

DOI: 10.3724/SP.J.1005.2010.00685

中国汉族个体 *HLA-A*、*-B* 基因全长序列的测定及调控区多态性

徐筠婷, 邓志辉, 邹红岩, 高素青, 王大明, 何柳媚, 魏天莉

深圳市血液中心, 深圳市组织配型与免疫遗传重点实验室, 深圳 518035

摘要: 文章利用 20 个中国汉族个体样本建立了稳定精确的 *HLA-A*、*-B* 基因全长序列的克隆测序方法, 获得 *HLA-A* 10 个等位基因 4.2 kb 序列, *HLA-B* 6 个等位基因 3.7 kb 序列, 序列涵盖了两个基因的所有外显子、所有内含子、5'启动子区以及 3'非翻译区(3'UTR)。A*1153 是文章发现的一个新等位基因, B*151101 的内含子序列、5 个 *HLA-A* 以及 2 个 *HLA-B* 等位基因的 5'启动子序列和 3'UTR 序列为国际上首次报道, 其他等位基因均延伸了 IMGT/HLA 数据库中释放的全长序列。文章首次在中国汉族个体中测定了 IMGT/HLA 数据库中没有覆盖的 *HLA-A*、*-B* 基因的上游 5'启动子以及下游 3'UTR 区域的多态性模式。*HLA-A* 基因 5'启动子延伸区域共发现 26 个 SNPs 和一处 3 bp(AAA/-)的插入/缺失, 3'UTR 延伸区域共发现 14 个 SNPs; *HLA-B* 基因 5'启动子延伸区域共发现 5 个 SNPs 和一处 1 bp(T/-)的插入/缺失, 3'UTR 延伸区域共发现 8 个 SNPs。通过对两个基因的 5'启动子、外显子以及 3'UTR 的系统发育树分析, 发现两个基因调控区与外显子的进化关系有所不同, *HLA-A* 基因除 A*24020101 外, 其他等位基因两端调控区与外显子连锁比较紧密, *HLA-B* 基因两端调控区与外显子之间则发生了较为频繁的重组事件。

关键词: 人类白细胞抗原; 基因组序列; 克隆测序; 调控区; 多态性

Cloning and sequencing *HLA-A* and *-B* genomic DNA and analyzing polymorphism in regulatory regions in Chinese Han individuals

XU Yun-Ping, DENG Zhi-Hui, ZOU Hong-Yan, GAO Su-Qing, WANG Da-Ming, HE Liu-Mei, WEI Tian-Li

Shenzhen Key Laboratory of Histocompatibility and Immunogenetics, ShenZhen Blood Center, Shenzhen 518035, China

Abstract: In the present study, a high-resolute method for cloning and sequencing genomic full-length *HLA-A* and *-B* using 20 Chinese Han individuals was established. We detected 10 *HLA-A* allele sequences 4.2 kb in length and 6 *HLA-B* allele sequences 3.7 kb in length, and the sequences included all exons, all introns, 5' promoter, and 3' UTR of the two genes. All sixteen sequences have been submitted to GenBank and IMGT/HLA database. A*1153 is a novel allele, and the introns of B*151101 are firstly reported here. The 5' promoter and 3' UTR sequences of 5 *HLA-A* alleles and 2 *HLA-B* alleles are also firstly disclosed, and all other alleles have extended the genomic full length sequences released in IMGT/HLA database. The polymorphic structures of upper 5' promoter and downstream 3' UTR, which were uncovered in IMGT/HLA database, are firstly depicted in Chinese Han individuals. Twenty-six single nucleotide polymorphisms (SNPs) and one 3 bp-insertion/deletion (Indel) were located in the upper 5' promoter and 14 SNPs were located in the 3' UTR of *HLA-A*. In addition, five SNPs and one 1 bp-indel were located in the upper 5' promoter and 5 SNPs were located in the 3' UTR of

收稿日期: 2010-03-01; 修回日期: 2010-04-19

基金项目: 广东省自然科学基金资助项目(编号: 9451803501004124)和广东省科技计划项目(编号: 2008B030301277)资助

作者简介: 徐筠婷(1982-), 女, 博士, 研究方向: 分子免疫遗传学。Tel: 0755-83242567; E-mail: yunpingxu1982@yahoo.com.cn

HLA-B. Through analyzing the phylogenetic trees of 5 promoter, exons and 3 UTR of the two genes, we found that the evolution history of regulatory regions and exons is different between the two genes. The regulatory regions are tightly linked with exons in most of *HLA-A* alleles excluding A*24020101. On the contrary, recombinant events may occur frequently between regulatory regions and exons in most *HLA-B* alleles.

Keywords: human leucocyte antigen (HLA); genomic sequence; cloning and sequencing; regulatory region; polymorphism

以往对人类白抗原(Human leukocyte antigen, HLA) 类基因多态性的研究主要是针对编码抗原结合肽的外显子(第 2~4 外显子)。众多研究者对这些外显子的多态位点的探测、疾病相关性以及进化分析研究做了大量工作^[1~3]。20 世纪 90 年代, 科学家们开始对 *HLA* 基因的调控区特别是 5'-启动子的多态性进行研究^[4~6], 并从进化、功能分析等多种角度揭示了这些位于调控区的 SNPs 在基因表达调控中的作用^[7, 8]。大量研究表明, *HLA* 基因调控区多态性蕴含了丰富的与基因进化^[7]、基因表达调控^[8]以及疾病相关联^[9]的信息。

非编码区中除了调控区, 内含子多态性数据也不容忽视, 它们对于测序分型试剂盒的开发及完善非常重要。*HLA* 基因测序分型的 PCR 引物和测序引物, 通常位于目的外显子(*HLA*- 类基因的第 2、3、4 外显子)的侧翼即调控区及内含子内。某些等位基因的调控区及内含子中存在未知的变异, 可能刚好位于现有分型试剂盒的引物结合位点, 导致分型时对该等位基因漏检和丢失现象(Alele dropout)^[10, 11]。

完整的 *HLA-A*、*-B* 基因的全长序列, 应包含 8 个外显子、7 个内含子、5 启动子和 3 UTR。截止 2010 年 1 月, 世界卫生组织(World Health Organization, WHO)HLA 因子命名委员会命名的 *HLA-A* 等位基因达 965 个, *HLA-B* 等位基因有 1 543 个(IMG/HLA Database: <http://www.ebi.ac.uk/imgt/hla/align.html>), 但是 IMG/HLA 数据库中已公布包含 5 启动子和 3 UTR 全长序列的 *HLA-A* 等位基因仅 31 种, 占 3.2% (31/965); *HLA-B* 等位基因有 56 种, 占 3.6% (56/1543), 相对于 *HLA-A*、*-B* 等位基因总数来说是微乎其微的, 完整的外显子、内含子特别是 5 启动子及 3 UTR 两端调控区序列的缺乏, 难以系统地进行 *HLA-A*、*-B* 基因全长序列多态性分析, 以及基因表达调控的研究和临床应用, 大量 *HLA-A*、*-B* 等位基因的全长序

列有待填补。

基于上述需求, 本文利用中国汉族个体样本建立稳定可靠的 *HLA-A*、*-B* 基因全长序列分子克隆及测序方法, 并对基因调控区的多态性进行初步分析, 以应用于后续的中国人群 *HLA-A*、*-B* 基因全长序列多态性以及基因表达调控等研究。

1 材料和方法

1.1 样本

从 100 份已进行 *HLA-A*、*-B* 基因高分辨测序分型(AleleSEQ *HLA-A*、*-B* 测序分型试剂盒, 美国 Ariza 公司)、结果已知的深圳汉族造血干细胞捐献者的血样中选择 12 份互相之间包含相同 *HLA-A* 等位基因的样本, 8 份互相之间包含相同 *HLA-B* 等位基因的样本进行本文的研究, 保证每种等位基因全长序列的检出次数在 2 次以上。每份血样 5 mL, 用 5%EDTA 抗凝, 使用时 DNA 终浓度调节到 50 ng/ μ L。

1.2 方法

1.2.1 长距离 PCR 扩增策略

根据 GenBank(<http://www.ncbi.nlm.nih.gov/genbank>)中公布的人类基因组 *HLA* 全序列的 NC_000006(GI: 89161210)参考序列以及 dbSNP 数据库公布的 *HLA-A*、*-B* 基因周围单核苷酸多态性信息(<http://www.ncbi.nlm.nih.gov/snp>), 采用 Oligo 6 (Molecular Biology Insights, Inc., 2000)软件, 分别设计 *HLA-A*(AF: 5'-ACTCAGAGCTAWGGAATGATG-3', AR: 5'-ATGGAGCAGCAAAGATGAC-3'), *-B*(BF: 5'-CCAGTTCARGGACAGGGATTTC-3', BR: 5'-AACAGACTCAGCACAGCRAAC-3')基因全长扩增简并引物, 特异性地扩增 *HLA-A* 基因目的片段 4.2 kb 以及 *HLA-B* 基因目的片段 3.7 kb, 涵盖了两个基因所有

的外显子和内含子, 以及上游 5' 调控区和下游 3' UTR。

为保证长距离 PCR 产物的准确性, 采用高保真聚合酶(*Pfu* ultra Fusion HS DNA polymerase, 美国 Stratagene 公司)进行扩增。扩增体系包括: 2×GC 缓冲液 25.0 μL, dNTP(25 mmol/L)4.0 μL, PCR 引物 (10 μmol/L)各 2 μL, 基因组 DNA 100 ng, *Pfu* 酶 2.5 U, 加 ddH₂O 至 50 μL。扩增程序: 95℃ 3 min; 94℃ 30 s, 58℃ 30 s, 72℃ 4 min, 共 35 个循环; 最后再 72℃ 延伸 15 min, 4℃ 保存。

1.2.2 PCR 产物的纯化及分子克隆

PCR 产物克隆前需进行两步纯化。第一步纯化的目的是去除长距离 PCR 过程中产生的非特异性产物, 采用切胶回收法(胶纯化试剂盒, 美国 Axygen 公司)。由于 *Pfu* 酶扩增产物末端不带‘A’, 故克隆前应先对产物进行末端加‘A’反应, 该反应体系及程序如下: PCR 纯化产物 17 μL, dATP(25 mmol/L)0.5 μL, 10×*Taq* Buffer 2 μL, *Taq* 酶 0.5 μL, 总体积 20 μL, 于 70℃ 反应 15~30 min。加‘A’反应后利用 EDTA/醋酸钠以及乙醇进行第二步纯化去除多余的缓冲液、

酶以及 dATP, 纯化后进行连接反应。连接反应体系为 10 μL, 内含 pGEM T easy vector (美国 Promega 公司)15 ng, 2×连接 Buffer 5 μL, T4 连接酶 1.5 U, 纯化的 *HLA-A*、*B* 基因 PCR 目的片段 45~120 ng。

连接反应于 4℃ 过夜后, 转化入 JM109 感受态细胞, 抗氨苄培养基中 37℃ 培养过夜。每份样本随机挑选 12 个克隆子, 于 2YT 培养基中扩大培养 16~20 h 后, 提取质粒(质粒小量提取试剂盒, 美国 Axygen 公司), 凝胶电泳法鉴定阳性克隆。

1.2.3 克隆测序

根据目前 IMGT/*HLA* 中已有的全长序列, 尽量从 *HLA-A*、*-B* 基因保守区域由正反方向各设计 5 条步移测序引物(表 1), 加上 pGEM-Teasy 载体两端通用测序引物 T7、SP6, 进行 *HLA-A*、*-B* 各等位基因全长序列的双向测序。

1.2.4 全长序列拼接及确认

每份样本挑出 6 个以上阳性克隆, 保证两种等位基因均有 3 个以上阳性克隆。每一个阳性克隆所测出的 7 条序列, 用 DNASTAR 软件包中的 Seqman

表 1 *HLA-A*、*-B* 基因的全长单倍体测序采用的正向与反向测序引物及其结合位点

| 引物名称 | 引物序列(5' 3') | 结合位置 |
|------|----------------------|-----------------|
| T7 | TAATACGACTCACTATAGGG | 重组质粒末端 |
| AFS1 | CCCAGGCGTGGCTCTCAG | NT-266~NT-249 |
| AFS2 | CTACTACAACCAGAGCGA | NT-451~NT-468 |
| AFS3 | CCTCCCTCTGGTCCTGAG | NT-1091~NT-1108 |
| AFS4 | CCAGACCCAGGACACGGA | NT-1691~NT-1708 |
| AFS5 | AGCAGTCACAAGTCACAG | NT-2288~NT-2305 |
| ARS1 | TAAACAGCCCATCGCATG | NT-2840~NT-2823 |
| ARS2 | AGGTGCTTTACAAAAGAG | NT-2230~NT-2213 |
| ARS3 | GCCAGGTCAGTGTGATCT | NT-1674~NT-1657 |
| ARS4 | AATTGTCTCCCCTCCTTG | NT-1068~NT-1051 |
| ARS5 | GKCCTCGCTCTGGTTGTA | NT-472~NT-455 |
| BFS1 | GCCCTGACCGAGACCTG | NT-52~NT-68 |
| BFS2 | CGGTTTCATTTTCAGTTG | NT-625~NT-642 |
| BFS3 | CCCTCTCTCTCTAGGAC | NT-1146~NT-1163 |
| BFS4 | CAGACCAGCAGGAGATAG | NT-1720~NT-1737 |
| BFS5 | ATGAGTCAGGGAAGGTC | NT-2296~NT-2313 |
| BRS1 | GGATTTTGCCCTCAACTG | NT-2711~NT-2694 |
| BRS2 | GGACTGTCTTCCCCTCCA | NT-2209~NT-2192 |
| BRS3 | TGGTCTGGTCTCCACAAG | NT-1726~NT-1709 |
| BRS4 | GTGCTGGCACACAGGGTC | NT-1208~NT-1191 |
| BRS5 | AGACCCCAAGAATCTCAC | NT-654~NT-637 |
| SP6 | GATTTAGGTGACACTATAG | 重组质粒末端 |

模块进行拼接(Lasergene version 5; DNA Star, Inc., Madison, WI), 人工核查拼接的准确性。拼接后的 4 kb 序列由 Seqman 输出后用 Clustal 软件包中的 Alignment 程序进行比对^[12]。每种等位基因 3 个阳性克隆 4 kb 序列若一致则确认为该条等位基因的序列。最后得到的所有准确序列输入到 Bioedit (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>)以及 Dnasp4.0^[13]软件中进行核苷酸定位及多态性分析。

1.2.5 系统发育树构建

对本文中获得的 *HLA-A*、*-B* 基因全长序列中 5 启动子区、外显子以及 3 UTR 序列分别构建系统发育树, 通过对各序列进化树的拓扑结构进行比较, 可以判断这些区域的进化关系, 是否有重组事件发生。使用 MEGA3.1^[14]中的 Neighbor-Joining 方法构建。其中, 距离矩阵通过 Kimura 双参数法(Kimura's two-parameter method)计算。树的可靠性由 Bootstrapping 重抽样 1 000 次来估计。

2 结果与分析

2.1 长距离 PCR 扩增

采用本文设计的 PCR 扩增引物, 成功扩增了 12 份样本的 *HLA-A* 基因 4.2 kb 序列以及 8 份样本 *HLA-B* 基因 3.7 kb 序列, 每份样本克隆后随机挑选 12 个克隆子, 采用凝胶电泳法鉴定阳性克隆, 每个样本均得到 6 个以上阳性克隆。

2.2 *HLA-A*、*-B* 等位基因全长序列特点及其与 IMGT/HLA 数据库比较

IMGT/HLA 数据库中释放的 *HLA-A*、*-B* 全长序列分为两类: 一类只包含了所有的外显子和内含子序列(*HLA-A*: NT1~NT2903; *HLA-B*: NT1~NT2689), 而 5 上游及 3 下游调控序列并不清楚; 另一类除了包含所有外显子和内含子序列外, 还包含部分 5 启动子(*HLA-A*: NT-300~NT-1; *HLA-B*: NT-284~NT-1)以及部分 3 UTR 序列(*HLA-A*: NT2904~NT3203; *HLA-B*: NT2690~NT3043)。采用本文的方法, 可测得 *HLA-A* 基因 5 启动子的 NT-819 至 3 UTR 下游 NT3435 共 4 254 bp 序列 (以 A*01010101 作为参比序列, 起始密码子 ATG 中的 'A' 为 NT1), 以及 *HLA-B* 基因 5 启动子的 NT-410 至 3 UTR 下游

NT3298 共 3 708 bp 序列(以 B*350101 为参比序列)。所有序列的 GenBank 及 IMGT/HLA 数据库登录号见表 2。与 IMGT/HLA 数据库中现有数据比较, 所有等位基因均延伸了 5 启动子及 3 UTR 序列。其中 A*020601、A*0207、A*110101、A*330301 以及 B*550201 的 5 启动子及 3 UTR 序列为首次报道。A*1153 是本研究中发现的一个新等位基因, 由世界卫生组织(World Health Organization, WHO)HLA 命名委员会命名。与最相似的等位基因 A*110201 相比, 在第五外显子上有一个突变(NT1951 C T, codon276 CCG CTG), 导致氨基酸发生改变(Pro Leu), 该等位基因所有内含子、5 启动子及 3 UTR 序列均与 A*110101 一致。等位基因 B*151101 在 IMGT/HLA 数据库中只有外显子的序列, 本文中首次报道了其内含子序列, 发现它与最接近的 B*150101 序列内含子一致。每个等位基因在样本中出现的次数统计见表 2, 不同样本中相同等位基因序列没有变异。

2.3 *HLA-A*、*-B* 基因调控区多态性分析以及与外显子的进化关系

本文中 A*020601、A*0207、A*110101、A*1153、A*330301 以及 B*151101、B*550201 的 5 启动子序列及 3 UTR 序列为首次报道, 通过与 IMGT/HLA 数据库中相近等位基因序列比较发现, A*020601、A*0207 与 A*02010101 的 5 启动子(NT-300~NT-1)及 3 UTR(NT2904~NT3203)序列均一致; 同样, A*110101、A*1153 与 A*110201 三者之间的 5 启动子及 3 UTR 序列均一致; A*330301 与最相近的 A*330201 及 A*3307 相比在 5 启动子的 NT-274 发现一个新突变(C A); B*151101 和 B*550201 的 5 启动子(NT-284~NT-1)及 3 UTR(NT2690~NT3043)分别与最相近的 B*150101 和 B*550101 相比均无差异。

值得注意的是, 本研究首次揭示了在 IMGT/HLA 数据库中没有覆盖到的 5 启动子及 3 UTR 区域的多态性模式。在本文发现的 *HLA-A* 基因 10 条等位基因全长序列中, 上游 5 启动子区域(NT-819~NT-301)共发现 26 个 SNPs, 并且在-545 与-544 位之间有一处 3 个碱基(AAA/-)的插入/缺失, 下游 3 UTR 区域(NT3204~NT3435)共发现 14 个 SNPs。*HLA-B* 基因 6 条等位基因全长序列中, 上游 5 启动子

表 2 本研究发现的 *HLA-A*、*-B* 等位基因信息

| 等位基因 | 长度(bp) | 检出次数 | GenBank 登录号 | IMGT/HLA 数据库登录号 |
|------------|--------|------|-------------|-----------------|
| A*01010101 | 4 254 | 2 | GU812295 | HWS10010355 |
| A*02010101 | 4 269 | 2 | GQ996941 | HWS10006796 |
| A*020601 | 4 271 | 2 | GU812298 | HWS10010358 |
| A*0207 | 4 269 | 3 | GQ996942 | HWS10006797 |
| A*03010101 | 4 256 | 2 | GU812296 | HWS10010356 |
| A*110101 | 4 252 | 3 | GQ996938 | HWS10006792 |
| A*1153 | 4 252 | 3 | GQ996939 | HWS10006794 |
| A*24020101 | 4 254 | 3 | GQ996940 | HWS10006795 |
| A*260101 | 4 271 | 2 | GU812297 | HWS10010357 |
| A*330301 | 4 270 | 2 | GQ996943 | HWS10006798 |
| B*151101 | 3 717 | 3 | GU812304 | HWS10010364 |
| B*350101 | 3 708 | 3 | GU812299 | HWS10010359 |
| B*400102 | 3 705 | 3 | GU812300 | HWS10010360 |
| B*480101 | 3 705 | 2 | GU812301 | HWS10010361 |
| B*510101 | 3 708 | 3 | GU812302 | HWS10010362 |
| B*550201 | 3 714 | 2 | GU812303 | HWS10010363 |

区域(NT-410~NT-285)共发现 5 个 SNPs 以及一处单碱基(T/-)插入/缺失, 下游 3 UTR 区域(NT3044~NT3298)共发现 8 个 SNPs(表 3)。从表 3 中可以看出, 属于同一家系的等位基因 5 启动子与 3 UTR 序列非常相似, 如 A2 家系中 3 个等位基因 A*02010101、A*020601 与 A*0207 的 5 启动子序列在 NT-819~NT-301 位置及下游 3 UTR 区域在 NT3204~NT3435 位置的序列; A11 家系中的两个等位基因 A*110101 与 A*1153 两端调控区序列也非常相似。

但不同家系等位基因调控区之间的远近关系是否与外显子亲缘关系一致? 我们将所获得的 A 基因 10 条序列以及 B 基因 6 条序列各划分出 3 个区域: 5 启动子、外显子 2~4 以及 3 UTR, 分别作系统发育树进行比较分析。由图 1a 可以看出, *HLA-A* 基因 3 个区域的进化树均包含两个大分支, 5 启动子与 3 UTR 之间树形以及位于两大分支上的等位基因基本一致: A*01010101、A*110101、A*1153 与 A*03010101 聚在同一大支且相互之间的亲缘关系在 3 个区域中几乎无变化, A*02010101、A*020601、A*0207、A*24020101、A*260101 以及 A*330301 聚在另一大支系。A*02010101 所在分支在 5 启动子与 3 UTR 之间有 3 个等位基因 A*24020101、A*260101 以及 A*330301 出现一些排列变化: A*24020101 与 A*260101 的 5 启动子相近但两者与 A*330301 相差较远, 而在 3 UTR 树中 A*260101 与 A*330301 距离

更近, 这个结果也可以通过表 3 中各等位基因之间的序列差异程度判断出。奇怪的是, 外显子树形与两端调控区相比差异较大, A*24020101 的外显子并没有聚类在 A*02010101 所在的大支上, 而是明显分布在 A*01010101 所在的分支上, 说明 A*24020101 的外显子与调控区在进化过程中与其他等位基因之间发生过重组事件。

相对于 *HLA-A* 基因来说, *HLA-B* 基因在 3 个区域之间发生的重组事件更为频繁, 由图 1b 所示, 3 个区域的进化树同样都包含两个大分支, 但是互相之间的树形与分支上的等位基因排列都不一致, 如 B*350101 与 B*151101 的外显子亲缘关系最近且两者与 B*400102 及 B*480101 关系较远, 但是在 3 UTR 系统发育树及表 3 中看到, B*151101 的 3 UTR 与 B*400102 及 B*480101 的 3 UTR 非常相近。

3 讨论

2003 年, Cox 等^[15]在国际上首先报道了 *HLA-B*、*-C* 基因全长的克隆测序研究, 序列长约 3.3 kb, 包含了部分 5 启动子和部分 3 UTR 序列。Zhu 等^[16]最近报道了 15 个 *HLA-A* 等位基因全长序列的多态性, 但是序列仅包括所有的外显子及内含子。本文所建立的 *HLA-A*、*-B* 全长序列的克隆及测序方法, 集简并扩增引物的设计、高保真长距离 PCR 扩增、克隆及双向步移测序等要素, 保证了方法的通用性

以及所得序列的可靠性。利用该方法测出的序列,不仅包含了 *HLA-A*、*-B* 基因所有的外显子和内含子,还延伸了两端调控区 5 启动子和 3 UTR 的序列。所有序列均得到国际 IMGT/HLA 数据库的认可并公布。

HLA-I 类基因表达调控起关键作用的 Enhancer A、ISRE (Interferin-stimulated response element)、Site α 、Enhancer B 等调控元件,以丛集的方式存在于基因起始密码子的上游 300 bp 以内(即 5 启动子的 NT-300~NT-1),协同进行转录活性的调控^[17]。IMGT/HLA 数据库释放的全长数据只包含了 300 bp 调控功能比较清晰的近端启动子。*HLA-I* 类基因 3 UTR 的调控功能偶见报道^[18],但一直没有明确。对动物基因 3 UTR 的研究表明,3 UTR 在转录后调控中扮演着重要角色,它的影响包括翻译效率、mRNA 的稳定性、mRNA 的核质运输等^[19]。近年来,由于 microRNA 的发现^[20],3 UTR 作为其靶标序列在基因表达调控中的地位进一步受到关注,3 UTR 中产生的新突变,有可能形成 miRNA 作用的靶基序^[21]。完整的 *HLA-A* 基因 3 UTR 序列应该是从 NT2904~NT3324,*HLA-B* 基因 3 UTR 序列应该是从 NT2690~NT3287,而 IMGT/HLA 释放的两个基因 3 UTR 序列仅 300 bp。本研究首次测定了 IMGT/HLA 数据库中没有覆盖的 *HLA-A*、*-B* 基因的上游 5 启动子及下游 3 UTR 区域的多态性模式。*HLA-A* 基因 5 启动子延伸区域共发现 26 个 SNPs 和一处 3 bp(AAA/-)的插入/缺失,3 UTR 延伸区域共发现 14 个 SNPs;*HLA-B* 基因 5 启动子延伸区域共发现 5 个 SNPs 和一处 1 bp(T/-)的插入/缺失,3 UTR 延伸区域共发现 8 个 SNPs,表明两端调控区延伸区域仍然存在较为丰富的多态性,它们是否参与基因的表达调控,还有待功能验证。

HLA 基因的表达与其生物学功能密切相关,研究表明 *HLA* 等位基因在不同细胞、不同时期(如正常、刺激及病理状态)存在水平不一致的特异性表达^[22]。这种特异性表达影响着 T 细胞受体以及抗原的识别^[23],从而产生免疫应答的个体差异。不同等位基因的特异性表达及其对免疫应答的调控是极其复杂精细的过程,由调控区多态性、外显子多态性以及多种细胞因子共同决定。对全长序列单倍型的完整测定,可以清晰地判断出调控区与外显子的多

态性及相互之间的关联,为等位基因表达的时空特异性解释提供最原始的分子基础。*HLA-A* 基因调控区与外显子的系统发育树分析发现,除 A*24010101 外,其他等位基因两端调控区均与外显子紧密连锁,从表 3 可知,不同家系等位基因的调控区具有家系特异性,表明 *HLA-A* 等位基因的特异性表达与其调控区有非常紧密的关联。而 *HLA-B* 基因调控区与外显子的关系较 *HLA-A* 来说连锁并不紧密,从系统发育树分析可以看出,*HLA-B* 基因无论是 5 启动子还是 3 UTR 与外显子都发生过较为频繁的重组事件,说明 *HLA-B* 基因的调控区与外显子并不是协同进化的。

本文共获得了 10 条 *HLA-A* 基因 4.2 kb 全长序列,6 条 *HLA-B* 基因 3.7 kb 全长序列。A*1153 是本文发现的一个新等位基因,它与最近的 A*110201 差异在第五外显子,常规测序分型由于只针对第 2~4 外显子,所以很容易将其忽略而判定为 A*110201。然而在本研究中,12 个个体出现 3 次该等位基因,表明它是一个常见等位基因,在后续的测序分型中应该增加第五外显子的检测及判定。B*151101 的内含子序列、5 个 *HLA-A* 以及 2 个 *HLA-B* 等位基因的 5 启动子序列和 3 UTR 序列为国际上首次报道,其他等位基因均延伸了 IMGT/HLA 数据库中释放的全长序列。这些数据大大丰富了 *HLA-A*、*-B* 基因的全长序列多态性资源。

参考文献(References):

- [1] Hughes AL, Nei M. Pattern of nucleotide substitution at MHC class I loci reveals overdominant selection. *Nature*, 1988, 335(6186): 167-170.
- [2] Hughes AL, Yeager M. Natural selection at major histocompatibility complex loci of vertebrates. *Annu Rev Genet*, 1998, 32: 415-435.
- [3] Gao X, Nelson GW, Karacki P, Martin MP, Phair J, Kaslow R, Goedert JJ, Buchbinder S, Hoots K, Vlahov D, O'Brien SJ, Carrington M. Effect of a single amino acid change in MHC class I molecules on the rate of progression to AIDS. *N Engl J Med*, 2001, 344(22): 1668-1675.
- [4] Vallejo AN, Pease LR. Structure of the MHC A and B locus promoters in hominoids. Insights on the evolution of the class I MHC multigene family. *J Immunol*, 1995, 154(8): 3912-3921.
- [5] Xu Y, Hu Q, Liu Z, Shen Y, Liu X, Lin B, Wu Y, Chen S,

- Xu A. Sequence variations in the transcriptional regulatory region and intron1 of HLA-DQB1 gene and their linkage in southern Chinese ethnic groups. *Immunogenetics*, 2005, 57(7): 465–478.
- [6] Yao Z, Volgger A, Scholz S, Albert E. Polymorphism of the HLA-C promotor region. *Immunogenetics*, 1997, 45(6): 428–431.
- [7] Tan Z, Shon AM, Ober C. Evidence of balancing selection at the HLA-G promoter region. *Hum Mol Genet*, 2005, 14(23): 3619–3628.
- [8] Griffioen M, Ouwerkerk IJ, Harten V, Schrier PI. HLA-B locus-specific downregulation in human melanoma requires enhancer A as well as a sequence element located downstream of the transcription initiation site. *Immunogenetics*, 2000, 52(1–2): 121–128.
- [9] Thomas R, Apps R, Qi Y, Gao X, Male V, O'hUigin C, O'Connor G, Ge D, Fellay J, Martin JN, Margolick J, Goedert JJ, Buchbinder S, Kirk GD, Martin MP, Telenti A, Deeks SG, Walker BD, Goldstein D, McVicar DW, Moffett A, Carrington M. HLA-C cell surface expression and control of HIV/AIDS correlate with a variant upstream of HLA-C. *Nat Genet*, 2009, 41(12): 1290–1294.
- [10] Delfino L, Morabito A, Ferrara GB. HLA-C sequence based typing: nucleotide analysis from exon 1 through exon 8. Identification of a new allele: Cw*0718. *Tissue Antigens*, 2003, 62(5): 418–425.
- [11] Heinold A, Schaller-Suefling E, Opelz G, Scherer S, Tran TH. Identification of two novel HLA-alleles, HLA-A*02010103 and HLA-B*4455, and characterization of the complete genomic sequence of HLA-A*290201. *Tissue Antigens*, 2008, 72(4): 397–400.
- [12] Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequences alignment aided by quality analysis tools. *Nucleic Acids Res*, 1997, 25(24): 4876–4882.
- [13] Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, 2003, 19(18): 2496–2497.
- [14] Kumar S, Tamura K, Nei M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform*, 2004, 5(2): 150–163.
- [15] Cox ST, McWhinnie AJ, Robinson J, Marsh SGE, Parham P, Madrigal JA, Little AM. Cloning and sequencing full-length HLA-B and -C genes. *Tissue Antigens*, 2003, 61(1): 20–48.
- [16] Zhu F, He Y, Zhang W, He J, He J, Xu X, Yan L. Analysis of the complete genomic sequence of HLA-A alleles in the Chinese Han population. *Int J Immunogenet*, 2009, 36(6): 351–360.
- [17] Van den Elsen PJ, Gobin SJ, Van Eggermond MC, Peijnenburg A. Regulation of MHC class I and II gene transcription: differences and similarities. *Immunogenetics*, 1998, 48(3): 208–221.
- [18] Snyder SR, Waring JF, Zhu SZ, Kaplan S, Schultz J, Ginder GD. A 3'-transcribed region of the HLA-A2 gene mediates posttranscriptional stimulation by IFN-gamma. *J Immunol*, 2001, 166(6): 3966–3974.
- [19] Decker CJ, Parker R. Diversity of cytoplasmic functions for the 3' untranslated region of eukaryotic transcripts. *Curr Opin Cell Biol*, 1995, 7(3): 386–392.
- [20] 盛熙晖, 杜立新. MicroRNA 及其在人和动物上的研究进展. *遗传*, 2007, 29(6): 651–658.
- [21] 孟详人, 郭军, 赵倩君, 马月辉, 关伟军, 刘娣, 狄冉, 乔海云, 那日苏. 11 个绵羊品种 MSTN 基因非翻译区的变异. *遗传*, 2008, 30(12): 1585–1590.
- [22] Masuda K, Hiraki A, Fujii N, Watanabe T, Tanaka M, Matsue K, Ogama Y, Ouchida M, Shimizu K, Ikeda K, Tanimoto M. Loss or down-regulation of HLA class I expression at the allelic level in freshly isolated leukemic blasts. *Cancer Sci*, 2007, 98(1): 102–108.
- [23] Liu H, Fu J, Bouvier M. Allele- and locus-specific recognition of class I MHC molecules by the immunomodulatory E3-19K protein from adenovirus. *J Immunol*, 2007, 178(7): 4567–4575.