Original Paper

Genome-Wide Network-Based Analysis of Colorectal Cancer Identifies Novel Prognostic Factors and an Integrative Prognostic Index

Xiaolin Hou^a Xuelai He^b Kang Wang^b Nengyi Hou^b Junwen Fu^b
Guiqing Jia^b Xiaofei Zuo^b Haibo Xiong^b Minghui Pang^b

^aDepartment of Internal Medicine, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, ^bDepartment of Gastrointestinal Surgery, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, China

Key Words

Colorectal cancer • Prognosis • Prognostic factor • Genome-wide • Network analysis

Abstract

Background/Aims: Colorectal cancer (CRC) is one of leading cancers in both incidence and mortality rate. The 5-year survival rate varies considerably depending on the pathological stage of the tumor. Although prominent progress has been made through screening for survival-associated factors from a certain type of genetic or epigenetic modifications, few attempts have been made to apply a network-based approach in prognostic factor identification, which could prove valuable for a complex, multi-faceted disease such as CRC.

Methods: In this study, a TCGA dataset of 379 CRC patients was subjected to a network-based analysis strategy consisting of multivariate regression, co-expression network and gene regulatory network analyses, and survival analyses. Both genetic and epigenetic aberrations, including those in gene expression and DNA methylation at specific sites, were screened for significant association with patient survival. A prognostic index (PI) integrating all potential prognostic factors was subsequently validated for its prognostic value. **Results:** A collection of six miRNAs, eleven mRNAs, and nine DNA methylation sites were identified as potential prognostic factors. The low- and high-risk patient groups assigned based on PI level showed significant difference in overall survival (hazard ratio = 1.32, 95% confidence interval 1.29-1.36, $p < 0.0001$). Patients in the low- and high-risk groups can be further divided into a total of four subgroups, based on pathological staging. In the two high-risk subgroups ($PI > 0$), there was significant difference (Cox $p < 0.0001$) in OS between the earlier (stages I/II) and later stages (stages III/IV). However, in the two low-risk subgroups ($PI < 0$), earlier (stages I/II) and later stages (stages III/IV) showed no significant difference in OS (Cox $p = 0.185$). On the

X. Hou, X. He, K. Wang contributed equally to this work.

Minghui Pang

Department of Gastrointestinal Surgery, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731 (China)
Tel. +86-13402821952, E-Mail pang461886583@126.com

other hand, there were significant differences in OS between the low- and high-risk subgroups when both subgroups were of earlier stages (Cox $p < 0.001$) or of later stages (Cox $p < 0.0001$).

Conclusion: The novel network-based, integrative analysis adopted in this study was efficient in screening for prognostic predictors. Along with PI, the set of 6 miRNAs, 11 mRNAs, and 9 DNA methylation sites could serve as the basis for improved prognosis estimation for CRC patients in future clinical practice.

© 2018 The Author(s)
Published by S. Karger AG, Basel

Introduction

Colorectal cancer (CRC) is one of the most frequent cancers worldwide. A global incidence of more than one million new cases is reported worldwide each year [1]. CRC ranks as the third most common cancer in incidence and cancer-related death in the United States. In Europe, this disease is the third and second most commonly diagnosed cancer in men and women, respectively, and ranks second in both sexes in cancer-related mortality [2]. Therefore, more efficient preventative and therapeutic interventions of this disease are pivotal to the battle against this major public health threat.

Advances in early diagnosis and treatment modalities have reduced mortality rate over the past two decades [3]. However, depending on the tumor stage, patient survival varies considerably among CRC patients, thereby posing challenges to targeted treatment and prognosis. According to the SEER database (1973-2014, 2017 release), the 5-year relative survival rates are approximately 90% for CRC patients with tumors of stages I, IIA, and IIIA, but decreases to approximately 65% for stage IIB and IIIB patients, and further to 53% for stage IIIC. Furthermore, for metastatic, stage IV patients, the 5-year survival rate shows a drastic fall to 11% [4].

Most CRC cases (approx. 70%) occur sporadically as a result of genetic and epigenetic modifications of a number of genes, including prominent oncogenes and tumor suppressors such as *APC*, *TP53*, and *KRAS*, as well as genes involved in DNA mismatch repair [5-9]. Moreover, a panel of genetic (e.g. gene expression) and epigenetic (e.g. DNA methylation) aberrations have been associated with patient survival [5, 10-12]. For instance, overexpression of immune checkpoint regulator PD-L2 was recently reported to be associated with poor survival [13]. Additionally, a study of primary tumors found expression of several epithelial (claudin-1 and E-cadherin) and mesenchymal markers (snail-1 and vimentin) to be independent predictors of recurrence of stage II CRC [14].

In addition to protein-coding genes (PCGs), the significance of microRNAs (miRNAs) in CRC has been increasingly highlighted [15, 16]. miRNAs are a class of small, non-coding RNA molecules that exert their functions through binding to the 3' untranslated region of target mRNAs. This in turn results in degradation of or translation repression of the bound mRNAs and therefore downregulation of the corresponding genes [17]. Using transcriptome profiling techniques, a flurry of studies has proposed miRNAs as prognostic indicators for CRC patients. For instance, serum miR-200c level was found to be a better predictor of reoccurrence than pathological staging for stage II and III patients [18]. Hansen et al. [19] also reported significant association between miR-126 upregulation and a longer progression-free survival.

Besides genetic features, epigenetic modifications such as DNA methylation are drawing increasing interest in CRC research. Observed predominantly at CpG dinucleotides in repetitive sequences, DNA methylation usually occurs early in carcinogenesis [20]. Through characterizing methylomes, studies have also proposed a number of methylation sites as potential prognosis factors. Methylation of genes *p14ARF*, *RASSF1A* and *APC1A* were suggested as unfavorable prognostic factors independent of tumor stage and differentiation [21]. Hypermethylation of *p16* was also proposed as a predictive factor for poor prognosis [22]. Inversely, hypomethylation of genes such as *insulin growth factor II (IGF2)* has also been associated with poor prognosis [23].

Together, these studies have provided multiple lines of evidence to pursue to establish

a robust panel of predictive factors for CRC prognosis. However, a distinct limitation is that most investigations were conducted without taking into consideration the network formed by aberrations at different levels of regulation. For colorectal cancer, a multi-faceted disease during which dysregulations abound at the genome, transcriptome, and proteome levels, a network-based, systems biology approach can lend more comprehensive insights into diagnosis estimation.

In this study, we applied a battery of network-based analyses to identifying potential prognostic factors in CRC. Using a TCGA dataset of 379 CRC patients, we have found significant positive or negative associations between patient survival and the levels of six miRNAs, 11 mRNAs, and nine methylated DNA species. Furthermore, we proposed a prognostic index (PI) as an integrated measure of all these factors, and validated its efficiency in differentiating between groups of different lengths of survival. Interestingly, combination of PI and tumor stage resulted in more precision in distinguishing overall survival between different patient subgroups, suggesting PI as a promising indicator to be incorporated into prognosis estimation for colorectal cancer patients.

Materials and Methods

Data acquisition and preprocessing

A dataset containing information of 379 CRC patients was downloaded from the TCGA (The Cancer Genome Atlas) database, including expression levels of 1, 881 mRNAs and 60, 484 miRNAs, genome-wide level 3 data of 295 specimens with 482, 421 DNA methylation sites (Illumina methylation 450), and survival. Afterwards, a filter was applied and removed genes whose transcript (mRNA or miRNA) or DNA methylation level was absent in more than 50% of all samples.

Statistical analysis

All statistical analyses were performed with R, a statistical software program, and related packages survival, limma, glmnet. The effect of age, gender, pathological stage and TNM stage on survival rate was assessed by Cox regression, respectively, and considered statistically significant when $p \leq 0.05$. Kaplan-Meier survival analysis was used to plot the survival curve. Additionally, log-rank test was used to evaluate the difference of survival curves between different groups, where $p \leq 0.05$ is considered statistically significant.

Screening for genes and DNA methylation sites closely associated to patient survival

Cox regression analysis was used to evaluate the association of the level of each gene and methylated DNA species with survival length. Due to the large number of genes and methylation sites included in the dataset, we used an analytical strategy that combines univariate Cox regression with Lasso-Cox (Least Absolute Shrinkage and Selection Operator-Cox) to identify survival-associated factors. First, univariate Cox regression was applied to each gene and methylation site to evaluate its effect on patient survival, followed by a filter which retains those with $p \leq 0.1$ for subsequent analysis. Subsequently, Lasso-Cox was performed to further screen and shrink the data. Afterwards, association of gene expression and methylation level with patient survival was further analyzed with multivariate Cox regression, which adjusts for other patient characteristics (e.g. age and sex), as well as medical factors (e.g. pathological stage).

Screening for potential miRNA target genes

Next, for miRNAs whose expression levels showed significant association with patient survival, we retrieved their potential target genes from miRTarBase (the experimentally validated microRNA-target interactions database, release 7.0). miRTarBase is a manually collected, experimentally validated, and actively used database of miRNA targets. An miRNA-target co-expression network was then constructed. High-confidence ($p \leq 0.05$) miRNA targets were identified and subjected to Lasso-Cox regression to screen for high-confidence targets. Finally, an miRNA-target interaction network was constructed by Cytoscape (version 3.5.1).

Construction of gene regulatory network and identification of prognostic factors

After identifying survival-associated mRNAs, miRNAs and DNA methylation sites, we constructed an interaction network among these features using Ingenuity Pathway Analysis (www.ingenuity.com). Univariate Cox regression was applied to further screen for significant nodes ($p \leq 0.1$) in the network, in order to detect potentially important mRNAs that may have been missed in previous analyses. Subsequently, effect of each candidate on patient survival was evaluated with multivariate Cox regression, and a final list of candidate prognostic factors was established by retaining those with $p < 0.05$.

Validation of the prognostic value of the candidate prognostic factors

To evaluate the prognostic value of the candidate prognostic factors obtained in the last section, a prognostic index (PI) was proposed as an integrated indicator. In brief, PI was calculated as a weighted sum of expression levels of the transcripts (mRNA/miRNA) and methylated species present in a given sample (denoted as x in the following equation), using the Cox regression coefficient β as the corresponding weights. Therefore, for specimen j , PI was calculated as follows:

$$PI = \sum_{i=1}^n \beta_i x_i$$

A greater value of PI is indicative of worse prognosis. Base on the corresponding PI value, patients were divided into high-risk ($PI > 0$) and low-risk groups ($PI < 0$). Afterwards, Kaplan-Meier survival analysis and log-rank test were used to evaluate the differences in patient survival between the two groups.

Results

Patient information

Of the 379 CRC patients, 206 (54.35% in all patients) were male and 173 (45.65%) female. The average age was 63.7 ± 11.7 years. Locations of the primary tumors encompassed a wide range, include the sigmoid colon (98 cases, 25.86% in all patients), ascending colon (81, 21.37%), cecum (68, 17.94%), transverse colon (16, 4.22%), descending colon (15, 3.95%), colon (10, 2.64%), and splenic flexure (3 cases, 0.79%).

Histologically, the majority of the cases were categorized as adenoma (M8140/3/1) or mucinous adenoma (M8480/3/1), accounting for 325 (85.75%) and 50 (13.19%) cases respectively. In terms of pathological stages, 62 cases (16.35%) were determined at stage I, 142 cases at stage II (37.47%), 110 cases at stage III (29.02%), and 55 cases of stage IV (14.51%). The average overall survival (OS) was 902 days (approx. 30 months), during which 77 patients passed away after follow-up.

We first examined the association between OS and each individual patient characteristics (e.g. age and sex) and medical

Table 1. Risk factors for the 379 colorectal cancer cases analyzed in this study

OS	Univariate analysis			Multivariate analysis		
	HR	95% CI	P value	HR	95% CI	P value
Age						
<=65						
>65	1.360	0.860, 2.151	0.189	1.795	1.099, 2.935	0.0195
Gender						
Female						
Male	1.123	0.715, 1.764	0.613	1.01	0.629, 1.623	0.966
Histology						
Adenoma						
Myxadenoma	1.132	0.596, 2.147	0.705			
Pathological stage						
Stage I						
Stage II	2.004	0.499, 0.589	0.266	1.89	0.554, 6.448	0.3092
Stage III	4.119	0.243, 1.240	0.021	4.462	1.341, 14.848	0.0148
Stage IV	11.556	0.086, 3.493	<0.001	12.48	3.767, 41.351	<0.001
TNM(T)						
T1						
T2	0.277	0.046, 1.675	0.162			
T3	0.941	0.228, 3.873	0.933			
T4	3.169	0.728, 13.784	0.124			
TNM(N)						
N0						
N1	2.023	1.165, 3.513	0.012			
N2	4.272	2.496, 7.310	<0.001			
TNM(M)						
M0						
M1	5.173	3.111, 8.600	<0.001			
M2	2.209	1.059, 4.608	0.035			

Fig. 1. Overall survival was compared among patient groups with different age (B), sex (C), and pathological stage (D). While groups of different ages and sexes did not show significant difference, patients with tumors at stages III and IV showed significantly different OS with those with tumors at stages I and II. Dashed lines indicate 90% CI.

Table 2. miRNAs and mRNA whose expression levels showed significant association with overall survival. HR: Hazard Ratio; CI: 95.0% confidence interval; SE: standard errors of Coefficient; z value: Wald z-statistic value

Molecular	ID	Coefficient	HR	95% CI	SE	z value	p value
miRNA	hsa-mir-552	-0.299	0.74176	0.568, 0.968	0.136	-2.193	0.028
	hsa-mir-193a	0.27	1.30978	0.910, 1.885	0.186	1.453	0.146
	hsa-mir-145	-0.295	0.74448	0.522, 1.062	0.181	-1.628	0.103
	hsa-mir-216a	0.247	1.2806	0.985, 1.664	0.134	1.849	0.065
	hsa-mir-144	-0.308	0.73475	0.457, 1.181	0.242	-1.272	0.203
	hsa-mir-149	-0.031	0.96925	0.698, 1.347	0.167	-0.187	0.852
	hsa-mir-503	0.364	1.43924	1.095, 1.891	0.139	2.615	0.009
	hsa-mir-26b	0.322	1.37924	0.965, 1.972	0.182	1.762	0.078
	hsa-mir-887	0.167	1.18119	0.835, 1.670	0.177	0.942	0.346
	hsa-mir-4442	-0.309	0.73439	0.559, 0.965	0.139	-2.218	0.027
	hsa-mir-486-2	0.012	1.01191	0.605, 1.692	0.262	0.045	0.964
	hsa-mir-891a	0.271	1.31071	0.999, 1.719	0.138	1.954	0.051
	hsa-mir-3677	-0.21	0.81072	0.614, 1.070	0.142	-1.48	0.139
	hsa-mir-328	0.043	1.0436	0.718, 1.516	0.191	0.224	0.823
	hsa-mir-6514	-0.24	0.78663	0.577, 1.072	0.158	-1.521	0.128
	hsa-mir-664b	0.167	1.18179	0.820, 1.703	0.187	0.895	0.371
	hsa-mir-34b	-0.038	0.96314	0.735, 1.262	0.138	-0.273	0.785
	hsa-mir-6854	-0.397	0.67243	0.509, 0.888	0.142	-2.791	0.005
	ENSG00000228126	0.271	1.311	1.063, 1.617	0.107	2.526	0.012
	ENSG00000230306	0.044	1.045	0.920, 1.188	0.065	0.683	0.494
mRNA	ENSG00000244242	0.255	1.291	1.096, 1.520	0.083	3.059	0.002
	ENSG00000237036	0.116	1.123	0.901, 1.400	0.112	1.035	0.301
	ENSG00000229320	0.114	1.121	0.883, 1.424	0.122	0.938	0.348
	ENSG00000217275	0.011	1.011	0.823, 1.242	0.105	0.108	0.914
	ENSG00000249001	0.272	1.312	1.012, 1.701	0.133	2.048	0.041

Table 3. DNA methylation sties whose methylation levels showed significant association with overall survival. HR: Hazard Ratio; CI: 95.0% confidence interval; SE: standard errors of Coefficient; z value: Wald z-statistic value

Probe	Ref Gene	Coefficient	HR	95% CI	SE	z value	P value
cg00147009	KRR1	-0.458	0.633	0.405, 0.987	0.227	-2.016	0.044
cg00199007	SLC17A9	0.450	1.569	1.196, 2.057	0.138	3.255	<0.001
cg01467592	VPS28	-0.272	0.762	0.572, 1.015	0.146	-1.861	0.063
cg03895094	TAP2	-0.437	0.646	0.533, 0.783	0.098	-4.467	<0.001
cg04138846	MRNIP	0.858	2.358	1.541, 3.609	0.217	3.952	<0.001
cg05821816	PCCA	-0.078	0.925	0.761, 1.123	0.099	-0.789	0.430
cg06328127	NXPH1	0.644	1.904	1.034, 3.509	0.312	2.066	0.039
cg06882340	EPHA2	-0.127	0.881	0.667, 1.164	0.142	-0.893	0.372
cg07206725	CDKL3	-0.547	0.579	0.351, 0.954	0.255	-2.145	0.032
cg08021727	ADCY9	0.276	1.317	0.897, 1.936	0.196	1.404	0.160
cg10248231	PRDM16	0.316	1.371	1.008, 1.866	0.157	2.010	0.044
cg10584300	CPE	-0.259	0.772	0.608, 0.980	0.122	-2.124	0.034

factors (tumor histology, pathological stage, and TNM stage, etc.). Univariate Cox regression analysis revealed that pathological stage and TNM stage were most significantly associated with OS (Table 1). Additionally, Kaplan-Meier survival analysis and log-rank test showed that there was a significant difference in OS among patients with different tumors of pathological stages (Fig. 1D), but not between different age (> 65 and ≤ 65 years old) or sex groups (Fig. 1B & 1C). In particular, patients with specimens at stages III and IV showed significant decreases in OS (Fig. 1D), whereas specimens at stages I and II did not. A second analysis with the multivariate Cox regression model was further performed, in order to examine each factor after adjusting for all others, and validated pathological stage as the most significant survival-association factor (Table 1).

Screening miRNAs, mRNAs and DNA methylation sites for potential prognostic factors

From 1, 181 miRNAs, 60, 484 mRNAs and 394, 353 methylation sites, a combination of univariate Cox and Lasso-Cox analyses identified a series of factors closely associated with patient survival, including 18 miRNAs, 11 mRNAs and 12 methylation sites (Tables 2 and 3). These features were further validated with a multivariate Cox regression analysis. Among these features, worse prognosis was shown to be significantly associated with high expression of hsa-mir-503, ENSG00000228126, ENSG00000244242, and ENSG00000249001 (Table 2), as well as hypermethylation of cg00199007, cg04138846, cg06328127, and cg10248231 (Table 3). A better prognosis, on the other hand, showed strong association with elevated expression of hsa-mir-4442, hsa-mir-552, and hsa-mir-6854 (Table 2), as well as hypermethylation of cg00147009, cg03895094, cg07206725, and cg10584300 (Table 3).

Next, we set to identify high-confidence targets for the survival-associated miRNAs in Table 2. First, a list of 1, 749 target genes for these miRNAs were retrieved from the miRTarBase database. Then, a Lasso-Cox strategy was used to further select 38 target genes showing significant co-expression with the corresponding miRNAs in our TCGA dataset. A co-expression network of these miRNAs and their potential targets was constructed, visualizing the strong, diverse correlation between the expression levels of each miRNA and its targets (Fig. 2).

Identification of candidate prognostic factors

In the next step, a regulatory network was constructed, showing regulations reported thus far between the miRNAs and PCGs from Table 2, as well as the 38 high-confidence miRNA targets identified in the last section. As shown in Fig. 3, the resulting network is one of considerable connectivity, suggesting close regulations between the high-confidence miRNA

Fig. 2. A co-expression network of survival-associated miRNAs and potential target genes. Yellow boxes indicate miRNAs; orange boxes target genes. Solid lines indicate significantly co-expressed miRNA-target pairs ($p \leq 0.05$).

targets and the survival-associated miRNAs and PCGs. In particular, the highly connected nodes have been established to play roles in a plethora of cellular processes, including carcinogenesis, tumor proliferation, invasion, metastasis, and apoptosis (Table 4). Among others, mir-145, mir-26, and mir-34 have been implicated in adenocarcinoma growth, EPK2, mir-145 and mir-26 in tumor cell migration. In addition, CDK5R1, EPHA2, FOSL1, mir-328, mir-34, mir-503, and TGFB2 have been proposed roles in CRC metastasis. The high relevance of these survival-associated genes and miRNA targets in cancer biology suggest the power of the network-based analyses employed in this study in identifying essential factors related to patient prognosis.

As a final round of identification and validation, all the nodes in this regulatory network was subjected to a multivariate Cox regression analysis, in order to examine their association with patient survival. Compared with the previous multivariate Cox regression, which validated the 18 miRNAs and seven mRNAs in Table 2, this analysis also examined the association between patient survival and the 38 high-confidence miRNA targets. As miRNAs can exert their functions through downregulating the mRNA levels of their targets, addition of these 38 targets allowed the examination potential relevant factors that could have been missed in previous screenings due to a p value slightly higher than the cut-off threshold. Also, the new multivariate analysis took into account mRNA levels of these 38 targets in correcting the significance level of each survival-associated genes and methylated sites, thereby allowing screening for significant prognostic factors from a densely connected functional network.

This analysis yielded a final list consisting of 6 miRNAs, 11 mRNA (four from the 38 miRNA targets) and 9 methylation sites (Table 5), which were considered as candidate prognostic factors in colorectal cancer.

Validation of the candidate prognostic factors

To validate the prognostic value of all candidate prognostic factors identified in the last section, a prognostic index (PI) was proposed as an integrative prognostic predictor. A PI

Table 5. Candidate prognostic factors identified in this study, including 6 miRNAs, 11 mRNAs, and 9 methylation sites. HR: Hazard Ratio; CI: 95.0% confidence interval; SE: standard errors of Coefficient; z value: Wald z-statistic value

Molecular	ID	Ref Gene	Coefficient	HR	95% CI	SE	z value	p value
miRNA	hsa-mir-552	-	-0.298	0.743	0.586, 0.941	0.121	-2.467	0.014
	hsa-mir-216a	-	0.348	1.416	1.107, 1.812	0.126	2.765	0.006
	hsa-mir-26b	-	0.271	1.311	1.032, 1.666	0.122	2.217	0.027
	hsa-mir-4442	-	-0.387	0.679	0.535, 0.862	0.122	-3.175	0.002
	hsa-mir-891a	-	0.339	1.404	1.128, 1.747	0.112	3.036	0.002
	hsa-mir-6854	-	-0.333	0.717	0.556, 0.924	0.129	-2.57	0.01
mRNA	ENSG00000228126	FALEC	0.366	1.442	1.173, 1.772	0.105	3.473	0.001
	ENSG00000011275	RNF216	0.24	1.271	1, 1.616	0.122	1.962	0.05
	ENSG00000163792	TCF23	0.202	1.224	1.085, 1.382	0.062	3.287	0.001
	ENSG00000122378	FAM213A	0.219	1.244	1.073, 1.442	0.075	2.898	0.004
	ENSG00000102962	CCL22	-0.597	0.55	0.356, 0.852	0.223	-2.678	0.007
	ENSG00000204390	HSPA1L	0.331	1.392	1.109, 1.746	0.116	2.855	0.004
	ENSG00000138685	FGF2	0.331	1.392	1.171, 1.654	0.088	3.751	<0.001
	ENSG00000244242	IFITM10	0.219	1.245	1.041, 1.489	0.091	2.396	0.017
	ENSG00000159958	TNFRSF13C	0.385	1.47	1.225, 1.764	0.093	4.141	<0.001
	ENSG00000214357	NEURL1B	-0.476	0.621	0.451, 0.856	0.163	-2.911	0.004
	ENSG00000175592	FOSL1	0.298	1.348	1.064, 1.707	0.121	2.472	0.013
Methy	cg00147009	KRR1	-0.402	0.669	0.467, 0.958	0.183	-2.192	0.028
	cg00199007	SLC17A9	0.447	1.564	1.263, 1.936	0.109	4.101	<0.001
	cg01467592	VPS28	-0.373	0.688	0.55, 0.862	0.115	-3.254	0.001
	cg03895094	TAP2	-0.455	0.635	0.526, 0.766	0.096	-4.737	<0.001
	cg04138846	MRNIP	0.869	2.385	1.558, 3.65	0.217	4.003	<0.001
	cg06328127	NXPH1	0.64	1.896	1.052, 3.42	0.301	2.127	0.033
	cg07206725	CDKL3	-0.476	0.622	0.391, 0.988	0.237	-2.011	0.044
	cg10248231	PRDM16	0.381	1.464	1.094, 1.961	0.149	2.561	0.01
	cg10584300	CPE	-0.317	0.728	0.581, 0.912	0.115	-2.765	0.006

was calculated for each specimen as the weighted sum of expression/methylation level of each candidate gene/methylation site, with the corresponding Cox regression coefficients as weights. Consequently, a higher PI is a result of high expression levels of the miRNAs, mRNAs, or methylations whose expression associated with poor prognosis, as well as low expression levels of those whose expression associated with favorable prognosis. Therefore, a higher PI should be indicative of poor prognosis.

To test the efficiency of PI in grouping patients by overall survival, the 379 patients in the dataset were divided into two groups and subjected to survival analysis. The low-risk group (PI < 0) included 150 cases, whereas the high-risk group (PI > 0) included 145. There were five tumor-related deaths in the low-risk group, and 52 in the high-risk group. In addition, median OS for the low and high-risk groups were 783 and 514 days, respectively (Fig. 4). The former showed an average 1-, 3-, and 5-year survival rates 89.3%, 54.9%, and 35.9%, respectively, while those of the latter group were 70.4%, 37.6%, and 18.4%, respectively.

Furthermore, Kaplan-Meier survival analysis and log-rank test showed a significant difference in the OS between the two risk groups, with the high-risk group exhibiting poorer survival (log-rank test $p < 0.0001$, Fig. 5A).

Interestingly, difference in the OS between patients grouped by tumor stage appeared to depend on PI level (Fig. 5B). Patients in the low- and high-risk groups can be further divided into a total of four subgroups, based on pathological staging. In the two high-risk subgroups

Fig. 4. Median survival for patients in the low-risk (PI < 0) and high-risk (PI > 0) groups. Low-risk group showed a significantly greater survival time compared with the high-risk group ($p < 0.0001$).

Fig. 5. Kaplan-Meier curve showing OS of patients grouped by (A) PI, and (B) PI and pathological stage. (A) Low-risk (PI < 0) and high-risk (PI > 0) groups showed significantly different survival curves. (B) Patients in the low- and high-risk groups were further divided based on pathological stages of the tumor (stage I/II vs stage III/IV), and analyzed for difference in the corresponding survival curves.

were further divided based on pathological stages of the tumor (stage I/II vs stage III/IV), and analyzed for difference in the corresponding survival curves.

($PI > 0$), there was significant different (Cox $p < 0.0001$) in OS between the earlier (stages I/II) and later stages (stages III/IV). This observation was consistent with known fact. However, the two low-risk subgroups ($PI < 0$) showed no significant difference in OS (Cox $p = 0.185$). On the other hand, PI level was efficient in grouping patients by OS in manner that is independent of pathological staging. There were significant differences in OS between the low- and high-risk subgroups when both subgroups were of earlier stages (Cox $p < 0.001$) or of later stages (Cox $p < 0.0001$).

Together, these results suggest high value of the candidate prognostic factors identified in this study. Moreover, incorporation of PI into the conventional practice of patient grouping by pathological stage increased the precision with which to distinguish patients with different lengths of survival, thereby suggesting the use of tumor stage and PI to achieve a more accurate prognosis for CRC patients.

Discussion

Rapid progress of next-generation sequencing has provided a number of novel tools and thereby enabled the scientific community to approach cancers from previously impractical perspectives and to gain deeper insights. A first attempt was launched by The Cancer Genome Atlas, which conducted a genome-wide characterization of somatic mutations in CRC [24]. A variety of transcriptomic tools, such as mRNA and miRNA expression and promoter methylation, were applied. Based on the extent of mutations, a total of 276 CRC specimens were divided into two subsets, one showing hypermutation (16% of all cases), of which the majority exhibited high microsatellite instability and mutations in mismatch repair genes. The remaining 84% specimens showed similar patterns of mutations in 24 genes, including *APC*, *TP53*, and *KRAS*. In addition, recurrent alterations in copy-number (e.g. *ERBB2* and *IGF2*) and chromosomal translocations were also identified. These alterations serve as the basis for CRC diagnosis and subtyping, and warranted investigation to establish the clinical practice of applying the subtyping in improving prognosis estimation.

Later, Guinney et al. [25] orchestrated a consortium and analyzed a large of number of CRC patients with a network-based approach. Among more than 4000 cases, the majority presented distinctly distinguishable features. Samples with mixed features accounted for only 13% of all cases. Four robust consensus molecular subtypes (CMSs) were identified, including a hypermutated, microsatellite unstable CMS1 subtype, a CMS2 marked by WNT and MYC signaling activation, a CMS3 showing evident metabolic dysregulation, and a CMS4 with marked activation of transforming growth factor- β signaling. Through characterizing the subtypes of CRC with transcriptomics approaches, this study revealed convergent pathway dependencies in thousands of cases, and proposed a classification system with clear biological interpretability, which could prove useful for clinical stratification and targeted interventions [26, 27].

In the present study, through a comprehensive analysis of miRNA and mRNA expression, DNA methylation, and patient survival of 379 CRC patients documented in the TCGA database, we identified a set of genetic and epigenetic modifications as potential prognostic factors, and validated the value of a prognostic index (PI) that takes into consideration all these factors. Furthermore, based on the association between pathological stage and patient survival, we propose a novel subgrouping system combining PI and pathological staging, for predicting prognosis for CRC patients.

We started with screening for clinical features associated with prognosis. A combination of Cox regression and Kaplan-Meier survival analysis revealed that overall survival and tumor-free survival showed little association with patient characteristics such as age and sex, but were significantly different between patients at different pathological stages (Fig. 1 & Fig. 2). These findings were consistent with known fact. Further screening of genetic and epigenetic aberrations highlighted a collection of 18 miRNAs, seven mRNAs, and 12 DNA methylation sites whose expression/methylation levels were significantly associated with

overall survival (Tables 1 & 2). In order to identify potential target genes of the survival-associated 18 miRNAs, we retrieved targets from miRTarBase, a curated database of experimentally validated miRNA-target interactions [28], and performed a screening using a Lasso-Cox approach. Thirty-eight high-confidence targets were selected and showed highly correlated co-expression with their corresponding miRNAs (Fig. 3), suggesting potential regulations between the miRNA-target pairs [29], and thereby provide hints for downstream effectors through which the 18 miRNAs affect CRC prognosis.

As miRNAs exert their roles in CRC by downregulating the mRNA levels of their targets, it was possible that some survival-associated mRNAs have been missed during the first screening due to a slightly higher p value. To validate the functional relevance of these 38 targets with the survival-associated miRNAs and PCGs, a regulatory network was first constructed with these genes to visualize the reported biological relationships of the 38 targets in this network. Subsequently, a multivariate Cox regression was performed to screen all the nodes from this network and survival-associated methylated DNA species for factors significantly associated with OS. This analysis produced a list of candidate prognostic factors consisting of six miRNAs, 11 mRNA (four from the 38 miRNA targets) and nine methylation sites.

In order to validate the prognostic value of these candidates, a prognostic index (PI) was computed as an integrative indicator. Based on which patients were divided into low-risk ($PI < 0$) and high-risk ($PI > 0$) groups. Kaplan-Meier survival analysis revealed significant differences between OS of these two groups, suggesting the efficiency of the PI as a potential prognostic indicator. Furthermore, after combining PI with tumor pathological stages, a more precise prognosis was possible. On one hand, for patients in the same stage group (stages I & II), prognosis can be significantly different depending on the PI level. On the other, in the low-risk group, there was no significant difference in OS in patients at different stage groups (Cox $p = 0.185$), suggesting prognosis as the outcome of a comprehensive panel of factors, and therefore require a more well-defined approach in predicting prognosis.

These findings can be of considerable clinical importance, especially for patients at the later stages (stages III & IV). Evaluation of PI could provide assistance to making a more accurate prognosis, but also to planning therapeutic interventions. For instance, a low PI could help avoid prescribing a patient intense yet potentially unnecessary radio- and/or chemotherapies, and, equally importantly, provide mental relief for the patient.

However, our prognostic index model didn't take into consideration risk factors like lifestyle (diet, physical activity, tobacco smoking, etc) and therapeutic regimen. Many studies support the association between lifestyle and the development of colorectal cancer [30-34]. Various energy balance host factors, including obesity, physical inactivity and certain dietary factors can affect patient's outcome [33]. And, smoking has detrimental effects on survival even after CRC diagnosis [34]. The NCCN guideline recommend different therapeutic regimen according to disease status, like operation for early stage patients, targeted therapy for advanced patients without RAF/RAS driver gene mutation, chemotherapy for advanced patients with RAF/RAS driver gene mutation. Constrained by the limited lifestyle and therapeutic regimen information in our data, we have to assume that these factors are balanced between low-risk and high-risk groups identified by our prognostic index model. We will try to validate this model with patients in our hospital and think carefully the influence of lifestyle and therapeutic regimen in the future trail.

Conclusion

In summary, in this study, starting from a TCGA dataset of 379 colorectal cancer cases, we applied an integrative analysis combining Cox regression, co-expression and gene regulatory network analyses, and survival analysis to identify factors significantly associated to patient survival. By screening incorporating patient characteristics, genetic and epigenetic aberrations, and survival data, we identified a collection of 6 miRNAs, 11

mRNAs, and 9 DNA methylation sites whose expression levels were significantly associated with overall survival. An integrative prognostic index was proposed and used to divide patients to groups of different risk levels, which showed significantly different survival, thereby suggesting prognostic value of the PI and its underlying potential prognostic factors. Importantly, we found that combination of PI and tumor stage provided additional precision in subtyping patients with significantly different overall survival, suggesting the promise of a more comprehensive subtyping system, using both PI and pathological staging in predicting prognosis. Further research is warranted to validate the efficiency of this subtyping system, as well as to elucidate the roles these genetic and epigenetic modifications play in colon cancer biology.

Acknowledgements

We thank all the doctors in our department for their help in patients care so that we can have enough time to conduct this study.

Disclosure Statement

The authors declare that they have no competing interests.

References

- 1 Harrison S, Benziger H: The molecular biology of colorectal carcinoma and its implications: a review. *Surgeon* 2011;9:200-210.
- 2 Dattatreya S: Metastatic colorectal cancer-prolonging overall survival with targeted therapies. *South Asian J Cancer* 2013; 2:179-185.
- 3 Ting WC, Chen LM, Pao JB, Yang YP, You BJ, Chang TY, Lan YH, Lee HZ, Bao BY: Common genetic variants in Wnt signaling pathway genes as potential prognostic biomarkers for colorectal cancer. *PLoS One* 2013;8:e56196.
- 4 Duggan MA, Anderson WF, Altekruse S, Penberthy L, Sherman ME: The Surveillance, Epidemiology, and End Results (SEER) Program and Pathology: Toward Strengthening the Critical Relationship. *Am J Surg Pathol* 2016;40:e94-e102.
- 5 Coppede F, Lopomo A, Spisni R, Migliore L: Genetic and epigenetic biomarkers for diagnosis, prognosis and treatment of colorectal cancer. *World J Gastroenterol* 2014;20:943-956.
- 6 Peluso G, Incollingo P, Calogero A, Tammaro V, Rupealta N, Chiacchio G, Sandoval Sotelo ML, Minieri G, Pisani A, Riccio E, Sabbatini M, Bracale UM, Dodaro CA, Carlomagno N: Current Tissue Molecular Markers in Colorectal Cancer: A Literature Review. *Biomed Res Int* 2017;2017:2605628.
- 7 Marmol I, Sanchez-de-Diego C, Pradilla Dieste A, Cerrada E, Rodriguez Yoldi MJ: Colorectal Carcinoma: A General Overview and Future Perspectives in Colorectal Cancer. *Int J Mol Sci* 2017;18:p11 E197.
- 8 Hu H, Cai W, Zheng S, Ge W: SPARCL1, a Novel Prognostic Predictive Factor for GI Malignancies: a Meta-Analysis. *Cell Physiol Biochem* 2017;44:1485-1496.
- 9 Li Q, Li Y, Xu J, Wang S, Xu Y, Li X, Cai S: Aldolase B Overexpression is Associated with Poor Prognosis and Promotes Tumor Progression by Epithelial-Mesenchymal Transition in Colorectal Adenocarcinoma. *Cell Physiol Biochem* 2017;42:397-406.
- 10 Song BR, Xiao CC, Wu ZK: Predictors of Lymph Node Metastasis and Prognosis in pT1 Colorectal Cancer Patients with Signet-Ring Cell and Mucinous Adenocarcinomas. *Cell Physiol Biochem* 2017;41:1753-1765.
- 11 Zhou X, Xie S, Yuan C, Jiang L, Huang X, Li L, Chen Y, Luo L, Zhang J, Wang D, Liu L, Shi W, Han L, Tang N, Ji Y: Lower Expression of SPRY4 Predicts a Poor Prognosis and Regulates Cell Proliferation in Colorectal Cancer. *Cell Physiol Biochem* 2016;40:1433-1442.
- 12 Jia H, Wang Z: Telomere Length as a Prognostic Factor for Overall Survival in Colorectal Cancer Patients.

Cell Physiol Biochem 2016;38:122-128.

- 13 Wang H, Yao H, Li C, Liang L, Zhang Y, Shi H, Zhou C, Chen Y, Fang JY, Xu J: PD-L2 expression in colorectal cancer: Independent prognostic effect and targetability by deglycosylation. *Oncoimmunology* 2017;6:e1327494.
- 14 Rashed HE, Hussein S, Mosaad H, Abdelwahab MM, Abdelhamid MI, Mohamed SY, Mohamed AM, Fayed A: Prognostic significance of the genetic and the immunohistochemical expression of epithelial-mesenchymal-related markers in colon cancer. *Cancer Biomark* 2017;20:107-122.
- 15 Mazeh H, Mizrahi I, Ilyayev N, Halle D, Brucher B, Bilchik A, Protic M, Daumer M, Stojadinovic A, Itzhak A, Nissan A: The Diagnostic and Prognostic Role of microRNA in Colorectal Cancer - a Comprehensive review. *J Cancer* 2013;4:281-295.
- 16 Shao Y, Gu W, Ning Z, Song X, Pei H, Jiang J: Evaluating the Prognostic Value of microRNA-203 in Solid Tumors Based on a Meta-Analysis and the Cancer Genome Atlas (TCGA) Datasets. *Cell Physiol Biochem* 2017;41:1468-1480.
- 17 Ha M, Kim VN: Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol* 2014;15:509-524.
- 18 Toiyama Y, Hur K, Tanaka K, Inoue Y, Kusunoki M, Boland CR, Goel A: Serum miR-200c is a novel prognostic and metastasis-predictive biomarker in patients with colorectal cancer. *Ann Surg* 2014;259:735-743.
- 19 Hansen TF, Christensen R, Andersen RF, Sorensen FB, Johnsson A, Jakobsen A: MicroRNA-126 and epidermal growth factor-like domain 7-an angiogenic couple of importance in metastatic colorectal cancer. Results from the Nordic ACT trial. *Br J Cancer* 2013;109:1243-1251.
- 20 Feinberg AP, Vogelstein B: Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* 1983;301:89-92.
- 21 Nilsson TK, Lof-Ohlin ZM, Sun XF: DNA methylation of the p14ARF, RASSF1A and APC1A genes as an independent prognostic factor in colorectal cancer patients. *Int J Oncol* 2013;42:127-133.
- 22 Xing X, Cai W, Shi H, Wang Y, Li M, Jiao J, Chen M: The prognostic value of CDKN2A hypermethylation in colorectal cancer: a meta-analysis. *Br J Cancer* 2013;108:2542-2548.
- 23 Baba Y, Noshio K, Shima K, Huttenhower C, Tanaka N, Hazra A, Giovannucci EL, Fuchs CS, Ogino S: Hypomethylation of the IGF2 DMR in colorectal tumors, detected by bisulfite pyrosequencing, is associated with poor prognosis. *Gastroenterology* 2010;139:1855-1864.
- 24 Cancer Genome Atlas N: Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;487:330-337.
- 25 Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot BM, Morris JS, Simon IM, Gerster S, Fessler E, De Sousa EMF, Missiaglia E, Ramay H, Barras D, Homicsko K et al.: The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015;21:1350-1356.
- 26 Dienstmann R, Vermeulen L, Guinney J, Kopetz S, Tejpar S, Tabernero J: Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat Rev Cancer* 2017;17:79-92.
- 27 Chou CH, Shrestha S, Yang CD, Chang NW, Lin YL, Liao KW, Huang WC, Sun TH, Tu SJ, Lee WH, Chiew MY, Tai CS, Wei TY, Tsai TR, Huang HT, Wang CY, Wu HY, Ho SY, Chen PR, Chuang CH et al.: miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res* 2018;46:D296-D302.
- 28 de la Fuente A: From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends Genet* 2010;26:326-333.
- 29 Buset M: Primary prevention of colorectal cancer. *Acta Gastroenterol Belg* 2003;66:20-27.
- 30 Vijayvergia N, Denlinger CS: Lifestyle Factors in Cancer Survivorship: Where We Are and Where We Are Headed. *J Pers Med* 2015;5:243-263.
- 31 Campbell PT, Newton CC, Dehal AN, Jacobs EJ, Patel AV, Gapstur SM: Impact of body mass index on survival after colorectal cancer diagnosis: the Cancer Prevention Study-II Nutrition Cohort. *J Clin Oncol* 2012;30:42-52.
- 32 Schmid D, Leitzmann MF: Association between physical activity and mortality among breast cancer and colorectal cancer survivors: a systematic review and meta-analysis. *Ann Oncol* 2014;25:1293-1311.
- 33 Walter V, Jansen L, Hoffmeister M, Ulrich A, Chang-Claude J, Brenner H: Smoking and survival of colorectal cancer patients: population-based study from Germany. *Int J Cancer* 2015;137:1433-1445.
- 34 Walter V, Jansen L, Hoffmeister M, Brenner H: Smoking and survival of colorectal cancer patients: systemativ review and meta-analysis. *Ann Oncol* 2014;25:1517-1525.