Observational Study

Medicine

OPEN

Systematic analysis competing endogenous RNA coexpression network as a potentially prediction prognostic biomarker for colon adenocarcinoma

Jiaxi Xi, MM[a], Huajun Zhang, MM[a], Yan Li, MM[a], Henghai Su, MM[a], Xiaoyu Chen, MD[a], Xueyan Liang, MM[a,*]

Abstract

Colon adenocarcinoma (COAD) is one of the most common types of colon cancer, represents a major public health issue due to its high incidence and mortality. Competing endogenous RNAs (ceRNAs) hypothesis has generated a great interest in the study of molecular biological mechanisms of cancer progression. The aim of this study was to identify potential prediction prognostic biomarker associated with progression of COAD and illuminate regulatory mechanisms. Two RNA sequencing datasets downloaded from the Genotype-Tissue Expression and TCGA. The differentially expressed RNAs were analyzed. Weighted correlation network analysis was used to analyze the similarity of genes model with a trait in the network. Interactions between lncRNAs, miRNAs, and target mRNAs were predicted by MiRcode, starBase, miRTarBase, miRDB, and TargetScan, and the risk score of mRNAs was established. Based on the identified prognostic signature and independent clinical factors, then the nomogram survival model was built. Totally, we identified 3537 differentially expressed mRNAs, 2379 lncRNAs, and 449 microRNAs. Based on the 8 prognosis-associated mRNAs (CCNA2 + CEBPA + NEBL + SOX9 + DLG4 + RIMKLB + TCF7L1 + TUB), the risk score was proposed. After the independent clinical prognostic factors were identified, the nomogram survival model was built. LncRNA-miRNA-mRNA ceRNA network was built by 68 lncRNAs, 4 miRNAs, and 6 mRNAs, which might serve as prognostic biomarkers of COAD. These findings suggest several genes in ceRNA network might be novel important prognostic biomarkers and potential targets for COAD. CeRNA networks could provide further insight into the mRNA-related regulatory mechanism and COAD prognosis.

Abbreviations: BP = biological process,ceRNAs = competing endogenous RNAs, COAD = colon adenocarcinoma, GO = gene ontology, GSEA = Gene Set Enrichment Analysis, KEGG = Kyoto Encyclopedia of Genes and Genomes, lncRNAs = long non-coding RNAs, mRNAs = messenger RNAs, miRNAs = microRNAs, ROC = receiver operating characteristic, TCGA = The Cancer Genome Atlas, WGCNA = weighted correlation network analysis.

Keywords ceRNA coexpression network, colon adenocarcinoma, nomogram survival model, prognostic markers

1. Introduction

Colon cancer is a leading cause of mortality globally and the third most frequently diagnosed malignancy.[1] Approximate 98% of pathological type colon cancer is colon adenocarcinoma (COAD).[2] Previous studies have demonstrated that colon cancer can be successfully treated when diagnosed and identified at an early stage.[3] With the developments in diagnosis and treatment for COAD in recent years, however, it remains poor in the prognosis of COAD and the choice of therapeutic options.[4,5] Hence, identifying potential effective prognostic biomarkers and therapeutic targets for the treatment of COAD is urgently needed.

A competing endogenous RNA (ceRNA) hypothesis is the pool of long non-coding RNAs (lncRNAs) and messenger RNAs (mRNAs) compete and bind to microRNAs (miRNAs) by miRNA response elements, regulating their activity and play an important role in various cancer biological processes (BPs).[6–8] LncRNAs playing the regulatory role are considered to be the foundation of the hypothesis of ceRNA.[9] Hence, understanding the complex interaction among various ceRNA networks will lead to a profound understanding of gene regulatory networks and have indicated in cancer progress, diagnoses, and treatment.[10] However, a systematic analysis of COAD-associated ceRNA network and prediction prognostic biomarkers of mRNA is still lacking.

In this study, we conducted a systematic analysis to identify differentially expressed mRNAs (DEmRNAs), lncRNAs (DElncRNAs), and miRNA (DEmiRNAs) in COAD. Here we obtained RNA sequencing (RNA-Seq) data and the miRNA-seq

JX and HZ contributed equally to this work.

This study is funded by the National Natural Science Foundation of Guangxi Province (No. 2018GXNSFAA281159).

The authors have no conflicts of interest to disclose.

The datasets generated during and/or analyzed during the current study are publicly available.

[a] Department of Pharmacy, The People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, Guangxi, China.

**Correspondence: Xueyan Liang, Department of Pharmacy, The People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, Guangxi 530021, China (e-mail: liangxueyan20102010@outlook.com).*

How to cite this article: Xi J, Zhang H, Li Y, Su H, Chen X, Liang X. Systematic analysis competing endogenous RNA coexpression network as a potentially prediction prognostic biomarker for colon adenocarcinoma. Medicine 2022;101:39(e30681).

Received: 10 August 2020 / Received in final form: 21 August 2022 / Accepted: 22 August 2022

http://dx.doi.org/10.1097/MD.0000000000030681

data from The Cancer Genome Atlas (TCGA) and the Genotype-Tissue Expression (GTEx) database. Next, weighted correlation network analysis (WGCNA) were applied to enrich COAD relevance mRNAs and lncRNAs modules between the COAD patients and normal samples. And, the target mRNA was predicted by miRNA database. Univariate and multivariate Cox regression analysis was further established to identify prognosis-associated mRNA-based signature and establish a risk assessment system based on the regression coefficient. Furthermore, the nomogram survival model was constructed. Moreover, COAD-associated ceRNA network was successfully constructed and several molecules have been identified that might be novel important prediction prognostic biomarker and act as potential treatment targets for COAD.

2. Methods

The study was approved by the ethics institutional review board of the People's Hospital of Guangxi Zhuang Autonomous Region.

2.1. Data source

The RNA-Seq data, miRNA-seq data, and clinical data of COAD and a part of normal tissue samples were obtained from TCGA data repository (https://portal.gdc.cancer.gov/). Another part of normal tissue samples data was obtained from GTEx V8 release version (https://gtexportal.org/home/datasets). GTEx official annotation completely described the donor age, genders, multiple ethnicity groups, the biospecimen procurement methods, and sample fixation.

2.2. Identification of differentially expressed genes

The ensemble ID of TCGA and GTEx samples was annotated by using GENCODE Gene Set-11.2019 version. The miRNA ensemble ID of GTEx samples was converted by using Human Genome Organisation Gene Nomenclature Committee (http://www.genenames.org). We excluded lncRNAs and mRNAs ensemble ID which was not recorded in the GENCODE database. R package edgeR[11] was used to identify significant DEmRNAs, DElncRNAs, and DEmiRNAs between COAD and normal samples. Absolute log2 (fold change) ≥ 2 and false discovery rate < 0.05 were considered significant.[11–13] To visualization the distribution of obtained all DEmRNAs, DElncRNAs, and DEmiRNAs, volcano map was generated using the ggplot2[14] packages.

2.3. Functional enrichment analysis

ClusterProfiler was used to achieve enrichment analysis of gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Set Enrichment Analysis (GSEA).[15–17] The differentially expressed genes are often enriched with specific biological significance functions. To test the biological function of the identified different expression genes, information from different expression genes was applied to GO analysis. GO terms were used to describe gene functions, including BP, cellular component, and molecular function. The KEGG-GSEA pathways with the significance level set at P value < .05, which can further predict the biological function of DEmRNA.

2.4. Weighted correlation network analysis

WGCNA was used to develop in gene coexpression network. The R "WGCNA" package[18] was used to identify gene expression profiles with different modules, and evaluate the correlation of each module with cancer factors to find the most relevant mRNAs or lncRNAs with COAD patients. The adjacency matrix was calculated by the pairwise Pearson correlation analysis, and it was transformed into a topological overlap matrix to define the similarity of the coexpression gene. In this study, WGCNA was used to analyze mRNAs and lncRNAs to find the most relevant mRNAs or lncRNAs with COAD patients.

2.5. MiRNA regulatory network

Interactions between lncRNAs and miRNAs were predicted by MiRcode (http://www.mircode.org/). Target mRNAs implicated in the lncRNA-miRNA regulatory network was explored by miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/), TargetScan (http://www.targetscan.org/), miRDB (http://www.mirdb.org/), and StarBase (http://starbase.sysu.edu.cn/) databases.

2.6. Construction of a prognostic model for COAD

The R "caret" package[19] was used to randomly divide the TCGA cohort into training and testing sets. A univariate Cox regression analysis was used to identify the relationship between mRNAs expression level and the patient's overall survival. Following, multivariate Cox analysis was used to evaluate the corresponding coefficients of the selected genes. The risk score of each COAD patient can be calculated based on a prognostic gene signature. Based on the training set risk score formula, COAD patients were divided into low-risk and high-risk groups, respectively. Kaplan–Meier survival analyses were used to evaluate the differences in survival of COAD patients between these 2 groups. The receiver operating characteristic (ROC) curve for 1-, 3-, and 5-year were established to compare the specificity and sensitivity of the OS prediction based on the risk score. All analyses were conducted by the R package "glment," "survminer," "survival," and "survivalROC."

2.7. The nomogram establishing

Nomogram survival model based on the prognostic signature and clinical information for predicting the survival and risk information of COAD was performed using the R package "rms." The concordance index was used to assess the capability of prediction. The patients with COAD were separated to various risk clusters along with their risk scores.

3. Results

3.1. Identification of DEmRNAs, DElncRNAs, and DEmiRNAs

We explored the RNA expression levels in 426 COAD samples and 818 normal samples, and the trimmed mean of M values scaling method was used to normalize all read counts of mRNA, lncRNA, and miRNA by edgeR packages. Under the defined thresholds, we found that 6365 differentially expressed RNAs, including 3537 DEmRNAs (1391 down-regulated and 2146 up-regulated), 2379 DElncRNAs (1448 down-regulated and 931 up-regulated), and 449 DEmiRNAs (366 down-regulated and 83 up-regulated). Volcano map (Fig. 1A–C) showed the expression change of all the significantly DEmRNAs, DElncRNAs, and DEmiRNAs based on the two dimensions of defined thresholds, respectively. Up-regulated mRNAs were enriched in the regulation of transmembrane transporter activity, muscle contraction, and muscle system process in BP (Fig. 1D). The up-regulated mRNAs and their interactions in BP were shown in Figure 1E. Moreover, cytosolic DNA-sensing pathway and amyotrophic lateral sclerosis were up-regulated while IL-17 signaling pathway and viral protein interaction with cytokine were down-regulated by KEGG-GSEA (Fig. 1F).

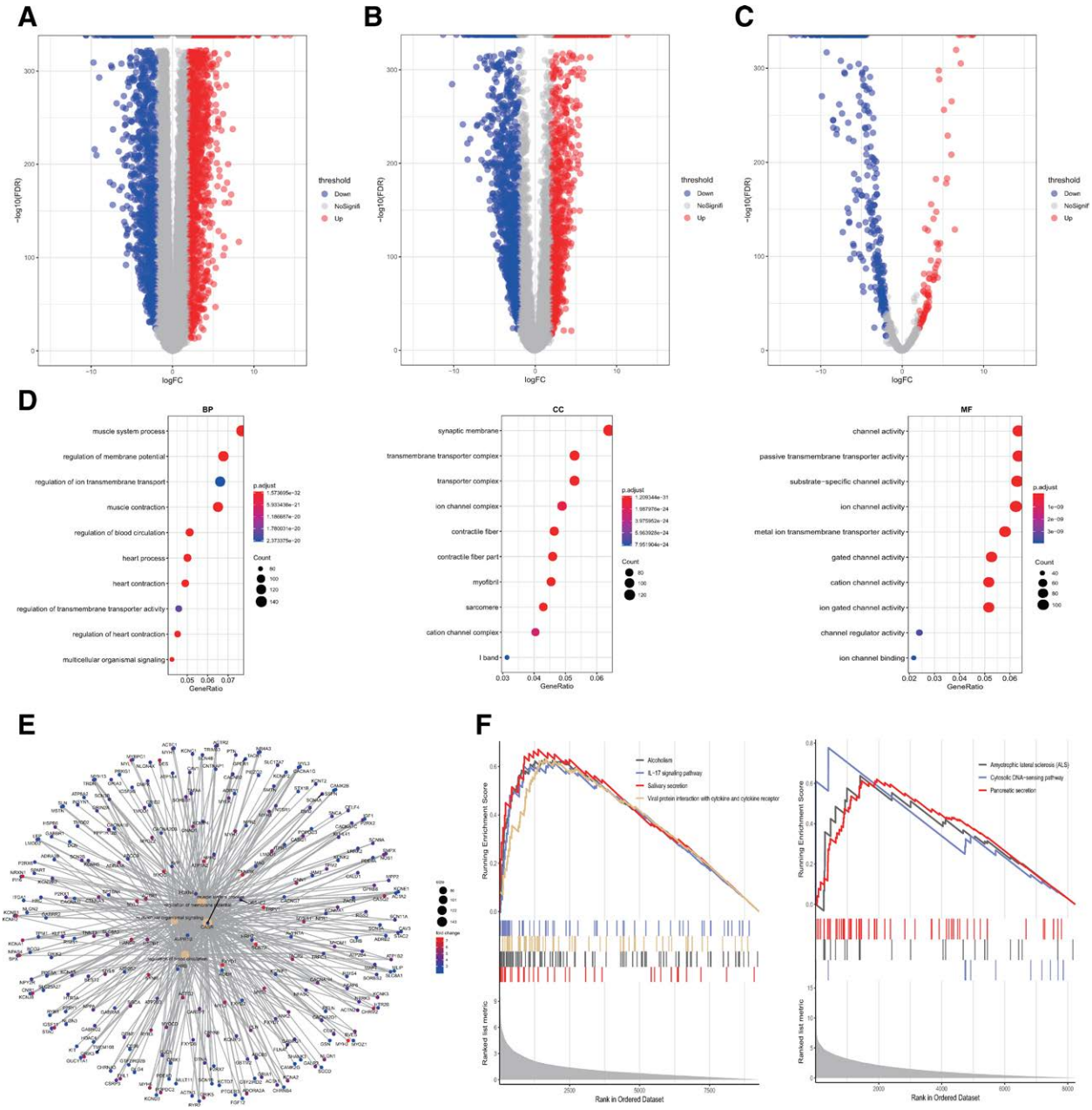

Figure 1. Different expression RNAs from data between TCGA and GTEx is analyzed. (A) Volcano map of significantly different expression of mRNAs. Red spots represent up-regulated genes, and blue spots represent down-regulated genes. (B) Volcano map of significantly different expression of lncRNAs. (C) Volcano map of significantly different expression of miRNAs. (D) Information from up-regulated genes was applied to GO analysis in BP, CC, and MF. (E) Gene symbols and interaction of the significantly up-regulated mRNAs in BP were shown. (F) KEGG-GSEA was applied for signaling pathway analysis. BP = biological process, CC = cellular component, GO = gene ontology, GTEx = the Genotype-Tissue Expression, KEGG = Kyoto Encyclopedia of Genes and Genomes, lncRNAs = long non-coding RNAs, MF = molecular function, mRNAs = messenger RNAs, miRNAs = microRNAs, TCGA = The Cancer Genome Atlas.

3.2. WGCNA is applied to analyze gene modules

Variance comparison methods were used to select the first 40% mRNAs, and the gene modules were analyzed using the WGCNA. Coexpressed gene modules were identified with power $\beta = 10$ (soft threshold), and optimal module size = 25 (Fig. 2A). We identified 11 gene color modules, and the adjacency and topological overlap matrix reflect gene coexpression similarity among the 11 color modules were analyzed between COAD and normal. The blue module included 3086 mRNAs that showed highly correlated with COAD (Fig. 2B). GO-GSEA function enrichment analysis was performed and the top interactions in BP terms were related to these 3086 mRNAs as shown in Figure 2C. The genes showed a high relationship with cell cycle, tissue development, and cell cycle process. Besides, genes were highly enriched in the proteasome, protein processing in the endoplasmic reticulum, RNA transport, and spliceosome by KEGG analysis (Fig. 2D).

3.3. LncRNAs modules are analyzed by WGCNA

Following, we investigated the lncRNAs coexpression network. Variance comparison methods were used to select the first 80% lncRNAs, and the gene modules were analyzed using the WGCNA. As shown in Figure 3A, 29 coexpressed lncRNA modules were identified with power $\beta = 10$ (soft threshold), and optimal module size = 25. Correlation analysis suggested that the blue

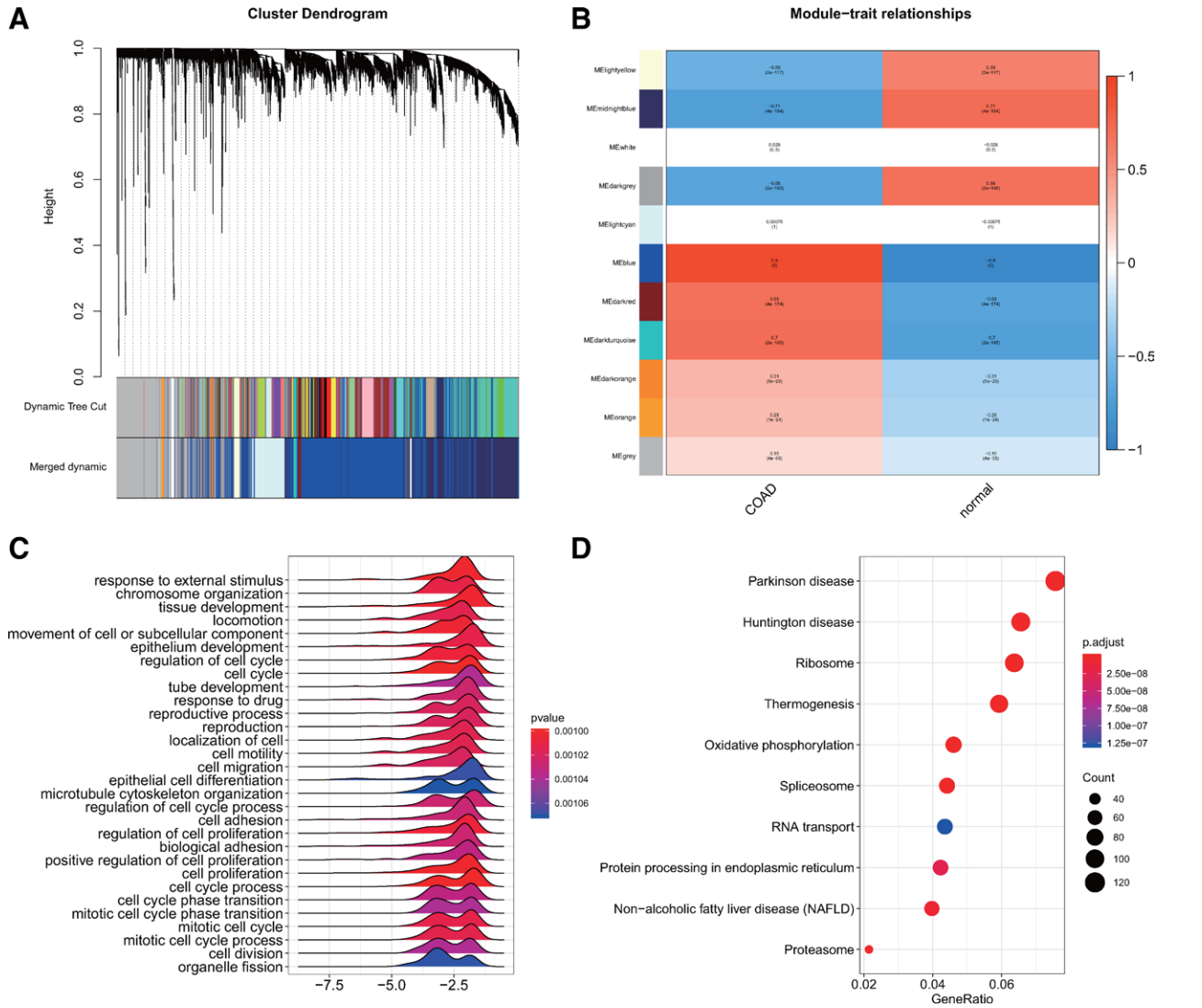

Figure 2. WGCNA is applied to analyze gene modules. (A) Cluster dendrogram of the coexpression network modules was produced based on topological overlap in the mRNAs. (B) The relation of genes in modules between COAD and normal samples was investigated. (C) GO-GSEA displayed the gene symbols and gene interaction in blue module. (D) KEGG analysis was used to investigate the pathway enrichment in blue module. COAD = colon adenocarcinoma, GO = gene ontology, GOSEA = Gene Set Enrichment Analysis, KEGG = Kyoto Encyclopedia of Genes and Genomes, mRNAs = messenger RNAs, WGCNA = weighted correlation network analysis.

module displayed highly correlation with COAD (Fig. 3B and C; $R = 0.86$). Then, miRNAs sponged by 1097 lncRNAs were predicted by miRcode to construct lncRNAs-miRcode-miRNAs relationship, and the first TCGA 400 miRNA-Seq with the highest expression was analyzed simultaneously. Then the miRNAs among 325 lncRNAs-miRcode-miRNAs and first 400 miRNAs were selected to obtain overlapped 72 lncRNAs-miRNAs.

The starBase, miRDB, miRTarBase, and Targetscan dataset were used to predict 9429 target mRNAs, which might be related with 72 miRNAs. Importantly, as shown in Figure 3D, overlapped target mRNAs were selected by analyzing the 9459 predicted target mRNAs, 3086 WGCNA blue module mRNAs, as well as 2146 up-regulated and 1391down-regulated DEmRNAs by edgeR. Finally, 66 up-regulated mRNAs and 201 down-regulated mRNAs were obtained. The heatmap of the expression of these 267 genes in 1244 samples was shown in Figure 3E.

3.4. Construction and evaluation of the prognostic mRNAs from the training set

We randomly divided the 426 TCGA COAD patients into a training (n = 214) set for the establishment of a prognostic model or a testing set (n = 212) for internal self-validation, respectively (Table 1). Based on the training set, we conducted a univariate Cox regression analysis to clarify the relationship between the expression levels of 268 genes and overall survival. Based on the univariable Cox regression analysis with the threshold of P value < .05, we obtained 32 genes to be significantly related to COAD. Then, these 32 genes were used for further multivariate Cox analysis (Table 2). Meanwhile, the risk score for COAD patients' survival prediction model based on the results of multivariate Cox analysis was constructed.

We then set up a survival model for prediction of COAD patients, and 8 genes are as following: CCNA2 + CEBPA + N EBL + SOX9 + DLG4 + RIMKLB + TCF7L1 + TUB. The results showed that CCNA2, CEBPA, NEBL, and SOX9 were up-regulated while DLG4, RIMKLB, TCF7L1, and TUB were down-regulated in COAD patients (Fig. 4A). Based on the multivariate Cox score, TCGA training set patients were divided into predicted low-risk or high-risk group (Fig. 4B). Furthermore, the heatmap of the expression of 8 genes in low-risk or high-risk group was shown in Figure 4B. We also evaluated the predictive accuracy of the 8 genes prognostic model on survival prediction. Kaplan–Meier survival curves demonstrated that patients with

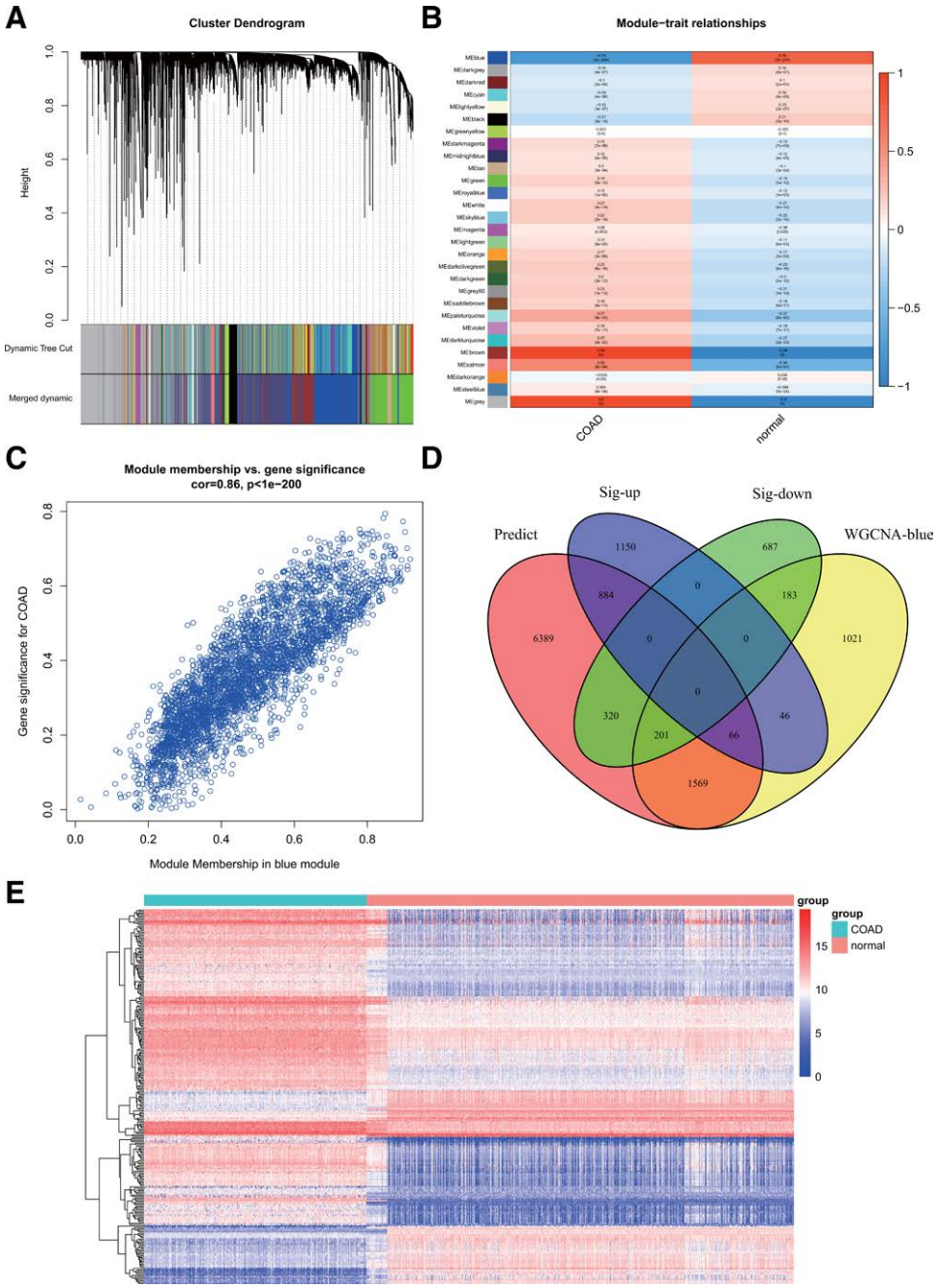

Figure 3. LncRNAs modules are analyzed by WGCNA. (A) Cluster dendrogram of the coexpression network modules was produced based on topological overlap in the lncRNAs. (B) The relation of lncRNAs in modules between COAD and normal samples was investigated. (C) Blue module showed highest relationship with COAD. (D) Overlapped target mRNAs were analyzed by the predicted target mRNAs, WGCNA-blue mRNAs, and the significantly up-regulated mRNAs and down-regulated mRNAs. (E) The expression of 267 selected target genes was displayed by heatmap. COAD = colon adenocarcinoma, lncRNAs = long non-coding RNAs, mRNAs = messenger RNAs, WGCNA = weighted correlation network analysis.

predicted low risk (n = 107) had significantly longer overall survival than those with high risk (n = 107, P = .002, Fig. 4C). We performed ROC analysis to evaluate the predictive sensitivity and specificity of models. TCGA training set demonstrated that the area under receiver operating characteristic curve of the 8 genes signature for 1-, 3-, and 5-year overall survival were 0.634, 0.744, and 0.784, respectively (Fig. 4D).

3.5. Validation of the 8-genes signature in testing set and the entire TCGA data set

Next, 8 genes signature in the testing set will be validated to confirm our findings. We calculated the risk score for each patient in the testing set based on the risk score formula of the training set. Then, we divided COAD patients into a low-risk group (n = 101) and a high-risk group (n = 111) using the same threshold, and the result was shown in Figure 4E. Furthermore, the heatmap of the 8 genes expression at low-risk or high-risk group was shown in Figure 4E. Similar results were shown in the testing set, patients in the low risk group had significantly longer overall survival than those in the higher risk group (P = .046, Fig. 4F). In the entire TCGA data set, the heatmap of the 8 genes expression was shown in Figure 4H. The consistent result was shown that patients in the low-risk group had significantly longer survival than those in the high-risk group (P = 1.46e-4; Fig. 4I). Time-dependent area under the ROC curves analysis

Table 1

Clinical pathological characteristics of patients in the training and testing set of TCGA.

Characteristics	TCGA training set (N = 214)	TCGA testing set (N = 212)	Entire TCGA set (N = 426)
Age at initial diagnosis (yr)	66.65 ± 12.48	66.36 ± 13.07	66.51 ± 12.76
Gender			
Male	119	109	228
Female	95	103	198
Pathologic M			
0	157	160	317
1	32	26	58
Not report	25	26	51
Pathologic N			
0	133	120	253
1	48	51	99
2	33	41	74
Pathologic T			
1	4	6	10
2	32	42	74
3	146	145	291
4	31	19	50
Not report	1	0	1
AJCC stage			
Stage I	32	41	73
Stage II	93	72	165
Stage III	51	68	119
Stage IV	32	26	58
Not report	6	5	11
Overall survival time (yr)	1.94 (1.04–3.00)	1.96 (1.17–3.02)	1.94 (1.09–3.00)
Overall survival status			
Alive	165	170	335
Dead	49	42	91

AJCC = American Joint Committee on Cancer, TCGA = The Cancer Genome Atlas.

Table 2

Multivariate Cox proportional hazard regression analysis of 32 genes.

Gene ID	Gene symbol	HR (95% CI)	*P*	HR (95%CI)	*P*
ENSG00000152284	TCF7L1	1.63 (1.23, 2.17)	.000791	1.91 (1.04, 3.50)	.03555*
ENSG00000166532	RIMKLB	1.32 (1.11, 1.58)	.002065	1.50 (1.06, 2.12)	.021575*
ENSG00000154277	UCHL1	1.22 (1.07, 1.38)	.002451		
ENSG00000078114	NEBL	0.75 (0.62, 0.91)	.003852	0.63 (0.45, 0.90)	.010261*
ENSG00000133216	EPHB2	0.73 (0.58, 0.92)	.006661		
ENSG00000039068	CDH1	0.64 (0.46, 0.89)	.007785		
ENSG00000124882	EREG	0.90 (0.83, 0.97)	.009143		
ENSG00000175538	KCNE3	0.75 (0.60, 0.94)	.010948		
ENSG00000154639	CXADR	0.74 (0.58, 0.93)	.011367		
ENSG00000243335	KCTD7	1.45 (1.08, 1.95)	.012434		
ENSG00000182481	KPNA2	0.61 (0.42, 0.90)	.012996		
ENSG00000116771	AGMAT	0.73 (0.56, 0.94)	.014053		
ENSG00000142279	WTIP	1.29 (1.05, 1.59)	.014225		
ENSG00000166402	TUB	1.20 (1.04, 1.40)	.015136	0.73 (0.54, 0.98)	.036628*
ENSG00000139625	MAP3K12	1.42 (1.07, 1.88)	.015574		
ENSG00000099864	PALM	1.21 (1.04, 1.41)	.01596		
ENSG00000132535	DLG4	1.36 (1.06, 1.74)	.016687	0.64 (0.39, 1.05)	.079441
ENSG00000064651	SLC12A2	0.77 (0.62, 0.96)	.019082		
ENSG00000168646	AXIN2	0.86 (0.75, 0.98)	.021272		
ENSG00000135525	MAP7	0.62 (0.41, 0.93)	.021735		
ENSG00000184992	BRI3BP	0.66 (0.47, 0.94)	.022954		
ENSG00000145386	CCNA2	0.72 (0.54, 0.96)	.02326	0.62 (0.39, 0.97)	.034915*
ENSG00000164109	MAD2L1	0.73 (0.55, 0.96)	.023297		
ENSG00000164398	ACSL6	0.93 (0.87, 0.99)	.027766		
ENSG00000136002	ARHGEF4	1.14 (1.01, 1.29)	.029676		
ENSG00000245848	CEBPA	0.80 (0.65, 0.98)	.031077	0.73 (0.50, 1.08)	.112413
ENSG00000094963	FMO2	1.12 (1.01, 1.24)	.031287		
ENSG00000117707	PROX1	0.85 (0.73, 0.99)	.039223		
ENSG00000198805	PNP	0.68 (0.47, 0.99)	.041559		
ENSG00000180817	PPA1	0.67 (0.46, 0.99)	.046401		
ENSG00000125398	SOX9	0.77 (0.59, 1.00)	.048372	0.68 (0.42, 1.10)	.114067
ENSG00000139998	RAB15	0.76 (0.58, 1.00)	.048525		

*significant difference.

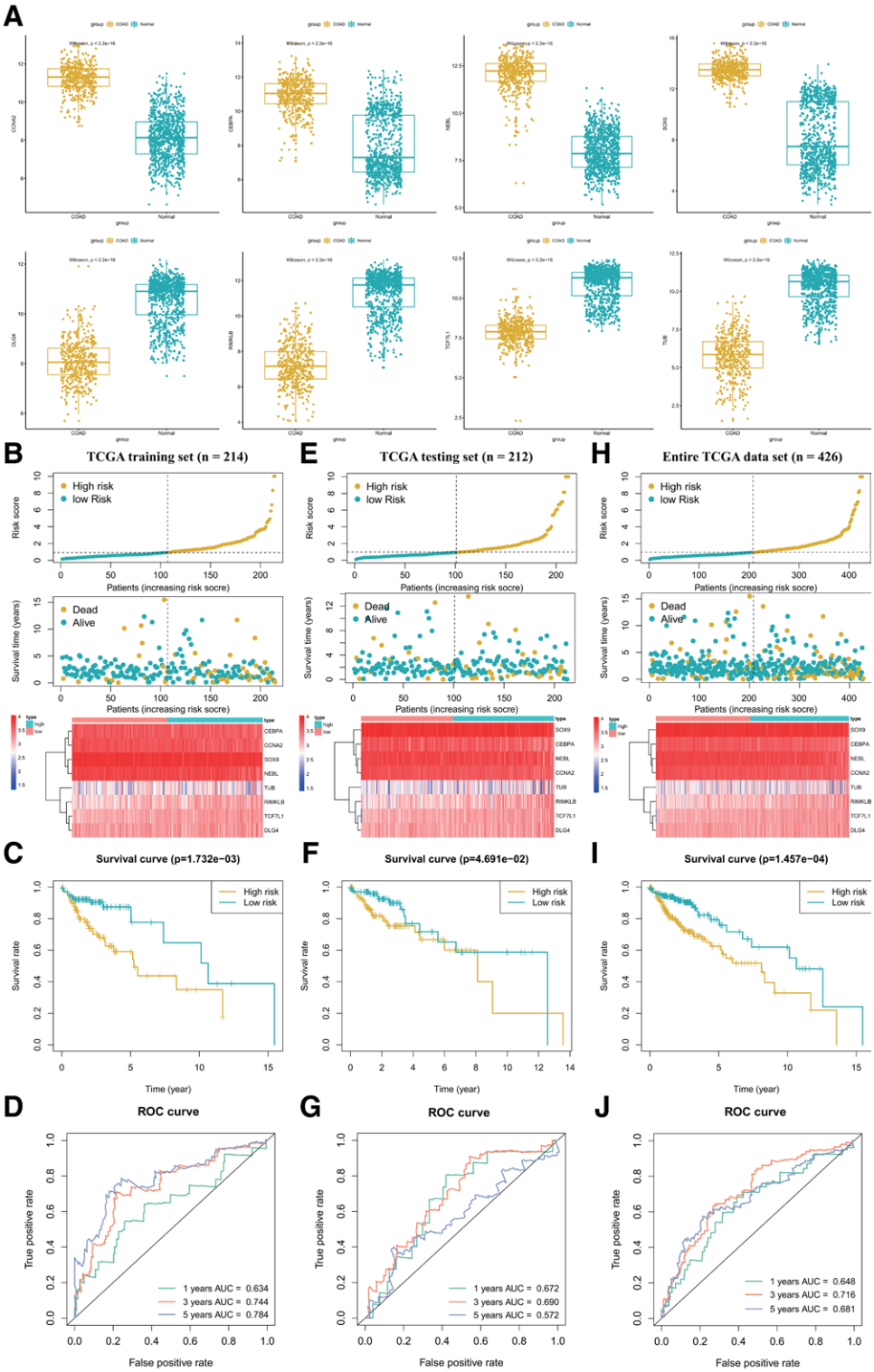

Figure 4. Survival analysis and development of the prognostic scoring model of the 8 genes in TCGA cohorts. (A) The expression of 8 selected genes between COAD and normal samples was shown. (B, E, H) Correlation between the prognostic signature and the overall survival of patients in the TCGA training set (B), TCGA testing set (E), and entire TCGA data set (H). (C, F, I) Kaplan–Meier survival curves of overall survival among risk stratification groups in the TCGA training set (C) and TCGA testing set (F), and entire TCGA data set (I). (D, G, J) ROC curves with calculated AUCs for risk prediction in 1-, 3-, 5-years in the TCGA training set (D) and TCGA testing set (G), and entire TCGA data set (J). AUC = area under receiver operating characteristic curve, COAD = colon adenocarcinoma, ROC = receiver operating characteristic, TCGA = The Cancer Genome Atlas.

for the 8 genes signature prediction model for 1-, 3-, and 5-year achieved area under receiver operating characteristic curve score of 0.672, 0.69, 0.572 and 0.648, 0.716, 0.681 in the testing set and the entire set (Fig. 4G and J), respectively.

3.6. Independent other clinical prognostic variables

By performing univariable and multivariable Cox regression analyses, the independent factors, such as age, gender, pathologic M, pathologic N, pathologic T, and American Joint Committee on

Cancer (AJCC) stage at diagnosis were selected to evaluate the predictive capacity of 8-genes signature. The univariate Cox regression results suggested that grade, stage, N stage, and risk score in the TCGA training set (pathologic M: $P < .001$, pathologic N: $P < .001$, pathologic T: $P < .05$, AJCC stage: $P < .001$, risk score: $P < .001$; Table 3). On the other hand, multivariate Cox regression relevant that age (HR = 1.06, 95% CI [1.03, 1.10]; $P < .001$; Table 3), and risk score (HR = 1.35; 95% CI [1.20, 1.52]; $P < .001$; Table 3) was significant independent risk factors in the training set. Similar results were observed in the TCGA testing set and entire set.

3.7. Construction and validation of the nomogram survival model

The nomogram survival model was established based on the three independent clinical variables (Fig. 5A). The nomogram can easily calculate 1-year, 3-year, and 5-year survival prediction value of patients based on their age, gender, pathologic M, pathologic N, pathologic T, and AJCC stage at diagnosis. Besides, the nomogram-predicted 1-year, 3-year, or 5-year survival prediction value was further compared to the actual 1-year, 3-year, or 5-year virtual survival probability of TCGA patients. The results showed that there is high consistency between nomogram predicted probability of survival and TCGA patient's virtual survival probability (Fig. 5B). Hence, the results indicated a good performance evaluation of the nomogram survival model.

3.8. Constructed lncRNA-miRNA-mRNA ceRNA network

Finally, we constructed the correlations network between the target genes and their corresponding miRNAs. The results showed that miR-206, miR-212-3p, miR-22-3p, and miR-429 could target the 6 mRNAs, respectively. Such as, miR-206 targeted DLG4 and SOX9, while miR-212-3p, miR-22-3p, and miR-429 regulated CCNA2 (Fig. 6A). The data of lncRNAs were also provided by TCGA and GTEx, edgeR was also used to analyze to obtain DElncRNA. We identified 1448 down-regulated and 931 up-regulated lncRNAs. Next, we overlapped 68 DElncRNAs between these 2379 DElncRNAs and 125 predicted lncRNAs from 4 miRNAs. Finally, we used 68 lncRNAs, 4 miRNAs, and 6 mRNAs to establish a lncRNA-miRNA-mRNA ceRNA network, as shown in Figure 6B.

4. Discussion

In our study, we identified a total of 6365 differentially expressed RNAs (including 3537 DEmRNAs, 2379 DElncRNAs, and 449 DEmiRNAs) between COAD and normal samples. After the univariate Cox regression analysis was used to select 32 prognosis-associated genes, and then, the multivariate Cox regression analysis was used to screen out 8 independent prognosis-associated genes (including CCNA2, CEBPA, NEBL, SOX9, DLG4, RIMKLB, TCF7L1, and TUB). On the other hand, pathologic M, pathologic N, pathologic T, and AJCC stage at diagnosis were identified to be the independent clinical prognostic factors, and those were used to establish the nomogram survival model. Meanwhile, Kaplan–Meier survival curves of our survival prognostic model showed that patients with predicted low risk had significantly longer OS time than those with high risk. Finally, we identified 68 lncRNAs, 4 miRNAs, and 6 mRNAs to establish a lncRNA-miRNA-mRNA ceRNA network.

RNA-Seq data tend to be understood from a clinical transformation perspective in the precision oncology medicine era.[20,21] It is necessary to obtain all the available information to identify and provide the most relevant biomarkers in critical and comprehensive analysis. WGCNA is a useful bioinformatics tool that identifies clusters of functional modules genes, investigates the molecular mechanisms of multiple malignancies, and therefore can identify clinically relevant markers.[22–25] WGCNA has been successfully used to identify hub module genes associated with prognosis and progression of pancreatic carcinoma[26] and to demarcate the transcriptional subtypes of glioblastoma.[27] LncRNA-miRNA-mRNA ceRNA network plays an important role in various BPs that can be predicted for cancer prognosis.[28] For instance, the previous study based on RNA-Seq data constructed a ceRNA regulatory network of acute myeloid leukemia,

Table 3

Univariate and multivariate Cox regression analyses of clinical factors associated with overall survival.

	Univariable Cox		Multivariable Cox	
Clinical characteristics	**HR (95% CI)**	***P***	**HR (95% CI)**	***P***
Training set (n = 214)				
Age	1.03 (1.00, 1.05)	.049988	1.06 (1.03, 1.10)	.000404
Gender	0.84 (0.47, 1.49)	.54573	0.98 (0.50, 1.90)	0.95014
Pathologic M	7.51 (3.98, 14.16)	4.64E-10	3.62 (0.69, 18.96)	.12782
Pathologic N	1.89 (1.33, 2.68)	.000363	0.63 (0.29, 1.35)	.236389
Pathologic T	2.26 (1.28, 3.98)	.004866	1.12 (0.52, 2.44)	.767589
AJCC stage	2.43 (1.71, 3.45)	6.17E-07	2.21 (0.69, 7.15)	.183987
Risk score	1.27 (1.17, 1.39)	6.29E-08	1.35 (1.20, 1.52)	9.64E-07
Testing set (n = 212)				
Age	1.01 (0.98, 1.04)	.42674	1.03 (1.00, 1.06)	.076459
Gender	1.11 (0.6, 2.07)	.742161	1.11 (0.55, 2.25)	.772347
Pathologic M	2.72 (1.23, 6.06)	.013969	1.17 (0.23, 5.96)	.8476
Pathologic N	2.52 (1.73, 3.68)	1.49E-06	2.42 (1.06, 5.50)	.035078
Pathologic T	3.55 (1.77, 7.14)	.000367	4.43 (1.64, 11.96)	.003345
AJCC stage	2.11 (1.45, 3.08)	9.82E-05	0.84 (0.21, 3.40)	.804537
Risk score	1.04 (0.96, 1.13)	.338998	1.04 (0.89, 1.22)	.603667
Entire set (n = 426)				
Age	1.02 (1.00, 1.04)	.052808	1.04 (1.02, 1.06)	.000317
Gender	0.96 (0.63, 1.46)	.859286	1.18 (0.74, 1.88)	.493178
Pathologic M	5.12 (3.19, 8.21)	1.25E-11	1.81 (0.61, 5.32)	.282214
Pathologic N	2.15 (1.67, 2.77)	2.63E-09	1.19 (0.73, 1.95)	.490903
Pathologic T	2.70 (1.76, 4.14)	5.02E-06	1.95 (1.11, 3.44)	.019919
AJCC stage	2.31 (1.79, 2.98)	1.48E-10	1.64 (0.72, 3.71)	.237539
Risk score	1.10 (1.04, 1.15)	.000222	1.20 (1.12, 1.30)	7.03E-07

AJCC = American Joint Committee on Cancer.

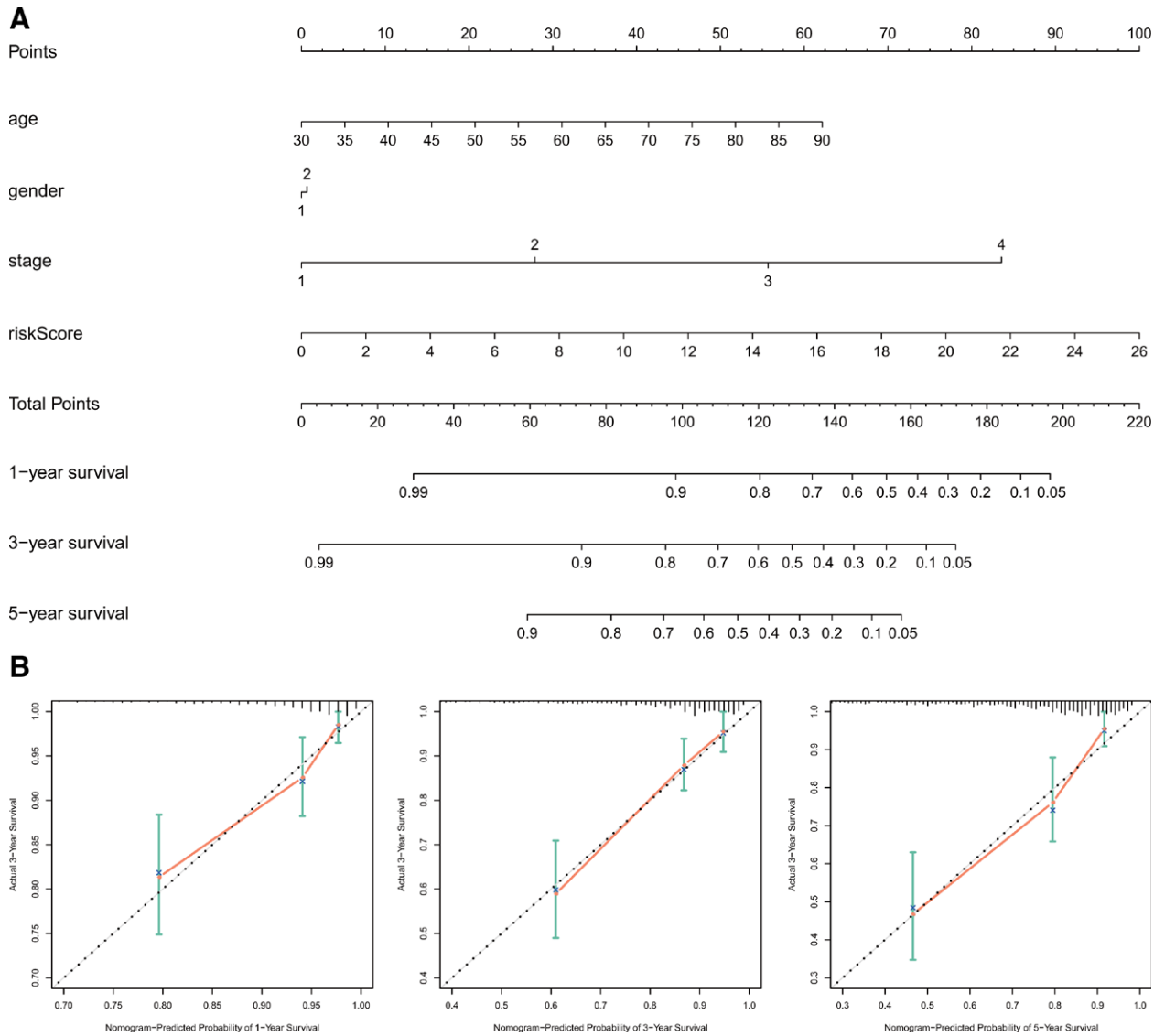

Figure 5. Nomogram survival model. (A) The nomogram consists of age, grade, stage, and the risk score based on the 8-genes signature. (B) Calibration curves of the nomogram for the estimation of survival rates at 1-, 3-, 5- years.

and the results suggested that 108 lncRNAs, 10 miRNAs, and 8 mRNAs were used to build a lncRNA-miRNA-mRNA ceRNA network which might be as prognostic markers of acute myeloid leukemia.[29] Another study compared recurrence and non-recurrence sample of COAD, high-throughput sequencing data of COAD from TCGA, a lncRNA-miRNA-mRNA ceRNA regulatory network was constructed, which comprised 3 lncRNAs, 4 miRNAs, and 77 mRNAs, and the results relevant that 3 of those lncRNAs had significant prognostic value based on multivariate Cox regression analysis.[30] For COAD, the previous study constructed ceRNA regulatory networks based on 133 DElncRNAs, 29 DEmiRNAs, and 55 DEmRNAs and evaluated those RNAs on overall survival.[31] However, the detail of the previous study did not be presented in the prognosis of COAD patients. On the other hand, nomogram is a useful tool for cancer prognostic, which can establish an individual probability by integrating diverse prognostic and determinant variables factors according to corresponding clinical prognostic characteristics. For example, in the study of RNA-Seq data of esophageal squamous cell carcinoma data from TCGA, a lncRNA prognosis prediction nomogram was established. Eight lncRNAs (including AP000487, AC011997, LINC01592, LINC01497, LINC01711, FENDRR, AC087045, AC137770) have been identified with significant prognostic value, and a nomogram based on clinical factors was built with good accuracy for predicting patients survival probability.[32]

Among the 8 genes in our study, CCNA2 and SOX9 had been certified to be essential in COAD pathogenesis and progression. Cyclin A2 (CCNA2) was identified as a novel target of miR-22 in colon cancer, with deeper research in CCNA2 regulatory COAD, there was a very interesting topic for COAD with CCNA2 overexpression.[33] For colorectal cancer, miR-22 even has a more profound function of tumor-suppressive. Compared to its adjacent normal mucosa, miR-22 has been identified as a significantly down-regulated microRNA in colorectal cancer tissue; on the other hand, it can improve the sensitivity of 5-FU and paclitaxel sensitivity in chemotherapy.[34] Hence, our study showed a similar relationship between CCNA2 and miR-22 in COAD and may provide attractive potential novel therapeutic targets. Those results are needed to be investigated for COAD prevention and treatment.

SOX9 is a high-mobility group box containing transcription factor that plays a key role in organ development, embryogenesis, and maintenance of stem or progenitor cells,[35–37] and present extensive studies showed that SOX family member primarily expressed at the bottom of the crypts such as in the stem or progenitor cell compartment[38–40] of the colon, small intestines,

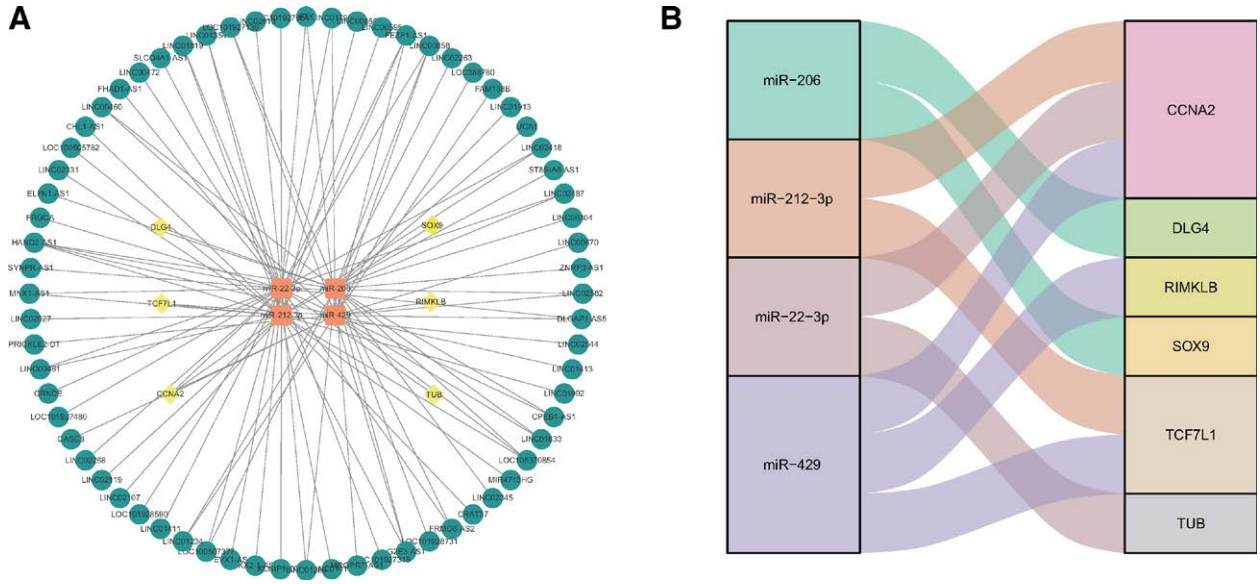

Figure 6. lncRNA-miRNA-mRNA ceRNA network. (A) A lncRNA-miRNA-mRNA ceRNA network was constructed by 68 lncRNAs, 4 miRNAs, and 6 mRNAs for COAD prognosis. (B) The relationship between the 6 target genes and their corresponding miRNA was shown. ceRNAs = competing endogenous RNAs, COAD = colon adenocarcinoma, lncRNAs = long non-coding RNAs, mRNAs = messenger RNAs, miRNAs = microRNAs.

and in the tuft cells along the villi of the small intestine.[41] SOX9 as an oncogene, the disorder of it has been further implicated in the progression of cancer, which promotes cell proliferation, facilitates transformation, and inhibits senescence.[42] Another previous study showed that lncRNA-miR-206-SOX9 regulatory network may suggest a novel therapeutic target for esophageal squamous cell carcinoma.[43] The main purposes of our study are to investigate mRNAs in the ceRNA regulatory network, and those mRNAs suggesting might be related to the prognosis prediction of COAD.

Some limitations of our study have been found. Firstly, the differential expression and prognostic prediction of the 8 genes were constructed from RNA-Seq data from TCGA and GTEx, and the results of these genes in COAD patients are urgently needed to be validated in further study by experiment validation. Secondly, for another result of our study, the genes included in the ceRNAs correlation network should also be validated in vivo and in vitro experimental studies.

In conclusion, 3537 DEmRNAs, 449 DEmiRNAs, and 2379 DElncRNAs were identified between COAD and normal samples. The risk score based on the involving 8 genes (CCNA2, CEBPA, NEBL, SOX9, DLG4, RIMKLB, TCF7L1, and TUB) for overall survival was identified. Based on the clinical factors and those genes prognostic signature, a nomogram was constructed. The nomogram survival model could be practical and reliable for COAD prediction of COAD progress. Finally, the ceRNA network was defined in our study from multiple dimensions, and provides potentially prognostic markers and molecular diagnostic, which will help us understand the potential mRNA-related regulatory mechanism about ceRNA network-mediated COAD progress. Further, more experiment studies are urgently needed to elucidate and evaluate the miRNA-related molecular mechanisms underlying COAD.

Author contributions

Data curation: Jiaxi Xi, Huajun Zhang.
Formal analysis: Jiaxi Xi.
Funding acquisition: Xiaoyu Chen.
Methodology: Jiaxi Xi, Huajun Zhang, Yan Li, Xiaoyu Chen, Xueyan Liang.
Software: Jiaxi Xi, Huajun Zhang, Yan Li, Henghai Su, Xueyan Liang.
Validation: Jiaxi Xi, Huajun Zhang, Yan Li, Henghai Su, Xiaoyu Chen, Xueyan Liang.
Writing – original draft: Jiaxi Xi, Huajun Zhang.
Writing – review & editing: Xiaoyu Chen, Xueyan Liang.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019;69:7–34.
[2] Corley DA, Jensen CD, Marks AR, et al. Adenoma detection rate and risk of colorectal cancer and death. N Engl J Med. 2014;370:1298–306.
[3] Radice E, Miranda V, Bellone G. Low-doses of sequential-kinetic-activated interferon-γ enhance the ex vivo cytotoxicity of peripheral blood natural killer cells from patients with early-stage colorectal cancer. A preliminary study. Int Immunopharmacol. 2014;19:66–73.
[4] Siegel RL, Miller KD, Fedewa SA, et al. Colorectal cancer statistics, 2017. CA Cancer J Clin. 2017;67:177–93.
[5] Liska D, Stocchi L, Karagkounis G, et al. Incidence, patterns, and predictors of locoregional recurrence in colon cancer. Ann Surg Oncol. 2017;24:1093–9.
[6] Qi X, Zhang DH, Wu N, et al. ceRNA in cancer: possible functions and clinical implications. J Med Genet. 2015;52:710–8.
[7] Ala U, Karreth FA, Bosia C, et al. Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. Proc Natl Acad Sci USA. 2013;110:7154–9.
[8] Conte F, Fiscon G, Chiara M, et al. Role of the long non-coding RNA PVT1 in the dysregulation of the ceRNA-ceRNA network in human breast cancer. PLoS One. 2017;12:e0171661.
[9] Salmena L, Poliseno L, Tay Y, et al. A ceRNA hypothesis: the rosetta stone of a hidden RNA language? Cell. 2011;146:353–8.
[10] Sanchez-Mejias A, Tay Y. Competing endogenous RNA networks: tying the essential knots for cancer biology and therapeutics. J Hematol Oncol. 2015;8:30.
[11] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
[12] Lai YA. Statistical method for the conservative adjustment of false discovery rate (q-value). BMC Bioinf. 2017;18:69.
[13] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40:4288–97.

[14] Wickham H. Getting started with ggplot2. In: ggplot2: Elegant Graphics for Data Analysis. Cham: Springer International Publishing; 2016.
[15] Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9.
[16] Yu G, Wang LG, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.
[17] Kanehisa M, Goto S, Furumichi M, et al. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010;38:D355–360.
[18] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf. 2008;9:559.
[19] Software MKJJoS. Building predictive models in R using the caret package. J Stat Softw. 2008;28:1–26.
[20] Van Allen EM, Wagle N, Stojanov P, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med. 2014;20:682–8.
[21] Roychowdhury S, Iyer MK, Robinson DR, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. Sci Transl Med. 2011;3:111ra121.
[22] Bailey P, Chang DK, Nones K, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature. 2016;531:47–52.
[23] Zhang XJ, Cheng X, Yan ZZ, et al. An ALOX12-12-HETE-GPR31 signaling axis is a key mediator of hepatic ischemia-reperfusion injury. Nat Med. 2018;24:73–83.
[24] Lai J, Chen B, Zhang G, et al. Identification of a novel microRNA recurrence-related signature and risk stratification system in breast cancer. Aging. 2019;11:7525–36.
[25] Zhou Z, Mo S, Dai W, et al. Development and validation of an autophagy score signature for the prediction of post-operative survival in colorectal cancer. Front Oncol. 2019;9:878.
[26] Zhou Z, Cheng Y, Jiang Y, et al. Ten hub genes associated with progression and prognosis of pancreatic carcinoma identified by co-expression analysis. Int J Biol Sci. 2018;14:124–36.
[27] Pan YB, Wang S, Yang B, et al. Transcriptome analyses reveal molecular mechanisms underlying phenotypic differences among transcriptional subtypes of glioblastoma. J Cell Mol Med. 2020;24:3901–16.
[28] Hu J, Xu L, Shou T, et al. Systematic analysis identifies three-lncRNA signature as a potentially prognostic biomarker for lung squamous cell carcinoma using bioinformatics strategy. Transl Lung Cancer Res. 2019;8:614–35.
[29] Wang JD, Zhou HS, Tu XX, et al. Prediction of competing endogenous RNA coexpression network as prognostic markers in AML. Aging. 2019;11:3333–47.
[30] Yang H, Lin HC, Liu H, et al. A 6 lncRNA-based risk score system for predicting the recurrence of colon adenocarcinoma patients. Front Oncol. 2020;10:81.
[31] Wang WJ, Li HT, Yu JP, et al. A competing endogenous RNA network reveals novel potential lncRNA, miRNA, and mRNA biomarkers in the prognosis of human colon adenocarcinoma. J Surg Res. 2019;235:22–33.
[32] Li W, Liu J, Zhao H. Identification of a nomogram based on long non-coding RNA to improve prognosis prediction of esophageal squamous cell carcinoma. Aging. 2020;12:1512–26.
[33] Yang F, Hu Y, Liu HX, et al. MiR-22-silenced cyclin A expression in colon and liver cancer cells is regulated by bile acid receptor. J Biol Chem. 2015;290:6507–15.
[34] Liu Y, Chen X, Cheng R, et al. The Jun/miR-22/HuR regulatory axis contributes to tumourigenesis in colorectal cancer. Mol Cancer. 2018;17:11.
[35] Sarkar A, Hochedlinger K. The sox family of transcription factors: versatile regulators of stem and progenitor cell fate. Cell Stem Cell. 2013;12:15–30.
[36] Guo W, Keckesova Z, Donaher JL, et al. Slug and Sox9 cooperatively determine the mammary stem cell state. Cell 2012;148:1015–28.
[37] Akiyama H, Chaboissier MC, Martin JF, et al. The transcription factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of SOX5 and SOX6. Genes Dev. 2002;16:2813–28.
[38] Bastide P, Darido C, Pannequin J, et al. Sox9 regulates cell proliferation and is required for Paneth cell differentiation in the intestinal epithelium. J Cell Biol. 2007;178:635–48.
[39] Blache P, van de Wetering M, Duluc I, et al. SOX9 is an intestine crypt transcription factor, is regulated by the Wnt pathway, and represses the CDX2 and MUC2 genes. J Cell Biol. 2004;166:37–47.
[40] Mori-Akiyama Y, van den Born M, van Es JH, et al. SOX9 is required for the differentiation of paneth cells in the intestinal epithelium. Gastroenterology 2007;133:539–46.
[41] Gerbe F, van Es JH, Makrini L, et al. Distinct ATOH1 and Neurog3 requirements define tuft cells as a new secretory cell type in the intestinal epithelium. J Cell Biol. 2011;192:767–80.
[42] Matheu A, Collado M, Wise C, et al. Oncogenicity of the developmental transcription factor SOX9. Cancer Res. 2012;72:1301–15.
[43] Wang L, Yu X, Zhang Z, et al. Linc-ROR promotes esophageal squamous cell carcinoma progression through the derepression of SOX9. J Exp Clin Cancer Res. 2017;36:182.