

· 论著 · 胸部 ·

孟德尔随机化分析：发现新的哮喘生物标志物

詹含韵 林 敏 王世彪*

福建医科大学妇产临床医学院儿童重症医学科(福建 福州 350000)

【摘要】目的 儿童哮喘(CA)时至今日仍是一种无法治愈的疾病,严重影响儿童健康。而CA患病率可能与鞘脂代谢失调(SM)有关,但其具体作用机制尚未完全了解。因此,本研究旨在利用转录组数据和孟德尔随机化(MR)分析来评估SM相关生物标志物对CA的影响。方法 本研究使用公共数据库中CA相关转录组数据,通过差异表达分析、加权基因共表达网络分析(WGCNA)和MR分析获得候选特征基因。通过两种机器学习(ML)算法筛选与CA相关的特征基因,并通过基因集富集分析(GSEA)等方法探索生物标志物影响CA的潜在机制。结果 PRSS33、NMRK1和GZMK被确定为CA的生物标志物。功能富集和探索发现NMRK1和GZMK共富集的KEGG通路单纯疱疹病毒1型感染相关, NMRK1和PRSS33共富集的KEGG通路与神经活性配体-受体相互作用相关。结论 本研究发现PRSS33、NMRK1和GZMK是CA中与SM相关的生物标志物,为儿童哮喘的病因和未来改进治疗和预防策略提供重要见解。

【关键词】儿童哮喘;鞘脂类代谢;孟德尔随机化;生物标志物

【中图分类号】R725.6; R34; R394.3

【文献标识码】A

DOI:10.3969/j.issn.1009-3257.2025.8.022

Mendelian Randomization Analysis: Discovery of New Asthma Biomarkers

ZHAN Han-yun, LIN Min, WANG Shi-biao*

Department of Pediatric Intensive Care Unit, College of Clinical Medicine for Obstetrics&Gynecology and Pediatrics, Fujian Medical University, Fuzhou 350000, Fujian Province, China

Abstract: Objective Childhood asthma (CA) remains an incurable disease, significantly impacting the health of children. Dysregulation of sphingolipid metabolism (SM) has been linked to CA risk, but the specific role is not yet fully understood. Hence, this study aimed to assess the impact of SM-related biomarkers on CA using transcriptome data and Mendelian randomization (MR) analysis. **Methods** In the study, candidate feature genes were obtained by differential expression analysis, weighted gene co-expression network analysis (WGCNA) and MR Analysis. Subsequently, two machine learning (ML) algorithms were proceeded to screen for feature genes related to CA. Ulteriorly, gene set enrichment analysis (GSEA) and other methods were implemented to explore the potential mechanism by which biomarkers influenced CA. **Results** PRSS33, NMRK1, and GZMK were identified as biomarkers for CA. Their functional enrichment and exploratory discoveries revealed that the KEGG pathways co-enriched by NMRK1 and GZMK are associated with herpes simplex virus type 1 infection, while the KEGG pathways co-enriched by NMRK1 and PRSS33 are related to neuroactive ligand-receptor interactions. **Conclusion** PRSS33, NMRK1, and GZMK were found as biomarkers linked to SM in CA, providing vital insights into its causes and potential for improved treatments and prevention strategies in the future.

Keywords: Childhood Asthma; Sphingolipid Metabolism; Mendelian Randomization; Biomarkers

1 背景

哮喘是一种气道慢性炎症性疾病^[1]。儿童哮喘(CA)发病率高,占发达国家卫生保健预算的1%~2%,具有沉重的社会经济负担。目前,CA的诊断主要依靠临床症状及肺功能、影像学检查等传统手段^[19-20]。调查显示,我国城市儿童哮喘的漏诊率高达30%,目前对于6岁以下儿童哮喘尚无明确的诊断方法^[2]。因此,亟需开发更有效的诊断工具。而鞘脂是一类含有鞘氨醇骨架的两亲性脂质。最新研究发现,不同亚类鞘脂在CA及其危险因素中发挥不同的作用^[3]。孟德尔随机化(MR)可以通过遗传数据评估疾病与暴露因素之间的因果关系。本研究基于CA相关转录组数据,通过加权基因共表达网络分析(WGCNA)、MR等生物信息学方法,鉴定CA中SM相关的生物标志物,评估和推断SMRGs在CA中的作用。有望在未来帮助开发更有效的治疗方法和预防策略。

2 方法

2.1 数据收集 本研究使用Gene Expression Omnibus数据库。其中GSE35571数据集(平台:GPL570)包括67名哮喘和64名非哮喘儿童(NC)微阵列测序数据。GSE40888数据集(平台:GPL6244)包括65名过敏性哮喘(AA)和40名正常人样本。从<http://www.innatedb.com>获得97个鞘脂代谢相关基因(SMRGs)^[4]。CA的全基因组关联研究(GWAS)数据通过IEU OpenGWAS数据库收集。共检索到16379865个单核苷酸多态性(SNPs)。

2.2 候选基因鉴定 首先,执行R软件包“limma”(版本3.52.4)^[5]获取GSE35571数据集中CA组和NC组之间的差异表达基因(DEGs) ($|\text{Log}_2\text{fold-change (FC)}| > 0, P < 0.05$)。通过火山图和热图分别显示前10个显著DEGs和所有DEGs。随后,通过R软件包“WGCNA”(版本1.70.3)^[6]进行加权基因共表达

【第一作者】詹含韵,女,住院医师,主要研究方向:儿科、重症医学科。E-mail: zhy49868449@163.com

【通讯作者】王世彪,男,主任医师,主要研究方向:儿科、呼吸及危重症医学。E-mail: wsb15005904839@163.com

网络分析(WGCNA), 获得与SM相关的关键模块基因。使用SMRGs为基因集, 计算GSE35571数据集各样本的SMRGs评分, 采用Wilcoxon秩和检验估计CA组与NC组的SMRGs评分差异($P<0.05$)。对GSE35571数据集中的样本进行聚类分析, 去除其中异常值。我们把平均连通性趋于0作为理想软阈值($R^2>0.90$), 计算基因间邻接关系, 据此引入基因间不相似系数, 得到基因间系统聚类树。然后根据混合动态树切割准则, 设置每个基因模块中最小的基因数为30, 并将MEDissThres设置为0.3, 得到聚类模块。选择与SMRGs评分显著相关的聚类模块作为关键模块($P<0.05$)。DEGs的交叉基因和关键模块基因作为候选基因。

2.3 MR分析 我们以候选基因对应的表达量性状位点(eQTL)为暴露因子, CA为结果。使用R软件包“TwoSampleMR”(版本0.5.6)^[7]进行MR分析。以 $P<5\times 10^{-6}$ 的阈值选择与候选基因显著相关的SNPs。然后对所有SNP进行F检验。使用逆方差加权法(IVW)($P<0.05$)对最终选择的暴露因子进行MR研究。使用比值比(OR)、95%置信区间(95%CI)和P值确定暴露与结果变量之间的因果关系。OR>1为危险因素。OR<1则为保护因素。通过散点图、森林图和漏斗图显示结果。通过敏感性分析评估MR研究的可靠性。采用Steiger检验评估暴露因素与CA之间的方向性, 最后筛选符合IVW方法($P<0.05$)且通过敏感性分析的基因, 将表达上调的危险因素和表达下调的保护因素作为候选特征基因。

2.4 筛选生物标志物、列线图建模及功能富集分析 使用R软件包“glmnet”(版本4.1-4)^[8]构建最小绝对收缩和选择算子(LASSO)回归模型。选取LASSO模型中最拟合的特征基因1(family=binomial, n-fold=10)。同时, 通过R包‘caret’(版本6.0-91)^[9]中的‘rfe’函数构建支持向量机-递归特征消除(SVM-RFE)模型, 筛选权重向量最大的特征基因2。特征基因1和2相交得到特征基因。将CA和NC中表达水平差异且表达趋势一致的基因鉴定为生物标志物($P<0.05$)。通过R包‘rms’(版本5.4-1)^[10]建立列线图评估其诊断性能。生成标定曲线和受试者工作特征(ROC)曲线, 评价列线图预测准确度[ROC曲线下面积(AUC)>0.7]。

对GSEA执行“clusterProfiler”软件包。计算生物标志物与GSE35571数据集中所有基因之间的相关系数并进行排

序, 生成执行GSEA的基因列表($\text{adj.}P<0.05$)。

本研究采用R软件(Ver4.2.0)进行生物信息学分析。两组间差异采用Wilcoxon秩和检验, $P<0.05$ 为有统计学意义。

3 结果

3.1 共鉴定出24个候选特征基因 在GSE35571数据集中共获得555个DEGs, 其中396个在CA中过表达, 159个在CA中下调表达(图1A~1B)。CA组SMRGs评分高于NC组($P<0.05$)。聚类分析显示GSE35571中没有异常样本(图1C)。当垂直无标度 R^2 超过阈值0.90且平均连接性趋于平稳时, 选择理想的软阈值为9。根据动态树切割标准得到24个模块, 对它们进行组合后生成17个聚类模块。按模块相关性绝对值排序, 其中MEroyalblue($\text{cor}=0.76$, $P<0.001$), MEbrown($\text{cor}=-0.42$, $P=0.001$), MERed($\text{cor}=0.40$, $P=0.002$), 和MEblack($\text{cor}=-0.81$, $P=0.002$)成为关键模块(图1D)。组合4个模块共获得14192个关键模块基因。将DEGs与关键模块基因交叉, 共筛选出252个候选基因。

共141个SNP对应252个候选基因纳入MR分析。其中PRSS33[比值比(OR)=1.526, $P=0.007$, 95%可信区间(CI): 1.112~2.075]是CA的危险因素, NMRK1(OR=0.977, $P=0.035$, 95%CI:0.957~0.998)和GZMK(OR=0.828, $P=0.0006$, 95%CI:0.744~0.922)是CA的保护因素。以PRSS33基因为例详细描述MR结果。根据IVW方法, PRSS33是CA的危险因素, OR值为1.526, P值为0.007(表1), 散点图的斜率大于0说明了PRSS33的危险性(IVW方法), PRSS33对CA的MR效应量在森林图上大于0。如散点图所示, PRSS33对于CA大概率不受混杂因素影响(图2A)。漏斗图上均匀分布的点证明MR分析符合孟德尔第二定律(图2B)。异质性和水平多效性检验表明, 部分暴露因素与CA数据集之间存在异质性($P<0.05$), 所有暴露因素与CA数据集之间不存在水平多效性($P<0.05$)(表2)。通过单独消除SNP的过程, 未发现对CA敏感的SNP, 表明单个SNP对因果关系无重大影响。共有44个基因符合IVW法($P<0.05$)和水平多效性($P<0.05$)。结合差异表达分析, 共筛选出24个上调的危险因素和下调的安全因素。Steiger检验评估表明它们对CA的因果影响不受反向因素的干扰, 因此确定其为候选特征基因(图3)。

表1 PRSS33 的MR结果

结果	暴露	方法	nSNP	b	se	P值	OR
eqtl-a-ENSG00000106733	NMRK1	MR Egger	108	-0.589	0.546	0.287	0.555
eqtl-a-ENSG00000106733	NMRK1	Weighted median	108	0.314	0.14	0.025	1.369
eqtl-a-ENSG00000106733	NMRK1	Simple mode	108	0.423	0.157	0.007	1.526
eqtl-a-ENSG00000106733	NMRK1	IVW	108	0.977	0.394	0.008	2.656
eqtl-a-ENSG00000106733	NMRK1	Weighted mode	108	0.869	0.394	0.034	2.385

表2 5个候选基因的遗传异质性和水平多效性分析结果

基因	暴露	方法	Q值	pleio_P
TFF3	eqtl-a-ENSG00000160180	IVW	0.00107809	0.09547
PRSS33	eqtl-a-ENSG00000103355	IVW	3.90E-22	0.061309
MYRIP	eqtl-a-ENSG00000170011	IVW	1	0.579532
CEBPE	eqtl-a-ENSG00000092067	IVW	8.81E-14	0.233397
GALM	eqtl-a-ENSG00000143891	IVW	0.99979312	0.059189

3.2 基于生物标志物的列线图可预测CA风险 当最小 λ 值为0.03126289时,确定了14个与CA最密切相关的特征基因1。SVM-RFE模型预测精度达到峰值时,共筛选出19个特征基因2。两者重叠后共得12个特征基因。特征基因中PRSS33在CA中过表达,而NMRK1和GZMK表达不足,这3个基因在两个数据集集中的表达趋势一致。故PRSS33、NMRK1和GZMK被确定为CA的生物标志物(图4A~4B)。根据Log2转换后生物标志物的表达建立列线图,CA的风险随着总分的增加而增加(图4C)。列线图C指数(ROC-AUC)为1.00,校准曲线验证了列线图

的精确性。同时,列线图的AUC为0.72,表明该模型可以准确预测CA的概率。

3.3 生物标志物可能参与细胞蛋白质合成 GSEA显示生物标志物在GO条目中共同富集,包括细胞质核糖体、核糖体结构成分、核糖体亚基,以及相同的KEGG途径,即核糖体。NMRK1和GZMK共富集的KEGG通路为单纯疱疹病毒1型感染相关,NMRK1和PRSS33共富集的KEGG通路为神经活性配体-受体相互作用相关(图5A~5B)。故生物标志物可能参与细胞内蛋白质合成过程。

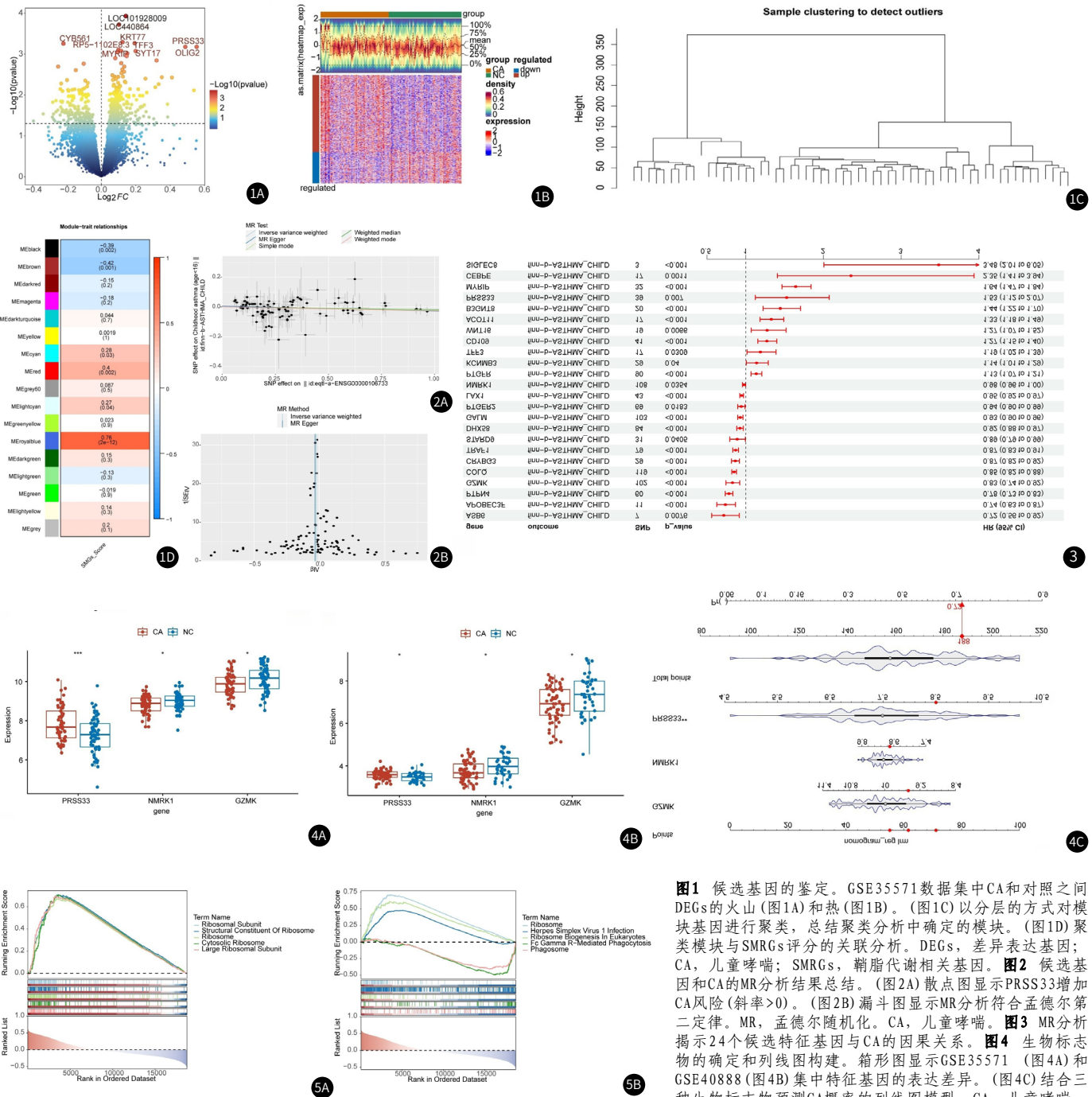


图1 候选基因的鉴定。GSE35571数据集中CA和对照之间DEGs的火山(图1A)和热(图1B)。(图1C)以分层的方式对模块基因进行聚类,总结聚类分析中确定的模块。(图1D)聚类模块与SMRGs评分的关联分析。DEGs, 差异表达基因; CA, 儿童哮喘; SMRGs, 鞘脂代谢相关基因。**图2** 候选基因和CA的MR分析结果总结。(图2A)散点图显示PRSS33增加CA风险(斜率>0)。(图2B)漏斗图显示MR分析符合孟德尔第二定律。MR, 孟德尔随机化。CA, 儿童哮喘。**图3** MR分析揭示24个候选特征基因与CA的因果关系。**图4** 生物标志物的确定和列线图构建。箱图显示GSE35571(图4A)和GSE40888(图4B)集中特征基因的表达差异。(图4C)结合三种生物标志物预测CA概率的列线图模型。CA, 儿童哮喘。**图5** 生物标志物功能探索。以GSEA为例显示GZMK(图5A)所富集的GO条目, 以及GZMK(图5B)参与的KEGG通路。GSEA, 基因集富集分析; GO, 基因本体; KEGG, 京都基因与基因组百科全书。

4 讨论

CA作为儿科疾病的一大类别,发病率高,误诊率高,幼儿诊断困难。目前,CA发生发展的机制尚不清楚。已有研究揭示了哮喘可能与鞘脂代谢有关,但对其基因调控知之甚少。利用MR筛选靶基因可以有效降低混杂因素的影响。本研究通过结合MR和基因转录组数据,鉴定出3个与鞘脂代谢相关的生物标志物(PRSS33、NMRK1、GZMK),发现它们可通过一些可能的机制影响CA。这对未来开发更有效的治疗和预防策略具有重要意义。

蛋白酶丝氨酸33(PRSS33)是一种编码丝氨酸蛋白酶的mRNA,在嗜酸性粒细胞中表达。活化的嗜酸性粒细胞经SIgA和GM-CSF激活后,可通过细胞表面PRSS33的表达诱导成纤维细胞外基质蛋白的合成,这部分解释了嗜酸性粒细胞浸润在II型炎症气道重构中的作用^[11]。

烟酰胺核苷激酶1(NMRK1/NRK1)可编码蛋白质,其翻译的烟酰胺核苷激酶可催化烟酰胺核苷酸(NMN)合成烟酰胺腺嘌呤二核苷酸(NAD⁺)^[12]。文献表明,NAD依赖性脱乙酰酶如SIRT1参与多种生物过程。其具有共同的NAD⁺结合和催化核心结构域。而SIRT1在哮喘中起保护作用^[13]。Colley等人发现哮喘患者的外周血单核细胞(PBMCs)脱乙酰酶活性和SIRT1表达降低。SIRT1活性的降低与PBMCs中IL-4和总IgE的转录呈负相关。人体细胞的体外研究表明,抑制SIRT1可以促进T细胞中2型辅助性T(Th2)样表型的出现。从机制上讲,SIRT1可以通过使GATA3去乙酰化来降低GATA3的活性及其与应答元件的结合,从而降低IL-4的表达和Th2细胞的分化^[14]。由此推断,NAD⁺可能与哮喘呈负相关,而催化NAD⁺产生的NMRK1可能是哮喘的保护因子。MR分析显示NMRK1是CA的保护因子,这些结果支持NMRK1可能作为CA的潜在治疗靶点或生物标志物的假设。

颗粒酶K(GZMK)的产物是一种丝氨酸蛋白酶,存在于细胞毒性淋巴细胞(主要是细胞毒性T淋巴细胞(CTL)和自然杀伤细胞(NK))的细胞质颗粒中^[15]。CD4⁺T2细胞在协调哮喘发病机制中起核心作用,其气道中存在的CD4⁺T细胞具有显著的异质性。与轻度哮喘患者(13%)相比,重度哮喘患者的CD4⁺T细胞亚群(表达CD103)显著增加,几乎占男性重度哮喘患者气道CD4⁺T细胞总数的65%。Y亚组富含与T细胞受体激活(HLA-DRB1,HLA-DPA1)和细胞毒性(GZMB和GZMA)相关的转录本,并在刺激后表达高水平的转录本,其可编码促炎性非T2细胞因子,可持续促进气道炎症和气道重塑^[16]。然而,GZMK在哮喘中的作用尚未被提及。本研究结果显示GZMK在CA的发生发展中起保护作用。

表达趋势上,基因差异分析显示CA组PRSS33上调,NMRK1和GZMK下调。MR中PRSS33对CA具有风险属性,推测其对CA的病理发展具有促进作用,NMRK1的OR为0.977,P值为0.035,表明其是CA的保护因子,散点图中的斜率小于0,也表明NMRK1具有保护作用。根据IVW方法,GZMK的OR为0.828,P值为0.0006,说明它对CA起保护作用,散点图的斜率小于0,也说明GZMK具有保护作用。在森林图中,GZMK对CA的MR效应量大于0。此外,从散点图可以看出,NMRK1对CA风险的影响大概率不受混杂因素的影响。这些结果表明

PRESS33、MNRK1和GZMK在CA中发挥关键作用,可能是CA潜在的关键治疗靶点。

通过GSEA评估显示,三种生物标志物在GO条目/KEGG途径中显著富集,包括细胞质核糖体、核糖体结构组分、核糖体亚基。核糖体是细胞内蛋白质合成的场所。有报道称,PRSS33诱导成纤维细胞外基质蛋白的合成,促进哮喘患儿气道重塑^[11]。本研究显示NMRK1和GZMK富集的KEGG途径与HSV-1感染有关。一项研究表明,特应性哮喘和AR儿童的HSV-1IgG阳性率(56.8%)显著高于同年龄的非特应性儿童(30.4%)($P<0.001$)。哮喘儿童的特应性状态可能与AR和HSV-1感染有关^[17]。另一项研究提示针对各种感染源的低水平抗体是特应性哮喘和AR的独立危险因素^[18]。

综上所述,我们通过MR结合基因转录组数据,鉴定出PRSS33、NMRK1、GZMK三个SMRGs作为生物标志物,通过功能探索及基因富集分析发现其可能通过参与细胞内蛋白质合成过程影响CA的发生发展。这不仅有助于了解鞘脂代谢与CA之间的潜在联系,也为未来开发与CA相关的生物标志物或临床预测工具奠定了基础。

参考文献

- [1] 辛晓峰,施毅. 支气管哮喘和慢性阻塞性肺疾病的相似关系[J]. 中华结核和呼吸杂志, 2006, 29(6): 428-430.
- [2] 中华医学会儿科学分会呼吸学组, 编辑委员会. 中华儿科杂志. 儿童支气管哮喘诊断与防治指南(2016年版)[J]. 中华儿科杂志, 2016, 54(3): 167-181.
- [3] 郭成林, 常春, 张丽娟, 等. 鞘脂在哮喘发病机制中的作用[J]. 国际呼吸杂志, 2019, 39(5): 367-372.
- [4] Chi H, Peng G, Yang J, et al. Machine learning to construct sphingolipid metabolism genes signature to characterize the immune landscape and prognosis of patients with uveal melanoma[J]. Front Endocrinol (Lausanne), 2022, 13: 1056310.
- [5] Ritchie M E, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies[J]. Nucleic Acids Res, 2015, 43(7): e47.
- [6] Langfelder P, Horvath S. Wgcna: an R package for weighted correlation network analysis[J]. BMC Bioinformatics, 2008, 9: 559.
- [7] Lu F, Wu B, Wang Y. Mendelian randomization indicates that atopic dermatitis contributes to the occurrence of diabetes[J]. BMC Med Genomics, 2023, 16(1): 132.
- [8] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent[J]. J Stat Softw, 2010, 33(1): 1-22.
- [9] Lopez-Diaz J, Mendez-Gonzalez J, Lopez-Serrano P M, et al. Dummy regression to predict dry fiber in agave lechuguilla torr. In two large-scale bioclimatic regions in Mexico[J]. PLoS One, 2022, 17(9): e274641.
- [10] Liu T T, Li R, Huo C, et al. Identification of cdk2-related immune forecast model and cerna in lung adenocarcinoma, a pan-cancer analysis[J]. Front Cell Dev Biol, 2021, 9: 682002.
- [11] Toyama S, Okada N, Matsuda A, et al. Human eosinophils constitutively express a unique serine protease, prss33[J]. Allergol Int, 2017, 66(3): 463-471.
- [12] Ratajczak J, Joffraud M, Trammell S A, et al. Nrkl controls nicotinamide mononucleotide and nicotinamide riboside metabolism in mammalian cells[J]. Nat Commun, 2016, 7: 13103.
- [13] 李炎玉, 麦宗炯, 文晓霞, 等. SIRT1在支气管哮喘中的研究进展[J]. 临床肺科杂志, 2018, 23(4): 745-747.
- [14] Colley T, Mercado N, Kunori Y, et al. Defective sirtuin-1 increases il-4 expression through acetylation of gata-3 in patients with severe asthma[J]. J Allergy Clin Immunol, 2016, 137(5): 1595-1597.
- [15] 郭玉明. 颗粒酶K介导肿瘤细胞杀伤机理研究[D]. 中国科学院生物物理研究所, 2008.
- [16] Herrera-De L M S, Ramirez-Suastegui C, Mistry H, et al. Cytotoxic CD4⁺ tissue-resident memory T cells are associated with asthma severity[J]. Med (NY), 2023, 4(12): 875-897.
- [17] Igde M, Igde F A, Yazici Z. Herpes simplex type 1 infection and atopy association in Turkish children with asthma and allergic rhinitis[J]. Iran J Allergy Asthma Immunol, 2009, 8(3): 149-154.
- [18] Janson C, Asbjornsdottir H, Birgisdottir A, et al. The effect of infectious burden on the prevalence of atopy and respiratory allergies in Iceland, Estonia, and Sweden[J]. J Allergy Clin Immunol, 2007, 120(3): 673-679.

(收稿日期: 2025-03-20)

(校对编辑: 翁佳鸿、姚丽娜)