



基于PCA-XGBoost方法的洲际人群 生物地理祖源推断模型研究*

姚昊天^{1,2)**} 江丽^{2)**} 王春年^{1,2)} 范虹^{1)***} 李彩霞^{2)***}

(¹⁾ 陕西师范大学计算机科学学院, 西安 710119;

(²⁾ 公安部鉴定中心, 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)

摘要 目的 通过DNA推断个体的生物地理祖源 (biogeographical ancestry, BGA) 在人类学、法医学等领域广受关注。目前常用方法是使用几十个祖先信息单核苷酸多态性 (single nucleotide polymorphism, SNP) 位点, 通过主成分分析 (principal component analysis, PCA)、似然比 (likelihood ratio, LR) 等方法判断个体的祖源。伴随高通量测序技术的发展, 批量获取人群样本的高密度SNP数据集变得容易, 同时计算机领域中机器学习等技术的引入, 使得BGA研究发展出新的变化。本研究旨在构建适应高密度SNP数据, 且具有高准确率和良好泛化能力的BGA推断模型。方法 首先基于307 866个SNP的数据, 使用机器学习领域中的监督学习模型XGBoost, 构建了基于多维度主成分 (principal component, PC) 的PCA-XGBoost推断模型, 其次基于LR对推断结果进行评估和优化模型, 确定了最佳PC数目和模型训练轮数, 最后在其他公共数据的测试集上进一步验证模型的表现。结果 基于LR的结果评估方法, 模型在参考集中人群预测准确率可以达到95%以上, 在测试集中准确率可以达到90%以上, 结论 PCA-XGBoost模型具有较高的洲际人群预测准确性, 基于LR的结果评估方法有助于对预测结果的可靠性进行进一步评估。该模型具有很好的泛化能力, 更换参考集的人群数据后, 有望实现更加精细的人群分析。

关键词 生物地理祖源推断, 监督学习, 主成分分析, XGBoost模型

中图分类号 TP312, R89, D919.2

DOI: 10.16476/j.pibb.2023.0453

生物地理祖源 (biogeographical ancestry, BGA) [1] 推断在群体遗传学、法医遗传学、流行病学和药物基因组学等多个领域有着重要的应用, 如: 解释族群扩张、迁移和混合 [2]; 识别嫌疑人和无名尸体的来源; 识别疾病易感性标记 [3] 以及校正人群分层 [4] 等。目前在法医DNA族群推断领域中已报道了大量针对洲际人群和部分亚人群筛选祖先信息标记 (ancestry informative markers, AIMs) [5-6]、构建检测体系和推断模型的研究 [7-10]。然而, 筛选AIMs需要基于特定人群, 且在人群区分程度上存在一定的限制。

近年来, 伴随高密度单核苷酸多态性 (single nucleotide polymorphism, SNP) 检测技术 (如下一代测序和SNP芯片等) 的成本下降, 群体基因组数据集规模的迅速增长, 空间群体遗传学 [11] 这一概念再次受到关注, 基于高密度SNP数据建立

的生物地理祖源推断模型应运而生。基于高密度位点的BGA推断常以基于模型假设的统计学方法为主, 利用已知的基因型数据进行参数估计 [12-13], 得到个体或群体生物学属性的时空分布。其中具有代表性的是基因频率空间分布模型, 通过最大似然或近似贝叶斯方法 [14] 估计, 得到基因型的空间概率分布。但是这类模型对于高维度生物学特征耗时较长, 更为复杂的模型假设使得拟合更加困难, 难

* 国家重点研发计划 (2022YFC3341004), 国家自然科学基金 (82171870), 陕西省自然科学基金 (2022ZJ-39), 法医遗传学公安部重点实验室开放课题 (2023FGKFKT01) 和公安部鉴定中心基本科研业务费专项资金 (2022JB020) 资助项目。

** 并列第一作者。

*** 通讯联系人。

范虹 Tel: 15929807273, E-mail: fanhong@snnu.edu.cn

李彩霞 Tel: 010-83752706, E-mail: licaixia@tsinghua.org.cn

收稿日期: 2023-11-16, 接受日期: 2024-06-05

以实现推断精度的提升。而基于机器学习 (machine learning, ML) 的方法可以避免使用理想化的参数模型, 通过数据学习和优化模型, 不需要过多的先验知识, 这类方法对于高维度基因组数据也有着很好的适应性。群体遗传学中常用的机器学习方法是将非监督学习 (如: 主成分分析 (principal component analysis, PCA)、Structure 方法) 探索到的经验与监督学习方法 (如线性回归方法) 结合, 计算遗传距离或预测祖源类别^[2, 11], 然而这些方法往往只保留了较少维度的特征 (如 2~3 维的主成分 (principal component, PC)), 其数据的利用率和预测准确率 (约 50% 概率准确定位到 400 km 之内; 洲内亚人群约 77% 的预测准确率) 还有待提升。近来, 深度神经网络 (又名深度学习) 作为新一代的机器学习算法在群体遗传研究等领域有所应用^[15], 然而由于其“黑盒特性”, 这类模型的可解释性存在不足^[16]。为此, 本研究结合监督学习与非监督学习方法, 在模型构建中同时发挥两者的优势。

本研究建立了一种结合非监督学习的 PCA 算法和监督学习的 XGBoost 算法^[17] 的机器学习模型, 即 PCA-XGBoost 生物地理祖源推断模型。该

模型使用了高密度的 SNP 基因型数据, 处理更多维的输入特征, 在洲际和人群层面验证 BGA 的推断效能。本研究基于千人基因组计划三期 (The 1000 Genomes Project Phase 3, 1KGPhase3) 数据集^[18] 的常染色体 SNP 构建参考集 (共 307 866 个 SNP 位点, 26 个人群, 2 504 个样本), 并采用 10 次 10 折交叉方法评估模型的准确性和性能, 最后收集公开数据 (共 76 个人群, 700 个样本) 测试该模型的预测能力。

1 材料与方法

1.1 样本信息

参考样本共计 5 个大洲 26 个人群的 2 504 人份, 均来源于千人基因组计划三期的数据集 (1KGPhase3); 测试样本共计 76 个人群的 700 人份, 包括: The Allen Ancient DNA Resource (AADR) 中的 1240K Version v54.1.p1 数据集^[19] 中的现代人群个体, 排除来源为 1KGPhase3 的个体, 选取参考人群所在国家或地区的人群; 厦门大学王传超实验室 (C.C. Wang Lab) 发表的部分中国南北方汉族样本^[20-22]。详细样本信息见表 1。

Table 1 Samples of 26 reference populations and 76 test populations

World region	Population ¹⁾	Abbr.	Samp.	Source
Europe (EUR)	Utah residents (CEPH) with Northern and Western European ancestry	CEU	99	1KGPhase3
	Finnish in Finland	FIN	99	1KGPhase3
	British in England and Scotland	GBR	91	1KGPhase3
	Iberian populations in Spain	IBS	107	1KGPhase3
	Toscani in Italia	TSI	107	1KGPhase3
	<i>Polish in Poland</i>	<i>T_POL</i>	<i>1</i>	<i>AADR</i>
	<i>Czech in Czech Republic</i>	<i>T_CZE</i>	<i>1</i>	<i>AADR</i>
	<i>Finnish in Finland²⁾</i>	<i>T_FIN</i>	<i>3</i>	<i>AADR</i>
	<i>Saami in Finland³⁾</i>	<i>T_SMF</i>	<i>1(1)</i>	<i>AADR</i>
	<i>Saami in Utsjoki, Finland</i>	<i>T_SUF</i>	<i>2</i>	<i>AADR</i>
	<i>English in Kent, the United Kingdom (UK)²⁾</i>	<i>T_EKU</i>	<i>2</i>	<i>AADR</i>
	<i>Orcadian in the UK</i>	<i>T_OCU</i>	<i>2</i>	<i>AADR</i>
	<i>Orcadian in Orkney Islands, the UK</i>	<i>T_OOU</i>	<i>15</i>	<i>AADR</i>
	<i>Spanish in Castilla-La Mancha, Spain</i>	<i>T_SCS</i>	<i>2</i>	<i>AADR</i>
	<i>Bergamo in Italy</i>	<i>T_BGM</i>	<i>1</i>	<i>AADR</i>
	<i>Italian in Northern Division, Italy</i>	<i>T_INI</i>	<i>21</i>	<i>AADR</i>
	<i>Sardinian in Italy</i>	<i>T_SDI</i>	<i>32</i>	<i>AADR</i>
	<i>Tuscan in Italy²⁾</i>	<i>T_TSI</i>	<i>2</i>	<i>AADR</i>
East Asia (EAS)	Chinese Dai in Xishuangbanna, China	CDX	93	1KGPhase3
	Han Chinese in Beijing, China	CHB	103	1KGPhase3
	Han Chinese South	CHS	105	1KGPhase3
	Japanese in Tokyo, Japan	JPT	104	1KGPhase3

Continued to Table 1

World region	Population ¹⁾	Abbr.	Samp.	Source
	Kinh in Ho Chi Minh City, Vietnam	KHV	99	1KGPhase3
	<i>Dai in China</i> ²⁾	<i>T_DCN</i>	<i>14</i>	<i>AADR</i>
	<i>Han in China</i>	<i>T_HCH</i>	<i>47</i>	<i>AADR</i>
	<i>Mongolia in China</i>	<i>T_MGC</i>	<i>11</i>	<i>AADR</i>
	<i>Ami in Taiwan, China</i>	<i>T_AMT</i>	<i>2</i>	<i>AADR</i>
	<i>Atayal in Taiwan, China</i>	<i>T_ATT</i>	<i>1</i>	<i>AADR</i>
	<i>Han in Shanxi, China</i> ²⁾	<i>T_HSX</i>	<i>8</i>	<i>C.C. Lab</i>
	<i>Han in Guangdong, China</i> ²⁾	<i>T_HGD</i>	<i>7</i>	<i>C.C. Lab</i>
	<i>Han in Guangxi, China</i>	<i>T_HGX</i>	<i>37</i>	<i>C.C. Lab</i>
	<i>Han in Guizhou, China</i>	<i>T_HGZ</i>	<i>14</i>	<i>C.C. Lab</i>
	<i>Han in Sichuan, China</i>	<i>T_HSC</i>	<i>7</i>	<i>C.C. Lab</i>
	<i>Han in Yunnan, China</i>	<i>T_HYN</i>	<i>16</i>	<i>C.C. Lab</i>
	<i>Han in Chongqing, China</i>	<i>T_HCQ</i>	<i>8</i>	<i>C.C. Lab</i>
	<i>Japanese in Japan</i> ²⁾	<i>T_JPJ</i>	<i>29</i>	<i>AADR</i>
	<i>Japanese in Tokyo, Japan</i> ²⁾	<i>T_JPT</i>	<i>1</i>	<i>AADR</i>
	<i>Kinh in Vietnam</i> ²⁾	<i>T_KHV</i>	<i>2</i>	<i>AADR</i>
Africa (AFR)	African Caribbean in Barbados	ACB	96	1KGPhase3
	African Ancestry in Southwest US	ASW	61	1KGPhase3
	Esan in Nigeria	ESN	99	1KGPhase3
	Gambian in Western Division, The Gambia	GWD	113	1KGPhase3
	Luhya in Webuye, Kenya	LWK	99	1KGPhase3
	Mende in Sierra Leone	MSL	85	1KGPhase3
	Yoruba in Ibadan, Nigeria	YRI	108	1KGPhase3
	<i>Esan in Nigeria</i> ²⁾	<i>T_ESN</i>	<i>2</i>	<i>AADR</i>
	<i>Gambian in Western Division, The Gambia</i> ²⁾	<i>T_GWD</i>	<i>2</i>	<i>AADR</i>
	<i>Bantu in Kenya</i>	<i>T_BTK</i>	<i>13</i>	<i>AADR</i>
	<i>Luo in Bondo District, Kenya</i>	<i>T_LBK</i>	<i>2</i>	<i>AADR</i>
	<i>Somali in Garissa, Kenya</i>	<i>T_SGK</i>	<i>1</i>	<i>AADR</i>
	<i>Masai in Kinyawa, Kenya</i>	<i>T_MKK</i>	<i>2</i>	<i>AADR</i>
	<i>Luhya in Webuye, Kenya</i> ²⁾	<i>T_LWK</i>	<i>2</i>	<i>AADR</i>
	<i>Mende in Sierra Leone</i> ²⁾	<i>T_MSL</i>	<i>2</i>	<i>AADR</i>
	<i>Yoruba in Nigeria</i> ²⁾	<i>T_YRN</i>	<i>26</i>	<i>AADR</i>
Americas (AMR)	Colombian in Medellin, Colombia	CLM	94	1KGPhase3
	Mexican Ancestry in Los Angeles, California	MXL	64	1KGPhase3
	Peruvian in Lima, Peru	PEL	85	1KGPhase3
	Puerto Rican in Puerto Rico	PUR	104	1KGPhase3
	<i>Piapoco in Colombia</i>	<i>T_PPC</i>	<i>9</i>	<i>AADR</i>
	<i>Huichol in Mexico</i> ³⁾	<i>T_HCM</i>	<i>1(1)</i>	<i>AADR</i>
	<i>Mayan in Mexico</i>	<i>T_MYM</i>	<i>24</i>	<i>AADR</i>
	<i>Pima in Mexico</i>	<i>T_PMM</i>	<i>15</i>	<i>AADR</i>
	<i>Mixtec in San Andres Nuxino, Mexico</i>	<i>T_MSM</i>	<i>2</i>	<i>AADR</i>
	<i>Zapotec in San Juan Guelavia, Mexico</i>	<i>T_ZSM</i>	<i>2</i>	<i>AADR</i>
	<i>Mixe in Tamazulapan, Mexico</i>	<i>T_MTM</i>	<i>2</i>	<i>AADR</i>
	<i>Mixe in Oaxaca, Tamazulapan, Mexico</i>	<i>T_MOT</i>	<i>1</i>	<i>AADR</i>
	<i>Nahua in Zitala, Mexico</i>	<i>T_NZM</i>	<i>1</i>	<i>AADR</i>
	<i>Aymara in Peru</i> ³⁾	<i>T_AMP</i>	<i>1(1)</i>	<i>AADR</i>
	<i>Quechua in Peru</i>	<i>T_QCP</i>	<i>3</i>	<i>AADR</i>

Continued to Table 1				
World region	Population ¹⁾	Abbr.	Samp.	Source
South Asia (SAS)	Bengali in Bangladesh	BEB	86	1KGPhase3
	Gujarati Indian in Houston, TX	GIH	103	1KGPhase3
	Indian Telugu in the UK	ITU	102	1KGPhase3
	Punjabi in Lahore, Pakistan	PJL	96	1KGPhase3
	Sri Lankan Tamil in the UK	STU	102	1KGPhase3
	<i>Banglali in Dhaka, Bangladesh²⁾</i>	<i>T_BDB</i>	<i>2</i>	<i>AADR</i>
	<i>Indian in India³⁾</i>	<i>T_IDI</i>	<i>1(1)</i>	<i>AADR</i>
	<i>Irula in India³⁾</i>	<i>T_IRI</i>	<i>12(10)</i>	<i>AADR</i>
	<i>Khonda in India</i>	<i>T_KHI</i>	<i>1</i>	<i>AADR</i>
	<i>Bengali in India³⁾</i>	<i>T_BGL</i>	<i>1(1)</i>	<i>AADR</i>
	<i>Birhor in India³⁾</i>	<i>T_BHI</i>	<i>9(9)</i>	<i>AADR</i>
	<i>Jarawa in India³⁾</i>	<i>T_JRI</i>	<i>4(4)</i>	<i>AADR</i>
	<i>Onge in India³⁾</i>	<i>T_OGI</i>	<i>6(6)</i>	<i>AADR</i>
	<i>Punjabi in India³⁾</i>	<i>T_PJI</i>	<i>1(1)</i>	<i>AADR</i>
	<i>Rajput in India³⁾</i>	<i>T_RJI</i>	<i>10(10)</i>	<i>AADR</i>
	<i>Riang in India³⁾</i>	<i>T_RII</i>	<i>10(10)</i>	<i>AADR</i>
	<i>Uttar Pradesh Brahmins in India³⁾</i>	<i>T_UPB</i>	<i>10(10)</i>	<i>AADR</i>
	<i>Vellalar in India³⁾</i>	<i>T_VLI</i>	<i>9(9)</i>	<i>AADR</i>
	<i>Brahmin in Visakhapatnam, India</i>	<i>T_BVI</i>	<i>2</i>	<i>AADR</i>
	<i>Kapu in Visakhapatnam, India</i>	<i>T_KVI</i>	<i>2</i>	<i>AADR</i>
	<i>Madiga in Visakhapatnam, India</i>	<i>T_MDV</i>	<i>2</i>	<i>AADR</i>
	<i>Mala in Visakhapatnam, India</i>	<i>T_MLV</i>	<i>2</i>	<i>AADR</i>
	<i>Relli in Visakhapatnam, India</i>	<i>T_RVI</i>	<i>2</i>	<i>AADR</i>
	<i>Yadava in Visakhapatnam, India</i>	<i>T_YVI</i>	<i>2</i>	<i>AADR</i>
	<i>Balochi in Pakistan</i>	<i>T_BLP</i>	<i>26</i>	<i>AADR</i>
	<i>Brahui in Pakistan</i>	<i>T_BHP</i>	<i>28</i>	<i>AADR</i>
	<i>Burusho in Pakistan</i>	<i>T_BRP</i>	<i>26</i>	<i>AADR</i>
	<i>Hazara in Pakistan³⁾</i>	<i>T_HZP</i>	<i>24(3)</i>	<i>AADR</i>
	<i>Kalash in Pakistan</i>	<i>T_KLP</i>	<i>24</i>	<i>AADR</i>
	<i>Makrani in Pakistan</i>	<i>T_MKP</i>	<i>27</i>	<i>AADR</i>
	<i>Pathan in Pakistan</i>	<i>T_PTP</i>	<i>26</i>	<i>AADR</i>
	<i>Punjabi in Lahore, Pakistan²⁾</i>	<i>T_PJL</i>	<i>4</i>	<i>AADR</i>
	<i>Sindhi in Pakistan</i>	<i>T_SDP</i>	<i>25</i>	<i>AADR</i>
Total			3 204	

¹⁾Test populations for model validation are highlighted in bold and italic, others are reference populations. ²⁾The test populations overlap with the reference population. ³⁾The population includes pseudo-diploid samples, with the specific number of samples noted in parentheses.

1.2 位点筛选

从参考数据集提取 Global Screening Array (GSA)、Chinese Genotyping Array (CGA) 芯片交集位点 (共553 849个SNPs) 中的常染色体SNP位点, 获得480 698个SNP位点。使用Plink v1.9软件^[23-24]进行位点过滤, 过滤条件为位点检出率 (geno 参数, 小于1%)、次要等位基因频率 (minor allele frequency, MAF) (maf 参数, 小于0.01)、Hardy-Weinberg 平衡 (Hardy-Weinberg

equilibrium, HWE) 检验 ($P>10^{-10}$)、连锁不平衡 (linkage disequilibrium, LD) 检验 (indep-pairwise 参数, 滑动窗口500、步长50、 $r^2>0.8$), 最终获得307 866个SNP位点。

1.3 人群遗传结构分析

使用GCTA v1.94软件进行参考数据集人群的PCA。使用ADMIXTURE v1.3.0软件进行参考数据集人群的族群成分分析, 10折交叉验证 (-CV=10) 运行^[25], K值为2~10。使用R v3.14的ggplot2

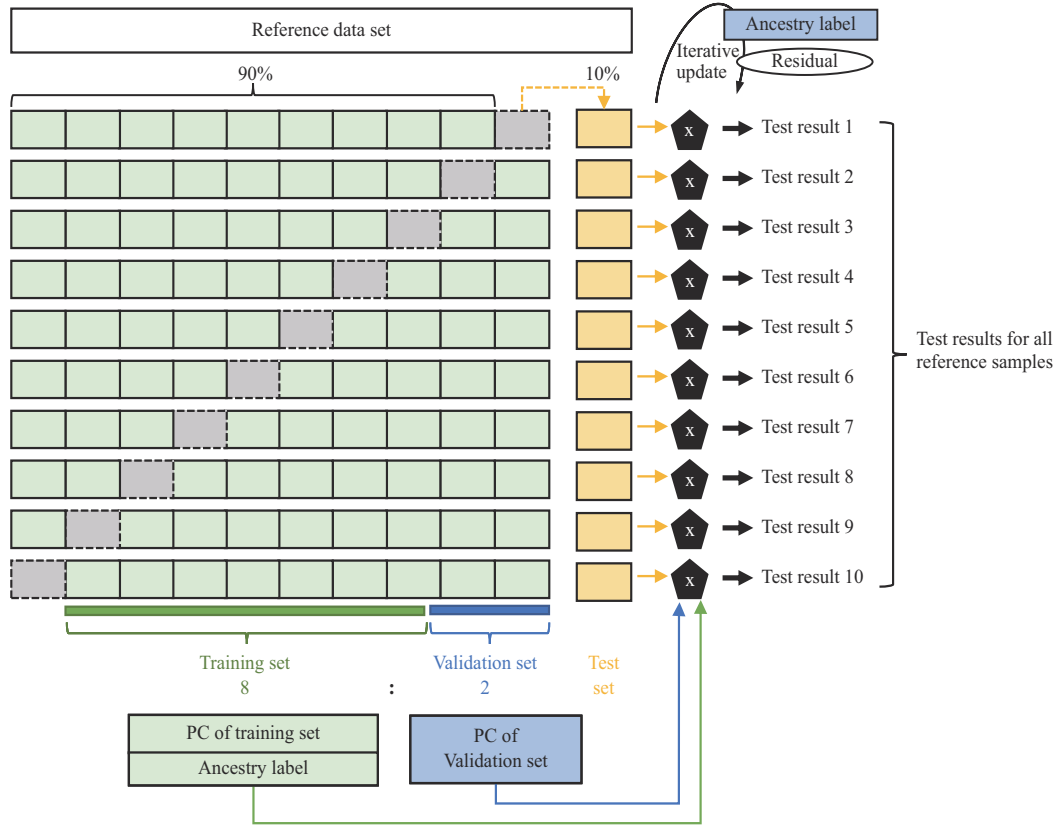


Fig. 2 Sample division of reference dataset and model testing based on reference dataset

According to the proportion of the population, 10% of the samples are selected as the test set; the 10% test result covers every reference sample.

(xthPred) 对应的概率值 (xthProb) 之比值, 每个测试样本的 LR_x 值随 x 单调递增, LR 值可以反映出模型对每一位预测结果的确定性。

本研究基于 LR 值的评价结果有 3 种结论 (设第 k 位预测结果 $kthPred$ 与真实祖源一致, 设 $m = \max_x (LR_x \leq 10)$, 即 $\forall x \leq m, LR_x \leq 10$; $x > m, LR_x > 10$): a. 一致性结论 (consistent conclusion, CC), $k=m=1$, 即 $LR_2 > 10$, 结论为“1stPred 即为正确类别且可以排除其他类别”; b. 不排除结论 (inconclusive conclusion, IC), $m \geq 2$, 且 $k \leq m$, 则表明预测结果排名前 m 位的类别中包括了正确类别, 结论为“预测正确但不排除前 m 个类别”; c. 错误结论 (error conclusion, EC), $k > m$, 则表明模型预测的前 m 个可能的结果中不含正确类别, 结论为“预测错误”。

采用的评价指标如下。

a. 第一位预测准确率 (accuracy of the 1stPred, 1stAcc), 即传统的准确率计算方式。其计算公式为: $1stAcc = \frac{\text{Number of test sample whoes 1stpred is right ancestry}}{\text{Total number of test samples}}$,

相当于阳性预测值 (positive predictive value, PPV), 只衡量 1stPred 的准确率;

b. 一致率 (consistency rate based on LR, $Cons_{LR}$), $Cons_{LR} = \frac{\text{Number of CC}}{\text{Total number of test samples}}$, 该值越高则越说明 1stPred 的准确率和确定性越高。

c. 不排除率 (inconclusive based on LR, $Incc_{LR}$), $Incc_{LR} = \frac{\text{Number of IC}}{\text{Total number of test samples}}$, 该值可以反映对于 1stPred 的不确定程度, 同时可以反映模型对于某种类别的区分困难度。

d. 准确率 (accuracy based on LR, Acc_{LR}), $Acc_{LR} = \frac{\text{Number of CC and IC}}{\text{Total number of test samples}}$, 即 $Acc_{LR} = Cons_{LR} + Incc_{LR}$; 该值基于一定合理怀疑度 (由 LR 阈值决定), 衡量前若干位预测结果的准确率。

e. 错误率 (error rate based on LR, Err_{LR}), $Err_{LR} = \frac{\text{Number of EC}}{\text{Total number of test samples}}$, 即 $Err_{LR} = 1 - Acc_{LR}$ 。

1.6 PCA-XGBoost模型的优化

为了研究不同数目PC下预测模型的准确性,分别用前5、10、20、40、80、160、1 233个PC作为XGBoost分类模型的输入特征,基于参考数据集按照1.4、1.5步骤进行模型构建和评估,计算每个模型的预测效能,以确定最优的PC数目 S 。然后,在原本round_num=1 000, early_stopping=100的基础上,增加训练轮数为100、300、500、2 000、3 000和4 000的数量梯度(early_stopping设置为对应round_num数量的1/10),比较不同训练轮数对于模型准确性和运行时间的影响。

1.7 PCA-XGBoost模型的验证

使用EIGENSOFT v8.0软件包中的convertf工具,将AADR数据库1240K Version v54.1.p1数据集中geno格式文件转换为bfile格式文件。使用Plink v1.9软件将所有测试样本与参考数据集合并(取SNP位点交集),使用优化后的模型对所有共700个测试样本进行预测和评估。

按照与参考人群族源所在地及民族是否相同的标准,将700个测试样本划分为两类,分析其预测结果的准确性。a. A类。108个与参考人群的祖源所在国家或地区相同,且人群类型(如民族、区域、城市等)高度相似的样本,采用1.5中的评估方法和指标;b. B类。592个仅与参考人群祖源所在国家或地区相同的测试样本,按照以下3种情况进行准确性统计:被预测到正确的国家或地区、大洲内其他参考人群以及其他大洲的参考人群。

2 结 果

2.1 参考集世界人群遗传结构

参考集人群包含了来自5个大洲、26个人群的2 504个样本,基于307 866个SNP位点,首先进行PCA分析人群遗传聚类 and 遗传关系。图3展示了前5个PC的组合空间中的PCA结果。总的来说,PC可以在一定程度上反映出人群之间地理距离、基因交流等多种因素影响下形成的遗传距离和遗传关系。如:人群在PC1-PC2空间中的分布可以大致反映其在全球地理空间中的分布,地理上相距较远的人群通常在PCA图中位于远处,而地理上接近的人群在PCA图中通常也表现为临近。综合前5维的PC空间中的结果可以发现,美洲人群的中美加勒比海的波多黎各人(PUR)和南美北部的哥伦比亚麦德林的哥伦比亚人(CLM)与欧洲人群靠近,而南美洲的秘鲁人群(PEL)则与之相距最远。美

国西南部的非裔人群(ASW)和加勒比海域的巴巴多斯非裔加勒比人群(ACB)向欧洲人群对应的簇延伸,形成了细长的簇,位于欧洲人群和非洲其他主要群体中间。另外,美洲人群和非洲人群之间的边界显得较为模糊,有少量PUR样本(约4%)明显地偏离了聚类的主要部分,与ASW和ACB相近。

考察PC对人群的区分效能可以发现,在各个二维PC空间中,5个洲际人群的区分比较容易,然而,当被进一步分成26个亚人群时,某些人群之间难以得到清晰有效的区分边界。但在不同的PC空间中,人群的重叠和分离情况可以有很大差异,即不同的PC对特定人群有不同的区分能力,如:使用PC3的维度区分南亚和其他大洲人群十分有效,没有出现类似PC1-PC2空间中的重叠;另外,相较于PC1至PC4,PC5对于东亚的5个亚人群有较好的区分能力。这表明不同PC对于人群模式的识别各有侧重,有的维度主要反映洲际人群整体的相对地理位置,有的维度反映了某些亚人群之间的区分,故纳入更多的PC有助于更全面地反映人群信息,从而实现更为精细的人群区分。

为了有效区分洲际亚人群,同时避免纳入过多PC影响模型的运行效率,本研究分析了前40个PC维度的单个解释方差(explained variance, EV)和累计解释方差(cumulative explained variance, CEV),详见表S1,并绘出其能量碎石折线图(图S1)。可以发现自PC5之后EV较小且减少趋势平缓,故首先尝试利用前10维PC构建整个生物地理祖源推断模型。

为了验证参考集中人群之间的遗传差异度和人群混合情况,并分析人群的遗传亚结构,应用ADMIXTURE进行世界人群的遗传祖先成分分析($K=2$ 至10),选取人群内个体成分一致性高且CV误差较低的 K 值为5($CV_{error}=0.452\ 99$),得到的ADMIXTURE群体及个体族群成分见图4(其他 K 值下的结果图见图S2),5个祖先成分可以解释参考集人群的遗传祖先成分,分别为在欧洲、东亚、非洲、南亚群体中占据最大比例的红色、绿色、蓝色和紫色对应的祖先成分,以及主要存在于美洲尤其是秘鲁人群(PEL)中的黄色祖先成分。欧洲、东亚、非洲人群的主要成分占据各自的绝大多数比例(通常大于95%),且各自亚人群间的一致性比较高。美洲人群的混合情况最为明显,PEL中占据主要地位的黄色祖先成分或可解释为距今约

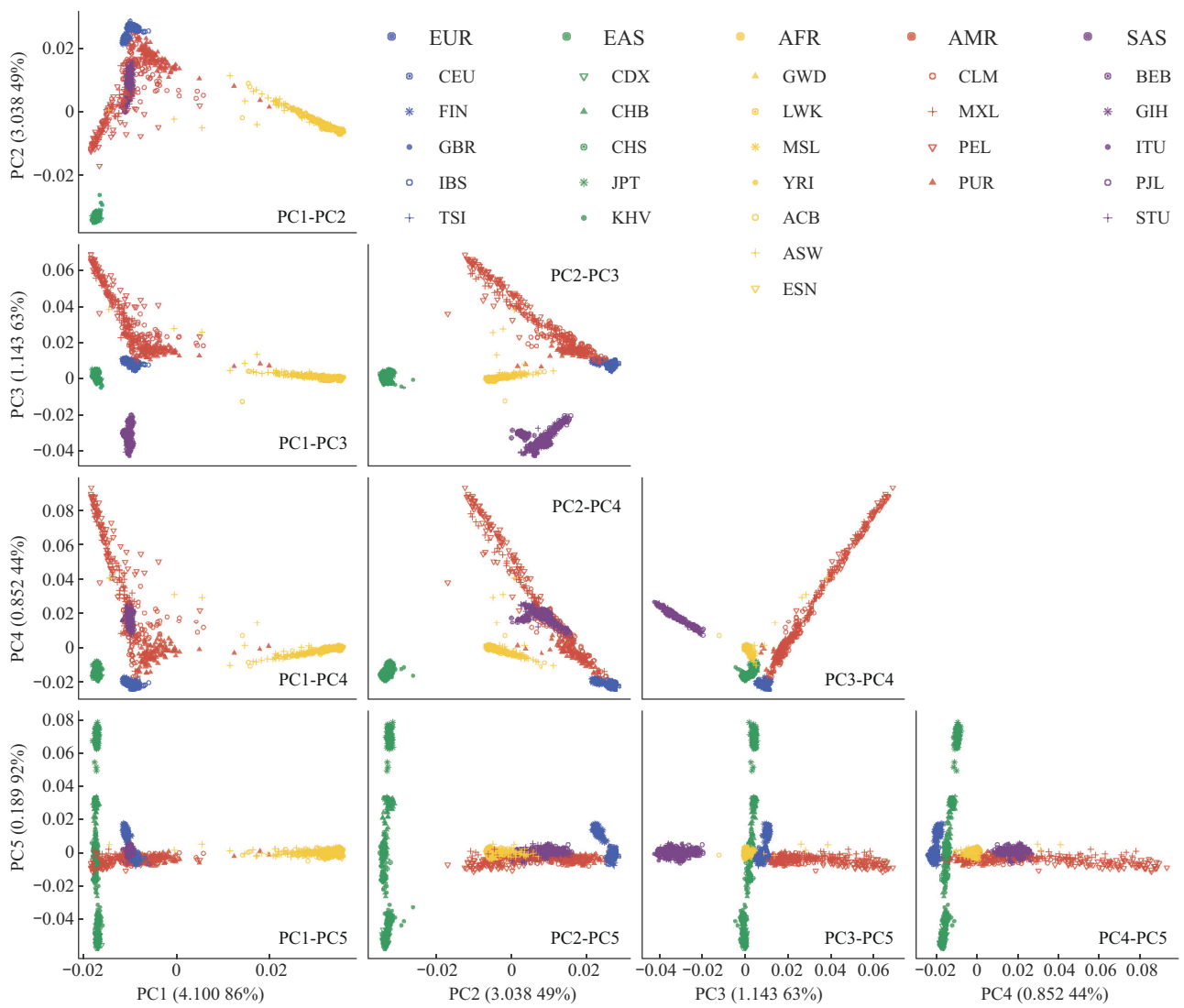


Fig. 3 PCA based on the first 5 PCs of 307 866 SNPs for 2 504 individuals from 26 reference populations
The explained variance (EV) of the corresponding dimension is noted in the brackets next to the coordinate axis.

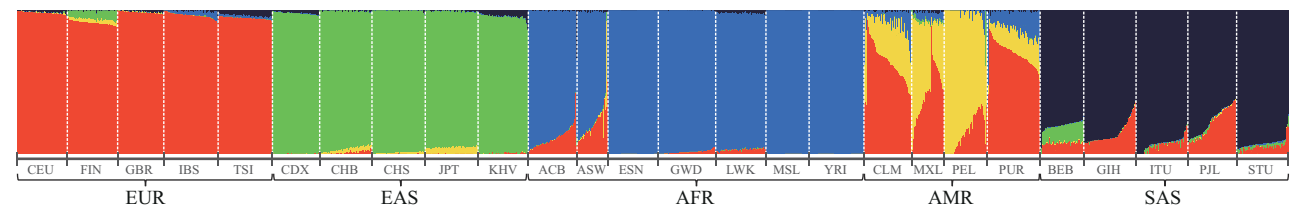


Fig. 4 ADMIXTURE results of 307 866 SNPs for 2 504 individuals from 26 reference populations
 $K=5$, $CV_{\text{error}}=0.452\ 99$.

16 500年前经白令海峡迁移至美洲大陆^[26]的美洲土著 (native American, NAMR) 成分。其他3个美洲人群都有较多的欧洲成分 (红色, 平均比例约为50%~70%) 和一部分的非洲祖先成分。不同美洲群体具有不同程度的两祖源 (EUR-NAMR) 或者三祖源 (EUR-AFR-NAMR) 混合模式的遗传背

景^[27]。非洲人群中的巴巴多斯的非裔加勒比人 (ACB) 与美国西南部有非裔美国人 (ASW) 相较于其他非洲人群混合了更多的欧洲成分 (占比约10%)。另外, 当 K 值增加时, 部分人群出现了特异性祖先成分, 例如FIN、JPT、LWK等 (图S2)。

2.2 PCA–XGBoost模型构建与评估

采用前 10 个 PC 初步构建模型（10PC-XGBoost模型），洲际和人群水平的测试准确性结

果分别见表 2 和表 3。洲际层次模型的第一位预测准确率（1stAcc）、一致率（Cons_{LR}）和准确率（Acc_{LR}）均达到 98% 以上。其中，美洲样本在 3 个

Table 2 Inferential results of 10PC–XGBoost model based on reference dataset at the continental level

TRUE/PRED	EUR	EAS	AFR	AMR	SAS	Total	1stAcc/%	Acc _{LR} /%		Err _{LR} /%
								Cons _{LR} /%	Incc _{LR} /%	
EUR	503 (2)					503	100	99.60	0.40	0
EAS		504 (0)				504	100	100	0	0
AFR			657 (1)	4 (2)		661	99.39	99.24	0.45	0.30
AMR	2 (2)		3 (1)	342 (1)		347	98.56	98.27	1.15	0.58
SAS					489 (0)	489	100	100	0	0
Total						2 504	99.64	99.48	0.36	0.16

The number in each grid represents the number of samples predicted for the corresponding continent, and the number in parentheses represents the number of samples in which the result belongs to inconclusive conclusion (IC).

Table 3 Inferential results of 10PC–XGBoost model based on reference dataset at the population level

TRUE/PRED		Accurate POP	Other POP within continent	POP in other continents	Total	1stAcc/%	Acc _{LR} /%		Err _{LR} /%
							Cons _{LR} /%	Incc _{LR} /%	
EUR	CEU	53 (48)	46 (43)	0	99	53.54	5.05	91.92	3.03
	FIN	99 (0)	0	0	99	100	100	0	0
	GBR	48 (46)	43 (41)	0	91	52.75	2.20	95.60	2.20
	IBS	105 (5)	2 (2)	0	107	98.13	93.46	6.54	0
	TSI	104 (5)	3 (3)	0	107	97.20	92.52	7.48	0
EAS	CDX	88 (7)	5 (2)	0	93	94.62	87.10	9.68	3.23
	CHB	80 (10)	23 (14)	0	103	77.67	67.96	23.30	8.74
	CHS	96 (47)	9 (6)	0	105	91.43	46.67	50.48	2.86
	JPT	104 (1)	0	0	104	100	99.04	0.96	0
	KHV	94 (7)	5 (5)	0	99	94.95	87.88	12.12	0
AFR	ACB	91 (13)	5 (3)	0	96	94.79	81.25	16.67	2.08
	ASW	53 (13)	5 (5)	3 (1)	61	86.89	65.57	31.15	3.28
	ESN	93 (14)	6 (2)	0	99	93.94	79.80	16.16	4.04
	GWD	112 (2)	1 (0)	0	113	99.12	97.35	1.77	0.88
	LWK	99 (0)	0	0	99	100	100	0	0
AMR	MSL	84 (1)	1 (1)	0	85	98.82	97.65	2.35	0
	YRI	96 (11)	12 (8)	0	98	88.89	78.70	17.59	3.70
	CLM	78 (34)	15 (12)	1 (0)	94	82.98	46.81	48.94	4.26
	MXL	34 (18)	29 (21)	1 (1)	64	53.13	25.00	62.50	12.50
	PEL	77 (14)	7 (5)	1 (1)	85	90.59	74.12	23.53	2.35
SAS	PUR	96 (8)	4 (2)	4 (2)	104	92.31	84.62	11.54	3.85
	BEB	81 (5)	5 (5)	0	86	94.19	88.37	11.63	0
	GIH	87 (9)	16 (9)	0	103	84.47	75.73	17.48	6.80
	ITU	67 (42)	35 (29)	0	102	65.69	24.51	69.61	5.88
	PJL	82 (32)	14 (12)	0	96	85.42	52.08	45.83	2.08
Total	STU	78 (54)	24 (19)	0	102	76.47	23.53	71.57	4.90
					2 504	87.02	69.21	27.96	2.84

The number in each grid represents the number of samples predicted for the corresponding category, and the number in parentheses represents the number of samples in which the result belongs to inconclusive conclusion (IC).

预测指标里数值最低, 有 3 个中美加勒比海域的 PUR 样本被误判为非洲祖源, 北美洲的 MXL 和南美北部的 CLM 人群各有 1 个样本被误判为欧洲祖源。另外, 有 4 个非洲祖源的 ASW 样本被误判为美洲祖源, 非洲和美洲的测试结果中少量相互预测混淆的结果与 PC 图中两者界限较为模糊的情况相一致。

人群层次的预测结果显示, 平均准确率较高 (97.16%), 比第一位预测准确率 (1stAcc=87.02%) 高出约 10%。同时, 参照表 S2 可以发现前 2 位累计的预测准确率 (Infir2P%) 高达 98.4%, 其中第 2 位预测准确率 (2ndP%) 占比约 11%, 而后续预测结果的准确率较低 (均小于 1%), 这说明模型对于人群的识别混淆多发生在前两位的预测结果之中。基于 LR 的准确率 ($Acc_{LR}=97.16\%$) 比第一位预测准确率 (87.02%) 高出约 10%, 接近 Infir2P% 的 98.4%, 即通过不排除 $LR<10$ 的人群类别 (主要是不排除 2ndPred) 便可减少预测错误。但 10PC-XGBoost 模型在人群层次较低的一致率 ($Incc_{LR}<70\%$) 和较高的不排除率 ($Incc_{LR}>25\%$) 表明该模型目前对于亚人群的区分能力不足。通过对测试结果的混淆矩阵 (表 S3) 的进一步分析也发现, 大洲内常有某两个亚人群之间相互预测混淆的情况, 例如: 对于欧洲的美国犹他州的北欧和西欧人后裔 (CEU) 和英格兰和苏格兰的英国人 (GBR) 人群, 模型难以区分彼此, 两者的不排除率大于 90%; 东亚中国北京汉族 (CHB) 中约有 10% 被错

误分类到中国南方汉族 (CHS) 等。

2.3 PCA-XGBoost模型的优化结果

上述 2.2 中的结果表明 10PC-XGBoost 模型可能存在选取的特征数量过少、欠拟合 (训练轮数不足等)、模型复杂度不够等需要优化的因素。

首先调整输入 XGBoost 模型的 PC 数量以优化模型参数。分别用前 5、10、20、40、80、160、1 233 个 PCs 训练 XGBoost 多分类模型 (前 1 233 维的 $CEV>60\%$)。图 5a 显示了每个模型在人群水平的预测准确性, 基于第一位预测准确率和一致率的折线可以发现随着 PC 数量的增加, 模型的准确率上升但增速显著变小, 并在 40 到 80 附近达到峰值, 而后降低。第一位预测准确率和一致率在 40 个 PC 以后没有明显提升, 且综合考虑模型训练的时间成本, 最终选择前 40 维 PC 作为模型特征。

增加特征 PC 维度数目后, 40PC-XGBoost 模型的洲际和人群水平的准确性评估结果分别见表 4 和表 5。洲际层次模型的第一位预测准确率、一致率和准确率改变不大 (虽然美洲人群的一致率有约 0.2% 的上升), 这反映了对洲际人群的区分和预测能力已经接近模型上限。人群层次的预测结果显示, 平均第一位预测准确率和平均一致率分别有约 5.3% 和 8.3% 的提升, 说明优化后模型的 1stPred 更加准确和可靠。结果表明更多维的 PC 中可能蕴含了更为精细的人群结构信息, 有助于区分和预测更加精细的亚人群。

Table 4 Inferential results of 40PC-XGBoost model based on reference dataset at the continental level										
TRUE\PRED	EUR	EAS	AFR	AMR	SAS	Total	1stAcc/%	Acc _{LR} /%		Err _{LR} /%
								Cons _{LR} /%	Incc _{LR} /%	
EUR	503 (2)					503	100	99.60	0.40	0
EAS		504 (0)				504	100	100	0	0
AFR			657 (1)	4 (2)		661	99.39	99.24	0.46	0.30
AMR	2 (2)		3 (1)	342 (1)		347	98.56	98.27	1.15	0.58
SAS					489 (0)	489	100	100	0	0
Total						2 504	99.64	99.48	0.36	0.16
The number in each grid represents the number of samples predicted for the corresponding continent, and the number in parentheses represents the number of samples in which the result belongs to inconclusive conclusion (IC).										

增加 PC 数目后, 模型处理的特征空间增大, 在训练充分度上可能变得不足, 容易导致模型未充分拟合。故在原先 1 000 次训练轮数 (round_num) 的基础上, 扩展为 100 到 4 000 的数量梯度, 比较模型训练轮数对于模型准确性的影响。

分析图 5b 不同训练轮数中 3 个指标的总体变化趋势, 可以发现模型的第一位预测准确率随训练轮数缓慢变大, 在大于 500 轮的实验中保持在 92%~93% 之间的水平。与之相反, 预测准确率在 97%~98% 的区间中缓慢下降。预测一致率在 1 000 轮以

Table 5 Inferential results of 40PC-XGBoost model based on reference dataset at the population level

TRUE\PRED		Accurate POP	Other POP within continent	POP in other continents	Total	1stAcc/%	Acc _{LR} /%		Err _{LR} /%
							Cons _{LR} /%	Incc _{LR} /%	
EUR	CEU	66 (57)	33 (30)	0	99	66.67	9.09	87.88	3.03
	FIN	99 (0)	0	0	99	100	100	0	0
	GBR	59 (47)	32 (29)	0	91	64.84	13.19	83.52	3.30
	IBS	107 (5)	0	0	107	100	95.33	4.67	0
	TSI	106 (4)	1 (1)	0	107	99.07	95.33	4.67	0
EAS	CDX	91 (4)	2 (0)	0	93	97.85	93.55	4.30	2.15
	CHB	81 (12)	22 (15)	0	103	78.64	66.99	26.21	6.80
	CHS	94 (47)	11 (9)	0	105	89.52	44.76	53.33	1.90
	JPT	104 (1)	0	0	104	100	99.04	0.96	0
	KHV	96 (4)	3 (3)	0	99	96.97	92.93	7.07	0
AFR	ACB	92 (14)	4 (3)	0	96	95.83	81.25	17.71	1.04
	ASW	56 (20)	2 (2)	3 (1)	61	91.80	59.02	37.70	3.28
	ESN	93 (18)	6 (2)	0	99	93.94	75.76	20.20	4.04
	GWD	112 (0)	1 (1)	0	113	99.12	99.12	0.88	0
	LWK	99 (0)	0	0	99	100	100	0	0
AMR	MSL	84 (4)	1 (1)	0	85	98.82	94.12	5.88	0
	YRI	101 (12)	7 (2)	0	98	93.52	82.41	12.96	4.63
	CLM	94 (9)	0	1 (0)	94	100	90.43	9.57	0
	MXL	62 (8)	0	2 (2)	64	96.88	84.38	15.63	0
	PEL	84 (3)	0	1 (1)	85	98.82	95.29	4.71	0
SAS	PUR	100 (1)	0	4 (4)	104	96.15	95.19	4.81	0
	BEB	85 (6)	1 (1)	0	86	98.84	91.86	8.14	0
	GIH	86 (11)	17 (13)	0	103	83.50	72.82	23.30	3.88
	ITU	87 (28)	15 (12)	0	102	85.29	57.84	39.22	2.94
	PJL	85 (25)	11 (11)	0	96	88.54	62.50	36.46	1.04
STU		89 (32)	13 (9)	0	102	87.25	55.88	40.20	3.92
Total					2 504	92.33	77.48	20.89	1.64

The number in each grid represents the number of samples predicted for the corresponding category, and the number in parentheses represents the number of samples in which the result belongs to inconclusive conclusion (IC).

前随着训练轮数增加显著变大，而后维持在78%~80%的水平。

在最小训练轮数（100轮）的实验中，一致率较低（18.8%），但准确率可达约为98.4%，第一位预测准确率也达到了90%以上。即在较低的训练轮数中，模型已经具备了一定的祖源分类预测能力，但是不排除率过高（约70%），增大了考虑LR时的结果空间和决策难度。另一方面，训练轮数的增加又会导致测试时间增加，一般情况下时间成本和训练轮数之间有着正相关关系，例如1 000轮的训练时间约是100轮的10倍。综合考虑模型的时间成本和准确率、不排除率的因素，本研究将在基于测试集的验证部分采用1 000轮训练次数（早停条

件为100）的40PC-XGBoost模型，予以进一步的泛化能力验证。

2.4 PCA-XGBoost模型在测试集中的验证

从测试数据集中提取307 866个SNP位点数据，其中来源为AADR的测试样本可提取到111 025个SNP位点；对于来源为C.C. Wang Lab的中国南北方样本：贵州汉族人群（T_HGZ）可提取到的位点数为306 500个，其余人群可提取到42 269或67 366个SNP位点。

700个测试样本在洲际层次的测试结果见表6，可以发现除了南亚之外的洲际人群的预测准确率均为100%。模型对于东亚人群和非洲人群的预测效能最佳，一致率均为100%；欧洲北部芬兰的萨米

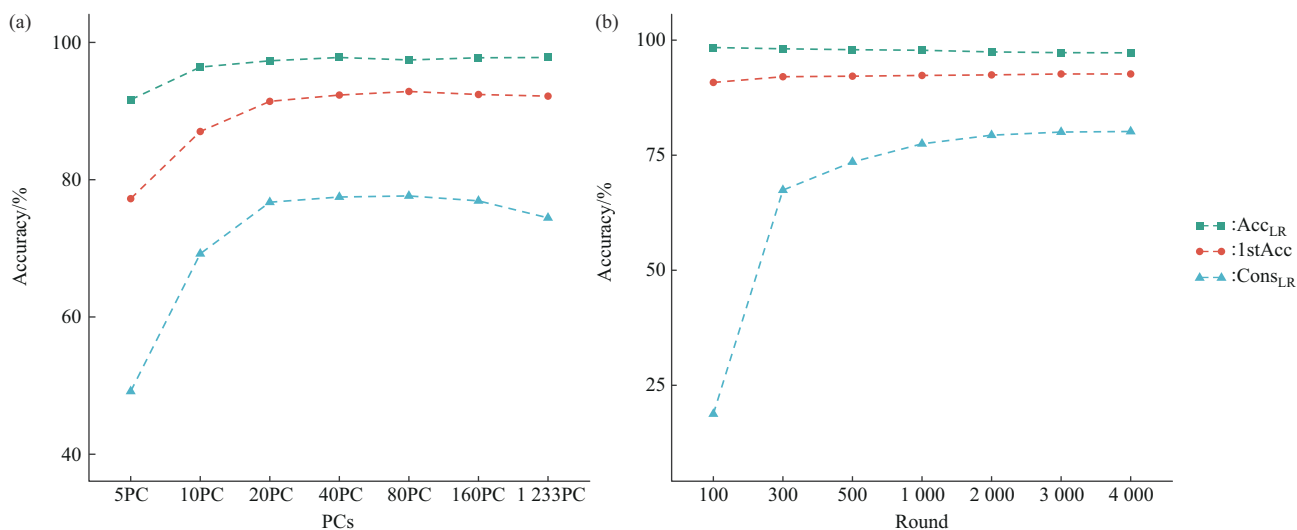


Fig. 5 The accuracy of the model based on different numbers of PCs (1 000 training rounds) and the 40PC-XGBoost model with different training rounds at the population level

Table 6 Inferential results of 40PC-XGBoost model based on test dataset at the continental level

TRUE\PRED	EUR	EAS	AFR	AMR	SAS	Total	1stAcc/%	Acc _{LR} /%		Err _{LR} /%
								Cons _{LR} /%	Incc _{LR} /%	
EUR	84 (3)			1 (1)		85	98.82	95.29	4.71	0
EAS		204 (0)				204	100	100	0	0
AFR			52 (0)			52	100	100	0	0
AMR				61 (5)		61	100	91.80	8.20	0
SAS		10 (0)	25 (23)	11 (11)	251 (56)	298	84.23	65.44	23.83	10.74
Total						700	93.14	84.00	11.43	4.57

The number in each grid represents the number of samples predicted for the corresponding continent, and the number in parentheses represents the number of samples in which the result belongs to inconclusive conclusion (IC).

人群 (T_{SMF} 和 T_{SUF}, 共 3 个样本) 的预测效果不佳, 难以排除祖源位于东亚等其他大洲的可能性, 而芬兰的芬兰人的全部 3 个样本均可得到一致性结论, 这可能是由于萨米人作为欧洲独特的土著人群与欧洲参考人群, 包括芬兰人人群存在遗传差异^[28-30]; 模型对美洲墨西哥的皮马人 (T_{PMM}) 中的 1/3 的测试样本 (5 个) 倾向于不排除其来自非洲祖源; 在南亚测试样本中, 印度东北部的里昂人 (T_{RII}) 全部 10 个的祖源均被预测到东亚, 这可能与以往文献中揭露的其蒙古人种部落背景相关^[31]。巴基斯坦的测试人群中, 哈扎拉人 (T_{HZP}) 的预测效果最差, 超过 90% 的样本不排除东亚祖源, 约 46% 的样本不排除欧洲祖源, 这可能与学者认为的高加索人种和蒙古人种的混合背

景相关^[32-33]。巴基斯坦南部沿海地区的马拉克尼人群 (T_{MKP}) 的美洲的不排除率较高, 可能与其多民族的混合背景相关。

108 个 A 类测试样本的测试结果见表 7, 可以发现模型在超半数 (8 个) 的测试人群中预测一致率可达 100%, 没有样本被错误预测到其他大洲。但是, 尼日利亚的约鲁巴人 (T_{YRN}) 的一致率和准确率分别为 35% 和 73.46%, 显著低于基于参考集的尼日利亚伊巴丹的约鲁巴人 (YRI) 的测试结果 (82.41% 和 95.37%), T_{YRN} 中有 11 个样本 (约占 42%) 被预测为其邻近国家的尼日利亚艾森人 (ESN)。

对于 592 个与参考人群祖源所在国家或地区相同的 B 类测试样本, 着重用于测试模型的泛化能

Table 7 Inferential results of 40PC–XGBoost model based on test dataset at the population level

TRUE\PRED		Target	Accurate	Other POP	POP in other	Total	1stAcc/%	Acc _{LR} /%		Err _{LR} /%
		POP	POP	within continent	continents			Cons _{LR} /%	Incc _{LR} /%	
EUR	<i>T_FIN</i>	FIN	3 (0)	0	0	3	100	100	0	0
	<i>T_EKU</i>	GBR	2 (0)	0	0	2	100	100	0	0
	<i>T_TSI</i>	TSI	2 (0)	0	0	2	100	100	0	0
EAS	<i>T_DCN</i>	CDX	14 (0)	0	0	14	100	100	0	0
	<i>T_HSX</i>	CHB	8 (0)	0	0	8	100	100	0	0
	<i>T_HGD</i>	CHS	7 (3)	0	0	7	100	57.14	42.86	0
	<i>T_JPJ+T_JPT</i>	JPT	30 (0)	0	0	30	100	100	0	0
	<i>T_KHV</i>	KHV	0	2 (1)	0	2	0	0	50	50
	<i>T_ESN</i>	ESN	2 (0)	0	0	2	100	100	0	0
AFR	<i>T_GWD</i>	GWD	2 (0)	0	0	2	100	100	0	0
	<i>T_LWK</i>	LWK	1 (1)	0	0	2	100	0	0	0
	<i>T_MSL</i>	MSL	1 (1)	1 (1)	0	2	50	0	100	0
	<i>T_YRN</i>	YRI	15 (6)	11 (4)	0	26	57.69	34.62	38.46	26.92
SAS	<i>T_BDB</i>	BEB	2 (1)	0	0	2	100	50	50	0
	<i>T_PJL</i>	PJL	3 (3)	0	0	4	75	0	100	0
Total						108	86.11	73.15	19.44	7.41

The number in each grid represents the number of samples predicted for the corresponding category, and the number in parentheses represents the number of samples in which the result belongs to Inconclusive Conclusion (IC).

力。按照测试人群的国家或地区统计其测试结果的第一位预测结果（1stPred），并分别按比例绘制在图6中。百分比堆积柱状图中每根柱描述了国家或地区级别的正确预测比例（浅色）、洲际级别的正确预测比例（深色）和其他大洲的错误预测比例（白色）。可以发现约有75%的测试样本可以预测到与参考人群相同的国家或地区，除了南亚复杂人群背景导致的错误预测之外，绝大多数的测试样本可以被预测到正确的大洲。南亚的印度和巴基斯坦人群约有16%和7%预测到其他大洲的情况，预测

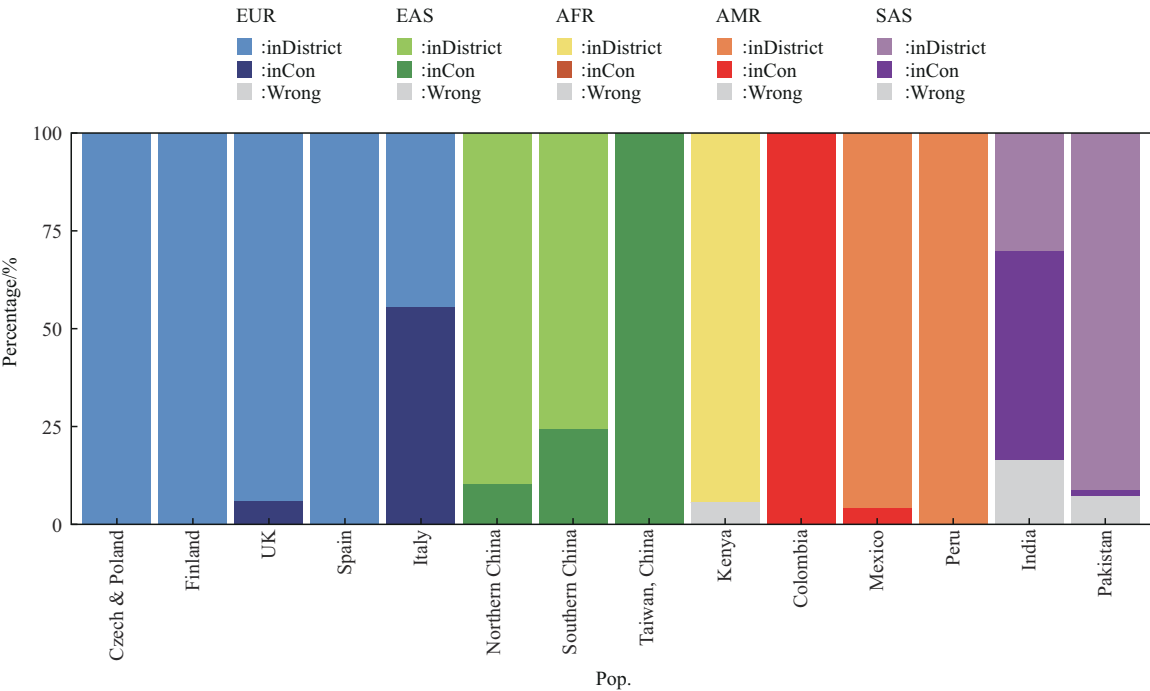


Fig. 6 Result of 1stAcc of test populations by 40PC–XGBoost model

Bars depicts correct results at country or regional level (light color), intercontinental level (dark) and incorrect results of other continents (white).

错误样本主要来自印度东北部的里昂人 (T_RII) 和巴基斯坦的哈扎拉人 (T_HZP)。

3 讨 论

3.1 位点选取与参考数据集

本研究的参考集是根据 1KPhase3 数据集和所筛选的 307 866 个 SNP 位点得到, PC1-PC2 图和 ADMIXTURE 的混合比例图表明该组位点分析的洲际人群遗传结构与已有研究结果基本一致^[18], 证明了所选位点的可用性和有效性。

本研究构建的 PCA-XGBoost 模型基于高密度的 SNP, 可以实现洲际和洲内亚人群的区分。但该模型的应用并不局限于本研究所质控的位点和选取的人群, 如果在一定研究范围内 (如在某个大洲内) 选定参考人群和合适的高密度 SNP 位点, 可使用相同思路完成模型构建。参考数据集覆盖的人群范围和细化程度决定了模型预测的能力范围, 在实际应用中, 参考人群的数量和比例分配应该尽量均匀, 防止样本不均衡引起的模型预测准确率的偏倚。

3.2 人群遗传结构与地理空间的关联性

PCA 作为一种无监督学习方法, 它通过数据在特征空间上的差异性来识别数据内在模式, 直观分析人群的遗传聚类特征。对 PC 维度的分析可以发现: 某些维度与现实地理空间的具有很好的对应性 (如经纬度和空间相对距离), 例如 Novembre 等^[11] 的文章中, 欧洲范围内 PC 空间中的 PC1 和 PC2 轴经过旋转一定角度后, 可较好地与欧洲范围内地图上的经度纬度相对应, 这与本研究的 PC1-PC2 空间的洲际人群相对关系大体一致。但是有些 PC 并不一定能找到现实中与之对应的维度, 某些 PC 可能蕴含了其他方面的信息, 例如: 空间地理距离远近、自然地形阻隔特征、人文交流和历史迁移因素共同导致的遗传距离。

但总的来说, 遗传距离与地理距离有很高的相关性 (主要体现在前若干维 PC 空间中), 这与 Elhaik 等^[2] 揭示的“遗传距离与地理距离在一定范围内呈线性关系”相一致。即 PC 空间上距离的远近一定程度上可以反应遗传距离的远近, 它往往与空间距离正相关, 当然, 也存在遗传距离与地理距离不对应的情况, 这与地理学两大定律 (空间相关性定律^[34] 和空间异质性定律^[35]) 反映出的规律有着内在一致性。

3.3 模型构建时的参数选择

PCA 可以识别数据的内在模式, 通过较少的特征保留较多的原始信息; 同时结合高效快速的 XGBoost 模型进行监督学习, 可达到人群分类和预测的目的。如果直接利用筛选后的 307 866 个 SNP 构建生物地理祖源推断模型, 将会耗费过多的计算资源和时间。而采用 PCA 的降维和特征提取的方法可以消除冗余信息, 有助于降低计算成本。

不同 PC 对于不同人群的区分能力存在差异, 使用少数几个 PC 不足以实现亚人群的有效识别, 若要提高人群的区分度需选取更多维度 PC。所以本研究模型初步构建时采用了前 10 个 PC, 相较于以往文献^[11] 仅使用前两维 PC 构建的线性回归模型, 区分度有所提升。本研究通过探索将 PC 数目增加到 40, 洲内人群层面的区分能力得到明显提升。

大洲内部分人群之间有着相互预测混淆的情况 (如 CEU 和 GBR、GIH 和 PJL、ITU 和 STU), 在 ADMIXTURE 当 K 为 7 和 8 等值分析结果中, 可以发现这些人群对具有相似的祖先成分, 由此导致了这些人群偏低的 1stAcc 结果。

10PC-XGBoost 模型在人群层面的第一位预测准确率只有 87.02%, 而 LR 的引入提高了模型上限, 基于 LR 的准确率指标 Acc_{LR} 由于纳入了后续位的预测结果而提高, 其中 LR 阈值就是控制标准。需要注意的是, 引入 LR 会难以避免地提高不排除率, 本研究选定的 LR 阈值为 10, 这是综合了较高一致率和较低不排除率确定的。

对比 10PC-XGBoost 和 40PC-XGBoost 的洲际推断结果, 可知少量 PC 已足够洲际人群区分, 增加 PC 反而增加时间成本。人群层次的优化过程表明增加 PC 数目对于预测效果有显著提升, 但是增加到一定数量后准确率反而下降, 例如: 图 5 中 PC 增加到 80 后准确率出现下降趋势, 原因可能是越靠后的 PC 可能包含了更多的数据噪音^[36]。不同训练轮数的结果对比表明, 增加训练轮数主要影响一致率, 可以增加模型对于第一位预测结果的确信度, 在实际应用中可根据待测样本的数量和时限要求调整训练轮数。

3.4 模型的泛化能力

模型验证的结果表明, 优化后的 40PC-XGBoost 模型有着良好的准确率和泛化能力, 与已有的参考人群国家、地区或民族相同的样本大概率 (约 75%) 会推断到相近的群体。对于 B 类测试人

群来说, 由于参考数据集的覆盖范围不够广、粒度不够细, 其测试结果的不确定性会比较大, 有时测试样本本身在遗传空间中处于某几个参考人群之间, 容易产生若干个不排除祖源的结果。

3.5 模型对比与不足

通过分析和对比其他生物地理祖源推断方法, 可以发现本研究模型在预测准确性方面有着较明显的提升。Yang 等^[13] 基于约 50 万个 SNP 位点, 通过显式地假设等位基因频率的分布函数, 构建了空间祖先分析 (spatial ancestry analysis, SPA) 方法, 在欧洲的 37 个人群中预测准确率为 (45±5)%; Novembre 等^[11] 同样基于约 50 万个 SNP 位点, 使用了一种联合 PC1 与 PC2 的线性回归模型, 在欧洲人群中的平均准确率为 (40±5)%。Elhaik 等^[2] 基于约 13 万个 SNP 位点, 通过 STRUCTURE 分析得到 9 个混合成分, 建立了地理人口结构分析 (geographic population structure, GPS) 方法, 在 34 个国家人群中的预测准确率达到 83%。Araghi 等^[37] 基于约 14 万个 SNP 位点, 基于 PC 特征建立了 LASSO 线性回归模型, 对 1000Genomes 数据集的人群预测准确率约为 92.1%。与上述预测方法对比, 本研究构建的模型在大洲内人群层次上可达 95% 以上, 在模型的推断准确性上有一定程度提升。

由于参考数据集选自公共数据集, 在洲际和亚人群的覆盖度和均匀度等方面存在一定的不足, 未来可进一步增加参考人群, 提高人群覆盖度和亚人群的数量, 从而进一步提升模型的泛化能力。本研究模型在较少类别的洲际人群区分上表现很好, 而在较多类别的亚人群水平上表现不足, 或可尝试借助模块划分思想, 先将测试样本定位到某个大洲, 再在洲内进行亚人群的细化推断, 从而将多分类的问题拆分成层次化的分类问题。此外, 模型对于遗传混合样本的推断能力相对不足, 后续可尝试加入其他生物学或群体遗传学特征作为分类模型的输入特征, 例如 ADMIXTURE 分析中的祖先成分特征等, 从而提升遗传混合人群的区分能力。

4 结 论

本研究构建了一个基于千人基因组参考数据集的生物地理祖源推断 PCA-XGBoost 模型, 并探索了不同 PC 维度、训练轮数等参数的优化条件, 对于群体区分和推断问题具有显著效用。对比以往的族群推断模型, 本研究优化后的模型有较高的洲际

人群预测准确性, 基于 LR 的方法有助于对预测结果的可靠性进行进一步评估。本研究模型具有高密度 SNP 的适用性, 不局限于具体区域内的人群, 具有很好的泛化能力, 在使用不同参考集的人群数据时有望实现更加精细的人群分析和推断。

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

PIBB_20230453_Figure_S1.pdf

PIBB_20230453_Figure_S2.pdf

PIBB_20230453_Table_S1.xlsx

PIBB_20230453_Table_S2.xlsx

PIBB_20230453_Table_S3.xlsx

参 考 文 献

- [1] 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
- [2] Elhaik E, Tatarinova T, Chebotarev D, *et al.* Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nat Commun*, 2014, **5**: 3513
- [3] Halder I, Kip K E, Mulukutla S R, *et al.* Biogeographic ancestry, self-identified race, and admixture-phenotype associations in the HeartSCORE Study. *Am J Epidemiol*, 2012, **176**(2): 146-155
- [4] Nelson M R, Bryc K, King K S, *et al.* The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet*, 2008, **83**(3): 347-358
- [5] Phillips C. Forensic genetic analysis of bio-geographical ancestry. *Forensic Sci Int Genet*, 2015, **18**: 49-65
- [6] Phillips C. Ancestry informative markers//Siegel J A, Saukko P J, Houck M M. *Encyclopedia of Forensic Sciences*. 2nd Ed. Waltham: Academic Press. 2013: 323-331
- [7] Kidd K K, Speed W C, Pakstis A J, *et al.* Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int Genet*, 2014, **10**: 23-32
- [8] Li C X, Pakstis A J, Jiang L, *et al.* A panel of 74 AISNPs: improved ancestry inference within Eastern Asia. *Forensic Sci Int Genet*, 2016, **23**: 101-110
- [9] Wei Y L, Wei L, Zhao L, *et al.* A single-tube 27-plex SNP assay for estimating individual ancestry and admixture from three continents. *Int J Legal Med*, 2016, **130**(1): 27-37
- [10] Wang Y, Lu D, Chung Y J, *et al.* Genetic structure, divergence and admixture of Han Chinese, Japanese and Korean populations. *Hereditas*, 2018, **155**: 19
- [11] Novembre J, Johnson T, Bryc K, *et al.* Genes mirror geography within Europe. *Nature*, 2008, **456**(7218): 98-101
- [12] Guillot G, Jónsson H, Hingee A, *et al.* Accurate continuous

- geographic assignment from low- to high-density SNP data. *Bioinformatics*, 2016, **32**(7): 1106-1108
- [13] Yang W Y, Novembre J, Eskin E, *et al.* A model-based approach for analysis of spatial structure in genetic data. *Nat Genet*, 2012, **44**: 725-731
- [14] Beaumont M A, Zhang W, Balding D J. Approximate Bayesian computation in population genetics. *Genetics*, 2002, **162**(4): 2025-2035
- [15] Battey C J, Ralph P L, Kern A D. Predicting geographic location from genetic variation with deep neural networks. *Elife*, 2020, **9**: e54507
- [16] Schrider D R, Kern A D. Supervised machine learning for population genetics: a new paradigm. *Trends Genet*, 2018, **34**(4): 301-312
- [17] Chen T, Guestrin C. XGBoost: a scalable tree boosting system// Association for Computing Machinery. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, United States: Association for Computing Machinery, 2016: 1603.02754
- [18] Genomes Project Consortium 1 0 0 0, Auton A, Brooks L D, *et al.* A global reference for human genetic variation. *Nature*, 2015, **526**(7571): 68-74
- [19] Mallick S, Micco A, Mah M, *et al.* The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. *Sci Data*, 2024, **11**(1): 182
- [20] 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
- [21] Wang Q, Zhao J, Ren Z, *et al.* Male-dominated migration and massive assimilation of indigenous east asians in the formation of muslim Hui people in southwest China. *Front Genet*, 2020, **11**: 618614
- [22] Yang M, He G, Ren Z, *et al.* Genomic insights into the unique demographic history and genetic structure of five Hmong-mien-speaking Miao and Yao populations in southwest China. *Front Ecol Evol*, 2022, **10**: 849195
- [23] Purcell S, Neale B, Todd-Brown K, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 2007, **81**(3): 559-575
- [24] 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
- [25] Alexander D H, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 2009, **19**(9): 1655-1664
- [26] Goebel T, Waters M R, O'Rourke D H. The late Pleistocene dispersal of modern humans in the Americas. *Science*, 2008, **319**(5869): 1497-1502
- [27] Adhikari K, Chacón-Duque J C, Mendoza-Revilla J, *et al.* The genetic diversity of the americas. *Annu Rev Genomics Hum Genet*, 2017, **18**: 277-296
- [28] Lahermo P, Sajantila A, Sistonon P, *et al.* The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. *Am J Hum Genet*, 1996, **58**(6): 1309-1322
- [29] Meiniälä M, Finnilä S, Majamaa K. Evidence for mtDNA admixture between the Finns and the Saami. *Hum Hered*, 2001, **52**(3): 160-170
- [30] Ikegaya H, Zheng H Y, Saukko P J, *et al.* Genetic diversity of JC virus in the Saami and the Finns: implications for their population history. *Am J Phys Anthropol*, 2005, **128**(1): 185-193
- [31] Kumar N, Sastry D B. A genetic survey among the Riang: a Mongoloid tribe of Tripura (North East India). *Zeitschrift für Morphologie und Anthropologie*, 1961, **51**(3): 346-355
- [32] Chen P, Adnan A, Rakha A, *et al.* Population background exploration and genetic distribution analysis of Pakistan Hazara via 23 autosomal STRs. *Ann Hum Biol*, 2019, **46**(6): 514-518
- [33] Perveen R, Ali Shahid A, Ahmad J. Forensic and phylogenetic characterization of 15 autosomal STRs in Hazara population of Pakistan. *Leg Med*, 2020, **47**: 101786
- [34] Tobler W R. A computer movie simulating urban growth in the Detroit Region. *Econ Geogr*, 1970, **46**: 234
- [35] Goodchild M F. The fundamental laws of GIScience//University Consortium for Geographic Information Science, University of California. Proceedings of the Invited Talk at University Consortium for Geographic Information Science, University of California. Pacific Grove, CA, USA: Summer Assembly of the University Consortium for Geographic Information Science, 2003: 127286192
- [36] Zhou Z H. Dimensionality reduction and metric learning//Zhou Z H. Machine Learning. Singapore: Springer Singapore. 2021: 241-264
- [37] Araghi S, Nguyen T. A hybrid supervised approach to human population identification using genomics data. *IEEE/ACM Trans Comput Biol Bioinform*, 2021, **18**(2): 443-454

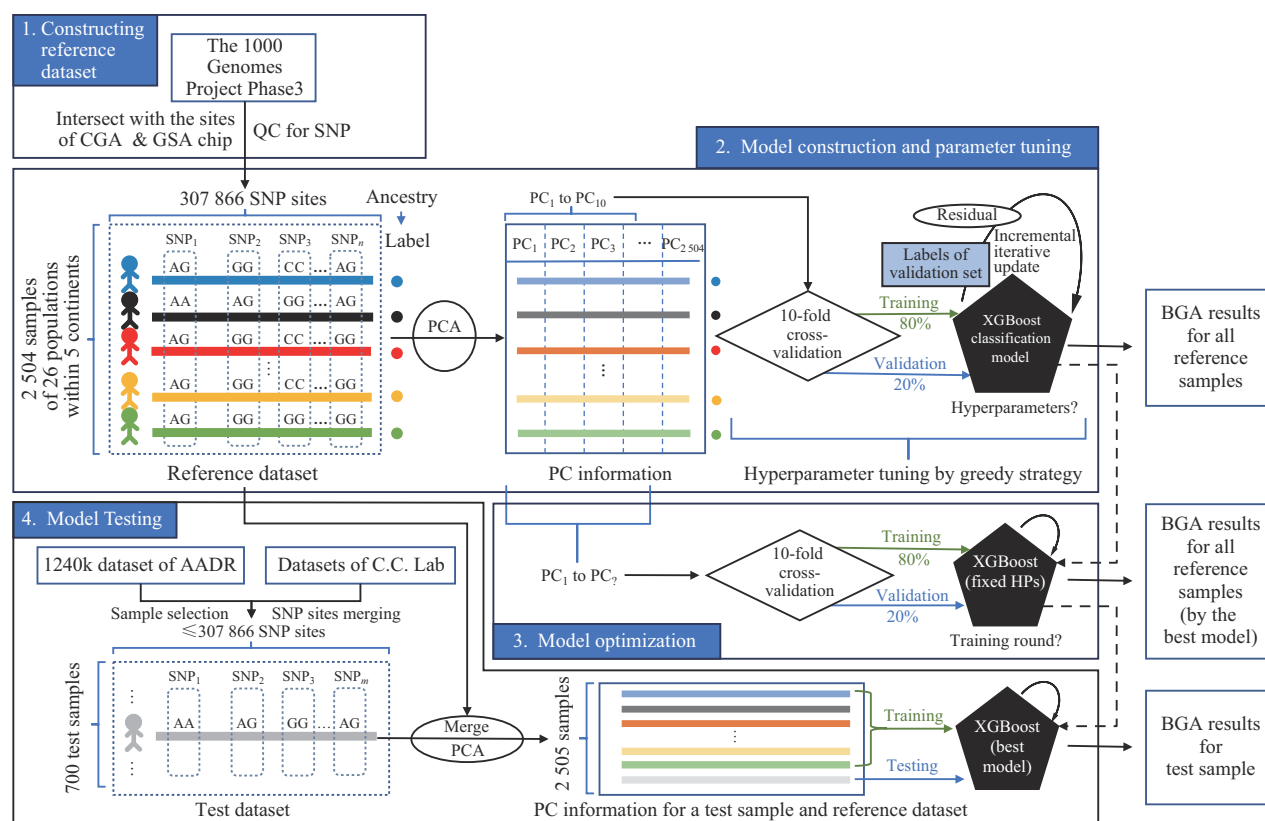
Research on The Intercontinental Population Biogeographic Ancestral Inference Model Based on PCA–XGBoost Method*

YAO Hao-Tian^{1,2)**}, JIANG Li^{2)**}, WANG Chun-Nian^{1,2)}, FAN Hong^{1)***}, LI Cai-Xia^{2)***}

⁽¹⁾School of Computer Science, Shaanxi Normal University, Xi'an 710119, 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)

Graphical abstract



* This work was supported by grants from National Key R&D Program of China (2022YFC3341004), The National Natural Science Foundation of China (82171870), Key Project of Natural Science Foundation of Shaanxi Province (2022ZJ-39), The Key Laboratory of Forensic Genetics Open Project (2023FGKFKT01), and The Fundamental Research Funds for Institute of Forensic Science (2022JB020).

** These authors contributed equally to this work.

*** Corresponding author.

FAN Hong. Tel: 86-15929807273, E-mail: fanhong@snnu.edu.cn

LI Cai-Xia. Tel: 86-10-83752706, E-mail: licaixia@tsinghua.org.cn

Received: November 16, 2023 Accepted: June 5, 2024

Abstract Objective The inference of biogeographical ancestry (BGA) using DNA is a significant focus within anthropology and forensic science. Current methods often utilize dozens of ancestry-informative SNPs, employing principal component analysis (PCA) and likelihood ratios (LR) to ascertain individual ancestries. Nonetheless, the selection of these SNPs tends to be population-specific and shows limitations in population differentiation. With the development of high-throughput sequencing technologies, acquiring high-density SNP datasets has become easier, challenging traditional statistical models which are often reliant on prior assumptions and struggle with high-density genetic data. The integration of machine learning, which prioritizes data learning and algorithmic iteration over prior knowledge, has propelled forward new developments in BGA research. This study aims to construct a BGA inference model suitable for high-density SNP data, characterized by broad population applicability, higher accuracy, and strong generalization capabilities. **Methods** Initially, intersection sites of autosomes from the phase III data of the 1000 Genomes Project and commonly used commercial chips were selected to build a reference dataset after thorough site quality control and filtering. This dataset was analyzed using PCA and ADMIXTURE to study population clustering, ancestral component mixing, and genetic substructures. Utilizing spaces of different principal component (PC), combinations, this study visually assessed the PCs' capabilities to differentiate between continental and intercontinental populations. Following this, the study employed the supervised learning classification model XGBoost, establishing a multidimensional PC-based PCA-XGBoost model with hyperparameters set through ten-fold cross-validation and a greedy strategy. Subsequently, the model was optimized and evaluated based on the LR, considering accuracy and runtime to determine the optimal number of PCs and training rounds, culminating in the study's optimal BGA inference model. Finally, the performance of the model was subsequently validated at national and regional levels using test sets from other public data to assess its post-optimization generalization capabilities. **Results** The reference dataset created contains 307 866 SNP sites. Top PCs reflect varying levels of population differentiation capabilities, with some PCs showing population specificity. Under smaller K values in ADMIXTURE results, genetic ancestral components between continents are elucidated, while larger K values reveal some specific ancestral components of certain populations within continents. The number of PCs and training rounds significantly affect the classification accuracy and efficiency of the XGBoost supervised model. With LR-based evaluation methods, the optimized PCA-XGBoost model achieved a continental prediction accuracy of over 98% in the reference set. For subcontinental population levels within the continents, the model achieved an accuracy of over 95% in the reference set and over 90% in the test set. **Conclusion** The reference dataset effectively represents the genetic substructures of populations at selected sites. Information derived from PC dimensions significantly aids in population differentiation and inference issues, and incorporating more PC dimensions as features in supervised learning models can increase the accuracy of BGA inference. The model of this study is suitable for high-density SNP data and is not confined to specific regional populations, offering enhanced population-wide applicability. Compared to previous ancestry inference models, the optimized PCA-XGBoost model demonstrates high intercontinental population predictive accuracy. LR-based evaluation methods further enhance the reliability of predictions. Additionally, the model's strong generalization capabilities suggest that updating the reference population data could enable more detailed population analysis and inference.

Key words biogeographic ancestral inference, supervised learning, principal component analysis, XGBoost

DOI: 10.16476/j.pibb.2023.0453