遗传 Hereditas (Beijing) 2020 年 3 月, 42(3): 296－308
www.chinagene.cn

研究报告

74-plex SNPs 复合检测体系在中国人群中的族群推断研究

刘杨[1,2]，孙昌春[1,2]，马咪[2,3]，王玲[2]，赵雯婷[2]，马泉[2]，季安全[2]，刘京[2]，李彩霞[1,2]

1. 山西医科大学，太原 036000
2. 公安部物证鉴定中心，法医遗传学公安部重点实验室，现场物证溯源技术国家工程实验室，北京 100038
3. 新疆生产建设兵团第七师公安局，奎屯 833200

摘要：使用一组祖源 SNP 可以分析某人群的遗传成分，推断某个体的族群来源。本课题组前期筛选出 74 个 SNP 位点实现了撒哈拉以南的非洲、北非、欧洲、美洲、大洋洲、南亚、西南亚、东亚、东北亚和东南亚等 10 个地理区域人群的推断，并基于 MassARRAY 质谱分析技术构建了 74-plex SNPs 复合检测体系。本研究利用该体系对 14 个中国人群 1371 份样本进行基因分型，验证评估该体系对中国人群的区分能力和法医学应用效能。首先，基于全球 57 个人群 3628 份个体构建参考人群分型库，采用 Structure 分析和等位基因频率热图等方法进行人群区分能力评估；然后，选取千人基因组计划中 3 个人群(不包含在参考人群分型库中)及本实验室检测的 14 个人群共计 1654 个体作为测试数据集，通过似然比和族群成分等统计分析，评估该体系对实际样本的族群来源推断能力。结果表明，DNA 的量最低为 1.5 ng 时，74 个 SNP 均可正确判型，适用于微量检材的检测；该体系对全球 10 个地理区域人群有区分能力，针对测试人群中欧洲、美洲、南部非洲个体族群来源推断的准确率为 95.4%、不排除率为 1.06%，东亚个体推断的准确率为 71.0%、不排除率为 17.9%，东南亚个体推断的准确率 66.4%、不排除率为 33.3%。该方法可以为实际案件侦察提供线索。

关键词：法医遗传学；祖源 SNP；族群推断；中国人群

收稿日期: 2019−11−13; **修回日期:** 2020−01−13

基金项目: 国家自然科学基金项目(编号：81772027)，国家重点研发计划资(编号：2017YFC0803501)，国家科技资源共享服务平台计划项目(编号：YCZYPT[2017]01-3)和中央公益类基本科研业务费专项资金项目(编号：2017JB025，2017JB026，2019JB011)资助[Supported by the National Natural Science Foundation of China (No.81772027), National Key R&D Program of China (No. 2017YFC0803501), National Science and Technology Resources Sharing Service Platform Project (No. YCZYPT[2017]01-3) and Fundamental Research Funds for Institute of Forensic Science (Nos. 2017JB025, 2017JB026, 2019JB011)]

作者简介: 刘杨，硕士研究生，研究方向：法医学。E-mail: 1979227825@qq.com

通讯作者: 李彩霞，博士，主任法医师，研究方向：法医遗传学。E-mail: licaixia@tsinghua.org.cn
刘京，硕士，主检法医师，研究方向：法医遗传学。E-mail: biojing@yeah.net

DOI: 10.16288/j.yczz.19-252

网络出版时间: 2020/2/10 16:00:49

URI: http://kns.cnki.net/kcms/detail/11.1913.R.20200210.1533.002.html

The ancestry inference of Chinese populations using 74-plex SNPs system

Yang Liu[1,2], Changchun Sun[1,2], Mi Ma[2,3], Ling Wang[2], Wenting Zhao[2], Quan Ma[2], Anquan Ji[2], Jing Liu[2], Caixia Li[1,2]

1. Shanxi Medical University, Taiyuan 036000, China
2. 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
3. Production and Construction Corps of Seventh Division Public Security Bureau, Kuitun 833200, China

Abstract: A panel of ancestry informative SNPs (AISNPs) can be used to analyze the genetic components of a population and infer the ancestral origin of a DNA sample. Previously, we have selected a 74-AISNPs panel and used it to infer the ancestry of unknown individuals in the following ten geographical regions: Sub-Saharan Africa, North Africa, Europe, Pacific, Americas, Southwest Asia, South Asia, North Asia, East Asia and Southeast Asia. We have also established a 74-plex SNPs assay based on SEQUENOM system. In the present study, we genotyped 1371 individuals from 14 populations of China using this multiplex assay, and validated its ability to infer the ancestry in Chinese populations. Firstly, based on the reference database of 3628 individuals from 57 world populations, Structure and Heatmap were employed to evaluate the population differentiation capacity. The training data include 1654 individuals from 14 Chinese populations and 3 populations from 1K Genome, which are not included in the reference database. Then the likelihood ratio and ancestry components were analyzed for individual ancestry assignment using the 74-plex SNPs. The minimum amount of DNA required for a full genotype of the 74 SNPs is 1.5 ng, which is applicable for forensic analysis. The results demonstrate that this system can be used in differentiating the population from ten geographical regions. The ancestry inference accuracy for EUR/SAFR/AME population is 95.4%, 71.0% for East Asia and 66.4% for Southeast Asia respectively. The ancestry inference inclusive rate for EUR/SAFR/AME population is 1.06%, 17.9% for East Asia and 33.3% for Southeast Asia respectively. The results suggest that this method can be used in forensic investigations of criminal cases.

Keywords: forensic genetics; AISNPs; ancestry inference; Chinese populations

DNA 供者的族群地域分析不仅对于生物医药、人类迁移进化等研究有重要参考价值，而且在法庭科学领域也具有重要应用价值，近年来被广泛关注[1~4]。当犯罪嫌疑人遗留在现场生物检材的 STR (short tandem repeat, STR)数据与 DNA 数据库或者某个嫌疑人没有比中时，如果能够对生物检材来源人的族群、地域进行推断，将有助于锁定嫌疑人范围，促进案件定性和明确侦查方向。通过测序技术获得个人基因组上的祖源 SNP (ancestry informative SNPs, AISNPs)分型信息，比较这些 SNP 分型数据与参考族群的相似性，可以计算族源成分，推断其族群来源[5~9]。

目前报道了大量洲际人群区分的 AISNPs 体系[10~14]。本课题组前期筛选出的 74 个 SNP 位点能够实现全球 10 个地理区域人群(撒哈拉以南的非洲、北非、欧洲、美洲、大洋洲，南亚、西南亚、东亚、东北亚、东南亚)的区分[15]，且基于质谱技术构建了 74-plex SNPs 复合检测体系[16]，实现了东亚人群的南北方遗传成分的进一步区分。但是，尚未进行该复合检测体系的性能验证及大规模样本的验证。本文利用 74-plex SNPs 复合检测体系对 14 个中国人群 1371 份样本进行基因分型，并对 74-plex SNPs 复合检测体系进行了体系性能验证和大规模样本的区分能力验证。本研究的成果可进一步丰富我国人群的 AISNPs 位点的数据，进而为中国不同语系人群特异性位点的筛选打下基础，并且可以为案件提供

侦查线索。

1 材料与方法

1.1 样本信息

参考数据库参照前期文献报道[17]，共计 57 个人群 3628 份个体。另外选取千人基因组 3 个人群(不包含在参考数据库中)和本实验室检测的 14 个中国人群，共 17 个人群 1654 份个体作为测试样本。基于参考数据库进行族群来源推断，评估体系在实际样本中的族群来源区分能力。本实验室样本均来源于国家科技资源共享服务平台计划项目(编号：YCZYPT[2017]01-3)。测试人群样本详细信息见表 1。本实验室检测的所有样本对象均签署知情同意书及自述其详细族群信息。本研究已通过公安部物证鉴定中心伦理委员会的审查批准。

1.2 DNA 的提取和定量

静脉血样本 DNA 的提取采用德国 QIAGEN 公司 QIAamp® DNA Blood Midi 试剂盒；用 NanoDrop 2000C 分光光度计(Thermo Scientific 公司，美国)进行定量。用 18.2MΩ 去离子灭菌水调整浓度至 5~10 ng/μL 备检。

1.3 SNP 位点来源

74 个 SNP 位点源于本课题组前期筛选[15]，基于 MassARRAY 质谱检测平台构建了复合检测体系[16]，74 个 SNP 位点在 3 个反应孔中检测，SNP 位点信息见表 2。

1.4 检测 SNP 分型

PCR 复合扩增及纯化：PCR 复合扩增反应体系为 5 μL，PCR 反应条件：95℃ 2 min，95℃ 30 s，56℃ 30 s，72℃ 1 min，循环 45 次，最后延伸 72℃ 5 min。纯化反应体系为 7 μL，充分振荡混匀后 37℃孵育 40 min，85℃ 5 min 灭活酶活性。

单碱基延伸反应：采用 9 μL 体系，94℃ 30 s；94℃ 5 s，(52℃ 5 s 和 80℃ 5 s，循环 5 次)，共 40 个循环；然后 72℃ 3 min。

树脂纯化：延伸后的体系加 15 mg 的 Clean

表 1 测试人群样本信息表

Table 1 The information of test populations

代码	人群信息	数量	样本来源
EIC	内蒙古鄂温克族人	35	本实验室
DIC	内蒙古达斡尔族人	34	本实验室
MIC	内蒙古蒙古族人	135	本实验室
TUQ	青海土族人	116	本实验室
CCT	西藏藏族人	143	本实验室
CHQ	青海汉族人	100	本实验室
CHL	青岛汉族人	94	本实验室
CHN	河南汉族人	107	本实验室
HGC	广西汉族人	79	本实验室
CHG	广东汉族人	46	本实验室
HCM	广东梅州客家汉族人	50	本实验室
DGC	广西侗族人	155	本实验室
CDY	云南傣族人	93	本实验室
KGC	广西京族人	184	本实验室
ESN	非洲尼日利亚人	99	千人基因组
FIN	芬兰人	99	千人基因组
PEL	秘鲁人	85	千人基因组

表 2　每个反应孔中的 SNP 位点信息

Table 2　The information of SNP loci in each well

反应孔	序号	SNP 编号	染色体	位置	等位基因	反应孔	序号	SNP 编号	染色体	位置	等位基因
Well 1	1	rs10496971	2	145769943	G/T	Well 2	38	rs1513056	12	17407792	A/G
Well 1	2	rs10511828	9	28628500	C/T	Well 2	39	rs174574	11	61600342	A/C
Well 1	3	rs10512572	17	69512099	A/G	Well 2	40	rs17822931	16	48258198	C/T
Well 1	4	rs10516441	4	100307167	A/G	Well 2	41	rs1876482	2	17362568	A/G
Well 1	5	rs12913832	15	28365618	A/G	Well 2	42	rs192655	6	90518278	A/G
Well 1	6	rs1426654	15	48426484	A/G	Well 2	43	rs1950993	14	58238687	G/T
Well 1	7	rs1572018	13	41715282	C/T	Well 2	44	rs2006996	9	117592638	C/T
Well 1	8	rs16891982	5	33951693	C/G	Well 2	45	rs2125345	17	73782191	C/T
Well 1	9	rs17028973	4	100322786	C/T	Well 2	46	rs2238151	12	112211833	C/T
Well 1	10	rs1800414	15	28197037	C/T	Well 2	47	rs3118378	1	68849687	A/G
Well 1	11	rs1871428	6	168665760	A/G	Well 2	48	rs37369	5	35037115	C/T
Well 1	12	rs2033111	17	53788280	A/G	Well 2	49	rs3814134	9	127267689	A/G
Well 1	13	rs2241894	4	100266133	C/T	Well 2	50	rs4833103	4	38815502	A/C
Well 1	14	rs2242480	7	99361466	C/T	Well 2	51	rs6451722	5	43711378	A/G
Well 1	15	rs2702414	4	179399523	A/G	Well 2	52	rs647325	1	18170886	A/G
Well 1	16	rs2899826	15	74734500	A/G	Well 2	53	rs6990312	8	110602317	G/T
Well 1	17	rs316598	5	2364626	C/T	Well 2	54	rs7226659	18	40488279	G/T
Well 1	18	rs3737576	1	101709563	C/T	Well 2	55	rs7238445	18	49781544	A/G
Well 1	19	rs3811801	4	100244319	A/G	Well 2	56	rs7554936	1	151122489	C/T
Well 1	20	rs3827760	2	109513601	A/G	Well 2	57	rs8035124	15	92105708	A/C
Well 1	21	rs385194	4	85309078	A/G	Well 2	58	rs917115	7	28172586	C/T
Well 1	22	rs459920	16	89730827	C/T	Well 2	59	rs9319336	13	27624356	C/T
Well 1	23	rs4908343	1	27931698	A/G	Well 3	60	rs1229984	4	100239319	C/T
Well 1	24	rs671	12	112241766	A/G	Well 3	61	rs174570	11	61597212	C/T
Well 1	25	rs6754311	2	136707982	C/T	Well 3	62	rs2024566	22	41697338	A/G
Well 1	26	rs734873	3	147750355	A/G	Well 3	63	rs2166624	13	42579985	A/G
Well 1	27	rs7997709	13	34847737	C/T	Well 3	64	rs2814778	1	159174683	C/T
Well 1	28	rs8003942	14	105971670	A/G	Well 3	65	rs2986742	1	6550376	C/T
Well 1	29	rs8113143	19	33652247	A/C	Well 3	66	rs310644	20	62159504	C/T
Well 1	30	rs870347	5	6845035	A/C	Well 3	67	rs4670767	2	37941396	G/T
Well 1	31	rs9522149	13	111827167	C/T	Well 3	68	rs6054605	20	744570	A/G
Well 2	32	rs10108270	8	4190793	A/C	Well 3	69	rs735480	15	45152371	C/T
Well 2	33	rs10236187	7	139447377	A/C	Well 3	70	rs7722456	5	170202984	C/T
Well 2	34	rs1040404	1	168159890	A/G	Well 3	71	rs7745461	6	21911616	A/G
Well 2	35	rs10513300	9	120130206	C/T	Well 3	72	rs798443	2	7968275	A/G
Well 2	36	rs11652805	17	62987151	C/T	Well 3	73	rs8021730	14	67886781	G/T
Well 2	37	rs13400937	2	79864923	G/T	Well 3	74	rs818386	16	65406708	C/T

Resin 树脂进行脱盐纯化。将 Clean Resin 树脂平铺到树脂板中，将干燥后的树脂倒入延伸产物板中，封膜，低速垂直旋转 25 min 使树脂与反应物充分接触，3000 r/min 离心 5 min 使树脂沉入孔底部。

芯片点样和质谱检测：用点样仪(MassARRAYTM Nanodispenser RS1000， 美国 Agena 公司)把纯化后的样本点到带有基质的芯片上(8~15 nL)。然后用质谱检测分析仪(MassARRAYTM Analyzer， 美国 Agena 公司)进行分型检测[18]。用 TYPER 4.0 软件对分型结果进行分析。

1.5 性能指标验证

分型准确性验证：选取 5 份样本：9947A、B0242、LCX、QEF、U144 送至生工生物工程有限公司进行 Sanger 测序，验证本研究检测体系的基因分型与测序结果一致性。

灵敏度验证：将 10 ng/μL 标准品 9947 做浓度梯度稀释，15 μL 体系中 DNA 模板最终量分别为 30、15、6、3、1.5 和 0.6 ng。使用构建的 74-plex SNPs 复合检测体系进行扩增和基因分型，每个浓度重复 3 次，用于验证该检测体系的灵敏度。

1.6 分析方法

1.6.1 Structure 分析

针对全球 10 大区域人群分型数据库，用 Structure 2.3.4[19]软件进行族群成分分析(*K*=3-10，run=15, 10000 burnins,10000 MCMC)，分析各人群的遗传结构。使用 Clumpak 软件绘制 Structure 结果人群聚类图，相似度的阈值设置为 0.9。

1.6.2 等位基因频率热图分析

用 Genepop 软件(http://www.genepop.curtin.edu.au/genepop_op5.html)计算每个位点的等位基因频率，使用 R v3.0.1 软件绘制等位基因频率热图。

1.6.3 群体匹配概率和似然比

用 DNA 族群推断系统软件(DAA)[20]计算 17 个人群 1654 份测试样本的群体匹配概率(AMP)和似然比(LR)，当 LR>10 时，AMP 排第一位的人群为未知个体的来源族群，当 LR≤10 时，AMP 排序前两位人群均不排除。

1.6.4 箱形图分析

用 Structure 软件分析 17 个人群 1654 份测试样本的族群成分(*K*=10, run=15)，基于每个个体族群成分的最大值、最小值、中位数和两个四分位数，用 EXCLE2016 软件绘制箱线图展示每个个体族群成分的分布。

2 结果与分析

2.1 74-plex SNPs 复合检测体系性能指标验证结果

分型准确性验证：5 份测序样本共获得 370 个 SNP 分型数据，经对比测序结果与本研究复合检测体系所获得的基因分型 100%一致。

灵敏度验证：使用构建的 74-plex SNPs 复合检测体系检测模板量为 30~0.6 ng 的 9947。3 次重复结果均显示，DNA 模板量最低为 1.5 ng 时 74 个位点等位基因均可正确判型(图 1)。

2.2 用全球 10 大区域人群分型数据库对体系效能进行评价

2.2.1 Structure 族群成分分析结果

图 2 展示了全球 57 个人群 3628 个个体的 Structure 分析结果(*K*=3~10)，图中展示的是每个 *K* 值多次运算结果中的最主要的聚类模式。当 *K*=10 时，57 个人群被聚类为撒哈拉以南的非洲、北非、西南亚、欧洲、南亚、东亚、东北亚、东南亚、大洋洲和美洲等 10 个区域。

2.2.2 等位基因频率热图

基于 57 个人群在 74 个 SNP 位点的等位基因频率分布，绘制等位基因频率聚类热图(图 3)。通过图 3 可以找出人群特异 SNP 位点，例如 rs10108270、rs2986742、rs7238445 和 rs451722 聚类在一起，且它们在南非人群中的频率明显高于其他人群，说明这些位点分型是南非人群特异位点。57 个人群在热图的左侧聚为 10 簇，分别为撒哈拉以南的非洲、北

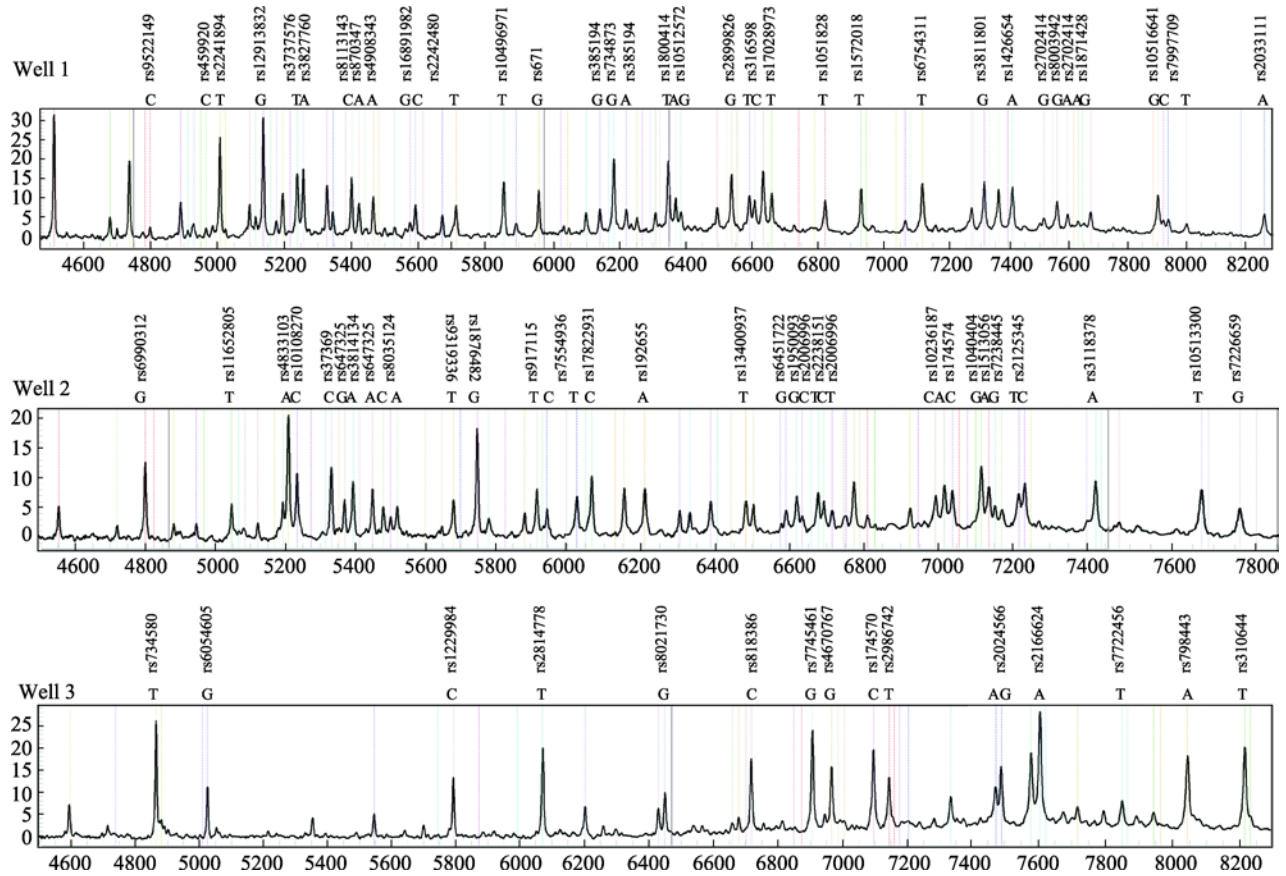

图 1　DNA 模板量为 1.5 ng 的分型结果

Fig. 1　The genotyping result of DNA template was 1.5 ng

非、欧洲、美洲、大洋洲、南亚、西南亚、东亚、东北亚和东南亚。

2.3　17 个人群 1654 份测试个体的族群来源推断

本文使用 17 个人群 1654 份个体作为测试数据集评估 74-plex SNPs 复合检测体系的族群来源推断能力，验证体系在实际样本中的应用效能。所有测试样本均不包括在参考数据库中。

2.3.1　似然比

对已知来源的 1654 份样本基于参考数据库进行随机人群匹配概率计算，基于似然比进行族群来源的统计如表 3。对测试人群整体的推断准确率为 74%，不排除率为 19%，错误率为 7%。针对测试人群中欧洲、美洲、南部非洲个体族群来源推断的准确率为 95.4%，不排除率为 1.06%；东亚个体推断的准确率为 71.0%，不排除率为 17.9%，错误率为 11.1%；东南亚个体推断的准确率 66.4%，不排除率为 33.3%，错误率为 0.2%。

2.3.2　族群成分

对已知来源的 1654 份样本基于参考数据库使用 Structure 2.3.4 软件计算其族群成分(K=10, run=10)。统计每个人群的平均族群成分见表 3，所有样本的族群成分绘制箱线图(图 4，A 和 B)。表 3 可见内蒙古蒙古族(MIC)、达斡尔族(DIC)、和鄂温克族(EIC)人群的东北亚成分的平均值分别为 0.56、0.45 和 0.31；东亚成分为 0.31、0.42 和 0.56。西藏藏族(CTT)和青海土族(TUQ)人群以东北亚成分为主，分别为 0.78 和 0.63。青海汉族(CHQ)表现为东北亚和东亚成分的混合，族群成分平均值分别 0.46 和 0.41。青岛汉族(CHL)和河南汉族(CHN)人群的以东亚成分为主,族群成分平均值分别为 0.61 和 0.65。广西汉族(HGC)、广东客家汉族(HCM)和广东汉族

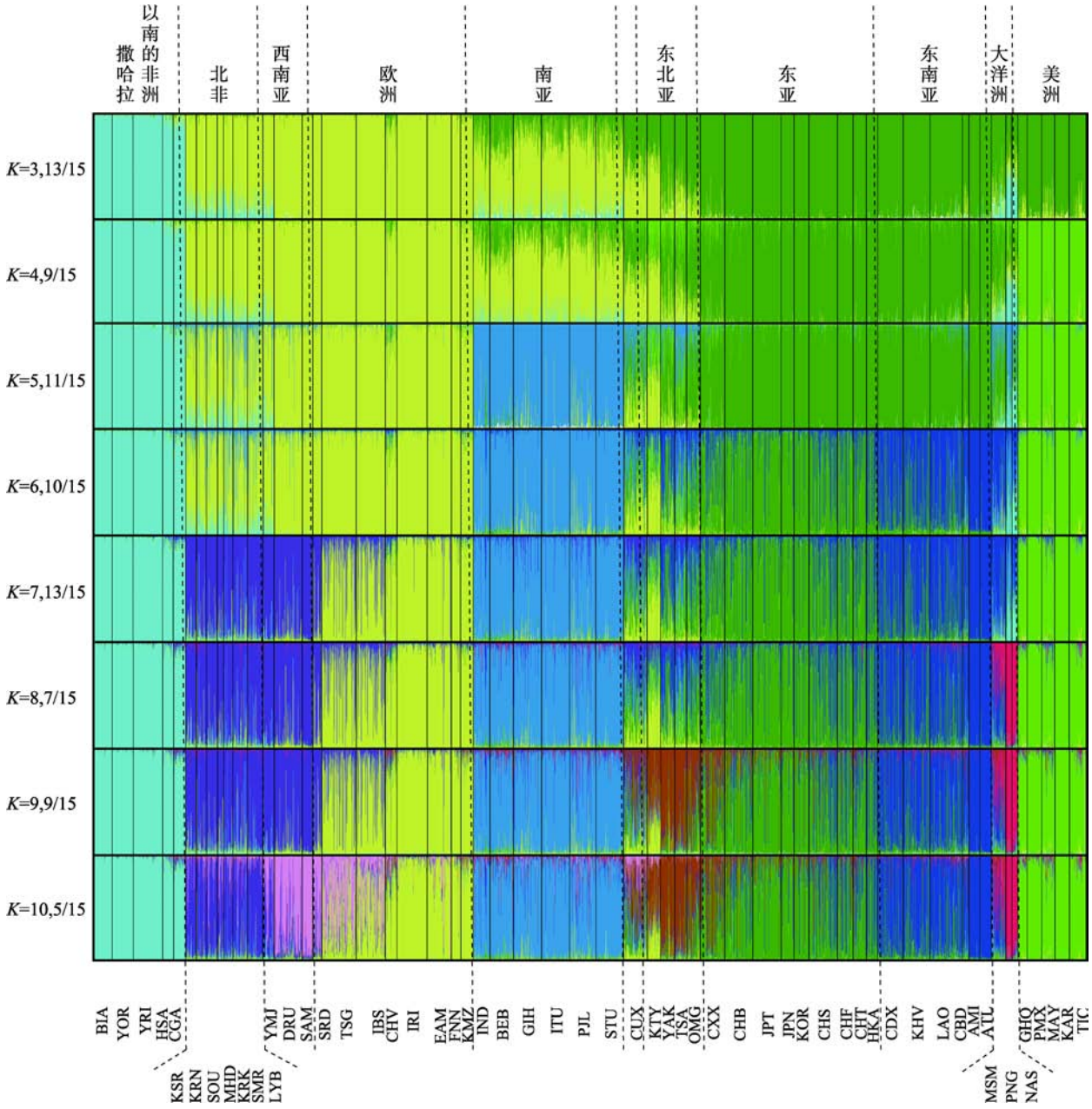

图 2　74 SNPs 的 57 人群 Structure 分析结果

Fig. 2　The Structure analysis of 74 SNPs for 57 populations

K 值是指运行 Structure 分析时，用户假定全部群体分为几个亚群，每个亚群在图中用一个单独颜色表示，图中展示的是每个 *K* 值多次运算结果中的最主要的聚类模式，例如：当 *K*=3 时，15 次运算中有 13 次结果中的祖先成分模式为图 2 中所示模式。

(CHG)等中国南方汉族人群中东亚成分的平均值分别为 0.31、0.38 和 0.49；东南亚成分的平均值分别为 0.6、0.53 和 0.45。广西京族(KGC)、广西侗族(DGC)和云南傣族(CDY)等中国南方少数民族以东南亚成分为主，3 个人群东南亚成分的平均值均大于 0.6。

3　讨论

本课题组前期建立了针对 3 大洲际人群(亚、欧、非)的 27-plex SNPs 和五大洲际人群的 28-plex SNPs，实现了洲际人群的区分及其遗传结构的分析，但对洲际人群内部亚人群的进一步区分效力不足。而近

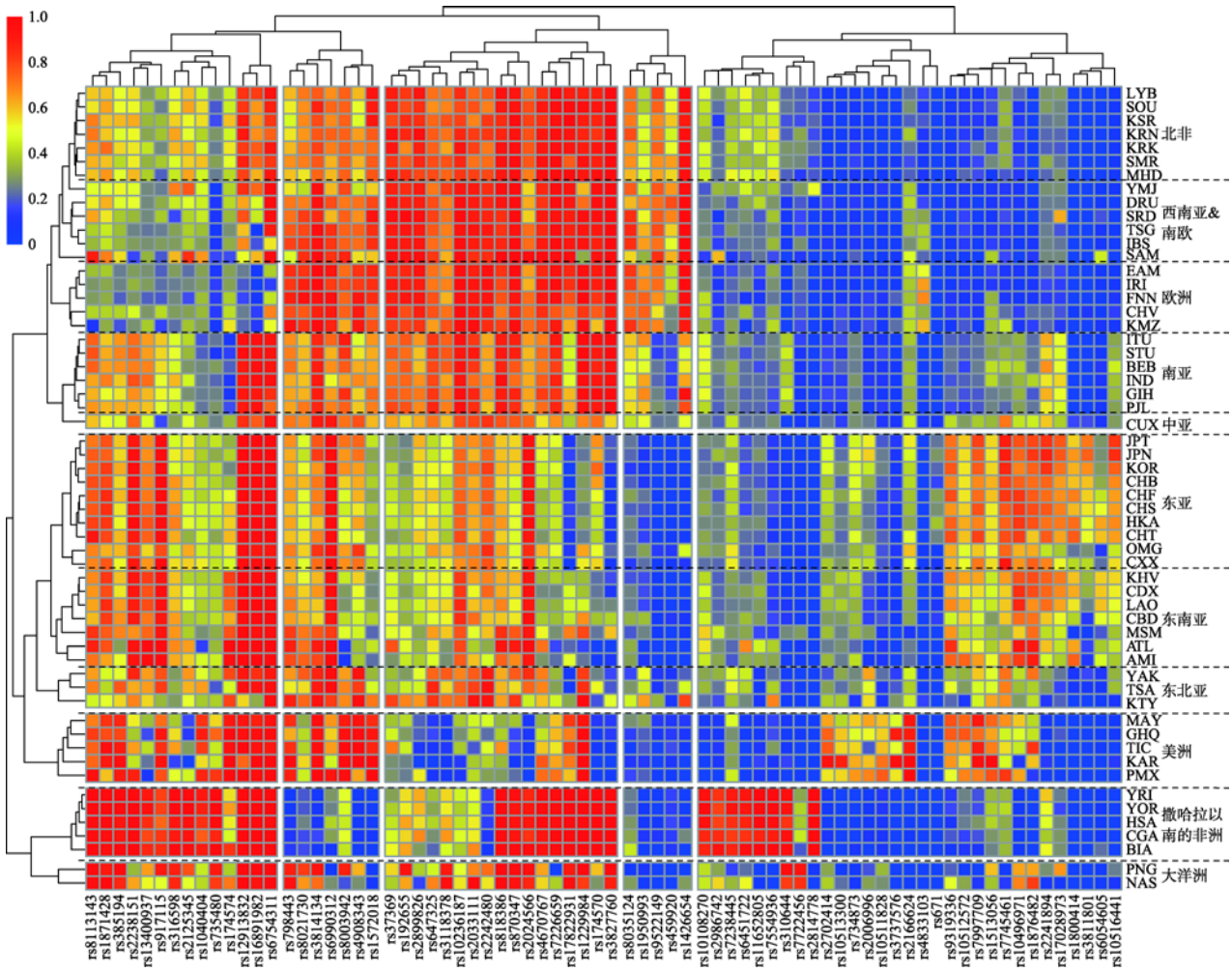

图 3　74 个 SNPs 的 57 个人群等位基因频率热图

Fig. 3　Heatmap of 74 SNPs based on the allele frequencies of 57 populations

颜色的深浅代表 SNP 位点的基因频率在不同群体中的相似性和差异性，红色表示最高等位基因频率，蓝色表示最低等位基因频率。

年来，相关研究已经逐步从洲际群体鉴别过渡到亚人群的鉴别研究，如 Phillips 的“MAPlex”体系[21]使用 164 个遗传标记实现了非洲、中东、欧洲、南亚、东亚、美洲、大洋洲人群的区分，但该组位点没有实现东亚人群的进一步细分；Sun 等[22]的 12 个 multi-indels 推断体系实现了喀拉拉人(Keralites)、老挝人(Laotians)、日本人(Japanese)、汉族(Han)和中国藏族(Chinese Tibetan)人群的区分。本研究体系实现了东南亚、东亚与东北亚人群的进一步区分，以及北非、西南亚与欧洲人群的进一步区分，且构建了适用于法医现场生物物证的检测体系，在实际应用中，可以使案件现场遗留的生物物证的族群来源进一步细化。在下一步研究中可以借鉴 12 个 Multi-indels 推断体系[22]等相关研究中的位点，构建更加精细的针对东亚人群的区分体系。

3.1　体系性能验证

5 份样本的 Sanger 测序结果与本研究的 74-plex SNPs 复合检测体系检测的 SNP 分型 100%一致。灵敏度结果显示，模板量最低为 1.5 ng 时 74 个位点等位基因均可正确判型，适用于微量检材的检测。该体系尚未进行检材适应性、组织统一性的验证，后期需要进行该两项的测试。

3.2　人群的遗传区分能力

本研究是基于全球十个区域 57 个人群为参考

表 3 测试样本的族群推断结果

Table 3 The ancestry inference result of test samples

测试人群（代码）	来源区域	预测第 1 位的人群(LR>10)						不排除(LR≤10)	错误	样本数量
		撒哈拉以南的非洲	北非 欧洲	东亚	东北亚	东南亚	美洲			
内蒙古达斡尔族(DIC)	EA(东亚)			19 (0.42)	8 (0.45)			4	3	34
内蒙古鄂温克族(EIC)	EA(东亚)			24 (0.56)	3(0.31)			3	5	35
内蒙古蒙古族(MIC)	EA(东亚)			59 (0.31)	37 (0.56)	2		26	13	135
青海土族(TUQ)	EA(东亚)			76 (0.26)	3 (0.63)			20	17	116
青海汉族(CHQ)	EA(东亚)			74 (0.41)	3 (0.46)			5	18	100
西藏藏族(CTT)	EA(东亚)			71 (0.14)	17 (0.78)	2		36	19	143
河南汉族(CHN)	EA(东亚)			87 (0.65)	1 (0.15)	2 (0.15)		1	19	107
青岛汉族(CHL)	EA(东亚)			85 (0.61)	(0.28)	1		0	9	94
广东汉族(CHG)	EA(东亚)			13 (0.49)		13 (0.45)		20		46
广西汉族(HGC)	EA(东亚)			13 (0.31)		30 (0.60)		36		79
广东客家汉族(HCM)	EA(东亚)			17(0.38)		15 (0.53)		17	1	50
广西侗族(DGC)	SEA(东南亚)			27 (0.32)		65 (0.62)		63		155
广西京族(KGC)	SEA(东南亚)			16 (0.21)		109 (0.73)		58	1	184
云南傣族(CDY)	SEA(东南亚)			6 (0.18)		64 (0.76)		23		93
非洲尼日利亚人(ESN)	S-AFR(撒哈拉以南的非洲)	99 (0.98)						0		99
芬兰人(FIN)	EUR(欧洲)		99 (0.93)					0		99
秘鲁人(PEL)	AME(美洲)	1			3		72(0.74)	3	10	85

括号中的数字表示每个人群对应族群成分的平均值。

数据库进行族群来源分析，与本课题组 2016 年研究的 61 个参考人群相比做了以下优化：(1)增加了维吾尔族(CUX)和锡伯族(CXX)，以评估该体系在新疆人群中的区分能力；(2)为避免人群样本数量不均一带来的结果偏差，将样本量较少且遗传结构相近的群体进行了合并，(比如，欧洲人群中的 TSI 和 GRK 人群合并为 TSG，南亚人群中的 KER、THT 和 KCH 人群合并为 IND，大洋洲人群中的 MLY、SMO 和 MCR 人群合并为 MSM，美洲人群中的 GHB 和 QUE 人群合并为 GHQ)。

用该体系对 57 个人群进行族群成分分析(图 2)，结果表明该体系可以对全球十大区域人群进行区分。当 *K*=3 时，可以对亚洲、欧洲、非洲进行明确区分，维吾尔族(CUX)、东北亚的汉特(KTY)等混合人群的遗传成分呈现在欧洲和东亚族群成分的连续分布，当 *K*=4 时，可以看出维吾尔族与汉特人群混合成分的差异，前者是欧洲(0.49)和东亚(0.44)成分的混合，而后者主要是欧洲(0.60)和美洲成分(0.31)的混合，这在实际应用中，有助于混合人群的进一步准确区分。随着 *K* 值增加，先后在美洲、南亚、东南亚、北非、大洋洲、东北亚、西南亚出现新的族群成分，*K*=10 时，该体系可以对全球十大区域人群有较好的区分效力。地中海沿岸人群由于存在着广泛的基因交流，北非和西南亚人群当 *K*=10 时才可以进行区分，并且南欧一些人群如由意大利和希腊人组成的 TSG 人群，有较多的西南亚成分。

通过图 3 可以找出人群特有的 SNP 位点，例如 rs10108270, rs2986742, rs7238445 和 rs451722 聚类在一起，且它们在南非人群中的频率明显高于其他人群，说明这些位点是南非人群特异。基因频率分布热图对所有人群的聚类结果与 Structure 分析 *K*=10 时的结果基本相同，二者可以相互印证。

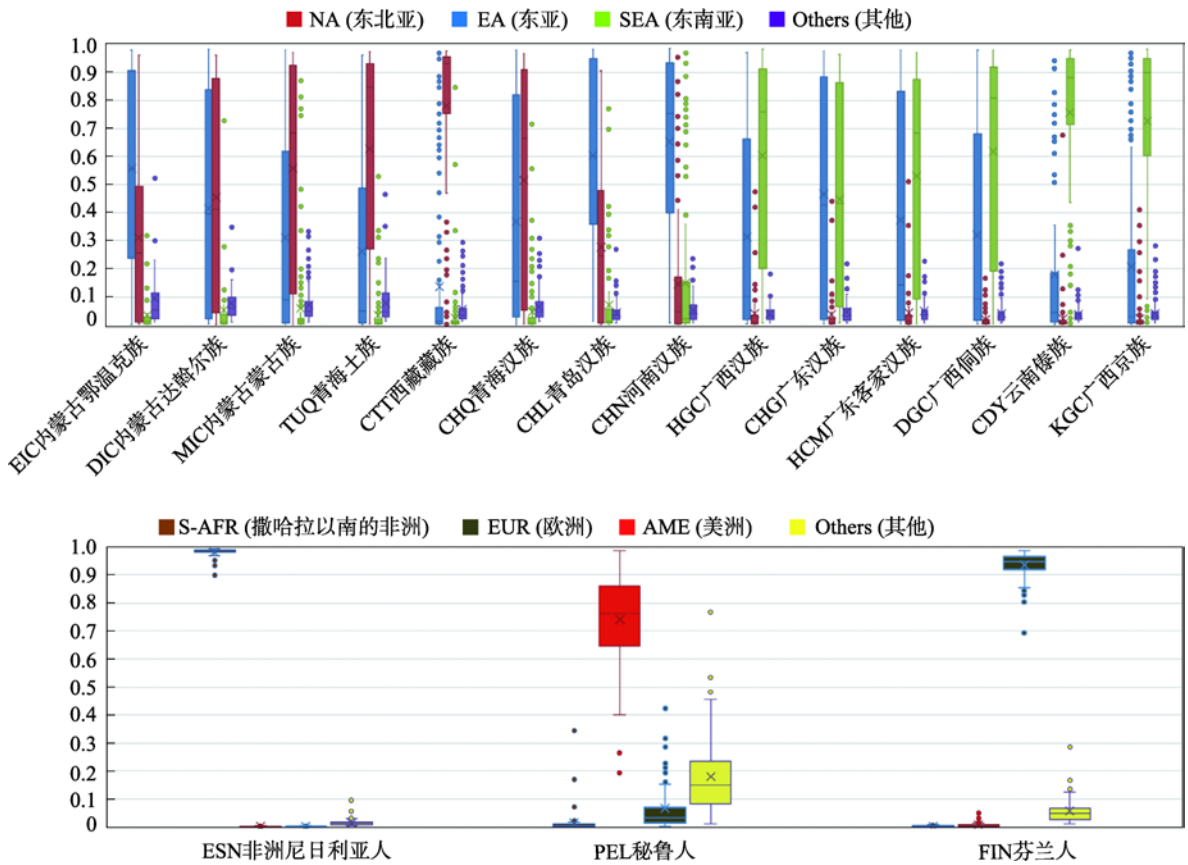

图 4　基于 17 个测试人群 1654 个体的族群成分的箱线图

Fig. 4　The box-plot of ancestry component for 1654 individuals of 17 test populations

3.3　未知个体族群来源推断

本研究使用 17 个人群 1654 份个体作为测试数据集，计算其随机人群匹配概率、似然比和族群成分，结果见表 3 和图 4。在所有测试样本中预测准确率较高的人群是遗传结构比较单纯的人群，如非洲尼日利亚人(ESN)的准确率为 100%，欧洲芬兰人(FIN)准确率为 100%，说明该体系对实际样本的区分能力较为稳定。

我国地处东亚，是一个多民族国家，中国南北方地区的一些少数民族人群在当地长期居住过程中形成了独特的体貌特征，随着战争、迁徙、通婚、融合等现象不断发生，不同人群之间出现基因交流，各地的汉族与当地的少数民族之间出现基因交流与融合，人群之间的差异是渐变的，中国地域人群的遗传结构复杂性在本研究测试人群中得到证实。

地处中国北方的达斡尔族(DIC)、鄂温克族(EIC)及蒙古族(MIC)人群是东胡后裔且都属于阿尔泰语系，在长期迁徙进化过程中与汉族人的基因交流等原因，部分个体被推断为东北亚人群或东亚和东北亚人群的混合[23,24](表 3)。比如 EIC-19 号样本的 AMP 第一位人群为东北亚，与第二位人群的 LR 值大于 10，该样本的东亚成分为 0.55，东北亚成分为 0.25，分析 EIC-19 号样本来源人遗传成分为东亚和东北亚人群混合。青海土族(TUQ)是鲜卑支系吐谷浑人后裔，在历史进程中不断吸收融合了羌、藏、汉、蒙古等民族的成分[25]，本研究中，基于似然比统计 TUQ 的 23 名个体被推断为东北亚或者东亚和东北亚的混合(见表 3)，比如 TUQ-71 号样本，AMP 第一位人群为东北亚，与第二位人群的 LR 值大于 10，其东北亚成分为 0.94，推断为东北亚，该结果与其历史起源相符。143 名西藏藏族(CTT)个体中 112 名表现出大于 0.7 的东北亚遗传成分，其原因可能是藏缅语族人群的北方起源，杜若甫等[26]和 Gayden 等[27]对藏族常染色体遗传标记的研究证明了其北方起源。

汉族是中国的主体民族，源于北方古老的华夏

部落[28,29]，前期研究表明汉族人群具有混合特征，基于常染色体 SNP 频率的主成分分析呈现明显的南北分化[30]。图 4A 可以看出汉族人群自北向南表现出：北方成分逐渐减少，南方成分逐渐增多的趋势。表 3 的似然比统计结果中青海汉族(CHQ)、山东青岛汉族(CHL)和河南汉族(CHN)中国北方汉族人群的样本被推断为东亚人群的比例分别为 74.0%、90.4%和 81.3%，证明其对中国北方汉族人群推断的准确率较高。广西汉族(HGC)、广东客家汉族(HCM)和广东汉族(CHG)等中国南方汉族人群表现东亚和东南亚成分的混合，与自秦以来汉族人群的南迁及在迁徙过程中不断与南方少数民族交流融合等现象相符[31]。

广西京族(KGC)[32]约在 16 世纪初从越南的涂山等地迁来中国，和陆续迁来的汉族、壮族等各族人群进行了基因交流[33]，广西侗族(DGC)和云南傣族(CDY)起源于南方的百越族[34,35]。在本研究中，KGC、DGC 和 CDY 人群的族群成分以东南亚为主，混有一定比例的东亚成分(表 3)。基于似然比的统计结果中部分个体被推断为东亚或者东亚和东南亚的混合，比如 DGC-28 号样本 AMP 第一位人群为东亚，与第二位人群的 LR 值大于 10，其东亚成分为 0.64，东南亚成分为 0.17，族群推断为东亚和东南亚人群的混合，这可能与它们在历史进程中与汉族通婚、基因融合等有关。另外，民族是文化层面的概念，不同民族人群长期迁移与融合，族群推断结果可能出现与户籍登记不符的情况。在实际案件应用中，应综合分析似然比和族群成分。

综上所述，本研究前期基于质谱检测平台构建的 74-plex SNPs 复合检测体系在模板 DNA 量最低为 1.5 ng 时均可正确判型，适用于微量检材的检测。该体系实现了全球十个区域人群的区分，对东亚人群的南北方遗传成分可以进一步区分。检测结果可为案件提供更加详细的侦查线索。

参考文献(References)：

[1] Tishkoff SA, Kidd KK. Implications of biogeography of human populations for 'race' and medicine. *Nat Genet*, 2004, 36(11 Suppl): S21–S27. [DOI]

[2] Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM. Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 2008, 319(5866): 1100–1104. [DOI]

[3] Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. A genetic atlas of human admixture history. *Science*, 2014, 343(6172): 747–751. [DOI]

[4] Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, Jorde LB, Posukh OL, Sahakyan H, Watkins WS, Yepiskoposyan L, Abdullah MS, Bravi CM, Capelli C, Hervig T, Wee JT, Tyler-Smith C, van Driem G, Romero IG, Jha AR, Karachanak-Yankova S, Toncheva D, Comas D, Henn B, Kivisild T, Ruiz-Linares A, Sajantila A, Metspalu E, Parik J, Villems R, Starikovskaya EB, Ayodo G, Beall CM, Di Rienzo A, Hammer MF, Khusainova R, Khusnutdinova E, Klitz W, Winkler C, Labuda D, Metspalu M, Tishkoff SA, Dryomov S, Sukernik R, Patterson N, Reich D, Eichler EE. Global diversity, population stratification, and selection of human copy-number variation. *Science*, 2015, 349(6253): aab3761. [DOI]

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

[6] Santos C, Phillips C, Oldoni F, Amigo J, Fondevila M, Pereira R, Carracedo Á, Lareu MV. Completion of a worldwide reference panel of samples for an ancestry informative Indel assay. *Forensic Sci Int Genet*, 2015, 17: 75–80. [DOI]

[7] Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. *Hum Mutat*, 2010, 29(5): 648–658. [DOI]

[8] Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat*, 2010, 30(1): 69–78. [DOI]

[9] Nievergelt CM, Maihofer AX, Shekhtman T, Libiger O, Wang X, Kidd KK, Kidd JR. Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investig Genet*, 2013, 4(1): 13. [DOI]

[10] Frudakis T, Venkateswarlu K, Thomas MJ, Gaskin Z, Ginjupalli S, Gunturi S, Ponnuswamy V, Natarajan S, Nachimuthu PK. A classifier for the SNP-based inference of ancestry. *J Forensic Sci*, 2003, 48(4): 771–782. [DOI]

[11] Phillips C, Salas A, Sánchez JJ, Fondevila M, Gómez-Tato

A, Álvarez-Dios J, Calaza M, de Cal MC, Ballard D, Lareu MV, Carracedo A. Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet*, 2007, 1(3–4): 273–280. [DOI]

[12] Wei YL, Wei L, Zhao L, Sun QF, Jiang L, Zhang T, Liu HB, Chen JG, Ye J, Hu L, Li CX. A single-tube 27-plex SNP assay for estimating individual ancestry and admixture from three continents. *Int J Legal Med*, 2015, 130(1): 27–37. [DOI]

[13] Wei L, Wei YL, Sun QF, Wang YY, Li CX. The development of a 27-plex SNP multiplex system. *Chin J Foren Med*, 2016, 31(1): 13–17.
魏丽, 魏以梁, 江丽, 孙启凡, 王英元, 李彩霞. 27-plex SNPs 复合扩增检测体系构建与应用评价. 中国法医学杂志, 2016, 31(1): 13–17. [DOI]

[14] Jiang L, Sun QF, Ma Q, Zhao WT, Liu J, Zhao L, Ji AQ, Li CX. Optimization and validation of analysis method based on 27-plex SNP panel for ancestry inference. *Hereditas (Beijing)*, 2017, 39(2): 166–173.
江丽, 孙启凡, 马泉, 赵雯婷, 刘京, 赵蕾, 季安全, 李彩霞. 27-plex SNP 种族推断方法的优化及验证. 遗传, 2017, 39(2): 166–173. [DOI]

[15] 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]

[16] Ma M, Liu J, Hu S, Zhang T, Zhou H, Feng BQ, Liu HB, Li B, Li CX. The validation study of 74-plex SNP assay for ancestry inference. *Chin J Foren Med*, 2019, 34(4): 324–329.
马咪, 刘京, 胡胜, 张涛, 周浩, 冯保强, 刘海渤, 李蓓, 李彩霞. 74 重 SNP 族群来源推断体系准确性验证研究. 中国法医学杂志, 2019, 34(4): 324–329. [DOI]

[17] Ren P, Liu J, Zhao H, Fan XP, Xu YC, Li CX. Construction of a rapid microfluidic-based SNP genotyping (MSG) chip for ancestry inference. *Forensic Sci Int Genet*, 2019, 41: 145–151. [DOI]

[18] Clendenen TV, Rendleman J, Ge W, Koenig KL, Wirgin I, Currie D, Shore RE, Kirchhoff T, Zeleniuch-Jacquotte A. Genotyping of single nucleotide polymorphisms in DNA isolated from serum using sequenom MassARRAY technology. *PLoS One*, 2015, 10(8): e0135943. [DOI]

[19] Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes*, 2007, 7(4): 574–578. [DOI]

[20] Liu J, Li S, Jang L, Zhao L, Zhao WT, Feng L, Liu HB, Ji AQ, Li CX. DNA Ancestry Analyzer: an automatic program for ancestry inference of unknown individuals. *Life Sci Res*, 2018, 22(1): 3–7, 41.
刘京, 李盛, 江丽, 赵蕾, 赵雯婷, 丰蕾, 刘海渤, 季安全, 李彩霞. 对于未知来源个体进行族群推断的自动分析系统. 生命科学研究, 2018, 22(1): 3–7, 41. [DOI]

[21] Phillips C, McNevin D, Kidd KK, Lagacé R, Wootton S, de la Puente M, Freire-Aradas A, Mosquera-Miguel A, Eduardoff M, Gross T, Dagostino L, Power D, Olson S, Hashiyada M, Oz C, Parson W, Schneider PM, Lareu MV, Daniel R. MAPlex - A massively parallel sequencing ancestry analysis multiplex for Asia-Pacific populations. *Forensic Sci Int Genet*, 2019, 42: 213–226. [DOI]

[22] Sun K, Yun LB, Zhang C, Shao CC, Gao TZ, Zhao ZQ, Hou YP, Xie JH, Tang QQ. Evaluation of 12 Multi-InDel markers for forensic ancestry prediction in Asian populations. *Forensic Sci Int Genet*, 2019, 43: 102155. [DOI]

[23] Xu Y, Zhang XL, Zhang QC, Cui YQ, Zhou H, Zhu H. Genetic relationship between ancient Khitan and modern Daur. *J Jilin Univ(Sci Ed)*, 2006, 44(6): 997–1000.
许月, 张小雷, 张全超, 崔银秋, 周慧, 朱泓. 古代契丹与现代达斡尔遗传关系分析. 吉林大学学报(理学版), 2006, 44(6): 997–1000. [DOI]

[24] Zhu H. The ethnic type and related issues of the Khitay. *Acta Sci Natl Univ Neimongol(Hum Soc Sci)*, 1991(2): 36–41.
朱泓. 契丹族的人种类型及其相关问题. 内蒙古大学学报(哲学社会科学版), 1991, (2): 36–41. [DOI]

[25] Fan H. Relationship among 28 Chinese populations in western and southern of China based on STR loci. *J Kunming Med Univ*, 2006.
范浩. 应用 STR 位点研究中国西、南部 28 个民族群体族源关系. 昆明医学院, 2006. [DOI]

[26] Du R, Xiao C, Cavalli-Sforza LL. Genetic distances between Chinese populations calculated on gene frequencies of 38 loci. *Sci China C Life Sci*, 1997, 40(6): 613–621. [DOI]

[27] Gayden T, Mirabal S, Cadenas AM, Lacau H, Simms TM, Morlote D, Chennakrishnaiah S, Herrera RJ. Genetic insights into the origins of Tibeto-Burman populations in the Himalayas. *J Hum Genet*, 2009, 54(4): 216–223. [DOI]

[28] Zhao YB, Yu CC, Zhou H. Study on the origin and development of the Han Chinese. *Jilin Norm Univ J(Nat Sci Ed)*, 2012, 33(4): 45–49.
赵永斌, 于长春, 周慧. 汉族起源与发展的遗传学探索. 吉林师范大学学报(自然科学版), 2012, 33(4): 45–49. [DOI]

[29] Huang YZ. The historical migration of the Han population

and the color pattern of the southern Han folk songs. *Musicol China*, 1989(4): 36–48.
黄允箴. 汉族人口的历史迁徙与南方汉族民歌的色彩格局. 中国音乐学, 1989, (4): 36–48. [DOI]

[30] Xu SH, Yin XY, Li LS, Jin WF, Lou HY, Yang L, Gong XH, Wang HY, Shen YP, Pan XD, He YG, Yang YJ, Wang Y, Fu WQ, An Y, Wang JC, Tan JZ, Qian J, Chen XL, Zhang X, Sun YF, Zhang XJ, Wu BL, Jin L. Genomic dissection of population substructure of Han Chinese and its implication in association studies. *Am J Hum Genet*, 2009, 85(6): 762–774. [DOI]

[31] Wang YS. Historical evolution of the Han population and its survival area. *Hist Teach*, 2010, (6): 3–7.
王跃生. 汉族人口及其生存区域的历史演变. 历史教学(下半月刊), 2010, (6): 3–7. [DOI]

[32] Mo L, Wang CL. Analysis on the characteristics of Guangxi Jing population. *Stud Ethn Guangxi*, 1990, (3): 17–22.
莫龙, 王春林. 广西京族人口特点浅析. 广西民族研究, 1990, (3): 17–22. [DOI]

[33] Jin TB, Gao Y, Chen T, Yan HX, Li SB. Genetic relationships of 15 populations of Guangxi province. *J Xi'an Jiaotong Univ (Med Sci)*, 2004, 25(5): 422–424, 429.
金天博, 高雅, 陈腾, 阎春霞, 李生斌. 广西地区 15 个不同民族人群的群体遗传学关系. 西安交通大学学报(医学版), 2004, 25(5): 422–424, 429. [DOI]

[34] Tang JP, Yu X, Jiang FH, Yu XJ. Analyzing population differentiation between Hall and other population of Guangxi. *Int J Genet* , 2008, 31(6): 409–412.
唐剑频, 于昕, 蒋丰慧, 于晓军. 广西汉族群体与其他群体的群体差异分析. 国际遗传学杂志, 2008, 31(6): 409–412. [DOI]

[35] Qian YP. Rearch on genetic diversity of five Yunnan ethnic groups in China[Dissertation]. *Peking Union Medical College*, 1999.
钱亚屏, 中国云南 5 个民族的遗传多样性研究[学位论文]. 中国协和医科大学, 1999. [DOI]

(责任编委: 赖江华)