(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 114424288 A
(43)申请公布日 2022.04.29

(21)申请号 202080065982.6
(22)申请日 2020.09.29
(30)优先权数据
1914064.9 2019.09.30 GB
(85)PCT国际申请进入国家阶段日
2022.03.21
(86)PCT国际申请的申请数据
PCT/GB2020/052358 2020.09.29
(87)PCT国际申请的公布数据
WO2021/064365 EN 2021.04.08
(71)申请人 朗斯科技有限公司
地址 澳大利亚新南威尔士州悉尼
(72)发明人 艾伦·厄尔·达尔林
(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
代理人 王达佐 洪欣
(51)Int.Cl.
G16B 30/10(2019.01)
G16B 30/20(2019.01)
G16B 40/00(2019.01)

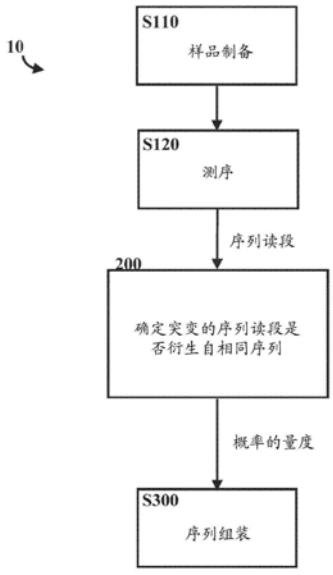
权利要求书4页 说明书27页
序列表1页 附图5页

(54)发明名称

用于确定与两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度的方法

(57)摘要

本发明公开了一种用于确定与两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度的计算机实现的方法。该方法包括:接收突变的序列读段,与不包含突变的序列相比,每个突变的序列读段对应于包含突变的序列的亚序列;将公共极小函数应用于每个突变的序列读段,以为每个突变的序列读段确定极小元;确定每个突变的序列读段中的该一个或多个极小元的位置;确定每个突变的序列读段中突变的位置;以及对于具有公共极小元的至少两个突变的序列读段,当比对所述相应的极小元时,对具有匹配位置和/或具有错配位置的突变的数量计数。本发明还公开了一种用于确定至少一个靶模板核酸分子的序列的至少一部分的对应方法。



CN 114424288 A

1.一种用于确定与两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度的计算机实现的方法,所述方法包括:

接收多个突变的序列读段,每个突变的序列读段对应于包含突变的序列的亚序列,其中与不包含突变的序列相比,所述包含突变的序列包含突变;

将公共极小函数应用于每个突变的序列读段,从而为每个突变的序列读段确定一个或多个相应的极小元;

确定每个突变的序列读段中的所述一个或多个相应的极小元的位置;

确定每个突变的序列读段中的一个或多个突变的位置;以及

对于具有公共极小元的至少两个突变的序列读段,当比对所述相应的极小元时,对具有匹配位置和/或具有错配位置的突变的数量计数。

2.根据权利要求1所述的方法,所述方法还包括接收多个非突变的序列读段,每个非突变的序列读段对应于所述不包含突变的序列的亚序列。

3.根据权利要求1或2所述的方法,其中所述将公共极小函数应用于每个突变的序列读段的步骤包括:i)识别所述相应的突变的序列读段中的在可能的k-mer的有序列表中首先列出的一个或多个k-mer,或者ii)识别存在于可能的k-mer的预先确定集合中的所述一个或多个k-mer,其中为所述相应的突变的序列读段确定的所述一个或多个极小元是所述已识别的一个或多个k-mer。

4.根据权利要求3所述的方法,其中,i)在可能的k-mer的所述有序列表中,基于所述k-mer存在于所述包含突变的序列中并且不存在于所述不包含突变的序列中的概率,对所述k-mer进行排序,或者ii)可能的k-mer的所述预先确定集合包含相对可能存在于所述包含突变的序列中并且不存在于所述不包含突变的序列中的k-mer,任选地,其中可能的k-mer的所述预先确定集合不包含相对不可能存在于所述包含突变的序列中的k-mer。

5.根据权利要求2和权利要求3或4所述的方法,其中可能的k-mer的所述有序列表或可能的k-mer的所述预先确定集合由k-mer组成,所述k-mer在所述多个突变的序列读段中比在所述多个非突变的序列读段中更经常存在,任选地,其中在所述多个突变的序列读段中比在所述多个非突变的序列读段中更经常存在的k-mer相对可能存在于所述包含突变的序列中。

6.根据权利要求2和权利要求3至5中任一项所述的方法,其中可能的k-mer的所述预先确定集合由k-mer组成,所述k-mer在所述多个突变的序列读段中存在n次或更多次并且在所述多个非突变的序列读段中存在少于n次,其中n为大于或等于1的整数,任选地,其中在所述多个突变的序列读段中存在n次或更多次并且在所述多个非突变的序列读段中存在少于n次的k-mer相对可能存在于所述包含突变的序列中。

7.根据权利要求6所述的方法,其中可能的k-mer的所述预先确定集合由不存在于所述多个非突变的序列读段中的k-mer组成。

8.根据权利要求6或7所述的方法,其中n为2。

9.根据权利要求2和权利要求3至8中任一项所述的方法,所述方法还包括基于所述多个突变的序列读段中的所述k-mer与所述多个非突变的序列读段中的所述k-mer的比较,生成可能的k-mer的所述有序列表或可能的k-mer的所述预先确定集合。

10.根据前述权利要求中任一项所述的方法,其中每个极小元均为长度大于5,优选地

大于10的k-mer。

11.根据前述权利要求中任一项所述的方法,所述方法还包括在一个或多个极小元箱中将所述突变的序列读段分箱,使得每个极小元箱含有具有公共极小元的突变的序列读段并且不含有不具有公共极小元的突变的序列读段,并且

其中,仅对相同的极小元箱中的突变的序列读段执行所述对具有匹配位置和/或具有错配位置的突变的数量计数的步骤。

12.根据权利要求2和权利要求2至11中任一项所述的方法,其中所述确定每个突变的序列读段中所述一个或多个突变的所述位置的步骤包括:

通过将一个或多个种子模式中的每个种子模式应用于所述多个非突变的序列读段中的k-mer,获得非突变的种子掩蔽的k-mer集合;

对于每个突变的序列读段,将所述一个或多个种子模式应用于所述相应的突变的序列读段中的k-mer,以获得多个突变的种子掩蔽的k-mer,并且通过识别所述突变的序列读段中的被所有的所述种子模式掩蔽的所述一个或多个位置来确定所述一个或多个突变的所述位置,所述种子模式对应于存在于所述非突变的种子掩蔽的k-mer集合中的所述多个突变的种子掩蔽的k-mer中的所述突变的种子掩蔽的k-mer。

13.根据权利要求12所述的方法,其中选择所述一个或多个种子模式,使得通过将所述一个或多个种子模式中的至少一个种子模式应用于所述多个突变的序列读段中的任何k-mer,并且应用于所述多个非突变的序列读段中对应的k-mer,获得相同的种子掩蔽的k-mer的概率大于90%,优选地大于99%。

14.根据权利要求12或13所述的方法,其中与所述不包含突变的序列相比,所述包含突变的序列包含转换突变;并且

其中所述一个或多个种子模式由一个或多个转换种子模式组成。

15.根据权利要求12至14中任一项所述的方法,其中所述多个突变的序列读段中的每个突变的序列读段对应于包含与多个样品中的一个样品相关联的突变的序列的亚序列,并且所述多个非突变的序列读段中的每个非突变的序列读段对应于不包含与所述多个样品中的一个样品相关联的突变的序列的亚序列,其中与相应的不包含突变的序列相比,每个包含突变的序列包含突变;

其中获得非突变的种子掩蔽的k-mer集合包括获得每个样品的相应的非突变的种子掩蔽的k-mer集合;

所述方法还包括创建非突变样品位向量的集合,每个非突变样品位向量针对所述非突变的种子掩蔽的k-mer集合中的相应k-mer定义,在所述多个样品的所述非突变的种子掩蔽的k-mer集合中存在所述相应k-mer;并且

其中确定所述一个或多个突变的所述位置包括:对于每个突变的序列读段,以及对于非突变的种子掩蔽的k-mer集合中的每个集合和/或每个组合,识别所述突变的序列读段中的被所有的所述种子模式掩蔽的所述一个或多个位置,所述种子模式对应于存在于非突变的种子掩蔽的k-mer集合的所述相应集合或组合中的所述多个突变的种子掩蔽的k-mer中的所述突变的种子掩蔽的k-mer;以及将所识别的一个或多个位置与所述一个或多个样品相关联,所述一个或多个样品与非突变的种子掩蔽的k-mer集合的所述相应集合或组合相关联。

16.根据权利要求2至15中任一项所述的方法,其中所述确定每个突变的序列读段中所述一个或多个突变的所述位置的步骤包括:

对于所述突变的序列读段中的一个或多个突变的序列读段,将所述相应的突变的序列读段与参考组装进行比对;以及

通过识别所述相应的突变的序列读段与所述参考组装之间的差异在所述相应的突变的序列读段中的位置,确定所述相应的突变的序列读段中所述一个或多个突变的位置。

17.根据从属于权利要求12至15中任一项的权利要求16所述的方法,其中所述对于每个突变的序列读段确定每个突变的序列读段中所述一个或多个突变的所述位置的步骤包括:

若将所述相应的突变的序列读段中的位置与所述参考组装比对,如果所述相应的突变的序列读段中的所述位置是所述相应的突变的序列读段不同于所述参考组装的位置,则将所述相应的突变的序列读段中的所述位置确定为所述相应的突变的序列读段中突变的位置;以及

若没有将所述相应的突变的序列读段中的位置与所述参考组装比对,如果所述相应的突变的序列读段中的所述位置是被所有的所述种子模式掩蔽的位置,所述种子模式对应于存在于所述非突变的种子掩蔽的k-mer集合中的所述多个突变的种子掩蔽的k-mer中的所述突变的种子掩蔽的k-mer,则将所述相应的突变的序列读段中的所述位置确定为所述相应的突变的序列读段中突变的位置。

18.根据前述权利要求中任一项所述的方法,所述方法包括基于具有匹配位置和/或具有错配位置的突变的数量,确定与所述至少两个突变的序列读段衍生自包含突变的相同序列的概率相关的所述量度。

19.根据权利要求18所述的方法,其中与所述至少两个突变的序列读段衍生自包含突变的相同序列的概率相关的所述量度为以下两者中的一者:i)所述至少两个突变的序列读段衍生自包含突变的相同序列的概率密度,和ii)与所述至少两个突变的序列读段衍生自包含突变的相同序列的所述概率密度相关的分数函数。

20.根据前述权利要求中任一项所述的方法,所述方法还包括从所述多个突变的序列读段创建无向加权图,

其中所述无向加权图包括对应于所述多个突变的序列读段的节点,并且其中所述节点之间的边缘与相应的边缘权重相关联,其中每个边缘权重基于具有匹配位置和/或具有错配位置的突变的数量来确定,所述突变的数量针对对应于与所述相应的边缘相关联的所述两个节点的两个突变的序列读段来确定。

21.根据权利要求20所述的方法,其中所述边缘权重对应于与所述至少两个突变的序列读段衍生自包含突变的相同序列的概率相关的所述量度,所述至少两个突变的序列读段对应于与所述相应的边缘相关联的所述两个节点。

22.根据权利要求20或21所述的方法,所述方法还包括对所述无向加权图执行图聚类,从而创建预期衍生自包含突变的相同序列的突变的序列读段的簇。

23.根据权利要求22所述的方法,其中所述图聚类包括马尔可夫聚类或Infomap。

24.根据权利要求22或23所述的方法,所述方法还包括通过组装所述簇中的所述突变的序列读段来重建所述包含突变的序列的至少一部分。

25.根据权利要求2和24所述的方法，所述方法还包括通过从所述包含突变的序列的所述重建部分推断所述可能的不包含突变的序列的至少一部分，任选地通过使用所述多个非突变的序列读段，来重建所述不包含突变的序列的至少一部分。

26.一种用于生成靶模板核酸分子的序列的至少一部分的方法，所述方法包括根据权利要求20至25中任一项所述的方法。

27.一种用于确定至少一个靶模板核酸分子的序列的至少一部分的方法，所述方法包括

对包含突变的至少一个靶模板核酸分子的区域测序，以提供多个突变的序列读段，

对获得的多个突变的序列读段进行根据前述权利要求中任一项所述的方法。

28.根据权利要求27所述的方法，其中所述测序步骤包括

(a)提供一对样品，每个样品包含至少一个靶模板核酸分子；

(b)对所述一对样品的第一样品中的所述至少一个靶模板核酸分子的区域测序，以提供所述多个非突变的序列读段；

(c)将突变引入所述一对样品的第二样品中的所述至少一个靶模板核酸分子中，以提供至少一个突变的靶模板核酸分子；

(d)对所述至少一个突变的靶模板核酸分子的区域测序，以提供所述多个突变的序列读段。

29.根据权利要求28所述的方法，其中所述引入突变的步骤包括将转换突变引入所述一对样品的所述第二样品中的所述至少一个靶模板核酸分子中。

用于确定与两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度的方法

技术领域

[0001] 本发明涉及用于确定与两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度的计算机实现的方法、用于生成序列的至少一部分的方法，以及用于确定至少一个靶模板核酸分子的序列的方法。

背景技术

[0002] 对核酸分子测序的能力是在无数不同应用中非常有用的工具。然而，可能难以确定包含问题结构的核酸分子的准确序列，诸如包含重复区域的核酸分子的准确序列。也可能难以解析结构特征，诸如二倍体和多倍体生物体的单倍型结构以及这些生物体基因组中的结构变体。

[0003] 许多更现代的技术(所谓的下一代测序技术)仅能够准确地对短核酸分子测序。下一代测序技术可用于对较长核酸序列测序，但这通常是困难且昂贵的。下一代测序技术可用于生成对应于核酸分子的部分的序列的短序列读段，并且可以从短序列读段组装完整序列。在核酸分子包含重复区域的情况下，用户可能不清楚具有相似序列的两个序列读段是对应于较长序列中的两个重复的序列，还是对应于相同序列的两个重复。相似地，用户可能希望同时对两个相似的核酸分子测序，并且可能难以确定具有相似序列的两个序列读段是对应于相同的原始核酸分子的序列，还是对应于两个不同的原始核酸分子的序列。

[0004] 可使用突变辅助测序(SAM)技术来辅助从短序列读段组装序列。一般来讲，SAM涉及将突变引入靶模板核酸序列中。引入的突变模式可帮助该方法的用户从短序列读段组装核酸分子的序列。

[0005] 例如，在模板核酸分子含有重复区域的情况下，可通过不同的突变模式将重复彼此区分开，从而使重复区域能够被正确解析和组装。

[0006] 一般来讲，SAM技术涉及使靶模板核酸分子的拷贝突变以产生突变的靶模板核酸分子和/或包含突变的一个或多个序列，对包含突变的一个或多个序列测序以创建包含突变的序列读段的SAM数据，并且随后基于它们的突变模式从突变的序列读段组装序列。由于不同的突变拷贝将包含在不同位置处的突变，因此组装的序列可代表原始模板核酸分子。

[0007] 然而，仍然需要更可靠和/或计算效率更高的方法来处理SAM数据。

发明内容

[0008] 本发明人已经开发了用于处理包含突变的序列读段的SAM数据的新改进方法。因此，在本发明的一个方面，提供了一种用于确定与两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度的计算机实现的方法。该方法包括接收多个突变的序列读段。每个突变的序列读段对应于包含突变的序列的亚序列。与不包含突变的序列相比，该包含突变的序列包含突变。该方法还包括，将公共极小函数应用于每个突变的序列读段，从而为每个突变的序列读段确定一个或多个相应的极小元。该方法还包括确定每个突变的序列读

段中的一个或多个相应的极小元的位置。该方法还包括确定每个突变的序列读段中的一个或多个突变的位置。该方法还包括,对于具有公共极小元的至少两个突变的序列读段,当比对相应的极小元时,对具有匹配位置和/或具有错配位置的突变的数量计数。

[0009] 在本发明的另一个方面,提供了用于生成靶模板核酸分子的序列的至少一部分的方法。

[0010] 在本发明的另一个方面,提供了用于确定至少一个靶模板核酸分子的序列的至少一部分的方法。

[0011] 在从属权利要求和具体实施方式中提供了本发明另外的方面。

附图说明

[0012] 图1示出了用于确定根据本发明的至少一个靶模板核酸分子的至少一部分的方法的实施方案。

[0013] 图2示出了用于确定根据本发明的与两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度的方法的实施方案。

[0014] 图3示出了确定突变的序列读段中一个或多个突变的位置的步骤的实施例。

[0015] 图4A示出了不使用本发明方法的2.3Mbp Arcobacter butzlerii基因组的短读段组装的比较例。

[0016] 图4B示出了使用本发明方法的组装2.3Mbp Arcobacter butzlerii基因组的实施例。

[0017] 图5示出了关于每个长模板的短读段覆盖率深度对本发明方法结果的影响的实验数据。

具体实施方式

[0018] 一般定义

[0019] 除非另有定义,否则本文所用的技术和科学术语都具有本发明所属技术领域的技术人员通常理解的含义。

[0020] 一般来讲,术语“包括”旨在表示包括但不限于。例如,短语“包括[某些步骤]的方法”应解释为意指该方法包括所述步骤,但是可执行附加步骤。

[0021] 在本发明的一些实施方案中,词语“包括”可被替换为短语“由……组成”。术语“由……组成”旨在进行限制。例如,短语“由[某些步骤]组成的方法”应理解为意指该方法包括所述步骤,并且不执行附加步骤。

[0022] 在一些方面,本发明提供了用于确定或生成至少一个靶模板核酸分子的序列的至少一部分的方法。该方法可用于确定或生成至少一个靶模板核酸分子的完整序列。另选地,该方法可用于确定或生成部分序列,即该至少一个靶模板核酸分子的一部分的序列。例如,如果不可能或不直接确定完整序列,则用户可决定至少一个靶模板核酸分子的一部分的序列对于其目的是有用的或甚至是足够的。

[0023] 出于本发明的目的,“核酸分子”(或“非突变核酸分子”)是指任何长度的核苷酸的聚合形式。核苷酸可以是脱氧核糖核苷酸、核糖核苷酸或它们的类似物。优选地,至少一个核酸分子由脱氧核糖核苷酸或核糖核苷酸构成。甚至更优选地,至少一个核酸分子由脱氧

核糖核苷酸构成,即该至少一个核酸分子是DNA分子。

[0024] “靶模板核酸分子”可以是用户想要测序的任何核酸分子。至少一个“靶模板核酸分子”可以是单链的,或者可以是双链复合物的一部分。如果该至少一个靶模板核酸分子由脱氧核糖核苷酸构成,则其可形成双链DNA复合物的一部分。在这种情况下,一条链(例如编码链)将被认为是至少一个靶模板核酸分子,并且另一条链是与该至少一个靶模板核酸分子互补的核酸分子。该至少一个靶模板核酸分子可以是对应于基因的DNA分子,可以包含内含子,可以是基因间区域,可以是基因内区域,可以是跨越多个基因的基因组区域,或者实际上可以是生物体的整个基因组。

[0025] 出于本发明的目的,“突变核酸分子”或“突变的靶模板核酸分子”是指已引入突变的“核酸分子”或“靶模板核酸分子”。突变可以是取代突变,任选地是转换突变。出于本发明的目的,术语“取代突变”应解释为意指核苷酸被不同核苷酸取代。例如,序列ATCC到序列AGCC的转换引入了单个取代突变。出于本发明的目的,术语“转换突变”应解释为意指核苷酸A被核苷酸G取代,并且反之亦然,(即,突变A G),或者核苷酸C被核苷酸T取代,并且反之亦然,(即,突变C T)。

[0026] 短语“将突变引入至少一个靶模板核酸分子中”是指将该对样品的第二样品中的至少一个靶模板核酸分子暴露于其中该至少一个靶模板核酸分子发生突变的条件下。可使用任何合适的方法来实现这一点。例如,可通过化学诱变和/或酶促诱变引入突变。

[0027] 出于本发明的目的,“包含突变的序列”对应于“突变核酸分子”或“突变的靶模板核酸分子”中的核苷酸序列的至少一部分。“包含突变的序列”也可以称为“突变序列”。“包含突变的序列”在本文中表示为μ^i,并且多个(即许多个)“包含突变的序列”表示为M,其中$\mu^1\cdots\mu^n\in M$。“不包含突变的序列”对应于“核酸分子”或“靶模板核酸分子”中核苷酸序列的至少一部分。“不包含突变的序列”也可以称为“非突变序列”。“不包含突变的序列”在本文中表示为S^i,并且多个(即许多个)“不包含突变的序列”表示为S,其中$S^1\cdots S^n\in S$。因此“包含突变的序列”和“不包含突变的序列”可对应于核苷酸(nt)腺嘌呤(A)、胸腺嘧啶(T)、鸟嘌呤(G)和胞嘧啶(C)的核酸分子亚序列的至少一部分。此类染色体序列的长度尺寸范围可从10^3个核苷酸(nt)至10^9个核苷酸,甚至更长。

[0028] 出于本发明的目的,“突变的序列读段”对应于“包含突变的序列”的亚序列,即,“突变的序列读段”可与“包含突变的序列”的至少一个亚序列基本相同,但与包含突变的序列相比,“突变的序列读段”包含突变,并且可包含由于读段错误而产生的附加小差异。“突变的序列读段”表示为ρ^i,并且多个(即许多个)“突变的序列读段”表示为P,其中$\rho^1\cdots\rho^n\in P$。“非突变的序列读段”对应于“不包含突变的序列”的亚序列,即,除了测序期间的读段错误,“非突变的序列读段”可与“不包含突变的序列”的亚序列基本相同。“非突变的序列读段”表示为r^i,并且多个(即许多个)“非突变的序列读段”表示为R,其中$r^1\cdots r^n\in R$。“突变的序列读段”可通过对“突变的靶模板核酸分子”的区域测序来获得,并且“非突变的序列读段”可通过对“靶模板核酸分子”的区域测序来获得。序列读段的长度可小于序列,例如为约150nt的长度。

[0029] 序列分析方法10

[0030] 图1示出了根据本发明确定至少一个靶模板核酸分子的至少一部分的方法10。

[0031] 确定至少一个靶模板核酸分子的至少一部分的方法10可包括样品制备步骤S110。

样品制备步骤S110可包括提供一对靶模板核酸分子，并且将突变引入该对靶模板核酸分子中的一个靶模板核酸分子中，以便产生突变的靶模板核酸分子。样品制备步骤S110可包括用于提供靶模板核酸分子和突变的靶模板核酸分子的任何已知技术。

[0032] 确定至少一个靶模板核酸分子的至少一部分的方法10还可包括测序步骤S120。测序步骤S120包括对包含突变的至少一个靶模板核酸分子的区域测序，从而提供多个突变的序列读段P。此外，测序步骤S120可包括对该至少一个(非突变)靶模板核酸分子(对应于突变的靶模板核酸分子的靶模板核酸分子)的区域测序，从而提供多个非突变的序列读段R。步骤S120可包括用于生成多个突变的序列读段P的任何已知技术。

[0033] 确定至少一个靶模板核酸分子的至少一部分的方法10包括确定两个突变的序列读段ρ^i、ρ^j是否衍生(或源)自包含突变的相同序列μ^i的步骤200或方法200。确定两个突变的序列读段ρ^i、ρ^j是否衍生(或源)自包含突变的相同序列μ^i包括确定两个突变的序列读段ρ^i、ρ^j是否衍生(或源)自包含突变的序列μ^i的相同或相似或重叠部分，即确定两个突变的序列读段ρ^i、ρ^j两者是否均包括对应于包含突变的序列μ^i的相同部分的亚序列。方法200是计算机实现的方法，并且可由计算机的处理器执行。方法200产生与两个突变的序列读段ρ^i、ρ^j衍生自包含突变的相同序列μ^i的概率相关的量度。

[0034] 确定至少一个靶模板核酸分子的至少一部分的方法10还可包括序列组装步骤S300。序列组装步骤S300包括组装或重建序列μ^i、S^i的至少一部分。可通过基于与相应的两个突变的序列读段ρ^i、ρ^j衍生自包含突变的相同序列μ^i的概率相关的量度来组装该多个突变的序列读段P，从而获得包含突变的序列μ^i。例如，可通过将该多个突变的序列读段P分组为对应于包含突变序列μ^i的组，然后分别组装每个组以重建部分或全部包含突变的单个序列μ^i来实现这一点。不包含突变的序列S^i可通过对包含突变的序列μ^i进行错误校正来获得，例如通过使用该多个非突变的序列读段R从包含突变的序列μ^i推断最可能不包含突变的序列S^i来获得。序列组装步骤S300可包括用于基于与相应的两个突变的序列读段ρ^i、ρ^j衍生自包含突变的相同序列μ^i的概率相关的量度，从多个突变的序列读段P组装包含突变的序列μ^i的任何已知方法。。

[0035] 图2示出了根据本发明确定两个突变的序列读段ρ^i、ρ^j是否衍生自包含突变的相同序列μ^i的方法200。

[0036] 该方法200包括接收多个突变的序列读段$\rho^1 \cdots \rho^n \in P$的步骤S210。每个突变的序列读段$\rho^i$对应于包含突变的序列$\mu^i$的亚序列。与不包含突变的序列$S^i$相比，包含突变的序列$\mu^i$包含突变，例如取代突变、任选地转换突变。包含突变的序列μ^i可以是突变的靶模板核酸分子的序列的至少一部分，并且不包含突变的序列可以是(非突变的)靶模板核酸分子的至少一部分，其中突变的靶模板核酸分子通过将突变，例如取代突变，任选地转换突变引入靶模板核酸分子中获得。包含突变的序列μ^i的每个亚序列可以是突变的靶模板核酸分子的片段的序列的至少一部分。不包含突变的序列S^i的每个亚序列可以是靶模板核酸分子的片段的序列的至少一部分。接收该多个突变的序列读段P的步骤S210可包括直接从用于对突变的靶模板核酸分子测序的测序机器接收该多个突变的序列读段P，或者从存储该多个突变的序列读段P的数据存储装置接收该多个突变的序列读段P。

[0037] 方法200还包括将公共极小函数应用于每个突变的序列读段ρ^i的步骤S220。应用公共极小函数为每个突变的序列读段ρ^i确定一个或多个相应的极小元。方法200还包括确

定每个突变的序列读段ρ^i中的一个或多个相应的极小元的位置的步骤S222。

[0038] 在优选的实施方案中，方法200包括通过它们相应的极小元将突变的序列读段P分箱的步骤S224。可在多个相应的极小元箱中提供已经确定了多于一个极小元的突变的序列读段ρ^i。

[0039] 方法200还包括确定每个突变的序列读段ρ^i中一个或多个突变的位置的步骤S230。确定每个突变的序列读段ρ^i中的一个或多个突变的位置的步骤S230可在与公共极小函数有关的步骤S220、S222和S224之前、之后或同时执行。

[0040] 方法200还包括，对于具有公共极小元的至少两个突变的序列读段ρ^i、ρ^j，当比对相应的极小元时，即当一个突变的序列读段ρ^i的核苷酸位置相对于另一个突变的序列读段ρ^j的核苷酸位置偏移时，使得一个突变的序列读段ρ^i的极小元的位置与另一个突变的序列读段ρ^j的极小元的位置相同，对具有匹配位置和/或具有错配位置的突变的数量计数。具有匹配位置和/或具有错配位置的突变的数量可以是与两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度。另选地，方法200可包括基于具有匹配位置和/或具有错配位置的突变的数量，确定与两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度的另外的步骤S242。

[0041] 接收多个突变的序列读段的步骤S210

[0042] 步骤S210包括接收多个突变的序列读段$\rho^1\cdots\rho^n\in P$。步骤S210还可包括接收多个非突变的序列读段$r^1\cdots r^m\in R$。每个突变的序列读段ρ^i可对应于包含突变的序列μ^i的亚序列。每个非突变的序列读段r^i可对应于不包含突变的序列S^i的亚序列。

[0043] 包含突变的序列μ^i可通过将突变引入不包含突变的序列S^i中获得。因此每个突变的序列读段ρ^i可包含突变，即对应于突变的靶模板核酸分子的包括突变的区域，即包含突变的序列的亚序列。在实施方案中，每个突变的序列读段ρ^i包含取代突变，即对应于突变的靶模板核酸分子的包括取代突变的区域。在优选的实施方案中，取代突变是转换突变，因此每个突变的序列读段ρ^i包含转换突变，即对应于突变的靶模板核酸分子的包括转换突变的区域。

[0044] 每个序列读段ρ^i、r^i的每个核苷酸优选地使用两位以二进制格式编码。有利的是，特别是当该多个突变的序列读段P包括转换突变(A G和C T)时，两位中的一位(例如第一位)定义核苷酸是嘌呤(A或G)还是嘧啶(T或C)。例如，核苷酸可使用以下格式以二进制形式编码：A:00、G:01、C:10和T:11。本说明书中将使用该编码。然而，显而易见的是，本发明不限于该编码，并且本发明可使用核苷酸的任何其他编码容易地执行。

[0045] 每个序列读段ρ^i、r^i可被编码以解释序列读段ρ^i、r^i中的均聚物错误。当相同核苷酸的运行以错误的长度被误读时，就会出现均聚物错误，例如，序列TAAAAGC可能被误读为TAAGC，因为测序机器难以区分多个A的运行中A的数量。为了解释此类均聚物错误，多个相同核苷酸的运行可被编码为核苷酸的单个实例。另选地，可在序列读段ρ^i、r^i的后续处理(即不是初始编码)期间解释均聚物错误。例如，通过对方法200中使用的任何k-mer和/或步骤S230中使用的任何种子模式进行编码，使得多个相同核苷酸的运行被编码为核苷酸的单个实例。

[0046] 步骤S220和S222：公共极小函数

[0047] 极小元是来自k-mer集合中的k-mer，它满足该k-mer集合上的公共极小函数min

()。

[0048] 出于本申请的目的,k-mer是长度为k的核苷酸亚序列。长度为n的序列S＝[S_1,S_2,…,S_{n-1},S_n]中从位置i开始的k-mer表示为k(S_i),其中k(S_i)＝[S_i,S_{i+1},…,S_{i+k-1}]。序列S中起始位置在i与j之间的k-mer集合表示为k(S_i...S_j)。起始位置在序列S的i至j范围内的所有k-mer的极小元将被表示为min(k(S_i...S_j))。

[0049] 公共极小函数min()用于从由序列读段ρ^i、r^i包含的k-mer集合,优选地所有或基本上所有的k-mer(即k-mer,优选地为存在于序列读段ρ^i、r^i中的所有或基本上所有的k-mer)中确定一个或多个极小元(即一个或多个代表性k-mer)。出于本发明的目的,存在于序列读段ρ^i、r^i中的k-mer集合可包括序列读段ρ^i、r^i的反向互补的k-mer。优选地,每个极小元是长度等于或大于5的k-mer(即5-mer或更长),优选地等于或大于10(即10-mer或更长),进一步优选地等于或大于15(即15-mer或更长)。每个极小元可以是长度小于50、任选地小于30、进一步任选地小于25的k-mer。当使用公共极小函数min(•)来确定更长的极小元时,确定的极小元更可能代表序列的特定部分,即极小元不太可能出现在序列的多个单独且不相关部分中。设置极小元的尺寸的上限降低了极小元含有测序错误的风险。

[0050] 应用公共极小函数min()的步骤S220可包括识别相应的突变的序列读段ρ^i中的一个或多个k-mer,该一个或多个k-mer在可能的k-mer的有序列表中首先列出。为相应的突变的序列读段ρ^i确定的一个或多个极小元可以是所识别的一个或多个k-mer。可能的k-mer的有序列表可以预先确定的顺序包括所有或一些可能的k-mer。步骤S220可包括生成可能的k-mer的有序列表,或者可不包括生成可能的k-mer的有序列表(例如,在不需要与列表直接比较来确定极小元的情况下,如下面的一些示例)。

[0051] 例如,公共极小函数min()可确定作为极小元的k-mer,该k-mer具有突变的序列读段ρ^i中所有k-mer的两位二进制编码的整数最小值。换句话讲,公共极小函数min()可识别首先在k-mer列表中列出的k-mer,这些k-mer按照它们的两位二进制编码的整数值进行排序。例如,基于二进制编码A:00、G:01、C:10和T:11,公共极小函数可识别突变的序列读段中的5-mer,该5-mer首先在示例性有序列表AAAAA、AAAAG、AAAAC、AAAAT、AAAGA、AAAGG、…、CTTTC、CTTTT和TTTTT中列出。例如,示例性突变的序列读段:

[0052] ACGGAAAGCGCTACGAGCGACTGATATTGAGCTACGTTCAGAGCC

[0053] 包含5-mer ACGGA、CGGAA、GGAAA、…、AGAGC和GAGCC。在上述示例性有序列表中首先列出5mer AAAGC,并且公共极小函数min()将识别AAAGC为该示例性突变的序列读段的极小元。应当理解,对于该公共极小函数min(),实际上不必生成可能的k-mer的有序列表以确定k-mer集合的极小元。

[0054] 确定突变的序列读段ρ^i中所有k-mer的两位二进制编码的整数最小值只是可应用于突变的序列读段ρ^i以确定极小元的公共极小函数min()的一个示例。可使用任何其他公共极小函数min()。例如,将整数最小值函数的排序随机化有利于公共极小函数min()。实现此类随机化的一种方式是首先将具有任意位向量的按位逻辑XOR应用于突变的序列读段ρ^i所包含的每个k-mer,之后可使用整数最小值函数。

[0055] 另选地,可使用可能的k-mer的预先确定集合代替可能的k-mer的有序列表,并且应用公共极小函数min()包括识别存在于可能的k-mer的预先确定集合中的一个或多个k-mer。为相应的突变的序列读段ρ^i确定的一个或多个极小元可以是所识别的一个或多个k-

mer。可能的k-mer的预先确定集合可以是有序的或无序的。可能的k-mer的预先确定集合可以是仅包括合适的或旨在用作极小元的k-mer的k-mer集合。应用公共极小函数min()的步骤S220可包括创建可能的k-mer的预先确定集合。

[0056] 在优选的实施方案中，在可能的k-mer的有序列表中，基于k-mer存在于包含突变的序列μ^i中并且不存在于不包含突变的序列S^i中的概率，对k-mer进行排序，即在包含突变的序列中而非不包含突变的序列中相对可能存在的k-mer在有序的列表中较早被列出，在包含突变的序列中而非不包含突变的序列中相对不可能存在的k-mer在有序的列表中则较晚被列出。在另选的优选的实施方案中，可能的k-mer的预先确定集合包含在包含突变的序列中而非不包含突变的序列中相对可能存在的k-mer，并且任选地不包含在包含突变的序列中而非不包含突变的序列中相对不可能存在的k-mer。步骤S220可包括确定该多个突变的序列读段P所包含的哪些k-mer在包含突变的序列中而非不包含突变的序列中相对可能存在，例如通过比较该多个突变的序列读段P中k-mer的出现(或观察)次数与该多个非突变的序列读段R中k-mer的出现次数来确定。这可包括对该多个突变的序列读段P中k-mer的出现次数计数，并且对该多个非突变的序列读段R中k-mer的出现次数计数。

[0057] 在两个优选的实施方案中，选择公共极小函数min()以便优先确定作为一个或多个极小元的k-mer，这些k-mer在突变的序列读段ρ^i中比在非突变的序列读段r^i中更可能出现。这增加了每个极小元包含突变的可能性。

[0058] 在更优选的实施方案中，可能的k-mer的有序列表仅含有(即组成有)在多个突变的序列读段P中比在多个非突变的序列读段R中更经常存在(或者在包含突变的序列中比在不包含突变的序列中更经常存在)的k-mer，即在该多个突变的序列读段P中的出现次数大于在该多个非突变的序列读段R中的出现次数的k-mer。在另选的更优选的实施方案中，可能的k-mer的预先确定集合仅含有(即组成有)在多个突变的序列读段P中比在多个非突变的序列读段R中更经常存在(或者在包含突变的序列中比在不包含突变的序列中更经常存在)的k-mer，即在该多个突变的序列读段P中的出现次数大于在该多个非突变的序列读段R中的出现次数的k-mer。优选地，可能的k-mer的有序列表或可能的k-mer的预先确定集合仅含有(即组成有)在多个突变的序列读段中存在n次或更多次并且在多个非突变的序列读段中存在少于n次的k-mer，即在多个突变的序列读段P中的出现次数等于或大于n并且在多个非突变的序列读段R中的出现次数小于n的k-mer。n可以为大于或等于1的整数。n可以为大于或等于2的整数。优选地，n为2。进一步优选地，可能的k-mer的有序列表或可能的k-mer的预先确定集合仅含有(即组成有)在多个非突变的序列读段中不存在的k-mer，即在多个非突变的序列读段R中的出现次数为0的k-mer。

[0059] 例如，可能的k-mer的有序列表或可能的k-mer的预先确定集合可仅含有在多个突变的序列读段P的k-mer集合中存在至少两次，但是在多个非突变的序列读段R的k-mer集合中从不存在(或很少存在)的k-mer。这以高概率确保了可能的k-mer的有序列表或可能的k-mer的预先确定集合包括含有存在于多个突变的序列读段P中的两个或更多个突变的序列读段的突变的极小元。任选地，在多个突变的序列读段中比在多个非突变的序列读段中更经常存在的k-mer相对可能存在于包含突变的序列中。任选地，其中在多个序列读段中存在n次或更多次并且在多个非突变的序列读段中存在少于n次的k-mer相对可能存在于包含突变的序列中。

[0060] 该可能的k-mer的预先确定集合可通过建立突变极小元集合U_M来创建，其中U_M包含k-mer，优选地所有或基本上所有的k-mer，对于这些k-mer，在多个突变的序列读段P中的出现或观察次数大于或等于n(优选地，其中n≥2，更优选地，其中n为2)，并且在多个非突变的序列读段P中的出现或观察次数小于n(优选地，其中n为0或1，更优选地，其中n为0)。该突变极小元集合U_M可通过对每个k-mer在多个非突变的序列读段R中和在多个突变的序列读段P中存在的频率计数来创建。该突变极小元集合U_M可使用概率数据结构，诸如计数布隆过滤器或相关布谷鸟过滤器和商过滤器方法从多个非突变的序列读段R和多个突变的序列读段P进行有效地计算。该可能的k-mer的有序列表可从整个突变极小元集合U_M创建。

[0061] 该突变极小元集合U_M可用作可能的k-mer的预先确定集合。另选地，可进一步处理该突变极小元集合U_M以创建可能的k-mer的预先确定集合。在优选的实施方案中，将突变极小元集合U_M的子集W_M用作可能的k-mer的预先确定集合。该子集W_M可通过将每个突变读段$\rho^i \in P$分成两个或更多个非重叠区段(任选地具有基本上相等的尺寸)，例如尺寸为L_w的非重叠k-mer起始位置集合，例如$\{1\cdots L_w\}$、$\{L_w+1\cdots 2L_w\}$等。当使用长度为150的突变的序列读段时，L_w的典型值可能为50，从而将可能的k-mer起始位置分成3组。然后对于起始位置的每个集合，子集W_M可如下表示：

[0062] $$W_M = \bigcup_{\rho \in P} \bigcup_{j=1\ldots\lceil \mathcal{L}/\mathcal{L}_w \rceil} min(k(\rho_{(j-1)\mathcal{L}_w+1} \cdots \rho_{j\mathcal{L}_w}) \cap U_M)$$

[0063] 实际上，多个突变的序列读段P中的每个突变的序列读段可分成两个或更多个区段(例如，3个区段)，并且可找到极小元来表示每个区段。通过如下步骤确定极小元：首先经由在相应的突变的序列读段的那个部分中的k-mer与突变极小元集合U_M的交集找到候选极小元，然后将公共极小函数应用于该集合以识别每个区段的一个极小元。

[0064] 因此，在优选的实施方案中，将公共极小函数min()应用于每个突变的序列读段的步骤S220包括：

[0065] 创建突变极小元集合U_M，该突变极小元集合由在多个突变的序列读段P中的k-mer(优选地所有或基本上所有的k-mer)组成，这些k-mer在多个突变的序列读段P中存在n次或更多次，并且在多个非突变的序列读段R中存在小于n次，其中n为大于或等于2的整数；

[0066] 任选地通过以下方式来创建突变极小元集合U_M的子集W_M：将多个突变的序列读段P中的每个突变的序列读段分成两个或更多个区段，识别在多个突变的序列读段P中的每个突变的序列读段的每个区段中的k-mer(优选地所有或基本上所有的k-mer)，这些k-mer存在于突变极小元集合U_M中，以及向子集W_M添加为多个突变的序列读段P中的每个突变的序列读段的每个区段识别的k-mer中的一个k-mer，任选地，其中通过将公共极小函数min()(例如，找到整数最小值或任何其他已知的极小函数)应用于为多个突变的序列读段P中的每个突变的序列读段的每个区段识别的k-mer，来选择为多个突变的序列读段P中的每个突变的序列读段的每个区段识别的k-mer中的一个k-mer；以及

[0067] 使用该突变极小元集合U_M或该突变极小元集合U_M的子集(诸如子集W_M)作为可能的k-mer的预先确定集合，并且，对于多个突变的序列读段P中的每个突变的序列读段，在相应的突变的序列读段μ^1中识别k-mer(优选地所有或基本上所有的k-mer)，这些k-mer存在于可能的k-mer的预先确定集合中，其中为相应的突变的序列读段确定的一个或多个极小元是所识别的k-mer。

[0068] 方法200还包括确定每个突变的序列读段ρ^i中的一个或多个相应的极小元的位置j的步骤S222。每个相应的突变的序列读段ρ^i中的每个极小元的位置j可与相应的极小元相关联地(例如,在与之相同的位置或极小元箱中)存储为整数位值。

[0069] 步骤S224:极小元分箱

[0070] 在优选的实施方案中,方法200包括在一个或多个极小元箱中将突变的序列读段P分箱的步骤S224。在一个或多个极小元箱中将突变的序列读段P分箱包括在一个或多个极小元箱中将突变的序列读段ρ^i的索引i特征分箱。每个极小元箱可含有具有公共极小元的突变的序列读段P,并且不含有不具有公共极小元的突变的序列读段P。仅对相同的极小元箱中的突变的序列读段P执行对具有匹配位置和/或具有错配位置的突变的数量计数的步骤S240。这提高了执行步骤S240的计算效率。

[0071] 换句话讲,一个或多个极小元可用作哈希键,以在公共哈希桶(本文称为极小元箱)中收集含有极小元的序列读段,例如准备对这些序列读段进行一些额外处理(例如步骤S240)。

[0072] 通过将公共极小函数min()应用于突变的序列读段P来确定的每个极小元可用于在该一个或多个极小元箱中将突变的序列读段P分箱的目的。在一个实施方案中,可能的k-mer的有序列表中的每个极小元,或可能的k-mer的预先确定集合中的每个极小元(例如,该突变极小元集合U_M中的每个极小元,或其子集,诸如子集W_M)可用于在该一个或多个极小元箱中将突变的序列读段P分箱的目的。

[0073] 在一个或多个极小元箱中将该突变的序列读段P分箱的步骤S224可包括创建一个或多个极小元箱。这可包括通过公共极小函数min()为每个已确定的极小元创建一个极小元箱,或者为在可能的k-mer的预先确定集合U_M中的每个极小元(或k-mer)创建一个极小元箱,或者为在子集W_M中的每个k-mer创建一个极小元箱。每个极小元箱可实现为连续的RAM嵌段。优选地,极小元箱的集合被实现为计算机存储介质(诸如计算机磁盘,例如,旋转磁盘或固态盘)上的文件,从而允许每个箱存储大量数据(在序列分析适当时)。

[0074] 在一个或多个极小元箱中将突变的序列读段P分箱的步骤S224可包括在相应的极小元箱中存储突变的序列读段ρ^i,或突变的序列读段ρ^i的索引i特征。在每个突变的序列读段ρ^i中确定一个或多个相应的极小元的位置j的步骤S222可包括将相应的极小元的位置j存储在相应的极小元箱中。此外,如通过确定每个突变的序列读段中的一个或多个突变的位置α的步骤S230所确定的,可将每个突变的序列读段ρ^i中的一个或多个突变的位置α=morphomuts(ρ^i,V_R)存储在相应的极小元箱中。任选地,如果任意附加值,诸如突变的序列读段ρ^i的序列、关于该序列的准确度的质量信息或其他信息用于下游处理,则可将这些信息存储在极小元箱中。与每个突变的序列读段ρ^i相关联的这些值可作为元组存储在每个极小元箱中。出于符号化目的,将第z个极小元箱的第y个元素$b_{z,y}$的元组元素表示为$b_{z,y}$.i、$b_{z,y}$.j和$b_{z,y}$.α。可将每个突变的序列读段ρ^i添加到多个极小元箱。

[0075] 步骤S230:突变的位置

[0076] 方法200包括确定每个突变的序列读段ρ^i中一个或多个突变的位置α的步骤S230。确定每个突变的序列读段ρ^i中的一个或多个突变的位置α的步骤S230可使用无比对方法来执行。

[0077] 确定每个突变的序列读段ρ^i中的一个或多个突变的位置α的步骤S230可包括获得

非突变的种子掩蔽的k-mer集合V_R，即非突变的序列读段R的k-mer集合，已对该集合应用了一个或多个种子模式ψ。获得非突变的种子掩蔽的k-mer集合V_R可包括创建或生成该非突变的种子掩蔽的k-mer集合V_R。可通过将一个或多个种子模式中的每个种子模式应用于不包含突变的序列中的每个k-mer，例如非突变的序列读段中的每个k-mer，来获得或创建该非突变的种子掩蔽的k-mer集合V_R。将种子模式应用于k-mer可包括确定种子模式和(两位编码的)k-mer的按位逻辑AND。将种子模式应用于k-mer创建了种子掩蔽的k-mer。非突变的种子掩蔽的k-mer集合V_R可表示为

[0078] $$V_R = \{\forall_{r(i)\in R} \forall_{j=1\ldots k} \forall_{\psi\in\Psi} : \psi(k(r_j^i))\},$$

[0079] 即，对于多个非突变的序列读段R中的所有(或基本上所有)的非突变读段r^i，进而对于每个非突变读段r^i中的k-mer的所有(或基本上所有)的位置j(即，从1至-k)，通过将种子家族Ψ的一个或多个种子模式ψ中的每个种子模式应用于每个k-mer $k(r_j^i)$，创建非突变的种子掩蔽的k-mer集合V_R。

[0080] 种子模式可用于修改k-mer彼此比较的过程。种子模式被定义为两个k-mer内的位置(即核苷酸)集合，在这两个k-mer中的位置需要相同，以便认为种子掩蔽的k-mer是匹配的。种子模式可包括掩蔽位置和非掩蔽位置。将种子模式应用于k-mer创建了种子掩蔽的k-mer，其中在任何进一步处理(诸如比较)中，忽略与对应的种子模式的掩蔽位置对应的种子掩蔽的k-mer的位置，其中在任何进一步处理(诸如比较)中不忽略与对应的种子模式的非掩蔽位置对应的种子掩蔽的k-mer的位置。例如，种子模式{1,2,4,6,7}需要在比较下相同的两个k-mer $k(S_i)$和$k(S_j)$中的第一、第二、第四、第六和第七位置(或核苷酸)，以便将它们视为匹配的(对于k＝7)。两个k-mer中的第三和第五位置可以是任意核苷酸。这意味着两个种子掩蔽的k-mer中的第三和第五位置被种子模式掩蔽。

[0081] 任选地，一个或多个种子模式可以是一个或多个转换种子模式。特别是当与不包含突变的序列S相比，包含突变的序列M包含突变时，即多个突变的序列读段P中的每个突变的序列读段含有一个或多个转换突变时，这是有利的。

[0082] 转换种子模式是一种特殊类型的种子模式，其中位置分为三类而不仅仅是两类：每个位置可以(1)需要精确匹配，或者(2)都是嘌呤或嘧啶，或者(3)四种核苷酸中的任一种核苷酸，以便匹配。当包含突变的序列包括转换突变时，转换种子模式是特别有利的。当使用上文引入的两位核苷酸编码在计算机上实现时，需要精确匹配的位置可被实现为位掩码11，而只允许转换突变的位置被实现为10，允许任何核苷酸的位置被实现为00。种子模式{1,2,4,6,7}可被写成位掩码11110011001111。转换种子模式{1,2,4,6,7}可被写成位掩码11111011101111。可通过分别计算k-mer的位掩码和两位编码的按位逻辑AND来评估两个k-mer是否匹配，然后检查两个所得的种子掩蔽的k-mer的同一性。为方便起见，经由按位逻辑AND将种子模式应用于k-mer $k(S_i)$的函数将表示为函数ψ(k(S_i))。

[0083] 在实施方案中，选择一个或多个种子模式，使得通过将一个或多个种子模式的至少一个种子模式应用于多个突变的序列读段P(或包含突变的序列)中的任何k-mer，并且应用于多个非突变的序列读段R(或不包含突变的序列)中对应的k-mer，获得相同的种子掩蔽的k-mer的概率大于90％，优选地大于95％，进一步优选地大于98％，最优选地大于99％。

[0084] 一个或多个种子模式可构成种子家族Ψ。种子家族Ψ是两个或更多个种子模式的集合，当一起使用时，这些种子模式能够以高概率，例如大于90％，优选地大于95％，进一步

优选地大于98％，最优选地大于99％的概率在特定百分比的核酸同一性处识别k-mer之间的匹配。种子家族Ψ表示为n个不同函数集合以应用种子模式$\psi_1 \cdots \psi_n \in \Psi$。种子模式的权重w(ψ)被定义为种子需要相同的位置数，以便两个k-mer被认为是匹配的，其中w(ψ)≤k。

[0085] 确定每个突变的序列读段ρ^i中的一个或多个突变的位置α的步骤S230可包括：对于每个突变的序列读段，将一个或多个种子模式ψ_i中的每一个种子模式应用于相应的突变的序列读段ρ^i中的k-mer(任选地每个k-mer)以获得多个突变的种子掩蔽的k-mer。一个或多个突变的位置可通过识别突变的序列读段ρ^i中被所有的种子模式掩蔽的一个或多个位置来确定，这些种子模式对应于存在于非突变的种子掩蔽的k-mer集合V_R中的多个突变的种子掩蔽的k-mer中的突变的种子掩蔽的k-mer。这意味着在突变的序列读段ρ^i中不是突变的位置可被识别为突变的序列读段ρ^i中未被任何一个种子模式掩蔽的一个或多个位置，这些种子模式对应于存在于非突变的种子掩蔽的k-mer集合V_R中的多个突变的种子掩蔽的k-mer中的突变的种子掩蔽的k-mer。

[0086] 例如，每个突变的序列读段ρ^i的一个或多个突变的位置α可通过以下步骤确定：

[0087] • 创建长度为2L的位向量α并将位向量α初始化为0

[0088] • 创建长度为2k的位向量b，将位向量b初始化为仅1

[0089] • 对于每个$\psi \in \Psi$，并且对于在1与-k之间的读段中的每个位置j，计算$\psi(k(\rho_j^i))$。如果$\psi(k(\rho_j^i)) \in V_R$，则分配$\alpha \leftarrow \alpha | (\psi(b) >> 2j)$，其中运算符“|”表示按位逻辑OR运算符，并且运算符“>>”表示位右移运算符。V_R中集合成员的查询既可完全使用哈希表之类的东西来实现，也可近似地使用高效的概率数据结构(如布隆过滤器、商过滤器或类似方法)来实现。

[0090] • 任选地，为了便于进一步处理，通过去除奇数位置，例如$\alpha \leftarrow \{\alpha 2, \alpha 4, \alpha 6, \ldots \alpha 2L\}$，将位向量α从二进制两位突变的序列读段编码的长度转换为突变的序列读段本身的长度。

[0091] • 任选地，为了便于进一步处理，应用逻辑NOT来翻转位，使得1表示从未找到种子匹配的位置。

[0092] 上述程序的结果将是位向量α，其中含有1的每个位置对应于具有高概率的突变的位置。出于符号化目的，计算突变的序列读段ρ^i的位向量α的函数表示为$\alpha = \text{morphomuts}(\rho^i, V_R)$。

[0093] 图3示出了实施例，该实施例说明了如何使用单个种子模式ψ＝1110110011为示例性突变的序列读段ρ＝ACGCAAAGCGCTACGAGCGACTGATATT获得位向量α。突变的序列读段ρ的第4、第8、第11、第12和第16位置对应于突变的序列读段ρ中的突变，即在不包含突变的序列中，这些位置中的核苷酸将不同。在实施过程中，突变的序列读段ρ可以两位二进制格式编码，并且种子模式ψ的每个位置可覆盖两个位(即，种子模式ψ中的每个1将实现为两个二进制1，并且种子模式ψ中的每个0将实现为二进制00或二进制10)。先前在该实施例中生成了种子掩蔽的非突变的k-mer集合V_R。

[0094] 如图3的实施例所示，将种子模式ψ应用于突变的序列读段ρ中的每个k-mer，从而为突变的序列读段ρ中的每个k-mer创建一个种子掩蔽的k-mer。然后检查种子掩蔽的k-mer是否存在于非突变的种子掩蔽的k-mer集合V_R中。在所示的实施例中，第1、第5、第13、第17、第18和第19种子掩蔽的k-mer全部存在于非突变的种子掩蔽的k-mer集合V_R中。这些种子掩蔽的k-mer不含有未被种子模式掩蔽的突变位置。

[0095] 然后使用第1、第5、第13、第17、第18和第19种子掩蔽的k-mer来识别被对应于这些种子掩蔽的k-mer的所有种子模式掩蔽的位置。突变的序列读段ρ的第4位置被所有这些种子模式掩蔽，注意当处理第13、第17、第18和第19种子掩蔽的k-mer时，忽略突变的序列读段ρ的第4位置，即突变的序列读段ρ的第4位置被第13、第17、第18和第19种子掩蔽的k-mer的种子模式掩蔽。这些种子模式都没有掩蔽突变的序列读段ρ的第4位置。因此突变的序列读段ρ的第4位置被识别为突变的位置。相比之下，虽然突变的序列读段ρ的第7位置被对应于第1、第13、第17、第18和第19种子掩蔽的k-mer的所有种子模式掩蔽，但突变的序列读段ρ的第7位置未被对应于第5种子掩蔽的k-mer的种子模式掩蔽。因此，突变的序列读段ρ的第7位置未被识别为突变的位置。相反，突变的序列读段ρ的第7位置被识别为不是突变的位置。

[0096] 实际上，对应于第1、第5、第13、第17、第18和第19种子掩蔽的k-mer的所有种子模式都使用逻辑OR进行组合。可翻转(例如，使用逻辑NOT操作)所得位向量的位以获得突变的序列读段ρ中突变的位置作为位向量α。

[0097] 使用参考组装的步骤230的另选具体实施

[0098] 在上述实施方案中，基于将种子模式应用于每个突变的序列读段ρ^i，确定每个突变的序列读段ρ^i中的一个或多个突变的位置α的步骤230使用多个突变的序列读段P和多个非突变的序列读段R来执行。

[0099] 在大型和复杂基因组中，诸如人类基因组，基因组的大量部分由重复序列构成。例如，超过一半的人类基因组被认为是重复序列的一部分。这些重复序列被归类到类似重复序列的“家族”中。人类基因组中最常见的家族是SINE(短散布核元件)的Alu家族，这些Alu约300nt长，并且存在约100万个拷贝。另一个常见的家族是长散布核元件(LINE)的L1家族，其中元件的尺寸范围为1kbp至6.5kbp，并且拷贝数为10,000。

[0100] 基因组内重复序列的各种拷贝可能是不相同的，例如含有单个碱基差异。由于突变的生物学，这些差异通常为转换差异，在一些情况下，这些差异可能看起来类似于由于在多个突变的序列读段P和多个非突变的序列读段R之间引入突变而导致的差异。这对于某些用于诱变的基于聚合酶的方法尤其如此，这些方法用于将突变引入至少一个靶模板核酸分子中，作为产生多个突变的序列读段P的一部分，因为这些方法通常会引入转换突变。

[0101] 结果，多个非突变的序列读段R可含有大量的k-mer，这些k-mer彼此仅通过一些数量的转换差异而不同。因此，尽管与非突变的序列读段R相比存在突变，但多个突变的序列读段P可包括与多个非突变的序列读段R中的k-mer相同的一个或多个k-mer。在一些情况下，可能的是，在不同非突变的序列读段r^i中重复序列的不同拷贝之间的这些自然出现的差异会部分“掩蔽”在多个突变的序列读段P中引入的突变。这在SINE的Alu家族中尤为明显。

[0102] 因此，如果在诸如此类的情况下，提供可以更好地区分有意引入的突变与重复序列拷贝之间的自然差异的方法的实施方案，那将是有利的。

[0103] 提高该方法区分有意引入的突变与重复序列拷贝之间的自然差异的能力的第一方法是使用具有更高权重的种子模式，以便突变的种子掩蔽的k-mer更可能包括含有区分重复序列的拷贝的差异的一个或多个位置。在采用第一方法的实施方案中，每个种子模式ψ的权重w(ψ)在50至100的范围内，优选地在70至90的范围内。对于人类基因组，约80的权重将足以用于第一方法。

[0104] 然而，第一方法可能并非在所有情况下都是理想的。具有权重80的种子模式将是非常长的，可能比突变的序列读段ρ^i的典型长度更长。此外，确保高灵敏度所需的种子家族Ψ的尺寸可能变得非常大，因此需要大量额外的计算资源来处理所有种子模式。最后，种子模式覆盖插入和缺失错误的概率会增加，并且需要额外的算法复杂性来适应插入和缺失错误的可能性。因此，在一些情况下，该第一方法可能不是优选的。

[0105] 提高该方法区分有意引入的突变与重复序列拷贝之间的自然差异的能力的第二方法是使用基于将多个突变的序列读段P与参考组装(或参考基因组)比对(或映射)的方法。参考组装可以独立生成，诸如由基因组参考联盟(GRC)生产的hg38人类基因组，或者基于多个非突变的序列读段R的从头组装。在第二方法中，确定每个突变的序列读段中的一个或多个突变的位置的步骤包括对于一个或多个突变的序列读段，将相应的突变的序列读段与参考组装比对。

[0106] 当突变的序列读段ρ^i是双端序列读段时，该方法可能特别适用。将双端突变的序列读段与参考组装(特别相对于SINE重复序列)比对的优势是短读段鸟枪库的片段长度通常大于重复序列的长度。双端测序的典型片段尺寸为400bp至600bp，其中约150bp从片段的每个末端测序。因此，如果一对双端序列读段中的一个双端序列读段落在重复序列中，那么这对双端序列读段中的另一个双端序列读段可能落在重复序列之外的独特序列中。因此，标准双端比对程序(例如，Burrows-Wheeler比对器，诸如BWA-MEM)能够将这对包括重复序列的正确拷贝的双端序列读段与参考组装中的正确位置确信地比对。然后可以记录比对的突变的序列读段ρ^i和参考组装之间的任何差异的位置，并且将这些位置存储在位向量α中，类似于使用基于将种子模式应用于每个突变的序列读段ρ^i的方法得出的位向量。因此，通过识别相应的突变的序列读段与参考组装之间的差异在相应的突变的序列读段中的位置实现了确定相应的突变的序列读段中一个或多个突变的位置。

[0107] 然而，在一些情况下，将多个突变的序列读段P与参考组装比对可能不是理想的，因为任何给定靶模板核酸分子通常将具有在参考组装中未表示的区域。因此，不可能将突变的序列读段ρ^i与在参考组装中未表示的那些区域比对，并且不可能从比对的突变的序列读段ρ^i和参考组装之间的差异导出位向量α。此外，在参考组装中未表示的区域通常具有临床意义，因为它们表示相对于参考组装的结构变体插入。除了大的插入区域之外，基于将多个突变的序列读段P与参考组装比对的方法也将遗漏出现在相对于参考组装的小插入中的任何突变。

[0108] 因此，提高该方法区分有意引入的突变与重复序列拷贝之间的自然差异的能力的第三混合方法是将基于将多个突变的序列读段P与参考组装比对的方法与基于将种子模式应用于每个突变的序列读段ρ^i的方法相结合。该第三方法可用作本发明方法的步骤230的另选具体实施。

[0109] 在第三方法中，使用基于将多个突变的序列读段P与参考组装比对的方法和基于将种子模式应用于每个突变的序列读段ρ^i的方法来确定每个突变的序列读段中一个或多个突变的位置。若将相应的突变的序列读段中的位置与参考组装比对，如果相应的突变的序列读段中的位置是相应的突变的序列读段不同于参考组装的位置，则将相应的突变的序列读段中的位置确定为相应的突变的序列读段中突变的位置。若没有将相应的突变的序列读段中的位置与参考组装比对，如果相应的突变的序列读段中的位置是被所有的种子模式

掩蔽的位置，这些种子模式对应于存在于非突变的种子掩蔽的k-mer集合中的多个突变的种子掩蔽的k-mer中的突变的种子掩蔽的k-mer，则将相应的突变的序列读段中的位置确定为相应的突变的序列读段中突变的位置。

[0110] 为了实现这一点，通过基于比对的方法和基于应用种子模式的方法独立地导出上述类型的位向量α。来自基于将种子模式应用于每个突变的序列读段ρ^i的方法的位向量表示为α_{mmd}并且来自基于将多个突变的序列读段P与参考组装比对的方法的位向量表示为α_{map}。还构建了表示为α_{amask}的另外的比对掩蔽位向量，该位向量记录了与参考组装成功比对的每个突变的序列读段的位置。比对掩蔽位向量α_{amask}将在与参考组装成功比对的每个位置中具有1，并且在未成功与参考组装比对的位置中具有0。

[0111] 然后构建了最终混合位向量α_{hybrid}，该混合位向量结合了来自基于将种子模式应用于每个突变的序列读段ρ^i的方法的位向量α_{mmd}和来自基于将多个突变的序列读段P与参考组装比对的方法的位向量α_{map}，如下所示：

[0112] $$\alpha_hybrid = \alpha_map \mid (\alpha_mmd \,\&\, \sim\alpha_amask)$$

[0113] 其中“|”表示按位逻辑OR运算符，“&”表示按位逻辑AND，而“～”表示按位NOT。

[0114] 因此，第三方法使用通过在突变的序列读段的位置(成功比对)与参考组装比对而确定的突变位置，以及通过在所有其他位置应用种子模式而确定的突变位置。这具有使高质量参考组装能够并入到分析中的优势，同时还可处理与参考组装相关的所有类型的插入。针对独立的高质量参考组装，诸如人类参考基因组的比对可比针对从头短读段组装的比对准确得多。使用通过与参考组装比对确定的突变位置可提供对突变位置的更准确估计，尤其是在重复序列区域中，而基于无比对种子模式的方法可识别参考中未表示的区域中的突变位置。后者可在不必计算组装的情况下发生，这是计算要求的任务。因此，相对于单独使用任一方法，混合方法在识别突变位置的准确度和计算效率方面提供了改进。

[0115] 还可用来自特定靶模板核酸分子的变体和局部组装的区域“扩充”参考组装，以产生对该靶模板核酸分子特异的组装图。来自基于将多个突变的序列读段P与参考组装比对的方法的位向量(表示为α_{map})可通过将突变的序列读段与扩充的组装图比对得到，然后由于技术或其他原因，对于仍然难以比对的靶模板核酸分子的任何区域，结合基于将种子模式应用于每个突变的序列读段ρ^i方法。

[0116] 步骤S240和S240：确定与两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度

[0117] 方法200包括步骤S240：对于具有公共极小元的至少两个突变的序列读段，当比对相应的极小元时，对具有匹配位置和/或具有错配位置的突变的数量计数。

[0118] 这可通过首先确定在步骤S222中为两个突变的序列读段中的每个突变的序列读段确定的极小元的位置j中的差异来实现。例如，在极小元箱b_z中存储为a＝$b_{z,y}$.i和c＝$b_{z,x}$.i的两个突变的序列读段ρ^a和ρ^c中的每个突变的序列读段的极小元的位置j中的差异可被确定为d＝$b_{z,y}$.j-$b_{z,x}$.j。

[0119] 对具有匹配位置的突变的数量计数可包括当为两个突变的序列读段中的一个突变的序列读段确定的突变的位置右移d时，确定在步骤S230中确定的突变位置的集合交集的尺寸。例如，对于存储为$b_{z,y}$和$b_{z,x}$的两个突变的序列读段ρ^x和ρ^y，具有匹配位置的突变的

数量可被确定为：

[0120] $\lambda_{x,y}=|\Omega(b_{z,x}.\alpha)\cap(\Omega(b_{z,y}.\alpha)-d)|$

[0121] 其中Ω(α)被定义为α中非零的位置索引集合(即相应的突变的序列读段ρ^i中的突变的位置集合)，并且其中$\Omega(b_{z,y}.\alpha)-d$被理解为从$\Omega(b_{z,y}.\alpha)$减去d的元素级减法。使用位偏移和弹出计数CPU指令，可在计算机上有效地实现集合交集。

[0122] 对具有错配位置的突变的数量计数可包括当为两个突变的序列读段中的一个突变的序列读段确定的突变的位置右移d时，确定在步骤S230中确定的突变位置的对称集差的尺寸。例如，对于存储为$b_{z,y}$和$b_{z,x}$的两个突变的序列读段ρ^x和ρ^y，具有错配位置的突变的数量可被确定为：

[0123] $\delta_{x,y}=|(\Omega(b_{z,x}.\alpha)\setminus(\Omega(b_{z,y}.\alpha)-d))\cup((\Omega(b_{z,y}.\alpha)-d)\setminus\Omega(b_{z,x}.\alpha))|$。

[0124] 确定与该至少两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度的步骤S242可基于具有匹配位置$\lambda_{x,y}$和/或具有错配位置$\delta_{x,y}$的突变的数量。在实施方案中，与该至少两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度对应于具有匹配位置$\lambda_{x,y}$的突变的数量。具有匹配位置$\lambda_{x,y}$的突变的数量越高，该至少两个突变的序列读段衍生自包含突变的相同序列的概率也越高。在另选的实施方案中，与该至少两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度对应于具有错配位置$\delta_{x,y}$的突变的数量。具有错配位置$\delta_{x,y}$的突变的数量越低，该至少两个突变的序列读段衍生自包含突变的相同序列的概率也越高。

[0125] 在优选的实施方案中，与该至少两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度为以下两者中的一者：i)该至少两个突变的序列读段衍生自包含突变的相同序列的概率密度，和ii)与该至少两个突变的序列读段衍生自包含突变的相同序列的概率密度相关的分数函数。

[0126] 例如，具有匹配位置$\lambda_{x,y}$的突变的数量和具有错配位置$\delta_{x,y}$的突变的数量可用于计算在两个读段衍生自包含突变M的相同序列的模型下的概率密度，或与此类概率密度相关的分数函数。一个此类分数函数可以为$\omega_{a,c}=\Delta(\lambda_{x,y})-\Delta(\delta_{x,y})$，其中对于$a=b_{z,x}.i$并且$c=b_{z,y}.i$，$\Delta(n)=(0.5n)(n+1)$。以此方式，$\omega_{a,c}$表示两个突变的序列读段$\rho^a$和$\rho^c$之间的链路的分数或权重。可以在相应极小元箱$b_z$中为所有对突变的序列读段$\rho^i$产生此类链路的集合，或者如果在极小元箱$b_z$中有大量的条目，则可在极小元箱$b_z$中，将链路权重的计算或报告二次抽样到随机选择的条目对。

[0127] 步骤S300：序列组装或序列重建

[0128] 方法10还可包括步骤S300，该步骤组装或重建序列或序列的至少一部分，例如包含突变的序列或不包含突变的序列。组装或重建的序列可以是包含突变的序列或不包含突变的序列。

[0129] 方法200，例如重建或组装序列的步骤S300可包括从多个突变的序列读段创建无向加权图。无向加权图包括对应于多个突变的序列读段的节点。例如，每个节点可对应于相应的突变的序列读段，因为每个节点由相应的突变的序列读段的读段索引i表示，或者由相应的突变的序列读段的序列表示。节点之间的边缘与相应的边缘权重相关联，其中每个边缘权重可基于具有匹配位置和/或具有错配位置的突变的数量来确定，该突变的数量针对对应于与相应的边缘相关联的两个节点的两个突变的序列读段来确定。每个边缘权重可对

应于与至少两个突变的序列读段(即,两个突变的序列读段对应于与边缘相关联的节点,这些边缘与边缘权重相关联)衍生自包含突变的相同序列的概率相关的量度。因此,连接两个突变的序列读段(节点)的边缘权重表示那些两个突变的序列读段衍生自包含突变的相同序列的概率,或与该概率相关的一些其他任意函数。

[0130] 可通过顺序地或并行地处理每个极小元箱来构建无向加权图,从而计算在每个极小元箱中突变的序列读段之间的边缘。边缘权重可以是分数函数$\omega_{a,c}$。

[0131] 然后包括边缘权重$\omega_{a,c}$的无向加权图可用于SAM数据(例如,突变的序列读段)的处理,例如将任何已知或未知的技术用于使用此类无向加权图来组装序列。从无向加权图组装序列可包括,例如产生突变的序列读段的簇,并且组装每个簇中的突变的序列读段以重建对应于包含突变的序列的至少一部分的模板。

[0132] 例如,重建或组装序列的至少一部分的方法200或步骤S300可包括对无向加权图执行图聚类,从而创建预期衍生自包含突变的相同序列的突变的序列读段的簇。可使用任何标准的基于流的图聚类算法(诸如马尔可夫聚类(MCL)或Infomap)来执行图聚类。另选地,可在一些最小权重阈值上过滤无向加权图的边缘,并且然后可将图的连接分量用于表示衍生自包含突变的相同序列的突变的序列读段。

[0133] 重建或组装序列的至少一部分的步骤S300还可包括通过组装簇中的突变的序列读段来重建包含突变的序列的至少一部分。例如,可对簇中的突变的序列读段进行标准从头组装方法,以重建包含突变的序列。此类从头组装方法包括例如"IDBA-UD:用于具有高度不均匀深度的单细胞和宏基因组测序数据的从头组装器"的IDBA-UD算法,Peng Y等人,《生物信息学》,2012年6月1日,第28卷(第11期):第1420-1428页,doi:10.1093/bioinformatics/bts174(Epub电子档出版于2012年4月11日);或"SPAdes:一种新的基因组组装算法及其在单细胞测序中的应用"的SPAdes方法,Benkevich A等人,《计算机生物信息学杂志》,2012年5月,第19卷(第5期):第455-477页;或"A5-miseq:从Illumina MiSeq数据组装微生物基因组的更新pipeline工具"的A5-miseq方法,Coil D等人,《生物信息学》,2015年2月15日,第31卷(第4期):第587-589页,doi:10.1093/bioinformatics/btu661(Epub电子档出版于2014年10月22日)。

[0134] 重建或组装序列的步骤S300还可包括使用对包含突变的序列的重建部分的错误校正来重建不包含突变的序列的至少一部分,即,通过使用多个非突变的序列读段从包含突变的序列的重建部分推断最可能的不包含突变的序列。用于此类错误校正的方法包括例如"FMLRC:使用FM索引的混合长读段错误校正"的FMLRC方法,Jeremy R.Wang等人,《BMC生物信息学》,第19卷,文章编号:50(2018年)。例如,可使用非突变的序列读段对包含突变的序列进行错误校正,以去除引入的突变,从而重建不包含突变的序列的部分。错误校正可包括,例如确定包含突变的序列的可能编辑集合,这将需要将包含突变的序列转换成与非突变的序列读段相容的不包含突变的序列,从可能的编辑集合确定具有最小尺寸(即包含最少编辑)的编辑集合,并且将确定的具有最小尺寸的编辑集合应用于包含突变的序列,从而得出不包含突变序列的可能估计。然后可使用标准从头长读段组装工具(诸如Canu、Flye或PEREGRINE)、或使用工具(如Unicycler或MaSuRCA)与R中的短读段混合,从而组装不包含突变的序列来组装不包含突变的序列的部分。

[0135] 样品池的处理

[0136] 当处理包括多个样品的样品批次时，可针对每个样品以定义的序列标签的形式引入样品条形码。如果方法200的用户希望对多个样品使用该方法，其中每个样品包括一个或多个突变的靶模板核酸分子，一种可能性是在实验室中单独处理(例如，使其突变或片段化)每个样品，并且然后仅在测序前的最终步骤引入样品条形码。另一个另选的实施方案是仅在靶模板核酸分子的末端处引入样品条形码，在这种情况下，可能在样品制备过程的早期汇集所有条形码的靶模板核酸分子，从而大大降低试剂和实际操作的劳动成本(所谓的早期样品汇集方法)。因此，样品制备可包括在每个样品中的靶模板核酸分子的末端引入相应的样品条形码，使得每个样品包含靶模板核酸分子，这些靶核酸分子具有与其他样品中的靶模板核酸分子不同的样品条形码。样品制备还可包括汇集样品以创建样品池，使样品池中的靶模板核酸分子突变和任选地片段化，以及对样品池中突变的靶模板核酸分子的部分测序。

[0137] 然而，因为所得的多个突变的序列读段P含有来自许多不同样品的突变的序列读段P的未标记混合物，所以早期样品汇集方法在数据处理中引入了另外的挑战。可单独处理样品以构建非突变的序列读段R，在这种情况下，非突变的序列读段R包含多个非突变的序列读段集合$R^1 \cdots R^\zeta$，其中ζ是批次中已处理的样品数量。非突变的序列读段的每个集合可与相应的样品相关联。方法200可包括在多个非突变的序列读段集合$R^1 \cdots R^\zeta$中接收非突变的序列读段R，每个非突变的序列读段集合$R^1 \cdots R^\zeta$与相应的一个或多个样品相关联。

[0138] 因此，多个突变的序列读段中的每个突变的序列读段可以是包含与多个样品中的一个样品相关联的突变的序列的亚序列。多个非突变的序列读段中的每个非突变的序列读段可对应于不包含与多个样品中的一个样品相关联的突变的序列的亚序列。与相应的不包含突变的序列相比，每个包含突变的序列可包含突变。获得非突变的种子掩蔽的k-mer集合可包括获得每个样品的相应的非突变的种子掩蔽的k-mer集合。

[0139] 处理来自ζ个样品的数据的简单方法将是对ζ个样品中的每个样品应用一次方法200。另选的方法是扩展方法200，使得可同时处理所有的ζ个样品。这一点可如下所述实现。

[0140] 方法200(例如，步骤S230)可包括创建非突变样品位向量的集合，每个非突变样品位向量针对非突变的种子掩蔽的k-mer集合V_R中的相应k-mer定义，在多个样品的该非突变的种子掩蔽的k-mer集合中存在(或存在至少x次，其中x为大于或等于1的整数)相应k-mer，并且在多个样品的该非突变的种子掩蔽的k-mer集合中不存在(或存在少于x次)相应k-mer。可以上文已经描述的方式从多个非突变k-mer创建非突变的种子掩蔽的k-mer集合V_R。该多个非突变k-mer可被定义为在多个样品$R^1 \cdots R^\zeta$中的每个样品中的所有k-mer的并集，即，多个非突变k-merR可被定义为$R = \cup R^1 \cdots R^\zeta$。

[0141] 例如，方法200可包括将V_R的满射定义到位向量的集合上，该集合含有用于每个样品中种子掩蔽的k-mer的存在的二进制指示符。如果多个样品中的第i个样品含有k-mer(或含有k-mer至少X次)，则每个位向量在位置i可具有1，否则在位置i具有0。在软件具体实施中，可使用无序映射数据结构(诸如哈希映射)或近似成员查询结构(诸如计数商过滤器)来存储该满射。满射可表示为$Z:V_R \rightarrow v$，其中v为长度为ζ的位向量的空间。

[0142] 确定每个突变的序列读段中一个或多个突变的位置的步骤S230可扩展为同时为多个样品构建位向量α。确定一个或多个突变的位置可包括：对于每个突变的序列读段，以及对于非突变的种子掩蔽的k-mer集合中的每个集合和/或每个组合，识别突变的序列读段

中的被所有的种子模式掩蔽的一个或多个位置，这些种子模式对应于存在于非突变的种子掩蔽的k-mer集合的相应集合或组合中的多个突变的种子掩蔽的k-mer中的突变的种子掩蔽的k-mer；以及将所识别的一个或多个位置与一个或多个样品相关联，这些样品与非突变的种子掩蔽的k-mer集合的相应集合或组合相关联。这一点可例如通过函数morphomuts(ρ^i,V_R)的多样品变体morphomutsMS(ρ^i,V_R)来实现，该变体包括以下步骤：

[0143] 1.初始化位向量α的集合A，其中一个初始位向量a_0长度为2且仅含有0；初始化长度为2k且仅含有1的位向量b；初始化位向量c的集合C，其中一个初始成员c_0长度为ζ且仅含有1；初始化映射$\Gamma: C \Leftrightarrow A;$

[0144] 2.对于1与-k之间的读段中的每个位置j：

[0145] a.对于每个种子模式ψ∈Ψ，确定ψ(k(ρ_j^i))

[0146] b.如果ψ(k(ρ_j^i))∈V_R，则执行以下步骤：

[0147] i.对于C的每个元素(即，对于每个c∈C)，计算d←c∧Z(ψ(k(ρ_j^i))；

[0148] ii.如果d仅含有0，则返回到2b.i以处理C的下一个元素(或如果下一个元素不存在，则处理下一个种子模式ψ或下一个位置j)，否则继续执行2b.iii；

[0149] iii.分配α←Γ(c)|(ψ(b)>>2j)，其中“|”表示按位逻辑OR，并且“>>”表示位右移运算符，并且从C中去除c；

[0150] iv.将d添加至C且将α添加至A，以及定义Γ中的映射d→a；

[0151] v.如果$\acute{d}\wedge c$为非零数，然后则将$\acute{d}\wedge c$添加至C且将Γ(c)添加至A，以及定义Γ中的映射$\acute{d}\wedge c \rightarrow \Gamma(c);$

[0152] vi.返回2b.i以处理C的下一元素c。如果C中不再存在c，则返回2a以处理下一个种子模式ψ。如果Ψ中不再存在ψ，则返回2以处理下一个位置j。否则，继续执行3，否则：

[0153] 3.在函数morphomuts()中使用所应用的转换来创建α，以此转换A中的位向量；并且

[0154] 4.作为函数的结果，返回C、A和映射Γ。

[0155] 任选地，当在A中的位向量中匹配的位置太少(即，少于预先确定的数量y，其中y为大于或等于1，优选地大于或等于2、3、4或5的整数)时，可丢弃C和A中的相应条目。因为此类条目可能由于输入样品之间的随机相似性而出现，并且由此所得的位向量是与错误样品的错误匹配的结果，所以丢弃此类条目是有利的。通过在进一步处理之前丢弃这些位置，可避免不必要的计算。方法200可包括将所识别位置的数量与预先确定的数量y进行比较，其中y为大于或等于1的整数，优选地大于或等于2，并且如果所识别位置的数量小于预先确定的数量，则丢弃(或在进一步处理中忽略)所识别的一个或多个位置以及所识别的一个或多个位置与一个或多个样品的关联。

[0156] 存储在极小元箱中的元组可扩展以包括在C中的样品位向量信息。具体地，所存储的元组可以是读段索引i、突变的序列读段中的极小元的位置j，和c和α，其中c为样品位向量并且α是由morphomutsMS(ρ^i,V_R)函数计算的突变位向量。

[0157] 随后，当处理极小元箱以计算边缘权重时，可用对应的样品集注释每个边缘权重。如果与一对突变的序列读段相关联的样品位向量的按位逻辑AND为零，则可丢弃对应的边缘。如果边缘分数小于预先确定的阈值分数，则可丢弃边缘。当一对突变的序列读段之间存

在多个可能的边缘时，可以仅保留最高的边缘权重，并且对于该边缘，可将相关联的样品位向量集合计算为样品位向量之间的按位逻辑AND。这种方法的优点在于，可将不同样品序列中自然出现的差异与样品处理过程中引入的突变区分开来。如果仅在与为两个突变的序列读段识别的突变的一个或多个位置的相应对相关联的样品中存在重叠，即如果为两个突变的序列读段识别的突变的一个或多个位置的仅一对位置与至少一个共同样品相关，当两个突变序列的极小元进行比对时，则对于为两个突变的序列读段识别的突变的一个或多个位置的任何一对位置，可执行对具有匹配位置和/或具有错配位置的突变的数量计数的步骤S240。

[0158] 如果仅突变的靶模板核酸分子的末端含有样品标签，则多个突变的序列读段P中的一些突变的序列读段将携带该样品标签。特别地，通过对突变的靶模板核酸分子的末端测序而创建的突变的序列读段将携带该样品标签。在对突变序列读段进行聚类之后，可简单地通过评估每个簇中样品标签的读段的存在来将读段簇分配给样品。当在簇中仅出现单个样品标签时，对样品的分配是直接且明确的。对无向加权图执行图聚类可包括将由相应的簇中的突变的序列读段中的至少一个突变的序列读段包含的样品标签关联到突变的序列读段的每个簇。

[0159] 有时，由于测序或数据分析过程中的干扰或错误，多个样品标签可能会出现在单个簇中。在这种情况下，如果与其他标签相比，一个样品标签存在大量过量，则可靠的样品分配仍然是可能的。在不可能进行单个分配的情况下，仍然可以通过使用半监督图分解程序来消除样品的歧义，该程序将多样品簇分解为多个较小的簇，每个样品标签对应一个簇。即使当簇不含有携带样品标签的读段时，如果与读段链路相关联的大多数样品掩码都指示单个样品，则仍然可以将簇可靠地分配给样品。对无向加权图执行图聚类可包括在突变的序列读段的每个簇中识别由突变的序列读段的相应簇中的突变的序列读段包含的一个或多个样品标签。突变的序列读段的每个簇可与在相应簇中的突变的序列读段中最频繁出现的样品标签相关联。任选地，如果在突变的序列读段的簇中识别两个或更多个不同的样品标签，则突变的序列读段的簇可分成两个或更多个簇，该两个或更多个簇中的每个簇与该两个或更多个不同样品标签中的相应一个标签相关联并且含有突变的序列读段中的不同序列读段。

[0160] 样品制备和测序

[0161] 用于确定至少一个靶模板核酸分子的序列的至少一部分的方法10可包括对包含突变的至少一个靶模板核酸分子的100个区域测序，以便提供多个突变的序列读段。用于确定至少一个靶模板核酸分子的序列的至少一部分的方法10还可包括执行方法200，基于通过测序100获得的多个突变的序列读段，该方法确定与两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度。

[0162] 测序步骤可包括：

[0163] a)提供一对样品，每个样品包含至少一个靶模板核酸分子；

[0164] (b)对该对样品的第一样品中的该至少一个靶模板核酸分子的区域测序，以提供多个非突变的序列读段；

[0165] (c)将突变引入该对样品的第二样品中的至少一个靶模板核酸分子中，以提供该至少一个突变的靶模板核酸分子；

[0166] (d)对该至少一个突变的靶模板核酸分子的区域测序,以提供该多个突变的序列读段。

[0167] 在优选的实施方案中,引入突变的步骤包括将转换突变引入该对样品的该第二样品中的至少一个靶模板核酸分子中。

[0168] 测序步骤可包括:

[0169] (a)提供多对样品,每对样品包括第一样品和第二样品,每个样品包含至少一个靶模板核酸分子;

[0170] (b)将样品条形码引入每对样品的该至少一个靶模板核酸分子中,使得每对样品与相应的样品条形码相关联;

[0171] (c)对每个第一样品中的该至少一个靶模板核酸分子的区域测序,以提供该多个非突变的序列读段,其中对每个第一样品单独测序,从而为每个第一样品提供相应的非突变的序列读段集合;

[0172] (d)汇集该第二样品,从而创建第二样品的样品池;

[0173] (e)将突变引入该样品池中的靶模板核酸分子中,以提供突变的靶模板核酸分子;

[0174] (d)对该突变的靶模板核酸分子的区域测序,以提供该多个突变的序列读段。

[0175] 任选地,测序步骤还可包括在引入样品条形码之后以及在对至少一个靶模板核酸分子的区域测序之前,使每个第一样品中的至少一个靶模板核酸分子片段化。任选地,测序步骤还可包括在对突变的靶模板核酸分子的区域测序之前,使样品池中的至少一个靶模板核酸分子或突变的靶模板核酸分子片段化。

[0176] 在本发明的方法中,可同时对任何数量的至少一个靶模板核酸分子测序。因此,在本发明的实施方案中,至少一个靶模板核酸分子包含多个靶模板核酸分子。任选地,该至少一个靶模板核酸分子包含至少10个、至少20个、至少50个、至少100个或至少250个靶模板核酸分子。任选地,至少一个靶模板核酸分子包含10个与1000个之间、20个与500个之间或50个与100个之间的靶模板核酸分子。

[0177] 步骤S110:样品制备

[0178] 由于该对样品的第一样品和该对样品的第二样品均包含至少一个靶模板核酸分子,因此该对样品可衍生自相同的靶生物体或从相同的原始样品采集。

[0179] 例如,如果用户期望对样品中的至少一个靶模板核酸分子测序,则用户可从相同原始样品采集一对样品。任选地,用户可在从原始样品采集该对样品之前复制原始样品中的至少一个靶模板核酸分子。用户可能期望对来自特定生物体(诸如大肠杆菌(E.coli))的各种核酸分子测序。如果是这种情况,该对样品的第一样品可能是来自一个来源的大肠杆菌样品,而该对样品的第二样品可能是来自第二来源的大肠杆菌的样品。

[0180] 该对样品可源自包含或疑似包含该至少一个靶模板核酸分子的任何来源。该对样品可包括衍生自人类的核酸分子的样品,例如从人类患者的皮肤拭子中提取的样品。另选地,该对样品可衍生自其他来源,诸如水供应。此类样品可含有数十亿个模板核酸分子。使用本发明的方法可以同时对这些数十亿个靶模板核酸分子中的每个靶模板核酸分子测序,因此对可用于本发明方法的靶模板核酸分子的数量没有上限。

[0181] 在一个实施方案中,可提供多对样品。例如,可提供至少2对、3对、4对、5对、6对、7对、8对、9对、10对、11对、15对、20对、25对、50对、75对、100对、500对、1000对或5000对样品。

另选地，提供小于10000对、小于5000对、小于1000对、小于100对、小于75对、小于50对、小于25对、小于20对、小于15对、小于11对、小于10对、小于9对、小于8对、小于7对、小于6对、小于5对或小于4对样品。另选地，提供2对与100对、2对与75对、2对与50对之间、2对与25对之间、5对与15对之间或7对与15对之间的样品。

[0182] 在提供多对样品的情况下，不同对样品中的至少一个靶模板核酸分子可用不同的样品标签(本文也称为样品条形码)进行标记。例如，如果用户期望提供2对样品，则第一对样品中的所有或基本上所有的至少一个靶模板核酸分子可用样品标签A进行标记，并且第二对样品中的所有或基本上所有的至少一个靶模板核酸分子可用样品标签B进行标记。

[0183] 用于扩增至少一个靶模板核酸分子的合适方法是本领域已知的。例如，通常使用PCR。PCR在以下标题“将突变引入至少一个靶模板核酸分子中”更详细地描述。

[0184] 将突变引入至少一个靶模板核酸分子中

[0185] 该方法可包括将突变引入该对样品的第二样品中的至少一个靶模板核酸分子中，以提供该至少一个突变的靶模板核酸分子的步骤。

[0186] 突变可以是取代突变、插入突变或缺失突变。出于本发明的目的，术语“取代突变”应解释为意指核苷酸被不同核苷酸取代。例如，序列ATCC到序列AGCC的转换引入了单个取代突变。出于本发明的目的，术语“插入突变”应解释为意指将至少一个核苷酸添加至序列。例如，序列ATCC到序列ATTCC的转换是插入突变的示例(其中插入了另外的T核苷酸)。出于本发明的目的，术语“缺失突变”应解释为意指从序列移除至少一个核苷酸。例如，序列ATTCC到ATCC的转换是缺失突变的示例(其中移除了T核苷酸)。优选地，突变是取代突变。

[0187] 短语“将突变引入至少一个靶模板核酸分子中”是指将该对样品的第二样品中的至少一个靶模板核酸分子暴露于其中该至少一个靶模板核酸分子发生突变的条件下。可使用任何合适的方法来实现这一点。例如，可通过化学诱变和/或酶促诱变引入突变。

[0188] 任选地，将突变引入至少一个靶模板核酸分子中的步骤在至少一个靶模板核酸分子的1％与50％之间、3％与25％之间、5％与20％之间或约8％的核苷酸发生突变。任选地，至少一个突变的靶模板核酸分子包含1％与50％之间、3％与25％之间、5％与20％之间或约8％突变。

[0189] 用户可确定至少一个突变的靶模板核酸分子内包含多少突变，和/或通过对已知序列的核酸分子执行引入突变的步骤，将突变引入至少一个靶模板核酸分子中的步骤使至少一个靶模板核酸分子突变的程度，对所得的核酸分子测序，并确定与原始序列相比已经发生变化的核苷酸总数的百分比。

[0190] 任选地，将突变引入至少一个靶模板核酸分子中的步骤使至少一种靶模板核酸分子以基本上随机的方式突变。任选地，至少一个突变的靶模板核酸分子包含基本上随机的突变模式。

[0191] 如果至少一个突变的靶模板核酸分子在整个长度上含有基本上相似水平的突变，则该至少一个突变的靶模板核酸分子包含基本上随机的突变模式。例如，用户可以通过使已知序列的测试核酸分子突变以提供突变的测试核酸分子，以确定至少一个突变的靶模板核酸分子是否包括基本上随机的突变模式。可以将突变的测试核酸分子的序列与测试核酸分子进行比较，以确定每个突变的位置。然后，用户可以通过以下方式确定突变是否在突变的测试核酸分子的整个长度内以基本相似的水平出现：

[0192] (i)计算突变中的每个突变之间的距离；

[0193] (ii)计算这些距离的平均值；

[0194] (iii)在不替换的情况下，将距离细分为较小的数量，诸如500或1000；

[0195] (iv)从几何分布构建一个500个或1000个距离的模拟集合，其平均值由矩量法给出，以匹配先前根据观察到的距离计算的值；并且

[0196] (v)计算两个分布上的Kolmolgorov-Smirnov。

[0197] 如果D<0.15、D<0.2、D<0.25或D<0.3，则可认为至少一个突变的靶模板核酸分子包含基本上随机的突变模式，这取决于非突变读段的长度。

[0198] 相似地，如果所得的至少一个突变的靶模板核酸分子包含基本上随机的突变模式，则将突变引入至少一个靶模板核酸分子中的步骤使至少一个靶模板核酸分子以基本上随机的方式突变。将突变引入至少一个靶模板核酸分子中的步骤是否使至少一个靶模板核酸分子以基本上随机的方式突变，可以通过进行将突变引入已知序列的测试核酸分子上的至少一个靶模板核酸分子中以提供突变的测试核酸分子的步骤来确定。然后，用户可以对突变的测试核酸分子测序，以识别哪些突变已经被引入，并确定突变的测试核酸分子是否包含基本上随机的突变模式。

[0199] 任选地，至少一个突变的靶模板核酸分子包括非偏置的突变模式。任选地，将突变引入至少一个靶模板核酸分子中的步骤以非偏置方式引入突变。如果引入的突变类型是随机的，则至少一个突变的靶模板核酸分子包含非偏置的突变模式。如果引入的突变是取代突变，然后如果引入了相似比例的A(腺嘌呤)、T(胸腺嘧啶)、C(胞嘧啶)和G(鸟嘌呤)核苷酸，则引入的突变是随机的。短语“引入了相似比例的A(腺嘌呤)、T(胸腺嘧啶)、C(胞嘧啶)和G(鸟嘌呤)核苷酸”，我们的意思是，引入的腺嘌呤数量、胸腺嘧啶数量、胞嘧啶数量和鸟嘌呤核苷酸数量在彼此的20％以内(例如，可以引入20个A核苷酸、18个T核苷酸、24个C核苷酸和22个G核苷酸)。

[0200] 将突变引入至少一个靶模板核酸分子中的步骤是否使至少一个靶模板核酸分子以非偏置的方式突变，可以通过进行将突变引入已知序列的测试核酸分子上的至少一个靶模板核酸分子中以提供突变的测试核酸分子的步骤来确定。然后，用户可以对突变的测试核酸分子测序，以识别哪些突变已经被引入，并确定突变的测试核酸分子是否包含非偏置的突变模式。

[0201] 可用的，甚至当将突变引入至少一个靶模板核酸分子中的步骤引入不均匀分布的突变时，也可以使用生成至少一个靶模板核酸分子的序列的方法。因此，在一个实施方案中，至少一个突变的靶模板核酸分子包含不均匀分布的突变。任选地，将突变引入至少一个突变的靶模板核酸分子中的步骤引入不均匀分布的突变。如果以偏置方式引入突变，则突变被认为是“不均匀分布”，即引入的腺嘌呤数量、胸腺嘧啶数量、胞嘧啶数量和鸟嘌呤核苷酸数量不在彼此的20％以内。无论至少一个突变的靶模板核酸分子是否包含不均匀分布的突变，还是将突变引入至少一个靶模板核酸分子中的步骤引入不均匀分布的突变，可以与上文所描述的方式相似的方式确定，以确定将突变引入至少一个靶模板核酸分子中的步骤是否以非偏置的方式引入突变。

[0202] 相似地，甚至当突变的序列读段和/或非突变的序列读段包括不均匀分布的测序错误时，也可以使用生成至少一个靶模板核酸分子的序列的方法。因此，在一个实施方案

中,突变的序列读段和/或非突变的序列读段包括不均匀分布的测序错误。相似地,在一个实施方案中,对至少一个靶模板核酸分子的区域测序和/或对至少一个突变的靶模板核酸分子的区域测序的步骤引入不均匀分布的序列错误。

[0203] 对至少一个靶模板核酸分子的区域测序和/或对至少一个突变的靶模板核酸分子的区域测序的特定步骤是否引入不均匀分布的序列错误,将可能取决于测序仪器的准确性并且将可能为用户所知。然而,通过对已知序列的核酸分子执行测序的方法,并将产生的序列读段与已知序列的原始核酸分子的序列读段进行比较,用户可以研究至少一个靶模板核酸分子的区域的测序和/或至少一个突变的靶模板核酸分子的区域的测序步骤是否引入了不均匀分布的序列错误。然后,用户可以应用实施例6中讨论的概率函数,并且确定M和E的值。如果E和矩阵模型的值不相等或基本上不相等(在彼此的10%内),则对至少一个靶模板核酸分子的区域测序的步骤引入不均匀分布的序列错误。

[0204] 可通过将至少一个靶模板核酸暴露于化学诱变来实现经由化学诱变将突变引入至少一个靶模板核酸分子中。合适的化学诱变剂包括丝裂霉素C(MMC)、N-甲基-N-亚硝基脲(MNU)、亚硝酸(NA)、二环氧丁烷(DEB)、1,2,7,8-二环氧辛烷(DEO)、甲磺酸乙酯(EMS)、甲磺酸甲酯(MMS)、N-甲基-N'-硝基-N-硝基胍(MNNG)、4-硝基喹啉-1-氧化物(4-NQO)、2-甲氧基-6-氯-9(3-[乙基-2-氯乙基]-氨丙基氨基)-吖啶二盐酸盐(ICR-170)、2-氨基嘌呤(2A)、亚硫酸盐和羟胺(HA)。例如,当核酸分子暴露于亚硫酸氢盐时,亚硫酸氢盐将胞嘧啶脱氨基形成尿嘧啶,从而有效地引入C-T取代突变。

[0205] 如上所述,将突变引入至少一个靶模板核酸分子中的步骤可以通过酶诱变进行。任选地,使用DNA聚合酶进行酶诱变。例如,一些DNA聚合酶是易错的(是低保真度聚合酶),并且使用易错的DNA聚合酶复制至少一个靶模板核酸分子将引入突变。Taq聚合酶是低保真度聚合酶的示例,并且将突变引入至少一个靶模板核酸分子中的步骤可以通过使用Taq聚合酶复制至少一个靶模板核酸分子来进行,例如通过PCR。

[0206] DNA聚合酶可以是低偏置的DNA聚合酶。

[0207] 如果使用DNA聚合酶进行将突变引入至少一个靶模板核酸分子中的步骤,则可以在适合DNA聚合酶的条件下将至少一个靶模板核酸分子与DNA聚合酶和合适的引物一起温育以催化至少一个突变的靶模板核酸分子的生成。

[0208] 合适的引物包含与至少一个靶模板核酸分子侧翼的区域或与至少一个靶模板核酸分子互补的核酸分子侧翼区域互补的短核酸分子。例如,如果该至少一个靶模板核酸分子是染色体的一部分,则引物将与该至少一个靶核酸分子的接近3'至3'端的染色体区域互补,并且与该至少一个靶核酸分子的接近5'至5'端的染色体区域互补,或者引物将与核酸分子的接近3'至3'端的染色体区域互补,该核酸分子与该至少一个靶模板核酸分子互补,并且与核酸分子的接近5'至5'端的染色体区域互补,该核酸分子与该至少一个靶模板核酸分子互补。

[0209] 合适的条件包括DNA聚合酶可以复制至少一个靶模板核酸分子的温度。例如,40℃与90℃之间、50℃与80℃之间、60℃与70℃之间或约68℃的温度。

[0210] 将突变引入至少一个模板核酸分子中的步骤可以包括多轮复制。例如,将突变引入至少一个靶模板核酸分子中的步骤优选地包括:

[0211] i)一轮复制至少一个靶模板核酸分子,以提供与至少一个靶模板核酸分子互补的

至少一个核酸分子;并且

[0212] ii)一轮复制至少一个靶模板核酸分子,以提供至少一个靶模板核酸分子的复制。

[0213] 任选地,将突变引入至少一个靶模板核酸分子中的步骤包括至少2轮、至少4轮、至少6轮、至少8轮、至少10轮、少于10轮、少于8轮、约6轮、2轮与8轮之间或1轮与7轮之间复制至少一个靶模板核酸分子。用户可以选择使用低数量的复制轮次来减少引入扩增偏差的可能性。

[0214] 任选地,将突变引入至少一个靶模板核酸分子中的步骤包括在60℃与80℃之间的温度,至少2轮、至少4轮、至少6轮、至少8轮、至少10轮、少于10轮、少于8轮、约6轮、2轮与8轮之间或1轮与7轮之间的复制轮次。

[0215] 任选地,使用聚合酶链反应(PCR)进行将突变引入至少一个靶模板核酸分子中的步骤。PCR是涉及用于复制核酸分子的多轮以下步骤的过程:

[0216] a)熔融;

[0217] b)退火;和

[0218] c)延伸和伸长。

[0219] 将核酸分子(诸如至少一个靶模板核酸分子)与合适的引物和聚合酶混合。在熔融步骤中,将核酸分子加热至高于90℃的温度,使得双链核酸分子将变性(分离成两条链)。在退火步骤中,将核酸分子冷却至低于75℃的温度,例如55℃与70℃之间、约55℃或约68℃,以允许引物与核酸分子退火。在延伸和伸长步骤中,将核酸分子加热至大于60℃的温度以允许DNA聚合酶催化引物延伸,添加与模板链互补的核苷酸。

[0220] 任选地,将突变引入至少一个靶模板核酸分子中的步骤包括使用Taq聚合酶在易错的反应条件下复制至少一个靶模板核酸分子。例如,将突变引入至少一个靶模板核酸分子中的步骤可以包括在存在Mn^{2+}、Mg^{2+}或不相等的dNTP浓度(例如过量的胞嘧啶、鸟嘌呤、腺嘌呤或胸腺嘧啶)的情况下使用Taq聚合酶的PCR。

[0221] 步骤S120:测序

[0222] 获得包括非突变的序列读段和突变的序列读段的数据

[0223] 本发明的方法可以包括接收突变的序列读段并且任选地接收非突变的序列读段的步骤。可以从任何来源获得非突变的序列读段和突变的序列读段。

[0224] 任选地,通过对一对样品的第一样品中的至少一个靶模板核酸分子的区域测序来获得非突变的序列读段。任选地,通过将突变引入一对样品的第二样品中的至少一个靶模板核酸分子中以提供至少一个突变的靶模板核酸分子,并且通过对至少一个靶模板核酸分子的区域测序获得突变的序列读段。

[0225] 任选地,非突变的序列读段包括一对样品的第一样品中的至少一个靶模板核酸分子的区域的序列,突变的序列读段包括一对样品的第二样品中的至少一个突变的靶模板核酸分子的区域的序列,并且该对样品采集自相同的原始样品或衍生自相同的生物体。

[0226] 对至少一个靶模板核酸分子或至少一个突变的靶模板核酸分子的区域测序

[0227] 用于确定至少一个靶模板核酸分子的序列的方法可以包括对该对样品的第一样品中的至少一个靶模板核酸分子的区域测序以提供非突变的序列读段的步骤,和/或对至少一个突变的靶模板核酸分子的区域测序以提供突变的序列读段的步骤。

[0228] 可使用任何测序方法进行测序步骤。可能的测序方法的示例包括Maxam Gilbert

测序、Sanger测序、包括桥式扩增(诸如桥式PCR)的测序或任何高通量测序(HTS)方法,如以下文献中所描述的:Maxam AM,Gilbert W(1977年2月),“一种新的DNA测序方法”,《美国国家科学院院刊》,第74卷(第2期):第560-564页;Sanger F,Coulson AR(1975年5月),“用DNA聚合酶引物合成快速确定DNA序列的方法”,《分子生物学杂志》,第94卷(第3期):第441-448页;以及Bentley DR,Balasubramanian S等人(2008年),“使用可逆终止子化学进行准确的全人类基因组测序”,《自然》,第456卷(第7218期):第53-59页。

[0229] 在典型的实施方案中,测序步骤中的至少一个步骤或优选地两个步骤都涉及桥式扩增。任选地,使用大于5、大于10、大于15或大于20秒的延伸时间进行桥式扩增步骤。使用桥式扩增的示例是使用Illumina公司的测序仪Genome Analyzer。优选地使用双端测序。

[0230] 任选地,使用相同的测序方法进行对该对样品的第一样品中的至少一个靶模板核酸分子的区域测序,以提供非突变的序列读段的步骤(i)和对至少一个突变的靶模板核酸的区域测序,以提供突变的序列读段的步骤(ii)。

[0231] 任选地,使用不同的测序方法进行对该对样品的第一样品中的至少一个靶模板核酸分子的区域测序,以提供非突变的序列读段的步骤(i)和对至少一个突变的靶模板核酸的区域测序,以提供突变的序列读段的步骤(ii)。

[0232] 任选地,可以使用多于一种的测序方法进行对该对样品的第一样品中的至少一个靶模板核酸分子的区域测序,以提供非突变的序列读段的步骤(i)和对至少一个突变的靶模板核酸的区域测序,以提供突变的序列读段的步骤(ii)。例如,可以使用第一测序方法对该对样品的第一样品中的至少一个靶模板核酸分子的一部分测序,并且可以使用第二测序方法对该对样品的第一样品中的至少一个靶模板核酸分子的一部分测序。相似地,可以使用第一测序方法对至少一个突变的靶模板核酸分子的一部分测序,并且可以使用第二测序方法对至少一个突变的靶模板核酸分子的一部分测序。

[0233] 任选地,在不同的时间进行对该对样品的第一样品中的至少一个靶模板核酸分子的区域测序,以提供非突变的序列读段的步骤(i)和对至少一个突变的靶模板核酸的区域测序,以提供突变的序列读段的步骤(ii)。另选地,步骤(i)和步骤(ii)可以相当同时地进行,诸如在彼此的1年内。该对样品的第一样品和该对样品的第二样品不需要彼此同时采集。在两个样品衍生自相同生物体的情况下,可以在基本上不同的时间提供它们,甚至相隔数年,并且因此两个测序步骤也可以相隔数年。此外,即使该对样品的第一样品和该对样品的第二样品衍生自相同的原始样品,也可以将生物样品存储一段时间,并且因此不需要同时进行测序步骤。

[0234] 突变的序列读段和/或非突变的序列读段可以是单端或双端序列读段。

[0235] 任选地,突变的序列读段和/或非突变的序列读段大于50nt、大于100nt、大于500nt、小于200,000nt、小于15,000nt、小于1,000nt、介于50nt与200,000nt之间、介于50nt与15,000nt之间或介于50nt与1,000nt之间。

[0236] 任选地,测序步骤使用每至少一个靶模板核酸分子的每个核苷酸的0.1个与500个读段之间、0.2个与300个读段之间或0.5个与150个读段之间的测序深度进行。测序深度越大,确定/生成的序列的准确度就将越大,但组装可能更困难。

[0237] 选择参数

[0238] 优选地,方法200中使用的参数选择如下。

[0239] 在优选的实施方案中，每个种子模式ψ的权重在5至50，优选地10至30的范围内，进一步优选地13至23的范围内。这确保每个种子模式ψ足够大以确保被每个种子模式ψ掩蔽的每个k-mer高概率是唯一的。例如，对于具有5百万个核苷酸的典型长度的细菌基因组，每个种子模式ψ的权重w(ψ)优选在13至19的范围内，注意4^{13}>5百万。对于具有约30亿个核苷酸的典型长度的人类大小的基因组，每个种子模式ψ的权重w(ψ)优选在19至23的范围内，注意4^{19}>3×10^{9}。

[0240] 在优选的实施方案中，在确定每个突变的序列读段中的一个或多个突变的位置的步骤S230中使用的每个k-mer的尺寸k大于每个种子模式ψ的权重w(ψ)。每个k-mer的尺寸k可以小于每个种子模式ψ的权重w(ψ)的5倍、小于4倍、小于3倍或小于2倍。在确定每个突变的序列读段中的一个或多个突变的位置的步骤S230中使用的每个k-mer的尺寸k可以在10至250，优选地13至100，进一步优选地15至50，最优选地20至40的范围内。这确保尺寸k足够小以确保任何k-mer包括插入或缺失测序错误的概率低，这在方法200的上下文中是不利的。

[0241] 示例性种子家族Ψ包括具有权重w(ψ)＝16和k＝27的种子模式ψ如下所示：

[0242] ψ_1＝{0,1,2,3,5,6,9,12,13,14,16,18,20,21,22,23}，

[0243] ψ_2＝{0,1,2,4,5,9,10,11,13,18,19,21,23,24,25,26}，

[0244] ψ_3＝{0,1,2,3,4,5,7,8,9,10,13,15,16,18,19,20}，

[0245] ψ_4＝{0,1,2,4,6,7,12,14,16,17,20,21,23,24,25,26}，

[0246] 在一个实施方案中，在应用公共极小函数的步骤S220中使用的k-mer(即为每个突变的序列读段确定的一个或多个极小元)具有与确定每个突变的序列读段中的一个或多个突变的位置的步骤S230中使用的k-mer不同的尺寸k。每个极小元的尺寸k可以在5至50，优选地10至30，进一步优选13至23的范围内。每个极小元的尺寸k可以基于与种子模式的权重w(ψ)的选择相同的考虑来选择。对于细菌，每个极小元的尺寸k可以在13至19的范围内，并且对于人类大小的基因组，每个极小元的尺寸k在19至23的范围内。

[0247] 方法200的具体实施

[0248] 可以多种方式实施方法200。优选的方法是首先在初始通过一些或所有的突变的序列读段P和非突变的序列读段R时计算集合U_M，然后在第二通过突变的序列读段P和非突变的序列读段R时计算W_M。有了W_M，在第三通过突变的序列读段P时，可以计算极小元位置以及一个或多个突变的位置，并且这些位置可以存储在RAM或固定存储装置(例如磁盘)中的极小元箱中。任选地，多个排序的或无序的极小元箱可以存储在单个文件中。然后可以顺序地或并行地读取每个极小元箱(或每个文件)，其中处理极小元箱以计算边缘权重。因为每个突变的序列读段可以在多个极小元箱中出现，所以一对突变的序列读段可能具有计算的它们的权重的多个估计。当这种情况发生时，必须使用一些措施来选择优选的权重，通常是最大值。最后，当测序化学已经产生双端读段，一对双端读段中的每个双端读段共享公共极小元，然后两个端的分数可以相加以得到该对双端读段的单个分数。

[0249] 实验数据

[0250] 方法200被用于处理若干实际SAM数据集，每个SAM数据集包括非突变的序列读段和突变的序列读段。

[0251] 处理Arobacter butzleri JV22的SAM数据集。这种生物有2.3Mbp的作为单条环状

染色体存在的基因组。在Amazon AWS实例上运行方法200的C++具体实施。SAM数据集由956133参考读段对(未突变)和衍生自大约8000突变的长模板的2154909突变读段对组成。突变读段的2087506(96.9%)衍生自突变的长模板的内部部分,而67403(3.1%)衍生自长模板的端部并且含有样品条形码。每个单独的读段长度为150nt或更小nt。读段对先前已经过质量和衔接子修整。方法200需要12分钟CPU时间和1.2GB RAM来处理数据集,从而在读段之间产生30033939候选链路。然后使用马尔可夫聚类(mcl)对这些链路进行图聚类,并且使用MEGAHIT(参见Dinghua Li,Chi-Man Liu,Ruibang Luo,Kunihiko Sadakane和Tak-Wah Lam,"MEGAHIT:经由简洁的de Bruijn图进行大型复杂宏基因组组装的超快速单节点解决方案",《生物信息学》,英国牛津,第31卷(第10期):第1674-1676页,2015年5月)从头组装所得的6779组读段,以产生长突变模板的重建。最后,长突变模板与未突变读段一起用于由Unicycler软件计算的杂交基因组组装中。与单独从短读段获得的组装相比(图4A所示),所得的组装如图4B所示。图4A示出了在执行方法200之前使用Shovill组装管道对2.3Mbp Arcobacter butzlerii基因组的短读段组装。这产生了78个支架的组装,其中最大的支架覆盖342kbp,并且支架N50约为127kbp。图4B示出了使用方法200的2.3Mbp Arcobacter butzlerii基因组的组装。环状染色体在很大程度上已解析为单个重叠群,其中<200nt的小片段的拷贝数仍未被解析。

[0252] 使用模拟数据测量方法200中实施的方法的可扩展性和解析能力。来自CFTR基因的50kbp序列被用于模拟增加数量的突变长模板覆盖率和来自这些模板的相应的突变短读段。使用新开发的首先生成长突变模板的脚本进行模拟,然后调用著名的Illumina读段模拟器artsim来模拟来自突变模板的短读段测序。除了突变数据之外,还使用artsim模拟了未突变序列的30X覆盖率。我们以数量级的增量模拟了10^1至10^6范围内的长突变模板覆盖率。突变率固定在6%。对于每个长模板,模拟10X短读段覆盖率。通过测量由方法200报告的假阳性链路率来评估模拟数据的结果。

[0253] 图5示出了每个长模板的短读段覆盖率的深度的影响。每个长模板生成的短读段数据的量在x轴上示出,y轴示出了来自方法200的结果的各种性能度量。可以看出,当每个模板的短读段覆盖率低时,例如<4x,获得了原始长模板的不良和不完整重建。然而,当每个突变模板的覆盖率处于5x至10x的范围内时,可以获得良好的重建。

[0254] • 链路num:由方法200报告的突变读段之间的链路数量。链路fp:报告的假阳性链路数量。

[0255] • 链路fp率:所有报告的链路中为假阳性的链路的分数。

[0256] • mcl num:通过对经mmdreaming报告的图进行马尔可夫聚类产生的簇数。

[0257] • idba scaf num:通过组装突变短读段的簇来重建的支架序列的数量。

[0258] • idba scaf bp:组装的所有支架的长度总和。

序列表

<110> 朗斯科技有限公司

<120> 用于确定与两个突变的序列读段衍生自包含突变的相同序列的概率相关的量度的方法

<130> N416858WO

<150> GB1914064.9

<151> 2019-09-30

<160> 2

<170> PatentIn 3.5版

<210> 1

<211> 45

<212> DNA

<213> 人工序列(Artificial Sequence)

<220>

<221>

<222>

<223> 示例性突变的序列读段

<400> 1

acggaaagcg ctacgagcga ctgatattga gctacgttca gagcc 45

<210> 2

<211> 28

<212> DNA

<213> 人工序列(Artificial Sequence)

<220>

<221>

<222>

<223> 示例性突变的序列读段

<400> 2

acgcaaagcg ctacgagcga ctgatatt 28

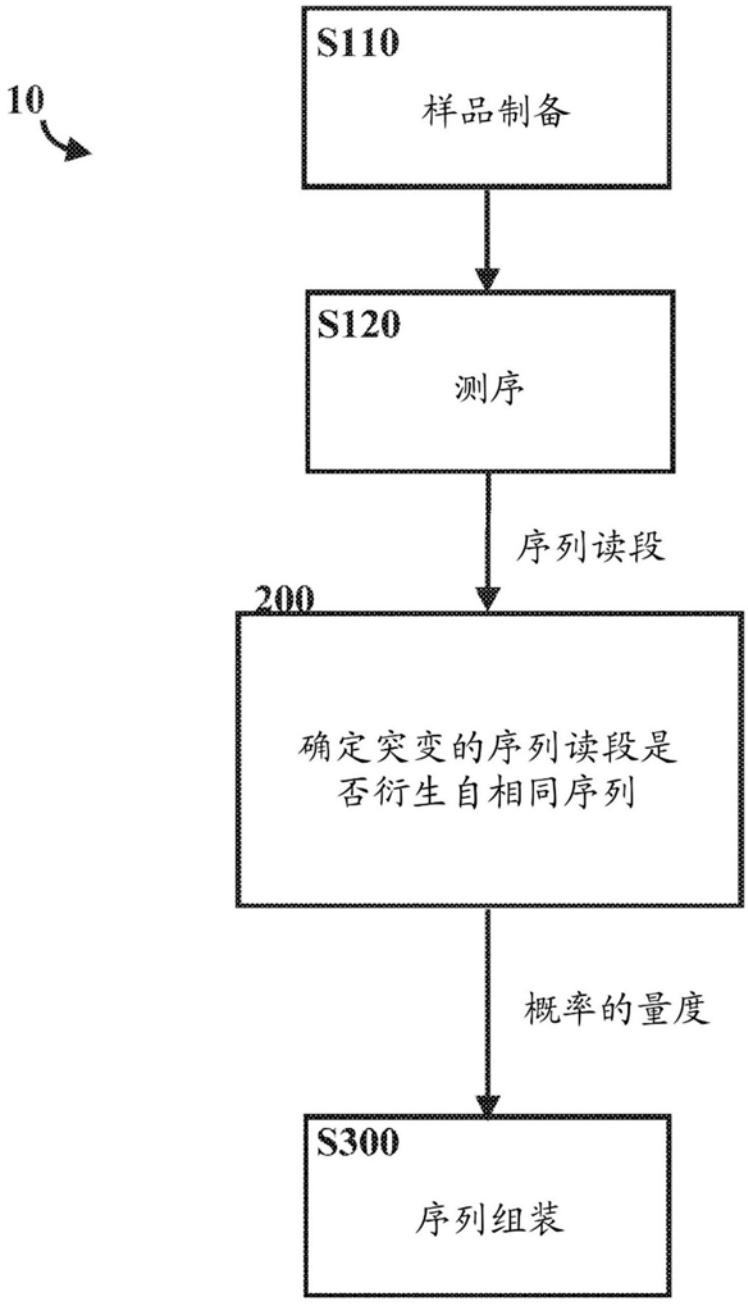

10
S110
样品制备
S120
测序
序列读段
200
确定突变的序列读段是
否衍生自相同序列
概率的量度
S300
序列组装

图1

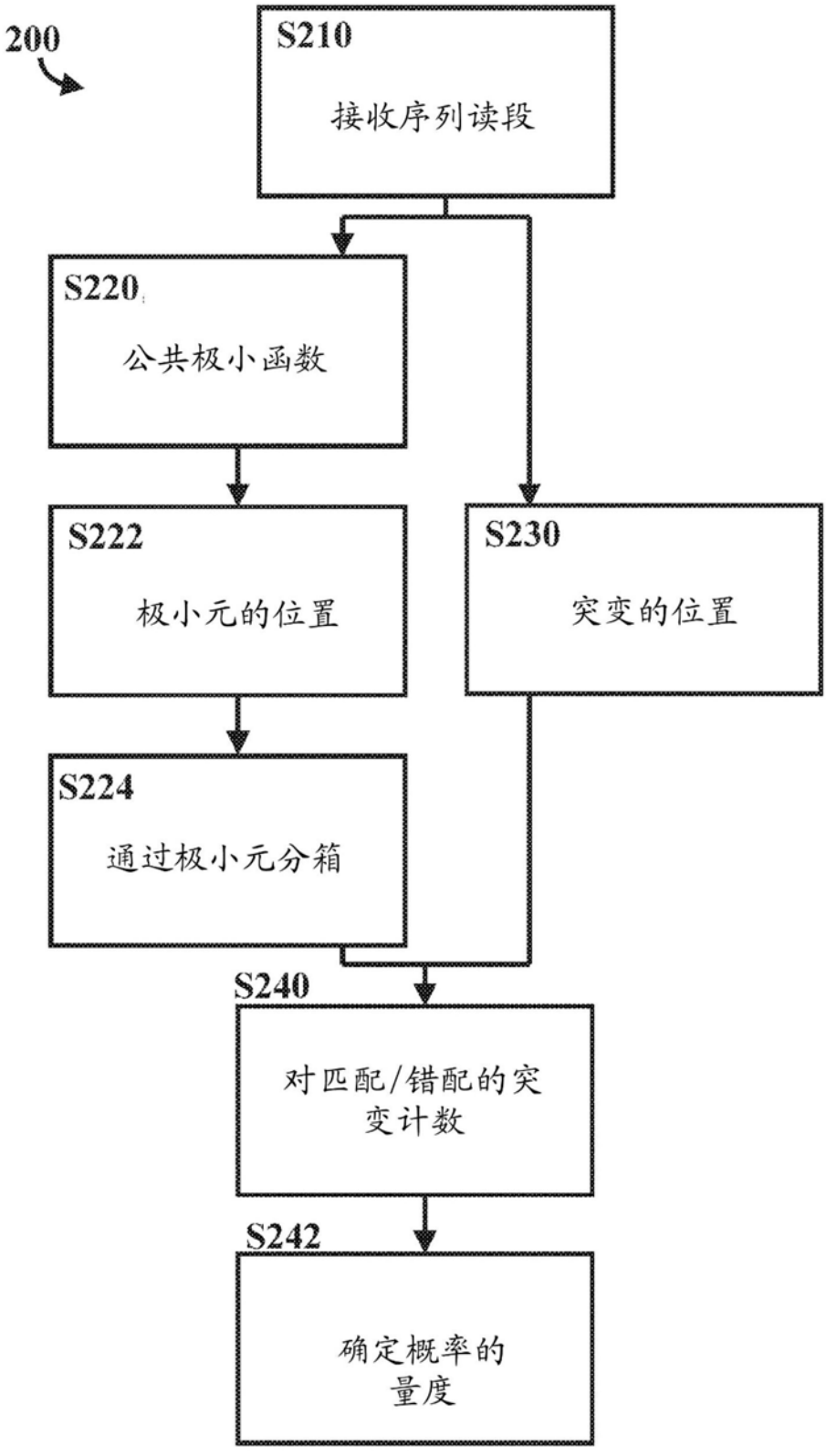

200
S210
接收序列读段
S220
公共极小函数
S222
极小元的位置
S230
突变的位置
S224
通过极小元分箱
S240
对匹配/错配的突变计数
S242
确定概率的量度

图2

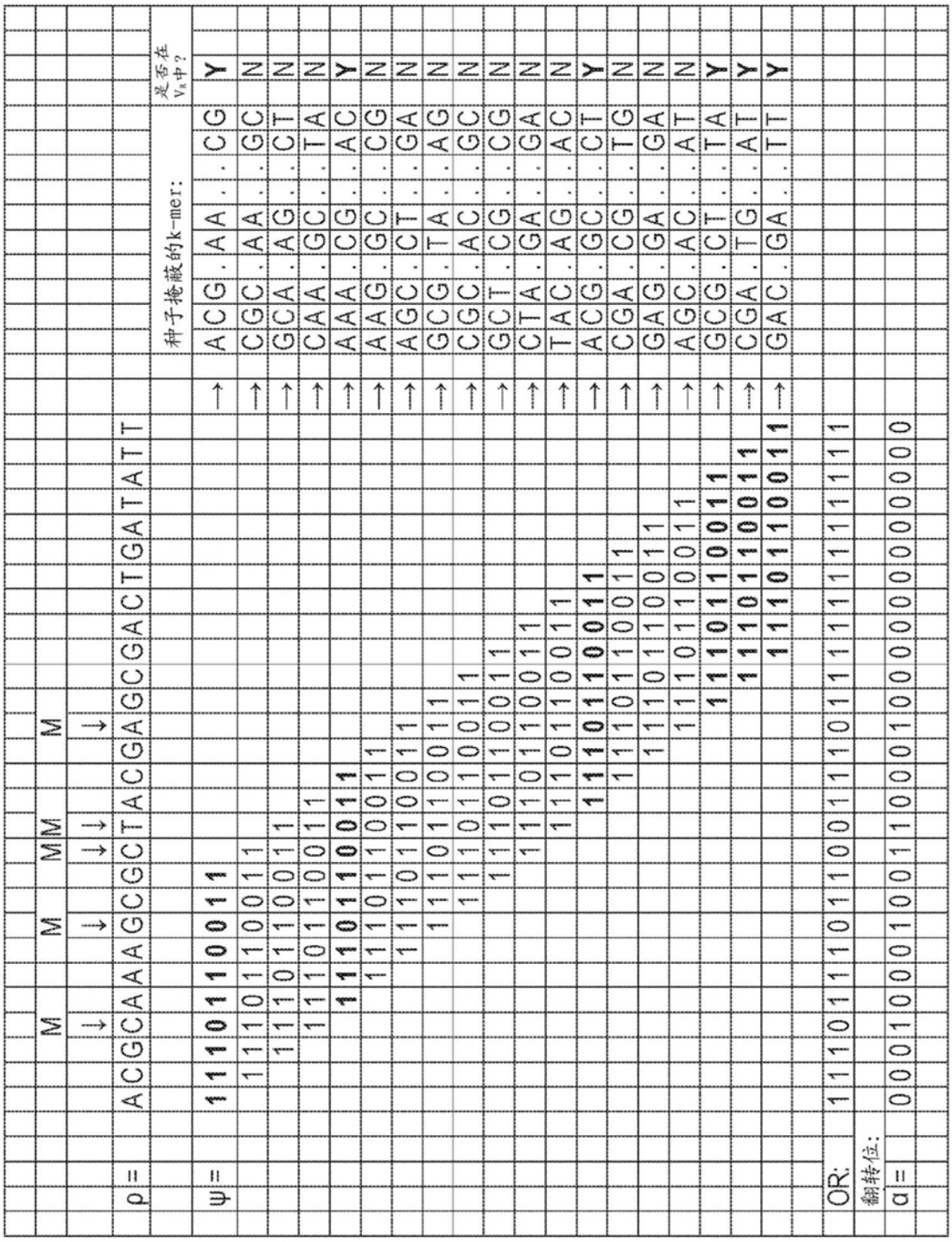

p= ACGCAAGCGCTACGAGCGACTGATATT
种子掩蔽的k-mer:
是否在 V_R 中?
Ψ= 1110110011
OR: 1110111011001111001111111111
翻转位:
q= 0001001001100010000000000

图3

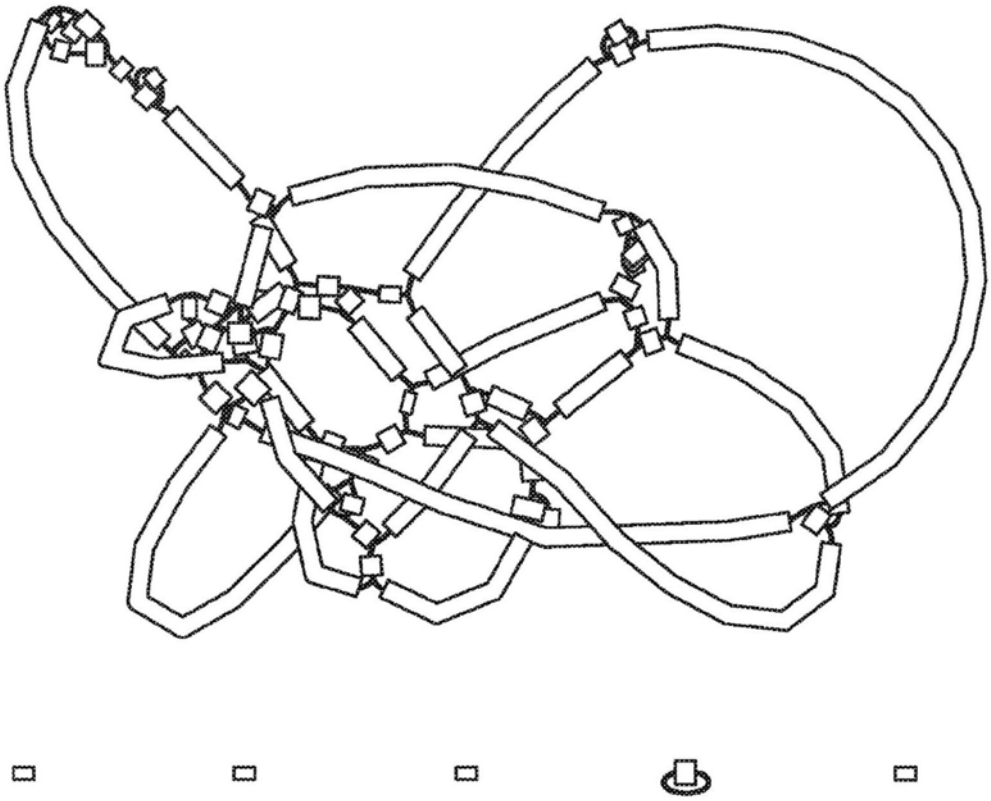


图4A

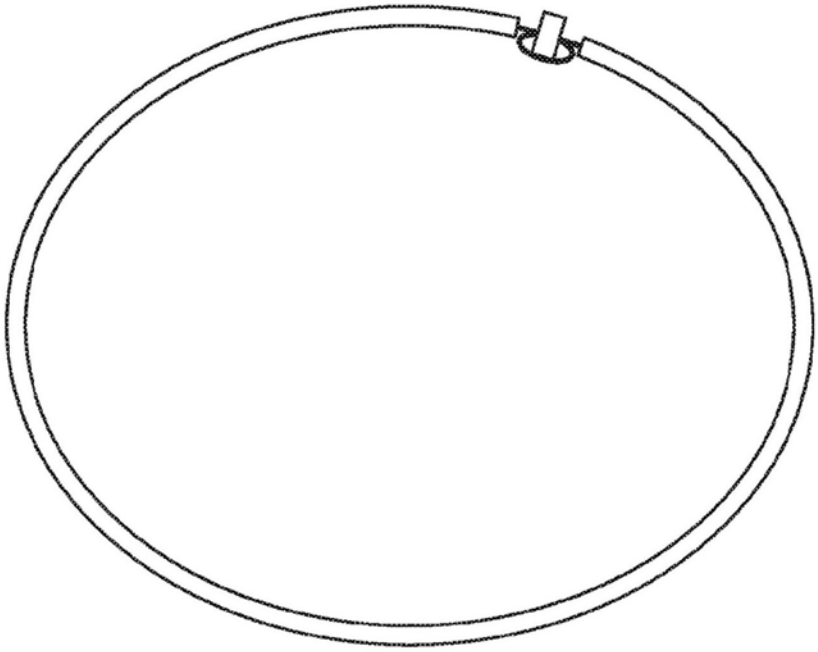


图4B

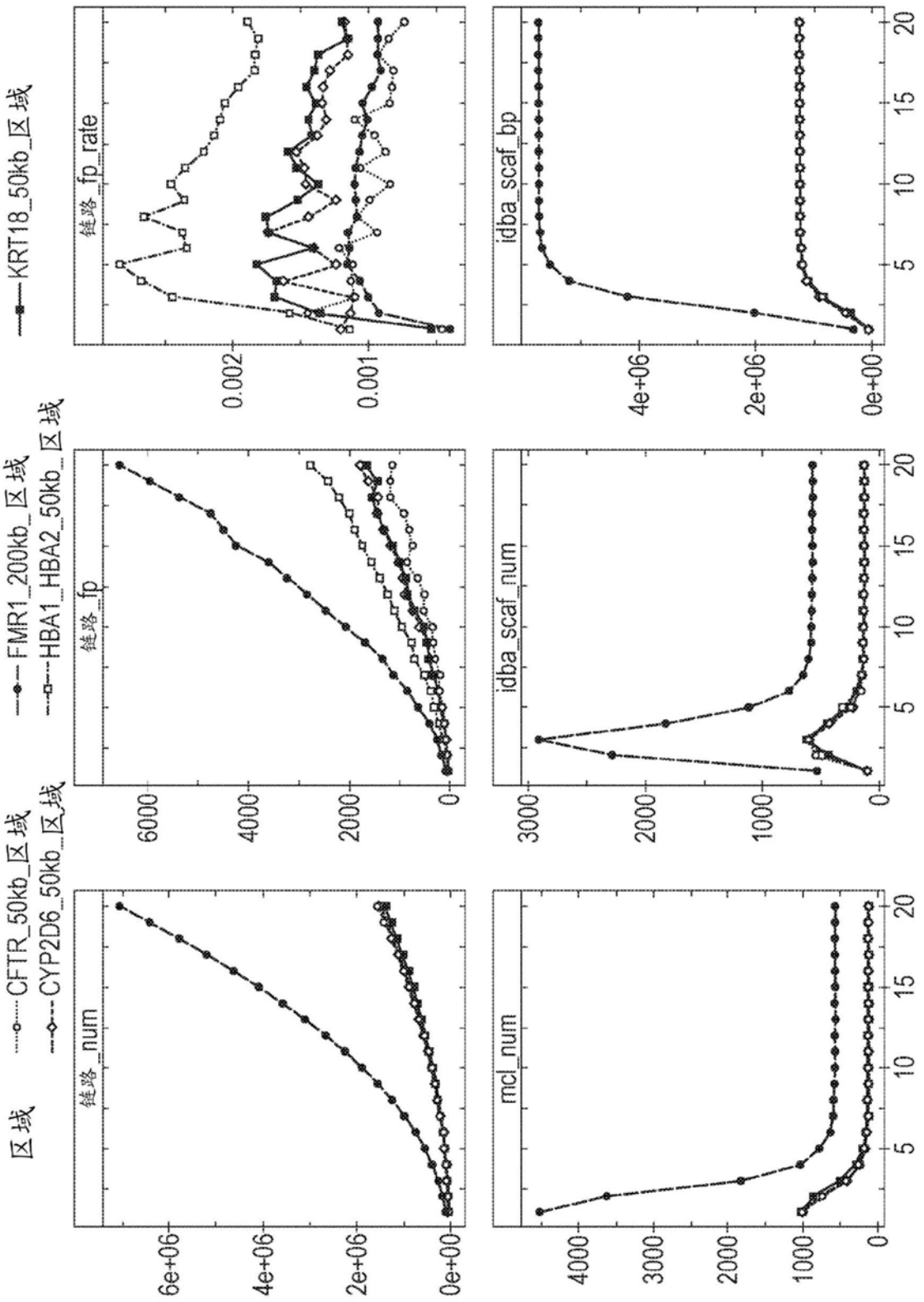

区域
CFTR_50kb_区域
CYP2D6_50kb_区域
FMR1_200kb_区域
HBA1_HBA2_50kb_区域
KRT18_50kb_区域
链路_num
链路_fp
链路_fp_rate
idba_scaf_num
idba_scaf_bp
mcl_num

图5