PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2022,49(7):1348~1357
www.pibb.ac.cn

基于DNA变异的中国汉族人群脱发表型推断及预测模型评估*

薛思瑶[1,2] 李彩霞[2] 贠克明[1] 丛 斌[3]** 赵雯婷[2]**

([1] 山西医科大学法医学院，太原 030001；
[2] 公安部物证鉴定中心，现场物证溯源技术国家工程实验室，法医遗传学公安部重点实验室，北京 100038；
[3] 河北医科大学法医学院，石家庄 050017）

摘要 **目的** 男性型脱发（male pattern baldness，MPB），又称为雄激素性脱发（AGA），是一种常见的男性脱发类型，大约80%的表型差异可以用遗传因素解释。目前的MPB遗传推断研究主要基于欧洲人群，东亚人群相关研究较少。本研究在中国人群中对欧洲人群MPB关联位点进行验证分析，并建立遗传推断模型。**方法** 本研究调查了486个与欧洲人群MPB相关单核苷酸多态性（SNP）位点在312名中国汉族男性中的关联性，分别使用逐步回归和Lasso回归方法对关联出的位点进行筛选。使用逻辑回归算法构建预测模型，通过十折交叉验证的方法评估。之后进一步比较了逻辑回归、k近邻分类器、随机森林、支持向量机4种常用分类器模型对MPB的预测准确性。**结果** 有174个SNP位点与中国汉族男性的MPB显著相关（$P < 0.05$）。通过不同的筛选方法，分别得到了22个SNP和25个SNP的位点集合。基于上述位点集合建立了22-SNP和25-SNP两种逻辑回归预测模型。以*AUC*（ROC曲线下方的面积大小，area under curve）来衡量，两种模型对MPB预测的准确性分别为0.85和0.84；经十折交叉验证后预测准确性分别下降至0.81和0.77。当加入年龄作为预测因子后，两种模型的*AUC*均达到最大值0.89。从运行结果来看，逻辑回归预测模型较本研究中的其他分类器模型具有明显优势。**结论** 总体而言，虽然预测模型的准确性尚未达到临床期望水平，但SNP在MPB的遗传预测方面仍具备很大的潜力，可以为MPB的早期诊断、临床干预和法庭科学应用提供参考。

关键词 男性型脱发，预测模型，单核苷酸多态性，汉族人群
中图分类号 R89，D919 **DOI**：10.16476/j.pibb.2021.0329

脱发问题是近年来社会各界关注的热点问题，尤其在中青年人群中的发病率一直居高不下，对患者的心理、生活社交造成明显影响。人类最常见的脱发形式是男性型脱发（male pattern baldness，MPB），其特点是头皮上依赖雄激素的进行性脱发表现。MPB严重程度与年龄、脱发部位等密切相关，发病率随年龄以平均每10年提高10%的增速增长[1]，其在欧洲男性中的患病率很高，可达到80%[2]，而一项针对3 519名上海男性脱发情况的研究显示脱发患病率在19.9%左右[3]。有多项研究表明，与高加索人相比，中国人、日本人和非裔美国人的患病率较低[4]。

人群遗传学研究显示，MPB是一种高度遗传的多基因疾病[5]。早期针对双胞胎的研究表明[6]，MPB的遗传力约为81%；Liu等[7]基于单核苷酸多态性（single nucleotide polymorphisms，SNPs）常见变异的分子遗传学方法估计MPB的遗传力可达50%。近年来，随着基因分型技术和DNA测序技术的快速发展，尤其是全基因组关联分析（genome-wide association study，GWAS）的应用，MPB的遗传学研究取得了突破性进展，欧洲人群GWAS研究发现的MPB显著关联SNP位点已达

* 中央级公益性科研院所基本科研业务费专项资金（2018JB046）和国家科技资源共享服务平台计划（YCZYPT[2017]01-3）资助项目。
** 通讯联系人。
赵雯婷 Tel：010-83752706，E-mail：wtzhao@sibs.ac.cn
丛斌 Tel：13315998962，E-mail：hbydcongbin@126.com
收稿日期：2021-10-28，接受日期：2022-01-06

1 000个以上。比如，2017年针对8个独立的欧洲血统人群队列22 518个样本的荟萃分析[8]确定了63个MPB显著关联位点（6个位于X染色体上，57个位于常染色体上），同时揭示了脱发不是孤立的特征，而是可与许多其他人类表型具有相关性的，例如前列腺癌和神经退行性疾病等。迄今人群规模最大的MPB遗传分析来自2018年Visscher等[9]对UK Biobank 205 327个欧洲男性的研究，通过GWAS关联出了624个近独立的位点（598个位于常染色体上，26个位于X染色体上）。同年一项针对7万欧洲人群的GWAS研究关联出了71个独立遗传位点[10]，可解释总遗传力的38%。可见，MPB虽然是多基因复杂表型，但与身高等表型相比，可以用相对较少的SNP来解释较大比例的遗传力。因此，通过SNP位点建立准确性较高的MPB遗传预测模型是可行的。

已有的MPB遗传预测模型大多采用了逻辑回归算法。Hagenaars等[11]使用287个SNP位点建立多元逻辑回归模型，重度脱发的*AUC*（ROC曲线下方的面积大小，area under curve）为0.78，但轻度脱发和中度脱发的*AUC*仅能达到0.68和0.61。Liu等[7]针对2 725个德国和荷兰男性的研究尝试建立了25个SNP的逻辑回归模型，*AUC*=0.74。Marcińska等[12]使用305个50岁及以上的欧洲人群样本构建了20个SNP的模型，对脱发的遗传解释力为35%，*AUC*=0.86。

与欧洲人群MPB的遗传预测研究相比，针对中国人群的研究报道相对较少。在本实验室的前期研究中，潘思宇等[13]针对中国的欧亚混合人群建立了两种MPB预测模型，一种以年龄、BMI和25个SNP为预测因子，*AUC*=0.82；另一种是以年龄、BMI和68个SNP为预测因子，*AUC*=0.89。这两种预测模型虽然展现出良好性能，但在仅将年龄作为预测因子的情况下*AUC*值就可以达到0.77。可见该模型年龄依赖性过强，SNP的独立预测能力有待提高。

本研究选取了近十余年发表的关于MPB研究的16篇文献中486个欧洲人群关联SNP位点[7-8，10-12，14-24]，在312名中国汉族人群样本中进行关联验证分析，并基于筛选后的具有显著关联性的SNP位点建立了MPB逻辑回归预测模型，同时对k近邻分类器（k-nearest neighbor classifier）、随机森林（random forest）、支持向量机（support vector machine，SVM）等常见的分类器模型[25]在MPB遗传预测中的性能进行了比对评估，力求找到MPB预测准确性最高的建模方法。

1　材料与方法

1.1　男性型脱发表型的获取及分类标准

Hamilton-Norwood（H-N）脱发分级标准[1]根据发际线后移程度以及头顶部毛发稀疏程度将MPB划分为不脱发（I类）、6种MPB主类型（II至VII类）和5种亚类型（IIA至VA以及III vertex）。参照该标准，本研究将表型分为两组（图1）：a. MPB表型组，即头顶部可见明显脱发且发际线严重后移（IV、IVa、V、Va、VI和VII）；b. 对照表型组，即完全没有脱发或轻微发际线后移（I和II）。表型读取时，由3名评分者同时观看照片，并独立对每一位志愿者的MPB等级进行评级，排除表型判断有困难的样本，以3个评分者对每个志愿者分级结果的众数作为最终的MPB等级。

1.2　样本及DNA提取

按照**1.1**的表型分组标准，本研究共收集了中国不同地域的汉族男性个体312例，除7例样本为南方汉族（四川6、江西1）外，其余均为北方汉族（山东4、山西296、河南5）群体，其中MPB表型组143例，对照表型组169例，且所有研究个体无内分泌功能障碍类疾病、未接受过毛发相关治疗。考虑到年龄因素对MPB的影响[12]，MPB表型组年龄在28~69岁之间，平均年龄约53，而对照表型组选取了高龄不脱发的志愿者，年龄在55岁以上，平均年龄59岁左右。详细组内信息和外观概览见表1和图1。使用Canon EOS 5D Mark II（佳能，日本）高清照相机分别采集志愿者头部左侧、正面及右侧3张二维照片。本研究通过公安部物证鉴定中心伦理委员会审查，所有参与者均签署了书面知情同意书。

1.3　基因分型及质量控制

使用Illumina HiSeq X Ten测序平台（Illumina，美国）对样本进行3X低深度全基因组测序，每个样本得到平均10G Raw data。对经过变异检测（variant calling）处理后的数据，使用本实验室中国人群低深度测序2 510份样本进行基因填补。使用PLINK v1.9[26]对SNP进行质量控制，包括分型成功率（call rate）>0.97，哈迪温伯格平衡（Hardy-Weinberg equilibrium，HWE）$P > 0.000\,1$和次等位基因频率（minor allele frequency，MAF）>0.01。个体样本质量控制包括性别检查，亲缘关系检测及杂

Table 1 Sample information

Sample information	MPB	Controls
n	143	169
Age（mean±SD）	52.7±8.3	59.2±3.8
Age range	20–69	55–73
Level of MPB	IV（9），IVa（31），V（30），Va（8），VI（37），VII（28）	II（169）

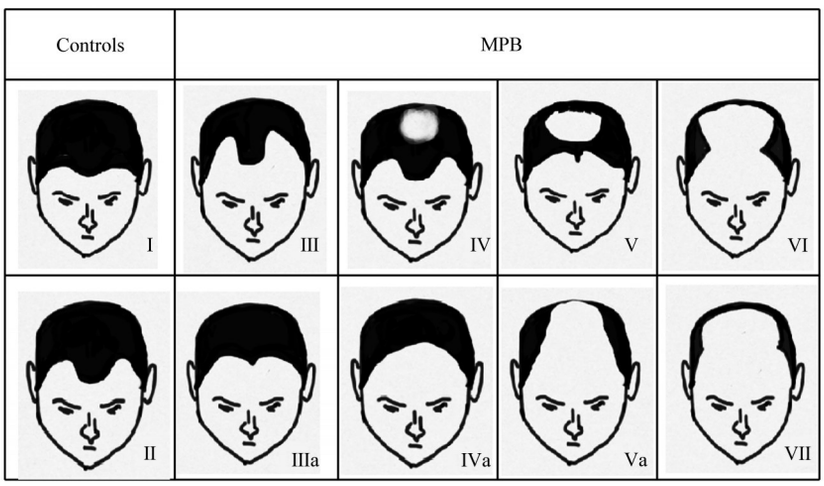


Fig. 1 Diagram of MPB

合性判断。以千人基因组数据第三阶段（1000 Genomes Project Phase 3）数据作为参考基因组，使用 IMPUTE [27] 对常染色体进行基因填补并过滤填补质量分数小于0.6的SNP，并再次重复上面质量控制标准，最终共获得20 681 872个SNP位点。

1.4 统计分析

1.4.1 遗传关联分析

本研究选用基于欧洲人群关联出的486个SNP位点，均通过了质量控制，详细位点信息见附件表S1。使用Plink v1.9软件（哈佛大学，波士顿，马萨诸塞州，美国）分别进行了一般线性回归（general linear model，GLM）和二元逻辑回归分析，测试486个SNP与MPB的相关性。基因型的赋值为加性模型，假设个体携带的次等位基因的数量与表型特征有累积效应。估计了所有SNP的优势比（odds ratio，OR）、相应的95%可信区间（confidence interval，CI）和*P*值。将*P*<0.05认为在关联分析中具有统计学意义。同时通过将所获得的*OR*与*OR*=1时相比，从而估计脱发风险增加倍数。使用wANNOVAR [28] 对与MPB相关性最高的前20个SNP进行相关基因区域识别。多重假设检验校正后没有达到显著关联性的位点，故而在本研究中没有应用多重假设检验的校正。

1.4.2 预测建模

将在关联分析中具有统计学意义的SNP位点作为建立预测模型的初始位点集合。首先对数据进行预处理，先将因变量的编码分为“1”（MPB表型）和“0”（对照表型），再依据次要等位基因数目对SNP基因型进行编码：具有2个次要等位基因编码为“2”，只有1个次要等位基因编码为“1”，不含次要等位基因编码为“0”。然后，采用两种方法对位点进行筛选，一种是基于R软件STEP函数对AIC信息标准进行逐步分析，另一种是通过R软件glmnet包建立Lasso回归模型，从而对SNP预测因子进行最终选择和排序。

逻辑回归适用于二值响应变量（即0和1），故选用二元逻辑回归对预测模型进行训练。模型假设因变量服从二项分布，模型的拟合形式为：

$$\log_e\left(\frac{\pi}{1-\pi}\right)=\beta_0+\sum_{j=1}^{P}\beta_j X_j \quad (1)$$

其中$\left(\frac{\pi}{1-\pi}\right)$为因变量为1时的*OR*，$\log\left(\frac{\pi}{1-\pi}\right)$为对数*OR*。

由于本研究样本量较小，采用十折交叉验证法来防止过度拟合。将MPB的预测概率与观察到的MPB状态进行比较，将AUC作为预测准确性的总

体衡量标准。*AUC*值的范围从0.5到1.0，0.5表示随机预测，1.0表示完全准确的预测。如果预测概率＞0.5，则定义受试者为MPB，否则为不脱发。使用混淆矩阵比较预测和观察的脱发状态，并得出灵敏度和特异值，两者的范围都在0到1之间。所有候选SNP分析和预测分析都在R v4.0.2（http://www.r-project.org/）中进行。

1.4.3　多模型对比评估

在R软件中分别对k近邻分类器、随机森林、支持向量机3种机器学习算法进行建模-验证，获得不同模型的预测准确性从而对比模型的预测性能。建模过程中使用的R包主要包括class包、kknn包、randomForest包、e1027包等。每种机器学习算法共运行10次，求其平均值。对于二分类任务，可将验证样本的真实情况作为金标准，对所有验证样本的模型分类结果和金标准结果分别计数，从而获得分类器性能混淆矩阵。分别计算模型的正确率、敏感度、特异度、阳性预测值、阴性预测值以及五折交叉验证的预测准确性。以上分类器性能衡量标准的取值范围均为0~1，值越大，表示分类性能越高。

2　结果与分析

2.1　遗传关联分析结果

通过对312个样本的486个SNP进行关联分析，发现174个SNP与MPB显著相关，相关性最显著的20个见表2。在与MPB相关的SNP中，位于chr20上的位点最多，有145个，chr5和chr6上各有6个SNP，chr1上有4个SNP，chr9和chr10上各有3个SNP，chr2上有2个SNP，chr3、chr7、chr8、chr15、chr19上各有1个SNP。与前20个显著位点有关的基因分别为EBF1、TFAP2C、PAX1以及RUNX3。位于EBF1的rs17643057在chr5上的分布具有最高的统计学意义（逻辑回归关联分析*OR*=0.479，95% *CI*=0.321~0.714，*P*=3.42×10^{-4}）。值得注意的是，当应用一般线性回归关联分析时，前9个SNP的显著性更高。据估计，携带rs985546-C等位基因的男性患MPB的风险是携带T等位基因男性的3.4倍。从*OR*来看，其余3个与MPB易感性相关最显著的SNP是rs17643057-G（chr5）、rs1422798-G（chr5）和rs6113382-A（Chr20），使MPB的风险分别增加2.1、2.0和1.9倍。

Table 2　Top 20 SNPs most significantly associated with MPB in Han Chinese （*P*＜0.05）

Chr	SNP ID	Genes	Effect/ Other	MAF		*OR*（95% *CI*）		*P*	
				Cases	Controls	LR	GLM	LR	GLM
5	rs17643057	EBF1	G/A	0.15	0.28	0.479（0.321–0.714）	0.465（0.312–0.693）	3.09E–04	1.42E–04
20	rs985546	TFAP2C	C/T	0.04	0.12	0.288（0.146–0.569）	0.309（0.159–0.599）	3.42E–04	2.69E–04
5	rs1422798	EBF1	G/C	0.15	0.26	0.508（0.341–0.757）	0.482（0.321–0.724）	8.89E–04	3.72E–04
20	rs6137444	PAX1	C/T	0.27	0.39	0.547（0.384–0.779）	0.558（0.397–0.785）	8.18E–04	7.54E–04
20	rs6047683	PAX1	A/C	0.30	0.42	0.595（0.425–0.833）	0.594（0.462–0.828）	2.49E–03	2.02E–03
20	rs6047684	PAX1	G/A	0.29	0.41	0.587（0.416–0.828）	0.595（0.426–0.832）	2.39E–03	2.27E–03
20	rs2180439	PAX1	C/T	0.30	0.42	0.607（0.435–0.848）	0.603（0.433–0.841）	3.46E–03	2.76E–03
1	rs11249243	RUNX3	T/C	0.15	0.24	0.544（0.360–0.821）	0.537（0.356–0.811）	3.80E–03	2.82E–03
20	rs6113382	PAX1	A/C	0.15	0.24	0.537（0.354–0.815）	0.537（0.356–0.811）	3.46E–03	2.82E–03
20	rs6113393	PAX1	C/T	0.29	0.40	0.592（0.418–0.839）	0.607（0.434–0.849）	3.23E–03	3.45E–03
20	rs4815081	PAX1	G/A	0.29	0.42	0.592（0.417–0.840）	0.610（0.437–0.853）	3.29E–03	3.67E–03
20	rs2328645	PAX1	C/G	0.30	0.42	0.604（0.430–0.849）	0.608（0.436–0.848）	3.64E–03	3.29E–03
20	rs2328646	PAX1	G/A	0.30	0.42	0.604（0.430–0.849）	0.608（0.436–0.848）	3.64E–03	3.29E–03
20	rs6047682	PAX1	C/A	0.30	0.41	0.617（0.442–0.860）	0.611（0.438–0.851）	4.42E–03	3.50E–03
20	rs1535199	PAX1	G/A	0.29	0.41	0.596（0.422–0.844）	0.610（0.437–0.853）	3.51E–03	3.67E–03
20	rs6082524	PAX1	C/T	0.29	0.41	0.596（0.422–0.844）	0.610（0.437–0.853）	3.51E–03	3.67E–03
20	rs6082532	PAX1	C/T	0.29	0.41	0.596（0.422–0.844）	0.610（0.437–0.853）	3.51E–03	3.67E–03
20	rs6035970	PAX1	C/T	0.29	0.41	0.596（0.422–0.844）	0.610（0.437–0.853）	3.51E–03	3.67E–03
20	rs6075849	PAX1	T/C	0.29	0.41	0.596（0.422–0.844）	0.610（0.437–0.853）	3.51E–03	3.67E–03
20	rs6047664	PAX1	G/T	0.29	0.40	0.597（0.421–0.848）	0.615（0.440–0.860）	3.91E–03	4.36E–03

2.2　位点筛选

逐步回归不仅可以从备选因子中筛选出最终预测变量，还可以防止模型过度拟合。本研究通过双向逐步回归的方法，根据提前设定的赤池信息准则（Akaike information criterion，AIC），将直接纳入模型的174个MPB相关SNP精简至22个SNP用于下游预测模型的建立。此时的*AIC*达到最小值，跨度区间为322.38~305.81（图2）。每一预测因子的方差膨胀系数（variance inflation factor，VIF）均小于10，不存在多重共线性问题。

Lasso回归基于惩罚系数λ对备选因子进行筛选，随着惩罚系数λ的增大，模型回归系数β逐渐趋近于0，最终变为0（图3a，b）。图3a左侧虚线对应使模型估计误差最小的λ，右侧虚线对应使模型估计误差在可接受范围内的λ，根据最高效原则确定纳入模型的最优变量组合，最终筛选出25个SNP位点。

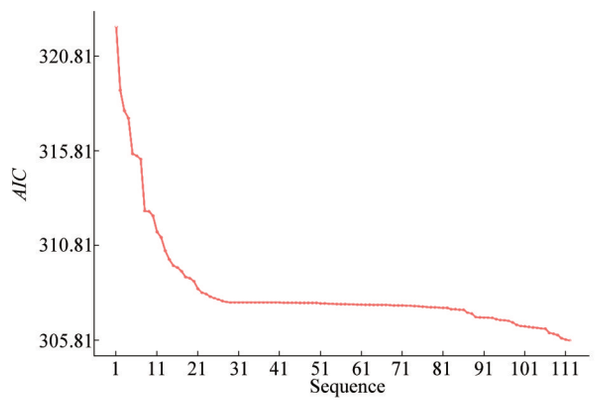


Fig. 2　Characteristic variable screening based on stepwise regression

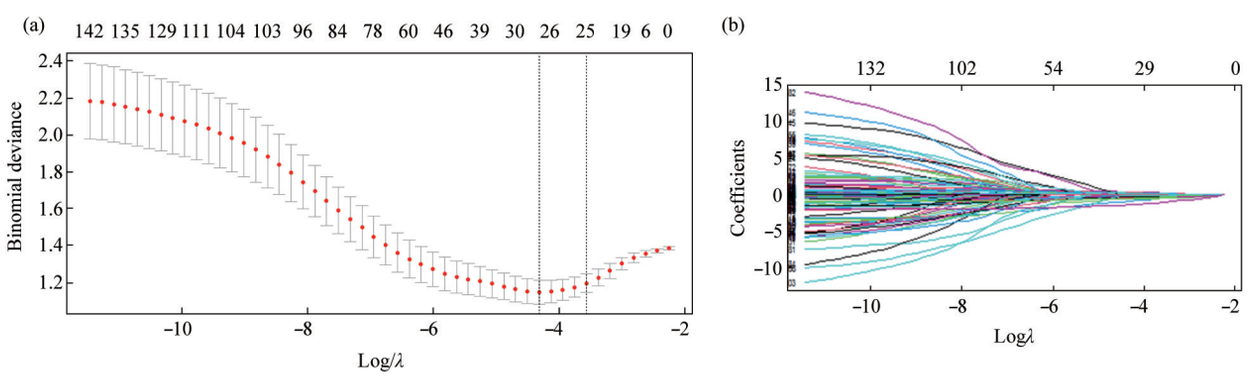


Fig. 3　Characteristic variable screening based on Lasso regression

The figure shows the process of selecting the most appropriate value of parameter λ in the Lasso model by cross-validation. The numbers on the upper horizontal axis represent the amounts of SNPs that can be incorporated into the model. (a) Cross-validation plot for the penalty term; (b) Plots for Lasso regression coefficients over different values of the penalty parameter.

2.3　预测模型

根据上述筛选得到的两种位点集合，建立了两个预测模型。两模型所包含位点信息见表3，详细位点信息见附件表S2。

第一个模型包括通过逐步回归分析筛选出的22个SNP，该模型解释了患MPB总风险的48%（R^2=0.48）。第二个模型包括通过Lasso回归筛选出的25个SNP，该模型解释了患MPB总风险的45%（R^2=0.45）。MPB预测模型具有总体预测精度，区分度指标分别为*AUC*=0.85和*AUC*=0.84，ROC曲线见图4。应用50%的概率阈值，22-SNP预测模型正确预测的总数为76%（236/309），有3个不确定结果。然而，65%的概率阈值的正确预测降低到75%（234/311），有1个不确定结果。同样，应用50%的概率阈值，25-SNP预测模型正确预测的总数为74%（228/309），有3个不确定结果。而65%的概率阈值的正确预测保持不变，仍为74%（226/307），有5个不确定结果。两模型均通过十折交叉验证的方法进行验证，验证后的*AUC*分别为0.81和0.77。在加入年龄作为预测因子之一后，预测准确性分别提升到了80%（251/312）和81%（252/312），没有不确定结果。相比较而言，通过Lasso回归筛选出来的位点在十折交叉验证过程中*AUC*有一定程度的下滑，且有个别位点存在多重共线性问题。22-SNP预测模型和25-SNP预测模型在18个SNP上相同，仅存在4~7个位点差异，但22-SNP预测模型在各项指标上均优于25-SNP预测模型。在加入年龄作为预测因子后，两模型的预测准确率等各指标均有提升，整体表现*AUC*均为0.89（表4）。

Table 3　Information of 22− and 25−SNP used in predictive model building

22-SNP		25-SNP	
rs11249243	rs6982226	rs11249243	rs9398035
rs17185996	**rs12686549**	rs17185996	rs17350355
rs13405699	**rs2416699**	rs13405699	rs6982226
rs9878451	**rs3118470**	**rs16862069**	**rs3781452**
rs335145	**rs3781458**	rs9878451	rs2028122
rs17643057	rs2028122	rs335145	rs17318596
rs77239429	rs17318596	rs17643057	**rs6137444**
rs11243290	rs6113382	rs77239429	rs6113382
rs79032435	rs75434917	**rs4959410**	**rs6047683**
rs9398035	rs199791	rs11243290	rs75434917
rs17350355	rs985546	**rs9380830**	rs199791
		rs79032435	**rs6072223**
			rs985546

Non-repeating SNPs showed in bold text.

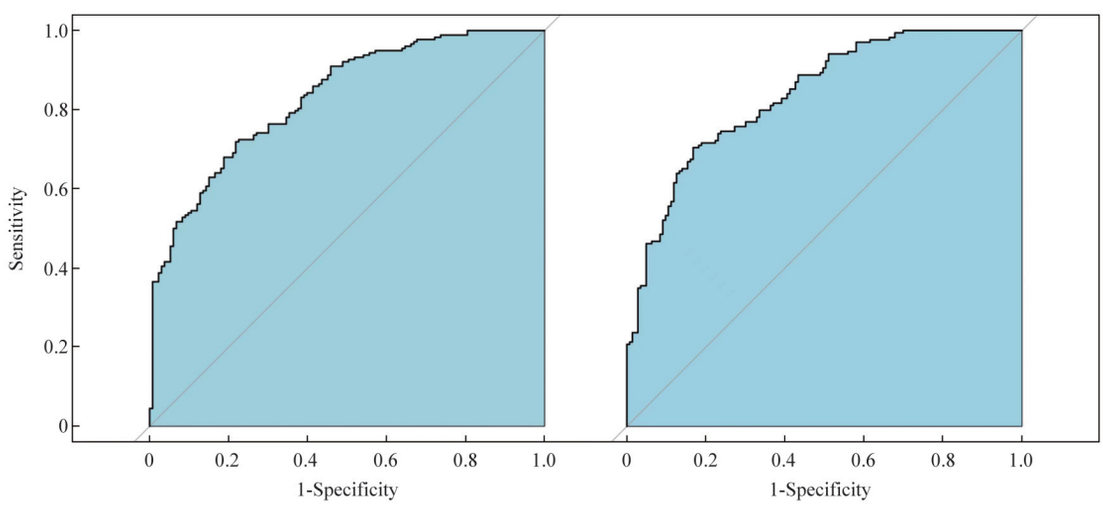


Fig. 4　Receiver operating characteristic (ROC) curves for 22−SNP (left) and 25−SNP (right) MPB prediction models

The ROC curves have sensitivity as the ordinate.

Table 4　Prediction performance for MPB with different SNP−sets and factors

Model type	Age[1)]	Accuracy/%	*AUC*	Sensitivity	Specificity	PPV (positive predictive value)	NPV (negative predictive value)	*AUC* of 10-fold cross-validation	R^2
22-SNP	0	76%	0.85	0.71	0.80	0.75	0.77	0.81	0.48
	1	80%	0.89	0.77	0.83	0.80	0.83	0.89	0.57
25-SNP	0	74%	0.84	0.71	0.76	0.71	0.76	0.77	0.45
	1	81%	0.89	0.79	0.82	0.79	0.82	0.89	0.56

[1)] The 0 and 1 in the second column represent age was exclude or included in the model.

2.4　分类器模型性能评估

通过混淆矩阵获得的分类器模型性能评价见表5。在3种分类器模型中，最高的准确率是基于22-SNP的支持向量机分类器模型，但也仅能到达68%，其预测效能和预测准确性远不如逻辑回归模型。

Table 5 Performance comparison of k−NN, random forest and SVM for MPB prediction

	Model type	Accuracy	Sensitivity	Specificity	PPV	NPV	prediction accuracy of validation (k=5)
k-Nearest neighbor classifier (k-NN)	22-SNP	0.59	0.57	0.60	0.63	0.55	0.60
	25-SNP	0.60	0.64	0.55	0.63	0.56	0.59
Random forest	22-SNP	0.63	0.44	0.78	0.59	0.66	0.70
	25-SNP	0.59	0.52	0.64	0.55	0.62	0.63
SVM	22-SNP	0.68	0.51	0.80	0.65	0.69	0.72
	25-SNP	0.60	0.50	0.69	0.57	0.63	0.69

3 讨 论

本研究首次在中国汉族人群中进行较为系统的MPB相关位点验证分析，并初步筛选出与中国人群MPB表型相关的SNP位点，同时构建出性能较高的非年龄依赖MPB预测模型。

从关联分析结果来看，chr20上的多个SNP与MPB具有强关联性，这说明chr20不仅是欧洲人群MPB的主要危险区域[29]，也是中国汉族人群MPB的主要危险区域，这提示了在不同祖源人群中MPB可能存在相似的遗传机制。本研究中关联性最显著的SNP位点（rs17643057）所在基因区域已被欧洲研究证实与毛发生长特征有关[8]，受早期B细胞因子1（early B cell factor 1，EBF1）调控。EBF1是早期B细胞发育和脂肪形成所必需的转录因子，动物研究表明其在小鼠成熟、生长的毛囊中表达。除EBF1外，在本研究前20个显著关联位点中，有75%以上位点与PAX1（paired Box 1，PAX1）这一基因区域有关。PAX1在皮肤、头发和头皮中表达，是典型的MPB易感位点[17, 29]。这提示了将PAX1作为中国汉族人群MPB候选基因的必要性。值得一提的是，本研究关联出的显著位点（rs2180439）在另一项基于中国汉族人群的研究中[14]同样被证实与脱发显著相关，效应方向与本研究一致，超过了统计意义的关联阈值（$P \leqslant 3.13\times10^{-3}$）。对于那些关联性较低的SNP，本文暂时无法验证SNP是否与MPB存在真实关联，需要进一步扩大样本量来提升结果的准确性。

为了进一步优化MPB相关SNP位点集合以建立预测模型，本研究采用了两种不同的位点筛选方法，并获得22-SNP和25-SNP两组位点集合。这样做的目的一方面是为了比较两种位点筛选办法所获得的SNP对模型的预测性能所造成的差异，另一方面是为了防止模型过度拟合。若模型过度拟合，其在外部验证中的表现就会变差。在仅使用SNP作为预测因子的情况下，基于22-SNP和25-SNP脱发的二分类预测模型均表现出了良好的性能。在加入年龄作为预测因子后，模型的预测性虽有小幅提升，但不能排除在高龄对照样本的影响下，年龄所产生的虚假相关性。在实验室前期研究成果中，不加入年龄作为预测因子的前提下，模型*AUC*低于0.7[13]。说明本研究所采用的表型组、对照组样本筛选方法，显著降低了年龄对关联结果的影响，筛选出的位点对表型的影响效力更强，所解释的遗传力度相较前期研究的不足30%也有显著提升。

已有的MPB预测模型大多基于逻辑回归算法，本研究进一步探索了不同分类器模型对MPB表型的预测性能。从逻辑回归、k近邻分类器、随机森林、SVM这4种常用分类器模型在本研究人群的运行结果来看，逻辑回归模型具有明显优势。

4 结 论

本研究通过将欧洲人群MPB关联位点在中国人群的验证分析，为了解中国汉族人群MPB的遗传机制奠定了基础。同时，所构建的预测模型，能够在不依赖年龄作为预测因子的条件下，达到较为优良的预测性能。在后续的研究工作中，通过扩大样本量、采用全基因组关联分析、引入表观遗传分析等方法，有望得到更优的MPB相关遗传位点集合，建立更为精准的MPB预测模型，应用到临床医学诊断和法庭科学领域中。

附件 PIBB_20210329_S1.pdf见本文网络版（http://www.pibb.ac.cn或http://www.cnki.net）。

参 考 文 献

[1] Norwood O T. Male pattern baldness: classification and incidence. South Med J,1975, **68**(11): 1359

[2] Hamilton J B. Patterned loss of hair in man: types and incidence. Ann N Y Acad Sci, 1951, **53**(3): 708-728

[3] Xu F,Sheng Y Y,Mu Z L *et al*. Prevalence and types of androgenetic alopecia in Shanghai, China: a community-based study. Br J Dermatol, 2009, **160**: 629-632

[4] Otberg N, Finner A M, Shapiro J. Androgenetic alopecia. Endocrinology Metab Clin North Am, 2007, **36**(2): 379-398

[5] Lakhani K, Constantinovici N, Purcell W, *et al*. Internal carotid-artery response to 5% carbon dioxide in women with polycystic ovaries. Lancet, 2000, **356**(9236): 1166-1167

[6] Heath A C, Nyholt D R, Gillespie N A, *et al*. Genetic basis of male pattern baldness. J Invest Dermatol, 2003, **121**(6): 1561-1564

[7] Liu F, Hamer M A, Heilmann S, *et al*. Prediction of male-pattern baldness from genotypes. Eur J Hum Genet, 2016, **24**(6): 895-902

[8] Heilmann-Heimbach S, Herold C, Hochfeld L M, *et al*. Meta-analysis identifies novel risk loci and yields systematic insights into the biology of male-pattern baldness. Nat Commun, 2017, **8**(1): 14694

[9] Yap C X, Sidorenko J, Wu Y, *et al*. Dissection of genetic variation and evidence for pleiotropy in male pattern baldness. Nat Commun, 2018, **9**(1): 5407

[10] Pirastu N, Joshi P K, De Vries P S, *et al*. GWAS for male-pattern baldness identifies 71 susceptibility loci explaining 38% of the risk. Nat Commun, 2017, **8**(1): 1584

[11] Hagenaars S P, Hill W D, Harris S E, *et al*. Genetic prediction of male pattern baldness.PLoS Genet, 2017, **13**(2): e1006594

[12] Marcińska M, Pośpiech E, Abidi S, *et al*. Evaluation of DNA variants associated with androgenetic alopecia and their potential to predict male pattern baldness. PLoS One, 2015, **10**(5): e0127852

[13] 潘思宇,赵雯婷,冯锐,等.基于DNA变异在亚欧混合人群中预测男性型脱发.生物化学与生物物理进展, 2020, **47**(10):1069-1079
Pan S Y, Zhao W T, Feng R, *et al*. Prog Biochem Biophys, 2020, **47**(10):1069-1079

[14] Liang B, Yang C, Zuo X, *et al*. Genetic variants at 20p11 confer risk to androgenetic alopecia in the Chinese han population. PLoS One, 2013, **8**(8): e71771

[15] Heilmann S, Kiefer A K, Fricker N, *et al*. Androgenetic alopecia: identification of four genetic risk loci and evidence for the contribution of WNT signaling to its etiology. J Invest Dermatol, 2013, **133**(6): 1489-1496

[16] Li R, Brockschmidt F F, Kiefer A K, *et al*. Six novel susceptibility loci for early-onset androgenetic alopecia and their unexpected association with common diseases. PLoS Genet, 2012, **8**(5): e1002746

[17] Hillmer A M, Brockschmidt F F, Hanneken S, *et al*. Susceptibility variants for male-pattern baldness on chromosome 20p11. Nat Genet, 2008, **40**(11): 1279-1281

[18] Brockschmidt F F, Heilmann S, Ellis J A, *et al*. Susceptibility variants on chromosome 7p21.1 suggest HDAC9 as a new candidate gene for male-pattern baldness. Br J Dermatol, 2011, **165**(6): 1293-1302

[19] Megiorni F, Mora B, Maxia C, *et al*. Cytotoxic T-lymphocyte antigen 4 (CTLA4) +49AG and CT60 gene polymorphisms in Alopecia Areata: a case-control association study in the Italian population. Arch Dermatol Res, 2013, **305**(7): 665-670

[20] Lee S, Paik S H, Kim H J, *et al*. Exomic sequencing of immune-related genes reveals novel candidate variants associated with Alopecia Universalis. PLoS one, 2013, **8**(1): e53613

[21] Jagielska D, Redler S, Brockschmidt F F, *et al*. Follow-up study of the first genome-wide association scan in Alopecia Areata: IL13 and KIAA0350 as susceptibility loci supported with genome-wide significance. J Invest Dermatol, 2012, **132**(9): 2192-2197

[22] Forstbauer L M, Brockschmidt F F, Moskvina V, *et al*. Genome-wide pooling approach identifies SPATA5 as a new susceptibility locus for alopecia areata. Eur J Hum Genet, 2012, **20**(3): 326-332

[23] Redler S, Brockschmidt F F, Tazi-Ahnini R, *et al*. Investigation of the male pattern baldness major genetic susceptibility loci AR/EDA2R and 20p11 in female pattern hair loss. Br J Dermatol, 2012, **166**(6):1314-1318

[24] Redler S, Brockschmidt F F, Forstbauer L, *et al*. The TRAF1/C5 locus confers risk for familial and severe alopecia areata: TRAF1/C5 locus in alopecia areata. Br J Dermatol, 2010, **162**(4): 866-869

[25] Kotsiantis S B. Supervised Machine learning: a review of classification techniques. Information, 2007, **31**: 249-268

[26] 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

[27] Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 (Bethesda), 2011, **1**(6): 457-470

[28] Chang X, Wang K. wANNOVAR: annotating genetic variants for personal genomes *via* the web. J Med Genet, 2012, **49**(7):433-436

[29] Richards J B, Yuan X, Geller F, *et al*. Male-pattern baldness susceptibility locus at 20p11. Nat Genet, 2008, **40**(11): 1282-1284

Phenotypic Prediction of Male-pattern Baldness in Chinese Han Population Based on DNA Variants*

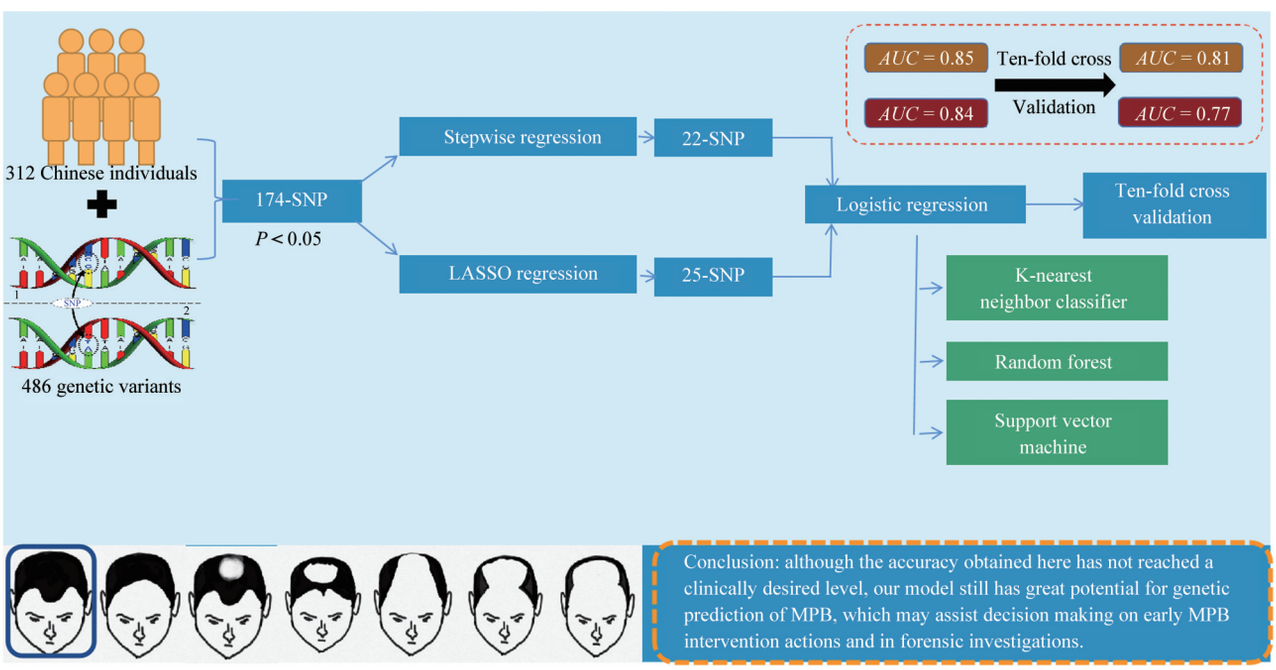
XUE Si-Yao[1,2)], LI Cai-Xia[2)], YUN Ke-Ming[1)], CONG Bin[3)**], ZHAO Wen-Ting[2)**]

([1)]*Istitute of Forensic Medicine, Shanxi Medical University, Taiyuan* 030001, China;

[2)]*National Engineering Laboratory for Forensic Science, Key Laboratory of Forensic Genetics, Institute of Forensic Science, Beijing* 100038, China;

[3)]*College of Forensic Medicine, Hebei Medical University, Shijiazhuang* 050017, China)

Graphical Abstract



Abstract　Objective　Male pattern baldness (MPB), or androgenetic alopecia (AGA), is a common type of hair loss in men, with an estimation that approximately 80% of the phenotypic variance can be explained by genetic factors. Most prediction models were developed in European and few MPB associated (single nucleotide polymorphisms, SNPs) have been validated in East Asian population. In this study, MPB associated SNPs in European were verified in Chinese population, and MPB risk prediction models were built based on those SNP data. **Methods**　We examined 486 genetic variants previously reported associated with MPB, and assessed their impacts on hair loss in 312 Chinese individuals. Different sets of SNPs were selected by stepwise regression and

* This work was supported by a grant from Central Public-Interest Scientific Institution Basal Research Fund (2018JB046) and National Science and Technology Resource Sharing Service Platform (YCZYPT[2017]01-3).

** Corresponding author.

ZHAO Wen-Ting. Tel：86-10-83752706，E-mail：wtzhao@sibs.ac.cn

CONG Bin. Tel：86-13315998962，E-mail：hbydcongbin@126.com

Received：October 28，2021　Accepted：January 6，2022

Lasso regression. Logistic regression algorithm was used to construct the prediction models and the evaluations were conducted by the method of 10-fold cross validation. We further compared the prediction accuracy among logistic regression, k-nearest neighbor classifier, random forest and support vector machine. **Results** 174 SNPs demonstrated significant associations with MPB (P<0.05). Among those SNP markers, 22 SNPs and 25 SNPs were selected by different screening methods. Two logistic regression model considering the genotypes of 22 and 25 SNPs demonstrate that the risk of MPB were predictable at *AUC* (area under curve) level of 0.85 and 0.84. Prediction accuracy was slightly reduced after performing 10-fold cross validation, 0.81 and 0.77 respectively. Moreover, the *AUC* of both models reaches maximum (0.89) when age was added as a predictive factor. From the running results, the logistic regression prediction model had obvious advantages. **Conclusion** Overall, although the accuracy obtained here has not reached a clinically desired level, our model still has great potential for genetic prediction of MPB, which may assist decision making on early MPB intervention actions and in forensic investigations.

Key words male pattern baldness (MPB), prediction model, single nucleotide polymorphism, Han Chinese
DOI：10.16476/j.pibb.2021.0329