遗传 Hereditas (Beijing) 2021 年 10 月, 43(10): 938－948
www.chinagene.cn

研究报告

基于全基因组数据的 AI-SNPs 筛选及大陆次级区域内群体遗传结构差异研究

王浩宇，胡渝涵，曹悦岩，朱强，黄雨果，李茜，张雳

四川大学华西基础医学与法医学院，成都 610041

摘要: 在涉及多群体样本的医学研究中，群体遗传结构差异是不容忽视的影响因素之一。利用族源信息单核苷酸多态性遗传标记(ancestry-informative single nucleotide polymorphism, AI-SNP)，通过分析群体遗传成分、推断个体遗传背景并对群体样本进行预筛选，可以有效降低群体遗传结构差异对医学研究影响。鉴于已发表的研究多为解析大陆间、大陆次级区域间的群体遗传结构差异，本研究拟基于千人基因组计划(GRCh37.p13)中东亚五群体：日本东京群体(Japanese in Tokyo, JPT)、北京汉族(Han Chinese in Beijing, CHB)、南方汉族(Southern Han Chinese, CHS)、西双版纳傣族(Chinese Dai in Xishuangbanna, CDX)、越南京族(Kinh in Ho Chi Minh City, KHV)的数据，以 F_{ST} 值为标准筛选 AI-SNP 并分析大陆次级区域内群体遗传结构差异。结果表明，研究涉及的东亚群体可分为三簇：JPT、CHB 和 CHS、CDX 和 KHV。利用 AI-SNP 可成功解析个体的遗传背景，而群体代表性遗传成分占比超过 80%的个体具有良好的群体代表性。本研究表明，基于 F_{ST} 值筛选一组 AI-SNP 用于核验样本遗传背景、筛选群体代表性样本的方法在降低大陆次级区域内群体遗传结构差异对群体相关医学研究的影响中具有实际应用价值。

关键词: 族源信息遗传标记；单核苷酸多态性；东亚群体；遗传结构差异

AI-SNPs screening based on the whole genome data and research on genetic structure differences of subcontinent populations

Haoyu Wang, Yuhan Hu, Yueyan Cao, Qiang Zhu, Yuguo Huang, Xi Li, Ji Zhang

West China School of Basic Medical Sciences and Forensic Medicine, Sichuan University, Chengdu 610041, China

Abstract: The genetic structure differences in population is one of the key elements in medical research involving multi-population samples. A set of ancestry-informative single nucleotide polymorphisms (AI-SNPs) can be utilized to

收稿日期: 2021−05−26; **修回日期:** 2021−07−23

基金项目: 国家自然科学基金项目(编号：81571861, 81630054)资助[Supported by the National Natural Science Foundation of China (Nos. 81571861, 81630054)]

作者简介: 王浩宇，在读硕士研究生，专业方向：法医物证学。E-mail: wanghy0707@gmail.com
胡渝涵，在读硕士研究生，专业方向：法医物证学。E-mail: huyuhan28@163.com
王浩宇和胡渝涵并列第一作者。

通讯作者: 张雳，博士，教授，研究方向：法医物证学。E-mail: zhangj@scu.edu.cn

DOI: 10.16288/j.yczz.21-185
网络出版时间: 2021/8/4 17:50:29
URI: https://kns.cnki.net/kcms/detail/11.1913.R.20210804.1141.001.html

analyze genetic component of a population, infer ancestral origin of individuals and pre-filter samples to reduce the impact of population genetic structure differences on medical research. However, most of the published studies were focused on revealing the differences between populations of continents or regions of a continent. In this paper, AI-SNPs were screened by calculating F_{ST} value in each pair of five East Asian populations: Japanese in Tokyo (JPT), Han Chinese in Beijing (CHB), Southern Han Chinese (CHS), Chinese Dai in Xishuangbanna (CDX) and Kinh in Ho Chi Minh City (KHV) in the 1000 Genomes Project phase 3 (GRCh37.p13) to analyze differences in subcontinent populations. The results demonstrate that the five East Asian populations in our study were assigned to three clusters: JPT, CHB and CHS, CDX and KHV. A set of AI-SNPs can be used for analysis of individual genetic composition and selection of representative individuals. Individuals with over 80% population representative genetic components have good representativeness of a population. This paper demonstrated the practical value of the method, which was performed to verify the ancestral composition and select representative samples with a panel of screened AI-SNPs by F_{ST} value, thereby reducing the influence of genetic structure differences in subcontinent populations on population-related medical research.

Keywords: ancestry-informative marker; single nucleotide polymorphism (SNP); East Asian populations; genetic structure differences

不同群体间遗传结构的差异受到种群迁移、隔离、混合等人口学因素，以及基因突变、重组、自然选择、随机遗传漂变等遗传学因素影响[1,2]。涉及群体的医学领域研究中，往往需考虑由群体遗传结构差异带来的影响。如关联分析中，需排除与目标基因无关、由群体间结构差异导致的等位基因频率差异，才能提供标记与疾病间的真实关联[3,4]。而明确药物反应相关基因变异[5]在群体中的差异则有利于针对不同人群进行靶向药物的筛选并提供精准个性化用药建议。此外，族源信息遗传标记也被法医遗传学家用于生物样本的生物地理起源推断，并用于案件侦破[6]。

在排除群体结构差异对医学研究的影响时，需对纳入研究的个体和生物样本进行遗传背景分析以核验声明血统和实际血统的一致性，并选择具有群体代表性的样本进行后续研究。常用的遗传背景分析工具包括基因芯片[7]、全基因组测序[8]和使用族源信息遗传标记(ancestry informative marker, AIM)[9]。尽管基因组测序可得到最精确的分析结果，但其数据分析量巨大且成本较高。在当前大数据时代下，诸如国际基因组样本资源库(The International Genome Sample Resource, IGSR)[10]中千人基因组计划(1000 Genomes Project)[11]、人类基因组多样性计划(Human Genome Diversity Project)等数据库提供了大量不同人群的基因组参考数据。依托于公开数据库的大规模数据，以 AIM 为基础的族源分析可解析个体遗传背景，并作为应用基因芯片或全基因组测序前进行群体代表性样本预筛选的有效手段[12]。

分析个体遗传背景常用的方法包括主成分分析(principal component analysis, PCA)[13]、基因组控制(genomic control)[14]及结构化关联(structured association)[15]等。PCA 分析是校正全基因组关联研究中群体分层的标准方法，但对如东亚群体这类遗传结构复杂的群体敏感性较差[16]。STRUCTURE[17]、ADMIXTURE[18]等结构化关联方法可依据族源成分和等位基因频率提供个体族源的最大似然估计，STRUCTURE 还提供了基于相关等位基因频率的混合祖先模型用于复杂遗传结构群体的分析。同时，预筛选仅分析一组 AIM，避免了结构化关联方法难以计算大型数据集的缺点[19]，故结构化关联方法可在样本预筛选中发挥关键作用。

族源推断分析最初多以区分大陆群体为目标[20]。近来也有不少研究者针对大陆内特定区域群体的区分开发了多类次级体系。以亚洲为例，主要包括亚洲内次级区域群体[21]、大陆次级区域内群体与全球其他群体区分[22]、亚洲内次级区域群体间的区分[23]和国家内民族的区分[24]，而大陆次级区域内群体间区分的相关研究则相对较少[25]。由于大陆次级区域内群体间遗传结构的相似性，以及人口迁移、通婚带来的基因流动等因素，此类区分最为困难，但也

是最为必要的。

本研究拟以 F_{ST} 值大小为标准，从千人基因组计划东亚五群体的数据中筛选一组 AIM 对东亚五个群体进行群体结构分析，从各个群体中找到具有群体代表性的个体。并以结果评估使用 AIM 方法对遗传结构复杂群体中个体遗传背景的解析能力，为其实际应用于核实样本的声明血统和实际血统、准确排除群体遗传结构对群体相关医学研究的影响提供理论依据和方法参考。

1 材料与方法

1.1 研究对象

本研究使用的东亚五个群体共 504 个无关个体均来自千人基因组计划第三阶段(GRCh37.p13)数据库(http://www.1000genomes.org)[11]，包括 104 个日本东京(Japanese in Tokyo, JPT)个体、103 个中国北京汉族(Han Chinese in Beijing, CHB)个体、105 个中国南方汉族(Southern Han Chinese, CHS)个体、93 个中国西双版纳傣族(Chinese Dai in Xishuangbanna, CDX)个体和 99 个越南胡志明市京族(Kinh in Ho Chi Minh City, KHV)个体。

1.2 位点筛选

基于千人基因组数据库(GRCh37.p13)的整体数据，使用 VCFtools[26]筛选 1~22 号常染色体上最小等位基因频率大于 0.01、P>0.05 阈值下满足 Hardy-Weinberg 平衡的二等位基因 SNP。按 Weir 和 Cockerham 等[27]的方法计算东亚五个群体两两之间，即 10 个群体对中所有保留 SNP 的 F_{ST} 值。本研究保留 F_{ST}>0.05 的 SNP，并使用 VCFtools 进行同染色体上的连锁不平衡计算。目前在族源推断体系中加入连锁不平衡位点是否会对体系区分具体群体的效能产生影响尚无定论，但研究者们在进行 AIM 筛选时会避免使用强连锁不平衡的基因座[21]。此外，STRUCTURE 软件也建议在体系中尽可能只使用弱连锁不平衡的位点[28]。因此本研究进行连锁不平衡计算时根据前人经验将阈值设置为 r^2>0.2，并将检测窗口设置为 50 Mb。当一个位点与任意位点满足 r^2>0.2 时，将它们分为同一连锁不平衡组，否则分至弱连锁不平衡组。

1.3 数据集构建

基于前述分组结果，10 个群体对分别建立数据集。各个群体对的每个连锁不平衡组中仅保留 F_{ST} 值最高的 SNP，将连锁不平衡组中筛选出来的 SNP 与该弱连锁不平衡组合并后确定最终的数据集 A_1~A_{10}。

各数据集分别从 F_{ST} 值最高的 10 个 SNP 开始，使用 Snipper 在线分析应用套件进行分析(后简称为 Snipper 分析)。自该体系开始，每次按 F_{ST} 值大小逐步顺序增加 10 个 SNP 并进行 Snipper 分析。为了保证结果的稳定性，此步骤将持续到连续三组体系(如分别由 60、70、80 个 SNP 组成的体系)均能将两个群体的个体均正确分配至原所属群体，也即分配正确率达到 100%时停止。经 STRUCTURE 分析和 PCA 分析验证后，认为该三组体系中的第一组(上述例子中由 60 个 SNP 组成的体系)所包含的 SNP 数是完全区分该群体对所需的最少 SNP 数。基于此结果，本研究筛选了包含尽可能多 SNP(975 个)的数据集 B 分析东亚五个群体的遗传结构。在筛选数据集 B 时，综合考虑了 SNP 的如下信息：在 10 个数据集中出现的次数、在各数据集中对应的 F_{ST} 值大小、F_{ST} 值在该数据集中的排序、是否涉及较难区分的群体(数据集内 SNP 数目较少或 SNP 的 F_{ST} 值普遍较低)等因素。

依据数据集 B 的 STRUCTURE 分析结果，筛选群体代表性遗传成分占个体总遗传成分分别达到 70%~80% (C_7)、80%~90% (C_8)和 90% (C_9)以上的个体作为数据集 C，各数据集内群体则按照群体编号(如数据集 C_7 中 JPT 编号为 JPT7)。对数据集 C 进行 STRUCTURE 分析和 PCA 分析，验证筛选群体代表性个体的可靠性、评估群体代表性遗传成分占比对判断群体代表性个体的影响。

1.4 群体遗传结构分析

使用 STRUCTURE v2.3.4[17]基于相关等位基因频率的混合祖先模型对每个数据集进行群体基因结构分析，数据集 A_1~A_{10} 设置 K=2，数据集 B、C 设置 K=2~7，均运行 10 次。利用 STRUCTURE HARVESTER[29]计算最佳 K 值，CLUMPP v.1.1.2[30]

和 Distruct v.1.1[31]用于构建结果图。Python 脚本用于基于个体基因型的 PCA 分析和结果图构建。Snipper 2.5 在线贝叶斯二分类分析应用套件(http://mathgene.usc.es/snipper/)用于基于训练集和测试集的交叉验证计算，各群体训练集和测试集的个体均按 7∶3 的比例随机设置，每组体系重复三次，最终测试集的结果取均值。

2　结果与分析

2.1　数据集 A、B 中 SNP 概况

数据集 A_1~A_{10} 中 SNP 的 F_{ST} 值分布情况见表 1。F_{ST} 值最高的 20 个 SNP 分别来自 A_3 (JPT-CDX，12 个)、A_4(JPT-KHV，3 个)、A_2(JPT-CHS，3 个)和 A_6(CHB-CDX，2 个)；而 F_{ST} 值最小的 20 个 SNP 均来自 A_5 (CHB-CHS)。除 A_3 (JPT-CDX)、A_4 (JPT-KHV)中 SNP 的 F_{ST} 值主要分布于 0.15~0.25 外，其余数据集内绝大多数 SNP 的 F_{ST} 值均小于 0.15，其中 A_5 (CHB-CHS)所有 SNP 的 F_{ST} 值均小于 0.15。

数据集 B 中 975 个 SNP 在染色体上的分布情况如图 1 所示，整体分布较为均匀。来自 1 号染色体和 6 号染色体的 SNP 最多，分别为 109 个和 95 个，而来自 22 号染色体的 SNP 最少，为 12 个。此外，本研究也统计了此 975 个 SNP 在 10 个群体对中出现的情况，结果如图 2 所示。975 个 SNP 中，大多数 SNP 只在一个(470/975)、两个(296/975)或三个(132/975)群体对中出现，只有极少数 SNP 在五个及

表 1　数据集 A 中 SNP 的 F_{ST} 值分布情况

Table 1　The distribution of F_{ST} value of SNPs in dataset A

数据集	最大 F_{ST} 值	最小 F_{ST} 值	总位点数	$F_{ST}\geqslant 0.25$ 位点数	$0.25>F_{ST}\geqslant 0.15$ 位点数	$0.15>F_{ST}\geqslant 0.05$ 位点数
$A_{1(JPT-CHB)}$	0.266925	0.095475	591	1	19	571
$A_{2(JPT-CHS)}$	0.407479	0.111107	598	10	87	501
$A_{3(JPT-CDX)}$	0.788409	0.16243	630	46	584	0
$A_{4(JPT-KHV)}$	0.583637	0.141417	623	19	435	169
$A_{5(CHB-CHS)}$	0.146048	0.05001	723	0	0	723
$A_{6(CHB-CDX)}$	0.517659	0.109475	563	9	84	470
$A_{7(CHB-KHV)}$	0.25611	0.087147	670	1	18	651
$A_{8(CHS-CDX)}$	0.310909	0.081595	787	3	18	766
$A_{9(CHS-KHV)}$	0.192399	0.069052	631	0	7	624
$A_{10(CDX-KHV)}$	0.256551	0.067479	461	1	6	454

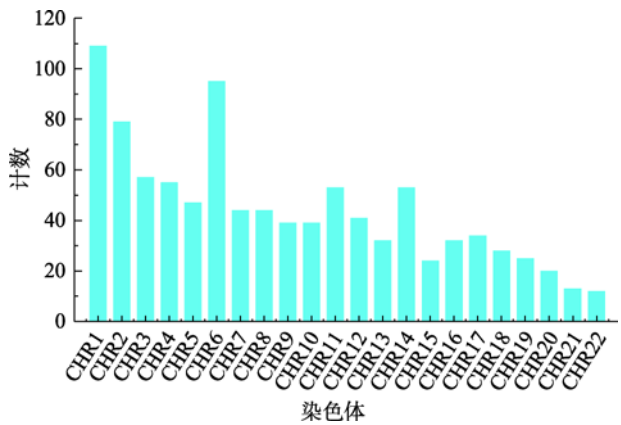

图 1　数据集 B 中 SNP 在染色体上的分布情况

Fig. 1　The chromosome distribution of SNPs in dataset B

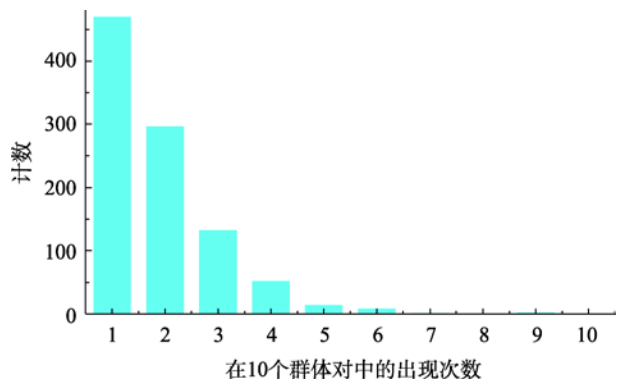

图 2　数据集 B 中 SNP 在 10 个群体对中出现次数的分布情况

Fig. 2　The distribution of SNPs in dataset B in ten paired population

以上(25/975)群体对中出现。其中 rs11850206 和 rs28558239 在除了 CHS-KHV 以外的九个群体对中均有出现，rs28498529 则在除了 JPT-CHB、CHS-KHV、CDX-KHV 以外的七个群体对中出现。此三个 SNP 均来自于 14 号染色体。

此外，本研究将数据集 B 与部分此前研究东亚群体遗传结构差异的文献[21,23,25]所使用的 SNP 进行了比较，发现数据集 B 未包含此三文献中报道的任一 SNP。

2.2 东亚五群体的遗传结构差异分析

对数据集 A_1~A_{10} 进行 Snipper 交叉验证分析，测试集分配完全正确所需最少 SNP 数结果见表 2。群体对中个体祖先分配完全正确所需的最少位点数可反映出群体两两之间遗传关系的远近。结果表明 JPT-CDX、JPT-KHV 群体对最易区分，而 CHB-CHS、CHS-KHV、CDX-KHV 较难区分。各群体对中的群体与 STRUCTURE 计算得到的聚类高度符合，而 PCA 分析中各个群体对均能在使用最少位点数时分别聚类且彼此分离(结果未列出)。

使用数据集 B 对东亚五群体进行 STRUCTURE 分析的结果如图 3 所示。*K* 值设置为 2~7，STRUCTURE HARVESTER 计算得到的最佳 K 值为 3。各个 K 值下 JPT 均表现出与其余群体不同的遗传成分。在最佳 *K* 值时，各群体均表现为混合遗传成分，975 SNPs 可将东亚五群体分为三簇：JPT 一簇、CHB 和 CHS 一簇、CDX 和 KHV 一簇，其中 CHB 和 CHS 还可依据遗传成分的比例区分。自 *K*=4 开始，CDX 和 KHV 也表现出主要遗传成分的差异，这一差异在 *K*=5 时更加显著。而自 *K*=6 开始，各群体混合遗传成分中的主要遗传成分各不相同，即主要遗传成分可与 STRUCTURE 计算得到的聚类匹配，可据此将五个群体分为五簇。

使用数据集 B 对东亚五群体进行 PCA 分析的结果如图 4 所示。前三个主成分分别占总方差的 3.21%、2.12%、1.36%。JPT、CHB、CDX 群体的个体紧密聚集，而 CHS、KHV 群体的聚类较分散。整体上，JPT、CHB、CHS 之间较为接近，其可与互相接近的 CDX、KHV 区分。PC1 维度可进一步将 JPT 与 CHB、CHS 区分，其中 CHB 和 CHS 个体相互重叠，表明二者的遗传关系十分接近(图 4)，而 PC3 维度可将 CDX 和 KHV 区分(图 4B)。

2.3 东亚五群体代表性个体筛选及分析

以数据集 B 进行 STRUCTURE 分析时 *K*=6 的结果为参考，按 1.3 的方法判断五个群体的群体代表性遗传成分并构建数据集 C (表 3)。数据集 C 中共包括 317 个个体，JPT 中群体代表性遗传成分占总体遗传成分超过 70%的个体最多，达 93%，其次是 CDX 和 KHV，分别为 78%和 59%，CHB 和 CHS 均

表 2 数据集 A 中两两群体完全区分所需最少 SNP 数

Table 2 The minimum number of SNPs required for a complete distinction between paired populations in dataset A

子数据集	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
SNP 数目	40	110	30	40	150	50	90	100	280	230

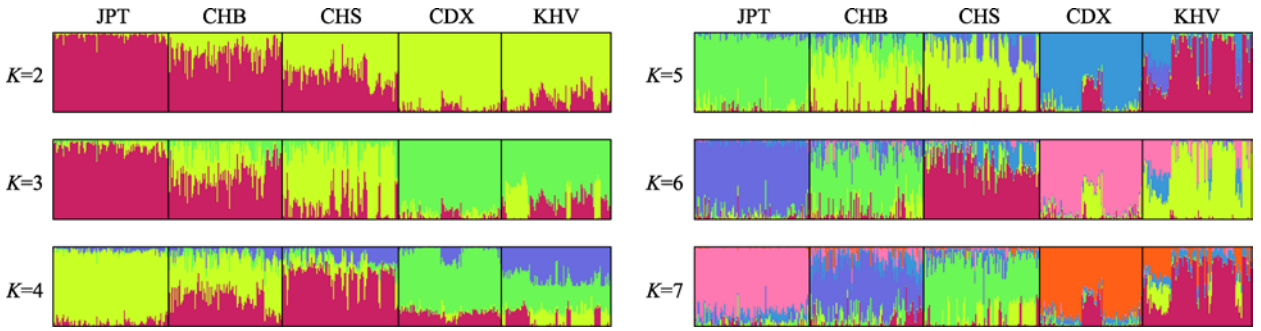

图 3 975 SNPs (数据集 B)的东亚五群体 STRUCTURE 分析结果

Fig. 3 The STRUCTURE analysis of 975 SNPs (dataset B) for five East Asian populations

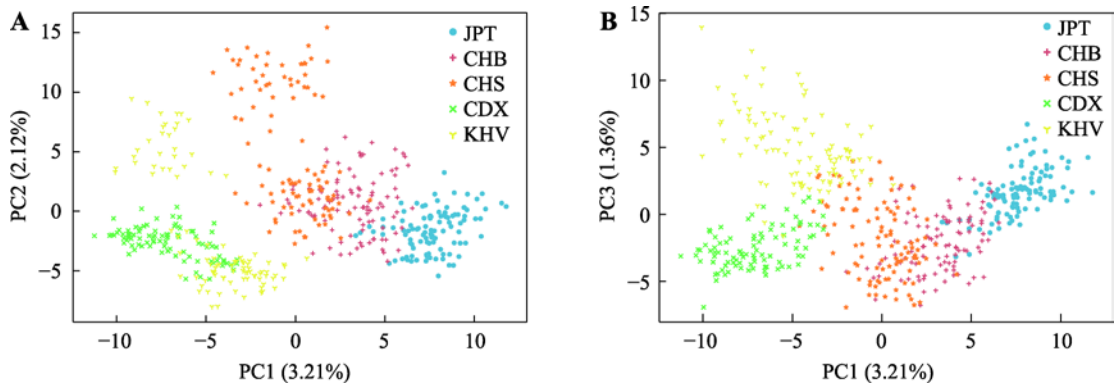

图 4　975 SNPs (数据集 B)的东亚五群体 PCA 分析结果

Fig. 4　The PCA analysis of 975 SNPs (dataset B) for five East Asian populations

各颜色代表群体：JPT(蓝色)，CHB(红色)，CHS(橙色)，CDX(绿色)，KHV(黄色)。A：975 SNPs 的东亚五群体 PCA 分析(PC1-PC2)，PC1=3.21%，PC2=2.12%；B：975 SNPs 的东亚五群体 PCA 分析(PC1-PC3)，PC1=3.21%，PC3=1.36%。

表 3　数据集 C 中 C_7、C_8、C_9 组个体数目

Table 3　The number of individuals in C_7, C_8, C_9 of dataset C

组别	JPT	CHB	CHS	CDX	KHV
C_7	18	20	13	9	6
C_8	41	20	21	20	11
C_9	38	6	9	44	41
合计/总样本数	97/104	46/103	43/105	73/93	58/99

未超过 50%。JPT、CDX、KHV 的筛选个体中大部分群体代表性遗传成分占比超过 80%，CHB 和 CHS 只有较少个体的群体代表性遗传成分占比超过 90%。

使用数据集 B 的 975 个 SNP 对筛选个体进行 STRUCTURE 分析的结果如图 5 所示。在各个 *K* 值下，筛选个体均表现为混合遗传成分。计算得到的最佳 *K* 值为 4，此时筛选出的个体可被分为四簇：JPT 一簇、CHB 和 CHS 一簇、CDX 一簇、KHV 一簇。自 *K*=5 开始，317 个个体可被分为五簇，各簇几乎都完全由其主要遗传成分组成，且其比例随着群体代表性遗传成分占比的增加而增加，但占比达到 80%后趋于稳定。STRUCTURE 的结果表明体系能够很好地区分筛选出的个体，即筛选个体能有效代表其所属群体。此外，群体代表性遗传成分占比更高的个体具有更强的群体代表性。

将数据集 C_7、C_8、C_9 在前述 PCA 分析中分别高亮表示的结果如图 6 所示。在全部个体中，筛选个体之间区分度更高，并随着个体的群体代表性遗

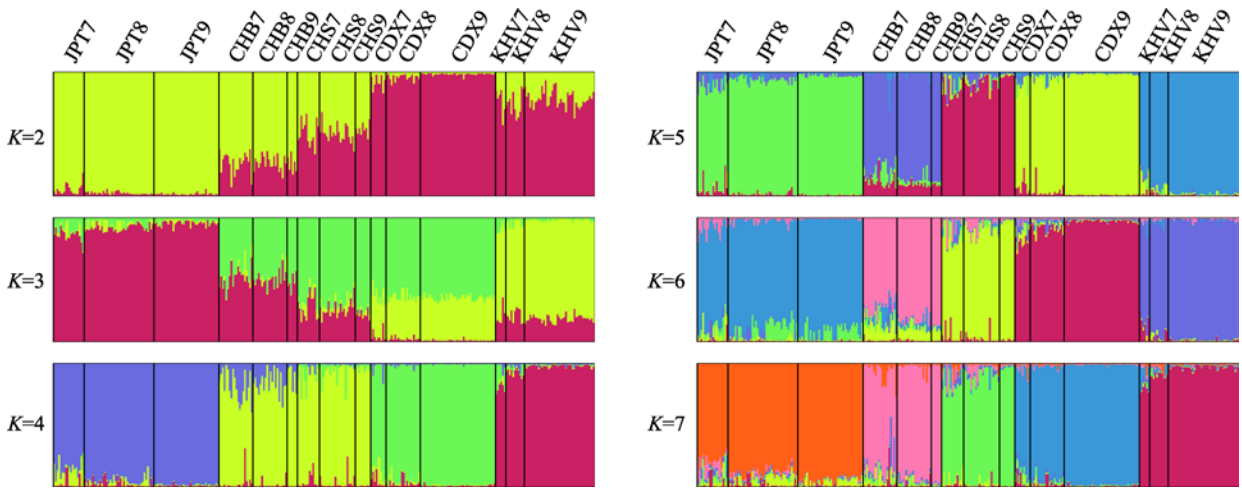

图 5　975 SNPs 的数据集 C STRUCTURE 分析结果

Fig. 5　The STRUCTURE analysis of 975 SNPs for dataset C

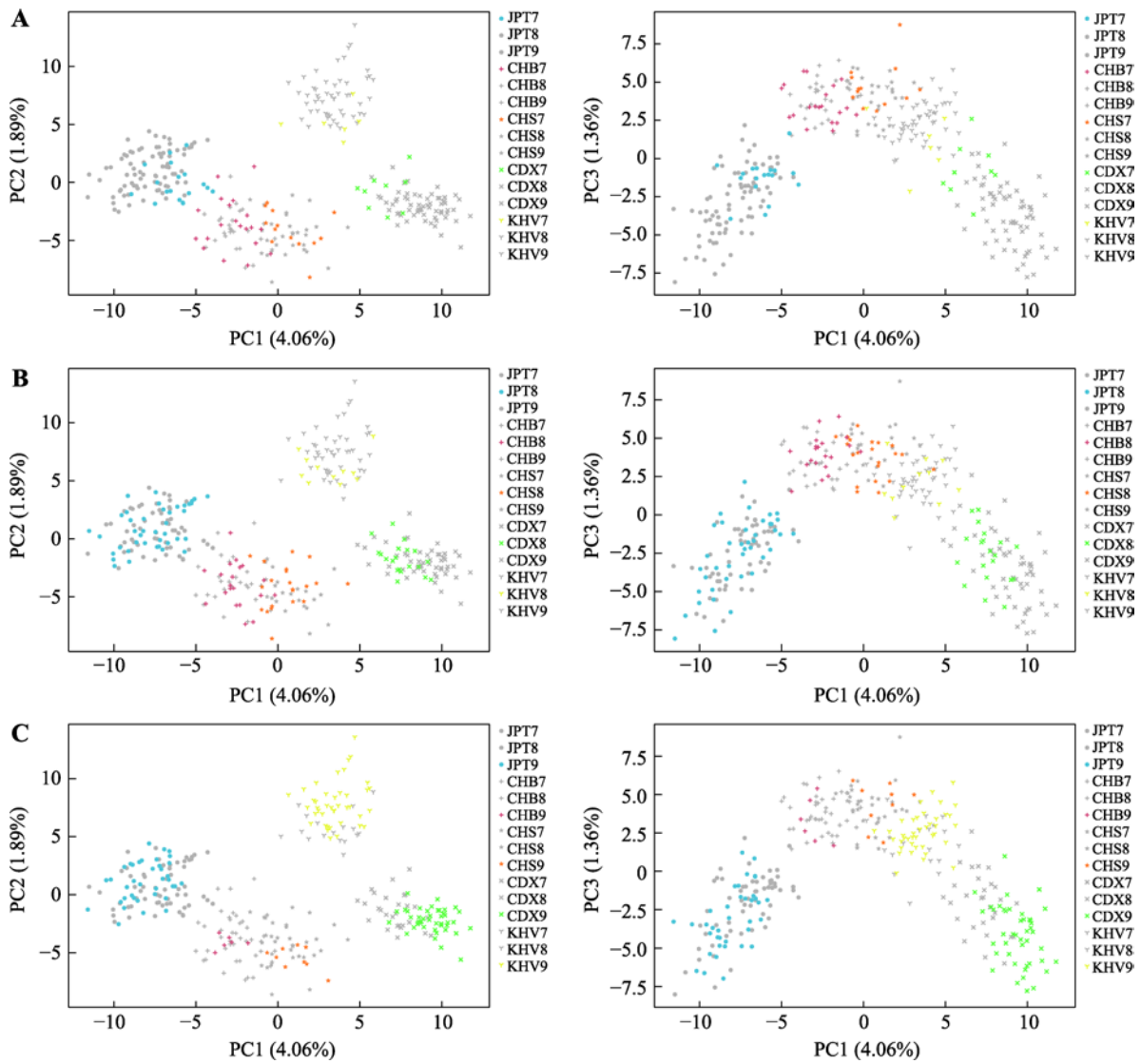

图 6　975 SNPs 的数据集 C PCA 分析结果

Fig. 6　The PCA analysis of 975 SNPs for dataset C

数据集 C_7、C_8、C_9 中个体分别依次标记为彩色，各数据集以外的个体标记为灰色。各颜色代表群体：JPT(蓝色)，CHB(红色)，CHS(橙色)，CDX(绿色)，KHV(黄色)。前三个主成分分别为：PC1=3.21%，PC2=2.12%，PC3=1.36%。A：标记数据集 C_7；B：标记数据集 C_8；C：标记数据集 C_9。

传成分增加而增强。数据集 C_7 (图 6A)和 C_8 (图 6B)中的五个群体聚类为四簇，数据集 C_7 中仅 JPT 和 CHB、CHB 和 CHS 的个体仍有少部分重叠，数据集 C_8 中仅有个别 CHB、CHS 的个体重叠。群体代表性遗传成分增加至 90%以上后(图 6C)五个群体可分别单独聚类。

依据上述 STRUCTURE 分析和 PCA 分析结果，本研究认为群体代表性遗传成分超过个体总遗传成分 80%的个体具有很好的群体代表性，可用于排除群体结构对医学研究的影响。

3　讨论

涉及群体的医学研究中，群体遗传结构的差异可影响结果的正确性和准确性，进行研究时需排除这种影响。而明确采集的样本能否真正代表群体、反映群体遗传结构则是准确排除这种影响的关键。因此，对采集的样本进行遗传结构分析、判断个体

声明血统和实际血统的吻合度、筛选群体代表性个体对于获取正确、准确的研究结果十分必要。

一般而言，研究者们多直接在研究过程中对样本的群体遗传结构进行质控。此方法在有较少特定目标基因片段的研究[32]中十分合理且高效。然而，对于目标基因片段较多，或应用基因芯片或全基因组测序进行大规模基因筛查的研究[33]，不合格的样本可能会导致测序成本的损耗。近年来，公开的多群体全基因组数据库为研究者们提供了新的思路：通过对大量数据进行分析、按照一定标准(如本研究所使用的 F_{ST} 值)进行筛选，找到一组可以反映特定群体之间遗传结构差异、区分群体来源的 AIM，将其作为测序前对群体样本进行预筛选的手段。

本研究使用 F_{ST} 值作为筛选 AI-SNP 的标准。Wright[34]提出的 F_{ST} 值是最常用于表征群体间遗传分化程度的指标之一[27]，其也可应用于控制遗传结构对关联分析的影响[35]。一组高 F_{ST} 值的 AIM 是进行群体遗传结构和遗传关系分析的有力工具。基于 F_{ST} 值筛选的 SNP 进行 Snipper 分析、STRUCTURE 分析和 PCA 分析的结果揭示了东亚群体中的亚结构。结果表明，虽然东亚五个群体两两之间遗传结构复杂，遗传分化程度并不显著，但仍可使用一组包含较多 AIM 的体系加以解析。

STRUCTURE 分析可计算各个聚类中每个个体的遗传成分比例。当定义的群体与其计算得到的聚类十分匹配(或相似)时，各聚类中的血统比例可看作群体的血统比例[36]。此时，STRUCTURE 聚类对应的遗传成分在整个群体的总体成分中占比最大，在每个个体中稳定存在，且与其他群体无关，这种成分可看作该群体的群体代表性遗传成分。高群体代表性遗传成分的个体遗传背景相对单一，可作为该群体一种较固定的遗传背景模式。同时，本研究中具有这类遗传背景模式的个体出现频率也较高，具有一定的群体代表性。综上，本研究设定此类个体作为潜在的群体代表性样本，按群体代表性遗传成分的占比设定了三个阈值：70%、80%、90%，并筛选出相应个体进行 STRUCTURE 分析和 PCA 分析验证。PCA 分析是目前最常用于校正研究中群体分层的方法[13]，可用于验证基于 STRUCTURE 筛选的群体代表性个体是否可靠，同时评估并确定筛选标准。结果表明筛选的个体具有群体代表性，群体代表性遗传成分超过个体总遗传成分 80%可作为筛选群体代表性个体的标准。

需要注意的是，筛选 AIM、分析群体遗传结构以及筛选群体代表性个体依赖于实际群体样本的组成。本研究的样本来自被广泛应用于各类研究的千人基因组数据库，分析这些群体、筛选具有群体代表性的个体可提供更大的实际应用价值。而为了弥补在大陆次级区域内 AIM 分析群体间遗传结构差异研究的缺失，同时证明使用 AIM 核验样本血统的实际应用可行性，本研究选取遗传结构非常复杂的东亚群体作为研究对象。在分析时，尽可能使用更多的 AIM 以得到更准确的群体结构信息，以夯实后续筛选群体代表性个体的数据基础。与既往区分全球群体的研究[20]相比，本研究所使用的 AIM 数量更多，但与同样对大陆次级区域内(欧洲)人口亚结构进行的研究[7]相比，本研究所使用 AIM 的数量则要更少。研究结果表明，即使是遗传背景高度混杂的多个群体，也可使用一组 AIM 解析群体遗传结构并成功筛选出具有群体代表性的个体，这充分说明了本研究方法的可行性，也证明了其应用于各类涉及群体的医学研究中以排除群体结构对医学研究影响的实际价值。

如上所述，此类研究的结论高度依赖于实际群体样本的组成。本研究证明了基于公开数据库中东亚五群体数据筛选的一组 AI-SNP 能在理论上解析遗传结构复杂的群体间遗传结构的差异，并成功依据个体血统差异筛选出群体代表性个体。然而，受到众多的族群种类、庞大的人口基数，以及复杂的人口流动等因素的影响，东亚地区实际的群体遗传结构极端复杂。因此，使用更多不同来源的族群个体真实样本对研究东亚群体间遗传结构的差异是十分迫切且必要的。对于本研究中筛选出的此组 AI-SNP，后续将构建体系并进一步使用来源于各个群体的真实样本进行验证。此外，今后的研究也将基于该体系尽可能补充更多不同群体的样本，以进一步将研究结果扩大化，使其能真正在实际应用中发挥价值。

综上所述，本研究使用 F_{ST} 值筛选的一组 AI-SNP 对遗传结构复杂的东亚五群体进行了遗传结构分析，

基于 STRUCTURE 的结果成功从各个群体中筛选了具有潜在群体代表性的个体。经 STRUCTURE 分析和 PCA 分析的验证，群体代表性遗传成分占个体总遗传成分超过 80%的个体具备良好的群体代表性。本研究的结果表明，使用一组筛选的 AIM 可对研究群体中个体的遗传结构进行解析，可核实样本的声明血统和实际血统的吻合度并成功筛选具有群体代表性的个体，这一方法在排除群体遗传结构差异对医学研究的影响时具备实际应用价值。

参考文献(References):

[1] Hellwege JN, Keaton JM, Giri A, Gao XY, Velez Edwards DR, Edwards TL. Population stratification in genetic association studies. *Curr Protoc Hum Genet*, 2017, 95: 1.22.1–1.22.23. [DOI]

[2] Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, Li S, De Jongh M, Singleton A, Blum MG, Soodyall H, Jakobsson M. Genomic variation in seven Khoe-San groups reveals adaptation and complex African History. *Science*, 2012, 338(6105): 374–379. [DOI]

[3] Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*, 2010, 11(7): 459–463. [DOI]

[4] Gong X, Zhang C, Yiliyasi A, Shi Y, Yang XW, Nuersimanguli A, Guan YQ, Xu SH. A comparative analysis of genetic diversity of candidate genes associated with type 2 diabetes in worldwide populations. *Hereditas (Beijing)*, 2016, 38(6): 544–565.
弓弦, 张超, 伊利亚斯·艾萨, 时瑛, 杨雪唯, 努尔斯曼古丽·奥斯曼, 关亚群, 徐书华. 2 型糖尿病易感候选基因在世界不同人群中的多样性比较分析. 遗传, 2016, 38(6): 544–565. [DOI]

[5] Dai R, Zhang C, Cheng YJ, Chen WL, Li Q, Wang YM. Pharmacogenomics genetic differences between Wa and Blang ethnic groups in Yunnan. *J Kunming Med Univ*, 2020, 41(5): 33–40.
代润, 张婵, 程瑜静, 陈婉璐, 李琦, 王玉明. 云南佤族和布朗族人群药物基因组学基因遗传差异. 昆明医科大学学报, 2020, 41(5): 33–40. [DOI]

[6] Phillips C, Prieto L, Fondevila M, Salas A, Gómez-Tato A, Alvarez-Dios J, Alonso A, Blanco-Verea A, Brión M, Montesino M, Carracedo A, Lareu MV. Ancestry analysis in the 11-M Madrid bomb attack investigation. *PLoS One*, 2009, 4(8): e6583. [DOI]

[7] Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, Selmi C, Klareskog L, Pulver AE, Qi LH, Gregersen PK, Seldin MF. Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genet*, 2008, 4(1): e4. [DOI]

[8] Enoch MA, Shen PH, Xu K, Hodgkinson C, Goldman D. Using ancestry-informative markers to define populations and detect population stratification. *J Psychopharmacol*, 2006, 20(4): 19–26. [DOI]

[9] Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet*, 2000, 67(1): 170–181. [DOI]

[10] Clarke L, Fairley S, Zheng-Bradley X, Streeter I, Perry E, Lowy E, Tassé AM, Flicek P. The international genome sample resource (IGSR): a worldwide collection of genome variation incorporating the 1000 genomes project data. *Nucleic Acids Res*, 2017, 45(D1): D854–D859. [DOI]

[11] 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. *Nature*, 2015, 526(7571): 68–74. [DOI]

[12] Qin PF, Li ZQ, Jin WF, Lu DS, Lou HY, Shen JW, Jin L, Shi YY, Xu SH. A panel of ancestry informative markers to estimate and correct potential effects of population stratification in Han Chinese. *Eur J Hum Genet*, 2014, 22(2): 248–253. [DOI]

[13] Severe Covid-19 GWAS Group, Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, Invernizzi P, Fernández J, Prati D, Baselli G, Asselta R, Grimsrud MM, Milani C, Aziz F, Kässens J, May S, Wendorff M, Wienbrandt L, Uellendahl-Werth F, Zheng TH, Yi XL, de Pablo R, Chercoles AG, Palom A, Garcia-Fernandez AE, Rodriguez-Frias F, Zanella A, Bandera A, Protti A, Aghemo A, Lleo A, Biondi A, Caballero-Garralda A, Gori A, Tanck A, Carreras Nolla A, Latiano A, Fracanzani AL, Peschuck A, Julià A, Pesenti A, Voza A, Jiménez D, Mateos B, Nafria Jimenez B, Quereda C, Paccapelo C, Gassner C, Angelini C, Cea C, Solier A, Pestaña D, Muñiz-Diaz E, Sandoval E, Paraboschi EM, Navas E, García Sánchez F, Ceriotti F, Martinelli-Boneschi F, Peyvandi F, Blasi F, Téllez L,

Blanco-Grau A, Hemmrich-Stanisak G, Grasselli G, Costantino G, Cardamone G, Foti G, Aneli S, Kurihara H, ElAbd H, My I, Galván-Femenia I, Martín J, Erdmann J, Ferrusquía-Acosta J, Garcia-Etxebarria K, Izquierdo-Sanchez L, Bettini LR, Sumoy L, Terranova L, Moreira L, Santoro L, Scudeller L, Mesonero F, Roade L, Rühlemann MC, Schaefer M, Carrabba M, Riveiro-Barciela M, Figuera Basso ME, Valsecchi MG, Hernandez-Tejero M, Acosta-Herrera M, D'Angiò M, Baldini M, Cazzaniga M, Schulzky M, Cecconi M, Wittig M, Ciccarelli M, Rodríguez-Gandía M, Bocciolone M, Miozzo M, Montano N, Braun N, Sacchi N, Martínez N, Özer O, Palmieri O, Faverio P, Preatoni P, Bonfanti P, Omodei P, Tentorio P, Castro P, Rodrigues PM, Blandino Ortiz A, de Cid R, Ferrer R, Gualtierotti R, Nieto R, Goerg S, Badalamenti S, Marsal S, Matullo G, Pelusi S, Juzenas S, Aliberti S, Monzani V, Moreno V, Wesse T, Lenz TL, Pumarola T, Rimoldi V, Bosari S, Albrecht W, Peter W, Romero-Gómez M, D'Amato M, Duga S, Banales JM, Hov JR, Folseraas T, Valenti L, Franke A, Karlsen TH. Genomewide association study of Severe Covid-19 with respiratory failure. *N Engl J Med*, 2020, 383(16): 1522–1534. [DOI]

[14] Foo JN, Tan LC, Irwan ID, Au WL, Low HQ, Prakash KM, Ahmad-Annuar A, Bei JX, Chan AY, Chen CM, Chen YC, Chung SJ, Deng H, Lim SY, Mok V, Pang H, Pei Z, Peng R, Shang HF, Song K, Tan AH, Wu YR, Aung T, Cheng CY, Chew FT, Chew SH, Chong SA, Ebstein RP, Lee J, Saw SM, Seow A, Subramaniam M, Tai ES, Vithana EN, Wong TY, Heng KK, Meah WY, Khor CC, Liu H, Zhang F, Liu J, Tan EK. Genome-wide association study of Parkinson's disease in East Asians. *Hum Mol Genet*, 2017, 26(1): 226–232. [DOI]

[15] Setakis E, Stirnadel H, Balding DJ. Logistic regression protects against population structure in genetic association studies. *Genome Res*, 2006, 16(2): 290–296. [DOI]

[16] Gaspar HA, Breen G. Probabilistic ancestry maps: a method to assess and visualize population substructures in genetics. *BMC Bioinformatics*, 2019, 20(1): 116. [DOI]

[17] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*, 2000, 155(2): 945–959. [DOI]

[18] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 2009, 19(9): 1655–1664. [DOI]

[19] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 2006, 38(8): 904–909. [DOI]

[20] Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato A, Alvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, Carracedo A; SNPforID Consortium. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet*, 2007, 1(3–4): 273–80. [DOI]

[21] Li CX, Pakstis AJ, Jiang L, Wei YL, Sun QF, Wu H, Bulbul O, Wang P, Kang LL, Kidd JR, Kidd KK. A panel of 74 AISNPs: improved ancestry inference within Eastern Asia. *Forensic Sci Int Genet*, 2016, 23: 101–110. [DOI]

[22] Liu j, Liu CC, Ma M, Wang L, Zhao WT, Ma Q, Ji AQ, Liu J, Li CX. The ancestry inference of Chinese populations using 74-plex SNPs system. *Hereditas (Beijing)*, 2020, 42(3): 296–308.
刘杨, 孙昌春, 马咪, 王玲, 赵雯婷, 马泉, 季安全, 刘京, 李彩霞. 74-plex SNPs 复合检测体系在中国人群中的族群推断研究. 遗传, 2020, 42(3): 296–308. [DOI]

[23] Qu SQ, Zhu J, Wang YJ, Yin L, Lv ML, Wang L, Jian H, Tan Y, Zhang RR, Liu YQ, Li F, Huang SC, Liang WB, Zhang L. Establishing a second-tier panel of 18 ancestry informative markers to improve ancestry distinctions among Asian populations. *Forensic Sci Int Genet*, 2019, 41: 159–167. [DOI]

[24] Bulbul O, Speed WC, Gurkan C, Soundararajan U, Rajeevan H, Pakstis AJ, Kidd KK. Improving ancestry distinctions among Southwest Asian populations. *Forensic Sci Int Genet*, 2018, 35: 14–20. [DOI]

[25] Shi CM, Liu Q, Zhao SL, Chen H. Ancestry informative SNP panels for discriminating the major East Asian populations: Han Chinese, Japanese and Korean. *Ann Hum Genet*, 2019, 83(5): 348–354. [DOI]

[26] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 2011, 27(15): 2156–2158. [DOI]

[27] Weir BS, Cockerham CC. Estimating F‐statistics for the analysis of population structure. *Evolution*, 1984, 38(6): 1358–1370. [DOI]

[28] Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*.

2003, 164(4): 1567–87. [DOI]

[29] Earl DA, vonHoldt BM. Structure Harvester: a website and program for visualizing structure output and implementing the Evanno method. *Conserv Genet Resour*, 2012, 4(2): 359–361. [DOI]

[30] Jakobsson M, Rosenberg NA. Clumpp: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 2007, 23(14): 1801–1806. [DOI]

[31] Rosenberg NA. Distructd: a program for the graphical display of population structure. *Mol Ecol Notes*, 2004, 4(1): 137–138. [DOI]

[32] Zhou CX, Li M, Huai C, He L, Qin SY. Study on hereditary susceptibility genetic markers to anti-tuberculosis drug induced liver injury in Chinese population. *Hereditas (Beijing)*, 2020, 42(4): 374–379.
周晨希, 李沫, 怀聪, 贺林, 秦胜营. 中国人群中抗结核药物引发肝损伤的易感基因标记研究. 遗传, 2020, 42(4): 374–379. [DOI]

[33] Sun YD, Tian ZZ, Zhou W, Li M, Huai C, He L, Qin SY. Genome-wide association study on liver function tests in Chinese. *Hereditas(Beijing)*, 2021, 43(3): 249–260.
孙一丹, 田子钊, 周伟, 李沫, 怀聪, 贺林, 秦胜营. 中国人群肝功能检测指标全基因组关联分析研究. 遗传, 2021, 43(3): 249–260. [DOI]

[34] Wright S. The genetical structure of populations. *Nature*, 1951, 15(4): 323–354. [DOI]

[35] Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting *F(ST)*. *Nat Rev Genet*, 2009, 10(9): 639–650. [DOI]

[36] Santos C, Phillips C, Gomez-Tato A, Alvarez-Dios J, Carracedo Á, Lareu MV. Inference of ancestry in forensic analysis II: analysis of genetic data. *Methods Mol Biol*. 2016, 1420: 255–285. [DOI]

(责任编委: 朱波峰)