

RESEARCH

Open Access



Identification and validation of a novel signature as a diagnostic and prognostic biomarker in colorectal cancer

Di Wang^{1†}, Junye Liufu^{1†}, Qiyuan Yang¹, Shengqun Dai¹, Jiaqi Wang² and Biao Xie^{1*}

Abstract

Background Colorectal cancer (CRC) is one of the most common malignant neoplasms worldwide. Although marker genes associated with CRC have been identified previously, only a few have fulfilled the therapeutic demand. Therefore, based on differentially expressed genes (DEGs), this study aimed to establish a promising and valuable signature model to diagnose CRC and predict patient's prognosis.

Methods The key genes were screened from DEGs to establish a multiscale embedded gene co-expression network, protein-protein interaction network, and survival analysis. A support vector machine (SVM) diagnostic model was constructed by a supervised classification algorithm. Univariate Cox analysis was performed to construct two prognostic signatures for overall survival and disease-free survival by Kaplan–Meier analysis, respectively. Independent clinical prognostic indicators were identified, followed by univariable and multivariable Cox analysis. GSEA was used to evaluate the gene enrichment analysis and CIBERSORT was used to estimate the immune cell infiltration. Finally, key genes were validated by qPCR and IHC.

Results In this study, four key genes (DKC1, FLNA, CSE1L and NSUN5) were screened. The SVM diagnostic model, consisting of 4-gene signature, showed a good performance for the diagnostic (AUC = 0.9956). Meanwhile, the four-gene signature was also used to construct a risk score prognostic model for disease-free survival (DFS) and overall survival (OS), and the results indicated that the prognostic model performed best in predicting the DFS and OS of CRC patients. The risk score was validated as an independent prognostic factor to exhibit the accurate survival prediction for OS according to the independent prognostic value. Furthermore, immune cell infiltration analysis demonstrated that the high-risk group had a higher proportion of macrophages M0, and T cells CD4 memory resting was significantly higher in the low-risk group than in the high-risk group. In addition, functional analysis indicated that WNT and other four cancer-related signaling pathways were the most significantly enriched pathways in the high-risk group. Finally, qRT-PCR and IHC results demonstrated that the high expression of DKC1, CSE1L and NSUN5, and the low expression of FLNA were risk factors of CRC patients with a poor prognosis.

[†]These two authors contributed equally to this article.

*Correspondence:

Biao Xie
xiebiaohxq@163.com

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Conclusion In this study, diagnosis and prognosis models were constructed based on the screened genes of DKC1, FLNA, CSE1L and NSUN5. The four-gene signature exhibited an excellent ability in CRC diagnosis and prognostic prediction. Our study supported and highlighted that the four-gene signature is conducive to better prognostic risk stratification and potential therapeutic targets for CRC patients.

Keywords Colorectal cancer, Differentially expressed genes, Diagnostic model, Prognostic model

Background

Colorectal cancer (CRC) is a common malignant tumor worldwide and also one of the leading causes of cancer-related mortality [1]. The mortality rate of CRC is second and third highest-ranking cancer in male and female patients, respectively, in the USA and fifth in China [2, 3]. The 5-year survival rate of CRC patients is exceed 90% when diagnosed at early stages. Due to the lack of adequate diagnostic methods, CRC is often diagnosed at an advanced stage, and the 5-year survival rate for CRC patients diagnosed with metastasis is low at approximately 12% [4]. At present, despite the significant improvements in diagnosis and treatment, the early diagnosis of CRC continues to be a global problem [5]. It is urgently needed that more effective diagnosis and prognostic evaluation systems to provide personalized medicine and improve outcome for CRC patients [6]. With the development of biology, biomarkers exhibited an increasingly important role in the early diagnosis, prognostication, survival, and clinical treatment monitoring of cancers. They have driven the development of personalized therapy and had a positive impact on patient outcomes [7]. Therefore, it is significant to identify and explore sensitive diagnostic and prognostic biomarkers for CRC.

With the advancement of gene chips and high-throughput second-generation sequencing technologies, the amount of publicly available high-throughput data is stored in global databases. Therefore, the combination of gene expression data with bioinformatics methods can be used to elucidate the expression of differentially expressed genes (DEGs) in the development and progression of CRC, as well as identify potential targets for the treatment of CRC. Recent studies have proposed DEGs as potential diagnostic and prognostic markers in CRC. Huang et al. identified hundreds of CRC-associated DEGs based on the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) database. Thereinto, five could be used as diagnostic biomarkers for CRC patients [8]. Hou et al. identified a cluster of DEGs and DNA methylation aberrations in CRC. The results indicated that the combination of DEGs, DNA methylation aberrations, and tumor stages results in effective prognostic evaluation of patients with CRC [9]. Although some multigene-based prognostic signatures that can assess the prognostic risk of CRC had been established over the past 20 years, the robustness of most markers

is less than expected and rarely could be used for CRC diagnosis, due to the inherent genetic heterogeneity of CRC [7, 10]. Therefore, finding effective and reliable signatures for the diagnosis and prognosis assessment of CRC patients is paramount and urgent.

In this study, four hub genes were identified from the DEGs by constructing a multiscale embedded gene co-expression network analysis (MEGENA), protein-protein interaction (PPI) network, and survival analysis. Then, to explore whether these key genes related to the diagnosis and prognosis of CRC, we constructed an SVM model and established a multi-gene signature for the prognosis of CRC patients by univariate and multivariable Cox regression analyses. This model adequately predicted the overall survival (OS) and disease-free survival (DFS) of CRC patients and provided a theoretical basis for the future study of CRC diagnosis and the underlying molecular mechanism.

Results

Identification of DEGs

The current study is illustrated in the flow chart (Fig. 1). In order to determine the DEGs between the CRC tumors and adjacent tissues, $p < 0.01$ and $|\log_2\text{FC}| > 2$ were used as the threshold. The DEGs were displayed in the heatmap and volcano plot (Fig. 2 A and B). The results showed that compared with paracancerous tissue, 3409 genes were significantly upregulated and 3410 genes were significantly downregulated in CRC tissue.

Identification of node genes

To identify the compact gene modules based on the 6819 DEGs, MEGENA method was used to construct a gene-gene causal network. The results showed that a total of 237 genes were cluster analyzed to identify the DEGs (Fig. 2 C). Finally, these gene networks identified 337 highly connected DEGs for downstream analyses.

PPI network analysis of DEGs

The STRING platform was used to construct a PPI network of 337 DEGs (Fig. 2D). A total of 61 nodes were included in the network based on the interaction score criteria ($\text{degree} > 10$). A total of 61 genes (MCM7, AURKA, MCM2, KIF23, PLK1, TPX2, ANLN, BIRC5, BUB1, CD4, CDC20, CDK4, WDR12, RUVBL1, KIF2C, NCAPH, CDCA, GTPBP4, TTK, DKC1, PKMYT1, NOP58, RRP9, DDX56, DCAF13, METTL, FTSJ1,

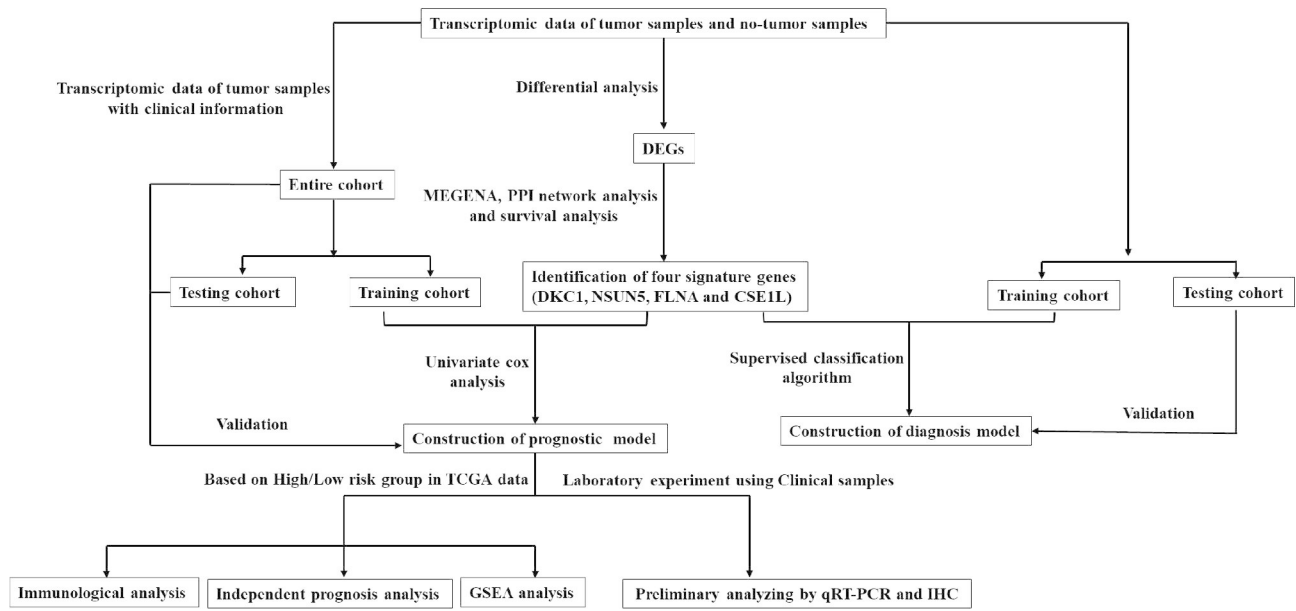


Fig. 1 Flowchart of the study

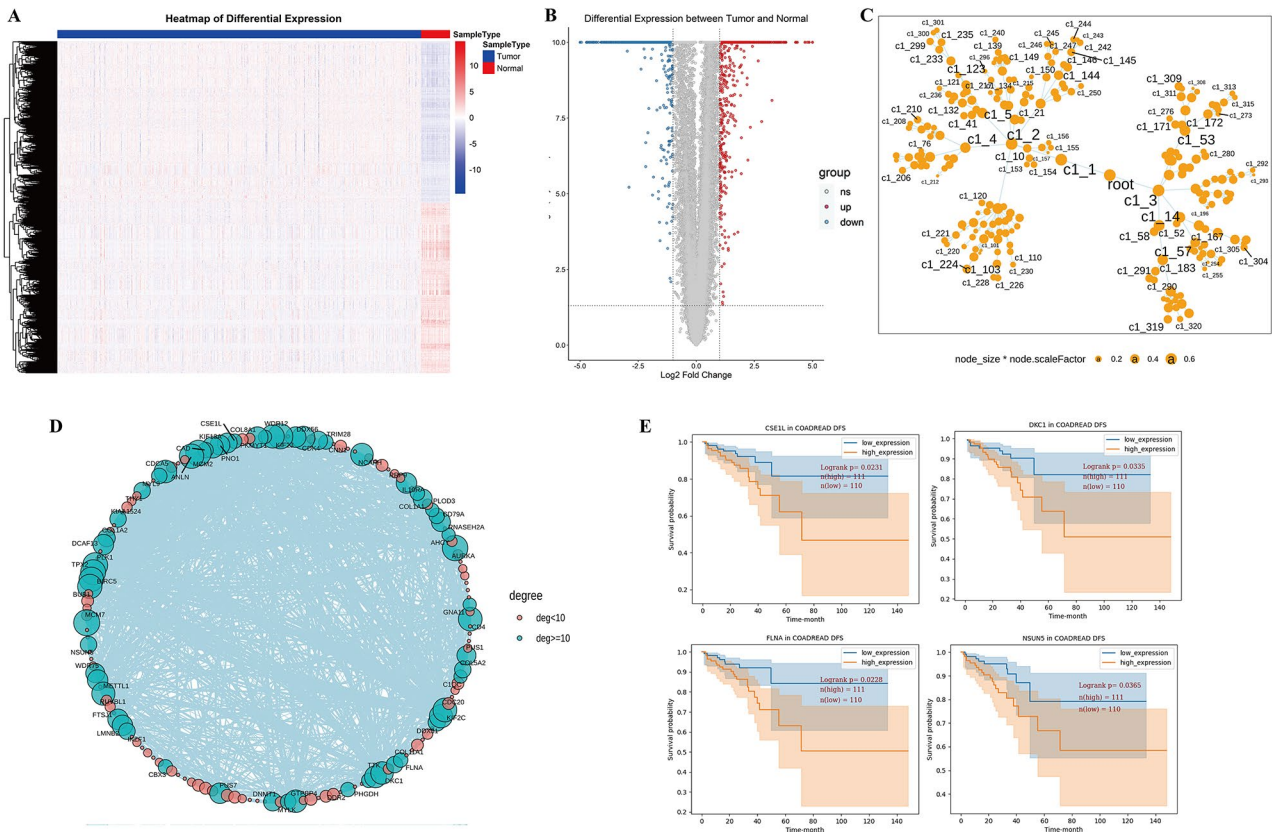


Fig. 2 Identification of a four-gene signature in CRC patients. (A) Heatmap of significant DEGs based on the expression level; (B) The volcano figure to show the upregulated and downregulated genes; (C) MEGENA among differentially expressed genes; (D) PPI network among node genes from the MEGENA; (E) Kaplan–Meier curves with the log-rank test were performed on disease free survival analysis of hub genes (DKC1, NSUN5, FLNA and CSE1L) from the PPI network

LMNB2, PUS7, KIF18A, PNO1, RNASEH2A, COL1A1, WDR75, PUS1, TRIM28, DNMT1, COL1A2, DDX31,

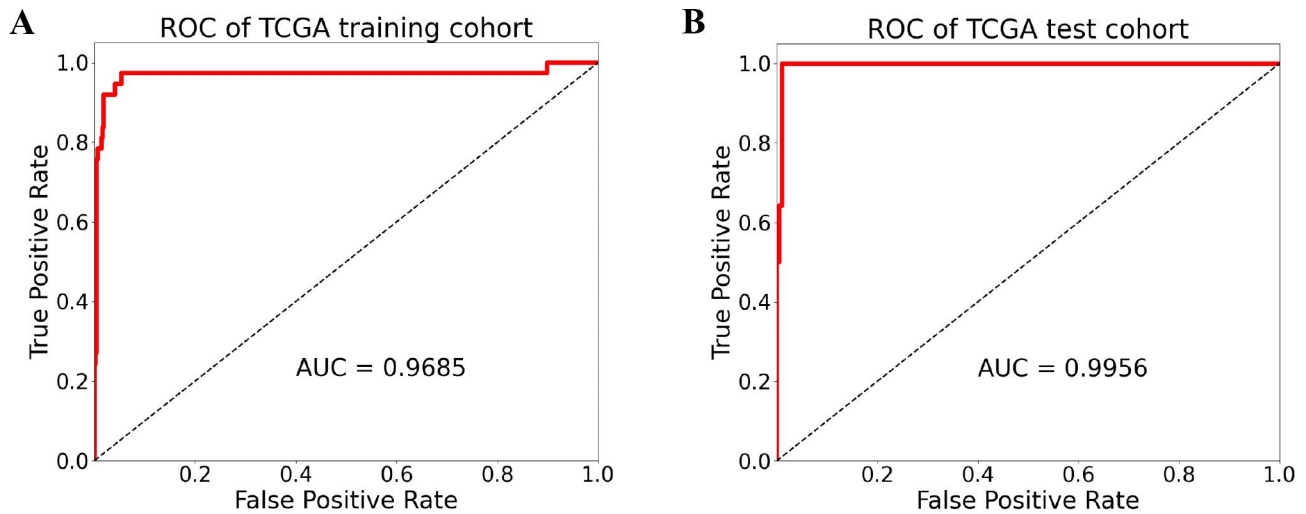


Fig. 3 Establishment and validation of the diagnostic model of SVM. (A) ROC curve of SVM based on four genes as features based on training data. The x-axis indicates the false positive rate, and the y-axis indicates the true positive rate. The five-fold cross-validation is represented. The final fitted average is denoted by the black dotted line; (B) The ROC curve analysis for model efficacy based on testing data. The x-axis represented the FPR, and the y-axis represented the TPR. ROC: receiver operating characteristic, FPR: false-positive rate, TPR, true-positive rate

THY1, KIAA1524, NSUN5, IKZF1, FLNA, COL5A2, CD79A, CAD, MYL9, MYLK, IL10RA, CNN1, CSE1L, CBX3, DDR2, PHGDH, COL11A1, GNA11, AHCY- PLOD3, COL8A1 and C1QC) were selected as the hub genes.

Survival analysis of hub genes

To further identify the key genes, we analyzed the correlation between hub gene expression and bioinformatics of CRC, and found that the DFS rate of high-expressing DKC1 was lower than that of the low-expressing molecule ($p=0.0335$). Compared to the low-expressing NSUN5, the DFS rate of high-expressing NSUN5 was lower ($p=0.0365$). Compared to the low-expressing FLNA, the DFS rate of the high-expression group of FLNA was lower ($p=0.0228$). Meanwhile, the DFS rate of the high-expression of CSE1L was lower than the low-expressing CSE1L ($p=0.0231$). No significant correlation was established between the expression of the other 57 hub genes and CRC prognosis ($p>0.05$). These results identified DKC1, NSUN5, FLNA and CSE1L as the four key genes (Fig. 2E).

Establishment and validation of the diagnostic model

All four genes were statistically significant ($p<0.05$) in the COADREAD dataset from TCGA after survival analysis. To construct and validate the four-gene diagnostic model, a total of 689 CRC samples in COADREAD datasets from TCGA were divided into the training ($n=482$) and test sets ($n=207$). A python 3.8 package scikit-learn was employed to construct the four-gene diagnostic model of SVM using the training set. The overall accuracy (fraction of correctly classified samples) of the binomial

Table 1 Univariate regression analysis of four-signature genes for DFS

Symbol	Beta	Hazard ratio (95% CI)	Wald Test	p-value
DKC1	0.6374	1.8916(0.7266–4.9248)	1.3057	0.1917
NSUN5	0.7798	2.1811(0.8753–5.4346)	1.6741	0.0941
FLNA	0.2717	1.3122(0.9170–1.8777)	1.4860	0.1373
CSE1L	-0.2354	0.7902(0.3783–1.6507)	-0.6264	0.5311

classifier for cross-validation on the training datasets was 0.9685 (Fig. 3 A), and the optimal parameter combination of the diagnostic model was determined (“C”: 23.2676, “gamma”: 0.4498). To further validate the current model, we evaluated the classifier on the test datasets; the area under curve (AUC) was 0.9956 (Fig. 3B). These results indicated that the constructed SVM-based classifier had high diagnostic values for the CRC patients.

Establishment and validation of prognostic model

A total of 221 CRC patients with the DFS data collected from the merged cohort of the TCGA and clinical information (DFS) were recruited, the detailed clinical features of ccRCC patients were listed in Supplementary material Table S1. Four genes were selected by survival analysis for constructing the prognostic model of DFS. The risk score formula consisting of the four DEGs was established for the prognostic model as follows: Risk score = $(0.6374 \times \text{DKC1}) + (0.7798 \times \text{NSUN5}) + (0.2717 \times \text{FLNA}) + (-0.2354 \times \text{CSE1L})$ (Table 1). Then, 221 CRC patients were divided into the training set ($n=111$) and

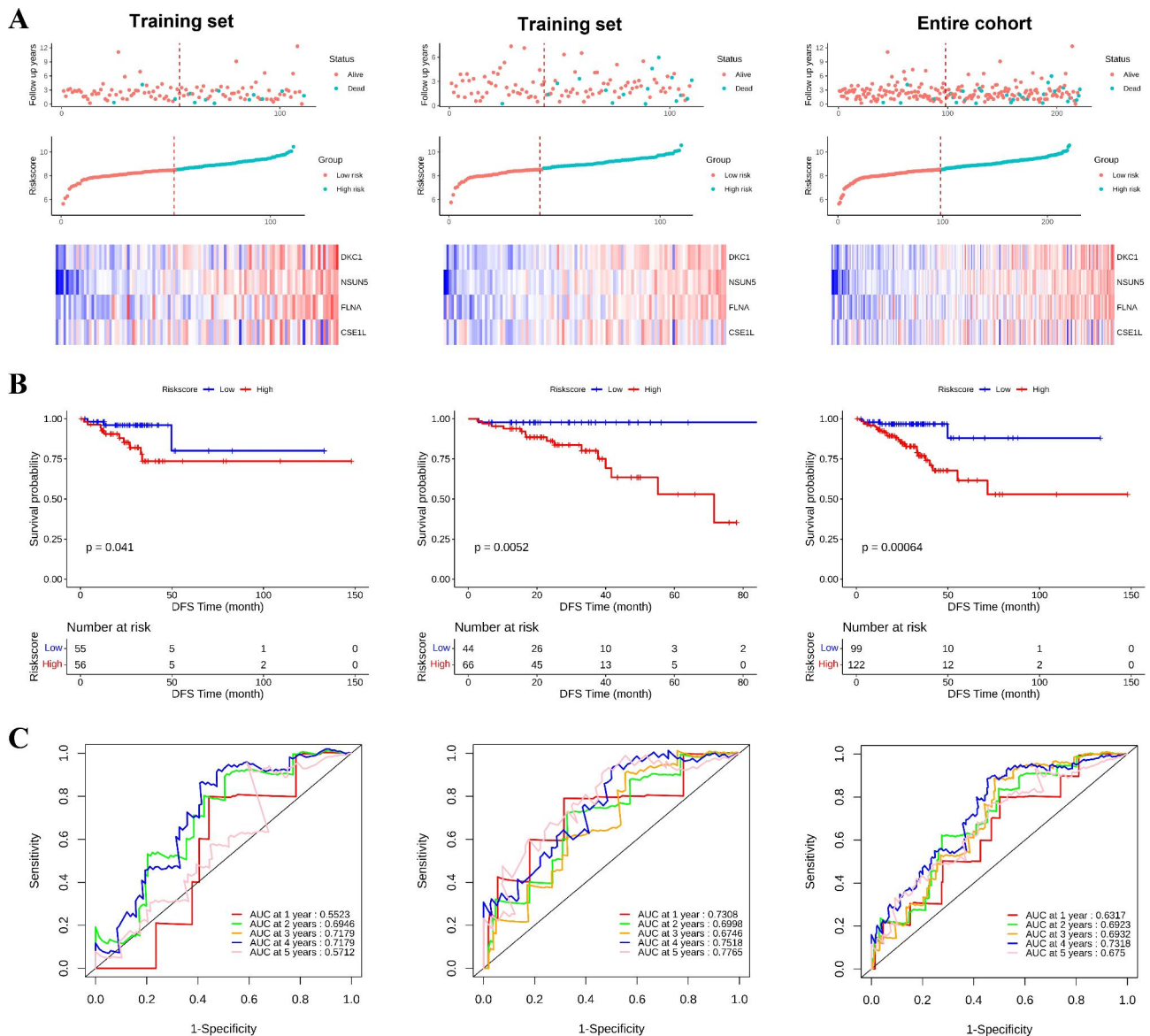


Fig. 4 Establishment and validation of the prognostic model for DFS of patients in the training set, the test set, and the entire cohort. (A) The distribution of risk scores, gene expression levels, and patient relapse status; (B) Kaplan–Meier curves of DFS of the low-risk and high-risk groups; (C) ROC curve for the 1-, 2-, 3-, 4-, and 5-year survival prediction by the four-gene signature, respectively. The black dotted line represents the median risk score cutoff dividing the patients into low- and high-groups

test set (n=110). Next, we presented the risk score distribution curve and survival status of the training set (Fig. 4 A). The results showed that there were more deaths in the high-risk group, while most of the patients in the low-risk group stayed alive during the follow-up. The heatmap revealed the expression patterns of four DEGs between the two risk groups (Fig. 4 A). Based on the median prognostic risk score, the patients in the training set were divided into either the high-risk group (n=56) or the low-risk group (n=55). The K–M survival curve suggested that the patients of high prognostic risk group had poor DFS than the low risk patients ($p < 0.05$, Fig. 4B). The time-dependent ROC curve analysis was

used to evaluate the predictive value of the signature. The AUCs of the training set at 1, 2, 3, 4, and 5 years were 0.5523, 0.6946, 0.7179, 0.7179, and 0.5712, respectively (Fig. 4 C). The formula mentioned above was used to calculate the risk scores, and the patients were also classified into high- and low-risk groups in the test and total sets based on the median value of 8.5343. The risk score distribution curve, survival status, and four genes expression heatmap of the test and total sets are shown in Fig. 4 A. Similarly, the predictive capability and clinical utility of the prognostic model was validated in both datasets. The results showed that the DFS of the high-risk group was significantly shorter than that of the low-risk

Table 2 Univariate regression analysis of four-signature genes for OS

Symbol	Beta	Hazard ratio (95% CI)	Wald test	p-value
DKC1	-0.1051	0.9002(0.5513–1.4700)	-0.4202	0.6744
NSUN5	0.3303	1.3915(0.9472–2.0441)	1.6835	0.0923
FLNA	0.0960	1.1007 (0.9529–1.2714)	1.3046	0.1920
CSE1L	-0.2391	0.7873 (0.5377–1.1529)	-1.2286	0.2192

group in both the test and totals set, according to the K–M survival analysis ($p < 0.05$, Fig. 4B), and the AUC for 1-, 2-, 3-, 4-, and 5-year DFS was 0.7308, 0.6998, 0.6746, 0.7518, and 0.7765 in the test set, respectively (Fig. 2E). In the total set, the AUCs for 1-, 2-, 3-, 4-, and 5-year DFS were 0.6317, 0.6923, 0.6932, 0.7318, and 0.675, respectively (Fig. 4 C). These results indicated the four-gene prognostic signature had the ability to predict DFS.

Simultaneously, to determine the association between the gene expression signatures of these four genes and the CRC patient OS, the risk score of the four DEGs was calculated using univariate Cox regression analysis as follows: Risk score = $(-0.1051 \times \text{DKC1}) + (0.3303 \times \text{NSUN5}) + (0.0960 \times \text{FLNA}) + (-0.2391 \times \text{CSE1L})$ (Table 2). The detailed clinical features of ccRCC patients were listed in Supplementary material Table S1. A total of 587 patients with OS data were divided into the training set ($n = 294$) and test set ($n = 293$). Based on the median risk score of -0.0477 , the training set ($n = 294$), the test set ($n = 293$), and the total set ($n = 587$) were divided into the high- and low-risk groups. The risk score distribution curve, survival status, and the four genes' expression heatmap of the training, test, and total sets are shown in Fig. 5 A. The results showed that the patients of high prognostic risk group had a higher mortality than the patients of low prognostic risk group. The heatmap revealed the expression patterns of four DEGs between the two risk groups. Compared to the low-risk group, a significantly poor OS was observed in the high-risk group by the K–M survival curve analysis ($p < 0.05$, Fig. 5B). We also performed the time-dependent ROC curve analysis, and the AUCs for 1-, 2-, 3-, 4-, and 5-year DFS were 0.6532, 0.6479, 0.5693, 0.598, and 0.5723, respectively, in the test set (Fig. 5 C). In the total set, the AUCs for 1-, 2-, 3-, 4-, and 5-year DFS were 0.641, 0.6793, 0.5992, 0.57, and 0.5557, respectively (Fig. 5 C). These results indicated that the prognostic model performed best in predicting the OS risk of CRC patients.

Gene set enrichment analyses (GSEA) of the prognostic signature

We performed GSEA on the two risk groups to further explore the critical KEGG pathways. The results showed that 164 KEGG enriched pathways were active in the high-risk group, while 13 were active in the low-risk group. Furthermore, 101 statistically significant KEGG

pathways ($p < 0.05$, $\text{FDR} < 0.25$) were screened, among which 100 were active in the high-risk group, while only one was active in the low-risk group. The top 10 pathways with the highest NES (normalized enrichment score) in the high-risk group and one pathway in the low-risk group (Table 3) were selected for visualization analysis (Fig. 6 A–B). The results showed that focal adhesion [11], ECM receptor interaction [12], AXON guidance [13], basal cell carcinoma [14], and WNT signaling pathway [15], which were considered to be associated with the poor prognosis of CRC, were the most significantly pathways enriched in the high-risk group.

Independent prognosis analysis in the TCGA dataset

Independent prognosis analyses of clinical characteristics (age, gender, tumor stage and pathological TNM (tumor-node-metastasis)) and risk score determined the patient survival. Univariate and multivariate Cox regression analyses were performed on the clinical characteristics and risk scores of the overall samples from the TCGA dataset. As shown in Table 4, the univariate Cox regression analyses evaluated the prognostic value of the signature (high risk/low risk, $p = 0.0022$), N stage (N0/N1, $p = 0.0328$), and gender (female/male). However, the multivariate Cox regression analysis showed that risk score and gender were significantly related to prognosis for OS. Therefore, the results of univariate Cox and further multivariate Cox analyses validated that risk score was an independent prognostic factor for CRC. These results suggested that the prognostic signature was an accurate model for predicting prognosis of patients with CRC.

Analysis of immune cell infiltration of the prognostic signature

Increasing evidence shows that the prognosis and clinical outcome of CRC patients are strongly influenced by immune cell infiltration in the tumor microenvironment [16]. Comprehensive evaluation of immune infiltrates integrating both the quantity and variety of tumor-infiltrating immune cells and incorporation of composite scores encompassing clinically biomarkers, is emerging as a promising strategy potentially capable of risk stratification and optimizing patient selection [17]. Thus, to further investigate the correlation between immune cell infiltration and the prognostic signature, we used the CIBERSORT algorithm to calculate the content of 22 immune cell populations in each CRC sample by setting $p < 0.05$ as the threshold for screening. The results showed that macrophages M0 were the most abundant infiltrating cells, followed by CD4 memory T cells (Fig. 7 A, B). We also observed differences in immune cell infiltration between the two risk groups. As the results showed in Fig. 7 C, the proportion of T cells gamma delta, eosinophils, neutrophils, macrophages M2, and CD4 memory

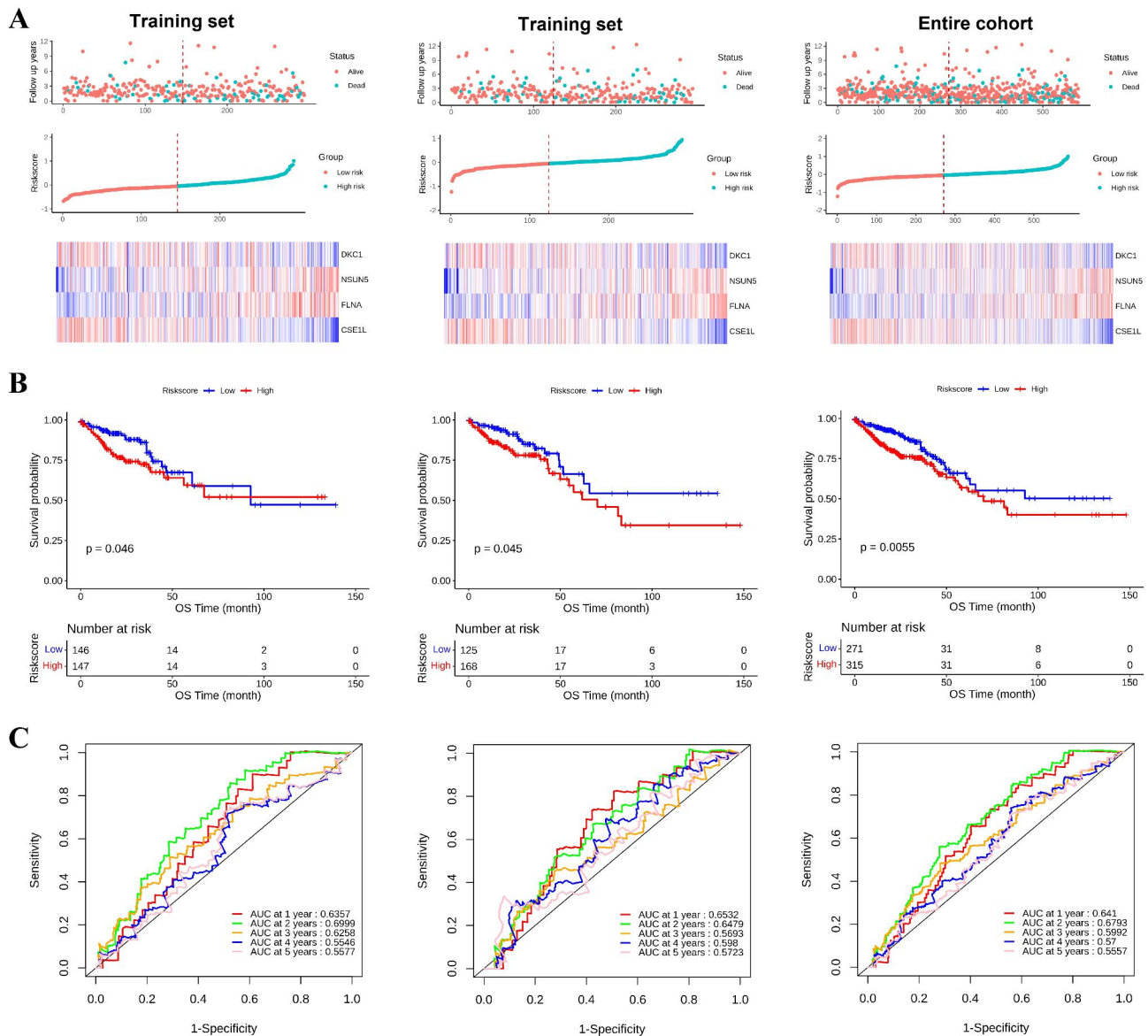


Fig. 5 Establishment and validation of the prognostic model for OS of patients in the training set, the test set, and the entire cohort. (A) The distribution of risk scores, gene expression levels, and patient relapse status; (B) Kaplan–Meier curves of DFS of the low- and high-risk groups; (C) ROC curve for the 1-, 2-, 3-, 4-, and 5-year survival prediction by the four-gene signature, respectively. The black dotted line represents the median risk score cutoff dividing patients into low- and high- groups

T cells were significantly enriched in the low-risk group compared to the high-risk group ($p < 0.05$), while a high proportion of regulatory T cells (Tregs), naïve B cells, and M0 macrophages were significantly in the high-risk group than in the low-risk group ($p < 0.05$).

Experimental verification of key genes by qRT-PCR and IHC

In order to further analyze the expression of the hub genes in patients, CRC and adjacent normal tissues of patients were collected (Supplemental material Table S2). To analyze the mRNA levels of DKC1, FLNA, CSE1L and NSUN5, qRT-PCR was performed. The mRNA levels of DKC1, CSE1L and NSUN5 were higher in the colon

cancer tissues than in the paired adjacent normal tissues, while the level of FLNA was lower ($p < 0.05$, Fig. 8 A). IHC staining of the five pairs of CRC and adjacent normal tissues showed that DKC1, CSE1L and NSUN5 expression levels were significantly higher in the CRC tissues compared to the paired adjacent normal tissue, while FLNA protein level was significantly lower in the CRC tissues (Fig. 8B). Taken together, these results suggested that DKC1, CSE1L and NSUN5 were dramatically overexpressed while FLNA was significantly downregulated in colorectal cancer.

Table 3 Ten KEGG pathways in the high-group and one in the low-group

Name	NES	p-val	FDR
High-risk group			
KEGG_FOCAL_ADHESION	2.9114	0	0
KEGG_ECM_RECEPTOR_INTERACTION	2.8482	0	0
KEGG_AXON_GUIDANCE	2.6994	0	0
KEGG_VASCULAR_SMOOTH_MUSCLE_CONTRACTION	2.6843	0	0
KEGG_BASAL_CELL_CARCINOMA	2.6754	0	0
KEGG_SPLICEOSOME	2.6402	0	0
KEGG_WNT_SIGNALING_PATHWAY	2.5723	0	0
KEGG_MELANOGENESIS	2.5571	0	0
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	2.5567	0	0
KEGG_PATHWAYS_IN_CANCER	2.5515	0	0
Low-risk group			
KEGG_ASCORBATE_AND_ALDARATE_METABOLISM	-	0.0143	0.172
	1.5729		

Discussion

CRC is the most common cancer worldwide, with the third-highest mortality rate [18]. It has been one of the leading causes of cancer-related deaths, and hence, the prognosis of CRC has always been a major concern. Over the past few decades, although the traditional therapies and combination chemotherapy has had a significant effect on the CRC, the treatment of recurrent and metastatic CRC patients did not improve [19–21]. Meanwhile, there is a lack of research on the molecular mechanism associated with metastasis of CRC [22–24]. Therefore, developing new biomarkers with high prognostic values to estimate the treatment response and survival outcomes of CRC patients is crucial. In this study, bioinformatics was used to screen the CRC-related genes based on the gene expression data of CRC patients in the TCGA database. A total of 6819 DEGs were explored from the TCGA database, followed by MEGENA and PPI to deduce the gene co-expression networks. A total of 337 candidate genes were identified. Finally, based on the survival analysis of the 337 candidate genes, we screened

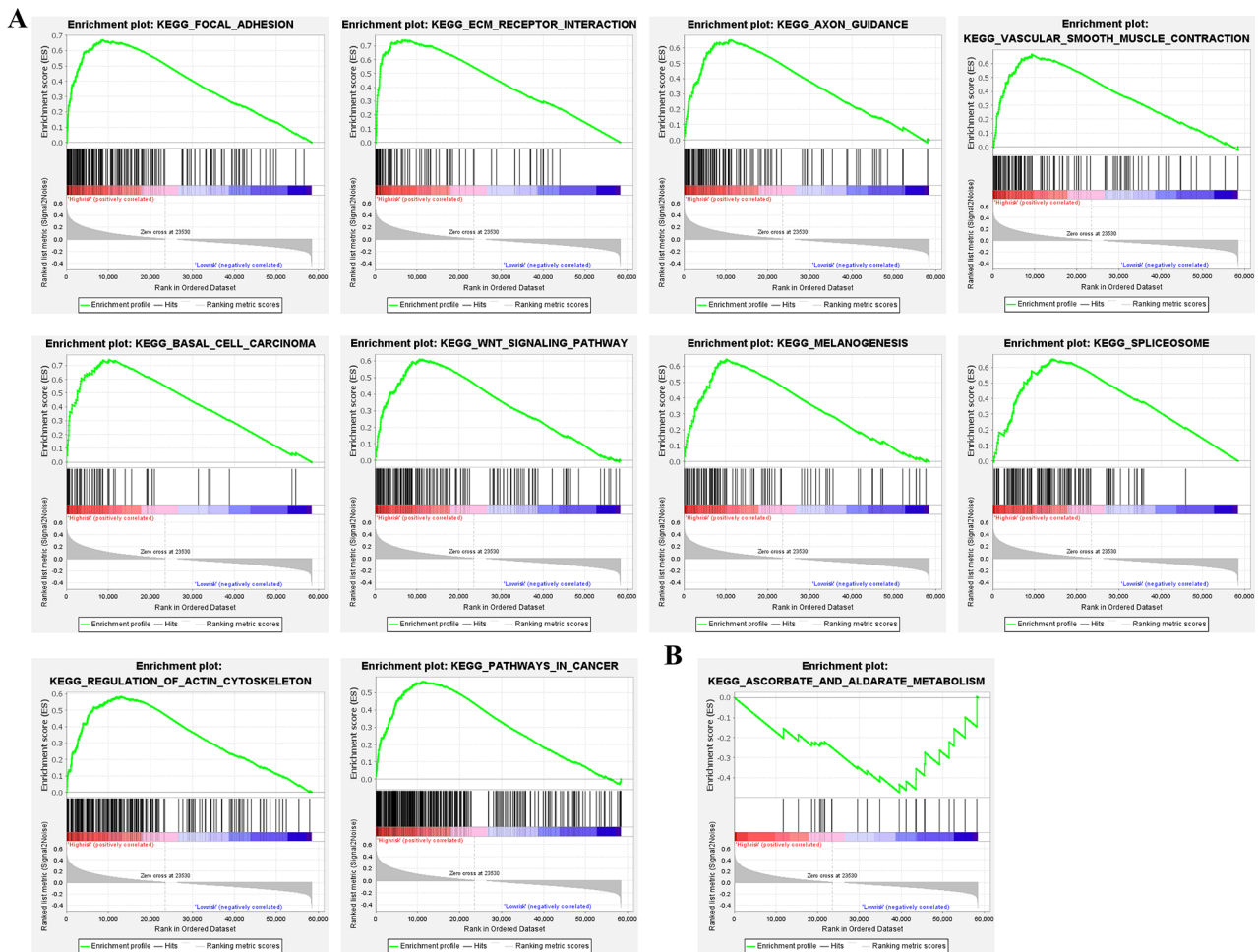


Fig. 6 Functional enrichment analysis of genes correlated with signature genes in the high- and low-risk groups via GSEA. (A) Top 10 of KEGG enrichment analysis of signature genes in the high-risk group; (B) Only one in the low-risk group

Table 4 Analysis of clinical characteristics and risk-score for the effectiveness of the prognostic signature

Variables	Univariate analysis			Multivariable analysis		
	HR	95% CI of HR	p	HR	95% CI of HR	p
Risk score						
Low-risk						
High-risk	5.1908	1.8055–14.924	0.0022	4.838	1.6668–14.044	0.0037
Age						
< 60						
≥ 60	0.8458	0.3945–1.8134	0.667			
Stage						
I-II						
III-IV	1.1476	0.526–2.5039	0.7294			
N stage						
N0						
N1	2.662	1.0831–6.5424	0.0328	2.09	0.8423–5.184	0.1119
T stage						
T1-T2						
T3-T4	2.1384	0.7408–6.1725	0.1599			
gender						
Female						
Male	2.2676	1.0336–4.9749	0.0411	2.378	1.0811–5.232	0.0313

four-gene expression signatures, including DKC1, NSUN5, FLNA and CSE1L.

Diagnosis and prediction are the most important steps in the management of CRC patients. To screen the candidate biomarkers that may be helpful for the diagnosis and prognosis for CRC, a four-gene signature was identified by our data processing system. Then, these genes were used to construct a diagnostic model. The AUCs of the four-gene signature showed a perfect diagnostic ability in CRC gene expression samples in the training and test sets from the TCGA database. Next, to provide a robust indicator for the prognostic evaluation on the DFS and OS with CRC, we also constructed a prognostic model and illustrated the impact of the prognostic signature for CRC patients, based on a combination of the risk score distribution, survival status scatter plot, and gene expression heatmap. In addition, the expressions of DKC1, NSUN5, FLNA and CSE1L were validated by qRT-PCR and IHC, and found that DKC1, CSE1L and NSUN5 were highly expressed and FLNA was underexpressed in CRC tumor tissues, consistent with the data in TCGA database. These results suggested that expression changes of DKC1, NSUN5, FLNA and CSE1L could be associated with the prognosis of CRC patients. Because the number of samples collected in this study is limited, it can be used as a basis for prognosis and diagnosis of the hub genes in CRC patients. Large clinical specimens to further evaluate the value of the four hub genes as a prognostic and diagnostic evaluation indicator of CRC is our next major study objectives.

DKC1 is an X-linked gene which encodes dyskeratosis protein on the X chromosome (Xq28) and participates

in the occurrence of congenital keratosis [25]. DKC1 is a major component of telomerase ribonucleoprotein complex and has a major impact on the functional stability of telomerase ribonucleoprotein complex. The upregulation of DKC1 is closely related to poor prognosis in prostate cancer, neuroblastoma, and hepatocellular carcinoma [26–28]. FLNA is located on the X chromosome and is a critical signal transduction scaffold protein [29, 30]. Previous studies have shown that FLNA promotes cell proliferation, migration, and invasion [31]. However, some studies have reported that FLNA is significantly upregulated in cervical cancer, and has a good predictive effect, which can be used as a prognostic signature of cervical cancer [32]. The involvement of chromosome 20 in human cancers is well-documented. CSE1L is mapped to 20q13, which encodes a protein that mediates the nuclear export of importin- α [33, 34]. CSE1L is differentially expressed in various malignant tumors and is related to the ability of tumor invasion, metastasis, and proliferation [35–37]. NSUN5 is generally upregulated in human cancers, including CRC, which may be attributed to the hypomethylation of NSUN5 promoter. The cancer patients with an overexpressed NSUN5 have a poorer prognosis, and this condition is positively correlated with NSUN5 translation. Epidemiologic study demonstrated that high expression of NSUN5 was associated with advanced tumor stages (III, IV), which possibly relate to its promotion of cell proliferation through regulating cell cycle in CRC [38]. Our study provides a different approach to establishing the prediction model and selecting the candidate oncogenes or biomarkers that exhibit a marked capacity for diagnosis and prognosis of CRC.

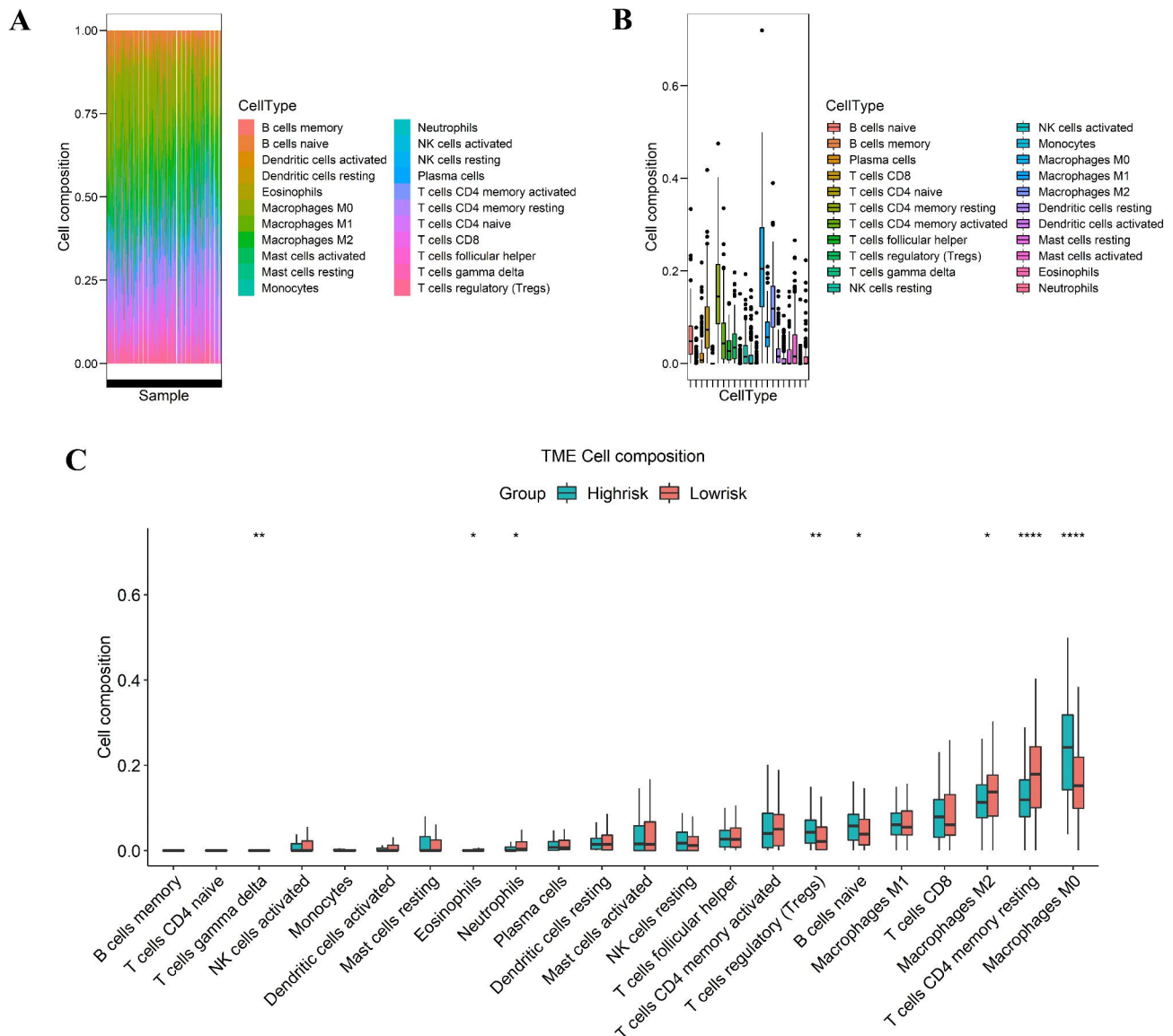


Fig. 7 Immune cell subtypes in CRC were analyzed using CIBERSORT. (A) Bar chart displaying the proportion of immune cell subsets. The x-axis shows sample names, and the y-axis shows the percentage of 22 immune cell subsets; (B) Box plot of the percentage of 22 immune cell subsets; (C) Proportion of 22 immune cell subsets in the high- and low-risk groups. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

In order to elucidate the mechanisms underlying the signature, we performed GSEA to analyze the KEGG pathways between the two risk groups. The most significantly enriched pathways in the high-risk group were focal adhesion, ECM receptor interaction, AXON guidance, basal cell carcinoma, and WNT signaling pathway. These pathways were considered to be associated with the poor prognosis of CRC [11–15]. These results suggested that the interaction of these pathways and the four hub genes (DKC1, NSUN5, FLNA and CSE1L) perhaps an important reason for the poor prognosis in the high-risk group.

Furthermore, the clinical features of CRC are vital factors that influence the prognosis of patients. In this study,

we established a robust prognostic signature, which could categorize CRC patients into high- and low-risk groups with statistically different OS outcomes. Next, we systematically analyzed the patient risk score and clinical information in TCGA (including age, gender, tumor stage, T stage, and N stage) for the OS by univariate and multivariate Cox regression analysis. The results indicated that our signature could be used as an independent prognostic factor to predict the prognosis of patients. In terms of clinical relevance, the prognostic signature was significantly correlated with gender.

Additionally, epidemiologic studies had revealed that the immune microenvironment affect the development and prognosis of colorectal cancer [39]. Zhang et

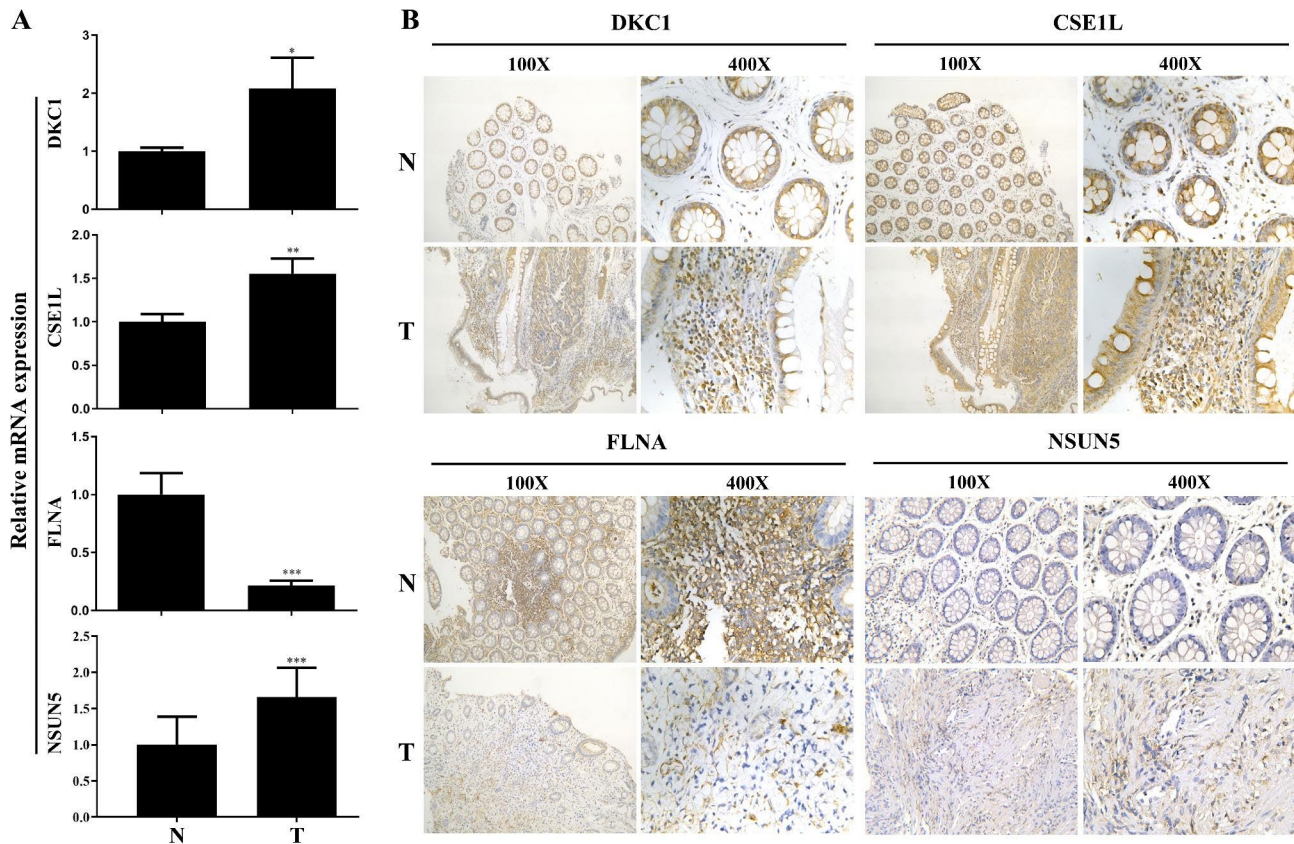


Fig. 8 The mRNA and protein expression of the four hub genes in CRC tissues and normal tissues. (A) The mRNA expression of DKC1, CSE1L, FLNA and NSUN5 were analyzed by RT-PCR, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$; (B) Protein levels of the four genes in CRC tissues and normal tissues were analyzed by IHC assay. N: normal tissues, T: CRC tissues

al. found that the patient with a higher degree of tumor immune invasion had a better clinical prognosis [40]. Therefore, exploring the correlation between prognosis and tumor immune function has guiding significance for the diagnosis and treatment of colorectal cancer. The current signature identified an association between the increased number of M0 macrophages in the high-risk patients and CD4 memory resting T cells, which was significantly higher in the low-risk group than in the high-risk group. We also observed differences in immune cell infiltration between the two risk groups. However, additional data are required to draw a conclusion. Previous studies focused on macrophages and Tregs in the immunological landscape of CRC. Tregs may contribute to cancer development, and macrophages may be associated with cancer metastasis and immune suppression. Zhang et al. found that CD4 memory resting T cells were associated with patients with advanced CRC, indicating that these cells predicted survival [41]. Jiang et al. indicated that patients with high numbers of M0 macrophages in the tumor environment have an increased risk of mortality [42]. A possible explanation is that M0 macrophages, together with other suppressor cells, such as Tregs, contribute to an immunosuppressive environment. Taken

together, the infiltration of immune cells in CRC indicates the immune status of patients, which might underlie the difference in survival outcomes between the two risk groups.

Nevertheless, the present study has some limitations. Firstly, our findings are based on public databases with a limited number of patients. Secondly, this is a retrospective study, and prospective studies are needed to verify our signature. Finally, we initially used a small number of clinical samples to verify the genes screened by the model. However, more clinical samples, basic experiments and the molecular mechanisms for this signature need to be further substantiated in future studies.

Conclusion

In conclusion, we constructed a diagnostic model based on the SVM. Our results revealed that this model has a relatively high diagnostic efficiency for CRC. Moreover, we identified a 4-gene prognostic signature as potential prognostic predictor for CRC patients. These findings would provide a theoretical reference for the future exploring the potential biomarkers for diagnosis and prognosis prediction of CRC patients.

Materials and methods

Collection of clinical samples

Five human CRC and adjacent normal tissues were obtained from patients diagnosed with CRC and received surgery at the People's Hospital of Longhua, Shenzhen, from October to November 2021. No patient had received preoperative chemotherapy or radiotherapy. The procedure of this study was approved by the Ethics Committee of the People's Hospital of Longhua, Shenzhen (approval number: Ethical review of the People's Hospital of Longhua (Institute) [2021] No. 120), and the utilization of clinical samples followed the guidelines of the Ethics Committee of the hospital. Informed consent was obtained from all participants.

Data download

We downloaded CRC RNA-seq raw count data from the TCGA database (<https://portal.gdc.cancer.gov>), which were preprocessed and normalized using the R package “Deseq2” and R package “pre-process Core.” The clinical information of CRC patients was downloaded from cBioPortal (<http://www.cbioportal.org/>).

Data preprocessing and DEGs analysis

Differential analysis between CRC tissues and non-cancerous adjacent tissues was performed using the R package “limma.” Then, the Benjamini–Hochberg method was utilized for differential analysis to obtain DEGs. Fold-change (FC) and adjusted p-values (adj. p-value) were used to screen the DEGs. $|\log(\text{FC})| \geq 1$ and $\text{adj. p-value} < 0.05$ were defined as the screening criteria for DEGs [43].

MEGENA of DEGs

MEGENA is an algorithm for mining module information from expression spectrum data [44]. According to the algorithm, modules are defined as a group of genes with similar expression profiles. If some genes always have similar expression changes in a physiological process or different tissues, they are defined as a module. MEGENA consists of three major steps: (1) The correlation between any two genes is calculated, and the genes are sorted according to the correlation; (2) Fast Planar Filtered Network construction (FPFNC) by introducing parallelization, early termination, and prior quality control; (3) PFNs are iteratively processed by three criteria: the shortest path distance, local path index, and overall modularity to obtain accurate gene classification.

PPI analysis of node genes

The node genes were obtained by MEGENA algorithm. The PPI analysis of the node genes was performed using String10.0. The results of the analysis were established using a gene expression network and mapped by

R package “igraph” and R package “ggraph”. The interaction correlation ($\text{degree} \geq 10$) was taken as the criterion to screen out the hub genes with high connectivity in the gene expression network.

Survival analysis of hub genes

The R package survival was applied to explore the roles of hub genes in disease-free survival (DFS) based on the tumor sample from the TCGA database. CRC patients were divided into two groups based on the median expression values of the hub genes. Then, the Kaplan–Meier (K–M) survival curve was established by combining the survival information of CRC patients in the high- and low-expression groups. $p < 0.01$ was considered statistically significant by the log-rank test.

Construction and validation of the diagnostic model by SVM

The significant risk genes were screened out through survival analysis. An SVM model was trained using ten-fold cross-validation [45]. The SVM model is a supervised classification algorithm of machine learning using python (version 3.8) package scikit-learn, which distinguishes and predicts the samples through Eigenvalues of the candidate genes in each sample and evaluates the probability of belonging to a specific category, thereby realizing the prediction between CRC tumor and normal samples. The true and false-positive rates were estimated, and the area under the ROC curve (AUC) was employed to estimate the performance of the model on the training and test sets from the TCGA database.

Establishment and validation of the prognostic model

Patients with clinical information (age, gender, stage, N stage, T stage, and OS time) were contained in the subsequent prognostic analysis. A total of 589 patients from the TCGA database were analyzed for the clinical correlation. These patients were divided into the training and test sets at a ratio of 1:1 using the R package “caret.” The training set was used to identify the survival-related DEGs and establish a prognostic signature, while the test and the total sets were used as internal validation sets.

The formula of the risk score for the prediction of CRC patients' prognosis was as follows: $\text{risk score} = \sum \beta_i * \text{expgene}_i$ (β_i : the coefficient of expgene_i . expgene_i : the expression level of gene_i). CRC patients were divided into high-risk and low-risk groups based on the median risk score of the training set. The survival analysis of the two groups was based on DFS and OS. The K–M survival curves were drawn using the R package “survival” and “survminer.” Next, we calculated the values of AUC to verify the feasibility and accuracy of our prognostic model and the clinical characteristics in predicting the patients' prognosis. Finally, both univariate and

multivariate Cox regression analyses were conducted using clinical parameters and risk scores to evaluate the independent prognostic value of the signature.

Gene set enrichment analysis

Patient samples were divided into high- and low-risk groups based on the risk score. Then, GSEA were performed to identify the KEGG pathways in the two groups using the Molecular Signatures Database (<http://www.gsea-msigdb.org/gsea/downloads.jsp>). At $p < 0.05$, normalized enrichment scores (NES) > 1.0 , and a false-discovery rate (FDR) $q < 0.25$, the pathway was considered significant.

Estimation of infiltrating immune cells

Tumor-infiltrating immune cells (TIICs) were estimated between the two risk groups using the CIBERSORT algorithm [46]. The algorithm utilized normalized gene expression data and the annotated gene signature matrix (LM22) to determine 22 immune cell subtypes. The LM22 file was downloaded from the CIBERSORT web portal (<https://cibersort.stanford.edu/>). Samples from CIBERSORT ($p < 0.05$) were considered significant and screened for further analysis.

Quantitative real-time polymerase chain reaction (qRT-PCR) analysis

Total RNA was extracted from clinical tissue samples (cancer and paired normal tissue) using TRIzol reagent (Invitrogen, Carlsbad, CA, USA). An equivalent of 1 μg of RNA was reverse transcribed into cDNA with oligo (dT) primers using the cDNA synthesis kit (Takara, Dalian, China). qRT-PCR was performed using SYBR Green qPCR Master Mix (Roche, Shanghai, China) according to manufacturer's instructions. β -actin was used as a house-keeping gene for normalization, and the sequences of the qRT-PCR primers are listed in Supplementary material Table S3. The qRT-PCR amplification reaction was as follows: initial denaturation at 95 $^{\circ}\text{C}$ for 10 min, followed by 40 cycles of 95 $^{\circ}\text{C}$ for 15 s and 60 $^{\circ}\text{C}$ for 45 s. The reactions were carried out in triplicates, and the relative quantification was performed using the comparative CT ($2^{-\Delta\Delta\text{CT}}$) method.

Immunohistochemistry (IHC) staining analysis

Tumor and paired normal tissue from the CRC patients were fixed in 10% formalin. Following standard protocols, 3- μm -thick sections were prepared and stained. The IHC staining was conducted using the UJltraSensitive™ SP (Mouse/Rabbit) IHC kit (MX Biotechnologies, Fuzhou, China), which contained endogenous peroxidase blocking solution, serum, secondary antibody, streptavidin-peroxidase, and DAB substrate-chromogen. The tissue sections were incubated with the rabbit

polyclonal antibodies of anti-DKC1 (1:100, Servicebio, Wuhan, China), anti-FLNA (1:100, Abcam, MA, USA), anti-NSUN5 (1:200, Abcam, MA, USA) and anti-CSE1L (1:100, Abcam, MA, USA) overnight at 4 $^{\circ}\text{C}$.

Data analysis and statistics

Statistical analyses of this study were conducted using the R software. The ROC curve analysis with the AUC was utilized to assess the predictive performance of the diagnostic and prognostic model. K–M curves with the log-rank test were plotted using the R package survival program. Additionally, univariate and multivariate Cox regression analyses were utilized to confirm the independent prognostic factors within clinicopathological characteristics. $P < 0.05$ indicated statistical significance for all analyses [47].

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13062-022-00342-w>.

Supplementary Material 1: Table S1: Characteristics of CRC patients included the prognostic model of DFS and OS in TCGA cohort

Acknowledgements

We thank all members of the Central Laboratory of People's Hospital of Longhua Shenzhen for excellent experimental environment. We thank all members of the Gastroenterology Department of People's Hospital of Longhua Shenzhen for their support.

Authors' contributions

DW performed experiments, analyzed data, and wrote the manuscript, JYLF and QYY established prognostic and diagnostic models, SQD and JQW collected clinical samples and patient materials, BX designed the study, commented on the manuscript, provided administrative support, and supervised the study. All authors contributed to the article and approved the submitted manuscript.

Funding

This study was supported by the National Natural Science Foundation of China (82100586), Shenzhen Science and Technology Program (RCBS20210609103823044), Scientific Research Projects of Medical and Health Institutions of Longhua District, Shenzhen (2021033), Medical Key Discipline of Longhua, Shenzhen (MKD202007090209), and Fund by Department of Innovation of Science and Technology of Longhua, Shenzhen (Public platform of Technique service of molecular Immunology and Molecular Diagnostics).

Availability of data and materials

The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding authors.

Declarations

Ethics approval and consent to participate

Human CRC and adjacent normal tissues were obtained from patients diagnosed with CRC and received surgery at the People's Hospital of Longhua, Shenzhen. The procedure of this study was approved by the Ethics Committee of the People's Hospital of Longhua, Shenzhen (approval number: Ethical review of the People's Hospital of Longhua (Institute) [2021] No. 120), and the utilization of clinical samples followed the guidelines of the Ethics Committee of the hospital. Informed consent was obtained from all participants.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Gastroenterology, People's Hospital of Longhua, NO.38 Jinglong Construction Road, Longhua District, 518109 Shenzhen, P.R. China

²Department of Gastroenterology, Guangzhou First People's Hospital, 511458 Guangzhou, P.R. China

Received: 19 July 2022 / Accepted: 14 October 2022

Published online: 02 November 2022

References

1. Siegel RL, Miller KD, Goding Sauer A, Fedewa SA, Butterly LF, Anderson JC, et al. Colorectal cancer statistics, 2020. *CA Cancer J Clin.* 2020;70(3):145–64. doi:<https://doi.org/10.3322/caac.21601>.
2. Siegel RL, Miller KD, Jemal A. Cancer Statistics. 2017. *CA Cancer J Clin.* 2017;67(1):7–30; doi: <https://doi.org/10.3322/caac.21387>.
3. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. *Cancer J Clin.* 2016;66(2):115–32. doi:<https://doi.org/10.3322/caac.21338>.
4. Siegel RL, Miller KD, Jemal A. Cancer statistics. 2015. *CA Cancer J Clin.* 2015;65(1):5–29; doi: <https://doi.org/10.3322/caac.21254>.
5. Siegel RL, Miller KD, Jemal A. Cancer statistics. 2019. *CA Cancer J Clin.* 2019;69(1):7–34; doi: <https://doi.org/10.3322/caac.21551>.
6. Qian Y, Wei J, Lu W, Sun F, Hwang M, Jiang K, et al. Prognostic Risk Model of Immune-Related Genes in Colorectal Cancer. *Front Genet.* 2021;12:619611. doi:<https://doi.org/10.3389/fgene.2021.619611>.
7. Ogunwobi OO, Mahmood F, Akingboye A. Biomarkers in Colorectal Cancer: Current Research and Future Prospects. *Int J Mol Sci.* 2020;21(15); doi:<https://doi.org/10.3390/ijms21155311>.
8. Huang Z, Yang Q, Huang Z. Identification of Critical Genes and Five Prognostic Biomarkers Associated with Colorectal Cancer. *Med Sci Monit.* 2018;24:4625–33. doi:<https://doi.org/10.12659/MSM.907224>.
9. Hou X, He X, Wang K, Hou N, Fu J, Jia G, et al. Genome-Wide Network-Based Analysis of Colorectal Cancer Identifies Novel Prognostic Factors and an Integrative Prognostic Index. *Cell Physiol Biochem.* 2018;49(5):1703–16. doi:<https://doi.org/10.1159/000493614>.
10. Wang L, Jiang X, Zhang X, Shu P. Prognostic implications of an autophagy-based signature in colorectal cancer. *Med (Baltim).* 2021;100(13):e25148. doi:<https://doi.org/10.1097/MD.00000000000025148>.
11. Machackova T, Vyčhítlová-Faltejsková P, Součková K, Trachtová K, Brchanelová D, Svoboda M, et al. MiR-215-5p Reduces Liver Metastasis in an Experimental Model of Colorectal Cancer through Regulation of ECM-Receptor Interactions and Focal Adhesion. *Cancers (Basel).* 2020;12(12); doi:<https://doi.org/10.3390/cancers12123518>.
12. Lascorz J, Bevier M, WVS, Kalthoff H, Aselmann H, Beckmann J, et al. Association study identifying polymorphisms in CD47 and other extracellular matrix pathway genes as putative prognostic markers for colorectal cancer. *Int J Colorectal Dis.* 2013;28(2):173–81. doi:<https://doi.org/10.1007/s00384-012-1541-4>.
13. Wang Z, Li P, Wu T, Zhu S, Deng L, Cui G. Axon guidance pathway genes are associated with schizophrenia risk. *Exp Ther Med.* 2018;16(6):4519–26. doi:<https://doi.org/10.3892/etm.2018.6781>.
14. Verkouteren BJA, Wakkee M, van Geel M, van Doorn R, Winnepenninckx VJ, Korpershoek E, et al. Molecular testing in metastatic basal cell carcinoma. *J Am Acad Dermatol.* 2021;85(5):1135–42. doi:<https://doi.org/10.1016/j.jaad.2019.12.026>.
15. Menck K, Wlochowicz D, Wachter A, Conradi LC, Wolff A, Scheel AH, et al. High-Throughput Profiling of Colorectal Cancer Liver Metastases Reveals Intra- and Inter-Patient Heterogeneity in the EGFR and WNT Pathways Associated with Clinical Outcome. *Cancers (Basel).* 2022;14(9); doi:<https://doi.org/10.3390/cancers14092084>.
16. Guo L, Wang C, Qiu X, Pu X, Chang P. Colorectal Cancer Immune Infiltrates: Significance in Patient Prognosis and Immunotherapeutic Efficacy. *Front Immunol.* 2020;11:1052. doi:<https://doi.org/10.3389/fimmu.2020.01052>.
17. Dieci MV, Miglietta F, Guarneri V. Immune Infiltrates in Breast Cancer: Recent Updates and Clinical Implications. *Cells.* 2021;10(2):223. doi:<https://doi.org/10.3390/cells10020223>.
18. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics. 2021. *CA Cancer J Clin.* 2021;71(1):7–33; doi: <https://doi.org/10.3322/caac.21654>.
19. McQuade RM, Stojanovska V, Bornstein JC, Nurgali K. Colorectal Cancer Chemotherapy: The Evolution of Treatment and New Approaches. *Curr Med Chem.* 2017;24(15):1537–57. doi:<https://doi.org/10.2174/0929867324666170111152436>.
20. Weng W, Feng J, Qin H, Ma Y. Molecular therapy of colorectal cancer: progress and future directions. *Int J Cancer.* 2015;136(3):493–502. doi:<https://doi.org/10.1002/ijc.28722>.
21. Wen F, Li Q. Treatment dilemmas of cetuximab combined with chemotherapy for metastatic colorectal cancer. *World J Gastroenterol.* 2016;22(23):5332–41. doi:<https://doi.org/10.3748/wjg.v22.i23.5332>.
22. Fessler E, Medema JP. Colorectal Cancer Subtypes: Developmental Origin and Microenvironmental Regulation. *Trends Cancer.* 2016;2(9):505–18. doi:<https://doi.org/10.1016/j.trecan.2016.07.008>.
23. Bijlsma MF, Sadanandam A, Tan P, Vermeulen L. Molecular subtypes in cancers of the gastrointestinal tract. *Nat Rev Gastroenterol Hepatol.* 2017;14(6):333–42. doi:<https://doi.org/10.1038/nrgastro.2017.33>.
24. Thanki K, Nicholls ME, Gajjar A, Senagore AJ, Qiu S, Szabo C, et al. Consensus Molecular Subtypes of Colorectal Cancer and their Clinical Implications. *other.* 2017;3(3).
25. Kirwan M, Dokal I. Dyskeratosis congenita: a genetic disorder of many faces. *Clin Genet.* 2008;73(2):103–12. doi:<https://doi.org/10.1111/j.1399-0004.2007.00923.x>.
26. Sieron P, Hader C, Hatina J, Engers R, Wlazlinski A, Muller M, et al. DKC1 overexpression associated with prostate cancer progression. *Br J Cancer.* 2009;101(8):1410–6. doi:<https://doi.org/10.1038/sj.bjc.6605299>.
27. Liu B, Zhang J, Huang C, Liu H. Dyskerin overexpression in human hepatocellular carcinoma is associated with advanced clinical stage and poor patient prognosis. *PLoS ONE.* 2012;7(8):e43147. doi:<https://doi.org/10.1371/journal.pone.0043147>.
28. O'Brien R, Tran SL, Maritz MF, Liu B, Kong CF, Purgato S, et al. MYC-Driven Neuroblastomas Are Addicted to a Telomerase-Independent Function of Dyskerin. *Cancer Res.* 2016;76(12):3604–17. doi:<https://doi.org/10.1158/0008-5472.CAN-15-0879>.
29. Bedolla RG, Wang Y, Asuncion A, Chamie K, Siddiqui S, Mudryj MM, et al. Nuclear versus cytoplasmic localization of filamin A in prostate cancer: immunohistochemical correlation with metastases. *Clin Cancer Res.* 2009;15(3):788–96. doi:<https://doi.org/10.1158/1078-0432.CCR-08-1402>.
30. Horimoto M. Expression of uncoupling protein-2 in human colon cancer. *Clin Cancer Res Official J Am Association Cancer Res.* 2004;10(18):6203–7.
31. Savoy RM, Ghosh PM. The dual role of filamin A in cancer: can't live with (too much of) it, can't live without it. *Endocr Relat Cancer.* 2013;20(6):R341–56. doi:<https://doi.org/10.1530/ERC-13-0364>.
32. Jin YZ, Pei CZ, Wen LY. FLNA is a predictor of chemoresistance and poor survival in cervical cancer. *Biomark Med.* 2016;10(7):711–9. doi:<https://doi.org/10.2217/bmm-2016-0056>.
33. Brinkmann U, Brinkmann E, Gallo M, Pastan I. Cloning and characterization of a cellular apoptosis susceptibility gene, the human homologue to the yeast chromosome segregation gene CSE1. *Proc Natl Acad Sci U S A.* 1995;92(22):10427–31. doi:<https://doi.org/10.1073/pnas.92.22.10427>.
34. Cook A, Fernandez E, Lindner D, Ebert J, Schlenstedt G, Conti E. The structure of the nuclear export receptor Cse1 in its cytosolic state reveals a closed conformation incompatible with cargo binding. *Mol Cell.* 2005;18(3):355–67. doi:<https://doi.org/10.1016/j.molcel.2005.03.021>.
35. Peiró G, Diebold J, Baretton GB, Kimmig R, Löhres U. Cellular apoptosis susceptibility gene expression in endometrial carcinoma: correlation with Bcl-2, Bax, and caspase-3 expression and outcome. *Int J Gynecol Pathol.* 2001;20(4):359–67. doi:<https://doi.org/10.1097/00004347-200110000-00008>.
36. Peiró G, Diebold J, Löhres U. CAS (cellular apoptosis susceptibility) gene expression in ovarian carcinoma: Correlation with 20q13.2 copy number and cyclin D1, p53, and Rb protein expression. *Am J Clin Pathol.* 2002;118(6):922–9. doi:<https://doi.org/10.1309/xycb-uw8u-5541-u4qd>.
37. Alnabulsi A, Agouni A, Mitra S, Garcia-Murillas I, Carpenter B, Bird S, et al. Cellular apoptosis susceptibility (chromosome segregation 1-like, CSE1L) gene is a key regulator of apoptosis, migration and invasion in colorectal cancer. *J Pathol.* 2012;228(4):471–81. doi:<https://doi.org/10.1002/path.4031>.
38. Jiang Z, Li S, Han MJ, Hu GM, Cheng P. High expression of NSUN5 promotes cell proliferation via cell cycle regulation in colorectal cancer. *Am J Transl Res.* 2020;12(7):3858–70.
39. Cao W, Chen HD, Yu YW, Li N, 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(7):783–91. doi:<https://doi.org/10.1097/CM9.0000000000001474>.
40. Zhao Z, McGill J, Gamero-Kubota P, He M. Microfluidic on-demand engineering of exosomes towards cancer immunotherapy. *Lab Chip*. 2019;19(10):1877–86. doi:<https://doi.org/10.1039/c8lc01279b>.
 41. Zhang X, Quan F, Xu J, Xiao Y, Li X, Li Y. Combination of multiple tumor-infiltrating immune cells predicts clinical outcome in colon cancer. *Clin Immunol*. 2020;215:108412. doi:<https://doi.org/10.1016/j.clim.2020.108412>.
 42. Jiang X, Wang M, Cyrus N, Yanez DA, Lacher RK, Rhebergen AM, et al. Human keratinocyte carcinomas have distinct differences in their tumor-associated macrophages. *Heliyon*. 2019;5(8):e02273. doi:<https://doi.org/10.1016/j.heliyon.2019.e02273>.
 43. Noble WS. How does multiple testing correction work? *Nat Biotechnol*. 2009;27(12):1135–7. doi:<https://doi.org/10.1038/nbt1209-1135>.
 44. Song WM, Zhang B. Multiscale Embedded Gene Co-expression Network Analysis. *PLoS Comput Biol*. 2015;11(11):e1004574. doi:<https://doi.org/10.1371/journal.pcbi.1004574>.
 45. Halloran JT, Rocke DM. A Matter of Time: Faster Percolator Analysis via Efficient SVM Learning for Large-Scale Proteomics. *J Proteome Res*. 2018;17(5):1978–82. doi:<https://doi.org/10.1021/acs.jproteome.7b00767>.
 46. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12(5):453–7. doi:<https://doi.org/10.1038/nmeth.3337>.
 47. Wang K, Liu J, Yan ZL, Li J, Shi LH, Cong WM, et al. Overexpression of aspartyl-(asparaginyloxy)-beta-hydroxylase in hepatocellular carcinoma is associated with worse surgical outcome. *Hepatology*. 2010;52(1):164–73. doi:<https://doi.org/10.1002/hep.23650>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.