

(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 5 月 22 日 (22.05.2014)



(10) 国际公布号
WO 2014/075274 A2

(51) 国际专利分类号:
C12Q 1/68 (2006.01)

(21) 国际申请号: PCT/CN2012/084689

(22) 国际申请日: 2012 年 11 月 15 日 (15.11.2012)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人 (对除美国外的所有指定国): **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: 及

(71) 申请人 (仅对美国): **张涛 (ZHANG, Tao)** [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)。 **曹红志 (CAO, Hongzhi)** [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)。 **王煜 (WANG, Yu)** [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)。 **全欣 (TONG, Xin)** [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)。 **刘小敏 (LIU, Xiaomin)** [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)。 **王俊 (WANG, Jun)** [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)。 **汪建 (WANG,**

Jian) [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)。 **杨焕明 (YANG, Huanming)** [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)。

(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:

— 包括按条约第 17 条(2)(a)所作的宣布, 不包括摘要, 发明名称未经国际检索单位审查。

(54) Title: MAJOR HISTOCOMPATIBILITY COMPLEX (MHC) TYPING METHOD AND APPLICATION THEREOF

(54) 发明名称: 一种主要组织相容性复合体 MHC 分型方法及其应用

(57) Abstract:

(57) 摘要:



WO 2014/075274 A2

一种主要组织相容性复合体MHC分型方法及其应用

技术领域

本发明属于生物信息学领域，具体地，本发明涉及一种主要组织相容性复合体MHC分型方法及其应用。

背景技术

主要组织相容性复合体（major histocompatibility complex, MHC）是存在于脊椎动物某一染色体上编码主要组织相容性抗原的一组紧密连锁的基因群，与免疫应答、免疫调节和移植排斥等有关。由于人类主要组织相容性抗原首先在白细胞表面被发现，故称其为人类白细胞抗原（human leucocyte antigen, HLA），并将人类的 MHC，即编码 HLA 的基因群称为 HLA 复合体。

人类白细胞抗原（Human leukocyte antigen, HLA）是人体内与免疫最相关的一段基因组区域。它位于人类6号染色体的短臂，由一系列紧密连锁的基因座构成。HLA基因是人类基因组中多态性最高，迄今为止人类最复杂的遗传系统之一。HLA基因也在识别自体与非体，调节免疫应答等方面起至关重要的作用。

HLA分型即确定HLA基因每个基因座上的等位基因的型别。目前 HLA分型的方法有多种，最早的为HLA血清学分型、细胞学分型的方法。现在主要为基于DNA水平的分型方法，包括单链构象多态性（PCR-SSCP）、序列特异性寡核苷酸探针（PCR-SSO(P)）、限制性片段长度多态性（PCR-RFLP）、基因芯片、序列特异性引物（PCR-SSP）以及基于序列分型法（sequence-based typing, SBT）。

综上所述，本领域迫切需要开发操作简单、准确性高的HLA分型方法。

发明内容

本发明的目的就是提供一种主要组织相容性复合体MHC分型方法及其应用。

30

在本发明的第一方面，提供了一种主要组织相容性复合体MHC型别数据库的构建方法，包括步骤：

(1) 将目标MHC型别序列和参考序列进行比对, 获得目标MHC型别相对于参考序列的差异位点, 所述的差异位点为SNP和/或InDel位点; 和

(2) 对已获得所述的差异位点信息的各个目标MHC型别进行汇总, 构建得到主要组织相容性复合体MHC型别数据库。

5 在另一优选例中, 所述的参考序列来自于hg18或hg19。

在另一优选例中, 所述的目标MHC型别序列为来自于数据库(如IMGT数据库)的已知序列。

在本发明的第二方面, 提供了一种主要组织相容性复合体MHC型别数据库, 所述的数据库是使用第一方面所述的方法构建的。

10 在另一优选例中, 所述数据库的结构为: 第一列表示MHC的不同型别, 第二列表示MHC型别所对应的序列, 第三列表示MHC型别序列比对参考序列后的SNP, 第四列表示MHC型别比对后的InDel。

在本发明的第三方面, 提供了一种构建主要组织相容性复合体MHC型别数据库的单元, 所述单元包括模块:

15 (1) 比对模块, 用于比对目标MHC型别序列和参考序列; 和

(2) 输出模块, 用比对模块获得的目标MHC型别序列和参考序列的比对结果, 输出MHC型别相对于参考序列的SNP和InDel位点信息。

在另一优选例中, 所述单元还包括: (3) 序列获取模块, 用于获得目标MHC型别序列和参考序列。

20 在本发明的第四方面, 提供了一种SNP和InDel 检测方法, 包括步骤:

(1) 获得经比对的不匹配的(unmap) SAM文件和目标基因区的SAM文件, 并且合并所述的SAM文件;

(2) 重新比对步骤(1)所获得的SAM文件, 获得每条读序比对的最佳型别、对应型别的初始位置、和与最佳型别差异的不匹配的数目; 和

25 (3) 过滤步骤(2)所获得的读序, 结合MHC数据库获得可信读序的SNP和InDel的信息。

在另一优选例中, 步骤(1)使用BWA软件进行比对。

在另一优选例中, 步骤(1)和步骤(2)之间还包括: 将SAM文件转换为FQ文件的步骤。

在另一优选例中，步骤(2)使用BWA软件进行比对。

在另一优选例中，步骤(2)使用目标基因对应的MHC型别序列作为参考序列。

在另一优选例中，步骤(2)所述的MHC数据库是用第一方面所述的方法制备的。

5 在本发明的第五方面，提供了一种SNP和InDel检测单元，所述单元包括模块：

(1) 比对模块，用于比对读序；

(2) 文件合并模块，用于合并经比对模块获得的比对的不匹配SAM文件和目标基因的SAM文件；

10 (3) 文件转换模块，用于转换文件合并模块获得的SAM文件和FQ文件；和

(4) 输出模块，用于输出SNP和/或InDel的信息。

在本发明的第六方面，提供了一种SNP和InDel的拼接方法，包括步骤：

(1) 将目标基因的一条读序所对应的SNP和InDel作为起始序列；

(2) 从目标基因中选择与步骤(1)获得的序列完全匹配的读序；

15 (3) 将步骤(2)选择的读序与步骤(1)的序列拼接，获得更长的序列；

(4) 从目标基因中提取与步骤(3)获得的序列完全匹配的读序，直到没有匹配的读序，从而获得拼接的序列；

(5) 将步骤(4)获得的拼接的序列与MHC型别数据库进行比对和过滤，对没有跨过整个外显子的序列进行随机组合和过滤，从而获得过滤的序列；和

20 (6) 对步骤(5)获得的过滤的序列按照读序支持数进行排序。

在本发明的第七方面，提供了一种SNP和InDel的拼接单元，所述单元包括模块：

(1) 序列获取模块，用于选择和/或获得读序；

(2) 拼接模块，用于对序列获取模块获得的读序进行读序与序列的拼接；

25 (3) 比对和过滤模块，用于比对和过滤拼接模块获得的读序；

(4) 排序模块，用于对比对和过滤模块获得的序列进行排序；和

(5) 输出模块，用于输出SNP和InDel的拼接信息。

在本发明的第八方面，提供了一种主要组织相容性复合体MHC分型方法，包括步骤：

(1) 获得目的基因的序列 (hap) 及其对应的MHC型别, 所述序列如式I所示:

$$(G_i-j)_n$$

式I

其中, G为基因类型, i为外显子编号, j为对应的序列排序数, n为外显子的数

5 目, i、j和n均为正整数;

(2) 对各个外显子序列对应的排序数进行打分, 确定最可信单体型 (type), 打分规则如下:

当j=1时, 分数=0;

当j=2时, 分数=0.5;

10 当j=其他时, 分数=1;

当没有外显子排序数时, 分数=2;

计算各个外显子对应的分数的总和, 分数最低的为最可信单体型;

(3) 分别比较其余单体型与最可信单体型的差异度, 差异度计算规则如下:

比较其余单体型与最可信单体型对应外显子的序列排序数 (即j值),

15 当j不同时, 分数=2;

当j相同时, 分数=1;

当其余单体型中不存在相应的j, 分数= -2;

最终总分最大的单体型与最可信单体型的组合, 为最佳组合; 和

(4) 基于步骤(3)的最佳组合, 得出MHC的分型信息。

20 在本发明的第九方面, 提供了一种主要组织相容性复合体MHC分型单元, 所述的单元包括模块:

(1) 序列获取模块, 用于选择和/或获得读序;

(2) 排序模块, 用于从序列获取模块获得的读序中, 确定最可信单体型和最佳单体型组合; 和

25 (3) 输出模块, 用于输出MHC的分型信息。

在本发明的第十方面, 提供了一种主要组织相容性复合体MHC分型系统, 所述系统包括单元:

(1) 本发明第九方面所述的主要组织相容性复合体MHC分型单元;

(2) 本发明第三方面所述的构建主要组织相容性复合体MHC型别数据库的

单元；

(3)本发明第五方面所述的SNP和InDel 检测单元；和

(4)本发明第七方面所述的SNP和InDel的拼接单元。

- 5 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之间都可以互相组合，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

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

图 1 显示了 MHC 的分型流程。

具体实施方式

- 15 本发明人经过广泛而深入的研究，首次建立了一种主要组织相容性复合体MHC分型方法。具体地，本发明提供了主要组织相容性复合体MHC型别数据库及其构建方法和构建单元、SNP和InDel检测方法和检测单元、SNP和InDel的拼接方法及拼接单元、以及主要组织相容性复合体MHC分型方法及其单元和系统。本发明方法和系统准确性高，对待测的数据要求比较低、相对于现有的分
- 20 型方法，大大提高了分型区域。

主要组织相容性复合体（MHC）

- 如本文所用，术语“主要组织相容性复合体”与“MHC”可以互换使用，都是指存在于脊椎动物某一染色体上的、编码主要组织相容性抗原的一组紧密连锁的基
- 25 因群，MHC 与免疫应答、免疫调节和移植排斥等有关。

人类白细胞抗原（HLA）

- 由于人类主要组织相容性抗原首先在白细胞表面被发现，故称其为人类白细胞抗原 (human leucocyte antigen, HLA)，并将人类的 MHC，即编码 HLA 的
- 30 基因群称为HLA 复合体。

人类白细胞抗原HLA是人体内与免疫最相关的一段基因组区域。它位于人类6号染色体的短臂，由一系列紧密连锁的基因座构成。HLA基因是人类基因组中多态性最高，迄今为止人类最复杂的遗传系统之一。HLA基因也在识别自体与非体，调节免疫应答等方面起至关重要的作用。

5

基因、外显子

如本文所用，术语“基因”是指是生物遗传的基本单位，存在于基因组上的基因区域内。在真核生物中，基因由内含子和外显子组成。基因一般拥有多个外显子。在很多情况下，基因拥有多个转录本，每个转录本是该基因的外显子的不同组合，甚至在外显子边界向外显子内缩减若干碱基，或者向内含子扩展若干碱基，这称为可变剪接。由于这些原因，一个基因可以拥有多个的转录本。生物在不同的环境不同的时间，可以获得不同的转录本。

10

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。

15

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。

20

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。

25

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

30

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三种第二代测序平台进行高通量测序：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进行，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达12种不同的样品。

MHC 型别数据库及其构建

- 5 本发明提供一种 MHC 型别数据库。所述数据库包括的内容用表 1 的形式表示。

表 1

MHC 型别	序列	SNP	InDel
G*01:05N	GTTCTCATACCCTCCAGTGGA TGATTGGCT……	29796327:T	29796435-D-C
G*01:01:05	GTTCTCACACCCTCCAGTGGA TGATTGGCT……	29796369:T	None

- 在表 1 中，第一列表示 MHC 的不同型别，第二列表示 MHC 型别所对应的序列，第三列表示 MHC 型别序列比对参考序列后的 SNP，第四列表示 MHC 型别比对后的 InDel。

例如，29796327:T 表示 29796327 位置 SNP 为 T；29796435-D-C 表示 29796436 位置缺失碱基 C；None 表示没有这种类型的变异。

本发明还提供了 MHC 型别数据库构建方法，在本发明的一个优选例中，所述方法包括步骤 a 和 b：

- 15 步骤 a) 下载已知的 MHC 型别序列，本领域的普通技术人员可以使用常规得到获得这些数据，例如，从 IMGT 数据库获得，网址为 <http://www.ebi.ac.uk/imgt/hla>。

MHC 型别序列的内容可以用表 2 的形式表示。

表 2

cDNA	1020	1030	1040
A*01:01:01:01	ATAGAAAAGGAGGGAGTTACACTCAGGCTGCAA		

- 20 在表 2 中，第一列表示型别，第二列表示型别的序列。

选取对照样本 hg19 或 hg18 中目标基因坐标对应的序列作为参考序列，例如 A 基因，选取 A*03:01:01:01 对应的序列作为参考序列。

步骤 b) 将 MHC 型别序列与参考序列进行比对, 获得与参考序列的差异位点, 即得到每个型别相对于参考序列的 SNP 和 InDel 关系, 构建 MHC 型别数据库。

5 **SNP&InDel 检测方法**

本发明提供了一种基于重新比对策略得到每条读序的 SNP 和 InDel 信息的方法。在一个优选例中, 所述方法包括步骤:

- a) 将 BWA 软件包比对得到的不匹配的 SAM 文件和目标基因区的 SAM 文件合并, 作为原始比对文件;
- 10 b) 将合并后的比对文件 (也就是 SAM 文件) 转为 FQ 文件;
- c) 将合并的 FQ 文件重新用 BWA 软件包比对, 参考序列选为目标基因对应的 MHC 型别序列, 如表 3 所示:

表 3

型别	型别对应的序列
>A*29:01:01:02N	ATAGAAAAGGAGGGAGCTACTCTCAGGCTGCAA
>A*29:02:01:02	ATAGAAAAGGAGGGAGCTACTCTCAGGCTGCAA

将型别作为表头, 而型别对应的序列作为参考序列;

- 15 d) 重新比对后, 得到每条读序比对的最佳型别, 以及对应型别的初始位置, 以及与最佳型别差异的 mismatch 数目。

通过设置 mismatch 数目过滤不可信的读序, 然后结合 MHC 型别数据库获得可信读序的 SNP 和 InDel 信息。

20 **基于连锁的 SNP 和 InDel 拼接**

本发明提供了一种基于连锁的 SNP 和 InDel 的拼接方法, 在一个优选例中, 所述方法包括步骤:

- a) 选取目标基因的一条读序对应的 SNP 和 InDel 作为起始序列, 用下面形式表示:

序列 1	29910242*29910331*29910276:G-29910286:T*None
------	--

- 25 b) 从目标基因中挑选与上述序列完全匹配的读序, 如下:

读序 1	29910282*29910371*29910286:T-29910358:G*None
读序 2	29910287*29910386*29910286:T-29910358:G-29910378:T*None
读序 3	29910282*29910371*29910286:T-29910371:T*None

c) 将挑选的读序与原始的序列进行拼接成更长的序列，如下：

序列 1	29910242*29910386*29910276:G-29910286:T-29910358:G-29910378:T*None
序列 2	29910242*29910371*29910276:G-29910286:T-29910371:T*None

d) 将得到的序列重新从目标基因中提取与序列完全匹配的读序直到没有匹配的读序；

e) 将没有 SNP 和 InDel 的读序按照同样的方法单独拼接；

- 5 f) 将拼接完成的序列比对到型别数据库进行过滤，将保留下来的序列，但是又没有跨过整个外显子的序列进行随机组合后重新过滤数据库，将过滤得到的 hap 排序，排序规则：首先选取读序支持数最多的一条 hap，然后将剩余的 hap 与这条 hap 取并集并去重复后，按照并集的读序支持数排序。

10 MHC分型方法

本发明提供一种MHC分型方法，在一个优选例中，步骤如下：

a) 将获得的hap以及hap对应的MHC型别转化为如表4所示的格式：

表4

外显子组合	型别
G1-1*G2-1*G3-2*G4-1*G5-2	G*01:04:03
G1-1*G2-2*G3-2*G4-1*G5-2	G*01:04:01
G1-2*G2-1*G3-1*G4-1*G5-1	G*01:01:01:01G*01:01:01:02 G*01:01:01:03
G1-2*G3-1*G4-1*G5-1	G*01:01:01:01G*01:01:01:02G*01:01:01:03 G*01:02:01

- 15 以 G*01:01:01:01G*01:01:01:02 G*01:01:01:03 为例，头两位数字表示的是在物种水平上的分型，第 3、4 位上的数字表示的是外显子非同义突变，第 5、6 位表示的同义突变，第 7、8 位上表示的是内显子上的突变的分型，但是由于在内显子上的突变的研究意义不大，所以一般只做前三部分的分型研究。

其中 G1 表示 G 基因 1 号外显子, -1 表示对应 hap 的排序数;

b) 确定最可信的 type 型别, 定义为 type 型别 1, 具体做法: 对所有的 type 型别进行打分, 选出最可信 type 型别作为 type 型别 1, 打分规则如下:

如果 hap 对应的排序数为 1, 则分数不加; 如果 hap 对应的排序数为 2, 则
5 分数加 0.5; 如果为其他数据, 则分数加 1; 如果部分外显子没有排序数, 则分数加 2; 如此分数最低的就定义为单体型 1。

例如:

$$G1-1 * G2-1 * G3-2 * G4-1 * G5-2 = 1$$

$$G1-1 * G2-2 * G3-2 * G4-1 * G5-2 = 1.5$$

$$10 \quad \underline{G1-2 * G2-1 * G3-1 * G4-1 * G5-1} = 0.5$$

$$G1-2 * G3-1 * G4-1 * G5-1 = 2.5$$

选择分数最低为 0.5 的 $G1-2 * G2-1 * G3-1 * G4-1 * G5-1$ 作为最可信 type, 从表 4 可以知道, 其对应的型别为 G*01:01:01:01 G*01:01:01:02 G*01:01:01:03。

c) 从剩余的 type 中选出 type2, 规则如下:

15 将剩下的 type 与步骤 b) 中获得的 type1 取差异度, 差异度表示同一个基因同一个 exon 他们对应的 hap 序数不一样, 这就是不同的 hap 取自原始读序的不同部分, 所以 hap 差异数越大, 表示这组型别组合可以得到原始读序的最大部分读序, 也就越可信。

计算规则, 如果外显子对应的 hap 序数不一样, 则加 2; 如果对应一样, 则
20 加 1; 如果 type1 存在而 type2 不存在, 则减 1; 最后得分最大的就是最佳组合。

以表 4 的数据为例, 差异度结果如下:

$$G1-1 * G2-1 * G3-2 * G4-1 * G5-2 \& \underline{G1-2 * G2-1 * G3-1 * G4-1 * G5-1} = 8;$$

$$\underline{G1-2 * G2-1 * G3-1 * G4-1 * G5-1} \& G1-1 * G2-2 * G3-2 * G4-1 * G5-2 = 9;$$

$$\underline{G1-2 * G2-1 * G3-1 * G4-1 * G5-1} \& G1-2 * G3-1 * G4-1 * G5-1 = 3。$$

25 因此, 最后选择 type 组合为 G*01:01:01:01 G*01:01:01:02 G*01:01:01:03 与 G*01:04:01。

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；
- (3) 相对于现有的分型方法，大大提高了分型区域。

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

4 个样品总计 102 个基因的分型

本实施例的目的：通过高通量测序技术对重测序数据或给予目标区域捕获数据进行 MHC 分型，确定 SNP 连锁性。

1: 构建 MHC 型别数据库

构建方法：

从网址 <http://www.ebi.ac.uk/imgt/hla> 下载最新的 MHC 型别对应的序列。

选取与对照样本 hg19/hg18 对应坐标的序列作为参考序列，对 MHC 型别序列进行比对，得到相对于参考序列的差异位点即 InDel 和 SNP。

2: 生成比对文件

利用 illumina 测序平台测序后得到下机数据，经过过滤低质量读序数和 adapter 污染的读序数后得到可利用的读序数。使用 BWA 比对软件和 samtools 软件为例来说明。

通过 BWA 比对软件，将这些序列与 hg19 或 hg18 作为参考序列进行序列比对，经过 aln 和 sampe 两步，得到比对结果*.sam 文件后，利用 samtools 工具包对*.sam 文件处理，包括排序，去重复，建立索引等处理得到*.bam 文件。

3: 挑出目标区域 SNP&InDel 信息

从 bam 文件中得到目标基因的比对信息，同时，与不匹配的读序合并，作为目标基因原始的比对信息，将比对信息重新转为 FQ 文件。

将 FQ 文件重新进行比对，比对参考序列改为目标基因对应的 MHC 型别，因为 MHC 多态性很高，部分读序在比对时被丢失掉，所以回收不匹配的读序重新进行比对，解决比对多态性问题。

从比对后的最佳型别对应的 SNP 和 InDel 中得到每条读序的比对信息。

5 4: 根据读序的 SNP 和 InDel 信息进行连锁

利用读序之间的 overlap 进行连锁，连锁原则，将读序上没有 SNP 和 InDel 的 reads 单独连锁。

5: 根据 MHC 型别数据库对连锁的 Hap 进行过滤以及连接

10 利用 MHC 型别数据库对拼接完的序列进行整体过滤，如果不能匹配数据库，则所有的读序都被丢失。而剩下的 hap 被定义为可信的 hap。

6: 对过滤得到的 hap 进行排序

首先取出读序支持数最多的一条 hap，定义为 hap1，然后将剩余的 hap 与 hap1 取并集后排序。

7: 结合所有外显子得到最后型别

15 将之前得到的 hap 以及 hap 对应的 MHC 型别转化为如下表 5 所示的格式：

表5

外显子组合	型别
G1-1*G2-1*G3-2*G4-1*G5-2	G*01:04:03
G1-1*G2-2*G3-2*G4-1*G5-2	G*01:04:01
G1-2*G2-1*G3-1*G4-1*G5-1	G*01:01:01:01G*01:01:01:02 G*01:01:01:03
G1-2*G3-1*G4-1*G5-1	G*01:01:01:01G*01:01:01:02G*01:01:01:03 G*01:02:01

其中 G1 表示 G 基因 1 号外显子，-1 表示对应 hap 的排序数。

首先确定最可信的 type，定义为 type1，利用打分对所有的 type 进行打分选出最可信 type 作为 type1，结果如下：

20 G1-1*G2-1*G3-2*G4-1*G5-2=1
 G1-1*G2-2*G3-2*G4-1*G5-2=1.5
 G1-2*G2-1*G3-1*G4-1*G5-1 =0
 G1-2*G3-1*G4-1*G5-1=2

选择 G1-2*G2-1*G3-1*G4-1*G5-1 对应的型别 G*01:01:01:01G*01:01:01:02 G*01:01:01:03 作为最可信 type（型别）。

从剩余的 type 中选出 type2，规则是将剩下的 type 与 type1 取差异度，差异度表示同一个基因同一个 exon 他们对应的 hap 序数不一样，这就是不同的 hap 取自原始读序的不同部分，所以 hap 差异数越大的表示这组型别组合可以得到原始读序的最大部分读序，也就是最可信了，差异度结果如下：

$$G1-1*G2-1*G3-2*G4-1*G5-2 \text{ \& } G1-2*G2-1*G3-1*G4-1*G5-1 = 8;$$

$$G1-2*G2-1*G3-1*G4-1*G5-1 \text{ \& } G1-1*G2-2*G3-2*G4-1*G5-2 = 9;$$

$$G1-2*G2-1*G3-1*G4-1*G5-1 \text{ \& } G1-2*G3-1*G4-1*G5-1 = 3;$$

10 所以最后选择 type 组合为 G*01:01:01:01G*01:01:01:02 G*01:01:01:03 与 G*01:04:01。

8. 综合上述步骤，4 个样品总计 102 个基因的分型结果见表 6。

表 6

样本	NA18532		NA18555		YH	
基因	PCR sanger 结果	本方法结果	PCR sanger 结果	本方法结果	PCR sanger 结果	本方法结果
A	30:01	30:01:01	02:06	02:06:01	02:01	02:01:01
	68:01	68:01:02	11:01	11:01:01	33:03	33:03:01
B	13:02	13:02:01	07:06	07:06	54:01	54:01:01
	15:18	15:18:01	54:01	54:01:01	58:01	58:01:01
C	06:02	06:02:01	01:02	01:02:01	01:02	01:02:01
	07:04	07:04:01	07:02	07:02:01	03:02	03:02:02
DRB1	04:01	04:01	04:05	04:05	03:01	03:01:01
	07:01	07:01	11:01	11:01	15:01	15:01:01
DQB1	02:02	02:02	03:01	03:01:01	02:01	02:01:01
	03:01	03:01:01	04:01	04:01	06:02	06:02:01
DPA1	01:03	01:03:01	02:02	02:02:02	01:03	01:03:01
	02:01	02:01:01	02:02	02:02:02	02:02	02:02:02
DPB1	03:01	03:01:01	05:01	05:01:01	--	04:01:01
	17:01	17:01	05:01	05:01:01	--	05:01:01

DQA1	02:01	02:01	03:03	03:03:01	01:02	01:02:01
	03:03	03:03:01	05:05	05:05:01	05:01	05:01:01
G	01:01	01:05N	01:01	01:01:03	01:01	01:01:01
	01:05N	01:05N	01:01	01:01:01	01:04	01:04:01

结果表明，本方法的正确率达到 98%以上。

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
5 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. 一种主要组织相容性复合体MHC型别数据库的构建方法，其特征在于，包括步骤：

5 (1) 将目标MHC型别序列和参考序列进行比对，获得目标MHC型别相对于参考序列的差异位点，所述的差异位点为SNP和/或InDel位点；和

(2) 对已获得所述的差异位点信息的各个目标MHC型别进行汇总，构建得到主要组织相容性复合体MHC型别数据库。

2. 一种主要组织相容性复合体MHC型别数据库，其特征在于，所述的数据
10 库是使用权利要求1所述的方法构建的。

3. 一种构建主要组织相容性复合体MHC型别数据库的单元，其特征在于，所述单元包括模块：

(1) 比对模块，用于比对目标MHC型别序列和参考序列；和

(2) 输出模块，用比对模块获得的目标MHC型别序列和参考序列的比对结果，
15 输出MHC型别相对于参考序列的SNP和InDel位点信息。

4. 一种SNP和InDel检测方法，其特征在于，包括步骤：

(1) 获得经比对的不匹配的（unmap）SAM文件和目标基因区的SAM文件，并且合并所述的SAM文件；

(2) 重新比对步骤(1)所获得的SAM文件，获得每条读序比对的最佳型别、对
20 应型别的初始位置、和与最佳型别差异的不匹配的数目；和

(3) 过滤步骤(2)所获得的读序，结合MHC数据库获得可信读序的SNP和InDel的信息。

5. 一种SNP和InDel检测单元，其特征在于，所述单元包括模块：

(1) 比对模块，用于比对读序；

25 (2) 文件合并模块，用于合并经比对模块获得的比对的不匹配SAM文件和目标基因的SAM文件；

(3) 文件转换模块，用于转换文件合并模块获得的SAM文件和FQ文件；和

(4) 输出模块，用于输出SNP和/或InDel的信息。

6. 一种SNP和InDel的拼接方法，其特征在于，包括步骤：

- (1) 将目标基因的一条读序所对应的SNP和InDel作为起始序列;
- (2) 从目标基因中选择与步骤(1)获得的序列完全匹配的读序;
- (3) 将步骤(2)选择的读序与步骤(1)的序列拼接, 获得更长的序列;
- (4) 从目标基因中提取与步骤(3)获得的序列完全匹配的读序, 直到没有匹配的

5 读序, 从而获得拼接的序列;

(5) 将步骤(4)获得的拼接的序列与MHC型别数据库进行比对和过滤, 对没有跨过整个外显子的序列进行随机组合和过滤, 从而获得过滤的序列; 和

(6) 对步骤(5)获得的过滤的序列按照读序支持数进行排序。

7. 一种SNP和InDel的拼接单元, 其特征在于, 所述单元包括模块:

- 10 (1) 序列获取模块, 用于选择和/或获得读序;
- (2) 拼接模块, 用于对序列获取模块获得的读序进行读序与序列的拼接;
- (3) 比对和过滤模块, 用于比对和过滤拼接模块获得的读序;
- (4) 排序模块, 用于对比对和过滤模块获得的序列进行排序; 和
- (5) 输出模块, 用于输出SNP和InDel的拼接信息。

15 8. 一种主要组织相容性复合体MHC分型方法, 其特征在于, 包括步骤:

(1) 获得目的基因的序列及其对应的MHC型别, 所述序列如式I所示:

$$(G_i-j)_n$$

式I

20 其中, G为基因类型, i为外显子编号, j为对应的hap排序数, n为外显子的数目, i、j和n均为正整数;

(2) 对各个外显子序列对应的排序数进行打分, 确定最可信单体型 (type), 打分规则如下:

当j=1时, 分数=0;

当j=2时, 分数=0.5;

25 当j=其他时, 分数=1;

当没有外显子排序数时, 分数=2;

计算各个外显子对应的分数的总和, 分数最低的为最可信单体型;

(3) 分别比较其余单体型与最可信单体型的差异度, 差异度计算规则如下:

比较其余单体型与最可信单体型对应外显子的序列排序数 (即j值),

当j不同时，分数=2；

当j相同时，分数=1；

当其余单体型中不存在相应的j，分数=-2；

最终总分最大的单体型与最可信单体型的组合，为最佳组合；和

5 (4) 基于步骤(3)的最佳组合，得出MHC的分型信息。

9. 一种主要组织相容性复合体MHC分型单元，其特征在于，所述的单元包括模块：

(1) 序列获取模块，用于选择和/或获得读序；

(2) 排序模块，从序列获取模块获得的读序中，确定最可信单体型和最佳单体

10 型的组合；和

(3) 输出模块，用于输出MHC的分型信息。

10. 一种主要组织相容性复合体MHC分型系统，其特征在于，所述系统包括单元：

(1) 权利要求9所述的主要组织相容性复合体MHC分型单元；

15 (2) 权利要求3所述的构建主要组织相容性复合体MHC型别数据库的单元；

(3) 权利要求5所述的SNP和InDel 检测单元；和

(4) 权利要求7所述的SNP和InDel的拼接单元。

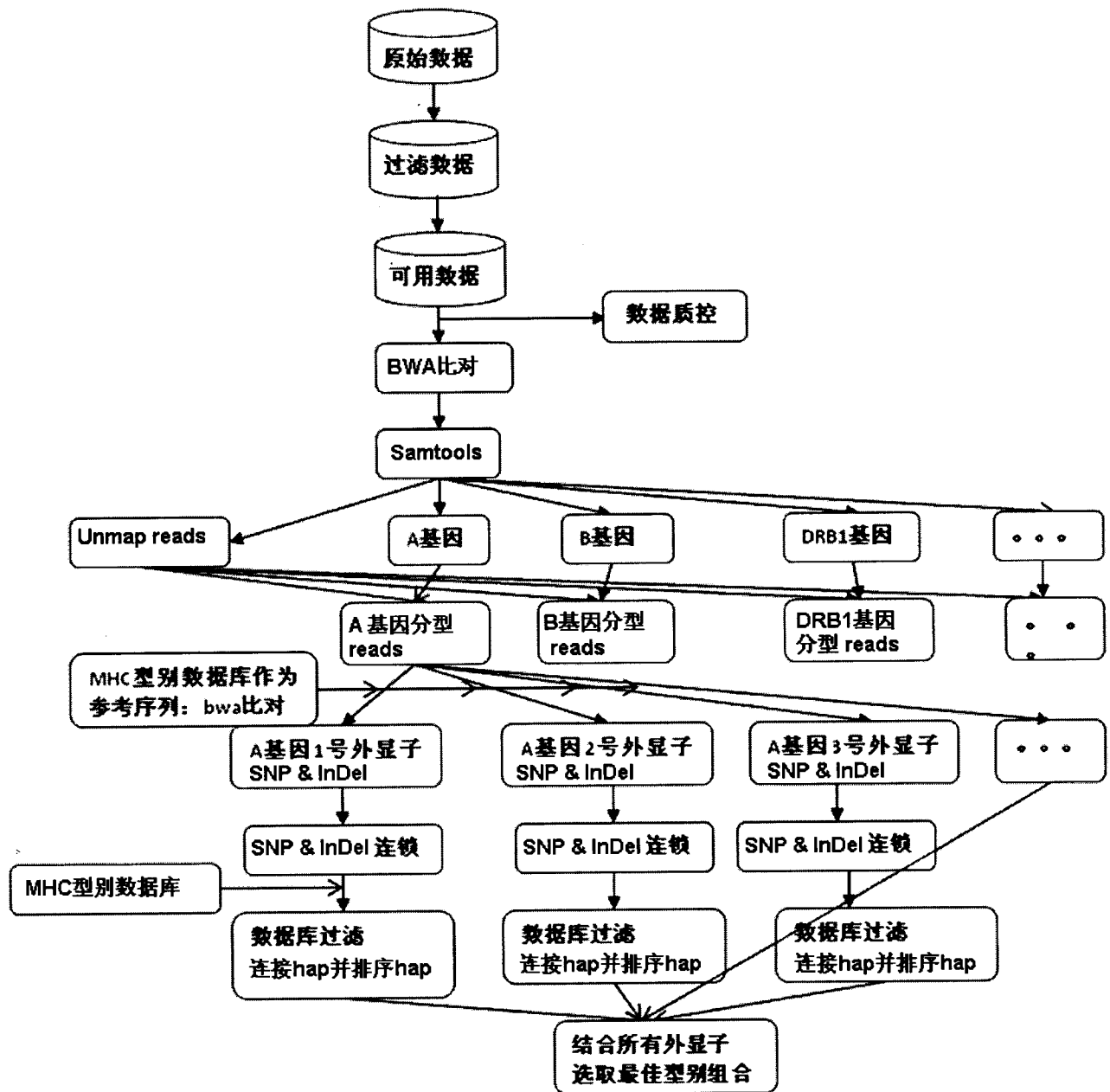


图 1

PATENT COOPERATION TREATY

PCT

DECLARATION OF NON-ESTABLISHMENT OF INTERNATIONAL SEARCH REPORT

(PCT Article 17(2)(a), Rules 13ter.1(c) and (d) and 39)

Applicant's or agent's file reference P2012-0912	IMPORTANT DECLARATION	Date of mailing (<i>day/month/year</i>) 29 August 2013 (29.08.2013)
International application No. PCT/CN2012/084689	International filing date (<i>day/month/year</i>) 15 November 2012 (15.11.2012)	(Earliest) Priority Date (<i>day/month/year</i>)
International Patent Classification (IPC) or both national classification and IPC C12Q 1/68 (2006.01) i		
Applicant BGI SHENZHEN CO., LIMITED et al		

This International Searching Authority hereby declares, according to Article 17(2)(a), that **no international search report will be established** on the international application for the reasons indicated below.

1. ☒ The subject matter of the international application relates to:
 - a. ☐ scientific theories
 - b. ☐ mathematical theories
 - c. ☐ plant varieties
 - d. ☐ animal varieties
 - e. ☐ essentially biological processes for the production of plants and animals, other than microbiological processes and the products of such processes
 - f. ☐ schemes, rules or methods of doing business
 - g. ☒ schemes, rules or methods of performing purely mental acts
 - h. ☐ schemes, rules or methods of playing games
 - i. ☐ methods for treatment of the human body by surgery or therapy
 - j. ☐ methods for treatment of the animal body by surgery or therapy
 - k. ☐ diagnostic methods practised on the human or animal body
 - l. ☒ mere presentations of information
 - m. ☐ computer programs for which this International Searching Authority is not equipped to search prior art
2. ☒ The failure of the following parts of the international application to comply with prescribed requirements prevents a meaningful search from being carried out:

☐ the description
☒ the claims
☐ the drawings
3. ☐ A meaningful search could not be carried out without the sequence listing; the applicant did not, within the prescribed time limit:

☐ furnish a sequence listing on paper complying with the standard provided for in Annex C of the Administrative Instructions, and such listing was not available to the International Searching Authority in a form and manner acceptable to it.

☐ furnish a sequence listing in electronic form complying with the standard provided for in Annex C of the Administrative Instructions, and such listing was not available to the International Searching Authority in a form and manner acceptable to it.

☐ pay the required late furnishing fee for the furnishing of a sequence listing in response to an invitation under Rule 13ter.1(a) or (b).
4. ☒ Further comments: (See extra sheet)

Name and mailing address of the ISA/CN The State Intellectual Property Office, the P.R.China 6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088 Facsimile No. 86-10-62019451	Authorized officer LI, Yang Telephone No. (86-10) 62411031
---	--

**DECLARATION OF NON-ESTABLISHMENT
OF INTERNATIONAL SEARCH REPORT**

International application No.
PCT/CN2012/084689

Claim 1 relates to an MHC type database construction method; claims 4 and 6 respectively relate to an SNP and an InDel testing and splicing method; and claim 8 relates to an MHC type-categorization method. All of these claims relate to methods or rules for purely mathematical operations, i.e. essentially forms of abstract thinking, and are thus considered methods of performing purely mental acts; therefore, said claims fall within the ambit of PCT Rule 39.1(iii).

Claim 3 relates to an MHC type database construction unit; claims 5 and 7 respectively relate to an SNP and an InDel testing and splicing unit; and claim 9 relates to an MHC type-categorization unit. Although these claims include functional modules, said modules by no means truly constitute components of physical devices. The substance of said claims, then, still relates to methods or rules for purely mathematical operations, and is regarded as methods of performing purely mental acts; therefore, said claims fall within the ambit of PCT Rule 39.1(iii).

Claim 10 relates to an MHC type-categorization system comprising all of the units described. Therefore, on the basis of the reasoning provided above, claim 10 also falls within the ambit of PCT Rule 39.1(iii).

Claim 2 relates to an MHC type database, but is a mere presentation of information and, thus, falls within the ambit of PCT Rule 39.1(v).

In sum, all of the claims of the present application relate to subject matter that does not require a search; hence, no international search has been conducted.

专 利 合 作 条 约

PCT

宣布不制定国际检索报告

(PCT 条约 17(2)(a)、细则 13 之三 1(c)和(d)和 39)

申请人或者代理人的档案号： P2012-0912	重 要 宣 布	发文日（日/月/年） 29.8 月 2013 (29.08.2013)
国际申请号 PCT/CN2012/084689	国际申请日（日/月/年）： 15.11 月 2012(15.11.2012)	(最早的) 优先权日（日/月/年）
国际专利分类（IPC）或者国家分类和 IPC 分类 C12Q 1/68 (2006.01) i		
申请人： 深圳华大基因科技有限公司 等		

本国际检索单位根据条约 17(2)(a)宣布，对于国际申请将不制定国际检索报告，其理由如下：

- ☒ 国际申请的主题涉及：
 - ☐ 科学理论。
 - ☐ 数学理论。
 - ☐ 植物品种
 - ☐ 动物品种。
 - ☐ 本质上是生物学方法的动植物生产方法，微生物方法以及用所述方法获得的产品除外。
 - ☐ 经营的方案、规则或方法。
 - ☒ 纯智力活动的方案、规则或方法。
 - ☐ 游戏的方案、规则或者方法。
 - ☐ 处理人体的外科手术或治疗方法。
 - ☐ 处理动物体的外科手术或治疗方法。
 - ☐ 在人体或动物体上实施的诊断方法。
 - ☒ 单纯的信息提供。
 - ☐ 本国际检索单位不具备检索现有技术的计算机程序。
- ☒ 国际申请的下列部分不符合规定的要求，以致不能进行有意义的检索：

☐说明书 ☒权利要求书 ☐附图
- ☐ 没有序列表,不能进行有意义的检索；申请人在规定的期限内：

☐ 没有提交符合《行政规程》附录 C 规定标准的纸件形式的序列表，并且国际检索单位也未获得形式和方式可以接受的序列表。

☐ 没有提交符合《行政规程》附录 C 规定标准的电子形式的序列表，并且国际检索单位也未获得形式和方式可以接受的序列表。

☐ 在答复根据细则 13 之三.1(a)或(b)的通知提交序列表时，未缴纳后提交费用。
- ☒ 补充意见：
(见附加页)

中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号： (86-10)62019451	受权官员 李洋 电话号码：(86-10) 62411031
--	---

权利要求 1 涉及 MHC 型别数据库的构建方法，权利要求 4 和 6 分别涉及 SNP 和 InDel 的检测和拼接方法，权利要求 8 涉及 MHC 分型方法，这些权利要求均涉及计算机程序进行的纯数学运算方法或规则，本质上属于人的抽象思维方式，是进行纯粹智力行为的方法，因此属于 PCT 细则 39.1 (iii)所列情形；权利要求 3 涉及构建 MHC 型别数据库的单元，权利要求 5 和 7 分别涉及 SNP 和 InDel 的检测和拼接单元，权利要求 9 涉及 MHC 分型单元，这些权利要求中虽然包括功能模块，然而其并不是能够真正构成有形设备的部件，因此其实质仍然是计算机程序进行的纯数学运算方法或规则，是进行纯粹智力行为的方法，属于 PCT 细则 39.1 (iii)所列情形；权利要求 10 涉及 MHC 分型系统，其包括上述各单元，因此基于同样的理由，也属于 PCT 细则 39.1 (iii)所列情形；权利要求 2 涉及 MHC 型别数据库，是单纯的信息表达，属于 PCT 细则 39.1 (v)所列情形。综上所述，本申请的全部权利要求均属于无需进行检索的主题，因此未进行国际检索。