

基于DNA变异在亚欧混合人群中
预测男性型脱发*潘思宇^{1)**} 赵雯婷^{2)**} 冯锐³⁾ 李琮³⁾ 景晓溪¹⁾ 高行健¹⁾陈燕¹⁾ 李祎¹⁾ 陶现明¹⁾ 刘天资^{1)***} 李彩霞^{2)***}⁽¹⁾ 中国科学院北京基因组研究所国家生物信息中心, 中国科学院精准基因组医学重点实验室, 中国科学院大学生命科学学院, 北京 100101;⁽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⁾ 湖南省公安厅刑侦总队, 长沙 410001)

摘要 男性型脱发 (male pattern baldness, MPB) 是一种雄激素依赖性的遗传性脱发疾病, 对个人形象、心理产生较大的消极影响. 近期欧美人群中进行的大样本全基因组关联分析 (genome wide association studies, GWAS) 已报道大量与 MPB 相关的遗传易感性单核苷酸多态性 (single nucleotide polymorphisms, SNPs) 位点, 但这些位点在东亚人群中的遗传效应尚不清楚. 本研究基于我国 684 个亚欧混合人群 (Eurasian) 样本, 对在 UK Biobank 205 327 个欧洲男性 GWAS 分析发现的 624 个与 MPB 相关的 SNPs 进行人群异质性分析, 基于多基因风险打分 (polygenic risk scores, PRS) 建立预测模型, 并对预测因子数量与模型预测性能的关系进行了研究. 通过质控的 467 个 SNPs 中 6.9% 与 MPB 显著相关 ($P < 0.05$). 结合年龄、体质指数 (body mass index, BMI) 和 25 个 SNPs 建立的线性回归和逻辑回归模型, 效果较好 ($R^2 = 28.9\%$, $AUC = 0.82$). 年龄对模型效果影响较大 ($R^2 = 22.9\%$, $AUC = 0.77$), 结合 BMI 及 68 个 SNPs 时 AUC 达到最大 (约 0.89). 本研究表明 MPB 在欧洲和东亚人群中存在较强的遗传异质性, 选取 SNPs 子集能达到与全集相近的预测准确性, 预测模型有助于东亚人群 MPB 遗传机制的理解及疾病的早期诊断和预防.

关键词 男性型脱发, 脱发预测模型, 人群异质性, 亚欧混合人群**中图分类号** Q3, R89**DOI:** 10.16476/j.pibb.2020.0095

男性型脱发 (male pattern baldness, MPB), 也称雄激素性脱发 (androgenetic alopecia, AGA), 是最常见的雄激素依赖性的遗传性脱发疾病, 典型的临床表现为发际线衰退, 前额及头顶毛发逐渐稀疏、脱落的过程^[1-2]. 其严重程度与年龄、脱发部位等相关, 发病率随年龄的增长而增长, 平均每 10 年约增长 10%^[3]. 该疾病可能始发于青春期后的任何时期, 高达 80% 的欧洲男性都会有一定程度脱发的风险^[4]; 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^[5]. 男性发病率明显高于女性^[6]. 然而, 男性型脱发并不同于传统疾病, 虽然这是一种外貌无自觉症状, 但却可能对患者的心理、社会认知及生活质量造成严重影响^[7-8]. 除此之外, 研究发现, 严重 MPB 的 BMI 比轻中度 MPB 的 BMI 有显著增高^[9], MPB 也与许多

疾病发生风险密切相关, 主要包括帕金森病^[10]、前列腺癌^[11]、冠心病及心血管疾病^[12] 和胰岛素抵抗相关疾病 (如肥胖、高血压、血脂紊乱)^[13]. 但是, 该病的遗传易感性和遗传模式尚未研究清楚,

* 国家重点研发项目 (2017YFC0803501), 现场物证溯源技术国家工程实验室开放课题 (2017NELKFKT05), 国家自然科学基金 (91651507), 中央级公益性科研院所基本科研业务费项目 (2018JB046), 中国科学院基因组与精准医学重点实验室开放课题, “万人计划”科技创新领军人才经费 (SQ2018RA23120013) 和国家科技资源共享服务平台计划 (YCYPT[2017]01-3 和 2017JB025) 资助项目.

** 并列第一作者.

*** 通讯联系人.

刘天资. Tel: 010-84097824, E-mail: liutz@big.ac.cn

李彩霞. Tel: 010-63264166, E-mail: licaixia@tsinghua.org.cn

收稿日期: 2020-04-11, 接受日期: 2020-07-07

对疾病的防控和治疗尚无很好的临床方案. 因此, 如果可以从DNA预测脱发程度, 不仅有利于对MPB遗传机制的了解, 也能为疾病的早期诊断提供参考依据.

MPB是一种高度遗传的多基因疾病^[14], 早期的双胞胎研究估测其遗传力约为81% (95%置信区间为0.77~0.85)^[15], 最近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测为50%^[16], 表明遗传效应是影响脱发过程的主要因素. 已知的MPB相关遗传位点主要位于X染色体的Xq12 (EDAR/AR)^[17]、常染色体的20p11^[18-19]、7p21.1 (HDAC9)^[20]等^[10, 21]. 2017年发表的3个较大规模人群样本集的荟萃分析 (meta analysis)^[22-24]发现了大量MPB相关位点. 目前最大人群样本的MPB遗传分析来自于Peter Visscher实验室对英国生物样本库 (UK Biobank) 205 327个欧洲男性的研究, 对狭义遗传力、多基因效应等进行了较全面的分析, 发现624个相关SNPs, 可以解释25%的表型差异^[25]. 这些相关SNPs位点的发现, 为MPB的预测提供了重要的遗传数据参考. 几个研究尝试通过多基因风险打分 (polygenic risk scores, PRS) 对男性的脱发程度进行预测. Marcinska等^[26]基于候选基因筛选, 使用多元逻辑回归分别构建的5-SNP和扩展的20-SNP预测模型, 对早发性脱发 (年龄<50岁) 具有较好的预测效果, 但对整个年龄范围的预测效果并不理想. Liu等^[16]构建的20-SNP预测模型在3个欧洲群体中进行评估, 对早发性脱发群体预测的AUC为0.74, 但对两个晚发性脱发群体预测能力相对较弱 (AUC分别为0.69和0.71). Hagenaaers等^[23]基于大型全基因组关联分析 (genome wide association studies, GWAS) 发现与MPB相关联的多个常见变异, 构建了287-SNP预测模型, 对严重脱发预测效果较好 (AUC为0.78), 但对轻微及中度脱发预测能力相对较弱 (AUC分别为0.61和0.68), 且该模型位点数量较多, 不易检测及应用. 这些研究主要集中在欧美人群, 已发现的位点是否在其他族群仍然具有适用性尚且未知.

遗传混合人群兼具多人群的表型特征, 是国际公认的研究人类复杂性状遗传结构的理想人群. 因此, 本研究选用东亚学者常用的亚欧混合人群 (Eurasian) 作为研究对象^[27], 其独特的遗传背景适合进行欧洲已知MPB遗传位点在亚洲人群中的验证评估研究. 通过对Visscher实验室624个SNPs

位点的分析, 本研究探索了人群异质性对MPB的影响, 建立了PRS预测模型, 评估了位点的预测效力, 并进一步研究预测因子数量与模型预测效能的关系, 且在千人基因组数据库模拟了MPB的全球地理分布情况.

1 材料与方法

1.1 样本及DNA提取

本研究共收集我国新疆维吾尔自治区亚欧混合人群无亲缘关系的男性个体684例. 所有研究个体同时满足以下几个条件: a. 父母和祖父母均为亚欧混合人群; b. 未接受过毛发修复手术或激素治疗; c. 无内分泌功能障碍类疾病, 如甲状腺疾病、脑垂体疾病和性腺疾病等. 使用Canon EOS 5D Mark II (佳能, 日本) 高清照相机分别采集志愿者左侧、正面及右侧3张二维照片. 使用外周血基因组DNA提取试剂盒QIAamp® DNA Blood Mini Kit (Qiagen, 希尔登, 德国), 按照生产商推荐的指导手册提取样本的DNA, EDTA-Vacutainer试管收集并置于-20℃冰箱中保存. 本研究通过公安部物证鉴定中心伦理委员会审查, 且所有参与者均签署了书面知情同意书.

1.2 男性型脱发表型的获取

脱发表型按照经Hamilton-Norwood (H-N)^[3]分级标准改编的4分级脱发系统^[25] (图1a) 评估: 第一类为没有或轻微脱发 (H-N等级I和II); 第二类为额部脱发 (H-N等级IIa、III、IIIa和IVa); 第三类为头顶脱发 (H-N等级IIIv); 第四类为额部和头顶均脱发 (IV、V、Va、VI和VII). 3名评分者随机选取150张图片, 同时观察每个样本的3张图片, 通过与分类系统的相似程度, 进行表型打分训练, 待对表型分级标准理解一致后, 共同观察并独立对所有样本进行1~4打分记录. 排除无法判断表型类别的样本, 以3个评分者对每个样本表型分类打分的众数作为最终的脱发等级.

1.3 基因分型及质量控制

按照生产商的推荐, 使用Illumina Infinium Global Screening Array 650K (GSA) 型号芯片 (Illumina, 圣地亚哥, 美国) 对样本DNA进行分型. 使用PLINK v1.9^[28]对SNPs进行质量控制, 包括分型成功率 (call rate) >0.9, 哈迪温伯格平衡 (Hardy-Weinberg equilibrium, HWE) $P>0.0001$ 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 IMPUTE2^[29-30] 分别对常染色体和 X 染色体进行基因填补并过滤填补质量分数小于 0.6 的 SNPs, 并再次重复上面质量控制标准, 最终共获得 20 681 872 个 SNPs.

1.4 统计分析

1.4.1 遗传关联分析与GWAS

本研究选用 Visscher 实验室公布的 624 个 SNPs 位点, 其中 467 个位点通过了质量控制 (附件表 S1). 使用一般线性回归模型 (general linear model, GLM) 进行 467 个 SNPs 与男性型脱发表型的关联分析. 基因型的赋值按照加性模型, 该模型假设个体携带的次等位基因数量与表型特征有累积效应. 建立模型时, 首先将表型与年龄和 BMI 建立线性回归, 再将获得的残差与 SNPs 做线性回归. 研究年龄和 BMI 对表型的影响, 单独做线性回归. 模型如下:

$$Y = \beta_{AGE} AGE + \beta_{BMI} BMI + \varepsilon_Y \quad (1)$$

$$\varepsilon_Y = \beta_{ADD} G + e \quad (2)$$

其中 Y 表示男性型脱发表型分类向量, β_{ADD} 表示基因型对表型的效应向量, β_{BMI} 表示 BMI 的效应向量, β_{AGE} 表示年龄效应向量, ε_Y 表示线性回归的残差向量, e 残差效应向量, $e \sim N(0, I\sigma_e^2)$, I 表示单位阵, σ_e^2 表示随机残差方差. 基因型数据表示模式:

$$\begin{cases} G_i = 0, \text{ SNP位点含有0个次等位基因} \\ G_i = 1, \text{ SNP位点含有1个次等位基因} \\ G_i = 2, \text{ SNP位点含有2个次等位基因} \end{cases} \quad (3)$$

关联分析 $P < 0.05$ 认为具有统计学意义. 此外, 在全基因组水平, 使用 PLINK v1.9 进行 GWAS 分析. 其中, 协变量增加了基因分型数据的前 3 个主成分 (principal components, PCs), $P < 5 \times 10^{-8}$ 认为在全基因组水平具有统计学意义. GWAS 分析结果的膨胀因子 λ 为 1.11, 因此我们使用基因组控制 (genomic control) 对 P 值进行了校正. 使用 R 包 qqman^[31] 绘制曼哈顿图 (manhattan plot) 和 Q-Q 图 (quantile-quantile plot).

1.4.2 人群异质性分析

本研究统计了受试人群 467 个效应等位基因的频率分布、大小和方向, 并与 UK Biobank 人群、千人基因组中的欧洲人群 (EUR) 及东亚人群 (EAS) 中的这些 SNPs 的频率分布进行相关性比较和配对 t 检验. 将效应等位基因的大小 (β) 和频率

(f) 的乘积作为一个新的统计量 ($\beta \times f$), 在 UK Biobank 人群和本研究人群中通过四象限图对人群异质性进一步比较.

1.4.3 预测分析

PRS 是用来评估个体患某种疾病风险的方法, 主要将被纳入分析中多个效应等位基因的效应值与其基因分型的乘积之和来计算, 计算公式如下:

$$PRS = \sum_{i=1}^n \beta_i G_i \quad (4)$$

其中, G_i 是个体 SNP 的基因型 (0, 1, 2), β_i 是关联分析中 SNP 的效应值, 连续性变量的表型为 beta 值, 分类变量的表型为 OR 值, n 为纳入的 SNP 数量.

基于 PRS、年龄、BMI 我们分别构建了线性回归模型和逻辑回归模型, 在构建逻辑回归模型时, 将 MPB 的前两类作为非脱发对照组, 后两类作为严重脱发病例组, 模型如下:

$$MPB \sim PRS + AGE + BMI \quad (5)$$

$$\ln\left(\frac{P}{1-P}\right) \sim PRS + AGE + BMI \quad (6)$$

其中, P 为严重脱发的概率, 其他符号与之前相同.

在构建模型的过程中, 随机选取整个样本的 80% 作为训练集, 20% 作为预测集, 且该过程重复进行 1 000 次, 分别使用线性回归的决定系数 R^2 和逻辑回归中的 AUC 值 (R 包 ROCR^[32]) 作为模型性能的评估标准. 其中, AUC 值变化范围为 0.5~1, 其值越接近 1 表明预测的准确性越高. 随后, 我们按照留一法交叉验证 (leave one out cross validation, LOOCV) 的方法, 将年龄、BMI 和 SNPs 的 PRS 作为自变量, 训练 MPB 二分类逻辑回归模型, 预测 684 个样本的脱发程度.

1.4.4 SNP 的数量与预测准确性关系

为了进一步研究 SNP 数量对模型预测效果的影响, 在年龄和 BMI 的基础上, 对通过质控的 467 个 SNPs 每次累计添加一个 SNP 位点构建 MPB 二分类逻辑回归模型. SNPs 位点添加的顺序由 MPB 与年龄、BMI 和 SNPs 逐步向前回归中的赤池信息量 (akaike information criterion, AIC) 确定, 每次计算中 AIC 值最小的 SNP 优先加入模型, AIC 的近似计算公式如下:

$$AIC \sim 2k + n \ln\left(\sum_{i=1}^n (y_i - E(y_i))^2\right) \quad (7)$$

其中, k 为 SNP 的数量, n 为样本量的大小, $E(y_i)$ 为拟合值.

建模时, 随机选取 80% 的样本作为训练集, 剩下 20% 的样本作为测试集, 重复 100 次, 使用平均 AUC 值作为模型预测性能评价标准. 将预测概率大于 0.5 作为脱发类型, 获得典型的预测样本类别和实际样本类别的 2×2 混淆矩阵, 并分别计算敏感性、特异性、阳性预测率、阴性预测率和准确率平均值. 以上分析过程, 除特别说明外, 均在 R 3.6.1 中完成.

1.4.5 千人基因组中PRS及MPB分布

由于 UK Biobank 研究中报道的 624 个 SNPs 在千人基因组中缺少 2 个, 因此使用 622 个 SNPs 对其 5 个人群的 1 233 名男性个体计算 PRS, 观察 PRS 在人群中的分布. 以所有人群 PRS 的均值作为参照将其分为二类, 并使用 Map Viewer 7 软件展示了预估 MPB 二分类在全球 5 大人群中的地理分布图.

2 结果与分析

2.1 人群样本信息的统计概况

本研究总共包括 684 例中国亚欧混合人群样本, 年龄范围为 16~59 岁, 平均年龄 35.6 岁, BMI 范围为 16.1~37.6 kg/m², 均值为 24.2 kg/m². MPB 按脱发的严重程度分为 4 类, 分别为 53.1%、30.3%、5.4% 和 11.2%, 脱发等级较低的个体倾向于比等级高的个体更年轻 (图 1). MPB 各等级 BMI 均值分别为 23.4、24.7、25.5、25.8 kg/m², 表明 MPB 等级越高可能伴随着较高的 BMI. 线性回归显示年龄 ($P=2\times 10^{-16}$) 与 BMI ($P=4.98\times 10^{-9}$) 均和脱发严重程度呈显著正相关, 这与之前的报道吻合 [9, 25].

2.2 人群异质性分析

UK Biobank 研究中报道的 624 个独立 SNPs 位点, 在本研究中共有 467 个通过质控. 这些 SNPs 有 6.9% (32 个) 展示了与 MPB 程度上的相关性 ($P<0.05$), 其中 78.1% (25 个) 与 UK Biobank 研究中效应等位基因的方向一致 (表 1, 附件表 S1).

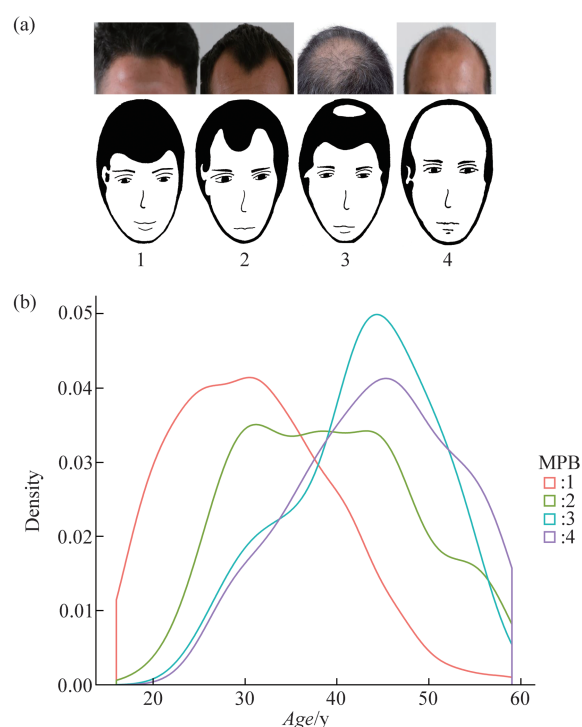


Fig. 1 The summary of MPB in Eurasian population

(a) Diagram of MPB and corresponding scores adapted from the Hamilton-Norwood scale. MPB scores 1~4 respectively correspond to different hair loss grades, including no balding, crown balding, vertex balding and the combination of crown and vertex balding, and also indicate severity. (b) Density plots of the relative age distributions in four categories. Abbreviations: MPB=male pattern baldness.

GWAS 中未发现在全基因组水平上显著关联的位点 ($P<5\times 10^{-8}$) (附件图 S1). UK Biobank、千人基因组的 EUR 和 EAS、本研究的 Eurasian, 这 4 个不同族群的 467 个 SNPs 的异质性分析表明, 除 UK Biobank 与 EUR 的效应等位基因频率的相关系数最高 ($r=0.99$), 无显著性差异 ($P>0.8$) 外, 其他人群之间均存在明显不同的遗传结构 ($1\times 10^{-14}<P<1\times 10^{-6}$) (图 2a). 将效应等位基因的大小 (β) 和频率 (f) 的乘积作为一个新的统计量 ($\beta\times f$), 四

Table 1 The comparison of SNP effect between the Eurasian and the UK Biobank study

Threshold (P)	N (SNP)	Proportion/%	Consistent/%	$r(\beta)$	$r(f)$	$r(\beta\times f)$
1	467	100	54.8	0.43	0.80	0.38
0.05	32	6.9	78.1	0.77	0.77	0.56

Consistent represents the proportion of effect alleles with the same direction in the UK Biobank study. $r(\beta)$ represents the correlation coefficient of the MPB regression β values between UK Biobank and the Eurasian population under corresponding P value. $r(f)$ represents the correlation coefficient of the effect allele frequencies between UK Biobank and the Eurasian population under corresponding P value.

象限图显示: 第一、三象限效应等位基因方向一致的数目 (256 个) 与第二、四象限效应等位基因方向不一致的数目 (211 个), 在统计学上无明显差异 ($P=0.16>0.05$) (图 2b)。以上结果表明人群异质性

对 MPB 遗传结构的解释具有重要作用。这些包含高度遗传信息的 SNPs, 可以为脱发程度的预测提供一定的预测信息。

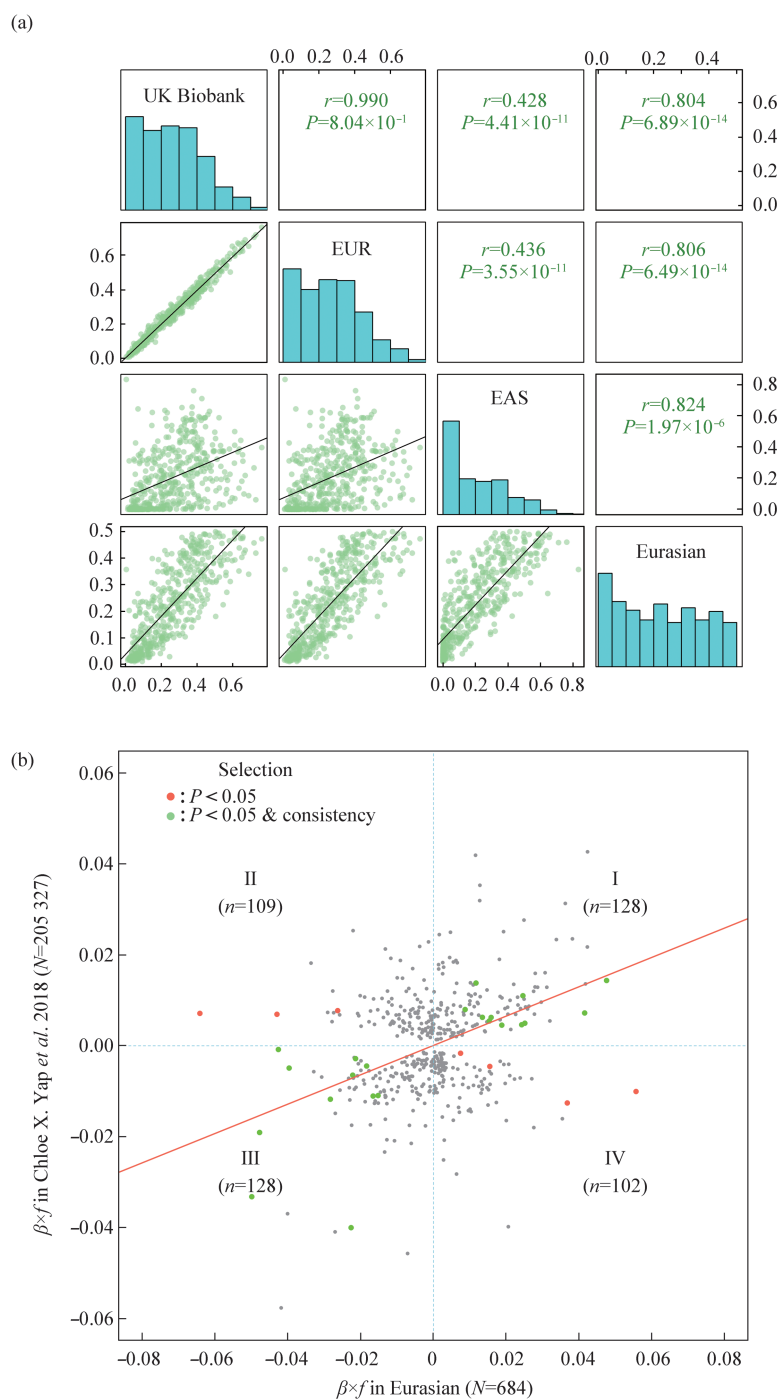


Fig. 2 Investigation of population heterogeneity

(a) The scatter plot of the effect allele frequency for 467 MPB-associated SNPs in the UK Biobank, EUR (from phase 3 of the 1000 Genomes), EAS (from phase 3 of the 1000 Genomes) and the Eurasian population. Diagonal is the frequency distribution of the corresponding population. Pearson correlation coefficients (r) were estimated between each pair of the populations. P values were derived from paired t tests. (b) The scatter plot of $\beta \times f$ comparing the Eurasian and the UK Biobank study. Orange line is the regression line. The orange dots represent alleles with P value less than 0.05. The green dots represent alleles with P value less than 0.05 and the effect allele in the same direction.

2.3 MPB的预测分析

对MPB的预测分析,主要使用年龄、BMI和PRS构建预测模型.由于我们的样本量有限,在计算PRS时,效应值采用UK Biobank中已经报道的效应等位基因的大小(β).几种建模方法比较发现,加入年龄和BMI均比未加入二者的 R^2 和AUC高4倍左右(表2).MPB的4等级分类与年龄、BMI和467个SNPs的PRS线性回归中,交叉验证的平均 R^2 为31.1%.其中,年龄和BMI可解释表型方差分别为22.9%和4.9%,而SNP解释表型方差平均为0.2%,其中最大的是位于EDA2R/AR区域的X:66358695:T:TA(约为2.7%);相应的MPB的二分类逻辑回归预测的平均AUC值为0.82.我们进一步探索了SNPs的子集对脱发预测效能的影响,根据关联分析中效应等位基因 P 值的阈

值和与UK Biobank中报道的 β 方向是否一致,选取不同SNPs的子集,利用PRS分别建立了线性回归模型和逻辑回归模型.随着 P 值减小, β 方向条件的控制,虽然SNPs位点的数量减少,但模型预测准确性指标 R^2 及AUC并无明显下降.其中,选取 P 值小于0.05且与UK Biobank研究中 β 方向一致的25个SNPs时,交叉验证的平均 R^2 为28.97%(CI(95%)=18.9%~39.4%),平均AUC值为0.82(CI(95%)=0.74~0.89)(表2).同理,采用留一法进行预测时,交叉验证的平均 R^2 为28.6%,AUC值为0.82.尽管这些AUC值没有达到理想的临床检验水平(0.85),但该模型预测严重脱发概率小于等于0.2和大于等于0.8的部分个体有较高的预测准确性(图3).

Table 2 The accuracy for predicting MPB in Eurasian

Method	N (SNP)	Age and BMI	R^2 / %	CI (95%)	AUC	CI (95%)
All	467	0	8.33	2.05–17.04	0.68	0.57–0.79
		1	31.09	20.83–42.07	0.82	0.74–0.89
$P < 0.05$	32	0	4.95	1.23–9.74	0.65	0.54–0.76
		1	28.50	18.69–38.83	0.82	0.74–0.88
$P < 0.05$ and consistent β direction	25	0	5.85	1.57–11.32	0.67	0.54–0.79
		1	28.97	18.89–39.42	0.82	0.74–0.89

The 0 and 1 in the third column represent age and BMI were exclude or included in the model. CI: Confidence interval, AUC: Area under curve.

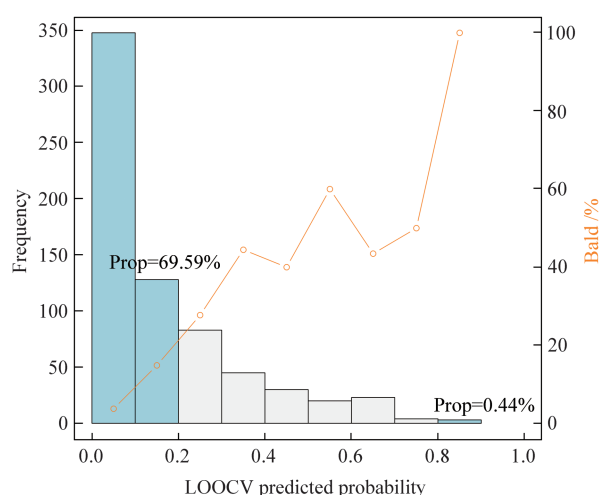


Fig. 3 Histogram of predicted probability overlaid with percentage of baldness in each probability bin

The blue area represents the percentage of individuals in the corresponding predicted probability interval (<0.2 or >0.8). Bald represents percentage of individuals with hair loss under the corresponding predicted probability bin. Prop: Proportion, LOOCV: Leave-one-out-cross-validated.

2.4 预测因子数量与预测准确性关系

我们探索了年龄、BMI和SNPs三个因子对预测脱发准确性的影响.仅使用年龄进行预测,AUC值可以达到约0.77,BMI的加入AUC值几乎无变化(0.77).随着467个通过质控的SNPs按照顺序依次添加至模型中(具体顺序参照方法1.4.4),累积AUC曲线呈上升趋势,添加68个SNPs后达到最大AUC值(约0.89),敏感度(0.59)小于特异度(0.92),阳性预测率(0.61)小于阴性预测率(0.92),整体的预测准确性为0.88(表3),模型对

Table 3 The accuracy for predicting binary MPB using different number of predictors

N (SNP)	AUC	CI (95%)	Sensitivity	Specificity	PPV	NPV	ACC
Age and BMI	0.77	0.69–0.85	0.12	0.96	0.40	0.84	0.82
68	0.89	0.79–0.94	0.59	0.92	0.61	0.92	0.87
464	0.57	0.42–0.71	0.45	0.66	0.20	0.86	0.62

非严重脱发样本的预测较好. 继续添加 SNP, *AUC* 值总体呈现下降趋势, 当添加 464 个 SNPs 时 *AUC* 达到最小值, 约为 0.57, 其他统计量均有所下降 (图 4, 表 3). 所有预测因子的详细结果见附件表 S2.

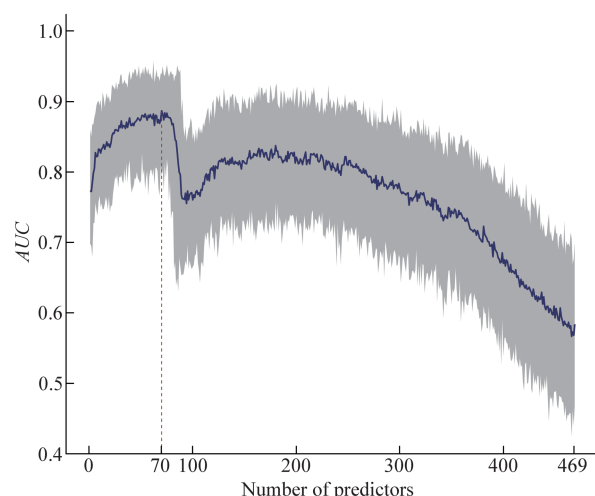


Fig. 4 Relationship between number of predictors and *AUC* for binary MPB degree prediction

Age and BMI are the first two predictors. The gray shadow is CI(95%) under corresponding *AUC*.

2.5 MPB在不同群体中的估测分布

随后, 我们通过计算千人基因组数据库中的 5 个洲际人群的 PRS, 观察 MPB 可能的分布状况. 在 5 个人群的 1 233 名男性 PRS 分布中, 平均分数最高的是南亚人群, 欧洲人群平均 PRS 高于非洲、美洲人和东亚人群 (图 5a). MPB 的二分类在全球 5 大人群的分布也观察到了类似分布结果 (图 5b).

3 讨 论

男性型脱发是男性中常见的一种复杂表型. 其受遗传、激素、年龄、环境等多方面因素的影响. 尽管 MPB 的发病率较高, 但当前对 MPB 的遗传机制尚未完全解析. 因而, 研究比较不同人群之间脱发的遗传异质性、从 DNA 预测患者有无脱发风险等有助于进一步了解 MPB 遗传机制, 有助于患者的早期诊断及预防, 也可与法庭科学研究中的 DNA 表型特征刻画相关研究成果结合, 如虹膜颜色^[33]、头发形态的曲直^[34]和颜色^[35]、年龄的推断^[36]、成人的身高预测^[37]等. 因此, 本研究基于当前男性型脱发研究中最大 GWAS 分析鉴定独立相关的 624 个 SNPs 位点, 在 Eurasian 中进行人群异

质性分析, 构建了 MPB 的预测模型, 对预测因子数量与预测模型性能的关系进行研究, 并通过模型预测几大人群的脱发程度, 以观察 MPB 在全球可能的地理分布状况.

Eurasian 是典型的亚欧混合人群, 其东亚和欧洲祖先成分约 50 : 50 的独特遗传组成^[27], 使之成为研究欧洲人群和东亚人群中人类复杂性状遗传结构的独特资源. 例如, 关于中国汉族和 Eurasian 的 GWAS 研究表明: EDAR 是影响东亚人头发曲直的决定性基因, 而在 Eurasian 中 EDAR 的效应相对较小, 和欧洲人特异的毛透明蛋白基因 (trichohyalin, TCHH) 等基因共同影响头发曲直^[38]. 在 MPB 相关效应等位基因的频率分布中, Eurasian 与 UK Biobank 的相关系数很高, 与东亚人群的却很低. 除此之外, 效应等位基因的大小和方向也存在较大的差异. 这些都表明人群异质性的对 MPB 潜在遗传机制有重要作用.

本研究对 684 例 Eurasian 样本的 MPB 与基因型进行了 GWAS, 但并没有发现显著关联的位点 ($P < 5 \times 10^{-8}$), 这可能与有限的样本量有关, 也可能是 MPB 表型分布偏倚、脱发等级低以及个体平均年龄较小的缘故. X 染色体 Xq12 上 EDAR/AR 是首先被报道与 MPB 最显著相关的位点^[17, 39], 它在毛发和牙齿的生长过程中具有重要作用, 对早发性脱发有较大的影响^[40]. 在本研究的关联分析中, 位点 23 : 66358695 : T : TA 在基因组水平上有最显著的 *P* 值 (1.75×10^{-5}) 且有最大的效应值, 验证了 EDAR/AR 是与 MPB 显著相关的区域. MPB 严重程度与高 BMI 有较强的相关性, 高 BMI 往往伴随着严重的脱发^[9], 体重大或者个子矮的男生 MPB 风险高^[22], 这与本研究中不同 MPB 等级的平均 BMI 趋势相一致.

当前关于 MPB 程度预测的研究并不多, 已有模型对早发性脱发预测效果较好, 但这些预测模型都是基于欧美人群, 由于遗传异质性, 应用时要注意适应的人群及年龄范围. Liu 等^[16]基于 20-SNP 在 3 个欧洲群体中分别构建多元逻辑回归模型, 除早发性群体外, 另两个群体的 MPB 各分类的平均年龄都在 44 岁以上. 早发性脱发群体的预测准确性 *AUC* 值为 0.74, 另两个群体构建的模型中增加了年龄, *AUC* 值分别为 0.69、0.71. 其发现的新位点 rs4959410 在我们的研究中也得到了验证 ($P < 0.05$). Hagenaars 等^[23]基于 UK Biobank 40~69 岁 4 000 名男性个体构建的 287-SNP 预测模型, 其预

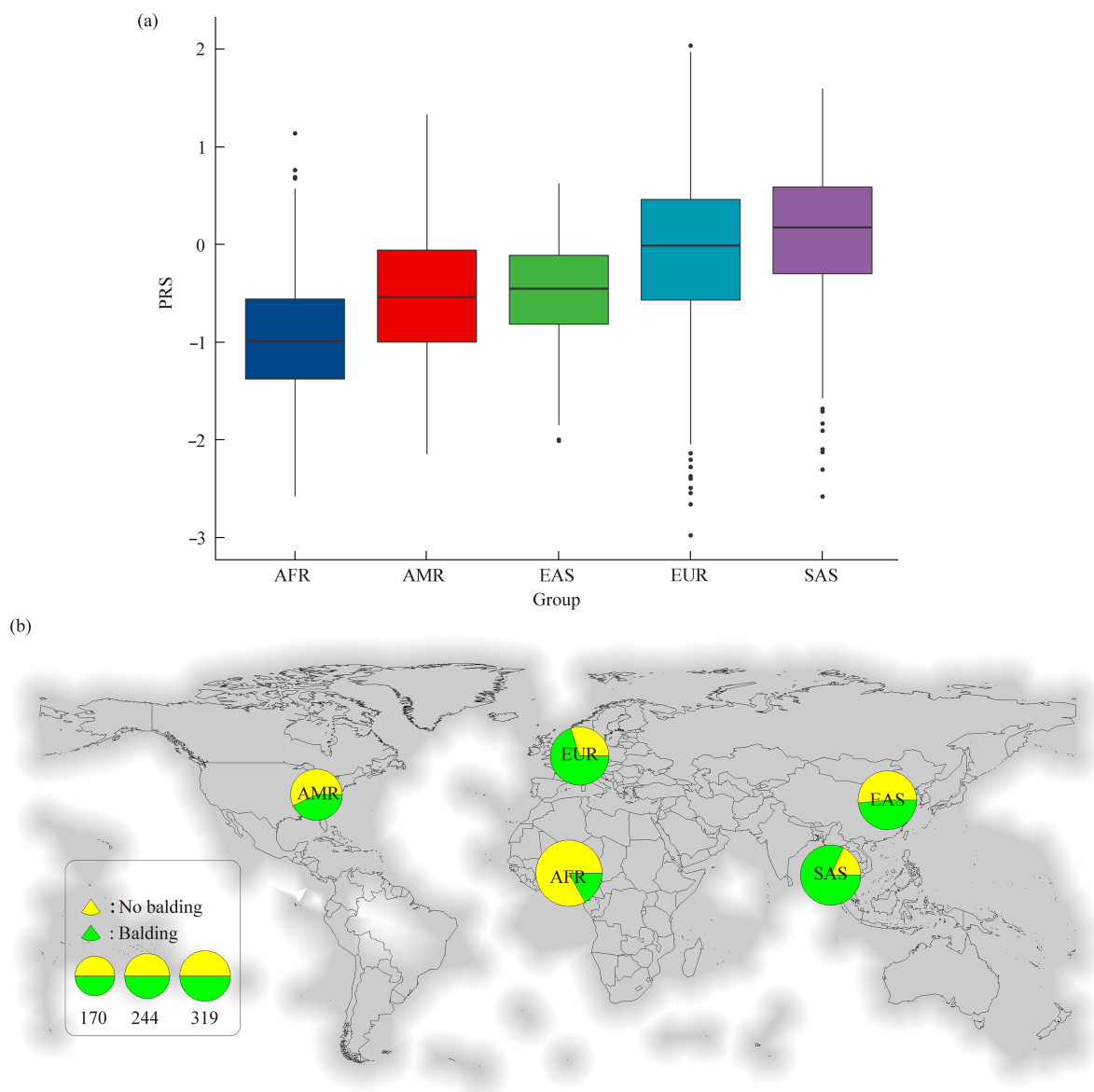


Fig. 5 The distribution of polygenic risk scores of males from the phase 3 of the 1 000 Genomes Project

(a) The PRS distribution of males grouped into five large populations including sub-Saharan Africans (AFR), Native Americans (AMR), East Asians (EAS) and central Europeans (EUR). (b) The geographic distribution of predicted above average PRS in five large populations. PRS: Polygenic risk scores.

测二分类严重脱发 AUC 值可达到 0.78, 模型中增加年龄时 AUC 为 0.79. 我国相关研究主要集中在 MPB 的致病机制、因素及治疗^[41-42], 而相关遗传位点的发掘及预测模型的研究鲜有报道.

本研究样本的年龄跨度相对较大、平均年龄较小, 年龄的回归参数约 0.05 ($P < 2 \times 10^{-16}$) 表明, 年龄会使 MPB 等级每年约增长 0.05, 即使将年龄的平方加入到模型中, 也不能对表型方差解释更多的变异, 这表明随年龄的增长脱发将变成男性群体一

种常态化表型. 同时年龄对 MPB 预测也有重要作用, 如线性回归中, 年龄对 MPB 的解释程度和预测模型中 AUC 值均远大于 BMI 和相关 SNPs. 仅使用 SNPs 构建的预测模型, 预测效果与 Liu 和 Hagenaaers 的预测模型相比, 相对较弱. 而结合年龄、BMI 和 SNPs 子集的 PRS 构建的预测模型, 与前两者相比, AUC 值略有提高. 本研究还发现, 增加 MPB 相关 SNPs 的数量虽然可以在一定程度上提高模型预测的准确性, 但两者并非线性关系, 当

SNPs达到一定数量时,继续增加数量,ROC曲线出现低谷区域及下降趋势,模型预测性降低,这可能是模型参数过多而样本量有限导致模型拟合不佳所致.而且在法庭科学、医院检验等实际应用场景中,DNA样本量有限,预测因子数量过多也不易实现快速的复合检测.总体来说,我们的模型在年龄、BMI的基础上使用SNP子集的PRS实现了与全集相近的预测准确性,但因预测准确性与年龄有较强相关性,应慎重使用,而且可能因人群差异产生进一步的变化.模型对70%的个体预测效果非常准确,但多数是非脱发个体,这可能是样本中非脱发样本数量多于脱发样本数量的原因,后期通过加大样本量,有望提升对严重脱发个体的预测效果.

本研究模型对千人数据库中5大洲际群体PRS的预测分布结果与当前流行病学对MPB的相关研究及认识大体一致:即非洲人群的发病率最低,欧洲人群发病率较东亚、美洲土著、非洲等人群高^[2, 5].但南亚人群MPB的预测发病率最高与我们的先验知识不相符,这可能是由于模型基于欧源样本的参数建立,应用在全球人群时存在预估偏倚.全球人群真实分布结果还需进一步开展相关研究进行验证.

人类复杂性状种类多样.从当前研究较多的表型来看,各表型的预测准确性差别较大,这与表型是否存在主效基因和是否受非遗传因素影响有关.比如,虹膜颜色有一个很强的主效基因(HER2)^[43],使用几个相关SNPs就可以较准确地预测蓝或棕色虹膜,AUC值达到0.9以上^[33, 44].类似的,预测白、黑皮肤颜色模型的AUC值分别为0.99和0.96^[45].然而,目前最大的25万样本的身高相关GWAS研究也未能发现身高的主效基因,上千个身高相关SNPs也只能解释身高表型方差不到30%^[46].对MPB来说,尽管已发现了主效基因,但当前报道的最大效应值(β)(1.2)^[24]与虹膜颜色(2.8)^[47]相比,相对较弱.同时,MPB表型与年龄相关性较强,会随着年龄增长明显变化,较难确定一个个体的准确脱发等级.这些都为MPB的遗传研究和准确预测带来了困难.近年来,我国启动了十万人的基因组计划,将绘制国人精细的基因组图谱,有助于发现更多与MPB相关的位点,有望在将来使MPB预测模型性能有明显提升,从而应用于医疗等领域,为患者早期诊断和干预提供帮助.

附件 20200095_图 S1.tif、20200095_表 S1.xlsx 和 20200095_表 S2.xlsx 见本文网络版 (<http://www.pibb.ac.cn> 或 <http://www.cnki.net>)

参 考 文 献

- [1] Gupta M, Mysore V. Classifications of patterned hair loss: a review. *J Cutan Aesthet Surg*, 2016, **9**(1): 3-12
- [2] Lolli F, Pallotti F, Rossi A, *et al*. Androgenetic alopecia: a review. *Endocrine*, 2017, **57**(1): 9-17
- [3] Norwood O T. Male pattern baldness: classification and incidence. *Southern Medical Journal*, 1975, **68**(11): 1359
- [4] Hamilton J B. Patterned loss of hair in man; types and incidence. *Annals of the New York Academy of Sciences*, 2010, **53**(3): 708-728
- [5] Otberg N, Finner A M, Shapiro J. Androgenetic alopecia. *Endocrinol Metab Clin North Am*, 2007, **36**(2): 379-398
- [6] Gold M H. Fractionated aminolevulinic acid-photodynamic therapy (PDT) provides additional evidence for the use of PDT for non-melanoma skin cancer. *J Eur Acad Dermatol Venereol*, 2009, **23**(5): 571-572
- [7] Cash T F. The psychosocial consequences of androgenetic alopecia: a review of the research literature. *Br J Dermatol*, 1999, **141**(3): 398-405
- [8] Alfonso M, Richter-Appelt H, Tosti A, *et al*. The psychosocial impact of hair loss among men: a multinational European study. *Curr Med Res Opin*, 2005, **21**(11): 1829-1836
- [9] Yang C C, Hsieh F N, Lin L Y, *et al*. Higher body mass index is associated with greater severity of alopecia in men with male-pattern androgenetic alopecia in Taiwan: a cross-sectional study. *J Am Acad Dermatol*, 2014, **70**(2): 297-302 e291
- [10] 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
- [11] Zhou C K, Levine P H, Cleary S D, *et al*. Male pattern baldness in relation to prostate cancer-specific mortality: a prospective analysis in the NHANES I Epidemiologic Follow-up Study. *Am J Epidemiol*, 2016, **183**(3): 210-217
- [12] Trieu N, Eslick G D. Alopecia and its association with coronary heart disease and cardiovascular risk factors: a meta-analysis. *Int J Cardiol*, 2014, **176**(3): 687-695
- [13] Matilainen V, Koskela P, Keinänen-Kiukaanniemi S. Early androgenetic alopecia as a marker of insulin resistance. *The Lancet*, 2000, **356**(9236): 1165-1166
- [14] Küster W, Happle R. The inheritance of common baldness: Two B or not two B?. *Journal of the American Academy of Dermatology*, 1984, **11**(5): 921-926
- [15] Nyholt D R, Gillespie N A, Heath A C, *et al*. Genetic basis of male pattern baldness. *Journal of Investigative Dermatology*, 2003, **121**(6): 1561
- [16] 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
- [17] Hillmer A M, Hanneken S, Ritzmann S, *et al.* Genetic variation in the human androgen receptor gene is the major determinant of common early-onset androgenetic alopecia. *Am J Hum Genet*, 2005, **77**(1): 140-148
- [18] 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
- [19] Richards J B, Yuan X, Geller F, *et al.* Male-pattern baldness susceptibility locus at 20p11. *Nat Genet*, 2008, **40**(11): 1282-1284
- [20] 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
- [21] 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
- [22] 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**: 14694
- [23] Hageaars S P, Hill W D, Harris S E, *et al.* Genetic prediction of male pattern baldness. *Plos Genet*, 2017, **13**(2): e1006594
- [24] 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
- [25] 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
- [26] Prokunina-Olsson L, Marcińska M, Pośpiech E, *et al.* Evaluation of DNA variants associated with androgenetic alopecia and their potential to predict male pattern baldness. *Plos One*, 2015, **10**(5): e0127852
- [27] Xu S, Huang W, Qian J, *et al.* Analysis of genomic admixture in Uyghur and its implication in mapping strategy. *Am J Hum Genet*, 2008, **82**(4): 883-894
- [28] 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
- [29] Howie B N, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *Plos Genet*, 2009, **5**(6): e1000529
- [30] Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3 (Bethesda)*, 2011, **1**(6): 457-470
- [31] Turner S D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *Biorxiv*, 2014(DOI: <https://doi.org/10.1101/005165>)
- [32] Sing T, Sander O, Beerenwinkel N, *et al.* ROCr: visualizing classifier performance in R. *Bioinformatics*, 2005, **21**(20): 3940-3941
- [33] Walsh S, Liu F, Ballantyne K N, *et al.* IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci Int Genet*, 2011, **5**(3): 170-180
- [34] Liu F, Chen Y, Zhu G, *et al.* Meta-analysis of genome-wide association studies identifies 8 novel loci involved in shape variation of human head hair. *Hum Mol Genet*, 2018, **27**(3): 559-575
- [35] Chaitanya L, Breslin K, Zuñiga S, *et al.* The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. *Forensic Science International: Genetics*, 2018, **35**: 123-135
- [36] Zbiec-Piekarska R, Spolnicka M, Kupiec T, *et al.* Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. *Forensic Sci Int Genet*, 2015, **14**: 161-167
- [37] Liu F, Zhong K, Jing X, *et al.* Update on the predictability of tall stature from DNA markers in Europeans. *Forensic Science International: Genetics*, 2019, **42**: 8-13
- [38] Wu S, Tan J, Yang Y, *et al.* Genome-wide scans reveal variants at EDAR predominantly affecting hair straightness in Han Chinese and Uyghur populations. *Hum Genet*, 2016, **135**(11): 1279-1286
- [39] Prodi D A, Pirastu N, Maninchedda G, *et al.* EDAR is associated with androgenetic alopecia. *J Invest Dermatol*, 2008, **128**(9): 2268-2270
- [40] Bubulya A, Wise S C, Shen X Q, *et al.* c-Jun can mediate androgen receptor-induced transactivation. *Journal of Biological Chemistry*, 1996, **271**(40): 24583-24589
- [41] 叶萌, 辛李宁, 程宝文, 等. 雄激素性脱发相关研究进展. *国际遗传学杂志*, 2015, **38**(3): 174-178
Ye M, Xin L N, Cheng B W, *et al.* *International Journal of Genetics*, 2015, **38**(3): 174-178
- [42] 李苏瑶, 高慧. 雄激素性秃发治疗的最新研究进展. *临床皮肤科杂志*, 2019, **48**(04): 66-70
Li S Y, Gao H. *Journal of Clinical Dermatology*, 2019, **48**(04): 66-70
- [43] Visser M, Kayser M, Palstra R J. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Res*, 2012, **22**(3): 446-455
- [44] Liu F, Van Duijn K, Vingerling J R, *et al.* Eye color and the prediction of complex phenotypes from genotypes. *Curr Biol*, 2009, **19**(5): R192-193
- [45] Maronas O, Phillips C, Sochtig J, *et al.* Development of a forensic skin colour predictive test. *Forensic Sci Int Genet*, 2014, **13**: 34-44
- [46] Wood A R, Esko T, Yang J, *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*, 2014, **46**(11): 1173-1186
- [47] Zhang M, Song F, Liang L, *et al.* Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. *Hum Mol Genet*, 2013, **22**(14): 2948-2959

Predicting Male Pattern Baldness From DNA Variants in a Eurasian Population*

PAN Si-Yu^{1)**}, ZHAO Wen-Ting^{2)**}, FENG Rui³⁾, LI Qiong³⁾, JING Xiao-Xi¹⁾, GAO Xing-Jian¹⁾,
CHEN Yan¹⁾, LI Yi¹⁾, TAO Xian-Ming¹⁾, LIU Tian-Zi^{1)***}, LI Cai-Xia^{2)***}

⁽¹⁾CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics (China National Center for Bioinformation),

Chinese Academy of Science, Life Sciences College, University of Chinese Academy of Sciences, Beijing 100101, China;

⁽²⁾National Engineering Laboratory for Forensic Science, Key Laboratory of Forensic Genetics of Ministry of Public Security, Beijing Engineering Research
Center of Crime Scene Evidence Examination, Institute of Forensic Science, Ministry of Public Security, Beijing 100038, China;

⁽³⁾Criminal Investigation Department, Public Security Department of Hunan Province, Changsha 410001, China)

Abstract Male pattern baldness is the most common type of progressive hair loss, which has a considerable negative impact on personal image and mental health. Recently, several large sized genome-wide association studies (GWASs) in European and American populations have reported a large number of MPB-associated genetic susceptibility sites, but the genetic effects of these loci in East Asian populations remain unclear. This study was based on a sample of 684 Eurasian males in China, and conducted a population heterogeneity analysis on 624 MPB-related single nucleotide polymorphisms (SNPs) found in a large GWAS consisting of 205 327 European males in the UK Biobank. Polygenic risk scores (PRS) were used to build predictive models, and the relationship between the number of predictive factors and the predictive performance of the model was investigated. A total of 467 SNPs passed quality control, among which 6.9% showed nominally significant association with MPB ($P < 0.05$). Linear and logistic models established by including the effects of age, body mass index (BMI), and 25 SNPs demonstrated fairly accurate prediction accuracy ($R^2 = 28.9\%$, $AUC = 0.82$). Age had the greatest impact on the model performance ($R^2 = 22.9\%$, $AUC = 0.77$), and the AUC reached the maximum when BMI and 68 SNPs were added cumulatively (about 0.89). This study indicates that MPB has strong genetic heterogeneity in European and East Asian populations. Selecting a subset of SNPs can achieve prediction accuracy close to that of the complete set. The prediction model constructed here may help understand the genetic mechanism, early diagnosis and prevention of MPB in East Asian populations.

Key words male pattern baldness, hair loss prediction model, population heterogeneity, Eurasian population

DOI: 10.16476/j.pibb.2020.0095

* This work was supported by grants from National Key R&D Program of China (2017YFC0803501), Open Projects of National Engineering Laboratory for Forensic Science (2017NELKFKT05), The National Natural Science Foundation of China (91651507), Central Public-Interest Scientific Institution Basal Research Fund (2018JB046), The Open Project of Key Laboratory of Genomic and Precision Medicine, Chinese Academy of Sciences, National High-level personnel of special support program (SQ2018RA23120013) and the National Science and Technological Resources Platform (YCZYPT [2017]01-3 and 2017JB025).

** These authors contributed equally to this work.

*** Corresponding author.

LIU Tian-Zi. Tel: 86-10-84097824, E-mail: liutz@big.ac.cn

LI Cai-Xia. Tel: 86-10-63264166, E-mail: licaixia@tsinghua.org.cn

Received: April 11, 2020 Accepted: July 7, 2020