

## (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  
2012 年 5 月 31 日 (31.05.2012)



(10) 国际公布号  
**WO 2012/068701 A2**

- (51) 国际专利分类号:  
*C12Q 1/68* (2006.01)
- (21) 国际申请号: PCT/CN2010/001879
- (22) 国际申请日: 2010 年 11 月 23 日 (23.11.2010)
- (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 号北山工业区综合楼 11F-3, Guangdong 518083 (CN)。
- (72) 发明人: 及
- (75) 发明人/申请人 (仅对美国): **曹红志 (CAO, Hongzhi)** [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山工业区综合楼, Guangdong 518083 (CN)。 **张伟 (ZHANG, Wei)** [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, Guangdong 518083 (CN)。 **王煜 (WANG, Yu)** [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山工业区综合楼, Guangdong 518083 (CN)。
- (74) 代理人: **中国国际贸易促进委员会专利商标事务所 (CCPIT PATENT AND TRADEMARK LAW OFFICE)**; 中国北京市西城区阜成门外大街 2 号万通新世界广场 8 层, Beijing 100037 (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, 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, PE, PG, PH, PL, PT, RO, RS, RU, 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, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, 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。
- 包括说明书序列表部分(细则 5.2(a))。

(54) Title: HLA GENOTYPE-SNP LINKAGE DATABASE, ITS CONSTRUCTING METHOD, AND HLA TYPING METHOD

(54) 发明名称: HLA 基因型别一 SNP 连锁数据库、其构建方法、以及 HLA 分型方法

(57) Abstract:

(57) 摘要:



WO 2012/068701 A2

## HLA 基因型别-SNP 连锁数据库、其构建方法、 以及 HLA 分型方法

### 技术领域

- 5 本发明属于基因组学和生物信息领域。涉及 HLA 基因型别-SNP 连锁数据库、其构建方法，确定 HLA 基因的 SNP 连锁关系的方法，HLA 分型方法，以及 HLA 分型装置。

### 技术背景

- 10 人类白细胞抗原，即 HLA(human leukocyte antigen, HLA)，是迄今为止发现的多态性最高的基因系统之一，它是调控人体特异性免疫应答和决定疾病易感性个体差异的主要基因系统，与同种异体器官移植的排斥反应密切相关。研究证明，在器官移植时，供体和受体双方的 HLA 匹配程度越高，移植的成功率越高(U. Shankarkumar. The Human Leukocyte Antigen  
15 (HLA) System. Int J Hum Genet, 4(2): 91-103 (2004))。

- HLA 基因位于 6 号染色体短臂上，长约 4000Kb。HLA 有几十个基因座位(基因座)，每个基因座位又有几十个等位基因，且呈共显性表达。由于 HLA 基因位于同一条染色体上，其多基因座位上的基因型组合相对稳定，很少发生同源染色体间交换，这就构成了以单元型(HAPLOTYPE，即在同一条染色体上紧密连锁的一系列等位基因的特殊组合)为特征的遗传。  
20

- HLA 基因有多种型别(即等位基因)，多种型别表示 HLA 每个基因座有多个等位基因。目前收录在 EBI 中的 HLA-A 型别有 1381 种，HLA-B 有 1927 种，HLA-C 有 960 种，HLA-DRB1 有 31 种，HLA-DQB1 有 127 种。不同型别序列间的差异很小，一般为几个 SNP(Single Nucleotide Polymorphisms，  
25 单核苷酸多态性)的差异。

- HLA 分型即检测每个基因座的等位基因。HLA 分型并不只是一种应用性的临床检测指标。HLA 分型方法有多种，最早的 HLA 血清学分型、细胞学分型方法，再后来出现基于 DNA 的分型方法，包括单链构象多态性(PCR-SSCP)、限制性片段长度多态性(PCR-RFLP)、序列特异性寡核苷酸  
30 探针(PCR-SSO(P))、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)(何丽，魏茂提，王世鑫. H L A

分型方法的研究进展. 免疫学杂志, 2006, 03(s)-0090-04; 王振雷, 何路军, 张 颀等. 人类白细胞抗原分型技术的进展. 中国组织工程研究与临床复, 2007, 11(37): 7457-7460)。

5 PCR-RFLP 操作繁琐, 结果解释复杂。PCR-SSP 对引物的设计和 PCR 条件要求很严, 且容易造成污染, 产生假阳性。基因芯片分型是用多态性的寡核苷酸为探针, 在芯片上固定多种分型的探针与目标序列杂交而进行判断, 这种分型方法价格比较昂贵且稳定性不高 (何丽, 魏茂提, 王世鑫. H L A 分型方法的研究进展. 免疫学杂志, 2006, 03(s)-0090-04; 王振雷, 何路军, 张 颀等. 人类白细胞抗原分型技术的进展. 中国组织工程研究与临床  
10 复, 2007, 11(37): 7457-7460)。

HLA-SBT (Sequence Based Typing, 基于 DNA 序列的分型方法) 是当前 HLA 高分辨率分型的主要方法。一般是通过基因的几个外显子的序列特征去分型, 如 HLA-A/B 的 2, 3, 4 号外显子, DRB1 的 2 号外显子。首先用 PCR 扩增获得 DNA 片段, 根据序列的信息进行 HLA 分型, 具有直观、高分辨且  
15 能检测新的等位基因的特点。HLA-SBT 方法主要基于 Sanger 测序分型和 454 测序分型方法。基于 Sanger 测序法的优点是准确性比较高, 缺点是测序通量小, 耗时长, 价格比较昂贵, 并且软件分型时导入的测序峰图质量的好坏对分型软件的峰图识别能力影响很大, 当软件识别错误时, 要求分型人员能及时发现并改正错误。软件分型没有达到自动化, 降低了分型效率,  
20 很难应用于大规模 HLA 高分辨分型项目。454 测序分型主要是将测得的样本序列与已有的 HLA 型别序列比对, 而通过软件预测的结果有一部分不能确定, 需人工去分析才能最后确定, 精确度还需提高 (G. Bentley, R. Higuchi, B. Hoglund. High-resolution, high-throughput HLA genotyping by ext-generation sequencing. Tissue  
25 Antigens, 2009, 3 - 403)。

## 发明内容

本发明通过选择参考序列, 将目前已知型别的 HLA 基因与参考序列进行比对, 构建 HLA 基因型别-SNP 连锁 (关系) 数据库。将实验测得的样本  
30 序列与参考序列比对, 得到相对于参考序列的 SNP 连锁关系, 通过这个 SNP 连锁关系确定出型别。具体来说, 用比对软件 (如 BWA、SOAP、BLAST、MAQ 等) 将样本序列与参考序列进行比对, 再用 samtools 软件工具对比对的结

果文件格式转换,得到一致性序列文件,然后通过一些判断条件确定两个位点的 SNP 连锁关系,之后进一步确定出整体的 SNP 连锁关系,最后将整体的 SNP 连锁关系与构建好的型别-SNP 连锁关系数据库比较,得到 HLA 的型别,有时还需对型别结果过滤。由此提供了下述发明:

5 本发明的一个方面涉及构建 HLA 基因型别-SNP 连锁(关系)数据库的方法,包括如下步骤:

a) 选择一个或多个 HLA 基因座的序列作为参考序列;

b) 将现有 HLA 数据库中已知型别的 HLA 基因与参考序列比对,找出与参考序列的差异位点即 SNP 位点,得到每个型别相对于参考序列的 SNP 连锁关系,构建 HLA 基因型别-SNP 连锁(关系)数据库。

10 HLA 总共有 200 个左右的基因座,关于步骤 a),所述参考序列优选是已发现的、序列比较准确和完善的、长度较长的。具体地,所述参考序列是常见型别,而非罕见型别,并且整个基因序列都已知(因为有的 HLA 型别只在几个外显子区有已知序列,其它外显子部分序列没有)。在本发明的一个实施方案中,具体地,将 SEQ ID NO: 1-5 作为参考序列,如下面的表 1 所示:

表 1: 选择的参考序列

| 参考序列     | 基因座名称            | SEQ ID NO: |
|----------|------------------|------------|
| HLA-A    | A*01: 01: 01: 01 | 1          |
| HLA-B    | B*07: 02: 01     | 2          |
| HLA-C    | C*01: 02: 01     | 3          |
| HLA-DRB1 | DRB1*01: 01: 01  | 4          |
| HLA-DQB1 | DQB1*02: 01: 01  | 5          |

在本发明的一个实施方案中,关于步骤 b),所述现有 HLA 数据库可以是 EBI (European Bioinformatics Institute) 的 HLA 数据库 (http://www.ebi.ac.uk/imgt/hla/)。

下面的表 2 示出了型别-SNP 连锁关系的部分数据。

表 2: 部分 HLA-A 的型别-SNP 连锁关系

| 型别名称              | 对应的 SNP 连锁关系                                                                        |
|-------------------|-------------------------------------------------------------------------------------|
| A*02: 01: 01: 01  | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722: C-729: T-731: G-771: C |
| A*02: 01: 01: 02L | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722: C-729: T-731: G-771: C |
| A*02: 01: 01: 03  | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722: C-729: T-731: G-771: C |

|              |                                                                                               |
|--------------|-----------------------------------------------------------------------------------------------|
| A*02: 01: 02 | 574: A-633: G-670: T-686: G-687: G-700: A-701: G-722: C-729:<br>T-731: G-771: C               |
| A*02: 01: 03 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-724: T-729: T-731: G-771: C |
| A*02: 01: 04 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 05 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 06 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 07 | 508: T-553: T-574: A-633: G-670: T-686: G-687: G-700: A-701:<br>G-722: C-729: T-731: G-771: C |
| A*02: 01: 08 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 09 | 508: T-633: G-670: T-686: G-687: G-700: A-701: G-722: C-729:<br>T-731: G-771: C               |
| A*02: 01: 10 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 11 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 12 | 508: T-556: A-574: A-633: G-670: T-686: G-687: G-700: A-701:<br>G-722: C-729: T-731: G-771: C |
| A*02: 01: 13 | 508: T-574: A-633: G-670: T-686: G-687: G-691: A-700: A-701:<br>G-722: C-729: T-731: G-771: C |
| A*02: 01: 14 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 15 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 17 | 508: T-574: A-633: G-686: G-687: G-700: A-701: G-722: C-729:<br>T-731: G-771: C               |
| A*02: 01: 18 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 19 | 508: T-574: A-577: C-633: G-670: T-686: G-687: G-700: A-701:<br>G-722: C-729: T-731: G-771: C |
| A*02: 01: 21 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 22 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 23 | 508: T-541: T-574: A-633: G-670: T-686: G-687: G-700: A-701:<br>G-722: C-729: T-731: G-771: C |

|              |                                                                                               |
|--------------|-----------------------------------------------------------------------------------------------|
| A*02: 01: 24 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 25 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 26 | 508: T-574: A-633: G-652: A-670: T-686: G-687: G-700: A-701:<br>G-722: C-729: T-731: G-771: C |
| A*02: 01: 27 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 28 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 29 | 508: T-547: T-574: A-633: G-670: T-686: G-687: G-700: A-701:<br>G-722: C-729: T-731: G-771: C |
| A*02: 01: 30 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 31 | 508: T-574: A-633: G-670: T-686: G-687: G-700: G-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 32 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 33 | 508: T-571: T-574: A-633: G-670: T-686: G-687: G-700: A-701:<br>G-722: C-729: T-731: G-771: C |
| A*02: 01: 34 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 35 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 36 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |
| A*02: 01: 37 | 508: T-574: A-633: G-670: T-686: G-687: G-700: A-701: G-722:<br>C-729: T-731: G-771: C        |

本发明的还一方面涉及根据上面本发明的构建 HLA 基因型别-SNP 连锁（关系）数据库的方法构建的 HLA 基因型别-SNP 连锁（关系）数据库。

5 本发明的还一方面涉及确定 HLA 基因的 SNP 连锁关系的方法，包括如下步骤：

1) 找出 SNP 位点：将样本序列与上述的参考序列进行比对，统计分析，得到样本序列与参考序列的差异位点，即 SNP 位点；

其中，样本序列可以通过多种测序方法得到，例如 SOLEXA、454、SOLID、

单分子测序技术等，优选地，采用单分子测序技术。术语“单分子测序技术”是基于纳米孔（nanopore）的单分子读取技术，有着更快的数据读取速度，如 Helicos 测序仪。

2) 初步确定 SNP 连锁关系：先根据经过 SNP 位点的 read 数目，确定出两个 SNP 位点的连锁关系，根据经过 SNP 位点的序列支持数，将两两 SNP 连锁关系连接起来，不断延伸，直到确定出整条 SNP 连锁关系（如 Fig. 1 所示）；如果相邻两 SNP 位点没有连接起来，使这两个 SNP 位点前后两部分的 SNP 连锁关系在此无法连接，则对这两个 SNP 位点前后两部分的 SNP 连锁关系自由组合连接。由于人是二倍体生物，最多只存在两种等位基因。最多保留两种连锁关系。

这种确定 SNP 连锁关系的方法效果很好，能够非常准确的确定出 SNP 连锁关系。

优选地，本发明的确定 HLA 基因的 SNP 连锁关系的方法，还包括如下步骤：

3) 通过聚类进一步确定 SNP 的连锁关系：从步骤 2) 中选择两条 SNP 连锁关系，然后对 reads 聚类，通过聚类，将 reads 分为两类，分别属于两条 SNP 连锁。

本发明还一方面涉及 HLA 分型方法，包括上面本发明的确定 HLA 基因的 SNP 连锁关系的方法，并且还包括如下步骤：

4) 型别确定：将得到的 SNP 连锁关系与本发明的型别-SNP 连锁数据库比较，得到与 SNP 连锁关系对应的 HLA 基因型别。

优选地，本发明的 HLA 分型方法还包括下述步骤：

5) 结果过滤：对每个基因判断出来的型别结果进一步进行过滤，去除假阳性；

过滤的方法可以是对预测的型别结果进行逻辑关系分析，以去掉自相矛盾的结果。

优选地，本发明的 HLA 分型方法还包括如下步骤：

6) 结果评价：对得到的结果进行评价，根据一定的条件判断结果是正确的还是错误的。具体地，需要综合考虑多个外显子，只有在多个外显子同时满足要求的条件，才认为是正确的。

在本发明的一个实施方案中，关于步骤 1)，首先将样本的所有序列与参考序列进行比对，序列比对软件可以用 BWA、SOAP、BLAST、MAQ 等；接着对序列比对的结果进行统计分析并进行处理，构建出一致性序列，显示与参考序列比对的所有位点信息，每个位点最多只有两个碱基；然后从一致性序列中提取出是 SNP 的位点信息，包括每个位点的位置、该位点上的比对序列的碱基、序列支持数等。上述过程可以用 samtools 或 soapnp 等软件进行操作。

在本发明的一个实施方案中，关于步骤 2)，SNP 连锁关系的确定需两步才能实现。第一步确定出两两 SNP 位点的连锁关系。在两个位点中，如果有杂合 SNP 位点，两个 SNP 连锁会有多种连接情况，但因为一个 SNP 位点有两个碱基，则这两个碱基应该同时存在于最终的 SNP 连锁关系中，并且是在不同的 SNP 连锁关系中，根据这种逻辑关系分析，将两个位点 SNP 连锁关系分成两类。再根据经过每个位点的序列支持数和同时通过这两个位点的序列支持数等信息去分析判断，确定出两两 SNP 位点的连锁关系，最多只保留两条连锁关系（因为人为二倍体，SNP 连锁最多只有两种，所以每两个位点的连锁关系最多也只有两种情况）。这里尽量多确定出两 SNP 位点的连锁关系，理想的情况下能够确定出任意两位点的连锁关系，但由于测序深度或测序错误等原因，可能部分两位点的 SNP 连锁关系无法确定。第二步是整合上一步得到的两 SNP 位点连锁关系，将相邻两 SNP 位点依次连接起来，不断延伸，确定出整个外显子的 SNP 连锁关系（如 Fig.1 所示）。这是本发明非常关键的一步，确定的 SNP 连锁关系越多，会带来越来越多的假阳性，本发明用了一个很好的方法确定 SNP 的连锁。

在本发明的一个实施方案中，关于步骤 3)，对 reads 聚类是为进一步优化 SNP 的连锁关系。因为在前面的步骤中确定 SNP 时会出现一定的错误：有的是假 SNP 位点，有的把纯合 SNP 确定为杂合或把杂合 SNP 确定为纯合；再者，在上一步 SNP 连锁关系确定时，由于所测的序列比较短等原因，个别的 SNP 位点在连锁时无法确定或连锁错误。基于以上原因，聚类这一步显得很重要，它能在很大程度上解决以上问题。从上一步中选择两条最有可能的 SNP 连锁关系，根据两条连锁关系上的 SNP 对 reads 聚类，将 reads 分成两类，再分别对这些 reads 的比对结果构建两条一致性序列，由一致性序列得到新的 SNP 连锁关系。经过这一步后，新 SNP 连锁关系的准确度已经非常的高了。在该步骤中，如何选择两条最优的 SNP 连锁关系去作聚类是很关键的，

可以根据上一步得到的初步 SNP 关系和每个 SNP 位点序列支持数等信息，综合各种条件进行考虑。对于那种经各条件判断后还无法精确选择两条连锁关系的，可以给出多组连锁关系去聚类。

在本发明的一个实施方案中，关于步骤 5)，因为只考虑外显子的序列，所以还要提取出基因中多个外显子共有的型别，由于多个外显子的型别可以自由组合导致假阳性，序列或其它方面也可能出现一定的假阳性，所以在最后要对结果进一步过滤。过滤的方法是通过逻辑关系分析滤掉假阳性。逻辑关系分析方法指的是根据型别结果由各外显子对应的 SNP 连锁关系去判断，如果在一外显子上存在两条 SNP 连锁关系，并且都有对应的型别结果，那么在最终的结果中每条 SNP 连锁关系只有一种对应的型别，若有多种对应型别，只有一种型别结果是正确的。每个外显子确定出的 SNP 连锁关系都对应型别，最初的型别结果是各外显子共有的型别，分析型别对应的外显子 SNP 连锁来源，在逻辑上，有些型别是不应存在的，即该型别为假阳性结果。具体分析可见具体实施例。在具体实施例中可以发现，用本发明的分型方法分型得到的初步结果大部分都只有两种正确型别结果，假阳性很少，只有比较少的一部分需要用逻辑关系去除假阳性，通过这种方法来分析，几乎可以去掉所有的假阳性结果。

在本发明的一个实施方案中，关于步骤 6)，对结果的评价是为了确保给出的结果是正确的。对于给定型别的结果，如果这些型别的 SNP 信息和经聚类后得到的 SNP 完全一样，则认为给定的型别结果是正确的，否则是错误的。这样的评价需要综合考虑多个外显子，只有在多个外显子同时满足要求的条件下，才认为是正确的。A, B, C 位点需要同时考虑 3 个外显子的情况，DQB1 需要考虑 2 个外显子的情况。这样的评价标准效果很好，准确度非常的高。

关于本发明的 HLA 分型的方法，具体流程图如 Fig. 2 所示。

本发明的又一方面涉及 HLA 分型装置，可以实现本发明的 HLA 分型的方法，其包括如下单元：

- 1) SNP 连锁关系确定单元；
- 2) 型别确定单元。

在本发明的一个实施方案中，所述 SNP 连锁关系确定单元包括：找出 SNP 位点的单元、初步确定 SNP 连锁关系的单元；优选地，还包括通过聚类进一步确定 SNP 的连锁关系的单元。所述型别确定单元包括型别-SNP 连锁数据库

单元。

在本发明的一个实施方案中，所述 HLA 分型装置，还包括结果过滤单元；并且优选地，还包括结果评价单元。

5        在本发明中，术语“SNP 连锁（关系）”是指这样一种关系，在同一条 DNA 链上的，可以将 SNP 位点按该关系顺序连接起来。

术语“一致性序列文件”是指一段 DNA 相关序列，序列反应了每个位置最可能出现的一个或几个碱基。

10        术语“read”或“reads”是指测序的结果，即测得的序列。例如，通过 solexa 测得的一条序列可以称为 read。

术语“序列支持数”是指经过某个位点的 read 的数目。

术语“聚类”是指将物理或抽象对象的集合分成由类似的对象组成的多个类的过程，例如将 reads 分为两类，分别属于两条 SNP 连锁。对本领域技术人员而言，可以通过简单的程序实现聚类。

15

#### 发明的有益效果

1) 本发明用生物信息的方法，找出样本序列与特定参考序列的 SNP 位点，进一步确定出基因的 SNP 连锁关系，与先准备好的型别-SNP 连锁关系数据库比较，可以快速、精确地对 HLA 分型。

20        由于 HLA 基因序列的一个特点是有高度的同源性，不同型别序列之间的差异度很小，有些只有几个碱基差异，通过 SNP 连锁关系确定型别准确而且简洁。

25        2) 相比于传统方法，本发明大大的提高了分型的准确性。本发明方法能够准确的将测序得到的短片段序列信息转变成完整的 HLA 基因的完整序列信息，如果此基因是杂合子，则能准确的区分出 HLA 基因的两个不同单倍体型别。

3) 本发明能同时处理大量的样本，耗时很短，经实验测序后，用该方法可以直接得到 HLA 基因的型别，不需要人工分析判断，减少了人为操作造成的错误，并实现了自动化分型。

30        总之，本发明的分型方法实现了 HLA 的低成本，高通量、高准确率、高分辨率的分型。

## 附图说明

Fig. 1: 表示由大量两 SNP 位点的连锁关系整合出来的一条完整的 SNP 连锁关系。图中圆点表示 SNP 位点, 通过首先确定两位点的连锁关系, 然后将所有位点连接起来, 得到最终的 SNP 连锁关系, 每个样本最多能确定两条

5 SNP 连锁关系。

Fig. 2: 基于 SNP 连锁关系的 HLA 基因高分辨率分型方法的流程图。

Fig. 3: 过滤并去除假阳性的流程示意图。

## 具体实施方式

下面将结合实施例对本发明的实施方案进行详细描述。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者, 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。

10

15

### 实施例 1: 94 个样本的 HLA 分型

#### 1. 实验目的:

用 SOLEXA 测序技术对获自 94 人的 94 个样本的 HLA-A/B/C 的 2, 3, 4 号外显子、HLA-DRB1 的 2 号外显子、HLA-DQB1 的 2, 3 号外显子分别进行了测序, 通过这些外显子的序列信息, 对 HLA-A/B/C/DRB1/DQB1 进行 HLA 分型。

20

#### 2. 实验方法:

##### (1) 构建 HLA 基因型别-SNP 连锁数据库构

按照本发明的方法构建 HLA 基因型别-SNP 连锁 (关系) 数据库构, 其中, 所用参考序列为 SEQ ID NO: 1-5, 现有数据库为 EBI 数据库 (EBI: <http://www.ebi.ac.uk/imgt/hla/>)。

25

##### (2) 找出 SNP 位点

在这里用 BWA 比对软件和 samtools 软件为例来说明。通过 BWA 比对软件 (Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. Vol. 25 no. 14 2009, pages 1754-1760), 将这些序列与特定的参考序列进行序列比对, 经过 aln 和

30

5 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, 优选地, 可以用 samtools 软件工具的 view 将文件转化为\*.bam 格式, 分别经过 sort、index、pileup 步骤后, 得到含有每个位点比对信息的一致性序列\*.consence 文件, 该文件包括位点坐标、参考序列的碱基、样本序

列的碱基、碱基质量等信息。  
在比对后处理的结果中提取 SNP 的位点信息, 每个 EXON 可以得到一个 SNP 位点的文件, 如表 3 所示。

表 3: 样本外显子的 SNP 位点信息

| 位点坐标 | 样本在位点上的碱基简写 | 样本在位点上的碱基 |
|------|-------------|-----------|
| 2215 | Y           | C/T       |
| 2216 | W           | A/T       |
| 2228 | S           | C/G       |
| 2247 | Y           | C/T       |
| 2253 | G           | G/G       |
| 2266 | Y           | C/T       |
| 2267 | W           | A/T       |
| 2273 | C           | C/C       |
| 2278 | T           | T/T       |
| 2279 | A           | A/A       |
| 2295 | G           | G/G       |
| 2299 | K           | G/T       |
| 2300 | A           | A/A       |
| 2303 | Y           | C/T       |
| 2304 | R           | A/G       |
| 2327 | T           | T/T       |
| 2330 | A           | A/A       |
| 2331 | T           | T/T       |

### (3) SNP 初步连锁关系确定

10 这一步是本发明非常关键的步骤。对于上一步统计出的 SNP 位点信息, 首先确定两个突变位点的连锁关系, 每两个位点的连锁关系最多只有两种情况, 然后通过序列支持数等信息将各 SNP 位点一一连接起来。在确定两个 SNP 位点的连锁时, 先通过逻辑关系分组, 如两个 SNP 位点的碱基分别为 A/C 和 G/T, 可以得到 A-G、A-T、C-G、C-T 四种关系, 但因为在第一个 SNP 位点有 A 和 C 出现, 那么一定有两种 SNP 连锁关系, 在该位点上, 一种连锁含 A, 另一种含 C, 后一个位点的情况类似, 所以可以分成 A-G、C-T 和 A-T、C-G 两类, 这两类是不可能同时出现的, 只有其中一种情况存在。

然后结合 BWA 比对出的 \*.sam 文件信息, 根据两个位点分别拥有的序列支持数和同时通过两个位点的支持数, 确定出最多两种连锁关系。如表 4 为其中一个示例:

表 4: 确定 SNP 连锁关系的示例

| 编号 | SNP 位点                                                          |
|----|-----------------------------------------------------------------|
| 1  | 5433: A-5453: C-5497: A-5529: C-5579: G-5593: A-5624: A-5633: T |
| 2  | 5433: G-5453: G-5497: G-5529: C-5579: A-5593: A-5624: G-5633: C |

#### 5 (4) 聚类

从上一步中确定两条 SNP 连锁关系, 根据连锁关系对 reads 进行聚类, reads 被分成两大部分, 根据各部分 reads 的比对结果用 samtools 的 pileup 分别构建新的一致性序列, 然后挑出 SNP 并构成新的 SNP 连锁关系。经过这一步聚类, SNP 和 SNP 的连锁关系都得到了进一步的校正。如表 5 为其中的一个示例:

表 5: 通过聚类确定 SNP 连锁关系的示例

| 编号 | SNP 位点                                                                       |
|----|------------------------------------------------------------------------------|
| 1  | 528: A-532: T-633: G-687: G-689: A-691: C-701: G-712: G-729: T-731: G-771: C |
| 2  | 633: G-701: G-712: G-729: T-731: G-771: C                                    |

#### 10 (5) 型别最终确定

将上一步得到的 SNP 连锁关系与 HLA 基因型别-SNP 连锁关系库比较, 得到对应的型别, 然后提取出基因中多个外显子共有的型别。如表 6-1 为某样本 B 基因的 2、3、4 号外显子对应的型别结果, 从表中可以得出, 该样本 B 基因的型别为 B\*35: 43、B\*35: 67、B\*46: 01: 01、B\*46: 13: 01。

表 6-1: 示例性样本 B 基因的 2、3、4 号外显子对应的型别结果

| EXON2 型别     | EXON3 型别     | EXON4 型别         |
|--------------|--------------|------------------|
| B*35: 43     | B*35: 43     | B*35: 43         |
| B*78: 02: 01 | B*46: 13: 01 | B*15: 05: 01     |
| B*35: 67     | B*15: 01: 04 | B*46: 01: 01     |
| B*46: 01: 01 | B*46: 01: 01 | B*46: 13: 01     |
| B*35: 68: 01 | B*15: 175    | B*15: 17: 01: 01 |
| B*46: 13: 01 | B*15: 125    | B*35: 67         |
|              | B*35: 67     | B*15: 25: 01     |

最后对得到的型别结果还需根据逻辑关系进一步过滤。

首先用图表说明逻辑关系分析去假阳性, 如表 6-2 为上例中某样本 HLA-B 得到的初步结果, 有 B\*35: 43、B\*35: 67、B\*46: 01: 01、B\*46: 13: 01 四种结果, 表中后三列为型别结果在 2, 3, 4 号外显子分别对应的 SNP 连锁

情况，可以看出，每个外显子都有两种 SNP 连锁关系，2 号外显子有 2 种 SNP 连锁关系（e21 和 e22），3 号外显子为 e31 和 e32，4 号外显子为 e41 和 e42，3 个外显子的 SNP 连锁关系会发生自由组合情况，得到假阳性的结果，具体如 Fig. 3 所示。因为人为二倍体，若外显子有两种 SNP 连锁关系，那么这两种 SNP 连锁都应该是最终 HLA 型别对应的 SNP 连锁组成部分，都应该包括进去，每条连锁关系只能对应一种型别。基于这个原理，从 Fig. 3 看出，针对 2 号外显子来说，型别 B\*35: 67 和 B\*35: 43 是不能同时存在的，B\*46: 13: 01 和 B\*46: 01: 01 也是不能同时存在，3 号、4 号外显子情况类似。通过这样的逻辑分析，同时符合 2、3、4 号外显子要求的只有 B\*35: 67 和 B\*46: 13: 01，B\*35: 43 和 B\*46: 01: 01 是假阳性型别。若外显子只有一条 SNP 连锁关系，那么情况会更简单些，分析方法和上面的类似。

表 6-2: 示例性样本 HLA-B 得到的初步结果

| 型别           | EXON2 的 SNP 连锁 | EXON3 的 SNP 连锁 | EXON4 的 SNP 连锁 |
|--------------|----------------|----------------|----------------|
| B*35: 43     | e21            | e31            | e42            |
| B*35: 67     | e21            | e31            | e41            |
| B*46: 01: 01 | e22            | e31            | e41            |
| B*46: 13: 01 | e22            | e32            | e42            |

如果通过逻辑关系分析后结果多于 2 种型别，可以参考 HLA 罕见型列表 (<http://bioinformatics.nmdp.org/HLA/Biannual-Rare-Allele-List/Version-2-28-0/index.htm>)，最终给出两种最有可能的型别。

在具体实施例中发现，有流程判断出的型别多于两种的只是少数，逻辑关系分析方法几乎过滤了所有的假阳性。

#### (6) 结果评价

对给定的结果作一个评价，根据条件判断，如果符合条件，则认为是给定的结果是正确的，用 TRUE 标记，否则认为给定的结果是错误的，用 FALSE 标记。如表 7 为其中的一个实例：

表 7: 结果评价示例

| 样本编号 | A 基因型别   | A 基因型别   | 准确度判断 | B 基因型别   | B 基因型别   | 准确度判断 | 罕见型标记 |
|------|----------|----------|-------|----------|----------|-------|-------|
| D001 | A*11: 01 | A*11: 01 | TRUE  | B*39: 39 | B*39: 39 | FALSE | r-r   |
| D003 | A*03: 02 | A*68: 01 | TRUE  | B*44: 06 | B*51: 01 | FALSE |       |

### 3. 结果及分析:

在本发明的具体实施例中，通过分析 HLA-A、HLA-B、HLA-C 的 2、3、

4 号外显子, HLA-DRB1 的 2 号外显子, HLA-DQB1 的 2 和 3 号外显子, 对 94 个人样本 HLA-A/B/C/DRB1/DQB1 进行 HLA 分型, 首先确定出基因中各外显子对应的型别, 然后整合多个外显子所属的型别类型, 提取出各外显子共有的型别 (其中 30 个样本的具体结果表 9-1 至表 9-5 所示)。在 94 个样本中, 有 90% 的分型结果与基于 sanger 法测序分型的结果完全相符。其中 94 个样本的基于 sanger 法测序分型的具体方法可以参考 Adams SD, Barracchini KC, Simonis TB, et al. High throughput HLA sequence-based typing (SBT) utilizing the ABI Prism 3700 DNA Analyzer. Tumori, 2001;87:S40-43, 其中 30 个样本的具体结果表 8-1 和 8-2 所示。

表 8-1: 基于 sanger 法测序分型的 HLA-A/B/C 型别 (部分)

| 样本编号 | A 基因<br>型别 | A 基因<br>型别 | B 基因<br>型别 | B 基因<br>型别 | C 基因<br>型别 | C 基因<br>型别 |
|------|------------|------------|------------|------------|------------|------------|
| D076 | A*11: 01   | A*33: 03   | B*13: 01   | B*15: 01   | C*03: 04   | C*04: 01   |
| D083 | A*11: 01   | A*31: 01   | B*15: 02   | B*46: 01   | C*01: 02   | C*08: 01   |
| D085 | A*11: 53   | A*30: 01   | B*13: 02   | B*15: 32   | C*06: 02   | C*12: 03   |
| D086 | A*02: 07   | A*33: 03   | B*46: 01   | B*58: 01   | C*01: 02   | C*03: 02   |
| D087 | A*02: 01   | A*30: 01   | B*13: 02   | B*67: 01   | C*06: 02   | C*07: 02   |
| D089 | A*11: 01   | A*11: 01   | B*07: 02   | B*15: 58   | C*01: 02   | C*07: 02   |
| D090 | A*01: 01   | A*32: 01   | B*40: 01   | B*57: 01   | C*06: 02   | C*07: 02   |
| D091 | A*24: 02   | A*30: 01   | B*35: 01   | B*52: 01   | C*03: 03   | C*12: 02   |
| D092 | A*11: 01   | A*24: 02   | B*15: 02   | B*40: 01   | C*08: 01   | C*15: 02   |
| D093 | A*01: 01   | A*11: 01   | B*08: 01   | B*37: 01   | C*06: 02   | C*07: 02   |
| D095 | A*11: 01   | A*24: 02   | B*35: 02   | B*58: 01   | C*03: 02   | C*04: 01   |
| D099 | A*11: 01   | A*33: 03   | B*15: 02   | B*58: 01   | C*03: 02   | C*08: 01   |
| D100 | A*02: 01   | A*24: 02   | B*13: 01   | B*51: 01   | C*03: 04   | C*14: 02   |
| D101 | A*02: 07   | A*11: 01   | B*39: 01   | B*46: 01   | C*01: 02   | C*07: 02   |
| D103 | A*02: 10   | A*24: 02   | B*35: 01   | B*40: 06   | C*03: 03   | C*08: 01   |
| D106 | A*03: 01   | A*32: 01   | B*44: 02   | B*50: 01   | C*05: 01   | C*06: 02   |
| D110 | A*11: 01   | A*11: 53   | B*15: 02   | B*27: 04   | C*08: 01   | C*12: 02   |
| D111 | A*11: 01   | A*33: 03   | B*15: 02   | B*58: 01   | C*03: 02   | C*08: 01   |
| D113 | A*01: 01   | A*02: 07   | B*07: 02   | B*46: 01   | C*01: 02   | C*07: 02   |
| D115 | A*11: 53   | A*24: 02   | B*40: 01   | B*40: 02   | C*03: 04   | C*15: 02   |
| D116 | A*11: 01   | A*31: 01   | B*35: 01   | B*51: 01   | C*03: 03   | C*15: 02   |
| D117 | A*01: 01   | A*02: 01   | B*13: 02   | B*15: 11   | C*03: 03   | C*06: 02   |
| D118 | A*24: 02   | A*24: 02   | B*07: 02   | B*35: 02   | C*04: 01   | C*07: 02   |
| D119 | A*24: 02   | A*30: 01   | B*13: 02   | B*46: 01   | C*01: 02   | C*06: 02   |

|      |          |          |          |          |          |          |
|------|----------|----------|----------|----------|----------|----------|
| D120 | A*11: 01 | A*30: 04 | B*14: 01 | B*40: 01 | C*03: 04 | C*08: 02 |
| D123 | A*24: 02 | A*30: 01 | B*13: 02 | B*40: 06 | C*06: 02 | C*08: 01 |
| D124 | A*03: 01 | A*11: 01 | B*40: 01 | B*58: 01 | C*03: 04 | C*07: 02 |
| D125 | A*03: 01 | A*24: 02 | B*44: 02 | B*51: 01 | C*05: 01 | C*14: 02 |
| D129 | A*02: 03 | A*31: 01 | B*07: 02 | B*55: 02 | C*01: 02 | C*07: 02 |
| D130 | A*11: 01 | A*30: 01 | B*13: 02 | B*35: 03 | C*04: 01 | C*06: 02 |

表 8-2: 基于 sanger 法测序分型的 HLA-DRB1/DQB1 型别 (部分)

| 样本编号 | DRB1 型别     | DRB1 型别     | DQB1 型别     | DQB1 型别     |
|------|-------------|-------------|-------------|-------------|
| D076 | DRB1*04: 06 | DRB1*11: 01 | DQB1*03: 01 | DQB1*03: 02 |
| D083 | DRB1*09: 01 | DRB1*12: 02 | DQB1*03: 01 | DQB1*03: 03 |
| D085 | DRB1*07: 01 | DRB1*13: 02 | DQB1*02: 02 | DQB1*06: 09 |
| D086 | DRB1*12: 01 | DRB1*13: 02 | DQB1*03: 01 | DQB1*06: 09 |
| D087 | DRB1*07: 01 | DRB1*16: 02 | DQB1*05: 02 | DQB1*02: 02 |
| D089 | DRB1*01: 01 | DRB1*11: 01 | DQB1*05: 01 | DQB1*03: 01 |
| D090 | DRB1*07: 01 | DRB1*15: 01 | DQB1*03: 03 | DQB1*06: 02 |
| D091 | DRB1*11: 01 | DRB1*15: 01 | DQB1*03: 01 | DQB1*06: 02 |
| D092 | DRB1*04: 06 | DRB1*09: 01 | DQB1*03: 02 | DQB1*03: 03 |
| D093 | DRB1*03: 01 | DRB1*10: 01 | DQB1*05: 01 | DQB1*02: 01 |
| D095 | DRB1*03: 01 | DRB1*11: 04 | DQB1*02: 01 | DQB1*03: 01 |
| D099 | DRB1*03: 01 | DRB1*15: 01 | DQB1*02: 01 | DQB1*06: 01 |
| D100 | DRB1*09: 01 | DRB1*11: 01 | DQB1*03: 01 | DQB1*03: 03 |
| D101 | DRB1*08: 03 | DRB1*09: 01 | DQB1*03: 03 | DQB1*06: 01 |
| D103 | DRB1*07: 01 | DRB1*11: 01 | DQB1*03: 01 | DQB1*03: 03 |
| D106 | DRB1*03: 01 | DRB1*11: 01 | DQB1*02: 01 | DQB1*03: 01 |
| D110 | DRB1*12: 02 | DRB1*16: 02 | DQB1*05: 02 | DQB1*03: 01 |
| D111 | DRB1*03: 01 | DRB1*15: 01 | DQB1*02: 01 | DQB1*06: 01 |
| D113 | DRB1*04: 04 | DRB1*09: 01 | DQB1*03: 02 | DQB1*03: 03 |
| D115 | DRB1*04: 03 | DRB1*15: 01 | DQB1*03: 02 | DQB1*06: 02 |
| D116 | DRB1*09: 01 | DRB1*15: 01 | DQB1*03: 03 | DQB1*06: 02 |
| D117 | DRB1*07: 01 | DRB1*09: 01 | DQB1*02: 02 | DQB1*03: 03 |
| D118 | DRB1*11: 04 | DRB1*15: 01 | DQB1*03: 01 | DQB1*06: 02 |
| D119 | DRB1*07: 01 | DRB1*09: 01 | DQB1*02: 02 | DQB1*03: 03 |
| D120 | DRB1*11: 01 | DRB1*15: 02 | DQB1*03: 01 | DQB1*06: 01 |
| D123 | DRB1*07: 01 | DRB1*09: 01 | DQB1*02: 02 | DQB1*03: 03 |
| D124 | DRB1*04: 05 | DRB1*13: 02 | DQB1*04: 01 | DQB1*06: 09 |
| D125 | DRB1*13: 01 | DRB1*14: 05 | DQB1*05: 03 | DQB1*06: 03 |
| D129 | DRB1*01: 01 | DRB1*12: 02 | DQB1*05: 01 | DQB1*03: 01 |
| D130 | DRB1*07: 01 | DRB1*11: 01 | DQB1*02: 02 | DQB1*03: 01 |

表 9-1: 本发明测得的 HLA-A 型别

| 样本编号 | A 基因型别            | A 基因型别            | 准确度判断 | 罕见型标记  |
|------|-------------------|-------------------|-------|--------|
| D076 | A*11: 01          | A*33: 03          | TRUE  |        |
| D083 | A*11: 01          | A*31: 01          | TRUE  |        |
| D085 | A*11: 02/A*11: 53 | A*30: 01          | TRUE  | 2: R - |
| D086 | A*02: 07          | A*33: 03          | TRUE  |        |
| D087 | A*02: 01          | A*30: 01          | TRUE  |        |
| D089 | A*11: 01          | A*11: 01          | TRUE  |        |
| D090 | A*01: 01          | A*32: 01          | TRUE  |        |
| D091 | A*24: 02          | A*30: 01          | TRUE  |        |
| D092 | A*11: 01          | A*24: 02          | TRUE  |        |
| D093 | A*01: 01          | A*11: 01          | TRUE  |        |
| D095 |                   |                   |       |        |
| D099 | A*11: 01          | A*11: 01          | FALSE |        |
| D100 | A*02: 01          | A*24: 02          | TRUE  |        |
| D101 | A*02: 07          | A*11: 01          | TRUE  |        |
| D103 | A*02: 10          | A*24: 02          | TRUE  |        |
| D106 |                   |                   |       |        |
| D110 | A*11: 01          | A*11: 02/A*11: 53 | TRUE  | -2: R  |
| D111 | A*11: 01          | A*33: 03          | TRUE  |        |
| D113 | A*01: 01          | A*02: 07          | TRUE  |        |
| D115 | A*11: 02/A*11: 53 | A*24: 02          | TRUE  | 2: R - |
| D116 | A*11: 01          | A*31: 01          | TRUE  |        |
| D117 | A*01: 01          | A*02: 01          | TRUE  |        |
| D118 | A*24: 02          | A*24: 02          | TRUE  |        |
| D119 | A*24: 02          | A*30: 01          | TRUE  |        |
| D120 | A*11: 01          | A*30: 04          | TRUE  |        |
| D123 | A*24: 02          | A*30: 01          | TRUE  |        |
| D124 | A*03: 01          | A*11: 01          | TRUE  |        |
| D125 | A*03: 01          | A*24: 02          | TRUE  |        |
| D129 | A*02: 03          | A*31: 01          | TRUE  |        |
| D130 | A*11: 01          | A*30: 01          | TRUE  |        |

表 9-2: 本发明测得的 HLA-B 型别

| 样本编号 | B 基因型别 | B 基因型别 | 准确度 | 罕见型 |
|------|--------|--------|-----|-----|
|------|--------|--------|-----|-----|

|      |                   |          | 判断    | 标记     |
|------|-------------------|----------|-------|--------|
| D076 | B*13: 01          | B*15: 01 | TRUE  |        |
| D083 | B*15: 02          | B*46: 01 | TRUE  |        |
| D085 | B*13: 02          | B*15: 32 | TRUE  |        |
| D086 | B*46: 01          | B*58: 01 | TRUE  |        |
| D087 | B*13: 02          | B*67: 01 | TRUE  |        |
| D089 | B*07: 02/B*07: 61 | B*15: 58 | TRUE  | 2: R - |
| D090 | B*40: 01          | B*57: 01 | TRUE  |        |
| D091 | B*35: 01/B*35: 42 | B*52: 01 | TRUE  | 2: R - |
| D092 | B*15: 02          | B*40: 01 | TRUE  |        |
| D093 | B*08: 01          | B*37: 01 | TRUE  |        |
| D095 | B*35: 02          | B*58: 01 | TRUE  |        |
| D099 | B*15: 02          | B*58: 01 | TRUE  |        |
| D100 |                   |          |       |        |
| D101 | B*39: 01          | B*46: 01 | TRUE  |        |
| D103 | B*35: 01/B*35: 42 | B*40: 06 | TRUE  | 2: R - |
| D106 | B*50: 01          | B*50: 01 | FALSE |        |
| D110 | B*15: 02          | B*27: 04 | TRUE  |        |
| D111 | B*15: 02          | B*58: 01 | TRUE  |        |
| D113 | B*07: 02/B*07: 61 | B*46: 01 | TRUE  | 2: R - |
| D115 | B*40: 01          | B*40: 02 | TRUE  |        |
| D116 | B*35: 01/B*35: 42 | B*51: 01 | TRUE  | 2: R - |
| D117 | B*13: 02          | B*15: 11 | TRUE  |        |
| D118 | B*07: 02/B*07: 61 | B*35: 02 | TRUE  | 2: R - |
| D119 | B*13: 02          | B*46: 01 | TRUE  |        |
| D120 | B*14: 01          | B*40: 01 | TRUE  |        |
| D123 | B*13: 02          | B*40: 06 | TRUE  |        |
| D124 | B*40: 01          | B*58: 01 | TRUE  |        |
| D125 |                   |          |       |        |
| D129 | B*07: 02/B*07: 61 | B*55: 02 | TRUE  | 2: R - |
| D130 | B*13: 02          | B*35: 03 | TRUE  |        |

表 9-3: 本发明测得的 HLA-C 型别

| 样本编号 | C 基因型别   | C 基因型别            | 准确型判断 | 罕见型标记 |
|------|----------|-------------------|-------|-------|
| D076 | C*03: 04 | C*04: 01/C*04: 30 | TRUE  | -2: R |
| D083 | C*01: 02 | C*01: 02          | FALSE |       |
| D085 | C*06: 02 | C*12: 03          | TRUE  |       |
| D086 | C*01: 02 | C*03: 02          | TRUE  |       |

|      |                   |                   |       |               |
|------|-------------------|-------------------|-------|---------------|
| D087 | C*06: 02          | C*07: 02/C*07: 50 | TRUE  | -2: R         |
| D089 | C*01: 02          | C*07: 02/C*07: 50 | TRUE  | -2: R         |
| D090 | C*06: 02          | C*07: 02/C*07: 50 | TRUE  | -2: R         |
| D091 |                   |                   |       |               |
| D092 | C*08: 01/C*08: 22 | C*15: 02          | TRUE  | 2: r -        |
| D093 | C*06: 02          | C*07: 02/C*07: 50 | TRUE  | -2: R         |
| D095 | C*03: 02          | C*04: 01/C*04: 30 | TRUE  | -2: R         |
| D099 | C*03: 02          | C*03: 02          | FALSE |               |
| D100 | C*03: 04          | C*14: 02          | TRUE  |               |
| D101 | C*01: 02          | C*07: 02/C*07: 50 | TRUE  | -2: R         |
| D103 | C*03: 03          | C*08: 01/C*08: 22 | TRUE  | -2: r         |
| D106 | C*05: 01          | C*06: 02          | TRUE  |               |
| D110 | C*08: 01/C*08: 22 | C*12: 02          | TRUE  | 2: r -        |
| D111 | C*03: 02          | C*08: 01/C*08: 22 | TRUE  | -2: r         |
| D113 | C*01: 02          | C*07: 02/C*07: 50 | TRUE  | -2: R         |
| D115 | C*03: 04          | C*15: 02          | TRUE  |               |
| D116 | C*03: 03          | C*15: 02          | TRUE  |               |
| D117 | C*03: 03          | C*06: 02          | TRUE  |               |
| D118 | C*04: 01/C*04: 30 | C*07: 02/C*07: 50 | TRUE  | 2: R<br>-2: R |
| D119 | C*01: 02          | C*06: 02          | TRUE  |               |
| D120 | C*03: 04          | C*08: 02          | TRUE  |               |
| D123 | C*06: 02          | C*08: 01/C*08: 22 | TRUE  | -2: r         |
| D124 | C*03: 04          | C*07: 02/C*07: 50 | TRUE  | -2: R         |
| D125 | C*05: 01          | C*14: 02          | TRUE  |               |
| D129 | C*01: 02          | C*07: 02/C*07: 50 | TRUE  | -2: R         |
| D130 | C*04: 01/C*04: 30 | C*06: 02          | TRUE  | 2: R -        |

表 9-4: 本发明测得的 HLA- DRB1 型别

| 样本编号 | DRB1 型别                                      | DRB1 型别     | 准确性判断 | 罕见型标记                  |
|------|----------------------------------------------|-------------|-------|------------------------|
| D076 | DRB1*04: 06/<br>DRB1*04: 49                  | DRB1*11: 01 | TRUE  | 2: R -                 |
| D083 | DRB1*09: 01                                  | DRB1*12: 02 | TRUE  |                        |
| D085 | DRB1*07: 01                                  | DRB1*13: 02 | TRUE  |                        |
| D086 | DRB1*12: 01/<br>DRB1*12: 06/<br>DRB1*12: 10/ | DRB1*13: 02 | TRUE  | 4: R<br>3: r<br>2: r - |

|      |                         |                             |      |        |
|------|-------------------------|-----------------------------|------|--------|
|      | DRB1*12: 17             |                             |      |        |
| D087 | DRB1*07: 01             | DRB1*16: 02                 | TRUE |        |
| D089 |                         |                             |      |        |
| D090 | DRB1*07: 01             | DRB1*15: 01                 | TRUE |        |
| D091 | DRB1*11: 01             | DRB1*15: 01                 | TRUE |        |
| D092 | DRB1*04: 06/DRB1*04: 49 | DRB1*09: 01                 | TRUE | 2: R - |
| D093 | DRB1*03: 01/DRB1*03: 50 | DRB1*10: 01                 | TRUE | 2: R - |
| D095 | DRB1*03: 01/DRB1*03: 50 | DRB1*11: 04                 | TRUE | 2: R - |
| D099 | DRB1*03: 01/DRB1*03: 50 | DRB1*15: 01                 | TRUE | 2: R - |
| D100 | DRB1*09: 01             | DRB1*11: 01                 | TRUE |        |
| D101 | DRB1*08: 03             | DRB1*09: 01                 | TRUE |        |
| D103 | DRB1*07: 01             | DRB1*11: 01                 | TRUE |        |
| D106 | DRB1*03: 01/DRB1*03: 50 | DRB1*11: 01                 | TRUE | 2: R - |
| D110 | DRB1*12: 02             | DRB1*16: 02                 | TRUE |        |
| D111 | DRB1*03: 01/DRB1*03: 50 | DRB1*15: 01                 | TRUE | 2: R - |
| D113 | DRB1*04: 04             | DRB1*09: 01                 | TRUE |        |
| D115 | DRB1*04: 03             | DRB1*15: 01                 | TRUE |        |
| D116 | DRB1*09: 01             | DRB1*15: 01                 | TRUE |        |
| D117 | DRB1*07: 01             | DRB1*09: 01                 | TRUE |        |
| D118 | DRB1*11: 04             | DRB1*15: 01                 | TRUE |        |
| D119 | DRB1*07: 01             | DRB1*09: 01                 | TRUE |        |
| D120 | DRB1*11: 01             | DRB1*15: 02/<br>DRB1*15: 19 | TRUE | -2: R  |
| D123 | DRB1*07: 01             | DRB1*09: 01                 | TRUE |        |
| D124 | DRB1*04: 05             | DRB1*13: 02                 | TRUE |        |
| D125 | DRB1*13: 01             | DRB1*14: 05                 | TRUE |        |
| D129 | DRB1*01: 01             | DRB1*12: 02                 | TRUE |        |
| D130 | DRB1*07: 01             | DRB1*11: 01                 | TRUE |        |

表 9-5: 本发明测得的 HLA- DQB1 型别

| 样本编号 | DQB1 型别     | DRB1 型别     | 准确性判断 |
|------|-------------|-------------|-------|
| D076 | DQB1*03: 01 | DQB1*03: 02 | TRUE  |
| D083 | DQB1*03: 01 | DQB1*03: 03 | TRUE  |
| D085 | DQB1*02: 02 | DQB1*06: 09 | TRUE  |
| D086 | DQB1*03: 01 | DQB1*06: 09 | TRUE  |
| D087 | DQB1*02: 02 | DQB1*05: 02 | TRUE  |
| D089 | DQB1*03: 01 | DQB1*05: 01 | TRUE  |
| D090 | DQB1*03: 03 | DQB1*06: 02 | TRUE  |
| D091 | DQB1*03: 01 | DQB1*06: 02 | TRUE  |

|      |             |             |      |
|------|-------------|-------------|------|
| D092 | DQB1*03: 02 | DQB1*03: 03 | TRUE |
| D093 | DQB1*02: 01 | DQB1*05: 01 | TRUE |
| D095 | DQB1*02: 01 | DQB1*03: 01 | TRUE |
| D099 | DQB1*02: 01 | DQB1*06: 01 | TRUE |
| D100 | DQB1*03: 01 | DQB1*03: 03 | TRUE |
| D101 | DQB1*03: 03 | DQB1*06: 01 | TRUE |
| D103 | DQB1*03: 01 | DQB1*03: 03 | TRUE |
| D106 | DQB1*02: 01 | DQB1*03: 01 | TRUE |
| D110 | DQB1*03: 01 | DQB1*05: 02 | TRUE |
| D111 | DQB1*02: 01 | DQB1*06: 01 | TRUE |
| D113 | DQB1*03: 02 | DQB1*03: 03 | TRUE |
| D115 | DQB1*03: 02 | DQB1*06: 02 | TRUE |
| D116 | DQB1*03: 03 | DQB1*06: 02 | TRUE |
| D117 | DQB1*02: 02 | DQB1*03: 03 | TRUE |
| D118 | DQB1*03: 01 | DQB1*06: 02 | TRUE |
| D119 | -           | -           | -    |
| D120 | DQB1*03: 01 | DQB1*06: 01 | TRUE |
| D123 | DQB1*02: 02 | DQB1*03: 03 | TRUE |
| D124 | DQB1*04: 01 | DQB1*06: 09 | TRUE |
| D125 | DQB1*05: 03 | DQB1*06: 03 | TRUE |
| D129 | DQB1*03: 01 | DQB1*05: 01 | TRUE |
| D130 | DQB1*02: 02 | DQB1*03: 01 | TRUE |

\*注释：在所示的结果列表中，“/”表示前后的型别序列在我们研究的外显子部分是完全一样，所以根据这些外显子序列信息无法将这些分开，如 C\*08: 01/C\*08: 22，表示 C\*08: 01 和 C\*08: 22 在 2、3 和 4 号外显子上的序列完全一样，无法区分。TRUE 表示对结果的评价是正确的，FALSE 是对结果评价是错误的，R 和 r 表示罕见型。

从上面随机给出的 30 个样本的结果看，没有给出结果和判断为 FALSE 的 HLA-A 有 3 个，HLA-B 有 3 个，HLA-C 有 3 个，HLA-DRB1 有 1 个，HLA-DQB1 有 1 个，判断为 TRUE 的都是正确的结果，没有出现假阳性。在正确的结果中还有一小部分的“/”，即用本方法是无法区分的。从这个结果看，本发

采用本发明的分型方法，对 94 份已知分型结果的样本进行了 HLA-A/B/C/DRB1/DQB1 位点的基因分型。如果本发明方法判断为 TRUE，并且真实结果是正确的，则定义为一次性分型成功；如果本发明方法判断为 TRUE，并且真实结果是错误的，则定义为假阳性；如果本发明方法判断为

FALSE 或未给出结果的, 则定义为假阴性。通过结果发现: 在各样本的分型中, HLA-A 一次分型成功率达到 93.6%, 假阳性为 0, 假阴性为 6.4%; HLA-B 一次分型成功率达到 86.14%, 假阳性为 1.06%, 假阴性为 12.8%; HLA-C 一次分型成功率达到 85.14%, 假阳性为 1.06%, 假阴性为 13.8%; HLA-DRB1 一次分型成功率达到 95.7%, 假阳性为 0, 假阴性为 4.3%; HLA-DQB1 一次分型成功率达到 98.9%, 假阳性为 0, 假阴性为 1.1%; 但在这些正确的结果中包括了 “/”, 即在研究的外显子区域是无法区分的, 总体来说, 这样的型别结果占 10%左右。从本实施例的分型结果看, 说明本发明的分型方法对 HLA 分型准确度非常高。

10

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。

15

## 权利要求

1. 一种构建 HLA 基因型别-SNP 连锁数据库的方法, 包括如下步骤:
  - a) 选择一个或多个 HLA 基因座的序列作为参考序列;
  - 5 b) 将现有 HLA 数据库中已知型别的 HLA 基因与参考序列比对, 找出与参考序列的差异位点即 SNP 位点, 得到每个型别相对于参考序列的 SNP 连锁关系, 构建 HLA 基因型别-SNP 连锁数据库。
2. 根据权利要求 1 所述的方法, 其中, 步骤 a) 中选取 5 种非罕见型 HLA 基因座的序列, 并且每个序列为已知的完整序列; 具体地, 选取  
10 SEQ ID NO: 1-5 作为参考序列。
3. 根据权利要求 1 所述的方法, 其中, 步骤 b) 中所述现有 HLA 数据库为 EBI 的 HLA 数据库。
4. 一种 HLA 基因型别-SNP 连锁数据库, 其根据权利要求 1-3 中任一项所述的方法构建。
- 15 5. 一种确定 HLA 基因的 SNP 连锁关系的方法, 包括如下步骤:
  - 1) 找出 SNP 位点: 将样本序列与权利要求 1 中所述的参考序列进行比对, 统计分析, 得到样本序列与参考序列的差异位点, 即 SNP 位点;
  - 2) 初步确定 SNP 连锁关系: 先根据经过 SNP 位点的 read 数目, 确定出两个 SNP 位点的连锁关系, 根据经过 SNP 位点的序列支持数, 将两  
20 两 SNP 连锁关系连接起来, 不断延伸, 直到确定出整条 SNP 连锁关系; 如果相邻两 SNP 位点没有连接起来, 使这两个 SNP 位点前后两部分的 SNP 连锁关系在此无法连接, 则对这两个 SNP 位点前后两部分的 SNP 连锁关系自由组合连接; 最多保留两种连锁关系。
6. 根据权利要求 5 所述的方法, 其中, 还包括步骤 3): 通过聚类  
25 进一步确定 SNP 连锁关系: 从步骤 2) 中选择两条 SNP 连锁关系, 然后对 reads 聚类, 通过聚类将 reads 分为两类, 分别属于两条 SNP 连锁。
7. 一种 HLA 分型方法, 包括权利要求 5 或 6 所述的确定 SNP 连锁关

系的方法的步骤，并且还包括如下步骤 4)：

4) 型别确定：将得到的 SNP 连锁关系与权利要求 4 所述的型别-SNP 连锁数据库比较，得到与 SNP 连锁关系对应的 HLA 基因型别；

优选地，还包括如下步骤 5)：

- 5 5) 结果过滤：每个基因判断出来的型别结果进一步进行过滤，去除假阳性；具体地，对预测的型别结果进行逻辑关系分析，以去掉自相矛盾的结果；

优选地，还包括如下的步骤 6)：

- 10 6) 结果评价：对得到的结果进行评价，根据一定的条件判断结果是正确的还是错误的；具体地，需要综合考虑多个外显子的，只有在多个外显子同时满足要求的条件，才认为是正确的。

8. 一种 HLA 分型装置，包括如下单元：

1) SNP 连锁关系确定单元；

2) 型别确定单元。

- 15 9. 根据权利要求 8 所述的装置，其中，所述 SNP 连锁关系确定单元包括：找出 SNP 位点的单元、初步确定 SNP 连锁关系的单元；优选地，还包括通过聚类进一步确定 SNP 的连锁关系的单元。

10. 根据权利要求 8 所述的装置，还包括结果过滤单元；并且优选地，还包括结果评价单元。

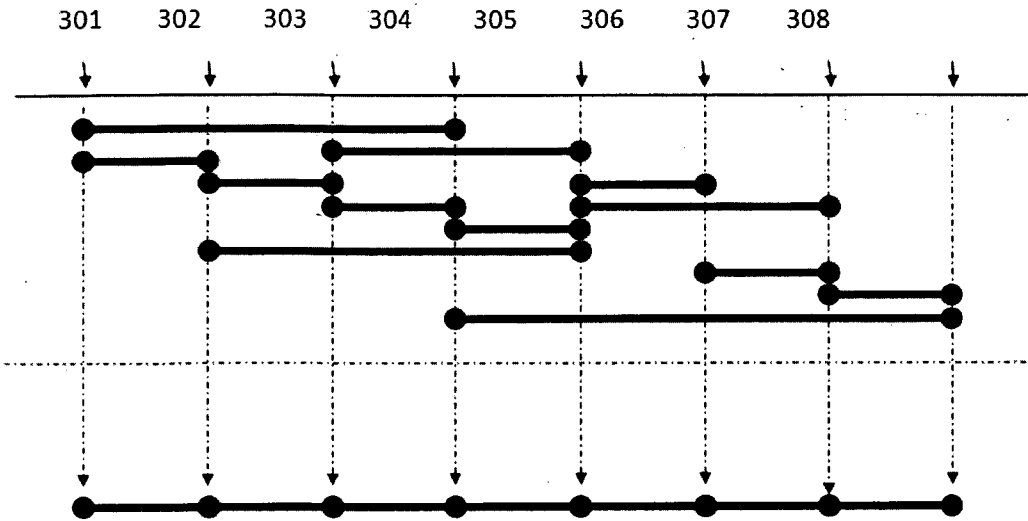


Fig. 1

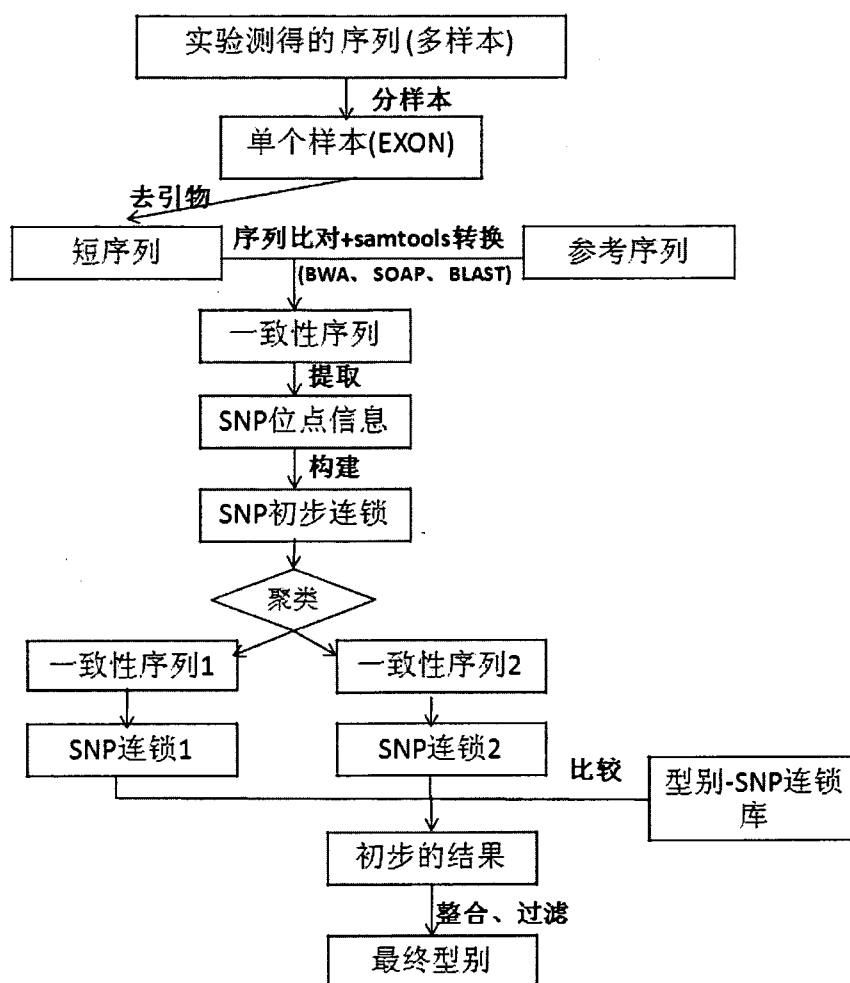


Fig. 2

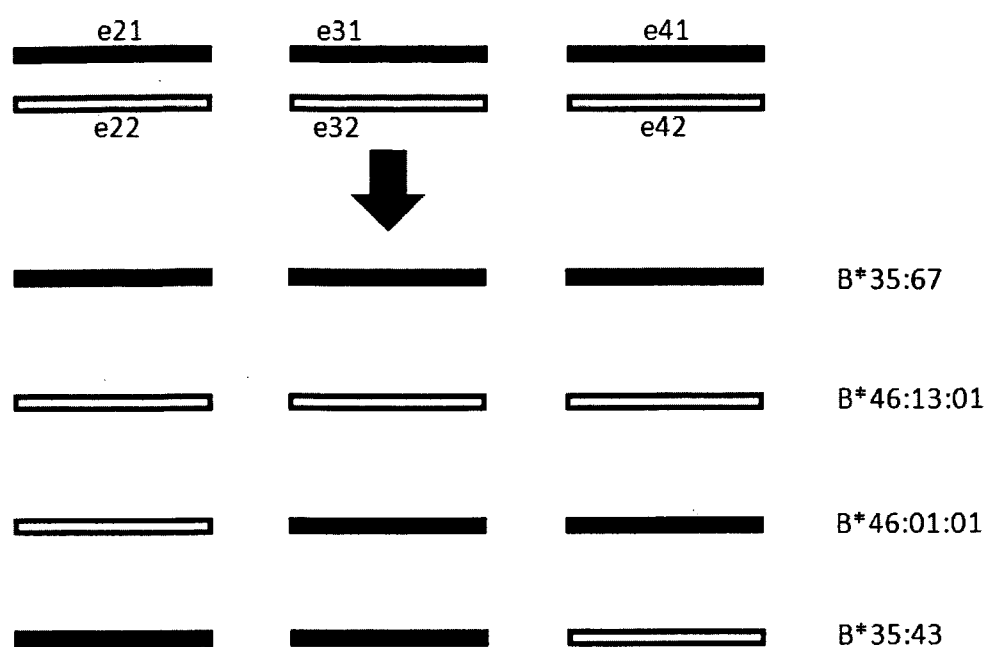


Fig. 3

# PATENT COOPERATION TREATY

# PCT

## DECLARATION OF NON – ESTABLISHMENT OF INTERNATIONAL SEARCH REPORT

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

|                                                                                                            |                                                                                 |                                                                        |
|------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------|------------------------------------------------------------------------|
| Applicant's agent's reference<br>IEC100078PCT                                                              | <b>IMPORTANT DECLARATION</b>                                                    | Date of mailing ( <i>day/month/year</i> )<br>01 Sep. 2011 (01.09.2011) |
| International application No.<br>PCT/CN2010/001879                                                         | International filing date ( <i>day/month/year</i> )<br>23 Nov. 2010(23.11.2010) | (Earliest)Priority date( <i>day/month/year</i> )                       |
| International Patent Classification (IPC) or both national classification and IPC<br>C12Q 1/68 (2006.01) i |                                                                                 |                                                                        |
| Applicant<br><br>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:  
 Claims 1-3 are directed to methods for constructing a database, claims 5-6 are directed to methods for determining the linkage relationship, and claims 7-10 are directed to HLA typing methods and devices thereof, thus all of these claims belong to methods of performing purely mental acts. Moreover, claim 4 is directed to a database and belongs to mere presentations of information. Therefore, all claims of the present application belong to the subject matter listed in PCT Rule 39.1 (iii) and (v) that is not required to be searched, and no international search has been conducted.

|                                                                                                                                                                                                                   |                                                                                                                                 |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| Name and mailing address of the ISA/CN<br><br>The State Intellectual Property Office, the P.R.China<br><br>6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088<br>Facsimile No. 86-10-62019451 | Authorized officer<br><br><div style="text-align: center; font-size: 1.2em;">SU, Lin</div><br><br>Telephone No. (86-10)62411030 |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|

# 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：<br><b>IEC100078PCT</b>                      | 重 要 宣 布                                          | 发文日（日/月/年）<br><b>01.9 月 2011 (01.09.2011)</b> |
| 国际申请号<br><b>PCT/CN2010/001879</b>                         | 国际申请日（日/月/年）：<br><b>23.11 月 2010(23.11.2010)</b> | (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<br><b>C12Q 1/68 (2006.01) i</b> |                                                  |                                               |
| 申请人：<br><b>深圳华大基因科技有限公司 等</b>                             |                                                  |                                               |

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：

1. ☒ 国际申请的主题涉及：

- a. ☐ 科学理论。
- b. ☐ 数学理论。
- c. ☐ 植物品种
- d. ☐ 动物品种。
- e. ☐ 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。
- f. ☐ 经营的方案、规则或方法。
- g. ☒ 纯智力活动的方案、规则或方法。
- h. ☐ 游戏的方案、规则或者方法。
- i. ☐ 处理人体的外科手术或治疗方法。
- j. ☐ 处理动物体的外科手术或治疗方法。
- k. ☐ 在人体或动物体上实施的诊断方法。
- l. ☒ 单纯的信息提供。
- m. ☐ 本国际检索单位不具备检索现有技术的计算机程序。

2. ☒ 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

3. ☐ 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。

4. ☒ 补充意见：

权利要求 1-3 涉及构建数据库的方法，权利要求 5-6 涉及确定连锁关系的方法，权利要求 7-10 涉及 HLA 分型方法及其装置，所有这些权利要求属于进行纯粹智力行为的方法的范畴。此外，权利要求 4 涉及一种数据库，属于单纯的信息表达的范畴。因此，本申请的全部权利要求都属于 PCT 细则 39.1 (iii)和(v)中所列的无需进行检索的主题。因此未进行国际检索。

|                                                                              |                                                        |
|------------------------------------------------------------------------------|--------------------------------------------------------|
| 中华人民共和国国家知识产权局(ISA/CN)<br>中国北京市海淀区蓟门桥西土城路 6 号 100088<br>传真号： (86-10)62019451 | 受权官员<br><br><b>苏林</b><br>电话号码： (86-10) <b>62411030</b> |
|------------------------------------------------------------------------------|--------------------------------------------------------|