

(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
2014 年 2 月 13 日 (13.02.2014)



(10) 国际公布号
WO 2014/023076 A1

- (51) 国际专利分类号:
C12Q 1/68 (2006.01) *C40B 50/00* (2006.01)
G06F 19/22 (2011.01) *C40B 40/06* (2006.01)
- (21) 国际申请号: PCT/CN2012/087645
- (22) 国际申请日: 2012 年 12 月 27 日 (27.12.2012)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201210284823.2 2012 年 8 月 10 日 (10.08.2012) CN
- (71) 申请人: 深圳华大基因科技有限公司 (BGI SHENZHEN CO., LIMITED) [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山工业园区综合楼 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园区综合楼, Guangdong 518083 (CN)。
- (72) 发明人: 曹红志 (CAO, Hongzhi); 中国广东省深圳市盐田区北山工业园区综合楼, Guangdong 518083 (CN)。 张涛 (ZHANG, Tao); 中国广东省深圳市盐田区北山工业园区综合楼, Guangdong 518083 (CN)。 柴相花 (CHAI, Xianghua); 中国广东省深圳市盐田区北山工业园区综合楼, Guangdong 518083 (CN)。 陈仕平 (CHEN, Shiping); 中国广东省深圳市盐田区北山工业园区综合楼, Guangdong 518083 (CN)。 李剑 (LI, Jian); 中国广东省深圳市盐田区北山工业园区综合楼, Guangdong 518083 (CN)。 王俊 (WANG, Jun); 中国广东省深圳市盐田区北山工业园区综合楼, Guangdong 518083 (CN)。 汪建 (WANG, Jian); 中国广东省深圳市盐田区北山工业

园区综合楼, Guangdong 518083 (CN)。 杨焕明 (YANG, Huanming); 中国广东省深圳市盐田区北山路 146 号北山工业园区综合楼 11F-3, Guangdong 518083 (CN)。

- (74) 代理人: 上海一平知识产权代理有限公司 (XU & PARTNERS, LLC.); 中国上海市普陀区真北路 958 号天地科技广场 1 号楼 106 室, Shanghai 200333 (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: THALASSEMIA TYPING METHOD AND USE THEREOF

(54) 发明名称: 一种地中海贫血的分型方法及其应用

(57) Abstract: Provided are a thalassemia typing method and uses thereof. Specifically provided is a method for detecting the copy number variation of the HBB gene and/or HBA gene of a to-be-detected sample, the method comprising: acquiring a target gene sequence, the sequence corresponding to the sample and/or the gene; correcting and/or counting a read sequence to obtain the absolute data value of a relevant gene; calculating value A, or value R1, value R2 and value R3, and acquiring the copy number variation information of the HBB gene and/or HBA gene of the to-be-detected sample according to the value A, or value R1, value R2 and value R3. Also provided are a typing system and analysis method for detecting thalassemia.

(57) 摘要: 本发明提供了一种地中海贫血的分型方法及其应用。具体地, 本发明提供了一种检测待测样本 HBB 基因和/或 HBA 基因拷贝数变异的方法, 该方法包括: 获得目的基因序列, 该序列与样品和/或基因相对应; 对所述读序进行修正和/或统计, 获得相关基因的绝对数据值; 计算 A 值或 R1 值、R2 值、R3 值; 根据 A 值或 R1 值、R2 值、R3 值获得待测样本 HBB 和/或 HBA 基因拷贝数变异信息。本发明还提供了检测地中海贫血的分型系统和分析方法。



WO 2014/023076 A1

一种地中海贫血的分型方法及其应用

技术领域

本发明属于基因组生物信息分析领域，具体地，本发明涉及一种地中海贫血的分型方法及其应用。

背景技术

地中海贫血是一种常见的常染色体隐性遗传病。它是由于珠蛋白基因变异，使珠蛋白生物合成受阻、产量不足或缺失所致。地中海贫血在我国长江以南各省发病率极高，是我国影响最大的遗传病之一。

组成珠蛋白的肽链有 4 种，即 α 、 β 、 γ 和 δ 链，分别由其相应的基因编码。这些基因的缺失或点突变可造成肽链的合成障碍，致使血红蛋白的组分改变。最常见有两种类型：由于 α 链变异导致的 α 地中海贫血和由于 β 链变异导致的 β 地中海贫血。

α 珠蛋白基因位于第 16 号染色体短臂末端的 α 珠蛋白基因簇中，该基因簇包括二个重复的 α 基因($\alpha 2$ 和 $\alpha 1$)、一个胚胎期类 α 基因($\zeta 2$)、三个假基因($\Psi \zeta 1$ 、 $\Psi \alpha 2$ 和 $\Psi \alpha 1$)和一个功能未知基因 ($\theta 1$)，其排列顺序为：5'-2 ζ 22 $\Psi \zeta$ 12 $\Psi \alpha$ 22 $\Psi \alpha$ 12 α 22 α 12 θ 12-3'，全长约 30kb。 β 珠蛋白基因位于 11 号染色体短臂(11p15)，基因排列顺序为 5'-2 ϵ 2G γ 2A γ 2 $\Psi \beta$ 2 δ 2 β 2-3'，长度约 60kb。

地中海贫血的分型即检测 α 链和 β 链相关基因的变异，包括点突变，插入缺失和拷贝数变异。地中海贫血检测方法有很多种，包括反向点杂交法，PCR2 单链构象多态性分析方法，基因芯片法等。

反向点杂交法：反向点杂交是检测点突变的一种新技术，它与是将膜上固定探针取代了固定靶 DNA 方式，从而改变了传统杂交法一次只能检测一种突变的方式。用固化了多种特异性探针的膜条与扩增靶序列杂交，使一次杂交即可同时筛查出被检测出来。但是反向点杂交法只能检测部分点突变，不具有普遍性。

PCR2 单链构象多态性分析(single strand conformation polymerase, SSCP)其原理是 DNA 单链上即使只有一个碱基发生改变，就可使该 DNA 单链在中性聚丙烯酰胺凝胶上的电泳迁移率发生改变。其过程是：先用 PCR 方法扩增靶 DNA 片段，然后将 PCR 产物经 95℃变性后在中性聚丙烯酰胺凝胶上进行

电泳，DNA 电泳带可用放射性核素及银染色等方法显示。由于银染法简便易行，且无放射性核素污染，目前已成为首选方法用于显示 SSCP 电泳带。但是这种方法分型慢，费用高，不适宜推广应用。

基因芯片法：1997 年 Affymetrix 公司开始将 DNA 芯片技术用于诊断地中海贫血，其基本原理是用 2 种不同荧光染料标记的靶序列同时与同 1 个 cDNA 芯片杂交，通过不同颜色的荧光信号强度分析即可反映出基因表达的变化。由于用该技术可以将极其大量的探针同时固定于支持物上，所以一次可以对大量的生物分子进行检测分析，从而解决了传统 PCR 和核酸印迹杂交等检测项目单一的缺陷，但是该方法存在低通量、费用昂贵不足等缺点。

目前本领域还没有一种准确简便的一次性检测 SNP、InDel、CNV 的方法，因此迫切需要开发适合于实际应用的地中海贫血的分型方法。

发明内容

本发明的目的是提供一种地中海贫血型别关系数据库及其建立方法。

本发明的另一目的是提供一种检测 SNP 和 Indel 变异的技术及其应用。

本发明的另一目的是提供一种拷贝数变异的检测方法及其应用。

本发明的另一目的是提供一种地中海贫血分型技术及其应用。

在本发明的第一方面，提供了一种检测待测样本 HBB 基因拷贝数变异的方法，包括步骤：

(1) 获得待测样本的 HBB 基因序列，所述序列与样品和/或基因相对应；

(2) 对步骤(1)获得的读序进行修正和/或统计，获得下述绝对数据值：

待测样本 HBB 基因的读序数、待测样本对照基因的读序数、正常样本 HBB 基因的读序数、正常样本对照基因的读序数；

(3) 计算 A 值，

$$A \text{ 值} = \frac{\text{待测样本 HBB 基因的读序数} / \text{待测样本对照基因的读序数}}{\text{正常样本 HBB 基因的读序数} / \text{正常样本对照基因的读序数}};$$

正常样本 HBB 基因的读序数/正常样本对照基因的读序数

(4) 根据 A 值，获得待测样本 HBB 基因拷贝数变异的检测结果。

在另一优选例中，用高通量测序技术对待测样本的 HBB 基因测序，获得基因序列。

在另一优选例中，(1)中获得的待测样本的 HBB 基因序列长度为 70-100bp，

较佳地为80-90bp，更佳地为87bp。

在另一优选例中，利用标签（index）以及引物的特异性将待测样本的HBB基因序列与样品和/或基因相对应。

在另一优选例中，在(2)中，根据SNP和/或InDel对骤(1)获得的读序进行修正。

在本发明的第二方面，提供了一种检测待测样本HBA基因拷贝数变异的方法，包括步骤：

(1) 获得待测样本的HBA1基因序列和HBA2基因序列，所述的序列与样品和/或基因相对应；

(2) 对步骤(1)获得的读序进行修正和/或统计，获得下述绝对数据值：

正常样本HBA1基因的读序数，正常样本HBA2基因的读序数，正常样本对照基因的读序数，待测样本HBA1基因的读序数，待测样本HBA2基因的读序数，待测样本对照基因的读序数；

(3) 计算下述的比值：

$$\begin{aligned} R1值 &= \frac{\text{待测样本HBA1基因的读序数/待测样本对照基因的读序数}}{\text{正常样本HBA1基因的读序数/正常样本对照基因的读序数}} \\ R2值 &= \frac{\text{待测样本HBA2基因的读序数/待测样本对照基因的读序数}}{\text{正常样本HBA2基因的读序数/正常样本对照基因的读序数}} \\ R3值 &= \frac{\text{待测样本HBA2基因的读序数/待测样本HBA1基因的读序数}}{\text{正常样本HBA2基因的读序数/正常样本HBA1基因的读序数}} ; \end{aligned}$$

(4) 根据R1值、R2值和R3值，获得待测样本HBA基因拷贝数变异检测结果。

在另一优选例中，(1)中获得的待测样本的HBA1基因或HBA2基因序列长度为100-150bp，较佳地为120-140bp，更佳地为128bp。

在另一优选例中，用高通量测序技术对待测样本的HBA1基因或HBA2基因测序，获得基因序列。

在另一优选例中，利用标签（index）以及引物的特异性将待测样本的HBA1基因或HBA2基因与样品和基因相对应。

在另一优选例中，在(2)中，根据SNP和/或InDel对骤(1)获得的读序进行修正。
在另一优选例中，所述的(4)包括步骤：

- (i) 对一个文库下所有样品的R1值、R2值、R3值进行大概归类；
- (ii) 计算每个类别中的方差和偏离度（偏离度=均值—类别理论值）；
- (iii) 综合考虑每个类别个数大于 15 个值的方差与偏离度，构建整个文库的方差和偏离度；
- (iv) 计算每个值与每类别的理论值之间的距离；
- (v) 选取每个标签（index）的 R1 值、R2 值、R3 值的最短距离作为最佳选项。

在本发明的第三方面，提供了一种检测待测样本 HBB 基因拷贝数变异的检测单元，所述单元包括模块：

- (1) 序列获取模块，用于获得待测样本的HBB基因序列，所述序列与样品和/或基因相对应；
- (2) 读序修正和/或统计模块，用于对模块(1)获得的读序进行修正和/或统计；
- (3) 比值计算模块，用于计算 A 值；和
- (4) 输出模块，用于输出待测样本 HBB 基因拷贝数变异的信息。

在本发明的第四方面，提供了一种检测待测样本HBA基因拷贝数变异的检测单元，所述单元包括模块：

- (1) 序列获取模块，用于获得待测样本的HBA1基因序列和HBA2基因序列，所述的序列与样品和/或基因相对应；
- (2) 读序修正和/或统计模块，用于对模块(1)获得的序列读序进行修正和/或统计；
- (3) 比值计算模块，用于计算 R1 值、R2 值、R3 值；
- (4) 型别判断模块，用于判断 HBA 基因拷贝数变异的型别；和
- (5) 输出模块，用于输出待测样本 HBA 基因拷贝数变异信息。

在本发明的第五方面，提供了一种检测地中海贫血的分型系统，包括：

- (1) 第三方面所述的检测待测样本HBB基因拷贝数变异的检测单元；和
- (2) 第四方面所述的检测待测样本HBA基因拷贝数变异的检测单元。

在另一优选例中，所述的系统还包括：(3) SNP和/或InDel的检测单元。

在本发明的第六方面，提供了一种检测待测样本SNP和Indel变异的方法，包括步骤：

- (1) 获得待测样本的基因测序数据；

(2) 将(1)中的测序数据与参考序列进行比对, 获得相对于参考序列的SNP和InDel连锁关系数据;

(3) 根据(2)中所述的连锁关系数据确定待测样本SNP和Indel变异信息。

在另一优选例中, 用选自下组的比对软件对(1)中的测序数据与参考序列进行比对: BWA、SOAP、BLAST、MAQ等。

在另一优选例中, 所述的待测样本为人的样本。

在另一优选例中, 所述的参考序列来自hg19。

在本发明的第七方面, 提供了一种对待测样本进行地中海分型的方法, 包括步骤:

(1) 获得待测样本SNP和InDel变异数据, 以及HBB基因、HBA基因拷贝数变异数据;

(2) 结合步骤(1)的所有变异数据, 获得待测样本的完整变异情况, 从而对待测样本进行地中海分型。

在另一优选例中, 用第六方面所述的方法获得待测样本SNP和InDel变异数据。

在另一优选例中, 用第一方面所述的方法对待测样本HBB基因拷贝数变异进行检测。

在另一优选例中, 用第二方面所述的方法对待测样本HBA基因拷贝数变异进行检测。

在本发明的第八方面, 提供了一种构建地中海贫血型别SNP和InDel关系数据库的方法, 包括步骤:

(1) 将现有地中海贫血数据库中已知型别的HBA1基因、HBA2基因和HBB基因与作为参考序列的已知正常无病基因序列进行比对, 获得差异位点, 所述的差异位点为SNP和InDel位点;

(2) 获得每个型别相对于参考序列的SNP和InDel关系, 得到地中海贫血型别SNP-InDel数据库。

在本发明的第九方面, 提供了一种SNP和Indel数据库, 是用第八方面所述的方法构建的。

应理解, 在本发明范围内中, 本发明的上述各技术特征和在下文(如实施例)中具体描述的各技术特征可以互相组合, 从而构成新的或优选的技术方案。限于篇幅, 在此不再一一累述。

附图说明

下列附图用于说明本发明的具体实施方案，而不适用于限定由权利要求书所界定的本发明范围。

图 1 显示了地中海贫血 SNP 和 InDel 信息分析流程。

图 2 显示了 HBB 基因拷贝数检测信息流程。

图 3 显示了 HBA 基因拷贝数信息分析流程。

图 4 显示了引物（primer）的设计思路。

图 5 显示了单个位点深度分布图。

具体实施方式

本发明人经过广泛而深入的研究，提供了一种检测待测样本 HBB 基因和/或 HBA 基因拷贝数变异的方法，所述方法包括步骤：获得目的基因序列，所述序列与样品和/或基因相对应；对所述读序进行修正和/或统计，获得相关基因的绝对数据值；计算 A 值或计算 R1 值、R2 值、R3 值；根据 A 值或 R1 值、R2 值、R3 值获得待测样本 HBB 和/或 HBA 基因拷贝数变异信息。本发明还提供了检测地中海贫血的分型系统和分析方法。在此基础上完成了本发明。

术语

SNP

如本文所用，术语“SNP”或“单核苷酸多态性”可以互换使用，是指在基因组水平上由单个核苷酸的变异所引起的DNA序列的多态性。SNP是可遗传变异中最常见的一种，占有已知多态性的绝大多数。

SNP 所表现的多态性只涉及到单个碱基的变异，这种变异可由单个碱基的转换(transition)或颠换(transversion)所引起，也可由碱基的插入或缺失所致。

InDel

如本文所用，术语“InDel”和“插入缺失突变”可以互换使用，是指涉及核苷酸插入和/或缺失的突变。

CNV

如本文所用，术语“拷贝数变异”和“CNV”可以互换使用，都是指 DNA 片

段大小范围从 kb 到 Mb 的亚微观突变，作为一种染色体内的变异，CNV 意味着某个细胞内的某些 DNA 片段过多或者不足。

双末端测序

对基因片段(包括 DNA 和 cDNA)进行测序，其测序对象都是一段物理连续的碱基序列片段，该片段称为插入片段，其长度称为插入片段长度(insertsize)。

如本文所用，术语“双末端测序”是对该片段的两侧碱基序列从边缘向内部的测序，测得的序列称为读序(read)，长度称为读长(read-length)。两侧测得的读序是来自于同一个插入片段，并且其末端距离为 insertsize，故两侧读序的配对关系确定。这两个读序被称为配对读序(Pair-end reads)。

高通量测序

基因组的高通量测序使得人类能够尽早地发现与疾病相关基因的异常变化，有助于对个体疾病的诊断和治疗进行深入研究。本领域技术人员通常可以采用常规方法进行高通量测序，包括但不限于：454FLX(Roche公司)、Solexa Genome Analyzer(Illumina公司)和Applied Biosystems 公司的SOLID等。这些平台共同的特点是极高的测序通量，相对于传统测序的96道毛细管测序，高通量测序一次实验可以读取40万到400万条序列，根据平台的不同，读取长度从25bp到450bp不等，因此不同的测序平台在一次实验中，可以读取1G到14G不等的碱基数。

Solexa 高通量测序包括DNA簇形成和上机测序两个步骤：PCR扩增产物的混合物与固相载体上固定的测序探针进行杂交，并进行固相桥式PCR扩增，形成测序簇；对所述测序簇用“边合成-边测序法”进行测序，从而得到样本中核酸分子的序列。

DNA簇的形成是使用表面连有一层单链引物(primer)的测序芯片(flow cell)，单链状态的DNA片段通过接头序列与芯片表面的引物通过碱基互补配对的原理被固定在芯片的表面，通过扩增反应，固定的单链DNA变为双链DNA，双链再次变性成为单链，其一端锚定在测序芯片上，另一端随机和附近的另一个引物互补从而被锚定，形成“桥”；在测序芯片上同时有上千万个DNA单分子发生以上的反应；形成的单链桥，以周围的引物为扩增引物，在扩增芯片的表面再次扩增，形成双链，双链经变性成单链，再次成为桥，称为下一轮扩增的模板继续扩增；反复进行了30轮扩增后，每个单分子得到1000倍扩增，称为单克

隆的DNA簇。

DNA簇在Solexa测序仪上进行边合成边测序，测序反应中，四种碱基分别标记不同的荧光，每个碱基末端被保护碱基封闭，单次反应只能加入一个碱基，经过扫描，读取该次反应的颜色后，该保护集团被除去，下一个反应可以继续反应，如此反复，即得到碱基的精确序列。在Solexa多重测序(Multiplexed Sequencing)过程中会使用Index(标签)来区分样品，并在常规测序完成后，针对Index部分额外进行测序，通过Index的识别，可以在1条测序甬道中区分多种不同的样品。

地中海贫血型别 SNP&InDel 数据库

本发明提供了一种地中海贫血型别 SNP&InDel 数据库及其构建方法。所述的数据库的内容如下所示：

71	HBA1:c.1A>G	Initiation codon (A->G)
542	HBA1:c.355G>T	Normal_542

如上标所示，第一列是突变的相对位置，第二列是具体突变信息：如第一行 HBA1 表示基因，A>G 表示参考序列由 A 突变为 G，最后一列代表型别，指出 HBA1 基因 71bp 位置由 A 突变为 G，地中海贫血型别为 Initiation codon (A->G)。

在不断的测试中，可对数据库的型别进行补充，如第二行，在测试样本中发现 HBA1 第 542 位置参考序列有 G 变为 T，属于同义突变，最后的型别给出 Normal。

本发明还提供了一种构建地中海贫血型别-SNP&InDel 关系数据库的方法，在一个优选例中，包括步骤：a) 选择多种正常无病基因的序列作为参考序列(如 hg19 中对应坐标的序列提取出来作为地中海贫血参考序列)；地中海贫血相关基因包括 HBB, HBA1, HBA2; b) 将现有地中海贫血数据库中已知型别的 HBA1 基因、HBA2 基因和 HBB 基因与参考序列比对，获得参考序列的差异位点，即 SNP 位点和 InDel 位点，得到每个型别相对于参考序列的 SNP 和 InDel 关系，得到地中海贫血型别 SNP-InDel 数据库。

SNP 和 InDel 的检测

本发明提供了一种检测 SNP 和 InDel 变异的方法，包括步骤：将测序数据

与对应的参考序列做比对，得到相对于参考序列的 SNP&InDel 连锁关系，通过所述的 SNP 和 InDel 连锁关系确定出型别。

具体地，用比对软件（优选 BWA、SOAP、BLAST、MAQ 等工具）将样本序列与参考序列进行比对；再将比对的结果文件进行格式转换（优选 samtools 软件工具进行转换），得到一致性序列文件；最后与 SNP&InDel 连锁关系数据库比较，获得地中海贫血的型别结果。

在一个优选例中，所述的 SNP&InDel 变异检测方法包括步骤：

a) 寻找 SNP 和 InDel 位点：对样本进行高通量测序，获得测序数据；将所述的测序数据与参考序列进行比对，得到 SNP 位点和 InDel 位点，获得深度分布（图 5）结果；对比对结果进行质量控制，过滤 SNP，InDel 支持数中质量值低于 20 的位点；

b) 校正 SNP 位点和 InDel 位点：由于利用 bwa 比对的罚分策略，InDel 附近如果有一样的重复序列，会导致 InDel calling 出现偏差：比如 ACTCT 缺失为 CT，无论是前面 CT 缺失还是后面 CT 缺失，bwa 都展示为前面的 CT 缺失，导致得到的 InDel 跟数据库比对不上，对于这些重复序列，都会基于数据库将缺失位置调整过来，然后利用校正好的 InDel 对 SNP 再进行校正。例如 97 位置缺失 C，98 位置参考序列由 T 突变为 G，这时将调整为 CT 被替换 G。

c) 寻找异常变异：利用 bwa 比对中的 smith-waterman 算法，将 bwa 比对中出现的 S 来记录复杂变异；利用 paired 关系，一条链比对很好，第二条链的一部分不能很好比对到参考序列，这时就会出现 S，例如 28S72M 表示前面 28bp 碱基没有比对到这个位置，利用更多的 reads 在同一位置出现的 S 来定义异常变异。本领域的普通技术人员可以使用常规方法，如 Shin Suzuki, Tomohiro Yasuda, Yuichi Shiraishi, Satoru Miyano, Masao Nagasaki, ClipCrop 等人在文献 BMC Bioinformatics 2011, 12(Suppl 14):S7: A tool for detecting structural variations with single-base resolution using soft-clipping information 中描述的那样，来寻找大片段的插入缺失和插入后缺失等复杂变异。将得到的变异位点作为错误输出，即得到复杂变异的信号数据，利用 sanger 测序法重新测序直接得到正确的变异位点。

d) 为了判定变异是否是同义突变，将得到的 SNP 和 InDel 做注释；然后将 SNP 和 InDel 比对到地中海型别数据库后得到地中海型别，同时补充数据库中不存在但通过样本测试得到的型别。

图 1 显示了地中海贫血 SNP 和 InDel 信息分析流程。

拷贝数变异的检测

本发明提供了一种检测拷贝数变异的方法，所述方法包括检测 HBB 基因的拷贝数变异部分和检测 HBA 基因的拷贝数变异部分。

在本发明的一个优选例中，包括如下步骤：

a) 计算参考基因与待测基因的读序 (reads) 数：CNV 部分分为 HBB 基因检测部分和 HBA 基因检测部分。首先利用 index+primer 将读序 (reads) 分到每个样品中。因为 HBA1 和 HBA2 基因同源性很高，设计时只设计一对引物，对两个基因进行捕获，利用 HBA1 和 HBA2 中的差异部分对实验得到的数据再次进行过滤得到 HBA1 和 HBA2 的 reads 数。在 HBB 基因部分中，计算 HBB 基因，control 基因的 reads 数；在 HBA 基因部分中，检测 HBA1 基因，HBA2 基因和 control 基因的 reads 数。

b) 计算相对比例：HBA 基因检测部分和 HBB 基因检测部分引入正常样本，排除标签 (index) 的扩增问题。

在 HBB 基因检测部分中，

检测待测样本 reads (HBB) /reads (control)：正常样本 reads (HBB) /reads (control)。

在 HBA 基因检测部分中，

待测样本 reads (HBA1) /reads(control)：正常样本 reads (HBA1) /reads(control)；

待测样本 reads (HBA2) /reads(control)：正常样本 reads (HBA2) /reads(control)；

待测样本 reads (HBA1) /reads(HBA2)：正常样本 reads (HBA1) /reads(HBA2)。

因为 HBA 和 HBB 的 CNV 检测很相似，但 HBA 更为复杂，以 HBA 为例描述 CNV 检测：HBA 定义 $R1 = \text{待测样本 reads (HBA1) /reads(control)}$ ：正常样本 reads (HBA1) /reads(control)； $R2 = \text{待测样本 reads (HBA2) /reads(control)}$ ：正常样本 reads (HBA2) /reads(control)， $R3 = \text{待测样本 reads (HBA1) /reads(HBA2)}$ ：正常样本 reads (HBA1) /reads(HBA2)。

而 HBB 定义为 $R1 = \text{检测待测样本 reads (HBB) /reads (control)}$ ：正常样本 reads (HBB) /reads (control)。因为一个文库有数个（一般 96 个）样品，将这 96 个样品结合在一起进行分型，以排除实验对整个文库的影响。

c) 对一个文库下所有样品的 R1, R2, R3 进行大概归类, 因为 R1, R2, R3 的理论值是 0、0.5、1、1.5 等, 将距离最近的值归于这一类, 例如 0.6 归于 0.5 类, 表示只有一个拷贝, 1.1 归于 1 类, 表示拷贝数正常, 对于中间值, 例如 0.75, 可以分到 0.5 类和 1 类中, 表示这个值暂时无法确认属于哪一类。

d) 计算每个类别中的方差和偏离度 (偏离度=均值—类别理论值), 计算均值与理论值之间的差距。例如理论值是 0.5, 而属于这类别的均值是 0.48, 那么这一类的偏离度为 0.02。

e) 综合考虑每个类别个数大于 15 个值的方差与偏离度, 构建整个文库的方差和偏离度。例如 0.5 类别中有 20 个样品, 偏离度为 0.02, 方差为 0.1, 类别 1 中有 70 个样品, 偏离度为 0.03, 方差为 0.03。那么整体的均值偏离度为 $(0.02*20+0.03*70)/90$, 整体的方差为 $(19*0.1+69*0.03)/88$ 。得到的均值偏离度和方差作为由于实验误差导致的固定误差, 重新得到每类别的理论值。例如 0.5 类别的理论值为 $0.5 - (0.02*20+0.03*70)/90$ 。

f) 计算每个值与每类别的理论值之间的距离, 计算公式: $(X_i - \text{ave}(X))/\sqrt{\text{var}(x)}$, 其中 $\text{ave}(X)$ 表示理论值, $\sqrt{\text{var}(x)}$ 表示方差。计算每个值的最短距离作为这个值应该属于的类别。例如最原始的值 0.75 无法分开, 但 1 类实际值为 0.85, 0.5 类值的实际值为 0.35, 这时, 0.75 判定距离是, 就距离 1 类更近了。

g) 选取每个 index 的 R1, R2, R3 的最短距离作为最佳选项。同时设定距离限制, 防止部分波动。同时利用贝叶斯思想, 即正常样品的概率很大, 所以, 对于距离限制时设置可以距离 1 类更远些。例如 0.8 到 1 的距离算出来为 0.2, 然后 0.7 到 0.5 的距离也为 0.2, 然后设置 0.15 为限制距离, 认为 0.15 以外的距离都不可靠, 但如果 0.8 修改为 0.85, 这时 0.8 属于 1 类的距离就是 1.5, 在允许之内, 而 0.7 就不属于 0.5 类。

HBB 基因拷贝数检测信息流程和 HBA 基因拷贝数信息分析流程分别见图 2 和图 3。

地中海贫血分型

如本文所用, 术语“地中海贫血的分型”是指检测 α 链和 β 链相关基因的变异, 包括点突变, 插入缺失和拷贝数变异等。

本发明提供了一种地中海贫血分型方法。具体包括如下步骤:

将 SNP, InDel 与型别数据库对应起来, 将 HBB 基因拷贝数变异与数据集

型别对应起来，将 HBA 基因拷贝数变异与数据集型别对应起来；同时考虑读序的深度问题，标记读序太少的标签；将三部分变异结果根据统一的标签总结到一起，获得每一个样品的完整变异情况。

本发明的主要优点包括：

- (1) 本发明方法和系统准确性高，对待测的数据要求比较低；
- (2) 本发明方法和系统对于信息的分析过程快速，操作简单。

下面结合具体实施例，进一步阐述本发明。应理解，这些实施例仅用于说明本发明而并不用于限制本发明的范围。下列实施例中未注明具体条件的实验方法，通常按照常规条件如 Sambrook 等人，分子克隆：实验室手册(New York: Cold Spring Harbor Laboratory Press, 1989)中所述的条件，或按照制造厂商所建议的条件。

实施例 1

960个样品的地中海贫血分型

本实施例的目的：通过捕获地中海贫血相关基因HBA基因和HBB基因，用高通量测序技术对SNP和InDel与CNV部分序列进行测序，利用重测序技术和CNV分析技术对变异进行检测，最后得出地中海贫血型别。图4显示了测序中引物设计的思路。

实验流程：

1. SNP&InDel检测

1.1 构建地中海贫血型别-SNP&InDel关系数据库

构建方法：

a) 从 hg19 的对应坐标中提取出 HBA1、HBA2、HBB 的序列作为参考序列。

b) 将现有地中海贫血数据库中已知型别的HBA1基因、HBA2基因和HBB基因与参考序列比对，获得参考序列的差异位点，即SNP位点和InDel位点，得到每个型别相对于参考序列的SNP和InDel关系，得到地中海贫血型别SNP-InDel数据库，地中海贫血已知的型别是从EBI（EBI: <http://www.ebi.ac.uk>）数据库下载的。同时将数据库中目前并不存在的但发明人在此过程中得到的型别添加进地中海贫血数据库中。

1.2 得到下机数据

小片段的插入缺失和点突变的检测包括HBB基因和HBA基因。利用引物（primer）特异性捕获基因后，在不同的样品上加上不同的标签（index）进行区分，随机打断回收200-500bp的胶片段，用高通量测序技术对其进行测序。最后用index+primer的特异性将reads分到不同的样品和基因中。

1.3 SNP和InDel初步检测

本实施例使用BWA比对软件和samtools软件，GATK软件包为例来说明。通过BWA比对软件，将这些序列与特定的参考序列进行序列比对，经过aln和sampe两步，得到比对结果*.sampe文件。

为了加快后续程序的运行速度，使用samtools软件工具的view将文件转化为*.bam格式，分别经过sort、index得到建立好index的bam文件。

然后利用GATK软件包对bam文件进行重新校正，bwa比对后同一位点的不同reads的比对并不相同，利用GATK软件包对InDel附近的低质量的SNP重新进行校正，更有利准确的InDel和SNP。

最后经过samtools的pileup得到原始的一致性序列，利用PERL语言编写的程序重新对质量值，深度进行过滤，得到可信的SNP和InDel。

1.4 SNP和InDel校正

由于全基因组存在很多重复序列，例如AGCCGCC，如果缺失为最后3bp的GCC，最后consensus文件会展现为A后面缺失GCC，所以要对这些重复区域附近的InDel重新纠正。

同时对InDel附近的SNP进行纠正，例如HBB97 C G，表示HBB基因97位置，ref的C突变为G。HBB 97 C -T，表示HBB基因98位置，缺失了一个T。这样的变异可以替换成delins，即缺失后插入变异。对于这种变异情况进行修正为CT/G，表示G替换了ref的CT碱基。

1.5 比对型别数据库

整理SNP和InDel格式，以“基因-位置-变异”为识别单位，比对到型别数据库，给出地中海贫血型别，变异型别不在数据库中的，标记为“Unknown”。

1.6 找复杂变异

性别数据库中存在一些复杂的变异情况，大片段的插入缺失，大片段的缺失后插入，bwa比对软件不能更好的展现这些变异情况。利用bwa比对算法，局部偶联算法中产生的soft chipping位置，soft chipping是由于pair end reads在比对时由于一条reads可以Uniq比对，而第二条出现了很多的mismatch，第二条reads

不会被丢弃，而是被保留下来，并标记S。

利用S可以找到断裂位点，这些断裂位点的产生可能是由于大片段插入，也可以是由于缺失导致的，可以通过收集S位点找到大片段的插入，缺失和缺失后插入等复杂的变异。

2. HBB基因拷贝数检测

2.1 HBB基因测序

基于引物特异性扩增，捕获长度为87bp的HBB基因片段，在两端加上标签(index)特异性标记每个样品后，利用高通量测序技术对HBB基因测序。

对下机后数据利用index+primer的特异性将读序(reads)分到每个样品和每个不同的基因中，分别同时进行检测。

2.2 修正识别序列并计算读序(reads)数

根据SNP和InDel对识别序列进行修正，由于存在识别序列区域发生突变，所以根据同一个样品中SNP情况对识别序列进行修正。对于多个SNP和InDel的情况，给出错误信息，因为多个SNP存在连锁关系，不能确定，InDel会严重影响序列，所以这两种情况都输出。利用识别序列对分到每个样品中的reads重新进行计数，得到4个绝对数据值，即：待测样本HBB基因reads数，待测样本控制基因reads数，正常样本HBB基因reads数，正常样本控制基因reads数。

2.3 计算比值

在检测HBB基因中加入正常样本排除了index对于不同的样品的影响，加入control基因作为一个绝对参考，排除不同的样品取样的绝对数量的不同。比值=待测样本HBB基因reads数/待测样本控制基因reads数：正常样本HBB基因reads数/正常样本控制基因reads数。

3. HBA基因拷贝数变异检测

3.1 HBA1基因和HBA2基因测序

HBA基因中HBA1和HBA2非常相似，所以只设计一对引物将HBA1，HBA2同时捕获。捕获长度为128bp的HBA基因片段，在两端加上index特异性标记每个样品后利用hiseq测序技术对HBA基因测序。对下机后数据利用index+primer的特异性将reads分到每个样品和每个不同的基因中，分别同时进行检测。

3.2 修正识别序列并统计reads数

修正识别序列与HBB基因部分一致。从HBA中利用HBA1与HBA2的差异序列将HBA分为HBA1和HBA2。按照本发明HBA基因拷贝数检测的流程，需要统计6个原始reads数，分别是正常样本HBA1基因reads，正常样本HBA2基因reads，

正常样本control基因reads，待测样本HBA1基因reads数，待测样本HBA2基因reads数，待测样本control基因reads数。

3.3 计算比值

HBA部分比较复杂，需要计算3个比值，定义为R1，R2，R3。

$R1 = \text{待测样本HBA1基因reads数} / \text{待测样本control基因reads数}$ ：正常样本HBA1基因reads数/正常样本control基因reads数。

$R2 = \text{待测样本HBA2基因reads数} / \text{待测样本control基因reads数}$ ：正常样本HBA2基因reads数/正常样本control基因reads数。

$R3 = \text{待测样本HBA2基因reads数} / \text{待测样本HBA1基因reads数}$ ：正常样本HBA2基因reads数/正常样本HBA1基因reads数。

R1和R2用来得到HBA1基因拷贝数绝对数值和HBA2基因拷贝数绝对数值，最后利用R3进行检测。

3.4 判定CNV型别

首先利用类似聚类的方法，对一个文库下所有的比值即R1，R2，R3进行大概聚类，判定规则为距离哪个值最近就归于哪类。然后求出每个类别中的方差和偏离度。偏离度表针的是由于实验误差导致的理论值的一个偏差。然后综合考虑每个类别有大于15个样品，用于构建出实验对整个文库的影响，包括方差和偏离度。计算公式 $(X_i - \text{ave}(X)) / \sqrt{\text{var}(x)}$ 也称为数据值的标准化。选取每个index的R1，R2，R3的最短距离作为最佳选项。同时利用贝叶斯思想，对属于正常样本的距离进行了减缩。

4. 整理所有变异

将SNP，InDel，HBB基因拷贝数变异，HBA拷贝数变异情况以index作为唯一标记，将他们整合到一起，同时整合一些质控信息，例如CNV primer区域的变异，reads绝对数据值过少等信息，最后给出完整的每个样品的变异信息。

在本发明提及的所有文献都在本申请中引用作为参考，就如同每一篇文献被单独引用作为参考那样。此外应理解，在阅读了本发明的上述讲授内容之后，本领域技术人员可以对本发明作各种改动或修改，这些等价形式同样落于本申请所附权利要求书所限定的范围。

权 利 要 求

1. 一种检测待测样本HBB基因拷贝数变异的方法,其特征在于,包括步骤:

(1) 获得待测样本的HBB基因序列,所述序列与样品和/或基因相对应;

(2) 对步骤(1)获得的读序进行修正和/或统计,获得下述绝对数据值:

待测样本HBB基因的读序数、待测样本对照基因的读序数、正常样本HBB基因的读序数、正常样本对照基因的读序数;

(3) 计算A值,

$$A值 = \frac{\text{待测样本HBB基因的读序数/待测样本对照基因的读序数}}{\text{正常样本HBB基因的读序数/正常样本对照基因的读序数}};$$

(4) 根据A值,获得待测样本HBB基因拷贝数变异的检测结果。

2. 一种检测待测样本HBA基因拷贝数变异的方法,其特征在于,包括步骤:

(1) 获得待测样本的HBA1基因序列和HBA2基因序列,所述的序列与样品和/或基因相对应;

(2) 对步骤(1)获得的读序进行修正和/或统计,获得下述绝对数据值: 正常样本HBA1基因的读序数, 正常样本HBA2基因的读序数, 正常样本对照基因的读序数, 待测样本HBA1基因的读序数, 待测样本HBA2基因的读序数, 待测样本对照基因的读序数;

(3) 计算下述的比值:

$$R1值 = \frac{\text{待测样本HBA1基因的读序数/待测样本对照基因的读序数}}{\text{正常样本HBA1基因的读序数/正常样本对照基因的读序数}}$$

$$R2值 = \frac{\text{待测样本HBA2基因的读序数/待测样本对照基因的读序数}}{\text{正常样本HBA2基因的读序数/正常样本对照基因的读序数}}$$

$$R3值 = \frac{\text{待测样本HBA2基因的读序数/待测样本HBA1基因的读序数}}{\text{正常样本HBA2基因的读序数/正常样本HBA1基因的读序数}};$$

(4) 根据R1值、R2值和R3值，获得待测样本HBA基因拷贝数变异检测结果。

3. 一种检测待测样本 HBB 基因拷贝数变异的检测单元，其特征在于，所述单元包括模块：

(1) 序列获取模块，用于获得待测样本的HBB基因序列，所述序列与样品和/或基因相对应；

(2) 读序修正和/或统计模块，用于对模块(1)获得的读序进行修正和/或统计；

(3) 比值计算模块，用于计算 A 值；和

(4) 输出模块，用于输出待测样本 HBB 基因拷贝数变异的信息。

4. 一种检测待测样本HBA基因拷贝数变异的检测单元，其特征在于，所述单元包括模块：

(1) 序列获取模块，用于获得待测样本的HBA1基因序列和HBA2基因序列，所述的序列与样品和/或基因相对应；

(2) 读序修正和/或统计模块，用于对模块(1)获得的序列读序进行修正和/或统计；

(3) 比值计算模块，用于计算 R1 值、R2 值、R3 值；

(4) 型别判断模块，用于判断 HBA 基因拷贝数变异的型别；和

(5) 输出模块，用于输出待测样本 HBA 基因拷贝数变异信息。

5. 一种检测地中海贫血的分型系统，其特征在于，包括：

(1) 权利要求3所述的检测待测样本HBB基因拷贝数变异的检测单元；和

(2) 权利要求4所述的检测待测样本HBA基因拷贝数变异的检测单元。

6. 如权利要求5所述的系统，其特征在于，还包括：

(3) SNP和/或InDel的检测单元。

7. 一种检测待测样本SNP和Indel变异的方法，其特征在于，包括步骤：

(1) 获得待测样本的基因测序数据；

(2) 将(1)中的测序数据与参考序列进行比对，获得相对于参考序列的SNP和InDel连锁关系数据；

(3) 根据(2)中所述的连锁关系数据确定待测样本SNP和Indel变异信息。

8. 一种对待测样本进行地中海分型的方法，其特征在于，包括步骤：

(1) 获得待测样本SNP和InDel变异数据，以及HBB基因、HBA基因拷贝数变异数据；

(2) 结合步骤(1)的所有变异数据，获得待测样本的完整变异情况，从而对待

测样本进行地中海分型。

9. 一种构建地中海贫血型别SNP和InDel关系数据库的方法，其特征在于，包括步骤：

(1) 将现有地中海贫血数据库中已知型别的HBA1基因、HBA2基因和HBB基因与作为参考序列的已知正常无病基因序列进行比对，获得差异位点，所述的差异位点为SNP和InDel位点；和

(2) 获得每个型别相对于参考序列的SNP和InDel关系，得到地中海贫血型别SNP-InDel数据库。

10. 一种SNP和Indel数据库，其特征在于，是用权利要求9所述的方法构建的。

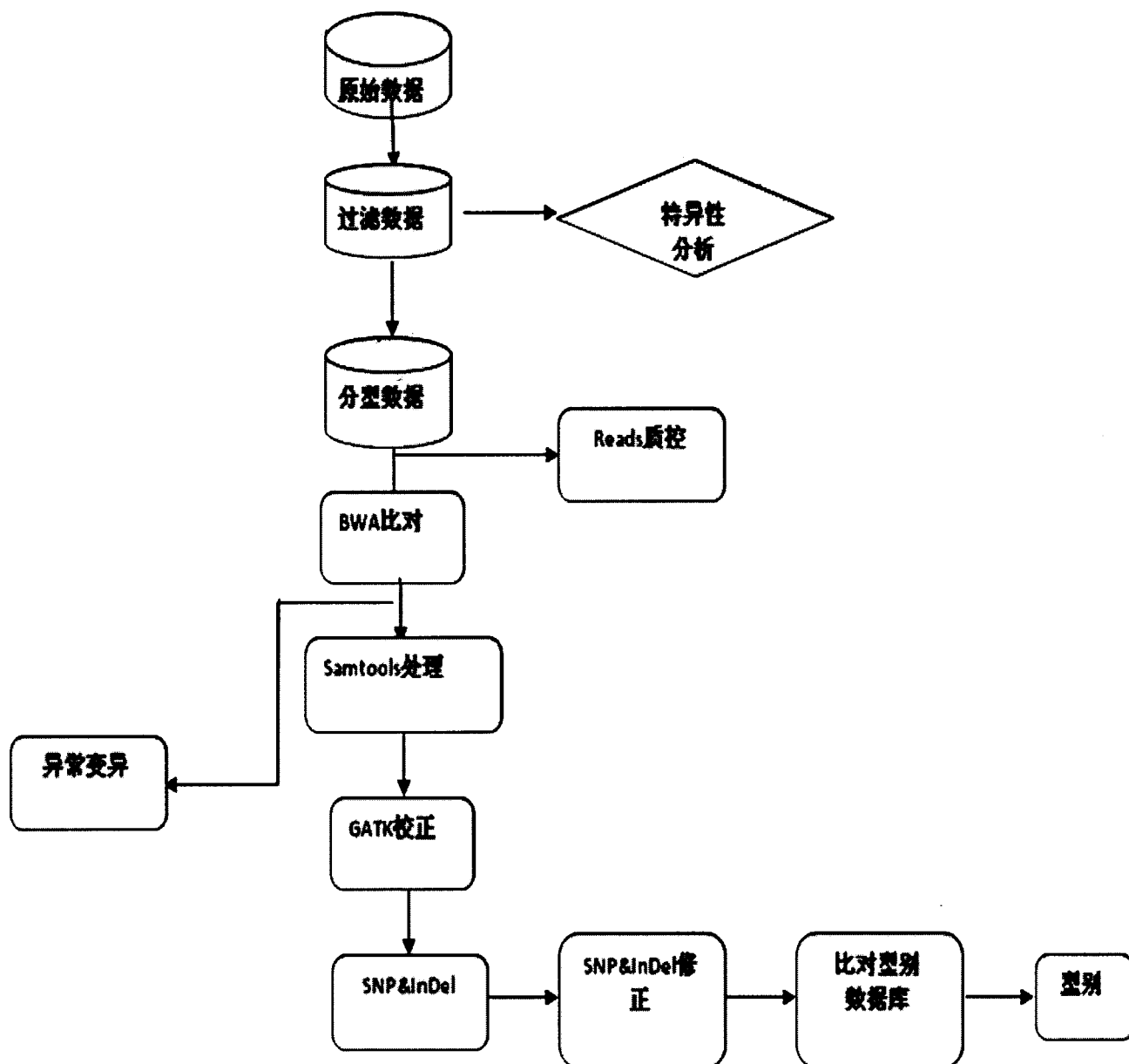


图 1

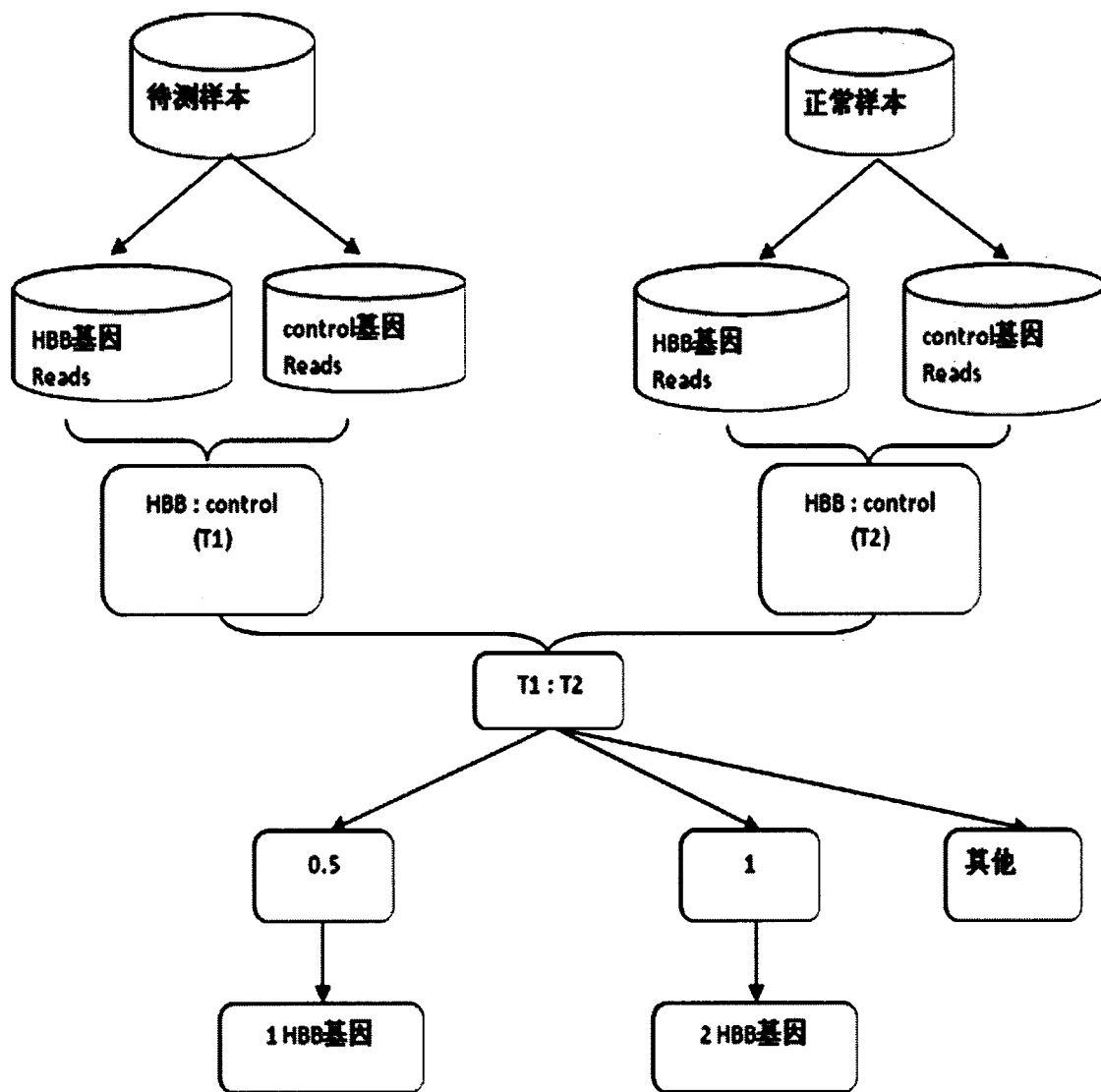


图 2

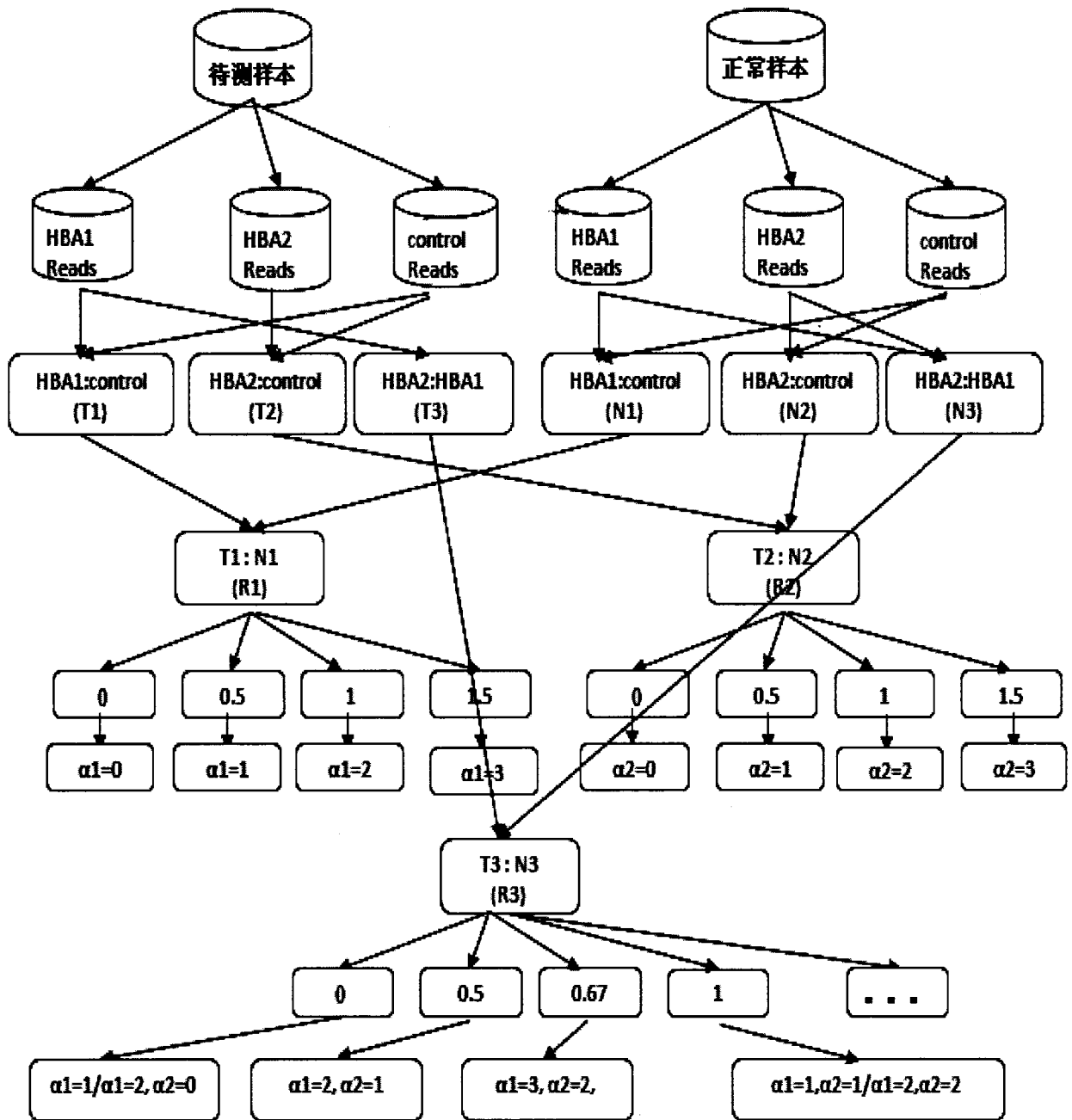


图 3

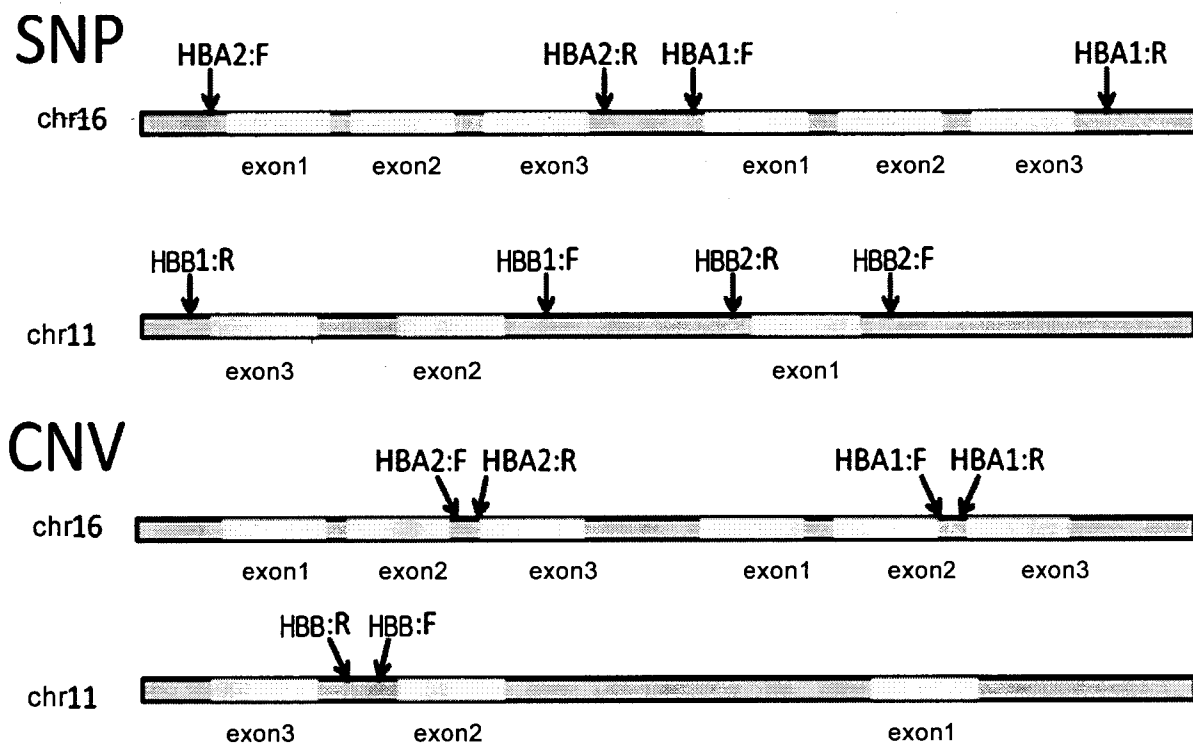


图 4

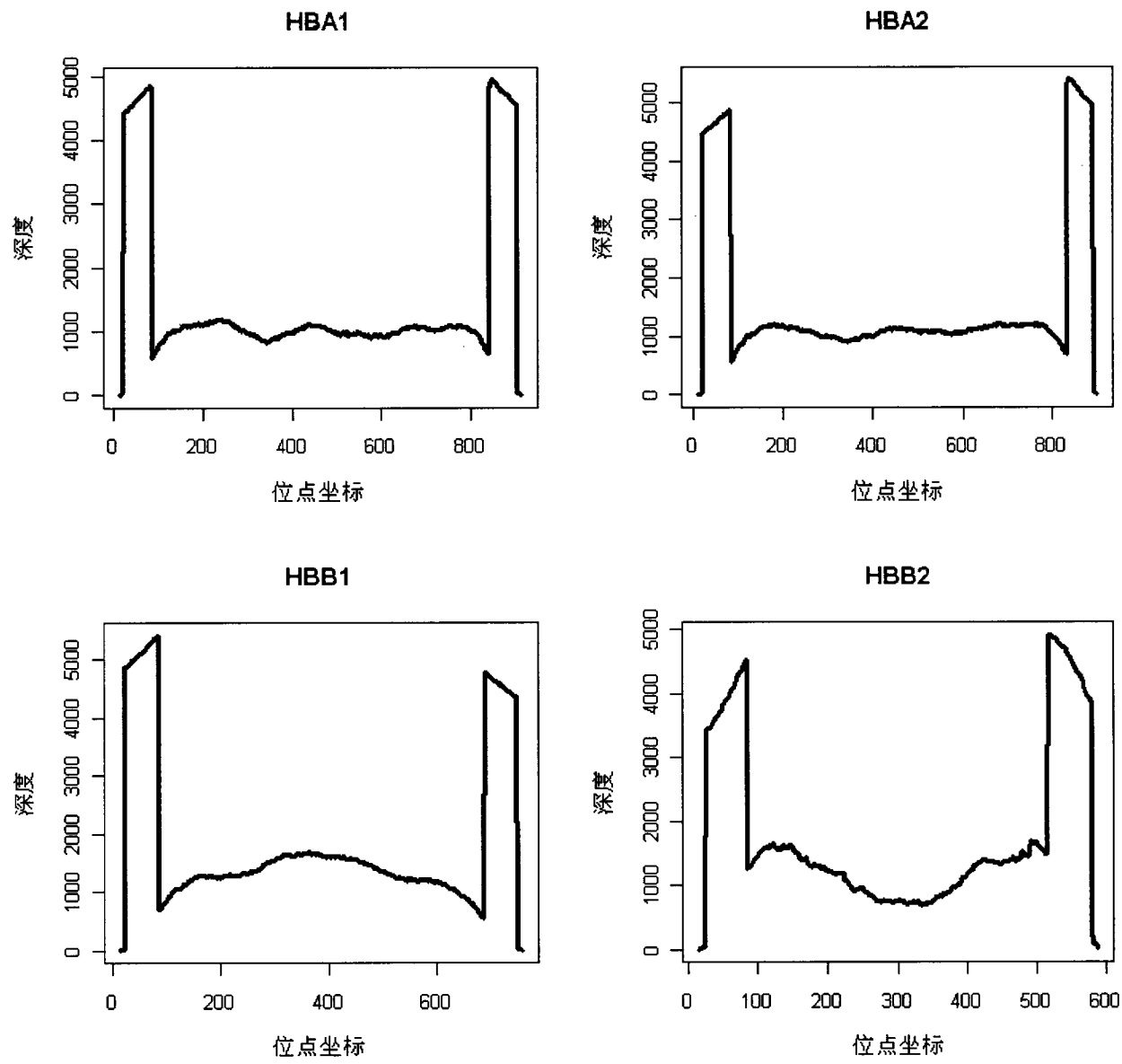


图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/087645

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: C12Q; G06F; C40B

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

DWPI, SIPOABS, CPRSABS, CPEA, TWABS, CNABS, JPABS, MOABS, HKABS, CNMED, TWMED, KRABS, AUABS, CNKI, ISI web of Knowledge, EMBASE, Google Scholar: mediterranean anemia, thalassemia, cooley anemia, type, typing, insertion, deletion, InDel, single nucleotide polymorphism, SNP, hemoglobin α/β , HBA, HBB, copy number variation/variant, CNV, read?, control, value, algorithm, formula, sequenc+, compar+, database

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SATHIRAPONGSASUTI, J. F. et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. BIOINFORMATICS. 09 August 2011, vol. 27, no. 19, pages 2648-2654 see the abstract, segments 2.2, 2.5 and 2.6	1-6
X	YANG, Xu et al. New-generation high-throughput technologies based 'omics' research strategy in human disease. HEREDITAS. August 2011, vol. 33, no. 8, pages 829-846 see the abstract, segment 1.2	7, 8

☒ 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 03 May 2013 (03.05.2013)	Date of mailing of the international search report 23 May 2013 (23.05.2013)
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 XING, Ying Telephone No. (86-10) 62411044

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/087645

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:
The claimed subject matter of claim 10 belongs 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/CN2012/087645

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	KITTS, A. and SHERRY, S. Chapter 5: The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation. THE NCBI HANDBOOK [Internet]. Last Update: 02 February 2011 see the abstract, page 1, line 1 to page 2, line 17, page 7, line 23 to page 8, line 25	9

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2012/087645

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68 (2006.01) i

G06F 19/22 (2011.01) i

C40B 50/00 (2006.01) i

C40B 40/06 (2006.01) i

国际检索报告

国际申请号

PCT/CN2012/087645

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: C12Q, G06F, C40B

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

DWPI, SIPOABS, CPRSABS, CPEA, TWABS, CNABS, JPABS, MOABS, HKABS, CNMED, TWMED, KRABS, AUABS, CNKI, ISI web of Knowledge, EMBASE, Google Scholar: 地中海贫血, 珠/球蛋白生成障碍性贫血, 库利(氏)贫血, 海洋性贫血, 海贫, 地贫, 分型, 插入, 缺失, 单核苷酸多态性, 血红蛋白, 拷贝数变异, 读数, 读序数, 序列数, 对照, 值, 公式, 序列, 测序, 比对, 对比, 数据库, mediterranean anemia, thalassemia, cooley anemia, type, typing, insertion, deletion, InDel, single nucleotide polymorphism, SNP, hemoglobin α/β , HBA, HBB, copy number variation/variant, CNV, read?, control, value, algorithm, formula, sequenc+, compar+, database

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	Jarupon Fah Sathirapongsasuti, 等. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. BIOINFORMATICS. 09.8 月 2011, 第 27 卷, 第 19 期, 第 2648-2654 页 参见摘要, 第 2.2、2.5、2.6 节	1-6

☒ 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

03.5 月 2013 (03.05.2013)

国际检索报告邮寄日期

23.5 月 2013 (23.05.2013)

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

幸颖

电话号码: (86-10) 62411044

C(续). 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	杨旭, 等. 基于新一代高通量技术的人类疾病组学研究策略. 遗传. 8 月 2011, 第 33 卷, 第 8 期, 第 829-846 页 参见摘要, 第 1.2 节	7-8
X	Adrienne Kitts, Stephen Sherry. Chapter 5: The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation. THE NCBI HANDBOOK [Internet]. Last Update: 02.2 月 2011 参见摘要, 第 1 页第 1 行至第 2 页第 17 行, 第 7 页第 23 行至第 8 页第 25 行	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：

权利要求10要求保护的主体属于单纯的信息表述（PCT实施细则39.1(v)）。

2. ☐ 权利要求：

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

具体地说：

3. ☐ 权利要求：

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

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

C12Q 1/68 (2006.01) i

G06F 19/22 (2011.01) i

C40B 50/00 (2006.01) i

C40B 40/06 (2006.01) i