(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
2020 年 3 月 5 日 (05.03.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/041946 A1

(51) 国际专利分类号:
C12Q 1/68 (2018.01)

(21) 国际申请号: PCT/CN2018/102546

(22) 国际申请日: 2018 年 8 月 27 日 (27.08.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: 深圳华大生命科学研究院 (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: 张海萍 (ZHANG, Haiping); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 杨林 (YANG, Lin); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 黄国栋 (HUANG, Guodong); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 曾鹏 (ZENG, Peng); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 高雅 (GAO, Ya); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 陈芳 (CHEN, Fang); 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); 中国广东省深圳市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,

(54) **Title:** METHOD AND DEVICE FOR DETECTING HOMOLOGOUS SEQUENCES ON BASIS OF HIGH-THROUGHPUT SEQUENCING

(54) 发明名称: 基于高通量测序检测同源序列的方法和装置

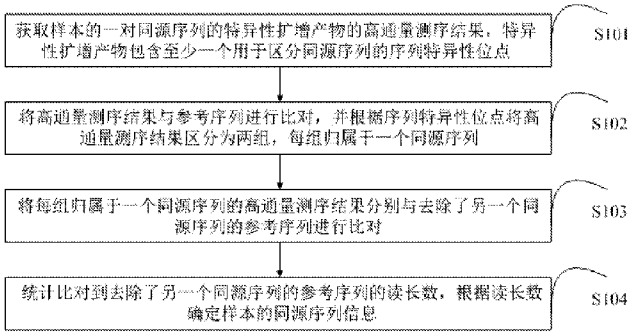

图 1

S101 Obtain high-throughput sequencing results of specific amplification products of a pair of homologous sequences of a sample, the specific amplification products comprising at least one sequence-specific site used for distinguishing homologous sequences

S102 Compare the high-throughput sequencing results with a reference sequence, and dividing the high-throughput sequencing results into two groups according to sequence-specific sites, each group belonging to a homologous sequence

S103 Compare each group of high-throughput sequencing results belonging to one homologous sequence with a reference sequence from which another homologous sequence has been removed

S104 Count the number of reads compared to the reference sequence from which another homologous sequence has been removed, and determining homologous sequence information of the sample according to the number of reads

(57) **Abstract:** A method and device for detecting homologous sequences on the basis of high-throughput sequencing, the method comprising: obtaining high-throughput sequencing results of specific amplification products of a pair of homologous sequences of a sample, the specific amplification products comprising at least one sequence-specific site used for distinguishing homologous sequences; comparing the high-throughput sequencing results with a reference sequence, and dividing the high-throughput sequencing results into two groups according to sequence-specific sites, each group belonging to a homologous sequence; comparing each group of

[见续页]

WO 2020/041946 A1

GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, 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, 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))。

— 包括说明书序列表部分(细则5.2(a))。

high-throughput sequencing results belonging to one homologous sequence with a reference sequence from which another homologous sequence has been removed; and counting the number of reads compared to the reference sequence from which another homologous sequence has been removed, and determining homologous sequence information of the sample according to the number of reads. The present invention may solve the problem of the accurate positioning of homologous sequence sources and achieve the purpose of accurately detecting mutations.

(57) 摘要：一种基于高通量测序检测同源序列的方法和装置，所述方法包括：获取样本的一对同源序列的特异性扩增产物的高通量测序结果，特异性扩增产物包含至少一个用于区分同源序列的序列特异性位点；将高通量测序结果与参考序列进行比对，并根据序列特异性位点将高通量测序结果区分为两组，每组归属于一个同源序列；将每组归属于一个同源序列的高通量测序结果分别与去除了另一个同源序列的参考序列进行比对；和统计比对到去除了另一个同源序列的参考序列的reads数目，根据reads数确定样本的同源序列信息。本发明能够解决同源序列来源的精确定位问题，实现准确检测突变的目的。

基于高通量测序检测同源序列的方法和装置

技术领域

本发明涉及生物信息学技术领域，具体涉及一种基于高通量测序检测同源序列的方法和装置。

背景技术

同源基因是指序列相似度大于 80%的 2 个或多个基因。基于高通量测序的数据结果，同源基因区域的 reads 在比对时，无法正确比对到合适位置而导致出现多重比对的情况，大多数情况下这样的 reads 没法正确反映目标位置的碱基情况。因此对于同源基因的正确比对会遇到一定困难，这使得无法使用现有的分析流程对下机数据直接进行突变分析。例如，临床上需要基于高通量的方法对 RHD 血型进行鉴定，由于 RHD 基因存在高度同源的 RHCE 基因（96%相似性），下机数据无法正确比对到 RHD 基因上。这导致一些包含同源基因的遗传病无法进行准确的检测。

Rh 血型 D 抗原(D 抗原在红细胞膜表达与否定义为 Rh 阳性或 Rh 阴性)是引起严重新生儿溶血病的主要红细胞抗原，编码 D 抗原的基因为 RHD 基因，一般一名 Rh 阳性个体拥有一条 RHD 基因[RHD 杂合子，RHD(+)/RHD(-)]或二条 RHD 基因[RHD 纯合子，RHD(+)/RHD(+)]，Rh 阴性个体则缺失 RHD 基因[RHD 缺失纯合体，RHD(-)/RHD(-)]，有些复杂的 Rh 阴性个体往往出现基因融合或者出现某几个外显子缺失的现象。以往 RHD 基因数目或 RHD 合子型测定主要是根据 Rh 小因子表型估计，或通过复杂的家系调查，或根据 RhD 抗原的量等间接方法鉴别，2000 年国际上建立了第一个 RHD 基因合子型直接测定技术，为限制性片段长度多态性(RFLP)方法，该方法采用一对 PCR 引物同时扩增下游和融合 Rh 盒子，然后采用限制性内切酶进行酶切，再电泳分析片段大小多态性，一次实验结果可以判断三种 RHD 合子型，但是该方法操作复杂、耗时较长且针对的是白种人特有的 RHD 阴性型别。2001 年和 2002 年，中国香港学者和奥地利学者分别独立报道了采用扩增突变阻滞检测系统和实时聚合酶链式反应(Real-time PCR)技术检测已知 Rh 阳性表型个体的 RHD 基因数目，该技术同样较为复杂且要求特殊的仪器，而且不能区分 RHD(+)/RHD(-)和 RHD(-)/RHD(-)合子型以及缺失外显子个数及类别的详细信息。

听力语言残疾居各类残疾之首，每年 1000 个新生儿中约有 1~3 个聋儿，其中 60%的聋病与遗传因素相关。中国听力障碍者 2780 万人，耳聋基因突变携带者约为 7800 万人，药物敏感性突变携带者(高危人群)约 400 万人，在我国每年新增的聋儿中有近半数为药物性耳聋。药物性耳聋还会在母系家族人群中遗传。因此，尽早发现药物致聋基因，避免用药损伤听力，防止家人和孩子用药致聋是急需解决的问题。与药物致聋相关的基因为 CYP2D6，如果对其检测可以预防绝大部分药物导致的耳聋，但由于 CYP2D6 有一个相似度高达 94%的基因 CYP2D7，使用现有的信息分析比对算法很难确定测序 reads 究竟是来源于 CYP2D6 还是 CYP2D7，因此需要一种方法来区分 CYP2D6 和 CYP2D7 的测序 reads，从而准确地检测发生

在 CYP2D6 上的突变。

对于药物致聋基因 CYP2D6 用药位点检测，现有的方法可以采用质谱或 sanger 测序来检测，但这两种方法都遇到通量低等问题，而高通量测序能够一次对多个位点多个基因进行检测，但是高通量测序遇到的问题是无法准确区分来源于 CYP2D6 和 CYP2D7 的 reads，因此检测的准确性遇到挑战。

发明内容

本发明提供一种基于高通量测序检测同源序列的方法和装置，能够解决同源序列来源的精确定位问题，实现准确检测突变的目的。

根据第一方面，一种实施例中提供一种基于高通量测序检测同源序列的方法，包括：

获取样本的一对同源序列的特异性扩增产物的高通量测序结果，上述特异性扩增产物包含至少一个用于区分同源序列的序列特异性位点；

将上述高通量测序结果与参考序列进行比对，并根据上述序列特异性位点将上述高通量测序结果区分为两组，每组归属于一个同源序列；

将每组归属于一个同源序列的高通量测序结果分别与去除了另一个同源序列的参考序列进行比对；和

统计比对到去除了另一个同源序列的参考序列的 reads 数目，根据上述 reads 数目确定上述样本的同源序列信息。

优选地，上述特异性扩增产物是使用靶向一对同源序列的多个区域的多对引物进行特异性扩增得到的产物。

优选地，上述同源序列是同源基因。

优选地，上述同源基因是 RHD 和 RHCE 基因，或 CYP2D6 和 CYP2D7 基因。

优选地，上述特异性扩增产物是使用靶向同源基因的多个外显子区域的多对引物进行特异性扩增得到的产物。

优选地，上述序列特异性位点是单核苷酸变异（SNV）位点。

优选地，上述去除了另一个同源序列的参考序列是另一个同源序列全部替换为 N 碱基序列的参考序列。

优选地，上述根据上述 reads 数目确定上述样本的同源序列信息，具体为：根据上述 reads 数目确定上述样本的同源序列的突变信息。

优选地，上述根据上述 reads 数目确定上述样本的同源序列信息，具体为：根据对比到上述同源序列的两个不同来源的 reads 数目的差异，确定上述同源序列在基因组上属于正常、发生缺失或重复的情况。

优选地，上述将上述高通量测序结果与参考序列进行比对后，根据 CIGAR 值矫正比对后的结果，以准确地根据上述序列特异性位点将上述高通量测序结果进行区分。

优选地，上述将每组归属于一个同源序列的高通量测序结果分别与去除了另一个同源序列的参考序列进行比对后，根据 CIGAR 值矫正比对后的结果。

优选地，上述将上述高通量测序结果与参考序列进行比对之前，还包括：去除上述高通量测序结果两端的接头序列。

优选地，上述将每组归属于一个同源序列的高通量测序结果分别与去除了另一个同源序列的参考序列进行比对之后，还包括：

过滤掉插入片段长度大于预设值或两端序列比对到不同染色体的结果；和/或

去除测序质量值低于预设值的碱基位点。

根据第二方面，一种实施例中提供一种基于高通量测序检测同源序列的装置，包括：

获取单元，用于获取样本的一对同源序列的特异性扩增产物的高通量测序结果，上述特异性扩增产物包含至少一个用于区分同源序列的序列特异性位点；

第一比对单元，用于将上述高通量测序结果与参考序列进行比对，并根据上述序列特异性位点将上述高通量测序结果区分为两组，每组归属于一个同源序列；

第二比对单元，用于将每组归属于一个同源序列的高通量测序结果分别与去除了另一个同源序列的参考序列进行比对；和

统计单元，用于统计比对到去除了另一个同源序列的参考序列的 reads 数目，根据上述 reads 数目确定上述样本的同源序列信息。

根据第三方面，一种实施例中提供一种计算机可读存储介质，包括程序，该程序能够被处理器执行以实现如第一方面的方法。

本发明的方法，通过序列特异性位点区分同源序列，将来源于同源序列的特异性扩增产物的高通量测序结果区分为两组，然后将每组数据分别比对去除了另一个同源序列的参考序列，从而得到分别属于同源序列中每一种类型序列的测序 reads 的数量，实现了准确区分测序序列来源，进而实现准确检测突变的目的。

附图说明

图 1 为本发明实施例的基于高通量测序检测同源序列的方法流程图；

图 2 为本发明实施例的基于高通量测序检测同源序列的装置结构框图；

图 3 为本发明实施例的 RHD 血型鉴定分析流程图；

图 4 为本发明实施例的 RHD 血型鉴定中引物均一性检测结果图；

图 5 为本发明实施例的 RHD 血型鉴定中特异性位点区分到不同基因的测序 reads 数目统计图；

图 6 为本发明实施例的药物致聋 CYP2D6 基因检测流程图；

图 7 为本发明实施例的药物致聋 CYP2D6 基因检测中引物均一性检测结果图；

图 8 为本发明实施例的药物致聋 CYP2D6 基因检测中特异性位点区分到不同基因的测序 reads 数目统计图。

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。

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。

现有技术中，基于高通量测序的数据结果，同源基因区域的测序 reads 在比对时，无法正确比对到合适位置而导致出现多重比对的情况，大多数情况下这样的 reads 没法正确反映目标位置的碱基情况。因此对于同源基因的正确比对会遇到一定困难，这使得无法使用现有的分析流程对下机数据直接进行突变分析，造成对于存在同源基因的相关遗传病无法进行准确检测。

针对现有技术中的问题，本发明提供一种基于高通量测序检测同源序列的方法，该方法包括使用特异性引物扩增一对（2 个）同源序列，该引物扩增区域包含至少一个序列特异性位点，用于区分一对同源序列，通过第一次比对找到测序 reads 可能存在的位置，通过序列特异性位点区分不同的同源序列，将区分好的测序 reads 进行重新比对，从而对突变进行精确的定位，通过比对结果进行突变的检测。

如图 1 所示，一种实施例中提供一种基于高通量测序检测同源序列的方法，包括：

S101：获取样本的一对同源序列的特异性扩增产物的高通量测序结果，上述特异性扩增产物包含至少一个用于区分同源序列的序列特异性位点。

本发明实施例中，“样本”即本发明的检测方法所针对的样本，可以是临床上的样本，包括健康人样本和病人样本，例如来源于健康人或病人的血液、脑脊液样本等。这些样本经本领域公知的技术进行核酸（如 DNA）提取，获取样本中的核酸序列片段，使用特异性靶向同源序列的引物扩增目标片段，得到特异性扩增产物，测序后得到高通量测序结果。测序平台不限，可以是任何第二代高通量测序平台，包括但不限于 Illumina、Ion Torrent、BGISEQ 或 MGISEQ 测序平台等。

本发明实施例中，“同源序列”一般是指序列相似度大于 80%的序列，但这仅是一种示例性的界定同源序列的方式。本发明实施例中，同源序列的类型没有限制，可以是同源基因序列（例如，含有可读编码框的序列），也可以是非基因类型的同源序列。典型但非限定性的同源序列的例子是 RHD 和 RHCE 基因，以及 CYP2D6 和 CYP2D7 基因。

本发明实施例中，“特异性扩增产物”是通过特异性扩增引物对同源序列的相应位置进行靶向扩增得到的。针对同源序列设计特异性扩增引物，每对引物同时扩增一对同源序列的相应位置，该位置至少包含一个序列特异性位点，该位点用于准确区分序列来源，准确定位测序 reads 来源，达到准确检测突变的目的。

在一些实施例中，特异性扩增产物是多重扩增的结果，即使用靶向一对同源序列的多个区域的多对引物进行特异性扩增得到的产物。例如，在同源序列是同源基因的情况下，使用靶向同源基因的多个外显子区域的多对引物进行特异性扩增得到特异性扩增产物。

本发明实施例中，“序列特异性位点”是指在一对同源序列的对应位置上彼此不同的位点，该序列特异性位点至少包含 1bp 特异性碱基，例如单核苷酸变异（SNV）位点。在其它实施例中，序列特异性位点还可以是碱基插入、缺失或拷贝数变异等。

在一些实施例中，作为本发明的方法的输入数据，高通量测序结果是指下机数据经过一定预处理的结果，例如对下机数据去除测序 reads 两端的接头序列，能够提高比对准确率和数据有效性。

S102：将高通量测序结果与参考序列进行比对，并根据序列特异性位点将高通量测序结果区分为两组，每组归属于一个同源序列。

本发明实施例中，“参考序列”一般是指同源序列对应的物种的基因组序列等，例如人类参考基因组序列等，尤其是人类参考基因组 hg19 等。

由于同源序列在序列特异性位点上碱基不同，因此序列特异性位点上的碱基型可以作为区分高通量测序结果的标志。经过该步骤，来源于一对同源序列的多条（可能成千上万条）测序 reads 就被分到了每种同源序列类型中。

作为本发明的优选方案，将高通量测序结果与参考序列进行比对后，根据 CIGAR 值矫正比对后的结果，以准确地根据序列特异性位点将高通量测序结果进行区分。CIGAR 值矫正在本发明中能够提高测序 reads 区分到各自所属组别中的准确度，因此是一种优选的实施方式。

S103：将每组归属于一个同源序列的高通量测序结果分别与去除了另一个同源序列的参考序列进行比对。

本发明实施例中，从参考序列中去除另一个同源序列的目的是，为了得到某一同源序列（例如 RHD 基因）比对到参考序列上的绝对位置，避免序列比对到另一同源序列（例如 RHCE 基因）而导致测序 reads 产生错误的比对位置。

在一些实施例中，通过将另一个同源序列全部替换为 N 碱基序列，来实现从参考序列中去除另一个同源序列。

类似于步骤 S102，在步骤 S103 中，根据 CIGAR 值矫正比对后的结果，能够提高测序 reads 的比对准确度，因此是一种优选的实施方式。

在一些实施例中，考虑到非正常测序结果对后续统计准确性的影响，在步骤 S103 的比对之后还包括如下任一项或多项：（a）过滤掉插入片段长度大于预设值（例如 500）或两端序列比对到不同染色体的结果；（b）去除测序质量值低于预设值（例如 10）的碱基位点。经过该处理再进行步骤 S104 的统计能够提高结果准确性。

S104：统计比对到去除了另一个同源序列的参考序列的 reads 数目，根据 reads 数目确定样本的同源序列信息。

本发明实施例中，比对到参考序列上的 reads 数目，反映了一个同源序列基因型在参考基因组中的存在情况，例如剂量。因此，统计比对到去除了另一个同源序列的参考序列的 reads 数目，就能够确定样本的同源序列信息，例如同源序列的突变信息，诸如同源序列在基因组上属于正常、发生缺失或重复的情况。

在一些实施例中，所谓“根据 reads 数目确定样本的同源序列信息”，具体为：根据 reads

数目确定样本的同源序列的突变信息。在一些实施例中，所谓“根据 reads 数目确定样本的同源序列信息”，具体为：根据对比到同源序列的两个不同来源的 reads 数目的差异，确定同源序列在基因组上属于正常、发生缺失或重复的情况。

如图 2 所示，对应于本发明的基于高通量测序检测同源序列的方法，本发明一种实施例中提供一种基于高通量测序检测同源序列的装置，包括：获取单元 201，用于获取样本的一对同源序列的特异性扩增产物的高通量测序结果，上述特异性扩增产物包含至少一个用于区分同源序列的序列特异性位点；第一比对单元 202，用于将上述高通量测序结果与参考序列进行比对，并根据上述序列特异性位点将上述高通量测序结果区分为两组，每组归属于一个同源序列；第二比对单元 203，用于将每组归属于一个同源序列的高通量测序结果分别与去除了另一个同源序列的参考序列进行比对；和统计单元 204，用于统计比对到去除了另一个同源序列的参考序列的 reads 数目，根据上述 reads 数目确定上述样本的同源序列信息。

相应地，本发明一种实施例中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本发明的基于高通量测序检测同源序列的方法。

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。

以下通过实施例详细说明本发明的技术方案，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 1：针对 RHD 样本进行检测

针对RHD基因和RHCE基因的10个外显子设计10对引物，每对引物分别扩增RHD基因和RHCE基因的一个外显子区域，通过扩增分别得到20个产物，对20个产物进行比对分类，根据特异性位点精确区分测序reads的来源，然后再一次进行比对，最后对比对后的结果进行统计，统计所有外显子上的测序reads覆盖深度。通过对比RHD基因和RHCE基因某一外显子上的覆盖度，来判断RHD基因的这个外显子是否发生缺失和重复，相同地，将所有外显子缺失和重复的情况进行统计，得到所有外显子的缺失情况，最后得到检测个体的合子类型RHD(+)/RHD(+)、RHD(+)/RHD(-)或RHD(-)RHD(-)，达到准确进行RHD血型鉴定的目的。

本实施例，包括实验部分和生物信息分析部分。实验部分包括：针对同源基因设计特异性引物并进行多重PCR扩增，完成高通量测序文库的制备，引物扩增得到的同源基因区域至少包含1bp差异序列，用于区分同源基因序列。生物信息分析部分包括：针对引物扩增得到的序列进行比对，通过第一次比对找到序列可能存在的位置，利用差异性位点区分不同的同源

基因的序列，将区分好的序列进行重新比对，通过比对的结果进行突变的检测，通过对比2个同源基因某个区域的测序reads覆盖深度，来判断该区域是否存在缺失或重复。通过精准定位的测序reads来进行突变检测，从而对遗传病突变进行准确检测。

具体而言，如图3所示，RHD血型鉴定分析流程包括：

（1）RHD基因和RHCE基因分别包含10个外显子区域，设计10对引物分别同时扩增RHD基因和RHCE基因对应的外显子序列，每个外显子序列包含至少1bp的特异性碱基，用于准确区分RHD基因和RHCE基因。

（2）通过PCR扩增得到二代测序上机文库，通过高通量测序以进行测序，分别得到10个外显子的测序reads信息。

（3）将测序得到的reads比对参考基因组序列，根据指定位置的特异性位点（RHD和RHCE分别对应的碱基），精准区分RHD和RHCE的序列，将含有RHD特异性位点的reads归类为RHDreads集合，将含有RHCE特异性位点的reads归类为RHCEreads集合。

（4）将分别得到的RHD集合的reads再一次比对参考基因组序列，此时RHCE基因在人类参考基因组上的碱基序列替换为N，这样可以保证RHD的序列无法比对到RHCE上，而只能比对到RHD上，从而得到最准确的比对结果，并得到每条read在参考基因组上正确的位置信息，由于后续SNV的检测，避免RHD序列比对到RHCE基因的现象；同理得到RHCE集合的reads也再一次比对参考基因组序列，此时RHD基因上在人类参考基因组上的碱基序列全部替换为N。

（5）根据引物的起始位置和序列的起始位置以及引物的长度，来对引物进行处理，保证最准确的对引物序列进行去除，从而保留最准确的真实序列信息。

（6）通过去除低质量的位点，可以降低错误率或者其他噪音对四碱基统计结果的影响。

以下是本实施例的具体实验部分和生物信息分析部分。

采用10对特异性引物分别对已知基因型别的RHD纯合阳性个体1（2条染色体10个外显子都正常）、RHD杂合阳性个体2（一条染色体10个外显子完全缺失）、RHD阴性个体3（2条染色体完全缺失）进行PCR扩增，扩增得到的产物上机测序，对测序结果分析来判断RHD型别。

目标区域引物设计：针对 RHD 基因进行引物设计，10 对引物覆盖 RHD 基因 10 个外显子区域，每对引物同时扩增 RHD、RHCE 的同一个同源外显子，得到的扩增产物至少包含 1bp 特异性序列，用于后续测序区分序列的来源。引物序列如表 1。

表 1 特异性引物池 1

引物编号	序列
RHD-RHCE-C01F	GACCGCTTGGCCTCCGACTTCGGCGCTGCCTGCCCCTCT（SEQ ID NO：1）
RHD-RHCE-C02F	GACCGCTTGGCCTCCGACTTTTGGCTTGGGCTTCCTC（SEQ ID NO：2）
RHD-RHCE-C03F	GACCGCTTGGCCTCCGACTTGTGGAGGTGACAG

	CTTTAGG（SEQ ID NO：3）
RHD-RHCE-C04F	GACCGCTTGGCCTCCGACTTGCCTGCCAAAGCCT CTACC（SEQ ID NO：4）
RHD-RHCE-C05F	GACCGCTTGGCCTCCGACTTCTGCTCTGCTGAGA AGTCCA（SEQ ID NO：5）
RHD-RHCE-C06F	GACCGCTTGGCCTCCGACTTGTCTTGTGGCTGGG CTGATCT（SEQ ID NO：6）
RHD-RHCE-C07F	GACCGCTTGGCCTCCGACTTTGTTGTAACCGAGT GCTGGGGATTC（SEQ ID NO：7）
RHD-RHCE-C08F	GACCGCTTGGCCTCCGACTTCTTGGCCATCGTGA TAGCTCTC（SEQ ID NO：8）
RHD-RHCE-C09F	GACCGCTTGGCCTCCGACTTTCTTAAAATATGGAA AGCACCTCATG（SEQ ID NO：9）
RHD-RHCE-C10F	GACCGCTTGGCCTCCGACTTACGCTCATGACAGCA AAGTCTC（SEQ ID NO：10）
RHD-RHCE-C01R	GACATGGCTACGATCCGACTTTTGATCCTCTAAGG AAGCGTCA（SEQ ID NO：11）
RHD-RHCE-C02R	GACATGGCTACGATCCGACTTATTGCCCACTGCAC ACCAAGCG（SEQ ID NO：12）
RHD-RHCE-C03R	GACATGGCTACGATCCGACTTCTTTTCTCCCAGGT CCCTC（SEQ ID NO：13）
RHD-RHCE-C04R	GACATGGCTACGATCCGACTTTTGTCCTTACCCAG CATGG（SEQ ID NO：14）
RHD-RHCE-C05R	GACATGGCTACGATCCGACTTCCTGAGATGGCTGT CACCAC（SEQ ID NO：15）
RHD-RHCE-C06R	GACATGGCTACGATCCGACTTAGTTGTCTAGTTTCT TACC（SEQ ID NO：16）
RHD-RHCE-C07R	GACATGGCTACGATCCGACTTATCTCTCCAAGCAG ACCCAGCAAGC（SEQ ID NO：17）
RHD-RHCE-C08R	GACATGGCTACGATCCGACTTTGTCCTGGCAATGG TGGAAGA（SEQ ID NO：18）
RHD-RHCE-C09R	GACATGGCTACGATCCGACTTTCATGCACTCAAAAT CTATCACG（SEQ ID NO：19）
RHD-RHCE-C10R	GACATGGCTACGATCCGACTTATGGTGAGATTCTCC TCAAAGAGT（SEQ ID NO：20）

特异性引物池1由上述引物等摩尔数混合得到。

实验部分：

1. 第一轮PCR扩增

PCR扩增酶采用美国kapa公司的KAPA2G Fast Multiplex PCR Kit产品（货号KK5801）：

PCR反应体系如下表2：

表2

试剂	体积
2Xkapa聚合酶混合液	25μl
特异性引物池1（10μM）	5μl
DNA	20μl
总共	50μl

扩增体系如下表3：

表3

步骤1	98℃，2min
步骤2	98℃，10s
步骤3	62℃，2min
步骤4	72℃，30s
步骤5	重复步骤2-4，15个循环
步骤6	72℃，5min

加入 1 倍体积的 Agencourt AMPure XP 磁珠（美国贝克曼库尔特有限公司）50μl，按照说明书进行纯化，纯化后用 20μl 蒸馏水溶解 DNA。

2. 第二轮 PCR 扩增

PCR 扩增酶采用美国 kapa 公司的 KAPA2G Fast Multiplex PCR Kit 产品（货号 KK5801）。

PCR反应体系如下表4：

表4

试剂	体积
2Xkapa聚合酶混合液	25μl
通用引物1（10μM）	2.5μl
通用引物2（10μM）	2.5μl
DNA	20μl
总共	50μl

通用引物如表5。

表5 通用引物

引物编号	引物序列(5'-3')
*通用引物1	Phos/GAACGACATGGCTACGATCCGACTT（SEQ ID NO：21）
通用引物2	TGTGAGCCAAGGAGTTGCTGCGTACATT TGTCTTCCTAAGACCGCTTGGCCTCCGACTT（SEQ ID NO：22）

*通用引物 1 的 5'端进行了磷酸化修饰，用于后续 BGISEQ-500 平台上的单链环化。

扩增体系如下表6：

表6

步骤1	98℃，2min
步骤2	98℃，10s
步骤3	62℃，2min
步骤4	72℃，30s
步骤5	重复步骤2-4，15个循环
步骤6	72℃，5min

加入 1 倍体积的 Agencourt AMPure XP 磁珠（美国贝克曼库尔特有限公司）50μl，按照说明书进行纯化，纯化后用 20μl 蒸馏水溶解 DNA。

3. 上机测序

文库质检合格后采用华大基因 BGISEQ-500 平台进行测序，测序类型双端 50bp。

数据分析部分：

1. 使用 cutadapt 去除两端含有接头的序列。

2. 通过 bwtsw 算法对人类参考基因组构建比对索引，bwa 版本为 0.7.15。

3. 通过 BWA-ALN 算法将目标序列比对到人类参考基因组 hg19，bwa 版本为 0.7.15，samtools 版本为 0.1.18。

将比对到 hg19 的数据进行提取同时将生成的*. map.bam 转化为*. map.flag，其目的是将第二列的 FLAG 值数值表达形式转变为字母表达形式，可以用于区分 R1 或 R2，从而用于计算各目标区域的引物富集程度。

4. 对每个样本的原始测序 reads 数、去除测序接头后的 reads 数目、比对率、目标区域数据所占比例等基本信息进行统计（表 7）。同时对每个目标区域的富集程度进行统计，从而评估引物的均一性，图 4 结果显示：10 对引物测序得到的深度差别不大，所有区域的深度都在

平均深度的 0.4X 以上，最低深度和最高深度的差值在 5 倍以内。

表 7 下机数据统计

样本编号	原始reads数目	去接头reads数目	比对reads数目	目标区域reads数目
1	10987	10664	10598	10464
2	13231	12876	12784	12643
3	12751	12404	12301	12131

显示：下机数据的数据利用率达到 97%，比对率达到 99%，目标区域比例 98%。

在计算引物富集程度时，如果当前这条序列有任一 1bp 在目标区域内，就认为此条序列就是这个目标区域的数据。

5. 通过 sam 文件的第六列内容，即 CIGAR 值将比对后的结果进行矫正，将 reads 还原成最原始的状态，具体矫正实例如下，结果如表 8 所示。

矫正实例：3M1I46M：与参考基因组相比，此条序列前 3 个碱基可以比对到参考基因组，第 4 个碱基为多出的碱基，第 5 个碱基开始可以比对到参考基因组，因此需要将第 4 个碱基删除；3M1D47M：与参考基因组相比，此条序列前 3 个碱基可以比对到参考基因组，第四个位置缺失一个碱基，第 4 个碱基开始可以比对到参考基因组，因此需要在第四个位置添加字母 D；48M2S：与参考基因组相比，此条序列前 48 个碱基可以比对到参考基因组，最后 2 个碱基无法比对到参考基因组，因此需要将最后 2 个碱基删除；3S47M：与参考基因组相比，此条序列前 3 个碱基无法比对到参考基因组，从第 4 个碱基开始可以比对到参考基因组，因此需要将开始 3 个碱基删除。

表 8 矫正前和矫正后结果

CIGAR 值	矫正前	矫正后
3M1I46M	GTCTTGTCCGGCGCTGCCTGCCCCTCTGGGCCCTAACACTGGAAGCAGCT（SEQ ID NO：23）	GTCTGTCCGGCGCTGCCTGCCCCTCTGGGCCCTAACACTGGAAGCAGCT（SEQ ID NO：24）
3M1D47M	GTCGTCCGGCGCTGCCTGCCCCTCTGGGCCCTAACACTGGAAGCAGCTCT（SEQ ID NO：25）	GTCDGTCCGGCGCTGCCTGCCCCTCTGGGCCCTAACACTGGAAGCAGCTCT（SEQ ID NO：26）
48M2S	AGCTGCTTCCAGTGTTAGGGCCCAGAGGGGCNGGCAGCGCCGGNCAGAAC（SEQ ID NO：27）	AGCTGCTTCCAGTGTTAGGGCCCAGAGGGGCNGGCAGCGCCGGNCAGA（SEQ ID NO：28）
3S47M	GTTCTGCTTCCAGTGTTGGGGCCCAGAGGGGCAGGCAGCGCCGGACAGAC（SEQ ID NO：29）	CTGCTTCCAGTGTTGGGGCCCAGAGGGGCAGGCAGCGCCGGACAGAC（SEQ ID NO：30）

6. 用blast对RHD和RHCE两个基因的10个外显子区域进行比对，利用比对后的测序reads差异位点的序列区分RHD/RHCE，并对应生成只包含RHD数据和只包含RHCE数据的两个文件。

用于区分的绝对位置以及RHD和RHCE对应的碱基型别如下表9所示。

表9 区分RHD和RHCE来源的特异性位点及碱基

外显子编号	CHR	位置	RHD 特异性位点	RHCE 特异性位点
RHD-EX1	chr1	25599086	G	C
RHD-EX2	chr1	25611116	G	A
RHD-EX3	chr1	25617251	A	C
RHD-EX4	chr1	25627527	G	A
RHD-EX5	chr1	25628073	G	C
RHD-EX6	chr1	25629927	G	A
RHD-EX7	chr1	25633115	C	A
RHD-EX8	chr1	25643553	T	C
RHD-EX9	chr1	25648419	A	T
RHD-EX10	chr1	25655520	A	T
RHCE-EX10	chr1	25688913	T	A
RHCE-EX9	chr1	25696992	T	A
RHCE-EX8	chr1	25701857	A	G
RHCE-EX7	chr1	25712307	G	T
RHCE-EX6	chr1	25715490	C	T
RHCE-EX5	chr1	25717344	C	G
RHCE-EX4	chr1	25718542	C	T
RHCE-EX3	chr1	25729118	T	G
RHCE-EX2	chr1	25735308	C	T
RHCE-EX1	chr1	25747230	A	G

例如，一条序列覆盖chr1：25599086这个位置，如果测序reads比对到这个位置出现的碱基为G，认为这条测序read属于RHD基因，如果出现的碱基是C，认为这条测序read属于RHCE基因。

7. 对参考基因组序列进行处理，将RHCE基因处的碱基替换为N，记为HG19 RHCE$^-$参考集；对参考基因组序列进行处理，将RHD基因处的碱基替换为N，记为HG19 RHD$^-$参考集。

将步骤6中得到的RHD文件和RHCE文件分别比对HG19 RHCE$^-$参考集和HG19 RHD$^-$参考集。这样可以保证RHD的序列无法比对到RHCE上，而只能比对到RHD上，同理RHCE的序列

也无法比对到RHD上，从而得到最准确的比对结果，并得到每条测序read在参考基因组上正确的位置信息，由于后续SNV的检测，避免RHD序列比对到RHCE基因的现象。

8. 过滤掉插入片段长度大于500或者PE测序reads比对到不同染色体的序列。

9. 通过重新比对后生成的sam文件的第六列内容，即CIGAR值将比对后的结果进行矫正，将测序reads还原成最原始的状态。

10. 去除低质量（测序质量值小于10）的碱基位点，同时进行标记，表10示出了一个实例。

将每个碱基对应的质量值ASCII转化为对应的十进制数值，然后减去33即可得到对应的质量值，如果这个值小于10，则将此碱基用*号代替。在python中可使用公式：碱基质量值=ord(ASCII) – 33。

表10 去低质量点实例

转化前序列	转化前碱基质量	转化后序列	转化后碱基质量
TCTCTTATTGGCTTCAACGCCTAGTGAGGGATCCATCCTGGCACGGTGGC（SEQ ID NO：31）	AFFFCF:EFDFCF+GFDGCAFFEDGFEF?F:F4GF59EEFFF(FF>DFFD	TCTCTTATTGGCT*CAACGCCTAGTGAGGGATCCATCCTGGC*CGGTGGC（SEQ ID NO：32）	AFFFCF:EFDFCF+GFDGCAFFEDGFEF?F:F4GF59EEFFF(FF>DFFD

11. 对目标位点或区域进行统计。

对目标区域序列进行计数，并比较两者的数量差异（图5）。结果显示：RHD纯合阳性个体1中RHD和RHCE10个外显子的深度覆盖差别不大，与RHCE相比RHD的10个外显子都正常，因此可以判断该RHD基因是纯合的RHD（+）/RHD（+）；RHD杂合阳性个体2中RHD外显子的覆盖度是RHCE的一半左右，与RHCE相比RHD缺失一半，因此可以判断该RHD基因是杂合的RHD（+）/RHD（-）；RHD阴性个体3中RHD外显子的覆盖度基本没有，与RHCE相比几乎不存在测序reads覆盖，因此可以判断该RHD基因10个外显子缺失，且是纯合缺失RHD（-）/RHD（-）。

可见，基于多重PCR捕同源基因的特异性序列，结合二代测序和信息分析方法，来准确对某一同源基因区域进行剂量分析，通过同源序列的差异可以得到基因是否发生缺失和重复。

实施例2：针对耳聋用药位点进行检测

CYP2D6基因存在几个SNP位点与药物致聋相关，不同的SNP碱基信息和药物代谢能力相关，但CYP2D6有一个相似度为94%的同源基因CYP2D7，基于PCR的方法很难避免扩增到CYP2D7。

采用5对特异性引物分别对2个已知用药位点碱基信息的样本（样本1、样本2）进行检测，扩增得到的产物上机测序，对测序结果分析，用特异性位点区分CYP2D6和CYP2D7的测序reads，然后根据区分结果准确检测CYP2D6和药物相关代谢位点的碱基信息。检测流程如图6所示。

目标区域引物设计：针对CYP2D6基因进行引物设计，5对引物覆盖CYP2D6基因5个用药相关的SNP位点（如表11），每对引物同时扩增CYP2D6、CYP2D7的同一个位点，得到的扩增产物至少包含1bp特异性序列（表12），用于后续测序区分序列的来源。

表11 CYP2D6基因5个用药相关的SNP位点

rs号	碱基
rs35742686	-/A
rs3892097	A/G
rs5030865	A/T/C/G
rs1065852	C/T
rs28624811	A/T

表12 用于区分CYP2D6和CYP2D7的绝对位置及对应的碱基型别

外显子编号	CHR	位置	RHD 特异性位点	RHCE 特异性位点
CYP2D6-1	Chr22	42523776	A	T
CYP2D6-2	Chr22	42524218	G	G
CYP2D6-3	Chr22	42524982	C	T
CYP2D6-4	Chr22	42525038	A	C
CYP2D6-5	Chr22	42526686	G	A
CYP2D7-1	Chr22	42537476	T	A
CYP2D7-2	Chr22	42537920	G	G
CYP2D7-3	Chr22	42538687	T	C
CYP2D7-4	Chr22	42538743	C	A
CYP2D7-5	Chr22	42540369	A	G

实验部分：

1. 第一轮 PCR 扩增

PCR 扩增酶采用美国 kapa 公司的 KAPA2G Fast Multiplex PCR Kit 产品（货号 KK5801）。PCR 反应体系如下表 13：

表 13

试剂	体积
2Xkapa聚合酶混合液	25μl
特异性引物池2（10μM）	5μl
DNA	20μl
总共	50μl

特异性引物池2如表14：

表14

引物编号	序列
CYP2D6-CYP2D7-1F	GACCGCTTGGCCTCCGACTTGAGCCCCGGGTGTC CCAGC（SEQ ID NO：33）
CYP2D6-CYP2D7-2F	GACCGCTTGGCCTCCGACTTAGGAAGGCCTCAGT CAGGTCTCGG（SEQ ID NO：34）
CYP2D6-CYP2D7-3F	GACCGCTTGGCCTCCGACTTCTCACGGCTTTGTCC AAGAGA（SEQ ID NO：35）
CYP2D6-CYP2D7-4F	GACCGCTTGGCCTCCGACTTGCCCTTCTGCCCATC ACCC（SEQ ID NO：36）
CYP2D6-CYP2D7-5F	GACCGCTTGGCCTCCGACTTAGGTTGCCCAGCCCG GGCAGTG（SEQ ID NO：37）
CYP2D6-CYP2D7-1R	GACATGGCTACGATCCGACTTGGATGTGCAGCGTG AGCCCA（SEQ ID NO：38）
CYP2D6-CYP2D7-2R	GACATGGCTACGATCCGACTTATCCCAGCGCTGGC TGGCAAGGT（SEQ ID NO：39）
CYP2D6-CYP2D7-3R	GACATGGCTACGATCCGACTTGGGCACAAAGCGG GAACTGGGAA（SEQ ID NO：40）
CYP2D6-CYP2D7-4R	GACATGGCTACGATCCGACTTTCCGTCTCCACCTT GCGCAACTTG（SEQ ID NO：41）
CYP2D6-CYP2D7-5R	GACATGGCTACGATCCGACTTGGGGCTAGAAGCA CTGGTGCCCCT（SEQ ID NO：42）

特异性引物池2由上述引物等摩尔数混合组成。

扩增体系如下表15：

表15

步骤1	98℃，2min
步骤2	98℃，10s
步骤3	62℃，2min
步骤4	72℃，30s
步骤5	重复步骤2-4，15个循环
步骤6	72℃，5min

加入 1 倍体积的 Agencourt AMPure XP 磁珠（美国贝克曼库尔特有限公司）50μl，按照

说明书进行纯化，纯化后用 20μl 蒸馏水溶解 DNA。

2. 第二轮 PCR 扩增

PCR 扩增酶采用美国 kapa 公司的 KAPA2G Fast Multiplex PCR Kit 产品（货号 KK5801）。反应体系如下表 16 所示：

表 16

试剂	体积
2Xkapa 聚合酶混合液	25μl
通用引物1（10μM）	2.5μl
通用引物2（10μM）	2.5μl
DNA	20μl
总共	50μl

通用引物如表5所示。

扩增体系如下表17所示：

表17

步骤1	98℃，2min
步骤2	98℃，10s
步骤3	62℃，2min
步骤4	72℃，30s
步骤5	重复步骤2-4，15个循环
步骤6	72℃，5min

加入 1 倍体积的 Agencourt AMPure XP 磁珠（美国贝克曼库尔特有限公司）50μl，按照说明书进行纯化，纯化后用 20μl 蒸馏水溶解 DNA。

3. 上机测序

文库质检合格后采用华大基因 BGISEQ-500 平台进行测序，测序类型双端 50bp。

数据分析部分：

1. 使用 cutadapt 去除两端含有接头的序列。

2. 通过 bwtsw 算法对人类参考基因组构建比对索引，bwa 版本为 0.7.15。

3. 通过 BWA-ALN 算法将目标序列比对到人类参考基因组 hg19，bwa 版本为 0.7.15，samtools 版本为 0.1.18。

将比对到 hg19 的数据进行提取；同时将生成的*.map.bam 转化为*.map.flag，其目的是为了将第二列的 FLAG 值数值表达形式转变为字母表达形式，可以用于区分 R1 或 R2，从而用

于计算各目标区域的引物富集程度。

4. 对每个样本的原始测序 reads 数、去除测序接头后的 reads 数目、比对率、目标区域数据所占比例等基本信息进行统计（表 18）。同时对每个目标区域的富集程度进行统计，从而评估引物的均一性（图 7），结果显示：5 对引物测序得到的深度差别不大，所有区域的深度都在平均深度的 0.5X 以上，最低深度和最高深度的差值在 3 倍以内。

表 18 CYP2D6 基因检测下机数据基本信息统计

样本编号	原始reads数目	去接头reads数目	比对reads数目	目标区域reads数目
1	17698	17468	17369	17287
2	18672	18453	18296	18211

显示：下机数据的数据利用率达到 98%，比对率达到 99%，目标区域比例 99%。

在计算引物富集程度时，如果当前这条序列有任一 1bp 在目标区域内，就认为此条序列就是这个目标区域的数据。

5. 通过 sam 文件的第六列内容，即 CIGAR 值将比对后的结果进行矫正，将 reads 还原成最原始的状态。

6. 通过比对后的 reads 差异位点的序列区分 CYP2D6 和 CYP2D7，并对应生成只包含 CYP2D6 数据和只包含 CYP2D7 数据的两个文件。

对参考基因组序列进行处理，将 CYP2D6 基因处的碱基替换为 N，记为 HG19 $CYP2D6^{-}$ 参考集；对参考基因组序列进行处理，将 CYP2D7 基因处的碱基替换为 N，记为 HG19 $CYP2D7^{-}$ 参考集。

7. 将步骤 6 中得到的 CYP2D6 文件和 CYP2D7 文件分别比对 HG19 $CYP2D7^{-}$ 参考集和 HG19 $CYP2D6^{-}$ 参考集。

8. 过滤掉插入片段长度大于 500 或者 PE 测序 reads 比对到不同染色体的序列。

9. 通过重新比对后生成的 sam 文件的第六列内容，即 CIGAR 值将比对后的结果进行矫正，将测序 reads 还原成最原始的状态。

10. 去除低质量（测序质量值小于 10）的碱基位点，同时进行标记。

将每个碱基对应的质量值 ASCII 转化为对应的十进制数值，然后减去 33 即可得到对应的质量值，如果这个值小于 10，则将此碱基用*号代替。在 python 中可使用公式：碱基质量值= ord(ASCII) – 33。

11. 对目标位点或区域进行统计。

对目标区域序列进行计数，并比较两者的数量差异（图 8）。结果显示：在一个样本中，区分到 CYP2D6 和 CYP2D7 的测序 reads 数目大致相当。

表 19 示出了两个样本位点结果检测。

表 19 位点结果检测（CYP2D6 基因检测）

Rs 号	样本信息		检测结果	
	样本 1	样本 2	样本 1	样本 2

rs35742686	A	A	A	A
rs3892097	G	A/G	G	A/G
rs5030865	A	A	A	A
rs1065852	C	C	C	C
rs28624811	A	A	A	A

结果表明：先通过基因特异性位点区分不同的测序 reads 来源后，根据区分后的测序 reads 来得到目标位置碱基信息，在这两个样本中，可以正确鉴定目标位点的碱基信息。

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. 一种基于高通量测序检测同源序列的方法，其特征在于，所述方法包括：

获取样本的一对同源序列的特异性扩增产物的高通量测序结果，所述特异性扩增产物包含至少一个用于区分同源序列的序列特异性位点；

将所述高通量测序结果与参考序列进行比对，并根据所述序列特异性位点将所述高通量测序结果区分为两组，每组归属于一个同源序列；

将每组归属于一个同源序列的高通量测序结果分别与去除了另一个同源序列的参考序列进行比对；和

统计比对到去除了另一个同源序列的参考序列的 reads 数目，根据所述 reads 数目确定所述样本的同源序列信息。

2. 根据权利要求 1 所述的方法，其特征在于，所述特异性扩增产物是使用靶向一对同源序列的多个区域的多对引物进行特异性扩增得到的产物。

3. 根据权利要求 1 所述的方法，其特征在于，所述同源序列是同源基因。

4. 根据权利要求 1 所述的方法，其特征在于，所述同源基因是 RHD 和 RHCE 基因，或 CYP2D6 和 CYP2D7 基因。

5. 根据权利要求 3 或 4 所述的方法，其特征在于，所述特异性扩增产物是使用靶向所述同源基因的多个外显子区域的多对引物进行特异性扩增得到的产物。

6. 根据权利要求 1 所述的方法，其特征在于，所述序列特异性位点是单核苷酸变异（SNV）位点。

7. 根据权利要求 1 所述的方法，其特征在于，所述去除了另一个同源序列的参考序列是另一个同源序列全部替换为 N 碱基序列的参考序列。

8. 根据权利要求 1 所述的方法，其特征在于，所述根据所述 reads 数目确定所述样本的同源序列信息，具体为：根据所述 reads 数目确定所述样本的同源序列的突变信息。

9. 根据权利要求 1 所述的方法，其特征在于，所述根据所述 reads 数目确定所述样本的同源序列信息，具体为：根据对比到所述同源序列的两个不同来源的 reads 数的差异，确定所述同源序列在基因组上属于正常、发生缺失或重复的情况。

10. 根据权利要求 1 所述的方法，其特征在于，所述将所述高通量测序结果与参考序列进行比对后，根据 CIGAR 值矫正比对后的结果，以准确地根据所述序列特异性位点将所述高通量测序结果进行区分。

11. 根据权利要求 1 所述的方法，其特征在于，所述将每组归属于一个同源序列的高通量测序结果分别与去除了另一个同源序列的参考序列进行比对后，根据 CIGAR 值矫正比对后的结果。

12. 根据权利要求 1 所述的方法，其特征在于，所述将所述高通量测序结果与参考序列进行比对之前，还包括：去除所述高通量测序结果两端的接头序列。

13. 根据权利要求 1 所述的方法，其特征在于，所述将每组归属于一个同源序列的高通量测序结果分别与去除了另一个同源序列的参考序列进行比对之后，还包括：

过滤掉插入片段长度大于预设值或两端序列比对到不同染色体的结果；和/或

去除测序质量值低于预设值的碱基位点。

14. 一种基于高通量测序检测同源序列的装置，其特征在于，所述装置包括：

获取单元，用于获取样本的一对同源序列的特异性扩增产物的高通量测序结果，所述特异性扩增产物包含至少一个用于区分同源序列的序列特异性位点；

第一比对单元，用于将所述高通量测序结果与参考序列进行比对，并根据所述序列特异性位点将所述高通量测序结果区分为两组，每组归属于一个同源序列；

第二比对单元，用于将每组归属于一个同源序列的高通量测序结果分别与去除了另一个同源序列的参考序列进行比对；和

统计单元，用于统计比对到去除了另一个同源序列的参考序列的 reads 数目，根据所述 reads 数目确定所述样本的同源序列信息。

15. 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 1-13 中任一项所述的方法。

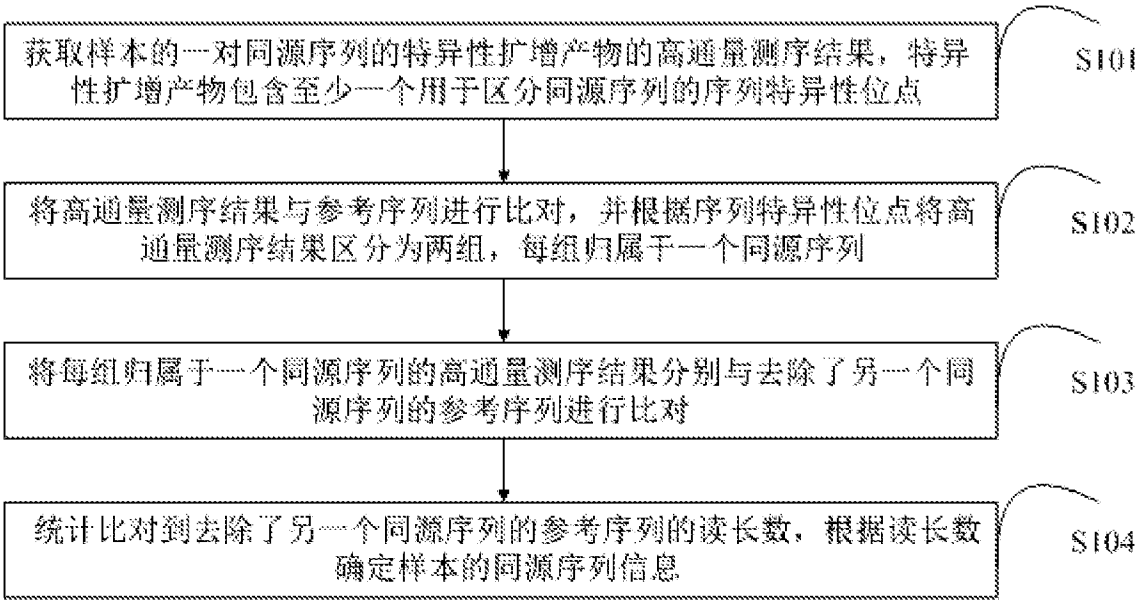
获取样本的一对同源序列的特异性扩增产物的高通量测序结果，特异性扩增产物包含至少一个用于区分同源序列的序列特异性位点
S101
将高通量测序结果与参考序列进行比对，并根据序列特异性位点将高通量测序结果区分为两组，每组归属于一个同源序列
S102
将每组归属于一个同源序列的高通量测序结果分别与去除了另一个同源序列的参考序列进行比对
S103
统计比对到去除了另一个同源序列的参考序列的读长数，根据读长数确定样本的同源序列信息
S104

图 1

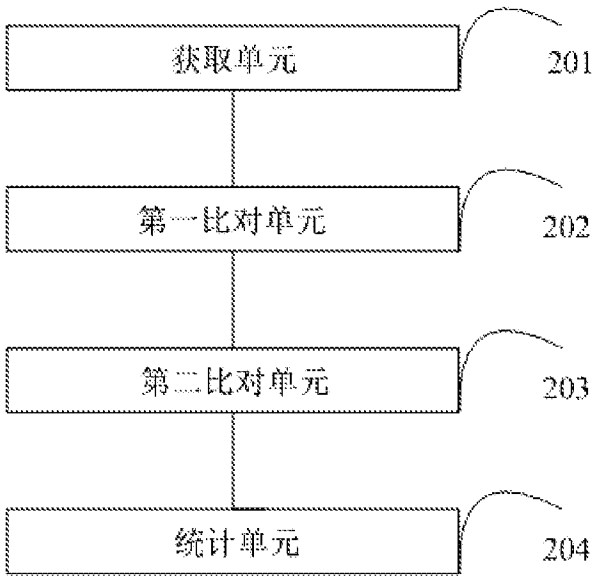
获取单元
201
第一比对单元
202
第二比对单元
203
统计单元
204

图 2

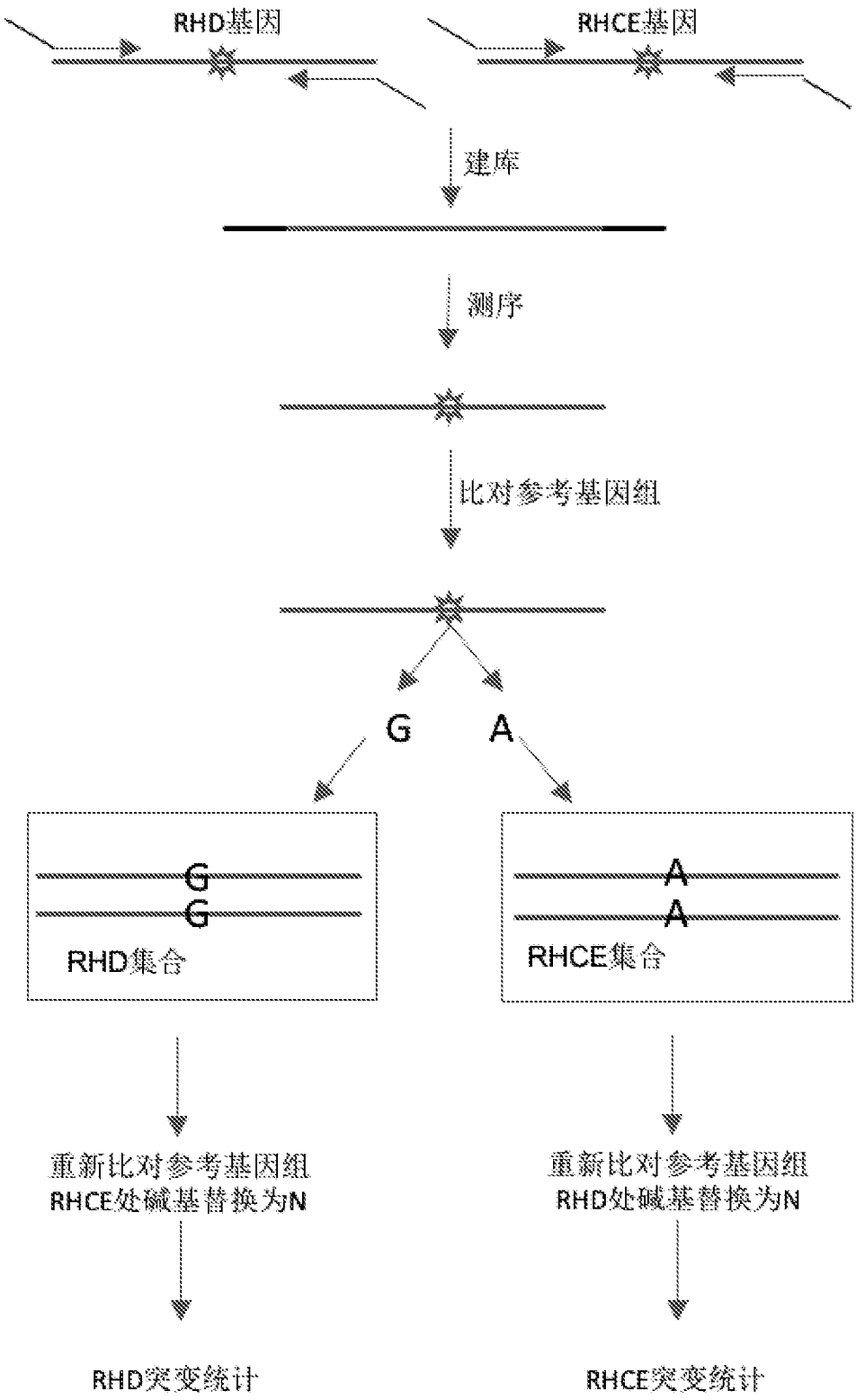
RHD基因
RHCE基因
建库
测序
比对参考基因组
G
A
G
G
RHD集合
A
A
RHCE集合
重新比对参考基因组
RHCE处碱基替换为N
重新比对参考基因组
RHD处碱基替换为N
RHD突变统计
RHCE突变统计

图 3

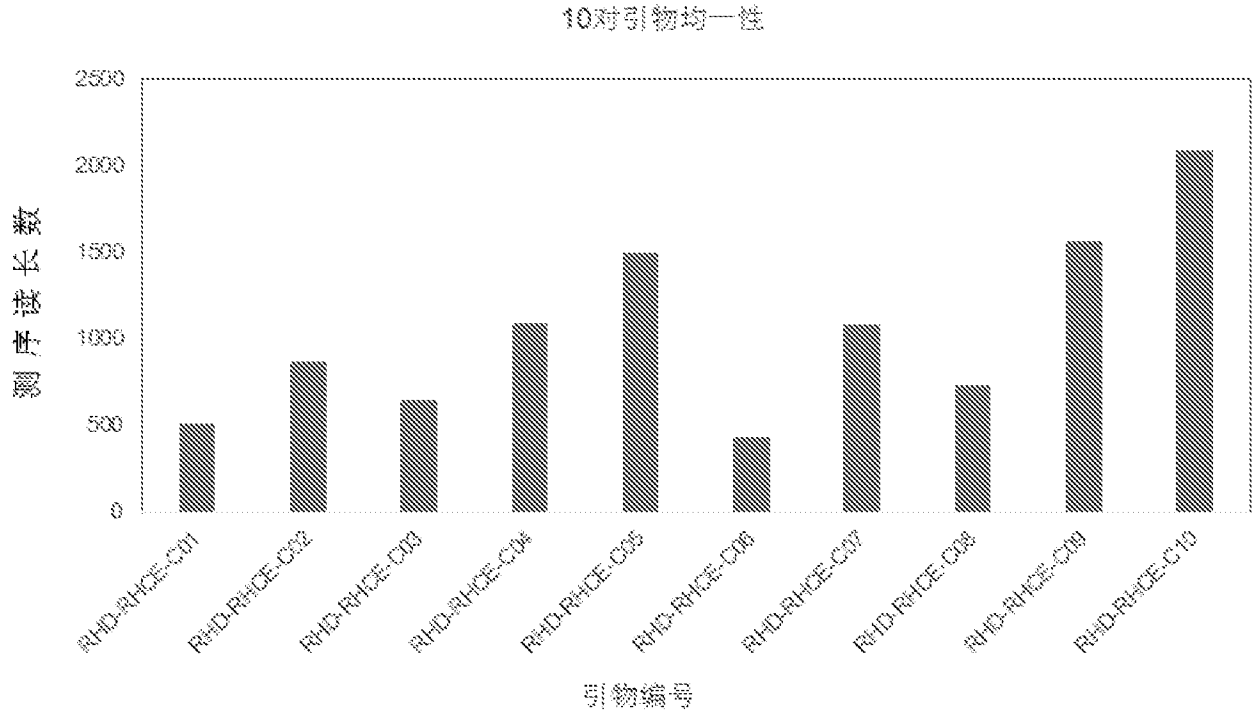
10对引物均一性
测序深度
2500
2000
1500
1000
500
0
RHD-RHCE-C01
RHD-RHCE-C02
RHD-RHCE-C03
RHD-RHCE-C04
RHD-RHCE-C05
RHD-RHCE-C06
RHD-RHCE-C07
RHD-RHCE-C08
RHD-RHCE-C09
RHD-RHCE-C10
引物编号

图 4

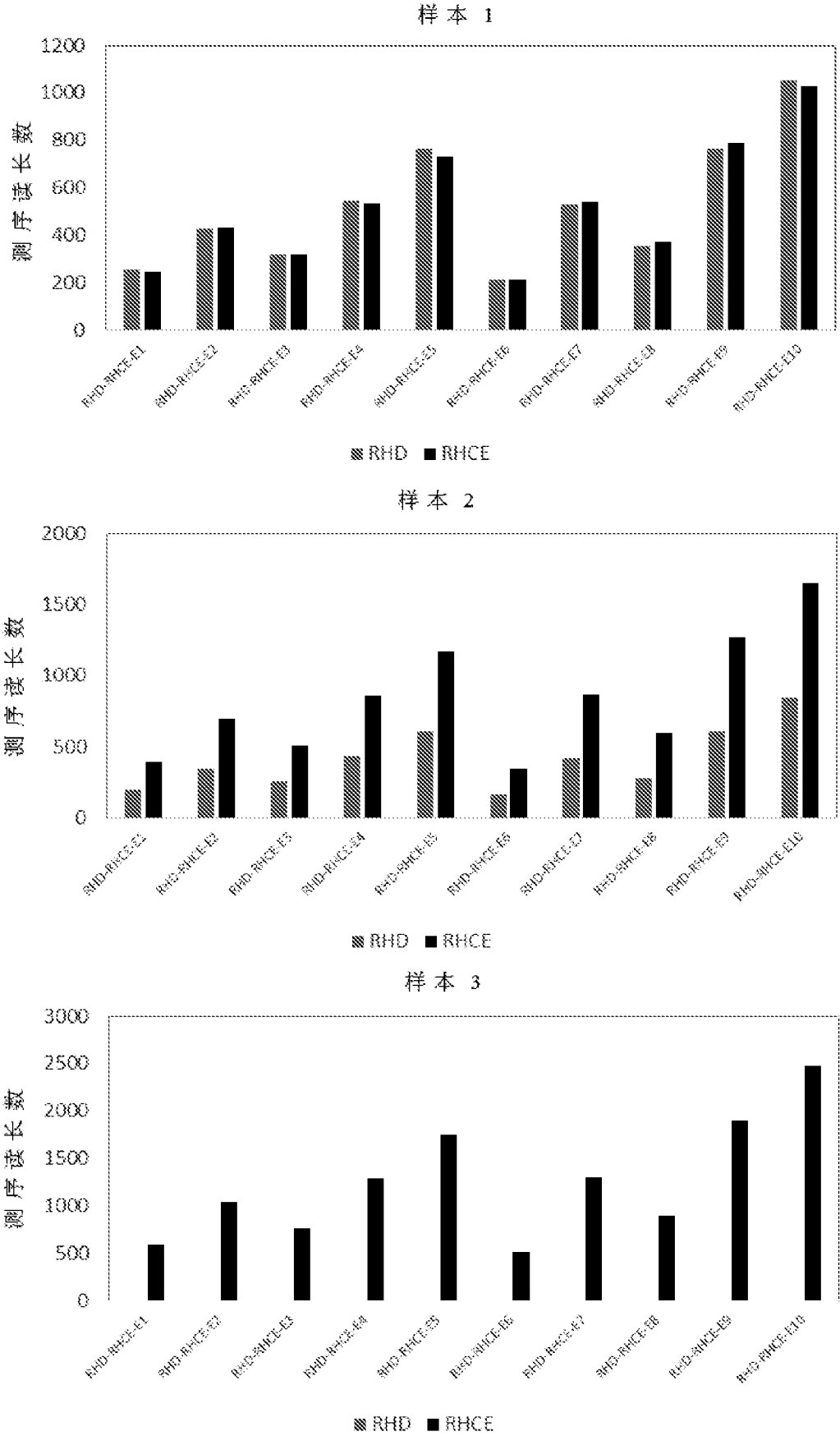
样本 1
测序读长数
1200
1000
800
600
400
200
0
RHD-RHCE-E1
RHD-RHCE-E2
RHD-RHCE-E3
RHD-RHCE-E4
RHD-RHCE-E5
RHD-RHCE-E6
RHD-RHCE-E7
RHD-RHCE-E8
RHD-RHCE-E9
RHD-RHCE-E10
RHD RHCE
样本 2
测序读长数
2000
1500
1000
500
0
RHD-RHCE-E1
RHD-RHCE-E2
RHD-RHCE-E3
RHD-RHCE-E4
RHD-RHCE-E5
RHD-RHCE-E6
RHD-RHCE-E7
RHD-RHCE-E8
RHD-RHCE-E9
RHD-RHCE-E10
RHD RHCE
样本 3
测序读长数
3000
2500
2000
1500
1000
500
0
RHD-RHCE-E1
RHD-RHCE-E2
RHD-RHCE-E3
RHD-RHCE-E4
RHD-RHCE-E5
RHD-RHCE-E6
RHD-RHCE-E7
RHD-RHCE-E8
RHD-RHCE-E9
RHD-RHCE-E10
RHD RHCE

图 5

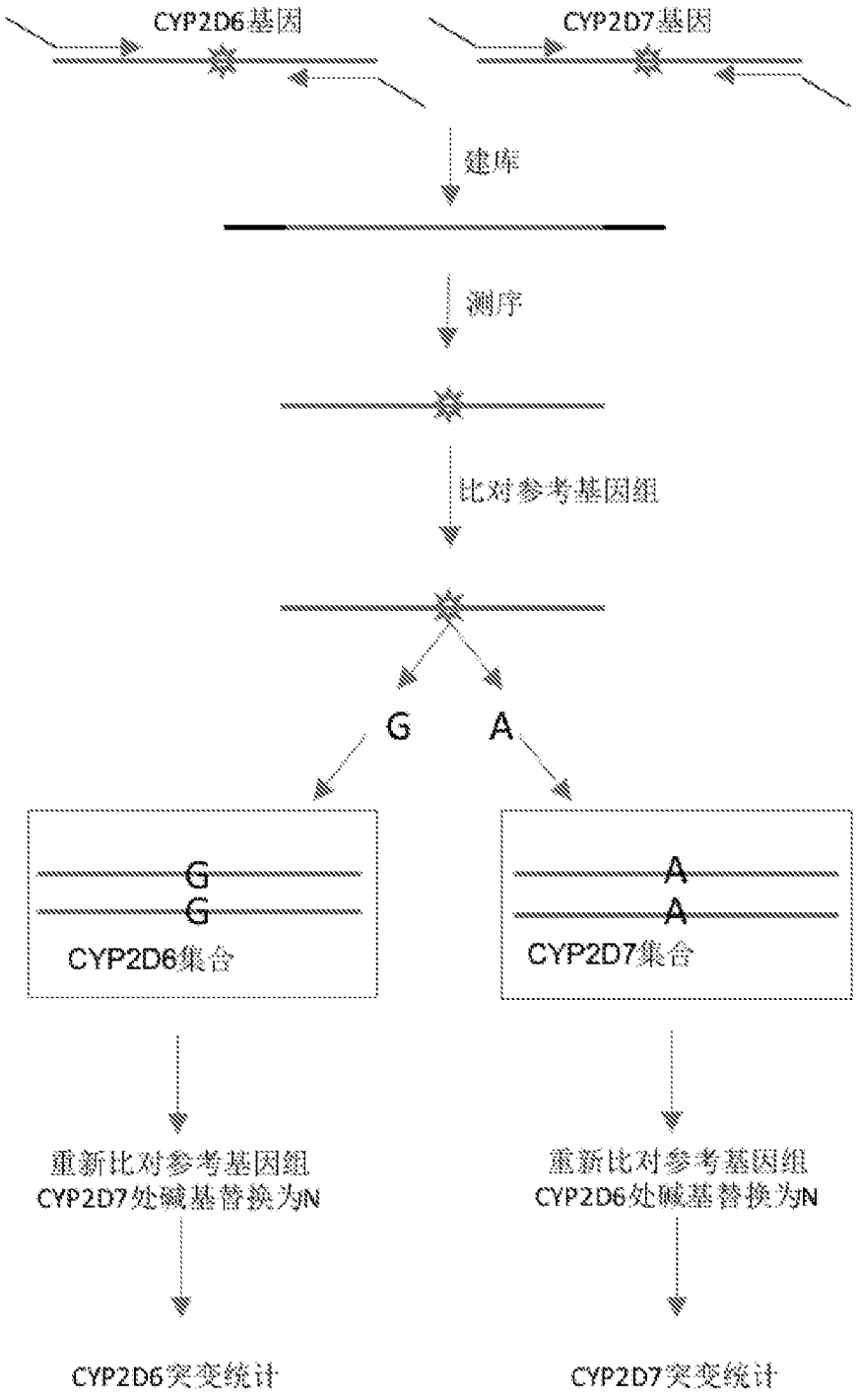
CYP2D6基因
CYP2D7基因
建库
测序
比对参考基因组
G
A
G
G
CYP2D6集合
A
A
CYP2D7集合
重新比对参考基因组
CYP2D7处碱基替换为N
重新比对参考基因组
CYP2D6处碱基替换为N
CYP2D6突变统计
CYP2D7突变统计

图 6

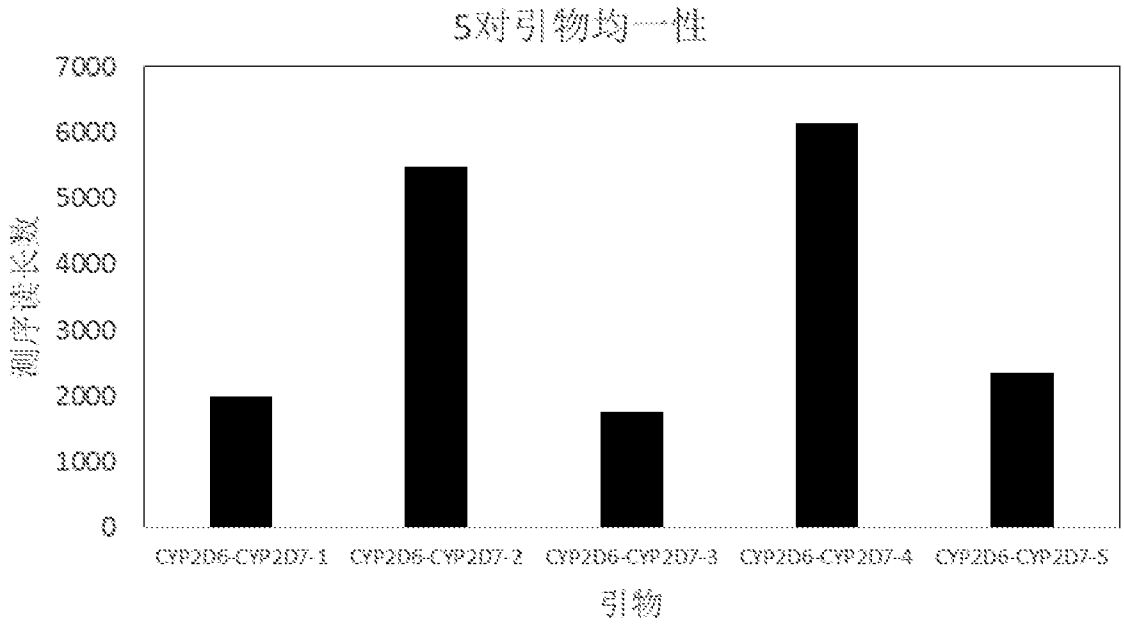
5对引物均一性
7000
6000
5000
4000
3000
2000
1000
0
测序读长数
CYP2D6-CYP2D7-1
CYP2D6-CYP2D7-2
CYP2D6-CYP2D7-3
CYP2D6-CYP2D7-4
CYP2D6-CYP2D7-5
引物

图7

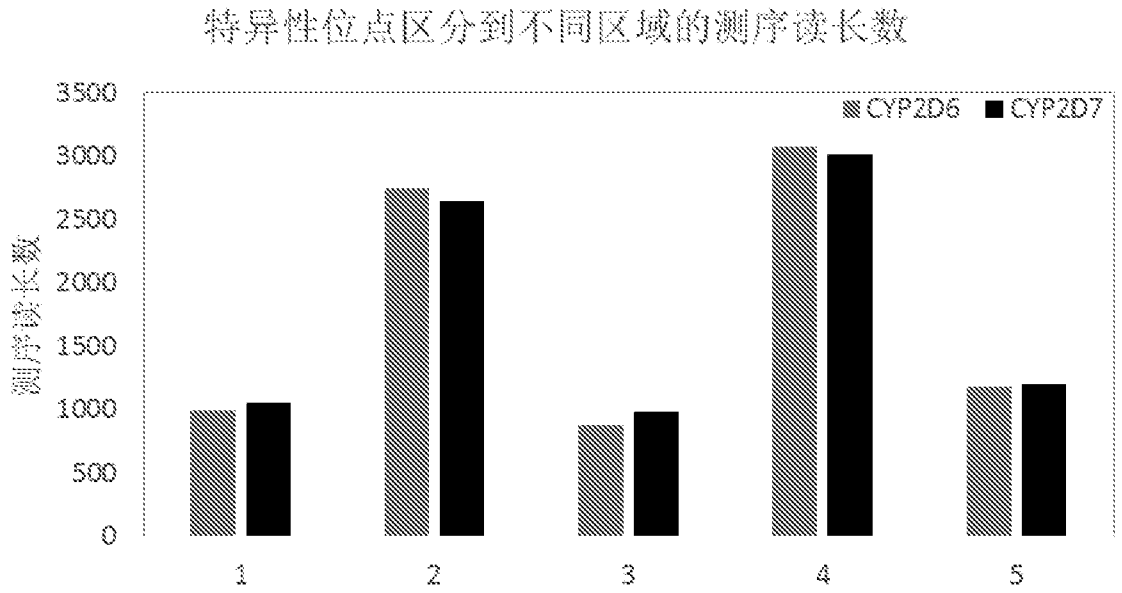
特异性位点区分到不同区域的测序读长数
3500
3000
2500
2000
1500
1000
500
0
测序读长数
CYP2D6
CYP2D7
1
2
3
4
5

图8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/102546

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68(2018.01)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, CPRSABS, SIPOABS, DWPI, CNTXT, WOTXT, EPTXT, USTXT, CNKI, 百度学术搜索, BAIDU XUESHU SEARCH, WEB OF SCIENCE, PubMed: 参考序列, 比对, 位点, 同源序列, 高通量, 扩增, 参考序列, 信息, 特异性, 对照, 同源, reads, homologous, orthologous, CYP2D6, CYP2D7, RHD, RHCE

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2016023962 A1 (PROGENIKA BIOPHARMA S.A.) 18 February 2016 (2016-02-18) see entire document	1-15
A	CN 102965367 A (COTTON RESEARCH INSTITUTE, CHINESE ACADEMY OF AGRICULTURAL SCIENCES) 13 March 2013 (2013-03-13) see entire document	1-15
A	CN 105112569 A (INSTITUTE OF PATHOGEN BIOLOGY, CHINESE ACADEMY OF MEDICAL SCIENCES) 02 December 2015 (2015-12-02) see entire document	1-15

☐ 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
15 May 2019	**31 May 2019**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/102546

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:
   a. ☑ forming part of the international application as filed:
      ☑ in the form of an Annex C/ST.25 text file.
      ☐ on paper or in the form of an image file.
   b. ☐ furnished together with the international application under PCT Rule 13*ter*.1(a) for the purposes of international search only in the form of an Annex C/ST.25 text file.
   c. ☐ furnished subsequent to the international filing date for the purposes of international search only:
      ☐ in the form of an Annex C/ST.25 text file (Rule 13*ter*.1(a)).
      ☐ on paper or in the form of an image file (Rule 13*ter*.1(b) and Administrative Instructions, Section 713).

2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that forming part of the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/102546

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
WO	2016023962	A1	18 February 2016	US	2018305756	A1	25 October 2018
				GB	201414350	D0	24 September 2014
				EP	3180442	A1	21 June 2017
CN	102965367	A	13 March 2013	CN	102965367	B	18 June 2014
CN	105112569	A	02 December 2015	CN	105112569	B	21 November 2017

国际检索报告

国际申请号
PCT/CN2018/102546

A. 主题的分类

C12Q 1/68(2018.01)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,CPRSABS,SIPOABS,DWPI,CNTXT,WOTXT,EPTXT,USTXT,CNKI,百度学术搜索,WEB OF SCIENCE,PubMed:参考序列，比对，位点，同源序列，高通量，扩增，参考序列，信息，特异性，对照，同源，reads，homologous，orthologous，CYP2D6，CYP2D7，RHD，RHCE

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	WO 2016023962 A1 (PROGENIKA BIOPHARMA SA) 2016年 2月 18日 (2016-02-18) 参见全文	1-15
A	CN 102965367 A (中国农业科学院棉花研究所) 2013年 3月 13日 (2013-03-13) 参见全文	1-15
A	CN 105112569 A (中国医学科学院病原生物学研究所) 2015年 12月 2日 (2015-12-02) 参见全文	1-15

☐ 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
2019年 5月 15日	2019年 5月 31日
ISA/CN的名称和邮寄地址 中国国家知识产权局(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	受权官员 赵彦豪 电话号码 62411043

国际检索报告

国际申请号
PCT/CN2018/102546

第I栏　核苷酸和/或氨基酸序列(续第1页第1.c项)

1. 关于国际申请中所公开的任何核苷酸和/或氨基酸序列,国际检索是基于下列序列表进行的：

a. ☑ 作为国际申请的一部分提交的：
   - ☑ 附件C/ST.25文本文件形式
   - ☐ 纸件或图形文件形式

b. ☐ 根据细则13之三.1(a)仅为国际检索目的以附件C/ST.25文本文件形式与国际申请同时提交的：

c. ☐ 仅为国际检索目的在国际申请日之后提交的：
   - ☐ 附件C/ST.25文本文件形式（细则13之三.1(a)）
   - ☐ 纸件或图形文件形式（细则13之三.1(b)和行政规程第713段）

2. ☐ 另外，在提交/提供了多个版本或副本的序列表的情况下，提供了关于随后提交的或附加的副本中的信息与申请时提交的作为申请一部分的序列表的信息相同或未超出申请时提交的申请中的信息范围（如适用）的所需声明。

3. 补充意见：

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/102546

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	2016023962	A1	2016年 2月 18日	US	2018305756	A1	2018年 10月 25日
				GB	201414350	D0	2014年 9月 24日
				EP	3180442	A1	2017年 6月 21日
CN	102965367	A	2013年 3月 13日	CN	102965367	B	2014年 6月 18日
CN	105112569	A	2015年 12月 2日	CN	105112569	B	2017年 11月 21日