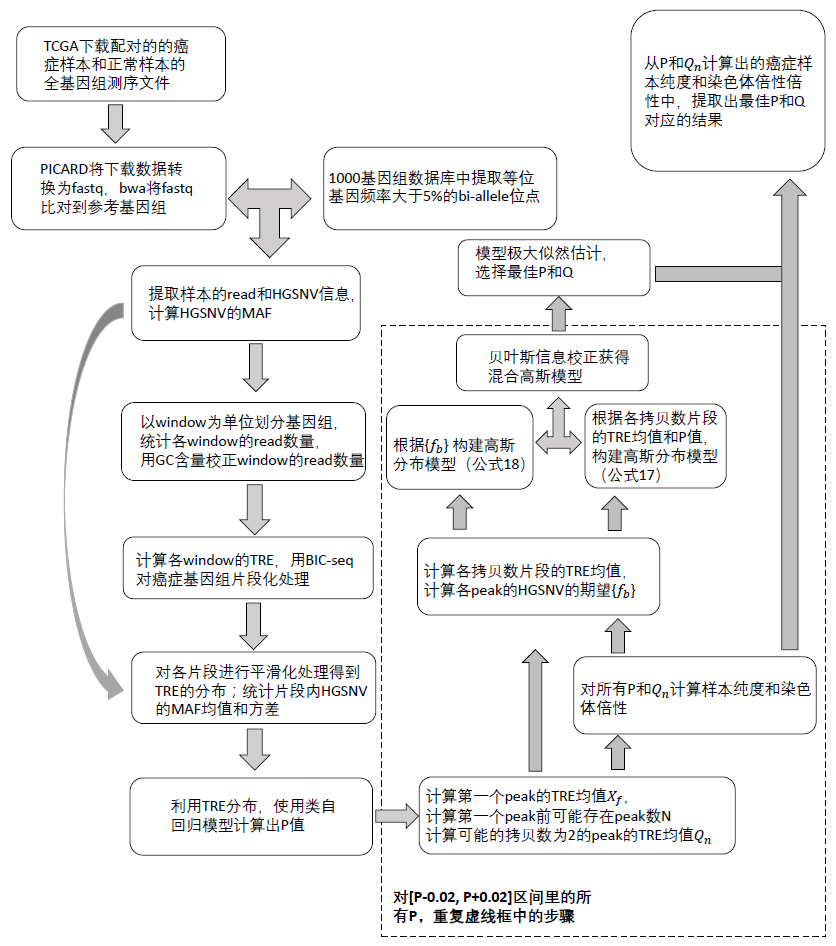
本发明提供了一种全自动、高效率、高准确性的计算癌症样本纯度和染色体倍性的方法和装置。通过本发明提供的层次混合高斯模型，实现了对癌症样本纯度和染色体倍性的快速和准确计算，节约了纯度估算的时间和经济成本，同时提高了计算结果的准确性。本发明在癌症样本纯度和染色体倍性计算上具有广阔的运用前景。



1、一种用于计算癌症样本中癌症细胞纯度和染色体倍性的方法，所述方法包括以下步骤：

步骤A：

获取配对的癌症组织样本和正常组织样本的全基因组测序数据，并将测序数据比对到参考基因组；

步骤B：

从步骤A得到的比对结果文件中，提取read位置和长度信息，HGSNV位点和覆盖该位点的read数量信息，计算所有HGSNV的MAF，其中，计算公式如（1.1）所示：

公式（1.1）中，为包含与参考基因组相同等位基因的read数量，为包含另一种等位基因的read的数量，表示覆盖该HGSNV位点的总read数量， 为该HGSNV的MAF值；

步骤C：

根据步骤B得到的read位置和长度信息，以window为单位统计各window内包含的read 数量，使用基因组GC含量校正所有window内read 数量；

步骤D：

使用步骤C校正后的read数量，使用公式（1）计算每一个window的TRE，然后运用TRE，通过BIC-seq软件对基因组进行片段化，获得以拷贝数划分的基因组片段：

（1）

公式（1）中，和分别表示在癌症样本中覆盖片段*s*的read数量和在正常样本中覆盖片段*s*的read数量，表示癌症样本总read数量，表示相应正常样本总read数量，为TRE值；

步骤E：

以步骤D中BIC-seq处理后的基因组片段为单位，统计片段内所有window的TRE的均值、方差和该片段内window数量，根据均值和方差对基因组每个片段的window数量进行平滑化处理，使TRE的分布更均匀，然后将平滑化处理后所有片段的window分布汇总，得到基因组上window随TRE变化的分布结果；同时以片段为单位，计算片段中所有HGSNV的MAF的均值和方差；

步骤F：

使用如公式（12）、（13）所示的类自回归模型，计算相邻拷贝数片段内TRE的差值即，其中，遍历一定范围的*P，*计算*Y()*， 在*Y()*的分布中，选择第二高峰内*Y()*的最大值对应的作为*P*的计算结果：

公式（12）和（13）中，*Xt*表示0到之间的TRE值；*t*表示扩大了1000倍的TRE值；表示TRE的最大值；变量表示两个TRE位点的间隔；表示在TRE为的位点，对应的window数量；*C(*)表示在TRE为的位点，对应的window数量；表示在变量下，类自回归模型的函数值；

步骤G：

根据步骤F得到的，计算TRE分布中第一个实际观测peak的TRE均值，然后计算在第一个实际peak之前最多可能存在理论peak的数量*N*，最后当第一个实际peak之前存在n个理论peak时，计算*Q*的值，以表示，其中步骤G包括：

G1：

根据步骤F计算的，使用公式（13.1），选取使公式（13.1）取最大值的作为第一个实际观测peak的TRE均值：

公式(13.1)中，*i*表示第*i*个peak，表示在TRE为的位点，对应的window数量，*n*表示以内peak的最大数量，表示TRE的最大值；

G2：

使用公式（13.2），根据步骤F计算的和步骤G1计算的，计算在 之前最多可能存在的peak数量：

公式（13.2）中，表示第一个peak的均值，表示相邻拷贝数片段对应的peak之间的间距，floor表示向下取整数；

G3：

利用步骤G2计算的*N*值，当*n*取0到*N*之间的整数时，使用公式（13.3）计算的值：

公式（13.3）中，*n*表示之前peak的数量，取值范围是0到*N*之间的整数，*P*表示相邻拷贝数片段对应的peak之间的间距，表示第一个实际观测peak的TRE均值，表示在之前理论上存在*n*个peak时的*Q*值；

步骤H：

使用步骤F计算的*P*与步骤G计算的，使用公式（10）、（11）计算癌症样本纯度和染色体倍性：

公式（10）、（11）中，表示样本纯度，表示染色体倍性，由此对（，）得到对应的（）；

步骤I：

当n取[0,*N*]之间的某个整数值时，使用公式（13.4）计算第*i*个peak的TRE均值：

公式（13.4）中，*n*表示之前peak的数量，取值范围是0到*N*之间的整数，*P*表示相邻拷贝数片段对应的peak之间的间距，表示第一个实际观测peak的TRE均值，表示第*i*个peak的TRE均值，

对于落在附近的片段，认为该片段具有拷贝数*i*；对于没有落在附近的片段，将其归类为亚克隆片段，在后续分析中剔除所有亚克隆片段；然后根据步骤H计算的癌症样本纯度和peak对应的拷贝数，计算peak的MAF的期望，不同peak的MAF期望不同，对基因组上的所有peak，最终得到MAF期望的集合{}；同时计算各个peak的TRE均值和方差或标准差；

步骤J：

根据步骤F计算的*P*和步骤I计算的构建如公式（19）所示的用“贝叶斯信息准则”校正后的混合高斯分布模型，然后对模型极大似然估计；其中，步骤J包括如下几步：

J1：

以步骤F计算的P构建如公式（17）所示的高斯分布模型：

公式（17）中，表示基因组片段TRE的似然函数，*N*表示基因组上的所有window的数量，*I*表示基因组中所有片段的最大的拷贝数，表示拷贝数为*i*的所有片段的TRE的标准差由步骤*I*得到，为第*s*个window的TRE观测值，表示第*i*个peak的TRE均值即步骤I中的，表示第*s*个window的拷贝数为*i*的权重，对所有的*i*，均取值为1；

J2：

以步骤I计算的构建如公式（18）所示的高斯分布模型：

(18)

公式（18）中，表示HGSNV的似然函数，*M*表示基因组中所有HGSNV数量，*S*表示第*S*个HGSNV，*I*表示基因组中所有片段的最大的拷贝数表示拷贝数为*i*，主要等位基因的拷贝数为*j*的片段内HGSNV的MAF期望值，由步骤I得到； 表示该片段内所有HGSNV的MAF的观测值均值，由步骤E得到；表示该片段内所有HGSNV的MAF观测值的标准差，由步骤E得到；表示在主要等位基因的拷贝数为*j*时，高斯分布的权重，对所有的*i*和*j*，取值均为1，表示第*S*个HGSNV所在片段的拷贝数为*i*的权重，对所有的*i*，取值均为1；

J3：

将（17）与（18）相加得到混合高斯模型，然后对混合模型进行BIC（Bayesian Information Criterion）校正得到最终混合模型如公式（19）：

公式（19）中，表示混合模型的似然函数，*I*表示基因组中所有片段的最大的拷贝数，*J*是公式（18）中*j*的取值个数，*N*是基因组中window的数量，*M*是基因组中HGSNV的个数，

对[*0, N*]范围内的每一个整数值*n*，通过步骤G得到，或者通过步骤I得到所有peak的MAF期望的集合{}，由一对（*P*，{}）构建一个公式（19）所示的模型；

步骤K：

以0.001为分辨率，对[*P-m, P +m*]区间的所有*P*值，重复步骤G~J，得到一系列不同的（*P，*）与对应的似然函数值，取最大的似然函数值对应的(*P，*作为最合适的*P*和*Q*值，*m*是0到0.5之间的一个值；

步骤L：

查询步骤H的结果，找到在步骤K得到的（*P，Q*）下，对应的癌症样本纯度和染色体倍性。

2、一种用于计算癌症样本中癌症细胞纯度和染色体倍性的装置，其包括处理器，所述处理器用于运行程序，所述程序运行时执行以下步骤：

步骤A：

获取配对的癌症组织样本和正常组织样本的全基因组测序数据，并将测序数据比对到参考基因组；

步骤B：

从步骤A得到的比对结果文件中，提取read位置和长度信息，HGSNV位点和覆盖该位点的read数量信息，计算所有HGSNV的MAF，其中，计算公式如（1.1）所示：

公式（1.1）中，为包含与参考基因组相同等位基因的read数量，为包含另一种等位基因的read的数量，表示覆盖该HGSNV位点的总read数量， 为该HGSNV的MAF值；

步骤C：

根据步骤B得到的read位置和长度信息，以window为单位统计各window内包含的read 数量，使用基因组GC含量校正所有window内read 数量；

步骤D：

使用步骤C校正后的read数量，使用公式（1）计算每一个window的TRE，然后运用TRE，通过BIC-seq软件对基因组进行片段化，获得以拷贝数划分的基因组片段：

（1）

公式（1）中，和分别表示在癌症样本中覆盖片段*s*的read数量和在正常样本中覆盖片段*s*的read数量，表示癌症样本总read数量，表示相应正常样本总read数量，为TRE值；

步骤E：

以步骤D中BIC-seq处理后的基因组片段为单位，统计片段内所有window的TRE的均值、方差和该片段内window数量，根据均值和方差对基因组每个片段的window数量进行平滑化处理，使TRE的分布更均匀，然后将平滑化处理后所有片段的window分布汇总，得到基因组上window随TRE变化的分布结果；同时以片段为单位，计算片段中所有HGSNV的MAF的均值和方差；

步骤F：

使用如公式（12）、（13）所示的类自回归模型，计算相邻拷贝数片段内TRE的差值即，其中，遍历一定范围的*P，*计算*Y()*， 在*Y()*的分布中，选择第二高峰内*Y()*的最大值对应的作为*P*的计算结果：

公式（12）和（13）中，*Xt*表示0到之间的TRE值；*t*表示扩大了1000倍的TRE值；表示TRE的最大值；变量表示两个TRE位点的间隔；表示在TRE为的位点，对应的window数量；*C(*)表示在TRE为的位点，对应的window数量；表示在变量下，类自回归模型的函数值；

步骤G：

根据步骤F得到的，计算TRE分布中第一个实际观测peak的TRE均值，然后计算在第一个实际peak之前最多可能存在理论peak的数量*N*，最后当第一个实际peak之前存在n个理论peak时，计算*Q*的值，以表示，其中步骤G包括：

G1：

根据步骤F计算的，使用公式（13.1），选取使公式（13.1）取最大值的作为第一个实际观测peak的TRE均值：

公式(13.1)中，*i*表示第*i*个peak，表示在TRE为的位点，对应的window数量，*n*表示以内peak的最大数量，表示TRE的最大值；

G2：

使用公式（13.2），根据步骤F计算的和步骤G1计算的，计算在 之前最多可能存在的peak数量：

公式（13.2）中，表示第一个peak的均值，*P*表示相邻拷贝数片段对应的peak之间的间距，floor表示向下取整数；

G3：

利用步骤G2计算的*N*值，当*n*取0到*N*之间的整数时，使用公式（13.3）计算的值：

公式（13.3）中，*n*表示之前peak的数量，取值范围是0到*N*之间的整数，*P*表示相邻拷贝数片段对应的peak之间的间距，表示第一个实际观测peak的TRE均值，表示在之前理论上存在*n*个peak时的*Q*值；

步骤H：

使用步骤F计算的*P*与步骤G计算的，使用公式（10）、（11）计算癌症样本纯度和染色体倍性：

公式（10）、（11）中，表示样本纯度，表示染色体倍性，由此对（*P，*）得到对应的（）；

步骤I：

当n取[0, *N*]之间的某个整数值时，使用公式（13.4）计算第*i*个peak的TRE均值：

公式（13.4）中，*n*表示之前peak的数量，取值范围是0到*N*之间的整数，*P*表示相邻拷贝数片段对应的peak之间的间距，表示第一个实际观测peak的TRE均值，表示第*i*个peak的TRE均值，

对于落在附近的片段，认为该片段具有拷贝数*i*；对于没有落在附近的片段，将其归类为亚克隆片段，在后续分析中剔除所有亚克隆片段；然后根据步骤H计算的癌症样本纯度和peak对应的拷贝数，计算peak的MAF的期望，不同peak的MAF期望不同，对基因组上的所有peak，最终得到MAF期望的集合{}；同时计算各个peak的TRE均值和方差或标准差；

步骤J：

根据步骤F计算的*P*和步骤I计算的构建如公式（19）所示的用“贝叶斯信息准则”校正后的混合高斯分布模型，然后对模型极大似然估计；其中，步骤J包括如下几步：

J1：

以步骤F计算的P构建如公式（17）所示的高斯分布模型：

公式（17）中，表示基因组片段TRE的似然函数，*N*表示基因组上的所有window的数量，*I*表示基因组中所有片段的最大的拷贝数，表示拷贝数为*i*的所有片段的TRE的标准差由步骤*I*得到，为第*s*个window的TRE观测值，表示第*i*个peak的TRE均值即步骤I中的，表示第*s*个window的拷贝数为*i*的权重，对所有的*i*，均取值为1；

J2：

以步骤I计算的构建如公式（18）所示的高斯分布模型：

(18)

公式（18）中，表示HGSNV的似然函数，*M*表示基因组中所有HGSNV数量，*S*表示第*S*个HGSNV，*I*表示基因组中所有片段的最大的拷贝数表示拷贝数为*i*，主要等位基因的拷贝数为*j*的片段内HGSNV的MAF期望值，由步骤I得到；表示该片段内所有HGSNV的MAF的观测值均值，由步骤E得到，表示该片段内所有HGSNV的MAF观测值的标准差，由步骤E得到；表示在主要等位基因的拷贝数为*j*时，高斯分布的权重，对所有的*i*和*j*，取值均为1，表示第*S*个HGSNV所在片段的拷贝数为*i*的权重，对所有的*i*，取值均为1；

J3：

将（17）与（18）相加得到混合高斯模型，然后对混合模型进行BIC校正得到最终混合模型如公式（19）：

公式（19）中，表示混合模型的似然函数，*I*表示基因组中所有片段的最大的拷贝数，*J*是公式（18）中*j*的取值个数，*N*是基因组中window的数量，*M*是基因组中HGSNV的个数，

对[*0, N*]范围内的每一个整数值*n*，通过步骤G得到，或者通过步骤I得到所有peak的MAF期望的集合{}，由一对（*P*，{}）构建一个公式（19）所示的模型；

步骤K：

以0.001为分辨率，对[*P-m, P +m*]区间的所有*P*值，重复步骤G~J，得到一系列不同的（*P，*）与对应的似然函数值，取最大的似然函数值对应的(*P，*作为最合适的*P*和*Q*值，*m*是0到0.5之间的一个值；

步骤L：

查询步骤H的结果，找到在步骤K得到的（*P，Q*）下，对应的癌症样本纯度和染色体倍性。

3、根据权利要求1所述的方法或者权利要求2所述的装置，其中，所述步骤A中，采用1000基因组计划第三期（phase 3）项目使用的参考基因组hs37d5 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2\_reference\_assembly\_sequence/hs37d5.fa.gz)作为所述参考基因组；和/或，比对软件使用Burrows-Wheeler Aligner (BWA)，比对方法使用其中的bwa mem，最终获得癌症和正常样本的比对结果bam格式文件。

4、根据权利要求1所述的方法或者权利要求2所述的装置，其中，所述步骤B中，采用samtools软件提取read的位置和长度信息，HGSNV的位点和覆盖该位点的read数量信息，其中，使用samtools view命令提取read信息时，使用参数 -q 31过滤掉序列比对质量（MAPQ）低于31的序列，其中q表示过滤掉测序质量差的序列，同时使用参数 -f 0x2 -F 0x18过滤掉未能正确匹配的read，其中f表示提取符合一定要求的序列，F表示过滤符合一定要求的序列，使用samtools mpileup命令提取HGSNV信息时，使用参数 -q 20过滤掉序列比对质量低于20的序列，并使用参数 -Q 20过滤掉碱基质量小于20的序列， 其中Q表示过滤掉碱基质量差的序列；选取等位基因频率时，使用samtools mpileup的-l参数；使用该参数需要提前准备一个包含SNP位点信息的bed格式文件。

5、根据权利要求1所述的方法或者权利要求2所述的装置，其中，

所述步骤C包括4步：

C1、将全基因组按照一定碱基长度的window为单位进行划分，对每个window统计覆盖该window的read数量，统计时以每条read的中点代表该read的位置；

C2、对参考基因组创建索引文件，提高GC含量的统计速度；

C3、以每个window的GC含量为自变量，以每个window的read数量为因变量，拟合read数量随GC含量变化的函数；

C4、使用拟合出的模型对全基因组read数量进行调整。

6、根据权利要求5所述的方法或者装置，其中，所述步骤C2中，为参考基因组创建GC含量索引文件，对每一条染色体分别统计1、5、25、125个碱基间隔的区域内，鸟嘌呤（G）和胞嘧啶（C）的累积数量，其中，在统计某一个window中的GC含量时，用a\* 125 + b\*25 + c\*5 +d\*1的快速算法提取，其中a,b,c,d表示系数变量。

7、根据权利要求5所述的方法或者装置，其中，所述步骤C3中，使用步骤C1和步骤C2提取的各window的 GC含量，通过如下弹性网络模型拟合read数量随GC含量变化，其中，使用window的GC含量为变量*x*，使用作为弹性网络模型的输入变量，以read数量为输出变量，构建弹性网络模型如公式（20）所示：

公式（20）中，*y*表示window内观测到的read数量，*X*表示输入变量矩阵，表示变量系数矩阵，*j*表示变量系数下标，*P*表示系数总数，和表示罚分系数。

8、根据权利要求5所述的方法或者装置，其中，所述步骤C4中，使用步骤C3中的模型预测每一个window理论上的read数量，基因组的平均GC含量定义为，window内观测到的read数量定义为，window内校正后的read数量为*Y*，那么校正公式如下（21）所示：

。

9、根据权利要求1所述的方法或者权利要求2所述的装置，其中，所述步骤E中，使用步骤D中BIC-seq处理后的基因组片段为单位，计算片段所包含的window数量，TRE的平均值以及方差，然后对片段的TRE进行平滑化处理，处理方式如公式（22）所示，针对每一个基因组片段，以TRE的均值作为正态分布的均值，以TRE的方差作为正态分布的方差，计算出TRE在[]范围内window数量的分布，定义*v*为TRE坐标，取值范围为[]，分辨率为0.001，为该片段分配到*v*位点的window数量，表示该片段内window的总数，将所有片段的window根据TRE值平滑化后，可使片段内的window数量呈现正态分布，对所有片段各TRE位点对应的window数求和汇总，得到基因组范围的window随TRE变化的分布：

。

10、根据权利要求1所述的方法或者权利要求2所述的装置，其中，所述步骤F中，以0.001为分辨率，遍历[0,1]范围内的所有*P*，使用类自回归模型，计算*Y(P)*的值，*Y(P)*表现为多峰分布，使用第二高峰内*Y()*的最大值对应的作为的计算结果，是TRE的最大取值，这里将设置为3。

11、根据权利要求1所述的方法或者权利要求2所述的装置，其中，所述步骤G中，步骤G包括3个步骤，步骤G1中，遍历[0,1]的TRE区间作为，过滤掉C()小于1000的TRE位点，计算使公式（13.1）取最大值时的作为第一个实际观测peak的均值。

12、根据权利要求1所述的方法或者权利要求2所述的装置，其中，所述步骤I中，然后根据步骤H计算的癌症样本纯度和peak对应的拷贝数，计算peak的MAF的期望，其中步骤I包括：



I1，使用公式（14）计算peak内HGSNV的MAF理论值：

公式（14）中，表示主要等位基因的拷贝数，表示peak的整体拷贝数，由步骤I得到，表示该peak内MAF的理论值，可见当较大时，有多种不同的可能值；

I2，利用负二项分布估计覆盖每个HGSNV位点的read总数的概率，使用公式（15）计算负二项分布的概率*p*和失败次数*r*：

公式（15）中，*m*是peak内所有window中read数量的均值，*v*是peak内所有window中read数量的方差，所求得的*p*是用于负二项分布的随机变量成功的概率，*r*为随机变量失败的次数，随机变量为覆盖某个HGSNV中的read数量；

I3，利用二项分布求得的覆盖某个HGSNV的read数的概率，结合在一定read数量下，HGSNV只有两种基因型，服从二项分布规律，利用公式（16）计算的校正值，同一个peak中，不同的计算得到不同的，选择与该peak的MAF观测均值最接近的作为该peak的：

公式（16）中，*k*表示在某个HGSNV位点，某一种等位基因A或B的数量，*d*为覆盖该HGSNV的read数量，*r*为随机变量失败的次数，*p*是用于负二项分布的随机变量成功的概率；

对每一个，可推断获得基因组所有peak对应的拷贝数和癌症样本纯度，从而对每一个peak求，进而得到所有peak的MAF的期望值得集合{}。

13、根据权利要求1所述的方法或者权利要求2所述的装置，其中，所述步骤K中，*m*取0.02作为*P*值得遍历区间为[*P-0.02, P+0.02*]。

**用于计算癌症样本纯度和染色体倍性的方法和装置**

**技术领域**

本发明属于癌症研究领域，具体涉及用于计算癌症样本中癌症细胞纯度和细胞内染色体倍性的方法和装置。

**背景技术**

癌症的研究是生命医学中的重要研究领域，并对人类健康生活有重大影响。癌症是一类细胞恶性增殖的疾病，因其病理十分复杂，人类尚无法攻克这类疾病。二代测序（next generation sequencing）为快速检测病人遗传信息提供了可能。然而测序需要从病人组织中提取样本，但通常癌症组织并不是只单纯地包含癌症细胞，它还有非常丰富的微环境。癌症细胞微环境指包围或伴随癌症细胞的正常细胞(non-cancerous cells)环境。癌细胞样本提取时，这些微环境会和癌细胞一起被提取，并会伴随癌细胞一起被测序[1]。癌症细胞在癌症样本中的比例被定义为癌症样本的纯度。癌症基因组通常包含着大量体细胞序列拷贝数变异，这些变异主要由基因组片段扩增或删除造成。识别特定肿瘤基因组的基因组片段拷贝数变化，是癌症基因组研究的一个重要课题。要准确鉴定基因组片段拷贝数具有一定挑战，因为癌症片段拷贝数主要由两个因素混合决定，一是癌症样本纯度，即癌症细胞在癌症样本中所占比例，二是染色体倍性[2,3]。传统中鉴定癌症样本纯度和染色体倍性的方法是使用实验技术，如定量图像分析[4]或单细胞测序[5]。但是在大型项目中，这样的方法会耗费大量人力、资金和时间。随着测序技术地发展，测序数据地快速增长，以及测序数据分析技术的积累，各种各样的癌症样本纯度算法被提出，并开发出了相对应的软件。

基于基因组片段拷贝数变异和基于等位基因频率（突变位点的B-等位基因（B-allele））的计算方法被相继提出。基于等位基因频率的方法有PurityEst[6]和PurBayes[7]，主要是依赖于随着肿瘤样本纯度和肿瘤基因组倍性的不同，等位基因的频率会有所不同。基于拷贝数变异的方法有CNAnorm[8]、THetA[9]、和ABSOLUTE[10]等。然而这两种方法都有不同程度的问题，使用等位基因频率的方法由于数据量的问题会有较大的误差，而运用拷贝数变异的方法虽然较稳定，但却无法区分样本纯度和染色体倍性的补偿效应，即存在识别问题。以上基于片段拷贝数的软件都没有解决这一问题，CNAnorm倾向于选择染色体倍性离二倍体最近的解决方案，ABSOLUTE结合了其他的经验数据，THetA则直接将所有可能结果都列出来了。

更优的方案应该是结合等位基因频率信息和片段拷贝数信息共同计算肿瘤样本纯度。PyLOH[11]，patchwork[12]使用了基因组上杂合SNV（单核苷酸变异）位点的频率信息，和基因组片段的拷贝数。PyLOH一定程度上解决了“识别困难”的问题，可以更合理给出唯一解决方案。但是其准确性较差，特别是遇到基因组中存在亚克隆（subclone）的情况下。Patchwork同时使用了两种信息，但是在计算基因型的中间步骤中，需要人工识别，人工判断的结果缺乏准确性，并且这种半自动化软件给应用带来很多不便。

如何充分利用现有的二代测序数据准确计算癌症样本纯度和癌症细胞基因组倍性问题仍然是一项具有挑战性的工作。

**参考文献**

1、Junttila M R, de Sauvage F J. Influence of tumour micro-environment heterogeneity on therapeutic response [J]. Nature, 2013, 501(7467):346-354.

2、Carter S L, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer [J]. Nature Biotechnology, 2012, 30(5):413-21.

3、Oesper L, Mahmoody A, Raphael B J. Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data [J]. Genome Biology, 2013, 14(7):R80.

4、Yuan Y, Failmezger H, Rueda O M, et al. Quantitative Image Analysis of Cellular Heterogeneity in Breast Tumors Complements Genomic Profiling [J]. Science Translational Medicine, 2012, 4(157):157ra143.

5、Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing [J]. Nature, 2011, 472(7341):90-4.

6、Su X, Zhang L, Zhang J, et al. PurityEst: estimating purity of human tumor samples using next-generation sequencing data.[J]. Bioinformatics, 2012, 28(17):2265-2266.

7、Larson N B. PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data [J]. Bioinformatics, 2013, 29(15):1888-9.

8、Gusnanto A, Wood H M, Pawitan Y, et al. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data [J]. Bioinformatics, 2012, 28(1):40-47.

9、Oesper L, Mahmoody A, Raphael B J. Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data [J]. Genome Biology, 2013, 14(7):R80.

10、Carter S L, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer [J]. Nature Biotechnology, 2012, 30(5):413-21.

11、Li Y, Xie X. Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity [J]. Bioinformatics, 2015, 30(4):2121.

12、Mayrhofer M, Dilorenzo S, Isaksson A. Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue [J]. Genome Biology, 2013, 14(3):R24.

**发明内容**

**术语定义**

为了更好地理解本发明，下面提供相关的解释和说明：

**Whole Genome Sequencing(WGS)：** 使用二代测序技术的全基因组测序。

**read：** 高通量测序平台产生的测序序列。

**测序深度：**测序得到的碱基（bp）总量与基因组（Genome）大小的比值，它是评价测序量的指标之一。

**window（窗口）**：按照一定长度划分的基因组片段，该长度代表window大小。本方法中window大小可由使用者自由设置，通常设置为几百碱基。一个大基因组片段*S*可以包含大量window。

**Tumor Read Enrichment（TRE）：** 癌症片段读长富集程度，指癌症样本中某一片段S内read数量与相应正常样本中对应片段read数量的比值，定义公式如下：

（1）

公式（1）中，和分别表示在癌症样本中覆盖片段*s*的read数量和相匹配的正常样本中覆盖片段*s*的read数量，表示癌症样本全基因组测序获得read总数量，表示相应正常样本全基因组测序获得read总数量。

**Heterozygous Germline Single Nucleotide Variants（HGSNV）：**杂合生殖系细胞单碱基变异，由于人类染色体属于二倍体，体细胞均由胚胎细胞发育而来，而生殖细胞中HGSNV位点只有两种碱基类型A和B，其中一种来源于父本，另一种来源于母本。

**Major Allele Fraction(MAF)：** 主要等位基因(allele)分数，本发明中使用的HGSNV只有两种等位基因，一种等位基因与参考基因组相同，另一种与参考基因组不同。这两种等位基因型分数的计算方法为覆盖某一等位基因的read数量除以覆盖该位点总read数量的比值，MAF就是两种等位基因分数中的较大值。计算公式如（1.1）所示，为包含与参考基因组相同等位基因的read数量，为包含另一种等位基因的read的数量，表示覆盖该HGSNV位点的总read数量， 为该HGSNV的MAF值。MAF是相对于HGSNV的概念，本发明中“片段的MAF”指片段内所有HGSNV的MAF均值，“peak的MAF”指peak中所有片段包含的HGSNV的MAF均值。

**major allele copy number：** 主要等位基因拷贝数，指在拷贝数为*i*的片段中，主要等位基因拷贝数的取值，它的取值范围为大于等于 的整数。

**peak：** 指基因组所有window的TRE分布中，聚集在一起的TRE簇。如附图1所示，图A表示基因组上所有window的TRE分布，纵轴表示对应某TRE位点的window总数量，该图为基因组GC含量校正之前的TRE分布，图B表示GC含量校正之后的TRE分布，图B中可以看到window明显以簇聚集，本方法将通过类自回归模型鉴定出来的TRE簇定义为peak，本质上是具有相同拷贝的基因组片段内window的聚集。

**癌症样本**： 指从某患癌症的个体身上取下的癌症组织，它包含了一部分癌症细胞和一部分正常细胞。

***P***： 指两个相邻peak之间的间距，由于peak是一个簇，这里peak的TRE由peak的TRE均值来表示，所以实际上是两个相邻peak的TRE均值的差。由于peak是具有相同拷贝数基因组片段内window的聚集，所以这里也表述为相邻拷贝数片段TRE的差值。

**发明目的**

本发明的目的在于克服现有技术的缺陷，提供一种全自动、高效率、高准确性的计算癌症样本纯度和染色体倍性的方法和装置。本发明在癌症样本纯度和染色体倍性计算上具有广阔的运用前景。

**技术方案**

为实现上述发明目的，本发明采取的技术方案为：通过癌症样本和匹配的正常样本的全基因组测序数据，对不同拷贝数片段的TRE和HGSNV的MAF分布构建混合高斯模型，计算癌症样本纯度和染色体倍性。

本发明主要运用了全基因组测序数据的TRE信息和HGSNV的MAF信息。TRE基本反映了癌症样本拷贝数变异情况，HGSNV的MAF信息基本反映了癌症样本的基因型。

TRE的差别主要来源于基因组片段的拷贝数差异，高拷贝数基因组片段内测序获得的read数量一定大于低拷贝数基因组片段测序获得的read数量，通过片段内read数量差异计算片段拷贝数差异是基因组拷贝数检测中的常用方法。但是大多数研究中，直接使用癌症样本片段内read数量除以正常样本该片段内read数量的比值（ratio）来进行read数量差异评估。本发明使用公式（1）所示的TRE来评估片不同片段的read数量差异。传统方法计算所得ratio不仅受癌症样本纯度与染色体倍性的影响，还受到癌症样本和正常样本测序深度的影响，而TRE不会受到样本测序深度的影响。

单独依赖read数量差异，无法确定各拷贝数片段的基因型，更重要的是无法区分样本纯度与样本倍性的补偿效应。而结合拷贝数差异片段内的HGSNV可以提供基因型信息，并帮助解决纯度与倍性的补偿效应，然而在前人的研究中，并没有高效利用HGSNV的方法，大部分方法采用枚举的方式将不同拷贝数片段可能对应的基因型一一列出，然后对排列组合的结果进行计算，挑选最可信的结果。这些方法共同的特点是，方法计算时间长，准确性差，对拷贝数很高或基因组变化较大的样本效果很差。本发明依据HGSNV的MAF与TRE的混合高斯模型计算癌症样本纯度和染色体倍性，能显著减少计算时间，并提高计算结果准确率。

假设某癌症样本的纯度为，那么癌症样本中的正常细胞比例为。癌症样本中正常细胞的染色体倍性为2，癌症细胞的染色体倍性为。那么癌症样本的染色体倍性如下公式（2）所示。

假设在癌症细胞中某一片段*S*的拷贝数为。那么癌症样本的片段*S*的拷贝数应该为如下公式（3）所示。

对于基因组片段*S*，TRE的计算方式如公式（1）所示。而TRE的期望值（expectation）的推导公式如下公式（4）所示，式中的、、和与公式（1）中含义相同。

为了更进一步引出，本方法定义了一些帮助理解的参数。片段S的长度，人类参考基因组的长度，癌症样本的测序深度， 正常样本的测序深度。那么片段S在癌症样本中测序深度为，片段S在正常样本中测序深度为。是指与片段S特性（如GC含量等引起测序深度偏好的特性）有关的参数，所以在癌症和正常样本中是一样的。进一步通过来表示，如公式（5）所示。

公式（5）中表示在癌症细胞中片段S的拷贝数，那么当片段S的拷贝数为*i*时和*i*+1时对应的TRE均值和分别如公式（6）和公式（7）所示：

通过公式（6）和公式（7），对于相邻的拷贝数对应的片段，它们的TRE的差值*P*如公式（8）所示，可见*P*值的大小与片段具体拷贝数没有关系，它只决定于癌症样本纯度和染色体倍性。通过附图2可以直观的看到在TRE分布图中，peak之间的距离是恒定的。

此外，对于*i* = 2即拷贝数为2的片段，它们的TRE值*Q*如公式9所示。附图2中*Q*对应的peak的TRE值略大于1。

通过上述公式（8）和（9），可以解得癌症样本的纯度（）和染色体倍性() 分别为：

通过以上分析可以得知，通过确定P和Q可以计算出癌症样本纯度和染色体倍性。

如附图2所示，计算出全基因组所有片段的TRE分布后，可以计算peak间的间距得到*P*。在前人的研究方法中，patchwork[12]使用相邻拷贝数片段的对应的read数量的比值的间距来辅助计算癌症样本纯度，但是该研究无法自动识别read数量比值之间的间距，需要人工识别图像确定read数量比值间距来进行下一步计算，效率和准确性都比较低。本发明开创性的使用类自回归模型鉴定TRE之间的间距，如公式（12）、（13）所示。公式（12）和（13）中，*Xt*表示0到之间的TRE值；*t*表示扩大了1000倍的TRE值；表示TRE的最大值；表示两个TRE位点的间隔；表示在TRE为的位点，对应的window数量；C()表示在TRE为的位点，对应的window数量；表示在下，类自回归模型的函数值；显而易见当=0时，的取值最大，但这时候的并不是实际peak之间的间距。

以0.001为分辨率，遍历0到1之间的所有值，然后求。的值分布如附图3所示。根据公式（13）的特点，我们可以知道，当等于0时，的值会是最大的，但此时的并不是peak之间的间距。我们选择图中第二高峰中的最大值对应的x轴坐标值作为peak之间的间距的计算结果。

如图1中B图所示的peak之所以簇状分布，是因为具有相同拷贝数的基因组片段的TRE值（指片段内所有window的TRE的均值）并不完全相等，同拷贝数片段TRE相互之间存在误差。该误差服从高斯分布，所以图B中的簇状分布被认为是高斯分布。

如附图2所示，*P*确定以后，peak会被识别出来，但是有部分基因组片段没有落在识别出的peak上，这些片段被称作亚克隆片段（subclone segmentation）。在考虑亚克隆片段的情况下，会对后面公式（17）和（18）所示的高斯模型取值有影响，进而会影响最终混合高斯模型的取值。由于在后续的分析中，本发明只需考虑落在peak位点的片段，由此排除了亚克隆片段的干扰。

在如图2所示的TRE分布图中， *Q*位点表示拷贝数为2的片段对应的TRE值。首先我们可以推测，如果癌症细胞基因组中，存在部分片段的拷贝数为1，部分片段的拷贝数为0，那么在*Q*位点之前应该存在两个peak分别对应拷贝数1和0。如果不存在拷贝数为1的片段，只存在拷贝数为0的片段，那么在*Q*位点之前距离*2P*的位点存在一个peak，而在*Q*位点之前距离*P*的位点peak的window数为0，这也就是图2所示的情形。另一种情况假若拷贝数为1和0的片段都没有，那么在*Q*位点之前距离*P*和距离*2P*的位点对应的window数都为0。那么对于TRE的分布图，对于，即第一个出现的peak，它可能对应的拷贝数为2（拷贝数1和0的peak对应的window就是0），也可能对应的拷贝数是1（拷贝数为0的片段对应peak的window为0），也有可能对应的拷贝数是0。

通过上述分析，我们可以知道，图2中的第一个peak即对应的片段的拷贝数有几种不同的可能，而每一种可能都会使*Q*对应不同的peak。本发明通过混合高斯模型计算出了最可能的对应的片段的拷贝数，从而确定了*Q*的取值，最终得到了癌症样本纯度和染色体倍性。首先我们需要鉴定出对应片段的拷贝数的所有可能值。本发明通过如下公式（13.1）来确定(13.1)中，表示在TRE为的位点，对应的window数量，*n*表示以内peak的最大数量。当为第一个peak的TRE均值。

然后使用公式（13.2）求之前最多可能有几个peak。其中表示第一个peak的TRE均值，*P*表示相邻拷贝数片段对应的peak之间的间距，floor表示向下取整数，当*N*=0时，表示之前没有peak，对应的片段拷贝数为0；当*N* = 1时，表示之前最多可能有1个peak，也可能没有peak；当*N* = 2时，表示之前最多可能有2个peak，也可能只有一个peak或者没有peak；

对于之前可能有（1，2，3…*N*）个peak的情形，每一种情形下都可以通过如下公式（13.3）计算出一个对应的*Q*值。根据*Q*的定义，我们知道*Q*为拷贝数为2的片段的peak对应的TRE值。首先可以推断拷贝数为0的片段对应的TRE值为，其中*n*表示之前peak的个数，取值范围是0到*N*之间的整数，*P*表示相邻拷贝数片段对应的peak之间的间距，含义与公式（13.1）中相同，公式（13.3）如下所示，其中表示在之前存在*n*个peak时，*Q*的取值。

根据以上分析，可以得到对于之前可能有（0，1，2，3…*N*）个peak的情形时，取值可能为（…）。而前面的类自回归模型已经计算出了*P*，那么对于每一个可能的*Q*，我们可以通过公式（10）和（11）计算出相应的和。本发明通过混合高斯模型计算出之前最可能的peak数量n，从而确定了对应的片段的拷贝数，进而确定了*Q*的取值，最终得到了癌症样本纯度和染色体倍性。具体方法如下说明。

对于的每一个可能的取值，结合*P*，我们可以计算得到相应的，然后计算各个拷贝数片段内HGSNV的MAF的理论值。其理论计算方式如公式（14）所示。表示主要等位基因的拷贝数（major allele copy number），表示peak的整体拷贝数。表示该peak内MAF的理论值。

但是在实际情况下当测序深度比较低时，与真实值（期望值）会有较大的误差。这里需要进一步校正，校正方法如公式（15）、（16）所示。式（15）中*m*是某peak内所有window中read数量的均值，*v*是peak内所有window中read数量的方差，所求得的*p*是用于负二项分布的随机变量成功的概率，*r*为随机变量失败的次数，这里的随机变量*d*为测序获取到的测序深度（read coverage）。

在某一个测序深度即*d*下，MAF实际上是服从以为概率、*d*为实验次数的二项分布。本发明以如下公式（16）对进行校正，得到MAF的期望值。公式中*k*表示在某个HGSNV位点，某一种等位基因（A或B）的数量，测得的等位基因总量为*d*（与测序深度相等）。

公式（13.2）表明每个基因组片段的拷贝数有*N*种可能。公式（14）表明，某一拷贝数片段可以有多种主要等位基因拷贝数，于是对每个基因组peak，可以算出多个，也可以算出多个，取距离peak实际观测MAF均值最近的作为该peak的MAF期望。而全基因组有多个peak，每个peak的MAF期望不同，对应算出多个MAF期望值{。考虑到某一peak内HGSNV的MAF会存在一定误差但也近似服从高斯分布，peak内所有HGSNV的MAF期望值可以由实际数据直接计算获得。假设某peak的基因型一定，那么通过比较peak的MAF观测值与peak的{的值就可以判断该peak的拷贝数和基因型。也就可以计算出拷贝数为2的peak对应的TRE值即*Q*的位置。同时也为了进一步校正*P*，本发明提出了一种混合高斯模型对TRE和HGSNV的观测数据进行拟合。

通过前面的分析可以得知因为公式（12）中的存在，并不能十分准确的代表各个peak的TRE均值。TRE的高斯分布模型如公式17所示：

其中，表示基因组片段TRE的似然函数。*N*表示基因组上的所有window的数量。*I*表示基因组中所有片段的最大的拷贝数。 表示拷贝数为*i*的所有片段的TRE的标准差。为第*s*个window的TRE观测值，表示第*i*个peak的TRE均值。表示第*s*个window的拷贝数为*i*的权重，本公式中对所有的*i*，均取值为1。该公式表明似然函数的大小与的取值相关，当与越接近，似然函数的取值越大，同时也表示*P*值越接近真实值。使用的极大似然估计可以计算出较合理的*P*值。

然而在部分情况下，*P*在较小的区间内波动时（如[-0.005, 0.005]），对应的似然函数值可能相同。本发明通过结合HGSNV的高斯分布模型如公式（18）所示，来更近一步确定*P*值，同时确定*Q*值。

(18)

其中，表示HGSNV的似然函数。*M*表示基因组中所有HGSNV。*S*表示第*S*个HGSNV。*I*表示基因组中所有片段的最大的拷贝数。为拷贝数为*i*、主要等位基因的拷贝数为*j*的片段内HGSNV的MAF期望值，即公式（16）计算出的。表示该片段内MAF的观测值均值，表示该片段内HGSNV的MAF观测值的标准差。表示在主要等位基因的拷贝数为*j*时，高斯分布的权重，对所有的*i*和*j*，取值均为1。表示第*S*个HGSNV所在片段的拷贝数为*i*的权重，对所有的*i*，取值均为1。公式（18）表明，似然函数的大小与值相关，当与越接近，似然函数越大，表明越准确，同时表明公式（14）中的越准确，从而可以得到各片段对应的与。于是确定了*Q*的取值。为了得到最准确的*P*与*Q*，本方法将公式（17）和公式（18）相加，得到混合高斯模型。

但对该混合模型统计计算容易发生模型过拟合现象。本发明进一步使用了贝叶斯信息准则（Bayesian Information Criterion，BIC）方法，给混合高斯模型一个罚分函数，用于控制模型的过拟合，最终混合高斯模型如公式（19）所示：

其中，表示混合模型的似然函数，*I*是公式（17）中高斯分布的个数，*J*是公式（18）中高斯分布的个数。*N*是基因组中窗口的数量，*M*是基因组中HGSNV的个数。

通过遍历[*P-0.02, P+0.02*]的区间，对所有的（P，）求极大似然估计，可以得到最合适的*P*值与*Q*值然后根据公式（10）和（11）可计算癌症样本的纯度和染色体倍性。

因此，本发明一方面提供了一种用于计算癌症样本中癌症细胞纯度和染色体倍性的方法，所述方法包括以下步骤：

步骤A：

获取配对的（来自同一癌症病人的）癌症组织样本和正常组织样本的全基因组测序（WGS）数据，并将测序数据比对到参考基因组；

步骤B：

从步骤A得到的比对结果文件中，提取read位置和长度信息，HGSNV位点和覆盖该位点的read数量信息，计算所有HGSNV的MAF，其中，计算公式如（1.1）所示：

公式（1.1）中，为包含与参考基因组相同等位基因的read数量，为包含另一种等位基因的read的数量，表示覆盖该HGSNV位点的总read数量， 为该HGSNV的MAF值；

步骤C：

根据步骤B得到的read位置和长度信息，以window为单位统计各window内包含的read 数量，使用基因组GC含量校正所有window内read 数量；

步骤D：

使用步骤C校正后的read数量，使用公式（1）计算每一个window的TRE，然后运用TRE，通过BIC-seq软件对基因组进行片段化，获得以拷贝数划分的基因组片段：

（1）

公式（1）中，和分别表示在癌症样本中覆盖片段*s*（这里指window）的read数量和在正常样本中覆盖片段*s*的read数量，表示癌症样本总read数量，表示相应正常样本总read数量，为TRE值；

步骤E：

以步骤D中BIC-seq处理后的基因组片段为单位，统计片段内所有window的TRE的均值、方差和该片段内window数量，根据均值和方差对基因组每个片段的window数量进行平滑化（smooth）处理，使TRE的分布更均匀，然后将平滑化处理后所有片段的window分布汇总，得到基因组上window随TRE变化的分布结果；同时以片段为单位，计算片段中所有HGSNV的MAF的均值和方差；

步骤F：

使用如公式（12）、（13）所示的类自回归模型，计算相邻拷贝数片段内TRE的差值即，具体方法为遍历一定范围的*P，*计算*Y()*， 在*Y()*的分布中，选择第二高峰内*Y()*的最大值对应的作为*P*的计算结果：

公式（12）和（13）中，*Xt*表示0到之间的TRE值；*t*表示扩大了1000倍的TRE值；表示TRE的最大值；变量表示两个TRE位点的间隔；表示在TRE为的位点，对应的window数量；*C(*)表示在TRE为的位点，对应的window数量；表示在变量下，类自回归模型的函数值；

步骤G：

根据步骤F得到的，计算TRE分布中第一个实际观测peak的TRE均值，然后计算在第一个实际peak之前最多可能存在理论peak的数量*N*，最后当第一个实际peak之前存在n个理论peak时，计算*Q*的值，以表示，其中步骤G可以包括：

G1：

根据步骤F计算的，使用公式（13.1），选取使公式（13.1）取最大值的作为第一个实际观测peak的TRE均值：

公式(13.1)中，*i*表示第*i*个peak，表示在TRE为的位点，对应的window数量，*n*表示以内peak的最大数量，表示TRE的最大值；

G2：

使用公式（13.2），根据步骤F计算的和步骤G1计算的，计算在 之前最多可能存在的peak数量：

公式（13.2）中，表示第一个peak的均值，表示相邻拷贝数片段对应的peak之间的间距，floor表示向下取整数；

G3：

利用步骤G2计算的*N*值，当*n*取0到*N*之间的整数时，使用公式（13.3）计算的值：

公式（13.3）中，*n*表示之前peak的数量，取值范围是0到*N*之间的整数，*P*表示相邻拷贝数片段对应的peak之间的间距，表示第一个实际观测peak的TRE均值，表示在之前理论上存在*n*个peak时的*Q*值；

步骤H：

使用步骤F计算的*P*与步骤G计算的所有可能的，使用公式（10）、（11）计算癌症样本纯度和染色体倍性：

公式（10）、（11）中，表示样本纯度，表示染色体倍性，那么对所有的（*P，*）都可以得到对应的（）；

步骤I：

当*n*取[0,*N*]之间的某个整数值时，使用公式（13.4）计算第*i*个peak的TRE均值：

公式（13.4）中，*n*表示之前peak的数量，取值范围是0到*N*之间的整数，*P*表示相邻拷贝数片段对应的peak之间的间距，表示第一个实际观测peak的TRE均值，表示第*i*个peak的TRE均值，

对于落在附近的片段，认为该片段具有拷贝数*i*；对于没有落在附近的片段，将其归类为亚克隆片段，在后续分析中剔除所有亚克隆片段；然后根据步骤H计算的癌症样本纯度和peak对应的拷贝数，可计算peak的MAF的期望，不同peak的MAF期望不同，对基因组上的所有peak，最终得到MAF期望的集合{}；同时计算各个peak的TRE均值和方差(或标准差)；

步骤J：

根据步骤F计算的*P*和步骤I计算的构建如公式（19）所示的用“贝叶斯信息准则”校正后的混合高斯分布模型，然后对模型极大似然估计；其中，步骤J可以包括如下几步：

J1：

以步骤F计算的*P*构建如公式（17）所示的高斯分布模型：

公式（17）中，表示基因组片段TRE的似然函数，*N*表示基因组上的所有window的数量，*I*表示基因组中所有片段的最大的拷贝数，表示拷贝数为*i*的所有片段的TRE的标准差由步骤*I*得到，为第*s*个window的TRE观测值，表示第*i*个peak的TRE均值即步骤I中的，表示第*s*个window的拷贝数为*i*的权重，对所有的*i*，均取值为1；

J2：

以步骤I计算的构建如公式（18）所示的高斯分布模型：

(18)

公式（18）中，表示HGSNV的似然函数，*M*表示基因组中所有HGSNV数量，*S*表示第*S*个HGSNV，*I*表示基因组中所有片段的最大的拷贝数表示拷贝数为*i*，主要等位基因的拷贝数为*j*的片段内HGSNV的MAF期望值，由步骤I得到； 表示该片段内所有HGSNV的MAF的观测值均值，由步骤E得到，表示该片段内所有HGSNV的MAF观测值的标准差，由步骤E得到；表示在主要等位基因的拷贝数为*j*时，高斯分布的权重，对所有的*i*和*j*，取值均为1，表示第*S*个HGSNV所在片段的拷贝数为*i*的权重，对所有的*i*，取值均为1；

J3：

将（17）与（18）相加得到混合高斯模型，然后对混合模型进行BIC（Bayesian Information Criterion）校正得到最终混合模型如公式（19）：

公式（19）中，表示混合模型的似然函数，*I*表示基因组中所有片段的最大的拷贝数，*J*是公式（18）中*j*的取值个数，*N*是基因组中window的数量，*M*是基因组中HGSNV的个数，

对[*0, N*]范围内的每一个整数值*n*，可以通过步骤G得到，也可以通过步骤I得到所有peak的MAF期望的集合{}，而一对（*P*，{}）可以构建一个公式（19）所示的模型，实质上是对每一对（*P*），可以构建一个公式（19）所示的模型；

步骤K：

以0.001为分辨率，对[*P-m, P +m*]区间的所有*P*值，重复步骤G~J，可以得到一系列不同的（*P，*）与对应的似然函数值，取最大的似然函数值对应的(*P，*作为最合适的*P*和*Q*值，*m*是0到0.5之间的一个值；

步骤L：

查询步骤H的结果，可以找到在步骤K得到的（*P，Q*）下，对应的癌症样本纯度和染色体倍性。

另一方面，本发明提供了一种用于计算癌症样本中癌症细胞纯度和染色体倍性的装置，其包括处理器，所述处理器用于运行程序，所述程序运行时执行以下步骤：

步骤A：

获取配对的（来自同一癌症病人的）癌症组织样本和正常组织样本的全基因组测序（WGS）数据，并将测序数据比对到参考基因组；

步骤B：

从步骤A得到的比对结果文件中，提取read位置和长度信息，HGSNV位点和覆盖该位点的read数量信息，计算所有HGSNV的MAF，其中，计算公式如（1.1）所示：

公式（1.1）中，为包含与参考基因组相同等位基因的read数量，为包含另一种等位基因的read的数量，表示覆盖该HGSNV位点的总read数量， 为该HGSNV的MAF值；

步骤C：

根据步骤B得到的read位置和长度信息，以window为单位统计各window内包含的read 数量，使用基因组GC含量校正所有window内read 数量；

步骤D：

使用步骤C校正后的read数量，使用公式（1）计算每一个window的TRE，然后运用TRE，通过BIC-seq软件对基因组进行片段化，获得以拷贝数划分的基因组片段：

（1）

公式（1）中，和分别表示在癌症样本中覆盖片段*s*（这里指window）的read数量和在正常样本中覆盖片段*s*的read数量，表示癌症样本总read数量，表示相应正常样本总read数量，为TRE值；

步骤E：

以步骤D中BIC-seq处理后的基因组片段为单位，统计片段内所有window的TRE的均值、方差和该片段内window数量，根据均值和方差对基因组每个片段的window数量进行平滑化（smooth）处理，使TRE的分布更均匀，然后将平滑化处理后所有片段的window分布汇总，得到基因组上window随TRE变化的分布结果；同时以片段为单位，计算片段中所有HGSNV的MAF的均值和方差；

步骤F：

使用如公式（12）、（13）所示的类自回归模型，计算相邻拷贝数片段内TRE的差值即，具体方法为遍历一定范围的*P，*计算*Y()*， 在*Y()*的分布中，选择第二高峰内*Y()*的最大值对应的作为*P*的计算结果：

公式（12）和（13）中，*Xt*表示0到之间的TRE值；*t*表示扩大了1000倍的TRE值；表示TRE的最大值；变量表示两个TRE位点的间隔；表示在TRE为的位点，对应的window数量；*C(*)表示在TRE为的位点，对应的window数量；表示在变量下，类自回归模型的函数值；

步骤G：

根据步骤F得到的，计算TRE分布中第一个实际观测peak的TRE均值，然后计算在第一个实际peak之前最多可能存在理论peak的数量*N*，最后当第一个实际peak之前存在n个理论peak时，计算*Q*的值，以表示，其中步骤G可以包括：

G1：

根据步骤F计算的，使用公式（13.1），选取使公式（13.1）取最大值的作为第一个实际观测peak的TRE均值：

公式(13.1)中，*i*表示第*i*个peak，表示在TRE为的位点，对应的window数量，*n*表示以内peak的最大数量，表示TRE的最大值；

G2：

使用公式（13.2），根据步骤F计算的和步骤G1计算的，计算在 之前最多可能存在的peak数量：

公式（13.2）中，表示第一个peak的均值，*P*表示相邻拷贝数片段对应的peak之间的间距，floor表示向下取整数；

G3：

利用步骤G2计算的*N*值，当*n*取0到*N*之间的整数时，使用公式（13.3）计算的值：

公式（13.3）中，*n*表示之前peak的数量，取值范围是0到*N*之间的整数，*P*表示相邻拷贝数片段对应的peak之间的间距，表示第一个实际观测peak的TRE均值，表示在之前理论上存在*n*个peak时的*Q*值；

步骤H：

使用步骤F计算的*P*与步骤G计算的所有可能的，使用公式（10）、（11）计算癌症样本纯度和染色体倍性：

公式（10）、（11）中，表示样本纯度，表示染色体倍性，那么对所有的（*P，*）都可以得到对应的（）；

步骤I：

当n取[0,*N*]之间的某个整数值时，使用公式（13.4）计算第*i*个peak的TRE均值：

公式（13.4）中，*n*表示之前peak的数量，取值范围是0到*N*之间的整数，*P*表示相邻拷贝数片段对应的peak之间的间距，表示第一个实际观测peak的TRE均值，表示第*i*个peak的TRE均值。

对于落在附近的片段，认为该片段具有拷贝数*i*；对于没有落在附近的片段，将其归类为亚克隆片段，在后续分析中剔除所有亚克隆片段；然后根据步骤H计算的癌症样本纯度和peak对应的拷贝数，可计算peak的MAF的期望，不同peak的MAF期望不同，对基因组上的所有peak，最终得到MAF期望的集合{}；。同时计算各个peak的TRE均值和方差(或标准差)；

步骤J：

根据步骤F计算的*P*和步骤I计算的构建如公式（19）所示的用“贝叶斯信息准则”校正后的混合高斯分布模型，然后对模型极大似然估计；其中，步骤J可以包括如下几步：

J1：

以步骤F计算的P构建如公式（17）所示的高斯分布模型：

公式（17）中，表示基因组片段TRE的似然函数，*N*表示基因组上的所有window的数量，*I*表示基因组中所有片段的最大的拷贝数，表示拷贝数为*i*的所有片段的TRE的标准差由步骤*I*得到，为第*s*个window的TRE观测值，表示第*i*个peak的TRE均值即步骤I中的，表示第*s*个window的拷贝数为*i*的权重，对所有的*i*，均取值为1；

J2：

以步骤I计算的构建如公式（18）所示的高斯分布模型：

(18)

公式（18）中，表示HGSNV的似然函数，*M*表示基因组中所有HGSNV数量，*S*表示第*S*个HGSNV，*I*表示基因组中所有片段的最大的拷贝数表示拷贝数为*i*，主要等位基因的拷贝数为*j*的片段内HGSNV的MAF期望值，由步骤I得到； 表示该片段内所有HGSNV的MAF的观测值均值，由步骤E得到，表示该片段内所有HGSNV的MAF观测值的标准差，由步骤E得到；表示在主要等位基因的拷贝数为*j*时，高斯分布的权重，对所有的*i*和*j*，取值均为1，表示第*S*个HGSNV所在片段的拷贝数为*i*的权重，对所有的*i*，取值均为1；

J3：

将（17）与（18）相加得到混合高斯模型，然后对混合模型进行BIC（Bayesian Information Criterion）校正得到最终混合模型如公式（19）：

公式（19）中，表示混合模型的似然函数，*I*表示基因组中所有片段的最大的拷贝数，*J*是公式（18）中*j*的取值个数，*N*是基因组中window的数量，*M*是基因组中HGSNV的个数。

对[*0, N*]范围内的每一个整数值*n*，可以通过步骤G得到，也可以通过步骤I得到所有peak的MAF期望的集合{}，而一对（*P*，{}）可以构建一个公式（19）所示的模型，实质上是对每一对（*P*），可以构建一个公式（19）所示的模型；

步骤K：

以0.001为分辨率，对[*P-m, P +m*]区间的所有*P*值，重复步骤G~J，可以得到一系列不同的（*P，*）与对应的似然函数值，取最大的似然函数值对应的(*P，*作为最合适的*P*和*Q*值，*m*是0到0.5之间的一个值；

步骤L：

查询步骤H的结果，可以找到在步骤K得到的（*P，Q*）下，对应的癌症样本纯度和染色体倍性。

作为一种优选方案，上述计算癌症样本纯度和染色体倍性的方法和装置中，所述步骤A中，采用1000基因组计划第三期（phase 3）项目使用的参考基因组hs37d5 (<ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz>)作为本发明的参考基因组,它包含了 [GRCh37](http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/index.shtml)中的所有染色体和零散序列（decoy sequences）。比对软件使用Burrows-Wheeler Aligner (BWA)， 比对方法使用其中的bwa mem，最终获得癌症和正常样本的比对结果bam格式文件。

作为一种优选方案，上述计算癌症样本纯度和染色体倍性的方法和装置中，所述步骤B中，采用了samtools软件提取read的位置和长度信息，HGSNV的位点和覆盖该位点的read数量信息。使用samtools view命令提取read信息时，过滤掉序列比对质量（MAPQ）低于31的序列（参数 -q 31, q表示过滤掉测序质量差的序列），同时过滤掉未能正确匹配的read（参数 -f 0x2 -F 0x18， f表示提取符合一定要求的序列，F表示过滤符合一定要求的序列）。使用samtools mpileup命令提取HGSNV信息时，过滤掉序列比对质量（MAPQ）低于20的序列（参数 -q 20），并过滤掉碱基质量小于20的序列（参数 -Q 20， Q表示过滤掉碱基质量差的序列）。选取等位基因频率（allele frequence）时，本发明使用samtools mpileup的-l参数。使用该参数需要提前准备一个包含SNP位点信息的bed格式文件。本发明方法提前收集了1000 基因组（genome）计划（http://www.internationalgenome.org/）中，根据大量样本统计出来的杂合等位基因位点，并且过滤掉B-等位基因频率（B-allele frequence）小于0.05的位点，然后做成bed文件。使用“ - l”参数在确保能提供充足的HGSNV位点基础上，大大加快了HGSNV位点的提取速度，提高了装置运行效率。

作为一种优选方案，上述计算癌症样本纯度和染色体倍性的方法和装置中，所述步骤C中，步骤C可以包括4步：

C1、将全基因组按照一定碱基长度的window为单位进行划分，对每个window统计覆盖该window的read数量，统计时以每条read的中点代表该read的位置；

C2、对参考基因组创建索引文件，提高GC含量的统计速度；

C3、以每个window的GC含量为自变量，以每个window的read数量为因变量，拟合read数量随GC含量变化的函数；

C4、使用拟合出的模型对全基因组read数量进行调整。

作为一种优选方案，上述计算癌症样本纯度和染色体倍性的方法和装置中，所述步骤C2中，本发明为参考基因组创建GC含量索引文件。对每一条染色体分别统计1、5、25、125个碱基间隔的区域内，鸟嘌呤（G）和胞嘧啶（C）的累积数量。那么在统计某一个window中的GC含量时，可以用a\* 125 + b\*25 + c\*5 +d\*1（其中a,b,c,d表示系数变量）的快速算法提取。例如想要统计某380bp区域内的GC含量时，可以分解为3\*125 + 1\*5形式，那么只需要读取特定索引文件的某个区域5碱基中的GC含量和某个区域125碱基中的GC含量即可。同时本发明将索引文件存储为二进制的格式，极大的加快了对特定区域的GC含量的提取。

作为一种优选方案，上述计算癌症样本纯度和染色体倍性的方法和装置中，所述步骤C3中，本发明使用步骤C1和步骤C2提取的各window GC含量，通过如下弹性网络模型拟合read数量随GC含量变化。本发明使用window的GC含量为变量*x*， 使用作为弹性网络模型的输入变量，以read数量为输出变量，构建弹性网络模型如公式（20）所示。式中，*y*表示window内观测到的read数量，*X*表示输入变量矩阵，表示变量系数矩阵，*j*表示变量系数下标，*P*表示系数总数，和表示罚分系数。

作为一种优选方案，上述计算癌症样本纯度和染色体倍性的方法和装置中，所述步骤C4中，使用步骤C3中的模型预测每一个window理论上的read数量，基因组的平均GC含量定义为，window内观测到的read数量定义为，window内校正后的read数量为*Y*。那么校正公式如下（21）所示：

作为一种优选方案，上述计算癌症样本纯度和染色体倍性的方法和装置中，所述步骤D中，本发明使用公式（1）计算每个window的TRE取值。然后运用TRE的值，使用BIC-seq软件对全基因组进行片段化（segmentation）。BIC-seq的思路是使用Bayesian Information Criterion（BIC）算法，统计相邻窗口的BIC值，值越小说明两个窗口越相似，然后将BIC值小于0的window合并，最终BIC-seq会按照片段拷贝数的差异，将全基因组分割为不同片段。每一个片段与相邻片段有不同的TRE均值，即拷贝数存在差异。

作为一种优选方案，上述计算癌症样本纯度和染色体倍性的方法和装置中，所述步骤E中，使用步骤D中BIC-seq处理后的基因组片段为单位，计算片段所包含的window数量，TRE的平均值以及方差。然后对片段的TRE进行smooth处理。处理方式如公式（22）所示。针对每一个基因组片段，以TRE的均值作为正态分布的均值，以TRE的方差作为正态分布的方差，计算出TRE在[]范围内window数量的分布，定义*v*为TRE坐标，取值范围为[]，分辨率为0.000，为该片段分配到*v*位点的window数量，表示该片段内window的总数。将所有片段的window根据TRE值smooth后，可使片段内的window数量呈现正态分布，对所有片段各TRE位点对应的window数求和汇总，可以得到基因组范围的window随TRE变化的分布。

作为一种优选方案，上述计算癌症样本纯度和染色体倍性的方法和装置中，所述步骤F中，以0.001为分辨率，遍历范围内的所有，使用类自回归模型，计算*Y()*的值。*Y()*表现为多峰分布，类似图3所示，图中横轴为，纵轴表示*Y()*，本发明使用第二高峰内*Y()*的最大值对应的作为的计算结果，是TRE的最大取值，这里将设置为3。

作为一种优选方案，上述计算癌症样本纯度和染色体倍性的方法和装置中，所述步骤G中，步骤G包括3个步骤，步骤G1中，[0,1]的TRE区间作为变量的取值范围，过滤掉C()小于1000的TRE位点，计算使公式（13.1）取最大值时的作为第一个peak的均值点。

作为一种优选方案，上述计算癌症样本纯度和染色体倍性的方法和装置中，所述步骤I中，根据步骤H计算的癌症样本纯度和peak对应的拷贝数，可计算peak的MAF的期望 。其中步骤I可以包含I1、I2、I3三个步骤。

I1，使用公式（14）计算peak内HGSNV的MAF理论值，公式（14）中，表示主要等位基因的拷贝数（major allele copy number），表示peak的整体拷贝数，由步骤I得到，表示该peak内MAF的理论值,可见当较大时，有多种不同的可能值。

I2，利用负二项分布估计覆盖每个HGSNV位点的read总数的概率，使用公式（15）计算负二项分布的概率*p*和失败次数*r*。公式（15）中，*m*是片段内所有window中read数量的均值，*v*是片段内所有window中read数量的方差，所求得的*p*是用于负二项分布的随机变量成功的概率，*r*为随机变量失败的次数，随机变量为覆盖某个HGSNV中的read数量。

I3，利用二项分布求得的覆盖某个HGSNV的read数的概率。结合在一定read数量下，HGSNV只有两种基因型，服从二项分布规律，利用公式（16）计算的校正值（即的期望）。同一个peak中，不同的可以计算得到不同的，选择与该peak的MAF观测均值最接近的作为该peak的。公式（16）中，*k*表示在某个HGSNV位点，某一种等位基因（A或B）的数量，*d*为覆盖该HGSNV的read数量，*r*为随机变量失败的次数，*p*是用于负二项分布的随机变量成功的概率；

对每一个，可推断获得基因组所有peak对应的拷贝数和癌症样本纯度，从而对每一个peak可求，进而可以得到所有peak的MAF的期望值得集合{}。

作为一种优选方案，上述计算癌症样本纯度和染色体倍性的方法和装置中，所述步骤K中，*m*取0.02，*P*值的遍历区间为[*P-0.02, P+0.02*]。

通过本发明提供的层次混合高斯模型，实现了对癌症样本纯度的快速和准确计算，节约了纯度估算的时间和经济成本，同时提高了计算结果的准确性。

**附图说明**

图1表示全基因组中window数量在TRE上的分布。其中，图A表示的是未进过GC校正的TRE分布，图B表示经过GC含量校正后的TRE分布图。

图2表示一种癌症细胞中的TRE分布的模型，图为smooth处理以后，图中的peak满足以*P*为周期的分布，少量不满足周期性分布的小peak被认为是亚克隆片段。*Q*表示拷贝数为2的peak，不存在拷贝数为1的片段，所以在大约0.6的位置的peak的window数量为0。

图3表示横轴为*P*纵轴为类自回归模型计算值得分布。

图4表示本发明方法和装置的流程图。

**具体实施方式**

为更好地说明本发明的目的、技术方案和优点，下面将结合附图和具体实施例对本发明作进一步说明。但是提供实施例仅用于说明目的，而本发明的范围不限于实施例。

使用本发明所述的装置计算癌症样本纯度和染色体倍性的流程图如图4所示。

实施例中，所使用的实验材料为TCGA（https://cancergenome.nih.gov/）数据库下载的样本（TCGA-AD-A5EJ）的正常组织TCGA-AD-A5EJ-10A和癌症组织TCGA-AD-A5EJ-01A的全基因组测序数据。计算平台为ubuntu 16.04，方法的具体实现为C++，python，R程序。

**实施例：根据样本**TCGA-CM-4746**的癌症组织和正常组织的全基因组测序数据，使用层次性混合高斯模型计算癌症样本的纯度和染色体倍性。**

1. 收集样本数据，在TCGA中下载TCGA-CM-4746-01A的肿瘤样本和正常样本的全基因组测序数据。癌症样本bam文件大小为12.6G，正常样本bam文件大小为10.1G。将bam文件用PICARD软件处理为fastq文件。将fastq使用bwa mem比对到参考基因组hs37d5得到新的癌症样本和正常样本bam文件，文件大小分别为12.4G和9.9G。
2. 下载1000 genome项目提供的1到22号染色体的vcf文件（ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/），使用GATK的SelectVariants方法，提取参考基因组hs37d5.fa中的等位基因频率大于5%的BIALLELIC位点作为潜在的HGSNV位点，最终得到5633774个biallele位点。
3. 提取正常样本和癌症样本的read信息，同时提取癌症样本的HGSNV信息。使用samtools提取癌症样本的序列覆盖度和HGSNV，得到HGSNV 67732个。提取HGSNV时，采用上述步骤一获得的biallele位点作为备选位点表单。使用samtools view方法，直接从备选表单中提取HGSNV，加快提取速度。
4. 以500bp为window对参考基因组建立GC含量的索引文件，对上述步骤一下载得到的参考基因组hs37d5.fa文件，建立1, 5, 25, 125区段中的GC含量索引文件。存储为二进制格式。
5. 以500bp为一个窗口，统计全基因组范围内每个窗口中的read数量。同时使用步骤四中产生的索引文件计算每个窗口中的GC含量。通过弹性网络模型对read数量进行GC含量校正。
6. 对每一个窗口，使用矫正后的read数量计算TRE。并依据TRE，通过BIC-seq对基因组进行片段化。片段化的结果如表一所示，每一列数据表示了一个基因组片段的位置信息和TRE的均值，方差和片段内window的数量。

**表1 BIC-seq对基因组片段化后的结果**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **染色体号** | **起始** | **终止** | **TRE 均值** | **TRE 方差** | **Windows数量** |
| chr1 | 13001 | 45265500 | 0.944508 | 0.0014742 | 86128 |
| chr1 | 45265501 | 85978000 | 0.945454 | 0.00133201 | 80970 |
| chr1 | 85978001 | 86011500 | 1.27362 | 0.0981321 | 68 |
| chr1 | 86011501 | 116069000 | 0.94915 | 0.00153058 | 58775 |
| chr1 | 116069001 | 120339500 | 1.01323 | 0.00437891 | 8488 |
| chr1 | 120339501 | 143744000 | 1.07492 | 0.016337 | 2442 |
| chr1 | 143744001 | 144707500 | 1.36461 | 0.0514154 | 469 |
| chr1 | 144707501 | 145290000 | 1.46903 | 0.0252744 | 887 |
| chr1 | 145290001 | 145833000 | 1.73666 | 0.0262944 | 936 |
| chr1 | 145982001 | 148248000 | 1.33895 | 0.0131223 | 2306 |
| chr1 | 148248001 | 149200000 | 1.68214 | 0.0381542 | 725 |
| chr1 | 149200001 | 249240500 | 1.3383 | 0.00114954 | 196613 |
| chr2 | 10001 | 41403000 | 0.948769 | 0.00134201 | 81865 |
| chr2 | 41403001 | 51317000 | 0.563251 | 0.00183306 | 19717 |
| chr2 | 51317001 | 91701500 | 0.950325 | 0.00145809 | 75550 |
| chr2 | 91701501 | 91824500 | 1.23428 | 0.0605257 | 188 |
| chr2 | 91824501 | 233387500 | 0.952652 | 0.000744377 | 271202 |
| chr2 | 233387501 | 243186500 | 0.936487 | 0.00308593 | 19032 |
| chr3 | 60001 | 197956500 | 0.952134 | 0.000613478 | 386567 |
| chr4 | 10001 | 6890000 | 0.959657 | 0.00417921 | 13520 |
| chr4 | 6890001 | 191044500 | 0.952703 | 0.000633657 | 358360 |
| chr5 | 11501 | 12616000 | 0.55999 | 0.00180406 | 24895 |
| chr5 | 12616001 | 180901000 | 0.948663 | 0.000676533 | 323787 |
| chr6 | 124001 | 24827000 | 0.941509 | 0.00171221 | 49041 |
| chr6 | 24827001 | 25982500 | 0.612532 | 0.00544556 | 2311 |
| chr6 | 25982501 | 171051000 | 0.953312 | 0.000725761 | 280819 |
| chr7 | 10001 | 5630500 | 0.954414 | 0.00461892 | 10976 |
| chr7 | 5630501 | 38306500 | 0.955796 | 0.00149665 | 64732 |
| chr7 | 38306501 | 38394000 | 1.24215 | 0.0368986 | 176 |
| chr7 | 38394001 | 55087500 | 0.950529 | 0.00212248 | 33130 |
| chr7 | 55087501 | 73283500 | 0.947545 | 0.00272214 | 27156 |
| chr7 | 73283501 | 142340000 | 0.955659 | 0.00107445 | 134626 |
| chr7 | 142340001 | 142491000 | 1.14391 | 0.0259809 | 303 |
| chr7 | 142491001 | 159128500 | 0.965812 | 0.00251871 | 31914 |
| chr8 | 11501 | 46857500 | 1.32635 | 0.00175707 | 84373 |
| chr8 | 46857501 | 47744500 | 1.01383 | 0.015617 | 1735 |
| chr8 | 47744501 | 146304000 | 1.33053 | 0.00112019 | 194805 |
| chr9 | 10501 | 89624000 | 0.956959 | 0.00120825 | 119638 |
| chr9 | 89624001 | 90043500 | 1.25975 | 0.0167381 | 836 |
| chr9 | 90043501 | 92343000 | 1.91959 | 0.010442 | 4544 |
| chr9 | 92343001 | 93352000 | 3.24803 | 0.0257886 | 1498 |
| chr9 | 93352001 | 93696000 | 3.54267 | 0.0399231 | 683 |
| chr9 | 93696001 | 94232000 | 2.92273 | 0.0283735 | 1068 |
| chr9 | 94232001 | 95076500 | 3.23526 | 0.0249389 | 1668 |
| chr9 | 95076501 | 95080000 | 1.47383 | 0.2343 | 8 |
| chr9 | 95080001 | 95099000 | 0.580555 | 0.0461102 | 39 |
| chr9 | 95099001 | 124413000 | 0.94853 | 0.00161715 | 58162 |
| chr9 | 124413001 | 124419500 | 1.57438 | 0.201191 | 14 |
| chr9 | 124419501 | 141128000 | 0.944109 | 0.00248041 | 32615 |
| chr10 | 66001 | 135525000 | 0.946217 | 0.000787218 | 255445 |
| chr11 | 113001 | 134946500 | 0.950948 | 0.000777276 | 259377 |
| chr12 | 60501 | 93500 | 1.88481 | 0.206704 | 52 |
| chr12 | 93501 | 100378000 | 0.949564 | 0.000874433 | 193005 |
| chr12 | 100378001 | 133841500 | 0.943986 | 0.00155671 | 65852 |
| chr13 | 19020501 | 115110000 | 0.984116 | 0.000893338 | 189923 |
| chr14 | 19000001 | 22533500 | 0.981344 | 0.00661204 | 5638 |
| chr14 | 22533501 | 23038000 | 1.12146 | 0.0151437 | 1009 |
| chr14 | 23038001 | 74253500 | 0.948794 | 0.00119784 | 101769 |
| chr14 | 74253501 | 74254500 | 3.63901 | 1.12718 | 3 |
| chr14 | 74254501 | 107289500 | 0.943291 | 0.00152386 | 65750 |
| chr15 | 20000001 | 29346000 | 0.940128 | 0.00397412 | 13933 |
| chr15 | 29346001 | 102521500 | 0.946193 | 0.00104706 | 141616 |
| chr16 | 60001 | 32646500 | 0.950261 | 0.00184498 | 60296 |
| chr16 | 32646501 | 33796500 | 0.889339 | 0.0178243 | 1172 |
| chr16 | 33796501 | 90282000 | 0.945044 | 0.00135985 | 89305 |
| chr17 | 1 | 6109000 | 0.91987 | 0.00396313 | 11836 |
| chr17 | 6109001 | 6125500 | 1.97935 | 0.163898 | 34 |
| chr17 | 6125501 | 16653500 | 0.922942 | 0.00286745 | 20945 |
| chr17 | 16653501 | 16738500 | 0.739932 | 0.0500593 | 142 |
| chr17 | 16738501 | 22262500 | 0.959384 | 0.00557573 | 9958 |
| chr17 | 22262501 | 27097500 | 1.71012 | 0.0106513 | 3647 |
| chr17 | 27097501 | 34432500 | 0.93298 | 0.00339514 | 14570 |
| chr17 | 34432501 | 34508500 | 1.28216 | 0.050695 | 121 |
| chr17 | 34509001 | 36215500 | 0.91458 | 0.00729353 | 2890 |
| chr17 | 36215501 | 36415000 | 1.0867 | 0.0380984 | 263 |
| chr17 | 36415001 | 39421500 | 0.949891 | 0.00582802 | 5984 |
| chr17 | 39421501 | 39432000 | 0.249188 | 0.0386297 | 20 |
| chr17 | 39432001 | 51996000 | 0.937382 | 0.00264461 | 24210 |
| chr17 | 51996001 | 52037000 | 1.3725 | 0.0543202 | 83 |
| chr17 | 52037001 | 55322000 | 0.938207 | 0.00465989 | 6539 |
| chr17 | 55322001 | 55632500 | 0.607987 | 0.0102 | 622 |
| chr17 | 55632501 | 68609500 | 0.942895 | 0.00246878 | 25575 |
| chr17 | 68609501 | 68745000 | 1.26374 | 0.0235519 | 272 |
| chr17 | 68745001 | 81195000 | 0.943562 | 0.00287259 | 24285 |
| chr18 | 10001 | 9988500 | 1.32762 | 0.00351405 | 19817 |
| chr18 | 9988501 | 78017500 | 0.944828 | 0.00106534 | 128567 |
| chr19 | 89001 | 23964500 | 0.940128 | 0.00211755 | 46618 |
| chr19 | 23964501 | 24028000 | 0.547803 | 0.0243662 | 128 |
| chr19 | 24028001 | 24621000 | 0.915363 | 0.0118983 | 1155 |
| chr19 | 24628501 | 59119000 | 1.27099 | 0.00216296 | 61940 |
| chr20 | 60001 | 29423000 | 0.944634 | 0.00166596 | 51929 |
| chr20 | 29423001 | 62965500 | 1.6413 | 0.0024125 | 66170 |
| chr21 | 9411001 | 40067000 | 0.938475 | 0.00164672 | 52525 |
| chr21 | 40067001 | 40442000 | 0.605613 | 0.00897434 | 747 |
| chr21 | 40442001 | 48120000 | 0.958554 | 0.00369137 | 14996 |
| chr22 | 16050001 | 51235000 | 0.944787 | 0.00174455 | 66591 |

1. 通过步骤六得到了每个片段的TRE的均值和方差，以及该片段内包含的窗口数量。使用正态分布的方法，以每个片段的TRE均值和方差作为正态分布的均值和方差，将片段中的窗口按照正态分布进行平滑化。汇总所有片段smooth后的TRE以及对应窗口数的信息。
2. 对smooth后的TRE的窗口数进行自回归分析，得到*P*的取值为0.386。
3. *P*等于0.386时，第一个实际观测peak的TRE均值为0.562，第一个实际观测peak前最多可以存在1个理论peak，即N=1。可能的*Q*为：，这两种Q的混合高斯模型的似然函数值分别为1.77E+07，1.78E+07。
4. 计算*P*在取值范围[*P-0.02, P+0.02*]内，BIC校正后的混合高斯模型的极大似然值，计算结果如表2所示。

表2 在*P*的取值范围内混合高斯模型的结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***P*值** | ***Q*值** | **似然函数值** | ***P*值** | ***Q*值** | **似然函数值** |
| 0.366 | 0.932 | 1.12E+06 | 0.386 | 1.334 | 1.77E+07 |
| 0.366 | 1.298 | 1.12E+06 | 0.387 | 0.948 | 1.78E+07 |
| 0.367 | 0.933 | 849906 | 0.387 | 1.335 | 1.77E+07 |
| 0.367 | 1.3 | 847125 | 0.388 | 0.948 | 1.84E+07 |
| 0.368 | 0.934 | 795735 | 0.388 | 1.336 | 1.84E+07 |
| 0.368 | 1.302 | 792858 | 0.389 | 0.948 | 1.88E+07 |
| 0.369 | 0.935 | 832922 | 0.389 | 1.337 | 1.87E+07 |
| 0.369 | 1.304 | 830037 | 0.39 | 0.948 | 1.87E+07 |
| 0.37 | 0.936 | 1.05E+06 | 0.39 | 1.338 | 1.87E+07 |
| 0.37 | 1.306 | 1.04E+06 | 0.391 | 0.948 | 1.88E+07 |
| 0.371 | 0.937 | 1.18E+06 | 0.391 | 1.339 | 1.87E+07 |
| 0.371 | 1.308 | 1.17E+06 | 0.392 | 0.949 | 1.84E+07 |
| 0.372 | 0.938 | 1.72E+06 | 0.392 | 1.341 | 1.83E+07 |
| 0.372 | 1.31 | 1.71E+06 | 0.393 | 0.95 | 1.84E+07 |
| 0.373 | 0.939 | 3.74E+06 | 0.393 | 1.343 | 1.83E+07 |
| 0.373 | 1.312 | 3.73E+06 | 0.394 | 0.951 | 1.80E+07 |
| 0.374 | 0.94 | 5.34E+06 | 0.394 | 1.345 | 1.79E+07 |
| 0.374 | 1.314 | 5.31E+06 | 0.395 | 0.952 | 1.75E+07 |
| 0.375 | 0.941 | 7.56E+06 | 0.395 | 1.347 | 1.74E+07 |
| 0.375 | 1.316 | 7.52E+06 | 0.396 | 0.953 | 1.63E+07 |
| 0.376 | 0.942 | 8.71E+06 | 0.396 | 1.349 | 1.62E+07 |
| 0.376 | 1.318 | 8.67E+06 | 0.397 | 0.954 | 1.53E+07 |
| 0.377 | 0.943 | 1.08E+07 | 0.397 | 1.351 | 1.52E+07 |
| 0.377 | 1.32 | 1.07E+07 | 0.398 | 0.955 | 1.34E+07 |
| 0.378 | 0.944 | 1.58E+07 | 0.398 | 1.353 | 1.33E+07 |
| 0.378 | 1.322 | 1.57E+07 | 0.399 | 0.956 | 1.17E+07 |
| 0.379 | 0.945 | 1.69E+07 | 0.399 | 1.355 | 1.17E+07 |
| 0.379 | 1.324 | 1.68E+07 | 0.4 | 0.957 | 8.63E+06 |
| 0.38 | 0.946 | 1.78E+07 | 0.4 | 1.357 | 8.57E+06 |
| 0.38 | 1.326 | 1.77E+07 | 0.401 | 0.958 | 6.79E+06 |
| 0.381 | 0.947 | 1.88E+07 | 0.401 | 1.359 | 6.74E+06 |
| 0.381 | 1.328 | 1.87E+07 | 0.402 | 0.959 | 1.77E+06 |
| 0.382 | 0.948 | 1.93E+07 | 0.402 | 1.361 | 1.76E+06 |
| 0.382 | 1.33 | 1.92E+07 | 0.403 | 0.96 | 831069 |
| 0.383 | 0.948 | 1.92E+07 | 0.403 | 1.363 | 826993 |
| 0.383 | 1.331 | 1.91E+07 | 0.404 | 0.961 | 412408 |
| 0.384 | 0.948 | 1.92E+07 | 0.404 | 1.365 | 411139 |
| 0.384 | 1.332 | 1.91E+07 | 0.405 | 0.962 | 351299 |
| 0.385 | 0.948 | 1.88E+07 | 0.405 | 1.367 | 350311 |
| 0.385 | 1.333 | 1.87E+07 | 0.406 | 0.963 | 352988 |
| 0.386 | 0.948 | 1.78E+07 | 0.406 | 1.369 | 352002 |

1. 步骤十中的结果显示，*P*为0.382时，混合模型取极大值，此时的*Q*为0.948，据此可计算获得癌症样本纯度为0.80，癌症细胞染色体倍性为2.14。

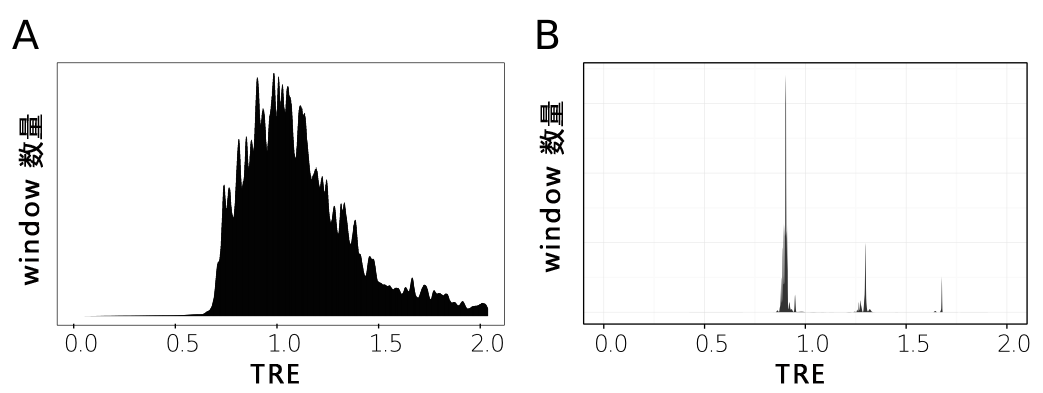


图1

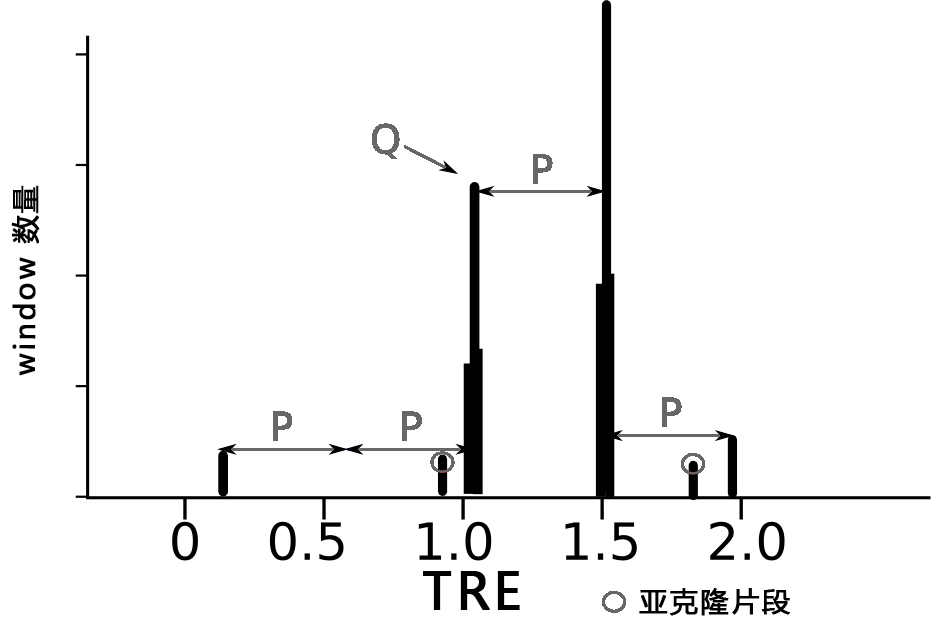


图2

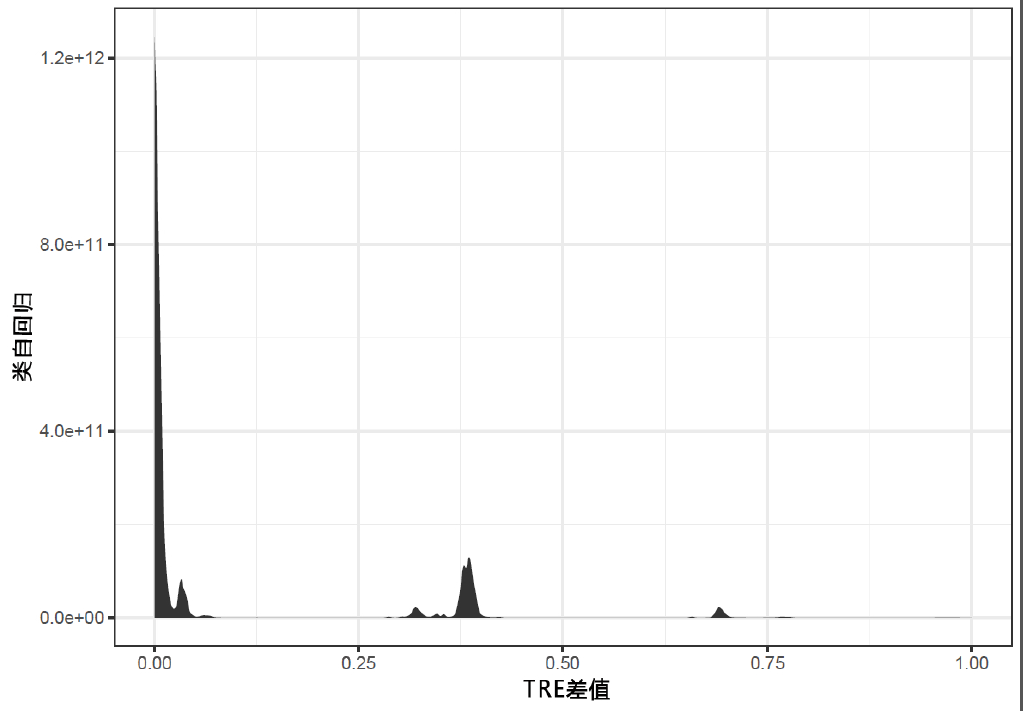


图3

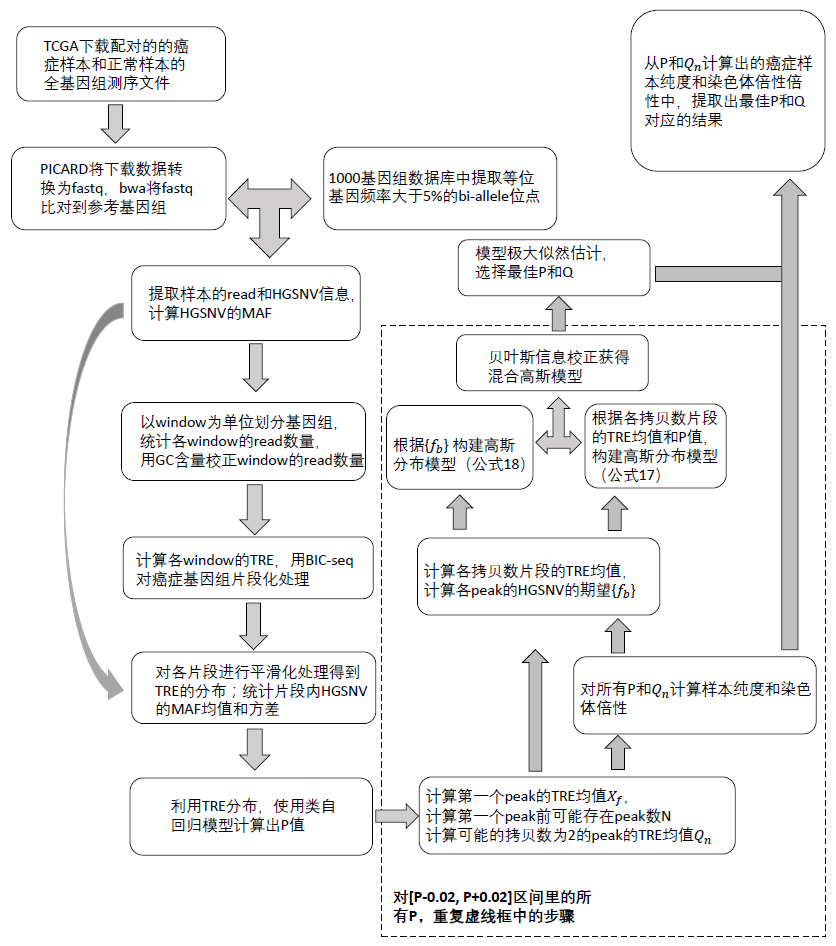


图4