Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks

Endre Sebestyén[1,*], Babita Singh[1,*], Belén Miñana[1,2], Amadís Pagès[1], Francesca Mateo[3], Miguel Angel Pujana[3], Juan Valcárcel[1,2,4], Eduardo Eyras[1,4,5]

[1]Universitat Pompeu Fabra, Dr. Aiguader 88, E08003 Barcelona, Spain

[2]Centre for Genomic Regulation, Dr. Aiguader 88, E08003 Barcelona, Spain

[3]Program Against Cancer Therapeutic Resistance (ProCURE), Catalan Institute of Oncology (ICO), Bellvitge Institute for Biomedical Research (IDIBELL), E08908 L'Hospitalet del Llobregat, Spain.

[4]Catalan Institution for Research and Advanced Studies, Passeig Lluís Companys 23, E08010 Barcelona, Spain

[*]Equal contribution

[5]Correspondence to: eduardo.eyras@upf.edu

Abstract

Alternative splicing is a molecular mechanism regulated by RNA-binding proteins and affecting most eukaryotic genes. However, its role in human diseases, including cancer, is only starting to be unveiled. We systematically analyzed the mutation, copy number and gene expression patterns of 1348 RNA-binding protein (RBP) genes in 11 solid tumor types, together with alternative splicing changes in these tumors and the enrichment of binding motifs in the alternatively spliced sequences. Our comprehensive study reveals widespread alterations in the expression of RBP genes as well as novel mutations and copy number variations that are associated with multiple alternative splicing changes in cancer drivers and oncogenic pathways. Remarkably, breast and other tumors recapitulate splicing patterns similar to undifferentiated cells. These patterns, mainly controlled by MBNL1, involve multiple cancer drivers, including the mitotic gene *NUMA1*. We show that *NUMA1* alternative splicing contributes to enhanced cell proliferation and induces centrosome amplification in non-tumorigenic mammary epithelial cells. Our study uncovers novel splicing networks that potentially contribute to cancer development and progression.

Keywords:

Alternative splicing, RNA binding proteins, splicing factors, cancer

Introduction

Alternative splicing alterations are emerging as important signatures to understand tumor formation and uncover new therapeutic strategies[1]. Specific alternative splicing changes that provide a selective advantage to the tumor cells[2] may be caused by mutations in splicing regulatory sequences[3] and/or regulatory factors[4]. Various splicing factors have been described to be mutated in tumors, including *SF3B1*, *SRSF2*, *ZRSR2*, *U2AF1* in myelodysplastic syndromes and lymphoid leukemias[5], *RBM10* and *U2AF1* in lung tumors[4 6] or *SF3B1* in breast tumors[7]. These mutations generally impair the recognition of key regulatory sites, thereby affecting the splicing of multiple genes, including oncogenes and tumor suppressors[8]. On the other hand, increasing evidence shows that changes in the relative concentration of splicing factors can also trigger oncogenic processes. For instance, splicing factors from the SR-protein[9 10] and hnRNP[11 12] families are overexpressed in multiple tumor types and induce splicing changes that contribute to cancer proliferation. Similarly, downregulation of splicing factors has also been observed, like *RBM4*[13] and *QKI*[14], which have been proposed to act as tumor suppressors.

Importantly, specific alternative splicing events can substantially recapitulate cancer-associated phenotypes linked to mutations or expression alterations of splicing factors. This is the case of *NUMB*, for which the reversal of the splicing change induced by *RBM10* mutations in lung cancer cells can revert the proliferative phenotype[2]. A similar example is *S6K1*, where expression of isoform-2 is sufficient to reverse the transformation of immortal rodent fibroblasts caused by the overexpression of SRSF1 in vitro and in vivo[9]. Events that contribute to cancer are often controlled by multiple factors, like the exon skipping event of *MST1R* involved in cell invasion, which is controlled by SRSF1[15], hnRNPA2B1[12], hnRNPH1 and SRSF2[16]. Furthermore, some events may be affected by both mutations and expression changes in splicing factors. For instance, mutations in *RBM10*[2] or downregulation of *QKI*[14] lead to the same splicing change in *NUMB* that leads to cell proliferation. Alternative splicing changes that potentially characterize and contribute to the pathophysiology of cancer are thus triggered by alterations in a complex network of RNA binding proteins. However, the complete set of these alterations and how they may globally affect alternative splicing in cancer remain to be comprehensively described.

To elucidate the alterations in regulatory factors leading to the alternative splicing changes that may contribute to cancer, we analyzed RNA and DNA sequencing data from The Cancer Genome Atlas (TCGA) project for 11 solid tumor types. We performed a systematic analysis of the expression, mutation, and copy number alterations for 1348 genes encoding known and putative RNA binding proteins, and analyzed the alternative splicing changes potentially associated to these alterations. Our study reveals novel splicing networks involving RNA binding protein genes, whose alterations are predicted to trigger multiple alternative splicing changes in different tumor types that are potentially relevant for the development and progression of cancer.

Results

Expression changes in splicing factors separate tumor types and suggest new subtypes

Using data from TCGA for 11 solid tumors (Supp. Table S1) (Methods), we analyzed the differential mRNA expression between normal and tumor sample pairs of 1348 genes encoding known and predicted RNA binding proteins (RBPs) (Supp. Table S2) (Methods). The majority of them (1143, 84,8%) show significant differential expression in at least one tumor type (Supp. Fig. 1) (Supp. File 1). Examining in detail a subset of 162 RBPs annotated as known or putative splicing factors (SFs), 132 (80%) of them are differentially expressed in at least one tumor type, with approximately the same number showing up- and downregulation (Fig. 1a). A number of SFs show frequent downregulation, including *KHDRBS2*, *MBNL1*, *RBFOX*, *RBMS3*, *SRRM4*, and *QKI* (Fig. 1a); whereas *ELAVL2*, *IGF2BP*, *PABPC1*, *PABPC3*, *RBM28*, *SNRPA* and *SRRM3* and show frequent upregulation. Additionally, *WT1* is strongly up or downregulated in the majority of tumors studied. Expression changes are similar between breast invasive carcinoma (BRCA) and prostate adenocarcinoma (PRAD), between colon adenocarcinoma (COAD) and lung squamous cell carcinoma (LUSC), and between lung adenocarcinoma (LUAD) and head and neck squamous cell carcinoma (HNSC) (Fig. 1a). Kidney renal clear cell (KIRC) and papillary cell (KIRP) carcinomas also show similar patterns, which are frequently opposite to the other tumor types.

Notably, only 29 (21.9%) of the 132 differentially expressed SFs were previously associated with oncogenic or tumor suppressor activities (labeled in red in Fig. 1a) (Supp. Table S3), and some of the previously reported patterns were not detected. For instance, although *SRSF5* was described as oncogenic[17], it is downregulated in 6 tumor types; and *TRA2B*, reported as oncogenic in breast cancer[18], is upregulated in LUSC but downregulated in KICH and thyroid carcinoma (THCA). Moreover, the oncogenic *SRSF2*, *SRSF3* and *SRSF6*[10,19,20] are downregulated in KICH, while the oncogenic *SRSF1*[16] and the tumor suppressor *RBM4*[22] do not show any significant expression changes. Interestingly, new patterns emerge, including upregulation of genes from the RBM family, *RBM28*, *RBM15*, *RBM39* and *RBM41*, and downregulation of the genes from the *MBNL* family, *MBNL1*, *MBNL2* and *MBNL3* (Fig. 1a).

Unsupervised clustering of the entire set of 4442 tumor samples using normalized expression values per sample for SFs (Fig. 1b, Supp. Fig. 2) or for all RBPs (Supp. Fig. 3) (Methods) largely separates samples by tumor type. In fact, a similar result is achieved when using a different gene set of similar size (Supp. Fig. 3), indicating that the *RBP* expression patterns reflect intrinsic properties of the tumor types and their tissue of origin[23]. LUAD and LUSC samples separate into two large subgroups, with the majority of LUSC samples showing frequent upregulation of *TRA2B* (Fig. 1b, Supp. Fig. 2). KIRC and KIRP samples tend to cluster together and separately from KICH. The prostate adenocarcinoma (PRAD) and HNSC samples cluster closely and show a general pattern of low expression variation. Interestingly, a group of BRCA samples cluster separately from the rest of BRCA samples and close to LUSC and COAD tumors. Closer inspection reveals that the expression patterns of SFs largely reproduce BRCA and COAD subtypes (Supp. Figs. 4 and 5). *MBNL1* is frequently downregulated in COAD and BRCA samples, whereas *MBNL2* is specifically downregulated in BRCA samples. Moreover, *ESRP1* shows specific upregulation in BRCA basal samples (Fisher test p-value = 7.887E-08), which may be related to the general worse prognosis of basal tumors, as *ESRP1* expression promotes a CD44 isoform that induces metastasis of breast tumor cells to the lung[24]. Collectively, expression analyses indicate that RBP genes are frequently and specifically altered in human cancer.

Patterns of RBP mutations and copy-number variation across tumors

To further define the extent to which RBP genes are altered in human cancer, the TCGA data was analyzed for protein-affecting mutations and copy number variations (CNVs) (Methods). As for gene expression, most of the RBP genes are mutated in tumors, with only 10 (0.7%) showing no mutations in any of the samples tested. However, they are mutated in a smaller proportion of samples compared to cancer drivers and to other genes (Fig. 1c) (Supp. File 1) (Methods).

We confirmed 7,6% (35/458) of LUAD tumor samples to have protein-affecting mutations for *RBM10*, in agreement with previous studies[6]. Using this case as reference, 182 (13,5%) of all RBPs (10 (6%) of the 162 SFs) were mutated in more than 7% of samples in a given tumor type (Table 1) (Supp. File 1). In general, there is a weak correlation between mutations and expression changes of the corresponding genes (Table 1) (Supp. Table S4) (Methods). On the other hand, we tested the mutual exclusion of mutations and expression changes, as they both may have similar functional impact. The top cases include *SYNE1* in COAD and *NBPF10* in LUAD (Table 1) (Supp. Table S4) (Methods). In contrast to mutations, CNVs are more recurrent across samples; with gains more frequent than losses (Fig. 1c). CNV gains show frequent association with upregulation, the strongest ones including *TRA2B* in LUSC, *ESRP1* in BRCA, and *RBM39* and *SRSF6* in COAD (Table 1) (Supp. Table S5). *SRSF6* amplifications were observed before in colon tumors[25] and *RBM39* overexpression was linked before to breast cancer progression[26]. On the other hand, deletions show weaker associations with downregulation, the most frequent ones being *RBFOX1* in COAD and *RBMS3* in LUSC (Table 1) (Supp. Table S6). These analyses highlight the potential relevance of expression alterations of RBPs, besides mutations and copy number variation, in shaping the tumor phenotypes.

Patterns of alternative splicing alteration in tumors

Next, we investigated the patterns of differential splicing that may occur as a consequence of the alterations described above. To determine those possibly related to expression alterations in RBPs, we first evaluated the significant splicing changes in tumors compared to normal tissue classified into 5 major event types: skipping exon (SE), alternative 5' splice-site (A5), alternative 3' splice-site (A3), mutually exclusive exon (MX) and retained intron (RI) events (Fig. 2a) (Supp. File 2) (Methods). There

was no clear relationship between the number of differentially spliced events and the number of paired samples used and the proportions for each event type are similar across tumor types with SE events being the most abundant (average 58%) (Supp. Table S7).

As alternative splicing in cancer drivers may contribute to the oncogenic process[1], we investigated their patterns of differential splicing. We collected a list of 937 genes, collectively called drivers, which either show alternative splicing in relation to cancer (82 cases) (Supp. Table S8) or are predicted as drivers based on mutations and CNVs (889 cases, 34 in common with previous set) (Supp. Table S9) (Methods). 653 (69.7%) of these 937 drivers have annotated alternative splicing and 292 (31.2%) have at least one differentially spliced event in these tumors (Supp. File 2). Moreover, comparing the splicing changes in drivers and non-drivers, 7 of the 11 tumors (Fig. 2b, in red) show enrichment of differentially spliced events in drivers, with a number of them occurring in multiple tumor types (Supp. Table S10) (Supp. Fig. 6). To further characterize the differentially spliced events, we evaluated whether specific cancer hallmarks from MSigDB[28] are enriched in differentially spliced events (Methods). Interestingly, various hallmarks are enriched according to differential splicing but not to differential expression, including Mitotic spindle, WNT/Beta-catenin, UV response, and Notch signaling (Fig. 2c) (Supp. Fig. 7) (Supp. Table S11). These results suggest possible relevant mechanisms of alternative splicing in tumors involving genes and pathways that are independent of expression alterations.

To determine the splicing changes possibly related to mutations in RBPs, we compared the inclusion levels (percent spliced in, PSI) of the events between samples with or without protein-affecting mutations for each RBP (Supp. Fig. 8) (Methods). In LUAD, we found 53 (enrichment z-score = 52.59) and 83 (z-score = 82.59) differentially spliced events associated to *RBM10* and *U2AF1*, respectively (Figs. 2d and 2e) (Supp. Fig. 9). Additionally, we found *HNRNPL* (42 events, z-score = 41.45) in COAD, with 13 of the 16 mutations consisting of deletions or insertions in a specific position of the second RNA recognition motif that cause a frameshift (Fig. 2f) (Supp. File 3). In LUAD we also found the transcription factor and predicted RBP *TCF20*[29] (50 events, z-score = 49.59), with most of the mutations consisting of an insertion in a Glycine-rich region at the N-terminus (Supp. Fig. 9). In COAD we also found the predicted RBP *MACF1*[30] (84 events, z-score = 83.45) with mutations along

the entire protein (Supp. Fig. 9). We analyzed *SF3B1*[7] as comparison and found 64 differentially spliced events (z-score = 63.66) in BRCA, despite being mutated in only 1.7% of the tested samples (16 out of 956 for which we had mutation and RNA-Seq data).

Notably, compared with the proportions of the different event types, there is a significant enrichment of A3 (Fisher's test p-value = 1.45E-11) events associated with *SF3B1*, of A5 (p = 2.09E-6) events for *TCF20*, and of SE events for *RBM10* (p = 0.003) (Fig. 2g) (Supp. Fig. 9). There is also a significant depletion of SE (p = 0.001) events for *TCF20*, SE events (p = 2.24E-8) for *SF3B1*, and of A5 events for *U2AF1* (p = 0.001) and *HNRNPL* (p = 0.005) (Fig. 2g) (Supp. Fig. 9) (Supp. Table S12). Furthermore, we confirmed 17 (20%) of the previously detected differentially spliced genes for *SF3B1*[7], 32 (38%) for *U2AF1*[4] and 21 (30%) for *RBM10*[4] (Supp. Table S13). Although we found no clear enrichment of any particular cancer hallmark from MSigDB[28] (Supp. Table S11), some of the alternatively spliced events associated with mutations in RBPs occur in cancer drivers (Table 2) (Supp. File 3). *HNRNPL* mutations are associated with an A5 event in *CASP8* (Fig. 2h), a gene involved in programmed cell death[31]. As hnRNPL was related before to the regulation of CASP9 splicing[32], this result suggests a possible role in apoptosis. Interestingly, *MACF1* and *NBPF10* mutations associate to the same event of *RAC1* in the direction of skipping of the Rac1b isoform associated to high WNT-pathway activity in colon tumors[33] (Supp. Fig 9). *MACF1*, which also appears differentially spliced in lung tumors[34], is itself a component of the WNT-pathway, which is also known to affect RNA processing and splicing[35]. On the other hand, *NBPF10* was predicted before to be a tumor driver[36], but its role in tumors is not yet known. Our analyses reveal a rich source of new information about alternative splicing events associated with RBP alterations with potential relevance in cancer.

Common and specific patterns of differential splicing in tumors

Our results suggest that mutations in RBPs may not be the main cause of splicing changes in tumors. We thus decided to characterize further the splicing changes between tumor and normal samples to determine their association to RBPs. We first identified common patterns of splicing change between tumors, by selecting those

events with a strong correlation with a differentially expressed SF in at least two tumor types (Methods). The changes in PSI (ΔPSI) for these common events show high correlation between tumor pairs (Fig. 3a) and indicate potential common regulators (Fig. 3b) (Supp. Fig. 10) (Supp. Table S14). For example, BRCA and LUAD have 229 common events associated to various factors, including *SRSF5* and *QKI* (Fig. 3b, upper left panel); whereas HNSC and LUSC have 141 common events, 63 of them associated to *RBM38* (Fig. 3b, upper right panel). PRAD and KIRC share 78 anti-correlating events associated to *ESRP2* and *MBNL1* (Fig. 3b. lower left panel). Interestingly, *RBM47*, recently described as a tumor suppressor[37], appears as the main common SF between KIRC and HNSC with 61 associated events (Fig. 3b, lower left panel). These results suggest an association between some of the RBPs in tumors and the splicing changes detected. On the other hand, these results also raise the question of whether there may be tumor specific events. To address this, we used an entropy-based feature selection approach to identify events that separate between tumor types according to the PSI values (Methods). This produced 380 events that largely separate the 4442 tumor samples by type (Fig. 3a) (Supp. Fig. 11) (Supp. Table S15). These splicing changes may be indicative of oncogenic mechanisms that are tumor-type specific.

Enriched RBP motifs in differentially spliced events

To further understand the link between the observed alternative splicing patterns and RBPs, we tested the enrichment of putative binding motifs in differentially spliced events. We assigned binding motifs from RNAcompete[38] to 104 of the analyzed RBPs and tested enrichment by comparing the motif frequencies between differentially and non-differentially spliced events on and around the variable region (Methods) (see specific diagram in Supp. Fig. 12). Nearly all differentially expressed RBPs from this set show motif enrichment (average 92.8% across tumor types) (Supp. Fig. 12). Motifs associated to non-differentially expressed factors are also frequently enriched, probably due the similarities between motifs (Supp. Fig. 13). We thus considered the enriched motifs of differentially expressed *RBPs* in the same tumor type to be the functionally important ones (Fig. 4a). CELF, MBNL, RBFOX families are among the most frequently enriched in the different tumor types, as well as in luminal breast tumors (Supp. Figs. 14-18). Moreover, downregulated RBPs in inclusion events and upregulated *RBPs* in skipping events are more frequently enriched in upstream and

exonic regions, consistent with the positional effects proposed for splicing factors [39]. On the other hand, downregulated *RBPs* in skipping events and upregulated *RBPs* in inclusion events are most frequently enriched on exons, suggesting that RBPs more often enhance the inclusion of exonic regions to which they bind.

To define candidate target events for RBPs, we selected differentially spliced events whose PSI correlates with *RBP* expression ($|R| > 0.5$, Spearman) and that contain the corresponding RNA binding motif for differentially expressed RBPs (Methods). We could assign significantly 20-80% of the differentially spliced events to at least one RBP (Fig. 4c, upper panel) (Supp. File 4). Some of the RBPs with assigned targets were previously associated to cancer: RBFOX2 in PRAD (23.3% of events); ESRP1 in KIRC (24.3% events) and PRAD (17.8% events), QKI in KICH (8.9%) and LUAD (5.5%), PTBP1 in COAD (23.4%) and LUSC (16.1%), and RBM47 in THCA (18.4%) (Fig. 4c, lower panel) (Supp. Table S16). Interestingly, TRA2B, whose motif is enriched in LUSC, is linked to 6% of the differentially spliced events, including an SE event in the DNA damage response gene CHEK1, reported recently to be controlled by Tra2 proteins[40]. Additionally, MBNL1 appears relevant in COAD (15.3% events), PRAD (14%) and BRCA (8.8%). We confirm the role of MBNL1, QKI, RBFOX2, PTBP1, RBM47 and ESRP1 in some of these tumors by comparing the ΔPSI values of differentially spliced events with those obtained in knockdown or overexpression experiments of the individual RBPs in different cell lines (Supp. Figs. 19-21) (Supp. Table S17). As MBNL1 and MBNL2 depletion induces an undifferentiated state in cells[41] [42], we compared the tumor events with those differentially spliced between human embryonic stem cells (hESC) and differentiated cell lines or tissues[42]. We found a high positive correlation of ΔPSI values in BRCA, PRAD and LIHC, as well as in breast luminal tumors; to a lesser extent in COAD and LUAD, and an anti-correlation in KIRC, in agreement with the *MBNL1* and *MBNL2* expression patterns in these tumors (Fig. 4d, upper panels) (Supp. Fig. 22) (Supp. Table S18). Additionally, the majority of correlating events contain the MBNL binding motif. The potential major role of MBNL1 and MBNL2 in cancer is further highlighted by the fact that they both target a considerable proportion of cancer drivers in BRCA, PRAD (Fig. 4d, lower panels) (Supp. Fig. 23).

To further explore the role of *RBP* expression changes in splicing in tumors, we evaluated the potential association of RBP motifs on differentially spliced events,

taking into account motif redundancies (Methods). Motifs for IGF2BP2, IGF2BP3, PABPC3, PABPC5, RBM46, HNRNPC, CELF5, CPEB2 and CPEB4 associate frequently with each other in 4 or more tumor types (Fig. 4d). Similarly, motifs for CELF5 and RBFOX1 associate on differentially spliced events in 4 tumor types. In contrast, motifs for SRSF7, ZNF638, RBMS3, DAZAP1, among others, do not show any association. As these are differentially expressed and show motif enrichment, it suggests that they alone may control multiple splicing changes in cancer (Fig. 4d). These results highlights the prominent role of the expression changes in RBPs and splicing factors in the alterations of splicing in cancer and suggests new mechanisms of regulation.

Network analysis uncovers overlapping regulatory modules in cancer

To further investigate the synergy between RBPs in the splicing control of relevant pathways in cancer, we built clusters with the set of 162 SFs using the correlation between gene expression and event PSI values and linked these clusters to differentially spliced genes in the enriched hallmarks (Supp. Fig. 24) (Methods). Based on these clusters, we defined modules of splicing regulation with potential relevance in cancer (Figs. 5a and 5b) (Supp. Figs. 25 and 26). This analysis reveals that one or two factors as main regulator of a hallmark across different tumor types. Potential main regulators of myogenesis and epithelial-mesenchymal transition (EMT) include *ESRP1* (BRCA, LUAD, LUSC) and *MBNL1;* potential regulators of the apical junction complex include the genes from the *RBFOX* family (BRCA, KIRP and PRAD); and potential regulators of mitotic spindle include *MBNL* genes (THCA, PRAD) as well as *PTBP1* (LUSC), which is also related to myogenesis (COAD, LUSC) and EMT (LUSC). BRCA shows 3 modules (Fig. 5a): two mainly associated to EMT and myogenesis, and a third one including *RBFOX2*, mainly related to the apical junction complex. These regulatory modules also control two relevant hallmarks: the G2 checkpoint (G2M), which includes *NUMA1*, a gene involved in spindle formation during cell division[43]; and the WNT/Beta-catenin pathway, which includes *NUMB*, and inhibitor of the NOTCH pathway whose alternative splicing has been linked to cell proliferation[2] (ref) (Figs. 5c and 5d). COAD has 6 modules (Fig. 5b) with *PTBP1* as a potential main regulator of myogenesis. Additionally, angiogenesis, which is an enriched hallmark in COAD for splicing but not for gene expression, includes an event in *SERPINA5*, an inhibitor of serine proteases involve in

homeostasis and thrombosis[44], which we predict to be controlled by RBM47, PTBP1 and RBM28 (Fig. 5e). This analysis reveals new roles of RBPs and splicing in cancer-relevant processes.

MBNL1 contributes to cell proliferation through alternative splicing regulation of NUMA1

MBNL1 emerges as a relevant regulator of splicing for multiple cancer drivers, particularly in luminal breast tumors (Supp. Fig. 27). Events potentially controlled by MBNL1 include an exon skipping in *NUMA1*, which correlates with *MBNL1* expression (Spearman R = 0.65 in LumA and R = 0.66 in LumB) and has a MBNL1 motif downstream of the alternative spliced exon (Fig. 6a) (Supp. Fig. 28). The same event is significantly more included in KIRC (ΔPSI = 0.11, corrected p-value = 3.11e-06), where *MBNL1* is upregulated compared to normal tissues, providing further support for the dependence of *NUMA1* splicing on MBNL1. We detected MBNL1 protein in the breast epithelial cell line MCF10A, and the triple-negative cell lines MDA-MB-231, MDA-MB-436 and MDA-MB-468 but not in the luminal-like MCF7 (Supp. Fig. 29). To test the effect of the downregulation of MBNL1, we used siRNAs to deplete *MBNL1* in MCF10A cells. *MBNL1* depletion with two different siRNAs targeting exons 3 and 5 induces skipping of exon 16 in *NUMA1* measured by semi-quantitative RT-PCR, recapitulating the splicing pattern observed in the tumor samples (Fig. 6b, upper panels) and in MCF7 (Supp. Fig. 29). We also tested *NUMB* alternative splicing of exon 9, which we predict to be dependent on MBNL1 in BRCA luminal tumors. The depletion of *MBNL1* recapitulates the *NUMB* splicing pattern in luminal samples (Fig. 6b). For comparison, we evaluated the effect of *QKI*, which we also observe downregulated in BRCA luminal tumors and is detected in MCF10A cells but not in MCF7 cells (Supp. Fig. 29). Upon *QKI* depletion, *NUMA1* exon 16 inclusion changes in the direction opposite to that with *MBNL1* depletion to a small but reproducible extent (Fig. 6b, upper panel). Although we did not find a QKI motif on the *NUMA1* event, this is consistent with the low and negative correlation found with *QKI* gene expression (R = -0.11) in BRCA. We also tested *NUMB* alternative splicing of exon 9, which is regulated by QKI in lung tumors[14] and which we predict to be also affected by QKI in BRCA luminal tumors. The depletion of *QKI* induces

exon 9 inclusion and recapitulates the *NUMB* splicing pattern in luminal samples (Fig. 6b, middle panels).

To measure whether the depletion of *MBNL1* or the splicing change in *NUMA1* had any effect on cell proliferation, we designed 2'-O-methyl phosphorothioate-modified antisense oligonucleotides (AONs) targeting specifically the 5' and 3' splice-sites of *NUMA1* exon 16. As expected, these AONs promote exon skipping, recapitulating in the MCF10A epithelial cell line the splicing pattern observed in BRCA luminal tumors, with the AON against the 5'ss being more efficient (Supp. Fig. 30). We used a Resazurin-based assay[45] to measure the proliferation/viability of MCF10A cells transfected with the AONs targeting *NUMA1* exon 16, or with siRNAs against MBNL1 and QKI. We observed a significant increase in cell proliferation/viability at 72, 96 and 120 hours upon depletion of *MBNL1* or QKI compared with controls (t-test p-value < 0.05) (Fig. 6c) (Supp. Table S19), and when transfecting cells with the AON against the 5' splice site (t-test p-values < 0.05) (Fig. 6c). Using only the 3' splice site or both AONs we also measured certain increase, albeit not statistically significant.

We decided to study the possible effects of the alternative splicing of *NUMA1* exon 16 on centrosome amplification (Methods) (Supp. Table S20). Using the AON against the 5' splice-site that recapitulates the splicing form of *NUMA1* in luminal types, we observed a significant increase in number of cells with centrosome amplification compared with controls in MACF10 cells (Fig. 6d). Using the siRNA against *MBNL1* we could not detect any significant difference, which may be explained by the superposition of indirect effects. To further relate *NUMA1* alternative splicing to the fidelity of centrosome formation, we compared the PSI values with an expression signature for chromosome instability and aneuploidy[46]. We observe an inverse correlation between this signature and the inclusion of *NUMA1* event in luminal tumors (Fig. 6e), which is higher than for any other tumor type (Supp. Fig. 31). The exon skipping described in *NUMA1* is the only coding difference between the tumoral and the normal alternative splicing isoforms. While it is unclear whether the 14 amino acid change between the isoforms could explain the observed effects (e.g. we could not detect any protein domain or disordered regions[47]) using GPS[48] we predict loss of a high scoring threonine phosphorylation site (FDR $\leq$ 2%) upon exon skipping, suggesting a possible mechanism for the differential activities of the two isoforms

(Supp. Fig. 31) (Supp. Table S21).

Discussion

Our study reveals that expression changes in RBP genes are pervasive in cancer and characterize the different tumor types. Most of these expression changes are not related to deletions or amplifications, suggesting that they may originate from genetic alterations in the pathways that control the regulation of RBP genes. Despite the fact that most RBPs show mutations, they are mutated in a low proportion of samples and only a few are associated to genome-wide effects on alternative splicing. Additionally, the number of events potentially affected is low compared with all the splicing alterations that occur in tumors. Thus, expression changes in RBP genes may be the major contributor of the alternative splicing changes observed in tumors.

Some of the described RBPs were predicted before to be possible mutation drivers, but their mechanisms in cancer have not yet been described. One possibility is that they contribute to cancer through some of the alternative splicing changes predicted here, consistent with other studies[2,9]. An important implication for prognostic and clinical studies is that the definition of functional impact of somatic mutations should be expanded to include alterations in the alternative splicing of the gene targets.

Considering splicing alterations in predicted cancer drivers as a measure of the tumorigenic impact of alterations in RBPs, we identified various potentially relevant genes, including *MBNL1* in breast and prostate tumors, and *TRA2B* in lung squamous tumors. Squamous carcinomas (LUSC and HNSC) show frequent amplifications of *TRA2B*, but only LUSC shows overexpression of *TRA2B*, which would explain the differences in splicing between these two tumor types. The potential relevance of *TRA2B* in LUSC is further highlighted by the enrichment of its binding motif in differentially spliced events. Thus, despite the similar genetic alterations between the squamous tumors[49], their expression phenotypes are very different, partly due to the *TRA2B* overexpression.

Our analyses provide possible new roles for RBPs that had no clear involvement in splicing. For instance, *IGF2BP* genes are upregulated in multiple tumor types and their motifs are commonly enriched in differentially spliced events, but interestingly only on exonic regions. Although this does not provide rigorous evidence of their

involvement in splicing, it does support a role for IGF2BP in most of the tumors analyzed[50], an observation that deserves further mechanistic analyses. Similarly, *RBMS3*, which appears frequently deleted and downregulated, has no known role in splicing. *RBMS3* modulates the TGFβ signaling pathway post-transcriptionally[52] and was found depleted in esophageal squamous cell carcinoma[53] and lung squamous cell carcinoma[54]. Interestingly, *RBMS3* also has a nuclear role in regulating *MYC* [53], which has recently been related to the maintenance of splicing fidelity[55]. Its frequent downregulation in tumors, often in the absence of DNA deletions, suggest a general role for *RBMS3* in cancer, possibly related to splicing.

The splicing alterations detected are predicted to have an impact on many cancer hallmarks, some of them independently of changes in gene expression. This is for instance the case of angiogenesis in colon tumors, mitotic spindle in lung tumors, NOTCH signaling and WNT/Beta-catenin pathways in kidney tumors. This agrees with previous reports on the impact of splicing on cancer hallmarks on the basis of literature searches[56] and highlights the relevance of alternative splicing as a complementary molecular mechanism to explain tumor development.

We found that splicing changes in several tumor types and especially in breast luminal tumors recapitulate the splicing pattern of undifferentiated cells, mainly linked to *MBNL1* and *MBNL2*, in agreement with recent studies[41,42]. It was proposed recently that RBPs involved in development and differentiation could act as master splicing regulators[57]. These included *MBNL1*, *RBFOX2*, *RBM24*, *RBM38*, *RBM20*, *RBFOX1*, *ZNF638*, and *RBMS3*, which we found frequently downregulated in tumors. We hypothesize that the reversal of RNA splicing to an undifferentiated pattern through the deactivation of one or more RNA binding proteins may be a general mechanism of tumors.

MBNL1 potentially controls multiple genes that participate in cancer-related pathways, including the mitotic gene *NUMA1*, whose alternative splicing correlates with cell differentiation. Although NUMA1 locus was related before to breast cancer risk[58], a clear mechanism explaining its relevance in cancer is still lacking and its exon skipping event has not been characterized so far. We observed that *NUMA1* alternative splicing leads to higher proliferation and increased centrosome amplification in normal cells. *NUMA1* produces a protein component of the nuclear

matrix, which is dependent on threonine-phosphorylation to regulate the orientation of mitotic spindles and ensure symmetric cell division[43 59 60]. *NUMA1* alternative splicing removes a strong predicted threonine phosphorylation site; hence, one attractive possibility is this splicing change affects its phosphorylation, thereby impairing correct spindle positioning, leading to increased genome instability.

The results of this study provide a rich resource of information about novel networks of splicing factors and RBPs that trigger common and specific alternative splicing changes in several solid tumors and provide candidate alternative splicing changes that may be relevant to understand the molecular basis for - and potentially reverse - the oncogenic properties of tumor cells.

Methods

Datasets

Tumor types were selected from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) according to whether they had a sufficient number of RNA-Seq samples from tumor and paired normal samples at the time of start of the study: breast invasive carcinoma (BRCA)[61], colon adenocarcinoma (COAD)[62], head and neck squamous cell carcinoma (HNSC)[63], kidney chromophobe (KICH)[64], kidney renal clear cell carcinoma (KIRC)[65], kidney renal papillary carcinoma (KIRP) (https://tcga-data.nci.nih.gov/), liver hepatocellular carcinoma (LIHC) (https://tcga-data.nci.nih.gov/), lung adenocarcinoma (LUAD)[66], lung squamous cell carcinoma (LUSC)[67], prostate adenocarcinoma (PRAD) (https://tcga-data.nci.nih.gov/) and thyroid carcinoma (THCA)[68]. Processed RNA-Seq data for tumor and normal samples was downloaded for RNA-Seq version 2 Level 3 data with RSEM estimated read counts for genes and isoforms and the TCGA annotation (hg19, June 2011) were used. Mutation and copy number variation (CNV) data was downloaded from the TCGA data portal for all tumor types. For the mutation data, MAF files containing Level 2 somatic mutation calls from whole exome sequencing was used. For the CNV data, Level 3 SNP array data, containing the normalized copy number variation and purity/ploidy analysis results for each sample was used, excluding germline copy number variations. BRCA subtypes were determined using the classification from TCGA[61]. COAD subtypes were determined by counting the number of somatic

mutations per sample[62]. A sample was classified as hypermutated if it contained more than 250 mutations in total, and as non-hypermutated otherwise. To assess sample quality, tissue types were predicted with URSA[69] from the RSEM estimated read counts per gene, keeping only those predictions with posterior $P \geq 0.1$. Samples that did not cluster with the rest from the same class (either normal or tumor) were removed.

The 1336 genes coding for RNA binding proteins (RBPs) analyzed includes those with high confidence for RNA binding[29,30,70] and those with known RNA-binding activity from Ensembl[71] (Supp. Table S2). From this set, a subset of 162 known and potential splicing factors (SFs) was selected (Supp. Table S2). SF3B1 and the kinases SRPK1 and SRPK2 were included in the analyses for comparison. Additionally, 1426 genes coding for transcription factors (TFs) from the Animal Transcription Factor Database[72] were analyzed. TFs not present in the TCGA annotation or in the latest ENSEMBL release (v80), based on their HUGO or Entrez ids, were removed; and only TFs showing differential expression (absolute log fold change > 0.5 and adjusted p-value < 0.05) in at least one tumor type were kept for analysis.

Differential expression

Quantile normalization and voom transformation was performed on gene-level read-count data[73]. Differential expression analysis was performed using the empirical Bayes function from the limma package[74] and p-values were corrected for multiple testing using the Benjamini-Hochberg method. Genes were considered differentially expressed if they had an absolute $\log_2$-fold change > 0.5 and corrected p-value < 0.05. A robust Z-score per gene and per tumor sample was calculated using the quantile normalized and mean-variance relationship corrected voom transformed read-counts, based on the values in the tumor sample (n) and the median (m) and median absolute deviation (MAD) values in the normal samples for the same tumor type:

$$Z-score = \frac{n-m}{1.486 \times MAD}$$

or using mean absolute deviation (MeanAD) of the normal samples if the MAD value was equal to zero:

$$Z-score = \frac{n-m}{1.253314 \times MeanAD}$$

The genes RBMY1 and RBMY2 were discarded, as they show almost no variability

in most samples.

Mutation and Copy number variation analysis

The frequency of somatic mutations across all samples with available data was calculated per gene and for each tumor type, and genes were ranked according to this frequency. The top 5% of somatic mutation frequencies were considered significant. For samples with copy number variation (CNV) data available, the overlap of each RBP with the annotated CNVs was calculated, requiring a CNV score > $\log_2(3)$ or < $\log_2(1)$ for gain or loss, respectively. The frequency of CNV gain/loss across all samples with available data was calculated per gene and tumor type, and genes were ranked based on this frequency. The top 5% of CNV gain/loss were considered significant. For each RBP the robust Z-score defined above was used to calculate the association between up regulation (Z-score > 1.96) or down regulation (Z-score < -1.96) with CNV gain or loss, respectively, was calculated using a Jaccard index based on the presence or absence of each of these two features per tumor sample. Mutual exclusion was measured in the following way: given the number of samples having an RBP mutation and no expression change, n_{10}, and the number of those having an expression change but no RBP mutation, n_{01}, a mutual-exclusion score, mx, with values between 0 and 1, was defined as:

$$mx = 2\frac{\min(n_{10}, n_{01})}{N},$$

where N is the total number of samples with both mutation and expression data available.

Alternative splicing events

Alternative splicing events were calculated from the gene annotation using SUPPA [75], producing a total of 30820 events: 16232 exon skipping (SE) events, 4978 alternative 5' splice-site (A5) events, 6336 alternative 3' splice-site (A3) events, 1478 mutually exclusive exon (ME) events, and 1787 retained intron (RI) events. Alternative first and last exons were not included in the analysis. The percent spliced-in (PSI) value for each alternative splicing event and per sample was calculated with SUPPA from the quantification of the transcripts isoforms in transcript per million (TPM) units. Only those events with a total TPM value for the transcripts defining the event higher

than 0.1 were considered. Differentially spliced events were obtained by comparing the PSI value distributions between the normal and tumor samples using a Wilcoxon signed rank test, removing samples with missing PSI values, using at least 10 paired samples, and correcting for multiple testing using the Benjamini-Hochberg method.

Events were considered as differentially spliced if the absolute difference between the tumor and normal median PSI values was > 0.1 and corrected p-value <0.05. Non-regulated events were defined as having an absolute difference of median PSI values (ΔPSI) $\leq$ 0.01 and corrected p-value > 0.05. The same approach was used to calculate differentially spliced events associated with mutations in RBPs. Tumor samples were separated according to whether they have any mutation in a given RBP, and the two sample groups were compared using a Wilcoxon signed rank test and correcting for multiple testing using the Benjamini-Hochberg method. The tests were performed using protein-affecting mutations (Frame_Shift_Del, Frame_Shift_Ins, In_Frame_Del, In_Frame_Ins, Missense_Mutation, Nonsense_Mutation, Nonstop_Mutation). Only RBPs that had mutations in at least 10 samples were tested. A median-based enrichment z-score was calculated per RBP and tumor type by comparing the number of events changing significantly in relation to mutations with the median value obtained using all RBPs tested.

Tumor type specific alternative splicing events were calculated by comparing the PSI values for events between each pair of tumor types. An equal number of samples from each tumor pair were subsampled 100 times and compared using information gain (IG). An average IG per event was then calculated from the 100 iterations, keeping only events differentially spliced in at least one of the tumor types. The events were ranked based on the average IG and the top 1% was chosen as the most discriminating events between the pair of tumor types. All events obtained this way were finally combined into a final non-redundant set. The list of alternative splicing events changing splicing patterns between embryonic stem cells (ESCs) and differentiated cells and tissues, or upon the knockdown of various proteins were obtained from the literature[76 77 37 42]. These events were matched to our events and the correlations of ΔPSI values for each tumor type were calculated.

Gene sets

Annotations for 50 cancer hallmarks were obtained from the Molecular Signatures Database v4.0[28]. A Fisher exact test was performed using genes with annotated events and genes with differentially spliced events in each tumor type. A list of 82 genes whose alternative splicing was linked before to angiogenesis, apoptosis, metastasis, cancer therapy, proliferation and DNA damage was collected from the literature (Supp. Table S8). Additionally, a set of 889 cancer drivers based on mutations and CNVs (34 in common with the previous set) (Supp. Table S9) was obtained from a combination of COSMIC (October 2014)[78], the genes tested in Jelinic et al., 2014[79], 291 high-confidence tumor drivers (HCD)[36] and 260 drivers from the set Cancer5000[80]. These genes were labeled as oncogenes or tumor suppressors based on the annotations from COSMIC, from Vogelstein et al 2013[81], and from the TSGene database[82]. For unlabeled cases in the HCD and Cancer5000 lists, OncodriveROLE[83] was used to assign a classification with cutoffs of 0.3 (loss-of-function class) and 0.7 (activating class). Mutational/CNV drivers that remained without annotation were labeled as unknown. This resulted in a total set of 937 genes that we collectively call drivers.

RNA binding motif enrichment

The RNA binding motif collection from Ray et al., 2013[38] was used for motif analysis. For RBPs with more than one motif, the matrix model with the highest relative entropy was used. When the RNA binding motif was missing for an RBP, the motif model from another member of the same protein family with reported similar binding affinities was used (Supp. Table S2). For a number of RBPs, the motif from a different species was used after confirming that the RNA binding domain is conserved between the human and the other species: RBM47 (chicken), SF1 (Drosophila), SRP54 (Drosophila), TRA2 (Drosphila), and PCBP3 (mouse) (Supp. Table S2). The tool fimo[84] was used to scan the motifs in the event regions using p-value < 0.001 as cut-off. Motif enrichment analysis was performed by comparing the frequency of regions in regulated events with an RNA binding motif with 100 random subsamples of the same size from equivalent regions in non-differentially spliced (DS) events. Motif enrichment was performed separately for the two directions of splicing change (ΔPSI > 0.1 or ΔPSI < -0.1). An enrichment z-score per RNA binding motif, region and direction of regulation was calculated by normalizing the observed frequency in the DS events set with the mean and standard deviation of the 100 random control

sets. Random controls were sampled from non-differentially spliced events for each region and direction of regulation controlling for G+C content. The G+C content distribution for each percentile (0-100) was calculated for DS events, then the G+C percentage for each event in the control set was calculated and a probability assigned based on the distribution of the DS events. Finally, as many control events as DS events were sampled 100 times using the assigned probabilities.

We considered a differentially spliced event to be a potential target of a differentially expressed RNA binding protein gene if the correlation between the event PSI value and the gene expression robust Z-score was $|R| > 0.5$ (Spearman) and the event contained a binding motif. We assessed the significance of the number of events associated to a certain RBP in a tumor type in the following way: The same number of differentially spliced events in a tumor type was randomly selected from the set of non-differentially spliced events 100 times, and events associated to the RBPs calculated each time as described previously. A mean and standard deviation were calculated, from which a z-score was calculated. Only cases with z-score > 1.96 were considered significant.

RBP motif associations and networks

To study the association of RNA binding motifs, a hypergeometric test p-value was calculated to test the co-occurrence of motifs on DS events for a given tumor type. Only motif co-occurrences with adjusted hypergeometric test p-value < 0.05 were kept. To take into account the similarities between motifs, STAMP[85] was used to calculate a motif dissimilarity. The association between motifs was then measured using the geometric mean of the motif dissimilarity and the Jaccard score of the association, multiplied by the $\log_{10}$ of the number of DS events involved, to give more relevance to motifs that occur in many events.

Networks of RBPs and events were built based on the correlations between RBPs through events. A correlation between a pair of RBPs was calculated using the Spearman correlation values with all differentially spliced events in the same tumor type. RBP clusters were built by calculating an inverse covariance matrix of the correlations using the glasso algorithm[86] and then searching for dense, highly connected sub-graphs with a greedy algorithm[87]. Events were associated to a network

if they had |R| > 0.8 (Spearman) or |R| > 0.5 plus motif for any of the RBPs in an RBP cluster.

Cell culture and siRNA transfection

The MCF10A cell line (ATCC, CRL-10317) was sub-cultured in DMEM-F12 (Life Technologies, 31330038) containing 2.5 mM glutamine, 15 mM HEPES, 10% FBS (Life Technologies, 10270), Pen-Strep (Life Technologies, 15070063), human insulin solution at 10 ng/ml concentration (SIGMA, I9278), Hydrocortisone at 0.50 µg/ml (Merck, 386698) and human recombinant EGF from *E. coli* at 25 ng/ml (Merck, 324831). 250,000 MCF10A cells were plated in 6-well plates for cell transfection, and transfected in triplicate using 2 µl Lipofectamine RNAiMax (Life Technologies, 13778150) per 1 ml of total volume of transfection in OPTIMEM (Life Technologies, 13778150). Media was replaced to DMEM-F12 containing 10% FBS and Pen-Strep five hours after treatment. Total RNA was extracted 72 hours after cell transfection, using Maxwell Simply RNA Tissue kit (PROMEGA, AS1280). RNA quality was assessed by Nanodrop spectrophotometer, and in parallel, protein extracts were prepared with RIPA buffer (1mM EDTA, 1.5 mM MgCl2, 20 mM TrisHCl pH7.5, 150 mM NaCl, 1% NP40) with 1x Complete protease inhibitor (ROCHE, 11697498001). MBNL1 siRNA (Life Technologies, s8553 and s8555) and QKI siRNA (Life Technologies, s18084 and s18085) were used at 20, 60, 100 nM concentrations, as well as Silencer® Select Negative Control No. 1 siRNA (Life Technologies, 4390843) at 20 and 100 nM concentrations.

Antisense oligonucleotides treatment

2'-O-Methyl RNA oligos were designed with full phosphorothioate linkage, antisense to the 5' or 3' splice sites of NUMA1 alternative exon 16 (UCSC genome browser hg19 coordinates chr11:71723447-71723488) optimizing GC content to 45-60 %. Custom modified and HPLC purified RNA oligos were ordered in a 0.2 µM scale from SIGMA-ALDRICH.

NUMA1_ex16_5'ss: 5'- ggcauuacCUGCUUAGUUUGC-3'

NUMA1_ex16_3'ss: 5'- CCUCUAGCUGCUCCACcugu-3'

RANDOM 2'-O-Methyl RNA oligo: 5'-GCAAUGGCGUCAAGUGUGUCG-3'

Antisense RNA oligos were transfected in triplicate at 20 nM final concentration using 2 µl Lipofectamine RNAiMax (Life Technologies, 13778150) per 1 ml of total

volume of transfection in OPTIMEM (Life Technologies, 13778150). After five hours of treatment, media was replaced by DMEM-F12 containing 10% FBS and Pen-Strep.

Western Blot analysis

Protein extracts were fractionated by electrophoresis in 10% native acrylamide:bis-acrylamide 30:0.8% gels, and semi-dry transferred to a 0.45 μM nitrocellulose membrane (Protran BA85 10401196, Whatman). MBNL1 monoclonal antibody (M02), clone 3E7 (ABNOVA, H00004154-M02), QKI MaxPab mouse polyclonal antibody (B01) (ABNOVA, H00009444-B01), monoclonal anti-β-tubulin (SIGMA, T4026) and ECL rabbit or mouse IgG, HRP-Linked Whole Ab (GE Healthcare, NA9340 or NA931) were incubated with the membranes and after extensive washes the bound antibodies detected by Western Lightning Plus ECL chemiluminescence reagent (PERKIN-ELMER, NEL105001EA) and exposed to Kodak BioMax MR film (SIGMA, Z353949).

Cell proliferation/viability assay

2500 MCF10A cells/well were seeded the night before treatment in 96-well plates (NUNC, 167008) in 100 μl complete DMEM-F12 medium. Wells with none, half or double amount of cells were also seeded for fluorescence calibration. Cells were transfected with siRNA or AON oligos as described. Resazurin (SIGMA, R7017) treatment was performed 72, 96 and 120 hours after transfection, in 7 replicates and incubated for 4 hours in a 37ºC incubator. Fluorescence was measured after 4 hours of incubation, using a TECAN infinite m200 device with 530 nm excitation wavelength, 590 nm emission wavelength, 30 nm emission bandwidth, and set to optimal gain. The medium was replaced by complete DMEM-F12 after measurements.

Semi quantitative RT-PCR

500 ng of total RNA was reverse-transcribed with Superscript III (Life Technologies, 18080085) with a mix of random primers and oligo-dT (18-mer), and 1 μl of cDNA was analyzed by PCR, using specific primers complementary to the constitutive exons flanking the alternative exon and GoTaq flexi DNA polymerase (Promega, M7806). PCR products were analyzed by 6% native acrylamide gel electrophoresis in 1x TBE and Sybr safe staining (Life Technologies, S33102). The ratio between exon inclusion and skipping isoforms was quantified from biological triplicates using ImageJ 1.47v

(NIH, USA). The list of primers used for the semi-quantitative RT-PCR can be found in Supp. Table S22.

Centrosome count and aneuploidy signature

The number of centrosomes was determined by immunofluorescence assays using an anti-γ-tubulin (TUBG1) antibody (clone GTU-88, Sigma-Aldrich; dilution 1:1,000). The expected immunostaining pattern of this centrosomal marker in normal cells is one or two foci proximal to the nucleus. The cells were fixed in methanol cold for 10 minutes and washed in phosphate-buffered saline. The secondary antibody was Alexa Fluor 488 (Molecular Probes, Life Thecnologies) and the cells were mounted using VECTASHIELD® with DAPI. The results correspond to at least five independent fields and > 200 cells analyzed. The significance of the results was assessed using the one-sided Mann-Whitney test (Supp. Table S20). The chromosome instability signature (CIN25) from [46] was used by calculating the mean value of the normalized expression robust Z-score values for the 25 genes from the signature in each sample.

Supplementary Data

Supp. File 1: Information about the differential expression, mutations and copy number variations of the all the RBP genes analyzed.

Supp. File 2: Differentially spliced events in the comparison of tumor vs. normal.

Supp. File 3: Differentially spliced events in the comparison of samples with and without mutations in RNA binding proteins.

Supp. File 4: Candidate target differentially spliced events for each tumor and for each differentially expressed RBP.

Supp. File 5: Information per tumor type for each gene: whether it is differentially expressed, has alternative splicing events, has differentially spliced events in tumor vs normal, or is a cancer driver.

Acknowledgements

We are thankful to P. Papasaikas, B. Blencowe, M. Irimia and Q. Morris for comments and discussions. ES, BS, AP and EE were supported by grants BIO2011-

23920 and Consolider RNAREG (CSD2009-00080) from the MINECO (Spanish Government), by AGAUR and by the Sandra Ibarra Foundation for Cancer (FSI2013). JV and BM were supported by Fundación Botín, by Banco de Santander through its Santander Universities Global Division and by Consolider RNAREG (CSD2009-00080), MINECO and AGAUR.

Tables

Protein affecting mutations

Tumor	RBP	logFC	Pval	Frequency	Assoc. Freq.	Jac | Mex
COAD	*HNRNPL*	0.4393	2.97E-10	7.14	5.71	0.06 | 0.03
COAD	*RALY*	0.3369	0.00624	9.52	5.24	0.09 | 0.09
COAD	*SRRM2*	0.5345	0.00531	8.09	0	0 | 0.05
KICH	*PABPC1*	-0.0069	1	11.29	3.23	0.10 | 0.16
KICH	*PABPC3*	1.2009	3.42E-7	14.52	8.06	0.14 | 0.13
KICH	*RBMXL1*	-0.6954	1.62E-5	14.52	6.45	0.13 | 0.16
LIHC	*SRRM2*	0.1716	0.263	8.42	4.21	0.08 | 0.08
LIHC	*ZNF638*	-0.0726	0.438	8.95	2.11	0.07 | 0.14
LUAD	*RBM10*	0.1471	0.0573	7.64	5.90	0.14 | 0.03
LUSC	*SRRM2*	-0.2196	0.0836	7.87	3.37	0.06 | 0.09

CNV Gains

Tumor	RBP	logFC	Pval	Frequency	Assoc. Freq.	Jaccard
BRCA	*CELF3*	2.11	8.82E-14	20.21	15.29	0.22
BRCA	*ESRP1*	2.07	1.03E-13	23.66	20.21	0.50
BRCA	*HNRNPU*	0.56	5.50E-29	20.52	14.14	0.32
COAD	*HNRNPA1L2*	0.71	2.24E-10	16.67	14.29	0.17
COAD	*PABPC1*	1.21	2.68E-12	10.00	10.00	0.13
COAD	*PABPC3*	1.05	4.21E-10	18.57	15.24	0.22
COAD	*RBM39*	0.81	3.31E-08	36.67	35.71	0.54
COAD	*SRSF6*	0.52	3.05E-05	38.10	31.43	0.52
KIRP	*IGF2BP3*	2.86	1.98E-05	13.94	9.70	0.12
KIRP	*RBM28*	0.69	6.84E-10	16.36	15.15	0.29
KIRP	*SRRM3*	2.35	0.000300325	13.94	12.73	0.17
LIHC	*HNRNPU*	0.57	1.23E-08	20.00	12.11	0.22
LIHC	*PABPC1*	1.05	3.26E-09	28.42	28.42	0.37
LUAD	*ESRP1*	0.96	4.34E-18	8.08	7.86	0.10
LUAD	*PABPC1*	0.94	3.68E-15	9.39	9.17	0.13
LUAD	*SRP54*	0.63	4.42E-15	11.35	11.14	0.16
LUSC	*FXR1*	1.44	6.35E-25	57.87	57.87	0.60
LUSC	*HNRNPL*	0.71	1.97E-20	7.87	7.87	0.08
LUSC	*IGF2BP2*	2.29	1.37E-12	55.06	49.44	0.57
LUSC	*TRA2B*	0.64	6.57E-12	53.93	53.37	0.66

CNV Losses

Tumor	RBP	logFC	Pval	Frequency	Assoc. Freq.	Jaccard
COAD	*RBFOX1*	-2.19	0.000197793	11.90	4.76	0.07
KIRC	*RBFOX1*	-4.20	1.45E-30	1.71	1.71	0.02
LIHC	*RBMS3*	-1.51	6.75E-09	1.05	1.05	0.02
LUSC	*CELF2*	-2.91	4.41E-29	1.12	1.12	0.01
LUSC	*CPEB2*	-0.87	1.98E-07	1.12	1.12	0.02
LUSC	*KHDRBS2*	-5.78	4.49E-32	1.12	1.12	0.01
LUSC	*RBM47*	-0.97	9.45E-14	2.25	1.69	0.02
LUSC	*RBMS3*	-2.09	2.18E-16	2.25	2.25	0.03

Table 1. Association of mutations and CNVs with expression changes. For each RBP gene and each tumor type we give the log-fold change (logFC) and adjusted p-value (Pval) of the differential expression analysis between tumor and normal samples, the frequency of the alteration (Frequency), the association of the alteration with expression Z-score (Assoc. Freq.) and the Jaccard score of the association. For mutations we show those cases with mutation frequency >7% and report the Jaccard (Jac) and mutual exclusion (Mex) score. For CNV gains we show those cases that show a significant upregulation and association frequency >7%. For CNV losses, we show those cases with significant downregulation and association frequency > 1%. We only show known or putative splicing factors. Other RBPs are described in the Supplementary Material.

Tumor type	Gene	Cancer driver	event type	event	ΔPSI	Pval
BRCA	*SF3B1*	*BCL2L1*	A3	chr20:30310151-30310421:30310133-30310421:-	0.23	4.14E+07
BRCA	*SF3B1*	*MEF2A*	SE	chr15:100106309-100138635:100138716-100173183:+	-0.21	2.59E-03
LUAD	*RBM10*	*BLM*	SE	chr15:91260671-91267265:91267374-91290619:+	0.13	3.40E-02
LUAD	*RBM10*	*CTNND1*	A3	chr11:57529518-57558857:57529518-57558966:+	-0.34	6.43E-03
LUAD	*RBM10*	*MUC1*	SE	chr1:155160052-155160198:155160334-155160639:-	0.13	3.33E-02
LUAD	*RBM10*	*WNK1*	A3	chr12:980514-987378:980514-987381:+	0.22	3.28E-03
LUAD	*U2AF1*	*BCOR*	SE	chrX:39930412-39930890:39930943-39931602:-	-0.12	3.51E-02
LUAD	*U2AF1*	*CHCHD7*	A3	chr8:57127226-57128948:57127226-57128992:+	-0.11	5.45E-04
LUAD	*U2AF1*	*CTNNB1*	A3	chr3:41280845-41281151:41280845-41281310:+	0.26	1.51E-04
LUAD	*U2AF1*	*CTNNB1*	RI	chr3:41280625:41280845-41281310:41281939:+	0.19	4.05E-04
LUAD	*U2AF1*	*MUC1*	MX	chr1:155159850-155159931:155160052-155160484:155159850-155160198:155160334-155160484:-	-0.26	4.72E-02
LUAD	*U2AF1*	*PATZ1*	A3	chr22:31724910-31731678:31724845-31731678:-	-0.15	3.47E-02
LUAD	*U2AF1*	*PCM1*	SE	chr8:17838264-17840742:17840798-17842956:+	0.12	1.60E-02
LUAD	*U2AF1*	*RIPK2*	SE	chr8:90770461-90775057:90775210-90777569:+	-0.26	7.98E-04
LUAD	*U2AF1*	*RIT1*	SE	chr1:155880297-155880447:155880595-155881034:-	-0.13	1.82E-03
COAD	*HNRNPL*	*CASP8*	A5	chr2:202098345-202098739:202098277-202098739:+	0.17	2.71E-02
COAD	*MACF1*	*CCNB1IP1*	MX	chr14:20784719-20785954:20786133-20793698:20784719-20786415:20786629-20793698:-	0.23	3.95E-02
COAD	*MACF1*	*MDM4*	A3	chr1:204501374-204506558:204501374-204506587:+	0.33	4.00E-02
COAD	*MACF1*	*RAC1*	SE	chr7:6431672-6438293:6438349-6439757:+	-0.15	2.45E-03
COAD	*MACF1*	*TJP2*	SE	chr9:71863140-71865951:71866280-71867731:+	0.15	4.67E-02
COAD	*NBPF10*	*KDM6A*	SE	chrX:44919401-44919854:44920009-44920569:+	-0.11	3.16E-02
COAD	*NBPF10*	*RAC1*	SE	chr7:6431672-6438293:6438349-6439757:+	-0.10	1.56E-02
COAD	*YLPM1*	*CD44*	SE	chr11:35211612-35218293:35218421-35219668:+	0.12	2.68E-02
COAD	*ZC3H18*	*ACSL6*	MX	chr5:131309093-131310451:131310528-131312341:131309093-131310586:131310642-131312341:-	-0.84	1.14E-02
COAD	*ZC3H18*	*MBD1*	SE	chr18:47796188-47797839:47797910-47799047:-	-0.11	1.73E-02

Table 2. Event in cancer drivers associated to protein-affecting mutations in RBPs. For each tumor and for each RBP (Gene**)** we indicate the cancer driver gene and event predicted to have a significant splicing change in association to protein-affecting mutations. We also provide the PSI change between mutated and non-

mutated samples (ΔPSI) and the p-value of the comparison after correcting for multiple testing (Pval). SF3B1 is included for comparison. Events are defined according to the format used by SUPPA[75].

Figure captions

Figure 1. Cancer alterations in splicing factors. (a) Up- (red) and downregulation (blue) patterns of 162 splicing factors (x-axis) in the different tumor types (y-axis) compared to normal samples (Methods). The color intensity indicates the $\log_2$-fold change (log2 FC). The bar plot above indicates the frequency of tumor types with up- (red) or down- (blue) regulation for each factor. Splicing factors previously described to have oncogenic or tumor-suppressing activities (Supp. Table S3) are indicated in red on the x-axis. Dendrogram was built with Ward clustering after transforming to +1 or -1 significant cases for log2 FC>0.5 or <0.5, respectively, and using Gower distance. **(b)** Dendrogram for the unsupervised hierarchical clustering of the 4442 tumor samples across the 11 tumor types using the expression robust Z-score expression relative to normal samples. **(c)** Percentage of samples (y axis) in which RBPs (upper panel), driver genes (middle panel) and the rest of genes (lower panel) show mutations in each tumor type (x axis). Distributions are represented as violin plots, where the width indicates the density at a given y-axis value. Drivers were extracted from the literature (Methods). **(d)** Copy number gains (left panel) and losses (right panel) of the tested splicing factors. Only those with a frequency of amplification or deletion in the top 5% of all genes using all tumor samples are shown. SRPK1 and SRPK2 are included for comparison.

Figure 2. Differentially spliced events in tumors. **(a)** Upper panel: number of paired-samples used per tumor type. Lower panel: number of differentially spliced events per tumor type compared to normal samples, split according to the number of each type of event: alternative 3' splice-site (A3), alternative 5' splice-site (A5), mutually exclusive exon (MX), retained intron (RI) and skipping exon (SE) (Supp. Table S4). **(b)** Proportion of driver and non-driver genes with differentially spliced

events. We indicated in red those tumors for which the enrichment is significant. **(c)** Cancer hallmarks (x-axis) that are enriched (Fisher test p-value < 0.05) in differentially spliced events in each tumor type (y-axis). The color indicates the odds ratio of the enrichment. Hallmarks that are also enriched according to gene expression are indicated with a black dot. **(d)** Proportion of samples with mutations in each tumor type for RBP genes with at least 10 associated differentially spliced events. *SF3B1* is included for comparison. **(e)** Number of differentially spliced events related to the mutations in (d) color-labeled by tumor type. Only cases with at least 10 associated differentially spliced events are shown. **(f)** Number of protein-affecting mutations (y-axis) along the HNRNPL protein (x-axis), color-labeled according to whether they are substitutions, insertions or deletions. Protein domains are indicated in light red. **(g)** Enrichment or depletion of specific event types in association to mutations in HNRNPL (red bars) compared to the overall proportions of events (black bars). Significant differences ($p < 0.05$, Fisher test) are labeled in red. Contingency tables are provided as Supplementary Tables. **(h)** Distribution of PSI values for the A5 event in *CASP8* associated to the mutations of *HNRNPL* in COAD, separated into normal samples, tumor samples without protein-affecting mutations (Tumor – NM), and tumor samples with protein-affecting mutations (Tumor – M).

Figure 3. Common and specific events in tumors. (a) Common events and splicing factors between pairs of tumor types. For each pair of tumor types and for each splicing factor differentially expressed in both tumor types, we plot the correlation of ΔPSI values for events that have a correlation of $|R| > 0.5$ (Spearman) with these splicing factors in both tumor types. Only factors with more than 50 associated events in both tumor types are shown. Each event is only plotted once and the color of the plot corresponds to the most common correlating splicing factor. Correlations between ΔPSI values are indicated. In red or green, we highlight those higher than 0.8 or lower than -0.8, respectively. **(b)** ΔPSI correlations for the pairs LUAD - BRCA, PRAD - KIRC, KIRC – HNSC, and LUSC – HNSC, for the common events separated according to their potential splicing factor regulators. Events associated to more than one factor are represented with jitter. **(c)** Principal Component Analysis (PCA) plot of 380 tumor specific alternative splicing events colored by tumor type.

Figure 4. **Enriched RNA binding motifs in differentially spliced events. (a)** Enriched RNA binding motifs in differentially spliced skipping exon events in each tumor type, separated by inclusion (upper panels) or skipping (lower panels) events, and by upstream (left), exonic (middle) or downstream (right) regions. Only enriched motifs for splicing factors that are differentially expressed in each tumor type are indicated. RBP gene up- and downregulation is indicated in red and blue, respectively. The color intensity indicates the Z-score of the motif enrichment. Similar plots for the other event types are given in the Supp. Material. **(b)** Proportion of enriched motifs in inclusion (ΔPSI > 0.1) (left panel) and skipping (ΔPSI < 0.1) (right panel) events, in each of the event regions (x-axis): upstream (Upstr.), exon and downstream (Downstr.). Proportions are separated according to whether the RBP gene is up- (red) or down- (blue) regulated. **(c)** Upper panel: Total proportion (y-axis) of differentially spliced events in each tumor type (x-axis) that are assigned as potential targets of one or more differentially expressed RBPs with significance z-score > 1.96. Lower panel: Proportion of differentially spliced events (marked in green) that are assigned as potential targets of each RBP (y-axis) in each tumor type (x-axis), with significance z-score > 1.96. **(d)** Top panels: Correlation (Spearman R) of ΔPSI values in breast tumors (BRCA) and prostate tumors (PRAD) with the ΔPSI obtained from the comparison of stem cells (ESCs) with differentiated cells (CL). Events with a predicted MBNL binding motif are indicated in blue. Lower panels: for each RBP (y axis), it shows proportion of cancer drivers (x-axis) with differentially spliced events, whose PSI correlates ($|R| > 0.5$ Spearman) with the RBP gene expression and contains the corresponding RNA binding motif. Only the top 10 RBPs are shown. **(e)** Circos plot for the association of RNA binding motifs on differentially spliced events (Methods). Only significant links are shown (association score > 1.5 and hypergeometric p-value < 0.05). The thickness of the links indicates the number of tumor types for which a significant association has been found.

Figure 5. Networks of splicing regulation. Modules of alternative splicing regulation according to cancer hallmarks in breast **(a)** and colon **(b)** tumors. For each cluster of splicing factors (x-axis) we indicate in gray the total number of genes

associated to these factors in each hallmark (y-axis). Only enriched hallmarks are shown. We indicate in red the number of cancer drivers associated to each factor. Splicing factors in each cluster are ordered according to the total number of genes they are associated to. Representation of the regulatory modules for G2M checkpoint **(c)** and WNT/Beta-catenin **(d)** hallmarks in breast tumors, and for the angiogenesis hallmark **(e)** in colon tumors. Splicing factors are indicated as square boxes in red or blue depending of whether they are up- or downregulated. Target genes are presented as white diamonds for cancer drivers and white boxes for the rest. Connections indicate predicted splicing regulation by a splicing factor.

Figure 6. Regulation of NUMA1 alternative splicing by MBNL1 in breast luminal tumors. (a) PSI value distributions in tumor and paired normal sample for luminal A (LA) and luminal B (LB) breast tumors for the events in NUMA1 (LA: ΔPSI = -0.22, p-value = 7.81e-07, LB: ΔPSI = -0.23, p-value = 0.037) and NUMB (LA: ΔPSI = 0.28 p-value = 0.0001, LB: ΔPSI = 0.28, p-value = 0.016). All p-values given are corrected for multiple testing. (b) RT-PCR isoform analysis upon knockdowns of *MBNL1* (lanes 2-10) and *QKI* (lanes 12-14) and their respective controls with scrambled siRNAs (lanes 1 and 11). The diagrams to the right indicate the position of the alternatively spliced exons. **(c)** Resazurin-based assays of cell viability/proliferation. Measurements were performed in triplicate at 72, 96 and 120 hours. The plot shows measurements upon knockdowns of *MBNL1* (siMBNL1) and *QKI* (siQKI), upon transfection of AONs targeting the 3' and 5' splice-sites independently and both together, and the corresponding controls (scrambled siRNA and random AON). **(d)** Left panel: graph showing the results of the evaluation of centrosome amplification upon knockdown of *MBNL1* (siMBNL1) or upon transfection of AONs targeting 5' splice-sites (5'ss AON), compared to the corresponding controls siScrambled (p=0,4271) and random AON (p=0,04356), respectively (one-sided Mann-Whitney test). Right panels: representative merged (TUBG1 and DAPI) images of immunofluorescence assays. **(e)** Correlation of *NUMA1* PSI (x-axis) with the CIN25 signature of aneuploidy (y-axis) across the tumor (red) and normal (blue) samples for luminal A (upper panel) (R=-0.4 Spearman) and B (lower panel) (R=-0.33 Spearman).

References

1. Grosso, A. R. & Carmo-Fonseca, M. in *Nuclear Signaling Pathways and Targeting Transcription in Cancer* (ed. Kumar, R.) 313–336 (2014). doi:10.1007/978-1-4614-8039-6

2. Bechara, E. G., Sebestyén, E., Bernardis, I., Eyras, E. & Valcárcel, J. RBM5, 6, and 10 Differentially Regulate NUMB Alternative Splicing to Control Cancer Cell Proliferation. *Mol. Cell* **52,** 720–733 (2013).

3. Dorman, S. N., Viner, C. & Rogan, P. K. Splicing mutation analysis reveals previously unrecognized pathways in lymph node-invasive breast cancer. *Sci. Rep.* **4,** 7063 (2014).

4. Brooks, A. N. *et al.* A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events. *PLoS One* **9,** e87361 (2014).

5. Yoshida, K. *et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478,** 64–69 (2011).

6. Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150,** 1107–1120 (2012).

7. Maguire, S. L. *et al.* SF3B1 mutations constitute a novel therapeutic target in breast cancer. *J. Pathol.* **235,** 571–580 (2015).

8. Kim, E. *et al.* SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. *Cancer Cell* **27,** 617–630 (2015).

9. Karni, R. *et al.* The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nat. Struct. Mol. Biol.* **14,** 185–93 (2007).

10. Xiao, R. *et al.* Splicing regulator SC35 is essential for genomic stability and cell proliferation during mammalian organogenesis. *Mol. Cell. Biol.* **27,** 5393–5402 (2007).

11. Roychoudhury, P. & Chaudhuri, K. Evidence for heterogeneous nuclear ribonucleoprotein K overexpression in oral squamous cell carcinoma. *Br. J. Cancer* **97,** 574–575 (2007).

12. Golan-Gerstl, R. *et al.* Splicing factor hnRNP A2/B1 regulates tumor suppressor gene splicing and is an oncogenic driver in glioblastoma. *Cancer Res.* **71,** 4464–4472 (2011).

13. Wang, Y. *et al.* The Splicing Factor RBM4 Controls Apoptosis, Proliferation, and Migration to Suppress Tumor Progression. *Cancer Cell* **26,** 374–389 (2014).

14. Zong, F.-Y. *et al.* The RNA-Binding Protein QKI Suppresses Cancer-Associated Aberrant Splicing. *PLoS Genet.* **10,** e1004289 (2014).

15. Ghigna, C. *et al.* Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. *Mol. Cell* **20,** 881–890 (2005).

16. Moon, H. *et al.* SRSF2 promotes splicing and transcription of exon 11 included isoform in Ron proto-oncogene. *Biochim. Biophys. Acta* **1839,** 1132–1140 (2014).

17. Huang, C.-S., Shen, C.-Y., Wang, H.-W., Wu, P.-E. & Cheng, C.-W. Increased expression of SRp40 affecting CD44 splicing is associated with the clinical outcome of lymph node metastasis in human breast cancer. *Clin. Chim. Acta* **384,** 69–74 (2007).

18. Watermann, D. O. *et al.* Splicing factor Tra2-β1 is specifically induced in breast cancer and regulates alternative splicing of the CD44 gene. *Cancer Res.* **66,** 4774–4780 (2006).

19. Jia, R., Li, C., McCoy, J. P., Deng, C. X. & Zheng, Z. M. SRp20 is a proto-oncogene critical for cell proliferation and tumor induction and maintenance. *Int. J. Biol. Sci.* **6,** 806–826 (2010).

20. Jensen, M. A., Wilkinson, J. E. & Krainer, A. R. Splicing factor SRSF6 promotes hyperplasia of sensitized skin. *Nat. Struct. Mol. Biol.* **21,** 189–197 (2014).

21. Sampath, J. *et al.* Human SPF45, a splicing factor, has limited expression in normal tissues, is overexpressed in many tumors, and can confer a multidrug-resistant phenotype to cells. *Am. J. Pathol.* **163,** 1781–1790 (2003).

22. Shitashige, M. *et al.* Increased susceptibility of Sf1+/- mice to azoxymethane-induced colon tumorigenesis. *Cancer Sci.* **98,** 1862–1867 (2007).

23. Katz, Y. *et al.* Musashi proteins are post-transcriptional regulators of the epithelial-luminal cell state. *Elife* **3,** e03915 (2014).

24. Yae, T. *et al.* Alternative splicing of CD44 mRNA by ESRP1 enhances lung colonization of metastatic cancer cell. *Nat. Commun.* **3,** 883 (2012).

25. Cohen-Eliav, M. *et al.* The splicing factor SRSF6 is amplified and is an oncoprotein in lung and colon cancers. *J. Pathol.* **229,** 630–9 (2013).

26. Mercier, I. *et al.* CAPER, a novel regulator of human breast cancer progression. *Cell Cycle* **13,** 1256–1264 (2014).

27. Ghigna, C., Riva, S. & Biamonti, G. in *RNA and Cancer* (ed. Wu, J. Y.) **158,** 95–117 (Springer Berlin Heidelberg, 2013).

28. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).

29. Castello, A. *et al.* Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* **149,** 1393–1406 (2012).

30. Baltz, A. G. *et al.* The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Mol. Cell* **46,** 674–90 (2012).

31. Yang, M. *et al.* Functional variants in cell death pathway genes and risk of pancreatic cancer. *Clin. Cancer Res.* **14,** 3230–3236 (2008).

32. Goehe, R. W. *et al.* hnRNP L regulates the tumorigenic capacity of lung cancer xenografts in mice via caspase-9 pre-mRNA processing. *J. Clin. Invest.* **120,** 3923–3939 (2010).

33. Gonçalves, V., Matos, P. & Jordan, P. Antagonistic SR proteins regulate alternative splicing of tumor-related Rac1b downstream of the PI3-kinase and Wnt pathways. *Hum. Mol. Genet.* **18,** 3696–707 (2009).

34. Misquitta-Ali, C. M. *et al.* Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer. *Mol. Cell. Biol.* **31,** 138–150 (2011).

35. Bordonaro, M. Crosstalk between Wnt signaling and RNA processing in colorectal cancer. *J. Cancer* **4,** 96–103 (2013).

36. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3,** 2650 (2013).

37. Vanharanta, S. *et al.* Loss of the multifunctional RNA-binding protein RBM47 as a source of selectable metastatic traits in breast cancer. *Elife* (2014). doi:10.7554/eLife.02734

38. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499,** 172–177 (2013).

39. Witten, J. T. & Ule, J. Understanding splicing regulation through RNA splicing maps. *Trends Genet.* **27,** 89–97 (2011).

40. Best, A. *et al.* Human Tra2 proteins jointly control a CHEK1 splicing switch among alternative and constitutive target exons. *Nat. Commun.* **5,** 4760 (2014).

41. Venables, J. P. *et al.* MBNL1 and RBFOX2 cooperate to establish a splicing programme involved in pluripotent stem cell differentiation. *Nat. Commun.* **4,** 2480 (2013).

42. Han, H. *et al.* MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498,** 241–245 (2013).

43. Zheng, Z. *et al.* LGN regulates mitotic spindle orientation during epithelial morphogenesis. *J. Cell Biol.* **189,** 275–288 (2010).

44. Suzuki, K. The multi-functional serpin, protein C inhibitor: Beyond thrombosis and hemostasis. *J. Thromb. Haemost.* **6,** 2017–2026 (2008).

45. Anoopkumar-Dukie, S. *et al.* Resazurin assay of radiation response in cultured cells. *Br. J. Radiol.* **78,** 945–947 (2005).

46. Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.* **38,** 1043–1048 (2006).

47. Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21,** 3433–3434 (2005).

48. Xue, Y. *et al.* GPS 2.1: Enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. *Protein Eng. Des. Sel.* **24,** 255–260 (2011).

49. Hoadley, K. A. *et al.* Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell* **158,** 929–944 (2014).

50. Bell, J. L. *et al.* Insulin-like growth factor 2 mRNA-binding proteins (IGF2BPs): Post-transcriptional drivers of cancer progression? *Cell. Mol. Life Sci.* **70,** 2657–2675 (2013).

51. Sengupta, N. *et al.* Analysis of colorectal cancers in British Bangladeshi identifies early onset, frequent mucinous histotype and a high prevalence of RBFOX1 deletion. *Mol. Cancer* **12,** 1 (2013).

52. Jayasena, C. S. & Bronner, M. E. Rbms3 functions in craniofacial development by posttranscriptionally modulating TGF-beta signaling. *J. Cell Biol.* **199,** 453–466 (2012).

53. Li, Y. *et al.* Downregulation of RBMS3 Is Associated with Poor Prognosis in Esophageal Squamous Cell Carcinoma. *Cancer Res.* **71,** 6106–6115 (2011).

54. Liang, Y.-N. *et al.* RBMS3 is a tumor suppressor gene that acts as a favorable prognostic marker in lung squamous cell carcinoma. *Med. Oncol.* **32,** (2015).

55. Koh, C. M. *et al.* MYC regulates the core pre-mRNA splicing machinery as an essential step in lymphomagenesis. *Nature* **523,** 96–100 (2015).

56. Oltean, S. & Bates, D. O. Hallmarks of alternative splicing in cancer. *Oncogene* 1–8 (2013). doi:10.1038/onc.2013.533

57. Jangi, M. & Sharp, P. A. Building Robust Transcriptomes with Master Splicing Factors. *Cell* **159,** 487–498 (2014).

58. Kammerer, S. *et al.* Association of the NuMA region on chromosome 11q13 with breast cancer susceptibility. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 2004–2009 (2005).

59. Kotak, S., Busso, C. & Gönczy, P. NuMA phosphorylation by CDK1 couples mitotic progression with cortical dynein function. *EMBO J.* **32,** 2517–2529 (2013).

60. Bergstralh, D. T. & St Johnston, D. Spindle orientation: What if it goes wrong? *Semin. Cell Dev. Biol.* **34,** 140–145 (2014).

61. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

62. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).

63. The Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517,** 576–582 (2015).

64. The Cancer Genome Atlas Network. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26,** 319–330 (2014).

65. The Cancer Genome Atlas Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499,** 43–49 (2013).

66. The Cancer Genome Atlas Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511,** 543–550 (2014).

67. The Cancer Genome Atlas Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).

68. The Cancer Genome Atlas Network. Integrated Genomic Characterization of Papillary Thyroid Carcinoma. *Cell* **159,** 676–690 (2014).

69. Lee, Y., Krishnan, A., Zhu, Q. & Troyanskaya, O. G. Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. *Bioinformatics* **29,** 3036–3044 (2013).

70. Kwon, S. C. *et al.* The RNA-binding protein repertoire of embryonic stem cells. *Nat. Struct. Mol. Biol.* **20,** 1122–30 (2013).

71. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** 662–669 (2014).

72. Zhang, H. M. *et al.* AnimalTFDB: A comprehensive animal transcription factor database. *Nucleic Acids Res.* **40,** 144–149 (2012).

73. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15,** R29 (2014).

74. Smyth, G. K. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* 397–420 (Springer New York, 2005). doi:10.1007/0-387-29362-0_23

75. Alamancos, G. P., Pagés, A., Trincado, J. L. & Eyras, E. Leveraging transcript quantification for fast computation of alternative splicing profile. *RNA J.* **51,** 769–784 (2015).

76. Brosseau, J.-P. *et al.* Tumor microenvironment-associated modifications of alternative splicing. *RNA* **20,** 189–201 (2014).

77. Shen, S. *et al.* MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res.* **40,** e61 (2012).

78. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43,** D805–D811 (2015).

79. Jelinic, P. *et al.* Recurrent SMARCA4 mutations in small cell carcinoma of the ovary. *Nat. Genet.* **46,** 424–6 (2014).

80. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).

81. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).

82. Zhao, M., Sun, J. & Zhao, Z. TSGene: A web resource for tumor suppressor genes. *Nucleic Acids Res.* **41,** 970–976 (2013).

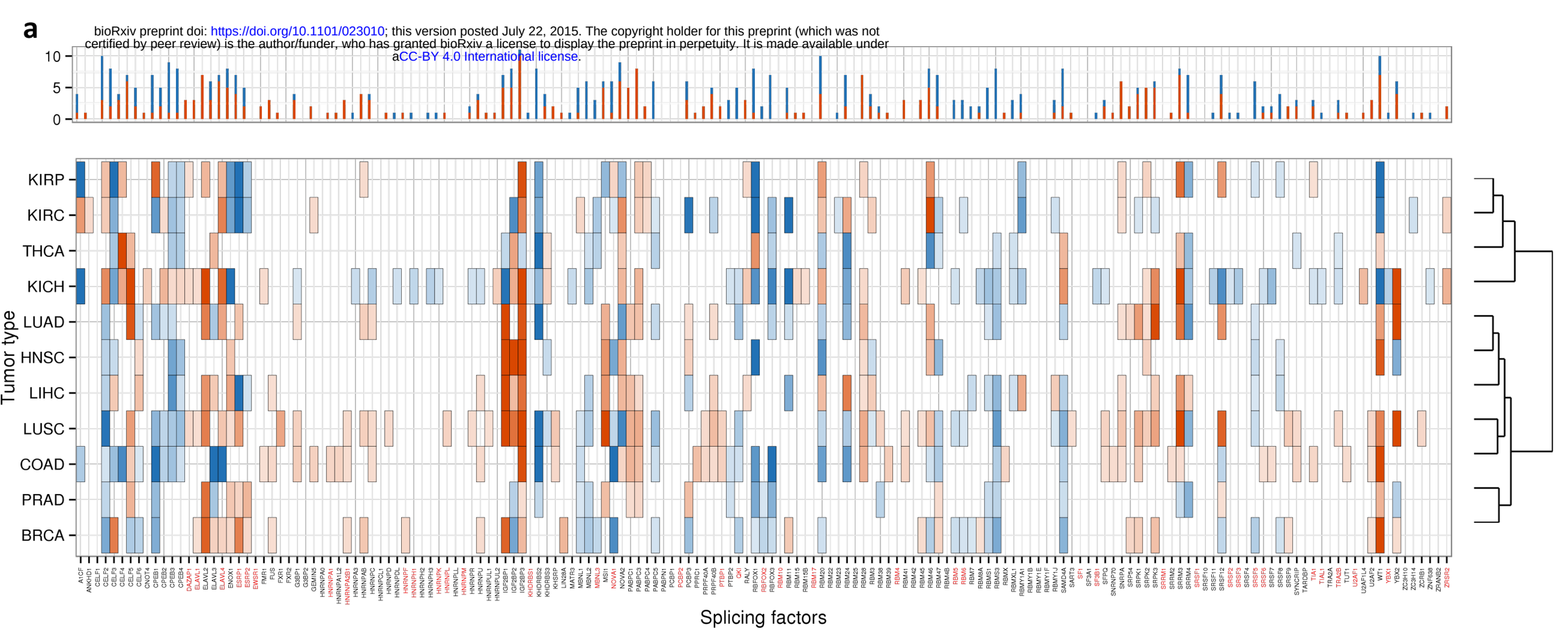
83. Schroeder, M. P., Rubio-Perez, C., Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveROLE classifies cancer driver genes in loss of function and activating mode of action. *Bioinformatics* **30,** i549–i555 (2014).

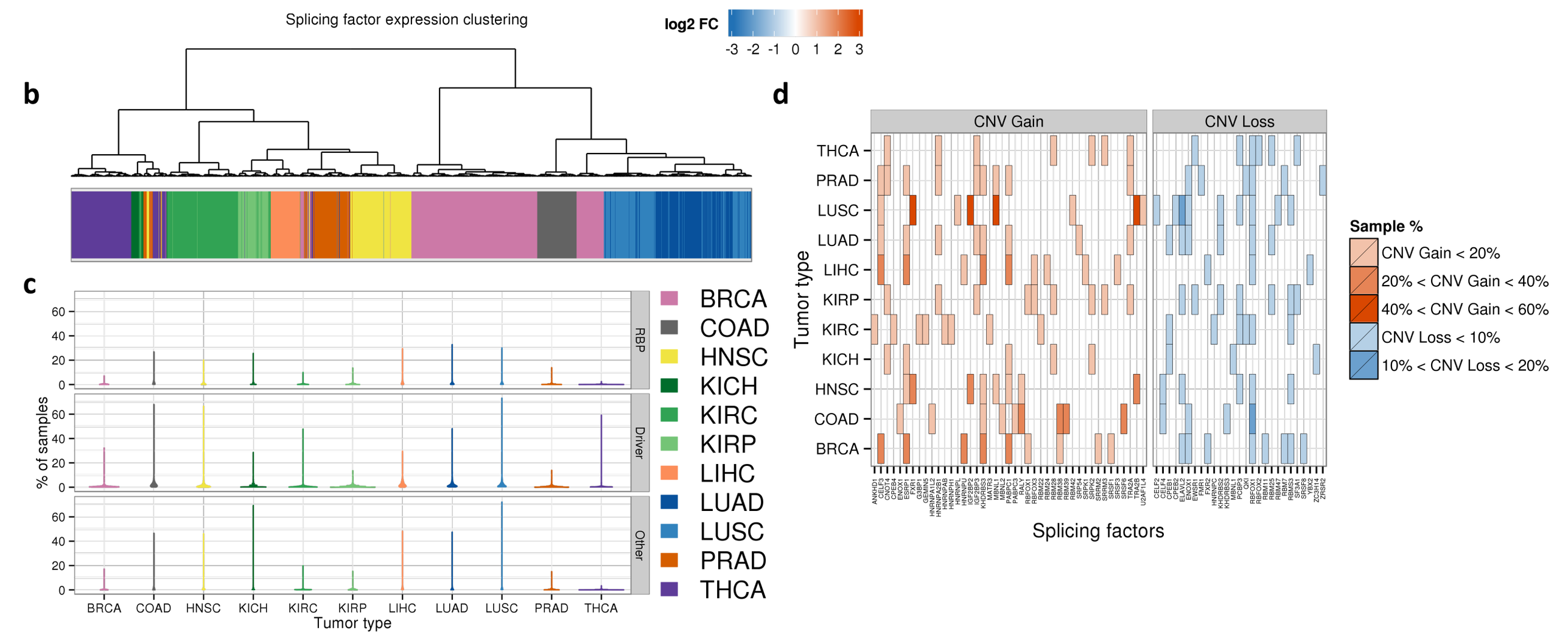
84. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Second Int. Conf. Intell. Syst. Mol. Biol.* **2,** 28–36 (1994).

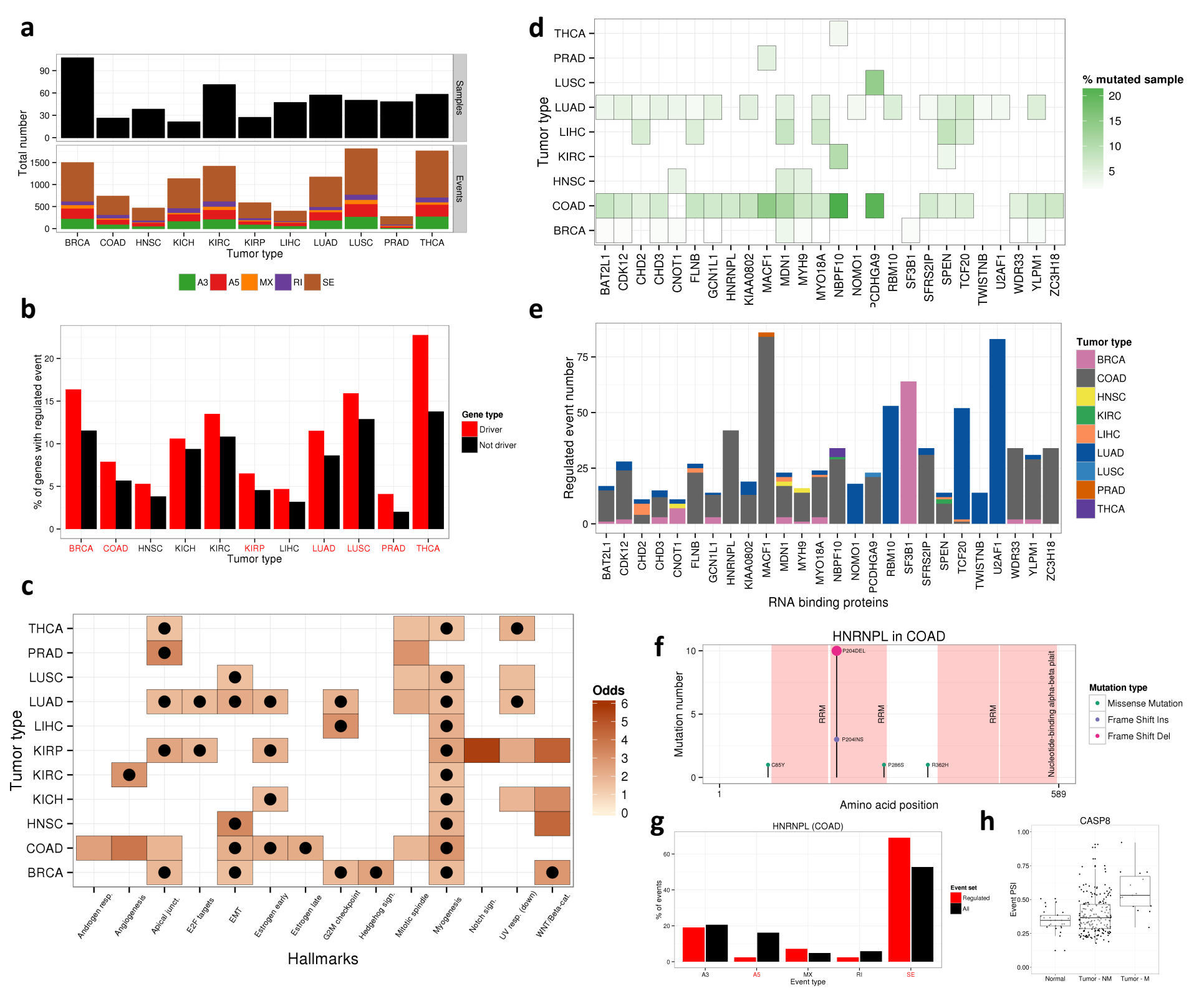
85. Mahony, S. & Benos, P. V. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.* **35,** W253–W258 (2007).

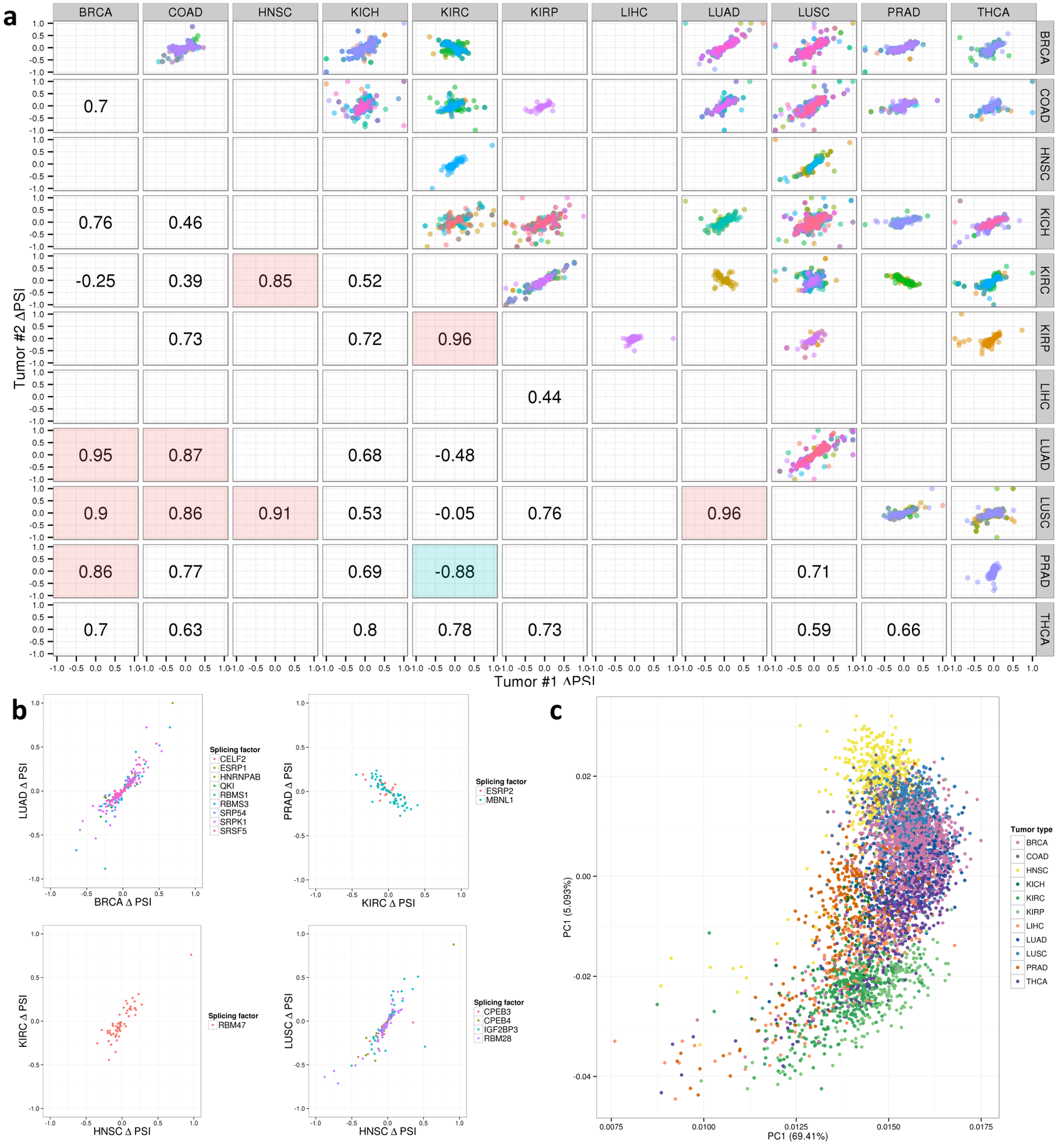
86. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9,** 432–41 (2008).

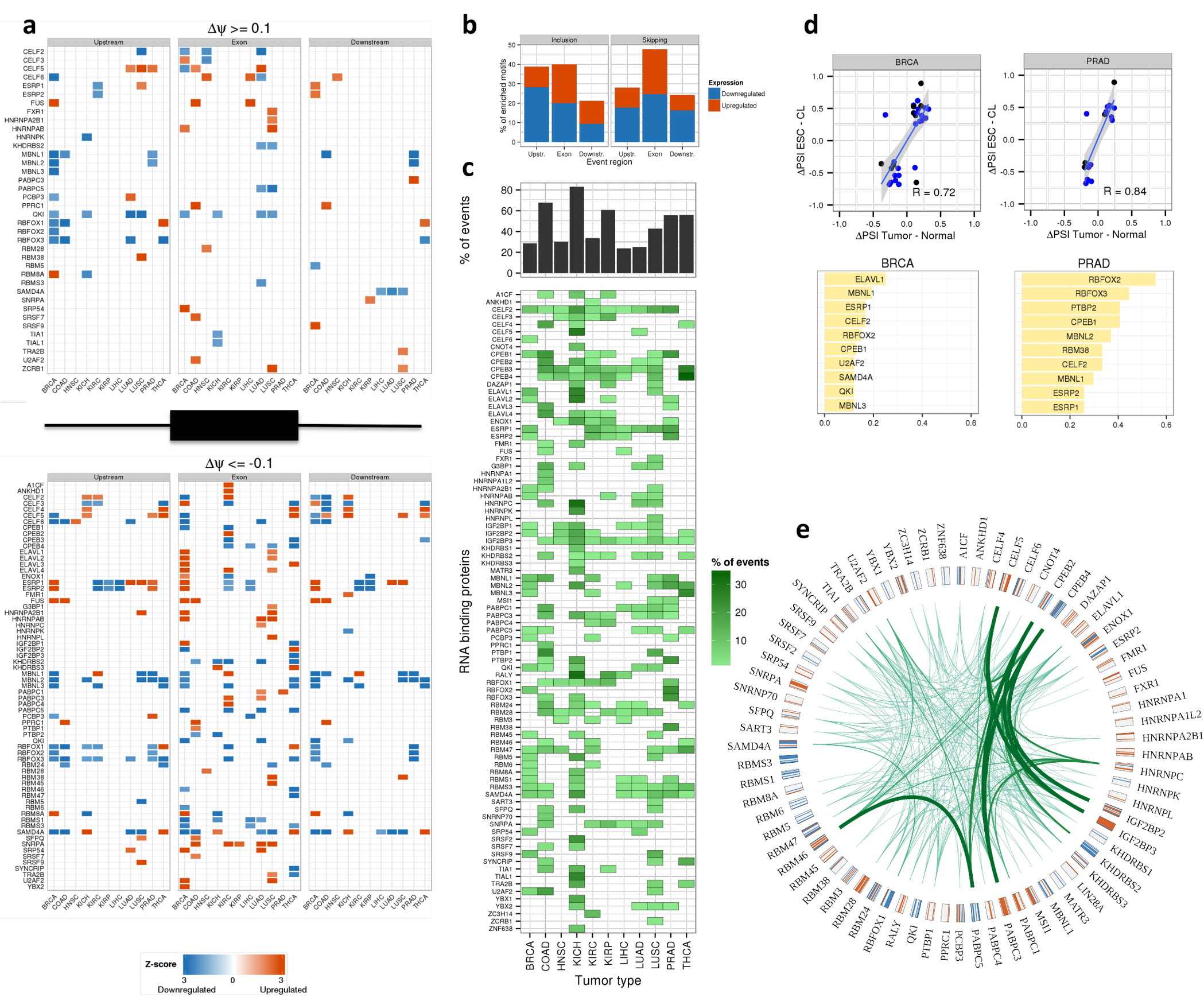
87. Clauset, A., Newman, M. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **70,** 066111 (2004).

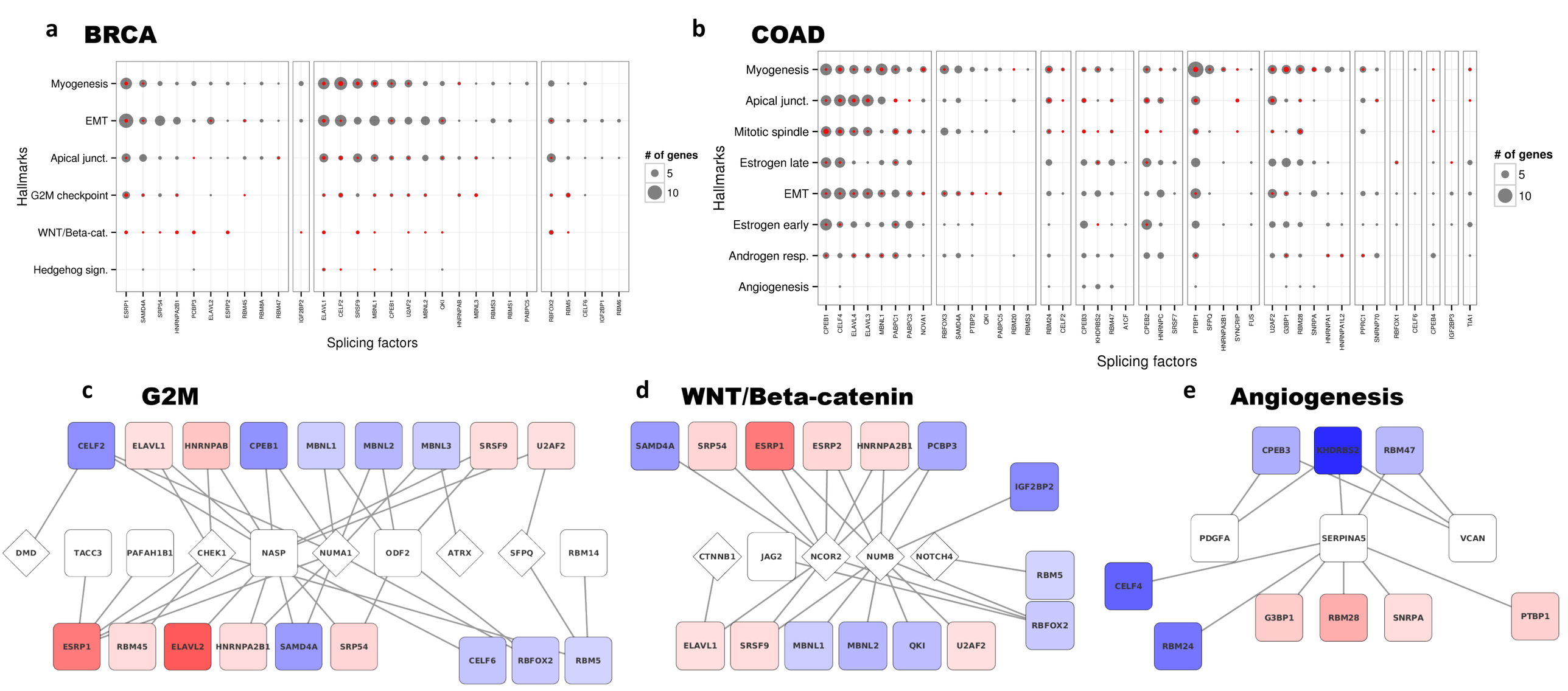


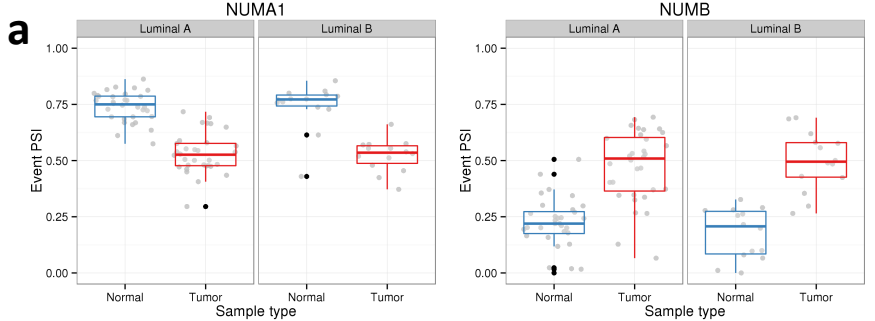


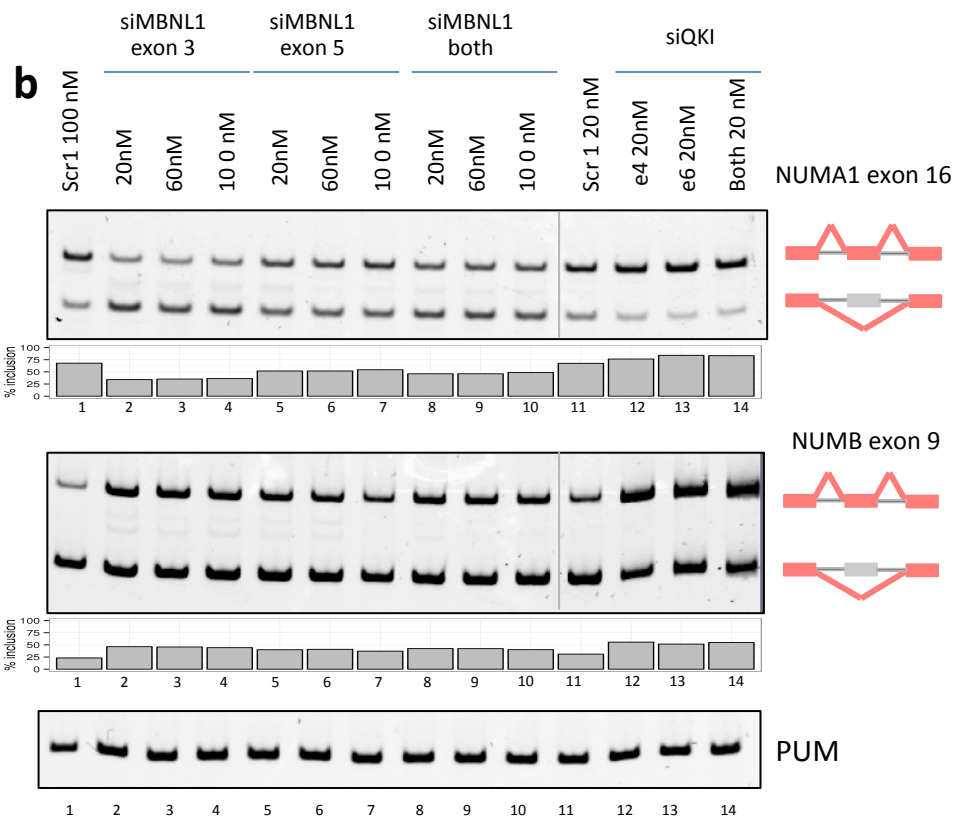

a
Total number
Samples
Events
BRCA COAD HNSC KICH KIRC KIRP LIHC LUAD LUSC PRAD THCA
Tumor type
A3 A5 MX RI SE
b
% of genes with regulated event
Gene type
Driver
Not driver
Tumor type
c
Tumor type
Odds
Hallmarks
Androgen resp.
Angiogenesis
Apical junct.
E2F targets
EMT
Estrogen early
Estrogen late
G2M checkpoint
Hedgehog sign.
Mitotic spindle
Myogenesis
Notch sign.
UV resp. (down)
WNT Beta-cat.
d
% mutated sample
Tumor type
BAT2L1 CDK12 CHD2 CHD3 CNOT1 FLNB GCN1L1 HNRNPL KIAA0802 MACF1 MDN1 MYH9 MYO18A NBPF10 NOMO1 PCDHGA9 RBM10 SF3B1 SFRS2IP SPEN TCF20 TWISTNB U2AF1 WDR33 YLPM1 ZC3H18
e
Regulated event number
Tumor type
RNA binding proteins
f
HNRNPL in COAD
Mutation number
Amino acid position
RRM
Nucleotide-binding alpha-beta plait
P204DEL
P204INS
C85Y
P286S
R362H
Mutation type
Missense Mutation
Frame Shift Ins
Frame Shift Del
g
HNRNPL (COAD)
% of events
Event type
Event set
Regulated
All
h
CASP8
Event PSI
Normal
Tumor - NM
Tumor - M

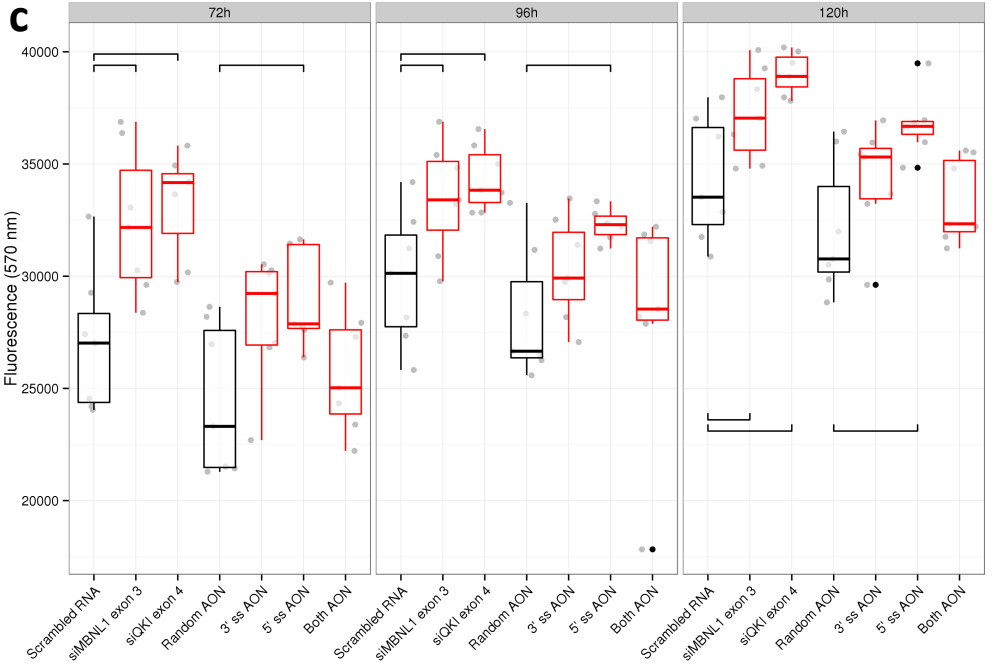


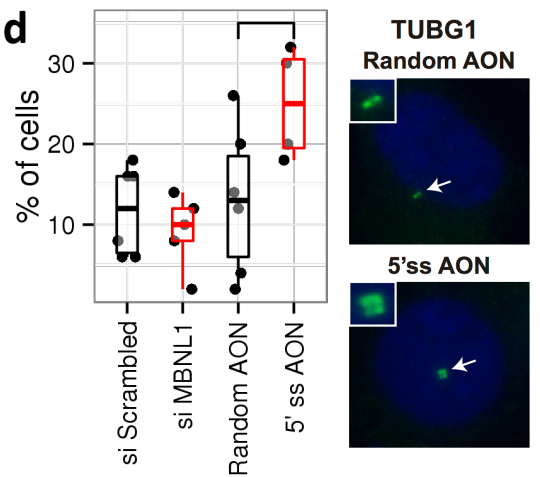

a
Δψ >= 0.1
Upstream
Exon
Downstream
CELF2
CELF3
CELF5
CELF6
ESRP1
ESRP2
FUS
FXR1
HNRNPA2B1
HNRNPAB
HNRNPK
KHDRBS2
MBNL1
MBNL2
MBNL3
PABPC3
PABPC5
PCBP3
PPRC1
QKI
RBFOX1
RBFOX2
RBFOX3
RBM28
RBM38
RBM5
RBM8A
RBMS3
SAMD4A
SNRPA
SRP54
SRSF7
SRSF9
TIA1
TIAL1
TRA2B
U2AF2
ZCRB1
BRCA COAD HNSC KICH KIRC KIRP LIHC LUAD LUSC PRAD THCA
Δψ <= -0.1
A1CF
ANKHD1
CELF4
CELF5
CELF6
CPEB1
CPEB2
CPEB3
CPEB4
ELAVL1
ELAVL2
ELAVL3
ELAVL4
ENOX1
FMR1
G3BP1
HNRNPA1
HNRNPC
HNRNPL
IGF2BP1
IGF2BP2
IGF2BP3
KHDRBS1
KHDRBS3
MATR3
MSI1
PABPC1
PABPC4
RALY
RBM24
RBM3
RBM45
RBM46
RBM47
RBM6
SFPQ
SRSF2
SYNCRIP
YBX1
YBX2
ZC3H14
ZNF638
Z-score
3
0
3
Downregulated
Upregulated
b
Inclusion
Skipping
% of enriched motifs
Upstr.
Exon
Downstr.
Event region
Expression
Downregulated
Upregulated
c
% of events
RNA binding proteins
Tumor type
d
BRCA
PRAD
ΔPSI ESC - CL
ΔPSI Tumor - Normal
R = 0.72
R = 0.84
ELAVL1
MBNL1
ESRP1
CELF2
RBFOX2
CPEB1
U2AF2
SAMD4A
QKI
MBNL3
RBFOX3
PTBP2
MBNL2
RBM38
ESRP2
e
% of events

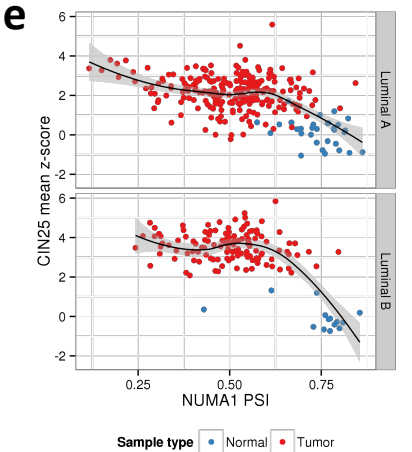

a
BRCA
Myogenesis
EMT
Apical junct.
G2M checkpoint
WNT/Beta-cat.
Hedgehog sign.
Hallmarks
ESRP1 SAMD4A SRP54 HNRNPA2B1 PCBP3 ELAVL2 ESRP2 RBM45 RBM8A RBM47 IGF2BP2 ELAVL1 CELF2 SRSF9 MBNL1 CPEB1 U2AF2 MBNL2 QKI HNRNPAB MBNL3 RBMS3 RBMS1 PABPC5 RBFOX2 RBM5 CELF6 IGF2BP1 RBM6
Splicing factors
# of genes
5
10
b
COAD
Myogenesis
Apical junct.
Mitotic spindle
Estrogen late
EMT
Estrogen early
Androgen resp.
Angiogenesis
Hallmarks
CPEB1 CELF4 ELAVL4 ELAVL3 MBNL1 PABPC1 PABPC3 NOVA1 RBFOX3 SAMD4A PTBP2 QKI PABPC5 RBM20 RBMS3 RBM24 CELF2 CPEB3 KHDRBS2 RBM47 A1CF CPEB2 HNRNPC SRSF7 PTBP1 SFPQ HNRNPA2B1 SYNCRIP FUS U2AF2 G3BP1 RBM28 SNRPA HNRNPA1 HNRNPA1L2 PPRC1 SNRNP70 RBFOX1 CELF6 CPEB4 IGF2BP3 TIA1
Splicing factors
# of genes
5
10
c
G2M
CELF2 ELAVL1 HNRNPAB CPEB1 MBNL1 MBNL2 MBNL3 SRSF9 U2AF2
DMD TACC3 PAFAH1B1 CHEK1 NASP NUMA1 ODF2 ATRX SFPQ RBM14
ESRP1 RBM45 ELAVL2 HNRNPA2B1 SAMD4A SRP54 CELF6 RBFOX2 RBM5
d
WNT/Beta-catenin
SAMD4A SRP54 ESRP1 ESRP2 HNRNPA2B1 PCBP3
IGF2BP2
CTNNB1 JAG2 NCOR2 NUMB NOTCH4
RBM5
RBFOX2
ELAVL1 SRSF9 MBNL1 MBNL2 QKI U2AF2
e
Angiogenesis
CPEB3 KHDRBS2 RBM47
PDGFA SERPINA5 VCAN
CELF4
G3BP1 RBM28 SNRPA PTBP1
RBM24


a
NUMA1
NUMB
Luminal A
Luminal B
Event PSI
Normal
Tumor
Sample type


b
siMBNL1 exon 3
siMBNL1 exon 5
siMBNL1 both
siQKI
Scr1 100 nM
20nM
60nM
10 0 nM
Scr 1 20 nM
e4 20nM
e6 20nM
Both 20 nM
NUMA1 exon 16
NUMB exon 9
PUM
% inclusion


c
72h
96h
120h
Fluorescence (570 nm)
Scrambled RNA
siMBNL1 exon 3
siQKI exon 4
Random AON
3' ss AON
5' ss AON
Both AON


d
% of cells
si Scrambled
si MBNL1
Random AON
5' ss AON
TUBG1
Random AON
5'ss AON


e
CIN25 mean z-score
NUMA1 PSI
Luminal A
Luminal B
Sample type
Normal
Tumor