基于数据库的儿童急性髓系白血病DNA甲基化亚型及其预后分析*

张敬晨[1] 丛雅迪[2] 赵宇柯[3] 玄立田[4] 张胜发[1] 周 伟[1△]

【提 要】 **目的** 基于多组学数据解析儿童急性髓系白血病(pediatric acute myeloid leukemia, PAML)的甲基化特征及其分子机制。**方法** 整合TARGET和GEO数据库,基于PAML患者与正常样本间的差异甲基化位点,将PAML患者分为三个甲基化亚型。利用基因表达数据筛选差异表达基因,然后用LASSO-Cox筛选PAML的预后相关关键基因,最后基于XGBoost的Cox比例风险模型构建预后得分模型。**结果** 共筛选出*SPEF2*、*HCG11*和*SLC16A1-AS1*等9个PAML预后相关关键基因,成功构建出基于PAML甲基化亚型的预后评分模型。生存分析显示甲基化亚型与患者预后显著相关,提示异常甲基化可能通过沉默抑癌基因或激活原癌基因影响预后。**结论** PAML的分子机制和潜在治疗靶点,为改善PAML患者预后提供了新的研究方向。

【关键词】 儿童急性髓系白血病 甲基化 基因表达 关键基因 预后得分模型
【中图分类号】 R733.7 【文献标志码】 **DOI** 10.11783/j.issn.1002-3674.2026.04.017

急性髓系白血病(acute myeloid leukemia, AML)是一种具有显著复发性和遗传易感性的高危恶性肿瘤。AML的发病机制源于基因突变导致的髓系干细胞(红细胞、血小板和白细胞)前体肿瘤化和克隆增殖[1]。儿童急性髓系白血病(pediatric acute myeloid leukemia, PAML)的发病率相对较低,但高危型患者死亡率高。目前FAB(French-American-British)将AML通过细胞形态学和细胞化学划分为M0~M7[2]。这种传统分类具有主观性,不能准确预测AML和PAML预后和不同个体的治疗需求。高通量测序技术的发展推动了分子分层方法(通过基因组学、转录组学等多组学数据对疾病进行分子特征分型)的应用,为疾病分类提供了新的视角。该方法能够更深层次地揭示不同患者的分子生物学特征。例如,Midostaurin(*FLT3*抑制剂)对PAML基因组中高频突变*FLT3*相关突变展现出较好的治疗效果[3-4]。然而,与其他常见的白血病相比,PAML DNA甲基化水平的研究仍相对不足。

DNA甲基化作为与基因表达密切相关的表观遗传修饰方式,被公认为是癌症发展中的关键因素之一。本研究旨在通过生物信息学分析确定PAML的甲基化亚型并评估其预后意义,为PAML的诊断提供新的方向和策略。

材料和方法

1. 数据获取

本研究数据来源于儿童肿瘤数据库TARGET(Therapeutically Applicable Research to Generate Effective Treatments, https://ocg. cancer. gov/programs/target)的TARGET-AML项目,包含280例PAML样本的DNA甲基化数据(含临床信息),274例对应样本的转录组数据和70例作为模型独立测试集的转录组数据。健康儿童对照数据从GEO(Gene Expression Omnibus, https://www. ncbi. nlm. nih. gov/geo/)的GSE36054数据集获取(n=175),用于差异甲基化位点分析(表1)。

表1 样本信息

数据集	数据类型	样本数量(个)		分析项目
		健康对照	疾病	
TARGET-AML	DNA甲基化数据		280	差异甲基化分析
TARGET-AML	基因表达数据		274	GSEA,差异表达分析
TARGET-AML	基因表达数据		70	预后得分模型测试集
GSE36054	DNA甲基化数据	175		差异甲基化分析

2. PAML甲基化分型

(1)筛选差异甲基化位点

用R 4. 3. 1中ChAMP(v2. 32. 0)对DNA甲基化数据预处理,去除X/Y染色体上的探针、与单核苷酸多态性(single nucleotide polymorphism, SNP)相关探针、非CpG位点探针、存在多重比对(multi-hit)的探针和基因间区探针。使用limma(v3. 38. 1)中normalizeBetweenArrays和removeBatchEffect对数据进行标准化和去除批次效应。随后进行差异甲基化分析,筛选位点标准为:Benjamini-Hochberg(BH)校正后P<0. 01。PAML样本β值(甲基化/未甲基化等位基因强度比)均值大于正常样本β值均值;正常样本β值均值小于0. 05[5]。该策略可以筛选正常样本中未甲基化而在肿瘤中发生异常甲基化的位点。对差异甲基化位

* 基金项目:国家重点研发计划常见多发病防治研究专项(2023YFC2508801)
1. 中国医学科学院北京协和医学院,国家人口健康科学数据中心(100730)
2. 爱丁堡大学
3. 北京协和医院肾内科
4. 首都儿科研究所附属儿童医院血液内科
△通信作者:周伟,E-mail: zhouwei_ncmi@126. com

点进行单因素Cox回归分析（P<0.01），筛选与患者生存期显著性相关的位点。然后结合白细胞计数、骨髓和血浆中白血病细胞比例进行多因素Cox回归（P<0.01）对位点进行验证，避免高估独立效应。

（2）PAML DNA甲基化亚型的聚类分析

用R 4.3.1中ConsensusClusterPlus（v1.66.0）采用K-means算法（重抽样比例0.8，重复500次）对PAML预后相关甲基化位点进行无监督一致性聚类。基于模糊聚类比例（proportion of ambiguous clustering，PAC）原理并结合轮廓系数曲线将所有样本分为3个亚型。然后使用Kaplan-Meier分析三种亚型的生存差异。

（3）PAML甲基化亚型与临床特征的相关性

使用R 4.3.1中gmodels（v2.9.1）对甲基化亚型与临床特征的关联性进行卡方检验，评估其与性别和急性髓系白血病的FAB分型的关系。为确保统计准确性，分析FAB分型时去除了分型不明确的样本（n=33）。

3. 基因集富集分析（gene set enrichment analysis，GSEA）

GSEA用于评估基因集在表型相关基因中的分布趋势，揭示基因集对表型的贡献。本研究使用GSEA（v4.3.3）比较三种甲基化亚型的基因表达数据。采用c2.all.v2024.1.Hs.symbols基因集，使用weighted富集算法，执行1000次置换检验，基于Signal2Noise对基因进行排序。结果筛选标准为P<0.01且错误发现率（FDR）<25%。

4. 建立PAML预后得分模型

（1）差异基因表达矩阵构建

使用R 4.3.1中Deseq2（v1.42.0）计算三种甲基化亚型间的差异表达基因。用BH校正P<0.01和|logFC|>1作为筛选标准，logFC（$\log_2$ fold change）表示基因表达在不同亚型间的倍数变化。

（2）GO和KEGG富集分析

使用R 4.3.1中clusterProfiler（v4.10.0）和enrichplot（v1.22.0）对差异表达基因进行GO富集分析，阈值为BH矫正P<0.01且Q（FDR矫正P）<0.01。然后进行KEGG富集分析，阈值为P<0.01且Q<0.01。

（3）关键基因的选取和模型构建

对差异表达基因进行LASSO-Cox回归分析（P<0.01），筛选出9个PAML预后相关关键基因。结合FAB分型信息，用R 4.3.1中caret（v6.0.94）构建基于XGBoost的Cox比例风险模型，XGBoost算法具有出色的特征重要性排序能力和抗过拟合特性[6]。Cox模型的风险函数如下：

$$h(t|\boldsymbol{x})=h_0(t)\cdot\exp(\boldsymbol{\beta x})$$

其中，$h(t|\boldsymbol{x})$表示个体在t时间的风险函数，$h_0(t)$表示风险基线函数，x表示基因表达值，β表示特征权重系数（通过最大部分似然估计）。XGBoost通过最大化分裂增益（gain）选择特征，gain代表某基因在所有树的分裂中带来的累积损失减少量占比。

首先用五折交叉验证法筛选较优参数范围，用负对数似然作为评估指标，采用调参的方法确定模型的最优超参数（学习率0.1，树深6，最小样本权重5，样本比例0.7，colsample_bytree 0.8，gamma 0.1，lambda 1，alpha 0.1，迭代48次）。按样本预后评分的中位数将样本分为高风险和低风险组，进行Kaplan-Meier生存分析（log-rank检验，P<0.01）。

（4）模型评估及验证

通过计算训练集和测试集关键临床时间点（0.5年、1年、3年和5年）的时间依赖性AUC和C-index（一致性指数）来评估模型。对测试集样本进行Kaplan-Meier生存分析（log-rank检验，P<0.01），检验模型的分层效果。

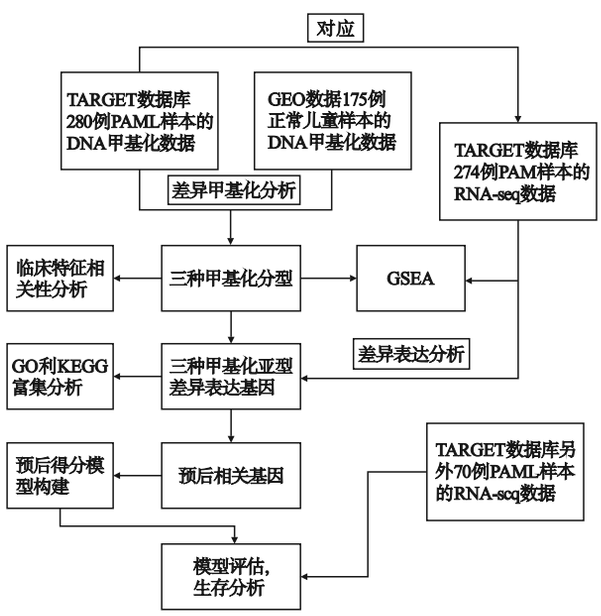


图1 分析流程

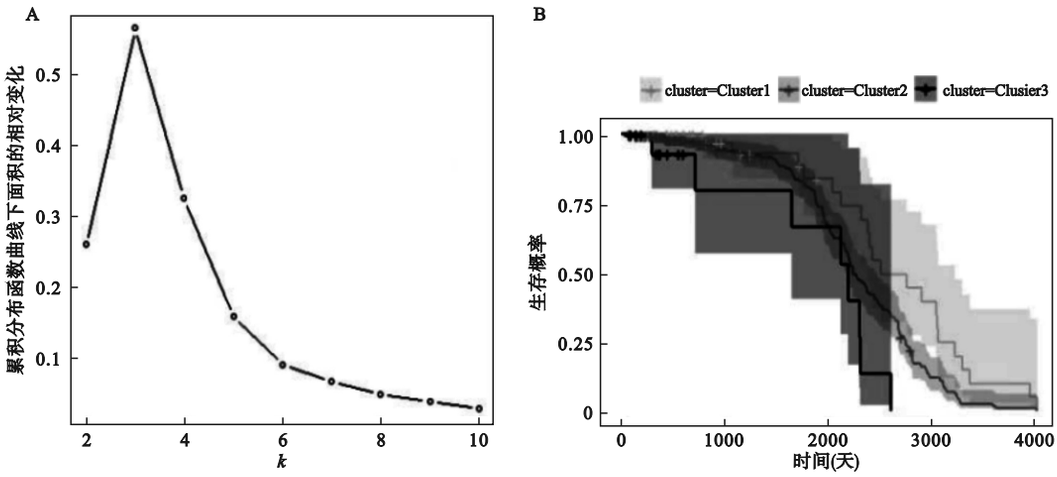
结 果

1. PAML中预后相关的DNA甲基化差异位点

基于TARGET-AML中PAML患者和GSE36054中正常儿童DNA甲基化数据筛选出117个预后相关甲基化位点。用K-means算法进行无监督聚类，并结合PAC值和轮廓系数曲线确定最佳聚类数为k=3（图2A）。将三个甲基化亚型命名为Cluster1，Cluster2和Cluster3。

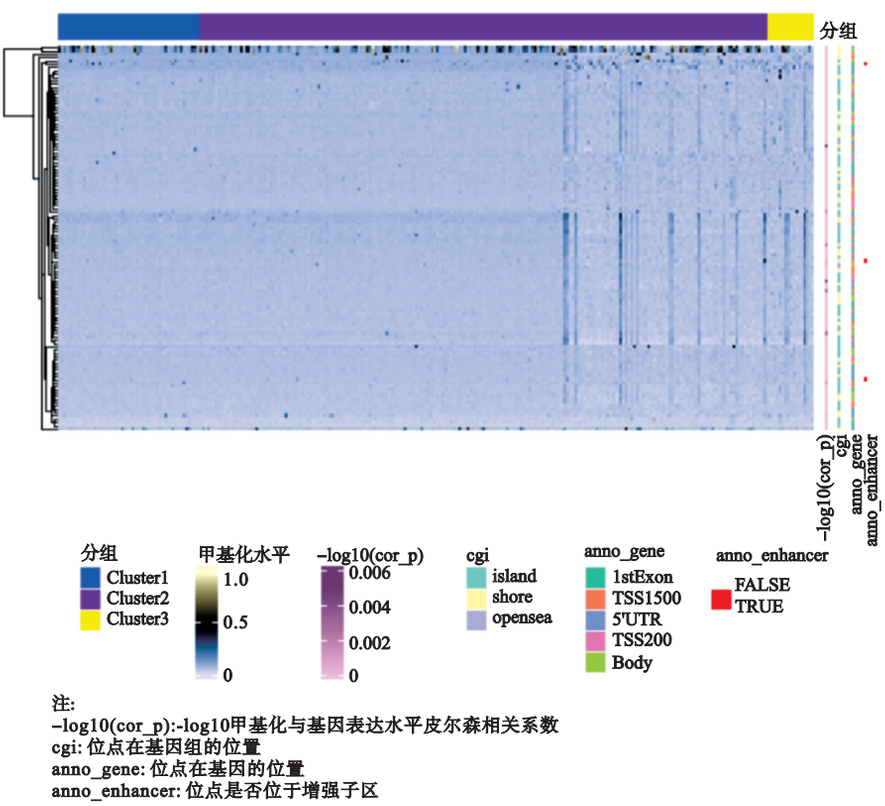
生存分析显示Cluster1预期生存时间最长，预后最好；Cluster3预期生存时间最短，预后最差（图2B）。甲基化分析表明，Cluster1的总体甲基化水平最低，Cluster3最高，Cluster2样本甲基化水平分布跨越较大

（图 3）。三种亚型差异甲基化位点主要富集于 CpG 岛，且多数位于启动子区域（TSS1500 和 TSS200）。



（A）不同分类系数下的轮廓系数折线图；（B）三种甲基化亚型的生存曲线

图 2 PAML 预后相关 DNA 甲基化位点的一致性聚类分析



注：
–log10(cor_p):-log10甲基化与基因表达水平皮尔森相关系数
cgi: 位点在基因组的位置
anno_gene: 位点在基因的位置
anno_enhancer: 位点是否位于增强子区

图 3 PAML 预后相关的 DNA 甲基化位点热图

相关性分析显示，PAML 患者的甲基化亚型与性别无显著相关性，但与 FAB 分型显著相关（表 2）。说明不同 FAB 分型的 PAML 患者在预后方面存在显著差异，且与 DNA 甲基化修饰密切相关。

2. GSEA

GSEA 显示不同甲基化亚型转录存在显著异质性：Cluster1 特异性富集于 *NPM1* 突变相关基因集（如 VERHAAK_AML_WITH_NPM1_MUTATED_DN）。Cluster2 差异表达基因显著富集于 *FLT3* 相关通路并呈显著上调趋势。Cluster3 未在 c2. all. v2024. 1. Hs. symbols 数据库中发现显著富集通路。

表 2 卡方检验 PAML DNA 甲基化亚型与性别和 FAB 分型的关系

类目		样本个数	卡方检验 *P* 值
性别	女	131	0. 93
	男	149	
FAB 分型	M0	11	0. 02
	M1	35	
	M2	65	
	M4	67	
	M5	49	
	M6	6	
	M7	14	

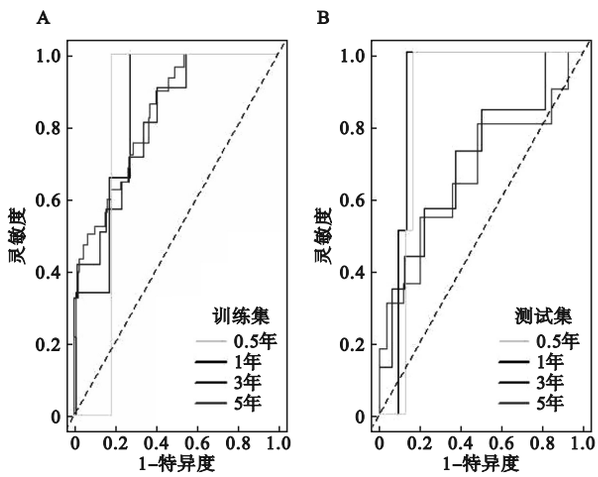
3. PAML预后关键基因筛选

基于三种甲基化亚型，从274例样本中筛选出1883个差异表达基因。两两比较结果显示，Cluster1和Cluster3的差异表达基因最多，Cluster1和Cluster2的差异表达基因最少。结合差异表达基因和患者FAB分型，用LASSO−Cox回归筛选出与生存期显著相关的基因，然后用基于XGBoost的Survival−Cox方法构建PAML预后得分模型。此过程中筛选出的9个基因为PAML预后相关关键基因（表3）。此外，与生存期显著相关的FAB分型为M1（急性粒细胞白血病未分化型）、M2（急性粒细胞白血病部分分化型）、M4（急性粒单核细胞白血病）和M5（急性单核细胞白血病）。

通过预后得分模型计算样本分数后，用AUC及*C*−index评估模型性能。训练集患者0.5年、1年、3年和5年生存期的AUC值分别为0.817、0.849、0.818、0.834（图4A），*C*−index值为0.827。独立测试集患者0.5年、1年、3年和5年生存期的AUC值分别为0.853、0.887、0.700、0.665（图4B），*C*−index是0.702。结合ROC曲线可知，测试集的5年生存率准确性较差，但在短期（0.5年，1年，3年）预测中表现较好。

表3　PAML预后得分模型基因列表

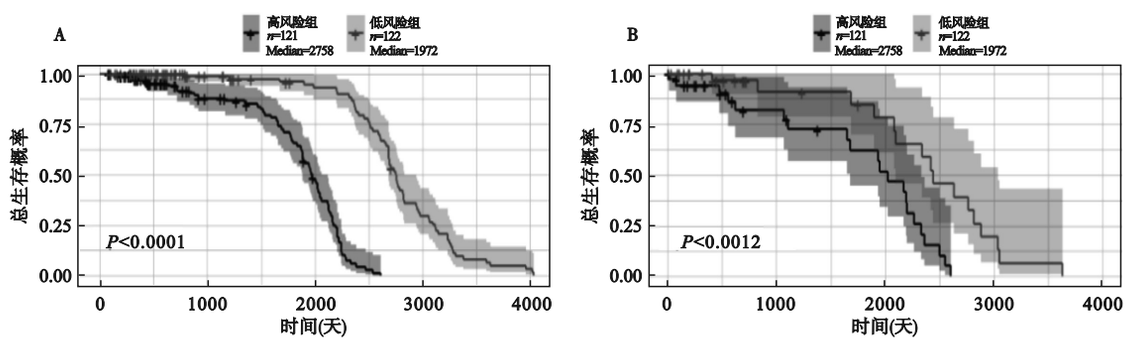
gene	gain	cover	frequency
SPEF2	0.153672647	0.168746431	0.148148148
SLC16A1−AS1	0.132278416	0.126542815	0.126705653
AC131212.2	0.131140756	0.104490994	0.101364522
AC000065.1	0.117241979	0.114184715	0.103313840
HCG11	0.100114684	0.087197339	0.087719298
UBE2D1	0.091929017	0.087325973	0.089668616
ARRDC5	0.084027589	0.091386204	0.093567251
SOCS6	0.081821601	0.087553452	0.107212476
TAF10	0.063605701	0.073647405	0.083820663



（A）训练集的ROC曲线；（B）测试集的ROC曲线

图4　预后得分模型的ROC曲线

训练集和测试集样本按中位数分为高风险组和低风险组后，进行Kaplan−Meier生存分析，结果显示两组患者的总体生存率存在显著差异（图5）。训练集和测试集生存曲线约在1000天逐渐分离，表示该模型具有良好的风险分层能力。



（A）训练集的K−M曲线；（B）测试集的K−M曲线

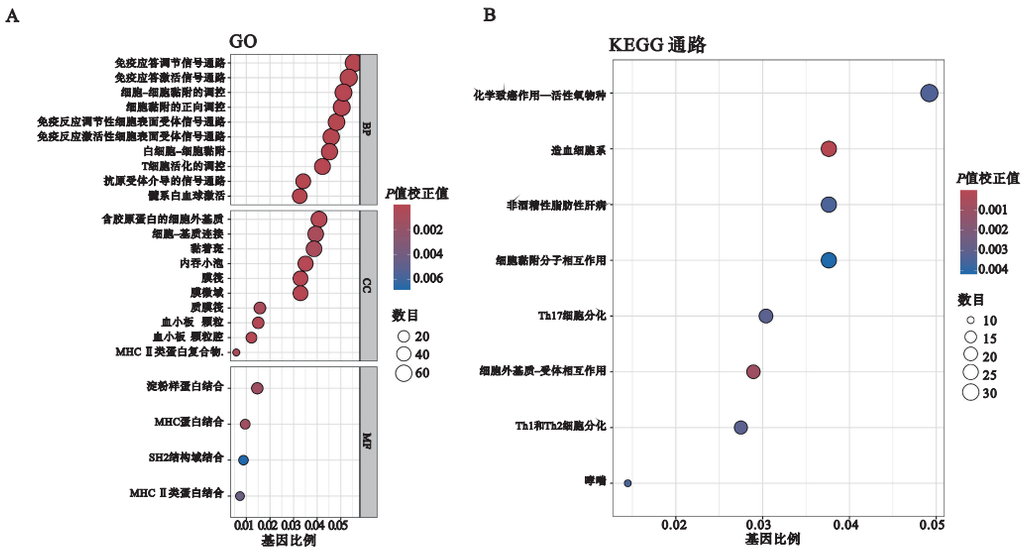
图5　预后得分模型区分的高低风险组的K−M曲线

4. GO和KEGG富集功能分析

对PAML不同甲基化亚型的1883个差异表达基因进行GO和KEGG富集分析。GO功能富集得到70个生物过程：免疫反应激活细胞表面受体信号传导通路，抗原受体介导的信号传导通路和髓系白细胞激活等；29个细胞组分：膜筏，血小板α颗粒和内吞小泡等；4个分子功能结果：MHC蛋白结合作用，MHC Ⅱ类分子与抗原肽相互作用，*β*淀粉样蛋白结合作用和SH2结构域结合作用（图6A）。KEGG富集结果有8条：造血细胞系，细胞外基质−受体相互作用，Th1和Th2细胞分化等（图6B）。

讨　论

PAML是一种在青少年中发生的恶性急性髓系白血病，本文对PAML发展的分子机制进行了系统性分析，将患者分为三种甲基化亚型：Cluster1，Cluster2和Cluster3。GSEA结果显示Cluster1亚型患者表现出发生*NPM1*突变相关的基因表达特征。*NPM1*突变是AML中最常见的获得性分子异常，不伴有*FLT3*内部串联重复（*FLT3* ITD）的*NPM1*突变患者具有更好的临床预后[7]。*NPM1*突变型AML在2022年国际共识分类和WHO第五版髓系肿瘤分类中被标记为单独的一类，可结合ELN分型指导治疗[8]。这说明Cluster1



（A）GO 富集结果；（B）KEGG 富集结果

图 6 功能富集图

患者预后较好，可能对 *NPM*1 靶向治疗有良好反应。Cluster2 亚型患者表现出 *FLT*3 信号通路激活特征，提示可能存在 *FLT*3 ITD，该突变在约 1% 的儿童患者中发现与疾病表型恶化相关[9]。*FLT*3 的激活通过 RAS-ERK 信号通路促进肿瘤发生发展，Cluster2 型患者可能通过 *FLT*3 抑制剂治疗。Cluster3 亚型患者的高甲基化可能通过沉默抑癌基因导致预后不良。研究表明，B 细胞慢性淋巴细胞白血病的发病机制与启动子区域异常高甲基化引起的抑癌微小 RNA（miRNA）功能失调相关[10]。因此，靶向 DNA 甲基化修饰可能成为改善 Cluster3 亚型患者预后的治疗策略。

本研究筛选出 9 个 PAML 预后相关关键基因：*SPEF*2，*UBE*2*D*1，*ARRDC*5，*SLC*16*A*1-*AS*1，*AC*131212.2，*HCG*11，*AC*000065.1，*TAF*10 和 *SOCS*6。这些基因通过不同机制参与疾病进程：*SPEF*2 编码环状 RNA circ-*SPEF*2，诱导其表达能阻碍弥漫性大 B 细胞淋巴瘤的增殖[11]，提示其可能成为 PAML 的治疗靶点。*UBE*2*D*1 和 *ARRDC*5 通过影响泛素化过程参与肿瘤细胞的增殖和转移。*UBE*2*D*1 的异常表达会影响 TGF-*β* 通路[12]，该通路早期抑癌，晚期促癌。*SLC*16*A*1-*AS*1、*HCG*11 和 *AC*000065.1 都是长链非编码 RNA（lncRNA），参与基因表达调控、信号转导和表观遗传修饰等过程。*SLC*16*A*1-*AS*1 的表达量与肿瘤突变负荷正相关[13]，且在预后不良的 Cluster3 亚型中高表达；*HCG*11 在 Cluster1 至 Cluster3 亚型中表达量递减，与其在骨肉瘤和卵巢癌中低表达预示不良预后的报道一致[14-15]；*AC*000065.1 的功能尚未明确，未来将通过实验手段进一步验证其潜在作用。*AC*131212.2 是脂质分解代谢相关基因 *LRCOL*1 的源序列[16]。癌细胞利用脂质代谢获取能量，生物膜组分和信号分子[17]，调控 *AC*131212.2 表达可以影响肿瘤微环境。*TAF*10 是 RNA 聚合酶Ⅱ转录因子的组成部分，其异常表达可能通过影响细胞转录过程促进肿瘤进展。*SOCS*6 编码的蛋白质是细胞因子信号抑制物，其在慢性粒细胞白血病和红细胞白血病中高表达，而在胃肠道基质瘤中，其低表达与患者预后不良有关[18]。本研究在 PAML 中观察到与胃肠道基质瘤中相似的 *SOCS*6 表达模式，提示可以进一步比对 PAML 与胃肠道基质瘤在治疗策略上的相似性与差异性，探讨 *SOCS*6 作为潜在治疗靶点的可行性。

GO 分析结果显示，不同甲基化亚型的差异表达基因在生物过程层面主要富集于免疫相关生物过程，细胞组分层面主要富集于膜微结构域和运输系统，分子功能层面主要富集于蛋白结合与信号转导。KEGG 分析结果显示，PAML 可能通过造血细胞分化调控等参与肿瘤微环境重塑。这表明 PAML 的甲基化亚型可能反映差异化的免疫微环境，对揭示发病机制和靶点开发具有重要意义。

本研究构建的 PAML 预后得分模型在疾病分层中表现出良好的性能，并证实 PAML 甲基化亚型的预后价值：低风险组可以保守治疗，而高风险组需要尽快进行强化疗等措施。未来可以通过以下策略继续优化模型：整合更多研究数据；结合蛋白质组和 ChIP-seq 数据分析；整合 WHO 分型和 ELN 分型验证模型的准确度并优化精确度；开展功能验证实验解析关键基因的病理作用。此外，各甲基化亚型的特异性分子标志物（如 *SLC*16*A*1-*AS*1、*HCG*11）的表达特征，可通过动态监测指导调整治疗方案，实现"诊断-治疗-预后评估"的全周期管理。

结　论

PAML 进展迅速，及时诊断对 PAML 的治疗十分

重要。PAML的靶向治疗开发被列为儿童肿瘤组骨髓委员会的最高临床研究优先事项[19]。因此，PAML分子层面的分析是临床的迫切需求。本研究整合多组学数据，成功识别出与预后相关的9个关键基因，并构建了基于甲基化亚型的预后得分模型。同时，深入分析了不同甲基化亚型在预后和治疗方案上存在的差异，提示甲基化特征可作为个体化治疗的重要依据。未来可进一步开展功能验证实验，整合更多数据进一步验证和优化模型，为儿童PAML的精准医疗提供理论基础和临床指导。

参 考 文 献

[1] Pelcovits A, Niroula R. Acute myeloid leukemia: a review [J]. Rhode Island Medical Journal, 2020, 103(3): 38-40.

[2] Hayhoe FGJ. Classification of acute leukaemias [J]. Blood Reviews, 1988, 2(3): 186-193.

[3] Marceau-Renaut A, Duployez N, Ducourneau B, et al. Molecular profiling defines distinct prognostic subgroups in childhood AML: a report from the French ELAM02 study group [J]. HemaSphere, 2018, 2(1): e31.

[4] Zhao JC, Agarwal S, Ahmad H, et al. A review of FLT3 inhibitors in acute myeloid leukemia [J]. Blood Reviews, 2022, 52(1): 100905.

[5] Jian J, Yuan C, Ji C, et al. DNA methylation-based subtypes of acute myeloid leukemia with distinct prognosis and clinical features [J]. Clinical and Experimental Medicine, 2023, 23(6): 2639-2649.

[6] 卢娅欣,李康,黄月 . XGBoost算法在类别非平衡高维数据分析中的应用[J]. 中国卫生统计,2021,38(1):21-23.

[7] Verhaak RG, Goudswaard CS, vanPutten W, et al. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance [J]. Blood, 2005, 106(12): 3747-3754.

[8] Falini B, Dillon R. Criteria for diagnosis and molecular monitoring of NPM1-mutated AML [J]. Blood Cancer Discovery, 2024, 5(1): 8-20.

[9] Shih LY, Lin TL, Wang PN, et al. Internal tandem duplication of fms-like tyrosine kinase 3 is associated with poor outcome in patients with myelodysplastic syndrome [J]. Cancer, 2004, 101(5): 989-998.

[10] Wang LQ, Chim CS. DNA methylation of tumor-suppressor miRNA genes in chronic lymphocytic leukemia [J]. Epigenomics, 2015, 7(3): 461-473.

[11] Zhou J, Xu M, Chen ZZ, et al. circ_SPEF2 Regulates the Balance of Treg Cells by Regulating miR-16-5p/BACH2 in Lymphoma and Participates in the Immune Response [J]. Tissue Engineering and Regenerative Medicine, 2023, 20(7): 1145-1159.

[12] Guan XQ, Yuan XN, Feng KX, et al. IGF2BP2-modified UBE2D1 interacts with Smad2/3 to promote the progression of breast cancer [J]. American Journal of Cancer Research, 2023, 13(7): 2948-2968.

[13] Zhao S, Liu M, Zhou H. Exploring the multifaceted role of key lncRNA in glioma: From genetic expression to clinical implications and immunotherapy potential [J]. Heliyon, 2024, 10(17): e36822.

[14] Gu J, Dai B, Shi X, et al. lncRNA HCG11 suppresses human osteosarcoma growth through upregulating p27 Kip1 [J]. Aging (Albany NY), 2021, 13(17): 21743-21757.

[15] Chen X, Yang Y, Sun J, et al. LncRNA HCG11 represses ovarian cancer cell growth via AKT signaling pathway [J]. Journal of Obstetrics and Gynaecology Research, 2022, 48(3): 796-805.

[16] Alliance of Genome Resources. LRCOL1 [EB/OL]. (2024-03-24) [2025-09-23]. https://www. alliancegenome. org/gene/HGNC: 44160.

[17] Bian X, Liu R, Meng Y, et al. Lipid metabolism and cancer [J]. Journal of Experimental Medicine, 2020, 218(1): e20201606.

[18] Ouyang J, An T, Wang Y, et al. Down-regulation of SOCS6: an unfavorable prognostic factor for gastrointestinal stromal tumor proven by survival analysis [J]. Diagnostic Pathology, 2021, 16 (1): 113.

[19] Gamis AS, Alonzo TA, Perentesis JP, et al. Children's Oncology Group's 2013 blueprint for research: acute myeloid leukemia [J]. Pediatric Blood & Cancer, 2013, 60(6): 964-971.

（责任编辑：张　悦）