昆明医科大学学报　2024, 45(4):9~16　　DOI: 10.12259/j.issn.2095-610X.S20240402
Journal of Kunming Medical University　　CN 53－1221/R

基于单细胞测序技术分析上皮细胞相关基因与卵巢癌患者预后的关系

赵丽珠，董　莹，邓　玥，杨丽华
（昆明医科大学第二附属医院妇科，云南 昆明　650101）

［摘要］**目的**　基于上皮细胞标志物的表达构建1个多基因风险评分来评估卵巢癌患者的预后。**方法**　对卵巢癌单细胞测序数据进行降维、聚类，识别上皮细胞标记物、恶性和非恶性标记物。使用回归分析筛选与预后相关的上皮细胞标记基因以构建风险评分模型，基于风险评分将患者分为高、低风险(H.Risk、L.Risk)组，用于预测卵巢癌患者的预后。**结果**　构建了1个4个基因(EPCAM、CLDN4、CXCR4和TIMP3)的风险评分模型。生存分析表明在试验队列和验证队列中H.Risk组患者的OS均比L.Risk组患者差($P<0.05$)。途径富集分析显示，高、低风险组之间的差异基因与免疫抑制和恶性进展相关，包括细胞粘附、细胞外基质、神经活性配体-受体相互作用、钙信号通路、转化生长因子-β等。**结论**　通过bulkRNA-seq和scRNA-seq数据的综合分析提出了1种基于上皮细胞亚群标记基因的风险评分模型，并可能为卵巢癌患者提供潜在的治疗靶点。

［关键词］卵巢癌；上皮细胞；单细胞测序；生物信息学；预后
［中图分类号］R711.75；R737.31　［文献标志码］A　［文章编号］2095－610X(2024)04－0009－08

Correlation between Epithelial Cell Related Genes and Prognosis of Patients with Ovarian Cancer based on Single Cell Sequencing

ZHAO Lizhu，DONG Ying，DENG Yue，YANG Lihua
（*Dept. of Gynecology*，*The 2nd Affiliated Hospital of Kunming Medical University*，*Kunming Yunnan* 650101，*China*）

［**Abstract**］ Objective　To construct a polygenic risk score based on the expression of epithelial cell markers to evaluate the prognosis of patients with ovarian cancer. Methods　The single cell sequencing data of ovarian cancer were reduced and clustered to identify epithelial cell markers，malignant and non-malignant markers. Regression analysis was used to screen epithelial marker genes related to prognosis to construct a risk score model. Based on the risk score，patients were divided into high risk group and low risk group（H.Risk，L.Risk）to predict the prognosis of patients with ovarian cancer. Results　A risk scoring model with four genes（EPCAM，CLDN4，CXCR4 and TIMP3）was constructed. Survival analysis showed that the OS of patients in H.Risk group was worse than that in L.Risk group in trial cohort and verification cohort（$P<0.05$）. Pathway enrichment analysis showed that the differential genes between high and low risk groups were associated with immunosuppression and malignant progression，including cell adhesion，extracellular matrix，neuroactive ligand-receptor interaction，calcium

［收稿日期］2024－01－04
［基金项目］国家自然科学基金资助项目(82360579)；云南省万人计划名医专项基金资助项目(YNWR-MY-2019-037)；昆明医科大学创新团队基金资助项目(CXTD202008)；昆明医科大学第二附属医院对外合作研究基金资助项目(2022dwhz06)；云南省科技厅-昆明医科大学应用基础研究联合专项基金资助项目(202401AY070001-053)
［作者简介］赵丽珠(1997～)，女，白族，云南剑川人，医学硕士，住院医师，主要从事妇科肿瘤研究工作。
［通信作者］杨丽华，E-mail：lihuazhang33@sina.com

signal pathway, transforming growth factor-β. **Conclusion** Through the comprehensive analysis of bulkRNA-seq and scRNA-seq data, a risk scoring model based on epithelial cell subsets marker genes is proposed, which may provide potential therapeutic targets for patients with ovarian cancer.

[**Key words**] Ovarian cancer; Epithelial cells; Single cell RNA sequencing; Bioinformatics; Prognosis

卵巢癌(ovarian cancer, OC)是女性生殖系统常见的恶性肿瘤之一[1]。原发性卵巢癌的组织学类型中90%以上为上皮性卵巢癌(epithelial ovarian cancer, EOC), EOC的发展和进展与上皮组织密切相关[2-3]。正常上皮细胞具有抗肿瘤活性,并能够通过调节细胞骨架蛋白来消除致癌转化细胞[4]。上皮细胞是大多数人类肿瘤的来源,其恶性转化通常与细胞极性的丧失和解体密切相关,并且上皮细胞极性的破坏促进上皮-间质转化(epithelial-mesenchymal transition, EMT),这是上皮肿瘤细胞侵入周围基质的关键步骤[5-6]。随着对上皮细胞研究的不断深入,上皮细胞相关生物标志物成为近年来的研究热点。单细胞转录组测序(single cell RNA sequencing, scRNA-seq)是在单细胞水平对转录组进行测序的1项新技术,可以研究单个细胞内的基因表达情况,能解决用组织样本测序无法解决的细胞异质性难题[7-8]。

为了探索上皮细胞相关基因与卵巢癌的关系,笔者使用生物信息学方法,基于卵巢癌的scRNA-seq数据对卵巢癌上皮细胞进行了更精确的分析,并联合bulk RNA-seq构建了1个基于上皮细胞标记基因的风险评分模型,可能为OC患者提供潜在的治疗靶点。

1 材料与方法

1.1 获取卵巢癌单细胞测序的转录组和临床数据

从GEO数据库(https://www.ncbi.nlm.nih.gov/geo/)获取OC单细胞测序数据(GSE118828)进行单细胞分析(包含18个样本);获得具有完整临床信息的数据集GSE140082作为模型的验证队列(包含380个OC的bulk RNA测序样本)。从TCGA数据库(https://portal.gdc.cancer.gov/)获得379份具有生存时间及生存状态的OC样本(TCGA-OV)作为模型的试验队列,从GTEx数据库获取88份非肿瘤卵巢组织的转录组数据用于差异分析。

1.2 单细胞数据处理及细胞聚类

使用R语言(4.1.2版)中的“Seurat”和“Harmony”包对GSE118828数据集进行整理并去除批次效应。“PercentageFeatureSet”包用于确定每个细胞中线粒体基因的百分比后删除线粒体基因比例>15%的细胞,避免线粒体基因的表达影响细胞分群,同时删除状态差的细胞。以上数据标准化后使用“FindVariableFeatures”函数识别前1 500个高度可变基因(highly variable features, HVGs)用于细胞分群。使用“JackStraw”函数基于HVGs进行PCA以降低维度,选择前10个PC对细胞进行聚类。应用“FindClusters”函数,根据已报道的细胞特异性标记基因对细胞簇进行注释[9-15]。

1.3 单细胞数据的拷贝组变异(copy number variation, CNV)分析

使用“copyKAT”包进行CNV分析,识别每个样本的染色体拷贝数变异情况,鉴定良恶性细胞。选取在良恶性细胞中表达差异具有统计学意义($P<0.05$)的基因作为恶性细胞标记基因。

1.4 上皮细胞相关评分模型的构建和验证

单因素Cox分析用于识别TCGA-OV队列中与生存显著相关的恶性上皮细胞标志基因。使用“Glmnet”包整合生存时间、生存状态和基因表达量,利用Lasso-cox方法进行回归分析。设置10折交叉验证以获得最优模型。Lambda值为0.002 356 510 330 766 21,确定模型公式为:RiskScore=-0.169 036 513 490 011*EPCAM+0.194 701 577 619 064*CLDN4-0.139 298 669 105 292*CXCR4+0.155 318 080 556 463*TIMP3。使用“Maxstat”包计算RiskScore的最佳截断值,基于此截断值将患者分成高、低风险2组,使用“Survival”包分析2组的预后差异。

获取患者的OS及模型风险评分,利用pROC的ROC函数进行1、3、5 a的ROC分析,并评估AUC和置信区间以获得最终的AUC结果。

1.5 TCGA队列中的功能分析

使用“Limma”包在高低风险组(H.Risk和L.Risk)间进行差异表达分析,并得到差异表达基因(differentially expressed genes, DEGs)(|logFC|≥1且$P<0.05$)。随后,使用“clusterProfiler”包和“org.Hs.eg.db”包确定了DEGs的潜在生物学机制。基于基因本体论(gene ontology, GO)和基因组百科全书(kyoto encyclopedia of genes and genomes, KEGG)对高低风险组的DEGs进行功能富集分析,确定

差异具有统计学意义的生物学功能及通路($P<0.05$)。

2　结果

2.1　高质量细胞的识别及主成分分析

通过计算，发现 GSE118828 数据集中测序深度与基因数目呈正相关，所有细胞中均未测到线粒体基因，见图 1A～1B。消除在<10 个细胞中表达的基因和表达<200 个基因的细胞，共得到 3 066 个高质量细胞。获取了在样本中波动最大的 1 500 个高变基因(HVGs)，见图 1C。使用 HVGs 进行主成分分析，可视化结果显示前 15 个 PC 均在虚线上方且靠近 y 轴，越靠近 y 轴表明实际基因与理论基因的差值越小($P<0.05$)，见图 1D。

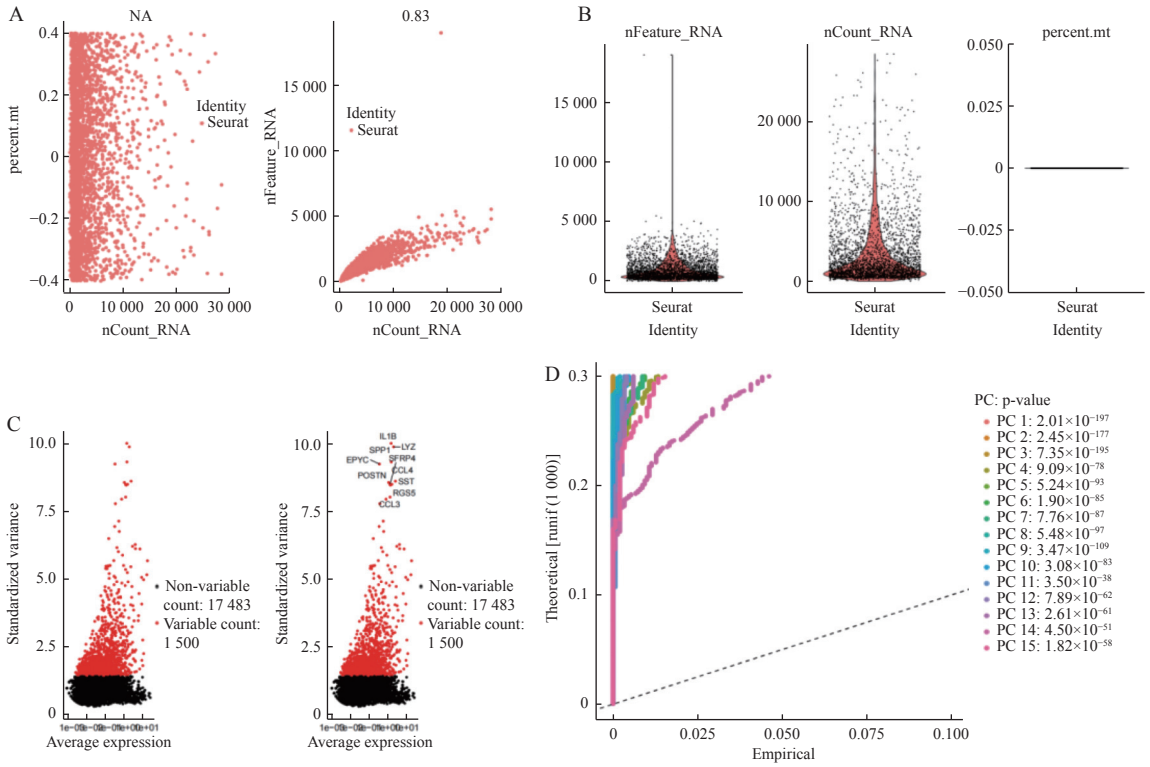

图 1　卵巢癌的单细胞 RNA 测序分析

Fig. 1　Single cell RNA sequencing analysis of ovarian cancer

A：线粒体基因与测序深度的关系(线粒体基因含量为0)和测序深度与基因数目呈正相关关系；B：线粒体基因含量；C：前 1 500 个高变基因；D：前 15 个主成分。

2.2　鉴定上皮细胞类群

将 3 066 个细胞样本根据前 10 个 PC 进行 T-SNE 聚类，共得到 13 个亚群，见图 2A。差异分析表明各亚群之间基因表达具有显著差异 P，见图 2C。对细胞亚群进行注释共得到 8 个主要细胞群(上皮细胞、T 细胞、B 细胞、单核细胞、成纤维细胞、组织干细胞、平滑肌细胞和内皮细胞)，上皮细胞在其中占了相对较大的比例，见图 2B。筛选出 56 个在上皮细胞中表达显著的基因用于后续分析($P<0.05$)。

2.3　鉴定良恶性细胞并筛选恶性基因集

CNV 将细胞样本区分为 1 530 个恶性细胞和 1 536 个非恶性细胞，两类细胞间基因表达具有明显差异 P，见图 3A。筛选出恶性和非恶性细胞间的 657 个差异基因作为恶性细胞基因集($P<0.05$)。恶性细胞基因集与上皮细胞基因集取交集共得到 45 个基因。对以上基因在 TCGA-OV 和 GTEx 中的表达进行差异分析，得到 36 个在肿瘤组织和正常卵巢组织中差异显著的恶性上皮细胞标记基因(MECRGs)，见图 3B。

2.4　筛选预后相关基因

对 MECRGs 进行单因素 Cox 回归分析，得到 4 个与预后显著相关的基因($P<0.05$)，分别为 EPCAM、CLDN4、CXCR4、TIMP3，其中 EPCAM、CXCR4 为保护性因素，CLDN4、TIMP3 为预后危险因素，见图 3C。

2.5　预后模型的构建及验证

TCGA-OV 作为试验组，Lasso-cox 分析显示以上 4 个基因均参与模型构建，见图 4A～4B。其中 EPCAM、CXCR4 的表达量随着风险评分的增

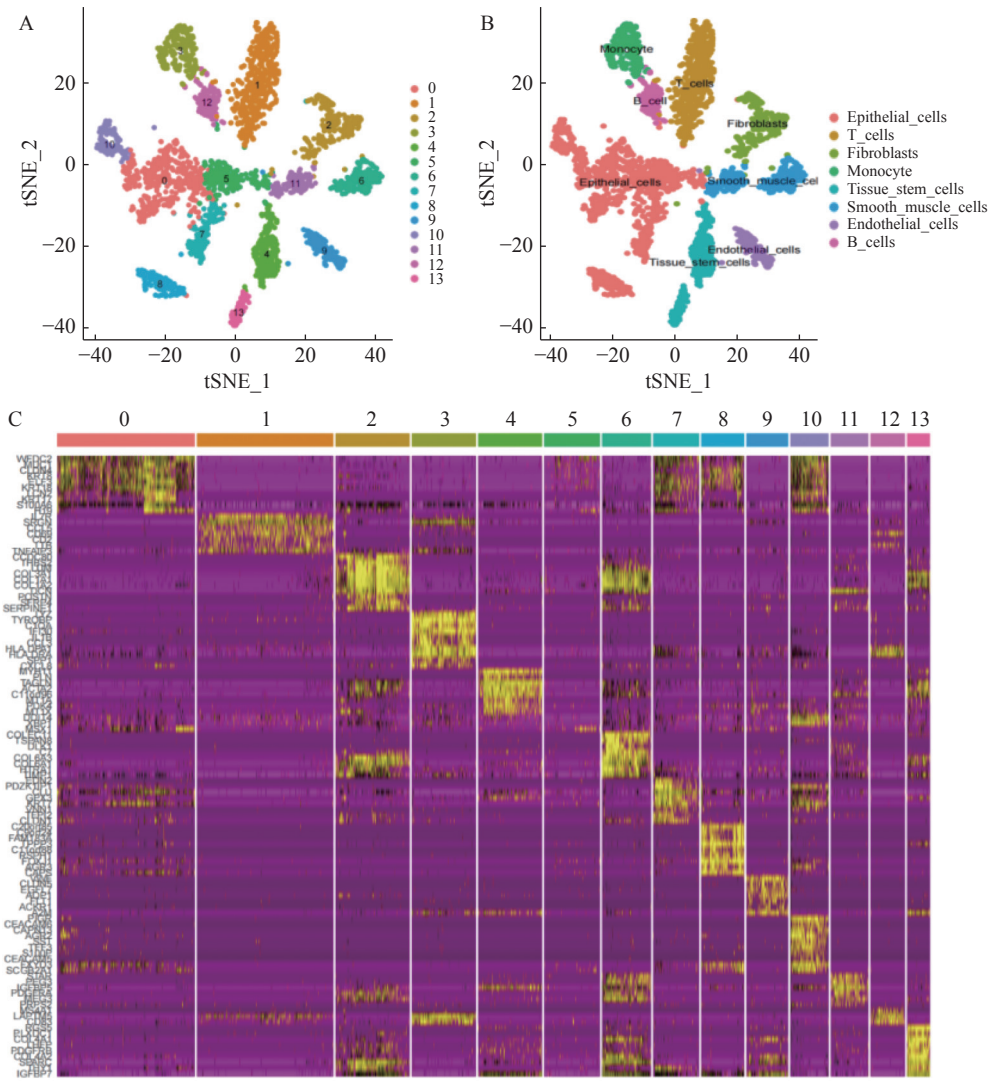

图 2 单细胞数据的降维和聚类

Fig. 2 Dimensionality reduction and clustering of single cell data

A：细胞的 T-SNE 图，显示细胞群；B：细胞注释的 T-SNE 图；C：细胞群标记物的相对表达热图(仅显示前10 名)。

加呈现下调趋势，CLDN4、TIMP3 的表达量随着风险评分的增加呈现上调趋势，见图 4C。K-M 生存分析显示 H.Risk 组患者的 OS 明显低于 L.Risk 组患者($P < 0.05$)，见图 4D。ROC 曲线显示 AUC 值在 1 a、3 a、5 a 分别达到 0.55、0.60、0.62，见图 4E。

模型基因在 GSE140082 验证集中的上下调趋势与在试验集中一致，见图 4F。K-M 生存分析显示 H.Risk 组患者的 OS 明显低于 L.Risk 组($P < 0.05$)，见图 4G。ROC 曲线显示 AUC 值在 1a、3a 分别为 0.51、0.58，见图 4H，表明随着观察时间的延长，模型准确性逐渐提高。

2.6 TCGA 队列中上皮细胞相关标记的功能富集分析

根据风险评分的最佳截断值-0.450 193 847 032 549 将 TCGA-OV 为高、低风险 2 组，差异分析显示 H.Risk 组中 311 个基因上调，174 个基因下调。DEGs 的 GO、KEGG 富集分析表明，2 组之间的差异与细胞粘附、细胞外基质、神经活性配体-受体相互作用、钙信号通路、转化生长因子-β 相关，见图 5A ~ 5B。

3 讨论

3.1 上皮相关风险模型具有一定预后价值

由于诊断和治疗的延误，OC 患者不可避免地出现不良预后。因此，有效的预后生物标志物仍然是 OC 患者迫切需要的。OC 的发生和发展与上皮组织之间存在密切关联，因此，需要进一步研究上皮细胞相关基因和 OC 患者预后之间的潜在相关性。

这项研究首次建立了 1 个基于上皮细胞相关

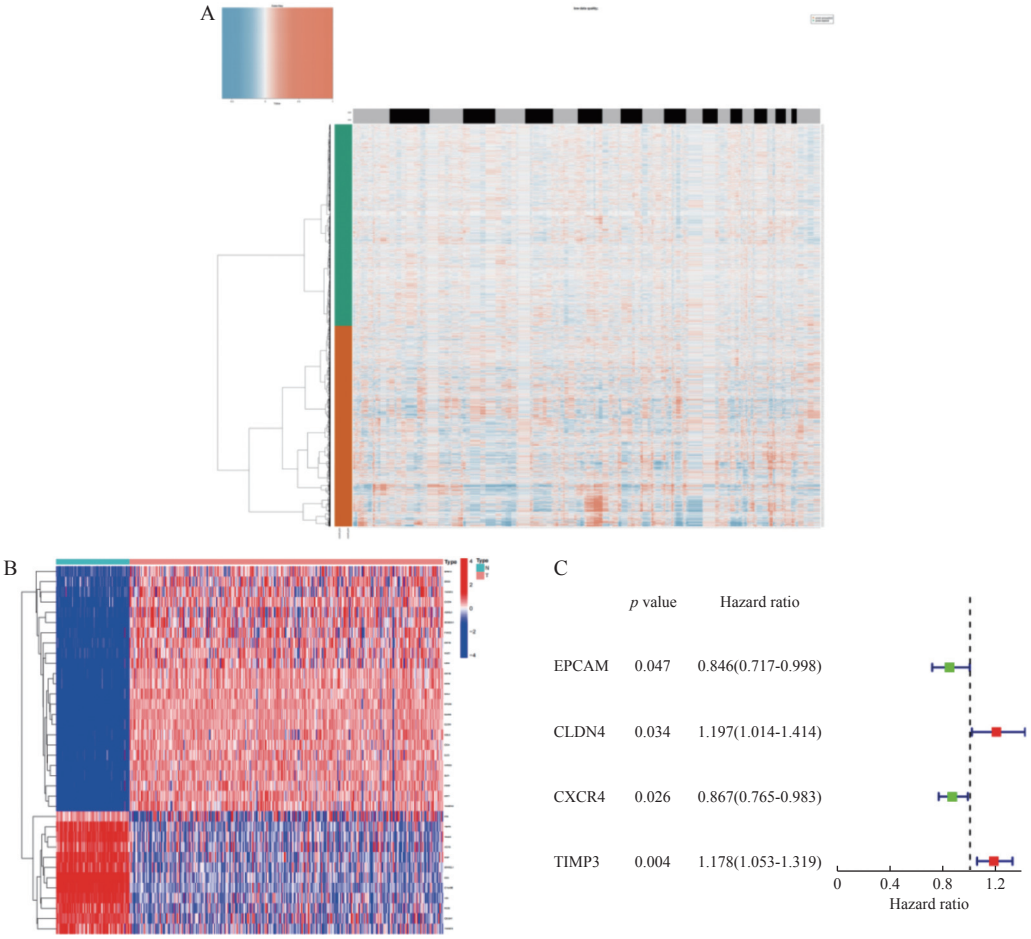

图 3　单细胞 CNV 分析及预后基因筛选

Fig. 3　Single cell CNV analysis and prognosis gene screening

A：拷贝组变异分析热图；B：MECRGs 在良恶性患者中的表达热图；C：基因表达和 OS 之间单变量 Cox 回归分析结果的森林图。

基因的 OC 风险评分模型，笔者用严格的标准和可信的算法创建了该模型。为证明风险模型的实用价值，本研究将 TCGA-OV 和 GSE140082 分别作为试验集和验证集，并根据风险评分将患者分为高、低风险组，以验证所构建的风险模型的预后价值。本究结果显示在试验集和验证集中低风险组患者的预后均优于高风险组。

3.2　模型基因可能成为卵巢癌的潜在靶点

本研究构建的模型包括 4 个基因（EPCAM、CLDN4、CXCR4 和 TIMP3）。EPCAM 表达能通过阻止细胞间黏附、促进免疫逃逸或激活致癌基因来促进癌症侵袭。既往研究表明，EPCAM 在原发食管癌中高表达，在早期食道癌中 EPCAM 表达减少会诱导 EMT，从而促进癌症进展；而在肝癌中，EPCAM 上调有促进血管生成的作用，EPCAM 阳性的肝细胞癌患者预后更差、复发风险更高[16-19]。CLDN4 在各种类型肿瘤中广泛出现高表达或表达缺失。当 CLDN4 相关蛋白发生变化时，将影响细胞间的通透性，可引起疾病的发生[20]。既往研究表明 CLDN4 在生殖系统肿瘤如卵巢癌、宫颈癌中呈现异常的高表达或低表达，CLDN4 基因表达增加可以促进卵巢癌细胞的 EMT 进程，与卵巢癌的不良预后相关[21-23]，该基因在卵巢癌中可能有着重要的应用价值。CXCR4 是 1 种已被证实的治疗靶点，当激活受体的信号途径调控失常时，CXCR4 可引起癌细胞生长及扩散。而且 CXCR4 的高表达还与许多癌症亚型的不良预后和化疗抵抗相关，部分是通过增强癌症和基质之间的相互作用[24]。TIMP3 是 1 种与细胞外基质紧密结合的蛋白质，TIMP3 对 MMP 的调控可以抑制肿瘤的生长、肿瘤细胞的侵袭和迁移。然而，TIMP3 的调控机制目前尚不清楚[25]。综上所述，EPCAM、CLDN4、CXCR4 和 TIMP3 基因与肿瘤的发生及进展密切相关，但在卵巢癌中的研究相对较少。因此，对以上 4 个基因继续深入研究有可能为卵巢癌的靶向治疗提供新的依据。

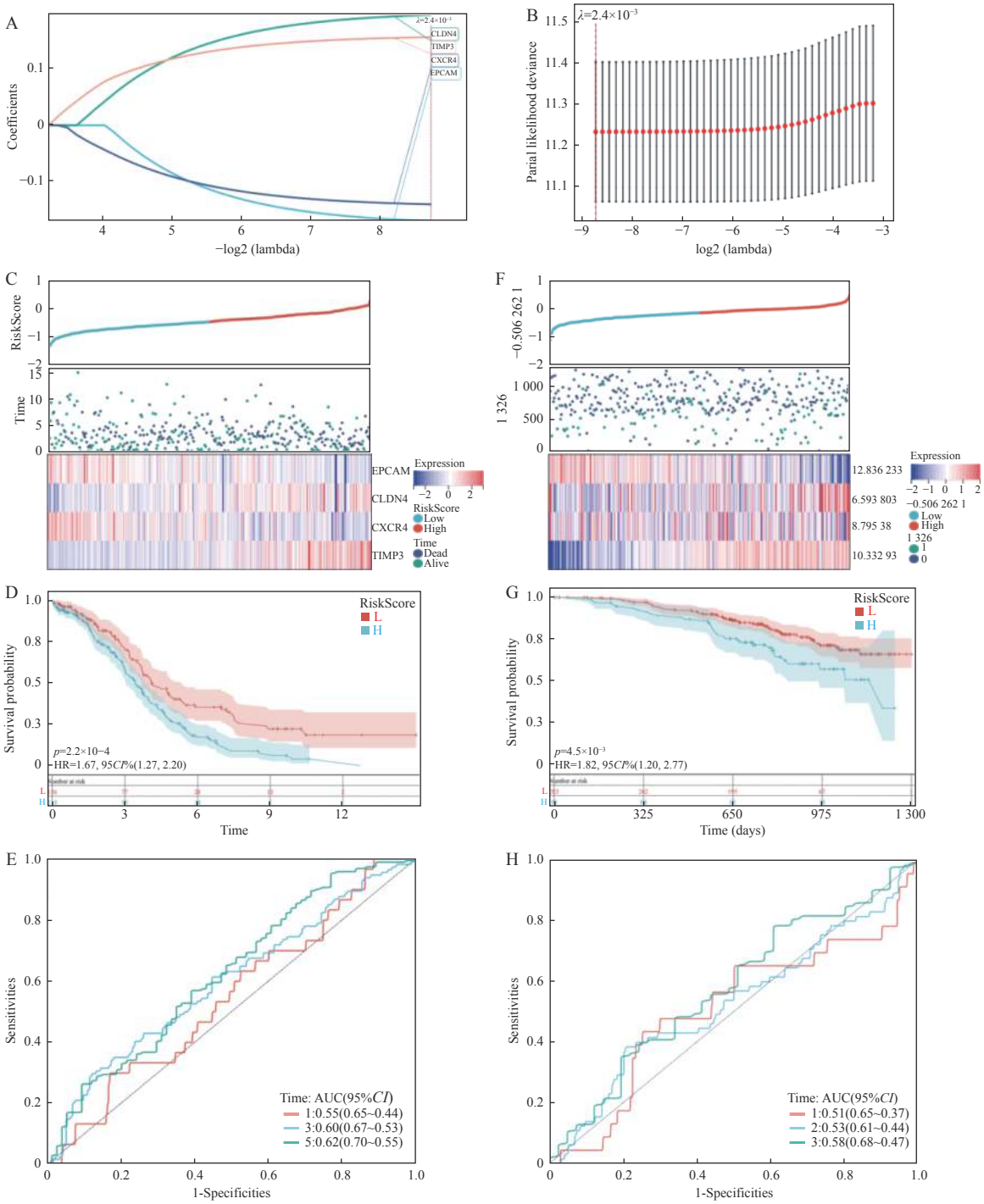

图 4　TCGA–OV 中的多基因风险评分构建与验证

Fig. 4　Construction and verification of polygene risk score in 4TCGA-OV

A ~ B：Lasso–cox 回归分析；C：TCGA 队列中 OS 状态、OS 和风险评分的分布、模型基因的表达热图；D：TCGA–OV 队列高、低风险组患者 OS 的 K–M 曲线（$P < 0.001$）；E：TCGA–OV 队列 ROC 曲线；F：GEO 队列中 OS 状态、OS 和风险评分的分布、模型基因的表达热图；G：GEO 队列高低风险组患者 OS 的 K–M 曲线（$P < 0.001$）；H：GEO 队列 ROC 曲线。

高、低风险组间差异基因的功能富集分析也表明，差异基因主要与细胞粘附、细胞外基质、神经活性配体–受体相互作用、钙信号通路、转化生长因子–β 相关。所有这些信号都被报道促进了肿瘤的发展。在前列腺癌中，细胞粘附性的增加会诱导肿瘤细胞发生上皮间质转化，促进细胞侵袭和转移[26]。在膀胱癌中，转化生长因子 β 1 在肿瘤早期抑制肿瘤增值，在晚期则通过促进上皮间质转化、诱导免疫抑制与肿瘤微环境形成，从而促进肿瘤进展[27]。因此，参与模型的基因也可能通过调节以上通路来影响卵巢癌的发生与进展。

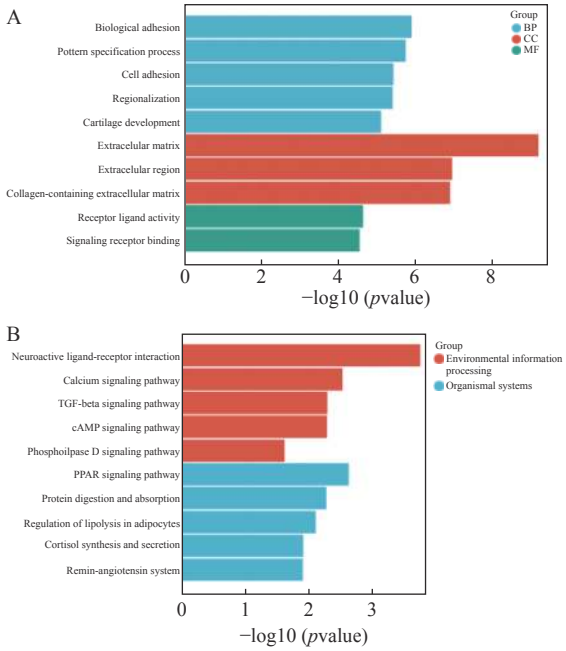

图 5　路径分析的主要结果

Fig. 5　Main results of path analysis

A ~ B：在高低风险组中差异基因的 GO 和 KEGG 富集分析。

3.3　本研究存在的局限性

本研究仍然存在一定的局限性。首先，一些临床信息的不可获得性阻碍了笔者充分探索模型和临床特征之间的关系，例如治疗的细节、肿瘤的病理学细节。其次，虽然这项研究使用了外部队列来验证模型的可靠性，但是仍需要更多的数据来进行验证。

［参考文献］

［1］ Menon U, Karpinskyj C, Gentry-Maharaj A. Ovarian cancer prevention and screening[J]. Obstetrics & Gynecology, 2018, 131(5): 909-927.

［2］ Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2008[J]. CA:A Cancer Journal for Clinicians, 2008, 58(2): 71-96.

［3］ Karantza V. Keratins in health and cancer: More than mere epithelial cell markers[J]. Oncogene, 2011, 30(2): 127-138.

［4］ Tanimura N, Fujita Y. Epithelial defense against cancer (EDAC)[C]. Seminars in Cancer Biology, 2020, 63(6): 44-48.

［5］ Royer C, Lu X. Epithelial cell polarity: A major gatekeeper against cancer?[J]. Cell Death & Differentiation, 2011, 18(9): 1470-1477.

［6］ Bai Z, Woodhouse S, Zhao Z, et al. Single-cell antigen-specific landscape of CAR T infusion product identifies determinants of CD19-positive relapse in patients with ALL[J]. Science Advances, 2022, 8(23): eabj2820.

［7］ Parker K R, Migliorini D, Perkey E, et al. Single-cell analyses identify brain mural cells expressing CD19 as potential off-tumor targets for CAR-T immunotherapies[J]. Cell, 2020, 183(1): 126-142.e17.

［8］ Varga J, Greten F R. Cell plasticity in epithelial homeostasis and tumorigenesis[J]. Nature Cell Biology, 2017, 19(10): 1133-1141.

［9］ Chen Z, Zhang H, Bai Y, et al. Single cell transcriptomic analysis identifies novel vascular smooth muscle subsets under high hydrostatic pressure[J]. Science China Life Sciences, 2021, 64(1): 1677-1690.

［10］ Pan J, Zhou H, Cooper L, et al. LAYN is a prognostic biomarker and correlated with immune infiltrates in gastric and colon cancers[J]. Frontiers in Immunology, 2019, 10(1): 6.

［11］ Lombardo G, Gili M, Grange C, et al. IL-3R-alpha blockade inhibits tumor endothelial cell-derived extracellular vesicle (EV)-mediated vessel formation by targeting the β-catenin pathway[J]. Oncogene, 2018, 37(9): 1175-1191.

［12］ Ichimiya H, Maeda K, Enomoto A, et al. Girdin/GIV regulates transendothelial permeability by controlling VE-cadherin trafficking through the small GTPase, R-Ras[J]. Biochemical and Biophysical Research Vommunications, 2015, 461(2): 260-267.

［13］ Gires O, Pan M, Schinke H, et al. Expression and function of epithelial cell adhesion molecule EpCAM: Where are we after 40 years?[J]. Cancer and Metastasis Reviews, 2020, 39(6): 969-987.

［14］ Corso G, Figueiredo J, De Angelis S P, et al. E - cadherin deregulation in breast cancer[J]. Journal of Cellular and Molecular Medicine, 2020, 24(11): 5930-5936.

［15］ Fang L, Yu G, Yu W, et al. The correlation of WDR76 expression with survival outcomes and immune infiltrates in lung adenocarcinoma[J]. Peer J, 2021, 9(10): 12277.

［16］ Liu Y, Wang Y, Sun S, et al. Understanding the versatile roles and applications of EpCAM in cancers: From bench to bedside[J]. Experimental Hematology & Oncology,

2022, 11(1): 1-19.

[17] Yahyazadeh Mashhadi S M, Kazemimanesh M, Arashkia A, et al. Shedding light on the EpCAM: An overview[J]. Journal of Cellular Physiology, 2019, 234(8): 12569-12580.

[18] Driemel C, Kremling H, Schumacher S, et al. Context-dependent adaption of EpCAM expression in early systemic esophageal cancer[J]. Oncogene, 2014, 33(41): 4904-4915.

[19] Yoon S M, Gerasimidou D, Kuwahara R, et al. Epithelial cell adhesion molecule (EpCAM) marks hepatocytes newly derived from stem/progenitor cells in humans[J]. Hepatology, 2011, 53(3): 964-973.

[20] Uthayanan L, El-Bahrawy M. Potential roles of claudin-3 and claudin-4 in ovarian cancer management[J]. Journal of the Egyptian National Cancer Institute, 2022, 34(1): 1-9.

[21] Hicks D A, Galimanis C E, Webb P G, et al. Claudin-4 activity in ovarian tumor cell apoptosis resistance and migration[J]. BMC Cancer, 2016, 16(1): 1-11.

[22] Yamamoto T M, Webb P G, Davis D M, et al. Loss of claudin-4 reduces DNA damage repair and increases sensitivity to PARP inhibitors[J]. Molecular Cancer Therapeutics, 2022, 21(4): 647-657.

[23] English D P, Santin A D. Claudins overexpression in ovarian cancer: potential targets for Clostridium Perfringens Enterotoxin (CPE) based diagnosis and therapy[J]. International Journal of Molecular Sciences, 2013, 14(5): 10412-10437.

[24] Jacobson O, Weiss I D. CXCR4 chemokine receptor overview: Biology, pathology and applications in imaging and therapy[J]. Theranostics, 2013, 3(1): 1.

[25] Zhou Y, Zhang T, Wang S, et al. Targeting of HBP1/TIMP3 axis as a novel strategy against breast cancer[J]. Pharmacological Research, 2023, 194(8): 106846.

[26] Pol M, Gao H, Zhang H, et al. Dynamic modulation of matrix adhesiveness induces epithelial-to-mesenchymal transition in prostate cancer cells in 3D[J]. Biomaterials, 2023, 299(16): 122180.

[27] 陈鑫磊, 余永波, 李鹏, 等. 转化生长因子β1对膀胱癌细胞增殖与迁移能力的影响及其机制[J]. 精准医学杂志, 2023, 38(2): 111-115.