与免疫浸润相关的肺腺癌驱动基因识别

张天宇，张璐强*

内蒙古大学　内蒙古呼和浩特

【摘要】作为发病率和死亡率较高的恶性肿瘤，肺腺癌由于异质性和早期诊断不足，导致其预后差。本文从癌症基因组图谱数据库下载了 576 个肺腺癌患者样本的临床信息和转录组数据，通过免疫分析和加权基因共表达网络分析，从与肺腺癌相关的差异表达基因中识别出 149 个与基质评分、免疫评分及肿瘤纯度高度关联的肿瘤免疫浸润相关基因。在此基础上，基于生存分析和 LASSO 回归分析，3 个与免疫浸润相关的肺腺癌驱动基因（IL16、P2RY13 和 HLA-DPB1）被识别，并用于构建风险评估模型。ROC 曲线显示，该模型可较好地模拟肺腺癌患者一年、三年和五年的生存率，预测的 AUC 值分别为 0.74、0.68 和 0.70。免疫分析显示，相较于高风险组，低风险组有更高的基质评分与免疫评分以及更低的肿瘤纯度，并且低风险组的免疫细胞富集程度显著高于高风险组。总之，这些结果有望为肺腺癌的临床研究提供理论帮助。

【关键词】肺腺癌；免疫浸润；肺腺癌驱动基因

【基金项目】国家自然科学基金（62161033）、内蒙古自治区自然科学基金（2021BS06001）、内蒙古自治区高等学校科学研究项目（NJZZ21002）

Identification of driver genes related to immune infiltration in lung adenocarcinoma

*Tianyu Zhang, Luqiang Zhang**

Inner Mongolia University, Hohhot Inner Mongolia

【**Abstract**】As a malignancy with high morbidity and mortality, lung adenocarcinoma (LUAD) has a poor prognosis due to its heterogeneity and inadequate early diagnosis. Clinical information and transcriptomic data of 576 LUAD patients were downloaded from the TCGA database, 149 immune-infiltrated genes highly associated with stromal score, immune score and tumor purity were extracted from the LUAD-related differentially expressed genes via performing immunoanalysis and weighted gene co-expression networkanalysis. On this basis, survival analysis and LASSO regression analysis were applied to these genes. Three LUAD driving genes (IL16, P2RY13 and HLA-DPB1) related to immune infiltration were obtained and subsequently transformed into a risk assessment model. ROC curve showed that the model could effectively simulate the survival rates of LUAD patients at one-year, three-year and five-year, and the predicted AUC results were 0.74, 0.68 and 0.70, respectively. Immunoanalysis displayed that the low-risk group got higher stromal and immune scores and lower tumor purity than those in the high-risk group, and the enrichment of immune cells in the low-risk group was significantly higher than that in the high-risk group. In summary, these results may provide theoretical guidance for the clinical studies of LUAD.

【**Keywords**】Lung adenocarcinoma; Immune infiltration; Lung adenocarcinoma driving gene

2020 年国际癌症研究机构（IARC）发布的数据显示，肺癌是全球致命性最高的癌症[1]。作为非小细胞肺癌的主要亚型[2]，肺腺癌（LUAD）的 5 年生存率不足 20%[3,4]。即使接受了早期的手术治疗，肺腺癌患者的预后依然较差，5 年内复发率为 40%[5]，生存率仅为 50-60%[6,7]。这些信息表明，处于肺癌早

*通讯作者：张璐强

期阶段的患者中存在部分高危个体。因此，寻找可信的预后生物标志物，进而利用它们对这些高危个体进行精准的前期诊断是必要的。

近年来，随着程序性死亡受体 1（PD-1）单抗在肺腺癌方面研究的深入，PD-L1（PD-1 配体）在肺腺癌免疫逃逸中的作用及预后应用价值逐渐凸显。然而，在临床实践中，仅依靠 PD-1 和 PD-L1 的表达水平来预测肺腺癌患者的生存率存在一定的局限[8]。而其它与免疫相关的基因，如 HLA 家族基因，它们的参与极大地提高了肺腺癌患者的预后[9]。这些研究提示，寻找新的、与免疫相关的生物标志物对肺腺癌的临床研究具有重要意义。

本文基于肺腺癌样本中基因的表达水平，计算了肺腺癌样本对应的基质评分（stromal score）、免疫评分（immune score）及肿瘤纯度（tumor purity）等免疫浸润得分。利用加权共表达网络算法，挖掘了与免疫浸润得分高度关联的基因。在此基础上，结合生存分析算法和 LASSO 回归算法，与免疫浸润相关的肺腺癌驱动基因被识别并用于风险评估模型构建。最后，针对构建的风险评估模型，我们执行了交叉验证和免疫分析，以验证其临床应用价值。

1 材料与方法

1.1 数据下载与预处理

517 个肺腺癌肿瘤组织样本和 59 个癌旁组织样本中的 *RNA*-seq 数据（count 格式）及临床信息数据从 TCGA 数据库下载。随后，20530 个基因在 576 个肺腺癌样本中的 count 数据被整合成 20530×576 维的数据矩阵。基于该数据矩阵和主成分分析算法，三个离群的肿瘤组织样本被剔除（图 1A）。为了获得与免疫浸润相关的临床特征，20530×573 维的 count 数据矩阵被提交给 Estimate 算法[10]以计算每一样本对应的免疫评分、基质评分和肿瘤纯度。最后，将上述三类免疫浸润得分与肺腺癌患者的年龄、TNM 分期、肿瘤 STAGE 分期、患者的吸烟史整合，作为临床特征用于后续分析。

1.2 肺腺癌差异表达基因的识别及功能富集分析

本文利用公式（1），计算了肺腺癌样本中基因的表达水平、癌变过程中基因表达水平的差异程度及相应的统计显著性，并定义 $\log_2(FC)>1$ 且 $P<0.01$ 的基因为上调差异表达基因；$\log_2(FC)<-1$ 且 $P<0.01$ 的基因为下调差异表达基因。随后，这些差异表达基因被用于执行 GO 注释，以获得这些差异表达基因的富集情况。

$$E_{i,m}=\frac{C_{i,m}\times 10^9}{C_m\times L_i}$$

$$\log_2(FC)=\log_2\left(\frac{\overline{E_i^c}}{\overline{E_i^n}}\right) \qquad (1)$$

$$P=(\overline{E_i^c}-\overline{E_i^n})/\sqrt{\frac{S_{i,c}^2}{N_c}+\frac{S_{i,n}^2}{N_n}}$$

其中，$E_{i,m}$ 表示第 i 个基因在第 m 个样本中的表达水平；$C_{i,m}$ 是 *RNA*-seq 数据 map 到第 m 个样本中第 i 个基因外显子区域的 read 数；C_m 指第 m 个样本内 *RNA*-seq 数据包含的 read 总数；L_i 为第 i 个基因外显子区域的长度。$\overline{E_i^c}$ 和 $\overline{E_i^n}$ 分别表示第 i 个基因在 N_C 个肿瘤样本和 N_n 个癌旁样本中的平均表达水平，$S_{i,c}^2$ 和 $S_{i,n}^2$ 表示第 i 个基因在肿瘤样本和癌旁样本中表达水平改变的方差。

1.3 免疫浸润相关基因的筛选

结合加权基因共表达网络（weighted gene co-expression network analysis，WGCNA）算法[11]本文评估了所有参考基因与肺腺癌临床特征间的关联度，流程如下：（i）剔除在>10%的样本中表达水平为 0 的基因；（ii）计算第 i 个基因和第 j 个基因间的 pearson 相关系数 ρ_{ij}，并构建关联矩阵 $\rho=\left[\rho_{ij}\right]$；（iii）选择最佳软阈值 β，以构建加权邻接矩阵 $a_{ij}=|\rho_{ij}|^{\beta}$；（iv）构建拓扑重叠矩阵（topological overlap measure, TOM）[12]；（v）计算 TOM 的相异度（1-TOM），并使用动态树切割方法将基因分配到模块中[13]，每个模块中基因的最小数量设置为 30；（vi）计算基因模块与临床表型间的相关性，以获取与免疫浸润得分密切相关的核心基因。

1.4 肺腺癌驱动基因的挖掘

为了挖掘与免疫浸润相关的肺腺癌驱动基因，以下策略被执行：（i）利用单 cox 生存分析评估每个基因与肺腺癌患者生存时间的关联性，对数秩检验 $P<0.05$ 的基因被识别为与免疫浸润相关的种子基因；（ii）Lasso 回归和逐步回归分析算法被应用，以评估种子基因和肺腺癌患者生存时间的关联性。其中，赤池信息准则和贝叶斯信息准则最小化的基

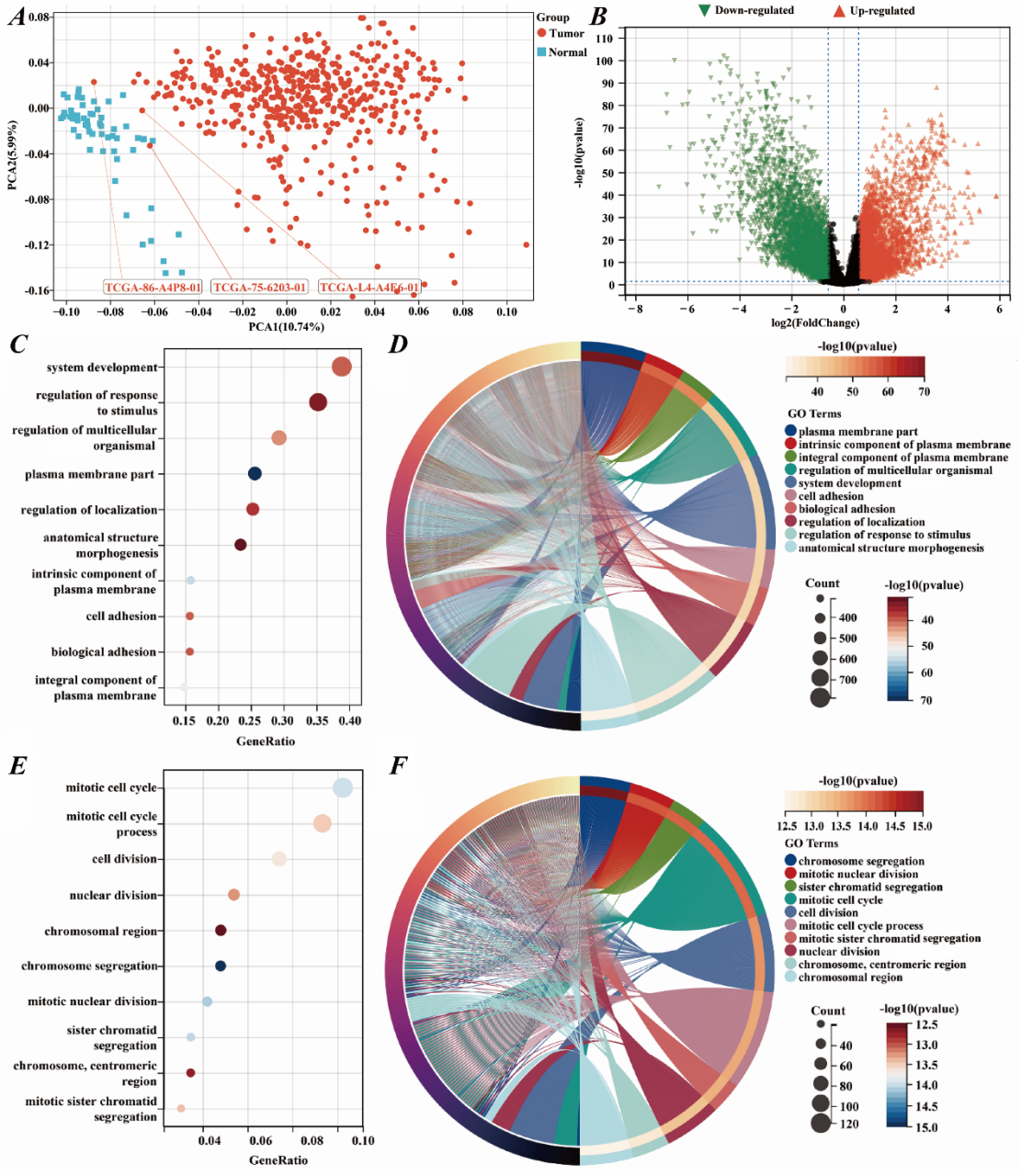
因被保留为与免疫浸润相关的肺腺癌驱动基因。

2 结果

2.1 肺腺癌样品中差异表达基因的鉴定及 GO 富集化分析

针对 576 个样本内的 20530 个基因，本文将它们的 count 数据整合成一个 20530×576 维的数据矩阵，并对该矩阵执行了主成分分析。结果显示，第一主成分（占总成份的 10.74%）和第二主成分（占总成份的 5.99%）即可很好地区分肿瘤组织样本和癌旁组织样本，根据聚类结果，TCGA-86-A4P8-01、TCGA-L4-A4E6-01 和 TCGA-75-6203-01 三个离群的肿瘤组织样本被剔除（图 1A）。

对于剩余的 514 个肿瘤组织样本和 59 个癌旁组织样本中的 count 数据，本文基于公式（1）计算了肺腺癌样本中基因的表达水平、癌变过程中基因表达水平的差异程度及相应的统计显著性。根据计算结果，1820 个上调表达基因和 2626 个下调差异表达基因分别被获得（图 1B）。



（A）肺腺癌肿瘤组织和癌旁组织的主成分分析结果。（B）差异表达基因的火山图。（C）和（E）下调和上调差异表达基因参与的前 10 个生物学过程；（D）和（F）下调和上调差异表达基因参与生物学过程间的互作用关系。

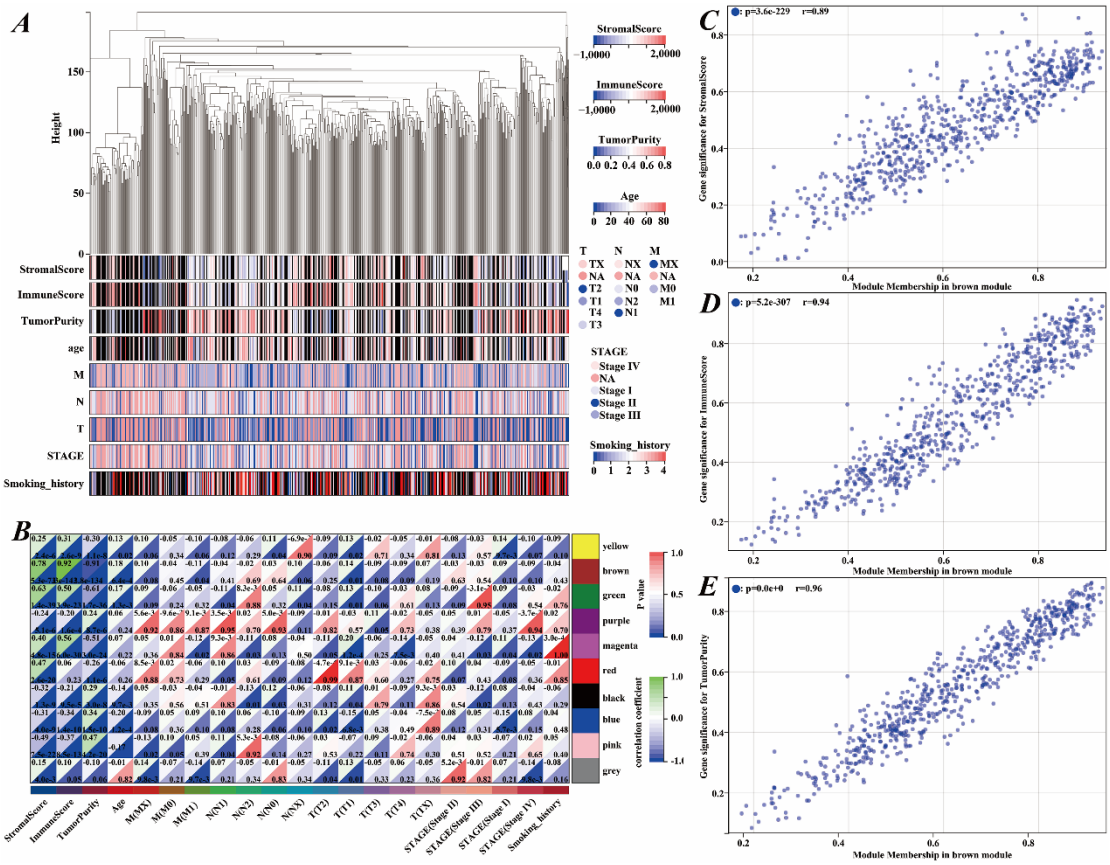
图 1 肺腺癌样品中差异表达基因的识别及功能富集化分析。

为了探究获得的差异表达基因是否参与了肺腺癌的发生和发展，针对这些差异表达基因，我们分别执行了 GO 功能富集分析。其中，下调差异表达基因主要参与生物膜系统调节、跨膜运输等过程（图 1C）；上调差异表达基因主要参与细胞分裂、细胞周期调节、染色体分离等生物学过程（图 1E）。相应生物学过程间的互作用关系见图 1D 和 1F 所示。这些与肺腺癌发生相关的生物学过程[14]表明，被识别的差异表达基因参与了肺腺癌的发生和发展。

2.2 *与免疫浸润相关的驱动基因鉴别*

为了检验所选特征是否具有临床意义，我们利用免疫评分、基质评分、肿瘤纯度、年龄、TNM 分期、肿瘤 STAGE 分期、患者的吸烟史等 9 类特征对 514 个肿瘤组织样本和 59 个癌旁组织样本进行了聚类分析（距离参数选择欧式距离），结果见图 2A 所示。从图中可知，所有样本被显著地分成了两类，表明这些特征与肺腺癌的发生和发展密切相关。

随后，利用 1.3 节所述方法计算了 4446 个差异表达基因与 9 类临床特征间的关联性，结果见图 2B。从图中，我们可以获知 brown 基因模块与基质评分、免疫评分以及肿瘤纯度间具有强相关性，pearson 相关系数分别达到 $\rho_{stromal}$=0.78（P=5.3×10^{-73}）、ρ_{immune}=0.92（P=1.3×10^{-143}）和 $\rho_{tumor\ purity}$=-0.91（P=2.8×10^{-134}）。在此基础上，我们进一步计算了 brown 模块对三种免疫浸润得分的 MM（module membership，即模块中基因表达水平与该模块的相关性，0 表示不相关，1 表示强相关）与 GS（gene significance，即模块中基因表达水平和基质评分、免疫评分以及肿瘤纯度等临床特征间的相关性）之间的 pearson 相关系数。结果分别为 $\rho_{stromal}$=0.89（P=3.6×10^{-229}）、ρ_{immune}=0.94（P=5.2×10^{-307}）和 $\rho_{tumor\ purity}$=0.96（P=0.0）。高强度的关联性表明 brown 基因模块中的基因和基质评分、免疫评分以及肿瘤纯度等临床特征密切关联。最后，通过设置截断阈值 |MM|>0.8 和|GS|>0.1，149 个与肿瘤免疫浸润相关的基因被识别。



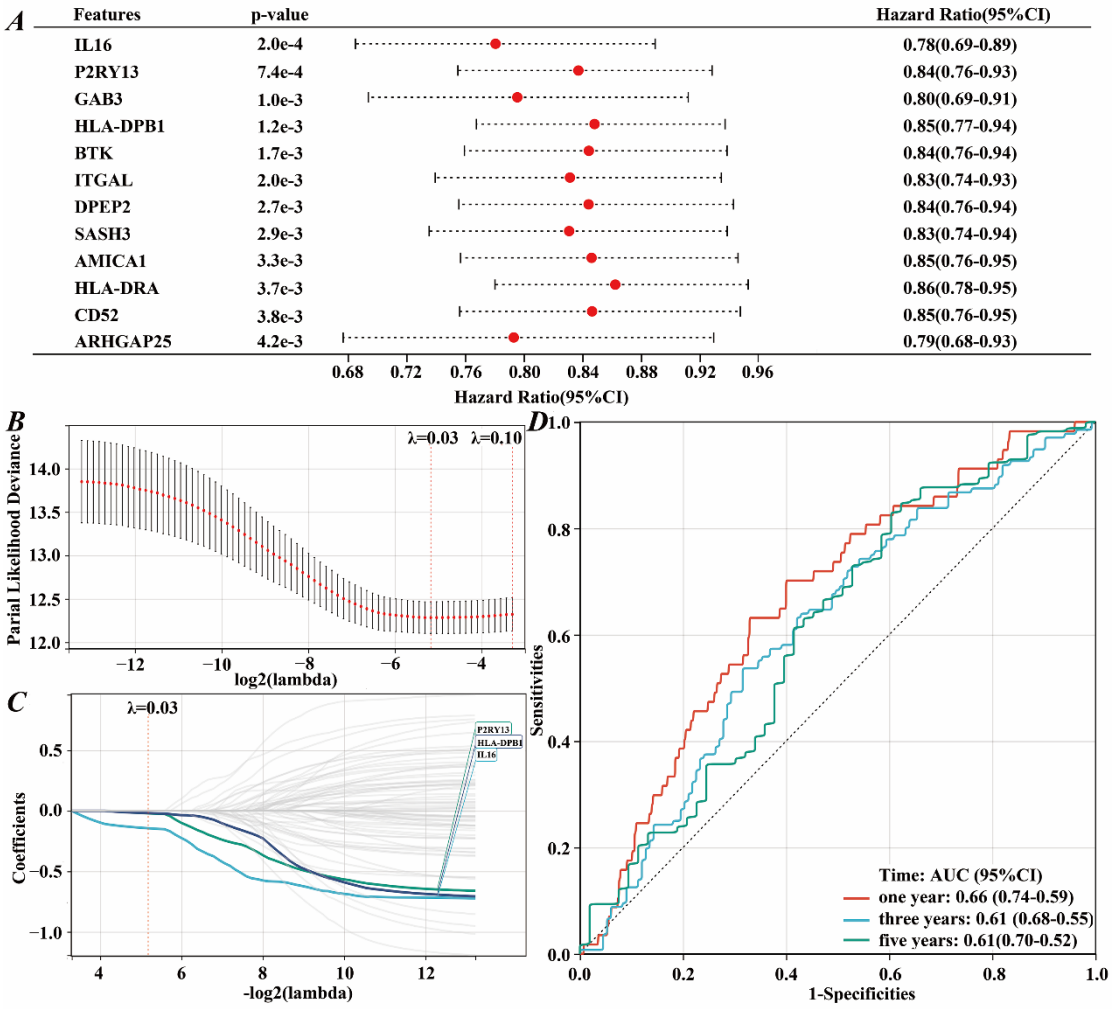
（A）样本聚类图。（B）基因模块与临床特征间的相关性热图。（C）brown 模块内样本基质得分的 GS 与 MM 间的相关性散点图。（D）brown 模块内样本免疫得分的 GS 与 MM 间的相关性散点图。（E）brown 模块内样本肿瘤纯度的 GS 与 MM 间的相关性散点图。

图 2 免疫浸润相关基因的识别

为了进一步挖掘与免疫浸润相关的肺腺癌驱动基因，上述 149 基因被提交到单变量 cox 分析算法中，以评估每个基因与肺腺癌患者生存时间的关联性。60 个对数秩检验 P<0.05 的基因被保留为与免疫浸润相关的种子基因。部分基因与肺腺癌患者生存间的关联性如图 3A 所示。之后，结合 LASSO 回归分析，本文利用 60 个种子基因在肺腺癌患者中的表达水平来模拟肺腺癌患者的生存状态。在保证 10 折交叉验证中赤池信息准则和贝叶斯信息准则最小化的情形下，本文设置 λ=0.03（图 3B）。最终，IL16、P2RY13 和 HLA-DPB1 被识别（图 3C）。其中，IL16 主要诱导人体淋巴细胞的 IL-2R 表达，而这一基因的低表达已被证实与肺腺癌的发生相关[15]；P2RY13 主要参与细胞外 ATP 的代谢，其低表达可使细胞外产生炎症环境，以抑制免疫细胞，最终帮助肿瘤细胞免疫逃逸。此外，Lin 等人[16]的研究已表明 P2RY13 可作为肺腺癌免疫相关的预后生物标记物。尽管没有研究表明，HLA-DPB1 与肺腺癌的发生有直接关系，但作为人类白细胞抗原的一种，HLA-DPB1 的主要作用是为 T 细胞提供抗原使其激活[17]。

2.3 风险评估模型的构建及临床意义检验

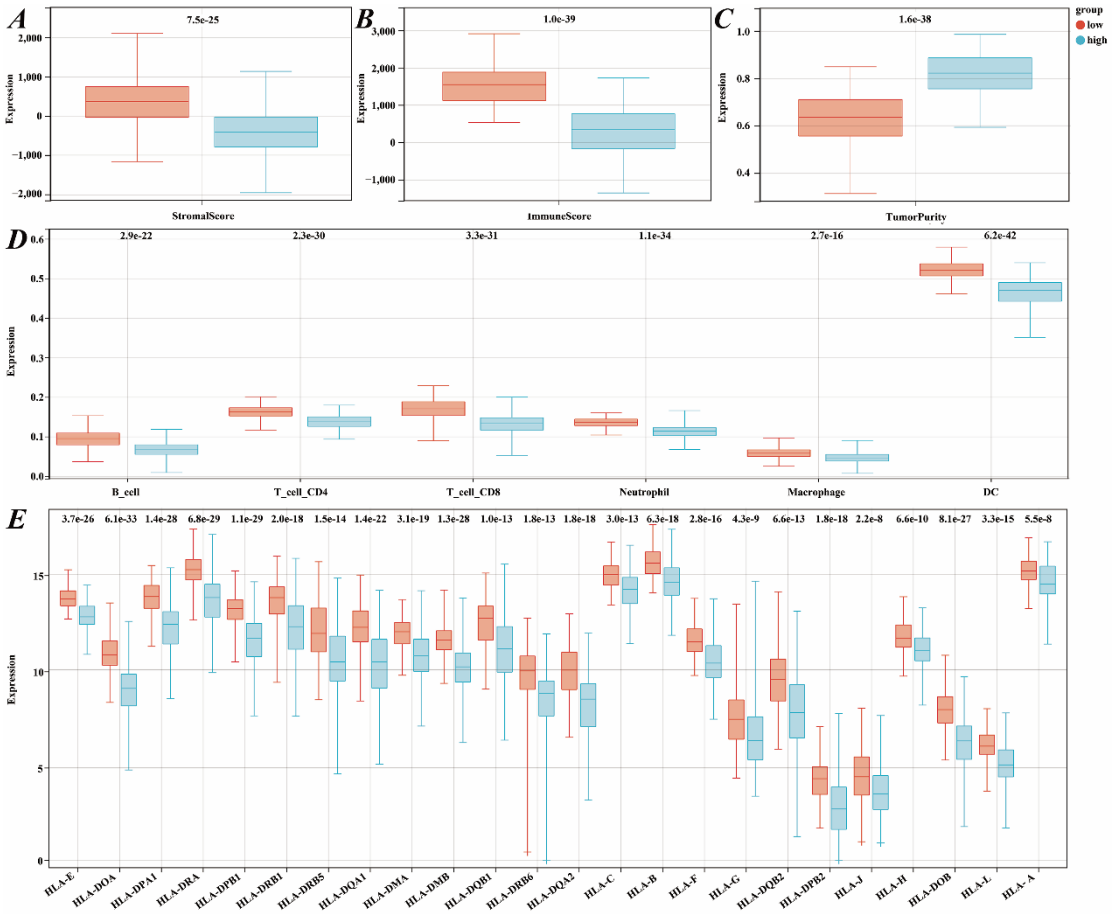
为了加速 IL16、P2RY13 和 HLA-DPB1 的临床应用，本文利用肺腺癌肿瘤组织中三者的平均表达水平及 LASSO 分析获得的回归系数，构建了如下风险评估模型：RiskScore=-0.144× $\bar{E}^{c}_{IL16}$ -0.021× $\bar{E}^{c}_{P2RY13}$ -0.019× $\bar{E}^{c}_{HLA-DPB1}$。为了评估该模型的稳健性，我们利用此模型来模拟肺腺癌患者一年、三年和五年的生存率，交叉验证结果如图 3D 所示。由图可知，该模型在一年、三年和五年的维度上均呈现出较好的效能，AUC 的最大值分别达到 0.74、0.68 和 0.70。



（A）与肺腺癌显著相关的部分基因森林图。（B）十折交叉验证结果，由图我们选择 λ=0.03。（C）LASSO 回归分析筛选肺腺癌驱动基因。（D）一年、三年和五年情形下风险得分模型的模拟结果。

图 3 肺腺癌驱动基因识别及风险评估模型构建

为了进一步检验该模型的临床应用价值，本文首先基于此模型计算了所有肺腺癌肿瘤组织的 risk score，并按照 risk score 的中值将样本分为高、低风险组。通过比较两组内样本的基质评分、免疫评分以及肿瘤纯度，我们发现，相较于高风险组，低风险组有更高的基质得分与免疫得分以及更低的肿瘤纯度（图 4A-4C）。其次，通过对高、低风险组执行 timer 分析，我们发现低风险组富集了更多免疫细胞，这些免疫细胞包括 B 细胞、CD4+T 细胞、CD8+T 细胞、中性粒细胞、巨噬细胞等（图 4D）。最后，我们进一步计算了 24 个 HLA 免疫检查点家族基因在高、低风险组的表达水平。Wilcoxon 检验显示，24 个 HLA 家族基因在高、低风险组呈现出显著的统计差异性，且 24 个 HLA 基因在低风险组表达水平更高（图 4E）。这些结果在揭示低风险组患者具有更高生存率的同时，也表明基于 IL16、P2RY13 和 HLA-DPB1 构建的风险得分模型可用于评估肺腺癌患者是否对免疫治疗更加敏感。



高、低风险组病例基质评分（A）、免疫评分（B）和肿瘤纯度（C）的分布。（D）高、低风险组内各种免疫细胞的富集程度。（E）高、低风险组内 HLA 家族基因的表达水平。

图 4 风险得分模型临床价值评估

3 结论

本文以肺腺癌为研究对象，通过对肺腺癌样本的转录组数据和临床数据进行免疫分析、加权基因共表达网络分析和生存分析，3 个与免疫浸润相关的肺腺癌驱动基因（IL16、P2RY13 和 HLA-DPB1）被识别。为了加快这些驱动基因的临床应用，一个评估肺腺癌患者风险的数理模型被构建，模型如下：RiskScore=-0.144× $\bar{E}^{c}_{IL16}$ -0.021× $\bar{E}^{c}_{P2RY13}$ -0.019× $\bar{E}^{c}_{HLA-DPB1}$。ROC 曲线显示，该模型可以很好地评估肺腺癌患者在一年、三年和五年的生存率，预测的 AUC 结果分别为 0.74、0.68 和 0.70。免疫分析显示，相较于高风险组，低风险组有更高的基质评分与免疫评分以及更低的肿瘤纯度，并且低风险组的免疫细胞富集程度显著高于高风险组。这些结果共同暗示，基于驱动基因 IL16、P2RY13 和 HLA-DPB1 的表达水平构建的风险得分模型可被用于评估患者是

否对免疫治疗具有较好疗效。

参考文献

[1] Sung H, Ferlay J, Siegel RL,et al. Global cancer statistics 2020: GLOBOCAN Estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. Ca-Cancer J Clin, 2021, **71**(3): 209-249.

[2] Herbst RS, Morgensztern D,Boshoff C. The biology and management of non-small cell lung cancer[J]. Nature, 2018, 553: 446-454.

[3] Siegel RL, Miller KD, Jemal A.Cancer statistics, 2018[J]. Ca-Cancer J Clin, 2018, 68(1): 7-30.

[4] Chen WQ, Zheng RS, Baade PD, et al. Cancer statistics in China, 2015[J]. Ca-Cancer J Clin, 2016, 66(2): 115-132.

[5] Hoffman P C,Mauer A M,Vokes EE. Lung cancer[J].Lancet, 2000, 355(9202):479-485.

[6] Nicholson, Andrew G, Crowley J, et al. The international association for the study of lung cancer lung cancer staging project: proposals for the revision of the clinical and pathologic staging of small cell lung cancer in the forthcoming eighth edition of the TNM classification for lung cancer[J]. J Thorac Oncol, 2016, 11(3): 300-311.

[7] Sawabata N, Asamura H, GoyaT, et al. Japanese lung cancer registry study: first prospective enrollment of a large number of surgical and nonsurgical cases in 2002[J]. J Thorac Oncol, 2010, 5(9): 1369-1375.

[8] Xu XL, Huang ZY, Zheng L, et al. The efficacy and safety of anti-PD-1/PD-L1 antibodies combined with chemotherapy or CTLA4 antibody as a first-linetreatmenttreatment for advanced lung cancer[J]. Int JCancer, 2018, 142(11): 2344-2354.

[9] Wu J, Li L, Zhang HB,et al. A risk model developed based on tumor microenvironment predicts overall survival and associates with tumor immunity of patients with lung adenocarcinoma[J]. Oncogene, 2021, 40(26): 1-12.

[10] Yoshihara K, Shahmoradgoli M,Martinez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data[J]. Nat Commun, 2013, 4: 2612-2623.

[11] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis[J]. Bmc Bioinformatics, 2008, 9(1): 559-572.

[12] Nakamura H, Fujii K, GuptaV,et al. Identification of key modules and hub genes for small-cell lung carcinoma and large-cell neuroendocrine lung carcinoma by weighted gene co-expression network analysis of clinical tissue-proteomes[J]. PloS one, 2019, 14(6):e0217105.

[13] Chen X, Hu L, Wang Y, et al. Single cell gene co-expression network reveals FECH/CROT signature as a prognostic marker[J]. Cells, 2019, 8(7): 698-711.

[14] 黄超.肺癌肿瘤微环境的系统解析及靶向中药发现[D]. 西北农林科技大学, 2021.

[15] 熊红.白细胞介素 16 研究进展[J].国外医学（免疫学分册）, 2002, 25(03):161-164.

[16] Lin J, Wu C, Ma D,et al. Identification of P2RY13 as an immune-related prognostic biomarker in lung adenocarcinoma: A public database-based retrospective study[J]. PeerJ, 2021, 9: 113-119.

[17] 易珊，廖娟，李寻亚，等.人 HLA-DPB1 基因及蛋白的生物信息学分析[J].基因组学与应用生物学，2019, 38(10): 4389-4394.

收稿日期：2022 年 7 月 1 日

出刊日期：2022 年 8 月 5 日

引用本文：张天宇，张璐强，与免疫浸润相关的肺腺癌驱动基因识别[J]. 国际肿瘤前沿杂志, 2022, 3(1): 1-7.

DOI: 10.12208/j.ijcan.20220001

检索信息：中国知网、万方数据、Google Scholar

版权声明：©2022 作者与开放获取期刊研究中心(OAJRC)所有。本文章按照知识共享署名许可条款发表。http://creativecommons.org/licenses/by/4.0/

OPEN ACCESS