Identification of novel alternative splicing isoform biomarkers and their association with overall survival in colorectal cancer

CURRENT STATUS: UNDER REVIEW

Hongtao Jia  jiaht@tianjiagenomes.com
Tianjia Genomes Tech CO., LTD.
Corresponding Author

Aili Wang
Binzhou Medical University Hospital

Haifeng Lian
Binzhou Medical University Hospital

Yuanyuan Shen
Binzhou Medical University Hospital

Qian Wang
Tianjia Genomes Tech CO. LTD.

Zhenru Zhou
Tianjia Genomes Tech CO. LTD.

Ranran Zhang
Binzhou Medical University Hospital

Kun Li
Binzhou Medical University Hospital

Chengxia Liu
Binzhou Medical University Hospital

DOI:
10.21203/rs.2.10736/v1

SUBJECT AREAS
Gastroenterology & Hepatology

KEYWORDS
Alternative splicing, colorectal cancer (CRC), metastasis, RNA-seq, TCGA
Abstract

Alternative splicing is an important mechanism of regulating eukaryotic gene expression. Understanding the most common alternative splicing events in colorectal cancer (CRC) will help developing diagnostic, prognostic or therapeutic tools in CRC. Publicly available RNA-seq data of 31 pairs of CRC and normal tissues and 18 pairs of metastatic and normal tissues were used to identify alternative splicing events using PSI and DEXSeq methods. The highly significant splicing events were used to search a database of The Cancer Genome Atlas (TCGA). We identified alternative splicing events in 10 genes marking the signature of CRC (more inclusion of CLK1-E4, COL6A3-E6, CD44v8-10, alternative first exon regulation of ARHGEF9, CHEK1, HKDC1 and HNF4A) or metastasis (decrease of SERPINA1-E1a, CALD-E5b, E6 and FBLN2-E9). Except for CHEK1, all other 9 splicing events were confirmed by TCGA data with 382 CRC tumors and 52 normal controls. Two splicing events (COL6A3 and HKDC1) were found to be significantly associated with patient overall survival. The alternative splicing signatures of the 10 genes are highly consistent with previous reports and/or relevant to cancer biology. The significant association of higher expression of the COL6A3 E5-E6 junction and HKDC1 E1-E2 with better overall survival was firstly reported. This study might be of significant value in the future biomarker, prognosis marker and therapeutics development of CRC.

Background

Colorectal cancer (CRC) is the third and second most common cancer in men and women worldwide [1] and it’s the second and third leading cause of death in men and women in developed countries [1]. A deep understanding of the genes involved in the tumorigenesis of CRC will eventually contribute to developing diagnosis, prognosis and therapeutic methods of CRC.
Alternative splicing is an important mechanism of regulating eukaryotic gene expression. Ninety-two to ninety-four percent of human genes undergo alternative splicing. Different protein isoforms produced by the same gene through alternative splicing may have related, distinct or even opposing functions [2]. Regulation of alternative splicing plays an important role in both normal and disease states of biological processes [3, 4]. In cancer research, defects of alternative splicing or mutation, misregulation of splicing factors were linked to tumorigenesis [5, 6], cancer metastasis [7] and cancer drug resistance [8]. Targeting alternative splicing or targeting splicing factors are new therapeutic approaches to fight against cancer [9, 10].

Aberrant splicing events in CRC and/or other cancer types were identified using either microarray or RNA-seq techniques in the past [11-13]. The roles of alternative splicing in CRC were reported for individual genes: ITGA6 [14], MAP4K4 [15], EPDR1, ZNF518B [16], FIR [17], BRAF [18], Rac1 [19], OCT4 [20], RON [21], CD44 [22, 23]. However, it’s still unclear what the most common alternative splicing events are in large CRC patient populations and what events are related to metastasis. With the publicly available RNA-seq data, we studied 31 pairs of CRC and normal tissues and 18 pairs of metastatic and normal tissues. As results, we identified alternative splicing events in 10 genes marking the signature of CRC or metastasis. Nine splicing events were confirmed by independent TCGA data with 382 CRC tumors and 52 normal controls. Finally, two splicing events were found to be significantly associated with patient overall survival.

Methods

**RNA-seq data analysis**

Raw reads in fastq format of the three GEO datasets (Figure 1, Supplementary table 1) were downloaded from SRA (https://www.ncbi.nlm.nih.gov/sra/) using SRA Toolkit. Reads
were mapped to the human hg19 genome using STAR (2.5.3a) [24]. Only the uniquely mapped reads with MAPQ value >10 and mismatch < 5% of the read length were used. The Sashimi Plot showing the junction read number was generated using Integrative Genomics Viewer (IGV) [25]. The uniquely mapped reads mapping from bam files were converted to bigwig format and visualized in UCSC genome browser (https://genome.ucsc.edu). The height of each coordinate in the genome represents the reads per million (RPM) value of the read coverage at that coordinate.

For splicing analysis of exons, we used Refseq defined exons and counted the reads mapped to them using a custom perl script. The exon read counts table were used as input for the R library DEXSeq [26]. The adjusted P-value and fold change of each exon of a gene were used to select significantly regulated exons.

**Percent-spliced-in (PSI) analysis**

We used three PSI calculation methods: PSI_exon, PSI_junc5’ and PSI_junc3’ as illustrated in Figure 2A, 3A, and 3B, respectively. We required a minimal total read of 5 to calculate a PSI value. The PSI values of two groups of samples (eg. CRC vs. NC) were compared using unpaired two-tailed Student’s t-test. For ΔPSI (the PSI change between two sample groups), the PSI values of each group were averaged and then the difference was calculated. All statistical analysis and plots were performed using R.

**Survival analysis**

All patient overall survival (OS) data and junction usage data were downloaded from TSVdb [27]. All patients with OS data were separated into two groups based on the median value of a particular junction usage. The survival curve was drawn using the R package “survminer” (https://cran.r-project.org/web/packages/survminer/). P-value is calculated based on the log-rank test.

**Results**
Identification of commonly regulated alternative splicing events from RNA-seq data using multiple CRC sample cohorts.

We collected three public RNA-seq data related to CRC from GEO database (Figure 1, Supplementary table 1): CRC18P, CRC10P, and CRC3P, in which paired tumor (CRC) and normal colon tissue (NC or NM) were collected from 18, 10 and 3 patients, respectively. In CRC18P, 18 matched liver metastatic (MC) tissue were also collected. We applied percent-spliced-in (PSI) analysis (see Methods for details) and DEXSeq [26] to identify splicing events, which are different in CRC or MC compared to NC tissues. CRC3P data has relatively lower sequencing depth and higher 3’ bias in two NM samples (Supplementary table 2). Therefore, we focused on the analysis of CRC18P and CRC10P. As results, three exons in three genes (CLK1, COL6A3 and CD44) were identified to show splicing changes in CRC vs. NC in both CRC18P and CRC10P data (Figure 2A-B, Supplementary table 3).

The exon 4 of CLK1 has a median inclusion level (PSI) of about 50% in NC samples and higher inclusion level (median 80%-90%) in CRC samples in CRC18P and CRC10P (Figure 2C-E, supplementary figure 1). In CRC18P, MC samples maintain the higher PSI compared to NCs (Figure 2D). The exon4-skipping isoform, predominantly expressed in NC samples, matches the isoform 3 of CLK1 transcripts from Refseq annotation. Interestingly, isoform 3 is annotated as a non-coding transcript because skipping the 91 nt exon 4 is predicted to cause out-of-frame translation and the nonsense mediated decay (NMD) of the transcript (Figure 2C). The increase of exon4-inclusion isoform (isoform 1) at the expense of isoform 3 in CRC and MC may thus increase the productive transcript level of CLK1 and potentially produce more proteins. CLK1 encodes a member of the CDC2-like family of protein kinases (CLKs). In the cell nucleus, CLKs phosphorylate serine/arginine-rich proteins, release them into the nucleoplasm and then regulate alternative splicing of genes. Small molecule inhibitors against these CLKs were developed and exhibited growth suppression and
apoptosis induction effect [28]. Here, we found that CLK1 itself can be alternatively spliced. CRC cells may increase the CLK1 expression by regulating the inclusion of exon 4. If the function of this splicing event can be validated by further evidence, exon 4 skipping could be a new target for cancer therapy.

The exon 6 of COL6A3 has a median inclusion level (PSI) of about 10% in NC samples and higher inclusion level (median 40%-50%) in CRC samples in CRC18P and CRC10P (Figure 2F-H, supplementary figure 2). In CRC18P, MC samples maintain the higher PSIs as CRCs compared to NCs (Figure 2G). This splicing change has been reported previously in colon, bladder, prostate and pancreatic cancers tissues compared to normal tissues [11, 29], indicating that the splicing change may play a role in multiple cancer types. High expression of COL6A3 in stroma were associated with poor prognosis in CRC [30]. COL6A3 was also found to be a key hub gene in the cell migration/extracellular matrix module that was associated with poor prognosis in CRC [31]. COL6A3 knockout decreases cell proliferation and invasion, increases cell apoptosis in cancer cell lines [31]. Taken together, the relevance and importance of COL6A3 to CRC tumorigenesis were highlighted. Our finding of exon 6 splicing change added more complexity of the role of COL6A3 in CRC and this may serve as an additional biomarker in the diagnosis of CRC.

**Alternative splicing of CD44 gene in CRC**

Exon v10 of the CD44 gene was found to be upregulated in CRC compared to NCs (Figure 2B, Supplementary table 3, supplementary figure 3). Exon v10 is the last exon of the 9 alternative exons between exon 5 and exon 16 (Figure 3C). We noticed multiple alternative exons were included in CRC tissues. The complexity of the alternative splicing in CD44 prompted us to use a modified PSI calculation method. As shown in Figure 3A-B, PSI_junc5’ represents the usage/expression of an exon-exon junction compared to all junctions sharing the same 5’ splice site. PSI_junc3’ represents the usage/expression of an
exon-exon junction compared to all junctions sharing the same 3’ splice site. Results showed an increased expression of junc5’_E5-v8 (Figure 3D top), junc3’_v10-E16 (Figure 3D bottom), and decreased expression of junc5’_E5-E16 (Figure 3D middle) in CRC tissues in all of the three cohorts. All other junctions showed less significant changes (data not shown). These data indicate an increase of CD44v8-10 (inclusion of exons v8, v9, and v10) and the decrease of CD44s (standard isoform that skips all 9 alternative exons) isoform in CRC tissues. Metastatic tissues also showed similar changes compared to NCs (Figure 3D left).

The expression of alternatively spliced CD44 adhesion molecules has been implicated in the pathogenesis and metastasis of colorectal cancer. mRNA expression, different splice isoform expression or protein isoform expression have been used in different studies. However, the results are usually conflicting. Either positive correlation [23, 32-35] or no correlation [36] to CRC or metastasis has been reported. Our study suggests that isoform CD44v8-10 is upregulated in CRC, while CD44s are relatively down-regulated in liver metastasis tissues. Furthermore, increasing of CD44v8-10 expression has not been shown in metastasis tissues as well. Therefore, CD44v8-10 or maybe the ratio of CD44v8-10/CD44s could be used as a CRC biomarker.

**Alternative first exon regulation in CRC**

Alternative terminal exon regulations including alternative first exons or alternative last exons are types of alternative splicing, which couples with alternative transcription start site and alternative polyadenylation, respectively. Regular method of calculating PSI for an exon (Figure 2A) does not apply to these situations since only one side of the terminal exon has splice junctions. We used DEXSeq, PSI_junc5’ and PSI_junc3’ (Figure 3A-B) to study these events. After manual inspection of the raw results (Supplementary table 4-6), we finally identified four events (Figure 4A-D, supplementary figure 4-7) regulated in CRC
compared to NC. All of the four events are alternative first exon events. In ARHGEF9 and CHEK1, the more upstream first exons showed downregulation (Figure 4A, C), whereas in HKDC1 and HNF4A the upstream first exons showed upregulation (Figure 4B, D), relative to the downstream first exons. In all of the four cases, changes of the first exon will alter the protein sequences thus will have a potential functional impact on the genes. Although some studies indicated that many of these genes linked to tumorigenesis or metastasis, the detailed functions of these splicing events in CRC have not been reported. ARHGEF9 has been shown to play a role in tumor cell migration, invasion and metastasis by linking oncogene CHD1L and Cdc42 pathway in hepatocellular carcinoma (HCC) [37]. HKDC1 encodes the hexokinase domain containing 1, which catalyzes the phosphorylation of glucose. Its high expression in hepatocarcinoma (HCC) is related to poor overall survival (OS) [38]. And it is also predicted to be a novel therapeutic target in lung cancer [39]. CHEK1 contributes to CDC25C-mediated Docetaxel resistance and can also be a therapeutic target in prostate cancer [40]. In HNF4A gene, the promoter P1 driven isoforms (expressing exon 1A) were decreased in CRCs and MCs compared to the promoter P2-driven isoforms (expressing exon 1D) (Figure 4D, supplementary figure 7). It was reported that P1-HNF4a is expressed primarily in the differentiated compartment of the mouse colonic crypt and P2-HNF4a in the proliferative compartment. The mice that could only produce P2-HNF4a experienced more colitis and developed more tumors than normal mice [41]. Taken together, the splicing changes of these genes identified in this study may contribute to tumorigenesis and can be better biomarkers than gene expression in CRC.

**Alternative splicing associated with metastasis of CRC**

Taking the advantage of the metastasis data from CRC18P, we sought to identify splicing events associated with metastasis of CRC. We selected a subset of splicing events with
the further increase or decrease of PSI values in metastatic tissues compared to CRC. We also required that the CRC and MC samples in CRC18P and the CRC samples in CRC10P data have the consistent trend of PSI change compared to the corresponding NC samples. Nine, eight and fifty events were identified from PSI_exon, PSI_junc5’ and PSI_junc3’ data, respectively (Figure 5A, Supplementary table 3, 5, 6). We manually selected three splicing events in SERPINA1, CALD1 and FBLN2 based on the fact that they were picked up by multiple analysis methods. The results indicated a higher magnitude of PSI change in metastatic samples (Figure 5A-D).

We found that the first exon of SERPINA1 switch from 1a to more downstream one 1b in metastatic samples (supplementary figure 8). Although the two splice isoforms have the same start codon in exon 4, their different 5’UTR sequences may contribute to different RNA decay or protein translation of the gene. Elevated expression of SERPINA1 has been associated with the advanced stage, lymph node metastasis, poor prognosis and shorter overall survival in CRC [42] and gastric cancer [43].

Splicing of a shorter version of exon 5 of CALD1 has been identified by using the alternative 5’ splice sites. And skipping of exon 6 of CALD1 in metastatic CRC samples has also been found (Figure 5C, supplementary figure 9). The same splicing event was reported to be present in colon, bladder and prostate tumors compared to normal tissues [11]. The isoform expressed in metastatic CRC samples matches the low-molecular-weight isoforms (L-CAD), specifically encoded by WI-38 L-CAD II (transcript variant 2). The same isoform was significantly associated with poorer prognosis in urothelial bladder carcinoma (BC) [44]. L-CaD was also linked to lymph node metastasis and poorer prognosis in oral cavity squamous cell carcinoma (OSCC) [45].

Skipping of exon 9 of FBLN2 was identified in metastasis CRC samples (Figure 5D, supplementary figure 10). Preferentially skipping of exon 9 of FBLN2 has been reported in
five cancer types [12]. The FBLN2 short isoform expression was suggested to drive malignant progression and metastasis in lung adenocarcinoma [46].

**Validation of splicing events using TCGA data**

To validate the splicing events identified in this study in an independent dataset, we used a recently developed tool TSVdb [27], which integrates and visualizes alternative splicing data based on TCGA samples for 33 tumor types. We searched all the 10 genes identified in this study (CLK1, COL6A3, CD44, ARHGEF9, HKDC1, CHEK1, HNF4A, SERPINA1, CALD1, FBLN2) against TSVdb and downloaded the junction usage value for colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ). Junction usage is calculated by dividing junction quantification value to mean junction quantification value of that person for a specific gene [27]. We combined junction usage for 379 primary solid tumor samples with 1 metastatic and 2 recurrent solid tumor samples and compared with 51 normal colon or rectal tissue samples. Results indicated that, except for one gene CHEK1 (not significant, P=0.2), all of other 9 genes showed significant splicing changes in tumors samples compared to normal tissues (Wilcoxon test, P-value range from $1 \times 10^{-16}$ to $1 \times 10^{-45}$, supplementary figure 11), consistent with what we observed in the three GEO datasets. ARHGEF9 E1-E3 junction and HNF4A E1b-E3 junction showed most significant P-values (Figure 6A-B). The highly consistent results among several cohorts of patient data indicate that the splicing events identified in this study are reliable markers of CRCs.

**Relevance of splicing events to the survival of patients**

To assess the clinical significance of these findings, we compared the overall survival (OS) data for patients with different splicing profile for the 10 genes. For each junction usage data, we separated all of patients with OS data into two equal sized groups (high and low junction usage groups). We found that two junctions (COL6A3 E5-E6 junction and HKDC1 E1-E2 junction) showed a significant difference in OS (Figure 6C-D). The patients with
higher expression of the junctions, which are also more expressed in CRC cancers, are associated with better survival. The high gene expression of COL6A3 in stroma has been linked to poor overall survival in CRC [30], while high expression of HKDC1 is related to poor overall survival in hepatocarcinoma [38]. However, it’s still unclear whether the splicing isoforms of these two genes have the same functions. It might suggest that the splice isoforms have variable functions in different cancer types. More data is required to illustrate the functional link between the isoforms and patient survival.

**Discussion**

Global splicing changes in cancer have been studied using microarray or RNA-seq techniques in the past [11-13]. Compared to microarray (eg. Affymatrix Exon Array), RNA-seq can more precisely measure the splicing changes because the reads covering the exon-exon junction directly represent the splicing choice. Usually, a PSI value can be calculated based on the read counts of the upstream junction, downstream junction and the junction skipping the exon (eg. Figure 2A). In this study, we used two additional approaches (PSI_junc5’ and PSI_junc3’) to calculate the PSI values (Figure 3A, 3B). The two approaches have the advantage to deal with more complex splicing events like in CD44, nine consecutive alternative exons can be partially included or all excluded. Using regular PSI exon method, only exon v10 was found to be significantly included in CRC. However, combining with the two additional PSI analysis, we concluded that the isoform CD44v8-10 is upregulated in CRC. The PSI_junc5’ and PSI_junc3’ approaches can also be used to study alternative last exons and alternative first exons, respectively. In this study, we identified 5 alternative first exon events in CRC or metastasis tissues. Except for HNF4A, none of the other four events was reported previously. The results indicated that the new approaches could identify novel splicing events. Interestingly, using the same approach, we did not identify any significant alternative last exon events. This may indicate that the
transcription-coupled alternative first exon choices play a more important role in CRC compared to alternative polyadenylation-coupled last exon choices.

Nine of ten splicing events identified in this study were verified by TCGA data. Altogether, 431 CRC or metastatic samples and 82 normal control samples were analyzed, which represent the most comprehensive analysis of splicing in CRC to date. Additional splicing events were listed in supplementary tables 3-6, which provide a resource for future studies.

In summary, we have identified significant alternative splicing of 7 genes in CRC and its metastasis tissue, and 3 genes with the stronger effect in metastatic tissue compared to normal tissue. Among them, the more inclusion of COL6A3-E6, CD44v8-10, and the more exclusion/decrease of HNF4A P1-driven isoform, CALD-E5b and E6, and FBLN2-E9 have been reported in CRC previously. Other 5 splicing events were newly identified in this study. Although the splicing events of these genes were not reported previously, their gene expression level was linked to tumorigenesis or prognosis of CRC (SERPINA1) or other cancer types (ARHGEF9, CHEK1, HKDC1). Targeting the kinase activity of CLK1 was suggested to be a therapeutic approach in cancer therapy. Considering the high consistence of alternative splicing events of the 10 genes identified in this study and previously studies, as well as their high relevance to cancer, it might suggest that the splicing signature of the 10 genes could serve as prognosis marker and facilitate drug development of CRC as well.

Declarations

Funding

This work was financially supported by grants from Natural Science Foundation of Shandong Province (ZR2015LH083)

Authors’ contributions
HJ, CL, AW, and HL conceived this idea. AW, HL, YS, QW, ZZ, RZ, and KL performed the analysis. AW, HL, YS, QW, ZZ, RZ, and KL drafted the manuscript and prepared the Figs. HJ and CL revised the manuscript. All authors contributed to this manuscript. All authors read and approved the final manuscript. All authors declare no conflict of interests in this study.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

**References**

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A: *Global cancer statistics, 2012*. *Ca A Cancer Journal for Clinicians* 2015, 65(2):87-108.

2. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: *Alternative isoform regulation in human tissue transcriptomes*. *Nature* 2008, 456(7221):470-476.

3. Kim HK, Pham M, Ko KS, Rhee BD, Han J: *Alternative splicing isoforms in health and disease*. *Pflugers Arch* 2018:1-22.

4. Su CH, Dhananjaya D, .., Tarn WY: *Alternative Splicing in Neurogenesis and Brain Development*. *Frontiers in Molecular Biosciences* 2018, 5.

5. Urbanski LM, Leclair N, Anczukō W O: *Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics*. *Wiley Interdiscip Rev Rna* 2018(9).

6. Wong A, Rasko J, Wong JJ: *We skip to work: alternative splicing in normal and
malignant myelopoiesis. *Leukemia* 2018.

7. Pradella D, Naro C, Sette C, Ghigna C: **EMT and stemness: flexible processes tuned by alternative splicing in development and cancer progression.**
   *Molecular Cancer* 2017, **16**(1):8.

8. Siegfried Z, Karni R: **The role of alternative splicing in cancer drug resistance.**
   *Curr Opin Genet Dev* 2018, **48**:16-21.

9. Lin JC: **Therapeutic Applications of Targeted Alternative Splicing to Cancer Treatment.** *International Journal of Molecular Sciences* 2018, **19**(1):75.

10. Martinezmontiel N, Rosasmurrieta NH, Anaya MR, Monjarazguzman E, Martinezconrreras R: **Alternative Splicing as a Target for Cancer Treatment.**
    *International Journal of Molecular Sciences* 2018, **19**(2):545.

11. Thorsen K, Sørensen KD, Brems-Eskildsen AS, Modin C, Gaustadnes M, Hein AM, Kruhøffer M, Laurberg S, Borre M, Wang K: **Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis.** *Molecular & Cellular Proteomics Mcp* 2008, **7**(7):1214.

12. Danangotthold M, Golangerstl R, Eisenberg E, Meir K, Karni R, Levanon EY: **Identification of recurrent regulated alternative splicing events across human solid tumors.** *Nucleic Acids Research* 2015, **43**(10):5130-5144.

13. Liu J, Li H, Shen S, Sun L, Yuan Y, Xing C: **Alternative splicing events implicated in carcinogenesis and prognosis of colorectal cancer.** *Journal of Cancer* 2018, **9**(10):1754-1764.

14. Groulx JF, Giroux V, Beauséjour M, Boudjadi S, Basora N, Carrier JC, Beaulieu JF: **Integrin α6A splice variant regulates proliferation and the Wnt/β-catenin pathway in human colorectal cancer cells.** *Carcinogenesis* 2014, **35**(6):1217.

15. Lin JC, Lee YC, Tan TH, Liang YC, Chuang HC, Fann YC, Johnson KR, Lin YJ: **RBM4-**
SRSF3-MAP4K4 splicing cascade modulates the metastatic signature of colorectal cancer cell. *Biochim Biophys Acta* 2017, **1865**(2).

16. Riffo-Campos ÁL, Castillo J, Vallet-Sánchez A, Ayala G, Cervantes A, López-Rodas G, Franco L: In silico RNA-seq and experimental analyses reveal the differential expression and splicing of EPDR1 and ZNF518B genes in relation to KRAS mutations in colorectal cancer cells. *Oncology Reports* 2016, **36**(6):3627.

17. Kajiwara T, Matsushita K, Itoga S, Tamura M, Tanaka N, Tomonaga T, Matsubara H, Shimada H, Habara Y, Matsuo M: **SAP155-mediated c-myc suppressor far-upstream element-binding protein-interacting repressor splicing variants are activated in colon cancer tissues.** *Cancer Science* 2013, **104**(2):149-156.

18. Hirschi B, Kolligs FT: **Alternative splicing of BRAF transcripts and characterization of C-terminally truncated B-Raf isoforms in colorectal cancer.** *Int J Cancer* 2013, **133**(3):590-596.

19. Pelisch F, Khauv D, Risso G, Stallingsmann M, Blaustein M, Quadrana L, Radisky DC, Srebrow A: **Involvement of hnRNP A1 in the matrix metalloprotease-3-dependent regulation of Rac1 pre-mRNA splicing.** *Journal of Cellular Biochemistry* 2012, **113**(7):2319-2329.

20. Gazouli M, Roubelakis MG, Theodoropoulos GE, Papailiou J, Vaiopoulou A, Pappa KI, Nikiteas N, Anagnou NP: **OCT4 spliced variant OCT4B1 is expressed in human colorectal cancer.** *Mol Carcinog* 2012, **51**(2):165-173.

21. Wang MH, Lao WF, Wang D, Luo YL, Yao HP: **Blocking tumorigenic activities of colorectal cancer cells by a splicing RON receptor variant defective in the tyrosine kinase domain.** *Cancer Biology & Therapy* 2007, **6**(7):1121-1129.

22. Wong LS, Cantrill JE, Morris AG, Fraser IA: **Expression of CD44 splice variants in colorectal cancer.** *British Journal of Surgery* 2015, **84**(3):363-367.
23. Takeuchi K, Yamaguchi A, Urano T, Goi T, Nakagawara G, Shiku H: Expression of CD44 Variant Exons 8-10 in Colorectal Cancer and Its Relationship to Metastasis. Jpn J Cancer Res 2010, 86(3):292-297.

24. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013, 29(1):15-21.

25. Helga T, Robinson JT, Mesirov JP: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 2013, 14(2):178-192.

26. Anders S, Reyes A, Huber W: Detecting differential usage of exons from RNA-seq data. Genome Research 2012, 22(10):2008-2017.

27. Sun W, Duan T, Ye P, Chen K, Zhang G, Lai M, Zhang H: TSVdb: a web-tool for TCGA splicing variants analysis. Bmc Genomics 2018, 19(1):405.

28. Araki S, Dairiki R, Nakayama Y, Murai A, Miyashita R, Iwatani M, Nomura T, Nakanishi O: Inhibitors of CLK Protein Kinases Suppress Cell Growth and Induce Apoptosis by Modulating Pre-mRNA Splicing. Plos One 2015, 10(1):e0116929.

29. Arafat H, Lazar M, Salem K, Chipitsyna G, Gong Q, Pan TC, Zhang RZ, Yeo CJ, Chu ML: Tumor-specific expression and alternative splicing of the COL6A3 gene in pancreatic cancer. Journal of Surgical Research 2011, 165(2):176-176.

30. Qiao J, Fang CY, Chen SX, Wang XQ, Cui SJ, Liu XH, Jiang YH, Wang J, Zhang Y, Yang PY: Stroma derived COL6A3 is a potential prognosis marker of colorectal carcinoma revealed by quantitative proteomics. Oncotarget 2015, 6(30):29929-29946.

31. Liu W, Li L, Ye H, Tao H, He H: Role of COL6A3 in colorectal cancer. Oncology Reports 2018, 39(6):2527-2536.
32. Ropponen K, Eskelinen M, Lipponen P, Alhava E, Kosma V: Expression of CD44 and variant proteins in human colorectal cancer and its relevance for prognosis. Scandinavian Journal of Gastroenterology 1998, 33(3):301-309.

33. Li XD, Ji M, Wu J, Jiang JT, Wu CP: Clinical significance of CD44 variants expression in colorectal cancer. Tumori 2013, 99(1):88-92.

34. Du L, Wang H, He L, Zhang J, Ni B, Wang X, Jin H, Cahuzac N, Mehrpour M, Lu Y: CD44 is of Functional Importance for Colorectal Cancer Stem Cells. Clinical Cancer Research 2008, 14(21):6751-6760.

35. Xia P, Xu XY: Prognostic significance of CD44 in human colon cancer and gastric cancer: evidence from bioinformatic analyses. Oncotarget 2016, 7(29):45538-45546.

36. Givehchian M, Wörner S, Sträter J, Zöller M, Heuschen U, Heuschen G, Lehnert T, Herfarth C, Doeberitz MVK: No evidence for cancer-related CD44 splice variants in primary and metastatic colorectal cancer. European Journal of Cancer 1998, 34(7):1099-1104.

37. Chen L, Chan TH, Yuan YF, Hu L, Huang J, Ma S, Wang J, Dong SS, Tang KH, Xie D: CHD1L promotes hepatocellular carcinoma progression and metastasis in mice and is associated with these processes in human patients. Journal of Clinical Investigation 2010, 120(4):1178-1191.

38. Zhang Z, Huang S, Wang H, Wu J, Chen D, Peng B, Zhou Q: High expression of hexokinase domain containing 1 is associated with poor prognosis and aggressive phenotype in hepatocarcinoma. Biochemical & Biophysical Research Communications 2016, 474(4):673-679.

39. Li GH, Huang JF: Inferring therapeutic targets from heterogeneous data: HKDC1 is a novel potential therapeutic target for cancer. Bioinformatics 2014,
30(6):748-752.

40. Nakouzi NA, Cotteret S, Commo F, Gaudin C, Rajpar S, Dessen P, Vielh P, Fizazi K, Chauchereau A: Targeting CDC25C, PLK1 and CHEK1 to overcome Docetaxel resistance induced by loss of LZTS1 in prostate cancer. Oncotarget 2014, 5(3):667-678.

41. Chellappa K, Deol P, Evans JR, Vuong LM, Chen G, Briançon N, Bolotin E, Lytle C, Nair MG, Sladek FM: Opposing roles of nuclear receptor HNF4α isoforms in colitis and colitis-associated colon cancer. Elife 2016, 5.

42. Kwon CH, Park HJ, Jin HC, Lee JR, Kim HK, Jo H, Kim HS, Oh N, Song GA, Park DY: Snail and serpinA1 promote tumor progression and predict prognosis in colorectal cancer. Oncotarget 2015, 6(24):20312-20326.

43. Kwon CH, Park HJ, Lee JR, Kim HK, Jeon TY, Jo HJ, Kim DH, Kim GH, Park DY: Serpin peptidase inhibitor clade A member 1 is a biomarker of poor prognosis in gastric cancer. Br J Cancer 2014, 111(10):1993-2002.

44. Myung-Shin L, Jisu L, Heon KJ, Tae KW, Wun-Jae K, Hanjong A, Jinsung P: Overexpression of caldesmon is associated with tumor progression in patients with primary non-muscle-invasive bladder cancer. Oncotarget 2015, 6(37):40370-40384.

45. Chang KP, Wang CL, Kao HK, Liang Y, Liu SC, Huang LL, Hseuh C, Hsieh YJ, Chien KY, Chang YS et al: Overexpression of caldesmon is associated with lymph node metastasis and poorer prognosis in patients with oral cavity squamous cell carcinoma. Cancer 2013, 119(22):4003-4011.

46. Baird BN, Schliekelman MJ, Ahn YH, Chen Y, Roybal JD, Gill BJ, Mishra DK, Erez B, O’Reilly M, Yang Y: Fibulin-2 Is a Driver of Malignant Progression in Lung Adenocarcinoma. Plos One 2013, 8(6):e67054.
### Figures

| Dataset name | GEO ID  | Description                                                                 |
|--------------|---------|-----------------------------------------------------------------------------|
| CRC18P       | GSE50760| normal colon (NC), primary CRC (CRC), and liver metastasis (MC) for 18 patients |
| CRC10P       | GSE95132| 10 pairs primary tumor (CRC) and matching normal colon tissue (NC)           |
| CRC3P        | GSE104178| 3 pairs of colorectal cancer (CRC) and matched normal mucosa (NM)           |

**Figure 1**

Diagram of CRC RNA-seq dataset and analysis method.
Figure 2

Cassette exon regulation in CRC. (A) Diagram of the method to calculate the PSI of an exon. The orange boxes and gray boxes are the alternative exons and

\[ \text{PSI of exon}_i = \frac{\text{average}(a,b)}{\text{average}(a,b) + c} \]
neighboring exons. Thick bars connected by a dotted line represent a read cover two exons (junction read). a, b and c are read counts for the three junctions. (B) Venn diagram showing the overlap of exon splicing events between CRC and NC in CRC18P and CRC10P datasets. P-value of non-paired, two-tailed t-test<0.001 and |ΔPSI|>20% were used as cutoffs to select significant events. Inc, exon inclusion in CRC; Exc, exon exclusion in CRC. (C) CLK1 gene structure (top) and read coverage (Sashimi Plot) for exon 3 to exon 5 region. (D) Boxplot of PSI of CLK1 exon4 in the CRC18P dataset. Dots in the boxplot represent individual patient data. P-value is based on Student’s t-test. (E) Similar to (D) except that the CRC10P dataset was shown. (F) COL6A3 gene structure (top) and read coverage (Sashimi Plot) for exon 5 to exon 7 regions. (G) Boxplot of PSI of COL6A3 exon6 in the CRC18P dataset. Dots in the boxplot represent individual patient data. P-value is based on Student’s t-test. (H) Similar to G except that the CRC10P dataset was shown.
CD44v8-10 showed up-regulation in CRC at the expense of other CD44 splicing
variants. (A) Diagram of the method to calculate PSI\textsubscript{junc5'}, which represents the usage of a junction among all junctions sharing the same 5’ splice site. The boxes are exons. Thick bars connected by a dotted line represent a read cover two exons (junction read). a, b and c are read counts for the three junctions. (B) Similar to (A) except that diagram of PSI\textsubscript{junc3'} was shown, which represents the usage of a junction among all junctions sharing the same 3’ splice site. (C) CD44 gene structure (bottom) and read coverage (Sashimi Plot) for exon 5 to exon 16 regions. (D) Boxplot of PSI of CD44 junc\textsubscript{5'} E5-v8 (top row), junc\textsubscript{5'} E5-E16 (middle row) and junc\textsubscript{3'} v10-E16 (bottom row) in CRC18P (left column), CRC10P (middle column) and CRC3P (right column) datasets. Dots in the boxplot represent individual patient data. P-value is based on Student’s t-test.
Alternative first exon regulation in CRC. (A) Read coverage of ARHGEF9 alternative first exons E1α and E1β. The height of the RNA-seq tracks represents the Read Per Million (RPM) values of the read coverage at each genomic location. The adjusted P-value and the log2 ratio (based on DEXSeq) of E1α were shown. (B) Similar to (A) except that gene HKDC1 was shown. (C) similar to A except that gene CHEK1 was shown. (D) Similar to (A) except that gene HNF4A was shown and t-test P-value and ΔPSI of junc_3’ E1α-E3 was shown.
Metastasis-related splicing events. (A) Heat map showing the PSI exon, PSI_junc5’ and PSI_junc3’ values for metastasis-related splicing events. The following criteria were used to select these events: MC vs. NC |ΔPSI|>20% and P<0.01 (t-test); MC vs. NC and CRC vs. NC in CRC18P and CRC10P data showing a consistent trend; MC vs. CRC |ΔPSI|>15%. Several exons or junctions were labeled on the right. (B-D) Read coverage for SERPINA1 alternative first exons (B), CALD1 E5 to E7 (C), FBLN2 E8 to E10 regions (D).
Figure 6

Splicing events confirmed by TCGA data. (A-B) Boxplots of junction usage of ARHGEF9 E1-E3 junction (A) and HNF4A E1b-E3 junction (B) in 51 normal tissue and 382 CRC or metastatic tissue (tumor). P-values are based on Wilcoxon Rank-Sum Test. Dots in the boxplot represent the individual patient in TCGA. (C-D) Survival curves of 357 patients with overall survival data equally separated into two groups (low and high) based on junction usage of COL6A3 E5-E6 (C) and HKDC1 E1-E2 (D). P-value is based on the log-rank test. Confidence intervals were shown as shaded areas.

Supplementary Files
This is a list of supplementary files associated with the primary manuscript. Click to download.

supp_tables.xlsx
supp-Figures.pptx