PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2021,48(8):973~981
www.pibb.ac.cn

东亚三族群SNP高分辨推断模型构建与效能评估*

文 豪[1] 魏以梁[3] 郭晓媛[4] 孙昌春[4] 薛思瑶[4] 刘 京[2] 范 虹[1]** 江 丽[2]**

([1] 陕西师范大学计算机科学学院，西安 710119；[2] 公安部物证鉴定中心，北京 100038；
[3] 江苏师范大学，江苏省系统发育与比较基因组学重点实验室，徐州 221116；[4] 山西医科大学法医学院，太原 030001)

摘要 单核苷酸多态性（single nucleotide polymorphism，SNP）是法医遗传学个体识别和族群推断常用的遗传标记. 本研究集合文献和公共库中祖先信息SNP位点（ancestry informative SNPs，AISNPs），应用softmax回归、支持向量机和随机森林3种算法，研究东亚北方的3个主体人群（中国北方汉族人、日本人和韩国人）的族群推断效果. 我们分析了来自千人基因组计划的103份中国北方汉族人样本、104份日本人样本和亚洲多样性计划的100份韩国人样本的428个AISNP位点分型，采用多元线性回归共线性诊断筛选出67个高信息量的AISNPs位点组合，构建了softmax回归和支持向量机算法的两种族群推断模型，采用随机森林平均降准分析筛选出42个高信息量的AISNPs位点组合，并构建了随机森林算法的族群推断模型，将softmax回归、支持向量机与随机森林3种模型用于北方汉族人、日本人、韩国人的族群推断，五次十折交叉验证（training∶testing=9∶1）测试3种模型的平均准确率分别为95.19%、95.77%、94.53%. 本研究建立的3种族群推断模型均可用于东亚北方三大人群的遗传推断，42 AISNPs组合的位点数目较少，更适于构建法医检测体系，具有较高的实际应用价值.

关键词 法医遗传学，祖源SNP，东亚北方，softmax回归，支持向量机，随机森林
中图分类号 TP312，R89，D919.2 **DOI**：10.16476/j.pibb.2020.0339

不同人群之间存在基因频率分布差异较大的遗传标记，这种遗传标记被称为祖先信息标记（ancestry informative markers，AIMs）[1-2]. 单核苷酸多态性（single nucleotide polymorphism，SNP）在人类基因组中含量丰富，常用于筛选AIMs[3-12]. 用做AIMs的SNP也称祖先信息SNPs（ancestry informative SNPs，AISNPs），受到了广泛的研究关注[13-16].

东亚是世界上人口最多的地区之一，约占亚洲人口的38%，世界人口的22%. 中国北方汉族人、日本人和韩国人是东亚北方主体人群，遗传背景相近，在外貌特征上有许多相似之处，如黄皮肤、黑眼睛、黑头发和鼻子短而扁平等特征. 研究表明，中国北方汉族、日本和韩国人群在全基因组水平上具有遗传差异[7-9].

常用的族群来源推断软件方法主要包括STRUCTURE v2.3.4软件[17]祖先成分分析、主成分分析（principle component analysis，PCA）以及Forensic Intelligence Version 1.0[18]计算群体匹配概率和似然比. 随着统计方法不断发展，新的机器学习算法不断涌现. Softmax回归、支持向量机（support vector machine）及随机森林（random forest）可用于分类、回归等任务，无需对模型进行必要的假设[19]. 基于国内外已发表的针对汉族、日本及韩国人群的遗传结构研究，本文集合了428个中、日、韩AISNP位点[7-8]，利用共线性诊断和随机森林平均降准分析筛选出最优的高分辨率AISNP组合，结合3种不同的机器学习算法建立族

* 陕西省重点研发计划项目（2018SF-251），国家自然科学基金（81772027），国家工程实验室开放课题（2018NELKFKT15），法医遗传学公安部重点实验室开放课题（2020FGKFKT01）和江苏省高校重大项目（17KJA180003）资助.
** 通讯联系人.
范虹. Tel：15929807273，E-mail：fanhong@snnu.edu.cn
江丽. Tel：18519073957，E-mail：jl@mail.bnu.edu.cn
收稿日期：2020-09-22，接受日期：2021-01-07

群推断预测模型.

1　材料与方法

1.1　样本信息

307份基因组分型数据包括：千人基因组计划（The 1 000 Genomes Project）[20] 的103份汉族人（CHB）和104份日本人（JPT），亚洲多样性计划（Asia Diversity Project；Affymetrix Genome-wide human SNP Array 6.0芯片）[21] 的100份韩国人（KOR），人群样本详细信息见表1.

Table 1　The information of populations

Abbr.	Population	Number	Source
CHB	Northern Han	103	The 1 000 Genomes Project
JPT	Japanese	104	The 1 000 Genomes Project
KOR	Korean	100	Asia Diversity Project

1.2　位点来源

a. 陈华团队的142个AISNP位点[7]；b. 徐书华团队的203个AISNP位点[8]；c. 本课题组前期筛选的88个AISNP位点[22-25]. 以上位点合并去重复后得到428个AISNP位点（附表S1）.

1.3　质量控制

1.3.1　样本的质量控制

提取307份数据中的428个AISNP位点数据，去除SNP数据缺失率大于10%的样本.

1.3.2　位点的质量控制

SNP位点的质量控制指标包括SNP缺失率（>10%）、哈-温平衡（Hardy-Weinberg equilibrium，HWE）检验及连锁不平衡（linkage disequilibrium，LD）分析. Haploview v4.1[26-27] 同时对一组SNP位点进行HWE检验和LD分析. HWE使用卡方检验，α值0.05，使用Bonferroni校正[28]，校正后的α'值0.000 12. LD分析中去除r^2大于0.8的被认为具有显著连锁关系的SNP[29].

1.4　基因型填充及编码

因计算模型要求样本基因分型数据完整，故首先对分型缺失数据进行填充. R v4.0.0软件imputeMissings包有3种填充方式：随机森林、中位数及众数. 基于AISNP特性，本研究选择利用众数填充样本的缺失基因分型.

将基因分型字符型数据编码为数值型数据，具体做法是将各SNP位点的所有纯合基因分型按照ACGT的排序先后分别编码为0和2，杂合基因分型编码为1，具体编码示例见表2.

Table 2　The table of coding example

Individual	SNP（before encoding）	SNP（after encoding）
1	CT	1
2	CC	0
3	TT	2

1.5　共线性诊断AISNPs筛选

Softmax回归模型对多重共线性敏感，即当变量之间的相关程度提高时，系数估计的标准误将会急剧增加，高度相关的变量直接进入模型必然会导致严重的多重共线性干扰. 在支持向量机的实际应用中，如果变量间存在较强的共线性，相当于某个变量的权重远远高于其他变量，其噪音的影响也会相应增加，从而降低支持向量机训练模型的精确度[30].

逐步回归是一种线性回归模型解释变量选择方法，基本思想是将变量逐个引入模型，每引入一个解释变量后都要进行F检验，并对已经选入的解释变量逐个进行t检验，当原来引入的解释变量由于后面解释变量的引入变得不再显著时，则将其删除. 以确保每次引入新的变量之前回归方程中只包含显著性变量. 这是一个反复的过程，直到既没有显著的解释变量选入回归方程，也没有不显著的解释变量从回归方程中剔除为止，以保证最后所得到的解释变量集是最优的.

共线性分析采用SPSS v20软件中线性回归的共线性诊断过程拟合实现[31].

1.6　随机森林AISNPs筛选

随机森林除了分类器外的另一常用功能是识别重要的变量，亦可筛选能够稳定区分东亚北方三大人群的高信息量AISNP组合. “平均降准值”（mean decrease accuracy，MDA）表示每个变量对随机森林预测准确性的降低程度，该值越大表示该变量对分类精度的贡献越大. 利用R v4.0.0软件randomForest包建立随机森林算法模型，通过importance函数基于预测模型计算得到各SNP的MDA值排序，对排序后的训练集利用replicate函数和rfcv函数选取不同的SNP组合训练集进行十折交叉验证，用曲线图展示模型误差与用于拟合的SNPs数量之间的关系，根据交叉验证曲线由高到底（MDA值）选择最佳SNP组合.

1.7　预测模型构建

1.7.1　Softmax回归预测模型构建

逻辑回归分析，是研究二分类（可扩展到多分

类）观察结果与一些影响因素之间关系的一种多变量分析方法. Softmax 回归模型是 logistic 回归模型在多分类问题上的推广，当分类数为 2 的时候会恢复为 logistic 分类. 在多分类问题中，类标签可以取两个以上的值.

R v4.0.0 软件 softmaxreg 包中的 trainModel 函数建立了具有多个隐藏层和 1 个 softmax 最终层的前馈神经网络，以共线性诊断筛选出的 AISNP 位点组合为自变量，族群类别为因变量，利用 trainModel 函数对训练集样本进行回归训练构建 softmax 回归预测模型.

1.7.2　支持向量机预测模型构建

支持向量机的基本思想是求解能够正确划分训练数据集并且几何间隔最大的分离超平面，其学习策略便是间隔最大化，最终化为一个凸二次规划问题的求解，它的基本模型是定义在特征空间上的间隔最大的线性分类器.

R v4.0.0 软件 e1071 包中的 svm 函数用于训练支持向量机，以共线性诊断筛选出的 AISNP 为自变量，族群类别为因变量，利用 svm 函数对训练集样本进行训练，构建线性核函数配置的支持向量机预测模型.

1.7.3　随机森林预测模型构建

随机森林算法是通过对对象和变量进行抽样构建预测模型，即生成多个决策树，并依次对对象进行分类，最后将各决策树的分类结果汇总，所有预测类别中的众数类别即为随机森林所预测的该对象的类别.

R v4.0.0 软件 randomForest 包实现 Breiman 的随机森林算法（基于 Breiman 和 Cutler 的原始 Fortran 代码）的分类或回归，以随机森林算法 MDA 筛选的 AISNP 位点组合为自变量，族群类别为因变量，利用 randomForest 函数对训练集样本进行训练，构建随机森林预测模型.

1.8　预测模型效能评估

本研究采用了两种模型效能评估方法.

其一是利用 R v4.0.0 软件 caret 包中的 createDataPartition 函数从每个标记人群中随机采样 80% 个体作为训练集建立预测模型，其余 20% 作为测试集，利用 pROC 包的 predict 函数预测分类，利用 caret 包中的 confusionMatrix 函数对预测结果创建混淆矩阵并对预测模型进行评价. Kappa 系数是一种衡量分类精度的指标，既可以用于一致性检验，也可以用于衡量分类精度，Kappa 系数的计算基于混淆矩阵，结果为［−1，1］，但通常 Kappa 是落在［0，1］之间，可分为 5 组来表示不同级别的一致性：0~0.20 为极低的一致性、0.21~0.40 为一般的一致性、0.41~0.60 为中等的一致性、0.61~0.80 为高度的一致性和 0.81~1 为几乎完全一致. 此外，模型的评价指标还采用准确率的 95% 置信区间（confidence interval）、灵敏度、特异性、阳性预测值及阴性预测值.

其二是采用十折交叉验证方法，将所有样本分成 10 份，轮流将其中 9 份作为训练样本，1 份作为测试样本，进行试验. 每次试验都会得出相应的正确率，多次结果的正确率平均值作为对算法精度的估计. 利用 R v4.0.0 软件 caret 包中的 createFolds 函数对数据集进行划分，所有预测模型均使用 5 次十折交叉验证进行测试，将统计得到的 5 次十折交叉验证准确率均值作为预测模型的效能评估指标.

共线性诊断和 MDA 值筛选出的位点组合，利用 Fst 值、In 值、STRUCTURE 与 PCA 计算分析评估其推断效果.

2　结　　果

2.1　AISNP筛选

针对 307 份样本 428 个 AISNP 的基因分型数据，删除同时满足如下指标的位点：位点检出率< 90%（去除 25 个 SNP）、HWE 平衡 P 值<0.000 12（去除 12 个 SNP）、LD 分析 r^2>0.8（去除 30 个 SNP），最终得到 361 个 AISNP 用于多元线性回归的共线性诊断与非线性模型随机森林的 AISNP 挑选. 首先进行 361 个 SNP 所有样本的基因分型数据基因型众数填充及编码，然后进行共线性诊断，多个 SNP（维度）特征值约为 0 证明存在多重共线性，最终筛选出 67 个 AISNP 位点（表 3）.

Table 3　67 AISNPs by collinearity diagnosis

rs#	Chromosome	Position	Alleles	Fst	In	Eigenvalue
rs17129041	1	66996037	T/C	0.052 9	0.008 8	0.424 9
rs11164354	1	102439329	G/A	0.043 6	0.006 9	0.214 9
rs12134013	1	164459913	C/T	0.058 6	0.010 2	0.314 0
rs12039715	1	242801261	G/C	0.172 6	0.032 6	1.016 9
rs4668680	2	10529961	A/G	0.019 9	0.003 5	0.154 2
rs2587694	2	120213497	A/G	0.002 6	0.001 1	0.099 6
rs1465759	2	142576915	T/C	0.036 5	0.006 3	0.590 8
rs1453054	2	181359907	T/C/G	0.011 9	0.002 3	0.324 3
rs4677399	3	74519384	T/C	0.030 7	0.005 0	0.178 8
rs9825713	3	100936248	C/A	0.045 0	0.007 2	0.282 9

Continued to Table 3

rs#	Chromosome	Position	Alleles	Fst	In	Eigenvalue
rs4678169	3	124543103	C/A	0.055 7	0.009 0	0.832 3
rs6599390	4	956047	A/G	0.058 4	0.009 2	0.432 0
rs4144800	4	38006916	G/A	0.022 8	0.003 8	0.137 0
rs17583068	4	38908616	G/C/T	0.055 8	0.009 1	0.108 9
rs4865142	4	57549583	C/T	0.037 2	0.006 2	0.376 8
rs7698051	4	60278484	A/G	0.044 5	0.007 3	0.044 0
rs17372695	4	155060317	A/G	0.003 0	0.000 4	0.092 0
rs11133144	4	177229719	A/C/T	0.044 6	0.007 2	0.231 0
rs4133446	5	86155113	T/C	0.064 5	0.010 1	0.528 3
rs17207681	5	138168527	A/G	0.082 3	0.013 2	0.244 0
rs1609763	5	142018424	T/G	0.023 2	0.004 1	0.265 2
rs11959012	5	167668843	C/T	0.069 4	0.010 4	0.942 2
rs1178148	7	18778113	A/C	0.079 5	0.012 2	0.464 9
rs258728	7	81663275	C/A	0.033 7	0.005 7	0.925 5
rs7802058	7	81697666	A/C	0.051 2	0.009 0	0.517 8
rs6465469	7	95179593	G/A	0.061 1	0.009 7	0.775 9
rs7006443	8	9290753	T/C	0.038 3	0.008 5	0.087 3
rs10088365	8	10097398	G/A	0.270 9	0.043 2	1.255 8
rs12676684	8	118857704	G/C/T	0.058 4	0.009 1	0.342 0
rs2976396	8	143764001	G/A	0.152 1	0.024 6	1.090 0
rs12351269	9	16806521	T/C	0.055 8	0.009 1	0.207 0
rs7038964	9	104423907	C/T	0.041 9	0.007 1	0.440 2
rs12768145	10	17610277	G/A	0.053 2	0.008 4	0.279 8
rs827287	10	72708276	A/T	0.043 5	0.007 4	0.456 8
rs7097617	10	77504287	A/G	0.044 7	0.007 0	0.065 7
rs10883736	10	104320029	G/T	0.036 5	0.005 8	0.399 0
rs1794072	11	61303803	C/T	0.044 6	0.007 3	0.227 5
rs6589072	11	109322914	C/G	0.050 1	0.008 3	0.072 8
rs4578397	11	131832284	G/T	0.073 7	0.012 4	0.075 6
rs10849181	12	638166	C/T	0.052 5	0.008 7	0.127 8
rs11174261	12	62362843	C/T	0.047 3	0.007 6	0.037 0
rs11177698	12	69906287	A/G	0.053 4	0.008 6	0.121 3

Continued to Table 3

rs#	Chromosome	Position	Alleles	Fst	In	Eigenvalue
rs10506725	12	77449514	T/C	0.044 4	0.007 7	0.047 7
rs2349093	12	128520953	T/A	0.000 4	0.000 7	0.170 5
rs9549212	13	41022093	C/T	0.070 8	0.011 6	0.106 3
rs1322944	13	53683163	A/G	0.074 0	0.011 7	0.677 3
rs7317643	13	108536028	C/T	0.166 5	0.027 9	0.734 7
rs12589835	14	30848932	C/T	0.061 5	0.009 3	0.002 1
rs12586912	14	34023334	G/T	0.042 2	0.007 0	0.398 3
rs11621121	14	65822493	C/T	0.099 0	0.014 6	1.058 7
rs174520	14	88006776	T/G	0.089 9	0.013 6	0.026 7
rs10483991	14	88682616	G/A	0.060 7	0.011 6	0.492 6
rs6576127	14	106194763	C/T	0.107 6	0.016 5	0.165 7
rs7173982	15	36685718	C/A/T	0.001 3	0.000 6	0.150 7
rs7173716	15	70123826	T/G	0.072 3	0.011 2	0.641 6
rs12598852	16	13329469	G/A	0.050 0	0.008 3	0.352 0
rs4280278	16	29439227	C/G	0.084 1	0.013 0	0.568 4
rs1420288	16	54477881	T/C	0.051 9	0.008 9	0.141 5
rs4325622	17	28526475	T/C	0.361 7	0.060 4	2.311 3
rs9303660	17	31420730	C/T	0.061 4	0.009 8	0.686 9
rs9896443	17	47125982	T/C	0.073 8	0.013 1	0.838 5
rs12459941	19	10666112	G/A	0.063 1	0.010 2	0.013 8
rs807959	19	22115835	A/G	0.056 1	0.009 1	0.366 8
rs7268940	20	9961532	C/T	0.046 1	0.008 2	0.295 5
rs6016226	20	38661387	C/T	0.018 1	0.005 3	0.269 3
rs229562	22	37599065	G/T	0.082 7	0.014 3	0.611 8
rs131864	22	47271217	T/G	0.045 0	0.007 2	0.064 7

Bold indicates 13 overlapping locis.

以随机森林算法评估361个位点的MDA值，通过5次十折交叉验证曲线确定最佳SNP位点组合. 使用MDA值最高的42个AISNP位点时，十折交叉验证错误率达到最低值（图1）. 据此筛选出

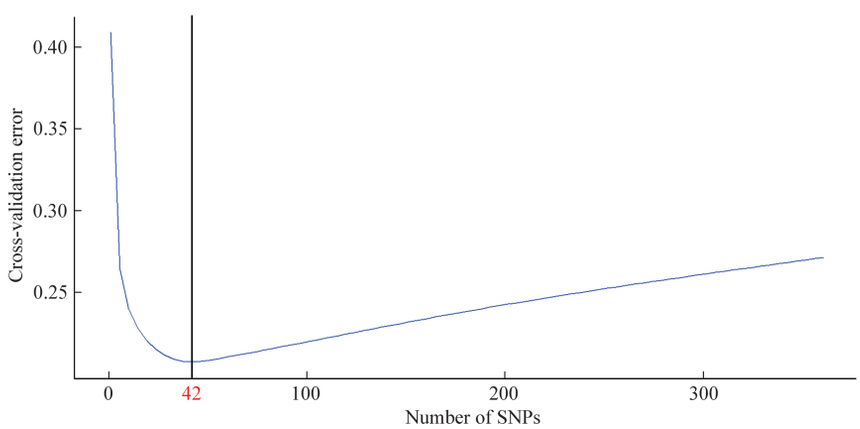

Fig. 1　The cross-validation plot of AISNP number

42个最佳AISNP位点（表4）. 42个位点在三大人群中LD分析r^2值为0~0.356. 上述筛选出的两组AISNPs组合中，共有13个重叠位点（图2）.

Table 4　42 AISNPs by random forest MDA

rs#	Chromosome	Position	Alleles	Fst	In	MDA
rs1325192	1	199244248	G/C/T	0.062 5	0.010 0	3.658 6
rs12039715	1	242801261	G/C	0.172 6	0.032 6	8.775 9
rs10171891	2	195971878	C/T	0.043 1	0.007 3	3.900 8
rs6436971	2	231582797	T/C	0.081 5	0.013 3	4.223 0
rs4353835	3	32446775	C/T	0.084 9	0.014 0	4.338 1
rs9860483	3	138730737	C/T	0.098 5	0.016 3	6.084 1
rs17631488	5	18777746	A/G	0.103 0	0.015 2	4.802 1
rs17207681	5	138168527	A/G	0.082 3	0.013 2	4.180 8
rs17451739	5	144030993	C/T	0.073 2	0.014 0	4.554 7
rs9321180	6	130102959	C/T	0.086 6	0.012 8	3.830 6
rs1178148	7	18778113	A/C	0.079 5	0.012 2	4.177 7
rs1465306	7	66699784	T/C	0.067 2	0.011 5	5.306 3
rs258728	7	81663275	C/A	0.033 7	0.005 7	4.444 0
rs10088365	8	10097398	G/A	0.270 9	0.043 2	12.634 3
rs2945733	8	134615750	T/G	0.060 2	0.009 4	3.715 2
rs2976396	8	143764001	G/A	0.152 1	0.024 6	6.929 8
rs12351269	9	16806521	T/C	0.055 8	0.009 1	3.721 1
rs7022178	9	29780195	G/A	0.072 0	0.012 8	4.098 5
rs12006467	9	35090720	T/C	0.099 3	0.016 3	6.322 9
rs10521076	9	108878562	C/T	0.077 5	0.012 8	4.231 5
rs7032231	9	117025771	C/A	0.070 6	0.011 2	3.675 1
rs17121800	10	108710873	C/T	0.100 7	0.015 0	5.520 9
rs11034709	11	38428289	A/G	0.108 3	0.016 9	5.864 1
rs11224765	11	101310590	C/T	0.098 1	0.015 3	3.816 8
rs7117447	11	101351383	G/A	0.111 8	0.019 2	6.648 9
rs10894034	11	129255618	C/T	0.052 1	0.008 5	3.695 2
rs4578397	11	131832284	G/T	0.073 7	0.012 4	4.724 4
rs1678537	12	57900341	G/A	0.088 5	0.014 5	4.732 1
rs7317643	13	108536028	C/T	0.166 5	0.027 9	10.211 7
rs8014475	14	65811537	T/C	0.093 7	0.013 9	5.141 6
rs10483991	14	88682616	G/A	0.060 7	0.011 6	4.922 3
rs6576127	14	106194763	C/T	0.107 6	0.016 5	4.098 3
rs4325622	17	28526475	T/C	0.361 7	0.060 4	16.487 7
rs9896443	17	47125982	T/C	0.073 8	0.013 1	5.237 8
rs2642066	17	65639014	G/T	0.091 0	0.013 9	3.917 4
rs9941426	18	487281	T/G	0.068 9	0.010 9	5.002 7
rs883433	19	39206288	C/T	0.058 0	0.009 2	3.809 6
rs6123723	20	37082145	C/T	0.089 1	0.015 1	5.957 8
rs6030932	20	42146320	T/G	0.065 9	0.011 1	3.768 8
rs760873	20	45393072	T/G	0.058 2	0.009 5	4.562 4
rs1524930	21	40968460	A/G	0.064 9	0.010 8	3.740 9
rs4820428	22	41537589	A/G	0.076 2	0.012 6	4.067 5

Bold indicates 13 overlapping locis.

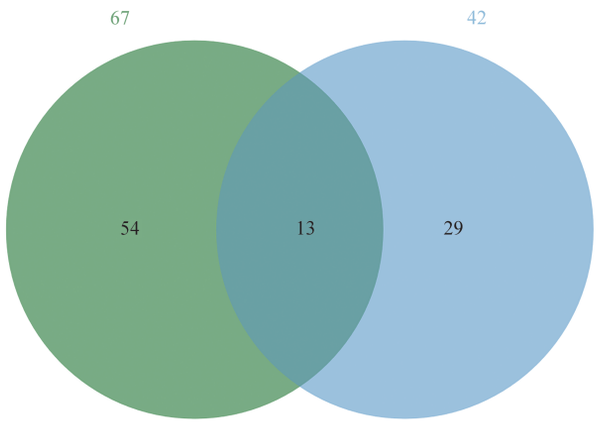

Fig. 2　Venn diagrams of 67 and 42 AISNPs

2.2　预测模型建立与效能评估

从每个人群中随机采样80%个体作为训练集构建softmax回归、支持向量机和随机森林3种算法预测模型，为了获得最佳拟合模型，对3种算法特定的不同参数进行了调试，包括softmax回归的学习率rate值（0.05）、支持向量机算法惩罚参数cost值（1）、随机森林决策树的数量ntree（500）. 其余20%个体作为测试集基于3种算法预测模型预测分类，对预测结果创建混淆矩阵并对预测模型进行评价，评价指标Kappa系数、准确率的95%置信区间、灵敏度、特异性、阳性预测值及阴性预测值见表5. 3种算法预测模型准确率均大于96%，softmax模型准确率和Kappa系数均高于支持向量机模型与随机森林模型.

3种预测模型5次十折交叉验证测试的平均准确率（表5）约为95%，基于更少位点的随机森林模型（42个AISNP位点，94.53%）与67个AISNP位点softmax模型（95.51%）及支持向量机模型（95.77%）准确性水平相同.

将307份样本的67 AISNPs组合和42 AISNPs组合的分型数据全部作为训练集，利用R v4.0.0软件对训练集样本进行训练分别建立softmax回归、支持向量机及随机森林预测模型，利用pROC包的predict函数即可实现对未知族群来源样本类别的预测.

2.3　STRUCTURE与PCA评估

将筛选出的67与42位点组合，分别进行STRUCTURE与PCA分析，进一步评价位点的族群划分效果. STRUCTURE结果K=3时，中国北方汉族人、日本人和韩国人分化生成三大主体祖先成分. PCA结果显示，这两组位点均实现了三个人群

Table 5　Performance assessment of three predictive models

Model	Population	80% training *vs* 20% testing						Average accuracy of 5 times 10-fold cross-validation
		Sensitivity	Specificity	PPV	NPV	Kappa coefficient	Accuracy（95% CI）	
Softmax regression（67AISNPs）	CHB	1	1	1	1	0.975	0.984（0.912，0.999）	95.51%
	JPT	0.947	1	1	0.977			
	KOR	1	0.976	0.952	1			
SVM（67AISNPs）	CHB	0.947	1	1	0.977	0.951	0.967（0.887，0.996）	95.77%
	JPT	0.955	0.974	0.955	0.974			
	KOR	1	0.976	0.952	1			
Random forest（42AISNPs）	CHB	0.958	1	1	0.974	0.950	0.967（0.887，0.996）	94.53%
	JPT	0.941	0.977	0.941	0.977			
	KOR	1	0.976	0.952	1			

的区分（图3）. 通过比较可以看出，共线性诊断筛选出的67个AISNP位点对人群的区分效果略优于随机森林筛选出的42个AISNP位点，但是随机森林筛选出的42个AISNP仍然可以实现三人群划分.

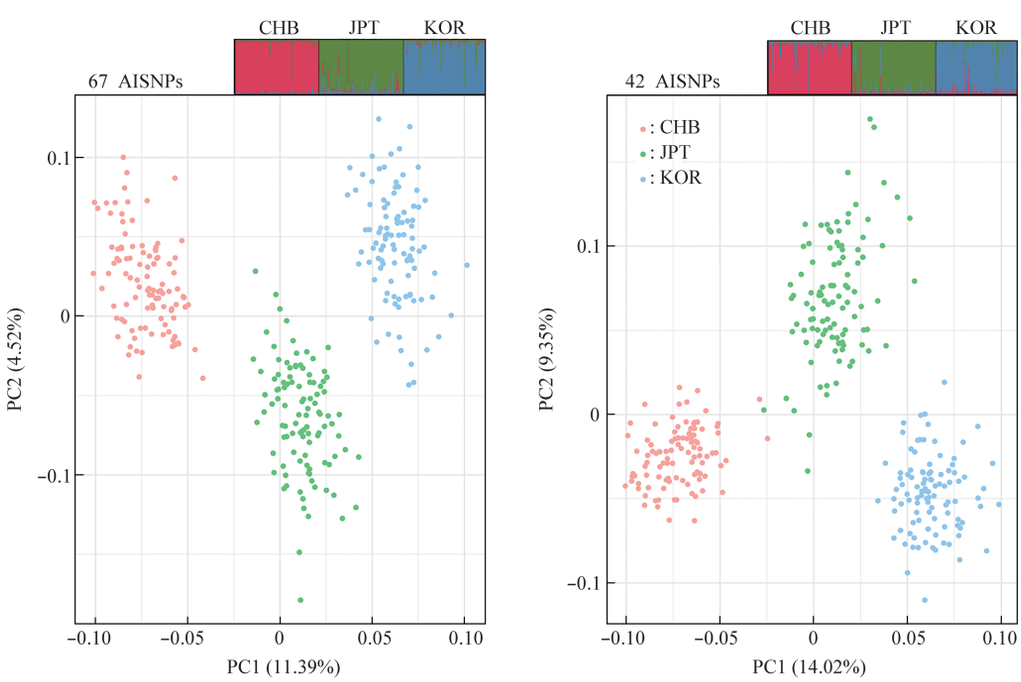

Fig. 3　The Structure result and PCA plots of 67 AISNPs and 42 AISNPs

3　讨　　论

DNA族群推断是近年来法医遗传领域的关注热点之一. 目前，国内外已发表大量洲际人群推断相关研究，但对东亚地区亚人群族群推断的研究报道较少. 徐书华团队从全基因组水平分析发现了汉族、日本和韩国人群之间的遗传差异[8]. 陈华团队建立了分别包含36、59、98及142个AISNP的四组嵌套体系，可实现汉族、日本和韩国人群的区分，总体平均区分准确率从90%到99%不等[7]. 本研究在前期研究报道基础上，筛选获得67个AISNP位点，将这些位点作为自变量，族群类别作为因变量，建立了softmax回归和支持向量机的预测模型. 这两种预测模型均可实现3个人群高精度

判别，平均准确率超过95%. 而利用随机森林算法也可识别重要预测变量. 以MDA值为筛选指标，通过十折交叉验证挑选出42个AISNP的最佳组合，并达到与67个位点同等水平的判别精度. 本研究筛选的位点组合与构建的算法方案，与之前的研究相比，位点数目更少，更易构建复合检测体系，适合法医学应用. 因此，较Softmax回归模型、SVM模型以及对应的共线性诊断方法，可优先考虑随机森林模型和MDA筛点方法.

一方面，本研究在位点筛选方面采用了两种方式. 一种是利用SPSS v20软件中线性回归的共线性诊断过程拟合实现softmax回归的共线性分析和支持向量机的共线性分析，从而确保筛选出的AISNPs具备显著性的同时位点之间又相互独立，为模型提供了最优的解释变量集；另一种是随机森林算法评估候选AISNPs的重要性后，按AISNPs重要性从大到小的顺序对数据集重新整理，对不同SNPs组合的数据集进行十折交叉验证获得交叉验证曲线图，根据交叉验证曲线对SNP进行取舍. 与传统筛点方式相比，本文的两种筛点方式更简单易实现，充分利用了机器学习的优势，更侧重于SNP位点组合的选择，而传统筛点方式更侧重于单个最优SNP的选择，特别对近距离人群划分效果不佳. 另一方面，本研究在算法模型方面选用了机器学习领域常用的softmax回归、支持向量机及随机森林等经典算法，在东亚族群推断上获得了很好的效果.

总体而言，本研究建立的67个AISNP的softmax回归与支持向量机预测模型及42个AISNP的随机森林预测模型均能够实现北方汉族、日本及韩国人群的区分. 由于本研究所使用的各人群数据来源单一（北方汉族和日本人数据来源于千人基因组计划，韩国人数据来源于亚洲多样性计划），且数量有限，因此，后续需扩大数据来源以验证模型的可靠性. 收集其他数据库中北方汉族、日本及韩国人数据或检测实际测试样本来进一步评估和确定哪种体系更适于实际应用，进而为东亚北方未知来源样本的族群推断提供有效手段，为案件侦查提供科学线索.

附件　PIBB20200339Sup1_Tables S1.xlsx见本文网络版（http://www.ibp.ac.cn或http://www.cnki.net）

参　考　文　献

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

[2] Phillips C. Ancestry informative markers//Siegel J A, Saukko P J, Houck M M. Encyclopedia of Forensic Sciences. 2nd. [S. I.]: Elsevier, 2013: 323-331(DOI: 10.1016/B978-0-12-382165-2.00060-X)

[3] 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

[4] 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

[5] 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, 2015, **130**(1): 27-37

[6] 江丽, 孙启凡, 马泉, 等. 27-plex SNP种族推断方法的优化及验证. 遗传, 2017, **39**(2): 166-173
Jiang L, Sun Q F, Ma Q, *et al*. Hereditas (Beijing), 2017, **39**(2): 166-173

[7] 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

[8] 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

[9] Sato T, Nakagome S, Watanabe C, *et al*. Genome-wide SNP analysis reveals population structure and demographic history of the ryukyu islanders in the southern part of the Japanese archipelago. Mol Biol Evol, 2014, **31**(11): 2929-2940

[10] Santos C, Fondevila M, Ballard D, *et al*. Forensic ancestry analysis with two capillary electrophoresis ancestry informative marker (AIM) panels: results of a collaborative EDNAP exercise. Forensic Sci Int Genet, 2015, **19**: 56-67

[11] 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

[12] Collins-Schramm H E, Chima B, Morii T, *et al*. Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. Hum Genet, 2004, **114**(3): 263-271

[13] Tishkoff S A, Kidd K K. Implications of biogeography of human populations for 'race' and medicine. Nat Genet, 2004, **36**(11 Suppl): S21-27

[14] Li J Z, Absher D M, Tang H, *et al*. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008, **319**(5866): 1100-1104

[15] Sudmant P H, Mallick S, Nelson B J, *et al*. Global diversity, population stratification, and selection of human copy-number variation. Science, 2015, **349**(6253): aab3761

[16] Hellenthal G, Busby GBJ, Band G, *et al*. A genetic atlas of human admixture history. Science, 2014, **343**(6172): 747-751

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

[18] 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

[19] Wang L. Support Vector Machines: Theory and Applications (Studies in Fuzziness and Soft Computing). New York: Springer-Verlag , 2005

[20] 1 000 Genomes Project Consortium, Auton A, Brooks L D, *et al*. A global reference for human genetic variation. Nature, 2015, **526**(7571): 68-74

[21] Liu X, Lu D, Saw W Y, *et al*. Characterising private and shared signatures of positive selection in 37 Asian populations. Eur J Hum Genet, 2017, **25**(4): 499-508

[22] Jinam T A, Kanzawa-Kiriyama H, Inoue I, *et al*. Unique characteristics of the Ainu population in Northern Japan. J Hum Genet, 2015, **60**(10): 565-571

[23] Qin P, Li Z, Jin W, *et al*. A panel of ancestry informative markers to estimate and correct potential effects of population stratification in Han Chinese. Eur J Hum Genet, 2014, **22**(2): 248-253

[24] Kim J J, Verdu P, Pakstis A J, *et al*. Use of autosomal loci for clustering individuals and populations of East Asian origin. Hum Genet, 2005, **117**(6): 511-519

[25] 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

[26] Guo SW. Linkage disequilibrium measures for fine-scale mapping: a comparison. Hum Hered, 1997, **47**(6): 301-314

[27] Barrett J C, Fry B, Maller J, *et al*. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics, 2005, **21**(2): 263-265

[28] Hwa H L, Wu M Y, Lin C P, *et al*. A single nucleotide polymorphism panel for individual identification and ancestry assignment in Caucasians and four East and Southeast Asian populations using a machine learning classifier. Forensic Sci Med Pathol, 2019, **15**(1): 67-74

[29] Carlson C S, Eberle M A, Rieder M J, *et al*. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet, 2004, **74**(1): 106-120

[30] 惠守博, 王文杰. 支持向量机分类算法中多元变量共线性问题的改进. 计算机工程与设计, 2006, **27**(8): 1385-1388
Hui S B, Wang W J. Computer Engineering and Design, 2006, **27**(8): 1385-1388

[31] 赵宇东, 刘嵘, 刘延龄, 等. 多元 logistic 回归的共线性分析. 中国卫生统计, 2000, **17**(5):259-261
Zhao Y D, Liu R, Liu Y L, *et al*. Chinese Journal of Health Statistics, 2000, **17**(5): 259-261

High-resolution SNP Ancestry Inference Model and Efficiency Evaluation in Three East Asian Populations*

WEN Hao[1)], WEI Yi-Liang[3)], GUO Xiao-Yuan[4)], SUN Chang-Chun[4)], XUE Si-Yao[4)], LIU Jing[2)], FAN Hong[1)**], JIANG Li[2)**]

([1)]*School of Computer Science, Shaanxi Normal University, Xi'an* 710119, China;
[2)]*Physical Evidence Evaluation Center of the Ministry of Public Security, Beijing* 100038, China;
[3)]*Key Laboratory of Phylogeny and Comparative Genomics of Jiangsu Province, Xuzhou* 221116, China;
[4)]*School of Forensic Medcine, Shanxi Medical University, Taiyuan* 030001, China)

Abstract　Single nucleotide polymorphism (SNP) profiling is a commonly used genetic tool for individual identification and ancestry inference in forensic genetics. This study collected ancestry informative SNPs (AISNPs) from literature and public libraries, and applied softmax regression, support vector machine and random forest, which were used to infer ancestry origins of Northern Han, Japanese and Korean, the three major populations in the North of East Asia. We analyzed 428 AISNPs in 103 northern Han samples and 104 Japanese samples from the 1 000 Genomes Project and 100 Korean samples from the Asian Diversity Project, using multiple linear regression collinearity diagnostics and random forest mean decrease accuracy to screen and optimize high-information AISNPs combinations which were used for ancestry inference linear and nonlinear prediction models, respectively. We constructed two discriminant models of softmax regression and support vector machine with 67-plex AISNPs and a random forest discriminant model with 42-plex AISNPs, achieving high-precision division of Northern Han, Japanese and Korean. The accuracy rates of the 5 times 10-fold cross-validation test of the softmax regression model, support vector machine model and random forest model were 95.19%, 95.77%, and 94.53%, respectively. The 67-plex and 42-plex AISNP prediction models established in this study can be used for genetic inference of the three major populations in the North of East Asia with high practical application value.

Key words　forensic genetics, ancestry informative SNPs, the North of East Asia, softmax regression, support vector machine, random forest
DOI：10.16476/j.pibb.2020.0339

* This work was supported by grants from the Key Research and Development Program of Shanxi Province (2018SF-251), The National Natural Science Foundation of China (81772027), Open Projects of National Engineering Laboratory (2018NELKFKT15), Open Projects of the Key Laboratory of Forensic Genetics of the Ministry of Public Security (2020FGKFKT01), and Major Projects of Universities in Jiangsu Province (17KJA180003).

** Corresponding author.
FAN Hong. Tel：86-15929807273，E-mail：fanhong@snnu.edu.cn
JIANG Li. Tel：86-18519073957，E-mail：jl@mail.bnu.edu.cn

Received：September 22，2020　Accepted：January 7，2021