

基于生物信息学方法的肺腺癌分子机制探索及预后风险模型构建

项黎¹ 胡果荣²

摘要 目的 通过数据挖掘和生物信息学分析探索影响肺腺癌(lung adenocarcinoma, LUAD)相关基因的表达谱,识别关键基因及其互作网络,探索其在LUAD发生发展中的潜在作用机制,并构建基于关键基因的LUAD预测模型。方法 通过基因表达综合数据库(Gene Expression Omnibus, GEO)获取LUAD相关基因数据集GSE227427,使用R软件“impute”包筛选差异表达的基因,导入STRING数据库构建蛋白质-蛋白质相互作用(protein-protein interaction, PPI)网络,在Cytoscape软件中进一步分析识别筛选出关键靶点。通过DAVID数据库进行基因本体(Gene Ontology, GO)、京都基因与基因组百科全书(Kyoto Encyclopedia of Genes and Genoms, KEGG)通路富集分析,使用R软件可视化富集结果,基于关键基因通过支持向量机(support vector machine, SVM)构建LUAD预测模型,采用五折交叉验证评估模型性能。结果 共筛选出784个LUAD相关基因,构建包含759个节点和31 955条边的PPI网络,进一步筛选出179个关键靶点,主要涉及细胞免疫、炎症反应及肿瘤坏死因子调控等生物过程。构建基于SVM的LUAD预测模型,高斯核函数总体准确率(accuracy, ACC)为96.98%,特异性(specificity, SP)为99.02%,线性核函数、多项式核函数以及sigmoid核函数对该模型的评价均较高,提示该模型具有很高的临床应用潜力和价值。结论 通过系统的生物信息学分析,成功识别了LUAD中的关键基因及其互作网络,揭示了这些基因在LUAD发生发展中的潜在作用机制。基于这些关键基因构建的LUAD预测模型表现出较高的分类性能,为LUAD的诊断和治疗提供了新的思路和潜在的靶点。

关键词 肺腺癌;预测模型;生物信息;分子机制

DOI: 10.3969/j.issn.1002-3208.2026.03.003.

中图分类号 R318.04 **文献标志码** A **文章编号** 1002-3208(2026)03-0239-08

本文著录格式 项黎,胡果荣. 基于生物信息学方法的肺腺癌分子机制探索及预后风险模型构建[J]. 北京生物医学工程, 2026, 45(3): 239-246. XIANG Li, HU Guorong. Exploring the molecular mechanisms of lung adenocarcinoma and construction of a prognostic risk model based on bioinformatics methods[J]. Beijing Biomedical Engineering, 2026, 45(3): 239-246.

Exploring the molecular mechanisms of lung adenocarcinoma and construction of a prognostic risk model based on bioinformatics methods

XIANG Li¹, HU Guorong²

1 Department of Medical Informatics, Wuhan Central Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430014;

2 Department of Respiratory and Critical Care Medicine, Wuhan Central Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430014

Corresponding author: XIANG Li (E-mail: 6524338@qq.com)

【Abstract】 Objective Through data mining and bioinformatics analysis, to explore the expression profiles of genes related to lung adenocarcinoma (LUAD), identify key genes and their interaction networks, investigate their potential mechanisms in the development and progression of LUAD, and construct a LUAD prediction model based on these key genes. **Methods** Obtained the LUAD-related gene dataset GSE227427 through the Gene Expression Omnibus (GEO) database. Using the "impute" package in R software, differentially expressed genes were screened, and then imported into the STRING database to construct a protein-protein interaction (PPI) network, which was further analyzed and refined in Cytoscape to identify key targets. Enrichment analyses of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genoms (KEGG) pathways were conducted through the DAVID database, and the enrichment results were visualized using R software. A LUAD prediction model was constructed based on the key genes using support vector machine (SVM), and the performance of the model was evaluated through five-fold cross-validation. **Results** A total of 784 LUAD-related genes were screened, and a PPI network with 759 nodes and 31 955 edges was constructed. Further screening identified 179 key targets, which were mainly involved in biological processes such as cellular immunity, inflammatory response, and tumor necrosis factor regulation. A LUAD prediction model based on SVM was constructed, with the Gaussian kernel function achieving an accuracy (ACC) of 96.98% and a specificity (SP) of 99.02%. The linear kernel, polynomial kernel, and sigmoid kernel all provided high evaluations for this model, suggesting that it possesses great potential and value for clinical application. **Conclusions** The present study successfully identified key genes and their interaction networks in LUAD through systematic bioinformatics analysis, revealing the potential mechanisms of these genes in the development and progression of LUAD. The LUAD prediction model constructed based on these key genes demonstrated high classification performance, offering new avenues and potential targets for the diagnosis and treatment of LUAD.

【Keywords】 lung adenocarcinoma; prediction model; bioinformatics; molecular mechanism

0 引言

肺癌作为当前全球范围内主要的致死性癌症之一,其发病率呈现出不断攀升的态势,且年轻化趋势愈发显著^[1-2]。肺腺癌(lung adenocarcinoma, LUAD)是肺癌的主要组织学类型之一,具有高发病率和晚期高转移复发率,严重威胁着患者的生命健康^[3-4]。因此,探索LUAD发生发展的分子机制,特别是关键基因的筛选与鉴定,对实现LUAD的早期诊断与精准治疗、提高患者预后、降低疾病负担具有不可估量的意义。近年来,生物信息学、基因表达谱以及微阵列技术的飞速发展,为癌症的深入研

究提供了强有力的技术支撑^[5-6]。作为高通量基因表达数据及其他形式功能基因组数据的公共存储平台,基因表达综合数据库(Gene Expression Omnibus, GEO, <http://www.ncbi.nlm.gov/geo/>)涵盖了多种物种的转录组学、蛋白组学以及表观遗传组学数据,在深入理解生命现象、探索疾病成因以及发现潜在治疗策略等方面具有不可替代的作用^[7-8]。基于上述背景,本研究聚焦于LUAD的发病机制,通过深度挖掘GEO数据库中LUAD相关基因的表达数据进行蛋白质-蛋白质相互作用(protein-protein interaction, PPI)关系分析以及基因本体/京都基因与基因组百科全书(Gene Ontology/ Kyoto Encyclopedia of Genes and Genoms, GO/KEGG)通路富集分析,旨在揭示影响LUAD发生发展的关键基因及其作用机制,增进对LUAD发病机制的理解,为LUAD的早期筛查、精确诊断以及个性化治疗提供新的靶点和策略,以期改善LUAD患者的治疗效果和生存质量,推动肿瘤学研究的进一步发展。

基金项目:武汉市卫健委医学科研项目(WX21B30)资助

作者单位:1 华中科技大学同济医学院附属武汉中心医院医学资讯科(武汉 430014)

2 华中科技大学同济医学院附属武汉中心医院呼吸与危重症医学科(武汉 430014)

通信作者:项黎。E-mail:6524338@qq.com

1 方法

1.1 数据来源

通过 GEO 数据库获得 LUAD 相关基因数据集 GSE227427,该数据集包含了 226 例 LUAD 样本以及 20 例对照样本。根据平台注释文件对 GEO 数据集进行注释,并将探针 ID 转换为基因符号,移除没有匹配基因符号的探针。使用 R 软件“impute”包估算基因表达矩阵中的缺失值, k 值采用默认值 10。使用“limma”包完成数据归一化处理与差异表达分析,筛选条件: $|\log_2FC| \geq 1.5$ 、 $P < 0.05$,共筛选出 784 个差异表达基因。

1.2 PPI 网络分析

将在 GEO 得到的 LUAD 相关基因处理后,导入 STRING 数据库(<https://string-db.org/>),并设定特定筛选参数:选择“人类(Homo sapiens)”作为物种范围,并将置信度阈值设定为 0.400 及以上。在 STRING 数据库生成 PPI 网络图后,隐藏在网络中没有连接的节点,并导出数据为“tsv”格式的文件以及“png”格式的图像。随后,将“tsv”格式的文件导入 Cytoscape 3.7.2 软件中,使用“分析网络”功能进行数据的深入分析,并运用“CytoNCA”插件中的“网络分析器”工具来计算网络中各个节点的度值,将度值超过中位数度值两倍的节点选为关键靶点,筛选后,再次利用 STRING 数据库对其中心靶点进行分析,并利用 Cytoscape 3.7.2 软件进行关键靶点的可视化处理,并进行美化。

1.3 LUAD 的 GO 功能富集分析

使用 DAVID 6.8 数据库(<https://david.ncifcrf.gov/>)进行 GO 分类富集分析,对富集结果进行分析整理,重要程度以 P 值为指标进行排序,选取前 20 位为重要通路,导入 R 软件(4.3.3)绘制气泡图,将结果进行可视化处理。

1.4 LUAD 的 KEGG 通路富集分析

使用 DAVID 6.8 数据库进行 KEGG 分类富集分析,深入分析富集结果,重要程度以 P 值为指标进行排序。选取前 20 位为重要通路并将这些数据导入 R 软件,创建气泡图,将这些通路的分析结果以可视化形式展现出来。

1.5 预测模型的建立及验证

完成关键基因的筛选后,将这些基因作为预测

基因,利用支持向量机(support vector machine, SVM)构建 LUAD 预测模型。为了保证模型预测的精确性和可靠性,采用五折交叉验证技术来确定最佳的核函数和参数设置。该技术涉及将整个数据集均分为 5 部分,每一部分轮流作为验证集,而其余 4 部分用作训练集,重复这一过程 5 次。最终,根据这 5 次的验证结果计算平均值,以此作为评估模型性能的标准。

模型训练完成后,采用生物信息学中常用的评价标准来全面衡量模型的分类性能。评价标准包括总体准确率(accuracy, ACC)、敏感性(sensitivity, SE)、特异性(specificity, SP)、马修斯相关系数(Mathew's correlation coefficient, MCC)。模型的整体准确率展示了其在区分样本正确性方面的表现;敏感性则衡量了模型识别实际为正类(LUAD)样本的效率;特异性衡量了模型正确识别负类(非 LUAD)样本的能力;而马修斯相关系数则是一个综合指标,它考虑了真阳性、真阴性、假阳性和假阴性的数量,为评估模型的整体效能提供了更全面的视角。

其中,ACC、SE、SP、MCC 的计算公式如下^[9]:

$$\begin{aligned} \text{ACC} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \\ \text{SE} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{SP} &= \frac{\text{TN}}{\text{TN} + \text{FP}} \\ \text{MCC} &= \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \end{aligned}$$

式中:TP 为正确预测的正例个数(true positives);TN 为正确预测的反例个数(true negatives);FP 为反例预测为正例的个数(false positives);FN 为正例预测为反例的个数(false negatives)

2 结果

2.1 PPI 网络及拓扑分析

PPI 网络包含 759 个节点和 31 955 条边,节点度值的中位数为 62,阈值即为 124,根据每个节点的度值进行排序,以颜色深浅表现靶点的重要程度,最终筛选出 179 个关键节点作为潜在的重要靶点,见图 1,关键靶点网络见图 2。

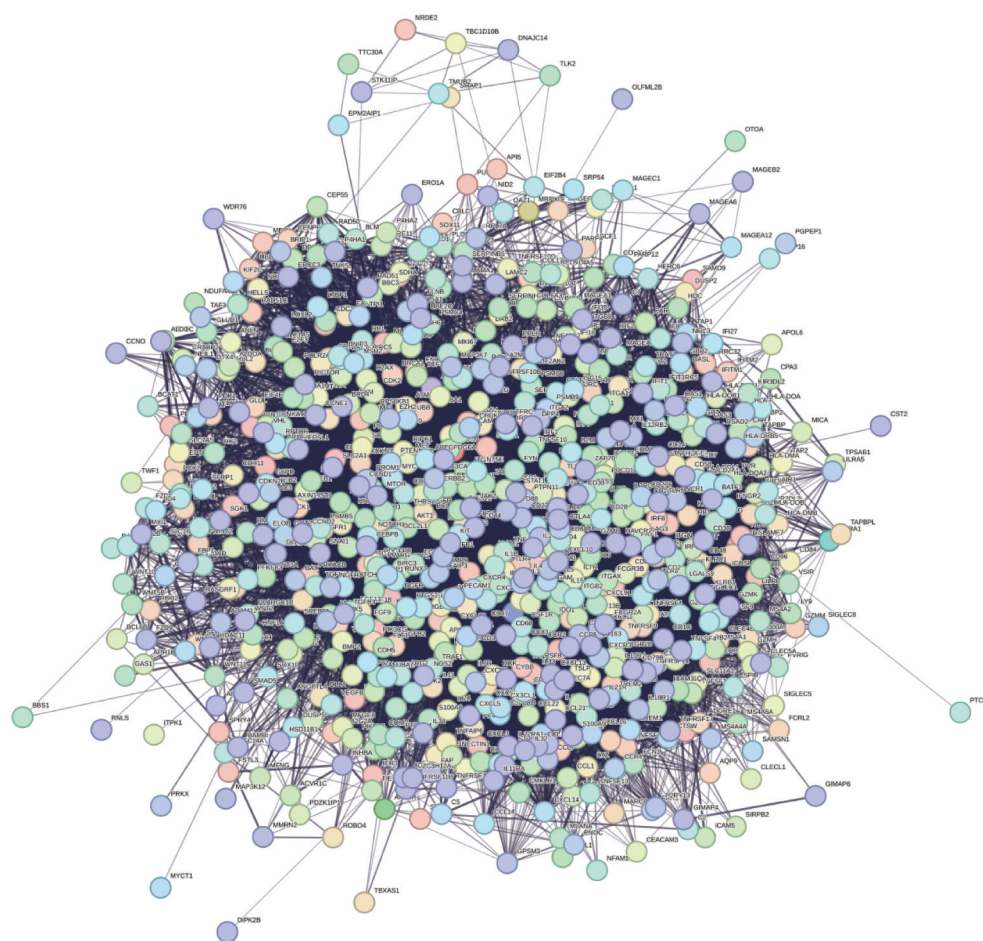


图 1 蛋白质-蛋白质相互作用网络

Figure 1 PPI network

2.2 相关靶点的GO通路富集分析

通过DAVID数据库对筛选出的179个关键基因进行GO富集分析,最终筛选出820个具有显著性的通路。这些通路按照生物学过程(biological process, BP)650个、细胞组分(cellular component, CC)85个、分子功能(molecular function, MF)129个进行分类。对每个类别的前20个通路进行排序,关键通路主要与细胞免疫及炎症反应、肿瘤坏死因子的正调控相关,与肿瘤期间细胞反应有密切关系,见图3。

2.3 相关靶点的KEGG通路富集分析

利用DAVID数据库对179个关键靶点进行KEGG富集分析,共筛选出139条通路,与LUAD的驱动、免疫微环境调节及靶向治疗等方面密切相关,前20条通路见图4。

2.4 预测模型性能评价

利用SVM构建LUAD预测模型,对数据集GSE227427进行测试,分别统计ACC、SE、SP、MCC值,高斯核函数在分类肺癌样本时表现最优,其特异性高达99.02%,准确率为96.98%,线性核函数和sigmoid核函数虽然在识别癌症样本上表现良好,但整体性能略逊于高斯核函数,见表1。

3 讨论

LUAD的发生发展是一个高度复杂的生物学过程,涉及多种基因与蛋白质表达模式的异常变化^[10]。深入分析与LUAD紧密相关的基因不仅能够为早期诊断和预后治疗评估提供更为丰富的理论依据,还有助于延长癌症患者的生存时间,提高生存质量^[11]。

本研究通过对GEO数据库中LUAD相关基因

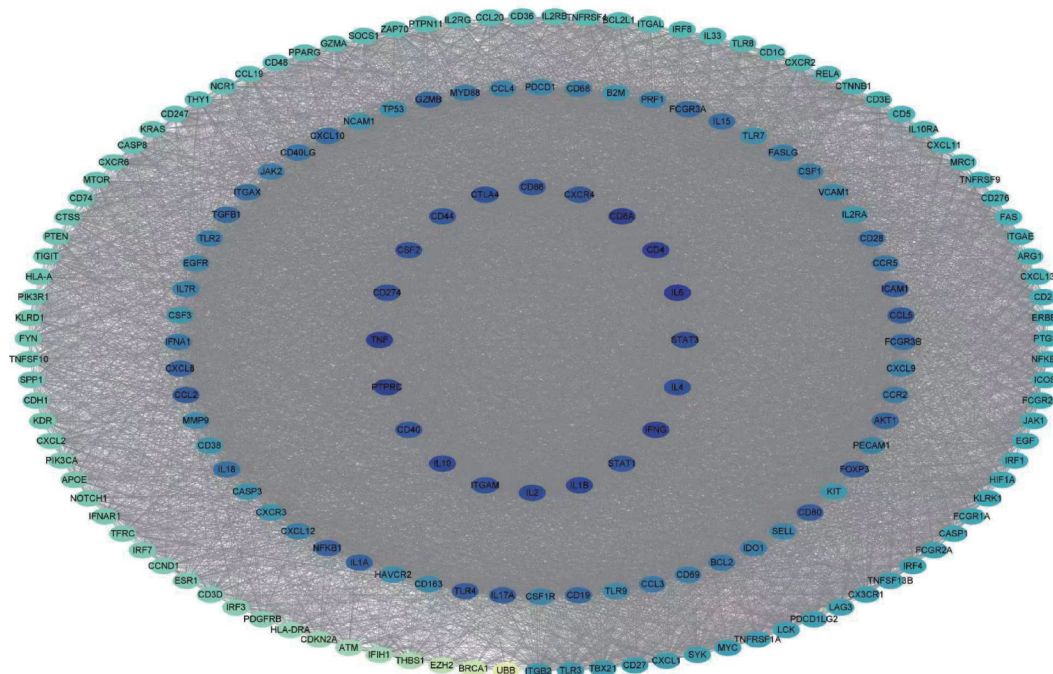


图2 关键靶点网络

Figure 2 Key target network

数据集 GSE31210 的深入分析,获得了一些新发现。该数据集涵盖了 784 个与 LUAD 相关的基因,通过 STRING 数据库与 Cytoscape 软件构建 PPI,最终筛选出 179 个关键靶点。经 GO 和 KEGG 富集分析发现这些关键靶点例如肿瘤坏死因子(tumor necrosis factor, TNF)、白细胞介素-6(interleukin-6, IL-6)、干扰素- γ 等,主要涉及细胞免疫及炎症反应、肿瘤坏死因子的正调控等关键生物过程^[12-13]。Zhang 等^[14]收集了来自 7 个不同队列的 1 300 例 LUAD 患者的临床信息,通过基因表达谱分析构建基于 TNF 家族的预后模型,发现该模型与不同肿瘤浸润免疫细胞类型以及炎症活动水平相关。IL-6 是一种功能广泛的多效性细胞炎症因子,在乳腺癌、肺癌等癌症中发挥促癌作用^[15-16],Islas-Vazquez 等^[17]发现 LUAD 患者 IL-6 水平和中性粒细胞与淋巴细胞计数比值系统性免疫炎症指数均为正相关,可作为监测化疗以及临床反应的标志物。在 KEGG 通路中,细胞因子-细胞因子受体相互作用以最高的基因计数和极低的 *P* 值位居榜首,细胞因子网络复杂,涉及免疫调节、炎症和肿瘤进展等多个方面,在 LUAD 的发病和治疗过程中起着关键的作用^[18-19]。癌症中的 PD-L1 表达与 PD-1 检查点通路和癌症中的信号通路显著

富集强调了免疫检查点抑制在 LUAD 治疗中的作用,对这些通路的探索有助于理解肿瘤逃避免疫监视的机制,可为免疫疗法的开发提供新思路^[20]。

尽管本研究分析结果显示这些关键基因在 LUAD 发生发展过程中具有潜在作用,但仍需通过进一步的实验验证来确认其准确性和可靠性。未来研究应聚焦于这些关键基因的分子功能验证,包括基因敲除、过表达实验以及临床样本的验证等,以全面评估这些基因在 LUAD 中的实际作用。

本研究采用SVM方法,通过高斯核函数构建LUAD预测模型,经五折交叉验证,结果显示,该模型SP高达99.02%,ACC为96.98%,MCC达到了96.28%,提示该模型在特异性、准确率、稳定性、可靠性和非线性处理能力等方面都展现出了显著的优势,具有很高的临床应用潜力和价值。线性核函数、多项式核函数以及sigmoid核函数对该模型的评价均较高,提示该模型具有很高的灵活性与稳定性,分类能力较强,适用性广。

本研究仍存在一些局限性,如基因表达数据可能受到样本来源、实验条件等其他因素影响,生物信息学预测结果的准确性需要进一步通过实验验证等。为了克服这些局限性,未来研究应更加注重

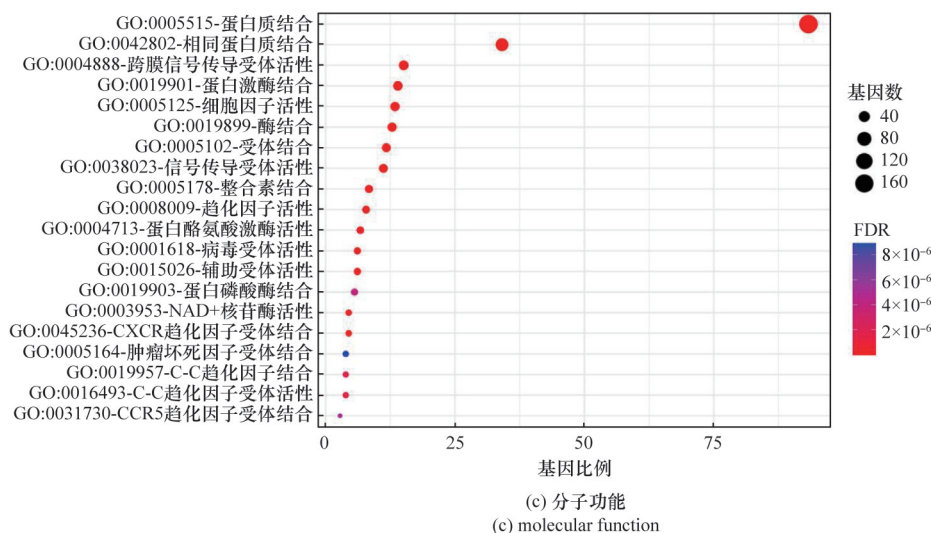
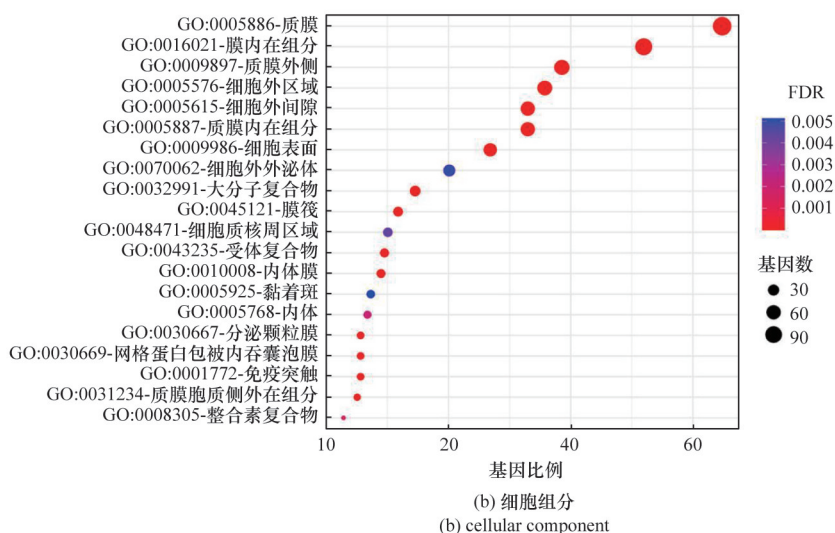
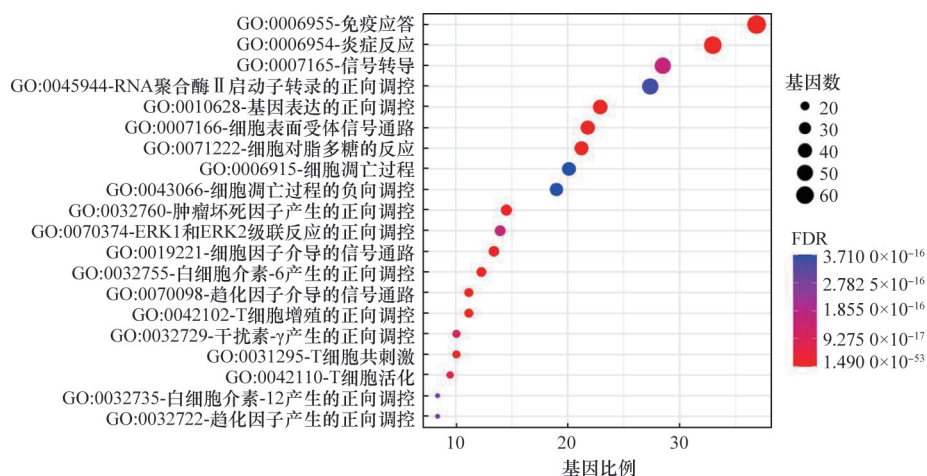


图3 GO富集分析结果

Figure 3 Results of GO enrichment analysis

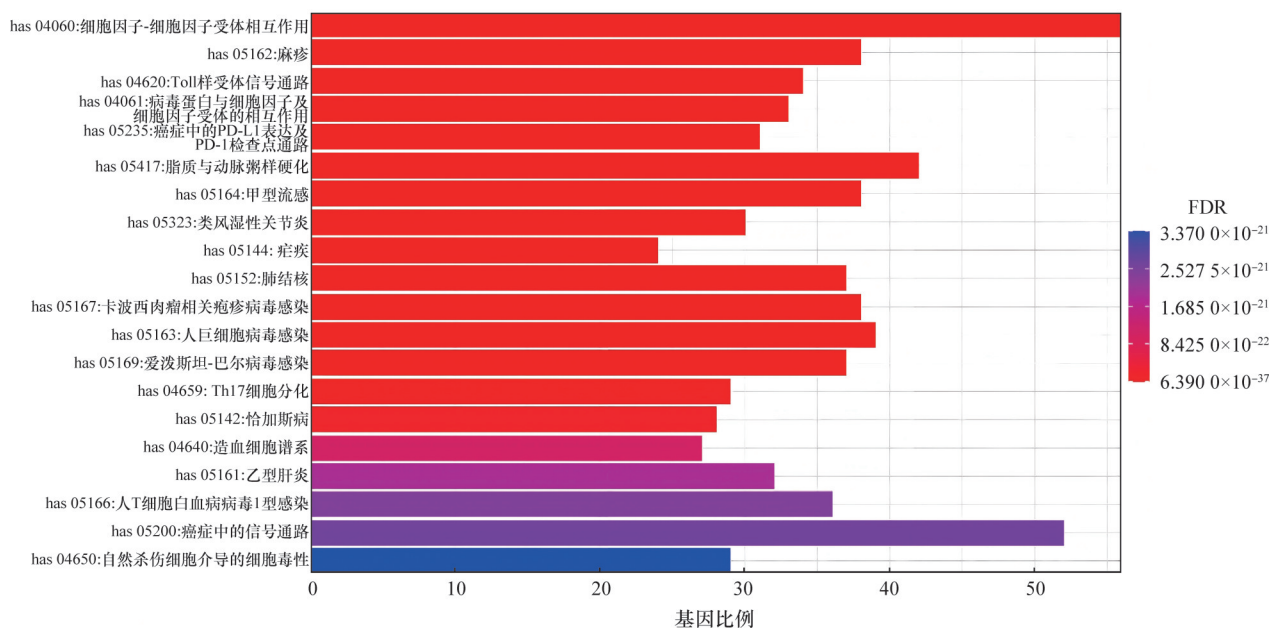


图 4 KEGG 富集分析结果

Figure 4 Results of KEGG enrichment analysis

表 1 预测模型性能评价

Table 1 Evaluation of prediction model performance

核函数	评价标准/%			
	ACC	SE	SP	MCC
线性	98.32	99.23	92.15	93.11
多项式	90.11	84.61	96.58	83.69
高斯	96.98	95.77	99.02	96.28
sigmoid	96.32	98.09	93.13	97.35

数据的质量控制,采用更为严格的筛选标准和验证方法,以提高结果的可靠性和准确性。

4 结论

本研究通过综合运用生物信息学分析方法,筛选出 179 个与 LUAD 紧密相关的关键基因,并通过功能富集分析揭示了这些基因在细胞免疫反应、炎症反应及肿瘤坏死因子调控等关键生物过程中的潜在作用。尽管研究结果仍需实验验证,但本研究为 LUAD 的分子机制探索提供了新的视角和线索,也为未来的肺癌诊断、治疗和预后评估提供了宝贵的理论依据。这些新发现不仅丰富了关于 LUAD 病理机制的认识,也为肺癌的精准治疗开辟了新的途径与策略。

参考文献

[1] 姜力文,南群.拉曼光谱技术在早期肺癌诊断中的应用及其

研究进展[J].北京生物工程,2023,42(6):643-647,653.

Jiang LW, Nan Q. The application and research status of Raman technology in the diagnosis of early lung cancer [J]. Beijing Biomedical Engineering, 2023,42(6):643-647,653.

[2] Filetti M, Lombardi P, Falcone R, et al. Comparing efficacy and safety of upfront treatment strategies for anaplastic lymphoma kinase-positive non-small cell lung cancer: a network meta-analysis[J]. Exploration of Targeted Anti-tumor Therapy, 2023, 4(6): 1136-1144.

[3] Fujikawa R, Muraoka Y, Kashima J, et al. Clinicopathologic and genotypic features of lung adenocarcinoma characterized by the international association for the study of lung cancer grading system [J]. Journal of Thoracic Oncology, 2022, 17(5): 700-707.

[4] Bertoglio P, Aprile V, Ventura L, et al. Impact of high-grade patterns in early-stage lung adenocarcinoma: A multicentric analysis[J]. Lung, 2022, 200(5): 649-660.

[5] Alharbi F, Vakanski A. Machine learning methods for cancer classification using gene expression data: A review [J]. Bioengineering (Basel, Switzerland), 2023, 10(2): 173-198.

[6] Schettini F, Brasó -Maristany F, Kuderer NM, et al. A perspective on the development and lack of interchangeability of the breast cancer intrinsic subtypes [J]. NPJ Breast Cancer, 2022, 8(1): 85-88.

[7] Wang GL, Wang FF, Meng ZS, et al. Uncovering potential genes in colorectal cancer based on integrated and DNA methylation analysis in the gene expression omnibus database [J]. BMC Cancer, 2022, 22(1): 138-150.

- [8] Wen SD, Peng WW, Chen YZ, et al. Four differentially expressed genes can predict prognosis and microenvironment immune infiltration in lung cancer: a study based on data from the GEO[J]. *BMC Cancer*, 2022, 22(1): 193–204.
- [9] 潘以红, 范雪萌, 朱平. 肺癌关键基因筛选及预测模型的研究[J]. *生物学杂志*, 2019, 36(2): 22–27, 40.
- Pan YH, Fan XM, Zhu P. Research on key gene and prediction model associated with lung cancer [J]. *Journal of Biology*, 2019, 36(2): 22–27, 40.
- [10] Xu JY, Zhang C, Wang X, et al. Integrative proteomic characterization of human lung adenocarcinoma [J]. *Cell*, 2020, 182(1): 245–261.
- [11] Ren J, Zhang H, Wang JN, et al. Transcriptome analysis of adipocytokines and their-related lncRNAs in lung adenocarcinoma revealing the association with prognosis, immune infiltration, and metabolic characteristics [J]. *Adipocyte*, 2022, 11(1): 250–265.
- [12] Kondoh N, Mizuno-Kamiya M. The role of immune modulatory cytokines in the tumor microenvironments of head and neck squamous cell carcinomas[J]. *Cancers*, 2022, 14(12): 2884–2901.
- [13] Martínez-Sabadell A, Arenas EJ, Arribas J. IFN γ signaling in natural and therapy-induced antitumor responses [J]. *Clinical Cancer Research*, 2022, 28(7): 1243–1249.
- [14] Zhang CQ, Zhang GC, Sun N, et al. Comprehensive molecular analyses of a TNF family-based signature with regard to prognosis, immune features, and biomarkers for immunotherapy in lung adenocarcinoma [J]. *EBioMedicine*, 2020, 59: 102959.
- [15] Felcher CM, Bogni ES, Kordon EC. IL-6 cytokine family: A putative target for breast cancer prevention and treatment [J]. *International Journal of Molecular Sciences*, 2022, 23(3): 1809–1823.
- [16] Liu CM, Yang L, Xu HY, et al. Systematic analysis of IL-6 as a predictive biomarker and desensitizer of immunotherapy responses in patients with non-small cell lung cancer [J]. *BMC Medicine*, 2022, 20(1): 187–201.
- [17] Islas-Vazquez L, Aguilar-Cazares D, Galicia-Velasco M, et al. IL-6, NLR, and SII markers and their relation with alterations in CD8 $^{+}$ T-lymphocyte subpopulations in patients treated for lung adenocarcinoma [J]. *Biology*, 2020, 9(11): 376–391.
- [18] Hu LT, Deng WJ, Chu ZS, et al. Comprehensive analysis of CXCR family members in lung adenocarcinoma with prognostic values [J]. *BMC Pulmonary Medicine*, 2022, 22(1): 259–269.
- [19] Wang H, Wang Y, Luo W, et al. Integrative stemness characteristics associated with prognosis and the immune microenvironment in lung adenocarcinoma [J]. *BMC Pulmonary Medicine*, 2022, 22(1): 463–473.
- [20] Ortega MA, Boaru DL, De Leon-Oliva D, et al. PD-1/PD-L1 axis: Implications in immune regulation, cancer progression, and translational applications [J]. *Journal of Molecular Medicine (Berlin, Germany)*, 2024, 102(8): 987–1000.

(2024-07-25 收稿, 2024-09-04 修回)

(上接第 228 页)

- [16] Isensee F, Jaeger PF, Kohl SA-A, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation [J]. *Nature Methods*, 2020, 18(2): 203–211.
- [17] He KY, Peng B, Yu WB, et al. A novel Mis-Seg-Focus loss function based on a two-stage nnU-Net framework for accurate brain tissue segmentation [J]. *Bioengineering*, 2024, 11(5): 427.
- [18] Marcos F, Bruno S, Sandro Q, et al. Exploring optical flow inclusion into nnU-Net framework for surgical instrument segmentation [C]//Life and Health Sciences Research Institute (Portugal); IPCA (Portugal); KARL STORZ SE Co. KG (Germany), 2024: 1292827.
- [19] Lunardo F, Baker L, Tan A, et al. How much data do you need? An analysis of pelvic multi-organ segmentation in a limited data context [J]. *Physical and Engineering Sciences in Medicine*, 2025, 48(1): 409–419.
- [20] Zhang X, Liu X, Wang ZY. DRA U-Net: An attention based U-Net framework for 2D medical image segmentation [C]//Proceedings of the 2021 IEEE International Conference on Big Data (Big Data). Los Angeles: IEEE, 2021: 3936–3942.
- [21] Chen Y, Wu Y, Wu JH, et al. Redefining contextual and boundary synergy: a boundary-guided fusion network for medical image segmentation [J]. *Electronics*, 2024, 13(24): 4986–4986.
- [22] Evgin G. Medical image data augmentation: techniques, comparisons and interpretations [J]. *Artificial Intelligence Review*, 2023, 56(11): 41–45.
- [23] Chen J, Chen R, Wang W, et al. TinyU-Net: Lighter yet better U-Net with Cascaded Multi-receptive fields [C]//Linguraru MG, Dou Q, Feragen A, et al, eds. Medical image computing and computer assisted intervention-MICCAI 2024. Cham: Springer, 2024: 626–635.
- [24] Liu XY, Qu LH, Xie ZY, et al. Towards more precise automatic analysis: a systematic review of deep learning-based multi-organ segmentation [J]. *BioMedical Engineering OnLine*, 2024, 23(1): 52–52.

(2025-12-22 收稿, 2026-02-25 修回)