

Characterization of RNA Processing Genes in Colon Cancer for Predicting Clinical Outcomes

Jianwen Hu^{1,2*}, Yingze Ning^{3*}, Yongchen Ma⁴,
Lie Sun¹ and Guowei Chen¹

¹Gastrointestinal Surgery Department, Peking University First Hospital, Beijing, China.

²Laboratory Department of Anzhen Hospital, Capital Medical University, Beijing, China.

³Department of Thoracic Surgery, Peking University Third Hospital, Beijing, China. ⁴Endoscopy Center, Peking University First Hospital, Beijing, PR China.

Biomarker Insights

Volume 19: 1–15

© The Author(s) 2024

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/11772719241258642



ABSTRACT

OBJECTIVE: Colon cancer is associated with multiple levels of molecular heterogeneity. RNA processing converts primary transcriptional RNA to mature RNA, which drives tumourigenesis and its maintenance. The characterisation of RNA processing genes in colon cancer urgently needs to be elucidated.

METHODS: In this study, we obtained 1033 relevant samples from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) databases to explore the heterogeneity of RNA processing phenotypes in colon cancer. Firstly, Unsupervised hierarchical cluster analysis detected 4 subtypes with specific clinical outcomes and biological features via analysis of 485 RNA processing genes. Next, we adopted the least absolute shrinkage and selection operator (LASSO) as well as Cox regression model with penalty to characterise RNA processing-related prognostic features.

RESULTS: An RNA processing-related prognostic risk model based on 10 genes including *FXR1*, *MFAP1*, *RBM17*, *SAGE1*, *SNRPA1*, *SRRM4*, *ADAD1*, *DDX52*, *ERI1*, and *EXOSC7* was identified finally. A composite prognostic nomogram was constructed by combining this feature with the remaining clinical variables including TNM, age, sex, and stage. Genetic variation, pathway activation, and immune heterogeneity with risk signatures were also analysed via bioinformatics methods. The outcomes indicated that the high-risk subgroup was associated with higher genomic instability, increased proliferative and cycle characteristics, decreased tumour killer CD8⁺ T cells and poorer clinical prognosis than the low-risk group.

CONCLUSION: This prognostic classifier based on RNA-edited genes facilitates stratification of colon cancer into specific subgroups according to TNM and clinical outcomes, genetic variation, pathway activation, and immune heterogeneity. It can be used for diagnosis, classification and targeted treatment strategies comparable to current standards in precision medicine. It provides a rationale for elucidation of the role of RNA editing genes and their clinical significance in colon cancer as prognostic markers.

KEYWORDS: Colon cancer, TCGA, GEO, RNA processing gene, risk score, nomogram, prognosis

RECEIVED: October 31, 2023. **ACCEPTED:** May 5, 2024.

TYPE: Research Article

CORRESPONDING AUTHORS: Lie Sun, Gastrointestinal Surgery Department, Peking University First Hospital, Beijing, China. Email: sunliemd@126.com

Guowei Chen, Gastrointestinal Surgery Department, Peking University First Hospital, Beijing, China. Email: guoweichen@263.net

Introduction

Colon cancer ranks third in morbidity and mortality among all cancer types mentioned in the 2020 Global Cancer Statistical Report, accounting for 6.0% and 5.8% of total cancer incidence and cancer-related deaths, respectively.¹ Local recurrence, chemotherapy resistance, lymph node metastasis, liver and lung metastasis are the key factors associated with poor prognosis in colon cancer.² Although, advances in systemic therapy have improved the overall prognosis of patients with colon cancer, significant differences in clinical outcomes exist among patients treated similarly. Current treatment decisions and prognoses are largely based on cancer cell-centric factors, such as TNM staging system. There is an urgent need to investigate a new classification protocol that provides comprehensive insights into the prognosis and treatment of colon cancer and improves the accuracy of traditional staging approaches such as TNM.

RNA processing genes are involved in mRNA transport, editing, and decay of messenger RNA.³ RNA process is an intermediate step linking genotype and phenotype, and facilitates conversion of the original RNA transcript into mature RNA.⁴ In both prokaryotes and eukaryotes, many RNAs require processing for functional maturation into RNA molecules. The processing of eukaryotic messenger RNAs is complex. Common processing events include: (1) The mRNA produced by polymerase II transcription undergoes a 5' end capping before leaving the nucleus; (2) Except for histone mRNA, most type II transcripts undergo 3' end processing. Coupled with a poly(A) tail measuring tens to hundreds of adenine nucleotides in length, the process entails cleavage of the 3' end of the pre-mRNA and polyadenylation; (3) The transcript undergoes editing including spacer deletion and splicing to form a functional mRNA. RNA sequencing analysis shows that more than 95% of human genes are regulated via alternative splicing, thus ensuring that a gene can produce multiple pre-mRNA or protein subtypes with different functions. RNA processing is

* Contributed equally



involved in several phenomena, including cellular apoptosis and maturation, tissue-specific expression, immune response, and tumor development and maintenance.⁵

Cancer-related shifts in RNA modification, RNA editing, and expression of non-coding RNA species such as long non-coding RNAs and micro-RNAs have been reported recently.^{6–9} RNA processing genes can influence the prognostic response of patients diagnosed with cancer.^{10–21} Differential analysis of RNA editing genes in tumors based on RNA processing factors facilitates tumor grading and treatment, and can serve as an important complementary marker of TNM staging system. However, changes in RNA splicing and their functional role in colon cancer development and maintenance, a feature of RNA processing, urgently needs to be fully elucidated. The rapid development and standardisation of high-throughput and low-cost next-generation sequencing protocols has facilitated clinical prognosis of tumors. In our study, we acquired eligible colon cancer samples to delineate the differences in tumour RNA-editing gene phenotypes.^{22,23} We used transcriptome data to identify various heterologous RNA-processing gene phenotypes and further investigate the underlying mechanism of each gene.

Finally, a 10-gene RNA processing-related prognostic classifier involving *FXR1*, *MFAP1*, *RBM17*, *SAGE1*, *SNRPA1*, *SRRM4*, *ADAD1*, *DDX52*, *ERI1*, and *EXOSC7* was identified in this study. As shown in the bioinformatics analysis above, the 10 genes not only suggest a significant impact on the prognosis of colon cancer, but several studies have also suggested that ten candidate RNA processing factors play important regulatory roles in the progression of colon cancer. For example, *FXR1* is a member of the RNA binding protein family and is highly amplified in many cancers.²⁴ *FXR1* can also stabilize target mRNAs such as *MYC*.²⁵ *MFAP1* is a member of microfiber related proteins involved in microfiber assembly, elastin generation, and tissue homeostasis.²⁶ By combining this 10-gene RNA processing-related prognostic classifier with other clinical variables, a composite prognostic nomogram was constructed to facilitate clinical practice based on TNM stage and age. The study first explains changes in RNA splicing in cancer and their role in the initiation as well as maintenance of cancer. Further, it distinguishes tumour RNA editing phenotypes, combined with transcriptome data to explore differences in heterologous RNA processing genotypes.

Materials and Methods

Colon cancer patient databases

Transcriptome sequencing data involving 446 colon cancer samples were acquired from The Cancer Genome Atlas colon cancer (TCGA-COAD) database (<https://portal.gdc.cancer.gov/>). A total of 585 microarray transcriptome data (GSE39582)²⁷ were obtained from the Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/geo/>). Clinical information and survival follow-up information

was also accessed from the corresponding websites. The expression profile (fragments per kilobase million [FPKM] value) of TCGA-COAD database was transformed into transcripts per kilobase million (TPMs) similar to the microarray data.²⁸ The “ComBat” algorithm in *sva* R package²⁹ was adopted to reduce batch impact due to non-biological technical bias across various databases.^{29,30} This study is a bioinformatics analysis and can be broadly classified as a retrospective study. This study mainly involved data analysis, which lasted about 3 months. Because all clinical samples and follow-up data involved in this study were obtained from the TCGA and GEO databases, we included all study samples with complete information.

Acquisition of RNA processing factors

RNA processing factors are genes involved in conversion of primary RNA transcripts into mature RNA molecules. The factors were obtained from the Gene Ontology term (GO:0006396) in the AmiGO database³¹ and 485 RNA processing factors were finally used for follow-up analysis. Gene expression data involving these RNA processing factors were obtained from the TCGA-COAD and GEO databases.

Unsupervised hierarchical cluster analysis for RNA processing factors

The study was carried out using an unsupervised clustering “Pam” method according to Euclidean and Ward chains and using “ConsensusClusterPlus” R package.³² In order to ensure the stability of the classification, a total of 1000 replicates were performed.³³ Different RNA processing modes were employed according to the expression of RNA processing factors for future investigation. Gene expression clustering was carried out with 80% of the resampling items, 50 resampling times and a maximum *K* value of 10.³³ The optimal *K* value was evaluated from uniform heat map and cumulative distribution function (CDF).³³

Identification of prognostic characteristics related to RNA processing factors

We adopted univariate Cox regression analysis to identify prognostic genes and a combined database including TCGA-COAD and GEO data was generated. Subsequently, the target gene with the greatest prognostic weight was identified using R package “glmnet” with the minimum absolute contraction and selection operator (LASSO).^{34–36} Then, a multivariate Cox analysis based on the Enter method was used to measure the effect of gene expression on prognosis. Finally, the prognostic risk features were established using multivariate Cox analysis combined with the identified gene expression and matching regression coefficients (β values).^{33,37} Using the median risk score as the critical value, gastric cancer was divided into low- and high-risk groups. The R “survival” package was used to assess differences in overall survival between the 2 groups.³⁸

Kaplan-Meier survival curves were generated, and differences were detected via log-rank test.

Development and verification of complex RNA processing-clinical prognostic nomogram

Based on the results of multivariate analysis, we integrated prognostic characteristics associated with age, sex, TNM stage, and RNA-treated genes to generate a composite prognostic model using cox proportional risk regression analysis of the combined TCGA-COAD and GEO data. The corresponding coefficients in the combined database were used to further verify TCGA-COAD and GEO databases. We compared the prognostic value of the composite prognostic model with the consistency index (C-index) of the TNM staging system, as indicated by RMS curve.³⁹ The C-index was used for discrimination and the U test was used to determine the unreliability of the calibration curve.⁴⁰ We also used RMS package to draw the nomogram and calibration curve in R software. Decision curve analysis (DCA) was used to quantify the net benefit at various threshold probabilities in validation sets to identify the clinical usefulness of the nomogram.^{41,42} The net benefit was assessed as follows:

$$\text{Net benefit} = \frac{\text{True positives}}{n} - \frac{\text{Pt}}{1 - \text{Pt}} \times \frac{\text{False positives}}{n}$$

where n is the total number of patients and Pt denotes the probability of a given threshold.

The differences in restricted mean survival (RMS) based on the risk score between the 2 risk groups were assessed.⁴³ RMS represents 60-month life expectancy of patients with various risk scores. Eventually, based on model visualization and clinical indicator analysis (calibration time curve graph), a modal graph was generated utilising the clinical application performance indicator (calibration curve) and decision curve analysis.^{39,41}

Bioinformatics analyses

We used principal component analysis (PCA) to identify the difference in expression between R package “extra factor” groups.⁴⁴ The gene set was annotated via GO and KEGG pathway enrichment analysis. GO and KEGG analyses were performed using clusterProfiler R Package.³² The differentially expressed RNA processing factors were identified with a statistically significant difference of $|\log_2\text{FC}| > 1$ and a false discovery rate (FDR) $< .05$ among different clusters in the combined database (combined TCGA-COAD and GEO data).³³ We counted the nonsynonymous mutations in colon cancer to identify the mutational burden. Somatic changes in colon cancer driver genes were assessed as high or low risk scores, followed by identification of colon cancer driver genes using “maftool” R Package.⁴⁵ The 25 most frequent cancer driver genes were analysed. The mutant landscape was created using the maftools package, initially deleting 100 frequently mutated genes (FLAGS).^{45,46} The levels of infiltration of different

immune cells in tcGA-COAD and GEO data were quantified using the “CIBERSORT” R package,⁴⁷ which carries an LM22 signature and 1000 permutations. The ESTIMATE algorithm can be used to estimate the immune and matrix contents (immune and matrix fractions)⁴⁸ of TCGA-COAD and GEO samples using the MCP counter package.⁴⁹ Different levels of immune cell infiltration are shown in heat maps and histograms. Gene Set Enrichment analysis (GSEA) was used to analyze the functional enrichment of genes related to risk score using cluster Profiler packages.^{32,50} The aforementioned data visualization uses Ggplot2 package.⁵¹

Based on protocol recommendations, WGCNA is used to identify gene modules associated with prognostic signals associated with RNA processing.^{52,53} The scale-free topological fitting index (R^2) $> .85$ was used as the threshold to build the weighted gene co-expression network. A minimum cluster size of 30 and a threshold of 0.25 were selected as the threshold for identifying co-expressed gene modules.⁵² The two-weight intermediate correlation coefficient $|R| \geq .15$ and a P -value $< .05$ represent the thresholds to identify gene modules associated with prognosis.⁵²

Statistical analyses

Statistical tests were carried out using SPSS 24.0 (IBM, Chicago, IL, USA) and R statistical software (version 4.1.1; <http://www.r-project.org/>). Kruskal-Wallis test and Wilcoxon test were used for comparison of more than 2 groups.³⁷ Continuous data were tested with Mann-Whitney test, and classified data were tested using Fisher’s exact test.³⁷ Kaplan-Meier plots were used to generate survival curves of each subgroup in each database, and the log-rank test facilitated the evaluation of statistically significant differences. Univariate cox regression analysis was used to screen genes with prognostic value, and multivariate cox regression analysis was used to evaluate the weight of candidate gene expression based on prognosis.³⁷ Spearman analysis facilitated the analysis of the relationship between WGCNA module and clinical characteristics.⁵³ The relationship between 2 continuous variables was measured with Pearson correlation coefficient.³⁷ The RMS curve and RMS time differences were estimated using survRM2 software package.^{39,41,43} COX regression and correlation analyses were performed using SPSS 24.0. Other statistical analyses and visualization were mainly implemented using R⁵⁴ and ggplot2 R Package.^{51,55} Two-tailed $P < .05$ was considered statistically significant. The data processing flow chart is provided in Supplemental Figure 1.

Results

Acquisition of RNA processing genes and identification of 4 different RNA processing modes in colon cancer

A total of 1033 samples (TCGA-COAD and GSE39582) were included in the study and 982 RNA editing genes were obtained from the AmiGO database (GO:0006396). The

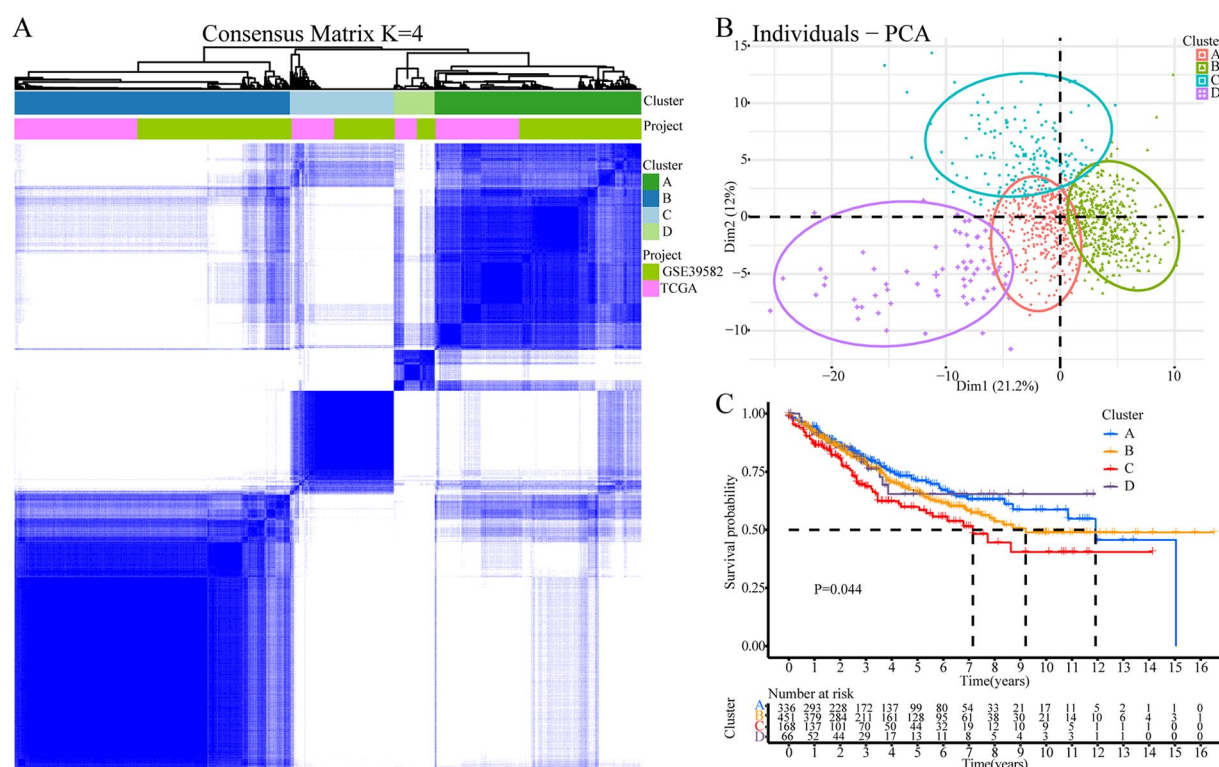


Figure 1. Identification of 4 different RNA processing modes in colon cancer. (A) Heatmap of 4 distinctive RNA processing modes defined by an unsupervised cluster analysis. (B) The principal component analysis (PCA) of the 4 subtypes were shown by the heatmap. (C) Kaplan-Meier survival curve analysis of overall survival for the 4 modes.

expression data of RNA editing genes detected in public database is only 485 genes, so a total of 485 genes were included in the sample with gene expression data (Supplemental Table 1). Thus, the data of 485 genes expressed were finally included in the prognostic model construction. Unsupervised cluster analysis was used to perform hierarchical clustering of data related to RNA editing genes in patients with colon cancer, with a cluster coefficient of $K=4$. Finally, 4 different modes of RNA editing were identified as shown in the heatmap (Figure 1A). The principal component analysis (PCA) plot showed tremendous variation in the expression profile of the 4 modes (Figure 1B). Significant differences in survival prognosis were detected among the 4 modes (Figure 1C). Clusters A and D were associated with prognostic advantage, while clusters B and C were related to prognostic disadvantage.

Identification of prognostic features related to RNA processing

We adopted univariate Cox regression analysis to identify the genes associated with prognostic significance in the combined database. Among 458 RNA processing genes, 51 were associated with overall survival including 20 genes ($HR > 1$, $P < .05$) associated with risk and 31 with a protective role ($HR < 1$, $P < .05$, Supplemental Table 2). Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) functional enrichment analysis suggested the genes associated with prognosis were related to key biological functions such as

spliceosome, RNA methyltransferase activity, U2-type precatytic spliceosome, catalytic activity involving RNA, U2-type spliceosome complex, RNA splicing, and ncRNA processing via transesterification involving a bulged adenosine used as the nucleophile (Figure 2A). A total of 10 genes containing non-zero parameters were determined (Figure 2B and C). The LASSO regression algorithm was used to screen 51 genes in the combined database to conveniently and effectively stratify the RNA processing genes with increased prognostic accuracy (Figure B and C). The genes finally incorporated into the survival prediction model included *FXR1*, *MFAP1*, *RBM17*, *SAGE1*, *SNRPA1*, *SRRM4*, *ADAD1*, *DDX52*, *ERI1*, and *EXOSC7* (Figure 2D). The effect of the selected genes on prognosis was measured via multivariate Cox analysis, based on the Enter method, visualized by dendrogram (Figure 2D). Prognostic risk profiles were established by combining the confirmed gene expression values with matching regression coefficients (β value) in multivariate Cox analysis. The corresponding risk scores for the databases were calculated based on the following formula: Risk score = $-2.224 \times ADAD1 - 0.309 \times DDX52 - 0.325 \times ERI1 - 0.501 \times EXOSC7 + 0.405 \times FXR1 + 0.508 \times MFAP1 + 0.495 \times RBM17 + 0.514$

$\times SAGE1 + 0.500 \times SNRPA1 + 0.997 \times SRRM4$. Using the median risk score as a cutoff, patients with colon cancer was divided into high-risk and low-risk groups (Supplemental Table 3). Kaplan-Meier survival curves showed that the survival time of patients in the high-risk group was less than that of patients in the high-risk group of the Union database

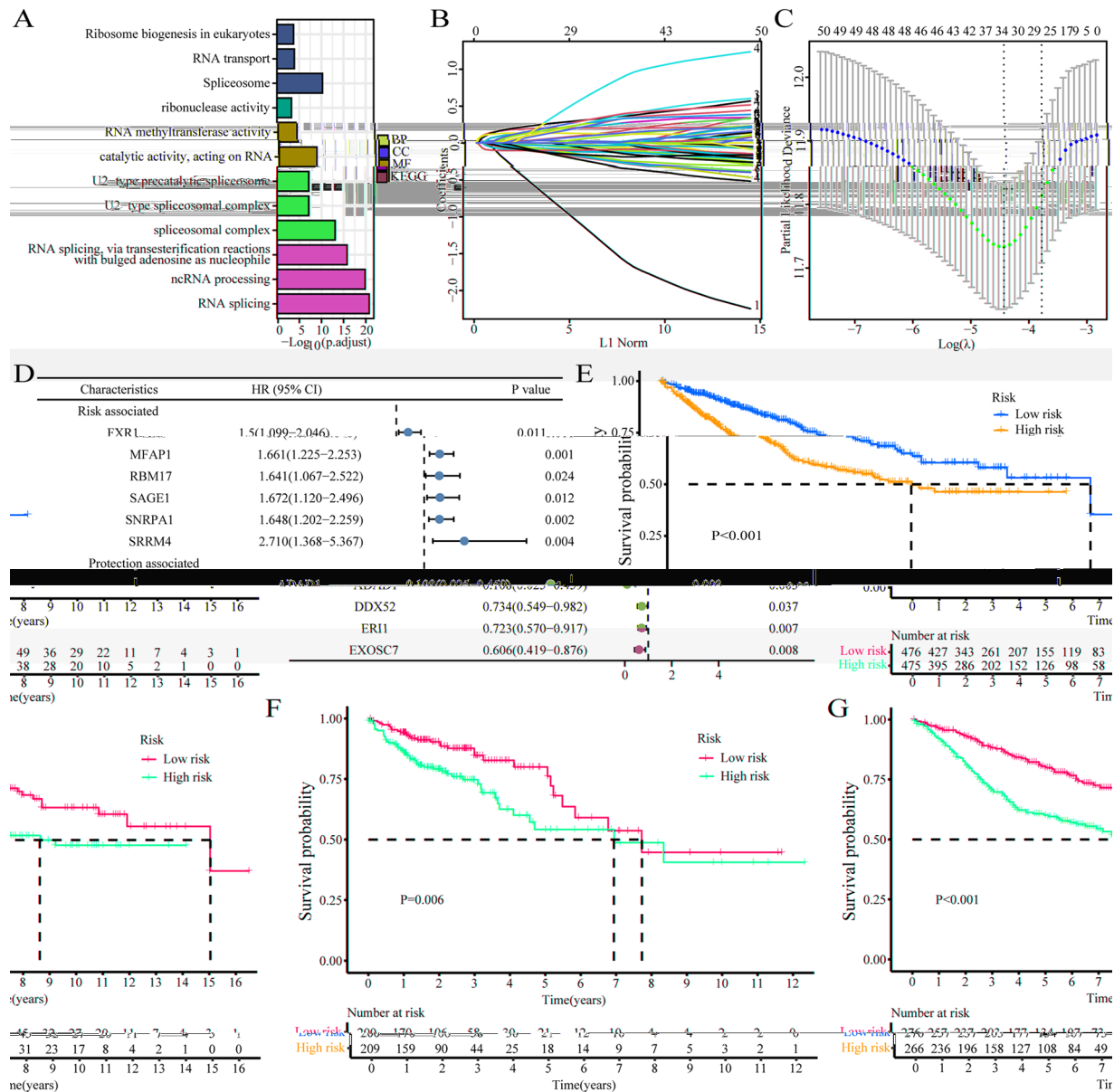


Figure 2. Identification of prognostic features related to RNA processing. (A) GO and KEGG functional enrichment analysis of differential genes among the 4 subgroups. (B) Lasso variable trajectory plot of prognosis-related RNA-processing genes. (C) Lasso coefficient filter plot of prognosis-related RNA-processing genes. (D) The corresponding hazard ratios of the contained 10 genes in the signature represented by the dendrogram. (E-G) Kaplan-Meier survival curve analysis by log-rank test of high-risk and low-risk groups in combined database (E), TCGA database (F) and GEO database (G).

(Figure 2E), TCGA (Figure 2F) and GEO databases (Figure 2G). Kaplan-Meier survival analysis showed that patients with low-risk scores showed significantly longer overall survival than those with high-risk scores. Multivariate Cox regression analysis showed the following hazard ratios for high-risk scores: combined database: $P < .001$, HR=2.646, 95% CI: 2.096-3.34; TCGA database: $P < .001$, HR=2.649, 95% CI: 1.835-3.825; GEO database: $P < .001$, HR=2.663, 95% CI: 1.957-3.623 (Table 1). In addition, subgroup analysis was performed based on age, gender, and TNM staging to explore the interaction between RNA processing-related risk signatures and clinical features in the combined database. The outcomes indicated that the group with the high-risk score was a distinct

element in the overall survival in different subgroups (Table 2). In addition, clinical features such as TNM stage and age were also distinct prognostic elements in colon cancer (Table 2).

Construction of overall survival nomogram for colon cancer patients

A combined nomogram containing the RNA processing gene-based risk score signatures and the significant clinical variables (TNM stage, sex, and age) was established to further improve the prognostic accuracy of the combined data (Figure 3A) to facilitate model visualization as well as clinical application. Meanwhile, the restricted mean survival time (RMST) of the 2

Table 1. Univariate and multivariate Cox analyses of the RNA processing-related signature in different databases.

DATABASE	FACTOR	UNIVARIATE		MULTIVARIATE	
		HR (95% CI)	P-VALUE	HR (95% CI)	P-VALUE
Combined database (n=951)	Age (>60 vs ≤60)	1.024 (1.013-1.034)	<.001	1.029 (1.019-1.04)	<.001
	Sex (Male vs Female)	1.232 (0.961-1.579)	.1	1.29 (1.006-1.654)	.045
	TNM stage (III and IV vs I and II)	2.199 (1.87-2.586)	<.001	2.295 (1.935-2.722)	<.001
	Risk score (High risk vs low risk)	2.923 (2.322-3.679)	<.001	2.646 (2.096-3.34)	<.001
TCGA database (n=409)	Age (>60 vs ≤60)	1.021 (1.002-1.04)	.034	1.035 (1.016-1.055)	<.001
	Sex (Male vs female)	1.1 (0.715-1.692)	.664	0.845 (0.542-1.317)	.457
	TNM stage (III and IV vs I and II)	2.276 (1.768-2.929)	<.001	2.386 (1.832-3.107)	<.001
	Risk score (High risk vs low risk)	2.774 (1.972-3.903)	<.001	2.649 (1.835-3.825)	<.001
GEO database (n=542)	Age (>60 vs ≤60)	1.025 (1.012-1.038)	<.001	1.029 (1.016-1.042)	<.001
	Sex (Male vs female)	1.3 (0.959-1.763)	.091	1.495 (1.098-2.034)	.011
	TNM stage (III and IV vs I and II)	2.124 (1.721-2.622)	<.001	2.254 (1.804-2.816)	<.001
	Risk score (High risk vs low risk)	2.923 (2.151-3.973)	<.001	2.663 (1.957-3.623)	<.001

groups in various databases was evaluated. Low risk group patients have better RMST at different time points in the combined database (Figure 3B), TCGA (Figure 3F) and GEO database (Figure 3J). Important differences in RMST between the high-risk and the low-risk groups were determined at various times; the difference in RMST increased with extended duration (Table 3). The differences in RMST between these groups were 0.427 (combined database), 0.552 (TCGA), and 0.231 (GEO) months for overall survival in the next year, reaching 7.816 (combined database), 7.601 (TCGA) and 7.780 (GEO) in the fifth year.

To validate the prognostic performance of the constructed nomogram, calibration curve, decision curve analysis (DCA), and concordance index (C-index) were used to test the prognostic performance of the combined database model, TCGA and GEO. Compared with the TNM stage model, the clinical models based on TNM stage, sex, and age and risk score, the composite model involving risk score, TNM stage, age and sex revealed a significant improvement in survival rates (Figure 3C, G, and K), which were assessed by C-index. The calibration curve showed the probabilities of observation and prediction of the nomograms in the combined database, TCGA and GEO (Figure 3D, H, and L). Eventually, we compared the net clinical benefits of the composite model with the models based on TNM stage, clinical and risk scores using the DCA curve. The nomogram based on the composite model showed higher net benefit (Figure 3E, I, and M). The foregoing results demonstrate the reliability, stability and enhanced prognostic performance of the nomogram-based composite model.

Functional enrichment analysis of RNA editing gene-related modules

RNA expression profiles were assessed based on prognostic features related to RNA processing considering that RNA processing genes control the life cycle of nuclear RNA. Using Pearson correlation analysis, we identified genes with expression levels related to trait risk scores, to establish a list ranked according to Pearson's correlation coefficients. The gene set enrichment analysis (GSEA) of genes with strong risk score correlation was performed. The gene expression map showed that the genes positively associated with risk score were largely expressed in pathways associated with cell proliferation cycle markers, such as reactome cell mitotic, reactome cell cycle checkpoints, reactome cell cycle, reactome mitotic prometaphase and reactome M phase, which were significantly enriched in colon cancer samples expressing higher risk (Figure 4A). By contrast, genes inversely associated with risk score were largely expressed in pathways associated with metabolic markers, including reactome defensins, reactome class C 3 metabotropic glutamate pheromone receptors, kegg allograft rejection, reactome antimicrobial peptides and kegg autoimmune thyroid disease in colon cancer samples (Figure 4A).

Next, we adopted weighted correlation network analysis (WGCNA) to obtain feature correlation modules based on approximately scale-free characters. The top 5000 most variable genes estimated via median absolute deviation (MAD) were used to perform WGCNA analysis. The optimal soft threshold power was selected to compute the adjacency matrix

Table 2. Subgroup analysis of the RNA processing-related signature.

DATABASE	FACTOR	SUBGROUP ANALYSIS			P-VALUE FOR INTERACTION
		SAMPLES	HR (95% CI)	P-VALUE	
Combined database (n=951)	Age				
	≤60	280	1.871 (1.096-3.192)	0.022	0.773
	>60	671	2.047 (1.534-2.733)	<0.001	
	Sex				
	Male	438	1.947 (1.309-2.894)	0.001	0.930
	Female	513	1.999 (1.436-2.781)	<0.001	
	TNM stage				
	I and II	526	1.740 (1.177-2.574)	0.006	0.557
	III and IV	425	1.944 (1.388-2.722)	<0.001	
TCGA database (n=409)	Age				
	≤60	125	1.804 (0.689-4.721)	0.229	0.801
	>60	284	1.975 (1.194-3.268)	0.008	
	Sex				
	Male	191	1.606 (0.837-3.083)	0.154	0.654
	Female	218	1.959 (1.061-3.616)	0.032	
	TNM stage				
	I and II	236	1.296 (0.624-2.690)	0.487	0.478
	III and IV	173	1.880 (1.050-3.368)	0.034	
GEO database (n=542)	Age				
	≤60	155	1.938 (1.016-3.697)	0.045	0.860
	>60	387	2.077 (1.459-2.958)	<0.001	
	Sex				
	Male	247	2.169 (1.314-3.581)	0.002	0.760
	Female	295	1.949 (1.313-2.892)	0.001	
	TNM stage				
	I and II	290	1.952 (1.224-3.114)	0.005	0.906
	III and IV	252	1.933 (1.278-2.924)	0.002	

and the largest 20 adjacency matrices, the largest complement of 20. The adjacency matrix was used to construct the cluster dendrogram; 8 color modules (green, blue, turquoise, brown, red, black, yellow, and gray) were recognized (Figure 4B). Uninvolved genes were transferred to the grey module for downstream investigation.

The correlation analysis of clinical traits was performed using the modular features constructed via WGCNA analysis to assess the modular-trait relationship. The 3 modules (brown,

black, and turquoise) were largely and positively related to the RNA processing gene-based risk score ($|R| > .15$, $P < .05$, Figure 4C), which suggests that the genes in these modules may perform basic biological roles associated with prognostic characteristics. Therefore, a gene functional enrichment investigation was performed in every module to determine the biological function of risk score-related modules. The most abundant terms in the brown module genes were homologous recombination, Fanconi anemia pathway, cell cycle, microtubule

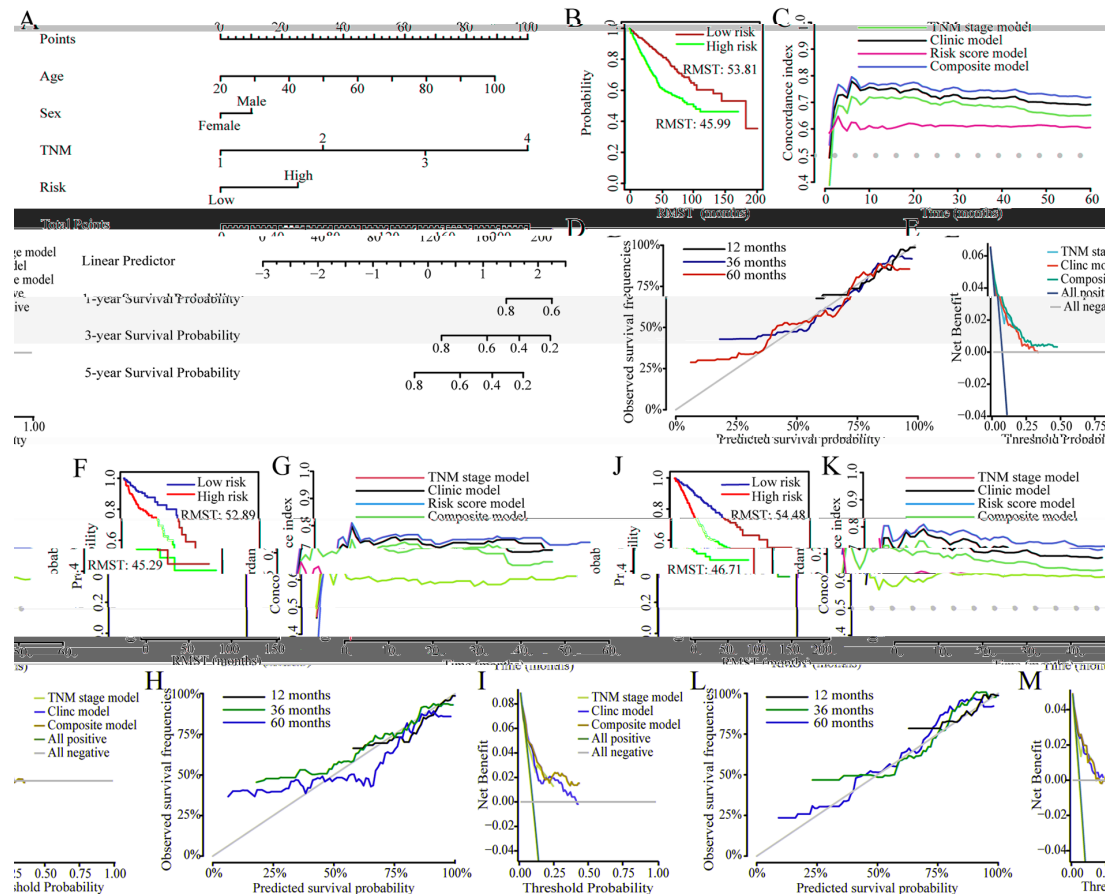


Figure 3. The construction of nomogram for overall survival in patients with colon cancer. (A) Prognostic nomogram for 1-year, 3-year, and 5-year overall survival of patients with colon cancer. (B, F, J) The restricted mean survival time (RMST) between the high-risk group and the low-risk group in the combined database, TCGA and GEO. (C, G, K) Concordance index (c-index) plots of TNM stage model, Clinic model, Risk score model and Composite model at different time points in the in the combined, TCGA and GEO databases respectively. (D, H, L) The calibration curve of the observation and prediction probabilities of the nomograms in the combined, TCGA and GEO databases respectively. (E, I, M) Decision curve analysis plots of TNM stage model, Clinic model and Composite model in combined, TCGA and GEO databases respectively. TNM stage model: involved in TNM stage only; Clinic model: involved in age, sex and TNM stage; Risk score model: involved in risk score only; Composite model: involved in risk score, TNM stage, age, and sex.

Table 3. RMST between the 2 risk groups at different time points.

DATABASE	TIME POINT (MONTHS)	RMST ^a		RMST DIFFERENCE ^b	P-VALUE
		LOW RISK (95% CI)	HIGH RISK (95% CI)		
Combined database (n=951)	12	11.751 (11.628, 11.874)	11.378 (11.182, 11.574)	-0.373(-0.604, -0.142)	0.002
	36	33.770 (33.121, 34.418)	30.644 (29.695, 31.593)	-3.126(-4.275, -1.976)	0.000
	60	53.810 (52.399, 55.221)	45.994 (44.055, 47.932)	-7.816(-10.214, -5.419)	0.000
TCGA database (n=409)	12	11.659 (11.435, 11.882)	11.107 (10.747, 11.466)	-0.552(-0.976, -0.128)	0.011
	36	33.220 (32.048, 34.393)	30.071 (28.451, 31.692)	-3.149(-5.149, -1.149)	0.002
	60	52.887 (50.246, 55.528)	45.286 (41.881, 48.691)	-7.601(-11.910, -3.292)	0.001
GSE39582 database (n=542)	12	11.817 (11.680, 11.953)	11.585 (11.379, 11.791)	-0.231(-0.478, 0.016)	0.060
	36	34.164 (33.422, 34.906)	31.224 (30.099, 32.348)	-2.940(-4.288, -1.593)	0.000
	60	54.485 (52.849, 56.121)	46.705 (44.365, 49.045)	-7.780(-10.635, -4.925)	0.000

The bold value means the outcome was statistically significant.

^aRestricted mean survival time (RMST), months.

^bRMST difference=RMST_{high risk}-RMST_{low risk}.

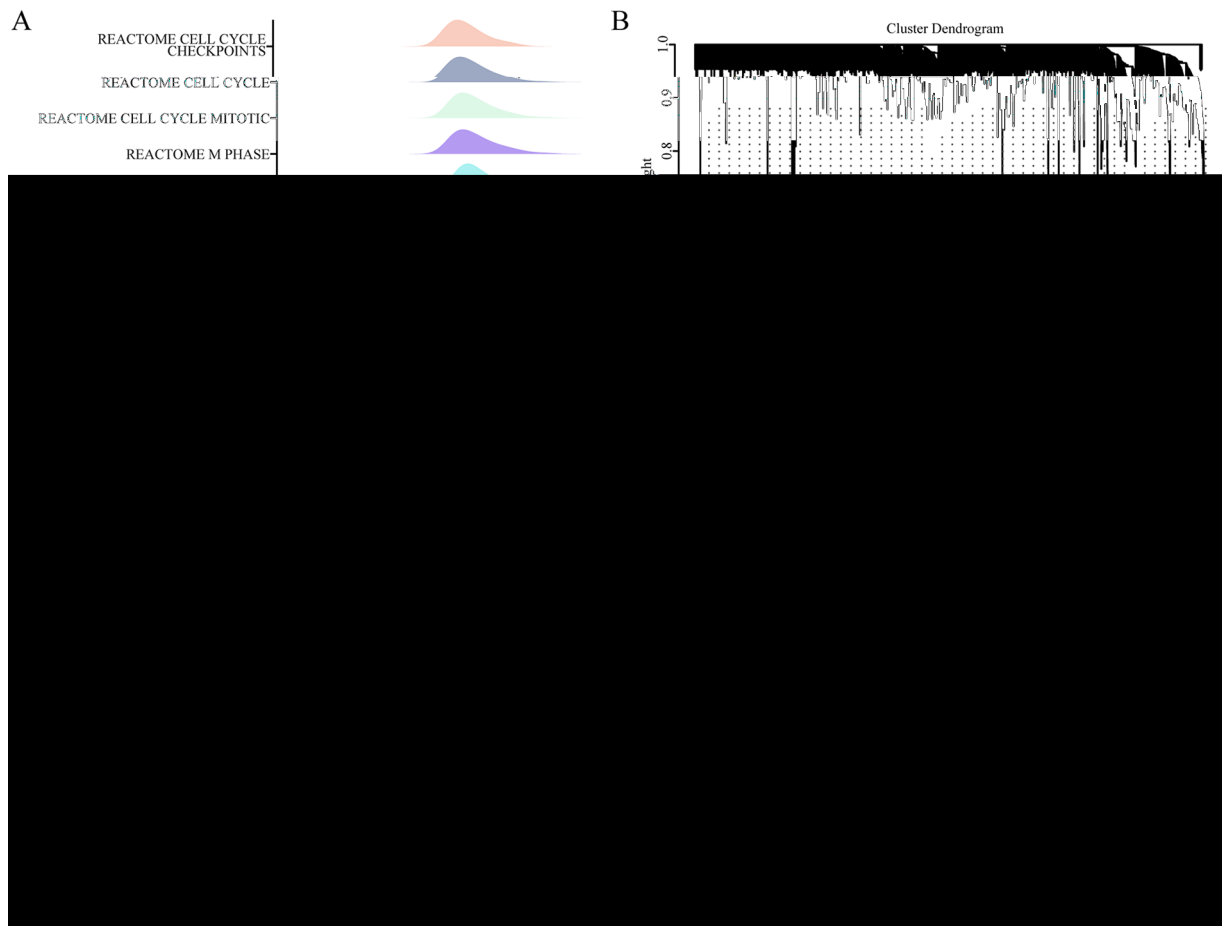


Figure 4. Functional enrichment analysis of RNA processing-based risk score-related genes. (A) Mountain map showed GSEA analysis results of genes associated with risk score. (B) The construction of a clustering dendrogram of the top 5000 most variable genes by an adjacency matrix. (C) Module-clinical traits relationship. Each column showed a module characteristic gene; each column corresponds to a clinical trait. Each cell contained the corresponding correlation (upper number) and *P*-value (lower number). (D) GO and KEGG functional enrichment analysis of genes in the brown module. (E) GO and KEGG functional enrichment analysis of genes in the turquoise module.

motor activity, DNA-dependent ATPase activity, ATPase activity, chromosomal region, spindle, condensed chromosome, organelle fission, nuclear division, chromosome segregation, indicating the role of the brown module in the cell cycle (Figure 4D). The genes in the turquoise module were associated with protein digestion and absorption, ECM-receptor interaction, collagen-containing extracellular matrix, focal adhesion, collagen binding, extracellular matrix component, extracellular matrix structural constituent, growth factor binding, endoplasmic reticulum lumen, extracellular matrix organization, extracellular structure organization, and cell-substrate adhesion (Figure 4E). In the absence of sufficient gene enrichment in the black modules, no functional enrichment analysis was performed. These findings suggest that RNA-processing genes based on risk scores were involved in multiple and important mechanisms associated with tumour malignancy phenotypes (proliferation, cell cycle, and metabolism).

Expression and clinical features of RNA processing-related prognostic features

All 1033 colon cancer samples were combined to identify the clinical characteristics and expression of prognostic signals

associated with RNA processing. The expression of all 10 prognostic gene markers identified via Cox regression analysis and LASSO varied significantly between the 2 risk groups (Figure 5A and B). Risk-related genes indicated higher levels of expression in patients with high risk score. In contrast, the expression of protective genes was higher in patients with low risk scores (Figure 5A and B).

In addition, the high-risk group had a higher proportion of patients with advanced tumor stages (stage III and IV) (Figure 5C). The proportions of clusters A and D with poor prognosis as well as stromal activation were higher in the high-risk group (Figure 5C).

Genetic variation and immune heterogeneity in prognostic characteristics related to RNA processing gene

Tumour mutational burden (TMB) in the TCGA database was analysed to identify the extent of genomic changes in the low- and high-risk subgroups. Figure 6 shows the TMB in different risk subgroups. In Figure 6A, the TMB of the high-risk group was significantly higher than in low-risk group ($P = .018$). First, the genomic data, including tumour mutations in the

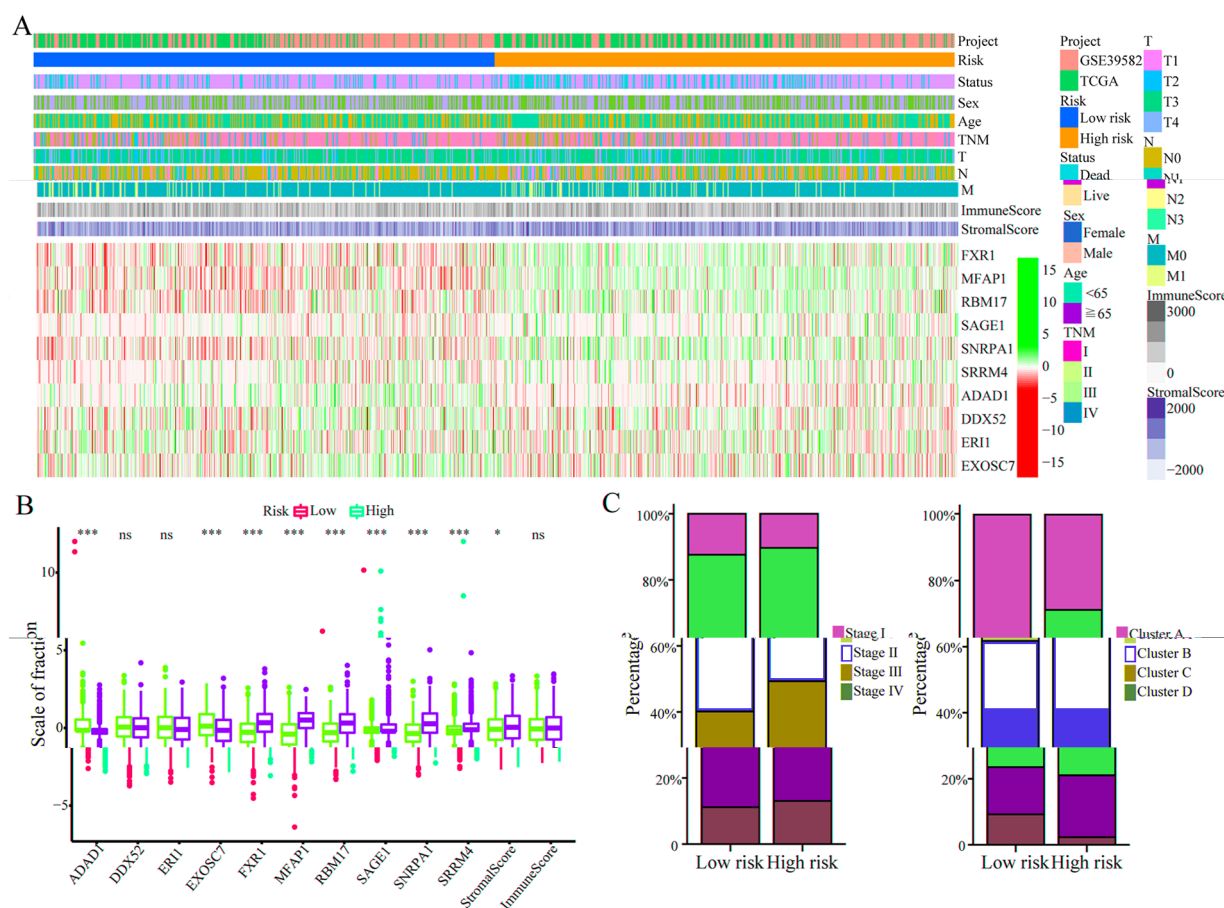


Figure 5. Expression and clinical features of RNA processing-related prognostic features. (A) The expression patterns of 10 prognostic-related RNA processing genes in the entire 1033 colon cancer samples shown by the heatmap. (B) The differential expression of 10 prognostic-related RNA processing genes between the low-risk group and the high-risk group. The P value was obtained by Mann-Whitney test. (C) The distribution of TNM stages and 4 different RNA processing modes between the low-risk group and the high-risk group shown by the histogram.

Abbreviation: ns, no statistical significance

* $P < .05$, ** $P < .01$, *** $P < 0.001$.

TCGA-COAD dataset were analysed to identify potential mechanisms of RNA processing-related prognostic features. Additional mutations are related to higher risk scores (Figure 6B). Kaplan-Meier survival analysis suggested that patients carrying low TMB (L-TMB) showed better survival than those with high TMB (H-TMB, $P = .043$, Figure 6C).

Next, the synergistic effects of TMB and risk scores were evaluated in TCGA-COAD prognostic stratification. Results indicated that TMB status did not influence prognosis based on risk score. Low- and high-risk subgroups showed enormous diversity in survival in both low- and high-TMB subgroups ($P = .006$; Figure 6D). Overall, the outcomes suggested that risk score stratification is a potential predictor distinct from TMB.

Furthermore, Maftools was used to access colon cancer driver genes. High-risk groups showed higher TMB in colon cancer. When genes were filtered out with low-frequency mutations (5% of colon cancer samples), the top 25 driver genes with the highest frequency of alteration were further analysed (Figure 6E and F). The mutation annotation file (MAF) analysis of the TCGA cohort showed that the altered frequency of *PIK3CA*, *APC*, *TP53*, *TIN*, *KRAS*, *SYNE1*, *RYR2*,

OBSCN, *MUC16*, *FAT4*, *ZFXH4*, *PCLO*, *FBXW7*, *LRP2*, *CSMD3*, *DNAH5*, *LRP1B*, *DNAH11*, *ABCA13*, and *USH2A* was not similar in the subgroups with low and high risk scores (Figure 6E and F). These results may provide new insights into the prognostic mechanisms of tumour risk score as well as gene mutations.

The relationship between tumour microenvironment and risk based on genes associated with RNA processing was explored to characterise immune heterogeneity, given the abundance of stromal as well as immune activation pathways in high-risk populations. The heatmap in Figure 7A shows immune cell infiltration in low- and high-risk groups. The MCP inverse algorithm was adopted to identify the proportion of immune and stromal cells in different risk groups. No substantial difference was found in stromal, estimated, and immune scores between the two groups (Figure 7B). Additionally, differential analysis of cellular components revealed higher percentages of CD8⁺ T cells, plasma cells, and eosinophils in samples with a low risk score than in those with a high risk score (Figure 7C). The high-risk group had a higher percentage of M0 and M1 macrophages.

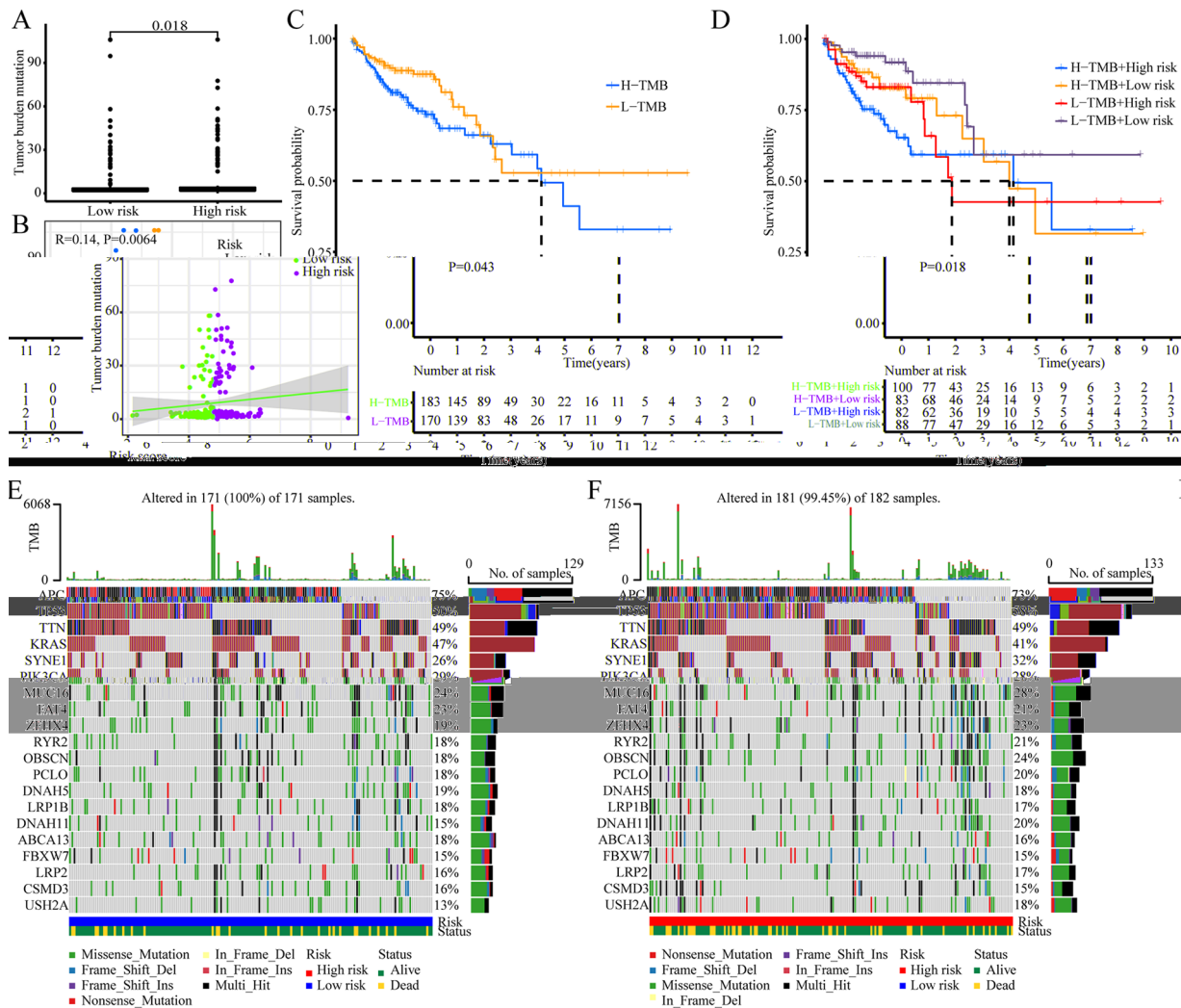


Figure 6. Genetic variation in prognostic characteristics related to RNA processing gene. (A) Violin chart of the tumor mutation burden (TMB) between low-risk and high-risk groups. (B) Scatter plot of correlation analysis between tumor mutational burden risk scores. (C) Kaplan-Meier survival curve analysis combining high and low tumor burden mutations and high and low risk scores. (D) Kaplan-Meier curves analysis for low and high TMB groups of the TCGA database. (E) TMB analysis in the low-risk group. (F) TMB analysis in the high-risk group.

Discussion

Although advances in systemic therapy have helped the overall prognosis of patients diagnosed with colon cancer, significant differences in clinical outcomes exist among patients treated similarly. Current treatment decisions and prognosis are mainly based on cancer cell-centric factors, such as the TNM staging system. RNA processing is involved in tissue-specific expression, apoptosis and maturation, immune responses, and tumor development and maintenance. Rapid advances in genomics and transcriptomics have facilitated systematic exploration of the heterogeneity of colon cancer. In this study, we identified transcriptional heterogeneity of RNA-editing genes in colon cancer based on an open database. Cluster analysis of colon cancer samples successfully identified and validated 4 RNA-edited gene-based molecules with molecular heterogeneity. Subtypes in each cluster were compared with clinical data to identify the unique molecular features of each cluster. Finally, a 10-gene RNA processing-related prognostic classifier involving *FXR1*,

MFAP1, *RBM17*, *SAGE1*, *SNRPA1*, *SRRM4*, *ADAD1*, *DDX52*, *ERI1*, and *EXOSC7* was identified. By combining this feature with other clinical variables, a composite prognostic nomogram was constructed to facilitate clinical practice based on TNM stage and age. The study first explains changes in RNA splicing in cancer and their role in the initiation as well as maintenance of cancer. Further, it distinguishes tumour RNA editing phenotypes, combined with transcriptome data to explore differences in heterologous RNA processing genotypes. Concurrently, by obtaining gene-like labels, the risk assessment classifier can be used to stratify the survival prognosis of patients with colon cancer in specific TNM subgroups, predict new sample categories and identify cancer subtypes. It can be used to provide classification and targeted treatment plans for patients, disease diagnosis and treatment according to specific transcriptome data, reduce cancer mortality, improve cancer life expectancy, and meet the current criteria for precision medicine. This study identified the differences in the phenotypes of RNA-edited

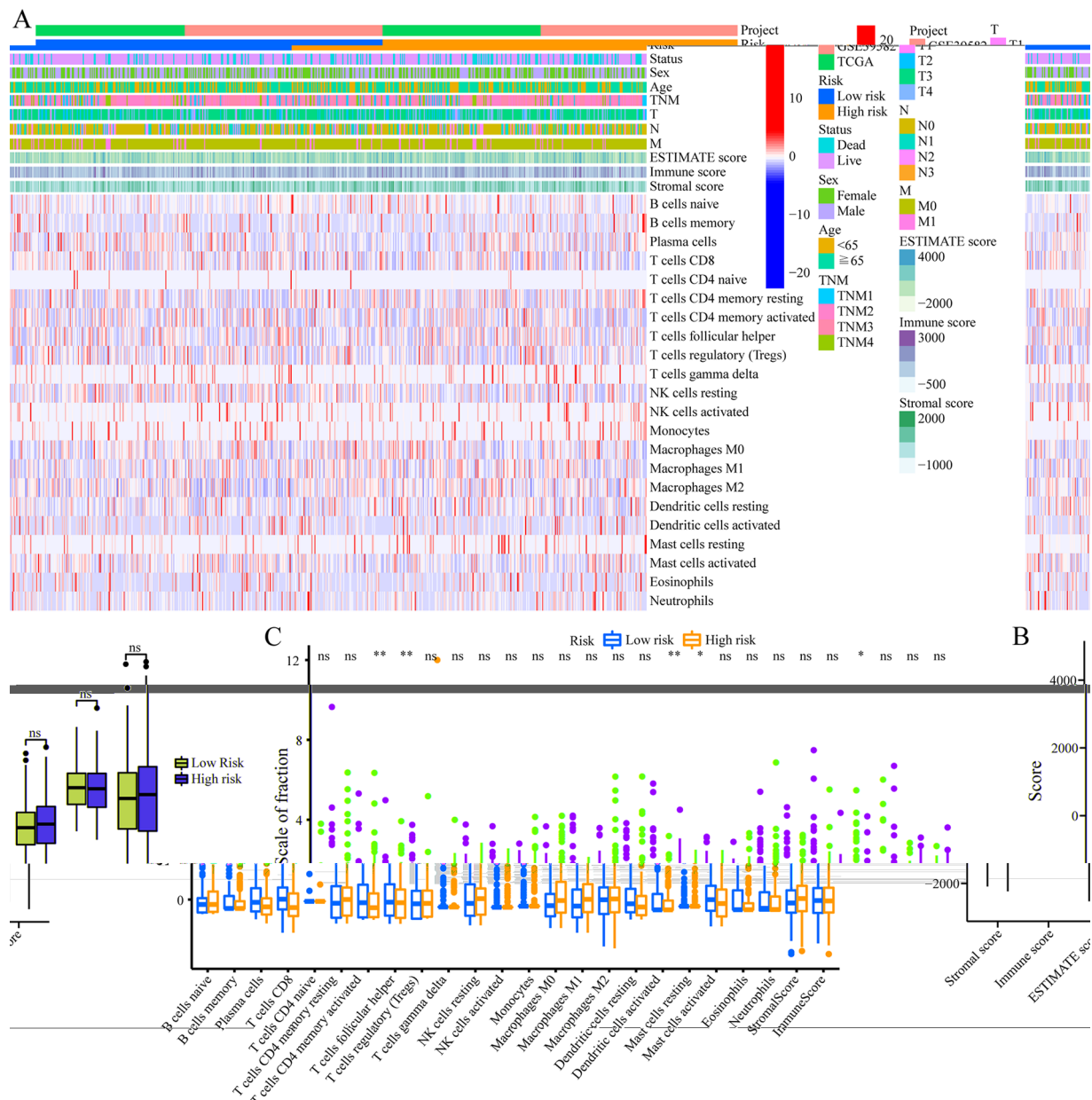


Figure 7. Immune heterogeneity in prognostic characteristics related to RNA processing gene. (A) Heat map showing immune cell infiltration between the high-risk and low-risk groups. (B) The differential expression of stromal score, immune score and ESTIMATE score between low-risk and high-risk groups. (C) Heat map showing immune cell infiltration between the high-risk and low-risk groups. The *P* value was obtained by the Mann Whitney test. Abbreviation: ns, non-statistical. **P* < .05; ***P* < .01; ****P* < .001.

genes in colon cancer, combined with transcriptome data to identify various phenotypes of heterologous RNA-edited genes, and provided further theoretical standards associated with potential pathogenic mechanisms of each RNA-edited gene. The new classification method combined with RNA-edited gene expression data provide detailed prognostic insights into colon cancer to predict treatment response and improve the role of traditional staging methods such as TNM.

This study provides a standard of reference for the discovery of highly sensitive and specific biomarkers for colon cancer treatment, which can predict tumour prognosis and facilitate the development of more effective anti-tumour drugs.

Currently, heterozygous mutations of RNA splicing elements *U2AF1*, *SRSF2*, and *SF3B1* have been reported in chronic lymphocytic leukemia, myeloid leukemia, uveal and mucosal melanoma.¹⁵⁻²¹ The study provides candidates for future research and improved treatment for colon cancer, adding to the growing understanding of RNA processing gene. Several valuable studies investigated the ten candidate RNA processing factors. Fragile X-related protein 1 (FXR1), which is highly amplified in many types of cancer,²⁴ is a member of the fragile X-related (FXR) family of RNA-binding proteins (RBPs). It is widely recognized that FXR1 binds to AU-rich elements (AREs) within the 3' untranslated region (3'UTR) and

enhances the stability of tumour necrosis factor alpha (TNF- α) and COX2 mRNAs.⁵⁶ Several mechanisms of FXR1 have been shown to stabilise the target mRNAs, such as MYC.²⁵ Microfibrillar-associated protein 1 (MFAP1) is a member of microfibrillar-associated proteins (MFAPs), which are extracellular matrix glycoproteins involved in microfibril assembly, elastinogenesis and tissue homeostasis.²⁶ It is presumed to be an ortholog of the *Saccharomyces cerevisiae* tri-snRNP protein Spp381 involved in the regulation of mRNA splicing.⁵⁷ RNA binding motif protein 17 (RBM17) binds to spliceosome and participates in the alternative splicing of mRNAs.⁵⁸ It is widely accepted that the expression of RBM17 is correlated with malignant tumors.⁵⁹ RBM17 is directly bound by tRNA-Gly and improves the malignant activities of cancer cells via RBM17-mediated alternative splicing.⁶⁰ Sarcoma antigen 1 (SAGE1) is expressed in different histological tumors.⁶¹ As one of the cancer/testis antigens, male germ cell proteins are expressed ectopically in different malignant tumours. SAGE1 is an ideal target in cancer immunotherapy.⁶² Small nuclear ribonucleoprotein polypeptide A1 (SNRPA1) is a protein-coding gene. It is associated with mRNA splicing-major pathway.⁶³ SNRPA1 plays an oncogenic role by interacting with RMRP to inactivate p53 in colorectal cancer.⁶⁴ Serine/Arginine Repetitive Matrix 4 (SRRM4) is generally acknowledged as a neural-specific splicing factor, with a very low basal expression outside of the brain.⁶⁵ Head et al.⁶⁶ reported that although SRRM4 expression is low in normal non-neural tissues, in malignant tumours, it is further silenced, leading to inhibition of normal microexon inclusion. Therefore, the SRRM4 splicing program acts as a proliferation inhibitor mediated via differentiation. Adenosine deaminase domain containing 1 (ADAD1) associated with CD4⁺ T cells was significantly related to the prognosis of colon cancer.⁶⁷ Dead-box RNA helicase 52 (DDX52) is involved in various RNA-based processes that bind to ATP. Previous studies have shown that DDX52 expression is down-regulated in prostate cancer, lung cancer and malignant melanoma.⁶⁸ Its function is mainly mediated via c-MYC pathway.⁶⁹ Exoribonuclease 1 (ERI1) is an RNA exonuclease involved in binding histone mRNA, playing a role in the decay of histone mRNA after replication. It also acts as a regulatory RNA interference (RNAi).

Immune cells are an important part of tumour microenvironment, and their number and status are important in the genesis, metastasis, and invasion of tumor. Our findings revealed no substantial difference in immune cell types among different risk groups, suggesting that the diversity of immune cell infiltration may not be the primary factor underlying the difference in prognosis among high-risk groups.

This study has limitations. First, because of the lack of clinical sample detection data specific to our own research centre, we can only provide validated results of the RNA-edited gene-based prognostic prediction model in the future to determine the prognostic efficiency of the model. Second, the role of

potential RNA-edited gene prognostic markers in tumourigenesis and maintenance requires additional experimental validation and in-depth study of related mechanisms. Our research group is conducting follow-up clinical study and related experiments.

Conclusion

We constructed an RNA-edited gene-based prognostic feature classifier of colon cancer to improve the prognosis of patients with colon cancer in well-defined TNM subgroups, and was characterised by immune heterogeneity, pathway activation, clinical outcomes, and genetic variation. The study provides a rationale for the elucidating the role of RNA-editing genes and suggests the clinical role of potentially meaningful colon cancer RNA-processing factors as prognostic markers.

Declarations

Ethics Approval and Consent to Participate

Since all data came from a public database, this entry was not applicable.

Consent for Publication

Not applicable.

Author Contributions

Jianwen Hu: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Supervision; Validation; Visualization; Writing – original draft. Yingze Ning: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft.

Yongchen Ma: Conceptualization; Funding acquisition; Methodology; Project administration; Software; Supervision; Visualization.

Lie Sun: Funding acquisition; Project administration; Supervision; Validation; Writing – review and editing.

Guowei Chen: Investigation; Methodology; Project administration; Resources; Supervision; Writing – review and editing.

Acknowledgements

The authors acknowledge Hang Zheng for his bioinformatics technical support.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Peking University Medicine Seed Funds for Interdisciplinary Research and the Fundamental Research Funds for the Central

Universities (No. BMU2022MX018), and the National Science and Beijing Natural Science Foundation of China (No. 7214259) supported the completion of this project.


Competing Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Availability of Data and Materials

All data and materials are freely accessible and can be obtained by contacting the corresponding author.

ORCID iD

Jianwen Hu  <https://orcid.org/0000-0003-3739-6996>

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

- Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71:209-249.
- Zhang YY, Chen SW, Wang PY, Liu YC. Research progress of conversion therapy in colorectal cancer liver metastases. *Zhonghua Wei Chang Wai Ke Za Zhi*. 2021;24:85-93.
- Obeng EA, Stewart C, Abdel-Wahab O. Altered RNA processing in cancer pathogenesis and therapy. *Cancer Discov*. 2019;9:1493-1510.
- Manning KS, Cooper TA. The roles of RNA processing in translating genotype to phenotype. *Nat Rev Mol Cell Biol*. 2017;18:102-114.
- Tollervey D, Caceres JF. RNA processing marches on. *Cell*. 2000;103:703-709.
- Han L, Diao L, Yu S, et al. The genomic landscape and clinical relevance of A-to-I RNA editing in human cancers. *Cancer Cell*. 2015;28:515-528.
- Schmitt AM, Chang HY. Long noncoding RNAs in cancer pathways. *Cancer Cell*. 2016;29:452-463.
- Vu LP, Cheng Y, Kharas MG. The biology of m(6)A RNA methylation in normal and malignant hematopoiesis. *Cancer Discov*. 2019;9:25-33.
- Peng Y, Croce CM. The role of MicroRNAs in human cancer. *Signal Transduct Target Ther*. 2016;1:15004.
- Geles KG, Zhong W, O'Brien SK, et al. Upregulation of RNA processing factors in poorly differentiated lung cancer cells. *Transl Oncol*. 2016;9:89-98.
- Lou S, Meng F, Yin X, Zhang Y, Han B, Xue Y. Comprehensive characterization of RNA processing factors in gastric cancer identifies a prognostic signature for predicting clinical outcomes and therapeutic responses. *Front Immunol*. 2021;12:719628.
- El-Sheikh NM, Abulsoud AI, Wasfey EF, Hamdy NM. Insights on the potential oncogenic impact of long non-coding RNA nicotinamide nucleotide transhydrogenase antisense RNA 1 in different cancer types; integrating pathway(s) and clinical outcome(s) association. *Pathol Res Pract*. 2022;240:154183.
- Emam O, Wasfey EF, Hamdy NM. Notch-associated lncRNAs profiling circulating epigenetic modification in colorectal cancer. *Cancer Cell Int*. 2022;22:316.
- Abd El, Fattah YK, Abulsoud AI, AbdelHamid SG, Hamdy NM. Interactome battling of lncRNA CCDC144NL-AS1: its role in the emergence and ferocity of cancer and beyond. *Int J Biol Macromol*. 2022;222:1676-1687.
- Papaemmanuil E, Cazzola M, Boulton J, et al. Somatic SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med*. 2011;365:1384-1395.
- Graubert TA, Shen D, Ding L, et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. *Nat Genet*. 2011;44:53-57.
- Wang L, Lawrence MS, Wan Y, et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N Engl J Med*. 2011;365:2497-2506.
- Furney SJ, Pedersen M, Gentien D, et al. SF3B1 mutations are associated with alternative splicing in uveal melanoma. *Cancer Discov*. 2013;3:1122-1129.
- Harbour JW, Roberson ED, Anbunathan H, Onken MD, Worley LA, Bowcock AM. Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma. *Nat Genet*. 2013;45:133-135.
- Martin M, Maßhöfer L, Temming P, et al. Exome sequencing identifies recurrent somatic mutations in *EIF1AX* and *SF3B1* in uveal melanoma with disomy 3. *Nat Genet*. 2013;45:933-936.
- Hintzsche JD, Gordon NT, Amato CM, et al. Whole-exome sequencing identifies recurrent SF3B1 R625 mutation and comutation of NF1 and KIT in mucosal melanoma. *Melanoma Res*. 2017;27:189-199.
- Abd El, Fattah YK, Abulsoud AI, AbdelHamid SG, AbdelHalim S, Hamdy NM. CCDC144NL-AS1/hsa-miR-143-3p/HMGA2 interaction: in-silico and clinically implicated in CRC progression, correlated to tumor stage and size in case-controlled study; step toward ncRNA precision. *Int J Biol Macromol*. 2023;253:126739.
- El-Sheikh NM, Abulsoud AI, Fawzy A, Wasfey EF, Hamdy NM. lncRNA NNT-AS1/hsa-miR-485-5p/HSP90 axis in-silico and clinical prospect correlated-to histologic grades-based CRC stratification: a step toward ncRNA precision. *Pathol Res Pract*. 2023;247:154570.
- Truitt ML, Ruggero D. New frontiers in translational control of the cancer genome. *Nat Rev Cancer*. 2017;16:288-304.
- George J, Li Y, Kadamberi I, et al. RNA-binding protein FXR1 drives cMYC translation by recruiting eIF4F complex to the translation start site. *Cell Rep*. 2021;37:109934.
- Zhu S, Ye L, Bennett S, Xu H, He D, Xu J. Molecular structure and function of microfibrillar-associated proteins in skeletal and metabolic disorders and cancers. *J Cell Physiol*. 2021;236:41-48.
- Marisa L, de Reyniès A, Duval A, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med*. 2013;10:e1001453.
- Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci*. 2012;131:281-285.
- Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28:882-883.
- Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8:118-127.
- Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. AmiGO: online access to ontology and annotation data. *Bioinformatics*. 2009;25:288-289.
- Yu G, Wang LG, Han Y, He QY. ClusterProfiler: an R package for comparing biological themes among gene clusters. *Omic*. 2012;16:284-287.
- Hu J, Yang Y, Ma Y, Ning Y, Chen G, Liu Y. Proliferation cycle transcriptomic signatures are strongly associated with gastric cancer patient survival. *Front Cell Dev Biol*. 2021;9:770994.
- Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33:1-22.
- Yang C, Delcher C, Shenkman E, Ranka S. Machine learning approaches for predicting high cost high need patient expenditures in health care. *Biomed Eng Online*. 2018;17:131.
- Dutta A, Batabyal T, Basu M, Acton ST. An efficient convolutional neural network for coronary heart disease prediction. *Expert Syst Appl*. 2019;159:113408.
- Hu J, Ma Y, Ma J, et al. M2 macrophage-based prognostic nomogram for gastric cancer after surgical resection. *Front Oncol*. 2021;11:690037.
- Therneau T. A package for survival analysis. *R package* 2.37-2; 2012.
- Eng KH, Schiller E, Morrel K. On representing the prognostic value of continuous gene expression biomarkers with the restricted mean survival curve. *Oncotarget*. 2015;6: 36308-36318.
- Alba AC, Agoritsas T, Walsh M, Hanna S, Guyatt G. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *JAMA*. 2017;318:1377.
- Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26:565-574.
- Fitzgerald M, Saville BR, Lewis RJ. Decision curve analysis. *JAMA*. 2015;313:409-410.
- Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol*. 2014;32:2380-2385.
- Kassambara A, Mundt F. *Factoextra: extract and visualize the results of multivariate data analyses*; 2017.
- Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res*. 2018;28:1747-1756.
- Shyr C, Tarailo-Graovac M, Gottlieb M, Lee JJ, van Karnebeek C, Wasserman WW. FLAGS, frequently mutated genes in public exomes. *BMC Med Genomics*. 2014;7:64.
- Newman AM, Liu CL, Green MR, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12:453-457.
- Yoshihara K, Shahmoradgoli M, Martínez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun*. 2013;4:2612.

49. Becht E, Giraldo NA, Lacroix L, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* 2016;17:218.
50. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005;102:15545-15550.
51. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer; 2011.
52. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* 2008;9:559.
53. Zheng H, Liu H, Li H, et al. Characterization of stem cell landscape and identification of stemness-relevant prognostic gene signature to aid immunotherapy in colorectal cancer. *Stem Cell Res Ther.* 2022;13:244.
54. Lee C, Grasso C, Sharlow MF. *R: a language and environment for statistical computing*; 2002.
55. Vivian J, Rao AA, Nothhaft FA, et al. Toil enables reproducible, open source, big biomedical data analyses. *Nat Biotechnol.* 2017;35:314-316.
56. Xiao-Cui L, Meng-Fan S, Feng S, et al. Fragile X-related protein 1 (FXR1) regulates cyclooxygenase-2 (COX-2) expression at the maternal-fetal interface. *Reprod Fert Develop.* 2018;30:1566-1574.
57. Ulrich AK, Wahl MC. Human MFAP1 is a cryptic ortholog of the *Saccharomyces cerevisiae* Spp381 splicing factor. *BMC Evol Biol.* 2017;17:91.
58. Lallena MJ, Chalmers KJ, Llamazares S, Lamond AI, Valcárcel J. Splicing regulation at the second catalytic step by sex-lethal involves 3' splice site recognition by SPF45. *Cell.* 2002;109:285-296.
59. Liu Y, Conaway L, Rutherford Bethard J, et al. Phosphorylation of the alternative mRNA splicing factor 45 (SPF45) by Clk1 regulates its splice site utilization, cell migration and invasion. *Nucleic Acids Res.* 2013;41:4949-4962.
60. Han L, Lai H, Yang Y, et al. A 5'-tRNA halve, tiRNA-Gly promotes cell proliferation and migration via binding to RBM17 and inducing alternative splicing in papillary thyroid cancer. *J Exp Clin Cancer Res.* 2021;40:222.
61. Ishihara M, Kageyama S, Miyahara Y, et al. MAGE-A4, NY-ESO-1 and SAGE mRNA expression rates and co-expression relationships in solid tumours. *BMC Cancer.* 2020;20:606.
62. Maheswaran E, Pedersen CB, Ditzel HJ, Gjerstorff MF. Lack of ADAM2, CALR3 and SAGE1 cancer/testis antigen expression in lung and breast cancer. *PLoS One.* 2015;10:e0134967.
63. Ram M, Najafi A, Shakeri MT. Classification and biomarker genes selection for cancer gene expression data using random forest. *Iran J Pathol.* 2017;12:339-347.
64. Chen Y, Hao Q, Wang S, et al. Inactivation of the tumor suppressor p53 by long noncoding RNA RMRP. *Proc Natl Acad Sci U S A.* 2021;118:e2026813118.
65. Irimia M, Weatheritt RJ, Ellis JD, et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell.* 2014;159:1511-1523.
66. Head SA, Hernandez-Alias X, Yang JS, et al. Silencing of SRRM4 suppresses microexon inclusion and promotes tumor growth across cancers. *PLoS Biol.* 2021;19:e3001138.
67. Yang X, Wu W, Pan Y, Zhou Q, Xu J, Han S. Immune-related genes in tumor-specific CD4(+) and CD8(+) T cells in colon cancer. *BMC Cancer.* 2020;20:585.
68. Yu W, Ma H, Li J, et al. DDX52 knockdown inhibits the growth of prostate cancer cells by regulating c-Myc signaling. *Cancer Cell Int.* 2021;21:430.
69. Wang Q, Qian L, Tao M, Liu J, Qi FZ. Knockdown of DEAD-box RNA helicase 52 (DDX52) suppresses the proliferation of melanoma cells in vitro and of nude mouse xenografts by targeting c-Myc. *Bioengineered.* 2021;12:3539-3549.