Contents lists available at ScienceDirect

Translational Oncology

journal homepage: www.elsevier.com/locate/tranon

Original Research

Identification of a 5-gene-based signature to predict prognosis and correlate immunomodulators for rectal cancer

Lin Yi [a,b,1], Ji Qiang [b,c,1], Peng Yichen [b], Yu Chunna [b], Zheng Yi [b,c], Kang Xun [b], Zheng Jianwei [a], Bai Rixing [a], Yan Wenmao [a], Wang Xiaomin [b,d,*], Li Parker [e], Li Wenbin [b,*]

[a] *Department of General Surgery, Beijing Tiantan Hospital, Capital Medical University, Beijing, China*
[b] *Department of Neuro-oncology, Cancer Center, China National Clinical Research Center for Neurological Diseases, Beijing Tiantan Hospital, Capital Medical University, Beijing, China*
[c] *Department of Epidemiology and Health Statistics, School of Public Health, Capital Medical University, Beijing, China*
[d] *Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Peking University Cancer Hospital & Institute, Beijing, China*
[e] *Clinical Medicine, Shanghai Jiao Tong University School of Medicine, Shanghai, China*

ARTICLE INFO

Keywords:
Rectal cancer
Prognosis
Bioinformatics
Differentially expressed genes
Biomarker

ABSTRACT

Background: Specific tumor markers have yet to be identified in rectal cancer. This study aims to identify a novel genetic signature in rectal cancer to provide clues for survival and immunotherapy.

Methods: DEGs were obtained from two GEO datasets of rectal cancer. By using data from TCGA and GSE133057, two cohorts of rectal cancer were applied to establish and evaluate the signature. A nomogram was constructed for training and validation. We integrated the risk-score with clinicopathological features and assessed its interplay with immune cells and molecules. Finally, our study performed functional annotations, gene-targeted miRNAs, and single-cell analysis.

Results: A total of 468 DEGs were identified, and a signature consisting of 5 genes (CLIC5, ENTPD8, PACSIN3, HGD, and GNG7) was selected to calculate the risk-score. The model exhibited high performance in time-dependent ROC and a nomogram. Further results showed that overall survival was significantly worse in the high-risk group. As an independent prognostic factor, the risk-score was associated with vascular invasion. There was a dramatic difference in nonregulatory $CD4^+$ and $CD8^+$ T cells between the high and low-risk groups, and the 5 genes were correlated with immune inhibitors. There was a considerable difference in autophagy, immune, cell cycle, infection, and apoptosis-associated terms and pathways in GO and KEGG. The functional states of differentiation, apoptosis, and quiescence were closely related to the 5-gene signature in single-cell analysis.

Conclusion: Our results suggest that the signature could serve as a novel prognostic biomarker in rectal cancer, which might benefit decision-making regarding immunotherapy.

Abbreviations

DEGs	differentially expressed genes
GEO	Gene Expression Omnibus
OS	overall survival
TCGA	The Cancer Genome Atlas
LASSO	least absolute shrinkage and selection operator
nCRT	neoadjuvant chemoradiotherapy
TME	total mesorectal excision
LARC	locally advanced rectal cancer
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
GSEA	Gene Set Enrichment Analysis
TMB	tumor mutation burden
MSI	microsatellite instability
ssGSEA	single-sample gene-set enrichment analysis
ROC	receiver operating characteristic
DCA	decision curve analysis

[*] Corresponding authors at: Department of Cancer Center, Beijing Tiantan Hospital, Capital Medical University, No.119 South Fourth Ring West Road, Fengtai District, Beijing 100070, China.

E-mail addresses: wangxiaomin@ihcams.ac.cn (W. Xiaomin), liwenbin@ccmu.edu.cn (L. Wenbin).

[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.tranon.2022.101529
Received 30 May 2022; Received in revised form 11 August 2022; Accepted 1 September 2022

Introduction

According to the GLOBOCAN 2018 updates, colorectal cancer (CRC) is the third most commonly diagnosed cancer and the second leading cause of cancer death worldwide [1,2]. In 2018, an estimated 1.85 million new cases were diagnosed, and 880,000 people died from this disease [1]. Rectal cancer accounts for more than 30% of CRCs in China and is associated with unfavorable clinical outcomes [3,4]. To date, the standard strategy for locally advanced rectal cancer (LARC) is neoadjuvant chemoradiotherapy (nCRT) followed by total mesorectal excision (TME) [4,5]. However, the evaluation of survival rates of rectal cancer is largely based on the TNM staging system, which has undergone changes, resulting in different editions, and the estimation of rates based on genetic markers has been limited due to the lack of convincing data.

Several studies have aimed to evaluate potential genetic prognostic factors as predictors of treatment response and disease outcome of LARC. In 2005, Ghadimi and his colleagues reported gene expression signatures from 30 LARC patients, in which a list of 54 differentially expressed genes (DEGs) between responders and nonresponders was identified and further used to generate expression profiles. This signature was able to successfully predict tumor response in 83% of patients ($P = 0.02$) [6]. In 2020, Ja Park et al. performed a gene expression study and reported a nine-gene signature (FGFR3, GNA11, H3F3A, IL12A, IL1R1, IL2RB, NKD1, SGK2, and SPRY2) to predict the response using biopsy samples from 156 LARC patients [7]. In 2022, Kim S et al. reported that CXCL12 levels were increased in LARC cells after nCRT, and that CXCL12 expression in the plasma membrane of LARC cells after nCRT correlated with a worse prognosis of LARC [8].

By collecting the increasing data regarding disease outcomes, immune factors and microRNAs (miRNAs) had been identified as potential prognostic factors for rectal cancer [9–12]. Accumulating evidence suggested that appropriate molecular predictors may be more crucial than clinicopathological features for understanding prognosis and making treatment decisions [11,13,14]. Nevertheless, unlike breast cancer, for which a number of prospective and retrospective studies have shown an association with prognosis using the 21-gene assay, no specific tumor marker has yet been identified for rectal cancer [15]. Thus, the lack of sufficiently informative biomarkers continues to hinder the molecular diagnosis and accurate prediction of prognosis for LARC.

Currently, published data on neoadjuvant immunotherapy in rectal cancer provides an innovative opportunity for improving LARC treatment. Cercek A et al. reported a prospective phase 2 study that utilized an anti-PD-1 monoclonal antibody in LARC patients with mismatch repair-deficient, and all 12 patients achieved a clinical complete response objectively [16]. Not long ago, Seo I et al. published a translational study of neoadjuvant chemoradiation therapy for LARC, while GSEA revealed six biological pathways significantly associated with dysregulated genes, including DNA replication, cell cycle, ribosome, base excision repair, mismatch repair (MMR) deficiency, and peroxisome [17]. Rizzo A et al. summarized several studies for effective biomarkers to predict immunotherapy responses in cancer, and also provided a further overview of the evidence that is currently available on effective biomarkers to predict immunotherapy responses in cancer, such as tumor mutation burden (TMB), microsatellite instability (MSI), MMR deficiency [18]. Thus, the objective of this study was to determine whether differentially expressed gene profiles obtained from rectal cancer and normal mucosa could offer insight into the prognosis and immunotherapy for LARC.

Materials and methods

Pre-procession of datasets and database

Two independent datasets of LARC were downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/geo/). GSE 15781, the platform of which was GPL2986, included 20 normal rectal tissues and 21 rectal cancer tissues. GSE 20842, which was performed on GPL4133, included 65 normal rectal mucosa and 65 rectal cancer tissues. In addition, the FPKM values of rectal cancer gene expression and related clinical phenotype were downloaded from the TCGA database (UCSC website: http://xena.ucsc.edu/) [19]. Moreover, GSE 133057 ($N = 33$) data from the GEO database (https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE133057) was applied for further prognosis validation.

Identification of differentially expressed genes

The R package "limma" was used to calculate statistical changes in expression levels between two datasets. The DEGs between normal tissues and rectal cancer tissues were extracted as described previously, and the Enhanced Volcano Plots were performed using the R package "EnhancedVolcano".The genes with standards of adjusted *P-value* <0.05 and $|\log2FC|\geq1$ were considered as DEGs. To evaluate the transcriptomic expression levels of two datasets, Heat Maps were constructed using the R package "pheatmap" across normal and cancer tissues. In addition to the heatmaps, the Venn diagram was used with the R package "Venn diagram" to determine the overlap of DEGs.

Construction of the prognostic model

For modeling, variables and characteristics (as shown in Table 1 and Fig. 4G) were abstracted from the TCGA data, and patients were stratified into high and low expression groups based on the median cut-off values of each gene expression. Then, Kaplan-Meier survival and Cox proportional hazard analyses were performed using the R package "survival" to determine overall survival. The R package "glmnet" was used for least absolute LASSO analyses, and the R package "survival" was used for multivariate Cox regression to obtain a risk classification score. The risk-score was calculated by using the R package "survival", and the mathematical model is as follows: Risk score = h0(t)*exp(b1X1 + b2X2 + … + bnXn) where n is the representative number of modeling genes; b and X are the correlation coefficient and expression level of model gene prediction, respectively; and h0(t) is derived from the "predict" function.

Analysis of clinicopathological features and survival

When it comes to a potential clinical significance, the interaction between 5 gene-based risk-score and characteristics of rectal cancer was investigated with the chi-squared test. Chi-squared analysis was performed using the "chisq.test" function in R software to calculate the association of risk subgroups with clinical data. Then, Kaplan-Meier survival and Cox regression were performed to investigate the prognostic value of integrated risk subgroups and clinicopathological features by using the R package "survival".

Establishment and validation of time-roc curve and predictive nomogram

Patients were assigned to the high-risk and low-risk group according to the risk-score calculated by the model, and the GSE133057 dataset was used as a validation cohort to evaluate the prognostic value of the 5-gene signature. A nomogram was constructed to evaluate a prognostic scoring system for survival prediction in rectal cancer patients. The R package"rms" was used to build the nomogram and the calibration chart. The calibration chart was used to validate the performance of the nomogram. The R package "survivalROC" was performed to draw the receiver operating characteristic (ROC) curve to evaluate the accuracy of the nomogram. Decision curve analysis (DCA) was then employed to evaluate the clinical performance of the nomogram by using the R package "ggdca".

Table 1
Association between the patient's clinicopathological features and risk scores in TCGA rectal cancer patients.

Clinicopathological Features		Risk Scores			
	No.	High (%)	Low (%)	χ^2	*P*-value
Gender				0.10717	0.743
Male	45	24 (61.5)	21 (55.3)		
Female	32	15 (38.5)	17 (44.7)		
Age				0.32765	0.567
≤60	23	10 (25.6)	13 (34.2)		
>60	54	29 (74.4)	25 (65.8)		
History of Polyps				0.010969	0.917
Absent	34	18 (54.5)	16 (59.3)		
Present	26	15 (45.5)	11 (40.7)		
No record[*]	17				
Tumor Depth				1.199	0.2735[†]
T1+T2	13	10	3		
T3+T4	61	29	33		
No record[*]	3				
Lymph-Node Metastasis				0	1
Absent	42	21 (53.8)	21 (55.3)		
Present	35	18 (46.2)	17 (44.7)		
Distant Metastasis				2.536	0.324[†]
M0	58	27 (69.2)	31 (81.6)		
M1	11	8 (20.5)	3 (7.9)		
Mx	8	4 (10.3)	4 (10.5)		
Stage				2.4232	0.489[†]
I	13	8 (21.6)	5 (13.9)		
II	28	11 (29.7)	16 (44.4)		
III	22	11 (29.7)	11 (30.6)		
IV	11	7 (18.9)	4 (11.1)		
No record[*]	3				
Lymphatic Invasion				2.4368	0.119
Absent	38	15 (44.1)	23 (65.7)		
Present	31	19 (55.9)	12 (34.3)		
No record[*]	8				
Vascular Invasion				4.2871	0.038
Absent	50	20 (60.6)	30 (85.7)		
Present	18	13 (39.4)	5 (14.3)		
No record[*]	9				
Neoadjuvant Treatment				0.40822	0.483
Responders	57	26 (74.3)	31 (83.8)		
Nonresponders	15	9 (25.7)	6 (16.2)		
No record[*]	5				
Microsatellite Test				0.88399	0.347
Microsatellite stable	63	34 (87.2)	29 (76.3)		
Microsatellite instability	13	5 (12.8)	9 (23.7)		
No record[*]	1				

Abbreviations:
[*] Data incomplete.
[†] Fisher's exact test.

Analysis of functional annotation and enrichment

By using the R package "clusterProfiler", the low- and high-risk groups were analyzed to estimate the discrepancy of the potential biological processes and functions. $P < 0.05$ was set as the cut-off value for both Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment. R-based package "GOplot" was used for visualization of the GO and KEGG analyses. Later, to estimate the significance of differences across the contrast of the HALLMARK gene set obtained from the MSigDB database, GSEA's analysis was carried out via "GSEA (v4.0.3)", and *P-value* <0.05 was considered to be enriched significantly.

Analysis of immune features with the 5-gene signature

To address the proportions of tumor-infiltrating immune cells, ssGSEA (single sample GSEA) analysis was performed by using the "gsva" package. Scores of each rectal cancer sample were calculated based on immune-related gene sets and estimated for relative proportions of tumor-infiltrating immune cells by CIBERSORT. To better understand the immune functions of these genes, the ssGSEA was performed to evaluate the enrichment degree of 547 immune-related genes. A heat map plot was generated to show the correlation between the 5-gene in our model and immune inhibitors.

Prediction of miRNAs and analysis of functional state

Targertscan is an online tool that allows the user to identify candidate targeted MicroRNAs (miRNAs) by gene name [20]. Then, the results of those predicted targets were intersected by the Venn diagram, and overlapping MicroRNAs indicated the common target of the input genes. Additionally, the co-expression of the 5 genes and MicroRNAs were analyzed by using the TCGA database. Kaplan-Meier analyses of TCGA datasets were performed to compare survival results of 5 genes and MicroRNAs. In addition, the CancerSEA tool provided analysis of genes' expression by single-cell and linked to functional states as previously described [21].

Statistical analysis

The chi-squared test (Fisher's exact test) was performed to analyze the correlations between genes' expression and clinicopathological variables. Overall survival was plotted and calculated using the Kaplan–Meier method, and differences between groups were compared by the log-rank test. The Multivariate Cox analysis was carried out on the statistically significant variables in the univariate Cox regression. *P-value* <0.05 (two-sided) were considered statistically significant. All analyses were performed by R (version 4.0.2) and visualized by R package "ggplot2".

Results

Differentially expressed genes between rectal cancer and normal tissues

The results of the analyses showed significant differences in unique DEGs when comparing rectal cancer and normal tissues with thresholds of |log2FC| >1.0 and adjusted *P-value* <0.05. In total, there were 1474 DEGs obtained from the GSE20842 dataset and 768 DEGs identified from the GSE15781 dataset, as displayed in the enhanced volcano map (Fig. 1A-B). To further elucidate the DEGs between cancer and normal tissues, heatmaps were grouped using hierarchical clustering. As illustrated in Fig. 1C-D, heatmap analyses revealed distinguishable trends of expression differences between rectal cancer and normal tissues. To gain insight into the number of shared genes, a Venn diagram was constructed to assess the expression levels. The summary data showed that there were 468 genes in common between the two datasets. The DEGs that were found to be significant in both datasets were candidates for further investigation into the prognostic evaluation of rectal cancer.

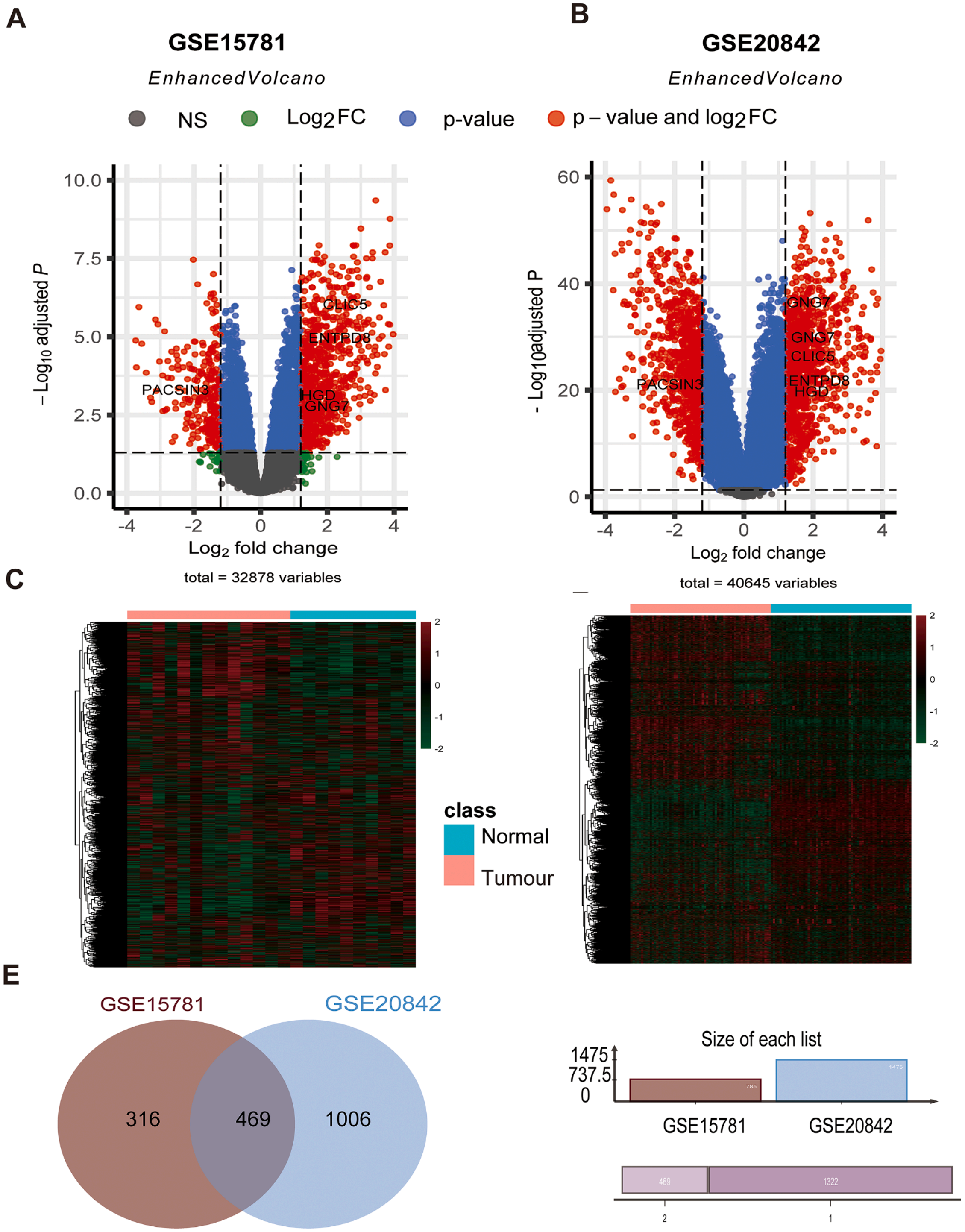


Fig. 1. Identification of differential expression genes (DEGs) between normal tissue and cancer tissue in two rectal cancer datasets. (A) Enhanced Volcano plot of DEGs for GSE15781 dataset. (B) Enhanced Volcano plot of DEGs for the GSE20842 dataset. (C) The heatmap plot of DEGs for the GSE15781 dataset. (D) The heatmap plot of DEGs in the GSE20842 dataset. The rows of the heatmap represented the z-score of log2 fold change (log2 FC) from high to low expression in the two datasets, and the columns represented the contrasts between normal and cancer tissues. In the column labels, green indicated low level of gene expression, whereas red indicated high level of expression. In the row labels, blue represented normal tissue, and red represented tumors. (E) The Venn diagram showed the number of DEGs and common DEGs identified by two profiling datasets. Blue represented the significant DEGs from the GSE20842 dataset, while red represented the significant DEGs from the GSE15781 dataset. (DEGs according to the value of $P < 0.05$ and $|\log2FC| < 1$; Red, high expression; green, low expression.).

Construction of prognostic gene signatures for rectal cancer

To investigate potential factors predictive of outcome in rectal cancer, survival data were obtained from the publicly available TCGA dataset to construct a training set. Univariable Cox regression was performed to assess the association between 468 DEGs and survival among rectal cancer patients. Then, 14 DEGs were identified by the survival analysis using the significance criterion of *P-value* < 0.05, and Fig. 2A visualized the results. The training set was used to construct a prognostic model for the survival of rectal cancer patients using least absolute shrinkage and selection operator (LASSO) logistic regression (Fig. 2B). Next, the 7 selected survival-related DEGs were further investigated by

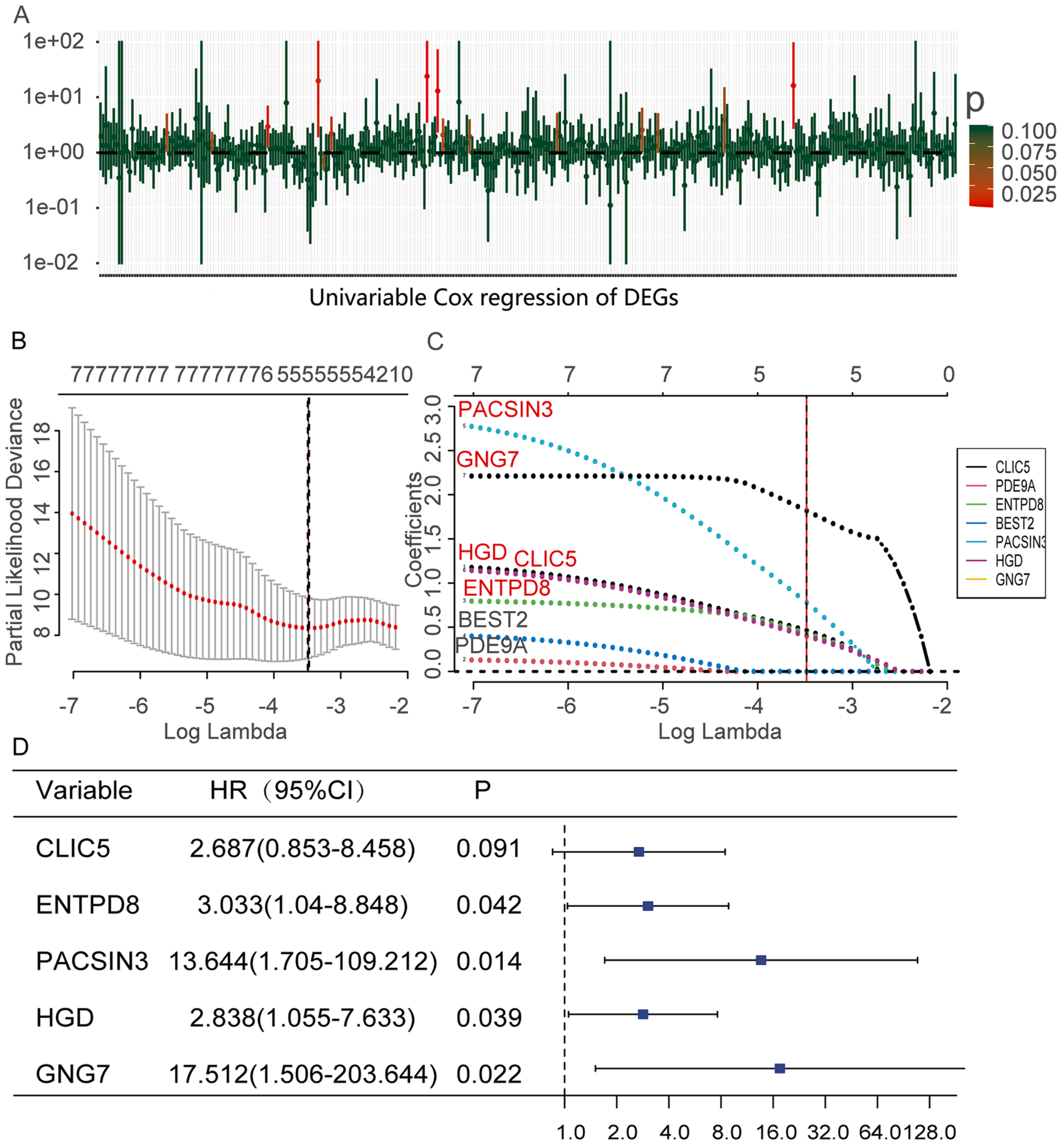


Fig. 2. Construction of prognostic genes signature for rectal cancer from DEGs. (A) Univariable Cox regression to evaluate the association between 468 DEGs and survival of rectal cancer patients. The x-axis indicates the value of *P* while the y-axis indicates the 95% Confidence Interval (CI). (B) Least absolute shrinkage and selection operator (LASSO) coefficient profiles (y-axis) of the genes. (C) The optimal penalization coefficient (λ) via 5-fold cross-validation is based on partial likelihood deviance. (D) Multivariate Cox regression to establish the best prognostic genes signature for rectal cancer patients. (Green, value of $P \geq 0.05$; Red, value of $P < 0.05$.).

comparing with theλ, and the hyperparameter L1 was optimized by using 5-fold cross-validation with the minimal partial likelihood deviance (Fig. 2B-C). Then, multivariate Cox regression was used to evaluate the risk-score of the 5 genes for the overall survival of rectal cancer patients (Fig. 2D). Cox regression was implemented under the proportional hazards (PH) assumption, and the risk-score was obtained for each patient using the product of the gene expression levels and estimated coefficients. As a result, a prognostic signature for rectal cancer was established, consisting of 5 genes (CLIC5, ENTPD8, PACSIN3, HGD, and GNG7). The formula used to calculate the risk-score of the 5-gene signature was as follows: CLIC5*0.99 + ENTPD8*1.11 + PACSIN3*2.61 + HGD*1.04 + GNG7*2.86.

Validating the risk-score in the time-ROC curves and survival analysis

To verify the prognostic value of the risk-score, the rectal cancer patients were divided into low-risk and high-risk groups according to the median risk-score across all rectal cancer patients (Fig. 3A-B). Thereafter, a time-dependent receiver operating characteristic (ROC) curve was generated to evaluate whether this risk-score had high predictive accuracy for prognosis, with an AUC of 0.926 at 3 years, an AUC of 0.881 at 4 years, and an AUC of 0.948 at 5 years for TCGA (Fig. 3C-D). The AUCs to predict survival for the GSE133057 dataset, as an independent validation shown in Fig. 3D, were 0.812 (3 years), 0.835 (4 years), and 0.828 (5 years). The model had a concordance index of 0.66 (95% CI 0.57 to 0.71). Finally, Kaplan–Meier survival curve analysis was used to assess the effectiveness of the 5 genes for the prognostic stratification of rectal cancer patients. In the TCGA data analysis, the high-risk group

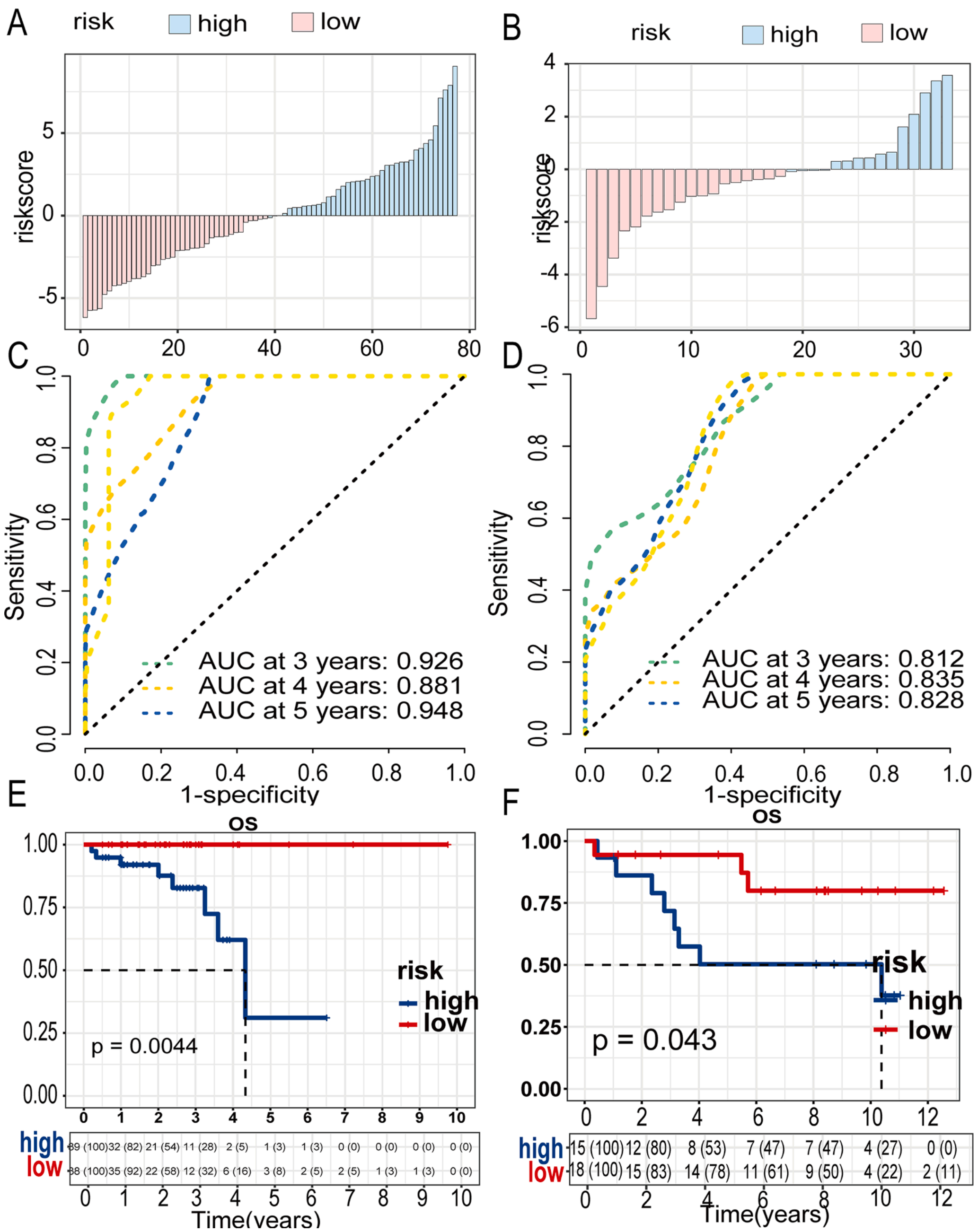


Fig. 3. Prognostic performance of the 5-gene signature model in TCGA and GSE133057 cohorts. (A-B) The distribution and median value of the risk-score in the TCGA and GSE133057 cohorts. (C-D) The ROC curves derived from models indicate the prognostic accuracy of the risk-score for the overall survival in TCGA and GSE133057 rectal cancer cohorts. (E-F) The Kaplan-Meier curves showed the overall survival in patients with rectal cancer according to the low and high-risk groups in TCGA and GSE133057 cohorts.

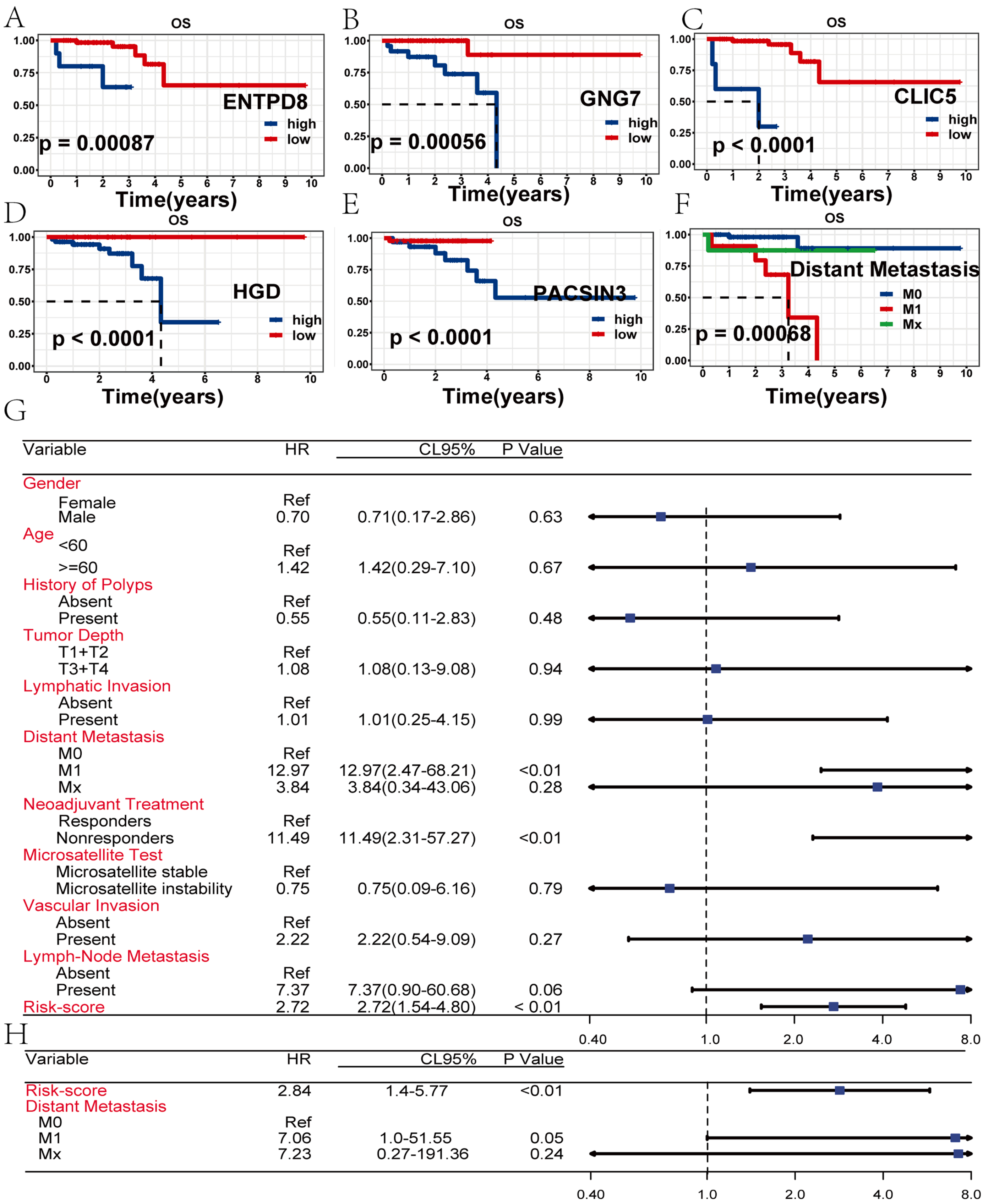
showed significantly unfavorable OS (P = 0.0044, log-rank = 6.328) compared with the low-risk group. A similar trend was observed in the GSE15781 dataset, where the high-risk group showed significantly poor OS (Fig. 3E-F).

The prognostic signature is correlated with clinicopathological features and survival

To gain insight into the association between risk-score and clinical variables, clinicopathological parameters were compared between the high- and low-risk groups. Clinical data downloaded from TCGA are summarized and analyzed in Table 1. Firstly, Kaplan–Meier analysis of TCGA datasets showed that rectal cancer tumors with high expression of the 5 genes showed worse outcomes (Fig. 4A-E). Then Kaplan–Meier curve was shown in Fig. 4F according to Distant Metastasis. A comparison suggested that prognostic values for CLIC5, ENTPD8, PACSIN3 HGD, and GNG7 were superior to Distant Metastasis (Fig. 4A-F). Secondly, the risk-score was associated with vascular invasion (P = 0.038). Associations between the risk-score and other clinicopathological features, such as stage and lymphatic invasion, failed to reach statistical significance in the TCGA dataset (Table 1). Thirdly, the results of Cox regression showed that after multivariable adjustments for clinicopathological factors, the risk-score remained significantly associated with patient OS (Fig. 4G). Our results also confirmed that the risk-score was an independent prognostic predictor of longer OS for rectal cancer patients (HR = 2.84; 95% CI, 1.4 to 5.77; P = 0.001, Fig. 4H). Distant metastasis, neoadjuvant treatment, and risk-score all had independent prognostic values in the multivariate analysis(Fig. 4G). Other clinicopathological parameters had no prognostic value in multivariate analyses.

Establishment of a predictive nomogram

To evaluate whether the 5-gene model was useful for survival



G

Variable	HR	CL95%	P Value
Gender			
Female	Ref		
Male	0.70	0.71(0.17-2.86)	0.63
Age			
<60	Ref		
>=60	1.42	1.42(0.29-7.10)	0.67
History of Polyps			
Absent	Ref		
Present	0.55	0.55(0.11-2.83)	0.48
Tumor Depth			
T1+T2	Ref		
T3+T4	1.08	1.08(0.13-9.08)	0.94
Lymphatic Invasion			
Absent	Ref		
Present	1.01	1.01(0.25-4.15)	0.99
Distant Metastasis			
M0	Ref		
M1	12.97	12.97(2.47-68.21)	<0.01
Mx	3.84	3.84(0.34-43.06)	0.28
Neoadjuvant Treatment			
Responders	Ref		
Nonresponders	11.49	11.49(2.31-57.27)	<0.01
Microsatellite Test			
Microsatellite stable	Ref		
Microsatellite instability	0.75	0.75(0.09-6.16)	0.79
Vascular Invasion			
Absent	Ref		
Present	2.22	2.22(0.54-9.09)	0.27
Lymph-Node Metastasis			
Absent	Ref		
Present	7.37	7.37(0.90-60.68)	0.06
Risk-score	2.72	2.72(1.54-4.80)	< 0.01

H

Variable	HR	CL95%	P Value
Risk-score	2.84	1.4-5.77	<0.01
Distant Metastasis			
M0	Ref		
M1	7.06	1.0-51.55	0.05
Mx	7.23	0.27-191.36	0.24

Fig. 4. The prognostic estimations of clinicopathological features and the risk-score among TCGA rectal cancer cohort. (A-E) The Kaplan–Meier analyses of single gene expression from the 5-Gene-Based Signature in TCGA rectal cancer cohort. (F) The prognostic value of a representative clinical variable (Distant Metastasis) in TCGA rectal cancer cohort. (G) The multivariate Cox analyses according to overall survival in the TCGA rectal cancer cohort. (H) The independent prognostic predictor of overall survival for TCGA rectal cancer patients.

prediction, a nomogram was developed to identify low- and high-risk patients based on the 3-year, 4-year, and 5-year OS rates. The nomogram demonstrated good discrimination of 5-year OS among patients with different clinical and pathologic parameters (Fig. 5A). Moreover, decision curve analysis (DCA) revealed that the model relative to the nomogram was associated with benefit gains (Fig. 5B). The calibration plot showed that the predicted power of both the training and validation set was near the ideal curve (Fig. 5C-D). Nonetheless, calculating the scores can be cumbersome, and the survival rate of patients at any time cannot be calculated conveniently. For this reason, the online predicted tool was made to provide access to the dynamic estimate of OS, and the final web-based nomogram is available online at https://xqccc.shinyapps.io/DynNoma/.

Functional enrichment analysis between the high and low-risk groups

To investigate the molecular function and signaling pathways between the high- and low-risk groups, further enrichment analyses were performed using GO terms and KEGG pathway annotations. According to GO analysis, the top 30 terms were visualized in Fig. 6A, and the most significant enrichments were autophagy (GO:0006914), vacuolar membrane (GO:0005774), and ubiquitin-like protein transferase activity (GO:0019787), which contained 227, 187, and 176 genes, respectively(Supplemental Table I). The top 30 enriched KEGG pathways are visualized in Fig. 6B, while the results of KEGG analysis indicated that the main enrichments were cell cycle (hsa04110), measles (hsa05162), apoptosis (hsa04210), p53 signaling pathway (hsa04115), and

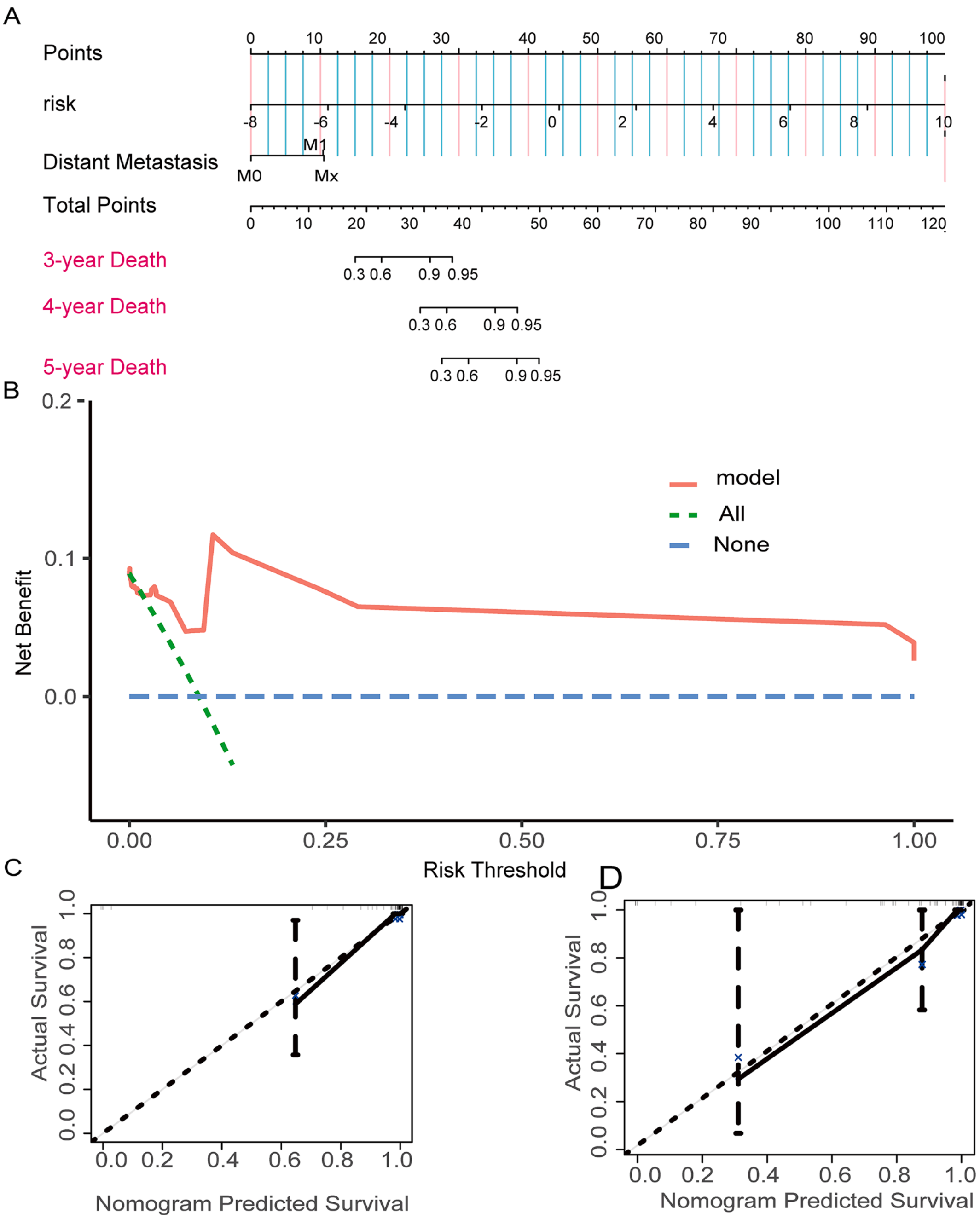


Fig. 5. Nomogram, calibration plots, and decision curves for the prediction of survival for patients with rectal cancer. (A) Nomogram for the prediction of survival possibility at 3, 4 and 5 years. (B) DCA for assessing the clinical utility of the nomogram. The x-axis indicated the percentage of threshold probability while the y-axis indicates the net benefit. (C, D) Calibration plots for predicting survival possibility at 3, 4 and 5 years, Diagonal line: ideal model, vertical bars: 95% confidence interval. (AUC: area under the receiver operating characteristic curve) .

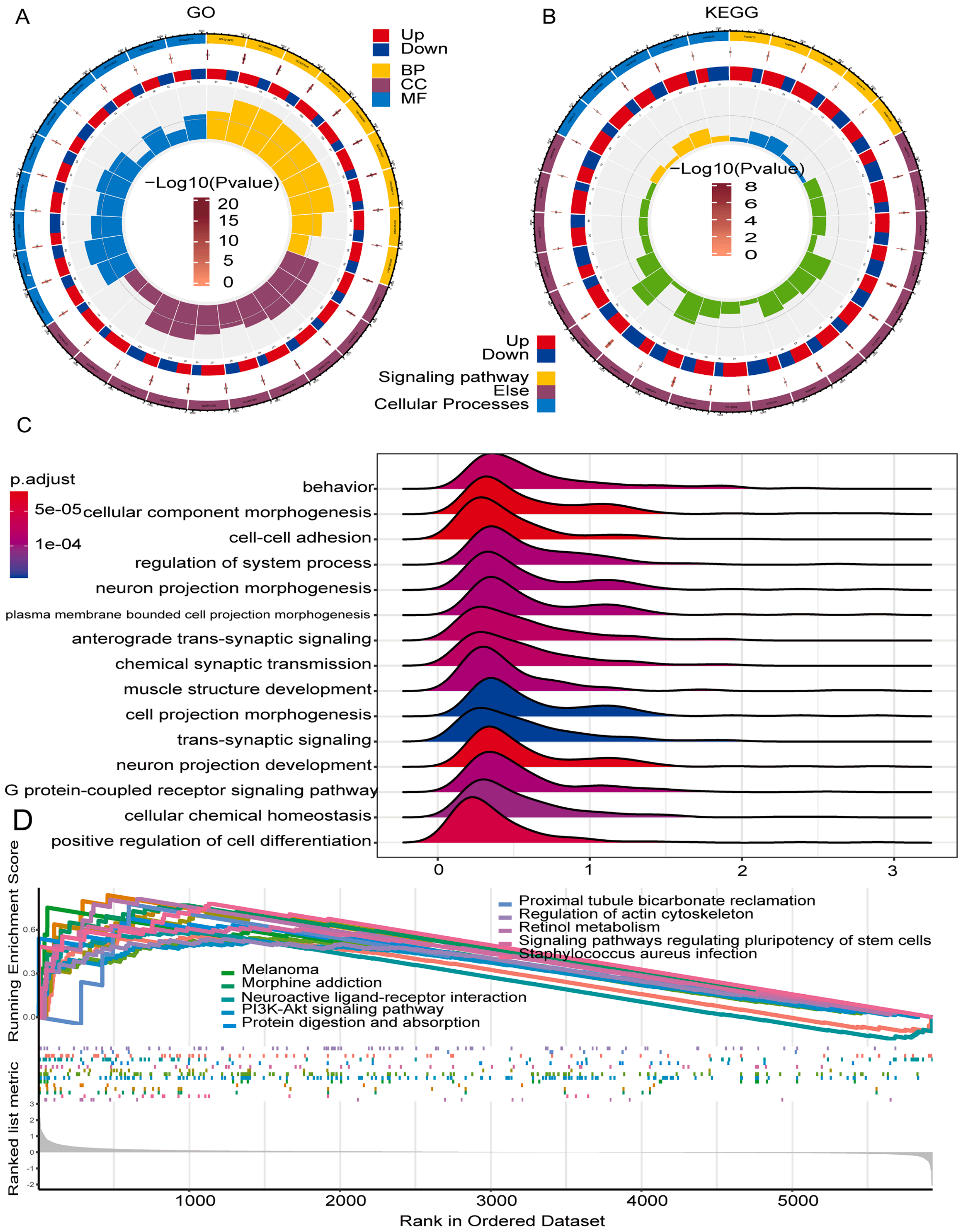


Fig. 6. Results of the enrichment analyses in the TCGA cohort. (A) The significant GO term analyses of differentially expressed genes between the high- and low-risk groups. (Red indicated up-regulation; Blue indicated down-regulation; Yellow indicated biological process class; blue indicated molecular function class; green indicated cellular component class.) (B) The significant KEGG term analyses of differentially expressed genes between the high- and low-risk groups. (Red indicated up-regulation; Blue indicated down-regulation; Yellow indicated Signaling Pathway class; Blue label belonged to cellular process class. Purple label belonged to else class). (C) The Ridge plot by GSEA between the high- and low-risk groups showed the top 15 significant GO terms. (D) The Enrichment plot by GSEA between the high- and low-risk groups showed the top 15 significant KEGG terms (Permutation tests $P < 0.05$, FDR < 0.25).

nucleotide excision repair (hsa03420). These results also indicated that infection (by human cytomegalovirus, Escherichia coli, or Epstein–Barr virus) likely played an important role in rectal cancer, which contained 95, 90, and 89 genes, respectively(Supplemental Table II). As shown in Fig. 6C-D, the Ridge plot and Enrichment plot showed the top 15 enrichments of GSEA, and these results confirmed that genes are predominantly involved in the cell cycle, autophagy, apoptosis, and immunity pathways between the high- and low-risk groups in rectal cancer.

Efficacy of the model with signature immunotherapeutic relevant genes

To assess the appropriateness of the 5 genes as clinically accepted biomarkers for immunotherapy, correlation analyses were carried out for microsatellite instability (MSI), tumor mutation burden (TMB), and the tumor microenvironment (TME) between the high- and low-risk groups. As shown, no significant differences were found in terms of microsatellite instability between the high- and low-risk groups (Fig. 7A). Nevertheless, the low-risk group had a higher proportion of MSI scores. In terms of the TMB, there were no significant differences between the two groups (Fig. 7B). Correlation analysis was evaluated between the 5-gene signature and immune cell infiltration. The tumor immune microenvironment consists of massive immune cell subsets surrounding cancer cells, including B cells, CD4$^+$ T cells, CD8$^+$ T cells, neutrophils, macrophages, and dendritic cells. Notably, the high- and low-risk groups showed significant differences in nonregulatory CD4$^+$ T cells and CD8$^+$ T cells (Fig. 7C-D). Myeloid dendritic cells were found to be almost significantly different between the two groups (Fig. 7C-D). More importantly, the Heat map plot showed that genes in our model were correlated with immune inhibitors, especially CLIC5 and GNG7 (Fig. 7E).

The prognostic signature correlated with differentiation, apoptosis, and quiescence

To understand how the potential molecular function of the 5-gene signature would impact the survival of rectal cancer patients, TargetScan was used to predict the targets of the 5 genes. As expected, it provided 5 sets of miRNAs targeted to 5 genes, and these were further used to evaluate the consensus prediction. As shown in the Venn diagram, the overlap of the miRNAs revealed one common microRNA (Fig. 8A). The TCGA rectal cancer dataset was stratified into high versus low hsa-miR-6887-expressing groups, and then, upon Kaplan–Meier survival analysis, high hsa-miR-6887-expressing tumors showed a tendency toward an improved outcome for patients, but this did not reach statistical significance (Fig. 8B). Thus, Analyses of TCGA cohorts showed that the expression of GNG7 was correlated with hsa-miR-6887 (Fig. 8C). The CancerSEA tool was used to identify genes correlated with functional state, in addition, differentiation, apoptosis, and quiescence were significantly related to the 5-gene signature in the single-cell dataset GSE81861 (Fig. 8D-F).

Discussion

In this study, we identified a 5-gene-based prognostic model by comparing normal tissues with rectal cancer and further validated the predictive power of the model in two independent rectal cancer datasets. Our findings indicated that OS rates increased with the risk-score in the TCGA and GSE133057 rectal cancer datasets. Strikingly, both the high- and low-risk groups showed significant association with the TME, especially non-regulatory CD4$^+$ T cells and CD8$^+$T cells. Moreover, the data showed that the risk-score was significantly associated with vascular invasion (Table 1). Based on these results, we propose the use of the 5-gene-based classification model as a novel molecular-based prognostic tool to evaluate the survival of rectal cancer patients. The model provides a starting point for further research into the role of genes in the development of rectal cancer.

Rectal cancer is a tumor with a relatively high prevalence but without convincing prognostic and predictive molecular markers. Biomarkers have been shown to classify patients with rectal cancer into specific subtypes, some of which may benefit from tailored therapy. Several important factors have been linked to rectal cancer survival, and these molecules are involved in key processes of cancer development, including cell migration and invasion [22–26]. Although these factors were considered to be significant in colorectal cancer, the findings were not verified in rectal cancer via validation in an independent dataset. Our results were validated in a separate cohort of patients with rectal cancer. These data supported the findings that the 5-gene-based signature served well as a predictor of survival in rectal cancer patients.

As our study showed, there is a strong correlation between the 5-gene-based tumor markers and clinical outcomes of rectal cancer. The molecular function of CLIC5, ENTPD8, PACSIN3, HGD, and GNG7 remains unclear, partly due to the lack of original research in rectal cancer. Chloride intracellular channel 5 (CLIC5) belongs to the family of chloride (Cl−) channels which are responsible for encoding chloride intracellular channel (CLIC) proteins. To date, there are six known members in the CLIC family (CLIC1-6), and increasing evidence supports their role in tumor biology, especially gastrointestinal cancer [27]. In a previous study, CLIC1 was found to be upregulated in colorectal cancer (CRC), which was associated with poor prognosis in CRC patients [28]. CLIC4 was also found to be overexpressed in CRC, and its upregulation was correlated with an unfavorable 5-year prognosis [29]. To date, there are no reports of an association between CLIC5 and rectal cancer. Ectonucleoside triphosphate diphosphohydrolase 8 (ENTPD8) is a member of the ectonucleoside triphosphate diphosphohydrolases (E-NTPDase) family, plays an essential role in ATP metabolism, and is mainly expressed in the intestine [30]. Although ENTPD8 is still poorly understood, there is evidence that it plays a crucial role in pancreatic cancer and exhibits metabolic activity toward gene-metabolite networks [31]. PACSIN3 has been identified as an intracellular adapter protein that regulates endocytosis, vesicle transport, membrane internalization, and actin reorganization. As reported, PACSIN3 is one of the mobility-related genes that is downregulated in ING5-overexpressing SGC-7901 gastric cancer cells. In another study, PACSIN3 was also found to be decreased in prostate cancer. To our knowledge, there is no report on PACSIN3 in rectal cancer. The HGD gene encodes one of the enzymes called homogentisate 1,2 dioxygenase, which is required for the catabolism of the amino acids tyrosine and phenylalanine, and is generally active in the kidneys and liver to catalyze oxidation–reduction reactions. The current study showed that high HGD mRNA expression ($\geq$3-fold) was associated with poorer survival, histological grade, advanced stage, and metastasis in patients with cholangiocarcinoma [32]. Previous results have shown that HGD is a potential key factor in the regulatory mechanism of BRAFV600E-mediated PTC. However, there was no significant discrepancy in overall survival [33]. G protein γ subunit 7 (GNG7), a component of the large G γ family, was first identified as a downregulated differentially expressed gene in pancreatic cancer [34] and then found in cancer of the gastrointestinal tract (including esophageal, gastric, and colorectal cancer) [35]. In a previous study, GNG7 was shown to act as a potential tumor suppressor both *in vitro* and *in vivo* [36]. A similar study also demonstrated that it was a tumor suppressor gene in clear cell renal cell carcinoma and lung adenocarcinoma [37]. Taken together, our research is the first work to evaluate the survival value of a 5-gene-based tumor marker in rectal cancer and might may help improve molecular prognosis [11].

Collectively, several studies have shown the clinical importance of immune infiltrates in colorectal cancer. Galon and his colleagues examined tumor-infiltrating lymphocytes (TILs) in approximately 400 colorectal cancer specimens and found that CD8$^+$ and CD45RO$^+$ T cells in the tumor were superior predictors to the histopathological staging methods [12]. Previous studies have also shown that CD4$^+$ and CD8$^+$T cells were promising survival predictors in colorectal cancer patients. A

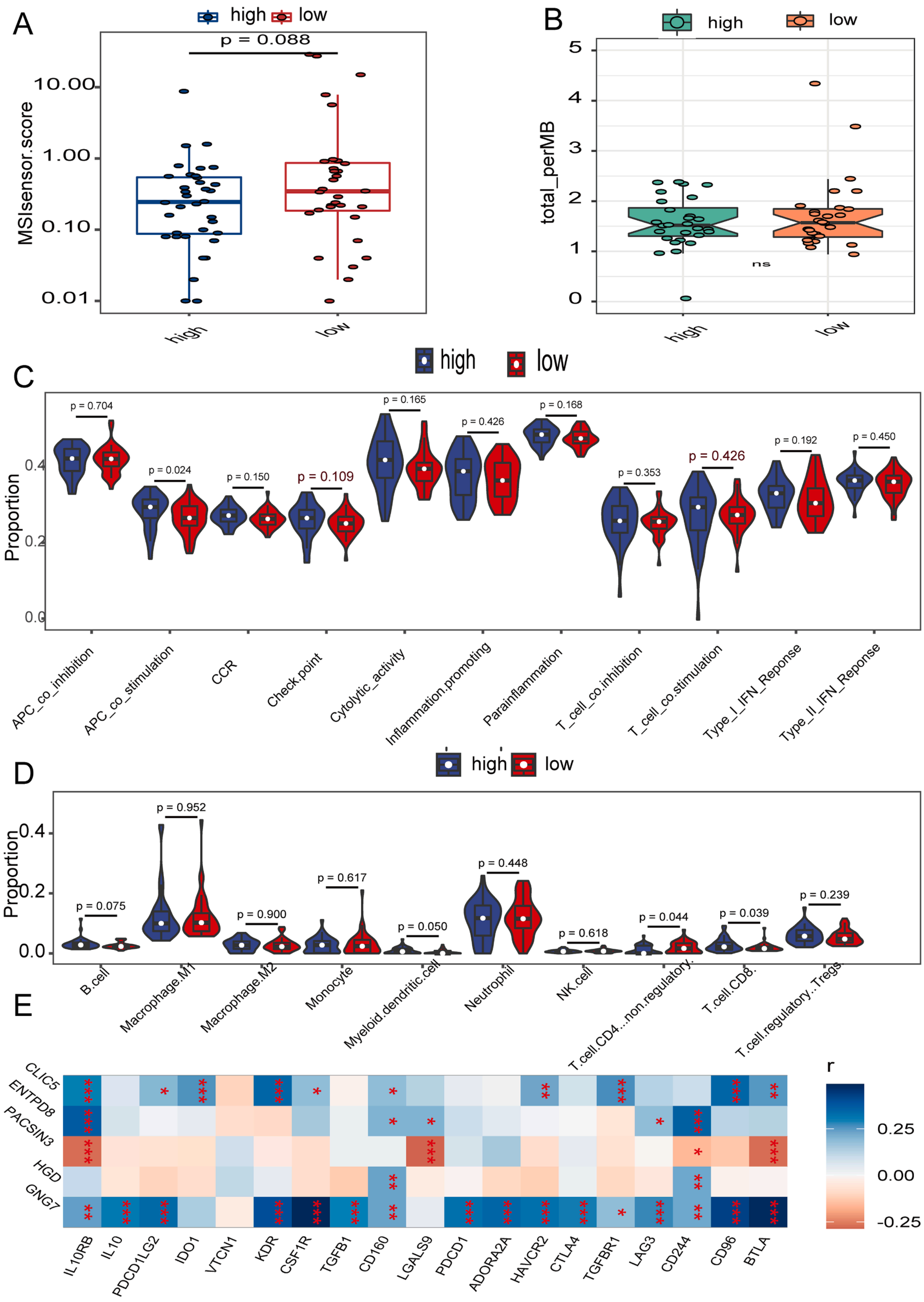


Fig. 7. Efficacy of the model with signature immunotherapeutic relevant genes. (A) The MIS scores in the boxplot between different risk groups in the TCGA rectal cancer cohort. (B) Differential TMB levels in the boxplot between high- and low-risk groups among TCGA rectal cancer cohort. (C) The ssGSEA scores of the 10 immune cells in the Violin Plot between high- and low-risk groups among TCGA rectal cancer cohort. (D) Comparison of the 11 immune-related functions in the Violin Plot between high- and low-risk groups among TCGA rectal cancer cohort. (E) Correlation of the 5-gene expression with immune checkpoint molecules among TCGA rectal cancer cohort. (ns, not significant; ***P* < 0.01; ****P* < 0.001.) .

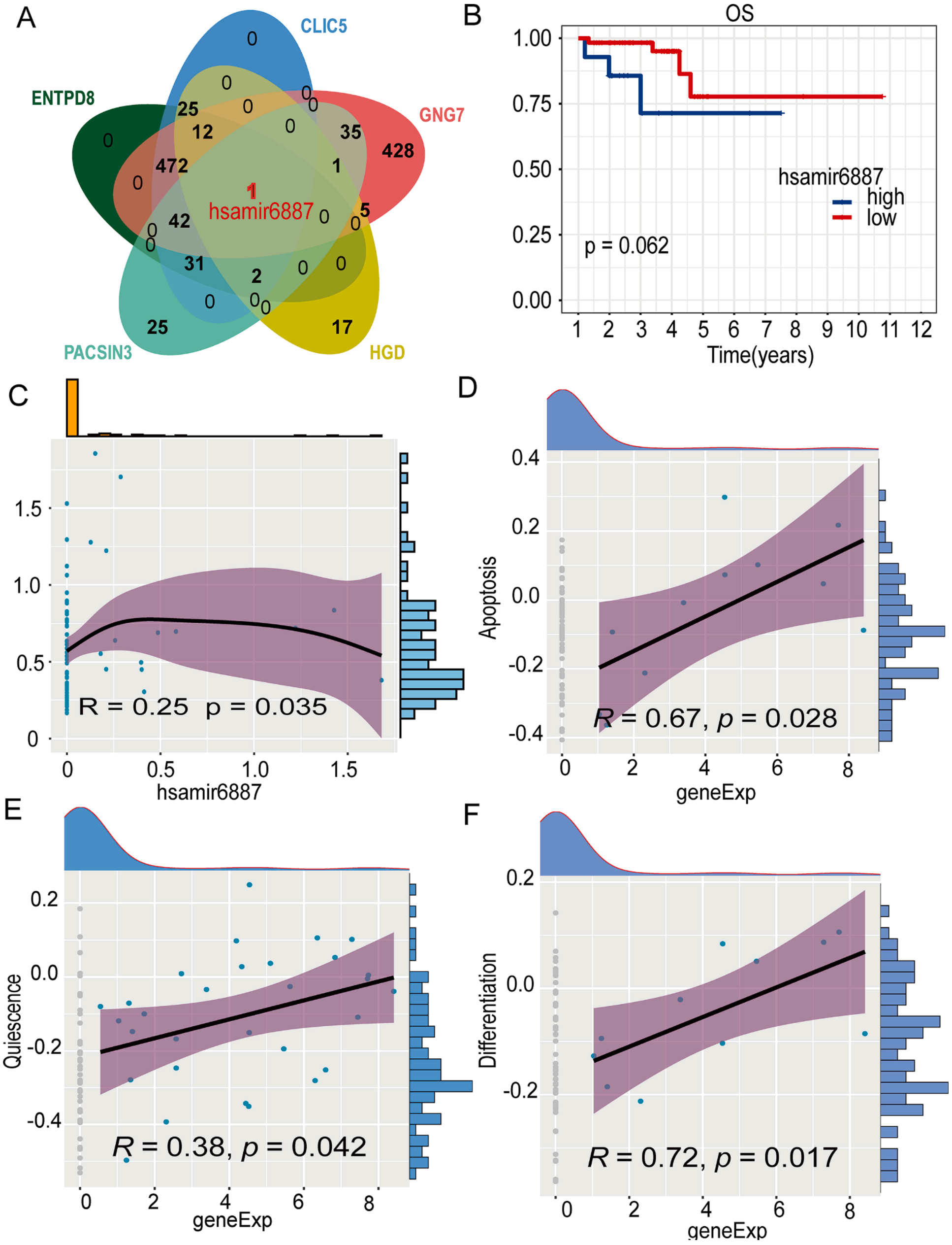


Fig. 8. Analyses of the function of the 5-gene signature. (A) Venn diagram to show the overlap of the miRNAs targeted to 5 genes predicted by TargetScan. (B) Kaplan-Meier survival analyses of high and low hsa-miR-6887-expressing groups. (C) The expression of GNG7 correlated with hsa-miR-6887 in TCGA rectal cancer cohort. (D-F) Visualization of correlations between the 5-gene signature and functional states (differentiation, apoptosis, and quiescence) originated from CancerSEA.

significant correlation was observed between the density of CD8$^+$T cells in the peritumoral region and a longer disease-free interval ($P = 0.009$), and Kaplan–Meier analysis later suggested that the percentage of CD8$^+$ T cells may have clinical application in stratifying patients' risk of recurrence ($P = 0.006$) [38]. Yasuda K et al. reported that tumor-infiltrating lymphocytes (TILs), especially density CD4$^+$ T cells and CD8$^+$ TILs, were strongly associated with tumor treatment response of rectal cancer after neoadjuvant chemoradiotherapy (nCRT) [12]. As mentioned earlier, our study found that the 5-gene signature was strongly correlated with tumor-specific CD4$^+$ and CD8$^+$T cells. Obviously, these data were consistent with studies of TILs in colorectal cancer.

Another interesting finding of our study was that the 5-gene signature was predicted to be intersected in hsa-miR-6887-3p, whereas only GNG7 correlated with hsa-miR-6887-3p. MicroRNAs (miRNAs) are critical mediators of tumorigenesis in many human cancers. The role of miRNAs as clinical biomarkers in colorectal cancer research is promising [10]. Many studies have documented aberrant miRNA levels as biomarkers of colorectal cancer and reported the evaluation of their potential roles as diagnostic and prognostic indicators. Li H demonstrated that hsa-miR-6887-3p inhibited the tumorigenesis of colorectal cancer by downregulating Mex3a expression and functioned as an important regulator in the hsa-miR-6887-3p/Mex3a/RAP1GAP signaling axis [39]. However, the clinical application of miRNAs as predictive biomarkers in rectal cancer remains to be seen [9]. Our data also suggest the need for further studies to investigate the biological roles of hsa-miR-6887-3p and GNG7 in rectal cancer.

It is noteworthy that the 5-gene signature was correlated with the vascular invasion ($P = 0.038$), but not the lymphatic invasion. The vascular invasion has been associated with an increased risk of regional and distant metastasis in colorectal cancer patients [40]. Some authors have reported that vascular invasion was a strong prognosticator for rectal cancer affecting disease progression and survival [40–42]. However, a few analyses have failed to indicate the prognostic value of vascular invasion for the survival of colon, colorectal, and rectal cancers [43,44]. Therefore, it is important to identify biomarkers of vascular invasion to better diagnose and classify patients with rectal cancer. Until now, no validated tumor markers have been identified to date [44–46]. Our findings suggest that vascular invasion is crucial for the progression of colorectal cancer, which warrants further investigation. It has also been reported that lymphatic invasion is associated with an unfavorable prognosis of colorectal cancer [41]. And yet, our data did not show any statistically significant association between lymphatic invasion and prognosis. Based on our results, further exploration of the function of the 5-gene signature in vascular invasion is warranted.

One potential limitation of our research was that it was predominantly based on bioinformatics results. Our study was based on two independent datasets, but the sample number is limited. The majority of the research focused on colorectal cancer, however, unfortunately, many public studies had no prognostic information. As a result, there were 110 cases in total that fitted the research model exactly. Moreover, we designed to decrease bias and increase the repeatability of the analytic results. Nevertheless, our findings needed to be validated in a more prospective study, and the correlation between gene markers and other factors requires further investigation. Another key question is whether the 5 genes indeed affect the progression of rectal cancer. Consequently, future research will be essential to understand the biological association between the expression of these 5 genes and rectal cancer.

To summarize, as highlighted in this paper, the 5-gene-based signature was a robust prognosticator for rectal cancer patients. More importantly, it was an independent prognostic factor based on the Cox regression model. Our research aimed to uncover the complex interaction between 5 genes and clinical outcomes in rectal cancer through the use of microRNA and single-cell analysis tools. In summary, our results indicated that the 5-gene-based signature could contribute to the prognostic evaluation of rectal cancer and might pave the way for novel therapeutic strategies in the foreseeable future.

Authors' contributions

W. Li and X. Wang conceived the idea, supervised the project and revised the paper.

Y. Lin, Q. Ji, Y. Peng and Y. Zheng performed the bioinformatics analyses and generated the data.

C. Yu, J. Zheng, R. Bai, W. Yan, and X. Kang assisted with the data analyses, interpretation and paper discussion.

Y. Lin, Q. Ji, Y. Peng and Parker Li wrote and revised the manuscript.

W. Li, X. Wang and Y. Lin funded the research and contributed to the manuscript preparation.

All the authors read and approved the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Our study was supported by funds from the National Natural Science Foundation of China (Li Wenbin, Grant No. 81972338); the National Natural Science Foundation of China (Wang Xiaomin, Grant No. 82070169); and the Natural Foundation of Capital Medical University (Lin Yi, Grant No. PYZ2017160).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tranon.2022.101529.

References

[1] F. Bray, et al., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 68 (6) (2018) 394–424.
[2] W. Chen, et al., National cancer incidence and mortality in China, 2012, Chin. J. Cancer Res. 28 (1) (2016) 1–11.
[3] W. Cao, et al., Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020, Chin. Med. J. 134 (7) (2021) 783–791.
[4] A.W. Wu, et al., Pattern and management of recurrence of mid-low rectal cancer after neoadjuvant intensity-modulated radiotherapy: single-Center Results of 687 cases, Clin. Colorectal Cancer 17 (2) (2018) e307–e313.
[5] C. Dong, et al., Update in version 2021 of CSCO guidelines for colorectal cancer from version 2020, Chin. J. Cancer Res. 33 (3) (2021) 302–307.
[6] B.M. Ghadimi, et al., Effectiveness of gene expression profiling for response prediction of rectal adenocarcinomas to preoperative chemoradiotherapy, J. Clin. Oncol. 23 (9) (2005) 1826–1838.
[7] I.J. Park, et al., A nine-gene signature for predicting the response to preoperative chemoradiotherapy in patients with locally advanced rectal cancer, Cancers 12 (4) (2020).
[8] S. Kim, et al., Elevated CXCL12 in the plasma membrane of locally advanced rectal cancer after neoadjuvant chemoradiotherapy: a potential prognostic marker, J. Cancer 13 (1) (2022) 162–173.
[9] R.M. Waldron, et al., MicroRNAs as biomarkers of multimodal treatment for rectal cancer, Br. J. Surg. 108 (8) (2021) e260–e261.
[10] S. Ghafouri-Fard, et al., MicroRNAs as important contributors in the pathogenesis of colorectal cancer, Biomed. Pharmacother. 140 (2021), 111759.
[11] J. Galon, et al., Type, density, and location of immune cells within human colorectal tumors predict clinical outcome, Science 313 (5795) (2006) 1960–1964.
[12] K. Yasuda, et al., Density of CD4(+) and CD8(+) T lymphocytes in biopsy samples can be a predictor of pathological response to chemoradiotherapy (CRT) for rectal cancer, Radiat. Oncol. 6 (2011) 49.
[13] H.K.S. Hamid, et al., Prognostic and predictive value of neutrophil-to-lymphocyte ratio after curative rectal cancer resection: a systematic review and meta-analysis, Surg. Oncol. 37 (2021), 101556.
[14] K. Zhao, et al., A prognostic five long-noncoding RNA signature for patients with rectal cancer, J. Cell. Biochem. (2019).
[15] K. Kalinsky, et al., 21-Gene assay to inform chemotherapy benefit in node-positive breast cancer, N. Engl. J. Med. 385 (25) (2021) 2336–2347.
[16] A. Cercek, et al., PD-1 blockade in mismatch repair-deficient, locally advanced rectal cancer, N. Engl. J. Med. 386 (25) (2022) 2363–2376.
[17] I. Seo, et al., Neoadjuvant chemoradiation alters biomarkers of anticancer immunotherapy responses in locally advanced rectal cancer, J. Immunother. Cancer 9 (3) (2021).
[18] A. Rizzo, A.D. Ricci, PD-L1, TMB, and other potential predictors of response to immunotherapy for hepatocellular carcinoma: how can they assist drug clinical trials? Expert Opin. Investig. Drugs 31 (4) (2022) 415–423.
[19] M.J. Goldman, et al., Visualizing and interpreting cancer genomics data via the Xena platform, Nat. Biotechnol. 38 (6) (2020) 675–678.
[20] V. Agarwal, et al., Predicting effective microRNA target sites in mammalian mRNAs, Elife 4 (2015).
[21] H. Yuan, et al., CancerSEA: a cancer single-cell state atlas, Nucleic Acids. Res. 47 (D1) (2019) D900–D908.
[22] F. Pages, et al., Effector memory T cells, early metastasis, and survival in colorectal cancer, N. Engl. J. Med. 353 (25) (2005) 2654–2666.
[23] D. Jeong, et al., RhoA is associated with invasion and poor prognosis in colorectal cancer, Int. J. Oncol. 48 (2) (2016) 714–722.
[24] B. Aykut, et al., EMX2 gene expression predicts liver metastasis and survival in colorectal cancer, BMC Cancer 17 (1) (2017) 555.
[25] M. Paauwe, et al., Endoglin expression on cancer-associated fibroblasts regulates invasion and stimulates colorectal cancer metastasis, Clin. Cancer Res. 24 (24) (2018) 6331–6344.
[26] Y. Tian, et al., Expression and clinical significance of POLR1D in colorectal cancer, Oncology 98 (3) (2020) 138–145.
[27] K.J. Anderson, R.T. Cormier, P.M. Scott, Role of ion channels in gastrointestinal cancer, World J. Gastroenterol. 25 (38) (2019) 5732–5772.
[28] D.T. Petrova, et al., Expression of chloride intracellular channel protein 1 (CLIC1) and tumor protein D52 (TPD52) as potential biomarkers for colorectal cancer, Clin. Biochem. 41 (14–15) (2008) 1224–1236.
[29] Y.J. Deng, et al., CLIC4, ERp29, and Smac/DIABLO derived from metastatic cancer stem-like cells stratify prognostic risks of colorectal cancer, Clin. Cancer Res. 20 (14) (2014) 3809–3817.
[30] M. Uhlen, et al., Proteomics. Tissue-based map of the human proteome, Science 347 (6220) (2015), 1260419.
[31] Y. An, et al., Identification of ENTPD8 and cytidine in pancreatic cancer by metabolomic and transcriptomic conjoint analysis, Cancer Sci. 109 (9) (2018) 2811–2821.
[32] R. Aukkanimart, et al., Altered expression of oxidative metabolism related genes in cholangiocarcinomas, Asian Pac. J. Cancer Prev. 16 (14) (2015) 5875–5881.
[33] X. Yu, et al., Key candidate genes associated with BRAF(V600E) in papillary thyroid carcinoma on microarray analysis, J. Cell. Physiol. 234 (12) (2019) 23369–23378.

[34] K. Shibata, et al., Identification and cloning of human G-protein gamma 7, down-regulated in pancreatic cancer, Biochem. Biophys. Res. Commun. 246 (1) (1998) 205–209.

[35] K. Shibata, et al., G-protein gamma 7 is down-regulated in cancers and associated with p 27kip1-induced growth arrest, Cancer Res. 59 (5) (1999) 1096–1101.

[36] M. Ohta, et al., Clinical significance of the reduced expression of G protein gamma 7 (GNG7) in oesophageal cancer, Br. J. Cancer 98 (2) (2008) 410–417.

[37] H. Zheng, et al., G protein gamma 7 suppresses progression of lung adenocarcinoma by inhibiting E2F transcription factor 1, Int. J. Biol. Macromol. 182 (2021) 858–865.

[38] S.T. Makkai-Popa, et al., Corelation of lymphocytic infiltrates with the prognosis of recurrent colo-rectal cancer, Chirurgia 108 (6) (2013) 859–865.

[39] H. Li, et al., Mex3a promotes oncogenesis through the RAP1/MAPK signaling pathway in colorectal cancer and is inhibited by hsa-miR-6887-3p, Cancer Commun. 41 (6) (2021) 472–491.

[40] M.J. Krasna, et al., Vascular and neural invasion in colorectal carcinoma. Incidence and prognostic significance, Cancer 61 (5) (1988) 1018–1023.

[41] J.A. de Ridder, et al., Lymphatic invasion is an independent adverse prognostic factor in patients with colorectal liver metastasis, Ann. Surg. Oncol. 22 (suppl 3) (2015) S638–S645.

[42] A. Sternberg, et al., Conclusions from a study of venous invasion in stage IV colorectal adenocarcinoma, J. Clin. Pathol. 55 (1) (2002) 17–21.

[43] G. Bianchi, et al., Three distinct outcomes in patients with colorectal adenocarcinoma and lymphovascular invasion: the good, the bad, and the ugly, Int. J. Colorectal Dis. (2021).

[44] R.G. Campanati, et al., Primary tumor lymphovascular invasion negatively affects survival after colorectal liver metastasis resection? Arq. Bras. Cir. Dig. 34 (1) (2021) e1578.

[45] H. Tao, et al., Construction of a ceRNA network and a prognostic lncRNA signature associated with vascular invasion in hepatocellular carcinoma based on weighted gene co-expression network analysis, J. Cancer 12 (13) (2021) 3754–3768.

[46] O.S. Guner, L.V. Tumay, Persistent extramural vascular invasion positivity on magnetic resonance imaging after neoadjuvant chemoradiotherapy predicts poor outcome in rectal cancer, Asian J. Surg. 44 (6) (2021) 841–847.