Am J Cancer Res 2023;13(1):45-65
www.ajcr.us /ISSN:2156-6976/ajcr0146924

Original Article

In silico development and in vitro validation of a novel five-gene signature for prognostic prediction in colon cancer

Qiankun Zhu[1,2,3,4,5], Benqiang Rao[1,2,3,4,5], Yongbing Chen[1,2,3,4,5], Pingping Jia[1,2,3,4,5], Xin Wang[1,2,3,4,5], Bingdong Zhang[1,2,3,4,5], Lin Wang[1,2,3,4,5], Wanni Zhao[1,2,3,4,5], Chunlei Hu[1,2,3,4,5], Meng Tang[1,2,3,4,5], Kaiying Yu[1,5], Wei Chen[5,6], Lei Pan[5,7], Yu Xu[8], Huayou Luo[8], Kunhua Wang[9], Bo Li[10], Hanping Shi[1,2,3,4,5]

[1]Department of General Surgery, Beijing Shijitan Hospital, Capital Medical University, Beijing 100038, The People's Republic of China; [2]Department of Clinical Nutrition, Beijing Shijitan Hospital, Capital Medical University, Beijing 100038, The People's Republic of China; [3]Beijing International Science and Technology Cooperation Base for Cancer Metabolism and Nutrition, Beijing 100038, The People's Republic of China; [4]Key Laboratory of Cancer FSMP for State Market Regulation, Beijing 100038, The People's Republic of China; [5]Ninth School of Clinical Medicine, Peking University, Beijing 100038, The People's Republic of China; [6]Department of Intensive Care Unit, Beijing Shijitan Hospital, Capital Medical University, Beijing 100038, The People's Republic of China; [7]Department of Respiratory and Critical Care, Beijing Shijitan Hospital, Capital Medical University, Beijing 100038, The People's Republic of China; [8]Department of General Surgery, The First Affiliated Hospital of Kunming Medical University, Kunming 650032, Yunnan, The People's Republic of China; [9]Yunnan University, Kunming 650091, Yunnan, The People's Republic of China; [10]Department of General Surgery, The Affiliated Hospital of Yunnan University, Kunming 650091, Yunnan, The People's Republic of China

Received October 10, 2022; Accepted December 24, 2022; Epub January 15, 2023; Published January 30, 2023

Abstract: Colon cancer is one of the most common cancers in digestive system, and its prognosis remains unsatisfactory. Therefore, this study aimed to identify gene signatures that could effectively predict the prognosis of colon cancer patients by examining the data from the Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) database. LASSO-Cox regression analysis generated a five-gene signature (*DCBLD2*, *RAB11FIP1*, *CTLA4*, *HOXC6* and *KRT6A*) that was associated with patient survival in the TCGA cohort. The prognostic value of this gene signature was further validated in two independent GEO datasets. GO enrichment revealed that the function of this gene signature was mainly associated with extracellular matrix organization, collagen-containing extracellular matrix, and extracellular matrix structural constituent. Moreover, a nomogram was established to facilitate the clinical application of this signature. The relationships among the gene signature, mutational landscape and immune infiltration cells were also investigated. Importantly, this gene signature also reliably predicted the overall survival in IMvigor210 anti-PD-L1 cohort. In addition to the bioinformatics study, we also conducted a series of in vitro experiments to demonstrate the effect of the signature genes on the proliferation, migration, and invasion of colon cancer cells. Collectively, our data demonstrated that this five-gene signature might serve as a promising prognostic biomarker and shed light on the development of personalized treatment in colon cancer patients.

Keywords: Colon cancer, prediction, prognosis, gene signature, risk model

Introduction

Colon cancer is the fourth most common cancer and the third leading cause of cancer-related death worldwide [1]. In the United States, it was estimated that approximately 104,270 new cases and 52,980 deaths were diagnosed with and attributed to colon cancer in 2021 [2]. The main treatment modalities of colon cancer include surgery, chemotherapy, radiotherapy, targeted therapy, and immunotherapy [3]. Although combinational therapies have been widely implemented to achieve better therapeutic efficacy [4], the results are still not satisfactory, especially in the long-term survival of patients with quality of life [5]. Thus, it is crucial to investigate the mechanism of tumorigenesis and identify predictive biomarkers for the diagnosis and prognosis of patients with colon cancer.

The etiology of colon cancer comprises environmental exposures, genetic factors, intestinal flora, and family history of the disease [6]. Identifying genetic mutations that affect the prognosis of colon cancer has become a research hotspot over the last decade. For example, *TP53*, *KRAS* and *BRAF* mutations have been found to be associated with the prognosis or therapeutic response of colon cancer [7]. In addition, a consensus molecular subtype classification based on microarray or RNA sequencing data has been developed to predict the different clinical outcomes of colon cancer [8]; however, its reliability and clinical application need further validation. Hence, exploring novel molecular markers to identify high-risk subgroups and to improve the treatment response is highly desirable.

In recent years, immunotherapy is emerging as a promising treatment strategy for cancer patients [9]. Immune checkpoint inhibitor, such as programmed cell death 1 (PD-1) antibody, has been approved by the U.S. Food and Drug Administration (FDA) for the management of colon cancer in 2017 [10]. But the heterogeneity and complexity of tumor immune microenvironment compromise its therapeutic effects in patients with advanced colon cancer [11]. As a result, increasing attention has been focused on identifying patients who would benefit from immunotherapy. The advent of high-throughput sequencing and large-scale databases has provided unprecedented opportunities to accelerate such process from genetic perspectives.

In this study, we aimed to identify gene signatures that distinguish high-risk subgroup by analyzing the mRNA sequencing data of colon cancer patients from The Cancer Genome Atlas (TCGA). Moreover, we validated the gene signatures externally using Gene Expression Omnibus (GEO) datasets. Further bioinformatics analysis was also conducted to predict the molecular function and immunotherapeutic response of the gene signature. Finally, we conducted in vitro biochemical experiments to investigate the biological activity of each gene in the signature.

Materials and methods

Data acquisition and preprocessing

The mRNA expression data and the corresponding clinical characteristics of colon adenocarcinoma (COAD) patients were obtained from the publicly accessible datasets at the NCBI TCGA database (https://portal.gdc.cancer.gov/) and GEO database (https://www.ncbi.nlm.nih.gov/geo/). In TCGA COAD, we downloaded the raw counts of each patient and filtered the datasets by the following inclusion criteria: 1) samples are pathologically diagnosed as colon cancer; 2) samples with complete interested clinical information; 3) transcript is within the protein-coding region; 4) transcript is expressed in more than half the COAD samples; 5) reads per kilobase of exon model per million mapped reads (RPKM) has an average value of more than 0.1 across all the samples. At the end, a total of 394 patents in TCGA were enrolled. In GEO database, we selected GSE39582 and GSE17538 datasets due to their larger sample size (n = 542 and n = 238, respectively) among similar cohorts for our model validation. The gene expression and clinical annotation of patients from an immunotherapeutic cohort of advanced urothelial cancer (IMvigor210 cohort) treated with atezolizumab (anti-PD-L1 agent) were obtained from the official website (http://researchpub.gene.com/IMvigor210CoreBiologies) [12].

Risk model construction and external validation

The TCGA cohort was used as training cohort to construct the risk predicting model. Univariate Cox regression analysis, random forest selection, and LASSO-Cox regression analysis were conducted to identify the molecular signature. Then, the coefficient of genes in LASSO-Cox regression (glmnet package version 4.1) was utilized to calculate the risk score using the following formula [13]: risk score = $\sum_{i=1}^{n}$ Coefficient i * Expression i, whereas Coefficient i and Expression i denote the coefficient and expression of the ITH gene in LASSO model. Using the median risk score as a cutoff value, we assigned the patients into low- and high-risk groups. Subsequently, Kaplan-Meier estimates were used to assess the survival of patients between the high- and low-risk groups. Furthermore, multivariate Cox regression analysis (survival package version 3.2-7) was carried out to determine the independence of this risk score model from other clinical variables. Time-dependent receiver operating characteristic (ROC) curves (survivalROC package version 1.0.3) were plotted to illustrate the sensitivity and specificity of the survival prediction of the risk model. To verify the reliability of

this risk predictive model, GSE39582 and GSE17538 were used as the external validation cohorts; meanwhile, IMvigor210 cohort was used to evaluate its predictive value to immune therapeutic response. All statistical analyses were conducted using R software (version 4.0.3).

Expression analysis of genes in risk model

To further evaluate the mRNA and protein expression of the genes used in our risk signature, immunohistochemistry staining and mRNA expression data from the Human Protein Atlas (HPA) database were analyzed.

Gene set enrichment analysis (GSEA)

DESeq2 package (version 1.28.1) was first used to compare the differential gene expression between the high- and low-risk groups. *P* value <0.05 was considered statistically significant. The significantly differentially expressed genes were then subjected to gene set enrichment analysis using ClusterProfiler package (version 3.16.1). Kyoto Encyclopedia of Genes and Genomes (KEGG) background was set with c2.cp.kegg.v7.2.entrez which was downloaded from molecular signatures database (MSigDB) (http://www.gsea-msigdb.org/gsea/downloads.jsp). Gene Ontology (GO) comprising biological process (BP), molecular function (MF), and cellular component (CC) was conducted to achieve the functional annotation of the differentially expressed genes between the high- and low-risk groups.

Identification of hub genes by WGCNA

Weighted gene co-expression network analysis (WGCNA) was conducted by WGCNA R package (version 1.69). Soft threshold power value was calculated automatically, and a soft power of 8 was selected for subsequent analysis. The minimum module size was 30, and the type of topological overlap measure (TOM) was set as "signed". Each module was labeled with a unique color. Hub genes were defined when these two criteria were satisfied simultaneously: 1) gene trait significance (GS) >0.2; 2) gene module membership (MM) >0.8.

Development and validation of nomogram for survival prediction

Multivariate Cox regression analysis was first conducted for risk score and clinical variables including age, gender and TNM stage. Then, a nomogram containing the multivariate Cox results was visualized by the R package rms (version 6.1-0). Using a bootstrap with 1000 times of resample, we verified the performance of the nomogram by drawing calibration curve and time-dependent ROC curve. Decision curve analysis (DCA) (dca package version 0.1.0.9000) was used to evaluate the accuracy of the nomogram.

Significantly mutated genes between the high- and low-risk groups

TCGA data were analyzed by R package (Maftools, version 2.4.12) to extract gene mutations, including nonsense mutation, missense mutation, frame shift insertion, frame shift deletion, splice site mutation, in frame deletion, and translation start site mutation. Transition (C>T, and T>C) and transversion (C>G, C>A, T>A, and T>G) were also counted to show overall distribution of conversions. The significantly different gene mutations between the high- and low-risk groups were compared by Fisher exact tests.

Immune cell infiltration analysis

The single sample gene set enrichment analysis (ssGSEA) algorithm was used to quantify the infiltration level of 28 types of immune cell in TCGA COAD patients according to a recent publication [14]. The calculation of abundance was attained by the package GSVA (version 1.38.0) with Gaussian fitting model to generate an enrichment score.

Cell line and cell culture

Human colon cancer HCT116 cells were purchased from ATCC cell bank (USA) and cultured in DMED medium (Hyclone, USA) supplemented with 10% fetal bovine serum (GIBCO, USA). Cells were maintained in a 37°C humidified incubator with 5% CO_2.

Cell transfection

Three different small interfering RNAs (siRNAs) specifically targeting an indicated gene were synthesized by Hippobiotec company (Zhejiang, China). The siRNA sequences were shown in Table S1. Scramble siRNAs were constructed as the negative control. The siRNA with the maximal knockdown efficiency was selected for further functional assays. Overexpression plas-

mids for the selected genes were purchased from HTHealth company (Beijing, China). Approximately 1*10^5 cells were seeded into each well in 6-well plates, transfected with appropriated plasmid for 48 hours, and collected for further analysis.

Real time quantitative PCR (RT-qPCR)

Total RNA from cells was extracted using TRIzol reagent (Invitrogen, MA, USA) according to the manufacturer's instruction and quantified by Thermo Nanodrop lite Spectrophotometry. Reverse transcription was performed using SuperScript III reverse transcriptase (Invitrogen, MA, USA) to obtain cDNA. mRNA expression was analyzed by the SYBR green RT-PCR kit (Invitrogen, MA, USA). Beta-actin was used as normalization, and the comparative Ct method ($2^{-\Delta\Delta Ct}$) was used to evaluate relative expression. The sequences of primers used in this study were shown in Table S2.

Western blot

Cell lysates were collected using cell lysis buffer in the presence of protease inhibitors, and protein concentration was quantified by Bicinchoninic acid (BCA) quantification kit (MDL, China). Proteins were separated by 10% SDS-PAGE and then transferred onto polyvinylidene fluoride (PVDF) membranes (Millipore, USA). The membranes were blocked in 5% skimmed milk at room temperature for 1 h and subsequently incubated with specific primary antibodies at 4°C overnight followed by incubation in secondary antibodies at room temperature for 1 h. The signal was then developed by ECL reagent and detected using a chemiluminescence imaging system (Bio-rad, USA). The information of the primary antibodies used in this study were summarized in Table S3.

Cell viability assay

Cell viability was evaluated by CCK-8 kit (Fluorescence, China). Briefly, cells were seeded into 96-well plates and incubated for 12 hours. Then, 10 μL of CCK-8 solution was added to each well for 1 hour, and the optical density (OD) values at 450 nm was measured by a microtiter plate reader.

Transwell assay

Cell motility was measured by transwell assays with 8.0 μm transparent PET membrane (for cell migration assay). For cell invasion assays, the upper surface of the membrane on the insert was coated with Matrigel. Briefly, cells (1*10^5) in serum free medium were added to the upper chamber, while complete medium was added to the lower chamber. After 24 hours of incubation, cells that invaded into the lower side of the insert were fixed with 4% paraformaldehyde and stained with 0.1% crystal violet.

Wound healing assay

Cells were cultured in 6-well plates, and artificial wounds were created by scraping with a sterile 200 μl pipette tip. Then, the cells were washed with PBS for three times and incubated in medium containing 10% fetal bovine serum. Wound closure was determined via photographing at 24 hours after injury.

Statistical analysis

Statistical analysis was conducted by R software (version 4.0.3). For categorical variables, Fisher exact test was performed to compare the difference. For continuous data, student T test or Wilcox test was applied depending on the data distribution (skewed or normal) and the variances (unequal or homogeneous). Survival curves in Kaplan-Meier analysis were compared by two-sided log-rank test. A two-tailed *P* value of 0.05 was set for statistical significance.

Results

Patients' demographic and clinical characteristics

The workflow pertaining to the present study was shown in **Figure 1**. In the TCGA-COAD cohort, 394 patients were selected according to our inclusion criteria, while 542 and 238 patients from GSE39582 and GSE17538 datasets, respectively, were included. The patients' demographic and clinical characteristics were summarized in **Table 1**.

Identification of gene signature with prognostic values in the TCGA cohort

The TCGA cohort was subjected to univariate Cox regression analysis as well as the random forest selection and LASSO-Cox regression analysis. The LASSO-Cox model with a lambda of 0.001135068 was selected to develop the

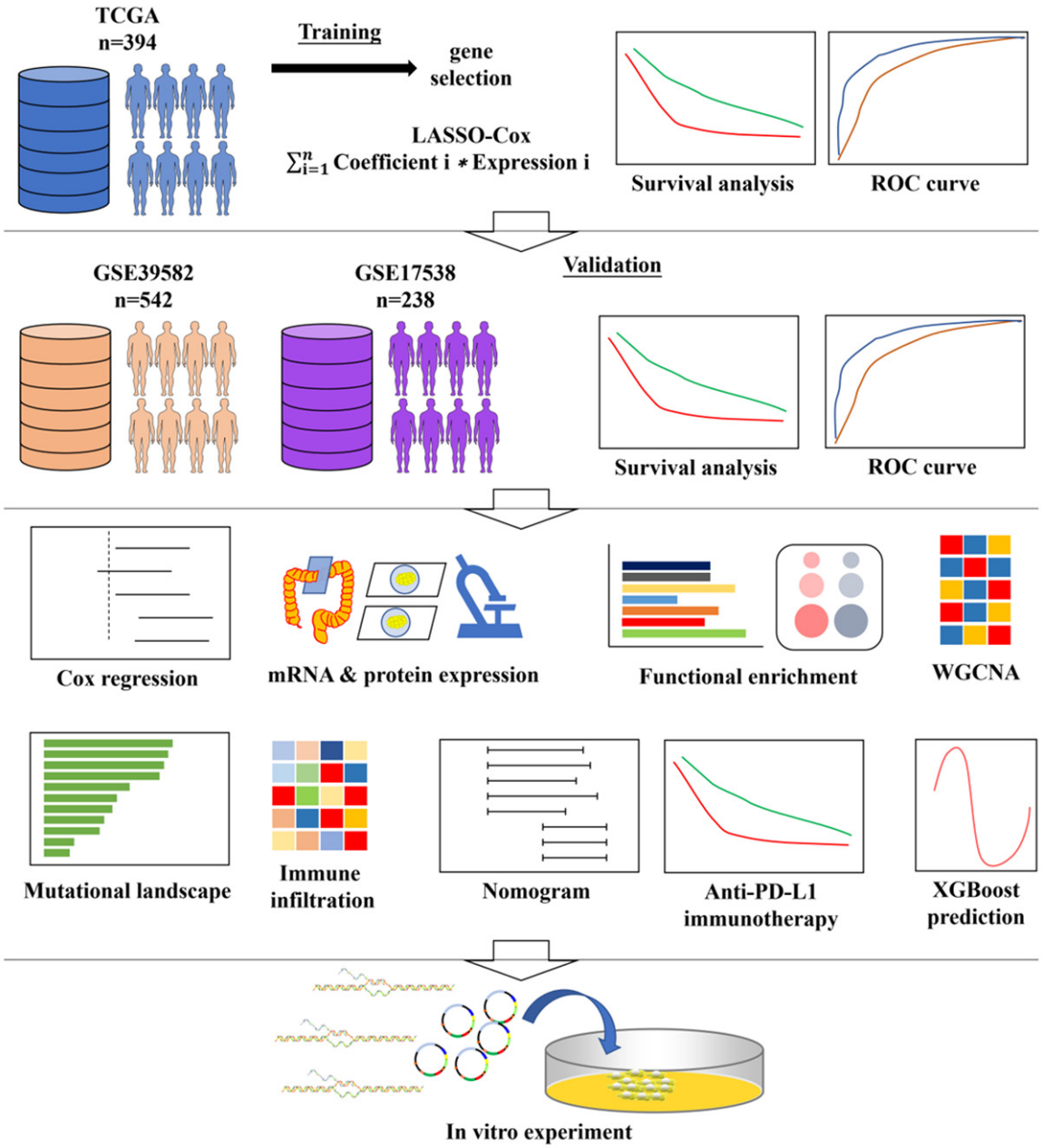


Figure 1. Workflow of the training and validation process in the present investigation.

risk score model, and subsequently, five genes were identified. Based on the coefficient, the model was constructed as follows: risk score = (0.1443***DCBLD2*) + (-0.0585***RAB11FIP1*) + (-0.3256***CTLA4*) + (0.1447***HOXC6*) + (0.0745***KRT6A*). Among the five genes, the coefficients of three genes were positive, suggesting the association of higher gene expressions with poor survival, while those of the other two genes were negative, suggesting their favorable effects on prognosis. We further calculated the five-gene signature risk score for each patient in the TCGA cohort, and by using the median risk score as the cutoff threshold, we divided the TCGA cohort into high-risk (n = 197) and low-risk (n = 197) groups. The distribution of risk score was shown in **Figure 2A-C**. In the TCGA cohort, patients with high-risk scores showed poorer prognosis in the survival analysis than those with low-risk scores (log-rank test, P = 0.037) (**Figure 2D**). Specifically, the survival rates of 50, 100, and 150 months

Table 1. Demographic and clinical characteristics of the TCGA and GEO cohorts

	TCGA (n = 394)	GSE39582 (n = 542)	GSE17538 (n = 238)
Age in yrs (mean ± SD)	67.26±13.01	66.74±13.34	64.56±13.49
Gender = Male (%)	206 (52.3%)	295 (54.4%)	124 (52.1%)
Status = Dead (%)	46 (11.7%)	170 (31.4%)	93 (40.1%)
Survival time in months (median [IQR])	61.00 [2.00, 301.75]	54.50 [28.00, 82.00]	46.72 [23.32, 63.96]
TNM stage (%)			
Stage 0	0 (0.0%)	1 (0.2%)	0 (0.0%)
Stage 1	71 (18.0%)	36 (6.6%)	28 (12.1%)
Stage 2	152 (38.6%)	255 (47.0%)	72 (31.0%)
Stage 3	117 (29.7%)	191 (35.2%)	76 (32.8%)
Stage 4	54 (13.7%)	59 (10.9%)	56 (24.1%)

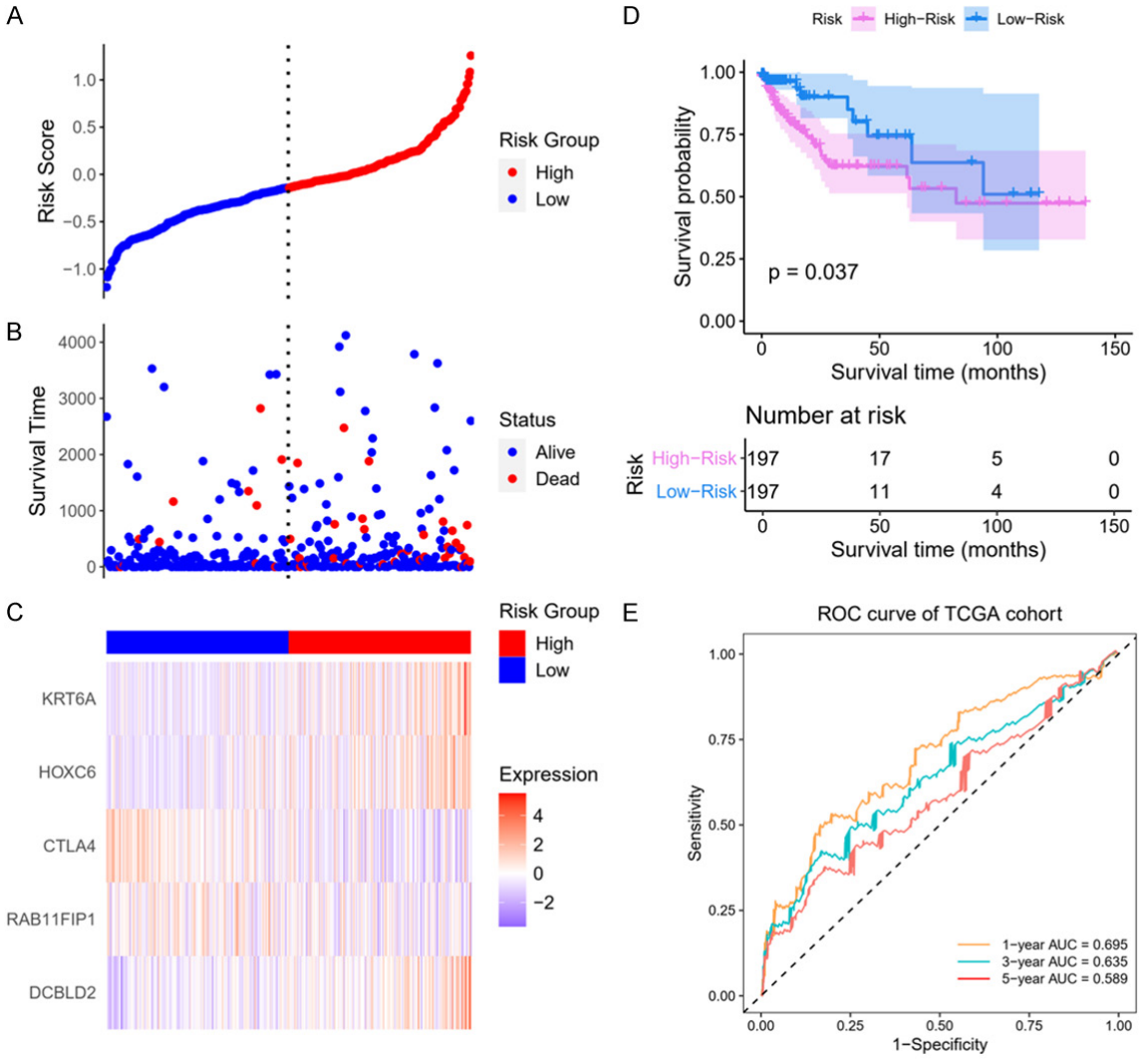


Figure 2. The five-gene signature predicting overall survival in patients of the TCGA cohort. A. The five-gene risk score distribution. B. The survival time and status arranged by increasing risk score. C. Heatmap of the five genes' expression profiles. D. Kaplan-Meier estimate of survival probability between high- and low-risk group. E. Receiver operating characteristic (ROC) curve at one-, three-, and five-year time point.

were 5.58%, 2.03%, and 0.00%, respectively, in the low-risk group, compared with those of 8.63%, 2.54%, and 0.00% in the high-risk group. Time dependent ROC analysis revealed that the prognosis model had achieved area under the curve (AUC) score of 0.695, 0.635, and 0.589 for one-, three-, and five-year survival, respectively (**Figure 2E**).

Validation of the five-gene signature in GSE39582 and GSE17538 datasets

To confirm our findings from TCGA cohort, we used GSE39582 and GSE17538 datasets to verify the reliability of our risk model. By applying the same formula, we calculated the risk score and classified patients into high- and low-risk groups based on the median score as the cutoff value (**Figure 3A**, **3E**). Consistent with the results from the TCGA cohort, Kaplan-Meier analysis also suggested that patients in the GSE39582 cohort with high-risk scores had significantly shorter overall survival time than that of patients with low-risk scores (log-rank test, P = 7E-4) (**Figure 3B**). Recurrent free survival curve also demonstrated similar pattern, as the disease in the high-risk patients of GSE39582 cohort progressed much faster (**Figure 3C**). ROC curve suggested that the five-year AUC was 0.64 (**Figure 3D**). Similarly, in the GSE17538 dataset, patients with high-risk scores demonstrated a remarkably unfavorable prognosis than those with low-risk scores (log-rank P<0.05 for overall survival, disease free or disease specific survival) (**Figure 3F-H**). The five-year AUC in ROC curve was 0.683 (**Figure 3I**).

Independence of the five-gene signature from other clinical variables

Furthermore, to evaluate the independence of the five-gene signature, we carried out univariate and multivariate Cox regression analyses on this risk signature as well as on other clinical variables, including age, gender, and TNM stage. Univariate Cox analysis indicated that the five-gene risk score was significantly associated with overall survival in the TCGA cohort (P<0.05) (**Table 2**). The significance also emerged in multivariate Cox analysis of TCGA when age, gender and TNM stage were taken into account simultaneously (Figure S1A). In GEO datasets, results demonstrated the independence of the risk score after adjusting the clinical variables as well (Figure S1B, S1C). Furthermore, we found that TNM stage IV was significantly associated with overall survival in either multivariate or univariate analysis.

Comparison of the current signature with previously published signatures

To demonstrate the significance of our risk model, we compared our prognostic signature with five published signatures: a two-gene signature (Liu) [15], a three-gene signature (Xu) [16], two seven-gene signatures (Sun; Zou) [17, 18], and a nine-gene signature (Yang) [19]. The GSE39582 and GSE17538 cohorts were also divided into high- and low-risk groups by using the median score as the cutoff value. ROC and survival curves were plotted (Figure S2). When comparing the AUC values for five-year survival, the five published signatures showed the values of 0.503, 0.496, 0.578, 0.611 and 0.506, respectively, in GSE39582 cohort and 0.461, 0.568, 0.606, 0.537, and 0.541, respectively, in GSE17538 cohort. However, the values of our signature were 0.640 in GSE39582 and 0.683 in GSE17538, higher than that of the previous models suggesting the higher accuracy of our signature. In addition, we calculated the concordance indexes (C-indexes) to evaluate the predictive performance of these signatures. Results unveiled that our signature had the highest C-index (0.6148 in GSE39582 and 0.6497 in GSE17538, Figure S2).

The mRNA and protein expression of risk predicting genes

As shown in **Figure 4**, we examined the mRNA and protein expressions of the five genes in our risk predicting signature by using information from HPA database. Known available mRNA expression and tissue staining were extracted, and we found a substantial dysregulation of both mRNA and protein expressions in colon cancer samples, suggesting their potential effects on the carcinogenesis of colon cancer.

Functional annotation with GSEA enrichment of high-risk and low-risk groups

To access the potential biological functions of the risk score signature, GSEA were performed using TCGA database. In KEGG analysis, the main pathways enriched in the high-risk group were PPAR signaling pathway, maturity onset

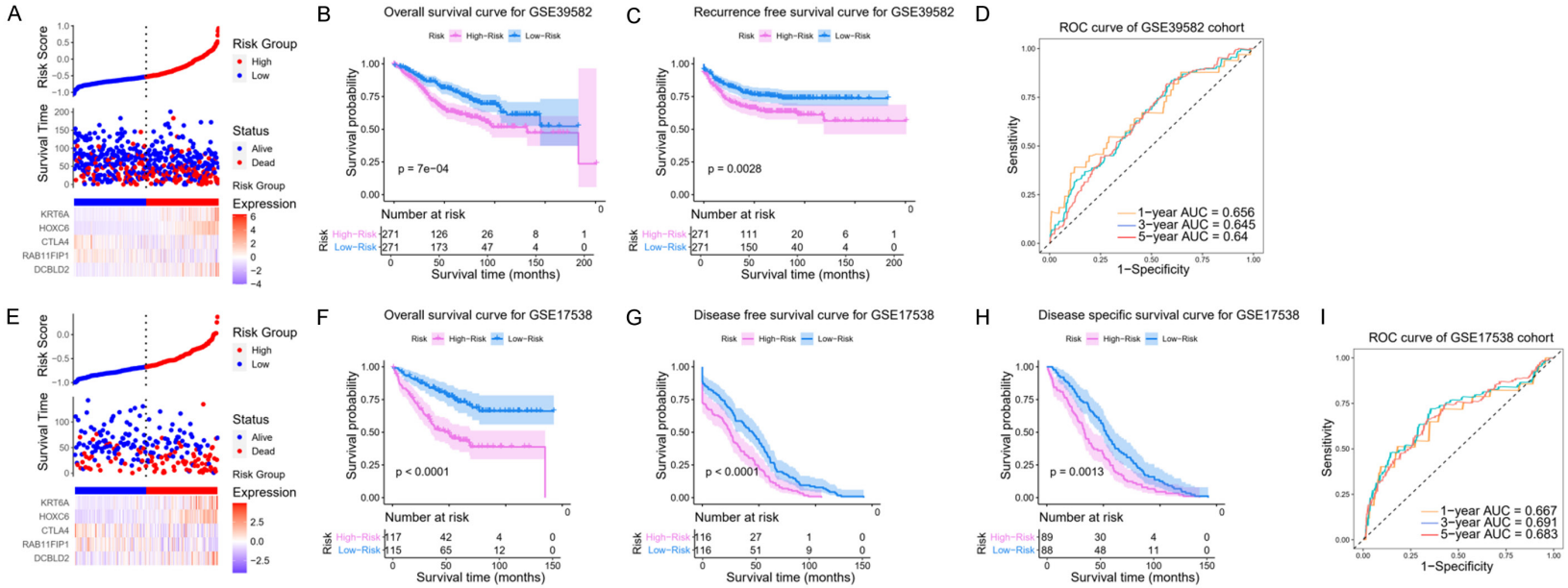


Figure 3. Validation of the gene signature in GSE39582 and GSE17538 datasets. A. Risk score distribution, overall survival status, and heatmap of five-gene signature in GSE39582. B. Overall survival in GSE39582 cohort. C. Recurrence free survival in GSE39582 cohort. D. Receiver operating characteristic (ROC) curve of GSE39582. E. Risk score distribution, overall survival status, and heatmap of five-gene signature in GSE17538. F. Overall survival in GSE17538 cohort. G. Disease free survival in GSE17538 cohort. H. Disease specific survival in GSE17538 cohort. I. ROC curve of GSE17538.

Table 2. Univariate Cox regression analysis in the TCGA cohort, GSE39582 cohort, and GSE17538 cohort

	Variable	Hazard ratio	Lower 95% CI	Upper 95% CI	*P* value
TCGA cohort (n = 394)	Age	1	1	1	0.01794
	Gender	1.342	0.7485	2.406	0.3234
	TNM stage				
	I	1 (ref)			
	II	1.287	0.3667	4.519	0.6934
	III	1.354	0.3825	4.792	0.6385
	IV	4.883	1.408	16.93	0.01243
	Risk score	2.909	1.539	5.498	0.00101
GSE39582 (n = 542)	Age	1.026	1.013	1.039	7.399E-05
	Gender	1.299	0.9561	1.765	0.0944
	TNM stage				
	I	1 (ref)			
	II	1.939	0.7818	4.808	0.153
	III	2.321	0.9294	5.796	0.07136
	IV	11.16	4.378	28.46	4.38E-07
	Risk score	2.697	1.807	4.028	1.224E-06
GSE17538 (n = 238)	Age	1.009	0.9923	1.025	0.304
	Gender	1.006	0.6685	1.515	0.9753
	TNM stage				
	I	1 (ref)			
	II	1.891	0.6293	5.68	0.2565
	III	3.186	1.101	9.22	0.03253
	IV	13.55	4.779	38.42	9.51E-07
	Risk score	5.987	3.086	11.61	1.2E-07

Age, gender and risk score were treated as continuous variables. TNM stage was evaluated as ordered categorical variable (stage I = reference).

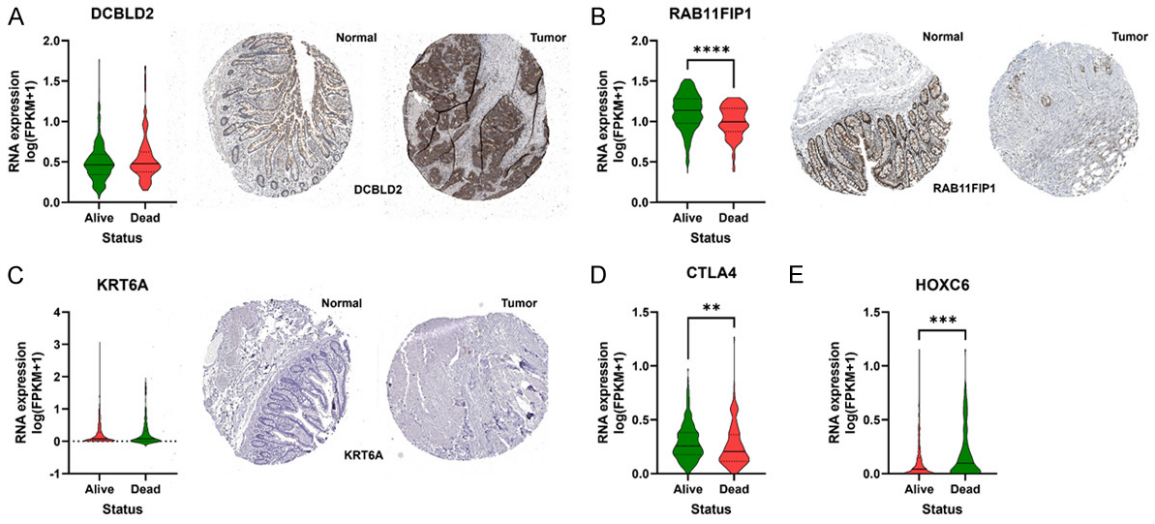


Figure 4. The mRNA and protein expression of the risk predicting genes in colon tissues. Data and images were obtained from the HPA database for risk genes including (A) *DCBLD2*, (B) *RAB11FIP1*, (C) *KRT6A*, (D) *CTLA4*, and (E) *HOXC6*. Three genes, except for *CTLA4* and *HOXC6*, have available pathology images.

diabetes of the young, cardiac muscle contraction, adipocytokine signaling pathway, and vasopressin regulated water reabsorption, while the main pathways enriched in the low-risk group were intestinal immune network for IgA production, allograft rejection, primary immunodeficiency, asthma, and graft versus host disease (Figure S3A). In GO enrichment analysis, the most enriched terms were related to extracellular matrix organization (biological process) (Figure S3B), collagen-containing extracellular matrix (cellular component) (Figure S3C), and extracellular matrix structural constituent (molecular function) (Figure S3D).

WGCNA analysis of high-risk and low-risk groups

To establish the key module associated with the high- and low-risk groups, WGCNA was performed using the TCGA expression profile, and 29 modules were identified (Figure S4A). Notably, compared with other modules, the black module was remarkably correlated with risk signature (correlation coefficient = 0.19, *P* value = 2E-04) (Figure S4B, S4C). Therefore, five genes in this black module were selected as hub genes (*PLAGL2*, *POFUT1*, *FITM2*, *TP53RK*, and *MOCS3*) using the cutoff value of GS >0.2 and MM >0.8. These hub genes were significantly correlated with our risk score genes (Figure S4D).

Construction of a prognosis predictive nomogram

We constructed a nomogram for predicting the prognosis of patients with colon cancer by integrating the risk scores with clinical factors including age, gender, and TNM stage by using TCGA database. A model with C-index of 0.718 (95% CI: 0.662-0.774) was established (**Figure 5A**), and the calibrations of nomogram were shown in **Figure 5B-D**. Furthermore, a predicted score was calculated for each patient according to the nomogram equation. Time-dependent ROC curve showed that the AUCs of nomogram prediction signature were 0.80, 0.75, and 0.70 for one-, three- and five-year survival, respectively (**Figure 5E**). Importantly, the decision curve analysis implicated a higher net benefit of our model compared to either the treat-all-patients schemes or the treat-none scheme in predicting the one-, three- or five-year survival (**Figure 5F**).

Mutation landscape between the high- and low-risk groups

The significantly mutated genes between the high- and low-risk groups were plotted in Figure S5A. Via the threshold *P* value of 0.01 and 95% CI not intersecting 1, 36 significantly mutated genes were identified between the high- and low-risk groups, of which four genes, *MAP2K7*, *AMPD2*, *KMT2B* and *KRAS*, were significantly enriched in the high-risk group ($P<0.01$) (Figure S5B), while the other 32 genes were significantly enriched in the low-risk group (Figure S5B). The stratified plots of the 36 genes by risk groups were shown in Figure S5C.

Immune signature associated with the risk score model

Since immune cell infiltration in the tumor microenvironment significantly affects the prognosis, we analyzed the 28 immune infiltrating subpopulations in the high- and low-risk groups with ssGSEA in TCGA database. As illustrated by heatmap in **Figure 6A**, several anti-tumor immune cells (activated $CD4^+$ T cell, activated $CD8^+$ T cell, activated dendritic cell, CD56 bright natural killer cell, natural killer T cell, and type 17 T helper cell) and pro-tumor immune cells (macrophage, plasmacytoid dendritic cell, regulatory T cell, and type 2 T helper cell) were statistically differentially enriched between the high- and low-risk groups ($P<0.05$) (**Figure 6B**).

Prediction of immunotherapeutic response

In recent years, the development of immune checkpoint inhibitors, such as anti-PD-1 or anti-PD-L1, has significantly improved the therapeutic effect and prolonged the survival time of cancer patients. However, the different patient response to immunotherapy limits its application. Therefore, in this study, we investigated the predictive value of our risk signature to immunotherapy. We reviewed the expression matrix and clinical characteristics of an anti-PD-L1 dataset for urothelial cancer (IMvigor210) published in *Nature* in 2018. By using the same risk formula and then the median score as cut-off value, we divided the patients in IMvigor210 cohort into high- and low-risk groups. Kaplan-Meier analysis suggested that patients with high-risk score had significantly shorter survival time than those with low-risk score ($P<0.05$) (**Figure 7A**). The violin plot further illustrated

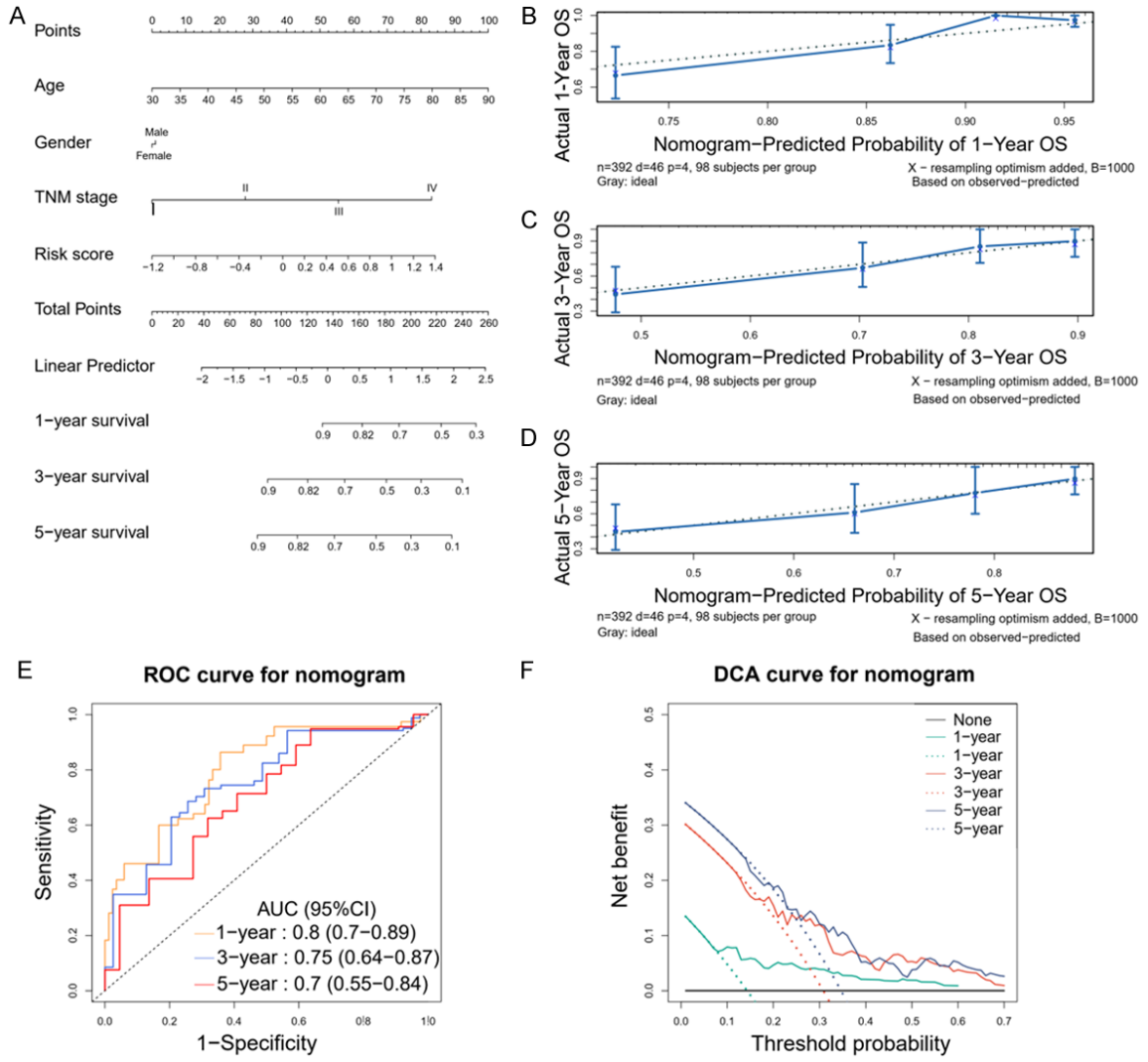


Figure 5. Nomogram predicting the prognosis of patients in the TCGA cohort. A. Constructed nomogram containing risk scores. B. Calibration of nomogram at one-year. C. Calibration of nomogram at three-year. D. Calibration of nomogram at five-year. E. Time-dependent ROC of the nomogram for one-, three-, and five-year survival. F. Decision curve analysis of the nomogram for one-, three-, and five-year risk. Black line indicates assumption that no patients die.

that the patients in the high-risk group had much lower mutation burden than those in the low-risk group (*P* value = 0.021) (**Figure 7B**). The neoantigen in the high-risk group was also significantly lowered than that in the low-risk group (**Figure 7C**). We further determined the immune response to anti-PD-L1 therapy. According to the response to neoadjuvant chemotherapy, the patients were categorized into four groups: progressive disease (PD), stable disease (SD), partial response (PR) and complete response (CR). The group of SD/PD exhibited higher risk scores compared with CR/PR group, though not statistically significant (**Figure 7E**). There was an increasing trend of risk score across the PR, SD, and PD group (*P*<0.05 for PD vs. PR) (**Figure 7F**). The high-risk group had higher percentage of SD/PD than the low-risk group, though not statistically significant (**Figure 7G**). After adjusting for gender, platinum treatment, tobacco history, baseline ECOG score and immune phenotype, the risk score remained statistically significant with respect to overall survival (HR = 1.73, 95% CI = 1.19-2.52, *P* = 0.004) (**Figure 7D**). The distribution of immune response across risk scores was shown in the waterfall plots (**Figure 7H**). The time-dependent AUC of ROC curve for the

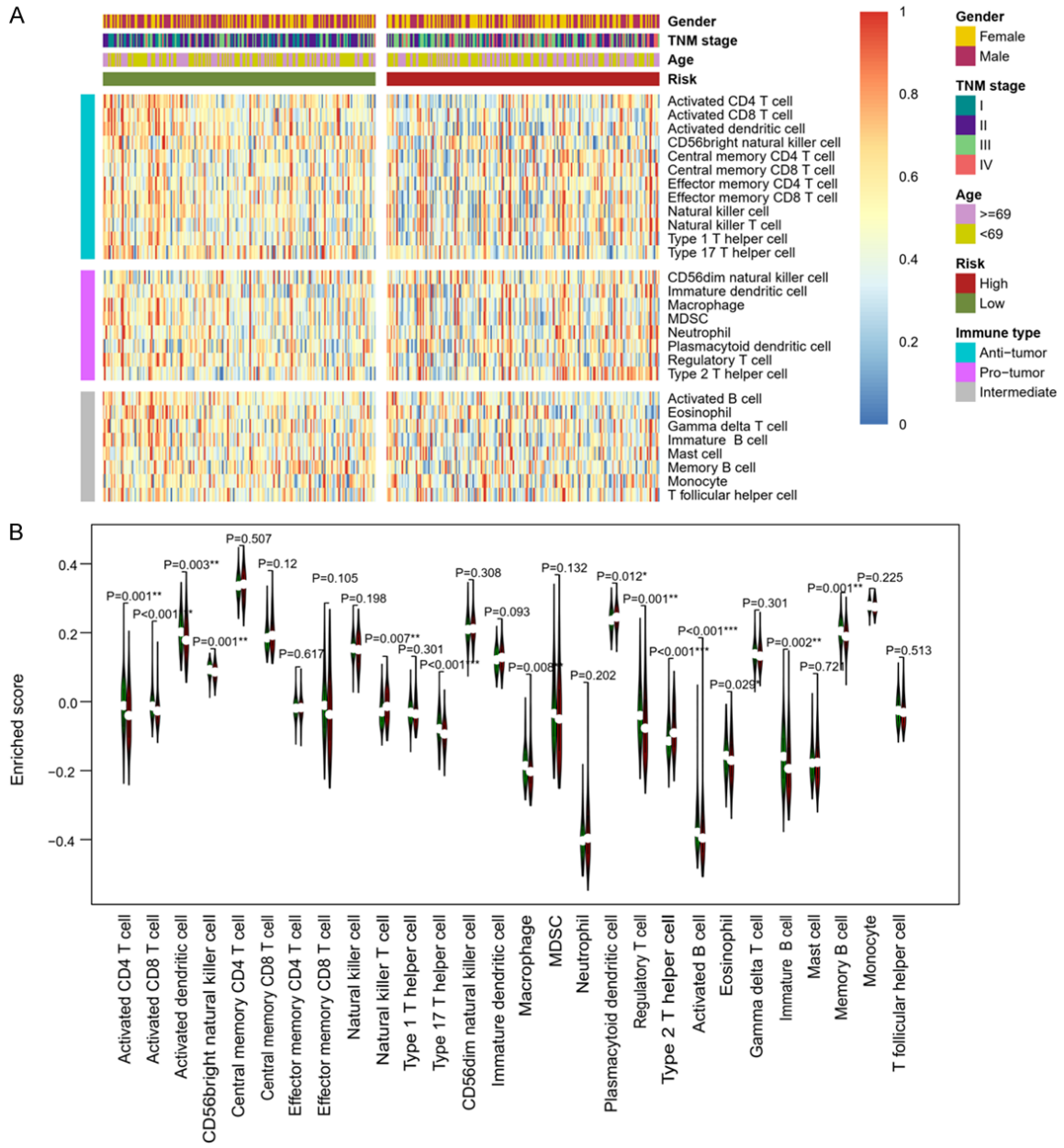


Figure 6. Immune cell infiltration analysis of the high- and low-risk signature. A. Heatmap representation of 28 immune infiltrating cell enrichment analysis. B. Comparison of 28 immune infiltrating cells between high- (red) and low-risk group (green). **P*<0.05, ***P*<0.01, ****P*<0.001.

risk signature in the cohort was presented in **Figure 7I**.

Construction of immune feature-related risk signature by extreme gradient boosting (XGBoost) algorithm

Extreme gradient boosting (XGBoost) is one of the cutting-edge machine learning algorithms under the gradient boosting framework that can solve big data issues in a highly efficient and accurate way. In principle, XGBoost improves the classification accuracy through iteratively optimizing the customized objective function step-by-step. To stratify the immune features that can predict the high- and low- risk status for colon cancer, we took advantage of the XGBoost algorithm to build a model with enriched score of 28 immune infiltrating subpopulations in the TCGA cohort. The cohort was

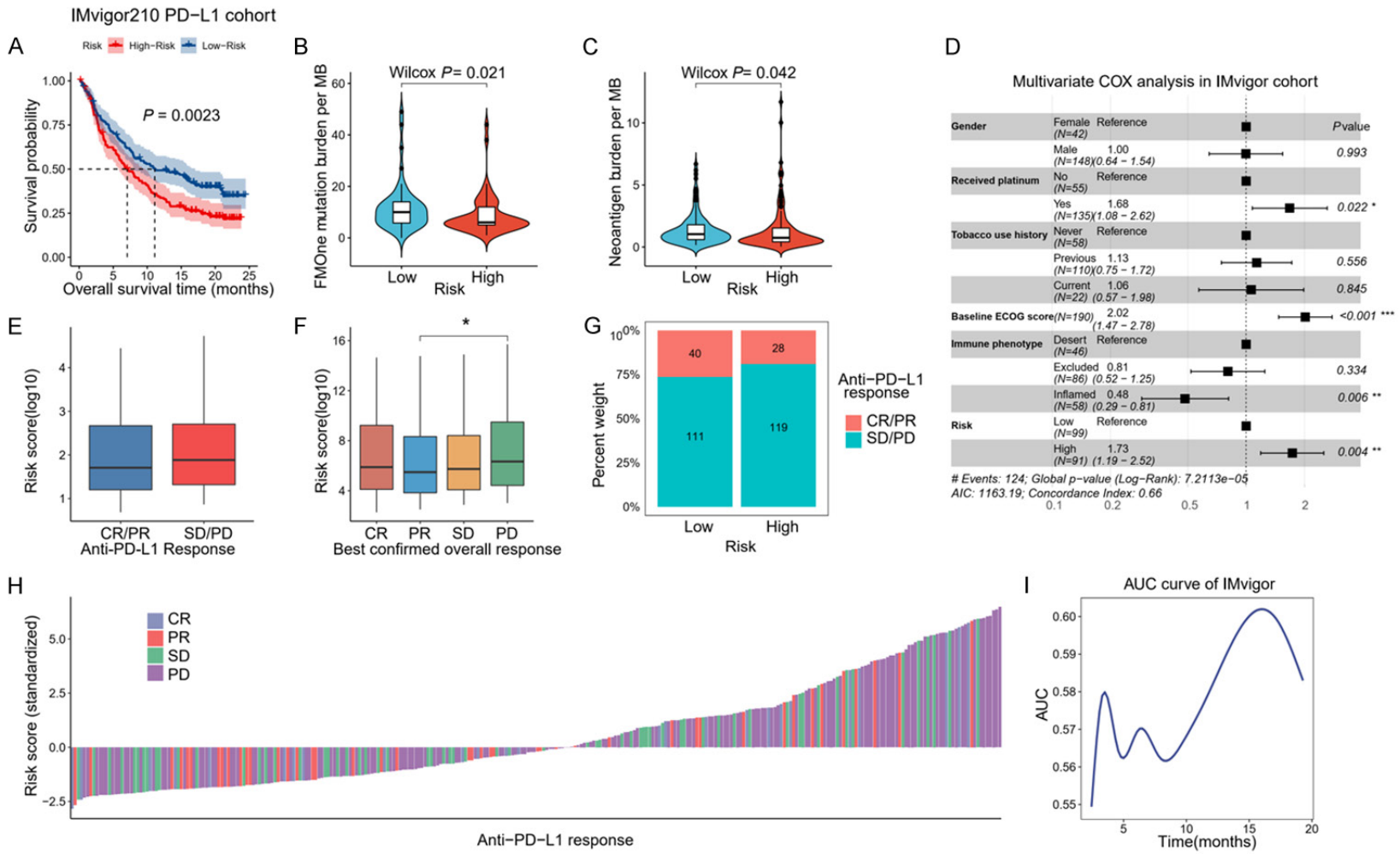


Figure 7. Immunotherapeutic response of the risk signature in IMvigor210 cohort. A. Kaplan-Meier analysis of risk signature. B. Violin plot of mutation burden. C. Violin plot of neoantigen burden. D. Forest plot of multivariate Cox analysis. E. Distribution of risk scores for patients with CR/PR vs. SD/PD. F. Risk score across CR, PR, SD, and PD response. G. The number of immunotherapy response in the high- and low-risk group. H. Waterfall plot of risk score and different immunotherapy responses. I. Time-dependent area under the curve (AUC) of receiver operating characteristic (ROC) analysis. Abbreviations: PD: Progressive Disease; SD: Stable Disease; PR: Partial Response; CR: Complete Response.

randomly equally divided into training and testing subgroups. In the final output, we selected a ten-feature model to predict the risk status. The Shapley additive explanation (SHAP) contribution dependency plots of the ten features were illustrated in **Figure 8A**, while the important features were ranked in **Figure 8B**. The overall accuracy and AUC in the training cohort were 81.73% and 0.909, respectively (**Figure 8C**). Then, the testing cohort was utilized to validate the performance of the XGBoost model. The AUC for the testing cohort was 0.717 (**Figure 8D**) with an accuracy of 63.45%. For the entire TCGA cohort, the accuracy and AUC were 72.59% and 0.814 (**Figure 8E**).

Overexpression and knockdown of risk genes in HCT116 cells

To study the function of each gene in the risk signature on colon cancer, we manipulated the gene expression level in HCT116 cells by transfection with siRNA or overexpression plasmid of each gene. qPCR and western blot analyses were first carried out to confirm the successful knockdown or overexpression of the target gene. The results showed that mRNA or protein level of the risk genes was significantly reduced after transfection with siRNAs (Figure S6A-J), while increased after transfection with overexpression plasmids (Figure S6K-T).

Effects of risk genes on the proliferation, migration, and invasion of colon cancer cells

After successful gene knockdown or overexpression as shown above, we investigated the effects of risk genes on the proliferation and viability of HCT116 cells by using CCK-8 assays. The results showed that, compared with siRNA vector controls, knockdown of *DCBLD2*, *HOXC6* or *KRT6A* by siRNAs significantly decreased (**Figure 9A**, **9J**, **9M**), while knockdown of *RAB11FIP1* or *CTLA4* significantly increased the cell proliferation (**Figure 9D**, **9G**). In the reciprocal experiment, overexpression of *DCBLD2*, *HOXC6* or *KRT6A* significantly enhanced (**Figure 9A**, **9J**, **9M**), while overexpression of *RAB11FIP1* or *CTLA4* clearly suppressed the cell viability (**Figure 9D**, **9G**). Wound healing assays showed that the wound closure was delayed after *DCBLD2*, *HOXC6* or *KRT6A* knockdown (**Figure 9B**, **9K**, **9N**), but was accelerated after *RAB11FIP1* or *CTLA4* knockdown (**Figure 9E**, **9H**). Consistently, overexpression of these genes exhibited the reversed effects. These results were further supported by transwell assays where the cell invasion and migration ability was dampened by *DCBLD2*, *HOXC6* or *KRT6A* siRNA transfection (**Figure 9C**, **9L**, **9O**) and increased by *RAB11FIP1* or *CTLA4* siRNA transfection (**Figure 9F**, **9I**). Conversely, overexpression of *DCBLD2*, *HOXC6* or *KRT6A* promoted while overexpression of *RAB11FIP1* or *CTLA4* decreased the invasion and migration of HCT116 cells.

Discussion

The etiology and pathophysiologic process underlying colon cancer development remain elusive, although current understandings indicate that it is a multifactorial disease encompassing genetic factors, environmental exposures, epigenetic alternations, and immune dysregulation [20]. Traditional predictive tools for prognosis involve age, gender, or TNM staging, which has tremendous limitations because of the molecular heterogeneity of cancer [21, 22]. As a result, molecular prognostic markers are being widely investigated. A prominent example of such markers, published in *Lancet*, derives from the consensus of immunoscore® as a reliable prediction system for colon cancer recurrence complementary to the TNM classification system [23]. The immune response is found to be a key determinator for the prognosis.

Here, we identified a five-gene signature, consisting of *DCBLD2*, *RAB11FIP1*, *CTLA4*, *HOXC6* and *KRT6A*, that could predict the prognosis of colon cancer patients. This signature was successfully validated in two independent GEO datasets, demonstrating its reliability and accuracy. In univariate Cox or multivariate Cox regression analysis, the gene signature still exhibited statistical significance after adjusting for age, gender or TNM stage, suggesting its independence from other clinical variables. The mRNA or protein expression level of each gene was also consistent with the prediction trend in the signature. We then functionally annotated the signature and found that they were mainly enriched in extracellular matrix organization, collagen-containing extracellular matrix, and extracellular matrix structural constituent terms. WGCNA generated five hub genes (*PLAGL2*, *POFUT1*, *FITM2*, *TP53RK*, and *MOCS3)* associated with the signature genes. To facili-

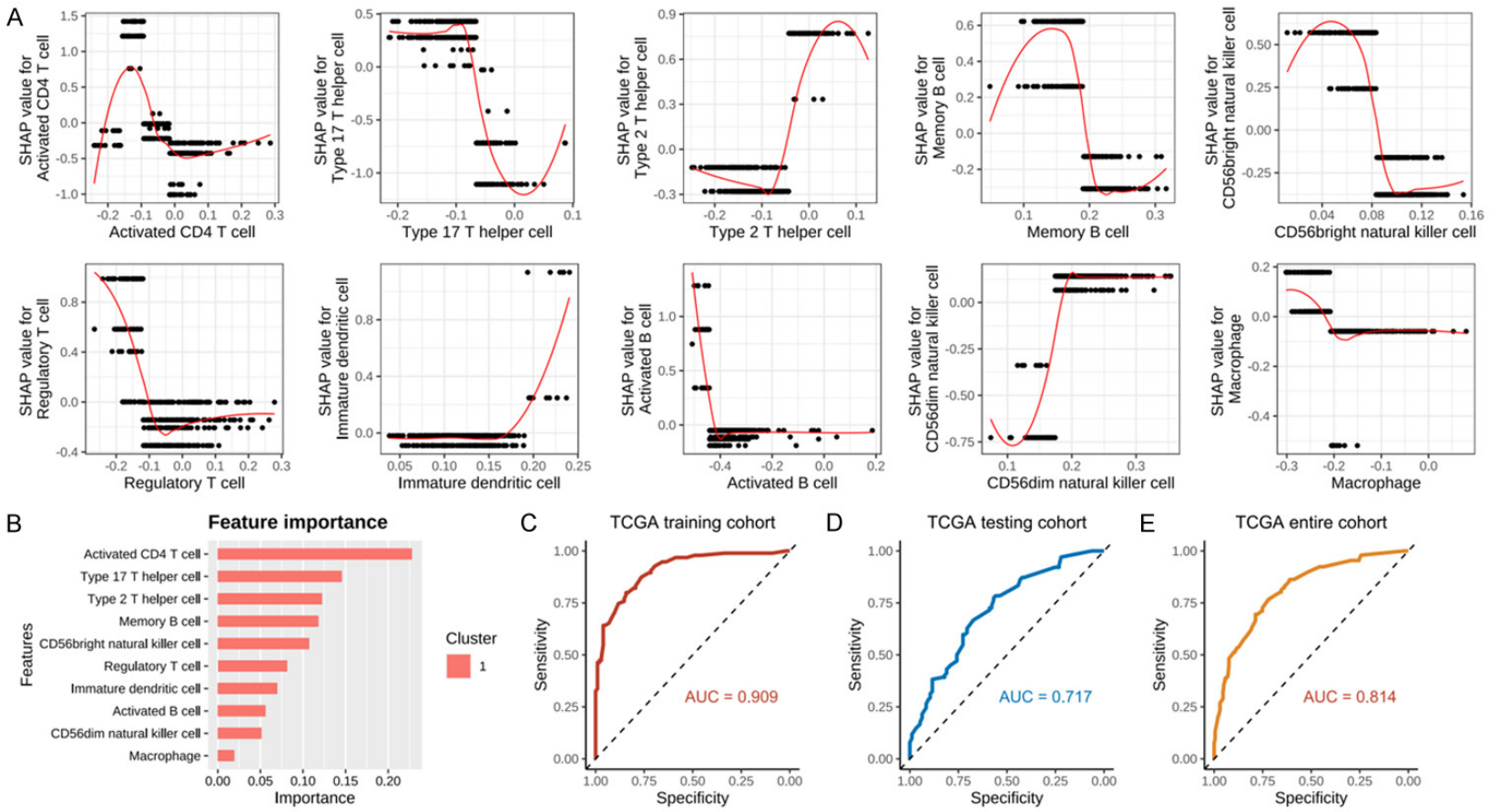


Figure 8. XGBoost prediction model for risk status. A. The Shapley additive explanation (SHAP) visualization plots for XGBoost. B. Ranking of ten features by importance. C. Receiver operating characteristic (ROC) curve of the TCGA training cohort. D. ROC curve of the TCGA testing cohort. E. ROC curve of the entire TCGA cohort.

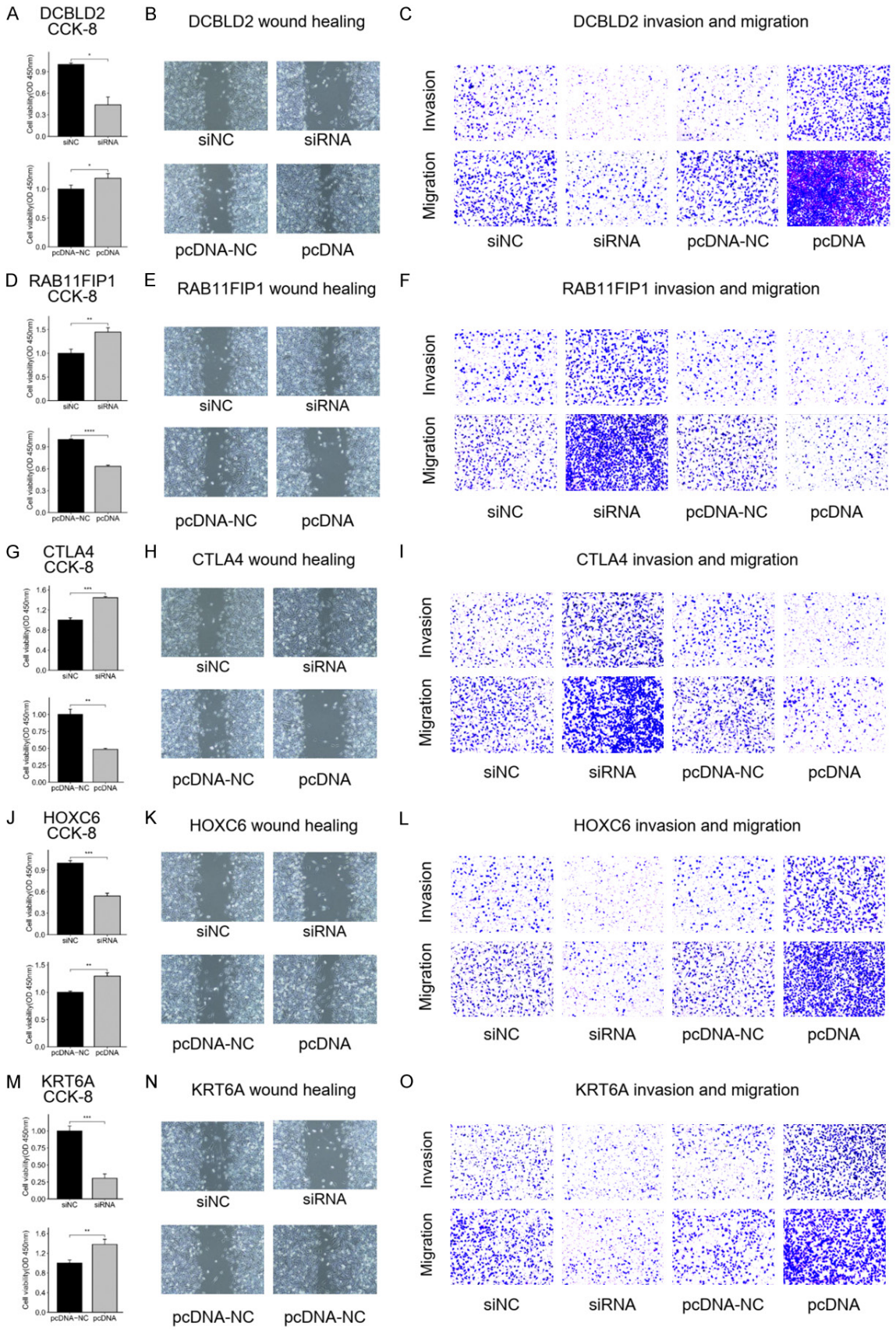

A
DCBLD2
CCK-8
B
DCBLD2 wound healing
C
DCBLD2 invasion and migration
siNC
siRNA
pcDNA-NC
pcDNA
Invasion
Migration
D
RAB11FIP1
CCK-8
E
RAB11FIP1 wound healing
F
RAB11FIP1 invasion and migration
G
CTLA4
CCK-8
H
CTLA4 wound healing
I
CTLA4 invasion and migration
J
HOXC6
CCK-8
K
HOXC6 wound healing
L
HOXC6 invasion and migration
M
KRT6A
CCK-8
N
KRT6A wound healing
O
KRT6A invasion and migration

Figure 9. CCK-8, wound healing, and transwell invasion and migration assays. Cell viability, invasion and migration experiments in HCT116 cells were conducted to evaluate the biological function of five risk genes. (A-C) CCK-8, wound healing, and transwell analysis of *DCBLD2*, and the corresponding analysis on (D-F) *RAB11FIP1*, (G-I) *CTLA4*, (J-L) *HOXC6*, and (M-O) *KRT6A*. **P*<0.05, ***P*<0.01, ****P*<0.001, *****P*<0.0001.

tate the use of this signature in clinical practice, we constructed a nomogram incorporating common clinical information and the gene risk score for prognosis prediction. Furthermore, mutation landscape profiling revealed that *KRAS* was the most frequently mutated genes (44%) significantly affected by the risk signature, followed by *KMT2B* (13%) and *PCDHGA2* (6%). Immune infiltration analysis indicated that myriad types of immune cells were involved in the functional modality in the high- and low-risk groups, including activated $CD4^+$ T cell, activated $CD8^+$ T cell, activated dendritic cell, CD56 bright natural killer cell, natural killer T cell, type 17 T helper cell, macrophage, plasmacytoid dendritic cell, regulatory T cell, and type 2 T helper cell. Importantly, the risk signature might be related to the immunotherapeutic benefits and prognosis, as evidenced by the results from anti-PD-L1 cohort analysis (PR vs PD $P<0.05$; Kaplan-Meier $P<0.05$). Moreover, in this study, XGBoost algorithm was applied to construct a predictive model using the immune infiltrating subpopulation features, which demonstrated desirable accuracy (72.59%) and AUC (0.814). Lastly, we conducted in vitro experiments to verify the function of the risk genes and found that *DCBLD2*, *HOXC6* and *KRT6A* promoted, while *RAB11FIP1* and *CTLA4* suppressed, the proliferation, migration, and invasion of colon cancer cells.

Our in vitro validation results were consistent with previous reports. For example, Xie et al. [24] investigated the function of *DCBLD2* by knocking down its expression via siRNA and found that downregulation of *DCBLD2* could inhibit the proliferation, migration, and invasion of colorectal cancer cells. Similarly, He et al. [25] also found that *DCBLD2* downregulation reduced colorectal cancer cell proliferation and invasion in vitro. *HOXC6*, another gene in the risk signature, has been reported to play an oncogenic role in colon cancer as *HOXC6* overexpression promoted the migration and invasion of HCT116 and RKO cells by activating Wnt/β-catenin pathway [26], while its downregulation inhibited cell viability and colony formation through mTOR pathway [27]. As for *KRT6A* gene, although there has been no in vitro study reported in colon cells, its oncogenic effect was observed in nasopharyngeal carcinoma [28] and non-small cell lung cancer [29]. In contrast, the other three genes in the risk signature exhibited opposite effects on malignancy. For example, *RAB11FIP1* gene has been reported to attenuate tumor progression in an ErbB2 dependent manner, indicating its tumor suppressing role [30]. Similarly, *CTLA4* has been found to be constitutively expressed in various tumor cells, and treatment of *CTLA4*-expressing cells with soluble recombinant ligands, r-CD80 and rCD86, induced apoptosis signals [31]. Consistent with these findings, higher expression of *CTLA4* was associated with less advanced TNM stage and well/moderately differentiated gastric adenocarcinoma [32]. In our study, we also found higher expression of *CTLA4* might be associated with favorable overall survival, evidenced by the negative coefficient in our risk model. Nevertheless, our finding conflicts with the widely accepted notion that *CTLA4* overexpression in T cells would contribute to a worse prognosis due to *CTLA4*-mediated downregulation of T cell activation. The possible explanation for this discrepancy is *CTLA4* expression in different target cells, i.e., tumor cells vs T cells, causes distinct phenotypes. The effect of *CTLA4* on the proliferation and apoptosis of tumor cells is dramatically different to those of T cells. *CTLA4* positive tumor cells may transduce negative extracellular signals into cells by interacting with ligands in the microenvironment, thus leading to the inhibition of cell proliferation or induction of apoptosis [31]. In contrast, *CTLA4* upregulation in conventional T cells after activation suppresses immune response, thereby leading to tumor growth and poor overall survival [33]. Our findings and hypothesis are in accordance with previous reports that *CTLA4*-overexpressing non-small cell lung cancer patients had a better prognosis compared to patients with low *CTLA4* expression [34]. In this regard, our current findings support the notion that *CTLA4* may serve as a potential anti-tumor intervention target in tumor cells, in addition to its classical immune roles in T cells.

Efforts have been undertaken to predict the prognosis of colon cancer through screening and establishing molecular signatures. For example, Liu et al. [15] established a two-gene model using TCGA data. Similarly using TCGA data, Xu et al. [16] generated a hypoxia-related three-gene signature for prognostic prediction. In addition, Sun et al. [17] established an immune-related gene model, and Zou et al. [18] built a seven-gene model to predict immunotherapy response. Furthermore, Yang et al. [19] constructed a nine-gene model to improve the prediction of prognosis and drug sensitivity. To demonstrate the superiority of our signature over the published signatures, we evaluated the performance of these models in both GSE39582 and GSE17538 datasets. The results showed that our five-gene signature had better predicative ability and higher AUC values. One of the possible reasons that the previously published signatures failed to perform well during external validation in GSE is because of the existence of high heterogeneity among different cancer cohorts [35]. Since our signature was successfully validated in external datasets, we speculated that the gene combination we used had overcome the heterogeneity in different cohorts, especially in anti-PD L1 cohort. Another possible explanation for the unsatisfactory performance of the reported signature might be common statistical cognition: the use of dichotomous threshold of *P* value of 0.05 as either insignificance or significance [36]. Although it has been widely accepted that *P* values less than 0.05 are statistically significant, many scholars have argued that "scientists rise up against statistical significance" and have raised the concern that the common threshold of statistical significance may be misleading due to human or cognitive restriction [36].

To our knowledge, this study was the first to combine the five genes as a signature to predict the prognosis of colon cancer. Although many studies have focused on oncogenes or tumor suppressor genes in colon cancer, most of them mainly adopted single gene strategy to study the gene function individually. Integrating different genes is expected to increase prognostic value since the combined signature is more likely in line with the multifactorial nature of cancer. Additionally, it is well known that the prognosis of colon cancer is highly related to TNM stage [37]; nevertheless, our risk signature was independent of TNM stage I, II or III, but not TNM IV. Therefore, we constructed a nomogram including TNM stage to facilitate the clinical use of our model. Our study has several advantages. First, the AUCs of the nomogram were 0.8, 0.75, and 0.7 for one-, three-, and five-year survival, respectively, demonstrating a better performance compared to the model reported by Han et al. [38], in which 17 clinical variables were extracted from SEER database to construct a survival-related nomogram for colorectal cancer, and the one-, three-, and five-year AUCs were 0.705, 0.675, and 0.648, respectively. Clearly, our nomogram has the advantage of remarkably larger AUC and fewer variables, indicating a higher predictive power. Second, statistical significance was observed in predicting several different survivals, including overall survival, recurrence free survival, disease free survival, and disease specific survival, across the three independent datasets we examined, which was scarcely reported previously. The statistical significance in overall survival was difficult to repeat in disease free survival or recurrence free survival within one cohort. It was even more difficult to repeat them in different cohorts, suggesting that our risk model was strongly correlated with tumor progression. The strong correlation would exceed the effects caused by inherent biases, e.g., tumor cell heterogeneity, distinctive sequencing platforms or normalization methods, that can influence the statistical significance. Nevertheless, the mechanisms underlying the function of our gene signature and patient survival require further investigation.

Although our study generated a reliable prognostic signature in colon cancer, there are some limitations in this study. First, the sample size of TCGA and GEO we used to generate the signature may not be sufficient to represent colon cancer. Thus, cohorts with larger sample sizes are warranted to further demonstrate its reliability and specificity. Second, the gene signature was selected mainly on the basis of mRNA expression data, thus, integrating biological information, such as protein or miRNA expression data, will enhance the accuracy of the model. Lastly, the mechanisms of the signature in immunotherapy are still unclear. Further studies are needed to elucidate the molecular mechanism underlying the role of the signature in immunotherapeutic effect.

Conclusion

In summary, we demonstrated the effective prediction of the clinical outcomes of colon cancer patients by using a five-gene panel. This five-gene signature could serve as a potential biomarker for the prognosis of colon cancer. Extensive clinical testing and basic research are required to validate our findings.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 81900499, 81860100, 82060525); National Key Research and Development Projects (Grant No. 2017YFC1309200); Yunling Scholar (Grant No. YLXL20170002); Yunnan province science and Technology Hall Youth Academic and technical leader Reserve Talents Project (Grant No. 202005AC160057); Reserve talents of high-level health technical talents in Yunnan Province (Grant No. H-2018062); Yunnan Province "high level talent training support plan" training plan-special program for young top talent (Grant No. YNWR-QNBJ-2019-243); Yunnan high level talent training support plan-"famous doctors" special project (Grant No. RLMY20-200019).

Disclosure of conflict of interest

None.

Address correspondence to: Hanping Shi, Department of General Surgery, Beijing Shijitan Hospital, Capital Medical University, No. 10th Tieyi Road, Haidian District, Beijing 100038, The People's Republic of China. Tel: +86-13802741263; Fax: +86-010-63926519; E-mail: shihp@ccmu.edu.cn; Bo Li, Department of General Surgery, The Affiliated Hospital of Yunnan University, No. 176 Qingnian Road, Wuhua District, Kunming 650091, Yunnan, The People's Republic of China. Tel: +86-137-08472035; E-mail: 2296591440@qq.com

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA and Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018; 68: 394-424.

[2] Siegel RL, Miller KD, Fuchs HE and Jemal A. Cancer statistics, 2021. CA Cancer J Clin 2021; 71: 7-33.

[3] Argilés G, Tabernero J, Labianca R, Hochhauser D, Salazar R, Iveson T, Laurent-Puig P, Quirke P, Yoshino T, Taieb J, Martinelli E and Arnold D; ESMO Guidelines Committee. Electronic address: clinicalguidelines@esmo.org. Localised colon cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann Oncol 2020; 31: 1291-1305.

[4] Abukar AA, Ramsanahie A, Martin-Lumbard K, Herrington ER, Winslow V, Wong S, Ahmed S and Thaha MA. Availability and feasibility of structured, routine collection of comorbidity data in a colorectal cancer multi-disciplinary team (MDT) setting. Int J Colorectal Dis 2018; 33: 1057-1061.

[5] Cao W, Chen HD, Yu YW, Li N and Chen WQ. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. Chin Med J (Engl) 2021; 134: 783-791.

[6] Song M, Chan AT and Sun J. Influence of the gut microbiome, diet, and environment on risk of colorectal cancer. Gastroenterology 2020; 158: 322-340.

[7] Dekker E, Tanis PJ, Vleugels JLA, Kasi PM and Wallace MB. Colorectal cancer. Lancet 2019; 394: 1467-1480.

[8] Fontana E, Eason K, Cervantes A, Salazar R and Sadanandam A. Context matters-consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials. Ann Oncol 2019; 30: 520-527.

[9] Billan S, Kaidar-Person O and Gil Z. Treatment after progression in the era of immunotherapy. Lancet Oncol 2020; 21: e463-e476.

[10] Tolba MF. Revolutionizing the landscape of colorectal cancer treatment: the potential role of immune checkpoint inhibitors. Int J Cancer 2020; 147: 2996-3006.

[11] Punt CJ, Koopman M and Vermeulen L. From tumour heterogeneity to advances in precision treatment of colorectal cancer. Nat Rev Clin Oncol 2017; 14: 235-246.

[12] Mariathasan S, Turley SJ, Nickles D, Castiglioni A, Yuen K, Wang Y, Kadel EE III, Koeppen H, Astarita JL, Cubas R, Jhunjhunwala S, Banchereau R, Yang Y, Guan Y, Chalouni C, Ziai J, Şenbabaoğlu Y, Santoro S, Sheinson D, Hung J, Giltnane JM, Pierce AA, Mesh K, Lianoglou S, Riegler J, Carano RAD, Eriksson P, Höglund M, Somarriba L, Halligan DL, van der Heijden MS, Loriot Y, Rosenberg JE, Fong L, Mellman I, Chen DS, Green M, Derleth C, Fine GD, Hegde PS, Bourgon R and Powles T. TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. Nature 2018; 554: 544-548.

[13] Sullivan LM, Massaro JM and D'Agostino RB Sr. Presentation of multivariate data for clinical

use: the Framingham study risk score functions. Stat Med 2004; 23: 1631-1660.
[14] Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, Hackl H and Trajanoski Z. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep 2017; 18: 248-262.
[15] Liu C, Liu D, Wang F, Liu Y, Xie J, Xie J and Xie Y. Construction of a novel choline metabolism-related signature to predict prognosis, immune landscape, and chemotherapy response in colon adenocarcinoma. Front Immunol 2022; 13: 1038927.
[16] Xu Y, Cao C, Zhu Z, Wang Y, Tan Y and Xu X. Novel hypoxia-associated gene signature depicts tumor immune microenvironment and predicts prognosis of colon cancer patients. Front Genet 2022; 13: 901734.
[17] Sun Y, Zhang Y, Guo Y, Yang Z and Xu Y. A prognostic model based on the immune-related genes in colon adenocarcinoma. Int J Med Sci 2020; 17: 1879-1896.
[18] Zou Z, Chai Y, Li Q, Lin X, He Q and Xiong Q. Establishment of lactate-metabolism-related signature to predict prognosis and immunotherapy response in patients with colon adenocarcinoma. Front Oncol 2022; 12: 958221.
[19] Yang Y, Feng M, Bai L, Liao W, Zhou K, Zhang M, Wu Q, Wen F, Lei W, Zhang P, Zhang N, Huang J and Li Q. Comprehensive analysis of EMT-related genes and lncRNAs in the prognosis, immunity, and drug treatment of colorectal cancer. J Transl Med 2021; 19: 391.
[20] Burnett-Hartman AN, Lee JK, Demb J and Gupta S. An update on the epidemiology, molecular characterization, diagnosis, and screening strategies for early-onset colorectal cancer. Gastroenterology 2021; 160: 1041-1049.
[21] Dienstmann R, Mason MJ, Sinicrope FA, Phipps AI, Tejpar S, Nesbakken A, Danielsen SA, Sveen A, Buchanan DD, Clendenning M, Rosty C, Bot B, Alberts SR, Milburn Jessup J, Lothe RA, Delorenzi M, Newcomb PA, Sargent D and Guinney J. Prediction of overall survival in stage II and III colon cancer beyond TNM system: a retrospective, pooled biomarker study. Ann Oncol 2017; 28: 1023-1031.
[22] Gaiani F, Marchesi F, Negri F, Greco L, Malesci A, de'Angelis GL and Laghi L. Heterogeneity of colorectal cancer progression: molecular gas and brakes. Int J Mol Sci 2021; 22: 5246.
[23] Pagès F, Mlecnik B, Marliot F, Bindea G, Ou FS, Bifulco C, Lugli A, Zlobec I, Rau TT, Berger MD, Nagtegaal ID, Vink-Börger E, Hartmann A, Geppert C, Kolwelter J, Merkel S, Grützmann R, Van den Eynde M, Jouret-Mourin A, Kartheuser A, Léonard D, Remue C, Wang JY, Bavi P, Roehrl MHA, Ohashi PS, Nguyen LT, Han S, MacGregor HL, Hafezi-Bakhtiari S, Wouters BG, Masucci GV, Andersson EK, Zavadova E, Vocka M, Spacek J, Petruzelka L, Konopasek B, Dundr P, Skalova H, Nemejcova K, Botti G, Tatangelo F, Delrio P, Ciliberto G, Maio M, Laghi L, Grizzi F, Fredriksen T, Buttard B, Angelova M, Vasaturo A, Maby P, Church SE, Angell HK, Lafontaine L, Bruni D, El Sissy C, Haicheur N, Kirilovsky A, Berger A, Lagorce C, Meyers JP, Paustian C, Feng Z, Ballesteros-Merino C, Dijkstra J, van de Water C, van Lent-van Vliet S, Knijn N, Mu□ină AM, Scripcariu DV, Popivanova B, Xu M, Fujita T, Hazama S, Suzuki N, Nagano H, Okuno K, Torigoe T, Sato N, Furuhata T, Takemasa I, Itoh K, Patel PS, Vora HH, Shah B, Patel JB, Rajvik KN, Pandya SJ, Shukla SN, Wang Y, Zhang G, Kawakami Y, Marincola FM, Ascierto PA, Sargent DJ, Fox BA and Galon J. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. Lancet 2018; 391: 2128-2139.
[24] Xie P, Yuan FQ, Huang MS, Zhang W, Zhou HH, Li X and Liu ZQ. DCBLD2 affects the development of colorectal cancer via EMT and angiogenesis and modulates 5-FU drug resistance. Front Cell Dev Biol 2021; 9: 669285.
[25] He J, Huang H, Du Y, Peng D, Zhou Y, Li Y, Wang H, Zhou Y and Nie Y. Association of DCBLD2 upregulation with tumor progression and poor survival in colorectal cancer. Cell Oncol (Dordr) 2020; 43: 409-420.
[26] Qi L, Chen J, Zhou B, Xu K, Wang K, Fang Z and Shao Y. HomeoboxC6 promotes metastasis by orchestrating the DKK1/Wnt/β-catenin axis in right-sided colon cancer. Cell Death Dis 2021; 12: 337.
[27] Ji M, Feng Q, He G, Yang L, Tang W, Lao X, Zhu D, Lin Q, Xu P, Wei Y and Xu J. Silencing homeobox C6 inhibits colorectal cancer cell proliferation. Oncotarget 2016; 7: 29216-29227.
[28] Chen C and Shan H. Keratin 6A gene silencing suppresses cell invasion and metastasis of nasopharyngeal carcinoma via the β-catenin cascade. Mol Med Rep 2019; 19: 3477-3484.
[29] Che D, Wang M, Sun J, Li B, Xu T, Lu Y, Pan H, Lu Z and Gu X. KRT6A promotes lung cancer cell growth and invasion through MYC-regulated pentose phosphate pathway. Front Cell Dev Biol 2021; 9: 694071.
[30] Boulay PL, Mitchell L, Turpin J, Huot-Marchand JÉ, Lavoie C, Sanguin-Gendreau V, Jones L, Mitra S, Livingstone JM, Campbell S, Hallett M, Mills GB, Park M, Chodosh L, Strathdee D, Norman JC and Muller WJ. Rab11-FIP1C Is a critical negative regulator in ErbB2-mediated mammary tumor progression. Cancer Res 2016; 76: 2662-2674.

[31] Contardi E, Palmisano GL, Tazzari PL, Martelli AM, Falà F, Fabbi M, Kato T, Lucarelli E, Donati D, Polito L, Bolognesi A, Ricci F, Salvi S, Gargaglione V, Mantero S, Alberghini M, Ferrara GB and Pistillo MP. CTLA-4 is constitutively expressed on tumor cells and can trigger apoptosis upon ligand interaction. Int J Cancer 2005; 117: 538-550.

[32] Kim JW, Nam KH, Ahn SH, Park DJ, Kim HH, Kim SH, Chang H, Lee JO, Kim YJ, Lee HS, Kim JH, Bang SM, Lee JS and Lee KW. Prognostic implications of immunosuppressive protein expression in tumors as well as immune cell infiltration within the tumor microenvironment in gastric cancer. Gastric Cancer 2016; 19: 42-52.

[33] Egen J, Ouyang W and Wu L. Human anti-tumor immunity: insights from immunotherapy clinical trials. Immunity 2020; 52: 36-54.

[34] Salvi S, Fontana V, Boccardo S, Merlo DF, Margallo E, Laurent S, Morabito A, Rijavec E, Dal Bello MG, Mora M, Ratto GB, Grossi F, Truini M and Pistillo MP. Evaluation of CTLA-4 expression and relevance as a novel prognostic factor in patients with non-small cell lung cancer. Cancer Immunol Immunother 2012; 61: 1463-1472.

[35] Papalexi E and Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol 2018; 18: 35-45.

[36] Amrhein V, Greenland S and McShane B. Scientists rise up against statistical significance. Nature 2019; 567: 305-307.

[37] Liu Q, Luo D, Cai S, Li Q and Li X. P-TNM staging system for colon cancer: combination of P-stage and AJCC TNM staging system for improving prognostic prediction and clinical management. Cancer Manag Res 2018; 10: 2303-2314.

[38] Han L, Dai W, Mo S, Xiang W, Li Q, Xu Y, Cai G and Wang R. Nomogram of conditional survival probability of long-term survival for metastatic colorectal cancer: a real-world data retrospective cohort study from SEER database. Int J Surg 2021; 92: 106013.

Table S1. siRNA sequences for the selected genes

siRNA		Sequences (5'-3')	
		Sense	Anti-sense
DCBLD2	si-1	GCAAGAGAACAGUUGGAAACCTT	GGUUUCCAACUGUUCUCUUGCTT
	si-2	CGGCCAAAUCAGUGUUGUAAUTT	AUUACAACACUGAUUUGGCCGTT
	si-3	GUGUGGAGCAAGAUAAGAUAUTT	AUAUCUUAUCUUGCUCCACACTT
RAB11FIP1	si-1	CGAUAAGCAAGAAGGAGUUTT	AACUCCUUCUUGCUUAUCGTT
	si-2	GGUUAAUGAUUACAAUUAATT	UUAAUUGUAAUCAUUAACCTT
	si-3	CGCACUCGCUAAUACAGUUTT	AACUGUAUUAGCGAGUGCGTT
CTLA4	si-1	CCCAAAUUACGUGUACUACTT	GUAGUACACGUAAUUUGGGTT
	si-2	GGUGGAGCUCAUGUACCCATT	UGGGUACAUGAGCUCCACCTT
	si-3	UGAGUUGACCUUCCUAGAUGATT	UCAUCUAGGAAGGUCAACUCATT
HOXC6	si-1	CUCGUUCUCGGCUUGUCUATT	UAGACAAGCCGAGAACGAGTT
	si-2	CCGUAUGACUAUGGAUCUATT	UAGAUCCAUAGUCAUACGGTT
	si-3	GCCAGAUCUACUCGCGGUATT	UACCGCGAGUAGAUCUGGCTT
KRT6A	si-1	CCAGCAGGAAGAGCUAUATT	UAUAGCUCUUCCUGCUGGTT
	si-2	GCAAGCUGCUGGAGGGUGATT	UCACCCUCCAGCAGCUUGCTT
	si-3	ACAAGGUUCUGGAAACAAATT	UUUGUUUCCAGAACCUUGUTT

Three siRNAs targeting different sites of each gene were designed and synthesized respectively.

Table S2. Specific primers for each gene used in this study

Gene	Forward primers (5'-3')	Reverse primers (5'-3')
DCBLD2	GGCCCAGTATGATACCCCGAA	ACATCACATTCCCCATCCCT
RAB11FIP1	AGAAAACCAAGAAGCGTGTGTCA	GCGTTTCCAGCAACAGACCATG
CTLA4	TGGAGCTCATGTACCCACC	ATTTTCACATAGACCCCTGTTGT
HOXC6	TTACCCCTGGATGCAGCGAAT	CCGCGTTAGGTAGCGATTGAAGT
KRT6A	GATCGCCACCTACCGCAAG	CTGCACCACAGAGATGTTGACT
beta-actin	TCCTCCTGAGCGCAAGTACTCC	CATACTCCTGCTTGCTGATCCAC

Table S3. Information of primary antibodies in western blot

Antibody	Company	Catalogue number
DCBLD2	proteintech	13168-1-AP
RAB11FIP1	proteintech	16778-1-AP
CTLA4	Santa Cruz Biotechnoligy	SC-376016
HOXC6	Abcam	ab151575
KRT6A	proteintech	10590-1-AP
anti-actin	MDL	MD6553

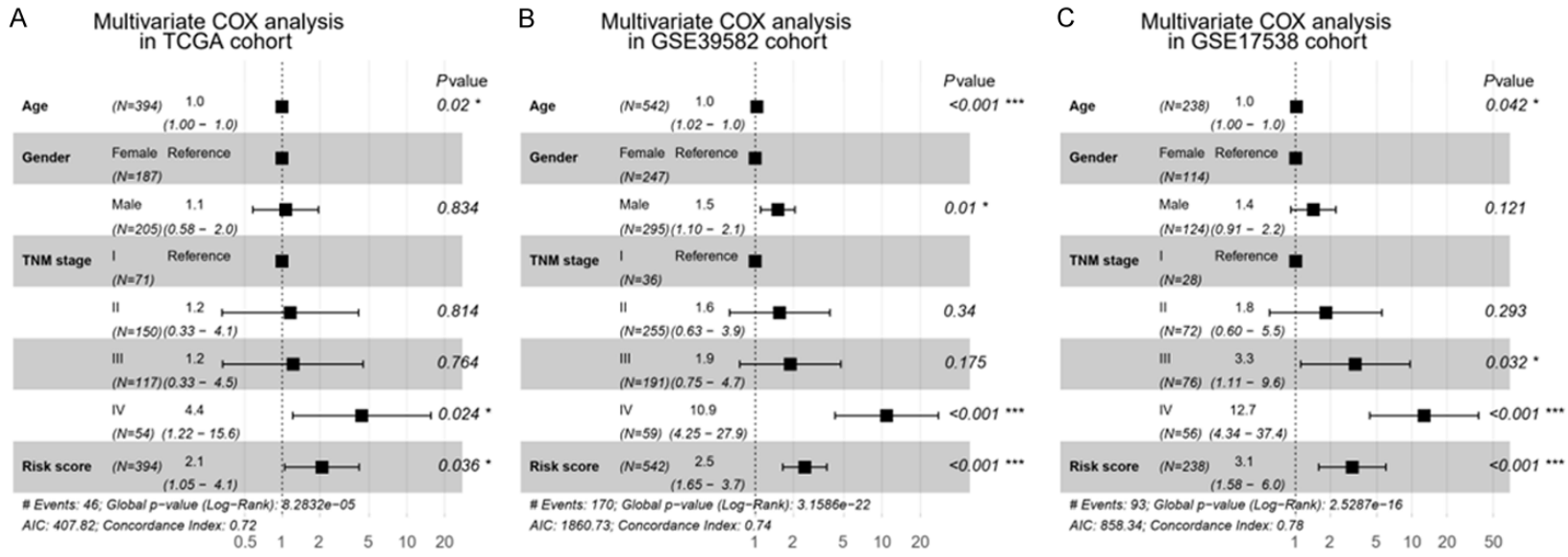


Figure S1. Independence of the five-gene risk model from other clinical variables. Multivariate Cox regression analysis was conducted in (A) the TCGA cohort, (B) the GSE39582 cohort and (C) the GSE17538 cohort.

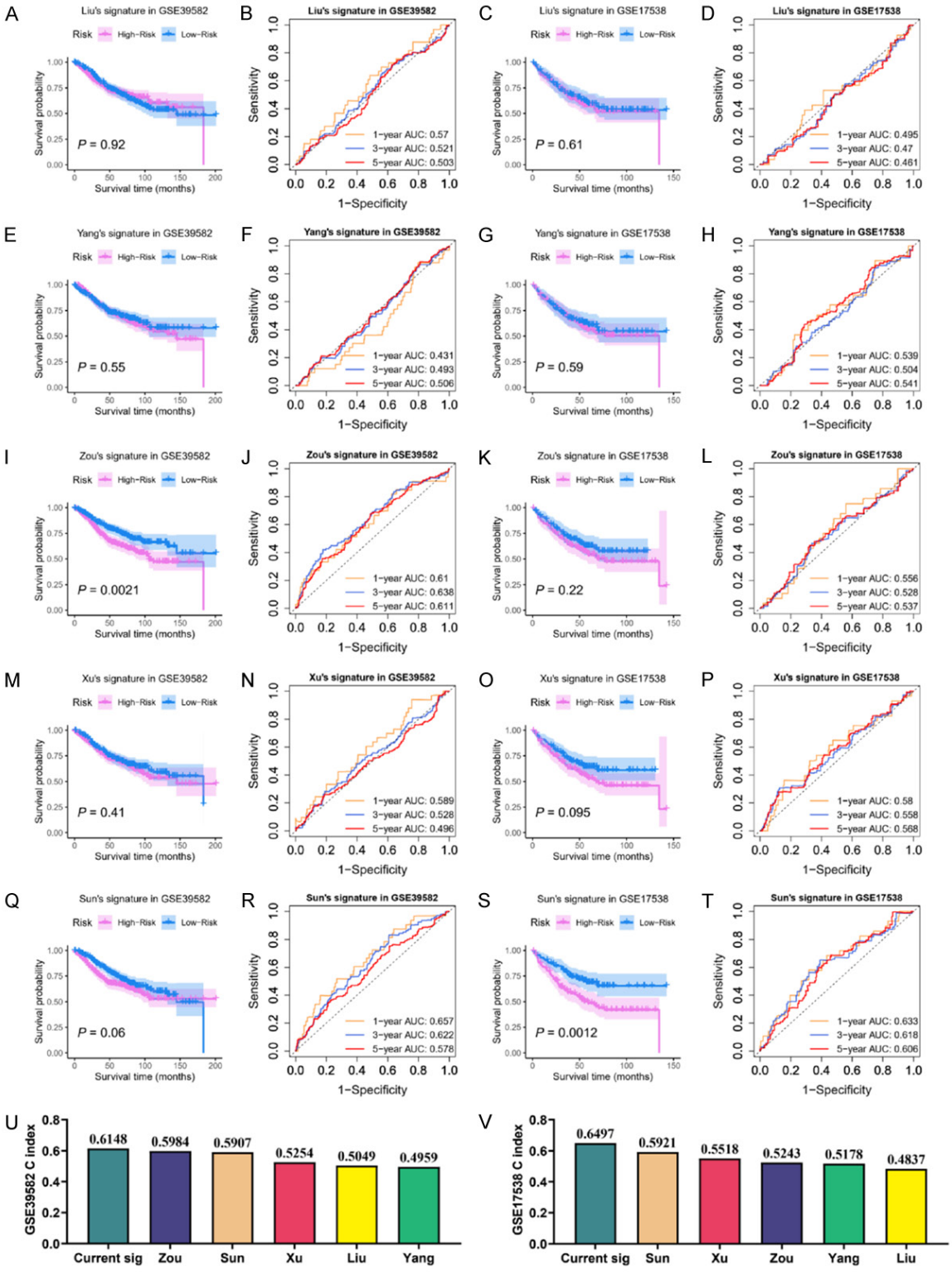
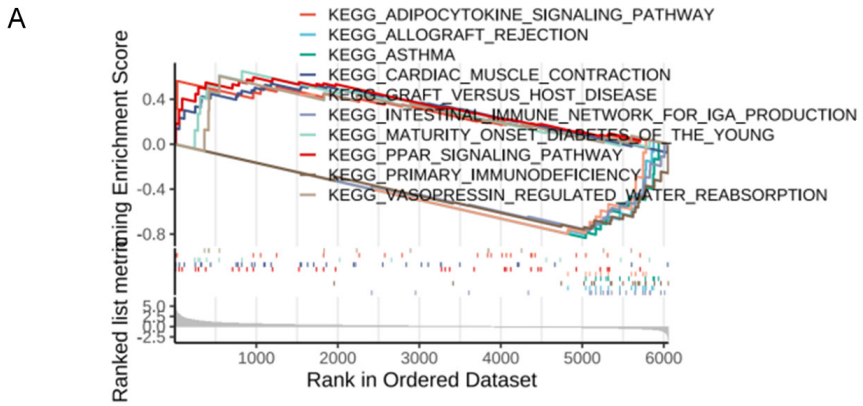
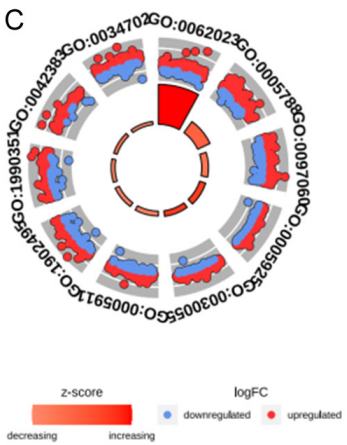
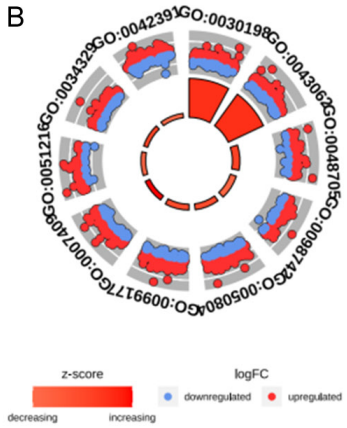


Figure S2. Comparison with previously published signatures using GSE39582 and GSE17538. A-D. Survival curve and ROC of Liu's signature; E-H. Survival curve and ROC of Yang's signature; I-L. Survival curve and ROC of Zou's signature; M-P. Survival curve and ROC of Xu's signature; Q-T. Survival curve and ROC of Sun's signature; U-V. C-indexes of each signature in the GSE39582 and the GSE17538.

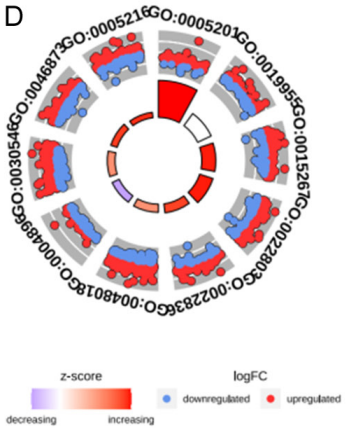




ID	Description
GO:0062023	Collagen-containing extracellular matrix
GO:0005788	Endoplasmic reticulum lumen
GO:0097060	Synaptic membrane
GO:0005925	Focal adhesion
GO:0030055	Cell-substrate junction
GO:0005911	Cell-cell junction
GO:1902495	Transmembrane transporter complex
GO:1990351	Transporter complex
GO:0042383	Sarcolemma
GO:0034702	Ion channel complex



ID	Description
GO:0030198	Extracellular matrix organization
GO:0043062	Extracellular structure organization
GO:0048705	Skeletal system morphogenesis
GO:0098742	Cell-cell adhesion via plasma-membrane adhesion molecules
GO:0050804	Modulation of chemical synaptic transmission
GO:0099177	Regulation of trans-synaptic signaling
GO:0007409	Axonogenesis
GO:0051216	Cartilage development
GO:0034329	Cell junction assembly
GO:0042391	Regulation of membrane potential



ID	Description
GO:0005201	Extracellular matrix structural constituent
GO:0019955	Cytokine binding
GO:0015267	Channel activity
GO:0022803	Passive transmembrane transporter activity
GO:0022836	Gated channel activity
GO:0048018	Receptor ligand activity
GO:0004896	Cytokine receptor activity
GO:0030546	Signaling receptor activator activity
GO:0046873	Metal ion transmembrane transporter activity
GO:0005216	Ion channel activity

Figure S3. GSEA analysis of high- and low-risk groups. A. The five most upregulated and the five most downregulated KEGG pathways between high- and low-risk groups by GESA. B. Biological process in Gene Ontology (GO) analysis. C. Cellular component in GO analysis. D. Molecular function in GO enrichment.

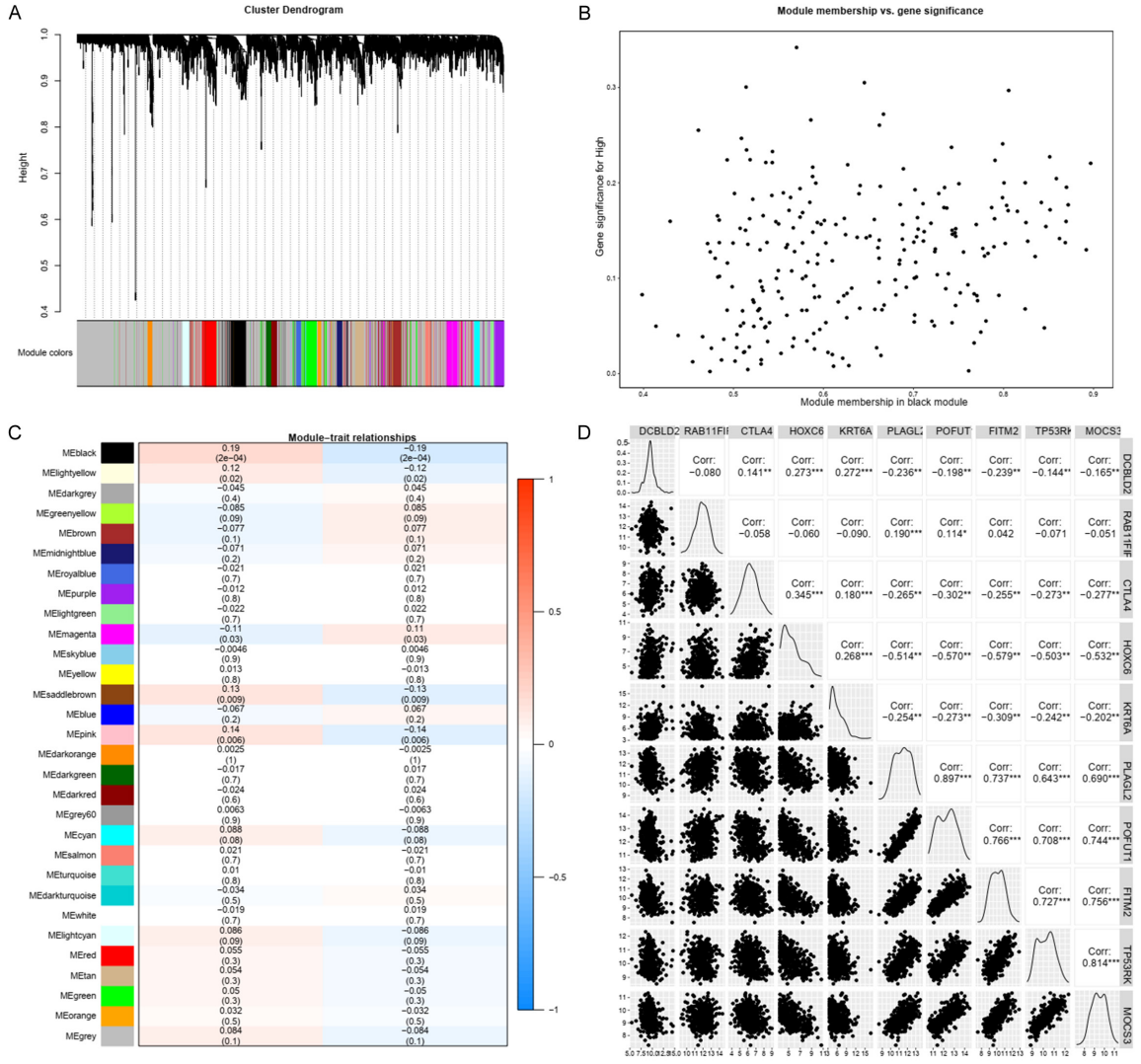


Figure S4. Weighted gene co-expression network analysis of high- and low-risk group. A. The identified module assigned with unique colors. B. The correlation map between module membership and gene significance in black module. C. Module-trait correlation heatmap with red and blue corresponding to positive and negative correlation respectively. D. The correlation between risk genes and hub genes. **P*<0.05, ***P*<0.01, ****P*<0.001.

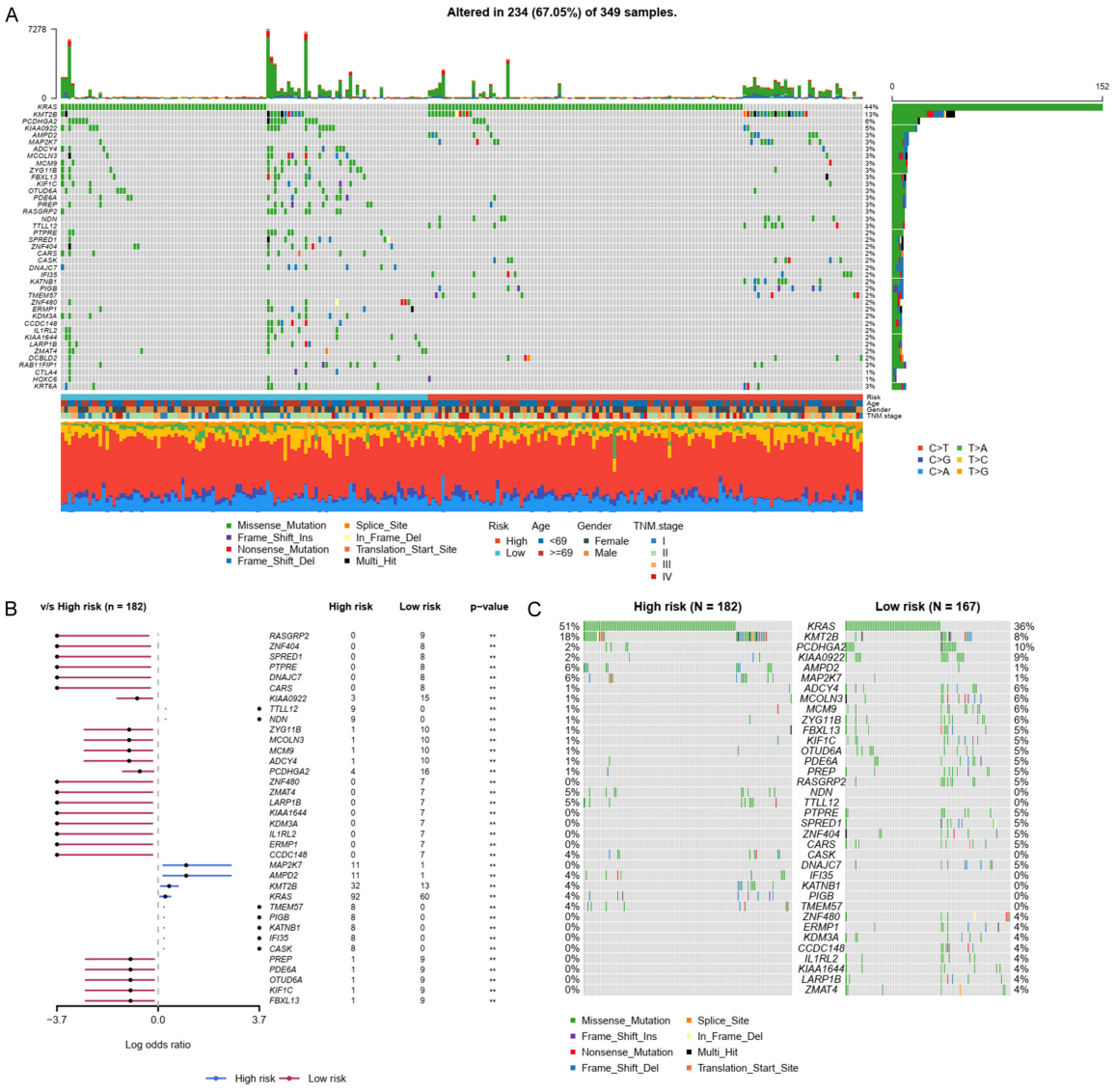


Figure S5. Significantly mutated genes in the TCGA cohort grouped by risk signatures. A. Significantly mutated genes between risk groups (upper) and risk panel genes (lower). B. Forest plot of the 36 significantly mutated genes between risk groups (***P*<0.01). C. Stratified landscape of the 36 significantly mutated genes.

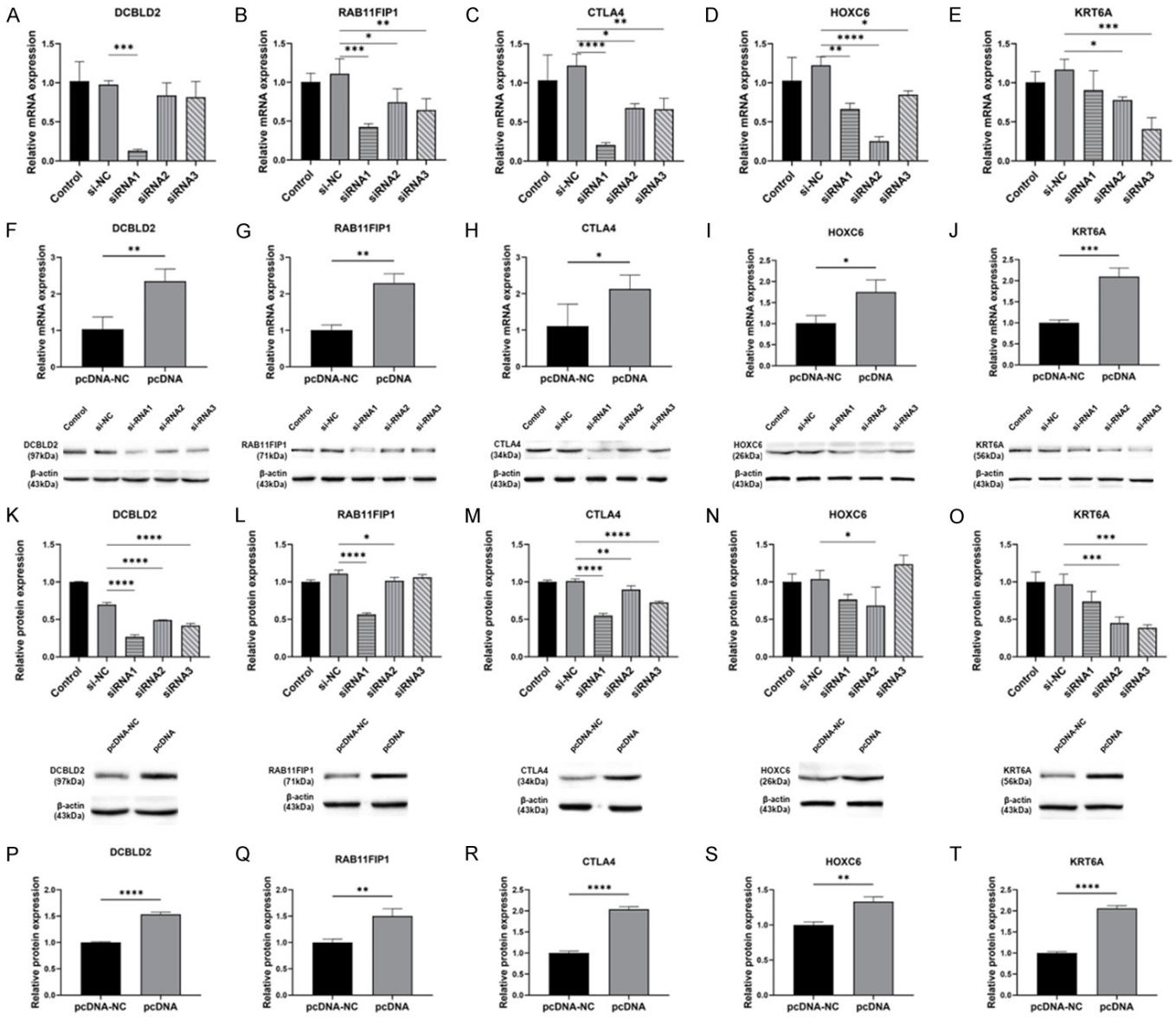


Figure S6. Transfection efficiency evaluation by qPCR and Western blot. The transfection efficiency of siRNAs and pcDNA-3.1-overexpression plasmids for five risk genes were evaluated by qPCR and western blot after 48 hours of transfection in HCT116 cells. A-E. qPCR analysis of transfection efficiency of siRNAs or vectors. F-J. qPCR analysis of transfection efficiency of pcDNA3.1-overexpression plasmids or pcDNA 3.1-vectors. K-O. Western blot analysis of transfection efficiency of siRNAs or vectors. P-T. Western blot analysis of transfection efficiency of pcDNA3.1-overexpression plasmids or pcDNA 3.1-vectors. **P*<0.05, ***P*<0.01, ****P*<0.001, *****P*<0.0001.