网络出版时间:2025 - 10 - 28 13:56:21 网络出版地址:https://link.cnki.net/urlid/34.1065.R.20251027.1506.003

◇基础医学研究◇

丙烯醛相关基因的肺癌预后预测模型

冯祎婷[1,2],任亮亮[2],娄丽娟[2],沈玉先[1],姜 颖[1,2]

[[1] 安徽医科大学基础医学院生物化学与分子生物学教研室,合肥 230032;[2] 医学蛋白质组全国重点实验室,北京蛋白质组研究中心,国家蛋白质科学中心(北京),北京生命组学研究所,北京 102206]

摘要 目的 通过生物信息学方法构建并验证基于丙烯醛相关基因的肺癌预后预测模型。**方法** 利用GEO数据库获取肺癌数据集GSE30219和GSE68465,同时从CTD数据库筛选丙烯醛相关基因集。首先,在GSE30219数据集中筛选癌与癌旁的差异表达基因(DEGs),与丙烯醛基因集取交集,获得候选基因。随后,采用基因集变异分析(GSVA)以评估其功能变化特征。基于STRING数据库构建蛋白质互作(PPI)网络,筛选核心枢纽基因(Hub Genes)。采用SVM-RFE和LASSO-Cox回归分析构建基于丙烯醛相关基因的肺癌预后预测模型,并使用GSE68465数据集进行独立验证。通过CIBERSORT方法分析高低风险组的免疫细胞浸润特征,同时对高低风险组的DEGs进行功能富集分析,进一步揭示基于丙烯醛相关基因的肺癌预后的潜在分子机制。**结果** 共筛选出361个丙烯醛相关的肺癌DEGs,进一步确定7个关键基因用于模型构建。Kaplan-Meier生存分析显示,高风险组患者的生存率显著低于低风险组($P<0.000\ 1$)。ROC曲线分析结果表明,该模型具有良好的预测性能。此外,免疫浸润分析显示,风险评分与多种免疫细胞亚群密切相关,揭示了丙烯醛相关基因在肺癌免疫微环境中的潜在作用。**结论** 基于丙烯醛相关基因的肺癌预后模型在肺癌的预后中展现出显著的应用价值,为揭示丙烯醛在肺癌发生与发展的潜在机制提供新的依据。

关键词 丙烯醛;肺癌;环境污染物;生物信息学;机器学习;预后模型

中图分类号 R 734.2

文献标志码 A **文章编号** 1000 - 1492(2025)11 - 1985 - 11

doi:10.19405/j.cnki.issn1000 - 1492.2025.11.001

丙烯醛是一种具有高度反应性、挥发性和持久性的环境污染物,广泛存在于工业排放、各种燃烧过程(如烟草烟雾、汽车尾气、厨房油烟、木材燃烧等)以及农业活动(如农药的使用)中。因此,丙烯醛在城市空气中普遍存在,尤其在雾霾严重的地区,对生物体及生态系统构成潜在危害[1]。

现有研究[2]表明,丙烯醛与多种人类疾病密切相关,国际癌症研究机构评估认为其可能具有人类致癌性,并与多种癌症存在潜在关联。由于主要通过吸入暴露,丙烯醛与哮喘、慢性阻塞性肺病及肺癌等呼吸系统疾病显著相关[2]。其中,丙烯醛与肺癌之间的潜在联系已引起了广泛关注。体外研究[3]表明,丙烯醛可诱导DNA加合物的形成,引发遗传毒性反应;动物实验进一步证实其可剂量依赖性地增加肺部肿瘤发生概率,机制涉及DNA损伤及修复抑制过程。此外,Liu et al[4]对大鼠/小鼠模型的荟萃分析也表明,丙烯醛对呼吸系统具有毒性,可能诱发肺癌及其他呼吸系统病变。因此,该研究通过构建丙烯醛相关肺癌预后模型,旨在探讨丙烯醛与肺癌患者预后的潜在联系,为评估肺癌风险、个体化治疗及公共卫生策略提供理论依据,并加强与人类肺癌相关的环境污染物研究。

1 材料与方法

1.1 数据收集 本研究从GEO(Gene Expression Omnibus,http://www.ncbi.nlm.nih.gov/geo/)数据库中下载了肺癌队列的基因芯片数据集GSE30219和GSE68465以及相应的样本临床信息。其中GSE68465作为外部验证集。本研究还利用CTD(Comparative Toxicogenomics Database,http://ctdbase.org)数据库,该数据库是一个专注于毒物及环境污染物影响人类健康的公开资源,提供了关于化学 - 基因/蛋白相互作用、化学 - 疾病关系以及基因 - 疾病关系的人工整理信息。基于该数据库下载了与丙烯醛相关的2 515个基因,筛选标准为以人类

2025 - 09 - 30 接收

基金项目:国家重点研发计划项目(编号:2020YFE0202200)

作者简介:冯祎婷,女,硕士研究生;

姜 颖,女,博士,研究员,博士生导师,通信作者,E-mail:jiangying@ncpsb.org.cn;

沈玉先,女,博士,教授,博士生导师,通信作者,E-mail:shenyx@ahmu.edu.cn

为研究对象的基因数据。

1.2 肺癌差异表达基因筛选 GEO 数据集的原始数据通过 R 软件(版本 4.4.1)进行预处理。对于存在重复的 Gene Symbol,仅保留其表达量的均值。利用“limma”包对表达矩阵进行四分位数归一化处理,并以差异表达倍数的绝对值($|\log_2 FC|>1$)和 Benjamini-Hochberg 矫正后 $P<0.05$ 为筛选标准,鉴定出正常肺组织与肺癌组织之间的差异表达基因(differentially expressed genes, DEGs)。通过绘制火山图对筛选出的 DEGs 进行可视化展示。

1.3 基于丙烯醛相关基因的肺癌交集基因筛选及其 GSVA 通路富集分析 通过 Sangerbox 在线工具,将训练集 GSE30219 筛选出的 DEGs 与丙烯醛相关的 2 515 个基因取交集,并绘制韦恩图展示交集结果。使用 R 语言“pheatmap”包对交集基因进行热图可视化。使用 R 语言的“GSVA”包对交集基因进行 GSVA(gene set variation analysis)通路富集分析。选择物种为人类(Homo sapiens),并采用 MSigDB 基因集中的 C2 基因集(即“category = 'C2' ”)进行注释。通过 GSVA 计算每条通路的富集分数,并利用“limma”包对富集结果进行差异分析,筛选出在正常肺组织与肺癌组织之间显著差异表达的通路,并将富集分析结果进行可视化。

1.4 蛋白质互作(Protein-protein interactions, PPI)网络构建与枢纽基因筛选 利用筛选得到的交集基因,通过 STRING(https://string-db.org/)构建 PPI 网络,并将置信度设置为 0.4,然后导入 Cytoscape 软件(版本 3.10.2)进行可视化分析,并通过 CytoHubba 插件计算基因的中间度,根据 Degree 值对基因进行排序,最终筛选出枢纽基因(Hub genes),并将这些枢纽基因联合起来。

1.5 基于机器学习构建预后风险模型 本研究采用 SVM-RFE(支持向量机-递归特征消除)和 Lasso-Cox 回归两种机器学习方法构建肺癌预后风险模型。首先,通过 R 语言的“e1071”包实现 SVM-RFE 算法,模型参数设置如下:SVM = rfeControl(function = caretFuncs, method = “cv”, number = 10, methods = “svmLinear”),采用 10 折交叉验证以筛选特征集合,选择纵轴上交叉验证最低的点对应的特征作为潜在生物标志物;然后,基于 R 语言的“glmnet”包,整合患者的生存时间、生存状态和基因表达数据,利用 Lasso-Cox 回归方法进行分析。此外还设置了 10 折交叉验证,建立可靠和稳健的模型。风险评分公式如下:风险评分 = $\sum_{i=1}^{\infty}(coef_i \times Exp_i)$。其中,coef 表示回归系数,Exp 表示基因的表达值。利用 R 语言的“maxstat”包计算最优截断值,并根据该截断值将患者分为高风险组和低风险组。随后,使用“Survminer”包绘制 Kaplan-Meier 生存曲线,若 log-rank 检验结果的 P 值小于 0.05,则认为高、低风险组患者的预后差异具有统计学显著性。时间依赖的接受者操作特性曲线(receiver operating characteristic, ROC)采用 R 语言的“pROC”包绘制,并计算曲线下面积(area under curve, AUC),以量化模型的预测准确性和稳健性。

1.6 预后风险模型的验证与评估 为验证所构建的基于丙烯醛相关基因的肺癌预后风险模型的稳健性,将基于 GSE30219 数据集生成的风险评分公式应用于 GSE68465 数据集,对于存在重复的 Gene Symbol,仅保留其表达量的均值。利用“limma”包对表达矩阵进行四分位数归一化处理,并以差异表达倍数的绝对值($|\log_2 FC|>1$)和 Benjamini-Hochberg 矫正后 $P<0.05$ 为筛选标准,鉴定出正常肺组织与肺癌组织之间的 DEGs,而后计算验证集中患者的风险评分。通过 R 包“Survminer”、“pROC”等绘制 Kaplan-Meier 生存曲线和 ROC 曲线评估预后模型的预测能力。

1.7 高低风险患者的临床特征与通路富集分析 基于 GSE30219 数据集,结合高低风险患者的临床特征,分析所筛选的 7 个关键基因与吸烟状态、组织学类型、TNM 分期、年龄和性别等临床特征之间的关联。对于关联性较强的临床特征(吸烟状态与组织学类型),采用卡方检验进行分析,并使用 R 语言的“ggplot2”包绘制柱状图,以直观展示基因表达与临床特征之间关系。然后利用 R 语言“GSVA”包将 GSE30219 数据集的高低风险人群分组进行 GSVA 通路富集。使用“limma”包对富集的通路进行差异分析,筛选出在低风险组和高风险组之间差异显著的通路。对筛选出的差异通路进行可视化展示,以进一步揭示丙烯醛相关的肺癌 DEGs 在不同风险人群中调控的分子机制及潜在功能通路。

1.8 免疫浸润分析及其相关性分析 采用 CIBERSORT 算法对 GSE30219 数据集的免疫细胞浸润水平进行分析,计算每种免疫细胞在高低风险组样本中的比例以及不同免疫细胞之间的相关性。利用 R 包“corrplot”生成相关性热图,直观呈现各免疫细胞之间的相互关联。使用 R 包“ggplot2”绘制箱型图,对 22 种免疫细胞在高低风险组中的表达差异进行可视化展示。使用 R 包“ggstatsplot”,采用 Spearman

相关性分析方法，评估浸润性免疫细胞之间的相关性，为揭示免疫细胞在丙烯醛相关肺癌患者预后中的潜在作用机制提供依据。

2 结果

2.1 丙烯醛相关肺癌差异表达基因的筛选及通路富集分析 采用 GSE30219 作为训练集，GSE68465 作为外部验证集。整体研究路线如图 1。对 GSE30219 数据集中的样本进行整理和预处理，其中包含 245 例肺癌患者样本和 14 例正常样本。在标准化处理后，共鉴定出 2 834 个 DEGs。如图 2A 所示，通过火山图对筛选结果进行可视化展示。

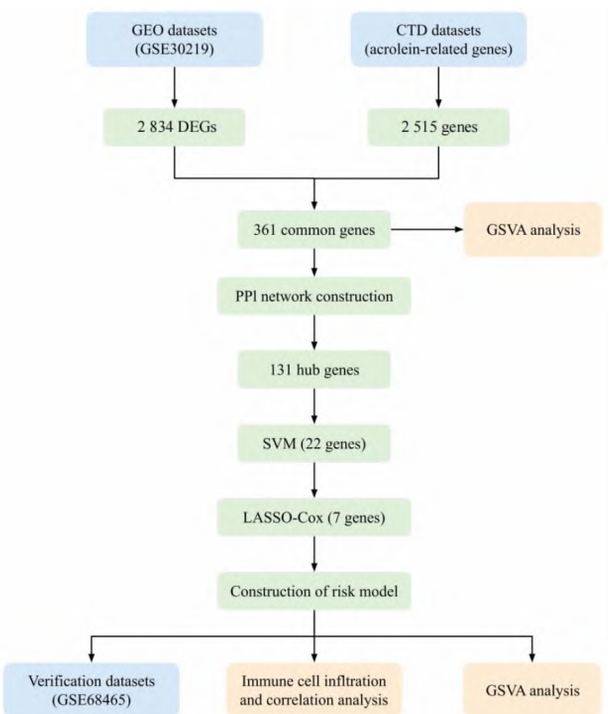

图 1 研究流程图

Fig. 1 Flow chart of the study

此外，通过 CTD 数据库筛选得到与丙烯醛相关的基因共 2 515 个。将 GSE30219 数据集中筛选的肺癌 DEGs 与丙烯醛相关基因取交集，获得 361 个交集基因（图 2B），利用热图对交集基因在不同样本中的表达情况进行可视化展示（图 2C）。随后，对这 361 个交集基因进行通路富集分析，计算每条通路的富集分数，再对富集的通路进行差异分析，筛选出 P 值 <0.05 的显著通路，并对结果进行可视化展示（图 2D）。选取了 20 条在肺癌中显著或抑制的信号通路，激活了如 TGFβ 信号通路、NOTCH 信号通路、p53 信号通路、细胞凋亡、DNA 修复及 DNA 损伤等信号通路，抑制了脂肪酸代谢、BCR 等信号通路。这些分析结果表明，丙烯醛相关基因在肺癌的发生发展中可能发挥关键作用，调控了 NOTCH、p53、细胞凋亡、TGFβ、脂肪酸代谢等信号通路。

2.2 PPI 网络的构建和枢纽基因的筛选 将筛选得到的 361 个交集基因导入 STRING 数据库进行在线分析，构建 PPI 网络。生成的 PPI 网络由 359 个节点和 2 925 条边组成（图 3A）。然后使用 Cytoscape 软件中的 CytoHubba 插件对 PPI 网络进行分析，并使用 Degree 算法对基因的重要性进行排序，最终筛选出 131 个枢纽基因（图 3B），并绘制 131 个基因的表达量热图（图 3C）。

2.3 肺癌中丙烯醛相关基因的预后风险预测模型的构建 利用 SVM-RFE 算法对 131 个枢纽基因进行筛选，鉴定出 22 个特征基因（图 4A）。随后，采用 Lasso-Cox 回归分析对特征基因进一步优化，构建了包含 7 个特征基因的预后风险模型（图 4B），其风险评分公式如下：风险评分 = 0.009 7 × CENPF + 0.015 0 × EXO1 + 0.280 4 × GAPDH + 0.058 6 × KIF18B + 0.050 9 × LMNB1 + 0.179 9 × MIF + 0.102 0 × TRIP13。使用 R 语言的“maxstat”包计算得到最优截断值为 7.675，据此将患者分为高风险组和低风险组。进一步分析患者的风险评分与生存时间、生存状态及 7 个基因的 mRNA 表达水平之间的关系。如图 4C 所示，随着风险评分的升高，患者的生存时间显著缩短，而 7 个基因的 mRNA 表达量则呈现出随风险评分升高而增加的趋势。此外，基于风险评分预测患者预后的 1、3、5 年生存率的 AUC 值分别为 0.74、0.69 和 0.71（图 4D），表明该模型具有良好的预测性能。Kaplan-Meier 生存曲线分析结果显示，与高风险组患者相比，低风险组患者的总体生存率（OS）提高（log-rank χ^2 = 13.0，P < 0.001）（图 4E）。其中，95% 置信区间为 1.54 ~ 3.26，表明 HR 值在该区间内有统计学意义，且完全大于 1，说明高风险组的生存风险高于低风险组（HR = 2.24，95% CI = 1.54 ~ 3.26，P < 0.001）。此外，P 值小于 0.05，进一步支持两组生存曲线之间差异的统计学显著性。

2.4 基于 GSE68465 数据集验证预后风险模型 将在 GSE30219 数据集中构建的风险评分公式应用于 GSE68465 数据集，计算验证集中肺癌患者的风险评分，其中 GSE68465 数据集中包含 443 例肺癌患者样本和 19 例正常样本。结果显示，随着风险评分的升高，患者的生存时间显著缩短，且 7 个基因的

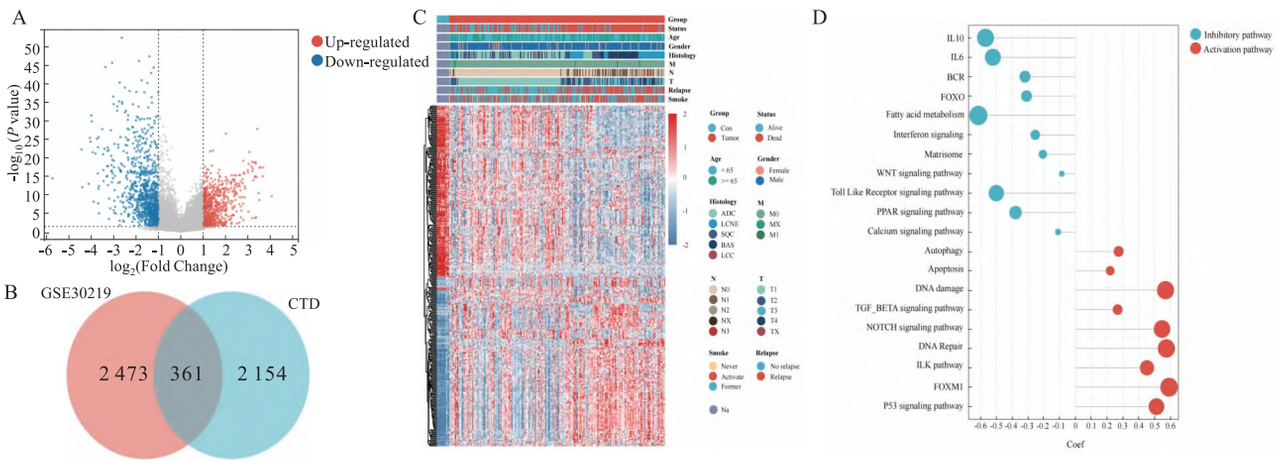

图 2　丙烯醛相关肺癌差异表达基因的筛选及 GSVA 通路富集结果

Fig. 2　Screening of differential expression genes related to acrolein in lung cancer and GSVA pathway enrichment results

A: Volcano plot depicting the differentially expressed genes between normal and lung cancer samples in the GSE30219 dataset; B: Venn diagram showing the overlap between differentially expressed genes from the GSE30219 dataset and acrolein-related genes from the CTD database; C: Heatmap representing the expression levels of the intersecting genes; D: GSVA pathway enrichment analysis of acrolein-related DEPs in lung cancer.

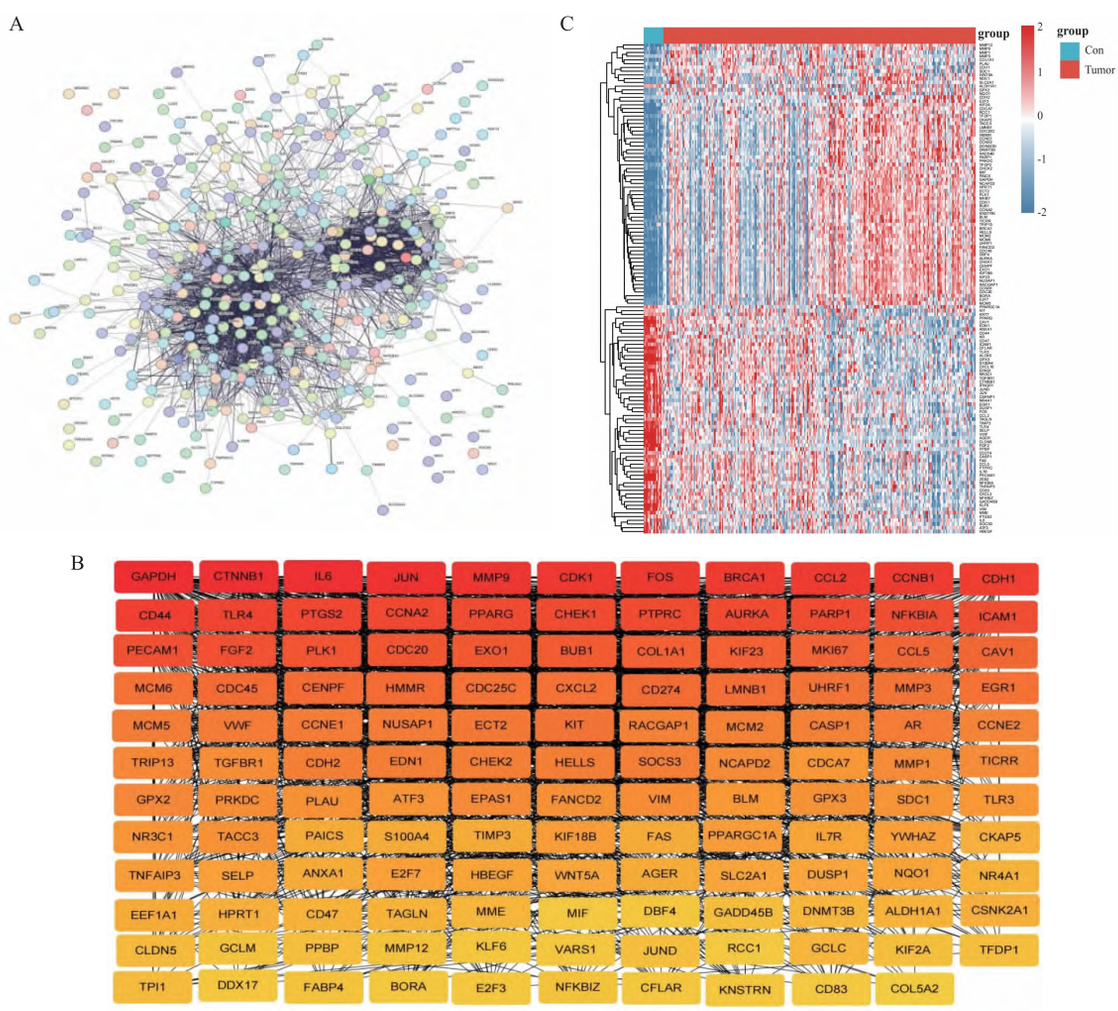

图 3　PPI 网络及枢纽基因筛选结果

Fig. 3　PPI network and hub genes selection results

A: PPI network constructed based on the intersecting genes; B: Hub genes identified using the Degree algorithm; C: Heatmap showing the expression levels of the hub genes.

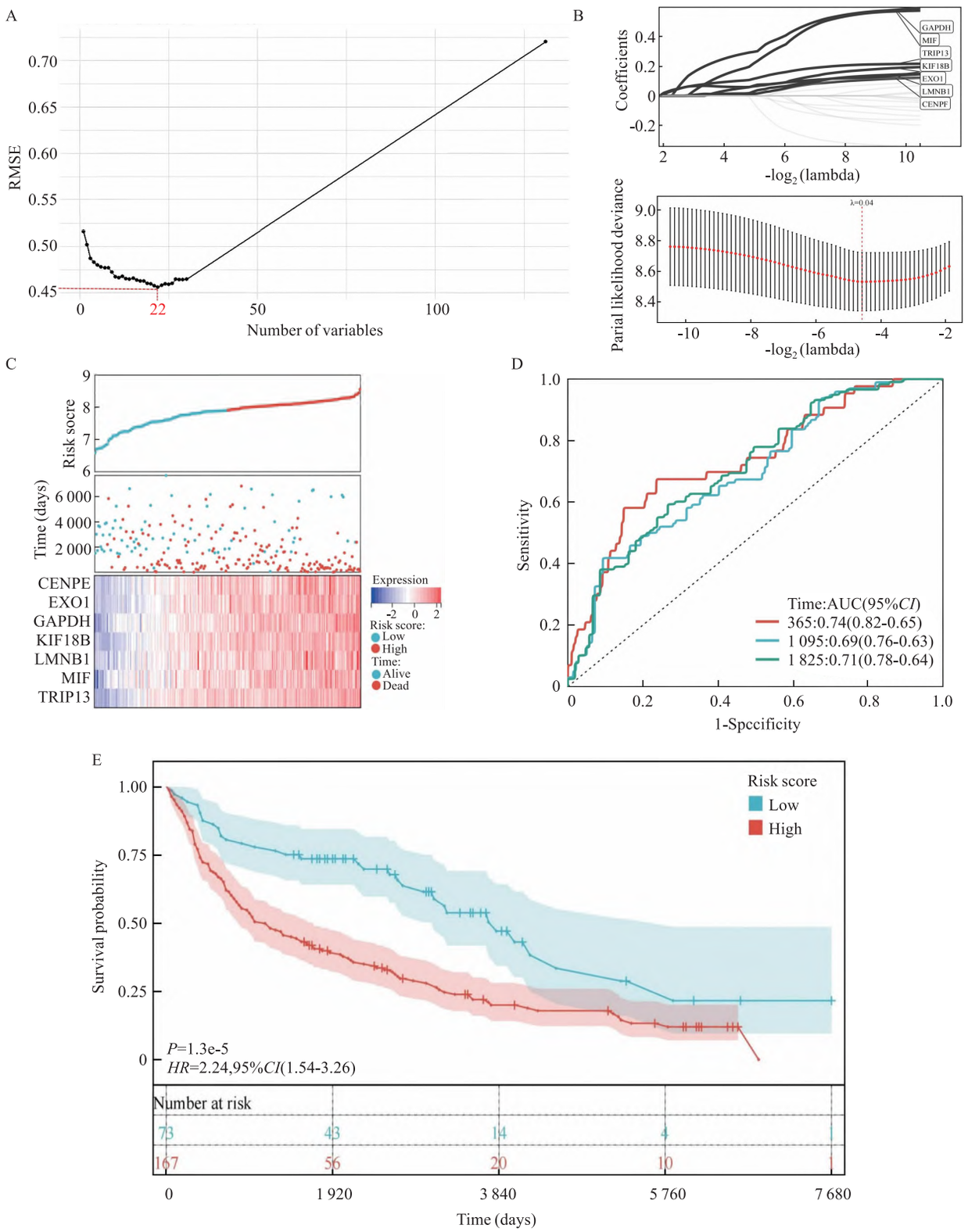

图 4　通过机器学习构建预后风险模型

Fig. 4　Construction of prognostic risk model using machine learning

A: Biomarkers selected using the SVM-RFE algorithm; B: Prognostic risk model constructed through LASSO-Cox regression analysis; C: Changes in patient survival status and the expression levels of 7 genes' mRNA with increasing risk scores in the GSE30219 dataset; D: ROC curves for 1-year, 3-year, and 5-year outcomes based on risk scores; E: Kaplan-Meier survival curves comparing high-risk and low-risk patient groups.

mRNA 表达趋势与 GSE30219 数据集中一致(图 5A)。基于风险评分预测患者 1、3、5 年生存率的 AUC 值分别为 0.65、0.62 和 0.59,Kaplan-Meier 生存曲线分析显示高、低风险组患者的生存率差异显著,且高风险组患者的总体生存期显著较短(图 5B、C)。其中,95% 置信区间为 1.35 ~ 2.21,表明 *HR* 值在该范围内具有统计学意义,且区间完全大于 1,说明高风险组患者的生存风险显著高于低风险组。此外,*P* 值小于 0.05,进一步验证了两组生存曲线之间的显著差异。结果表明,所构建的风险模型在验证数据集中有较好的预测能力和准确性。

2.5 高低风险患者的临床特征分析及 GSVA 通路富集分析 将 GSE30219 数据集中患者年龄、吸烟情况、性别、TNM 分期、组织学类型、复发情况、生存状态进行比较(图 6A)。采用卡方检验方法,对高风险与低风险患者的组织学类型进行了比较分析。结果表明,在低风险患者群体中,腺癌(ADC)所占的比例相对较高;而在高风险患者群体中,鳞状细胞癌(SQC)的占比则相对较高,高低风险患者之间差异具有统计学意义(*P* <0.001)(图 6B)。

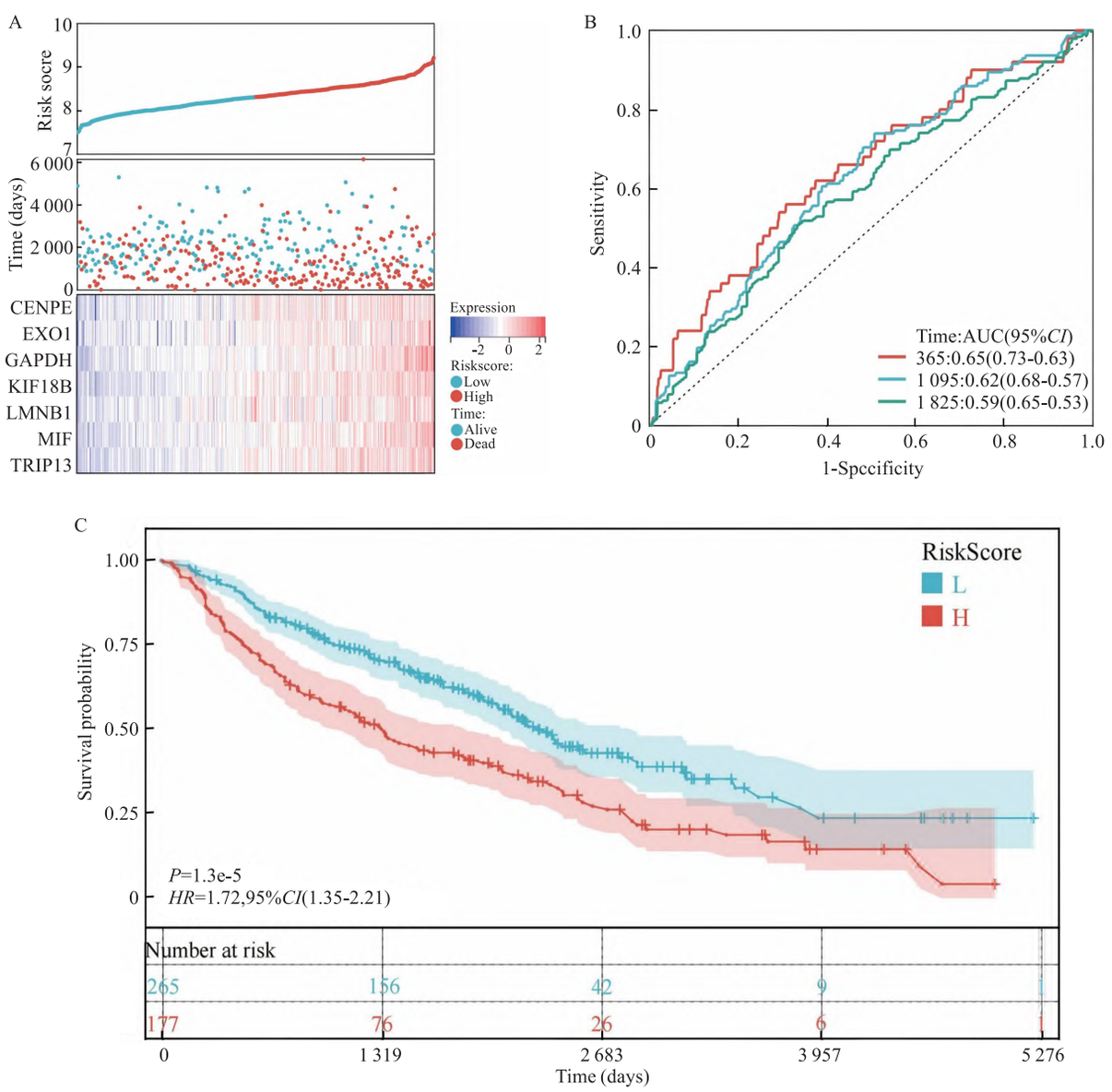

图 5 使用验证集验证预后风险模型

Fig. 5 Validation of the prognostic risk model using the validation cohort

A: Changes in patient survival status and the expression levels of 7 genes' mRNA with increasing risk scores in the GSE68465 validation cohort; B: ROC curves for 1-year, 3-year, and 5-year outcomes based on risk scores in the validation cohort; C: Kaplan-Meier survival curves comparing high-risk and low-risk patient groups in the validation cohort; H: high-risk; L: low-risk.

采用卡方检验方法对高风险与低风险患者的吸烟状况进行对比分析结果显示，在两个风险群体中，吸烟活跃人群的占比均相对较高。然而，高风险人群中吸烟活跃者的比例高于低风险人群，无吸烟史者的比例则低于低风险人群。这一发现支持吸烟与丙烯醛相关肺癌之间存在相关性的假设（图 6C）。

此外，将基于丙烯醛的肺癌相关差异蛋白按高低风险人群进行 GSVA 通路富集分析，筛选出显著的 20 条通路，结果显示，p53 信号通路、TGFβ 信号通路、DNA 修复、DNA 损伤等信号通路被激活，而 PPAR、BCR 等信号通路被抑制。结果表明，基于丙烯醛的肺癌相关 DEGs 在丙烯醛的生物学作用及肺癌的发生发展中发挥着关键作用。同时，与低风险人群相比，高风险人群中的这些通路的激活或抑制

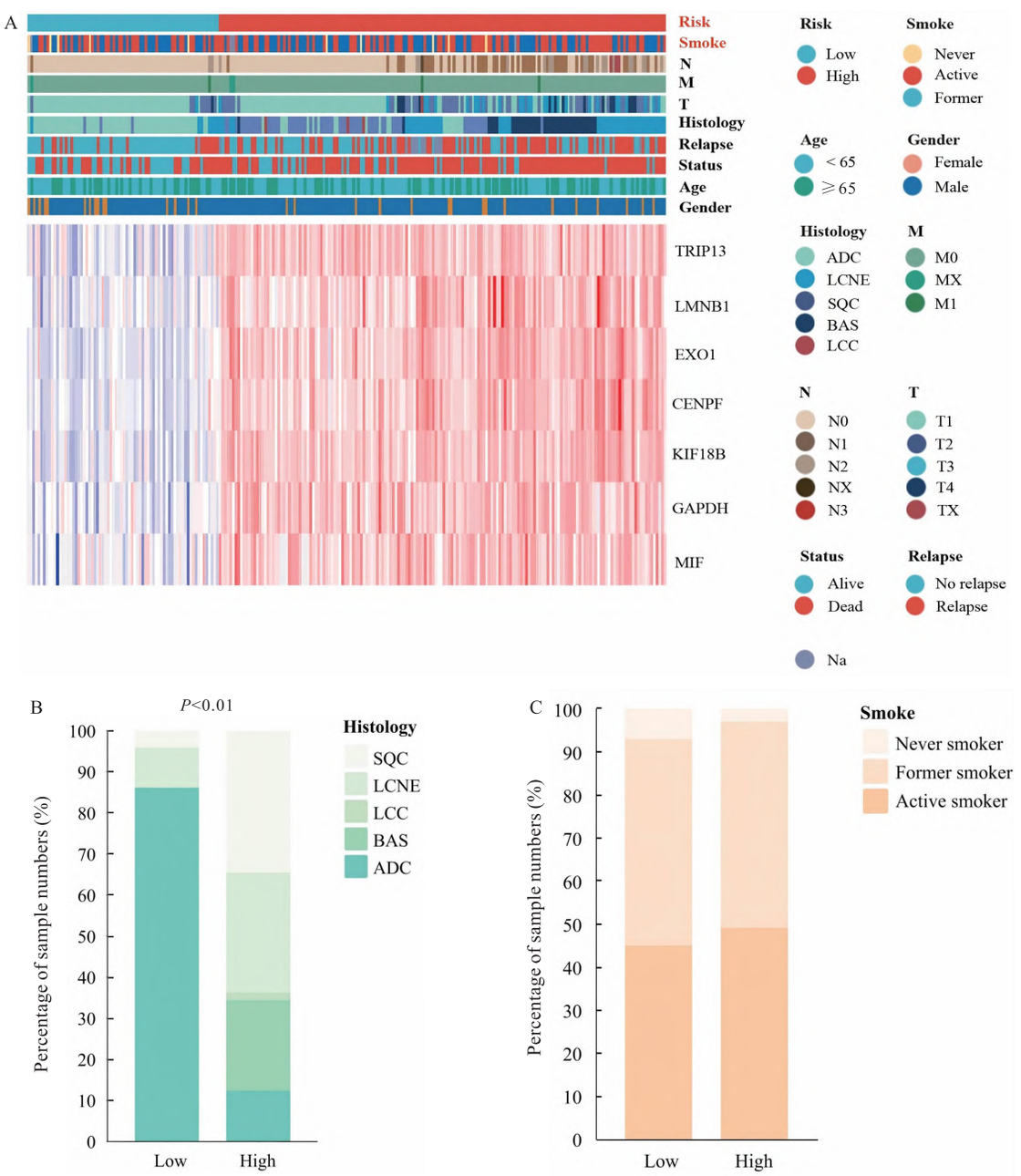

图 6　高低风险患者的临床特征与通路富集分析

Fig. 6　Clinical features and pathway enrichment analysis of high risk and low risk patients

A: Heatmap showing the expression levels of 7 genes in relation to clinical features in the GSE30219 dataset; B: Bar plot depicting the results of chi-square tests for different histological types; C: Bar plot illustrating the results of chi-square tests for smoking status.

程度更为显著,表明该预后风险模型具有良好的预后预测能力(图 7)。

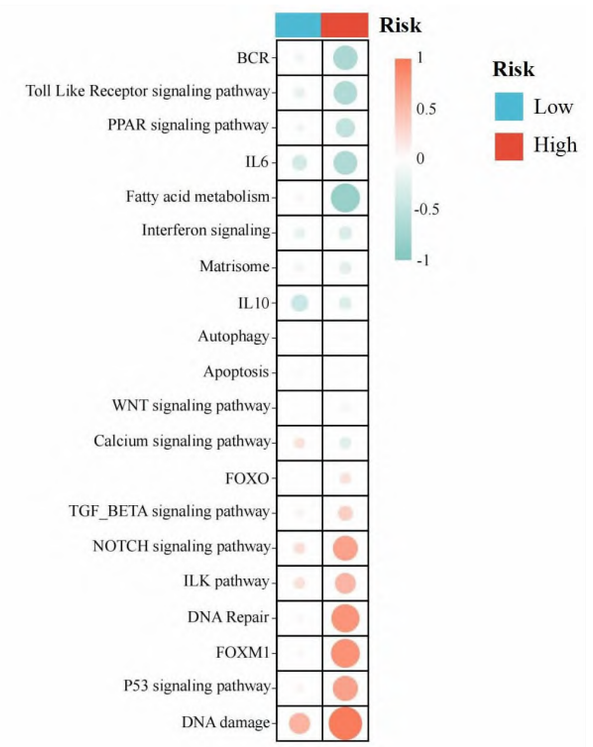

图 7 高低风险患者的 GSVA 通路富集分析结果

Fig. 7 GSVA pathway enrichment analysis results of high risk and low risk patients

2.6 高低风险患者的免疫浸润分析及相关性分析

使用 CIBERSORT 算法对 GSE30219 数据集中高、低风险患者样本进行免疫浸润分析(图 8A)。22 个免疫细胞相关性热图结果显示,嗜酸性粒细胞与调节性 T 细胞、休眠自然杀伤细胞、M0 型巨噬细胞、M2 型巨噬细胞的相关系数分别为 1.0、0.8、0.71、1.0,呈正相关;分别与 $CD4^+$ 初始 T 细胞、γδT 细胞的相关系数都为 -1.0,呈负相关(图 8B)。对高风险与低风险肺癌患者中的免疫细胞变化进行对比分析,结果显示,在高风险患者中,初始 B 细胞、$CD4^+$ 记忆 T 细胞激活状态、休眠自然杀伤细胞、M1 型巨噬细胞、静息髓样树突状细胞、静息肥大细胞的表达呈上调趋势,其中休眠自然杀伤细胞、静息髓样树突状细胞和静息肥大细胞的上调最为显著(P < 0.000 1);在低风险患者中,记忆 B 细胞、T 细胞滤泡辅助性 T 细胞、γδT 细胞、激活 NK 细胞、M2 型巨噬细胞、激活髓样树突状细胞、激活肥大细胞的表达呈下降趋势,其中记忆 B 细胞和激活肥大细胞的下调最为显著(P <0.000 1)(图 8C)。

3 讨论

丙烯醛是一种具有致突变性和遗传毒性的环境污染物[5]。丙烯醛与多种疾病密切相关,包括呼吸系统疾病(如哮喘、慢性阻塞性肺病、肺癌等)以及氧化应激相关疾病(如阿尔茨海默病、帕金森病和心血管疾病等)[6]。对于普通人群而言,烟草烟雾是丙烯醛最主要的暴露来源,其暴露剂量远高于其他环境来源[7]。烟草烟雾中的丙烯醛、尼古丁和乙醛是关键成分[8]。已有体内研究显示,尼古丁与乙醛主要与成瘾性相关,而丙烯醛则是引起烟草烟雾暴露所致呼吸反应及细胞毒性的主要因素[9]。丙烯醛的高接触人群包括吸烟者、二手烟接触者、工业工人及居住在汽车交通密集和高污染地区的居民,消防员也是丙烯醛的高危职业暴露人群[10]。因此,丙烯醛在日常生活中广泛存在。

研究[11]表明,丙烯醛可能在肺癌的发生发展中起到重要作用,丙烯醛与肺癌之间的潜在关联已引起广泛关注。本研究通过综合生物信息学分析,构建了基于丙烯醛相关基因的肺癌预后风险模型,并筛选出 7 个关键预后基因:*CENPF*、*EXO1*、*GAPDH*、*KIF18B*、*LMNB1*、*MIF* 和 *TRIP13*。这些基因与丙烯醛密切相关,并可能通过多种机制促进肺癌及其他肺部疾病的发生与发展[12]。

此外,免疫浸润分析结果显示,风险评分与免疫浸润评分显著相关。CIBERSORT 算法评估免疫细胞亚型比例结果显示,高低风险组在免疫细胞成分上存在显著差异,免疫细胞成分比例的失衡可能与癌症患者的不良预后和较低生存率相关,揭示了丙烯醛相关基因对肺癌患者免疫微环境的潜在影响。

本研究模型与既往基于全基因组筛选或传统临床病理特征(如 TNM 分期)构建的肺癌预后预测模型相比,首次以丙烯醛相关基因为切入点,构建了具有明确致癌机制背景的预后模型,且选取的核心基因多已被证实在肺癌发生发展或氧化应激通路中发挥关键作用。研究[13]表明,丙烯醛刺激可激活 p53 信号通路,转录因子 p53 作为一种肿瘤抑制因子,在低剂量丙烯醛刺激下,通过诱导细胞周期停滞、促进 DNA 修复与细胞存活发挥作用,在高剂量或长时间刺激的条件下,p53 则通过诱导细胞凋亡来触发细胞死亡。丙烯醛能够激活细胞凋亡及 DNA 修复等相关信号通路[14]。NOTCH信号通路的激活可能

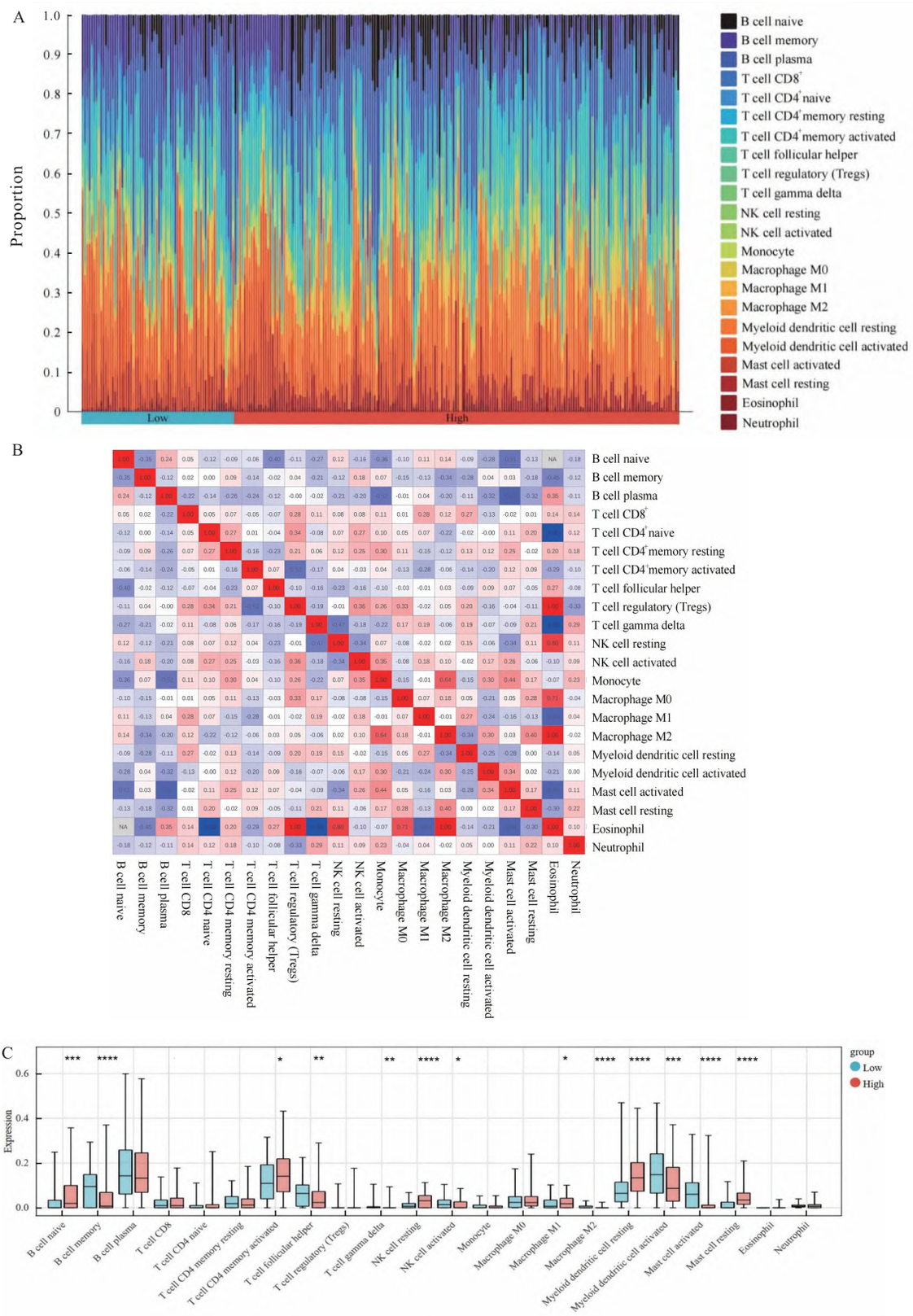

图 8　高低风险患者的免疫浸润分析及相关性分析

Fig. 8　Immune infiltration and correlation analysis of high risk and low risk patients

A: Immune infiltration analysis results of high risk and low risk patient groups; B: Correlation heatmap of 22 immune cell types; C: Differential expression of immune cells between high risk and low risk patients; $^{*}P<0.05$, $^{**}P<0.01$, $^{***}P<0.001$, $^{****}P<0.000\ 1$ *vs* Low-risk group.

促进肺癌细胞的增殖、存活及转移,并且 NOTCH 通路的异常激活被认为是肿瘤发生的一个关键因素[15]。因此,丙烯醛相关的 DEGs 在肺癌的发生和发展过程中起着至关重要的作用,为肺癌的干预提供了新的理论依据和潜在的靶向治疗靶点。

由于丙烯醛已被确认为是烟草烟雾中引发肺癌等呼吸系统疾病的首要因素[9],本模型将吸烟暴露相关因素纳入考量,可更精准预测吸烟导致的个体化风险增量,并将患者吸烟史纳入预后评估体系。与其他肺癌风险模型相比,本模型弥补了未充分考虑环境暴露因素的局限性,提升了预测准确性和特定人群的适用性。因此,本研究不仅建立了具有良好预测效能的模型,也从环境致癌物所介导的特异性损伤角度构建模型,为肺癌高危人群的精准评估和个体化干预策略的制定提供了新的思路。

本研究仍存在一定的局限性。首先,未构建丙烯醛与肺癌发生发展的直接关联模型,限制了对其作用机制的深入探讨。其次,训练集中正常样本数量较少,相较于癌症样本存在不均衡,可能会影响统计分析结果的稳健性和可靠性。此外,未来研究可进一步细化丙烯醛与肺癌之间的关系,重点分析吸烟作为丙烯醛暴露的主要来源之一在肺癌发生中的作用,明确吸烟作为危险因素的影响,从而为临床风险评估与个体化防控策略提供更为精准的依据。

综上所述,本研究通过筛选丙烯醛相关基因与肺癌 DEGs,成功构建了丙烯醛相关基因的肺癌预后预测模型。该模型在肺癌患者的预后评估中表现出良好的临床应用潜力,并为肿瘤微环境的深入评估提供了依据。结果进一步揭示了丙烯醛作为有毒环境污染物在肺癌发生与发展中的潜在影响,凸显其在促进肺癌进展中的毒性作用,强调减少丙烯醛暴露对于肺癌防治和公共健康保护的重要性。同时,提示要进一步关注环境暴露因素与疾病发生和进展之间的关系,研究为公共卫生干预和疾病防控策略提供科学依据。

参考文献

[1] Ou J, Zheng J, Huang J, et al. Interaction of acrylamide, acrolein, and 5-hydroxymethylfurfural with amino acids and DNA[J]. J Agric Food Chem, 2020, 68(18): 5039 - 48. doi:10.1021/acs.jafc.0c01345.

[2] Averill-Bates D A, Tanel A. Activation of cellular signalling pathways and apoptosis by the aldehyde acrolein - a major environmental hazard[J]. Redox Biochem Chem, 2024, 7:100019. doi:10.1016/j.rbc.2023.100019.

[3] Lee H W, Wang H T, Weng M W, et al. Cigarette side-stream smoke lung and bladder carcinogenesis: inducing mutagenic acrolein-DNA adducts, inhibiting DNA repair and enhancing anchorage-independent-growth cell transformation [J]. Oncotarget, 2015, 6(32): 33226 - 36. doi:10.18632/oncotarget.5429.

[4] Liu Q, Lou H, Zhang X, et al. Association between acrolein exposure and respiratory hazards: a systematic review and meta-analysis[J]. Atmos Pollut Res, 2023, 14(1): 101633. doi:10.1016/j.apr.2022.101633.

[5] Schieweck A, Uhde E, Salthammer T. Determination of acrolein in ambient air and in the atmosphere of environmental test chambers[J]. Environ Sci Process Impacts, 2021, 23(11): 1729 - 46. doi:10.1039/d1em00221j.

[6] Muguruma K, Pradipta A R, Ode Y, et al. Disease-associated acrolein: a possible diagnostic and therapeutic substrate for *in vivo* synthetic chemistry[J]. Bioorg Med Chem, 2020, 28(24): 115831. doi:10.1016/j.bmc.2020.115831.

[7] Hikisz P, Jacenik D. The tobacco smoke component, acrolein, as a major culprit in lung diseases and respiratory cancers: molecular mechanisms of acrolein cytotoxic activity[J]. Cells, 2023, 12(6): 879. doi:10.3390/cells12060879.

[8] Soleimani F, Dobaradaran S, De-la-Torre G E, et al. Content of toxic components of cigarette, cigarette smoke *vs* cigarette butts: a comprehensive systematic review[J]. Sci TotalEnviron, 2022, 813: 152667. doi:10.1016/j.scitotenv.2021.152667.

[9] Chen H C, Cheng S W, Chen N Y, et al. Characterization and quantification of acrolein-induced modifications in hemoglobin by mass spectrometry-effect of cigarette smoking[J]. Chem Res Toxicol, 2022, 35(12): 2260 - 70. doi:10.1021/acs.chemrestox.2c00262.

[10] Peterson L A, Seabloom D, Smith W E, et al. Acrolein increases the pulmonary tumorigenic activity of the tobacco-specific nitrosamine 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK)[J]. Chem Res Toxicol, 2022, 35(10): 1831 - 9. doi:10.1021/acs.chemrestox.2c00135.

[11] Park S L, Le Marchand L, Cheng G, et al. Quantitation of DNA adducts resulting from acrolein exposure and lipid peroxidation in oral cells of cigarette smokers from three racial/ethnic groups with differing risks for lung cancer[J]. Chem Res Toxicol, 2022, 35(10):1914 - 22. doi:10.1021/acs.chemrestox.2c00171.

[12] Gonzalez-Rivera J C, Baldridge K C, Wang D S, et al. Post-transcriptional air pollution oxidation to the cholesterol biosynthesis pathway promotes pulmonary stress phenotypes[J]. Commun Biol, 2020, 3(1): 392. doi:10.1038/s42003 - 020 - 01118 - 6.

[13] 郭 欣, 胡代菊, 梅晓冬. 烟雾暴露小鼠肺部氧化应激与炎症的变化及戒烟的影响[J]. 安徽医科大学学报, 2015, 50(6): 757 - 60. doi:10.19405/j.cnki.issn1000 - 1492.2015.06.009.

[13] Guo X, Hu D J, Mei X D. Oxidative stress and inflammatory changes in the lung caused by cigarette smoking exposure in mice and the effect of smoking cessation[J]. Acta Univ Med Anhui, 2015, 50(6): 757 - 60. doi:10.19405/j.cnki.issn1000 - 1492.2015.06.009.

[14] Liu D, Cheng Y, Tang Z, et al. Toxicity mechanism of acrolein on DNA damage andapoptosis in BEAS-2B cells: insights from cell biology and molecular docking analyses[J]. Toxicology, 2022, 466: 153083. doi:10.1016/j.tox.2021.153083.

[15] Owen D H, Giffin M J, Bailis J M, et al. DLL3: an emerging target in small cell lung cancer[J]. J Hematol Oncol, 2019, 12: 61. doi:10.1186/s13045-019-0745-2.

Construction of a prognostic model for lung cancer based on acrolein-related genes

Feng Yiting[1,2], Ren Liangliang[2], Lou Lijuan[2], Shen Yuxian[1], Jiang Ying[1,2]

[[1]*Dept of Biochemistry and Molecular Biology, School of Basic Medical Sciences, Anhui Medical University, Hefei* 230032; [2]*State Key Laboratory of Medical Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing* 102206]

Abstract ***Objective*** To construct and validate a prognostic model for lung cancer based on acrolein-related genes using bioinformatics methods. ***Methods*** Lung cancer datasets GSE30219 and GSE68465 were obtained from the GEO database, and acrolein-related gene sets were retrieved from the CTD database. Differentially expressed genes (DEGs) between cancer and adjacent tissues were identified in the GSE30219 dataset. The intersection of these DEGs and acrolein-related genes was then used to identify candidate genes. Gene set variation analysis (GSVA) was performed to assess functional alterations based on the intersection genes. A protein-protein interaction (PPI) network was constructed based on the STRING database to identify core hub genes. Subsequently, support vector machine recursive feature elimination (SVM-RFE) and LASSO-Cox regression analyses were employed to develop a prognostic model based on acrolein-related genes, which was independently validated using the GSE68465 dataset. The CIBERSORT algorithm was applied to evaluate the immune cell infiltration characteristics between high- and low-risk groups, and functional enrichment analysis of DEGs between the two groups was conducted to further explore the potential molecular mechanisms underlying the prognostic model. ***Results*** A total of 361 acrolein-related DEGs were identified in lung cancer, and 7 key genes were selected for model construction. Kaplan-Meier survival analysis revealed that patients in the high-risk group had significantly lower survival rates compared to those in the low-risk group ($P<0.000\ 1$). Receiver operating characteristic (ROC) curve analysis demonstrated that the model possessed good predictive performance. Moreover, immune infiltration analysis indicated that the risk score was closely associated with multiple immune cell subsets, suggesting a potential role of acrolein-related genes in modulating the lung cancer immune microenvironment. ***Conclusion*** The prognostic model for lung cancer based on acrolein-related genes demonstrates significant application value in predicting the prognosis of lung cancer, providing new insights into the potential mechanisms of acrolein in the onset and progression of lung cancer.

Key words acrolein; lung cancer; environmental pollutants; bioinformatics; machine learning; prognostic model

Fund program National Key Research and Development Program of China (No. 2020YFE0202200)

Corresponding authors Jiang Ying, E-mail: jiangying@ncpsb.org.cn; Shen Yuxian, E-mail: shenyx@ahmu.edu.cn