

确定样本基因组中是否存在拷贝数变异的方法、系统和计算机可读介质

优先权信息

无

5

技术领域

本发明涉及确定样本基因组中是否存在拷贝数变异的方法和适于执行该方法的系统和计算机可读介质。

10 背景技术

在科学研究及应用领域,时常遇到需要对单个细胞或几个细胞,或微量核酸样本进行分析的问题,比如在辅助生殖技术领域的植入前诊断(PGD)和植入前筛查(PGS),涉及对单个生殖细胞或单个卵裂球细胞或胚胎细胞进行分析;无创产前诊断技术领域,涉及通过母体外周血中微量的胎儿细胞进行检测的问题;在宏基因组学中,对环境中单个或微量的生物细胞进行分析;以及在疾病或生理研究中,涉及对组织或体液中单个细胞进行分析。

15

然而,目前确定拷贝数变异的方法仍有待改进。

发明内容

本发明旨在至少解决现有技术中存在的技术问题之一。

20

根据本发明的一个方面,本发明提出了一种确定样本基因组中是否存在拷贝数变异的方法。根据本发明的实施例,该方法包括以下步骤:对所述样本基因组进行测序,以便获得由多个测序序列构成的测序结果;将所述测序结果与参照基因组序列进行比对,以便确定所述测序序列在所述参照基因组序列上的分布;基于所述测序序列在参照基因组序列上的分布,在所述参照基因组序列上确定多个突破点,所述突破点两侧的测序序列数目存在显著差异;基于所述多个突破点,在所述参照基因组上确定检验窗口;基于落入所述检验窗口的测序序列,确定第一参数;以及基于所述第一参数与预定阈值的差异,确定所述样本基因组,针对所述检验窗口是否存在拷贝数变异。利用根据本发明实施例的确定样本基因组中是否存在拷贝数变异的方法,能够有效地确定样本基因组中是否存在拷贝数变异,并且适用于各种拷贝数变异,包括但不限于染色体非整倍性、染色体片段缺失、染色体片段增加、微缺失、微重复。

25

30

根据本发明的第二方面,本发明提出了一种确定样本基因组中是否存在拷贝数变异的系统。根据本发明的实施例,该系统包括:测序装置,所述测序装置用于对所述样本基因组进行测序,以便获得由多个测序序列构成的测序结果;分析装置,所述分析装置与所述测序装置相连,以便基于所述测序结果确定所述基因组中是否存在拷贝数变异,所述分析装置进一步包括:比对单元,所述比对单元适于将所述测序结果与参照基因组序列进行比对,以便确定所述测序序列在所述参照基因组序列上的分布;突破点确定单元,所述突破点确定单元与所述比对单元相连,并且适于基于所述测序序列在参照基因组序列上的分布,在所述参照基因组序列上确定多个突破点,所述突破点两侧的测序序列数目存在显著差异;检验窗口确定单元,所述检验窗口确定单元与所述突破点确定单元相连,并且适于基于所述多个突破点,在所述参照基因组上确定检验窗口;参数确定单元,所述参数确定单元与所述检验窗口确定单元相连,并且适于基于落入所述检验窗口的测序序列,确定第一参数;以及判断单元,所述判断单元与所述参数确定单元相连,并且适于基于所述第一参数与预定阈值的差异,确定所述样本基因组,针对所述检验窗口是否存在拷贝数变异。利用根据本发明实施例的确定样本基因组中是否存在拷贝数变异的系统,能够有效地实施根据本发明实施例的确定样本基因组中是否存在拷贝数变异的方法,从而能够有效地确定样本基因组中是否存在拷贝数变异,并且适用于各种拷贝数变异,包括但不限于染色体非整倍性、染色体片段缺失、染色体片段增加、微缺失、微重复。

35

40

45

根据本发明的第三方面,本发明提出了一种计算机可读介质。根据本发明的实施例,该计算机可读介质上存储有指令,所述指令适于被处理器执行以便通过下列步骤确定样本基因组中是否存在拷贝数变异:将测序结果与参照基因组序列进行比对,以便确定所述测序序列在所述参照基因组序列上的分布,其中所述测序结果是由通过对所述样本基因组进行测序所获得的多个测序序列构成的;基于所述测序序列在参照基因组序列上的分布,在所述参照基因组序列上确定多个突破点,所述突破点两侧的测序序列数目存在显著差异;基于所述多个突破点,在所述参照基因组上确定检验窗口;基于落入所述检验窗口

的测序序列，确定第一参数；以及基于所述第一参数与预定阈值的差异，确定所述样本基因组，针对所述检验窗口是否存在拷贝数变异。借助该计算机可读介质，能够有效地实施根据本发明实施例的确定样本基因组中是否存在拷贝数变异的方法，从而能够有效地确定样本基因组中是否存在拷贝数变异，并且适用于各种拷贝数变异，包括但不限于染色体非整倍性、染色体片段缺失、染色体片段增加、微缺失、微重复。

本发明的附加方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

本发明的上述和/或附加的方面和优点从结合下面附图对实施例的描述中将变得明显和容易理解，其中：

图 1 显示了根据本发明一个实施例的确定样本基因组中是否存在拷贝数变异的方法的流程示意图；

图 2 显示了根据本发明一个实施例的确定样本基因组中是否存在拷贝数变异的系统的结构示意图；

图 3 显示了根据本发明又一个实施例的确定样本基因组中是否存在拷贝数变异的方法的流程示意图；

图 4 显示了根据本发明的实施例，样品 S1 的染色体数字核型图。左图为胚胎单细胞经全基因组扩增后用本发明进行拷贝数变异检测的结果，右图为相同的胚胎提取 DNA 后直接测序(未经 WGA)的结果；以及

图 5 显示了根据本发明的实施例，样品 S2 染色体数字核型图。左图为胚胎单细胞经全基因组扩增后用本发明进行拷贝数变异检测的结果，右图为相同的胚胎提取 DNA 后直接测序(未经 WGA)的结果。

发明详细描述

下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能理解为对本发明的限制。

需要说明的是，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。进一步地，在本发明的描述中，除非另有说明，“多个”的含义是两个或两个以上。如果没有明确说明，在本文的公式或标识中，相同的字母代表相同的含义。

一、确定样本基因组中是否存在拷贝数变异的方法

根据本发明的一个方面，本发明提出了一种确定样本基因组中是否存在拷贝数变异的方法。在本发明中所使用的术语“拷贝数变异 (copy number variation, CNV)”的含义是染色体或染色体片段拷贝数的异常，包括但不限于染色体非整倍性、染色体片段缺失、染色体片段增加、微缺失、微重复。

参考图 1，根据本发明实施例的确定样本基因组中是否存在拷贝数变异的方法包括：

S100: 对样本基因组进行测序，以便获得由多个测序序列构成的测序结果

根据本发明的实施例，本发明的方法可以采用的样本基因组的类型并不受特别限制，既可以是全基因组，也可以是基因组的一部分，例如可以是染色体或其片段。另外，根据本发明的实施例，在对样本基因组进行测序之前，进一步包括从生物样本中提取样本基因组的步骤。由此，能够直接以生物样本作为原材料，获得关于该生物样本是否具有拷贝数变异的信息，从而反映生物体的健康状态。根据本发明的实施例，可以采用的生物样本并不受特别限制。根据本发明的一些具体示例，可以采用的生物样本为选自血液、尿液、唾液、组织、生殖细胞、受精卵、卵裂球和胚胎的任意一种。本领域技术人员能够理解的是，针对不同的疾病，可以采用不同的生物样本来进行分析。由此，可以方便地从生物体获取这些样本，并且能够具体地针对某些疾病采取不同的样本，从而针对某些特殊疾病采取特定的分析手段。例如，对于可能罹患特定癌症的测试对象，可以从该组织或其附近采集样本，并进一步分离细胞进行分析，由此，能够精确并且尽可能早地获知该组织是否发生癌变。根据本发明的具体实施例，可以采用单细胞作为生物样本。根据本发明的实施例，从生物样本分离单细胞的方法和设备不受特别限制。根据本发明的一些具体示例，可以采用选自稀释法、口吸管分离法、显微操作（优选显微切割）、流式细胞分离术、微流控法的至少一种从生物样本分离单细胞。由此，能够有效便捷地获得生物样本的单细胞，以便实施

后续操作,由此,可以进一步提高确定样本基因组中是否存在拷贝数变异的效率。

另外,根据本发明的实施例,对样本基因组进行测序的方法不受特别限制。根据本发明的一个实施例,对样本基因组进行测序进一步包括:首先,对样本基因组进行扩增得到经过扩增的基因组;接下来,利用经过扩增的样本基因组构建测序文库;最后,对所得到的测序文库进行测序,以便获得由多个测序序列构成的测序结果。由此,能够有效地获取样本基因组的测序结果的全基因组信息,并且能够对单细胞基因组或者微量核酸样本进行有效测序,从而进一步提高了确定样本基因组中是否存在拷贝数变异的效率。本领域技术人员可以根据采用的基因组测序技术的具体方案选择不同的构建测序文库的方法,关于构建基因组测序文库的细节,可以参见测序仪器的厂商例如 Illumina 公司所提供的规程,例如参见 Illumina 公司 Multiplexing Sample Preparation Guide (Part#1005361; Feb 2010) 或 Paired-End SamplePrep Guide (Part#1005063; Feb 2010),通过参照将其并入本文。

任选地,对于以单细胞作为生物样本提取基因组,根据本发明的实施例,可以进一步包括对单细胞进行裂解,以便释放所述单细胞的全基因组的步骤。根据本发明的一些示例,可以用于裂解单细胞并释放全基因组的方法不受特别限制,只要能够将单细胞裂解优选充分裂解即可。根据本发明的具体示例,可以利用碱性裂解液将所述单细胞裂解并释放所述单细胞的全基因组。发明人发现,这样能够有效地裂解单细胞并释放出全基因组,并且所释放的全基因组在进行测序时,能够提高准确率,从而进一步提高了确定样本基因组中是否存在拷贝数变异的效率。根据本发明的实施例,单细胞全基因组扩增的方法不受特别限制,可以采用基于 PCR 的方法例如可以采用 PEP-PCR、DOP-PCR、和 OmniPlex WGA,也可以采用非基于 PCR 的方法例如 MDA (多重链置换扩增)。根据本发明的具体示例,优选采用基于 PCR 的方法,例如 OmniPlex WGA 方法。可选用的商业化试剂盒包括但不限于 Sigma Aldrich 的 GenomePlex, Rubicon Genomics 的 PicoPlex, Qiagen 的 REPLI-g, GE Healthcare 的 illustra GenomiPhi 等。根据本发明的具体示例,在构建测序文库之前,可以采用 OmniPlex WGA 对单细胞全基因组进行扩增。由此,能够有效地对全基因组进行扩增,从而进一步提高了确定样本基因组中是否存在拷贝数变异的效率。根据本发明的实施例,可以采用选自第二代测序技术如 Illumina 公司的 Hiseq 系统, Miseq 系统, Genome Analyzer(GA)系统, Roche 公司的 454 FLX, Applied Biosystems 公司的 SOLiD 系统, Life Technologies 公司的 Ion Torrent 系统等至少一种对所述全基因组测序文库进行测序。由此,能够利用这些测序装置的高通量、深度测序的特点,进一步提高了确定单细胞染色体非整倍性的效率。当然,本领域技术人员能够理解的是,还可以采用其他的测序方法和装置进行全基因组测序,例如第三代测序技术即单分子测序技术如 Helicos BioSciences 公司的 HeliScope 系统, PacBio 公司的 RS 系统等任一种,以及以后可能开发出来的更先进的测序技术。根据本发明的实施例,通过全基因组测序所得到的测序数据的长度不受特别限制。根据本发明的一个具体示例,所述多个测序数据的平均长度为约 50bp。申请人惊奇地发现,当测序数据的平均长度为约 50bp 时,能够极大地方便对测序数据进行分析,提高分析效率,同时能够显著降低分析的成本。进一步提高了确定单细胞染色体非整倍性的效率,并且降低了确定单细胞染色体非整倍性的成本。这里所使用的术语“平均长度”是指各个测序数据长度数值的平均值。

S200: 将测序结果与参照基因组序列进行比对,以便确定测序序列在参照基因组序列上的分布

在完成对样本基因组进行测序之后,所得到的测序结果中包含了多个测序序列。将所得到的测序结果与参照基因组序列进行比对,从而可以确定所得到的测序序列在参照基因组序列上的定位。根据本发明的实施例,可以采用任何已知的方法对这些测序数据的总数目进行计算。例如,可以采用测序仪器的制造商所提供的软件进行分析。优选采用短寡核苷酸分析包 (Short Oligonucleotide Analysis Package, SOAP) 和 BWA 比对 (Burrows-Wheeler Aligner) 进行,将测序序列与参考基因组序列比对,得到测序序列在参考基因组上的位置。进行序列比对可以使用程序提供的默认参数进行,或者由本领域技术人员根据需要对参数进行选择。在本发明的一个实施方案中,所采用的比对软件是 SOAPaligner/soap2。

根据本发明的实施例,参照基因组序列是 NCBI 数据库中的标准人类基因组参考序列 (例如可以为 hg18, NCBI Build 36)。也可以是已知基因组序列的一部分,例如可以为选自人类 21 号染色体、18 号染色体、13 号染色体、X 染色体和 Y 染色体的至少一种的序列。

根据本发明的实施例,通过将测序结果与参照基因组序列进行比对,可以选择与参照基因组序列唯一比对的序列,进行后续分析,由此,能够避免重复序列对拷贝数变异分析的干扰,进一步提高确定样

本基因组中是否存在拷贝数变异的效率。

S300: 基于测序序列在参照基因组序列上的分布, 在参照基因组序列上确定多个突破点

在本文中所使用的术语“突破点”指的是基因组上这样一种位点, 在该位点两侧相同的区段之间测序序列数目存在显著差异。因为测序序列 (reads) 是来源于样本基因组的, 因而当样本基因组中特定区域出现拷贝数变异时, 与该区域对应的测序序列的数目也会有显著变化。由此, 在确定多个突破点后, 可以初步判断两个相邻突破点之间的区段可能存在拷贝数变异。

根据本发明的实施例, 在参照基因组上确定突破点可以进一步包括:

首先, 将参考基因组序列划分为多个预定长度的一级窗口, 并确定落入各一级窗口中的测序序列。根据本发明的具体实例, 可以通过常规的比对程序, 将所得到的测序结果中所包含的测序序列与参照基因组序列进行比对, 从而确定落入各一级窗口中的测序序列。例如可以在前面描述的 S200 步骤中完成。根据本发明的具体实例, 落入各一级窗口中的测序序列为唯一比对测序序列。由此, 能够避免重复序列对拷贝数变异分析的干扰, 进一步提高确定样本基因组中是否存在拷贝数变异的效率。

接下来, 针对参考基因组序列上的至少一个位点, 确定落入位点两侧相同数目一级窗口中的测序序列数目。根据本发明的实施例, 可以对参考基因组序列上的所有位点进行相关分析, 也可以对感兴趣的染色体, 例如人类 21 号染色体、18 号染色体、13 号染色体、X 染色体和 Y 染色体的至少一种上的所有位点进行该分析。根据本发明的实施例, 各一级窗口的长度可以相同或者不同, 并且一级窗口之间可以有重叠, 只要各个一级窗口的信息是已知的即可, 优选各一级窗口具有相同的长度。根据本发明的实施例, 各一级窗口的长度可以均为 100-200Kbp, 优选 150Kbp。根据本发明的实施例, 在位点两侧选择的一级窗口的数目并不受特别限制, 根据具体实例, 可以在位点两侧各取 100 个一级窗口。

接下来, 可以通过统计分析, 确定所研究位点两侧测序序列数据分布的 p 值, 该 p 值可以反映两侧测序数据数目的显著差异性。如果所述位点的 p 值小于终止 p 值, 判断该位点为突破点。根据本发明的实施例, 终止 p 值的范围可以通过对已知序列的样本进行平行分析而确定的, 根据本发明的一个具体实例, 终止 p 值为 1.1×10^{-50} 。

根据本发明的一个实施例, 确定位点两侧测序数据数目的显著差异性 p 值进一步包括:

针对选定的位点, 在该位点两侧各取相同数目的一级窗口, 并且计算每个一级窗口的相对测序序列数目 R_i , 其中 i 表示一级窗口的编号,

对所有一级窗口的相对测序序列数目 R_i 进行游程检验, 以便确定该位点的 p 值, 其中,

所述相对测序序列数目是通过下列公式确定的:

$$R_i = \log_2 \left(\frac{r_i}{\bar{r}} \right)$$

其中 r_i 表示落入第 i 一级窗口的测序序列数目,

$$\bar{r} = \frac{1}{n} \sum_{i=1}^n r_i, \quad n \text{ 表示一级窗口的总数目。}$$

具体地, 对所有一级窗口的相对测序序列数目进行游程检验进一步包括: 对每个一级窗口的相对测序序列数目 R_i 进行 GC 含量校正, 以便获得校正的相对测序序列数目 $\tilde{R}_i$; 基于校正的相对测序序列数目, 确定每个一级窗口的标准化的测序序列数目 Z_i ; 以及对所有一级窗口的标准化的测序序列数目 Z_i 进行游程检验。更具体地, 所述校正的相对测序序列数目 $\tilde{R}_i$ 是通过下列步骤获得的:

首先, 计算每个一级窗口的 GC 含量;

接下来, 将 GC 含量以预定数值为单位划分为多个区域, 并且统计每个区域中相对测序序列数目的平均值 M_s , 其中 s 为 GC 区域的编号, 根据本发明的实施例, 预定数值可以为在 0.0005-0.01 范围内的任意数值, 对应的区域大小为 50k-300k, 优选采用 0.001, 由此进行矫正的力度(power)最佳;

接下来, 根据下列公式确定所述校正的相对测序序列数目 $\tilde{R}_i$:

$$\tilde{R}_i = R_i - M_s;$$

5 最后, 通过下列公式确定标准化的测序序列数目 Z_i

$$Z_i = (R_i - \tilde{R}_i - mean) / SD$$

其中,

$$mean = \frac{1}{n} \sum_{i=1}^n (R_i - \tilde{R}_i)$$

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (R_i - \tilde{R}_i - mean)^2}。$$

由此, 可以通过 GC 含量对测序序列数目进行校验。由此, 可以消除基因组扩增的偏好所造成的干扰, 从而进一步提高确定样本基因组中是否存在拷贝数变异的精确性和效率。

10 在确定多个突破点后, 可以初步判断两个相邻突破点之间的区段可能存在拷贝数变异。因而可以将这些区段作为检验窗口, 用于进一步判断是否存在拷贝数变异。对于初步判断得到的突破点较多的情形, 可以对突破点进一步进行筛选。由此, 根据本发明的实施例, 基于多个突破点, 在参照基因组上确定检验窗口进一步包括:

- 1) 确定多个候选突破点, 候选突破点是指这样一种突破点, 在该突破点的前后均存在其他突破点;
- 2) 确定每个候选突破点的 p 值, 并剔除 p 值最大的候选突破点;
- 15 3) 对剩余的候选突破点重复步骤 2), 在剔除一部分候选突破点, 直到剩余候选突破点的 p 值均小于终止 p 值, 剩余候选突破点作为经过筛选的候选突破点; 以及
- 4) 确定相邻两个经过筛选的候选突破点之间的区域为检验窗口。

根据本发明的实施例, 可以通过下列步骤确定候选突破点的 p 值:

20 将候选突破点与相邻的上一个突破点之间的区域作为第一候选区域, 将所述候选突破点与相邻的下一个突破点之间的区域作为第二候选区域;

对第一候选区域和第二候选区域中所包含一级窗口的标准化的测序序列数目 Z_i 进行游程检验 (游程检验是一种非参数检验, 利用两个群体元素混合后的分布均匀状态评价此两个群体的差异显著性, 关于该检验的细节, 可以参见 Wald A. WJ. On a Test Whether Two Samples are from the Same Population. The Annals of Mathematical Statistics 1940; 11:147-162, 通过参照将其并入本文), 以便确定所述候选突破点的 p 值。

25 根据本发明的实施例, 终止 p 值是通过下列步骤确定的:

利用对样品品的测序结果, 重复在参照基因组上确定检验窗口的操作, 并记录每次被剔除候选突破点的 p 值, 直到候选突破点的数目为零, 这里所使用的术语对照样品指的是已知核酸序列中不存在拷贝数变异的样品; 以及

30 基于被剔除候选突破点的 p 值分布, 确定所述终止 p 值, 例如将被剔除的候选突破点的 p 值做分布图, 选取 p 值变化趋势最大的地方作为终止 p 值 (p_{final})。

根据本发明的具体示例, 终止 p 值可以为 1.1×10^{-50} 。

S400: 基于落入检验窗口的测序序列, 确定第一参数

在确定检验窗口后，可以通过对检验窗口中所包含的测序序列进行统计分析，从而确定该检验窗口是否存在拷贝数变异。根据本发明的一个实施例，基于落入所述检验窗口的测序序列，确定第一参数进一步包括：确定检验窗口中所包含的所有一级窗口的平均标准化的测序序列数目 $\bar{Z}$ ，该平均标准化的测序序列数目 $\bar{Z}$ 作为第一参数。关于标准化的测序序列数目，前面已经进行了详细描述，此处不再赘述。

5 S500: 基于第一参数与预定阈值的差异，确定样本基因组针对该检验窗口是否存在拷贝数变异

根据本发明的实施例，可以通过将前面所确定的第一参数与预定阈值进行比较，基于第一参数和预定阈值之间的差异，来确定关于特定的检验窗口，基因组样本是否具有拷贝数变异。基于基因组测序的测序结果中，针对某一特定窗口的测序序列的数目，是与全基因组中该窗口在染色体或基因组中的含量呈正相关的，因而，通过对测序结果中来源于某一特定窗口的测序序列进行统计分析，能够有效地确定关于该窗口，样本基因组是否具有拷贝数变异。在本文中所使用的术语“预定阈值”是指将已知基因组正常的样本基因组重复针对上述实施的操作和分析所得到的关于特定窗口的相关数据。本领域技术人员能够理解的是，可以采用相同的测序条件和数学运算方法，分别获得特定窗口的相关参数，以及正常细胞的相关参数。这里，可以将正常细胞的相关参数作为预定阈值。另外，本文中所使用的术语“预定”，应做广义理解，可以是预先通过实验确定的，也可以是在进行生物样本分析时，采用平行实验获得的。这里所使用的术语“平行实验”应作广义理解，既可以指的是同时进行未知样品和已知样品的测序和分析，也可以是先后进行在相同条件下的测序和分析。根据本发明的实施例，所述预定阈值采用第一阈值和第二阈值，通过将第一参数 $\bar{Z}$ 与第一阈值和第二阈值相比较，小于第一阈值为拷贝数减少（即缺失），大于第二阈值为拷贝数增加（即重复），由此可以确定拷贝数变异类型。根据本发明的具体示例，设定 $\alpha = 0.05$ 为显著界线，第一阈值为 -1.645，第二阈值为 1.645，由此，可以进一步有效地确定拷贝数变异类型。

利用根据本发明实施例的确定样本基因组中是否存在拷贝数变异的方法，能够有效地确定样本基因组中是否存在拷贝数变异，并且适用于各种拷贝数变异，包括但不限于染色体非整倍性、染色体片段缺失、染色体片段增加、微缺失、微重复。拷贝数变异是引起出生缺陷的主要因素，在体外培养的胚胎中也非常常见，是引起体外生殖失败的主要原因。拷贝数变异也是很多疾病如癌症的致病因素。全基因组扩增是对单个细胞、几个细胞或微量核酸样本进行全基因组范围扩增的技术，可以在尽量保持全基因组代表性前提下使样本量增加，达到所需的样本量。但是，通常而言，全基因组扩增都存在扩增偏向性的问题，有可能给后续分析带来误差。根据本发明实施例的确定样本基因组中是否存在拷贝数变异的方法，在对单细胞或微量核酸样本经全基因组扩增后，通过测序技术获得数据，进行拷贝数变异的分析，一方面通过全基因组扩增解决了单细胞或微量核酸样本分析难的问题，另一方面避免了全基因组扩增对拷贝数变异分析产生的偏差，使检测更准确更全面，尤其是通过 GC 含量校正，能够更进一步提高检测效率。另外，根据本发明的实施例，在不同样品的文库构建过程中引入不同的标签，由此可以同时多种样品进行检验，进一步提高确定样本基因组中是否存在拷贝数变异的效率。利用根据本发明实施例的确定样本基因组中是否存在拷贝数变异的方法，能够进行胚胎植入前拷贝数变异筛查和诊断或无创胎儿拷贝数变异筛查，有利于提供遗传咨询和临床决策依据；进行产前诊断可有效防止病变胚胎植入，防止患儿出生。

二、确定样本基因组中是否存在拷贝数变异的系统

根据本发明的第二方面，本发明提出了一种确定样本基因组中是否存在拷贝数变异的系统，利用该系统能够有效的实施前述确定样本基因组中是否存在拷贝数变异的方法，从而可以有效的确定样本基因组中是否存在拷贝数变异。

参考图 2，根据本发明的实施例，确定样本基因组中是否存在拷贝数变异的系统 1000 包括：测序装

置 100 和分析装置 200。

根据本发明的实施例，测序装置 100 用于对样本基因组进行测序，以便获得由多个测序序列构成的测序结果。根据本发明的实施例，确定样本基因组中是否存在拷贝数变异的系统 1000 可以进一步包括基因组提取装置（图中未示出），该基因组提取装置适于从生物样本中提取样本基因组，并且该基因组提取装置与测序装置相连以便为测序装置 100 提供样本基因组。由此，能够直接以生物样本作为原材料，获得关于该生物样本是否具有拷贝数变异的信息，从而反映生物体的健康状态。根据本发明的实施例，测序装置可以进一步包括：基因组扩增单元、测序文库构建单元以及测序单元。其中，基因组扩增单元适于对所述样本基因组进行扩增，测序文库构建单元与基因组扩增单元相连，并且适于利用经过扩增的样本基因组构建测序文库；以及测序单元，测序单元与所述测序文库构建单元相连，并且适于对所述测序文库进行测序。根据本发明的实施例，测序单元为选自第二代测序技术如 Illumina 公司的 Hiseq 系统，Miseq 系统，Genome Analyzer(GA)系统，Roche 公司的 454 FLX，Applied Biosystems 公司的 SOLiD 系统，Life Technologies 公司的 Ion Torrent 系统和单分子测序装置的至少一种。由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高了确定单细胞染色体非整倍性的效率。

根据本发明的实施例，分析装置 200 与测序装置 100 相连，以便基于测序结果确定基因组中是否存在拷贝数变异。根据本发明的实施例，分析装置 200 进一步包括：比对单元 201、突破点确定单元 202、检验窗口确定单元 203、参数确定单元 204 以及判断单元 205。其中，比对单元 201 适于将测序结果与参照基因组序列进行比对，以便确定测序序列在参照基因组序列上的分布。根据本发明的实施例，比对单元 201 内存储有参照基因组序列为已知的人类基因组序列，任选地，该参照基因组序列为选自人类 21 号染色体、18 号染色体、13 号染色体、X 染色体和 Y 染色体的至少一种的序列。突破点确定单元 202 与比对单元 201 相连，并且适于基于测序序列在参照基因组序列上的分布，在参照基因组序列上确定多个突破点，如前所述，突破点两侧的测序序列数目存在显著差异。检验窗口确定单元 203 与突破点确定单元 202 相连，并且适于基于这些突破点，在参照基因组上确定检验窗口。参数确定单元 204 与检验窗口确定单元 203 相连，并且适于基于落入检验窗口的测序序列，确定第一参数判断单元 205 与参数确定单元 204 相连，并且适于基于所得到的第一参数与预定阈值的差异，确定样本基因组针对所确定的检验窗口是否存在拷贝数变异。

根据本发明的实施例，突破点确定单元 202 可以进一步包括适于执行下列以确定突破点的模块：

将参考基因组序列划分为多个预定长度的一级窗口，并确定落入各一级窗口中的测序序列；

首先，将参考基因组序列划分为多个预定长度的一级窗口，并确定落入各一级窗口中的测序序列。

根据本发明的具体实例，可以通过常规的比对程序，将所得到的测序结果中所包含的测序序列与参照基因组序列进行比对，从而确定落入各一级窗口中的测序序列。根据本发明的实施例，各一级窗口的长度可以相同或者不同，并且一级窗口之间可以有重叠，只要各个一级窗口的信息是已知的即可，优选各一级窗口具有相同的长度。根据本发明的实施例，各一级窗口的长度可以均为 100-200Kbp，优选 150Kbp。根据本发明的实施例，在位点两侧选择的一级窗口的数目并不受特别限制，根据具体实例，可以在位点两侧各取 100 个一级窗口。

接下来，确定所述位点的 p 值，该 p 值可以反映两侧测序数据数目的显著差异性。以及如果所述位点的 p 值小于终止 p 值，判断该位点为突破点。根据本发明的实施例，终止 p 值的范围可以是通过对于已知序列的样本进行平行分析而确定的，根据本发明的一个具体实例，终止 p 值可以为 1.1×10^{-50} 。

根据本发明的实施例，突破点确定单元 202 可以进一步包括适于执行下列以确定 p 值的模块：

针对选定的位点，在该位点两侧各取相同数目的一级窗口，并且计算每个一级窗口的相对测序序列数目 R_i ，其中 i 表示一级窗口的编号

对所有一级窗口的相对测序序列数目 R_i 进行游程检验，以便确定所述位点的 p 值，

其中，

所述相对测序序列数目是通过公式确定的：

$$R_i = \log_2 \left(\frac{r_i}{\bar{r}} \right)$$

其中 r_i 表示落入第 i 一级窗口的测序序列数目,

$$\bar{r} = \frac{1}{n} \sum_{i=1}^n r_i, \quad n \text{ 表示一级窗口的总数目。}$$

- 5 根据本发明的实施例, 突破点确定单元 202 可以进一步包括适于执行下列以对所有一级窗口的相对测序序列数目进行游程检验的模块:

对每个一级窗口的相对测序序列数目 R_i 进行 GC 含量校正, 以便获得校正的相对测序序列数目 $\tilde{R}_i$;

基于所述校正的相对测序序列数目, 确定每个一级窗口的标准化的测序序列数目 Z_i ; 以及

对所有一级窗口的标准化的测序序列数目 Z_i 进行游程检验。其中, 根据本发明的实施例, 校正的相

- 10 对测序序列数目 $\tilde{R}_i$ 是通过适于执行下列步骤的模块获得的:

计算每个一级窗口的 GC 含量;

将 GC 含量以预定数值为单位划分为多个区域, 并且统计每个区域中相对测序序列数目的平均值 M_s , 其中 s 为 GC 区域的编号, 根据本发明的实施例, 预定数值可以为在 0.0005-0.01 范围内的任意数值, 对应的区域大小为 50k-300k, 优选采用 0.001, 由此进行矫正的力度(power)最佳;

- 15 根据公式下列公式确定所述校正的相对测序序列数目 $\tilde{R}_i$:

$$\tilde{R}_i = R_i - M_s;$$

所述标准化的测序序列数目 Z_i 是通过下列公式确定的

$$Z_i = (R_i - \tilde{R}_i - \text{mean}) / SD$$

其中,

$$\text{mean} = \frac{1}{n} \sum_{i=1}^n (R_i - \tilde{R}_i)$$

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (R_i - \tilde{R}_i - \text{mean})^2}。$$

- 20 在确定多个突破点后, 可以初步判断两个相邻突破点之间的区段可能存在拷贝数变异。因而可以将这些区段作为检验窗口, 用于进一步判断是否存在拷贝数变异。对于初步判断得到的突破点较多的情形, 可以对突破点进一步进行筛选。由此, 根据本发明的实施例, 基于多个突破点, 在参照基因组上确定检验窗口进一步包括适于执行下列的模块:

- 1) 确定多个候选突破点, 候选突破点是指这样一种突破点, 在该突破点的前后均存在其他突破点;
 - 2) 确定每个候选突破点的 p 值, 并剔除 p 值最大的候选突破点;
 - 3) 对剩余的候选突破点重复步骤 2), 在剔除一部分候选突破点, 直到剩余候选突破点的 p 值均小于终止 p 值, 剩余候选突破点作为经过筛选的候选突破点; 以及
 - 4) 确定相邻两个经过筛选的候选突破点之间的区域为检验窗口。
- 25

其中, 根据本发明的实施例, 通过下列步骤确定所述候选突破点的 p 值:

将候选突破点与相邻的上一个突破点之间的区域作为第一候选区域, 将所述候选突破点与相邻的下一个突破点之间的区域作为第二候选区域;

5 对第一候选区域和第二候选区域中所包含一级窗口的标准化的测序序列数目 Z_i 进行游程检验, 以便确定所述候选突破点的 p 值。

根据本发明的实施例, 终止 p 值是通过下列步骤确定的:

利用对照样品的测序结果, 重复在参照基因组上确定检验窗口的操作, 并记录每次被剔除候选突破点的 p 值, 直到候选突破点的数目为零, 这里所使用的术语对样品指的是已知核酸序列中不存在拷贝数变异的样品; 以及

10 基于被剔除候选突破点的 p 值分布, 确定所述终止 p 值, 例如将被剔除的候选突破点的 p 值做分布图, 选取 p 值变化趋势最大的地方作为终止 p 值 (p_{final})。

根据本发明的具体示例, 终止 p 值可以为 1.1×10^{-50} 。根据本发明的实施例, 参数确定单元 204 可以进一步包括适于执行下列的模块: 确定所述检验窗口中所包含的所有一级窗口的平均标准化的测序序列数目 $\bar{Z}$, 其中, 所得到的平均标准化的测序序列数目 $\bar{Z}$ 作为第一参数。进一步, 在判断单元 205 中存储
15 有预定阈值, 由此, 判断单元 205 可以参数确定单元 204 所确定的第一参数与该预定阈值进行比较, 从而判断所得到的检验窗口是否具有拷贝数变异, 其中, 根据本发明的实施例, 根据本发明的实施例, 所述预定阈值采用第一阈值和第二阈值, 通过将第一参数 $\bar{Z}$ 与第一阈值和第二阈值相比较, 小于第一阈值为拷贝数减少 (即缺失), 大于第二阈值为拷贝数增加 (即重复), 由此可以确定拷贝数变异类型。根据本发明的具体示例, 设定 $\alpha = 0.05$ 为显著界线, 第一阈值为 -1.645, 第二阈值为 1.645, 由此, 可以进一步
20 有效地确定拷贝数变异类型。

由此, 利用根据本发明实施例的确定样本基因组中是否存在拷贝数变异的系统, 能够有效地实施根据本发明实施例的确定样本基因组中是否存在拷贝数变异的方法, 从而能够有效地确定样本基因组中是否存在拷贝数变异, 并且适用于各种拷贝数变异, 包括但不限于染色体非整倍性、染色体片段缺失、染色体片段增加、微缺失、微重复。

25 需要说明的是, 本领域技术人员能够理解, 在前面所描述的确定样本基因组中是否存在拷贝数变异的方法的特征和优点也适合于确定样本基因组中是否存在拷贝数变异的系统, 为描述方便, 不再详述。

计算机可读介质

根据本发明的第三方面, 本发明提出了一种计算机可读介质。根据本发明的实施例, 该计算机可读介质上存储有指令, 所述指令适于被处理器执行以便通过下列步骤确定样本基因组中是否存在拷贝数变异:
30 将测序结果与参照基因组序列进行比对, 以便确定所述测序序列在所述参照基因组序列上的分布, 其中所述测序结果是由通过对所述样本基因组进行测序所获得的多个测序序列构成的; 基于所述测序序列在参照基因组序列上的分布, 在所述参照基因组序列上确定多个突破点, 所述突破点两侧的测序序列数目存在显著差异; 基于所述多个突破点, 在所述参照基因组上确定检验窗口; 基于落入所述检验窗口的测序序列, 确定第一参数; 以及基于所述第一参数与预定阈值的差异, 确定所述样本基因组, 针对所
35 述检验窗口是否存在拷贝数变异。借助该计算机可读介质, 能够有效地实施根据本发明实施例的确定样本基因组中是否存在拷贝数变异的方法, 从而能够有效地确定样本基因组中是否存在拷贝数变异, 并且适用于各种拷贝数变异, 包括但不限于染色体非整倍性、染色体片段缺失、染色体片段增加、微缺失、微重复。

40 需要说明的是, 本领域技术人员能够理解, 在前面所描述的确定样本基因组中是否存在拷贝数变异的方法的特征和优点也适合于该计算机可读介质, 为描述方便, 不再详述。

下面将结合实施例对本发明的方案进行解释。本领域技术人员将会理解，下面的实施例仅用于说明本发明，而不应视为限定本发明的范围。实施例中未注明具体技术或条件的，按照本领域内的文献所描述的技术或条件（例如参考 J. 萨姆布鲁克等著，黄培堂等译的《分子克隆实验指南》，第三版，科学出版社）或者按照产品说明书进行。所用试剂或仪器未注明生产厂商者，均为可以通过市购获得的常规产品，例如可以采购自 Illumina 公司。

一般方法

参考图 3，在实施例中采用的确定样本基因组中是否存在拷贝数变异的方法包括下列：

首先，对样品的全基因组进行扩增，并进行测序得到测序序列（测序数据）；

接下来，通过 SOAP2 将所得到的测序序列与 NCBI 数据库中的标准人类基因组参考序列进行 SOAP2 比对，得到所测序列在基因组上的位置信息。为避免重复序列对拷贝数变异分析的干扰，只选取与人类基因组参考序列唯一比对的测序序列（reads），进行后续分析。

接下来，寻找测试样本基因组上两侧测序序列数目在统计上有显著差异的位点，其包括下列步骤：

a) 计算测试样本（可以同时多个样本进行分析）的相对测序序列数：

在人类基因组参考序列上开长度为 w 的窗口（ w 可以是大于 1 的任意整数，例如 10K - 10M bp，优选 50K - 1M bp，更优选为 100K - 300K bp，例如约 150K bp），统计所得到的测序序列中落在每个窗口上的测序序列数 $r_{i,j}$ ，其中下标 i 和 j 分别代表窗口编号和样本编号，并计算每个窗口的 GC 含量

$GC_{i,j}$ ，计算相对测序序列数 $R_{i,j} = \log_2 \left(\frac{r_{i,j}}{\bar{r}_j} \right)$ ，其中平均测序序列数 $\bar{r}_j = \frac{1}{n} \sum_{i=1}^n r_{i,j}$ ，

b) 数据校正与标准化：

在 GC 含量为横坐标和相对测序序列数 R 为纵坐标的坐标系中，将 GC 从小到大划分为大小相等的区域，统计每个区域中 R 的平均值 M_s ， s 为 GC 区域的编号；

对样品中的每个窗口，计算校正的相对测序序列数 $\tilde{R}_{i,j} = R_{i,j} - M_s$ ，窗口 i 的 GC 含量在第 s GC 区域内；

对于样品中每个窗口，计算标准化的相对测序序列数 $Z_{i,j}$ ：

$Z_{i,j} = (R_{i,j} - \tilde{R}_{i,j} - mean_j) / SD_j$ ，其中 $mean_j = \frac{1}{n} \sum_{i=1}^n (R_{i,j} - \tilde{R}_{i,j})$ ，

$SD_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (R_{i,j} - \tilde{R}_{i,j} - mean_j)^2}$

c) 确定并筛选突破点

确定突破点：针对参考基因组序列上每个点，取其左右两侧各 n 个窗口（例如 100 个窗口）作为两个群体进行统计检验，每个点会计算得到一个代表该点两侧差异的 p 值，留下 p 值最小的 m 个点（例如 3000 个）作为突破点（Breakpoint）

筛选突破点：将所有排过序的突破点记为 $B_c = \{b_1, b_2, \dots, b_s\}$ ，每个突破点都存在左右两个片段，所述片段即上一个突破点到该突破点的区域以及该突破点到下一个突破点的区域，将这两个片段中所有 $Z_{i,j}$ 进行统计检验（例如进行游程检验，一种非参数检验，利用两个群体元素混合后的分布均匀状态评价此两个群体的差异显著性）所得的 p 值（ p_k ），视作“ b_k 作为突破点的显著性”，将 p_k 最大的候选突破点剔除，反复此步骤，直到所有 p 值都小于该染色体的终止 p 值（ p_{final} ）；

终止 p 值的获得：在测试过程中，将以对照样本作为测试样本进行上述步骤确定若干突破点步骤，将全基因组上所有排过序的候选突破点记为 $B_c = \{b_1, b_2, \dots, b_s\}$ ，每个候选突破点 b_k 都存在左右两个窗口，将这两个窗口中所有 $Z_{i,j}$ 进行游程检验所得的 p 值（ p_k ），视作“ b_k 作为突破点的显著性”，将最不显著的候选突破点剔除并记录该突破点对应的 p 值，直到候选突破点数为 0，将被剔除的候选突破点的 p 值做分布图，选取 p 值变化趋势最大的地方作为终止 p 值（ p_{final} ）；

确定检验窗口，并验证检验窗口：在获得经过筛选的突破点后，确定检验窗口。为了进一步对检验窗口进行判断，计算该片段中 $Z_{i,j}$ 的平均值，记为 $\bar{Z}$ ，如果片段的 $\bar{Z}$ 超出阈值范围，则该片段为拷贝数变异，其中阈值的确定具体如下：

对合并窗口后每个片段, 计算所有对照样品在该片段中的标准化相对测序序列数 $Z_{i,j}$ 的平均值和标准差; 因为每个片段中 Z 符合正态分布, 所以根据前面步骤中计算得到的平均值和标准差, 计算累积概率在 0.05 时该片段的阈值范围, 作为过滤该片段是否存在拷贝数变异的阈值。

实施例 1 对一例胚胎单细胞样品进行胎儿片段拷贝数变异检测, 以及对一例胚胎单细胞样品进行染色体非整倍性变异检测

1、全基因组扩增: 本实施例采用 Sigma Aldrich 公司的 GenomePlex® Single Cell Whole Genome Amplification Kit 对两例胚胎单细胞样本进行全基因组扩增。所述胚胎单细胞样本为第五天囊胚期的外滋养层单细胞, 通过激光捕获显微切割方法从囊胚中分离。两例胚胎单细胞经裂解后, 均按照制造商所提供的试剂盒说明书进行全基因组扩增操作。

2、测序: 本实施例中, 采用 Illumina 公司的 Hiseq2000 测序平台对于获自上述 2 例胚胎单细胞全基因组扩增的 DNA 进行测序, 按照 Illumina 公司所提供的说明书, 进行文库构建、上机测序, 使每个样本得到约 0.36G 数据量, 每个样本根据标签序列进行区分。利用比对软件 SOAP2, 将测序所得测序序列与 NCBI 数据库中版本 36 (hg18; NCBI Build36) 的人类基因组参考序列进行比对, 将得到测序序列在人类基因组参考序列上进行定位。

3、数据分析

a) 计算测试样本和对照样本 (对照样本为核型正常的样品) 的相对测序序列数:

在人类基因组参考序列上划分为多个长度为 150K bp 的窗口, 统计前面步骤 2) 所得到测序序列中落在每个窗口上的测序序列数 r_{ij} , 其中下标 i 和 j 分别代表窗口编号和样本编号, 并计算每个窗口的 GC 含量 GC_{ij} , 按照一般方法中提供的公式计算相对测序序列数

b) 数据校正与标准化:

在 GC 含量为横坐标和相对测序序列数 R 为纵坐标的坐标系中, 将 GC 从小到大划分为大小为 0.001 的区域, 统计每个区域中 R 的平均值 M_s , s 为 GC 区域的编号, 见表 1。按照一般方法中提供的公式对所得到的测序序列进行校正和标准化。

表 1 校正过程中每个 GC 区域 M_s 列表

s	样品 S1		样品 S2		S	样品 S1		样品 S2	
	GC	M_s	GC	M_s		GC	M_s	GC	M_s
1	0.255~0.256	2.45	0.255~0.256	2.74	118	0.433~0.434	-0.23	0.452~0.453	-0.06
2	0.314~0.315	0.04	0.336~0.337	-0.26	119	0.434~0.435	-0.21	0.453~0.454	-0.15
3	0.317~0.318	0.22	0.337~0.338	-0.21	120	0.435~0.436	-0.25	0.454~0.455	-0.22
4	0.319~0.32	0.01	0.338~0.339	-0.18	121	0.436~0.437	-0.25	0.455~0.456	-0.16
5	0.32~0.321	0.19	0.339~0.34	0.16	122	0.437~0.438	-0.24	0.456~0.457	-0.19
6	0.321~0.322	0.13	0.34~0.341	-0.73	123	0.438~0.439	-0.23	0.457~0.458	-0.14
7	0.322~0.323	0.11	0.341~0.342	-0.3	124	0.439~0.44	-0.29	0.458~0.459	-0.15
8	0.323~0.324	0.12	0.342~0.343	-0.28	125	0.44~0.441	-0.28	0.459~0.46	-0.21
9	0.324~0.325	-0.08	0.343~0.344	-0.36	126	0.441~0.442	-0.41	0.46~0.461	-0.1
10	0.325~0.326	0.02	0.344~0.345	-0.31	127	0.442~0.443	-0.28	0.461~0.462	-0.2
11	0.326~0.327	0.39	0.345~0.346	-0.19	128	0.443~0.444	-0.36	0.462~0.463	-0.19
12	0.327~0.328	0.15	0.346~0.347	-0.18	129	0.444~0.445	-0.33	0.463~0.464	-0.12
13	0.328~0.329	0.11	0.347~0.348	-0.25	130	0.445~0.446	-0.35	0.464~0.465	-0.3
14	0.329~0.33	0.22	0.348~0.349	-0.33	131	0.446~0.447	-0.36	0.465~0.466	-0.29
15	0.33~0.331	0	0.349~0.35	-0.28	132	0.447~0.448	-0.3	0.466~0.467	-0.18
16	0.331~0.332	-0.04	0.35~0.351	-0.33	133	0.448~0.449	-0.47	0.467~0.468	-0.27
17	0.332~0.333	0.12	0.351~0.352	-0.14	134	0.449~0.45	-0.38	0.468~0.469	-0.24
18	0.333~0.334	0.12	0.352~0.353	-0.24	135	0.45~0.451	-0.43	0.469~0.47	-0.28
19	0.334~0.335	0.06	0.353~0.354	-0.23	136	0.451~0.452	-0.4	0.47~0.471	-0.25
20	0.335~0.336	0.14	0.354~0.355	-0.15	137	0.452~0.453	-0.34	0.471~0.472	-0.24
21	0.336~0.337	0.1	0.355~0.356	-0.21	138	0.453~0.454	-0.5	0.472~0.473	-0.44
22	0.337~0.338	0.08	0.356~0.357	-0.19	139	0.454~0.455	-0.45	0.473~0.474	-0.37
23	0.338~0.339	0.08	0.357~0.358	-0.18	140	0.455~0.456	-0.5	0.474~0.475	-0.36
24	0.339~0.34	0.1	0.358~0.359	-0.14	141	0.456~0.457	-0.47	0.475~0.476	-0.31
25	0.34~0.341	0.15	0.359~0.36	-0.09	142	0.457~0.458	-0.49	0.476~0.477	-0.41
26	0.341~0.342	0.12	0.36~0.361	-0.15	143	0.458~0.459	-0.47	0.477~0.478	-0.41
27	0.342~0.343	0.11	0.361~0.362	-0.13	144	0.459~0.46	-0.52	0.478~0.479	-0.41
28	0.343~0.344	0.06	0.362~0.363	-0.13	145	0.46~0.461	-0.58	0.479~0.48	-0.36

29	0.344~0.345	0.17	0.363~0.364	-0.09	146	0.461~0.462	-0.61	0.48~0.481	-0.44
30	0.345~0.346	0.09	0.364~0.365	-0.13	147	0.462~0.463	-0.64	0.481~0.482	-0.37
31	0.346~0.347	0.14	0.365~0.366	-0.08	148	0.463~0.464	-0.55	0.482~0.483	-0.38
32	0.347~0.348	0.08	0.366~0.367	-0.06	149	0.464~0.465	-0.57	0.483~0.484	-0.46
33	0.348~0.349	0.11	0.367~0.368	-0.06	150	0.465~0.466	-0.68	0.484~0.485	-0.52
34	0.349~0.35	0.13	0.368~0.369	-0.08	151	0.466~0.467	-0.57	0.485~0.486	-0.57
35	0.35~0.351	0.08	0.369~0.37	-0.06	152	0.467~0.468	-0.78	0.486~0.487	-0.47
36	0.351~0.352	0.14	0.37~0.371	-0.09	153	0.468~0.469	-0.75	0.487~0.488	-0.55
37	0.352~0.353	0.13	0.371~0.372	-0.03	154	0.469~0.47	-0.64	0.488~0.489	-0.45
38	0.353~0.354	0.12	0.372~0.373	-0.01	155	0.47~0.471	-0.74	0.489~0.49	-0.74
39	0.354~0.355	0.13	0.373~0.374	-0.03	156	0.471~0.472	-0.57	0.49~0.491	-0.52
40	0.355~0.356	0.12	0.374~0.375	-0.06	157	0.472~0.473	-0.69	0.491~0.492	-0.59
41	0.356~0.357	0.15	0.375~0.376	-0.04	158	0.473~0.474	-0.73	0.492~0.493	-0.57
42	0.357~0.358	0.14	0.376~0.377	-0.04	159	0.474~0.475	-0.74	0.493~0.494	-0.59
43	0.358~0.359	0.16	0.377~0.378	-0.01	160	0.475~0.476	-0.84	0.494~0.495	-0.54
44	0.359~0.36	0.14	0.378~0.379	-0.01	161	0.476~0.477	-0.79	0.495~0.496	-0.63
45	0.36~0.361	0.14	0.379~0.38	0	162	0.477~0.478	-0.81	0.496~0.497	-0.69
46	0.361~0.362	0.14	0.38~0.381	-0.01	163	0.478~0.479	-0.78	0.497~0.498	-0.63
47	0.362~0.363	0.15	0.381~0.382	0.03	164	0.479~0.48	-0.71	0.498~0.499	-0.7
48	0.363~0.364	0.09	0.382~0.383	0	165	0.48~0.481	-0.94	0.499~0.5	-0.69
49	0.364~0.365	0.1	0.383~0.384	0.01	166	0.481~0.482	-0.8	0.5~0.501	-0.64
50	0.365~0.366	0.14	0.384~0.385	0	167	0.482~0.483	-0.74	0.501~0.502	-0.75
51	0.366~0.367	0.12	0.385~0.386	0.04	168	0.483~0.484	-0.78	0.502~0.503	-0.71
52	0.367~0.368	0.11	0.386~0.387	0.03	169	0.484~0.485	-0.95	0.503~0.504	-0.85
53	0.368~0.369	0.12	0.387~0.388	0.03	170	0.485~0.486	-0.81	0.504~0.505	-0.67
54	0.369~0.37	0.15	0.388~0.389	0.04	171	0.486~0.487	-0.96	0.505~0.506	-0.97
55	0.37~0.371	0.15	0.389~0.39	0.03	172	0.487~0.488	-1	0.506~0.507	-0.81
56	0.371~0.372	0.14	0.39~0.391	0.05	173	0.488~0.489	-0.91	0.507~0.508	-0.72
57	0.372~0.373	0.09	0.391~0.392	0.02	174	0.489~0.49	-0.86	0.508~0.509	-0.75
58	0.373~0.374	0.11	0.392~0.393	0.03	175	0.49~0.491	-0.85	0.509~0.51	-0.6
59	0.374~0.375	0.13	0.393~0.394	0.05	176	0.491~0.492	-1.01	0.51~0.511	-0.78
60	0.375~0.376	0.11	0.394~0.395	0.07	177	0.492~0.493	-1.11	0.511~0.512	-0.76
61	0.376~0.377	0.12	0.395~0.396	0.05	178	0.493~0.494	-0.94	0.512~0.513	-0.75
62	0.377~0.378	0.13	0.396~0.397	0.07	179	0.494~0.495	-1.01	0.513~0.514	-0.82
63	0.378~0.379	0.08	0.397~0.398	0.06	180	0.495~0.496	-0.95	0.514~0.515	-0.75
64	0.379~0.38	0.13	0.398~0.399	0.03	181	0.496~0.497	-0.99	0.515~0.516	-1.15
65	0.38~0.381	0.08	0.399~0.4	0.08	182	0.497~0.498	-1.09	0.516~0.517	-0.68
66	0.381~0.382	0.06	0.4~0.401	0.08	183	0.498~0.499	-1.17	0.517~0.518	-0.73
67	0.382~0.383	0.12	0.401~0.402	0.1	184	0.499~0.5	-0.96	0.518~0.519	-1.07
68	0.383~0.384	0.1	0.402~0.403	0.09	185	0.5~0.501	-1.02	0.519~0.52	-1
69	0.384~0.385	0.11	0.403~0.404	0.08	186	0.501~0.502	-1.06	0.52~0.521	-0.93
70	0.385~0.386	0.08	0.404~0.405	0.09	187	0.502~0.503	-1.13	0.521~0.522	-0.99
71	0.386~0.387	0.07	0.405~0.406	0.09	188	0.503~0.504	-1.48	0.522~0.523	-1
72	0.387~0.388	0.07	0.406~0.407	0.1	189	0.504~0.505	-1.16	0.523~0.524	-1.01
73	0.388~0.389	0.07	0.407~0.408	0.06	190	0.505~0.506	-0.8	0.524~0.525	-1.17
74	0.389~0.39	0.07	0.408~0.409	0.07	191	0.506~0.507	-1.22	0.525~0.526	-1.13
75	0.39~0.391	0.1	0.409~0.41	0.08	192	0.507~0.508	-1.06	0.526~0.527	-1.14
76	0.391~0.392	0.06	0.41~0.411	0.06	193	0.508~0.509	-1.31	0.527~0.528	-0.73
77	0.392~0.393	0.06	0.411~0.412	0.05	194	0.509~0.51	-1.27	0.528~0.529	-1.01
78	0.393~0.394	0.06	0.412~0.413	0.09	195	0.51~0.511	-1.05	0.529~0.53	-1.15
79	0.394~0.395	0.05	0.413~0.414	0.06	196	0.511~0.512	-1.37	0.53~0.531	-1.03

80	0.395~0.396	0.04	0.414~0.415	0.08	197	0.512~0.513	-1.39	0.531~0.532	-1.06
81	0.396~0.397	0.06	0.415~0.416	0.05	198	0.513~0.514	-1.43	0.532~0.533	-1.05
82	0.397~0.398	0.03	0.416~0.417	0.04	199	0.514~0.515	-1.45	0.533~0.534	-1.42
83	0.398~0.399	0.02	0.417~0.418	0.09	200	0.515~0.516	-1.3	0.534~0.535	-0.89
84	0.399~0.4	0.09	0.418~0.419	0.06	201	0.516~0.517	-1.38	0.535~0.536	-1.8
85	0.4~0.401	0.02	0.419~0.42	-0.01	202	0.517~0.518	-0.94	0.536~0.537	-0.81
86	0.401~0.402	0.01	0.42~0.421	0.09	203	0.518~0.519	-1.48	0.537~0.538	-0.89
87	0.402~0.403	0.03	0.421~0.422	0.08	204	0.519~0.52	-1.48	0.538~0.539	-0.91
88	0.403~0.404	0	0.422~0.423	0.06	205	0.52~0.521	-0.91	0.539~0.54	-0.96
89	0.404~0.405	0.03	0.423~0.424	0.08	206	0.521~0.522	-0.89	0.54~0.541	-1.98
90	0.405~0.406	0.02	0.424~0.425	0.03	207	0.522~0.523	-1.9	0.541~0.542	-0.29
91	0.406~0.407	0.03	0.425~0.426	0.06	208	0.523~0.524	-1.46	0.542~0.543	-1.28
92	0.407~0.408	0.02	0.426~0.427	0.05	209	0.524~0.525	-2.02	0.543~0.544	-1.84
93	0.408~0.409	-0.01	0.427~0.428	0.06	210	0.525~0.526	-1.39	0.544~0.545	-1.41
94	0.409~0.41	-0.06	0.428~0.429	0.03	211	0.526~0.527	-1.72	0.545~0.546	-0.54
95	0.41~0.411	-0.06	0.429~0.43	0.04	212	0.528~0.529	-1.08	0.547~0.548	-1.31
96	0.411~0.412	-0.04	0.43~0.431	0.05	213	0.529~0.53	-1.42	0.548~0.549	-1.11
97	0.412~0.413	-0.04	0.431~0.432	0.01	214	0.53~0.531	-1.71	0.549~0.55	-1.38
98	0.413~0.414	-0.02	0.432~0.433	0.04	215	0.531~0.532	-2.27	0.55~0.551	-1.5
99	0.414~0.415	-0.05	0.433~0.434	0	216	0.532~0.533	-1.78	0.551~0.552	-1.22
100	0.415~0.416	-0.07	0.434~0.435	-0.02	217	0.533~0.534	-1.55	0.552~0.553	-0.8
101	0.416~0.417	-0.08	0.435~0.436	0.01	218	0.535~0.536	-1.25	0.553~0.554	-1.32
102	0.417~0.418	-0.11	0.436~0.437	0.04	219	0.536~0.537	-1.09	0.554~0.555	-1.79
103	0.418~0.419	-0.07	0.437~0.438	0.01	220	0.537~0.538	-2.02	0.556~0.557	-1.3
104	0.419~0.42	-0.09	0.438~0.439	-0.01	221	0.54~0.541	-2.16	0.557~0.558	-1.48
105	0.42~0.421	-0.13	0.439~0.44	-0.01	222	0.541~0.542	-1.64	0.558~0.559	-1.7
106	0.421~0.422	-0.1	0.44~0.441	-0.01	223	0.544~0.545	-2.3	0.559~0.56	-1.55
107	0.422~0.423	-0.12	0.441~0.442	-0.01	224	0.546~0.547	-2.51	0.561~0.562	-1.62
108	0.423~0.424	-0.11	0.442~0.443	-0.06	225	0.548~0.549	-2.7	0.563~0.564	-1.68
109	0.424~0.425	-0.17	0.443~0.444	-0.04	226	0.549~0.55	-1.77	0.564~0.565	-1.47
110	0.425~0.426	-0.14	0.444~0.445	-0.07	227	0.55~0.551	-1.08	0.569~0.57	-1.42
111	0.426~0.427	-0.14	0.445~0.446	-0.11	228	0.551~0.552	-2.13	0.58~0.581	-1.74
112	0.427~0.428	-0.15	0.446~0.447	-0.13	229	0.553~0.554	-2.19	0.583~0.584	-2.43
113	0.428~0.429	-0.19	0.447~0.448	-0.08	230	0.555~0.556	-2.04	0.6~0.601	-1.79
114	0.429~0.43	-0.18	0.448~0.449	-0.11	231	0.556~0.557	-1.93		
115	0.43~0.431	-0.18	0.449~0.45	-0.07	232	0.562~0.563	-2.51		
116	0.431~0.432	-0.21	0.45~0.451	-0.16	233	0.572~0.573	-1.85		
117	0.432~0.433	-0.26	0.451~0.452	-0.08	234	0.574~0.575	-2.74		

c) 合并窗口

确定突破点, 针对参考基因组序列上每个点, 取其左右两侧各 100 个窗口作为两个群体进行游程检验, 每个点会计算得到一个代表该点两侧差异的 p 值, 留下 p 值最小的 3000 个点作为突破点(Breakpoint)

5 筛选突破点: 将所有排过序的突破点记为 $B_c = \{b_1, b_2, \dots, b_s\}$, 每个突破点都存在左右两个片段, 所述片段即上一个突破点到该突破点的区域以及该突破点到下一个突破点的区域, 将这两个片段中所有 $Z_{i,j}$ 进行游程检验所得的 p 值 (p_k), 视作 " b_k 作为突破点的显著性", 将 p_k 最大的候选突破点剔除, 反复此步骤, 直到所有 p 值都小于该染色体的终止 p 值 1.1×10^{-30} ;

10 d) 在得到筛选后的突破点后, 确定相邻两个突破点之间的区域为检验窗口, 从而对窗口进行合并, 为了进一步对合并窗口后获得的片段进行过滤, 计算该片段中 $Z_{i,j}$ 的平均值, 记为 $\bar{Z}$, 如果片段的 $\bar{Z}$ 超出阈值范围, 则该片段为拷贝数变异。采用 -1.645 作为第一阈值和采用 1.645 作为第二阈值。

4、结果。表 2 表示了在该实施例中, 各胚胎单细胞样品经 WGA 全基因组扩增后, 拷贝数变异的检测结果列表。

表 2. 实施例 1 各胚胎单细胞样品经 WGA 拷贝数变异检测结果列表

编号	染色体	CNV 起点	CNV 终点	CNV 大小	CNV 类型	涉及区带
S1	5	63,429	23,496,649	23.4M	缺失	4q34.3→q35.2
	12	16,037	18,926,068	18.9M	重复	7p21.1→p22.3
S2	21	1	46,944,323	46.9M	重复	21p13→q22.3

由表 2 的结果, 可以看出, 通过本发明的确定样本基因组中是否存在拷贝数变异的方法, 能够有效地确定各种类型的拷贝数变异。

实施例 2

使用与实施例 1 相同的胚胎, 重复实施例 1, 只是在提取 DNA 后直接进行测序(未经 WGA)。实施例 1 和实施例 2 结果比较见表 3、图 4 和图 5。

表 3. 实施案例各样品 WGA 与未经 WGA 样本测序数据拷贝数变异检测结果比较

编号	染色体	胚胎 DNA(未经 WGA)的测序结果	胚胎单细胞经 WGA 用本发明方法的检测结果	判断结果
S1	5	缺失: 10,002-23,312,155	缺失: 63,429-23,496,649	一致
	12	重复: 145,741-14,780,155	重复: 16,037-18,926,068	一致
S2	21	21 三体	21 三体	一致

从表 3 数据及染色体数字核型图图 4、图 5 可看出, 样品经 WGA 与未经 WGA 测序数据拷贝数变异检测结果是一致的。对于表 3 “缺失”或“重复”起始终止位置(发生拷贝数变异的边界)的差异, 由于拷贝数变异的边界难以精确确定, 一般地, 对于约 150K 的一级窗口, 可以判定两者的边界差异在 100-300Kb 为完全一致, 在 300Kb-1Mb 范围为较一致, 表 3 显示两方法确定的拷贝数变异边界的差异范围在 100-300Kb 或 300Kb-1Mb 之内, 判定两方法确定的发生拷贝数变异的边界一致。

工业实用性

本发明的确定样本基因组中是否存在拷贝数变异的方法、系统和计算机可读介质能够有效地用于确定样本基因组中是否存在拷贝数变异。

尽管本发明的具体实施方式已经得到详细的描述, 本领域技术人员将会理解。根据已经公开的所有教导, 可以对那些细节进行各种修改和替换, 这些改变均在本发明的保护范围之内。本发明的全部范围由所附权利要求及其任何等同物给出。

在本说明书的描述中, 参考术语“一个实施例”、“一些实施例”、“示意性实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中, 对上述术语的示意性表述不一定指的是相同的实施例或示例。而且, 描述的具体特征、结构、材料或者特点可以在任何一个或多个实施例或示例中以合适的方式结合。

1、一种确定样本基因组中是否存在拷贝数变异的方法，其特征在于，包括以下步骤：

对所述样本基因组进行测序，以便获得由多个测序序列构成的测序结果；

5 将所述测序结果与参照基因组序列进行比对，以便确定所述测序序列在所述参照基因组序列上的分布；

基于所述测序序列在参照基因组序列上的分布，在所述参照基因组序列上确定多个突破点，所述突破点两侧的测序序列数目存在显著差异；

基于所述多个突破点，在所述参照基因组上确定检验窗口；

10 基于落入所述检验窗口的测序序列，确定第一参数；以及

基于所述第一参数与预定阈值的差异，确定所述样本基因组，针对所述检验窗口是否存在拷贝数变异。

2、根据权利要求1所述的方法，其特征在于，进一步包括从生物样本中提取样本基因组的步骤。

15 3、根据权利要求2所述的方法，其特征在于，所述生物样本为孕妇样本或胎儿样本，任选地，所述生物样本为选自孕妇血浆、绒毛膜绒毛、羊水、脐带血、胎盘和胎儿足跟血的至少一种。

4、根据权利要求2所述的方法，其特征在于，所述生物样本为选自血液、尿液、唾液、组织、生殖细胞、受精卵、卵裂球和胚胎的至少一种，

任选地，所述生物样本为单细胞。

20 5、根据权利要求1所述的方法，其特征在于，对所述样本基因组进行测序进一步包括：

对所述样本基因组进行扩增；

利用所述经过扩增的样本基因组构建测序文库；以及

对所述测序文库进行测序。

6、根据权利要求4所述的方法，其特征在于，进一步包括对所述单细胞进行裂解，以便释放所述单细胞的全基因组的步骤。

25 7、根据权利要求6所述的方法，其特征在于，

利用碱性裂解液对所述单细胞进行裂解，以便释放所述单细胞的全基因组。

8、根据权利要求7所述的方法，其特征在于，利用基于PCR的全基因组扩增方法对所述全基因组进行扩增。

9、根据权利要求8所述的方法，其特征在于，

30 所述基于PCR的全基因组扩增方法为OmniPlex WGA方法。

10、根据权利要求5所述的方法，其特征在于，

利用选自HiSeq系统、Miseq系统、Genome Analyzer(GA)系统、454 FLX、SOLiD系统、Ion Torrent系统和单分子测序装置的至少一种对所述测序文库进行测序。

35 11、根据权利要求1所述的方法，其特征在于，所述拷贝数变异为选自染色体非整倍性、染色体片段缺失、染色体片段增加、微缺失、微重复的至少一种。

12、根据权利要求1所述的方法，其特征在于，在所述参照基因组上确定突破点进一步包括：

将参考基因组序列划分为多个预定长度的一级窗口，并确定落入各一级窗口中的测序序列；

针对参考基因组序列上的至少一个位点，确定落入所述位点两侧相同数目一级窗口中的测序序列数目；

40 确定所述位点的p值，所述p值表示两侧测序数据数目的显著差异性；以及

如果所述位点的p值小于终止p值，判断所述位点为突破点。

13、根据权利要求12所述的方法，其特征在于，所述落入各一级窗口中的测序序列为唯一比对测序序列。

14、根据权利要求12所述的方法，其特征在于，在所述位点两侧各取100个一级窗口。

45 15、根据权利要求12所述的方法，其特征在于，所述一级窗口的长度均为100-200Kbp，优选150Kbp。

16、根据权利要求 12 所述的方法，其特征在于，所述终止 p 值为至多 1.1×10^{-50} 。

17、根据权利要求 12 所述的方法，其特征在于，确定所述位点两侧测序数据数目的显著差异性 p 值进一步包括：

5 针对所述位点，在所述位点两侧各取相同数目的一级窗口，并且计算每个一级窗口的相对测序序列数目 R_i ，其中 i 表示一级窗口的编号

对所有一级窗口的相对测序序列数目 R_i 进行游程检验，以便确定所述位点的 p 值，

其中，

所述相对测序序列数目是通过下列公式确定的：

$$R_i = \log_2 \left(\frac{r_i}{\bar{r}} \right)$$

其中 r_i 表示落入第 i 一级窗口的测序序列数目，

10

$$\bar{r} = \frac{1}{n} \sum_{i=1}^n r_i, \quad n \text{ 表示一级窗口的总数目。}$$

18、根据权利要求 17 所述的方法，其特征在于，对所有一级窗口的相对测序序列数目进行游程检验进一步包括：

15 对每个一级窗口的相对测序序列数目 R_i 进行 GC 含量校正，以便获得校正的相对测序序列数目 $\tilde{R}_i$ ；
基于所述校正的相对测序序列数目，确定每个一级窗口的标准化的测序序列数目 Z_i ；以及
对所有一级窗口的标准化的测序序列数目 Z_i 进行游程检验。

19、根据权利要求 18 所述的方法，其特征在于，

所述校正的相对测序序列数目 $\tilde{R}_i$ 是通过下列步骤获得的：

20 计算每个一级窗口的 GC 含量；

将 GC 含量以 0.001 为单位划分为多个区域，并且统计每个区域中相对测序序列数目的平均值 M_s ，
其中 s 为 GC 区域的编号；

$$\tilde{R}_i = R_i - M_s;$$

根据下列公式确定所述校正的相对测序序列数目 $\tilde{R}_i$ ：

所述标准化的测序序列数目 Z_i 是通过下列公式确定的

25 其中，

$$Z_i = (R_i - \tilde{R}_i - \text{mean}) / SD$$

$$\text{mean} = \frac{1}{n} \sum_{i=1}^n (R_i - \tilde{R}_i)$$

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (R_i - \tilde{R}_i - \text{mean})^2}。$$

20、根据权利要求 19 所述的方法，其特征在于，基于所述多个突破点，在所述参照基因组上确定检验窗口进一步包括：

1) 确定多个候选突破点，其中在所述候选突破点的前后均存在其他突破点；

2) 确定每个候选突破点的 p 值, 并剔除 p 值最大的候选突破点;

3) 对剩余的候选突破点重复步骤 2), 直到剩余候选突破点的 p 值均小于终止 p 值, 所述剩余候选突破点作为经过筛选的候选突破点; 以及

4) 确定相邻两个经过筛选的候选突破点之间的区域为检验窗口,

5 其中, 通过下列步骤确定所述候选突破点的 p 值:

将所述候选突破点与相邻的上一个突破点之间的区域作为第一候选区域, 将所述候选突破点与相邻的下一个突破点之间的区域作为第二候选区域;

对所述第一候选区域和第二候选区域中所包含一级窗口的标准化的测序序列数目 Z_i 进行游程检验, 以便确定所述候选突破点的 p 值,

10 任选地, 所述终止 p 值是通过下列步骤确定的:

利用对照样品的测序结果, 重复在参照基因组上确定检验窗口的操作, 并记录每次被剔除候选突破点的 p 值, 直到候选突破点的数目为零; 以及

基于被剔除候选突破点的 p 值分布, 确定所述终止 p 值,

任选地, 所述终止 p 值为 1.1×10^{-50} 。

15 21、根据权利要求 20 所述的方法, 其特征在于, 基于落入所述检验窗口的测序序列, 确定第一参数进一步包括:

确定所述检验窗口中所包含的所有一级窗口的平均标准化的测序序列数目 $\bar{Z}$, 所述平均标准化的测序序列数目 $\bar{Z}$ 作为第一参数。

20 22、根据权利要求 1 所述的方法, 其特征在于, 所述预定阈值采用 -1.645 作为第一阈值和 1.645 作为第二阈值。

23、根据权利要求 1 所述的方法, 其特征在于, 所述参照基因组序列为选自人类 21 号染色体、18 号染色体、13 号染色体、X 染色体和 Y 染色体的至少一种的序列。

24、一种确定样本基因组中是否存在拷贝数变异的系统, 其特征在于, 包括:

25 测序装置, 所述测序装置用于对所述样本基因组进行测序, 以便获得由多个测序序列构成的测序结果;

分析装置, 所述分析装置与所述测序装置相连, 以便基于所述测序结果确定所述基因组中是否存在拷贝数变异, 所述分析装置进一步包括:

比对单元, 所述比对单元适于将所述测序结果与参照基因组序列进行比对, 以便确定所述测序序列在所述参照基因组序列上的分布;

30 突破点确定单元, 所述突破点确定单元与所述比对单元相连, 并且适于基于所述测序序列在参照基因组序列上的分布, 在所述参照基因组序列上确定多个突破点, 所述突破点两侧的测序序列数目存在显著差异;

检验窗口确定单元, 所述检验窗口确定单元与所述突破点确定单元相连, 并且适于基于所述多个突破点, 在所述参照基因组上确定检验窗口;

35 参数确定单元, 所述参数确定单元与所述检验窗口确定单元相连, 并且适于基于落入所述检验窗口的测序序列, 确定第一参数; 以及

判断单元, 所述判断单元与所述参数确定单元相连, 并且适于基于所述第一参数与预定阈值的差异, 确定所述样本基因组, 针对所述检验窗口是否存在拷贝数变异。

40 25、根据权利要求 24 所述的系统, 其特征在于, 进一步包括基因组提取装置, 所述基因组提取装置适于从生物样本中提取样本基因组。

26、根据权利要求 24 所述的系统, 其特征在于, 所述测序装置进一步包括:

基因组扩增单元, 所述基因组扩增单元适于对所述样本基因组进行扩增;

测序文库构建单元, 所述测序文库构建单元与所述基因组扩增单元相连, 并且适于利用所述经过扩增的样本基因组构建测序文库; 以及

测序单元, 所述测序单元与所述测序文库构建单元相连, 并且适于对所述测序文库进行测序。

27、根据权利要求 26 所述的系统, 其特征在于, 所述测序单元为选自 Hiseq 系统、Miseq 系统、Genome Analyzer(GA)系统、454 FLX、SOLiD 系统、Ion Torrent 系统和单分子测序装置的至少一种。

28、根据权利要求 24 所述的系统, 其特征在于, 所述突破点确定单元进一步包括适于执行下列以确定突破点的模块:

将参考基因组序列划分为多个预定长度的一级窗口, 并确定落入各一级窗口中的测序序列;

针对参考基因组序列上的至少一个位点, 确定落入所述位点两侧相同数目一级窗口中的测序序列数目;

确定所述位点的 p 值, 所述 p 值表示两侧测序数据数目的显著差异性; 以及

如果所述位点的 p 值小于终止 p 值, 判断所述位点为突破点。

29、根据权利要求 28 所述的系统, 其特征在于, 所述突破点确定单元进一步包括适于执行下列以确定 p 值的模块:

针对所述位点, 在所述位点两侧各取相同数目的一级窗口, 并且计算每个一级窗口的相对测序序列数目 R_i , 其中 i 表示一级窗口的编号

对所有一级窗口的相对测序序列数目 R_i 进行游程检验, 以便确定所述位点的 p 值,

其中,

所述相对测序序列数目是通过公式确定的:

$$R_i = \log_2 \left(\frac{r_i}{\bar{r}} \right)$$

其中 r_i 表示落入第 i 一级窗口的测序序列数目,

$$\bar{r} = \frac{1}{n} \sum_{i=1}^n r_i$$

n 表示一级窗口的总数目。

30、根据权利要求 29 所述的系统, 其特征在于, 所述突破点确定单元进一步包括适于执行下列以对一级窗口的相对测序序列数目进行游程检验的模块:

对每个一级窗口的相对测序序列数目 R_i 进行 GC 含量校正, 以便获得校正的相对测序序列数目 $\tilde{R}_i$; 基于所述校正的相对测序序列数目, 确定每个一级窗口的标准化的测序序列数目 Z_i ; 以及对一级窗口的标准化的测序序列数目 Z_i 进行游程检验。

31、根据权利要求 30 所述的系统, 其特征在于,

所述校正的相对测序序列数目 $\tilde{R}_i$ 是通过适于执行下列步骤的模块获得的:

计算每个一级窗口的 GC 含量;

将 GC 含量以 0.001 为单位划分为多个区域, 并且统计每个区域中相对测序序列数目的平均值 M_s , 其中 s 为 GC 区域的编号;

$$\tilde{R}_i = R_i - M_s;$$

根据公式下列公式确定所述校正的相对测序序列数目 $\tilde{R}_i$:

所述标准化的测序序列数目 Z_i 是通过下列公式确定的

$$Z_i = (R_i - \tilde{R}_i - \text{mean}) / SD$$

其中,

$$\text{mean} = \frac{1}{n} \sum_{i=1}^n (R_i - \tilde{R}_i)$$

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (R_i - \tilde{R}_i - \text{mean})^2}。$$

32、根据权利要求 31 所述的系统,其特征在于,基于所述多个突破点,检验窗口确定单元进一步包括适于执行下列的模块:

- 1) 确定多个候选突破点,其中在所述候选突破点的前后均存在其他突破点;
- 5 2) 确定每个候选突破点的 p 值,并剔除 p 值最大的候选突破点;
- 3) 对剩余的候选突破点重复步骤 2),直到剩余候选突破点的 p 值均小于终止 p 值,所述剩余候选突破点作为经过筛选的候选突破点;以及
- 4) 确定相邻两个经过筛选的候选突破点之间的区域为检验窗口,
- 其中,通过下列步骤确定所述候选突破点的 p 值:
- 10 将所述候选突破点与相邻的上一个突破点之间的区域作为第一候选区域,将所述候选突破点与相邻的下一个突破点之间的区域作为第二候选区域;
- 对所述第一候选区域和第二候选区域中所包含一级窗口的标准化的测序序列数目 Z_i 进行游程检验,以便确定所述候选突破点的 p 值,
- 任选地,所述终止 p 值是通过下列步骤确定的:
- 15 利用对照样品的测序结果,重复在参照基因组上确定检验窗口的操作,并记录每次被剔除候选突破点的 p 值,直到候选突破点的数目为零;以及
- 基于被剔除候选突破点的 p 值分布,确定所述终止 p 值,
- 任选地,所述终止 p 值为 1.1×10^{-50} 。

33、根据权利要求 32 所述的系统,其特征在于,参数确定单元进一步包括适于执行下列的模块:

- 20 确定所述检验窗口中所包含的所有一级窗口的平均标准化的测序序列数目 $\bar{Z}$,所述平均标准化的测序序列数目 $\bar{Z}$ 作为第一参数。

34、根据权利要求 24 所述的系统,其特征在于,所述确定单元中存储有预定阈值,所述预定阈值采用 -1.645 作为第一阈值和 1.645 作为第二阈值。

- 25 35、根据权利要求 24 所述的系统,其特征在于,所述比对单元内存储有参照基因组序列为已知的人类基因组序列,任选地,所述参照基因组序列为选自人类 21 号染色体、18 号染色体、13 号染色体、X 染色体和 Y 染色体的至少一种的序列。

36、一种计算机可读介质,其特征在于,所述计算机可读介质上存储有指令,所述指令适于被处理器执行以便通过下列步骤确定样本基因组中是否存在拷贝数变异:

- 30 将测序结果与参照基因组序列进行比对,以便确定所述测序序列在所述参照基因组序列上的分布,
- 其中所述测序结果是由通过对所述样本基因组进行测序所获得的多个测序序列构成的;
- 基于所述测序序列在参照基因组序列上的分布,在所述参照基因组序列上确定多个突破点,所述突破点两侧的测序序列数目存在显著差异;
- 基于所述多个突破点,在所述参照基因组上确定检验窗口;
- 基于落入所述检验窗口的测序序列,确定第一参数;以及
- 35 基于所述第一参数与预定阈值的差异,确定所述样本基因组,针对所述检验窗口是否存在拷贝数变异。

37、根据权利要求 36 所述的计算机可读介质,其特征在于,在所述参照基因组上确定突破点进一步

包括:

将参考基因组序列划分为多个预定长度的一级窗口, 并确定落入各一级窗口中的测序序列;

针对参考基因组序列上的至少一个位点, 确定落入所述位点两侧相同数目一级窗口中的测序序列数目;

- 5 确定所述位点的 p 值, 所述 p 值表示两侧测序数据数目的显著差异性; 以及
如果所述位点的 p 值小于终止 p 值, 判断所述位点为突破点。

38、根据权利要求 37 所述的计算机可读介质, 其特征在于, 所述落入各一级窗口中的测序序列为唯一比对测序序列。

39、根据权利要求 37 所述的计算机可读介质, 其特征在于, 在所述位点两侧各取 100 个一级窗口。

- 10 40、根据权利要求 37 所述的计算机可读介质, 其特征在于, 所述一级窗口的长度均为 100-200Kbp, 优选 150Kbp。

41、根据权利要求 37 所述的计算机可读介质, 其特征在于, 所述终止 p 值为至多 1.1×10^{-50} 。

42、根据权利要求 37 所述的计算机可读介质, 其特征在于, 确定所述位点两侧测序数据数目的显著差异性 p 值进一步包括:

- 15 针对所述位点, 在所述位点两侧各取相同数目的一级窗口, 并且计算每个一级窗口的相对测序序列数目 R_i , 其中 i 表示一级窗口的编号

对所有一级窗口的相对测序序列数目 R_i 进行游程检验, 以便确定所述位点的 p 值,

其中,

所述相对测序序列数目是通过公式确定的:

$$R_i = \log_2 \left(\frac{r_i}{\bar{r}} \right)$$

- 20 其中 r_i 表示落入第 i 一级窗口的测序序列数目,

$$\bar{r} = \frac{1}{n} \sum_{i=1}^n r_i, \quad n \text{ 表示一级窗口的总数目。}$$

- 25 43、根据权利要求 42 所述的计算机可读介质, 其特征在于, 在对所有一级窗口的相对测序序列数目进行游程检验进一步包括:

对每个一级窗口的相对测序序列数目 R_i 进行 GC 含量校正, 以便获得校正的相对测序序列数目 $\tilde{R}_i$;

基于所述校正的相对测序序列数目, 确定每个一级窗口的标准化的测序序列数目 Z_i ; 以及

对所有一级窗口的标准化的测序序列数目 Z_i 进行游程检验。

44、根据权利要求 43 所述的计算机可读介质, 其特征在于,

- 30 所述校正的相对测序序列数目 $\tilde{R}_i$ 是通过下列步骤获得的:

计算每个一级窗口的 GC 含量;

将 GC 含量以 0.001 为单位划分为多个区域, 并且统计每个区域中相对测序序列数目的平均值 M_s , 其中 s 为 GC 区域的编号;

$$\tilde{R}_i = R_i - M_s;$$

根据公式下列公式确定所述校正的相对测序序列数目 $\tilde{R}_i$:

- 35 所述标准化的测序序列数目 Z_i 是通过下列公式确定的

$$Z_i = (R_i - \tilde{R}_i - \text{mean}) / SD$$

其中,

$$\text{mean} = \frac{1}{n} \sum_{i=1}^n (R_i - \tilde{R}_i)$$

$$SD = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (R_i - \tilde{R}_i - \text{mean})^2}。$$

45、根据权利要求 43 所述的计算机可读介质,其特征在于,基于所述多个突破点,在所述参照基因组上确定检验窗口进一步包括:

- 1) 确定多个候选突破点,其中在所述候选突破点的前后均存在其他突破点;
- 2) 确定每个候选突破点的 p 值,并剔除 p 值最大的候选突破点;
- 3) 对剩余的候选突破点重复步骤 2),直到剩余候选突破点的 p 值均小于终止 p 值,所述剩余候选突破点作为经过筛选的候选突破点;以及
- 4) 确定相邻两个经过筛选的候选突破点之间的区域为检验窗口,

其中,通过下列步骤确定所述候选突破点的 p 值:

将所述候选突破点与相邻的上一个突破点之间的区域作为第一候选区域,将所述候选突破点与相邻的下一个突破点之间的区域作为第二候选区域;

对所述第一候选区域和第二候选区域中所包含一级窗口的标准化的测序序列数目 Z_i 进行游程检验,以便确定所述候选突破点的 p 值,

任选地,所述终止 p 值是通过下列步骤确定的:

利用对照样品的测序结果,重复在参照基因组上确定检验窗口的操作,并记录每次被剔除候选突破点的 p 值,直到候选突破点的数目为零;以及

基于被剔除候选突破点的 p 值分布,确定所述终止 p 值,

任选地,所述终止 p 值为 1.1×10^{-50} 。

46、根据权利要求 45 所述的计算机可读介质,其特征在于,基于落入所述检验窗口的测序序列,确定第一参数进一步包括:

确定所述检验窗口中所包含的所有一级窗口的平均标准化的测序序列数目 $\bar{Z}$,所述平均标准化的测序序列数目 $\bar{Z}$ 作为第一参数。

47、根据权利要求 46 所述的计算机可读介质,其特征在于,所述预定阈值采用 -1.645 作为第一阈值和 1.645 作为第二阈值。

48、根据权利要求 36 所述的计算机可读介质,其特征在于,所述参照基因组序列为选自人类 21 号染色体、18 号染色体、13 号染色体、X 染色体和 Y 染色体的至少一种的序列。