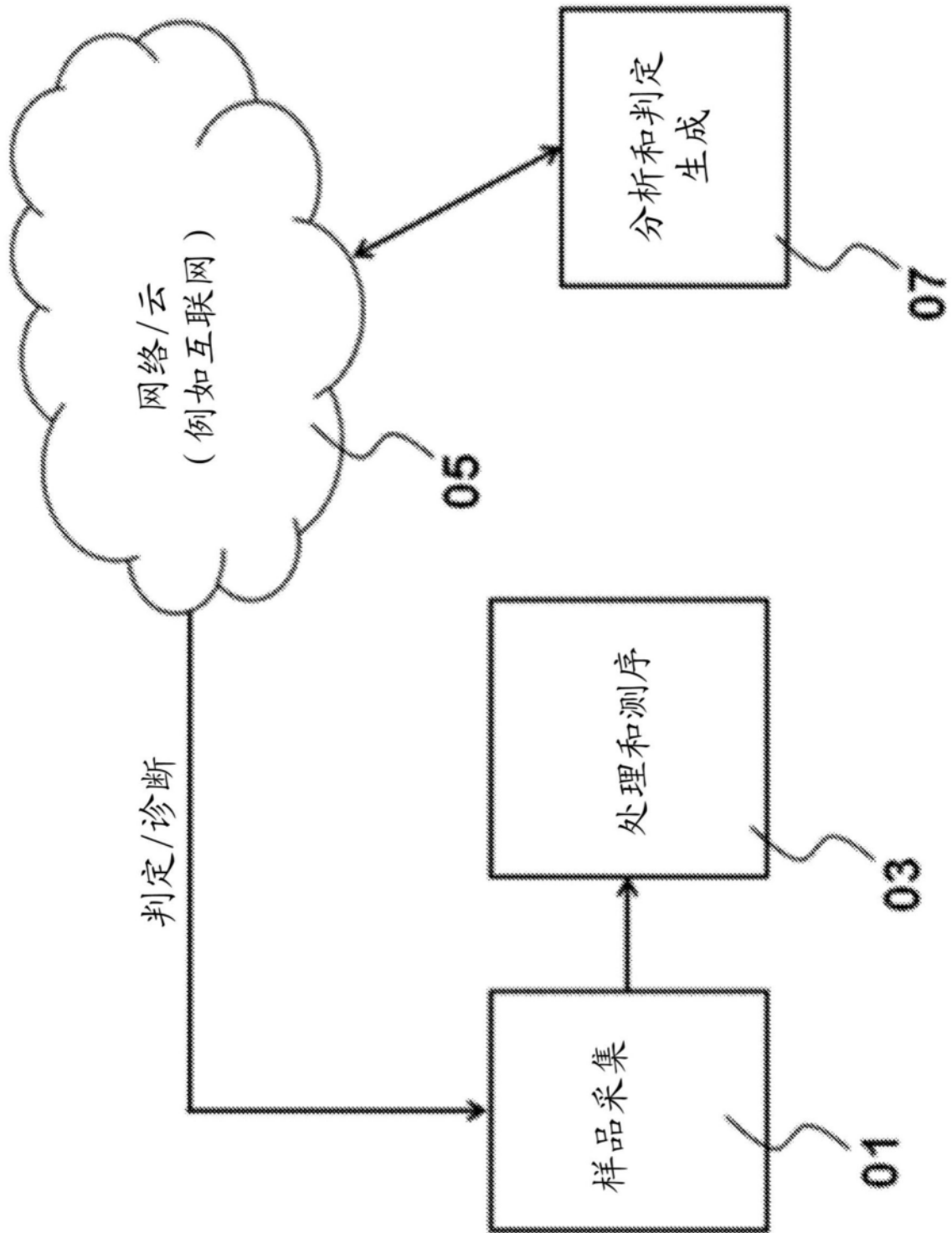


图17

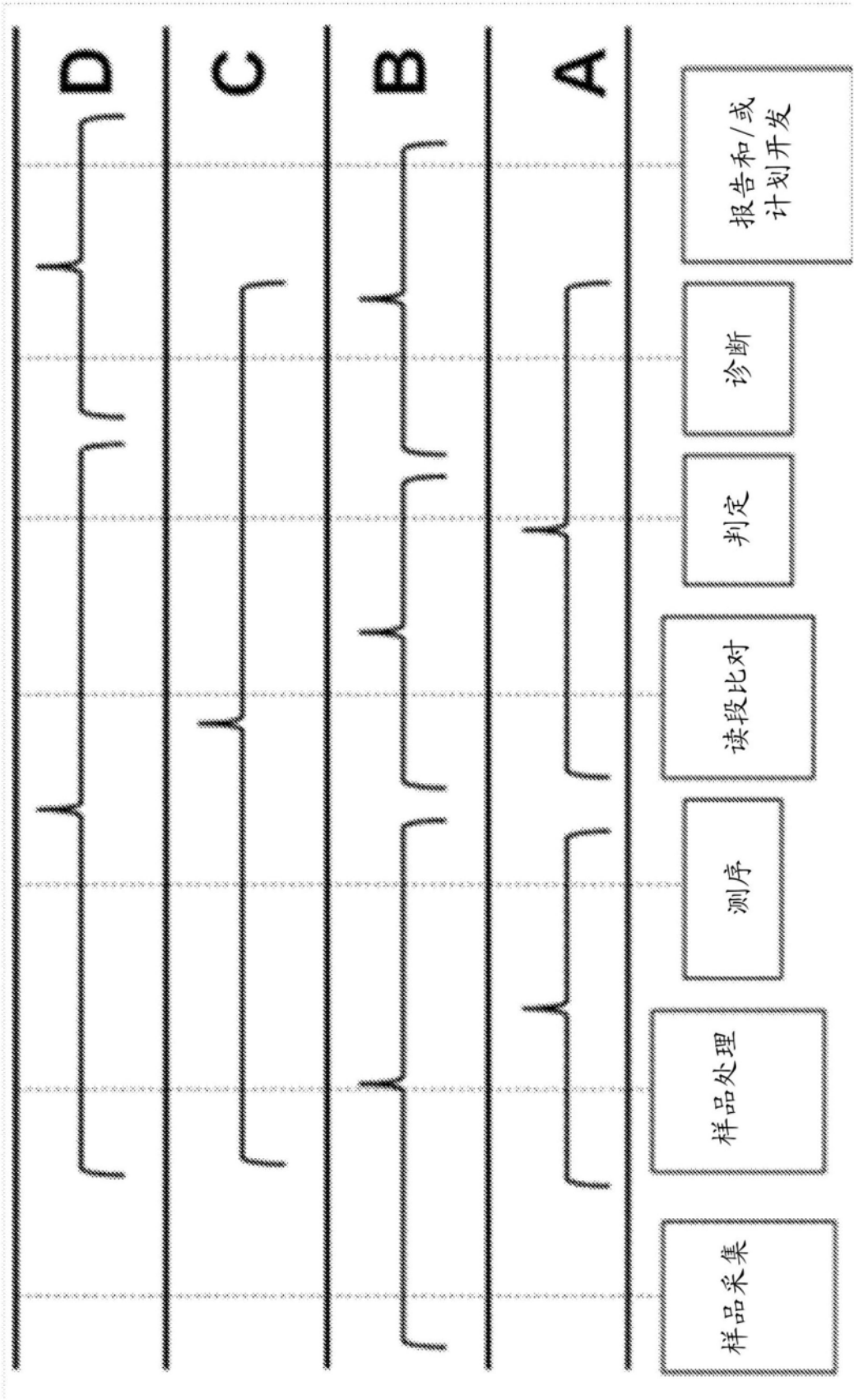


图18

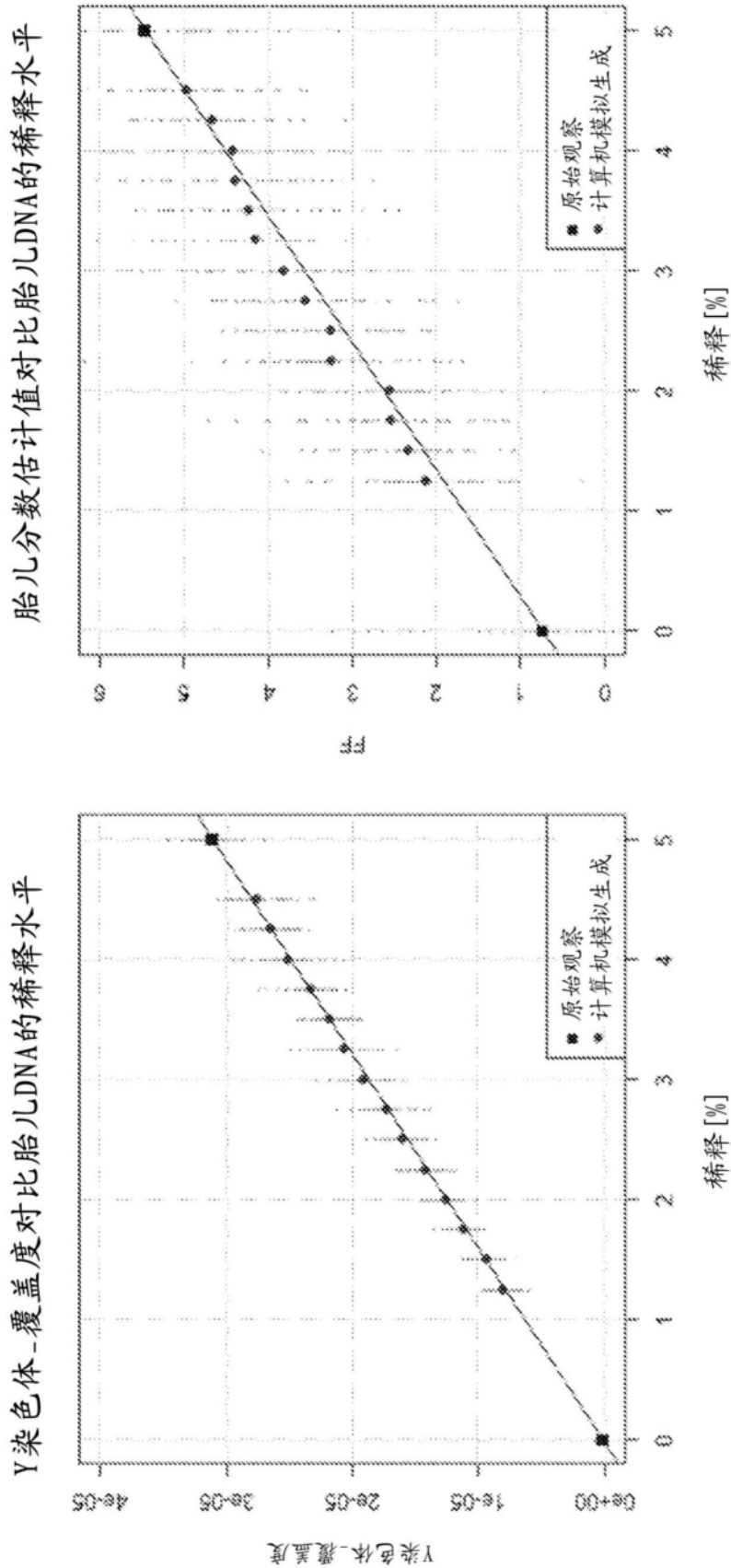


图19

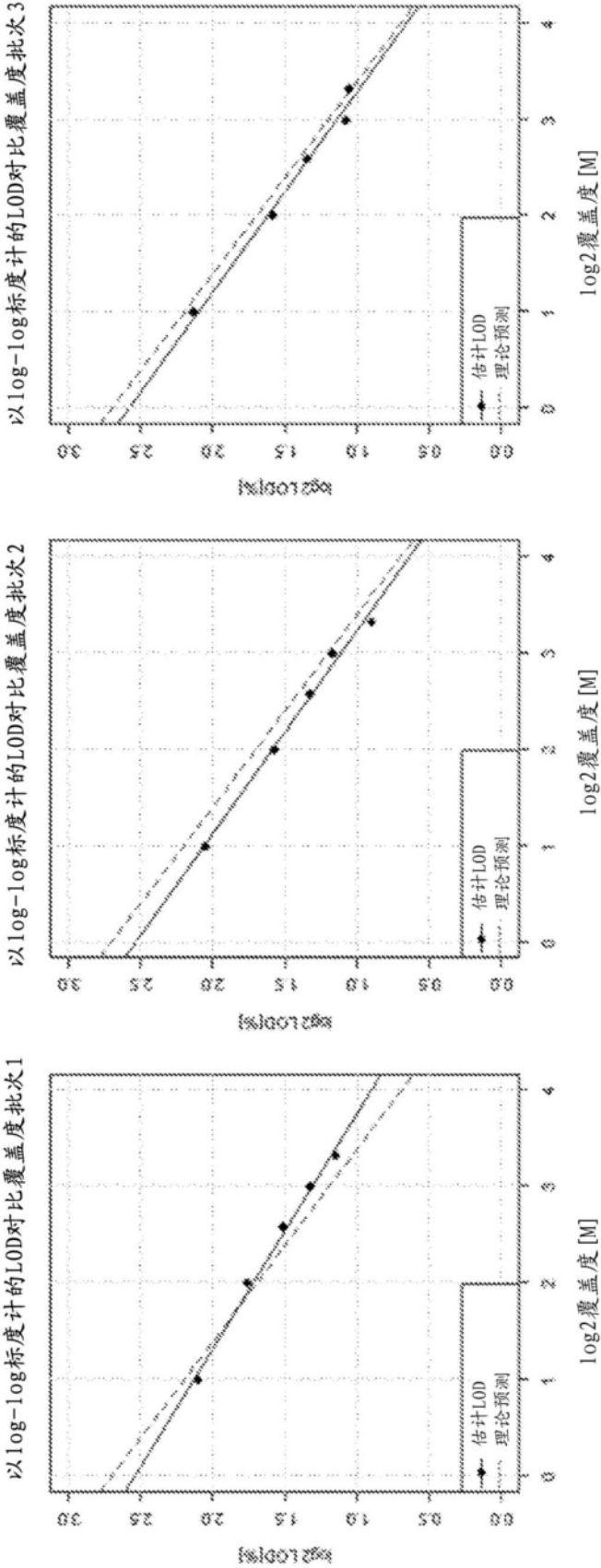


图20

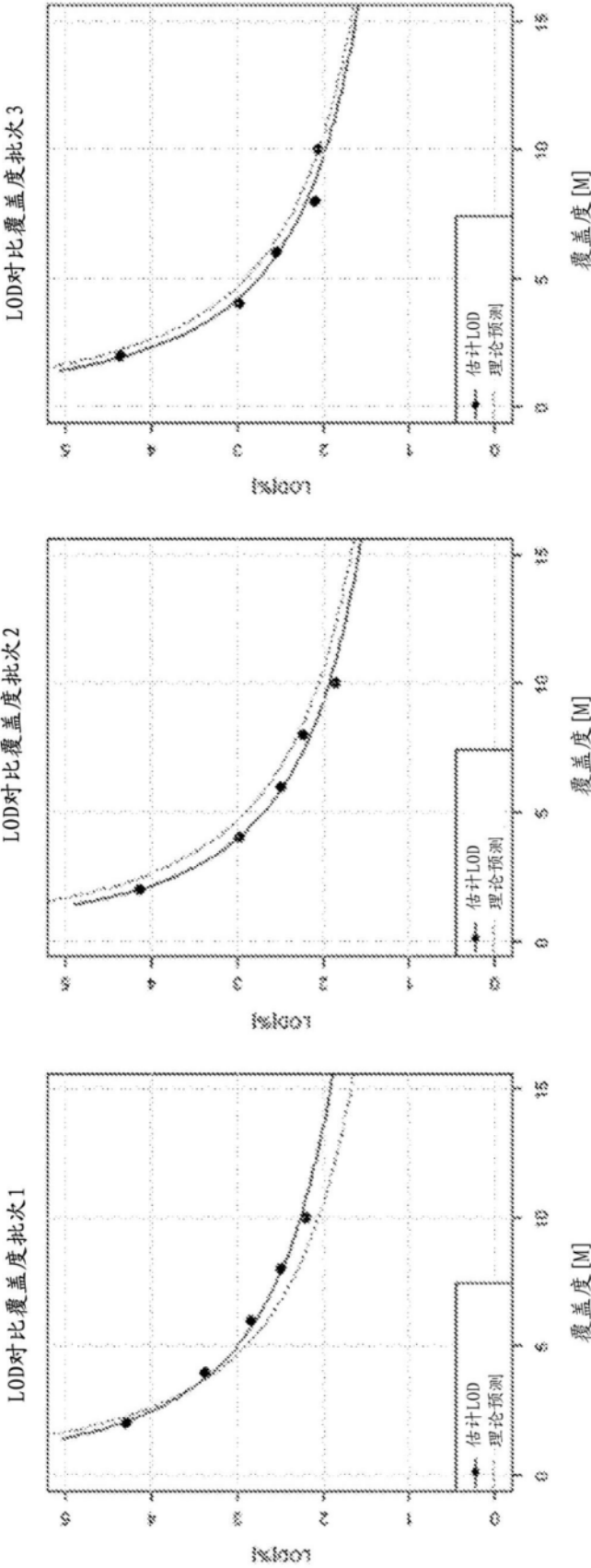


图21

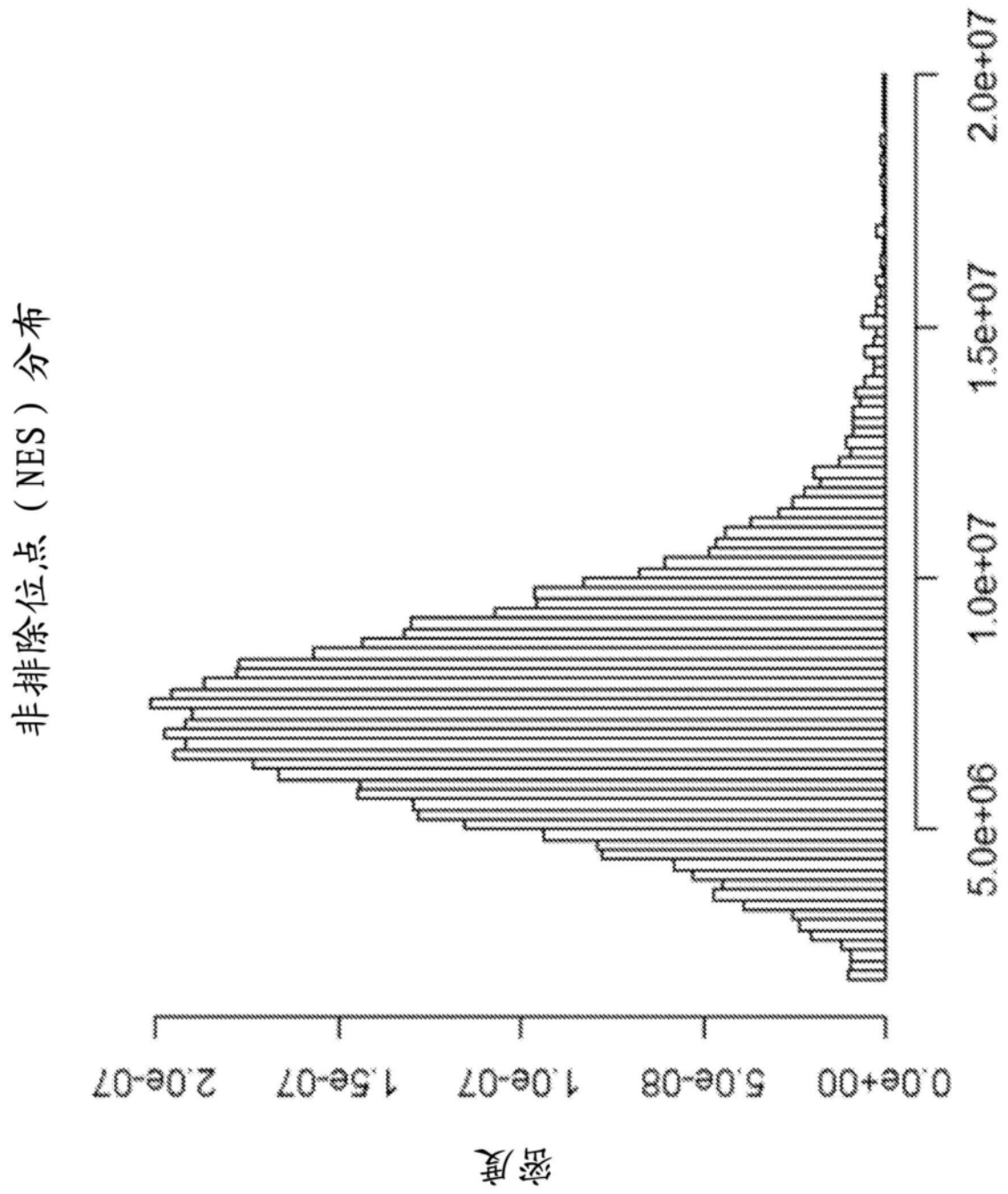


图22

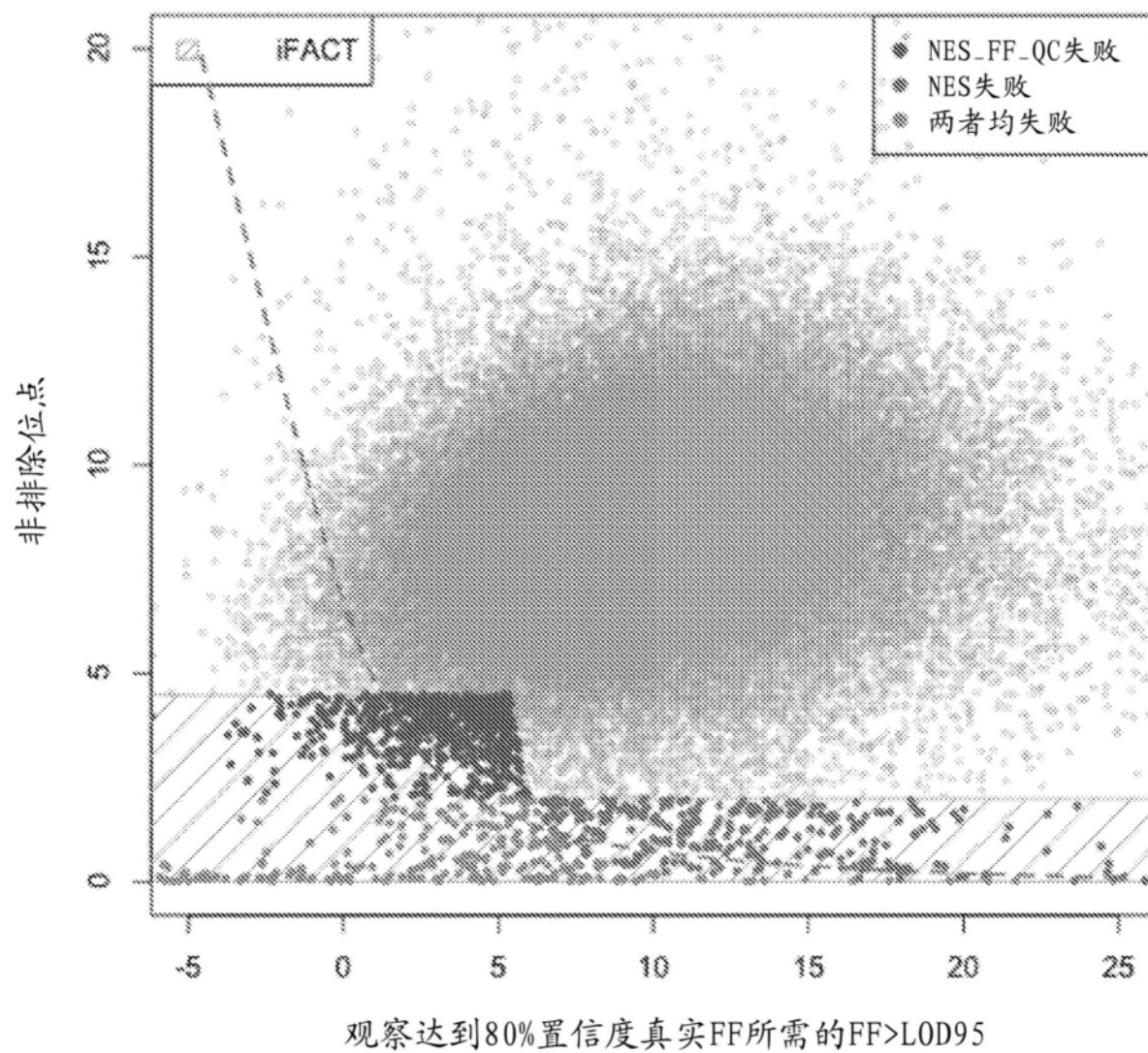


图23

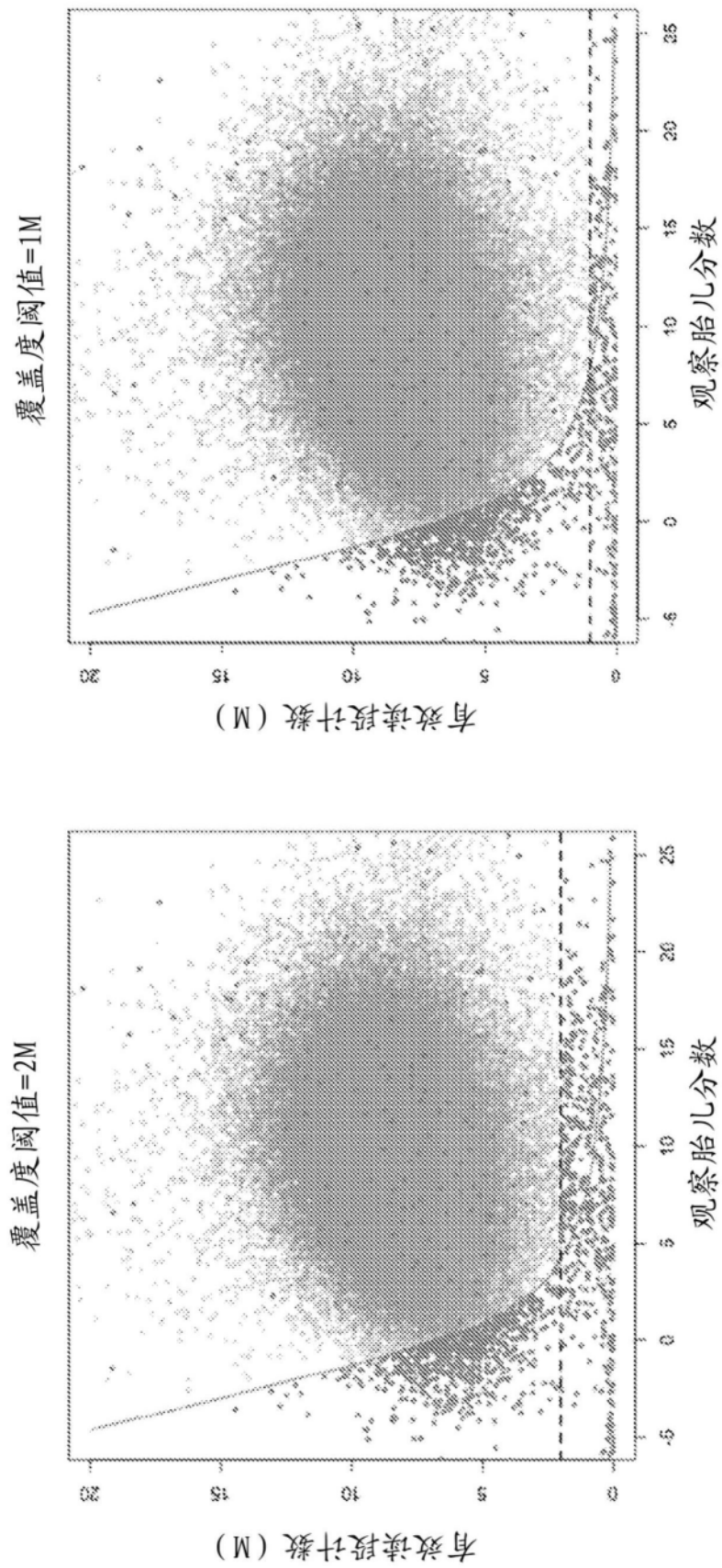
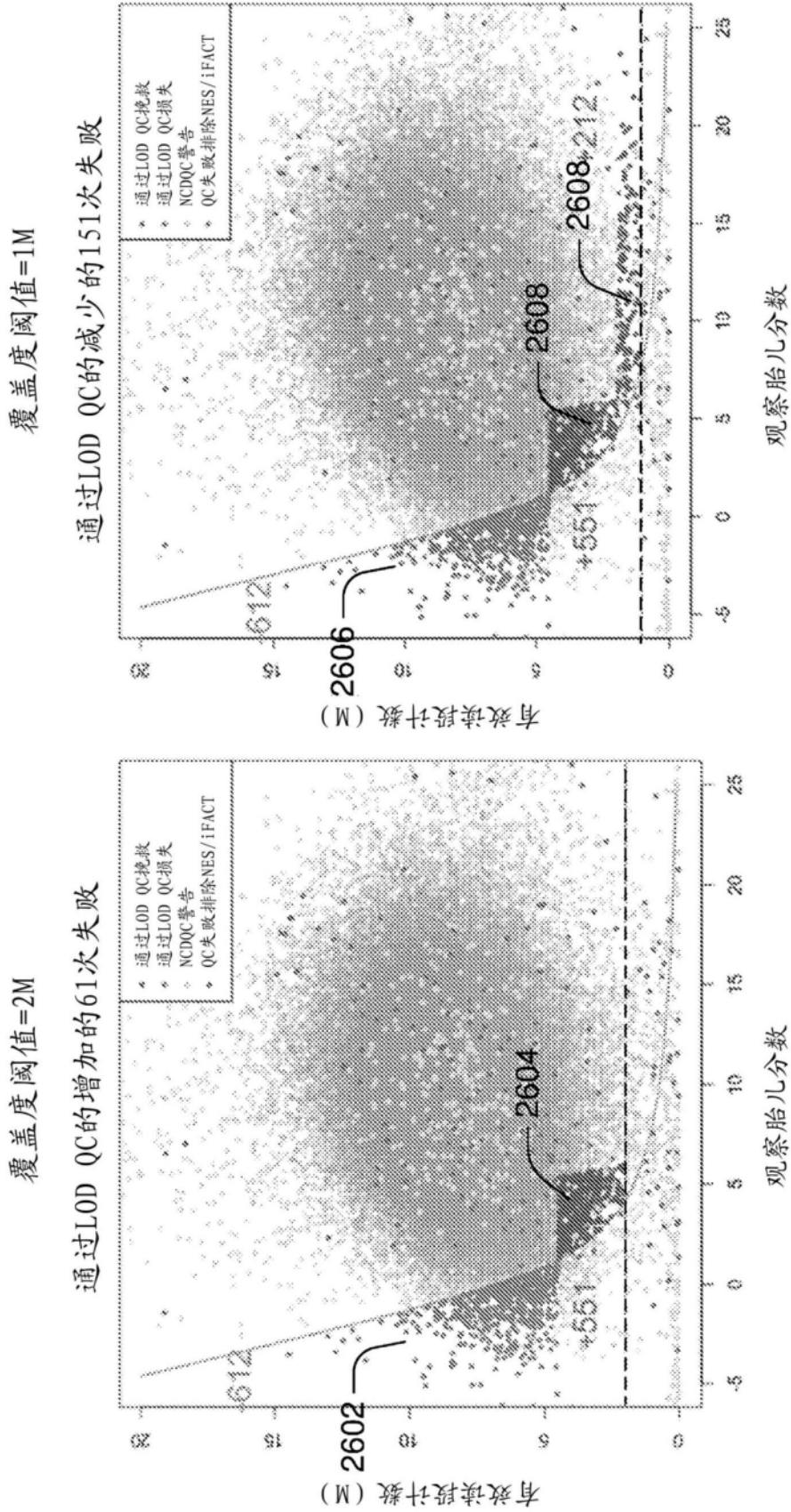


图24

	原始QC (具有NES FF QC)	新QC (具有LOD QC, 80%+1M)
第1次运行通过 (%)	97.75	97.88
第1次运行失败 (%)	2.25	2.12
重新运行的通过 (%)	81.31 (包括那些未重新运行的样品, 因为未重新运行意味着不可挽救)	81.87 (排除那些未重新运行的样品, 因为这些样品中的大多数由于通过原始QC而不重新运行)
最终通过 (%)	99.58	99.62
最终失败 (%)	0.42	0.38

图25



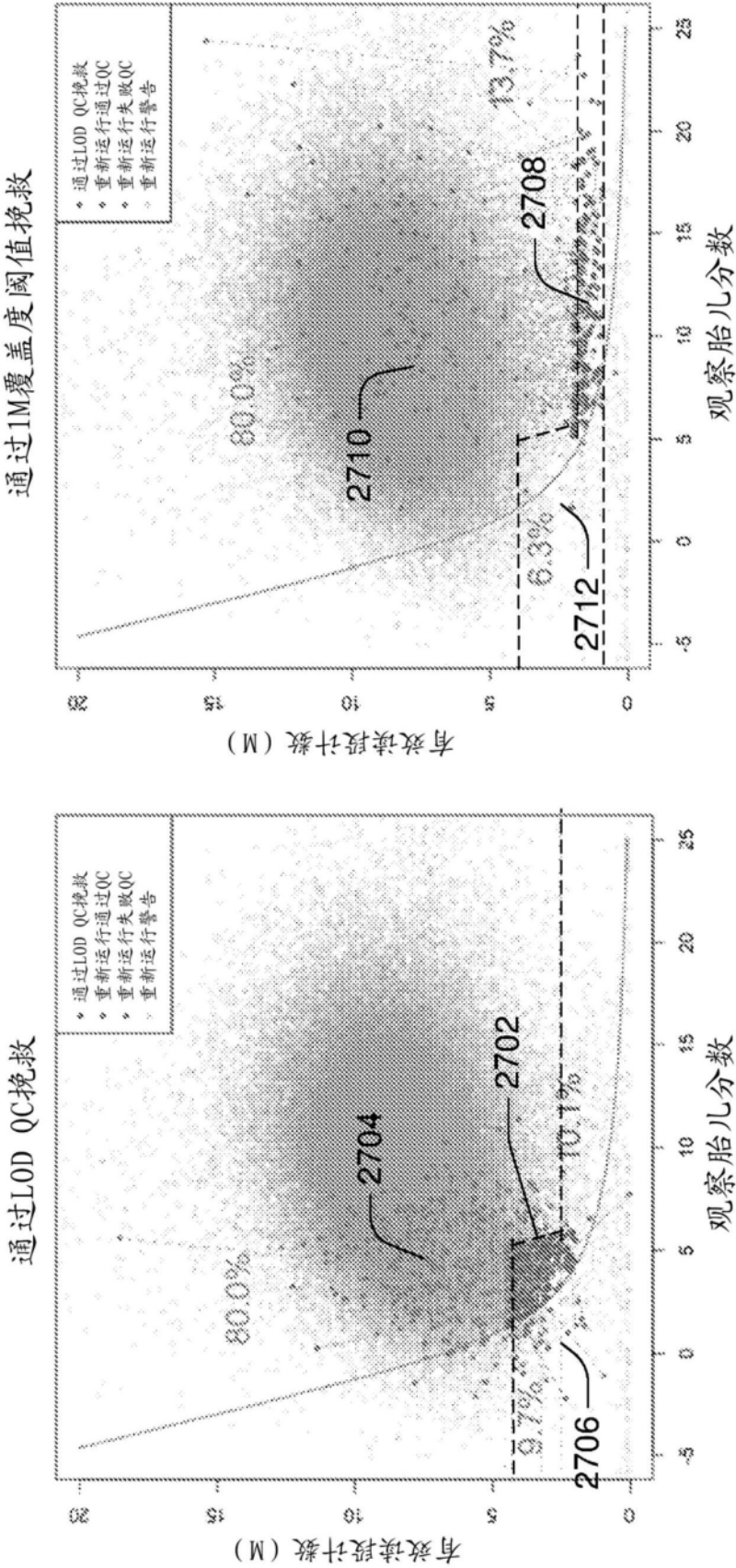


图27

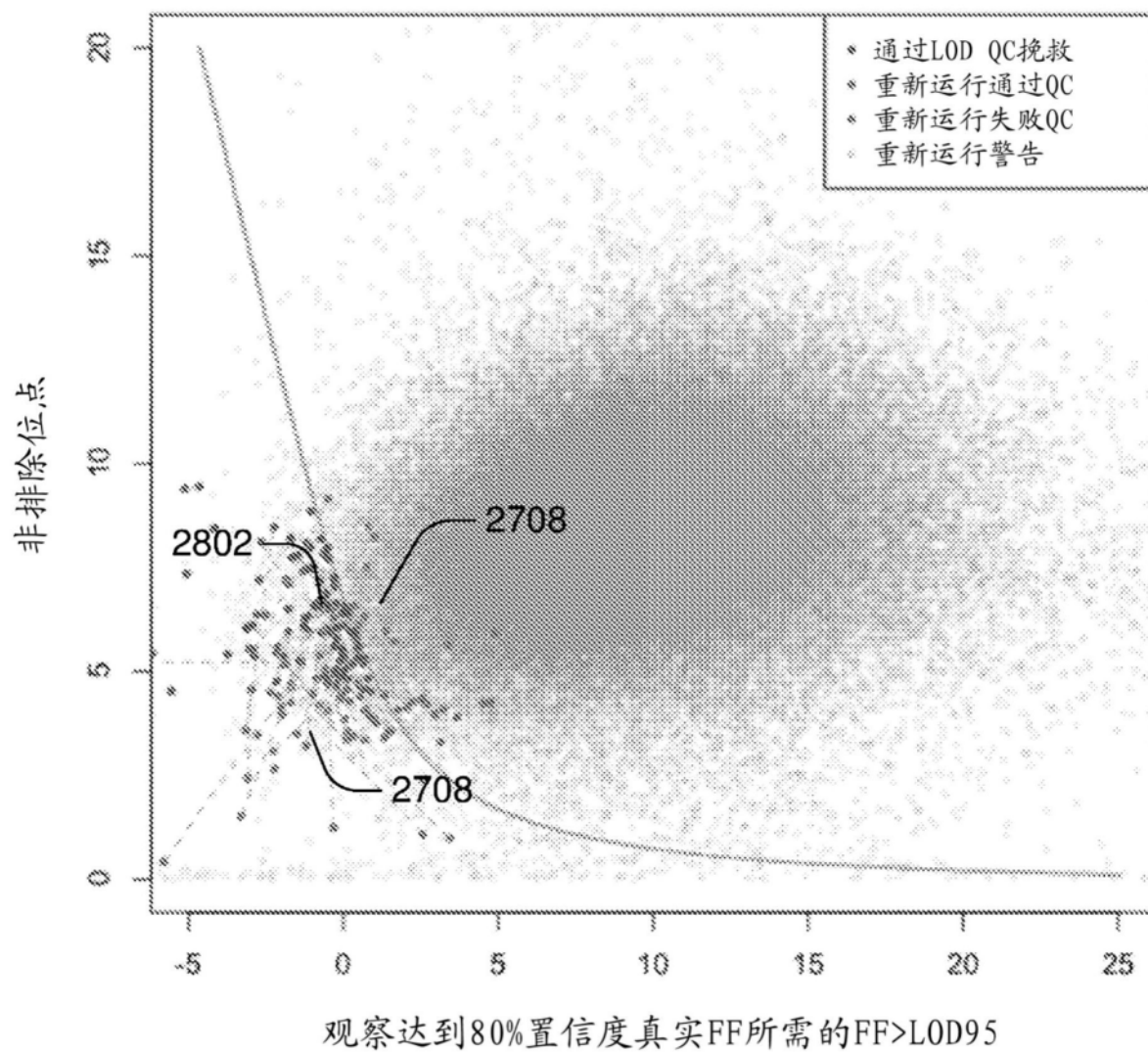


图28

	原始QC (具有NES FF QC)	新QC (具有LOD QC, 80%+1M)
第1次运行通过 (%)	97.75	97.88
第1次运行失败 (%)	2.25	2.12
重新运行的通过 (%)	81.31 (包括那些未重新运行 的样品, 因为未重新运行意 味着不可挽救)	81.71 (排除那些未重新运行 的样品, 因为这些样品中的 大多数由于通过原始QC而不 重新运行)
最终通过 (%)	99.58	99.62
最终失败 (%)	0.42	0.38

图29

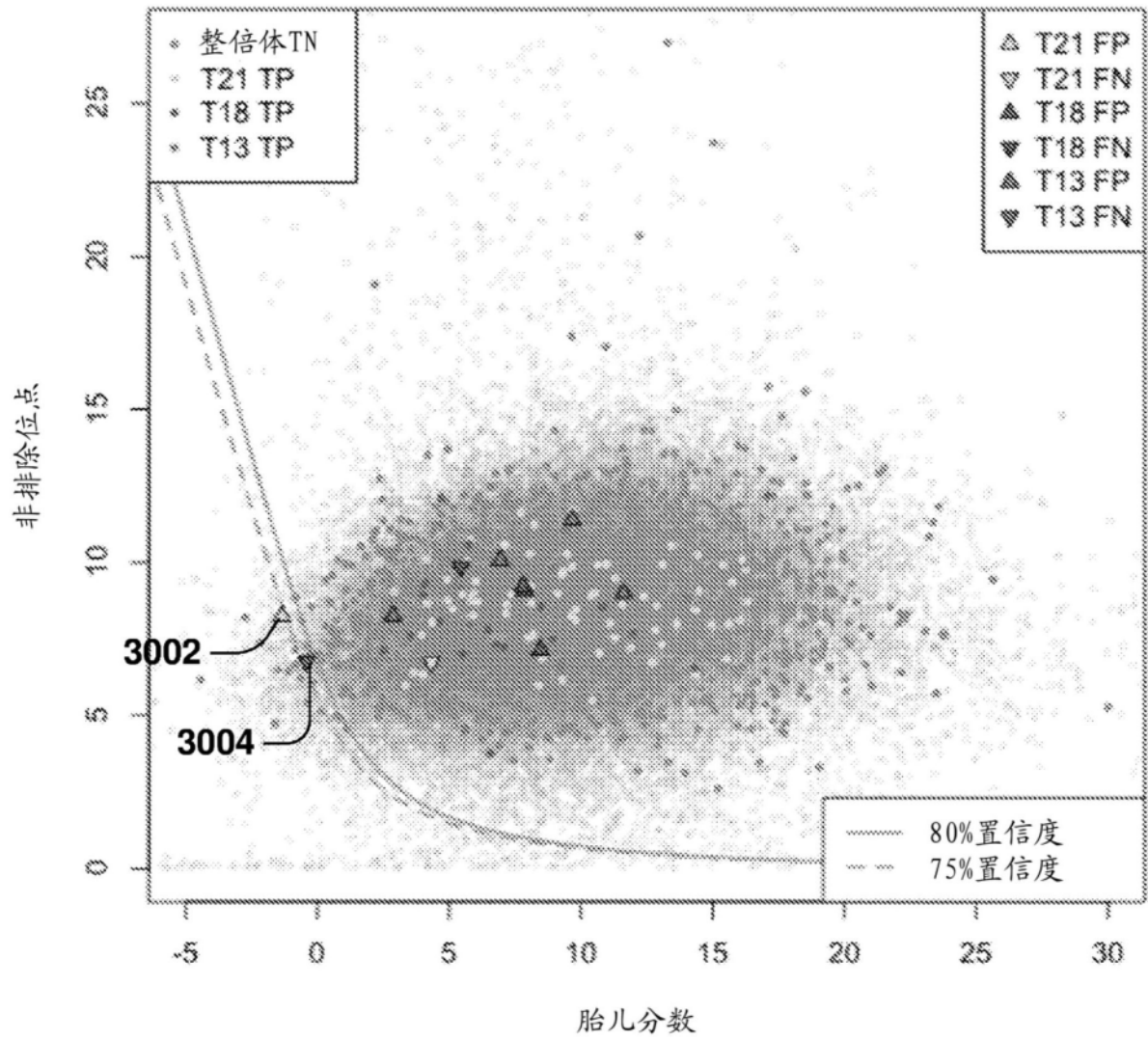
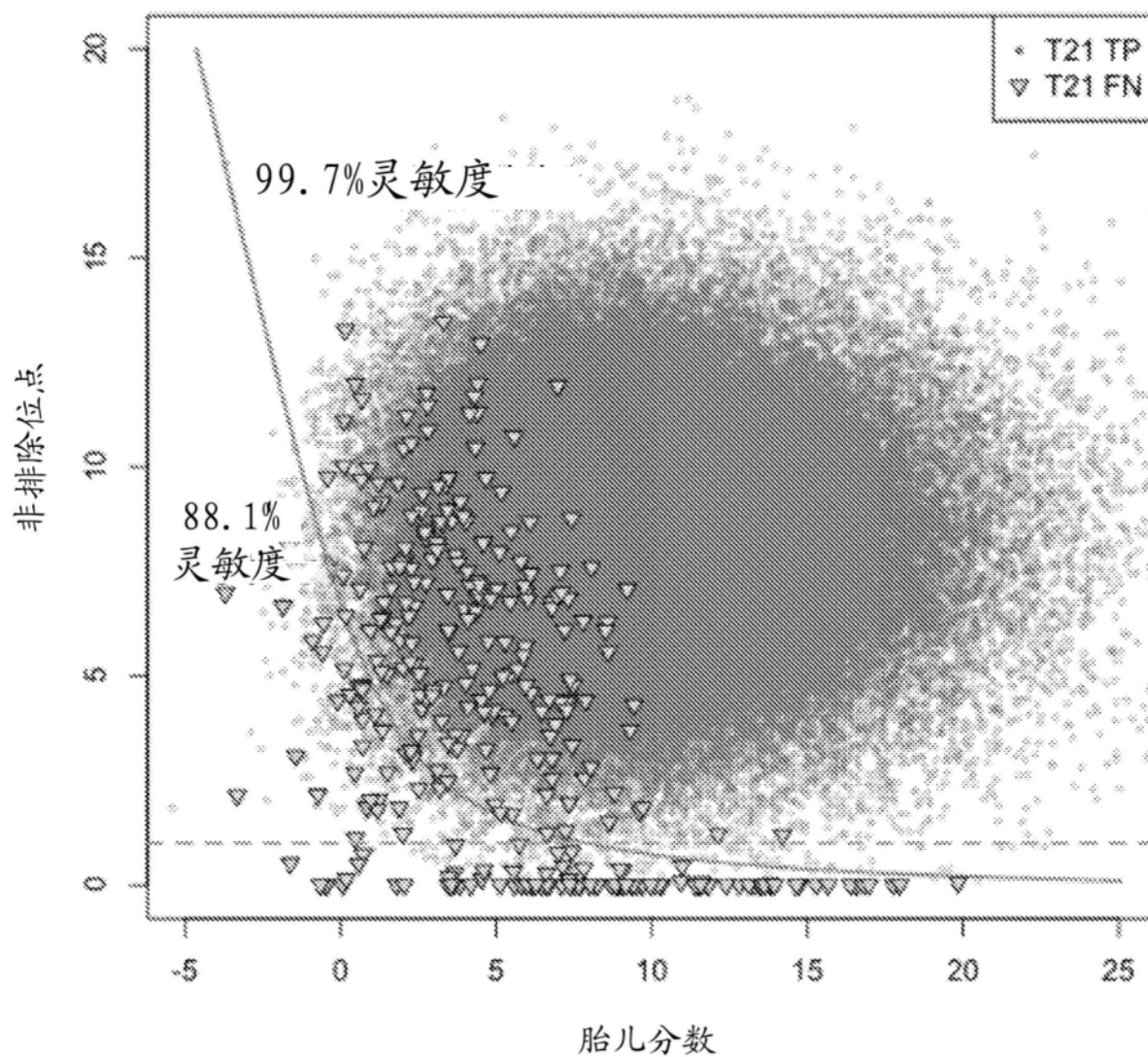


图30



	灵敏度 (%)
通过LOD QC	99.7 (99.7-99.8)
失败LOD QC	88.1 (85.3-90.9)
总体	99.7 (99.6-99.7)

图31