Am J Cancer Res 2025;15(7):3128-3141
www.ajcr.us /ISSN:2156-6976/ajcr0163090

Original Article

Transcripts of splicing factors with time-varying associations with survival outcomes in colorectal cancer

Yajun Yu[1*], Yingla Li[2*], Lu Ma[3*], Huilin Chen[4], Yiran Peng[2], Haiqiong Yang[5], Qiuting Li[1], Xiaoping Gao[1], Jianguo Feng[4], Yang Yu[5], Ying Wan[6], Yan Yang[1], Yan Deng[6], Chunxiang Zhang[1,5,6]

[1]Key Laboratory of Medical Electrophysiology, Ministry of Education, Institute of Cardiovascular Research and Institute of Metabolic Diseases, Southwest Medical University, Luzhou 646000, Sichuan, China; [2]School of Pharmacy, Southwest Medical University, Luzhou 646000, Sichuan, China; [3]Department of Neonatology, Children's Medical Center, The Affiliated Hospital, Southwest Medical University, Luzhou 646000, Sichuan, China; [4]Department of Anesthesiology, The Affiliated Hospital, Southwest Medical University, Luzhou 646000, Sichuan, China; [5]Department of Cardiology, The Affiliated Hospital, Southwest Medical University, Luzhou 646000, Sichuan, China; [6]School of Basic Medical Science, Southwest Medical University, Luzhou 646000, Sichuan, China. [*]Equal contributors

Received January 4, 2025; Accepted June 25, 2025; Epub July 15, 2025; Published July 30, 2025

Abstract: Identifying variables with time-varying associations can help guide patient stratification and treatment strategies. Time-varying associations refer to the associations between variables of interest and the outcomes that change over time. Such associations of gene transcripts have never been examined in colorectal cancer (CRC). In the present study, we examined the transcripts of splicing factors and their associations with prognosis in CRC. We performed analyses using Cox PH models and Cox models with time-varying coefficients when appropriate. A transcript of pre-mRNA processing factor 38A (*PRPF38A*) was identified to be significantly associated with the progression-free survival after two years post-diagnosis, but not before that, making this transcript a candidate late-outcome marker in CRC. Further explorative analysis showed that this transcript correlated with the alternative splicing of genes involved in RNA binding and mitochondrial protein synthesis. Overall, our study, for the first time, revealed the association patterns of splicing factor transcripts over time in CRC, identified a transcript of *PRPF38A* as a candidate late-outcome marker, and provided mechanistic insights into outcome heterogeneity in CRC.

Keywords: Colorectal cancer, prognosis, proportional hazards (PH) assumption, *PRPF38A*, splicing factor transcripts, time-varying associations

Introduction

Colorectal cancer (CRC) is one of the most common types of cancer globally and causes ~700,000 deaths annually [1]. Many patients with CRC experience local recurrence and metastasis [2, 3]. Predicting the early- and late-outcomes of patients and thus identifying patients with a high outcome risk within or after certain years post-diagnosis, is crucial, as this helps guide disease treatment and management. Unfortunately, prognostic markers for such a prediction in CRC are lacking. To date, only a few studies have been performed to identify early- or late-outcome markers of this disease [4-17], and these have mainly focused on clinicodemographic variables and genetic variations. Other forms of gene-related variations, including changes in gene transcript levels, have not yet been examined. Gene transcripts represent another layer of gene-related variation that contributes to phenotypes. Some transcripts (or corresponding isoforms) are reportedly associated with the prognosis of CRC patients [18-22]. However, whether transcripts can be early- or late-outcome markers of CRC remains unknown.

Transcripts of splicing factors originate from splicing factor-coding genes, often with multiple transcripts generated per gene. Splicing factors are components of the spliceosome, which is a protein-RNA complex that functions in the splicing of pre-mRNAs [23, 24]. Impaired function or

https://doi.org/10.62347/DWPA1730

altered expression levels of splicing factors can lead to altered gene splicing (e.g., exon skipping, intron retention, and 5' or 3' alternative splicing sites), which further affects gene function and related biological processes and phenotypes, including the outcomes of CRC patients [25-29]. Given the importance of splicing factors in gene function and phenotypes, it is of interest and worthwhile to investigate splicing factor transcripts for their prognostic value in CRC, especially their potential as early- and late-outcome markers, which has never been examined in previous studies.

Candidate early- or late-outcome markers can be identified by taking advantage of examining the time-varying associations of the variables of interest. Time-varying associations are associations between variables of interest and survival outcomes that change over time [5, 6, 30]. Unlike variables with constant associations with survival outcomes, variables with time-varying associations have their prognostic associations become stronger, weaker, appear, or diminish over time. Thus, variables that have associations within or after certain years post-diagnosis can be early- or late-outcome markers [5, 6].

In the present study, we examined 985 transcripts derived from 144 splicing factors to determine their potential as prognostic markers in CRC. Our main objectives were to a) examine the associations between transcripts of splicing factors and survival outcomes in CRC; b) identify time-varying associations of splicing factor transcripts; and c) explore potential targets of identified transcripts and their related mechanisms on patients' outcomes.

Materials and methods

Patient cohort, clinical data, and transcript data of splicing factors

The Cancer Genome Atlas (TCGA) transcript and survival data of CRC were downloaded from UCSC Xena (https://xena.ucsc.edu/). Transcript data for primary tumors were used in this study. The clinicodemographic variables and MSI information of patients with CRC were downloaded from the GDC portal (https://portal.gdc.cancer.gov/). Genes (n = 217) encoding splicing factors in the spliceosome were downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.genome.jp/dbget-bin/www_bget?pathway:hsa03040). After excluding patients with either no clinicodemographic variables or no transcript data of spliceosome genes, 376 patients with 985 transcripts (from 144 genes) were examined. The baseline characteristics of the patients are summarized in **Table 1**.

Outcome measures

The outcome measures examined in this study were overall survival (OS), disease-specific survival (DSS), disease-free survival (DFS), and progression-free survival (PFS) from TCGA survival data of UCSC Xena (https://xena.ucsc.edu/), and these outcome data were previously summarized and curated by Liu et al. [31]. OS measures the survival from the disease diagnosis till the death from any causes, and it is usually regarded as the golden standard endpoint in clinical trials of oncology, though it can be influenced by non-cancer deaths and some other factors [32]. DSS measures the survival from disease diagnosis till the deaths caused by the cancer under study. It is a good outcome measure to assess the effects/impacts of treatments/variables specifically on the disease-related survival and has greater relevance to cancer biology. However, in TCGA data, this outcome was approximated because the absolute verification of the cause of death was hard to achieve [31]. Compared to OS and DSS, DFS (or DFI [disease-free interval], as used in UCSC Xena) and PFS (or PFI [progression-free interval], as used in UCSC Xena) are outcome measures generally require fewer patients and shorter follow-up to obtain a given number of outcome events, and thus often lead to reduced costs while remain the ability to provide information about treatment/variable activity/effects. DFS describes the survival from the completion of curative treatments to tumor relapse or death from any causes, and it is normally used for early-stage patients with curable tumors. PFS represents the survival from the start of palliative treatment to tumor progression (e.g., tumor growth, metastasis) or death from any causes. It is generally used in locally advanced or metastatic diseases. It is of note that although the DFS and PFS times should be calculated from the time of treatments, in the TCGA survival data of UCSC Xena these two outcome times were calculated from the time

Table 1. Baseline characteristics of TCGA COAD-READ dataset (n = 376)

Variable	Measure/Category	Number of patients	Percentage (%)
Age at diagnosis	Median (range)	66 (31-90) years	
Sex	Male	204	54.26
	Female	172	45.74
Stage	I	56	14.89
	II	133	35.37
	III	115	30.59
	IV	53	14.10
	NA	19	5.05
Location	Colon	285	75.80
	Rectum	91	24.20
MSI status	MSS/MSI-L	319	84.84
	MSI-H	57	15.16
Race	Non-Africans	285	75.80
	Africans	59	15.69
	NA	32	8.51
Follow-up time	Median (range)	672 (0-4502) days	
Overall survival (OS)	Alive	291	77.40
	Death	85	22.61
Disease-free survival (DSS)	Death from other causes or alive	314	83.51
	Death from colorectal cancer	41	10.90
	NA	21	5.59
Disease-free survival (DFS)*	0	112	34.67
	1	21	6.50
	NA	190	58.82
Progression-free survival (PFS)	0	273	72.61
	1	103	27.39

*, DFS information was available for stage I-III patients, not stage IV patients in this study dataset. MSI, microsatellite instability; MSI-H, microsatellite instability-high; MSI-L, microsatellite instability-low; MSS, microsatellite stable; NA, not available.

of diagnosis. This was mainly because the time of completing the curative treatments and the time of initiating the treatments were not available/complete in the TCGA data, and thus the date of diagnosis was used as a surrogate [31]. Readers should keep this in mind while interpreting our results. Another major difference is the death-related endpoint event used in DFS and PFS. Instead of using death from any causes, the TCGA survival data of UCSC Xena took death from advancement of the same tumor and death from tumor as the death-related endpoint events for DFS and PFS, respectively. Authors summarizing the TCGA survival data hold this opinion because deaths from tumor are more relevant in cancer studies [31]. This difference should also be considered while interpreting the results of this study.

Baseline model construction

Baseline models were constructed using the covariates of age at diagnosis, sex, disease stage, disease location, MSI status, and race. The PH assumption was checked for all variables in the baseline model using the cox.zph function in R (version 4.4.1) [33], and those that violated the PH assumption were stratified in the models. Transcripts were then individually added to the baseline models and examined for their associations with survival outcomes.

Statistical analysis

Correlation among baseline variables: Baseline variables were calculated for their pairwise Pearson correlation coefficient (r) values.

All variable pairs had r < 0.8, indicating no collinearity among the variables.

Survival analysis for splicing factor transcripts: The Cox PH model and the Cox PH model with time-varying coefficients were used to examine the associations between splicing factor transcripts and the survival outcomes of patients with CRC. Age at diagnosis, sex, disease stage, disease location, MSI status, and race were used as covariates to adjust for the effects/associations of the transcripts of interest. Treatment regimens or therapies were not included as covariates in this study because they had more than 10% missing data in the study dataset (Supplementary Tables 1, 2).

The PH assumption was tested using the cox.zph function, and a *p*-value of 0.05 was considered the significant threshold. The transcripts that satisfied the PH assumption were examined for their associations with outcomes in the Cox PH model, and those that violated the PH assumption (P < 0.05 in the PH assumption test) were identified for their best cutoff time points and then examined in Cox PH models with time-varying coefficients. The best cutoff of time points for a given transcript was identified using the method described by Yu *et al.* [5] and Yu *et al.* [6]. Briefly, models with different cutoff time points (with an increment of one year) were constructed and compared. The cutoff time point that gave the model the smallest Akaike information criterion (AIC) value, no infinity in 95% confidence intervals, and the satisfaction of the PH assumption both before and after the cutoff time point was deemed to be the best cutoff time point. All transcripts that violated the PH assumption were identified for the best cutoff time points and were examined in models with time-varying coefficients. The hazard ratios (HRs) and *p*-values were obtained from the constructed models.

All analyses were performed using R software (version 4.4.1) [33]. A strict *p*-value of 5.08×10^{-05} (i.e., 0.05/985; Bonferroni method for multiple testing correction) was considered the significant threshold. Variables with *p*-values less than this threshold were deemed to be the variables that are significantly associated with survival outcomes in CRC.

Survival analysis for gene expression levels of PRPF38A: RNA-Seq by Expectation Maximization (RSEM) normalized count data of pre-mRNA processing factor 38A (*PRPF38A*) in TCGA COAD and READ cohorts were downloaded from UCSC Xena (https://xena.ucsc.edu/). The association between *PRPF38A* expression and PFS was analyzed using the Cox PH model. The method used was the same as that used for transcript analysis. Statistical significance was set at $P < 0.05$.

Correlation analysis between PRPF38A short-transcript levels and the transcript ratios of other genes: Given that long non-coding RNAs (lncRNAs) have been reported to affect the alternative splicing of other genes through multiple mechanisms [34-36], and considering that the *PRPF38A* short transcript ENST00000474048.1 is a lncRNA of a splicing factor (which participates in the splicing of genes), it is possible that this transcript can affect the splicing of other genes. As an exploratory practice, we performed a correlation analysis between the short transcript of *PRPF38A* and the transcript ratios of other genes (n = 198,619). A strong correlation may imply an impact of the splicing factor transcript on the splicing of the correlated genes. Correlation analyses were performed using the cor.test function in R (version 4.4.1) [33]. A correlation with the absolute value of the Pearson correlation coefficient ≥ 0.5 and the associated $P < 2.52 \times 10^{-07}$ (i.e., 0.05/198,619; Bonferroni method for multiple testing correction) was deemed to be a strong correlation.

Results

Transcripts of splicing factors having constant associations with survival outcomes

Of the 985 transcripts investigated, the majority satisfied the PH assumption in multivariable analyses (n = 903 for OS, 895 for DSS, 947 for DFS, and 884 for PFS). Among these transcripts, none reached the significance threshold of 5.08×10^{-05}.

The top 10 transcripts are shown in **Table 2**.

Transcripts of splicing factors having time-varying associations with survival outcomes

Transcripts that violated the PH assumption were further identified for their best-cutoff time points and then fit to the multivariable

Table 2. The top transcripts of splicing factors in survival analyses using Cox PH model

Outcome	Transcripts	Gene	HR (95% CI)	*P*-value	*P*-value of the PH assumption test
OS	ENST00000487160.1	*PRPF38A*	0.75	1.44×10^{-03}	0.31
	ENST00000509810.5	*TCERG1*	0.72	2.52×10^{-03}	0.87
	ENST00000562646.5	*RBMX*	0.88	2.54×10^{-03}	0.42
	ENST00000465245.1	*TRA2B*	0.77	3.54×10^{-03}	0.45
	ENST00000519349.5	*SLU7*	0.77	4.68×10^{-03}	0.32
	ENST00000555566.1	*ACIN1*	0.81	7.08×10^{-03}	0.94
	ENST00000472237.5	*CTNNBL1*	1.12	9.33×10^{-03}	0.18
	ENST00000493598.6	*U2SURP*	0.79	9.77×10^{-03}	0.47
	ENST00000566095.6	*SF3B3*	0.76	1.02×10^{-02}	0.48
	ENST00000505969.1	*THOC3*	1.10	1.11×10^{-02}	0.33
DSS	ENST00000495868.1	*MAGOH*	0.64	6.05×10^{-03}	0.65
	ENST00000509810.5	*TCERG1*	0.65	1.03×10^{-02}	0.24
	ENST00000613191.4	*PRPF18*	0.63	1.03×10^{-02}	0.22
	ENST00000587044.1	*DHX8*	0.64	1.14×10^{-02}	0.92
	ENST00000583741.1	*SRSF1*	0.83	1.28×10^{-02}	0.33
	ENST00000527197.5	*PUF60*	0.78	1.39×10^{-02}	0.86
	ENST00000496549.5	*CRNKL1*	1.28	1.44×10^{-02}	0.09
	ENST00000428425.1	*NCBP2*	1.92	1.66×10^{-02}	0.84
	ENST00000427214.5	*DDX39B*	0.85	1.66×10^{-02}	0.76
	ENST00000617260.4	*PRPF18*	1.58	1.74×10^{-02}	0.91
DFS	ENST00000469522.1	*SNU13*	2.53	4.54×10^{-03}	0.40
	ENST00000468923.5	*NCBP2*	2.33	6.61×10^{-03}	0.21
	ENST00000628103.2	*CTNNBL1*	1.66	7.72×10^{-03}	0.38
	ENST00000471497.5	*ISY1*	0.77	1.16×10^{-02}	0.23
	ENST00000524590.5	*HSPA8*	2.37	1.20×10^{-02}	0.34
	ENST00000465906.5	*RBM17*	2.57	1.24×10^{-02}	0.28
	ENST00000555547.5	*SRSF5*	1.94	1.25×10^{-02}	0.26
	ENST00000560383.5	*SNRPA1*	2.15	1.32×10^{-02}	0.57
	ENST00000497370.5	*PPIE*	1.52	1.35×10^{-02}	0.24
	ENST00000581497.1	*SRSF1*	2.08	1.38×10^{-02}	0.24
PFS	ENST00000476231.1	*SYF2*	0.71	5.00×10^{-03}	0.60
	ENST00000553182.5	*DDX23*	0.83	8.22×10^{-03}	0.08
	ENST00000247140.8	*PQBP1*	0.83	9.23×10^{-03}	0.06
	ENST00000622389.4	*PRPF18*	0.76	9.38×10^{-03}	1.00
	ENST00000402849.5	*SNRPD3*	0.91	1.50×10^{-02}	0.26
	ENST00000368932.5	*CDC40*	0.91	1.97×10^{-02}	0.70
	ENST00000464604.1	*THOC2*	1.15	2.02×10^{-02}	0.44
	ENST00000526110.5	*HSPA8*	0.91	2.23×10^{-02}	0.85
	ENST00000530251.1	*SART1*	0.60	2.26×10^{-02}	0.93
	ENST00000560496.5	*SNRPA1*	1.21	2.51×10^{-02}	0.60

CI, confidence interval; DFS, disease-free survival; DSS, disease-specific survival; OS, overall survival; PFS, progression-free survival; PH, proportional hazards. Age at diagnosis, sex, disease stage, disease location, microsatellite instability status, and race were included in the models adjusting the effects (i.e., HRs) of examined transcripts.

models. The results of the multivariable analyses showed that *PRPF38A* short-transcript ENST00000474048.1 (**Figure 1**) was significantly associated with PFS after two years

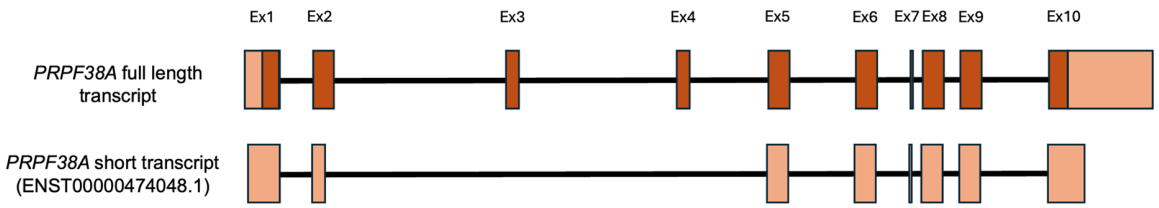

Figure 1. Schematic diagram of the *PRPF38A* short transcript ENST00000474048.1. *PRPF38A* short transcript ENST00000474048.1 is a non-coding transcript. Compared to the full-length transcript, the *PRPF38A*-short transcript skipped exons 3 and 4, had a different 5' splicing site for exons 1, and altered 3' splicing sites for exons 2 and 10. Boxes with dark ochre represent the coding regions, and those with light ochre are non-coding regions.

Table 3. The top transcripts of splicing factors in survival analyses using Cox model with time-varying coefficients

Outcome	Transcripts	Gene	Cutoff time point T (years)	Before or after T	HR (95% CI)	*P*-value	*P*-value of the PH assumption test
OS	ENST00000557515.5	*ACIN1*	4	Before	1.31 (0.91, 1.87)	1.46×10^{-01}	0.94
				After	0.31 (0.16, 0.58)	2.50×10^{-04}	0.42
	ENST00000428450.5	*DDX39B*	2	Before	1.23 (0.87, 1.76)	2.44×10^{-01}	0.66
				After	0.56 (0.40, 0.79)	8.51×10^{-04}	0.77
	ENST00000394082.8	*SNRPA1*	4	Before	1.55 (1.20, 2.01)	9.60×10^{-04}	0.85
				After	0.66 (0.49, 0.89)	6.86×10^{-03}	0.67
	ENST00000431461.1	*CDC40*	2	Before	0.73 (0.46, 1.16)	1.83×10^{-01}	0.41
				After	2.05 (1.33, 3.16)	1.24×10^{-03}	0.17
	ENST00000557201.5	*HNRNPC*	4	Before	1.02 (0.89, 1.17)	7.28×10^{-01}	0.14
				After	0.76 (0.64, 0.90)	1.59×10^{-03}	0.48
	ENST00000502781.5	*LSM6*	1	Before	2.17 (1.33, 3.56)	2.04×10^{-03}	0.48
				After	0.97 (0.72, 1.31)	8.40×10^{-01}	0.49
	ENST00000581135.5	*DDX42*	2	Before	1.20 (0.59, 2.43)	6.09×10^{-01}	0.11
				After	5.00 (1.78, 14.03)	2.26×10^{-03}	0.14
	ENST00000590105.1	*EFTUD2*	4	Before	1.05 (0.60, 1.87)	8.54×10^{-01}	0.55
				After	5.47 (1.79, 16.74)	2.88×10^{-03}	0.77
	ENST00000587196.2	*U2AF2*	4	Before	1.17 (0.89, 1.53)	2.60×10^{-01}	0.77
				After	0.32 (0.15, 0.68)	3.27×10^{-03}	0.56
	ENST00000507392.5	*DDX46*	4	Before	0.92 (0.76, 1.11)	3.71×10^{-01}	0.54
				After	0.52 (0.33, 0.80)	3.49×10^{-03}	0.70
DSS	ENST00000558059.5	*SNRPA1*	2	Before	0.75 (0.34, 1.63)	4.64×10^{-01}	0.45
				After	3.16 (1.71, 5.82)	2.25×10^{-04}	0.68
	ENST00000585392.2	*SNRPD2*	4	Before	1.39 (1.00, 1.91)	4.67×10^{-02}	0.68
				After	0.60 (0.43, 0.82)	1.61×10^{-03}	0.77
	ENST00000483853.1	*FUS*	4	Before	1.03 (0.69, 1.55)	8.84×10^{-01}	0.71
				After	0.28 (0.12, 0.62)	1.78×10^{-03}	0.90
	ENST00000584010.5	*DDX42*	5	Before	0.72 (0.51, 1.01)	5.42×10^{-02}	0.72
				After	0.27 (0.12, 0.63)	2.13×10^{-03}	0.42
	ENST00000486941.1	*SNRNP40*	4	Before	1.35 (1.04, 1.76)	2.68×10^{-02}	0.76
				After	0.35 (0.18, 0.70)	2.97×10^{-03}	0.47
	ENST00000577429.5	*THOC1*	4	Before	1.39 (0.99, 1.95)	5.53×10^{-02}	0.64
				After	0.32 (0.14, 0.69)	3.68×10^{-03}	0.82
	ENST00000579618.1	*SNRPD1*	1	Before	2.79 (1.38, 5.64)	4.45×10^{-03}	0.62
				After	0.80 (0.41, 1.55)	5.06×10^{-01}	0.45
	ENST00000504279.1	*DHX15*	4	Before	1.03 (0.84, 1.25)	7.98×10^{-01}	0.07
				After	0.28 (0.12, 0.67)	4.45×10^{-03}	0.21

	ENST00000556897.5	*HNRNPC*	5	Before	1.68 (1.18, 2.41)	4.47×10^{-03}	0.48
				After	0.34 (0.09, 1.33)	1.21×10^{-01}	0.75
	ENST00000491445.1	*NCBP1*	5	Before	1.67 (1.17, 2.36)	4.51×10^{-03}	0.30
				After	0.37 (0.06, 2.12)	2.64×10^{-01}	0.60
DFS	ENST00000547764.1	*PRPF40B*	2	Before	1.98 (1.25, 3.13)	3.49×10^{-03}	0.49
				After	0.68 (0.26, 1.83)	4.47×10^{-01}	0.91
	ENST00000367208.1	*SNRPE*	1	Before	0.56 (0.36, 0.87)	1.04×10^{-02}	0.92
				After	1.07 (0.79, 1.45)	6.59×10^{-01}	0.41
	ENST00000581576.1	*THOC1*	1	Before	0.20 (0.06, 0.70)	1.22×10^{-02}	0.23
				After	1.33 (0.70, 2.53)	3.81×10^{-01}	0.15
	ENST00000498663.5	*RBM8A*	2	Before	1.09 (0.58, 2.03)	7.91×10^{-01}	0.47
				After	5.48 (1.45, 20.74)	1.23×10^{-02}	0.63
	ENST00000464992.6	*THOC2*	1	Before	0.69 (0.30, 1.58)	3.81×10^{-01}	0.21
				After	3.64 (1.30, 10.18)	1.37×10^{-02}	0.45
	ENST00000496762.1	*RBM17*	3	Before	1.17 (0.63, 2.17)	6.18×10^{-01}	0.86
				After	13.60 (1.68, 109.96)	1.44×10^{-02}	0.58
	ENST00000493864.1	*TRA2B*	3	Before	0.29 (0.02, 3.55)	3.35×10^{-01}	0.70
				After	4.80 (1.33, 17.39)	1.68×10^{-02}	0.26
	ENST00000478050.1	*SF3B6*	2	Before	3.08 (1.20, 7.93)	1.94×10^{-02}	0.24
				After	0.77 (0.31, 1.89)	5.63×10^{-01}	0.84
	ENST00000533421.2	*SF3B2*	2	Before	3.13 (1.20, 8.18)	2.00×10^{-02}	0.75
				After	0.82 (0.41, 1.61)	5.58×10^{-01}	0.98
	ENST00000428425.1	*NCBP2*	4	Before	3.84 (1.23, 11.98)	2.05×10^{-02}	0.98
				After	0.04 (0.02, 7.70)	2.30×10^{-01}	0.41
PFS	ENST00000474048.1	*PRPF38A*	2	Before	1.07 (0.76, 1.53)	6.88×10^{-01}	0.40
				After	0.31 (0.18, 0.54)	3.69×10^{-05}	0.43
	ENST00000583741.1	*SRSF1*	1	Before	1.04 (0.89, 1.23)	6.08×10^{-01}	0.34
				After	0.80 (0.71, 0.90)	2.27×10^{-04}	0.89
	ENST00000559767.1	*AQR*	2	Before	1.36 (1.04, 1.78)	2.58×10^{-02}	1.00
				After	0.52 (0.36, 0.74)	3.47×10^{-04}	0.65
	ENST00000450554.6	*U2AF2*	2	Before	1.08 (0.70, 1.65)	7.33×10^{-01}	0.60
				After	0.36 (0.20, 0.64)	5.59×10^{-04}	0.58
	ENST00000419020.1	*DDX39B*	2	Before	1.12 (0.85, 1.45)	4.32×10^{-01}	0.96
				After	0.54 (0.37, 0.79)	1.38×10^{-03}	0.83
	ENST00000376281.8	*HNRNPK*	2	Before	1.12 (0.72, 1.76)	6.13×10^{-01}	0.83
				After	0.44 (0.26, 0.74)	2.11×10^{-03}	0.63
	ENST00000565990.2	*SF3B3*	2	Before	1.16 (0.82, 1.64)	4.03×10^{-01}	0.54
				After	0.47 (0.29, 0.77)	2.30×10^{-03}	0.78
	ENST00000432305.6	*THOC3*	3	Before	0.83 (0.67, 1.02)	7.15×10^{-02}	0.69
				After	0.42 (0.24, 0.73)	2.35×10^{-03}	0.32
	ENST00000320676.11	*RBMX*	2	Before	0.97 (0.66, 1.43)	8.86×10^{-01}	0.43
				After	0.36 (0.18, 0.70)	2.52×10^{-03}	0.55
	ENST00000330752.12	*HNRNPA1*	2	Before	1.07 (0.79, 1.45)	6.72×10^{-01}	0.45
				After	0.55 (0.36, 0.82)	3.35×10^{-03}	0.49

CI, confidence interval; DFS, disease-free survival; DSS, disease-specific survival; OS, overall survival; PFS, progression-free survival; PH, proportional hazards. Age at diagnosis, sex, disease stage, disease location, microsatellite instability status, and race were included in the models adjusting the effects (i.e., HRs) of examined transcripts.

post-diagnosis (HR [95% CI] = 0.31 [0.18, 0.54], $P = 3.69 \times 10^{-05}$), but not before that, indicating that this transcript is a candidate transcript of splicing factor with time-varying associations (**Table 3**). Patients with high expression levels of this transcript had longer progression-free survival (or better outcomes) than those with low expression levels after two

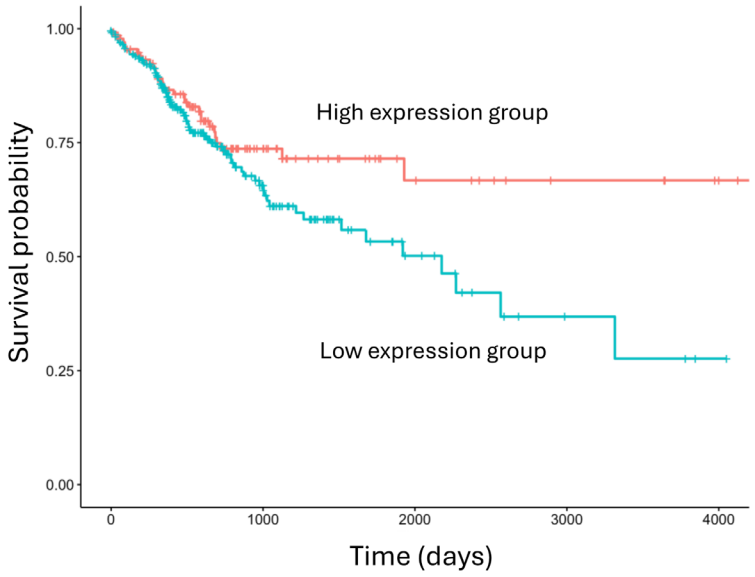

Figure 2. Kaplan-Meier curves of the *PRPF38A* short-transcript on progression-free survival. Patients with high expression levels of this transcript had higher survival probabilities (in other words, better outcomes) after around two years post-diagnosis, but not before that. This is consistent with the time-varying association analysis results of the *PRPF38A* short-transcript showing that this transcript was significantly associated with PFS only after two years post-diagnosis, during which time patients with a higher expression level of the transcript had better outcomes (HR = 0.31).

years post-diagnosis (**Table 3** and **Figure 2**). Within the first two years post-diagnosis, however, patients with different expression levels of ENST00000474048.1 had no difference in PFS (**Table 3** and **Figure 2**).

Other transcripts were not significantly associated with survival outcomes. The top transcripts are listed in **Table 3**.

Gene expression levels of PRPF38A and progression-free survival in CRC

The gene expression levels of *PRPF38A* were further examined for their associations with PFS in CRC, and the results showed that there was a nominal association between expression levels of *PRPF38A* and PFS (P = 0.086). This suggests that the *PRPF38A* short-transcript expression, rather than *PRPF38A* gene expression, is a candidate marker for predicting the PFS in CRC.

Correlation between PRPF38A short-transcript levels and transcript ratios of other genes

To further explore the possible mechanisms of the *PRPF38A* short-transcript on survival outcomes, we performed a correlation analysis between the *PRPF38A* short-transcript and transcript ratios of other genes. Results showed that the transcript ratios of RNA binding motif protein 28 (*RBM28*; encodes a RNA-binding protein) and mitochondrial ribosomal protein S30 (*MRPS30*; encodes a mitochondrial protein synthesis protein) had a strong negative correlation (Pearson correlation coefficient ≤ -0.5) with the levels of *PRPF38A* short-transcript (**Tables 4**, **5**), implying that the *PRPF38A* short-transcript may convey its effects on survival outcomes through the regulation of splicing and transcript levels of these genes.

The top 10 positively and negatively correlated transcripts are shown in **Table 4**.

Discussion

In the current study, we examined the transcripts of splicing factors with and without time-varying associations with survival outcomes (OS, DSS, DFS, and PFS) in CRC. This is the first study to investigate the prognostic value of splicing factor transcripts in this disease, providing a comprehensive view of the relationships between splicing factor transcripts and patient outcomes. This is also the first study to investigate the time-varying associations of transcripts and their potential as early- or late-outcome markers in CRC. Though most of the transcripts had no significant associations with survival outcomes, interestingly, a short transcript (i.e., ENST00000474048.1) of *PRPF38A* was identified to be significantly associated with PFS after two years post-diagnosis, but not before that, suggesting a potential late effect of this splicing factor transcript on patient outcomes. Further correlation analyses implied that this transcript may influence RNA binding (*RBM28*) and mitochondrial protein synthesis (*MRPS30*), which are important biological processes in cell survival, proliferation, and migration. The findings of this study

Table 4. The top positively and negatively correlated transcripts with the *PRPF38A* short-transcript

Correlation type	Correlated transcript	Gene	Pearson correlation coefficient r	*P* value
Positively correlated	ENST00000336733.10	*SLC25A26*	0.47	1.41×10^{-21}
	ENST00000376547.7	*STK24*	0.44	5.49×10^{-19}
	ENST00000410049.1	*C2orf49*	0.43	1.45×10^{-18}
	ENST00000407439.7	*MRE11*	0.43	1.54×10^{-18}
	ENST00000488746.1	*SELENOK*	0.43	1.91×10^{-18}
	ENST00000541489.5	*DDX51*	0.43	2.46×10^{-18}
	ENST00000479408.6	*PDE4DIP*	0.43	2.68×10^{-18}
	ENST00000538833.6	*DUS1L*	0.43	3.31×10^{-18}
	ENST00000491736.1	*SLC25A4*	0.43	3.60×10^{-18}
	ENST00000620209.4	*DDX52*	0.43	3.82×10^{-18}
Negatively correlated	ENST00000223073.6	*RBM28*	-0.53	3.98×10^{-28}
	ENST00000230914.4	*MRPS30*	-0.50	3.35×10^{-25}
	ENST00000317623.8	*PCNX4*	-0.49	1.29×10^{-24}
	ENST00000505587.5	*CCNH*	-0.49	7.32×10^{-24}
	ENST00000479950.5	*OARD1*	-0.48	1.21×10^{-22}
	ENST00000479950.5	*GLRX3*	-0.47	3.86×10^{-22}
	ENST00000368644.5	*ANKRD28*	-0.47	5.23×10^{-22}
	ENST00000461696.1	*CCNB2*	-0.47	1.29×10^{-21}
	ENST00000621385.1	*PTBP2*	-0.46	1.58×10^{-21}
	ENST00000609116.5	*PCBD2*	-0.46	1.21×10^{-20}

Table 5. Expression levels of *PRPF38A* short-transcript and transcript ratios of *RBM28* and *MRPS30*

PRPF38A short transcript	Mean of *RBM28* transcript ratio (%) (ENST00000223073.6/all transcripts)	Mean of *MRPS30* transcript ratio (%) (ENST00000230914.4/all transcripts)
High expression	6.47	13.66
Low expression	8.70	16.55

are novel and can help guide patient stratification/disease management, thereby contributing to precision medicine. Mechanistically, this study provides insights into the prognostic role of the *PRPF38A* short transcript in CRC.

The short transcript ENST00000474048.1 (1059 nt in length) is a lncRNA transcribed from *PRPF38A*, a gene encoding a component of the U4/U6-U5 tri-small nuclear ribonucleoprotein (snRNP) of spliceosome complex B [37]. Although this transcript has not been examined in previous studies, the gene *PRPF38A* was reported to affect disease prognosis. It was reported that the downregulation of *PRPF38A* in human breast cancer cell lines leads to intron retention for genes that participate in homeostasis, mitosis, and apoptosis [38]. Studies on osteosarcoma (preprint in Research Square; https://doi.org/10.21203/rs.3.rs-24106/v1) and recurrent liver cancer following surgical treatment or liver transplantation have shown that *PRPF38A* is a candidate biomarker or signature for these diseases [39]. The knockdown of *PRPF38A* has also been reported to influence osteogenesis [40]. Although the effects of *PRPF38A* on prognosis have not been reported in CRC, this gene is likely to affect patient outcomes given its regulatory role in splicing, which is a critical biological process affecting phenotypes, including disease progression. Regarding the short transcript of *PRPF38A*, it is possible that this transcript can also affect the prognosis of patients with CRC, considering that this transcript is a lncRNA of the splicing factor PRPF38A; lncRNAs have been reported to affect mRNAs transcribed from the same genes [41, 42] and may also directly interact with splicing factors [34-36]. Supported by these findings, the identifica-

tion of an association between the *PRPF38A* short transcript and patient survival in this study implies a potential role for this transcript in CRC progression. Whether this regulatory role is mediated through *PRPF38A* or not can be investigated in future studies.

Two genes with transcript ratios correlated with the short transcript of *PRPF38A* play important roles in cells and affect disease progression. *RBM28* encodes an RNA-binding protein that is a nucleolar component of spliceosome small nuclear ribonucleoproteins (snRNPs) [43] as well as a component involved in ribosome biogenesis [44]. Studies have shown that *RBM28* is upregulated in different types of cancer (including colon cancer) and is associated with poor patient outcomes [45]. In line with this, another study in 2024 reported that *RBM28* was overexpressed in liver cancer and enhanced tumor angiogenesis [46]. *RBM28* has also been reported to be associated with splicing alterations in cancers [47] and can interact with p53 to affect cell proliferation and metastasis [45, 48]. Although few studies have investigated the transcript ENST00000223073.6 of *RBM28*, this transcript is the MANE select transcript and encodes the full-length (and likely to represent a complete function) protein of *RBM28*; thus, the altered ratio of this transcript may impact patient outcomes. *MRPS30* encodes the mitochondrial ribosomal small subunit 30 [49]; thus, it is pivotal for the synthesis of mitochondrial proteins, which are essential for oxidative phosphorylation. Although little is known about the role of *MRPS30* transcript ENST00000230914.4, this transcript may influence the production of cell energy, given that MRPS30 is a key component of the mitochondrial ribosome. Future studies should be performed on the detailed roles of the *RBM28* and the *MRPS30* transcripts, as well as their mechanistic relationships with the *PRPF38A* short transcript in the prognosis of CRC.

This study has implications for the missing heritability of CRC prognosis. Missing heritability refers to the phenomenon in which identified genetic factors with prognostic associations explain only a portion of the expected genetic variance, leaving the rest of the variance unexplained. Large-scale studies on genetic variations (e.g., genome-wide association studies) may identify the genetic factors that explain the remaining variance [50, 51]. However, after years of work on such studies, heritability remains missing. Investigation of gene transcripts provided an opportunity to explain the remaining variance. The identification of a transcript with a prognostic association in this study suggests that transcripts, as an understudied layer of genetic architecture, can be a promising source of variation explaining missing heritability. Additionally, considering that the short transcript of *PRPF38A* can only be identified as a factor with time-varying associations (not as a factor with constant associations), examining factors for their time-varying associations can further contribute to the identification of prognosis-associated factors and thus help account for missing heritability. Our findings may inspire future studies to identify novel prognostic factors explaining missing heritability.

The method used to find time-varying associations in this study relies on checking the PH assumption of the Cox PH model (violation of this assumption implies a time-varying association). Checking the PH assumption of the Cox PH method is not a common practice in medical research [52-55], which may lead to bias in findings and even miss important associations (e.g., time-varying associations). By taking advantage of the violation of the PH assumption, we used appropriate models to identify non-constant associations over time. Thus, the results obtained in this study are more reliable than those of many other studies that have never checked/reported the PH assumption test results or simply assumed constant associations. Checking the PH assumption in the analyses using the Cox PH models increased our confidence in the findings of the study. Regarding the discovery of hidden associations, the identification of the late association transcript (the *PRPF38A* short transcript) in this study is a great example, showing that non-constant associations would be missed if the PH assumption were not checked ($P = 0.129$ if analyzed using the Cox PH model). This further means that the early- or late-outcome markers would not be identified if the PH assumption was not checked and the Cox PH model was arbitrarily used for all analyses, highlighting the

importance of checking the assumption and using appropriate models in survival analysis.

The strengths of this study include the examination of an uncommonly investigated type of variation, the transcripts, and their associations (including time-varying associations) with survival outcomes in CRC. The transcripts were RSEM-normalized data generated using the same pipeline, which increased the confidence in the study findings. We used appropriate models after checking the PH assumption for survival data analysis, which further increased the reliability of the results. This study also has some limitations. It examined the transcripts of splicing factors, leaving the majority of transcripts in cells to be investigated in future studies. The findings of this study need to be validated in other studies using larger cohorts. The small sample size (n = 376) in the current study may limit our capability to identify transcripts with small effect sizes, and studies with larger study powers can be performed to identify additional transcripts with and without time-varying associations.

In conclusion, this study presented a comprehensive view of the association patterns of splicing factor transcripts with survival outcomes in CRC. A short transcript of *PRPF38A* with a time-varying association with patient survival was also identified. Such an association, if validated in independent patient cohorts, can be a late-outcome marker to stratify patients into different outcome risk groups after two years post-diagnosis, which further helps with disease treatment and follow-up management in CRC.

Acknowledgements

We gratefully acknowledge that the results of this study are based on TCGA data generated by or stored at the TCGA Research Network (https://www.cancer.gov/tcga) and the UCSC Xena (https://xena.ucsc.edu/). This work was supported by National Natural Science Foundation of China (U23A20398, 82030007), Sichuan Science and Technology Program (2023NSFSC1894, 2022YFS0614), Southwest Medical University Technology Program (2023ZD006), and Southwest Medical University Undergraduate Training Programs for Innovation and Entrepreneurship (2023411), and we would like to thank funders that supported this study. We also thank Yuhao Luo M.D. for his help in determining and reviewing the treatment regimen information in the study dataset, and Xiong Li M.D. for his help in describing the outcome measures in section 2.2.

Disclosure of conflict of interest

None.

Address correspondence to: Yan Deng and Chunxiang Zhang, Southwest Medical University, 1 Xianglin Road, Luzhou 646000, Sichuan, China. E-mail: dengyan3691@swmu.edu.cn (YD); zhangchunxiang@swmu.edu.cn (CXZ)

References

[1] Klimeck L, Heisser T, Hoffmeister M and Brenner H. Colorectal cancer: a health and economic problem. Best Pract Res Clin Gastroenterol 2023; 66: 101839.

[2] Nors J, Iversen LH, Erichsen R, Gotschalck KA and Andersen CL. Incidence of recurrence and time to recurrence in stage I to III colorectal cancer: a nationwide Danish cohort study. JAMA Oncol 2024; 10: 54-62.

[3] Wang R, Su Q and Yan ZP. Reconsideration of recurrence and metastasis in colorectal cancer. World J Clin Cases 2021; 9: 6964-6968.

[4] Werdyani S, Yu Y, Skardasi G, Xu J, Shestopaloff K, Xu W, Dicks E, Green J, Parfrey P, Yilmaz YE and Savas S. Germline INDELs and CNVs in a cohort of colorectal cancer patients: their characteristics, associations with relapse-free survival time, and potential time-varying effects on the risk of relapse. Cancer Med 2017; 6: 1220-1232.

[5] Yu Y, Carey M, Pollett W, Green J, Dicks E, Parfrey P, Yilmaz YE and Savas S. The long-term survival characteristics of a cohort of colorectal cancer patients and baseline variables associated with survival outcomes with or without time-varying effects. BMC Med 2019; 17: 150.

[6] Yu Y, Werdyani S, Carey M, Parfrey P, Yilmaz YE and Savas S. A comprehensive analysis of SNPs and CNVs identifies novel markers associated with disease outcomes in colorectal cancer. Mol Oncol 2021; 15: 3329-3347.

[7] Pavelitz T, Renfro L, Foster NR, Caracol A, Welsch P, Lao VV, Grady WB, Niedzwiecki D, Saltz LB, Bertagnolli MM, Goldberg RM, Rabinovitch PS, Emond M, Monnat RJ Jr and Maizels N.

MRE11-deficiency associated with improved long-term disease free survival and overall survival in a subset of stage III colon cancer patients in randomized CALGB 89803 trial. PLoS One 2014; 9: e108483.
[8] Penney ME, Parfrey PS, Savas S and Yilmaz YE. A genome-wide association study identifies single nucleotide polymorphisms associated with time-to-metastasis in colorectal cancer. BMC Cancer 2019; 19: 133.
[9] Roncucci L, Fante R, Losi L, Di Gregorio C, Micheli A, Benatti P, Madenis N, Ganazzi D, Cassinadri MT, Lauriola P and Ponz de Leon M. Survival for colon and rectal cancer in a population-based cancer registry. Eur J Cancer 1996; 32A: 295-302.
[10] Bolard P, Quantin C, Esteve J, Faivre J and Abrahamowicz M. Modelling time-dependent hazard ratios in relative survival: application to colon cancer. J Clin Epidemiol 2001; 54: 986-996.
[11] Giorgi R, Abrahamowicz M, Quantin C, Bolard P, Esteve J, Gouvernet J and Faivre J. A relative survival regression model using B-spline functions to model non-proportional hazards. Stat Med 2003; 22: 2767-2784.
[12] Quantin C, Abrahamowicz M, Moreau T, Bartlett G, MacKenzie T, Tazi MA, Lalonde L and Faivre J. Variation over time of the effects of prognostic factors in a population-based study of colon cancer: comparison of statistical models. Am J Epidemiol 1999; 150: 1188-1200.
[13] Zahl PH. Regression analysis with multiplicative and time-varying additive regression coefficients with examples from breast and colon cancer. Stat Med 2003; 22: 1113-1127.
[14] Liu M, Li L, Yu W, Chen J, Xiong W, Chen S and Yu L. Marriage is a dependent risk factor for mortality of colon adenocarcinoma without a time-varying effect. Oncotarget 2017; 8: 20056-20066.
[15] van der Kruijssen DEW, van der Kuil AJS, Vink GR, Punt CJA, de Wilt JHW, Elias SG and Koopman M. Time-varying prognostic value of primary tumor sidedness in metastatic colorectal cancer: a population-based study and meta-analysis. Int J Cancer 2023; 152: 1360-1369.
[16] Roberts EK, Luo L, Mondul AM, Banerjee M, Veenstra CM, Mariotto AB, Schipper MJ, He K, Taylor JMG and Brouwer AF. Time-varying associations of patient and tumor characteristics with cancer survival: an analysis of SEER data across 14 cancer sites, 2004-2017. Cancer Causes Control 2024; 35: 1393-1405.
[17] Syn NL, Chua DW, Raphael Chen L, Tan YC, Goh BKP, Chung Cheow P, Jeyaraj PR, Koh Y, Chung A, Yee Lee S, Lucien Ooi L, Tai BC, Yip Chan C and Teo JY. Time-varying prognostic effects of primary tumor sidedness and grade after curative liver resection for colorectal liver metastases. Surg Oncol 2021; 38: 101586.
[18] Wu Y, Yang L, Zhao J, Li C, Nie J, Liu F, Zhuo C, Zheng Y, Li B, Wang Z and Xu Y. Nuclear-enriched abundant transcript 1 as a diagnostic and prognostic biomarker in colorectal cancer. Mol Cancer 2015; 14: 191.
[19] Yamada Y, Itano N, Narimatsu H, Kudo T, Morozumi K, Hirohashi S, Ochiai A, Ueda M and Kimata K. Elevated transcript level of hyaluronan synthase1 gene correlates with poor prognosis of human colon cancer. Clin Exp Metastasis 2004; 21: 57-63.
[20] Peng J, Ou Q, Guo J, Pan Z, Zhang R, Wu X, Zhao Y, Deng Y, Li C, Wang F, Li L, Chen G, Lu Z, Ding P, Wan D and Fang Y. Expression of a novel CNPY2 isoform in colorectal cancer and its association with oncologic prognosis. Aging (Albany NY) 2017; 9: 2334-2351.
[21] Wang C, Weng M, Xia S, Zhang M, Chen C, Tang J, Huang D, Yu H, Sun W, Zhang H and Lai M. Distinct roles of programmed death ligand 1 alternative splicing isoforms in colorectal cancer. Cancer Sci 2021; 112: 178-193.
[22] Sun C, Zeng B, Zhou J, Li N, Li M, Zhu C, Xie S, Wang Y, Wang S and Wang X. Analysis of SLC genes alternative splicing identifies the SLC7A6 RI isoform as a therapeutic target for colorectal cancer. Cancer Sci 2025; 116: 233-247.
[23] Tholen J and Galej WP. Structural studies of the spliceosome: Bridging the gaps. Curr Opin Struct Biol 2022; 77: 102461.
[24] Yang H, Beutler B and Zhang D. Emerging roles of spliceosome in cancer and immunity. Protein Cell 2022; 13: 559-579.
[25] Pan Q, Shai O, Lee LJ, Frey BJ and Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 2008; 40: 1413-1415.
[26] Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP and Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature 2008; 456: 470-476.
[27] Eilertsen IA, Sveen A, Stromme JM, Skotheim RI, Nesbakken A and Lothe RA. Alternative splicing expands the prognostic impact of KRAS in microsatellite stable primary colorectal cancer. Int J Cancer 2019; 144: 841-847.
[28] Cieply B and Carstens RP. Functional roles of alternative splicing factors in human disease. Wiley Interdiscip Rev RNA 2015; 6: 311-326.
[29] Yoshida K and Ogawa S. Splicing factor mutations and cancer. Wiley Interdiscip Rev RNA 2014; 5: 445-459.
[30] Zhang Z, Reinikainen J, Adeleke KA, Pieterse ME and Groothuis-Oudshoorn CGM. Time-vary-

ing covariates and coefficients in Cox regression models. Ann Transl Med 2018; 6: 121.
[31] Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, Omberg L, Wolf DM, Shriver CD and Thorsson V; Cancer Genome Atlas Research Network; Hu H. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell 2018; 173: 400-416, e11.
[32] Kilickap S, Demirci U, Karadurmus N, Dogan M, Akinci B and Sendur MAN. Endpoints in oncology clinical trials. J BUON 2018; 23: 1-6.
[33] Team RC. R: A language and environment for statistical computing. 2013.
[34] Chen J, Liu Y, Min J, Wang H, Li F, Xu C, Gong A and Xu M. Alternative splicing of lncRNAs in human diseases. Am J Cancer Res 2021; 11: 624-639.
[35] Liu Y, Liu X, Lin C, Jia X, Zhu H, Song J and Zhang Y. Noncoding RNAs regulate alternative splicing in Cancer. J Exp Clin Cancer Res 2021; 40: 11.
[36] Statello L, Guo CJ, Chen LL and Huarte M. Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol 2021; 22: 96-118.
[37] Fu X and Hoskins AA. Dynamics and evolutionary conservation of B complex protein recruitment during spliceosome activation. Nucleic Acids Res 2025; 53: gkaf124.
[38] Chan S, Sridhar P, Kirchner R, Lock YJ, Herbert Z, Buonamici S, Smith P, Lieberman J and Petrocca F. Basal-A triple-negative breast cancer cells selectively rely on RNA splicing for survival. Mol Cancer Ther 2017; 16: 2849-2861.
[39] Das T, Diamond DL, Yeh M, Hassan S, Bryan JT, Reyes JD and Perkins JD. Molecular signatures of recurrent hepatocellular carcinoma secondary to hepatitis C virus following liver transplantation. J Transplant 2013; 2013: 878297.
[40] Cousminer DL, Wagley Y, Pippin JA, Elhakeem A, Way GP, Pahl MC, McCormack SE, Chesi A, Mitchell JA, Kindler JM, Baird D, Hartley A, Howe L, Kalkwarf HJ, Lappe JM, Lu S, Leonard ME, Johnson ME, Hakonarson H, Gilsanz V, Shepherd JA, Oberfield SE, Greene CS, Kelly A, Lawlor DA, Voight BF, Wells AD, Zemel BS, Hankenson KD and Grant SFA. Genome-wide association study implicates novel loci and reveals candidate effector genes for longitudinal pediatric bone accrual. Genome Biol 2021; 22: 1.
[41] Shao M, Yang Q, Zhu W, Jin H, Wang J, Song J, Kong Y and Lv X. LncHOXA10 drives liver TICs self-renewal and tumorigenesis via HOXA10 transcription activation. Mol Cancer 2018; 17: 173.
[42] Williamson L, Saponaro M, Boeing S, East P, Mitter R, Kantidakis T, Kelly GP, Lobley A, Walker J, Spencer-Dene B, Howell M, Stewart A and Svejstrup JQ. UV irradiation induces a non-coding RNA that functionally opposes the protein encoded by the same gene. Cell 2017; 168: 843-855, e13.
[43] Damianov A, Kann M, Lane WS and Bindereif A. Human RBM28 protein is a specific nucleolar component of the spliceosomal snRNPs. Biol Chem 2006; 387: 1455-1460.
[44] Bryant CJ, Lorea CF, de Almeida HL Jr, Weinert L, Vedolin L, Pinto E Vairo F and Baserga SJ. Biallelic splicing variants in the nucleolar 60S assembly factor RBM28 cause the ribosomopathy ANE syndrome. Proc Natl Acad Sci U S A 2021; 118: e2017777118.
[45] Lin X, Zhou L, Zhong J, Zhong L, Zhang R, Kang T and Wu Y. RNA-binding protein RBM28 can translocate from the nucleolus to the nucleoplasm to inhibit the transcriptional activity of p53. J Biol Chem 2022; 298: 101524.
[46] Han H, Yuan Y, Li C, Liu L, Yu H, Han G, Wang Q, Lin M and Huang J. RNA-binding motif protein 28 enhances angiogenesis by improving STAT3 translation in hepatocellular carcinoma. Cancer Lett 2024; 604: 217191.
[47] Sebestyen E, Singh B, Minana B, Pages A, Mateo F, Pujana MA, Valcarcel J and Eyras E. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res 2016; 26: 732-744.
[48] Xu H, Wang T, Nie H, Sun Q, Jin C, Yang S, Chen Z, Wang X, Tang J, Feng Y and Sun Y. USP36 promotes colorectal cancer progression through inhibition of p53 signaling pathway via stabilizing RBM28. Oncogene 2024; 43: 3442-3455.
[49] Wu H, Zhu X, Zhou H, Sha M, Ye J and Yu H. Mitochondrial ribosomal proteins and cancer. Medicina (Kaunas) 2025; 61: 96.
[50] Mortezaei Z and Tavallaei M. Novel directions in data pre-processing and genome-wide association study (GWAS) methodologies to overcome ongoing challenges. Inform Med Unlocked 2021; 24: 100586.
[51] Matthews LJ and Turkheimer E. Three legs of the missing heritability problem. Stud Hist Philos Sci 2022; 93: 183-191.
[52] Carroll OU, Morris TP and Keogh RH. How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review. BMC Med Res Methodol 2020; 20: 134.
[53] Jachno K, Heritier S and Wolfe R. Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis

of randomised controlled trials? A review of current practice. BMC Med Res Methodol 2019; 19: 103.

[54] Altman DG, De Stavola BL, Love SB and Stepniewska KA. Review of survival analyses published in cancer journals. Br J Cancer 1995; 72: 511-518.

[55] Kuitunen I, Ponkilainen VT, Uimonen MM, Eskelinen A and Reito A. Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review. BMC Musculoskelet Disord 2021; 22: 489.

Supplementary Table 1. Summary of treatment regimens in the study dataset

Treatment regimen	Number of patients	Percentage (%)
FOLFOX	69	18.35
FOLFOX + Bevacizumab	9	2.39
FOLFOX + Cetuximab	3	0.80
FOLFOX + Dexamethasone	2	0.53
FOLFIRI	4	1.06
FOLFIRI + Bevacizumab	1	0.27
XELOX	1	0.27
XELOX + Raltitrexed	1	0.27
Fluorouracil	9	2.39
Fluorouracil + Oxaliplatin	7	1.86
Fluorouracil + Oxaliplatin + Bevacizumab	3	0.80
Fluorouracil + Leucovorin	8	2.13
Fluorouracil + Leucovorin + Etoposide	1	0.27
Fluorouracil + Leucovorin Calcium	3	0.80
Fluorouracil + Irinotecan	1	0.27
Floxuridine	2	0.53
Capecitabine	19	5.05
Capecitabine + Bevacizumab	1	0.27
Oxaliplatin	3	0.80
Oxaliplatin + Leucovorin	2	0.53
Irinotecan + Cetuximab	1	0.27
No treatment	176	46.81
Unknown	50	13.30

FOLFOX, fluorouracil + oxaliplatin + leucovorin; FOLFIRI, fluorouracil + Irinotecan + leucovorin; XELOX, capecitabine + oxaliplatin.

Supplementary Table 2. Summary of main treatment types in the study dataset

Received treatment or not	Chemotherapy		Radiation therapy	
	Number (n)	Percentage (%)	Number (n)	Percentage (%)
Yes	151	40.16	25	6.65
No	176	46.81	302	80.32
Not available/unknown	49	13.03	49	13.03