



中国南北方汉族人群DNA甲基化表观遗传差异研究*

孙昌春^{1,2)**} 许继臣^{2)**} 江丽²⁾ 郭晓媛^{1,2)} 赵雯婷²⁾ 叶健^{2)***} 李彩霞^{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, 法医遗传学公安部重点实验室, 现场物证溯源技术国家工程实验室, 北京 100038)

摘要 目的 族群地域、地貌特征等表型是基因型与环境共同作用的结果。大量基因组研究表明, 汉族人群具有混合特征, 内部存在明显的南北遗传差异。本研究旨在探索研究表观基因组在中国南北方汉族人群之间是否存在差异, 并筛选差异遗传位点。方法 使用GLINT软件对483份汉族样本的全基因组甲基化芯片数据进行EWAS分析, 使用Lasso回归方法筛选位点。使用多元逻辑回归算法构建南北方汉族人群预测模型, 通过十折交叉验证的方法评估。结果 筛选出一组南北方汉族之间差异显著的CpG位点, 准确性为99.03%, Kappa系数为0.979 6。结论 本研究表明南北方汉族人群之间存在表观遗传差异, 本研究为进一步开展不同地域汉族人群之间的表观遗传差异研究奠定了基础。

关键词 法医遗传学, 表观遗传学, DNA甲基化, 表观基因组关联研究, 汉族人群

中图分类号 R89, D919

DOI: 10.16476/j.pibb.2021.0091

饮食、气候等环境因素可能会影响人的表观遗传学结构^[1-3]。DNA甲基化是一种重要的表观遗传标记, 具有遗传稳定、含量丰富、随龄变化等特点^[4]。伴随高通量的DNA甲基化数据的不断涌现, 大量与肿瘤等疾病发生相关的DNA甲基化位点被发现^[5]。在法医学领域, DNA甲基化已被用于年龄推断^[6-9]、组织属性判别^[10]、同卵双胞胎的鉴别^[11-12]等。基于欧美等国外人群的研究证明DNA甲基化在族群地域间存在差异^[13-14]。

汉族是世界上人口最多的民族。研究表明, 汉族人群具有混合特征^[15], 呈现明显的南北分化, 在基于STR^[16]、线粒体DNA (mtDNA)^[17-18]、Y染色体SNP^[15]、常染色体SNP^[19]等遗传标记的研究均已表明汉族内部存在明显的南北遗传差异。北方人群由于受到来自中亚和欧洲遗传成分的影响, 呈现东西走向的变化趋势, 南北方人群遗传的差异以秦岭淮河到长江为地理分界^[20]。汉族人群与当地少数民族之间的遗传差异小于南北方汉族之间的遗传差异^[21]。已有研究多局限在DNA遗传标记层面。中国不同地域的气候、饮食等环境因素有很大差异, 目前缺乏不同地域人群之间是否存在表观遗传学差异相关研究。本文应用表观基因组关联分析

(epigenome-wide association study, EWAS) 技术研究了中国南北方汉族人群之间DNA甲基化的差异, 并结合机器学习等算法构建了DNA甲基化南北方汉族人群推断算法。

1 材料与方法

1.1 样本信息

本研究使用的483份汉族男性DNA样本来源于国家科技资源共享服务平台计划项目, 使用Illumina的甲基化芯片 Infinium MethylationEPIC BeadChip (简称850K芯片) 检测 (科技部备份号: *BF2020121803316), 850K芯片可以检测整个基因组860 000个CpG位点, 覆盖CpG岛、启动子区、编码区、开放染色质和增强子区域。样本的数量和年龄、性别信息分别见表1及附件表S1。本研究通过公安部物证鉴定中心伦理委员会审查 (编

* 国家自然科学基金 (81772027) 和首都科技领军人才培养工程 (lj201825) 资助项目。

** 并列第一作者。

*** 通讯联系人。

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

叶健 Tel: 010-66269548, E-mail: yejian77@126.com

收稿日期: 2021-06-30, 接受日期: 2021-09-02

号: 2017-001), 所有参与者均签署了书面知情同意书。

Table 1 Size of samples used in this study		
Abbr.	Population	Sample size (<i>n</i>)
CHL	Han Chinese in Shandong, China	50
CHY	Han Chinese in Shanxi, China	149
HCH	Han Chinese in Henan, China	97
CHJ	Han Chinese in Jiangxi, China	89
CHC	Han Chinese in Sichuan, China	98

1.2 质量控制

使用R软件的ChAMP包对低质量数据进行预处理, 根据以下原则过滤探针: 检测到原始探针的*P*值大于0.01的甲基化位点; 在大于等于5%的样本的beads数目小于3的探针; 非CpG探针; 是SNP或探针覆盖区域内存在SNP的探针^[22]; 被鉴定为交叉反应的探针 (cross-reactive or multi-hit probes)^[23], 以及性染色体上的探针。分析得到每个探针位点的甲基化 β 值, 然后应用BIMQ (beta-mixture quantile) 方法进行 β 值的归一化处理。使用奇异值分解 (singular value decomposition, SVD) 方法检测是否存在批次效应与甲基化水平的关系^[24]。

本研究中样本的DNA提取自外周血, 外周血中不同的细胞类型DNA甲基化特征都不相同, 细胞类型组成是EWAS分析中的一个潜在混杂因素。使用ReFACToR算法^[25]计算不同细胞类型组成的主成分并在GLINT软件下游分析时将其添加为协变量, 从而减少细胞类型对EWAS分析结果的影响。参考Teschendorff等^[26]研究, 设置参数*K*=7 (7种细胞类型, 嗜酸性粒细胞、中性粒细胞、单核细胞、B细胞、NK细胞、CD8+ T和CD4+ T细胞)。

1.3 位点筛选及评估

通过EWAS结合机器学习Lasso回归的方法筛选南北方汉族人群的差异甲基化位点。GLINT软件^[27-30]基于Python2.7开发, 可以实现组织异质性校正和EWAS分析两种功能。基于上述质量控制后的DNA甲基化矩阵文件, 首先对南方汉族和北方汉族人群进行二分类编码, 南方汉族编码为1, 包括江西汉族和四川汉族, 北方汉族编码为0, 包括山东汉族、山西汉族和河南汉族, 并以此编码用作表型文件, 通过GLINT软件的数据管理功能 (data management) 检测并删除异常值。然后使用

ReFACToR算法校正组织的异质性 (adjusting for tissue heterogeneity), 最后对生成的数据文件进行EWAS分析, 删除染色体X和Y中的位点, 删除非特异性的位点^[31]。年龄和细胞类型用作协变量。

使用Lasso分析方法进一步筛选位点, 使用R (version 4.0.2) 软件的glmnet包建立Lasso-logistic回归模型。该模型是通过构造惩罚函数实现变量选择和参数估计, 通过将其回归系数设置为0的方式剔除呈现共线性或者与因变量没有相关性的冗余变量^[32-33], 选择对因变量影响较大的自变量并计算出相应的回归系数, 最终得到一个预测模型。

使用R软件基于每个DNA甲基化差异位点在中国南北方汉族群体的甲基化水平表达特征进行可视化展示。使用R软件softmaxreg包构建多元逻辑回归模型, 通过十折交叉验证的方法评估模型的准确性。其中多元逻辑回归使用caret包中的createDataPartition函数从训练集的每个标记人群中随机采样70%个体构建模型, 其余30%个体测试模型准确性, 使用confusionMatrix函数对上述模型进行评价; 十折交叉验证则是使用caret包中的createFolds函数对数据集进行划分, 将数据集分成10份, 轮流将其中9份作为训练集构建模型, 1份作为测试集来测试模型的准确性。为了使结果更加精确, 每种算法运行10次, 使用seed函数设定随机数的初始值, 产生不同的样本组合, 10次结果的均值作为对算法精度的估计。模型评价指标包括Kappa系数、灵敏性 (sensitivity)、特异性 (specificity)、阳性预测值 (PPV) 和阴性预测值 (NPV)。

2 结果与分析

2.1 73个CpG位点的筛选

483例汉族男性样本数据的EWAS分析结果见图1。挑选355个差异性显著的CpG位点 ($P < 1 \times 10^{-6}$)。使用Lasso回归进一步筛选位点剔除冗余变量。Lasso回归的特点是在拟合广义线性模型的同时进行变量筛选和复杂度调整。变量筛选是指有选择的把变量放入模型从而得到更好的性能参数。复杂度调整的程度由参数 λ 来控制, λ 越大对变量较多的线性模型的惩罚力度就越大, 从而精简变量, 结果见图2a、b。根据图2a曲线最低点确定惩罚值 λ , 在图2b的相应惩罚值的位置确定出模型最后所纳入的变量, 最终筛选出73个CpG位点, 位点信息具体见表2。

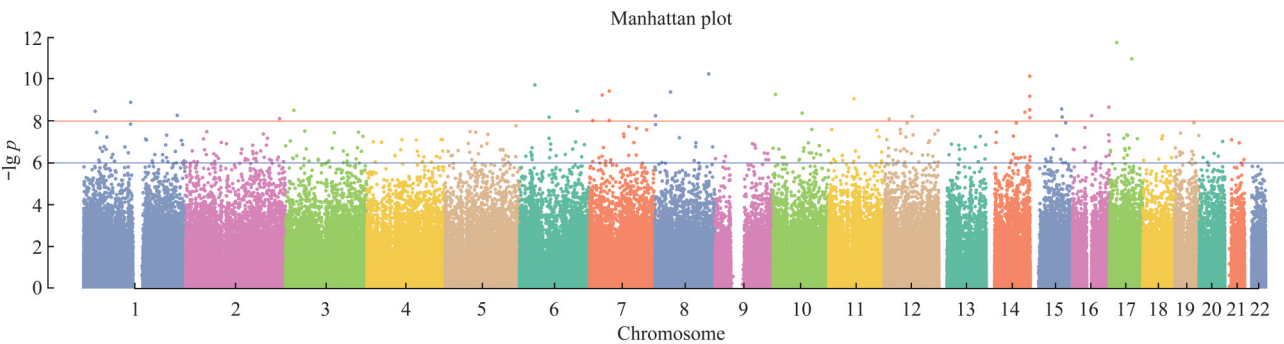


Fig. 1 Manhattan diagram of EWAS analysis results

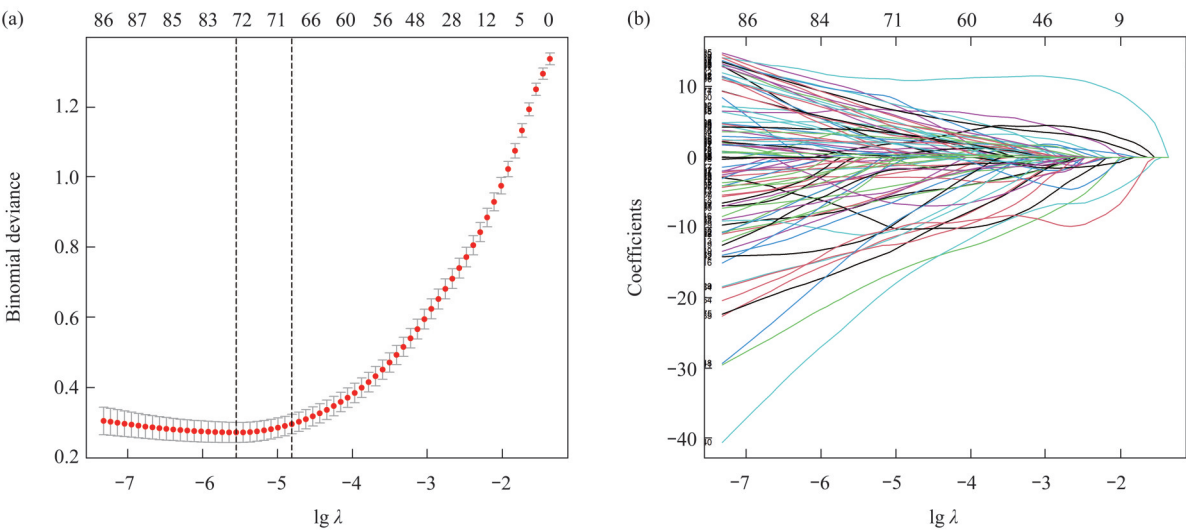


Fig. 2 Characteristic variable screening based on Lasso regression

The figure shows the process of selecting the most appropriate value of parameter λ in the Lasso model by cross-validation. (a) Lasso regression cross-validation of the optimal parameter atlas. (b) Sites in the model of regression coefficient.

Table 2 Information of the 73 CpGs

CpG ID	Chromosome	Position/bp	Genes	Classification ¹⁾
cg01485046	19	46 498 389	CCDC61	
cg01538166	17	17 743 987		Island
cg01606875	9	25 678 753	TUSC1	S_Shore
cg01775897	14	73 146 637	DPF3	
cg02250536	5	153 698 431	GALNT10	
cg02311879	2	86 388 045	IMMT	
cg02580250	10	134 807 902		S_Shore
cg03572525	16	4 332 683		
cg03691072	3	171 889 956	FNDC3B	
cg04017034	10	6 158 049	RBM17	
cg04264908	17	35 303 330		Island
cg04335779	16	89 386 882	LOC100287036; ANKRD11	
cg04531030	12	50 932 172	DIP2B	
cg04585778	17	54 395 292	ANKFN1	
cg04604773	15	74 218 217	LOXL1; LOXL1-AS1	N_Shore

Continued to Table 2				
CpG ID	Chromosome	Position/bp	Genes	Classification ¹⁾
cg04966714	14	106 071 722		
cg05237929	1	241 626 464		
cg06114546	16	88 645 532	ZC3H18	
cg06901759	20	47 001 709		Island
cg07163845	14	106 071 651		
cg07511934	16	89 386 912	ANKRD11	
cg07735461	8	95 523 890	KIAA1429	
cg07873154	6	30 122 544	TRIM10	
cg08103072	14	106 088 547		
cg08594888	12	53 183 773	KRT3	Island
cg09203577	16	46 995 848	DNAJA2	
cg09481056	16	2 339 260	ABCA3	Island
cg09485398	17	38 520 151	GJD3	Island
cg09794680	14	24 540 452	CPNE6	
cg09934216	9	1 050 222	DMRT2	Island
cg10016175	16	88 747 919	MGC23284; SNAI3	Island
cg10126715	1	37 500 195	GRIK3	Island
cg10321204	14	106 397 847	KIAA0125	
cg10321729	1	40 916 411	ZFP69B	S_Shore
cg11631063	4	184 679 835		S_Shore
cg11753363	6	37 626 255	MDGA1	Island
cg11815057	1	153 670 855		
cg12016107	1	202 311 492	UBE2T	S_Shore
cg12204171	10	115 935 565	C10orf118	S_Shore
cg12683946	3	45 521 859	LARS2	
cg13739942	3	46 977 532	CCDC12	
cg13764085	11	61 607 741	FADS2	
cg14749448	11	65 325 872	LTBP3	Island
cg15108641	10	99 263 320	UBTD1	S_Shelf
cg15303073	12	21 928 451	KCNJ8	
cg15451504	14	106 267 680		
cg15847527	7	47 611 885		
cg16063312	10	22 725 587		Island
cg16422365	14	106 041 833		S_Shore
cg16875104	7	30 635 889	GARS	S_Shore
cg17072244	8	140 418 352		
cg17326647	2	208 347 711		
cg17369140	3	150 066 466		
cg17620101	12	6 752 881	ACRBP	N_Shelf
cg17758808	8	37 979 743	ASH2L	
cg17867252	6	73 364 350	KCNQ5-IT1; KCNQ5	
cg18841660	11	128 936 723	ARHGAP32	
cg19610905	11	61 596 333	FADS2	Island
cg19962304	2	170 135 359	LRP2	
cg20329210	1	228 566 080	OBSCN	Island
cg20802826	6	30 782 080		
cg21314058	14	94 228 704	PRIMA1	
cg22052026	14	77 417 199		
cg22379462	17	2 615 549	CLUH	Island
cg23352146	11	12 701 929	TEAD1	
cg23742152	7	50 849 805	GRB10	Island

Continued to Table 2				
CpG ID	Chromosome	Position/bp	Genes	Classification ¹⁾
cg23964865	6	140 640 522		
cg24512321	2	199 179 548	LOC101927619	
cg25692791	17	43 019 644	KIF18B	
cg26455718	19	36 076 992		
cg26673722	4	68 776 652	TMPRSS11A	
cg27133716	17	37 897 822	GRB7	
cg27134251	8	1 365 049		N_Shore

¹⁾ Based on Illumina 850K annotation.

2.2 位点的评估

73个CpG位点在南北方汉族群体的甲基化水平表达特征如图3所示，蓝色代表北方汉族群体，红色代表南方汉族群体，图中的白点是中位数，小提琴图中的黑色粗条范围是上下四分位数，外部形

状即为分布密度。可以看出这些甲基化位点在两个群体中具有不同程度的差异。73个CpG位点的南北方汉族群体的主成分分析结果见图4。图4a可以看到，前两个主成解释变异的23%，基本可以区分南北方汉族人群。图4b中不同汉族人群用不同

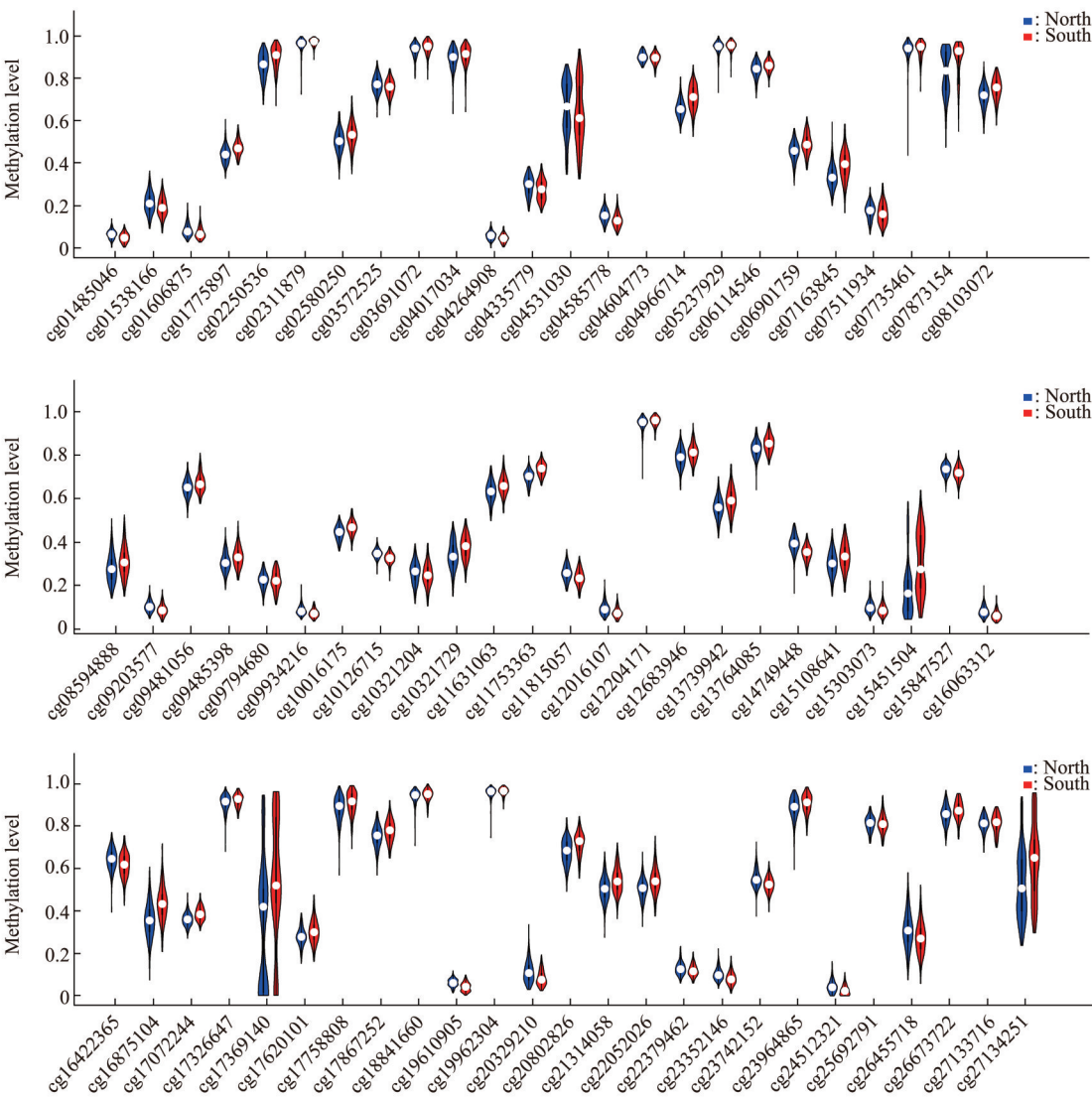


Fig. 3 Violin diagram of methylation levels of 73 CpG sites in northern and southern Han populations

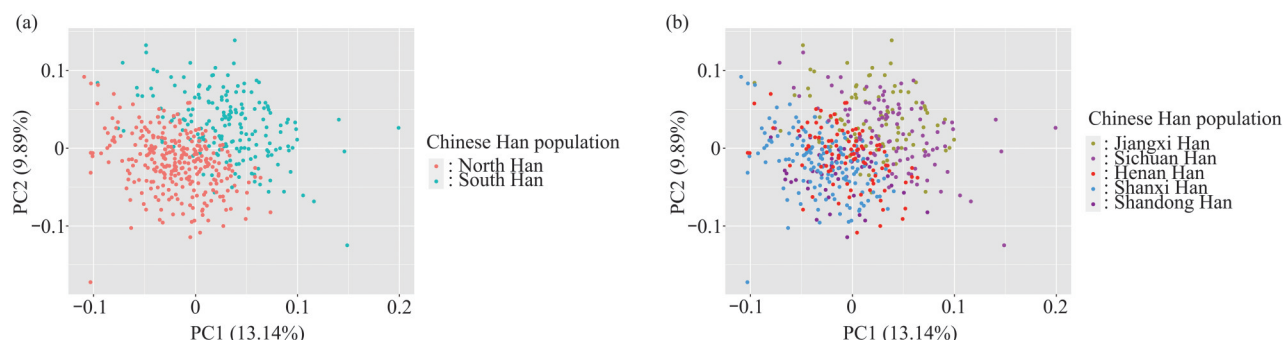


Fig. 4 Principal component analysis of northern and southern Han populations using 73 CpG sites

(a) Principal component analysis of northern and southern Han populations. (b) Principal component analysis of Han populations of different province.

颜色标注, 南北方汉族人群较各亚人群之间的差异更大, 北方人群中的河南汉族相对居中。绝大部分的江西汉族和四川汉族与山西汉族和山东汉族可以很好地区分开来。

模型构建与评估 10 次重复的结果具体见表 3, 最终多元逻辑回归 30% 测试集的预测准确率为 99.03%。Kappa 系数均值为 0.979 6; 灵敏性是两个

相关个体正确归类为相关的概率, 特异性反之, 分别为 0.990 7 和 0.989 5; 阳性预测值是判为相关个体中有多少是真的相关, 阴性预测值反之, 结果分别为 0.993 1 和 0.986 2。10 次十折交叉验证的结果均在 98% 以上, 最终平均准确率为 98.79%, 其余各项模型预测性能指标均超过 0.95。

Table 3 The performance statistics of the prediction model

Times	Seed	70% Training vs 30% Testing						Average accuracy of 10 times 10-fold cross-validation
		Accuracy	Kappa	Sensitivity	Specificity	PPV	NPV	
1	21	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0	0.987 5
2	30	0.993 1	0.985 0	1.000 0	0.981 1	0.989 1	1.000 0	0.989 6
3	41	0.993 1	0.984 8	0.989 4	1.000 0	1.000 0	0.980 4	0.985 5
4	44	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0	1.000 0	0.991 7
5	49	0.993 1	0.985 8	0.988 0	1.000 0	1.000 0	0.983 9	0.987 6
6	57	0.979 2	0.957 4	0.975 9	0.983 6	0.987 8	0.967 7	0.983 3
7	65	0.993 1	0.985 7	1.000 0	0.983 3	0.988 2	1.000 0	0.987 5
8	77	0.979 2	0.956 8	0.976 7	0.982 8	0.988 2	0.966 1	0.987 6
9	81	0.986 1	0.969 6	0.989 2	0.980 4	0.989 2	0.980 4	0.989 6
10	90	0.986 1	0.971 4	0.988 1	0.983 3	0.988 1	0.983 3	0.989 6
Average		0.990 3	0.979 6	0.990 7	0.989 5	0.993 1	0.986 2	0.987 9

3 讨 论

环境因素可通过表观遗传机制, 如诱导 DNA 甲基化模式改变等, 在不改变 DNA 序列前提下, 改变基因表达, 从而引发表型变化。法医学研究已经发现大量不同组织之间、不同年龄之间存在差异的 DNA 甲基化位点, 并构建预测模型^[6-10]。不同地域的气候、饮食等环境因素存在差异, 环境通过影响 DNA 甲基化水平进而影响基因表达, 使得生

物表型发生变化, 进而造成不同地域人群间的差异。Fraser 等^[13]研究北欧 (CEU) 和西非 (YRI) 多个家系的 DNA 甲基化数据, 发现在族群内和族群间存在显著的甲基化差异, 这种差异可能来自等位基因突变、上位效应, 以及基因跟环境之间的相互作用, 还发现在转录起始位点附近的 DNA 甲基化有群体特异性。Yuan 等^[14]研究了 509 份胎盘的 450K 甲基化芯片数据, 基于甲基化位点进行亚洲、非洲、及高加索人群的推断, 准确性为 0.938。东

亚是全球人口最多的区域,占全球人口的22%,汉族是东亚主体民族,大量基因组学研究发现汉族内部存在明显的南北遗传差异,然而目前缺乏针对不同地域汉族人群的表现遗传学差异研究。

本文研究了483份汉族个体的甲基化数据,虽然南方人群没有使用广东广西样本,而是相对靠北的江西和四川人群样本,仍然研究发现了南北方汉族人群之间的甲基化差异,并最终筛选出73个CpG位点。图4a、b展示了不同地域人群间的甲基化差异,这些差异可能归因于等位基因频率的差异以及上位效应或基因与环境的相互作用^[13]。地理学第一和第二定律指出地物之间的距离越近,相关性越大;空间隔离又造成地物之间的空间异质性^[34-36]。从图4b可以看出河南汉族居中,可能是因为河南在地理位置上属于华中地区,地理位置接近的人群之间的基因交流频繁,遗传距离接近,遗传差异度就小。对于距离相近的人群往往需要更多的位点和更大的参考人群数据才可以实现精确区分。模型预测性能验证采用多元逻辑回归随机抽取70%个体构建模型,30%个体进行测试以及十折交叉验证的方法。未来可增加样本量进一步验证这一组甲基化位点的人群区分准确性。

生物体的一些性状是由多基因共同控制的,其测量值可以用连续的数量进行表示,这些性状统称为数量性状(quantitative trait, QT)。遗传变异可以调节多种机体代谢功能,包括DNA甲基化在表现遗传学中所表现出的基因表达调节作用。目前比较重要的一种研究方法称为数量性状定位分析(quantitative trait locus, QTL),即定位控制数量性状的基因位点在基因组中的位置。DNA甲基化数量性状位点分析(DNA methylation quantitative trait locus, meQTLs)是以DNA甲基化作为数量性状,探讨DNA甲基化水平与基因表达之间关系^[37-38]。除了年龄、性别以及疾病等环境因素会影响个体的甲基化水平,DNA序列也会对甲基化产生影响,尤其是特异性SNP的改变。有研究发现非洲和欧洲祖先群体之间70%的差异甲基化位点至少与一个meQTL相关^[39],这表明很大一部分的DNA甲基化的群体差异可能主要是由DNA序列变异引起。而从表现遗传角度入手实现地域人群的区分关键在于寻找独立于SNP的甲基化位点,从而作为STR、SNP等基因组遗传标记的有效补充。

近距离人群的区分和推断是法医DNA领域的难点之一。目前已报道的AISNP体系初步实现了

东亚南北方人群的区分,未来结合表观遗传标记,有望实现更加精细的人群区分。本文探索研究了利用甲基化进行不同地域人群推断的可行性,研究表明南北方汉族人群之间存在表观遗传差异,未来需进一步增加人群数据获取更多地域相关甲基化位点,与族群相关SNP位点配合使用实现东亚人群的精细区分。

附件 PIBB_20210091_Table_S1.xlsx 见本文网络版 (<http://www.pibb.ac.cn>或<http://www.cnki.net>)。

参 考 文 献

- [1] Lam L L, Emberly E, Fraser H B, *et al.* Factors underlying variable DNA methylation in a human community cohort. *Proc Natl Acad Sci USA*, 2012, **109**(Suppl 2):17253-17260
- [2] Feil R, Fraga M F. Epigenetics and the environment: emerging patterns and implications. *Nat Rev Genet*, 2012, **13**(1): 97-109
- [3] Pujadas E, Feinberg A P. Regulated noise in the epigenetic landscape of development and disease. *Cell*, 2012, **148**(6): 1123-1131
- [4] 聂燕钗,俞丽娟,管桦,等.DNA甲基化检测方法及其法医学应用研究进展. *法医学杂志*, 2017, **33**(3):293-300
Nie Y C, Yu L J, Guan H, *et al.* *Journal of Forensic Medicine*, 2017, **33**(3):293-300
- [5] Kader F, Ghai M. DNA methylation-based variation between human populations. *Mol Genet Genomics*, 2017, **292**(1):5-35
- [6] Bocklandt S, Lin W, Sehl M E, *et al.* Epigenetic predictor of age. *PLoS One*, 2011, **6**(6):e1482
- [7] Koch C M, Wagner W. Epigenetic-aging-signature to determine age in different tissues. *Aging (AlbanyNY)*, 2011, **3**(10): 1018-1027
- [8] Hannum G, Guinney J, Zhao L, *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell*, 2013, **49**(2):359-367
- [9] Freire-Aradas A, Phillips C, Mosquera-Miguel A, *et al.* Development of a methylation marker set for forensic age estimation using analysis of public methylation data and the Agena Bioscience EpiTYPER system. *Forensic Sci Int Genet*, 2016, **24**:65-74
- [10] Forat S, Huettel B, Reinhardt R, *et al.* Methylation markers for the identification of body fluids and tissues from forensic trace evidence. *PLoS One*, 2016, **11**(2):e0147973
- [11] Boks M P, Derks E M, Weisenberger D J, *et al.* The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PLoS One*, 2009, **4**(8):e6767
- [12] Gervin K, Hammerø M, Akselsen H E, *et al.* Extensive variation and low heritability of DNA methylation identified in a twin study. *Genome Res*, 2011, **21**(11):1813-1821
- [13] Fraser H B, Lam L L, Neumann S M, *et al.* Population-specificity of human DNA methylation. *Genome Biol*, 2012, **13**(2):R8

- [14] Yuan V, Price E M, Del Gobbo G, *et al.* Accurate ethnicity prediction from placental DNA methylation data. *Epigenetics Chromatin*, 2019, **12**(1):51
- [15] Wen B, Li H, Lu D, *et al.* Genetic evidence supports demic diffusion of Han culture. *Nature*, 2004, **431**(7006):302-305
- [16] Chu J Y, Huang W, Kuang S Q, *et al.* Genetic relationship of populations in China. *Proc Natl Acad Sci USA*, 1998, **95**:11763-11768
- [17] Yao Y G, Kong Q P, Bandelt H J, *et al.* Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *Am J Hum Genet*, 2002, **70**(3):635-651
- [18] Xue F, Wang Y, Xu S, *et al.* A spatial analysis of genetic structure of human populations in China reveals distinct difference between maternal and paternal lineages. *Eur J Hum Genet*, 2008, **16**(6):705-717
- [19] 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
- [20] Zhang F, Su B, Zhang Y P, *et al.* Genetic studies of human diversity in East Asia. *Philos Trans R Soc Lond B Biol Sci*, 2007, **362**(1482):987-995
- [21] 江丽, 刘京, 李彩霞. 从"非洲起源说"研究到"DNA 族群地域推断"侦查应用. *法医学杂志*, 2019, **35**(5):519-524
Jiang L, Liu J, Li C X. *Journal of Forensic Medicine*, 2019, **35**(5):519-524
- [22] Zhou W, Laird P W, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res*, 2017, **45**(4):e22
- [23] Chen Y A, Lemire M, Choufani S, *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, 2013, **8**(2):203-209
- [24] Johnson W E, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 2007, **8**(1):118-127
- [25] Rahmani E, Zaitlen N, Baran Y, *et al.* Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat Methods*, 2016, **13**(5):443-445
- [26] Teschendorff A E, Breeze C E, Zheng S C, *et al.* A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. *BMC Bioinformatics*, 2017, **18**(1):105
- [27] Rahmani E, Shenhav L, Schweiger R, *et al.* Genome-wide methylation data mirror ancestry information. *Epigenetics Chromatin*, 2017, **10**:1
- [28] Rahmani E, Yedidim R, Shenhav L, *et al.* GLINT: a user-friendly toolset for the analysis of high-throughput DNA-methylation array data. *Bioinformatics*, 2017, **33**(12):1870-1872
- [29] Reinus L E, Acevedo N, Joerink M, *et al.* Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One*, 2012, **7**(7):e41361
- [30] Koestler D C, Jones M J, Usset J, *et al.* Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *BMC Bioinformatics*, 2016, **17**:120
- [31] McCartney D L, Walker R M, Morris S W, *et al.* Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. *Genom Data*, 2016, **9**:22-24
- [32] 李淑锦, 嵇晓佳. Lasso-Cox 模型在个人信用风险评估中的应用. *资源开发与市场*, 2021, **37**(2):129-135
Li S J, Ji X J. *Resource Development & Market*, 2021, **37**(2):129-135
- [33] 冯巩, 张燕, 严琴琴, 等. 基于 LASSO 回归的非酒精性脂肪性肝病进展性肝纤维化预测模型的构建及分析. *临床肝胆病杂志*, 2020, **36**(10):2214-2218
Feng G, Zhang Y, Yan Q Q, *et al.* *Journal of Clinical Hepatology*, 2020, **36**(10):2214-2218
- [34] Tobler W R. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 2016, **46**:234-240
- [35] 李小文, 曹春香, 常超一. 地理学第一定律与时空邻近度的提出. *自然杂志*, 2007, **29**(2):69-71
Li X W, Cao C X, Chang C Y. *Chinese Journal of Nature*, 2007, **29**(2):69-71
- [36] Zhu A X, Lu G N, Liu J, *et al.* Spatial prediction based on third law of geography. *Ann GIS*, 2018, **24**(4):225-240
- [37] 刘瑞. 生命早期饥荒暴露人群DNA甲基化数量性状位点与精神分裂症的关联性研究[D]. 长春: 吉林大学, 2018
Liu R. *Association Between DNA Methylation Quantitative Trait Loci and Schizophrenia with Famine in Early Life*[D]. Changchun: Jilin University, 2018
- [38] 梁成. 自噬相关基因遗传变异与非小细胞肺癌预后关联及功能研究[D]. 南京: 南京医科大学, 2017
Liang C. *Role of ATG10 Expression Quantitative Trait Loci in Non-small Cell Lung Cancer Survival*[D]. Nanjing: Nanjing Medical University, 2017
- [39] Husquin L T, Rotival M, Fagny M, *et al.* Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation. *Genome Biol*, 2018, **19**(8):66-78

Epigenetic Differences Research of DNA Methylation in Northern and Southern Chinese Han Populations*

SUN Chang-Chun^{1,2)**}, XU Ji-Chen^{2)**}, JIANG Li²⁾, GUO Xiao-Yuan^{1,2)}, ZHAO Wen-Ting²⁾,
YE Jian^{2)***}, LI Cai-Xia^{2)***}

⁽¹⁾School of Forensic Medicine, Shanxi Medical University, Taiyuan 030001, 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)

Abstract Objective The phenotypes of population, such as biogeographical and physical characteristics, are the result of the interaction between genotype and environment. A large number of genomic studies have shown that the Han Chinese are admixed population which displayed obvious genetic differences between the north and the south. The purpose of this study is to explore whether there are differences in epigenomes between northern and southern Han populations, and to screen biogeographical specific genetic loci. **Methods** Glim software was used to perform epigenome-wide association study (EWAS) analysis on genome-wide methylation microarray data of 483 Han Chinese individuals, and Lasso regression was used to screen the sites. Multiple logistic regression algorithm was used to construct the prediction model for the Han population in the south and the north, and the evaluation was conducted by the method of 10-fold cross validation. **Results** A group of CpG sites with significant differences between southern and northern Han populations were screened out, with accuracy of 99.03% and Kappa of 0.979 6. **Conclusion** This study shows that there are epigenetic differences between the southern and northern Han populations, which lays a foundation for further studies on epigenetic differences between the Han populations in different regions.

Key words forensic genetics, epigenetics, DNA methylation, EWAS, Han Chinese

DOI: 10.16476/j.pibb.2021.0091

* This work was supported by grants from The National Natural Science Foundation of China (81772027) and Beijing Leading Talent Program (lj201825).

** These authors contributed equally to this work.

*** Corresponding author.

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

YE Jian. Tel: 86-10-66269548, E-mail: yejian77@126.com

Received: June 30, 2021 Accepted: September 2, 2021