(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
2016 年 9 月 15 日 (15.09.2016)

WIPO | PCT

(10) 国际公布号
WO 2016/141516 A1

(51) 国际专利分类号:
C12Q 1/68 (2006.01)

(21) 国际申请号: PCT/CN2015/073816

(22) 国际申请日: 2015 年 3 月 6 日 (06.03.2015)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: **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: **孙宇辉 (SUN, Yuhui)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **曹红志 (CAO, Hongzhi)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **仝欣 (TONG, Xin)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **李振宇 (LI, Zhenyu)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **曹丹丹 (CAO, Dandan)**; 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: **北京清亦华知识产权代理事务所（普通合伙） (TSINGYIHUA INTELLECTUAL PROPERTY LLC)**; 中国北京市海淀区清华园清华大学照澜院商业楼 301 室, Beijing 100084 (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, BN, 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, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, 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。

(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, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, 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, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

(54) **Title**: METHOD FOR ACQUIRING SPECIFIC SEQUENCE OF OFFSPRING, AND METHOD AND DEVICE FOR DETECTING DENOVO MUTATION OF OFFSPRING

(54) **发明名称** : 获取子代特异性序列、检测子代新突变的方法和装置

(57) **Abstract**: Disclosed is a method for acquiring a specific sequence of an offspring. The method comprises: acquiring a first sequencing result and a second sequencing result, wherein the first sequencing result is a genome sequencing result of an offspring and the second sequencing result is a genome sequencing result of a parent of the offspring; acquiring a first subread set by using reads in the first sequencing result and acquiring a second subread set by using the second sequencing result, wherein the first subread set consists of multiple first subread subsets, the second subread set consists of multiple second subread subsets, each first subread subset is formed by all subreads that occur in a read in the first sequencing result and that have a length of K each, and each second subread subset is formed by all subreads that occur in a read in the second sequencing result and that have a length of K each; comparing subreads in the first subread set and the second subread set, and acquiring subreads that are only comprised in the first subread set, wherein the subreads that are only comprised in the first subread set are called specific subreads of the offspring; and extracting reads corresponding to the specific subreads of the offspring from the first sequencing result, wherein the extracted reads are specific sequences of the offspring. Also disclosed are a method and device for detecting a denovo mutation of an offspring.

(57) **摘要**: 本发明公开一种获取子代特异性序列的方法，包括：获取第一和第二测序结果，第一和第二测序结果分别为子代的和子代的父母的基因组测序结果；分别利用第一和第二测序结果中的读段，获得第一和第二亚读段集，第一和第二亚读段集各自由多个第一和第二亚读段子集构成，一个第一或第二亚读段子集分别由第一或第二测序结果中的一个读段中出现的全部长度为 K 的亚读段组成；比较第一和第二亚读段集中的亚读段，获得仅包含于第一亚读段集的亚读段，称之为子代特异性亚读段；从第一测序结果中提取子代特异性亚读段对应的读段，提取得的读段为子代特异性序列。本发明还公开检测子代新突变的方法及装置。

WO 2016/141516 A1

获取子代特异性序列、检测子代新突变的方法和装置

技术领域

本发明涉及生物信息领域，具体的，本发明涉及一种获取子代特异性序列的方法、一种检测子代新突变的方法以及一种检测子代新突变的装置。

背景技术

人类是二倍体生物，由父亲的精子和母亲的卵子结合发育而来。受精卵的遗传信息绝大多数都是来自于父母，但是不排除极个别 DNA 序列在受精卵分裂之前就发生了突变，从而在子代中存在既不同于父亲也不同于母亲的 DNA 突变序列，可称为子代新突变，也就是我们常说的”denovo mutation”（denovo 突变）。如果这些序列发生在基因区，尤其是功能蛋白的编码区，那么有可能影响子代的性状和健康状况，比如，自闭症、精神分裂症等。因此，能否精确地、高灵敏度检测 denovo mutation 对于医学研究，尤其是辅助疾病解读具有非常重大的意义。

高通量测序在近几年内得到极大发展，测序成本急剧下降，无论是欧美发达国家，还是中印等发展中国家都开始大规模的个体全基因组高深度测序获取大量个体基因组数据并分析解读，包括进行疾病研究。目前，检测子代个体 denovo 突变的思路，都是分别检测出子代及其父母或者其家系成员各自的变异信息，接着再比较判断出只存在于子代的真实变异。目前的这种思路或方法，直接又简单，但是严重依赖于测序获得的读段（reads）的长短及其比对定位时的准确度，对于 reads 较短，例如只有不到 100bp，利用这些 reads 检测在一些高突变、高杂合或者高重复区域发生的突变，检测精确度一般较低。

发明内容

依据本发明的一方面，本发明提供一种获取子代特异性序列的方法，该方法包括：获取第一测序结果和第二测序结果，第一测序结果为目标子代的基因组测序结果，所述第二

测序结果为该子代的父亲和母亲的基因组测序结果，第一测序结果包括多个长度为 K_1 的读段，第二测序结果包括多个长度为 K_2 的读段；分别利用第一测序结果和第二测序结果中的读段，获得第一亚读段集和第二亚读段集，第一亚读段集由多个第一亚读段子集构成，一个第一亚读段子集由第一测序结果中的一个读段中出现的全部长度为 K 的亚读段组成，第二亚读段集由多个第二亚读段子集构成，一个第二亚读段子集由第二测序结果中的一个读段中出现的全部长度为 K 的亚读段组成；将第一亚读段集的亚读段和第二亚读段集中的亚读段进行比较，获得仅包含于第一亚读段集的亚读段，将仅包含于第一亚读段集的亚读段称为子代特异性亚读段；从第一测序结果中提取子代特异性亚读段对应的读段，提取得的读段为所称的子代特异性序列；其中，$K < K_1$，$K < K_2$。图 1 示意该方法构思。获取基因组测序结果包括对至少一部分基因组序列进行测序文库构建，以及将构建得的测序文库进行上机测序，测序文库的类型和制备方法可根据所选用的测序方法进行，可选用的测序方法根据来自的测序平台包括但不限于 CG(Complete Genomics) CGA、Illumina/ Solexa、Life Technologies/ Ion Torrent 和 Roche 454，依据所选测序平台进行单端或多端、链状或环状测序文库的制备。所称的子代的父亲和母亲指子代的生物学父母，为子代遗传信息的来源个体。目标子代和其父母的基因组测序结果可以通过共同测序同时获得，也可以分别测序获得，例如，在文库构建时引入标签以区分来自子代和来自父母的核酸，将构建好的文库混合上机测序，再依据标签区分开子代的测序结果和其父母的测序结果。根据本发明的一个实施例，对子代测序文库及其父母测序文库分别进行双末末端测序，获得相应的测序结果，两个测序结果都包含多对读段对，一对读段对的两个读段分别来自一个插入片段的两端，所说的插入片段通常是将来自目标个体的基因组经过打断获得的或者为游离 DNA 片段，测序文库为适于在某一测序平台上测序的带测序接头的插入片段。所称的亚读段对应的读段指其上出现该亚读段的读段，亚读段与读段通常非唯一对应，即一般一个亚读段对应多个读段。

根据本发明的另一个实施例，所述第一测序结果和所述第二测序结果均不包括符合任一以下描述的读段，含 N 超过 10%、质量值低于 10、被接头污染，其中，N 指读段中的不确定的碱基，质量值为测序平台赋予读段的值，质量值为-10*lg (p)，这里，p 为测错的概率，当一条 reads 某位置出错概率为 0.1 时，其质量值为 10，被接头污染指 reads 中包含测序接头。去掉这些 reads，能使测序结果包含的 reads 整体质量提高，也利于减少后续快速检测分析对机器内存的需求。

所称一个读段中出现的每个长度为 K 的亚读段指出现在一个序列中每个亚序列，一个序列中的出现的亚序列的数目为（序列的长度-亚序列的长度+1），例如，长度 7bp 的读段 AGAGAGT 中出现的 3bp 的亚读段为 AGA、GAG、AGA、GAG 和 AGT，共 5（即 7-3+1）个 3bp 的亚读段，这 5 个亚读段即组成一个所说的一个亚读段子集。将长度为 K 的亚读段称为 Kmer。根据本发明的一个实施例，所述分别利用第一测序结果和第二测序结果中的读段，获得第一亚读段集和第二亚读段集，包括，利用第一测序结果中的全部读段，获得全部第一亚读段子集，全部第一亚读段子集构成所述第一亚读段集，利用第二测序结果中的全部读段，获得全部第二亚读段子集，全部第二亚读段子集构成所述第二亚读段集。再具体的，所述分别利用第一测序结果和第二测序结果中的读段，获得第一亚读段集和第二亚读段集，包括，从第一测序结果中的读段 i 的一端的第 j 个核苷酸开始，沿另一端方向拷贝该读段的 K 个连续核苷酸为一个 Kmer，i 取遍{1，2，…，m-1，m}中的数值，对应每个 i，j 取遍{1，2，…，（K_1-K），（K_1-K+1）}中的数值，以获得所述第一亚读段集，i 为第一测序结果中的读段的编号，m 为第一测序结果包含的读段的总数，第二亚读段集的获取也类似的，从第二测序结果中的读段 w 的一端的第 z 个核苷酸开始，沿另一端方向拷贝该读段的 K 个连续核苷酸为一个 Kmer，w 取遍{1，2，…，n-1，n}中的数值，对应每个 w，z 取遍{1，2，…，（K_2-K），（K_2-K+1）}中的数值，以获得所述第二亚读段集，w 为第二测序结果中的读段的编号，n 为第二测序结果包含的读段的总数。在机器上进行上述获取亚

读段的过程即包括输入读段，利用哈希算法（散列算法），把每个输入变换成固定长度的输出，该输出为散列值，即亚读段。这种转换是一种压缩映射，即散列值的空间通常远小于输入（预映射）的空间。将读段转换为 Kmer，利于快速比较及计算检出特异性 Kmer。基于 Kmer 的分析对内存的需求与 kmer 的数量和种类有关，如果 kmer 过长，kmer 的种类就很多，利用机器存储分析会占用较多机器内存，而如果 Kmer 过短，Kmer 的数目就很多，利用机器存储分析也会占用较多机器存储内存，kmer 的数量和种类与 Kmer 的长度相关，即与 K 的大小有关。在本发明的一个实施例中，第一测序结果和第二测序结果包含的读段的长度都不小于 50bp，即 $K_1 \geq 50bp$、$K_2 \geq 50bp$，较佳的，可设置亚读段的长度为 19~31bp，使对机器内存的要求较低，且使后续的比较分析能快速进行。

根据本发明的一个实施例，在提取对应的读段之前，对子代特异性 Kmer 进行过滤，其中包括，过滤掉频数小于 2 的子代特异性 Kmer。一亚读段在一个集合中的频数指该亚读段在这个集合中的个数。也可以采用以下实施方式达到与该实施例一样的效果，在获得第一亚读段集和第二亚读段集之后，过滤掉各 Kmer 集中的频数小于 2 的 Kmer，如此，除了能同样降低最终获得的特异性 reads 的假阳性，还能减少下一步的比较运行所需的时间。

根据本发明的一些实施例，该方法还包括：在获取子代特异性序列之后，对所述子代特异性序列进行过滤，其中包括，基于子代特异性 Kmer 频数，过滤掉在第一亚读段集中频数小于 2 的子代特异性亚读段对应的读段，和/或，过滤掉其上出现的子代特异性亚读段的数目少于 2 的读段。一亚读段在一个集合中的频数指该亚读段在这个集合中的个数。图 2 示意获取和过滤子代特异性 reads 的流程。对 Kmer 的计数可以直接利用哈希表进行，哈希表是一种以空间换取时间的数据结构，将第一测序结果建成一哈希表，哈希表中的关键字即为 Kmers，存储值（哈希值）即为 Kmer 的个数，也可以利用软件 Jellyfish [http://www.cbcb.umd.edu/software/jellyfish/] 、 Tallymer

[http://www.zbh.uni-hamburg.de/?id=211]和 Celera 中的 Meryl 模块 [Miller,J.R. et al. (2008) Aggressive assembly of pyrosequencing reads with mates. Bioinformatics, 24, 2818–2824.]等进行。所说的过滤掉其上出现的子代特异性亚读段的数目少于 2 的读段指，若一读段中出现的子代特异性 Kmer 的数量少于 2 则将其过滤掉，例如，若一子代特异性序列上出现 2 种子代特异性 Kmer，不论这两种特异性 Kmer 各自的频数，该子代特异性序列都得以保留；若一子代特异性序列上只出现一种子代特异性 Kmer，该子代特异性 Kmer 在该子代特异性序列上出现不小于 2 次，该子代特异性序列也得以保留。对获得的子代特异性序列进行上述至少之一的过滤，有利于减少获得假阳性子代特异性序列，有利于获得高质量的子代特异性序列。

本领域普通技术人员可以理解，上述实施例中各种方法的全部或部分步骤可以通过程序指令相关硬件完成，该程序可以存储于一机器可读存储介质中，例如存储于计算机可读存储介质中，存储介质可以包括：只读存储器、随机存储器、磁盘或光盘等。

依据本发明的另一方面，本发明提供一种检测子代新突变的方法，包括，利用上述本发明一方面或者任一实施例中的方法，获取子代特异性序列；基于所述子代特异性序列，检测所述子代新突变。所称的子代新突变指非遗传自其父母的变异，即其父母基因组上不带有的变异，即子代在该位置的基因型既不同于其父亲，也不同于其母亲，这里的父亲或者母亲 ,指其生物学意义上的父母。前述对本发明一方面的方法的技术特征和优点的描述，同样适用本发明这一方面的子代新突变的检测方法，在此不再赘述。

根据本发明的一个实施例，获取子代特异性序列包括，过滤掉在第一亚读段集中频数小于 2 的子代特异性亚读段对应的读段。根据本发明的再一个实施例，过滤掉在第一亚读段集中频数小于 2 的子代特异性亚读段对应的读段，也过滤掉其上出现的子代特异性亚读段的数目少于 2 的读段。一亚读段在一个集合中的频数指该亚读段在这个集合中的个数。对获得的子代特异性序列进行上述至少之一的过滤 ,利于减少获得假阳性子代特异性序列，

利于提高后续基于子代特异性序列的新突变检测的灵敏度。

根据本发明的一个实施例，所述基于子代特异性序列，检测所述子代新突变，包括：将所述子代特异性序列与参考序列比对，获得比对结果；基于所述比对结果，检测所述子代新突变，所述子代新突变包括 SNP、INDEL 和 SV 中的至少一种。比对可以选择利用但不限于 Samtools、SOAP、BWA 和 TeraMap 软件来进行。在本发明的一个实施例中，比对是利用 BWA 和 Samtools，依照其默认参数进行的。所使用的参考序列是已知序列，可以是预先获得的目标个体所属生物类别中的任意的参考模板。例如，若目标子代是人类，参考序列可选择美国国家生物技术信息中心（NCBI，national center for biotechnology information 提供的 HG19。SNP、INDEL 或 SV 的检测可选择但不限于利用软件 SOAPsnp、SOAPindel、GATK 和 SOAPdenovo 进行。

本领域普通技术人员可以理解，上述实施例中各种方法的全部或部分步骤可以通过程序来指令相关硬件完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：只读存储器、随机存储器、磁盘或光盘等。

依据本发明的再一方面，提供一种检测子代新突变的装置，包括：数据输入单元，用于输入数据；数据输出单元，用于输出数据；存储单元，用于存储数据，其中包括计算机可执行程序；处理器，与所述数据输入单元、所述数据输出单元和所述存储单元连接，用于执行所述程序，执行所述程序包括完成本发明一方面或者任一实施例中的检测子代新突变的方法的全部或部分步骤。前述对本发明一方面的子代新突变的检测方法的技术特征和优点的描述，同样适用本发明这一方面的子代新突变的检测装置，在此不再赘述。所称的处理器一般是负责运算和处理，存储单元中的一部分为内存，主要为交换数据。当程序或者操作者对处理器发出指令，这些指令和输入数据暂存在内存里，在处理器空闲时传送给处理器，处理器处理后把结果输出到诸如显示器、打印机等输出设备上。在输出完成之前，这些结果也保存在内存里，如果内存不足，结果数据读取的速度会变慢很多。执行该程序

能快速实现本发明一方面的方法的全部或部分步骤，对内存的需求比目前的方法低。

本发明这一方面的子代新突变的检测方法/装置，不再依赖重测序，而是另辟蹊径，利用分析 Kmer 找出子代特异性 reads，从而基于这些 reads 进行子代新突变（denovo mutation）的检测，原理简单，大幅降低了高杂合、高突变以及高重复区域的负面影响，检测结果的准确度和灵敏度都很高。

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 是本发明的一个实施例中的检测子代新突变的方法的技术原理示意图。

图 2 是本发明的一个实施例中的获取和过滤子代特异性 reads 的流程示意图。

图 3 是本发明的一个实施例中的子代 Kmer 的频数分布示意图。

图 4 是本发明的一个实施例中的父母 Kmer 的频数分布示意图。

图 5 是本发明的一个实施例中的子代 Kmer 集中的子代特异性 Kmer 的频数分布示意图，其中，在每个频数对应的 3 个条形柱，从左往右的依次为 21mer、25mer 和 29mer。

图 6 是本发明的一个实施例中的子代特异性 Kmer 的频数与检测灵敏度的关系的示意图。

图 7 是本发明的子代特异性 reads 的具体过滤条件与检测灵敏度的关系的示意图。

具体实施方式

术语解释

变异：

基因组 DNA 序列上的各种变化，包括单碱基突变（SNP），碱基的缺失或者插入（INDEL）以及大的染色体结构变异（SV），变异的分类是一个相对的概念，在本发明中的变异或者

突变，与核酸变异、基因变异、染色体变异等可通用，本发明中的 SNP、插入缺失（indel）和 SV 同通常定义，但本发明对后两者的大小不作特别限定，这样不同种变异之间有的有交叉，比如 SNP 为单核酸突变，包括单核苷酸的插入和/或缺失，这样与插入缺失变异有交叉；又比如，当发生的插入缺失较大时，也属于 SV。这些类型变异的大小交叉并不妨碍本领域普通技术人员通过上述描述执行实现本发明的方法和/或装置，并且达到所描述的结果。

新突变（Denovo mutation）：

变异中的一个分支，它不同于通过遗传获得的突变（Inherited mutation），也就是说，它与父母的遗传信息是不一致的，主要由于受精卵在分裂前发生了变化。

读段（Reads）：

对 DNA 片段进行测序，其测序对象都是一段物理连续的 DNA 序列，该片段称为插入片段，其长度称为插入片段长度(insert size)。测序过程中，是对该插入片段或者其上至少其一部分序列的碱基进行读取，例如双末端测序中，是对插入片段的两侧碱基序列从边缘向内部进行读取，读测得的序列称为“reads”，其长度可称为读长。

Kmer：

Kmer 是具有指定长度为 K（例如 K=17）的 DNA 序列，一条长度为 L 的 reads 产生的 Kmer 数量为 L-K+1。假定 K=17，L=100，则该 reads 产生 84 条 Kmer。总 Kmer 数与总碱基数的比率为(L-K+1)/L，在上述假定中，该比率为 0.84。由于测序错误不可避免，往往有很多低频的 Kmer 存在，这些低频 Kmer 的总类与测序错误率有关，在碱基错误率为 0.1%，Kmer 大小为 25，基因组大小为 3G，基因组测序深度为 50X 的情况下，由于错误碱基带来的新的 Kmer 数大约为 3.5G 个，对于人的基因组来说自身的无重复 25Kmer 数为 2.5G 左右，这样错误的碱基造成内存增加为原来的 2.4 倍。所以在组装中，例如在构建重叠群（Contigs）之前，可对低频 Kmer 进行删除或者校正。

下面详细描述本发明的实施例，所述实施例中的示例图，其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。需要说明的是在本文中所使用的术语“第一”、“第二”等仅用于方便描述目的，而不能理解为指示或暗示相对重要性，也不能理解为之间有先后顺序关系。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。

除另有交待，以下实施例中涉及的试剂、仪器或者软件，都是常规市售产品或者开源的，比如从 Illumina 公司购买测序文库制备试剂盒，依照试剂盒说明书进行建库等。

实施例

从医院获得一个自闭症孩童以及其父母的血液细胞。自闭症是起病于婴幼儿期的一种严重的广泛性发育障碍，其主要特征为社会交往障碍、语言交流技巧障碍，以及重复刻板的行为和狭窄的兴趣等。家系研究、双生子研究均提示该病具有遗传倾向。它是一种与遗传高度相关的疾病 ,其遗传相关性在 40% - 90%之间。其同卵双胞胎之间的一致性在 82%，而异卵双胞胎之间的一致性为 10%。目前已经发现与自闭症相关的基因组变异类型包括 SNP、INDEL 和 SV 等，但是自闭症的遗传解释度仍然只有 10 - 20%左右。随着自闭症在基因组学方面的相应研究的进行以及全基因组测序技术的运用，越来越多新的遗传相关位点被研究者所发现,相应自闭症的遗传易感性的解释度能达到 50%左右。

一般可采用的步骤如下：

1、利用 Illumina Hiseq2000 对这三个样本的核酸进行常规建库和高深度全基因组测序，双末端测序，获得子代和父母测序数据均为 90G（测序深度各约 30X）。同时，基于前人对该家系的研究结果(Yong-hui Jiang, *et al.* 2013. Detection of Clinically Relevant Genetic Variants in Autism Spectrum Disorder by Whole-Genome Sequencing)，我们挑选研究结果中 60 个被验证正确的 denovo mutation 用作最后的灵敏度测试。

2、过滤原始下机数据，对于含 N 超过 10%的序列、质量值低于 10 的序列、包含测序引物的序列，均予以删除。根据以下图 3 显示的，不同大小 Kmer 的频数分布曲线以及基于不同 reads 过滤条件的同一 Kmer 的频数分布曲线，可看出该步骤对检测结果的准确度和灵敏度基本不影响。但该步骤减少了数据量，利于提高机器运行检测方法的速度。

3、将父母测序数据看成是一个整体，用工具 jellyfish 对子代和父母的过滤完的原始测序数据分别进行 Kmer 分析。Jellyfish 参数如下："jellyfish count /dev/fd/0 -m 21 -s 3G -C -t 8 -Q T --bf-size 10G"，各参数的意思及 Jellyfish 的用法参照说明请见 http://www.cbcb.umd.edu/software/jellyfish/。单个个体的 Kmer 分析需求为 50G，计算时间 12 小时。我们的 Kmer 分析采用三档的 Kmer 长度——21bp、25bp 和 29bp，即对应 21mer、25mer 和 29mer，图 3 和图 4 分别显示子代和其父母的 Kmer 频数的分布，其中的"21mer-Q"表示基于上步过滤后的 reads 的 21mer 的频数分布曲线。从同样长度的 Kmer 的频数分布曲线，即从"21mer"和"21mer-Q"曲线可看出，在较低深度，如小于 5X，却有较高频率的 Kmer 的，这些 Kmer 很可能是由于测序错误包括低质量 reads 带来的。

4、获取、过滤子代特异性 Kmer。在上一步的基础上，过滤频数小于 2 的 Kmer，对子代和父母各自构建 Kmer 集合，逐个遍历，找出仅存在于子代 Kmer 集的特异性 Kmer。图 5 展示了子代 Kmer 集中的子代特异性 Kmer（UC-Kmer）的频数分布，对应于每个频数的三个条柱，从左往右依次为 21mer、25mer 和 29mer，从图中可看出，三种长度的 Kmer 集中的子代特异性 Kmer 频数绝大多数都分布在频数为 1 的位置。可靠的 UC-Kmer 需要一定频数的支持，基于频数过低很有可能由于测序错误导致的，我们这边将频数不大于 3 的子代特异性 Kmer 过滤掉。

基于已找出的特异性 Kmer，追踪它们对应的 reads，也就是特异性 reads，对于双末端测序，追踪获得成对的 reads。随后，利用这些特异性 reads 检测子代变异，例如，利

用重测序手段将这些 reads 比对到参考基因组，所选工具为 bwa 和 samtools，均为默认参数，检测出子代特异性变异（子代新突变），包括 SNP、INDEL 和 SV。

理论上，对子代特异性 Kmer 的频数进行要求，会对特异性 reads 的数量产生非常大的影响，所以我们采用梯度测试，即设置不同的 Kmer 频数过滤阈值，得到不同数目的特异性 reads 集合，从而获取不同的 denovo mutations，它们最终与 60 个已被验证正确的变异进行比较。图 6 显示子代特异性 Kmer 的最小频数与灵敏度的关系，其中，横标代表子代特异性 Kmer 的频数，纵坐标代表假阴性（FN）的比率，假阴性 FN 值越低表明灵敏度越高。由图可见，三种长度的 Kmer 在频数为 1-2 的阶段有较好的灵敏度，随着最小频数的上升，假阴性 FN 值有不同程度的上升，进而灵敏度有不同程度的下降。

5、由于每一条特异性 reads 一般都包含不同数量的特异性 Kmer，所以为了客观反映该方法检测的灵敏度和准确性，我们又采用独特的双重梯度测试，即基于特异性 Kmer 的频数以及每条特异性 reads 包含的特异性 Kmer 的数量，对特异性 reads 进行过滤，然后，将不同过滤参数下的特异性 reads 集合分别比回参考基因组检测 denovo mutation，与 60 个已被验证正确的变异进行比较，灵敏度结果如图 7 所示。图 7 显示了子代特异性 reads 的过滤条件与检测灵敏度的关系，子代特异性 reads 的过滤条件包括：子代特异性 Kmer 的频数（frequency，图上简单表示为 freq）和每个子代特异性 reads 包含的特异性 Kmer 的数量。

综上所述，可看出该方法具有如下优点：原理简单，大幅降低了高杂合、高突变以及高重复区域的负面影响；检测灵敏度可达到接近 100%，即假阴性接近为 0。

权 利 要 求 书

1．一种获取子代特异性序列的方法，其特征在于，包括，

获取第一测序结果和第二测序结果，所述第一测序结果为所述子代的基因组测序结果，所述第二测序结果为所述子代的父亲和母亲的基因组测序结果，所述第一测序结果包括多个长度为 K_1 的读段，所述第二测序结果包括多个长度为 K_2 的读段；

分别利用所述第一测序结果和所述第二测序结果中的读段，获得第一亚读段集和第二亚读段集，

所述第一亚读段集由多个第一亚读段子集构成，一个所述第一亚读段子集由第一测序结果中的一条读段中出现的每条长度为 K 的亚读段组成，

所述第二亚读段集由多个第二亚读段子集构成，一个所述第二亚读段子集由第二测序结果中的一条读段中出现的每条长度为 K 的亚读段组成；

将所述第一亚读段集的亚读段和所述第二亚读段集中的亚读段进行比较，获得仅包含于第一亚读段集的亚读段，将所述仅包含于第一亚读段集的亚读段称为子代特异性亚读段；

从所述第一测序结果中提取所述子代特异性亚读段对应的读段，所述提取得的读段为所述子代特异性序列；其中，

$K < K_1$，$K < K_2$。

2．权利要求 1 的方法，其特征在于，所述分别利用第一测序结果和第二测序结果中的读段，获得第一亚读段集和第二亚读段集，包括，

利用第一测序结果中的全部读段，获得全部第一亚读段子集，全部第一亚读段子集构成所述第一亚读段集，

利用第二测序结果中的全部读段，获得全部第二亚读段子集，全部第二亚读段子集构成所述第二亚读段集。

3．权利要求 2 的方法，其特征在于，所述分别利用第一测序结果和第二测序结果中的读段，获得第一亚读段集和第二亚读段集，包括，

从第一测序结果中的读段 i 的一端的第 j 个核苷酸开始，沿另一端方向拷贝该读段的 K 个连续核苷酸为一个亚读段，i 取遍{1，2，…，m-1，m}中的数值，对于每个 i，j 取遍{1，2，…，(K_1-K)，(K_1-K+1)}中的数值，以获得所述第一亚读段集，

i 为第一测序结果中的读段的编号，m 为第一测序结果包含的读段的总数，

从第二测序结果中的读段 w 的一端的第 z 个核苷酸开始，沿另一端方向拷贝该读段的 K 个连续核苷酸为一个亚读段，w 取遍{1，2，…，n-1，n}中的数值，对于每个 w，z 取遍{1，2，…，(K_2-K)，(K_2-K+1)}中的数值，以获得所述第二亚读段集，

w 为第二测序结果中的读段的编号，n 为第二测序结果包含的读段的总数。

4．权利要求 1-3 任一方法，其特征在于，在将所述第一亚读段集的亚读段和所述第二亚读段集中的亚读段进行比较之前，分别过滤掉第一亚读段集和第二亚读段集中的频数小于 2 的亚读段。

5．权利要求 1-3 任一方法，其特征在于，在获得子代特异性亚读段之后，过滤掉频数小于 2 的子代特异性亚读段。

6．权利要求 1 的方法，其特征在于，还包括，

对所述子代特异性序列进行过滤，其中包括，

过滤掉在第一亚读段集中频数小于 2 的子代特异性亚读段对应的读段，和/或，

过滤掉其上出现的子代特异性亚读段的数目少于 2 的读段。

7．权利要求 1-6 任一方法，其特征在于，所述第一测序结果和所述第二测序结果均不包括符合以下描述的读段，

含 N 超过 10%，或质量值低于 10，或被接头污染。

8．权利要求 1-7 任一方法，其特征在于，在 $K_1 \geq 50bp$、$K_2 \geq 50bp$ 时，$19bp \leq K \leq$

31bp。

9．一种检测子代新突变的方法，其特征在于，包括，

利用权利要求 1-8 任一方法，获取子代特异性序列；

基于所述子代特异性序列，检测所述子代新突变。

10．权利要求 9 的方法，其特征在于，所述基于子代特异性序列，检测所述子代新突变，包括，

将所述子代特异性序列与参考序列比对，获得比对结果；

基于所述比对结果，检测所述子代新突变，所述子代新突变包括 SNP、INDEL 和 SV 中的至少一种。

11．一种检测子代新突变的装置，其特征在于，包括，

数据输入单元，用于输入数据；

数据输出单元，用于输出数据；

存储单元，用于存储数据，其中包括计算机可执行程序；

处理器，与所述数据输入单元、所述数据输出单元和所述存储单元连接，用于执行所述程序，执行所述程序包括完成权利要求 9 或 10 的方法。

12．一种计算机可读介质，其特征在于，用于存储供计算机执行的程序，执行所述程序包括完成权利要求 9 或 10 的方法。

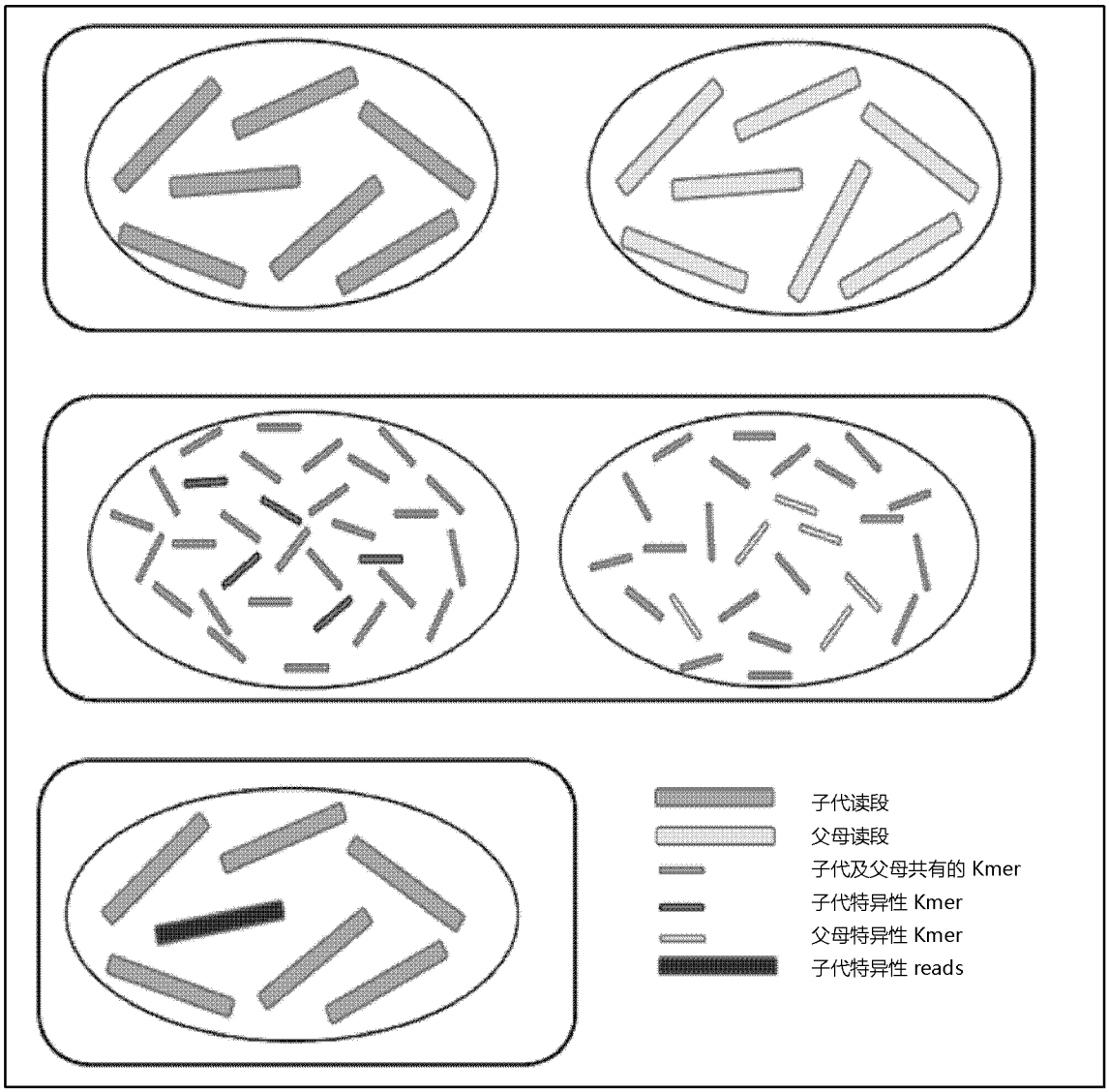


图 1

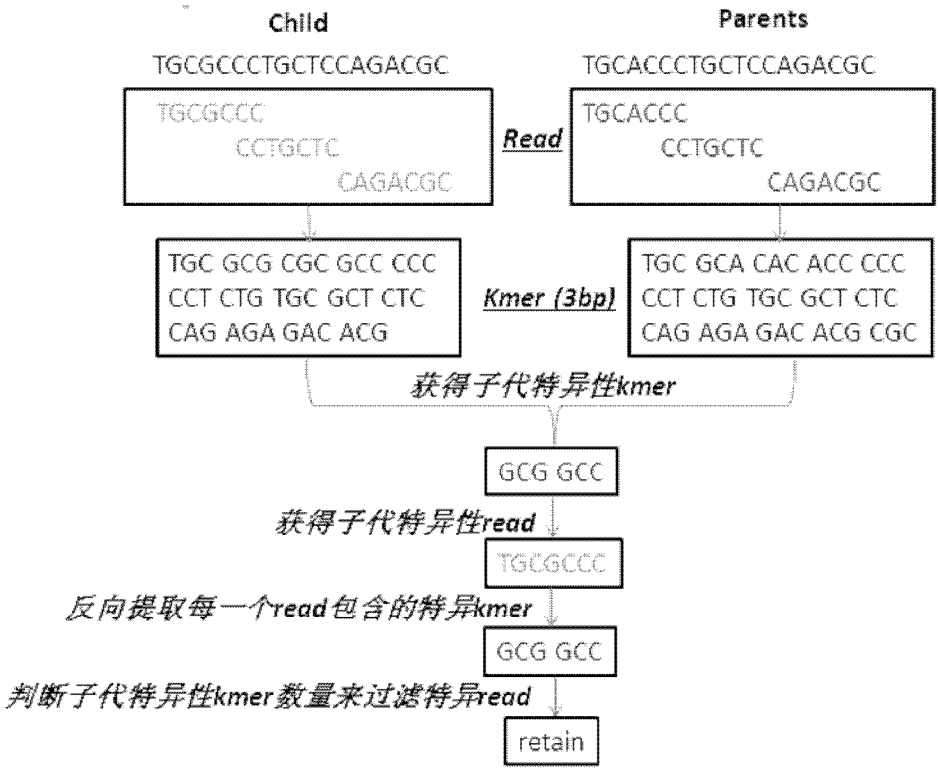

Child
Parents
TGCGCCCTGCTCCAGACGC
TGCACCCTGCTCCAGACGC
TGCGCCC
CCTGCTC
CAGACGC
Read
TGCACCC
CCTGCTC
CAGACGC
TGC GCG CGC GCC CCC
CCT CTG TGC GCT CTC
CAG AGA GAC ACG
Kmer (3bp)
TGC GCA CAC ACC CCC
CCT CTG TGC GCT CTC
CAG AGA GAC ACG CGC
获得子代特异性kmer
GCG GCC
获得子代特异性read
TGCGCCC
反向提取每一个read包含的特异kmer
GCG GCC
判断子代特异性kmer数量来过滤特异read
retain

图 2

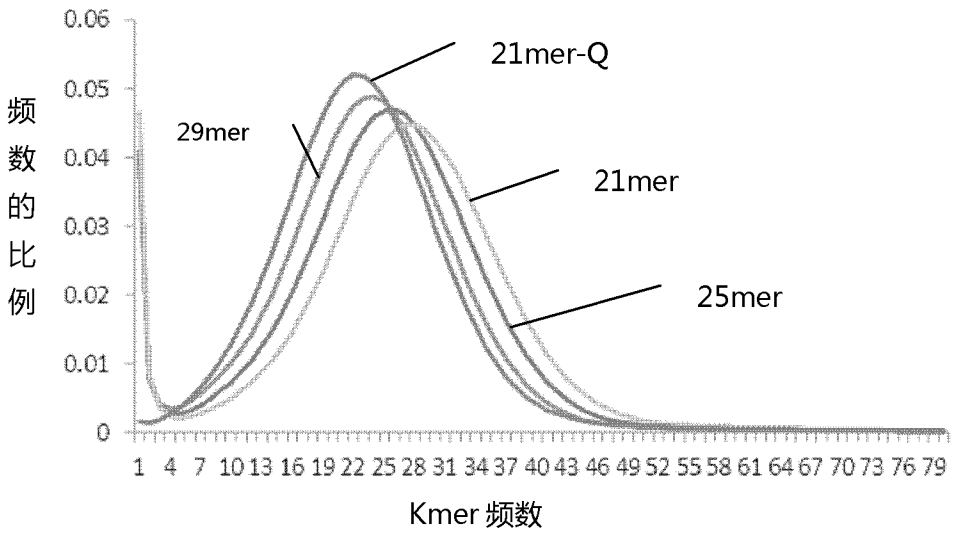

0.06
0.05
0.04
0.03
0.02
0.01
0
频数的比例
21mer-Q
29mer
21mer
25mer
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79
Kmer 频数

图 3

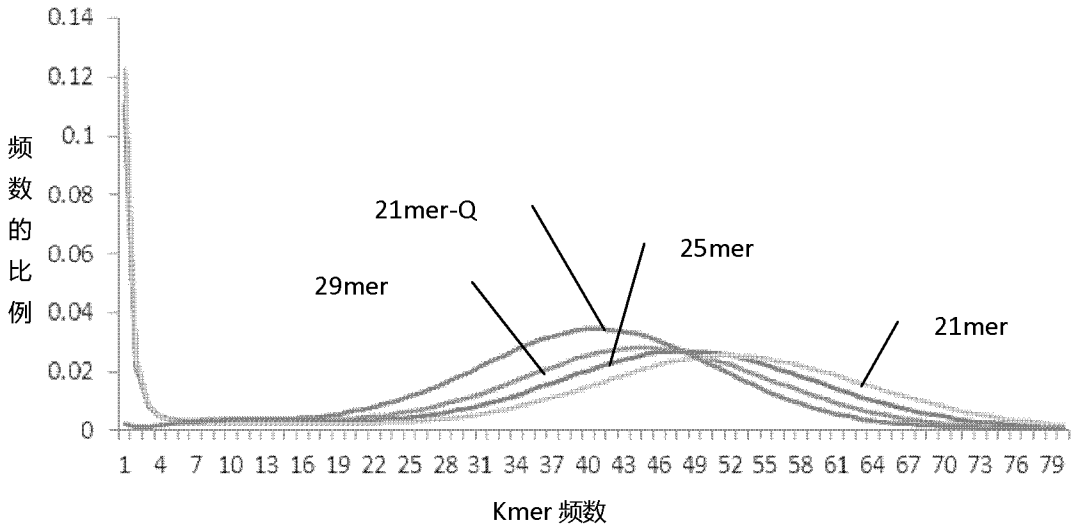

0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
频数的比例
21mer-Q
25mer
29mer
21mer
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79
Kmer 频数

图 4

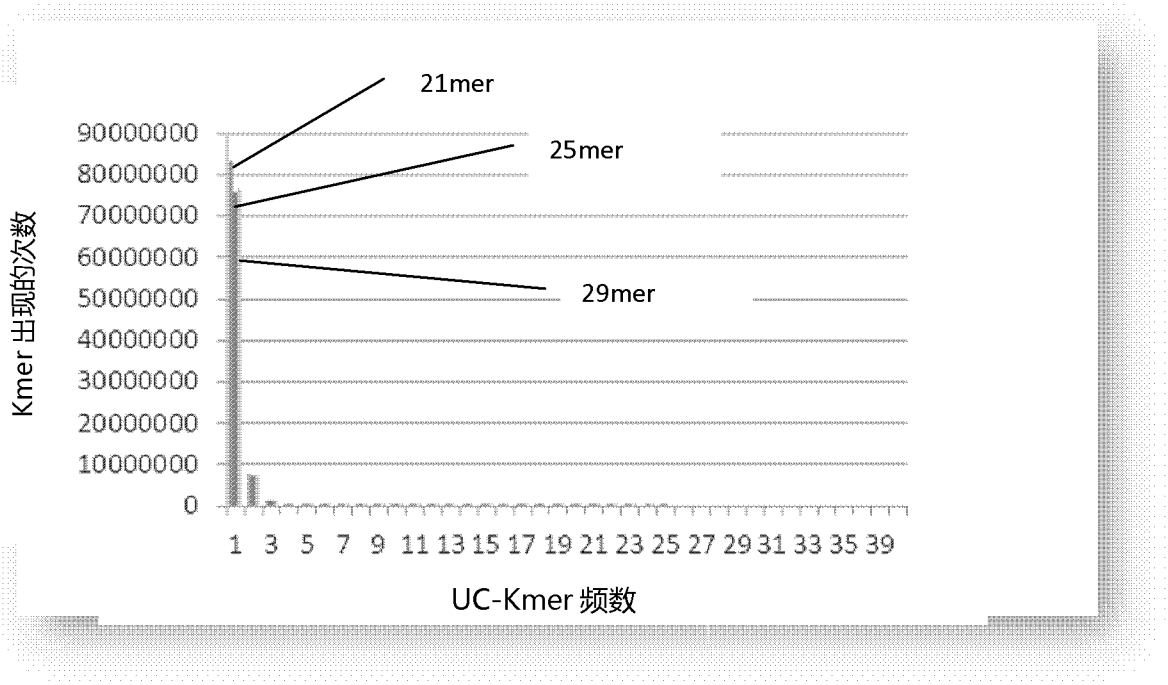

21mer
25mer
29mer
90000000
80000000
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
Kmer 出现的次数
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 39
UC-Kmer 频数

图 5

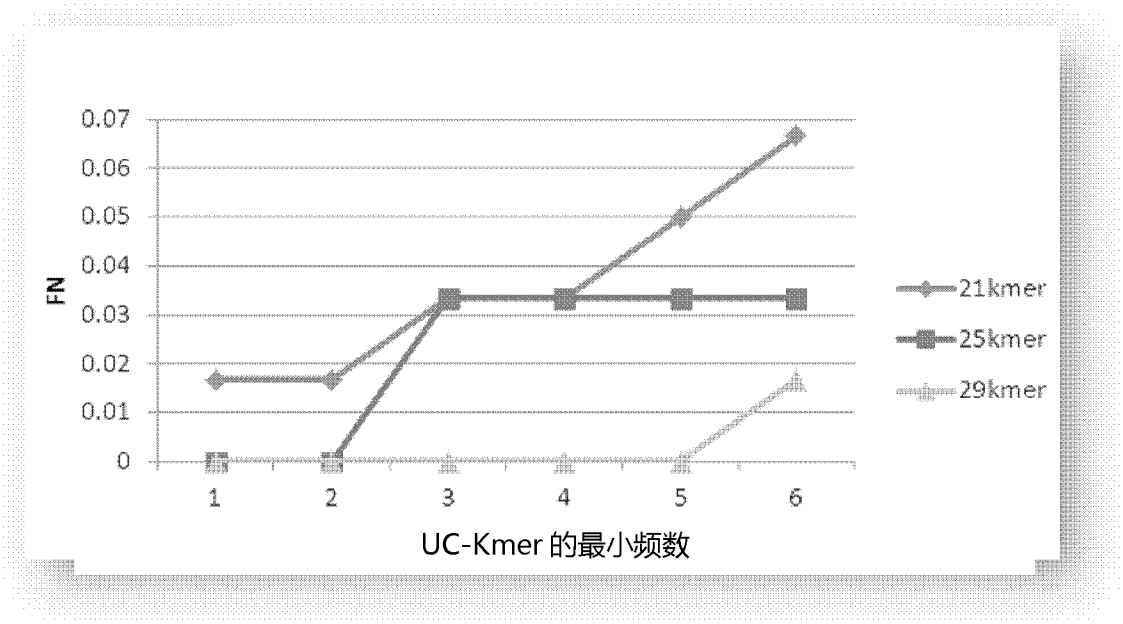

0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
FN
1
2
3
4
5
6
UC-Kmer 的最小频数
21kmer
25kmer
29kmer

图 6

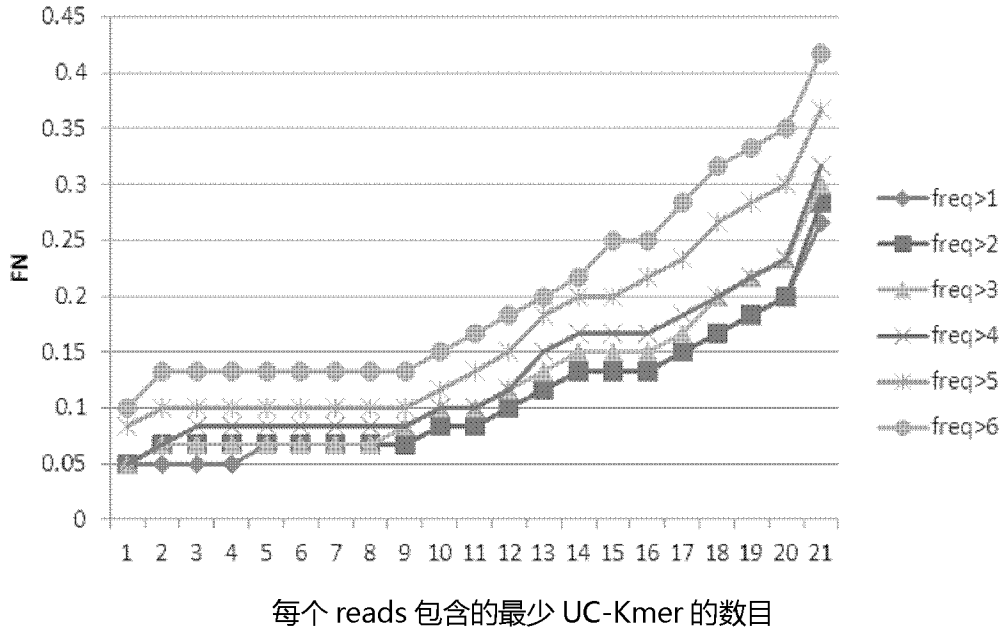

0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
FN
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
每个 reads 包含的最少 UC-Kmer 的数目
freq>1
freq>2
freq>3
freq>4
freq>5
freq>6

图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2015/073816

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)

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, CPEA, CNABS, JPABS, CNMED, CNKI, Google Scholar: sequencing, sequence, read, mutation, variation, parental, paternal, maternal, denovo, de novo

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 2013177581 A2 (UNIV WASHINGTON CT COMMERCIALI) 28 November 2013 (28.11.2013), see the abstract, description, paragraphs [0009], [0038] to [0040], and claim 23	1-11
Y	CN 103824001 A (BEIJING LUPHE ZHIYUAN BIOLOGICALINFORMAT) 28 May 2014 (28.05.2014), see the abstract	1-11
A	WO 2014019267 A1 (BGI SHENZHEN et al.) 06 February 2014 (06.02.2014), see the whole document	1-11
A	CN 104160391 A (CALLIDA GENOMICS INC.) 19 November 2014 (19.11.2014), see the whole document	1-11

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"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)

"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

"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

"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

"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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
02 December 2015	11 December 2015

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
XING, Ying
Telephone No. (86-10) 62411044

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2015/073816

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.:12
because they relate to subject matter not required to be searched by this Authority, namely:
Claim 12 is directed to mere presentations of information (PCT Rule 39. 1(v)).

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2015/073816

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
WO 2013177581 A2	28 November 2013	US 2015105267 A1	16 April 2015
		WO 2013177581 A3	26 June 2014
CN 103824001 A	28 May 2014	None	
WO 2014019267 A1	06 February 2014	WO 2014019180 A1	06 February 2014
CN 104160391 A	19 November 2014	WO 2013040583 A3	22 May 2014
		WO 2013040583 A2	21 March 2013
		US 2013110407 A1	02 May 2013
		HK 1199313 A1	26 June 2015

国际检索报告

国际申请号 PCT/CN2015/073816

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)

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, CPEA, CNABS, JPABS, CNMED, CNKI, Google Scholar: 测序, 读段, 读数, 突变, 变异, 父母, 亲代, 父本, 母本, sequencing, sequence, read, mutation, variation, parental, paternal, maternal, denovo, de novo

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
Y	WO 2013177581 A2 (UNIV WASHINGTON CT COMMERCIALI) 2013年 11月 28日 (2013-11-28) 参见摘要，说明书第[0009]、[0038]-[0040]段，权利要求23	1-11
Y	CN 103824001 A (北京诺禾致源生物信息科技有限公司) 2014年 5月 28日 (2014-05-28) 参见摘要	1-11
A	WO 2014019267 A1 (BGI SHENZHEN等) 2014年 2月 6日 (2014-02-06) 参见全文	1-11
A	CN 104160391 A (考利达基因组股份有限公司) 2014年 11月 19日 (2014-11-19) 参见全文	1-11

☐ 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	国际检索报告邮寄日期
2015年 12月 2日	2015年 12月 11日

ISA/CN的名称和邮寄地址	受权官员
中华人民共和国国家知识产权局(ISA/CN) 北京市海淀区蓟门桥西土城路6号 100088 中国 传真号 (86-10)62019451	幸颖 电话号码 (86-10)62411044

国际检索报告

国际申请号
PCT/CN2015/073816

第II栏　某些权利要求被认为是不能检索的意见(续第1页第2项)

根据条约第17条(2)(a)，对某些权利要求未做国际检索报告的理由如下:

1. ☑ 权利要求： 12
因为它们涉及不要求本单位进行检索的主题，即：

[1] 权利要求12涉及单纯的信息表达（PCT实施细则39.1(v)）。

2. ☐ 权利要求：
因为它们涉及国际申请中不符合规定的要求的部分，以致不能进行任何有意义的国际检索， 具体地说：

3. ☐ 权利要求：
因为它们是从属权利要求，并且没有按照细则6.4(a)第2句和第3句的要求撰写。

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2015/073816

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
WO	2013177581	A2	2013年 11月 28日	US	2015105267	A1	2015年 4月 16日
				WO	2013177581	A3	2014年 6月 26日
CN	103824001	A	2014年 5月 28日		无		
WO	2014019267	A1	2014年 2月 6日	WO	2014019180	A1	2014年 2月 6日
CN	104160391	A	2014年 11月 19日	WO	2013040583	A3	2014年 5月 22日
				WO	2013040583	A2	2013年 3月 21日
				US	2013110407	A1	2013年 5月 2日
				HK	1199313	A1	2015年 6月 26日