



(12) 发明专利申请

(10) 申请公布号 CN 112005306 A

(43) 申请公布日 2020. 11. 27

(21) 申请号 201980023369.5

(22) 申请日 2019.03.13

(30) 优先权数据

62/642,461 2018.03.13 US

(85) PCT国际申请进入国家阶段日

2020.09.28

(86) PCT国际申请的申请数据

PCT/US2019/022139 2019.03.13

(87) PCT国际申请的公布数据

W02019/178289 EN 2019.09.19

(71) 申请人 格里尔公司

地址 美国加利福尼亚州门洛帕克市

(72) 发明人 达莉亚·菲利波娃

安东·瓦卢耶夫 弗吉尔·尼古拉

卡蒂克·贾加迪什

M·赛勒斯·马厄 马修·H·拉森

莫妮卡·波特拉·朵斯·桑托斯·

皮门特尔

罗伯特·安倍·潘恩·卡列夫

(74) 专利代理机构 上海翼胜专利商标事务所

(普通合伙) 31218

代理人 翟羽

(51) Int.Cl.

G16B 30/10 (2019.01)

权利要求书9页 说明书53页 附图36页

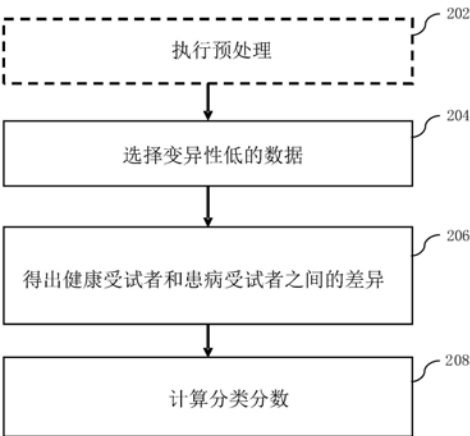
(54) 发明名称

选择、管理和分析高维数据的方法和系统

(57) 摘要

本发明公开一种用于分析高维数据(例如与一疾病状况有关的多个核酸样本的序列读数)的系统、方法和计算机程序产品。

200



1. 一种分析与一疾病状况有关的多个核酸样本的序列读数的方法,其特征在于:所述方法包括步骤:

根据来自多个健康受试者的一参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数,识别一参考基因组中的多个低变异区域,其中在来自每个健康受试者的多个核酸样本的所述第一组的序列读数中的每个序列读数可以与所述参考基因组中的一区域对齐;

从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与所述参考基因组中所述低变异性区域中的一区域对齐,其中所述训练组包括来自多个健康受试者的多个核酸样本的序列读数和来自已知患有所述疾病状况的多个患病受试者的多个核酸样本的序列读数,并且其中来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;

使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了所述训练群组内所述多个健康受试者的多个核酸样本的序列读数与来自所述多个患病受试者的多个核酸样本的序列读数之间的差异;

接收与来自一测试受试者的多个核酸样本相关的一测试组的序列读数,所述测试受试者的所述疾病状况为未知;以及

根据所述一个或多个参数,预测所述测试受试者患有所述疾病状况的一可能性。

2. 如权利要求1所述的方法,其特征在于:所述多个核酸样本包含多个游离核酸(cfDNA)片段。

3. 如权利要求1或2所述的方法,其特征在于:所述疾病状况为癌症。

4. 如权利要求1至3任一项所述的方法,其特征在于:所述疾病状况是选自下列群组中组成的一癌症类型:肺癌、卵巢癌、肾癌、膀胱癌、肝胆癌、胰腺癌、上消化道癌、肉瘤、乳腺癌、肝癌、前列腺癌、脑癌及其组合。

5. 如权利要求1至4任一项所述的方法,其特征在于:所述方法更包括:根据来自多个健康受试者的一基线群组的多个核酸样本的序列读数,对多个健康受试者的所述参考群组中每个健康受试者的多个核酸样本的所述第一组的序列读数进行初始数据处理,其中所述参考群组和所述基线群组不重迭,并且其中所述初始数据处理包括校正GC偏差或对与所述参考基因组的多个区域对齐的序列读数进行的标准化。

6. 如权利要求1至4任一项所述的方法,其特征在于:所述方法更包括:根据来自多个健康受试者的一基线群组的多个核酸样本的序列读数,所述训练群组中每个受试者的多个核酸样本的序列读数进行初始数据处理,其中所述基线群组和所述训练群组不重迭,并且其中所述初始数据处理包括校正GC偏差或对与所述参考基因组的多个区域对齐的序列读数进行的标准化。

7. 如权利要求1至6任一项所述的方法,其特征在于:识别所述参考基因组中的所述多个低变异区域的步骤还包括:

将来自多个健康受试者的所述参考群组中的每个健康受试者的多个核酸样本的所述第一组的序列读数的多个序列对准所述参考基因组的多个非重叠区域,所述参考群组包括一第一多数健康受试者;

对于所述参考群组中的每个健康受试者,推导与多个序列读数相关联的一数量,其与所述参考基因组的所述多个非重叠区域内的一区域对齐,从而呈现与所述区域相对应的一第一多数数量;

基于所述第一多数数量确定一第一参考量和一第二参考量;以及

当所述第一参考量和所述第二参考量满足一预定条件时,将所述区域识别为具有低变异性。

8.如权利要求7所述的方法,其特征在于:所述方法更包括:

对所述参考基因组的多个非重叠区域中的所有剩余的区域,重复所述确定和识别的步骤,从而识别所述参考基因组中的所述多个低变异区域。

9.如权利要求1至8任一项所述的方法,其特征在于:从所述训练群组的多个核酸样本的多个序列读数中选择所述训练组的序列读数的步骤还包括:从所述训练群组的多个核酸样本的多个序列读数中选择多个序列读数,所述序列读数与所述参考基因组中的多个低变异性区域对齐,从而生成所述训练组的序列读数。

10.如权利要求1至9任一项所述的方法,其特征在于:确定一个或多个参数的步骤还包括:

对于所述训练群组中的每个受试者,并且针对所述多个低变异性区域中的一区域,基于与所述区域对齐的序列读数,推导一个或多个数量;

对所有剩余的低变异性区域,重复所述推导的步骤,以呈现与所述训练群组中所有受试者的低变异性区域相对应的多个数量,其中所述多个数量包括:与多个健康受试者有关的多个数量的一第一子集,和与已知患有所述疾病状况的多个受试者有关的多个数量的一第二子集;以及

确定用于反映多个数量的一第一子集和一第二子集之间的差异的一个或多个参数。

11.如权利要求10所述的方法,其特征在于:所述一个或多个数量由一个数量组成,并对应于与所述区域对齐的序列读数的总数。

12.如权利要求10所述的方法,其特征在于:所述一个或多个数量包括多个数量,每个数量对应于与所述区域对齐的多个序列读数的一子集,其中在相同子集中的每个序列读数对应于具有相同预定片段尺寸或尺寸范围的核酸样本,其中多个不同子集中的多个序列读数对应于具有不同片段尺寸或尺寸范围的多个核酸样本。

13.如权利要求10所述的方法,其特征在于:所述一个或多个参数由主成分分析(PCA)确定。

14.如权利要求1至13任一项所述的方法,其特征在于:所述方法更包括:通过将所述训练组划分为一训练子集和一验证子集,在一多重交叉验证过程中改进所述一个或多个参数。

15.如权利要求14所述的方法,其特征在于:所述多重交叉验证过程中的一环节中的训练子集和验证子集不同于所述多重交叉验证过程的另一环节中的不同的训练子集和验证子集。

16.如权利要求1至15任一项所述的方法,其特征在于:所述方法更包括:从来自所述受试者的多个核酸样本的多个序列读数中选择多个序列读数,所述序列读数与所述参考基因组中的低变异区域对齐,从而生成所述测试组的序列读数;以及

基于所述测试组的序列读数和所述一个或多个参数,计算代表受试者患疾病状况的可能性的一分类分数。

17.如权利要求1至16任一项所述的方法,其特征在于:所述参考基因组中每个变异区域的大小在1万碱基对至10万碱基对之间。

18.如权利要求17所述的方法,其特征在于:所述参考基因组中每个变异区域的大小相同。

19.如权利要求17所述的方法,其特征在于:所述参考基因组中的多个所述变异区域的大小不相同。

20.如权利要求1至19任一项所述的方法,其特征在于:所述一个或多个参数基于所述训练组的序列读数的一子集来确定。

21.如权利要求1至20任一项所述的方法,其特征在于:

所述训练组的序列读数中的所述序列读数包括来自所述训练群组中多个受试者的多个核酸样本中游离细胞DNA(cfDNA)片段的序列读数;

来自所述训练群组中多个受试者的多个核酸样本包括比一第一阈值长度长的cfDNA片段,其中所述第一阈值长度小于160个核苷酸;以及

所述训练组的序列读数中的所述序列读数不包括大于第一阈值长度的多个cfDNA分子的序列读数。

22.如权利要求21所述的方法,其特征在于:所述第一阈值长度为140个核苷酸或更少。

23.如权利要求21或22所述的方法,其特征在于:所述训练组中的所述序列读数包括来自所述训练组中所述受试者的所述核酸样本中的cfDNA片段的序列读数,所述序列读数的长度介于一第二阈值长度和一第三阈值长度之间;

其中,

所述第二阈值长度是240个核苷酸至260个核苷酸;以及

所述第三阈值长度是290个核苷酸至310个核苷酸。

24.如权利要求21至23任一项所述的方法,其特征在于:排除长于所述第一阈值长度的cfDNA分子的序列读数是通过将来自所述训练群组的所述受试者中长于所述第一阈值长度的cfDNA分子与来自所述训练群组中的所述受试者中短于所述第一阈值长度的cfDNA分子进行物理分离来实现的。

25.如权利要求21至23任一项所述的方法,其特征在于:排除长于所述第一阈值长度的cfDNA分子的序列读数是通过从计算器模拟中筛选出来自所述训练群组的多个所述受试者的核酸样本中长度大于第一阈值长度的cfDNA片段的序列读数来实现的。

26.一种基于来自一参考群组中多个健康受试者的测序数据识别一参考基因组中多个低变异区域的方法,其特征在于:所述方法包括步骤:

将来自所述参考群组中每个健康受试者的多个核酸样本的多个序列读数的第一组的多个序列对准所述参考基因组的多个非重叠区域,所述参考群组具有第一多数健康受试者;

对于所述参考群组中的每个健康受试者,推导与序列读数相关联的一数量,其与所述参考基因组的所述多个非重叠区域内的一区域对齐,从而呈现与所述区域相对应的一第一多数数量;

基于所述第一多数数量,确定一第一参考量和一第二参考量;以及

当所述第一参考量和所述第二参考量满足一预定条件时,将所述区域识别为具有低变异性。

27. 如权利要求26所述的方法,其特征在于:所述方法更包括:

对所述参考基因组的多个非重叠区域中的所有剩余的区域,重复所述确定和识别的步骤,从而识别所述参考基因组中的多个低变异区域。

28. 如权利要求26或27所述的方法,其特征在于:所述数量对应于一健康受试者与所述区域对齐的多个序列读数的一总数。

29. 如权利要求28所述的方法,其特征在于:与所述区域对齐的多个序列读数中的每一个还包括一预定的遗传变异。

30. 如权利要求28所述的方法,其特征在于:与所述区域对齐的多个序列读数中的每一个还包括一表观遗传修饰。

31. 如权利要求30所述的方法,其特征在于:所述表观遗传修饰包括甲基化。

32. 如权利要求26至31任一项所述的方法,其特征在于:所述第一参考量选自下列所组成的群组:平均值、平均数、中位数、标准化平均值,标准化平均数,标准化中位数及其组合。

33. 如权利要求26至31任一项所述的方法,其特征在于:所述第二参考量选自下列所组成的群组:四分位间距、中位数绝对偏差、标准偏差及其组合。

34. 如权利要求26至33任一项所述的方法,其特征在于:所述预定条件包括反映出所述第一参考量和所述第二参考量之间的一差异低于一阈值。

35. 一种分析与一疾病状况有关的多个核酸样本的序列读数的方法,其特征在于:所述方法包括步骤:

从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与一参考基因组中多个低变异性区域中的一区域对齐,其中所述训练组包括多个健康受试者的序列读数和已知患有所述疾病状况的多个患病受试者的序列读数,并且其中来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;

使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了在所述训练群组中所述多个健康受试者的序列读数与多个患病受试者的序列读数之间的差异;

接收与来自一测试受试者的一核酸样本相关的一测试组的序列读数,所述测试受试者的所述疾病状况为未知;以及

根据所述一个或多个参数预测所述测试受试者患有所述疾病状况的一可能性。

36. 如权利要求35所述的方法,其特征在于:所述测试组的序列读数中的所述序列读数包括来自所述受试者的所述核酸样本中游离细胞DNA (cfDNA) 片段的序列读数;

来自所述测试受试者的所述核酸样本包括比一第一阈值长度长的cfDNA片段,其中所述第一阈值长度小于160个核苷酸;以及

所述训练组的序列读数中的所述序列读数不包括大于第一阈值长度的多个cfDNA分子的序列读数。

37. 如权利要求36所述的方法,其特征在于:所述第一阈值长度为140个核苷酸。

38. 如权利要求36或37所述的方法,其特征在于:所述测试组的序列读数中的所述序列读数包括来自所述测试受试者的所述核酸样本中的cfDNA片段的序列读数,所述序列读数的长度介于一第二阈值长度和一第三阈值长度之间;

其中,

所述第二阈值长度是240个核苷酸至260个核苷酸;以及

所述第三阈值长度是290个核苷酸至310个核苷酸。

39. 如权利要求36至38任一项所述的方法,其特征在于:排除长于所述第一阈值长度的cfDNA分子的序列读数是通过将来自所述测试受试者的长于所述第一阈值长度的cfDNA分子与来自所述测试受试者的短于所述第一阈值长度的cfDNA分子进行物理分离来实现的。

40. 如权利要求36至38任一项所述的方法,其特征在于:排除长于所述第一阈值长度的cfDNA分子的序列读数是通过从计算器模拟中筛选出来自所述测试受试者的核酸样本中长度大于第一阈值长度的cfDNA片段的序列读数来实现的。

41. 一种分析与一疾病状况有关的多个核酸样本的序列读数的方法,其特征在于:所述方法包括步骤:

根据来自多个健康受试者的一参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数,识别一参考基因组中的多个低变异区域,其中每个健康受试者的所述第一组的序列读数中的每个序列读数可以与所述参考基因组中的一区域对齐;

从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与所述参考基因组中所述低变异性区域中的一区域对齐,其中所述训练组包括多个健康受试者的序列读数和已知患有所述疾病状况的多个患病受试者的序列读数,并且其中,来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;以及

使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了在所述训练群组中所述多个健康受试者的序列读数与多个患病受试者的序列读数之间的差异。

42. 如权利要求41所述的方法,其特征在于:

所述训练组的序列读数中的所述序列读数包括来自所述训练群组中多个受试者的多个核酸样本中游离细胞DNA (cfDNA) 片段的序列读数;

来自所述训练群组中多个受试者的多个核酸样本包括比一第一阈值长度长的cfDNA片段,其中所述第一阈值长度小于160个核苷酸;以及

所述训练组的序列读数中的所述序列读数不包括大于第一阈值长度的多个cfDNA分子的序列读数。

43. 如权利要求42所述的方法,其特征在于:所述第一阈值长度为140个核苷酸或更少。

44. 如权利要求42或43所述的方法,其特征在于:所述训练组中的所述序列读数包括来自所述训练组中所述受试者的所述核酸样本中的cfDNA片段的序列读数,所述序列读数的长度介于一第二阈值长度和一第三阈值长度之间;

其中,

所述第二阈值长度是240个核苷酸至260个核苷酸;以及

所述第三阈值长度是290个核苷酸至310个核苷酸。

45. 如权利要求42至44任一项所述的方法,其特征在于:排除长于所述第一阈值长度的cfDNA分子的序列读数是通过将来自所述训练群组的所述受试者中长于所述第一阈值长度的cfDNA分子与来自所述训练群组中的所述受试者中短于所述第一阈值长度的cfDNA分子进行物理分离来实现的。

46. 如权利要求21至23任一项所述的方法,其特征在于:排除长于所述第一阈值长度的cfDNA分子的序列读数是通过从计算器模拟中筛选出来自所述训练群组的多组所述受试者的核酸样本中长度大于第一阈值长度的cfDNA片段的序列读数来实现的。

47. 一种计算器系统,其特征在于:所述计算器系统包括:

一个或多个处理器;以及

一非暂时性计算器可读介质,包括一个或多个指令序列,当由一个或多个所述处理器执行时,所述指令序列使所述处理器:

根据来自多个健康受试者的一参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数,识别一参考基因组中的多个低变异区域,其中在来自每个健康受试者的多个核酸样本的所述第一组的序列读数中的每个序列读数可以与所述参考基因组中的一区域对齐;

从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与所述参考基因组中所述低变异性区域中的一区域对齐,其中所述训练组包括来自多个健康受试者的多个核酸样本的序列读数和来自已知患有所述疾病状况的多个患病受试者的多个核酸样本的序列读数,并且其中,来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多组核酸样本的一类型相同或相似;

使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了所述训练群组内所述多个健康受试者的多个核酸样本的序列读数与来自所述多个患病受试者的多个核酸样本的序列读数之间的差异;

接收与来自一测试受试者的多个核酸样本相关的一测试组的序列读数,所述测试受试者的所述疾病状况为未知;以及

根据所述一个或多个参数,预测所述测试受试者患有所述疾病状况的一可能性。

48. 一种非暂时性计算器可读存储介质,其特征在于:所述非暂时性计算器可读存储介质存储有多个程序代码指令,当由一讯息管理服务器的一处理器执行所述程序代码指令时,所述程序代码指令使所述讯息管理服务器执行以下方法:

根据来自多个健康受试者的一参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数,识别一参考基因组中的多个低变异区域,其中,在来自每个健康受试者的多个核酸样本的所述第一组的序列读数中的每个序列读数可以与所述参考基因组中的一区域对齐;

从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与所述参考基因组中所述低变异性区域中的一区域对齐,其中所述训练组包括来自多个健康受试者的多个核酸样本的序列读数和来自已知患有所述疾病状况的多个患病受试者的多个核酸样本的序列读数,并且其中来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多组核酸样本

的一类型相同或相似；

使用从所述训练组的序列读数中得到的多个数量，确定一个或多个参数，所述参数反映了所述训练群组内所述多个健康受试者的多个核酸样本的序列读数与来自所述多个患病受试者的多个核酸样本的序列读数之间的差异；

接收与来自一测试受试者的多个核酸样本相关的一测试组的序列读数，所述测试受试者的所述疾病状况为未知；以及

根据所述一个或多个参数，预测所述测试受试者患有所述疾病状况的一可能性。

49. 一种计算器系统，其特征在于：所述计算器系统包括：

一个或多个处理器；以及

一非暂时性计算器可读介质，包括一个或多个指令序列，当由一个或多个所述处理器执行时，所述指令序列使所述处理器：

将来自所述参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数的多个序列对准所述参考基因组的多个非重叠区域，所述参考群组包括一第一多数健康受试者；

对于所述参考群组中的每个健康受试者，推导与多个序列读数相关联的一数量，其与所述参考基因组的所述多个非重叠区域内的一区域对齐，从而呈现与所述区域相对应的一第一多数数量；

基于所述第一多数数量确定一第一参考量和一第二参考量；以及

当所述第一参考量和所述第二参考量满足一预定条件时，将所述参考基因组的所述区域识别为具有低变异性。

50. 一种非暂时性计算器可读存储介质，其特征在于：所述非暂时性计算器可读存储介质存储有多个程序代码指令，当由一讯息管理服务器的一处理器执行所述程序代码指令时，所述程序代码指令使所述讯息管理服务器执行以下方法：

将来自所述参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数的多个序列对准所述参考基因组的多个非重叠区域，所述参考群组包括一第一多数健康受试者；

对于所述参考群组中的每个健康受试者，推导与多个序列读数相关联的一数量，其与所述参考基因组的所述多个非重叠区域内的一区域对齐，从而呈现与所述区域相对应的一第一多数数量；

基于所述第一多数数量确定一第一参考量和一第二参考量；以及

当所述第一参考量和所述第二参考量满足一预定条件时，将所述参考基因组的所述区域识别为具有低变异性。

51. 一种计算器系统，其特征在于：所述计算器系统包括：

一个或多个处理器；以及

一非暂时性计算器可读介质，包括一个或多个指令序列，当由一个或多个所述处理器执行时，所述指令序列使所述处理器：

从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数，其中在所述训练组中的每个序列读数与一参考基因组中多个低变异性区域中的一区域对齐，其中所述训练组包括多个健康受试者的序列读数和已知患有所述疾病状况的多个患病

受试者的序列读数,并且其中来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;

使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了在所述训练群组中所述多个健康受试者的序列读数与多个患病受试者的序列读数之间的差异;

接收与来自一测试受试者的一核酸样本相关的一测试组的序列读数,所述测试受试者的所述疾病状况为未知;以及

根据所述一个或多个参数,预测所述测试受试者患有所述疾病状况的一可能性。

52. 一种非暂时性计算器可读存储介质,其特征在于:所述非暂时性计算器可读存储介质存储有多个程序代码指令,当由一讯息管理服务器的一处理器执行所述程序代码指令时,所述程序代码指令使所述讯息管理服务器执行以下方法:

从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与一参考基因组中多个低变异性区域中的一区域对齐,其中所述训练组包括多个健康受试者的序列读数和已知患有所述疾病状况的多个患病受试者的序列读数,并且其中,来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;

使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了在所述训练群组中所述多个健康受试者的序列读数与多个患病受试者的序列读数之间的差异;

接收与来自一测试受试者的一核酸样本相关的一测试组的序列读数,所述测试受试者的所述疾病状况为未知;以及

根据所述一个或多个参数,预测所述测试受试者患有所述疾病状况的一可能性。

53. 一种计算器系统,其特征在于:所述计算器系统包括:

一个或多个处理器;以及

一非暂时性计算器可读介质,包括一个或多个指令序列,当由一个或多个所述处理器执行时,所述指令序列使所述处理器:

根据来自多个健康受试者的一参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数,识别一参考基因组中的多个低变异区域,其中,每个健康受试者的所述第一组的序列读数中的每个序列读数可以与所述参考基因组中的一区域对齐;

从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与所述参考基因组中所述低变异性区域中的一区域对齐,其中所述训练组包括多个健康受试者的序列读数和已知患有所述疾病状况的多个患病受试者的序列读数,并且其中来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;以及

使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了在所述训练群组中所述多个健康受试者的序列读数与多个患病受试者的序列读数之间的差异。

54. 一种非暂时性计算器可读存储介质,其特征在于:所述非暂时性计算器可读存储介质存储有多个程序代码指令,当由一讯息管理服务器的一处理器执行所述程序代码指令

时,所述程序代码指令使所述讯息管理服务器执行以下方法:

根据来自多个健康受试者的一参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数,识别一参考基因组中的多个低变异区域,其中,每个健康受试者的所述第一组的序列读数中的每个序列读数可以与所述参考基因组中的一区域对齐;

从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与所述参考基因组中所述低变异性区域中的一区域对齐,其中所述训练组包括多个健康受试者的序列读数和已知患有所述疾病状况的多个患病受试者的序列读数,并且其中,来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;以及

使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了在所述训练群组中所述多个健康受试者的序列读数与多个患病受试者的序列读数之间的差异。

55.一种用于分析与一疾病状况相关的多个核酸样本的序列读数的计算器程序产品,其特征在于:所述计算器程序产品包括:

一种非暂时性计算器可读介质,存储用于执行如权利要求1至46中任一项所述的方法的多个指令。

选择、管理和分析高维数据的方法和系统

[0001] 相关申请的交叉引用

[0002] 本申请主张于2018年3月13日提交的美国专利临时申请案申请号为62/642,461的优先权,其公开内容通过引用并入本文作为参考。

技术领域

[0003] 本发明是有关于一种用于选择和分析高维生物学数据的方法、系统和计算程序产品,特别是用于使用下一代测序技术获得的核酸测序数据。

背景技术

[0004] 生物学的现代发展,特别是下一代测序技术,已经产生了大量的数据。然而,对数据进行梳理以得到有用和有帮助的信息仍然是一个巨大的挑战,尤其是在疾病诊断和预后需要这种有用和有帮助的信息时。例如,人类基因组包括超过30亿个碱基对的核酸序列。尽管有可能获得整个人基因组的序列读数,但许多测序数据仍编码与疾病诊断和预后无关的信息。

[0005] 有必要提供一种处理大数据的方式,以便有效、准确地得出有用和相关的信息。

发明内容

[0006] 在一目的,本文公开了一种分析与一疾病状况相关的多个核酸样本的序列读数的方法。如本文所公开的,所述方法包括步骤:根据来自多个健康受试者的一参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数,识别一参考基因组中的多个低变异区域,其中在来自每个健康受试者的多个核酸样本的所述第一组的序列读数中的每个序列读数可以与所述参考基因组中的一区域对齐;从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与所述参考基因组中所述低变异性区域中的一区域对齐,其中所述训练组包括来自多个健康受试者的多个核酸样本的序列读数和来自已知患有所述疾病状况的多个患病受试者的多个核酸样本的序列读数,并且其中来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了所述训练群组内所述多个健康受试者的多个核酸样本的序列读数与来自所述多个患病受试者的多个核酸样本的序列读数之间的差异;接收与来自一测试受试者的多个核酸样本相关的一测试组的序列读数,所述测试受试者的所述疾病状况为未知;以及根据所述一个或多个参数,预测所述测试受试者患有所述疾病状况的一可能性。

[0007] 在一些实施例中,所述多个核酸样本包含多个游离核酸(cfDNA)片段。

[0008] 在一些实施例中,所述疾病状况为癌症。

[0009] 在一些实施例中,所述疾病状况是选自下列群组中组成的一癌症类型:肺癌、卵巢癌、肾癌、膀胱癌、肝胆癌、胰腺癌、上消化道癌、肉瘤、乳腺癌、肝癌、前列腺癌、脑癌及其组

合。

[0010] 在一些实施例中,所述方法更包括:根据来自多个健康受试者的一基线群组的多个核酸样本的序列读数,对多个健康受试者的所述参考群组中每个健康受试者的多个核酸样本的所述第一组的序列读数进行初始数据处理,其中所述参考群组和所述基线群组不重迭,并且其中所述初始数据处理包括校正GC偏差或对与所述参考基因组的多个区域对齐的序列读数进行的标准化。

[0011] 在一些实施例中,所述方法更包括:根据来自多个健康受试者的一基线群组的多个核酸样本的序列读数,所述训练群组中每个受试者的多个核酸样本的序列读数进行初始数据处理,其中所述基线群组和所述训练群组不重迭,并且其中所述初始数据处理包括校正GC偏差或对与所述参考基因组的多个区域对齐的序列读数进行的标准化。

[0012] 在一些实施例中,识别所述参考基因组中的所述多个低变异区域的步骤还包括:将来自多个健康受试者的所述参考群组中的每个健康受试者的多个核酸样本的所述第一组的序列读数的多个序列对准所述参考基因组的多个非重叠区域,所述参考群组包括一第一多数健康受试者;对于所述参考群组中的每个健康受试者,推导与多个序列读数相关联的一数量,其与所述参考基因组的所述多个非重叠区域内的一区域对齐,从而呈现与所述区域相对应的一第一多数数量;基于所述第一多数数量确定一第一参考量和一第二参考量;以及当所述第一参考量和所述第二参考量满足一预定条件时,将所述区域识别为具有低变异性。

[0013] 在一些实施例中,所述方法更包括:对所述参考基因组的多个非重叠区域中的所有剩余的区域,重复所述确定和识别的步骤,从而识别所述参考基因组中的所述多个低变异区域。

[0014] 在一些实施例中,从所述训练群组的多个核酸样本的多个序列读数中选择所述训练组的序列读数的步骤还包括:从所述训练群组的多个核酸样本的多个序列读数中选择多个序列读数,所述序列读数与所述参考基因组中的多个低变异性区域对齐,从而生成所述训练组的序列读数。

[0015] 在一些实施例中,确定一个或多个参数的步骤还包括:对于所述训练群组中的每个受试者,并且针对所述多个低变异性区域中的一区域,基于与所述区域对齐的序列读数,推导一个或多个数量;对所有剩余的低变异性区域,重复所述推导的步骤,以呈现与所述训练群组中所有受试者的低变异性区域相对应的多个数量,其中所述多个数量包括:与多个健康受试者有关的多个数量的一第一子集,和与已知患有所述疾病状况的多个受试者有关的多个数量的一第二子集;以及确定用于反映多个数量的一第一子集和一第二子集之间的差异的一个或多个参数。

[0016] 在一些实施例中,所述一个或多个数量由一个数量组成,并对应于与所述区域对齐的序列读数的总数。

[0017] 在一些实施例中,所述一个或多个数量包括多个数量,每个数量对应于与所述区域对齐的多个序列读数的一子集,其中在相同子集中的每个序列读数对应于具有相同预定片段尺寸或尺寸范围的核酸样本,其中多个不同子集中的多个序列读数对应于具有不同片段尺寸或尺寸范围的多个核酸样本。

[0018] 在一些实施例中,所述一个或多个参数由主成分分析(PCA)确定。

[0019] 在一些实施例中,所述方法更包括:通过将所述训练组划分为一训练子集和一验证子集,在一多重交叉验证过程中改进所述一个或多个参数。

[0020] 在一些实施例中,所述多重交叉验证过程中的一环节中的训练子集和验证子集不同于所述多重交叉验证过程的另一环节中的不同的训练子集和验证子集。

[0021] 在一些实施例中,所述方法更包括:从来自所述受试者的多个核酸样本的多个序列读数中选择多个序列读数,所述序列读数与所述参考基因组中的低变异区域对齐,从而生成所述测试组的序列读数;以及基于所述测试组的序列读数和所述一个或多个参数,计算代表受试者患疾病状况的可能性的一分类分数。

[0022] 在一些实施例中,所述参考基因组中每个变异区域的大小在1万碱基对至10万碱基对之间。在一些实施例中,所述参考基因组中每个变异区域的大小相同。在一些实施例中,所述参考基因组中的多个所述变异区域的大小不相同。

[0023] 在一些实施例中,所述一个或多个参数基于所述训练组的序列读数的一子集来确定。

[0024] 在一些实施例中,所述训练组的序列读数中的所述序列读数包括来自所述训练群组中多个受试者的多个核酸样本中游离细胞DNA (cfDNA) 片段的序列读数;来自所述训练群组中多个受试者的多个核酸样本包括比一第一阈值长度长的cfDNA片段,例如其中所述第一阈值长度小于160个核苷酸;以及所述训练组的序列读数中的所述序列读数不包括大于第一阈值长度的多个cfDNA分子的序列读数。在一些实施例中,所述第一阈值长度为140个核苷酸或更少。

[0025] 在一些实施例中,所述训练组中的所述序列读数包括来自所述训练组中所述受试者的所述核酸样本中的cfDNA片段的序列读数,所述序列读数的长度介于一第二阈值长度和一第三阈值长度之间;其中,所述第二阈值长度是240个核苷酸至260个核苷酸;以及所述第三阈值长度是290个核苷酸至310个核苷酸。

[0026] 在一些实施例中,排除长于所述第一阈值长度的cfDNA分子的序列读数是来自所述训练群组的所述受试者中长于所述第一阈值长度的cfDNA分子与来自所述训练群组中的所述受试者中短于所述第一阈值长度的cfDNA分子进行物理分离来实现的。

[0027] 在一些实施例中,排除长于所述第一阈值长度的cfDNA分子的序列读数是来自所述训练群组的所述受试者中短于所述第一阈值长度的cfDNA分子进行物理分离来实现的。

[0028] 尽管在此针对结合疾病状况分析核酸样本的序列读数的特定方法进行了描述,但是核酸序列读数(例如cfDNA序列读数)的尺寸选择可以结合以下方面进行应用:例如使用如方法200、210、300、310、400、500、600、1200、1300和1400中的一种或多种来实现本揭示。尺寸选择方法的进一步描述可以在例如美国专利临时申请案申请号为62/818,013中找到,申请日为2019年3月13日,所述申请案的名称为“使用片段尺寸扩增癌症衍生性片段的系统和方法”,所述专利的全部内容通过引用并入本文。

[0029] 在一目的,本文公开了一种基于来自一参考群组中多个健康受试者的测序数据识别一参考基因组中多个低变异区域的方法,例如,所述方法包括步骤:将来自所述参考群组中每个健康受试者的多个核酸样本的多个序列读数的第一组的多个序列对准所述参考基因组的多个非重叠区域,所述参考群组具有第一多数健康受试者;对于所述参考群组中的

每个健康受试者,推导与序列读数相关联的一数量,其与所述参考基因组的所述多个非重叠区域内的一区域对齐,从而呈现与所述区域相对应的一第一多数数量;基于所述第一多数数量,确定一第一参考量和一第二参考量;以及当所述第一参考量和所述第二参考量满足一预定条件时,将所述区域识别为具有低变异性。

[0030] 在一些实施例中,所述方法更包括:对所述参考基因组的多个非重叠区域中的所有剩余的区域,重复所述确定和识别的步骤,从而识别所述参考基因组中的多个低变异区域。

[0031] 在一些实施例中,所述数量对应于一健康受试者与所述区域对齐的多个序列读数的一总数。

[0032] 在一些实施例中,与所述区域对齐的多个序列读数中的每一个还包括一预定的遗传变异。

[0033] 在一些实施例中,与所述区域对齐的多个序列读数中的每一个还包括一表观遗传修饰。在一些实施例中,所述表观遗传修饰包括甲基化。

[0034] 在一些实施例中,所述第一参考量选自下列所组成的群组:平均值、平均数、中位数、标准化平均值,标准化平均数,标准化中位数及其组合。

[0035] 在一些实施例中,所述第二参考量选自下列所组成的群组:四分位间距、中位数绝对偏差、标准偏差及其组合。

[0036] 在一些实施例中,所述预定条件包括反映出所述第一参考量和所述第二参考量之间的一差异低于一阈值。

[0037] 在一目的,本文公开了一种分析与一疾病状况有关的多个核酸样本的序列读数的方法,例如,所述方法包括步骤:从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与一参考基因组中多个低变异性区域中的一区域对齐,其中所述训练组包括多个健康受试者的序列读数和已知患有所述疾病状况的多个患病受试者的序列读数,并且其中,来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了在所述训练群组中所述多个健康受试者的序列读数与多个患病受试者的序列读数之间的差异;接收与来自一测试受试者的一核酸样本相关的一测试组的序列读数,所述测试受试者的所述疾病状况为未知;以及根据所述一个或多个参数预测所述测试受试者患有所述疾病状况的一可能性。

[0038] 在一些实施例中,所述测试组的序列读数中的所述序列读数包括来自所述受试者的所述核酸样本中游离细胞DNA (cfDNA) 片段的序列读数;来自所述测试受试者的所述核酸样本包括比一第一阈值长度长的cfDNA片段,其中所述第一阈值长度小于160个核苷酸;以及所述训练组的序列读数中的所述序列读数不包括大于第一阈值长度的多个cfDNA分子的序列读数。在一些实施例中,所述第一阈值长度为140个核苷酸。

[0039] 在一些实施例中,所述测试组的序列读数中的所述序列读数包括来自所述测试受试者的所述核酸样本中的cfDNA片段的序列读数,所述序列读数的长度介于一第二阈值长度和一第三阈值长度之间;其中,所述第二阈值长度是240个核苷酸至260个核苷酸;以及所述第三阈值长度是290个核苷酸至310个核苷酸。

[0040] 在一些实施例中,排除长于所述第一阈值长度的cfDNA分子的序列读数是通过将来自所述测试受试者的长于所述第一阈值长度的cfDNA分子与来自所述测试受试者的短于所述第一阈值长度的cfDNA分子进行物理分离来实现的。

[0041] 在一些实施例中,排除长于所述第一阈值长度的cfDNA分子的序列读数是通过从计算器模拟中筛选出来自所述测试受试者的核酸样本中长度大于第一阈值长度的cfDNA片段的序列读数来实现的。

[0042] 尽管这里描述了用于分析与疾病状况相关的核酸样本的序列读数的特定方法,但是核酸序列读数的尺寸选择,例如cfDNA序列读数,可以与本发明的任何目的一起应用,例如,与方法200、210、300、310、400、500、600、1200、1300和1400中的一种或多种相结合。

[0043] 在一目的,本文公开了一种分析与一疾病状况有关的多个核酸样本的序列读数的方法,例如,所述方法包括步骤:根据来自多个健康受试者的一参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数,识别一参考基因组中的多个低变异区域,其中,每个健康受试者的所述第一组的序列读数中的每个序列读数可以与所述参考基因组中的一区域对齐;从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与所述参考基因组中所述低变异性区域中的一区域对齐,其中所述训练组包括多个健康受试者的序列读数和已知患有所述疾病状况的多个患病受试者的序列读数,并且其中,来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;以及使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了在所述训练群组中所述多个健康受试者的序列读数与多个患病受试者的序列读数之间的差异。

[0044] 在一些实施例中,所述训练组的序列读数中的所述序列读数包括来自所述训练群组中多个受试者的多个核酸样本中游离细胞DNA(cfDNA)片段的序列读数;来自所述训练群组中多个受试者的多个核酸样本包括比一第一阈值长度长的cfDNA片段,其中所述第一阈值长度小于160个核苷酸;以及所述训练组的序列读数中的所述序列读数不包括大于第一阈值长度的多个cfDNA分子的序列读数。在一些实施例中,所述第一阈值长度为140个核苷酸或更少。

[0045] 在一些实施例中,所述训练组中的所述序列读数包括来自所述训练组中所述受试者的所述核酸样本中的cfDNA片段的序列读数,所述序列读数的长度介于一第二阈值长度和一第三阈值长度之间;其中,所述第二阈值长度是240个核苷酸至260个核苷酸;以及所述第三阈值长度是290个核苷酸至310个核苷酸。

[0046] 在一些实施例中,排除长于所述第一阈值长度的cfDNA分子的序列读数是通过将来自所述训练群组的所述受试者中长于所述第一阈值长度的cfDNA分子与来自所述训练群组中的所述受试者中短于所述第一阈值长度的cfDNA分子进行物理分离来实现的。

[0047] 在一些实施例中,其特征在于:排除长于所述第一阈值长度的cfDNA分子的序列读数是通过对从计算器模拟中筛选出来自所述训练群组的多个所述受试者的核酸样本中长度大于第一阈值长度的cfDNA片段的序列读数来实现的。

[0048] 尽管这里描述了用于分析与疾病状况相关的核酸样本的序列读数的特定方法,但是核酸序列读数的尺寸选择,例如cfDNA序列读数,可以与本发明的任何目的一起应用,例如,与方法200、210、300、310、400、500、600、1200、1300和1400中的一种或多种相结合。

[0049] 在一目的中,本文公开了一种计算器系统,其包括:一个或多个处理器;以及一非暂时性计算器可读介质,包括一个或多个指令序列,当由一个或多个所述处理器执行时,所述指令序列使所述处理器:根据来自多个健康受试者的一参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数,识别一参考基因组中的多个低变异区域,其中,在来自每个健康受试者的多个核酸样本的所述第一组的序列读数中的每个序列读数可以与所述参考基因组中的一区域对齐;从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与所述参考基因组中所述低变异性区域中的一区域对齐,其中所述训练组包括来自多个健康受试者的多个核酸样本的序列读数和来自已知患有所述疾病状况的多个患病受试者的多个核酸样本的序列读数,并且其中,来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了所述训练群组内所述多个健康受试者的多个核酸样本的序列读数与来自所述多个患病受试者的多个核酸样本的序列读数之间的差异;接收与来自一测试受试者的多个核酸样本相关的一测试组的序列读数,所述测试受试者的所述疾病状况为未知;以及根据所述一个或多个参数,预测所述测试受试者患有所述疾病状况的一可能性。

[0050] 在一目的中,本文公开了一种非暂时性计算器可读存储介质,所述非暂时性计算器可读存储介质存储有多个程序代码指令,当由一信息管理服务器的一处理器执行所述程序代码指令时,所述程序代码指令使所述信息管理服务器执行以下方法:根据来自多个健康受试者的一参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数,识别一参考基因组中的多个低变异区域,其中,在来自每个健康受试者的多个核酸样本的所述第一组的序列读数中的每个序列读数可以与所述参考基因组中的一区域对齐;从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与所述参考基因组中所述低变异性区域中的一区域对齐,其中所述训练组包括来自多个健康受试者的多个核酸样本的序列读数和来自已知患有所述疾病状况的多个患病受试者的多个核酸样本的序列读数,并且其中,来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了所述训练群组内所述多个健康受试者的多个核酸样本的序列读数与来自所述多个患病受试者的多个核酸样本的序列读数之间的差异;接收与来自一测试受试者的多个核酸样本相关的一测试组的序列读数,所述测试受试者的所述疾病状况为未知;以及根据所述一个或多个参数,预测所述测试受试者患有所述疾病状况的一可能性。

[0051] 在一目的中,本文公开了一种计算器系统,其包括:一个或多个处理器;以及一非暂时性计算器可读介质,包括一个或多个指令序列,当由一个或多个所述处理器执行时,所述指令序列使所述处理器:将来自所述参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数的多个序列对准所述参考基因组的多个非重叠区域,所述参考群组包括一第一多数健康受试者;对于所述参考群组中的每个健康受试者,推导与多个序列读数相关联的一数量,其与所述参考基因组的所述多个非重叠区域内的一区域对齐,从而呈现与所述区域相对应的一第一多数数量;基于所述第一多数数量确定一第一参考量和一第二参

考量;以及当所述第一参考量和所述第二参考量满足一预定条件时,将所述参考基因组的所述区域识别为具有低变异性。

[0052] 在一目的中,本文公开了一种非暂时性计算器可读存储介质,所述非暂时性计算器可读存储介质存储有多个程序代码指令,当由一信息管理服务器的一处理器执行所述程序代码指令时,所述程序代码指令使所述信息管理服务器执行以下方法:将来自所述参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数的多个序列对准所述参考基因组的多个非重叠区域,所述参考群组包括一第一多数健康受试者;对于所述参考群组中的每个健康受试者,推导与多个序列读数相关联的一数量,其与所述参考基因组的所述多个非重叠区域内的一区域对齐,从而呈现与所述区域相对应的一第一多数数量;基于所述第一多数数量确定一第一参考量和一第二参考量;以及当所述第一参考量和所述第二参考量满足一预定条件时,将所述参考基因组的所述区域识别为具有低变异性。

[0053] 在一目的中,本文公开了一种计算器系统,其包括:一个或多个处理器;以及一非暂时性计算器可读介质,包括一个或多个指令序列,当由一个或多个所述处理器执行时,所述指令序列使所述处理器:从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与一参考基因组中多个低变异性区域中的一区域对齐,其中所述训练组包括多个健康受试者的序列读数和已知患有所述疾病状况的多个患病受试者的序列读数,并且其中来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了在所述训练群组中所述多个健康受试者的序列读数与多个患病受试者的序列读数之间的差异;接收与来自一测试受试者的一核酸样本相关的一测试组的序列读数,所述测试受试者的所述疾病状况为未知;以及根据所述一个或多个参数,预测所述测试受试者患有所述疾病状况的一可能性。

[0054] 在一目的中,本文公开了一种非暂时性计算器可读存储介质,所述非暂时性计算器可读存储介质存储有多个程序代码指令,当由一信息管理服务器的一处理器执行所述程序代码指令时,所述程序代码指令使所述信息管理服务器执行以下方法:从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与一参考基因组中多个低变异性区域中的一区域对齐,其中所述训练组包括多个健康受试者的序列读数和已知患有所述疾病状况的多个患病受试者的序列读数,并且其中,来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了在所述训练群组中所述多个健康受试者的序列读数与多个患病受试者的序列读数之间的差异;接收与来自一测试受试者的一核酸样本相关的一测试组的序列读数,所述测试受试者的所述疾病状况为未知;以及根据所述一个或多个参数,预测所述测试受试者患有所述疾病状况的一可能性。

[0055] 在一目的中,本文公开了一种计算器系统,其包括:一个或多个处理器;以及一非暂时性计算器可读介质,包括一个或多个指令序列,当由一个或多个所述处理器执行时,所述指令序列使所述处理器:根据来自多个健康受试者的一参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数,识别一参考基因组中的多个低变异区域,其中,每个

健康受试者的所述第一组的序列读数中的每个序列读数可以与所述参考基因组中的一区域对齐;从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与所述参考基因组中所述低变异性区域中的一区域对齐,其中所述训练组包括多个健康受试者的序列读数和已知患有所述疾病状况的多个患病受试者的序列读数,并且其中来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;以及使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了在所述训练群组中所述多个健康受试者的序列读数与多个患病受试者的序列读数之间的差异。

[0056] 在一目的中,本文公开了一种非暂时性计算机可读存储介质,所述非暂时性计算机可读存储介质存储有多个程序代码指令,当由一信息管理服务器的一处理器执行所述程序代码指令时,所述程序代码指令使所述信息管理服务器执行以下方法:根据来自多个健康受试者的一参考群组中的每个健康受试者的多个核酸样本的一第一组的序列读数,识别一参考基因组中的多个低变异区域,其中,每个健康受试者的所述第一组的序列读数中的每个序列读数可以与所述参考基因组中的一区域对齐;从一训练群组中多个受试者的多个核酸样本的序列读数中选择一训练组的序列读数,其中在所述训练组中的每个序列读数与所述参考基因组中所述低变异性区域中的一区域对齐,其中所述训练组包括多个健康受试者的序列读数和已知患有所述疾病状况的多个患病受试者的序列读数,并且其中,来自所述训练群组的所述多个核酸样本的一类型与来自多个健康受试者的所述参考群组的多个核酸样本的一类型相同或相似;以及使用从所述训练组的序列读数中得到的多个数量,确定一个或多个参数,所述参数反映了在所述训练群组中所述多个健康受试者的序列读数与多个患病受试者的序列读数之间的差异。

[0057] 在一目的中,使用一个或多个噪声模型来识别和排除属于噪声的序列读数;例如,可能由非癌源引起的序列读数。例如,复制次数变异可由白血球中的复制性造血而不是体细胞肿瘤细胞引起。在一些实施例中,基于使用来自基因组核酸样本的测序数据构建的一个或多个噪声模型来识别cfDNA样本中的复制次数畸变;例如,从白血球层中的白血球(WBCs)获得的基因组DNA:(gDNA)。

[0058] 在一些实施例中,基于白血球的噪声模型用于进一步消除与体细胞肿瘤细胞无关的复制次数变异。在一些实施例中,将基于白血球的噪声模型应用于图2A中的步骤202处的数据预处理的一部分。在一些实施例中,基于白血球的噪声模型在数据预处理步骤202之后应用。在一些实施例中,在图2A中的数据选择步骤204之后应用基于白血球的噪声模型。

[0059] 在一目的中,本文提供一种计算机程序产品,其包括:存储用于执行本文所公开的任何方法的指令的非临时计算机可读介质。

[0060] 应当理解的是,本文中公开的任何一个实施例都可以结合本发明的任何方面单独使用,或者与一个或多个其他实施例组合使用。

[0061] 应当理解,在适用的情况下,本文所公开的任何实施例可以单独或以任何组合应用于本发明的任何目的。

[0062] 一个或多个实施的细节将在附图和下面的描述中阐述。根据说明书和附图以及根据权利要求书,其他特征、目的和潜在优点将变得显而易见。

附图说明

[0063] 本领域技术人员将理解,下面描述的附图仅用于说明性目的。这些附图无意以任何方式限制本揭示的范围。

[0064] 图1A是处理高维数据的一样本系统示意图。

[0065] 图1B是一处理高维数据的软件平台实例示意图。

[0066] 图2A是一个实施例过程,说明了处理高维数据的总体方法流程示意图。

[0067] 图2B是一个示例实施例,说明了处理高维数据时的信息流程示意图。

[0068] 图3A描绘了用于数据选择的样本处理。

[0069] 图3B描绘了用于数据选择的样本处理。

[0070] 图3C描绘了用于数据选择的样本处理。

[0071] 图4描绘了用于分析数据以减少数据维数的示例过程。

[0072] 图5描绘了用于基于从具有减小的维度的数据中获知的信息来分析数据的示例过程。

[0073] 图6描绘了根据本揭示的用于数据分析的示例过程。

[0074] 图7描绘了用于实施图1至图6的特征和流程的示例系统架构示意图。

[0075] 图8描绘了本揭示方法的示例性结果。

[0076] 图9描绘了本揭示方法的示例性结果。

[0077] 图10描绘了本揭示方法的示例性结果。

[0078] 图11A描绘了本揭示方法的示例性结果。

[0079] 图11B描绘了本揭示方法的示例性结果。

[0080] 图12A描绘了用于识别在cfDNA样本中识别出的复制次数变异的来源的样本处理。

[0081] 图12B描绘了根据本揭示的实施例的用于识别衍生自cfDNA和gDNA样本的统计学上显著的宏基因组和统计学上显著的片段的样本处理过程。

[0082] 图12C描绘了根据本发明的一个实施例的用于一训练特征数据库的示例系统架构示意图。

[0083] 图13根据本发明的各种实施例提供了一种使用从受试者的生物样本中的体外尺寸选择的游离DNA来确定一受试者的一癌症状态的方法的流程示意图。

[0084] 图14提供了根据本发明的各种实施例中使用从受试者的生物样本中经由计算机仿真尺寸选择的游离DNA的选择序列读数来确定受试者的癌症状态的流程图。

[0085] 图15显示了受试者的多个游离DNA片段长度的平均分布,绘制为受试者肿瘤分数的函数,如实施例4所述。50%至100%肿瘤分数队列的数据来自于具有转移性癌症的患者的单个样本。

[0086] 图16A、图16B和图16C描绘了未经筛选(图16A)时,从cfDNA样本的全基因组测序中所获得的、在电子显微镜下筛选,仅包括来自大小为90个核苷酸到150个核苷酸的cfDNA片段的序列(图16B),如实施例5所述、以及在电子显微镜中筛选以仅包括来自大小为100个核苷酸或更小的cfDNA片段的序列(图16C)的测序覆盖直方图。

[0087] 图17A显示了使用完整(未筛选)CCGA WGS数据集、尺寸选择的(已筛选)CCGA WGS数据集、以及具有从完整(未筛选)CCG AWGS数据集中随机选择的序列读数的对照组数据集的癌症分类特异性的方框图,以匹配尺寸选择的(已筛选)CCG AWGS数据集的序列覆盖范

围,如实施例6所述。

[0088] 图17B显示了使用完整(未筛选)CCGA WGS数据集、尺寸选择的(已筛选)CCGA WGS数据集、以及具有从完整(未筛选)CCG AWGS数据集中随机选择的序列读数的对照组数据集的癌症分类特异性的方框图,以匹配尺寸选择的(已筛选)CCG AWGS数据集的序列覆盖范围,如实施例7所述。

[0089] 图17C和图17D显示了使用完整(未筛选)CCGA WGS数据集、尺寸选择的(已筛选)CCGA WGS数据集、以及对照组数据集,其具有从完整(未筛选)CCG AWGS数据集中随机选择的序列读数的癌症分类特异性的方框图,以匹配针对每个癌症阶段的尺寸选择的(已筛选)CCG AWGS数据集的序列覆盖,如实施例7所述。

[0090] 图17E显示了如实施例7中所述的使用计算器尺寸选择的序列读数进行的图16C和图16D所示分类的癌症阶段依赖性统计。

[0091] 图17F和图17G展示了使用完整(未筛选)CCGA WGS数据集、尺寸选择的(已筛选)CCGA WGS数据集,以及对照组数据集,其具有从完整(未筛选)CCG AWGS数据集中随机选择的序列读数的癌症分类特异性的方框图,以匹配尺寸选择的(已筛选)CCG AWGS数据集的序列覆盖范围,针对每一个癌症阶段流入的血液,如实施例7所述。

[0092] 图17H显示如实施例7中所述使用计算器模拟尺寸选择的序列读数进行图17F和图17G所示分类的癌症阶段依赖性统计。

[0093] 图18显示如实施例8中所述在cfDNA库的体外尺寸选择后产生的片段计数。

[0094] 图19显示了如实施例9所述在体外尺寸选择之前(x轴)和之后(y轴)样本中癌症来源的cfDNA片段的估计分数(肿瘤分数)。

[0095] 图20显示了使用从完整cfDNA样本和体外尺寸选择的cfDNA样本中的序列读数产生的分类分数,如实施例10中所述,绘制为样本原始肿瘤分数的函数。

[0096] 各附图中的类似参考标号表示相似的组件。

具体实施方式

[0097] 定义

[0098] 如本文所用术语“高维数据”是指如此庞大和复杂以至于传统数据处理应用软件不足以处理它们的数据集。例如,一个普通的人类基因组包括大约20,000个基因,每个单倍体基因组编码超过32亿个碱基对的核酸序列,每个二倍体基因组编码大约65亿个碱基对的核酸序列。在一些实施例中,即使从每个样本收集的数据可能受到限制,大量样本也可以导致高数据维度。如本文所用术语“高维数据”可包括靶向测序数据、全基因组测序数据、揭示表观遗传修饰(例如甲基化)的测序数据及其组合。在一些实施例中,“高维数据”可以包括核酸测序数据和蛋白质测序数据。在一些实施例中,“高维数据”可以包括非生物数据。在本发明的整个过程中,核酸测序数据被用作说明,这不应被解释为对本发明范围的限制。

[0099] 如本文所揭示术语“受试者”是指任何生物或非生物有机体,包括但不限于人类(例如人类男性、人类女性、胎儿、怀孕女性、儿童等)、非人类动物、植物、细菌、真菌或原生生物。可选择任何人类或非人类动物,包括但不限于哺乳动物、爬行动物、鸟类、两栖动物、鱼类、有蹄类动物、反刍动物、牛科动物(如牛)、马科动物(如马)、羊(绵羊、山羊)、猪(如猪)、骆驼科动物(如骆驼、骆驼、羊驼)、猴、猿(如大猩猩、黑猩猩)、熊科动物(如熊)、家禽、

狗、猫、小鼠类、大鼠、鱼、海豚、鲸鱼和鲨鱼。受试者可以是任何阶段的男性或女性(例如男人、女人或儿童)。

[0100] 如本文所揭示的术语“生物样本”是指从受试者身上提取的能够反映与所述受试者相关联的生物状态的任何样本。生物样本的实例包括但不限于组织样本、毛发样本、血液样本、血清样本、血浆样本、泪液样本、汗液样本、尿液样本、唾液样本等。在一些实施例中,生物样本包括核酸分子,例如DNA或RNA。在一些实施例中,生物样本包括蛋白质分子。

[0101] 如本文所用术语“核酸”和“核酸分子”可互换使用。这些术语是指任何组成形式的核酸,如脱氧核糖核酸(DNA,例如互补DNA(cDNA)、基因组DNA(gDNA)等)、核糖核酸(RNA,例如信息RNA(mRNA)、短抑制RNA(siRNA)、核糖体RNA(rRNA)、转移RNA(tRNA)、microRNA、胎儿或胎盘中高度表现的RNA,及/或DNA或RNA类似物(例如含有碱类似物、糖类似物及/或非天然主链等)、RNA/DNA杂交物和聚酰胺核酸(PNA),所有这都可以是单链或双链形式。除非另有限制,核酸可包含已知的天然核苷酸类似物,其中一些类似物的功能与自然产生的核苷酸类似。核酸可以是任何有助于在本文中进行过程的形式(例如线性、圆形、超螺旋、单链、双链等)。在某些实施例中,核酸可以是或可以来自质粒、噬菌体、自主复制序列(ARS)、着丝粒、人工染色体、染色体或其他能够在体外或在宿主细胞、细胞、细胞核或细胞浆中复制或复制的核酸。在一些实施例中,核酸可以来自单条染色体或其片段(例如,核酸样本可以来自二倍体生物体的样本的一个染色体)。在某些实施例中,核酸包含核小体,核小体的片段或部分或核小体类似结构。核酸有时包含蛋白质(例如,组蛋白,DNA结合蛋白等)。通过本文所述的方法分析的核酸有时基本上分离的,并且基本上不与蛋白质或其他分子缔合。核酸还包括由单链(“正义股”或“反义股”,“正”链或“负”链,“正向”读序或“反向”读序)和双链多核苷酸合成、复制或扩增的RNA或DNA的衍生物、变体和类似物。脱氧核糖核苷酸包括脱氧腺苷、脱氧胞苷、脱氧鸟苷和脱氧胸苷。对于RNA,碱基胞嘧啶被尿嘧啶替代,并且糖2'位置包括羟基部分。可以使用从受试者获得的核酸作为模板来制备核酸。

[0102] 如本文所公开术语“游离核酸”是指可以在细胞外、体液(如血液、汗液、尿液或唾液)中发现的核酸分子。游离细胞核酸可交换地用作循环核酸。游离细胞核酸的例子包括但不限于RNA、线粒体DNA或基因组DNA。

[0103] 如本文所公开的使用的术语“测序”、“序列测定”等通常指可用于确定诸如核酸或蛋白质等生物大分子的顺序的任何和所有生化过程。例如,测序数据可以包括核酸分子(例如DNA片段或RNA片段)中的全部或部分核苷酸碱基。

[0104] 如本文所公开术语“测序数据”是指确定序列信息的任何数据。测序数据可以通过多种技术获得,包括但不限于高通量测序系统,如罗氏454平台、应用生物系统SOLID平台、Helicos True单分子DNA测序技术、Affymetrix Inc.的单分子测序平台、Pacific Biosciences的实时(SMRT)技术、454Life Sciences、Illumina/Solexa和Helicos Biosciences的合成平台测序、以及应用生物系统公司的连接测序平台。ION TORRENT科技提供的生命科学技术和纳米孔测序技术也可用于高通量测序方法。例如,可以在本文描述的方法中使用的核酸测序技术是通过合成测序和基于可逆终止子的测序(例如Illumina的基因组分析仪;基因组分析仪II;HISEQ 2000;HISEQ 2500(Illumina,加州,圣地亚哥))。有了这项技术,数以百万计的核酸(如DNA)片段可以并行测序。在这种类型的测序技术的一个实例中,使用了流通池,所述流通池包含光学透明的载玻片,所述载玻片具有在其表面上结

合有寡核苷酸锚(例如衔接子引子)的8个独立通道。流通池通常是固体支持物,可以配置为保留及/或允许试剂溶液在结合的分析物上有序通过。流动池通常为平面形状,光学透明,通常为毫米或亚毫米级,并且通常具有发生分析物/试剂相互作用的通道或泳道。在一些实施例中,核酸样本可包括促进检测的信号或标签。测序数据包括通过各种技术的信号或标签的量化信息。例如流式细胞仪、定量聚合酶链反应(qPCR)、凝胶电泳、基因芯片分析、微数组、质谱、细胞荧光分析、荧光显微镜、共聚焦激光扫描显微镜、激光扫描细胞仪、亲和层析、手动批次模式分离、电场悬浮、定序及其组合。

[0105] 如本文所公开术语“序列读数”或“读数”是指由本文所述或本领域已知的任何测序过程产生的核苷酸序列。读数可以从核酸片段的一端生成(“单端读数”),有时也可以从核酸的两端生成(例如,成对末端读数、双端读数)。序列读数的长度通常与特定的测序技术有关。例如,高通量方法可提供大小从数十个碱基对(bp)到数百个碱基对不等的序列读数。在一些实施例中,序列读数的平均、中位数值或平均长度约为15个碱基对至900个碱基对(例如约20个碱基对、约25个碱基对、约30个碱基对、约35个碱基对、约40个碱基对、约45个碱基对、约50个碱基对、约55个碱基对、约60个碱基对、约65个碱基对、约70个碱基对、约75个碱基对、约80个碱基对、约85个碱基对、约90个碱基对、约95个碱基对、约100个碱基对、约110个碱基对、约120个碱基对、约130个碱基对、约140个碱基对、约150个碱基对、约200个碱基对、约250个碱基对、约300个碱基对、约350个碱基对、约400个碱基对、约450个碱基对、或约500个碱基对。在一些实施例中,序列读数的平均、中位数值或平均长度为约1000个碱基对或更长。例如,纳米孔测序可以提供序列读取,其大小可以从数十个,几百个到数千个碱基对变化。Illumina平行测序可以提供变化不大的序列读数,例如,大多数序列读数可以小于200个碱基对。

[0106] 如本文所公开术语“片段尺寸”是指生物样本中核酸分子的大小。在源自细胞材料的核酸样本中,核酸分子具有更大的尺寸。有时,需要应用诸如超声处理的方法将核酸分子分解成更小的片段。在游离的生物样本中,核酸分子的尺寸往往更小。

[0107] 如本文所公开术语“参考基因组”是指任何生物体或病毒的任何特定已知、测序或特征化的基因组,无论是部分还是完整的,这些基因组可用于参考受试者的已识别序列。例如,在国家生物技术信息中心(NCBI)或加州大学圣克鲁兹分校(UCSC)主持的在线基因组浏览器中,可以找到用于人类受试者以及许多其他生物体的参考基因组。术语“参考基因组”是指以核酸序列表示的有机体或病毒的完整遗传信息。如本文所用,参考序列或参考基因组通常是来自单个或多个个体的组装或部分组装的基因组序列。在一些实施例中,参考基因组是来自一个或多个人类个体的组装或部分组装的基因组序列。参考基因组可以被视为一个物种的一组基因的代表例子。在一些实施例中,参考基因组包含分配给染色体的序列。示例性人类参考基因组包括但不限于NCBI建构体34(UCSC等效物:hg16)、NCBI建构体35(UCSC等效物:hg17)、NCBI建构体36.1(UCSC等效物:hg18)、GRCh37(UCSC等效物:hg19)和GRCh38(UCSC等效物:hg38)。

[0108] 如本文所公开,术语“参考基因组的区域”,“基因组区域”或“染色体区域”是指参考基因组的连续或不连续的任何部分。例如,它也可以被称为宏基因组、分区、基因组读取数、参考基因组的一部分、染色体的一部分等。在一些实施例中,基因组部分基于基因组序列的特定长度。在一些实施例中,一种方法可以包括对多个基因组区域的多个映射序列读

数的分析。基因组区域的长度可以大致相同,也可以是不同的长度。在一些实施例中,基因组区域的长度大约相等。在一些实施例中,调整或加权不同长度的基因组区域。在一些实施例中,基因组区域为约10千碱基对(kb)至约500千碱基对、约20千碱基对至约400千碱基对、约30千碱基对至约300千碱基对、约40千碱基对至约200千碱基对、有时约50千碱基对至约100千碱基对。在一些实施例中,基因组区域约为100千碱基对至约200千碱基对。基因组区域并不局限于连续的序列。因此,基因组区域可以由连续及/或非连续序列组成。基因组区域不限于单一条染色体。在一些实施例中,基因组区域包括一条染色体的全部或部分或两个或多个染色体的全部或部分。在一些实施例中,基因组区域可以跨越一个、两个或更多个整条的染色体。另外,基因组区域可以跨越多条染色体的连接或不连接部分。

[0109] 如本文所公开术语“与参考基因组的区域的对齐”是指基于序列之间的完全或部分的一致性,将来自一个或多个序列读数的序列与参考基因组的序列进行比对的过程。可以手动进行对齐,也可以通过计算器算法进行对齐,示例包括作为Illumina基因组学分析管道的一部分衍生的核苷酸数据的有效局部对齐(ELAND)计算器程序。序列读数的对齐可以是100%的序列匹配。在一些实施例中,对齐是小于100%的序列匹配(即非完全匹配、部分匹配、部分对齐)。在一些实施例中,对齐约为99%、98%、97%、96%、95%、94%、93%、92%、91%、90%、89%、88%、87%、86%、85%、84%、83%、82%、81%、80%、79%、78%、77%、76%或75%的匹配。在一些实施例中,对齐包括不匹配。在一些实施例中,对齐包括1、2、3、4或5个错配。两个或多个序列可以使用其中一个链对齐。在一些实施例中,核酸序列与另一核酸序列的反向互补序列对齐。

[0110] 如本文所公开的术语“复制次数变异”是指与合格样本中存在的核酸序列的复制次数相比,测试样本中存在的核酸序列的复制次数的变化(例如与特定医疗状况有关的已知状态的对照样本)。在某些实施例中,复制次数变异发生在1千碱基或更小的核酸序列中。在某些实施例中,复制次数变异发生在1千碱基或更大的核酸序列中。在某些情况下,核酸序列是整个染色体或其重要部分。术语“复制次数变体”是指通过将测试样本中的相关序列与预期水平的相关序列进行比较,发现复制次数差异的核酸序列。例如,将测试样本中感兴趣的序列水平与合格样本中的序列水平进行比较。复制次数变体/变异包括缺失(包括微缺失)、插入(包括微插入)、重复、复制、倒位、易位和复杂的多位点变异。CNVs包括染色体非整倍体和部分非整倍体。

[0111] 关于相关技术和术语的其他细节可以在例如美国专利申请公开号US 2013/0325360、美国专利申请公开号US 2013/0034546、美国专利申请公告号US 5,235,038、美国专利申请公告号US 8,706,422、美国专利申请公开号US 2010/0112590中找到。其全部内容通过引用并入本文。

[0112] 本文所用术语“个体”是指人类个体。术语“健康个体”一词是指假定没有癌症或疾病的个人。术语“癌症受试者”一词是指已知患有或可能患有癌症或疾病的个人。

[0113] 本文所用术语“序列读数”是指从个体获得的样本中读取的核苷酸序列。序列读数可以通过本领域已知的各种方法获得。

[0114] 本文所用术语“游离核酸”、“游离DNA”或“cfDNA”是指在个体体内(例如血流)中循环并且源自一种或多种健康细胞及/或源自一种或多种癌细胞酸的核酸片段。

[0115] 本文所用术语“基因组核酸”、“基因组DNA”或“gDNA”是指核酸,包括来自一个或多

个健康(例如非肿瘤)细胞的染色体DNA。在各种实施例中,可从衍生自血细胞谱系的细胞(例如白血球)中提取gDNA。

[0116] 本文所用术语“复制次数畸变”或“CNAs”是指体细胞中复制次数的变异。例如,CNAs可以指实体瘤中复制次数的变异。

[0117] 本文所用术语“复制次数变异”或“CNVs”是指生殖系细胞或非肿瘤细胞体细胞复制次数变化引起的复制次数变化。例如,CNVs指的是由于克隆造血而引起的白血球复制次数变化。

[0118] 本文所用术语“复制次数事件”是指复制次数畸变和复制次数变异中的一种或两种。

[0119] 示例性系统实施例

[0120] 图1A描绘了用于处理高维数据的示例性系统。示例性系统100包括经由网络40彼此连接的一数据收集组件10、一数据库20以及装置数据智能组件30。可替代地,或者另外的,一个或多个组件可以在不依赖于网络连接的情况下本地连接到另一个组件。例如通过有线连接。游离核酸的测序数据被用来说明这些概念。然而,本领域技术人员将理解,本发明方法也可应用于其它材料的测序数据或非测序数据。

[0121] 如本文所公开的,数据收集组件10可以包括生成大数据或高维数据的设备或机器。在一些实施例中,数据收集组件10可以包括测序机或使用测序机来生成生物样本的核酸序列数据的设备。可使用任何适用的生物样本。在一些实施例中,生物样本基本上是细胞;例如,一种或多种类型的组织。在一些实施例中,生物样本是包括游离核酸片段的样本。生物样本的实例包括但不限于血液样本、血清样本、血浆样本、尿液样本、唾液样本等。

[0122] 测序数据的示例可以包括但不限于靶向基因组位置的序列读数数据、在游离或基于细胞的样本中由核酸片段表示的基因组的部分或全基因组测序数据、包括一种或多种表观遗传修饰(例如甲基化)的部分或全基因组测序数据,或其组合。

[0123] 由数据收集组件10获取的数据可以通过网络40或通过数据传输电缆本地传输到数据库20。在一些实施例中,可以由数据智能组件30经由本地或网络连接来分析收集的数据。图1B描绘了可以被实现以执行数据智能组件30的任务的示例性功能模块。

[0124] 在一个方面,本文公开了一种用于通过执行许多任务来处理和分析高维数据的系统,包括例如原始序列读数数据的初始处理(例如通过数据标准化,GC含量补偿等)丢弃高变异性的数据,确定一个模型来表示健康受试者的序列读数和癌症受试者的序列读数之间的差异,表示每个受试者的数据,并与远程设备通信(例如,另一台计算机或服务器)。

[0125] 图1B描绘了用于处理和分析高维数据的示例性计算机系统110。示例性实施例110通过在一个或多个计算机设备上实现用户输入和输出(I/O)模块120、存储器或数据库130、数据处理模块140、数据分析模块150、分类模块160、网络通信模块170来实现所述功能,以及执行特定任务所需的任何其他功能模块(例如除错或补偿模块、数据压缩模块等)。如本文所公开的,用户I/O模块120还可以包括输入子模块(例如键盘)和输出子模块(例如打印机、监视器或触摸板)。在一些实施例中,所有功能由一个计算机系统执行。在一些实施例中,功能由多台计算机执行。

[0126] 本文还公开了通过实现一个或多个功能模块来执行特定任务。具体地说,每个列举模块本身可以依次包括多个子模块。例如,数据处理模块140可以包括用于数据质量评估

的子模块(例如,用于排除非常短的序列读数或包括明显错误的序列读数)、用于标准化与参考基因组的区域对齐的序列读数的数目的子模块、用于补偿/校正GC偏差的子模块,等等。

[0127] 在一些实施例中,用户可以使用I/O模块120来操作本地设备上可用的数据,或者可以通过网络连接从远程服务设备或另一用户设备获得的数据。例如,I/O模块120可以允许用户通过图形用户界面(GUI)对键盘或触控板执行数据分析。在一些实施例中,用户可以通过语音控制操作数据。在一些实施例中,在授权用户访问所请求的数据之前需要用户认证。

[0128] 在一些实施例中,用户I/O模块120可用于管理各种功能模块。例如,当现有数据处理会话正在进行时,用户可以经由用户I/O模块120请求来要求输入数据。用户可以通过选择菜单选项或离散地键入命令来完成此操作,而不会中断现有进程。

[0129] 如本文所公开的,用户可以使用任何类型的输入来指导和控制经由I/O模块120的数据处理和分析。

[0130] 在一些实施例中,系统110还包括存储器或数据库130。在一些实施例中,数据库130包括可经由用户I/O模块120访问的本地数据库。在一些实施例中,数据库130包括可由用户I/O模块120经由网络连接访问的远程数据库。在一些实施例中,数据库130是存储从另一设备(例如用户设备或服务器)检索的数据的本地数据库。在一些实施例中,存储器或数据库130可以存储从因特网搜索实时检索的数据。

[0131] 在一些实施例中,数据库130可以向一个或多个其他功能模块发送数据并从中接收数据,包括但不限于数据收集模块(未示出)、数据处理模块140、数据分析模块150、分类模块160、网络通信模块170等。

[0132] 在一些实施例中,数据库130可以是其他功能模块的本地数据库。在一些实施例中,数据库130可以是远程本地数据库,其他功能模块可以通过有线或无线网络连接(例如,经由网络通信模块170)访问所述远程数据库。在一些实施例中,数据库130可以包括本地部分和远程部分。

[0133] 在一些实施例中,系统110包括数据处理模块140。数据处理模块140可以从I/O模块120或数据库130接收实时数据。在一些实施例中,数据处理模块140可以执行标准数据处理算法,例如噪声降低、信号增强、序列读数计数的标准化、GC偏差的校正等。

[0134] 在一些实施例中,数据处理模块140可以识别全局或局部系统错误。例如,测序数据可以与参考基因组内的区域比对。对于同一受试者,与不同基因组区域比对的序列读数的数目可能会有所不同。与相同基因组区域对齐的序列读数的数目在受试者之间可以不同。其中一些差异,特别是在健康受试者身上观察到的差异,可能是由于系统性错误而不是与一种或多种疾病有关。例如,如果与特定基因组区域相对应的测序数据显示健康受试者之间的广泛变化,则数据处理模块140可以将特定基因组区域分类为高噪声区域,并且可以从进一步分析中排除相应的数据。在一些实施例中,不使用排除,可以将权重分配给假设的高噪声区域以减少。在一些实施例中,可以由数据分析模块140执行对可能的系统错误的识别和处理,如下所示。

[0135] 在一些实施例中,系统10包括数据分析模块150。在一些实施例中,数据分析模块150包括识别和处理测序数据中的系统错误,如结合数据处理模块140所述。

[0136] 在一些实施例中,数据分析模块150可以将一个或多个机器学习算法应用于与单个受试者相关联的高维数据。通过这种方法,可以降低数据的维数,并且可以简化数据中嵌入的信息,然后将来自大量受试者的数据组合起来进行进一步的分析,如特征提取或模式识别。在一些实施例中,数据分析模块150可以同时实现用于数据降维和模式识别的一个或多个机器学习算法。

[0137] 在一些实施例中,数据分析模块150包括使用减小处理后的测序数据的维数。可以应用诸如主成分分析(PCA)之类的方法将高维数据转换为仍然可以代表原始测序数据主要特征的低维数据。例如,与受试者相关的大约20,000个序列读数计数,对应于20,000个低变异性的不同染色体区域,可以减少到1,000个参数或更少、500个参数或更少、200个参数或更少、100个参数或更少、90个参数或更少、80个参数或更少、70个参数或更少、60个参数或更少、50个参数或更少、40个参数或更少、30个参数或更少、20个参数或更少、10个参数或更少、8个参数或更少、5个参数或更少、4个参数或更少、3个参数或更少、2个参数或更少或单个参数。可以基于减少量对低变异训练数据进行转换,得到一个转换后的数据集,可以对其进行进一步的分析,从而得出说明健康受试者和患病受试者之间差异的关系。

[0138] 在一些实施例中,可以使用一个或多个监督学习算法来发现转换后的数据集中的模式或特征。如本文所揭示的,有监督学习问题可分为分类问题和回归问题。如本文所公开的,分类问题是当输出变量是诸如“红色”或“蓝色”或“疾病”和“没有疾病”之类的类别时。回归问题是输出变量是实际值(例如“元”或“权重”)时。可以采用两种方法来确定受试者是否患有特定的疾病状态。示例学习算法包括但不限于支持向量机(SVM)、线性回归、逻辑回归、朴素贝叶斯、决策树算法、线性判别分析、判别分析、最近邻分析(kNN)、基于特征点的方法、神经网络分析(多层感知器)、主成分分析(PCA)、线性判别分析(LDA)等。

[0139] 在一些实施例中,可以使用一个或多个无监督学习算法来发现转换数据集中的模式或特征。例如,无监督学习问题可以进一步分为聚类 and 关联问题。聚类问题是您想要发现数据中的固有分组,例如通过购买行为对客户进行分组。关联学习问题是您想发现描述数据大部分的规则,例如购买X的人也倾向于购买Y。示例无监督学习算法包括但不限于聚类算法,例如层次聚类、k-Means聚类、高斯混合模型、自组织映像和隐马尔可夫模型、用于异常检测的算法、基于神经网络的算法(例如自动编码器)、深层信念网,Hebbian学习、生成对抗网络、用于学习潜在变量模型的算法,例如期望最大化算法(EM)、矩量方法、盲信号分离技术(例如主成分分析(PCA),独立成分分析)、非负矩阵分解、奇异值分解等。

[0140] 在一些实施例中,可以使用半监督机器学习算法;例如,使用本文列举的或本领域已知的算法的任何组合。

[0141] 在一些实施例中,数据分析模块150基于训练数据为每个受试者导出一个或多个参数,其中可以进行或不进行数据降维。在一些实施例中,一个或多个参数用于对测试的受试者进行分类。例如,训练数据可用于计算二项式或多项式概率分数。在一些实施例中,系统10包括分类模块160,所述模块分析来自与医疗状况有关的状态未知的受试者的数据,并随后基于对象适合于特定类别的可能性对未知受试者进行分类。在一些实施例中,一个或多个参数包括基于逻辑回归分析计算的二项式概率分数。如本文所公开的,二项式概率分数可以对应于患有某种医疗状况例如癌症的受试者的可能性。例如,超过预定阈值的分数可以指示该受试者比没有癌症更可能患有癌症。在一些实施方案中,一个或多个参数可包

括与癌症的存在相关的测序数据分布模式。与癌症模式相似的受试者可能被诊断为患有癌症。在一些实施例中,序列数据分布模式可与特定类型的癌症相关地识别,从而允许用进一步的细节对未知受试者进行分类。

[0142] 如本文所公开的网络通信模块170可用于促进用户设备、一个或多个数据库以及通过有线或无线网络连接的任何其他系统或设备之间的通信。可以使用任何通信协议/设备,包括但不限于调制解调器、以太网连接、网卡(无线或有线)、红外通信设备、无线通信设备,及/或芯片组(如蓝牙™设备、802.11设备、WiFi设备、WiMax设备、行动通信设施等)、近场通信(NFC)、Zigbee通信、射频(RF)或射频识别(RFID)通信、PLC协议、基于3G/4G/5G/LTE的通信等。例如,具有用于处理/分析高维数据的用户接口平台的用户设备可以与具有相同平台的另一用户设备、没有相同平台的常规用户设备(例如,普通智能手机)、远程服务器、远程物联网本地网络的物理设备、可穿戴设备通信,可通信地连接到远程服务器等的用户设备。

[0143] 通过示例的方式提供了本文描述的功能模块。将理解的是,可以将不同的功能模块组合以创建不同的实用程序。还应该理解的是,可以创建附加的功能模块或子模块以实现某种实用程序。

[0144] 图2A描绘了示例过程,描绘了用于处理高维数据的总体方法流程200。在图2A中,突出显示了一些关键动作,包括但不限于定义可用于识别高质量数据的低变异筛选器(例如,步骤204)、基于训练样本的筛选数据建立差分或预测模型(例如,步骤206),基于测试样本的筛选数据计算分类分数,并预测测试样本具有特定医疗状况的可能性(例如,步骤208)。任选地,步骤202也是可能的,其中可以处理数据以提高质量。游离细胞核酸的测序数据被用来说明这些概念。然而,本领域技术人员应理解,本发明方法也可应用于其它材料的测序数据或非测序数据。

[0145] 在步骤202,可以执行可选的数据处理以提高数据质量。如本文所公开,数据处理可包括数据调整或校准;例如,基于从健康受试者的对照组获得的数据。例如,可以使用一种或多种方法对生物数据(例如测序数据)进行预处理以校正偏差或错误,例如标准化、校正GC偏差、校正PCR过度扩增引起的偏差等。

[0146] 在步骤204,制定一个或多个准则,以通过在数据收集期间识别和消除系统误差或其他类型的非疾病相关噪声来提高高维数据的质量。尽管高维数据可以被广泛地解释为包括非生物数据,但本发明的重点是高维生物数据,例如测序数据。测序数据的实例包括但不限于全基因组测序数据、靶向测序数据、表观遗传学分析数据等。如本文所公开,测序可包括但不限于核酸测序(例如,DNA、RNA或其杂交或混合物)、蛋白质测序、用于分析蛋白质-核酸相互作用的基于序列的表观遗传学分析(例如DNA或RNA甲基化分析、组蛋白修饰分析或其组合),或蛋白质序列修饰分析,例如乙酰化、甲基化、泛素化、磷酸化、苏甲酰化或其组合。

[0147] 在一些实施例中,经过步骤204分析的生物数据(例如测序数据)已经被预处理以使用一种或多种方法来校正偏差或错误,例如标准化、校正GC偏差、校正由PCR过度扩增引起的偏差等。

[0148] 在一些实施例中,建立一个或多个标准来排除在数据收集期间可能包含系统错误或其他类型的非疾病相关噪声的核酸测序数据。如本文所公开,序列数据可包括任何生物

样本的序列读数,包括但不限于游离细胞核酸样本。

[0149] 在一些实施例中,仅来自健康受试者的数据用于建立一个或多个标准,以避免来自与一个或多个疾病状况相关联的数据的干扰。在一些实施例中,可以针对基因组或染色体区域建立本文所公开的标准。例如,核酸序列读数可与参考基因组的区域对齐,并且序列读数的一个或多个特征可用于确定与特定基因组区域相关联的数据是否包含比有用信息更多的噪声,因此应排除在随后的分析中。示例性特征包括但不限于,例如,读数的数量、读数的可映射性等。

[0150] 在一些实施例中,基因组区域具有相同的尺寸。在一些实施例中,基因组区域可以具有不同的尺寸。在一些实施例中,基因组区域可由区域内核酸残基的数量来定义。在一些实施例中,基因组区域可通过其位置和区域内核酸残基的数量来定义。任何合适的尺寸都可以用来定义基因组区域。例如,基因组区域可以包括10千碱基对或以下、20千碱基对或以下、30千碱基对或以下、40千碱基对或以下、50千碱基对或以下、60千碱基对或以下、70千碱基对或以下、80千碱基对或以下、90千碱基对或以下、100千碱基对或以下、110千碱基对或以下、130千碱基对或以下、140千碱基对或以下、150千碱基对或以下、160千碱基对或以下、170千碱基对或以下、180千碱基对或以下、190千碱基对或更以下、200千碱基对或以下、或250千碱基对或以下。

[0151] 将染色体区域设置为较大的尺寸,可以扫描目标(例如,整个人类基因组)并快速进行分析。另一方面,将一个染色体区域设置为一个较小的尺寸,就可以更精确地确定一个参考基因组上更有可能是系统错误或噪音来源的位置。然而,这样的详细分析将更加耗时。

[0152] 在一些实施例中,可以使用较大的染色体区域来执行基因组的粗略扫描。在一些实施例中,可以在粗略扫描之后对较小的染色体区域执行更精细的扫描。

[0153] 步骤204本质上是数据选择步骤。当步骤204完成时,识别可能与系统误差相关联的区域。在一些实施例中,可以定义一个或多个准则来减少或消除与这些噪声区域相对应的数据。例如,可以创建高变异筛选器以允许丢弃数据变化高于阈值的与所有区域相对应的数据。在其他实施例中,可以创建低变异筛选器以将后续分析集中于数据变化低于阈值的数据。举例来说,人类单倍体参考基因组包括超过30亿个碱基,这些碱基可分为约30000个区域(或区段)。如果观察到每个区段的实验值,例如,与特定区域或区段对齐的序列读数总数,则每个受试者可以对应30000多次测量。在应用低或高变异筛选器之后,可以将与受试者相对应的测量数量减少很多。例如,包括但不限于约50%以下,约45%以下,约40%以下,约35%以下,约30%以下,约25%以下,20%以下,15%以下,10%以下或5%以下。在一些实施例中,对应于受试者的测量数量可以减少50%或更多,例如约55%,60%,65%或70%或更多。例如,最初具有30,000多个相应测量值的受试者在应用高变异或低变异筛选器之后,可以减少30%以上的测量值(例如约20,000)。

[0154] 在步骤206,从上一步骤建立的一个或多个标准可应用于训练群组的生物数据集(也称为“训练数据”)。如本文所公开,训练群组包括健康受试者和已知具有一种或多种医疗状况的受试者(也称为“患病受试者”)。例如,对于测序数据,将先前在步骤204中确定的一个或多个准则(例如,低变异性或高变异性筛选器)应用于训练群组的数据,以完全移除与在筛选器中定义的染色体区域相关联的数据部分。在一些实施例中,仅噪声数据仅被部分去除。在一些实施例中,未分配的可能有噪声的数据可以被分配一个加权系数以减少它

们在整个数据集中的重要性。

[0155] 一旦对训练群组的生物数据集进行了数据选择, 剩余的训练数据, 也称为“选定训练数据”或“已筛选训练数据”, “需要进一步分析, 以提取反映健康受试者和已知有一种或多种疾病的受试者之间差异的特征。如前所述, 原始训练数据包括健康受试者和患病受试者的数据。已筛选的训练数据构成原始训练数据的一部分, 因此也包括来自健康受试者和已知有医疗状况的受试者的数据。假设已筛选的训练数据中最大的差异来自于健康受试者的数据与患病受试者的数据之间的差异。本质上, 假设与健康受试者相关的数据应比来自任何患病受试者的数据更类似于另一健康受试者的数据; 反之亦然。

[0156] 与原始训练数据一样, 已筛选的训练数据也是高维的。在一些实施例中, 对已筛选的训练数据进行进一步的分析以降低数据维度, 并且基于降维定义健康和患病受试者之间的差异。对于一个给定的受试者, 大约20000个已筛选的测量值可以进一步简化为几个数据点。例如, 大约20000个已筛选的测量值可以基于一些提取的特征(例如, 许多主成分)进行转换, 以呈现多个数据点。在一些实施例中, 在降维之后, 有5个或更少的特征; 6个或更少的特征; 7个或更少的特征; 8个或更少的特征; 9个或更少的特征; 10个或更少的特征; 12个或更少的特征; 15个或更少的特征; 或20个或更少的特征。在一些实施例中, 已筛选的测量可以具有20个以上的特征。然后, 可以根据所选特征转换已筛选的测量值。例如, 具有两个20000个已筛选的测量值的样本可以被转换并减少到五个或更少的数据点。在一些实施例中, 具有两个20000个已筛选的测量值的样本可以被转换并减少为五个以上的数据点, 例如10、15、20等。

[0157] 如本文所揭示的, 对已筛选的训练数据集中所有受试者的转换数据点进行进一步分析, 以提取反映已筛选的训练数据集中各子群组之间差异的关系或模式。在一些实施例中, 进一步的分析包括二项式逻辑回归过程; 例如, 用于确定受试者患癌症与不患癌症的可能性。在一些实施例中, 进一步的分析包括多项式逻辑回归过程。例如, 除了确定受试者患癌症的可能性之外, 还用于确定癌症的类型。

[0158] 在步骤208, 为每个受试者计算分类分数。在一些实施例中, 分类分数是表示被分类为具有特定条件的受试者的可能性的概率分数; 例如, 正常与患癌症、或患肝癌与肺癌。

[0159] 图2B是一个示例实施例, 说明了处理高维数据时的信息流程示意图。示例性实施例210包括数据选择(例如单元220到单元250)、训练数据的处理和分析(例如单元260和单元270)以及测试数据的分类(例如单元280和单元290)。游离细胞核酸的测序数据被用来说明这些概念。然而, 本领域技术人员应理解, 本发明方法也可应用于其它材料的测序数据或非测序数据。

[0160] 在数据选择部分, 初始处理高维数据(例如, 单元220, 例如测序读数)以提高质量。在一些实施例中, 与参考基因组的特定区域对齐的序列读数的数量被标准化。例如, 数据220可以包括来自一组健康受试者(也称为基线受试者)的序列读数, 并且来自基线受试者的数据可用于建立标准化标准。在一些实施例中, 来自基线受试者的序列读数与已经划分为多个区域的参考基因组对齐。假设在测序过程中没有明显的偏差, 基因组中的不同区域应该被覆盖在大致相同的水平上。因此, 与特定区域对齐的序列读数的数量应该与和相同尺寸的另一区域对齐的序列读数的数量相同。

[0161] 在一个例子中, 一个基线受试者跨越不同基因组区域的序列读数的数量可以写成

“ 读数_i^j ”，其中整数*i*表示受试者，从1到*n*；整数*j*表示基因组区域，其值为1到*m*。如所公开的，参考基因组可被划分为任意数量的基因组区域，或任何尺寸的基因组区域。一个参考基因组可分为1000个区域、2000个区域、4000个区域、6000个区域、8000个区域、10000个区域、12000个区域、14000个区域、16000个区域、18000个区域、20000个区域、22000个区域、24000个区域、26000个区域、28000个区域、30000个区域、32000个区域、34000个区域、36000个区域、38 000个区域、40000个区域、42000万个区域、44000万个区域、46000万个区域、46000万个区域、50000个区域、55000个区域、60000个区域、65000个区域、70000个区域、80000个区域、90000个区域、或多达100000个区域。因此，*m*可以是与基因组区域数相对应的整数。在一些实施例中，*m*可以是大于100000的整数。

[0162] 在一些实施例中，可将受试者的序列读数规范化为受试者的所有染色体区域的平均读数的数量。当*i*保持不变时，从基因组区域1到*m*的序列读数以及相应区域的大小可用于计算受试者*i*的平均预期序列读数的数量，例如，基于以下等式：

$$[0163] \quad \overline{\text{读数}_i} = \sum_{j=1}^m (\text{读数}_i^j / \text{区域的大小}_i^j) / m \quad (1)。$$

[0164] 其中 区域的大小_i^j 表示序列读数 (读数_i^j) 对齐的特定染色体区域的大小（例如，以碱基对或千碱基对为单位）。此处， $\text{读数}_i^j / \text{区域的大小}_i^j$ 是序列读数密度值。因此，对于受试者*i*，可以使用以下方法计算与尺寸为 区域的大小_i^j 的给定染色体区域*j*对齐的预期序列读数的数量：

$$[0165] \quad \overline{\text{读数}_i} \times \text{区域的大小}_i^j \quad (2)。$$

[0166] 如本文所公开的，跨不同基因组区域的任何受试者的数据可以用作对照组以标准化基因组区域的序列读数。在这里，可以计算一个健康的对照组受试者、一组对照组受试者或一个受试者本身的平均读数，它被用作数据规范化的基础。

[0167] 在一些实施例中，可以根据来自一组受试者（例如，一组*n*个健康受试者）的总体平均计数对受试者的序列读数进行标准化。其他细节可以在图3的描述中找到。

[0168] 在一些实施例中，可以使用多种方法来标准化对应于特定区域的受试者的序列读数，利用来自受试者自身的不同区域的数据和跨不同对照受试者的数据。

[0169] 在一个方面，本文公开了基于从健康受试者（例如，基线健康受试者220和参考健康受试者230）收集的数据的模式，建立用于选择用于进一步分析的数据的模板的方法。在优选实施例中，参考健康受试者230与基线健康受试者220没有或仅具有最小重叠。游离细胞核酸的测序数据被用来说明这些概念。然而，本领域技术人员应理解，本发明方法也可应用于其它材料的测序数据或非测序数据。

[0170] 在一些实施例中，基线或参考健康受试者组中健康受试者的数量可以改变。在一些实施例中，基线和参考健康受试者组中的健康受试者的选择标准相同。在一些实施例中，基线和参考健康受试者组中的健康受试者的选择标准不同。

[0171] 在一些实施例中,使用来自健康参考受试者(例如单元230)的数据建立高变异或低变异筛选器。如本文所揭示的,来自健康参考受试者230的数据可以预处理(例如,经历各种标准化步骤);例如,基于来自健康受试者的基线对照组数据(例如,单元220)。例如,可以对来自健康和癌症受试者的训练数据进行预处理。在一些实施例中,原始序列读数的数据可直接用于设置高变异或低变异筛选器。

[0172] 在一些实施例中,可以将每个健康受试者的序列读数(例如,来自健康受试者数据230)与参考基因组的多个染色体区域比对。可以评估到达基因组区域的变异性;例如,通过比较对照组中所有健康受试者的特定基因组区域的序列读数的数量。作为说明,可以将预期不患有癌症的健康受试者作为参考对照组。健康受试者包括但不限于没有癌症家族史或健康年轻的受试者(例如35岁或30岁以下)。在一些实施例中,参考对照组中的健康受试者可满足其他条件;例如,只有健康妇女将被包括在用于乳腺癌分析的对照组中。只有男性被纳入前列腺癌分析的对照组。在一些实施例中,对于主要或仅在特定族群中发现的疾病,仅使用来自相同族群的人来建立参考对照群组。

[0173] 例如,对于一组健康对照受试者(n),如果我们计算与基因组区域一致的序列读数的数量,则每个基因组区域将有n个值。参数,例如平均或中等计数、标准偏差(SD)、中值绝对偏差(MAD)或四分位间距(IQR),可基于n个计数值计算,并用于确定基因组区域是否具有低或高变异性。任何计算这些参数的方法都可以使用。

[0174] 例如,对象1到n中区域j的序列读数的数量可以表示为 读数_i^j ,其中,j是一个整数,i

是一个介于1和n之间的整数。区域j的平均读数计数 $\overline{\text{读数}}^j$ 可以用 $\overline{\text{读数}}^j = \left(\sum_{i=1}^{i=n} \text{读数}_i^j / n \right)$ 表示。

在一些实施例中,可以计算IQR并将其与 $\overline{\text{读数}}^j$ 进行比较。若IQR和 $\overline{\text{读数}}^j$ 之间的差异大于一预定阈值,来自区域j的数据可能被认为具有高变异性,并且在后续分析之前将被丢弃。通过对参考基因组中的所有区域重复所述过程,可以建立全基因组范围的高变异性或低变异性筛选器(例如组件250)。例如,对于与受试者(优选不在参考对照组中)相关的任何测序数据,将丢弃与对应于高变异性筛选器的区域对齐的序列读数。低变异性筛选器将包括IQR和 $\overline{\text{读数}}^j$ 之间的差异小于一预定阈值的区域。

[0175] 在一些实施例中,可以仅对基因组的一部分创建高或低变异性筛选器。例如,仅针对特定染色体或其一部分。

[0176] 在一些实施例中,训练数据240包括来自健康受试者和已知具有医疗状况的受试者(也称为患病受试者)的生物数据(例如测序数据)。在一些实施例中,将从训练数据240中排除先前包括在基线对照组或参考对照组中的与数据相关联的健康受试者,以可能避免某些偏差。

[0177] 在一些实施例中,可以将使用健康受试者数据220和低变异性或高变异性筛选器250获得的标准化参数应用于训练数据240,以呈现新的和已筛选的训练数据260以供后续分析。

[0178] 在一些实施例中,已筛选的训练数据260包括健康和患病受试者的平衡数据;例如,健康和患病受试者的数量彼此之间在约5%到10%的范围内。在一些实施例中,已筛选

的训练数据260包括用于健康和患病受试者的不平衡数据；例如，健康和患病受试者的数量彼此相差超过10%。在后一种情况下，可以采用一些方法来减少不平衡数据的影响。

[0179] 在一些实施例中，对已筛选的训练数据260进行进一步分析以创建预测模型270。预测模型270用于预测受试者是否具有一定的医疗状况。在一些实施例中，预测模型270反映健康和患病受试者之间的差异。在一些实施例中，预测模型270中使用的差异可以通过例如对已筛选的训练数据260应用逻辑回归来获得。在一些实施例中，已筛选的训练数据260（例如，与参考基因组的某些区域对齐的序列读数的数量）可直接用于逻辑回归分析。在一些实施例中，已筛选后的训练数据260经历维数缩减以减小数据集并可能将其转换为更小的规模。例如，主成分分析（PCA）可用于将数据集的大小减少约100,000倍以下，约90,000倍以下，约80,000倍以下，约70,000倍以下，约60,000倍以下，约50,000倍以下，约40,000倍以下，约30,000倍以下，约20,000倍以下，约10,000倍以下，约9,000倍以下，约8,000倍以下，约7,000倍以下，约6,000倍以下，约5,000倍以下，约4,000倍以下，约3,000倍以下，约2,000倍以下，约1,000倍或更少，或约500倍或更少。在一些实施例中，数据集的大小可以减少超过100,000倍。在一些实施例中，数据集的大小可以减小几百倍或更少。如本文所公开的，尽管减小了数据集的大小，但是可以保留样本的数量。例如，在PCA之后，具有1,000个样本的数据集仍可以保留1,000个样本，但是每个样本的复杂度降低了（例如，从对应于25,000个特征减少到5个或更少特征）。这样，本文公开的方法可以提高数据处理的效率和准确性，同时大大减少所需的计算机存储空间。

[0180] 一旦建立了预测模型，就可以将其应用于测试数据280。测试数据280可取自关于医疗状况的状态未知的受试者。在一些实施例中，来自已知状态的测试受试者的数据也可用于验证目的。虽然图2B中没有描述，但是将处理测试数据；例如，使用单元220到单元250中描述的方案。在一些实施例中，将对测试数据280进行预处理，例如进行标准化、GC含量校正等。在一些实施例中，对测试数据280应用高变异性或低变异性筛选器250以移除可能对应于系统错误的染色体区域中的数据。在一些实施例中，预处理和高变异性或低变异性筛选器都可以应用于测试数据280，以呈现已筛选的测试数据以供进一步处理。

[0181] 在一些实施例中，当将预测模型260应用于已筛选的测试数据时，可以将分类分数计算为概率分数，以表示特定医疗状况在被分析的测试受试者中存在的可能性。在一些实施例中，概率分数可以是二项式分类分数，例如，非癌症与癌症。在一些实施例中，概率分数可以是多项式分类分数，例如非癌症、肝癌、肺癌、乳腺癌、前列腺癌等。

[0182] 本文所公开的方法和系统可用于提供与任何种系或体细胞突变相关的任何适当医疗状况的诊断或预后。具体而言，医疗条件包括但不限于美国国家癌症研究院定义的任何癌症或肿瘤，包括但不限于急性淋巴细胞白血病（ALL）、急性髓细胞白血病（AML）、青少年癌症、肾上腺皮质癌、儿童肾上腺皮质癌、艾滋病相关癌症、卡波西肉瘤、艾滋病相关淋巴瘤（淋巴瘤）、肛管癌、阑尾癌-参见胃肠道类癌、星形细胞瘤、儿童（脑癌）、非典型畸胎瘤/横纹肌瘤、儿童、中枢神经系统（脑癌）、皮肤基底细胞癌、胆管癌、膀胱癌、儿童膀胱癌、骨癌（例如，尤文氏肉瘤和骨肉瘤以及恶性纤维组织细胞瘤）、脑瘤、乳腺癌、儿童乳腺癌、儿童支气管肿瘤、伯基特淋巴瘤、类癌肿瘤（胃肠道）、儿童类癌、未知原发性癌、儿童原发性癌、儿童心脏（心脏）肿瘤、中枢神经系统（如脑癌，如儿童非典型畸胎/横纹肌样肿瘤、儿童胚胎性肿瘤、儿童生殖细胞肿瘤，宫颈癌、儿童宫颈癌等）、胆管癌、儿童脊索瘤、慢性淋巴细胞白血病

(CLL)、慢性粒细胞白血病(CML)、慢性骨髓增生性肿瘤、结直肠癌、儿童结直肠癌、儿童期颅咽管瘤、皮肤T细胞淋巴瘤(例如真菌肉芽肿和塞氏病综合征)、原位导管癌(DCIS)、儿童胚胎期肿瘤、子宫内膜癌(子宫癌)、儿童室管膜瘤、食管癌、儿童食管癌、嗅神经母细胞瘤(头颈部癌)、儿童颅外生殖细胞瘤、性腺外生殖细胞瘤、眼癌,包括儿童眼内黑色素瘤、眼内黑色素瘤、视网膜母细胞瘤等;输卵管癌、胆囊癌、胃癌、儿童胃癌、胃肠道类癌、胃肠道间质瘤(GIST)、儿童胃肠道间质瘤、生殖细胞肿瘤(如儿童中枢神经系统生殖细胞肿瘤、儿童颅外生殖细胞肿瘤、性腺外生殖细胞肿瘤、卵巢生殖细胞肿瘤或睾丸癌)、妊娠滋养细胞疾病、毛细胞白血病、头颈部癌、儿童心脏肿瘤、肝细胞癌(HCC)、郎格罕细胞组织细胞增生症、霍奇金淋巴瘤、眼内黑色素瘤、儿童眼内黑色素瘤、胰岛细胞肿瘤(胰腺神经内分泌肿瘤)、肾或肾细胞癌(RCC)、朗格汉斯细胞组织细胞增生症、喉癌、白血病、肝癌、肺癌(非小细胞和小细胞)、儿童肺癌、淋巴瘤、男性乳腺癌、恶性骨纤维组织细胞瘤和骨肉瘤、黑色素瘤、儿童黑色素瘤、眼内黑色素瘤、梅克尔细胞癌、恶性间皮瘤、儿童间皮瘤、转移性癌、具有隐匿性原发性的转移性鳞状上皮癌、NUT基因改变的中线癌、口腔癌(头颈部)、多发性内分泌肿瘤、多发性骨髓瘤/浆细胞瘤、蕈样肉芽肿(淋巴瘤)、骨髓增生异常综合征、骨髓增生异常/骨髓增生性肿瘤、慢性骨髓增生性肿瘤、鼻腔和鼻窦癌、鼻咽癌、神经母细胞瘤、非霍奇金淋巴瘤、非小细胞肺癌、口腔癌、唇癌、口腔癌、口咽癌、骨肉瘤、骨恶性纤维组织细胞瘤、卵巢癌、儿童卵巢癌、胰腺癌、儿童胰腺癌、乳头状瘤病(儿童喉癌)、副神经节瘤、儿童副神经节瘤、副鼻窦和鼻腔癌、甲状旁腺癌、阴茎癌、咽癌、嗜铬细胞瘤、儿童嗜铬细胞瘤、垂体瘤、浆细胞肿瘤/多发性骨髓瘤、胸膜肺母细胞瘤、妊娠和乳腺癌、原发性中枢神经系统(CNS)淋巴瘤、原发性腹膜癌、前列腺癌、直肠癌、复发癌、视网膜母细胞瘤、儿童横纹肌肉瘤、涎腺癌、肉瘤(如儿童血管肿瘤、骨肉瘤和子宫肉瘤)、塞扎里综合征(淋巴瘤)、皮肤癌、儿童皮肤癌、小细胞肺癌、小肠癌、皮肤鳞状细胞癌、鳞状颈癌伴隐匿性原发性转移(头颈部)、胃癌(胃癌)、儿童胃癌(胃癌)、皮肤t细胞淋巴瘤、睾丸癌、儿童睾丸癌、喉癌(如鼻咽癌、口咽癌、下咽癌)、胸腺瘤和胸腺癌、甲状腺癌、肾盂和输尿管的移行细胞癌、不明原发性癌、儿童罕见癌、输尿管肾盂癌、移行细胞癌(肾(肾细胞)癌)、尿道癌、子宫内膜癌、子宫肉瘤、阴道癌、儿童阴道癌、血管瘤、外阴癌、肾母细胞瘤等儿童肾肿瘤,青壮年癌症等等。

[0183] 在一些实施例中,本文公开的方法可用于检测与非癌性疾病有关的遗传变异,所述非癌性疾病包括但不限于例如阿尔茨海默氏症、雄激素不敏感、扁桃体发育不良、慢性婴儿神经皮肤关节(CINCA)、锁骨发育不良、CHARGE综合征、先天性中枢性通气不足、先天性水痘症候群、杜氏肌营养不良、EEC(直肠、外胚层发育不良和口面部裂痕)、单纯性大疱性表皮松解症、肩胛肌营养不良、血友病A、血友病B、遗传性痉挛性截瘫、亨特综合征、低钙血症、小儿脊髓性肌萎缩症、X染色体性联隐性遗传疾病、L-D症候群、马凡氏症、非肌性肌球蛋白重链9基因(MYH9)基因突变、肌阵挛癫痫、新生儿糖尿病、鸟氨酸转氨酶缺乏、成骨不全、耳腭指综合征、苯丙酮尿症、早熟症、色素性视网膜炎RPGR或RP2、视网膜母细胞瘤、Rett综合征、RSTS多重先天异常症候群、眼睑发育不良、林道症候群、先天性X连锁角化不良、X连锁性低磷血症、X连锁性智力低下(ARX或SLC6A8)、染色体异常(三倍体)、单体7异常、环7异常、环8异常、染色体异常(四倍体)、环17异常、等着丝粒Y异常、XXY异常、神经纤维瘤病、McCune-Albright综合征、色素失禁、阵发性夜间血红蛋白尿、变形杆菌综合征、变形杆菌(Klippel-Trenaunay和Maffucci),杜兴肌营养不良症等。附加信息可以在例如埃里克森R.,2003年,

突变研究期刊,543(2):第87-180页;Erickson R,2010,突变研究期刊;705(2):96-106;其中每一个都通过引用并入本文中。

[0184] 在一目的中,本文公开了一种用于从高维数据中选择用于后续分析的数据的有效方法。图3A描述了用于数据选择的示例流程。本质上,流程300示出了促进单元210、220、230和240之间的数据转换和处理的示例性步骤。特别地,来自两组健康受试者的数据(例如,步骤302)被用于导出一般的处理标准(例如,步骤304)和一组系统数据筛选器标准(例如,步骤306),然后再将这两组标准应用到训练数据上,以创建经过处理和筛选的训练数据以进行后续分析(例如,步骤308)。

[0185] 在步骤302,从两组健康受试者:基线健康受试者和参考健康受试者获得生物学数据,例如测序数据。在一些实施例中,基线健康受试者用于调整或校准数据。例如,在对数据进行测序的情况下,使用基线健康受试者从宏观上改善总体数据质量。在一些实施例中,参考健康受试者用于定义系统性错误并允许系统性排除可能与此类错误相对应的数据。

[0186] 在步骤304,预处理来自参考健康受试者的数据,以提高数据质量。例如,可以使用通过分析基线健康受试者的数据而建立的参数,对数据进行标准化或校正,以消除GC偏差。在一些实施例中,可以基于使用来自所有参考健康受试者的数据设置的参数来预处理来自每个参考健康受试者的数据。

[0187] 在步骤306,然后对来自参考健康受试者的处理数据进行进一步分析,以定义高变异或低变异筛选器。如本文所揭示的,处理后的数据将被细分为非重叠组。例如,可以基于序列数据如何与参考基因组中的区域对齐来将测序数据分为几类。

[0188] 可变性筛选器定义了参考基因组中可能存在的“噪声”基因组区域。即倾向于与系统错误相关联的区域。如上所述,使用参考健康受试者来识别系统性错误。例如,高变异筛选器指定任何具有高于设定阈值的误差的区域将被排除在进一步分析之外。另一方面,低变异筛选器指定将选择任何误差低于设定阈值的区域进行进一步分析。

[0189] 在步骤308,来自步骤306的高变异或低变异筛选器将应用于与训练组相关联的数据。如本文所公开,训练数据包括来自健康受试者和已知具有医疗状况的受试者(也称为“患病受试者”)的数据。训练数据将分为非重叠组;例如,与参照健康受试者的数据中定义的定义相同。例如,对于序列数据,与那些与“噪声”基因组区域相关联的数据(例如,在高变异性筛选器中指定的)将被丢弃。筛选后的数据将进行进一步分析。

[0190] 图3B和图3C描绘了用于数据选择的示例流程。再次,来自游离细胞核酸样品的测序数据用于说明样品处理流程310,并且与样品处理流程300(图3A)中概述的一般步骤相比提供了更多细节。然而,本领域技术人员将理解,当前的方法可以应用于其他材料的测序数据或非测序数据。

[0191] 在步骤312,接收来自基线健康受试者和参考健康受试者的游离细胞核酸样品的序列读数。如本文所公开的,健康受试者是尚未被诊断患有正在分析的医疗状况的受试者。如本文所公开的,健康受试者是没有直系亲属的受试者,这些直系亲属已经被诊断患有正在分析的医疗状况。如本文所公开的,健康受试者是尚未被诊断患有正在分析的医疗状况并且处于预定年龄限制内的受试者。例如,年龄在40岁以下,35岁以下或30岁以下。在一些实施例中,健康受试者高于预先确定的年龄限制,例如约15岁或以上、约18岁或以上、或约21岁或以上。在优选实施例中,基线健康组和参考健康组中的受试者不重叠。在一些实施例

中,可使用包括在两个组中的一个或多个健康受试者。在一些实施例中,定义健康受试者的标准对于两个组是相同的。在一些实施例中,定义健康受试者的标准对于两个组是不同的。

[0192] 在步骤314,参考基因组被划分为多个基因组区域。在这里,参考基因组包括代表该生物体的个体的所有序列信息。如本文所公开,参考受试者、健康受试者和训练组中的受试者均为同一生物体。在一些实施例中,多个区域具有相同的大小。在一些实施例中,多个区域可以具有不同的大小。在一些实施例中,基因组区域可由区域内核酸残基的数量来定义。在一些实施例中,参考基因组可被分割多次。例如,一个更宽或更大的基因组区域可以快速扫描或分析整个基因组。在一些实施例中,如果区域是感兴趣的,但似乎对应于高系统误差(因此应从进一步分析中丢弃),则所述区域(有时是调整区域)可重新分组并重新划分为更小的基因组区域。这样,可以更精确地描述假定的系统误差。例如,原始区域的一部分可能具有较低的变异性,可以保留下来以进行进一步分析。

[0193] 任何合适的尺寸可以用于定义基因组区域。例如,基因组区域可包括10,000个碱基或更少的碱基,20,000个碱基或更少的碱基,30,000个碱基或更少,40,000个碱基或更少,50,000个碱基或更少,60,000个碱基或更少,70,000个碱基或更少,80,000个或更少,90,000个碱基或更少,100,000个碱基或更少,110,000个碱基或更少,120,000个碱基或更少,130,000个碱基或更少,140,000个碱基或更少,150,000个碱基或更少,160,000个碱基或更少,170,000个碱基或更少,180,000个碱基或更少,190,000个碱基或更少,200,000个碱基或更少,220,000个碱基或更少,250,000个碱基或更少,270,000个碱基或更少,300,000个碱基或更少,350,000个碱基或更少,400,000个碱基或更少,500,000个碱基或更少,600,000个碱基或更少,700,000个碱基或更少,800,000个碱基或更少,900,000个碱基或更少,或1,000,000个碱基或更少。在一些实施例中,基因组区域可包括超过1,000,000个碱基。

[0194] 在步骤316,将来自基线健康受试者和参考健康受试者的测序数据(例如,序列读数)与参考基因组比对。

[0195] 在步骤318,可以通过其在参考基因组上的位置和反映与该区域比对的序列读数的数量的量来表征预先指定的基因组区域。在一些实施例中,特征可以是数量。例如,所述特征可以是与参考基因组上的特定区域相关的序列读数的计数。例如,可以将与特定区域相对应的测序数据减少为单一数量,例如与所述区域对齐的序列读数的总数或序列读数的密度值(例如,序列读数的总数除以区域的大小)。在一些实施例中,可以通过序列读数所对应的片段尺寸进一步细分计数。与表示与特定区域对齐的序列读数的总数的单个数量不同,与所述特定区域相关的测序数据可以由多个数量表示,每个数量对应于目标片段的长度或长度范围。例如,对应于150个碱基到155个碱基的目标片段的序列读数将被特征化为一个计数编号,而对应于155个碱基到160个碱基的目标片段的序列读数将被特征化为另一个计数编号。在一些实施例中,可以使用反映特定基因组区域特征的一数量。例如,可以使用指示特定区域的甲基化水平的序列读数的数量。在一些实施例中,在与特定基因组区域比对的所有序列读数中,仅将揭示一个或多个甲基化位点的那些序列读数的数量为表示为特定基因组区域。在一些实施例中,由序列读数揭示的甲基化位点的总数将用于代表特定的基因组区域。在一些实施例中,可以使用甲基化密度值(例如,甲基化位点的总数除以特定基因组区域的大小)。在一些实施方案中,可以制定一参数,所述参数可以代表与特定基

因组区域相关的一个或多个特征。

[0196] 在步骤320,可以使用来自对照健康受试者的数据来定义校准参数。在一些实施例中,可根据来自一组受试者(例如,一组n个基线健康受试者)的总体平均计数对受试者的序列读数进行标准化。例如,可以使用以下公式基于基线对照组中每个受试者的平均值 $\overline{\text{读数}}$ 来计算总体平均值:

$$[0197] \quad \overline{\text{读数}} = \frac{\sum_{i=1}^{i=n} \text{读数}_i}{n} \quad (3)$$

[0198] 此处, $\overline{\text{读数}}$ 是不同基因组区域中基线健康受试者的平均值,其中整数i指的是从1到n的受试者。例如,可以使用公式(1)来确定 $\overline{\text{读数}}$ 。

[0199] 在一些实施例中,总体平均值可用于例如使用以下等式将绑定到特定区域(x)的序列读数的数量标准化以供将来的任何受试者使用:

$$[0200] \quad \text{标准化读数} = \overline{\text{读数}} \times \text{区域大小}(x) = w_x \times \text{实际读数}(x) \quad (4)$$

[0201] 其实际读数(x)是与区域x对齐的序列读数的实际数量, w_x 是分配给所述区域的权重,用于将序列读数标准化为可以使用总体平均值获得的预期值。

[0202] 在一些实施例中,可以将针对特定区域的受试者的序列读数与一组健康受试者(例如,单元220的基线健康受试者)中相同区域的平均序列读数的平均数进行标准化。举例来说,受试者i的区域(j)的序列读数可以表示为 读数_i^j , 其中,受试者i可以是1到n的整数。区域(j)序列读数的平均数可以根据以下公式计算:

$$[0203] \quad \overline{\text{读数}}^j = \frac{\sum_{i=1}^{i=n} \text{读数}_i^j}{n} \quad (5)$$

[0204] 使用此交叉受试者平均值作为参考,任何受试者的区域(j)的序列读数可计算为:

$$[0205] \quad \text{标准化读数} = \overline{\text{读数}}^j = w_j \times \text{实际读数}(j) \quad (6)$$

[0206] 其中实际读数(j)是与区域j对齐的序列读数的实际数量, w_j 是分配给所述区域的权重,用于将序列读数标准化为可以使用平均读取 $\overline{\text{读数}}^j$ 获得的期望值。

[0207] 在一些实施例中,针对与特定区域相对应的受试者的序列读数可以通过任何可用的方法进行校准,包括使用多个方法,针对受试者本身利用来自不同区域的数据和跨不同的对照组受试者。在一些实施例中,可以在计算中使用相对序列读数的数量。例如,在后续

分析中,将使用 $\frac{\text{读数}_i^j}{\overline{\text{读数}}_i}$, 而不是序列读数观察到的值 读数_i^j 。示例性校准方法进一步包括

但不限于GC偏差校正,由于PCR过度扩增引起的偏差校正等。

[0208] 在步骤322,可以使用基于基线健康受试者而开发的参数来校准为每个参考健康受试者的每个基因组区域得出的数量;例如,在步骤320的描述中示出的那些。在一些实施例中,将对参考健康受试者的校准数据进行进一步分析。在一些实施例中,参考健康受试者

的数据可以基于基线健康受试者而无需校准而被进一步分析。

[0209] 在步骤324,可基于针对特定基因组区域的所有参考健康受试者的校准数量数据,针对参考基因组的基因组区域计算一个或多个参考量。如本文所公开,参考量可用于评估特定基因组区域的所有参考健康受试者之间的数量数据之间的变异性。例如,对于代表n个参考健康受试者的n个数量,可以将代表所有数量数据(例如,平均值、均值或中位数)的第一参考数量与至少反映所分析的所有数量数据的特征的第二参考数量进行比较(例如,标准偏差(SD),中位数绝对偏差(MAD)或四分位间距(IQR)。例如,如果参考量数据集的平均值与同一数据集的IQR大不相同,则可能指示高变异性。

[0210] 在步骤326,通过指定第一参考量和第二参考量之间的条件,可以确定对应于区域的数据是否具有高变异或低变异。例如,这可以通过建立阈值并将其与第一参考量和第二参考量之间的差值进行比较来实现。

[0211] 在步骤328,对参考基因组内的所有基因组区域重复步骤318至326的流程,以识别可能与高变异性相关的基因组区域。可以定义一个高变异性筛选器,以便在将筛选器应用于测试数据时,将识别与高变异性区域对齐的序列读数,并将其从进一步的分析中排除。另一方面,低变异性筛选器指定表现出低变异性的基因组区域。当将低变异性筛选器应用于测试数据时,将识别与低变异性区域对齐的序列读数,并从进一步的分析中进行选择。

[0212] 分析背后的理由是这样的假设,即健康受试者的生物学数据之间的正常变化将倾向于小于在生成生物学数据的过程中发生的系统误差。在一些实施例中,为了避免或减少可能的年龄相关变化,参考对照群组中的健康受试者是35岁或更年轻的健康年轻人。因此,如步骤308所示,通过定位可能与最显著变化相关联的那些区域,可以从随后的分析中消除相应的数据,以避免或减少系统误差。如前所述,筛选器可以以不同的方式建立,以允许排除高度变异的数据或包含低变异的数据。在一些实施例中,可以调整筛选器中的一个或多个阈值以改变将要进一步分析的数据量。

[0213] 在一些实施例中,将从随后的分析中去除的区域包括GC含量高于阈值的基因组区域。在一些实施例中,将从随后的分析中去除的区域包括GC含量低于阈值的基因组区域。

[0214] 在一些实施例中,整体高变异或低变异筛选器可以识别和应用,例如跨整个基因组区域。在一些实施例中,可以针对更特异性和更小的基因组区域确定更精细的筛选器。例如,如果可疑状况仅与一个或染色体或其一部分有关,则可以仅针对粒子区域设置筛选器。

[0215] 如本文所揭示,可使用与生物数据直接相关联的数量或从中导出的数量来导出筛选器。

[0216] 步骤330到步骤336示出一个示例,说明如何通过应用低变异筛选器来快速选择训练数据的一部分以进行进一步分析。

[0217] 在步骤330,提供来自一训练群组的游离细胞核酸样本的序列读数。训练群组既包括健康受试者,也包括已知患有某种疾病的受试者(也称为“患病受试者”)。

[0218] 在步骤332,将训练群组的序列读数与相同参考基因组的基因组区域比对。从每个区域,可以根据与特定区域对齐的序列读数的数量来导出数量。在一些实施例中,数量可以是与区域对齐的序列读数的总数。在一些实施例中,数量可以是序列读数密度值(例如序列读数的总数除以区域的大小)。在一些实施例中,可以计算相对计数值;例如,可以将观察到的序列读数编号标准化为相同的区域大小,然后除以对照组平均序列读数。

[0219] 在步骤334,可以基于来自基线健康受试者的数据校准步骤322中导出的数量数据(例如,校正GC偏差、标准化计数等等)。例如,可以应用步骤314到322中所示的流程。

[0220] 在步骤336,将步骤328之后定义的低变异性筛选器应用于校准数据。在一些实施例中,仅选择与低变异性区域相对应的数量数据以供进一步分析。在一些实施例中,与高变异性区域相对应的数据被分配不同的权重以反映其可能的重要性,而不是被完全丢弃。

[0221] 如本文所述,先前用于基线及/或参考健康受试者的区域名称可用于训练群组中的受试者。如本文所公开,在步骤334的校准期间,用于基线健康受试者的基因组区域指定优选应用于来自训练群组的数据。类似地,在步骤336的数据选择期间,用于定义低变异性筛选器的基因组区域指定应用于来自训练群组的数据。

[0222] 为了简单起见,可以将相同的区域名称用于来自基线健康受试者、参考健康受试者和训练群组的数据。

[0223] 在图3B和图3C所示的流程完成之后,对来自训练群组的数据进行预处理(例如标准化),低变异性被筛选,并准备进行进一步分析。如本文所公开的,这样的数据被称为已筛选的训练数据。

[0224] 一方面,本文公开了一种用于分析高维数据以建立表示数据的一个或多个特征的参数的方法。

[0225] 图4描绘了用于分析数据以减少数据维度的示例过程。样本数据分析流程400使用已筛选的训练数据(例如根据图3A至图3C所示的流程处理的训练数据)开始步骤405。例如,对于测序数据,在步骤405仅接收与假定的低变异性基因组区域相对应的数据。如本文所公开的,训练数据包括来自健康和患病受试者的数据。在应用高变异性或低变异性筛选器后,已筛选的训练数据仍应包括来自健康和患病受试者的生物学数据。在一些实施例中,患病受试者是已经被诊断出患有至少一种类型的癌症的患者。

[0226] 在步骤410,使用交叉验证方法分离已筛选的训练数据。在一些实施例中,交叉验证方法包括但不限于像leave-p-out交叉验证(LpO-CV)这样的穷举方法,其中p可以具有将创建有效分区的任何值,或者在 $p=1$ 的情况下leave-p-out交叉验证。在一些实施例中,交叉验证方法包括但不限于非穷举方法,例如holdout方法、重复随机子抽样验证方法或分层或非分层k倍交叉验证方法,其中k可以具有可以创建有效分区的任何值。如本文所公开的,交叉验证过程以预定的百分比分割,将已筛选的训练数据划分为训练子集和验证子集的不同对。例如,在步骤410描述的第一训练子集和第一验证子集表示在k倍交叉验证实验的一次倍数期间的80:20分割。在同一k倍交叉验证实验的另一个倍数中,已筛选的训练数据将以相同的百分比分成不同的训练子集和验证子集。在一些实施例中,应用多个交叉验证实验,其中一对训练和验证子集的分割比率可以在每个实验中变化。如本文所公开的,子集可以被随机创建。在一些实施例中,创建子集使得每个子集包括来自健康和患病对象的数据。在一些实施例中,子集中只有一个包括来自健康和患病受试者的数据。例如基本上是训练子集必须包括健康和患病的受试者。

[0227] 在一些实施例中,训练子集构成了大部分已筛选的训练数据;例如,高达60%、高达65%、高达70%、高达75%、高达80%、高达85%、高达90%或高达95%的已筛选的训练数据。在一些实施例中,超过95%的一组已筛选的训练数据可以用作训练子集。为了避免训练偏差,通常的优良作法是将至少5%的原始数据保存为测试子集;也就是说,由于此子集将

永远不会用作训练数据,而只会用于验证结果模型。

[0228] 在步骤415,可以使用来自第一训练子集的数据来导出获得健康和患病受试者的数据之间的一个或多个差异的关键特征。在一些实施例中,在可以从减少的数据集中导出关键特征之前,可以减少第一训练子集中每个受试者的数据(例如,序列读数的计数或从其衍生的数量)的维度。例如,已被识别为具有约10,000至约20,000的低变异性区域的样本可以具有10,000至20,000的对应计数值(或衍生量,例如相对计数值、对数计数值等)。通过使用诸如主成分分析(PCA)的方法,可以识别和选择代表第一训练子集中数据之间最大变化的主成分(PCs)。这些主成分(PCs)可用于将10000到20000个计数数据缩减到一个低维度特征空间,其中每个特征对应于所选PC中的一个。在一些实施例中,选择5个或更少的PCs。在一些实施例中,选择10个或更少的PCs。在一些实施例中,选择15个或更少的PCs。在一些实施例中,选择20个或更少的PCs。在一些实施例中,选择25个或更少的PCs。在一些实施例中,选择30个或更少的PCs。在一些实施例中,选择35个或更少的PCs。在一些实施例中,选择40个或更少的PCs。在一些实施例中,选择45个或更少的PCs。在一些实施例中,选择50个或更少的PCs。在一些实施例中,选择60个或更少的PCs。在一些实施例中,选择70个或更少的PCs。在一些实施例中,选择80个或更少的PCs。在一些实施例中,选择90个或更少的PCs。在一些实施例中,选择100个或更少的PCs。在一些实施例中,选择100个以上的PCs。

[0229] 在一些实施例中,提取的特征包括所选择的PCs。在一些实施例中,提取的特征是所选择的PCs。在一些实施例中,提取的特征是与其他特征相结合的所选择的PCs;例如,可以单独加权PCs。在一些实施例中,选择PCs以外的特征。

[0230] 在步骤420,可以获得一个或多个参数,以反映提取的每个特征对第一训练子集中的数据差异的相对贡献。例如,对于每个选定的主成分或PC,将权重分配给来自每个低变异性区域的数量数据,以反映数据各自的重要性。对观察到的差异贡献较大的区域将被赋予更大的权重;反之亦然。在某些实施例中,权重是PC特定和区域特定的,但对于所有受试者而言都是相同的,例如 w_k^j 可以代表与PC k 相关的区域 j 的权重,其中 j 是1至 m' 的整数, m' 是小于 m 的整数, m 是参考基因组中指定的基因组区域的原始数目。对于不同的受试者,此权重值相同。在一些实施例中,可以制定更多个别化的权重值以反映受试者之间的差异。例如,不同癌症类型的权重可能不同。对于同一区域和同一PC,不同族裔的人的权重可能有所不同。

[0231] 在步骤425,基于所提取的特征(例如,一些选定的PC)来变换第一训练子集中的数据。在一些实施例中,变换后的数据的维度远远小于已筛选的训练数据的维度,所述已筛选的训练数据的维度已经从原始未筛选的数据中减小了。这些概念如下所示。

$$[0232] \quad \text{受试者}_1(\text{读数}) = [\text{读数}_1^1, \text{读数}_1^2, \text{读数}_1^3, \dots, \text{读数}_1^m] \quad (7)$$

[0233] 公式(7)说明了应用低变异性筛选器之前的受试者1的数据,其中 m 是区域的总数。

$$[0234] \quad \text{受试者}_1(F\text{读数}) = [F\text{读数}_1^1, F\text{读数}_1^2, F\text{读数}_1^3, \dots, F\text{读数}_1^{m'}] \quad (8)$$

[0235] 公式(8)说明了在应用低变异性筛选器之前的受试者1的数据,其中 m' 是区域总数。应用低变异性筛选器后,基因组区域的总数减少到 m' ,可以明显小于 m 。例如,受试者的未筛选数据可以包括30,000个或更多的成分,每个成分与一个基因组区域相关联。应用低

变异性筛选器后,可以排除具有高变异性的大部分基因组区域。例如,针对同一受试者的已筛选数据可以包括20,000个或更少的成分,每个成分都与一个低变异性基因组区域相关联,如(8)所示。

[0236] 在步骤425,可以基于提取的特征的数量来进一步减小已筛选的数据的数据维度。例如,如果选择了k个主要成分,则可以将已筛选的数据的维度减小为k。如结合步骤415所描述的,所选择的PC的数量可以远小于已筛选的数据的维度。例如,当仅选择5个PC时,可以将受试者1的已筛选的读数数据(FRead)的数据维度进一步减小为5,例如下面的(9)中的表达式:

$$[0237] \quad F\text{读数}_{PC1} = \sum_{j=1}^{j=m'} (w_{PC1}^j \times F\text{读数}_1^j)$$

$$[0238] \quad F\text{读数}_{PC2} = \sum_{j=1}^{j=m'} (w_{PC2}^j \times F\text{读数}_1^j)$$

$$[0239] \quad F\text{读数}_{PC3} = \sum_{j=1}^{j=m'} (w_{PC3}^j \times F\text{读数}_1^j)$$

$$[0240] \quad F\text{读数}_{PC4} = \sum_{j=1}^{j=m'} (w_{PC4}^j \times F\text{读数}_1^j)$$

$$[0241] \quad F\text{读数}_{PC5} = \sum_{j=1}^{j=m'} (w_{PC5}^j \times F\text{读数}_1^j)$$

[0242] 因此,与大量低变异性区域相关联的数量数据(例如读数的数量)可以被减少并转换为少数数值。在一些实施例中,可以将权重分配给每个PC。在一些实施例中,可以基于与多个PC相关联的值来计算单个值。

[0243] 在步骤430,将分类方法应用于每个受试者的转换数据以提供分类分数。可以应用结合所述分析模块150和分类模块160描述的任何合适的算法。在一些实施例中,分类分数可以是二项或多项概率分数例如,在癌症的二项式分类中,逻辑回归可以用于计算概率分数,其中0表示没有癌症的可能性,而1表示患癌症的最高确定性。得分超过0.5分表明受试者患癌症的可能性比没有癌症的可能性大。逻辑回归生成公式的系数(及其标准误差和显著性水平),以预测感兴趣特征存在的概率的对数转换。使用相同的示例来说明通过逻辑回归确定概率的方法,可以将方程式(10)中的癌症患者的概率(p)写为:

$$[0244] \quad \text{逻辑概率}(p) = b_0 + b_1 \times F\text{读数}_{PC1} + b_2 \times F\text{读数}_{PC2} + b_3 \times F\text{读数}_{PC3} + b_4 \times F\text{读数}_{PC4} + b_5 \times F\text{读数}_{PC5} \quad (10)$$

[0245] 其中,从PC1导出的每个转换和简化的数据都被分配了一个权重。逻辑转换定义为方程式(11)中记录的机率:

$$[0246] \quad \text{机率} = \frac{p}{1-p} = \frac{\text{存在癌症的可能性}}{\text{不存在癌症的可能性}} \quad (11)$$

[0247] 和方程式 (12) 中的概率 p 。

$$[0248] \quad p = \frac{1}{1+e^{-\text{逻辑概率}(p)}} \quad (12)$$

[0249] p 的值可以用方程式 (12) 计算, 方法是将方程式 (10) 中的值代入。在一些实施例中, 可以在逻辑表中查找值。

[0250] 在一些实施例中, 可以采用多项式分类方法将受试者分类为不同的癌症类型。例如, 可以将现有的多项式分类技术分类为 (i) 转换为二进制 (ii) 从二进制扩展和 (iii) 层次分类。在二进制转换方法中, 可以将多类问题基于一对多或一对一方法转换为多个二元问题。二进制算法的示例性的延伸包括但不限于神经网络、决策树、 k -最接近值、朴素贝叶斯、支援向量机制和极限学习机制等。层次分类通过将输出空间划分为树来解决多项式分类问题。将每个父节点划分为多个子节点, 然后继续所述过程, 直到每个子节点仅代表一个类。已经基于分层分类提出了几种方法。在一些实施例中, 可以应用多项式逻辑回归。它用于预测给定一组自变量 (可能是实值、二进制值、范畴值等) 的范畴分布因变量的不同可能结果的概率。

[0251] 在步骤 435, 已筛选的训练数据被划分为第二训练子集和第二测试/验证子集, 并且 410 到 430 的步骤在一个或多个细化周期 (也称为“一个或多个交叉验证周期”) 中重复。如本文所揭示的, 在交叉验证程序中, 验证子集本身在不同倍数上几乎没有重叠 (例如, 在重复随机抽样中) 或根本没有重叠 ($L00CV$ 、 $Lp0$ CV 、 k -倍)。

[0252] 在细化周期中, 可以应用预定条件 (例如成本函数) 来优化分类结果。在一些实施例中, 在交叉验证过程的每一倍数期间, 使用训练数据子集完善分类函数中的一个或多个参数, 并通过验证或衍生的子集进行验证。在一些实施例中, 可以优化 PC 特定的权重及/或区域特定的权重以优化分类结果。

[0253] 在一些实施例中, 在交叉验证过程的任何倍数期间, 一小部分已筛选的训练数据可以被保留, 而不是作为训练子集的一部分, 以更好地估计过度拟合。

[0254] 在步骤 440, 使用改进的参数来计算分类分数。如本文所揭示的, 改进的参数可以作为癌症以及癌症类型的预测模型。可以使用多种生物数据构建预测模型; 包括但不限于核酸测序数据 (游离细胞与非细胞、全基因组测序数据、全基因组甲基化测序数据、 RNA 测序数据、目标平板测序数据)、蛋白质测序数据、组织病理学资料、家族史资料、流行病学资料等。

[0255] 一方面, 本文公开了一种基于使用训练数据建立的参数将受试者分类为具有某种医疗状况的方法。

[0256] 图 5 描绘了用于基于从具有减小的维度的数据中获知的信息来分析数据的示例流程。流程 500 示出了来自受试者的关于医疗状况的状态未知的测试数据如何可以用于计算分类分数并用作诊断受试者是否可能患有所述状况的基础。

[0257] 在步骤 510, 从状态未知的受试者的测试样本接收测试数据。在一些实施例中, 测试数据是与来自基线健康受试者的数据相同的类型。在一些实施方案中, 测试数据与来自

参考健康受试者的测试数据具有相同类型。样本数据类型包括但不限于用于检测目标突变的测序数据、全基因组测序数据、RNA测序数据和用于检测甲基化的全基因组测序数据。在一些实施例中,可以对测试数据进行校准和调整以提高质量(例如,标准化、GC含量校正等)。

[0258] 在步骤520,使用先前定义的低变异性筛选器执行数据选择。有利的是,基于筛选器的方法是直接的,并且可以通过改变为参考基因组中的基因组区域计算的参考量的阈值来容易地进行调整。

[0259] 在步骤530,可以基于先前基于训练数据确定的参数(例如,在步骤435获得的细化参数)来计算测试受试者的分类分数。先前确定的参数可以形成癌症和特定类型癌症的预测模型。

[0260] 在步骤540,可以基于分类分数向测试提供诊断。在一些实施例中,确定用于癌症与非癌症诊断的参数。在一些实施例中,确定用于癌症类型诊断的参数。

[0261] 如前所述,本揭示的方法可应用于任何合适的生物数据,尤其是核酸测序数据。在一些实施例中,可使用多种类型的数据来构建预测模型,包括但不限于核酸测序数据(游离细胞与非细胞、全基因组测序数据、全基因组甲基化测序数据、RNA测序数据、靶向面板测序数据)、蛋白质测序数据、组织病理学资料、家族史资料、流行病学资料等。

[0262] 一方面,本文公开了在多个层次上分析高维数据并使用这些分析的结果进行分类的方法。

[0263] 图6描述了根据本发明进行数据分析的示例流程。如图3和图4中详细描述,在数据分析的多个点期间可以进行维度的降低。

[0264] 在一些实施例中,在初始数据处理期间可以发生一定程度的数据选择:例如,在标准化、GC内容校正和其他初始数据校准步骤期间,可以拒绝明显有缺陷的序列读数,从而减少数据的数量。如图所示,在样本处理流程600中,可以通过应用低变异性或高变异性筛选器来进行数据维度缩减。例如,一个参考基因组可以分为许多区域。这些区域的大小可以相等也可以不相等(例如,图6中的单元610)。

[0265] 如本文所揭示的,低变异性筛选器指定将被选择用于进一步处理的610中的基因组区域的子集(例如,单元620中突出显示的区域)。使用突出显示的组合基因组区域,筛选器可以根据参考健康受试者对可能的系统错误进行的既定分析,对数据进行分类选择或拒绝。

[0266] 然后对所选数据进行转换,以进一步降低数据维度(例如图6中的单元630)。在一些实施例中,可以使用来自所有选择的基因组区域的数据来生成转换的数据630,但是数据的维度可以大大降低。例如,来自20000多个不同基因组区域的数据可以转换成少数几个值。在一些实施例中,可以生成单个值。

[0267] 在一些实施例中,从单元620选择的测序数据可以根据由测序数据表示的片段大小在子组中排序。例如,不是对绑定到特定区域的所有序列读数进行单一计数,而是可以导出多个分位数,每个分位数对应于一个尺寸或尺寸范围。例如,与140至150碱基片段相对应的序列读数将与对应于150至160碱基片段的序列读数分开分组,如图6中的单元640所示。因此,在数据用于分类之前,可以进行额外的细节和微调。

[0268] 如图6所示,可使用多种类型的数据进行分类(如图6中的单元650),包括但不限于

来自未经维度缩减的选择/已筛选基因组区域的数据、缩减数据、简化和测序的数据等。

[0269] 本文的方法和系统提供了优于现有已知方法的优点。例如,使用易于从原始测序数据得出的数量进行分类。本揭示不需要构建特定于染色体的分割图,因此消除了生成那些图的耗时过程。而且,本揭示方法允许更有效地利用计算机存储空间,因为它不再需要为大型分割图进行存储。

[0270] 图12A是根据一个实施例,用于识别在cfDNA样本中标识的复制次数事件的来源的示例性流程1200。具体地,图12A描绘用于检测个体中的CNA的示例性步骤。

[0271] 从测试样品中提取游离细胞DNA (cfDNA) 和基因组DNA (gDNA) 并测序(例如,使用全外显子组或全基因组测序) 以获得序列读数。分别分析cfDNA序列读数和gDNA序列读数,以鉴定每个相应样品中一个或多个复制次数事件的可能存在。

[0272] 在这里,来自cfDNA的复制次数事件的来源可以是生殖系来源、体细胞非肿瘤来源或体细胞肿瘤来源中的任何一个。源于gDNA的复制次数事件的来源可以是生殖系来源,也可以是体细胞非肿瘤来源。因此,在cfDNA中检测到但在gDNA中未检测到的复制次数事件很容易归因于体细胞肿瘤来源。

[0273] 在步骤1205中,获得源自cfDNA样品的比对序列读数(以下称为cfDNA序列读数)和源自gDNA样品的比对序列读数(以下称为gDNA序列读数)。

[0274] 在步骤1210,分析比对的cfDNA序列读数和gDNA序列读数,以分别识别cfDNA样品和gDNA样品各自在参考基因组上的统计上显著的读取数和片段。一个读取数包括一个基因组的一系列核苷酸碱基。一片段指一个或多个读取数。因此,每个序列读数被分类在包含与所述序列读数相对应的一系列核苷酸碱基的读取数及/或片段中。基因组的每个具有统计意义的读取数或片段包括在表示复制次数事件的读取数或片段中分类的序列读数的总数。通常,具有统计意义的读取数或片段包括序列读数计数,所述序列读取计数与所述读取数或片段的预期序列读数计数显著不同,即使在考虑可能的混杂因素时,混杂因素示例性包括处理偏差、读取数或片段中的变异、或样本中的总体噪声水平(例如,cfDNA样本或gDNA样本)。因此,具有统计学意义的读取数及/或具有统计意义的片段的序列读数计数可能指示生物异常,例如样本中存在复制次数事件。

[0275] 步骤1210包括用于识别统计上重要的仓的读取数级分析以及用于识别统计上重要的片段的片段级分析。在读取数和片段级别执行分析可以更准确地识别可能的复制次数事件。在一些实施例中,仅在读取数级别执行分析可能不足以获得跨越多个读取数的复制次数事件。在其他实施例中,仅在片段级别执行分析可能会产生不够细化的分析,无法获得其大小在各个读取数数量级上的复制次数事件。

[0276] 通常,cfDNA序列读数的分析和gDNA序列读数的分析彼此独立地进行。在各种实施例中,并行进行cfDNA序列读数和gDNA序列读数的分析。在一些实施例中,cfDNA序列读数和gDNA序列读数的分析在不同的时间进行,这取决于获得序列读数的时间(例如,当在步骤1205中获得序列读数时)。

[0277] 现在参考图12B,是根据实施例描述用于识别从cfDNA和gDNA样本导出的具有统计意义的读取数和具有统计意义的片段的分析的示例流程。具体而言,图12B描绘了在图12A所示的步骤1210中包括的步骤。因此,可以对cfDNA样本执行步骤1220到步骤1260,并且类似地,可以对gDNA样本单独执行步骤1220到步骤1260。

[0278] 在步骤1220,为参考基因组的每个读取数确定读取数序列读数计数。一般来说,每一个读取数代表基因组的一系列相邻的核苷酸碱基。一个基因组可以由多个读取数组成(例如数百个甚至数千个)。在一些实施例中,每个读取数中的核苷酸碱基的数量在基因组中的所有读取数中是恒定的。在一些实施例中,对于基因组中的每个读取数,每个读取数中的核苷酸碱基的数量不同。在一个实施例中,每个读取数中的核苷酸碱基的数量在25千碱基对和200千碱基对之间。在一个实施例中,每个读取数中的核苷酸碱基的数量在40千碱基对和100千碱基对之间。在一个实施例中,每个读取数中的核苷酸碱基的数量在45千碱基对和75千碱基对之间。在一个实施例中,每个读取数中的核苷酸碱基的数量为50千碱基对。实际上,也可以使用其他大小的读取数。

[0279] 回到图12B,在步骤1225,对每个读取数的读取数序列读数计数进行标准化以移除一个或多个不同的处理偏差。通常,基于先前为同一个存储单元确定的处理偏差,对一个读取数的存储单元序列读数计数进行标准化。在一个实施例中,标准化读取数序列读数计数涉及将读取数序列读数计数除以代表处理偏差的值。在一个实施例中,标准化读取数序列读数计数涉及从读取数序列读数计数减去代表处理偏差的值。读取数的处理偏差的示例可包括鸟嘌呤-胞嘧啶 (GC) 含量偏差,可映射性偏差或通过主成分分析捕获的其他形式的偏差。可以从图12C所示的处理偏差存储器1270访问用于读取数的处理偏差。

[0280] 在步骤1230,通过将所述读取数的读取数序列读数计数修改为所述读取数的预期的读取数序列读数计数,来确定每个读取数的读取数分数。步骤1230用于标准化所观察的读取数序列读数计数,以便如果特定的读取数在多个样本中一致地具有高序列读数计数(例如高预期的读取数序列读数计数),则观察的读取数序列读数计数的标准化说明了所述趋势。可以从训练特征数据库1265(参见图12C)中的读取数期望计数存储器1280访问读取数的预期序列读数计数。下面将进一步详细描述每个读取数的预期序列读数计数的生成。

[0281] 在一个实施例中,读取数的读取数分数可以表示为针对所述读取数的观察的序列读数计数与针对所述读取数的预期的序列读数计数的比率的函数。例如,读取数分数 b_i 和 b_{ini} 可以表示为:

$$[0282] \quad b_i = \log \left(\frac{\text{观察的读取数序列读数计数}}{\text{预期的读取数序列读数计数}} \right) \quad (13)$$

[0283] 在其他实施例中,读取数的读取数分数可以表示为读取数的观察的序列读数计数与读取数的预期的序列读数计数之间的比率(例如 $\frac{\text{观察的}}{\text{预期的}}$),比率的平方根(例如 $\sqrt{\frac{\text{观察的}}{\text{预期的}}}$),比率的广义对数变换(glog)(例如 $\log \left(\text{观察的} + \sqrt{\text{观察的}^2 + \text{预期的}} \right)$),或比率的其他方差稳定转换。

[0284] 返回到图12B,在步骤1235,为每个读取数确定读取数方差估计。这里,读取数方差估计表示读取数的预期方差,所述方差由表示样本中方差水平的膨胀系数进一步调整。换言之,读取数方差估计表示从先前的训练样本中确定的读取数的期望方差和当前样本(例如cfDNA或gDNA样本)的膨胀系数的组合,所述膨胀系数未计入读取数的预期方差中。

[0285] 举个例子,一个读取数 i 的读取数方差估计(var_i)可以表示为:

$$[0286] \quad \text{var}_i = \text{var}_{\text{预期}_i} * I_{\text{样本}} \quad (14)$$

[0287] 其中 $\text{var}_{\text{预期}_i}$ 表示从先前训练样本确定的读取数 i 的预期方差,而 $I_{\text{样本}}$ 表示当前样本的膨胀系数。通常,通过图12C中所示的读取数期望方差存储器1290来获得读取数的期望方差(例如 $\text{var}_{\text{预期}}$)。

[0288] 为了确定样本的膨胀系数 $I_{\text{样本}}$,确定样本的偏差并与从图12C所示的样本变异系数存储器1295检索到的样本变异系数相结合。样本变异系数是先前通过对从多个训练样本得出的数据进行拟合而得出的系数值。例如,如果执行线性拟合,则样本变异系数可以包括斜率系数和截距系数。如果执行更高阶拟合,则样本变异系数可以包括其他系数值。

[0289] 样本的偏差代表了整个样本的读取数中序列读数计数的变异性的量度。在一个实施例中,样本的偏差是中值绝对成对偏差(MAPD),并且可以通过分析相邻读取数的序列读数计数来计算。具体而言,MAPD表示整个样本中相邻读取数的读取数分数之间的绝对值差的中值。在数学上,MAPD可以表示为:

$$[0290] \quad \forall (\text{读取数}_i, \text{读取数}_{i+1}), \text{MAPD} = \text{中值}\{|(b_i) - (b_{i+1})|\} \quad (15)$$

[0291] 其中, b_i 和 b_{i+1} 分别是读取数 i 和读取数 $i+1$ 的读取数分数。

[0292] 通过组合样本变异系数和样本偏差(例如MAPD)来确定膨胀系数 $I_{\text{样本}}$ 。例如,样本的膨胀系数 $I_{\text{样本}}$ 可以表示为:

$$[0293] \quad I_{\text{样本}} = \text{斜率} * \sigma_{\text{样本}} + \text{截距} \quad (16)$$

[0294] 这里,“斜率”和“截距”系数中的每一个都是从样本变异系数存储器1295获取的样本变异系数,而 $\sigma_{\text{样本}}$ 代表样本的偏差。

[0295] 在步骤1240,分析每个读取数以基于读取数分数和读取数方差估计来确定读取数是否具有统计意义。对于每个读取数 i ,可以将所述读取数的读取数分数(b_i)和读取数方差估计(var_i)组合起来,以生成读取数的 z -分数。读取数 i 的 z -分数(z_i)示例可以表示为:

$$[0296] \quad z_i = \frac{b_i}{\text{var}_i} \quad (17)$$

[0297] 为了确定一个读取数是否为具有统计意义的读取数,将所述读取数的 z -分数与阈值进行比较。如果读取数的 z -分数大于阈值,则读取数被视为具有统计意义的读取数。相反,如果读取数的 z -分数小于阈值,则读取数不被视为具有统计意义的读取数。在一个实施例中,如果读取数的 z -分数大于2,则确定读取数具有统计意义。在其他实施例中,如果读取数的 z -分数大于2.5、3、3.5或4,则确定读取数具有统计意义。在一实施例中,如果读取数的 z -分数小于-2,则确定读取数具有统计意义。在其他实施例中,如果读取数的 z -分数小于-2.5、-3、-3.5或-4,则确定读取数具有统计意义。具有统计意义的读取数可表示样本中存在的一个或多个复制次数事件(例如,cfDNA或gDNA样本)。

[0298] 在步骤1245,生成参考基因组的多个片段。每个片段由一个或多个参考基因组组成,并有一个统计序列读数计数。统计序列读数计数的示例可以是平均读取数序列读数计数、中值读取数序列读数计数等。一般来说,参考基因组的每个生成片段具有与相邻片段的统计序列读数计数不同的统计序列读数计数。因此,第一片段可以具有与相邻第二片段的平均读取数序列读数计数显著不同的平均读取数序列读数计数。

[0299] 在各种实施例中,参考基因组片段的产生可包括两个分离的阶段。第一阶段可以包括基于每个片段中读取数的读取数序列读数计数的差异将参考基因组初始分割成多个初始片段。第二阶段可以包括重新分段过程,所述过程涉及将一个或多个初始片段重新组合成更大的片段。在这里,第二阶段考虑通过初始分段过程创建的片段的长度,以合并并在初始分段过程中发生的过度分段导致的假阳性片段。

[0300] 更具体地涉及初始分段过程,所述初始分段过程的一个例子包括执行循环二进制分段算法,以基于片段内读取数的读取数序列读数计数,将参考基因组的部分递归地分解为片段。在其他实施方案中,可以使用其他算法来执行参考基因组的初始分段。作为循环二进制分段过程的一个例子,所述算法识别参考基因组内的断点,使得由所述断点形成的第一片段包括第一片段中的读取数仓的统计读取数序列读数计数,其与由所述断点形成的第二片段中的读取数的统计读取数序列读数计数显着不同。因此,循环二进制分段过程产生许多片段,其中第一片段中的读取数的统计读取数序列读数计数与第二相邻片段中的读取数的统计读取数序列读数计数显着不同。

[0301] 初始分段过程可以在生成初始分段时进一步考虑每个读取数的读取数方差估计。例如,当计算片段中的读取数的统计读取数序列读数计数时,可以为每个读取数 i 分配一个权重,所述权重取决于所述读取数的读取数方差估计(例如 var_i)。在一个实施例中,分配给读取数的权重与所述读取数的读取数方差估计的大小成反比。具有较高的读取数方差估计值的读取数被赋予较低的权重,从而减小所述读取数的序列读数计数对所述片段中的读取数的统计读取数序列读数计数的影响。相反,具有较低的读取数方差估计值的读取数被赋予更高的权重,这增加了所述读取数的序列读数计数所述片段中的读取数的统计读取数序列读数计数的影响。

[0302] 现在参考重新分段过程,它分析由初始分段过程创建的片段,并标识要重新组合的成对的错误分离的片段。重新分段过程可以考虑初始分段过程中未考虑的分段的特征。作为示例,片段的特征可以是片段的长度。因此,一对错误地分开的片段可以指的是相邻片段,当从一对片段的长度考虑时,它们不具有明显不同的统计读取数序列读数计数。通常,较长的片段与统计读取数序列读数计数的较高变化相关。这样,通过考虑每个片段的长度,最初确定为每个片段具有统计读取数序列读数计数不同于其他片段的相邻片段可以被视为一对错误分离的片段。

[0303] 所述对中错误地被分开的片段被组合。因此,执行初始分段和重新分段过程导致参考基因组的生成片段,所述片段考虑了由每个片段的不同长度引起的差异。

[0304] 在步骤1250中,基于所述片段的观察的片段序列读数计数和所述片段的预期的片段序列读数计数,为每个片段确定片段分数。所述片段的观察的片段序列读数计数代表所述片段中分类的观察的序列读数的总数。因此,所述片段的观察的片段序列读数计数可以通过将包含在所述片段中的多个读取数的多个观察的读取序列读数计数相加来确定。类似地,预期的片段序列读数计数代表包含在所述片段中涵盖多个读取数的多个预期的序列读数计数。因此,可以通过量化片段中包含的多个读取数的多个预期的序列读数计数来计算片段的预期的片段序列读数计数。包含在片段中的读取数的预期的序列读数计数可从读取数预期计数存储器1280得到。

[0305] 片段的片段分数可以表示为所述片段的片段序列读数计数与所述片段的期望的

片段序列读数计数的比值。在一个实施例中,片段的片段分数可以表示为所述片段的观察的序列读数计数与所述片段的预期的序列读数计数的比率的函数。片段k的片段分数 s_k 可以表示为:

$$[0306] \quad s_k = \log \left(\frac{\text{观察的片段序列读数计数}}{\text{预期的片段序列读数计数}} \right) \quad (18)$$

[0307] 在其它实施例中,所述片段的片段分数可表示为比率的平方根(例如 $\sqrt{\frac{\text{观察的}}{\text{预期的}}}$)、比率的广义对数变换(例如 $\log \left(\text{观察的} + \sqrt{\text{观察的}^2 + \text{预期的}} \right)$),或比率的其他方差稳定变换之一。

[0308] 在步骤1255,为每个片段确定片段方差估计。通常,片段方差估计表示片段的序列读数计数有多偏离。在一个实施例中,可以通过使用包括在所述片段中的读取数的读取数方差估计并通过片段膨胀系数($I_{\text{片段}}$)进一步调整读取数方差估计来确定片段方差估计,举例来说,片段k的片段方差估计可以表示为:

$$[0309] \quad \text{var}_k = \text{平均值}(\text{var}_i) * I_{\text{片段}} \quad (19)$$

[0310] 其中,平均值(var_i)表示片段k中包含的读取数i的读取数方差估计的平均值。可以通过进入读取数期望方差存储器1290来获得读取数方差估计值。

[0311] 片段膨胀系数解释了片段水平上的增加的偏差,所述偏差通常比读取数等级上的偏差更高。在各种实施例中,片段膨胀系数可以根据片段的尺寸缩放。例如,一个由大量读取数组成的较大片段将被分配一个片段膨胀系数,所述片段膨胀系数大于分配给由较少读取数组成的较小片段的片段膨胀系数。因此,片段膨胀系数说明了较长片段中出现的更高水平的偏差。在各种实施例中,分配给第一样本的片段的片段膨胀系数不同于分配给第二样本的相同片段的片段膨胀系数。在各个实施例中,可以预先根据经验确定具有特定长度的片段的片段膨胀系数 $I_{\text{片段}}$ 。

[0312] 在各个实施例中,可以通过分析训练样本来确定每个片段的片段方差估计。例如,一旦在步骤1245中产生了多个片段,就分析来自训练样本的序列读数,以确定每个产生的片段的预期片段序列读数计数和每个片段的预期片段方差估计。

[0313] 每个片段的片段方差估计可以表示为使用训练样本通过膨胀系数调整后确定的每个片段的预期片段方差估计。例如,片段k的片段方差估计(var_k)可以表示为:

$$[0314] \quad \text{var}_k = \text{var}_{\text{预期}_k} * I_{\text{样本}} \quad (20)$$

[0315] 其中 $\text{var}_{\text{预期}_k}$ 是片段k的预期片段方差估计,而 $I_{\text{样本}}$ 是上面相对于步骤1235和公式(4)所述的样本膨胀系数。

[0316] 在步骤1260,基于片段分数和针对片段的片段方差估计,分析每个片段以确定所述片段在统计上是否有意义。对于每个片段k,可以将片段的片段分数(s_k)和片段方差估计(var_k)组合起来以生成所述片段的z-分数。片段k的z-分数(z_k)的示例可以表示为:

$$[0317] \quad z_k = \frac{s_k}{\text{var}_k} \quad (21)$$

[0318] 为了确定片段是否为统计上重要的片段,将所述片段的z-分数与阈值进行比较。如果所述片段的z-分数大于阈值,则将所述片段视为具有统计意义的片段。相反,如果片段的z-分数小于阈值,则所述片段不被视为统计上重要的片段。在一个实施例中,如果片段的z-分数大于2,则确定所述片段在统计上是有意义的。在其他实施例中,如果片段的z-分数大于2.5、3、3.5或4,则将所述片段确定为具有统计意义。在一些实施例中,如果片段的z-分数小于-2,则确定所述片段在统计上是有意义的。在其他实施例中,如果片段的z-分数小于-2.5、-3、-3.5或-4,则将所述片段确定为具有统计意义。统计上重要的片段可以指示样本(例如,cfDNA或gDNA样品)中存在的一个或多个复制次数事件。

[0319] 返回图12A,在步骤1215,确定由从cfDNA样本导出的具有统计意义的读取数(例如,在步骤1240中确定)及/或具有统计意义的片段(例如,在步骤1260中确定)所指出的复制次数事件的来源。具体而言,将cfDNA样本的具统计意义的读取数与gDNA样本的相应读取数进行比较。另外,将cfDNA样本的具统计意义的片段与gDNA样本的相应片段进行比较。

[0320] 将cfDNA样本的具有统计意义的片段和读取数与gDNA样本的相应片段和读取数进行比较,可以确定cfDNA样本的具有统计意义的片段和读取数是否与gDNA样本的相应片段和读取数对齐。如下文所用,对齐片段或读取数是指片段或读取数在cfDNA样本和gDNA样本中均具有统计意义。相反地,未对齐或不对齐的片段或读取数是指这些片段或读取数在一个样本(例如cfDNA样本)中具有统计意义,但在另一个样本(例如gDNA样本)中不具有统计意义。

[0321] 通常,如果cfDNA样本的具有统计意义的读取数和具有统计意义的片段与gDNA样本的相应的也具有统计意义的读取数和片段对齐,这表明在cfDNA样本和gDNA样本中都存在相同的复制次数事件。因此,因此复制次数事件的来源很可能是本身事件,即很可能是复制次数变异。

[0322] 相反地,如果cfDNA样本的具有统计意义的读取数和具有统计意义的片段与gDNA样本中不具统计意义的相应的读取数和片段对齐,则表明复制次数事件存在于cfDNA样本中,但不存在于gDNA样本中。在这种情况下,cfDNA样本中复制次数事件的来源是一个体细胞肿瘤事件引起的,复制次数事件是一复制次数畸变。

[0323] 识别在cfDNA样本中检测到的复制次数事件的来源有助于筛选出由种系或体细胞非肿瘤事件引起的复制次数事件。这提高了正确识别由实体瘤引起的复制次数畸变的能力。

[0324] 在一些目的中,结合本文公开的疾病状况,在分析核酸样本的序列读数的方法中使用尺寸选择的游离细胞DNA(cfDNA)序列读数。可以通过体外选择特定尺寸范围的cfDNA,即在生成测序数据之前,或通过计算器筛选序列读数的数据来实现尺寸选择。

[0325] 有利的是,发现使用尺寸选择的cfDNA测序数据提高了疾病分类器基于参考基因组中低变异区域获得的信息的疾病分类的灵敏度。例如,如下文所述,使用小于160个核苷酸的上限阈值选择来自癌症患者的cfDNA片段的测序数据,可显著增加数据集中源自癌症的序列读数的比例。此外,尽管尺寸选择显著降低了数据集的序列覆盖率,以及癌症衍生的cfDNA的序列读数的总数,当应用于基于人类基因组低变异区域的信息的癌症状态分类器时,使用尺寸选择的测序数据会产生更高的灵敏度。

[0326] 因此,本文描述了提高癌症分类可信度的各种方法。事实上,这些方法中的一些不

仅提高了癌症分类的可信度,而且减少了分类所需的DNA测序数据量,从而提高了分类过程的速度,同时降低了分析的成本和计算负担。

[0327] 在一目的中,本发明提供了一种改进的系统和方法,基于对来自受试者生物样本的游离细胞DNA序列读数的分析,对受试者进行癌症分类,并对其进行计算器筛选,以富集来自癌细胞衍生片段的序列读数。例如,通过去除大于阈值长度的游离细胞DNA片段的序列读数,所述阈值长度小于160个核苷酸。有利地,与从样本中对游离细胞DNA进行测序获得的完整序列读数相比,由于筛选后的序列读数集所包含的序列读数较少,因此减轻了处理数据集并将处理后的数据应用于分类器的计算负担,提高了用于对受试者的癌症状态进行分类的计算器系统的效率,并减少了总体时间。此外,出乎意料地发现,尽管通过筛选过程删除了大部分可用数据,但是通过使用筛选后的数据集可以提高进行分类的置信度。例如,如实施例6和实施例7中所述,利用经过计算器筛选的测序数据,去除长度超过150个核苷酸的cfDNA分子中的序列读数,使用一种基于在参考人类基因组中具有低变异性的预定数量基因读取数的复制次数畸变的分类器,提高了癌症检测的灵敏度。具体地说,图17A至图17D和图17F至图17G描述了使用来自1至100个核苷酸、0至140个核苷酸、90至140个和90至150个核苷酸的cfDNA片段的测序数据,分类的灵敏度增加了95%、98%和99%。

[0328] 一方面,本公开提供了改进的系统和方法,用于基于对来自受试者的生物样本的游离细胞DNA的序列读数的分析来对受试者进行癌症分类,所述DNA序列是在体外选择的,以移除大于阈值长度的游离细胞DNA片段,例如少于160个核苷酸。有利的是,由于来自生物样本的游离细胞DNA是尺寸选择的,需要测序的DNA总量减少了。反过来,可以在单个测序反应中合并更多样本,从而减少了测序成本和每个样本的时间。此外,由于从每个样本产生的序列读数较少,因此减少了处理数据集并将处理后的数据应用于分类器的计算负担,从而提高了用于对受试者癌症状态进行分类的计算器系统的效率,并降低了总体时间。而且,出乎意料地发现,尽管没有从样本中获得很大一部分潜在的测序数据,但是通过使用小的数据集可以提高进行分类的置信度。例如,如实施例9中所述,在30至140个核苷酸和30至150个核苷酸范围内选择cfDNA片段的尺寸之后,源自癌症来源的cfDNA片段的序列读数的部分丰富了从cfDNA样本产生的测序数据。具体而言,图19显示,体外尺寸选择增加了65个样本的肿瘤比例,这些样本来自于诊断为10种癌症之一的受试者,并且代表了所有癌症阶段的分布。此外,实施例10表明,利用从体外尺寸选择的cfDNA片段生成的测序数据,使用基于参考人类基因组中具有低变异性的预定数量的基因组读取数的复制次数畸变的分类器来提高癌症检测的灵敏度。具体而言,据表6所示,cfDNA片段的体外尺寸选择在95%、98%和95%的特异性下将分类器的灵敏度提高了约20%至30%。

[0329] 在一些方面,所公开的方法与癌症分类模型结合工作。例如,机器学习或深度学习模型(例如疾病分类器)可用于基于从由cfDNA片段产生的尺寸选择的序列读数确定的一个或多个特征的值来确定疾病状态。在各种实施例中,机器学习或深度学习模型的输出是疾病状态的预测分数或概率(例如,癌症预测分数)。因此,机器学习或深度学习模型会基于预测分数或概率生成疾病状态分类。

[0330] 参照图13和图14公开了根据本公开的各种实施例的关于方法和系统的流程和特征的细节。在一些实施例中,系统的此类流程和特征由示例系统700中描述的各种模块执行,如图7所示。

[0331] 以下描述的实施例涉及使用从生物样本例如血液样本获得的游离细胞DNA片段的序列读数进行的分析。通常,这些实施例是独立的,因此,不依赖于任何特定的测序方法。然而,在一些实施例中,下面描述的方法包括生成用于分析的序列读数的一个或多个步骤,及/或指定对正在执行的特定类型的分析有利的某些测序参数。在一些实施例中,如下文所述,与cfDNA的大尺寸选择序列读数相关的实施例与用于训练分类器及/或分类疾病状态的方法200、210、300、400和500中的任何一种结合使用。

[0332] 图13描述了分析一个样本的流程,所述流程基于从体外尺寸选择的cfDNA测序数据中获得的信息进行分析。流程1300说明了如何使用来自受试者的测试数据(其相对医疗状况(例如癌症)的状态未知)来计算分类分数并作为诊断受试者是否可能患有所述疾病的基础。

[0333] 在步骤1302,获得来自疾病状态的可能未知的个体的生物样本。在一些实施例中,样本包括受试者的体液,例如血液、全血、血浆、血清、尿液、脑脊液、粪便、唾液、汗液、眼泪、胸膜液、心包液、腹膜液、其他类型的体液或其任何组合。在一些实施例中,有利的是,用于提取流体样本的方法(例如,通过注射器或手指刺取血液样本)比用于获取组织活检的程序(可能需要手术)具有更小的侵入性。在一些实施例中,生物样本包括cfDNA。

[0334] 在一些实施例中,血液样本是全血样本,并且在从全血样品产生多个序列读数之前,从全血样本中去除白血球。在一些实施例中,白血球被收集为第二种类型的样本,例如,根据血沉棕黄层提取方法,可从中获得或不可获得额外的测序数据。用于白血球的血沉棕黄层提取方法在本领域中是已知的,例如,如在2018年6月1日提交的美国专利申请临时案申请序列号62/679,347中所述,其内容出于所有目的通过引用并入本文。

[0335] 在一些实施例中,所述方法还包括从全血样本中移除的白血球中获得基因组DNA的电子形式的多个第二序列读数。在一些实施例中,多个第二序列读数用于识别来自克隆造血的等位基因变体,而不是来自个体癌症的生殖系等位基因变体及/或等位基因变体。

[0336] 在一些实施例中,生物样本包括长于第一阈值长度的游离细胞DNA分子,其中第一阈值长度小于160个核苷酸。然而,在一些实施例中,本文描述的分类器训练、分类器验证和疾病分类方法中使用的序列读数的数据不包括长于第一阈值长度的游离细胞DNA分子的序列读数。因此,所使用的序列读数的数据代表生物样本中游离细胞DNA分子序列的降维空间。由于需要较少的计算,因此对减少的这组的序列读数的分析减少了处理测序数据的计算负担,从而减少了所需的时间并提高了执行分析的计算器系统的效率。

[0337] 在步骤1304,从样本中分离出作为测序反应模板的cfDNA。从生物样本中分离cfDNA的方法是本领域众所周知的。对于市售的游离细胞DNA分离试剂盒的比较,请参阅例如Sorber, L.等人,分子诊断杂志,19(1):162-68(2017)。其内容出于所有目的通过引用并入本文。

[0338] 在一些实施例中,例如,在cfDNA片段的尺寸选择之前,在步骤1304制备测序库。在库制备过程中,通过接合子连接将独特分子标识子(UMIs)添加到核酸分子(如DNA分子)中。在一些实施例中,UMIs是在接合子连接过程中添加到DNA片段末端的短核酸序列(例如4个碱基对至10个碱基对)。在一些实施例中,UMIs是简并碱基对,其用作可用于识别源自特定DNA片段的序列读取的唯一标记。在一些实施例中,例如,当将多重测序用于在单个测序反应中对来自多个受试者的cfDNA进行测序时,还将患者特异性索引序列添加至核酸分子。在

一些实施例中,患者特异性索引序列是在库构建过程中添加至DNA片段末端的短核酸序列(例如3至20个核苷酸),其可用作可用于鉴定源自于特定的患者样本的序列读数的独特标签。在接合子连接后的PCR扩增过程中,UMI与附着的DNA片段一起复制。这提供了一种在下游分析中鉴定来自相同原始片段的序列读数的方法。

[0339] 在一些实施例中,所述库的构建包括将具有固定长度 x 个核苷酸的核酸片段添加到受试者的游离细胞DNA分子中,其中核酸片段包括个体特有的独特分子辨识子。在一些实施例中,将核酸片段添加到游离细胞DNA分子的两端。因此,如本文所述, x 个核苷酸的固定长度是指添加到游离细胞DNA分子任一端的所有核酸片段的总长度。在一些实施例中,独特分子辨识子对选自于下列的集合进行编码—独特预定值: $\{1, \dots, 1024\}$, $\{1, \dots, 4096\}$, $\{1, \dots, 16,384\}$, $\{1, \dots, 65,536\}$, $\{1, \dots, 262,144\}$, $\{1, \dots, 1,048,576\}$, $\{1, \dots, 4,194,304\}$, $\{1, \dots, 16,777,216\}$, $\{1, \dots, 67,108,864\}$, $\{1, \dots, 268,435,456\}$, $\{1, \dots, 1,073,741,824\}$, 或 $\{1, \dots, 4,294,967,296\}$ 。在一些实施例中,独特分子辨识子定位于所添加的核酸片段内的一组相邻的寡核苷酸。在一些实施例中,相邻的寡核苷酸集合是 N -聚体,其中 N 是选自集合 $\{4, \dots, 20\}$ 的整数。在一些实施例中,核酸片段还包括UMI、启动子杂交序列(例如用于PCR扩增及/或测序)和用于聚类的互补序列中的一个或多个。在一些实施例中,所添加的核酸片段的固定长度 x 是从100个核苷酸到200个核苷酸。在其它实施例中,所添加核酸片段的固定长度 x 约为50、55、60、65、70、75、80、85、90、95、100、105、110、115、120、125、130、135、140、145、150、155、160、165、170、175、180、185、190、195、200、205、210、215、220、225、230、235、240、245、250或更多个核苷酸。

[0340] 在一些实施例中,来自多个受试者的cfDNA测序库,无论库是在尺寸选择之前还是之后准备的,在测序之前被汇集在一起。通过将样本集中在一起进行下一代测序可以获得一些优势。首先,由于下一代测序仪的高通量能力,单个反应需要大量的模板DNA。通过将cfDNA测序库集中在一起,测序反应需要每个患者更少的cfDNA。第二,由于单个测序反应的成本基本上是固定的,因此将cfDNA测序库合并在一起进行测序,可以将每个cfDNA库的测序成本降低一倍。

[0341] 在步骤1306,选择cfDNA分子的尺寸,例如,移除来自长于阈值长度(例如阈值长度小于160个核苷酸)的cfDNA片段的分子。尺寸选择核酸片段的方法是本领域已知的,例如洋菜胶电泳。在一些实施例中,尺寸选择发生在库准备之前,并且在其他实施例中发生在库准备之后。在一些实施例中,当在库准备之后发生尺寸选择时,在尺寸选择之前,来自多个受试者的cfDNA片段库被汇集在一起。在尺寸选择之前合并cfDNA库的一个优点是,由于尺寸选择技术的成本基本上是固定的,因此在单个反应(例如,基于洋菜胶电泳技术的一个孔)中选择cfDNA库可以降低每个样本的选择成本。

[0342] 一般而言,设定阈值长度是为了增加源于癌细胞的cfDNA片段(相对于来自体细胞或造血细胞的cfDNA片段)产生的序列读数百分比。例如,从图15中cfDNA片段长度分布随肿瘤分数的变化可以看出,平均而言,源自癌细胞的cfDNA片段的长度比源自体细胞或造血细胞的cfDNA片段的长度短。因此,给定片段衍生自癌细胞的可能性随着片段尺寸的减小而增加。因此,在一些实施例中,第一阈值长度被设定为小于160个核苷酸的值。在一些实施例中,第一阈值长度是150个核苷酸或更少。在一些实施例中,第一阈值长度为140个核苷酸或更少。在一些实施例中,第一阈值长度为130个核苷酸或更少。在一些实施例中,第一阈值长

度为159、158、157、156、155、154、153、152、151、150、149、148、147、146、145、144、143、142、141、140、139、138、137、136、135、134、133、132、131、130、129、128、127、126、125个核苷酸或更少的核苷酸。在一个实施例中，第一阈值长度为140个核苷酸。在一些实施例中，第一阈值长度在130核苷酸和150核苷酸之间。在一些实施例中，第一阈值长度在140核苷酸和150核苷酸之间。在一些实施例中，第一阈值长度在130核苷酸和140核苷酸之间。

[0343] 对于来自单核小体结构的cfDNA片段，从双核小体片段衍生的cfDNA片段也观察到类似的尺寸现象。也就是说，长度在大约220到340个核苷酸之间的游离细胞DNA片段通常来自双核小体结构。平均而言，来自于肿瘤细胞的双核小体结构的cfDNA片段比来自体细胞或造血细胞的双核小体结构体的cfDNA片段的长度短。因此，在一些实施例中，为了从富含癌源的cfDNA片段提供更多的测序数据，从来自双核小体结构的较短cfDNA分子生成的序列读数也包括在用于确定受试者的癌症状态的多个序列读数中。

[0344] 因此，在一些实施例中，长度在第二阈值长度和第三阈值长度之间的游离细胞DNA片段的序列读数被包括在已筛选的数据集中。在一些实施例中，第二阈值长度为240核苷酸至260核苷酸，第三阈值长度为290核苷酸至310核苷酸。在一些实施例中，第二阈值长度为250个核苷酸。在其它实施例中，第二阈值长度为240、241、242、243、244、245、246、247、248、249、250、251、252、253、254、255、256、257、258、259个核苷酸或260个核苷酸。在一些实施例中，第三阈值长度为300个核苷酸(3028)。在一些实施例中，第三阈值长度为290、291、292、293、294、295、296、297、298、299、300、301、302、303、304、305、306、307、308、309个核苷酸或310个核苷酸。

[0345] 当从合并的cfDNA测序库中选择片段尺寸时，所选长度根据原始cfDNA片段(w)的期望长度范围和接合子的长度(x，例如包含UMIs、启动子位点、患者特定指数等)的总和来确定，例如 $w+x$ 。

[0346] 在步骤1308，从尺寸选择的cfDNA库及/或库中产生序列读数。测序数据可以通过本领域已知的手段来获得。例如，下一代测序(NGS)技术包括合成技术(Illumina)、焦磷酸测序(454Life Sciences)、离子半导体技术(Ion Torrent测序)、单分子实时测序(Pacific Biosciences)、通过连接测序(固体测序)、纳米孔测序(Oxford Nanopore Technologies)或配对末端测序。

[0347] 在一些实施例中，使用具有可逆染料终止子的合成测序进行大规模平行测序。在一些实施方案中，在测序反应是来自一个以上样品和/或受试者的cfDNA的多重测序反应的情况下，随后基于识别出的cDNA，对测序数据进行解复用，以基于独特的UMI序列鉴定来自每个样本及/或受试者的序列读数。在一些实施例中，针对受试者的物种，涵盖参考基因组的序列读数的平均覆盖率是至少3倍。在一些实施例中，平均覆盖率是至少5倍。在一些实施例中，平均覆盖率是至少10倍。在一些实施例中，平均覆盖率在约0.1倍和约35倍之间，或在约2倍和约20倍之间，例如，约0.1倍、0.5倍、1倍、2倍、3倍、4倍、5倍、6倍、7倍、8倍、9倍、10倍、11倍、12倍、13倍、14倍、15倍、16倍、17倍、18倍、19倍、20倍、25倍、30倍、35倍等。在一些实施例中，在测序之前对cfDNA片段进行尺寸选择的情况下，测序反应的序列覆盖率在这个覆盖率的低端。例如，研究发现，在5倍覆盖率下对次采样cfDNA数据进行尺寸选择，导致平均序列覆盖率仍为0.09倍，与在5倍覆盖率下的非尺寸选择数据一样。因此，如果在测序前对cfDNA进行尺寸选择，那么低的序列覆盖率将有望提供高特异性的必要的诊断灵敏度。相

应地,在一些实施例中,在测序之前对cfDNA片段进行尺寸选择的情况下,平均覆盖率在大约0.1倍和大约5倍之间,或者在大约0.5倍和大约3倍之间,平均覆盖率介于大约0.1x大约0.1x、0.2x、0.3x、0.4倍、0.5倍、0.6倍、0.7倍、0.8倍、0.9倍、1倍、1.25倍、1.75倍、2倍、2.5倍、3倍、3.5倍、4倍、4.5倍、5倍、6倍、7倍、8倍、9倍,或者10倍。

[0348] 然后如上所述分别参考流程500的步骤520、530和540,分别执行步骤1310、1312和1314。实际上,在一些实施例中,如上所述,在步骤1308生成的测序数据用作流程200、210、300、400或500中的任何过程的输入。

[0349] 在一些实施例中,使用第一置信度来确定步骤1314中提供的诊断,如果在步骤1312中使用尚未尺寸选择的测序数据来计算分类分数,则第一置信度大于将提供的第二置信度。在一些实施例中,尽管使用尺寸选择的序列读数以更大的置信度确定与受试者中疾病(例如癌症)的阳性诊断相关的疾病状态(例如癌症类别),但是与使用一组非尺寸选择的序列读数的情况相比,不能更可靠地做出与受试者没有疾病的诊断相关的诊断状态。也就是说,在一些实施例中,本文提供的方法导致疾病分类,当受试者患有疾病时,以较高的置信度进行疾病分类,但是当受试者没有疾病时,以相似或较低的置信度进行疾病分类。当然,这并不需要实际执行带有从生物样本中读取的全套序列的分类,也不需要实际计算出第二置信度。然而,在一些实施例中,确定第二分类及/或置信度。

[0350] 图14描绘了用于基于从计算器上选择的具有减小的维数的cfDNA测序数据中学到的信息来分析数据的示例流程。流程1400示出了来自受试者的关于医疗状况(例如癌症)的状态未知的测试数据如何可以用于计算分类得分并用作诊断受试者是否可能患有所述状况的基础。

[0351] 在步骤1402,获得来自疾病状态可能未知的受试者的cfDNA测序数据。在一些实施例中,cfDNA包括从受试者的体液(例如血液、全血、血浆、血清、尿液、脑脊液、粪便、唾液、汗液、眼泪、胸膜液、心包液、腹膜液、其他类型的体液或其任何组合)的样本中分离。在一些实施例中,有利的是,用于提取流体样本的方法(例如,通过注射器或手指刺取血液样本)的侵入性小于用于获得组织活检的程序,其可能需要手术,例如参考流程1300中的步骤1302所述。在一些实施例中,测序是在汇集cfDNA库之后执行的,例如,参照过程1300中的步骤1304如上所述制备。在一些实施例中,如上所述参照过程1300中的步骤1308执行测序。

[0352] 在一些实施例中,使用具有可逆染料终止子的合成测序进行大规模平行测序。在一些实施方案中,在测序反应是来自一个以上样品和/或受试者的cfDNA的多重测序反应的情况下,随后基于识别出的cDNA,对测序数据进行解复用,以基于独特的UMI序列鉴定来自每个样本及/或受试者的序列读数。在一些实施例中,针对受试者的物种,涵盖参考基因组的序列读数的平均覆盖率是至少3倍。在一些实施例中,平均覆盖率是至少5倍。在一些实施例中,平均覆盖率是至少10倍。在一些实施例中,平均覆盖率在约0.1倍和约35倍之间,或在约2倍和约20倍之间,例如,约0.1倍、0.5倍、1倍、2倍、3倍、4倍、5倍、6倍、7倍、8倍、9倍、10倍、11倍、12倍、13倍、14倍、15倍、16倍、17倍、18倍、19倍、20倍、25倍、30倍、35倍等。在一些实施例中,在测序之前不对cfDNA片段进行尺寸选择,而是筛选所得的序列读数的情况下,测序反应的序列覆盖率不在所述覆盖率的低端。例如,发现对具有5倍序列覆盖率的二次采样CCGA数据集进行计算器筛选,会导致筛选后的数据集仅具有0.09倍序列覆盖率。因此,在一些实施例中,在测序后按尺寸选择cfDNA序列读数的情况下,平均覆盖率在约2倍至约35

倍之间,或约3倍至约10倍之间,例如,大约或至少2倍、3倍、4倍、5倍、6倍、7倍、8倍、9倍、10倍、11倍、12倍、13倍、14倍、15倍、16倍、17倍、18倍、19倍、20倍、25倍、30倍或35倍。

[0353] 在步骤1404,对测序数据进行筛选以从长度在期望范围内的cfDNA片段中选择序列读数,例如,通过对在步骤1402中获得的前驱序列读数应用长度筛选器。在一些实施例中,这包括筛选序列读数以从长于阈值长度(例如,阈值长度小于160个核苷酸)的cfDNA分子中排除序列读数。在一些实施例中,这是通过例如根据参考基因组中起始和终止核苷酸碱基的位置确定对应于一个或多个序列读数的cfDNA的长度,并仅选择对应于cfDNA片段的那些序列读数来实现的,其长度落在期望的长度范围内,使得所选的序列读数子集被富集用于对应于源自癌细胞的cfDNA片段的序列读数。

[0354] 一般而言,参考如上文流程1300的步骤1306所述,设置阈值长度以增加针对源自癌细胞的cfDNA片段生成的序列读数的百分比,而不是来自体细胞或造血细胞的cfDNA片段生成的序列读数的百分比。例如,从图15中cfDNA片段长度分布随肿瘤分数的变化可以看出,平均而言,源自癌细胞的cfDNA片段的长度比源自体细胞或造血细胞的cfDNA片段的长度短。因此,给定片段衍生自癌细胞的可能性随着片段尺寸的减小而增加。因此,在一些实施例中,第一阈值长度被设定为小于160个核苷酸的值。在一些实施例中,第一阈值长度是150个核苷酸或更少。在一些实施例中,第一阈值长度为140个核苷酸或更少。在一些实施例中,第一阈值长度为130个核苷酸或更少。在一些实施例中,第一阈值长度为159、158、157、156、155、154、153、152、151、150、149、148、147、146、145、144、143、142、141、140、139、138、137、136、135、134、133、132、131、130、129、128、127、126、125个核苷酸或更少的核苷酸。在一个实施例中,第一阈值长度为140个核苷酸。在一些实施例中,第一阈值长度在130核苷酸和150核苷酸之间。在一些实施例中,第一阈值长度在140核苷酸和150核苷酸之间。在一些实施例中,第一阈值长度在130核苷酸和140核苷酸之间。

[0355] 对于来自单核小体结构的cfDNA片段,从双核小体片段衍生的cfDNA片段也观察到类似的尺寸现象。也就是说,长度在大约220到340个核苷酸之间的游离细胞DNA片段通常来自双核小体结构。平均而言,来自于肿瘤细胞的双核小体结构的cfDNA片段比来自体细胞或造血细胞的双核小体结构体的cfDNA片段的长度短。因此,在一些实施例中,为了从富含癌源的cfDNA片段提供更多的测序数据,从来自双核小体结构的较短cfDNA分子生成的序列读数也包括在用于确定受试者的癌症状态的多个序列读数中。

[0356] 然后分别参照流程500的步骤520、530和540,如上所述执行步骤1410、1412和1414。事实上,在一些实施例中,如上文所述,在步骤1404生成的已筛选的测序数据被用作流程200、210、300、400或500中的任何一个的输入。

[0357] 在一些实施例中,在步骤1314中提供的诊断以第一置信度来确定,第一置信度大于第二置信度,如果步骤1312中的分类分数是使用未选择尺寸的排序数据来计算的,则会提供第二置信度。在一些实施例中,尽管使用尺寸选择的序列读数以更大的置信度确定与受试者中疾病(例如癌症)的阳性诊断相关的疾病状态(例如癌症类别),但是与使用一组非尺寸选择的序列读数的情况相比,不能更可靠地做出与受试者没有疾病的诊断相关的诊断状态。也就是说,在一些实施例中,本文提供的方法导致疾病分类,当受试者患有疾病时,以较高的置信度进行疾病分类,但是当受试者没有疾病时,以相似或较低的置信度进行疾病分类。当然,这并不需要实际执行带有从生物样本中读取的全套序列的分类,也不需要实际

计算出第二置信度。然而,在一些实施例中,确定第二分类及/或置信度。

[0358] 示例系统架构

[0359] 图7描绘了用于实现图1至图6的特征和流程的示例性系统架构的图。

[0360] 在一目的中,一些实施例可以使用计算机系统(例如计算机系统700)来执行根据本发明的各种实施例的方法。计算机系统700的示例性实施例包括总线702、一个或多个处理器712、一个或多个存储设备714、至少一输入设备716、至少一输出设备718、通信子系统720、工作存储器730,其包括:操作系统732、设备驱动程序,可执行数据库及/或其他代码(例如一个或多个应用程序734)。

[0361] 根据一组实施例,计算机系统700响应于处理器712执行包含在工作中的一个或多个指令(可并入操作系统732及/或其他代码,例如应用程序734)的一个或多个序列,执行这些方法的部分或全部过程包括在工作存储器730中。这样的指令可以从另一计算机可读介质(例如存储设备714中的一个或多个)读入工作存储器730中。仅作为示例,执行包含在工作存储器730中的指令序列可以使得处理器712执行本文所述方法的一个或多个过程。另外,或者可选地,本文所描述的方法的部分可以通过专用硬件来执行。仅作为示例,关于上述方法描述的一个或多个过程的一部分,例如方法200、方法210、方法300、方法310、方法400、方法500、方法600、以及图2至图6中所示的任何变化,可以由处理器712执行。在一些实施例中,处理器712可与系统100或系统110结合使用。在一些示例中,应用程序734可以是执行图2-6所示的迭代实时学习方法的应用的示例。

[0362] 在一些实施例中,计算机系统700可以进一步包括(及/或与其通信)一个或多个非临时存储设备714,其可以包括但不限于本地及/或网络可访问存储组件,及/或可以包括但不限于磁盘驱动器、驱动器阵列、光存储设备、固态存储器设备,例如随机存取存储器(“RAM”)及/或唯读存储器(“ROM”),其可编程、可闪存更新等。这种存储设备可以被配置成实现任何适当的数据存储,包括但不限于各种文件系统、数据库结构等。在一些实施例中,存储设备714可以是数据库130的示例。

[0363] 在一些实施例中,计算机系统700可以进一步包括一个或多个输入设备716,其可以包括但不限于允许计算机设备从用户、从另一计算机设备、从计算机设备的环境接收信息的任何输入设备,或来自与计算机设备可通信连接的功能部件。

[0364] 在一些实施例中,计算机系统700可以进一步包括一个或多个输入输出设备718,其可以包括但不限于任何输出设备,所述输出设备可以从计算机设备接收信息并将所述信息传送给用户、另一个计算机设备和计算机设备的环境,或与计算机设备可通信连接的功能部件。输入设备的示例包括但不限于显示器、扬声器、打印机、灯、传感器设备等。传感器设备可以接收和显示能够导致用户感觉的形式的数据。这些形式包括但不限于热、光、触、压、动等。

[0365] 可以理解的是,任何适用的输入/输出设备或组件,例如与系统100或110有关的那些,都可以应用于输入设备716和输出设备718。

[0366] 在一些实施例中,计算机系统700还可以包括通信子系统720,其可以包括但不限于调制解调器、以太网连接、网卡(无线或有线)、红外通信设备、无线通信设备及/或芯片组(例如蓝牙™、装置、802.11装置、WiFi装置、WiMax装置、移动通信设施等)、近场通信(NFC)、Zigbee通信、射频(RF)或射频识别(RFID)通信、PLC协议、基于3G/4G/5G/LTE的通信等。通信

子系统720可以包括一个或多个输入及/或输出通信接口,以允许与网络、其他计算机系统及/或任何其他电气设备/外围设备交换数据。在许多实施例中,如上文所述,计算机系统700还将包括工作存储器730,其可以包括RAM或ROM设备。

[0367] 在一些实施例中,计算机系统700还可以包括软件组件,显示为设置于工作存储器730内,包括操作系统732、设备驱动程序、可执行数据库及/或其他代码,例如一个或多个应用程序734,其可以包括由各种实施例提供的计算机程序,及/或可以设计为实现由其他实施例提供的方法及/或配置系统,如本文所述。仅作为示例,关于上述方法描述的一个或多个过程的一部分,例如相对于图2至图6,可实现为由计算机(及/或计算机内的处理单元)执行的代码及/或指令;在一个方面,则可以使用这些代码及/或指令来配置。在一些实施例中,通用计算机(或其他设备)可用于根据所述方法执行一个或多个操作。在一些实例中,工作存储器730可以是与系统100或系统110连接使用的任何设备的存储器的示例。

[0368] 这些指令及/或代码的集合可以存储在非临时计算机可读存储介质上,例如上述存储设备714。在一些情况下,存储介质可以并入计算机系统中,例如计算机系统600。在其他实施例中,存储介质可以与计算机系统(例如,可移动介质,例如光盘)分开,及/或提供在安装包中,使得存储介质可以用于对具有存储在其上的指令/代码的通用计算机进行编程、配置及/或调整。这些指令可以采取可执行代码的形式,可由计算机系统700执行及/或可以采用源代码及/或可安装代码的形式,在计算机系统700上编译及/或安装时(例如,使用各种通用的编译器、安装程序,压缩/解压实用程序等),然后采用可执行代码的形式。在一些实例中,存储设备730可以是设备102、220或240的存储器的例示。

[0369] 对于本领域技术人员来说,显而易见的是,可以根据特定要求进行实质性的变更。例如,还可以使用定制的硬件、及/或可以在硬件、软件(包括便携式软件,例如小应用程序等)中执行特定的组件,或者两者都执行。此外,可以使用到诸如网络输入/输出设备之类的其他计算设备的连接。

[0370] 本文使用的术语“机器可读介质”和“计算机可读介质”是指参与提供导致机器以特定方式操作的数据的任何介质。在使用计算机系统700执行的实施例中,各种计算机可读介质可用于向处理器712提供用于执行的指令/代码及/或可用于存储及/或携带此类指令/代码。在许多实例中,计算机可读介质是物理及/或有形存储介质。这种介质可以采取非易失性介质或易失性介质的形式。非易失性介质包括例如光盘及/或磁盘,例如存储设备714。易失性介质包括但不限于动态存储器,例如工作存储器730。

[0371] 物理及/或有形计算机可读介质的常见形式包括例如软盘、软盘、硬盘、磁带、闪存、闪存驱动器或任何其他磁性介质、CD-ROM、任何其他光学介质、具有孔图案的任何其他物理介质、RAM、PROM、EPROM、FLASH-EPROM,任何其他存储芯片或盒式磁带,或计算机可以从中读取指令及/或代码的任何其他介质。

[0372] 各种形式的计算机可读介质可涉及将一个或多个指令的一个或多个序列传送给处理器712以供执行。仅通过示例的方式,指令可以最初承载在远程计算器的磁盘及/或光盘上。远程计算器可以将指令加载到其动态存储器中,并通过传输介质将指令作为信号发送,以由计算机系统700接收及/或执行。

[0373] 通信子系统720(及/或其组件)通常将接收信号,然后总线702可以将信号(及/或信号所承载的数据、指令等)传送到工作存储器730,处理器712从工作存储器730检索和执

行指令。由工作存储器730接收的指令可以任选地在处理器712执行之前或之后存储在非临时存储设备714上。

[0374] 实施例

[0375] 提供以下非限制性例示以进一步说明本文所公开的本发明的实施例。本领域技术人员应当理解,在下面的例示中公开的技术表示已经被发现在本发明的实施中具有良好作用的方法,因此可以被认为构成用于其实施的模式的示例。然而,根据本发明,本领域技术人员应当理解,可以在所公开的特定实施例中进行许多改变,并且仍然获得相似或类似的结果,而不脱离本发明的精神和范围。

[0376] 实施例1

[0377] B-分数和Z-分数的比较

[0378] 图8包括一个表格,所述表格将本揭示的方法(b-分数)与现有已知的分段方法(z-分数)进行了比较。数据显示,在乳腺癌样本的所有阶段中,b-分数的总体预测能力始终高于z-分数的预测能力。

[0379] 图9提供了患有乳腺癌的个体受试者的更详细的示例比较分类分数(z分数:顶部;b分数:底部)。同样,使用b分数,可以在浸润性乳腺癌的所有不同阶段正确预测更多受试者的癌症状态。

[0380] 实施例2

[0381] 不同类型的癌症

[0382] 图10显示了对于所有类型的癌症,可以观察到使用b分数的改进的预测能力(顶部)。对早期癌症的预测能力得到改善,对晚期癌症的预测能力尤其出色(底部)。

[0383] 图11A显示可以观察到患有肺癌(上)和前列腺癌(下)的受试者的改善。图11B显示可以观察到患有结肠直肠癌的受试者的改善。

[0384] 实施例3

[0385] 用于早期癌症检测的血浆游离细胞DNA(cfDNA)分析方法的开发:循环游离细胞基因组图谱研究(CCGA)的初步见解

[0386] 以下实施例4-7所示分析中使用的数据是与斯隆-凯特林癌症纪念中心(MSKCC)一起收集的,作为CCGA临床研究的一部分。CCGA[NCT02889978]是基于cfDNA的早期癌症检测的最大研究;本文示出了从多个cfDNA检测中获得的第一个CCGA经验。这项前瞻性、多目标、观察性研究在141个地点的15000名人口均衡参与者中招收了9977人。血液的采集来自登记时诊断为未接受治疗的癌症患者(例C)和未诊断为癌症(非癌症,例NC,对照组)的参与者的血液。这项预先计划的亚研究包括878例病例,580个对照组和169个分析对照组(n=1627),涵盖20种肿瘤类型和所有临床阶段。所有样本的分析方法如下:1)配对的cfDNA和白血球(WBC)靶向测序(60,000X,507个基因组);联合主持者去除了白血球衍生的体细胞变异和残余的技术噪声;2)配对cfDNA和白血球全基因组测序(WGS;35X);一种新的机器学习算法生成癌症相关信号评分;联合分析识别共享事件;以及3)cfDNA全基因组亚硫酸氢盐测序(WGBS;34X);使用异常甲基化片段生成标准化分数。在靶向检测中,非肿瘤白血球匹配的cfDNA体细胞变体(SNVs/indels)占NC所有变异的76%,C占65%。与体细胞嵌合体(即克隆性造血)一致,白血球匹配的变异随着年龄的增长而增加;一些是以前未报道的非典型功能丧失突变。去除白血球变体后,典型驱动体细胞变体对C高度特异(例如在EGFR和PIK3CA

中,NC具有0个变体,C分别具有11和30个变体)。同样,在用WGS检测到的8例具有体细胞复制次数畸变(SCNAs)的NC中,4例来自白血球。WGBS数据展示了信息丰富的高片段和低片段水平的CpGs(1:2的比率);使用一个子集来计算甲基化分数。在所有试验中,只有不到1%的NC参与者观察到一致的“类似癌症”信号(代表潜在的未诊断出癌症)。在NC与I-III期和IV期之间观察到一个增加的趋势(异步。每Mb SNVs/指数[平均值±标准差]NC:1.01±0.86,I-III阶段:2.43±3.98;IV阶段:6.45±6.79;WGS评分NC:0.00±0.08,I-III:0.27±0.98;IV:1.95±2.33;甲基化分数NC:0±0.50;I-III:1.02±1.77;IV:3.94±1.70)。这些数据证明了对浸润性癌的特异性>99%的可行性,并支持cfDNA检测在早期癌症检测中的应用前景。其他数据将在检测到的血浆:组织变异一致性和多测定模型上提供。

[0387] CCGA研究中包括的癌症类型包括浸润性乳腺癌、肺癌、结直肠癌、DCIS、卵巢癌、子宫癌、黑色素瘤、肾癌、胰腺癌、甲状腺癌、胃癌、肝胆癌、食管癌、前列腺癌、淋巴瘤、白血病、多发性骨髓瘤、头颈部癌,还有膀胱癌。

[0388] 实施例4

[0389] 肿瘤患者体内游离DNA片段的分布

[0390] 通过全基因组测序(WGS)确定了游离细胞DNA片段长度的分布,研究对象是来自具有不同肿瘤分数的受试者的游离细胞DNA样本。简单地说,绘制了747名健康个体和1001名来自CCGA研究的确证癌症患者的WGS结果,作为受试者肿瘤分数的函数。图15显示了来自747名健康受试者(1502)的游离细胞DNA片段长度的平均分布,708名肿瘤分数小于1%的癌症患者(1504),136名肿瘤分数在1-5%之间的癌症患者(1506),61名癌症患者的肿瘤分数在5-10%之间(1508),73名癌症患者的肿瘤分数在10-25%之间(1510),22名癌症患者的肿瘤分数在25-50%之间(1512)和1名癌症患者肿瘤分数介于50%至100%之间(1514)。如图15所示,游离细胞DNA片段的长度分布随着患者肿瘤分数的变化而变短。也就是说,受试者的肿瘤分数与游离细胞DNA片段长度偏移的大小相关。这代表了癌症细胞和健康细胞在生物学上的不同,来源于癌细胞的游离细胞DNA的长度比来自健康细胞的游离细胞DNA的长度短。

[0391] 实施例5

[0392] 经计算机尺寸选择后的基因组序列覆盖率

[0393] 研究了通过游离细胞DNA样本的全基因组测序(WGS)产生的尺寸选择数据后获得的序列覆盖率。简单地说,从上述CCGA研究中获得的测序数据在计算机中被筛选,以仅包括从大小为90个核苷酸到150个核苷酸的cfDNA片段中获得的序列(图16B),或仅包括由尺寸为100个核苷酸或更小的cfDNA片段产生的序列(图16C)。然后计算未筛选的数据集的平均序列覆盖率,已筛选的数据集仅包括来自90到150个核苷酸的cfDNA片段的序列,并且已筛选的数据集仅包括来自100个核苷酸或更少的cfDNA片段的序列。如图16所示的直方图所示,未筛选的CCGA数据集的中值序列覆盖率约为34倍(图16A);从90个核苷酸到150个核苷酸的cfDNA片段中筛选出的数据集的中值序列覆盖率约为6倍(图16B);从100个核苷酸或更少的cfDNA片段中筛选出的数据集的中值序列覆盖率约为0.6倍(图16C)。图16还显示了每个分布的第五和第九十五个百分位。

[0394] 实施例6

[0395] 计算器尺寸选择后的癌症分类

[0396] 尽管从较小的cfDNA片段(例如少于150个核苷酸)产生的序列读数的选择应能丰富癌症患者样本中的癌症衍生的片段,但是由于某些原因,这种选择将导致关于癌症的信息的净损失,因为被移除的一些较大的cfDNA片段将来自于癌症。因此,尽管在尺寸选择的数据集中相对于非癌症衍生的片段而言,癌症衍生的片段得到了富集,但理论上来说,相对于完整的数据集,所述数据集的总体诊断能力将会降低。为了测试是否是这种情况,如本文所述,将如实施例5所述已筛选的数据集输入到针对涵盖多个预定基因组库的复制次数畸变而训练的癌症分类器中,每个预定基因组的读取数代表人类基因组的预定义部分。

[0397] 简言之,按照实施例5所述,筛选来自癌症患者和健康受试者(不包括子宫癌、甲状腺癌、前列腺癌、黑色素瘤、肾脏癌和HR+期I/II乳腺癌)序列读取的CCGA数据集,从长度为90至150个核苷酸的cfDNA中选择序列读数,或者从长度为100个核苷酸或更少核苷酸的cfDNA中选择序列读数。然后将已筛选的数据在已筛选的数据集中进行标准化,并将每个已筛选样本的标准化数据输入到逻辑回归分类器中,针对低方差基因组读取数的计数的特征进行训练。为了控制已筛选的数据集和未筛选的数据集之间的序列覆盖率差异,如在示例5中所呈现的,通过尺寸独立地从未筛选的数据集中随机选择序列读数来生成对照组数据集,以实现与相应的尺寸选择的数据集相同的序列覆盖。例如150至90个核苷酸的对照组数据集的中位数序列覆盖率约为6.2倍,0至100个核苷酸的对照组数据集的中位数序列覆盖率约为0.6倍。

[0398] 然后以95%的特异性和99%的特异性为每个未筛选的、对照组和尺寸选择的数据集生成分类。使用90至10个CCGA训练数据进行50轮分类,针对癌症数据集和非癌症数据集进行平衡,并使用针对多个预定基因组读取数中复制次数畸变进行训练的癌症分类器,每个读取数代表人类基因组的预定部分,在健康受试者的基因组中变异性低,如上所述。然后,根据CCGA中每个受试者的已知状态,生成每组分类的灵敏度。这些分类的结果如图17A所示,每个分组(例如1702)左侧显示全深度(未筛选)数据集的结果,每个分组中间显示序列覆盖对照组数据集的结果(例如1704),以及每个分组右侧显示尺寸选择的数据集的结果(例如1706)。

[0399] 与具有相同序列覆盖范围的对照组数据集相比,尺寸选择的数据集始终表现更好(将每个分组的右图与图17A中每个分组的中间图进行比较)。这与以下事实相符:尺寸选择的数据集应包含比对照组数据集更多的来自癌症的cfDNA片段的序列读数。然而,值得注意的是,两种类型的尺寸选择的数据集的性能也优于相应的完整数据集,尽管序列覆盖率比对照数据集少5倍到50倍,且从癌症衍生的cfDNA片段中读取的序列更少(将每个分组的右图与图17A中每个分组的左图进行比较)。

[0400] 实施例7

[0401] 计算器尺寸选择后的癌症分类

[0402] 在CCGA研究中,重复实施例5中概述的分析,以计算器方式选择与所有癌症类型中100个核苷酸或更少的cfDNA相对应的序列读数,例如,不排除子宫癌、甲状腺癌、前列腺癌、黑色素瘤、肾癌和HR+I/II期乳腺癌。再次,作为一种对照组,对全部数据集进行二次采样以匹配尺寸选择数据集的序列覆盖率的序列覆盖率。使用90至10个CCGA训练数据进行了50轮分类,对癌症和非癌症数据集进行了平衡。如图17B所示,相对于完整数据集和序列覆盖匹配的对照组数据集,在所有癌症类型中进行的计算器的尺寸选择均提供了95%特异性和

99%特异性的分类灵敏度改善。表1列出了此分析的分类统计信息。

[0403] 表1。在对CCGA数据集进行计算器筛选以对代表长度为100个核苷酸或更少的长度的cfDNA片段的序列读数进行计算器筛选后的癌症分类统计量。

[0404]	尺寸选择的	子样本	全深度	子样本	尺寸选择的
	特异性…时的 灵敏度	95%		99%	
	p值	2e-6	0.13	3e-8	4e-3
	平均变化 (2.5%, 97.5% CI)	0.051 (0.032, 0.069)	0.013 (-0.004, 0.031)	0.052 (0.036, 0.067)	0.023 (0.007, 0.038)

[0405] 接下来,根据癌症的阶段对上述所有癌症生成的分类数据进行分析。如图17C (95%特异性)和图17D (99%特异性)所示,在95%和98%的特异性下,尺寸选择的数据提供了相对于序列覆盖匹配的所有癌症阶段的对照组数据集同等或更好的灵敏度。值得注意的是,对于所有处于两个特异性阶段的癌症,均以尺寸选择的数据提供了相对于全部数据集相同或更好的灵敏度,但以95%特异性确定的1期癌症除外。此类分析的分类统计数据如图17E所示。

[0406] 接下来,对更可能以游离细胞核酸形式进入血液的癌症(例如不包括子宫癌、甲状腺癌、前列腺癌、黑素瘤、肾癌和HR+乳腺癌)生成的分类数据进行癌症分期分析。如图17F (95%特异性)和图17G (99%特异性)所示,在95%和98%的特异性下,尺寸选择的数据所提供的灵敏度均优于序列覆盖率匹配的对照组数据集。值得注意的是,对于所有处于两个特异性阶段的癌症,均以尺寸选择的数据提供了相同或优于全部数据集的敏感性,但以95%特异性确定的1期癌症除外。此分析的分类统计数据如图17H所示。

[0407] 实施例8

[0408] 体外尺寸选择

[0409] 接下来确定测序前DNA片段的体外的尺寸选择是否可以为测序后的计算机筛选的可行替代方案。简言之,从健康受试者获得的cfDNA样本中制备cfDNA库,其中肿瘤来源的cfDNA滴定为如上文参考图13所述。然后,使用PipenPrep® 洋菜凝胶电泳仪 (Sage Science)对cfDNA库中的DNA片段进行尺寸选择,将碱基对目标值设置为0至100+x个核苷酸,其中x是在所述库准备过程中添加到cfDNA片段的核苷酸数量。然后通过WGS对尺寸选择的片段进行测序。如图18所示,所产生的序列读数显示出适当的尺寸选择,并在约100个核苷酸处出现了截断。此外,当分析序列读数时,观察到肿瘤来源的cfDNA的序列读数的富集,类似于在计算机下的尺寸选择结果。此外,其他测序指标,如重复率和总读数失去量,与未经体外尺寸选择观察到的指标相似。

[0410] 实施例9

[0411] 体外尺寸选择后肿瘤分数的确定

[0412] 下一个问题是,在测序之前,DNA片段的体外的尺寸选择是否会提供与实施例5和实施例7中提供的序列读数的计算机下的尺寸选择相似的分类灵敏度改进。简言之,从收集的不同癌症和癌症阶段的CCGA样本中选择了65个癌症样本和29个非癌症样本,如表2和表3

所示。38个被选中的癌症被证明更多地流入血液。

[0413] 表2。用于体外尺寸选择研究的样本的癌症阶段。

[0414]	阶段	计数
	I	13
	II	15
	III	21
	IV	12
	非信息性	4

[0415] 表3。用于体外尺寸选择研究的样本的癌症类型。

[0416]	类型	计数	类型	计数
	乳癌	18	肝胆管癌	3
	结直肠癌	13	淋巴瘤	3
	肺癌	6	胰腺癌	3
	肾癌	5	子宫颈癌	2
	头部/颈部癌	3	其他	9

[0417] 简言之,从所选样本制备cfDNA库,如上文参考图13所述。

[0418] 将含有UMI序列、启动子杂交位点等的接合子添加到片段中,使cfDNA片段的长度增加到大约170个核苷酸。然后根据标准方案使用Pipen Prep®仪器(Sage Science)进行体外尺寸选择,设置为选择200至310个核苷酸或200至320个核苷酸的尺寸范围,代表30至140个核苷酸或30至150个核苷酸的cfDNA片段连接到3'和5'接合子上,总计170个核苷酸。然后,针对尺寸选择的库生成序列读数,如上文参考图13所述。

[0419] 然后使用癌症衍生的变体检测来估计从癌症衍生的cfDNA片段的序列读数的分数。如上所述,对65名癌症受试者中的每一个的cfDNA样本进行测序反应,无论是否在30至140个核苷酸或30至150个核苷酸之间进行体外尺寸选择。还针对来自35位癌症受试者的肿瘤匹配样品的基因组DNA制备物进行了测序反应。然后,通过比较在肿瘤匹配的样本中鉴定出的癌症衍生的变异等位基因与从尺寸选择的样本中或计算器仿真的尺寸选择后获得的游离细胞DNA的序列读数,来估计癌症衍生的序列读数的比例。如图19所示,在30至40个核苷酸(橙色)或30至150个核苷酸(蓝色)范围内的体外尺寸选择几乎增加了每个样本的估计肿瘤分数。在某些情况下,这种增加导致肿瘤分数超过了整个基因组测序分类器的检测水平,估计约为0.5%,在图19中用虚线表示。

[0420] 下列表4和表5分别显示了按原肿瘤分数和癌症阶段分组的样本的倍数改善情况。

[0421] 表4。体外尺寸选择后,按样本原始肿瘤分数分组的肿瘤分数的中位数改善。

[0422]	原始肿瘤分数	中位数改善
	≤ 0.005	2.24
	≤ 0.01	2.38
	≤ 0.05	2.00
	≤ 0.1	1.78
	≤ 0.2	1.45

≤ 0.5	1.41
------------	------

[0423] 表5.在体外尺寸选择后,按癌症阶段分组的肿瘤分数的中位数改善。

阶段	中位数改善
I	1.89
II	2.12
III	2.24
IV	1.58
NI	2.24

[0425] 实施例10

[0426] 体外尺寸选择后分类灵敏度的提高

[0427] 从所有65个癌症样本中进行体外尺寸选择后,在全cfDNA样本和cfDNA样本上生成的序列读数的数据集被输入到一个针对多个预定基因组读取数的复制次数畸变而训练的癌症分类器中,如上文所述,每个代表人类基因组的预定义部分,在健康受试者的基因组中具有低变异性,从而为每个数据集生成一个“B-分数”。然后将得出的分类分数(癌症=橙色;非癌症=蓝色)相对于样本的原始肿瘤分数制图,如图20所示。根据分类分数确定95% (2156)、98% (2154) 和99% (2152) 的特异性,并以这些分类分数将样本的肿瘤分数确定为LOD水平。然后,根据确定的LOD水平,在每种特异性下,估算使用完整数据集或体外尺寸选择数据集的分类灵敏度。如表6所示,在所有三种特异性下,cfDNA片段的体外尺寸选择将测定的灵敏度提高了约20%至30%。

[0428] 表6.体外尺寸选择后的分类灵敏度。

特异性	(p- δ , p+ δ) 中的肿瘤分 数中位数	原始灵敏度	体外灵敏度	比率
95%	0.008	0.46	0.56	1.22
98%	0.010	0.46	0.54	1.18
99%	0.014	0.39	0.51	1.30

[0431] 上述各种方法和技术提供了许多实施本发明的方法。当然,应当理解,根据本文描述的任何特定实施例,不一定可以实现所描述的所有目标或优点。因此,例如,本领域技术人员将认识到,可以以实现或优化本文所述的一个或一组优点的方式来执行所述方法,而不必实现本文中可能教授或建议的其他目标或优点。这里提到了各种有利和不利的实施方案。应理解的是,一些优选实施例具体包括一个、另一个或几个有利特征,而其他实施例具体排除一个、另一个或几个不利特征,而还有一些实施例通过包含一个、另一个或多个有利特征来具体减轻所面临的不利特征。

[0432] 此外,本领域技术人员将认识到来自不同实施例的各种特征的适用性。类似地,上

文讨论的各种组件、特征和步骤,以及每个此类组件、特征或步骤的其他已知等效物可以由本领域的普通技术人员混合和搭配,以执行根据本文所述原理的方法。在各种组件、特征和步骤中,一些将被具体地包括在不同的实施例中,而另一些将被具体地排除在不同的实施例中。

[0433] 虽然本发明已经在某些实施例和例示的上下文中公开,但是本领域的技术人员将理解,本发明的实施例扩展到其他替代实施例及/或其用途、修改和等效物。

[0434] 在本发明的实施例中已经公开了许多变化和替代组件。此外,对于本领域技术人员而言,其他变化和替代组件将是显而易见的。

[0435] 本文所公开的本发明的替代组件或实施例的分组不应被解释为限制。每个分组的组件可以单独地或以与组的其他组件或本文中找到其他组件的任何组合来引用和声明。出于方便及/或专利性的原因,可以将一个组中的一个或多个组件包括在组中或从组中删除。

[0436] 当发生任何此类涵盖或删除时,在此说明书被视为包含经修改的组,从而实现了所附权利要求中使用的所有马库西式的书面描述。

[0437] 最后,应理解的是,本文公开的本发明的实施例是对于本发明原理的说明。可用的其他修改可以包含在本发明的范围内。因此,仅是例示性但不限于,可以根据本文的教导利用本发明的替代配置。因此,本发明的实施例不限于所示和描述的精确实施例。

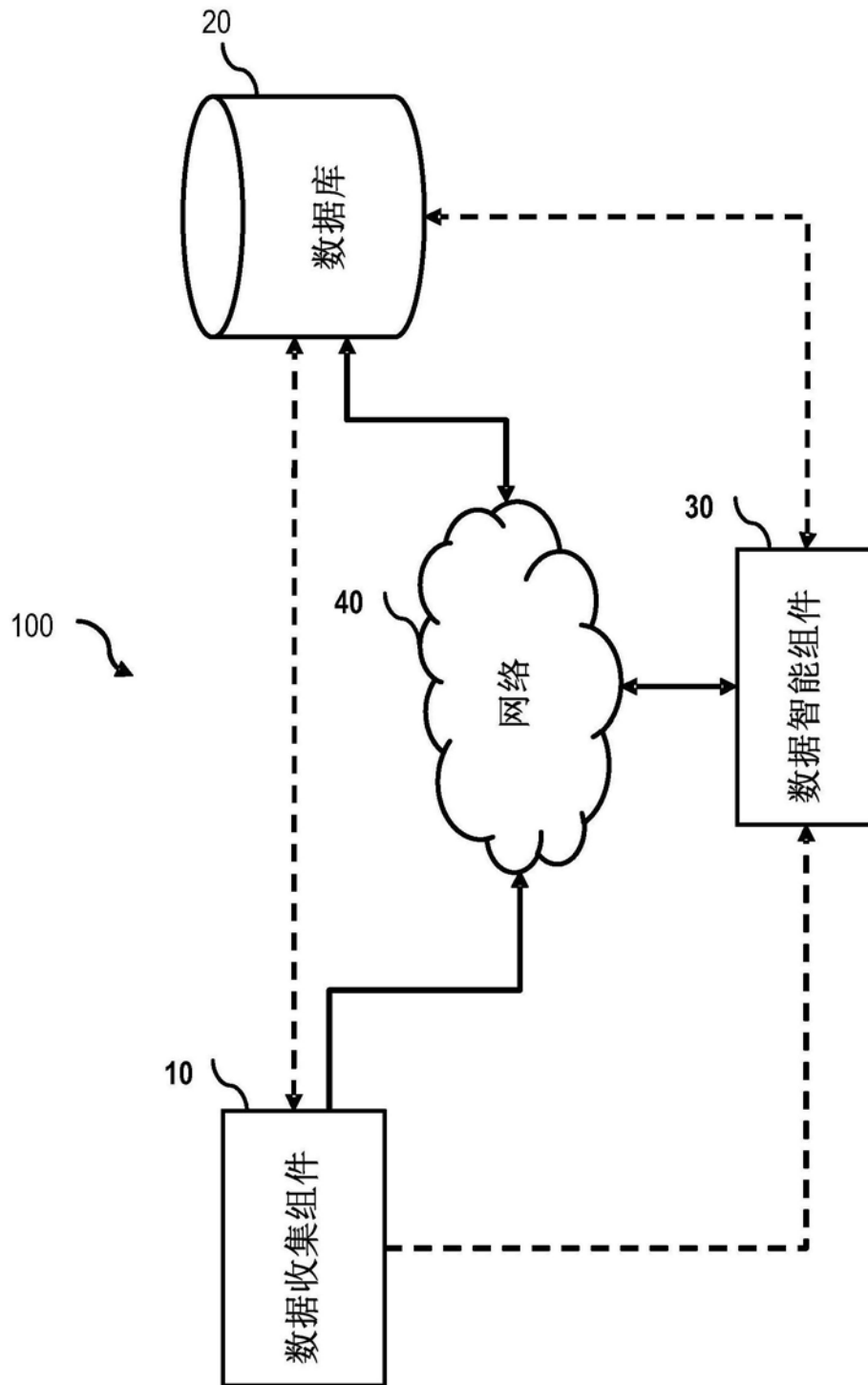


图1A

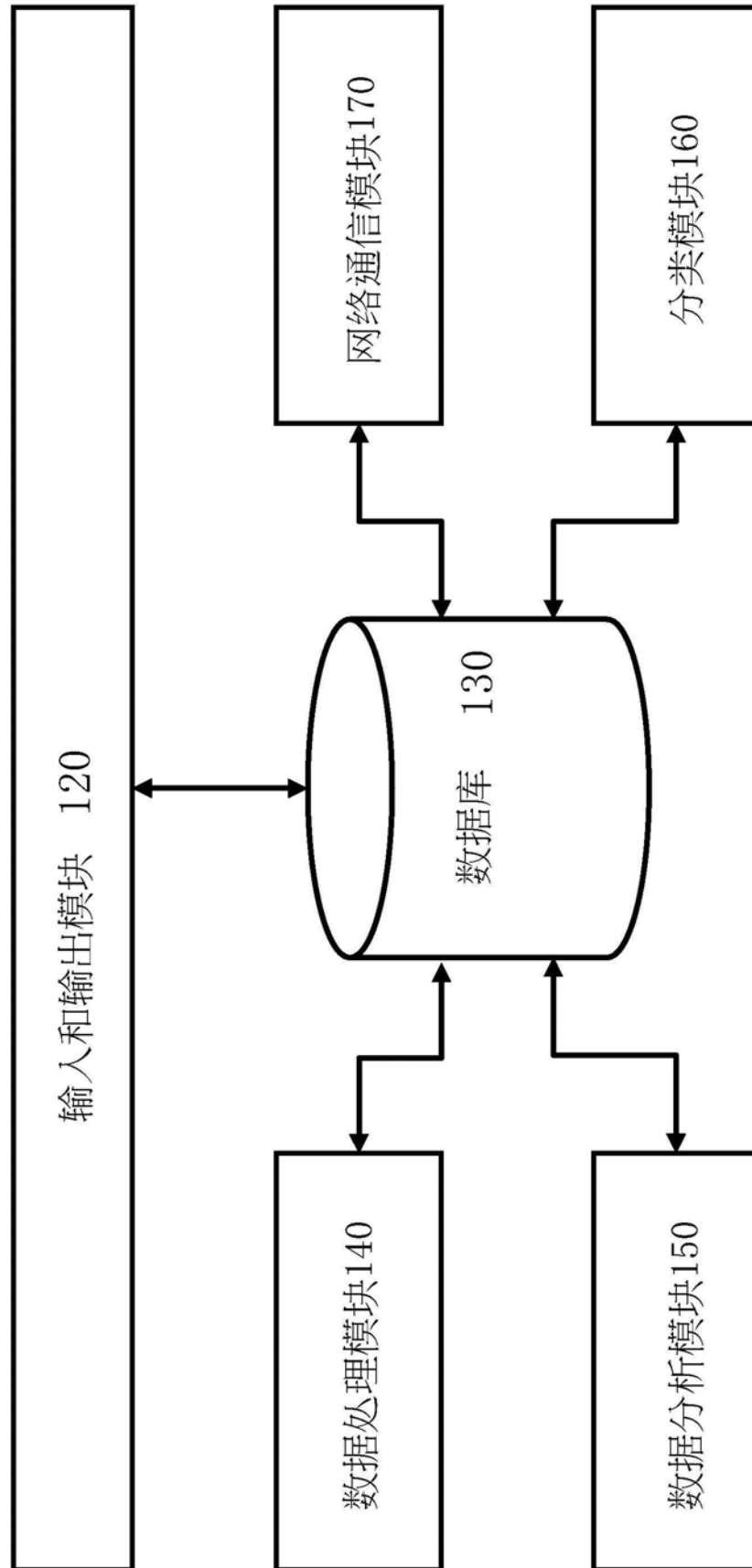


图1B

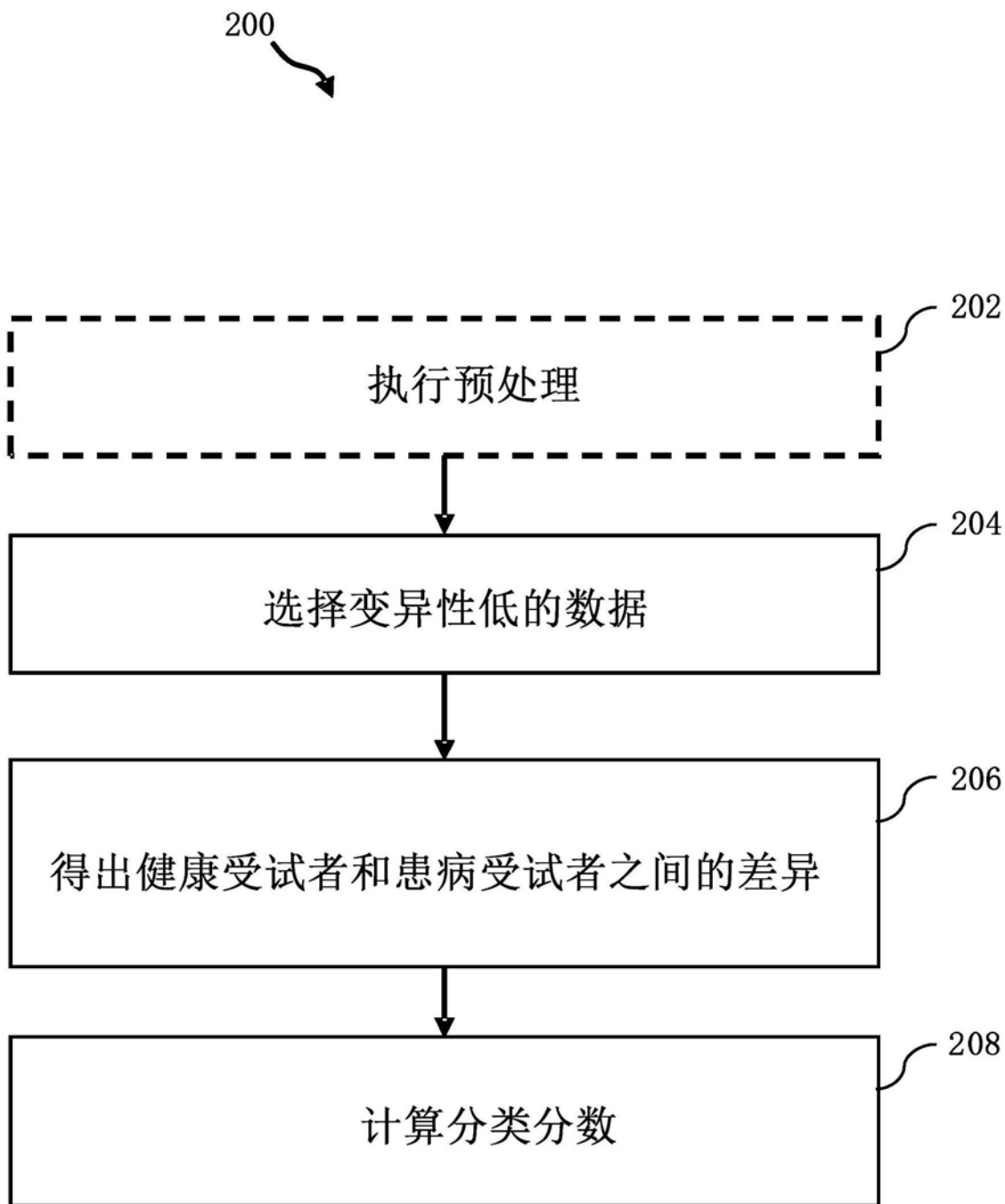


图2A

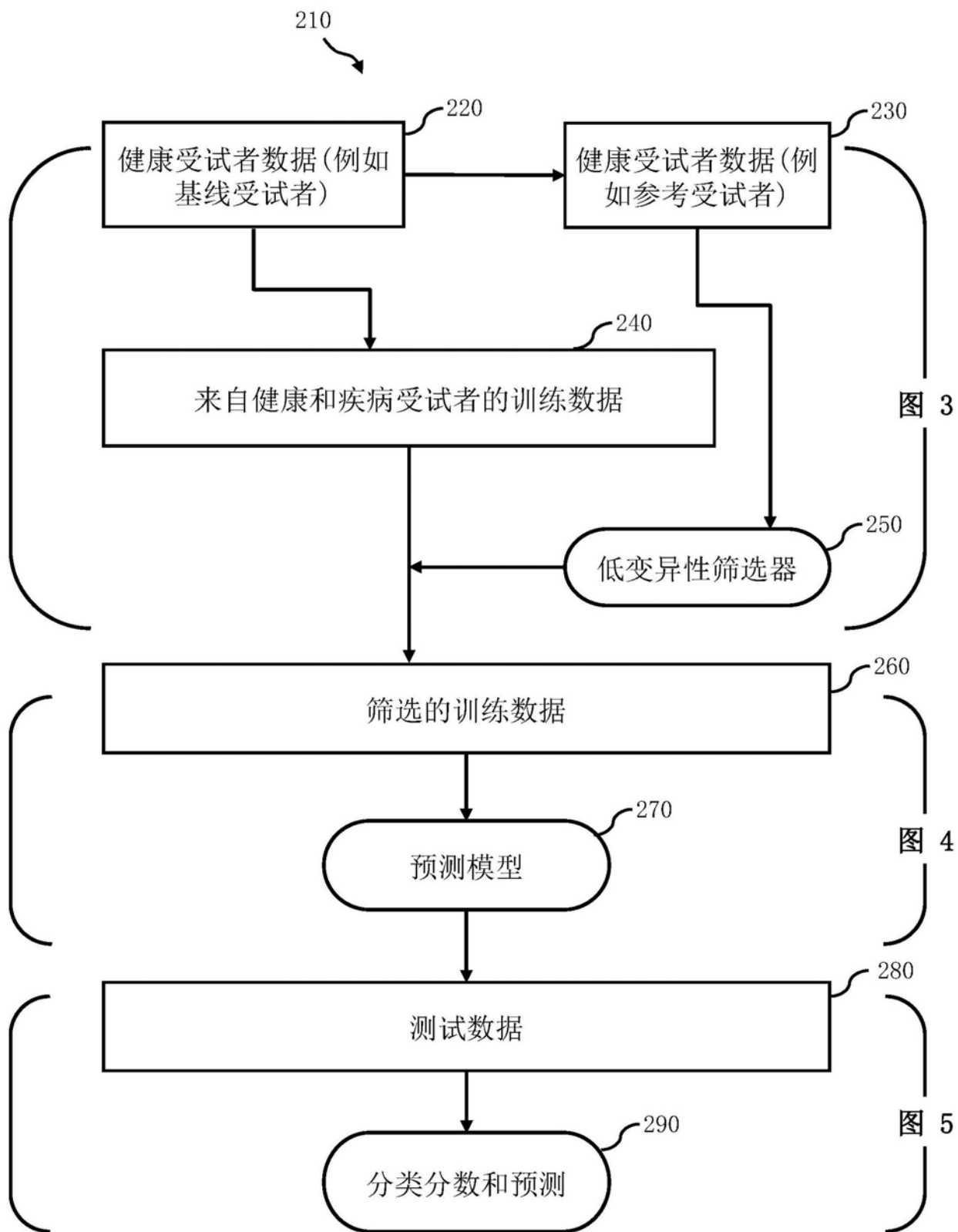


图2B

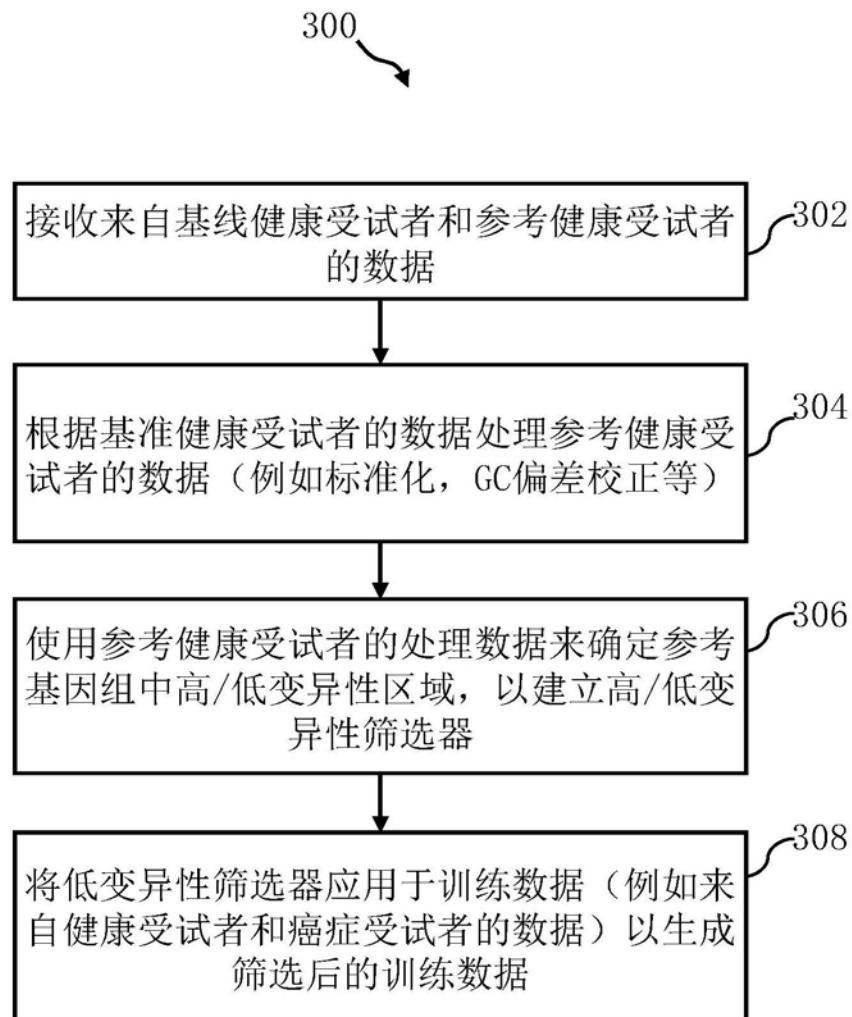


图3A

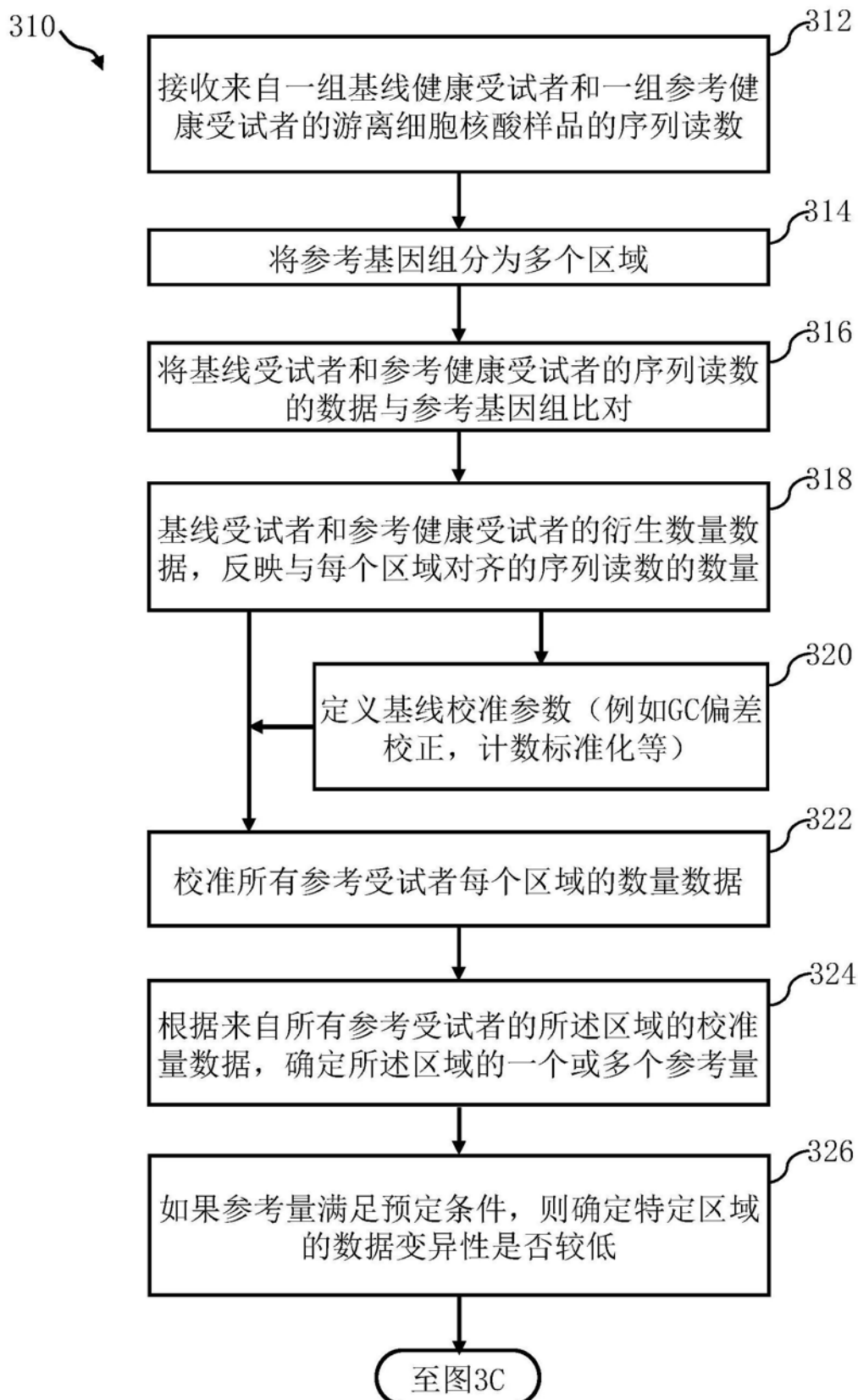


图3B

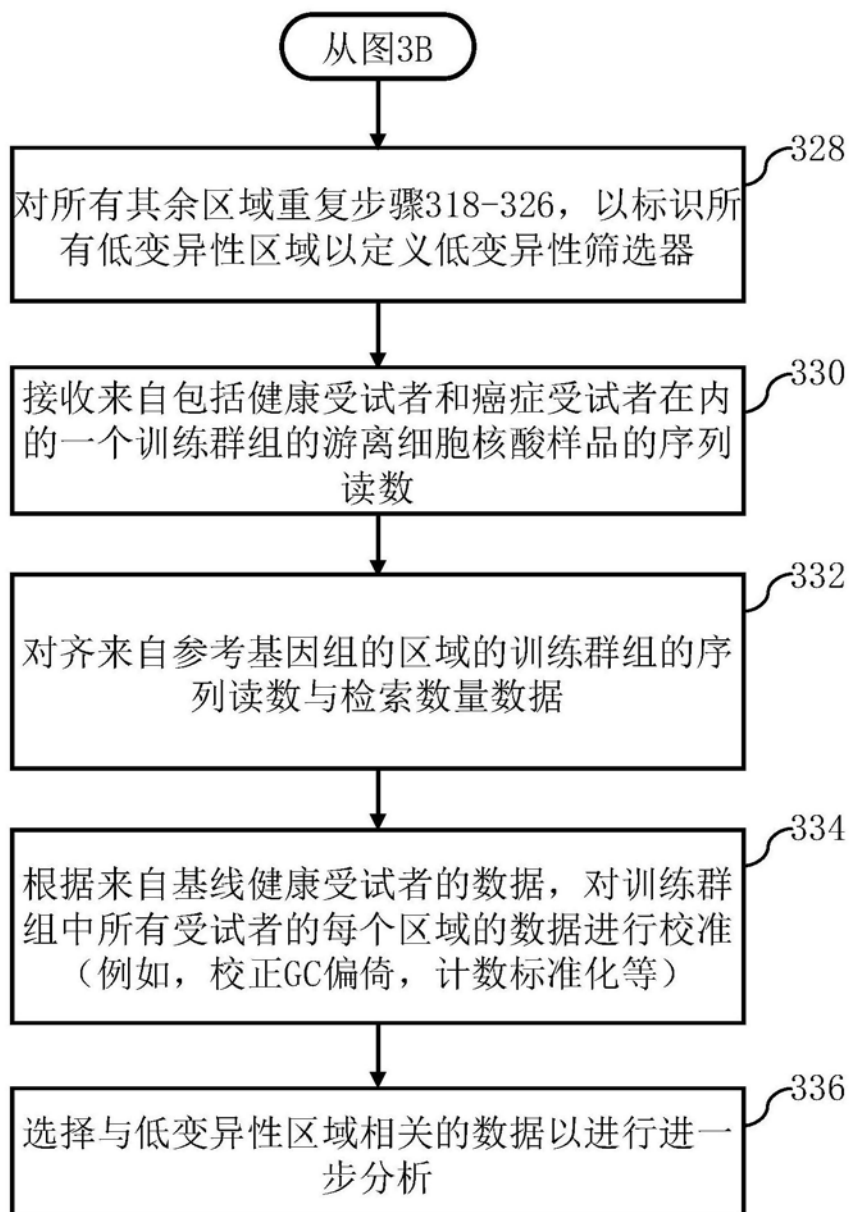


图3C

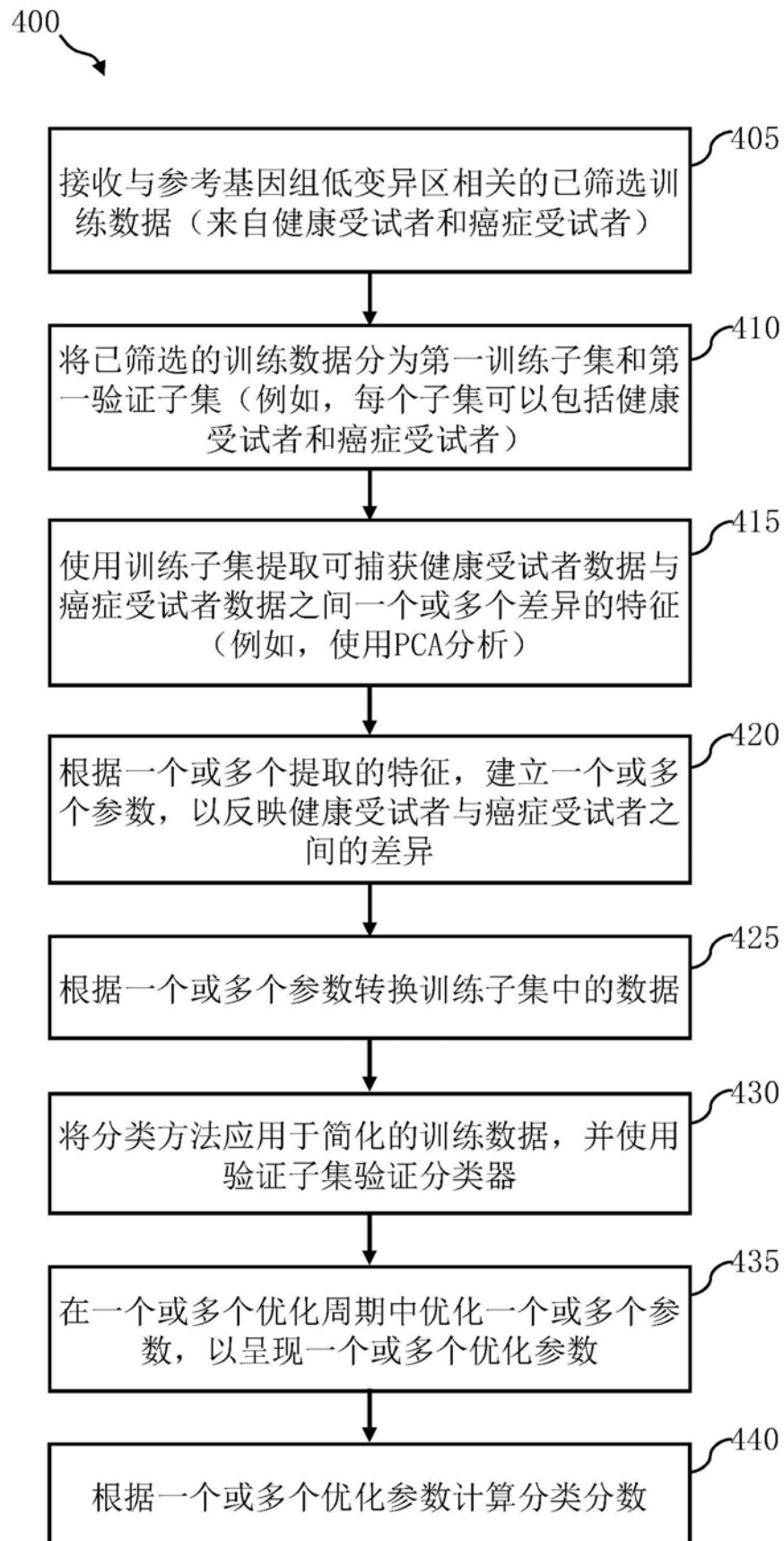


图4

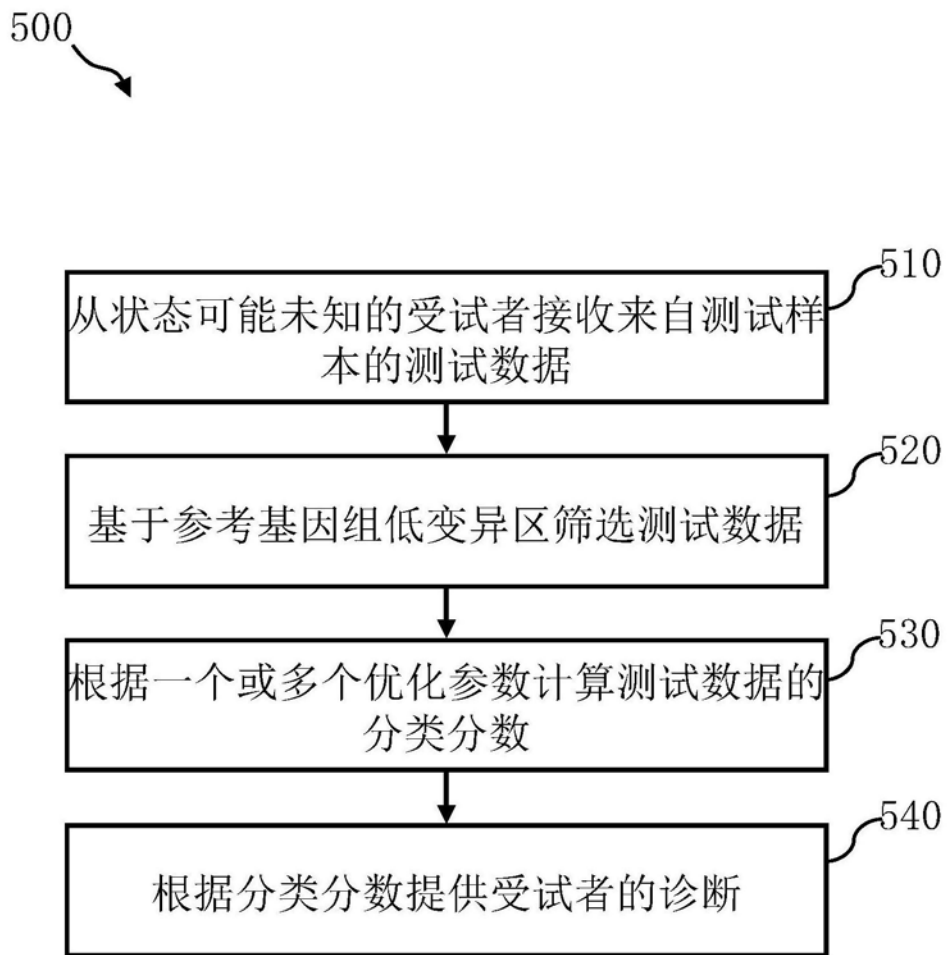


图5

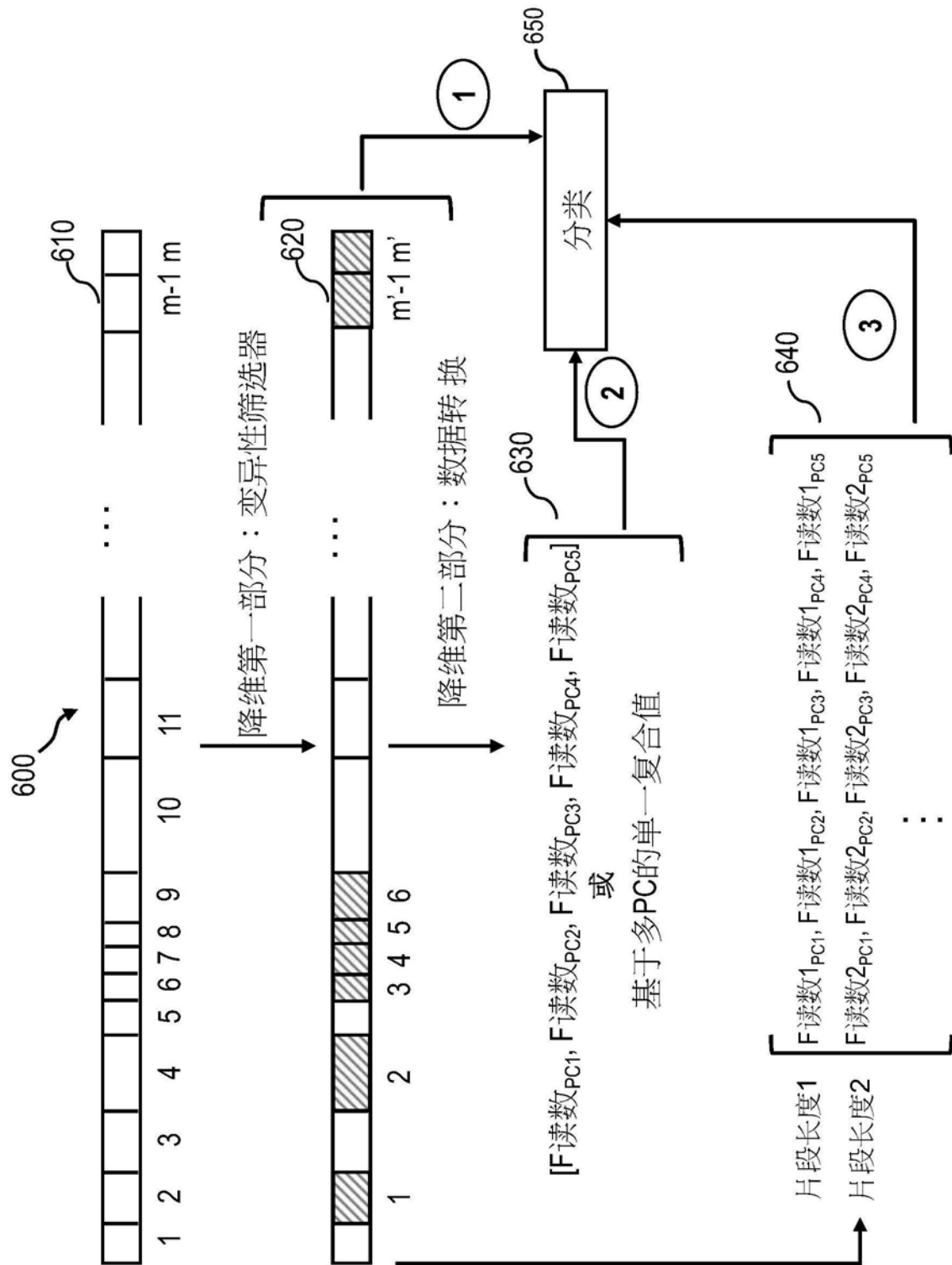


图6

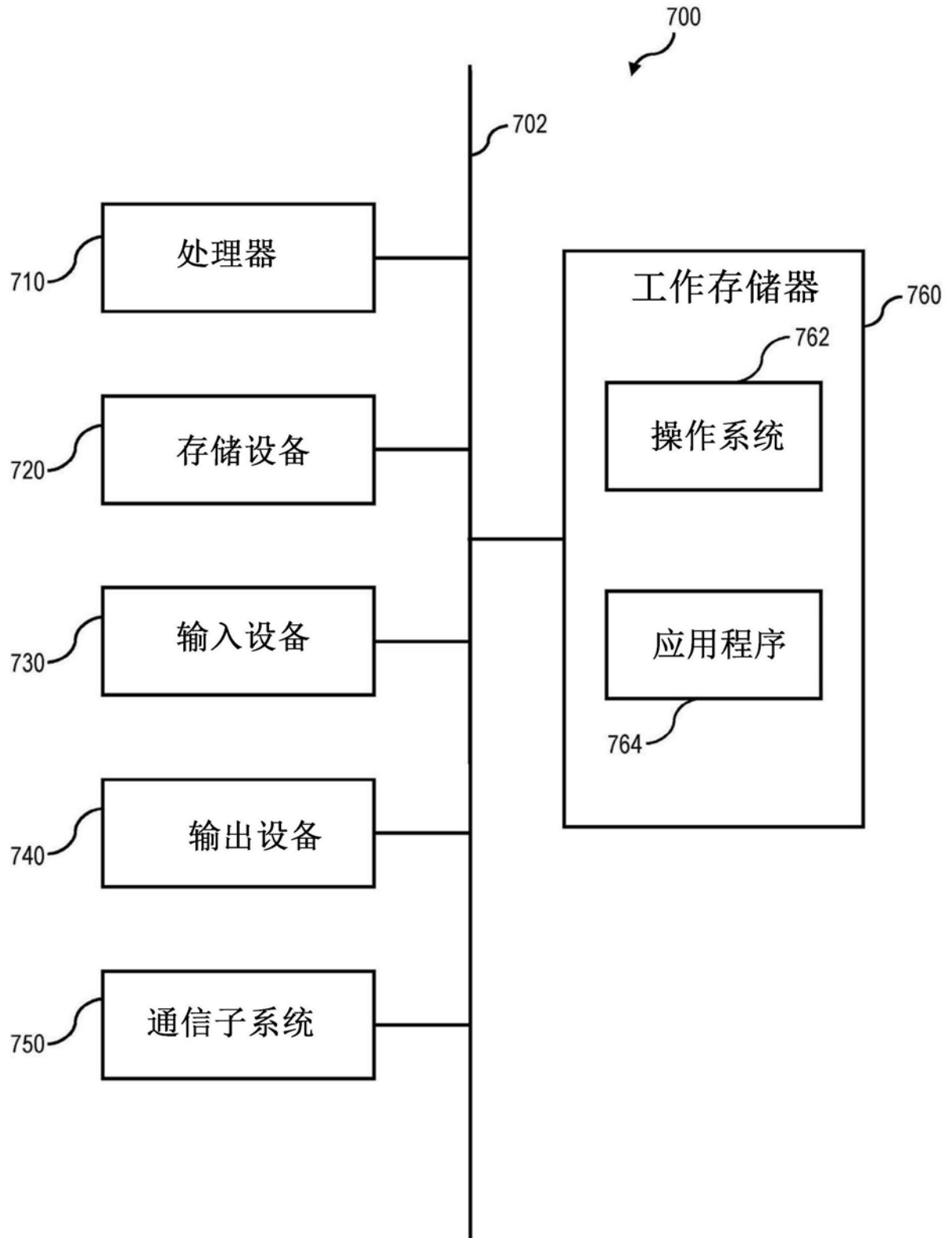
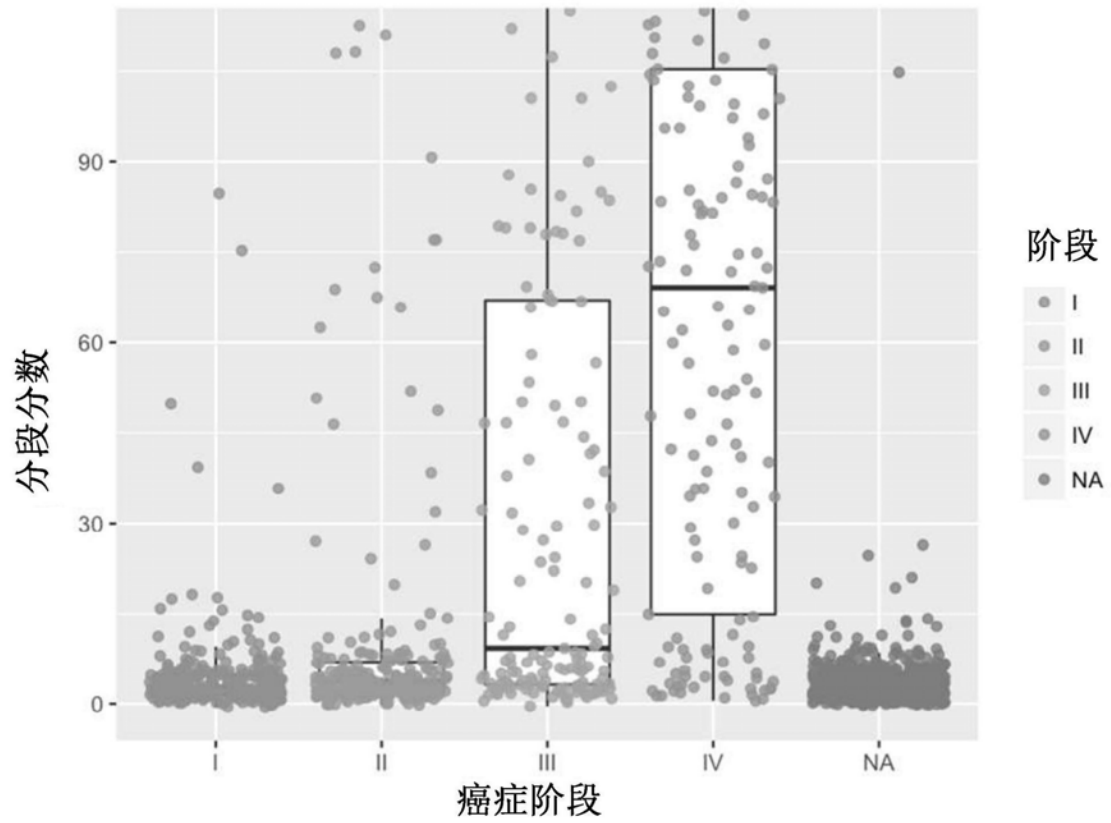


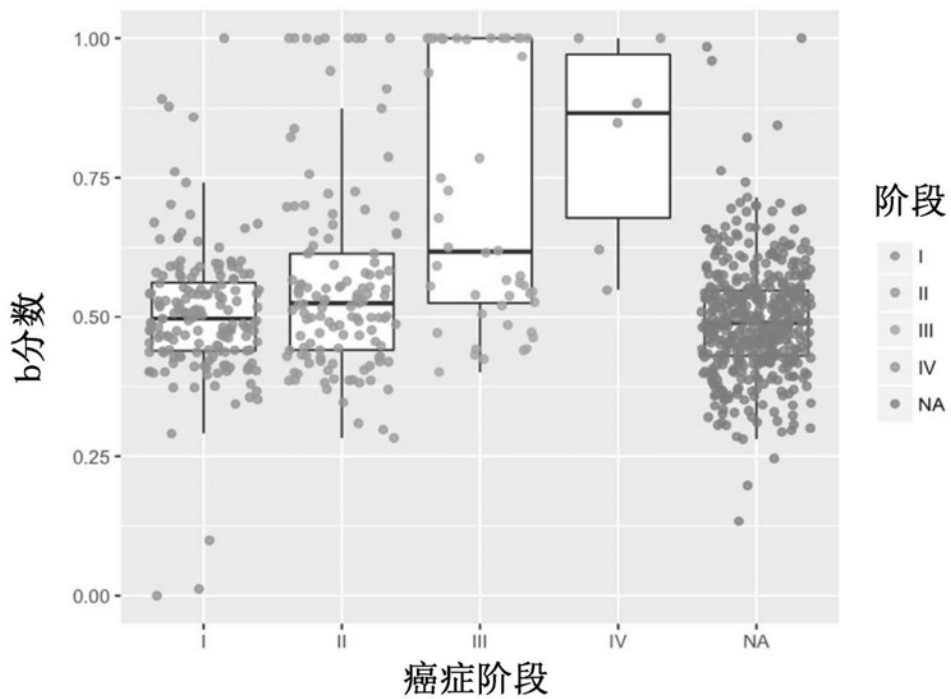
图7

类型	乳癌				全部
阶段	I	II	III	IV	N/A
样本编号	86	82	28	5	359
<u>z-分数</u>					
AUC	0.45	0.56	0.71	0.94	0.60
TPR @ 5% FPR	0.04	0.16	0.47	0.80	0.21
FPR @ 50% TPR	0.60	0.48	0.19	0.01	0.37
<u>b-分数</u>					
AUC	0.57	0.61	0.80	0.99	0.68
TPR @ 5% FPR	0.03	0.19	0.52	1.00	0.25
FPR @ 50% TPR	0.37	0.28	0.03	0.01	0.25

图8



SCNA分段分数/ z分数与乳腺癌阶段



浸润性乳腺癌的b分数表现

图9

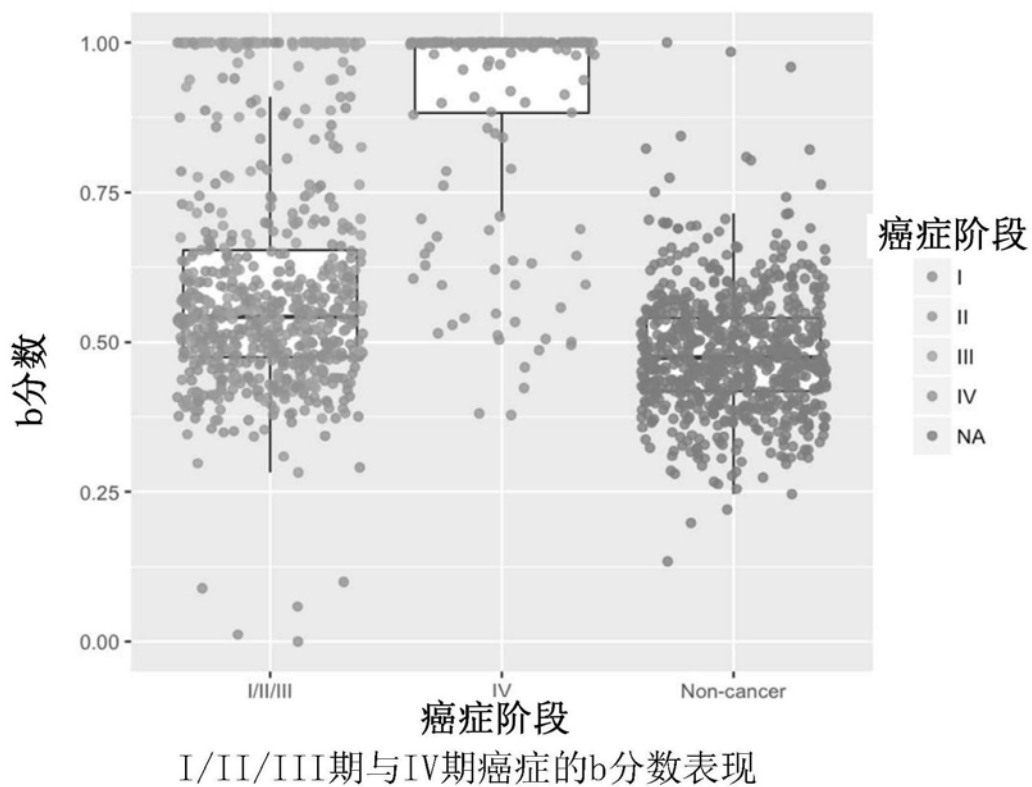
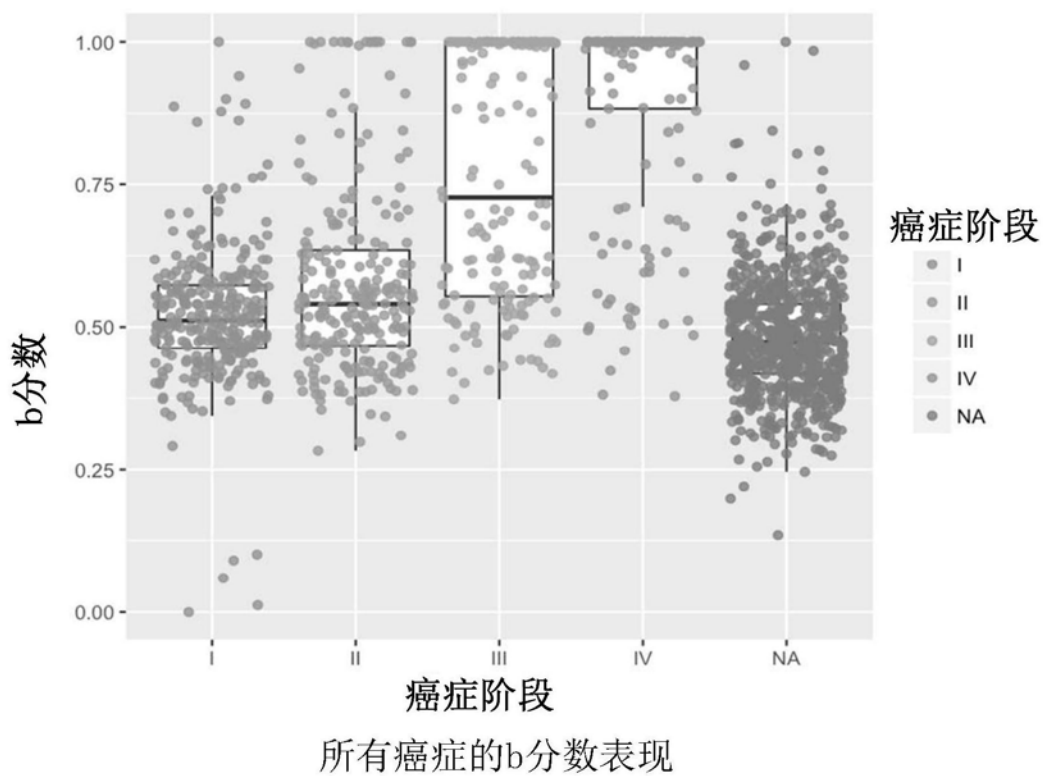


图10

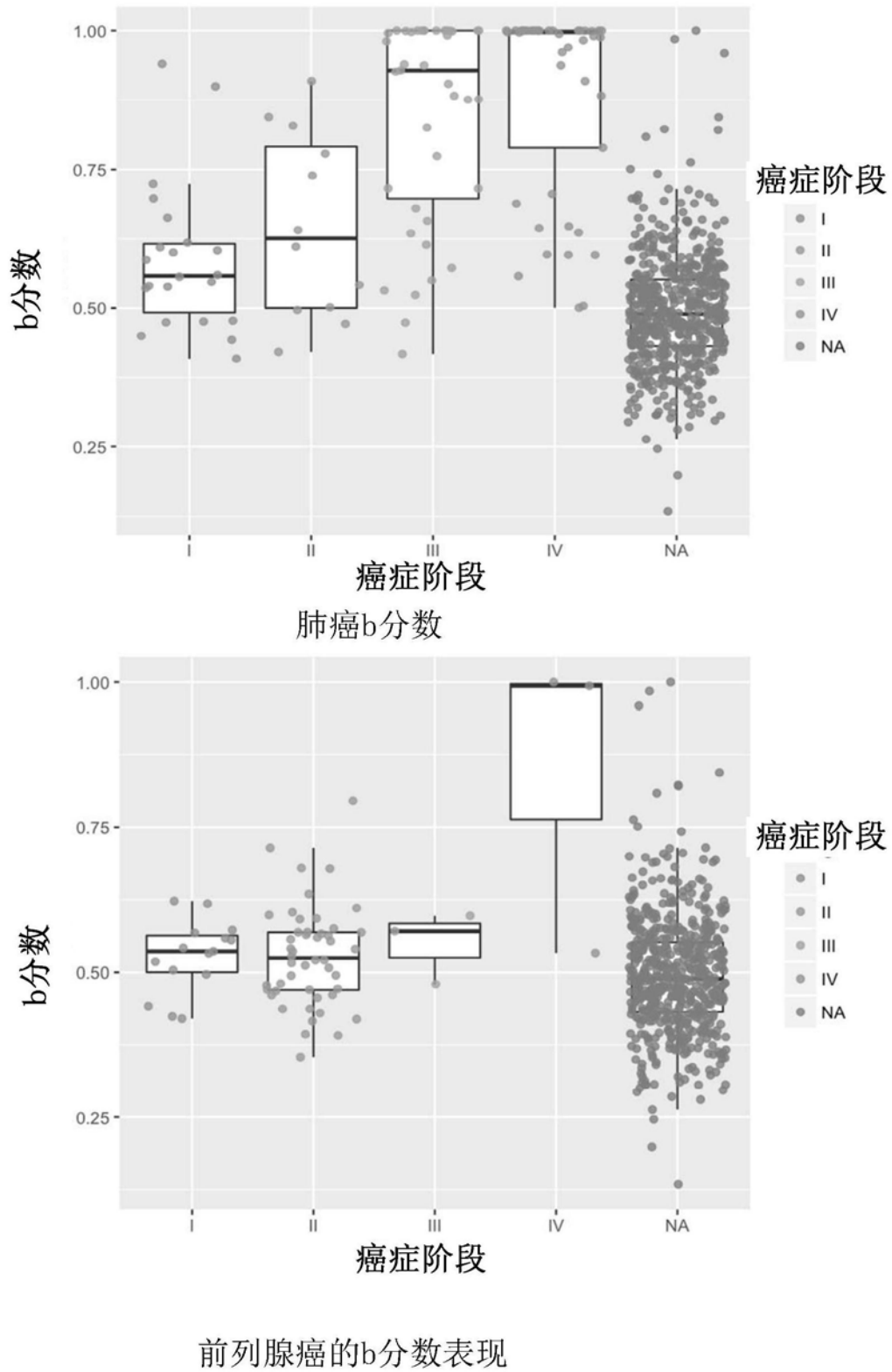


图11A

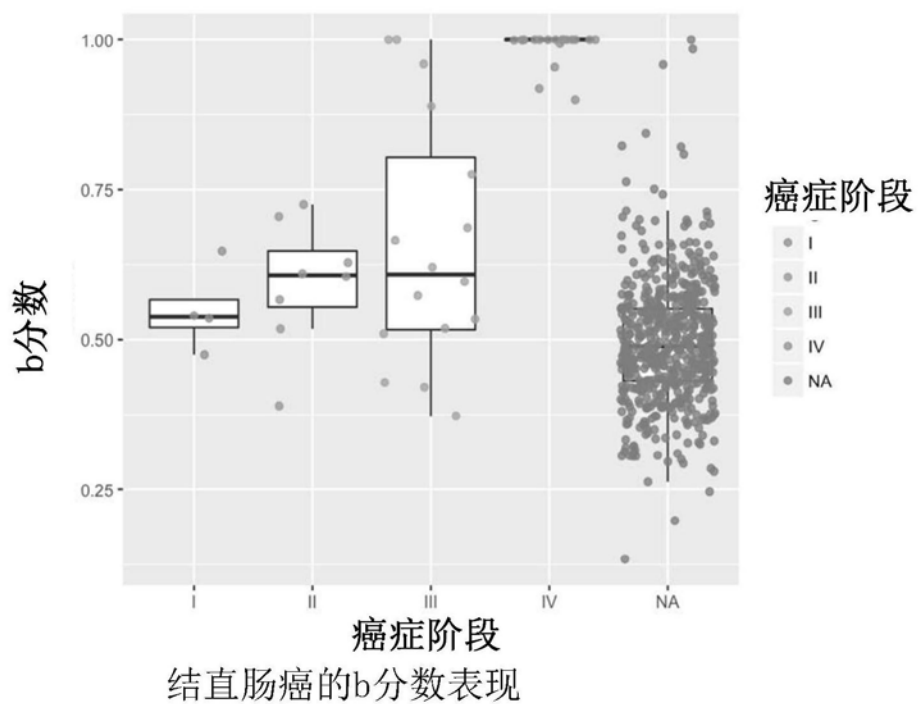


图11B

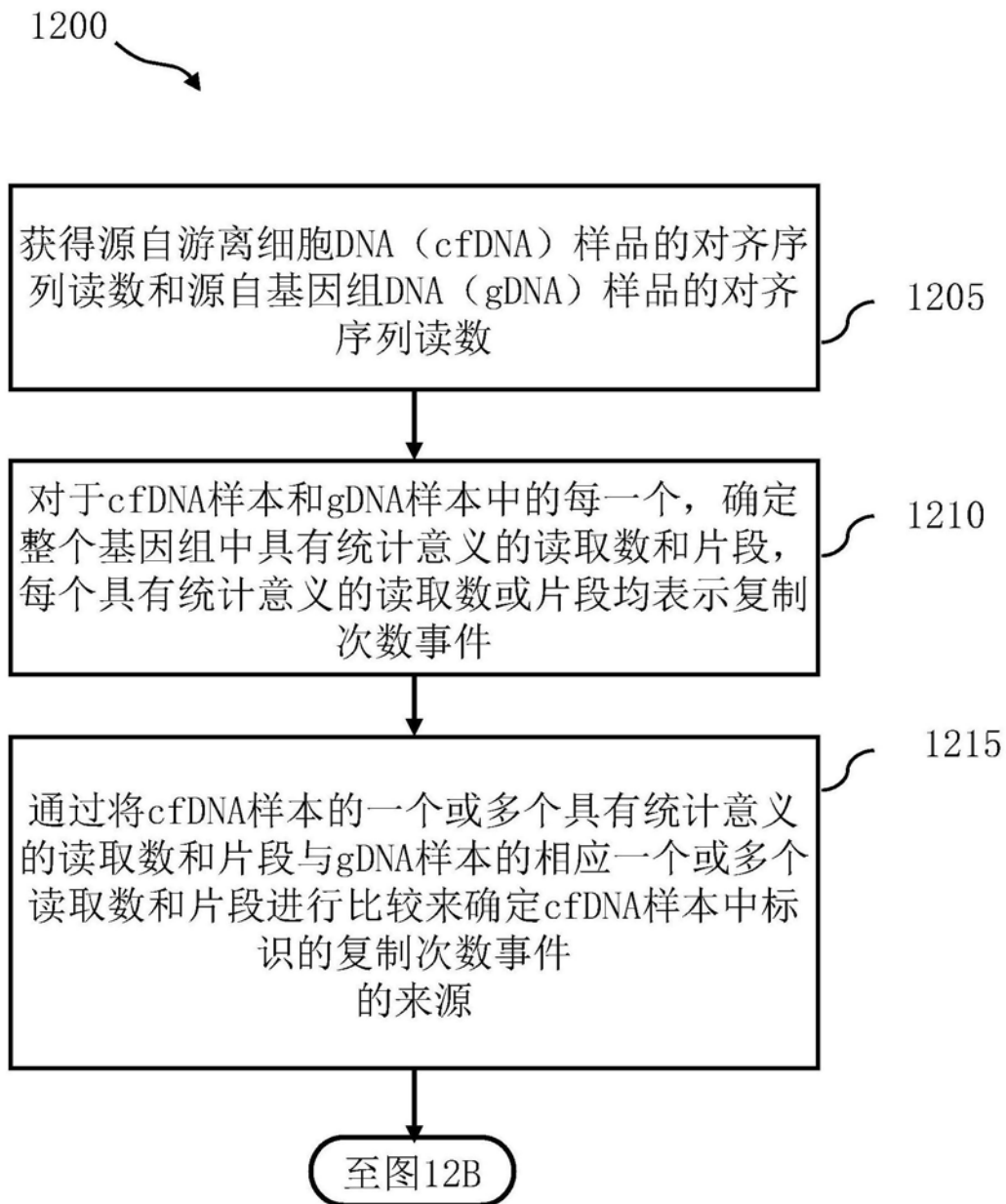


图12A

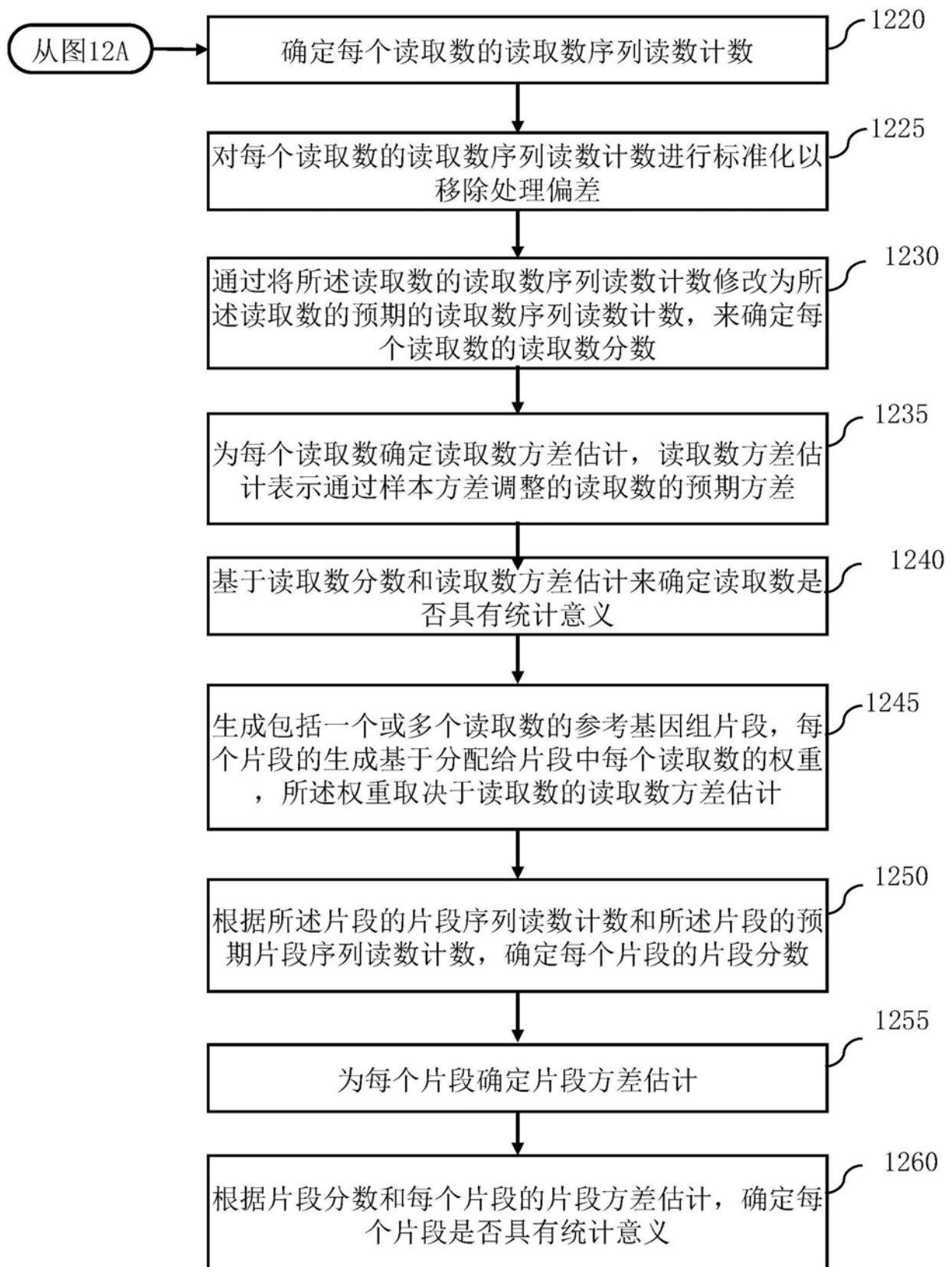


图12B

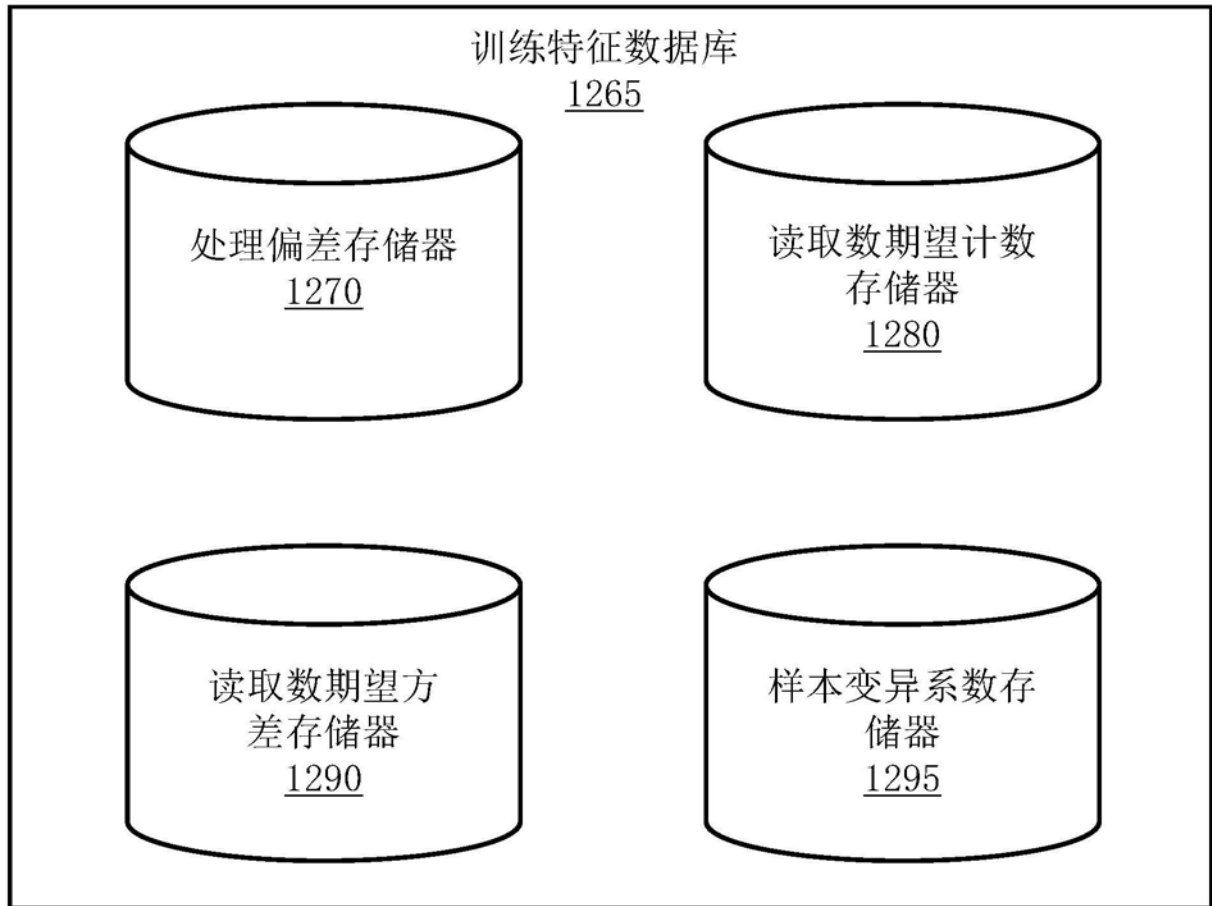


图12C

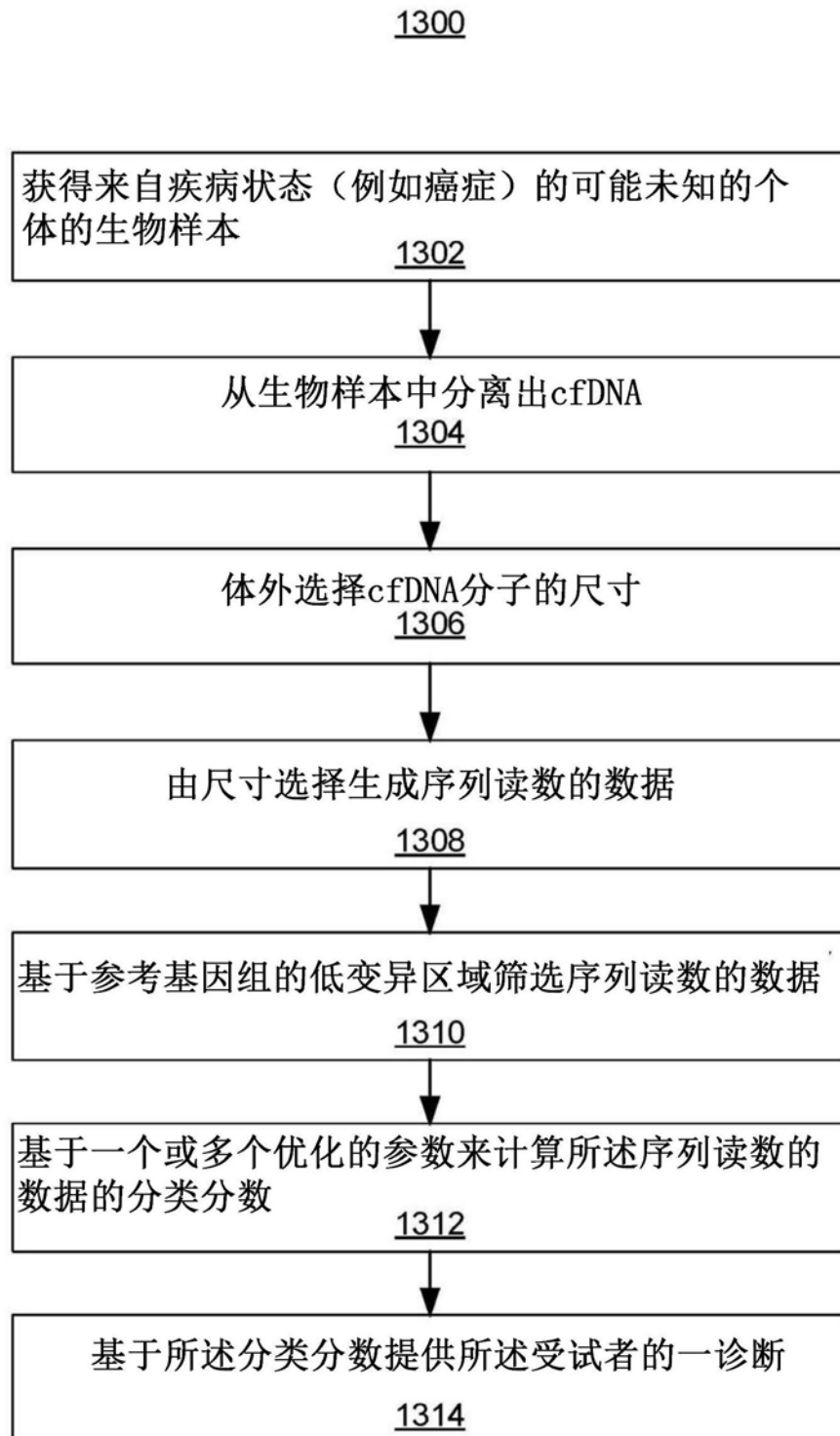


图13

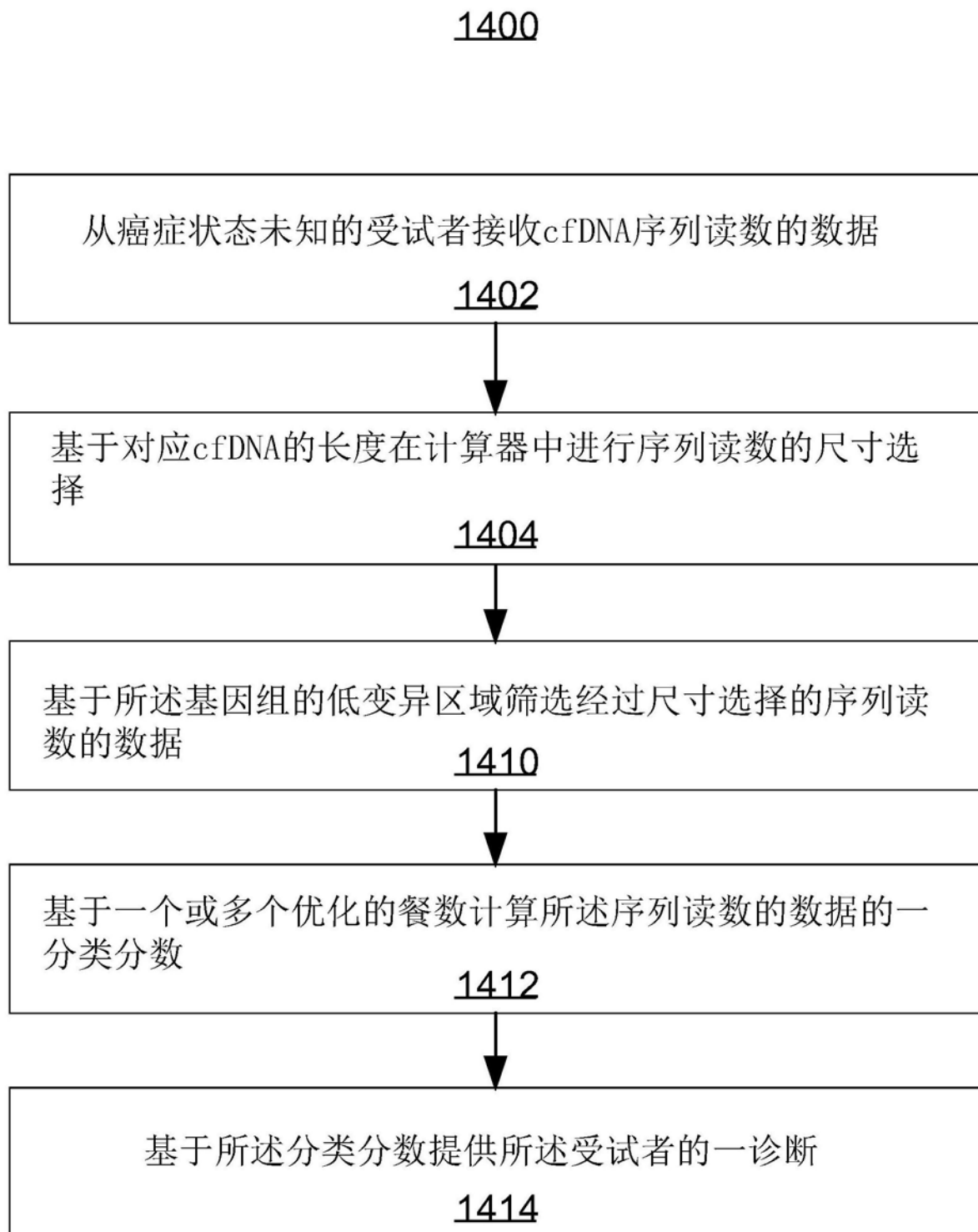


图14

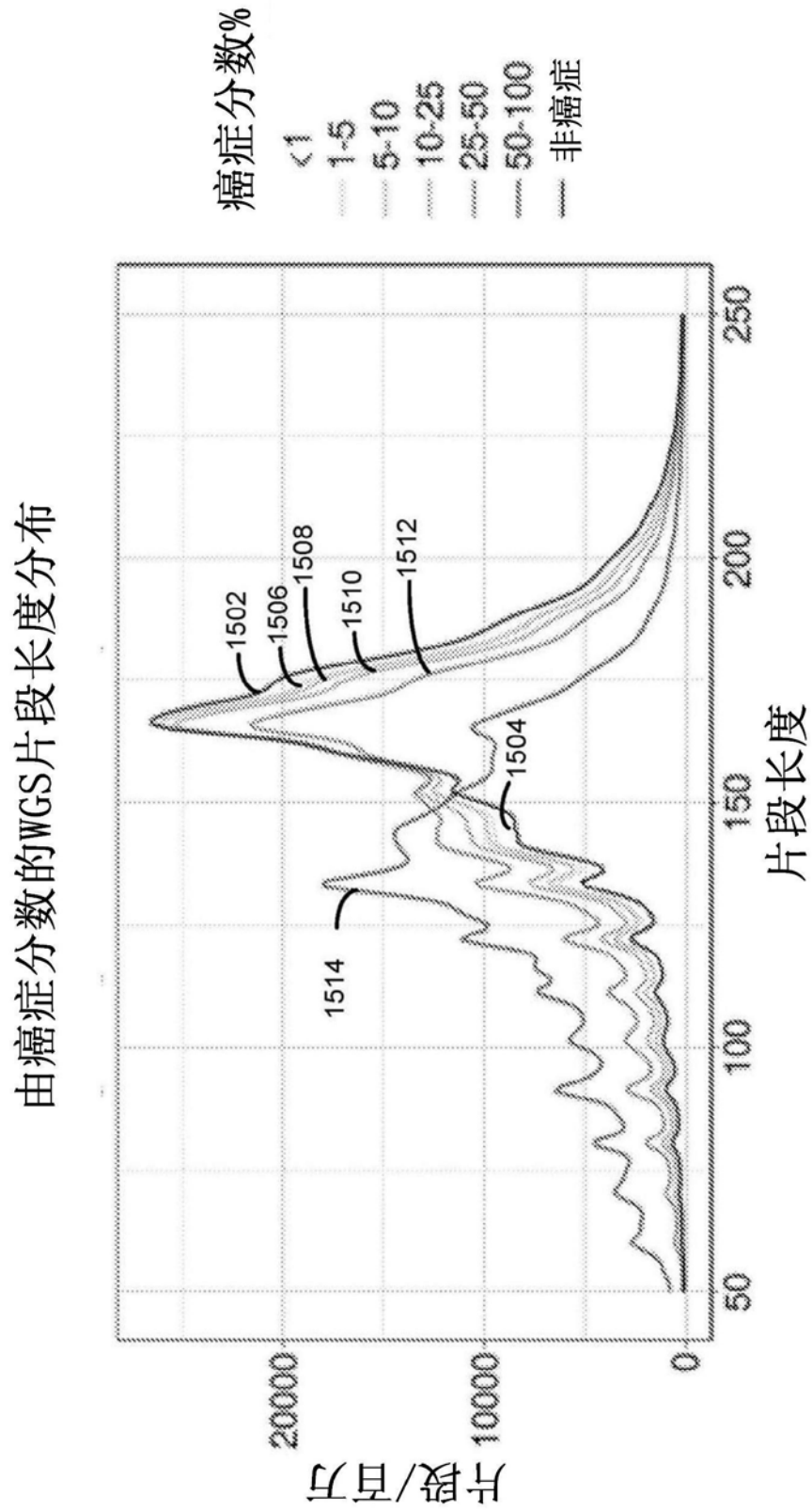


图15

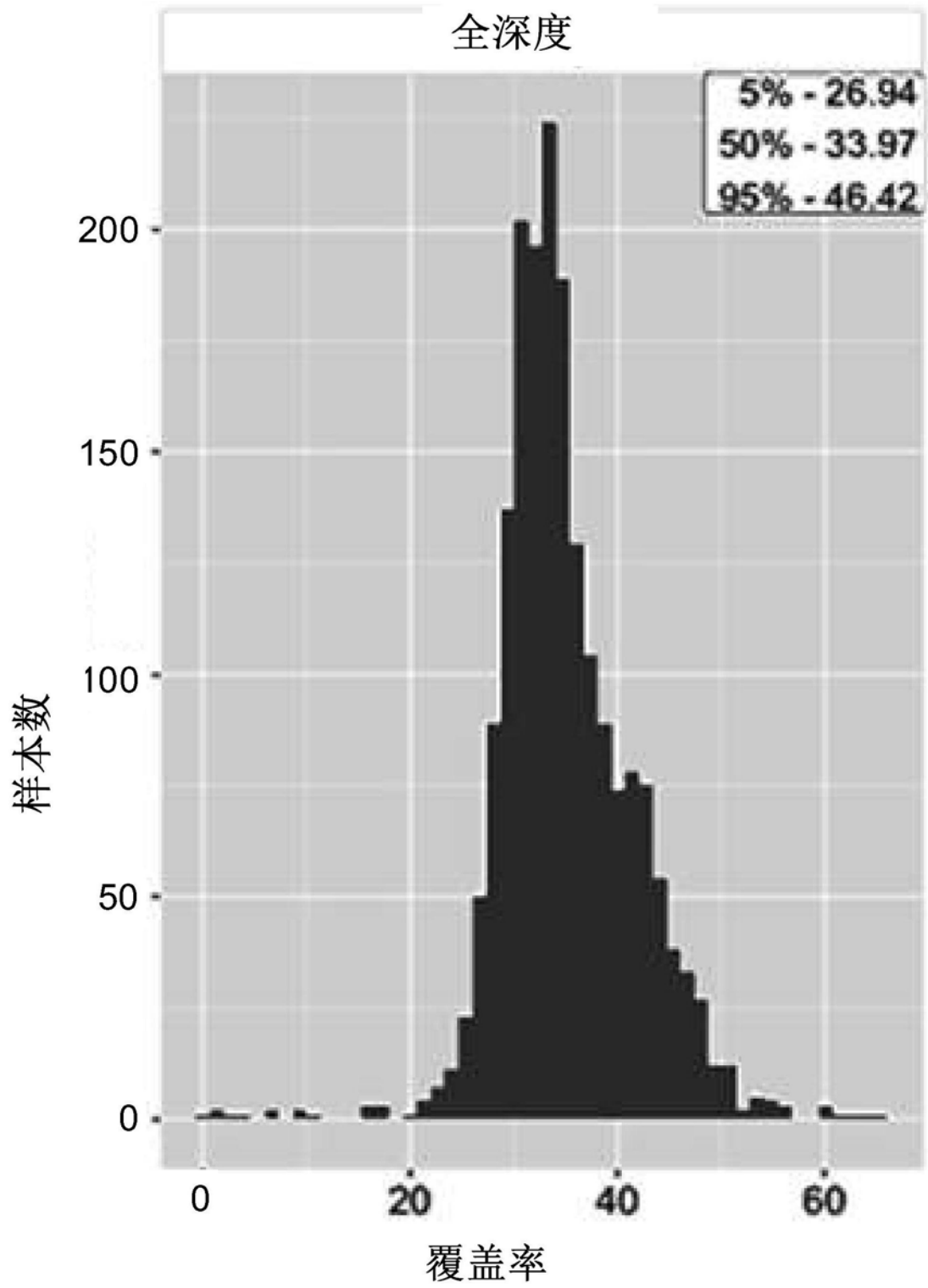


图16A

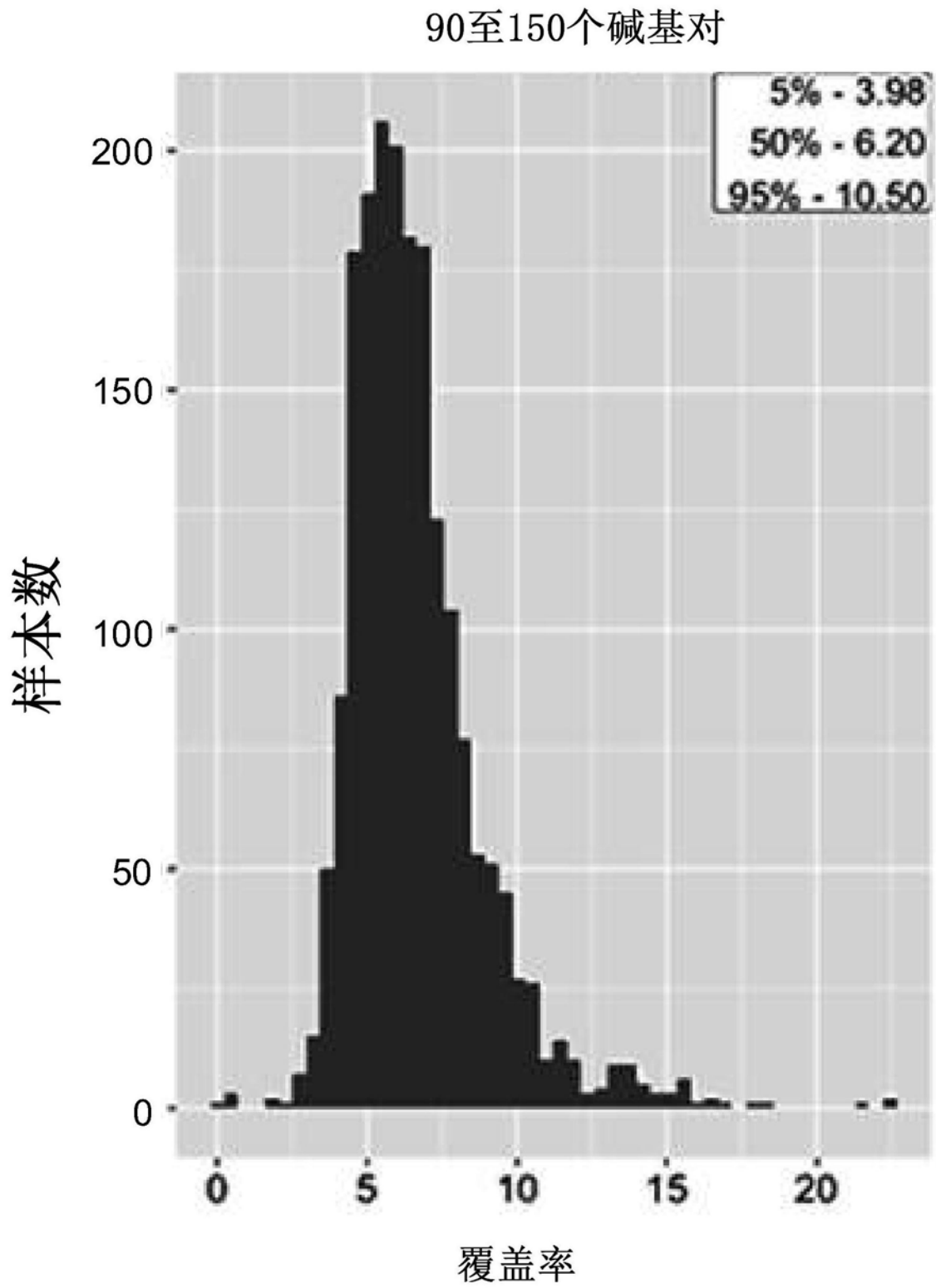


图16B

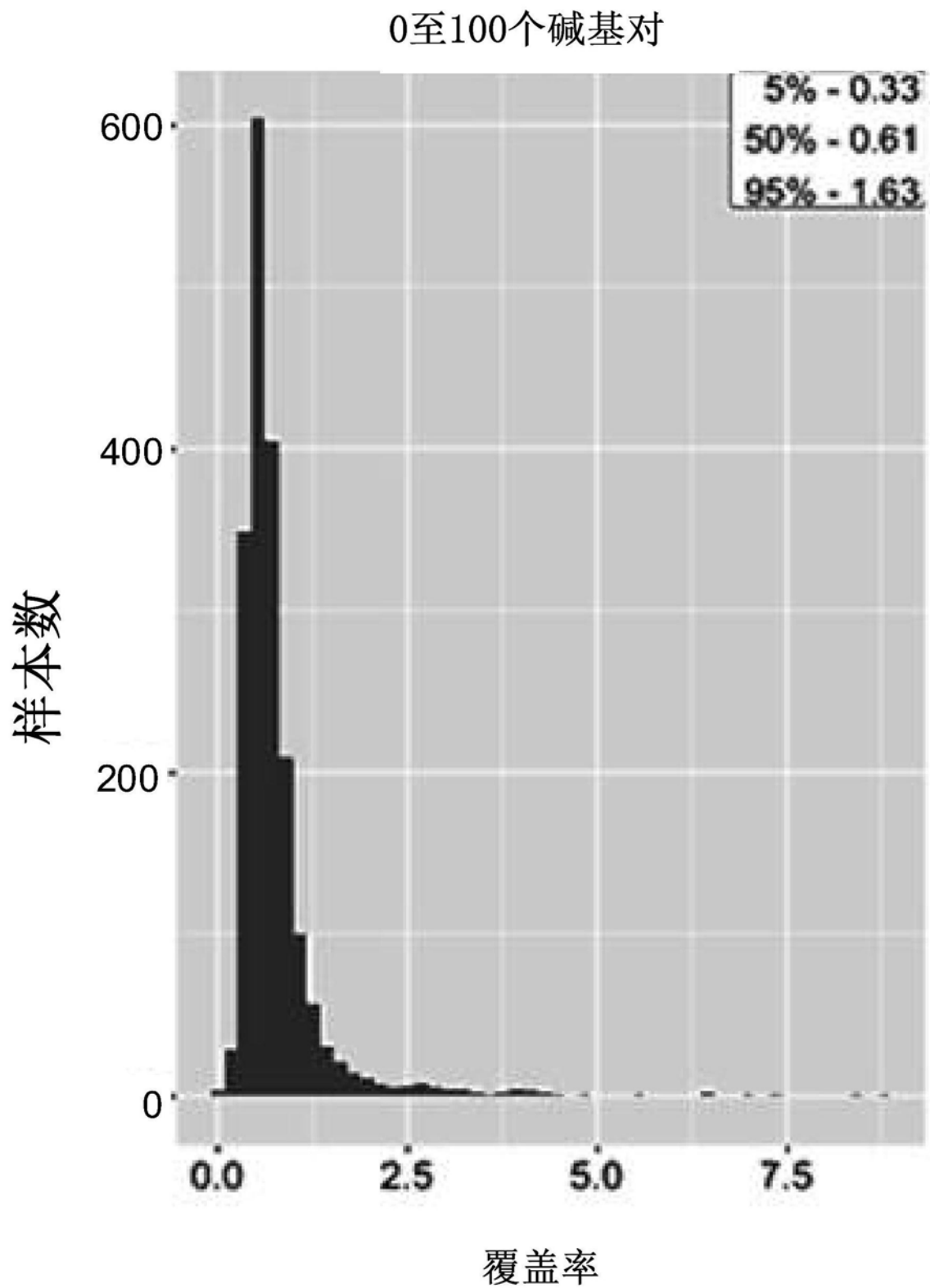


图16C

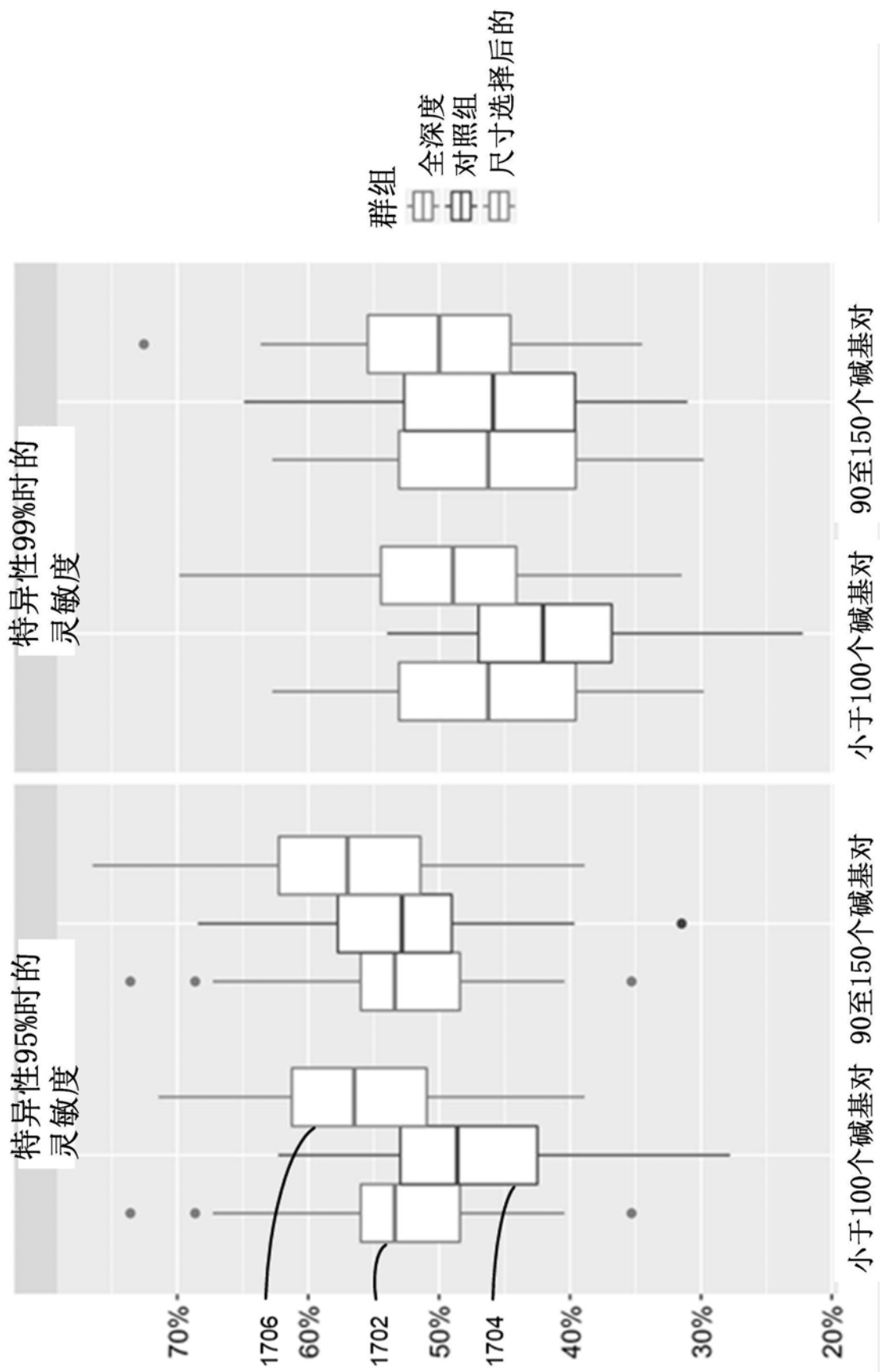


图17A

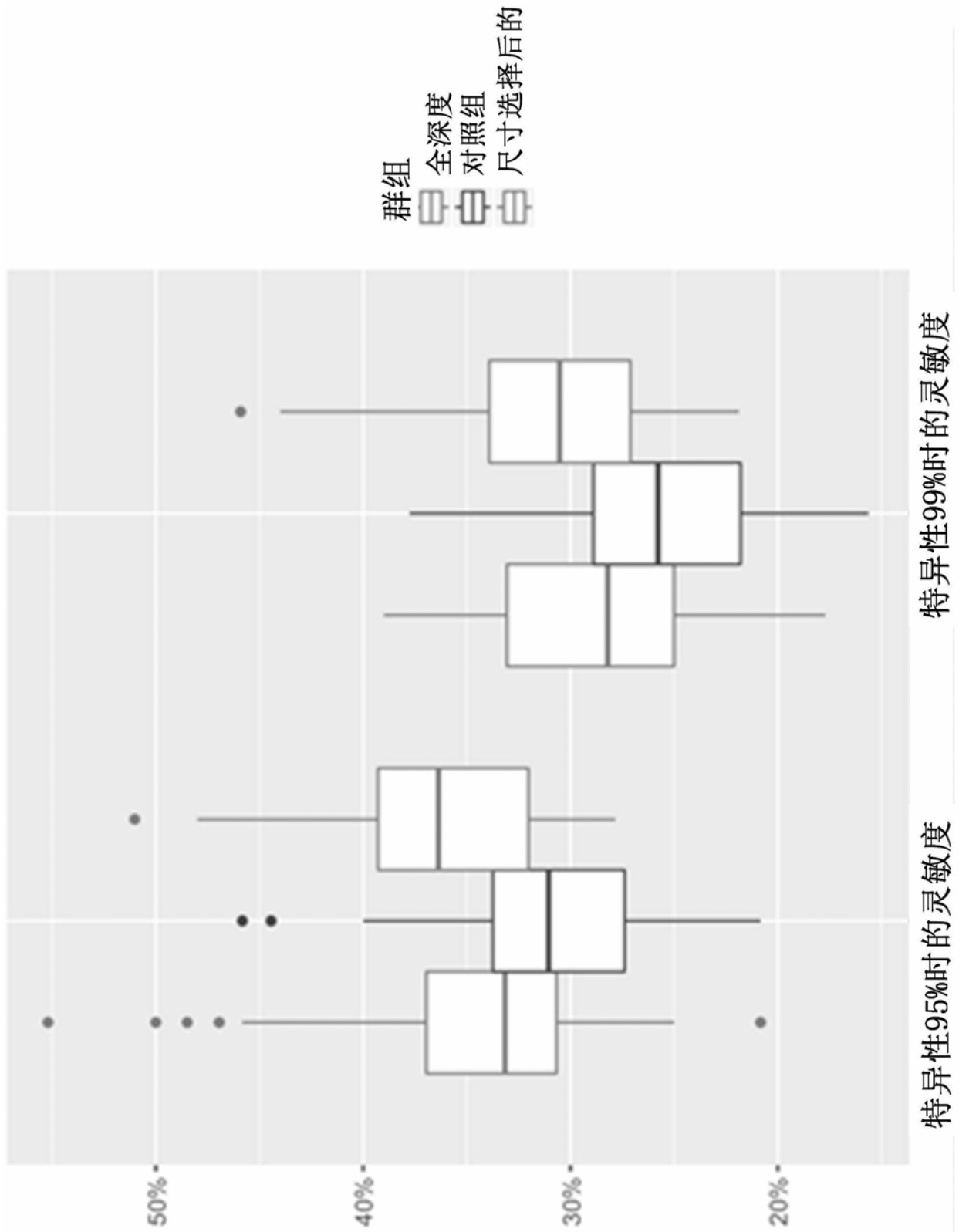


图17B

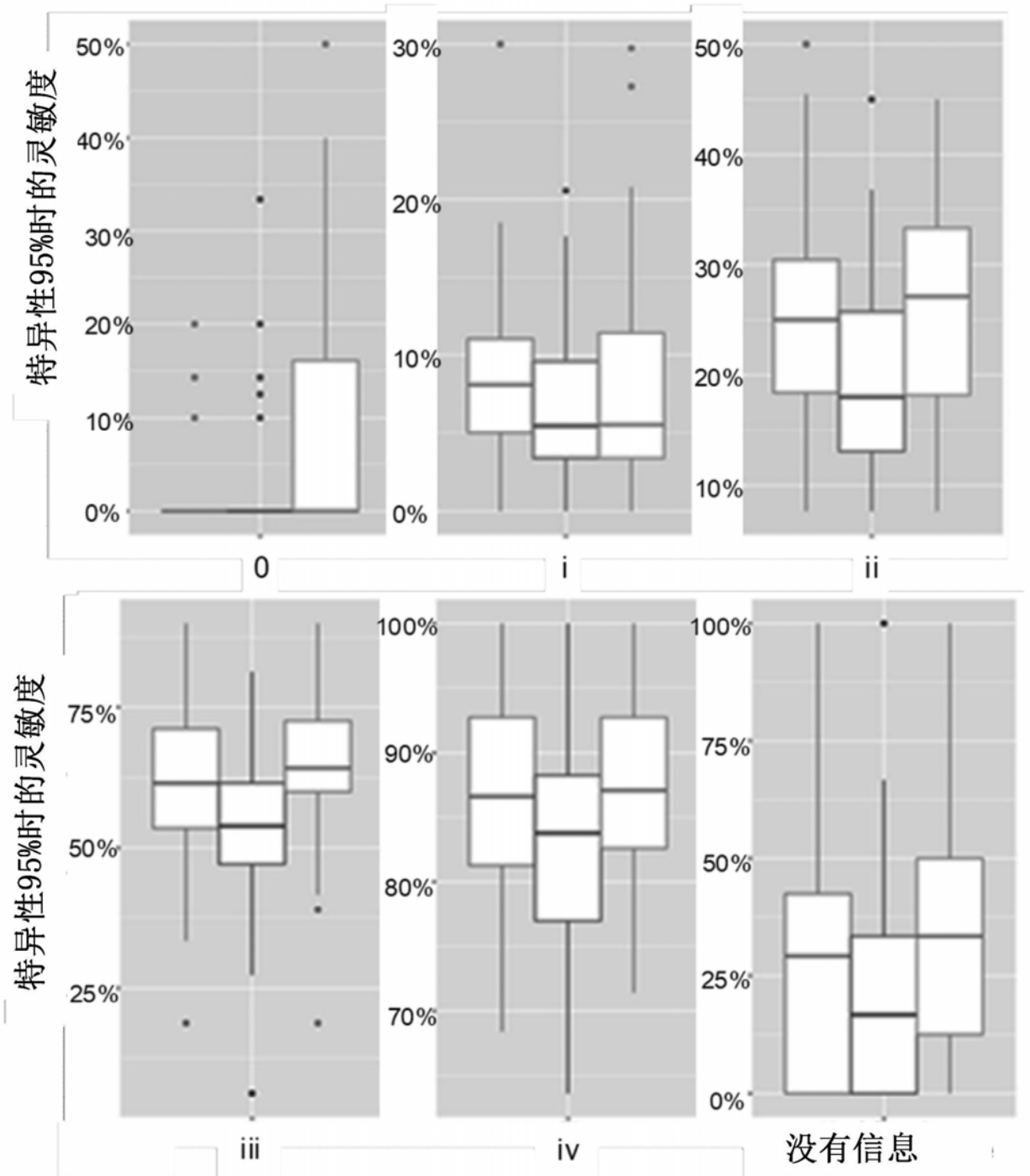


图17C

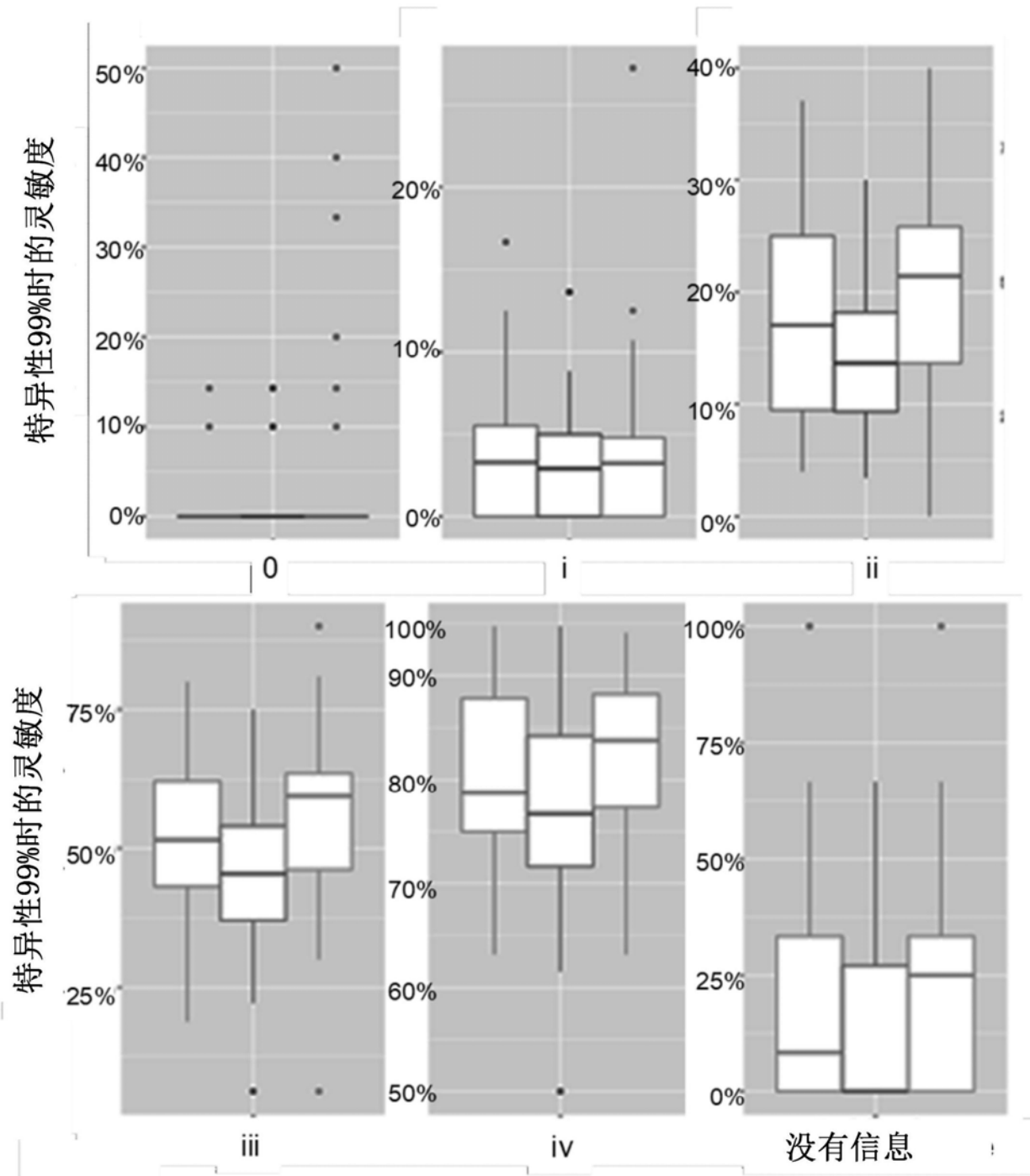


图17D

特异性95%时的灵敏度										
尺寸选择的阶段	子样本	全深度	子样本	全深度	子样本	全深度	子样本	全深度	子样本	全深度
p值	0		I		II		III		IV	
	0.08	0.01	0.24	0.85	2e-6	0.06	3e-7	8e-3	3e-4	0.5
平均变化 (2.5%, 97.5% CI)	0.036 (-0.005, 0.072)	0.044 (0.010, 0.079)	0.014 (-0.010, 0.039)	0.002 (-0.024, 0.028)	0.086 (0.055, 0.118)	0.026 (-0.001, 0.055)	0.105 (0.071, 0.141)	0.035 (0.009, 0.061)	0.047 (0.023, 0.071)	0.006 (-0.012, 0.024)
特异性99%时的灵敏度										
尺寸选择的阶段	子样本	全深度	子样本	全深度	子样本	全深度	子样本	全深度	子样本	全深度
p值	0		I		II		III		IV	
	0.08	0.07	0.04	0.48	6e-6	7e-3	4e-8	6e-3	2e-5	4e-3
平均变化 (2.5%, 97.5% CI)	0.025 (-0.003, 0.052)	0.025 (-0.003, 0.052)	0.015 (0.000, 0.027)	0.005 (-0.009, 0.018)	0.063 (0.038, 0.087)	0.031 (0.009, 0.053)	0.123 (0.086, 0.160)	0.059 (0.017, 0.099)	0.061 (0.035, 0.087)	0.030 (0.010, 0.051)

图17E

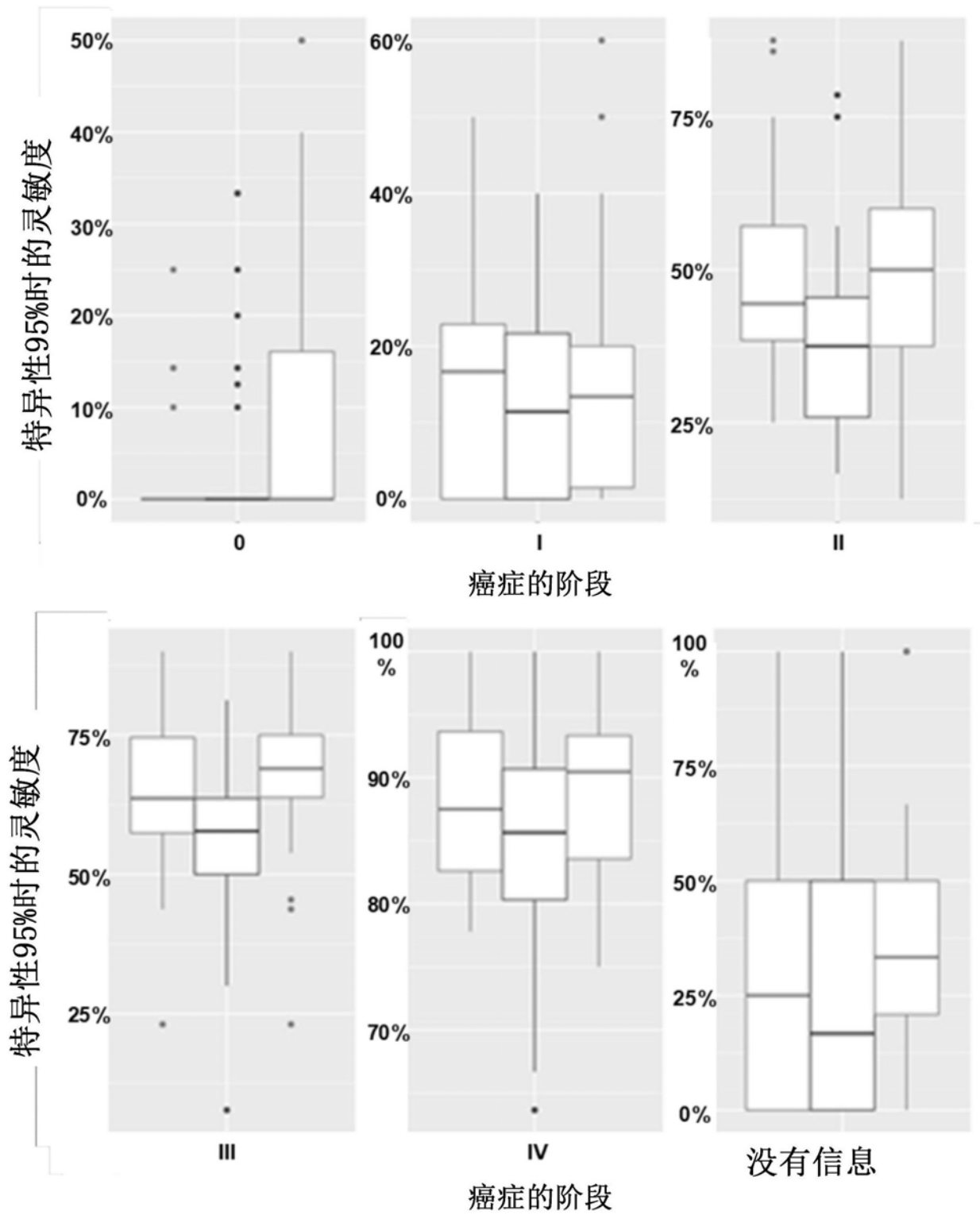


图17F

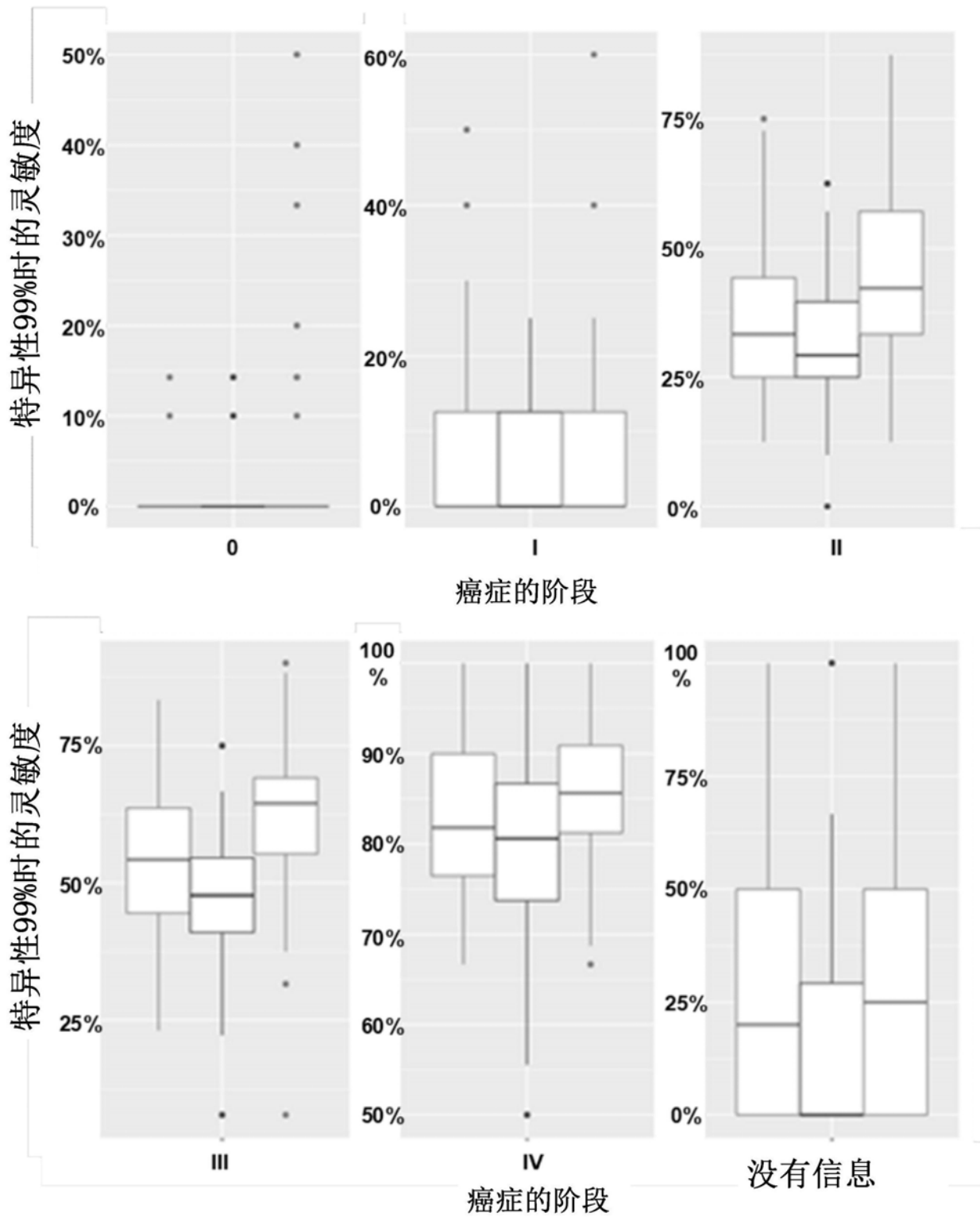


图17G

特异性95%时的灵敏度										
尺寸选择的	子样本	全深度	子样本	全深度	子样本	全深度	子样本	全深度	子样本	全深度
阶段	0		I		II		III		IV	
p值	0.34	0.01	0.15	0.82	3e-6	2e-3	1e-7	2e-3	8e-4	0.51
平均变化 (2.5%, 97.5% CI)	0.021 (-0.023, 0.065)	0.044 (0.010, 0.079)	0.028 (-0.010, 0.066)	0.005 (-0.039, 0.049)	0.144 (0.091, 0.198)	0.070 (0.028, 0.112)	0.115 (0.079, 0.151)	0.043 (0.017, 0.069)	0.047 (0.021, 0.074)	0.006 (-0.013, 0.025)

特异性99%时的灵敏度										
尺寸选择的	子样本	全深度	子样本	全深度	子样本	全深度	子样本	全深度	子样本	全深度
阶段	0		I		II		III		IV	
p值	0.05	0.05	0.02	0.26	2e-5	2e-4	5e-10	4e-4	3e-5	2e-4
平均变化 (2.5%, 97.5% CI)	0.029 (0.001, 0.059)	0.029 (0.001, 0.059)	0.035 (0.005, 0.065)	0.019 (-0.014, 0.052)	0.121 (0.070, 0.171)	0.085 (0.042, 0.128)	0.142 (0.107, 0.178)	0.075 (0.036, 0.114)	0.070 (0.040, 0.100)	0.040 (0.020, 0.059)

图17H

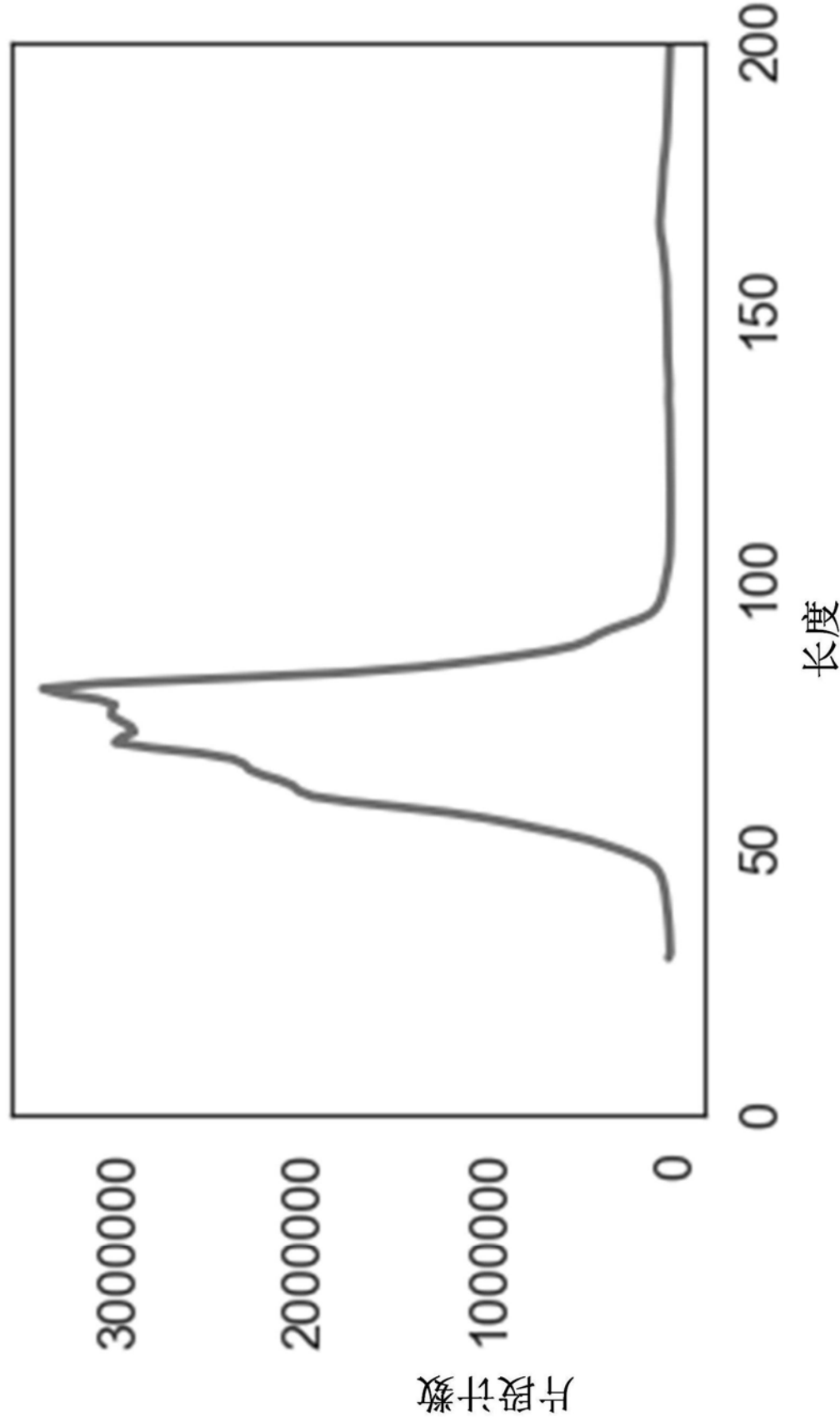


图18

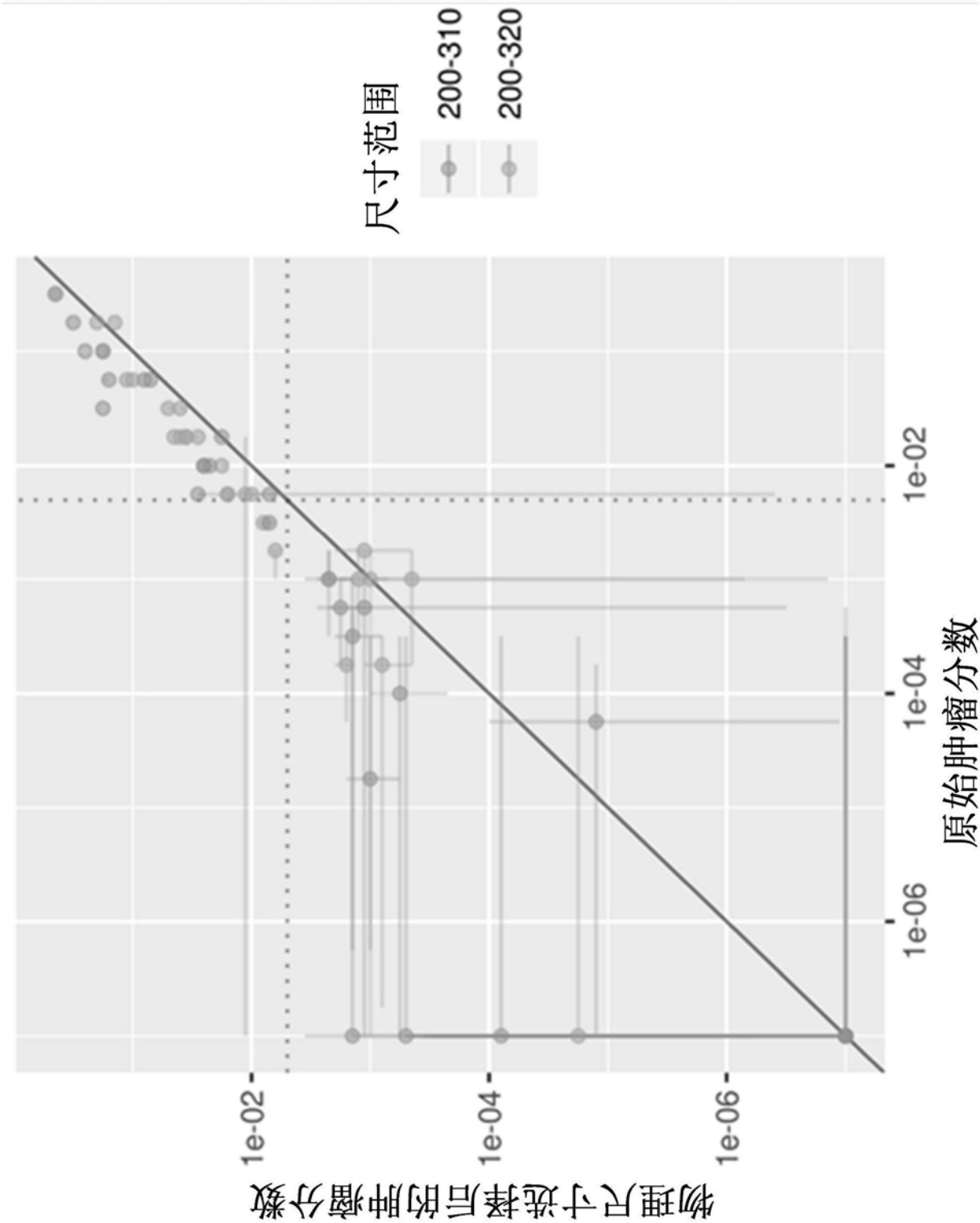


图19

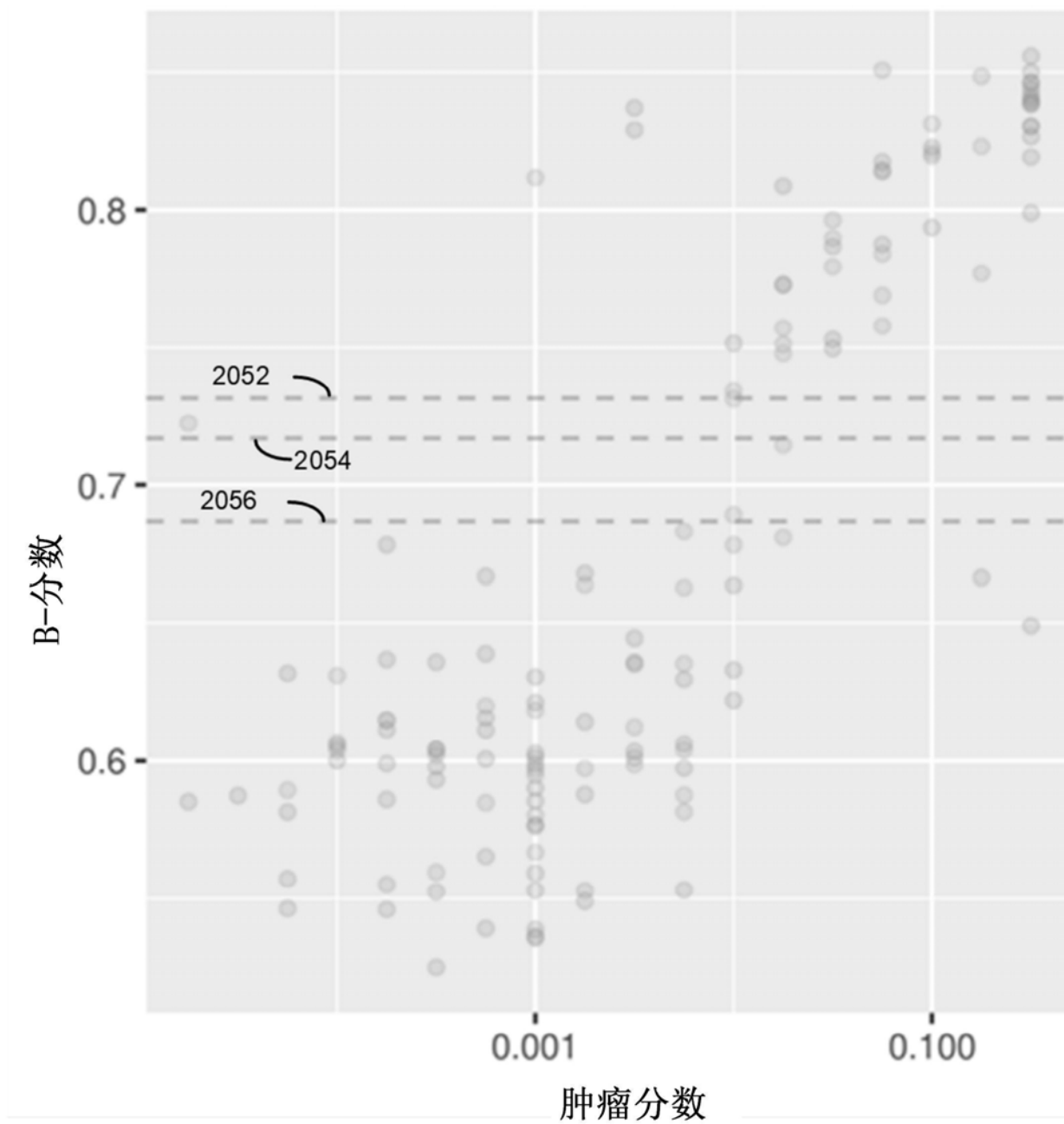


图20