PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2023,50(11):2739~2752
www.pibb.ac.cn

基于高密度SNP数据的东亚人群遗传结构研究*

顾佳琪[1,2)**] 江 丽[1)**] 徐景怡[1)] 王 寒[1)] 魏以梁[2)***] 李彩霞[1)***]

([1)] 公安部物证鉴定中心，法医遗传学公安部重点实验室，北京市现场物证检验工程技术研究中心，现场物证溯源技术国家工程实验室，北京 100038；

[2)] 江苏省系统发育与比较基因组学重点实验室，江苏师范大学生命科学学院，徐州 221116）

摘要 **目的** 东亚疆域辽阔，民族众多，有着广泛多样的语言。中国34个省级行政区可划分为7个地理分区，人群主要分属世界七大语系。已有研究主要集中在东亚人群的起源、迁徙、融合等遗传历史。本文基于5 147份世界人群个体的高密度单核苷酸多态性（SNP）数据，从地域及语言两个角度研究东亚人群尤其是中国人群与世界其他人群的遗传关系，研究中国人群的遗传关系和遗传结构。**方法** 收集了5 147份世界人群个体的高密度SNP数据，并对其进行质控、合并。通过频率差异分析方法对最终获得的32 789个SNP进行统计学检验，并进一步使用主成分分析、系统发育树、祖先成分分析和*D*检验统计等方法，对东亚人群与世界其他人群的遗传关系，以及中国人群的遗传关系和遗传结构进行研究。**结果** 研究发现东亚人群与非洲、美洲和欧洲人群存在显著差异。中国人群可分为7个亚群，不同人群间的遗传聚类与其地理分布、语系语族和族源历史有很强的相关性。**结论** 本文研究了中国人群与世界人群的遗传关系和差异，并系统研究了中国人群的遗传亚结构。这将丰富东亚人群的群体遗传学、法医遗传学等研究基础，为个体化医疗等工作提供数据支撑。

关键词 群体遗传分析，法医遗传学，东亚人群，遗传关系
中图分类号 Q3，TP312
DOI：10.16476/j.pibb.2022.0441

东亚一般包括中国、日本、韩国、朝鲜和蒙古五个国家。据2018年世界人口网统计，东亚人口约16亿，占全球人口的22%，而中国人口约14亿，是东亚人口的主体。“非洲起源说”认为现代人到达东亚的时间约为5~6万年前，然后经过“南线”为主的路线扩散至整个东亚地区[1-2]。传统遗传标记[3-4]、Y染色体单核苷酸多态性（Y-SNP）和线粒体DNA（mitochondrial DNA，mtDNA）[5]、常染色体SNP[6-7]等研究均表明东亚人群存在明显的南北分化。由于受到来自中亚和欧洲遗传成分的影响，北方人群遗传成分呈现东西走向的变化趋势[8]，并且南北方人群遗传的差异以秦岭淮河和长江为地理分界[9-11]。

中国作为东亚最主要的国家，人群主要分属七个语系：汉藏语系（Sino-Tibetan）、阿尔泰语系（Altaic）、侗台语系（Tai-Kadai）、苗瑶语系（Hmong-Mien）、南亚语系（Austro-Asiatic）、南岛语系（Austronesian）及印欧语系（Indo-European）[12]。Y-SNP单倍群和常染色体SNPs研究发现，东亚各个语系人群之间存在遗传差异，东亚人群的遗传结构与族源历史和语言结构具有对应关系，同一语系人群有聚类倾向[5, 8, 10]。最近有研究者对东亚不同语言人群的精细遗传多样性和混合历史的研究表明，南岛语系和侗台语系人群起源于中国南方，而生活在不同地域的侗台语系人群有着复杂的遗传亚结构[13]。阿尔泰语系人群也具有明显的亚结构，蒙古语族和北方汉族之间存在复杂的遗传混合史[14]。贵州地区的阿尔泰语系人群形成了独特的遗传梯度，其遗传结构受到贵州土著人群的影响并显著有别于居住在西伯利亚南部及东亚北部的阿尔泰语系人群，而毛南族与贵州周围的侗台

* 国家自然科学基金（82171870），法医遗传学公安部重点实验室开放课题（2020FGKFKT01）和江苏省研究生科研与实践创新计划（KYCX21_2597）资助项目。

** 并列第一作者。
*** 通讯联系人。
魏以梁 Tel：15332195528，E-mail：weiyiliang.2013@tsinghua.org.cn
李彩霞 Tel：13810666929，E-mail：licaixia@tsinghua.org.cn

收稿日期：2022-09-15，接受日期：2023-02-13

语系人群聚集在一起[15]。

利用祖先信息性单核苷酸多态性（AISNPs）进行群体遗传结构分析在医学全基因组关联研究（GWAS）和法医生物地理推断中发挥着重要作用[16-20]。主成分分析[21]、系统发育树[22]、频率差异[23-24]和群体祖先成分[25]分析常用于研究群体遗传结构。*f*3/*f*4/*D*检验等统计方法[26]常用于群体进化历史分析。本研究运用这些方法系统地探讨了东亚人群与世界人群的遗传关系，并从语系和地理分区两个角度研究了中国不同人群之间的遗传关系和遗传亚结构。

1 材料与方法

1.1 样本收集与基因分型

本研究使用的5 948名中国个体的DNA样本来自国家科技资源共享服务平台计划项目。本研究已获公安部物证鉴定中心伦理委员会批准（批准号：2021-006），所有参与者均签署了书面知情同意书。

根据Illumina测序仪的标准流程及Novogene（北京）的标准文库制备，使用Illumina Novaseq 6000对3 473份样本进行5×深度以及2 475份样本进行3.5×深度的全基因组测序。对读序（reads）质控后进行合并，并用Burrows-Wheeler算法[27]比对至人类基因组g1k_v37，通过bcftools和sentieon[28-29]对单个样本进行SNP和Indel检测。然后由华大基因的lowpass *v*0.4根据参考面板（REFPANEL_b37_KG）进行填补。最终所有样本的vcf文件由bcftools *v*1.10.2[28-29]合并，共79 556 172个SNP（5×深度）和32 812 390个SNP（3.5×深度）。

1.2 公共数据收集

通过IBM Aspera *v*3.0.0[30]从IGSR[31]（https://www.internationalgenome.org/data-portal/sample）下载了2 504份（7.4×）的vcf格式全基因组数据，包括约8千万个SNP。从文献中[15, 32-36]（厦门大学王传超课题组）获取了592份样本的plink和eigenstart格式数据，由3款Illumina和1款Affymetrix芯片检测，其中Illumina芯片含70万个SNP，Affymetrix芯片含50~60万个SNP。

1.3 数据处理

首先，使用Plink *v*1.9[37]对全基因组测序数据进行质控，参数分别为过滤检出率小于5%（geno95%）、Hardy-Weinberg平衡检验（Hardy-Weinberg equilibrium，HWE）*P*值大于10^{-6}（HWE 1×10^{-6}）及次要等位基因频率（MAF）小于0.02或0.06（根据测序数据中SNP频率分布所设）的SNP（表1）。接着，通过KING *v*2.2.7[38]计算全基因组测序样本间的亲缘关系系数Φ，共删除三级以内（Φ为0.044 2~0.088 4）亲缘关系样本790份，将本实验室检测数据集和公共数据集通过EIGENSOFT *v*6.1.4[21]的mergeit参数合并为世界人群数据集，根据主成分分析对样本进行二次质控，共删除偏离样本3 125份，最终世界人群数据集共5 147份样本（表S1）的32 789个SNP，其中东亚人群包含3 141份样本。由于SNP染色体位置相对固定，我们将所有数据集的SNP编号改成染色体位置信息。

Table 1 Number of SNP in two genotyped datasets

Technology	Number of SNP				Number of post-merger SNP			
	vcf/plink	Geno 95%	HWE (10^{-6})	MAF (threshold)				
Global Screening Array	717 228	/	/	/		543 273[1)]	543 258[2)]	57 428[3)]	32 789[4)]
Illumina wegene Array1	699 537	/	/	/	691 917[5)]			
Illumina wegene Array2	699 537	/	/	/				
Illumina Array	692 122	/	/	/				
Affymetrix Human Origin Array	597 573	/	/	/				
Low-sequencing 3.5×	32 812 390	31 864 701	31 616 923	5 299 581(0.06)	4 415 699[6)]	3 108 934[7)]		
Low-sequencing 5×	79 556 172	76 344 414	76 188 708	6 546 851(0.02)				
Low-sequencing 7.4×	81 370 339	81 357 018	77 741 088	6 851 772(0.02)				

[1)]Number of intersecting SNP in Global Screening Array and Illumina wegene Array. [2)]Number of intersecting SNP in Global Screening Array, Illumina wegene Array and Illumina Array. [3)]Number of intersecting SNP in Global Screening Array, Illumina wegene Array, Illumina Array and Affymetrix Human Origin Array. [4)]Number of intersecting SNP in two genotyped datasets. [5)]Number of intersecting SNP in two Illumina wegene Array. [6)]Number of intersecting SNP in low-sequencing 3.5× and low-sequencing 5×. [7)]Number of intersecting SNP in low-sequencing 3.5, low-sequencing 5× and low-sequencing 7.4×.

1.4　人群遗传多样性和遗传结构分析

为评估32 789个SNP是否在研究人群中存在遗传多样性，本研究使用Plink *v*1.9计算了这些SNP在世界亚人群中的MAF（表S2），并通过R *v*4.0.2进行单因素方差分析。随后，为了研究东亚人群与世界其他人群之间的遗传关系，使用R *v*4.0.2的ggplot2、gmodels包计算和可视化世界亚人群频率主成分图。

1.5　中国人群的聚类、系统发育及祖先成分分析

对东亚人群进行主成分分析，研究不同地理分区和语言背景人群间聚类情况。利用Plink *v*1.9计算东亚人群数据集中每个亚人群共32 789个SNP的等位基因频率，以Affymetrix Human Origins Array注释文件为标准，dbsnp版本为142。随后使用TreeView *v*1.6.6[39]可视化Phylip *v*3.695[40]基于SNP的等位基因频率绘制东亚人群系统发育树。使用EIGENSOFT *v*6.1.4的smartpca参数计算东亚3 143份和中国2 943份样本的特征向量与主成分特征值，并通过R *v*4.0.2的ggplot2包绘制主成分图。最后通过Plink *v*1.9的indep-pairwise参数，以滑动窗口200、步长20、r^2为0.4进行连锁不平衡分析，SNP数过滤为28 481个，利用ADMIXTURE *v*1.3.0[25]进行东亚人群的遗传混合和祖先成分研究，其中K值范围2~14，循环数随机，以Ancestry Painter *v*5[41]可视化分析结果。通过分析交叉验证错误率及不同K值时群体和个体层面的族群成分以确定最佳K值。

1.6　中国人群间基因交流的预测及特异SNP功能分析

使用ADMIXTOOLS *v*7.0.2[26]的默认参数对东亚人群进行*D*检验，研究人群间是否有基因交流事件的发生。此外，在分析时也关注了中国人群特异SNP，通过ANNOVAR[42]和DAVID *v*6.8[43]对东亚人群数据集进行基因注释和KEGG信号通路的富集。

2　结　　果

2.1　世界人群数据集

本研究中的世界人群数据集包括5 147份样本的32 789个SNP，人群来自非洲、美洲、欧洲、南亚和东亚，东亚人群包括日本、越南及中国20个省份人群，共计48个人群的3 141份样本（表S1）。

2.2　遗传多样性和人群遗传结构分析

首先，计算32 789个SNP在世界77个人群中的等位基因频率（表S2），去除MAF值为NA（缺失值）的SNP获得32 638个SNP数据，基于频率数据进行单因素方差分析（表2），研究人群间是否存在遗传多样性。结果显示，*P*-value远小于0.05的Bonferroni校准值[44]（0.05/77），且*F*-value大于*P*-value。说明在α为0.05/77的情况下，77个人群中至少有一组人群的32 638个SNP的等位基因频率存在显著差异。

Table 2　Analysis of variance of MAF in 77 world populations

	Df	Sum-Sq	Mean-Sq	*F*-value	*P*-value
Variable	76	119	1.569 9	82.09	$<2\times10^{-16}$***
Residuals	2 513 049	48 059	0.019 1		

其次，为了研究东亚人群与世界其他人群中的遗传关系，本研究进行了主成分分析（图1）。结果显示，东亚、非洲、欧洲三大洲际人群各自聚为一簇，南亚和美洲人群居于洲际人群之间，南亚的尼泊尔人群与东亚人群分布较近，其中，尼泊尔夏尔巴族、拉伊族与藏族人群分布较近，这与研究人群的地理分布是一致的。并且东亚人群沿PC1呈东西遗传差异，沿PC2呈南北遗传差异。

2.3　东亚人群的遗传分化研究

为了系统研究东亚人群之间的遗传距离与遗传关系，本研究汇集了来自日本、越南及中国20个省份的48个人群共3 141份样本的32 789个SNP。通过系统发育分析和主成分分析方法，对中国汉族、少数民族的遗传数据与语系语族、地理分区的相关性进行研究。本研究中的东亚人群来自汉藏语系、阿尔泰语系、侗台语系、南亚语系和苗瑶语系五个语系人群。因日语和朝鲜语的语系归属存在争议，故单独列出。

首先计算等位基因频率进行单因素方差分析（表3），结果显示*P*-value远小于0.05的Bonferroni校准值（0.05/48），且*F*-value远大于*P*-value。说明在α为0.05/48的情况下，48个人群中至少有一组人群的32 638个SNP的等位基因频率存在显著差异。随后根据上述人群等位基因频率绘制系统发育树（图2），探索东亚人群间的遗传关系。结果显示，东亚人群的遗传聚类与各自语系语族分类是一致的，汉藏语系的两个语族人群聚类且汉语族人群呈现南北之分，侗台语系的三个语族人群聚类，并与苗瑶和南亚语系人群有较近的遗传关系。其中，北方汉语族人群与中国北方的藏缅语族、阿尔

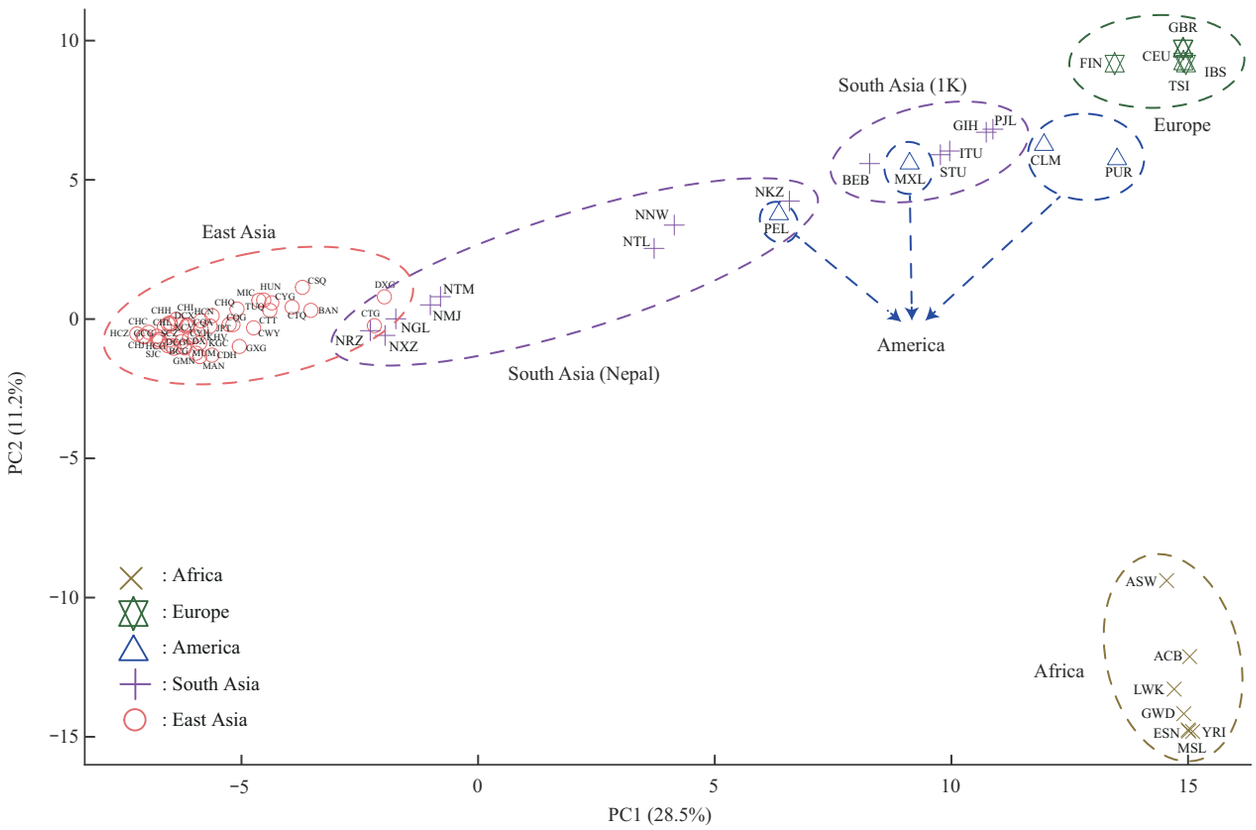

Fig. 1　Principal component analysis based on 32 638 SNP allele frequency of 77 world populations

Table 3　Analysis of variance of MAF in 48 East Asia populations

	Df	Sum-Sq	Mean-Sq	*F*-value	Pr (>F)
Variable	47	31	0.663 8	35.12	$<2\times10^{-16}$***
Residuals	1 566 576	29 607	0.018 9		

泰语系人群及日本人、朝鲜人相邻，而南方汉语族人群则与苗瑶、侗台和南亚语系人群相邻。此外，研究观察到相同民族分布在不同区域，表现出与当地人群更近的遗传关系，例如：四川甘孜羌族与藏缅语族其他人群相邻，而四川阿坝羌族与南方汉族和少数民族相邻。进一步*D*检验结果显示四川甘孜羌族中存在北方少数民族显著基因流，四川阿坝羌族则有南方汉族及少数民族的显著基因流信号（图S1a）。

分别对东亚3 141份样本和中国2 943份样本的遗传数据进行主成分分析，以揭示中国人群之间的遗传关系与其地理分区、语系语族的关系。研究发现少数民族与汉族人群存在遗传差异（图3，S2，S3），且不同人群之间的遗传聚类与其地理分区、语系语族有很强的相关性。少数民族可分为如下七个亚群（图3a，S2a，S3a）：南部、西南、东部、中部、东北、北部和西北。其中，中国南部、西南、东部和中部地区的苗瑶、侗台和南亚语系人群聚为一簇，而日本人和中国东北地区的通古斯语族人群聚为另一簇，与中国北部的阿尔泰语系人群以及中国西北、西南地区的藏缅语族人群相邻。贵州地区的仡佬族（仡央语族）、土家族（藏缅语族）和云南佤族（孟高棉语族），以及江西和浙江畲族、湖南瑶族（苗瑶语系）聚类，位于这两个遗传簇之间。另外，研究观察到同语族人群的遗传聚类受到其地理分区的影响（图2，3a，S2a，S3a），例如：生活在广西和贵州的毛南族、广西和越南的京族表现出了较近的遗传亲和力，但贵州和广西的仡佬族之间的遗传差异较大。以及研究发现同一地区人群的遗传分布与其语系语族有关（图3a，S3a），例如：云南佤族与北方藏缅语族人群相邻，云南傣族则与南方侗台语系人群相邻。进一步*D*检验也揭示贵州仡佬族与北方少数民族的遗传亲和力更强，广

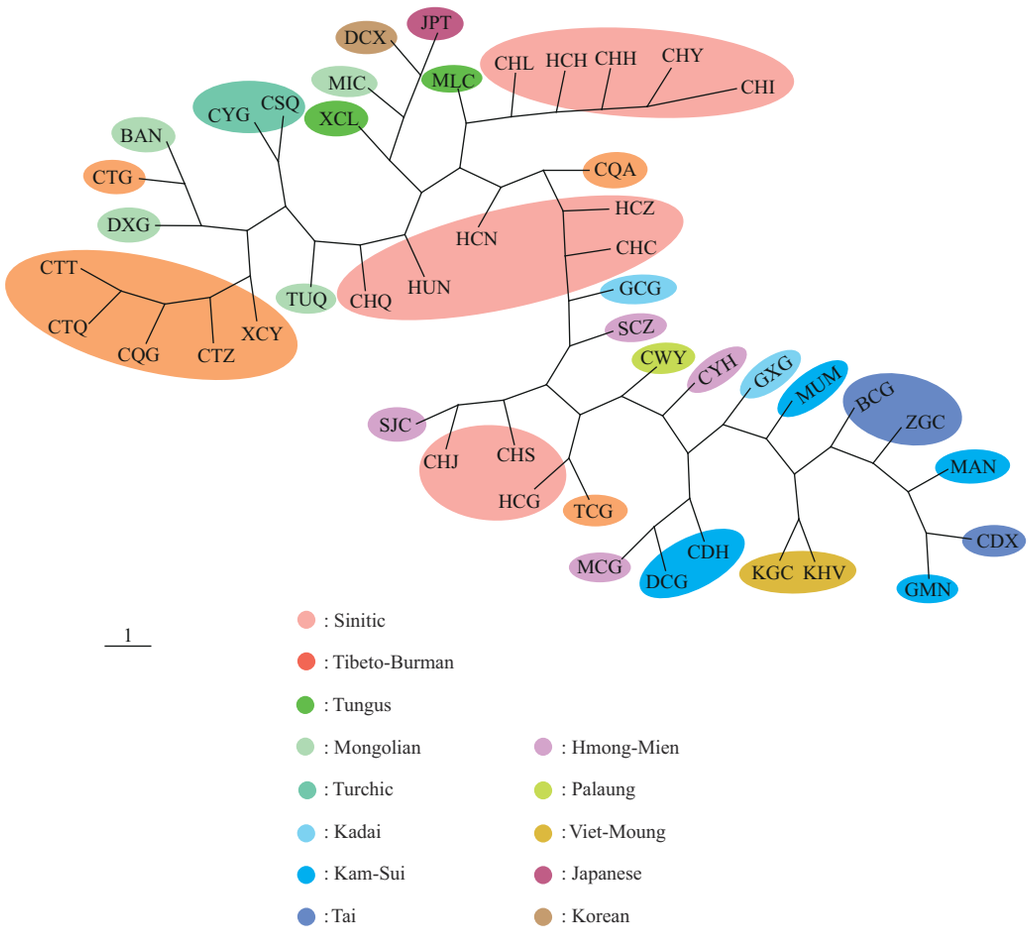

Fig. 2　Phylogenetic analysis based on allele frequencies of 32 638 SNP in 3 141 East Asians

西仫佬族与同语系南方人群的遗传亲和力更强（图S1b），并且云南佤族中存在藏缅语族人群的显著基因流，而云南傣族中存在南方汉族和少数民族的显著基因流信号（图S1d）。

汉族人群的分布较为居中，表现出与少数民族分布类似的三个遗传簇，可分为如下六个亚群（图3b，S2b，S3b）：南部、西南、东部、中部、北部和西北。其中，中国西南、东部和南部地区的汉族聚为一簇，中国北部和西北地区的汉族聚为另一簇，而居住在中国中部地区的汉族则位于这两个遗传簇之间。

2.4　东亚人群的遗传结构研究

为了进一步揭示我国人群之间的遗传差异度和遗传亚结构，本研究基于3 141份样本的32 789个SNP分型数据进行遗传祖先成分分析（ADMIXTURE）。本研究应用ADMIXTURE进行东亚人群的遗传祖先成分分析（K=2~14，图4b~f），其中，K=6时的结果出现了最低的交叉验证误差（0.566 3）（图4a），表明6个祖先成分可以解释研究人群的遗传祖先成分（图4f）。分别为：a. 主要存在于日本和朝鲜人群的红色遗传祖先成分；b. 主要存在于东亚北方阿尔泰语系人群的紫色遗传祖先成分；c. 主要存在于中国西北和西南地区藏缅语族人群的黄色遗传祖先成分；d. 主要存在于东亚南北方汉语族人群的湖蓝色遗传祖先成分；e. 主要存在于东亚南方侗台和南亚语系人群的绿色遗传祖先成分；f. 主要存在于苗瑶语系人群的蓝色遗传祖先成分。其中，中国西北、北部的蒙古语族和中国东北、北部的通古斯语族人群遗传祖先成分混合模拟最相似，由41.2%阿尔泰语系、20.3%藏缅语族、12.5%汉语族和12.1%侗台语系人群相关的成分构成，该结果与主成分分析结果相印证（图3a，S2a，S3a）。中国西北和西南地区藏缅语族三个语支人群的遗传祖先成分混合模拟最相似，由46.1%藏缅语族、23.5%汉语族、16.1%侗台语系和5.5%阿尔泰语系人群相关的遗传祖先成分组成，但藏语支和羌语支人群的阿尔泰语系人群相关遗传祖先成分多于彝语支人群，且侗台语系人群相关遗传祖先

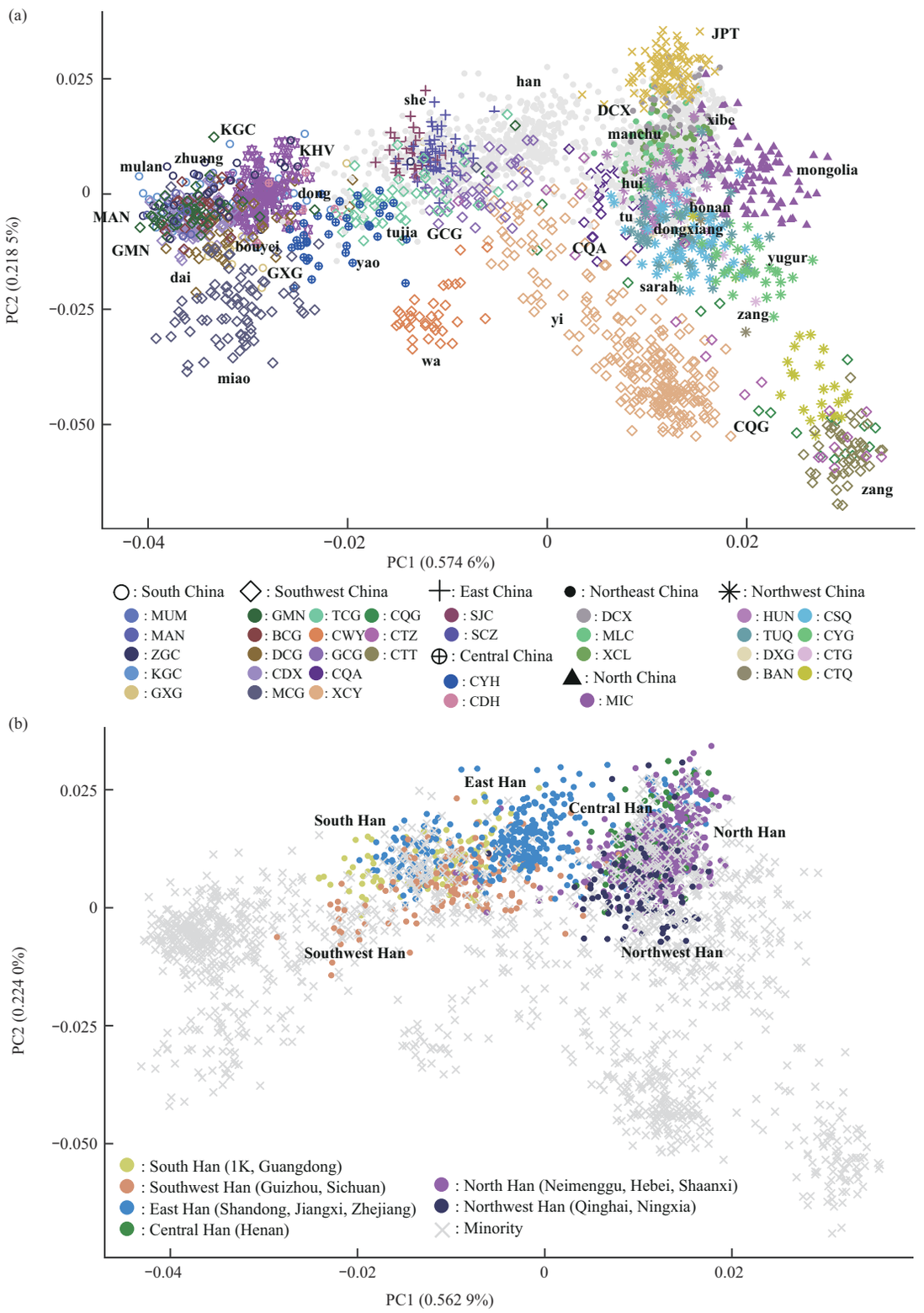

Fig. 3　Principal component analysis of East Asians

(a) Principal component analysis of 32 789 SNPs from 3 141 individuals in East Asia. (b) Principal component analysis of 32 789 SNPs from 2 943 individuals in China.

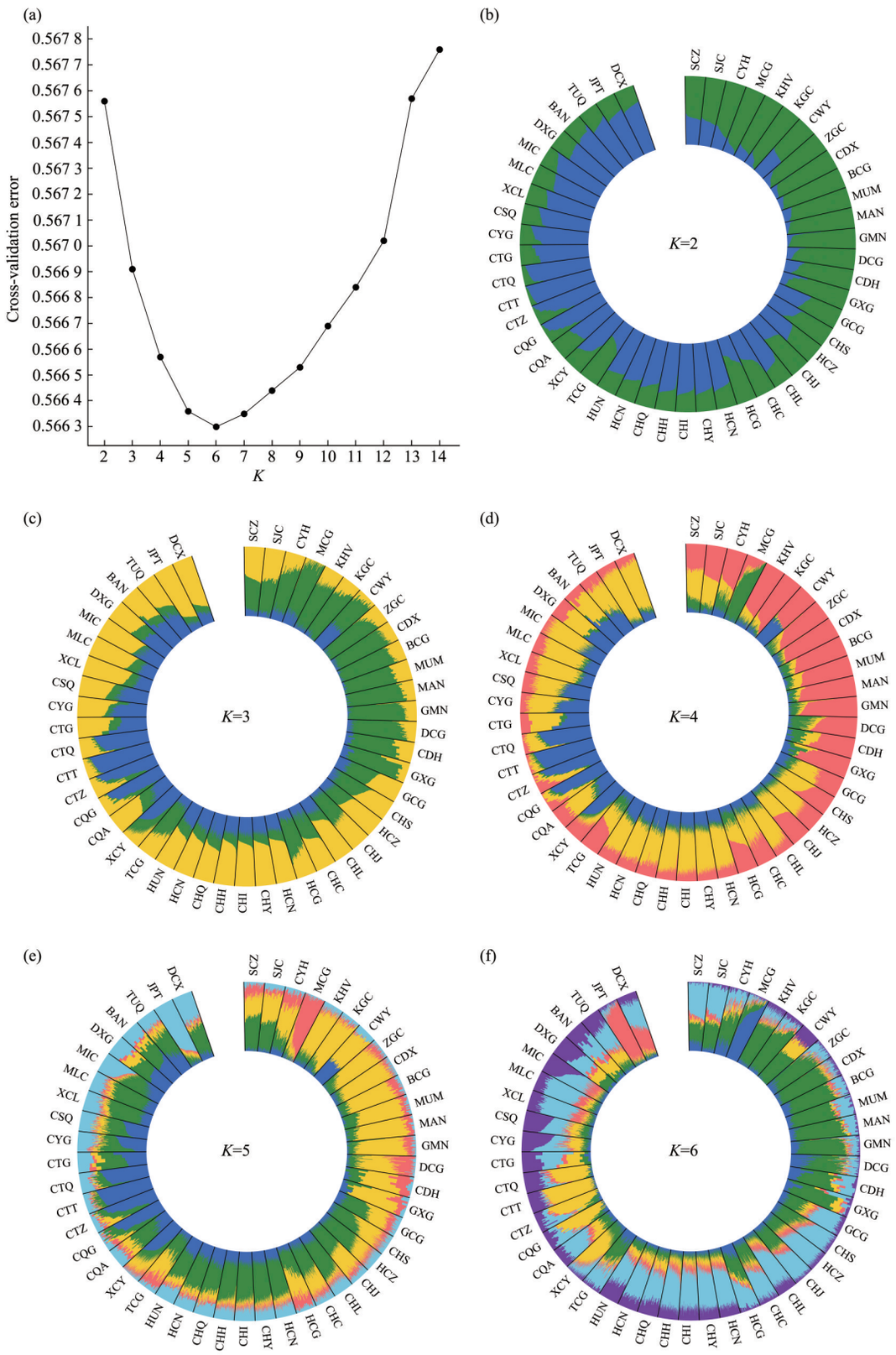

Fig. 4　ADMIXTURE results of 28 481 SNP in East Asians

(a) Estimation of the number of groups (ranging from 2 to 14) for *K* values in ADMIXTURE. (b) Distribution of the 2 ancestry components in 48 East Asia populations. (c) Distribution of the 3 ancestry components in 48 East Asia populations. (d) Distribution of the 4 ancestry components in 48 East Asia populations. (e) Distribution of the 5 ancestry components in 48 East Asia populations. (f) Distribution of the 6 ancestry components in 48 East Asia populations.

成分少于彝语支人群。并且，研究观察到汉语族人群存在南北方遗传结构差异，北方汉语族人群拥有更多阿尔泰语系人群相关成分，南方汉语族人群则拥有更多东亚南方的侗台、南亚等语系人群相关成分。中国西南和南部的侗台语系三个语族人群遗传祖先成分最相似，由59.7%侗台、19.3%汉语族和11.0%苗瑶语系人群相关遗传祖先成分构成，而侗水语族人群拥有更多苗瑶语系人群相关成分。此外，本研究还对东亚人群数据集的32 789个SNP进行了注释，从中选取外显子和剪接位置共1 061个（去重后为951个基因）并富集KEGG相关信号通路（表S3，图S4）。在选择的14个信号通路中有10个基因富集到黏合连接通路，11个基因富集到胆汁分泌通路，10个基因富集到药物代谢其他酶通路，9个基因富集到药物代谢-细胞色素P450通路以及92个基因富集到代谢通路。其中，研究发现与各种代谢有关的通路中均有*UGT1A6*基因，而*PTPRJ*基因在黏合连接通路中似乎发挥着更重要的作用，且11号染色体48145375位置和2号染色体234601669、234602191位置的SNP分别与*PTPRJ*和*UGT1A6*基因关联。通过计算这两个基因在东亚人群数据集中的等位基因频率，本研究发现与黏合链接通路相关的*PTPRJ*基因在中国东部汉族和西北少数民族中的突变率最低，而*UGT1A6*基因在贵州侗族和南方汉族中的突变频率最低，其次是浙江、江西畲族和西南地区人群，在湖南侗族和北方汉族中的突变频率最高（图5）。

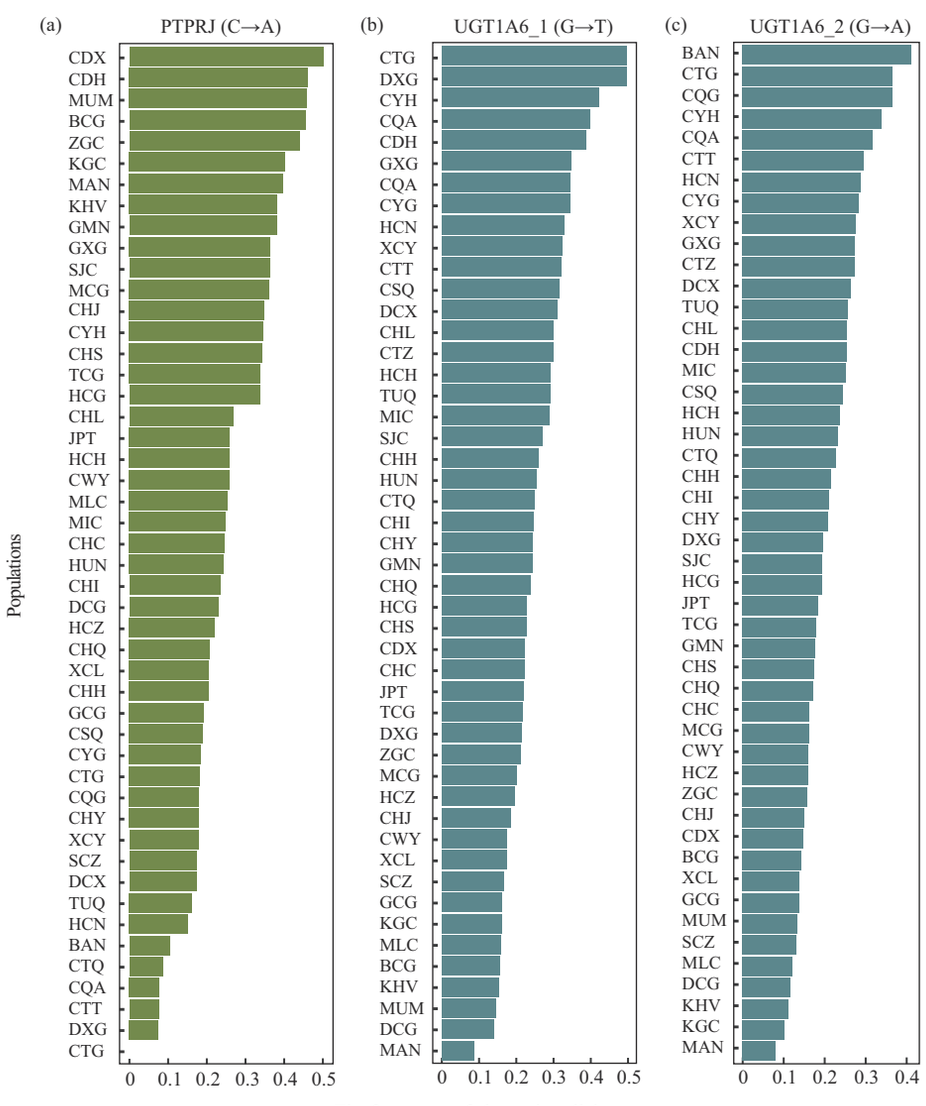

Fig. 5 The distribution of ALT in two genes in East Asians

(a) The frequency of alternative allele in *PTPRJ* in 48 East Asia populations. (b, c) The frequency of alternative allele in *UGT1A6* in 48 East Asia populations.

3 讨　论

群体遗传学、分子人类学等大量研究揭示了人类的演化和迁徙历史，以及自然选择对人类演化的影响。人类在不同环境中长期生活，适应进化、遗传漂变等因素使得不同人群之间产生遗传分化和差异，最终形成现代人群的遗传结构[45]。基于前人从地理分区角度[7-8, 46]和语系语族角度[13-15, 32, 36, 47-48]对东亚人群的遗传结构和历史演化等研究结果，本研究将实验室数据与公开发表的数据合并开展深入研究，共计获得5 147份样本的32 789个SNP，研究人群来自非洲、美洲、欧洲、南亚及东亚（日本、越南及中国七大地理分区），统计学检验结果表明这些SNP在上述人群中存在遗传多样性。进而，本研究从地理分区和语系语族两个角度对世界人群尤其是以中国为主的东亚人群进行了群体基因组学分析，发现东亚人群的遗传特征与非洲、美洲、欧洲、南亚人群之间存在较大差异，与南亚尼泊尔人群差异较小。同时，中国人群可分为七个亚群（即中国七大地理分区），汉族和少数民族表现出了不同的遗传分化和遗传亚结构，不同人群的遗传聚类不仅受到语系语族还有地理分布的影响，即同语系或同语族人群表现出相近的遗传关系，同地区人群之间也存在较强的遗传亲和力。

3.1 东亚人群与世界其他人群的遗传关系

现代人（*Homo sapiens sapiens*）大约20万年前起源于非洲东部，约6万年前走出非洲，逐渐分散到世界各地[49]。本研究从SNP多样性和世界人群遗传关系的分析中发现，东亚人群内部表现出较强的遗传相关性，人群呈南北和东西遗传分化（图1）。先前的Y染色体DNA研究表明在东亚男性中占比最大O-M175单倍群及下游支系来源中国南方和东南亚人群[50-55]，基于常染色体SNP的研究也证实越南人群对东亚人群的遗传贡献更大[8]。本研究还发现，散居在尼泊尔的夏尔巴人与中国北方藏族聚为一簇（图1），与藏族相比，夏尔巴人包含更多的南亚成分，这与已有研究结果[56]一致。

3.2 中国不同地域及语系人群间的遗传连续性与异质性

东亚位于欧亚大陆东部，是研究人类起源历史及民族演化的重要地区之一。中国作为东亚最主要的国家，主要分为七大地理分区。ChinaMAP一期研究显示出中国多区域人群遗传背景的多样性和复杂性[57]，本研究通过对东亚人群SNP数据进行群体遗传学分析，发现不同地域和语系人群的遗传变异特征与中国历史上的人口迁移和变迁有关。少数民族的遗传聚类与人群分布的地理位置一致，可聚为南部、西南、东部、中部、东北、北部和西北这七个地理分区，并表现为南、北和中部这三个遗传簇（图3a）。本研究观察到不同人群之间的遗传聚类与地理分区和语系语族有关，首先，同地区同语族人群之间遗传关系较为相近，例如：位于中国西南、南部和中部地区的侗台语系三个语族人群有着相似的遗传祖先成分并表现出了较强的遗传亲和力，且与同地区的苗瑶、南亚语系人群分布较近（图2，3a，4f，S2a，S3a）。Y染色体DNA研究表明单倍群D-M174在东亚北部藏缅群体中高频分布[50, 52, 58-59]，D下分支D1在藏语支、羌语支和彝语支人群中广泛分布[46, 60]。本研究在对东亚人群进行系统发育和主成分分析时也发现藏缅语族三个语支人群聚为一簇（图2，3a）。另有研究[46, 61]证实，现代藏缅语族人群是由旧石器时代携带单倍群D-M174人群与黄河流域中部携带单倍群O2a2b1a1a-F5的仰韶文化人群混合形成，随后仰韶人向西迁移，形成了居住在中国西北的氐羌人，并沿藏彝走廊大规模迁移至西藏、云南等地区[60]。本研究也发现生活在中国北方的汉藏语系人群之间有较近的遗传关系，例如：东亚人群系统发育结果显示（图2）中国西南和西北的藏缅语族人群与中国青海、宁夏汉族相邻。其次，同语族人群的遗传关系和结构会受到地理分布的影响，从而表现出不同的遗传模式。例如：四川阿坝羌族与南方人群有较近的遗传关系，而四川甘孜羌族却与北方少数民族表现出了较强的遗传亲和力（图2，S1a），这可能是因为人群所在地区的海拔以及与周围其他人群发生基因交流事件导致的。同时，*D*检验结果（图S1b）也揭示贵州仡佬族中存在中国西南和西北地区的汉藏和阿尔泰语系人群的显著基因流信号，而广西仡佬族与中国西南和南部的侗台及南亚语系人群发生基因交流事件，这与系统发育及主成分分析的结果一致（图2，3）。但藏族人群因地区差异表现出遗传亚结构（图3a，S2a，S3a），较甘肃藏族（CTG）来说，青海藏族（CTQ）与西藏、四川甘孜藏族的遗传亲和力更强（图S1c）。最后，同地区不同语族人群之间存在不同的遗传关系，例如：云南地区的佤族（孟高棉语族）与傣族（壮傣语族）分布较远（图2，3a），本研究在云南佤族中

检测到了西藏和青海藏族的基因流信号，在云南傣族中检测到了广西毛南和壮族的基因流信号（图S1d），这一点在人群主成分分析和祖先成分分析中也可以观察到（图3a，4f，S2a，S3a）；而甘肃保安族和东乡族、青海土族（蒙古语族）以及青海撒拉族、甘肃裕固族（突厥语族）聚类（图3a，S2a，S3a）且遗传祖先成分较为相似（图4f），拥有20%~30%与藏缅语族人群相关遗传祖先成分，内蒙蒙古族以及辽宁满族、锡伯族（通古斯语族）聚类（图3a）且遗传祖先成分较为相似（图4f），拥有10%~30%与日本和朝鲜人群相关遗传祖先成分。

线粒体和Y染色体DNA研究[13，35-36，48，62-64]表明，南北方汉族存在遗传差异，北方汉族优势单倍群为D4、A（线粒体）和O2-M122、C-M130（Y染色体），而南方优势单倍群为B4、F1（线粒体）和O2-M122、O1-F265（Y染色体），且常染色体DNA研究[7，57]也表明汉族存在南北方的遗传分化差异。本研究观察到汉族人群与少数民族存在遗传差异（图3），汉族的分布较为居中，并表现出与少数民族类似的三个遗传簇，可分为南部、西南、东部、中部、北部和西北这六个亚群（图3b）。同时，人群的迁徙还会导致人群遗传融合事件的发生，本研究中汉族人群的六个亚群簇分布较为分散（图3b，S2b，S3b），说明不同地区的汉族人群受到周围少数民族的遗传影响。相关研究[55，64-65]发现单倍群A在一些南方地区（安徽和江苏）以及单倍群F1在中国西北地区（青海）也有较高频率的分布，单倍群O2-M122在南北方汉族中均有较高的分布频率。此外，本研究发现不同人群之间的遗传关系与他们的地理分布有关，例如：南方汉语族人群与中国南方侗台语系和苗瑶语系人群分布较近，北方汉语族人群与中国北方的藏缅语族和阿尔泰语系人群、日本和韩国人分布较近（图2，3），这也印证了王传超等前期Y染色体DNA文献报道[13，35-36]的汉族人中有高频分布的O2-M122单倍群，其下游支系O2a2b1a1-M117在中国北方的藏缅语族人群中广泛分布，以及O2a2a1a2-M7在中国南方的苗瑶语系人群中高频出现[51，55]。并且，本研究在东亚人群遗传结构研究中观察到南方汉语族人群拥有更多侗台语系和苗瑶语系人群相关的遗传祖先成分，北方汉语族人群拥有更多藏缅语族和阿尔泰语系人群相关的遗传祖先成分（图4f），这与线粒体单倍群研究[13，48，65]证实的单倍群B4和F1在贵州侗族和苗族、广西布依族中有较高的分布频率，D4和A在甘肃东乡族、保安族（蒙古语族）和青海裕固族（突厥语族）中有较高的分布频率，以及单倍群F1在宁夏回族和青海撒拉族中有较高的分布频率是一致的。

3.3 中国不同人群的特异性SNP

SNP分析有助于解释群体的表型差异，不同群体和个体对疾病，特别是对复杂疾病的易感性以及对药物的敏感性。通过对东亚人群数据集32 789个SNP的注释和相关基因通路分析，本研究发现*PTPRJ*和*UGT1A6*这两个基因在东亚人群中的等位基因频率存在地区差异性，这与相关研究报道[66-67]酪氨酸磷酸酶*PTPRJ*基因突变率与血型类型相关，即O型血人群的*PTPRJ*基因突变率较低，尤其是中国北方少数民族及中国东部人群等地O型血的人最多[68]，以及早期研究发现*UGT1A6*基因在中国汉族、侗族和畲族人群中的分布存在差异[69]是一致的。近年来，频率差异分析、机器学习等算法[18-20，70-80]逐渐用于人群特异性SNP或祖先信息标记（ancestry informative marker）筛选和人群遗传推断模型构建。例如：陈华等[19]利用F_{st}值和AIM-SNPtag筛选了中国汉族人、日本人和韩国人的AISNP，并构建了人群遗传推断模型，Oscar Gaggiotti等[79]使用回归分析（逻辑回归和支持向量机）、决策树（随机森林和XGboost）等方法筛选AISNP并构建推断模型。由于本研究侧重东亚人群之间遗传关系和遗传结构的研究，并对合并质控后的SNP进行注释和功能分析，尚未进行系统的东亚人群AISNP的筛选和遗传推断模型构建研究。

总之，群体遗传结构研究对于医学和法医学都具有重要作用。在医学领域，了解人群遗传结构差异，可以避免筛选出假阳性的疾病关联基因位点。在法医领域，基于DNA的族群地域分析，可以缩小嫌疑人的范围，为侦查提供线索[81]。首先，本研究通过将东亚人群的遗传数据与其地理分区和语系语族分布进行亚人群的遗传关系和结构分析，并揭示了东亚与世界其他人群、东亚人群之间的遗传关系和遗传亚结构，这些将为后续东亚人群的群体遗传学、法医遗传学等研究奠定基础并提供数据支撑。其次，本研究所用数据集的人群未覆盖全部少数民族，而是以不同地理分区和语系语族的代表人群为主，未来需要继续增加人群数据，并尽量使用相同的检测平台。在研究东亚人群遗传关系时

（图3），主成分的前两个维度解释度有限，本研究通过结合主成分的多维度结果（图S2，S3）和*D*检验（图S1）进行进一步分析，后续需增加人群测序数据，增加数据合并后的位点数量，进一步提高主成分分析的差异解释度，并实现更加精细的遗传结构分析。最后，在后续不同人群AISNP挑选时，可以基于本研究获知的人群遗传结构和质控筛选的AISNP，采用传统的AISNP筛选方法和最新的机器学习算法[18-20, 70-80]，构建东亚人群遗传推断模型。

4 结　论

本研究通过对收集到的5 147份世界人群个体的高密度SNP数据进行群体基因组学分析，揭示了东亚人群尤其是中国人群与世界其他人群的遗传关系，并系统研究了中国人群的遗传关系和遗传结构，为丰富东亚人群法医遗传学等研究奠定了基础。同时，经过质控筛选的AISNP，实现了地理分区等层面的亚结构分析，可以将东亚人群按中国七大地理分区和五大语系分别区分开。在后续的研究工作中，将增加人群数据，结合多种群体遗传学分析方法，并应用最新机器学习算法，构建东亚人群遗传推断模型，为个体化医疗等工作提供数据支撑。

附件　见本文网络版（http://www.pibb.ac.cn或http://www.cnki.net）：

PIBB_20220441_Figure S1.pdf
PIBB_20220441_Figure S2.pdf
PIBB_20220441_Figure S3.pdf
PIBB_20220441_Figure S4.pdf
PIBB_20220441_Table S1.xlsx
PIBB_20220441_Table S2.xlsx
PIBB_20220441_Table S3.xlsx

参　考　文　献

[1] Su B, Xiao J, Underhill P, *et al*. Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age. Am J Hum Genet, 1999, **65**(6): 1718-1724

[2] Wang C C, Li H. Inferring human history in East Asia from Y chromosomes. Investig Genet, 2013, **4**(1): 11-21

[3] 杜若甫. 我国的人类群体遗传学研究. 生物学通报, 1997(7): 9-12
Du R F. Biol Bull, 1997(7): 9-12

[4] 杜若甫, 肖春杰. 从遗传学探讨中华民族的源与流. 中国社会科学, 1997(4): 138-145
Du R F, Xiao C J. Social Sciences in China, 1997(4): 138-145

[5] 文波. Y染色体、mtDNA多态性与东亚人群的遗传结构[D]. 上海: 复旦大学, 2004
Wen B. Y-Chromosome, mtDNA Polymorphism and Genetic Structure of East Asian Population[D]. Shanghai: Fudan University, 2004

[6] Chen J, Zheng H, Bei J X, *et al*. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am J Hum Genet, 2009, **85**(6): 775-785

[7] Xu S, Yin X, Li S, *et al*. Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am J Hum Genet, 2009, **85**(6): 762-774

[8] Abdulla M A, Ahmed I, Assawamakin A, *et al*. Mapping human genetic diversity in Asia. Science, 2009, **326**(5959): 1541-1545

[9] 肖春杰, 杜若甫, Cavalli-Sforza L L, 等. 中国人群基因频率的主成分分析. 中国科学C辑:生命科学, 2000(4): 434-442+449
Xiao C J, Du R F, Cavalli-Sforza L L, *et al*. Chinese Science Series C: Life Science, 2000(4): 434-442+449

[10] Zhang F, Su B, Zhang Y P, *et al*. Genetic studies of human diversity in East Asia. Philos Trans R Soc Lond B Biol Sci, 2007, **362**(1482): 987-995

[11] 任红雨. 秦岭—淮河如何分割中国人的南北意识. 科学大观园, 2018(4): 64-67
Ren H Y. Grand Garden of Science, 2018(4): 64-67

[12] 蔡晓云. Y染色体揭示的早期人类进入东亚和东亚人群特征形成过程[D]. 上海: 复旦大学, 2009
Cai X Y. Y Chromosome Reveals the Formation Process of Early Human Entering East Asia and East Asian Population Characteristics[D]. Shanghai: Fudan University, 2009

[13] Wang M G, He G, Zou X, *et al*. Reconstructing the genetic admixture history of Tai-Kadai and Sinitic people: insights from genome-wide SNP data from South China. J Syst Evol, 2023, **61**(1): 175-178

[14] He G, Wang M, Zou X, *et al*. Extensive ethnolinguistic diversity at the crossroads of North China and South Siberia reflects multiple sources of genetic diversity. J Syst Evol, 2023, **61**(1): 230-250

[15] Chen J, He G, Ren Z, *et al*. Fine-scale population admixture landscape of Tai-Kadai-speaking Maonan in Southwest China inferred from genome-wide SNP data. Front Genet, 2022, **13**: 815285-815298

[16] Tishkoff S A, Kidd K K. Implications of biogeography of human populations for 'race' and medicine. Nat Genet, 2004, **36**(11): 21-27

[17] Phillips C. Forensic genetic analysis of bio-geographical ancestry. Forensic Sci Int Genet, 2015, **18**: 49-65

[18] Jin Y, Schaffer A A, Feolo M, *et al*. GRAF-pop: a fast distance-based method to infer subject ancestry from multiple genotype datasets without principal components analysis. G3 (Bethesda, Md), 2019, **9**(8): 2447-2461

[19] Jung J Y, Kang P W, Kim E, *et al*. Ancestry informative markers

(AIMs) for Korean and other East Asian and South East Asian populations. Int J Legal Med, 2019, **133**(6): 1711-1719

[20] Shi C M, Liu Q, Zhao S, *et al*. Ancestry informative SNP panels for discriminating the major East Asian populations: Han Chinese, Japanese and Korean. Ann Hum Genet, 2019, **83**(5): 348-354

[21] Price A L, Patterson N J, Plenge R M, *et al*. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet, 2006, **38**(8): 904-909

[22] Morrison D A. Phylogenetic tree-building. Int J Parasitol, 1996, **26**(6): 589-617

[23] Boca S M, Rosenberg N A. Mathematical properties of Fst between admixed populations and their parental source populations. Theor Popul Biol, 2011, **80**(3): 208-216

[24] Shriver M D, Smith M W, Jin L, *et al*. Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet, 1997, **60**(4): 957-964

[25] Shriver M D, Smith M W, Jin L, *et al*. Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet, 1997, **60**(4): 957-964

[26] Kim K, Omori R, Ito K. Inferring epidemiological dynamics of infectious diseases using Tajima's D statistic on nucleotide sequences of pathogens. Epidemics, 2017, **21**: 21-29

[27] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), 2010, **26**(5): 589-595

[28] Depristo M A, Banks E, Poplin R, *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet, 2011, **43**(5): 491-498

[29] Mckenna A, Hanna M, Banks E, *et al*. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res, 2010, **20**(9): 1297-1303

[30] Kim T, Seo H D, Hennighausen L, *et al*. Octopus-toolkit: a workflow to automate mining of public epigenomic and transcriptomic next-generation sequencing data. Nucleic Acids Res, 2018, **46**(9): 53-58

[31] Auton A, Brooks L D, Durbin R M, *et al*. A global reference for human genetic variation. Nature, 2015, **526**(7571): 68-74

[32] Chen J, He G, Ren Z, *et al*. Genomic insights into the admixture history of Mongolic- and Tungusic-Speaking populations from Southwestern East Asia. Front Genet, 2021, **12**: 685285

[33] Luo T, Wang R, Wang C C. Inferring the population structure and admixture history of three Hmong-Mien-speaking Miao tribes from Southwest China based on genome-wide SNP genotyping. Ann Hum Biol, 2021, **48**(5): 418-429

[34] Wang C C, Yeh H Y, Popov A N, *et al*. Genomic insights into the formation of human populations in East Asia. Nature, 2021, **591**(7850): 413-419

[35] Yang X, Sarengaowa, He G, *et al*. Genomic insights into the genetic structure and natural selection of Mongolians. Front Genet, 2021, **12**: 735786-735801

[36] Tan H, Wang R, Wang C C, *et al*. Fine-scale genetic profile and admixture history of two Hmong-Mien-Speaking Miao tribes from Southwest China inferred from genome-wide data. Hum Biol, 2022, **93**(3): 179-199

[37] Chang C C, Chow C C, Tellier L C, *et al*. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience, 2015, **4**: 7-22

[38] Manichaikul A, Mychaleckyj J C, Rich S S, *et al*. Robust relationship inference in genome-wide association studies. Bioinformatics (Oxford, England), 2010, **26**(22): 2867-2873

[39] Page R D. Visualizing phylogenetic trees using TreeView. Curr Protoc Bioinformatics, 2002, **6**(2): 1-15

[40] Retief J D. Phylogenetic analysis using PHYLIP. Methods Mol Biol (Clifton, NJ), 2000, **132**: 243-258

[41] Feng Q, Lu D, Xu S. AncestryPainter: a graphic program for displaying ancestry composition of populations and individuals. Genomics Proteomics Bioinformatics, 2018, **16**(5): 382-385

[42] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res, 2010, **38**(16): 164-170

[43] Huang D W, Sherman B T, Tan Q, *et al*. The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol, 2007, **8**(9): 183-198

[44] Armstrong R A. When to use the Bonferroni correction. Ophthalmic Physiol Opt, 2014, **34**(5): 502-508

[45] Manica A, Prugnolle F, Balloux F. Geography is a better determinant of human genetic differentiation than ethnicity. Hum Genet, 2005, **118**(3-4): 366-371

[46] Shi H, Zhong H, Peng Y, *et al*. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. BMC Biol, 2008, **6**: 45-55

[47] Bin X, Wang R, Huang Y, *et al*. Genomic insight into the population structure and admixture history of Tai-Kadai-Speaking Sui people in Southwest China. Front Genet, 2021, **12**: 735084-735099

[48] Huang X, Xia Z Y, Bin X, *et al*. Genomic insights into the demographic history of the Southern Chinese. Front Ecol Evol, 2022, **10**: 853391-853408

[49] Li H, Jin W L. Human Origin and Migration. Shanghai: Shanghai Scientific and Technological Education Publishing House, 2021

[50] Su B, Jin L, Underhill P, *et al*. Polynesian origins: insights from the Y chromosome. Proc Natl Acad Sci USA, 2000, **97**(15): 8225-8228

[51] Shi H, Dong Y L, Wen B, *et al*. Y-chromosome evidence of Southern origin of the East Asian-specific haplogroup O3-M122. Am J Hum Genet, 2005, **77**(3): 408-419

[52] Hammer M F, Karafet T M, Park H, *et al*. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. J Hum Genet, 2006, **51**(1): 47-58

[53] Kayser M, Choi Y, Van Oven M, *et al*. The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. Mol Biol Evol, 2008, **25**(7): 1362-1374

[54] Cai X, Qin Z, Wen B, *et al*. Human migration through bottlenecks

from Southeast Asia into East Asia during Last Glacial Maximum revealed by Y chromosomes. PLoS One, 2011, **6**(8): e24282

[55] Yan S, Wang C C, Li H, *et al*. An updated tree of Y-chromosome haplogroup O and revised phylogenetic positions of mutations P164 and PK4. Eur J Hum Genet, 2011, **19**(9): 1013-1015

[56] Zhang C, Lu Y, Feng Q, *et al*. Differentiated demographic histories and local adaptations between Sherpas and Tibetans. Genome Biol, 2017, **18**(1): 115-132

[57] Cao Y, Li L, Xu M, *et al*. The ChinaMAP analytics of deep whole genome sequences in 10, 588 individuals. Cell Res, 2020, **30**(9): 717-731

[58] Karafet T, Xu L, Du R, *et al*. Paternal population history of East Asia: sources, patterns, and microevolutionary processes. Am J Hum Genet, 2001, **69**(3): 615-628

[59] Thangaraj K, Singh L, Reddy A G, *et al*. Genetic affinities of the Andaman Islanders, a vanishing human population. Curr Biol, 2003, **13**(2): 86-93

[60] Wen B, Xie X, Gao S, *et al*. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet, 2004, **74**(5): 856-865

[61] Yu X, Li H. Origin of ethnic groups, linguistic families, and civilizations in China viewed from the Y chromosome. Mol Genet Genomics, 2021, **296**(4): 783-797

[62] Chiang C W K, Mangul S, Robles C, *et al*. A comprehensive map of genetic variation in the world's largest ethnic group-Han Chinese. Mol Biol Evol, 2018, **35**(11): 2736-2750

[63] Zuo X, Lu H, Jiang L, *et al*. Dating rice remains through phytolith carbon-14 study reveals domestication at the beginning of the Holocene. Proc Natl Acad Sci USA, 2017, **114**(25): 6486-6491

[64] Li Y C, Ye W J, Jiang C G, *et al*. River valleys shaped the maternal genetic landscape of Han Chinese. Mol Biol Evol, 2019, **36**(8): 1643-1652

[65] Ma B, Chen J, Yang X, *et al*. The genetic structure and East-West population admixture in Northwest China inferred from genome-wide array genotyping. Front Genet, 2021, **12**: 795570-795584

[66] Dunne E, Qi Q M, Shaqfeh E S, *et al*. Blood group alters platelet binding kinetics to von Willebrand factor and consequently platelet function. Blood, 2019, **133**(12): 1371-1377

[67] Marconi C, Di Buduo C A, Levine K, *et al*. Loss-of-function mutations in PTPRJ cause a new form of inherited thrombocytopenia. Blood, 2019, **133**(12): 1346-1357

[68] 彭德仁 . 中国汉族人 ABO 血型的分布 . 中国输血杂志, 1991 (1): 20-23
Peng D R. Chinese Journal of Blood Transfusion, 1991(1): 20-23

[69] 邢一 . UGT1A6 基因在中国不同人群中的药物基因组学研究 [D]. 上海: 上海交通大学, 2009
Xing Y. Pharmacogenomics Study of UGT1A6 Gene in Different Chinese Populations[D]. Shanghai: Shanghai Jiaotong University, 2009

[70] Lavalley M P. Logistic regression. Circulation, 2008, **117**(18): 2395-2399

[71] Che D, Liu Q, Rasheed K, *et al*. Decision tree and ensemble learning algorithms with their applications in bioinformatics. Adv Exp Med Biol, 2011, **696**: 191-199

[72] Pandis N. Linear regression. Am J Orthod Dentofacial Orthop, 2016, **149**(3): 431-434

[73] Rigatti S J. Random Forest. J Insur Med, 2017, **47**(1): 31-39

[74] Huang S, Cai N, Pacheco P P, *et al*. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics Proteomics, 2018, **15**(1): 41-51

[75] Connor C W. Artificial intelligence and machine learning in anesthesiology. Anesthesiology, 2019, **131**(6): 1346-1359

[76] Hatwell J, Gaber M M, Atif Azad R M. Ada-WHIPS: explaining AdaBoost classification with applications in the health sciences. BMC Med Inform Decis Mak, 2020, **20**(1): 250-274

[77] Karalis G. Decision trees and applications. Adv Exp Med Biol, 2020, **1194**: 239-242

[78] Ye X, Huang Y, Lu Q. Automatic multichannel electrocardiogram record classification using XGBoost fusion model. Front Physiol, 2022, **13**: 840011-840023

[79] Qin X, Chiang C W K, Gaggiotti O E. KLFDAPC: a supervised machine learning approach for spatial genetic structure analysis. Brief Bioinform, 2022, **23**(4): bbac202

[80] Alladio E, Poggiali B, Cosenza G, *et al*. Multivariate statistical approach and machine learning for the evaluation of biogeographical ancestry inference in the forensic field. Sci Rep, 2022, **12**(1): 8974-8990

[81] Phillips C, Salas A, Sánchez J J, *et al*. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int Genet, 2007, **1**(3-4): 273-280

Genetic Structure of East Asians Based on High-density SNP Data*

GU Jia-Qi[1,2)**], JIANG Li[1)**], XU Jing-Yi[1)], WANG Han[1)], WEI Yi-Liang[2)***], LI Cai-Xia[1)***]

([1)]*Key Laboratory of Forensic Genetics, Beijing Engineering Research Center of Crime Scene Evidence Examination, National Engineering Laboratory for Forensic Science, Institute of Forensic Science, Beijing* 100038, China;

[2)]*Key Laboratory of Phylogeny and Comparative Genomics of Jiangsu Province, College of Life Sciences, Jiangsu Normal University, Xuzhou* 221116, China)

Abstract **Objective** East Asia harbors a vast territory, with many populations and a diversity of languages. China has 34 provincial-level administrative districts which distributed in seven geographical divisions, mainly populated by seven linguistic family-speaking populations. Previous studies have focused on the genetic history of origin, migration and fusion of East Asia populations. We collected and analyzed high-density SNP data of 5 147 individuals in the world, studied the genetic relationship and structure between East Asia populations, especially Chinese populations and other world populations from the perspective of geography and language. **Methods** We collected and carried out quality control of high-density SNP data of 5 147 individuals in the world. We studied the genetic structure of Chinese population. The final obtained 32 789 SNPs were statistically tested by allele frequency difference analysis. Meanwhile, we employed principal component analysis, phylogenetic tree, ancestry component analysis and *D*-test statistics to explore the genetic relationship between East Asia populations and other populations in the world, as well as the genetic relationship and structure of Chinese populations. **Results** We found that there were significant differences among East Asian, African, American and European. Chinese population can be divided into seven subgroups. The genetic clustering in different populations has a strong correlation with their geographical distribution, linguistic families and ethnic origin history. **Conclusion** We studied the genetic relationship and differences between Chinese population and world population, and systematically studied the genetic substructure of Chinese population. This will enrich the research foundation of population genetics and forensic genetics of East Asia population and provide data support for individualized medical work.

Key words population genetic analysis, forensic genetics, East Asians, genetic relationship

DOI：10.16476/j.pibb.2022.0441

---

* This work was supported by grants from The National Natural Science Foundation of China (82171870), The Key Laboratory of Forensic Genetics Open Project (2020FGKFKT01), and Postgraduate Research and Practice Innovation Program of Jiangsu Province (KYCX21_2597).

** These authors contributed equally to this work.

*** Corresponding author.

WEI Yi-Liang. Tel：86-15332195528，Email：weiyiliang.2013@tsinghua.org.cn

LI Cai-Xia. Tel：86-13810666929，E-mail：licaixia@tsinghua.org.cn

Received：September 15，2022 Accepted：February 13，2023