



(12) 发明专利申请

(10) 申请公布号 CN 118369727 A

(43) 申请公布日 2024. 07. 19

(21) 申请号 202280081951.9

(22) 申请日 2022.10.31

(30) 优先权数据

63/274,330 2021.11.01 US

(85) PCT国际申请进入国家阶段日

2024.06.11

(86) PCT国际申请的申请数据

PCT/US2022/078956 2022.10.31

(87) PCT国际申请的公布数据

W02023/077114 EN 2023.05.04

(71) 申请人 佩索纳里斯公司

地址 美国加利福尼亚州

(72) 发明人 扎伊德·M·鲁桑

尼古拉斯·A·菲利普斯

杰森·哈里斯 张宁

(74) 专利代理机构 北京安信方达知识产权代理

有限公司 11262

专利代理师 刘晓杰 郑霞

(51) Int.Cl.

G16B 25/10 (2006.01)

G16B 20/20 (2006.01)

G16B 30/10 (2006.01)

G16B 40/30 (2006.01)

权利要求书2页 说明书26页 附图17页

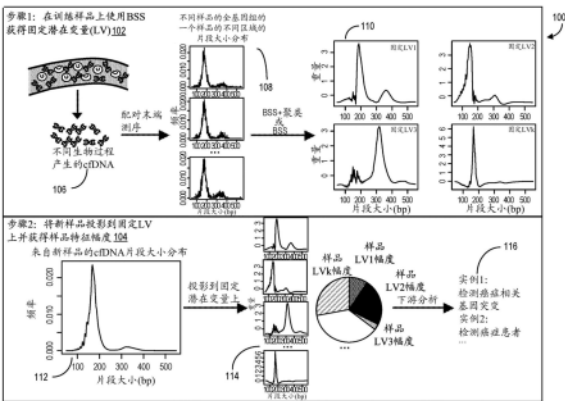
(54) 发明名称

基于核酸分子的潜在变量确定片段组学特征

征

(57) 摘要

一种基于片段组学特征预测受试者疾病分类的方法可以包括访问受试者生物样品的序列数据。所述方法还可以包括基于序列数据生成一组序列大小值。所述组中的每个序列大小值可以对应于序列数据的序列的大小。所述方法还可以包括通过将所述组序列大小值投影到片段组学特征的潜在变量上来确定受试者的片段组学特征幅度。可以通过将一个或多个信号分离算法应用于从一个或多个参考生物样品获得的其他序列大小值来生成潜在变量。所述方法也可以包括通过使用机器学习模型处理片段组学特征幅度来生成结果。结果可以包括预测受试者是否患有特定疾病的分类。



1. 一种方法,其包括:

访问受试者的生物样品的序列数据;

基于所述序列数据生成一组序列大小值,其中所述组中的每个序列大小值对应于所述序列数据的序列的大小;

通过将所述组序列大小值投影到片段组学特征的潜在变量上来确定受试者的片段组学特征幅度,其中所述潜在变量是通过将一个或多个信号分离算法应用于从一个或多个参考生物样品获得的其他序列大小值来生成的;

通过使用机器学习模型处理所述片段组学特征幅度来生成结果,其中所述结果包括预测所述受试者是否患有特定疾病的分类;以及

输出所述结果。

2. 根据权利要求1所述的方法,其中所述片段组学特征的所述潜在变量中的每个潜在变量包括表示所述一个或多个参考样品的其他序列大小值的大小分布的直方图或权重向量,并且其中所述生物样品的片段组学特征幅度是通过将多组序列大小值投影到所述潜在变量中的每个潜在变量上来确定的。

3. 根据权利要求1或权利要求2所述的方法,其中所述组序列大小值对应于与一个或多个基因组区域比对的序列数据的序列。

4. 根据权利要求1至3中任一项所述的方法,其中所述一个或多个信号分离算法包括一个或多个盲源分离算法。

5. 根据权利要求4所述的方法,其中所述一个或多个盲源分离算法还包括独立成分分析算法。

6. 根据权利要求4所述的方法,其中所述一个或多个盲源分离算法还包括非负矩阵分解算法。

7. 根据权利要求1至6中任一项所述的方法,其中所述组潜在变量中的第一个潜在变量的一个或多个图形分量可预测与核小体结合的DNA螺旋间距相关的DNA片段的逐渐消化。

8. 根据权利要求1至7中任一项所述的方法,还包括其中所述组潜变量中的第二个潜在变量的一个或多个图形分量可预测DNA结合蛋白的细胞间异质性。

9. 根据权利要求1至8中任一项所述的方法,其中由相应的序列大小值表示的每个序列包括大小在60bp至600bp之间的DNA片段。

10. 根据权利要求1至9中任一项所述的方法,其中所述序列数据包括与从所述生物样品中检测到的多个体细胞变异相对应的序列。

11. 根据权利要求1至10中任一项所述的方法,其中所述组序列大小值进一步是基于所述序列数据的序列生成的经验概率质量函数。

12. 根据权利要求1至11中任一项所述的方法,其中所述序列数据对应于所述生物样品的多个无细胞DNA分子,并且其中所述多个无细胞DNA分子包括循环肿瘤DNA分子。

13. 根据权利要求1至12中任一项所述的方法,其中所述特定疾病为癌症。

14. 根据权利要求1至13中任一项所述的方法,其中确定所述受试者的片段组学特征幅度包括将所述组序列大小值投影到所述潜在变量的子集中的每一个上。

15. 根据权利要求14所述的方法,其中所述片段组学特征的潜在变量的子集包括对所述潜在变量应用聚类算法。

16. 根据权利要求14所述的方法, 其中所述片段组学特征的潜在变量的子集用作另一个机器学习模型的预处理训练数据。

17. 根据权利要求14所述的方法, 其中所述片段组学特征的潜在变量组的子集用作后续主成分分析的成分。

18. 根据权利要求1至17中任一项所述的方法, 其还包括:

基于所述序列数据生成一组末端基序序列数据, 其中所述组中的每个末端基序序列数据标识具有对应于特定末端基序的末端序列的核酸分子的数量或相对频率; 以及

通过对所述组末端基序序列数据应用一个或多个信号分离算法来确定另一个片段组学特征的潜在变量。

19. 根据权利要求18中任一项所述的方法, 其中确定所述受试者的片段组学特征幅度包括将所述组末端基序序列数据投影到所述另一个片段组学特征的潜在变量上。

20. 一种系统, 其包括:

一个或多个数据处理器; 和

含有指令的非暂时性计算机可读存储介质, 当所述指令在一个或多个数据处理器上执行时, 使得所述一个或多个数据处理器执行本文公开的一个或多个方法的部分或全部。

21. 一种有形地体现在非暂时性机器可读存储介质中的计算机程序产品, 包括被配置为使得一个或多个数据处理器执行本文公开的一个或多个方法的部分或全部的指令。

基于核酸分子的潜在变量确定片段组学特征

[0001] 相关申请的交叉引用

[0002] 本申请要求2021年11月1日提交的标题为“基于核酸分子的潜在变量确定基因特征”的美国临时专利申请号63/274,330的优先权,出于所有目的,其内容特此通过引用整体并入。

背景技术

[0003] 下一代测序可用于识别受试者的遗传特征。例如,全基因组测序可用于揭示受试者序列数据中的体细胞变异,其中一些对应于肿瘤DNA。此外,最近对无细胞DNA片段组学特征诸如片段长度和末端基序的研究扩大了使用血浆样品准确诊断受试者的潜力。例如,此类研究促进发现与不同类型肿瘤相对应的片段长度特征。也一直在努力开发上述技术来检测受试者中的癌症。

[0004] 尽管做出了这些努力,但从血浆样品中准确检测出癌症仍然难以实现。造成这种困难的一个因素是,血浆样品中的肿瘤DNA数量可能会有很大变化,这取决于肿瘤类型、疾病进展阶段以及肿瘤细胞DNA释放进入循环系统(肿瘤“脱落”)的程度。由于这些不同的特征,识别肿瘤DNA的特定特征(例如片段长度、末端基序)一直具有挑战性。此外,血浆样品中的肿瘤DNA浓度可能很低,以至于难以对受试者进行准确诊断。

发明内容

[0005] 在一些实施方案中,提供了一种根据基于核酸分子的大小分布的片段组学特征来预测受试者的疾病分类的方法。所述方法可以包括访问受试者的生物样品的序列数据。所述方法还可以包括基于序列数据生成一组序列大小值。所述组中的每个序列大小值可以对应于序列数据的序列的大小。所述方法还可以包括通过将所述组序列大小值投影到片段组学特征的潜在变量上来确定受试者的片段组学特征幅度。可以通过将一个或多个信号分离算法应用于从一个或多个参考生物样品获得的其他序列大小值来生成潜在变量。所述方法可以通过使用机器学习模型处理片段组学特征幅度来生成结果。结果可以包括预测受试者是否患有特定疾病的分类。所述方法可以包括输出结果。

[0006] 在一些实施方案中,提供了一种根据基于末端基序频率分布的片段组学特征来预测受试者的疾病分类的方法。所述方法可以包括访问受试者的生物样品的序列数据。所述方法还可以包括基于序列数据生成一组末端基序序列数据。所述组中的每个末端基序序列数据标识具有对应于特定末端基序的末端序列的核酸分子的数量或相对频率。所述方法还可以包括通过将所述组末端基序序列数据投影到片段组学特征的潜在变量上来确定受试者的片段组学特征幅度。可以通过将一个或多个信号分离算法应用于从一个或多个参考生物样品获得的其他末端基序序列数据来生成潜在变量。所述方法可以通过使用机器学习模型处理片段组学特征幅度来生成结果。结果可以包括预测受试者是否患有特定疾病的分类。所述方法可以包括输出结果。

[0007] 本公开的一些实施方案包括一种系统,所述系统包括一个或多个数据处理器。在

一些实施方案中,所述系统包括含有指令的非暂时性计算机可读存储介质,当所述指令在一个或多个数据处理器上被执行时,使得一个或多个数据处理器执行本文公开的一个或多个方法的部分或全部和/或一个或多个过程的部分或全部。本公开的一些实施方案包括有形地体现在非暂时性机器可读存储介质中的计算机程序产品,包括被配置为使得一个或多个数据处理器执行本文公开的一个或多个方法的部分或全部和/或一个或多个过程的部分或全部的指令。

[0008] 所采用的术语和表述用作描述性术语而非限制性术语,并且这些术语和表述的使用无意排除所示和所述特征或其部分的任何等同物,但应认识到,在所要求保护的本发明范围内,各种修改是可能的。因此,应当理解,尽管所要求保护的本发明已通过一些实施方案和任选特征进行了具体公开,但是本领域技术人员可以对本文所公开的概念进行修改和变化,并且这种修改和变化被认为在所附权利要求所定义的本发明的范围内。

附图说明

[0009] 结合附图描述本公开。

[0010] 图1显示了根据一些实施方案的示意图,所述示意图说明从序列数据生成潜在变量并使用潜在变量预测癌症存在的过程。

[0011] 图2包括根据一些实施方案的流程图,所述流程图说明根据基于核酸分子的大小分布确定的片段组学特征来预测受试者的疾病分类的方法的实例。

[0012] 图3显示了根据一些实施方案的示意图,所述示意图说明用于生成序列大小值的实例技术。

[0013] 图4显示了根据一些实施方案的实例图,所述实例图说明使用信号分离算法生成一组未混合图像的过程。

[0014] 图5显示了根据一些实施方案的示意图,所述示意图说明用于生成一组序列大小潜在变量的实例技术。

[0015] 图6显示了根据一些实施方案的示意图,所述示意图说明对全外显子组序列数据使用独立成分分析算法以生成片段组学特征的实例技术。

[0016] 图7显示了通过对与正常受试者的生物样品相对应的序列大小值应用信号分离算法而生成的一组潜在变量。

[0017] 图8显示了通过对与另一正常受试者的生物样品相对应的序列大小值应用信号分离算法而生成的一组潜在变量。

[0018] 图9显示了通过对与被诊断为结直肠癌的受试者的生物样品相对应的序列大小值应用信号分离算法而生成的一组潜在变量。

[0019] 图10显示了通过对与被诊断为结直肠癌的另一受试者的生物样品相对应的序列大小值应用信号分离算法而生成的一组潜在变量。

[0020] 图11显示了根据一些实施方案的示意图,所述示意图说明基于片段组学特征预测源自具有不同表观遗传状态的基因座的DNA分子的富集的技术。

[0021] 图12显示了根据一些实施方案的示意图,所述示意图说明通过将原始序列大小分布投影到潜在变量上,将分布转换为一组片段组学特征幅度,并使用所述幅度检测癌症相关基因突变来预处理原始序列大小分布的过程。

[0022] 图13显示了根据一些实施方案的示意图,所述示意图说明对杂交捕获样品使用独立成分分析算法生成一组潜在变量的实例技术。

[0023] 图14显示了根据一些实施方案的示意图,所述示意图说明使用手工制作的特征或潜在变量特征监测癌症复发的实例技术。

[0024] 图15显示了一组接收者操作特性 (ROC) 曲线,其显示了使用片段组学特征对疾病进行分类的准确度水平。

[0025] 图16包括根据一些实施方案的流程图,所述流程图说明根据基于核酸分子的末端基序频率确定的片段组学特征来预测受试者的疾病分类的方法的实例。

[0026] 图17说明了用于实现本文公开的一些实施方案中的一些的计算机系统的实例。

具体实施方式

[0027] 片段组学通常是指对无细胞DNA的片段化模式的分析,包括片段大小和末端基序。这些片段化模式可能与组织类型和癌症特有的表观遗传特征有关。人们认为,正常细胞和癌细胞组蛋白以及控制染色质结构的相关调节蛋白以及相互关联的基因转录景观的差异在一定程度上可以表现为DNA片段长度分布差异。尽管取得了这些进展,但难以利用这些特征来开发有效的疗法(例如新抗原疫苗)。例如,血浆样品中的肿瘤DNA水平通常较低,这种有限的的数据可能无法使结果在预测受试者是否患有癌症方面达到足够的准确性。在另一个实例中,在无细胞DNA中发现的非肿瘤来源的体细胞变异,包括克隆造血的那些,会使变异调用和肿瘤DNA的检测复杂化。在又一个实例中,序列知识数据库的肿瘤序列数据的可用性可能是不一致的,使得它们不能充分用作训练机器学习模型来预测受试者是否患有癌症的训练数据。

[0028] 为了解决这些挑战,本技术可以包括基于片段组学特征预测受试者的疾病分类。如本文所用,片段组学特征是指可以预测疾病分类的核酸分子的大小和/或末端基序分布的一个或多个特征。例如,片段组学特征可以表示与一组基因组区域中的每个基因组区域比对的序列的片段长度(可互换称为“大小”)的估计分布,其可以预测疾病的分类。另外地或替代地,片段组学特征可以表示X-bp 5'和3'序列同一性或“末端基序”(例如CCCA)的估计分布,其可以预测疾病的分类。

[0029] 片段组学特征可以包括基于将盲源分离 (BSS) 算法应用于从参考队列获得的核酸分子的大小和/或末端基序分布而生成的一个或多个潜在变量。每个潜在变量可以鉴定在基因组区域组中可变地富集的序列的估计大小和/或末端基序分布,并为序列大小和/或末端基序数据定义新的基向量。在一些情况下,一个或多个潜在变量是基于从诊断有疾病的参考样品获得的序列数据确定的。在一些情况下,一个或多个潜在变量鉴定具有与一组特定末端基序相对应的末端序列的序列读数的估计分布。然后可以使用片段组学特征来预测受试者是否患有癌症(例如)。

[0030] 基于片段组学特征预测受试者疾病分类的技术可以通过访问受试者生物样品的序列数据来启动。在一些情况下,测序数据识别多个核酸序列,这些核酸序列是通过对生物样品的多个无细胞DNA分子进行测序获得的。多个无细胞DNA分子可以包括循环肿瘤DNA分子。另外地或替代地,序列数据还可以识别对应于从生物样品中检测到的多个体细胞变异的序列读数。可以通过将序列数据的每个序列读数与参考序列(例如,人参考基因组)比对

来检测多个体细胞变异。

[0031] 在一些情况下,参考序列包括从一个或多个受试者的健康血细胞(例如,白细胞)、颊细胞和/或根毛根细胞获得的“正常”或“健康”序列。细胞可以通过多种方式被鉴定为健康的,例如,当一个人之前被诊断为没有特定类型的癌症或可以从不太可能含有癌细胞或癌前细胞的组织中获得样品时。在一些情况下,正常序列是通过以下方式获得的:(i)从血液样品中分离血浆和包括白细胞和外周血单核细胞的血沉棕黄层;(ii)从血沉棕黄层中分离DNA;以及(iii)从分离的DNA中确定正常序列。美国专利号10,125,399进一步描述了从生物样品中确定正常序列的实例技术,出于所有目的,所述专利的内容通过引用整体并入本文。

[0032] 基于序列数据,可以生成一组序列大小值(例如,序列大小值的二维矩阵)。例如,序列大小值组可以包括序列大小值的二维矩阵,其中定义序列大小(例如,30bp)的第一维与鉴定与序列大小相对应的序列数(例如,50个计数)的第二维相关联。对于与一组基因组区域的相应基因组区域比对的序列数据的每个核酸序列,所述组序列大小值可以包括表示序列大小的序列大小值。由相应序列大小值表示的每个序列可以包括一定大小(bp)范围(例如,60bp与600bp之间的大小范围)内的DNA片段。在一些情况下,使用参考序列(例如,人参考基因组)来鉴定所述组基因组区域。另外地或替代地,所述组序列大小值可以转换为一组经验概率质量函数,其中每个经验概率函数可以基于与相应基因组区域比对的核酸序列的序列大小值生成。在一些情况下,代表所述组序列大小值的大小分布数据是根据受试者的序列数据确定的。例如,所述组序列大小值可以转换为经验概率质量函数(PMF)。如本文所用,PMF是指估计离散随机变量恰好等于某个值的概率的函数。所述组PMF可以用作输入来确定所述组潜在变量。

[0033] 为了使用生物样品的所述组序列大小值来预测疾病的分类,可以确定片段组学特征。为了确定片段组学特征,首先可以生成一组潜在变量。如本文所用,术语“潜在变量”是指序列大小的估计分布(编码为序列大小有符号权重的数字向量的模式)对应于与所述组基因组区域比对的序列数据背后的几个潜在变异来源之一。可以通过将一个或多个信号分离算法应用于从一个或多个参考样品获得的序列大小值来确定所述组潜在变量。在一些情况下,参考样品包括从具有疾病诊断(例如,癌症、健康)的其他受试者获得的生物样品。另外地或替代地,为了监测疾病的进展,参考样品可以包括从同一受试者但在不同时间点获得的生物样品。例如,参考样品可以对应于在获得当前生物样品的时间之前2年获得的受试者的生物样品。

[0034] 一个或多个信号分离算法可以包括盲源分离算法、独立成分分析算法和/或非负矩阵分解算法。如本文所用,盲源分离算法可以是一种将一组源信号(例如,潜在变量)与一组混合信号(例如,序列大小值)分离的技术,无需借助有关源信号或混合过程的信息(或仅借助很少的信息)。另外地或替代地,一个或多个信号分离算法可以是一个或多个其他无监督机器学习技术。在一些情况下,通过将聚类算法应用于所述组潜在变量并鉴定由聚类算法生成的每个聚类的质心或中心点来选择潜在变量的子集(或者称为“固定”潜在变量)。固定的潜在变量可用作片段组学特征,其可以表示一个或多个参考样品中反映的不同生物过程产生的核酸分子。

[0035] 然后可以将生物样品的序列大小值组投影到片段组学特征的潜在变量上。例如,

可以将从序列大小值组确定的大小分布数据投影到固定的潜在变量上以生成一个或多个潜在变量系数(可互换称为“幅度”)。片段组学特征幅度可以表示生物样品中不同组核酸分子的富集。

[0036] 可以通过使用机器学习模型处理片段组学特征幅度来生成结果,其中结果包括预测受试者是否患有特定疾病的分类。另外地或替代地,结果包括预测受试者是否患有特定类型癌症的分类,并且结果可用于预测受试者的治疗和/或预测应多久对受试者施用治疗。

[0037] 在一些情况下,片段组学特征可用于生成执行其他类型任务的结果。根据一些实施方案,实例任务可以包括基于片段组学特征,基于将某些潜在变量鉴定为由独立调节的分子组所携带的信号,测量与序列大小数据中的表观遗传状态相关的蛋白质分子组的富集。例如,已知与特定基因组区域或特定等位基因的染色体相关的核酸分子的生物样品的富集可用于预测染色体的结合和在有助于测序数据的相应细胞基因座处的沉默异染色质状态的富集。在另一个实例中,序列数据中与癌症相关的核酸分子的生物样品富集可用于预测特定等位基因是否来自癌细胞。在又一个实例中,序列数据中与基因表达相关的核酸分子的生物样品富集可用于预测有助于测序数据的特定等位基因是否正在表达。

[0038] 又另一项任务包括估计已知与诸如核小体和转录因子的蛋白质分子结合的生物样品的核酸分子的大小变异性和中心趋势。另一项任务包括通过比较一个或多个多峰潜在变量的峰值之间的大小差异来推断不同蛋白质的间距和接头DNA大小。其他任务包括预测一个或多个多峰潜在变量的短序列和长序列大小物种的相对丰度,以及从潜在变量空间的不同变化率推断片段组学特征的无细胞DNA降解动力学。

[0039] 另外地或替代地,片段组学特征可用于数据去噪,这一功能在低数据量的情况下尤其有利。将原始数据(例如,大小分布数据)投影到潜在变量空间上减少模型的下游过拟合,并通过降低噪声或其他被认为不感兴趣的数据来提高统计测试的准确性。又一个实例可以包括通过将原始数据投影到感兴趣的几个潜在变量上以生成几个数值幅度(例如,两个,适用于在2D图中描绘数据)来使用潜在变量进行降维。可以比较野生型等位基因支持读数和突变型等位基因支持读数之间的数值幅度,从而可以过滤掉由PCR或测序错误引起的假阳性突变。

[0040] 因此,本公开的一些实施方案通过使用信号分离和聚类算法来生成片段组学特征,以表示生物样品中序列大小和/或末端基序数据背后的独立变异来源的准确且可概括的模型,从而提供了优于传统技术的技术优势。例如,盲源分离算法可以从序列数据中识别隐藏但独立的因素,其可用于预测受试者的疾病或疾患。在一些情况下,片段组学特征可用于预测不同核酸或蛋白质分子的特征,其可用于不同表观遗传状态的诊断或预后。如上所述,片段组学可以识别核酸分子的固有特性,通过了解它们所代表的内容,这些特性可用于解释疾病表型。

[0041] 所有上述技术优势都会给计算效率以及监督机器学习模型解释、普遍性和性能带来下游优势。然后可以使用包括固定潜在变量的片段组学特征来预测生物样品是否包含肿瘤DNA和/或也可以用来预测癌症的特定阶段。此外,即使检测到的肿瘤DNA量较低,使用片段组学特征也可以准确预测受试者是否患有特定疾病。因此,本技术通过将机器学习模型应用于来自无细胞血浆样品的序列数据来促进准确可靠地检测受试者的疾病。

[0042] 提供以下实施例来介绍某些实施方案。在以下描述中,出于解释的目的,陈述了具

体细节以提供对本公开的实施例的深入理解。然而,显然的是,没有这些具体细节也可以实践各种实施例。例如,装置、系统、结构、组件、方法和其他部件可以以框图形式显示为部件,以免不必要的细节模糊实施例。在其他情况下,可以在没有必要细节的情况下显示众所周知的装置、过程、系统、结构和技术,以避免模糊实施例。附图和描述并非旨在限制。本公开中所使用的术语和表述用作描述而非限制的术语,并且这些术语和表述的使用无意排除所示和所述特征或其部分的任何等同物。本文使用的“实施例”一词意指“作为实例、示例或说明”。本文中描述为“实施例”的任何实施方案或设计不一定被解释为比其他实施方案或设计优选或有利。

[0043] I. 基于片段组学特征预测受试者疾病分类的概述

[0044] 对受试者的配对肿瘤活检和无细胞DNA血浆研究揭示了两个相应的体细胞变异调用组之间的不同特征。发现配对肿瘤和血浆体细胞变异调用组之间的差异具有很大的可变性,因为这些差异可能受到肿瘤类型、阶段、异质性和每个样品获取时间的影响。

[0045] 与正常无细胞DNA相比,肿瘤DNA片段长度的整体模式已被证明更短。通过富集已知肿瘤变体的生物样品,然后检查从正常和富集的生物样品对产生的序列读数推断出的片段长度,可以实现预测疾病状态区分特征的技术。尽管如此,仅根据其大小分布来区分肿瘤DNA和正常DNA可能具有挑战。此外,生物样品中肿瘤DNA的量可能有所不同,其中先前的技术无法准确预测特定的一组序列是否来自肿瘤DNA。由于这种困难,将肿瘤DNA的DNA片段长度整合到各种技术(例如,体细胞变异调用)中仍然难以实现。

[0046] 为了解决这些挑战,本技术可以包括基于BSS算法生成的潜在变量确定片段组学特征。片段组学特征可用于预测受试者的疾病分类。可以通过将一个或多个机器学习模型(例如,一个或多个无监督机器学习模型、BSS算法)应用于从参考样品获得的多组序列大小值来生成潜在变量。可以从诊断患有疾病(例如癌症)的受试者和/或健康受试者获得参考样品。可以测量每个序列的大小来鉴定序列大小值。多组序列大小值中的每一组可以与参考基因组中鉴定的一组基因组区域中的基因组区域相关联。多组序列大小值可以表示为矩阵或任何表示序列大小值的大小分布的数据结构。另外地或替代地,每组序列大小值可以转换为经验概率质量函数(PMF),使得多组序列大小值可以由一组PMF表示。每个潜在变量表示一组与独立生物过程、技术伪影或噪声源相关的加权片段大小。可以选择具有生物学意义的潜在变量(例如,固定潜在变量)并将其用作片段组学特征。在一些情况下,将聚类算法应用于潜在变量以选择形成片段组学特征的固定潜在变量。

[0047] 可以通过将特定受试者(例如,未诊断出疾病的受试者)的生物样品的序列的大小分布投影到所述组具有生物学意义的潜在变量上来确定所述特定受试者的片段组学特征幅度。在一些情况下,片段组学特征幅度可用于预测受试者的序列数据是否包括已知肿瘤来源的体细胞变异,这可以促进无细胞变异调用和早期癌症检测。在一些情况下,通过分类器模型处理片段组学特征幅度,以预测一个人是否患有某种疾病(例如癌症)。

[0048] A. 使用片段组学特征预测受试者疾病的实例技术

[0049] 图1显示了根据一些实施方案的示意图100,所述示意图说明从序列数据生成潜在变量并使用潜在变量预测癌症存在的过程。过程100包括:(i)第一阶段102,用于在参考样品上使用盲源分离算法生成所述组潜在变量;以及(ii)第二阶段104,用于将新样品的序列数据投影到固定潜在变量上。

[0050] i. 第一阶段

[0051] 在步骤106中,可以从不同的参考样品中获得无细胞DNA分子(或称为“cfDNA”)。cfDNA可由多种生物过程产生,诸如细胞凋亡和坏死。这些来自不同来源的cfDNA可能具有不同的片段大小分布,可以通过BSS算法进行分离。在一些情况下,参考样品包括从诊断患有疾病(例如癌症)的受试者获得的生物样品(例如血清或血浆样品)。因此,针对上述参考样品确定的潜在变量可用作与从另一个受试者的另一个生物样品确定的序列数据的大小分布进行比较的潜在变量(或“固定潜在变量”)。

[0052] 在步骤108中,可以确定无细胞DNA分子的大小分布。可以针对参考基因组的N个基因组区域的每个基因组区域鉴定一组序列大小值。所述组序列大小值中的每个序列大小值可以识别与基因组区域比对的相应序列读数的大小(例如,DNA衔接子序列之间的配对末端读数片段插入大小,片段长度单位为bp)。然后,对于每个参考样品,可以为这些组序列大小值生成片段大小分布。分析中仅使用50与550bp之间的序列大小值。在一些情况下,为每组序列大小值生成PMF来确定大小分布,其中所述组中的每个序列大小值都通过所述组中的序列总数进行归一化。

[0053] 在步骤110中,可以通过盲源分离算法处理序列大小值的大小分布以生成一组潜在变量。在一些情况下,所述组PMF被用作一个或多个盲源分离算法的输入数据(例如,包含 $N \times 501$ 维的矩阵)。在此实例中,BSS算法包括独立成分分析(ICA)算法(例如,fastICA)或非负矩阵分解(NMF)算法。根据BSS算法的类型,对输入数据进行额外的格式化。例如,输入到非负矩阵分解算法中的X矩阵是本文所述的原始PMF。在另一个实例中,输入到独立成分分析算法中的X矩阵首先进行均值中心化,然后缩放为单位方差。BSS算法执行多次,直到随机初始状态实现收敛(例如,达到解,满足最小化或最大化标准,ICA的最大化非高斯性)。

[0054] 对每个样品进行BSS之后,可以使用聚类算法(例如,K均值聚类算法)将来自J个参考样品的潜在变量形成保守聚类。可以选择每个聚类的质心或中心点,所述质心或中心点可以用作可以表示来自不同生物过程的cfDNA的固定潜在变量。另外地或替代地,可以针对J个参考样品中的每一个鉴定全基因组的序列大小PMF。这个 $J \times 501$ (50-550bp) 矩阵可用作一个或多个BSS算法的输入。BSS算法的输出潜在变量可以直接用作固定潜在变量。固定潜在变量可以共同用作片段组学特征来预测其他生物样品中的癌症。

[0055] ii. 第二阶段

[0056] 对于第二阶段104,可以从另一个生物样品中获得无细胞DNA分子。另一个生物样品可以从另一个疾病诊断未知的受试者获得。另一个生物样品可以从已接受癌症治疗和手术的受试者获得。在一些情况下,另一个生物样品是从获得参考样品的同一受试者获得的,但是在一段时间的不同时间点获得。可以将其他生物样品与片段组学特征的固定潜在变量进行比较,以预测受试者的癌症分类和/或监测其进展(例如)。在步骤112中,可以确定另一个受试者的cfDNA的大小分布数据。大小分布数据可以包括一个或多个PMF。图1的步骤108描述了用于确定大小分布的过程。

[0057] 在步骤114中,可以将表示另一个受试者的序列大小的大小分布数据投影到片段组学特征的固定潜在变量上,以生成另一个受试者的片段组学特征幅度。为了计算片段组学特征幅度,将表示另一个受试者的大小分布数据的PMF矩阵右乘以潜在变量矩阵的倒数。根据用于潜在变量生成的BSS算法,可以对输入PMF进行额外的格式化,诸如缩放、白化。另

外地或替代地,相似性测量(例如余弦相似性或Pearson相关性)也可用于测量片段大小PMF中每个固定潜在变量的富集。在一些情况下,另一个样品的片段大小PMF是由序列大小的大小分布生成的,此时PMF可以与k个固定潜在变量进行比较或投影到其上。在一些情况下,选择 $1 \times k$ 个特征幅度向量作为PMF的低维表示,用于下游分析。

[0058] 在步骤116中,可以进行下游分析,以基于另一个受试者的片段组学特征幅度来预测所述另一个受试者是否患有癌症。在一些情况下,使用潜在变量作为特征训练的机器学习模型被配置为执行一个或多个预测任务,包括:(i) 预测另一个受试者是否携带癌症相关基因突变;(ii) 预测另一个受试者是否患有疾病(例如,癌症);(iii) 预测另一个受试者是否患有特定类型的疾病(例如,肝癌);(iv) 预测疾病的阶段(例如,IV期癌症);以及(v) 预测另一个受试者是否已响应于特定治疗从疾病中康复。

[0059] 在一些实施方案中,机器学习模型包括一个以上的模型(例如,1、2、3、4、5、6、7、8、9或10个机器学习模型)。在一些情况下,经过训练的机器学习模型包括深度神经网络。深度神经网络可用于捕获越来越大的高维数据集(例如核酸序列数据)的内部结构。深度神经网络可以识别高级特征,提高传统统计模型的性能,增强可解释性,并提供对核酸序列数据结构的额外理解。

[0060] 其他类型的机器学习模型可以包括梯度提升决策树(例如XGBoost框架、LightGBM框架)、装袋过程、提升过程、支持向量机和/或随机森林算法中的一个或多个。例如,梯度提升可以对应于一种机器学习技术,所述技术可用于回归和分类问题,以及用于产生可以包括弱预测模型集合(例如决策树)的预测模型。在一些情况下,梯度提升决策树可以包括例如XGBoost框架或LightGBM框架。

[0061] 机器学习模型可以包括超参数。超参数可以是模型外部的配置,其值不能从数据(例如训练数据、输入数据)中估计。在一些情况下,超参数被调整,例如,被调整以解决给定的预测建模问题。在一些情况下,超参数用于帮助估计模型参数。超参数可以由用户指定。在一些情况下,可以使用一组启发式算法来确定超参数。

[0062] B. 基于从核酸分子的大小分布确定的片段组学特征预测受试者疾病分类的方法

[0063] 图2包括根据一些实施方案的流程图200,其示出了基于生物样品的核酸分子的大小分布确定受试者的片段组学特征的方法的实例。流程图200中描述的一些操作可以由计算机系统(例如,图17的计算机系统1700)执行。尽管流程图200可以将操作描述为顺序过程,但是在各种实施方案中,许多操作可以并行或同时执行。此外,操作的顺序可能会重新排列。操作可能具有图中未显示的额外步骤。此外,所述方法的一些实施方案可以通过硬件、软件、固件、中间件、微代码、硬件描述语言或其任何组合来实现。当在软件、固件、中间件或微代码中实现时,执行相关任务的程序代码或代码段可以存储在计算机可读介质(诸如存储介质)中。

[0064] 在步骤202中,可以访问受试者的生物样品的序列数据。在一些情况下,序列数据对应于生物样品的多个无细胞DNA分子,并且多个无细胞DNA分子包括循环肿瘤DNA分子。序列数据还可以包括与从生物样品检测到的多个体细胞变异相对应的序列。

[0065] 在步骤204中,基于序列数据,可以生成一组序列大小值(例如,序列大小值的二维矩阵)。对于与一组基因组区域的相应基因组区域比对的序列数据的每个序列,序列大小值可以包括与所述序列的大小相对应的序列大小值。由相应的序列大小值表示的每个序列包

括具有一定范围内的大小的DNA片段。在一些情况下,使用参考序列(例如,人参考基因组)来鉴定所述组基因组区域。所述组序列大小值可以是基于与相应基因组区域比对的序列数据的序列生成的经验概率质量函数(PMF)。在一些情况下,为所述组序列大小值生成PMF来确定序列数据的大小分布,其中所述组中的每个序列大小值都通过生物样品中的序列总数进行归一化。

[0066] 在步骤206中,可以将所述组序列大小值与片段组学特征的潜在变量进行比较或投影到其上。片段组学特征可以包括一个或多个可以预测疾病分类的核酸分子的大小分布特征。片段组学特征可以包括一个或多个固定潜在变量。一个或多个固定潜在变量可以通过以下方式确定:(i) 将一个或多个信号分离算法应用于从一个或多个参考样品获得的序列大小值以生成一组潜在变量;以及(ii) 应用聚类算法来选择所述组潜在变量的子集,其中所述子集对应于固定潜在变量。在一些情况下,参考样品包括从诊断患有疾病(例如癌症)的受试者获得的生物样品(例如组织、血浆样品)。另外地或替代地,参考样品可以包括从同一受试者但在另一时间点获得的生物样品。在一些情况下,所述组潜在变量中的每个潜在变量包括直方图或权重向量,其表示由不同生物过程产生的核酸分子的估计大小分布。一个或多个信号分离算法可以是盲源分离算法,其可以包括独立成分分析算法和/或非负矩阵分解算法。在一些情况下,将聚类算法应用于从不同参考样品生成的潜在变量,以确定所述组潜在变量的子集。潜在变量的子集可以包括从已鉴定的潜在变量聚类中选择的质心或中心点。

[0067] 另外地或替代地,可以通过应用变换(诸如缩放、平移、平均、频域转换(例如,快速傅立叶变换)和其他类似变换)来确定潜在变量的导数。

[0068] 在一些情况下,鉴于BSS算法具有估计物理上分离的实体所携带的信号的记录,所述组潜在变量可以预测结合和保护无细胞DNA的蛋白质分子组(例如,单核小体、双核小体、单染色体、双染色体和转录因子复合物)。在一些情况下,潜在变量可以基于其相关的DNA片段组学特征预测潜在的新型核酸和/或蛋白质实体及其相应的结构。

[0069] 此外,对于与表观遗传状态相对应的潜在变量,密度特征诸如方差和中心趋势(例如峰值序列大小)可以预测结合DNA的细胞间异质性程度,从而预测实现表观遗传状态的蛋白质。此外,可以通过比较多峰潜在变量的峰值之间的大小差异来预测潜在变量的不同蛋白质的相对间距和平均接头DNA大小。此外,可以通过测量多峰潜在变量的短和长(例如,分别与单核小体和双核小体相关)大小密度的相对比例,以及沿潜在变量空间中曲线长序列到短序列大小的变化率来预测无细胞DNA降解动力学。

[0070] 在步骤208中,可以基于将所述组序列大小值(例如,大小分布数据)投影到片段组学特征来确定生物样品的一个或多个片段组学特征幅度。可以通过将大小分布数据(例如PMF)投影到片段组学特征的潜在变量子集来确定片段组学特征幅度。

[0071] 在步骤210中,片段组学特征幅度可用作机器学习算法(例如逻辑回归分类器模型)的输入以生成结果。结果可以预测受试者是否患有特定疾病的分类。特定疾病可以包括癌症。在一些情况下,如果参考样品是从同一受试者但在不同时间点获得的,则结果可以预测特定疾病的进展或复发。所述结果可用于鉴定针对受试者的治疗和/或确定应多久对受试者实施治疗。在另一个实例中,基于确定与等位基因相关的单染色体和双染色体序列的显著富集的算法,结果可以预测感兴趣的等位基因处染色体结合的存在和沉默。因此,通过

利用由受试者的全外显子组测序或某些基因组区域生成的无细胞序列数据(例如),可以将所述基因组区域的每个基因组区域建模为等位基因表观遗传状态(例如,转录或抑制基因座)的线性混合物,其各自具有不同的DNA序列大小特征和幅度。

[0072] 在步骤214中,可以输出结果。例如,结果可以在本地呈现或传输到另一个装置。结果可以与受试者的标识符一起输出。此后过程200终止。

[0073] II. 序列数据

[0074] A. 受试者和样品

[0075] 为了基于受试者的生物样品的核酸分子的大小确定片段组学特征,可以从受试者的生物样品中获得代表多个核酸分子的核酸序列数据。受试者可以是人。受试者可能是男性或女性。受试者可能是胎儿、婴儿、儿童、青少年、少年或成年人。受试者可能是任何年龄的患者。例如,受试者可能是小于约10岁的患者。例如,受试者可能是至少约0、5、10、20、30、40、50、60、70、80、90或100岁的患者。受试者可以是正在接受治疗方案或正在评估治疗方案(例如,癌症治疗)的患者或其他个体。然而,在一些情况下,受试者并未接受治疗方案。

[0076] 在一些情况下,受试者可能是哺乳动物或非哺乳动物。在一些情况下,受试者是哺乳动物,诸如人、非人灵长类动物(例如,猿、猴子、黑猩猩)、猫、狗、兔、山羊、马、牛、猪、啮齿动物、小鼠、SCID小鼠、大鼠、豚鼠或绵羊。在一些实施方案中,这些基因的物种变体或同源物用于非人动物模型。物种变体可能是彼此具有最大序列同一性和功能特性相似性的不同物种的基因。许多此类物种变体人基因可能在Swiss-Prot数据库中列出。

[0077] 某些实施方案可以包括从受试者(诸如人受试者)获得样品。在一些情况下,获得患者的临床试样。例如,可以从患者抽取血液。某些实施方案可以包括特异性检测、分析或定量生物样品内的分子(例如,核酸、DNA、RNA等)。

[0078] 样品可以是组织样品或体液。在一些情况下,样品是组织样品或器官样品,诸如活检。在一些情况下,样品包括癌细胞。在一些情况下,样品包括癌细胞和正常细胞。在一些情况下,样品是肿瘤活检。体液可以是汗液、唾液、泪液、尿液、血液、月经、精液和/或脊髓液。在一些情况下,样品是血液样品。样品可能包括一个或多个外周血淋巴细胞。样品可能是全血样品。血液样品可能是外周血样品。在一些情况下,样品包括外周血单核细胞(PBMC);在一些情况下,样品包括外周血淋巴细胞(PBL)。样品可能是血清样品。样品可能是血浆样品。

[0079] 可以使用任何可提供适合于本文所述分析方法的样品的方法来获得样品。样品可以通过非侵入性方法获得,诸如咽拭子、口腔拭子、支气管灌洗、尿液收集、皮肤或宫颈刮擦、面颊擦拭、唾液收集、粪便收集、月经收集或精液收集。样品可以通过诸如抽血的微创方法获得。样品可以通过静脉穿刺获得。在其他情况下,样品是通过侵入性程序获得的,包括但不限于:活检、肺泡或肺灌洗或针吸。活检方法可以包括手术活检、切口活检、切除活检、穿刺活检、刮取活检或皮肤活检。样品可以是福尔马林固定的切片。针吸活检的方法还可以包括细针抽吸、芯针活检、真空辅助活检或大芯活检。在一些情况下,可通过本文的方法获得多个样品以确保足够量的生物材料。在一些情况下,样品不是通过活检获得的。

[0080] B. 生成序列数据

[0081] 在一些实施方案中,对样品进行处理以获得核酸序列数据。“核酸”或“核酸分子”可以对应于任意长度的核苷酸的聚合形式,无论是核糖核苷酸、脱氧核糖核苷酸还是肽核酸(PNA),其包括嘌呤和嘧啶碱基,或其他天然的、化学或生物化学修饰的、非天然的或衍生

的核苷酸碱基。多核苷酸的主链可以包括糖和磷酸基团,如通常在RNA或DNA中发现的,或者修饰或取代的糖或磷酸基团。多核苷酸可以包含修饰的核苷酸,诸如甲基化核苷酸和核苷酸类似物。核苷酸序列可以被非核苷酸组分中断。因此,术语核苷、核苷酸、脱氧核苷和脱氧核苷酸通常包括类似物,诸如本文所述的类似物。这些类似物是那些与天然存在的核苷或核苷酸具有某些共同结构特征的分子,使得当掺入核酸或寡核苷酸序列中时,它们允许与溶液中天然存在的核酸序列杂交。通常,这些类似物是通过替换和/或修饰碱基、核糖或磷酸二酯部分而从天然存在的核苷和核苷酸衍生的。可以根据需要定制改变以稳定或不稳定杂交体形成,或者增强与互补核酸序列杂交的特异性。核酸分子可以是DNA分子。核酸分子可以是RNA分子。

[0082] 样品处理包括核酸样品处理和后续的核酸样品测序。可以对部分或全部生物样品进行测序,以提供核酸序列数据,可将其存储或以其他方式维护在电子、磁性或光学存储位置中。可以借助计算机处理器来分析序列信息,并且可以将分析的序列信息存储在电子存储位置中。电子存储位置可以包括从核酸样品生成的序列信息以及分析的序列信息的池或集合。在一些实施方案中,生物样品取自患有或疑似患有癌症的受试者。

[0083] 在一些实施方案中,核酸序列数据从纯肿瘤和纯正常样品生成。可以从另一个来源(例如,美国典型培养物保藏中心)获得匹配对细胞系。每个匹配对可以包括来自同一受试者的肿瘤细胞系和正常细胞系。可以在体外培养和扩增细胞系,以获得用于DNA提取的合适数量的细胞。提取、处理DNA,并进行全外显子组或全基因组测序。序列读数可以进行质量控制处理(例如,通过FastQC)以提供FASTQ文件。

[0084] 在一些情况下,核酸序列数据是利用全基因组测序生成的。在一些情况下,全基因组测序用于识别人类的变异。全基因组测序可以包括对整个或几乎整个基因组的浅测序(1-5倍覆盖率)或深度测序(30-100倍覆盖率)。在一些情况下,测序包括对部分基因组进行测序。例如,部分基因组可以为至少约50;75;100;125;150;175;200;225;250;275;300;350;400;450;500;550;600;650;700;750;800;850;900;950;1,000;1100;1200;1300;1400;1500;1600;1700;1800;1900;2,000;3,000;4,000;5,000;6,000;7,000;8,000;9,000;10,000;15,000;20,000;30,000;40,000;50,000;60,000;70,000;80,000;90,000;100,000或更多个碱基或碱基对。在一些情况下,可以在100万、200万、300万、400万、500万、600万、700万、800万、900万、1000万或超过1000万个碱基或碱基对上进行基因组测序。在一些情况下,可以在整个外显子组上进行基因组测序(例如,全外显子组测序)。在一些情况下,深度测序可以包括在部分基因组上获取多个读数。例如,获取多个读数可以包括在部分基因组上至少2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000、10,000个读数或超过10,000个读数。

[0085] 另外地或替代地,可以使用基于面板的测序来生成核酸序列数据。基于面板的测序可用于同时评估疑似病症、疾病或表型的多种潜在遗传原因。基因面板可用于鉴定与核酸分子序列比对的靶向基因组区域。在一些情况下,靶向基因组区域的数量包括至少50、100、200、500、1000、1500、2000个基因组区域。基于面板的测序可以定义靶向基因组区域的足迹,而不是靶向基因组区域的数量。例如,足迹可能在175Kb与3Gb之间。在其他实施方案中,基于面板的测序可以靶向具有已知生物学意义的多个基因组区域,诸如与癌症驱动基因或肿瘤逃逸基因中的变体相关的基因组区域。

[0086] 在一些情况下,生成核酸序列数据包括通过深度测序检测低等位基因部分。在一些情况下,深度测序是通过下一代测序完成的。在一些情况下,深度测序是通过避开易出错的区域来进行的。在一些情况下,易出错的区域可能包括近序列重复的区域、%GC异常高或低的区域、近均聚物、二核苷酸和三核苷酸的区域以及近其他短重复的区域。在一些情况下,易出错的区域可能包括导致DNA测序错误的区域(例如,均聚物序列中的聚合酶滑动)。

[0087] 在一些情况下,生成核酸序列数据包括对样品中的一个或多个核酸分子进行一次或多次测序反应。某些实施方案可以包括对样品中的一个或多个核酸分子进行1次或更多次、2次或更多次、3次或更多次、4次或更多次、5次或更多次、6次或更多次、7次或更多次、8次或更多次、9次或更多次、10次或更多次、15次或更多次、20次或更多次、30次或更多次、40次或更多次、50次或更多次、60次或更多次、70次或更多次、80次或更多次、90次或更多次、100次或更多次、200次或更多次、300次或更多次、400次或更多次、500次或更多次、600次或更多次、700次或更多次、800次或更多次、900次或更多次或1000次或更多次测序反应。测序反应可以同时、顺序或其组合进行。测序反应可以包括全基因组测序、外显子组测序或小面板靶向测序。测序反应可以包括Maxim-Gilbert、链终止或高通量系统。替代地或另外地,测序反应可以包括Helioscope™单分子测序、Nanopore DNA测序、Lynx Therapeutics的大规模并行特征测序(MPSS)、454焦磷酸测序、单分子实时(RNAP)测序、Illumina(Solexa)测序、SOLiD测序、Ion Torrent™、Ion半导体测序、单分子SMRT(TM)测序、Polony测序、DNA纳米球测序、VisiGen Biotechnologies方法或其组合。替代地或另外地,测序反应可以包括一个或多个测序平台,包括但不限于Illumina提供的Genome Analyzer IIx、HiSeq、MiSeq和NovaSeq、单分子实时(SMRT™)技术,诸如Pacific Biosciences(加利福尼亚州)提供的PacBio RS系统和Solexa测序仪、True单分子测序(tSMSTM)技术,诸如Helicos Inc.(马萨诸塞州剑桥)提供的HeliScope™测序仪。测序反应还可以包括电子显微镜或化学敏感场效应晶体管(chemFET)阵列。在本公开的一些方面,测序反应包括毛细管测序、下一代测序、桑格测序、合成测序、连接测序、杂交测序、单分子测序或其组合。合成测序可以包括可逆终止子测序、进行性单分子测序、顺序流测序或其组合。顺序流测序可以包括焦磷酸测序、pH介导的测序、半导体测序或其组合。

[0088] 在一些情况下,生成核酸序列数据包括进行至少一次长读测序反应和至少一次短读测序反应。长读测序反应和/或短读测序反应可以在核酸分子子集的至少一部分上进行。长读测序反应和/或短读测序反应可以在两个或多个核酸分子子集的至少一部分上进行。长读测序反应和短读测序反应可以在一个或多个核酸分子子集的至少一部分上进行。

[0089] 对一个或多个核酸分子或其子集的测序可以包括至少约5;10;15;20;25;30;35;40;45;50;60;70;80;90;100;200;300;400;500;600;700;800;900;1,000;1,500;2,000;2,500;3,000;3,500;4,000;4,500;5,000;5,500;6,000;6,500;7,000;7,500;8,000;8,500;9,000;9,500;10,000;25,000;50,000;75,000;100,000;250,000;500,000;750,000;1,000,000;2,500,000;5,000,000;10,000,000;25,000,000;50,000,000;100,000,000;250,000,000;500,000,000;750,000,000;1,000,000,000或更多个测序读数。

[0090] 测序反应可以包括测序一个或多个核酸分子的至少约50;60;70;80;90;100;110;120;130;140;150;160;170;180;190;200;210;220;230;240;250;260;270;280;290;300;325;350;375;400;425;450;475;500;600;700;800;900;1,000;1,500;2,000;2,500;3,

000;3,500;4,000;4,500;5,000;5,500;6,000;6,500;7,000;7,500;8,000;8,500;9,000;10,000;20,000;30,000;40,000;50,000;60,000;70,000;80,000;90,000;100,000个或更多个碱基或碱基对。测序反应可以包括测序一个或多个核酸分子的至少约50;60;70;80;90;100;110;120;130;140;150;160;170;180;190;200;210;220;230;240;250;260;270;280;290;300;325;350;375;400;425;450;475;500;600;700;800;900;1,000;1,500;2,000;2,500;3,000;3,500;4,000;4,500;5,000;5,500;6,000;6,500;7,000;7,500;8,000;8,500;9,000;10,000;20,000;30,000;40,000;50,000;60,000;70,000;80,000;90,000;100,000个或更多个连续碱基或碱基对。

[0091] 在一些情况下,测序技术每次运行产生至少100个读数、每次运行至少200个读数、每次运行至少300个读数、每次运行至少400个读数、每次运行至少500个读数、每次运行至少600个读数、每次运行至少700个读数、每次运行至少800个读数、每次运行至少900个读数、每次运行至少1,000个读数、每次运行至少5,000个读数、每次运行至少10,000个读数、每次运行至少50,000个读数、每次运行至少100,000个读数、每次运行至少500,000个读数或每次运行至少1,000,000个读数。替代地,测序技术每次运行产生至少1,500,000个读数、每次运行至少2,000,000个读数、每次运行至少2,500,000个读数、每次运行至少3,000,000个读数、每次运行至少3,500,000个读数、每次运行至少4,000,000个读数、每次运行至少4,500,000个读数或每次运行至少5,000,000个读数。

[0092] 在一些情况下,测序技术每次读数产生至少约30个碱基对、至少约40个碱基对、至少约50个碱基对、至少约60个碱基对、至少约70个碱基对、至少约80个碱基对、至少约90个碱基对、至少约100个碱基对、至少约110个碱基对、至少约120个碱基对、每次读数至少约150个碱基对、至少约200个碱基对、至少约250个碱基对、至少约300个碱基对、至少约350个碱基对、至少约400个碱基对、至少约450个碱基对、至少约500个碱基对、至少约550个碱基对、至少约600个碱基对、至少约700个碱基对、至少约800个碱基对、至少约900个碱基对或至少约1,000个碱基对。另外地或替代地,测序技术可以产生长测序读数。在一些情况下,测序技术可以每次读数产生至少约1,200个碱基对、每次读数至少约1,500个碱基对、每次读数至少约1,800个碱基对、每次读数至少约2,000个碱基对、每次读数至少约2,500个碱基对、每次读数至少约3,000个碱基对、每次读数至少约3,500个碱基对、每次读数至少约4,000个碱基对、每次读数至少约4,500个碱基对、每次读数至少约5,000个碱基对、每次读数至少约6,000个碱基对、每次读数至少约7,000个碱基对、每次读数至少约8,000个碱基对、每次读数至少约9,000个碱基对、每次读数至少约10,000个碱基对、每次读数至少约20,000个碱基对、每次读数至少约30,000个碱基对、每次读数至少约40,000个碱基对、每次读数至少约50,000个碱基对、每次读数至少约60,000个碱基对、每次读数至少约70,000个碱基对、每次读数至少约80,000个碱基对、每次读数至少约90,000个碱基对或每次读数至少约100,000个碱基对。

[0093] 高通量测序系统可允许在测序核苷酸掺入生长链之后或掺入生长链时立即检测,即实时或基本上实时检测序列。在一些情况下,高通量测序每小时产生至少1,000、至少5,000、至少10,000、至少20,000、至少30,000、至少40,000、至少50,000、至少100,000或至少500,000个序列读数;其中每个读数为每次读数至少50、至少60、至少70、至少80、至少90、至少100、至少120、至少150、至少200、至少250、至少300、至少350、至少400、至少450或至少

500个碱基。可以使用本文所述的核酸诸如基因组DNA、mtDNA、源自RNA转录物的cDNA或RNA作为模板进行测序。

[0094] III. 用于确定潜在变量的输入数据

[0095] 片段组学特征可被定义为包括一个或多个潜在变量。每个潜在变量可以识别在所述组基因组区域中可变地富集的序列的估计大小和/或末端基序分布。潜在变量可以表示序列大小和/或末端基序数据的新的独立基向量,然后用于确定受试者的片段组学特征的富集。

[0096] A. 序列大小值

[0097] 图3显示了根据一些实施方案的示意图300,所述示意图说明用于生成序列大小值的实例技术。示意图300显示与参考序列310比对的序列数据305。在一些情况下,参考序列310是人参考基因组(例如,hg19基因组)。参考序列310可分为一组基因组区域,包括基因组区域315。在图3所示的实例中,序列数据305的一组序列可以与基因组区域315比对。对于基因组区域315,可以识别所述组序列的大小分布。

[0098] 基于序列数据305,可以生成多组序列大小值320。对于每个基因组区域,每组序列大小值可以包括与相应序列组的大小相对应的序列大小值。一组序列大小值可以对应于矩形内的一组整数。在一些情况下,使用全基因组的一组序列大小值。可以从与一组基因组区域(例如,已知体细胞变异与癌症相关的基因组区域)比对的序列确定所述组序列大小值。

[0099] 在一些情况下,由相应的序列大小值表示的每个序列包括大小在50bp与550bp之间的DNA片段。分析中未使用50与550bp之间的范围之外的序列大小值(例如,SAM/BAM文件中的模板片段长度值)。

[0100] 除了这些组序列大小值之外或作为其替代,每组序列大小值可用于生成PMF 325。PMF可以指一种函数,所述函数给出离散随机变量(例如,序列大小值)等于特定范围内的某个值(例如,167bp)的概率。

[0101] 在一些实施方案中,对每组序列大小值生成PMF 325,其中所述组的每个序列大小值以1bp分辨率分箱,然后通过总读数对计数进行归一化。所述组PMF用作BSS算法的输入数据(例如,包含 $N \times 501$ 维的矩阵),以识别可预测特定疾病的一个或多个成分。

[0102] B. 末端基序

[0103] 末端基序可被鉴定为DNA片段的核苷酸的末端序列,例如片段任一端的K个碱基的序列。末端序列可以是具有不同碱基数(例如1、2、3、4、5、6、7)的k聚体。在一些情况下,通过将序列读数与参考基因组比对并识别起始位置之前或终止位置之后的核苷酸碱基来确定末端基序。此类碱基仍将对应于DNA片段的末端,例如,因为它们是基于片段的末端序列来识别的。

[0104] 可利用各种技术从比对的序列读数中识别出末端基序。在一些情况下,k聚体末端基序直接由血浆DNA分子每个末端的第一个k-bp序列构建。例如,可以使用测序片段的前4个核苷酸或最后4个核苷酸。在另一个实例中,通过使用来自片段测序末端的(k-2)聚体序列和来自与所述片段末端相邻的基因组区域的另一个(k-2)聚体序列联合构建k聚体末端基序。可以使用各种长度的末端基序,例如,1-聚体、2-聚体、3-聚体、5-聚体、6-聚体、7-聚体末端基序。序列读数可以与末端基序比对,从而生成一组末端基序序列数据。所述组末端基序序列数据可应用于一个或多个信号分离算法,以生成相应的一组潜在变量。

[0105] 无细胞DNA末端特征中包含的核苷酸数量越多,基序的特异性越高。例如,基因组中6个碱基以精确配置排列的概率低于基因组中2个碱基以精确配置排列的概率。因此,末端基序的长度选择可由预期用途应用所需的敏感性和/或特异性决定。

[0106] IV. 使用信号分离算法确定片段组学特征的潜在变量

[0107] 为了确定片段组学特征的潜在变量,可以将一个或多个信号分离算法应用于多组序列大小值和/或末端基序序列数据。在一些情况下,所述组潜在变量中的每个潜在变量包括直方图或权重向量,其表示生物样品的一部分序列数据背后的大小分布。一个或多个信号分离算法可以是盲源分离算法,其可以使用独立成分分析算法和/或非负矩阵分解算法。

[0108] A. 信号分离算法

[0109] 信号分离算法可被配置并用于从一组观测信号中估计一组源信号,其中每个观测信号是源信号的混合。图4显示了实例图400,其说明了将信号分离算法应用于由线性求和图像源信号制成的图像混合组。在图4中,基于一组图像信号410的混合生成输入图像405。每个图像信号 S_i 可以按随机系数 k_i 进行缩放,以在图像混合概要中产生源图像幅度分集。至少一个图像信号可以包括每个图像混合唯一的随机噪声信号。在图4中,基于一组四个图像信号410的混合构建的两千个输入图像405可以用以下表达式表示:

$$[0110] \quad \sum_{i=1}^4 k_i s_i + 10s_{\text{噪声}}$$

[0111] 可以使用一个或多个信号分离算法来处理输入图像405,以输出在视觉上近似于所述组图像信号410所表示的图像的一组图像415。信号分离算法可以是盲源分离算法。盲源分离算法被称为“盲”,因为所述算法可以在不直接观察的情况下,而是通过观察源信号对可测量基本量的影响来估计源信号。

[0112] 盲源分离算法可以包括通过奇异值分解(SVD)算法420或ICA算法430(例如,fastICA、InfoMax)实现的主成分分析(PCA)。SVD算法420和ICA算法430都可以在一组数据上进行训练,并且可以将它们的输出应用于新的数据组。如本文所用,SVD算法420可以指从较大数据组中识别较少数量的称为主成分的不相关变量的技术。所述技术被广泛用于强调变化并捕获数据组(例如,对应于输入图像405的一组像素)中的强模式。在图4中,一组实例未混合图像425可以显示为通过将SVD算法420应用于整个组的两千个图像混合405而生成的输出。在一些情况下,在应用SVD算法420之前对输入图像405进行预处理。

[0113] ICA算法430可以包括用于寻找非高斯性的最大值的定点迭代方案,其中ICA算法可以迭代执行直到达到最大非高斯值。在一些情况下,在应用ICA算法430之前对输入图像405进行预处理,包括零中心化和缩放到单位方差。如图4所示,未混合图像的实例组435和440可以显示为通过将ICA算法430应用于不同量的输入图像405而生成的输出。第一组未混合图像435的视觉特征似乎与图像信号组410不匹配,这可能是由于使用了具有相对低的源图像幅度分集的低数量(10)的图像混合(I_1-I_{10})。相比之下,第二组图像440的视觉特征似乎与相应组的图像信号相匹配,这可能是由于使用了具有丰富的源图像幅度分集的大量(2000)的图像混合(I_1-I_{2000})。此外,虽然图4中未示出,但盲源分离算法可以包括NMF算法。

[0114] B. 确定片段组学特征的潜在变量的实例方案

[0115] 图5显示了根据一些实施方案的示意图500,其说明用于生成一组潜在变量的实例

技术。如图5所示,多组序列大小值505中的每一组均可用作BSS算法510的输入。BSS算法510可以包括ICA算法515或NMF算法520。根据BSS算法的类型,可以对多组序列大小值505进行额外格式化。例如,NMF算法520可以使用与序列大小值组相对应的原始PMF。在另一个实例中,ICA算法515可以使用首先进行均值中心化然后缩放到单位方差的PMF。在一些情况下,BSS算法510执行多次,直到随机初始状态实现收敛(例如,达到解,满足最小化或最大化标准,ICA的最大化非高斯性)。重复执行BSS算法以确保从多次运行的可重复性中捕获良好估计的潜在变量。

[0116] 尽管图5的实例技术讨论了从序列大小值组生成一组潜在变量,但是BSS算法510也可应用于一组末端基序序列数据以生成另一组潜在变量。在这种情况下,另一组潜在变量识别序列数据的末端基序分布。例如,对于核苷酸的每个k聚体末端基序(例如,CCCA、TAAA),末端基序分布可以包括具有与k聚体末端基序相对应的末端序列的DNA片段的数量或相对频率。通过将序列数据的末端基序分布分离为一组独立的权重向量,可以确定片段组学特征。

[0117] 另外地或替代地,BSS算法510还可应用于生成潜在变量组,使得每个潜在变量可识别具有与相应末端基序相对应的末端序列的序列读数的估计大小分布。

[0118] ICA算法515的输入是数据矩阵X,其输出包括:(i) S矩阵,其包括与所述组潜在变量相对应的直方图或有符号权重向量数据;以及(ii) A线性混合矩阵,其包括所述组基因组区域中潜在变量的估计幅度。类似地,NMF算法520的输入是数据矩阵Y,其输出包括:(i) W矩阵,其包括与所述组潜在变量相对应的非负直方图数据;以及(ii) H矩阵,其包括所述组基因组区域中潜在变量的估计幅度。这些数据矩阵可用于生成一组直方图或权重向量,其可表示为相应片段组学特征的潜在变量525。

[0119] 继续图5中所示的实例,潜在变量525对应于通过将BSS算法510应用于多组序列大小值505而生成的权重向量或直方图。每个潜在变量525中识别的峰值和幅度可以揭示生物样品的核酸分子的大小和/或序列数据中序列的序列片段化模式。潜在变量525可以表示受试者的片段组学特征。

[0120] V. 使用片段组学特征的潜在变量检测癌症相关基因突变

[0121] 许多癌症是由复发性基因突变引起的,诸如乳腺癌的BRCA和HER2突变。因此,准确检测这些癌症相关基因突变是癌症患者进行靶向治疗的关键。然而,NGS实验的各个步骤都会产生噪声,诸如PCR错误或测序错误。这些噪声对于cfDNA样品来说尤其成问题,因为真实突变的等位基因频率很低。先前的研究表明,来自癌症组织的cfDNA分子与来自健康组织的cfDNA分子相比具有不同的片段大小分布。通过研究突变支持读数的片段组学特征,我们可以确认观察到的突变是否与癌症相关或与噪声相关。

[0122] A. 从一组cfDNA全外显子组测序(WES)样品中生成片段组学特征

[0123] 图6显示了根据一些实施方案的示意图600,其说明了使用独立成分分析算法为参考特征生成一组潜在变量的实例技术。如图6所示,可以收集一组cfDNA样品605(n=10)。所述组cfDNA样品605可以包括4个以前缀“CRC”识别的结直肠癌样品cfDNA和6个以前缀“PON”识别的健康供体样品。可以使用全外显子组测序(WES)对所述组生物样品605进行测序以生成序列数据,从中可以生成序列大小值组。

[0124] 将序列数据分为309个均匀区间,可生成每个区域的片段大小分布。然后将ICA算

法应用于每个样品610的309x 541 (60-600bp) 矩阵,以生成对应于一组潜在变量的S矩阵。为了从所述组中选择最相关的潜在变量,可以将聚类算法615 (例如,k均值聚类算法) 应用于来自所有样品的所述组潜在变量,其中可以基于相似性将所述组的潜在变量聚类在一起。对于每个聚类,可以生成潜在变量质心。对于每个受试者,与聚类质心最相似的潜在变量可用于检测癌症突变。例如,可以从健康的cfDNA样品中选择与聚类质心相似的潜在变量620。在另一个实例中,可以从与结直肠癌相关的cfDNA样品中选择类似于聚类质心的潜在变量625。

[0125] B. 比较健康供体与癌症患者之间的潜在变量

[0126] 回顾图6,将正常生物样品的核酸分子的潜在变量620与具有肿瘤DNA (例如,TP53 肿瘤变体) 的生物样品的核酸分子的潜在变量625进行比较。潜在变量620与潜在变量625之间的差异可用于预测癌症相关基因突变。

[0127] 继续图6中的实例,比较可以包括将所述组潜在变量620中的每一个中所示的峰值的序列大小值与所述组潜在变量625的相应潜在变量中所示的峰值的序列大小值进行比较。在一个实例中,所述组潜在变量625 (结直肠癌样品) 中显示的峰值的序列大小值比所述组潜在变量620 (健康样品) 中显示的少大约10-30个碱基对。因此,相对较短的DNA分子的丰度可以预测受试者是否患有癌症。在一些情况下,计算来自癌症样品的潜在变量的峰值的序列大小值与来自健康样品的相应潜在变量的峰值的序列大小值之间的差异,以预测疾病的特定阶段。例如,如果差异指示相对小的值 (例如小于10bp), 则可以预测癌症处于早期阶段。相反,如果差异指示相对大的值 (例如大于30bp), 则可以预测癌症处于晚期。

[0128] 另外地或替代地,比较可以包括对每个潜在变量确定预定大小范围内的序列大小值的密度。可以将序列大小值的密度与预定阈值进行比较,以预测受试者是否患有特定疾病。例如,可以计算200-300bp大小范围内潜在变量的序列大小值的密度,并且可以将所述密度与阈值 (例如,2) 进行比较,以预测受试者是否患有癌症。

[0129] 另外地或替代地,潜在变量的比较可以包括比较所述组潜在变量620中每个潜在变量的幅度值和所述组潜在变量625中相应潜在变量的幅度值。在一些情况下,将每对幅度值之间的差异与阈值进行比较。

[0130] C. 单个样品的片段组学特征的实例

[0131] 图7-10说明了根据一些实施方案的单个样品的片段组学特征的实例组。在每个图中,我们强调了在样品和BSS算法的多次运行中重复识别的潜在变量。图7显示了通过对与正常受试者的生物样品相对应的序列大小值应用信号分离算法而生成的一组潜在变量700。可以从所述组潜在变量700中选择潜在变量子集705,其中所述潜在变量子集705识别代表相应生物样品片段组学特征的序列的大小分布。在一些情况下,基于潜在变量是否包括峰值宽度值大于特定阈值的峰值来选择子集705的潜在变量。另外地或替代地,可以通过将聚类算法应用于潜在变量以生成一组聚类并选择与所述组的特定聚类的质心相对应的潜在变量来选择子集705的潜在变量。

[0132] 图8显示了通过对与另一个正常受试者的生物样品相对应的序列大小值应用信号分离算法而生成的一组潜在变量800。可以从所述组潜在变量800中选择潜在变量子集805,其中所述潜在变量子集805识别代表相应生物样品片段组学特征的序列的大小分布。如图8所示,潜在变量子集805包括与图7的片段组学特征705的潜在变量具有相似性的潜在变量

子集。

[0133] 图9显示了通过对与被诊断为结直肠癌的受试者的生物样品相对应的序列大小值应用信号分离算法而生成的一组潜在变量900。可以从所述组潜在变量900中选择潜在变量子集905,其中所述潜在变量子集905识别代表相应生物样品片段组学特征的序列的大小分布。图10显示了通过对与被诊断为结直肠癌的另一个受试者的生物样品相对应的序列大小值应用信号分离算法而生成的一组潜在变量1000。潜在变量子集1005可以从所述组潜在变量1000中选择,其中潜在变量子集1005识别代表相应生物样品片段组学特征的序列的大小分布。

[0134] D. 影响潜在变量的生物过程

[0135] 在一些情况下,潜在变量的形状可用于预测无细胞DNA序列大小数据中蛋白质分子组(例如,单核小体、双核小体、单染色体、双染色体和转录因子复合物)的富集,从而基于将潜在变量作为代表控制染色质表观遗传状态的独立调控分子的统计对象,预测所述分子组在体内相应细胞基因座处的结合。在一些情况下,潜在变量还用于基于其片段长度模式预测潜在的新型核酸和蛋白质实体及其相应的结构。

[0136] 例如,序列数据中已知与染色体相关的潜在变量的富集可用于预测染色体的结合和相应细胞基因座处沉默异染色质状态的富集。在另一个实例中,序列数据中与癌症相关的潜在变量的富集可用于预测特定等位基因是否来自癌细胞。在又一个实例中,序列数据中与基因表达相关的潜在变量的富集可用于预测对测序数据有贡献的特定等位基因是否正在表达。另外地,或替代地,可以通过比较双峰潜在变量密度峰值,并从潜在变量空间的变化率推断无细胞DNA降解动力学来预测分子组内较短和较长尺寸类型的相对无细胞DNA丰度。

[0137] 图11显示了示意图1100,其说明了根据一些实施方案的用于预测序列大小数据中蛋白质分子组的富集的技术。

[0138] 一组潜在变量1105可以包括潜在变量1110a-e,每个潜在变量提供生物样品中核酸分子的估计大小分布。对于每个潜在变量,可以识别一组峰。所述组峰中的每个峰可用于预测特定蛋白质的结合。例如,潜在变量1110a可以表示为在序列大小约150bp处具有峰1115a的单峰大小分布。当在与基因组位置相关的序列大小数据中富集时,这种大小分布的峰值可以预测核小体结合的位置。除峰1115a之外,具有约10bp峰宽的峰的周期性模式可以通过考虑额外的DNA降解信号的影响来预测核小体结合,所述信号可能与核小体结合的DNA螺旋间距相关的离散片段的逐渐消化有关。这些峰可以与核酸分子在细胞凋亡或坏死细胞死亡过程中的行为以及随后脱落到和穿过循环系统的行为相关。例如,蛋白质结合的DNA分子(通常是与组蛋白或转录因子相关的分子)优先在损伤(例如消化)后存活下来并释放到血液循环中,而未结合的DNA分子则丢失。

[0139] 继续图11中的实例,一些潜在变量可以预测特定蛋白质的结合。例如,潜在变量1110e包括峰1120a-b,它们一起表示为多峰分布,并且当在与转录基因座相关的序列大小数据中富集时,峰1120a-b中每一个可用于预测转录因子和转录因子-双核小体复合物结合。此外,潜在变量1110e包括具有较大峰值宽度值的第三峰1120c。这种峰可用于预测转录因子-单核小体结合。除了峰1120a-b中每一个的宽度之外,可能指纹识别CTCF的约10bp的叠加周期性模式(例如,ATAC-seq数据中所见)可以预测控制转录基因座表观遗传状态的

DNA结合蛋白的细胞间异质性。

[0140] 图11中描述的技术可以应用于不同的一组潜在变量1125。预测的蛋白质结合可用于定义整个基因组的核小体足迹模式,然后可用于预测受试者是否患有癌症。

[0141] E. 使用癌症衍生的潜在变量来调用癌症相关的基因突变

[0142] 片段组学特征的潜在变量可用作预测生物样品的核酸分子是否携带癌症相关突变的特征。例如,TP53基因编码调节细胞分裂的肿瘤抑制蛋白。它是人类癌症中最常见的突变基因之一。在CRC10063cfDNA样品中鉴定出TP53_chr17_7578265_A>C突变。在配对的肿瘤样品中也发现了相同的突变,而在匹配的邻近健康组织或白细胞正常样品中没有发现,这表明来自cfDNA的突变支持读数来自肿瘤细胞。这些TP53突变支持读数的序列大小分布可以投影到固定的潜在变量上,从而得到1x 5向量,来表示序列大小数据中每个潜在变量的幅度或富集。作为阴性对照,使用高质量野生型等位基因支持读数进行相同的ICA投影。这两个向量可以提供一种量化片段大小差异的方法。

[0143] 图12显示了示意图1200,其说明根据一些实施方案,通过将原始序列大小分布投影到潜在变量上来预处理,将片段大小分布转换为一组幅度,并使用所述幅度检测癌症相关基因突变的过程。作为初始步骤,可以将BSS应用于从癌症参考样品获得的序列数据,并选择在各种参考样品中重复识别的潜在变量。在步骤1205中,在cfDNA样品中调用候选突变后,可以将突变(ALT等位基因)支持读数与野生型等位基因(REF等位基因)支持读数分离。在步骤1210中,ALT等位基因和REF等位基因的序列大小分布可以转换为它们各自的PMF并投影到固定的潜在变量上。得到的潜在变量幅度可用于衡量突变调用的质量。例如,与来自肿瘤组织的突变支持读数对应的潜在变量的幅度应不同于与野生型支持读数对应的潜在变量的幅度,而来自实验噪声或错误的突变读数的幅度应所述与野生型读数更相似。

[0144] 可以表明,潜在变量可以在血浆样品中循环肿瘤DNA(ctDNA)的序列大小分布中高度富集。例如,潜在变量包括通过将ICA算法应用于晚期结直肠癌患者的序列大小数据而生成的权重向量。已知来自匹配的肿瘤活检样品的与上述TP53突变相关的血浆cfDNA序列大小数据显示在140-150bp范围内的窄峰和在200-300bp范围内的宽峰,这与潜在变量“LV5”非常相似。

[0145] 为了开发用于过滤候选突变调用的模型,选择了对应于78个具有高质量突变的基因座的序列大小分布并将其投影到固定潜在变量上(参见步骤1210)。同时,使用相同的方法投影相应的高质量野生型读数。绘制每个基因座的LV1和LV5幅度。LV幅度图1215显示,对应于REF等位基因的一组结果一起聚类在第一区域1220中,而对应于ALT等位基因的一组结果大多远离所述聚类。在一些情况下,基于1215所示的绘制结果确定一个或多个截止值。例如,截止值1225为约0.4。任何超过截止值1225的结果都可以被认为具有肿瘤DNA特有的体细胞变异。作为说明性实例,从LV幅度图1215鉴定的结果1230鉴定具有已知肿瘤相关变体(TP53基因变体)的核酸分子的大小分布。

[0146] VI. 使用片段组学特征的潜在变量预测癌症

[0147] 片段组学特征所代表的潜在变量还可用于受试者的癌症早期检测以及治疗后癌症复发的监测。与健康对照受试者相比,癌症患者通常具有不同的cfDNA片段大小分布和末端基序频率。可以将特定受试者的大小分布数据投影到片段组学特征的潜在变量组上,以生成片段组学特征幅度,其中可以使用片段组学特征幅度(例如,通过机器学习)来预测特

定受试者是否患有疾病(例如,癌症)。例如,可以将片段组学特征幅度与健康供体的片段组学特征幅度进行比较,以预测受试者是否有异常。可以基于潜在变量的类型和用于机器学习模型的训练数据的类型来考虑其他类型的预测。如果训练数据包括患有特定阶段疾病的患者的参考样品,则可以使用基于潜在变量的模型来预测特定受试者是否患有所述特定阶段疾病(例如,癌症IV期)。如果训练数据包括患有特定类型疾病的参考样品,则可以使用机器学习模型来预测受试者是否患有所述特定类型疾病(例如,结直肠癌)。

[0148] A. 从一组由肿瘤知情小组富集的cfDNA样品中生成潜在变量cfDNA分析是一种检测治疗后微小残留疾病或癌症复发的新兴方法。然而,其灵敏度受到血浆中低肿瘤信号的限制。以下描述显示了一个实例,其中可以通过个性化杂交捕获来富集肿瘤衍生的DNA,然后可以使用肿瘤富集的样品来推导用于检测受试者癌症的潜在变量。

[0149] 图13显示了示意图,其说明根据一些实施方案,对杂交捕获样品使用独立成分分析算法生成一组潜在变量的实例技术。在图13中,使用了不同肿瘤类型的正常/肿瘤/血浆样品三合一。使用正常和肿瘤样品进行全基因组测序,以检测肿瘤中的体细胞突变。在步骤1305中,使用约1800个具有高等位基因频率的体细胞突变来为每个患者设计个性化的杂交捕获面板。然后可以使用所述肿瘤知情小组来富集来自血浆的肿瘤信号。对于每个面板设计,还使用一个或两个健康血浆样品作为参考进行杂交捕获。

[0150] 为了生成潜在变量,将测序读数映射到参考基因组,并将所有读数保持在50-550bp内。在步骤1310中,为每个血浆样品生成每个目标区域周围的片段大小PMF。在步骤1315中,识别片段的序列大小分布的1800x 501矩阵随后被用作BSS算法(例如ICA)的输入。在此实例中,对每个血浆样品进行ICA,并保留10个潜在变量(独立成分)。

[0151] 一旦从119个健康和101个患者血浆样品中收集到潜在变量,则使用K均值聚类来识别不同样品中的相似潜在变量(步骤1320)。使用不同的参数运行聚类算法,并根据几种聚类评估方法选择了8聚类模型。结果1325显示每个聚类的质心,其用作下游分析的潜在变量(例如,预测其他受试者的癌症)。

[0152] B. 使用潜在变量幅度检测癌症发生或复发

[0153] 以下描述显示了使用潜在变量监测手术后癌症复发的实例。图14显示了示意图,其说明根据一些实施方案的使用手工制作的特征或潜在变量特征监测癌症复发的实例技术。根据肿瘤样品的体细胞突变进行杂交捕获面板设计。在步骤1405中,在手术后的不同时间点收集血浆样品,然后使用杂交捕获技术来富集肿瘤样品中的靶cfDNA。作为参考,还使用健康血浆样品进行杂交捕获。在步骤1410中,为了监测不同时间点的可能癌症复发,可以提取患者和健康样品的片段大小分布并将其投影到8个固定潜在变量上。在步骤1415中,然后可以使用预先训练的机器学习模型来比较患者潜在变量幅度与对照潜在变量幅度,并为不同的时间点分配状态。在步骤1420中,为了将潜在变量方法与其他技术进行比较,还使用13个手工制作的特征来训练基线模型,例如50-250bp内的读取的累积概率、160-170bp内的读取的概率以及250-300bp读取与300-350bp读取的比率。

[0154] 关于机器学习模型,使用43对健康-健康对和59对健康-患者对的数据集来训练两个基于手工制作(步骤1420)或潜在变量特征(步骤1415)的逻辑回归模型。图15显示了一组接收者操作特征(ROC)曲线1500,其显示了使用片段组学特征进行疾病分类的准确度水平。在图15中,图1505和1510显示了识别两个模型性能的接收者操作特征曲线。与基于手工制

作的特征的模型(图1505)相比,基于潜在变量特征的模型(图1510)具有更好的性能:曲线下训练面积从95.8%增加到97%,曲线下交叉验证面积(平均值)从93.8%增加到97%。因此,BSS算法可以有效地用于提取数据驱动的潜在变量特征而无偏差。此外,在癌症检测中,潜在变量可以用来获得比手工制作的特征更好的性能。

[0155] VII. 基于由末端基序频率确定的片段组学特征预测受试者疾病分类的方法

[0156] 图16包括流程图1600,其说明了根据一些实施方案基于生物样品的核酸分子的末端基序频率确定受试者的片段组学特征的方法的实例。流程图1600中描述的一些操作可以由计算机系统执行(例如,图17的计算机系统1700)。尽管流程图1600可以将操作描述为顺序过程,但在各种实施方案中,许多操作可以并行或并发执行。此外,操作的顺序可能会重新排列。操作可能具有图中未显示的额外步骤。此外,所述方法的一些实施方案可以通过硬件、软件、固件、中间件、微代码、硬件描述语言或其任何组合来实现。当在软件、固件、中间件或微代码中实现时,执行相关任务的程序代码或代码段可以存储在计算机可读介质(诸如存储介质)中。

[0157] 在步骤1602中,可以访问受试者的生物样品的序列数据。在一些情况下,序列数据对应于生物样品的多个无细胞DNA分子,并且多个无细胞DNA分子包括循环肿瘤DNA分子。序列数据还可以包括与从生物样品检测到的多个体细胞变异相对应的序列。此外,每个无细胞DNA分子可以包括相应的末端基序。末端基序可以被识别为DNA片段的核苷酸的末端序列,例如片段任一末端的K个碱基序列。末端序列可以是具有不同碱基数(例如1、2、3、4、5、6、7)的k聚体。在一些情况下,通过将序列读数与参考基因组比对并识别起始位置之前或终止位置之后的核苷酸碱基来确定末端基序。此类碱基仍将对应于DNA片段的末端,例如,因为它们是基于片段的末端序列来识别的。本公开的III.B部分进一步描述了识别无细胞DNA分子中末端基序的方法。

[0158] 在步骤1604中,基于序列数据,可以生成一组末端基序序列数据。在一些情况下,所述组中的每个末端基序序列数据标识具有对应于特定末端基序的末端序列的核酸分子的数量或相对频率。例如,对于特定的4聚体末端基序(例如CCCA),所述组末端基序序列数据可以包括具有所述末端基序的序列读数计数。在一些情况下,具有特定末端基序的核酸分子的数量或相对频率可以被归一化(例如,使用生物样品中序列读数的总计数)。

[0159] 在步骤1606中,可以将所述组末端基序序列数据投影到片段组学特征的潜在变量上。片段组学特征可以包括一个或多个可以预测疾病分类的核酸分子的末端基序频率的分布特征。片段组学特征可以包括一个或多个固定潜在变量。固定潜在变量可通过以下方式确定:(i) 将一个或多个信号分离算法应用于从一个或多个参考样品获得的末端基序序列数据以生成一组潜在变量;以及(ii) 应用聚类算法来选择所述组潜在变量的子集,其中所述子集对应于固定潜在变量。在一些情况下,参考样品包括从诊断患有疾病(例如,癌症)的受试者获得的生物样品(例如,组织、血浆样品)。另外地或替代地,参考样品可以包括从同一受试者但在不同时间点获得的生物样品。在一些情况下,所述组潜在变量中的每个潜在变量包括表示参考样品的末端基序频率分布的直方图或权重向量。一个或多个信号分离算法可以是盲源分离算法,其可以包括独立成分分析算法和/或非负矩阵分解算法。

[0160] 另外地或替代地,可以通过应用变换(诸如缩放、平移、平均、频域转换(例如,快速傅立叶变换)和其他类似变换)来确定潜在变量的导数。

[0161] 血浆DNA核酸酶如DFFB、DNASE1L3和DNASE1参与cfDNA生成和清除。这些不同核酸酶的切割偏好会影响cfDNA末端基序频率。研究表明,多种疾病可以改变血浆DNA核酸酶的活性,例如癌症和红斑狼疮。在一些情况下,鉴于BSS算法具有估计物理上分离的实体携带的信号记录,所述组潜在变量可以预测不同受试者的血浆DNA核酸酶的活性。

[0162] 此外,cfDNA末端基序频率与DNA核酸酶对不同基因组区域的可及性相对应。因此,潜在变量可以预测结合DNA的细胞间异质性程度,从而预测实现表观遗传状态的蛋白质。

[0163] 在步骤1608中,可以基于将所述组末端基序序列数据投影到所述组潜在变量上来确定生物样品的一个或多个片段组学特征幅度。可以通过将所述组末端基序序列数据(例如PMF)投影到参考样品的所述组潜在变量的子集上来确定片段组学特征幅度。在一些情况下,将聚类算法应用于从不同参考样品生成的潜在变量,以确定所述组潜在变量的子集。潜在变量的子集可以包括从已鉴定的潜在变量聚类中选择的质心或中心点。

[0164] 在步骤1610中,片段组学特征幅度可用作机器学习算法(例如逻辑回归分类器模型)的输入以生成结果。结果可以预测受试者是否患有特定疾病的分类。特定疾病可以包括癌症。在一些情况下,如果参考样品是从同一受试者但在不同时间点获得的,则结果可以预测特定疾病的进展或复发。所述结果可用于鉴定针对受试者的治疗和/或确定应多久对受试者实施治疗。在另一个实例中,基于确定与等位基因相关的单染色体和双染色体序列的显著富集的算法,结果可以预测感兴趣的等位基因处染色体结合的存在和沉默。

[0165] 因此,通过利用由受试者的全外显子组测序或某些基因组区域(例如)生成的末端基序数据,可以将映射到所述组基因组区域的每个基因组区域的cfDNA建模为由不同的血浆DNA核酸酶处理的DNA分子的线性混合物。然后可以将BSS算法应用于所述组混合物来估计受试者的片段组学特征。

[0166] 在步骤1614中,可以输出结果。例如,结果可以在本地呈现或传输到另一个装置。结果可以与受试者的标识符一起输出。此后过程1600终止。

[0167] VIII. 疾病与治疗

[0168] 某些实施方案可以包括基于一个或多个生物学输出来预测、诊断和/或预报受试者的疾病或病症的状态或结果。预测、诊断和/或预报受试者的疾病状态或结果可以包括诊断疾病或病症、预测疾病或病症、预测疾病或病症的阶段、评估疾病或病症的风险、评估疾病复发的风险、评估药物的功效、评估不良药物反应的风险、预测最佳药物剂量、预测耐药性、或其组合。

[0169] 本文公开的样品可能来自孕妇。所述样品可以是包含胎儿核酸分子的母体血浆样品。胎儿可能携带染色体非整倍体。胎儿非整倍体可导致多种疾病,例如唐氏综合症(21三体综合症)、帕陶综合症(13三体综合症)和爱德华综合症(18三体综合症)。胎儿可能携带由基因突变和缺失引起的疾病,例如脊髓性肌萎缩和DiGeorge综合征。

[0170] 本文公开的样品可以来自患有癌症的受试者。样品可以包括恶性组织、良性组织、液体活检或其混合物。癌症可能是复发性或/或难治性癌症。癌症的实例包括但不限于肉瘤、癌、淋巴瘤或白血病。在一些情况下,获得包含癌症组织的样品,但没有获得匹配的正常样品。在一些情况下,没有匹配的正常样品可用。在一些情况下,获得匹配的正常样品(例如,用于本文公开的模型的训练和测试)。

[0171] 肉瘤是骨、软骨、脂肪、肌肉、血管或其他结缔或支持组织的癌症。肉瘤包括但不限

于骨癌、纤维肉瘤、软骨肉瘤、尤文氏肉瘤、恶性血管内皮瘤、恶性神经鞘瘤、双侧前庭神经鞘瘤、骨肉瘤、软组织肉瘤(例如,肺泡软部肉瘤、血管肉瘤、叶状囊肉瘤、皮肤纤维肉瘤、纤维样瘤、上皮样肉瘤、骨外骨肉瘤、纤维肉瘤、血管外细胞瘤、血管肉瘤、卡波西肉瘤、平滑肌肉瘤、脂肪肉瘤、淋巴管肉瘤、淋巴肉瘤、恶性纤维组织细胞瘤、神经纤维肉瘤、横纹肌肉瘤和滑膜肉瘤)。

[0172] 癌症是始于上皮细胞的癌症,上皮细胞是覆盖身体表面、产生激素并构成腺体的细胞。作为非限制性实例,癌症包括乳腺癌、胰腺癌、肺癌、结肠癌、结肠直肠癌、直肠癌、肾癌、膀胱癌、胃癌、前列腺癌、肝癌、卵巢癌、脑癌、阴道癌、外阴癌、子宫癌、口腔癌、阴茎癌、睾丸癌、食道癌、皮肤癌、输卵管癌、头颈癌、胃肠基质癌、腺癌、皮肤或眼内黑色素瘤、肛门区癌、小肠癌、内分泌系统癌、甲状腺癌、甲状旁腺癌、肾上腺癌、尿道癌、肾盂癌、输尿管癌、子宫内膜癌、宫颈癌、垂体癌、中枢神经系统(CNS)肿瘤、原发性CNS淋巴瘤、脑干神经胶质瘤和脊髓轴肿瘤。癌症可能是皮肤癌,例如基底细胞癌、鳞状细胞癌、黑色素瘤、非黑色素瘤或光化性(日光性)角化病。

[0173] 癌症可能是肺癌。肺癌可以始于从气管分支出来供应肺部的气道(支气管)或肺部的小气囊(肺泡)。肺癌包括非小细胞肺癌(NSCLC)、小细胞肺癌和间皮瘤。NSCLC的实例包括鳞状细胞癌、腺癌和大细胞癌。间皮瘤可能是肺和胸腔内膜(胸膜)或腹部内膜(腹膜)的癌性肿瘤。间皮瘤可能是由于接触石棉引起的。癌症可能是脑癌,例如胶质母细胞瘤。

[0174] 癌症可能是中枢神经系统(CNS)肿瘤。CNS肿瘤可分为胶质瘤或非胶质瘤。胶质瘤可以是恶性胶质瘤、高级别胶质瘤、弥漫性内在性脑桥胶质瘤。胶质瘤的实例包括星形细胞瘤、少突神经胶质瘤(或少突神经胶质瘤和星形细胞瘤成分的混合物)和室管膜瘤。星形细胞瘤包括但不限于低度星形细胞瘤、间变性星形细胞瘤、多形性胶质母细胞瘤、毛细胞星形细胞瘤、多形性黄色星形细胞瘤和室管膜下巨细胞星形细胞瘤。少突神经胶质瘤包括低度少突神经胶质瘤(或少突星形细胞瘤)和间变性少突神经胶质瘤。非胶质瘤包括脑膜瘤、垂体腺瘤、原发性中枢神经系统淋巴瘤和髓母细胞瘤。癌症可能是脑膜瘤。

[0175] 白血病可能是急性淋巴细胞白血病、急性髓细胞白血病、慢性淋巴细胞白血病或慢性髓细胞白血病。其他类型的白血病包括毛细胞白血病、慢性粒单核细胞白血病和幼年型粒单核细胞白血病。

[0176] 淋巴瘤是淋巴细胞的癌症,可能由B淋巴细胞或T淋巴细胞发展而来。淋巴瘤的两种主要类型是霍奇金淋巴瘤(以前称为霍奇金病)和非霍奇金淋巴瘤。霍奇金淋巴瘤的特征是Reed-Sternberg细胞的存在。非霍奇金淋巴瘤是所有非霍奇金淋巴瘤的淋巴瘤。非霍奇金淋巴瘤可能是惰性淋巴瘤和侵袭性淋巴瘤。非霍奇金淋巴瘤包括但不限于弥漫大B细胞淋巴瘤、滤泡性淋巴瘤、粘膜相关淋巴组织淋巴瘤(MALT)、小细胞淋巴细胞淋巴瘤、套细胞淋巴瘤、伯基特淋巴瘤、纵隔大B细胞淋巴瘤、瓦尔登斯特伦巨球蛋白血症、结节边缘区B细胞淋巴瘤(NMZL)、脾边缘区淋巴瘤(SMZL)、结外边缘区B细胞淋巴瘤、血管内大B细胞淋巴瘤、原发性渗出性淋巴瘤和淋巴瘤样肉芽肿。

[0177] 某些实施方案可以包括基于一个或多个生物学输出治疗来治疗和/或预防受试者的疾病或病症。一个或多个生物学输出可以推荐一种或多种疗法。一个或多个生物学输出可以建议、选择、指定、推荐或以其他方式确定疾病或疾患的治疗过程和/或预防。一个或多个生物学输出可以建议修改或继续一种或多种疗法。修改一种或多种疗法可以包括施

用、启动、减少、增加和/或终止一种或多种疗法。一种或多种疗法包括抗癌、抗病毒、抗菌、抗真菌、免疫抑制疗法或其组合。一种或多种疗法可以治疗、缓解或预防一种或多种疾病或适应症。

[0178] 抗癌疗法的实例包括但不限于手术、化学疗法、放射疗法、免疫疗法/生物疗法、光动力疗法。抗癌疗法可以包括化疗药物、单克隆抗体(例如利妥昔单抗、曲妥珠单抗)、癌症疫苗(例如治疗性疫苗、预防性疫苗)、基因疗法或其组合。

[0179] IX. 计算环境

[0180] 图17说明了用于实施本文公开的一些实施方案的计算机系统1700的实例。计算机系统1700可以包括分布式架构,其中一些组件(例如,存储器和处理器)是最终用户装置的一部分,而一些其他类似组件(例如,存储器和处理器)是计算机服务器的一部分。在一些情况下,计算机系统1700是用于基于核酸分子的大小分布确定片段组学特征的计算机系统,其至少包括处理器1702、存储器1704、存储装置1706、输入/输出(I/O)外围装置1708、通信外围装置1710和接口总线1712。接口总线1712被配置为在计算机系统1700的各个组件之间通信、传输和传递数据、控制和命令。处理器1702可以包括一个或多个处理单元,例如CPU、GPU、TPU、脉动阵列或SIMD处理器。存储器1704和存储装置1706包括计算机可读存储介质,例如RAM、ROM、电可擦除可编程只读存储器(EEPROM)、硬盘驱动器、CD-ROM、光学存储装置、磁存储装置、电子非易失性计算机存储器(例如Flash®存储器)和其他有形存储介质。任何这样的计算机可读存储介质都可以被配置为存储体现本公开的方面的指令或程序代码。存储器1704和存储装置1706还包括计算机可读信号介质。

[0181] 计算机可读信号介质包括传播数据信号,其中实施有计算机可读程序代码。这种传播信号采用多种形式中的任一种,包括但不限于电磁、光学或其任何组合。计算机可读信号介质包括任何计算机可读介质,所述计算机可读介质并非计算机可读存储介质且可通信、传播或传输用于与计算机系统1700结合使用的程序。

[0182] 此外,存储器1704包括操作系统、程序和应用程序。处理器1702被配置为执行存储的指令,并包括例如逻辑处理单元、微处理器、数字信号处理器和其他处理器。例如,计算机系统1700可以执行指令(例如,程序代码),所述指令将处理器1702配置为执行本文所述的一个或多个操作。程序代码包括例如实现分析序列数据的代码,和/或执行本文描述的一个或多个操作的任何其他合适的应用程序。指令可以包括由编译器或解释器根据以任何合适的计算机编程语言编写的代码生成的处理器特定指令,包括例如C、C++、C#、Visual Basic、Java、Python、Perl、JavaScript、R和ActionScript。

[0183] 程序代码可以存储在存储器1704或任何合适的计算机可读介质中,并且可以由处理器1702或任何其他合适的处理器执行。在一些实施方案中,计算机系统中用于执行本文所述的各种特征和过程的所有模块都存储在存储器1704中。在另外或替代实施方案中,上述计算机系统上的这些模块中的一个或多个存储在不同计算系统的不同存储装置中。

[0184] 存储器1704和/或处理器1702可以被虚拟化并且可以托管在例如云网络或数据中心的另一个计算系统内。I/O外围装置1708包括用户接口,例如键盘、屏幕(例如,触摸屏)、麦克风、扬声器、其他输入/输出装置和计算组件,例如图形处理单元、串行端口、并行端口、通用串行总线和其他输入/输出外围装置。I/O外围装置1708通过耦合到接口总线1712的任意端口连接到处理器1702。通信外围装置1710被配置为促进计算机系统1700与其他计算装

置之间通过通信网络进行通信,并且包括例如网络接口控制器、调制解调器、无线和有线接口卡、天线和其他通信外围装置。例如,计算系统1700能够使用通信外围装置1710的网络接口装置通过数据网络与一个或多个其他计算装置(例如,基于核酸分子的大小分布确定片段组学特征的计算装置、生成受试者的生物样品的序列数据的另一个计算装置)进行通信。

[0185] 尽管已参照其具体实施方案来详细描述本主题,但应理解的是,本领域的技术人员在理解前述内容之后可容易得出这些实施方案的替代物、变化和等同物。因此,应理解,本公开仅用于实例而非限制目的,并不排除对本主题进行本领域技术人员显而易见的修改、变化和/或添加。实际上,本文描述的方法和系统可以以各种其他形式体现;此外,在不背离本公开的精神的情况下,可以对本文描述的方法和系统的形式进行各种省略、替换和改变。所附权利要求及其等同物旨在涵盖落入本公开的范围和精神内的此类形式或修改。

[0186] 除非另有明确说明,应理解,在本说明书中,使用诸如“处理”、“计算”、“运算”、“确定”和“识别”等术语的讨论是指计算装置(例如一台或多台计算机或类似的电子计算装置)的动作或过程,其操纵或转换在计算平台的存储器、寄存器或其他信息存储装置、传输装置或显示装置内表示为物理电子量或磁量的数据。

[0187] 本文所讨论的一个或多个系统不限于任何特定的硬件架构或配置。计算装置可以包括任何合适的组件排列,以根据一个或多个输入提供结果。合适的计算装置包括访问存储的软件的基于多用途微处理器的计算系统,所述软件将计算系统从通用计算装置编程或配置为实现本主题的一个或多个实施方案的专用计算装置。任何合适的编程、脚本或其他类型的语言或语言组合均可用于在将用于编程或配置计算装置的软件中实现本文所包含的教导。

[0188] 本文公开的方法的某些实施方案可以在这样的计算装置的操作中执行。上述实例中呈现的块的顺序可以改变,例如,块可以重新排序、组合和/或分解为子块。某些块或过程可以并行执行。

[0189] 如“能够”、“可以”、“可能”、“也许”、“例如”等的本文中使用的条件语言,除非另有特别说明,或在使用的上下文中以其他方式理解,通常旨在传达:尽管其他实施例不包括,但某些实施例包括某些特征、元件和/或步骤。因此,此类条件语言通常并非意图暗示无论如何所述特征、元件和/或步骤都是一个或多个实施例必需的,或者并非暗示一个或多个实施例必须包括用于在借助和不借助作者输入或提示下决定是否包括这些特征、元件和/或步骤或者是否在任何特定实施例中实施这些特征、元件和/或步骤的逻辑。

[0190] 术语“包括”、“包含”、“具有”等是同义词并且以包容性、开放式的方式使用,并且不排除额外的元素、特征、动作、操作等。此外,术语“或”以其包含性含义使用(而非其排他性含义),因此,例如,当用于连接元素列表时,术语“或”是指列表中的一个、一些或所有元素。本文中使用的“适用于”或“配置为”意为开放和包容性的语言,并不排除适用于或配置为执行额外任务或步骤的装置。此外,“基于”的使用意在开放和包容,因为基于一个或多个所述条件或值的过程、步骤、计算或其他动作在实践中可能基于所述条件或值之外的额外条件或值。类似地,“至少部分基于”的使用意在开放和包容,因为至少部分基于一个或多个所述条件或值的过程、步骤、计算或其他动作在实践中可能基于所述条件或值之外的额外条件或值。本文所包含的标题、列表和编号仅用于方便解释,并非旨在限制。

[0191] 上面描述的各种特征和过程可以彼此独立地使用,或者可以以各种方式组合。所

有可能的组合和子组合均旨在落入本公开的范围。此外,在一些实施中可以省略某些方法或过程块。本文所述的方法和过程也不限于任何特定的顺序,并且与其相关的块或状态可以按照适当的其他顺序来执行。例如,所描述的块或状态可以按照与具体公开的顺序不同的顺序执行,或者可以将多个块或状态组合成单个块或状态。实例块或状态可以串行、并行或以某种其他方式执行。可以在所公开的实例中增加或删除块或状态。类似地,本文描述的实例系统和组件的配置可能与所描述的不同。例如,与公开的实施例相比,可以添加、删除或重新排列元素。

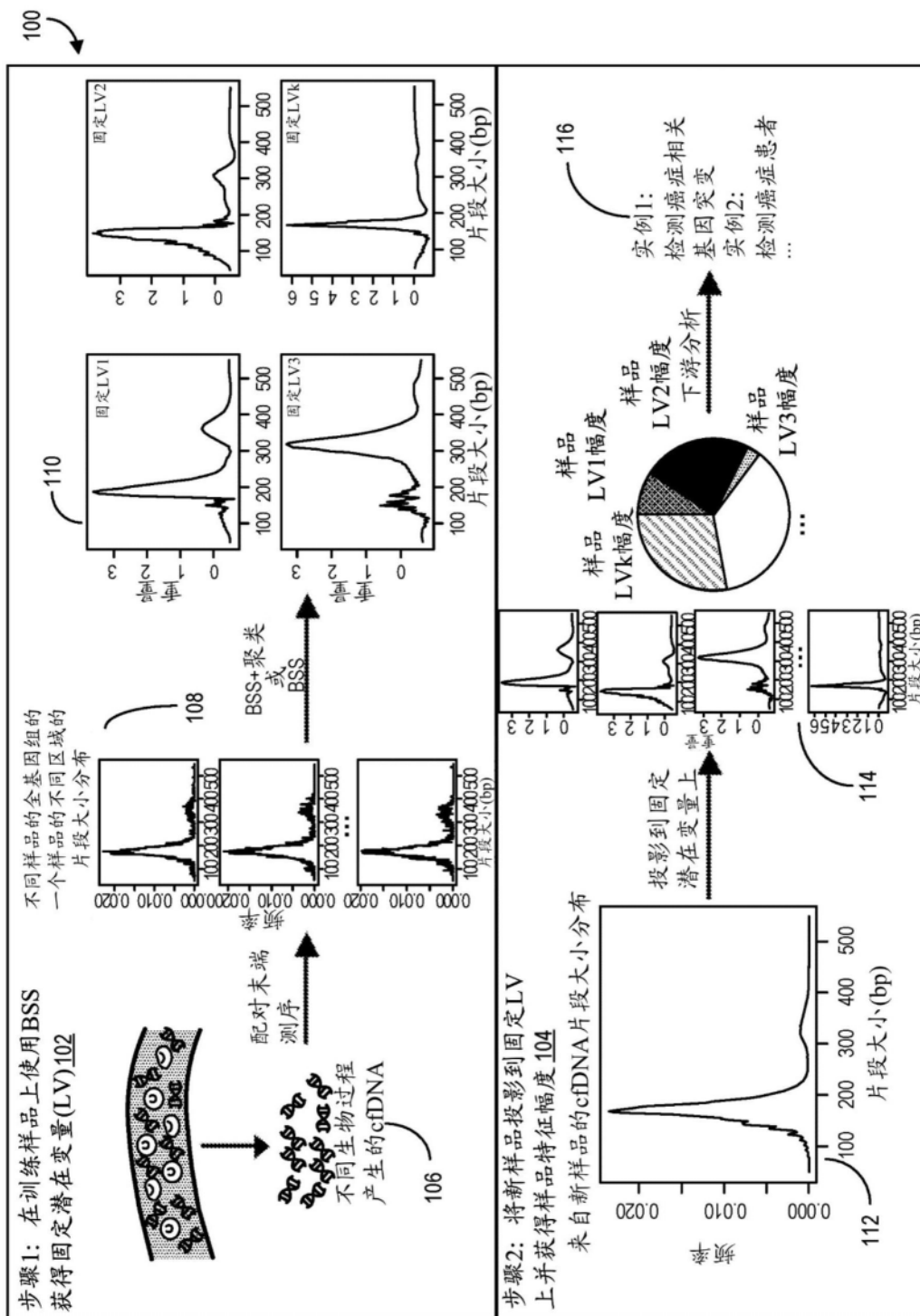


图1

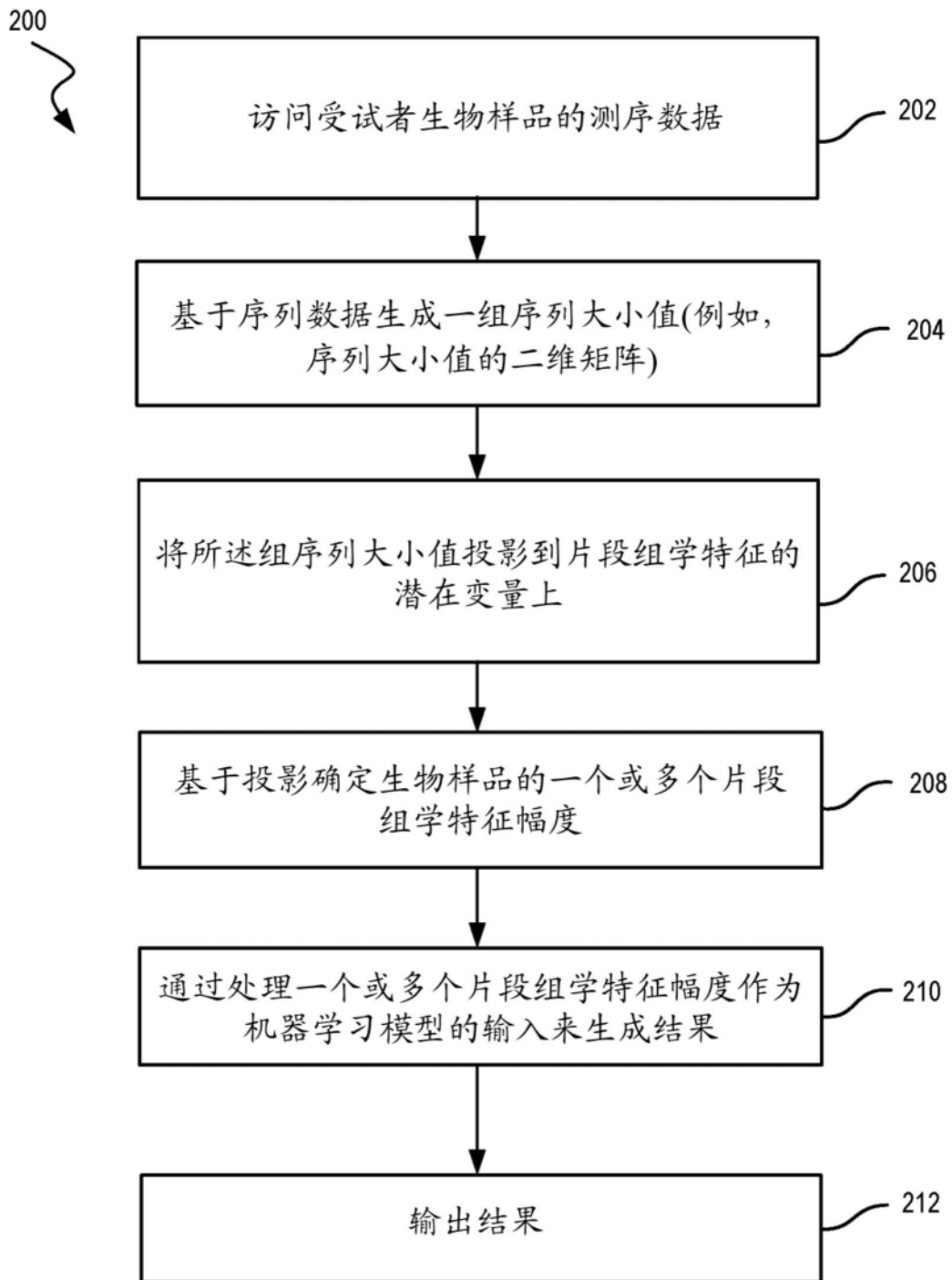


图2

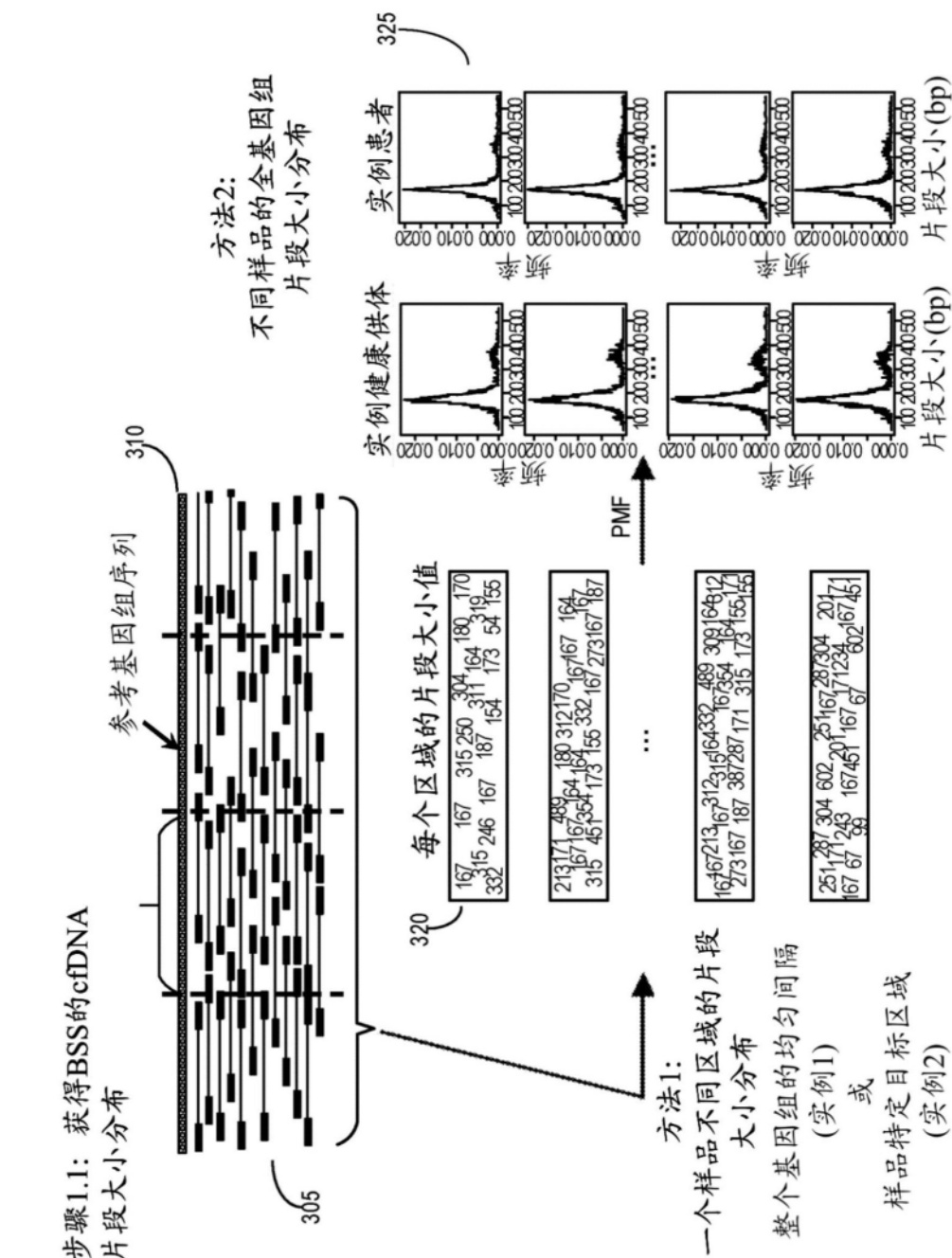


图3

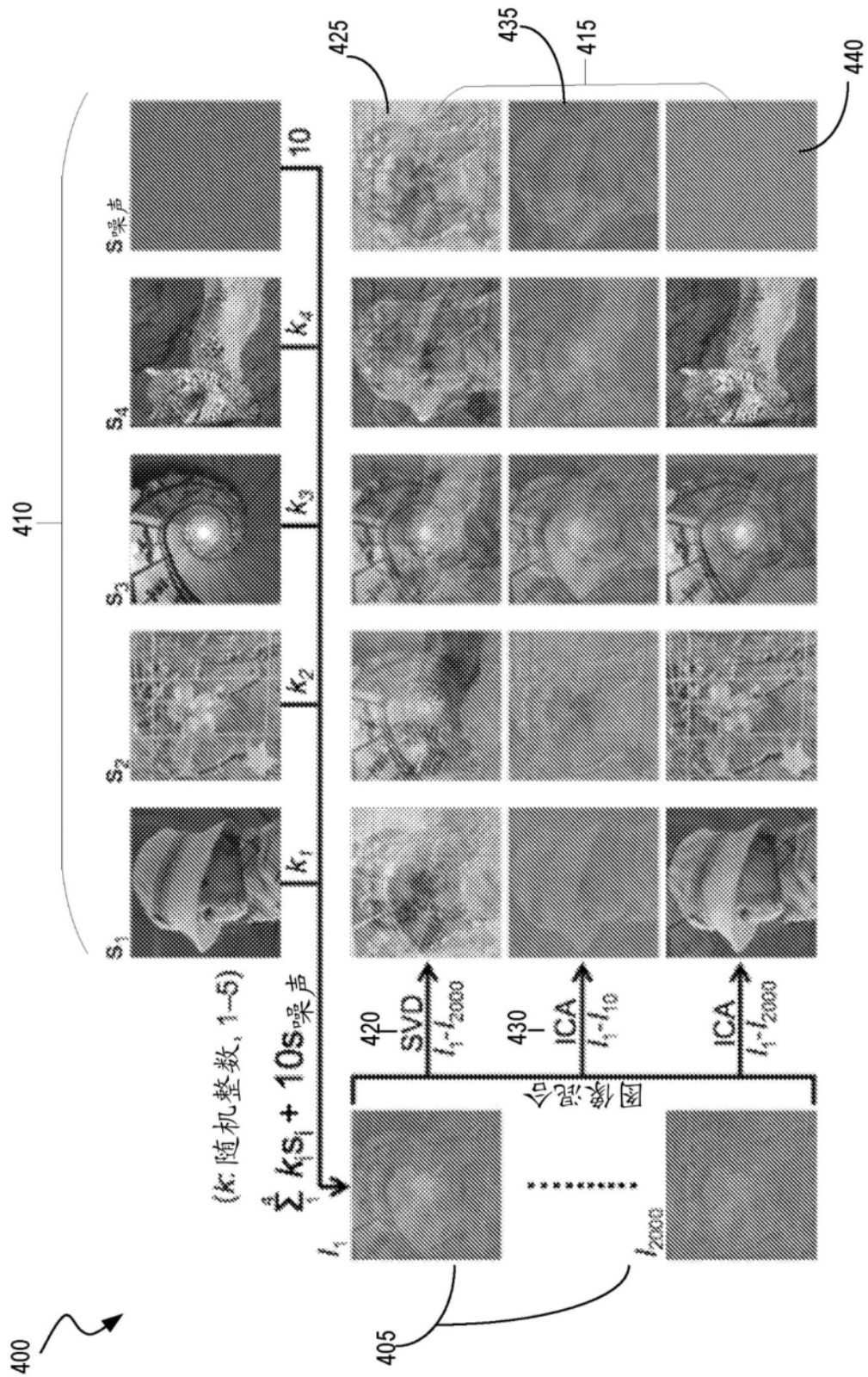


图4

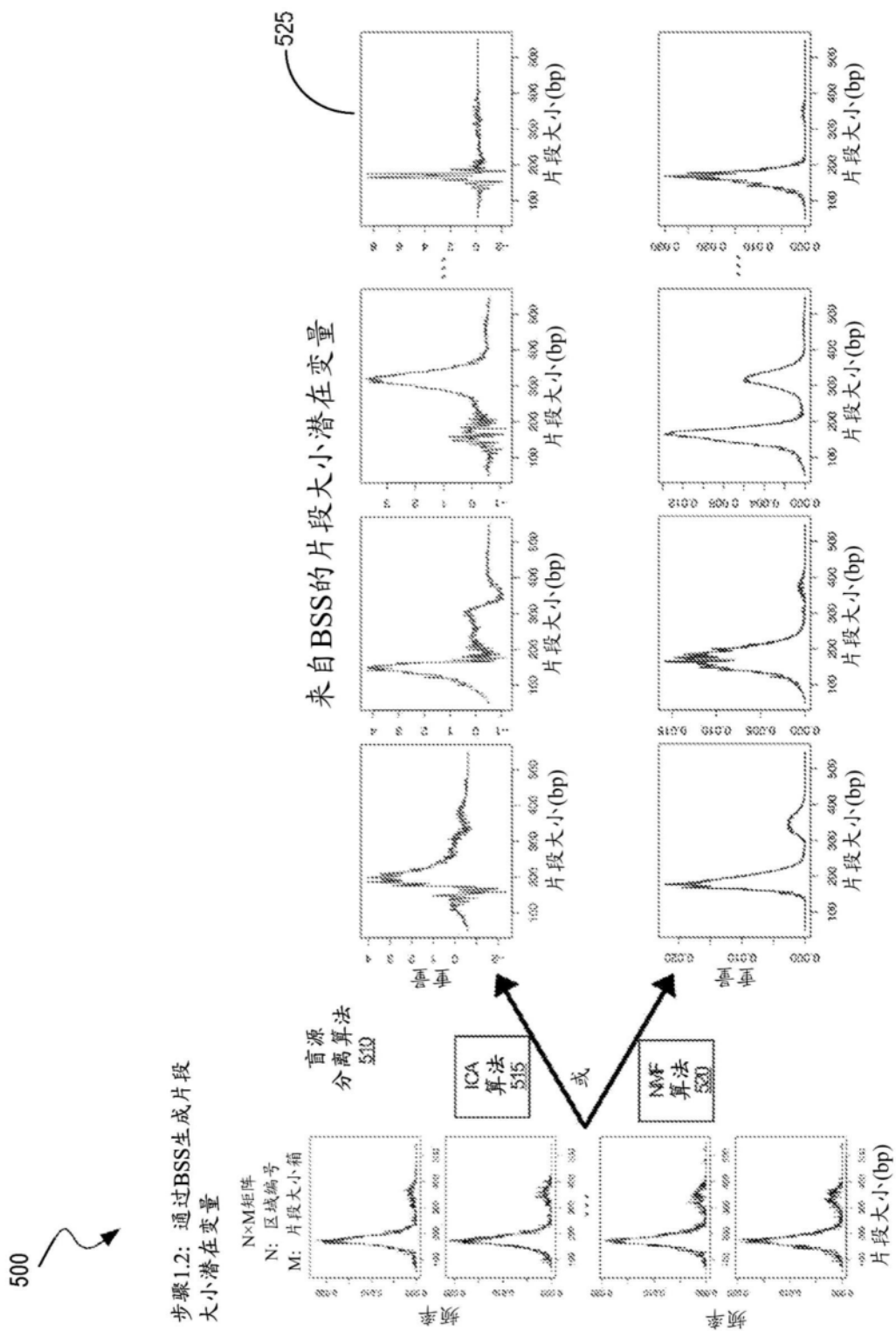


图5

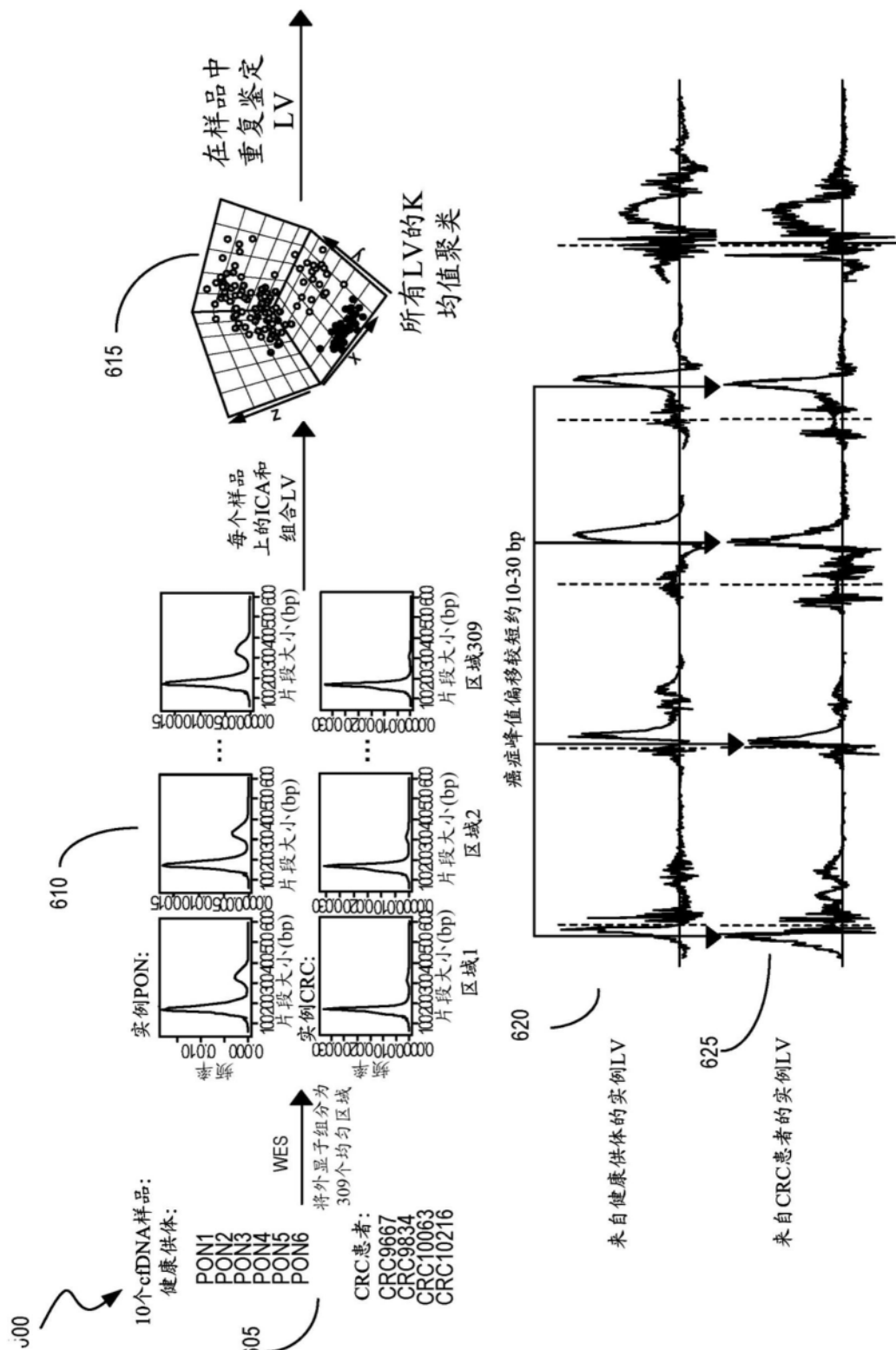


图6

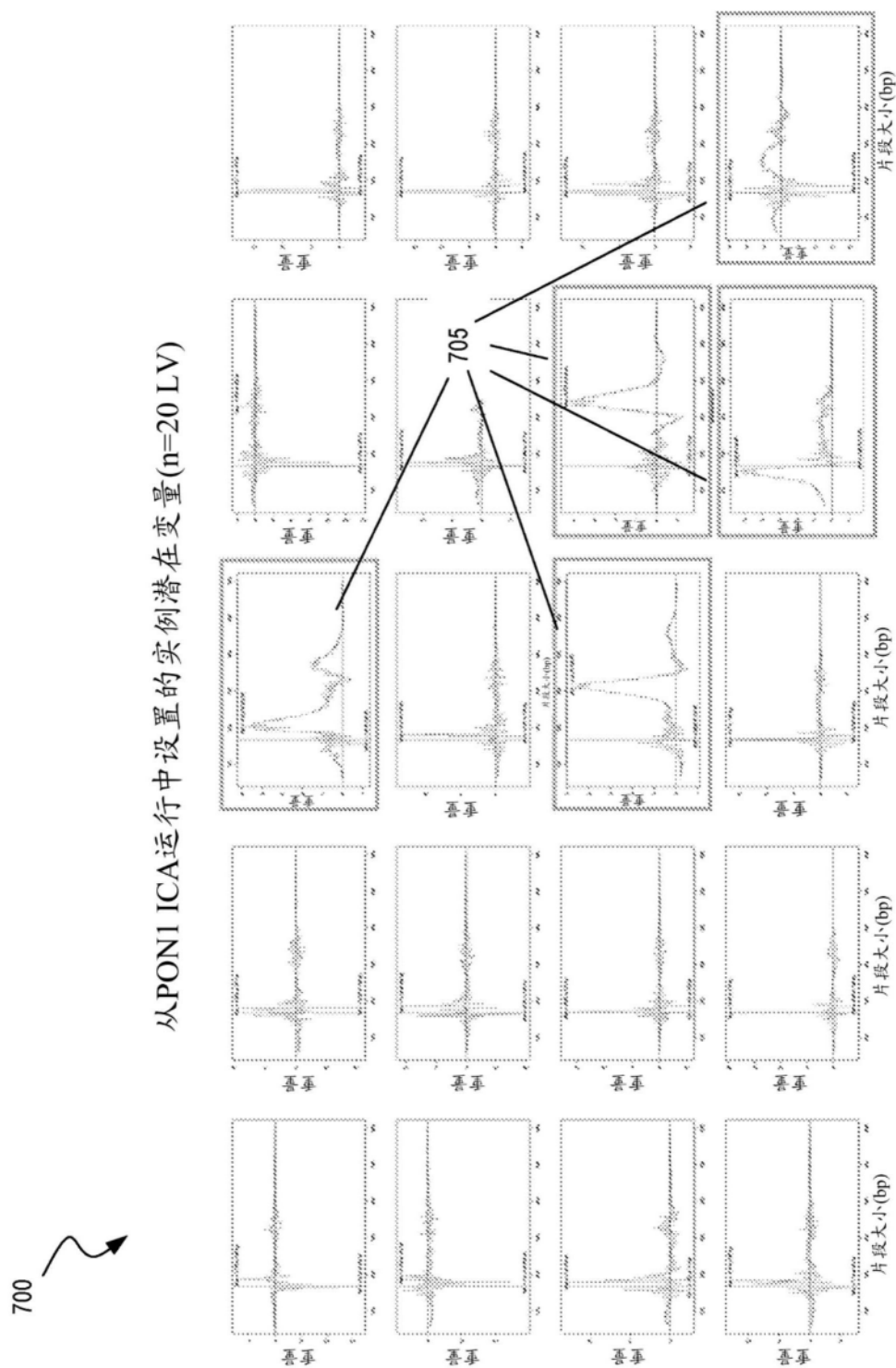


图7

800

从PON2 ICA运行中设置的实例潜在变量(n=20 LV)

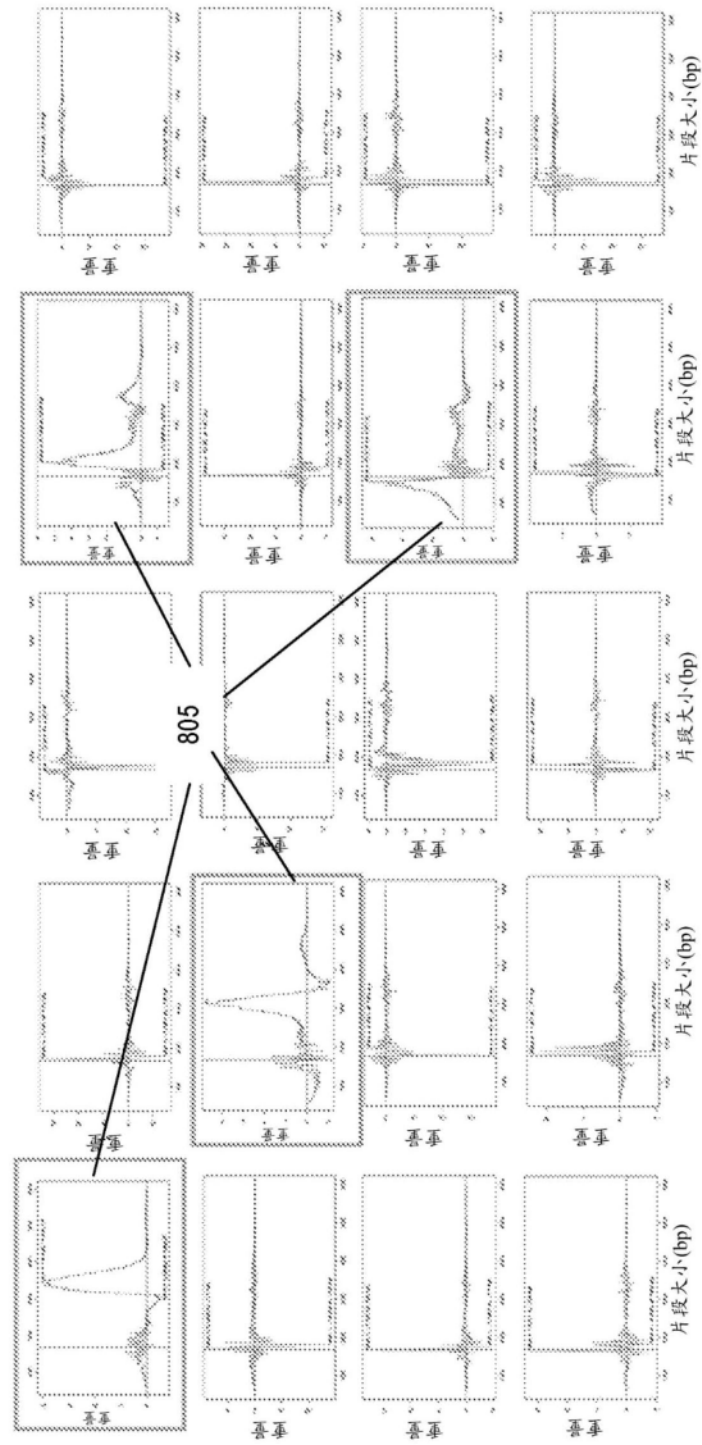


图8

900

从CRC10063 ICA运行中设置的实例潜在变量(n=20 LV)

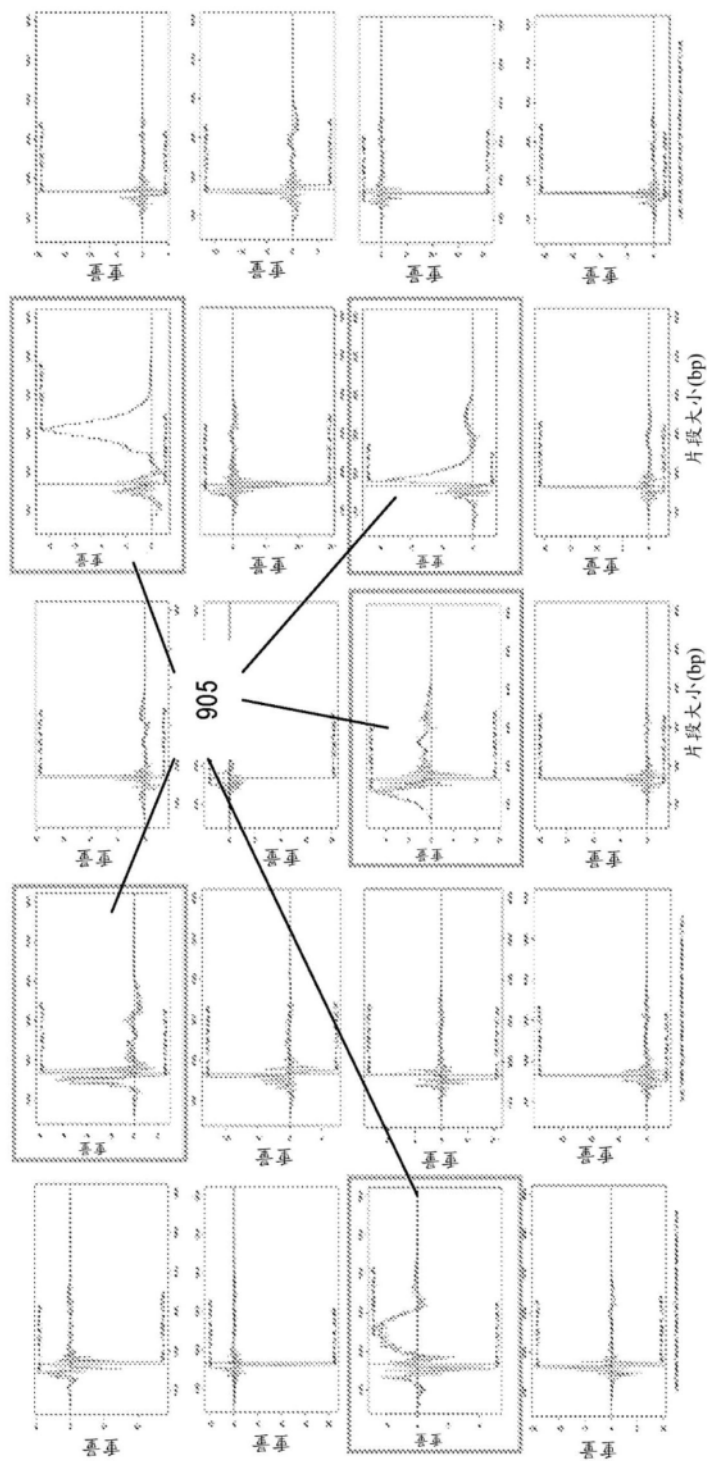


图9

1000

从CRC9834 ICA运行中设置的实例潜在变量(n=20 LV)

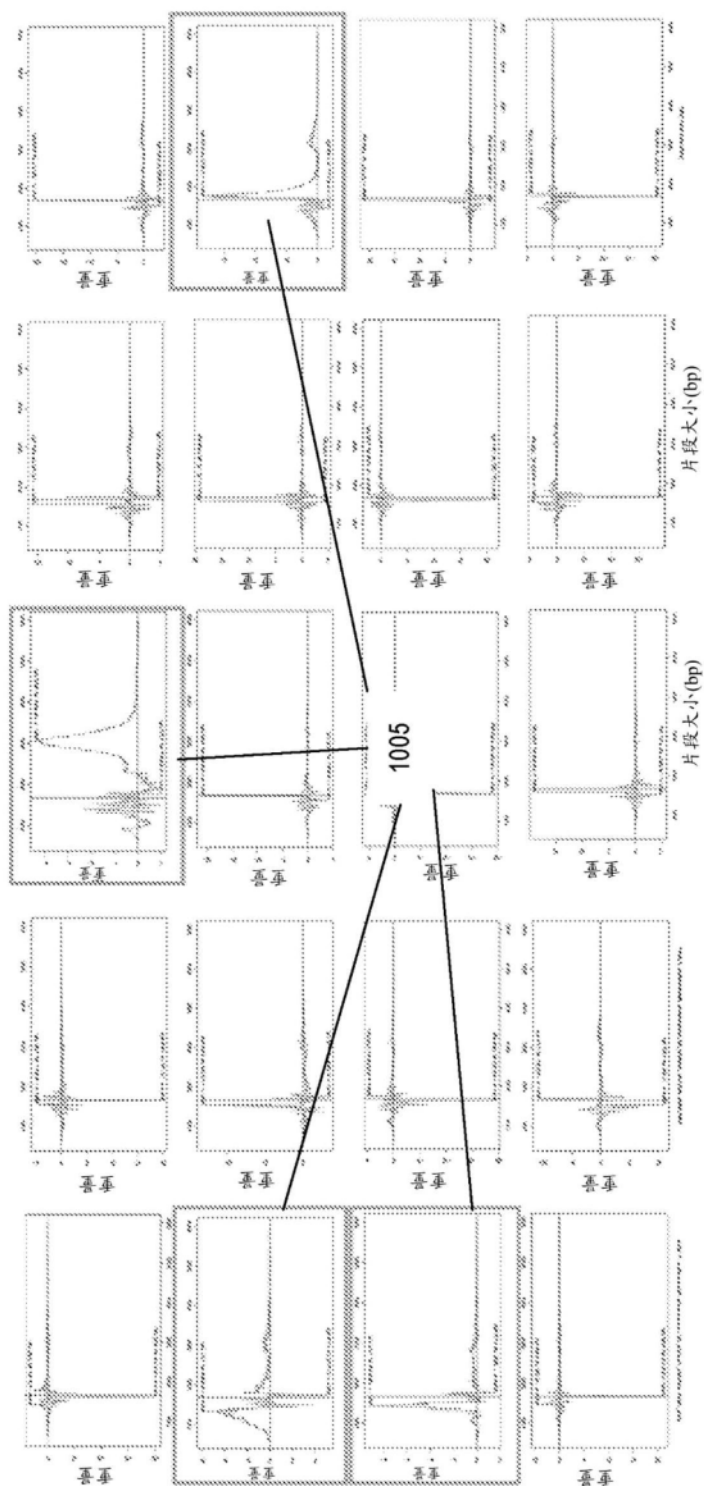


图10

1100

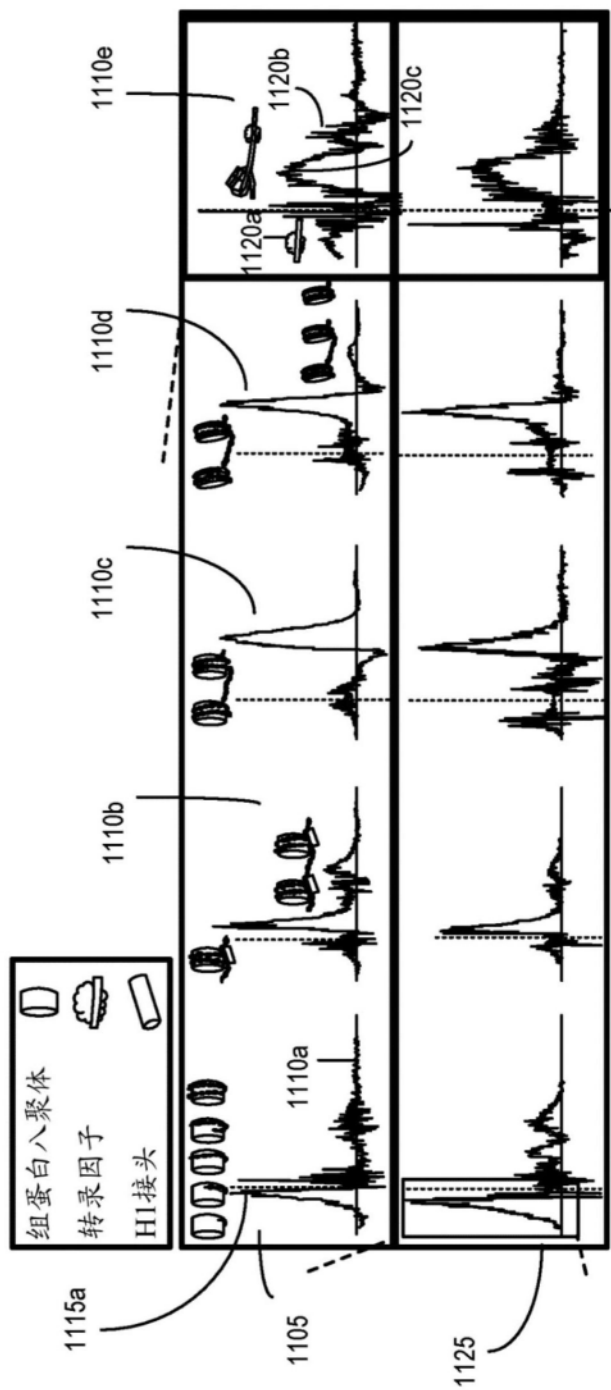


图11

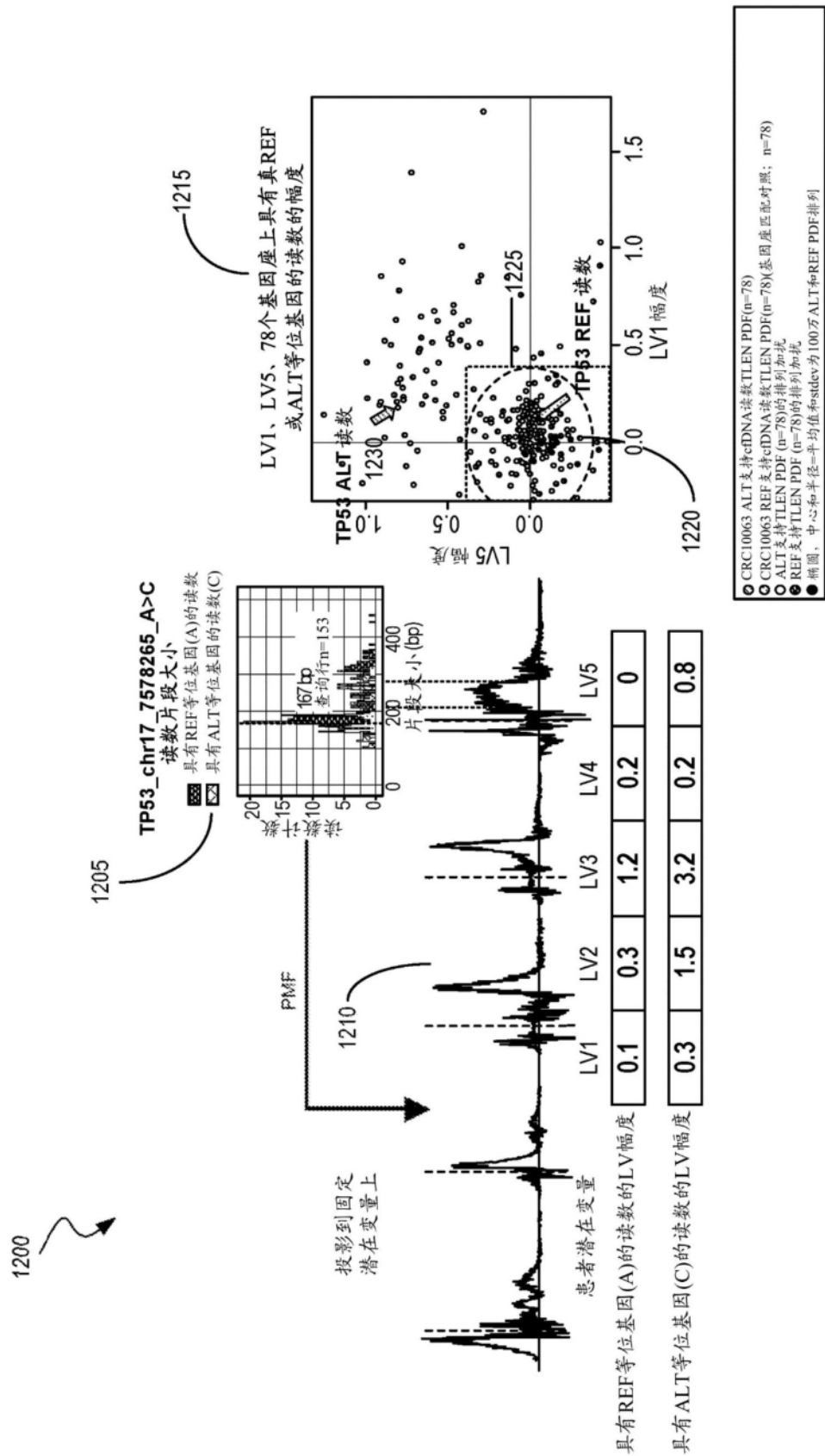


图12

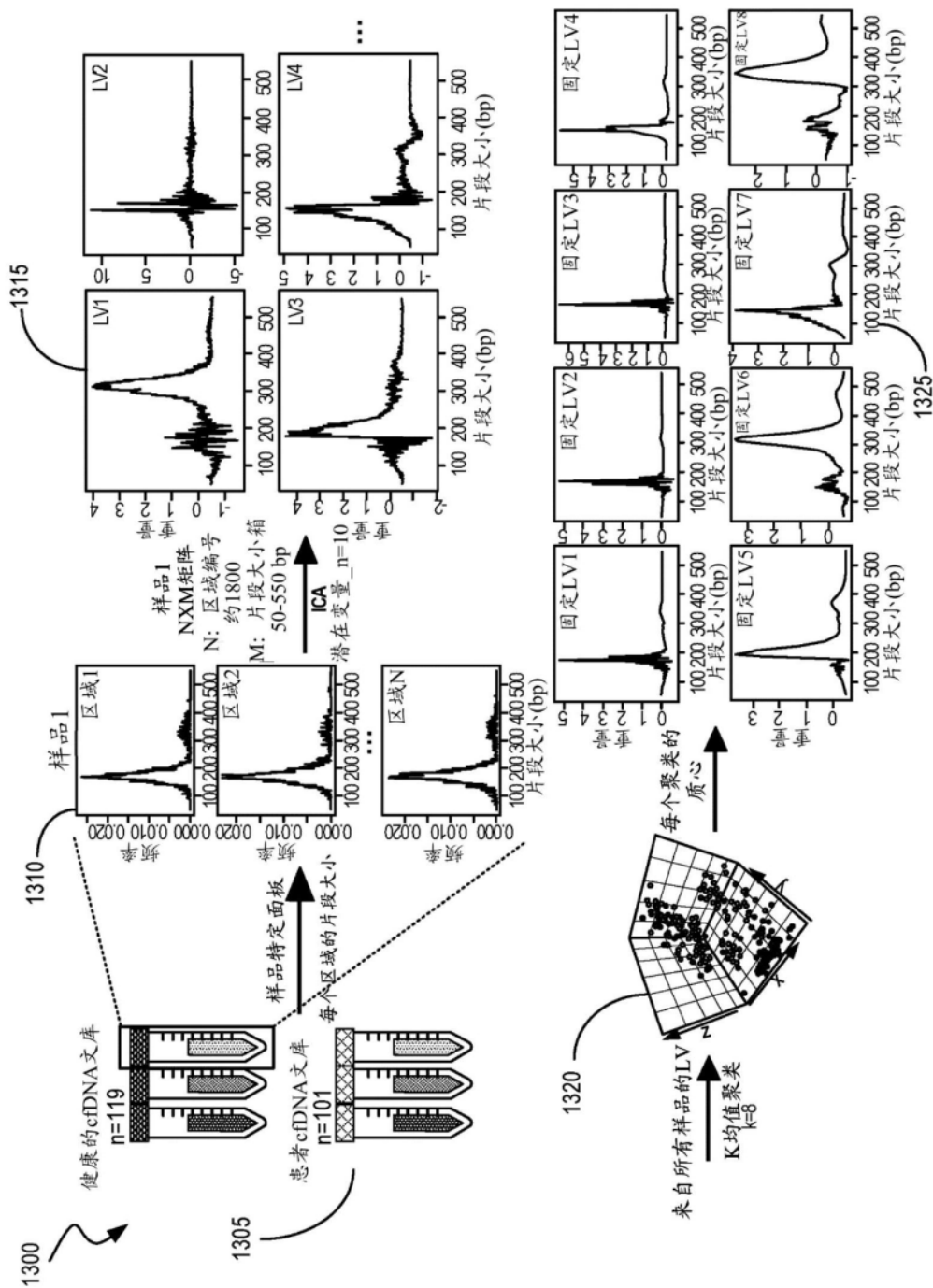


图13

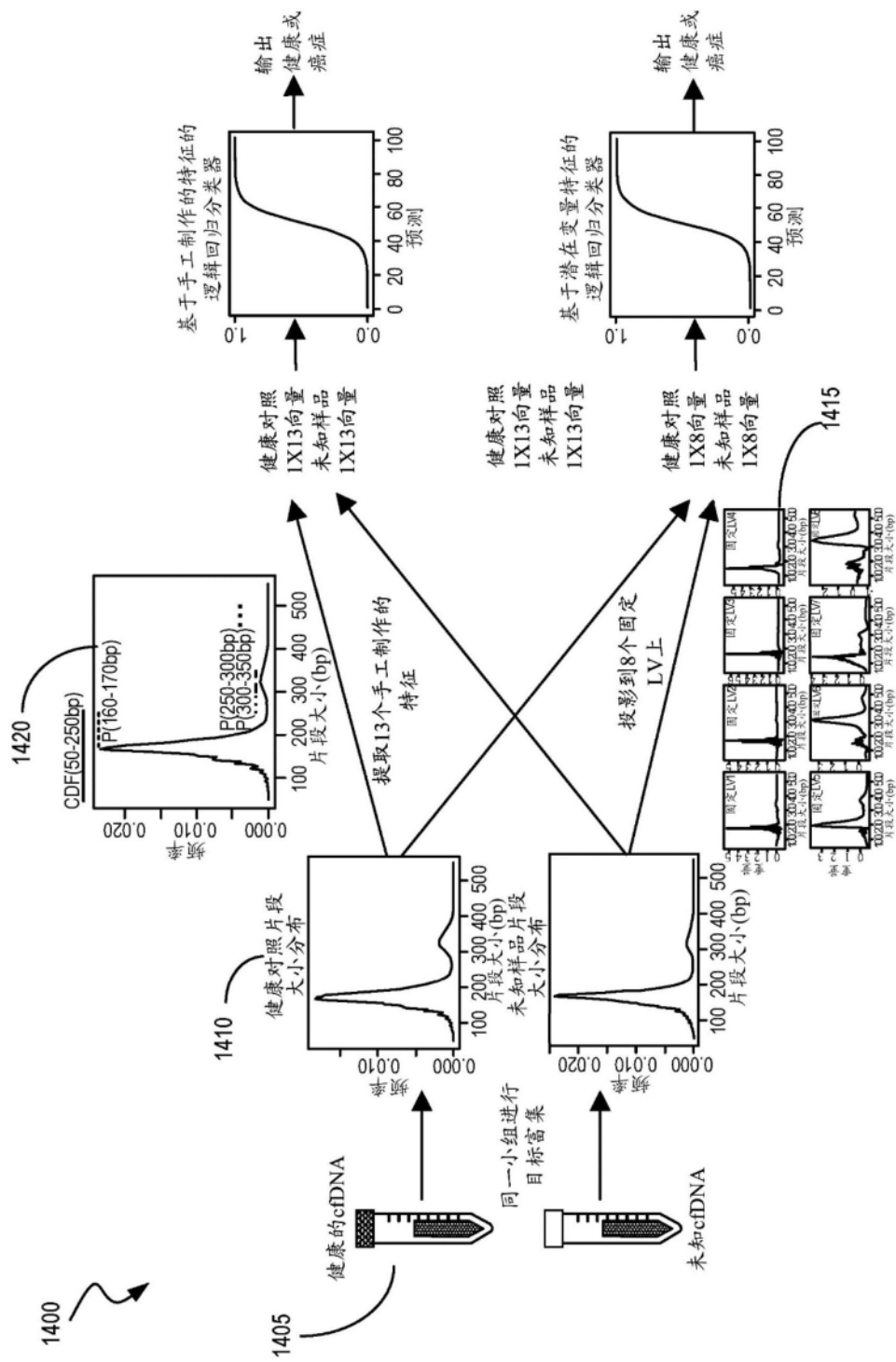


图14

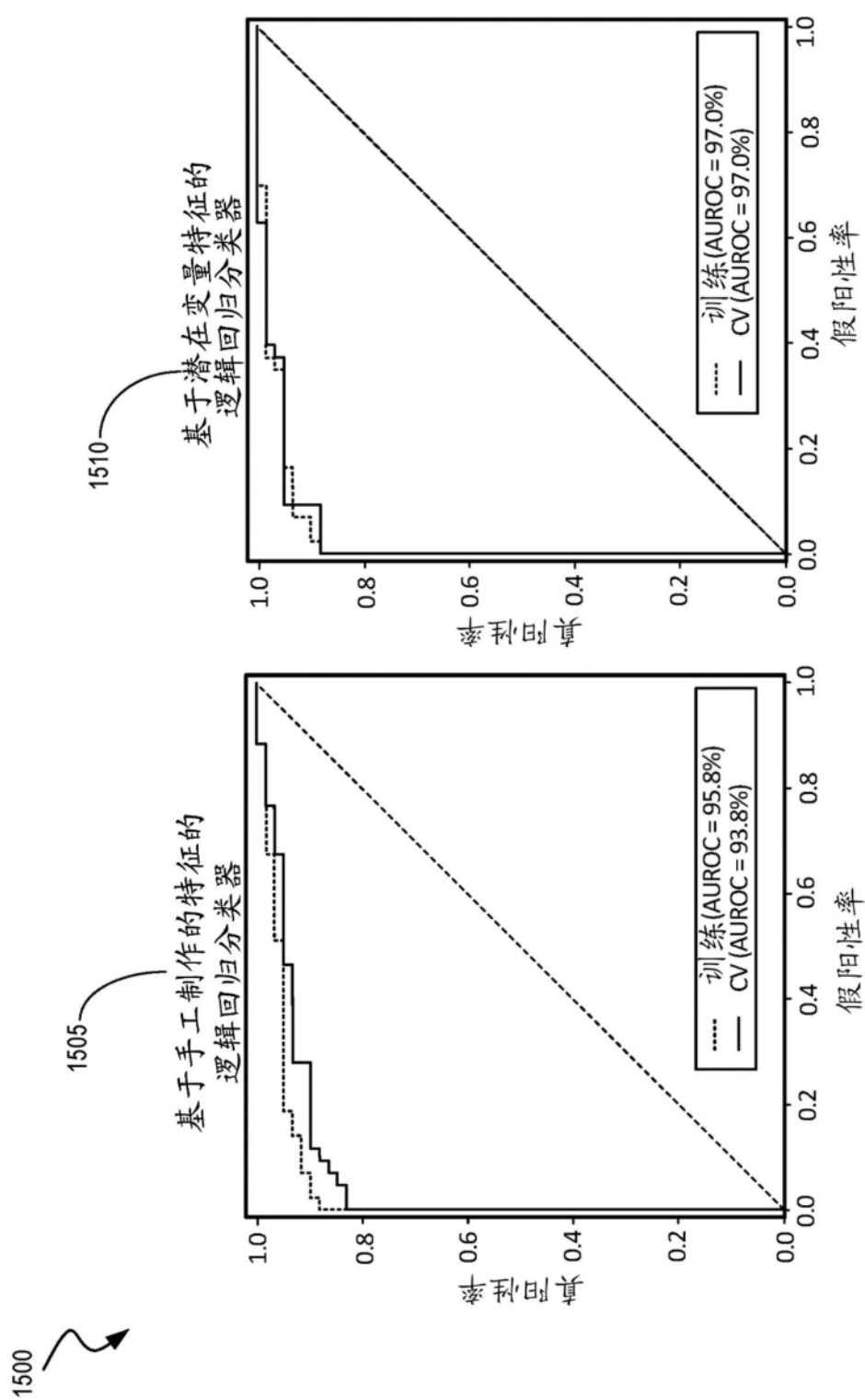


图15

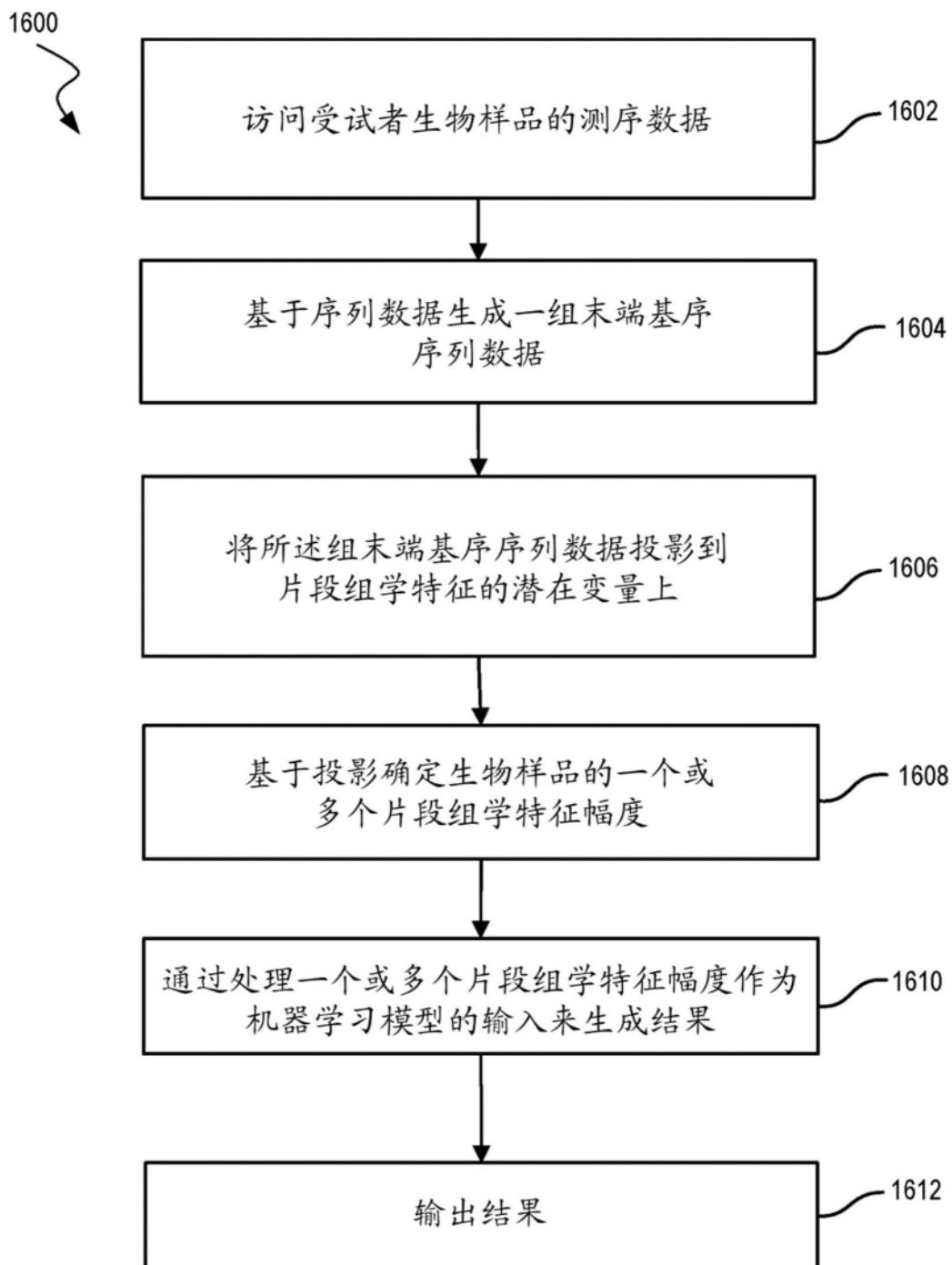


图16

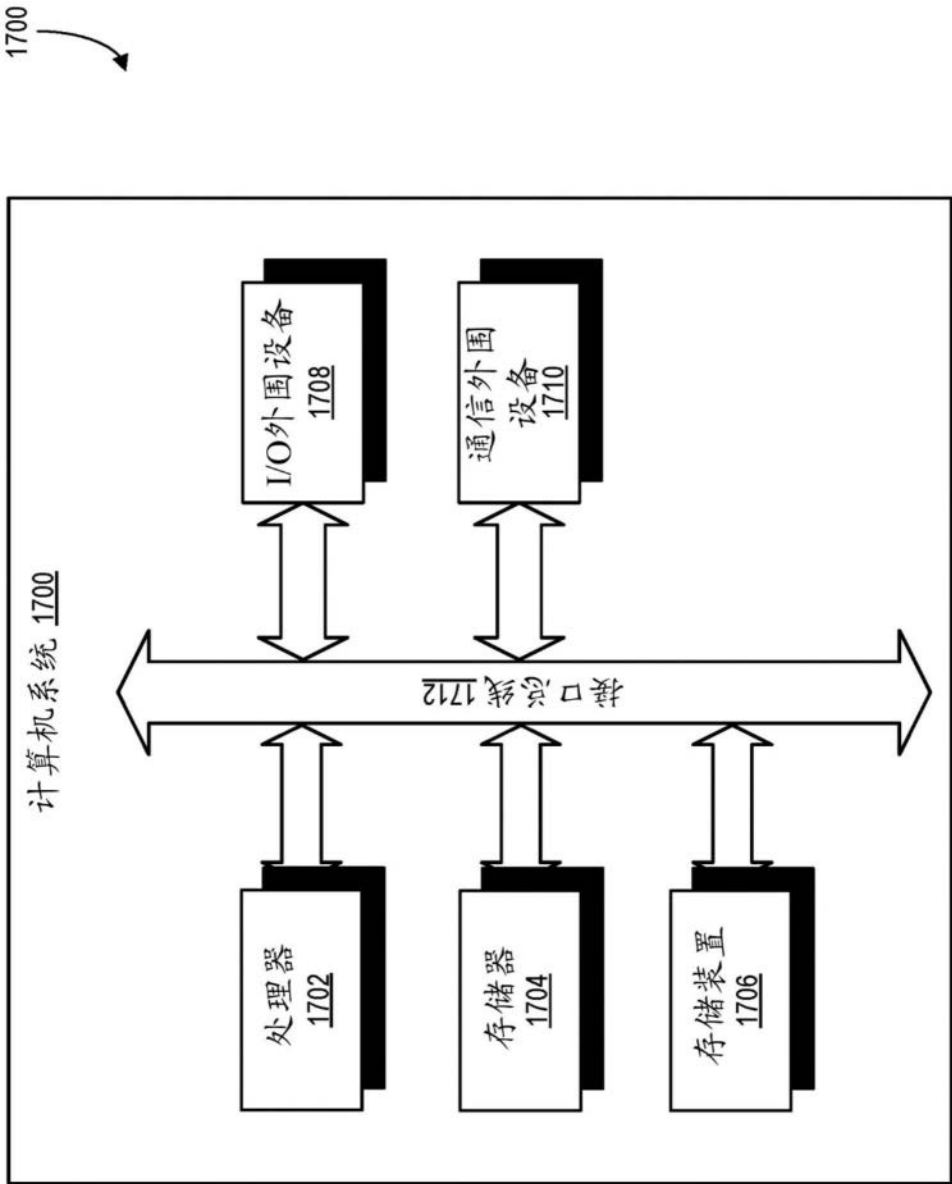


图17