一种拷贝数变异检测方法和系统

技术领域

本发明涉及生物信息学技术领域，尤其涉及一种拷贝数变异（Copy Number Variation，CNV）检测方法和系统。

背景技术

CNV 是基因组结构变异的一种，狭义的 CNV 通常指染色体上 DNA 片段的拷贝数发生了变化。基因组结构变异的类型及原因包括：1.缺失（末端缺失、中间缺失）；2.易位（相互易位、罗伯逊易位）；3.倒位；4.环状染色体 ；5.双着丝粒染色体；6.插入等。广义的 CNV 也包括例如染色体非整倍性和部分非整倍性的结构变异。

目前拷贝数变异的检测方法，主要有高分辨率染色体核型分析、FISH（荧光原位杂交）、Array CGH（比较基因组杂交）、MLPA（多重连接探针扩增技术）和 PCR（Polymerase Chain Reaction，多聚酶链式反应）的方法等，其中遗传学诊断以 FISH 检测为金标准，可以有效地检测出大部分已知染色体缺失或重复。但这些方法普遍存在效率较低，尤其在全基因组水平全面扫描的情况下，资源消耗较大或无法检测未知 CNV 等缺点。

因此，开发一种新的拷贝数变异检测方法，对已知的位点进行鉴定，并对未知的位点进行发现性探索，已经迫在眉睫。

发明内容

本发明要解决的一个技术问题是提供一种拷贝数变异检测方法和系统，能够精确检测包括微缺失/微重复的拷贝数变异。

根据本发明的一个方面，提供一种拷贝数变异检测方法，其特征在于，包括：获取受试样品中至少一部分核酸分子的读序信息；根据读序信息确定唯一比对至（基因组）参考序列的序列标

签；将基因组参考序列划分窗口,统计落入各个窗口的序列标签数目；对各个窗口的序列标签数目进行 GC 校正并根据以对照样本集修正的期望序列标签数目进行修正获得调整后的序列标签数目；以窗口的起点或终点作为分界点，计算两侧由调整后的序列标签数目组成数值群体的差异显著性值，选取显著性值较小的分界点为候选的 CNV 断点；对于每个 CNV 断点至在先 CNV 断点和在后 CNV 断点的两段序列，计算两段序列包含的窗口的调整后的序列标签数目组成的两个数值群体的差异显著性值，每次剔除最不显著的候选 CNV 断点并更新被剔除的候选 CNV 断点左右两个候选 CNV 断点的差异显著性值，循环迭代，直至所有候选 CNV 断点的差异显著性值都小于终止阈值，从而确定 CNV 断点。

可选地，还包括对受试样本中的至少一部分核酸分子进行测序以获得读序信息的步骤。

可选地，每个窗口具有相同的参考序列标签数目（reference unique reads），或者，每个窗口具有相同的长度。

可选地，终止阈值根据由正常样本组成对照样本集获得。

可选地，对各个窗口的序列标签数目进行 GC 校正包括以下的步骤：将窗口按 GC 含量分组，基于组内序列标签的平均数与所有窗口的序列标签平均数获得修正系数，对该窗口序列标签数进行修正获得 GC 校正的序列标签数。

可选地，对照样本集修正的期望序列标签数目由以下方法获得：计算对照集中每个窗口经过 GC 校正的序列标签数与标签总数的比值；基于该比值，获得所有对照样本对应窗口比值的平均数；基于上述平均数和待测样本的序列标签总数，计算待测样本各个窗口序列标签数的期望值。

可选地，在确定 CNV 断点后还包括:对根据 CNV 断点之间的片段进行置信选择的步骤；所述置信选择包括以下步骤：根据调整后的序列标签数目的分布规律，利用对照集确定调整后的序

列标签数目正常的置信区间；当片段内调整后的序列标签数目的均值处于置信区间之外时，认为该 CNV 断点之间的片段确实存在异常。

可选地，序列标签数目符合正态分布，所述置信区间为 95% 置信区间。

可选地，在选择候选的 CNV 断点时进行单条染色体环化或者全基因组水平环化处理。

可选地，还包括：所述受试样品是来自人类样本的样品，所述受试样品包括是羊膜腔穿刺获得的羊水、绒毛活检得到的绒毛、经腹脐静脉穿刺得到的脐带血、自然流产的胎儿组织或人体的外周血；和/或所述受试样品的基因组 DNA 分子采用盐析法、柱层析法、磁珠法、或 SDS 法的 DNA 提取方法获得；和/或对所述受试样品的基因组 DNA 分子采用酶切、雾化、超声、或者 HydroShear 法随机打断得到 DNA 片段；和/或对所述受试样品的基因组 DNA 分子片段进行单向测序或双向测序获得所述 DNA 片段读序信息。

可选地，该方法还包括：在每个受试样品的 DNA 片段上加上不同的标签序列以区别不同的受试样品。

根据本发明的另一方面，提供一种拷贝数变异检测系统，包括：读序获取单元，用于获取据受试样品中至少一部分核酸分子的读序信息；序列标签确定单元，用于根据读序信息确定唯一比对至（基因组）参考序列的序列标签；标签数目统计单元，用于将基因组参考序列划分窗口,统计落入各个窗口的序列标签数目；标签数目调整单元，用于对各个窗口的序列标签数目进行 GC 校正并根据以对照样本集修正的期望序列标签数目进行修正获得调整后的序列标签数目； 候选断点选择单元，用于以窗口的起点或终点作为分界点，计算两侧由调整后的序列标签数目组成数值群体的差异显著性值，选取显著性值较小的分界点为候选的 CNV

断点；断点确定单元，用于对于每个 CNV 断点至在先 CNV 断点和在后 CNV 断点的两段序列，计算所述两段序列包含的窗口的调整后的序列标签数目组成的两个数值群体的差异显著性值，每次剔除最不显著的候选 CNV 断点并更新被剔除的候选 CNV 断点左右两个候选 CNV 断点的差异显著性值，循环迭代，直至所有候选 CNV 断点的差异显著性值都小于终止阈值，从而确定 CNV 断点。

可选地，每个窗口具有相同的参考序列标签数目，或者，每个窗口具有相同的长度。

可选地，终止阈值根据由正常样本组成对照样本集获得。

可选地，标签数目调整单元包括：GC 校正模块，用于将窗口按 GC 含量分组，基于组内序列标签的平均数与所有窗口的序列标签平均数获得修正系数，对该窗口序列标签数进行修正获得 GC 校正的序列标签数；窗口修正模块，用于计算对照集中每个窗口经过 GC 校正的序列标签数与标签总数的比值；基于该比值，获得所有对照样本对应窗口比值的平均数；基于上述平均数和待测样本的序列标签总数，计算待测样本各个窗口序列标签数的期望值，对 GC 校正后的序列标签数以对照样本集修正的期望序列标签数目进行修正获得调整后的序列标签数目。

可选地，该系统还包括断点过滤单元，用于在所述断点确定单元确定 CNV 断点后根据序列标签数目的分布规律，利用对照集确定序列标签数目正常的置信区间；当片段内序列标签数目的均值处于置信区间之外时，认为该 CNV 断点之间的片段确实存在异常。

可选地，调整后的序列标签数目符合正态分布，所述置信区间为 95%置信区间。

可选地，候选断点选择单元在选择候选的 CNV 断点时进行单条染色体环化或者全基因组水平环化处理。

可选地，受试样品是来自人类样本的样品，所述受试样品包括是羊膜腔穿刺获得的羊水、绒毛活检得到的绒毛、经腹脐静脉穿刺得到的脐带血、自然流产的胎儿组织或人体的外周血；和/或所述受试样品的基因组 DNA 分子采用盐析法、柱层析法、磁珠法、或 SDS 法的 DNA 提取方法获得；和/或对所述受试样品的基因组 DNA 分子采用酶切、雾化、超声、或者 HydroShear 法随机打断得到 DNA 片段；

和/或

对所述受试样品的基因组 DNA 分子片段进行单向测序或双向测序获得所述 DNA 片段读序信息。

可选地，通过受试样品 DNA 片段上所加的不同标签序列以区别不同的受试样品。

本发明的拷贝数变异检测方法和系统的一个优点在于，具有临床可行性，可精确检测到包括较小微缺失/微重复区域的拷贝数变异。

附图说明

图 1 示出本发明的拷贝数变异检测方法的一个实施例的流程图；

图 2 示出本发明的拷贝数变异检测方法的另一个实施例的流程图；

图 3 示出本发明的拷贝数变异检测方法的又一个实施例的流程图；

图 4 示出本发明一个实现中对染色体 CNV 分析的简要流程图；

图 5 示出本发明的拷贝数变异检测系统的一个实施例的结构图；

图 6 示出本发明的拷贝数变异检测系统的另一个实施例的结

构图;

图 7A-7H 示出本发明的一个应用例中 8 个样品的检测结果示意图。

具体实施方式

对本文中用到的术语进行说明:

拷贝数变异（copy number variation, CNV）：是指待测样品核酸序列与正常样品核酸序列相比，大于 1kb 的核酸分子的拷贝数发生了变化。拷贝数变异的情形及原因包括：缺失，例如微缺失；插入，例如微插入，微重复，重复，倒置，转座和复杂的多位点变异。

非整倍体：是指与正常样本相比，遗传物质中存在染色体数目的增加或减少，进一步包括整条或部分染色体的增加或减少。本发明中拷贝数变异也包括非整倍体的情形。

测序：获得样品核酸序列信息的过程。测序可通过各种测序方法进行，包括但不限于双脱氧链终止法；优选高通量的测序方法，包括但不限于第二代测序技术或者是单分子测序技术。

第二代测序平台（Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan; 11(1):31-46）包括但不限于 Illumina-Solexa（GA™,HiSeq2000™ 等）、ABI-Solid 和 Roche-454（焦磷酸测序）测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术（True Single Molecule DNA sequencing）, Pacific Biosciences 公司单分子实时测序（single molecule real-time (SMRT™)），以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等（Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 2446 (4)。

测序类型可以为单向（single-end）测序和双向（Pair-end）

测序，测序长度可以为 50bp、90bp、或 100bp。在本发明的一个实施方案中，测序平台为 Illumina/Solexa，测序类型为 Pair-end 测序，得到具有双向位置关系的 100bp 大小的 DNA 序列分子。

本发明的一个实施方案中，测序的测序深度可以依据检测的受试样品染色体变异片段大小确定，测序深度越高，检测的灵敏度越高，即可检出的缺失和重复的片段越小。测序深度可以是 0.1-30×，即总数据量为人类基因组长度的 0.1-30 倍，例如在本发明的一个实施方案中，测序深度为 0.1×，（2.5×108bp）。

读序（reads）：一定长度的核酸序列（一般大于 20bp）,例如由测序仪产生的测序序列的结果；通过序列比对的方法，可定位至参考序列的 特定区域或位置。

序列比对（比对）：是指一个或多个核酸序列与参考序列进行比较的过程。具体常见为将一段较短的核酸序列（如读序）与参考基因组序列相比较，以确定较短核酸序列在参考基因组上的位置。在利用计算机进行序列比对时，序列比对可以通过任何一种序列比对程序，ELAND（efficient local alignment of nucleotide data），SOAP（Short Oligonucleotide Analysis Package）和 BWA（Burrows-Wheeler Aligner）等程序进行。比对成功的认定标准又分为不容错比对（100%匹配）和部分容错比对（小于 100%的匹配）。

序列标签：是指可定位参考序列（例如参考基因组序列）唯一位置的读序（reads）。

参考序列标签（reference unique reads）：是指具有固定长度，且在参考序列（通常是参考基因组）上有唯一比对位置的序列。获得参考序列标签的过程，例如包括，将参考基因组切割为固定长度的序列，将这些序列比对回参考基因组，选择唯一比对到参考基因组的序列为参考唯一比对序列。固定长度依测序仪的测序结果序列长度而定，具体可参考平均长度。不同测序仪得到

的测序结果长度是不同的，具体每一次测序，测序结果的长度也可能不同，该长度的选取存在一定主观和经验因素。

标签序列（index）：是一段特定长度的，起标识性作用的核酸序列。当待测的 DNA 分子来自多个受试样本时，每个样本可以被加上不同的标签序列，以用于在测序过程中进行样品的区分(Micah Hamady, Jeffrey J Walker, J Kirk Harris et al. Error-correcting barcoded primers forpyrosequencing hundreds of samples in multiplex. Nature Methods, 2008, March, Vol.5 No.3)，从而实现同时对多个样品进行测序。标签序列为了区分不同序列，但不影响添加标签序列的 DNA 分子的其他功能。

GC 校正：因为测序批次间/内存在一定的 GC 偏向性，会使基因组中高 GC 或低 GC 区域出现拷贝数偏差，对测序数据基于对照样本集进行 GC 校正得到每个窗口中校正后的相对测序序列数，可以去除此偏向性，提高拷贝数变异检测的精度。

平均数:本文所指平均数一般为算术平均数或中位数。

序列标签数目：序列标签数目可以基于最初的数量统计得到个数统计量，也可以是序列标签个数经其他参数修正得到的修正值，例如可以是一个比值，一些情况下与“拷贝率”可以互换。

受试样品：在一些情形下可被称为待测样品，是指含有核酸分子，而且核酸分子被怀疑发生变异的样品。核酸的类型并不受特别限制，可以是脱氧核糖核酸（DNA），也可以是核糖核酸（RNA），优选 DNA。对于 RNA，可以通过常规手段将其转换为具有相应序列的 DNA，进行后续检测和分析。

对照样本：是相对受试样本而言的，是被认为正常的样本，通常情况下，正常是指表型正常。

对照样本集（对照集）：是指对照样本组成的集合，本发明一个实施中，集合中对照样本的数目要求大于 30。

下面参照附图对本发明进行更全面的描述，其中说明本发明

的示例性实施例。

随着高通量测序技术的不断发展与测序成本的不断降低，测序技术在染色体变异检测方面得到了越来越广泛的应用。

为了提高目前临床上对拷贝数变异检测的技术，本公开设计了一种基于高通量测序技术的进行全基因组水平拷贝数变异筛查的技术方案，具有通量高、特异性高、定位准确的优点。通过获取受试者样品，提取 DNA，进行高通量测序，对获得的数据进行分析，得出检测结果。

图 1 示出本发明的拷贝数变异检测方法的一个实施例的流程图。

如图 1 所示，在步骤 102，获取受试样品中至少一部分核酸分子的读序（reads）信息。可以对受试样本中的至少一部分核酸分子或者全部核酸分子进行测序以获得读序信息。可以获取受试样品的部分核酸分子的读序信息，或者获取全部核酸分子的读序信息。例如，将来自受试样品的基因组 DNA 分子随机打断得到 DNA 片段，对 DNA 片段进行测序，然后获得一定长度的读序。获得的读序的长度可能在一定范围内，可以通过截取操作获得固定长度的读序。DNA 片段的长度可以在 50bp~1500bp，例如 50bp~150bp、150bp ~ 350bp、350bp~500bp、500bp~700bp、700bp ~ 1000bp、或 1000bp ~ 1500bp。例如，可以选择 50bp、90bp、100bp、150bp、300bp、350bp、500bp、700bp、1000bp、1500bp。在一个实施例中，优选 300bp ~ 700bp，更优选 350bp ~ 500bp。读序的长度鉴于测序仪的不同可能会有较大的差别，例如 illumina-solexa 和 life technologies-solid 仪器目前普遍读序长度在 300bp 以内，而 roche-454、传统的 sanger 测序、最新的单分子的测序系统得到的读序长度可接近或超过 1000bp。为了适应唯一比对的需要，在基于读序选择序列标签时，一般会选择读序中 20bp 以上的序列去比对，优选地，在 26bp 以上。

步骤 104，根据读序信息确定唯一比对至（基因组）参考序列的序列标签。例如，将读序的全部或部分序列与基因组参考序列进行比对以获得读序在基因组上的位点信息，得到读序在具体染色体上的位点信息。对于人类受试样本，人类基因组参考序列可以是 NCBI 数据库中的人类基因组参考序列。在本发明的一个实施例中，人类基因组序列是 NCBI 数据库中版本 36（hg18; NCBI Build 36）的人类基因组参考序列，所采用的比对软件是 SOAPaligner/soap2。选择与基因组参考序列唯一比对上的 DNA 片段读序，即只在人类基因组参考序列上比对上一次的读序，也就是唯一比对至（基因组）参考序列的序列标签。

步骤 106，将基因组参考序列划分窗口，统计落入各个窗口的序列标签数目。对于窗口采用以下方法确定：将参考基因组打断为与待检测样品测序长度相同的片段，采用相同的比对软件的相同参数进行比对，筛选出染色体上可被唯一比对的位置；然后每间隔一定长度的可比对位置划定一次窗口。窗口之间可以选择是否有交叉滑动。窗口内应包含的可被唯一比对的位点数与待测样品的测序数据量有关。一般要使待测样品落在每个窗口中的期望 reads 数在 300 条以上，从而保证落在窗口内的 reads 数分布符合泊松分布。例如，假设基因组可被唯一比对的点数为 N，待检测样品的有效 reads 数为 n，每个窗口的期望 reads 数为 E，则参考基因组的每个窗口中应包含 $\frac{N}{n}\times E$ 个可被唯一比对位点。

步骤 108，对各个窗口的序列标签数目进行 GC 校正并根据以对照样本集修正的期望序列标签数目进行修正获得调整后的序列标签数目。在一个实施例中，对各个窗口的序列标签数目进行 GC 校正包括以下的步骤:将窗口按 GC 含量分组，基于组内序列标签的平均数与所有窗口的序列标签平均数获得修正系数，对该窗口序列标签数进行修正获得 GC 校正的序列标签数；对照样本

集修正的期望序列标签数目通过以下方法获得：计算对照集中每个窗口经过 GC 校正的序列标签数与标签总数的比值；基于该比值，获得所有对照样本对应窗口比值的平均数；基于上述平均数和待测样本的序列标签总数，计算待测样本各个窗口序列标签数的期望值

步骤 110，以窗口的起点或终点作为分界点，计算两侧由调整后的序列标签数目组成数值群体的差异显著性值，选取显著性值较小（即差异显著性较大）的分界点为候选的 CNV 断点。例如，在全基因组范围内根据表征每个窗口两侧拷贝数变异的显著性水平的 p 值选取预定数量的窗口作为候选 CNV 断点，获得每个候选 CNV 断点的差异显著性值，即 p 值。

步骤 112，对于每个 CNV 断点至在先 CNV 断点和在后 CNV 断点的两段序列，计算所述两段序列包含的窗口的调整后的序列标签数目组成的两个数值群体的差异显著性值，每次剔除最不显著的候选 CNV 断点并更新被剔除的候选 CNV 断点左右两个候选 CNV 断点的差异显著性值，循环迭代，直至所有候选 CNV 断点的差异显著性值都小于终止阈值，从而确定 CNV 断点。终止阈值通常预先设定。例如，通过对由正常样本组成对照样本集进行分析处理获得该终止阈值。

上述实施例中，通过将获取的读序与基因组参考序列比对，选择唯一比对的读序进行窗口统计，对每个窗口的序列标签数目进行 GC 校正和对照集修正后进行基因显著性差异的循环迭代进行 CNV 断点选择，从而完成 CNV 的检测，能够精确检测到较小的包括微缺失/微重复区域的拷贝数变异。

对于人类样本，受试样品可以是羊膜腔穿刺获得的羊水、绒毛活检得到的绒毛、经腹脐静脉穿刺得到的脐带血、自然流产的胎儿组织或人体外周血中获取的基因组 DNA，可以采用盐析法、柱层析法、磁珠法、SDS 法等常规 DNA 提取方法。在一个实施

例中，优选采用柱层析法，该柱层析法是指血液、组织或细胞经过细胞裂解液和蛋白酶 K 的作用后得到裸露的 DNA 分子，DNA 分子在高盐条件下与硅胶膜结合，在低盐、高 pH 值时 DNA 分子从硅胶膜上洗脱下来。具体原理和方法参见 Tiangen TIANamp Micro DNA Kit （DP316）产品说明书））。

当待测的 DNA 片段来自多个受试样本的 DNA 分子时，每个受试样本的 DNA 片段可以被加上不同的标签序列（index），长度可选 4bp ~ 12bp，以用于在测序过程中进行受试样品的区分 (Micah Hamady, Jeffrey J Walker, J Kirk Harris et al. Error-correcting barcoded primers forpyrosequencing hundreds of samples in multiplex. Nature Methods, 2008, March, Vol.5 No.3)。通过这样的方式，能够同时处理多个受试样品的检测，提高了效率，减小了检测成本。

图 2 示出本发明的拷贝数变异检测方法的另一个实施例的流程图。

步骤 202，对受试样品的基因组 DNA 分子随机打断得到 DNA 片段。DNA 分子随机打断处理可以采用酶切、雾化、超声、或者 HydroShear 法。优选地，采用超声法，例如 Covaris 公司的 S-series（基于 AFA 技术，当由传感器释放的声能/机械能通过 DNA 样品时，溶解气体形成气泡。当能量移除后，气泡破裂并产生断裂 DNA 分子的能力。通过设置一定的能量强度和时间间隔等条件，可将 DNA 分子打断至一定范围的大小。具体原理和方法请参见 Covaris 公司的 S-series 说明书），将 DNA 分子打断为比较集中的一定大小的片段。

步骤 204，对 DNA 片段进行测序获得 DNA 片段测序序列，即读序。测序获得的读序可以在一定长度范围内，可以通过截取操作从 DNA 片段测序序列获得固定长度的读序，在本实施例下文中 DNA 片段测序序列指固定长度的读序。所采用的测序方法

可以为高通量测序方法 Illumina/Hiseq2000、ABI/SOLiD、Roche/454。测序类型可以为 Single-end（单向）测序或 Pair-end（双向）测序，测序长度可以为 50bp~1500bp。在本发明的一个实施例中，测序平台为 Illumina/Hiseq2000，测序类型为 Pair-end 测序，得到具有双向位置关系的 100bp 大小的 DNA 序列分子。对测序深度有一定要求，测序深度可以依据检测的染色体变异片段大小确定，测序深度越高，检测的灵敏度越高，即可检出的缺失和重复的片段越小。在本发明的一个实施例中，人类受试样本的测序量在 2~900 × 10^8 条测序片段之间。

步骤 206，将读序与基因组参考序列进行比对以获得读序在基因组上的位点信息。

步骤 206，选择与基因组参考序列唯一比对上的读序作为序列标签。

步骤 208，统计落入基因组上每个窗口的 DNA 片段唯一比对的序列标签的个数。对于每个受试样品，统计每个窗口所落的序列标签个数（记为 $n_{i,j}$，下标 i 和 j 分别代表窗口编号和样本编号，以作区分，下不复述）。

步骤 210，确定基因组上每个窗口的平均 GC 含量以确定该窗口修正系数，根据修正系数获得每个窗口的序列标签的修正个数。该步骤主要根据每个窗口的 GC 含量对窗口的序列标签个数进行修正，可称为批次修正或 GC 校正。

统计每个窗口所落的序列标签的平均 GC 含量（记为 $GC_{i,j}$）。平均 GC 含量（$GC_{i,j}$）的计算，即计算落在统计窗口内的所有序列标签的平均 GC 含量水平。在计算时，将所有序列标签中的 G 和 C 碱基的个数总和即为 N_{gc}，所有序列标签的总长度记为 L，则 $GC_{i,j}=\frac{N_{gc}}{L}$。

为了修正样品数据量的差异，将各个统计窗口按其 $GC_{i,j}$ 进行

分组，即$GC_{i,j}$相同的窗口划为一组。取各分组中序列标签数$n_{g,j}$的中位数或算术平均数$m_{g,j}$，除全基因组水平序列标签数的中位数或算术平均数m_j，得到修正系数$c_{g,j}$（$c_{g,j}=m_{g,j}/m_j$），下标 g 代表不同分组的 GC 含量。将原始的每个窗口的序列标签数$n_{i,j}$乘以修正系数$c_{g,j}$得到每个窗口内序列标签数的修正值（记为$n_{i,j}'$）

步骤 212，将每个窗口的序列标签的修正个数除以该窗口的个数期望值获得每个窗口的调整后的序列标签数，即拷贝率。其中每个窗口的个数期望值通过由正常样本组成对照集(control set)获得。该步骤主要根据正常样本数据进行每个窗口的序列标签的个数的校正，可以称为窗口校正。

在对照集样本中，定义相对序列标签数百分比（$P_{i,j}$）为窗口内序列标签数（$n_{i,j}'$）和全基因组总序列标签数（N_j）之比，即$P_{i,j}=\frac{n_{i,j}'}{N_j}$；然后计算得到对照集中每个窗口的平均百分比$\overline{P_i}$，$\overline{P_i}=\frac{1}{n}\sum_{j=1}^{n}P_{i,j}$。在受试样本中，每个窗口的拷贝率$r_{i,j}$等于修正序列标签数$n_{i,j}'$除以窗口内的期望值（全基因组总的序列标签数乘以窗口序列标签数百分比），即$r_{i,j}=\frac{n_{i,j}'}{\overline{P_i}\times N_j}$。

在对照集的选择中，建库方法、测序试剂及测序类型等应尽量与待测样品一致，从而提高对照样品对待测样品的校正效果。对照集中的样品应为正常样本，且样本量大于 30。

步骤 214，在全基因组范围内根据表征每个窗口两侧拷贝数变异的显著性差异的 p 值选取预定数量的窗口作为候选 CNV 断点，获得每个候选 CNV 断点的 p 值。

1）候选 CNV 断点选择：针对全基因组的所有窗口，计算每个窗口左右两侧一定数量窗口（窗口数量通常大于 30 或满足检验模型的最低样品量限制，以使得检验模型具有统计意义）的拷

贝率变化差异，根据全基因组范围的显著性差异水平（p 值从小到大），选取一定数量（例如，总窗口数量的 1%）的（与窗口对应）点作为候选 CNV 断点（Breakpoint，即每个 CNV 片段的边界点）。

2）初始化：将排序后的所有断点的集合记为 $B_c = \{b_1, b_2, ...b_k .., b_s\}$，则在每个断点 k 左右两侧，分别存在相邻的断点 k+1、k-1。通过计算 k-1 与 k 之间的拷贝数集合与 k 与 k+1 之间的拷贝率集合之间的显著性差异，得到代表每个断点两侧显著性差异的 p 值。例如，在一个实施例中，选用非参数检验中的游程检验，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，详见：Wald, A. & Wolfowitz, J. On a test whether two samples are from the same population. The Annals of Mathematical Statistics 11, 147-162 (1940)。

步骤 216，每次剔除最不显著的候选 CNV 断点并更新被剔除的候选 CNV 断点左右两个候选 CNV 断点的 p 值，循环迭代，直至所有候选 CNV 断点的 p 值都小于终止 p 值（即，终止阈值），终止 p 值根据对照集获得。

迭代合并：通过不断的循环迭代，每次将最不显著的候选突破点剔除，同时更新左右两个断点的 p 值，直到所有 p 值都小于终止 p 值为止。

终止 p 值的获得：例如，可以在将对照样品进行上述迭代合并的操作，记录每次迭代合并的最大 p 值，迭代合并至一条片段时终止。此时，根据最大 p 值的变化趋势，将最大 p 值变化最剧烈的点（即取 p 值的变化曲线中斜率变化最明显的点（曲率最大的点））对应或之前的那次合并过程的最大 p 值作为终止阈值。在一个实施例中，亦可将迭代合并的终止设置为迭代合并后的片段数等于预知的片段数，例如，当对全基因组进行分析时，对照样品迭代合并后的片段数为 24 个时终止，通过计算此时终止 p

值的平均值，可以有效地获得终止 p 值。

上述步骤 214 和 216 也可以称为片段化。注意：在步骤 214 中 1）和 2）中的窗口和断点选择时，可以考虑进行单条染色体环化或者全基因组水平环化。单条染色体环化是指：在计算染色体起点附近的窗口时，如果左侧的有效窗口数量不足以进行统计检验，则从本条染色体的尾部反向获得足够的窗口进行计算；同理，对于终点附近右侧无法获得足够有效窗口的位置则从染色体前端获得。这一操作使得位于染色体前部和尾部的窗口依然可以计算。全基因组水平环化则是在每条染色体前端有效窗口数不足时索引前一条染色体的尾端，而尾端不足时索引下一条染色体的前端，1 号染色体则和 Y 染色体相连。

在步骤 216 后还可以包括对根据 CNV 断点之间的片段进行置信选择的步骤：根据拷贝率的分布规律，利用对照集确定拷贝率正常的置信区间；当片段内拷贝率的均值处于置信区间之外时，认为该 CNV 断点之间的片段确实存在异常。在一个实施例中，拷贝率符合正态分布，置信区间为 95%置信区间。通过该步骤对片段化结果进行过滤，获得可靠的结果。如果片段的平均 $r_{i,j}$ 小于阈值下限或者大于阈值上限时，将作为阳性结果输出。

阈值选择：统计每个对照样品中窗口的拷贝率分布，根据中心极限定理，读序在窗口内是随机的，所以拷贝率 r 是服从正态分布的，取其显著性水平为 0.05 时的左右分位点。计算其在对照集中的均值，分别作为筛选拷贝率变异时的上下限阈值。

上述实施例中，通过批次修正和窗口校正，提高了检测结果的准确率。通过引入参照集，可以通过扩大参照集来提高精度，以减轻对起始 DNA 量的压力。

图 3 示出本发明的基因组拷贝数变异检测方法的又一个实施例的流程图。在图 3 中，包括了由正常样本组成的对照集的处理流程（3A）和受试样品的处理流程（3B）。对照集主要用于获取

对受试样品进行校正的数据以及受试样品处理的迭代终止条件的终止阈值。

如图 3 所示，流程 3A 包括:

步骤 310A，提取对照样品的 DNA 分子。

步骤 311A，对对照样品的 DNA 分子随机打断为 DNA 片段进行测序，获得对照样品的 DNA 片段测试序列数据，即读序。

步骤 312A，将对照样品的读序与参考基因组进行对比。

步骤 313A，统计窗口唯一比对读序数，即序列标签数目。

步骤 314A，对对照样品进行批次修正。

步骤 315A，通过对照样品获得窗口个数期望值，用于受试样品的窗口校正。

步骤 316A，断点选择与片段化。选择候选 CNV 断点，每次剔除最不显著的候选 CNV 断点并更新被剔除的候选 CNV 断点左右两个候选 CNV 断点的 p 值，循环迭代，使最终的片段剩余预定个数（例如 24 个）时停止。

步骤 317A，确定终止阈值。通过计算此时终止 p 值的平均值，可以有效地获得终止 p 值，作为受试样品迭代终止条件的终止阈值。

流程 3B 包括:

步骤 310B，提取受试样品的 DNA 分子。

步骤 311B，对受试样品的 DNA 分子随机打断为 DNA 片段进行测序，获得受试样品的 DNA 片段的读序。

步骤 312B，将受试样品的 DNA 片段的读序与参考基因组进行对比。

步骤 313B，去统计窗口唯一比对读序数，即序列标签数。

步骤 314B，对受试样品进行批次修正。

步骤 315B，根据对照样品获得的窗口个数期望值对受试样品进行窗口校正。

步骤 316B，断点选择与片段化。

步骤 317B，进行结果过滤。

在对照集的选择中，建库方法、测序试剂及测序类型等应尽量与受试样品一致，从而提高对照样品对受试样品的校正效果。对照集中的样品应为正常样本，且样本量大于 30。

图 4 示出本发明一个实现中染色体 CNV 分析的简要流程图。

如图 4 所示，步骤 401，DNA 提取及测序：根据 Tiangen DP327-02 Kit 操作手册提取基因组 DNA 后，基本按照 Illumina/Hiseq2000 标准建库流程进行建库。在这个过程中，本身集中于 500bp 的 DNA 分子两端被加上测序所用接头，每个样本被加上不同的标签序列（index），从而在一次测序得到的数据中可以使多个样本的数据区分开。

步骤 402，序列比对：利用测序方法 Illumina/Hiseq2000 测序（用其它测序方法如 ABI/SOLiD 能达到相同或相近的效果），每个样本得到一定长度片段的 DNA 读序，将其与 NCBI 数据库中的标准人类基因组参考序列进行 SOAP2 比对，得到读序定位于基因组相应位置的信息。为避免重复序列对 CNV 分析的干扰，只选取与人类基因组参考序列唯一比对的读序，即只在人类基因组参考序列上比对上一次的读序，也即序列标签数，作为后续 CNV 分析的有效数据。

步骤 403，PSCC 分析。运用本申请人自主开发的针对全基因组拷贝数变异检测的一系列生物信息学方法（下称 PSCC），在对受试样本进行批次性差异修正，根据对照集（control set），对受试样品进行窗口校正（correction）、标准化（Normalization）和拷贝数片段化（segmentation）。

步骤 404，基于步骤 403 中得到的拷贝数片段进行 CNV 分析，将受试样品拷贝率≤0.7 和≥1.3 分别作为片段缺失和重复的检测阈值，分析得到全基因组水平拷贝数变异片段，然后进行结果的可

视化。

上述实施例中，实现的软件算法是一种由深圳华大基因研究院开发针对全基因组拷贝数变异检测的一系列程序，统称为PSCC。它能够通过新一代测序技术产生的数据，将受试样本进行批次修正，然后和对照集合进行数据校正、标准化和片段化，估算出受试样本的拷贝数变异程度和大小。在较低测序深度（50M 条测序短序列）的时候可检测出 0.5Mb 左右的单个拷贝数变异（CNV）片段。

图 5 示出本发明的基因组拷贝数变异检测系统的一个实施例的结构图。如图 5 所示，该系统包括：读序获取单元 51，获取受试样品中至少一部分核酸分子的读序信息；序列标签确定单元 52，确定唯一比对至（基因组）参考序列的读序作为序列标签；标签数目统计单元 53，将基因组参考序列划分窗口，统计落入各个窗口的序列标签数目；标签数目调整单元 54，对各个窗口的序列标签数目进行 GC 校正并根据以对照样本集修正的期望序列标签数目进行修正获得调整后的序列标签数目；候选断点选择单元 55，以窗口的起点或终点作为分界点，计算两侧由调整后的序列标签数目组成数值群体的差异显著性值，选取显著性值较小（即差异显著性较大）的分界点为候选的 CNV 断点；断点确定单元 56，对于每个 CNV 断点至在先 CNV 断点和在后 CNV 断点的两段序列，计算所述两段序列包含的窗口的调整后的序列标签数目组成的两个数值群体的差异显著性值，每次剔除最不显著的候选 CNV 断点并更新被剔除的候选 CNV 断点左右两个候选 CNV 断点的差异显著性值，循环迭代，直至所有候选 CNV 断点的差异显著性值都小于终止阈值，从而确定 CNV 断点；该终止阈值可以根据由正常样本组成对照样本集获得。标签数目统计单元 53 划分窗口时，可以使得每个窗口具有相同的参考序列标签数目（reference unique reads），或者，使得每个窗口具有相同的长度。

在一个实施例中，候选断点选择单元 55 在选择候选的 CNV 断点时进行单条染色体环化或者全基因组水平环化处理

上述实施例中，序列标签确定单元根据读序确定唯一比对至（基因组）参考序列的序列标签，标签数目调整单元对各个窗口的序列标签数目进行修正，由候选断点选择单元和断点确定单元进行基因显著性差异的循环迭代进行 CNV 断点选择，从而完成 CNV 的检测，能够精确检测到包括较小的微缺失/微重复区域的拷贝数变异。

图 6 示出本发明的拷贝数变异检测系统的另一个实施例的结构图。如图 6 所示，该系统包括：读序获取单元 51、序列标签确定单元 52、标签数目统计单元 53、标签数目调整单元 64、候选断点选择单元 55 和断点确定单元 56。读序获取单元 51、序列标签确定单元 52、标签数目统计单元 53、候选断点选择单元 55、和断点确定单元 56 可以参见图 5 的具体描述，为简洁起见在此不再详细描述。标签数目调整单元 64 包括 GC 校正模块 641 和窗口修正模块 642。其中，GC 校正模块 641，将窗口按 GC 含量分组，基于组内序列标签的平均数与所有窗口的序列标签平均数获得修正系数，对该窗口序列标签数进行修正获得 GC 校正的序列标签数；窗口修正模块 642，计算对照集中每个窗口经过 GC 校正的序列标签数与标签总数的比值；基于该比值，获得所有对照样本对应窗口比值的平均数；基于上述平均数和待测样本的序列标签总数，计算待测样本各个窗口序列标签数的期望值，对 GC 校正后的序列标签数以对照样本集修正的期望序列标签数目进行修正获得调整后的序列标签数目，也称拷贝率。

根据本发明的一个实施例，该系统还包括断点过滤单元 67，在断点确定单元确定 CNV 断点后根据拷贝率的分布规律，利用对照集确定拷贝率正常的置信区间；当片段内拷贝率的均值处于置信区间之外时，认为该 CNV 断点之间的片段确实存在异常。

在一个实施例中，序列标签数目符合正态分布，所述置信区间为95%置信区间。

在一个实施例中，受试样品是来自人类样本的样品，所述受试样品的基因组 DNA 分子是羊膜腔穿刺获得的羊水、绒毛活检得到的绒毛、经腹脐静脉穿刺得到的脐带血、自然流产的胎儿组织或人体外周血中获取的基因组 DNA；受试样品的基因组 DNA 分子采用盐析法、柱层析法、磁珠法、或 SDS 法的 DNA 提取方法获得；对受试样品的基因组 DNA 分子为采用酶切、雾化、超声、或者 HydroShear 法随机打断得到 DNA 片段；DNA 片段测序序列通过对受试样品的基因组 DNA 分子片段进行单向测序或双向测序获得。可以通过受试样品 DNA 片段上所加的不同标签序列以区别不同的受试样品。

对于图 5 至图 6 中各个单元的功能，可以参考上文中关于本发明方法的实施例中对应部分的说明，为简洁起见，在此不再详述。

本领域技术人员将意识到硬件、固件和软件配置在这些情况下的可替换性，以及如何最好地实现每个特定应用地该功能。

下面将结合应用例对本发明的实施例进行详细描述，但是本领域技术人员将会理解，下列应用例仅用于说明本发明，而不应视为限定本发明的范围。应用例中未注明具体条件者，按照常规条件或制造商建议的条件进行。所用试剂或仪器未注明生产厂商者，均为可以通过市场获得的常规产品。以下括号内为各个试剂或试剂盒的厂家货号。所使用的测序用的接头和标签序列 index 来源于 Illumina 公司的 Multiplexing Sample Preparation Oligonutide Kit。

应用例一、2 个染色体数目异常及 6 个微缺失样品检测

1、DNA 提取:

按照 Tiangen 的 TIANamp Micro DNA Kit （DP316）操作

流程提取 8 例样品（以下简称 Sample1、Sample2、Sample3...Sample8）的 DNA，所提取 DNA 按照修改后的 Illumina/Hiseq2000 标准建库流程进行建库，在本身集中于 500bp 的 DNA 分子两端被加上测序所用接头，每个样本被加上不同的标签序列（index），然后与 Flowcell 表面互补接头杂交，在一定条件下使核酸分子成簇生长，然后在 Illumina Hiseq2000 上通过双末端测序，得到长度为 100bp 的 DNA 片段序列。

具体的，将获自上述羊水样品的约 100ng（Quant-IT dsDNA HS Assay kit 定量）的 DNA，进行修改后的 Illumina/Hiseq2000 标准流程建库，具体流程参照现有技术（参见 http://www.illumina.com/提供的 Illumina/Solexa 标准建库说明书）。经 2100Bioanalyzer (Agilent)确定 DNA 文库大小及插入片段为约 500bp，QPCR 精确定量后可上机测序。

2.测序：本实施例中，对于获自上述 8 例样品的 DNA 样本按照 Illumina/Solexa 官方公布的 ClusterStation 和 Hiseq2000（PEsequencing）说明书进行操作，使每个样品得到约 5G 数据量进行上机测序，每个样本根据所述 index 标签区进行分开。利用比对软件 SOAP2，将测序所得 DNA 序列与 NCBI 数据库中版本 36（hg18; NCBIBuild36）的人类基因组参考序列进行比对，得到所测 DNA 序列定位于基因组相应位置的信息。

3.数据分析

a）基本统计：统计每个窗口所落的序列标签个数（即唯一比对的 reads 数，记为 $n_{i,j}$，下标 i 和 j 分别代表窗口编号和样本编号，以作区分，下不复述）和平均 GC 含量（记为 $GC_{i,j}$）。

b) 批次修正：为了修正样品数据量的差异，将各个统计窗口按落入其中的 reads 平均 GC 分组，取各分组的中位数或算术平均数除全基因组水平的中位数或算术平均数之间得到修正系数 c_g，下标 g 代表不同分组的 GC 含量。将原始的每个窗口的 reads 个

数$n_{i,j}$乘以修正系数c_g得到每个窗口内序列标签数的修正值（记为$n_{i,j}^{'}$）。

c）数据校正：选取 90 例 YH 群体的测序数据作为对照集，定义相对序列标签数百分比（$P_{i,j}$）为窗口内测序序列标签数（$n_{i,j}^{'}$）和全基因组总序列标签数（N_j）之比，即$P_{i,j}=\frac{n_{i,j}^{'}}{N_j}$。然后计算得到每个窗口的平均百分比$\overline{P_i}$，$\overline{P_i}=\frac{1}{n}\sum_{j=1}^{n}P_{i,j}$。在受试样本中，每个窗口的拷贝率$r_{i,j}$等于修正修正序列标签数$n_{i,j}^{'}$除以窗口内的期望值（全基因组总序列标签数乘以窗口序列标签数百分比），即$r_{i,j}=\frac{n_{i,j}}{\overline{P_i}\times N_j}$。

d）片段化

①候选断点选择：针对全基因组的所有$r_{i,j}$，计算其左右两侧 100 个窗口的拷贝率变化差异，根据全基因组范围的显著水平（P 值从小到大），选取最显著的 10000 个点作为候选 CNV 断点。

②初始化：将排序后的所有断点的集合记为$B_c=\{b_1,b_2,....b_k..,b_s\}$，则在每个断点 k 左右两侧，分别存在相邻的断点 k+1、k-1。通过计算 k-1 与 k 之间的拷贝率集合与 k 与 k+1 之间的拷贝率集合之间的显著性差异（本实施例中选用非参数检验中的游程检验），得到代表每个断点两侧显著性差异的 p 值。

③迭代合并：通过不断的循环迭代，每次将最不显著的候选断点剔除，同时更新左右两个断点的 p 值，直到所有 p 值都小于 $1E^{-50}$ 时为止。

注意：在步骤①和②中的窗口和断点选择时，进行了全基因组水平环化，从而使得位于染色体前部和尾部的窗口依然可以计算。

e）阈值与过滤：对片段化过后的结果进行过滤。如果片段

的平均 $r'_{i,j}$ 小于 0.7 或者大于 1.3 时，作为阳性结果输出。

f）结果可视化。

4. 结果统计

对于 8 个样品，详细的检测结果和验证结果如下表 1 所示。

其中验证结果通过 CGH 芯片（比较基因组杂交芯片）得出。实验使用 Human Genome CGH Microarray Kit，（Agilent Technologies Inc.），完全按照厂商的使用说明进行操作。

样品编号	检测结果	长度	验证结果	判断结果
Sample1	T7	159.14M	T7	一致
Sample2	XYY	59.37M	XYY	一致
Sample3	chr5:1.. 36877903 缺失 chr18: 39024931.. 76038278 增加	36.88M 37.01M	5p15.3~p13.2 (183931~36816731)×1 ； 18p12.3~q23 (39086755~76067279)×3	一致
Sample4	chr5:1..17710089 缺失	17.71M	5p15.33~p15.1×1	一致
Sample5	chr15:21236149..26219186 缺失	4.98M	15q11.2~q13.1×1	一致
Sample6	chr1:1..5065299 增加	5.07M	1p36.33~p36.32×3	一致
Sample7	chr5:105562861..106156933 缺失	0.59M	5q21.3×1	一致
Sample8	chr9:6851755..7248416 缺失	0.40M	9p24.1×1	一致

表 1. 8 例样品的 CNV 结果 其中 chr 代表染色体，T7 代表 7 号染色体三体型。XYY 未性染色体三体异常。

图 7A-7H 示出 8 个样品的检测结果示意图。

从上表 1 和图 7A-7H 可以看到，小至 0.4M 的微缺失片段，大到整条染色体数目异常，本方法都可精确的进行检测和定位，证明其检测效率和精度都较为优秀。

与目前已报道的拷贝数变异检测的分析方法对比，本公开的优越性主要有一下几点:

（1）分辨率，使用 50M 左右的数据，可精确检测到 0.5M 的微缺失区域。

（2）可扩展性。除了通过增加测序量之外，可以通过扩大参照集来提高精度，以减轻对起始 DNA 量的压力。

（3）更稳定，更加全面。已报道文章中，并无明确指出自身的操作细节，而本发明涉及数据批次修正，群体校正，片段化条件优选等的各个方面。

本发明对适用人群进行全基因组拷贝数变异检测分析，有利于提供遗传咨询和提供临床决策依据，对微缺失综合征患者进行准确的病理判定。本发明适用人群可以是所有微缺失患者或潜在微缺失携带者，适用人群举例仅用于说明本发明，而不应为限定本发明的范围。

本发明的描述是为了示例和描述起见而给出的，而并不是无遗漏的或者将本发明限于所公开的形式。很多修改和变化对于本领域的普通技术人员而言是显然的。选择和描述实施例是为了更好说明本发明的原理和实际应用，并且使本领域的普通技术人员能够理解本发明从而设计适于特定用途的带有各种修改的各种实施例。

权 利 要 求

1. 一种拷贝数变异检测方法，其特征在于，包括:

获取受试样品中至少一部分核酸分子的读序信息;

根据所述读序确定唯一比对至基因组参考序列的序列标签;

将基因组参考序列划分窗口，统计落入各个窗口的序列标签数目;

对各个窗口的序列标签数目进行 GC 校正并根据以对照样本集修正的期望序列标签数目进行修正获得调整后的序列标签数目;

以窗口的起点或终点作为分界点，计算两侧由调整后的序列标签数目组成数值群体的差异显著性值，选取显著性值较小的分界点为候选的 CNV 断点;

对于每个 CNV 断点至在先 CNV 断点和在后 CNV 断点的两段序列，计算所述两段序列包含的窗口的调整后的序列标签数目组成的两个数值群体的差异显著性值，每次剔除最不显著的候选 CNV 断点并更新被剔除的候选 CNV 断点左右两个候选 CNV 断点的差异显著性值，循环迭代，直至所有候选 CNV 断点的差异显著性值都小于终止阈值，从而确定 CNV 断点。

2. 根据权利要求 1 所述的方法，其特征在于，还包括对受试样本中的至少一部分核酸分子进行测序以获得读序信息的步骤。

3. 根据权利要求 1 所述的方法，其特征在于，每个窗口具有相同的参考序列标签数目，或者，每个窗口具有相同的长度。

4. 根据权利要求 1 或 2 所述的方法，其特征在于，所述终止阈值根据由正常样本组成对照样本集获得。

5. 根据权利要求 1 所述的方法，其特征在于，所述对各个窗口的序列标签数目进行 GC 校正包括以下的步骤:

将窗口按 GC 含量分组，基于组内序列标签的平均数与所有窗口的序列标签平均数获得修正系数，对该窗口序列标签数进行

修正获得 GC 校正的序列标签数;

和/或

对照样本集修正的期望序列标签数目由以下方法获得:

计算对照集中每个窗口经过 GC 校正的序列标签数与序列标签总数的比值;基于该比值,获得所有对照样本对应窗口比值的平均数;基于上述平均数和待测样本的序列标签总数,计算待测样本各个窗口序列标签数的期望值。

6. 根据权利要求 5 所述的方法,其特征在于,在选择候选的 CNV 断点时进行单条染色体环化或者全基因组水平环化处理。

7. 根据权利要求 1 或 2 所述的方法,其特征在于,在确定 CNV 断点后还包括:

对根据 CNV 断点之间的片段进行置信选择的步骤;

所述置信选择包括以下步骤:根据调整后的序列标签数目的分布规律,利用对照集确定调整后的序列标签数目正常的置信区间;当片段内调整后的序列标签数目的均值处于置信区间之外时,认为该 CNV 断点之间的片段确实存在异常。

8. 根据权利要求 7 所述的方法,其特征在于,所述调整后的序列标签数目符合正态分布,所述置信区间为 95%置信区间。

9. 根据权利要求 1 至 8 中任意一项所述的方法,其特征在于,还包括:

所述受试样品是来自人类样本的样品,所述受试样品包括是羊膜腔穿刺获得的羊水、绒毛活检得到的绒毛、经腹脐静脉穿刺得到的脐带血、自然流产的胎儿组织或人体外周血;

和/或

所述受试样品的基因组 DNA 分子采用盐析法、柱层析法、磁珠法、或 SDS 法的 DNA 提取方法获得;

和/或

对所述受试样品的基因组 DNA 分子采用酶切、雾化、超声、

或者 HydroShear 法随机打断得到 DNA 片段;

和/或

对所述受试样品的基因组 DNA 分子片段进行单向测序或双向测序获得所述 DNA 片段读序。

10. 根据权利要求 1 所述的方法，其特征在于，还包括:

在每个受试样品的 DNA 片段上加上不同的标签序列以区别不同的受试样品。

11. 一种拷贝数变异检测系统，其特征在于，包括:

读序获取单元，用于获取受试样品中至少一部分核酸分子的读序信息;

序列标签确定单元，用于根据所述读序信息确定唯一比对至基因组参考序列的序列标签;

标签数目统计单元，用于将基因组参考序列划分窗口，统计落入各个窗口的序列标签数目;

标签数目调整单元，用于对各个窗口的序列标签数目进行 GC 校正并根据以对照样本集修正的期望序列标签数目进行修正获得调整后的序列标签数目;

候选断点选择单元，用于以窗口的起点或终点作为分界点，计算两侧由调整后的序列标签数目组成数值群体的差异显著性值，选取显著性值较小的分界点为候选的 CNV 断点;

断点确定单元，用于对于每 CNV 断点至在 CNV 断点和在后 CNV 断点的两段序列，计算所述两段序列包含的窗口的调整后的序列标签数目组成的两个数值群体的差异显著性值，每次剔除最不显著的候选 CNV 断点并更新被剔除的候选 CNV 断点左右两个候选 CNV 断点的差异显著性值，循环迭代，直至所有候选 CNV 断点的差异显著性值都小于终止阈值，从而确定 CNV 断点。

12. 根据权利要求 11 所述的系统，其特征在于，每个窗口具有相同的参考序列标签数目，或者，每个窗口具有相同的长度。

13. 根据权利要求 11 所述的系统，其特征在于，所述终止阈值根据由正常样本组成对照样本集获得。

14. 根据权利要求 11 所述的系统，其特征在于，所述标签数目调整单元包括:

GC 校正模块，用于将窗口按 GC 含量分组，基于组内序列标签的平均数与所有窗口的序列标签平均数获得修正系数，对该窗口序列标签数进行修正获得 GC 校正的序列标签数;

窗口修正模块，用于计算对照集中每个窗口经过 GC 校正的序列标签数与序列标签总数的比值；基于该比值，获得所有对照样本对应窗口比值的平均数；基于上述平均数和待测样本的序列标签总数，计算待测样本各个窗口序列标签数的期望值，对 GC 校正后的序列标签数以对照样本集修正的期望序列标签数目进行修正获得调整后的序列标签数目。

15. 根据权利要求 11 所述的系统，其特征在于，还包括断点过滤单元，用于在所述断点确定单元确定 CNV 断点后根据序列标签数目的分布规律，利用对照集确定序列标签数目正常的置信区间；当片段内序列标签数目的均值处于置信区间之外时，认为该 CNV 断点之间的片段确实存在异常。

16. 根据权利要求 15 所述的系统，其特征在于，所述序列标签数目符合正态分布，所述置信区间为 95%置信区间。

17. 根据权利要求 11 至 16 中任意一项所述的系统，其特征在于，所述受试样品是来自人类样本的样品，所述受试样品包括是羊膜腔穿刺获得的羊水、绒毛活检得到的绒毛、经腹脐静脉穿刺得到的脐带血、自然流产的胎儿组织或人体的外周血;

和/或

所述受试样品的基因组 DNA 分子采用盐析法、柱层析法、磁珠法、或 SDS 法的 DNA 提取方法获得;

和/或

对所述受试样品的基因组 DNA 分子采用酶切、雾化、超声、或者 HydroShear 法随机打断得到 DNA 片段;

和/或

对所述受试样品的基因组 DNA 分子片段进行单向测序或双向测序获得所述 DNA 片段的读序。

18. 根据权利要求 11 所述的系统，其特征在于，通过受试样品 DNA 片段上所加的不同标签序列以区别不同的受试样品。

19. 根据权利要求 11 所述的系统，其特征在于，候选断点选择单元在选择候选的 CNV 断点时进行单条染色体环化或者全基因组水平环化处理。