(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 106570350 B
(45)授权公告日 2019.04.05

(21)申请号 201510955436.0

(22)申请日 2015.12.17

(65)同一申请的已公布的文献号
申请公布号 CN 106570350 A

(43)申请公布日 2017.04.19

(73)专利权人 复旦大学
地址 200433 上海市杨浦区邯郸路220号

(72)发明人 金力 李士林 王一

(74)专利代理机构 上海元一成知识产权代理事务所(普通合伙) 31268
代理人 吴桂琴

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

(56)对比文件
CN 101210266 A,2008.07.02,
CN 103914631 A,2014.07.09,
CN 101539967 A,2009.09.23,
熊兴东等.DNA修复基因ERCC1 C19007T多态与宫颈癌.《实用妇产科杂志》.2010,第26卷(第4期),
熊兴东等.DNA修复基因ERCC1 C19007T多态与宫颈癌.《实用妇产科杂志》.2010,第26卷(第4期),

审查员 朱哲

权利要求书2页 说明书3页

(54)发明名称

单核苷酸多态位点分型算法

(57)摘要

本发明属于生物信息学领域,具体涉及用于从原始测序数据中对单核苷酸多态位点(Single nucleotide polymorphisms,SNP)进行精确分型的算法。该算法基于统计学和群体遗传学原理,可对样本的指定SNP位点进行分型,且对该分型结果进行相应的质量评估。本发明的标准质量分数能精确的评价分型的准确率,且非常容易在实际工作中使用。可进一步作为实际法医学工作中标准化的质量统计量。

CN 106570350 B

1.一种单核苷酸多态位点分型算法,其特征在于,所述算法是对样本的指定SNP位点进行精确分型的算法,其中,通过构建二项分布统计学模型,对SNP位点的等位基因在人群中的分布进行模拟,从而精确的推测出个体的基因分型;

所述的算法包括步骤:

(1)建立模型:给定一个SNP位点,分别提取每个样本的两个等位基因的有效乘数EBD:

$$EBD = \sum_{i=1}^{reads} (1 - 10^{-0.1 \times base\_quality_i})(1 - 10^{-0.1 \times mapping\_quality_i})$$

对于一个群体,第i个个体的参考等位基因(reference allele)与交互等位基因(alternative allele)的EBD分别为ri和ai;对三种可能的基因型RR、RA、AA,假设它们在测序中分别有一个固定的突变等位基因出现率,分别为p(RR)、p(RA)和p(AA);理想情况下p(RR)接近0,p(RA)接近0.5,p(AA)接近1;假设等位基因频率服从哈迪-温伯格平衡,同时有固定的交互等位基因频率(alternative allele frequency)fre,因此:

$f(RR) = (1-fre)^2$

$f(RA) = 2fre(1-fre)$

$f(AA) = fre^2$

实际样本由于其基因型未知,认为它是由三种等位基因叠加而成,因此SNP模型具有如下概率模型:

$$likelihood = const \times \prod_i \begin{matrix} (1-fre)^2(1-p(RR))^{r_i} p(RR)^{a_i} \\ +2\,fre(1-fre)(1-p(RA))^{r_i} p(RA)^{a_i} \\ +fre^2(1-p(AA))^{r_i} p(AA)^{a_i} \end{matrix}$$

(2)最大似然估计:引入隐变量:$w(RR)_i$、$w(RA)_i$、$w(AA)_i$表述所述个体的三种基因型概率;使用Expectation-Maximization(EM)算法进行最大似然估计,其E步骤和M步骤分别是:

E步骤:

$$w(geno)_i = \frac{f(geno)(1-p(geno))^{r_i} p(geno)^{a_i}}{\sum_{geno} f(geno)(1-p(geno))^{r_i} p(geno)^{a_i}}$$

M步骤:

$$fre = \frac{2\sum_i w(AA)_i + \sum_i w(RA)_i}{2N}$$

$$p(geno) = \frac{\sum_i w(geno)_i a_i}{\sum_i w(geno)_i a_i + \sum_i w(geno)_i r_i}$$

(3)样本基因型确定:对于第i个样本,取$w(RR)_i$、$w(RA)_i$、$w(AA)_i$中最大的基因型作为该样本的基因型。

2.如权利要求1所述的算法,其特征在于,所述算法中还包括对分型结果的质量评估,其包括:构建一个统计量描述,令所选的基因型对应的$1-w(geno)_i$为该基因型的标准质量分数,表征该位点基因分型的准确率,标准质量分数越高,分型准确率越低。

3.如权利要求1-2中任一项所述的算法在用于分析第二代测序技术产生的测序数据中的应用。

单核苷酸多态位点分型算法

技术领域

[0001] 本发明属于生物信息学领域，涉及单核苷酸多态位点分型算法，尤其涉及一个用于从原始二代测序数据中对SNP进行精确分型的算法。

背景技术

[0002] 对生物样本的DNA进行精确解读是进行分子生物学、遗传学及法医物证鉴定等领域的前提。

[0003] 单核苷酸多态遗传标记位点(Single nucleotide polymorphisms,SNP)是由单个碱基的突变造成。人类基因组上已经发现大约千万级的SNP位点，并且证明其与众多表型、疾病等相关联。

[0004] 第二代测序技术是目前最流行的DNA测序分型方法。通过使用第二代测序技术，大量的原始人类基因组测序数据在近年来被产生。因此，实践中，需要创建可对这些SNP位点进行精确分型的算法。

[0005] 基于现有技术的现状，本申请的发明人拟提供一种单核苷酸多态位点分型算法，尤其是能够进行从第二代测序数据中精确分型特定SNP位点的算法。

发明内容

[0006] 本发明的目的在于提供一种能够进行从第二代测序数据中精确分型特定SNP位点的算法。

[0007] 本发明提供了进行精确分型特定SNP位点的算法。

[0008] 本发明通过构建二项分布统计学模型，对SNP位点的等位基因在人群中的分布进行模拟，精确的推测出个体的基因分型。本算法提供了对分型结果的质量评估，从而提供了二代测序数据背景下的质量评价体系。

[0009] 本发明中，软件基于C/C++语言，适用于linux或windows系统。

[0010] 更具体的，本发明的单核苷酸多态位点分型算法，其包括步骤：

[0011] 给定一个SNP位点，本发明的实施例中，分别提取每个样本的两个等位基因的有效乘数(the effective base depth,简写EBD)：

[0012] $$EBD = \sum_{i=1}^{reads} (1 - 10^{-0.1 \times Base\_quality_i})(1 - 10^{-0.1 \times mapping\_quality_i})$$

[0013] 对于一个群体，第i个个体的参考等位基因(reference allele)与交互等位基因(alternative allele)的EBD分别为ri和ai。对三种可能的基因型RR、RA、AA，本发明中，假设它们在测序中分别有一个固定的突变等位基因出现率，分别为p(RR)、p(RA)和p(AA)；理想情况下p(RR)接近0，p(RA)接近0.5，p(AA)接近1；假设等位基因频率服从哈迪-温伯格平衡，同时有固定的交互等位基因频率(alternative allele frequency)fre，因此：

[0014] f(RR)＝(1-fre)2

[0015] f(RA)＝2fre(1-fre)

[0016] $f(AA)=fre^2$

[0017] 本发明中，实际样本由于其基因型未知，认为它是由三种等位基因叠加而成，因此，SNP模型有如下概率模型：

[0018] $$likelihood = const \times \prod_i \begin{matrix} (1-fre)^2(1-p(RR))^{r_i}\,p(RR)^{a_i} \\ +\,2\,fre(1-fre)(1-p(RA))^{r_i}\,p(RA)^{a_i} \\ +\,fre^2(1-p(AA))^{r_i}\,p(AA)^{a_i} \end{matrix}$$

[0019] 当上述模型建立完成后，引入隐变量：$w(RR)_i$、$w(RA)_i$、$w(AA)_i$来表述这个个体的三种基因型概率；使用Expectation-Maximization(EM)算法进行最大似然估计，其E步骤和M步骤分别是：

[0020] E步骤：

[0021] $$w(geno)_i = \frac{f(geno)(1-p(geno))^{r_i}\,p(geno)^{a_i}}{\sum_{geno} f(geno)(1-p(geno))^{r_i}\,p(geno)^{a_i}}$$

[0022] M步骤：

[0023] $$fre = \frac{2\sum_i w(AA)_i + \sum_i w(RA)_i}{2N}$$

[0024] $$p(geno) = \frac{\sum_i w(geno)_i a_i}{\sum_i w(geno)_i a_i + \sum_i w(geno)_i r_i}$$

[0025] 最后，对于第i个样本，取$w(RR)_i$、$w(RA)_i$、$w(AA)_i$中最大的基因型作为该个样本的基因型。

[0026] 本发明中，还对推测出的样本基因型进行对应的质量评估，其包括：构建一个统计量描述，令所选的基因型对应的$1-w(geno)_i$为该基因型的标准质量分数，表征该位点基因分型的准确率；标准质量分数越高，分型准确率越低。

[0027] 初步结果显示，所述标准质量分数能精确的评价分型的准确率，且非常容易在实际工作中使用。可进一步作为实际法医学工作中标准化的质量统计量。

[0028] 为了便于理解，以下将通过具体的实施例对本发明的进行详细地描述。需要特别指出的是，具体实例仅是为了说明，显然本领域的普通技术人员可以根据本文说明，在本发明的范围内对本发明做出各种各样的修正和改变，这些修正和改变也纳入本发明的范围内。

具体实施方式

[0029] 实施例1：对177个特定SNP位点进行分析，数据为729个中国样本上的原始二代测序数据

[0030] 使用二代测序中比对软件Burrows-Wheeler Aligner将原始测序数据映射至参考人类基因组上(human reference genome,hg19)；

[0031] 使用本发明算法对所有729个样本的177SNP位点进行分型，对于其中某个样本的

每个SNP位点而言：

[0032] 首先建立模型，分别提取两个等位基因的有效乘数EBD：

[0033] $$EBD = \sum_{i=1}^{reads} (1-10^{-0.1\times base\_quality_i})(1-10^{-0.1\times mapping\_quality_i})$$

[0034] 对于一个群体，第i个个体的参考等位基因与交互等位基因的EBD分别为ri和ai；对三种可能的基因型RR、RA、AA，假设它们在测序中分别有一个固定的突变等位基因出现率，分别为p(RR)、p(RA)和p(AA)；理想情况下p(RR)接近0，p(RA)接近0.5，p(AA)接近1；假设等位基因频率服从哈迪-温伯格平衡，同时有固定的交互等位基因频率fre，则：

[0035] $f(RR) = (1-fre)^2$

[0036] $f(RA) = 2fre(1-fre)$

[0037] $f(AA) = fre^2$

[0038] 实际样本由于其基因型未知，认为它是由三种等位基因叠加而成，因此SNP模型具有如下概率模型：

[0039] $$likelihood = const \times \prod_i \begin{matrix} (1-fre)^2(1-p(RR))^{r_i}\,p(RR)^{a_i} \\ +\,2\,fre(1-fre)(1-p(RA))^{r_i}\,p(RA)^{a_i} \\ +\,fre^2(1-p(AA))^{r_i}\,p(AA)^{a_i} \end{matrix}$$

[0040] 为了估计上述概率模型的参数，引物隐变量：$w(RR)_i$、$w(RA)_i$、$w(AA)_i$表述所述个体的三种基因型概率；使用Expectation-Maximization(EM)算法进行最大似然估计，其E步骤和M步骤分别是：

[0041] E步骤：

[0042] $$w(geno)_i = \frac{f(geno)(1-p(geno))^{r_i}\,p(geno)^{a_i}}{\sum_{geno} f(geno)(1-p(geno))^{r_i}\,p(geno)^{a_i}}$$

[0043] M步骤：

[0044] $$fre = \frac{2\sum_i w(AA)_i + \sum_i w(RA)_i}{2N}$$

[0045] $$p(geno) = \frac{\sum_i w(geno)_i a_i}{\sum_i w(geno)_i a_i + \sum_i w(geno)_i r_i}$$

[0046] 通过EM算法对$w(RR)_i$、$w(RA)_i$、$w(AA)_i$进行参数估计，i代表第i样本。取$w(RR)_i$、$w(RA)_i$、$w(AA)_i$中最大的基因型作为该样本的基因型，从而完成对该样本中特定SNP位点的基因分型。同时，令所选的基因型对应的$1-w(geno)_i$为该基因型的标准质量分数，表征该位点基因分型的准确率，标准质量分数越高，分型准确率越低。

[0047] 对所有729个样本的所有177个SNP位点，重复上述步骤，从而得到所有基因分型结果与相应的标准质量分数。

[0048] 通过对482个基因型进行Sanger法测序方法进行验证，证明分型结果正确。