(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 114005489 B
(45)授权公告日 2022.03.22

(21)申请号 202111616129.1
(22)申请日 2021.12.28
(65)同一申请的已公布的文献号
申请公布号 CN 114005489 A
(43)申请公布日 2022.02.01
(73)专利权人 成都齐碳科技有限公司
地址 610041 四川省成都市中国(四川)自由贸易试验区成都高新区天府五街200号7栋A区2楼
(72)发明人 郎继东 孙继国
(74)专利代理机构 北京东方亿思知识产权代理有限责任公司 11258
代理人 臧静
(51)Int.Cl.
G16B 30/10(2019.01)
G16B 20/50(2019.01)
(56)对比文件
CN 109033749 A,2018.12.18
CN 112309502 A,2021.02.02
CN 113096728 A,2021.07.09
Jacopo Oieni 等.Nano-ghosts: Novel biomimetic nano-vesicles for the delivery of antisense oligonucleotides.《Journal of Controlled Release》.2021,第333卷
审查员 高婕

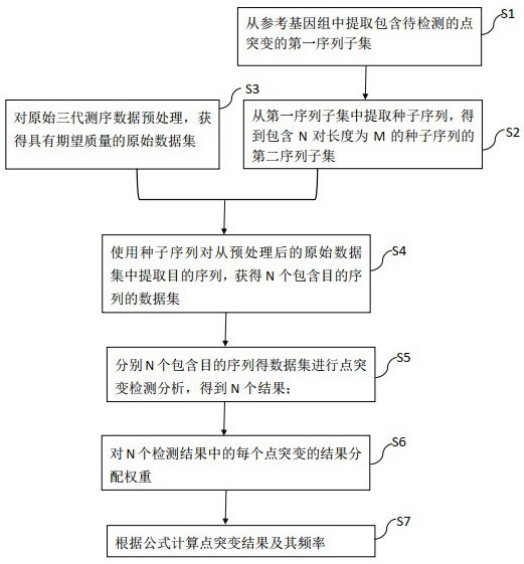
权利要求书3页 说明书10页 附图2页

(54)发明名称

基于三代测序数据检测点突变的分析方法和装置

(57)摘要

本发明提供了基于三代测序数据检测点突变的分析方法和装置。本发明的分析方法包括：1)提取包含待检测的点突变的第一序列子集；2)从第一序列子集中提取种子序列，获得第二序列子集；3)获得具有期望质量的原始数据集；4)使用第二序列子集的种子序列对，获得N个包含目的序列的数据集；5)对N个包含目的序列的数据集进行点突变检测分析；6)对N个检测结果中的每个点突变的结果分配权重W；7)根据公式计算点突变结果及其频率。本发明还提供了一种基于三代测序数据检测点突变的装置。使用本发明的方法，不仅从数据特征上有效地规避随机indel或较高测序错误导致的比对率不高引起的假阴性的问题，同时能更有效地控制假阳性的结果。



CN 114005489 B

1.一种基于三代测序数据检测点突变的分析方法,所述方法包括以下步骤:

1)从参考基因组中提取包含待检测的点突变的第一序列子集;

在所述参考基因组上进行固定长度L的短序列提取N次,所述短序列之间满足待检测的点突变在提取后的短序列上的位置与其在前一次提取的短序列上的位置之间具有固定距离D,并且$N\leqslant\frac{L-D}{D}$,其中,N、D、L均为整数,最终得到第一序列子集,其包含N个含有待检测的点突变的短序列;

2)从步骤1)的第一序列子集中提取种子序列,提取位置为每条短序列的首尾端各M个碱基,得到第二序列子集,其包含N对长度为M的种子序列;

3)对原始三代测序数据预处理,获得具有期望质量的原始数据集;

4)使用步骤2)获得的第二序列子集的种子序列对从步骤3)得到的原始数据集中提取目的序列,获得N个包含目的序列的数据集;

5)分别对步骤4)的N个包含目的序列的数据集进行点突变检测分析,得到N个结果;其中,每个结果包括待检测的位点的突变频率F,点突变的reads支持数AO,点突变位置的测序深度DP;

6)对步骤5)的N个检测结果中的每个点突变的结果分配权重W;

7)根据公式计算点突变结果及其频率;

$$F_{correct}=\frac{\text{(点突变支持数*权重)之和}}{\text{(点突变位置的测序深度*权重)之和}}$$

若$F_{correct}\geq 1\%$,则为阳性,反之为阴性,其中$F_{correct}$为最终该位点的检测突变频率。

2.根据权利要求1所述的方法,其中,在步骤1)中,$5\leqslant D\leqslant\frac{L}{2}$。

3.根据权利要求1所述的方法,其中,在步骤1)中,第一次提取的短序列中,待检测的点突变在短序列上的位置为D_0,第X次提取时,所述点突变在该第X次提取的短序列中的位置L_x满足$L_x=D_0+(X-1)D$;

其中,$M<D_0\leqslant\frac{L}{4}$。

4.根据权利要求1所述的方法,其中,L为76-151bp。

5.根据权利要求1所述的方法,其中,在步骤2)中,M≥5。

6.根据权利要求1所述的分析方法,其中,在步骤3)中,对原始三代测序数据进行数据预处理,包括过滤低质量以及过短的测序reads;

其中,所述低质量的阈值为Q5;和/或过短的测序reads的序列长度阈值为100bp。

7.根据权利要求1所述的分析方法,其中,在步骤4)中,所述目的序列的长度L'≤L+50。

8.根据权利要求1所述的分析方法,其中,在步骤5)中,所述分析使用GATK Best Practice分析流程。

9.根据权利要求1所述的分析方法,其中,在步骤6)中,对N个检测结果中的每个点突变的结果分配权重,包括:

权重W_1至W_N的总和为1;和

在步骤1)中获得的N条短序列中,点突变在所述短序列的固定长度L上的位置越邻近中

间，与所述短序列相关的检测结果分配的权重越大。

10.根据权利要求9所述的分析方法，其中，在步骤6)中，对N个检测结果中的每个点突变的结果分配权重，

其中，N为偶数时，第$\frac{N}{2}$个和第$\frac{N}{2}$+1个数据集具有最大的权重$W_{N/2}=W_{N/2+1}$，然后$W_N=W_1$，$W_{N-1}=W_2$，$W_{N-2}=W_3$，以此类推；

其中，N为奇数时，第$\frac{N+1}{2}$个数据集具有最大的权重$W_{N+1/2}$，然后$W_N=W_1$，$W_{N-1}=W_2$，$W_{N-2}=W_3$，以此类推。

11.一种基于三代测序数据检测点突变的分析方法，所述方法包括以下步骤：

1)从参考基因组中提取包含待检测的点突变的第一序列子集；

在所述参考基因组上进行固定长度L的短序列提取N次，第一次提取获得的短序列中，待检测的点突变所在位置为D_0，所述短序列之间满足待检测的点突变在提取后的短序列上的位置与其在前一次提取的短序列上的位置之间具有固定距离D，最终得到第一序列子集，其包含N个含有待检测的点突变的短序列；

其中，L为76-151bp之间的任意整数，D为8到15之间的任意整数，N为4到18之间的任意整数，D_0为5到14之间的任意整数；

2)从步骤1)得到的第一序列子集中的每条序列提取种子序列，提取位置分别为每条序列两端各M个碱基，最终得到N个种子序列对的第二序列子集，其中$5\leqslant M<D_0$；

3)对原始三代测序数据进行数据预处理，利用Porechop软件以及NanoFilt软件去除实验建库过程中加入的接头及条形码序列，过滤低质量以及过短的测序reads，得到具有期望质量的原始数据集；

4)根据步骤2)得到的种子序列对，从步骤3)得到的原始数据集中提取出相应的目的序列，所述目的序列长度$L'\leqslant L+50$，最终得到N个包含目的序列的数据集；

5)利用GATK Best Practice分析流程对步骤4)中得到的N个包含目的序列的数据集分别进行点突变检测分析，得到N个靶向位点检测的最终结果，记每个靶向位点检测的突变频率为F_N，该位点的突变reads支持数为AO_N，该位置的测序深度为DP_N；

6)步骤5)的N个检测结果中的每个点突变的结果分配权重，权重W_1至W_N的总和为1；

其中，N为偶数时，第$\frac{N}{2}$个和第$\frac{N}{2}$+1个数据集具有最大的权重$W_{N/2}=W_{N/2+1}$，然后$W_N=W_1$，$W_{N-1}=W_2$，$W_{N-2}=W_3$，以此类推；

其中，N为奇数时，第$\frac{N+1}{2}$个数据集具有最大的权重$W_{N+1/2}$，然后$W_N=W_1$，$W_{N-1}=W_2$，$W_{N-2}=W_3$，以此类推；

7)加权及纠错矫正步骤5)中得到的靶向点突变结果及其频率，定义

$$F_{correct}=\frac{W_1*AO_1+W_2*AO_2+\cdots....W_{N-1}*AO_{N-1}+W_N*AO_N}{W_1*DP_1+W_2*DP_2+\cdots....W_{N-1}*DP_{N-1}+W_N*DP_N}$$；$F_{correct}$为最终该位点的检测突变频率；

若$F_{correct}$≥1%,则为阳性,反之为阴性。

12.一种基于三代测序数据检测点突变的装置,其中,所述装置包括:

种子序列提取模块,用于从参考基因组中提取包含N个含有待检测的点突变的短序列的第一序列子集,然后从所述第一序列子集中提取包含种子序列对的第二序列子集;

预处理模块,用于对三代测序数据预处理,获得具有期望质量的原始数据集;

初级分析模块,用于使用第二序列子集的种子序列对从预处理后的原始数据集提取包含目的序列的数据集,获得N个包含目的序列的数据集,然后进行点突变检测分析并获得N个结果;其中,每个结果包括待检测的位点的突变频率F,点突变的reads支持数AO,点突变位置的测序深度DP;

高级分析模块,用于对得到的结果进一步加权及矫正,并获得最终的分析结果;及

报告模块,用于根据数据输出结果;

所述高级分析模块用于对N个检测结果中的每个点突变的结果分配权重W,根据公式计算点突变结果及其频率;

$$F_{correct}=\frac{\text{(点突变支持数*权重)之和}}{\text{(点突变位置的测序深度*权重)之和}}$$

若$F_{correct}$≥1%,则为阳性,反之为阴性,其中$F_{correct}$为最终该位点的检测突变频率;

所述报告模块用于输出点突变结果及其频率。

13.根据权利要求12所述的装置,其中,所述预处理模块用于过滤低质量以及过短的测序reads,包括Porechop软件以及NanoFilt软件。

14.根据权利要求12所述的装置,其中,所述初级分析模块包含GATK Best Practice分析流程。

15.根据权利要求12所述的装置,其中,所述高级分析模块包含用于对每个结果分配权重的程序或软件。

基于三代测序数据检测点突变的分析方法和装置

技术领域

[0001] 发明本属于测序技术和测序数据的生物信息技术分析领域，尤其涉及一种基于三代测序数据检测点突变的方法，本发明还涉及基于三代测序数据检测点突变的装置和系统。

背景技术

[0002] 点突变指只有一个碱基对发生改变。广义点突变可以是碱基替换，单碱基插入或碱基缺失；狭义点突变也称作单碱基替换(base substitution)。碱基替换又分为转换(transitions)和颠换(transversions)两类。目前常见的检测基因点突变的方法有PCR法、Sanger测序法(一代测序)和二代测序。PCR法具有敏感性高的特点，且技术已经成熟，但每对引物只能检测一种突变，无法同时检测太多样品和位点，通量较低。Sanger测序法的成本较低，但所需样品用量大，且对低频突变的检测敏感性低。二代测序具有通量高的特点，测序成本也在逐年下降，但目前检测点突变常用的方法工具检测特异性不高(如Varscan)，对低频的检测敏感性也偏低(如Mutect)，或者使用了局部组装步骤导致运行时间过长(如Mutect2)，不能很好地满足点突变检测的需求。

[0003] 第三代测序技术，又称三代测序技术(Third generation sequencing)或单分子实时DNA测序技术，是一种在DNA测序时，不需要经过PCR扩增即可实现对每一条DNA分子的单独测序的技术。目前第三代测序技术原理主要分为以Pacbio的SMRT技术为代表的单分子荧光测序以及以牛津纳米孔公司和齐碳科技公司的纳米孔电泳技术为代表的纳米孔测序。三代测序的主要的技术特点之一是实现了DNA聚合酶内在自身的反应速度，一秒可以测10个碱基，测序速度是化学法测序的2万倍；其二是实现了DNA聚合酶内在自身的延续性，一个反应就可以测非常长的序列；二代测序可以测到上百个碱基，但是三代测序就可以测几千个碱基。进一步地，三代测序对DNA或RNA分子进行实时测序时无需进行PCR扩增或化学标记，避免在操作过程中引入的错误突变，高度保真，并且测序速度可以达到DNA为450bp/秒，RNA为70nt/秒，整体可以达到几兆碱基级别的超长读长。

[0004] 目前基于三代测序检测点突变(包含胚系突变以及体细胞突变)的方法还不是很成熟，但全球范围内已经有一些研究课题组致力于开发一些算法来精确识别三代测序数据中的点突变(SNV和InDel)，例如发表于Nature Communications杂志上的加利福尼亚大学开发的结合隐马尔可夫链模型的Longshot方法(DOI: 10.1038/s41467-019-12493-y)，发表于Nature Machine Intelligence杂志上的香港大学开发的结合深度神经网络模型的Clair方法(doi: https://doi.org/10.1038/s42256-020-0167-4)，公开于bioRxiv上基于google团队的DeepVariant开发优化的PEPPER-Margin-DeepVariant方法(doi: https://doi.org/10.1101/2021.03.04.433952)等。这些研究成果不仅仅丰富了基于三代测序数据的突变检测手段，更重要的是为三代测序的广阔发展及广泛的实际应用提供了技术保障。

[0005] 然而，当前基于三代测序检测点突变在方法上还存在很大的挑战和问题。众所周知，三代测序的数据在单碱基识别的精准度上还存在一些问题，造成该问题的因素有很多，

比如样本质量,电流通过“motor”蛋白的稳定性及basecalling模型的精确度等,最终体现在数据层面上就是测序质量不高或测序错误的现状以及随机indel分布等的数据特征。故在基于三代测序的数据分析中,如何稳定地检出点突变并且还能较好地控制假阳性及假阴性的检测结果就显得尤为重要,其对检测算法的灵敏度及特异性的要求也提出了巨大的挑战。虽然现阶段有一些基于三代测序数据开发的检测点突变的方法(如上所述),但各自的缺点也非常明显,最主要的就是受限于测序质量以及依赖的比对算法或深度学习训练集的数据分布等,并且适用场景并不够广泛,鲁棒性(robust)不足。

[0006] 因此,对现有的基于三代测序数据检测点突变的分析方法进行进一步的改进,使其在稳定地检出点突变的同时,还能较好地控制假阳性及假阴性的问题,具有非常重要的意义。

发明内容

[0007] 因此,本发明的目的是针对现有技术的不足,提供一种基于三代测序数据检测点突变的分析方法,本发明提供的方法能够在数据分析层面上良好地解决了上述问题,不仅从数据特征上较为有效地规避掉随机indel或较高测序错误导致的比对率不高导致的假阴性的问题,同时设计结合碱基在测序序列位置上的“中间较准,两边较差”的理论观点、数据分析层面上的分子生物标签(UMI/UID)思想以及“权重”统计的方法对检测结果进行整体评估、纠错及矫正,更加有效地控制了假阳性的结果。

[0008] 本发明的目的是通过以下技术方案实现的:

[0009] 一方面,本发明提供了一种基于三代测序数据检测点突变的分析方法,所述方法包括以下步骤:

[0010] 1)从参考基因组中提取包含待检测的点突变的第一序列子集;

[0011] 在所述参考基因组上进行固定长度L的短序列提取N次,所述短序列之间满足待检测的点突变在提取后的短序列上的位置与其在前一次提取的短序列上的位置之间具有固定距离D, $N \leqslant \frac{L-D}{D}$;其中,N、D、L均为整数;最终得到第一序列子集,其包含N个含有待检测的点突变的短序列;

[0012] 2)从步骤1)的第一序列子集中提取种子序列,提取位置为每条短序列的首尾端各M个碱基,得到第二序列子集,其包含N对长度为M的种子序列,所述种子序列中不含待检测的点突变;

[0013] 3)对原始三代测序数据预处理,获得具有期望质量的原始数据集;

[0014] 4)使用步骤2)获得的第二序列子集的种子序列对从步骤3)得到的原始数据集中提取目的序列,获得N个包含目的序列的数据集;

[0015] 5)分别对步骤4)的N个包含目的序列的数据集进行点突变检测分析,得到N个结果;其中,每个结果包括待检测的位点的突变频率F,点突变的reads支持数AO,点突变位置的测序深度DP;

[0016] 6)对步骤5)的N个检测结果中的每个点突变的结果分配权重W;

[0017] 7)根据公式计算点突变结果及其频率;

[0018] $$F_{correct}=\frac{\text{（点突变支持数*权重）之和}}{\text{（点突变位置的测序深度*权重）之和}}$$

[0019] 若$F_{correct}$≥1%，则为阳性，反之为阴性。

[0020] 根据本发明所述的方法，其中，在步骤1)中，D表示在任意提取的序列中，点突变所处的位置之间的碱基距离。所述固定距离D可以为大于1的任意整数，不限于任何特定的理论，但是优选地距离D设置为$5\leqslant D\leqslant\frac{L}{2}$；无需任何理论的限制，本领域技术人员可以任选地设置D的数值，例如设置为5≤D≤20，8≤D≤15等，例如D可以为5到20之间的任意整数。

[0021] 本领域技术人员可以理解的是，如果第一次提取的短序列中，待检测的点突变在短序列上的位置为D_0，则第X次提取时，所述点突变在该提取短序列中的位置L_x满足$L_x=D_0+(X-1)D$。

[0022] 根据本发明所述的分析方法，其中，对于$L_x=D_0+(X-1)D$而言，D_0可以理解为第一次提取时，待检测的点突变位于提取短序列中的位置；例如D_0可以为第一次提取的短序列中的第一个碱基、第二个碱基、第三个碱基、第四个碱基，以此类推；在优选的实施方案中，D_0≤L/4和/或D_0≥D，例如D_0可以为D、D+1、D+2等。

[0023] 在一个具体的实施方案中，比如待检测的点突变的位置分别位于提取的短序列上的第11个碱基，第21个碱基，第31个碱基等；可以理解为D_0为11，D为10，X为1、2和3。

[0024] 在任选的实施方案中，$M<D_0\leqslant\frac{L}{4}$。

[0025] 根据本发明所述的分析方法，其中，在步骤1)中，提取次数N需要根据固定长度L和固定距离D决定。

[0026] 在优选的实施方案中，N为偶数时，所获得的N条短序列中，第$\frac{N}{2}$次和第$\frac{N}{2}$+1次提取的短序列中待检测的点突变与其在其他的短序列上的位置相比，可以位于该短序列的中间位置或最靠近中间的位置；N为奇数时，第$\frac{N+1}{2}$次提取得到的短序列中待检测的点突变与其在其他的短序列上的位置相比，位于该短序列的中间位置或最靠近中间的位置。

[0027] 根据本发明所述的分析方法，其中，在步骤1)中，每条序列的固定长度L可以是任选长度，并且该长度可以短至35bp，或长达250bp，优选地为76-151bp。

[0028] 根据本发明所述的分析方法，其中，在步骤2)中，M可以为任选的整数，但是基于现实考虑，M可以为2、3、4或5，优选地，M≥5。

[0029] 根据本发明所述的分析方法，其中，在步骤3)中，原始数据为经纳米孔测序获得的长读长数据。

[0030] 对原始三代测序数据进行数据预处理，包括利用例如Porechop软件以及NanoFilt软件去除实验建库过程中加入的接头及条形码(barcode)序列，过滤低质量以及过短的测序reads，得到期望的原始数据集(clean data)。

[0031] 在任选的实施方案中，所述低质量的阈值包括但不限于Q5，例如所述阈值可以为Q7或更高；其中，Q表示测序read的平均质量值，即测序read中的每一个碱基的准确率求和取平均后获得的值。本领域技术人员已知的是，该阈值可以根据实际情况进行调整，具体的

调整参数详见https://en.wikipedia.org/wiki/FASTQ_format,该处通过引用将其并入本文。

[0032] 在任选的实施方案中,过短的测序reads的序列长度阈值包括但不限于100bp;例如所述阈值可以为50bp、200bp、300bp等。本领域技术人员可以根据实际情况进行调整该阈值。

[0033] 根据本发明所述的分析方法,其中,在步骤4)中,考虑到三代测序数据的特征干扰,限制提取出相应的目的序列长度L'≤L+50。

[0034] 根据本发明所述的分析方法,其中,在步骤5)中,经过本申请的前述步骤处理之后的获得的N个包含目的序列的数据集,可以使用二代测序数据分析点突变的标准或成熟的主流分析流程,例如GATK Best Practice等。

[0035] 其中,N个包含目的序列的数据集进行点突变检测分析,得到N个结果;每个结果包括突变频率为F,点突变的reads支持数为AO,点突变位置的测序深度为DP。

[0036] 例如第一数据集的结果包括突变频率F_1,点突变的reads支持数AO_1,点突变位置的测序深度DP_1。

[0037] 第二数据集的结果包括突变频率F_2,点突变的reads支持数AO_2,点突变位置的测序深度DP_2。

[0038] ……

[0039] 例如第N数据集的结果包括突变频率F_N,点突变的reads支持数AO_N,点突变位置的测序深度DP_N。

[0040] 根据本发明所述的分析方法,其中,在步骤6)中,对N个检测结果中的每个点突变的结果分配权重(Weight),即W_1、W_2、W_3、……,W_{N-1},W_N,且$W_1+W_2+W_3+\cdots\cdots+W_{N-1}+W_N=1$,其中,在步骤1)中获得的N条短序列中,点突变在所述短序列的固定长度L上的位置越邻近中间,与所述短序列相关的检测结果分配的权重越大。

[0041] 在一个优选的实施方案中,N为偶数时,第$\frac{N}{2}$个和第$\frac{N}{2}$+1个数据集(可以理解为使用第$\frac{N}{2}$次和第$\frac{N}{2}$+1次提取的短序列获得的种子序列所得到的数据集)具有最大的权重$W_{N/2}=W_{N/2+1}$,然后$W_N=W_1$,$W_{N-1}=W_2$,$W_{N-2}=W_3$,以此类推。其中,NN为奇数时,第$\frac{N+1}{2}$个数据集(可以理解为使用第$\frac{N+1}{2}$次提取的短序列获得的种子序列所得到的数据集)具有最大的权重$W_{N+1/2}$,然后$W_N=W_1$,$W_{N-1}=W_2$,$W_{N-2}=W_3$,以此类推。

[0042] 根据本发明所述的分析方法,其中,在步骤7)中,所述公式为

[0043] $$F_{correct}=\frac{W_1*AO_1+W_2*AO_2+\cdots....W_{N-1}*AO_{N-1}+W_N*AO_N}{W_1*DP_1+W_2*DP_2+\cdots....W_{N-1}*DP_{N-1}+W_N*DP_N}$$。

[0044] 在所述公式中,发明人同时结合碱基在测序序列位置上的“中间较准,两边较差”的理论观点、数据分析层面上的分子生物标签(UMI/UID)思想以及“权重”统计的方法对检测结果进行整体评估、纠错及矫正,更加有效地控制了假阳性的结果。

[0045] 在一个具体的实施方案中，本发明的方法包括以下步骤：

[0046] 1)从参考基因组中提取包含待检测的点突变的第一序列子集；

[0047] 在所述参考基因组上进行固定长度L的短序列提取N次，第一次提取获得的短序列中，待检测的点突变所在位置为D_0，所述短序列之间满足待检测的点突变在提取后的短序列上的位置与其在前一次提取的短序列上的位置之间具有固定距离D，最终得到第一序列子集，其包含N个含有待检测的点突变的短序列；

[0048] 其中，L为76-151bp之间的任意整数，D为8到15之间的任意整数，N为4到18之间的任意整数，D_0为5到14之间的任意整数；

[0049] 2)从步骤1)得到的第一序列子集中的每条序列提取种子序列，提取位置分别为每条序列两端各M个碱基，最终得到N个种子序列对的第二序列子集，其中$5 \leq M < D_0$；

[0050] 3)对原始三代测序数据进行数据预处理，利用例如Porechop软件以及NanoFilt软件去除实验建库过程中加入的接头及barcode序列，过滤低质量以及过短的测序reads，得到具有期望质量的原始数据集；

[0051] 4)根据步骤2)得到的种子序列对，从步骤3)得到的原始数据集中提取出相应的目的序列，考虑到三代测序数据的特征干扰，限制提取出相应的目的序列长度$L' \leq L+50$，最终得到N个包含根据种子序列对提取出的目的序列数据集；

[0052] 5)对步骤4)得到的N个包含目的序列的数据集分别进行点突变检测分析，利用但不限于利用GATK Best Practice等分析流程，得到N个靶向位点检测的最终结果，记每个靶向位点检测的突变频率为F_N，该位点的突变reads支持数为AO_N，该位置的测序深度为DP_N；

[0053] 6)步骤5)的N个检测结果中的每个点突变的结果分配权重(Weight)，即W_1、W_2、W_3、……，W_{N-1}，W_N，N为偶数时，第$\frac{N}{2}$个和第$\frac{N}{2}$+1个数据集(可以理解为使用第$\frac{N}{2}$次和第$\frac{N}{2}$+1次提取的短序列获得的种子序列所得到的数据集)具有最大的权重$W_{N/2}=W_{N/2+1}$，然后$W_N=W_1$，$W_{N-1}=W_2$，$W_{N-2}=W_3$，以此类推。其中，N为奇数时，第$\frac{N+1}{2}$个数据集(可以理解为使用第$\frac{N+1}{2}$次提取的短序列获得的种子序列所得到的数据集)具有最大的权重$W_{N+1/2}$，然后$W_N=W_1$，$W_{N-1}=W_2$，$W_{N-2}=W_3$，以此类推。以此类推；

[0054] 7)加权及纠错矫正步骤5)中得到的靶向点突变结果及其频率，定义

[0055] $$F_{correct}=\frac{W_1*AO_1+W_2*AO_2+\cdots....W_{N-1}*AO_{N-1}+W_N*AO_N}{W_1*DP_1+W_2*DP_2+\cdots....W_{N-1}*DP_{N-1}+W_N*DP_N};$$

[0056] $F_{correct}$为最终该位点的检测突变频率；

[0057] 若$F_{correct} \geq 1\%$，则为阳性，反之为阴性。

[0058] 本发明还提供了一种基于三代测序数据检测点突变的装置，其中，所述装置包括：

[0059] 种子序列提取模块，用于获得包含种子序列对的第二序列子集；

[0060] 预处理模块，用于对三代测序数据预处理，获得具有期望质量的原始数据集；

[0061] 初级分析模块，用于使用第二序列子集的种子序列对从预处理后的原始数据集提取包含目的序列的数据集，然后进行点突变检测分析并获得数据；

[0062] 高级分析模块，用于对得到的结果进一步加权及矫正，并获得最终的分析结果；及

[0063] 报告模块，用于根据数据输出结果。

[0064] 根据本发明所述的装置,其中,所述种子序列提取模块用于从参考基因组中提取包含N个含有待检测的点突变的短序列的第一序列子集,然后从所述第一序列子集中提取包含种子序列对的第二序列子集;其中所述种子序列对根据本发明所述的数据处理方法获得。

[0065] 根据本发明所述的装置,其中,所述预处理模块用于过滤低质量以及过短的测序reads,可以包括例如Porechop软件以及NanoFilt软件。

[0066] 根据本发明所述的装置,其中,所述初级分析模块获得的数据具有与二代NGS测序数据类似的特征,可以使用NGS数据分析点突变的标准或成熟的主流分析流程,例如GATK Best Practice等

[0067] 根据本发明所述的装置,其中,所述高级分析模块包含用于对每个结果分配权重的程序或软件。其中,所述权重分配符合碱基在测序序列位置上的“中间较准,两边较差”的理论观点、数据分析层面上的分子生物标签(UMI/UID)思想以及“权重”统计的方法。

[0068] 本发明的发明人,基于三代测序的特有的数据特征,从数据分析层面上较好地解决了三代测序数据受限于测序质量以及依赖的比对算法或深度学习训练集的数据分布问题,以及适用场景并不够广泛,鲁棒性(robust)不足问题。使用本发明的方法,不仅从数据特征上有效地规避随机indel或较高测序错误导致的比对率不高导致的假阴性的问题,同时设计结合碱基在测序序列位置上的“中间较准,两边较差”的理论观点、数据分析层面上的分子生物标签(UMI/UID)思想以及“权重”统计的方法对检测结果进行整体评估、纠错及矫正,更加有效地控制了假阳性的结果。本发明的方法能够很好的兼容目前二代测序数据分析点突变的标准或成熟的主流分析流程,例如GATK Best Practice等,丰富了三代测序数据分析点突变的技术手段,很大程度上解决了三代测序检测点突变精准度不足的现状,在充分发挥了三代测序数据长读长的优势的同时,也进一步推动了三代测序在科研上的应用,特别适用于靶向相关热点panel的突变检测中。

附图说明

[0069] 为了更清楚地说明本发明实施例的技术方案,下面将对本发明实施例中所需要使用的附图作简单地介绍,显而易见地,下面所描述的附图仅仅是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0070] 图1示出为本发明的一个实施方案中基于三代测序数据检测点突变的分析方法的流程框架图;

[0071] 图2示出为本发明的一个实施方案中基于三代测序数据检测点突变的装置的结构框图。

具体实施方式

[0072] 下面将详细描述本发明的各个方面的特征和示例性实施例。在下面的详细描述中,提出了许多具体细节,以便提供对本发明的全面理解。但是,对于本领域技术人员来说很明显的是,本发明可以在不需要这些具体细节中的一些细节的情况下实施。下面对实施例的描述仅仅是为了通过示出本发明的示例来提供对本发明的更好的理解。

[0073] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将结合附图对实施例进行详细描述。

[0074] 在三代测序中，三代测序的数据在单碱基识别的精准度上还存在一些问题，体现在数据层面上就是测序质量不高或测序错误的现状以及随机indel分布等的数据特征。故在下游的数据分析中，如何稳定地检出点突变并且还能较好地控制假阳性及假阴性的检测结果就显得尤为重要。

[0075] 结合本发明的图1和图2，本发明提供了一种基于三代测序数据检测点突变的分析方法，所述方法包括以下步骤：

[0076] S1：从参考基因组中提取包含待检测的点突变的第一序列子集；

[0077] 在所述参考基因组上进行固定长度L的短序列提取N次，所述短序列之间满足待检测的点突变在提取后的短序列上的位置与其在前一次提取的短序列上的位置之间具有固定距离D，$N\leqslant\frac{L-D}{D}$；其中，N、D、L均为整数；最终得到第一序列子集，其包含N个含有待检测的点突变的短序列；

[0078] S2：从S1的第一序列子集中提取种子序列，提取位置为每条短序列的首尾端各M个碱基，得到第二序列子集，其包含N对长度为M的种子序列，所述种子序列中不含待检测的点突变；

[0079] S3：对原始三代测序数据预处理，获得具有期望质量的原始数据集；

[0080] S4：使用S2获得的第二序列子集的种子序列对从S3得到的原始数据集中提取目的序列，获得N个包含目的序列的数据集；

[0081] S5：分别对S4的N个包含目的序列的数据集进行点突变检测分析，得到N个结果；其中，每个结果包括待检测的位点的突变频率F，点突变的reads支持数AO，点突变位置的测序深度DP；

[0082] S6：对S5的N个检测结果中的每个点突变的结果分配权重W；

[0083] S7：根据公式计算点突变结果及其频率；

[0084] $$F_{correct}=\frac{\text{（点突变支持数*权重）之和}}{\text{（点突变位置的测序深度*权重）之和}}$$

[0085] 若$F_{correct}\geq 1\%$，则为阳性，反之为阴性。

[0086] 从上述方法可以获知，本发明的发明人通过制备种子序列，并结合测序的数据特点来进行多次采样提取，将三代测序的长读长的测序序列转变成短序列的序列，然后进行NGS数据类似的点突变分析，同时结合实验上单分子标签技术(UMI/UID)以及权重统计思想对多采样结果进行整合、评估、纠错及矫正来最终评判数据分析结果，能够有效地避免三代测序检测点突变精准度不足的问题。

[0087] 进一步地，如图2所示，本发明的一个实施方案中，提供了一种基于三代测序数据检测点突变的装置，其中，所述装置包括：种子序列提取模块101，用于获得包含种子序列对的第二序列子集；预处理模块102，用于对三代测序数据预处理，获得具有期望质量的原始数据集；初级分析模块103，用于使用第二序列子集的种子序列对从预处理后的原始数据集提取包含目的序列的数据集，然后进行点突变检测分析并获得数据；高级分析模块104，用于对得到的结果进一步加权及矫正，并获得最终的分析结果；及报告模块105，用于根据数

据输出结果。

[0088] 根据本发明所述的装置,其中,所述种子序列提取模块用于从参考基因组中提取包含N个含有待检测的点突变的短序列的第一序列子集,然后从所述第一序列子集中提取包含种子序列对的第二序列子集;其中所述种子序列对根据本发明所述的数据处理方法获得。

[0089] 根据本发明所述的装置,其中,所述预处理模块用于过滤低质量以及过短的测序reads,可以包括例如Porechop软件以及NanoFilt软件。

[0090] 根据本发明所述的装置,其中,所述初级分析模块获得的数据具有与二代NGS测序数据类似的特征,可以使用NGS数据分析点突变的标准或成熟的主流分析流程,例如GATK Best Practice等。

[0091] 根据本发明所述的装置,其中,所述高级分析模块包含用于对每个结果分配权重的程序或软件。其中,所述权重分配符合碱基在测序序列位置上的“中间较准,两边较差”的理论观点、数据分析层面上的分子生物标签(UMI/UID)思想以及“权重”统计的方法。

[0092] 实施例1 使用本发明的方法分析数据

[0093] 1.将含有*BRAF-V600E*、*EGFR-L858R*、*EGFR-T790M*、*KRAS-G13D*以及*AKT1-E17K*的标准品样本以及阴控样本 NA12878的标准品,通过实验文库制备且重复三次,利用QNome-9604的纳米孔测序仪进行测序,得到6个原始的长读长测序数据,其中HUM964、HUM965和HUM966为阳控数据,HUM967、HUM968和HUM969为阴控数据。

[0094] 2.分别对步骤1待检测的5个靶向位点根据其位置在基因组上进行固定长度101bp的短序列提取9次,其中靶向位点在提取后的短序列上的位置分别固定在第11个碱基,第21个碱基,第31个碱基,第41个碱基,第51个碱基,第61个碱基,第71个碱基,第81个碱基以及第91个碱基(即D=10bp),得到最终的9个包含5个靶向位点的短序列片段集合,且短序列片段长度为101bp。

[0095] 3.对每个短序列片段集合提取种子序列,提取位置分别为各个靶向位点短序列的首尾各10个碱基,最终得到9个包含靶向位点的短序列种子对序列的片段集合。

[0096] 4.对原始三代测序数据进行数据预处理,利用例如Porechop软件以及NanoFilt软件去除实验建库过程中加入的接头及barcode序列,过滤低质量Q7以及过短100bp以下的测序reads,得到clean data。

[0097] 5.从步骤4得到的clean data中,根据步骤3得到的短序列种子对序列提取出相应的目的序列,考虑到三代测序数据的特征干扰,限制提取出相应的目的序列长度$L'<151$,最终得到9个根据种子序列对提取出的目的序列数据集合。

[0098] 6.对步骤5中得到的9个数据集合分别进行点突变检测分析,本实施例中利用GATK Best Practice 进行点突变的检测,得到9个靶向位点检测的最终结果,记每个靶向位点检测的突变频率为F_N,该位点的突变reads支持数为AO_N,该位置的测序深度为DP_N。

[0099] 7.由于步骤5获得的包含长度L’的目的序列的数据集具有与二代测序获得的数据类似的特征,因此该步骤中假设步骤5中得到的目的短序列数据为二代测序平台数据并分配权重,根据碱基在二代测序的序列位置上的“中间较准,两边较差”的二代测序数据特点,对9个检测结果中的每个点突变的结果分配权重(Weight),即W_1、W_2、W_3、W_4、W_5、W_6、W_7、W_8、W_9,且$W_1+W_2+W_3+W_4+W_5+W_6+W_7+W_8+W_9=1$,$W_5=0.25$,$W_1=W_9=0.05$,$W_2=W_8=0.075$,$W_3=W_7=0.1$,$W_4=W_6=$

0.15。

[0100] 加权及纠错矫正步骤6中得到的靶向点突变结果及频率，定义

$$F_{correct}=\frac{W1*AO1+W2*AO2+W*AO3+W*AO4+W5*AO5+W6*AO6+W7*AO7+W8*AO8+W9*AO9}{W1*DP1+W*DP2+W*DP3+W*DP4+W5*DP5+W*DP6+W*DP7+W8*DP8+W9*DP9};$$

[0101] 且$F_{correct}$为最终该位点的检测突变频率；若$F_{correct}\geq 1\%$，则为阳性，反之为阴性。

[0102] 结果统计如表1所示，可见，本发明方法可以非常灵敏的将各个已知突变结果检出，与预期结论一致，且结果优于目前主流的分析三代测序点突变的算法及软件，有效地控制了假阴性及假阳性的结果，故本发明的方法可行。

[0103] 表1. 本发明方法检出各突变以及其频率的结果统计

[0104]

类型	突变	期望频率	Nano2NGS						iGDA					
			HUM964	HUM965	HUM966	HUM967	HUM968	HUM969	HUM964	HUM965	HUM966	HUM967	HUM968	HUM969
SNV	BRAF V600E	8.00%	10.63%	10.14%	10.33%	-	-	-	11.81%	10.90%	11.63%	-	-	-
SNV	EGFR L858R	5.00%	2.78%	2.59%	2.62%	-	-	-	2.78%	2.96%	2.47%	-	-	-
SNV	EGFR T790M	5.00%	5.86%	5.77%	4.69%	-	-	-	4.68%	5.13%	4.68%	-	-	-
SNV	KRAS G13D	5.00%	1.64%	1.57%	1.03%	-	-	-	9.47%	9.32%	10.88%	-	-	13.79%
SNV	AKT1 E17K	5.00%	3.09%	3.92%	4.30%	-	-	-	10.64%	9.41%	7.09%	6.02%	5.62%	6.52%
类型	突变	期望频率	Longshot						DeepVariant-Pepper					
			HUM964	HUM965	HUM966	HUM967	HUM968	HUM969	HUM964	HUM965	HUM966	HUM967	HUM968	HUM969
SNV	BRAF V600E	8.00%	-	-	-	-	-	-	-	-	-	-	-	-
SNV	EGFR L858R	5.00%	-	-	-	-	-	-	-	-	-	-	-	-
SNV	EGFR T790M	5.00%	-	-	-	-	-	-	-	-	-	-	-	-
SNV	KRAS G13D	5.00%	-	-	-	-	-	-	-	-	-	-	-	-
SNV	AKT1 E17K	5.00%	-	-	-	-	-	-	-	-	-	-	-	-

[0105] 其中Nano2NGS表示本发明所述的方法，通过表1的数据可以得知，使用本发明的方法，在三次重复中均检测到了*BRAF-V600E*、*EGFR-L858R*、*EGFR-T790M*、*KRAS-G13D*以及*AKT1-E17K*的突变，并且三次结果之间具有良好的重现性，与期望的频率之间没有显著差异。

[0106] Longshot方法例如发表于Nature Communications杂志(DOI: 10.1038/s41467-019-12493-y)，为加利福尼亚大学开发的结合隐马尔可夫链模型的得到的三代测序的点突变检测方法，由表1的数据可以，使用该方法分析，无法获得点突变的数据。

[0107] DeepVariant方法(公开于bioRxiv上基于google团队的DeepVariant开发优化的PEPPER-Margin-DeepVariant方法(doi: https://doi.org/10.1101/2021.03.04.433952))也无法直接用于三代测序的点突变检测方法。

[0108] iGDA方法虽然可以直接用于三代测序的点突变检测，但是在阴控样本中也检测出点突变，获得假阳性的检测结果。

[0109] 因此，本发明的方法不仅从数据特征上有效地规避随机indel或较高测序错误导致的比对率不高导致的假阴性的问题，同时设计结合碱基在测序序列位置上的“中间较准，两边较差”的理论观点、数据分析层面上的分子生物标签(UMI/UID)思想以及“权重”统计的

方法对检测结果进行整体评估、纠错及矫正，更加有效地控制了假阳性的结果。本发明的方法能够很好的兼容目前二代测序数据分析点突变的标准或成熟的主流分析流程，例如GATK Best Practice等，丰富了三代测序数据分析点突变的技术手段，很大程度上解决了三代测序检测点突变精准度不足的现状，在充分发挥了三代测序数据长度长的优势的同时，也进一步推动了三代测序在科研上的应用，特别适用于靶向相关热点panel的突变检测中。

[0110] 另外，本文中术语“和/或”，仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况。另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系。

[0111] 应理解，在本发明实施例中，“与A相应的B”表示B与A相关联，根据A可以确定B。但还应理解，根据A确定B并不意味着仅仅根据A确定B，还可以根据A和/或其它信息确定B。

[0112] 以上所述，仅为本发明的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，可轻易想到各种等效的修改或替换，这些修改或替换都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应以权利要求的保护范围为准。

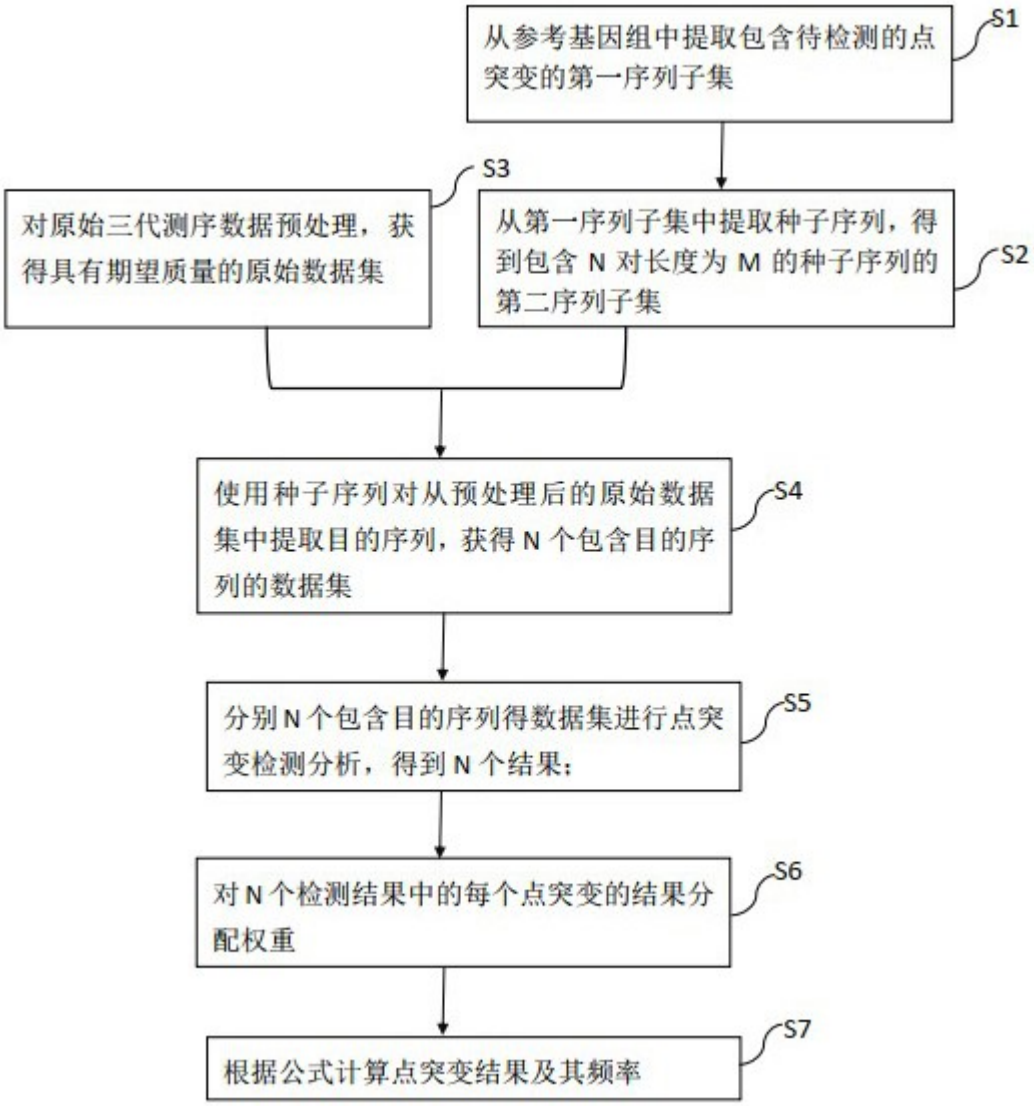

从参考基因组中提取包含待检测的点突变的第一序列子集
S1
对原始三代测序数据预处理，获得具有期望质量的原始数据集
S3
从第一序列子集中提取种子序列，得到包含 N 对长度为 M 的种子序列的第二序列子集
S2
使用种子序列对从预处理后的原始数据集中提取目的序列，获得 N 个包含目的序列的数据集
S4
分别 N 个包含目的序列得数据集进行点突变检测分析，得到 N 个结果；
S5
对 N 个检测结果中的每个点突变的结果分配权重
S6
根据公式计算点突变结果及其频率
S7

图1

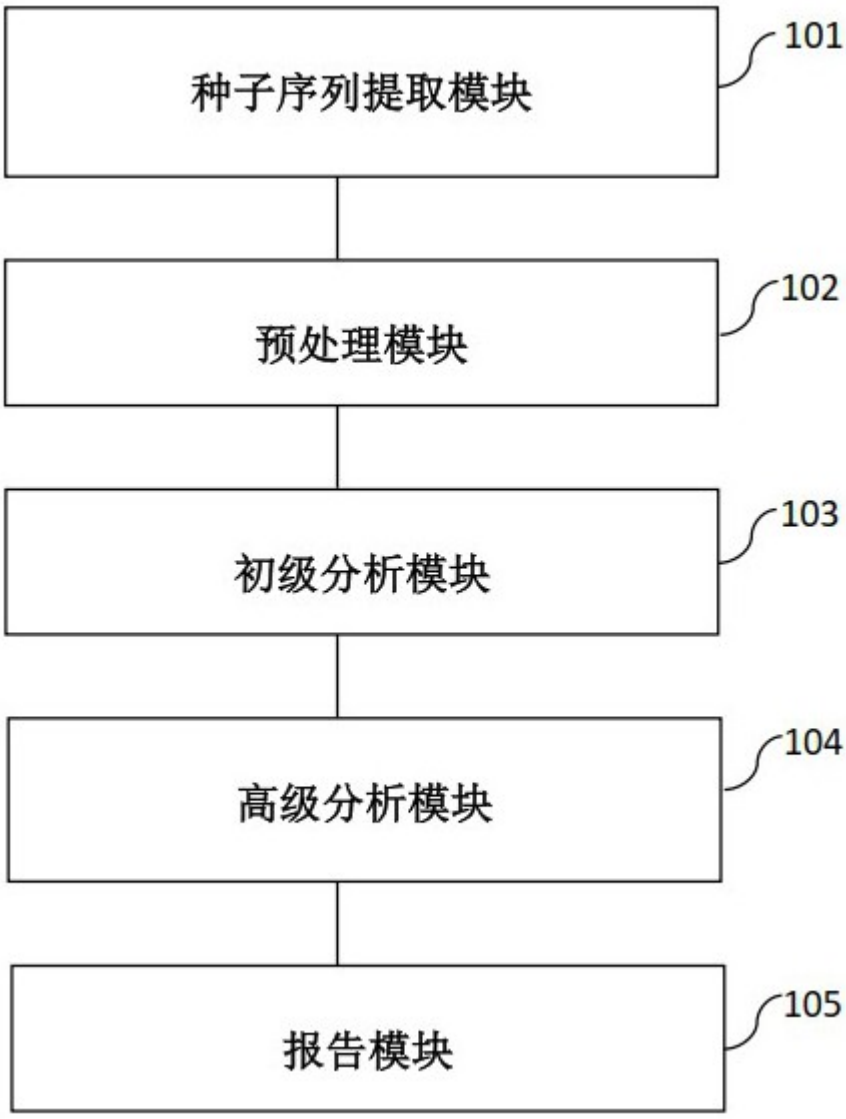


图2