

(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) 公布日

2013 年 7 月 11 日 (1.07.2013)



W I P O I P C T



国际公布号
W O 2013/102441 A 1

- (51) 国际分类号 :
G06F 19/22 (2011.01)
C12Q 1/68 (2006.01)
C12Q 1/26 (2006.01)
- (21) 国际申请号 : PCT/CN20 13/070080
- (22) 国际申请日 : 2013 年 1 月 5 日 (05.01.2013)
- (25) 申报语言 : 中文
- (26) 公布语言 : 中文
- (30) 优先权 :
2012 10002976.3 2012 年 1 月 6 日 (06.01.2012) CN
- (71) 申请人: 深圳华大基因科技有限公司 (BGI SHEN - ZHEN CO., LIMITED) [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山工业区综合楼 11F-3, Guangdong 518083 (CN)。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 刘晓 (LIU, Xia o); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。徐怀前 (XU, Huaiqian); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。张伟 (ZHANG, Wei); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。苏政 (SU, Zheng); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。王冠 (WANG, Guan); 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 PROP - ERTY 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, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, 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, 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, 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: CYP450 GENOTYPE DATABASE AND METHOD FOR GENOTYPING AND ASSESSMENT OF ENZYME ACTIVITY

(54) 发明名称 CYP450 基因型别数据库及基因分型、酶活性鉴定方法

(57) Abstract: Provided are a CYP450 standard genotype database, a construction method therefor, and a method for CYP450 genotyping and for assessment of enzyme activity. The method for constructing the CYP450 standard genotype database comprises the following steps: comparing the standard sequence of complete human genome with a specific sequence corresponding to mutation information of a CYP450 genotype, acquiring a correlation on each base position between the CYP450 specific sequence and the standard sequence of complete human genome; converting, on the basis of the correlation acquired, the CYP450 genotype into a genotype having the standard sequence of complete human genome as a reference sequence, and acquiring a standardized genotype of the CYP450 gene.

(57) 摘要: 提供了 CYP450 的标准基因型别数据库及其构建方法, 以及 CYP450 基因分型和酶活性鉴定方法。其中, 构建 CYP450 基因标准型别数据库的方法包括以下步骤: 将 CYP450 基因型别的突变信息对应的特定序列与人类全基因组标准序列进行比对, 获得 CYP450 特定序列与人类全基因组标准序列在每个碱基位置上的对应关系; 根据所获得的对应关系, 将 CYP450 基因型转换成以人类全基因组标准序列为参考序列的基因型, 获得 CYP450 基因的标准化基因型别。



WO 2013/102441 A1

CYP450 基因型别数据库及基因分型、酶活性鉴定方法

优先权信息

本申请请求 2012 年 01 月 06 日向中国国家知识产权局提交的、专利申请号为
5 201210002976.3 的专利申请的优先权和权益，并且通过参照将其全文并入此处。

技术领域

本发明涉及基因检测领域，特别涉及 CYP450 的标准基因型别数据库及其构建方法，以及 CYP450 基因分型和酶活性鉴定方法。

10

背景技术

细胞色素 P450 简称 CYP450，目前已经从人体中鉴定出了 57 种 CYP450 氧化酶，其中参与药物代谢主要集中于 CYP1、CYP2 及 CYP3 家族，参与代谢目前 90% 以上的药物。CYP450 在药物代谢中的重要作用导致 CYP450 基因多态性成为影响药物个体差异的最重要因素之一，其多态性包括点突变、插入或缺失、整个基因的缺失或复制，最终导致酶的活性增强、减弱或完全缺失。CYP450 多态性可导致在标准的药物剂量下可能产生个体严重的副反应或不起作用。比如，华法林（warfarin）作为目前的一线口服抗凝药，其预防和治疗血栓栓塞的作用效果和口服给药的便利性和经济性是显而易见的，但是华法林治疗窗口很窄，且给药各个体差异和种族差异很大，要达到同样的作用效果，高低剂量可相差 10 倍以上。研究表明，这与个体间的 CYP2C9 的基因多态性密切相关。再如，广泛用于治疗乳腺癌和卵巢癌的他莫昔芬（tamoxifen）需要经过一系列的 CYP450 酶类代谢，才最终形成有活性的产物发挥治疗效果，其中 CYP2D6 的基因型是影响他莫昔芬疗效的主要限制因素，其缺陷型可导致无病生存期变短，而 CYP2D6 强代谢型只需要较小剂量的他莫昔芬即可达到同样的效果。

越来越多的研究显示，CYP450 基因除了影响药物代谢外，其基因多态性与许多疾病的发生紧密相关。在先天性肾上腺皮质增生症中，CYP21B 基因缺陷占病因 90%-95% 左右，CYP11B1 基因突变约占 5%，CYP17AK CYP11A1 的缺陷型突变也可导致此疾病的发生。

CYP450 在代谢以及疾病发生中的重要作用，使得 CYP450 基因多态性的检测具有非常重要的社会意义和研究价值。

但是，目前对于 CYP450 突变检测的技术主要集中于单个或几个 P450 基因的已知突变；并且，各研究中的突变位点都是相对各自特定的序列而言的，彼此之间缺乏统一衡量比较的标准。另外，现有的 CYP450 检测方法对 CYP450 基因中仍然存在的未知突变位点检测效率不高；需要对大量样本进行检测时，耗时长，不能很好的满足研究和实践应用的需要。

因而，目前的 CYP450 突变检测及基因分型的方法仍有待改进。

发明内容

本发明旨在至少解决现有技术中存在的技术问题之一。为此，本发明的一个目的是提供一种 CYP450 基因型别的标准数据库及其构建方法，以及基于该数据库的快速检测 CYP450 基因型的方法和 CYP450 酶活性鉴定方法。

因而，根据本发明的一个方面，本发明提供了一种构建 CYP450 基因标准型别数据库的方法。根据本发明的实施例，该方法包括以下步骤：将 CYP450 基因型别的突变信息对应的特定序列与人类全基因组标准序列进行比对，获得 CYP450 特定序列与人类全基因组标准序列在每个碱基位置上的对应关系；根据所获得的对应关系，将 CYP450 基因型转换成以人类全基因组标准序列为参考序列的基因型，获得 CYP450 基因的标准化基因型别。发明人惊奇地发现，利用本发明的构建 CYP450 基因标准型别数据库的方法，能够有效地构建 CYP450 基因标准型别数据库，并且该数据库能够为各 CYP450 基因分型提供一个统一标准，从而，基于该 CYP450 基因标准型别数据库，对于已知基因型的 CYP450 基因，能够快速准确地确定待测样本的 CYP450 基因的基因型别，进一步，能够为针对待测样本的涉及 CYP450 基因的疾病诊断或药物治疗等提供更精确的判断依据。

根据本发明的另一方面，本发明还提供了一种 CYP450 基因标准型别数据库，其是采用前面所述的本发明的构建 CYP450 基因标准型别数据库的方法构建的。发明人发现，该 CYP450 基因标准型别数据库能够为各 CYP450 基因分型提供一个统一标准，从而，基于该 CYP450 基因标准型别数据库，对于已知基因型的 CYP450 基因，能够快速准确地确定待测样本的 CYP450 基因的基因型别，进一步，能够为针对待测样本的涉及 CYP450 基因的疾病诊断或药物治疗等提供更精确的判断依据。

根据本发明的另一方面，本发明还提供了一种 CYP450 基因分型的方法。根据本发明的实施例，该方法包括：获取待测样本 CYP450 基因的外显子序列，采用高通量测序平台测序并进行数据分析，将分析结果与前面所述的本发明的 CYP450 基因型别数据库进行比较，从而得到待测样本的基因型别。利用该方法，对于已知基因型的 CYP450 基因，能够快速准确地确定待测样本的 CYP450 基因的基因型别，进一步，能够为针对待测样本的涉及 CYP450 基因的疾病诊断或药物治疗等提供更精确的判断依据。

根据本发明的再一方面，本发明还提供了一种 CYP450 酶活性鉴定方法。根据本发明的实施例，该方法包括：获取待测样本 CYP450 基因的外显子序列，采用高通量测序平台测序并进行数据分析，将分析结果与前面所述的本发明提供的含有 CYP450 酶活性信息的 CYP450 基因标准型别数据库进行比较，得到待测样本的基因型别，并根据基因型别对应的酶活信息获得待测样本的 CYP450 酶活性结果。利用该方法，对于已知基因型的 CYP450 基因，能够快速准确地确定待测样本的 CYP450 基因的基因型别以及 CYP450 酶活性结果，进一步，能够为涉及 CYP450 的疾病或药物等提供更精确的判断依据。

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

附图说明

5 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为根据本发明一个实施例,本发明的CYP450基因分型及酶活性鉴定方法中,获取待测样本CYP450基因的外显子序列的步骤的具体流程示意图;

10 图2为根据本发明一个实施例,本发明的CYP450基因分型及酶活性鉴定方法中,数据分析的步骤的具体流程示意图。

发明详细描述

下面详细描述本发明的实施例。下面通过参考附图描述的实施例是示例性的,仅用于解释本发明,而不能理解为对本发明的限制。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。

15 根据本发明的一个方面,本发明公开提供了一种构建CYP450基因标准型别数据库的方法。根据本发明的实施例,该方法包括以下步骤:将CYP450基因型别的突变信息对应的特定序列与人类全基因组标准序列进行比对,获得CYP450特定序列与人类全基因组标准序列在每个碱基位置上的对应关系;根据所获得的对应关系,将CYP450基因型转换成以人类全基因组标准序列为参考序列的基因型,获得CYP450基因的标准化基因型别。发明人惊奇地发现,利用本发明的构建CYP450基因标准型别数据库的方法,能够有效地构建CYP450基因标准型别数据库,并且该数据库能够为各CYP450基因分型提供一个统一标准,从而,基于该CYP450基因标准型别数据库,对于已知基因型的CYP450基因,能够快速准确地确定待测样本的CYP450基因的基因型别,进一步,能够针对待测样本的涉及CYP450基因的疾病诊断或药物治疗等提供更精确的判断依据。

25 根据本发明的一个实施例,CYP450基因包括选自CYP11A1、CYP11B1、CYP11B2、CYP17AK、CYP1A1、CYP1A2、CYP1B1、CYP20A1、CYP21A2、CYP24A1、CYP26AK、CYP26B1、CYP26C1、CYP27A1、CYP27B1、CYP27C1、CYP2A13、CYP2A6、CYP2A7、CYP2B6、CYP2C18、CYP2C19、CYP2C8、CYP2C9、CYP2D6、CYP2E1、CYP2FK、CYP2J2、CYP2R1、CYP2S1、CYP2U1、CYP2W1、CYP39A1、CYP3A4、CYP3A43、CYP3A5、CYP3A7、CYP4A1、CYP4A11、CYP4A22、CYP4B1、CYP4F11、CYP4F12、CYP4F2、CYP4F22、CYP4F3、CYP4F8、CYP4V2、CYP4X1、CYP4Z1、CYP51A1、CYP5A1、CYP7A1、CYP7B1、CYP8A1、CYP8BK、POR等58个人类CYP450基因的至少一种。

35 根据本发明的另一个实施例,上述方法进一步包括将CYP450基因定位于人类全基

基因组标准序列上，确定 CYP450 基因编码序列的起始位置和终止位置，获得 CYP450 基因型别突变信息对应的特定序列。

根据本发明的一个实施例，所述 CYP450 基因型别突变信息对应的特定序列包含 CYP450 基因编码序列的起始位置上游 5000bp 至编码序列终止位置下游 500bp 区域的 DNA 片段。

根据本发明的一个优选实施例，所述人类全基因组标准序列为 hgl9。

根据本发明的另一方面，本发明还提供了一种 CYP450 基因标准型别数据库，其是采用前面所述的本发明的构建 CYP450 基因标准型别数据库的方法构建的。发明人发现，该 CYP450 基因标准型别数据库能够为各 CYP450 基因分型提供一个统一标准，从而，基于该 CYP450 基因标准型别数据库，对于已知基因型的 CYP450 基因，能够快速准确地确定待测样本的 CYP450 基因的基因型别，进一步，能够为针对待测样本的涉及 CYP450 基因的疾病诊断或药物治疗等提供更精确的判断依据。

根据本发明的一个优选实施例，在本发明的 CYP450 基因标准型别数据库中，CYP450 基因的各标准化基因型别对应酶活性信息。

根据本发明的另一个实施例，该 CYP450 基因包括下表 1 中所述的 58 个人类 CYP450 基因的至少一种。

表 1 人类 CYP450 基因

CYP11A1	CYP27B1	CYP2S1	CYP4F22
CYP11B1	CYP27C1	CYP2U1	CYP4F3
CYP11B2	CYP2A13	CYP2W1	CYP4F8
CYP17A1	CYP2A6	CYP39A1	CYP4V2
CYP19A1	CYP2A7	CYP3A4	CYP4X1
CYP1A1	CYP2B6	CYP3A43	CYP4Z1
CYP1A2	CYP2C18	CYP3A5	CYP51A1
CYP1B1	CYP2C19	CYP3A7	CYP5A1
CYP20A1	CYP2C8	CYP46A1	CYP7A1
CYP21A2	CYP2C9	CYP4A11	CYP7B1
CYP24A1	CYP2D6	CYP4A22	CYP8A1
CYP26A1	CYP2E1	CYP4B1	CYP8B1
CYP26B1	CYP2F1	CYP4F11	POR
CYP26C1	CYP2J2	CYP4F12	
CYP27A1	CYP2R1	CYP4F2	

根据本发明的另一方面，本发明还提供了一种 CYP450 基因分型的方法。根据本发明的实施例，该方法包括：获取待测样本 CYP450 基因的外显子序列，采用高通量测序平台测序并进行数据分析，将分析结果与前面所述的本发明的 CYP450 基因型别数据库

进行比较,从而得到待测样本的基因型别。利用该方法,对于已知基因型的 CYP450 基因,能够快速准确地确定待测样本的 CYP450 基因的基因型别,进一步,能够为针对待测样本的涉及 CYP450 基因的疾病诊断或药物治疗等提供更精确的判断依据。

根据本发明的再一方面,本发明还提供了一种 CYP450 酶活性鉴定方法。根据本发明的实施例,该方法包括:获取待测样本 CYP450 基因的外显子序列,采用高通量测序平台测序并进行数据分析,将分析结果与前面所述的本发明提供的含有 CYP450 酶活性信息的 CYP450 基因标准型别数据库进行比较,得到待测样本的基因型别,并根据基因型别对应的酶活信息获得待测样本的 CYP450 酶活性结果。利用该方法,对于已知基因型的 CYP450 基因,能够快速准确地确定待测样本的 CYP450 基因的基因型别以及 CYP450 酶活性结果,进一步,能够为涉及 CYP450 的疾病或药物等提供更精确的判断依据。

根据本发明的一个实施例,获取待测样本 CYP450 基因的外显子序列是通过以下步骤实现的:

A、制备能够捕获 CYP450 基因外显子序列的芯片,所述芯片上含有与 CYP450 基因外显子序列反向互补的寡核苷酸探针;

B、用待测样本的基因组 DNA 制备序列捕获文库,包括将待测样本基因组 DNA 打断为 200~500bp 大小的片段,进行末端处理后扩增得到序列捕获文库;

C、将步骤 B 制备得到的序列捕获文库与步骤 A 的芯片杂交,从而获取得到待测样本的 CYP450 基因外显子文库。

根据本发明的一个实施例,在上述步骤 A 中,芯片含有能分别与 58 个人类 CYP450 基因的所有外显子序列反向互补的寡核苷酸探针,寡核苷酸探针的长度为 55-105bp。

根据本发明的另一个实施例,在上述步骤 B 中,将待测样本基因组 DNA 打断为 200~300bp 大小的片段。

根据本发明的再一个实施例,在上述步骤 B 中,末端处理包括进行末端修复形成平末端磷酸化的 DNA 片段,并在该平末端磷酸化的 DNA 片段的 3' 末端加上碱基 "A",并进一步连接标签。

进一步,根据本发明的一个实施例,在上述步骤 C 中,在进行杂交之前,将来自多个不同待测样本的序列捕获文库混合后再同时与步骤 A 的芯片杂交,每个文库带有不同的标签 (Index) 碱基序列而相互区别,且该标签碱基序列长度优选为 6~8bp。

根据本发明的一个实施例,在本发明的 CYP450 基因分型和酶活性鉴定方法中,数据分析进一步包括:

i、过滤去掉影响信息分析的低质量测序序列;

ii、以人类全基因组标准序列为参考序列,将步骤 i 得到的序列用比对软件进行比对,比对软件优选 SOAP 或 BWA;

iii、选取比对到目标区域的序列进行后续分析,所述目标区域是指 CYP450 基因外显子序列所在区域;

iv、数据质控合格后进行变异分析，所述变异分析包括检测以下中的至少一种：单核苷酸多态性、插入和删除、结构性变异、拷贝数变异。

需要说明的是，由于采用了以上技术方案，使本发明至少具备以下有益效果：

1、本发明的构建 CYP450 基因标准型别数据库的方法，能够为各 CYP450 基因分型提供一个统一标准的数据库，基于该数据库，对于已知基因型的 CYP450 基因，能够快速准确地给出待测样本的 CYP450 基因基因型别，从而能够对涉及 CYP450 的疾病或药物等提供更精确的判断依据。进一步，通过在 CYP450 基因标准型别数据库中增加与基因型别对应的酶活性信息，能够在对待测样本进行基因分型的同时，直接得到其 CYP450 酶活结果。

2、本发明的 CYP450 基因分型方法，先获取所有 CYP450 基因的外显子序列后，再进行测序分析，与 CYP450 基因标准型别数据库进行比对后，能够有效地对 CYP450 基因中未知突变位点进行检测。

3、本发明的 CYP450 基因分型及酶活鉴定方法，利用芯片捕获具有高通量的性质，可以实现一次实验同时检测多达上百个样本，不仅提高了检测样本的数量，同时也大大降低了每个样本的检测费用。

4、本发明的 CYP450 基因分型及酶活鉴定方法，包含了目前人类中已经鉴定出的 57 个 CYP450 氧化酶基因及 1 个 P450 还原酶基因，覆盖范围全面广泛，极大地方便了专门针对 CYP450 基因的研究。

进一步，本发明还提供了构建 CYP450 基因型别标准化数据库、CYP450 基因分型及酶活性鉴定的整体技术方案，其是以目标区域经序列捕获后的高通量测序为基础进行的，具体地，可以包括以下步骤：

一、CYP450 基因型别标准化数据库的构建

收集 CYP450 全部功能基因，其中包含 57 个 CYP450 氧化酶基因和 1 个 CYP450 还原酶基因（见上述表 1），通过 BLAST（<http://blast.ncbi.nlm.nih.gov/Blast.cgi>）比对软件，以人类全基因组标准序列 hg19 为参考序列，将 58 个 CYP450 基因的所有基因型别序列与 hg19 参考序列比对，根据比对结果得到相对于 hg19 的突变位点信息，将 CYP450 基因的所有基因型别转换成统一的格式和标准。根据基因在全基因组上的注释信息，将基因型别转换为以 hg19 为标准的型别。具体包括以下步骤：

1. 收集 CYP450 基因型别相关突变及酶活性信息

收集现有的 58 个 CYP450 基因的所有基因型别的突变信息和型别与酶活性相关信息。这些信息主要包括基因型别的名称、基因型别对应的蛋白质编号、基因型别与特定序列的突变信息、基因型别对应的酶在活体中的活性、基因型别对应的酶在体外试验中的活性等。需要说明的是，本申请中所述“特定序列”，是指研究中所采用的作为参考的 DNA 序列片段或者一段 cDNA 序列。对收集的资料分析发现，每个型别的突变信息都是相对于其中一个特定序列给出的；也就是说，不同的研究资料中，58 个 CYP450 基因其基因型别的参考对象不同，而针对不同的参考对象，同一个基因的不同基因型别也存在差异。对于不同资

料上格式的不一致，需要改成统一的格式，以便后续的整理。

2. 收集基因在特定序列上的 CDS 区域，及基因在 hgl9 上的位置

在收集的资料中，很多基因型别突变信息是相对于给定的特定序列的，并且，突变位点信息是以 1998 公布的基因突变命名规则 (Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group) 为标准的，以基因的 CDS (编码序列) 起始位置为 +1 的标准来给出突变位置的。所以为了后续的分析，需要找出所有基因在特定序列上的 CDS 起始位置。又因为特定序列非常的长，有些序列上包括了多个基因，所以要确定哪一段区域是发明人需要的 CYP450 基因。发明人是先找出 CYP450 基因在 hgl9 上的位置，然后从 CDS 起始位置上游的 5000bp 到 CDS 终止位置下游的 500bp 作为 CYP450 基因的区域，但有些基因型别的突变位点离 CDS 区比较远，超出了上述的范围，对于这些基因，发明人会把这个基因的区域定得更长一些，以囊括上述突变位点为原则。

3. BLAST 比对

将特定序列与 hgl9 进行 BLAST 比对。如果特定序列是 cDNA，发明人用 BLAT 进行比对。

4. 确定特定序列与 hgl9 的突变信息

在比对结果中，特定序列可能会比对上 hgl9 的多个位置，选择比对最好的一个位置的比对结果，对每一个位置上的碱基进行分析，得到特定序列与 hgl9 在每一个位置上的碱基对应关系。需要注意的是，如果比对到染色上的负链上，需要将碱基转换成正链上的碱基。

5. 转换所有 CYP450 基因型别

根据特定序列与 hgl9 的比对情况，将所有 CYP450 基因型别转换为以 hgl9 为标准的突变位点信息。在进行坐标转换时，需要用到上面基因的 CDS 起始位置和定义的基因区域。有些基因型别上的突变位点信息都是负链的，在转换时需要将负链信息转换为正链。

6. 整理文件格式及检查

整理文件格式，将基因型别酶活性的信息也加入进来，具体例子如表格 2 所列。之后再检查结果的正确性。

表 2 CYP450 基因的标准化基因分型数据库信息(部分)

基因	基因型别	染色体	SNP 突变位点	INDEL 突变位点	其他突变信息	酶在体内活性	酶在体外实验中活性	参考文献
CYP2C9	CYP2C9*12	chr10	96748777 :C>T	-	-	-	降低	Blaisdell et al, 2004
CYP2C9	CYP2C9*13	chr10	96701715 T>C	-	-	降低	降低	Si et al, 2004; Guo et al, 2005a; Guo et al, 2005b

CYP2C 9	CYP2 C9* 1 4	chr10	9670199 1 G>A	-	-	-	降低	Zhao et al, 2004; Delozier et al, 2005
CYP2C 9	CYP2 C9* 1 5	chr10	96707539 C>A	-	-	-	无活性	Zhao et al, 2004; Delozier et al, 2005
CYP2C 9	CYP2 C9*2 5	chr10	-	9670 1970-9 6701979:Del AGAAATG GAA	-	-	无活性	Maekawa et al, 2006

二、CYP450 基因型别检测

首先，参照图 1 所示的获取待测样本 CYP450 基因的外显子序列的步骤的具体流程示意图，确定待测样本 CYP450 基因的外显子序列，具体地：

5 1、P450 外显子探针设计

发明人根据 57 个 CYP450 氧化酶基因及一个 CYP450 还原酶基因，以人类基因组 hg19 为参考序列，选取这 58 个基因的全部外显子区域作为靶序列，靶序列长度之总和约 276 kb。针对每一个外显子序列，设计与外显子序列反向互补的长度约为 55-105bp 的寡核苷酸捕获探针。将设计的捕获探针高密度的固定合成在芯片上，形成包含 58 个 CYP450 基因所有外显子捕获探针的捕获芯片。设计好的探针由 Roche-Nimblegen 生产并合成固定在捕获芯片上。

上述探针序列是参照 hg19 设计的，由于不同物种间基因组序列存在一定的差异，因此该探针优先适用于人源基因组 DNA 捕获，其它跟人类基因组同源性较高的物种的基因组可以适用，但捕获效果可能不如人源基因组理想。不同物种可以根据其参考序列设计跟本发明类似的探针，应用于不同物种靶区域的捕获。

2、基因组打断、纯化

以没有 RNA、蛋白质污染且没有降解的人基因组 DNA 作为实验材料，利用物理或化学的方法将 DNA 打断成 200~300bp 大小的片段，使用相关回收试剂盒回收 DNA 片段。

3、末端修复、纯化

回收纯化的片段化 DNA 通过 T4 DNA 聚合酶、Klenow 片段和 T4 多聚核苷酸激酶等酶的作用以 dNTP 为作用底物进行末端修复，形成补平的末端磷酸化的 DNA 片段，然后纯化。

4、3'加 "A"、纯化

利用 Klenow 片段 (3',-5',exo-)聚合酶及 dATP，将经过纯化的末端磷酸化的 DNA 片段的 3'末端加上碱基 "A"，然后纯化。

25 5、接头连接、纯化

利用 T4 DNA 连接酶，将经过纯化的末端加 "A" 的 DNA 片段与标签接头连接，并用试剂盒纯化接头连接产物。

6、连接产物 PCR、定量

以标签接头序列引物对加接头后的 DNA 文库进行扩增，扩增产物经纯化后经 Agilent 2100 和 NanoDrop 定量、质控合格后备用。

7、PCR 产物、接头封闭序列、CotI DNA 混合

将上述获得的 PCR 产物、接头封闭序列、CotI DNA 混合以构成外显子捕获文库。

然后，将建好的多个待测样本的文库混合，为了在测序中区别来自不同样本的文库，每个文库的 DNA 在连接标签接头时，其接头中都含有不同的 6bp 或 8bp 的标签碱基序列，每个文库 DNA 混合量可根据需要等量或按照一定比例混合。需要说明的是，等量即在需要每个样本测序数据量相同时，每个文库混合 DNA 量一致；有的研究不同样本测序数据量可能不同，文库使用量也就不同，混合比例按照本领域技术人员具体的研究目的或设计要求来确定。

8、探针与目标区域杂交

按照 Mmblegen 固相芯片杂交标准操作说明，将上述多个样本的混合文库与芯片进行杂交。

9、探针洗脱、纯化

将杂交后的 DNA 进行洗脱、纯化，然后以接头序列为引物进行扩增，以便获得扩增产物。

10、质控

利用 Agilent 2100 和 Q-PCR 将上述获得的扩增产物进行质控，经质控合格后备用。其中各扩增产物构成测序文库。

11、上机测序

使用 HiSeq2000 平台，采用边合成边测序的测序方法对上述获得的测序文库进行测序。

然后，以人类基因组 hg19 (UCSC) 为参考序列，对获得的测序数据进行数据分析。参照图 2，该数据分析步骤可以包括以下步骤：

步骤一 过滤：去掉质量值较低和有测序接头污染的序列

首先去掉影响信息分析的低质量测序序列：序列中每个碱基分别对应一个测序质量值，对于测序结果的一段序列，计算这段序列的平均质量值，若这条序列的平均质量值低于常规的经验阈值，这条序列会被过滤掉；另一方面，测序序列可能会被机器上的接头污染，这部分含有的序列也会被过滤掉。

步骤二 与参考序列 (hg19) 进行序列比对

以 hg19 (UCSC) 为参考序列，将经过步骤一过滤后的序列用比对软件 (如 SOAP, BWA) 进行序列比对。这些比对软件对于一段序列，能够选择一个最佳的比对位置。对于比对位置有多个的重复序列，软件会选择一个位置输出，并添加一个标签。

步骤三 选取比对到目标区域的序列

芯片杂交后会捕获到部分非目标区域的序列，步骤二中以 hg19 全基因组序列作为参考序列，非目标区域的序列就会根据最佳匹配原则比对到相应的位置，而不会比对的区域。选取比对到目标区域的序列用于后续分析，保证选取的序列都是目标区域序列。

步骤四 数据控制

- 5 数据控制（质控）包括多个方面，如比对上序列的百分比，unique reads（序列与参考序列比对时只有一个最佳比对位置）的百分比，duplication（相同的序列）的比例，测序深度，目标区域的覆盖度等。这些质控要符合常规的经验阈值才能进行下一步的分析，如测序深度与预期一致，单碱基深度覆盖图服从泊松分布。

步骤五 变异检测

- 10 数据质控合格后，才能进行变异分析，包括检测 SNP（单核苷酸多态性），INDEL（插入和删除），SV（结构性变异），CNV（拷贝数变异）等。每种变异检测可根据需要使用不同的方式来实现。

步骤六 CYP450 基因分型

- 15 当变异检测分析完之后，整理每个基因中的突变位点信息，与之前整理好的 CYP450 标准数据库中的相应基因型别比较，得到每个样本的基因型别。由于人是二倍体生物，每个基因的型别最多只有两种型别，最后 CYP450 基因的分型结果是一种纯合型别或者杂合型别。一些基因型别有相应酶活性信息，所以通过样本基因分型之后，同时也可以得到样本对酶活性的反应情况。

- 20 此外，还需要说明的是，现阶段，对于 CYP450 突变检测的技术主要集中于单个或几个 CYP450 基因的已知突变，对于未知突变或者大量样本检测存在耗时长、费用高等限制因素。而本发明的上述整体技术方案，明显具有以下几个优点：

- 25 一、本发明构建的 CYP450 基因标准型别数据库，包含了目前人类中已经鉴定出的所有有功能的 57 个 CYP450 氧化酶基因及 1 个 CYP450 还原酶基因，包含范围全面广泛，利用该数据库可以检测出待测样本这些 CYP450 基因所有已知和未知的多态性位点，极大的方便了专门针对 CYP450 基因的研究。

二、能够建立一个以 hg19 为参考序列的 CYP450 基因标准型别数据库，对于已知基因型的 CYP450 基因，能够快速准确的给出相应基因型别信息，这对于涉及 CYP450 的疾病或药物等提供更精确的判断依据。

- 30 三、利用芯片捕获具有高通量的性质，一次实验同时检测多达上百个样本，不仅提高了检测样本的数量，同时也大大降低了每个样本的检测费用。

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

实施例 1

本实施例根据本发明的构建 CYP450 基因型别标准化数据库的方法构建基因型别标准化数据库，具体地：发明人收集了 CYP450 全部功能基因，其中包含 57 个 CYP450 氧化酶基因和 1 个 CYP450 还原酶基因（见上述表 1），通过 BLAST（<http://blast.ncbi.nlm.nih.gov/Blast.cgi>）比对软件，以人类全基因组标准序列 hgl9 为参考序列，将 58 个 CYP450 基因的所有基因型别序列与 hgl9 参考序列比对，根据比对结果得到相对于 hgl9 的突变位点信息，将 CYP450 基因的所有基因型别转换成统一的格式和标准。根据基因在全基因组上的注释信息，将基因型别转换为以 hgl9 为标准的型别。具体包括以下步骤：

1. 收集 CYP450 基因型别相关突变及酶活性信息

收集现有的 58 个 CYP450 基因的所有基因型别的突变信息和型别与酶活性相关信息。这些信息主要包括基因型别的名称、基因型别对应的蛋白质编号、基因型别与特定序列的突变信息、基因型别对应的酶在活体中的活性、基因型别对应的酶在体外试验中的活性。

2. 收集基因在特定序列上的 CDS 区域，及基因在 hgl9 上的位置

找出收集的资料中所有基因在特定序列上的 CDS 起始位置，并确定哪一段区域是本发明需要的 CYP450 基因。具体地，先找出 CYP450 基因在 hgl9 上的位置，然后从 CDS 起始位置上游的 5000bp 到 CDS 终止位置下游的 500bp 作为 CYP450 基因的区域，其中有些基因型别的突变位点离 CDS 区比较远，超出了上述的范围，对于这些基因，将其区域定得更长一些，以囊括上述突变位点为原则。

3. BLAST 比对

将特定序列与 hgl9 进行 BLAST 比对。

4. 确定特定序列与 hgl9 的突变信息

在比对结果中，对于比对上 hgl9 的多个位置的特定序列，选择比对最好的一个位置的比对结果，对每一个位置上的碱基进行分析，得到特定序列与 hgl9 在每一个位置上的碱基对应关系。需要注意的是，如果比对到染色上的负链上，需要将碱基转换成正链上的碱基。

5. 转换所有 CYP450 基因型别

根据特定序列与 hgl9 的比对情况，将所有 CYP450 基因型别转换为以 hgl9 为标准的突变位点信息。在进行坐标转换时，需要用到上面基因的 CDS 起始位置和定义的基因区域。有些基因型别上的突变位点信息都是负链的，在转换时需要将负链信息转换为正链。

6. 整理文件格式及检查

整理文件格式，将基因型别酶活性的信息也加入进来，之后再检查结果的正确性。

由此，获得类似表 2 所示的包含 58 个 CYP450 基因性别及酶活信息的 CYP450 基因标准型别数据库。

实施例 2

本实施例实验流程部分描述为包括炎黄在内的 50 个样本建库杂交一张芯片，本实施例中的样本数用以解释本发明，而不是限制每张芯片可以杂交的样本数。

1、实验材料

本实施例中的试剂见表 3，其它试剂、耗材和仪器设备未在表 3 中注明者，均为可通过市场购买的通用产品。

表 3 本实施例所用试剂

试剂名称	型号	厂家
QIAquick PCR 纯化试剂盒	28106	QIAGEN
10× 多聚核苷酸激酶缓冲液	B904	Enzymatics
dNTP Solution Set	N201L	Enzymatics
T4 DNA 聚合酶	P708L	Enzymatics
Klenow 酶	P706L	Enzymatics
T4 多聚核苷酸激酶	Y904L	Enzymatics
10x blue 缓冲液	B011	Enzymatics
25 mM dATP	28406501V	GE
Klenow (3'-5' exo-)	P701-LC-L	Enzymatics
2x Rapid ligation 缓冲液	B101	Enzymatics
T4 DNA 连接酶(Rapid)	L603-HC-L	Enzymatics
PE/PE Index Adapter oligo mix (40 uM)		Invitrogen
MiniElute PCR 纯化试剂盒	28006	QIAGEN
Platinum® Pfx DNA 聚合酶	C11708-013	Invitrogen
PCR 引物 PE 1.0/PE 标签引物 1.0 (10pM)		Invitrogen
PCR 引物 PE 2.0/PE 标签引物 2.0 (index, 10 pM)		Invitrogen
Multiplexing Sample Preparation Oligonucleotide 试剂盒	PE-400-1001	PE-400-1001
2×SC 杂交缓冲液	5340721001	Nimblegen
SC Hybridization Component A	5340721001	Nimblegen
Phusion Mix	F-531L	NEB
Multiplexing 测序引物和 Phix Control 试剂盒	PE-400-1002	Illumina
Cot-1 DNA	15279-011	Invitrogen

2、序列捕获文库制备

(1) 基因组 DNA 片段化

以 3 μ g 无蛋白质、RNA 污染且没有降解的炎黄基因组 DNA 为材料，使用 Covaris-S2 超声打断仪（Covaris, US）仪器进行打断。打断参数设置如下：

Treatment1（处理 1）	duty/cycle（%）（负载比）	10
	Intensity（强度）	10
	Cycle/burst（循环/脉冲）	1000
	Time(min)（时间（秒））	60
Treatment2（处理 2）	Time(s)（时间（秒））	0
Treatment3（处理 3）	Time(s)（时间（秒））	0
Treatment4（处理 4）	Time(s)（时间（秒））	0
Cycles（循环）		4

打断后的片段经电泳检测合格（主带集中在 200bp-300bp 之间）后，使用 QIAquick PCR 纯化试剂盒回收纯化，样本溶于 75 μ L 洗脱缓冲液中。

5 （2）DNA 片段末端修复

将打断后回收纯化得到的 DNA 片段按下表在 1.5mL 的离心管中配制末端修复反应体系，形成补平的末端磷酸化的 DNA 片段。

样品 DNA	75 μ L
10 $\times$ 多聚核苷酸激酶缓冲液	10 μ L
dNTP Solution Set（每种 10mM）	4 μ L
T4 DNA 聚合酶	5 μ L
Klenow 片段	1 μ L
T4 多聚核苷酸激酶	5 μ L
总体积	100 μ L

将上述 100 μ L 反应混合物轻柔混匀后，在 Thermomixer（Eppendorf）中 20 $^{\circ}$ C 温浴 30 min 后用 QIAquick PCR 纯化试剂盒纯化，DNA 最后于 32 μ L ddH₂O 中充分溶解。

10 （3）3 末端加“A”碱基修饰

在末端补平修复后的 DNA 片段 3 末端加上“A”碱基，以便于下一步标签接头（Index Adapter）连接。末端加“A”碱基反应体系如下表。

DNA	32 μ L
10x blue 缓冲液	5 μ L
dATP(1mM)	10 μ L
Klenow (3'-5' exo-)	3 μ L
总体积	50 μ L

将上述 50 μ L 反应混合物轻微混匀后，在 Thermomixer（Eppendorf）中 37 $^{\circ}$ C 温浴 30min 后用 QIAquick PCR 纯化试剂盒纯化，DNA 最后于 15 μ L ddH₂O 中充分溶解。

(4) Index Adapter 接头连接

末端加 "A" 后的 DNA 片段纯化后在 T4 DNA Ligase 作用下与标签接头连接。在 1.5 ml 的离心管中配制标签接头连接反应体系：

末端需要加 "A" 的 DNA 片段	15 μ L
2x Rapid 连接缓冲液	25 μ L
PE/PE 标签接头寡核苷酸混合物(40 μ M)	5 μ L
T4 DNA 连接酶(Rapid)	5 μ L
总体积	50 μ L

上述 50 μ L 反应混合物轻微振荡混合均匀，瞬时离心后置于 Thermomixer（Eppendorf）中 20 $^{\circ}$ C 温浴 15min，反应完后用 MiniElute PCR 纯化试剂盒进行纯化，最后将样品溶于 25 μ L 洗脱缓冲液。

10 (5) 杂交前 PCR 扩增

以上述步骤 (4) 得到的 DNA 为模板，以含有接头序列的引物进行扩增，扩增体系和条件如下：

含接头序列的 DNA	25 μ L
10 $\times$ pfx 扩增缓冲液	10 μ L
MgSO ₄ (50mM)	4 μ L
dNTP mix (10mM)	4 μ L
Platinum® Pfx DNA 聚合酶	2 μ L
PCR 引物 PE 1.0/PE 标签引物 1.0 (10pM)	10 μ L
PCR 引物 PE 2.0/PE 标签引物 2.0 (index, 10pM)	10 μ L
ddH ₂ O	35 μ L
总体积	100 μ L

PCR 程序为 94 $^{\circ}$ C 2min; 4 个循环的 94 $^{\circ}$ C 15s, 62 $^{\circ}$ C 30s, 72 $^{\circ}$ C 30s; 72 $^{\circ}$ C 5min。PCR 产物用 QIAquick PCR 纯化试剂盒纯化，洗脱体积为 30 μ L。

15 (7) 样本文库混合

按照上述 DNA 打断、末端修复、加标签接头 (Index Adapter)、杂交前 PCR 等步骤，

构建其它 49 个样本文库,包括炎黄基 因组 DNA 样本文库共计 50 个文库(包含 4 个 HapMap 样本、1 个炎黄样本和 45 个正常人样本,其中 45 正常人样本用于测试一张芯片可以杂交的样本数目),从这 50 个文库中取等量的 DNA 均匀混合。为了在测序中区别来自不同样本的文库,在加 Index Adapter 时,每个文库的 DNA 末端都含有不同的 6bp 或 8bp 的 Index 碱基序列。需要说明的是,Index Adapter 包括两部分,分别为用于区分各文库的标签碱基序列和接头序列。

4、外显子文库构建

外显子文库的构建包括采用制备的序列捕获文库与捕获芯片杂交,将 58 个 CYP450 基因的全部外显子富集到捕获芯片上,洗脱杂交后的捕获芯片,洗脱产物即外显子序列,对外显子序列扩增处理得到外显子文库,具体如下:

(1) 芯片杂交

A) 在 1.5mL 离心管中加入 45 μ g 的 Cot-1 DNA、3 μ g 来自混合文库的 DNA、1mmol Index-adapter1-block 和 Index-adapter2-block (Multiplexing Sample Preparation Oligonucleotide Kit, Illumina), 混合物置于 SpeedVac (Thermo) 中蒸干,温度设置为 60 $^{\circ}$ C。

B) 在蒸干的离心管中加入 11.2 μ L 纯水,充分溶解 DNA 后加入 18.5 μ L 的 2 $\times$ SC 杂交缓冲液和 7.3 μ L 的 SC Hybridization, 充分混匀后将混合物转移至杂交仪 (Nimblegen) 上 95 $^{\circ}$ C 水浴器中 10 分钟使 DNA 变性。

C) 将样品取出震荡后置于离心机上全速离心 30 秒,置于杂交仪 (Nimblegen) 上 42 $^{\circ}$ C 位置,与外显子捕获芯片杂交。

D) 杂交方法参照 NimbleGen 公司芯片杂交方法 (NimbleGen Arrays User's Guide, Version 3.1, 7 Jul 2009, Roche NimbleGen, Inc., 通过参照将其全文并入本文)。样品上样量 35 μ L, 42 $^{\circ}$ C 杂交 64-72hr, 杂交完成并经过芯片的杂交后处理后,用 900 μ L 160mM NaOH 洗脱富集在芯片上的序列,洗脱产物用 MinElute PCR 纯化试剂盒纯化,最终用 80 μ L 洗脱缓冲液洗脱。

(2) 捕获后 PCR 扩增

以从捕获芯片上洗脱下来的序列为模板进行 PCR 扩增,体系为 Phusion Mix 150 μ L, 上下游引物各 4.2 μ L (Multiplexing 测序引物和 Phix Control 试剂盒),上述的 80 μ L 洗脱样品加 85 μ L ddH₂O, 混合后分 6 管进行 PCR。PCR 反应条件 94 $^{\circ}$ C, 1min; 16 个循环的 94 $^{\circ}$ C 30s, 58 $^{\circ}$ C 30s, 72 $^{\circ}$ C 30s; 72 $^{\circ}$ C 5min。PCR 反应后把 6 管混合并用 QIAquick PCR 纯化试剂盒磁珠纯化回收 300-450bp 大小的片段,洗脱体积为 50 μ L。

(3) 文库检测:

采用 Bioanalyzer analysis system (Agilent, Santa Clara, USA) 检测文库插入片段大小及含量;Q-PCR 精确定量文库的浓度。

5、序列测定

对上述经过纯化和质量检测合格的 PCR 扩增产物进行测序,测序方法参照 Illumina 公司 HiSeq2000 操作方法 (HiSeq 2000 User Guide. Catalog # SY-940-1 001 Part # 150 11190 Rev B, Illumina),。

6、数据分析

(1) 测序数据过滤

对测序获得的数据进行两方面的过滤，一是测序质量值，对整条序列，计算其碱基质量值，当整条序列的平均质量值低于 10 时，将其过滤掉；二是检测 Adapter 接头污染，如果序列中含有 Adapter 序列，也将其过滤掉。

测序数据过滤结果显示，被过滤掉的序列约占 7%，其余 93% 用于下一步的分析。

(2) 序列比对

以 hg19 为参考序列，用 BWA (Burrows-Wheeler Aligner) 比对软件对经过数据过滤的序列进行比对。比对时每条序列最多允许 5 个错配，开 gap (比对时允许有插入和删除) 的比对，当一条序列有多个最佳比对位置时，随机选择一个位置输出，但会有标记。在本实施例的测试中，样本比对上的序列占有进行比对的序列的约 97%。

(3) 选取比对到目标区域的序列

比对完之后，首先，根据比对的结果，去掉非 unique reads，只保留那些唯一比对到全基因组中的序列；再去 duplication，对于比对到参考序列上同一位置的配对 reads，去重复任意保留其中一对 reads，因为比对到同一位置的配对序列很可能是 PCR 过程引起的。

上面处理完后，根据 CYP450 芯片设计的目标区域，保留那些比对到参考序列上的目标区域的序列，进行下一步的分析。

(4) 数据质控

数据质控包括样本的数据量，过滤的数据量大小，序列比对时比对上序列的比例，样本的平均深度是否符合预期，单碱基深度覆盖图是否符合泊松分布，样本的目标区域覆盖度等。

统计分析结果显示，本实施例的 50 个样本均符合质控要求，部分结果见表 4。

具体地，数据质控包括两方面，一方面是看各样本之间是不是比较一致，如果各样本之间的数据都差不多，表示符合要求，如果有个别样本的数据其他大多数样本相差很多，说明这个样本很可能有问题；另一方面是每个样本的各质控数据，这些标准本领域技术人员都可根据经验来确定一个大概的范围，不同的测序区域可能会有些变化，具体来说，“数据过滤后剩余量”一般在 90% 以上，比对序列的比例 (%) 90% 以上，去重复后剩余数据量 60% 以上，unique reads 占的比例与具体的测序目标区域相关且 90% 以上，平均深度符合预期的实验设计要求，覆盖度要 95% 以上，都是可以接受的。

表 4 部分样本的数据质控结果

样本名称	原始数据量	数据过滤后剩余量 (%)	比对序列的比例 (%)	去重复后剩余数据量 (%)	unique reads 占的比例 (%)	目标区域数据量	平均深度 (%)	覆盖度 (%)
样本 1	83.47M	93.18	96.85	68.47	93.12	19.26M	70.44	99.04

样本 2	81.73M	92.66	96.77	68.04	93.01	18.25M	66.61	98.95
样本 3	72.84M	92.91	96.77	73.52	93.60	17.80M	65.06	98.75
样本 4	71.07M	93.11	96.74	71.23	93.28	16.73M	61.16	99.05
样本 5	72.71M	92.84	96.71	73.67	93.24	17.07M	62.33	98.87
样本 6	80.67M	92.69	96.64	69.05	93.23	17.51M	63.92	98.61
样本 7	63.65M	93.10	96.71	74.63	93.66	15.63M	57.27	98.77
样本 8	77.48M	92.97	96.79	70.90	93.22	18.45M	67.46	98.89
样本 9	66.75M	93.49	96.79	74.93	93.98	16.71M	61.01	98.76
样本 10	73.58M	92.95	96.79	70.25	93.32	16.97M	62.17	98.86

(5) SNP 分析

本实施例 中，SNP 是用 samtools 得到的，当选取 比对到 目标 区域的序列后，用 samtools 转换格式、排序之后，用其中的 mpileup 命令进行 SNP Callings 原始的 SNP 还会进行一些过滤，包括位点的深度、质量值等。通常，深度在 4-400 符合要求，质量值则是通过用统计的方法计算质量值的显著性，对显著性过滤。

在本实施例的样本中，包括 4 个 HapMap 样本（a、b、c、d）和 1 个炎黄样本（这 5 个样本已经有公布的基因组及分型数据），其中炎黄样本测了两次，对这五个样本的 SNP 进行了评价。4 个 HapMap 样本与已有的 HapMap 数据进行比较，炎黄样本的 SNP 与已有的炎黄样本 Genotyping 位点进行了比较，结果见下表 5 和表 6。

表 5 HapMap 样本的 SNP 分析结果

样本名称	样本的 SNP 数目	HapMap 在 目标区域 的位点数 目	SNP 真 阳性个 数	假阳性 (假阳性 率)	假阴性(假 阴性率)
样本 a	290	511	149	2(0.57%)	13(8.02%)
样本 b	253	512	117	3(0.78%)	11(8.59%)
样本 c	266	510	119	5(1.32%)	13(9.85%)
样本 d	285	510	140	5(1.40%)	14(9.09%)

表 6 炎黄样本的 SNP 分析结果

样本名称	炎黄 Genotyping 位 点数目	SNP 真阳 性个数	假阳性(假 阳性率)	假阴性 (假阴性 率)
单独测 炎黄样本	397	89	0(0.00%)	0(0.00%)
Pooling 中炎黄样 本	397	89	0(0.00%)	0(0.00%)

(6) CYP450 基因分型

做完变异检测后，根据每个基因在全基因组上的区域，提取出每个基因的突变位点信息。根据这些突变位点信息与实施例 1 中构建好的 CYP450 基因标准型别数据库进行比较，确定样本的基因型别信息和酶活性信息。部分样本的检测结果如表 7 所示。

5

表 7 部分样本的 CYP450 基因型和酶活信息

样本	基因	基因型别	酶在体内活性	酶在体外实验中活性	参考文献
样本 1	CYP2C19	CYP2C19*1A	正常	正常	Romkes et al, 1991
		CYP2C19*27	—	降低	Drgemller et al, 2010
样本 2	CYP2C9	CYP2C9*5	降低	降低	Dickmann et al, 2001;
		CYP2C9*8	降低	增加	Blaisdell et al, 2004;

分型结果显示，采用本发明的方法得到的基因型别信息及酶活性信息与现有参考文献记载一致。

工业实用性

10

本发明的构建 CYP450 基因型别标准化数据库、CYP450 基因分型及酶活性鉴定的方法，能够有效地用于 CYP450 基因分型和酶活性鉴定，并且省时省工、成本低、结果准确。

15

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

20

在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示意性实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

权利要求书

1、一种构建 CYP450 基因标准型别数据库的方法，其特征在于，包括以下步骤：

将 CYP450 基因型别突变信息对应的特定序列与人类全基因组标准序列进行比对，获得

5 CYP450 特定序列与所述人类全基因组标准序列在每个碱基位置上的对应关系；

根据所述的对应关系，将所述 CYP450 基因型别转换成以所述人类全基因组标准序列为参考序列的基因型，获得所述 CYP450 基因的标准化基因型别。

2、根据权利要求 1 所述的方法，其特征在于，所述 CYP450 基因包括选自 CYP11A1、CYP11BK、CYP11B2、CYP17A1、CYP1A1、CYP1A2、CYP1B1、CYP20A1、CYP21A2、
10 CYP24A1、CYP26A1、CYP26BK、CYP26C1、CYP27AK、CYP27B1、CYP27CK、CYP2A13、CYP2A6、CYP2A7、CYP2B6、CYP2C18、CYP2C19、CYP2C8、CYP2C9、CYP2D6、CYP2E1、CYP2F1、CYP2J2、CYP2R1、CYP2S1、CYP2U1、CYP2WK、CYP3A1、CYP3A4、CYP3A43、CYP3A5、CYP3A7、CYP46A1、CYP4A11、CYP4A22、CYP4B1、CYP4F11、CYP4F12、CYP4F2、CYP4F22、CYP4F3、CYP4F8、CYP4V2、CYP4X1、CYP4Z1、CYP51A1、CYP5A1、
15 CYP7A1、CYP7B1、CYP8A1、CYP8B1、POR 等 58 个人类 CYP450 基因的至少一种。

3、根据权利要求 1 所述的方法，其特征在于，进一步包括：

定位所述 CYP450 基因于所述人类全基因组标准序列上，确定所述 CYP450 基因编码序列的起始位置和终止位置，获得所述 CYP450 基因型别突变信息对应的特定序列。

4、根据权利要求 1 所述的方法，其特征在于，所述 CYP450 基因型别突变信息对应的
20 特定序列包含所述 CYP450 基因编码序列的起始位置上游 5000bp 至终止位置下游的 500bp 区域。

5、根据权利要求 1 所述的方法，其特征在于，所述人类全基因组标准序列为 hg19。

6、一种 CYP450 基因标准型别数据库，其特征在于，所述数据库采用权利要求 1-5 任一项所述的方法构建。

25 7、根据权利要求 6 所述的 CYP450 基因标准型别数据库，其特征在于，所述 CYP450 基因标准型别数据库中，CYP450 基因的各标准化基因型别对应酶活性信息。

8、根据权利要求 6 所述的 CYP450 基因标准型别数据库，其特征在于，所述 CYP450 基因包括选自 CYP11A1、CYP11BK、CYP11B2、CYP17AK、CYP1A1、CYP1A2、CYP1B1、CYP20A1、CYP21A2、CYP24AK、CYP26A1、CYP26BK、CYP26C1、CYP27A1、CYP27B1、
30 CYP27C1、CYP2A13、CYP2A6、CYP2A7、CYP2B6、CYP2C18、CYP2C19、CYP2C8、CYP2C9、CYP2D6、CYP2E1、CYP2FK、CYP2J2、CYP2R1、CYP2S1、CYP2U1、CYP2W1、CYP39AK、CYP3A4、CYP3A43、CYP3A5、CYP3A7、CYP46A1、CYP4A11、CYP4A22、CYP4B1、CYP4F11、CYP4F12、CYP4F2、CYP4F22、CYP4F3、CYP4F8、CYP4V2、CYP4X1、CYP4Z1、CYP51A1、CYP5A1、CYP7AK、CYP7B1、CYP8A1、CYP8BK、POR 基因的至少一种。
35

9、一种 CYP450 基因分型的方法，其特征在于，包括：

获取待测样本 CYP450 基因的外显子序列，采用高通量测序平台测序并进行数据分析，将分析结果与权利要求 6-8 任一项所述的 CYP450 基因标准型别数据库进行比较，从而得到所述待测样本的基因型别。

10、一种 CYP450 酶活性鉴定方法，其特征在于，包括：

5 获取待测样本 CYP450 基因的外显子序列，采用高通量测序平台测序并进行数据分析，将分析结果与权利要求 7 所述的 CYP450 基因标准型别数据库进行比较，得到待测样本的基因型别，并根据所述待测样本的基因型别对应的酶活信息获得所述待测样本的 CYP450 酶活性结果。

10 11、根据权利要求 9 或 10 所述的方法，其特征在于，所述获取待测样本 CYP450 基因的外显子序列是通过以下步骤实现的：

A、制备能够捕获 CYP450 基因外显子序列的芯片，所述芯片上含有与所述 CYP450 基因外显子序列反向互补的寡核苷酸探针；

B、用待测样本的基因组 DNA 制备序列捕获文库，包括将所述待测样本基因组 DNA 打断为 200~500bp 大小的片段，进行末端处理后扩增得到所述序列捕获文库；

15 C、将步骤 B 制备得到的序列捕获文库与步骤 A 的芯片杂交，从而获得所述待测样本的 CYP450 基因外显子文库。

12、根据权利要求 11 所述的方法，其特征在于，在所述步骤 A 中，所述芯片含有能分别与 58 个人类 CYP450 基因的所有外显子序列反向互补的寡核苷酸探针，所述寡核苷酸探针的长度为 55-105bp。

20 13、根据权利要求 11 所述的方法，其特征在于，在所述步骤 B 中，将所述待测样本基因组 DNA 打断为 200~300bp 大小的片段。

14、根据权利要求 11 所述的方法，其特征在于，在所述步骤 B 中，所述末端处理包括进行末端修复形成平末端磷酸化的 DNA 片段，并在所述平末端磷酸化的 DNA 片段的 3' 末端加上碱基 "A"，并进一步连接标签。

25 15、根据权利要求 11 所述的方法，其特征在于，在所述步骤 C 中，在进行所述杂交之前，将来自多个不同待测样本的序列捕获文库混合后再同时与步骤 A 的芯片杂交，每个文库带有不同的标签碱基序列而相互区别。

16、根据权利要求 11 所述的方法，其特征在于，所述标签碱基序列长度为 6~8bp。

17、根据权利要求 9 或 10 所述的方法，其特征在于，所述数据分析进一步包括：

30 i、过滤去掉影响信息分析的低质量测序序列；

ii、以人类全基因组标准序列为参考序列，将步骤 i 得到的序列用比对软件进行比对；

iii、选取比对到目标区域的序列进行后续分析，所述目标区域是指 CYP450 基因外显子序列所在区域；

35 iv、数据质控合格后进行变异分析，所述变异分析包括检测以下中的至少一种：单核苷酸多态性、插入和删除、结构性变异、拷贝数变异。

18、根据权利要求 17 所述的方法，其特征在于，所述比对软件为选自 SOAP 和 BWA

的至少一种。

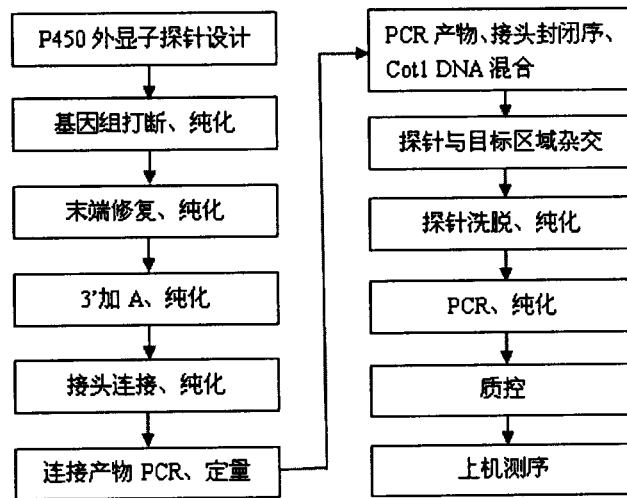


图 1

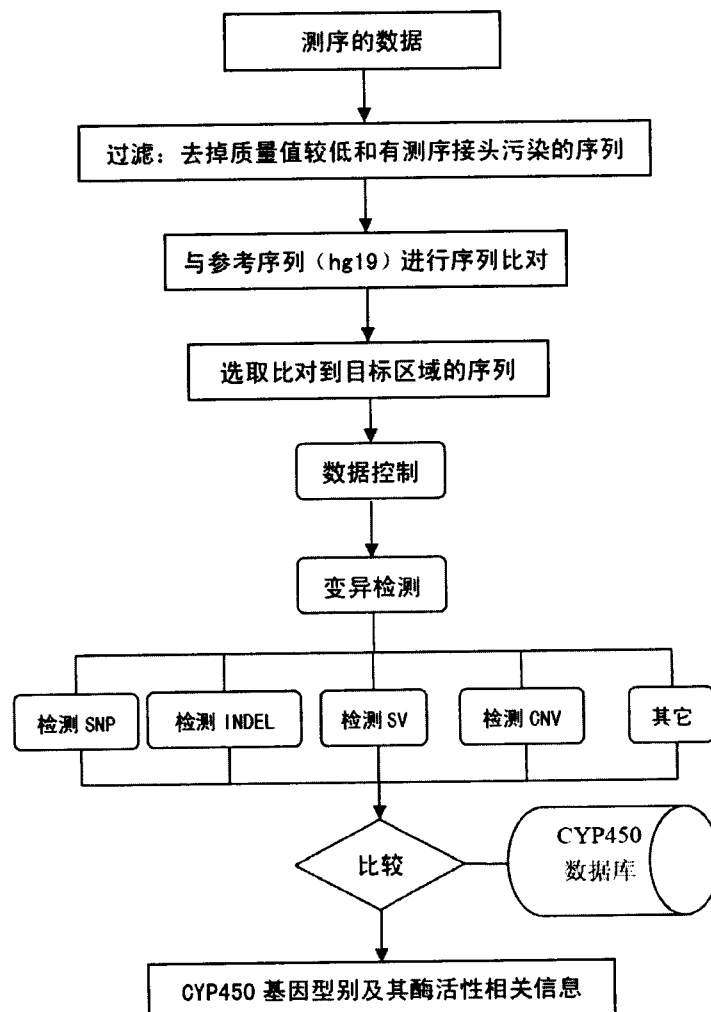


图 2

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2013/070080

A. CLASSIFICATION OF SUBJECT MATTER

See the extra sheet

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)

IPC: G06F, 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, CPEA, CNABS, CPRSABS, CNMED, CNKI, ISI Web of Knowledge, Google Scholar: CYP450, database, standard, align, compare, sequencing, chip, array, probe, capture, exon library.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Sim, S.C. et al. The human cytochrome P450 (CYP) allele nomenclature website: a	1-5
Y	peer-reviewed database of CYP variants and their associated effects, Human Genomics, April 2010, vol. 4, no. 4, pages 278-281	9-18
Y	JIANG Tao et al. High-performance single-chip exon capture allows accurate whole exome sequencing using the Illumina Genome Analyzer. SCIENTIA SINICA Vitae, 2011, vol. 41, no. 9, pages 714-721	9-18

☒ 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 01 April 2013 (01.04.2013)	Date of mailing of the international search report 25 April 2013 (25.04.2013)
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 TIAN, Tian Telephone No. (86-10) 62412276

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2013/070080

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Marsh, S. et al. SNP databases and pharmacogenetics: great start, but a long way to go., Human Mutation, 2002, vol. 20, pages 174-179	1-5, 9-18

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2013/070080

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.: 6-8
because they relate to subject matter not required to be searched by this Authority, namely:
Claims 6-8 relate 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

International application No.

PCT/CN2013/070080

A. CLASSIFICATION OF SUBJECT MATTER

G06F 19/22 (2011.01) i

C12Q 1/68 (2006.01) i

C12Q 1/26 (2006.01) i

A. 主题的分类

见附加页

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)

IPC: G06F, 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, CPEA, CNABS, CPRSABS, CNMED, CNKI, ISI Web of Knowledge, Google Scholar: 细胞色素 P450、数据库、标准、基因组、比对、测序、芯片、探针、捕获、外显子文库、CYP450、database, standard, align, compare、sequencings chip、array、probe、capture、exon library。

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
X	Sarah C. Sim 等, The human cytochrome P450 (CYP) allele nomenclature	1-5
Y	website: a peer-reviewed database of CYP variants and their associated effects, Human Genomics, 4 月 2010, 第 4 卷, 第 4 期, 第 278-281 页。	9-18
Y	蒋涛等, 一种结合单张芯片序列捕获和高通量测序技术测序外显子组的方法, 2011, 第 41 卷, 第 9 期, 第 714-721 页。	9-18
A	Sharon Marsh 等, SNP databases and Pharmacogenetics :Great start, but a long way to go, Human Mutation, 2002, 第 20 卷, 第 174-179 页。	1-5, 9-18

□ 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 3 个月公布在先申请或专利

"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

01. 4 月 2013 (01.04.2013)

国际检索报告邮寄日期

25.4 月 2013 (25.04.2013)

ISA/CN 的名称和邮寄地址:

中华人民共和国国家知识产权局

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

田甜

电话号码: (86-10) 62412276

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. 因 权利要求: 6-8

因为它们涉及不要求本单位进行检索的主题, 即:

权利要求 6-8 涉及单纯的信息表达 (PCT 细则第 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。

第 III 栏 缺乏发明单一性的意见 (续第 1 页第 3 项)

本国际检索单位在该国际申请中发现多项发明, 即:

1. □ 由于申请人按时缴纳了被要求缴纳的全部附加检索费, 本国际检索报告涉及全部可作检索的权利要求。

2. □ 由于无需付出有理由要求附加费的劳动即能对全部可检索的权利要求进行检索, 本单位未通知缴纳任何附加费。

3. □ 由于申请人仅按时缴纳了部分被要求缴纳的附加检索费, 本国际检索报告仅涉及已缴费的那些权利要求。
具体地说, 是权利要求:

4. □ 申请人未按时缴纳被要求缴纳的附加检索费。因此, 本国际检索报告仅涉及权利要求书中首先提及的发明; 包含该发明的权利要求是:

关于异议的说明: □ 申请人缴纳了附加检索费, 同时提交了异议书, 适用时, 缴纳了异议费。

□ 申请人缴纳了附加检索费, 同时提交了异议书, 但未在通知书规定的时间期限内缴纳异议费。

□ 缴纳附加检索费时未提交异议书。

A. 主题的分类

G06F 19/22 (201 1.01) i

C12Q1/68 (2006.01) i

C12Q1/26 (2006.01) i