

(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
2018 年 11 月 29 日 (29.11.2018)



(10) 国际公布号
WO 2018/214010 A1

- (51) 国际专利分类号:
C12Q 1/68 (2006.01) **C12Q 1/00** (2006.01)
- (21) 国际申请号: PCT/CN2017/085448
- (22) 国际申请日: 2017 年 5 月 23 日 (23.05.2017)
- (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: 莫晓东(MO, Xiaodong); 中国广东省深
圳市盐田区北山工业区综合楼, Guangdong
518083 (CN)。 杨焕杰(YANG, Huanjie); 中国
广东省深圳市盐田区北山工业区综合楼,
Guangdong 518083 (CN)。 张涛(ZHANG, Tao);
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: 深圳鼎合诚知识产权代理有限公司
(DHC IP ATTORNEYS); 中国广东省深圳福

田区金田路与福华路交汇处现代商务大
厦2201, Guangdong 518048 (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, DJ, 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, KH, KN, KP, KR, KW, 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,

(54) Title: METHOD, DEVICE, AND STORAGE MEDIUM FOR DETECTING MUTATION ON THE BASIS OF SEQUENCING DATA

(54) 发明名称: 一种基于测序数据的变异检测方法、装置和存储介质

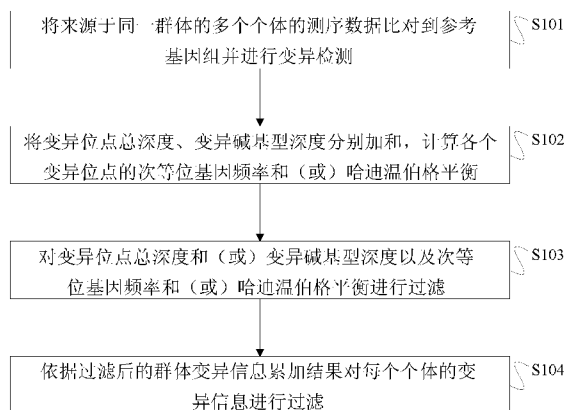


图 1

- S101 Align sequencing data of a plurality of individuals from the same population to a reference genome and conduct mutation detection
- S102 Respectively sum total mutated site depths and mutated base-type depths, and calculate sub-allelic frequencies and/or Hardy-Weinberg equilibriums of mutated sites
- S103 Filter the total mutated site depths and/or the mutated base-type depths as well as the sub-allelic frequencies and/or the Hardy-Weinberg equilibriums
- S104 Filter mutation information of each individual according to the filtered populational mutation information accumulation result

(57) Abstract: A method, a device, and a storage medium for detecting mutation on the basis of sequencing data. The method comprises: aligning sequencing data of a plurality of individuals from the same population to a reference genome and conducting mutation detection, to obtain alignment position and mutation information in read lengths; for mutated sites, respectively summing the total mutated site depths and mutated base-type depths of all of the individuals in the population, and then calculating sub-allelic frequencies and/or Hardy-Weinberg equilibriums of the mutated sites in the population; plotting a diagram of the relationship between the total mutated site depths and/or the mutated base-type depths of the mutated sites and dbSNP%, and filtering the total mutated site depths and/or the mutated base-type depths as well as the sub-allelic frequencies and/or the Hardy-Weinberg equilibriums of the mutated sites; and filtering the mutation information of each individual, to obtain final mutation information of each individual. The present application can use low-depth sequencing data of populations for accurate mutation detection, improving data utilization.

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))。

(57) 摘要：一种基于测序数据的变异检测方法、装置和存储介质，该方法包括：将来源于同一群体的多个个体的测序数据比对到参考基因组并进行变异检测，得到读长的比对位置和变异信息；针对各个变异位点，将群体中所有个体的变异位点总深度、变异碱基型深度分别加和，然后计算各个变异位点在群体中的次等位基因频率和/或哈迪温伯格平衡；绘制变异位点总深度和/或变异碱基型深度与dbSNP%关系图，并对变异位点总深度和/或变异碱基型深度以及次等位基因频率和/或哈迪温伯格平衡进行过滤；对每个个体的变异信息进行过滤，得到每个个体的最终变异信息。能够利用群体的低深度测序数据进行精准的变异检测，提高数据利用率。

一种基于测序数据的变异检测方法、装置和存储介质

技术领域

本发明涉及生物信息学技术领域，具体涉及一种基于测序数据的变异检测方法、装置和存储介质。

背景技术

尽管测序成本仍在下降，但是目前测序策略仍没有质的突破，高深度的测序数据成本仍然居高不下。因此，基于低深度的变异检测方法，将有效提高现有低深度全基因组测序数据的利用率。目前对低深度的变异检测，仍是利用常规软件进行，得到的变异信息往往由于深度不足而在质控中被过滤掉。

为了得到这些低深度区域的变异信息，现在主要采用基于基因连锁推断的基因型填补（Genotype imputation），以高深度的变异位点为标记位点，参考已有的基因集合（收集了一定群体的基因信息），根据已知的标记位点与临近未知位点的连锁关系来推断未知的变异信息。这种方法利用基因连锁关系进行推断，通过推断方法填补得到的未知位点，事实上在个体测序过程中也有覆盖，但是由于深度较低，一直被忽视，造成这些数据的浪费。通过推断方法进行基因型填补，没有考虑测序得到的真实碱基情况，单纯参考基因集合中的基因信息无法获得个体本身特有的变异，牺牲了准确度和低深度数据。不仅如此，由于基因型推断填补技术需要利用已知的高深度变异位点信息，所以对变异信息密度有所要求，如果已知的变异位点数目过少、分布密度过低，将难以开展基因型的推断填补。此外基因型填补技术耗时较长，时效性较差。

发明内容

本发明提供一种基于测序数据的变异检测方法、装置和存储介质，能够利用群体的低深度测序数据直接进行精准的变异检测，提高数据利用率。

根据第一方面，一种实施例中提供一种基于测序数据的变异检测方法，包括：

将来源于同一群体的多个个体的测序数据比对到参考基因组并进行变异检测，得到读长的比对位置和变异信息；

依据上述比对位置和变异信息，针对各个变异位点，将群体中所有个体的变异位点总深度、变异碱基型深度分别加和，然后计算各个变异位点在群体中的次等位基因频率和/或哈迪温伯格平衡，得到群体变异信息累加结果；

绘制变异位点总深度和/或变异碱基型深度与 dbSNP%关系图，并依据设定的过滤阈值对上述变异位点总深度和/或变异碱基型深度以及上述次等位基因频率和/或哈迪温伯格平衡进行过滤以滤除低于上述过滤阈值的值，得到过滤后的群体变异信息累加结果；

依据上述过滤后的群体变异信息累加结果对每个个体的变异信息进行过滤，得到每个个体的最终变异信息。

根据第二方面，一种实施例中提供一种基于测序数据的变异检测装置，包括：

比对与变异检测装置，用于将来源于同一群体的多个个体的测序数据比对到参考基因组并进行变异检测，得到读长的比对位置和变异信息；

加和与计算装置，用于依据上述比对位置和变异信息，针对各个变异位点，将群体中所有个体的变异位点总深度、变异碱基型深度分别加和，然后计算各个变异位点在群体中的次等位基因频率和/或哈迪温伯格平衡，得到群体变异信息累加结果；

群体变异信息过滤装置，用于绘制变异位点总深度和/或变异碱基型深度与 dbSNP%关系图，并依据设定的过滤阈值对上述变异位点总深度和/或变异碱基型深度以及上述次等位基因频率和/或哈迪温伯格平衡进行过滤以滤除低于上述

过滤阈值的值，得到过滤后的群体变异信息累加结果；

个体变异信息过滤装置，用于依据上述过滤后的群体变异信息累加结果对每个个体的变异信息进行过滤，得到每个个体的最终变异信息。

根据第三方面，一种实施例中提供一种基于测序数据的变异检测装置，包括：

存储器，用于存储程序；

处理器，用于通过执行上述存储器存储的程序以实现如下的方法：

将来源于同一群体的多个个体的测序数据比对到参考基因组并进行变异检测，得到读长的比对位置和变异信息；

依据上述比对位置和变异信息，针对各个变异位点，将群体中所有个体的变异位点总深度、变异碱基型深度分别加和，然后计算各个变异位点在群体中的次等位基因频率和/或哈迪温伯格平衡，得到群体变异信息累加结果；

绘制变异位点总深度和/或变异碱基型深度与 dbSNP%关系图，并依据设定的过滤阈值对上述变异位点总深度和/或变异碱基型深度以及上述次等位基因频率和/或哈迪温伯格平衡进行过滤以滤除低于上述过滤阈值的值，得到过滤后的群体变异信息累加结果；

依据上述过滤后的群体变异信息累加结果对每个个体的变异信息进行过滤，得到每个个体的最终变异信息。

根据第四方面，一种实施例中提供一种计算机可读存储介质，包括程序，上述程序能够被处理器执行以实现如下的方法：

将来源于同一群体的多个个体的测序数据比对到参考基因组并进行变异检测，得到读长的比对位置和变异信息；

依据上述比对位置和变异信息，针对各个变异位点，将群体中所有个体的变异位点总深度、变异碱基型深度分别加和，然后计算各个变异位点在群体中

的次等位基因频率和/或哈迪温伯格平衡，得到群体变异信息累加结果；

绘制变异位点总深度和/或变异碱基型深度与 dbSNP%关系图，并依据设定的过滤阈值对上述变异位点总深度和/或变异碱基型深度以及上述次等位基因频率和/或哈迪温伯格平衡进行过滤以滤除低于上述过滤阈值的值，得到过滤后的群体变异信息累加结果；

依据上述过滤后的群体变异信息累加结果对每个个体的变异信息进行过滤，得到每个个体的最终变异信息。

本发明能够利用群体的低深度测序数据，不再借助连锁推断而直接进行精准的变异检测，提高数据利用率，并且低深度区域获得的变异信息可作为标记位点，协助改善现有的连锁推断结果。

附图说明

图 1 为本发明一个实施例基于测序数据的变异检测方法的流程图；

图 2 为本发明一个实施例基于测序数据的变异检测装置的结构框图；

图 3 为本发明一个实施例中变异碱基型深度与 dbSNP%关系图，其中，各曲线代表不同染色体的情况，加粗曲线代表平均值，灰色部分代表标准差。

具体实施方式

下面通过具体实施方式结合附图对本发明作进一步详细说明。在以下的实施方式中，很多细节描述是为了使得本发明能被更好的理解。然而，本领域技术人员可以毫不费力的认识到，其中部分特征在不同情况下是可以省略的，或者可以由其他元件、材料、方法所替代。在某些情况下，本发明相关的一些操作并没有在说明书中显示或者描述，这是为了避免本发明的核心部分被过多的描述所淹没，而对于本领域技术人员而言，详细描述这些相关操作并不是必要

的，他们根据说明书中的描述以及本领域的一般技术知识即可完整了解相关操作。

另外，说明书中所描述的特点、操作或者特征可以以任意适当的方式结合形成各种实施方式。同时，方法描述中的各步骤或者动作也可以按照本领域技术人员所能显而易见的方式进行顺序调换或调整。因此，说明书和附图中的各种顺序只是为了清楚描述某一个实施例，并不意味着是必须的顺序，除非另有说明其中某个顺序是必须遵循的。

对于低深度测序数据的变异检测，传统方法都是针对单个样本，本发明的方法将群体的低深度变异合并在一起，看作是单个样本的变异信息，从而实现低深度群体转化为虚构的高深度个体，进而进行变异检测，得到精准的变异检测结果后再拆分得到单个个体本身的变异信息。

如图 1 所示，本发明一个实施例的基于测序数据的变异检测方法包括如下步骤：

步骤 S101：将来源于同一群体的多个个体的测序数据比对到参考基因组并进行变异检测，得到读长（Reads）的比对位置和变异信息。

本发明实施例中，测序数据属于同一群体，例如同一物种（人、猪等）。这些测序数据可以是同一批次的下机数据，也可以是虽然属于不同批次但是属于同一群体的下机数据。测序数据的格式例如 Fastq 格式。参考基因组可以是各群体（物种）的已公开的基因组数据，例如对人而言，参考基因组可以是人类参考基因组 hg19。比对软件可以是常用的 BWA 软件等，变异检测软件可以是常用的 GATK 软件等。得到的变异信息包含变异所在的染色体、位置、变异碱基型等元素。关于个体数量（即样本量），本发明实施例无严格要求，但样本量越多越有利于找到更多变异位点。

本发明实施例中，测序数据的平均测序深度没有特别限制，本发明实施例的方法特别适合于低深度测序数据，例如平均深度为 1×至 15×的测序数据。在

本发明一个实施例中，测序数据的平均测序深度是 $3.5\times$ 。

步骤 S102：依据得到的比对位置和变异信息，针对各个变异位点，将群体中所有个体的变异位点总深度、变异碱基型深度分别加和，然后计算各个变异位点在群体中的次等位基因频率（Minor allele frequency, MAF）和/或哈迪温伯格平衡（Hardy-Weinberg equilibrium, HWE），得到群体变异信息累加结果，所谓“群体变异信息累加结果”包括各个变异位点加和的变异位点总深度、变异碱基型深度，以及 MAF 和/或 HWE。

如本发明所使用的，术语“变异位点总深度”是指覆盖在该变异位点上的所有个体的测序读长（Reads）的数量；术语“变异碱基型深度”是指覆盖在该变异位点上某种特定的碱基类型的所有个体的测序读长的数量。可见，对特定变异位点而言，“变异位点总深度”是各种碱基类型的“变异碱基型深度”之和。例如，若某一变异位点存在 A、T 两种碱基类型，并且在所有个体的测序读长中，有 100 个在该变异位点为 A，有 100 个在该变异位点为 T，则 A、T 两种碱基类型的“变异碱基型深度”分别是 100，该变异位点的“变异位点总深度”是 200。

如本发明所使用的，术语“次等位基因频率”，指在给定群体中的不常见的等位基因的频率；术语“哈迪温伯格平衡”，是指各等位基因的频率和等位基因的基因型频率在遗传中稳定不变，即保持着基因平衡。

本发明实施例的方法特别适合于单核苷酸多态性（SNP）变异、插入/删除（Ins/Del）变异检测，因此在本发明的一个优选的实施例中，在将群体中所有个体的变异位点总深度、变异碱基型深度分别加和时，去除三碱基及其它多碱基变异而仅保留单核苷酸多态性变异。

步骤 S103：绘制变异位点总深度和/或变异碱基型深度与 dbSNP%关系图，并依据设定的过滤阈值对变异位点总深度和/或变异碱基型深度以及次等位基因频率和/或哈迪温伯格平衡进行过滤以滤除低于过滤阈值的值，得到过滤后的群体变异信息累加结果。

如本发明所使用的，术语“dbSNP%”是指发现的变异位点在寡核苷酸多态性数据库中的占比，通常用来衡量检测到的 SNP 的准确性。

在本发明实施例中，过滤阈值包括变异位点总深度的过滤阈值和/或变异碱基型深度的过滤阈值，以及 MAF 的过滤阈值和/或 HWE 的过滤阈值。也就是说，关键指标是变异碱基型深度（或变异位点总深度）、MAF 值和/或 HWE 值。某种碱基类型的变异碱基型深度与该变异的可信度直接相关，HWE 和 MAF 常用于除去低可信度的变异位点。

在本发明的一个实施例中，变异碱基型深度的过滤阈值是 $30\times$ ，MAF 的过滤阈值是 0.05。也就是说，变异碱基型深度大于等于 $30\times$ ，MAF 大于 0.05，认为达到统计学显著效果。

在本发明的其它实施例中，可以根据实际深度密度分布适当调整过滤深度（变异碱基型深度或变异位点总深度）。其中，“实际深度密度分布”可以是指不同变异位点的测序深度的分布。

在本发明的其它实施例中，样本量可以根据数据平均深度来设置，例如平均深度为 $1\times$ ，样本量应该为 30 例，使得变异位点类型深度可达到设定的大于等于 $30\times$ 的过滤条件。样本量越多，越有利于检测更多的变异位点。

步骤 S104：依据过滤后的群体变异信息累加结果对每个个体的变异信息进行过滤，得到每个个体的最终变异信息。

在本发明的一个具体实施例中，对每个个体的变异信息进行过滤具体包括：若在个体原始的变异信息中存在染色体、位置和变异碱基类型与过滤后的群体变异信息累加结果一致的结果，则保留该结果，否则过滤掉该结果。此外，还包括：若在一个变异位点上存在两种以上的变异，则过滤掉该变异位点。

通过以上步骤，可以得到每个个体的最终变异信息。然后，可通过 dbSNP% 来评估变异的准确性并且进行后续推断填补和研究分析。

本发明的方法提高了高通量测序数据中低深度区域的数据利用率，将本来

被忽视的数据转化为可利用的数据，同时显著提高了检测出 SNP 的准确性，而且方法简单易用，既可以提高准确率也可以增加数据利用率，还可以加入到传统变异填补的上一步，改善传统方法。

本领域技术人员可以理解，上述实施方式中各种方法的全部或部分功能可以通过硬件的方式实现，也可以通过计算机程序的方式实现。当上述实施方式中全部或部分功能通过计算机程序的方式实现时，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、光盘、硬盘等，通过计算机执行该程序以实现上述功能。例如，将程序存储在设备的存储器中，当通过处理器执行存储器中程序，即可实现上述全部或部分功能。另外，当上述实施方式中全部或部分功能通过计算机程序的方式实现时，该程序也可以存储在服务器、另一计算机、磁盘、光盘、闪存盘或移动硬盘等存储介质中，通过下载或复制保存到本地设备的存储器中，或对本地设备的系统进行版本更新，当通过处理器执行存储器中的程序时，即可实现上述实施方式中全部或部分功能。

因此，本发明的一种实施例还提供一种基于测序数据的变异检测装置，如图 2 所示，包括：

比对与变异检测装置 201，用于将来源于同一群体的多个个体的测序数据比对到参考基因组并进行变异检测，得到读长的比对位置和变异信息；加和与计算装置 202，用于依据比对位置和变异信息，针对各个变异位点，将群体中所有个体的变异位点总深度、变异碱基型深度分别加和，然后计算各个变异位点在群体中的次等位基因频率和/或哈迪温伯格平衡，得到群体变异信息累加结果；群体变异信息过滤装置 203，用于绘制变异位点总深度和/或变异碱基型深度与 dbSNP% 关系图，并依据设定的过滤阈值对变异位点总深度和/或变异碱基型深度以及次等位基因频率和/或哈迪温伯格平衡进行过滤以滤除低于过滤阈值的值，得到过滤后的群体变异信息累加结果；个体变异信息过滤装置 204，用于依据过滤后的群体变异信息累加结果对每个个体的变异信息进行过滤，得到每个

个体的最终变异信息。

本发明的另一种实施例还提供一种基于测序数据的变异检测装置，包括：

存储器，用于存储程序；

处理器，用于通过执行上述存储器存储的程序以实现如下的方法：

将来源于同一群体的多个个体的测序数据比对到参考基因组并进行变异检测，得到读长的比对位置和变异信息；依据比对位置和变异信息，针对各个变异位点，将群体中所有个体的变异位点总深度、变异碱基型深度分别加和，然后计算各个变异位点在群体中的次等位基因频率和/或哈迪温伯格平衡，得到群体变异信息累加结果；绘制变异位点总深度和/或变异碱基型深度与 dbSNP%关系图，并依据设定的过滤阈值对变异位点总深度和/或变异碱基型深度以及次等位基因频率和/或哈迪温伯格平衡进行过滤以滤除低于过滤阈值的值，得到过滤后的群体变异信息累加结果；依据过滤后的群体变异信息累加结果对每个个体的变异信息进行过滤，得到每个个体的最终变异信息。

本发明的又一种实施例还提供一种计算机可读存储介质，包括程序，该程序能够被处理器执行以实现如下的方法：

将来源于同一群体的多个个体的测序数据比对到参考基因组并进行变异检测，得到读长的比对位置和变异信息；依据比对位置和变异信息，针对各个变异位点，将群体中所有个体的变异位点总深度、变异碱基型深度分别加和，然后计算各个变异位点在群体中的次等位基因频率和/或哈迪温伯格平衡，得到群体变异信息累加结果；绘制变异位点总深度和/或变异碱基型深度与 dbSNP%关系图，并依据设定的过滤阈值对变异位点总深度和/或变异碱基型深度以及次等位基因频率和/或哈迪温伯格平衡进行过滤以滤除低于过滤阈值的值，得到过滤后的群体变异信息累加结果；依据过滤后的群体变异信息累加结果对每个个体的变异信息进行过滤，得到每个个体的最终变异信息。

以下通过实施例详细说明本发明的技术方案和效果，应当理解，实施例仅

是示例性的，不能理解为对本发明保护范围的限制。

实施例

本实施例中，样本为105例人类MHC区域（6号染色体上4.9M区域）捕获的测序数据，这批数据中集中在MHC区域的数据占约50%。也就是说还有约50%是伴随产生的基因组其他位置的数据，相对而言，这部分数据就是低深度的数据。利用本发明的方法对这50%低深度的区域进行变异检测，为对结果的准确性进行测评，还对这105例样本进行外显子测序，在外显子测序数据中，外显子区域是高深度的。通过比较低深度数据和外显子测序在外显子区域的一致性，便能确定变异的一致性情况，从而测评本发明方法的准确性。

本实施性的具体步骤如下：

（1）下载人类参考基因组hg19，利用BWA软件将测序读长比对到参考基因组hg19上，并用GATK软件对变异信息进行检测，将阈值调低以保留更多的变异信息。

（2）针对MHC区域捕获的测序数据变异位点，本实施例以低深度区域进行研究，本实施例中低深度区域的平均深度为 $3.5\times$ 。按照变异位点所在的染色体、位置、变异碱基型，将105个MHC样本的变异位点总深度、变异碱基型深度分别进行加和，去除三碱基及其它多碱基变异并且计算各个变异位点在群体中的MAF值，得到105个样本的群体变异信息累加结果。

（3）计算得到变异位点总深度、变异碱基型深度、MAF值（部分结果如表1所示）后，计算得到相关数值的数据分布形式。发现dbSNP%与变异碱基型深度在 $30\times$ 达到稳定，分析变异位点和MAF数值分布后，设定变异位点总深度的阈值 $30\times$ ，MAF的阈值为0.05，滤除低于阈值的值，得到过滤后的群体变异信息累加结果。图3示出了本实施例中变异碱基型深度与dbSNP%关系图。

表1

染色	位置	dbSNP 数据库	参考	突变	突变碱基	深度	MAF
----	----	-----------	----	----	------	----	-----

体	收录		型	型	型深度		值
chr1	10000400	rs1237370	T	A	10	10	0.06667
chr1	100008709	rs10875232	G	T	7	7	0.06667
chr1	100967839	rs10783133	A	G	150	154	0.3238
chr1	101033987	rs1846215	G	T	96	102	0.4476

(4) 根据105例样本过滤后的群体变异信息累加结果，对每个个体的变异结果进行过滤。若在个体原始的变异信息中存在染色体、位置和变异碱基类型与过滤后的群体变异信息累加结果一致的结果，则保留该结果，否则过滤掉该结果。此外，还包括：若在一个变异位点上存在两种以上的变异，则过滤掉该变异位点。

(5) dbSNP%评估变异的准确性，发现dbSNP%可以达到90.3%（若去掉Y染色体与覆盖度较差的9号染色体dbSNP%为93%）。另外，与对应的高深度外显子测序数据进行对比，比较后一致性可以达到96.45%，结果如表2所示，表中“编号”表示105例样本的编号，表中“未处理”表示未使用本发明的方法得到的结果与对应的高深度外显子测序数据对比的一致性数据，表中“本方法”表示使用本发明的方法得到的结果与对应的高深度外显子测序数据对比的一致性数据。

表2

编号	未处理	本方法	编号	未处理	本方法	编号	未处理	本方法
1	47.78%	96.25%	36	50.27%	92.96%	71	48.66%	93.75%
2	44.68%	98.28%	37	48.78%	86.27%	72	48.53%	91.04%
3	48.32%	100%	38	50.37%	89.80%	73	52.55%	94.00%
4	51.82%	100%	39	51.33%	97.18%	74	46.78%	98.15%
5	50.00%	93.42%	40	66.90%	100%	75	49.69%	91.36%
6	49.73%	90.48%	41	53.37%	100%	76	50.06%	94.81%
7	49.54%	100%	42	49.26%	96.15%	77	47.41%	96.23%

8	54.24%	97.50%	43	47.83%	94.23%	78	49.70%	98.36%
9	51.69%	98.33%	44	50.24%	94.37%	79	47.96%	90.54%
10	50.88%	91.53%	45	50.63%	88.14%	80	47.79%	100%
11	47.90%	98.36%	46	47.28%	97.10%	81	48.36%	98.18%
12	47.22%	98.08%	47	48.41%	90.77%	82	50.66%	96.15%
13	49.02%	100%	48	49.21%	90.91%	83	48.75%	91.94%
14	49.37%	100%	49	53.75%	97.73%	84	53.66%	96.15%
15	46.38%	98.51%	50	50.20%	96.47%	85	49.04%	98.46%
16	50.67%	95.45%	51	48.59%	92.54%	86	52.10%	95.31%
17	48.52%	98.08%	52	50.29%	95.77%	87	48.73%	98.51%
18	50.64%	100%	53	51.07%	90.57%	88	48.56%	92.42%
19	47.65%	100%	54	47.56%	100%	89	54.16%	91.23%
20	48.38%	89.86%	55	50.15%	100%	90	49.41%	89.09%
21	47.45%	98.33%	56	50.33%	89.47%	91	47.39%	95.24%
22	50.30%	100%	57	47.21%	92.54%	92	50.33%	90.91%
23	49.76%	95.65%	58	52.93%	93.24%	93	47.33%	100%
24	48.35%	100%	59	51.90%	94.20%	94	47.68%	92.45%
25	51.32%	94.59%	60	46.73%	91.07%	95	50.21%	96.15%
26	43.85%	96.47%	61	46.64%	94.92%	96	49.18%	85.96%
27	50.00%	91.86%	62	52.33%	91.04%	97	48.77%	100%
28	46.93%	97.22%	63	48.80%	97.22%	98	53.43%	98.18%
29	48.32%	91.18%	64	48.67%	97.30%	99	47.57%	100%
30	49.01%	98.21%	65	49.79%	93.94%	100	51.10%	97.50%
31	48.58%	100%	66	48.90%	94.12%	101	49.99%	100%

32	51.12%	86.79%	67	47.66%	89.80%	102	50.18%	94.81%
33	46.50%	98.41%	68	48.46%	100%	103	46.63%	93.33%
34	49.68%	90.70%	69	53.46%	95.12%	104	50.16%	100%
35	49.63%	92.19%	70	50.78%	95.31%	105	49.22%	98.11%
							平均值	49.01% 96.45%

以上应用了具体个例对本发明进行阐述，只是用于帮助理解本发明，并不用以限制本发明。对于本发明所属技术领域的技术人员，依据本发明的思想，还可以做出若干简单推演、变形或替换。

权 利 要 求 书

1. 一种基于测序数据的变异检测方法，其特征在于，包括：

将来源于同一群体的多个个体的测序数据比对到参考基因组并进行变异检测，得到读长的比对位置和变异信息；

依据所述比对位置和变异信息，针对各个变异位点，将群体中所有个体的变异位点总深度、变异碱基型深度分别加和，然后计算各个变异位点在群体中的次等位基因频率和/或哈迪温伯格平衡，得到群体变异信息累加结果；

绘制变异位点总深度和/或变异碱基型深度与 dbSNP%关系图，并依据设定的过滤阈值对所述变异位点总深度和/或变异碱基型深度以及所述次等位基因频率和/或哈迪温伯格平衡进行过滤以滤除低于所述过滤阈值的值，得到过滤后的群体变异信息累加结果；

依据所述过滤后的群体变异信息累加结果对每个个体的变异信息进行过滤，得到每个个体的最终变异信息。

2. 根据权利要求 1 所述的变异检测方法，其特征在于，所述测序数据的平均测序深度是 $1\times$ 至 $15\times$ 。

3. 根据权利要求 1 所述的变异检测方法，其特征在于，还包括：在将群体中所有个体的变异位点总深度、变异碱基型深度分别加和时，去除三碱基及其它多碱基变异而仅保留单核苷酸多态性变异。

4. 根据权利要求 1 所述的变异检测方法，其特征在于，所述过滤阈值包括所述变异位点总深度的过滤阈值和/或所述变异碱基型深度的过滤阈值，以及所述次等位基因频率的过滤阈值和所述哈迪温伯格平衡的过滤阈值。

5. 根据权利要求 4 所述的变异检测方法，其特征在于，所述变异碱基型深度的过滤阈值是 $30\times$ ，所述次等位基因频率的过滤阈值是 0.05，所述哈迪温伯格平衡的过滤阈值是 0.01。

6. 根据权利要求 1 所述的变异检测方法，其特征在于，所述对每个个体的

变异信息进行过滤具体包括：

若在所述个体原始的变异信息中存在染色体、位置和变异碱基类型与所述过滤后的群体变异信息累加结果一致的结果，则保留该结果，否则过滤掉该结果。

7. 根据权利要求 6 所述的变异检测方法，其特征在于，所述对每个个体的变异信息进行过滤具体还包括：

若在一个所述变异位点上存在两种以上的变异，则过滤掉该变异位点。

8. 一种基于测序数据的变异检测装置，其特征在于，包括：

比对与变异检测装置，用于将来源于同一群体的多个个体的测序数据比对到参考基因组并进行变异检测，得到读长的比对位置和变异信息；

加和与计算装置，用于依据所述比对位置和变异信息，针对各个变异位点，将群体中所有个体的变异位点总深度、变异碱基型深度分别加和，然后计算各个变异位点在群体中的次等位基因频率和/或哈迪温伯格平衡，得到群体变异信息累加结果；

群体变异信息过滤装置，用于绘制变异位点总深度和/或变异碱基型深度与 dbSNP% 关系图，并依据设定的过滤阈值对所述变异位点总深度和/或变异碱基型深度以及所述次等位基因频率和/或哈迪温伯格平衡进行过滤以滤除低于所述过滤阈值的值，得到过滤后的群体变异信息累加结果；

个体变异信息过滤装置，用于依据所述过滤后的群体变异信息累加结果对每个个体的变异信息进行过滤，得到每个个体的最终变异信息。

9. 一种基于测序数据的变异检测装置，其特征在于，包括：

存储器，用于存储程序；

处理器，用于通过执行所述存储器存储的程序以实现如权利要求 1 至 7 中任一项所述的方法。

10. 一种计算机可读存储介质，其特征在于，包括程序，所述程序能够被处理器执行以实现如权利要求 1 至 7 中任一项所述的方法。

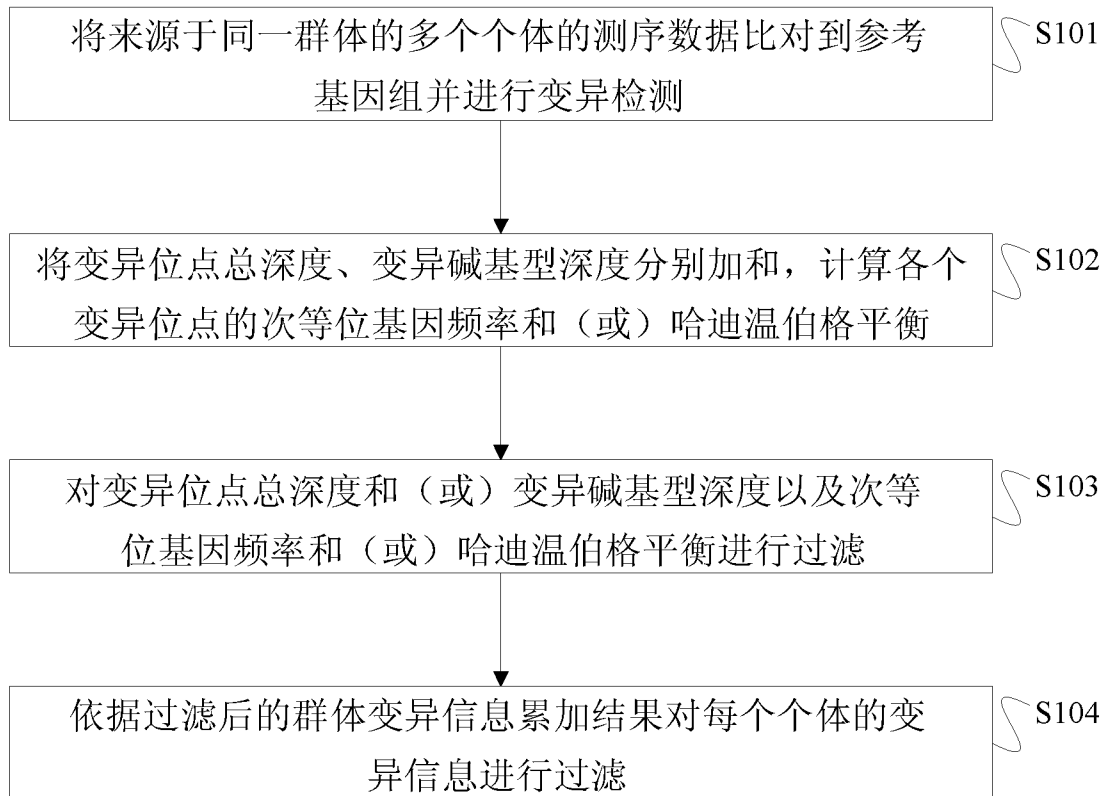


图 1

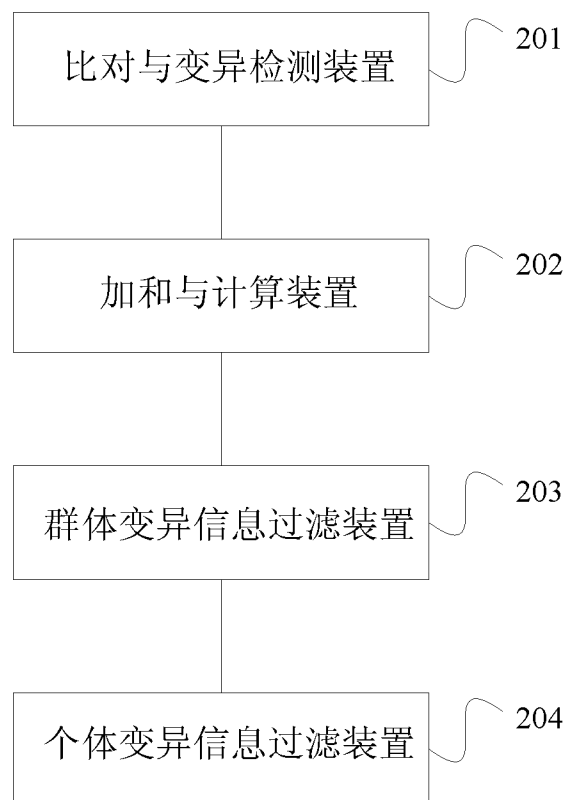


图 2

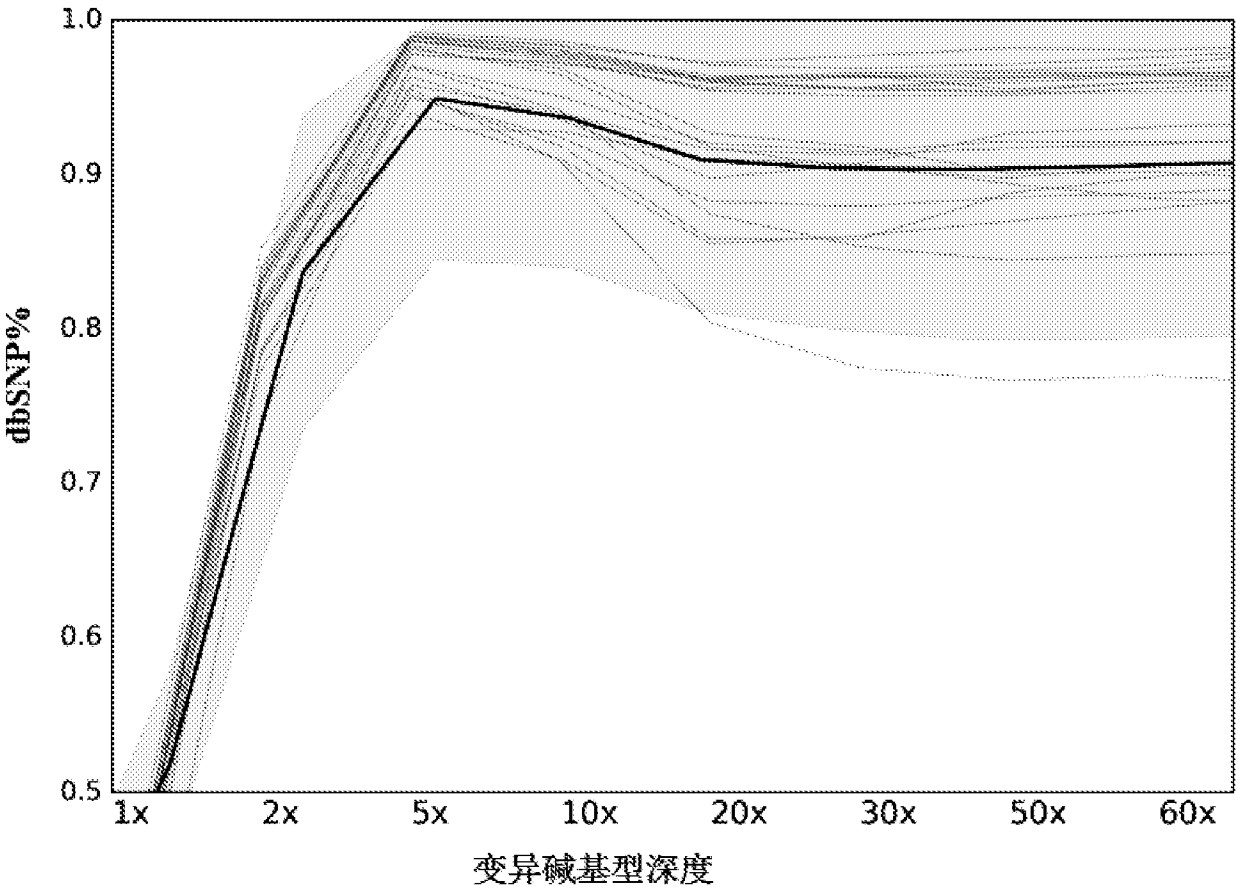


图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/085448

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68 (2006.01) i; C12Q 1/00 (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)

CNABS, SIPOABS, DWPI, CNTXT, VEN, USTXT, WOTXT, CNKI, NCBI, EBI, ISI Web of Knowledge, google, baidu, patentics, WANFANG, DUXIU, SUPERSTAR: 测序数据, 变异检测方法, 同一群体, 多个个体, 测序, 参考基因组, 变异检测, 比对位置, 变异信息, 比对, 变异, 变异位点, 变异位点总深度, 变异碱基型深度, 加和, 次等位基因频率, 哈迪温伯格平衡, 群体变异信息累加结果, dbSNP%关系图, 过滤阈值, 过滤, 滤除, 最终变异信息, 平均测序深度, 三碱基变异, 单核苷酸多态性变异, 染色体, 位置, 变异碱基类型, 群体变异信息, 深圳华大基因研究院, 莫晓东, 杨焕杰, 张涛, SNP, dbSNP, CV, CNV, Minor allele frequency, MAF, Hardy-Weinberg equilibrium, HWE, xiao dong mo, huan jie yang, tao zhang

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 104204220 A (SHENZHEN BGI DIAGNOSIS CO., LTD.) 10 December 2014 (10.12.2014), entire document	1-9
A	CN 103617256 A (BEIJING NOVOGENE BIOINFORMATICS TECHNOLOGY CO., LTD.) 05 March 2014 (05.03.2014), entire document	1-9
A	CN 105441432 A (TIANJIN BGI TECHNOLOGY CO., LTD. et al.) 30 March 2016 (30.03.2016), entire document	1-9

☒ 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 04 January 2018	Date of mailing of the international search report 02 February 2018
--	--

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 ZHU, Ning Telephone No. (86-10) 62414293
---	---

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/085448

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.: 10
because they relate to subject matter not required to be searched by this Authority, namely:
[1] Claim 10 relates the computer readable storage medium and belongs to the information expression simply. Therefore, claim 10 does not meet the requirement of 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 REPORTInternational application No.
PCT/CN2017/085448

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 105760712 A (XIDIAN UNIVERSITY) 13 July 2016 (13.07.2016), entire document	1-9
A	CN 105925685 A (MY HEALTH (TIANJIN) GENE TECHNOLOGY CO., LTD.) 07 September 2016 (07.09.2016), entire document	1-9
A	CN 105989246 A (SHENZHEN BGI RESEARCH) 05 October 2016 (05.10.2016), entire document	1-9
A	CN 106520940 A (SHENZHEN BGI RESEARCH) 22 March 2017 (22.03.2017), entire document	1-9
A	CN 103080333 A (SHENZHEN BGI TECHNOLOGY CO., LTD.) 01 May 2013 (01.05.2013), entire document	1-9
A	WO 2016049878 A1 (BGI SHENZHEN CO. LTD.) 07 April 2016 (07.04.2016), entire document	1-9

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/085448

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104204220 A	10 December 2014	US 2014370504 A1	18 December 2014
		EP 2772549 A4	18 March 2015
		CN 104204220 B	06 June 2017
		JP 2015502749 A	29 January 2015
		WO 2013097062 A1	04 July 2013
		JP 5993029 B2	14 September 2016
		EP 2772549 A1	03 September 2014
CN 103617256 A	05 March 2014	None	
CN 105441432 A	30 March 2016	HK 1218312 A1	10 February 2017
CN 105760712 A	13 July 2016	None	
CN 105925685 A	07 September 2016	None	
CN 105989246 A	05 October 2016	None	
CN 106520940 A	22 March 2017	None	
CN 103080333 A	01 May 2013	CN 103080333 B	24 June 2015
		WO 2012034251 A2	22 March 2012
WO 2016049878 A1	07 April 2016	CN 106715712 A	24 May 2017
		WO 2016049993 A1	07 April 2016

国际检索报告

国际申请号

PCT/CN2017/085448

A. 主题的分类 C12Q 1/68(2006.01)i; C12Q 1/00(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)) CNABS, SIPOABS, DWPI, CNTXT, VEN, USTXT, WOTXT, CNKI, NCBI, EBI, ISI Web of Knowledge, google, baidu, patentics, 万方, 读秀, 超星, 测序数据, 变异检测方法, 同一群体, 多个个体, 测序, 参考基因组, 变异检测, 比对位置, 变异信息, 比对, 变异, 变异位点, 变异位点总深度, 变异碱基型深度, 加和, 次等位基因频率, 哈迪温伯格平衡, 群体变异信息累加结果, dbSNP%关系图, 过滤阈值, 过滤, 滤除, 最终变异信息, 平均测序深度, 三碱基变异, 单核苷酸多态性变异, 染色体, 位置, 变异碱基类型, 群体变异信息, 深圳华大基因研究院, 莫晓东, 杨焕杰, 张涛, SNP, dbSNP, CV, CNV, Minor allele frequency, MAF, Hardy-Weinberg equilibrium, HWE, xiao dong mo, huan jie yang, tao zhang		
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
A	CN 104204220 A (深圳华大基因医学有限公司) 2014年 12月 10日 (2014 - 12 - 10) 全文	1-9
A	CN 103617256 A (北京诺禾致源生物信息科技有限公司) 2014年 3月 5日 (2014 - 03 - 05) 全文	1-9
A	CN 105441432 A (天津华大基因科技有限公司等) 2016年 3月 30日 (2016 - 03 - 30) 全文	1-9
A	CN 105760712 A (西安电子科技大学) 2016年 7月 13日 (2016 - 07 - 13) 全文	1-9
A	CN 105925685 A (万康源天津基因科技有限公司) 2016年 9月 7日 (2016 - 09 - 07) 全文	1-9
☒ 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
2018年 1月 4日		2018年 2月 2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) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		朱宁 电话号码 (86-10)62414293

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
A	CN 105989246 A (深圳华大基因研究院) 2016年 10月 5日 (2016 - 10 - 05) 全文	1-9
A	CN 106520940 A (深圳华大基因研究院) 2017年 3月 22日 (2017 - 03 - 22) 全文	1-9
A	CN 103080333 A (深圳华大基因科技有限公司) 2013年 5月 1日 (2013 - 05 - 01) 全文	1-9
A	WO 2016049878 A1 (BGI SHENZHEN CO. LTD) 2016年 4月 7日 (2016 - 04 - 07) 全文	1-9

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： 10
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10涉及一种计算机可读存储介质，属于单纯的信息表达，因此不符合PCT细则第39.1(v)的规定。
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/CN2017/085448

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)
CN	104204220	A	2014年 12月 10日	US	2014370504	A1	2014年 12月 18日
				EP	2772549	A4	2015年 3月 18日
				CN	104204220	B	2017年 6月 6日
				JP	2015502749	A	2015年 1月 29日
				WO	2013097062	A1	2013年 7月 4日
				JP	5993029	B2	2016年 9月 14日
				EP	2772549	A1	2014年 9月 3日
CN	103617256	A	2014年 3月 5日	无			
CN	105441432	A	2016年 3月 30日	HK	1218312	A1	2017年 2月 10日
CN	105760712	A	2016年 7月 13日	无			
CN	105925685	A	2016年 9月 7日	无			
CN	105989246	A	2016年 10月 5日	无			
CN	106520940	A	2017年 3月 22日	无			
CN	103080333	A	2013年 5月 1日	CN	103080333	B	2015年 6月 24日
				WO	2012034251	A2	2012年 3月 22日
WO	2016049878	A1	2016年 4月 7日	CN	106715712	A	2017年 5月 24日
				WO	2016049993	A1	2016年 4月 7日