(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109920480 B
(45)授权公告日 2020.02.21

(21)申请号 201910194839.6

(22)申请日 2019.03.14

(65)同一申请的已公布的文献号
申请公布号 CN 109920480 A

(43)申请公布日 2019.06.21

(73)专利权人 深圳市海普洛斯生物科技有限公司
地址 518000 广东省深圳市前海深港合作区前湾一路1号A栋201室(入驻深圳市前海商务秘书有限公司)

(72)发明人 周衍庆 陈亚如 尤沁 徐云

(74)专利代理机构 深圳鼎合诚知识产权代理有限公司 44281
代理人 李小焦

(51)Int.Cl.
G16B 20/00(2019.01)
G16B 30/00(2019.01)

(56)对比文件
US 2017240963 A1,2017.08.24,
CN 108229103 A,2018.06.29,
CN 108690871 A,2018.10.23,
WO 2011143231 A2,2011.11.17,
WO 2015073711 A1,2015.05.21,
CN 108280325 A,2018.07.13,
CN 109033749 A,2018.12.18,
CN 109295198 A,2019.02.01,
Paul A. Hohenlohe et al..“Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing”.《molecular ecology》.2014,
Heng Li et al..“The Sequence Alignment/Map format and SAMtools”.《BIOINFORMATICS》.2009,

审查员 郭悦

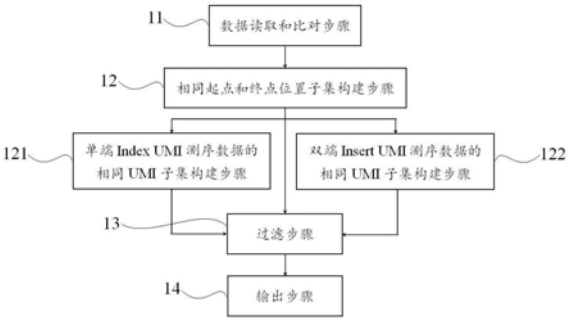
权利要求书2页 说明书9页 附图2页

(54)发明名称

一种校正高通量测序数据的方法和装置

(57)摘要

本申请公开了一种校正高通量测序数据的方法和装置。本申请的方法包括,将测序获得的read pair或read数据与参考基因组比对;将相同起点和终点位置的read pair或read分成一个Ai子集;比较每个子集中的read pair或read在基因组比对位置上的每一个碱基序列,根据预设的突变阈值去除重复和假阳性突变位点;最后输出高覆盖率的一致性数据,每一个子集只保留修正过的单一read pair或read。本申请的方法,能去除高通量测序中建库、杂交捕获和PCR产生的大量重复和假阳性突变,适用于去除癌症组织突变检测和液体活检等易产生假阳性突变的高深度测序,为提高检测质量和效率奠定了基础。



CN 109920480 B

1.一种校正高通量测序数据的方法,其特征在于:包括以下步骤,

数据读取和比对步骤,包括读取高通量测序数据,将测序获得的read pair或read数据与参考基因组比对;

相同起点和终点位置子集构建步骤,包括根据比对结果将具有相同起点和终点位置的read pair或read分成一个子集,标记为Ai子集,i为子集的编号;

过滤步骤,包括比较每个子集中的read pair或read在基因组比对位置上的每一个碱基序列,再根据预设的突变阈值去除重复和假阳性突变位点;

输出步骤,包括输出高覆盖率的一致性数据,每一个子集只保留修正过的单一read pair或read,即获得校正后的测序数据;

还包括相同UMI子集构建步骤,所述过滤步骤和输出步骤都以所述相同UMI子集构建步骤构建的子集为基础进行;

对于单端Index UMI测序数据,所述相同UMI子集构建步骤包括,根据所述相同起点和终点位置子集构建步骤构建的Ai子集,在一个Ai子集中将UMI序列相同的read pair或read分成一个Bi子集;并根据UMI代表的read pair或read数量将Bi子集进行降序排序,将数量最高的UMI依次与其它UMI比较,合并UMI序列相差不超过设定阈值的Bi子集;然后,对未合并的其它Bi子集重复前述排序、比较和合并过程,直到最后一个Bi子集;

对于双端Insert UMI测序数据,所述相同UMI子集构建步骤包括,根据所述相同起点和终点位置子集构建步骤构建的Ai子集,在一个Ai子集中将UMI序列相同或倒置序列相同的read pair或read分成一个Bi子集;并根据UMI代表的read pair或read数量将Bi子集进行降序排序,将数量最高的UMI依次与其它UMI比较,合并UMI序列或者倒置序列相差不超过设定阈值的Bi子集;然后,对未合并的其它Bi子集重复前述排序、比较和合并过程,直到最后一个Bi子集。

2.根据权利要求1所述的方法,其特征在于:所述过滤步骤,具体包括,将每个子集内的每条read pair或read与参考基因组比对,识别突变位点和基因型,并统计突变位点每种基因型出现的频率,将出现频率和所占比例大于预设的突变阈值的基因型作为突变位点的基因型,根据所确定的突变位点的基因型重构read pair或read。

3.根据权利要求2所述的方法,其特征在于:所述输出步骤,具体包括,根据每个子集中重构的read pair或read,计算每个read pair或read的质量值,及其与参考基因组的编辑距离,输出高质量的read pair或read。

4.一种校正高通量测序数据的装置,其特征在于:包括数据读取和比对模块、相同起点和终点位置子集构建模块、过滤模块和输出模块;

数据读取和比对模块,包括用于读取高通量测序数据,将测序获得的read pair或read数据与参考基因组比对;

相同起点和终点位置子集构建模块,包括用于根据比对结果将具有相同起点和终点位置的read pair或read分成一个子集,标记为Ai子集,i为子集的编号;

过滤模块,包括用于比较每个子集中的read pair或read在基因组比对位置上的每一个碱基序列,再根据预设的突变阈值去除重复和假阳性突变位点;

输出模块,包括用于输出高覆盖率的一致性数据,每一个子集只保留修正过的单一read pair或read,即获得校正后的测序数据;

还包括相同UMI子集构建模块;所述过滤模块和输出模块都以所述相同UMI子集构建模块构建的子集为基础进行;

对于单端Index UMI测序数据,所述相同UMI子集构建模块,包括用于根据所述相同起点和终点位置子集构建模块构建的Ai子集,在一个Ai子集中将UMI序列相同的read pair或read分成一个Bi子集;并根据UMI代表的read pair或read数量将Bi子集进行降序排序,将数量最高的UMI依次与其它UMI比较,合并UMI序列相差不超过设定阈值的Bi子集;然后,对未合并的其它Bi子集重复前述排序、比较和合并过程,直到最后一个Bi子集;

对于双端Insert UMI测序数据,所述相同UMI子集构建模块,包括用于根据所述相同起点和终点位置子集构建模块构建的Ai子集,在一个Ai子集中将UMI序列相同或倒置序列相同的read pair或read分成一个Bi子集;并根据UMI代表的read pair或read数量将Bi子集进行降序排序,将数量最高的UMI依次与其它UMI比较,合并UMI序列或者倒置序列相差不超过设定阈值的Bi子集;然后,对未合并的其它Bi子集重复前述排序、比较和合并过程,直到最后一个Bi子集。

5.根据权利要求4所述的装置,其特征在于:所述过滤模块,具体包括用于将每个子集内的每条read pair或read与参考基因比对,识别突变位点和基因型,并统计突变位点每种基因型出现的频率,将出现频率和所占比例大于预设的突变阈值的基因型作为突变位点的基因型,根据所确定的突变位点的基因型重构read pair或read。

6.根据权利要求5所述的装置,其特征在于:所述输出模块,具体包括用于根据每个子集中重构的read pair或read,计算每个read pair或read的质量值,及其与参考基因组的编辑距离,输出高质量的read pair或read。

7.一种校正高通量测序数据的装置,其特征在于:包括存储器和处理器;

所述存储器用于存储程序;

所述处理器用于通过执行所述存储器存储的程序实现权利要求1-3任一项所述的方法。

8.一种计算机可读存储介质,其特征在于:包括存储于其中的程序,所述程序能够被处理器执行以实现权利要求1-3任一项所述的方法。

一种校正高通量测序数据的方法和装置

技术领域

[0001] 本申请涉及高通量测序数据校正领域，特别是涉及一种校正高通量测序数据的方法和装置。

背景技术

[0002] 随着二代测序技术的发展，高深度测序在肿瘤突变检测、液体活检领域应用越来越广泛。尤其是以外周血游离DNA(缩写cfDNA)为主的突变检测成为癌症早期筛查和癌症临床治疗的重要辅助手段。虽然，随着肿瘤进展，癌症患者的外周血游离肿瘤DNA(缩写ctDNA)含量明显升高，但是大部分患者ctDNA含量的比例在0.5-5％之间，加之高通量测序在建库实验和测序过程中会引入大量的错误，导致目前检测肿瘤来源的体细胞突变难度依旧极大。

[0003] 目前能够进行ctDNA检测的方法包括基于聚合酶链式(缩写PCR)反应的BEAMing方法和微滴式数字PCR(缩写ddPCR)，以及高深度测序和通过加入UMI(即unique molecular identifier单分子编码)提高准确性和敏感性的深度测序技术。

[0004] 其中，高深度测序和UMI深度测序技术都是依赖于高通量测序进行ctDNA检测；特别是通过给每个原始的DNA模板加入特殊的分子标签序列进行高通量测序，能够提高后续数据分析的准确性，加强基因检测在临床实践的指导作用。

[0005] 但是，如前面提到的，ctDNA含量较低，需要采用PCR扩增富集建库，这个过程中会引入大量的PCR重复和假阳性，影响检测结果的准确性和重复性。因此，目前亟需一种对高深度测序或添加了分子标签的高深度测序结果进行校正的方法，以去除突变检测中PCR重复和建库实验过程中引入的假阳性。

发明内容

[0006] 本申请的目的是提供一种新的校正高通量测序数据的方法和装置。

[0007] 为了实现上述目的，本申请采用了以下技术方案：

[0008] 本申请的一方面公开了一种校正高通量测序数据的方法，包括以下步骤，

[0009] 数据读取和比对步骤，包括读取高通量测序数据，将测序获得的read pair或read数据与参考基因组比对；

[0010] 相同起点和终点位置子集构建步骤，包括根据比对结果将具有相同起点和终点位置的readpair或read分成一个子集，标记为Ai子集，i为子集的编号；

[0011] 过滤步骤，包括比较每个子集中的readpair或read在基因组比对位置上的每一个碱基序列，再根据预设的突变阈值去除重复和假阳性突变位点；

[0012] 输出步骤，包括输出高覆盖率的一致性数据，每一个子集只保留修正过的单一readpair或read，即获得校正后的测序数据。

[0013] 需要说明的是，本申请的关键在于将测序数据分成若干子集，并分别对子集进行过滤，去除重复和假阳性突变位点，使得最终输出的测序数据具有覆盖率高、一致性好等优

点。通过本申请的方法，去除了大量的PCR重复和假阳性，提高了高通量测序检测的准确性和重复性。可以理解，本申请的方法尤其适用于去除癌症组织突变检测和液体活检等易产生假阳性突变的高深度测序。

[0014] 优选的，本申请的方法还包括相同UMI子集构建步骤，过滤步骤和输出步骤都以该相同UMI子集构建步骤构建的子集为基础进行；对于单端Index UMI测序数据，该相同UMI子集构建步骤包括，根据相同起点和终点位置子集构建步骤构建的Ai子集，在一个Ai子集中将UMI序列相同的readpair或read分成一个Bi子集；并根据UMI代表的readpair或read数量将Bi子集进行降序排序，将数量最高的UMI依次与其它UMI比较，合并UMI序列相差不超过设定阈值的Bi子集；然后，对未合并的其它Bi子集重复前述排序、比较和合并过程，直到最后一个Bi子集；

[0015] 对于双端Insert UMI测序数据，该相同UMI子集构建步骤包括，根据相同起点和终点位置子集构建步骤构建的Ai子集，在一个Ai子集中将UMI序列相同或倒置序列相同的readpair或read分成一个Bi子集；并根据UMI代表的read pair或read数量将Bi子集进行降序排序，将数量最高的UMI依次与其它UMI比较，合并UMI序列或者倒置序列相差不超过设定阈值的Bi子集；然后，对未合并的其它Bi子集重复前述排序、比较和合并过程，直到最后一个Bi子集。

[0016] 需要说明的是，相同UMI子集构建步骤是针对加入UMI的深度测序技术获得的高通量测试数据而言的，如果高通量测序中没有添加UMI，则不需要该步骤。可以理解，加入UMI的深度测序技术又分为单端Index UMI测序技术和双端Insert UMI测序技术，因此，相应的相同UMI子集构建步骤也会有所区别。

[0017] 还需要说明的是，对于双端Insert UMI测序数据，UMI倒置序列相同是指前后两端的Insert标签互换的情况，例如ATC...GGA和GGA...ATC，如果将“GGA...ATC”前后两端的Insert标签互换过来，即将“ATC”置于前面，将“GGA”置于后面，就和“ATC...GGA”相同了。

[0018] 优选的，过滤步骤，具体包括，将每个子集内的每条read pair或read与参考基因组比对，识别突变位点和基因型，并统计突变位点每种基因型出现的频率，将出现频率和所占比例大于预设的突变阈值的基因型作为突变位点的基因型，根据所确定的突变位点的基因型重构readpair或read。

[0019] 优选的，输出步骤，具体包括，根据每个子集中重构的read pair或read，计算每个readpair或read的质量值，及其与参考基因组的编辑距离，输出高质量的readpair或read。

[0020] 本申请的另一面公开了一种校正高通量测序数据的装置，包括数据读取和比对模块、相同起点和终点位置子集构建模块、过滤模块和输出模块；

[0021] 数据读取和比对模块，包括用于读取高通量测序数据，将测序获得的read pair或read数据与参考基因组比对；

[0022] 相同起点和终点位置子集构建模块，包括用于根据比对结果将具有相同起点和终点位置的readpair或read分成一个子集，标记为Ai子集，i为子集的编号；

[0023] 过滤模块，包括用于比较每个子集中的readpair或read在基因组比对位置上的每一个碱基序列，再根据预设的突变阈值去除重复和假阳性突变位点；

[0024] 输出模块，包括用于输出高覆盖率的一致性数据，每一个子集只保留修正过的单一readpair或read，即获得校正后的测序数据。

[0025] 优选的，本申请的装置还包括相同UMI子集构建模块；过滤模块和输出模块都以相同UMI子集构建模块构建的子集为基础进行；

[0026] 对于单端Index UMI测序数据，相同UMI子集构建模块，包括用于根据相同起点和终点位置子集构建模块构建的Ai子集，在一个Ai子集中将UMI序列相同的readpair或read分成一个Bi子集；并根据UMI代表的read pair或read数量将Bi子集进行降序排序，将数量最高的UMI依次与其它UMI比较，合并UMI序列相差不超过设定阈值的Bi子集；然后，对未合并的其它Bi子集重复前述排序、比较和合并过程，直到最后一个Bi子集；

[0027] 对于双端Insert UMI测序数据，相同UMI子集构建模块，包括用于根据相同起点和终点位置子集构建模块构建的Ai子集，在一个Ai子集中将UMI序列相同或倒置序列相同的readpair或read分成一个Bi子集；并根据UMI代表的read pair或read数量将Bi子集进行降序排序，将数量最高的UMI依次与其它UMI比较，合并UMI序列或者倒置序列相差不超过设定阈值的Bi子集；然后，对未合并的其它Bi子集重复前述排序、比较和合并过程，直到最后一个Bi子集。

[0028] 优选的，过滤模块，具体包括用于将每个子集内的每条readpair或read与参考基因比对，识别突变位点和基因型，并统计突变位点每种基因型出现的频率，将出现频率和所占比例大于预设的突变阈值的基因型作为突变位点的基因型，根据所确定的突变位点的基因型重构readpair或read。

[0029] 优选的，输出模块，具体包括用于根据每个子集中重构的readpair或read，计算每个readpair或read的质量值，及其与参考基因组的编辑距离，输出高质量的readpair或read。

[0030] 需要说明的是，本申请校正高通量测序数据的装置，实际上就是通过各个模块实现本申请校正高通量测序数据的方法的各个步骤，以实现自动化校正。因此，本申请装置中的特征可以参考本申请的校正高通量测序数据的方法。

[0031] 可以理解，本申请校正高通量测序数据的方法，其全部或部分功能可以通过硬件的方式实现，也可以通过计算机程序的方式实现。当通过计算机程序的方式实现时，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：只读存储器、随机存储器、磁盘、光盘、硬盘等，通过计算机执行该程序以实现本申请的方法。例如，将程序存储在设备的存储器中，当通过处理器执行存储器中程序，即可实现本申请的方法。当本申请的方法中全部或部分功能通过计算机程序的方式实现时，该程序也可以存储在服务器、另一计算机、磁盘、光盘、闪存盘或移动硬盘等存储介质中，通过下载或复制保存到本地设备的存储器中，或对本地设备的系统进行版本更新，然后在处理器执行存储器中的程序时，即可实现本申请校正高通量测序数据的方法的全部或部分功能。

[0032] 因此，本申请的再一面还公开了一种校正高通量测序数据的装置，其包括存储器和处理器；存储器用于存储程序；处理器用于通过执行存储器存储的程序实现本申请的校正高通量测序数据的方法。

[0033] 本申请的再一面还公开了一种计算机可读存储介质，包括存储于其中的程序，该程序能够被处理器执行以实现本申请的校正高通量测序数据的方法。

[0034] 由于采用以上技术方案，本申请的有益效果在于：

[0035] 本申请校正高通量测序数据的方法，能够去除高通量测序中建库、杂交捕获和PCR

产生的大量重复和假阳性突变，提高了高通量测序检测的准确性和重复性，尤其适用于去除癌症组织突变检测和液体活检等易产生假阳性突变的高深度测序，为提高检测质量和效率奠定了基础。

附图说明

[0036] 图1是本申请实施例中校正高通量测序数据的方法的流程框图；

[0037] 图2是本申请实施例中校正高通量测序数据的装置的结构框图；

[0038] 图3是本申请实施例中校正处理之前的测序数据质量分析图；

[0039] 图4是本申请实施例中校正处理之后的测序数据质量分析图。

具体实施方式

[0040] 对高通量测序数据进行研究发现，将readpair或read与参照基因组mapping的过程当中，来自于相同原始DNA模板的readpair或read在参照基因组的比对起始和终止位置应该是相同的。在单端IndexUMI测序中，来自于相同原始DNA模板的readpair或read在PCR过程中携带的UMI也应该相同的，或者，在宽松的判定中可以容许阈值以内的错误碱基。在双端InsertUMI测序中，来自于相同原始双链DNA模板的readpair或read在PCR过程中携带的UMI也应该相同或者顺序相反序列相同的，或者可以容许阈值以内的错误碱基。因此，通过比较readpair或read比对的起点和终点，以及UMI的序列特征，可以识别出哪些readpair或read来自于同一条原始DNA模板。然后对来自于同一个DNA模板或者模板双链的readpair或read序列进行校正，可以有效的去除建库以及实验过程中引入的假阳性。

[0041] 基于以上研究和认识，本申请提出了一种校正高通量测序数据的方法，如图1所示，图1展示了三种方案，即针对非UMI测序技术获得的测序数据的校正方法、针对单端Index UMI测序技术获得的测序数据的校正方法和针对双端InsertUMI测序技术获得的测序数据的校正方法。

[0042] 如图1所示，针对非UMI测序技术获得的测序数据的校正方法包括数据读取和比对步骤11、相同起点和终点位置子集构建步骤12、过滤步骤13和输出步骤14；针对单端IndexUMI测序技术获得的测序数据的校正方法同样包括数据读取和比对步骤11、相同起点和终点位置子集构建步骤12、过滤步骤13和输出步骤14，并且在相同起点和终点位置子集构建步骤12之后增加了单端Index UMI测序数据的相同UMI子集构建步骤121，然后再根据构建的Bi子集进行过滤步骤13和输出步骤14；针对双端InsertUMI测序技术获得的测序数据的校正方法，与单端Index UMI测序数据类似，在相同起点和终点位置子集构建步骤12之后增加了双端InsertUMI测序数据的相同UMI子集构建步骤122，然后再根据构建的Bi子集进行过滤步骤13和输出步骤14。

[0043] 以上三种方案中，数据读取和比对步骤11，包括读取高通量测序数据，将测序获得的readpair或read数据与参考基因组比对；该步骤主要是对每一个read pair或read进行分析，识别readpair或read比对的染色体、起点、终点，以便于后续Ai子集的建立。

[0044] 相同起点和终点位置子集构建步骤12，包括根据比对结果将具有相同起点和终点位置的readpair或read分成一个子集，标记为Ai子集，i为子集的编号。在非UMI测序中，每个Ai子集中的readpair或read来自于同一个DNA分子模板或者原始DNA分子双链。

[0045] 单端Index UMI测序数据的相同UMI子集构建步骤121，包括根据相同起点和终点位置子集构建步骤构建的Ai子集，在一个Ai子集中将UMI序列相同的readpair或read分成一个Bi子集；并根据UMI代表的readpair或read数量将Bi子集进行降序排序，将数量最高的UMI依次与其它UMI比较，合并UMI序列相差不超过设定阈值的Bi子集；然后，对未合并的其它Bi子集重复前述排序、比较和合并过程，直到最后一个Bi子集。对于单端Index UMI测序数据，在与参考基因组比对时，还会识别readpair或read上携带的UMI标签序列，UMI为单端Index标签如ATCGACGT；在同一个Ai子集中根据每条readpair或read所带有的UMI是否一样将其分成子集Bi，其中i是Bi子集的编号，例如i＝1,2,3,4,……；在单端IndexUMI测序数据中，合并后的每个Bi子集中的所有readpair或read来自于同一个原始DNA分子。

[0046] 双端Insert UMI测序数据的相同UMI子集构建步骤122，包括根据相同起点和终点位置子集构建步骤构建的Ai子集，在一个Ai子集中将UMI序列相同或倒置序列相同的readpair或read分成一个Bi子集；并根据UMI代表的readpair或read数量将Bi子集进行降序排序，将数量最高的UMI依次与其它UMI比较，合并UMI序列或者倒置序列相差不超过设定阈值的Bi子集；然后，对未合并的其它Bi子集重复前述排序、比较和合并过程，直到最后一个Bi子集。同样的，对于双端InsertUMI测序数据，在与参考基因组比对时，也会识别readpair或read上携带的UMI标签序列，UMI是双端Insert标签如ATC...TCC，在同一个Ai子集中根据每条readpair或read所带有的UMI是否一样或者倒置后一样将其分成子集Bi，其中i是Bi子集的编号，例如i＝1,2,3,4,……；在双端InsertUMI测序数据中，合并后的每个子集Bi中的所有readpair或read来自于同一个原始DNA分子。

[0047] 过滤步骤13，如果是非UMI测序技术，则直接对每个Ai子集中的readpair或read进行处理；如果是单端或双端UMI测序技术，则对每个合并后的Bi子集中的readpair或read进行处理；过滤步骤13包括比较每个子集中的readpair或read在基因组比对位置上的每一个碱基序列，再根据预设的突变阈值去除重复和假阳性突变位点。

[0048] 具体的，本申请的一种实现方式中，对每个Ai子集中的readpair或read进行处理，包括以下步骤：

[0049] 首先判断该子集readpair或read统一的比对起点和终点，然后读取该比对区间范围内的参考基因组序列；

[0050] 对该子集内每条readpair或read进行处理，识别哪些位点与参考基因组相比发生了突变，突变后的基因型是什么；同时统计出现突变的位点每种基因型出现的频率；

[0051] 对该比对区间内发生突变的位点进行处理，比较该位点每种基因型出现的频率，当该某种突变基因型出现频率大于设定的阈值时，例如基因型出现次数大于等于2，并且所占的比例大于等于50％，则认为该位点基因型为该突变型，否则设为野生型；

[0052] 重新构建高质量readpair或read序列：随机选取子集中的一对readpair或read，分别进行处理，在read pair或read比对的区间范围内，除了在上一步识别为突变型的位点，其他位点均置为野生型，重新构建read序列，计算该read比对的质量值，CIGAR值，和参考基因组编辑距离。本申请的一种实现方式中无论校正前碱基质量值为多少，经过校正后的碱基质量均置为最高值40。

[0053] 本申请的一种实现方式中，对每个Bi子集中的readpair或read进行处理，包括以下步骤：

[0054] 首先判断该子集readpair或read统一的比对起点和终点，然后读取该比对区间范围内的参考基因组序列；

[0055] 对该子集内每条readpair或read进行处理，识别哪些位点与参考基因组相比发生了突变，突变后的基因型是什么；同时统计出现突变的位点每种基因型出现的频率；

[0056] 对该比对区间内发生突变的位点进行处理，比较该位点每种基因型出现的频率，当该某种突变基因型出现频率大于设定的阈值时，例如基因型出现次数大于等于2，并且所占的比例大于等于80％，则认为该位点基因型为该突变型，否则设为野生型；

[0057] 重新构建高质量readpair或read序列：随机选取子集中的一对readpair或read，分别进行处理，在read pair或read比对的区间范围内，除了在上一步识别为突变型的位点，其他位点均置为野生型，重新构建read序列，计算该read比对的质量值，CIGAR值，和参考基因组编辑距离。

[0058] 输出步骤14，包括输出高覆盖率的一致性数据，每一个Ai子集或Bi子集只保留修正过的单一readpair或read，即获得校正后的测序数据。

[0059] 可以理解，每个Ai子集中的readpair或read来自于同一个DNA分子模板或者原始DNA分子双链；同样的，合并后的每个Bi子集中的所有read pair或read来自于同一个原始DNA分子；因此，每个Ai子集或Bi子集只保留一个覆盖率最高的校正后的readpair或read，就可以去除大量的PCR重复，并去除假阳性突变。

[0060] 基于本申请的校正高通量测序数据的方法，本申请进一步提出了一种校正高通量测序数据的装置，如图2所示，该装置包括数据读取和比对模块21、相同起点和终点位置子集构建模块22、过滤模块23和输出模块24。而针对单端Index UMI测序技术获得的测序数据的校正方法和双端Insert UMI测序技术获得的测序数据的校正方法，本申请的装置进一步的还包括相同UMI子集构建模块221。本申请装置中的各模块分别用于执行本申请校正高通量测序数据的方法中相应的各个步骤。具体的，数据读取和比对模块21，包括用于读取高通量测序数据，将测序获得的readpair或read数据与参考基因组比对；相同起点和终点位置子集构建模块22，包括用于根据比对结果将具有相同起点和终点位置的readpair或read分成一个子集，标记为Ai子集，i为子集的编号；过滤模块23，包括用于比较每个子集中的readpair或read在基因组比对位置上的每一个碱基序列，再根据预设的突变阈值去除重复和假阳性突变位点；输出模块24，包括用于输出高覆盖率的一致性数据，每一个子集只保留修正过的单一readpair或read，即获得校正后的测序数据。其中，相同UMI子集构建模块221，在处理单端Index UMI测序数据时，用于根据相同起点和终点位置子集构建模块构建的Ai子集，在一个Ai子集中将UMI序列相同的readpair或read分成一个Bi子集；并根据UMI代表的readpair或read数量将Bi子集进行降序排序，将数量最高的UMI依次与其它UMI比较，合并UMI序列相差不超过设定阈值的Bi子集；然后，对未合并的其它Bi子集重复前述排序、比较和合并过程，直到最后一个Bi子集。相同UMI子集构建模块221，在处理双端Insert UMI测序数据时，用于根据相同起点和终点位置子集构建模块构建的Ai子集，在一个Ai子集中将UMI序列相同或倒置序列相同的readpair或read分成一个Bi子集；并根据UMI代表的readpair或read数量将Bi子集进行降序排序，将数量最高的UMI依次与其它UMI比较，合并UMI序列或者倒置序列相差不超过设定阈值的Bi子集；然后，对未合并的其它Bi子集重复前述排序、比较和合并过程，直到最后一个Bi子集。

[0061] 本申请中涉及的名词解释如下：

[0062] readpair或read:是pair end文库的测序结果,read pair中read1是从核苷酸序列的5′端的测序结果,read2是从核苷酸序列的3′端的测序结果。

[0063] cfDNA:外周血当中游离的DNA。

[0064] ctDNA:外周血中游离的肿瘤细胞释放的DNA。

[0065] 单端IndexUMI:用来标记每一个测序分子的标签,在单端测序引物的Index中。

[0066] 双端InsertUMI:用来标记每一个测序分子的标签,在DNA分子的两端。

[0067] 下面通过具体实施例和附图对本申请作进一步详细说明。以下实施例仅对本申请进行进一步说明,不应理解为对本申请的限制。

[0068] 实施例

[0069] 本例分别针对非UMI测序技术、单端IndexUMI测序技术和双端InsertUMI测序技术,详细介绍了针对不同技术的校正高通量测序数据的方法,具体如下：

[0070] 方法一:针对非UMI测序技术获得的高通量测序数据的校正方法

[0071] 1.读取测序数据与参考基因组比对排序之后的结果文件,同时读取参考基因组序列文件。

[0072] 2.对每一个readpair或read进行分析,识别readpair或read比对的染色体、起点、终点。

[0073] 3.根据readpair或read是否具有相同的起始和终止位置分成不同的集Ai其中,i＝1,2,3,4,……。在非UMI测序中,每个Ai中的readpair或read来自于同一个DNA分子模板或者原始DNA分子双链。

[0074] 4.对每个Ai子集中的readpair或read进行处理。

[0075] 4.1首先判断该子集readpair或read统一的比对起点和终点,然后读取该比对区间范围内的参考基因组序列。

[0076] 4.2对该子集内每条readpair或read进行处理,识别哪些位点与参考基因组相比发生了突变,突变后的基因型是什么;同时统计出现突变的位点每种基因型出现的频率。

[0077] 4.3对该比对区间内发生突变的位点进行处理,比较该位点每种基因型出现的频率,当该某种突变基因型出现频率大于设定的阈值时,例如基因型出现次数大于等于2,并且所占的比例大于等于50％,则认为该位点基因型为该突变型,否则设为野生型。

[0078] 4.4重新构建高质量readpair或read序列。随机选取子集中的一对readpair或read,分别read进行处理,在read比对的区间范围内,除了在上一步识别为突变型的位点,其他位点均置为野生型,重新构建read序列,计算该read比对的质量值,CIGAR值,和参考基因组编辑距离。

[0079] 4.5输出经过校正的高质量readpair或read。

[0080] 方法二:针对单端Index UMI测序技术获得的高通量测序数据的校正方法

[0081] 1.读取测序数据与参考基因组比对排序之后的结果文件,同时读取参考基因组序列文件。

[0082] 2.对每一个readpair或read进行分析,识别readpair或read比对的染色体、起点、终点。该步骤同时会识别readpair或read上携带的UMI标签序列。UMI为单端Index标签如ATCGACGT。

[0083] 3.根据readpair或read是否具有相同的起始和终止位置分成不同的集Ai,其i＝1,2,3,4,……。然后在同一个Ai集中根据每条readpair或read所带有的UMI是否一样将其分成子集Bi,其i＝1,2,3,4,……;并根据每条UMI代表的readpair或read数量将Bi集(i＝1,2,3,4,……)进行降序排序,将数量最高的UMI依次与其他UMI比较,合并UMI序列相差不超过设定阈值的子集,然后在合并的子集之外重复上述排序、比较、合并过程,直到最后一个UMI子集;在Index单分子编码测序中,合并后的每个子集Bi中的所有readpair或read来自于同一个原始DNA分子。

[0084] 4.对每个合并后的Bi子集中的readpair或read进行处理。

[0085] 4.1首先判断该子集readpair或read统一的比对起点和终点,然后读取该比对区间范围内的参考基因组序列。

[0086] 4.2对该子集内每条readpair或read进行处理,识别哪些位点与参考基因组相比发生了突变,突变后的基因型是什么;同时统计出现突变的位点每种基因型出现的频率。

[0087] 4.3对该比对区间内发生突变的位点进行处理,比较该位点每种基因型出现的频率,当该某种突变基因型出现频率大于设定的阈值时,例如基因型出现次数大于等于2,并且所占的比例大于等于80％,则认为该位点基因型为该突变型,否则设为野生型。

[0088] 4.4重新构建高质量readpair或read序列。随机选取子集中的一对readpair或read,分别read进行处理,在read比对的区间范围内,除了在上一步识别为突变型的位点,其他位点均置为野生型,重新构建read序列,计算该read比对的质量值,CIGAR值,和参考基因组编辑距离。

[0089] 4.5输出经过校正的高质量readpair或read。

[0090] 方法三:针对双端InsertUMI测序技术获得的高通量测序数据的校正方法

[0091] 1.读取测序数据与参考基因组比对排序之后的结果文件,同时读取参考基因组序列文件。

[0092] 2.对每一个readpair或read进行分析,识别readpair或read比对的染色体、起点、终点。该步骤同时会识别readpair或read上携带的UMI标签序列。UMI是双端Insert标签如ATC_TCC。

[0093] 3.根据read pair或read是否具有相同的起始和终止位置分成不同的集Ai(i＝1,2,3,4,……)。然后在同一个Ai集中根据每条readpair或read所带有的UMI是否一样或者倒置一样,例如(ATC_GGA和GGA_ATC,将其分成子集Bi;并根据UMI代表的readpair或read数量将Bi集(i＝1,2,3,4,……)进行降序排序,将数量最高的UMI依次与其他UMI比较,合并UMI序列或者倒置UMI序列相差不超过设定阈值的子集,然后在合并的子集之外重复上述排序、比较、合并过程,直到最后一个UMI子集;在Insert单分子编码测序中,合并后的每个子集Bi中的所有readpair或read来自于同一个原始DNA分子。

[0094] 4.对每个合并后Bi子集中的readpair或read进行处理。

[0095] 4.1首先判断该子集readpair或read统一的比对起点和终点,然后读取该比对区间范围内的参考基因组序列。

[0096] 4.2对该子集内每条readpair或read进行处理,识别哪些位点与参考基因组相比发生了突变,突变后的基因型是什么。同时统计出现突变的位点每种基因型出现的频率。

[0097] 4.3对该比对区间内发生突变的位点进行处理,比较该位点每种基因型出现的频

率，当该某种突变基因型出现频率大于设定的阈值时，例如基因型出现次数大于等于2，并且所占的比例大于等于80％，则认为该位点基因型为该突变型，否则设为野生型。

[0098] 4.4重新构建高质量readpair或read序列。随机选取子集中的一对readpair或read，分别read进行处理，在read比对的区间范围内，除了在上一步识别为突变型的位点，其他位点均置为野生型，重新构建read序列，计算该read比对的质量值，CIGAR值，和参考基因组编辑距离。

[0099] 4.5输出经过校正的高质量readpair或read。

[0100] 本例采用以上方法，具体对Horizon公司HD778标准品数据进行了校正，并对比分析了校正前后的数据比对结果中错误背景噪声。结果如图3和图4所示，图3是校正前的测序比对结果，图4是校正后输出的测序数据比对结果。对比图3和图4的结果可见，经过本例校正的高通量测序数据比对结果中，即图4中，错误背景噪声几乎全部被去除。

[0101] 以上内容是结合具体的实施方式对本申请所作的进一步详细说明，不能认定本申请的具体实施只局限于这些说明。对于本申请所属技术领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干简单推演或替换，都应当视为属于本申请的保护范围。

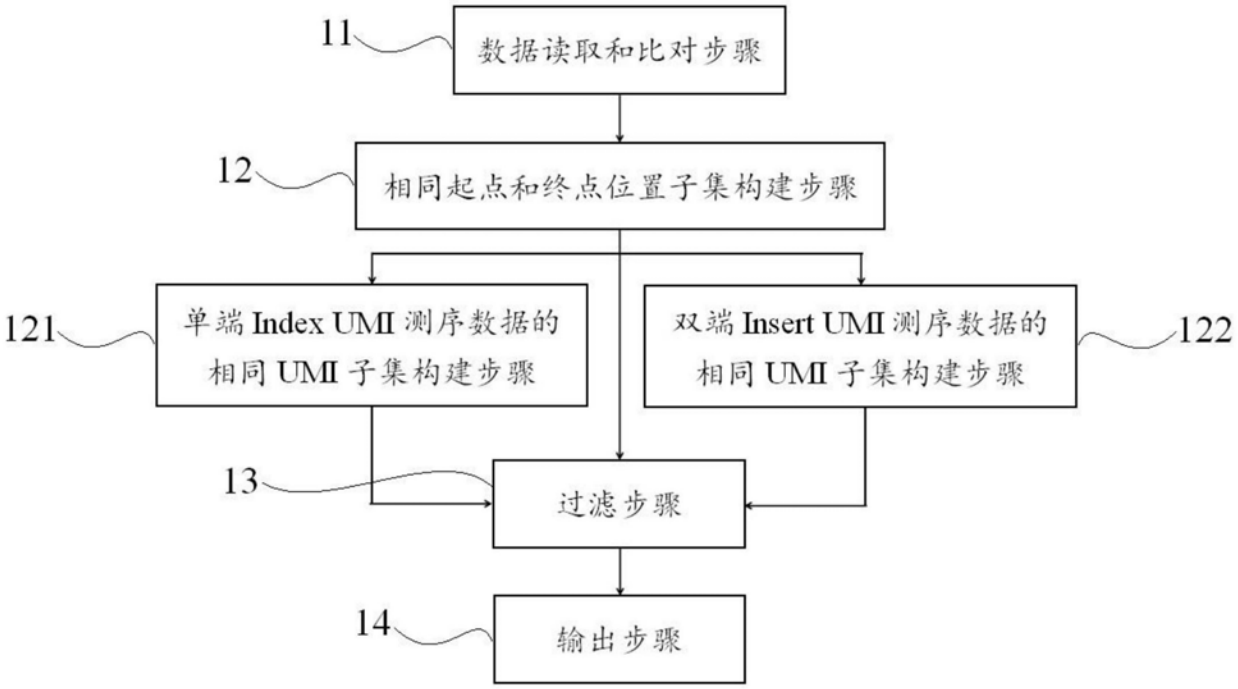

11
数据读取和比对步骤
12
相同起点和终点位置子集构建步骤
121
单端 Index UMI 测序数据的
相同 UMI 子集构建步骤
双端 Insert UMI 测序数据的
相同 UMI 子集构建步骤
122
13
过滤步骤
14
输出步骤

图1

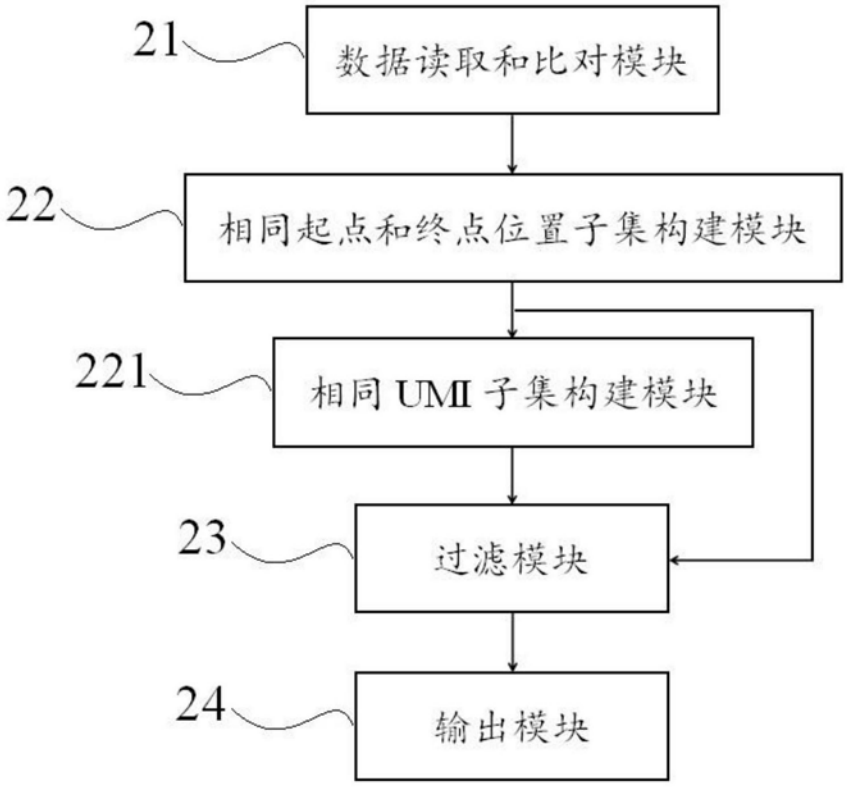

21
数据读取和比对模块
22
相同起点和终点位置子集构建模块
221
相同 UMI 子集构建模块
23
过滤模块
24
输出模块

图2

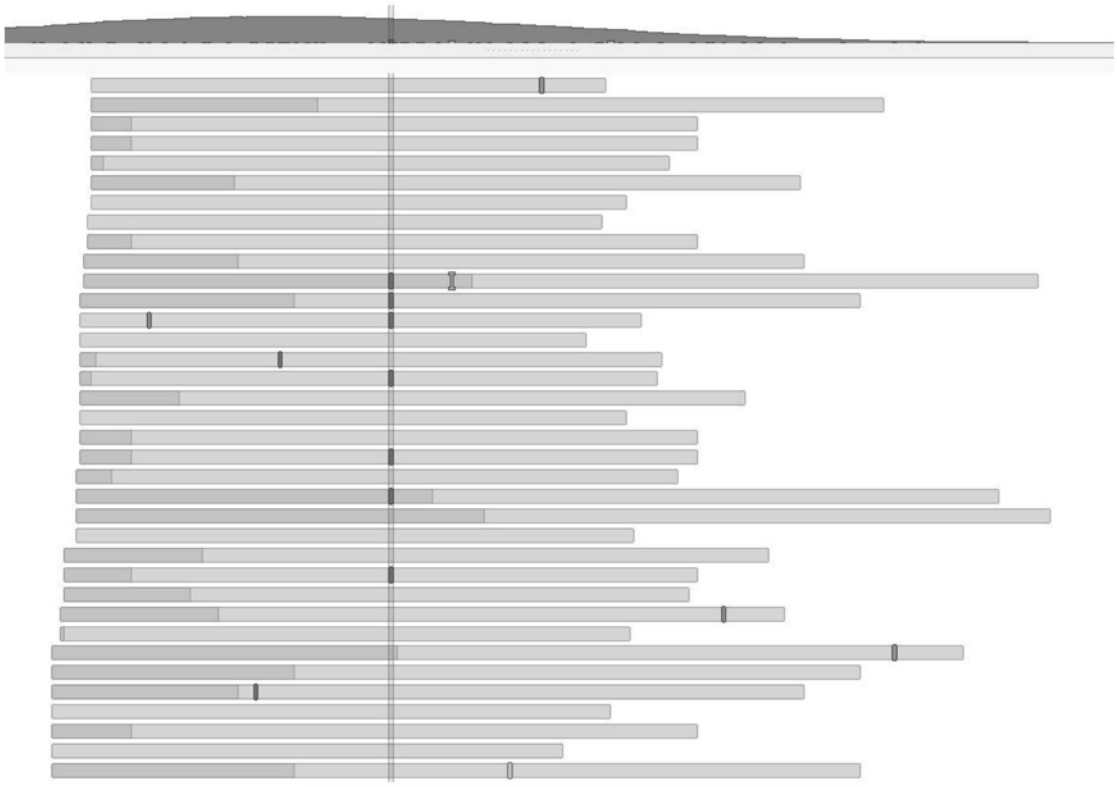


图3

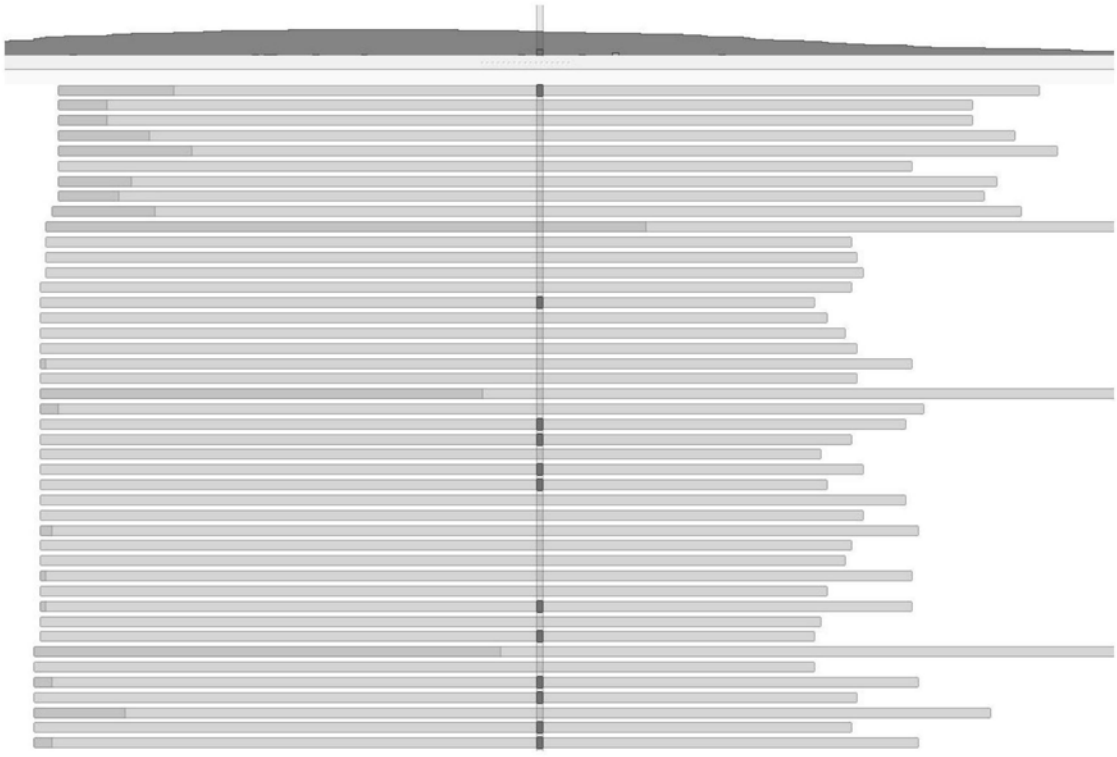


图4