Tumor heterogeneity and prognostic impact of the aberrant splicing burden of colorectal cancers

Ina A. Eilertsen[1], Jonas Langerud[1,2], Bjarne Johannessen[1], Arild Nesbakken[2,3], Ragnhild A. Lothe[1,2], and Anita Sveen[1,2]

[1]Department of Molecular Oncology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway; [2]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway; [3]Department of Gastrointestinal Surgery, Oslo University Hospital, Oslo, Norway.

Table of Contents

Supplementary Methods

DNA/RNA extraction, mutation and MSI status analyses

Total RNA and DNA extraction using the AllPrep DNA/RNA/miRNA Universal kit (Qiagen GmBH, Hilden, Germany), Sanger sequencing of mutation hotspots in *KRAS* (exon 2: codons 12 and 13; exon 3: codon 61; and exon 4: codons 117 and 146) and *BRAF* exon 15 (including codon 600), and determination of MSI status by analysis of the Bethesda marker panel have previously been published for a subset of samples (1-4). The remaining tumor samples (n = 294) were analyzed using the same protocols, but with determination of MSI status based on the BAT-25/26 loci or the MSI Analysis System v1.2 (Promega, Madison, WI, USA). Additional somatic mutations in 40 of the tumors have previously been analyzed by whole-exome sequencing (1, 5).

Visualization of detected splicing events from splicing-sensitive microarrays

Detected splicing events were visualized with the Integrative Genomics Viewer (IGV) v2.4.3 software (6), using two custom GTF files defining the structure of each event and the location of intra-exonic and junction probes supporting each path of each event (two mutually exclusive paths and a shared reference path). The GTF files were generated using the EventPointer_IGV function in the R package EventPointer v2.2.4 (7, 8). Gene-wise splice graphs were visualized with the plotSpliceGraph function in the R package SGSeq v1.18.0 (9). The feature annotation file Homo_sapiens.GRCh37.74.gtf from Ensembl was applied for both EventPointer and SGSeq.

Pre-processing of gene-level expression data from splicing-sensitive microarrays

For gene-level analysis, raw intensity CEL files were processed by the RMA algorithm using the aroma.affymetrix framework with a custom CDF file from Brainarray (v18; ENSG (10)), built using feature annotation from Ensembl (GRCh37 v74). Gene expression values were on log2-scale and adjusted for batch effect with ComBat (Fig. S2), as for the alternative splicing analysis. Ensembl IDs were translated to Entrez IDs using the R packages AnnotationDbi v1.46.0 and org.Hs.eg.db v3.8.2, and outdated gene symbols were corrected using the checkGeneSymbols function in the R package HGNChelper v0.8.1.

Pre-processing of splicing-level expression data from RNA sequencing

Raw sequencing reads were trimmed using Trimmomatic v0.38 (11), and alignment of trimmed reads to the human reference genome GRCh37 was performed with STAR v2.5.3a (2-pass mapping (12)), setting the outSAMattributes to XS for alternative splicing analysis. This generated a custom flag indicating the direction of transcription for spliced alignments. Transformation of the flagged aligned reads into gene-wise splice graphs and feature counts was done using the PrepareBam_EP function in the R package EventPointer (8), with a default FPKM requirement of 2 to call splice junctions. The EventDetection function in EventPointer was used to identify and classify all detected splicing events.

PSI values were estimated using the getPSI_RNASeq function, and splicing events with any missing values were omitted. The feature annotation file Homo_sapiens.GRCh37.74.gtf from Ensembl was applied for both STAR and EventPointer.

Splicing subtype discovery

Unsupervised classification of tumors according to mRNA splicing profiles was based on PSI values of aberrantly spliced events and using NMF as implemented in the R package NMF v0.23.0 (13), with 100 repetitions and the Brunet method (14). Features for NMF were selected by differential splicing analysis between tumor and normal samples ($|\Delta PSI| > 0.10$ and P-value < 0.001). This retained 1,091 events (Table S6), of which 98% overlapped with the cancer-specific events from divergence analysis. The predefined rank (K) was 2 to 6, and the optimal number of sample clusters was determined by selecting the rank with the highest cophenetic correlation and average silhouette width.

Assumption of proportional hazards in survival analyses

The assumption of proportional hazards was met in all analyses, except for cancer stage in the analyses using 5-year overall survival as endpoint. The multivariable model was therefore also evaluated by stratification according to cancer stage, providing similar results for the prognostic value of all other covariables (data not shown).

Statistical analyses

Tumor heterogeneity on the splicing and gene expression levels was calculated in the 98 tumors with multi-regional samples as pairwise Euclidean distances using the dist function in the R package stats v3.6.1 with the 5,000 splicing events or genes with the largest cross-sample variance as features, respectively. The mean of all pairwise Euclidean distances between any two samples from the same tumor was used as a measure for intra-tumor heterogeneity, whereas the pairwise distance between any two samples from different tumors was used as a measure for inter-tumor heterogeneity. PCA was done using the prcomp function in the R package stats with the 5,000 splicing events or genes with the largest cross-sample variance as features.

Spearman's rank correlation, Fisher's exact and Mann-Whitney tests were done using the cor.test, fisher.test and wilcox.test functions in the R package stats. Kruskal-Wallis test was performed using the kruskal.test function, followed by Dunn's multiple comparisons test with Benjamini-Hochberg correction using the dunnTest function in the R package FSA v0.8.26. Adjustments for multiple testing were done using the p.adjust function in the R package stats (method: false discovery rate, FDR).

Supplementary Figures

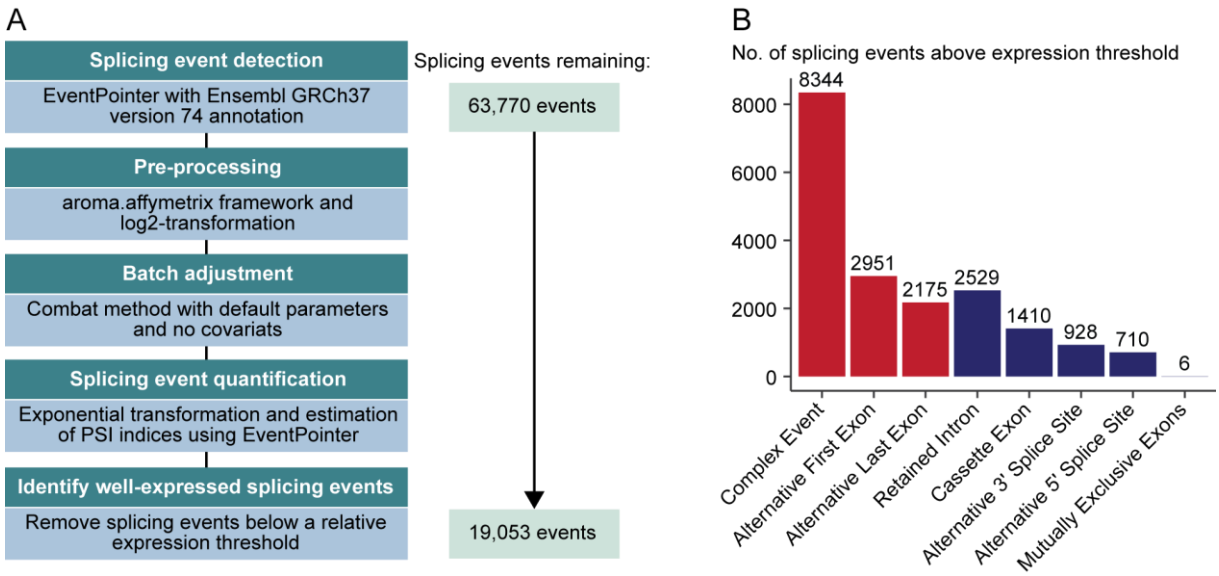


Fig. S1. Splicing event detection, quantification and filtering of the in-house dataset

A) The in-house microarray data was pre-processed according to a custom CDF file storing probe set information for 63,770 splicing events from 17,493 unique genes, identified by mapping the probes on the GeneChip Human Transcriptome Array 2.0 to feature annotation from Ensembl (GRCh37 version 74). After log2-transformation and adjustment for a technical batch effect, the adjusted probe set-level expression values were exponentially transformed prior to estimation of PSI indices. Next, a relative expression threshold was applied to exclude lowly expressed splicing events, resulting in 19,053 events for downstream analysis. **B)** Bar plot showing the number of splicing events above the relative expression threshold, grouped according to splicing event category. Although the mechanism of alternative first and last exons (in red) are different from that of typical alternative splicing (in blue), these splicing modes were retained. This was also done for events annotated as complex (not assigned any of the standard categories; also in red).

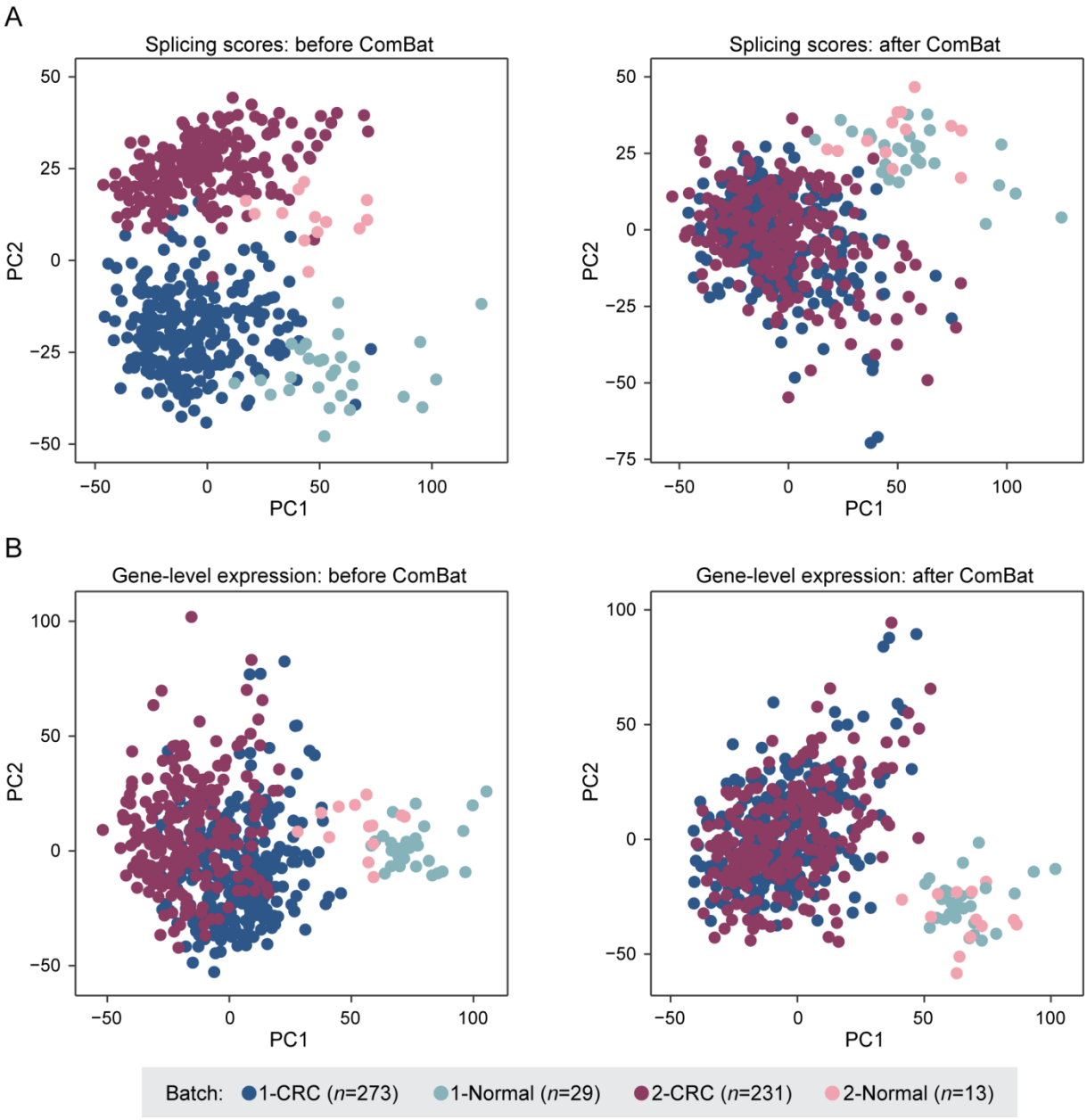


Fig. S2. ComBat adjustment of splicing scores and gene-level expression profiles

PCA of **A)** splicing scores and **B)** gene-level expression values of the in-house series (*n* = 504 CRCs and 42 normal colonic mucosa samples) showing the impact of reagent lot number (batch) on sample separation before (left panels) and after (right panels) ComBat adjustment.

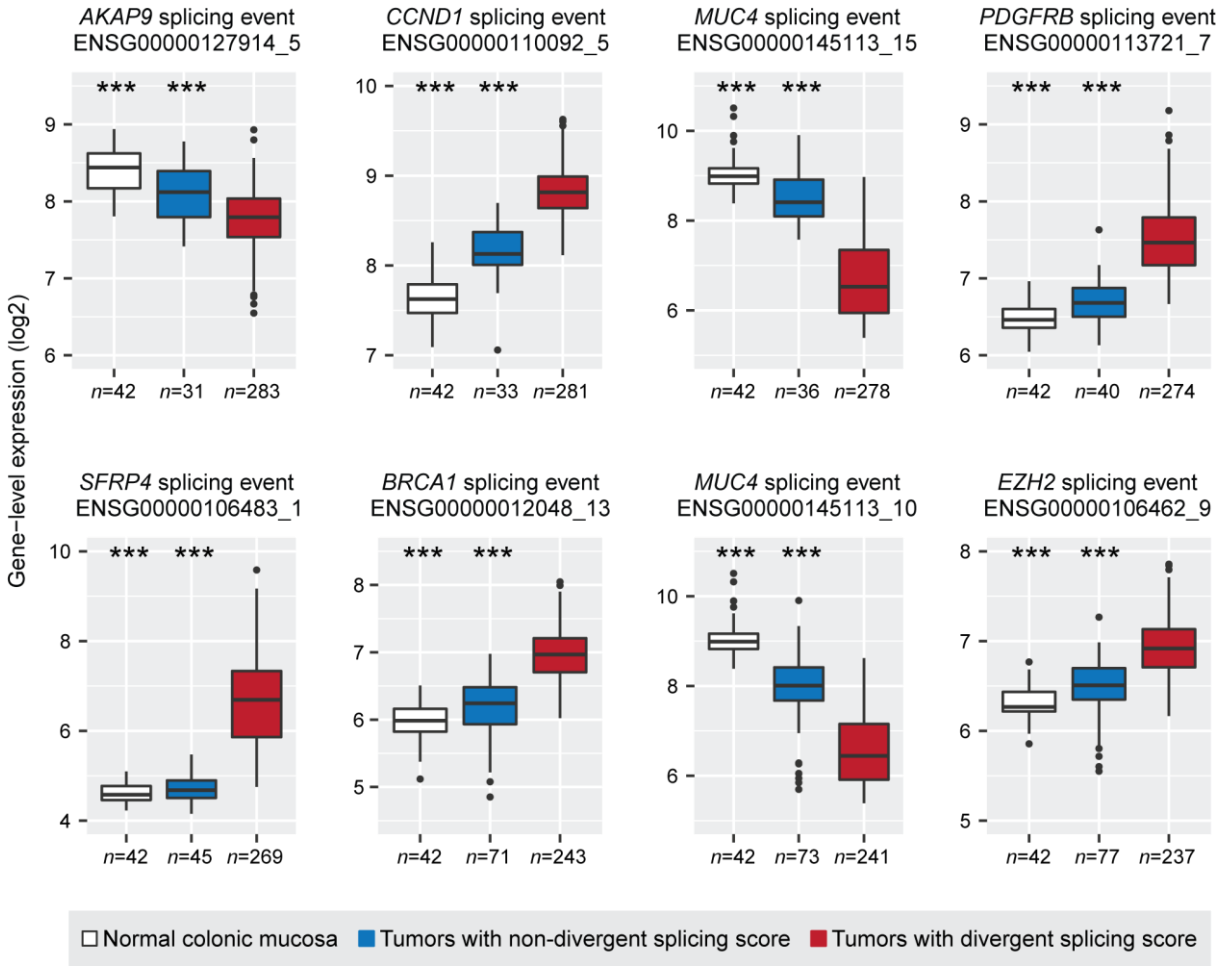


Fig. S3. Impact of aberrant splicing on the expression level of cancer-critical genes

Box plots of frequent aberrant splicing events in cancer-critical genes (sample set prevalence > 75%) showing the target gene expression level in tumor and normal samples. The tumors are stratified into groups according to the splicing status of each splicing event, indicating whether the splicing score in each tumor was within (non-divergent) or outside (divergent) the estimated range of splicing scores in the normal reference population. ***$P < 0.001$ compared to the divergent group from Dunn's multiple comparisons test with Benjamini-Hochberg correction. All analyses are based on single tumor samples from each patient.

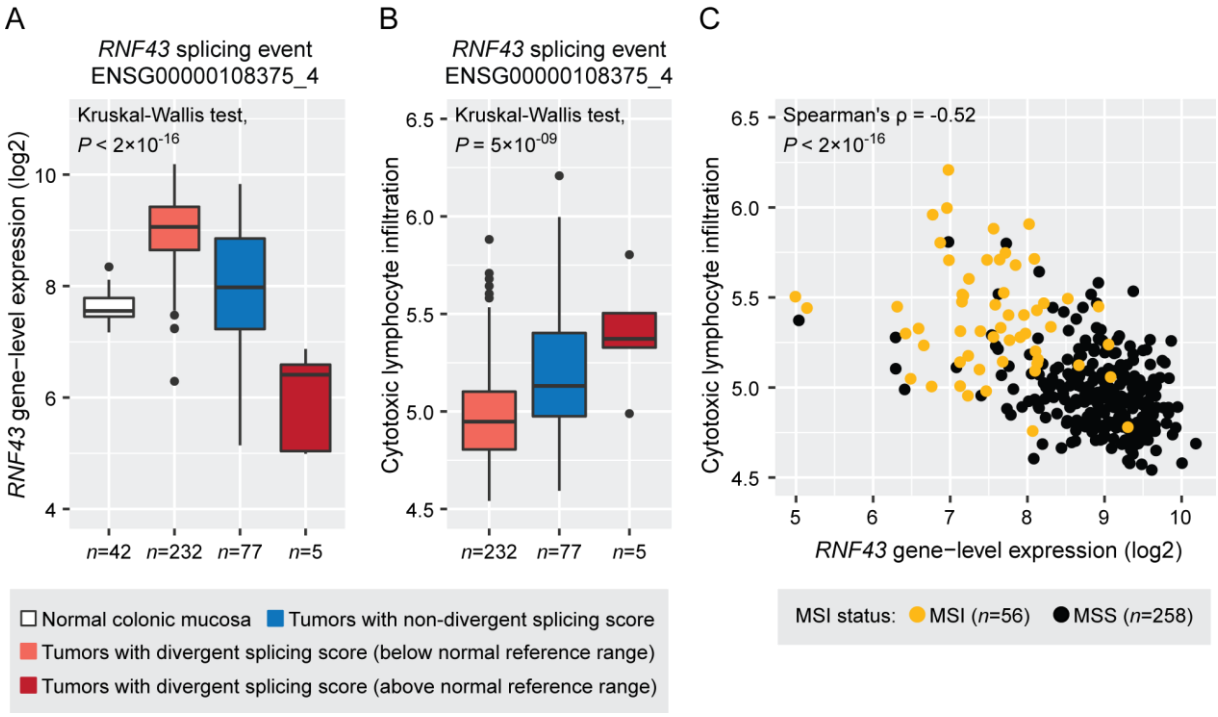
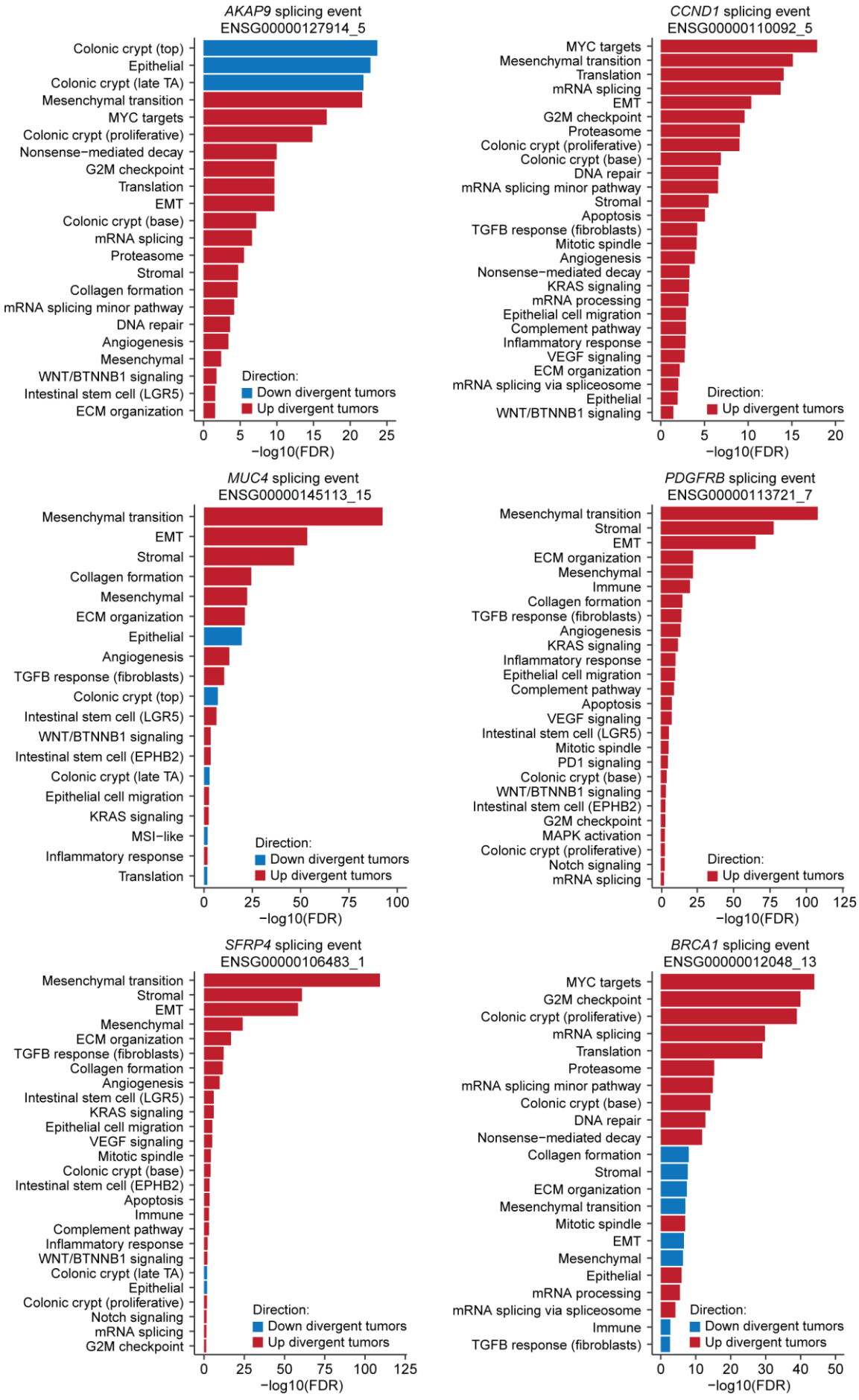


Fig. S4. Aberrant splicing of the cancer-critical *RNF43* gene

Box plots showing **A)** the expression level of the *RNF43* gene in tumor and normal samples and **B)** the level of cytotoxic lymphocyte infiltration in each tumor sample. The tumors are stratified into groups according to the splicing status of a single event in the *RNF43* gene, indicating whether the splicing score in each tumor was non-divergent or divergent by being either below or above the estimated range of splicing scores in the non-malignant reference population. **C)** Scatter plot of *RNF43* gene-level expression versus cytotoxic lymphocyte infiltration, colored according to MSI status of each tumor. All analyses are based on single tumor samples from each patient.


AKAP9 splicing event
ENSG00000127914_5
Colonic crypt (top)
Epithelial
Colonic crypt (late TA)
Mesenchymal transition
MYC targets
Colonic crypt (proliferative)
Nonsense−mediated decay
G2M checkpoint
Translation
EMT
Colonic crypt (base)
mRNA splicing
Proteasome
Stromal
Collagen formation
mRNA splicing minor pathway
DNA repair
Angiogenesis
Mesenchymal
WNT/BTNNB1 signaling
Intestinal stem cell (LGR5)
ECM organization
Direction:
Down divergent tumors
Up divergent tumors
0 5 10 15 20 25
−log10(FDR)
CCND1 splicing event
ENSG00000110092_5
MYC targets
Mesenchymal transition
Translation
mRNA splicing
EMT
G2M checkpoint
Proteasome
Colonic crypt (proliferative)
Colonic crypt (base)
DNA repair
mRNA splicing minor pathway
Stromal
Apoptosis
TGFB response (fibroblasts)
Mitotic spindle
Angiogenesis
Nonsense−mediated decay
KRAS signaling
mRNA processing
Epithelial cell migration
Complement pathway
Inflammatory response
VEGF signaling
ECM organization
mRNA splicing via spliceosome
Epithelial
WNT/BTNNB1 signaling
Direction:
Up divergent tumors
0 5 10 15 20
−log10(FDR)
MUC4 splicing event
ENSG00000145113_15
Mesenchymal transition
EMT
Stromal
Collagen formation
Mesenchymal
ECM organization
Epithelial
Angiogenesis
TGFB response (fibroblasts)
Colonic crypt (top)
Intestinal stem cell (LGR5)
WNT/BTNNB1 signaling
Intestinal stem cell (EPHB2)
Colonic crypt (late TA)
Epithelial cell migration
KRAS signaling
MSI−like
Inflammatory response
Translation
Direction:
Down divergent tumors
Up divergent tumors
0 25 50 75 100
−log10(FDR)
PDGFRB splicing event
ENSG00000113721_7
Mesenchymal transition
Stromal
EMT
ECM organization
Mesenchymal
Immune
Collagen formation
TGFB response (fibroblasts)
Angiogenesis
KRAS signaling
Inflammatory response
Epithelial cell migration
Complement pathway
Apoptosis
VEGF signaling
Intestinal stem cell (LGR5)
Mitotic spindle
PD1 signaling
Colonic crypt (base)
WNT/BTNNB1 signaling
Intestinal stem cell (EPHB2)
G2M checkpoint
MAPK activation
Colonic crypt (proliferative)
Notch signaling
mRNA splicing
Direction:
Up divergent tumors
0 25 50 75 100 125
−log10(FDR)
SFRP4 splicing event
ENSG00000106483_1
Mesenchymal transition
Stromal
EMT
Mesenchymal
ECM organization
TGFB response (fibroblasts)
Collagen formation
Angiogenesis
Intestinal stem cell (LGR5)
KRAS signaling
Epithelial cell migration
VEGF signaling
Mitotic spindle
Colonic crypt (base)
Intestinal stem cell (EPHB2)
Apoptosis
Immune
Complement pathway
Inflammatory response
WNT/BTNNB1 signaling
Colonic crypt (late TA)
Epithelial
Colonic crypt (proliferative)
Notch signaling
mRNA splicing
G2M checkpoint
Direction:
Down divergent tumors
Up divergent tumors
0 25 50 75 100 125
−log10(FDR)
BRCA1 splicing event
ENSG00000012048_13
MYC targets
G2M checkpoint
Colonic crypt (proliferative)
mRNA splicing
Translation
Proteasome
mRNA splicing minor pathway
Colonic crypt (base)
DNA repair
Nonsense−mediated decay
Collagen formation
Stromal
ECM organization
Mesenchymal transition
Mitotic spindle
EMT
Mesenchymal
Epithelial
mRNA processing
mRNA splicing via spliceosome
Immune
TGFB response (fibroblasts)
Direction:
Down divergent tumors
Up divergent tumors
0 10 20 30 40 50
−log10(FDR)

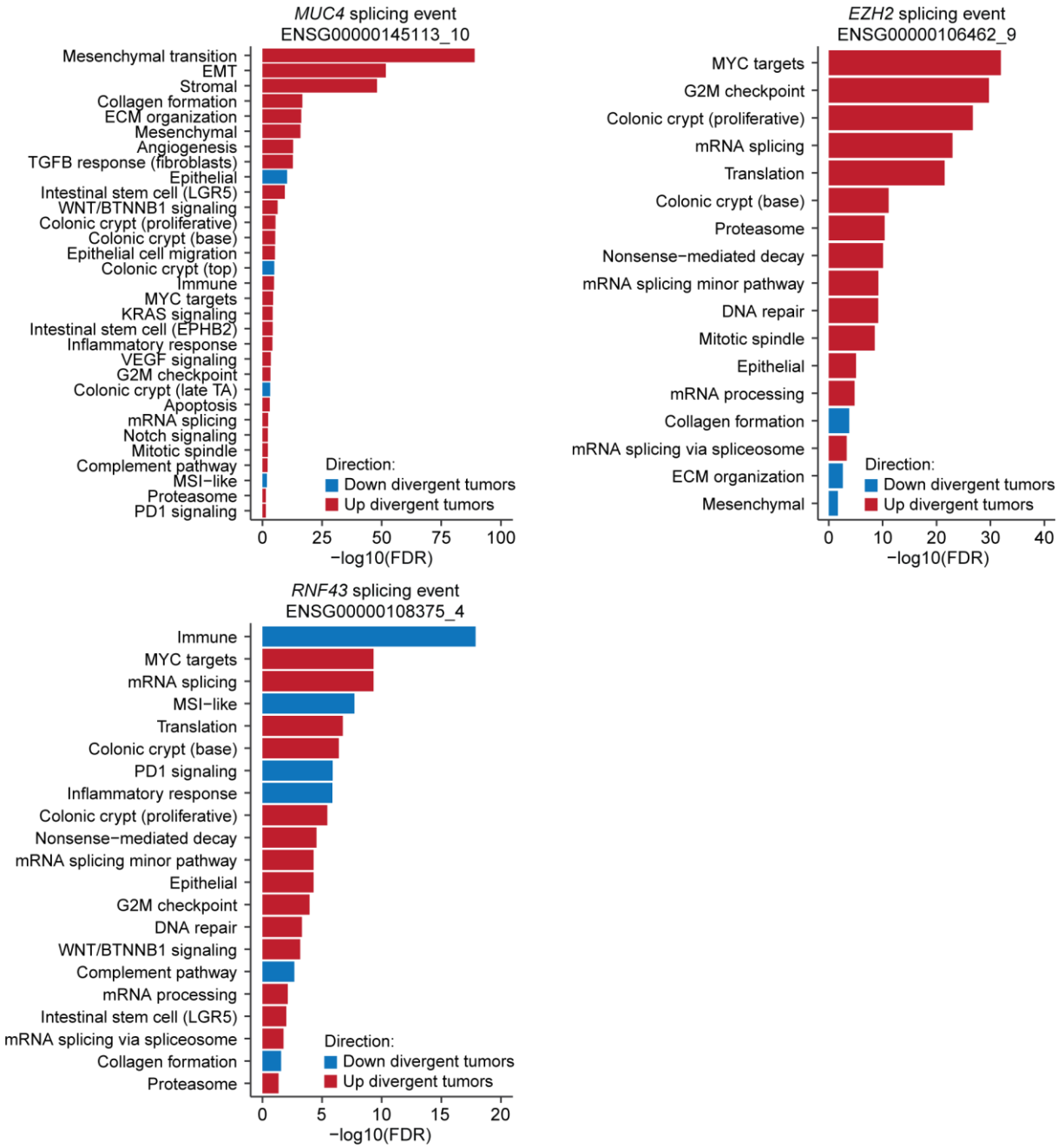


Fig. S5. Gene set enrichment analyses for splicing events in cancer-critical genes

Gene set enrichment analyses illustrating a custom gene set collection with FDR < 0.05 from comparisons of tumors with or without aberrant (divergent versus non-divergent) splicing of events in cancer-critical genes (9 events with prevalence > 75%). Epithelial-mesenchymal distinction was among the top two gene sets for events in *AKAP9*, *CCND1*, *MUC4*, *PDGFRB* and *SFRP4*, and mRNA splicing was among the top four for events in *CCND1*, *BRCA1*, *EZH2* and *RNF43*. Notably, for the event in *RNF43*, tumors with aberrant (divergent) splicing were grouped together irrespective of splicing-direction variability. All analyses are based on single tumor samples from each patient. TA: Transit-amplifying; ECM: Extracellular matrix.

Fig. S6. Relationship between splicing scores and gene-level expression profiles

Scatter plots illustrating the correspondence of PC1 (left) and PC2 (right) values from PCA of splicing scores and gene level expression among single tumor samples from each patient.

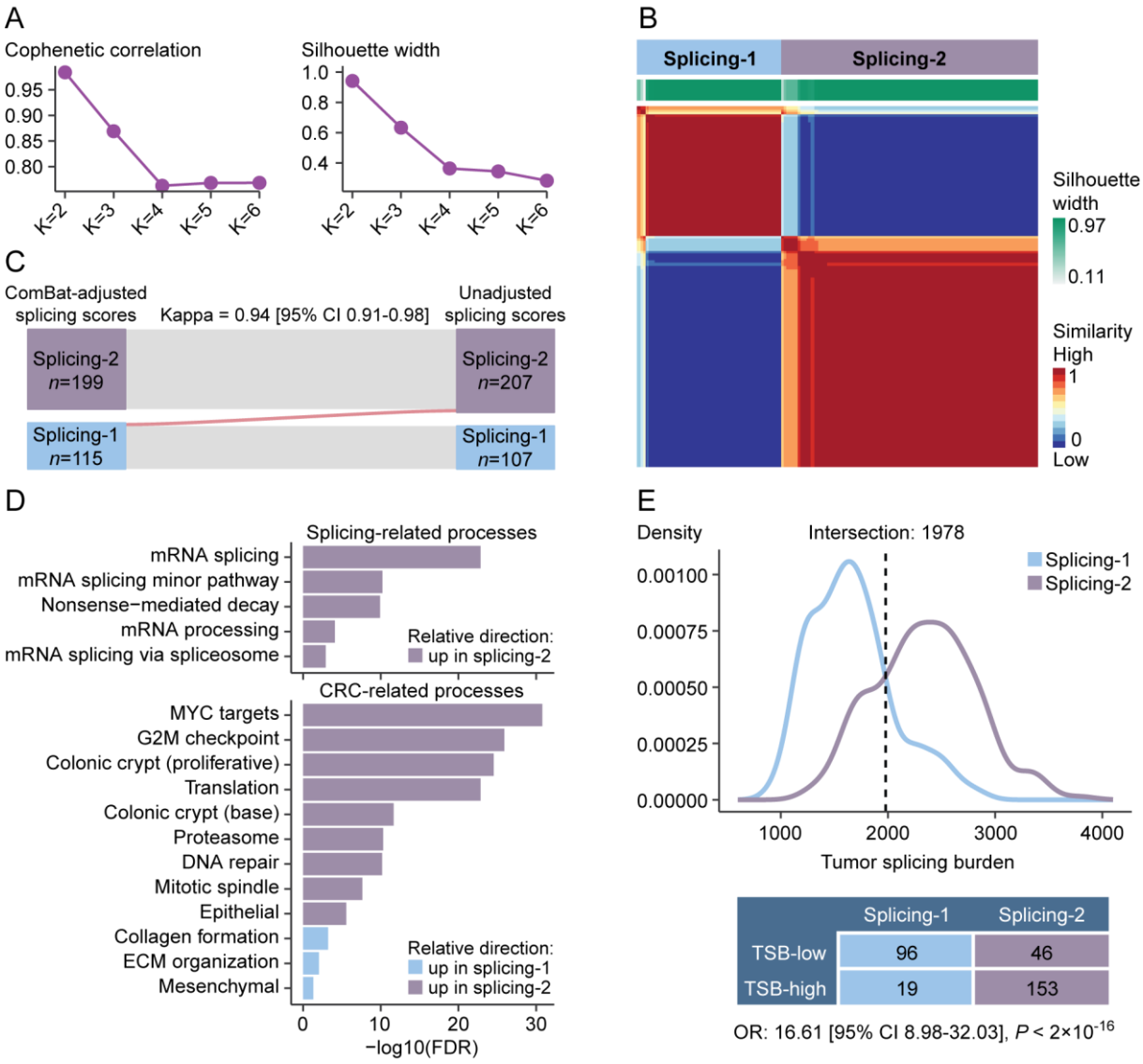


Fig. S7. Unsupervised *de novo* subtyping of CRCs based on aberrant splicing events

A) Quality metrics from unsupervised NMF classification of tumors (single samples per patient) based on the splicing scores of differentially spliced events between tumor and normal samples, indicating that the optimal number of stable sample clusters (K) was 2. **B)** Heatmap of NMF clustering at K = 2, with the color intensity representing the within-cluster similarity of each tumor. The top annotation bars indicate sample clusters and sample-wise silhouette width. **C)** Alluvial diagram comparing the splicing subtypes with results from analysis of splicing scores not adjusted for batch effect. Features for NMF classification based on unadjusted splicing scores were selected by differential splicing analysis between tumor and normal samples with reagent lot number included as blocking factor. **D)** Gene set enrichment analysis showing a custom gene set collection with FDR < 0.05 from comparison of the splicing-1 and splicing-2 subtypes. **E)** Density distribution (top) of the tumor splicing burden in the splicing-1 and splicing-2 subtypes. Cross-table (bottom) compares tumor subgroups dichotomized according to the tumor splicing burden density intersection (dotted line) between the two subtypes. ECM: extracellular matrix, OR: odds ratio, TSB: tumor splicing burden.

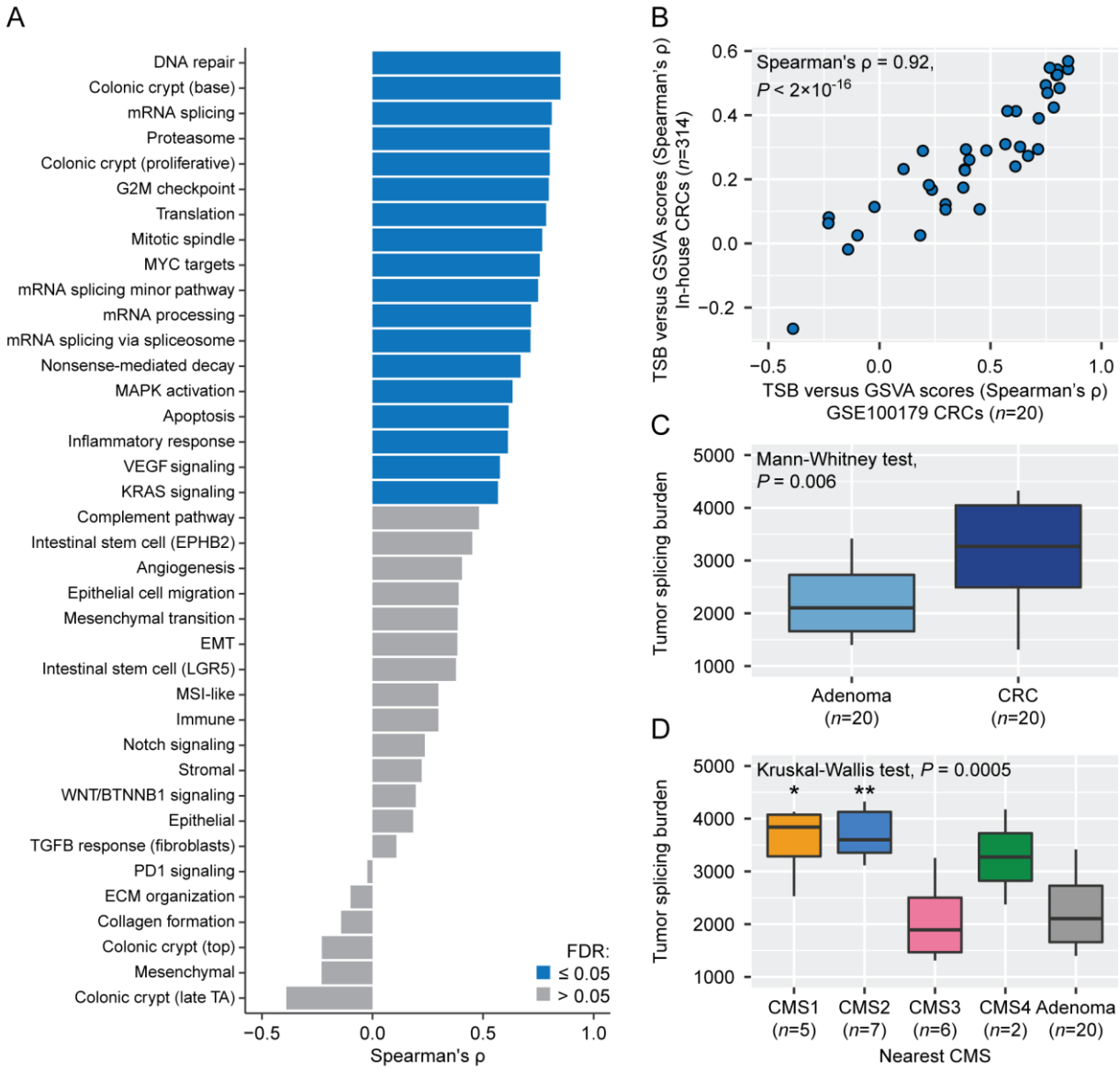


Fig. S8. Tumor splicing burdens of an external series of primary CRCs and colonic adenomas (GEO: GSE100179)

A) Bar plot of Spearman correlation coefficients between the tumor splicing burden and single-sample enrichment (GSVA) scores for a custom collection of 38 CRC- and splicing-related gene sets among 20 primary CRCs in the external GSE100179 series, ordered by decreasing correlation coefficients and colored according to significance level (FDR). **B)** Scatter plot of the external GSE100179 versus in-house series comparing correlation coefficients between the tumor splicing burden and enrichment scores of the same 38 gene sets. Box plots comparing the tumor splicing burden **C)** between colonic adenomas and CRCs, and **D)** among the four CMS groups. Tumors were classified according to the nearest CMS group (no posterior probability threshold). *P < 0.05 and **P < 0.01 compared to the CMS3 group from Dunn's multiple comparisons test with Benjamini-Hochberg correction. CMS4 tumors showed the same trend as CMS1 and CMS2 tumors, although not statistically significant.

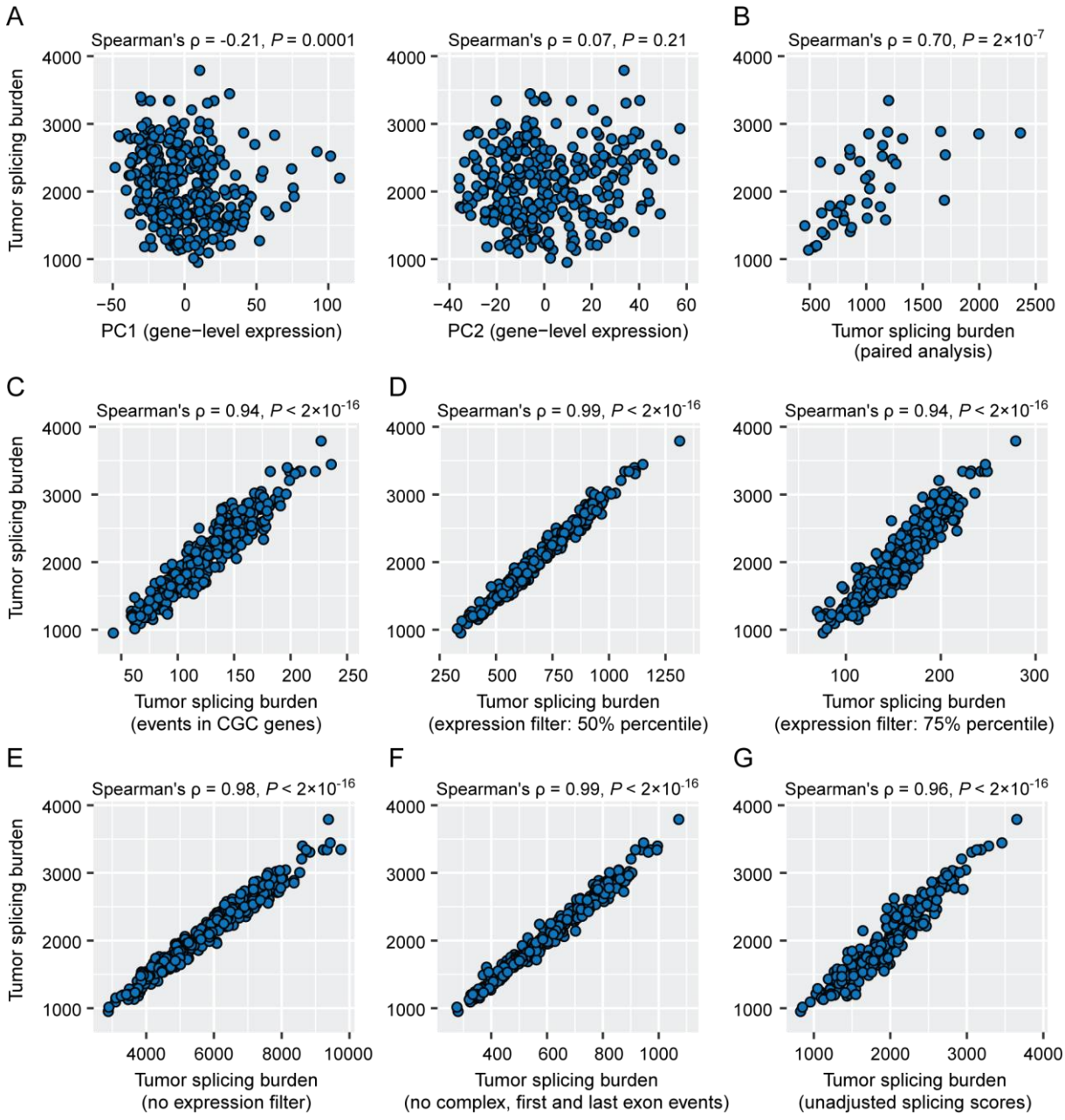


Fig. S9. Associations between the tumor splicing burden and selected parameters

A) Scatter plot of single tumor samples per patient comparing the tumor splicing burden with PC1 (left) and PC2 (right) from gene-level PCA analysis. **B)** Scatter plot of the 42 tumors with patient-matched normal samples comparing the tumor splicing burden with results from paired analysis. Scatter plots of single tumor samples per patient showing the impact of different event filtering criteria on the tumor splicing burden, including filtering based on **C)** target genes included in the Cancer Gene Census (CGC), **D)** more stringent relative expression thresholds, **E)** no expression threshold, and **F)** splicing event category. **G)** Scatter plot of single tumor samples per patient comparing the tumor splicing burden with results from analysis of splicing scores not adjusted for batch effect.

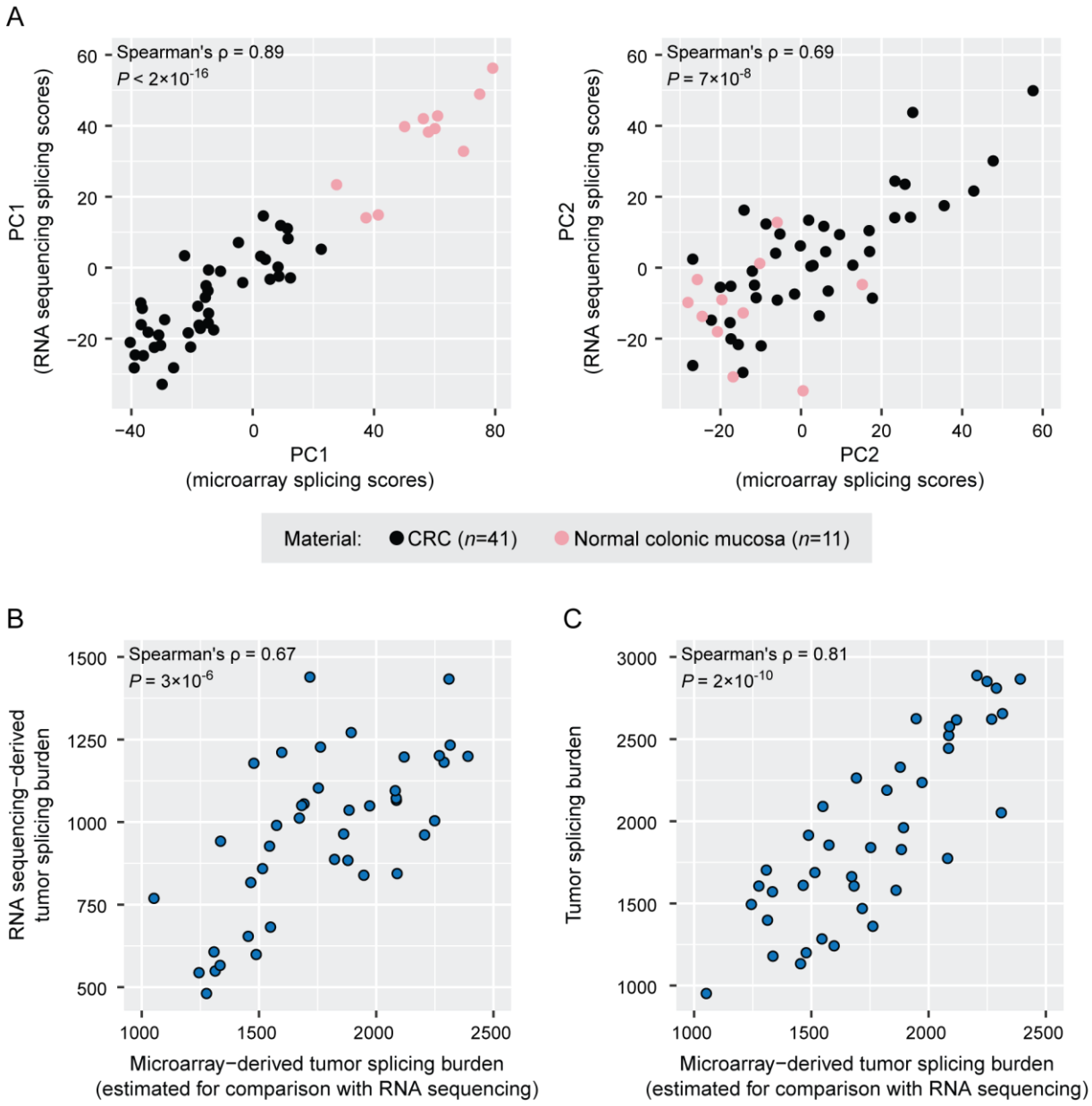


Fig. S10. Cross-platform correspondence of the tumor splicing burden

Scatter plots of the subset of tumor and normal samples analyzed by both microarrays and RNA sequencing showing the cross-platform correspondence in **A)** PC1 (left) and PC2 (right) values from PCA of splicing scores, as well as **B)** tumor splicing burdens estimated from divergence analyses with default beta parameter (0.95) and the 11 matched normal samples as reference. **C)** Scatter plot of microarray-derived tumor splicing burdens calculated with default beta parameter (for cross-platform comparison) versus non-default beta parameter (1.0, in the larger sample series) among the 41 CRCs analyzed on microarrays and RNA sequencing.

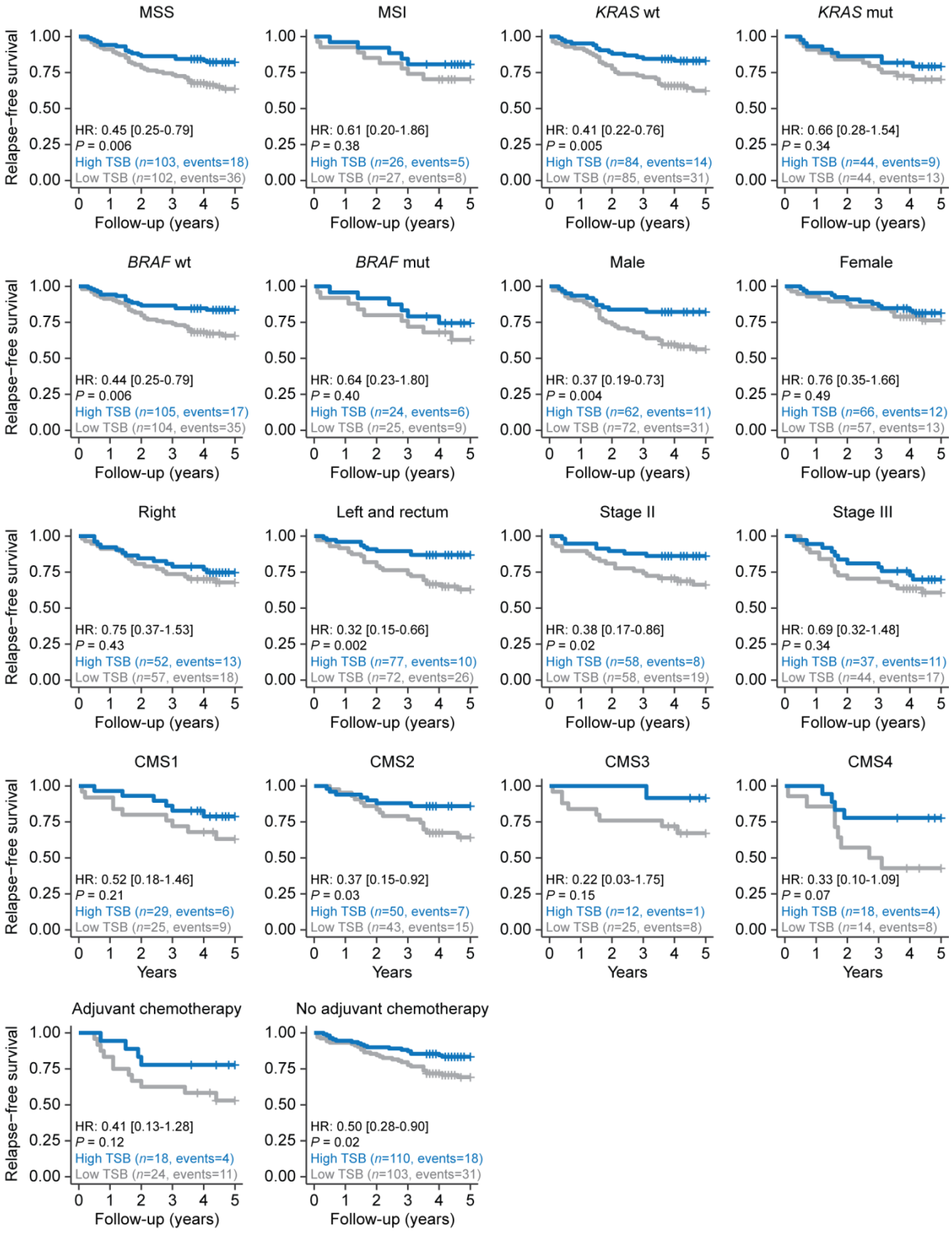


Fig. S11. Association of the tumor splicing burden with 5-year relapse-free survival within subgroups of CRC

Kaplan-Meier plots for different subsets of stage I-III CRC patients showing the association between the tumor splicing burden (TSB) and 5-year relapse-free survival.

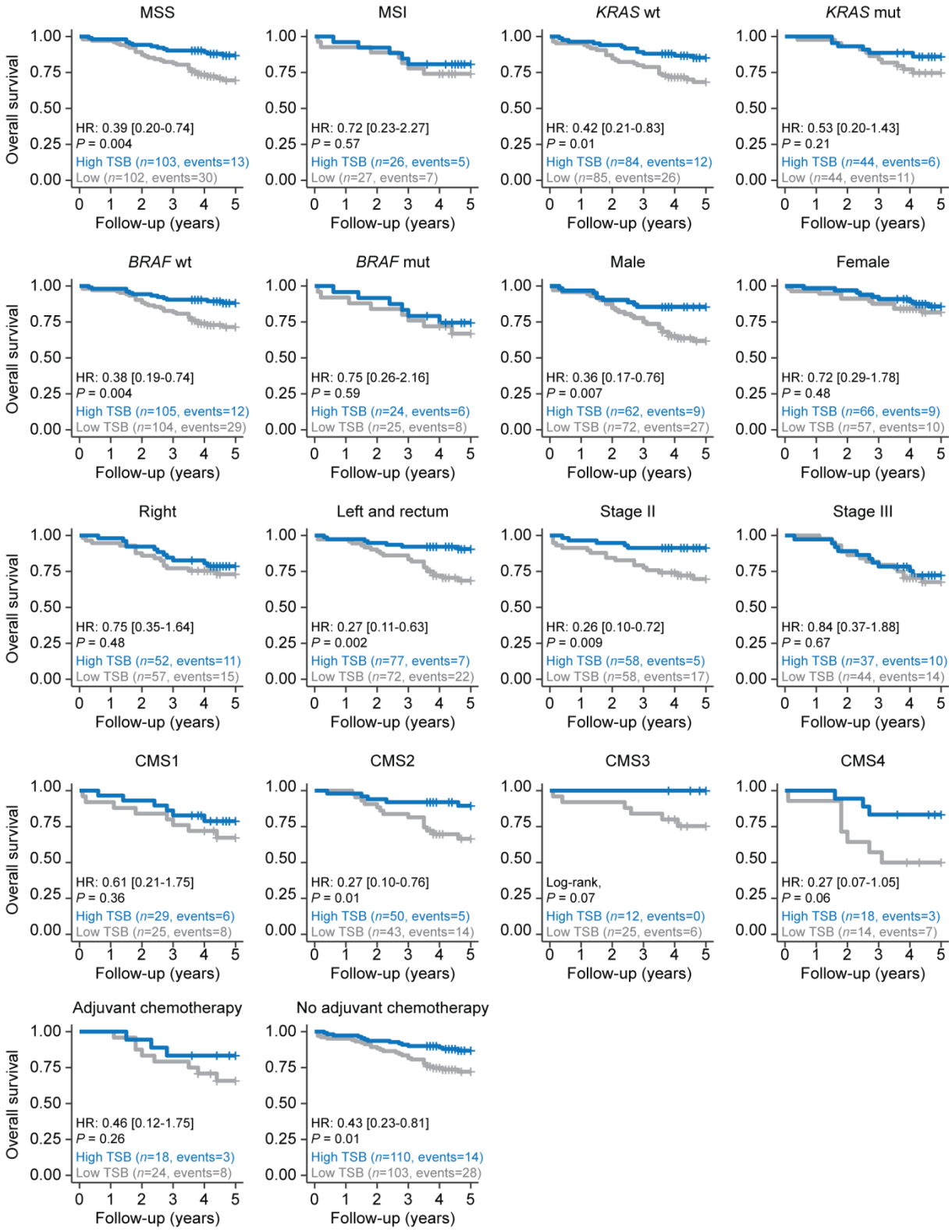


Fig. S12. Association of the tumor splicing burden with 5-year overall survival within subgroups of CRC

Kaplan-Meier plots for different subsets of stage I-III CRC patients showing the association between the tumor splicing burden (TSB) and 5-year overall survival.

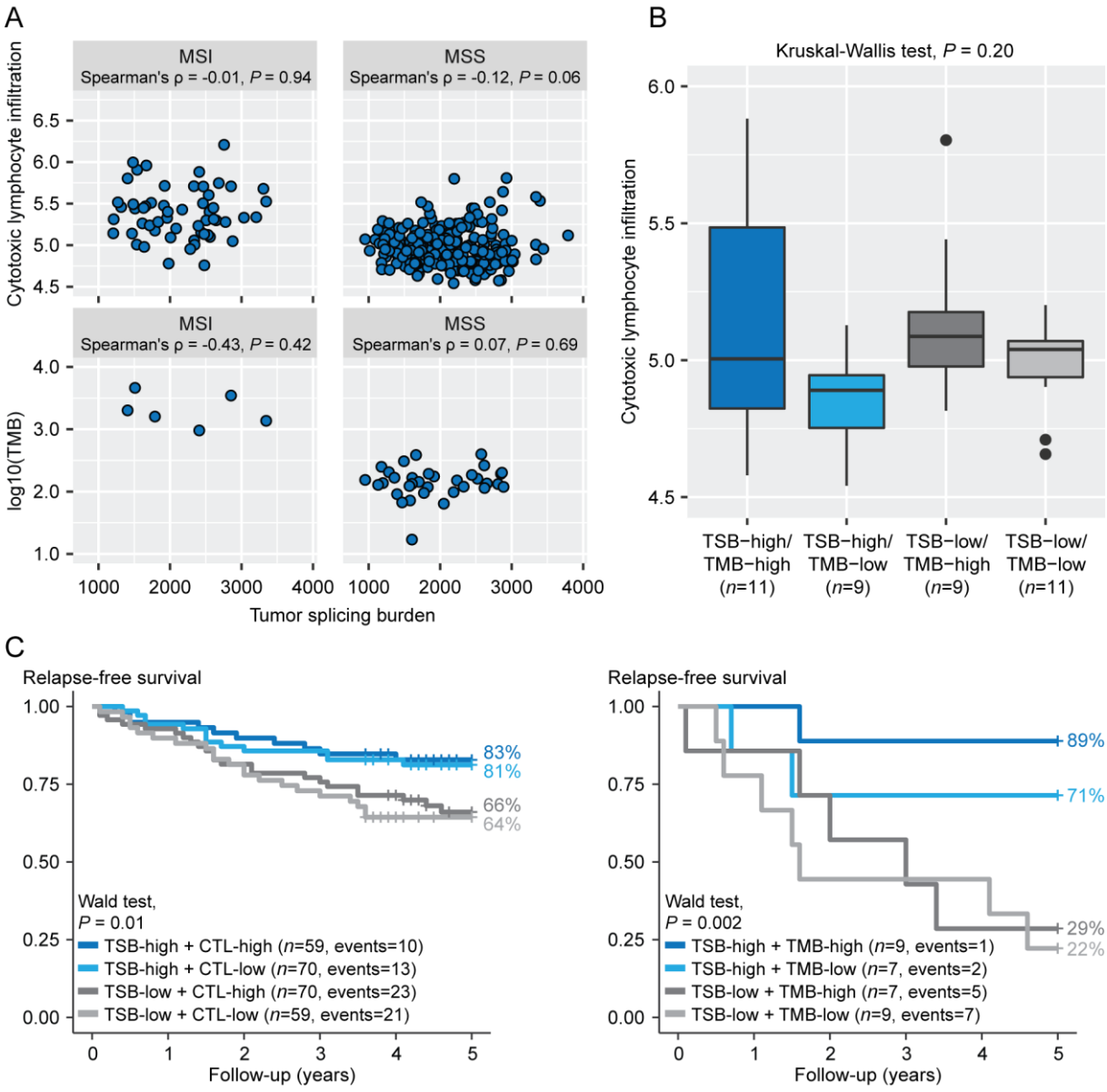


Fig. S13. Tumor splicing burden, immune cell infiltration and tumor mutation burden

A) Scatter plots (top) of the tumor splicing burden versus cytotoxic lymphocyte infiltration in MSI and MSS tumors in the in-house series (single samples per patient). The bottom scatter plots compare the tumor splicing and mutation burden in the subset of MSI (n = 6) and MSS (n = 34) tumors analyzed by whole-exome sequencing. **C)** Box plot of tumors analyzed by whole-exome sequencing showing the level of cytotoxic lymphocyte infiltration in each sample, grouped according to their median tumor splicing and mutation burden. **C)** Kaplan-Meier plots of 5-year relapse-free survival for stage I-III CRCs stratified by the tumor splicing burden and either cytotoxic lymphocyte infiltration (left) or the tumor mutation burden (right). All variables were dichotomized according to their median within the patient subsets. The *P*-values are from Wald tests for the multivariable Cox models including the tumor splicing burden and either cytotoxic lymphocyte scores or tumor mutation burdens, all as continuous variables. CTL: cytotoxic T lymphocytes, TMB: tumor mutation burden, TSB: tumor splicing burden.

Fig. S14. Splicing heterogeneity among multiregional samples from each of 98 CRCs

A) Box plot of inter- and intra-tumor splicing heterogeneity estimates (Euclidean distances) among tumors with multiregional sampling. **B)** Scatter plot of intra-tumor heterogeneity estimates (Euclidean distances) on the splicing versus gene expression levels. Intra-tumor heterogeneity on the splicing and gene expression levels was calculated as the mean of all pairwise Euclidean distances between any two samples from the same tumor, whereas the pairwise distance between any two samples from different tumors was used as a measure of inter-tumor heterogeneity.

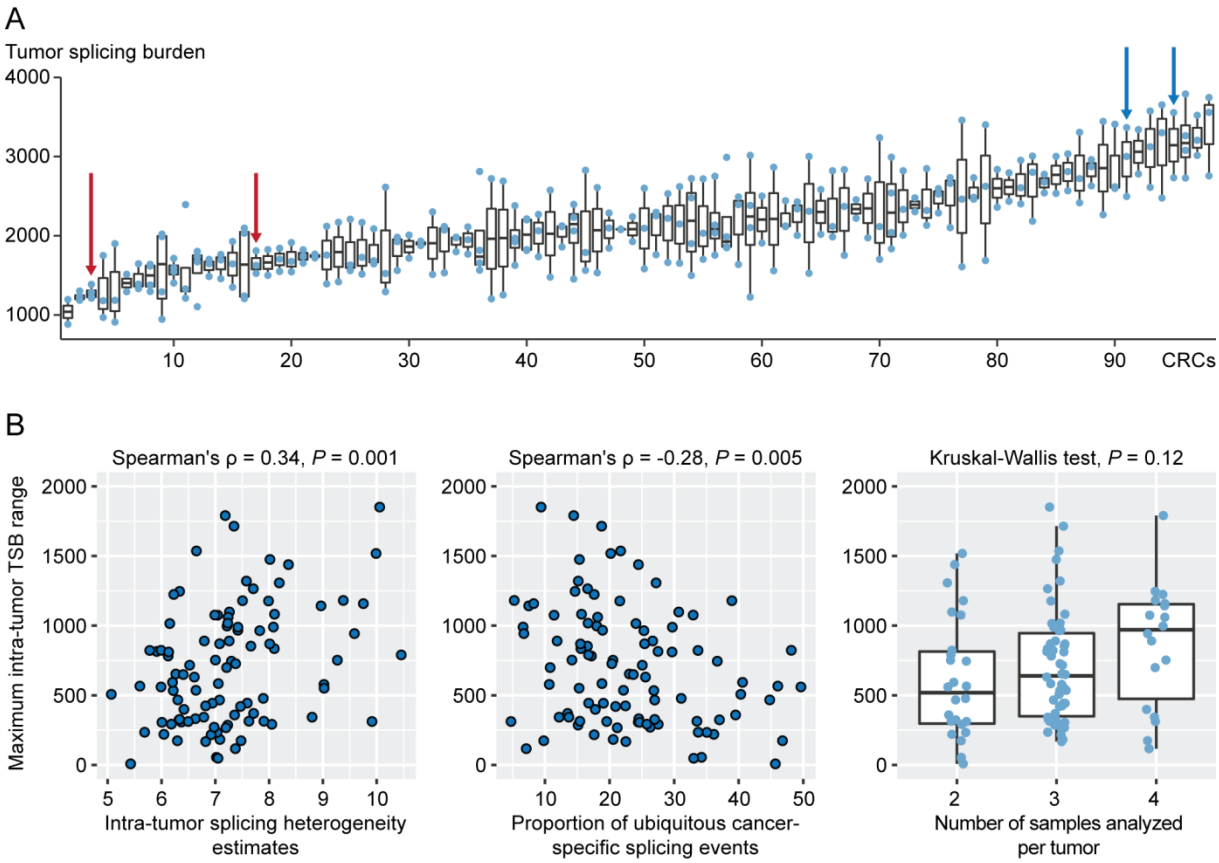


Fig. S15. Intra-tumor heterogeneity of the tumor splicing burden among the 98 CRCs with multiregional sampling

A) Box plot of the tumor splicing burden of multiregional tumor samples. Each box represents one tumor and each light blue dot represents one sample. Two pairs of synchronous tumors are indicated by red and blue arrows, respectively. **B)** Scatter plots and box plot showing the association of the maximum range in the tumor splicing burden among multiregional samples with intra-tumor splicing heterogeneity estimates (Euclidean distances; left), the proportion of ubiquitously expressed cancer-specific events (present in all samples of a given tumor; middle), and the number of samples analyzed per tumor (right).

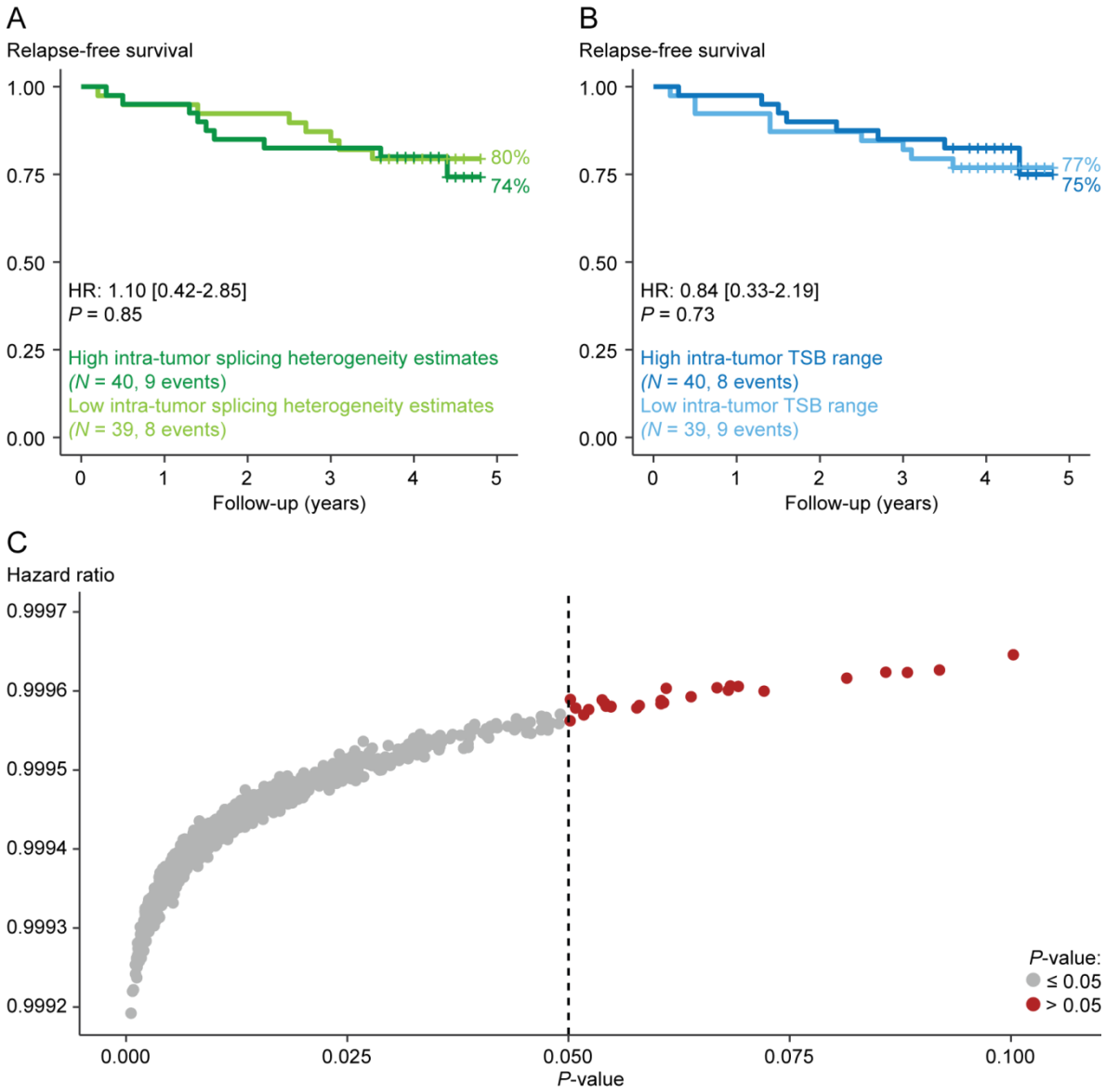


Fig. S16. Association of intra-tumor splicing heterogeneity with 5-year relapse-free survival

Among patients with stage I-III CRCs with multiregional sampling, neither **A)** the intra-tumor splicing heterogeneity estimates (Euclidian distances) nor **B)** the intra-tumor TSB range were associated with patient outcome. Both splicing heterogeneity measures were dichotomized into a high and low group according to their median within the patient subset. **C)** Scatter plot of the univariable hazard ratio and associated *P*-value of 1,000 survival models of the tumor splicing burden, analyzed as a continuous variable among all patients treated by complete resection of stage I-III CRC (n = 258), with random sampling of one tumor sample from each patient in the multiregional sample set. The tumor splicing burden was significantly associated with a favorable 5-year relapse free survival rate ($P \leq 0.05$) in 973 of 1,000 survival models. TSB: tumor splicing burden.

Supplementary References

1. Sveen A, Johannessen B, Tengs T, Danielsen SA, Eilertsen IA, Lind GE, et al. Multilevel genomics of colorectal cancers with microsatellite instability-clinical impact of JAK1 mutations and consensus molecular subtype 1. Genome Med. 2017;9:1-16.
2. Sveen A, Bruun J, Eide PW, Eilertsen IA, Ramirez L, Murumagi A, et al. Colorectal cancer consensus molecular subtypes translated to preclinical models uncover potentially targetable cancer cell dependencies. Clin Cancer Res. 2018;24:794-806.
3. Eilertsen IA, Sveen A, Stromme JM, Skotheim RI, Nesbakken A, Lothe RA. Alternative splicing expands the prognostic impact of KRAS in microsatellite stable primary colorectal cancer. Int J Cancer. 2019;144:841-7.
4. Eilertsen IA, Moosavi SH, Stromme JM, Nesbakken A, Johannessen B, Lothe RA, et al. Technical differences between sequencing and microarray platforms impact transcriptomic subtyping of colorectal cancer. Cancer Lett. 2020;469:246-55.
5. Sveen A, Johannessen B, Eilertsen IA, Røsok BI, Gulla M, Eide PW, et al. The expressed mutational landscape of microsatellite stable colorectal cancers. Genome Med. 2021;13:1-18.
6. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24-6.
7. Romero JP, Muniategui A, De Miguel FJ, Aramburu A, Montuenga L, Pio R, et al. EventPointer: an effective identification of alternative splicing events using junction arrays. BMC Genomics. 2016;17:1-18.
8. Romero JP, Ortiz-Estevez M, Muniategui A, Carrancio S, de Miguel FJ, Carazo F, et al. Comparison of RNA-seq and microarray platforms for splice event detection using a cross-platform algorithm. BMC Genomics. 2018;19:1-14.
9. Goldstein LD, Cao Y, Pau G, Lawrence M, Wu TD, Seshagiri S, et al. Prediction and Quantification of Splice Events from RNA-Seq Data. PLoS One. 2016;11:1-18.
10. Sandberg R, Larsson O. Improved precision and accuracy for microarrays using updated probe set definitions. BMC Bioinformatics. 2007;8:1-8.
11. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114-20.
12. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15-21.
13. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics. 2010;11:1-9.
14. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A. 2004;101:4164-9.