Clinics of Oncology

Research Article ISSN: 2640-1037 | Volume 6

Construction and Validation of Prognostic Signature Model Based on Metastatic Features for Colorectal Cancer

Zhao Z[1*], Chen H[1*], Yang Y[2*], Guan X[1], Jiang Z[1], Yang M[1], Liu H[1], Chen T[1], Lv J[1], Zou S[3#], Liu Z[1#] and Wang X[1#]

[1]Department of Colorectal Surgery, National Cancer Center/ National Clinical Research Center for Cancer/ Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[2]Department of Laboratory, National Center for Children's Health/ Beijing Children's Hospital, Capital Medical University, Beijing, China

[3]Department of Pathology, National Cancer Center/ National Clinical Research Center for Cancer/ Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[*]**Corresponding author:**

Xishan Wang,
Department of Colorectal Surgery, National Cancer Center/ National Clinical Research Center for Cancer/ Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China,
E-mail: wxshan1208@126.com

Zheng Liu,
Department of Colorectal Surgery, National Cancer Center/ National Clinical Research Center for Cancer/ Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China,
E-mail: zheng.liu@cicams.ac.cn

Shuangmei Zou,
Department of Pathology, National Cancer Center/ National Clinical Research Center for Cancer/ Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, E-mail: zousm@cicams.ac.cn

Received: 15 Sep 2022
Accepted: 24 Sep 2022
Published: 29 Sep 2022
J Short Name: COO

Citation:

Wang X, Zou S and Liu Z. Construction and Validation of Prognostic Signature Model Based on Metastatic Features for Colorectal Cancer Clin Onco. 2022; 6(12): 1-13

Keywords:

Colorectal Cancer; Liver Metastasis; Prognostic Signature; Risk Model; Infiltrating Immune Cells.

Authors Contribution:

Zhao Z, Chen H, Yang Y and all these authors are equally contributed to this work.

1. Abstract

1.1. Background: Colorectal Cancer (CRC) is a common malignant cancer with a poor prognosis. Liver metastasis is the dominant cause of death in CRC patients, and it often involves changes in various gene expression profiling. This study proposed to construct and validate a risk model based on differentially expressed genes between the primary and liver metastatic tumors from CRC for prognostic prediction.

1.2. Methods: Transcriptomic and clinical data of CRC were downloaded from The Cancer Genome Atlas database (TCGA) and Gene Expression Omnibus database (GEO). Identification and screening of candidate differentially expressed genes (DEGs) between liver metastatic tissues and corresponding primary tumors were conducted by R package "limma" and univariate Cox analysis in the GSE50760 and TCGA cohort. Last absolute shrinkage and selection operator (LASSO) Cox regression was carried out to shrink DEGs and develop the risk model. CRC patients from the GSE161158 cohort were utilized for validation. Functional enrichment, CIBERSORT algorithm, and ESTIMATE algorithm for further analysis.

1.3. Results: An 8-gene signature risk model, including HPD, C8G, CDO1, FGL1, SLC2A2, ALDOB, SPINK4, and ITLN1, was developed and classified the CRC patients from TCGA and GEO cohorts into high and low-risk groups. The high-risk group has

a worse prognosis compared with the low-risk group. The model was verified as an independent indicator for prognosis. Moreover, tumor immune infiltration analyses demonstrated that monocytes ($P = 0.006$), macrophage M0 ($P < 0.001$) and macrophage M1 ($P < 0.001$) were enriched in the high-risk group, while plasma cells ($P = 0.010$), T cells CD4 memory resting ($P < 0.001$) and dendritic cells activated ($P = 0.006$) were increased in the low-risk group.

1.4. Conclusions: We developed and validated a risk predictive model on the DEGs between liver metastases and primary tumor of CRC, which can be utilized for the clinical prognostic indicator in CRC.

2. Introduction

Colorectal Cancer (CRC), a major malignancy of the digestive system, ranks third among malignant cancers in terms of morbidity worldwide. [1] Approximately 30–50% of patients with primary colon cancer relapse and die from metastases, especially for liver metastasis. [2] The mechanisms of CRC metastasis have been investigated for a long time, which are involved in epithelial-mesenchymal transition, tumor motion, invasion, proliferation, and metabolism. [3] Due to the tumor heterogeneity, there are certain differences in the expression profile between the primary and the metastatic tumor. The existence of differences in expression leads to a series of changes in biological behaviors, and ultimately causes the occurrence of metastasis, which is related to a poor prognosis.

Meanwhile, given the poor prognosis of CRC, the identification of prognostic biomarkers in CRC patients will ultimately facilitate appropriate individualized treatments for patients with a high risk of tumor progression. Therefore, there is an urgent need to identify highly robust biomarkers to enable individualized treatment decisions, which may then guide drug development and the use of combination therapies, targeted therapies, and immunotherapies. Therefore, finding prognostic markers is critical for better patient management.

In addition, the composition and function of Tumor-Infiltrating Immune Cells (TIICs) have potential prognostic performance. CIBERSORT is a gene expression-based deconvolution algorithm that uses a set of barcode gene expression values to characterize immune cell composition. [4] The relative proportion of 22 types of infiltrating immune cells in tumors could be inferred by CIBERSORT algorithm, and a series of researches have taken advantage of this algorithm to investigate the relationship between tumor microenvironments (TME) and prognosis [5-7].

In current study, Differentially Expressed Genes (DEGs) between primary and liver metastases of CRC were screened out from GSE50760, and prognosis-related genes were identified from the public databases of adenocarcinoma and rectal adenocarcinoma dataset from The Cancer Genome Atlas Colon (TCGA-COREAD). A prognostic model was developed by using the lasso cox regression method and verified the performance by using the TCGA-COREAD and GSE161158 gene expression cohort. Analyses of functional enrichment and tumor microenvironment were also conducted to explore the potential mechanisms.

3. Materials and Methods

3.1. Data Collection and Preprocessing

Differentially Expressed Genes (DEGs) screening dataset GSE50760 was retrieved from the Gene Expression Omnibus (GEO) database. This cohort included RNA-seq data of 54 samples (normal colon, primary CRC, and liver metastasis) which were generated from 18 CRC patients. Sequencing was performed in paired end reads (2x100 bp) using Hiseq-2000 (Illumina). The gene expression and clinical information of colon adenocarcinoma (COAD) and rectal adenocarcinoma (READ) samples were downloaded from the UCSC Xena Browser (https://xenabrowser.net/) [8]. Totally 641 colorectal adenocarcinoma (COREAD) samples with corresponding expression and clinical data were obtained after combining the information of COAD and READ datasets. Another validation cohort GSE161158 was also downloaded from GEO database, which was performed on the microarray platform of [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. In GSE161158, totally 250 AJCC-TNM staging II or staging III CRC patients were enrolled.

3.2. Identification of Differentially Expressed and Prognostic Genes

In GSE50760 cohort, Differentially Expressed Genes (DEGs) between liver metastatic tissues and corresponding primary tumor were identified by R package “limma”, according to false discovery rate ($FDR < 0.05$ and $|\log2FoldChange| > 2$. Through univariate Cox analysis, the association between expression levels of DEGs between primary and liver metastatic tissues and CRC patients’ Overall Survival (OS) was explored. DEGs with prognostic value in the GSE50760 cohort were subjected to construct a prognostic model.

3.3. Construction and Validation Of The Prognostic Model

According to the expression of prognostic DEGs and survival data, the LASSO Cox regression analysis by R package “glmnet” was performed to further select the most useful prognostic markers and the penalty regularization parameter lambda was chosen based on 5 cross-validations. Through multiplying the expression level of a gene by its corresponding Cox regression coefficient, the risk score for each patient was calculated using the following formula: risk score = esum (each gene’s expression × corresponding coefficient). The patients were separated into high- and low-risk groups based on the median value of the risk score. The “Rtsne” package and the “prcomp” function in the “stats” package were used to perform the t-SNE and PCA analysis to explore the distribution of high- and low-risk groups. Kaplan–Meier survival curves and a time-dependent ROC curve analysis were applied to compare the

survival between the above two groups and evaluate the model's predictive ability using the "survivalROC" package in R, respectively.

3.4. Functional Enrichment Analysis

The enrichment analysis of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) was carried out according to the DEGs to explore different molecular mechanisms and between high- and low-risk patients by utilizing the "clusterProfiler" R package. The P values are adjusted using the BH method to control the FDR.

3.5. Inference of Immune Infiltration in Samples and Calculation of Immune Score, Stromal Score, and ESTIMATE Score

A deconvolution algorithm called CIBERSORT was used in the study, which can quantify the percentage of different types of tumor-infiltrating inflammatory cells (TIICs) accurately, under the complex "gene signature matrix" based on 547 genes. In the current study, we illustrated the immune infiltration of each sample with the LM22 signature file, which can define 22 subtypes of immune cells, including naïve B cells, memory B cells, plasma cells, CD8+ T cells, naïve CD4+ T cells, resting memory CD4+ T cells, activated memory CD4+ T cells, follicular helper T cells, regulatory T cells (Treg cells), gamma delta T cells, resting NK cells, activated NK cells, monocytes, M0 macrophages, M1 macrophages, M2 macrophages, resting dendritic cells, activated dendritic cells, resting mast cells, activated mast cells, eosinophils, and neutrophils, with the preset signature matrix at 1000 permutations. After using the CIBERSORT program, the distribution of 22 subtypes of TIICs was presented, along with the results of correlation coefficient, P-value and root mean squared error (RMSE), which can evaluate the accuracy of the results in each sample. The P-value ≤ 0.05 reflects a statistical connotation of the results of deconvolution across all cell subsets for each sample and is useful for excluding results with less accuracy. Finally, 18 primary tumors, 18 liver metastatic tissues and 18 control samples were selected for later analysis because they met the required P-value.

ESTIMATE (Estimation of Stromal and Immune cells in Malignant Tumor tissues using expression) algorithm was used to evaluate the ratio of the immune- stromal component in the tumor microenvironment (TME) through utilizing "estimate" R package, which generates three scores including Immune Score (reflecting the level of immune cells infiltrations), Stromal Score (reflecting the presence of stroma), and ESTIMATE Score (reflecting the sum of both). The higher the respective score is, the larger the ratio of the corresponding component in TME exists.

3.6. Statistical Analysis

Student's t-test was applied to identify the differentially expressed genes between tumor tissues and adjacent tissues and evaluate the difference of Immune Score, Stromal Score, and ESTIMATE Score between risk groups. The Chi-squared test was used to compare the difference of proportion composition. The OS between groups was compared by using the Kaplan–Meier analysis with the log-rank test. And the identification of an independent classifier of OS was managed by the analysis of univariate and multivariate Cox regression. All statistical analyses were completed with R software (Version 4.1.0). All P values are two-tailed with a P value less than 0.05 was considered statistically significant.

4. Results

4.1. Identification and Functional Enrichment Analysis of Prognostic Degs Between Primary and Liver Metastatic Tissues

A total of 158 DEGs were identified between primary and liver metastatic tissues in GSE50760, which were visualized by volcano map and heatmap (Figure 1A-B, Figure-S1). According to the univariate Cox regression analysis, 13 of the above DEGs were correlated with OS in the COREAD TCGA cohort, including 8 protective genes and 5 risk genes (Figure-2a and Figure-2b). To verify the correlation of biological functions and pathways with the prognostic model, the GO enrichment and KEGG pathway analyses were carried out according to the DEGs between the high-risk and low-risk groups in TCGA-GOREAD cohort. The genes were mainly enriched in the small molecule catabolic process and carbohydrate transmembrane transport between two groups in GO enrichment analysis (Figures 2E-G). KEGG pathway analysis also revealed that the amino metabolism and pentose phosphate pathway was enriched in the COREAD cohorts (Figures 2H).

4.2. Construction of A Risk Score Model in the TCGA Cohort

LASSO regression analysis was used to develop a risk score by analyzing the expression level of the 13 DEGs mentioned above. 8 genes most contributing to the OS of CRC patients were identified, according to the minimal value of lambda (Figure 3A-B, Figure S2) , and a risk formula was constructed with the expression levels of 8 genes: risk score = e(-0.168 ∗ expression level of FGL1+0.466∗ expression level of HPD + -0.333∗ expression level of SLC2A2 +0.218∗ expression level of C8G + 0.121∗ expression level of CDO1+-0.119∗ expression level of ALDOB +-0.038∗ expression level of SPINK4 +-0.006∗ expression level of ITLN1) . The patients in the COREAD cohort were divided into high- and low-risk groups, using the median risk score as the cut-off in both training set and test set. The expression pattern of risk genes in the high- and low-risk groups was visualized by heatmap, patients with high-risk demonstrated upregulation of HPD, C8G, CDO1, in contrast to patients with low-risk scores expressed upregulation of FGL1, SLC2A2, ALDOB, SPINK4, ITLN1 (Figure 2A-B). Besides, the risk signature for OS was ranked (Figure 2C-D). The OS status of individual patients with CRC was demonstrated by dot plots (Figure 2E-F). Meanwhile, significant differences were observed in the expression of all 8 DEGs between groups (high- vs. low-risk scores; $p < 0.001$; Figures 2D-I). As for the whole

patients from TCGA-COREAD cohort, the overall survival of the high-risk group was poorer than that in the low-risk group (Figure 3I, P<0.001). Moreover, the ROC curves were utilized to make an evaluation of the model, and the Area Under The Curve (AUC) reached values of 0.735, 0.745, and 0.753 at 1, 3, and 5 years, respectively (Figure 3J).

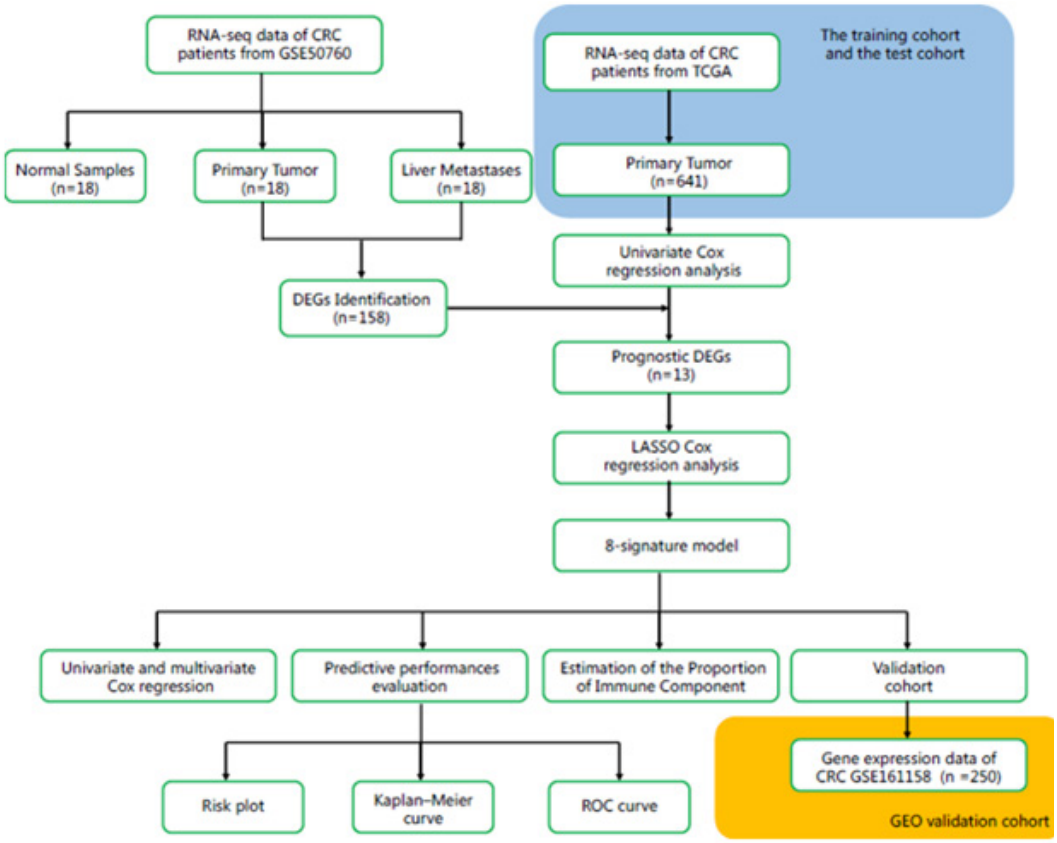
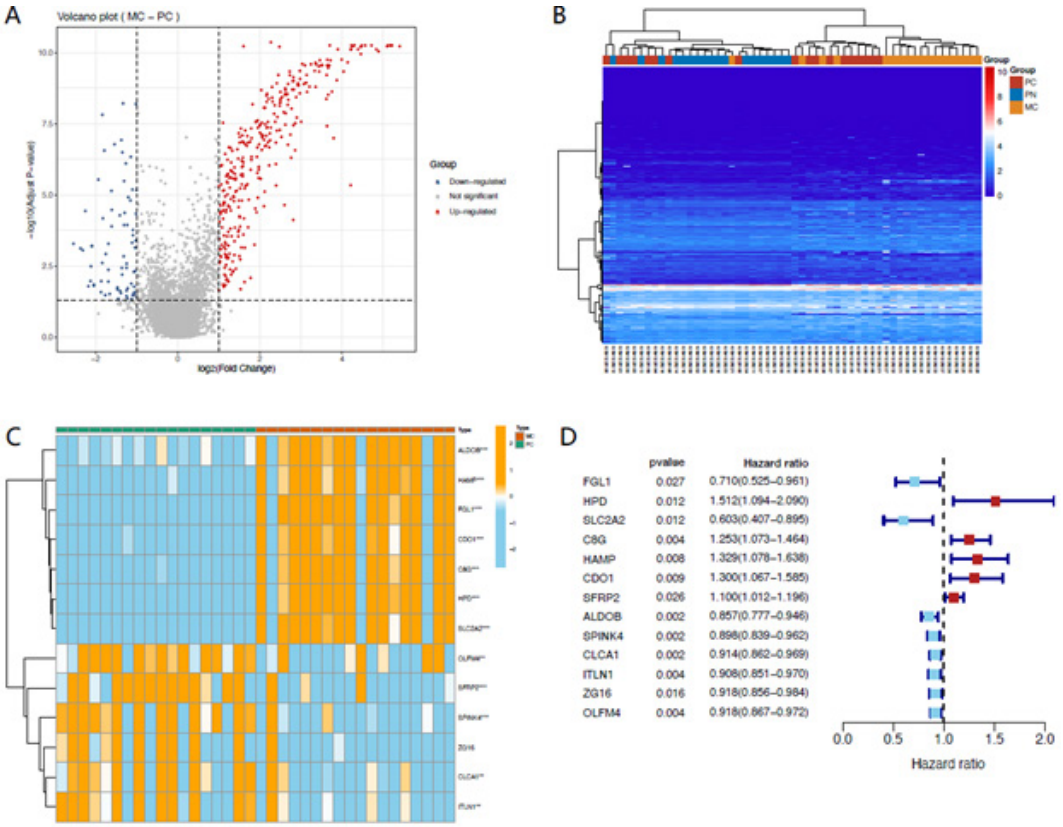


Figure 1: The flow diagram of data collection and analysis in the study



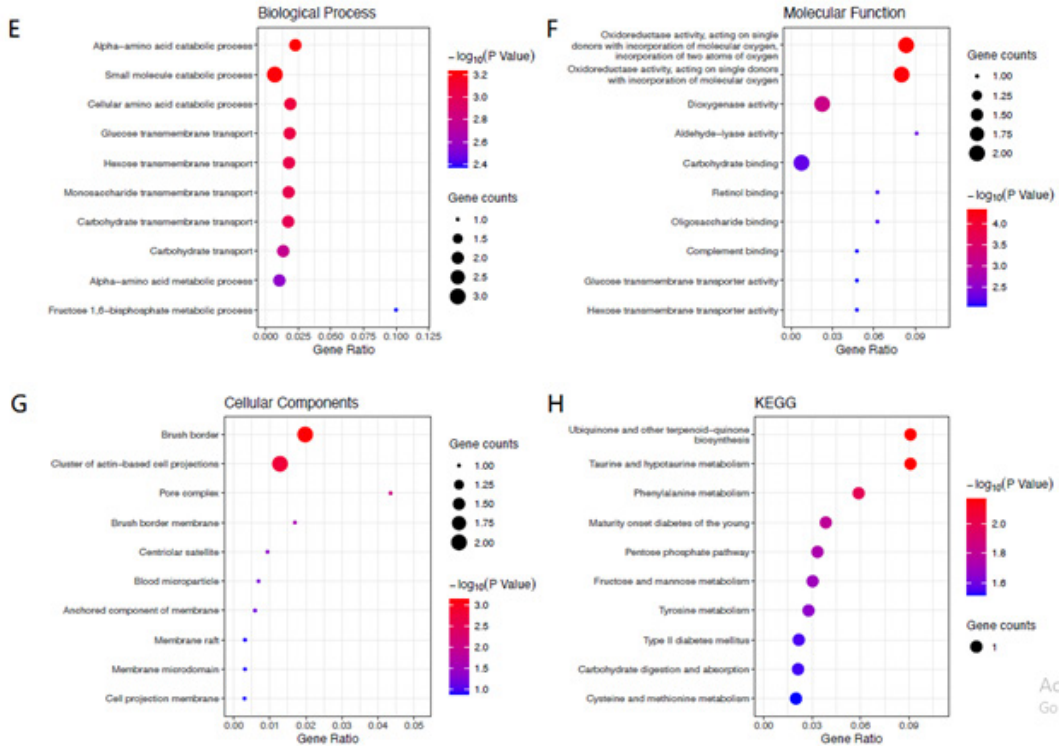


Figure 2: Volcano map(A) and heatmap(B) show DEGs between primary tumor and corresponding liver metastases in CRC were screening from GSE50760. Heatmap of the expression of prognostic DEGs between primary tumor and corresponding liver metastases in CRC (C). The depth of red represents the level of high expression, and the depth of green represents the level of low expression * P<0.05, **<0.01, ***<0.001, ****<0.0001. The effect of DEGs on the prognosis of CRC (D). GO terms and KEGG for mRNAs with 13 prognostic DEGS between primary CRC and corresponding liver metastases, including Cellular component (E), biological process (F), molecular function (G) and KEGG (H).

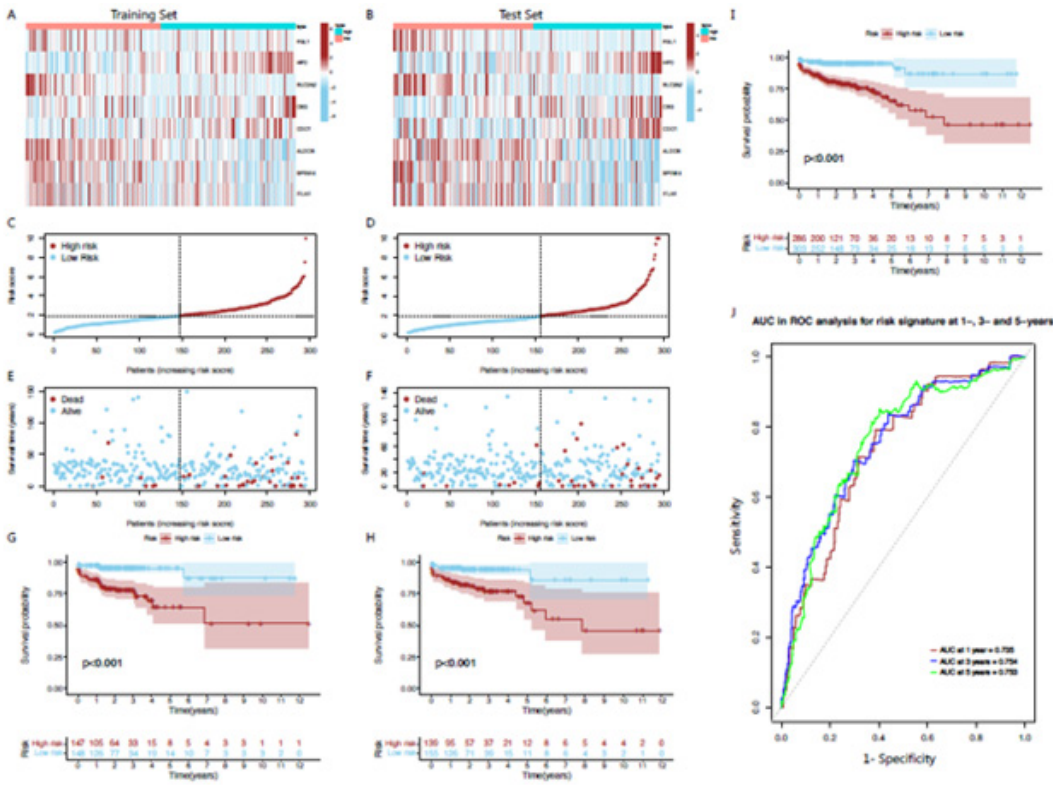


Figure 3: Heatmap demonstrates the expression of 8-signature risk model genes in the high- and low-risk score groups in training set and test set (A-B). Risk score distribution (C-D), scatter plot (E-F) and Kaplan–Meier OS curves (G-H) of high-risk group and low-risk group. The Kaplan-Meier plot of the risk score model to show the for the CRC prognostic status (I). The ROC curves of risk score model for the prognostic accuracy of CRC patients in 1-year, 3-years and 5-years (J).

4.3. Clinicopathological Relationship and Independent Prognostic Value of the Risk Signature Model

The association between the risk signature and different clinical characteristics was analyzed, and gender, age, tumor location, TNM pathological stage, T stage, N stage, and M stage were included (Figure 4 and Table 1). The tumor with a high-risk score tended to locate on the colon and have more invasive tumor traits including the later pathological stage, greater tumor depth, lymphatic node metastasis, and distant organ metastasis. Moreover, we plotted the Kaplan-Meier curves for different clinicopathological features based on the stratification of the risk signature model. As shown in Figure, except for stage IV (P=0.054), under the condition of other clinical factors, the 8-gene risk signature model was all closely associated with patient prognosis (Figure 5).

To further study the significance of the model in risk stratification, we carried out univariate and multivariate Cox regression analyses in training set and validation set of TCGA-COREAD cohort. As demonstrated in Figure 4, there were significant relationship with OS in the TCGA cohort training set (HR =1.382, 95% CI = 1.144–1.670, P < 0.001, Figure 6A) and test set (HR =1.296, 95% CI = 1.154–1.455, P < 0.001, Figure 6B) respectively. Further multivariate analyses revealed the risk score could be taken as an independent predictor for OS (training set: HR =1.367, 95% CI = 1.199–1.559, P < 0.001, Figure 6C; test set: HR =1.292, 95% CI = 1.161–1.439, P < 0.001, Figure 6D).

3.4 Verification of Prognostic Signature in colorectal GEO Cohort.

According to the model developed by COREAD TCGA cohort, the prognostic score of each patient from GSE161158 were figured out. Afterward, the patients were divided into the high-risk and low-risk group based on the median value of the risk score (Figure 4A-B). Considering the prognosis-related data of GSE 161158 were disease-free survival, the TNM stage IV were excluded. In consistent with the test set and validation set, the validation set patients with high-risk scores proved a poorer DFS (P=0.025) and the AUC of 3-year DFS was 0.603 (Figure 4D-E).

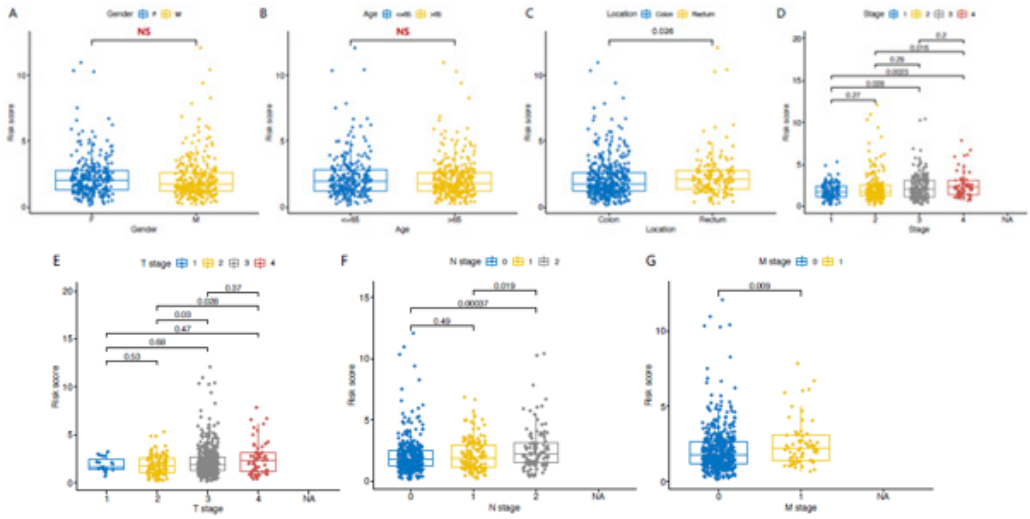


Figure 4: Stratified prognostic analysis of the clinicopathological parameters for CRC patients. The plot-boxes represent the Risk score of high and low risk patients according to gender (A), age (B), tumor location (C), TNM stage (D), T stage (E), N stage (F), and M stage (G). A two-sided Log-Rank and Wilcoxon test; P < 0.05 was considered significant.

Table 1: Clinical information of the high- and low risk groups

	Level	High-risk	Low-risk	P
		286	303	
Gender (%)	FEMALE	149 (52.1)	129 (42.6)	0.026
	MALE	137 (47.9)	174 (57.4)	
Age (%)	<=65	128 (44.8)	129 (42.6)	0.652
	>65	158 (55.2)	174 (57.4)	
Location (%)	Colon	197 (68.9)	233 (76.9)	0.036
	Rectum	89 (31.1)	70 (23.1)	
TNM Stage (%)	I	45 (16.4)	61 (20.6)	0.085
	II	100 (36.4)	123 (41.6)	
	III	89 (32.4)	84 (28.4)	
	IV	41 (14.9)	28 (9.5)	
T.stage (%)	T1	9 (3.1)	10 (3.3)	0.289
	T2	45 (15.7)	62 (20.5)	
	T3	197 (68.9)	205 (67.7)	
	T4	35 (12.2)	26 (8.6)	
N.stage (%)	N0	154 (54.2)	191 (63.0)	0.027
	N1	68 (23.9)	70 (23.1)	
	N2	62 (21.8)	42 (13.9)	
M.stage (%)	M0	209 (83.6)	243 (90.0)	0.042
	M1	41 (16.4)	27 (10.0)	

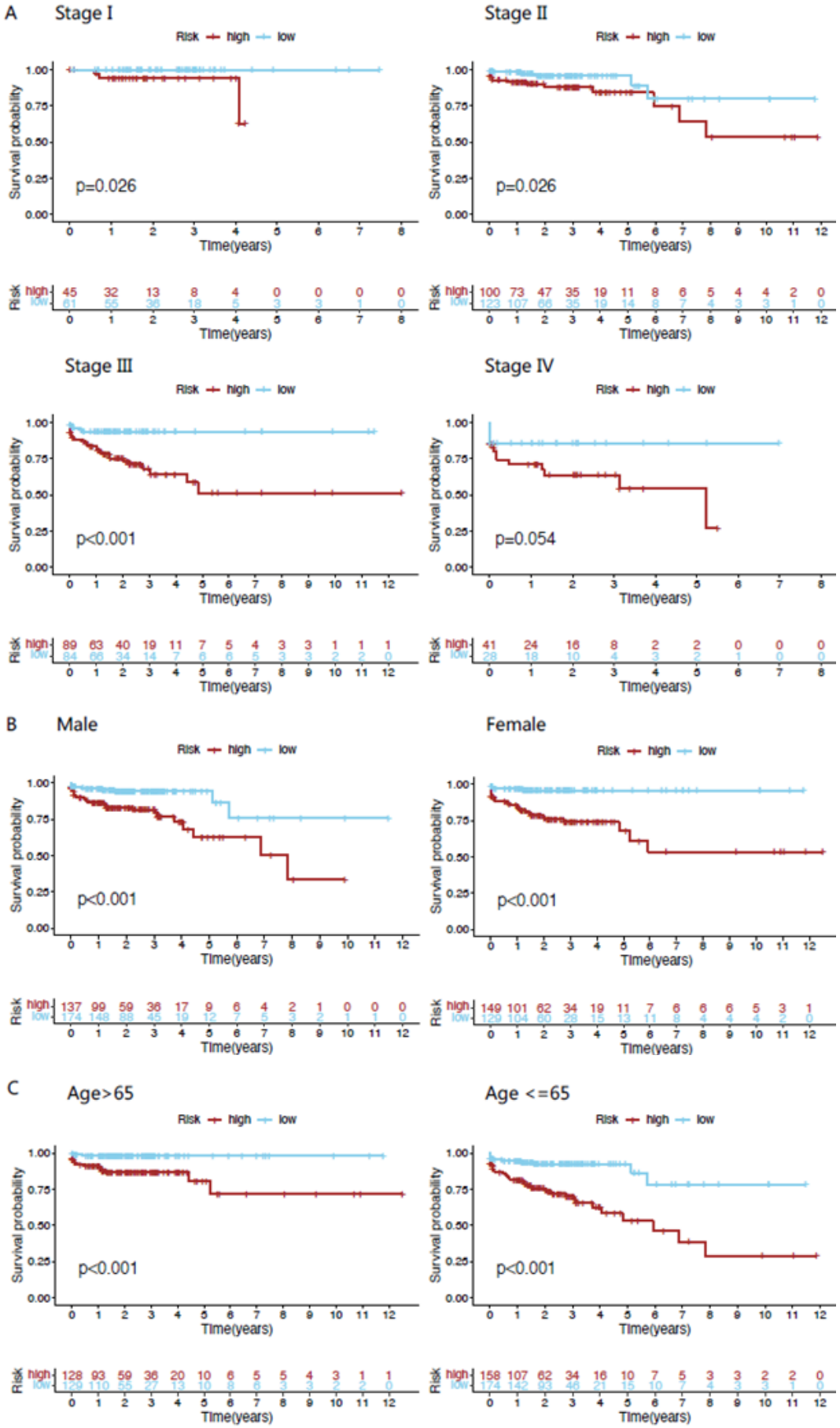


Figure 5: (A)Heatmap and clinicopathological characteristics of the subgroup classified by the 8-gene prognostic signature in COREAD. Kaplan–Meier OS curves for patients with TNM stage(B), gender (C), age

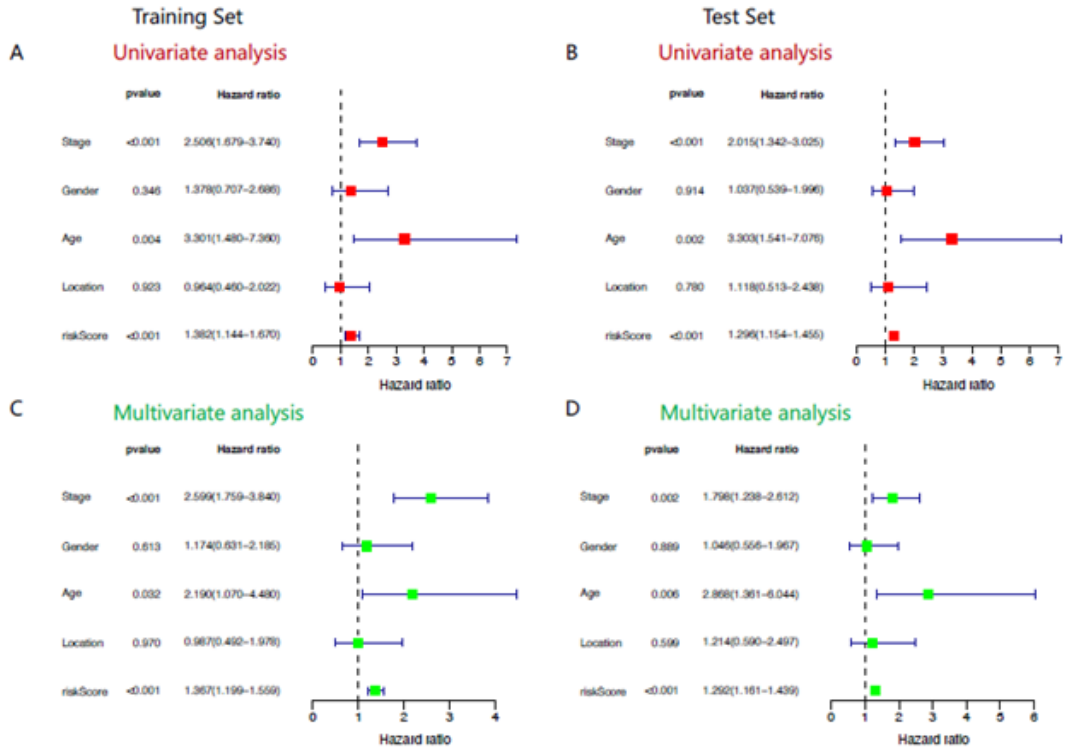


Figure 6: Cox's proportional hazard model of correlative factors in CRC patients. Univariate Cox regression analysis for the clinicopathological parameters affecting the overall survival in training set and test set of TCGA-COREAD cohort (A-B). Multivariate Cox regression analysis for the clinicopathological parameters affecting the overall survival in training set and test set of TCGA-COREAD cohort (C-D).

4.5. The Relevance of Risk Signature Model with Immune Infiltration

To distinguish the variance of the distribution of 22 TIICs between the two risk groups, we analyzed the CRC cases through the CIBERSORT algorithm on the basis of expression profiles from TCGA. The violin plot exhibited the ratio differentiation of 22 TIICs between high- and low-risk group expression (Figure 6A). Monocytes ($P = 0.006$), macrophage M0 ($P < 0.001$) and macrophage M1 ($P < 0.001$) were enriched in the high-risk group, while plasma cells ($P = 0.010$), T cells CD4 memory resting ($P < 0.001$) and dendritic cells activated ($P = 0.006$) were increased in the low-risk group. Next, Immune Score, Stromal Score, and ESTIMATE Score were analyzed to assess the contents of the immune-stromal components in tumor microenvironments (TME). As manifested in Figure, Stromal Score and ESTIMATE Score were significantly higher in high-risk groups ($P < 0.001$ and $P = 0.0019$, respectively), which demonstrated that more immune-stromal components existed in TME of the high-risk group (Figure 6B-D).

5. Discussions

Metastases are the main reasons for the poor prognosis of CRC, especially for liver metastases. It is of great significance to predict the progression risk of patients and carry out appropriate intervention for the high-risk group. In the current study, we established and validated an 8-signature risk model based on DEGs between the primary CRC and corresponding liver metastases, which could be taken as an independent risk factor. Firstly, we utilized the R limma package to screen out 158 DEGs between the primary and corresponding metastatic lesions. Subsequently, 13 of these genes were identified in the univariate Cox analysis in TCGA-COREAD cohort. Enrichment analysis was conducted to these genes that were significantly correlated with pathways of small molecule catabolic process and transport, including the ubiquinone and other terpenoid-quinone biosynthesis, alpha-amino acid catabolic process (Taurine and hypotaurine metabolism, phenylalanine metabolism, tyrosine metabolism, cysteine and methionine metabolism), and monosaccharide transmembrane transport. The stepwise regression algorithm was used to develop an 8-gene signature as a prognostic risk model by using Lasso regression. There were 8 signatures constituting the risk model, including with high-risk demonstrated upregulation of HPD, C8G, CDO1, in contrast to patients with low-risk scores expressed upregulation of FGL1, SLC2A2, ALDOB, SPINK4, ITLN1.

Notably, all eight DEGs have prognostic significance in tumor development. CDO1 plays a role as a methylation-specific gene in human cancer and the methylation abnormalities in CDO1 has been reported as a prognostic factor in various cancers, including colorectal cancer, breast cancer [9], gallbladder cancer [10], esophageal squamous cell carcinoma [11], and lung cancer [12]. The methylation abnormalities in CDO1 reflect the accumulation changes with progression and the degree of malignancy. Primary CRC cancers with liver metastasis harbored significantly higher methylation of CDO1 than those without liver metastasis [13]. It was demonstrated that the protein encoded by HPD is an enzyme

in the catabolic pathway of tyrosine. This kind of protein catalyzes the conversion of 4-hydroxyphenylpyruvate to homogentisate, which is a cause of tyrosinemia type 3 and hawkinsinuria. GO annotations related to this gene include oxidoreductase activity, acting on single donors with incorporation of molecular oxygen and 4-hydroxyphenylpyruvate dioxygenase activity [14]. The protein encoded by C8G belongs to the lipocalin family, which is one of the three subunits that constitutes complement component 8 (C8). C8 participates in the formation of the Membrane Attack Complex (MAC) on bacterial cell membranes. Gene Ontology (GO) annotations related to this gene include complement binding [14]. Gene expression analyses indicated that the expression levels of FGL1 were increased in human solid tumors, such as colorectal cancer, lung cancer, and breast cancer [15]. FGL1 promotes the development of tumor, which is involved in EMT process tumor proliferation, apoptosis, radiation and drug sensitivity [16-19]. Moreover, the FGL1 could combine with LAG-3, which weakens the cytotoxicity of CD8+ T cells and contributes to tumor growth [15] [20]. Lin and colleagues demonstrated that high levels of SLC2A2 in human colon cancer tissues are associated with advanced stages and poor prognosis. The inhibition of SLC2A2 can be effectively used for colon cancer chemoprevention [21]. Metabolic and transcriptomic analyses showed that ALDOB was upregulated in the liver metastases compared with the primary tumor [22]. Targeting ALDOB inhibition significantly reduces liver metastatic growth but has little effect on the primary tumor [23]. Mechanically, silencing ALDOB activated epithelial markers and repressed mesenchymal markers, indicating inactivation of ALDOB may lead to inhibition of Epithelial-Mesenchymal Transition (EMT) [24]. SPINK4 expression was downregulated in CRC compared with that in normal tissues, and low level of SPINK4 expression was associated with poor prognosis in CRC patients. The lower SPINK4 expression was significantly related to higher TNM stage. [25]. As for rectal cancer, the high SPINK4 expression is associated with advanced clinicopathological features and a poor therapeutic response among the patients undergoing neoadjuvant concurrent chemoradiotherapy [26]. ITLN1 acts as a tumor suppressor in various cancers, such as gastric cancer, colon cancer and ovarian cancer [27, 28]. Katsuya showed that the CRC cases with reduced ITLN1 expression had higher M grades than CRC cases in which ITLN1 was retained, and patients with retained ITLN1 expression tended to have more favorable prognoses than those with reduced ITLN1 expression [29]. Increased ITLN1 expression in CRC cells significantly inhibited local pre-existing vessels sprouting, and the infiltration of immunosuppressive myeloid-derived suppressor cells into tumor tissues without affecting the behavior of CRC cells [26].

Furthermore, a LASSO Cox regression was performed to establish an 8-gene biomarker as a novel prognostic model. The prognostic efficiency of the signature was investigated in TCGA dataset and GEO validation by Kaplan-Meier survival curves and ROC curves, which proved a good predictive performance of risk model. We carried out a Cox regression analysis on risk score, age, gender, tumor location and TNM stages. Results showed that the immune risk score model was an independent factor for predicting the prognosis of CRC. Considering the significance of clinical factors, Kaplan-Meier curves were applied to different clinical pathological parameters. As shown in the results, the model could obviously distinguish the difference in survival in aspect of TNM stage I-III, gender (both male and female), age (both >65 and ≤ 65), and tumor location (both colon and rectum). Importantly, the risk score tent to improve with the progression of tumor in TCGA cohort. The CRC with higher TNM stage, deeper invasion, more positive lymph node and distant metastasis tend to have higher score, which may implement the risk score may predict the tumor progression and patients' clinical outcomes.

Accelerated cancer deterioration is not only related to malignant cells, but also affected by the TME [30]. Tumor-infiltrating immune cells in the TME play a central role in a series of tumor behaviors, such as tumorigenesis, tumor proliferation, metastases and even tumor suppression. Thus, we evaluated the infiltrating immune cells in two risk group to reflect the TME of CRC. In the present study, the high-risk group has a higher level of monocytes and macrophages, but low-risk group has a higher level of plasma cells, T cells CD4 memory resting, and dendritic cells activated. Monocytes have been previously shown to promote metastasis [31]. Previous studies have shown that monocytes from patients with advanced cancer secreted higher level tumor necrosis factor-α (TNF-α) than those from patients at early stage. Monocytes in CRC are prone to produce TNF-α after stimulation, which is related to the survival risk [32]. Of note, we also found that high infiltration of M1 macrophages were associated with poor prognosis. In previous studies, M1 macrophages were often known to inhibit cancer progression. Zhang et al. reported that low infiltration of M1 macrophages was associated with poor progress of the prostate cancer patients in TCGA cohort [32, 33]. In our study, high infiltration of high level of M1 macrophages was found in high-risk tissues with reduced survival. There are two possible reasons for this opposite conclusion: one is that our model does not include enough samples and leads to result shifts, and another one is that M1 macrophages infiltrating into the CRC microenvironment are polarized into M2 macrophages. However, the mechanism needs to be further studied by experiments. Dendritic cells play a central role in the adaptive anti–tumor immune response. They act as sentinels, detect tumor antigens, present them to CD8+ T-cells, and supply necessary signals for both activation and suppression of CD8+ T-cells [34]. Studies investigating the clinical value of tumor-associated DC in CRC have found associations with improved outcome [35]. Notably, it was demonstrated that the number of infiltrating mature DC was higher in the CRC samples, while the DC density in metastases was markedly lower than in CRC primary tumors [35]. Similar results showed that a significant reduction of the DC number in

total and advanced stage-CRC patients compared to healthy controls and reported that this reduction was totally recovered after complete tumor resection [36].

The limitations of present study are as follow. The limited sample size and follow-up data may have led to selection bias. There was no OS data in validation set but given the sample qualities and size of the set, the DFS was used to evaluate the model value instead. For better clinical application value, further studies with larger sample sizes are needed to support results, and more function explorations should be conducted on the 8 genes in this research. Moreover, basic studies and clinical trials ought to be carried out to verify the predictive efficiency of our model and to identify potential liver metastasis-related mechanisms.

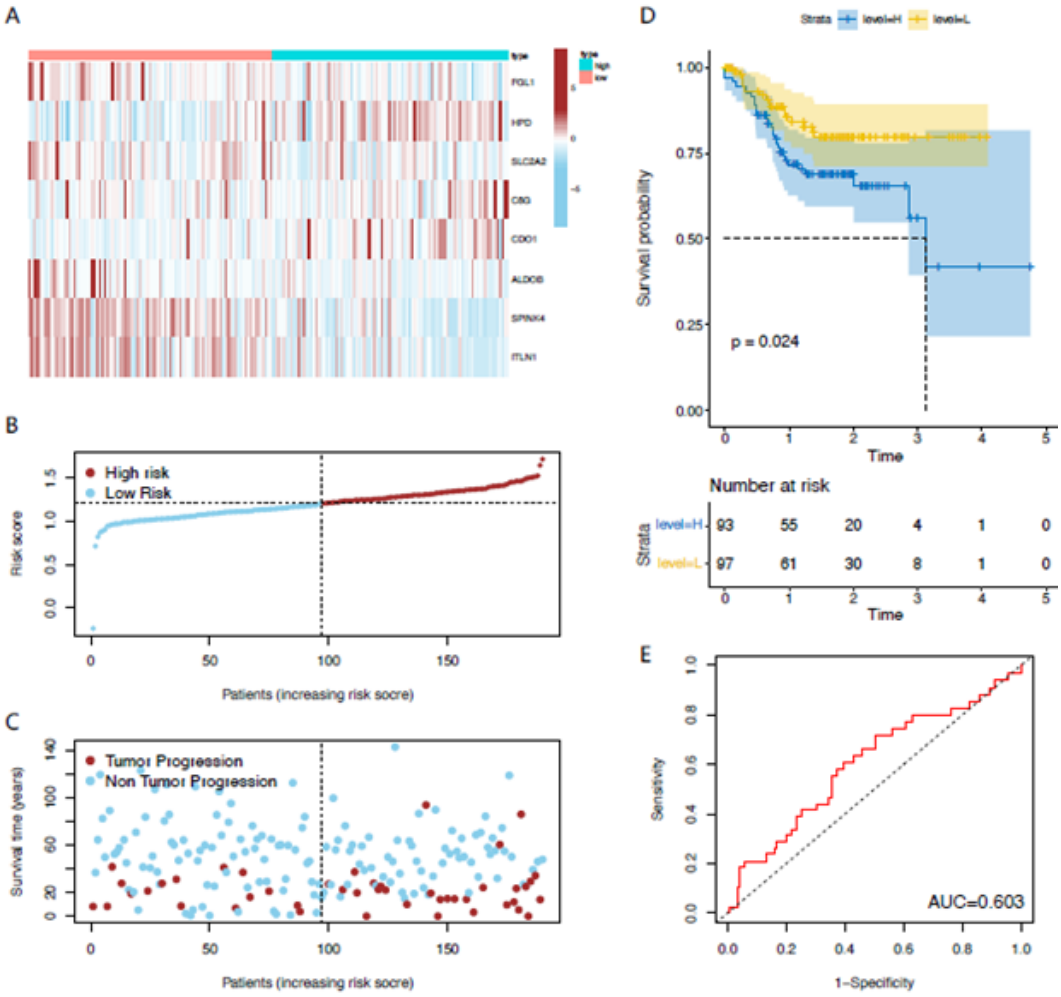


Figure 7: Heatmap demonstrates the expression of 8-signature risk model genes in the high- and low-risk score groups in validation set (A). Risk score distribution (B), scatter plot (C) and Kaplan–Meier OS curves (D) of high-risk group and low-risk group. The ROC curves of risk score model for the prognostic accuracy of CRC patients (E).

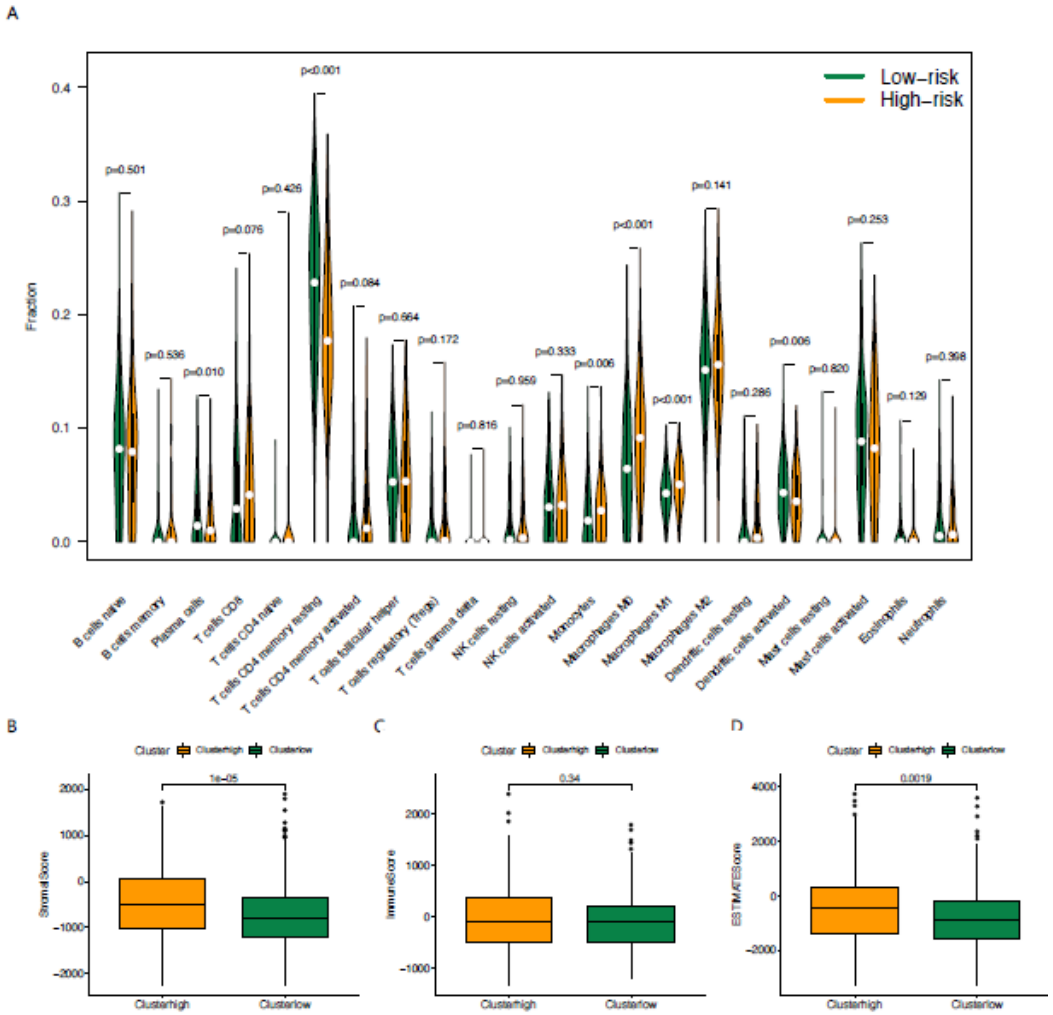
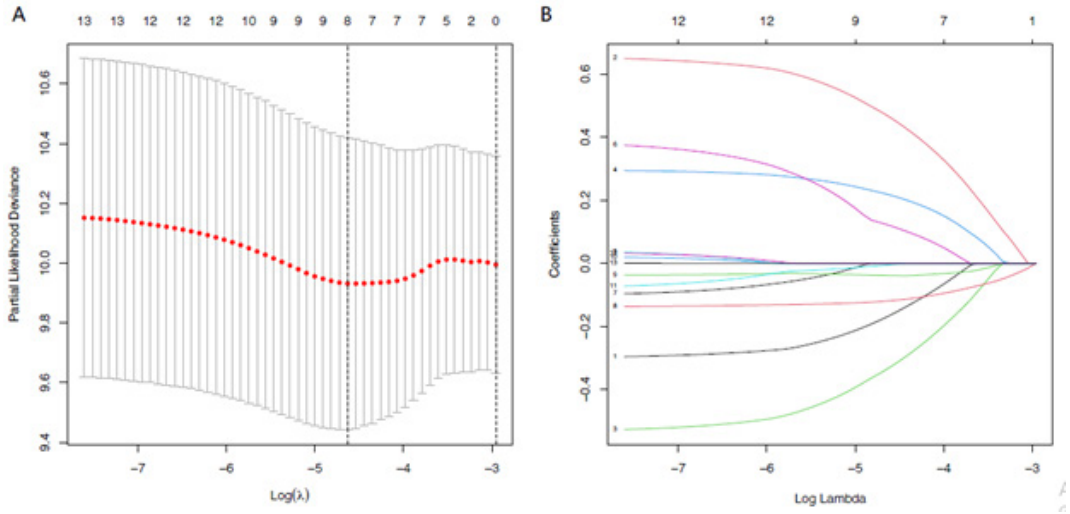


Figure 8: The violin map showed statistical differences between the immune cells of 8-signature risk groups (A). Estimation of the proportion of immune-stromal component. Immune Score, Stromal Score, and ESTIMATE Score (the sum of them) between different risk groups in the TCGA-COREAD cohort (B-D).



Supplementary Figure 1: LASSO Cox regression model to 8 prognostic factors used to construct risk score model in the TCGA cohort. (A) LASSO coefficient profiles of the ex- pression of 5 overlapping genes. (B) Selection of the penalty parameter (Lambda) in the LASSO model via 5-fold cross-validation.

6. Conclusions

In this study, we constructed an 8-gene signature (HPD, C8G, CDO1, FGL1, SLC2A2, ALDOB, SPINK4, ITLN1.) prognostic stratification system based on the DEGs between primary and liver metastasis lesions of CRC, and evaluated the value of the signature. The risk model had better AUC in both the training cohort and the independent validation cohort and was independent of clinical features. Therefore, we recommend this classifier as a molecular diagnostic test to assess the prognostic risk in patients with CRC. And it may have potential significance for new anti-tumor diagnosis and treatment strategies.

7. Funding

This study was supported by Beijing Science and Technology Program [D17110002617004], and Beijing Hope Run Special Fund of Cancer Foundation of China [LC2017A07, LC2019B14].

References

1. Sung H, Ferlay J, Siegel RL. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021; 71: 209-49.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019; 69: 7-34.
3. Yilmaz M, Christofori G. EMT, the cytoskeleton, and cancer cell invasion. Cancer Metastasis Rev. 2009; 28: 15-33.
4. Newman AM, Liu CL, Green MR. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015; 12: 453-7.
5. Zhang S, Zhang E, Long J. Immune infiltration in renal cell carcinoma. Cancer Sci. 2019; 110: 1564-72.
6. Zhou R, Zhang J, Zeng D. Immune cell infiltration as a biomarker for the diagnosis and prognosis of stage I-III colon cancer. Cancer Immunol Immunother. 2019; 68: 433-42.
7. Zeng D, Zhou R, Yu Y. Gene expression profiles for a prognostic immunoscore in gastric cancer. Br J Surg. 2018; 105: 1338-48.
8. Goldman M, Craft B, Swatloski T. The UCSC Cancer Genomics Browser: update 2013. Nucleic Acids Res. 2013; 41: 949-54.
9. Minatani N, Waraya M, Yamashita K. Prognostic Significance of Promoter DNA Hypermethylation of cysteine dioxygenase 1 (CDO1) Gene in Primary Breast Cancer. PLoS One. 2016; 11: 0144862.
10. Igarashi K, Yamashita K, Katoh H. Prognostic significance of promoter DNA hypermethylation of the cysteine dioxygenase 1 (CDO1) gene in primary gallbladder cancer and gallbladder disease. PLoS One. 2017; 12: 0188178.
11. Ushiku H, Yamashita K, Katoh H. Promoter DNA methylation of CDO1 gene and its clinical significance in esophageal squamous cell carcinoma. Dis Esophagus. 2017; 30: 1-9.
12. Ooki A, Maleki Z, Tsay JJ. A Panel of Novel Detection and Prognostic Methylated DNA Markers in Primary Non-Small Cell Lung Cancer and Serum DNA. Clin Cancer Res. 2017; 23: 7141-52.
13. Brait M, Ling S, Nagpal JK. Cysteine dioxygenase 1 is a tumor suppressor gene silenced by promoter methylation in multiple human cancers. PLoS One. 2012; 7: 44951.
14. Schreck SF, Parker C, Plumb ME, Sodetz JM. Human complement protein C8 gamma. Biochim Biophys Acta. 2000; 1482: 199-208.
15. Qian W, Zhao M, Wang R, Li H. Fibrinogen-like protein 1 (FGL1): the next immune checkpoint target. J Hematol Oncol. 2021; 14: 147.
16. Chen G, Feng Y, Sun Z. mRNA and lncRNA Expression Profiling of Radiation-Induced Gastric Injury Reveals Potential Radiation-Responsive Transcription Factors. Dose Response. 2019; 17: 1559325819886766.
17. Sun C, Gao W, Liu J, Cheng H, Hao J. FGL1 regulates acquired resistance to Gefitinib by inhibiting apoptosis in non-small cell lung cancer. Respir Res. 2020; 21: 210.
18. Chiu CF, Hsu MI, Yeh HY. Eicosapentaenoic Acid Inhibits KRAS Mutant Pancreatic Cancer Cell Growth by Suppressing Hepassocin Expression and STAT3 Phosphorylation. Biomolecules. 2021; 11: 370.
19. Son Y, Shin NR, Kim SH, Park SC, Lee HJ. Fibrinogen-Like Protein 1 Modulates Sorafenib Resistance in Human Hepatocellular Carcinoma Cells. Int J Mol Sci. 2021; 22: 5330.
20. Andrews LP, Marciscano AE, Drake CG, Vignali DA. LAG3 (CD223) as a cancer immunotherapy target. Immunol Rev. 2017; 276: 80-96.
21. Lin ST, Tu SH, Yang PS. Apple Polyphenol Phloretin Inhibits Colorectal Cancer Cell Growth via Inhibition of the Type 2 Glucose Transporter and Activation of p53-Mediated Signaling. J Agric Food Chem. 2016; 64: 6826-37.
22. Leong I. ALDOB promotes liver metastases development. Nat Rev Endocrinol. 2018; 14: 380.
23. Bu P, Chen KY, Xiang K. Aldolase B-Mediated Fructose Metabolism Drives Metabolic Reprogramming of Colon Cancer Liver Metastasis. Cell Metab. 2018; 27: 1249-62.
24. Li Q, Li Y, Xu J. Aldolase B Overexpression is Associated with Poor Prognosis and Promotes Tumor Progression by Epithelial-Mesenchymal Transition in Colorectal Adenocarcinoma. Cell Physiol Biochem. 2017; 42: 397-406.
25. Wang X, Yu Q, Ghareeb WM. Downregulated SPINK4 is associated with poor survival in colorectal cancer. BMC Cancer. 2019; 19: 1258.
26. Chen TJ, Tian YF, Chou CL. High SPINK4 Expression Predicts Poor Outcomes among Rectal Cancer Patients Receiving CCRT. Curr Oncol. 2021; 28: 2373-84.
27. Au-Yeung CL, Yeung TL, Achreja A. ITLN1 modulates invasive potential and metabolic reprogramming of ovarian cancer cells in omental microenvironment. Nat Commun. 2020; 11: 3546.
28. Kawashima K, Maeda K, Saigo C, Kito Y, Yoshida K, Takeuchi T, et al. Adiponectin and Intelectin-1: Important Adipokine Players in Obesity-Related Colorectal Carcinogenesis. Int J Mol Sci. 2017; 18: 886.
29. Katsuya N, Sentani K, Sekino Y. Clinicopathological significance of intelectin-1 in colorectal cancer: Intelectin-1 participates in tumor suppression and favorable progress. Pathol Int. 2020; 70: 943-52.
30. Giraldo NA, Sanchez-Salas R, Peske JD. The clinical role of the TME in solid cancer. Br J Cancer. 2019; 120: 45-53.

31. Qian BZ, Li J, Zhang H. CCL2 recruits inflammatory monocytes to facilitate breast-tumour metastasis. Nature. 2011; 475: 222-5.

32. Wu Z, Chen H, Luo W. The Landscape of Immune Cells Infiltrating in Prostate Cancer. Front Oncol. 2020; 10: 517637.

33. Zhang E, Dai F, Mao Y. Differences of the immune cell landscape between normal and tumor tissue in human prostate. Clin Transl Oncol. 2020; 22: 344-50.

34. Bottcher JP, Reis e Sousa C. The Role of Type 1 Conventional Dendritic Cells in Cancer Immunity. Trends Cancer. 2018; 4: 784-92.

35. Gulubova MV, Ananiev JR, Vlaykova TI, Yovchev Y, Tsoneva V, Manolova IM, et al. Role of dendritic cells in progression and clinical outcome of colon cancer. Int J Colorectal Dis. 2012; 27: 159-69.

36. Orsini G, Legitimo A, Failli A. Defective generation and maturation of dendritic cells from monocytes in colorectal cancer patients during the course of disease. Int J Mol Sci. 2013; 14: 22022-41.