创新团队成果专栏

基于转录组异常表达构建结直肠癌特征基因预后风险评分模型

包汝娟，陈慧芳，董　宇，叶幼琼#，苏　冰#

上海交通大学医学院上海市免疫学研究所，上海　200025

［摘要］目的·构建结直肠癌（colorectal cancer，CRC）预后风险评分模型，分析不同评分CRC患者间显著差异的肿瘤特征信号通路或生物过程，并预测该模型对其他癌症患者的免疫治疗效果。方法·从公共数据库中收集8个独立的CRC微阵列数据集和2个CRC RNA-seq数据集，筛选每个CRC数据集中的差异表达基因（differentially expressed genes，DEGs）。基于数据集共有的DEGs，采用单因素Cox回归模型筛选与不良预后相关的基因，采用套索（LASSO）回归和多因素Cox回归模型构建CRC预后风险评分模型。依据风险评分，将患者分为高风险组和低风险组。使用受试者操作特征曲线的曲线下面积（area under the curve，AUC）和Kaplan-Meier（KM）生存分析对模型性能进行评价。采用多因素Cox回归模型分析风险评分是否为CRC的独立预后因素。利用基因集富集分析（gene set enrichment analysis，GSEA）探究高、低风险组CRC患者在肿瘤特征基因集相关通路中的差异。通过KM生存分析和χ^2检验预测其他癌症患者的免疫治疗效果，以评估模型的应用价值。结果·单因素Cox回归分析，从不同数据集共有的DEGs中获得16个与不良预后相关的基因；以此为基础，构建了包含8个特征基因的CRC预后风险评分模型。该模型在训练集（AUC_{max}=0.788）、内外部验证集（AUC均值>0.600）中展现了中等程度的准确性，其低风险组患者的生存率均高于高风险组。多因素Cox回归分析显示，风险评分可作为CRC的独立预后因素。GSEA结果显示，肿瘤特征基因集相关通路在高风险组患者中显著富集。KM生存分析和χ^2检验结果显示，低风险组的其他癌症患者具有更高的生存率及更好的免疫治疗效果。结论·成功构建了含8个特征基因的CRC风险评分预后模型，可为改善CRC患者预后、预测其他癌症患者的免疫治疗效果提供参考。

［关键词］结直肠癌；套索回归；Cox回归模型；特征基因；预后模型

［DOI］10.3969/j.issn.1674-8115.2021.03.001　［中图分类号］R574.6　［文献标志码］A

Construction of prognostic risk score model of colorectal cancer gene signature based on transcriptome dysregulation

BAO Ru-juan, CHEN Hui-fang, DONG Yu, YE You-qiong#, SU Bing#

Shanghai Jiao Tong University School of Medicine, Shanghai Institute of Immunology, Shanghai 200025, China

[Abstract] Objective·To construct colorectal cancer (CRC) prognostic risk score model, analyze the significant differences of cancer hallmark signaling pathway or biological process among CRC patients with different scores, and predict the immunotherapy effect of the model on other cancer patients. Methods·Eight independent CRC microarray datasets and two CRC RNA-seq datasets were collected from a public database. Differentially expressed genes (DEGs) in each CRC dataset were screened. Based on DEGs with intersection from different datasets, univariate Cox regression model was used to screen the genes associated with adverse prognosis. LASSO regression and multivariate Cox regression models were used to construct CRC prognostic risk score model. According to the risk scores, the patients were divided into high risk group and low risk group. The area under the curve (AUC) of receiver operator characteristic curve and Kaplan-Meier (KM) survival analysis were used to evaluate the model performance. Multivariate Cox regression model was used to analyze whether risk score was an independent prognostic factor for CRC. Gene set enrichment analysis (GSEA) was used to analyze the differences of cancer hallmark gene sets-related pathways between the CRC patients in the high risk group and low risk group. KM survival analysis and chi-square test were used to predict the immunotherapy effect of other cancer patients, so as to evaluate the application value of CRC prognostic risk score model. Results·Univariate Cox regression analysis showed that 16 genes associated with adverse prognosis were obtained from DEGs with intersection from different datasets. Based on this, a CRC prognostic risk score model containing 8 gene signatures was constructed. In the training set (AUC_{max}=0.788) and internal/external validation sets (AUC_{mean}>0.600), the model displayed moderate accuracy, and the patients in the low risk group of all the above sets had significantly higher survival rate than those in the high risk group. Multivariate Cox regression analysis showed that risk score was an independent prognostic factor for CRC. GSEA results showed that cancer hallmark gene sets-related pathways were significantly enriched in CRC patients of the high risk group. KM survival analysis and chi-square test showed that other cancer patients in the low risk group had higher survival rate and better immunotherapy effect. Conclusion·The CRC risk score prognosis model containing 8 gene signatures is successfully constructed, which can provide reference for improving the prognosis of CRC patients and predicting the immunotherapy effect on other cancer patients.

[Key words] colorectal cancer (CRC); LASSO regression; Cox regression model; gene signature; prognosis model

[基金项目] 上海交通大学医学院高水平地方高校创新团队（SSMU-ZDCX20180300）。

[作者简介] 包汝娟（1991—），女，硕士；电子信箱：brj@shsmu.edu.cn。

[通信作者] 叶幼琼，电子信箱：youqiong.ye@shsmu.edu.cn。苏　冰，电子信箱：bingsu@sjtu.edu.cn。#为共同通信作者。

[Funding Information] Innovative Research Team of High-Level Local Universities in Shanghai (SSMU-ZDCX20180300).

[Corresponding Author] YE You-qiong, E-mail: youqiong. ye@shsmu.edu.cn. SU Bing, E-mail: bingsu@sjtu.edu.cn. #Co-corresponding authors.

[网络首发] https://kns.cnki.net/kcms/detail/31.2045.r.20210315.1323.002.html (2021-03-16 12:32:31)。

结直肠癌（colorectal cancer，CRC）是最常见的恶性肿瘤之一，发病率和死亡率均较高[1]。目前，CRC患者的预后主要取决于诊断时的疾病进程和肿瘤分期[2]。然而，由于缺乏适当的诊断方法，预后指标并不能满足实际的临床需求[3]。因此，开发能够有效预测患者预后的评价指标十分必要。目前，以靶向程序性死亡受体-1（programmed death-1，PD-1）为代表的免疫检查点抑制剂是CRC免疫治疗的主要研究方向，但仅有一小部分CRC患者受益，且在其他癌症中免疫治疗也并非适合于所有患者[4-5]。因此，找到能够预测癌症患者免疫治疗效果的关键特征基因，将有助于指导癌症的临床治疗。近年来，随着微阵列技术[6]和RNA测序技术的飞速发展，由此产生的转录组数据成为了探索疾病机制、挖掘各类疾病预后标志物的丰富来源[7-9]。因此，整合来源于不同技术平台的数据集，将会为全转录组水平分析肿瘤相关异常基因提供更综合的数据基础。

因此，本研究利用生物信息学技术整合公共数据库的CRC数据集，对CRC预后相关的关键基因进行筛选并构建模型，评估模型的预测性能，探索模型在预测免疫治疗效果方面的应用，为CRC患者的预后管理和个体化精准免疫治疗提供参考。

1　资料与方法

1.1　CRC相关数据的获取及整理

RNA-seq数据集：①从基因型-组织表达（genotype-tissue expression，GTEx）数据库（https：//commonfund.nih.gov/gtex）中下载308例正常结直肠组织样本的基因表达数据及对应的捐赠者临床信息（包含捐赠者的性别、组织来源等），即为GTEx数据集，将基因表达数据中每千个碱基的转录每百万映射读取的片段数（fragments per kilobase of exon model per million mapped fragments，FPKM）值进行$\log_2(x+0.001)$转换。②在癌症基因组图谱（the cancer genome atlas，TCGA）数据库（https://portal.gdc.cancer.gov）中下载471例CRC组织样本及41例正常结直肠组织样本的基因表达数据以及对应的患者临床信息（包含患者的年龄、性别、生存状态、生存时间等），即为TCGA数据集，对基因表达数据的FPKM值进行$\log_2(x+1)$转换。随后，使用sva R包的ComBat函数将GTEx数据集和TCGA数据集整合为一个数据集，即为RNA-seq数据集。

微阵列数据集：在基因表达综合（gene expression omnibus，GEO）数据库[10]（https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi）中下载10例CRC患者的基因表达数据及对应的患者临床信息（包含患者的年龄、性别、生存状态、生存时间等），包括GSE8671[11]、GSE18105[12]、GSE20916[13]、GSE23878[14]、GSE37364[15]、GSE21510[16]、GSE33113[17]、GSE39582[18]、GSE17536[19]和GSE17537[19]。使用oligo R包（http://www.bioconductor.org/packages/release/bioc/html/oligohtml）的Robust Multiarray Averaging函数对原始数据进行处理。由于前8个数据集同时包含了CRC组织和正常结直肠组织的样本，故用于后续差异分析；后3个数据集包含了模型研究部分所需的临床信息，故而作为模型验证集进行后续研究。

1.2　差异表达基因筛选及功能分析

使用R语言limma包[20]分别对每个微阵列数据集中的差异表达基因（differentially expressed genes，DEGs）进行筛选，再通过R语言RobustRankAggreg（RRA）包[21]获得前8个微阵列数据集中共有的DEGs。基因表达的差异用P值和差异倍数（fold change，FC）的对数（$\log_2$FC）表示。将$P<0.05$且$|\log_2\text{FC}|>1$的基因视为DEGs。随后，分别对RNA-seq数据集、TCGA数据集中的DEGs进行筛选，方法及筛选标准同上。使用R语言clusterProfiler包[22]分别对微阵列数据集、RNA-seq数据集获得的DEGs进行基因本体数据库（Gene Ontology，GO）功能分析，结果以$P<0.01$为入选标准。采用R语言GOplot包[23]呈现GO功能分析的结果。

1.3　预后风险评分模型的构建和性能评估

1.3.1　用于构建及评估模型的数据集说明　因建模构建需要，结合数据集的基因表达数据及临床信息中的生存状态、生存时间，从TCGA数据集中选择了438例符合上述要求的样本开展研究，其中随机抽取219例样本作为训练集，排除训练集的剩余219例样本作为内部验证集；同时，从微阵列数据集中选择GSE39582（556例样本）、GSE17536（177例样本）和GSE17537（55例样本）数据集作为外部验证集。

1.3.2　预后风险评分模型的构建及分组标准　对“1.2”中已筛选获得的DEGs进行二次筛选，获得GSE39582、TCGA数据集这2个数据共有的DEGs。采用R语言survival包[24]进行单因素Cox回归分析，筛选与不良预后相关的基因［风险比（hazard ration，HR）>1，$P<0.05$］，并将其作为构建模型的候选基因。基于训练集，通过R语言glmnet包[25]将候选基因作为参数进行套索（LASSO）回归分析，确定最佳惩罚值，而后选择其对应的回归系数

不为0的基因作为建模基因，再行多因素Cox回归分析，构建预后风险评分模型。风险评分计算公式为：风险评分=基因1表达量×多因素Cox回归系数1+…+基因N表达量×多因素Cox回归系数N（其中N代表基因数）。计算每例患者的风险评分，通过R语言survival包[24]的surv_cutpoint函数将患者划分为高风险组和低风险组。

1.3.3 在训练集中评估预后风险评分模型的性能 采用上述计算公式对训练集中每例患者的风险评分进行计算，并据此将其分为高风险组和低风险组。通过R语言ggplot2包绘制2组患者的风险评分和生存状态的分布图，以及建模基因表达量图谱。通过R语言survivalROC包[26]绘制时间依赖性受试者操作特征曲线（receiver operator characteristic curve，ROC曲线），计算训练集中患者生存时间分别在1、2、3年的曲线下面积（area under the curve，AUC）；通过R语言survminer包（https://cran.r-project.org/web/packages/survminer/index. html）绘制Kaplan-Meier（KM）生存曲线，计算2组患者的生存率。根据AUC值和组间的KM生存率的差异，评估模型在训练集中的预测性能。同时，结合训练集中CRC患者的年龄、性别、肿瘤分期等临床信息，采用多因素Cox回归模型分析风险评分是否为判断CRC不良预后的独立因素。

1.3.4 在内、外部验证集中评估预后风险评分模型的性能 采用上述计算公式对内、外部验证集中每例患者的风险评分进行计算，并据此将每个验证集患者分为高风险组和低风险组。根据内、外部验证集的ROC曲线和KM生存曲线分析结果，评估该模型在内、外部验证集中的预测性能。

1.4 基因集富集分析

为验证GSE39582数据集和TCGA数据集中低风险和高风险组间的功能差异，我们利用基因集富集分析（gene set enrichment analysis，GSEA）[27]方法比较肿瘤特征基因集（hallmark gene sets）在2组间的富集度。其中，该肿瘤特征基因集来源于Molecular Signatures Database（MSigDB）数据库（https：//www. gsea-msigdb. org/gsea/msigdb/index.jsp），包含炎症反应、低氧等50组基因集。分析输出结果为标准化富集分数（normalized enrichment score，NES），依据NES的高低衡量基因集在高、低风险组间的富集程度，从而揭示高、低风险组间差异显著的肿瘤特征信号通路或生物过程。即在P<0.05前提下，NES>1代表基因集在高风险组中显著富集，且NES越大富集程度越高；NES<-1代表基因集在低风险组中显著富集，且|NES|越大富集程度越高。

1.5 预后风险评分模型评估免疫治疗效果的应用分析

本研究依据相关参考文献，下载2个包含生存时间、生存状态和免疫治疗效果等临床信息的其他癌症的数据集，分别为348例接受程序性死亡配体-1（programmed death ligand-1，PD-L1）免疫治疗的转移性尿路上皮癌患者的基因表达数据和临床信息[28]及49例接受PD-1免疫治疗的黑色素瘤患者的基因表达数据和临床信息[29]，探究CRC预后风险评分模型在免疫治疗效果评估中的应用：①分析该模型在免疫数据集中是否成立。根据“1.3.2”中的计算公式对该2个数据集中的每例患者的风险评分进行计算，并据此将该2个数据集分别分为高风险组和低风险组。通过KM生存曲线对2组患者的生存率进行分析。②采用χ^2检验统计2个数据集中不同组患者免疫治疗效果间的差异，及其与风险评分之间的关系。

1.6 统计学方法

采用R软件（3.6.2版本）对所有数据进行统计分析。定性资料以频数（百分比）表示，采用χ^2检验进行比较。采用对数秩检验（Log-Rank法）比较所有数据集中高、低风险组患者KM生存分析间的差异。P<0.05表示差异具有统计学意义。

2 结果

2.1 CRC数据统计及分析

经整合，RNA-seq数据集的正常结直肠组织样本有349例、CRC组织样本有471例。来自GEO数据库的10个CRC数据集的基本信息见表1，因前8个数据集同时包含了CRC组织和正常结直肠组织的样本，故用于后续DEGs分析。

表1 GEO数据库的10个CRC数据集的基本信息

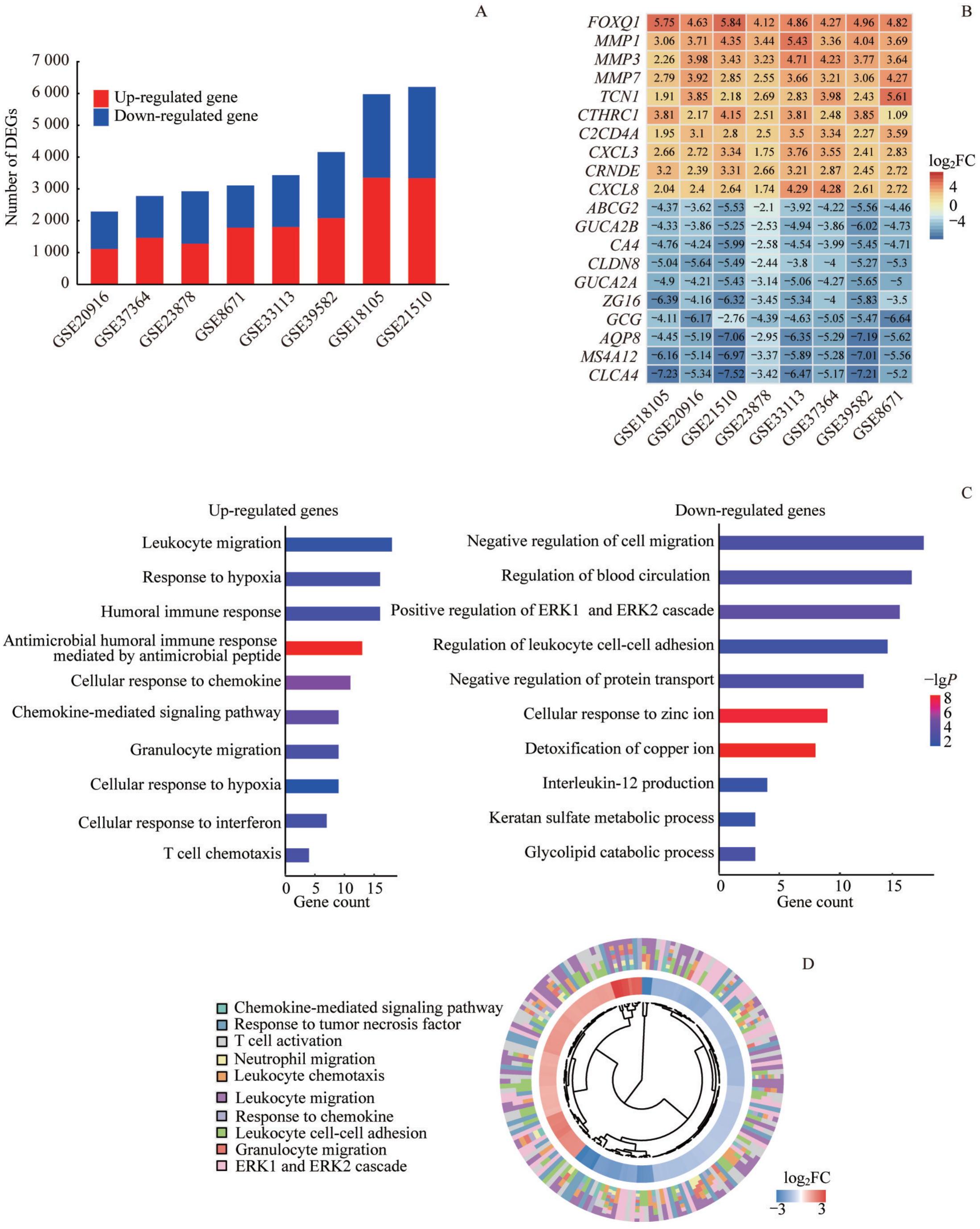
Tab 1 Basic information of ten CRC datasets from GEO database

GEO ID	Platform	Tumor (#)	Normal (#)	PMID	Country
GSE8671	GPL570	32	32	18171984	Switzerland
GSE18105	GPL570	94	17	20162577	Japan
GSE20916	GPL570	111	34	20957034	Poland
GSE23878	GPL570	35	24	21281787	Saudi Arabia
GSE37364	GPL570	56	38	25405986	Hungary
GSE21510	GPL570	123	25	21270110	Japan
GSE33113	GPL570	90	6	22496204	Netherland
GSE39582	GPL570	566	19	23700391	France
GSE17536	GPL570	177	0	19914252	USA
GSE17537	GPL570	55	0	19914252	USA

Note: (#) means number; PMID—PubMed unique identifier.

2.2 DEGs筛选及GO功能分析

经筛选，GEO数据库中前8个CRC数据集的DEGs数量如图1A所示；经第2次筛选后8个CRC数据集共有的DEGs为962个（上调427个，下调535个），其中最为显著的10个上调、10个下调的DEGs如图1B所示。GO功能分析结果（图1C）显示，上调基因主要富集在免疫细胞迁移、趋化因子介导的信号通路等，下调基因主要富集在负向调节细胞迁移等。



Note：A. Number of DEGs in the 8 CRC datasets from GEO database. B. Heatmap of the top 10 significantly DEGs with intersection in the 8 CRC datasets from GEO database. Shades of color—the degree of change in the expression. *FOXQ1*—forkhead box Q1; *MMP1*—matrix metallopeptidase 1; *TCN1*—transcobalamin 1; *CTHRC1*—collagen triple helix repeat containing 1; *C2CD4A*—C2 calcium dependent domain containing 4A; *CXCL3*—C-X-C motif chemokine ligand 3; *CRNDE*—colorectal neoplasia differentially expressed; *ABCG2*—ATP binding cassette subfamily G member 2; *GUCA2B*—guanylate cyclase activator 2B; *CA4*—carbonic anhydrase 4; *CLDN8*—claudin 8; *ZG16*—zymogen granule protein 16; *GCG*—glucagon; *AQP8*—aquaporin 8; *MS4A12*—membrane spanning 4-domains A12; *CLCA4*—chloride channel accessory 4. C. GO function analysis of DEGs with intersection in the 8 CRC datasets from GEO database. ERK1—extracellular signal-regulated kinase 1. D. GO function analysis of RNA-seq datasets. The inner dendrogram indicates the hierarchical clustering of the gene expression profiles, the outer circle represents the $\log_2$FC of each DEG, with the color corresponding to the gene level, and the outermost circle represents the biological process terms assigned to the gene.

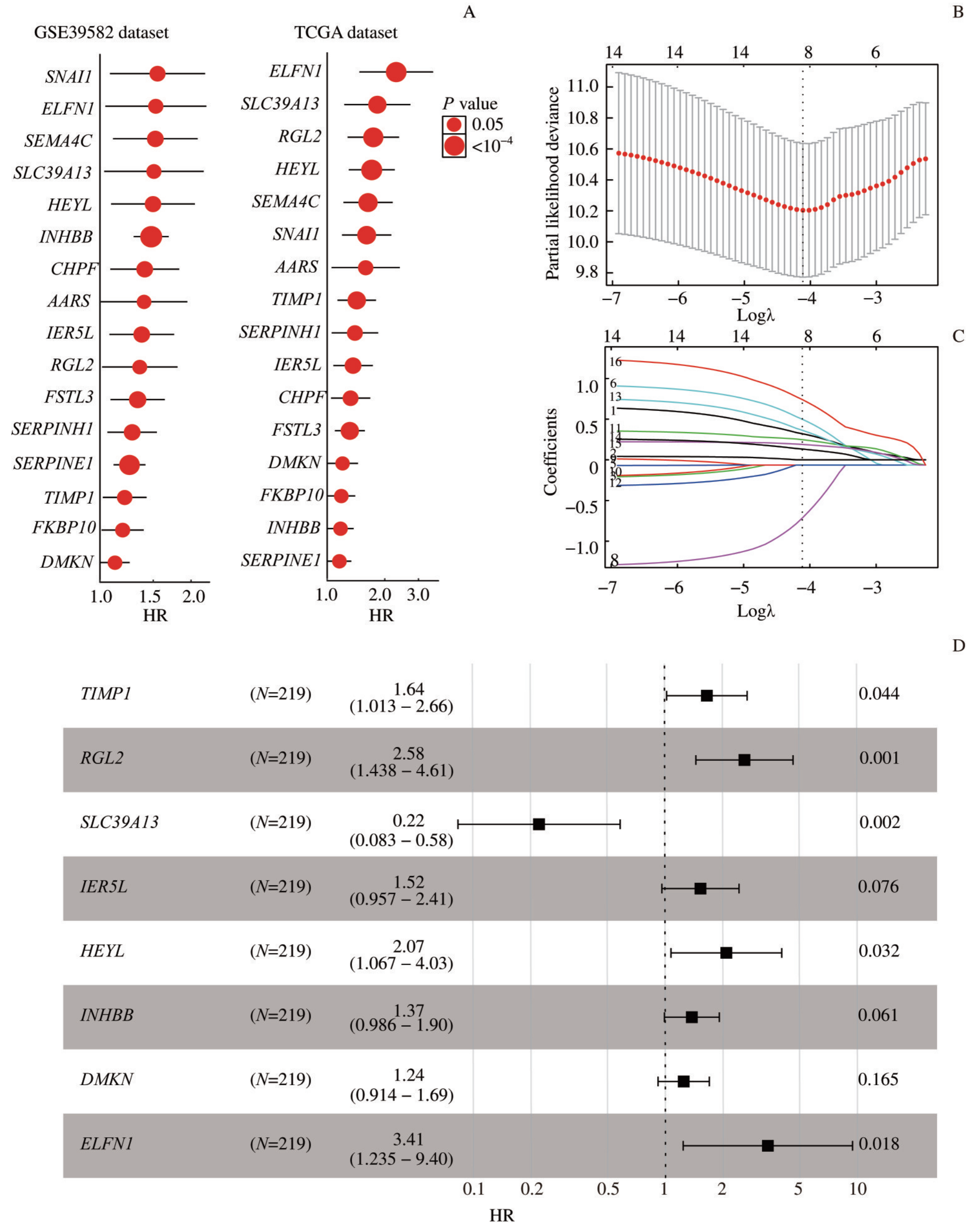
图1 CRC数据集中的DEGs筛选及其GO功能分析

Fig 1 Screening and GO functional analysis of DEGs from the CRC datasets

RNA-seq数据集中共筛选出1 749个DEGs（上调800个、下调949个）。GO功能分析结果（图1D）显示，多数DEGs富集在趋化因子介导的信号通路、免疫细胞迁移等。

2.3 CRC预后风险评分模型构建及其性能评价

采用单因素Cox回归模型对TCGA和GSE39582数据集共有的1 927个DEGs进行分析，得到16个与不良预后相关的基因（图2A）。以上述基因作为参数行LASSO回归分析，当logλ达到最小值（-4.11，图2B，即黑色虚线标注所示）时，对应的参数为最佳建模参数，而此时有8个最佳建模参数的LASSO回归系数均不为0（图2C，黑色虚线标注所示）；对该8个最佳建模基因［*TIMP1*（TIMP metallopeptidase inhibitor 1）、*RGL2*（ral guanine nucleotide dissociation stimulator like 2）、*SLC39A13*（solute carrier family 39 member 13）、*IER5L*（immediate



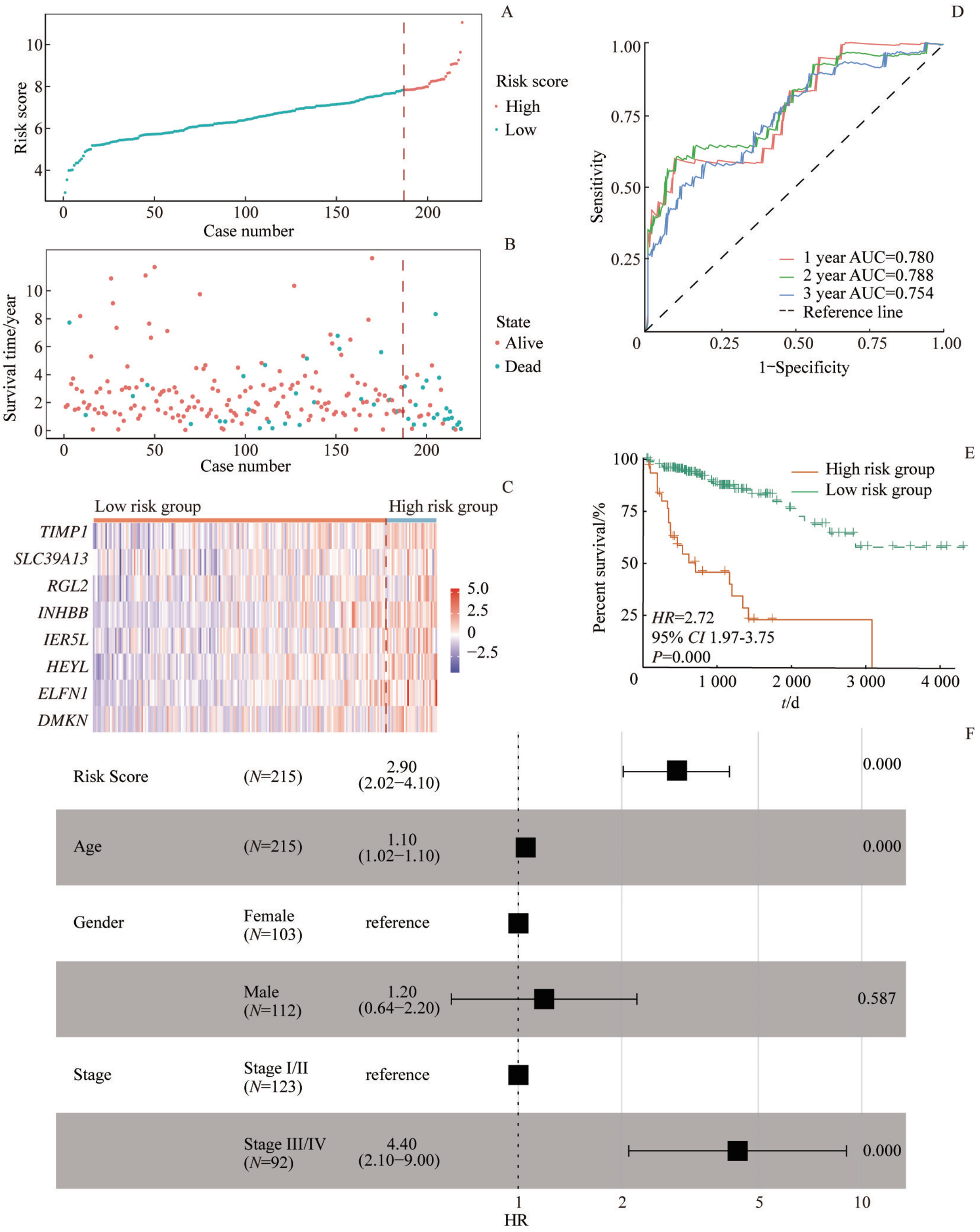
Note： A. HRs of 16 prognostic DEGs from TCGA dataset and GSE39582 dataset. *SEMA4C*—semaphorin 4C; *SNAI1*—snail family transcriptional repressor 1; *AARS*—alanyl-tRNA synthetase; *SERPINH1*—serpin family H member 1; *CHPF*—chondroitin polymerizing factor; *FSTL3*—follistatin like 3; *FKBP10*—FKBP prolyl isomerase 10; *SERPINE1*—serpin family E member 1. B. Optimal parameter selection in the LASSO regression model. C. LASSO coefficient profiles of the 16 prognostic DEGs. D. Forest plots of the 8 genes based on multivariate Cox regression model. N means patient number.

图2　CRC预后风险评分模型的构建

Fig 2　Construction of CRC prognostic risk score model

early response 5 like）、*HEYL*（hes related family bHLH transcription factor with YRPW motif like）、*INHBB*（inhibin subunit beta B）、*DMKN*（dermokine）和 *ELFN1*（extracellular leucine rich repeat and fibronectin type Ⅲ domain containing 1）］进行多因素Cox回归分析，最终得到由8个特征基因构成的CRC预后风险评分模型（图2D）。风险评分计算公式为：风险评分=*TIMP1*表达值×0.495+*RGL2*表达值×0.946+*SLC39A13*表达值×（−1.517）+*IER5L*表达值×0.418+*HEYL*表达值×0.729+*INHBB*表达值×0.315+*DMKN*表达值×0.218+*ELFN1*表达值×1.226。

通过上述计算公式，获得训练集中每例CRC患者的风险评分，并据此将患者分为高风险组（*N*=32）和低风险组（*N*=187）；不同组别的患者评分分布如图3A所示，红色虚线表示分组的分界线。与高风险组相比，低风险组患者的不良生存状态较少（图3B），且8个建模基因的表达量更低（图3C）。训练集的AUC均值大于0.750（图



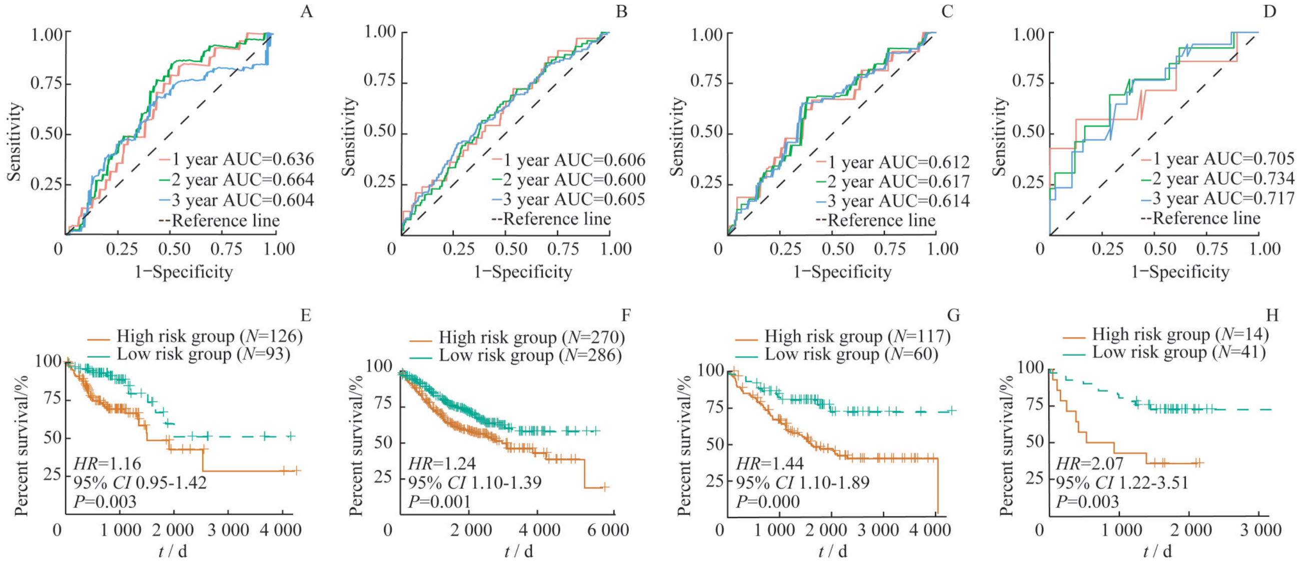
Note：A–C. The distribution of the risk scores (A), survival state (B) and modeling gene expression (C) in training set. D. ROC curves for predicting one-, two- and three-year survival in the training set. E. KM survival analysis of the two groups in the training set. F. Multivariate Cox regression analysis of patients′ clinical information in the training set.

图3　CRC预后风险评分模型在训练集中的性能评估

Fig 3　Performance evaluation of CRC prognostic risk score model in the training set

3D）；KM分析结果显示，与低风险组相比，高风险组患者的总体生存率更低（图3E）。多因素Cox回归分析的结果显示，风险评分、年龄、肿瘤分期可以作为CRC患者预后的独立因素（图3F）。以上结果证明，该模型在训练集中对CRC预后具有较高的预测价值。

随后，在内、外部验证集中评估该模型的性能。内、外部验证集的AUC均值大于0.600（图4A~D）；且该4个验证集的KM生存分析显示，与低风险组相比，高风险组患者的总体生存率更低（图4E~H）。以上结果表明，该模型在内、外部验证集中的预测能力具有中等准确度。



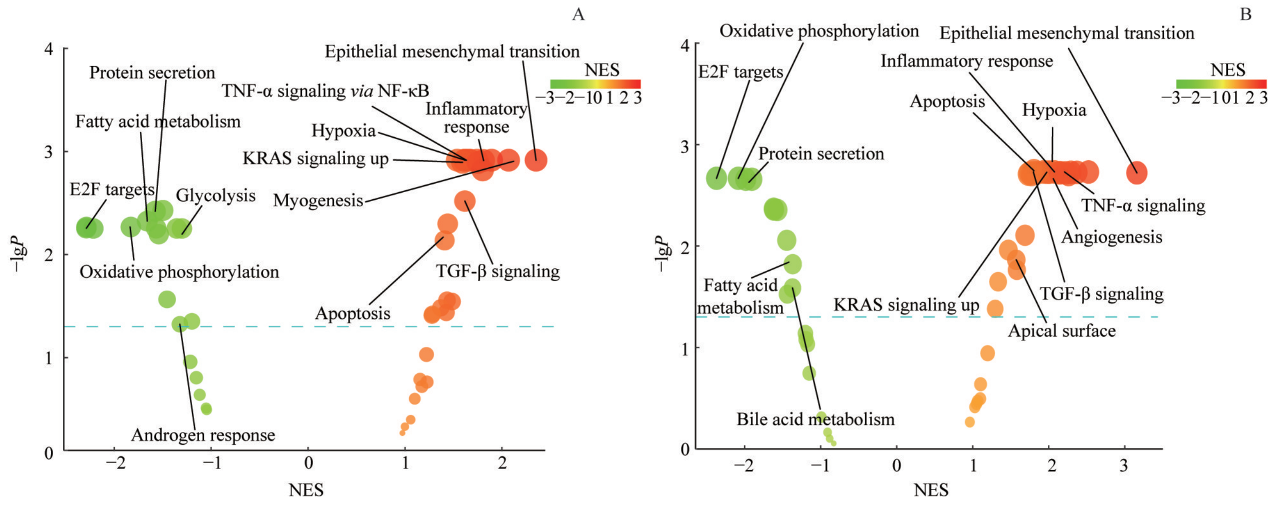
Note: A–D. ROC curves for predicting one-, two- and three-year survival in internal validation set (A), GSE39582 (B), GSE17536 (C) and GSE17537 (D). E–H. KM survival analyses of the two groups in internal validation set (E), GSE39582 (F), GSE17536 (G) and GSE17537 (H).

图4　CRC预后风险评分模型在内、外部独立验证集中的性能评估

Fig 4　Performance evaluation of CRC prognostic risk score model in internal/external validation sets

2.4　高、低风险组肿瘤特征基因集的相关通路富集分析

GSEA结果显示，在TCGA数据集（图5A）和GSE39582数据集（图5B）中，上皮细胞-间充质转化（epithelial-mesenchymal transition，EMT）生物过程、转化生长因子β（transforming growth factor-β，TGF-β）信号通路、Kirsten大鼠肉瘤病毒原癌基因同源产物（Kirsten rat sarcoma viral oncogene homolog，KRAS）信号通路、炎症反应等在高风险组中具有较高的NES，而蛋白分泌、氧化磷酸化等生物过程在低风险组中具有较高的NES；从而表明，高、低风险组间的肿瘤特征基因集相关信号通路或生物过程在2个数据集中的富集均具有显著差异。



Note: E2F—transcription factor E2F; TNF—tumor necrosis factor; NF-κB—nuclear factor-kappa B. Blue dashed line marks P=0.05.

图5　TCGA数据集(A)和GSE39582数据集(B)中高、低风险组肿瘤特征基因集相关信号通路或生物过程的NES

Fig 5　NES of cancer hallmark signaling pathway or biological process between the high risk group and low risk group in the TCGA dataset (A) and GSE39582 dataset (B)

2.5 预后风险评分模型的应用

通过KM生存曲线进行分别对2个数据集中高、低风险组患者进行分析，结果（图6）显示，与低风险组相比，高风险组患者的生存率均较低；继而说明，在免疫治疗数据集中风险评分的高低会影响患者最终的生存状态。因此，该CRC预后模型在接受免疫治疗的其他癌症患者中也具有较好的评估效果。

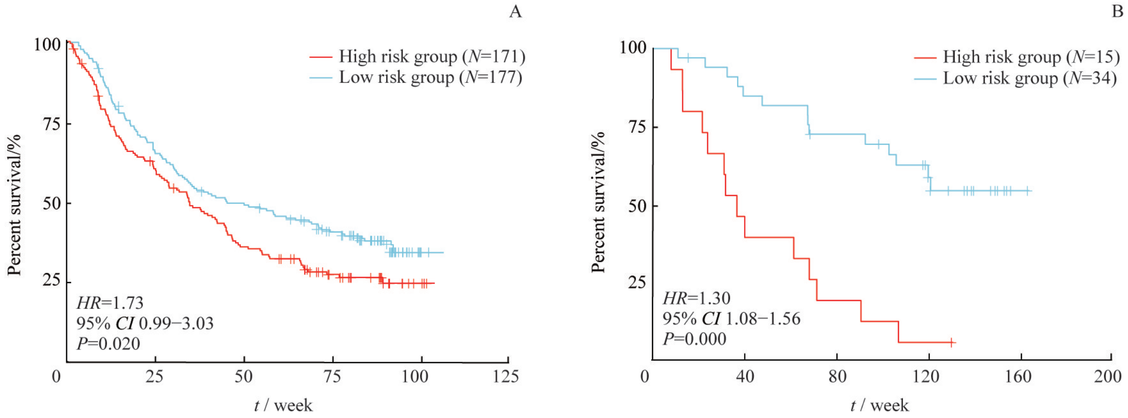


图6 Anti-PD-L1数据集(A)和anti-PD-1数据集(B)中高、低风险组患者的生存分析

Fig 6 KM survival analyses of the patients in the high and low risk group treated with anti-PD-L1 (A) and anti-PD-1 (B)

为了进一步分析患者的风险评分与免疫治疗效果间的相关性，排除临床信息中没有免疫治疗效果的患者后，分别对2个数据集中剩余的高、低风险组患者的免疫治疗效果进行分析，结果（表2）显示低风险组患者中免疫治疗效果为部分缓解（partial response，PR）和完全缓解（complete response，CR）的患者之和的占比较高，且经χ^2检验分析显示高、低风险组患者的上述免疫治疗效果间差异均具有统计学意义（anti-PD-L1数据集：P=0.007；anti-PD-1数据集：P=0.047）。

表2 2个数据集中高、低风险组患者的不同免疫治疗效果分析

Tab 2 Analysis of immunotherapy effect between the high risk group and low risk group in the two datasets

Immunotherapy effect	Anti-PD-L1 dataset		Anti-PD-1 dataset	
	High risk group (*N*=147)	Low risk group (*N*=151)	High risk group (*N*=15)	Low risk group (*N*=34)
PR+CR/*n* (%)	23 (15.65)	45 (29.80)	0 (0)	10 (29.41)
SD/*n* (%)	30 (20.41)	33 (21.85)	8 (53.33)	10 (29.41)
PD/*n* (%)	94 (63.95)	73 (48.34)	7 (46.67)	14 (41.18)

Note: SD—stable disease; PD—progressive disease.

3 讨论

由于缺乏适当的诊断方法，CRC患者往往在晚期才能得到确诊，因此针对该疾病的临床治疗及预后研究一直是学者们关注的焦点[30]。目前，部分研究已就CRC患者预后模型的构建进行报道，如Chen等[31]构建包含9个特征基因的CRC预后模型，Zuo等[32]构建6个特征基因的CRC预后模型，Pagès等[33]报道含3个特征基因的CRC预后模型，Zhao等[34]构建包含9个特征基因的CRC预后模型等。与之相比，本研究的CRC预后风险评分模型存在如下优势：①本研究的模型构建综合了来自不同数据库的转录组数据。相比Chen等[31]使用的3个GEO数据集［GSE32323(癌和癌旁样本各17例)、GSE74602(癌和癌旁样本各30例)、GSE113513（癌和癌旁样本各14例)］，本研究使用了更为充足的GEO样本量；相比Zuo等[32]基于TCGA数据集（647例CRC和51例正常结直肠样本）构建的模型，本研究不仅综合了GEO和TCGA的数据，还整合了正常组织样本数据库GTEx中数据使癌（471例）和癌旁（349例）样本数量更为均衡。②本研究的训练集AUC高于以往模型的训练集AUC。Chen等[31]得出的训练集5年AUC为0.741，Pagès等[32]分析的训练集3年AUC为0.711、5年AUC为0.683，Zhao等[34]获得的训练集3~6年的AUC分别为0.627、0.632、0.630、0.626，上述AUC值均低于本研究训练集AUC值（1年为0.780、2年为0.788、3年为0.754）。同时，本研究通过3个外部验证集对预后模型的预测效能进行检验（AUC均值>0.600），并证明该模型在其他独立数据集中也具有中等程度的准确性，而上述的前人研究并

未在多个数据集中加以验证。③本研究对预后模型在预测癌症患者免疫治疗效果方面的应用做了初步探索，这是其他CRC预后模型未涉足的领域。

构建可靠的预后模型离不开对建模参数的逐步筛选，本研究整合了8个GEO数据集，并获得了其共有的DEGs。在显著上调的共有DEGs中，*FOXQ1*可在CRC中促进肿瘤相关巨噬细胞的募集引发EMT[35]；显著下调的共有DEGs中，*ABCG2*可减轻氧化应激和炎症反应，在CRC中发挥潜在的保护作用[36]。GO功能分析显示，微阵列和RNA-seq数据集筛选的DEGs均富集在免疫细胞迁移、趋化因子介导的信号通路等。而上述富集的生物过程或信号通路与已有文献研究相一致。如Waugh等[37]发现，在CRC细胞中促炎趋化因子的表达较高；同时，Ackermann等[38]也证实，促炎趋化因子可促进CRC中嗜中性粒细胞的迁移，导致免疫细胞浸润增加，且这些过程对肿瘤微环境的改变具有十分重要的作用。

本研究通过DEGs筛选、单因素Cox回归分析、LASSO回归和多因素Cox回归分析，得到8个用于建模的特征基因，即*TIMP1*、*RGL2*、*SLC39A13*、*IER5L*、*HEYL*、*INHBB*、*DMKN*、*ELFN1*。研究[39]表明，敲低*TIMP1*可抑制结肠癌细胞的增殖、迁移和侵袭，并抑制CRC中的肿瘤发生和转移。Vigil等[40]发现，*RGL2*的异常过表达会促进胰腺癌的生长。文献[41]证实，与正常乳腺组织相比，乳腺癌组织中的*SLC39A13*异常高表达，且与不良的生存状态显著相关。Ruan等[42]发现，*IER5L*在CRC细胞中的表达较低。Liu等[43]报道，敲低*HEYL*可显著减少胃癌细胞的增殖和迁移。研究[44]发现，CRC组织中*INHBB*表达可发生异常改变。Morris等[45]对CRC患者的结肠黏膜组织进行测序，结果显示相比正常的结直肠组织，CRC患者的结肠黏膜组织中*DMKN* mRNA的非翻译区发生了异常选择性剪接和聚腺苷酸化修饰，并推测该基因可能是CRC的新型生物标志物。Lei等[46]报道*ELFN1*可在CRC细胞中发挥促增殖、抗凋亡和促迁移的功能。综合上述研究结果，我们发现本研究建立的模型可能与肿瘤的发生和迁移密切相关，且这些建模基因可能是肿瘤免疫微环境中发生异常改变的关键环节。

同时，本研究GSEA结果显示高风险组患者在EMT生物过程、TGF-β和KRAS信号通路中具有较高的富集评分。有文献报道，EMT通路与PD-L1之间具有双向调节的作用[47]，且该通路和PD-1/PD-L1联合靶向治疗也是一种新的免疫治疗策略[48]；另有研究发现，TGF-β[49]和KRAS信号通路[50]均参与了EMT的调节。综合上述结果我们推测，高风险组患者富集评分较高的肿瘤特征通路之间可能存在相互调节，而参与这些通路的基因或许可作为高风险组患者的生物标志物。

在预后模型的应用领域，Friedlander等[51]基于210例未接受过任何免疫治疗的黑色素瘤样本构建了包括*TIMP1*在内的15个特征基因的模型，在一定程度上可预测黑色素瘤患者的CTLA-4免疫治疗效果。本研究构建了8个特征基因的预后风险评分模型，根据风险评分将尿路上皮癌和黑色素瘤的免疫治疗患者分为高、低风险组，结果显示，高风险组患者的总体生存较差，而低风险组患者接受免疫治疗的效果更好，表现为免疫治疗效果为PR、CR的患者数量之和占比更多；上述结果表明，本研究构建的CRC预后风险评分模型也适用于其他癌症，风险评分不同的患者的免疫治疗效果间存在较大差异，而这种差异可能影响癌症患者最终的生存率。因此，未来或可将本研究的预后风险评分模型用于癌症患者的免疫效果的预测，以节约高风险患者的时间成本，优化临床治疗方案。

综上所述，本研究成功构建了包含8个特征基因的CRC预后模型，对CRC患者的预后具有一定的预测能力，可为挖掘CRC免疫治疗新靶点、优化肿瘤免疫治疗方案提供一定的参考。

参·考·文·献

[1] Testa U, Pelosi E, Castelli G. Colorectal cancer: genetic abnormalities, tumor progression, tumor heterogeneity, clonal evolution and tumor-initiating cells[J]. Med Sci (Basel), 2018, 6(2): 31.

[2] Bosch LJW, Carvalho B, Fijneman RJA, et al. Molecular tests for colorectal cancer screening[J]. Clin Color Cancer, 2011, 10(1): 8-23.

[3] Galon J, Costes A, Sanchez-Cabo F, et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome[J]. Science, 2006, 313(5795): 1960-1964.

[4] Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy[J]. Nat Rev Cancer, 2012, 12(4): 252-264.

[5] Golshani G, Zhang Y. Advances in immunotherapy for colorectal cancer: a review[J]. Therap Adv Gastroenterol, 2020, 13: 1756284820917527.

[6] Salem H, Attiya G, El-Fishawy N. Classification of human cancer diseases by gene expression profiles[J]. Appl Soft Comput, 2017, 50: 124-134.

[7] Xue J, Schmidt SV, Sander J, et al. Transcriptome-based network analysis reveals a spectrum model of human macrophage activation[J]. Immunity, 2014, 40(2): 274-288.

[8] Wang JD, Zhou HS, Tu XX, et al. Prediction of competing endogenous RNA coexpression network as prognostic markers in AML[J]. Aging (Albany NY), 2019, 11(10): 3333-3347.

[9] Lu Y, Beeghly-Fadiel A, Wu L, et al. A transcriptome-wide association study among 97, 898 women to identify candidate susceptibility genes for epithelial ovarian cancer risk[J]. Cancer Res, 2018, 78(18): 5419-5430.

[10] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene

expression and hybridization array data repository[J]. Nucleic Acids Res, 2002, 30(1): 207-210.

[11] Sabates-Bellver J, van der Flier LG, de Palo M, et al. Transcriptome profile of human colorectal adenomas[J]. Mol Cancer Res, 2007, 5(12): 1263-1275.

[12] Matsuyama T, Ishikawa T, Mogushi K, et al. MUC12 mRNA expression is an independent marker of prognosis in stage Ⅱ and stage Ⅲ colorectal cancer[J]. Int J Cancer, 2010, 127(10): 2292-2299.

[13] Skrzypczak M, Goryca K, Rubel T, et al. Modeling oncogenic signaling in colon tumors by multidirectional analyses of microarray data directed for maximization of analytical reliability[J]. PLoS One, 2010, 5(10): e13091.

[14] Uddin S, Ahmed M, Hussain A, et al. Genome-wide expression analysis of middle eastern colorectal cancer reveals FOXM1 as a novel target for cancer therapy[J]. Am J Pathol, 2011, 178(2): 537-547.

[15] Galamb O, Wichmann B, Sipos F, et al. Dysplasia-carcinoma transition specific transcripts in colonic biopsy samples[J]. PLoS One, 2012, 7(11): e48547.

[16] Tsukamoto S, Ishikawa T, Iida S, et al. Clinical significance of osteoprotegerin expression in human colorectal cancer[J]. Clin Cancer Res, 2011, 17(8): 2444-2450.

[17] de Sousa E Melo F, Colak S, Buikhuisen J, et al. Methylation of cancer-stem-cell-associated Wnt target genes predicts poor prognosis in colorectal cancer patients[J]. Cell Stem Cell, 2011, 9(5): 476-485.

[18] Marisa L, de Reyniès A, Duval A, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value[J]. PLoS Med, 2013, 10(5): e1001453.

[19] Smith JJ, Deane NG, Wu F, et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer[J]. Gastroenterology, 2010, 138(3): 958-968.

[20] Ritchie ME, 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.

[21] Kolde R, Laur S, Adler P, et al. Robust rank aggregation for gene list integration and meta-analysis[J]. Bioinformatics, 2012, 28(4): 573-580.

[22] Yu GC, Wang LG, Han YY, et al. clusterProfiler: an R package for comparing biological themes among gene clusters[J]. Omics, 2012, 16(5): 284-287.

[23] Walter W, Sánchez-Cabo F, Ricote M. GOplot: an R package for visually combining expression data with functional analysis[J]. Bioinformatics, 2015, 31(17): 2912-2914.

[24] Lin HQ, Zelterman D. Modeling survival data: extending the Cox model[J]. Technometrics, 2002, 44(1): 85-86.

[25] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models *via* coordinate descent[J]. J Stat Softw, 2010, 33(1): 1-22.

[26] Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker[J]. Biometrics, 2000, 56(2): 337-344.

[27] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles[J]. Proc Natl Acad Sci U S A, 2005, 102(43): 15545-15550.

[28] Mariathasan S, Turley SJ, Nickles D, et al. TGF- β attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells[J]. Nature, 2018, 554(7693): 544-548.

[29] Riaz N, Havel JJ, Makarov V, et al. Tumor and microenvironment evolution during immunotherapy with nivolumab[J]. Cell, 2017, 171(4): 934-949. e16.

[30] Siegel RL, Miller KD, Goding Sauer A, et al. Colorectal cancer statistics, 2020[J]. CA Cancer J Clin, 2020, 70(3): 145-164.

[31] Chen L, Lu D, Sun K, et al. Identification of biomarkers associated with diagnosis and prognosis of colorectal cancer patients based on integrated bioinformatics analysis[J]. Gene, 2019, 692: 119-125.

[32] Zuo S, Dai G, Ren X. Identification of a 6-gene signature predicting prognosis for colorectal cancer[J]. Cancer Cell Int, 2019, 19: 6.

[33] Pagès F, Mlecnik B, Marliot F, et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study[J]. Lancet, 2018, 391(10135): 2128-2139.

[34] Zhao X, Liu J, Liu S, et al. Construction and validation of an immune-related prognostic model based on TP53 status in colorectal cancer[J]. Cancers (Basel), 2019, 11(11): 1722.

[35] Wei C, Yang CG, Wang SY, et al. Crosstalk between cancer cells and tumor associated macrophages is required for mesenchymal circulating tumor cell-mediated colorectal cancer metastasis[J]. Mol Cancer, 2019, 18(1): 64.

[36] Nie S, Huang YQ, Shi MY, et al. Protective role of ABCG2 against oxidative stress in colorectal cancer and its potential underlying mechanism[J]. Oncol Rep, 2018, 40(4): 2137-2146.

[37] Waugh DJJ, Wilson C. The interleukin-8 pathway in cancer[J]. Clin Cancer Res, 2008, 14(21): 6735-6741.

[38] Ackermann A, Lafferton B, Plotz G, et al. Expression and secretion of the pro-inflammatory cytokine IL-8 is increased in colorectal cancer cells following the knockdown of non-erythroid spectrin α Ⅱ [J]. Int J Oncol, 2020, 56(6): 1551-1564.

[39] Song GH, Xu SF, Zhang H, et al. TIMP1 is a prognostic marker for the progression and metastasis of colon cancer through FAK-PI3K/AKT and MAPK pathway[J]. J Exp Clin Cancer Res, 2016, 35(1): 148.

[40] Vigil D, Martin TD, Williams F, et al. Aberrant overexpression of the Rgl2 Ral small GTPase-specific guanine nucleotide exchange factor promotes pancreatic cancer growth through Ral-dependent and Ral-independent mechanisms[J]. J Biol Chem, 2010, 285(45): 34729-34740.

[41] Ding BS, Lou WY, Xu L, et al. Analysis the prognostic values of solute carrier (SLC) family 39 genes in gastric cancer[J]. Am J Transl Res, 2019, 11(1): 486-498.

[42] Ruan W, Zhu S, Wang H, et al. IGFBP-rP1, a potential molecule associated with colon cancer differentiation[J]. Mol Cancer, 2010, 9: 281.

[43] Liu HN, Ni SJ, Wang HB, et al. Charactering tumor microenvironment reveals stromal-related transcription factors promote tumor carcinogenesis in gastric cancer[J]. Cancer Med, 2020, 9(14): 5247-5257.

[44] Badic B, Hatt M, Durand S, et al. Radiogenomics-based cancer prognosis in colorectal cancer[J]. Sci Rep, 2019, 9(1): 9743.

[45] Morris AR, Bos A, Diosdado B, et al. Alternative cleavage and polyadenylation during colorectal cancer development[J]. Clin Cancer Res, 2012, 18(19): 5256-5266.

[46] Lei R, Feng LC, Hong D. ELFN1-AS1 accelerates the proliferation and migration of colorectal cancer *via* regulation of miR-4644/TRIM44 axis[J]. Cancer Biomarkers, 2020, 27(4): 433-443.

[47] Jiang YY, Zhan HX. Communication between EMT and PD-L1 signaling: new insights into tumor immune evasion[J]. Cancer Lett, 2020, 468: 72-81.

[48] Malek R, Wang H, Taparra K, et al. Therapeutic targeting of epithelial plasticity programs: focus on the epithelial-mesenchymal transition[J]. Cells Tissues Organs, 2017, 203(2): 114-127.

[49] Zhu LY, Fu X, Chen X, et al. M2 macrophages induce EMT through the TGF-β/Smad2 signaling pathway[J]. Cell Biol Int, 2017, 41(9): 960-968.

[50] Yoon C, Till J, Cho SJ, et al. KRAS activation in gastric adenocarcinoma stimulates epithelial-to-mesenchymal transition to cancer stem-like cells and promotes metastasis[J]. Mol Cancer Res, 2019, 17(9): 1945-1957.

[51] Friedlander P, Wassmann K, Christenfeld AM, et al. Whole-blood RNA transcript-based models can predict clinical response in two large independent clinical studies of patients with advanced melanoma treated with the checkpoint inhibitor, tremelimumab[J]. J Immunother Cancer, 2017, 5(1): 67.

［收稿日期］ 2020-11-16 ［本文编辑］ 邢宇洋

特约创新团队介绍

创新团队名称

肠道免疫学

团队主要成员

苏冰(教授/博士)　**钟捷**(主任医师/博士)　**陆爱国**(主任医师/博士)
邹强(研究员/博士)　**李华兵**(研究员/博士)　**蒋玉辉**(研究员/博士)
叶菱秀(研究员/博士)　**王静**(研究员/博士)　**陈磊**(研究员/博士)
叶幼琼(研究员/博士)　**刘兆远**(副研究员/博士)

团队负责人介绍

苏冰 SU Bing

博士、教授、博士生导师

Ph.D, Professor, Doctoral Supervisor

ORCID ID: 0000-0003-0871-7666

苏冰（1963—），国家高层次人才，担任上海交通大学医学院上海市免疫学研究所所长、上海交通大学基础医学院免疫学与微生物学系主任，上海交通大学医学院-耶鲁大学免疫代谢研究院主任，上海交通大学王宽诚讲席教授，耶鲁大学医学院免疫生物学系客座教授。现任《现代免疫学》杂志主编、*JMCB*副主编、*Sci China Life Sci*编委等。苏冰教授长期致力于细胞内的信号转导研究，尤其是丝裂原激活的蛋白激酶（mitogen-activated protein kinase，MAPK）和哺乳动物雷帕霉素靶蛋白（mammalian target of rapamycin，mTOR）信号通路在免疫反应和血管功能/发育中介导的信号转导的生物学功能和意义方面，取得了多项突破性成果。自2012年回国以来，受到国家自然科学基金重大研究计划、重点项目、国际合作项目、面上项目等多项资助。

SU Bing (1963—), National High-Level Talent, director of Shanghai Institute of Immunology, Shanghai Jiao Tong University School of Medicine, director of the Department of Immunology and Microbiology of Shanghai Jiao Tong University College of Basic Medical Sciences, director of Shanghai Jiao Tong University School of Medicine-Yale University Institute for Immune Metabolism, KC Wong Chair Professor of Shanghai Jiao Tong University and Adjunct Professor of the Department of Immunobiology of Yale School of Medicine. Currently, he is editor-in-chief of *Current Immunology*, associate editor of *JMCB*, editorial board member of *Sci China Life Sci*, etc. Prof. SU's research focuses on the intracellular signal transduction pathways controlled by the mitogen-activated protein kinase (MAPK) and mammalian target of rapamycin (mTOR) and the roles of these intracellular signaling cascades in immune regulation and vascular function, making significant progress in this field. After returning to China in 2012, he has received a number of grants from National Natural Science Foundation of China, including major research programs, key projects, international cooperation projects, general projects, etc.

主要研究方向

苏冰教授领导的创新团队致力于MAPK/mTOR介导的（肠道）黏膜抗感染/肿瘤免疫研究，利用包括单细胞/空间多组学测序技术、先进组织成像、高维流式等技术，结合全新的遗传谱系示踪模型、小鼠肠道疾病模型及临床肠炎与肠癌样本，研究诸如固有淋巴样细胞、髓系细胞（巨噬、中性粒、肥大细胞等）和间充质基质细胞等特定细胞亚群在肠道疾病发生发展过程中的时空变化趋势，揭示其转录与代谢调控规律。一系列原创性研究为炎症性肠病和结直肠癌的治疗提供了新的治疗靶点及干预策略，研究成果相关论文在*Cell*、*Nature*、*Nat Genet*、*Nat Immunol*、*Immunity*、*Mol Cell*及*EMBO J*等杂志上发表100余篇，他引15 000余次。

Prof. SU's group has a keen interest in the biology of signal transduction mediated by the MAPK and mTOR pathways, particularly their regulation and function in mucosal/intestinal immunity against pathogens, tumor and other immunogens. We are employing cutting-edge technologies, including single cell and spatial multi-omic profiling, whole mount tissue imaging, Hi-parameter flow cytometry, combined with newly established lineage tracing mouse models, mouse intestinal disease models, and clinical biopsy of enteritis and colon cancer to investigate spatiotemporal distribution and the transcriptomic/metabolic regulation of certain cell types such as innate lymphocytes (ILCs), myeloid cells (macrophages, neutrophils, mast cells, etc.) and mesenchymal stromal cells in homeostatic and disease condition. A series of original scientific research provides novel therapeutic targets and intervention strategies for inflammatory bowel diseases (IBD) and colorectal cancer (CRC) treatment. Prof SU's group has published more than 100 SCI-indexed papers on *Cell*, *Nature*, *Nat Genet*, *Nat Immunol*, *Immunity*, *Mol Cell*, *EMBO J*, etc., with more than 15 000 citations.

近两年代表性成果

1) Wu, N, Sun, H, Zhao, X. et al. MAP3K2-regulated intestinal stromal cells define a distinct stem cell niche[J]. Nature, 2021. https://doi.org/10.1038/s41586-021-03283-y.

2) Hu ZL, Teng XL, Zhang T, et al. SENP3 senses oxidative stress to facilitate STING-dependent dendritic cell antitumor function[J]. Mol Cell, 2021: S1097-S2765(20)30945-X.

3) Yang XD, Li WG, Zhang SY, et al. PLK4 deubiquitination by Spata2-CYLD suppresses NEK7-mediated NLRP3 inflammasome activation at the centrosome[J]. EMBO J, 2020, 39(2): e102201.

4) Zhao Q, Zheng K, Ma CM, et al. PTPS facilitates compartmentalized LTBP1 S-nitrosylation and promotes tumor growth under hypoxia[J]. Mol Cell, 2020, 77(1): 95-107. e5.

5] Liu Z, Gu Y, Chakarov S, et al. Fate mapping *via* Ms4a3-expression history traces monocyte-derived cells[J]. Cell, 2019, 178(6): 1509-1525. e19.

6) Hu ZL, Qu GJ, Yu XY, et al. Acylglycerol kinase maintains metabolic state and immune responses of CD8$^+$ T cells[J]. Cell Metab, 2019, 30(2): 290-302. e5.