(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111278993 A
(43)申请公布日 2020.06.12

(21)申请号 201880070782.2

(22)申请日 2018.09.14

(30)优先权数据
62/559,366 2017.09.15 US

(85)PCT国际申请进入国家阶段日
2020.04.29

(86)PCT国际申请的申请数据
PCT/US2018/051160 2018.09.14

(87)PCT国际申请的公布数据
WO2019/055835 EN 2019.03.21

(71)申请人 加利福尼亚大学董事会
地址 美国加利福尼亚州

(72)发明人 向红·婕思敏·周 李硕 李文渊

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 郑斌 刘振佳

(51)Int.Cl.
C12Q 1/6886(2018.01)
G16B 20/20(2019.01)
G16B 20/50(2019.01)

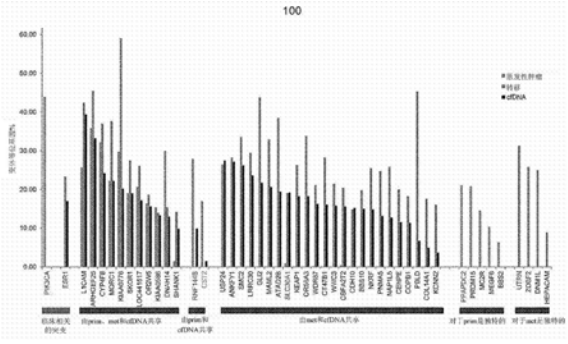
权利要求书9页 说明书22页 附图15页

(54)发明名称

从无细胞核酸中检测体细胞单核苷酸变体并应用于微小残留病变监测

(57)摘要

本公开内容提供了用于在无细胞核酸样品中进行准确且灵敏的体细胞单核苷酸变体(SNV)检测的概率模型,所述样品包含序列数据集。可对于序列数据集中的每个基因座确定联合基因型,并且可固有地去除种系突变。可应用一组过滤来消除低质量的体细胞变体调用。此外,可考虑总体肿瘤无细胞脱氧核糖核酸(cfDNA)分数和重叠读取伴侣,从而能够从具有低肿瘤cfDNA分数的样品中进行准确的SNV检测和变体等位基因频率估计。通过使用概率模型和机器学习模型以区分真实变体与测序误差,从而设计出微小残留病变(MRD)的灵敏早期检测。



CN 111278993 A

1.用于从无细胞核酸例如脱氧核糖核酸(cfDNA)和核糖核酸(cfRNA)中检测体细胞单核苷酸变体(SNV)的方法,所述方法包括:

对于包含序列数据集的cfDNA样品,估计总体肿瘤cfDNA分数;

对于测序数据集中的至少一个基因座k,确定基因型似然;

从包含序列数据集的所述cfDNA样品中消除种系多态性;

通过一组过滤器过滤SNV候选;以及

使用并入了所估计的总体肿瘤cfDNA分数的概率模型分析所述cfDNA样品,以确定所述cfDNA样品中肿瘤来源的DNA(ctDNA)的分数。

2.权利要求1所述的方法,其中对于所述cfDNA样品,估计所述总体肿瘤cfDNA分数包括:

组合来自所述cfDNA样品中所有潜在SNV位点的信息,以降低由在一个或更多个所述潜在SNV位点处发生的测序误差引起的噪声。

3.权利要求1所述的方法,其中确定所述基因型似然包括:

在给定不同联合基因型的情况下计算所观察到的读取覆盖所述基因座的似然。

4.权利要求1所述的方法,其中确定所述基因型似然包括:

估计基因座处的基因型,其使得后验概率最大化。

5.权利要求4所述的方法,其中估计所述基因座处的基因型包括:

确定所述序列数据集中每个基因座的联合基因型。

6.权利要求1所述的方法,其中过滤所述SNV候选包括以下中的至少一项:

基于链偏倚过滤器过滤SNV候选;

基于碱基质量过滤器过滤SNV候选;

基于读取伴侣过滤器过滤测序读取;

基于读取伴侣过滤器过滤SNV候选;

基于对真实变体与测序误差进行分类的机器学习模型过滤测序读取;

基于序列背景过滤器过滤SNV候选;以及

基于公共数据库过滤SNV候选。

7.权利要求6所述的方法,其中基于所述对真实变体与测序误差进行分类的机器学习模型过滤所述测序读取包括:

建立包含真实变体或测序误差的测序读取的基准真实训练数据;

创建每个测序读取的特征谱,其具有包含以下的信息:所述读取中每个碱基的测序质量,读取比对信息,序列背景(例如读取序列和插入/缺失)以及双端测序数据的插入大小;

基于训练数据训练分类器,以通过使用每个读取的特征谱对具有真实变体的测序读取和具有测序误差的读取进行分类;以及使用经训练的分类器将每个cfDNA测序读取分类为具有真实变体的读取或具有测序误差的读取。

8.用于从无细胞核酸例如脱氧核糖核酸(cfDNA)和核糖核酸(cfRNA)中检测体细胞单核苷酸变体(SNV)的系统,所述系统包含:

计算机存储器;

通信偶连至所述计算机存储器的一个或更多个计算机处理器,所述一个或更多个计算机处理器被配置为实现包括以下的方法:

对于包含序列数据集的cfDNA样品，估计总体肿瘤cfDNA分数；

对于测序数据集中的至少一个基因座k，确定基因型似然；

从包含序列数据集的所述cfDNA样品中消除种系多态性；

通过一组过滤器过滤SNV候选；以及

使用并入了所估计的总体肿瘤cfDNA分数的概率模型分析所述cfDNA样品以确定所述cfDNA样品中肿瘤来源的DNA(ctDNA)的分数。

9.权利要求8所述的系统，其中对于所述cfDNA样品，估计所述总体肿瘤cfDNA分数包括：

组合来自所述cfDNA样品中所有潜在SNV位点的信息，以降低由在一个或更多个所述潜在SNV位点处发生的测序误差引起的噪声。

10.权利要求8所述的系统，其中确定所述基因型似然包括：

在给定不同联合基因型的情况下计算所观察到的读取覆盖所述基因座的似然。

11.权利要求10所述的系统，其中确定所述基因型似然包括：

估计基因座处的基因型，其使得后验概率最大化。

12.权利要求8所述的系统，其中估计所述基因座处的基因型包括：

确定所述序列数据集中每个基因座的联合基因型。

13.权利要求12所述的系统，其中过滤所述SNV候选包括以下中的至少一个：

基于链偏倚过滤器过滤SNV候选；

基于碱基质量过滤器过滤SNV候选；

基于读取伴侣过滤器过滤测序读取；

基于读取伴侣过滤器过滤SNV候选；

基于对真实变体与测序误差进行分类的机器学习模型过滤测序读取；

基于序列背景过滤器过滤SNV候选；以及

基于公共数据库过滤SNV候选。

14.权利要求13所述的系统，其中基于对真实变体与测序误差进行分类的机器学习模型过滤所述测序读取包括：

建立包含真实变体或测序误差的测序读取的基准真实训练数据；

创建每个测序读取的特征谱，其具有包含以下的信息：所述读取中每个碱基的测序质量，读取比对信息，序列背景(例如读取序列和插入/缺失)以及双端测序数据的插入大小；

基于训练数据训练分类器以通过使用每个读取的特征谱对具有真实变体的测序读取和具有测序误差的读取进行分类；以及

使用经训练的分类器将每个cfDNA测序读取分类为具有真实变体的读取或具有测序误差的读取。

15.用于从手术之前和之后收集的血浆样品、白细胞和切除的肿瘤样品(如果有的话)中检测微小残留病变(MRD)的方法，所述方法包括：

从手术前血液样品和切除的肿瘤样品中的至少一种中鉴定一种或更多种截短突变和所述一种或更多种截短突变的突变谱；以及

在所述手术之后使用随访血浆cfDNA样品检测MRD。

16.权利要求15所述的方法，其中在所述手术之后使用随访血浆cfDNA样品检测MRD包

括：

提取覆盖截短突变位置的读取;以及

使用被分类为具有真实变体的读取来计算MRD预测得分。

17.权利要求16所述的方法,其中使用所述被分类为具有真实变体的读取来计算所述MRD预测得分包括：

对基因组中的k个位点进行采样,所述位点不包含已鉴定的突变,但匹配k个截短突变的特征;

过滤被鉴定为包含误差的读取;以及

生成所述MRD预测得分。

18.用于从手术之前和之后收集的血浆样品、白细胞和切除的肿瘤样品(如果有的话)中检测微小残留病变(MRD)的系统,所述系统包含：

计算机存储器;

通信偶连至所述计算机存储器的一个或更多个计算机处理器,所述一个或更多个计算机处理器被配置为实现包括以下的方法：

从手术前血液样品和切除的肿瘤样品中的至少一种中鉴定一种或更多种截短突变和所述一种或更多种截短突变的突变谱;以及

在所述手术之后使用随访血浆cfDNA样品检测MRD。

19.权利要求18所述的系统,其中在所述手术之后使用随访血浆cfDNA样品检测MRD包括：

提取覆盖截短突变位置的读取,以及使用被分类为具有真实变体的读取来计算MRD预测得分。

20.权利要求19所述的系统,其中使用所述被分类为具有真实变体的读取来计算所述MRD预测得分包括：

对基因组中的k个位点进行采样,所述位点不包含已鉴定的突变但匹配那些k个截短突变的特征;

过滤被鉴定为包含误差的读取;以及

生成所述MRD预测得分。

21.存储指令集的非暂时性存储介质,当执行所述指令时,其使得一个或更多个计算机处理器从无细胞核酸例如脱氧核糖核酸(cfDNA)和核糖核酸(cfRNA)中检测体细胞单核苷酸变体(SNV),所述指令集包含以下指令：

对于包含测序数据集的cfDNA/cfRNA样品,合并重叠读取伴侣;

对于包含序列数据集的cfDNA/cfRNA样品,估计总体肿瘤cfDNA分数;

对于所述测序数据集中的基因座,确定基因型似然;

从包含所述序列数据集的cfDNA/cfRNA样品中消除种系多态性;

使用并入了所估计的总体肿瘤cfDNA分数的概率模型分析所述cfDNA/cfRNA样品,以确定所述cfDNA/cfRNA样品中肿瘤来源的DNA(ctDNA/ctRNA)的分数;

使用并入了cfDNA/cfRNA样品特性的一组过滤器,消除低质量的体细胞SNV候选;

从包含所述测序数据集的cfDNA/cfRNA样品中消除不一致的重叠读取伴侣;

使用区分测序误差与真实变体的机器学习模型消除具有测序误差的读取;以及

使用截短突变谱和所述机器学习模型从所述cfDNA/cfRNA样品中确定早期微小残留病变(MRD)。

22.用于从无细胞核酸例如脱氧核糖核酸(cfDNA)和核糖核酸(cfRNA)中检测体细胞单核苷酸变体(SNV)的方法,所述方法包括:

对于包含测序数据集的cfDNA/cfRNA样品,合并重叠读取伴侣;

对于包含序列数据集的cfDNA/cfRNA样品,估计总体肿瘤cfDNA分数;

对于所述测序数据集中的基因座,确定基因型似然;

从包含所述序列数据集的cfDNA/cfRNA样品中消除种系多态性;

使用并入了所估计的总体肿瘤cfDNA分数的概率模型分析所述cfDNA/cfRNA样品,以确定所述cfDNA/cfRNA样品中肿瘤来源的DNA(ctDNA/ctRNA)的分数;

使用并入了cfDNA/cfRNA样品特性的一组过滤器,消除低质量的体细胞SNV候选;

从所述包含测序数据集的cfDNA/cfRNA样品中消除不一致的重叠读取伴侣;

使用区分测序误差与真实变体的机器学习模型消除具有测序误差的读取;以及

使用截短突变谱和所述机器学习模型从所述cfDNA/cfRNA样品中确定早期微小残留病变(MRD)。

23.用于从对象的多个无细胞核酸(cfNA)分子中检测体细胞单核苷酸变体(SNV)的方法,其包括:

(a)调取由测序仪产生的多个序列读取,其中所述多个序列读取的至少一个子集包含来自所述多个cfNA分子或其衍生物的序列;

(b)将概率模型应用于多个遗传基因座中的每一个处的所述多个序列读取,以估计所述多个cfNA分子的总体肿瘤负荷,其中所述估计的总体肿瘤负荷包含所述多个cfNA分子中肿瘤来源的cfNA分子的定量测量,其中所述多个遗传基因座包含潜在的SNV位点;

(c)对于所述多个遗传基因座中的每一个,至少部分地基于所述总体肿瘤负荷,确定所述对象的一种或更多种基因型的似然,其中所述一种或更多种基因型选自正常基因型、肿瘤基因型和联合正常肿瘤基因型;

(d)对于所述多个遗传基因座中的每一个,至少部分地基于(c)中确定的所述一种或更多种基因型的所述似然和所述多种cfNA分子中肿瘤来源的cfNA分子的所述定量测量,检测一种或更多种SNV;以及

(e)从(d)中检测的所述一种或更多种SNV中过滤出一种或更多种种系多态性,从而获得一种或更多种体细胞SNV。

24.权利要求23所述的方法,其还包括:

(f)使用选自以下的一种或更多种过滤器过滤出在(e)中获得的所述一种或更多种体细胞SNV,从而获得经过滤的一组体细胞SNV:链偏倚过滤器、碱基质量过滤器、读取伴侣过滤器、测序误差过滤器、插入或缺失(插失)和均聚物引发的误差过滤器和公共数据库过滤器。

25.权利要求24所述的方法,其还包括:

(g)将所述概率模型应用于所述多个遗传基因座中的每一个处的所述多个序列读取,以重新估计所述总体肿瘤负荷,其中所述多个遗传基因座包含在(f)中获得的所述经过滤的一组体细胞SNV。

26.权利要求23所述的方法,其中所述多个无细胞核酸(cfNA)分子包含无细胞脱氧核糖核酸(cfDNA)分子或无细胞核糖核酸(cfRNA)分子。

27.权利要求23所述的方法,其中估计所述总体肿瘤负荷包括组合来自跨越所述潜在SNV位点的所述多个序列读取的信息,以降低所述多个序列读取中由在一个或更多个所述潜在SNV位点处的所述测序中的误差引起的噪声。

28.权利要求25所述的方法,其中重新估计所述总体肿瘤负荷包括组合来自跨越所述经过滤的一组体细胞SNV的所述多个序列读取的信息,以降低所述多个序列读取中由在所述经过滤的一组体细胞SNV的一个或更多个处的所述测序中的误差引起的噪声。

29.权利要求27所述的方法,其中组合所述信息包括在给定所述总体肿瘤负荷的情况下,计算使在多个预定SNV热点中的每一个处观察到所述多个序列读取的似然最大化的值。

30.权利要求23所述的方法,其中确定所述遗传基因座的所述似然包括在给定一个或更多个基因型的情况下,确定观察到覆盖所述遗传基因座的所述多个序列读取中的数个序列读取的似然。

31.权利要求23所述的方法,其中确定所述遗传基因座的所述似然包括计算最大后验概率估计。

32.权利要求23所述的方法,其中所述概率模型包含机器学习模型,所述机器学习模型被配置为对具有真实变体的序列读取与具有测序误差的序列读取进行分类。

33.用于从对象的多个无细胞核酸(cfNA)分子中检测体细胞单核苷酸变体(SNV)的系统,其包含:

存储多个序列读取的数据库,其中所述多个序列读取的至少一个子集包含来自所述多个cfNA分子或其衍生物的序列;和

有效地偶连至所述数据库的一个或更多个计算机处理器,其中所述一个或更多个计算机处理器被单独地或共同地编程为:

(1)从所述数据库中调取所述多个序列读取;

(2)将概率模型应用于多个遗传基因座中的每一个处的所述多个序列读取,以估计所述多个cfNA分子的总体肿瘤负荷,其中所述估计的总体肿瘤负荷包含所述多个cfNA分子中肿瘤来源的cfNA分子的定量测量,其中所述多个遗传基因座包含潜在的SNV位点;

(3)对于所述多个遗传基因座中的每一个,至少部分地基于所述总体肿瘤负荷,确定所述对象的一种或更多种基因型的似然,其中所述一种或更多种基因型选自正常基因型、肿瘤基因型和联合正常肿瘤基因型;

(4)对于所述多个遗传基因座中的每一个,至少部分地基于(3)中确定的所述一种或更多种基因型的所述似然和所述多个cfNA分子中肿瘤来源的cfNA分子的所述定量测量,检测一种或更多种SNV;以及

(5)从(4)中检测的所述一种或更多种SNV中过滤出一种或更多种种系多态性,从而获得一种或更多种体细胞SNV。

34.权利要求33所述的系统,其中所述一个或更多个计算机处理器被单独地或共同地编程为另外地:

(6)使用选自以下的一种或更多种过滤器过滤出在(5)中获得的所述一种或更多种体细胞SNV,从而获得经过滤的一组体细胞SNV:链偏倚过滤器、碱基质量过滤器、读取伴侣过

滤器、测序误差过滤器、插入或缺失(插失)和均聚物引发的误差过滤器和公共数据库过滤器。

35.权利要求34所述的系统,其中所述一个或更多个计算机处理器被单独地或共同地编程为另外地:

(7)将所述概率模型应用于所述多个遗传基因座中的每一个处的所述多个序列读取以重新估计所述总体肿瘤负荷,其中所述多个遗传基因座包含在(6)中获得的所述经过滤的一组体细胞SNV。

36.权利要求33所述的系统,其中所述多个无细胞核酸(cfNA)分子包含无细胞脱氧核糖核酸(cfDNA)分子或无细胞核糖核酸(cfRNA)分子。

37.权利要求33所述的系统,其中估计所述总体肿瘤负荷包括组合来自跨越所述潜在SNV位点的所述多个序列读取的信息,以降低所述多个序列读取中由在一个或更多个所述潜在SNV位点处的所述测序中的误差引起的噪声。

38.权利要求35所述的系统,其中重新估计所述总体肿瘤负荷包括组合来自跨越所述经过滤的一组体细胞SNV的所述多个序列读取的信息,以降低所述多个序列读取中由在所述经过滤的一组体细胞SNV的一个或更多个处的所述测序中的误差引起的噪声。

39.权利要求37所述的系统,其中组合所述信息包括在给定所述总体肿瘤负荷的情况下,计算使在多个预定SNV热点中的每一个处观察到所述多个序列读取的似然最大化的值。

40.权利要求33所述的系统,其中确定所述遗传基因座的所述似然包括在给定一个或更多个基因型的情况下,确定观察到覆盖所述遗传基因座的所述多个序列读取中的数个序列读取的似然。

41.权利要求33所述的系统,其中确定所述遗传基因座的所述似然包括计算最大后验概率估计。

42.权利要求33所述的系统,其中所述概率模型包含机器学习模型,所述机器学习模型被配置为对具有真实变体的序列读取与具有测序误差的序列读取进行分类。

43.用于治疗患有癌症的患者的方法,其包括在基于从包含脱氧核糖核酸(cfDNA)和/或核糖核酸(cfRNA)的无细胞核酸中检测体细胞单核苷酸变体(SNV)而确定患者患有癌症之后,向所述患者施用癌症治疗,其中检测SNV包括:

对于包含序列数据集的cfDNA样品,估计总体肿瘤cfDNA分数;

对于测序数据集中的至少一个基因座k,确定基因型似然;

从包含序列数据集的所述cfDNA样品中消除种系多态性;

通过一组过滤器过滤SNV候选;以及

使用并入了所估计的总体肿瘤cfDNA分数的概率模型分析所述cfDNA样品,以确定所述cfDNA样品中肿瘤来源的DNA(ctDNA)的分数。

44.权利要求43所述的方法,其中对于所述cfDNA样品,估计所述总体肿瘤cfDNA分数包括:

组合来自所述cfDNA样品中所有潜在SNV位点的信息,以降低由在一个或更多个所述潜在SNV位点处发生的测序误差引起的噪声。

45.权利要求43或44所述的方法,其中确定所述基因型似然包括:

在给定不同联合基因型的情况下计算所观察到的读取覆盖所述基因座的似然。

46.权利要求43至45中任一项所述的方法,其中确定所述基因型似然包括:

估计基因座处的基因型,其使得后验概率最大化。

47.权利要求46所述的方法,其中估计所述基因座处的基因型包括:

确定所述序列数据集中每个基因座的联合基因型。

48.权利要求43至47中任一项所述的方法,其中过滤所述SNV候选包括以下中的至少一个:

基于链偏倚过滤器过滤SNV候选;

基于碱基质量过滤器过滤SNV候选;

基于读取伴侣过滤器过滤测序读取;

基于读取伴侣过滤器过滤SNV候选;

基于对真实变体与测序误差进行分类的机器学习模型过滤测序读取;

基于序列背景过滤器过滤SNV候选;以及

基于公共数据库过滤SNV候选。

49.权利要求48所述的方法,其中基于所述对真实变体与测序误差进行分类的机器学习模型过滤所述测序读取包括:

建立包含真实变体或测序误差的测序读取的基准真实训练数据;

创建每个测序读取的特征谱,其具有包含以下的信息:所述读取中每个碱基的测序质量,读取比对信息,序列背景(例如读取序列和插入/缺失)以及双端测序数据的插入大小;

基于训练数据训练分类器,以通过使用每个读取的特征谱对具有真实变体的测序读取和具有测序误差的读取进行分类;以及使用经训练的分类器将每个cfDNA测序读取分类为具有真实变体的读取或具有测序误差的读取。

50.权利要求43至49中任一项所述的方法,其中所述癌症治疗包含化学治疗、放射、手术、免疫治疗、细胞治疗、质子治疗或其组合。

51.用于治疗患有微小残留病变(MRD)的患者的方法,其包括对确定患有MRD的患者施用癌症治疗,其中从包括以下的步骤中确定所述患者患有MRD:从手术前血液样品和切除的肿瘤样品中的至少一种中鉴定一种或更多种截短突变和所述一种或更多种截短突变的突变谱;以及

在所述手术之后使用随访血浆cfDNA样品检测MRD。

52.权利要求51所述的方法,其中在所述手术之后使用随访血浆cfDNA样品检测MRD包括:

提取覆盖截短突变位置的读取;以及

使用被分类为具有真实变体的读取来计算MRD预测得分。

53.权利要求52所述的方法,其中使用被分类为具有真实变体的所述读取来计算所述MRD预测得分包括:

对基因组中的k个位点进行采样,所述位点不包含已鉴定的突变,但匹配k个截短突变的特征;

过滤被鉴定为包含误差的读取;以及

生成所述MRD预测得分。

54.权利要求51至77中任一项所述的方法,其中所述癌症治疗包含化学治疗、放射、手

术、免疫治疗、细胞治疗、质子治疗或其组合。

55.用于治疗患有癌症的患者的方法,其包括对在从包含来自所述患者的生物样品的脱氧核糖核酸(cfDNA)和/或核糖核酸(cfRNA)的无细胞核酸中检测体细胞单核苷酸变体(SNV)之后确定患有癌症的患者施用癌症治疗,其中SNV通过包括以下的方法检测:

对于包含测序数据集的cfDNA/cfRNA样品,合并重叠读取伴侣;

对于包含序列数据集的cfDNA/cfRNA样品,估计总体肿瘤cfDNA分数;

对于所述测序数据集中的基因座,确定基因型似然;

从包含所述序列数据集的所述cfDNA/cfRNA样品中消除种系多态性;

使用并入了所估计的总体肿瘤cfDNA分数的概率模型分析所述cfDNA/cfRNA样品以确定所述cfDNA/cfRNA样品中肿瘤来源的DNA(ctDNA/ctRNA)的分数;

使用并入了cfDNA/cfRNA样品特性的一组过滤器,消除低质量的体细胞SNV候选;

从包含所述测序数据集的所述cfDNA/cfRNA样品中消除不一致的重叠读取伴侣;

使用区分测序误差与真实变体的机器学习模型消除具有测序误差的读取;以及

使用截短突变谱和所述机器学习模型从所述cfDNA/cfRNA样品中确定早期微小残留病变(MRD)。

56.权利要求55所述的方法,其中所述癌症治疗包含化学治疗、放射、手术、免疫治疗、细胞治疗、质子治疗或其组合。

57.用于从对象的多个无细胞核酸(cfNA)分子中检测体细胞单核苷酸变体(SNV)的方法,其包括:

(a)调取由测序仪产生的多个序列读取,其中所述多个序列读取的至少一个子集包含来自所述多个cfNA分子或其衍生物的序列;

(b)将概率模型应用于多个遗传基因座中的每一个处的所述多个序列读取,以估计所述多个cfNA分子的总体肿瘤负荷,其中所述估计的总体肿瘤负荷包含所述多个cfNA分子中肿瘤来源的cfNA分子的定量测量,其中所述多个遗传基因座包含潜在的SNV位点;

(c)对于所述多个遗传基因座中的每一个,至少部分地基于所述总体肿瘤负荷,确定所述对象的一种或更多种基因型的似然,其中所述一种或更多种基因型选自正常基因型、肿瘤基因型和联合正常肿瘤基因型;

(d)对于所述多个遗传基因座中的每一个,至少部分地基于(c)中确定的所述一种或更多种基因型的所述似然和所述多种cfNA分子中肿瘤来源的cfNA分子的所述定量测量,来检测一种或更多种SNV;以及

(e)从(d)中检测的所述一种或更多种SNV中过滤出一种或更多种种系多态性,从而获得一种或更多种体细胞SNV。

58.权利要求57所述的方法,其还包括:

(f)使用选自以下的一种或更多种过滤器过滤出在(e)中获得的所述一种或更多种体细胞SNV,从而获得经过滤的一组体细胞SNV:链偏倚过滤器、碱基质量过滤器、读取伴侣过滤器、测序误差过滤器、插入或缺失(插失)和均聚物引发的误差过滤器和公共数据库过滤器。

59.权利要求58所述的方法,其还包括:

(g)将所述概率模型应用于所述多个遗传基因座中的每一个处的所述多个序列读取以

重新估计所述总体肿瘤负荷，其中所述多个遗传基因座包含在(f)中获得的所述经过滤的一组体细胞SNV。

60.权利要求57至59中任一项所述的方法，其中所述多个无细胞核酸(cfNA)分子包含无细胞脱氧核糖核酸(cfDNA)分子或无细胞核糖核酸(cfRNA)分子。

61.权利要求57至60中任一项所述的方法，其中估计所述总体肿瘤负荷包括组合来自跨越所述潜在SNV位点的所述多个序列读取的信息，以降低所述多个序列读取中由在一个或更多个所述潜在SNV位点处的所述测序中的误差引起的噪声。

62.权利要求57至61中任一项所述的方法，其中重新估计所述总体肿瘤负荷包括组合来自跨越所述经过滤的一组体细胞SNV的所述多个序列读取的信息，以降低所述多个序列读取中由在所述经过滤的一组体细胞SNV的一个或更多个处的所述测序中的误差引起的噪声。

63.权利要求57至62中任一项所述的方法，其中组合所述信息、包括在给定所述总体肿瘤负荷的情况下，计算使在多个预定SNV热点中的每一个处观察到所述多个序列读取的似然最大化的值。

64.权利要求57至63中任一项所述的方法，其中确定所述遗传基因座的所述似然包括在给定一个或更多个基因型的情况下，确定观察到覆盖所述遗传基因座的所述多个序列读取中的数个序列读取的似然。

65.权利要求57至64中任一项所述的方法，其中确定所述遗传基因座的所述似然包括计算最大后验概率估计。

66.权利要求57至65中任一项所述的方法，其中所述概率模型包含机器学习模型，所述机器学习模型被配置为对具有真实变体的序列读取与具有测序误差的序列读取进行分类。

67.权利要求57至66中任一项所述的方法，其中所述癌症治疗包含化学治疗、放射、手术、免疫治疗、细胞治疗、质子治疗或其组合。

从无细胞核酸中检测体细胞单核苷酸变体并应用于微小残留病变监测

[0001] 相关申请的交叉引用

[0002] 本申请要求于2017年9月15日提交的美国临时专利申请No.62/559,366的权益,其通过引用整体明确地并入本文。

[0003] 政府权益声明

[0004] 本发明是在国立卫生研究院(National Institutes of Health)授予的HL108634的政府支持下完成的。政府拥有本发明的某些权利。

背景技术

[0005] 体细胞突变可在对象的一生中自始至终在细胞中积累。尽管大多数这样的突变可具有极少明显的作用或没有明显的作用,但一些可改变基因和/或关键的细胞功能,并因此产生表型变化。体细胞突变的产物可以是癌症,这是由于细胞克隆扩增以及从正常体细胞行为的内置(in-built)程序和细胞增殖的外源性限制二者的逃逸所致。触发癌症进展的体细胞突变可被称为“驱动突变(driver mutation)”,并且不导致表型或生物学后果的体细胞突变可被称为“过客突变(passenger mutation)”。

[0006] 分析可常见于肿瘤中的驱动突变对于分析癌症病理学、癌症诊断、精确肿瘤学和预后可能是必不可少的。因此,能够获得肿瘤的遗传谱在医学治疗和临床研究二者中均可以是重要的。根据一些方法,可从手术试样或活检试样获得肿瘤的遗传谱。然而,由于这些方法可能是昂贵的、侵入性的并且可危及对象或患者的健康,因此不能总是执行这些方法。

[0007] 值得注意的是,体细胞和肿瘤二者均可在细胞破坏(例如凋亡和坏死)期间将其核和线粒体的脱氧核糖核酸(DNA)和/或核糖核酸(RNA)释放到对象的血流中。作为结果,可在对象的血流中发现肿瘤细胞的遗传信息,并从血浆中提取肿瘤细胞的遗传信息。

[0008] 最近,可在血浆中发现的无细胞DNA(cell-free DNA,cfDNA)和无细胞RNA(cell-free RNA,cfRNA)已被认为是在癌症诊断和预后中具有巨大潜力的生物标志物。但是,由于cfDNA的混合(例如,包含肿瘤的和非肿瘤的二者)的性质,因此从cfDNA测序数据中检测与癌症相关的等位基因可能是困难的。例如,可能需要专门的分析来检测肿瘤来源的DNA(ctDNA)及其突变,因为其在cfDNA中的数量和质量可显著变化。例如,ctDNA中ctDNA的分数可为0.01%至97%,而对于早期癌症,该分数通常可低于10%。由于cfDNA的非侵入性特性,可轻松且顺次地获得血浆样品用于诊断和预后,从而在一段时间内提供与癌症相关的大量信息。因此,检测ctDNA中的癌症突变可为早期癌症诊断、监测和微小残留病变监测提供有前景的方案。

[0009] 关于检测单核苷酸变体(single nucleotide variant,SNV)或体细胞突变,针对不纯的实体瘤样品设计的许多现有方法(例如如VarScan2和MuTect)可能不容易用于肿瘤分数与cfDNA一样低的样品。不考虑cfDNA的特殊特性,即使在低肿瘤分数下有充分的证据,这样的方法也可能无法鉴定体细胞SNV。此外,开发用于检测cfDNA中的体细胞SNV的方法包括具有iDES和SiNVICT的Capp-Seq。具有iDES的Capp-Seq可指基于靶向的测序组的实验性

误差抑制技术,其可以灵敏地检测对象(例如中至晚期癌症患者)的血液cfDNA样品中的癌症突变。该方法可通过深度测序以捕获cfDNA中的肿瘤信号来解决低肿瘤分数的问题。但是,由于深度测序的昂贵成本,它可仅对一小部分基因组区域进行测序,从而限制了其对不同癌症类型的检测能力以及对个性化治疗和结果预测的应用。SiNVICT可指计算SNV调用者(caller),其使用泊松函数(Poisson function)模拟cfDNA中的等位基因分布。此概率模型可能过于简单,并且可能无法考虑cfDNA的不纯性质,从而导致大量假阳性。此外,这些方法可能无法考虑cfDNA测序数据中的可为检测cfDNA中的突变提供有价值信息的重叠读取伴侣(overlapping read mate)。

[0010] 发明概述

[0011] 鉴于前述内容,本公开内容提供了新的概率方法,以从无细胞脱氧核糖核酸(cfDNA)或无细胞的核糖核酸(cfRNA)中检测体细胞单核苷酸变体(SNV),所述方法不仅能解决cfDNA/cfRNA测序数据中的独特挑战(例如,低肿瘤分数和低基因组覆盖和高重叠率(对于双端测序数据)),而且还估计cfDNA/cfRNA(以下通常称为cfDNA或cfRNA)中的肿瘤分数用于癌症诊断、预后和手术之后的微小残留病变(minimal residual disease,MRD)检测。具体地,本公开内容提供了利用cfDNA检查进行非侵入性癌症检测和诊断的方法。在这样做时,本公开内容的方法和系统可使用基于贝叶斯(Bayesian-based)的概率框架来从测序数据集中估计某种基因型的似然。后验估算最大化可用于确定对象(例如患者)的基因型并调用体细胞SNV候选。然后,可用在链偏倚、变体质量、序列背景和已知的单核苷酸多态性(single nucleotide polymorphism,SNP)方面适应cfDNA特性的一组标准过滤体细胞SNV候选。此外,可考虑血浆样品中的测序误差和低的肿瘤变体等位基因频率二者。这种方法可有助于检测体细胞SNV并确保良好的性能。这种方法可包括如下所述的一个或更多个步骤。

[0012] 根据一个实施方案,第一步骤包括通过将基因座与测序数据中观察到的变体等位基因总体组合来估计肿瘤分数。总体肿瘤分数估计可有效抑制测序噪声,并且避免来自采样波动和测序误差(尤其是低或中深度测序)的严重影响。这不仅可改善准确的体细胞SNV检测,而且自身还可以是在癌症诊断和MRD检测中具有临床价值的标志物。

[0013] 根据一个实施方案,第二步骤包括通过在给定总体肿瘤分数的情况下使联合肿瘤正常基因型的后验最大化来确定SNV。这种方法可与对包含参考等位基因和变体等位基因的读取进行比较的其他方法形成对比。

[0014] 根据一个实施方案,第三步骤包括利用就链偏倚、变体质量、序列背景和已知的SNP而言的一组cfDNA特征标准来过滤体细胞SNV候选。这样的调用后(post-call)过滤确保了cfDNA中变体调用的高质量。

[0015] 根据一个实施方案,第四步骤包括在过滤之后重新估计体细胞SNV。如关于第一步骤所述的,这可使用总体肿瘤分数来完成。

[0016] 本公开内容的方法和系统可以是特别有利的,因为(1)在第一步骤中获得的总体肿瘤分数即使在低肿瘤分数下也可在单独基因座处提供肿瘤读取的预期水平,使得具有低突变体频率的体细胞SNV仍可与随机测序误差区分开;(2)在存在混合正常的和肿瘤来源的cfDNA的情况下,联合基因型模型可更好地拟合cfDNA测序数据,因为其通过同时对正常的和肿瘤cfDNA进行基因分型而充分利用正常信号和肿瘤信号二者;(3)通过机器学习模型进行的计算测序误差抑制可基于单独读取信息有效区分真实变体和测序误差;以及(4)可基

于cfDNA特性特别地设计体细胞SNV候选的后过滤(Post-filtration)。

[0017] 鉴于这些有利特征,本公开内容的方法和系统不仅可应用于深靶向的cfDNA测序数据,而且可应用于中等深度的cfDNA测序数据,例如整个外显子组测序数据。因此,本公开内容的方法和系统可基于输入数据从cfDNA提供灵敏的突变调用和癌症的突变谱。

[0018] 作为本发明构思的示例性应用,使用本公开内容的方法和系统得到的突变谱可用于手术之后微小残留病变(MRD)的早期检测。该检测可基于已知的体细胞突变,所述体细胞突变是从患者的匹配肿瘤正常样品中或在患者手术之前从患者血浆中的cfDNA中检测到的。可基于成对的突变调用结果来选择截短的突变以进行监测。与随访(follow-up)血浆样品中的背景相比,MRD可通过截短突变统计上显著的累积发生来确定。

[0019] 根据一个方面,本公开内容提供了用于从无细胞核酸例如脱氧核糖核酸(cfDNA)和核糖核酸(cfRNA)中检测体细胞单核苷酸变体(SNV)的方法,所述方法包括:对于包含序列数据集的cfDNA样品,估计总体肿瘤cfDNA分数;对于测序数据集中的至少一个基因座k,确定基因型似然;从包含序列数据集的cfDNA样品中消除种系多态性;通过一组过滤器过滤SNV候选;以及使用并入了所估计的总体肿瘤cfDNA分数的概率模型分析cfDNA样品以确定cfDNA样品中肿瘤来源的DNA(ctDNA)的分数。

[0020] 在一些实施方案中,对于cfDNA样品,估计总体肿瘤cfDNA分数包括:组合来自cfDNA样品中所有潜在SNV位点的信息以降低由在一个或更多个潜在SNV位点处发生的测序误差引起的噪声。在一些实施方案中,确定基因型似然包括:在给定不同联合基因型的情况下计算观察到覆盖基因座的读取的似然。在一些实施方案中,确定基因型似然包括:估计基因座处的基因型,其使得后验概率最大化。在一些实施方案中,估计基因座处的基因型包括:确定序列数据集中每个基因座的联合基因型。在一些实施方案中,过滤SNV候选包括以下中的至少一个:基于链偏倚过滤器过滤SNV候选;基于碱基质量过滤器过滤SNV候选;基于读取伴侣过滤器过滤测序读取;基于读取伴侣过滤器过滤SNV候选;以及基于对真实变体与测序误差进行分类的机器学习模型过滤测序读取。在一些实施方案中,基于对真实变体与测序误差进行分类的机器学习模型过滤测序读取包括:建立包含真实变体或测序误差的测序读取的基准真实训练数据(ground-truth training data);创建每个测序读取的特征谱,其具有包含以下的信息:读取中每个碱基的测序质量,读取比对信息,序列背景(例如读取序列和插入/缺失)以及双端测序数据的插入大小;基于训练数据训练分类器以通过使用每个读取的特征谱对具有真实变体的测序读取和具有测序误差的读取进行分类;以及使用经训练的分类器将每个cfDNA测序读取分类为具有真实变体的读取或具有测序误差的读取。

[0021] 根据另一方面,本公开内容提供了用于从无细胞核酸例如脱氧核糖核酸(cfDNA)和核糖核酸(cfRNA)中检测体细胞单核苷酸变体(SNV)的系统,所述系统包含:计算机存储器;通信偶连至计算机存储器的一个或更多个计算机处理器,所述一个或更多个计算机处理器被配置为实现包括以下的方法:对于包含序列数据集的cfDNA样品,估计总体肿瘤cfDNA分数;对于测序数据集中的至少一个基因座k,确定基因型似然;从包含序列数据集的cfDNA样品中消除种系多态性;通过一组过滤器过滤SNV候选;以及使用并入了所估计的总体肿瘤cfDNA分数的概率模型分析cfDNA样品以确定cfDNA样品中肿瘤来源的DNA(ctDNA)的分数。

[0022] 在一些实施方案中，对于cfDNA样品，估计总体肿瘤cfDNA分数包括：组合来自cfDNA样品中所有潜在SNV位点的信息以降低由在一个或更多个潜在SNV位点处发生的测序误差引起的噪声。在一些实施方案中，确定基因型似然包括：在给定不同联合基因型的情况下计算观察到覆盖基因座的读取的似然。在一些实施方案中，确定基因型似然包括：估计基因座处的基因型，其使得后验概率最大化。在一些实施方案中，估计基因座处的基因型包括：确定序列数据集中每个基因座的联合基因型。在一些实施方案中，过滤SNV候选包括以下中的至少一个：基于链偏倚过滤器过滤SNV候选；基于碱基质量过滤器过滤SNV候选；基于读取伴侣过滤器过滤测序读取；基于读取伴侣过滤器过滤SNV候选；以及基于对真实变体与测序误差进行分类的机器学习模型过滤测序读取。在一些实施方案中，基于对真实变体与测序误差进行分类的机器学习模型过滤测序读取包括：建立包含真实变体或测序误差的测序读取的基准真实训练数据；创建每个测序读取的特征谱，其具有包含以下的信息：读取中每个碱基的测序质量，读取比对信息，序列背景(例如读取序列和插入/缺失)以及双端测序数据的插入大小；基于训练数据训练分类器以通过使用每个读取的特征谱对具有真实变体的测序读取和具有测序误差的读取进行分类；以及使用经训练的分类器将每个cfDNA测序读取分类为包含具有真实变体的读取或包含具有测序误差的读取。

[0023] 根据另一方面，本公开内容提供了用于从手术之前和之后收集的血浆样品、白细胞和切除的肿瘤样品(如果有的话)中检测微小残留病变(MRD)的方法，所述方法包括：从手术前血液样品和/或切除的肿瘤样品中鉴定一种或更多种截短突变和一种或更多种截短突变的每一种的突变谱；以及在手术之后使用随访血浆cfDNA样品检测MRD。

[0024] 在一些实施方案中，在手术之后使用随访血浆cfDNA样品检测MRD包括：提取覆盖截短突变位置的读取，以及使用被分类为具有真实变体的读取来计算MRD预测得分。在一些实施方案中，使用被分类为具有真实变体的读取来计算MRD预测得分包括：对基因组中的k个位点进行采样，所述位点不包含已鉴定的突变但匹配那些k个截短突变的特征；过滤被鉴定为包含误差的读取；以及生成MRD预测得分。

[0025] 根据另一方面，本公开内容提供了用于从手术之前和之后收集的血浆样品、白细胞和切除的肿瘤样品(如果有的话)中检测微小残留病变(MRD)的系统，所述系统包含：计算机存储器；通信偶连至计算机存储器的一个或更多个计算机处理器，所述一个或更多个计算机处理器被配置为实现包括以下的方法：从手术前血液样品和/或切除的肿瘤样品中鉴定一种或更多种截短突变和一种或更多种截短突变的每一种的突变谱；以及在手术之后使用随访血浆cfDNA样品检测MRD。

[0026] 在一些实施方案中，在手术之后使用随访血浆cfDNA样品检测MRD包括：提取覆盖截短突变位置的读取，以及使用被分类为具有真实变体的读取来计算MRD预测得分。在一些实施方案中，使用被分类为具有真实变体的读取来计算MRD预测得分包括：对基因组中的k个位点进行采样，所述位点不包含已鉴定的突变但匹配那些k个截短突变的特征；过滤被鉴定为包含误差的读取；以及生成MRD预测得分。

[0027] 根据另一方面，本公开内容提供了用于从手术之前和之后收集的血浆样品、白细胞和切除的肿瘤样品(如果有的话)中检测微小残留病变(MRD)的系统，所述系统包含：计算机存储器；通信偶连至计算机存储器的一个或更多个计算机处理器，所述一个或更多个计算机处理器被配置为实现包括以下的方法：从手术前血液样品和/或切除的肿瘤样品中鉴

定一种或更多种截短突变和一种或更多种截短突变的每一种的突变谱;以及在手术之后使用随访血浆cfDNA样品检测MRD。

[0028] 在一些实施方案中,在手术之后使用随访血浆cfDNA样品检测MRD包括:提取覆盖截短突变位置的读取,以及使用被分类为具有真实变体的读取来计算MRD预测得分。在一些实施方案中,使用被分类为具有真实变体的读取来计算MRD预测得分包括:对基因组中的k个位点进行采样,所述位点不包含已鉴定的突变但匹配那些k个截短突变的特征;过滤被鉴定为包含误差的读取;以及生成MRD预测得分。

[0029] 根据另一方面,本公开内容提供了存储指令集的非暂时性存储介质,当执行所述指令时使一个或更多个计算机处理器从无细胞核酸例如脱氧核糖核酸(cfDNA)和核糖核酸(cfRNA)中检测体细胞单核苷酸变体(SNV),所述指令集包含以下指令:对于包含测序数据集的cfDNA/cfRNA样品,重叠读取伴侣;对于包含序列数据集的cfDNA/cfRNA样品,总体肿瘤cfDNA分数;对于测序数据集中的基因座,基因型似然;从包含序列数据集的cfDNA/cfRNA样品中消除种系多态性;使用并入了所估计的总体肿瘤cfDNA分数的概率模型分析cfDNA/cfRNA样品以确定cfDNA/cfRNA样品中肿瘤来源的DNA(ctDNA/ctRNA)的分数;使用并入了cfDNA/cfRNA样品特性的一组过滤器,消除低质量的体细胞SNV候选;从包含测序数据集的cfDNA/cfRNA样品中消除不一致的重叠读取伴侣;使用区分测序误差与真实变体的机器学习模型消除具有测序误差的读取;以及使用截短突变谱和机器学习模型从cfDNA/cfRNA样品中确定早期微小残留病变(MRD)。

[0030] 在一些实施方案中,对于cfDNA/cfRNA样品,估计总体肿瘤cfDNA分数包括:组合来自cfDNA/cfRNA样品中所有潜在SNV位点的信息以降低由在一个或更多个潜在SNV位点处发生的测序误差引起的噪声。在一些实施方案中,确定基因型似然包括:在给定不同联合基因型的情况下计算观察到覆盖基因座的读取的似然。在一些实施方案中,确定基因型似然包括:估计基因座处的基因型,其使得后验概率最大化。在一些实施方案中,估计基因座处的基因型包括:确定序列数据集中每个基因座的联合基因型。在一些实施方案中,消除具有测序误差的读取包括:使用多次测序的样品训练机器学习模型;从测序数据集中的真实变体或测序误差中确定碱基。在一些实施方案中,从cfDNA/cfRNA样品中确定早期微小残留病变(MRD)包括:从从cfDNA/cfRNA样品调用的突变谱和变体等位基因频率确定截短突变;组合多个变体基因座以增强超低肿瘤分数样品中的肿瘤信号;以及通过样品内(within-sample)统计测试确定MRD状态。

[0031] 根据另一方面,本公开内容提供了用于从无细胞核酸例如脱氧核糖核酸(cfDNA)和核糖核酸(cfRNA)中检测体细胞单核苷酸变体(SNV)的方法,所述方法包括:对于包含测序数据集的cfDNA/cfRNA样品,合并重叠读取伴侣;对于包含序列数据集的cfDNA/cfRNA样品,估计总体肿瘤cfDNA分数;对于测序数据集中的基因座,确定基因型似然;从包含序列数据集的cfDNA/cfRNA样品中消除种系多态性;使用并入了所估计的总体肿瘤cfDNA分数的概率模型分析cfDNA/cfRNA样品以确定cfDNA/cfRNA样品中肿瘤来源的DNA(ctDNA/ctRNA)的分数;使用并入了cfDNA/cfRNA样品特性的一组过滤器,消除低质量的体细胞SNV候选;从包含测序数据集的cfDNA/cfRNA样品中消除不一致的重叠读取伴侣;使用区分测序误差与真实变体的机器学习模型消除具有测序误差的读取;以及使用截短突变谱和机器学习模型从cfDNA/cfRNA样品中确定早期微小残留病变(MRD)。

[0032] 在一些实施方案中，对于cfDNA/cfRNA样品，估计总体肿瘤cfDNA分数包括：组合来自cfDNA/cfRNA样品中所有潜在SNV位点的信息以降低由在一个或更多个潜在SNV位点处发生的测序误差引起的噪声。在一些实施方案中，确定基因型似然包括：在给定不同联合基因型的情况下计算观察到覆盖基因座的读取的似然。在一些实施方案中，确定基因型似然包括：估计基因座处的基因型，其使得后验概率最大化。在一些实施方案中，估计基因座处的基因型包括：确定序列数据集中每个基因座的联合基因型。在一些实施方案中，消除具有测序误差的读取包括：使用多次测序的样品训练机器学习模型；从测序数据集中的真实变体或测序误差中确定碱基。在一些实施方案中，从cfDNA/cfRNA样品中确定早期微小残留病变(MRD)包括：从从cfDNA/cfRNA样品调用的突变谱和变体等位基因频率确定截短突变；组合多个变体基因座以增强超低肿瘤分数样品中的肿瘤信号；以及通过样品内统计测试确定MRD状态。

[0033] 根据另一方面，本公开内容提供了用于从对象的多个无细胞核酸(cfNA)分子中检测体细胞单核苷酸变体(SNV)的方法，其包括：(a)调取(retrieve)由测序仪产生的多个序列读取，其中所述多个序列读取的至少一个子集包含来自所述多个cfNA分子或其衍生物的序列；(b)将概率模型应用于多个遗传基因座中的每一个处的所述多个序列读取，以估计所述多个cfNA分子的总体肿瘤负荷，其中所述估计的总体肿瘤负荷包含所述多个cfNA分子中肿瘤来源的cfNA分子的定量测量，其中所述多个遗传基因座包含潜在的SNV位点；(c)对于所述多个遗传基因座中的每一个，至少部分地基于所述总体肿瘤负荷，确定所述对象的一种或更多种基因型的似然，其中所述一种或更多种基因型选自正常基因型、肿瘤基因型和联合正常肿瘤基因型；(d)对于所述多个遗传基因座中的每一个，至少部分地基于(c)中确定的所述一种或更多种基因型的所述似然和所述多种cfNA分子中肿瘤来源的cfNA分子的所述定量测量，检测一种或更多种SNV；以及(e)从(d)中检测的所述一种或更多种SNV中过滤出一种或更多种种系多态性，从而获得一种或更多种体细胞SNV。

[0034] 在一些实施方案中，所述方法还包括(f)使用选自以下的一种或更多种过滤器过滤出在(e)中获得的所述一种或更多种体细胞SNV，从而获得经过滤的一组体细胞SNV：链偏倚过滤器、碱基质量过滤器、读取伴侣过滤器、测序误差过滤器、插入或缺失(插失(indel))和均聚物引发的误差过滤器和公共数据库过滤器。在一些实施方案中，所述方法还包括(g)将所述概率模型应用于所述多个遗传基因座中的每一个处的所述多个序列读取以重新估计所述总体肿瘤负荷，其中所述多个遗传基因座包含在(f)中获得的所述经过滤的一组体细胞SNV。在一些实施方案中，所述多个无细胞核酸(cfNA)分子包含无细胞脱氧核糖核酸(cfDNA)分子。在一些实施方案中，所述多个无细胞核酸(cfNA)分子包含无细胞核糖核酸(cfRNA)分子。在一些实施方案中，估计所述总体肿瘤负荷包括组合来自跨越所述潜在SNV位点的所述多个序列读取的信息，以降低由在一个或更多个所述潜在SNV位点处的所述测序中的误差引起的所述多个序列读取中的噪声。在一些实施方案中，重新估计所述总体肿瘤负荷包括组合来自跨越所述经过滤的一组体细胞SNV的所述多个序列读取的信息，以降低由在所述经过滤的一组体细胞SNV的一个或更多个处的所述测序中的误差引起的所述多个序列读取中的噪声。在一些实施方案中，组合所述信息包括计算在所述多个序列读取中包含多个预定SNV热点中的SNV的序列读取的总比例。在一些实施方案中，组合所述信息包括在给定所述总体肿瘤负荷的情况下，计算使在多个预定SNV热点中的每一个处观察到所

述多个序列读取的似然最大化的值。在一些实施方案中，组合所述信息包括计算多个预定的SNV热点中变体等位基因频率值的最大值。在一些实施方案中，所述信息包含碱基调用、碱基质量、映射质量或其组合。在一些实施方案中，确定所述遗传基因座的所述似然包括在给定一个或更多个基因型的情况下，确定观察到覆盖所述遗传基因座的所述多个序列读取中的数个序列读取的似然。在一些实施方案中，确定所述遗传基因座的所述似然包括计算最大后验概率估计。在一些实施方案中，确定所述遗传基因座的所述似然包括确定所述遗传基因座的联合正常肿瘤基因型。在一些实施方案中，使用选自以下的一种或更多种过滤器来执行所述过滤：链偏倚过滤器、碱基质量过滤器、读取伴侣过滤器、序列背景过滤器和测序误差过滤器。在一些实施方案中，所述概率模型包含机器学习模型，所述机器学习模型被配置为对具有真实变体的序列读取与具有测序误差的序列读取进行分类。

[0035] 在一些实施方案中，所述方法还包括使所述多个cfNA分子经历扩增。在一些实施方案中，所述扩增包括聚合酶链反应(polymerase chain reaction，PCR)。在一些实施方案中，所述方法还包括针对参考来处理所述检测到的体细胞SNV。在一些实施方案中，所述参考包含从一个或更多个另外的对象的多个cfNA分子中检测到的第二组体细胞SNV。在一些实施方案中，所述多个cfNA分子获自所述对象的身体样品。在一些实施方案中，所述身体样品选自血浆、血清、骨髓、脑脊液、胸膜液、唾液、粪便和尿。在一些实施方案中，所述方法还包括处理所述检测到的体细胞SNV以产生所述对象患有或被怀疑患有疾病或病症的似然。在一些实施方案中，所述疾病或病症是选自以下的癌症：胰腺癌、肝癌、肺癌、结直肠癌、白血病、膀胱癌、骨癌、脑癌、乳腺癌、宫颈癌、子宫内膜癌、食管癌、胃癌、头颈癌、黑素瘤、卵巢癌、睾丸癌、肾癌、肉瘤、胆管癌和前列腺癌。在一些实施方案中，所述方法还包括对所述多个cfNA分子或其衍生物进行测序以产生所述多个序列读取。在一些实施方案中，所述体细胞SNV包含癌症驱动的突变。

[0036] 根据另一方面，本公开内容提供了用于从对象的多个无细胞核酸(cfNA)分子中检测体细胞单核苷酸变体(SNV)的系统，其包含：存储多个序列读取的数据库，其中所述多个序列读取的至少一个子集包含来自所述多个cfNA分子或其衍生物的序列；和有效地偶连至所述数据库的一个或更多个计算机处理器，其中所述一个或更多个计算机处理器被单独地或共同地编程为：(1)从所述数据库中调取所述多个序列读取；(2)将概率模型应用于多个遗传基因座中的每一个处的所述多个序列读取，以估计所述多个cfNA分子的总体肿瘤负荷，其中所述估计的总体肿瘤负荷包含所述多个cfNA分子中肿瘤来源的cfNA分子的定量测量，其中所述多个遗传基因座包含潜在的SNV位点；(3)对于所述多个遗传基因座中的每一个，至少部分地基于所述总体肿瘤负荷，确定所述对象的一种或更多种基因型的似然，其中所述一种或更多种基因型选自正常基因型、肿瘤基因型和联合正常肿瘤基因型；(4)对于所述多个遗传基因座中的每一个，至少部分地基于(3)中确定的所述一种或更多种基因型的所述似然和所述多个cfNA分子中肿瘤来源的cfNA分子的所述定量测量，检测一种或更多种SNV；以及(5)从(4)中检测的所述一种或更多种SNV中过滤出一种或更多种种系多态性，从而获得一种或更多种体细胞SNV。

[0037] 在一些实施方案中，所述一个或更多个计算机处理器被单独地或共同地编程为另外地(6)使用选自以下的一种或更多种过滤器过滤出在(5)中获得的所述一种或更多种体细胞SNV，从而获得经过滤的一组体细胞SNV：链偏倚过滤器、碱基质量过滤器、读取伴侣过

滤器、测序误差过滤器、插入或缺失(插失)和均聚物引发的误差过滤器和公共数据库过滤器。在一些实施方案中,所述一个或更多个计算机处理器被单独地或共同地编程为另外地(7)将所述概率模型应用于所述多个遗传基因座中的每一个处的所述多个序列读取以重新估计所述总体肿瘤负荷,其中所述多个遗传基因座包含在(6)中获得的所述经过滤的一组体细胞SNV。在一些实施方案中,所述多个无细胞核酸(cfNA)分子包含无细胞脱氧核糖核酸(cfDNA)分子。在一些实施方案中,所述多个无细胞核酸(cfNA)分子包含无细胞核糖核酸(cfRNA)分子。在一些实施方案中,估计所述总体肿瘤负荷包括组合来自跨越所述潜在SNV位点的所述多个序列读取的信息,以降低由在一个或更多个所述潜在SNV位点处的所述测序中的误差引起的所述多个序列读取中的噪声。在一些实施方案中,重新估计所述总体肿瘤负荷包括组合来自跨越所述经过滤的一组体细胞SNV的所述多个序列读取的信息,以降低由在所述经过滤的一组体细胞SNV的一个或更多个处的所述测序中的误差引起的所述多个序列读取中的噪声。在一些实施方案中,组合所述信息包括计算在所述多个序列读取中包含多个预定SNV热点中的SNV的序列读取的总比例。在一些实施方案中,组合所述信息包括在给定所述总体肿瘤负荷的情况下,计算使在多个预定SNV热点中的每一个处观察到所述多个序列读取的似然最大化的值。在一些实施方案中,组合所述信息包括计算多个预定的SNV热点中变体等位基因频率值的最大值。在一些实施方案中,所述信息包含碱基调用、碱基质量、映射质量或其组合。在一些实施方案中,确定所述遗传基因座的所述似然包括在给定一个或更多个基因型的情况下,确定观察到覆盖所述遗传基因座的所述多个序列读取中的数个序列读取的似然。在一些实施方案中,确定所述遗传基因座的所述似然包括计算最大后验概率估计。在一些实施方案中,确定所述遗传基因座的所述似然包括确定所述遗传基因座的联合正常肿瘤基因型。在一些实施方案中,使用选自以下的一种或更多种过滤器来执行所述过滤:链偏倚过滤器、碱基质量过滤器、读取伴侣过滤器、序列背景过滤器和测序误差过滤器。在一些实施方案中,所述概率模型包含机器学习模型,所述机器学习模型被配置为对具有真实变体的序列读取与具有测序误差的序列读取进行分类。

[0038] 在一些实施方案中,所述一个或更多个计算机处理器被单独地或共同地编程为针对参考来处理所述检测到的体细胞SNV。在一些实施方案中,所述参考包含从一个或更多个另外的对象的多个cfNA分子中检测到的第二组体细胞SNV。在一些实施方案中,所述多个cfNA分子获自所述对象的身体样品。在一些实施方案中,所述身体样品选自血浆、血清、骨髓、脑脊液、胸膜液、唾液、粪便和尿。在一些实施方案中,所述一个或更多个计算机处理器被单独地或共同地编程为处理所述检测到的体细胞SNV以产生所述对象患有或被怀疑患有疾病或病症的似然。在一些实施方案中,所述疾病或病症是选自以下的癌症:胰腺癌、肝癌、肺癌、结直肠癌、白血病、膀胱癌、骨癌、脑癌、乳腺癌、宫颈癌、子宫内膜癌、食管癌、胃癌、头颈癌、黑素瘤、卵巢癌、睾丸癌、肾癌、肉瘤、胆管癌和前列腺癌。在一些实施方案中,所述体细胞SNV包含癌症驱动的突变。

[0039] 根据另一方面,本公开内容提供了包含机器可执行代码的非暂时性计算机可读介质,当所述代码由一个或更多个计算机处理器执行时,实施了用于从对象的多个无细胞核酸(cfNA)分子中检测体细胞单核苷酸变体(SNV)的方法,所述方法包括:(a)调取由测序仪产生的多个序列读取,其中所述多个序列读取的至少一个子集包含来自所述多个cfNA分子或其衍生物的序列;(b)将概率模型应用于多个遗传基因座中的每一个处的所述多个序列

读取，以估计所述多个cfNA分子的总体肿瘤负荷，其中所述估计的总体肿瘤负荷包含所述多个cfNA分子中肿瘤来源的cfNA分子的定量测量，其中所述多个遗传基因座包含潜在的SNV位点；(c)对于所述多个遗传基因座中的每一个，至少部分地基于所述总体肿瘤负荷，确定所述对象的一种或更多种基因型的似然，其中所述一种或更多种基因型选自正常基因型、肿瘤基因型和联合正常肿瘤基因型；(d)对于所述多个遗传基因座中的每一个，至少部分地基于(c)中确定的所述一种或更多种基因型的所述似然和所述多种cfNA分子中肿瘤来源的cfNA分子的所述定量测量，检测一种或更多种SNV；以及(e)从(d)中检测的所述一种或更多种SNV中过滤出一种或更多种种系多态性，从而获得一种或更多种体细胞SNV。

[0040] 在一些实施方案中，所述方法还包括(f)使用选自以下的一种或更多种过滤器过滤出在(e)中获得的所述一种或更多种体细胞SNV，从而获得经过滤的一组体细胞SNV：链偏倚过滤器、碱基质量过滤器、读取伴侣过滤器、测序误差过滤器、插入或缺失(插失)和均聚物引发的误差过滤器和公共数据库过滤器。在一些实施方案中，所述方法还包括(g)将所述概率模型应用于所述多个遗传基因座中的每一个处的所述多个序列读取以重新估计所述总体肿瘤负荷，其中所述多个遗传基因座包含在(f)中获得的所述经过滤的一组体细胞SNV。在一些实施方案中，所述多个无细胞核酸(cfNA)分子包含无细胞脱氧核糖核酸(cfDNA)分子。在一些实施方案中，所述多个无细胞核酸(cfNA)分子包含无细胞核糖核酸(cfRNA)分子。在一些实施方案中，估计所述总体肿瘤负荷包括组合来自跨越所述潜在SNV位点的所述多个序列读取的信息，以降低由在一个或更多个所述潜在SNV位点处的所述测序中的误差引起的所述多个序列读取中的噪声。在一些实施方案中，重新估计所述总体肿瘤负荷包括组合来自跨越所述经过滤的一组体细胞SNV的所述多个序列读取的信息，以降低由在所述经过滤的一组体细胞SNV的一个或更多个处的所述测序中的误差引起的所述多个序列读取中的噪声。在一些实施方案中，组合所述信息包括计算在所述多个序列读取中包含多个预定SNV热点中的SNV的序列读取的总比例。在一些实施方案中，组合所述信息包括在给定所述总体肿瘤负荷的情况下，计算使在多个预定SNV热点中的每一个处观察到所述多个序列读取的似然最大化的值。在一些实施方案中，组合所述信息包括计算多个预定的SNV热点中变体等位基因频率值的最大值。在一些实施方案中，所述信息包含碱基调用、碱基质量、映射质量或其组合。在一些实施方案中，确定所述遗传基因座的所述似然包括在给定一个或更多个基因型的情况下，确定观察到覆盖所述遗传基因座的所述多个序列读取中的数个序列读取的似然。在一些实施方案中，确定所述遗传基因座的所述似然包括计算最大后验概率估计。在一些实施方案中，确定所述遗传基因座的所述似然包括确定所述遗传基因座的联合正常肿瘤基因型。在一些实施方案中，使用选自以下的一种或更多种过滤器来执行所述过滤：链偏倚过滤器、碱基质量过滤器、读取伴侣过滤器、序列背景过滤器和测序误差过滤器。在一些实施方案中，所述概率模型包含机器学习模型，所述机器学习模型被配置为对具有真实变体的序列读取与具有测序误差的序列读取进行分类。

[0041] 在一些实施方案中，所述方法还包括针对参考来处理所述检测到的体细胞SNV。在一些实施方案中，所述参考包含从一个或更多个另外的对象的多个cfNA分子中检测到的第二组体细胞SNV。在一些实施方案中，所述多个cfNA分子获自所述对象的身体样品。在一些实施方案中，所述身体样品选自血浆、血清、骨髓、脑脊液、胸膜液、唾液、粪便和尿。在一些实施方案中，所述方法还包括处理所述检测到的体细胞SNV以产生所述对象患有或被怀疑

患有疾病或病症的似然。在一些实施方案中，所述疾病或病症是选自以下的癌症：胰腺癌、肝癌、肺癌、结直肠癌、白血病、膀胱癌、骨癌、脑癌、乳腺癌、宫颈癌、子宫内膜癌、食管癌、胃癌、头颈癌、黑素瘤、卵巢癌、睾丸癌、肾癌、肉瘤、胆管癌和前列腺癌。在一些实施方案中，所述体细胞SNV包含癌症驱动的突变。

[0042] 前面已经相当广泛地概述了本发明的特征和技术优点，以便可以更好地理解以下对本发明的详细描述。下文将描述形成本发明权利要求书主题的本发明的另外的特征和优点。本领域技术人员应理解，所公开的概念和具体实施方案可以容易地用作修改或设计其他结构以实施本发明的相同目的的基础。本领域技术人员还应认识到，这样的等效构造不脱离如所附权利要求书中所阐述的本发明的精神和范围。当结合附图考虑时，就本发明的组织和操作方法二者而言，与另外的目的和优点一起，将从以下描述中更好地理解被认为是本发明特征的新特征。但是，应清楚地理解，每个附图仅出于说明和描述的目的而提供，并不旨在作为对本发明限制的限定。

[0043] 附图简述

[0044] 为了更完整地理解本公开内容，现在结合附图参考以下描述，其中：

[0045] 图1示出了根据一个实施方案的在不同样品类型中的一组变体基因及其变体频率的小组(panel)；

[0046] 图2示出了根据一个实施方案的用于从血浆无细胞脱氧核糖核酸(cfDNA)中检测单核苷酸变体(SNV)的方法的流程图；

[0047] 图3示出了根据一个实施方案的用于估计总体肿瘤cfDNA分数的特定概念；

[0048] 图4示出了根据一个实施方案的研究基因座的特定概念；

[0049] 图5示出了根据一个实施方案的提取机器学习模型特征的方法，所述机器学习模型可基于单独读取信息来区分真实变体和序列误差；

[0050] 图6示出了根据一个实施方案的用于机器学习模型的训练数据生成的方法，所述机器学习模型可基于单独读取信息来区分真实变体和序列误差；

[0051] 图7示出了根据一个实施方案的执行微小残留病变(MRD)的早期检测的时间线；

[0052] 图8示出了根据一个实施方案的用于执行微小残留病变的早期检测的系统的某些组件；

[0053] 图9示出了根据一个实施方案的使用所公开方法来检测微小残留病变的一些方面；并且

[0054] 图10示出了根据一个实施方案的用于执行所公开方法的系统的某些组件。

[0055] 发明详述

[0056] 本公开内容提供了用于在无细胞脱氧核糖核酸(cfDNA)样品和/或无细胞核糖核酸(cfRNA)样品中进行准确且灵敏的体细胞单核苷酸变体(SNV)检测的概率模型，所述样品包含序列数据集。cfDNA样品可取自血浆样品。该模型可从cfDNA和匹配的正常样品(例如从血液中的白细胞)接收测序数据作为输入。此外，由于可对于序列数据集中的每个基因座确定联合基因型，因此可去除种系突变。另外，由于可考虑总体肿瘤cfDNA分数，因此可从具有低肿瘤分数的样品中进行准确的SNV检测，所述样品用其他方法可能无法使用。

[0057] 如本文中所述的，来自对象(例如，癌症患者)的cfDNA可包含来自患病器官的细胞(例如，肿瘤细胞)的DNA或RNA与来自正常细胞的DNA或RNA的混合物。因此，cfDNA的基因型

可建模为正常肿瘤联合基因型。对于给定的基因座，表示为θ的肿瘤基因型的先验变体等位基因频率(先验VAF)可近似为突变读取的分数。(例如肿瘤cfDNA分数)。早期癌症患者中的肿瘤cfDNA分数通常可低于10％，因此为了准确估计肿瘤细胞中的体细胞SNV，可首先估计肿瘤cfDNA分数。可假设在单个样品中肿瘤cfDNA分数在整个基因组中是均匀的。

[0058] 本公开内容的方法和系统可使得观察到测序读取的似然P(X|θ)最大化，以估计肿瘤cfDNA分数或使用癌症中预定义的、经常观察到的变体的变体等位基因频率的平均值。由于可假设肿瘤cfDNA分数是均匀的，因此可利用来自热点的数据以估计肿瘤cfDNA分数。根据一个实施方案，可使用以下标准来选择热点：(1)主要等位基因可以是参考碱基；(2)匹配的正常样品中的基因座可以是纯合的，并且可与参考相同；(3)支持次要等位基因(minor allele)的读取的测序误差概率可小于观察到的等位基因频率；(4)可观察到具有变体等位基因的至少一个读取；以及(5)覆盖基因座的读取必须通过链偏倚过滤器。因此，可消除测序噪声。观察到读取或碱基的概率可用读取的映射质量(其为错位的概率)和测序碱基的碱基质量(其为测序误差的概率)来计算。为了确定基因座的基因型，可在给定不同联合基因型的情况下计算观察到覆盖基因座的读取的似然。

[0059] 根据一个实施方案，从序列数据集计算联合基因型的后验概率。具有最大后验概率的联合基因型可被视为给定基因座处的真实基因型。可将后过滤应用于检测到的体细胞变体，这可以排除具有链偏倚、低覆盖或串联重复的基因座。可将最终的肿瘤分数估计为通过过滤的体细胞变体的最大变体等位基因频率。

[0060] 输入数据可包含来自对象(例如患者)的cfDNA的原始测序读取。而且，实施方案可在分析之后提供体细胞SNV报告。当从cfDNA检测SNV时，实施方案可接受这样的小组：其中评价包含肿瘤来源的点突变的具有最高似然的基因座。例如，图1示出了小组100，其可通过分析公共癌症数据库或收集公开文献中的已知突变来获得。

[0061] 从cfDNA检测SNV的主要挑战可以是血流中肿瘤来源的低cfDNA分数，这使得难以将突变与测序误差区分开。无需特别关注肿瘤cfDNA分数，检测突变可能是不可靠的。本公开内容的方法和系统可通过将低肿瘤cfDNA分数并入到概率框架中来解决挑战，例如，明确考虑低肿瘤来源cfDNA分数。为此，可通过组合来自所有潜在SNV位点，并且尤其是已知引起突变的位点的信息来估计总体肿瘤cfDNA分数θ。以这种方式组合位点可消除由单个位点处的测序误差引起的噪声。然后可将估计的θ用于在单独位点处精确地调用SNV。在一些实施例中，可以以至少80％、90％、95％、96％、97％、98％、99％或更高的准确度来调用单独位点处的SNV。

[0062] 图2示出了根据所公开的实施方案的用于从血浆cfDNA检测SNV的方法200的一般流程图。根据方法200，原始测序数据201可包含血浆cfDNA样品和匹配的正常样品(例如来自同一对象的白细胞)，并且可输入到框202。原始测序数据201可包含输入文件，例如比对文件(.sam文件或.bam文件)和堆积文件(.pileup文件)。从原始测序文件201中提取沿着小组(例如图1中所示的小组100)的区域中的数据。

[0063] 原始测序数据可通过对血浆cfDNA样品或其衍生物(例如扩增产物或富集产物)进行测序而获得。测序方法可包括但不限于高通量测序、焦磷酸测序、合成测序(sequencing-by-synthesis)、单分子测序、纳米孔测序、基于半导体的测序、连接测序(sequencing-by-ligation)、杂交测序(sequencing-by-hVbridization)、RNA-Seq(例如，Illumina)、数字基

因表达(Helicos)、下一代测序(例如,Illumina,Pacific Biosciences of California,Ion Torrent)、单分子合成测序(SMSS)(例如,Helicos)、大规模并行测序(massively-parallel sequencing)、克隆单分子阵列(例如Solexa)、鸟枪法测序(shotgun sequencing)、Maxim-Gilbert测序和引物步移(primer walking)。

[0064] 原始测序数据可被预处理以提高映射质量。具有低质量映射的那些读取可基于参考基因组被重新校准和局部重新排列。还可通过设置硬阈值来消除具有低碱基质量的碱基,这可确保低的测序误差率。

[0065] 肿瘤cfDNA分数估计框202可使用预处理的测序数据来估计总体肿瘤cfDNA分数,所述预处理的测序数据包含从SAM或BAM文件和/或PILEUP文件提取的信息。即,在步骤202处,通过将基因座与测序数据中观察到的变体等位基因总体组合来进行肿瘤分数估计。总体肿瘤分数估计可有效抑制测序噪声,并且避免来自采样波动和测序误差(尤其是低或中深度测序)的严重影响。这不仅可改善准确的体细胞SNV检测,而且自身还可以是在癌症诊断和MRD检测中具有临床价值的标志物。

[0066] 步骤203可包括通过在给定总体肿瘤分数的情况下使联合肿瘤正常基因型的后验最大化来确定SNV。这可与对包含参考等位基因和变体等位基因的读取进行比较的其他方法形成对比。

[0067] 步骤204可包括使用就链偏倚、变体质量、序列背景和已知的SNP而言的一组cfDNA特征标准来过滤体细胞SNV候选。这样的调用后过滤确保了cfDNA中变体调用的高质量。

[0068] 步骤205可包括在过滤之后重新估计体细胞SNV。如关于第一步骤所述的,这可使用总体肿瘤分数来完成。在框205处,在检查cfDNA样品及其匹配的正常样品中,可特别考虑变体等位基因频率。例如,由于真实变体等位基因频率可与存在于血流中的肿瘤来源cfDNA的量密切相关,因此变体等位基因频率不能简单地视为对于杂合变体为0.5并且对于纯合变体为1.0。这种方法可代表相对于其他SNV检测方法的改进。

[0069] 传统的SNV方法可假设所有观察到的等位基因均来自二倍体染色体,其中第一染色体拷贝拥有基因型g_1,并且第二染色体拷贝拥有g_2。因此,基因座处的双等位基因型可由对(couplet)G=(g1,g2)表示,其中g_1和g_2的先验等位基因频率为50%和50%。这种抽象可假设被检查的肿瘤样品是纯的或接近纯的(例如,肿瘤基因型在样品中占主导地位)。然而,如本文中所讨论的,本公开内容的方法和系统可解释以下事实:cfDNA实际上可包含肿瘤来源的DNA和非肿瘤来源的DNA的混合物。因此,这样的方法可仅关注没有种系多态性的基因座,并将cfDNA的基因型建模为$G=(g_n,g_t)$,其中$gn=(gn_1,gn_2)$是来自正常cfDNA的基因型,并且$gt=(gt_1,gt_2)$是来自肿瘤来源的cfDNA的基因型。对于特定基因座,表示为θ的gt先验变体等位基因频率(先验VAF)可近似为突变读取的分数或肿瘤cfDNA分数。那么(1-θ)是gn的先验频率。但是,在给定的非常低比例的肿瘤来源cfDNA的情况下,在单一基因座处θ的这样的近似可能非常嘈杂。为了解决这样的挑战,可假设整个样品共享相同的先验等位基因频率,因此使用总体肿瘤cfDNA分数(例如总体θ)进行操作。

[0070] 图2中示出的方法图2可以是有利的,因为(1)在第一步骤中获得的总体肿瘤分数即使在低肿瘤分数下也可在单独基因座处产生肿瘤读取的预期水平,使得具有低突变体频率的体细胞SNV仍可与随机测序误差区分开;(2)在存在混合正常的和肿瘤来源的cfDNA的情况下,通过同时对正常的和肿瘤cfDNA进行基因分型而充分利用正常信号和肿瘤信号二

者，联合基因型模型可更好地拟合cfDNA测序数据；(3)通过机器学习模型进行的计算测序误差抑制可基于单独读取信息有效区分真实变体和测序误差；以及(4)可基于cfDNA特性特别地设计体细胞SNV候选的后过滤。鉴于这些有利特征，本公开内容的方法和系统不仅可应用于深靶向的cfDNA测序数据，而且可应用于中等深度的cfDNA测序数据，例如整个外显子组测序数据。因此，本公开内容的方法和系统可基于输入数据从cfDNA提供灵敏的突变调用和癌症的突变谱。

[0071] 参照图3的图解300。可使用所有可能热点的突变读取的总体分数来近似θ，所述所有可能热点可包含极可能具有肿瘤变体的基因座。即，图3示出了这样的概念：其中，为了估计总体肿瘤cfDNA分数，数据测序包含在所有热点处的测序读取，这被认为具有较低的测序噪声效应，例如测序误差。

[0072] 可在考虑可能的测序误差的情况下应用概率方法。例如，可将θ估计为使似然P(X|θ)最大化的值，即在给定总体肿瘤cfDNA分数θ的情况下，在K个热点处观察到测序数据X＝($X^{(1)}$,…,$X^{(K)}$)的概率。

[0073] 可通过将基因型G边缘化来计算似然，

[0074] $$P(X|\theta)=\prod_{r\in R_H}\sum_{G_{h(r)}}P(X_r|G_{h(r)},\theta)\,P(G_{h(r)})$$

[0075] 其中RH表示覆盖至少一个热点的读取的集合(pool)；X_r表示读取r中包含的信息；Gh(r)表示读取r覆盖的热点的基因型；并且h(r)是读取r覆盖的热点集(set)。然后，可通过进一步指定读取起点(例如，肿瘤细胞或正常细胞)来分解概率P(Xr|Gh(r)，θ)。此外，在总体肿瘤cfDNA分数为θ的情况下，根据以下公式，来自测序数据的随机读取可具有为肿瘤来源的θ的概率：

[0076] $$P(X_r|G_{h(r)},\theta)=\theta P\left(X_r|g_{T_{h(r)}}\right)+(1-\theta)P\left(X_r|g_{N_{h(r)}}\right),$$

[0077] 其中$g_{T_{h(r)}}$表示读取r上热点的联合肿瘤基因型；而$g_{N_{h(r)}}$表示读取r上热点h(r)的联合正常基因型。

[0078] 在一些情况下，RH中的不同读取可覆盖不同的热点(相互排斥)或具有共享的热点。考虑到最大似然估计框架，可将对于所有热点的基因型的联合组合边缘化以消除基因型的影响。因为该步骤可仅包含最可疑的基因座，所以单一读取可覆盖大量的热点，从而使这样的方法对于计算而言是合理的。

[0079] 覆盖多于一个热点的单一读取可被认为是独立的基因座，而不是考虑可使问题进一步复杂化的连锁不平衡。同样，读取中的信息也可扩展到碱基调用、碱基质量和映射质量。这可通过以下等式表示：

[0080] $$P\left(X_r|g_{T_{h(r)}}\right)=\prod_{k=1}^{K}P\left(X_r^{(k)}|\,g_T^{(k)}\right)=\prod_{k=1}^{K}P\left(B_r^{(k)},Q_r^{(k)},M_r|g_T^{(k)}\right)$$

[0081] 其中读取r包含K个热点；$h(r)=(g_T^{(1)},\ldots,g_T^{(K)})$；$X_r^{(k)}$表示读取r上覆盖热点k的碱基的信息；$B_r^{(k)}$、$Q_r^{(k)}$和$M_r$分别表示热点k处的碱基调用、碱基质量和读取r的映射质量，并且$g_T^{(k)}$表示热点k处的肿瘤基因型。对于正常情况，这可适用。这可通过以下等式表示：

[0082] $$P(X_r|g_{N_r}) = \prod_{k=1}^{K} P(B_r^{(k)}, Q_r^{(k)}, M_r|g_{N_r}^{(k)})$$

[0083] 其中$g_{N_r}^{(k)}$表示热点k处的正常基因型。

[0084] 根据前述内容，可计算在给定基因型的情况下在测序数据中观察到碱基的概率。可假设，对于给定的热点，观察到具有碱基A的读取，而给定的基因型为AB。然后碱基可由等位基因A以概率$\frac{1}{2}$P(正确测序)来确定，并且以概率$\frac{1}{2}\left(\frac{1}{3}P(\text{错误测序})\right)$来自等位基因B，其中值1/3表示所有三(3)个可能的错误碱基调用。但是，不是简单地使用1/3，而是可使用不均匀概率来说明测序误差的不同趋势。这可通过以下等式表示：

[0085] $$P\left(\mathbf{A}, Q_r^{(k)}, M_r|\mathbf{AB}\right) = \frac{1}{2}P\left(\mathbf{A}, Q_r^{(k)}, M_r|\mathbf{A}\right) + \frac{1}{2}P\left(\mathbf{A}, Q_r^{(k)}, M_r|\mathbf{B}\right) = \frac{1}{2}(1-\epsilon) + \frac{1}{2}\left(\frac{1}{3}\epsilon\right)$$

[0086] 其中∈表示碱基来自测序误差的概率。同样，P(A,$Q^{(k)}$,M_r|A)表示如果碱基来自对等位基因A进行测序，观察到具有质量$Q_r^{(k)}$,M_r的A碱基的概率，而P(A,$Q^{(k)}$,M_r|B)表示如果碱基来自对等位基因B进行测序，观察到具有质量$Q_r^{(k)}$,M_r的A碱基的概率。因此，上述等式可说明碱基A的两个可能来源。碱基质量和映射质量可指示在Phred标度中错误测序的碱基或错误放置的读取的概率，因此其可被转化回概率。这可通过以下等式表示：

[0087] $$\epsilon\left(Q_r^{(k)}, M_r\right) = 1 - \left(1 - 10^{-\frac{M_r}{10}}\right)\left(1 - 10^{-\frac{Q_r^k}{10}}\right)$$

[0088] 通过求解以上等式，可以以高准确度(例如，至少80％、90％、95％、96％、97％、98％、99％或更高的准确度)进行总体肿瘤cfDNA分数的估计。

[0089] 假设读取和热点彼此独立，则可通过量化测序误差和映射误差的概率来计算观察到这样的测序数据集的概率。或者，可使用简化的方法，在所述方法中将总体肿瘤cfDNA分数估计为基因座中变体等位基因频率的最大值。即使未优化肿瘤cfDNA分数估计的精度，只要所估计的肿瘤cfDNA分数标度正确，则SNV调用的准确度也可不显著降低。

[0090] 基因型计算框203可使用从肿瘤cfDNA分数估计框202接收的信息来计算正常基因型、肿瘤基因型和联合基因型的概率。用于计算基因型(或调用SNV)的框架可包括将基因座k处的基因型似然P($X^{(k)}$|$G^{(k)}$)计算为$P(X^{(k)}|G^{(k)}) = \prod_{r=1}^{n} P(X_r^{(k)}|G^{(k)})$，其中$X^{(k)}$表示观察到的覆盖基因组上特定位点的n个测序读取$X^{(k)} = \left(X_1^{(k)}, \cdots, X_n^{(k)}\right)$的集。

[0091] 该概念部分地由图4示出。参照图4，当鉴定体细胞SNV时，可仅考虑覆盖读取上所研究的基因座的碱基。如所示出的，对于每个测序读取$X^{(k)}$，可研究肿瘤基因型$g_{T_r}^{(k)}$，其中测序等位基因A由实线示出，并且测序等位基因B由虚线示出。

[0092] 根据一个实施方案，执行最大化后验以预测联合正常肿瘤基因型。在给定观察到的测序数据的情况下，具有最高后验概率的联合正常肿瘤基因型可被认为是样品的真实基因型。

[0093] 当分析cfDNA时，基因型似然可需要通过添加新参数θ来并入低肿瘤DNA频率。这可通过以下等式表示：

[0094] $$P(X^{(k)}|G^{(k)},\theta)=\prod_{r=1}^{n}P(X_r^{(k)}|G^{(k)},\theta)$$

[0095] 通常，可将覆盖基因座k的每个读取r的似然分解为各自的似然：$g_N^{(k)}$和$g_T^{(k)}$。这可通过以下等式表示：

[0096] $$P(X_r^{(k)}|G,\theta)=P(\text{读取}r\text{来自正常}\,\text{cfDNA})P(X_r^{(k)}|g_N^{(k)})+P(\text{读取}r\text{来自肿瘤}\,\text{cfDNA})P(X_r^{(k)}|g_T^{(k)})$$

[0097] 注意，先验概率P(读取r来自肿瘤来源的cfDNA)实际上可以是当前位点k处的$g_T^{(k)}$的先验等位基因频率θ。这种方法可产生：

[0098] $$P(X^{(k)}|G,\theta)=(1-\theta)P(X_r^{(k)}|g_N^{(k)})+\theta P(X_r^{(k)}|g_N^{(k)})$$。

[0099] 在前述等式中，$P(X_r^{(k)}|g_N^{(k)})$或$P(X_r^{(k)}|g_T^{(k)})$可与前述讨论中计算出的相同。

[0100] 然而，与估计总体肿瘤cfDNA分数相反，当进行基因分型(例如，调用SNV)时，仅基因分型的基因座处的碱基可在单一读取上相关。然后，使用贝叶斯定理，可估计使后验概率最大化的基因座k处的后验基因型$G^{(k)}$。这可通过以下等式表示：

[0101] $$P(G^{(k)}|X^{(k)},\theta)\propto P(X^{(k)}|G^{(k)},\theta)P(G^{(k)})$$

[0102] $$P(g_T^{(k)},g_N^{(k)}|X^{(k)},\theta)\propto P(g_T^{(k)},g_N^{(k)})\prod_{\text{读取}\,r}\{(1-\theta)P(X_r^{(k)}|g_N^{(k)})+\theta P(X_r^{(k)}|g_T^{(k)})\}$$

[0103] 可在计算之前确定联合正常肿瘤基因型G的先验分布。例如，如果在公共癌症突变数据库例如单核苷酸多态性数据库(dbSNP，https://www.ncbi.nlm.nih.gov/projects/SNP/)或癌症中的体细胞突变目录(COSMIC，https://cancer.sanger.ac.uk/cosmic)中注释了感兴趣的基因座，则可直接获取一般群体中的变体等位基因频率。或者，可探索公共数据来源中的测序数据，例如癌症基因组图谱数据库(TCGA，https://cancergenome.nih.gov/)、1000基因组数据库(http://www.internationalgenome.org/)和国际癌症基因组联合数据库(ICGC，https://icgc.org/)。通过考虑相对于一组样品的特定基因座，可确定群体部分或对象(例如，患者)的组的等位基因频率。

[0104] 为了从cfDNA中检测肿瘤来源的SNV，可消除种系多态性。这可通过考虑正常基因型和肿瘤基因型二者从而考虑种系多态性来进行。

[0105] 再次参照图2，在框201处，来自同一对象的正常血细胞(例如白细胞)的信息可用于将正常读取并入到分析框架中。即，该信息可用于进一步计算正常基因型、肿瘤基因型或联合基因型的似然。这可以是与匹配的肿瘤正常组织样品中的种系SNV去除相似的情况。后验概率可变化为：

[0106] $$P(X_w^{(k)},X_p^{(k)}|G^{(k)},\theta)=P(X_p^{(k)}|G^{(k)},\theta)P(X_w^{(k)}|g_N^{(k)})$$

[0107] 其中$X_w^{(k)}$是来自匹配的正常样品(例如白细胞)的数据，并且$X_p^{(k)}$表示从血浆测序的cfDNA读取。

[0108] 过滤框204可执行调用后过滤。可开发一组过滤器以去除cfDNA数据的不可靠突变候选。碱基质量和映射质量对于去除包含测序误差的低质量读取可以是重要的。概率模型

可并入这样的质量得分。除了可通过质量得分来区分的误差类型之外，还可存在不能通过仅使用质量得分来消除的系统误差。实施方案可组合突变候选附近的读取信息，并且可通过基因座处观察到的所有读取来过滤低质量突变候选。

[0109] 如果突变候选满足以下标准之一，则链偏倚过滤器可通过过滤突变候选来应用：(1)来自单链的读取的百分比大于阈值；(2)包含变体等位基因的读取从单链观察到；(3)在所有非参考读取中，包含变体等位基因的读取的比例高于给定阈值；以及(4)在链上包含非参考等位基因的读取的比例高于给定阈值。

[0110] 碱基质量过滤器可通过比较候选及其相邻基因座之间的碱基质量来应用。如果突变候选满足以下条件之一，则可过滤突变候选：(1)基因座及其邻近基因座处碱基质量的T统计量高于给定阈值；以及(2)基因座及其邻近基因座处的非参考碱基质量的T统计量高于给定阈值。

[0111] 如果读取伴侣在其重叠序列处不一致，则可通过在随后的过滤中排除读取伴侣来应用读取伴侣过滤器。

[0112] 序列背景也可影响测序误差率。因此，可通过鉴定接近区域或在均聚物内的突变候选来应用序列背景过滤器。

[0113] 可基于机器学习模型来应用测序误差过滤器。参照图2，在框206处，训练机器学习模型以对具有真实变体的读取和具有测序误差的读取进行分类。在以下过滤中可消除被分类为包含测序误差的读取。可训练分类器，以基于包含肿瘤来源的突变和测序误差的读取在多个方面(例如碱基质量、测序背景、PHRED得分等)的内在差异对所述读取进行分类。特别是，可提取读取中等位基因的不同特征，例如：读取比对相关的特征(例如，在SAM文件中以CIGAR串(string)编码的每个周围碱基的比对质量和比对信息)，围绕该等位基因的基因组背景(例如读取序列和插入/缺失)，读取中每个碱基的测序质量，以及双端测序数据的插入大小。

[0114] 图5示出了提取机器学习模型的特征的方法。参照图5，该特征谱被输入到分类器以区分信号和误差。由于cfDNA可自然地片段化，其中长度分布集中在约166个碱基对(bp)附近，因此双端读取伴侣在映射至参考之后可具有重叠的碱基。重叠的碱基可从同一cfDNA片段的正向链和反向链二者进行测序，因此这样的碱基可提供额外的信息或证实它们位于变体基因座处。因此，从读取伴侣中提取的特征可组合在一起。为了评价此特征谱的有效性，可使用常规的分类方法(例如随机森林)以评价所提取的特征的区分性如何。结果可说明，这样的特征谱实现良好的性能。

[0115] 除了传统的分类器，不同的分类器，这样的更复杂的分类器(例如深度学习分类器)可用于进一步提高真实突变和测序误差的鉴别性能。深度学习模型可包括深度人工神经网络的任何下降结构，例如基于卷积神经网络(Convolutional Neural Network，CNN)的分类器。可使用基于CNN的分类器，其包含(1)卷积层，(2)可以以不同标度提取信息性和抽象的特征的子采样层，以及(3)可使用提取的特征用于真正突变和测序误差的最终预测的输出预测层。这些层可堆叠形成深度神经网络，并且可通过基于随机梯度的优化算法来学习它们的参数。例如，可将这样的基于CNN的结构应用于成像分类应用。

[0116] 为了训练机器学习模型，可需要高质量的基准真实训练数据，其包含已被验证具有序列误差或真实变体的读取。可通过使用来自具有足够覆盖的同一样品的不同测序数据

来获得这样的基准真实数据，例如，整个外显子组测序数据和高覆盖无PCR的整个基因组测序数据获自1000基因组项目中的NA12878和转移性乳腺癌患者(MBC315)。整个外显子组测序的目标区域中的变体可分别从两个样品中调用。包含在两个样品中均以高可信度被鉴定的变体的那些读取可被视为具有真实变体的读取的训练数据，而包含在一个样品中以低可信度被鉴定的变体的那些读取可被视为具有测序误差的读取的训练数据。

[0117] 图6示出了使用NA12878收集训练数据的一些方面。即，图6示出了用于机器学习模型的训练数据生成的方法。

[0118] 已知的SNP过滤器可通过过滤在公共数据库(例如dbSNP，外显子组聚集联合数据库(ExAC，http://exac.broadinstitute.org/)和基因组聚集数据库(gnomAD，http://gnomad.broadinstitute.org/))中观察到的突变候选来应用。可保留在COSMIC中观察到的候选。

[0119] 在实体瘤组织(在手术或活检中切除)和正常血液样品之间调用体细胞SNV之后，可使用从上述步骤：框202、203和204中鉴定的体细胞SNV进一步完善总体肿瘤分数的估计(图2的框205)。特别地，可从调用的体细胞突变中选择截短的突变以进行监测，其中计算了调用的体细胞突变的变体等位基因频率。在该步骤中可不需要实体瘤组织。可组合许多截短突变，以确保小分数肿瘤来源cfDNA的灵敏检测。

[0120] 检测MRD的应用：基于cfDNA的监测方法

[0121] 如框207处所示的，该方法可具有广泛的应用，例如癌症检测、监测、预后，以及更具体地血液肿瘤突变负荷(blood tumor mutation burden，bTMB)计算、MRD检测和抗性监测。例如，MRD检测是可证明本公开内容的方法和系统的临床用途的应用。

[0122] 图7示出了进行微小残留病变(MRD)的早期检测的时间线。如图7中所示，使用本公开内容的方法和系统用于MRD检测的时间线可包括两个步骤。在框701处，可鉴定来自手术前血浆cfDNA样品(“基线”样品)的截短突变和突变谱。由于血浆cfDNA包含源自多个肿瘤部位的DNA片段，因此可使用对象手术前血浆cfDNA样品中计算出的全面肿瘤突变谱(SNV及其变体等位基因频率)来鉴定“截短突变”，其通常指在所有检测到的变体中具有高变体等位基因频率的那些变体。

[0123] 特别地，可将cfDNA SNV检测方法应用于手术前血浆cfDNA样品及其匹配的正常样品(例如，来自同一对象的白细胞)以鉴定所有体细胞SNV及其变体频率。然后可基于来自cfDNA和匹配的正常样品(例如，来自白细胞)的突变调用结果选择截短突变，并且可以计算其变体等位基因频率。

[0124] 或者，从实体瘤组织(在手术或活检中切除)及其匹配的正常样品中调用的体细胞SNV可用于确定一组截短突变。可使用许多不同的突变调用工具(例如VarScan2、MuTect、SomaticSniper等)来调用体细胞SNV。然而，不能总是从取自单一肿瘤部位的实体肿瘤组织中确定截短突变。因此，可不需要实体瘤组织以使本公开内容的方法和系统有效。此外，可仅基于来自血液样品(例如血浆cfDNA和白细胞DNA)的突变调用结果来确定截短突变。

[0125] 在框702处，使用随访血浆cfDNA样品检测微小残留病变(MRD)。即，在手术之后，随访血浆样品可用于监测和检测MRD。因为已经治疗或切除了肿瘤，所以随访血浆中的肿瘤分数可低于基线血浆中的肿瘤分数。变体等位基因频率的标度可与测序误差率相同，或甚至更低。因此，MRD检测可需要灵敏且可靠地检测包含真实突变的读取。例如，使用机器学习方

法将真实突变与测序误差区分开。(如由图2的框206所描述的)。特别地,这种机器学习方法可应用于提取覆盖已鉴定的截短突变的位置并且被分类为携带真实变体的那些读取。然后,仅被分类为包含真实变体的那些读取可用于计算MRD预测得分(或称为MRD指标得分)。该得分可被定义为覆盖截短突变的读取中包含截短突变的读取的比例。

[0126] 图8更详细地说明了前述内容。即,图8示出了用于执行微小残留病变(MRD)的早期检测的系统的某些组件。

[0127] 为了评估MRD预测得分在确定肿瘤存在中的统计显著性,可使用样品内背景分布产生方法来对正常背景进行建模,而不引入额外的实验偏差并且不需要额外的样品组。可假设已鉴定出k个截短突变,并且可在基因组中对k个位点随机采样,所述位点不包含已鉴定的突变但匹配那些k个截短突变的特征(例如,突变数目、核苷酸分布和覆盖的深度)。可提取覆盖随机采样的k个位点的读取,并随后将其输入到测序误差抑制工作流中以过滤出测序误差。可为k个采样的位点生成MRD预测得分。可重复采样N次(例如,其中N为至少约5、约10、约25、约50、约100或大于约100),并产生N个MRD得分。这些N个MRD预测得分可包含背景分布以评价k个截短突变的MRD预测得分的统计显著性。可随时间重复该步骤以检测微小残留病变(MRD)。

[0128] 核酸测序

[0129] 可使用多种核酸测序方法对本公开内容的样品进行测序。这样的样品可在测序之前进行处理,例如通过进行纯化、分离、富集、核酸扩增(例如,聚合酶链反应(PCR))。测序可使用例如以下来进行:Sanger测序、高通量测序、焦磷酸测序、合成测序、单分子测序、纳米孔测序、半导体测序、连接测序、杂交测序、RNA-Seq(Illumina)、数字基因表达(Helicos)、下一代测序(例如,Illumina,Pacific Biosciences of California,Ion Torrent)、单分子合成测序(SMSS)(Helicos)、大规模并行测序、克隆单分子阵列(Solexa)、鸟枪法测序、Maxim-Gilbert测序、引物步移、使用PacBio,SOLiD,Ion Torrent或Nanopore平台进行的测序以及本领域中已知的任何其他测序方法。可使用多重测序进行同时测序反应。

[0130] 测序可产生可由计算机处理的测序读取(“读取”)。在一些实施例中,可针对一个或更多个参考来处理读取以鉴定单核苷酸变体(SNV)。

[0131] 在一些实施例中,可在可包含多种不同类型核酸的无细胞多核苷酸上进行测序。核酸可以是多核苷酸或寡核苷酸。核酸包括但不限于脱氧核糖核酸(DNA)或核糖核酸(RNA)、单链或双链DNA、互补DNA(cDNA)或RNA/cDNA对。

[0132] 结果

[0133] 在相同的测试条件下,将本公开内容的方法和系统在检测体细胞SNV中的性能与四种所选择的方法(SiNVICT、VarScan 2、MuTect和MuTect 2)的结果进行比较。

[0134] 由Wgsim生成的数据的模拟结果

[0135] 第一模拟数据集中的测序读取是由wgsim生成的。将包含参考序列的读取和包含变体等位基因的读取以预定义比例(0.1%、0.5%、1%和5%)混合。低变体等位基因频率可模拟cfDNA测序数据中遇到的条件。

[0136] 在第一模拟中,在模拟的数据上,针对VarScan2和SiNVICT比较使用本公开内容的方法和系统获得的结果,所述模拟的数据是通过将wgsim应用于人参考基因组(hg19)获得的。特别地,提取具有非小细胞肺癌(non-small cell lung cancer,NSCLC)的已知致病性

突变和模拟的超深度测序的区域(chr19:10602252-10602938)。十三种体细胞变体被添加至该区域。

[0137] 在四个不同的肿瘤来源cfDNA分数(5%、1%、0.5%、0.1%)中获得的平均读取深度为约1448。为了更好地接近真实读取,将测序误差和插失引入到读取中。如表1所示的,对于具有低肿瘤cfDNA分数的样品,本公开内容的方法和系统比其他两种方法表现更好。即,模拟结果表明,与其他方法形成对比,本公开内容的方法和系统能够从具有非常低的ctDNA分数的血浆DNA中有意义地检测SNV。

[0138] 表1

[0139]

肿瘤cfDNA分数	0.10%	0.50%	1%	5%
所描述的实施方案	0%	46%	62%	100%
VarScan2(p＜0.05)	0%	15%	62%	85%
SiNVICT	0%	0%	46%	85%

[0140] 表1说明了三种方法对具有不同肿瘤cfDNA分数的模拟测序样品的精度。注意,表1中未示出召回率,因为所有方法均证明了模拟读取的完美召回率。

[0141] 由转移性癌症患者产生的数据的模拟结果

[0142] 第二模拟数据集包含来自来自同一对象的血沉棕黄层样品和转移性肿瘤样品的整个外显子组测序数据的加标(spike-in)数据集。设置样品混合比例以确保对象基因组中体细胞SNV的变体等位基因频率约为20%、10%、5%和1%。

[0143] 在第二模拟中,在来自转移性癌症患者的真实的整个外显子组测序数据上测试了本公开内容的方法和系统。真实样品可包含比模拟的读取的更高的复杂性。提取被报道在cfDNA中具有体细胞突变的7个基因中的外显子区域,其中7个中的6个是在肝转移样品中鉴定的,并且剩余的1个仅在原发性乳腺癌中检测到。

[0144] 为了获得具有不同肿瘤来源cfDNA分数的样品,将血沉棕黄层样品与具有不同加标比例的肝转移样品混合。假设所有体细胞变体均是杂合的,则第一加标样品为六个不同的肿瘤来源cfDNA分数(约20%、10%、5%和1%)中的。突变基因座处的读取深度为46至195。因此,仅在原发性乳腺癌样品中出现的体细胞变体被认为是不应在加标样品中被鉴定出的对照。

[0145] 如表2中所示,在具有10%肿瘤cfDNA分数的样品中,本公开内容的方法和系统未检测到具有1%变体等位基因频率的体细胞变体。在真实的测序样品中,尽管总的肿瘤cfDNA分数可高达10%,但对于给定的基因座,变体等位基因频率可能较低。当肿瘤cfDNA分数较低时,加标样品在大多数体细胞变体基因座处均不包含支持变体等位基因的任何读取。

[0146] SiNVICT在模拟的测序数据上表现适中,但是在真实测序数据上表现不佳。其将所有体细胞变体鉴定为种系变体。

[0147] 表2

[0148]

	肿瘤 cfDNA 分数	本公开内容的方法	VarScan2	SiNVICT	MuTect	MuTect2
精度	20%	100%	83%	0%	100%	100%
召回率	20%	100%	100%	0%	100%	100%
精度	10%	83%	67%	0%	17%	67%
召回率	10%	100%	100%	0%	100%	100%
精度	5%	67%	33%	0%	0%	33%
召回率	5%	100%	0%	0%	0%	100%
精度	1%	17%	0%	0%	0%	0%
召回率	1%	9%	0%	0%	0%	0%

[0149] 表2显示了在具有不同肿瘤cfDNA分数的真实测序数据上三种方法的精度和召回率。读取深度为46至195。

[0150] 被设计用于从cfDNA中检测体细胞SNV的SiNVICT的计算方法在大部分早期癌症检测中(例如,当肿瘤cfDNA分数相对较低时)均失败了。结合了实验设计和计算方法二者的另一种方法CAPP,即使在基因分型的基因座处有10k的覆盖,在早期癌症检测中也至少失败了一半。

[0151] 相反,本公开内容的方法和系统以7至70倍更少的覆盖实现了可比较的体细胞SNV检测。此外,可不需要特殊的测序设计或目标小组来进行分析。因此,本公开内容的方法和系统可应用于一般的癌症检测或筛选,而不是单一癌症类型的特异性检测。这是因为可接受整个外显子组测序数据作为输入。如果指定特定的癌症类型用于使用本公开内容的方法和系统进行检查,则可将基因分型小组设计用于体细胞SNV检测。

[0152] 从被诊断为患有已经扩散到淋巴结的ER+HER2+乳腺癌的41岁女性中获得样品。在进行肝活检之前约30分钟收集血液以产生转移样品。根据测序结果,在cfDNA中获得的平均变体等位基因百分比为14%,表明约28%的cfDNA是肿瘤来源的。血液样品是在癌症晚期获得的,因此肿瘤DNA的百分比相对来说不低。

[0153] 表3和表4中示出了所分析样品的测序统计。

[0154] 表3

[0155]

组织	输入 DNA（ng）	读取（百万）	映射的读取（%）	成对的读取（%）
血沉棕黄层	412	182	181(99.6)	180(98.8)
原发性肿瘤	301	112	110.8(99.2)	99.1(88.8)
转移	341	173	171.8(99.5)	170(98.6)
cfDNA	155	286	284.8(99.5)	253(88.5)

[0156] 表4

[0157]

组织	靶向的映射的读取（%）	PCR 复本	平均测序深度
血沉棕黄层	154(84.8)	0.2	201
原发性肿瘤	92(82.8)	0.52	118
转移	140(81.6)	0.22	183
cfDNA	239(83.8)	37%	309

[0158] 如本公开内容中所述的，cfDNA为癌症诊断、预后和精确肿瘤学提供了独特的机会。但是，目前用于分析突变(特别是血浆中ctDNA)的计算工具可能不是最佳准确的。在给定测序数据集的情况下，本公开内容的方法和系统可使用基于贝叶斯的概率框架来估计基因型的似然。使后验估算最大化可用于确定对象(例如患者)的基因型并调用体细胞SNV。可考虑血浆样品中的测序误差和低的肿瘤变体等位基因频率二者，从而有助于体细胞SNV的检测，并确保良好的性能。

[0159] 如从模拟结果(例如，使用加标真实测序数据或模拟的测序数据)所看出的，本公开内容的方法和系统可胜过其他方法，因为它们可解释cfDNA的特殊生物学特性。由于本公开内容的方法和系统可从低肿瘤cfDNA分数血浆样品中检测体细胞SNV，因此它们可应用于早期癌症诊断以及液体活检和癌症治疗后监测。在检测到某些驱动突变的情况下，临床医生可确定癌症类型，并为患者提供个性化治疗。

[0160] 通过混合两个种系样品在数据上生成的模拟结果

[0161] 为了模拟血浆中ctDNA的不同稀释，将从公共数据库(例如，正常(NA24385)和肿瘤(NA12878))中获得的两个高质量种系样品进行混合。体细胞突变可包含679个高质量突变，其在NA12878(作为肿瘤)中由VarScan鉴定但在NA24385(作为正常的)中没有。

[0162] 根据模拟，可产生20个样品进行稀释(例如0.001％、0.002％、0.005％、0.008％、0.01％、0.02％、0.05％、0.08％、0.1％、0.2％、0.5％、0.8％和1％)。例如，如果测试的p值不大于约0.05，则可将样品鉴定为指示抗性。

[0163] 图9示出了使用本公开内容的方法和系统的微小残留病变(MRD)检测的性能900。如由图9所示出的，在具有低至约0.008％的肿瘤分数(tumor fraction，TF)的样品中可完美地检测MRD。因此，本公开内容的方法和系统可应用于MRD的早期检测。

[0164] 图10示出了根据本公开内容的方法适用于从无细胞核酸例如脱氧核糖核酸(cfDNA)和核糖核酸(cfRNA)中检测体细胞单核苷酸变体(SNV)的示例性系统。电子装置1010可包含装置的多种配置。例如，电子装置1010可包含计算机、膝上计算机(laptop computer)、平板装置(tablet computer)、服务器、专用空间处理组件或装置、智能手机、个人数字助理(personal digital assistant，PDA)、物联网(IoT或IOTA)装置、网络设备(例如，路由器、接入点、毫微微小区(femtocell)、微微蜂窝(Pico cell)等)等。

[0165] 电子装置1010可包含可操作以促进根据本公开内容方法的电子装置1010的功能的任何数目的组件，例如所示出的实施方案的处理器1011、系统总线1012、存储器1013、输入接口1014、输出接口1015和编码器1016。处理器1011可包含一个或更多个处理单元，例如中央处理单元(central processing unit，CPU)(例如，来自Intel CORE多处理器单元系列的处理器)、现场可编程门阵列(field programmable gate array，FPGA)和/或专用集成电

路(application specific integrated circuit,ASIC),其在一个或更多个定义逻辑模块的指令集的控制下是可操作的,所述逻辑模块被配置为提供本文中所述的操作。系统总线1012将多种系统组件例如存储器1013、输入接口1014、输出接口1015和/或编码器1016偶连至处理器1011。因此,一些实施方案的系统总线1012可以是多种类型的总线结构中的任一种,例如使用多种总线结构中的任一种的存储器总线或存储器控制器、外围总线和/或本地总线。作为补充或替代,可利用另一些接口和总线结构,例如并行端口、游戏端口或通用串行总线(universal serial bus,USB)。存储器1013可包含易失性和/或非易失性计算机可读存储介质的多种配置,例如RAM、ROM、EPSOM、闪速存储器或其他存储技术、CD-ROM、数字通用盘(digital versatile disk,DVD)或其他光盘存储、盒式磁带、磁带、磁盘存储或其他磁性存储装置,或可用于存储所需信息的其他有形和/或非暂时性介质。输入接口1014有助于将一个或更多个输入组件或装置偶连至处理器1011。

[0166] 例如,使用者可通过偶连至输入接口1014的一个或更多个输入装置(例如,小键盘、传声器、数字指向装置、触摸屏等)将命令和信息输入到电子装置1010中。图像捕获装置例如照相机、扫描仪、3-D成像装置等可偶连至实施方案的输入接口1014,例如以在本文中提供源视频。输出接口1015有助于将一个或更多个输出组件或装置偶连至处理器1011。例如,可通过偶连至输出接口1015的一个或更多个输出装置(例如,显示监视器、触摸屏、打印机、扬声器等)向使用者提供来自电子装置1010的数据、图像、视频、声音等的输出。实施方案的输出接口1015可提供至其他电子组件、装置和/或系统(例如,存储器、视频解码器、无线电发射机、网络接口卡、装置例如计算机、膝上计算机、平板装置、服务器、专用空间处理组件或装置、智能手机、PDA、IOTA装置、网络设备、机顶盒、电缆头端系统、智能TV等)的接口。

[0167] 尽管已经详细描述了本发明及其优点,但是应理解,在不脱离由所附权利要求书限定的本发明的精神和范围的情况下,可在本文中进行多种变化、替换和改变。此外,本申请的范围不旨在限于说明书中描述的过程、机器、制造、物质组成、手段、方法和步骤的一些具体实施方案。如本领域的普通技术人员将从本发明的公开内容中容易地理解的,根据本发明可利用目前存在或将要被开发的执行与本文中所述的相应实施方案基本上相同的功能或实现与其基本上相同的结果的过程、机器、制造、物质组成、手段、方法或步骤。因此,所附权利要求书旨在将这样的过程、机器、制造、物质组成、手段、方法或步骤包含在其范围内。

[0168] 此外,本申请的范围不旨在限于说明书中描述的过程、机器、制造、物质组成、手段、方法和步骤的一些具体实施方案。

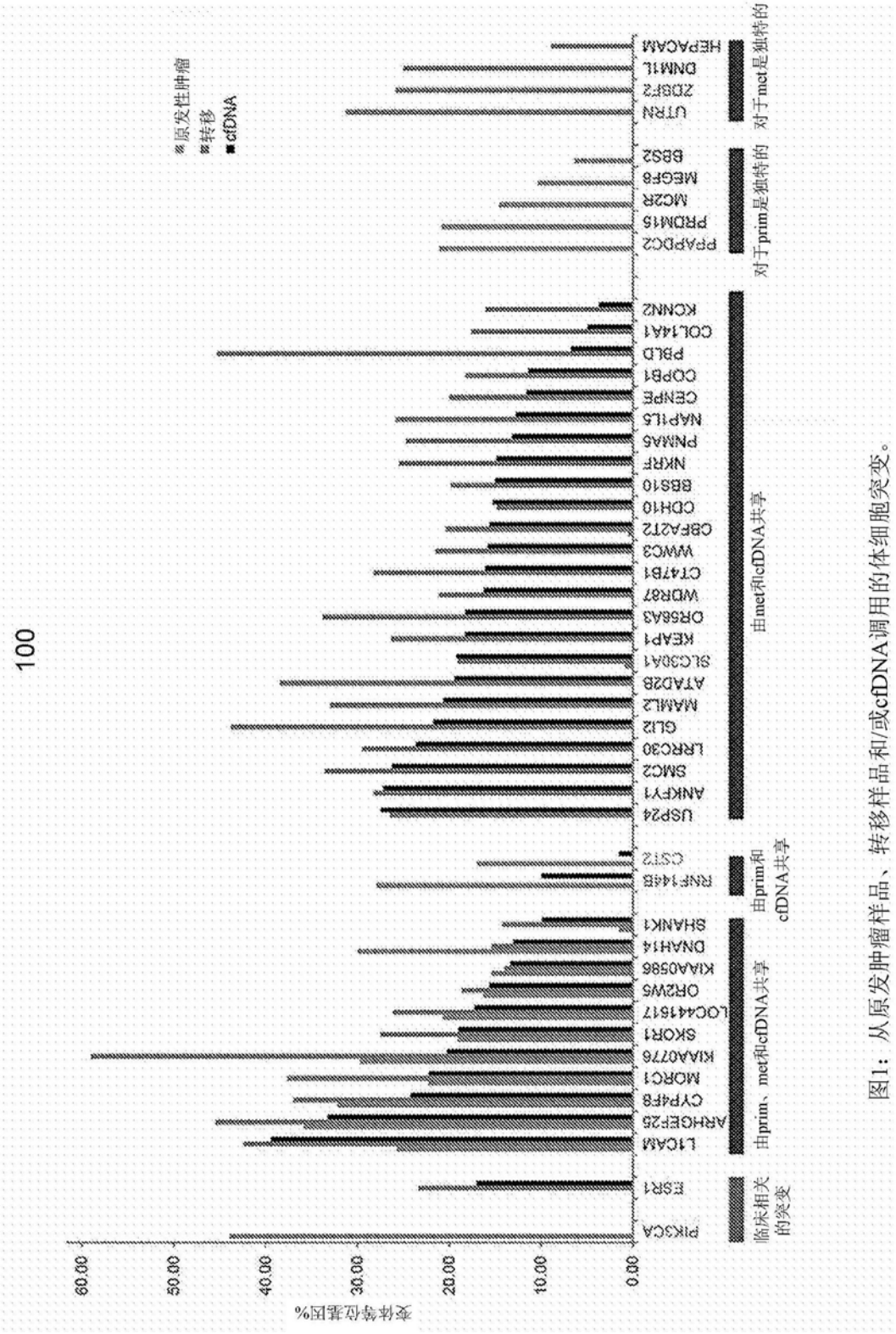


图1：从原发肿瘤样品、转移样品和/或cfDNA调用的体细胞突变。

图1

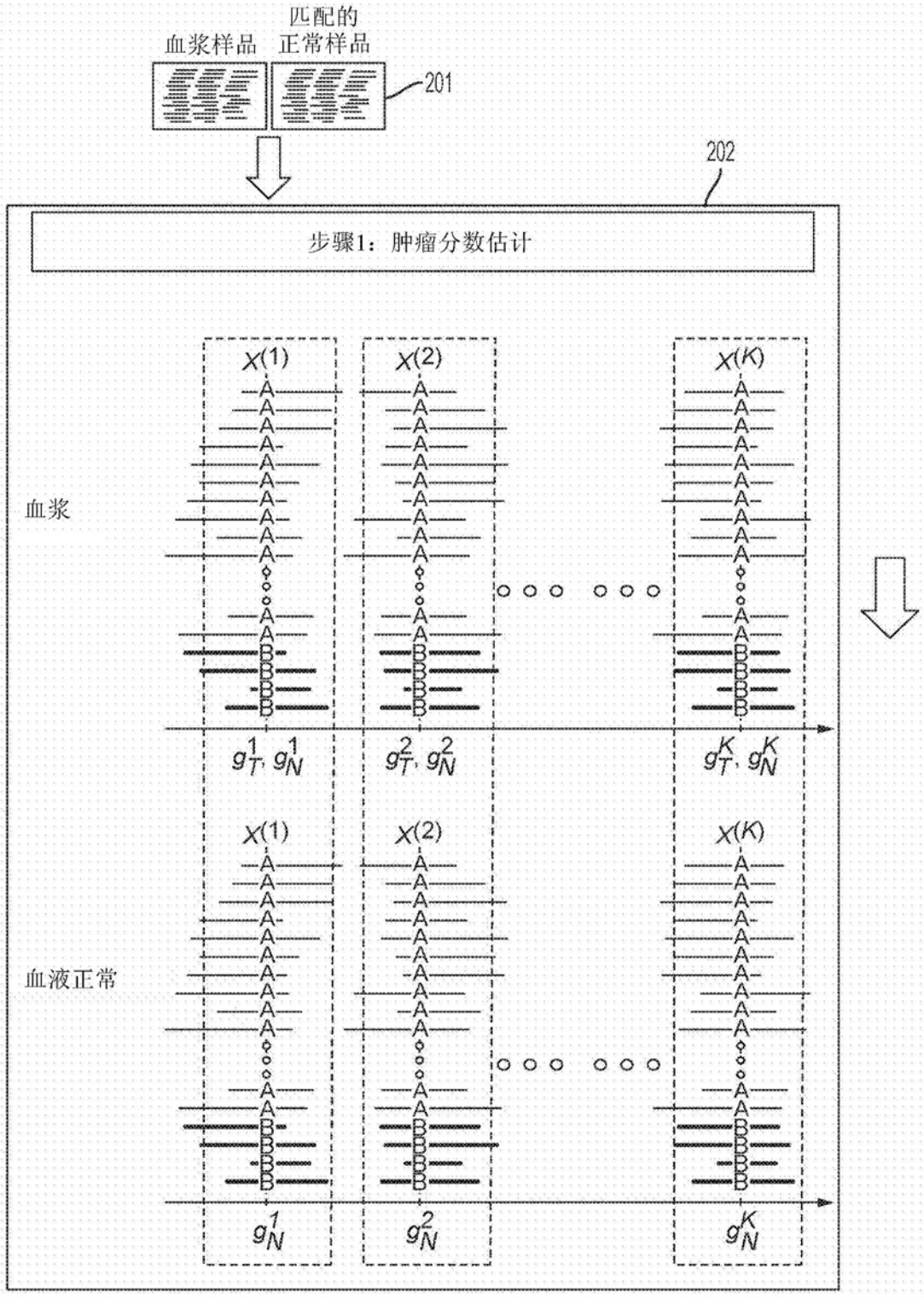
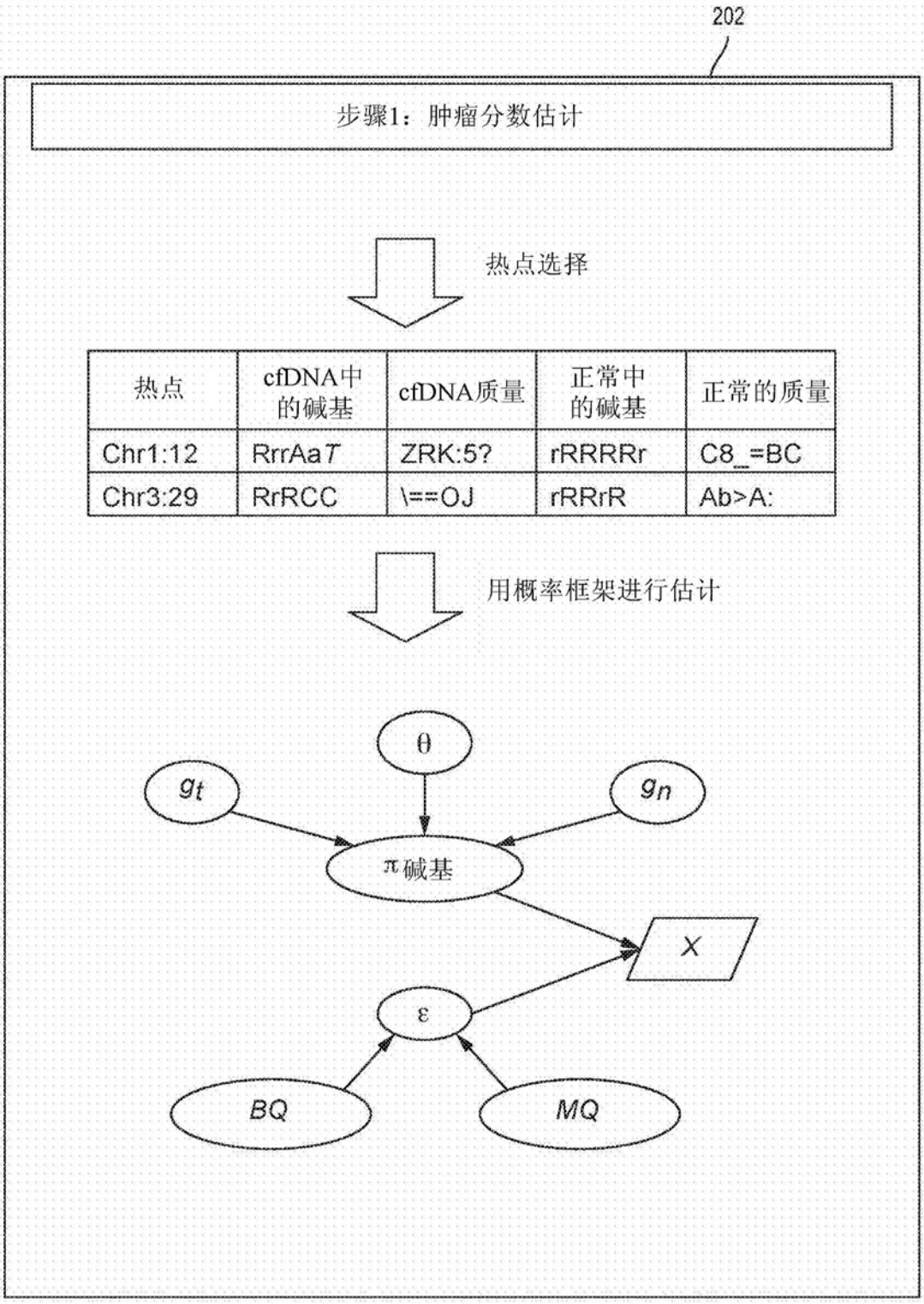


图2



热点	cfDNA中的碱基	cfDNA质量	正常中的碱基	正常的质量
Chr1:12	RrrAa*T*	ZRK:5?	rRRRRr	C8_=BC
Chr3:29	RrRCC	\==OJ	rRRrR	Ab>A:

图2(续)

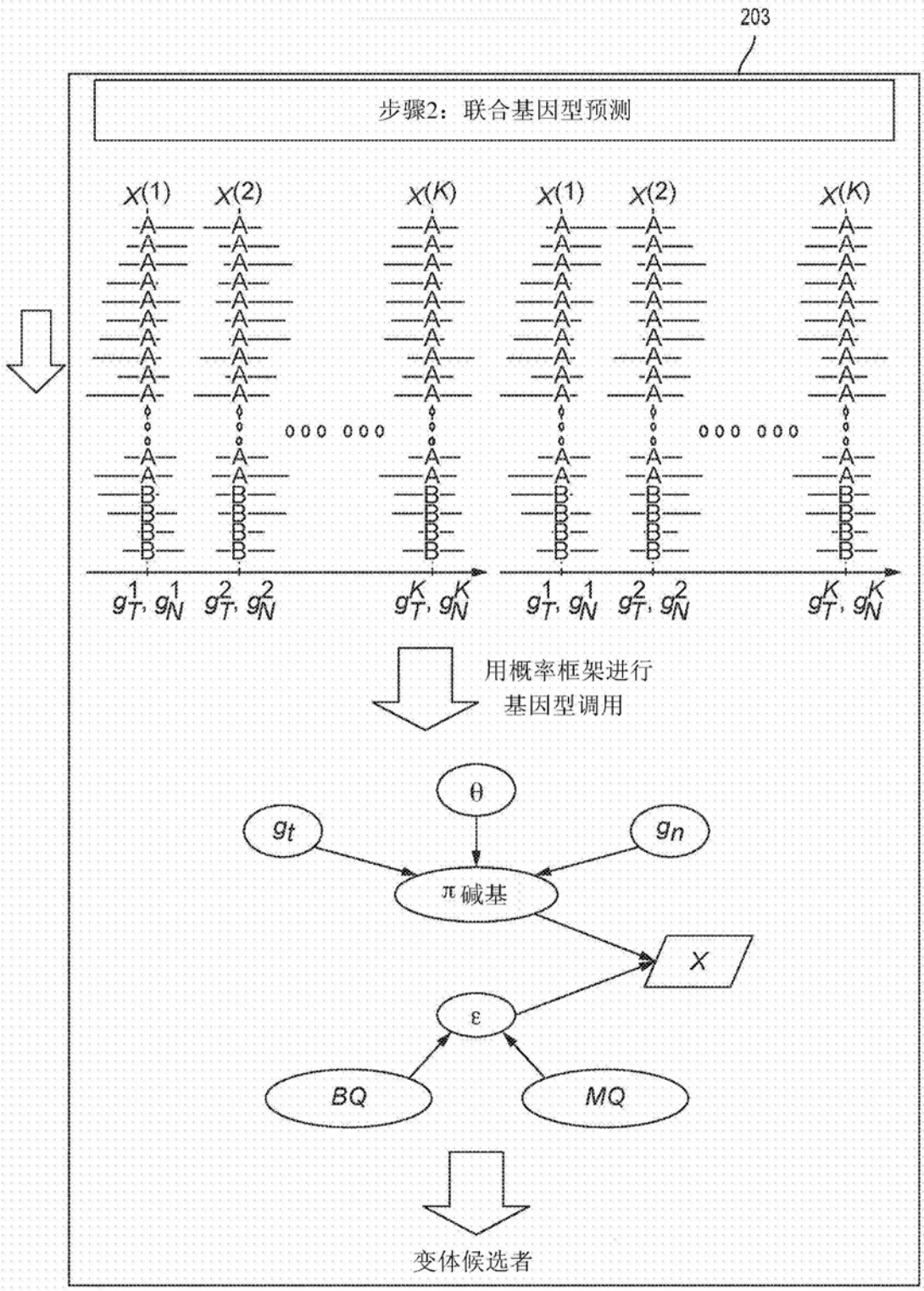

203
步骤2：联合基因型预测
X(1)
X(2)
X(K)
X(1)
X(2)
X(K)
g_T^1, g_N^1
g_T^2, g_N^2
g_T^K, g_N^K
g_T^1, g_N^1
g_T^2, g_N^2
g_T^K, g_N^K
用概率框架进行
基因型调用
θ
g_t
g_n
π碱基
X
ε
BQ
MQ
变体候选者

图2(续)

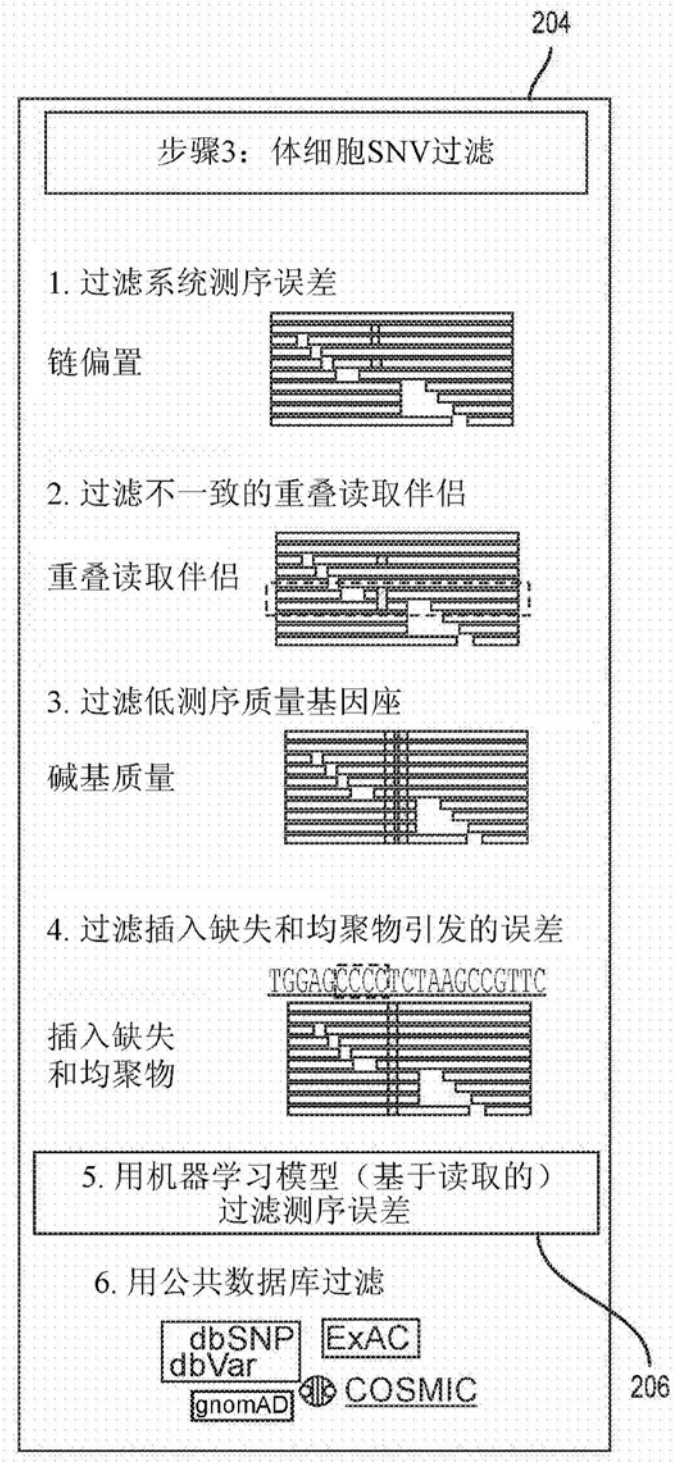

204
步骤3：体细胞SNV过滤
1. 过滤系统测序误差
链偏置
2. 过滤不一致的重叠读取伴侣
重叠读取伴侣
3. 过滤低测序质量基因座
碱基质量
4. 过滤插入缺失和均聚物引发的误差
TGGAGCCCCTCTAAGCCGTTC
插入缺失
和均聚物
5. 用机器学习模型（基于读取的）
过滤测序误差
6. 用公共数据库过滤
dbSNP
dbVar
ExAC
gnomAD
COSMIC
206

图2(续)

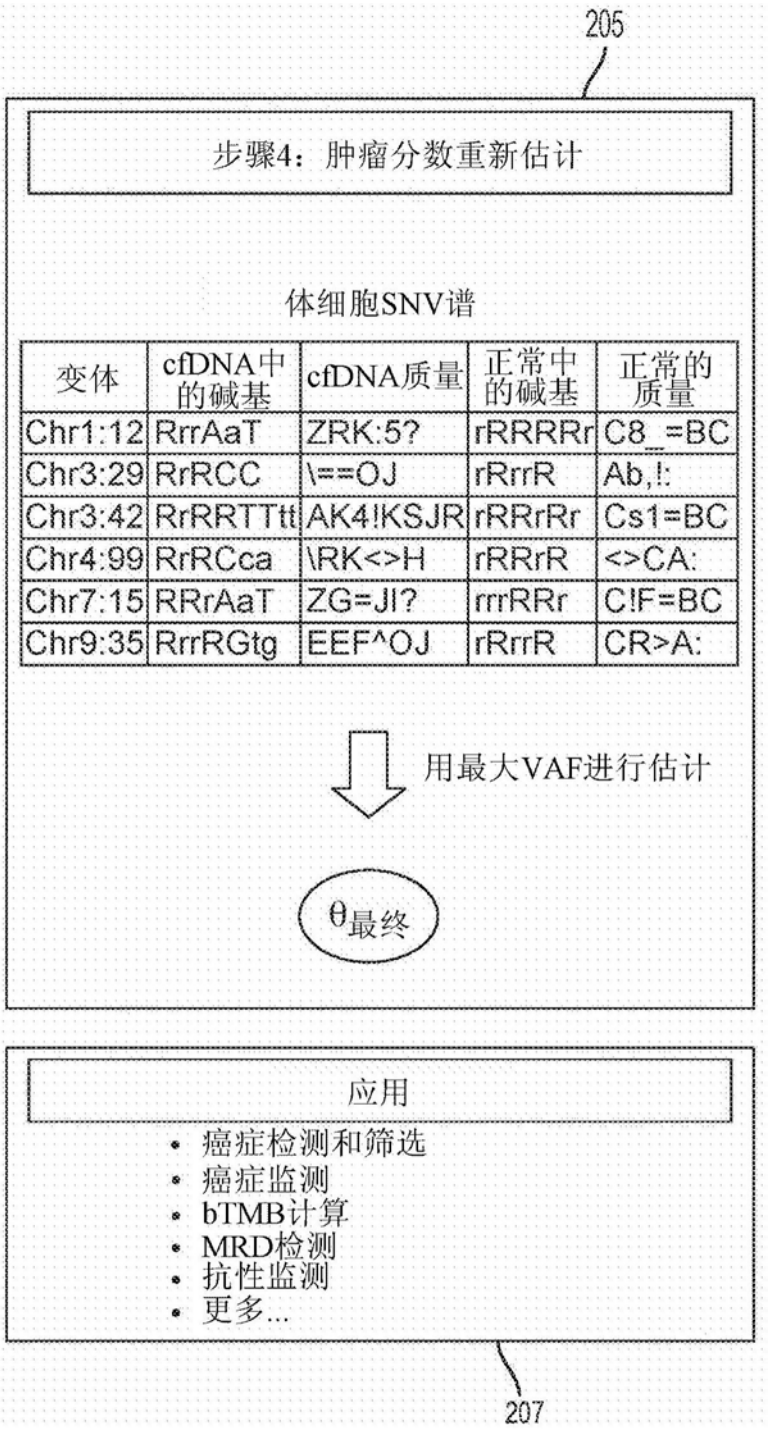

205
步骤4：肿瘤分数重新估计
体细胞SNV谱
变体 | cfDNA中的碱基 | cfDNA质量 | 正常中的碱基 | 正常的质量
Chr1:12 | RrrAaT | ZRK:5? | rRRRRr | C8_=BC
Chr3:29 | RrRCC | \==OJ | rRrrR | Ab,!:
Chr3:42 | RrRRTTtt | AK4!KSJR | rRRrRr | Cs1=BC
Chr4:99 | RrRCca | \RK<>H | rRRrR | <>CA:
Chr7:15 | RRrAaT | ZG=JI? | rrrRRr | C!F=BC
Chr9:35 | RrrRGtg | EEF^OJ | rRrrR | CR>A:
用最大VAF进行估计
θ最终
应用
癌症检测和筛选
癌症监测
bTMB计算
MRD检测
抗性监测
更多...
207

图2(续)

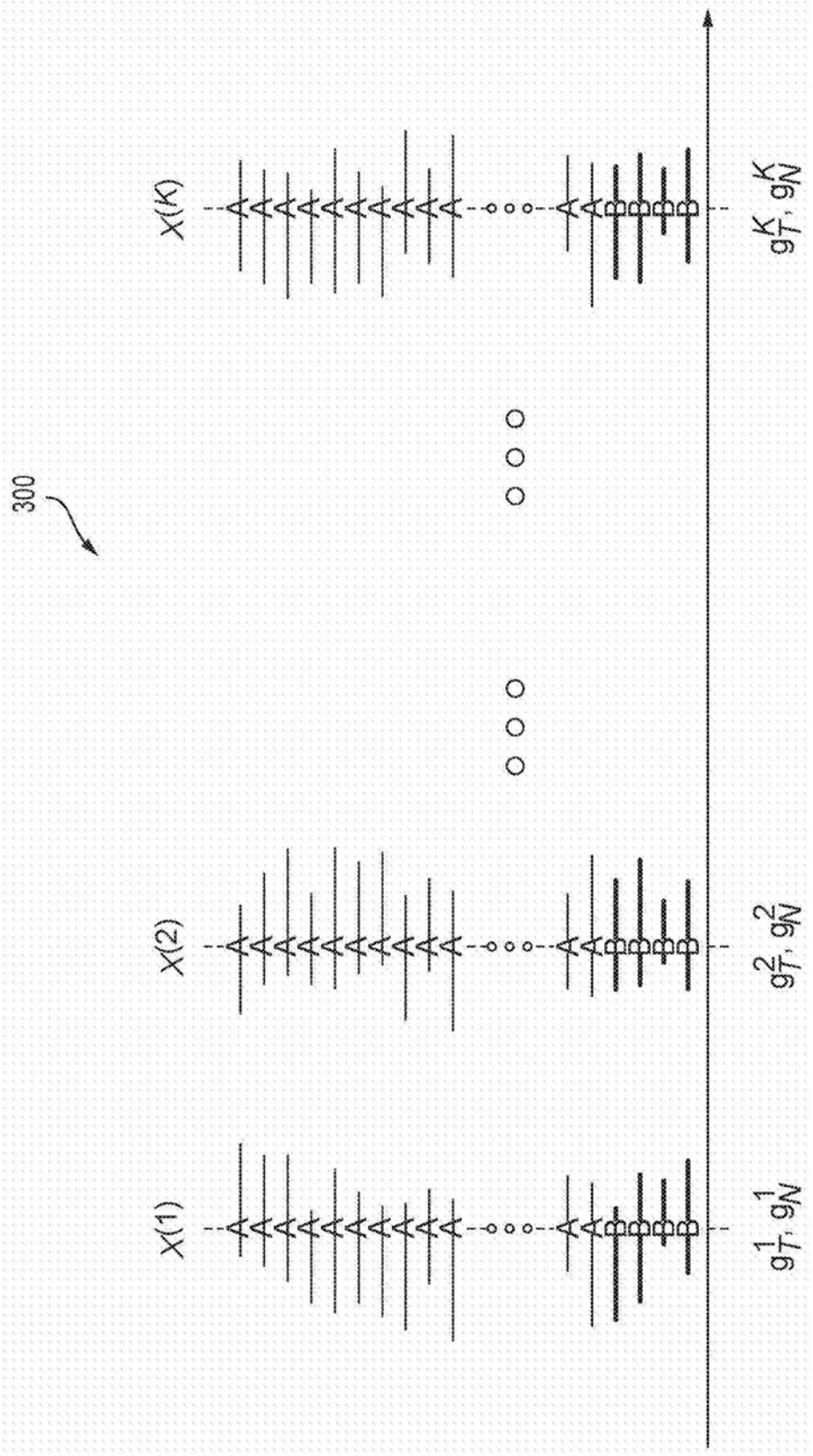


图3

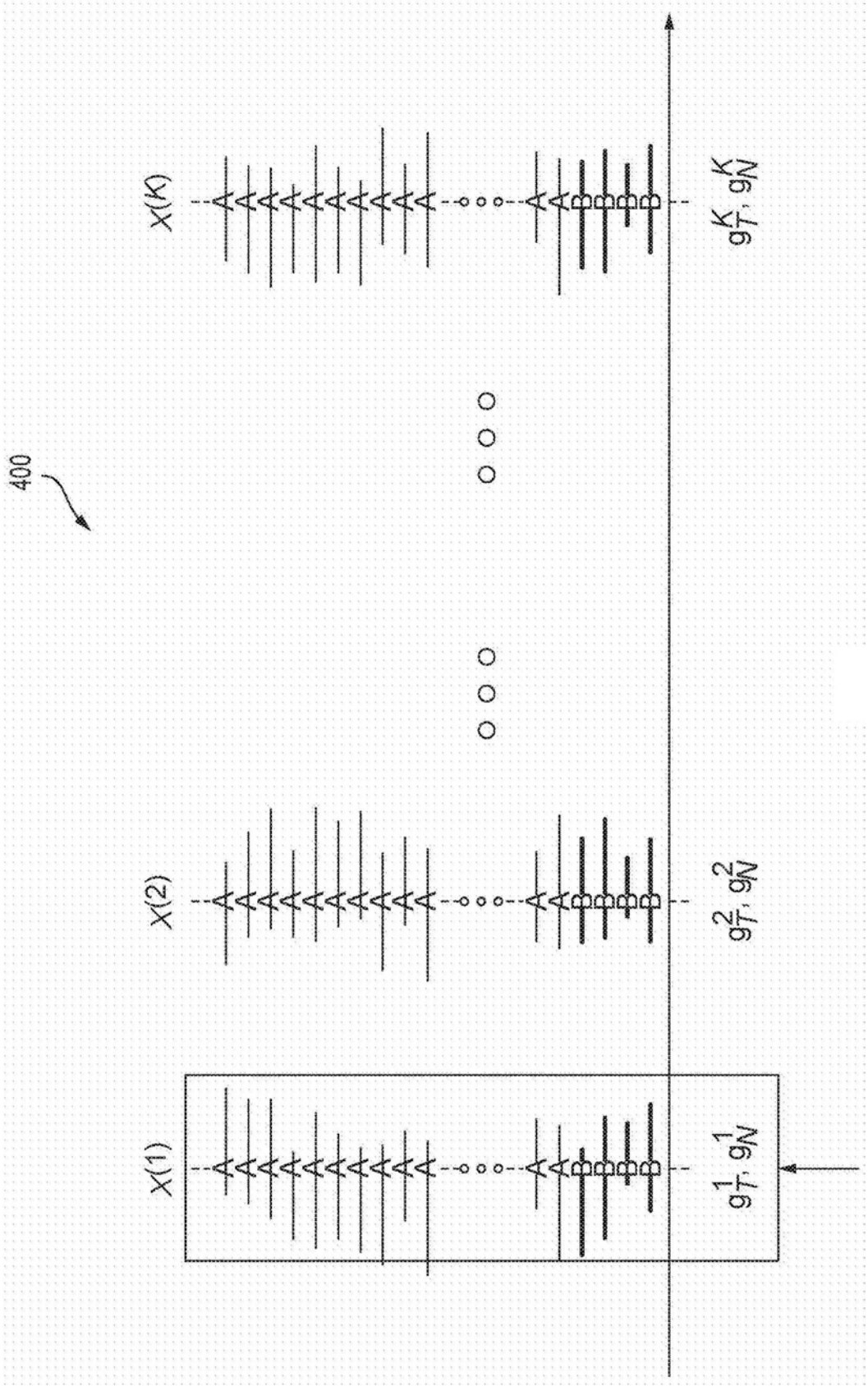


图4

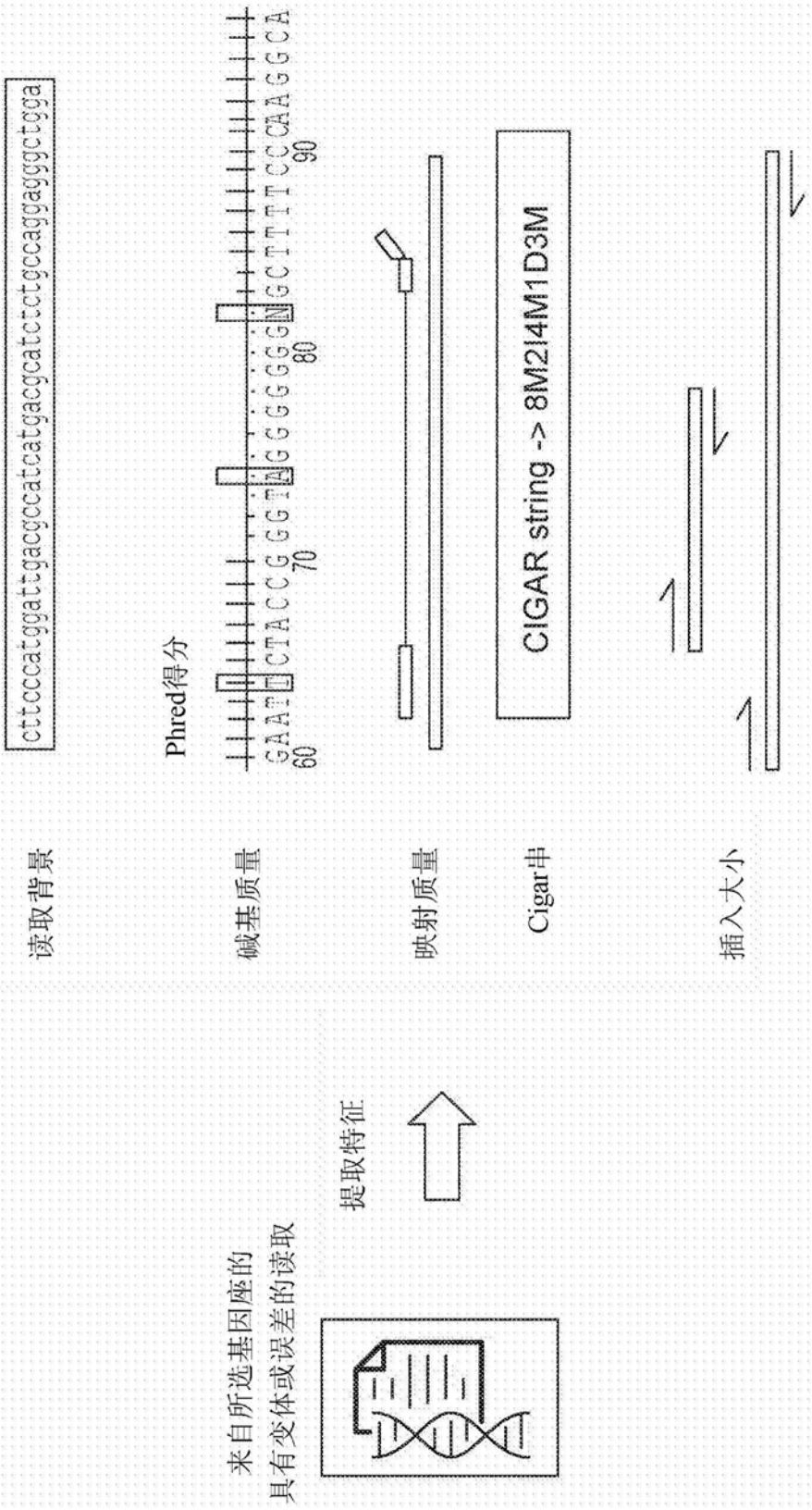

ctticccatggattgacgccatcatgacgcatcccctgccaggagggctggga
Phred得分
60
70
80
90
CIGAR string -> 8M2I4M1D3M
读取背景
碱基质量
映射质量
Cigar串
插入大小
来自所选基因座的
具有变体或误差的读取
提取特征

图5

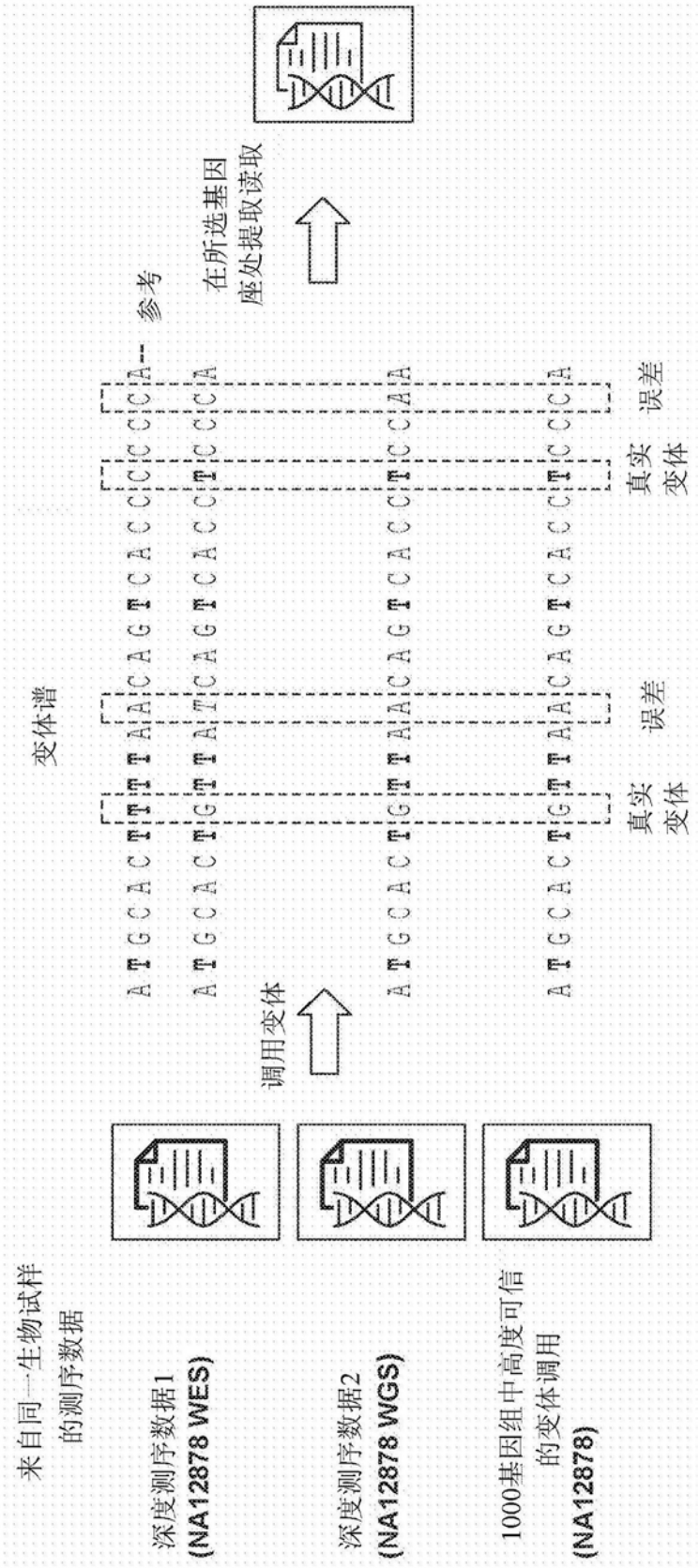

来自同一生物试样
的测序数据
深度测序数据1
(NA12878 WES)
深度测序数据2
(NA12878 WGS)
1000基因组中高度可信
的变体调用
(NA12878)
调用变体
变体谱
参考
在所选基因
座处提取读取
真实
变体
误差
真实
变体
误差

图6

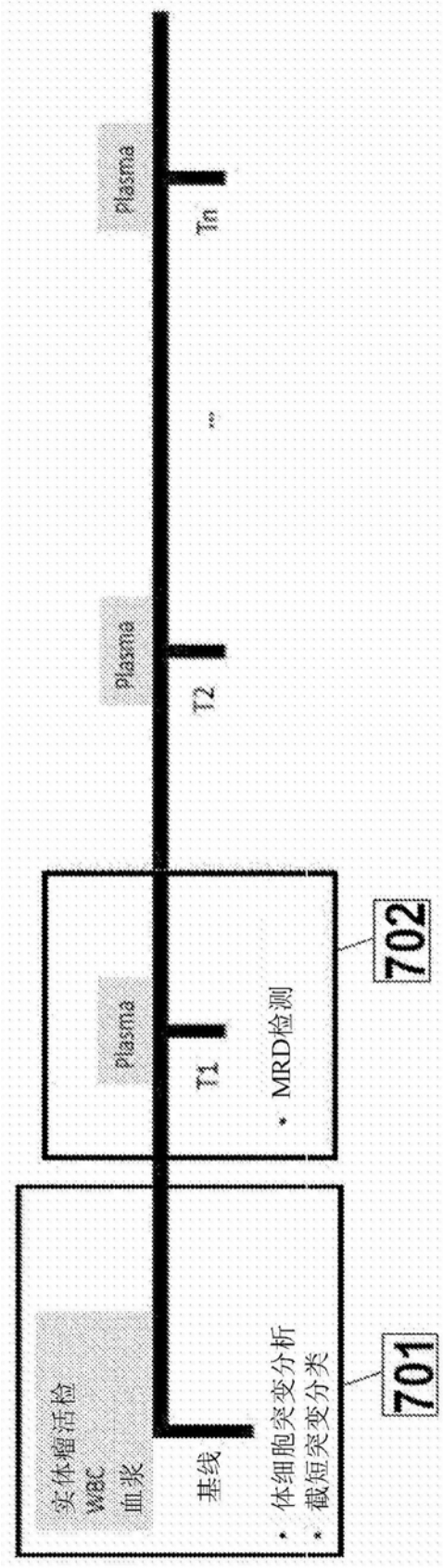

实体瘤活检
WBC
血浆
基线
• 体细胞突变分析
• 截短突变分类
701
Plasma
T1
• MRD检测
702
Plasma
T2
...
Plasma
Tn

图7

基线处的截短体细胞SNV谱

变体	cfDNA中的碱基	cfDNA质量	正常中的碱基	正常的质量
Chr1:12	RrrAa*T*	ZRK:5?	rRRRRr	C8_=BC
Chr3:29	RrRCC	\==OJ	rRrrR	Ab,!:
Chr3:42	RrRR*TTtt*	AK4!KSJR	rRRrRr	Cs1=BC
Chr4:99	RrRCca	\RK<>H	rRRrR	<>CA:
Chr7:15	RRrAa*T*	ZG=JI?	rrrRRr	C!F=BC
Chr9:35	RrrRGtg	EEF^OJ	rRrrR	CR>A:

用ML模型
（基于读取对的）
过滤误差

T1处的血浆WES

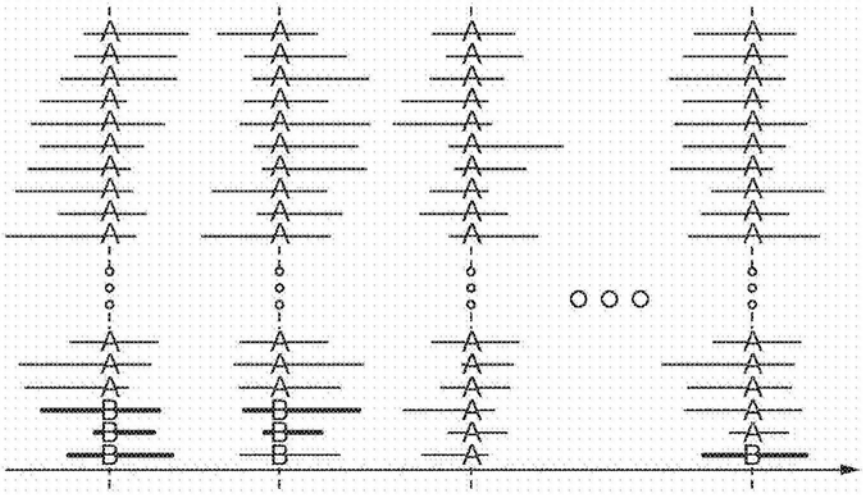


图8

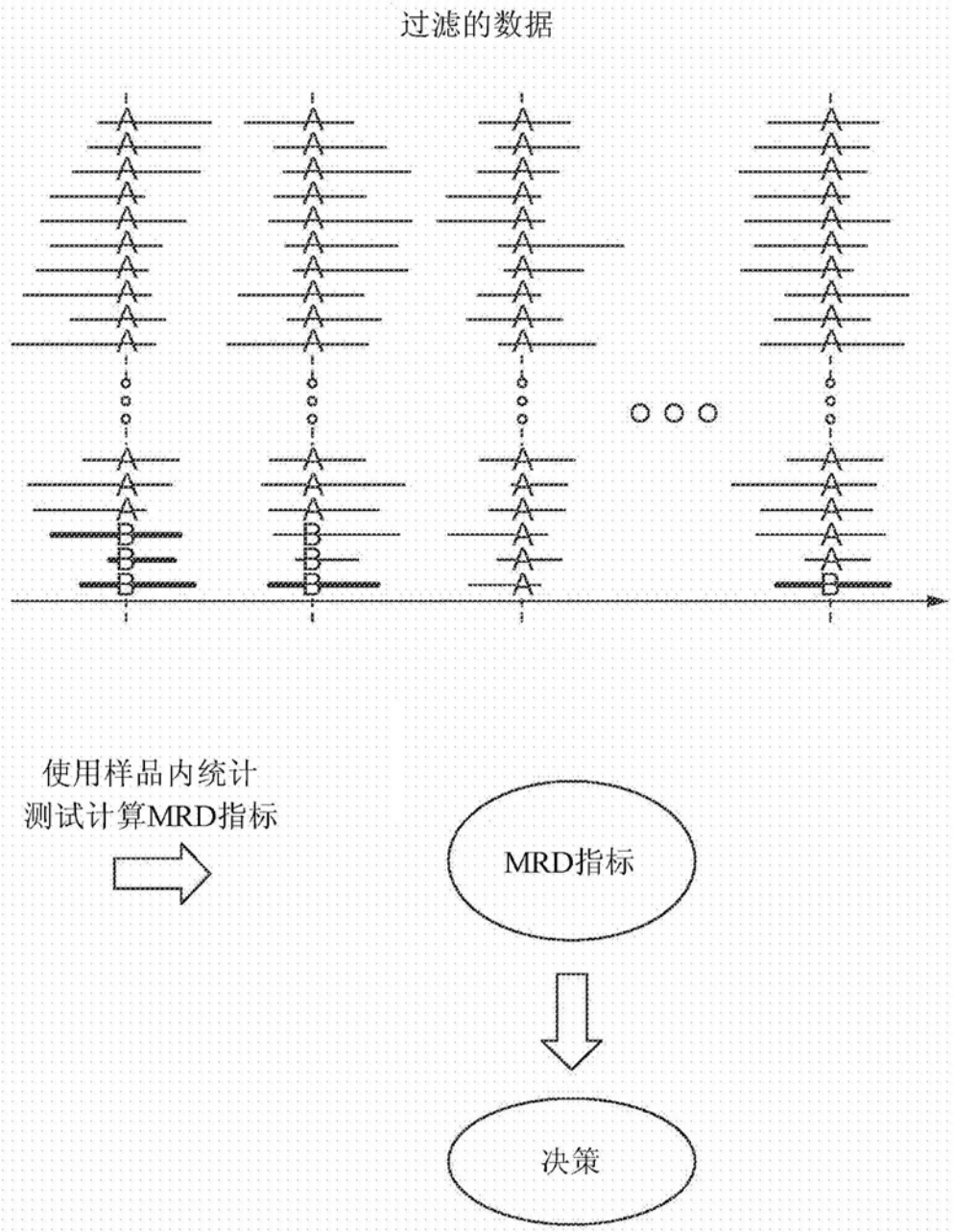


图8(续)

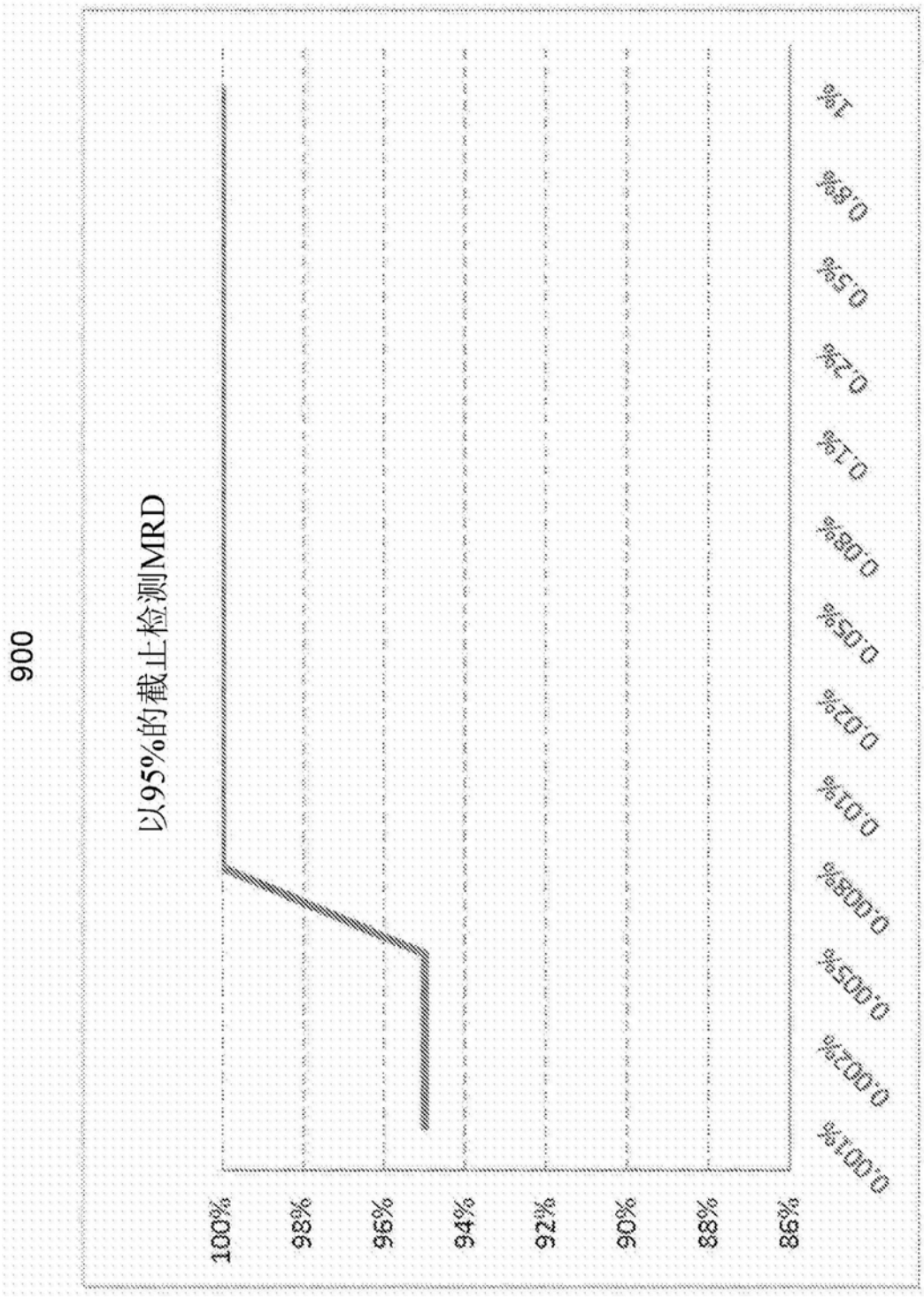

900
以95%的截止检测MRD
100%
98%
96%
94%
92%
90%
88%
86%
1%
0.8%
0.5%
0.2%
0.1%
0.08%
0.05%
0.02%
0.01%
0.008%
0.005%
0.002%
0.001%

图9

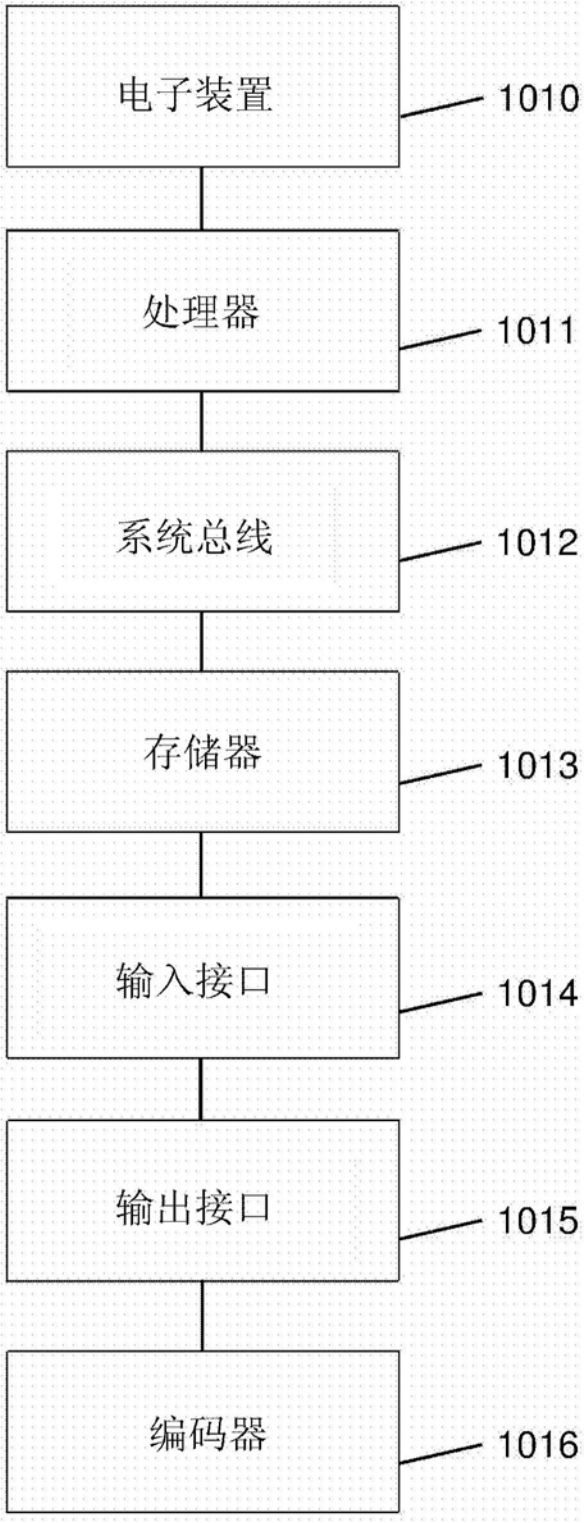

电子装置
1010
处理器
1011
系统总线
1012
存储器
1013
输入接口
1014
输出接口
1015
编码器
1016

图10