Establishment of a 7-microRNA prognostic signature and identification of hub target genes in colorectal carcinoma

Shengying Jiang^{1,2}, Xiaoli Xie¹, Huiqing Jiang^{1^}

¹Department of Gastroenterology, The Second Hospital of Hebei Medical University, Hebei Key Laboratory of Gastroenterology, Hebei Institute of Gastroenterology, Hebei Clinical Research Center for Digestive Diseases, Shijiazhuang, China; ²Endoscope Room, Department of General Surgery, Cangzhou Central Hospital, Cangzhou, China

Contributions: (I) Conception and design: S Jiang, H Jiang; (II) Administrative support: None; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: X Xie; (V) Data analysis and interpretation: S Jiang, X Xie; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Huiqing Jiang. Department of Gastroenterology, The Second Hospital of Hebei Medical University, Hebei Key Laboratory of Gastroenterology, Hebei Institute of Gastroenterology, Hebei Clinical Research Center for Digestive Diseases, No. 215, Heping West Road, Shijiazhuang 050000, China. Email: jianghq@aliyun.com.

Background: Colorectal cancer (CRC) is a malignant tumor with high morbidity and mortality, but there is still no recognized prognostic prediction model to better predict and intervene its prognosis. Our aim is to establish a novel microRNA (miRNA) signature and identify hub target genes for simply and accurately predicting survival risk for CRC patients and to provide therapeutic targets.

Methods: The miRNA expression profiles along with clinical data of 512 CRC patients were downloaded from the Cancer Genome Atlas (TCGA) database and randomly divided into training set and validation set. The signature was generated from the training set after a series of Cox regression analyses, including least absolute shrinkage and selection operator (LASSO)-Cox regression, and verified in the test set and the whole set. Furthermore, the signature was compared with clinical risk factors. Interaction network of target genes of the seven microRNAs was established. Functional enrichment analysis was performed to reveal the biological processes and pathways. GEPIA2 was used for prognostic analysis.

Results: A 7-microRNA prognostic signature was generated from the training set with the areas under the receiver operating characteristic (ROC) curve (AUC) of 5-year survival rate was 0.889. Its performance was well verified both in the test set and the entire set by Kaplan-Meier analysis (P value <0.05). Further analysis demonstrated that the signature was an independent prognostic risk factor for CRC patients and its predictive ability was superior to age and tumor-node-metastasis (TNM) stage. Interaction network found two major gene modules, and they might be involved in the activation of PI3K-Akt-mTOR and p53 signaling pathways, which related to epidermal growth factor receptor (EGFR) resistance. The GEPIA2 revealed that CDKN1A, eIF4E and SNAI1 were associated with CRC prognosis.

Conclusions: Our study demonstrated the potential of this novel 7-microRNA signature to independently predict overall survival in patients with CRC and provided potential therapeutic targets.

Keywords: Biomarker; TCGA (the Cancer Genome Atlas); colorectal cancer (CRC); microRNA (miRNA); prognosis

Submitted Sep 17, 2021. Accepted for publication Dec 22, 2021.

doi: 10.21037/tcr-21-1992

View this article at: <https://dx.doi.org/10.21037/tcr-21-1992>

[^] ORCID: 0000-0001-8706-0943.

Introduction

Colorectal cancer (CRC) has high morbidity and mortality, colorectal adenocarcinoma is the main form and accounting for more than 95% of CRC patients (1). Due to its high heterogeneity, traditional predictors such as age and tumor-node-metastasis (TNM) stage are not sufficient to accurately predict the survival risk of CRC patients. Exploring novel biomarkers is necessary to provide effective and personalized predictive tools. For the past few years, investigators have carried out a series of explorations in this field, and several prognostic gene signatures (2,3), transcriptional signatures and noncoding RNA signatures have been published (4-6). However, there is still no recognized prognostic prediction model, and further research is required.

MicroRNAs (miRNAs) are a class of endogenous non-coding single-stranded RNA molecules about 20–25 nucleotides with regulatory functions, and participates in a series of important processes in life, including early development, cell proliferation, apoptosis, cell death, fat metabolism and cell differentiation. Many miRNAs expression profiles related to particular malignancies have been found to have tumor-suppressive or oncogenic roles in different cancer types and further affect the prognosis of patients. Moreover, the functions of miRNAs are involved in the occurrence, development and metastasis of tumors (7). For instance, Mirzaei *et al.* (8) reported that miR-29b has significant tumor-suppressive effects on chronic lymphocytic leukemia (CLL). In addition, Zhou *et al.* (9) discovered that miR-130a acts as an oncogenic miRNA in gastric cancer.

In our study, we sought to develop and validate a miRNAs prognostic signature through data mining of the Cancer Genome Atlas (TCGA) database. The prognostic model can identify high-risk CRC patients with lower survival rates to allow intervention can be initiated earlier to improve quality of life, and find prognostic related target genes through the interaction study and functional analysis of target genes, so as to provide new ideas for targeted therapy.

We present the following article in accordance with the TRIPOD reporting checklist (available at <https://tcr.amegroups.com/article/view/10.21037/tcr-21-1992/rc>).

Methods

Acquisition and processing of the data sets

MiRNAs expression data along with clinical information for 521 colorectal adenocarcinoma patients were downloaded from the TCGA database at <http://cancergenome.nih.gov/> (up to March 2020). For miRNAs expression data, the miRNAs with minimal expression levels (<1) were first filtered out, and the data were then processed via a standard procedure using the “edgeR” package (version 3.6.2). Finally, 385 differentially expressed miRNAs were identified between cancer tissues and normal tissues in CRC patients, using $|\log_2FC|$ of >1 and an adjusted P value of <0.05 as cutoff criteria. The heatmap of the top significantly differentially expressed miRNAs was generated with the “ggplot2” package (version 3.6.3). Patients with incomplete clinical information (including sex, age, TNM stage, survival status and overall survival time) were excluded from further analysis. Overall survival time and survival state were predicted outcomes. After the above exclusion steps, 415 CRC patients were enrolled in and divided into training set (n=208) and test set (n=207) randomly in R-3.6.1. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Screening the prognosis-related miRNAs and modeling in the training set

Firstly, the regression analysis of univariate Cox proportional hazards was performed in the training set to analyze the correlation between the expression level of each miRNA and the patient's survival time and status via the R package “survival”. Secondly, least absolute shrinkage and selection operator (LASSO)-Cox regression was conducted on the miRNAs with a P value of <0.05 in the first analysis. LASSO is a compression estimate method that can produce a more refined model by constructing a penalty function. In this study, we used the R package “glmnet”. To determine the value of penalty parameter λ , cross-validation was carried out first, and the value of λ resulting in the minimum mean cross-validated error was selected (10). Thirdly, the selected miRNAs were included in the multivariate Cox regression analysis to calculate their respective coefficients. Lastly, the following microRNA

prognostic signature was constructed: microRNA signature = $\sum(\beta_i \times \text{Exp}_i)$, where i is the name of the prognostic miRNA and β is the coefficient of that miRNA in the third analysis. MiRNAs with a positive coefficient were considered as oncogenic miRNAs, while those with a negative coefficient were considered as protective miRNAs.

Efficacy verification of the miRNA signature in the three data sets

To further assess the efficacy of the microRNA signature, we calculated the score of each patient in the three data sets (the training set, the test set and the entire set) based on this microRNA signature. According to the median, each set was divided into a high-risk group and a low-risk group, and then the Kaplan-Meier survival analysis was applied to compare the overall survival difference between the high-risk group and the low-risk group using the R package “survival”. To verify the efficacy of the microRNA signature, time-dependent receiver operating characteristic (ROC) curves and the areas under the ROC curves (AUCs) were obtained via the R package “timeROC”.

Combination of clinical factors to evaluate the efficacy of the microRNA signature

In the entire data set, the microRNA signature combining clinical factors (including age, gender, race, tumor site and TNM stage) was analyzed by univariate Cox regression and multivariate Cox regression to identify associations between these miRNAs and overall patient survival. The variables with a P value of <0.05 were included in further horizontal and vertical comparisons. ROC curve analysis was performed with the R packages “plotROC” and “ggplot2” to horizontally compare the microRNA signature with clinical factors related to the prognosis of CRC. Kaplan-Meier survival curves were used for stratified longitudinal analysis.

MiRNAs target genes prediction and their interaction network

Target genes of the selected miRNAs were predicted via the following three miRNA databases: miRTarBase (<http://mirtarbase.mbc.nctu.edu.tw/>, version: 7.0), TargetScan (<http://www.targetscan.org/>, version: Human 7.2) and

miRDB (<http://mirdb.org/>). The intersection of the results obtained from the three databases was considered the set of miRNA target genes. MiRNA target genes interaction network was achieved using the STRING database (<https://string-db.org/>, version: 11.0). Cytoscape (version: 3.7.2) was used to screen out the Top10 target genes, and MCODE plug-in was applied to select the important gene modules.

Functional enrichment analysis and survival analysis of target genes

Functional enrichment analysis of these miRNA-related genes was download from the STRING database after interaction network analysis, including Gene Ontology-biological process (GO-BP) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. The results with a false discovery rate (FDR) of <0.05 were visualized using the R packages “Cairo” (version 3.6.3) and “ggplot2” (version 3.6.3). The online analysis site GEPIA2 (<http://gepia2.cancer-pku.cn/>) was used to perform ROAD and READ prognostic analysis of overall survival for Top10 target genes in the TCGA database.

Statistical analysis

Univariate Cox, LASSO-COX and multivariable Cox were used to select the prognostic miRNAs in R-3.6.1. LASSO-COX was conducted by R package “glmnet”. The penalty parameter λ was determined by cross-validation, and the value of λ resulting in the minimum mean cross-validated error was selected. Survivals were evaluated with the Kaplan-Meier method and log-rank test. $P < 0.05$ was considered statistically significant.

Results

Patient data

MiRNA expression files and clinical information for 521 CRC patients (comprising 529 tumor samples and 11 normal tissues) were downloaded from the TCGA database. A total of 415 CRC patients with complete clinical data were enrolled in further analysis. All enrolled patients had primary adenocarcinoma, did not have a previous or concurrent malignancy, and received no chemotherapy or radiotherapy before surgery. After differential miRNA

expression analysis, the 415 CRC patients were randomly divided into two sets (Table 1). The detailed research design could be seen in Figure 1A.

Differentially expressed miRNAs between cancer tissues and normal tissues in CRC patients

Before analysis, data normalization and log2 transformation were conducted in the miRNA expression files firstly using R software (version 3.6.1). We finally identified 385 differentially expressed (231 upregulated and 154 downregulated) miRNAs ($|\log_2\text{FC}| > 1$ and the adjusted P value < 0.05) using the “DESeq” package (Figure S1). The heatmap of top25 differentially expressed miRNAs could clearly distinguish the CRC samples from the controlling normal samples (Figure S2).

Signature identification of the prognosis-related miRNA in the training set

Univariate Cox regression analysis with the LASSO method was used in the training set for selection of CRC prognosis-related miRNAs among the 385 differentially expressed miRNAs. Eleven miRNAs with the best fit were screened by LASSO-Cox regression method (Figure 1B,1C). Then, through a stepwise multivariate Cox regression analysis to maximize accuracy and validity, seven miRNAs were finally screened, and the prognosis-related miRNA signature was constructed based on the expression levels of these seven miRNAs. The signature formula was as follows: 7-miRNA signature = $\text{hsa-miR-144} \times -0.255 + \text{hsa-miR-153-2} \times 0.472 + \text{hsa-miR-505} \times 0.475 + \text{hsa-miR-887} \times 0.565 + \text{hsa-miR-3199-2} \times 0.509 + \text{hsa-miR-561} \times 0.452 + \text{hsa-miR-3684} \times 0.516$. The higher concordance index (C index), the better model performance, and the C index of the 7-miRNA signature was 0.772.

Among these miRNAs, hsa-miR-144 had a negative coefficient in the risk formula, indicating that it might be considered a protective miRNA; that is, the higher the expression level of this miRNA, the longer was the survival time of the patient. The remaining six miRNAs in this risk formula had positive coefficients, pointing that higher expression levels of these miRNAs were associated with shorter patient survival times.

Evaluation of the survival prediction capability of the 7-miRNA signature in the training set

The 208 samples in the training set were scored according to the risk scoring formula and divided into the high-risk group (n=104) and the low-risk group (n=104), with a median value of 1.009 as the cutoff point. The Kaplan-Meier overall survival curves of the two groups showed statistically significant differences between patient survival (log-rank test $P=5e-08$) (Figure 2A). ROC curves along with AUC values were further generated to evaluate the prognostic ability of the 7-miRNA signature (Figure 2B). The AUCs for 3- and 5-year survival in the training set were 0.769 and 0.889, respectively, indicating that the signature had good performance in predicting CRC patients' survival risk. In addition, the distributions of the risk score, survival status and expression profile of the seven miRNAs in the training set were illustrated to visualize the analysis results (Figure 2C). Patients with higher risk scores had poorer overall survival than patients with lower risk scores, as shown in Figure 2C. The expression levels of the protective miRNA hsa-miR-144 were significantly decreased in patients with high-risk scores, whereas those of the remaining six oncogenic miRNAs (miR-153-2, miR-505, miR-887, miR-3199-2, miR-561, and miR-3684) were increased.

Performance verification of the 7-miRNA signature in the test set and entire set

To evaluate the performance of the 7-miRNA signature for predicting the survival of CRC patients, Kaplan-Meier survival analysis was conducted and ROC curves with AUC values were generated in the test set (n=207) and the entire set (n=415). Patients in the two validation sets were classified into a high-risk group and a low-risk group according to the constructed 7-miRNA signature. Similar results were obtained in Kaplan-Meier survival analysis of the two set, in which the patients in the high-risk group had significantly shorter overall survival times than those in the low-risk group (Figures S3,S4). These results confirmed the power of the 7-miRNA signature for predicting the prognosis of CRC patients. Besides, the overall survival times of patients in the high-risk group was significantly shorter than those of patients in the low-risk

Table 1 Clinical data for the CRC patients in each data set

Clinical information	Training set, n (%)	Testing set, n (%)	Entire set, n (%)	P value
N	208	207	415	
Age (years)				0.905
<65	97 (46.6)	92 (44.4)	189 (45.5)	
≥65	111 (53.4)	115 (55.6)	226 (54.5)	
Gender				0.360
Female	105 (50.5)	90 (43.5)	195 (47.0)	
Male	103 (49.5)	117 (56.5)	220 (53.0)	
Race				0.771
Asian	7 (3.4)	2 (1.0)	9 (2.2)	
Black	108 (51.9)	105 (50.7)	213 (51.3)	
White	66 (31.7)	74 (35.7)	140 (33.7)	
Not reported	27 (13.0)	26 (12.6)	53 (12.8)	
Tumor site				0.814
Rectum	27 (13.0)	36 (17.4)	63 (15.2)	
Rectosigmoid junction	28 (13.5)	26 (12.6)	54 (13.0)	
Colon	153 (73.6)	145 (70.0)	298 (71.8)	
AJCC-T				0.898
Tis	1 (0.5)	0 (0.0)	1 (0.2)	
T1	7 (3.4)	7 (3.4)	14 (3.4)	
T2	36 (17.3)	38 (18.4)	74 (17.8)	
T3	141 (67.8)	140 (67.6)	281 (67.7)	
T4	23 (11.1)	22 (10.6)	45 (10.8)	
AJCC-N				0.532
N0	118 (56.7)	118 (57.0)	236 (56.9)	
N1	55 (26.4)	47 (22.7)	102 (24.6)	
N2	35 (16.8)	42 (20.3)	77 (18.5)	
AJCC-M				0.839
M0	155(74.6)	158 (76.3)	314 (75.7)	
M1	32(15.4)	33 (15.9)	64 (15.4)	
MX	19 (9.1)	15 (7.2)	34 (8.2)	
Unknown	2 (1.0)	1 (0.5)	3 (0.7)	
Pathological stage				0.999
I	39 (18.8)	40 (19.3)	79 (19.0)	
II	74 (35.6)	76 (36.7)	150 (36.1)	
III	63 (30.3)	57 (27.5)	120 (28.9)	
IV	32 (15.4)	34 (16.4)	66 (15.9)	
Survival status				0.841
Alive	164 (78.8)	168 (81.2)	332 (80.0)	
Dead	44 (21.2)	39 (18.8)	83 (20.0)	

CRC, colorectal cancer; AJCC, American Joint Committee on Cancer.

Figure 1 Research flowchart and construction of the seven-miRNA signature. (A) The research flowchart; (B) LASSO coefficient profiles of the miRNAs associated with the prognosis of patients with CRC in the training set; (C) the λ value was determined by lambda.min. CRA, colorectal adenocarcinoma; miRNA, microRNA; TCGA, the Cancer Genome Atlas; DEG, differentially expressed gene; LASSO, least absolute shrinkage and selection operator; ROC, receiver operating characteristic; CRC, colorectal cancer.

group; moreover, one protective miRNA, hsa-mir-144, was highly expressed in the low-risk group, while the remaining six oncogenic miRNAs were highly expressed in the high-risk group, reconfirming the results in the training set.

Combination of routine clinical factors for efficacy evaluation of the 7-miRNA signature

Clinical factors play an important role in tumor prognosis. To construct a more stable and effective prognostic model, the 7-miRNA signature combined with conventional clinical parameters (including age, race, sex, tumor site and TNM stage) was analyzed by univariate and multivariate Cox regression along with stratification analysis to evaluate

the associations of these parameters with the overall survival in the entire data set. Variables with a P value of <0.05, the 7-miRNA signature, age and TNM stage, were considered to be associated with prognosis, as shown in Table 2. In addition, the conclusion can be made that the 7-miRNA signature is an independent risk factor for the prognosis of CRC patients.

Then, ROC curves and AUC values were used to horizontally compare the efficacy of the 7-miRNA signature with those of patient age and TNM stage in predicting the prognosis of patients with CRC. First, the 7-miRNA signature, age, and a new variable combining both of these variables were included in the analysis. As shown in Figure 3A, the 7-miRNA signature showed a

Figure 2 Capability evaluation and risk score analysis in the training set. (A) The time-dependent ROC curves for 3-year survival and 5-year survival; (B) Kaplan-Meier curves showing that the overall survival rate in the low-risk group was significantly higher than that in the high-risk group; (C) distribution of risk scores, patient's survival time and status and the heat map of the miRNAs in the prognostic signature. The red dotted line indicates the optimum cutoff value for dividing patients into the low-risk and high-risk groups. ROC, receiver operating characteristic; AUC, areas under the ROC curve; miRNA, microRNA.

higher predictive power than the age. Then, the 7-microRNA signature, TNM stage, and a new variable combining both of these variables were included; similarly, the predictive power of the 7-miRNA signature was superior to that of the TNM stage (Figure 3B).

Finally, analysis was performed on the 415 patients stratified by age and TNM stage to compare the risk score of the 7-microRNA signature. The patients were divided by the average age in years into the younger group (<65, n=189) and the aged group (≥ 65 , n=226). In the younger group, the 7-microRNA signature divided the patients into a high-risk group (n=95) and a low-risk group (n=94)

according to the median risk score. Kaplan-Meier survival curves showed that the difference between the two risk groups was statistically significant ($P=0.00323695$, Figure 3C). The corresponding values were also statistically significant ($P=1e-07$, Figure 3D) in the aged group, which was also divided into the high- and low-risk groups (n=112 and 114, respectively). Then, the patients were classified by TNM stage into the early-stage group (stages I + II, n=229) and the advanced-stage group (stages III + IV, n=186). In both stage groups, the 7-microRNA signature separated the patients into a high-risk group and a low-risk group independently. Similarly, the Kaplan-Meier survival

Table 2 Univariate and multivariate Cox regression analysis results for the survival risk of CRC patients across the entire set

Variables	Univariate Cox			Multivariate Cox		
	HR	95% CI of HR	P value	HR	95% CI of HR	P value
Gender (female vs. male)	0.995	0.646–1.533	0.983	−0.279	0.608–1.454	0.780
Race (Asian vs. Black vs. White reported)	1.003	0.732–1.375	0.983	−0.874	0.602–1.215	0.382
Age (<65 vs. ≥65 years)	2.101	1.338–3.299	0.001*	4.017	1.634–4.164	5.89E−05*
Tumor site (rectum vs. colon)	1.054	0.775–1.435	0.736	0.580	0.799–1.514	0.562
TNM stage (I + II vs. III + IV)	3.296	2.060–5.273	6.56E−07*	4.751	1.983–5.185	2.02E−06*
7-miRNA signature (low vs. high)	4.365	2.557–7.453	6.66E−08*	4.838	2.218–6.565	1.31E−06*

*, $P < 0.05$. CRC, colorectal cancer; TNM, tumor-node-metastasis; HR, hazard ratio.

Figure 3 Comparison of the prognostic value with clinical parameters and stratification analysis by age and TNM visualized by Kaplan-Meier overall survival curves in the entire set. (A) The time-dependent ROC curves for follow-up were plotted to assess the prognostic efficacy of age, miRNA signature, and a new variable combining both of these variables; (B) time-dependent ROC curves for follow-up were plotted to assess the prognostic efficacy of the TNM stage, miRNA signature, and a new variable combining both of these variables; (C) the younger group; (D) the aged group; (E) the early-stage group (TNM I & II); (F) the advanced-stage group (TNM III & IV). miRNA, microRNA; AUC, areas under the ROC curve; TNM, tumor-node-metastasis; ROC, receiver operating characteristic.

curves for both stage groups indicated that the overall survival times of patients in the high-risk groups were significantly shorter than those of patients in the low-risk groups ($P = 5.891 \times 10^{-5}$ & 0.00011855 , Figure 3E,3F).

Target gene prediction of the seven prognostic miRNAs and their interaction network

To investigate the cellular processes and biological pathways

Figure 4 Prediction of target genes of the seven miRNAs and the top 10 hub genes and two major gene modules of the interaction network. (A) The predicted target genes of the miRNAs in the 7-miRNA signature were defined as the intersection of the genes predicted by miRTarBase, TargetScan and miRDB; (B) the top 10 hub genes screened out from the interaction network; (C) two major gene modules were selected using MCODE plug-in. miRNA, microRNA.

associated with the seven prognostic miRNAs, miRNA target prediction was separately performed with miRDB, miRTarBase7.0 and TargetScan. MiRDB identified 6,267 miRNA-related genes (11), while TargetScan (12), the most comprehensive miRNA target gene database, identified 14,034 mRNAs. MiRTarBase (13), the experimentally validated miRNA-target interaction database, identified 149 mRNAs associated with the seven prognostic miRNAs. The intersection of the results from these three databases was identified as the set of miRNA targeted genes ($n=67$, Figure 4A) for further analyses. The interaction network

of the sixty-seven target genes were constructed by the STRING database (14) (Figure S5). The interactions between target genes were download, which was input to software Cytoscape (version: 3.7.2). In addition, the top10 hub genes including *PTEN*, *CDKN1A*, *NOTCH1*, *MTOR*, *EIF4E*, *HNRNPC*, *FOXO1*, *MCL1*, *SNAI1*, *DDX3X* were screened out (Figure 4B), and two major gene modules were selected using MCODE plug-in (Figure 4C). Except *DDX3X*, other nine top10 target genes were all in the first module, which had a score of 6.154. The second module, consisting of *DDX3X*, *DDX6*, and *YTHDC2*, has a score of 3.

Functional enrichment and survival analysis of the target genes

For functional enrichment, GO BP and KEGG pathway enrichment analyses were download from STRING database. Then, the terms with FDR <0.05 were visualized using the R packages “Cairo” and “ggplot2”. The GO analysis highlighted a 30- to 60-fold enrichment mainly in cell biological processes, metabolic processes, development and differentiation, as well as responses to stimulus (*Figure 5A*). The GO analysis results indicated the potential roles of the seven prognostic miRNAs in the regulation of CRC tumorigenesis and progression via targeting of their associated mRNAs. The KEGG analysis in biological pathways indicated that EGFR tyrosine kinase inhibitor resistance was a major factor affecting the prognosis of CRC patients, and the target genes involved mainly included *MTOR*, *PTEN*, *BCL2*, and *PRKCA* (*Figure 5B*, *Table 3*). We also found that MTOR signaling pathway was a key pathway in the downstream pathways of EGFR signaling pathway, including PI3K-Akt, HIF-1, JAK-STAT and ErbB signaling pathway. CDKN1A, which related to CRC, participated in the regulation of almost all the downstream EGFR signaling pathways with p53 signaling pathway included. Kaplan-Meier survival curves of the top10 target genes were plotted using the online analysis site GEPIA2 based on TCGA database, including COAD and READ (15). Finally, the results showed that CDKN1A, EIF4E and SNAI1 were associated with prognosis in patients with colorectal adenocarcinoma (*Figure 6A-6C*).

Discussion

CRC is the world's third most common cancer (16), accounting for approximately 10% of all new cancer cases annually worldwide (17). Adenocarcinoma is the most common histological type of CRC. Although CRC screening for earlier detection has somewhat reduced morbidity and mortality in recent years, the 5-year survival rate of CRC patients remains unsatisfactory (18), and the molecular mechanisms underlying CRC progression are still elusive. CRC is a highly heterogeneous tumor, and developing novel predictive biomarkers is important and needed for obtaining a better understanding of CRC and providing diagnostic and therapeutic strategies. In this study, we conducted a series of analyses on miRNA expression and corresponding clinical data for training and validation set

generated from 415 colorectal adenocarcinoma patients in the TCGA database, with an aim to develop a meaningful predictive model related to CRC prognosis.

MiRNAs, an important subset of noncoding RNAs, have been recognized to play multifaceted roles in controlling cellular functions by repressing their target genes (19). For example, let-7 family members can regulate cancer stem cells by targeting HRAS and HMGA2 (20), and miR-21, which is highly expressed in breast cancer, is correlated with poor patient survival (21). Virtually all CRCs exhibit altered miRNA expression (22); for instance, miR-22 contributes to tumorigenesis in CRC (23), and downregulation of miR-34 results in CRC metastasis via an increase in IL6 signaling pathway activity (24). These findings suggest the potential of miRNAs as excellent prognostic biomarkers for CRC survival prediction. The heat map generated in our study intuitively indicates that compared with the normal tissues, the CRC tissues exhibited obvious miRNA expression abnormalities, proving the previous statement by Okugawa that almost all CRCs exhibit altered miRNA expression.

Through a comprehensive analysis including Cox regression with the LASSO method on the training set, seven prognostic miRNAs were finally screened. Model evaluation is dependent on the values of the C index and AUC. The higher the C index and AUC values, the better is the model (25). Our risk score formula, termed the 7-miRNA signature, was established with a C index of 0.772. The AUC for 5-year survival was 0.889 in the training set and 0.816 in the entire TCGA set, indicating that the signature had good predictive power. Several types of CRC prognostic score models have recently been developed and verified, including an immunoscore with a C index of 0.612 and AUC of 0.630 for 5-year survival (26), a lncRNA score with an AUC of 0.733 for 5-year survival (27), and so on. However, prognostic models comprising miRNAs for use with CRC patients have not been reported, and our signature filled this gap. Moreover, the performance of the 7-miRNA signatures was evaluated in the training set and verified in the test set and the entire set using ROC curve and Kaplan-Meier survival analysis. The 7-miRNA signature performed well in distinguishing whether patients with CRC had a high or low prognostic risk.

The heat map for the three data set clearly shows the expression patterns of the seven miRNAs in CRC samples. The expression of miR-144 was high in the low-risk group but low in the high-risk group, and the remaining 6 miRNAs exhibited the opposite pattern. These results indicate that miR-144 is the only negatively regulated

Figure 5 Results of functional enrichment analysis for the target genes. (A) Biological processes in GO-BP; (B) KEGG. FDR, false discovery rate; EGFR, epidermal growth factor receptor; GO-BP, Gene Ontology-biological process; KEGG, Kyoto Encyclopedia of Genes and Genomes.

Table 3 Main KEGG terms enriched with the seven miRNAs in the signature

Term	Strength	FDR	Matching target genes in the network
miRNAs in cancer	1.2	8.93E-06	<i>NOTCH1, MTOR, MCL1, PTEN, BCL2, CDKN1A, PRKCA, ZEB2</i>
EGFR tyrosine kinase inhibitor resistance	1.18	0.0053	<i>MTOR, PTEN, BCL2, PRKCA</i>
Cellular senescence	0.97	0.0054	<i>MTOR, PTEN, FOXO1, PPP1CB, CDKN1A</i>
Endocrine resistance	1.09	0.0067	<i>NOTCH1, MTOR, BCL2, CDKN1A</i>
HIF-1 signaling pathway	1.08	0.0067	<i>MTOR, BCL2, CDKN1A, PRKCA</i>
Pathways in cancer	0.66	0.0067	<i>NOTCH1, MTOR, PTEN, FOXO1, NFE2L2, BCL2, CDKN1A, PRKCA</i>
Insulin resistance	1.04	0.0075	<i>MTOR, PTEN, FOXO1, PPP1CB</i>
Thyroid hormone signaling pathway	1.01	0.0089	<i>NOTCH1, MTOR, FOXO1, PRKCA</i>
Autophagy - animal	0.97	0.0097	<i>SNAP29, MTOR, PTEN, BCL2</i>
Platelet activation	0.98	0.0097	<i>PRKCI, FGA, FG, PPP1CB</i>
PI3K-Akt signaling pathway	0.7	0.0116	<i>MTOR, MCL1, PTEN, BCL2, CDKN1A, PRKCA</i>
Insulin signaling pathway	0.94	0.0116	<i>PRKCI, MTOR, FOXO1, PPP1CB</i>
p53 signaling pathway	1.11	0.0144	<i>PPM1D, PTEN, CDKN1A</i>
Jak-STAT signaling pathway	0.86	0.0174	<i>MTOR, MCL1, BCL2, CDKN1A</i>
ErbB signaling pathway	1.02	0.0221	<i>MTOR, CDKN1A, PRKCA</i>
CRC	1.01	0.0225	<i>MTOR, BCL2, CDKN1A</i>
Focal adhesion	0.77	0.029	<i>PTEN, PPP1CB, BCL2, PRKCA</i>
Proteoglycans in cancer	0.78	0.029	<i>MTOR, PPP1CB, CDKN1A, PRKCA</i>
Sphingolipid signaling pathway	0.88	0.0424	<i>PTEN, BCL2, PRKCA</i>

KEGG, Kyoto Encyclopedia of Genes and Genomes; miRNA, microRNA; FDR, false discovery rate; EGFR, epidermal growth factor receptor; CRC, colorectal cancer.

Figure 6 The target genes related to the prognosis of CRC. (A) Kaplan-Meier survival curves of CDKN1A; (B) Kaplan-Meier survival curves of EIF4E; (C) Kaplan-Meier survival curves of SNAI1. HR, hazard ratio; CRC, colorectal cancer.

miRNA with an antitumor effect; thus, miR-144 may be a protective factor for CRC and may become a new therapeutic target. This possibility needs further study.

Furthermore, among the factors related to the prognosis of CRC patients, clinical parameters, such as TNM stage (28), which is currently a prognostic indicator favored by clinicians, are essential. Cox regression analysis revealed three prognostic factors—age, TNM stage and the 7-miRNA signature, among which the 7-miRNA signature was an independent prognostic risk factor. To gain insight into the prognosis of patients with CRC, we performed a transverse ROC analysis and a stratified longitudinal analysis on the entire TCGA set. The AUC of the 7-miRNA signature was superior to that of TNM stage and age for predicting prognosis. Then, the stratification analysis results suggested that the 7-miRNA signature could accurately distinguish high-risk patients from low-risk patients stratified by TNM stage and age, which again demonstrated the predictive power of the 7-miRNA signature.

In addition, from the interaction network of the target genes connected with the 7-miRNA signature and major gene modules, we could see that PTEN ranked the first and CDKN1A ranks the second among the top10 key genes, they played a critical role in the network. PTEN is a tumor suppressor gene closely associated with tumorigenesis, an important negative regulator of PI3K-Akt signaling pathway, and remains one of the main challenges for CRC treatment (29). Functional enrichment analysis further revealed that EGFR tyrosine kinase inhibitor resistance was the leading reason affecting the prognosis of patients with CRC, and related pathways were mainly PI3K-Akt, p53 and JAK-STAT signal pathway, involving the target genes of *MTOR*, *MCL1*, *PTEN*, *BCL2*, *CDKN1A*, *PRKCA*, among which *MTOR*, *BCL2*, *CDKN1A* were associated with CRC. This states that PI3K-Akt-mTOR signaling pathway is a major factor affecting the prognosis of CRC patients. MTOR, a target protein of the PI3K-Akt signaling pathway, plays an important regulatory role in cell metabolism, development and survival in response to stimulus, which making it a crucial and validated target in the treatment of cancer (30). Kaplan-Meier survival curves of the top10 target genes showed that CDKN1A, EIF4E and SNAI1 were associated with prognosis in patients with CRC from TCGA database. CDKN1A, eIF4E and SNAI1, as downstream regulators of the PI3K-Akt-mTOR signaling pathway, play a vital role in drug resistance and prognosis of EGFR inhibitors for CRC.

To the best of our knowledge, this study is the first to report the relationship between a 7-miRNA signature and survival in CRC. However, this study has some limitations. First, although the biological functions of the prognostic miRNAs were annotated through functional enrichment analysis, *in vivo* and *in vitro* studies are still needed to further reveal the mechanism by which these prognostic miRNAs and related target genes including PTEN, MTOR, CDKN1A, eIF4E and SNAI1 mediate CRC tumor development and prognosis. Second, in the sample data obtained from the TCGA database, white ethnic groups were the most common, and data for Asian ethnic groups were scarce. Although our study showed that the prognosis of CRC patients was not correlated with the ethnic group, whether differences exist within ethnic groups is unknown. More data from Asian ethnic groups are needed for further investigation.

In conclusion, we performed scientific data mining for miRNAs from a TCGA dataset of patients with CRC and identified a 7-miRNA signature with good predictive performance and superior to other conventional clinical variables. Functional enrichment analysis results suggested that the miRNAs in the 7-miRNA signature may be related to EGFR tyrosine kinase inhibitor resistance for affecting the prognosis of patients with CRC, and PI3K-Akt-MTOR was the mainly related pathway. PTEN as its upstream molecule, CDKN1A, eIF4E and SNAI1 as its downstream molecules, respectively, play a significant regulatory role, and are potential therapeutic targets.

Acknowledgments

Funding: This work was supported by the National Natural Science Foundation of China (81702324, 81602529); Natural Science Foundation of Hebei Province (H2018206176, H2017206141); and the Medical Science Research Project of Hebei Province (20190510).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at <https://tcr.amegroups.com/article/view/10.21037/tcr-21-1992/rc>

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at <https://tcr.amegroups.com/article/view/10.21037/tcr-21-1992/coif>). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- Thrumurthy SG, Thrumurthy SS, Gilbert CE, et al. Colorectal adenocarcinoma: risks, prevention and diagnosis. *BMJ* 2016;354:i3590.
- Okugawa Y, Toiyama Y, Toden S, et al. Clinical significance of SNORA42 as an oncogene and a prognostic biomarker in colorectal cancer. *Gut* 2017;66:107-17.
- Hayama T, Hashiguchi Y, Okamoto K, et al. G12V and G12C mutations in the gene KRAS are associated with a poorer prognosis in primary colorectal cancer. *Int J Colorectal Dis* 2019;34:1491-6.
- Zheng H, Song K, Fu Y, et al. A qualitative transcriptional signature for determining the grade of colorectal adenocarcinoma. *Cancer Gene Ther* 2020;27:680-90.
- Kamal Y, Schmit SL, Hoehn HJ, et al. Transcriptomic differences between primary colorectal adenocarcinomas and distant metastases reveal metastatic colorectal cancer subtypes. *Cancer Res* 2019;79:4227-41.
- He Z, Dang J, Song A, et al. Identification of LINC01234 and MIR210HG as novel prognostic signature for colorectal adenocarcinoma. *J Cell Physiol* 2019;234:6769-77.
- Mollaei H, Safaralizadeh R, Rostami Z. MicroRNA replacement therapy in cancer. *J Cell Physiol* 2019;234:12369-84.
- Mirzaei H, Fathollahzadeh S, Khanmohammadi R, et al. State of the art in microRNA as diagnostic and therapeutic biomarkers in chronic lymphocytic leukemia. *J Cell Physiol* 2018;233:888-900.
- Zhou Y, Li R, Yu H, et al. microRNA-130a is an oncomir suppressing the expression of CRMP4 in gastric cancer. *Onco Targets Ther* 2017;10:3893-905.
- Birnbaum DJ, Finetti P, Lopresti A, et al. A 25-gene classifier predicts overall survival in resectable pancreatic cancer. *BMC Med* 2017;15:170.
- Liu W, Wang X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biol* 2019;20:18.
- Agarwal V, Bell GW, Nam JW, et al. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 2015.
- Chou CH, Shrestha S, Yang CD, et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res* 2018;46:D296-302.
- Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;47:D607-13.
- Tang Z, Kang B, Li C, et al. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res* 2019;47:W556-60.
- Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat Rev Gastroenterol Hepatol* 2019;16:713-32.
- Wong SH, Yu J. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nat Rev Gastroenterol Hepatol* 2019;16:690-704.
- Katona BW, Weiss JM. Chemoprevention of colorectal cancer. *Gastroenterology* 2020;158:368-88.
- Rupaimoole R, Calin GA, Lopez-Berestein G, et al. miRNA deregulation in cancer cells and the tumor microenvironment. *Cancer Discov* 2016;6:235-46.
- Sung SY, Liao CH, Wu HP, et al. Loss of let-7 microRNA upregulates IL-6 in bone marrow-derived mesenchymal stem cells triggering a reactive stromal response to prostate cancer. *PLoS One* 2013;8:e71637.
- Frankel LB, Christoffersen NR, Jacobsen A, et al. Programmed cell death 4 (PDCD4) is an important functional target of the microRNA miR-21 in breast cancer cells. *J Biol Chem* 2008;283:1026-33.
- Okugawa Y, Grady WM, Goel A. Epigenetic alterations in colorectal cancer: emerging biomarkers. *Gastroenterology* 2015;149:1204-1225.e12.
- Liu Y, Chen X, Cheng R, et al. The Jun/miR-22/HuR regulatory axis contributes to tumorigenesis in colorectal cancer. *Mol Cancer* 2018;17:11.
- Rokavec M, Öner MG, Li H, et al. IL-6R/STAT3/

- miR-34a feedback loop promotes EMT-mediated colorectal cancer invasion and metastasis. *J Clin Invest* 2014;124:1853-67.
25. Li B, Cui Y, Diehn M, et al. Development and validation of an individualized immune prognostic signature in early-stage nonsquamous non-small cell lung cancer. *JAMA Oncol* 2017;3:1529-37.
 26. Zhao X, Liu J, Liu S, et al. Construction and validation of an immune-related prognostic model based on TP53 status in colorectal cancer. *Cancers (Basel)* 2019;11:1722.
 27. Fan Q, Liu B. Discovery of a novel six-long non-coding RNA signature predicting survival of colorectal cancer patients. *J Cell Biochem* 2018;119:3574-85.
 28. Lea D, Håland S, Hagland HR, et al. Accuracy of TNM staging in colorectal cancer: a review of current culprits, the modern role of morphology and stepping-stones for improvements in the molecular era. *Scand J Gastroenterol* 2014;49:1153-63.
 29. De Mattia E, Cecchin E, Toffoli G. Pharmacogenomics of intrinsic and acquired pharmacoresistance in colorectal cancer: toward targeted personalized therapy. *Drug Resist Updat* 2015;20:39-70.
 30. Alqurashi N, Hashimi SM, Alowaidi F, et al. Dual mTOR/PI3K inhibitor NVP BEZ235 arrests colorectal cancer cell growth and displays differential inhibition of 4E BP1. *Oncol Rep* 2018;40:1083-92.

Cite this article as: Jiang S, Xie X, Jiang H. Establishment of a 7-microRNA prognostic signature and identification of hub target genes in colorectal carcinoma. *Transl Cancer Res* 2022;11(2):367-381. doi: 10.21037/tcr-21-1992

Figure S1 Volcano plot generated by the “DESeq” package for filtrating the miRNA expression data. FDR, false discovery rate; miRNA, microRNA.

Figure S2 Heatmap generated by the top25 differentially expressed miRNAs between cancer tissues and normal tissues in CRA patients. CRA, colorectal adenocarcinoma; miRNA, microRNA.

Figure S3 Performance verification in the test set. (A) ROC curve analysis of the seven-miRNA signature. (B) Kaplan-Meier overall survival analysis of the CRA patients using the seven-miRNA signature. (C) Distributions of the risk scores and survival statuses of the seven miRNAs in the CRA patients. ROC, receiver operating characteristic; AUC, areas under the ROC curve; miRNA, microRNA; CRA, colorectal adenocarcinoma.

Figure S4 Performance verification in the entire set. (A) Kaplan-Meier overall survival analysis of CRA patients using the seven-miRNA signature. (B) ROC curve analysis of the seven-miRNA signature. (C) Distributions of the risk scores, survival statuses and expression profiles of the seven miRNAs in the CRA patients. ROC, receiver operating characteristic; AUC, areas under the ROC curve; miRNA, microRNA; CRA, colorectal adenocarcinoma.

Figure S5 Interaction network of the target genes from the STRING dataset.