

基于测序数据的碱基突变检测方法、装置及存储介质

技术领域

本申请实施例涉及基因变异检测技术，例如涉及一种基于测序数据的碱基突变检测方法、装置及存储介质。

背景技术

自2008年第二代测序技术走向成熟且被推向商业化以来，测序数据持续增长，基于测序数据进行多种下游应用的算法也处于活跃开发状态中。其中一类具有广泛使用前景的算法叫做基于测序数据的碱基突变检测方法。基因组变异包括单碱基突变，小插入缺失和结构性变异，不同的基因组变异各自有不同的检测算法，本申请专注于单碱基突变的变异检测。针对第二代测序技术中单碱基突变检测已经有较为稳定的开源算法，比如基因组分析工具包（Genome Analysis ToolKit, GATK）。

但是，相关技术中的单碱基突变检测算法仍存在如下主要缺陷：第一，无法检测出三碱基和四碱基突变位点；第二，不适用于测序深度偏低且样本数据很大的场景。例如，由于内存、速度等问题，单碱基突变检测算法甚至无法稳定地分析出大于5万的样本数据。然而，无创产前检测就是一种全基因组超低深度的测序类型，平均测序深度低至0.06x，基因组覆盖率约为6%。虽然测序深度很低，但是大量无创产前测序数据比如说十万，百万甚至千万例的测序数据，构成了变异检测一个极好的原始数据集。且，全球累积了大概720万人的无创产前测序数据，通过对这些累积起来的群体基因组测序数据获得的人群特异变异频谱对于疾病研究，药物研发都有重要的指引作用。

在这一超大规模样本但是测序深度偏低的场景中，相关技术中的单碱基突变检测算法已经无法正常运行，因此，开发一种新的单碱基突变检测算法以适应上述应用场景则显得非常有必要。

发明内容

本申请公布一种基于测序数据的碱基突变检测方法、装置及存储介质，通过所述方法可实现对测序深度低，数据量大的测序数据进行碱基突变检测。

本申请实施例公布了一种基于测序数据的碱基突变检测方法，包括：

确定多个待检测样本的测序数据在研究位点为特定碱基的初始频率；

基于所述初始频率计算每个待检测样本在研究位点为特定碱基的期望值；

利用每个期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率进行更新；

利用更新后的初始频率继续计算每个待检测样本在研究位点为特定碱基的期望值，并利用每个新的期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率继续进行更新，重复上述迭代过程，直到每个待检测样本在研究位点为特定碱基的期望值收敛；

根据每个收敛的期望值确定每个待检测样本在研究位点的碱基突变类型以及变异置信度；

其中，所述特定碱基包括腺嘌呤 A 碱基、胸腺嘧啶 T 碱基、胞嘧啶 C 碱基或者鸟嘌呤 G 碱基。

本申请实施例还公布了一种基于测序数据的碱基突变检测装置，包括：

初始频率确定模块，设置为确定多个待检测样本的测序数据在研究位点为特定碱基的初始频率；

期望值计算模块，设置为基于所述初始频率计算每个待检测样本在研究位点为特定碱基的期望值；

更新模块，设置为利用每个期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率进行更新；

迭代模块，设置为利用更新后的初始频率继续计算每个待检测样本在研究位点为特定碱基的期望值，并利用每个新的期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率继续进行更新，重复上述迭代过程，直到每个待检测样本在研究位点为特定碱基的期望值收敛；

变异类型确定模块，设置为根据每个收敛的期望值确定每个待检测样本在研究位点的碱基突变类型以及变异置信度；

其中，所述特定碱基包括腺嘌呤 A 碱基、胸腺嘧啶 T 碱基、胞嘧啶 C 碱基或者鸟嘌呤 G 碱基。

本申请实施例还公布一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行以实现如上所述的任意一种基于测序数据的碱基突变检测方法。

本申请实施例公布的一种基于测序数据的碱基突变检测方法，通过确定多个待检测样本的测序数据在研究位点为特定碱基的初始频率；基于所述初始频率计算每个待检测样本在研究位点为特定碱基的期望值；利用每个期望值对所述待检测样本的测序数据在研究位点为特定碱基的初始频率进行更新；利用更新后的初始频率继续计算每个待检测样本在研究位点为特定碱基的期望值，并利用每个新的期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率继续进行更新，重复上述迭代过程，直到每个待检测样本在研究位点为特定碱基的期望值收敛；根据每个收敛的期望值确定每个待检测样本在研

究位点的碱基突变类型以及变异置信度的技术手段，实现了对测序深度低，数据量大的测序数据的碱基突变检测。

附图说明

图 1 为本申请一实施例提供的一种基于测序数据的碱基突变检测方法的流程示意图；

图 2 为本申请一实施例提供的一种基于测序数据的碱基突变检测装置的结构示意图；

图 3 为本申请一实施例提供的上述碱基突变检测方法在不同样本数据的条件下对不同突变频率的检出率示意图；

图 4 为本申请一实施例提供的在样本数目为 14 万的情况下，上述碱基突变检测方法在不同突变频率下对不同突变类型的突变检出率的示意图；

图 5 为本申请一实施例提供的上述碱基突变检测方法在不同突变频率下对突变位点的突变频率估算与真实值之间的差异示意图；

图 6 为本申请另一实施例提供的上述碱基突变检测方法在不同突变频率下对突变位点的突变频率估算与真实值之间的差异示意图。

具体实施方式

以下是对本文描述的主题的概述。本概述并非是为了限制权利要求的保护范围。

除非另有定义，本文所使用的所有的技术和科学术语与本技术领域的技术人员通常理解的含义相同。本文中所使用的术语是为了描述具体的实施例，不用于限制本申请。本文所使用的术语“和/或”包括一个或多个相关的所列项

目的任意的和所有的组合。

本申请实施例公布的基于测序数据的碱基突变检测方法可应用于产前检测或者其他超大规模的人群测序数据中，通常测序深度都会偏低，当然本申请提供的基于测序数据的碱基突变检测方法也可用于测序深度较高的情况。虽然每个样本的测序深度很低，但是当样本很多时，每个位点都有很多条碱基序列（以下简称 reads）的覆盖。每个覆盖到当前位点的来自独立 reads 的碱基都对应有一个由测序仪产生的碱基质量值，该碱基质量值记为 $p(d_i | b_{i,j})$ ，具体表示覆盖到当前位点的来自样本 i 独立 reads d_i 的碱基 j 的质量值。所述碱基质量值反应了对应碱基测错的概率。例如第一个位点 $L1=\{A, T\}$ 中第一条 reads 上覆盖的碱基是一个 A，A 的碱基质量值为 30，则表示第一条 reads 在第一个位点上检测到碱基 A 的错误概率为 $10^{-(30/10)}=10^{-3}$ ，对应的正确概率为 $1-10^{-3}$ ；其中，A 表示腺嘌呤碱基，T 表示胸腺嘧啶碱基，C 表示胞嘧啶碱基，G 表示鸟嘌呤碱基。如果样本是产前样本，且当前位点有多于一条 reads 的覆盖，为了避免胎儿和母亲 DNA 混杂导致 reads 之间不能独立的问题，每个样本只抽取一条 reads 序列。

图 1 为本申请实施例提供的基于测序数据的碱基突变检测方法的流程示意图。如图 1 所示，所述方法包括如下步骤。

在步骤 110 中，确定多个待检测样本的测序数据在研究位点为特定碱基的初始频率。

其中，所述研究位点指待检测是否存在碱基突变的位点；所述特定碱基包括腺嘌呤 A 碱基、胸腺嘧啶 T 碱基、胞嘧啶 C 碱基或者鸟嘌呤 G 碱基。

在一实施例中，确定多个待检测样本的测序数据在研究位点为特定碱基的初始频率，包括：统计多个待检测样本的测序数据中携带的特定碱基的数量以及多个待检测样本的测序数据中四种碱基的总数量；将所述特定碱基的数量与

所述四种碱基的总数量的商作为多个待检测样本的测序数据在研究位点为特定碱基的初始频率。

利用公式表示上述确定所述初始频率的过程为：

$$p_j = \frac{\sum_{i=1}^N b_{ij}}{n} \quad (1)$$

其中， b_{ij} 表示待检测样本 i 的测序数据中携带的特定碱基 j 的数量， N 表示待检测样本的总数， $\sum_{i=1}^N b_{ij}$ 表示多个待检测样本的测序数据中携带的特定碱基 j 的总数， n 表示多个待检测样本的测序数据中四种碱基的总数量， p_j 表示多个待检测样本的测序数据在研究位点为特定碱基 j 的初始频率， $j=\{0,1,2,3\}$ ， p_0 表示多个待检测样本的测序数据在研究位点为碱基 A 的初始频率， p_1 表示多个待检测样本的测序数据在研究位点为碱基 C 的初始频率， p_2 表示多个待检测样本的测序数据在研究位点为碱基 G 的初始频率， p_3 表示多个待检测样本的测序数据在研究位点为碱基 T 的初始频率。

在步骤 120 中，基于所述初始频率计算每个待检测样本在研究位点为特定碱基的期望值。

在一实施例中，通过如下公式 (2) 计算每个待检测样本在研究位点为特定碱基的期望值：

$$p(b_{i,j} | p_j, d_i) = \frac{p(b_{i,j} | p_j) p(d_i | b_{i,j})}{\sum_{b_{i,j}=0}^3 p(b_{i,j} | p_j) p(d_i | b_{i,j})} \quad (2)$$

其中， $b_{i,j}$ 表示待检测样本 i 在研究位点为特定碱基 j ， p_j 表示多个待检测样本的测序数据在研究位点为特定碱基 j 的初始频率， d_i 表示待检测样本 i 在研究位点覆盖到的碱基序列中的碱基集合， $p(b_{i,j} | p_j, d_i)$ 表示待检测样本 i 在研

究位点为特定碱基 j 的期望值, $p(b_{i,j}|p_j)$ 表示在给定 P_j 的情况下, 待检测样本 i 在研究位点为特定碱基 j 的先验概率, $p(d_i|b_{i,j})$ 表示待检测样本 i 在研究位点覆盖到的碱基序列的碱基质量值。

在步骤 130 中, 利用每个期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率进行更新。

在一实施例中, 通过如下公式 (3) 利用每个期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率进行更新:

$$\hat{p}_j = \frac{\sum_{i=1}^N p(b_{i,j}|p_j, d_i)}{n} \quad (3)$$

其中, $\hat{p}_j$ 表示多个待检测样本的测序数据在研究位点为特定碱基更新后的初始频率, $p(b_{i,j}|p_j, d_i)$ 表示待检测样本 i 在研究位点为特定碱基 j 的期望值, N 表示待检测样本的数量, n 表示多个待检测样本的测序数据中四种碱基的总数量, 所述四种碱基指碱基 A、T、C 和 G。

在步骤 140 中, 利用更新后的初始频率继续计算每个待检测样本在研究位点为特定碱基的期望值, 并利用每个新的期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率继续进行更新, 重复上述迭代过程, 直到每个待检测样本在研究位点为特定碱基的期望值收敛。

在一实施例中, 利用上述公式 (3) 的结果 $\hat{p}_j$ 替换公式 (2) 中的 P_j , 通过公式 (2) 得到每个待检测样本在研究位点为特定碱基的新的期望值, 并利用新的期望值按照上述公式 (3) 对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率继续进行更新, 直到公式 (2) 得出的期望值收敛。期望值收

敛的条件包括：相邻次迭代得到期望值之间的差连续三次小于万分之一；例如，第 10 次迭代得到的期望值为 q_{10} ，第 11 次迭代得到的期望值为 q_{11} ，第 12 次迭代得到的期望值为 q_{12} ，第 13 次迭代得到的期望值为 q_{13} ，若 $(q_{11}-q_{10}) < 0.0001$ ，且 $(q_{12}-q_{11}) < 0.0001$ ，且 $(q_{13}-q_{12}) < 0.0001$ ，则可确定期望值达到了收敛条件，期望值 q_{13} 为收敛的期望值。

直到所述期望值收敛时，可得到最优的碱基频率，即所述期望值收敛时对应的上述公式 (3) 的值，即研究位点为特定碱基 j 的频率。

在步骤 150 中，根据每个收敛的期望值确定每个待检测样本在研究位点的碱基突变类型以及变异置信度。

在一实施例中，根据每个收敛的期望值确定每个待检测样本在研究位点的碱基突变类型，包括：根据每个收敛的期望值计算多个待检测样本在研究位点属于四种特定碱基突变类型中每种特定碱基突变类型的最大似然估计值；计算两种相邻特定碱基突变类型的最大似然估计值的比值；对所述比值按照预设规则进行处理，以得到所述比值对应的概率；在所述概率小于设定阈值的情况下，确定每个待检测样本在研究位点的碱基突变类型为当前分母对应的特定碱基突变类型；其中，所述四种特定碱基突变类型包括：单碱基突变、二碱基突变、三碱基突变和四碱基突变。

所述根据每个收敛的期望值计算多个待检测样本在研究位点属于四种特定碱基突变类型中每种特定碱基突变类型的最大似然估计值，包括：按照如下公式 (4) 计算多个待检测样本在研究位点属于四种特定碱基突变类型中每种特定碱基突变类型的最大似然估计值：

$$p(D|p_j) = \prod_{i=1}^N \sum_{j=0}^3 \left[p(b_{i,j}|p_j) p(d_i|b_{i,j}) \right] \quad (4)$$

其中，D 表示所有待检测样本在研究位点覆盖到的碱基序列中的碱基集合构成的观察数据， p_j 表示根据每个收敛的期望值得到的多个待检测样本的测序数据在研究位点为特定碱基 j 的频率， $p(D|p_j)$ 表示待检测样本在研究位点为与 j 对应的碱基突变类型的最大似然估计值；在 j=0 的情况下，对应的碱基突变类型为单碱基突变，在 j=1 的情况下，对应的碱基突变类型为二碱基突变，j=2 时，对应的碱基突变类型为三碱基突变，在 j=3 的情况下，对应的碱基突变类型为四碱基突变。

在一实施例中，所述计算两种相邻特定碱基突变类型的最大似然估计值的比值，包括：假设研究位点是四碱基突变的最大似然估计值为 f_4 ，研究位点是三碱基突变的最大似然估计值为 f_3 ；将 $\frac{f_3}{f_4}$ 确定为两种相邻特定碱基突变类型的最大似然估计值的比值。

假设研究位点是二碱基突变的最大似然估计值为 f_2 ，三碱基突变中四种突变组合的最大似然估计值中的最小值为 $f_3 \min$ ；三碱基突变中四种突变组合分别为：{A,T,C}、{A,T,G}、{T,C,G}和{A,C,G}，每种突变组合对应一个上述公式 (4) 所述的最大似然估计值。将 $\frac{f_2}{f_3 \min}$ 确定为两种相邻特定碱基突变类型的最大似然估计值的比值。

假设研究位点是单碱基突变的最大似然估计值为 f_1 ，二碱基突变中 16 种突变组合的最大似然估计值中的最小值为 $f_2 \min$ ；将 $\frac{f_1}{f_2 \min}$ 确定为两种相邻特定碱基突变类型的最大似然估计值的比值。

对应的，对所述比值按照预设规则进行处理，以得到所述比值对应的概率，包括：对所述比值进行取自然对数操作，得到第一结果；将得到的第一结果乘

以-2，得到第二结果；通过查找卡方值分布表得到与所述第二结果对应的概率。

由于三碱基突变中，假设任意一个碱基的突变频率为 0，二碱基突变中，假设任意两个碱基的突变频率为 0，单碱基突变中假设任意三个碱基的突变频率为 0，因此，上述比值 ($\frac{f_3}{f_4}$ 、 $\frac{f_2}{f_3 \min}$ 以及 $\frac{f_1}{f_2 \min}$) 的分子都比分母少了一个频率参数 p_j ，因此，通过对上述比值进行取自然对数操作，并乘以-2 之后得到的统计量服从一个自由度的卡方分布，故可以通过卡方值分布表求出每个比值对应的概率。

对上述比值进行取自然对数操作，并乘以-2 之后得到的统计量为：

$$LRT_{4vs3} = -2 \ln\left(\frac{f_3}{f_4}\right)$$

$$LRT_{3vs2} = -2 \ln\left(\frac{f_2}{f_3 \min}\right)$$

$$LRT_{2vs1} = -2 \ln\left(\frac{f_1}{f_2 \min}\right)$$

当所述概率小于设定阈值时，例如，若 LRT_{4vs3} 对应的概率小于 10^{-6} ，则确定假设不成立，即确定研究位点不属于三碱基突变，而是四碱基突变；若 LRT_{3vs2} 对应的概率小于 10^{-6} ，则确定研究位点不属于二碱基突变，而是三碱基突变；若 LRT_{2vs1} 对应的概率小于 10^{-6} ，则确定研究位点不属于单碱基突变，而是二碱基突变。

在一实施例中，根据每个收敛的期望值确定每个待检测样本在研究位点的变异置信度，包括：将 $\frac{f_1}{f_2 \min}$ 对应的概率进行常规的 Phred-scale 转化，得到 Phred 质量值；将所述 Phred 质量值确定为每个待检测样本在研究位点的变异置信度。

通过变异置信度可进一步确定研究位点是否真实存在碱基突变。

本申请实施例公开的一种基于测序数据的碱基突变检测方法，与相关技术中的变异检测方法截然不同，且在时间复杂度上具备明显的优势，通过利用样本数据多但是单个样本测序深度低的数据特点，直接使用了基于等位基因突变型观察到数据的似然函数的方法（而不是使用基于基因型观察到数据的似然函数），使得整个检测方法速度变得很快，能完成大于十万甚至百万样本（例如百万例产前检测数据）的分析；另外，本申请实施例没有预先假设研究位点是二碱基突变，并且首次运用了多个似然检验的办法对不同碱基组合的最大似然估计值进行了检验，因此，本申请实施例的碱基突变检测方法首创除了能进行二碱基突变的检测，还能进行单碱基以及多碱基突变的检测。当运用本申请实施例的碱基突变检测方法对十万、百万例产前检测数据进行分析时，可获得高精度的人群突变位点和频率信息，这种信息同时具有科研和产业价值。

图 2 为本申请实施例提供的基于测序数据的碱基突变检测装置的结构示意图。如图 2 所示，所述装置包括：初始频率确定模块 210、期望值计算模块 220、更新模块 230、迭代模块 240 和变异类型确定模块 250；

本实施例中，初始频率确定模块 210，设置为确定多个待检测样本的测序数据在研究位点为特定碱基的初始频率；期望值计算模块 220，设置为基于所述初始频率计算每个待检测样本在研究位点为特定碱基的期望值；更新模块 230，设置为利用每个期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率进行更新；迭代模块 240，设置为利用更新后的初始频率继续计算每个待检测样本在研究位点为特定碱基的期望值，并利用每个新的期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率继续进行更新，重复上述迭代过程，直到每个待检测样本在研究位点为特定碱基的期望值收敛；变异类型确定模块 250，设置为根据每个收敛的期望值确定每个待检测样本在研

究位点的碱基突变类型以及变异置信度；其中，所述特定碱基包括腺嘌呤 A 碱基、胸腺嘧啶 T 碱基、胞嘧啶 C 碱基或者鸟嘌呤 G 碱基。

本申请实施例公布的一种基于测序数据的碱基突变检测装置，通过确定多个待检测样本的测序数据在研究位点为特定碱基的初始频率；基于所述初始频率计算每个待检测样本在研究位点为特定碱基的期望值；利用每个期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率进行更新；利用更新后的初始频率继续计算每个待检测样本在研究位点为特定碱基的期望值，并利用每个新的期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率继续进行更新，重复上述迭代过程，直到每个待检测样本在研究位点为特定碱基期望值收敛；根据每个收敛的期望值确定每个待检测样本在研究位点的碱基突变类型以及变异置信度的技术手段，实现了对测序深度低，数据量大的测序数据的碱基突变检测。

需要说明的是，装置实施例中，每个组成部分的具体工作原理，请参见方法实施例对应部分，此次不再赘述。

在一实施例中，为了验证上述碱基突变检测方法的效果，现提供一份利用上述碱基突变检测方法进行分析的实例，实例中利用上述碱基突变检测方法对大规模模拟数据进行分析，目的是检验上述碱基突变检测方法的突变检出率，假阳率，以及突变位点频率估算和真实值的差异。

实例

模拟数据概述

模拟数据一共模拟了100个单碱基位点(monopolymic loci)，50000个二碱基位点(di-allelic loci)，50000个三碱基位点(tri-allelic loci)以及50000个四碱基位点(tetra-allelic loci)。对于上述三组50000位点，都按照万分之一的间隔，设置最小

等位基因突变的频率，举例如表一所示：

表一：模拟数据的突变频率分布

等位基因位点	突变频率 (A)	突变频率 (C)	突变频率 (G)	突变频率 (T)	等位基因种类
1	1	0	0	0	1
2	1	0	0	0	1
3	1	0	0	0	1
...					
98	1	0	0	0	1
99	1	0	0	0	1
100	1	0	0	0	1
...					
101	1.00E-05	0.99999	0	0	2
102	2.00E-05	0.99998	0	0	2
103	3.00E-05	0.99997	0	0	2
...					
50098	0.49998	0.50002	0	0	2
50099	0.49999	0.50001	0	0	2
50100	0.5	0.5	0	0	2
...					
50101	1.00E-05	0.499995	0.499995	0	3
50102	2.00E-05	0.49999	0.49999	0	3
50103	3.00E-05	0.499985	0.499985	0	3
...					
100098	0.49998	0.25001	0.25001	0	3
100099	0.49999	0.250005	0.250005	0	3
100100	0.5	0.25	0.25	0	3
...					
100101	1.00E-05	0.33333	0.33333	0.33333	4
100102	2.00E-05	0.333326667	0.333326667	0.333326667	4
100103	3.00E-05	0.333323333	0.333323333	0.333323333	4
...					
150098	0.49998	0.166673333	0.166673333	0.166673333	4
150099	0.49999	0.16667	0.16667	0.16667	4
150100	0.5	0.166666667	0.166666667	0.166666667	4

在设定了表一所示的碱基突变频率的分布后，给定样本测序深度为0.06x，测序错误率在0.01左右。模拟数据的主要目的在于观察本申请实施例提供的上述基于测序数据的碱基突变检测方法在不同条件下(不同样本数目，不同突变频率，不同突变类型)的突变检出率，假阳率，以及突变位点频率估算和真实值的差异。参见图3所示的本申请实施例提供的上述基于测序数据的碱基突变检测方法在不同样本数据的条件下对不同突变频率的检出率示意图，横轴表示突变频率，纵轴表示检出率，图3中示出了在不同突变频率下样本数分别为44000(44k)人、140000(140k)人以及1百万(1M)人的检出率。通常会关注在一定样本数量下上述基于测序数据的碱基突变检测方法能够检出(检出率>0时)的最低突变频率以及检出率达到100%时的最低突变频率。从图3中可以看到，随着样本数目从44k升至140k至1M，检出率>0时的最低检出突变频率分别为0.002，0.001和0.00002，而检出率达到100%时的最低突变频率(完全检测突变频率)分别为：0.015，0.005，0.002，可见上述基于测序数据的碱基突变检测方法的性能较好。特别的，100个单碱基突变都未被错误检出，即假阳率为0。图3的结果表明了本申请实施例提供的基于测序数据的碱基突变检测方法能够灵敏且准确地从大量深度低至0.06x的测序数据中探测到单碱基突变。

同时参见图4所示的在样本数目为14万时，上述碱基突变检测方法在不同突变频率下对不同突变类型的突变检出率的示意图，其中，横轴表示突变频率，纵轴表示检出率。从图4中可以看出，对于二碱基突变，能够检出(检出率>0时)的最低突变频率以及检出率达到100%时的最低突变频率分别是0.002和0.005。较难的三碱基突变以及四碱基突变虽然检出率比二碱基突变稍低，但是效果相差不多，对于三碱基突变，能够检出(检出率>0时)的最低突变频率以及检出率

达到100%时的最低突变频率分别是0.002和0.008,对于四碱基突变,能够检出(检出率>0时)的最低突变频率以及检出率达到100%时的最低突变频率分别是0.001和0.008。同样地,无一非突变位点被检出,说明了本方法的准确度较高。图4的结果表明本申请实施例提供的基于测序数据的碱基突变检测方法能够灵敏且准确地探测到三碱基或者四碱基突变。

同时参见图5和图6所示的上述碱基突变检测方法在不同突变频率下对突变位点的突变频率估算与真实值之间的差异示意图,其中,横轴表示突变频率,图5的纵轴表示均方根误差,图6的纵轴表示均方根误差变异系数。图5反映的是每一个频段频率和真实突变频率的平均差异,直观反映了检出频率与真实频率地大致差异,但因为不同频率段度量不同,所以不能横向对不同频率差异进行比较。图6统一了不同频率段的度量,可以比较不同频率段下方法的检出频率和真实频率的差异大小。从图5可以看到,随着模拟突变频率的增加,检出频率与真实频率存在一定差异但较小。从图6可以看到,低频频率(<0.003)与真实频率段的度量差异在1-1.5倍左右,而频率大于0.003的突变小于度量差异。通过分析可以看到对于探测到的突变位点,上述方法能够准确地探测到真实的突变频率。

本申请实施例还公开一种包含计算机可执行指令的存储介质,所述计算机可执行指令在由计算机处理器执行时用于执行一种基于测序数据的碱基突变检测方法,该方法包括:确定多个待检测样本的测序数据在研究位点为特定碱基的初始频率;基于所述初始频率计算每个待检测样本在研究位点为特定碱基的期望值;利用所述期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率进行更新;利用更新后的初始频率继续计算每个待检测样本在研究位点为特定碱基的期望值,并利用新的期望值对所述多个待检测样本的测

序数据在研究位点为特定碱基的初始频率继续进行更新，重复上述迭代过程，直到所述期望值收敛；根据收敛的期望值确定每个待检测样本在研究位点的碱基突变类型以及变异置信度；其中，所述特定碱基包括腺嘌呤 A 碱基、胸腺嘧啶 T 碱基、胞嘧啶 C 碱基或者鸟嘌呤 G 碱基。

最后，还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本申请。

权利要求书

1、一种基于测序数据的碱基突变检测方法，包括：

确定多个待检测样本的测序数据在研究位点为特定碱基的初始频率；

基于所述初始频率计算每个待检测样本在研究位点为特定碱基的期望值；

利用每个期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率进行更新；

利用更新后的初始频率继续计算每个待检测样本在研究位点为特定碱基的期望值，并利用每个新的期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率继续进行更新，重复上述迭代过程，直到每个待检测样本在研究位点为特定碱基的期望值收敛；

根据每个收敛的期望值确定每个待检测样本在研究位点的碱基突变类型以及变异置信度；

其中，所述特定碱基包括腺嘌呤 A 碱基、胸腺嘧啶 T 碱基、胞嘧啶 C 碱基或者鸟嘌呤 G 碱基。

2、根据权利要求 1 所述的方法，其中，所述确定多个待检测样本的测序数据在研究位点为特定碱基的初始频率，包括：

统计多个待检测样本的测序数据中特定碱基的数量以及所述多个待检测样本的测序数据中四种碱基的总数量；其中，所述四种碱基包括：腺嘌呤 A 碱基、胸腺嘧啶 T 碱基、胞嘧啶 C 碱基和鸟嘌呤 G 碱基；

将所述特定碱基的数量与所述四种碱基的总数量的商作为所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率。

3、根据权利要求 1 所述的方法，其中，所述基于所述初始频率计算每个待检测样本在研究位点为特定碱基的期望值，包括：

通过如下公式计算每个待检测样本在研究位点为特定碱基的期望值：

$$p(b_{i,j}|p_j, d_i) = \frac{p(b_{i,j}|p_j)p(d_i|b_{i,j})}{\sum_{b_{i,j}=0}^3 p(b_{i,j}|p_j)p(d_i|b_{i,j})}$$

其中, $b_{i,j}$ 表示待检测样本 i 在研究位点为特定碱基 j , p_j 表示多个待检测样本的测序数据在研究位点为特定碱基 j 的初始频率, d_i 表示待检测样本 i 在研究位点覆盖到的碱基序列中的碱基集合, $p(b_{i,j}|p_j, d_i)$ 表示待检测样本 i 在研究位点为特定碱基 j 的期望值, $p(b_{i,j}|p_j)$ 表示在给定 p_j 的情况下, 待检测样本 i 在研究位点为特定碱基 j 的先验概率, $p(d_i|b_{i,j})$ 表示待检测样本 i 在研究位点覆盖到的碱基序列的碱基质量值。

4、根据权利要求 3 所述的方法, 其中, 利用每个期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率进行更新, 包括:

通过如下公式利用每个期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率进行更新:

$$\hat{p}_j = \frac{\sum_{i=1}^N p(b_{i,j}|p_j, d_i)}{n}$$

其中, $\hat{p}_j$ 表示多个待检测样本的测序数据在研究位点为特定碱基更新后的初始频率, $p(b_{i,j}|p_j, d_i)$ 表示待检测样本 i 在研究位点为特定碱基 j 的期望值, N 表示待检测样本的数量, n 表示多个待检测样本的测序数据中四种碱基的总数量。

5、根据权利要求 1 或 4 所述的方法, 其中, 根据每个收敛的期望值确定每个待检测样本在研究位点的碱基突变类型, 包括:

根据每个收敛的期望值计算多个待检测样本在研究位点属于四种特定碱基

突变类型中每种特定碱基突变类型的最大似然估计值；

计算两种相邻特定碱基突变类型的最大似然估计值的比值；

对所述比值按照预设规则进行处理，以得到所述比值对应的概率；

在所述概率小于设定阈值的情况下，确定每个待检测样本在研究位点的碱基突变类型为当前分母对应的特定碱基突变类型；

其中，所述四种特定碱基突变类型包括：单碱基突变、二碱基突变、三碱基突变和四碱基突变。

6、根据权利要求5所述的方法，其中，根据每个收敛的期望值计算多个待检测样本在研究位点属于四种特定碱基突变类型中每种特定碱基突变类型的最大似然估计值，包括：

按照如下公式计算多个待检测样本在研究位点属于四种特定碱基突变类型中每种特定碱基突变类型的最大似然估计值：

$$p(D|p_j) = \prod_{i=1}^N \sum_{j=0}^3 [p(b_{i,j}|p_j)p(d_i|b_{i,j})]$$

其中，D表示所有待检测样本在研究位点覆盖到的碱基序列中的碱基集合构成的观察数据， p_j 表示根据每个收敛的期望值得到的多个待检测样本的测序数据在研究位点为特定碱基j的频率， $p(D|p_j)$ 表示多个待检测样本在研究位点为与j对应的碱基突变类型的最大似然估计值， $p(b_{i,j}|p_j)$ 表示在给定 p_j 的情况下，待检测样本i在研究位点为特定碱基j的先验概率， $p(d_i|b_{i,j})$ 表示待检测样本i在研究位点覆盖到的碱基序列的碱基质量值；在j=0的情况下，对应的特定碱基突变类型为单碱基突变，在j=1的情况下，对应的特定碱基突变类型为二碱基突变，在j=2的情况下，对应的特定碱基突变类型为三碱基突变，在j=3的

情况下，对应的特定碱基突变类型为四碱基突变。

7、根据权利要求6所述的方法，其中，所述计算两种相邻特定碱基突变类型的最大似然估计值的比值，包括：

在研究位点是四碱基突变的最大似然估计值为 f_4 ，研究位点是三碱基突变的最大似然估计值为 f_3 的情况下；

将 $\frac{f_3}{f_4}$ 确定为两种相邻特定碱基突变类型的最大似然估计值的比值。

8、根据权利要求6所述的方法，其中，所述计算两种相邻特定碱基突变类型的最大似然估计值的比值，包括：

在研究位点是二碱基突变的最大似然估计值为 f_2 ，三碱基突变中四种突变组合的最大似然估计值中的最小值为 $f_3 \min$ 的情况下；

将 $\frac{f_2}{f_3 \min}$ 确定为两种相邻特定碱基突变类型的最大似然估计值的比值。

9、根据权利要求6所述的方法，其中，所述计算两种相邻特定碱基突变类型的最大似然估计值的比值，包括：

在研究位点是单碱基突变的最大似然估计值为 f_1 ，二碱基突变中16种突变组合的最大似然估计值中的最小值为 $f_2 \min$ 的情况下；

将 $\frac{f_1}{f_2 \min}$ 确定为两种相邻特定碱基突变类型的最大似然估计值的比值。

10、根据权利要求7-9任一项所述的方法，其中，对所述比值按照预设规则进行处理，以得到所述比值对应的概率，包括：

对所述比值进行取自然对数操作，得到第一结果；

将得到的第一结果乘以-2，得到第二结果；

通过查找卡方值分布表得到与所述第二结果对应的概率。

11、根据权利要求 9 所述的方法，其中，根据每个收敛的期望值确定每个待检测样本在研究位点的变异置信度，包括：

将 $\frac{f_1}{f_2 \min}$ 对应的概率进行常规的 Phred-scale 转化，得到 Phred 质量值；

将所述 Phred 质量值确定为每个待检测样本在研究位点的变异置信度。

12、一种基于测序数据的碱基突变检测装置，包括：

初始频率确定模块，设置为确定多个待检测样本的测序数据在研究位点为特定碱基的初始频率；

期望值计算模块，设置为基于所述初始频率计算每个待检测样本在研究位点为特定碱基的期望值；

更新模块，设置为利用每个期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率进行更新；

迭代模块，设置为利用更新后的初始频率继续计算每个待检测样本在研究位点为特定碱基的期望值，并利用每个新的期望值对所述多个待检测样本的测序数据在研究位点为特定碱基的初始频率继续进行更新，重复上述迭代过程，直到每个待检测样本在研究位点为特定碱基的期望值收敛；

变异类型确定模块，设置为根据每个收敛的期望值确定每个待检测样本在研究位点的碱基突变类型以及变异置信度；

其中，所述特定碱基包括腺嘌呤 A 碱基、胸腺嘧啶 T 碱基、胞嘧啶 C 碱基或者鸟嘌呤 G 碱基。

13、一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行以实现如权利要求 1-11 中任一所述的一种基于测序数据的碱基突变检测方法。