

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2013 年 9 月 6 日 (06.09.2013)



(10) 国际公布号
WO 2013/127049 A1

- (51) 国际专利分类号:
C12Q 1/68 (2006.01)
- (21) 国际申请号: PCT/CN2012/071648
- (22) 国际申请日: 2012 年 2 月 27 日 (27.02.2012)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (71) 申请人 (对除美国外的所有指定国): **深圳华大基因科技有限公司 (BGI SHENZHEN CO., LIMITED)** [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山工业区综合楼 11F-3, Guangdong 518083 (CN)。 **深圳华大基因研究院 (BGI SHENZHEN)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。
- (72) 发明人: 及
- (75) 发明人/申请人 (仅对美国): **刘晓 (LIU, Xiao)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **张俊杰 (ZHANG, Junjie)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **徐怀前 (XU, Huaiqian)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **苏政 (SU,**

Zheng) [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **张瑞芳 (ZHANG, Ruifang)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **王俊 (WANG, Jun)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **汪建 (WANG, Jian)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **杨焕明 (YANG, Huanming)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

(74) 代理人: **深圳市威世博知识产权代理事务所 (普通合伙) (CHINA WISPRO INTELLECTUAL PROPERTY LLP.)**; 中国广东省深圳市南山区高新区粤兴三道 8 号中国地质大学产学研基地中地大楼 A806, Guangdong 518057 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ,

[见续页]

(54) Title: METHOD AND DEVICE FOR DETECTING MICRODELETION IN CHROMOSOME STS AREA

(54) 发明名称: 一种检测染色体 STS 区域微缺失的方法及其装置

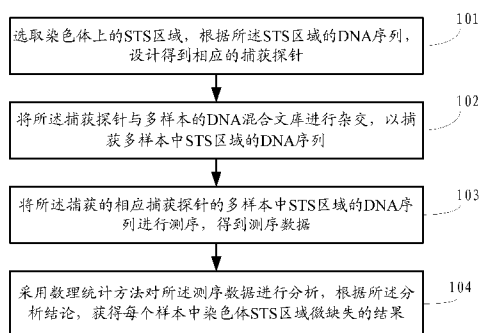


图 1 / Fig.1

101 Select an STS area on chromosome, and design according to a DNA sequence in the STS area to obtain a corresponding capture probe
102 Hybridize the capture probe and a multi-sample DNA hybrid library, so as to capture the DNA sequence in the STS area of the multi-sample
103 Sequence the captured DNA sequence in the STS area in the multi-sample of the corresponding capture probe, and obtain sequencing data
104 Analyze the sequencing data through a mathematical statistics method, and obtain a result regarding microdeletion in the chromosome STS area of each sample according to the analysis conclusion

(57) Abstract: A method and a device for detecting microdeletion in a chromosome sequence tagged site (STS) area. The method comprises: selecting an STS area on chromosome, and designing according to a DNA sequence in the STS area to obtain a corresponding capture probe; hybridizing the capture probe and a multi-sample DNA hybrid library, so as to capture the DNA sequence in the STS area in the multi-sample; sequencing the captured DNA sequence in the STS area in the multi-sample of the corresponding capture probe, and obtaining sequencing data; analyzing the sequencing data through a mathematical statistics method, and obtaining a result regarding microdeletion in the chromosome STS area of each sample according to the analysis conclusion.

(57) 摘要: 一种检测染色体序列标签位点 (STS) 区域微缺失的方法及其装置。该方法包括: 选取染色体上的 STS 区域, 根据 STS 区域的 DNA 序列, 设计得到相应的捕获探针; 将捕获探针与多样本的 DNA 混合文库进行杂交, 以捕获多样本中 STS 区域的 DNA 序列; 将捕获的相应捕获探针的多样本中 STS 区域的 DNA 序列进行测序, 得到测序数据; 采用数理统计方法对测序数据进行分析, 根据分析结论, 获得每个样本中染色体 STS 区域微缺失的结果。



WO 2013/127049 A1



NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

- (84) **指定国** (除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), 欧洲 (AL, AT,

本国际公布:

- 包括国际检索报告(条约第 21 条(3))。

一种检测染色体STS区域微缺失的方法及其装置

【技术领域】

本发明涉及基因工程技术领域，特别是涉及一种基于染色体序列标签位点 STS 区域微缺失检测的方法及其装置。

【背景技术】

缺失是染色体组中的染色体或 DNA 分子发生部分丢失的现象，它是导致基因突变的一个重要原因。

目前对染色体 STS 区域的微缺失检测主要是应用 PCR (Polymerase Chain Reaction) 技术。PCR 技术是通过模拟体内 DNA 复制的方式，在体外选择性地将 DNA 某个特殊区域扩增出来的技术。当对少量位点检测时，PCR 检测具有快速、方便的特点；另外，引物设计需要预先知道 DNA 某个特殊区域两端的序列，因此要求这些染色体 STS 区域的微缺失是预先已经报道的。

但是，在面对大量样本或未经报道的微缺失时，PCR 检测有很大的局限性；另外，当需要检测的 STS 位点较多、尤其是对整个染色体缺失进行检测时，传统的 PCR 已经不能满足这种需求，需要一种新的技术来进行研究。

【发明内容】

本发明主要解决的技术问题是提供一种基于染色体序列标签位点 STS 区域微缺失检测的方法及其装置，能够在有限的成本上，进行大量的染色体 STS 区域的微缺失的检测；也能够检测染色体 STS 区域未经报道的微缺失。

为解决上述技术问题，本发明采用的一个技术方案是：提供一种基于染色体序列标签位点 STS 区域微缺失检测的方法，包括：选取染色体上的 STS 区域，

根据所述 STS 区域的 DNA 序列,设计得到相应的捕获探针;将所述捕获探针与多样本的 DNA 混合文库进行杂交,以捕获多样本中 STS 区域的 DNA 序列;将所述捕获的相应捕获探针的多样本中 STS 区域的 DNA 序列进行测序,得到测序数据;采用数理统计方法对所述测序数据进行分析,根据所述分析结论,获得每个样本中染色体 STS 区域微缺失的结果。

其中,所述采用数理统计方法对所述测序数据进行分析,根据所述分析结论,获得每个样本中染色体 STS 区域微缺失的结果的步骤包括:将样本的 STS 区域的测序深度值进行均一化,得到均一化的深度值;根据得到的样本的 STS 区域的均一化的深度值,采用数理统计方法,检测所述样本 STS 区域的深度值异常值,并获得所述样本 STS 区域微缺失的结果。

其中,所述将样本的 STS 区域的测序深度值进行均一化的步骤包括:将所有样本中相同区域的深度值除以每个样本深度值的平均值,得到所述样本区域均一化的深度值。

其中,所述根据得到的样本的 STS 区域的均一化的深度值,采用数理统计方法,检测所述样本 STS 区域的深度值异常值,并获得所述样本 STS 区域微缺失的结果的步骤包括:根据得到的所有样本的同一区域的均一化的深度值,计算所述所有样本的同一区域的均一化深度值的平均值以及方差;根据所述所有样本的同一区域的均一化深度值的平均值以及方差,获得所述同一区域所有非离群样本的正态分布曲线;根据所述正态分布曲线,计算每个样本在每个区域在特定深度值时的概率值;根据每个样本在相应区域在特定深度值时的概率值,设置第一概率值阈值,若所述样本所在区域在特定深度值时的概率值小于所述第一概率值阈值,则获得所述样本区域有微缺失的结果 R1。

其中,所述根据每个样本在相应区域在特定深度值时的概率值,设置第一概率值阈值,若所述样本所在区域在特定深度值时的概率值小于所述第一概率值阈值,则获得所述样本区域有微缺失的结果 R1 的步骤之后还包括:对所述样本区域有微缺失的结果 R1 进行实验验证,根据实验验证结果,设置第二概率值

阈值，其中，所述第二概率值阈值小于第一概率值阈值；若所述样本区域在特定深度值时的概率值小于所述第二概率值阈值，则获得所述样本区域有微缺失的结果 R2。

其中，所述根据得到的样本的 STS 区域的均一化的深度值，采用数理统计方法，检测所述样本 STS 区域的深度值异常值，并获得所述样本 STS 区域微缺失的结果的步骤包括：根据得到的样本的 STS 区域的均一化的深度值，计算所述样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S ；根据得到的样本的 STS 区域的均一化的深度值，计算所述样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R ；将所述比值 D/S 通过 ID3 算法训练出第一比值阈值，将所述比值 D/R 通过 ID3 算法训练出第二比值阈值；若样本区域的比值 D/S 大于第一比值阈值，则获得所述样本区域没有微缺失的结果；若样本区域的比值 D/S 小于第一比值阈值，并且样本区域的比值 D/R 大于第二比值阈值，则获得所述样本区域没有微缺失的结果；若样本区域的比值 D/S 小于第一比值阈值，并且样本区域的比值 D/R 小于第二比值阈值，则获得所述样本区域有微缺失的结果。

其中，所述根据得到的样本的 STS 区域的均一化的深度值，采用数理统计方法，检测所述样本 STS 区域的深度值异常值，并获得所述样本 STS 区域微缺失的结果的步骤包括：根据得到的所有样本的同一区域的均一化的深度值，计算所述所有样本的同一区域的均一化深度值的平均值以及方差；根据所述所有样本的同一区域的均一化深度值的平均值以及方差，获得所述同一区域所有非离群样本的正态分布曲线；根据所述正态分布曲线，计算每个样本在每个区域在特定深度值时的概率值；根据每个样本在每个区域在特定深度值时的概率值，设置第三概率值阈值，若所述样本区域在特定深度值时的概率值小于所述第三概率值阈值，则获得所述样本区域有微缺失的结果 R3；计算所述结果 R3 中样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S ；计算所述结果 R3 中样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R ；将所

述比值 D/S 通过 ID3 算法训练出第三比值阈值, 将所述比值 D/R 通过 ID3 算法训练出第四比值阈值; 若样本区域的比值 D/S 大于第三比值阈值, 则获得所述样本区域没有微缺失的结果; 若样本区域的比值 D/S 小于第三比值阈值, 并且样本区域的比值 D/R 大于第四比值阈值, 则获得所述样本区域没有微缺失的结果; 若样本区域的比值 D/S 小于第三比值阈值, 并且样本区域的比值 D/R 小于第四比值阈值, 则获得所述样本区域有微缺失的结果。

其中, 所述选取染色体上的 STS 区域, 根据所述 STS 区域的 DNA 序列, 设计并合成相应的捕获探针的步骤包括: 在基因组数据库中查找染色体上的 STS 区域的 DNA 序列; 在所述查找到的 STS 区域的 DNA 序列中挑选符合捕获探针设计条件的序列; 根据所述挑选到的符合捕获探针设计条件的序列, 设计并合成得到捕获探针。

其中, 所述多样本的 DNA 混合文库的制备的步骤包括: 制备多个带有不同接头的质量控制合格的单样本的 DNA 文库; 将所述多个单样本的 DNA 文库按照预定比例混合; 检验所述混合的多样本的 DNA 文库的质量是否合格, 若是, 即为制备的多样本的 DNA 混合文库。

其中, 所述单样本的 DNA 文库的制备的步骤包括: 利用物理或化学的方法将基因组 DNA 打断成预定大小的 DNA 片段, 回收所述打断的 DNA 片段; 利用酶对所述回收的 DNA 片段进行末端修复, 形成补平的末端磷酸化的 DNA 片段, 回收所述补平的末端磷酸化的 DNA 片段; 利用酶对所述回收的补平的 DNA 片段的 3' 末端加上“A”碱基, 回收所述 3' 末端加上“A”碱基的 DNA 片段; 使所述回收的 3' 末端加上“A”碱基的 DNA 片段在酶的作用下与标签接头 Index Adapter 连接, 并回收带有标签接头的 DNA 片段; 以标签接头序列的引物做为引物, 对所述带有标签接头的 DNA 片段进行扩增, 回收所述扩增的产物; 检验所述扩增的产物的质量控制是否合格, 若是, 即为制备的单样本的 DNA 文库。

其中, 所述将捕获的相应捕获探针的多样本中 STS 区域的 DNA 序列进行测序, 得到测序数据的步骤之后还包括: 对所述对多样本中 STS 区域的 DNA 序列

的测序数据进行质量控制。

其中,所述对多样本中 STS 区域的 DNA 序列的测序数据进行质量控制的步骤包括:对所述多样本中 STS 区域的 DNA 序列的测序数据中不合格的数据进行过滤,得到合格的多样本的测序数据;通过短序列对比软件,将所述合格的多样本的测序数据与参考基因组序列进行对比,并统计每个样本的测序深度的相关参数以及不同样本之间相同的 STS 区域的测序深度的相关参数;根据所述统计得到的每个样本的测序深度的相关参数,过滤掉不合格的样本的测序数据,得到合格的样本的测序数据;根据所述统计得到的不同样本之间相同的 STS 区域的测序深度的相关参数,过滤掉不合格 STS 区域的测序数据,得到合格 STS 区域的测序数据。

其中,所述对多样本中 STS 区域的 DNA 序列的测序数据中不合格的数据进行过滤,得到合格的多样本的测序数据的步骤包括:通过测序数据中低质量值碱基的比例进行测序质量过滤,若低质量值碱基个数超过整条序列碱基个数的预定比例,则判断为是不合格的数据,将所述不合格的测序数据过滤掉,获得初步合格的第一测序数据集合;若所述初步合格的第一测序数据集合中测序结果不确定的碱基个数超过整条序列碱基个数的 10%,则判断为是不合格的数据,将所述不合格的测序数据过滤掉,获得初步合格的第二测序数据集合;将所述初步合格的第二测序数据集合中所有测序数据与测序接头序列库进行比对,若所述初步合格的第二测序数据集合中存在测序接头序列,则判断为是不合格的数据,将所述不合格的测序数据过滤掉,获得初步合格的第三测序数据集合;将所述初步合格的第三测序数据集合中所有测序数据与试验中引入的所有外源序列比对,若所述初步合格的第三测序数据集合中存在外源序列,则判断为是不合格的数据,将所述不合格的测序数据过滤掉,获得合格的多样本的测序数据的序列。

其中,所述根据统计得到的每个样本的测序深度的相关参数,过滤掉不合格的样本的测序数据,得到合格的样本的测序数据的步骤包括:将所有样本的

测序深度值按照从小到大的顺序进行排序，利用四分位函数确定所述排序后的所有样本的测序深度值的下四分位数 $Q1$ 、上四分位数 $Q3$ 以及四分位数间距 IQR ；将所有样本的测序深度值在 $Q1$ 减去 1.5 倍 IQR 和 $Q3$ 加上 1.5 倍 IQR 范围之外的不合格的样本的测序数据过滤掉，得到合格的样本的测序数据。

其中，所述根据统计得到的不同样本之间相同的 STS 区域的测序深度的相关参数，过滤掉不合格 STS 区域的测序数据，得到合格 STS 区域的测序数据的步骤包括：将不同样本之间相同的 STS 区域的测序深度值按照从小到大的顺序进行排序，利用四分位函数确定所述排序后的不同样本之间相同的 STS 区域的测序深度值的中位数、上四分位数 $Q3$ 以及四分位数间距 IQR ；将不同样本之间相同的 STS 区域的测序深度值中位数为 0 或者中位数大于 $Q3$ 加上 1.5 倍 IQR 的不合格的 STS 区域的测序数据过滤掉，得到合格的 STS 区域的测序数据。

为解决上述技术问题，本发明采用的另一个技术方案是：提供一种基于染色体序列标签位点 STS 区域微缺失检测的装置，包括：捕获探针获得模块，用于选取染色体上的 STS 区域，根据所述 STS 区域的 DNA 序列，设计得到相应的捕获探针；杂交模块，用于将所述捕获探针与多样本的 DNA 混合文库进行杂交，以捕获多样本中 STS 区域的 DNA 序列；测序数据获得模块，用于将所述捕获的相应捕获探针的多样本中 STS 区域的 DNA 序列进行测序，得到测序数据；微缺失结果获得模块，用于采用数理统计方法对所述测序数据进行分析，根据所述分析结论，获得每个样本中染色体 STS 区域微缺失的结果。

其中，所述微缺失结果获得模块包括：深度值均一化单元，用于将样本的 STS 区域的测序深度值进行均一化，得到均一化的深度值；微缺失结果获得单元，用于根据得到的样本的 STS 区域的均一化的深度值，采用数理统计方法，检测所述样本 STS 区域的深度值异常值，并获得所述样本 STS 区域微缺失的结果。

其中，所述深度值均一化单元具体用于将所有样本中相同区域的深度值除以每个样本深度值的平均值，得到所述样本区域均一化的深度值。

其中，所述微缺失结果获得单元包括：平均值方差获得单元，用于根据得到的所有样本的同一区域的均一化的深度值，计算所述所有样本的同一区域的均一化深度值的平均值以及方差；正态分布曲线获得单元，用于根据所述所有样本的同一区域的均一化深度值的平均值以及方差，获得所述同一区域所有非离群样本的正态分布曲线；概率值计算单元，用于根据所述正态分布曲线，计算每个样本在每个区域在特定深度值时的概率值；第一判断单元，用于根据每个样本在相应区域在特定深度值时的概率值，设置第一概率值阈值，若所述样本所在区域在特定深度值时的概率值小于所述概率值第一概率值阈值，则获得所述样本区域有微缺失的结果 R1。

其中，所述微缺失结果获得单元还包括：概率值阈值确定单元，用于对所述样本区域有微缺失的结果 R1 进行实验验证，根据实验验证结果，设置第二概率值阈值，其中，所述第二概率值阈值小于第一概率值阈值；第二判断单元，用于若所述样本区域在特定深度值时的概率值小于所述第二概率值阈值，则获得所述样本区域有微缺失的结果 R2。

其中，所述微缺失结果获得单元包括：比值 D/S 获得单元，用于根据得到的样本的 STS 区域的均一化的深度值，计算所述样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S；比值 D/R 获得单元，用于根据得到的样本的 STS 区域的均一化的深度值，计算所述样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R；第一、二比值阈值获得单元，用于将所述比值 D/S 通过 ID3 算法训练出第一比值阈值，将所述比值 D/R 通过 ID3 算法训练出第二比值阈值；第一判断单元，用于若样本区域的比值 D/S 大于第一比值阈值，则获得所述样本区域没有微缺失的结果；第二判断单元，用于若样本区域的比值 D/S 小于第一比值阈值，并且样本区域的比值 D/R 大于第二比值阈值，则获得所述样本区域没有微缺失的结果；第三判断单元，用于若样本区域的比值 D/S 小于第一比值阈值，并且样本区域的比值 D/R 小于第二比值阈值，则获得所述样本区域有微缺失的结果。

其中，所述微缺失结果获得单元包括：平均值方差获得单元，用于根据得到的所有样本的同一区域的均一化的深度值，计算所述所有样本的同一区域的均一化深度值的平均值以及方差；正态分布曲线获得单元，用于根据所述所有样本的同一区域的均一化深度值的平均值以及方差，获得所述同一区域所有非离群样本的正态分布曲线；概率值计算单元，用于根据所述正态分布曲线，计算每个样本在每个区域在特定深度值时的概率值；第一判断单元，用于根据每个样本在每个区域在特定深度值时的概率值，设置第三概率值阈值，若所述样本区域在特定深度值时的概率值小于所述第三概率值阈值，则获得所述样本区域有微缺失的结果 R3；比值 D/S 获得单元，用于计算所述结果 R3 中样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S；比值 D/R 获得单元，用于计算所述结果 R3 中样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R；第三、四比值阈值获得单元，用于将所述比值 D/S 通过 ID3 算法训练出第三比值阈值，将所述比值 D/R 通过 ID3 算法训练出第四比值阈值；第二判断单元，用于若样本区域的比值 D/S 大于第三比值阈值，则获得所述样本区域没有微缺失的结果；第三判断单元，用于若样本区域的比值 D/S 小于第三比值阈值，并且样本区域的比值 D/R 大于第四比值阈值，则获得所述样本区域没有微缺失的结果；第四判断单元，用于若样本区域的比值 D/S 小于第三比值阈值，并且样本区域的比值 D/R 小于第四比值阈值，则获得所述样本区域有微缺失的结果。

其中，所述捕获探针获得模块包括：区域查找单元，用于在基因组数据库中查找染色体上的 STS 区域的 DNA 序列；序列挑选单元，用于在所述查找到的 STS 区域的 DNA 序列中挑选符合捕获探针设计条件的序列；捕获探针获得单元，用于根据所述挑选到的符合捕获探针设计条件的序列，设计并合成得到捕获探针。

其中，所述装置还包括多样本 DNA 混合文库制备模块，所述多样本 DNA 混合文库制备模块包括：单样本 DNA 文库制备单元，用于制备多个带有不同接

头的质量控制合格的单样本的 DNA 文库；单样本文库混合单元，用于将所述多个单样本的 DNA 文库按照预定比例混合；多样本 DNA 混合文库获得单元，用于检验所述混合的多样本的 DNA 文库的质量是否合格，若是，即为制备的多样本的 DNA 混合文库。

其中，所述装置还包括测序数据质控模块，所述测序数据质控模块包括：合格序列获得单元，用于对所述多样本中 STS 区域的 DNA 序列的测序数据中不合格的数据进行过滤，得到合格的多样本的测序数据；测序深度统计单元，用于通过短序列对比软件，将所述合格的多样本的测序数据与参考基因组序列进行对比，并统计每个样本的测序深度的相关参数以及不同样本之间相同的 STS 区域的测序深度的相关参数；合格样本获得单元，用于根据所述统计得到的每个样本的测序深度的相关参数，过滤掉不合格的样本的测序数据，得到合格的样本的测序数据；合格区域获得单元，用于根据所述统计得到的不同样本之间相同的 STS 区域的测序深度的相关参数，过滤掉不合格 STS 区域的测序数据，得到合格 STS 区域的测序数据。

为解决上述技术问题，本发明采用的又一个技术方案是：提供一种计算机可读介质，所述介质承载一系列指令以控制计算机处理器执行如上所述的方法。

本发明的有益效果是：区别于现有技术的情况，本发明的基于染色体序列标签位点 STS 区域微缺失检测的方法及其装置，根据 STS 区域的 DNA 序列，设计获得捕获探针，探针涵盖了整个染色体上 STS 区域，与多样本的 DNA 混合文库进行杂交后，捕获到的多样本中 STS 区域的 DNA 序列，能够大量、高效、准确地检测出染色体上已经报道或者未经报道的 STS 相关区域的微缺失，另外，本发明的数理统计信息分析流程科学、稳定，灵敏度高、假阳性低，可以有效的针对微缺失进行分析。

【附图说明】

图 1 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法一实施

例的流程图；

图 2 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法另一实施例的流程图；

图 3 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例的流程图；

图 4 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例的流程图；

图 5 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例的流程图；

图 6 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例的流程图；

图 7 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例的流程图；

图 8 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例中过滤掉深度值离群的样本的示意图；

图 9 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例中深度值正常的样本的示意图；

图 10 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例的流程图；

图 11 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例中过滤掉深度值离群区域的示意图；

图 12 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例的流程图；

图 13 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例的流程图；

图 14 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实

施例的流程图；

图 15 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例的流程图；

图 16 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例中举例说明使用 ID3 算法所构造的决策树的示意图；

图 17 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例中决策树分析法的部分流程图；

图 18 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例中决策树分析法的流程图；

图 19 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例的流程图；

图 20 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例中样品测序深度中位数的柱状统计图以及箱线图；

图 21 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例中区域测序深度中位数的柱状统计图以及箱线图；

图 22 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例中获得区域深度为某个特定值的概率值的示意图；

图 23 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例中概率值阈值以及概率值阈值的期望值与观察值的关系示意图；

图 24 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法又一实施例中真、假阳性比例的示意图；

图 25 是本发明基于染色体序列标签位点 STS 区域微缺失检测的装置一实施例的结构示意图；

图 26 是本发明基于染色体序列标签位点 STS 区域微缺失检测的装置又一实施例的结构示意图；

图 27 是本发明基于染色体序列标签位点 STS 区域微缺失检测的装置又一实

施例的结构示意图；

图 28 是本发明基于染色体序列标签位点 STS 区域微缺失检测的装置又一实施例的结构示意图；

图 29 是本发明基于染色体序列标签位点 STS 区域微缺失检测的装置又一实施例的结构示意图。

【具体实施方式】

下面结合附图和实施例对本发明进行详细说明。

图 1 是本发明基于染色体序列标签位点 STS 区域微缺失检测的方法一实施例的流程图，如图 1 所示，所述方法包括：

步骤 101：选取染色体上的 STS 区域，根据所述 STS 区域的 DNA 序列，设计得到相应的捕获探针。

STS (sequence-tagged site) 序列标签位点，是基因组上定位明确、作为界标并能够通过 PCR 扩增、被唯一操作的短的、单拷贝的 DNA 序列，用于产生作图位点，即测定一系列 STS 的次序即可作出基因组区域的图谱。探针是一小段单链 DNA 或者 RNA 片段（大约是 20 到 500bp），用于检测与其互补的核酸序列。

步骤 102：将所述捕获探针与多样本的 DNA 混合文库进行杂交，以捕获多样本中 STS 区域的 DNA 序列。

DNA 文库是指某一特定来源 DNA 通过细胞-DNA 克隆技术构建呈含有所用 DNA 片段的重组 DNA 分子，并转化至细菌内，构成 DNA 文库。此处的 DNA 混合文库是指所有样本混合在一起的所用 DNA 片段。

步骤 103：将所述捕获的相应捕获探针的多样本中 STS 区域的 DNA 序列进行测序，得到测序数据。

步骤 104：采用数理统计方法对所述测序数据进行分析，根据所述分析结论，获得每个样本中染色体 STS 区域微缺失的结果。

数理统计是伴随着概率论的发展而发展起来的一个数学分支，研究如何有效的由集、整理和分析受随机因素影响的数据，并对所考虑的问题作出推断或预测，为采取某种决策和行动提供依据或建议。

在一实施例中，如图 2 所示，所述选取染色体上的 STS 区域，根据所述 STS 区域的 DNA 序列，设计并合成相应的捕获探针的步骤包括：

步骤 201：在基因组数据库中查找染色体上的 STS 区域的 DNA 序列。

找出 Y 染色体上 STS 序列在 UCSC 数据库中，在人类基因组参考序列 Hg19 (<http://genome.ucsc.edu/>) 的位置坐标，根据位置坐标即可找到 Y 染色体上 STS 区域的 DNA 序列。

步骤 202：在所述查找到的 STS 区域的 DNA 序列中挑选符合捕获探针设计条件的序列。

序列上的重复性及 GC 含量会影响到芯片的捕获效率，甚至发生捕获错误，所以挑选符合捕获探针设计条件的序列很重要。

步骤 203：根据所述挑选到的符合捕获探针设计条件的序列，设计并合成得到捕获探针。

为了捕获 DNA 片段，探针的长度一般为 60-150bp，GC 含量在 40%-70%之间。在对同一段上的 DNA 序列进行探针设计时，需要多条探针才能覆盖整段 DNA 序列，而且探针与探针之间存在重叠序列 (overlap)，其中，重叠序列的长度一般为 20bp。

在具体实施例中，在基因组数据库中查找染色体上的 STS 区域的 DNA 序列的位置坐标，将此坐标提交给提供 DNA 捕获服务的公司，由这些公司完成捕获探针的设计与合成。

其中，如图 3 所示，多样本的 DNA 混合文库的制备的步骤包括：

步骤 301：制备多个带有不同接头的质量控制合格的单样本的 DNA 文库。

每个样本带有一个不同的标签接头，为了在测序中区别来自不同样本的文库，每个样本文库的 DNA 末端都含有不同的 6bp 或 8bp 的 Index 碱基序列。

步骤 302: 将所述多个单样本的 DNA 文库按照预定比例混合。

每个样本文库 DNA 混合量可根据需要等量或按照一定比例混合。

步骤 303: 检验所述混合的多样本的 DNA 文库的质量是否合格, 若是, 即为制备的多样本的 DNA 混合文库。

对混合的多样本的 DNA 文库进行定量, 检测是否引入外源杂质等质量控制指标。

其中, 如图 4 所示, 单样本的 DNA 文库的制备的步骤包括:

步骤 401: 利用物理或化学的方法将基因组 DNA 打断成预定大小的 DNA 片段, 回收所述打断的 DNA 片段。

一般选择打断的 DNA 片段大小为 200~300bp, 探针长度一般为 80bp 左右, 片段长度在 200~300bp 会有较高的捕获效率; 另外捕获之后采用 PE 测序, 测通长度也在 200~300bp。

步骤 402: 利用酶对所述回收的 DNA 片段进行末端修复, 形成补平的末端磷酸化的 DNA 片段, 回收所述补平的末端磷酸化的 DNA 片段;

步骤 403: 利用酶对所述回收的补平的 DNA 片段的 3' 末端加上“A”碱基, 回收所述 3' 末端加上“A”碱基的 DNA 片段;

步骤 404: 使所述回收的 3' 末端加上“A”碱基的 DNA 片段在酶的作用下与标签接头 Index Adapter 连接, 并回收带有标签接头的 DNA 片段;

步骤 405: 以标签接头序列的引物做为引物, 对所述带有标签接头的 DNA 片段进行扩增, 回收所述扩增的产物;

步骤 406: 检验所述扩增的产物的质量控制是否合格, 若是, 即为制备的单样本的 DNA 文库。对单样本的 DNA 文库进行定量, 检测是否引入外源杂质等质量控制指标。

其中, 将捕获的相应捕获探针的多样本中 STS 区域的 DNA 序列进行测序, 得到测序数据的步骤之后还包括: 对所述对多样本中 STS 区域的 DNA 序列的测序数据进行质量控制。如图 5 所示, 对多样本中 STS 区域的 DNA 序列的测序数

据进行质量控制的步骤包括:

步骤 501: 对所述多样本中 STS 区域的 DNA 序列的测序数据中不合格的数据进行过滤, 得到合格的多样本的测序数据; 其中高通量测序技术可以为 Illumina Hiseq 2000 测序技术, 当然也可以采用现有的其他高通量测序技术。

步骤 502: 通过短序列对比软件, 将所述合格的多样本的测序数据与参考基因组序列进行对比, 并统计每个样本的测序深度的相关参数以及不同样本之间相同的 STS 区域的测序深度的相关参数;

步骤 503: 根据所述统计得到的每个样本的测序深度的相关参数, 过滤掉不合格的样本的测序数据, 得到合格的样本的测序数据;

步骤 504: 根据所述统计得到的不同样本之间相同的 STS 区域的测序深度的相关参数, 过滤掉不合格 STS 区域的测序数据, 得到合格 STS 区域的测序数据。

其中, 如图 6 所示, 在步骤 501 中, 对多样本中 STS 区域的 DNA 序列的测序数据中不合格的数据进行过滤, 得到合格的多样本的测序数据的步骤包括:

步骤 601: 通过测序数据中低质量值碱基的比例进行测序质量过滤, 若低质量值碱基个数超过整条序列碱基个数的预定比例, 则判断为是不合格的数据, 将所述不合格的测序数据过滤掉, 获得初步合格的第一测序数据集。不同的测序设备所用的质量值计算方法不同, 低质量值序列的标准可咨询测序设备提供公司或参考领域内的一般标准。在本实施例中所用的 Illumina 公司的 HiSeq2000 测序仪所用的质量值的计算公式为 $Q = A - 64$, 其中 Q 是某碱基的测序质量值, A 是 HiSeq2000 测序仪输出的 FQ 文件中该碱基对应质量字符的 ASCII 码。质量值低于 5 的碱基在此实施例中被定义为低质量值碱基, 若低质量值碱基个数超过整条序列碱基个数的 50%, 则判断为是不合格的数据, 将所述不合格的测序数据过滤掉, 获得初步合格的第一测序数据集。

步骤 602: 若所述初步合格的第一测序数据集中测序结果不确定的碱基个数超过整条序列碱基个数的 10%, 则判断为是不合格的数据, 将所述不合格的

测序数据过滤掉，获得初步合格的第二测序数据集合；

步骤 603: 将所述初步合格的第二测序数据集合中所有测序数据与测序接头序列库进行比对，若所述初步合格的第二测序数据集合中存在测序接头序列，则判断为是不合格的数据，将所述不合格的测序数据过滤掉，获得初步合格的第三测序数据集合；

步骤 604: 将所述初步合格的第三测序数据集合中所有测序数据与试验中引入的所有外源序列比对，若所述初步合格的第三测序数据集合中存在外源序列，则判断为是不合格的数据，将所述不合格的测序数据过滤掉，获得合格的多样本的测序数据。在本发明实施例中，试验中引入的外源序列为人参考基因组序列。

在实际应用中，高通量测序技术可以是 Illumina Hiseq 2000 测序技术，也可以是现有的其它高通量测序技术。不同的测序仪器或条件可有不同的不合格序列的标准，如 Illumina Hiseq 2000 进行测序时可用的某个标准：测序质量低于某一阈值的碱基个数超过整条序列碱基个数的 50% 则认为是不合格序列，其中，低质量阈值由具体测序技术及测序环境而定；序列中测序结果不确定的碱基（如 Illumina Hiseq 2000 测序结果中的 N）个数超过整条序列碱基个数的 10% 则认为是不合格序列；除样本接头序列外，与其它实验引入的外源序列比对，如各种接头序列，若序列中存在外源序列则认为是不合格序列。

其中，如图 7 所示，根据统计得到的每个样本的测序深度的相关参数，过滤掉不合格的样本的测序数据，得到合格的样本的测序数据的步骤包括：

步骤 701: 将所有样本的测序深度值按照从小到大的顺序进行排序，利用四分位函数确定所述排序后的所有样本的测序深度值的下四分位数 Q1、上四分位数 Q3 以及四分位数间距 IQR。

四分位数 (Quartile)，即在统计学中，把所有数值由小到大排列并分成四等份，处于三个分割点位置的得分就是四分位数。第一四分位数 (Q1)，又称“较小四分位数”，即下四分位数，等于该样本中所有数值由小到大排列后第 25% 的

数字。第二四分位数 (Q2), 又称“中位数”, 等于该样本中所有数值由小到大排列后第 50% 的数字。第三四分位数 (Q3), 又称“较大四分位数”, 即上四分位数, 等于该样本中所有数值由小到大排列后第 75% 的数字。第三四分位数与第一四分位数的差距又称四分位距 (InterQuartile Range, IQR)。不论 Q1, Q2, Q3 的变异量数数值为何, 均视为一个分界点, 以此将总数分成四个相等部份, 可以通过 Q1, Q3 比较, 分析其数据变量的趋势。四分位数在统计学中的箱线图绘制方面应用也很广泛。所谓箱线图就是由一组数据 5 个特征绘制的一个箱子和两条线段的图形, 这种直观的箱线图不仅能反映出一组数据的分布特征, 而且还可以进行多组数据的分析比较。这五个特征值, 即数据的最大值、最小值、中位数和两个四分位数。

步骤 702: 将所有样本的测序深度值在 Q1 减去 1.5 倍 IQR 和 Q3 加上 1.5 倍 IQR 范围之外的不合格的样本的测序数据过滤掉, 得到合格的样本的测序数据。

在本发明一实施例中, 如图 8、9 所示, 横坐标为区域深度分布, 纵坐标为相同深度区域频数, 图 8 是将深度值离群的样本过滤掉, 留下深度值正常的样本, 如图 9 所示, 为深度值正常的样本的示意图。

其中, 如图 10 所示, 根据统计得到的不同样本之间相同的 STS 区域的测序深度的相关参数, 过滤掉不合格 STS 区域的测序数据, 得到合格 STS 区域的测序数据的步骤包括:

步骤 1001: 将不同样本之间相同的 STS 区域的测序深度值按照从小到大的顺序进行排序, 利用四分位函数确定所述排序后的不同样本之间相同的 STS 区域的测序深度值的中位数、上四分位数 Q3 以及四分位数间距 IQR;

步骤 1002: 将不同样本之间相同的 STS 区域的测序深度值中位数为 0 或者中位数大于 Q3 加上 1.5 倍 IQR 的不合格的 STS 区域的测序数据过滤掉, 得到合格的 STS 区域的测序数据。

在本发明一实施例中, 如图 11 所示, 横坐标为不同样品的相同区域的深度

分布，纵坐标为相同深度区域频数，图 11 是将深度值离群的区域过滤掉后的示意图。

其中，如图 12 所示，采用数理统计方法对所述测序数据进行分析，根据所述分析结论，获得每个样本中染色体 STS 区域微缺失的结果的步骤包括：

步骤 1201：将样本的 STS 区域的测序深度值进行均一化，得到均一化的深度值；其中，将样本的 STS 区域的测序深度值进行均一化的步骤包括：将所有样本中相同区域的深度值除以每个样本深度值的平均值，得到所述样本区域均一化的深度值。

步骤 1202：根据得到的样本的 STS 区域的均一化的深度值，采用数理统计方法，检测所述样本 STS 区域的深度值异常值，并获得所述样本 STS 区域微缺失的结果。

在一优选实施例中，如图 13 所示，步骤 1202 的具体步骤包括：

步骤 1301：根据得到的所有样本的同一区域的均一化的深度值，计算所述所有样本的同一区域的均一化深度值的平均值以及方差；

步骤 1302：根据所述所有样本的同一区域的均一化深度值的平均值以及方差，获得所述同一区域所有非离群样本的正态分布曲线；

步骤 1303：根据所述正态分布曲线，计算每个样本在每个区域在特定深度值时的概率值；

步骤 1304：根据每个样本在相应区域在特定深度值时的概率值，设置第一概率值阈值，若所述样本所在区域在特定深度值时的概率值小于所述概率值第一概率值阈值，则获得所述样本区域有微缺失的结果 R1。

在另一优选实施例中，如图 14 所示，在步骤 1304 之后还包括：

步骤 1401：对所述样本区域有微缺失的结果 R1 进行实验验证，根据实验验证结果，设置第二概率值阈值，其中，所述第二概率值阈值小于第一概率值阈值；

步骤 1402：若所述样本区域在特定深度值时的概率值小于所述第二概率值

阈值，则获得所述样本区域有微缺失的结果 R2。

在实际应用中，在大部分样本都无微缺失的情况下，对某个区域而言，微缺失样本的深度值应该为无没有微缺失样本的深度得到的正态分布的异常值，采取适当的数理统计方法，将每个区域的深度的异常值检出，判读深度为异常值的区域为微缺失区域。

判断区域的深度值是否为异常值的方法为：对每个深度值离群过滤之后的区域，取深度离群深度值离群过滤之后的样本的数据，根据其平均值和方差，得到该区域所有非离群样本数据的正态分布曲线。根据该曲线，得到该区域深度为某个特定值的概率值（p.value），对这些概率值设置一个合适的阈值（p.value-cutoff），对概率值小于该阈值的深度的样本，判断该区域有微缺失。

对于概率值阈值的确定，可采用以下方法：通过以上方法算出每个样本每个区域深度在所得到的正态分布中的概率值，采取比较宽松的阈值，先判断更多的样本区域有微缺失，再对这些样本区域进行实验验证，验证其是否为微缺失。实验验证的方法可以为设计合适的引物，对该样本区域的 DNA 进行 PCR 反应，从 PCR 产物情况判断其是否有正常的 PCR 扩增产物，以判断其是否有微缺失。有这些样本区域是否是微缺失的信息之后，就可选择合适的概率值阈值，使之具有最好的假阳性假阴性指标。

在又一优选实施例中，如图 15 所示，根据得到的样本的 STS 区域的均一化的深度值，采用数理统计方法，检测所述样本 STS 区域的深度值异常值，并获得所述样本 STS 区域微缺失的结果的步骤包括：

步骤 1501：根据得到的样本的 STS 区域的均一化的深度值，计算所述样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S；

步骤 1502：根据得到的样本的 STS 区域的均一化的深度值，计算所述样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R；

步骤 1503：将所述比值 D/S 通过 ID3 算法训练出第一比值阈值，将所述比值 D/R 通过 ID3 算法训练出第二比值阈值；

步骤 1504: 若样本区域的比值 D/S 大于第一比值阈值, 则获得所述样本区域没有微缺失的结果;

步骤 1505: 若样本区域的比值 D/S 小于第一比值阈值, 并且样本区域的比值 D/R 大于第二比值阈值, 则获得所述样本区域没有微缺失的结果;

步骤 1506: 若样本区域的比值 D/S 小于第一比值阈值, 并且样本区域的比值 D/R 小于第二比值阈值, 则获得所述样本区域有微缺失的结果。

在实际应用中, 当有多个样本和区域需检测, 且微缺失的样本区域所占的比例较小时, 微缺失区域的深度和所有样本深度的中位数的比例一般很小, 且微缺失区域的深度和所有区域深度的中位数的比例一般也很小, 因此采用适当的参数, 可用决策树的方法进行深度异常值的检验。

决策树检验的具体步骤如下: 算出所要检验的样本区域的深度与所有样本的深度的中位数的比例, 如该比例大于一定的阈值, 则可判断为无微缺失, 但如该比例小于该阈值, 则要进一步判断: 算出所要检验的样本区域的深度与所有区域的深度的中位数的比例, 如该比例大于一定的阈值, 则可判断为无微缺失, 但如该比例小于该阈值, 则判断为有微缺失。

决策树检验中阈值的确定是通过决策树的迭代二叉树三代 (ID3) 算法进行计算的 (Mitchell, Tom M. *Machine Learning*. McGraw-Hill, 1997); ID3 算法的核心是: 在决策树各级结点上选择属性时, 用信息增益 (information gain) 作为属性的选择标准, 以使得在每一个非叶结点进行测试时, 能获得关于被测试记录最大的类别信息。

决策树 ID3 算法的基本思路是: 第一步, 选取一个属性作为决策树的根节点, 然后就这个属性的所有取值创建树的分支; 第二步, 用这棵树来对训练数据进行分类, 如果一个叶节点的所有实例属于同一类, 则以该类标记此节点, 如果所有的叶节点都有类标记, 则算法终止; 第三步, 若还有叶节点没有标记, 则选取一个从该节点到根路径中没有出现过的属性标记该结点, 然后就这个属性所有的取值继续创建树的分支; 重复算法步骤第二步。

在第一步选择不同的属性时会生成不同的决策树, 因此, 选择合适的属性将会生成一棵简单的决策树。在 ID3 算法中, 通常采用一种基于信息的启发式

的方法来决定如何选取属性。启发式方法选取具有最高信息量的属性，也就是生成最少分支决策树的属性。在此发明中有两个属性需要选择，样本区域深度与样本区域深度中位数的比值 $T1 (D/S)$ ，不同样品的相同区域深度与不同样品相同区域深度的中位数的比值 $T2 (D/R)$ 。

属性最高信息量增益以信息增益作为度量标准；信息增益计算方法如下：

设 D 为用类别对训练元组进行的划分，则 D 的熵 (entropy) 表示为：

$$info(D) = - \sum_{i=1}^m p_i \log_2(p_i)$$

其中 p_i 表示第 i 个类别在整个训练元组中出现的概率，可以用属于此类别元素的数量除以训练元组元素总数量进行估计。熵的实际意义是表示 D 中元组的类标号所需要的平均信息量。

现在我们假设将训练元组 D 按属性 A 进行划分，则 A 对 D 划分的期望信息为：

$$info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} info(D_j)$$

而信息增益即为两者的差值：

$$gain(A) = info(D) - info_A(D)$$

ID3 算法就是在每次需要分裂时，计算每个属性的增益率，然后选择增益率最大的属性进行分裂。

下面我们某社区中不真实账号检测的例子说明如何使用 ID3 算法构造决策树。为了简单起见，我们假设训练集合包含 10 个元素，如表 1：

日志密度	好友密度	是否使用真实头像	账号是否真实
s	s	no	no
s	l	yes	yes
l	m	yes	yes
m	m	yes	yes
l	m	yes	yes
m	l	no	yes
m	s	no	no

l	m	no	yes
m	s	no	yes
s	s	yes	no

表 1 某社区中不真实账号检测

其中 s、m 和 l 分别表示小、中和大。设 L、F、H 和 R 表示日志密度、好友密度、是否使用真实头像和账号是否真实，下面计算各属性的信息增益。

$$\text{info}(D) = -0.7\log_2 0.7 - 0.3\log_2 0.3 = 0.7 * 0.51 + 0.3 * 1.74 = 0.879$$

$$\text{gain}(L) = 0.879 - 0.603 = 0.276$$

因此日志密度的信息增益是 0.276。用同样方法得到 H 和 F 的信息增益分别为 0.033 和 0.553。因为 F 具有最大的信息增益，所以第一次分裂选择 F 为分裂属性，分裂后的结果如图 16 表示。

而为了解决与 T1、T2 相类似的具有连续性属性的数据，先对数据进行离散化，简单的做法是把属性值分为 A_i 小于等于 N 和 A_i 大于 N 两段。对于任何一个属性，其所有的取值在一个数据集中是有限的。假设该属性取值为 (v_1 、 v_2 、... v_n)，则在这个集合中，一共存在 $n-1$ 个分段值，然后进一步构建决策树。

其中离散化具体的方法是：

- 1) 寻找该连续型属性的最小值 MIN，寻找该连续型属性的最大值 MAX；
- 2) 设置区间 (MIN, MAX) 中的 N 个等分断点 A_i ，它们分别是 $A_i = \text{MIN} + ((\text{MAX} - \text{MIN}) / N) * i$ ，其中 $i = 1, 2, 3, \dots$ ；
- 3) 分别计算把 (MIN, A_i) 和 (A_i , MAX) 作为区间值时的增益值，进行比较；
- 4) 选取增益值最大的 A_k 作为该连续型属性的断点，把属性值设为 (MIN, A_k) 和 (A_k , MAX) 两个区间。

在本发明实施例中，通过决策树确定阈值的具体步骤是：

第一，首先对选定的几个样本做部分区域的 PCR 验证得到微缺失结果；用“+”代表缺失阳性，即有微缺失，“-”代表缺失阴性，即没有微缺失；计算熵： $\text{Info}(D) = -(P+) \log_2 (P+) - (P-) \log_2 (P-)$ 。

其中 $p+$ ($P-$) 表示在整个训练组中阳 (阴) 性缺失出现的概率，可以用属于此类别元素的数量除以训练组元素总数量作为估计。熵的实际意义表示是 D

中元组的类标号所需要的平均信息量。

第二, 计算比值 $T1$, $T1$ 等于样本区域均一化的深度值与所有样本的深度值的中位数的比值; 对得到的比值 $T1$ 从小到大排序, 得到 $V1$ 、 $V2$ 、 $V3$ 、...、 Vn ; 依次取一对相邻值的平均值作为阈值 $a1'=(V1+V2)/2$ 、 $a2'=(V2+V3)/2$ 、...、 $a(n-1)'=(Vn-1+Vn)/2$; 得到阈值最小值 $a1'$ 与阈值最大值 $a(n-1)'$ 。

第三, 计算比值 $T2$, $T2$ 等于样本区域均一化的深度值与所有区域的深度值的中位数的比值, 对得到的比值 $T2$ 从小到大排序, 得到 $U1$ 、 $U2$ 、 $U3$ 、...、 Un ; 依次取一对相邻值的平均值作为阈值 $b1'=(U1+U2)/2$ 、 $b2'=(U2+U3)/2$ 、...、 $b(n-1)'=(Un-1+Un)/2$; 得到阈值最小值 $b1'$ 与阈值最大值 $b(n-1)'$;

第四, 对于循环得到的分割值 $a1'$ 、 $a2'$ 、..., 计算得到不同分割值的熵:

$$info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} info(D_j)$$

而信息增益即为两者的差值:

$$gain(A) = info(D) - info_A(D)$$

第五, 对于循环得到的分割值 $b1'$ 、 $b2'$ 、..., 计算得到不同分割值的熵:

$$info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} info(D_j)$$

而信息增益即为两者的差值:

$$gain(A) = info(D) - info_A(D)$$

第六, 比较两组的信息增益量的最大增益, 较大者的属性 (a) 则为树根; 将增益量最大的阈值对信息进行分类, 如图 17 所示;

第七, 然后通过如上步骤, 计算两组数据的深度除以区域中位数深度的最大增益, 将增益量最大的阈值作为阈值对信息进行分类, 如图 18 所示, 即为在本发明中最终确定的决策树。

在又一优选实施例中, 如图 19 所示, 根据得到的样本的 STS 区域的均一化的深度值, 采用数理统计方法, 检测所述样本 STS 区域的深度值异常值, 并获得所述样本 STS 区域微缺失的结果的步骤包括:

步骤 1901: 根据得到的所有样本的同一区域的均一化的深度值, 计算所述所有样本的同一区域的均一化深度值的平均值以及方差;

步骤 1902: 根据所述所有样本的同一区域的均一化深度值的平均值以及方差, 获得所述同一区域所有非离群样本的正态分布曲线;

步骤 1903: 根据所述正态分布曲线, 计算每个样本在每个区域在特定深度值时的概率值;

步骤 1904: 根据每个样本在每个区域在特定深度值时的概率值, 设置第三概率值阈值, 若所述样本区域在特定深度值时的概率值小于所述第三概率值阈值, 则获得所述样本区域有微缺失的结果 R3;

步骤 1905: 计算所述结果 R3 中样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S;

步骤 1906: 计算所述结果 R3 中样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R;

步骤 1907: 将所述比值 D/S 通过 ID3 算法训练出第三比值阈值, 将所述比值 D/R 通过 ID3 算法训练出第四比值阈值;

步骤 1908: 若样本区域的比值 D/S 大于第三比值阈值, 则获得所述样本区域没有微缺失的结果;

步骤 1909: 若样本区域的比值 D/S 小于第三比值阈值, 并且样本区域的比值 D/R 大于第四比值阈值, 则获得所述样本区域没有微缺失的结果;

步骤 1910: 若样本区域的比值 D/S 小于第三比值阈值, 并且样本区域的比值 D/R 小于第四比值阈值, 则获得所述样本区域有微缺失的结果。

在实际应用中, 为了更精确地检测出微缺失区域, 可以采用了结合以上两种方法的策略; 先预设一个较为宽松的 p.value-阈值, 然后对得到的结果再进行决策树的过滤, 得到染色体 STS 区域微缺失的结果。

本发明基于染色体序列标签位点 STS 区域微缺失检测的方法, 根据 STS 区域的 DNA 序列, 设计获得捕获探针, 探针涵盖了整个染色体上 STS 区域, 与多样本的 DNA 混合文库进行杂交后, 捕获到的多样本中 STS 区域的 DNA 序列, 能够大量、高效、准确地检测出染色体上已经报道或者未经报道的 STS 相关区

域的微缺失，另外，本发明的数理统计信息分析可以按照正态分布的方法，也可以按照决策树的分析方法，或者将两种方法结合起来，最后还通过实验进行验证，这种信息分析流程科学、稳定，灵敏度高、假阳性低，可以有效的针对微缺失进行分析。

以下实施例用于解释本发明，而不用来限定本发明。本实施例中的操作为本领域的人员可以理解的内容。本实施例中所用试剂和耗材未注明生产商者，均为可通过市场购买的通用产品。本实施例以检测 Y 染色体 STS 区域微缺失的情况为例，但不限于 Y 染色体。本实施例采用 10 个不孕不育样本以及 1 个健康人样本合计 11 个样本，一起建库后杂交同一张 Nimblegen (Roche) 芯片，本实施例样本数用于说明而不是限定一次实验的样本数。

实施例中所使用试剂如表 2:

试剂名称	型号	厂家
QIAquick PCR Purification Kit	28106	QIAGEN
10×Polynucleotide Kinase Buffer	B904	Enzymatics
dNTP Solution Set	N201L	Enzymatics
T4 DNA Polymerase	P708L	Enzymatics
Klenow Enzyme	P706L	Enzymatics
T4 Polynucleotide Kinase	Y904L	Enzymatics
10x blue buffer	B011	Enzymatics
25 mM dATP	28406501V	GE
Klenow (3'-5' exo-)	P701-LC-L	Enzymatics
2x Rapid ligation buffer	B101	Enzymatics
T4 DNA Ligase (Rapid)	L603-HC-L	Enzymatics
PE/PE Index Adapter oligo mix (40 uM)		Invitrogen
MiniElute PCR Purification Kit	28006	QIAGEN
Platinum® Pfx DNA Polymerase	C11708-013	Invitrogen
PCR Primer PE 1.0/PE Index Primer 1.0 (10pM)		Invitrogen
PCR Primer PE 2.0/PE Index Primer 2.0 (index, 10 pM)		Invitrogen
Multiplexing Sample Preparation Oligonucleotide Kit	PE-400-1001	PE-400-1001
2×SC Hybridization Buffer	5340721001	Nimblegen
SC Hybridization Component A	5340721001	Nimblegen
Phusion Mix	F-531L	NEB
Multiplexing Sequencing Primers and Phix Control Kit	PE-400-1002	Illumina

Cot-1 DNA	15279-011	Invitrogen
-----------	-----------	------------

表 2 本发明实施例中所使用的试剂

本发明实施例的实验流程包括:

(一) 基因组 DNA 片段化, 就是单样本的 DNA 文库的制备的步骤中: 利用物理或化学的方法将基因组 DNA 打断成预定大小的 DNA 片段, 回收所述打断的 DNA 片段。

以重量为 3 μ g 的无蛋白质、RNA 污染且没有降解的炎黄基因组 DNA (<http://yh.genomics.org.cn/>) 为起始材料, 使用 Covaris-S2 超声打断仪 (Covaris, US) 仪器进行打断。打断参数设置如表 3:

处理 1	负载比 (%)	10
	强度	10
	循环/脉冲	1000
	时间 (秒)	60
处理 2	时间 (秒)	0
处理 3	时间 (秒)	0
处理 4	时间 (秒)	0
循环		4

表 3 Covaris-S2 超声打断仪参数设置

打断后的片段经电泳检测合格后, 使用 QIAquick PCR Purification Kit 回收纯化, 样本溶于 75 μ L Elution Buffer 中。此处片段经电泳检测合格主要是指主带集中在 200bp~300bp 之间。

(二) 片段 DNA 末端修复, 也就是单样本的 DNA 文库的制备的步骤中: 利用酶对所述回收的 DNA 片段进行末端修复, 形成补平的末端磷酸化的 DNA 片段, 回收所述补平的末端磷酸化的 DNA 片段。

将上一步得到的 DNA 按表 4 在 1.5ml 的离心管中配制末端修复反应体系:

样品 DNA	75 μ L
10 $\times$ Polynucleotide Kinase Buffer	10 μ L
dNTP Solution Set (10mM each)	4 μ L
T4 DNA Polymerase	5 μ L
Klenow Fragment	1 μ L
T4 Polynucleotide Kinase	5 μ L
Total volume	100 μ L

表 4 末端修复反应体系

将上述 100 μL 反应混合物轻微混匀后,于 20 $^{\circ}\text{C}$ 温浴 30 分钟后,用 QIAquick PCR Purification Kit 纯化回收,回收的 DNA 最后于 32 μL ddH₂O 中充分溶解。

(三) DNA 片段末端加“A”,也就是单样本的 DNA 文库的制备的步骤中:利用酶对所述回收的补平的 DNA 片段的 3'末端加上“A”碱基,回收所述 3'末端加上“A”碱基的 DNA 片段。

末端修复后的 DNA 片段按表 5 在 1.5ml 离心管中配制加“A”反应体系:

DNA	32 μL
10x blue buffer	5 μL
dATP(1mM)	10 μL
Klenow (3'-5' exo-)	3 μL
Total volume	50 μL

表 5 末端修复后的 DNA 片段加“A”反应体系

将上述 50 μL 反应混合物轻微混匀后,于 37 $^{\circ}\text{C}$ 温浴 30 分钟后,用 QIAquick PCR Purification Kit 纯化回收,回收的 DNA 最后于 15 μL ddH₂O 中充分溶解。

(四) 标签接头 Adapter 的连接,也就是单样本的 DNA 文库的制备的步骤中:使所述回收的 3'末端加上“A”碱基的 DNA 片段在酶的作用下与标签接头 Index Adapter 连接,并回收带有标签接头的 DNA 片段。

在 1.5 ml 的离心管中配制 Adapter 连接反应体系,如表 6 所示:

末端加“A”的 DNA	15 μL
2x Rapid ligation buffer	25 μL
PE/PE index Adapter oligo mix(40 μM)	5 μL
T4 DNA Ligase (Rapid)	5 μL
Total volume	50 μL

表 6 标签接头 Adapter 连接反应体系

上述 50 μL 反应混合物轻微振荡混合均匀,瞬时离心后置于 20 $^{\circ}\text{C}$ 温浴 15 分钟,反应完后用 MiniElute PCR Purification Kit 进行纯化回收,最后将回收的样品溶于 25 μL Elution Buffer。

(五) 杂交前 PCR, 也就是单样本的 DNA 文库的制备的步骤中: 以标签接头序列的引物做为引物, 对所述带有标签接头的 DNA 片段进行扩增, 回收所述扩增的产物。

以步骤(四)中 DNA 为模板扩增, 以含有接头序列的引物进行扩增, 扩增体系如表 7:

含接头序列的 DNA	25 μ L
10 $\times$ pfx amplification buffer	10 μ L
MgSO ₄ (50mM)	4 μ L
dNTP mix (10mM)	4 μ L
Platinum® Pfx DNA Polymerase	2 μ L
PCR Primer PE 1.0/PE Index Primer 1.0 (10pM)	10 μ L
PCR Primer PE 2.0/PE Index Primer 2.0 (index, 10pM)	10 μ L
ddH ₂ O	35 μ L
Total volume	100 μ L

表 7 PCR 扩增体系

PCR 程序为 94℃ 2 分钟; 4 个循环的 94℃ 15 秒, 62℃ 30 秒, 72℃ 30 秒; 72℃ 5 分钟。PCR 产物用 QIAquick PCR Purification Kit 纯化, 洗脱体积为 30 μ L。

(六) 混合文库构建, 也就是所述多样本的 DNA 混合文库的制备的步骤中: 将所述多个单样本的 DNA 文库按照预定比例混合, 比如可以是相同比例。在实际应用中, 可以根据构建文库的需要, 确定合适的比例。

炎黄文库 (<http://yh.genomics.org.cn>) 与按照步骤(一)至(五)构建的其他 10 个文库取等量的 DNA 混合在一起。

(七) 目标区域与探针的杂交: 也就是一种基于染色体序列标签位点 STS 区域微缺失检测的方法中: 将所述捕获探针与多样本的 DNA 混合文库进行杂交, 以捕获多样本中 STS 区域的 DNA 序列。

1) 在 1.5ml 离心管中加入 450 μ g 的 COT-1 DNA、3 μ g 来自步骤(六)的文库混合产物、1nmol Index-adpater1-block 和 Index-adpater2-block (Multiplexing Sample Preparation Oligonucleotide Kit, Illumina), 混合物置于 SpeedVac(Thermo)

中蒸干，温度设置为 60℃。

2) 在蒸干的离心管中加入 11.2μL 纯水，充分溶解 DNA 后加入 18.5μL 的 2×SC Hybridization Buffer 和 7.3μL 的 SC Hybridization，充分混匀后将混合物转移至杂交仪 (Nimblegen) 上 95℃ 干浴器中，经过 10 分钟使 DNA 变性。

3) 将样品取出震荡后置于离心机上全速离心 30 秒，置于杂交仪 (Nimblegen) 上 42℃ 离心管放置位置，准备杂交。

4) 杂交方法参照 NimbleGen 公司芯片杂交方法 (NimbleGen Arrays User's Guide, Version 3.1, 7 Jul 2009, Roche NimbleGen, Inc.)。样品上样量 35μL，42℃ 杂交 64-72 小时，用 900μL 160mM NaOH 洗脱，洗脱产物用 MinElute PCR Purification Kit 纯化，最终用 80μL Elution Buffer 洗脱。

(八) 捕获后 PCR:

捕获后的文库进行 PCR 扩增，体系为 Phusion Mix 150μL，上下游引物各 4.2μL (Multiplexing Sequencing Primers and Phix Control Kit)，上述的 80μL 洗脱样品加 85μL ddH₂O，混合后分 6 管进行 PCR，PCR 循环数为 16。PCR 反应后把 6 管混合，并用 QIAquick PCR Purification Kit 磁珠纯化回收 300~450bp 大小的片段，洗脱体积为 50μL。

(九) 文库检测:

Bioanalyzer analysis system (Agilent, Santa Clara, USA) 检测文库插入片段大小及含量；Q-PCR 精确定量文库的浓度。

(十) 测序及数据分析：也就是一种基于染色体序列标签位点 STS 区域微缺失检测的方法中：将所述捕获的相应捕获探针的多样本中 STS 区域的 DNA 序列进行测序，得到测序数据；采用数理统计方法对所述测序数据进行分析，根据所述分析结论，获得每个样本中染色体 STS 区域微缺失的结果。

步骤 (九) 中检测合格后的文库上机测序，测序方法参照 Illumina 公司 HiSeq2000 操作方法 (HiSeq 2000 User Guide. Catalog # SY-940-1001 Part # 15011190 Rev B, Illumina)。

本发明实施例的信息分析流程如下：

1. 接收高通量测序技术得到的测序数据，对测序数据进行质控：也就是所述对多样本中 STS 区域的 DNA 序列的测序数据进行质量控制的步骤中：对所述多样本中 STS 区域的 DNA 序列的测序数据的序列中不合格的序列进行过滤，得到合格的多样本的测序数据的序列。

在本发明实施例中，采用 Illumina Hiseq 2000 高通量测序技术。接收到测序序列后，对测序序列进行过滤，去除不合格的序列。不合格序列包括：测序质量值低于 5 的碱基个数超过整条序列碱基个数的 50%则认为是不合格序列；序列中测序结果中 N 的个数超过整条序列碱基个数的 10%则认为是不合格序列；与测序接头序列库进行比对，若序列中存在测序接头序列则认为是不合格序列。

2. 样本区域深度统计：也就是所述对多样本中 STS 区域的 DNA 序列的测序数据进行质量控制的步骤中：通过短序列对比软件，将所述合格的多样本的测序数据与参考基因组序列进行对比，并统计每个样本的测序深度的相关参数以及不同样本之间相同的 STS 区域的测序深度的相关参数。

在本实施例中，采用 SOAPaligner 比对程序，将高通量测序技术得到的测序数据比对到人参考基因组序列上，人参考基因组序列采用 HG19 (<http://genome.ucsc.edu/>)。比对后，进行样品区域测序深度进行统计。

3. 对每个样品进行深度统计，也就是所述对多样本中 STS 区域的 DNA 序列的测序数据进行质量控制的步骤中：根据所述统计得到的每个样本的测序深度的相关参数，过滤掉不合格的样本的测序数据，得到合格的样本的测序数据。

对深度离群深度值离群的样本进行过滤，深度处于 $Q1 - 1.5 \times IQR$ 和 $Q3 + 1.5 \times IQR$ ($Q1$ 、 $Q3$ 为下、上四分位数， IQR 为四分位数间距) 范围之外的深度的样本定义为离群样本，不用这些样本进行下一步正态分布曲线的构建。如图 20 所示，过滤掉深度异常的样品，在图 20 中，左边的图例是样品测序深度中位数的柱状统计图，右边以箱线图来表示样品测序深度中位数结果，由柱状图与箱线图可以看出样品测序深度中位数主要落在 42~60X。

4. 对每个区域深度进行统计,也就是所述对多样本中 STS 区域的 DNA 序列的测序数据进行质量控制的步骤中:根据所述统计得到的不同样本之间相同的 STS 区域的测序深度的相关参数,过滤掉不合格 STS 区域的测序数据,得到合格 STS 区域的测序数据。

把为 0 或大于 $Q3$ 加上 1.5 倍 IQR($Q3$ 为上四分位数,IQR 为四分位数间距)的区域当作离群点去掉,不对其进行微缺失的检验。如图 21 所示,左边的图例是区域测序深度中位数的柱状统计图,右边以箱线图来表示样品区域测序深度中位数结果,由柱状图与箱线图可以看出样品区域测序深度中位数主要落在 35~75X。

5.进行数据质控去掉深度值离群的区域或样本之后,对每个样本的深度值进行均一化处理,每个区域深度值除以改样本深度的平均值,得到标准化之后的深度值,然后利用正态分布曲线进行极端值检验。也就是所述根据得到的样本的 STS 区域的均一化的深度值,采用数理统计方法,检测所述样本 STS 区域的深度值异常值,并获得所述样本 STS 区域微缺失的结果的步骤包括:根据得到的所有样本的同一区域的均一化的深度值,计算所述所有样本的同一区域的均一化深度值的平均值以及方差;根据所述所有样本的同一区域的均一化深度值的平均值以及方差,获得所述同一区域所有非离群样本的正态分布曲线;根据所述正态分布曲线,计算每个样本在每个区域在特定深度值时的概率值;根据每个样本在相应区域在特定深度值时的概率值,设置第一概率值阈值,若所述样本所在区域在特定深度值时的概率值小于所述概率值第一概率值阈值,则获得所述样本区域有微缺失的结果 R1;对所述样本区域有微缺失的结果 R1 进行实验验证,根据实验验证结果,设置第二概率值阈值,其中,所述第二概率值阈值小于第一概率值阈值;若所述样本区域在特定深度值时的概率值小于所述第二概率值阈值,则获得所述样本区域有微缺失的结果 R2。

对每个深度值离群过滤之后的区域,取深度值离群过滤之后的样本的数据,根据其平均值和方差,得到该区域所有非离群样本数据的正态分布曲线。根据

该曲线，得到该区域深度为某个特定值的概率值 $p.value$ 。如图 22 所示，图形横坐标为样品，纵坐标为 $-\log_{10}(p.value)$ ；如图 23 所示的图形是说明 $p.value$ 值得期望值与观察值的结果。通过这个方法，求得测序得到的每个样本每个区域深度的概率值。对概率值较低的 20 个样本区域进行 PCR 反应验证，根据验证的结果对微缺失检验的概率值阈值进行确定，并对微缺失检验的准确性进行评估，具体步骤是：

首先选定一些样品的部分区域，其中包括了缺失阳性及缺失阴性且 $p.value$ 值跨度比较大的区域；然后对不同区域的两端进行引物设计，对样品库做 PCR 扩增，最后做电泳分析得到微缺失的结果；最终通过统计得到具有统计意义的 $P.value$ 的阈值。如表 8 所示，即为得到不同的深度对应的真阳性率和假阳性率，验证得到这个基于高通量测序的微缺失的检验方法的 AUC(area under curve)值可以达到 0.9968254，如图 24 所示，横坐标是假阳性比例，纵坐标是真阳性比例，图 24 是根据不同的真阳性率和假阳性率要求可选择不同的概率值的阈值。

阈值	真阳性	假阳性
0	1	1
1	1	0.20952381
2	1	0.123809524
3	1	0.076190476
4	1	0.038095238
5	1	0.028571429
6	0.974358974	0.028571429
7	0.897435897	0.019047619
8	0.871794872	0.00952381
9	0.846153846	0.00952381
10	0.846153846	0
11	0.820512821	0
12	0.512820513	0
13	0.512820513	0
14	0.512820513	0
15	0.512820513	0
16	0.512820513	0
17	0.512820513	0
18	0.512820513	0
19	0.512820513	0
20	0.487179487	0
21	0.256410256	0

22	0.256410256	0
23	0	0

表 8 p-value 取-log10 的对数后采用不同阈值得到不同的真、假阳性

6.另外借助样品区域深度值，通过决策树（如图 18），判定样品区域缺失阳性还是阴性。也就是所述根据得到的样本的 STS 区域的均一化的深度值，采用数理统计方法，检测所述样本 STS 区域的深度值异常值，并获得所述样本 STS 区域微缺失的结果的步骤包括：根据得到的样本的 STS 区域的均一化的深度值，计算所述样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S；根据得到的样本的 STS 区域的均一化的深度值，计算所述样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R；将所述比值 D/S 通过 ID3 算法训练出第一比值阈值，将所述比值 D/R 通过 ID3 算法训练出第二比值阈值；若样本区域的比值 D/S 大于第一比值阈值，则获得所述样本区域没有微缺失的结果；若样本区域的比值 D/S 小于第一比值阈值，并且样本区域的比值 D/R 大于第二比值阈值，则获得所述样本区域没有微缺失的结果；若样本区域的比值 D/S 小于第一比值阈值，并且样本区域的比值 D/R 小于第二比值阈值，则获得所述样本区域有微缺失的结果。

在本实施范例中，预先设置了相对宽松 p-value 阈值 10^{-6} ，然后对这部分阳性结果进行决策树分类，从而进一步降低假阳性率，最终得到样品区域缺失位点。也就是将上面所述两种方法结合起来进行数理统计学分析，即所述根据得到的样本的 STS 区域的均一化的深度值，采用数理统计方法，检测所述样本 STS 区域的深度值异常值，并获得所述样本 STS 区域微缺失的结果的步骤包括：根据得到的所有样本的同一区域的均一化的深度值，计算所述所有样本的同一区域的均一化深度值的平均值以及方差；根据所述所有样本的同一区域的均一化深度值的平均值以及方差，获得所述同一区域所有非离群样本的正态分布曲线；根据所述正态分布曲线，计算每个样本在每个区域在特定深度值时的概率值；根据每个样本在每个区域在特定深度值时的概率值，设置第三概率值阈值，若所述样本区域在特定深度值时的概率值小于所述第三概率值阈值，则获得所述

样本区域有微缺失的结果 R3；计算所述结果 R3 中样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S；计算所述结果 R3 中样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R；将所述比值 D/S 通过 ID3 算法训练出第三比值阈值，将所述比值 D/R 通过 ID3 算法训练出第四比值阈值；若样本区域的比值 D/S 大于第三比值阈值，则获得所述样本区域没有微缺失的结果；若样本区域的比值 D/S 小于第三比值阈值，并且样本区域的比值 D/R 大于第四比值阈值，则获得所述样本区域没有微缺失的结果；若样本区域的比值 D/S 小于第三比值阈值，并且样本区域的比值 D/R 小于第四比值阈值，则获得所述样本区域有微缺失的结果。

本发明还提供一种计算机可读介质，所述介质承载一系列指令以控制计算机处理器执行如上所述的方法，在此也不再赘述。

图 25 是本发明基于染色体序列标签位点 STS 区域微缺失检测的装置一实施例的结构示意图。如图 25，所述装置包括：捕获探针获得模块 2501、杂交模块 2502、测序数据获得模块 2503 以及微缺失结果获得模块 2504。

捕获探针获得模块 2501 用于选取染色体上的 STS 区域，根据所述 STS 区域的 DNA 序列，设计得到相应的捕获探针。

STS (sequence-tagged site) 序列标签位点，是基因组上定位明确、作为界标并能够通过 PCR 扩增、被唯一操作的短的、单拷贝的 DNA 序列，用于产生作图位点，即测定一系列 STS 的次序即可作出基因组区域的图谱。探针是一小段单链 DNA 或者 RNA 片段（大约是 20 到 500bp），用于检测与其互补的核酸序列。

杂交模块 2502 用于将所述捕获探针与多样本的 DNA 混合文库进行杂交，以捕获多样本中 STS 区域的 DNA 序列。

DNA 文库是指某一特定来源 DNA 通过细胞-DNA 克隆技术构建呈含有所用 DNA 片段的重组 DNA 分子，并转化至细菌内，构成 DNA 文库。此处的 DNA 混合文库是指所有样本混合在一起的所用 DNA 片段。

测序数据获得模块 2503 用于将所述捕获的相应捕获探针的多样本中 STS 区域的 DNA 序列进行测序，得到测序数据。

微缺失结果获得模块 2504 用于采用数理统计方法对所述测序数据进行分析，根据所述分析结论，获得每个样本中染色体 STS 区域微缺失的结果。

其中，捕获探针获得模块 2501 包括：区域查找单元，用于在基因组数据库 (GDB) 中查找染色体上的 STS 区域的 DNA 序列；序列挑选单元，用于在所述查找到的 STS 区域的 DNA 序列中挑选符合捕获探针设计条件的序列；捕获探针获得单元，用于根据所述挑选到的符合捕获探针设计条件的序列，设计并合成得到捕获探针。

所述基因组数据库 (GDB, <http://www.gdb.org/>) 为人类基因组计划 (HGP) 保存和处理基因组图谱的数据库。GDB 是构建关于人类基因组的百科全书，除了构建基因组图谱之外，还开发了描述序列水平的基因组内容的方法，包括序列变异和其它对功能和表型的描述。GDB 数据库以对象模型来保存数据，提供基于 Web 的数据对象检索服务，用户可以搜索各种类型的对象，并以图形方式观看基因组图谱。例如，找出 Y 染色体上 STS 序列在 UCSC 数据库中，在人类基因组参考序列 Hg19 (<http://genome.ucsc.edu/>) 的位置坐标，根据位置坐标即可找到 Y 染色体上 STS 区域的 DNA 序列。在具体实施例中，在基因组数据库中查找染色体上的 STS 区域的 DNA 序列的位置坐标，将此坐标提交给 Roche-NimbleGen 或其他的提供 DNA 捕获服务的公司，由这些公司完成捕获探针的设计与合成。

其中，所述装置还包括多样本 DNA 混合文库制备模块，所述多样本 DNA 混合文库制备模块包括：单样本 DNA 文库制备单元，用于制备多个带有不同接头的质量控制合格的单样本的 DNA 文库；单样本文库混合单元，用于将所述多个单样本的 DNA 文库按照预定比例混合；多样本 DNA 混合文库获得单元，用于检验所述混合的多样本的 DNA 文库的质量是否合格，若是，即为制备的多样本的 DNA 混合文库。

其中, 所述装置还包括测序数据质控模块, 所述测序数据质控模块包括: 合格序列获得单元, 用于对所述多样本中 STS 区域的 DNA 序列的测序数据中不合格的数据进行过滤, 得到合格的多样本的测序数据; 测序深度统计单元, 用于通过短序列对比软件, 将所述合格的多样本的测序数据与参考基因组序列进行对比, 并统计每个样本的测序深度的相关参数以及不同样本之间相同的 STS 区域的测序深度的相关参数; 合格样本获得单元, 用于根据所述统计得到的每个样本的测序深度的相关参数, 过滤掉不合格的样本的测序数据, 得到合格的样本的测序数据; 合格区域获得单元, 用于根据所述统计得到的不同样本之间相同的 STS 区域的测序深度的相关参数, 过滤掉不合格 STS 区域的测序数据, 得到合格 STS 区域的测序数据。

其中, 所述微缺失结果获得模块包括: 深度值均一化单元, 用于将样本的 STS 区域的测序深度值进行均一化, 得到均一化的深度值; 微缺失结果获得单元, 用于根据得到的样本的 STS 区域的均一化的深度值, 采用数理统计方法, 检测所述样本 STS 区域的深度值异常值, 并获得所述样本 STS 区域微缺失的结果。

其中, 所述深度值均一化单元具体用于将所有样本中相同区域的深度值除以每个样本深度值的平均值, 得到所述样本区域均一化的深度值。

其中, 如图 26 所示, 所述微缺失结果获得单元包括:

平均值方差获得单元 2601, 用于根据得到的所有样本的同一区域的均一化的深度值, 计算所述所有样本的同一区域的均一化深度值的平均值以及方差;

正态分布曲线获得单元 2602, 用于根据所述所有样本的同一区域的均一化深度值的平均值以及方差, 获得所述同一区域所有非离群样本的正态分布曲线;

概率值计算单元 2603, 用于根据所述正态分布曲线, 计算每个样本在每个区域在特定深度值时的概率值;

第一判断单元 2604, 用于根据每个样本在相应区域在特定深度值时的概率值, 设置第一概率值阈值, 若所述样本所在区域在特定深度值时的概率值小于

所述概率值第一概率值阈值, 则获得所述样本区域有微缺失的结果 R1。

在一优选实施例中, 如图 27 所示, 所述微缺失结果获得单元还包括:

概率值阈值确定单元 2701, 用于对所述样本区域有微缺失的结果 R1 进行实验验证, 根据实验验证结果, 设置第二概率值阈值, 其中, 所述第二概率值阈值小于第一概率值阈值;

第二判断单元 2702, 用于若所述样本区域在特定深度值时的概率值小于所述第二概率值阈值, 则获得所述样本区域有微缺失的结果 R2。

如图 28 所示, 在又一优选实施例中, 所述微缺失结果获得单元包括:

比值 D/S 获得单元 2801, 用于根据得到的样本的 STS 区域的均一化的深度值, 计算所述样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S;

比值 D/R 获得单元 2802, 用于根据得到的样本的 STS 区域的均一化的深度值, 计算所述样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R;

第一、二比值阈值获得单元 2803, 用于将所述比值 D/S 通过 ID3 算法训练出第一比值阈值, 将所述比值 D/R 通过 ID3 算法训练出第二比值阈值;

第一判断单元 2804, 用于若样本区域的比值 D/S 大于第一比值阈值, 则获得所述样本区域没有微缺失的结果;

第二判断单元 2805, 用于若样本区域的比值 D/S 小于第一比值阈值, 并且样本区域的比值 D/R 大于第二比值阈值, 则获得所述样本区域没有微缺失的结果;

第三判断单元 2806, 用于若样本区域的比值 D/S 小于第一比值阈值, 并且样本区域的比值 D/R 小于第二比值阈值, 则获得所述样本区域有微缺失的结果。

如图 29 所示, 在又一优选实施例中, 所述微缺失结果获得单元包括:

平均值方差获得单元 2901, 用于根据得到的所有样本的同一区域的均一化的深度值, 计算所述所有样本的同一区域的均一化深度值的平均值以及方差;

正态分布曲线获得单元 2902, 用于根据所述所有样本的同一区域的均一化深度值的平均值以及方差, 获得所述同一区域所有非离群样本的正态分布曲线;

概率值计算单元 2903, 用于根据所述正态分布曲线, 计算每个样本在每个区域在特定深度值时的概率值;

第一判断单元 2904, 用于根据每个样本在每个区域在特定深度值时的概率值, 设置第三概率值阈值, 若所述样本区域在特定深度值时的概率值小于所述第三概率值阈值, 则获得所述样本区域有微缺失的结果 R3;

比值 D/S 获得单元 2905, 用于计算所述结果 R3 中样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S;

比值 D/R 获得单元 2906, 用于计算所述结果 R3 中样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R;

第三、四比值阈值获得单元 2907, 用于将所述比值 D/S 通过 ID3 算法训练出第三比值阈值, 将所述比值 D/R 通过 ID3 算法训练出第四比值阈值;

第二判断单元 2908, 用于若样本区域的比值 D/S 大于第三比值阈值, 则获得所述样本区域没有微缺失的结果;

第三判断单元 2909, 用于若样本区域的比值 D/S 小于第三比值阈值, 并且样本区域的比值 D/R 大于第四比值阈值, 则获得所述样本区域没有微缺失的结果;

第四判断单元 2910, 用于若样本区域的比值 D/S 小于第三比值阈值, 并且样本区域的比值 D/R 小于第四比值阈值, 则获得所述样本区域有微缺失的结果。

本发明基于染色体序列标签位点 STS 区域微缺失检测的装置, 根据 STS 区域的 DNA 序列, 设计获得捕获探针, 探针涵盖了整个染色体上 STS 区域, 与多样本的 DNA 混合文库进行杂交后, 捕获到的多样本中 STS 区域的 DNA 序列, 能够大量、高效、准确地检测出染色体上已经报道或者未经报道的 STS 相关区域的微缺失, 另外, 本发明的数理统计信息分析可以按照正态分布的方法, 也可以按照决策树的分析方法, 或者将两种方法结合起来, 最后还通过实验进行验证, 这种信息分析流程科学、稳定, 灵敏度高、假阳性低, 可以有效的针对微缺失进行分析。

以上所述仅为本发明的实施例，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。

权利要求

1.一种基于染色体序列标签位点 STS 区域微缺失检测的方法,其特征在于:
所述方法包括:

选取染色体上的 STS 区域,根据所述 STS 区域的 DNA 序列,设计得到相应的捕获探针;

将所述捕获探针与多样本的 DNA 混合文库进行杂交,以捕获多样本中 STS 区域的 DNA 序列;

将所述捕获的相应捕获探针的多样本中 STS 区域的 DNA 序列进行测序,得到测序数据;

采用数理统计方法对所述测序数据进行分析,根据所述分析结论,获得每个样本中染色体 STS 区域微缺失的结果。

2.根据权利要求 1 所述的方法,其特征在于:所述采用数理统计方法对所述测序数据进行分析,根据所述分析结论,获得每个样本中染色体 STS 区域微缺失的结果的步骤包括:

将样本的 STS 区域的测序深度值进行均一化,得到均一化的深度值;

根据得到的样本的 STS 区域的均一化的深度值,采用数理统计方法,检测所述样本 STS 区域的深度值异常值,并获得所述样本 STS 区域微缺失的结果。

3.根据权利要求 2 所述的方法,其特征在于:所述将样本的 STS 区域的测序深度值进行均一化的步骤包括:将所有样本中相同区域的深度值除以每个样本深度值的平均值,得到所述样本区域均一化的深度值。

4.根据权利要求 2 所述的方法,其特征在于:所述根据得到的样本的 STS 区域的均一化的深度值,采用数理统计方法,检测所述样本 STS 区域的深度值异常值,并获得所述样本 STS 区域微缺失的结果的步骤包括:

根据得到的所有样本的同一区域的均一化的深度值,计算所述所有样本的同一区域的均一化深度值的平均值以及方差;

根据所述所有样本的同一区域的均一化深度值的平均值以及方差，获得所述同一区域所有非离群样本的正态分布曲线；

根据所述正态分布曲线，计算每个样本在每个区域在特定深度值时的概率值；

根据每个样本在相应区域在特定深度值时的概率值，设置第一概率值阈值，若所述样本所在区域在特定深度值时的概率值小于所述概率值第一概率值阈值，则获得所述样本区域有微缺失的结果 R1。

5.根据权利要求 4 所述的方法，其特征在于：所述根据每个样本在相应区域在特定深度值时的概率值，设置第一概率值阈值，若所述样本所在区域在特定深度值时的概率值小于所述概率值第一概率值阈值，则获得所述样本区域有微缺失的结果 R1 的步骤之后还包括：

对所述样本区域有微缺失的结果 R1 进行实验验证，根据实验验证结果，设置第二概率值阈值，其中，所述第二概率值阈值小于第一概率值阈值；

若所述样本区域在特定深度值时的概率值小于所述第二概率值阈值，则获得所述样本区域有微缺失的结果 R2。

6.根据权利要求 2 所述的方法，其特征在于：所述根据得到的样本的 STS 区域的均一化的深度值，采用数理统计方法，检测所述样本 STS 区域的深度值异常值，并获得所述样本 STS 区域微缺失的结果的步骤包括：

根据得到的样本的 STS 区域的均一化的深度值，计算所述样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S；

根据得到的样本的 STS 区域的均一化的深度值，计算所述样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R；

将所述比值 D/S 通过 ID3 算法训练出第一比值阈值，将所述比值 D/R 通过 ID3 算法训练出第二比值阈值；

若样本区域的比值 D/S 大于第一比值阈值，则获得所述样本区域没有微缺失的结果；

若样本区域的比值 D/S 小于第一比值阈值, 并且样本区域的比值 D/R 大于第二比值阈值, 则获得所述样本区域没有微缺失的结果;

若样本区域的比值 D/S 小于第一比值阈值, 并且样本区域的比值 D/R 小于第二比值阈值, 则获得所述样本区域有微缺失的结果。

7. 根据权利要求 2 所述的方法, 其特征在于: 所述根据得到的样本的 STS 区域的均一化的深度值, 采用数理统计方法, 检测所述样本 STS 区域的深度值异常值, 并获得所述样本 STS 区域微缺失的结果的步骤包括:

根据得到的所有样本的同一区域的均一化的深度值, 计算所述所有样本的同一区域的均一化深度值的平均值以及方差;

根据所述所有样本的同一区域的均一化深度值的平均值以及方差, 获得所述同一区域所有非离群样本的正态分布曲线;

根据所述正态分布曲线, 计算每个样本在每个区域在特定深度值时的概率值;

根据每个样本在每个区域在特定深度值时的概率值, 设置第三概率值阈值, 若所述样本区域在特定深度值时的概率值小于所述第三概率值阈值, 则获得所述样本区域有微缺失的结果 R3;

计算所述结果 R3 中样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S ;

计算所述结果 R3 中样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R ;

将所述比值 D/S 通过 ID3 算法训练出第三比值阈值, 将所述比值 D/R 通过 ID3 算法训练出第四比值阈值;

若样本区域的比值 D/S 大于第三比值阈值, 则获得所述样本区域没有微缺失的结果;

若样本区域的比值 D/S 小于第三比值阈值, 并且样本区域的比值 D/R 大于第四比值阈值, 则获得所述样本区域没有微缺失的结果;

若样本区域的比值 D/S 小于第三比值阈值, 并且样本区域的比值 D/R 小于第四比值阈值, 则获得所述样本区域有微缺失的结果。

8. 根据权利要求 1 所述的方法, 其特征在于: 所述选取染色体上的 STS 区域, 根据所述 STS 区域的 DNA 序列, 设计并合成相应的捕获探针的步骤包括: 在基因组数据库中查找染色体上的 STS 区域的 DNA 序列; 在所述查找到的 STS 区域的 DNA 序列中挑选符合捕获探针设计条件的序列; 根据所述挑选到的符合捕获探针设计条件的序列, 设计并合成得到捕获探针。

9. 根据权利要求 1 所述的方法, 其特征在于: 所述多样本的 DNA 混合文库的制备的步骤包括:

制备多个带有不同接头的质量控制合格的单样本的 DNA 文库;

将所述多个单样本的 DNA 文库按照预定比例混合;

检验所述混合的多样本的 DNA 文库的质量是否合格, 若是, 即为制备的多样本的 DNA 混合文库。

10. 根据权利要求 8 所述的方法, 其特征在于: 所述单样本的 DNA 文库的制备的步骤包括:

利用物理或化学的方法将基因组 DNA 打断成预定大小的 DNA 片段, 回收所述打断的 DNA 片段;

利用酶对所述回收的 DNA 片段进行末端修复, 形成补平的末端磷酸化的 DNA 片段, 回收所述补平的末端磷酸化的 DNA 片段;

利用酶对所述回收的补平的 DNA 片段的 3' 末端加上“A”碱基, 回收所述 3' 末端加上“A”碱基的 DNA 片段;

使所述回收的 3' 末端加上“A”碱基的 DNA 片段在酶的作用下与标签接头 Index Adapter 连接, 并回收带有标签接头的 DNA 片段;

以标签接头序列的引物做为引物, 对所述带有标签接头的 DNA 片段进行扩增, 回收所述扩增的产物;

检验所述扩增的产物的质量控制是否合格,若是,即为制备的单样本的 DNA 文库。

11.根据权利要求 1 所述的方法,其特征在于:所述将捕获的相应捕获探针的多样本中 STS 区域的 DNA 序列进行测序,得到测序数据的步骤之后还包括:对所述对多样本中 STS 区域的 DNA 序列的测序数据进行质量控制。

12.根据权利要求 10 所述的方法,其特征在于:所述对多样本中 STS 区域的 DNA 序列的测序数据进行质量控制的步骤包括:

对所述多样本中 STS 区域的 DNA 序列的测序数据中不合格的数据进行过滤,得到合格的多样本的测序数据;

通过短序列对比软件,将所述合格的多样本的测序数据与参考基因组序列进行对比,并统计每个样本的测序深度的相关参数以及不同样本之间相同的 STS 区域的测序深度的相关参数;

根据所述统计得到的每个样本的测序深度的相关参数,过滤掉不合格的样本的测序数据,得到合格的样本的测序数据;

根据所述统计得到的不同样本之间相同的 STS 区域的测序深度的相关参数,过滤掉不合格 STS 区域的测序数据,得到合格 STS 区域的测序数据。

13.根据权利要求 11 所述的方法,其特征在于:所述对多样本中 STS 区域的 DNA 序列的测序数据中不合格的数据进行过滤,得到合格的多样本的测序数据的步骤包括:

通过测序数据中低质量值碱基的比例进行测序质量过滤,若低质量值碱基个数超过整条序列碱基个数的预定比例,则判断为是不合格的数据,将所述不合格的测序数据过滤掉,获得初步合格的第一测序数据集合;

若所述初步合格的第一测序数据集合中测序结果不确定的碱基个数超过整条序列碱基个数的 10%,则判断为是不合格的数据,将所述不合格的测序数据过滤掉,获得初步合格的第二测序数据集合;

将所述初步合格的第二测序数据集合中所有测序数据与测序接头序列库进

行比对, 若所述初步合格的第二测序数据集合中存在测序接头序列, 则判断为是不合格的数据, 将所述不合格的测序数据过滤掉, 获得初步合格的第三测序数据集合;

将所述初步合格的第三测序数据集合中所有测序数据与试验中引入的所有外源序列比对, 若所述初步合格的第三测序数据集合中存在外源序列, 则判断为是不合格的数据, 将所述不合格的测序数据过滤掉, 获得合格的多样本的测序数据。

14. 根据权利要求 11 所述的方法, 其特征在于: 所述根据统计得到的每个样本的测序深度的相关参数, 过滤掉不合格的样本的测序数据, 得到合格的样本的测序数据的步骤包括:

将所有样本的测序深度值按照从小到大的顺序进行排序, 利用四分位函数确定所述排序后的所有样本的测序深度值的下四分位数 $Q1$ 、上四分位数 $Q3$ 以及四分位数间距 IQR ;

将所有样本的测序深度值在 $Q1$ 减去 1.5 倍 IQR 和 $Q3$ 加上 1.5 倍 IQR 范围之外的不合格的样本的测序数据过滤掉, 得到合格的样本的测序数据。

15. 根据权利要求 11 所述的方法, 其特征在于: 所述根据统计得到的不同样本之间相同的 STS 区域的测序深度的相关参数, 过滤掉不合格 STS 区域的测序数据, 得到合格 STS 区域的测序数据的步骤包括:

将不同样本之间相同的 STS 区域的测序深度值按照从小到大的顺序进行排序, 利用四分位函数确定所述排序后的不同样本之间相同的 STS 区域的测序深度值的中位数、上四分位数 $Q3$ 以及四分位数间距 IQR ;

将不同样本之间相同的 STS 区域的测序深度值中位数为 0 或者中位数大于 $Q3$ 加上 1.5 倍 IQR 的不合格的 STS 区域的测序数据过滤掉, 得到合格的 STS 区域的测序数据。

16. 一种基于染色体序列标签位点 STS 区域微缺失检测的装置, 其特征在于: 所述装置包括:

捕获探针获得模块，用于选取染色体上的 STS 区域，根据所述 STS 区域的 DNA 序列，设计得到相应的捕获探针；

杂交模块，用于将所述捕获探针与多样本的 DNA 混合文库进行杂交，以捕获多样本中 STS 区域的 DNA 序列；

测序数据获得模块，用于将所述捕获的相应捕获探针的多样本中 STS 区域的 DNA 序列进行测序，得到测序数据；

微缺失结果获得模块，用于采用数理统计方法对所述测序数据进行分析，根据所述分析结论，获得每个样本中染色体 STS 区域微缺失的结果。

17.根据权利要求 15 所述的装置，其特征在于：所述微缺失结果获得模块包括：

深度值均一化单元，用于将样本的 STS 区域的测序深度值进行均一化，得到均一化的深度值；

微缺失结果获得单元，用于根据得到的样本的 STS 区域的均一化的深度值，采用数理统计方法，检测所述样本 STS 区域的深度值异常值，并获得所述样本 STS 区域微缺失的结果。

18.根据权利要求 16 所述的装置，其特征在于：所述深度值均一化单元具体用于将所有样本中相同区域的深度值除以每个样本深度值的平均值，得到所述样本区域均一化的深度值。

19.根据权利要求 16 所述的装置，其特征在于：所述微缺失结果获得单元包括：

平均值方差获得单元，用于根据得到的所有样本的同一区域的均一化的深度值，计算所述所有样本的同一区域的均一化深度值的平均值以及方差；

正态分布曲线获得单元，用于根据所述所有样本的同一区域的均一化深度值的平均值以及方差，获得所述同一区域所有非离群样本的正态分布曲线；

概率值计算单元，用于根据所述正态分布曲线，计算每个样本在每个区域在特定深度值时的概率值；

第一判断单元,用于根据每个样本在相应区域在特定深度值时的概率值,设置第一概率值阈值,若所述样本所在区域在特定深度值时的概率值小于所述概率值第一概率值阈值,则获得所述样本区域有微缺失的结果 R1。

20.根据权利要求 19 所述的装置,其特征在于:所述微缺失结果获得单元还包括:

概率值阈值确定单元,用于对所述样本区域有微缺失的结果 R1 进行实验验证,根据实验验证结果,设置第二概率值阈值,其中,所述第二概率值阈值小于第一概率值阈值;

第二判断单元,用于若所述样本区域在特定深度值时的概率值小于所述第二概率值阈值,则获得所述样本区域有微缺失的结果 R2。

21.根据权利要求 16 所述的装置,其特征在于:所述微缺失结果获得单元包括:

比值 D/S 获得单元,用于根据得到的样本的 STS 区域的均一化的深度值,计算所述样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S;

比值 D/R 获得单元,用于根据得到的样本的 STS 区域的均一化的深度值,计算所述样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R;

第一、二比值阈值获得单元,用于将所述比值 D/S 通过 ID3 算法训练出第一比值阈值,将所述比值 D/R 通过 ID3 算法训练出第二比值阈值;

第一判断单元,用于若样本区域的比值 D/S 大于第一比值阈值,则获得所述样本区域没有微缺失的结果;

第二判断单元,用于若样本区域的比值 D/S 小于第一比值阈值,并且样本区域的比值 D/R 大于第二比值阈值,则获得所述样本区域没有微缺失的结果;

第三判断单元,用于若样本区域的比值 D/S 小于第一比值阈值,并且样本区域的比值 D/R 小于第二比值阈值,则获得所述样本区域有微缺失的结果。

22.根据权利要求 16 所述的装置,其特征在于:所述微缺失结果获得单元包括:

平均值方差获得单元, 用于根据得到的所有样本的同一区域的均一化的深度值, 计算所述所有样本的同一区域的均一化深度值的平均值以及方差;

正态分布曲线获得单元, 用于根据所述所有样本的同一区域的均一化深度值的平均值以及方差, 获得所述同一区域所有非离群样本的正态分布曲线;

概率值计算单元, 用于根据所述正态分布曲线, 计算每个样本在每个区域在特定深度值时的概率值;

第一判断单元, 用于根据每个样本在每个区域在特定深度值时的概率值, 设置第三概率值阈值, 若所述样本区域在特定深度值时的概率值小于所述第三概率值阈值, 则获得所述样本区域有微缺失的结果 R3;

比值 D/S 获得单元, 用于计算所述结果 R3 中样本区域均一化的深度值与所有样本的深度值的中位数的比值 D/S;

比值 D/R 获得单元, 用于计算所述结果 R3 中样本区域均一化的深度值与所有区域的深度值的中位数的比值 D/R;

第三、四比值阈值获得单元, 用于将所述比值 D/S 通过 ID3 算法训练出第三比值阈值, 将所述比值 D/R 通过 ID3 算法训练出第四比值阈值;

第二判断单元, 用于若样本区域的比值 D/S 大于第三比值阈值, 则获得所述样本区域没有微缺失的结果;

第三判断单元, 用于若样本区域的比值 D/S 小于第三比值阈值, 并且样本区域的比值 D/R 大于第四比值阈值, 则获得所述样本区域没有微缺失的结果;

第四判断单元, 用于若样本区域的比值 D/S 小于第三比值阈值, 并且样本区域的比值 D/R 小于第四比值阈值, 则获得所述样本区域有微缺失的结果。

23. 根据权利要求 15 所述的装置, 其特征在于: 所述捕获探针获得模块包括:

区域查找单元, 用于在基因组数据库中查找染色体上的 STS 区域的 DNA 序列;

序列挑选单元, 用于在所述查找到的 STS 区域的 DNA 序列中挑选符合捕获探针设计条件的序列;

捕获探针获得单元,用于根据所述挑选到的符合捕获探针设计条件的序列,设计并合成得到捕获探针。

24.根据权利要求 15 所述的装置,其特征在于:所述装置还包括多样本 DNA 混合文库制备模块,所述多样本 DNA 混合文库制备模块包括:

单样本 DNA 文库制备单元,用于制备多个带有不同接头的质量控制合格的单样本的 DNA 文库;

单样本文库混合单元,用于将所述多个单样本的 DNA 文库按照预定比例混合;

多样本 DNA 混合文库获得单元,用于检验所述混合的多样本的 DNA 文库的质量是否合格,若是,即为制备的多样本的 DNA 混合文库。

25.根据权利要求 15 所述的装置,其特征在于:所述装置还包括测序数据质控模块,所述测序数据质控模块包括:

合格序列获得单元,用于对所述多样本中 STS 区域的 DNA 序列的测序数据中不合格的数据进行过滤,得到合格的多样本的测序数据的序列;

测序深度统计单元,用于通过短序列对比软件,将所述合格的多样本的测序数据与参考基因组序列进行对比,并统计每个样本的测序深度的相关参数以及不同样本之间相同的 STS 区域的测序深度的相关参数;

合格样本获得单元,用于根据所述统计得到的每个样本的测序深度的相关参数,过滤掉不合格的样本的测序数据,得到合格的样本的测序数据;

合格区域获得单元,用于根据所述统计得到的不同样本之间相同的 STS 区域的测序深度的相关参数,过滤掉不合格 STS 区域的测序数据,得到合格 STS 区域的测序数据。

26.一种计算机可读介质,其特征在于,所述介质承载一系列指令以控制计算机处理器执行如权利要求 1 至 15 中任一项所述的方法。

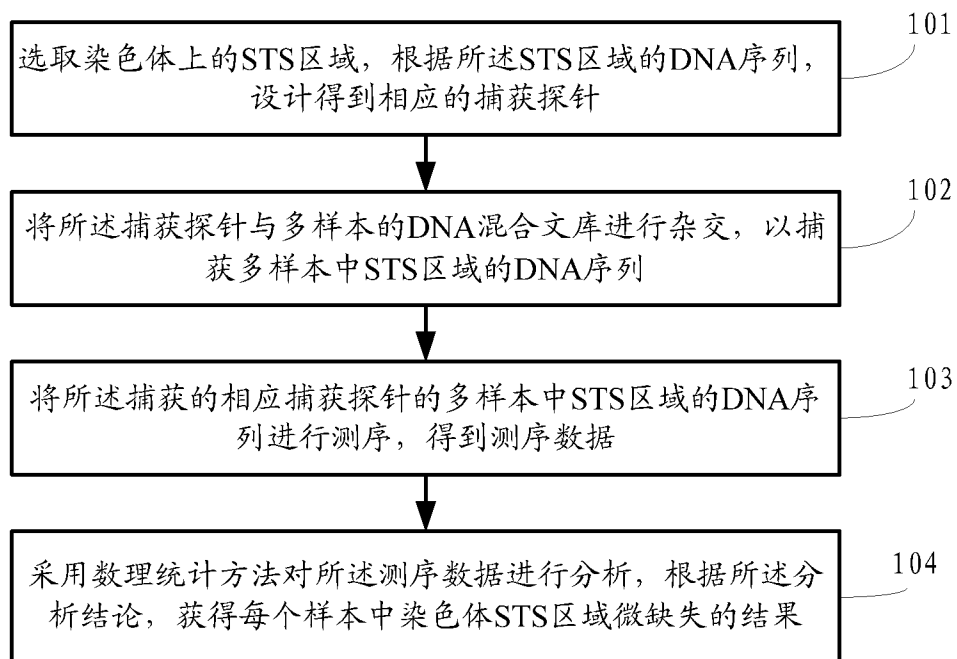


图 1

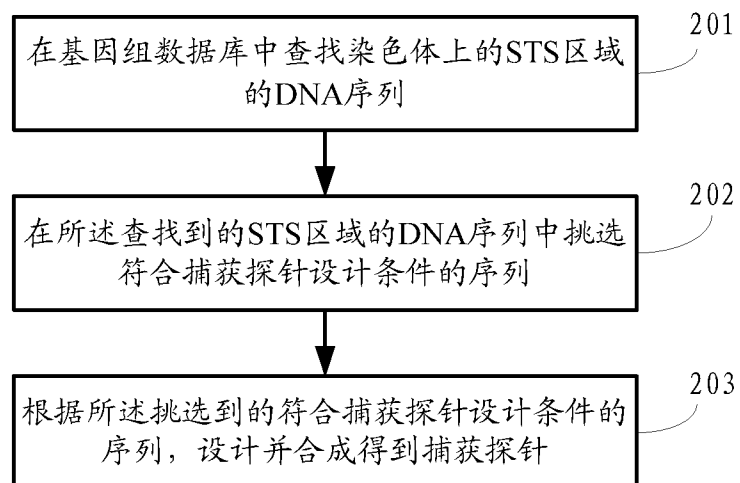


图 2

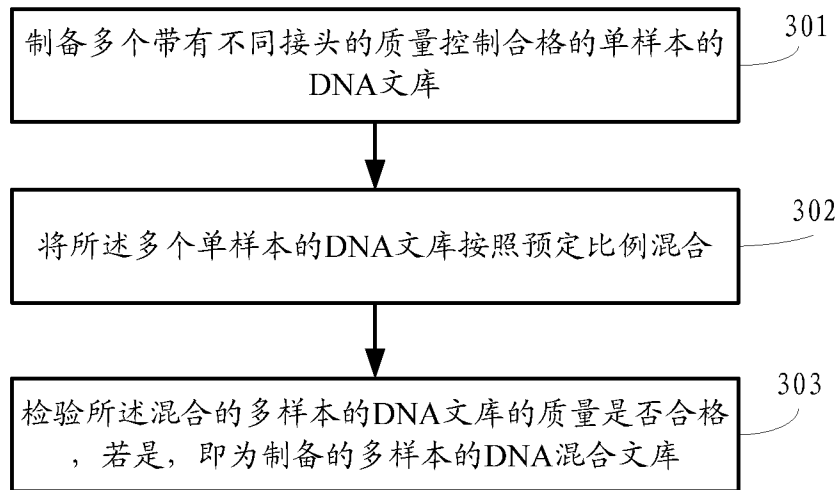


图 3

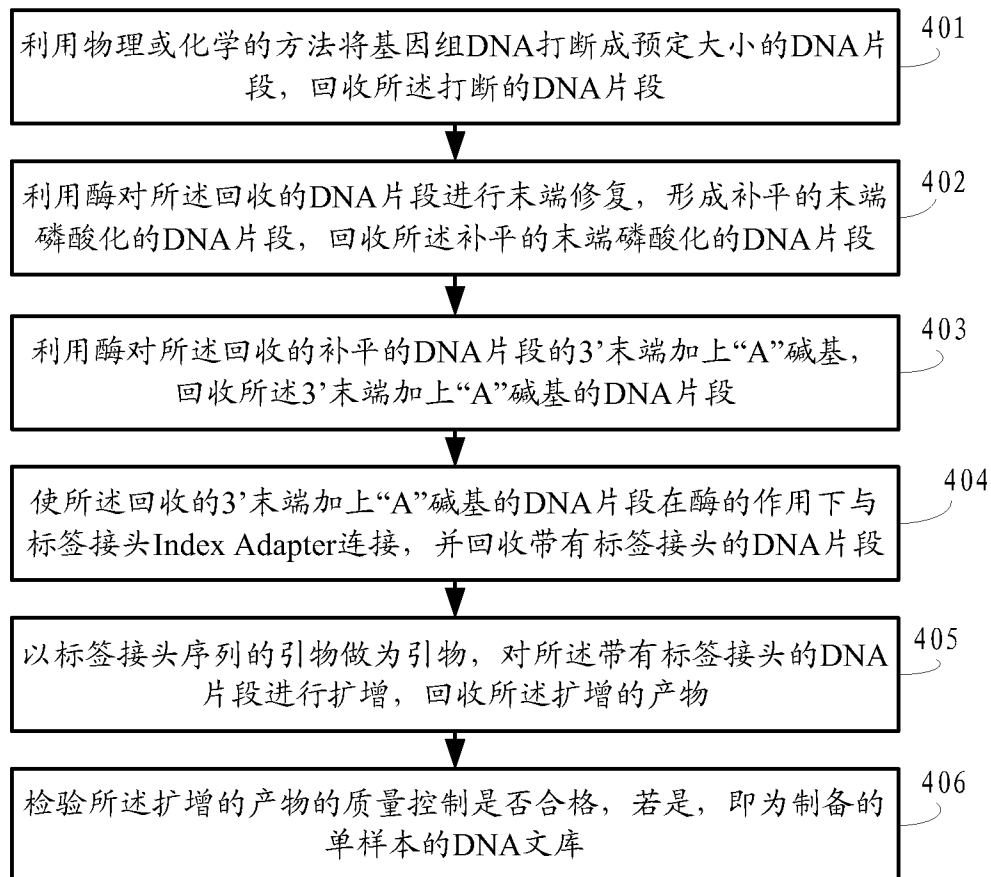


图 4

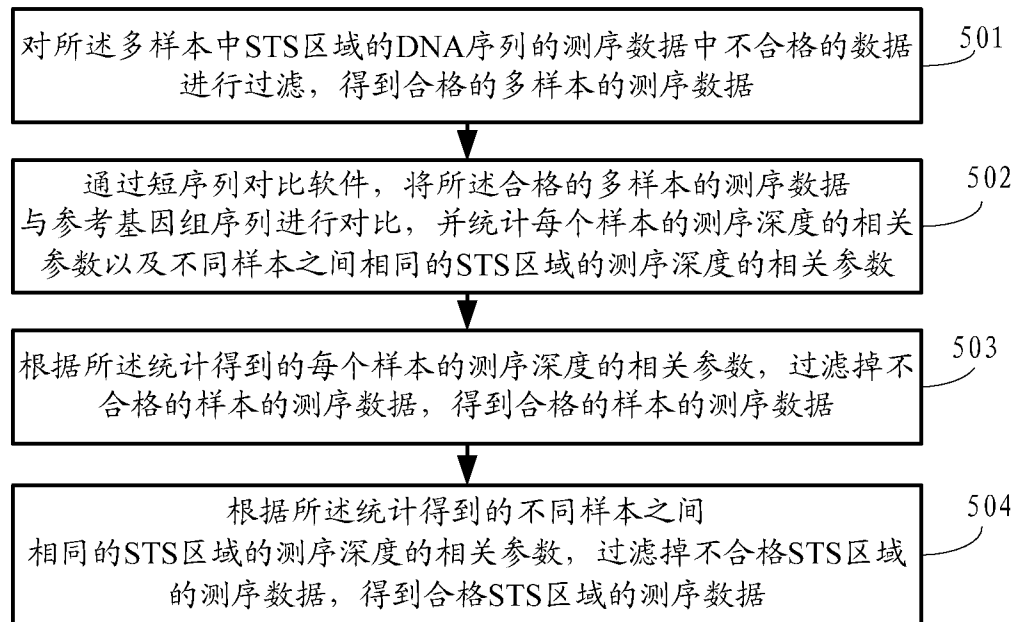


图 5

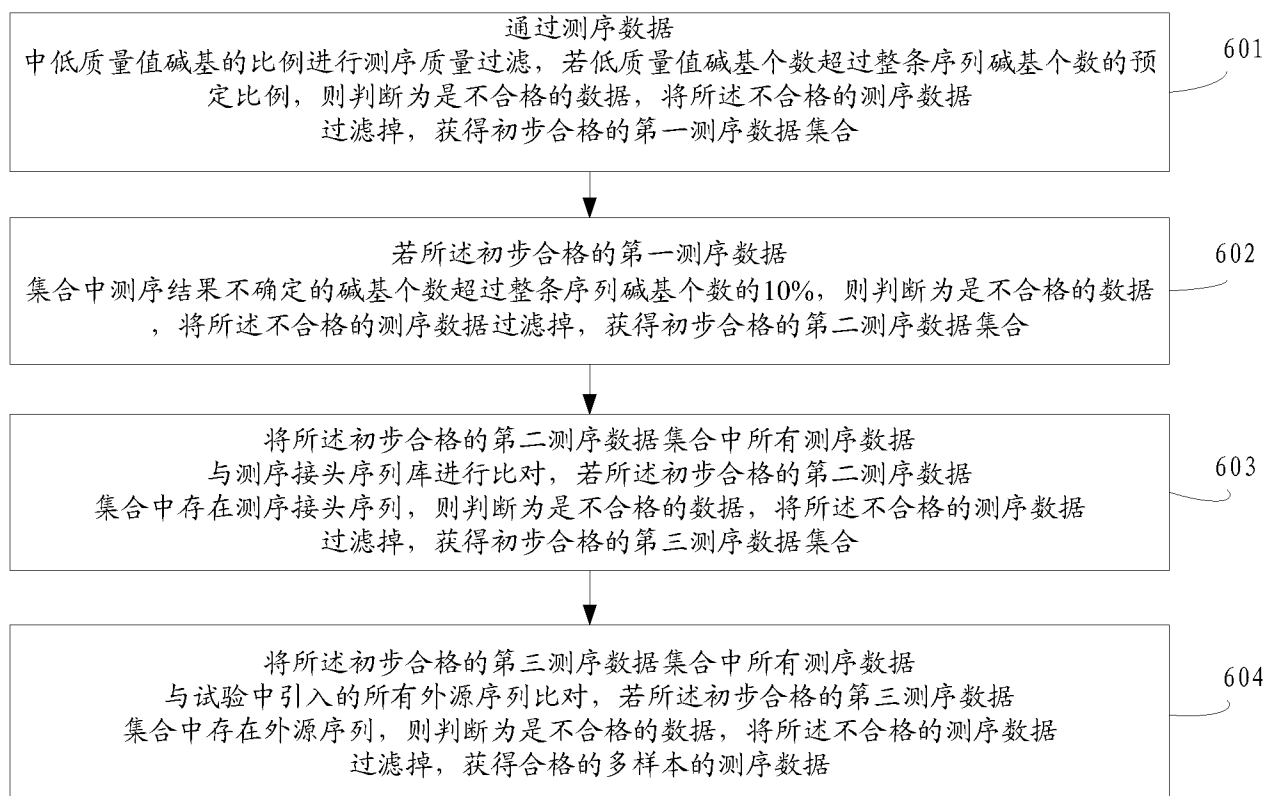


图 6

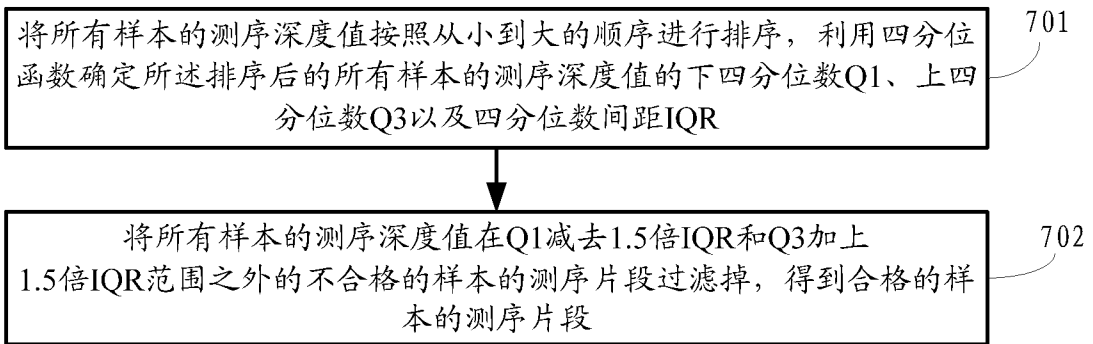


图 7

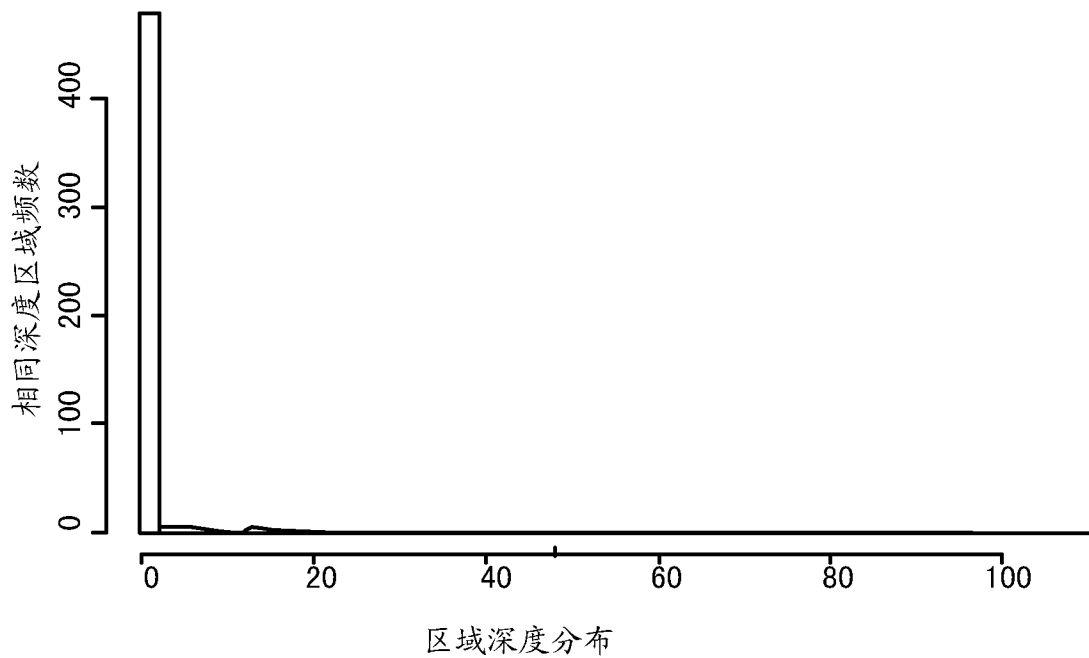


图 8

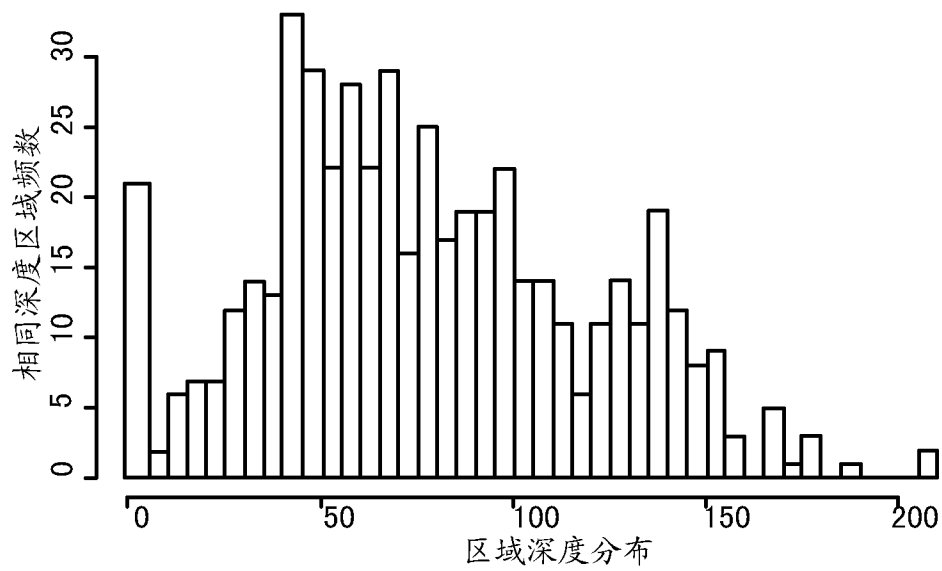


图 9

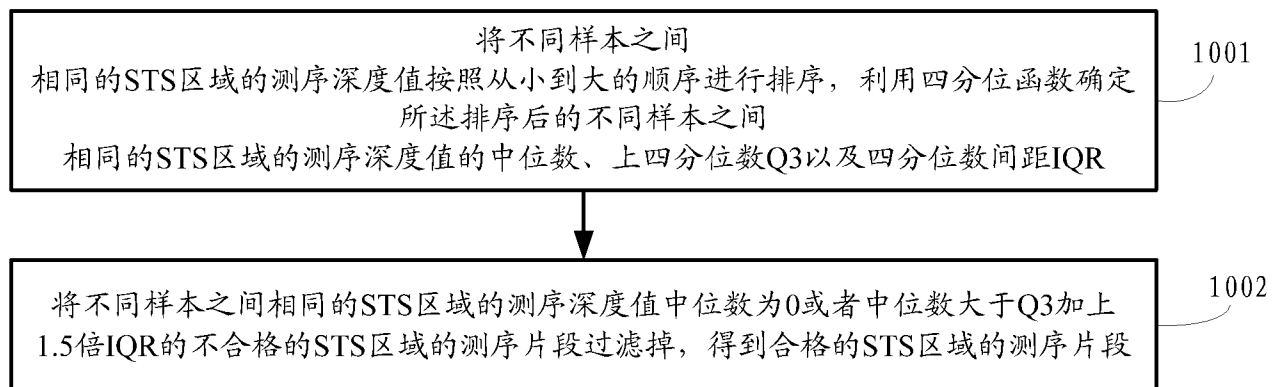


图 10

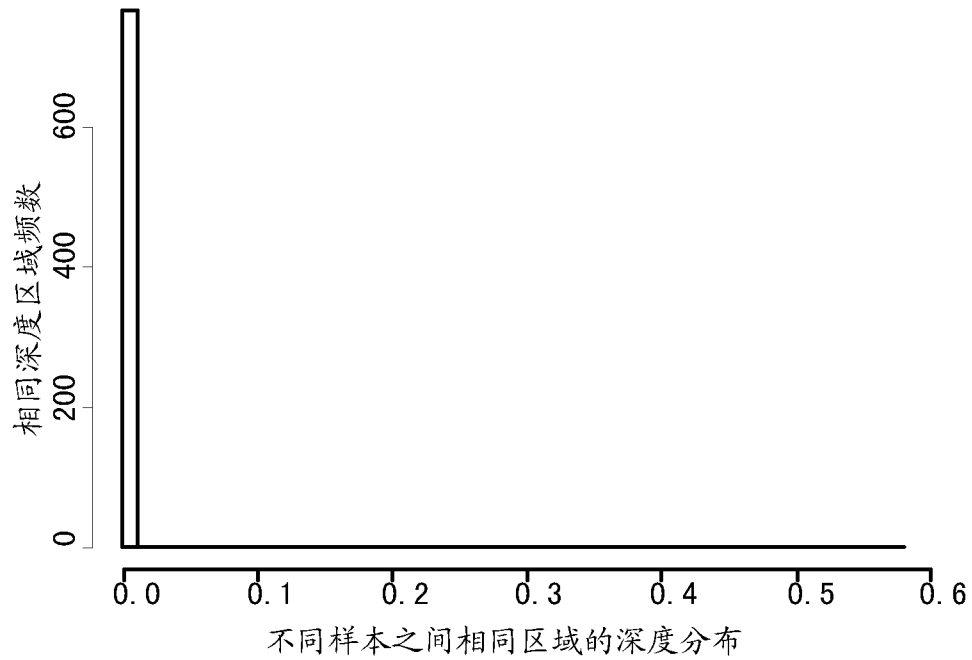


图 11

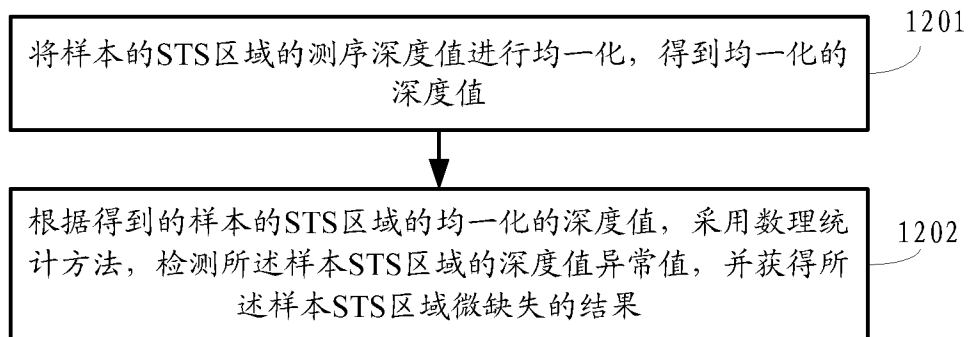


图 12

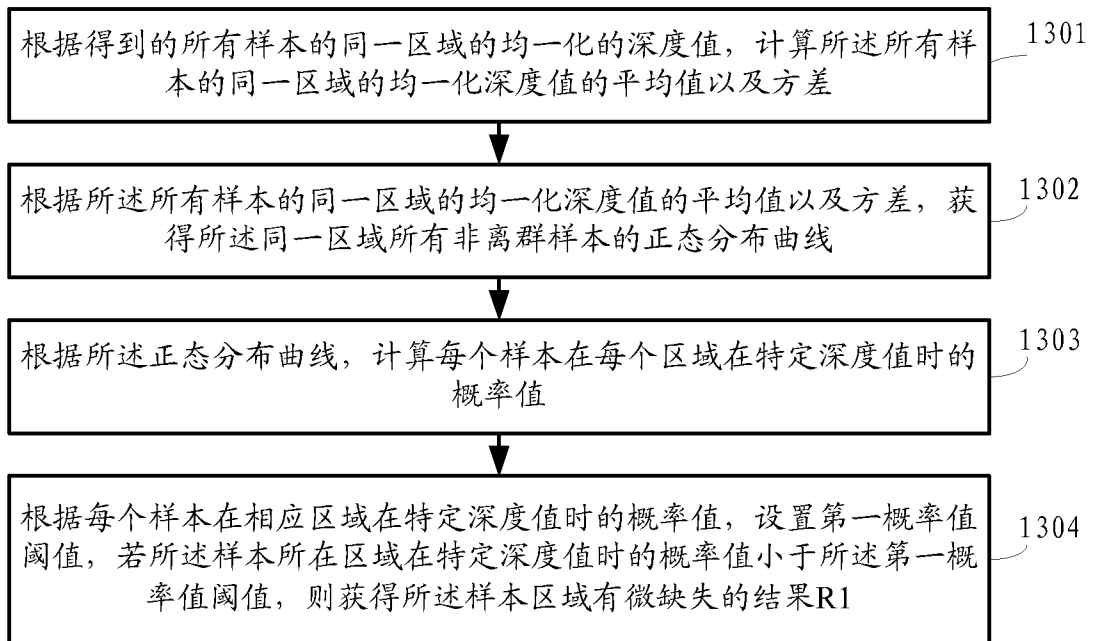


图 13

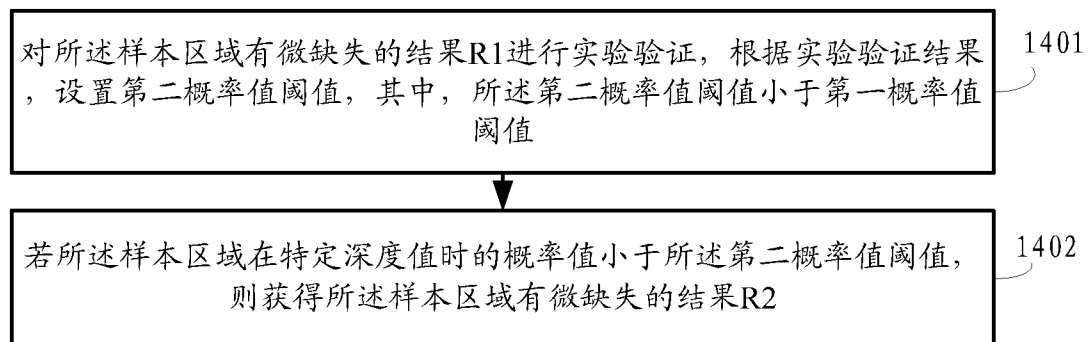


图 14

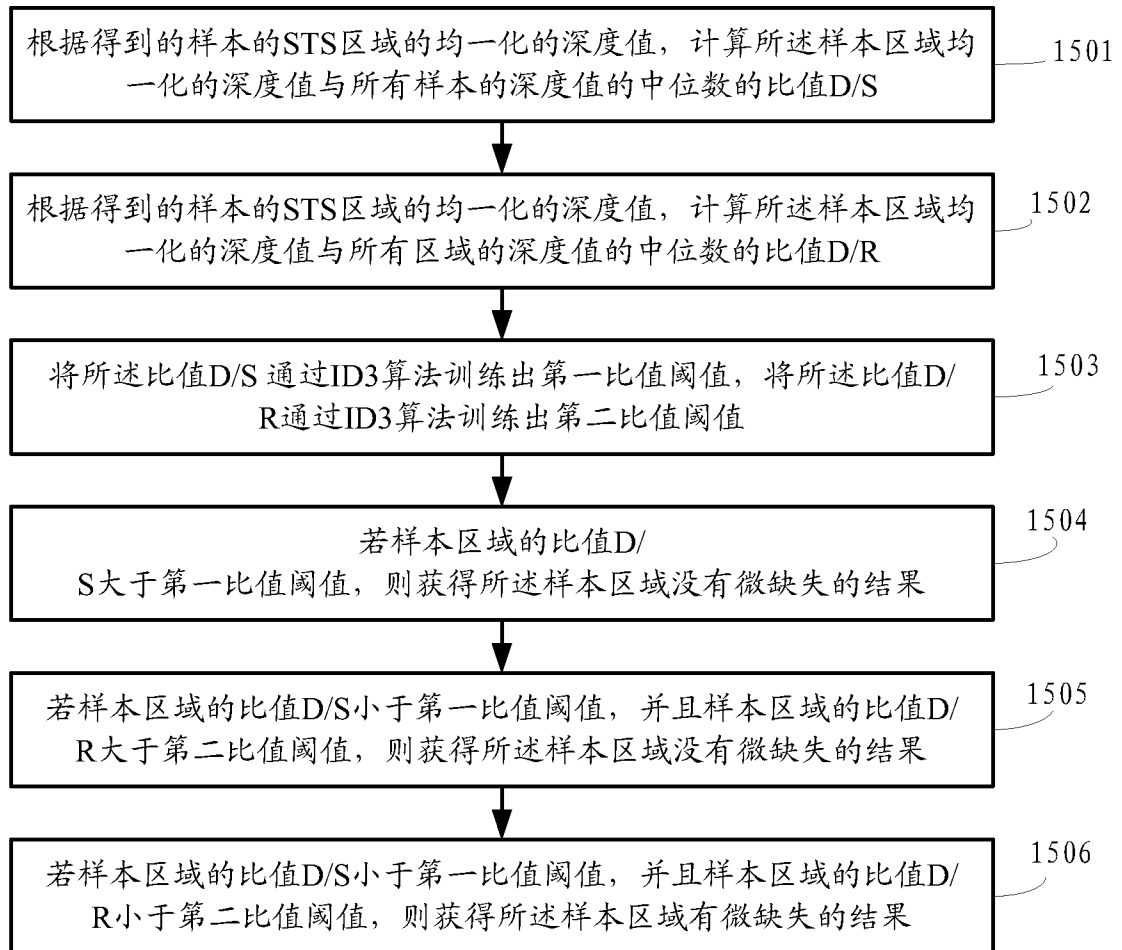


图 15

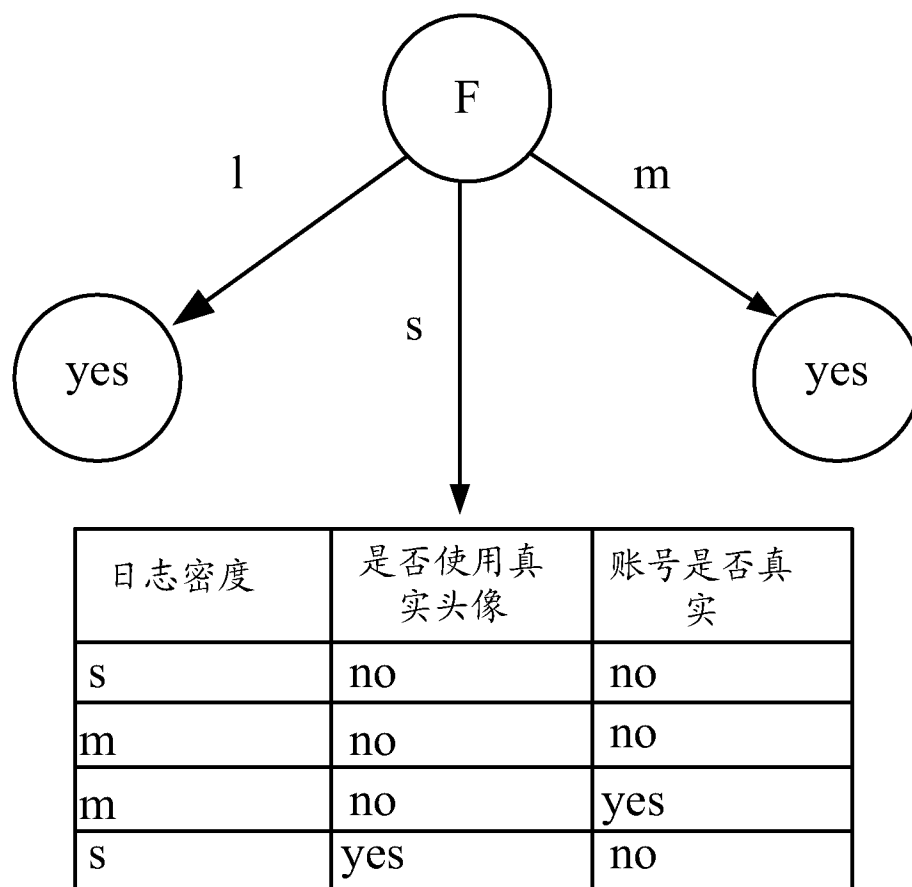


图 16

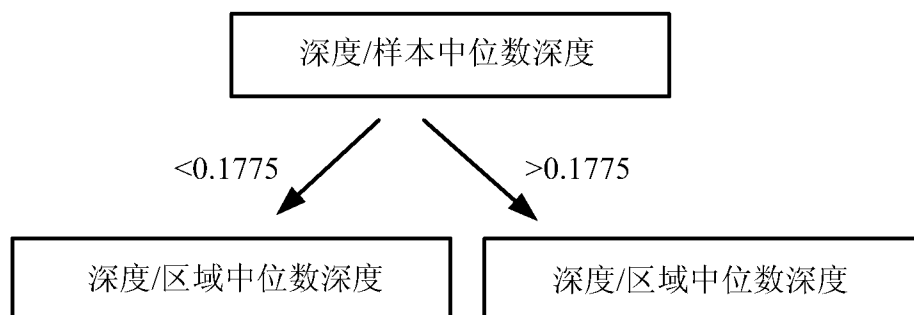


图 17

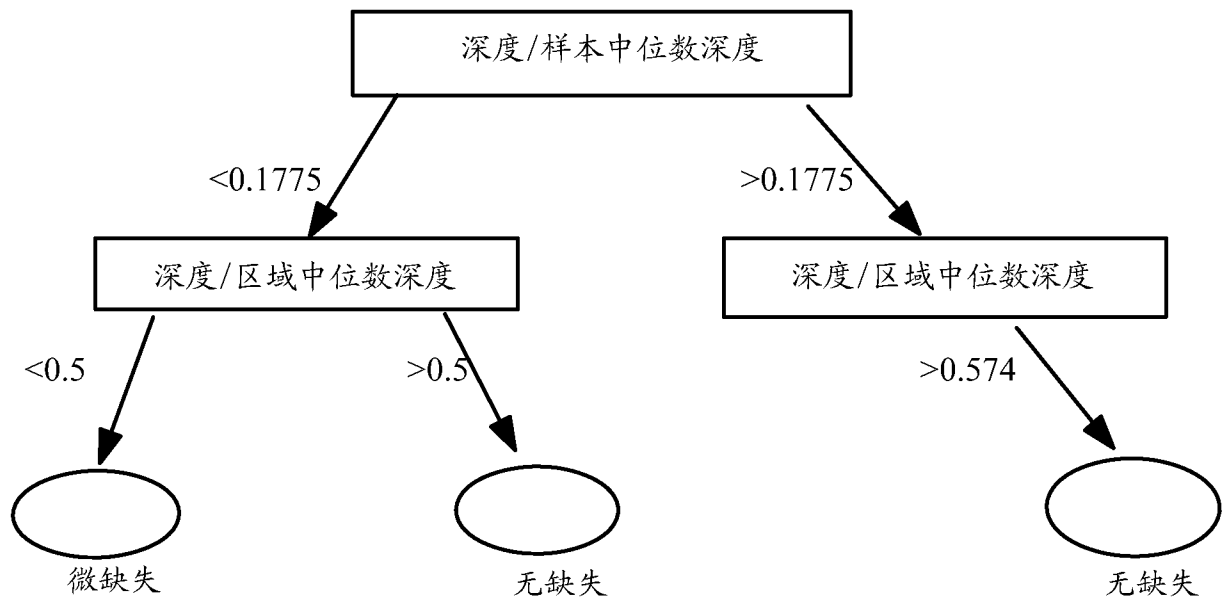


图 18

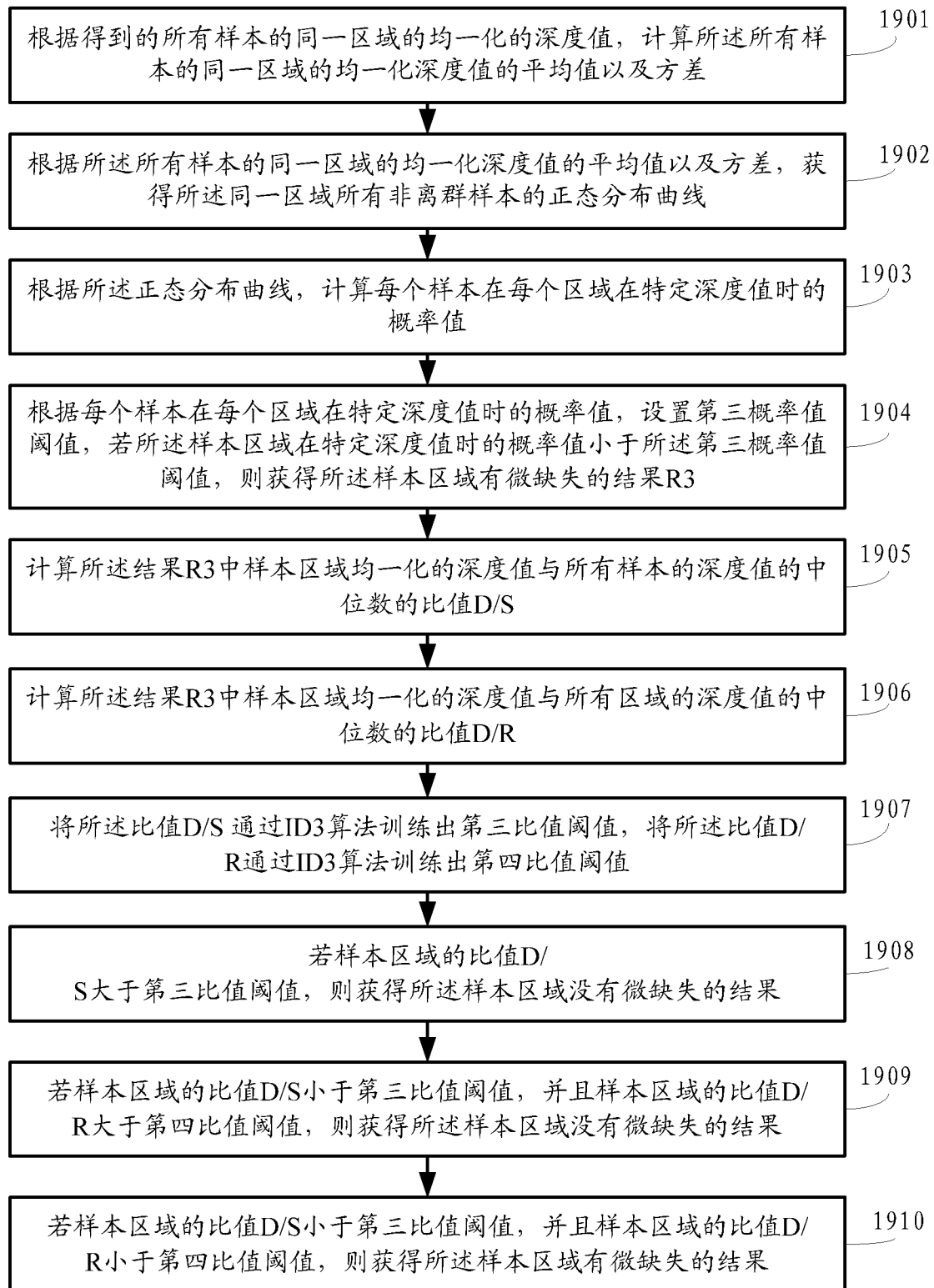


图 19

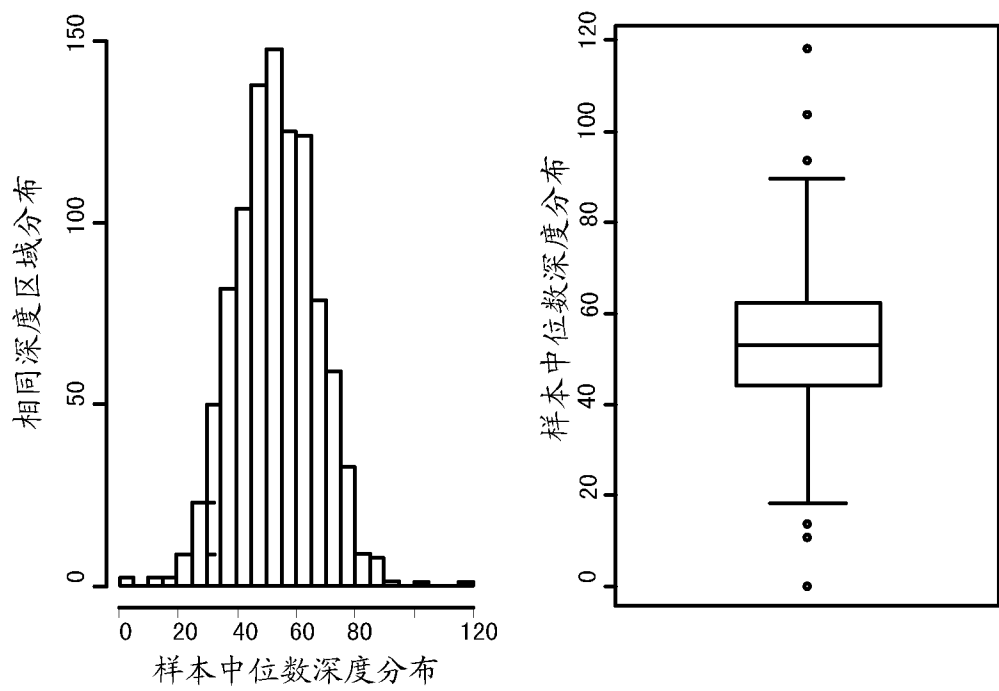


图 20

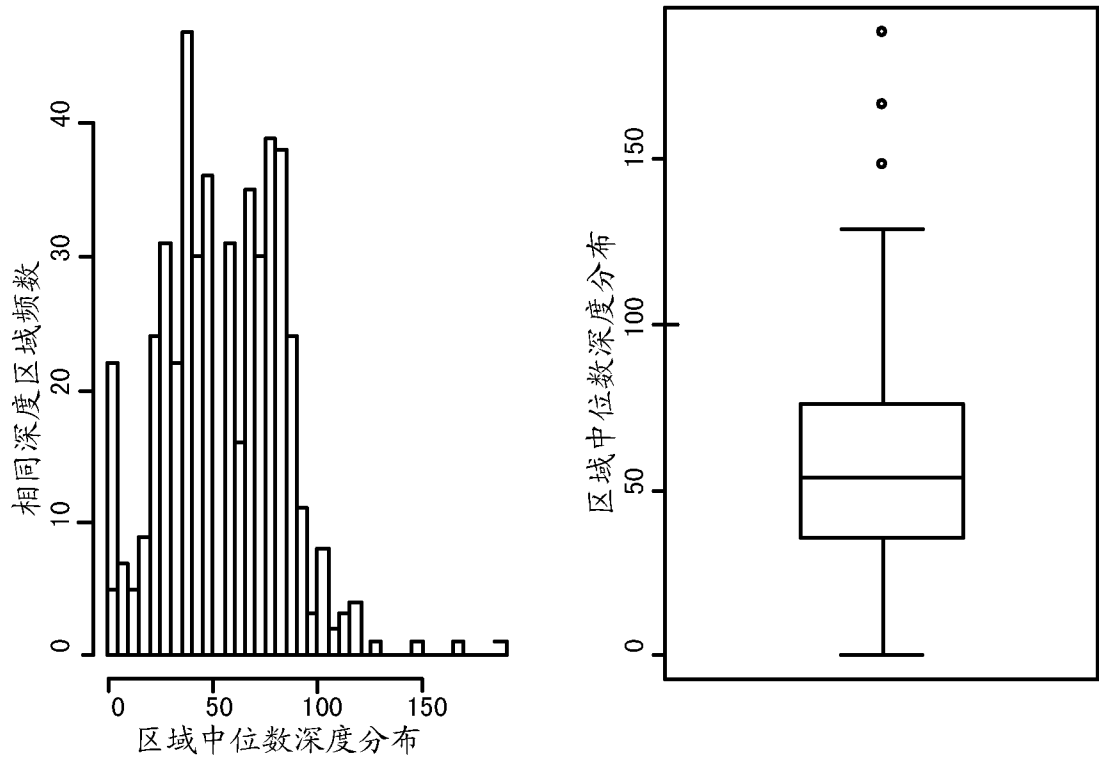


图 21

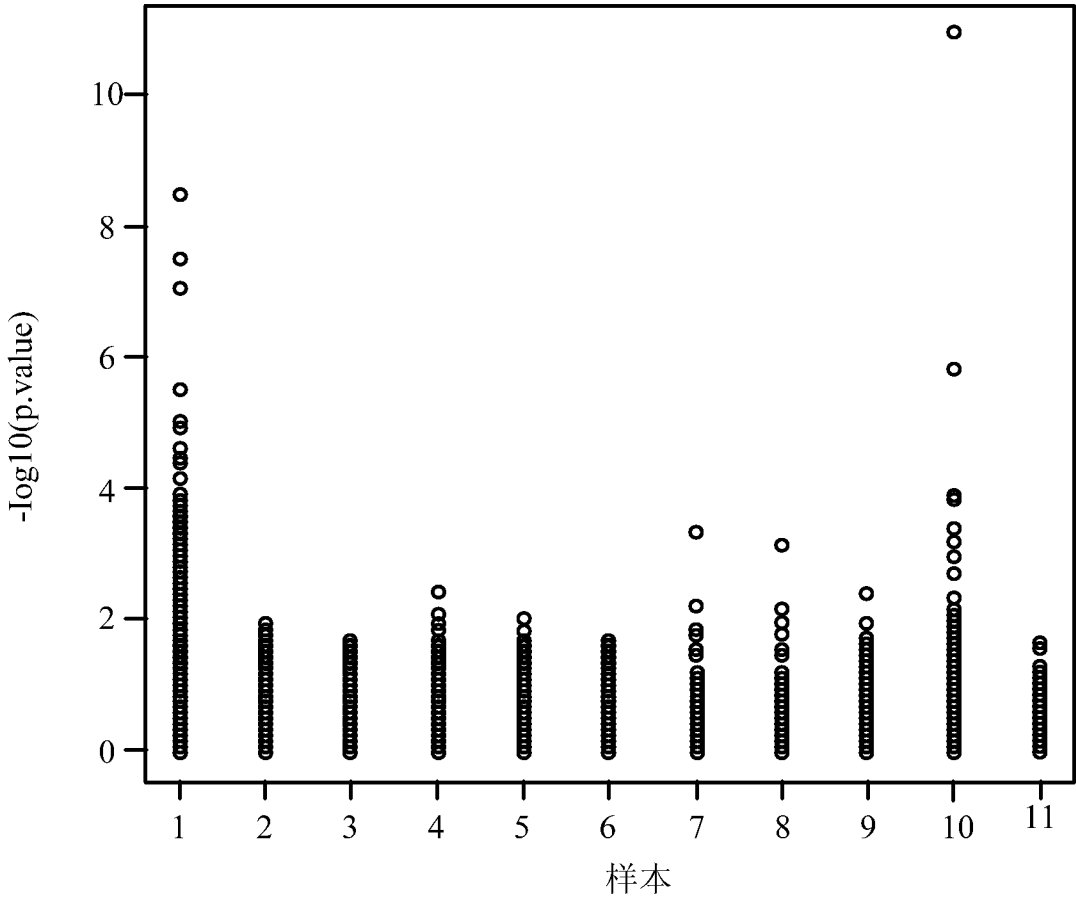


图 22

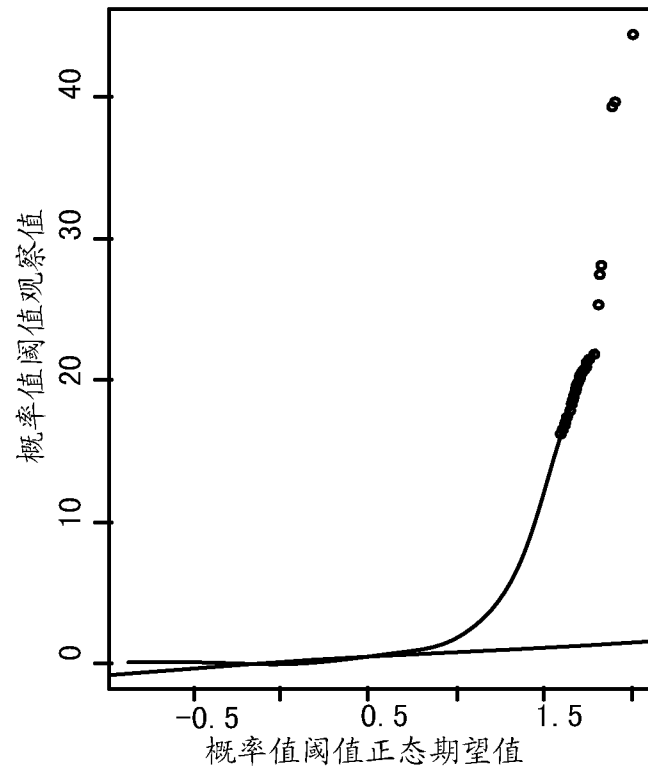


图 23

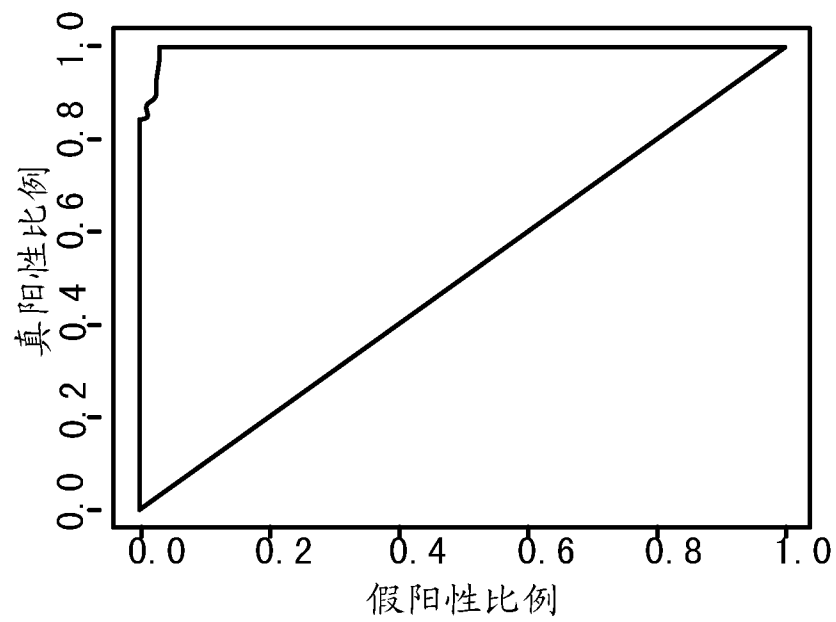


图 24



图 25

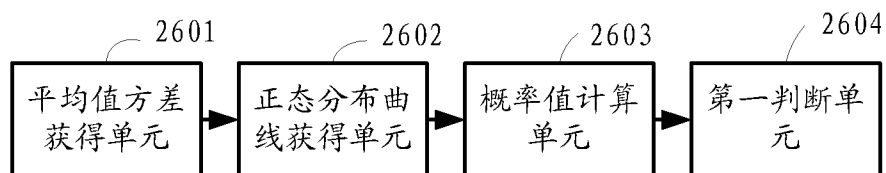


图 26

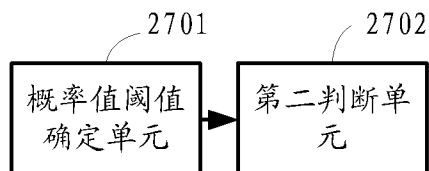


图 27

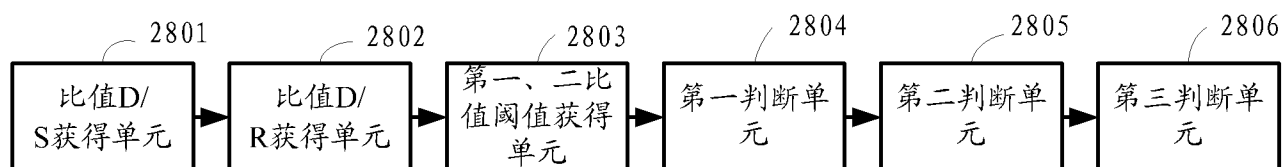


图 28

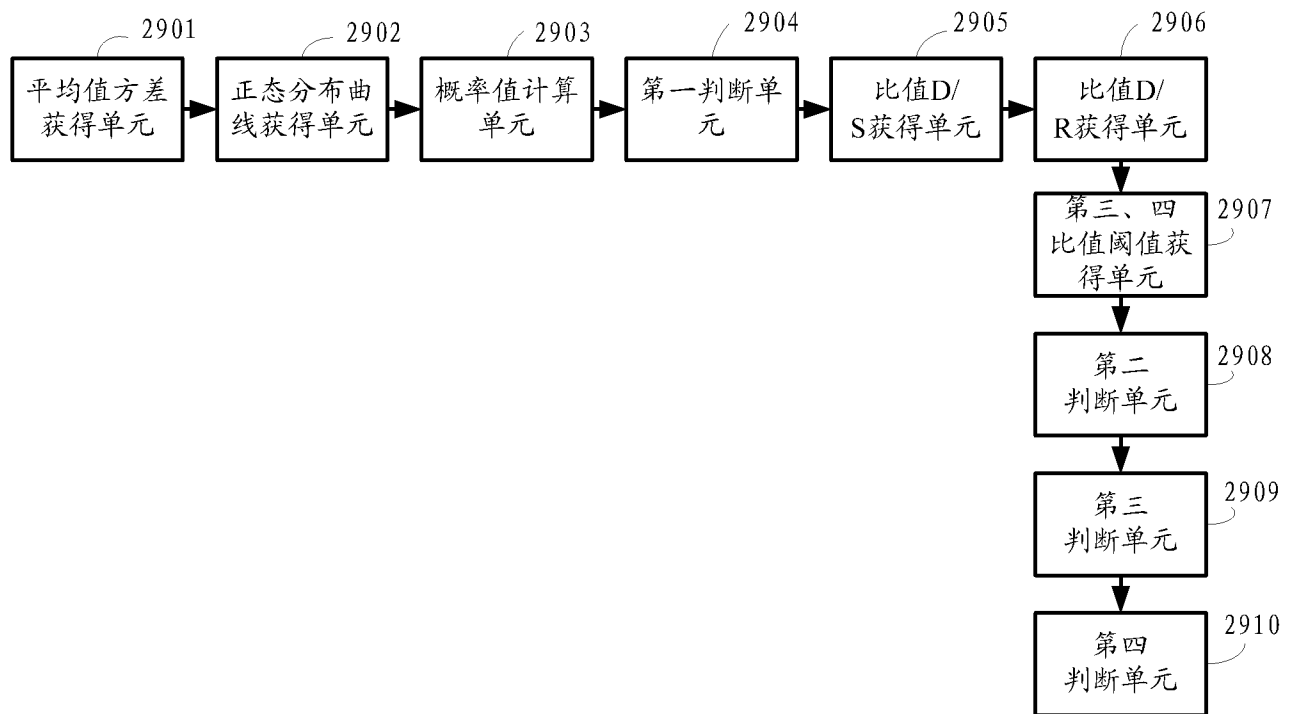


图 29

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/071648

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC: C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

DWPI, SIPOABS, CPRSABS, CNKI, ISI Web of Knowledge:
STS, Sequence-Tagged Site, Sequence Tag Site, microdeletion, probe?

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	KR 20090112235 A (PARK-I), 28 October 2009 (28.10.2009), see the abstract.	1-9, 11-26 10
Y	Title: "Meta-Genomics DNA sequencing", publication date: 20 December 2009 (20.12.2009), URL: http://biolab.blogbus.com/logs/54461787.html	10

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier application or patent but published on or after the international filing date	"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&" document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means	
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 23 November 2012 (23.11.2012)	Date of mailing of the international search report 13 December 2012 (13.12.2012)
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer LU, Chun Telephone No. (86-10) 62411048

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2012/071648

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
KR 20090112235 A	28.10.2009	None	

国际检索报告

国际申请号

PCT/CN2012/071648

A. 主题的分类

C12Q 1/68 (2006.01) i

按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: C12Q

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

DWPI, SIPOABS, CPRSABS, CNKI, ISI Web of Knowledge:

序列标签位点, 序列标记位点, 序列靶位点, 微缺失, 探针, STS, Sequence-Tagged Site, Sequence Tag Site, microdeletion, probe?

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
X	KR20090112235A (PARK-I), 28.10 月 2009 (28.10.2009), 见摘要。	1-9, 11-26
Y		10
Y	文章题目: “Meta-Genomics DNA 测序”, 发表时间: 20.12 月 2009 (20.12.2009), 获自网址: http://biolab.blogbus.com/logs/54461787.html	10



其余文件在 C 栏的续页中列出。



见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

23.11 月 2012 (23.11.2012)

国际检索报告邮寄日期

13.12 月 2012 (13.12.2012)

ISA/CN 的名称和邮寄地址:

中华人民共和国国家知识产权局

中国北京市海淀区蓟门桥西土城路 6 号 100088

传真号: (86-10)62019451

受权官员

陆纯

电话号码: (86-10) **62411048**

国际检索报告

关于同族专利的信息

国际申请号

PCT/CN2012/071648

检索报告中引用的 专利文件	公布日期	同族专利	公布日期
KR20090112235A	28.10.2009	无	