(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116364183 A
(43)申请公布日 2023.06.30

(21)申请号 202310245533.5

(22)申请日 2023.03.02

(71)申请人 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)经营地址:深圳市南山区西丽街道松坪山路3号奥特迅电力大厦8楼

(72)发明人 陈实富 耿宇涵 杨博 曹慧雅 锁培苏

(74)专利代理机构 深圳市世纪恒程知识产权代理事务所 44287
专利代理师 何秋石

(51)Int.Cl.
G16B 20/50(2019.01)
G16B 30/00(2019.01)

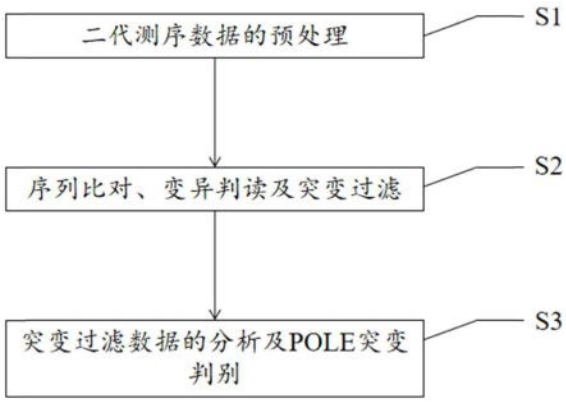
权利要求书2页 说明书14页 附图1页

(54)发明名称

一种基于二代测序结果的POLE突变判别模型的构建方法及其应用

(57)摘要

本发明涉及生物信息分析技术领域,尤其涉及一种基于二代测序结果的POLE突变判别模型的构建方法及其应用,所述构建方法包括S1、二代测序数据的预处理,S2、序列比对、变异判读及突变过滤,S3、突变过滤数据的分析及POLE突变判别。利用本发明的构建方法构建的POLE突变判别模型不需要明确POLE突变作为超扩增标志物,通过设置一定的突变数和最大突变丰度阈值,过滤突变位点过少导致的假阳性,并结合POLE突变判别得分标准,辅助POLE突变的确定,提高真阳性的检出率。同时,本发明的突变检测范围涵盖680个癌症相关基因,较大的检测覆盖度使得给出的得分标准更加接近预估假设,因此较为可靠。



CN 116364183 A

1.一种基于二代测序结果的POLE突变判别模型的构建方法,其特征在于,包括如下步骤:

S1、二代测序数据的预处理

对获得的待测样本和对照样本的二代测序数据进行预过滤,分别获得pileup文件;

所述预过滤包括质控、比对、去重;

所述待测样本为肿瘤患者待测组织DNA或血液游离DNA,所述对照样本为肿瘤患者唾液DNA或血液基因组DNA;

S2、序列比对、变异判读及突变过滤

利用突变检测工具对步骤S1获得pileup文件进行分析,分析数据中存在的与癌症相关基因的突变信息,注释点突变、插入/缺失突变的位点位置与具体碱基变化,标注POLE突变,获取样本所有突变的突变数、最大突变丰度;

根据突变数,过滤假阳性,过滤掉突变数小于等于50的测序数据对应的DNA样本;

S3、突变过滤数据的分析及POLE突变判别

根据步骤S2过滤后的突变数据及设定的阈值,判断POLE突变的致病性。

2.如权利要求1所述的构建方法,其特征在于,所述突变过滤数据的分析具体为:对步骤S2过滤后的数据进行分析,获取碱基C突变为碱基A突变数、碱基T突变为碱基G突变数、碱基C突变为碱基G突变数、点突变总数、插入/缺失突变总数和总突变数;

根据步骤S2注释的POLE突变,获取COSMIC数据库中POLE突变在样本对应肿瘤中的出现次数;

根据癌症相关基因的测序区域BED文件大小,获取肿瘤负荷数据;肿瘤负荷=总突变数/测序区域BED文件大小。

3.如权利要求1所述的构建方法,其特征在于,所述POLE突变判别具体为:根据权利要求2中突变过滤数据的分析的结果,设置6个整数型计分标准:碱基C突变为碱基A突变数占点突变总数的百分比大于20%,记1分,反之不计分;碱基T突变为碱基G突变数占点突变总数的百分比大于4%,记为1分,反之不计分;碱基C突变为碱基G突变数占点突变总数的百分比小于0.6%,记为1分,反之不计分;插入/缺失位点占总突变数的百分比小于5%记为1分,反之不计分;肿瘤负荷大于100muts/mb,记为1分,反之不计分;POLE突变在肿瘤中的出现次数大于1,记为1分,反之不计分;

根据计分标准统计样本POLE突变分数,总分数大于等于4,判断为致病性POLE突变。

4.如权利要求1所述的构建方法,其特征在于,根据获取的突变数和最大突变丰度,过滤假阳性;所述待测样本为肿瘤患者待测组织DNA,过滤掉突变数小于等于50、最大突变丰度小于等于5%的测序数据对应的样本。

5.如权利要求1所述的构建方法,其特征在于,根据获取的的突变数和最大突变丰度,过滤假阳性;所述待测样本为肿瘤患者待测血液游离DNA,过滤掉突变数小于等于50、最大突变丰度小于等于1%的测序数据对应的样本。

6.如权利要求1所述的构建方法,其特征在于,所述突变检测工具为Mutscan或Varscan。

7.如权利要求1至6任一所述的构建方法,其特征在于,所述肿瘤为子宫内膜癌、肺癌、结直肠癌、黑色素瘤、胶质瘤、卵巢癌或尿路上皮癌。

8.如权利要求7所述的构建方法,其特征在于,所述子宫内膜癌为POLE分型子宫内膜癌。

9.一种POLE突变判别模型,其特征在于,利用权利要求1至8任一所述构建方法构建所得。

10.一种计算机可读存储介质,其上存储有计算机程序/指令,其特征在于,该计算机程序/指令被处理器执行时实现权利要求1至8任一所述构建方法的步骤。

一种基于二代测序结果的POLE突变判别模型的构建方法及其应用

技术领域

[0001] 本发明涉及生物信息分析技术领域，尤其涉及一种基于二代测序结果的POLE突变判别模型的构建方法及其应用。

背景技术

[0002] 子宫内膜癌(endometrial carcinoma，EC)是一种常见的妇科癌症，是发生于子宫内膜的一组上皮性恶性肿瘤。其中，POLE超突变子宫内膜癌分型定义为DNA聚合酶ε(POLE)外切核酸酶结构域的体细胞突变引起的肿瘤。POLE分型的子宫内膜癌通常对于化学疗法与免疫疗法较敏感，临床治疗效果好。

[0003] 高肿瘤负荷(TMB>100)与COSMIC特征POLE体细胞突变可以作为一种重要的生物标记物，用于POLE超扩增(突变)肿瘤的诊断。同时，根据文献报道，不同种类碱基替换的分布、点突变和插入缺失的比值也都与POLE超扩增肿瘤的发生存在关联。然而，目前的主流判定方法都基于已知阳性POLE突变位点(P286R，V411L，S297F，A456P，S459F，M444K)，已有的方法在临床诊断上存在缺陷，即在缺少明确POLE突变作为超扩增标志物的情况下无法对样本是否阳性进行判断。同时，在已知高肿瘤负荷与特定碱基替换分布的情况下，没有具有统计学意义、根据大量样本数据验证过后的参数对样本的阳性概率进行评价，也没有可被量化的评价标准，现有的方法进行判断会存在一定比例的假阳性。

[0004] 目前，尚未有更加标准、多维度的判定阈值与流程对肿瘤是否属于POLE突变进行判断。

发明内容

[0005] 本发明的主要目的在于提供一种基于二代测序结果的POLE突变判别模型的构建方法及其应用，旨在提高POLE突变真阳性的检出率。

[0006] 本发明的目的还在于提供一种POLE突变判别模型，利用所述构建方法构建所得。

[0007] 本发明的目的还在于提供一种计算机可读存储介质。

[0008] POLE突变(DNA聚合酶ε[POLE]外切核酸酶结构域的体细胞突变)出现于多种肿瘤疾病中，POLE突变尤其是致病性POLE突变常常作为判别肿瘤疾病的依据，现有的POLE突变的判别需要明确POLE突变作为超扩增标志物，但其判断的假阳性概率偏高。本发明通过特定阈值的设定，先排除一部分假阳性样本，进而根据设定的评分标准判断致病性及潜在致病性的POLE突变，提高了POLE突变真阳性的检出率。

[0009] 为实现上述目的，本发明提供一种基于二代测序结果的POLE突变判别模型的构建方法，包括如下步骤：

[0010] S1、二代测序数据的预处理

[0011] 对获得的待测样本和对照样本的二代测序数据进行预过滤，分别获得pileup文件；pileup文件是指通过BAM文件每个位置重叠的read对比对结果进行的总结，可用于判断

各个位点突变的可能性；

[0012] 所述预过滤包括质控、比对、去重；

[0013] 所述待测样本为肿瘤患者待测组织DNA(ffpedna)或血液游离DNA(cfdna)，所述对照样本为肿瘤患者唾液DNA(sladna)或血液基因组DNA(gdna)；

[0014] 所述质控包括质量评估、去接头序列、去低质量序列、去重复序列。

[0015] 所述比对是将质控后的read与人参考基因对比，获得BAM文件；将通过质控后每一条read与参考基因组进行比对，回贴到基因组上最佳位置。Reads比对前需去掉接头序列。Reads通过双端的模式比对到GRCh37(hg19)版本的人类基因组上，将比对上的reads按染色体位置排序并转换为二进制BAM文件。

[0016] 所述去重是去除PCR重复，校正测序错误，生成用于检测的pileup文件。

[0017] S2、序列比对、变异判读及突变过滤

[0018] 利用突变检测工具对步骤S1获得pileup文件进行分析，分析数据中存在的与癌症相关基因的突变信息，注释点突变、插入/缺失突变的位点位置与具体碱基变化，标注POLE突变，获取样本所有突变的突变数、最大突变丰度；

[0019] 根据突变数，过滤假阳性，过滤掉突变数小于等于50的测序数据对应的DNA样本；

[0020] 分析肿瘤突变时，需采集两组基因组数据：肿瘤基因组(组织DNA[ffpedna]或血液游离DNA[cfdna])与对照基因组(唾液DNA[sladna]或血液基因组DNA[gdna])共同分析(这两个基因组的获取都需要经过质控、对比、去重)。若突变同时存在于肿瘤基因组与对照基因组，则为胚系突变；若突变只存在于肿瘤基因组，则为体细胞突变。本发明后续所讨论、分析的突变均为体细胞突变，因为只有体细胞突变才与癌症的产生有关。

[0021] 所述突变检测工具为Mutscan或Varscan。Mutscan或Varscan能够基于去重后所得的pileup文件对每个位点进行变异鉴定，变异类型包括SNV，Indel。

[0022] S3、突变过滤数据的分析及POLE突变判别

[0023] 根据步骤S2过滤后的突变数据及设定的阈值，判断POLE突变的致病性。

[0024] 优选地，所述突变过滤数据的分析具体为：对步骤S2过滤后的数据进行分析，获取碱基C突变为碱基A(C>A)突变数、碱基T突变为碱基G(T>G)突变数、碱基C突变为碱基G(C>G)突变数、点突变总数、插入/缺失突变总数和总突变数；

[0025] 根据步骤S2注释的POLE突变，获取COSMIC数据库中POLE突变在样本对应肿瘤中的出现次数；COSMIC是在人类癌症中发现的体细胞突变的在线数据库；

[0026] 根据癌症相关基因的测序区域BED文件大小(即癌症相关基因覆盖区域的BED格式的文件)，获取肿瘤负荷(TMB)数据；肿瘤负荷(TMB)＝总突变数/测序区域BED文件大小，单位：muts/mb；

[0027] 优选地，所述POLE突变判别具体为：根据突变过滤数据的分析的结果，设置6个整数型计分标准：碱基C突变为碱基A(C>A)突变数占点突变总数的百分比大于20％，记1分，反之不计分；碱基T突变为碱基G(T>G)突变数占点突变总数的百分比大于4％，记为1分，反之不计分；碱基C突变为碱基G(C>G)突变数占点突变(SNV)总数的百分比小于0.6％，记为1分，反之不计分；插入/缺失位点(Indel)占总突变数的百分比小于5％记为1分，反之不计分；肿瘤负荷(TMB)大于100muts/mb，记为1分，反之不计分；POLE突变在肿瘤中的出现次数大于1，记为1分，反之不计分；

[0028] 根据计分标准统计样本POLE突变分数，总分数大于等于4，判断为致病性POLE突变。

[0029] 优选地，根据计分标准统计样本POLE突变分数，总分数等于3，判断为未知致病性POLE突变；总分数小于3，判断为无致病性POLE突变。

[0030] 优选地，所述癌症相关的基因为癌症相关的680个基因。

[0031] 优选地，所述癌症相关基因的测序区域BED文件大小为680个基因的测序区域BED文件大小(即680基因覆盖区域的BED格式的文件)。

[0032] 优选地，根据获取的突变数和最大突变丰度，过滤假阳性；所述待测样本为肿瘤患者待测组织DNA(ffpedna)，过滤掉突变数小于等于50、最大突变丰度小于等于5％的测序数据对应的样本；

[0033] 进一步优选地，根据获取的的突变数和最大突变丰度，过滤假阳性；所述待测样本为肿瘤患者待测血液游离DNA(cfdna)，过滤掉突变数小于等于50、最大突变丰度小于等于1％的测序数据对应的样本；

[0034] 优选地，所述肿瘤为子宫内膜癌、肺癌、结直肠癌、黑色素瘤、胶质瘤、卵巢癌、尿路上皮癌。

[0035] 进一步优选地，所述肿瘤为子宫内膜癌、肺癌、结直肠癌。

[0036] 进一步优选地，所述肿瘤为子宫内膜癌。

[0037] 进一步优选地，所述子宫内膜癌为POLE分型子宫内膜癌。

[0038] 为实现上述目的，本发明提供一种POLE突变判别模型，利用所述构建方法构建所得。所述POLE突变判别模型能够更准确的判别出POLE突变，并分辨出致病性与非致病性POLE突变，发现潜在的致病性POLE突变。

[0039] 为实现上述目的，本发明的目的还提供一种计算机可读存储介质，其上存储有计算机程序/指令，该计算机程序/指令被处理器执行时实现上述构建方法的步骤。

[0040] 本发明在缺少明确POLE突变作为超扩增标志物的情况下，通过设置一定的突变数和最大突变丰度阈值，过滤突变位点过少导致的假阳性，并结合POLE突变判别得分标准，辅助POLE突变的确定，提高真阳性的检出率。同时，本发明的突变检测范围涵盖680个癌症相关基因，较大的检测覆盖度使得给出的得分标准更加接近预估假设，因此较为可靠。

附图说明

[0041] 图1为本发明实施例方案涉及的基于二代测序结果的POLE突变判别模型的构建方法的流程示意图。

[0042] 图2为本发明实施例方案涉及的计算机可读存储介质结构示意图。

[0043] 本发明目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

[0044] 应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0045] 本发明通过设置的突变数和最大突变丰度的阈值，过滤突变位点过少导致的假阳性，对真阳性的样本测序数据中的POLE突变进行评分，不仅可以分辨出致病性与非致病性POLE突变，还能发现潜在的POLE突变，提高了POLE突变真阳性的检出率。

[0046] 本发明的突变检测范围涵盖680个癌症相关基因，较大的检测覆盖度使得给出的得分标准更加接近预估假设，因此较为可靠。

[0047] 本发明提供一种基于二代测序结果的POLE突变判别模型的构建方法，包括如下步骤：

[0048] S1、二代测序数据的预处理

[0049] 对获得的待测样本和对照样本的二代测序数据进行预过滤，分别获得pileup文件；

[0050] 所述预过滤包括质控、比对、去重；

[0051] 所述待测样本为肿瘤患者待测组织DNA(ffpedna)或血液游离DNA(cfdna)，所述对照样本为肿瘤患者唾液DNA(sladna)或血液基因组DNA(gdna)；

[0052] 所述质控包括质量评估、去接头序列、去低质量序列、去重复序列；

[0053] 所述比对是将质控后的read与人参考基因对比，获得BAM文件；将通过质控后每一条read与参考基因组进行比对，回贴到基因组上最佳位置；Reads比对前需去掉接头序列；Reads通过双端的模式比对到GRCh37(hg19)版本的人类基因组上，将比对上的reads按染色体位置排序并转换为二进制BAM文件；

[0054] 所述去重是去除PCR重复，校正测序错误，生成用于检测的pileup文件。

[0055] S2、序列比对、变异判读及突变过滤

[0056] 利用突变检测工具对步骤S1获得pileup文件进行分析，分析数据中存在的与癌症相关基因的突变信息，注释点突变(SNV)、插入/缺失(Indel)突变的位点位置与具体碱基变化、标注POLE突变，获取样本所有突变的突变数、最大突变丰度；

[0057] 所述突变检测工具为Mutscan或Varscan；

[0058] 根据突变数，过滤假阳性，过滤掉突变数小于等于50的测序数据对应的DNA样本；

[0059] S3、突变过滤数据的分析及POLE突变判别

[0060] 根据步骤S2过滤后的突变数据及设定的阈值，判断POLE突变的致病性。

[0061] 优选地，所述突变过滤数据的分析具体为：对步骤S2过滤后的数据进行分析，获取碱基C突变为碱基A(C>A)突变数、碱基T突变为碱基G(T>G)突变数、碱基C突变为碱基G(C>G)突变数、点突变总数、插入/缺失突变总数和总突变数；

[0062] 根据步骤S2注释的POLE突变，获取COSMIC数据库中POLE突变在样本对应肿瘤中的出现次数；

[0063] 根据癌症相关基因的测序区域BED文件大小，获取肿瘤负荷(TMB)数据；肿瘤负荷(TMB)＝总突变数/测序区域BED文件大小，单位：muts/mb。

[0064] 优选地，所述POLE突变判别具体为：根据突变过滤数据的分析的结果，设置6个整数型计分标准：碱基C突变为碱基A(C>A)突变数占点突变总数的百分比大于20％，记1分，反之不计分；碱基T突变为碱基G(T>G)突变数占点突变总数的百分比大于4％，记为1分，反之不计分；碱基C突变为碱基G(C>G)突变数占点突变总数的百分比小于0.6％，记为1分，反之不计分；插入/缺失位点(Indel)占总突变数的百分比小于5％记为1分，反之不计分；肿瘤负荷(TMB)大于100muts/mb，记为1分，反之不计分；POLE突变在肿瘤中的出现次数大于1，记为1分，反之不计分；

[0065] 根据计分标准统计样本POLE突变分数，总分数大于等于4，判断为致病性POLE突

变。

[0066] 优选地，根据计分标准统计样本POLE突变分数，总分数等于3，判断为未知致病性POLE突变；总分数小于3，判断为无致病性POLE突变。

[0067] 优选地，所述癌症相关的基因为癌症相关的680个基因。

[0068] 优选地，所述癌症相关基因的测序区域BED文件大小为680个基因的测序区域BED文件大小(即680基因覆盖区域的BED格式的文件)。

[0069] 分析肿瘤突变时，需采集两组基因组数据：肿瘤基因组(组织DNA[ffpedna]或血液游离DNA[cfdna])与对照基因组(唾液DNA[sladna]或血液基因组DNA[gdna])共同分析(这两个基因组的获取都需要经过质控、对比、去重)。若突变同时存在于肿瘤基因组与对照基因组，则为胚系突变；若突变只存在于肿瘤基因组，则为体细胞突变。本发明后续所讨论、分析的突变均为体细胞突变，因为只有体细胞突变才与癌症的产生有关。

[0070] Mutscan或Varscan等突变检测工具能够基于去重后所得的pileup文件对每个位点进行变异鉴定，变异类型包括点突变(SNV)、插入/缺失(Indel)突变。

[0071] 优选地，本发明根据获取的突变数和最大突变丰度，过滤假阳性；所述待测样本为肿瘤患者待测组织DNA(ffpedna)，过滤掉突变数小于等于50、最大突变丰度小于等于5％的测序数据对应的样本；

[0072] 进一步优选地，本发明根据获取的的突变数和最大突变丰度，过滤假阳性；所述待测样本为肿瘤患者待测血液游离DNA(cfdna)，过滤掉突变数小于等于50、最大突变丰度小于等于1％的测序数据对应的样本；

[0073] 优选地，所述肿瘤为子宫内膜癌、肺癌、结直肠癌、黑色素瘤、胶质瘤、卵巢癌、尿路上皮癌。

[0074] 进一步优选地，所述肿瘤为子宫内膜癌、肺癌、结直肠癌。

[0075] 进一步优选地，所述肿瘤为子宫内膜癌。

[0076] 进一步优选地，所述子宫内膜癌为POLE分型子宫内膜癌。

[0077] 利用本发明的基于二代测序结果的POLE突变判别模型的构建方法构建POLE突变判别模型。所述POLE突变判别模型能够更准确的判别出POLE突变，并分辨出致病性与非致病性POLE突变，发现潜在的致病性POLE突变。所述模型能够适用于子宫内膜癌、肺癌、结直肠癌、黑色素瘤、胶质瘤、卵巢癌、尿路上皮癌等多种肿瘤的POLE突变的判别。

[0078] 本发明还提供一种计算机可读存储介质，其上存储有计算机程序/指令，该计算机程序/指令被处理器执行时实现本发明所述构建方法的步骤。

[0079] 实施例1一种基于二代测序结果的POLE突变判别模型的构建方法及POLE突变判别模型

[0080] 一种基于二代测序结果的POLE突变判别模型的构建方法，所述构建方法的流程示意图如图1所示。

[0081] 所述构建方法包括如下步骤：

[0082] S1、二代测序数据的预处理

[0083] 对获得的待测样本和对照样本的二代测序数据进行预过滤，分别获得pileup文件；

[0084] 所述待测样本为肿瘤患者待测组织DNA(ffpedna)或血液游离DNA(cfdna)，所述对

照样本为肿瘤患者唾液DNA(sladna)或血液基因组DNA(gdna)；

[0085] 所述预过滤包括质控、比对、去重；

[0086] 所述质控包括质量评估、去接头序列、去低质量序列、去重复序列；

[0087] 所述比对是将质控后的read与人参考基因对比，获得BAM文件；将通过质控后每一条read与参考基因组进行比对，回贴到基因组上最佳位置；Reads比对前需去掉接头序列；Reads通过双端的模式比对到GRCh37(hg19)版本的人类基因组上，将比对上的reads按染色体位置排序并转换为二进制BAM文件；

[0088] 所述去重是去除PCR重复，校正测序错误，生成用于检测的pileup文件。

[0089] S2、序列比对、变异判读及突变过滤

[0090] 利用突变检测工具Mutscan或Varscan对步骤S1获得pileup文件进行分析，分析数据中存在的与癌症相关的680个基因的突变信息，注释SNV、Indel突变的位点位置与具体碱基变化，标注POLE突变，获取样本所有突变的突变数、最大突变丰度；

[0091] 根据突变数过滤假阳性，过滤掉突变数小于等于50的测序数据对应的DNA样本；

[0092] S3、突变过滤数据的分析及POLE突变判别

[0093] 根据步骤S2过滤后的突变数据及设定的阈值，判断POLE突变的致病性。

[0094] 所述突变过滤数据的分析具体为：对步骤S2过滤后的数据进行分析，获取碱基C突变为碱基A(C>A)突变数、碱基T突变为碱基G(T>G)突变数、碱基C突变为碱基G(C>G)突变数、点突变总数、插入/缺失突变总数和总突变数；

[0095] 根据步骤S2注释的POLE突变，获取COSMIC数据库中POLE突变在样本对应肿瘤中的出现次数；

[0096] 根据680个基因的测序区域BED文件大小，获取肿瘤负荷(TMB)数据；肿瘤负荷(TMB)＝总突变数/测序区域BED文件大小，单位：muts/mb；

[0097] 所述POLE突变判别具体为：根据突变过滤数据的分析的结果，设置6个整数型计分标准：碱基C突变为碱基A(C>A)突变数占点突变总数的百分比大于20％，记1分，反之不计分；碱基T突变为碱基G(T>G)突变数占点突变总数的百分比大于4％，记为1分，反之不计分；碱基C突变为碱基G(C>G)突变数占点突变总数的百分比小于0.6％，记为1分，反之不计分；插入/缺失位点(Indel)占总突变数的百分比小于5％记为1分，反之不计分；肿瘤负荷(TMB)大于100muts/mb，记为1分，反之不计分；POLE突变在肿瘤中的出现次数大于1，记为1分，反之不计分；

[0098] 根据计分标准统计样本POLE突变分数，总分数大于等于4，判断为致病性POLE突变；

[0099] 总分数等于3，判断为未知致病性POLE突变；

[0100] 总分数小于3，判断为无致病性POLE突变。

[0101] 利用上述构建方法构建即得POLE突变判别模型。

[0102] 680个基因的信息如下表1所示。

[0103] 表1

[0104]

名称	转录本号	名称	转录本号	名称	转录本号	名称	转录本号
ABCB1	NM_001348946.2	ELAC2	NM_018127.7	KLLN	NM_001126049.2	RAF1	NM_002880.4
ABCC1	NM_004996.4	ELOC	NM_005648.4	KMT2A	NM_001197104.2	RARA	NM_000964.4
ABCC2	NM_000392.5	EML4	NM_019063.5	KMT2B	NM_014727.3	RB1	NM_000321.3
ABCC4	NM_005845.5	ENOSF1	NM_017512.7	KMT2C	NM_170606.3	RBFOX1	NM_018723.4
ABCC5	NM_005688.4	EP300	NM_001429.4	KMT2D	NM_003482.4	RBM10	NM_005676.5
ABCC11	NM_001370497.1	EPAS1	NM_001430.5	KRAS	NM_004985.5	RECK	NM_021111.3
ABCG1	NM_016818.3	EPCAM	NM_002354.3	KRT5	NM_000424.4	RECQL	NM_002907.4
ABCG2	NM_004827.3	EPHA2	NM_004431.5	KRT14	NM_000526.5	RECQL4	NM_004260.4
ABL1	NM_005157.6	EPHA3	NM_005233.6	KRT15	NM_002275.4	REL	NM_001291746.2
ABRAXAS1	NM_139076.3	EPHA5	NM_001281766.3	LARP4	NM_052879.5	RET	NM_020975.6
ACSS2	NM_018677.4	EPHA7	NM_004440.4	LATS1	NM_004690.4	REV3L	NM_001372078.1
ACTL6A	NM_004301.5	EPHB1	NM_004441.5	LATS2	NM_014572.3	RFC1	NM_002913.5
ACVR1	NM_001111067.4	EPHX1	NM_001136018.4	LBR	NM_002296.4	RFC4	NM_002916.5
ADCY2	NM_020546.3	ERBB2	NM_004448.4	LGALS8	NM_201544.4	RGS5	NM_003617.4
ADH1B	NM_000668.6	ERBB3	NM_001982.4	LGR5	NM_003667.4	RHBDF2	NM_001005498.4
ADH1C	NM_000669.5	ERBB4	NM_005235.3	LIG3	NM_013975.4	RHEB	NM_005614.4
AIP	NM_003977.4	ERCC1	NM_001983.4	LIG4	NM_206937.2	RHOA	NM_001664.4
AKR1C3	NM_003739.6	ERCC2	NM_000400.4	LMO1	NM_002315.3	RICTOR	NM_152756.5
AKT1	NM_001382430.1	ERCC3	NM_000122.2	LRIG3	NM_153377.5	RIF1	NM_018151.5
AKT2	NM_001626.6	ERCC4	NM_005236.3	LRMDA	NM_001305581.2	RILP	NM_031430.3
AKT3	NM_005465.7	ERCC5	NM_000123.4	LRP1B	NM_018557.3	RINT1	NM_021930.6
ALDH2	NM_000690.4	ERCC6	NM_000124.4	LRP2	NM_004525.3	RIT1	NM_006912.6
ALK	NM_004304.5	EREG	NM_001432.3	LYN	NM_002350.4	RNASEL	NM_021133.4
ALOX12	NM_000697.3	ERG	NM_182918.4	LZTR1	NM_006767.4	RNF43	NM_017763.6
AMER1	NM_152424.4	ERRFI1	NM_018948.4	MAD1L1	NM_001013836.2	ROBO2	NM_002942
ANXA5	NM_001154.4	ESR1	NM_000125.4	MAP2K1	NM_002755.4	ROS1	NM_001378902.1
APC	NM_000038.6	ESR2	NM_001437.3	MAP2K2	NM_030662.4	RPA4	NM_013347.4

[0105]

名称	转录本号	名称	转录本号	名称	转录本号	名称	转录本号
APLF	NM_173545.3	ETV1	NM_004956.5	MAP2K4	NM_003010.4	RPS6KB1	NM_003161.4
APOB	NM_000384.3	ETV4	NM_001079675.5	MAP3K1	NM_005921.2	RPS20	NM_001023.4
AR	NM_000044.6	ETV6	NM_001987.5	MAP3K6	NM_004672.5	RPTOR	NM_020761.3
ARAF	NM_001654.5	EWSR1	NM_005243.4	MAP4K4	NM_004834	RRAS2	NM_012250.6
AREG	NM_001657.4	EXT1	NM_000127.3	MAPK1	NM_002745.5	RRM1	NM_001033.5
ARID1A	NM_006015.6	EXT2	NM_207122.2	MAPK3	NM_002746.3	RSF1	NM_016578.4
ARID1B	NM_001374828.1	EZH2	NM_004456.5	MAPKBP1	NM_014994.3	RUNX1	NM_001754.5
ARID2	NM_152641.4	FANCA	NM_000135.4	MAX	NM_002382.5	SBDS	NM_016038.4
ARMS2	NM_001099667.3	FANCB	NM_001018113.3	MCL1	NM_021960.5	SCN10A	NM_006514.4
ASNS	NM_001673.5	FANCC	NM_000136.3	MDH2	NM_005918.4	SDHA	NM_004168.4
ASPH	NM_004318.4	FANCG	NM_004629.2	MDM2	NM_002392.6	SDHAF2	NM_017841.4
ASXL1	NM_015338.6	FANCI	NM_001113378.2	MDM4	NM_002393.5	SDHB	NM_003000.3
ATM	NM_000051.4	FANCL	NM_018062.4	MED12	NM_005120.3	SDHC	NM_003001.5
ATP7B	NM_000053.4	FANCM	NM_020937.4	MEF2B	NM_001145785.2	SDHD	NM_003002.4
ATR	NM_001184.4	FAT1	NM_005245.4	MEN1	NM_001370259.2	SELE	NM_000450.2
ATRIP	NM_130384.3	FBN3	NM_032447.5	MET	NM_000245.4	SELL	NM_000655.5
ATRX	NM_000489.6	FBXW7	NM_001349798.2	MGAT4A	NM_012214.3	SEMA3C	NM_006379.5
AURKA	NM_198437.3	FCGR2A	NM_001136219.3	MITF	NM_001354604.2	SERPINB3	NM_006919.3
AURKB	NM_004217.4	FCGR3A	NM_000569.8	MKI67	NM_002417.5	SERPINB4	NM_002974.4
AXIN1	NM_003502.4	FGF1	NM_000800.5	MLH1	NM_000249.4	SETBP1	NM_015559.3
AXIN2	NM_004655.4	FGF2	NM_001361665.2	MLH3	NM_001040108.2	SETD2	NM_014159.7
AXL	NM_021913.5	FGF3	NM_005247.4	MPL	NM_005373.3	SETD7	NM_030648.4
B2M	NM_004048.4	FGF4	NM_002007.4	MRE11	NM_005591.4	SF3B1	NM_012433.4
BABAM2	NM_199191.3	FGF5	NM_004464.4	MSH2	NM_000251.3	SH2B3	NM_005475.3
BAP1	NM_004656.4	FGF6	NM_020996.3	MSH3	NM_002439.5	SHMT1	NM_004169.5
BARD1	NM_000465.4	FGF7	NM_002009.4	MSH6	NM_000179.3	SHOX	NM_000451.4
BCL2	NM_000633.3	FGF8	NM_033163.5	MST1R	NM_002447.4	SHPRH	NM_001042683.3
BCL2L1	NM_138578.3	FGF9	NM_002010.3	MTHFR	NM_005957.5	SLC15A2	NM_021082.4
BCL2L11	NM_138621.5	FGF10	NM_004465.2	MTOR	NM_004958.4	SLC19A1	NM_194255.4
BCL6	NM_001706.5	FGF19	NM_005117.3	MTUS1	NM_001363059.2	SLC22A1	NM_003057.3
BCOR	NM_001123385.2	FGF23	NM_020638.3	MUTYH	NM_001048174.2	SLC22A2	NM_003058.4
BCORL1	NM_001379451.1	FGFR1	NM_023110.3	MYC	NM_002467.6	SLC22A4	NM_003059.3
BLM	NM_000057.4	FGFR2	NM_000141.5	MYCL	NM_001033081.3	SLC22A5	NM_003060.4
BMPR1A	NM_004329.3	FGFR3	NM_000142.5	MYCN	NM_005378.6	SLC22A16	NM_033125.4
BRAF	NM_004333.6	FGFR4	NM_213647.3	MYD88	NM_002468.5	SLC28A1	NM_004213.5
BRCA1	NM_007294.4	FH	NM_000143.4	MYOD1	NM_002478.5	SLC28A2	NM_004212.4

[0106]

名称	转录本号	名称	转录本号	名称	转录本号	名称	转录本号
BRCA2	NM_000059.4	FLCN	NM_144997.7	NAB2	NM_005967.4	SLC29A1	NM_001372327.1
BRD2	NM_005104.4	FLT1	NM_002019.4	NAT2	NM_000015.3	SLC31A1	NM_001859.4
BRD4	NM_001379291.1	FLT3	NM_004119.3	NBN	NM_002485.5	SLC34A2	NM_006424.3
BRIP1	NM_032043.3	FLT4	NM_182925.5	NCOA1	NM_003743.5	SLCO1B1	NM_006446.5
BTK	NM_000061.3	FNTB	NM_002028.4	NCOA3	NM_181659.3	SLCO1B3	NM_019844.4
BUB1	NM_004336.5	FOLR3	NM_000804.4	NCOA4	NM_001145263.2	SLX4	NM_032444.4
C8orf34	NM_052958.4	FOXA1	NM_004496.5	NF1	NM_001042492.3	SMAD2	NM_005901.6
CACNA1C	NM_000719.7	FOXK2	NM_004514.4	NF2	NM_000268.4	SMAD3	NM_005902.4
CADM2	NM_001167675.2	FOXL2	NM_023067.4	NFE2L2	NM_006164.5	SMAD4	NM_005359.6
CALR	NM_004343.4	FOXM1	NM_021953.4	NFKBIA	NM_020529.3	SMARCA4	NM_003072.5
CAMTA1	NM_015215.4	FOXP1	NM_001349338.3	NKX2-1	NM_001079668.3	SMARCB1	NM_003073.5
CAPN2	NM_001748.5	FOXP2	NM_014491.4	NOS3	NM_000603.5	SMO	NM_005631.5
CARD11	NM_032415.7	FRS2	NM_001278356.2	NOTCH1	NM_017617.5	SOCS1	NM_003745.2
CASP7	NM_001227.5	FUBP1	NM_003902.5	NOTCH2	NM_024408.4	SOCS6	NM_004232.4
CASP8	NM_001372051.1	FUS	NM_004960.4	NOTCH3	NM_000435.3	SOD2	NM_000636.4
CBFB	NM_022845.3	GAB2	NM_080491.3	NOVA1	NM_002515.3	SOX2	NM_003106.4
CBL	NM_005188.4	GALNT12	NM_024642.5	NPM1	NM_002520.7	SOX4	NM_003107.3
CBLB	NM_170662.5	GALNT14	NM_024572.4	NQO1	NM_000903.3	SOX9	NM_000346.4
CBR1	NM_001757.4	GATA1	NM_002049.4	NQO2	NM_000904.6	SOX10	NM_006941.4
CBR3	NM_001236.4	GATA2	NM_032638.5	NR1I2	NM_003889.4	SPEN	NM_015001.3
CCDC6	NM_005436.5	GATA3	NM_001002295.2	NR4A3	NM_006981.4	SPINK1	NM_001379610.1
CCL18	NM_002988.4	GATA6	NM_005257.6	NRAS	NM_002524.5	SPOP	NM_001007228.2
CCND1	NM_053056.3	GEMIN6	NM_024775.10	NRG1	NM_013964.5	SRC	NM_198291.3
CCND2	NM_001759.4	GEN1	NM_001130009.3	NSD1	NM_022455.5	SRD5A2	NM_000348.4
CCND3	NM_001760.5	GGH	NM_003878.3	NT5C2	NM_001351169.2	SRSF2	NM_001195427.2
CCNE1	NM_001238.4	GK5	NM_001039547.3	NTHL1	NM_002528.7	SS18	NM_001007559.3
CD74	NM_001025159.3	GLI1	NM_005269.3	NTRK1	NM_002529.4	STAG2	NM_001042750.2
CD79A	NM_001783.4	GLIPR1	NM_006851.3	NTRK2	NM_006180.6	STAT3	NM_139276.3
CD79B	NM_000626.4	GLRX	NM_001118890.2	NTRK3	NM_001012338.3	STK11	NM_000455.5
CD274	NM_014143.4	GMEB1	NM_001319674.2	NUP93	NM_014669.5	SUFU	NM_016169.4
CDA	NM_001785.3	GNA11	NM_002067.5	NUTM1	NM_001284292.2	SULT1A1	NM_001055.4
CDC25C	NM_001790.5	GNAQ	NM_002072.5	OPRM1	NM_000914.5	SUZ12	NM_015355.4

[0107]

名称	转录本号	名称	转录本号	名称	转录本号	名称	转录本号
CDC73	NM_024529.5	GNAS	NM_000516.7	OTOS	NM_148961.4	SWI5	NM_001318089.2
CDH1	NM_004360.5	GPER1	NM_001098201.3	PAK1	NM_002576.5	SYK	NM_003177.7
CDK2	NM_001798.5	GPRIN2	NM_001385282.1	PALB2	NM_024675.4	SYNE1	NM_182961.4
CDK4	NM_000075.4	GPX5	NM_001509.3	PALLD	NM_001166108.2	TACSTD2	NM_002353.3
CDK6	NM_001145306.2	GREM1	NM_013372.7	PARD3B	NM_001302769.2	TBX3	NM_005996.4
CDK8	NM_001260.3	GRIN2A	NM_001134407.3	PARK7	NM_007262.5	TENT5C	NM_017709.4
CDK12	NM_016507.4	GRM1	NM_001278064.2	PARP1	NM_001618.4	TERT	NM_198253.3
CDKN1A	NM_000389.5	GSK3B	NM_001146156.2	PAX3	NM_181458.4	TET1	NM_030625.3
CDKN1B	NM_004064.5	GSR	NM_000637.5	PAX5	NM_016734.3	TET2	NM_001127208.3
CDKN1C	NM_001122630.2	GSTA1	NM_145740.5	PAX7	NM_001135254.2	TFE3	NM_006521.6
CDKN2A	NM_000077.5	GSTM3	NM_000849.5	PAX8	NM_003466.4	TGFB1	NM_000660.7
CDKN2B	NM_004936.4	GSTP1	NM_000852.4	PBRM1	NM_018313	TGFBR2	NM_003242.6
CDKN2C	NM_078626.3	H3-3A	NM_002107.7	PCBP1	NM_006196.4	TMEM127	NM_017849.4
CEBPA	NM_004364.5	HDAC2	NM_001527.4	PCLO	NM_033026.6	TMPRSS2	NM_005656.4
CES1	NM_001025195.2	HERC2	NM_004667.6	PDCD1	NM_005018.3	TNF	NM_000594.4
CFD	NM_001928.4	HFM1	NM_001017975.6	PDCD1LG2	NM_025239.4	TNFAIP3	NM_001270508.2
CFH	NM_000186.4	HGF	NM_000601.6	PDGFB	NM_002608.4	TNFRSF11B	NM_002546.4
CFTR	NM_000492.4	HIF1A	NM_001530.4	PDGFRA	NM_006206.6	TNFRSF14	NM_003820.4
CHD4	NM_001273.5	HJV	NM_213653.4	PDGFRB	NM_002609.4	TNFRSF19	NM_148957.4
CHEK1	NM_001114122.3	HKDC1	NM_025130.4	PDPK1	NM_002613.5	TNFSF8	NM_001244.4
CHEK2	NM_007194.4	HLA-A	NM_002116.8	PGR	NM_000926.4	TNFSF11	NM_003701.4
CIC	NM_015125	HLA-B	NM_005514.8	PIAS4	NM_015897.4	TOP1	NM_003286.4
CLK2	NM_001294338.2	HLA-C	NM_002117.6	PIGB	NM_004855.5	TOP2A	NM_001067.4
CMPK1	NM_016308.3	HLA-DPB1	NM_002121.6	PIK3CA	NM_006218.4	TP53	NM_000546.6
CNTNAP5	NM_001367498.1	HLA-DQB1	NM_002123.5	PIK3CB	NM_006219.3	TP53BP1	NM_001141980.3
CREBBP	NM_004380.3	HLA-DRB1	NM_002124.4	PIK3CD	NM_005026.5	TPMT	NM_000367.5
CRKL	NM_005207.4	HLA-G	NM_001384290.1	PIK3CG	NM_001282426.2	TRAF1	NM_005658.5
CRLF2	NM_022148.4	HMGA2	NM_003483.6	PIK3R1	NM_181523.3	TREX2	NM_080701.4
CSDE1	NM_001007553.3	HMGCR	NM_000859.3	PIK3R2	NM_005027.4	TSC1	NM_000368.5
CSF1R	NM_001288705.3	HNF1A	NM_000545.8	PIM1	NM_002648.4	TSC2	NM_000548.5
CSF3R	NM_000760.4	HNF1B	NM_000458.4	PLAUR	NM_002659.4	TSHR	NM_000369.5
CSMD3	NM_198123.2	HOXB13	NM_006361.6	PLCG2	NM_002661.5	TSPAN31	NM_005981.5

[0108]

名称	转录本号	名称	转录本号	名称	转录本号	名称	转录本号
CSNK1A1	NM_001892.6	HPSE2	NM_021828.5	PLIN2	NM_001122.4	TTK	NM_003318.5
CSNK2A1	NM_177559.3	HRAS	NM_005343.4	PMS1	NM_000534.5	TUBB1	NM_030773.4
CTCF	NM_006565.4	HSD3B2	NM_000198.4	PMS2	NM_000535.7	TYMS	NM_001071.4
CTNNA1	NM_001903.5	HSD17B3	NM_000197.2	POLD1	NM_002691.4	U2AF1	NM_006758.3
CTNNB1	NM_001904.4	HSP90AA1	NM_005348.4	POLE	NM_006231.4	UBE2A	NM_003336.4
CUL3	NM_003590.5	HSPA5	NM_005347.5	POLH	NM_006502.3	UBE2I	NM_003345.5
CXCR4	NM_003467.3	HTRA1	NM_002775.5	POLQ	NM_199420.4	UBE2V2	NM_003350.3
CXXC4	NM_025212.4	IDH1	NM_005896.4	POR	NM_000941	VEGFA	NM_003376.6
CYLD	NM_001378743.1	IDH2	NM_002168.4	PPIB	NM_000942.5	VEGFC	NM_005429.5
CYP1A1	NM_001319217.2	IFNGR1	NM_000416.3	PPP2R1A	NM_014225.6	VHL	NM_000551.4
CYP1A2	NM_000761.5	IFNGR2	NM_005534.4	PPP2R2A	NM_002717.4	WAS	NM_000377.3
CYP1B1	NM_000104.4	IFNL2	NM_172138.2	PPP4R4	NM_058237.2	WIF1	NM_007191.5
CYP2B6	NM_000767.5	IFNLR1	NM_170743.4	PRDM1	NM_001198.4	WNK2	NM_006648.4
CYP2C8	NM_000770.3	IGF1R	NM_000875.5	PRDX4	NM_006406.2	WNT5B	NM_032642.3
CYP2C9	NM_000771.4	IGF2	NM_000612.6	PREX2	NM_024870.4	WRN	NM_000553.6
CYP2C19	NM_000769.4	IGFBP3	NM_000598.5	PRKACA	NM_002730.4	WT1	NM_024426.6
CYP2D6	NM_000106.6	IKBKE	NM_014002.4	PRKACB	NM_182948.4	XBP1	NM_001079539.2
CYP2E1	NM_000773.4	IKZF1	NM_006060.6	PRKAR1A	NM_002734.5	XPA	NM_000380.4
CYP3A4	NM_017460.6	IL1B	NM_000576.3	PRKCI	NM_002740.6	XPC	NM_004628.5
CYP3A5	NM_000777.5	IL7R	NM_002185.5	PRKDC	NM_006904.7	XPO1	NM_003400.4
CYP19A1	NM_000103.4	IL13	NM_002188.3	PRKN	NM_004562.3	XRCC1	NM_006297.3
DAXX	NM_001141969.2	IL16	NM_172217.5	PRSS1	NM_002769.5	XRCC2	NM_005431.2
DDB2	NM_000107.3	IL23R	NM_144701.3	PSME2	NM_002818.3	XRCC4	NM_003401.5
DDIT3	NM_004083.6	INHBA	NM_002192.4	PTCH1	NM_000264.5	YAP1	NM_001130145.3
DDR2	NM_006182.4	INPP4B	NM_001101669.3	PTCH2	NM_003738.5	YES1	NM_005433.4
DDX3X	NM_001356.5	IRF1	NM_002198.3	PTEN	NM_000314.8	ZBTB16	NM_006006.6
DDX51	NM_175066.4	IRF4	NM_002460.4	PTGER4	NM_000958.3	ZNF367	NM_153695.4
DHFR	NM_000791.4	IRS2	NM_003749.3	PTGES	NM_004878.5	ZNF423	NM_001379286.1
DICER1	NM_177438.3	JAK1	NM_002227.4	PTGS2	NM_000963.4	ZNF750	NM_024702.3
DNMT1	NM_001130823.3	JAK2	NM_004972.4	PTN	NM_002825.7	ZRSR2	NM_005089.4
DNMT3A	NM_022552.5	JAK3	NM_000215.4	PTPN11	NM_002834.5	CASC8	NR_024393.1
DOT1L	NM_032482.3	JUN	NM_002228.4	PTPRD	NM_002839.4	CCAT2	NR_109834.1
DPYD	NM_000110.4	KCNJ5	NM_000890.5	PTPRT	NM_007050.6	H19	NR_002196.2
DSCAM	NM_001389.5	KDM5A	NM_001042603.3	PZP	NM_002864.3	HOTAIR	NR_003716.3

[0109]

名称	转录本号	名称	转录本号	名称	转录本号	名称	转录本号
DYNC2H1	NM_001377.3	KDM5C	NM_004187.5	RAC1	NM_006908.5	MALAT1	NR_002819.4
E2F7	NM_203394.3	KDM6A	NM_001291415.2	RAD21	NM_006265.3	MIR3936HG	NR_110997.1
ECT2L	NM_001077706.3	KDR	NM_002253.4	RAD50	NM_005732.4	MIR4713HG	NR_146310.1
EDN1	NM_001955.5	KEAP1	NM_203500.2	RAD51	NM_002875.5	TERC	NR_001566.1
EED	NM_003797.5	KIF1B	NM_001365951.3	RAD51B	NM_133510.4	TRAM2-AS1	NR_103446.1
EEF1A1	NM_001402.6	KIF5B	NM_004521.3	RAD51C	NM_058216.3	UGT1A1	NM_000463.3
EGF	NM_001963.6	KIT	NM_000222.3	RAD51D	NM_002878.4	UGT1A4	NM_007120.3
EGFR	NM_005228.5	KLC1	NM_005552	RAD52	NM_134424.4	UGT1A6	NM_001072.4
EIF3A	NM_003750.4	KLF4	NM_004235.6	RAD54L	NM_003579.4	UGT1A9	NM_021027.3

[0110] 实施例2POLE突变判别模型测试

[0111] 利用实施例1的POLE突变判别模型进行测试。对深圳市海普洛斯生物科技有限公司的子宫内膜癌患者数据库中共计410个可能存在POLE超突变致病性的样本进行测试，利用本发明的模型进行假阳性过滤及POLE突变分数计算。

[0112] 根据本发明的模型过滤条件，410个可能存在POLE超突变样本中过滤掉160个阴性样本。160例样本中，158例按照计分标准(6分制)均给出3分及以下的得分；剩余的2例按照计分标准(6分制)给出4分及以上的得分。综合来说，该过滤去除了158例真阴性和2例假阳性。

[0113] 经过滤后的250个样本，根据计算的分数判断的致病性，并与真实临床诊断结果进行对比，并绘制四格表，并根据四格表数据，计算诊断实验评估指标中的灵敏度和特异性，计算公式：

[0114] 灵敏度TP/(FN+TP)×100%；

[0115] 特异性计算公式:TN/(TN+FP)×100%。

[0116] 得到如下表2所示的四格表，

[0117] 表2

[0118]

	标准集-真	标准集-假
测试集-真	24(TP true positive)	1(FN false negative)
测试集-假	1(FP false positive)	224(TN true negative)

[0119] (注:标准集-真即阳性，标准集-假即阴性，测试集-真即评分大于等于4，测试集-假即评分小于4)

[0120] 根据表2的数据计算可得本发明模型的灵敏度为96%，特异性为99.5%。即本发明的模型灵敏度高，特异性强。

[0121] 表2中，一例标准集为假但其POLE突变得分为4的位点(A1814T)，在常见的检测范围(exons 9,11,13,14)之外，可能具有尚未报道的潜在致病性。

[0122] 对比例1

[0123] 利用对比模型对二代测序数据进行分析，计算POLE突变分数。

[0124] 对比模型与本发明的区别仅在于，缺少假阳性过滤。

[0125] 对比模型的构建方法，包括如下步骤：

[0126] S1、二代测序数据的预处理

[0127] 对获得的待测样本和对照样本的二代测序数据进行预过滤，分别获得pileup文件；

[0128] 所述待测样本为肿瘤患者待测组织DNA(ffpedna)或血液游离DNA(cfdna)，所述对照样本为肿瘤患者唾液DNA(sladna)或血液基因组DNA(gdna)；

[0129] 所述预过滤包括质控、比对、去重；

[0130] 所述质控包括质量评估、去接头序列、去低质量序列、去重复序列；

[0131] 所述比对是将质控后的read与人参考基因对比，获得BAM文件；将通过质控后每一条read与参考基因组进行比对，回贴到基因组上最佳位置；Reads比对前需去掉接头序列；Reads通过双端的模式比对到GRCh37(hg19)版本的人类基因组上，将比对上的reads按染色体位置排序并转换为二进制BAM文件；

[0132] 所述去重是去除PCR重复,校正测序错误,生成用于检测的pileup文件。

[0133] S2、序列比对、变异判读及突变过滤

[0134] 利用突变检测工具Mutscan或Varscan对步骤S1获得pileup文件进行分析,分析数据中存在的与癌症相关的680个基因的突变信息,注释SNV、Indel突变的位点位置与具体碱基变化,标注POLE突变,获取碱基C突变为碱基A(C>A)突变数、碱基T突变为碱基G(T>G)突变数、碱基C突变为碱基G(C>G)突变数、点突变总数、插入/缺失突变总数和总突变数;

[0135] 根据步骤S2注释的POLE突变,获取COSMIC数据库中POLE突变在样本对应肿瘤中的出现次数;

[0136] 根据680个基因的测序区域BED文件大小(即680基因覆盖区域的BED格式的文件),获取肿瘤负荷(TMB)数据;肿瘤负荷(TMB)=总突变数/测序区域BED文件大小,单位:muts/mb;

[0137] S3、POLE突变判别

[0138] 设置6个整数型计分标准:步骤S2中,碱基C突变为碱基A(C>A)突变数占点突变总数的百分比大于20%,记1分,反之不计分;碱基T突变为碱基G(T>G)突变数占点突变总数的百分比大于4%,记为1分,反之不计分;碱基C突变为碱基G(C>G)突变数占点突变总数的百分比小于0.6%,记为1分,反之不计分;插入/缺失位点(indel)占总突变数的百分比小于5%记为1分,反之不计分;肿瘤负荷(TMB)大于100muts/mb,记为1分,反之不计分;POLE突变在肿瘤中的出现次数大于1,记为1分,反之不计分;

[0139] 根据计分标准统计样本POLE突变分数,总分数大于等于4,判断为致病性POLE突变;总分数等于3,判断为未知致病性POLE突变;总分数小于3,判断为无致病性POLE突变。

[0140] 利用对比模型1,未经过过滤,可能存在仅含2个或少量突变为:C>A,T>G的样本:C>A,1个突变;T>G,1个突变;C>G,0个突变;indels,0个突变;其获得四分并判别为阳性。然而,考虑POLE超扩增高肿瘤负荷(突变数多)的特性,该例样本显然为阴性。因此在实际判别时,应当加入过滤环节用于判别此类假阳性。

[0141] 利用对比模型1进行实验测试,未经过滤的结果中存在两例假阳性。因此,相较于实施例1模型,对比模型1假阳性率增加。

[0142] 对比例2

[0143] 利用对比模型2对二代测序数据进行分析,计算POLE突变分数。

[0144] 对比模型2与本发明的区别仅在于,调整过滤假阳性的条件,根据突变数,过滤假阳性,过滤掉突变数小于等于25的测序数据对应的样本。

[0145] 根据理论模型,如果以12种snv类型的突变率的背景频率都是1:1做假设,在0.05显著条件下检出至少一个特定点需要17个样本,而检出C>A百分比的H1假设与H0假设的差异性至少需要30个样本。

[0146] 利用对比模型2,过滤条件降低阈值为25,实验结果中存在一例阴性样本(编号:73871),其突变总数为36且总分为4,为假阳性。因此,相较于实施例1模型,对比模型2假阳性率增加。

[0147] 对比例3

[0148] 利用对比模型3对二代测序数据进行分析,计算POLE突变分数。

[0149] 对比模型3与本发明的区别仅在于,调整过滤假阳性的条件,根据突变数,过滤假

阳性,过滤掉突变数小于等于65的测序数据对应的样本。

[0150] 根据理论模型,如果以12种snv类型的突变率的背景频率都是1:1做假设,在0.05显著条件下检出至少一个特定点需要17个样本,而检出C>A百分比的H1假设与H0假设的差异性至少需要30个样本。

[0151] 利用对比模型3,过滤条件阈值为65,实验结果中存在一例阴性样本(编号:107878),其突变总数为59且总分为5,为真阳性。因此,相较于实施例1模型,对比模型3假阴性率增加。

[0152] 实施例3一种计算机可读存储介质

[0153] 一种计算机可读存储介质,结构示意图如同2所示,其上存储有计算机程序/指令,该计算机程序/指令被处理器执行时实现本发明实施例1所述构建方法的步骤。

[0154] 需要说明的是,在本文中,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、物品或者系统不仅包括那些要素,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、物品或者系统所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括该要素的过程、方法、物品或者系统中还存在另外的相同要素。

[0155] 上述本发明实施例序号仅仅为了描述,不代表实施例的优劣。

[0156] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件,但很多情况下前者是更佳的实施方式。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在如上述的一个计算机可读存储介质(如ROM/RAM、磁碟、光盘)中,包括若干指令用以使得校正双端测序数据的装置执行本发明各个实施例所述的方法。

[0157] 以上仅为本发明的优选实施例,并非因此限制本发明的专利范围,凡是利用本发明说明书及附图内容所作的等效结构或等效流程变换,或直接或间接运用在其他相关的技术领域,均同理包括在本发明的专利保护范围内。

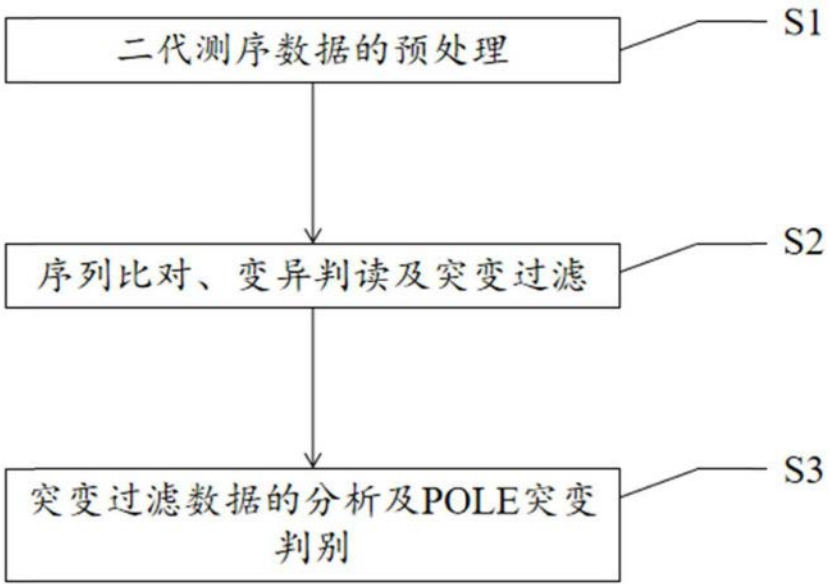

二代测序数据的预处理
S1
序列比对、变异判读及突变过滤
S2
突变过滤数据的分析及POLE突变判别
S3

图1

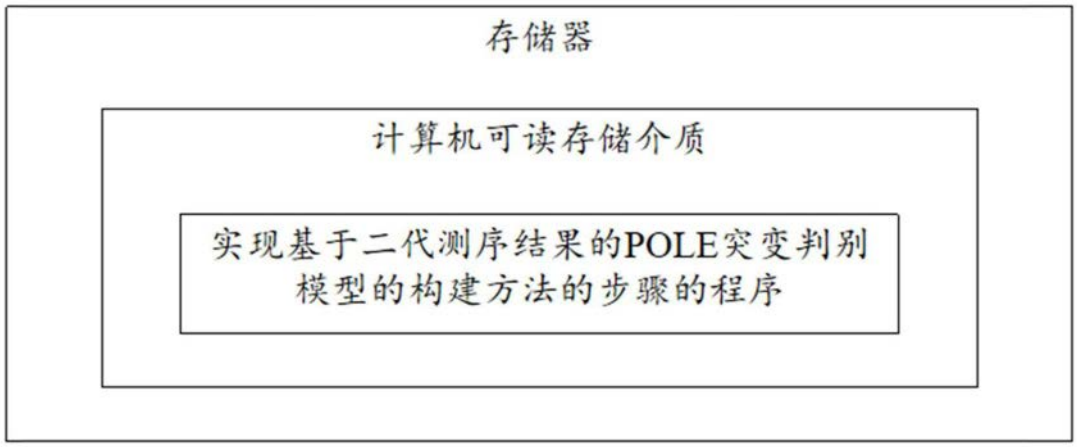

存储器
计算机可读存储介质
实现基于二代测序结果的POLE突变判别模型的构建方法的步骤的程序

图2