Supplemental Materials

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$^{1,2}$, Amadis 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
Supplemental 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 the study (Supp. Table S1). Estimated read counts for genes and isoforms for the TCGA gene annotation (hg19, June 2011) were downloaded (version 2 Level 3). Mutation and copy number variation (CNV) data was downloaded from the TCGA data portal for all tumor types. MAF files containing Level 2 somatic mutation calls from whole exome and Level 3 SNP array data, containing the normalized copy number variation and purity/ploidy analysis results for each sample excluding germline copy number variations were used for the mutation and CNV data. BRCA subtypes were determined using the classification from TCGA (The Cancer Genome Atlas Network 2012b). COAD samples were classified as hypermutated if they had more than 250 mutations in total and as non-hypermutated otherwise (The Cancer Genome Atlas Network 2012a). Sample quality was assessed as previously (Sebestyén et al. 2015).

The 1348 genes coding for RNA binding proteins (RBPs) analyzed includes those with high confidence for RNA binding (Baltz et al. 2012; Castello et al. 2012; Kwon et al. 2013) and those annotated as RNA-binding in Ensembl (Cunningham et al. 2014) (Supp. Table S2). From this set, a subset of 162 known and potential auxiliary splicing factors (SFs) was selected (Supp. Table S2).

Differential expression

Quantile normalization and voom transformation was performed on gene read counts (Law et al. 2014). Differential expression was analyzed using the limma package (Smyth 2005) and p-values were corrected for multiple testing using the Benjamini-Hochberg method. Genes were considered differentially expressed if log₂-fold change > 0.5 in absolute value and corrected p-value < 0.05. An expression Z-score per gene and per tumor sample was calculated using the quantile normalized and voom transformed read-counts, $Z$-score = $(n-m)/(1.486 \cdot MAD)$, where $n$ is the value in the tumor sample, and $m$ and $MAD$ are the median and median absolute deviation, respectively, in the normal samples for the same tumor type. If the MAD was equal to
zero, the mean absolute deviation (MeanAD) of the normal samples was used: $Z$-score $= (n-m)/(1.253314 \cdot \text{MeanAD})$.

**Mutation and Copy number variation analysis**

The frequency of somatic mutations across all samples with available data was calculated per gene and tumor type. The top 5% genes according to mutation frequencies were considered for further analyses. The robust $Z$-score was used to calculate the Jaccard index between any regulation up ($Z$-score $> 1.96$) or down ($Z$-score $< -1.96$) with mutations. Additionally, mutual exclusion was measured using the number of samples having an RBP mutation and no expression change, $n_{10}$, and the number of samples having expression change but no RBP mutation, $n_{01}$. A mutual-exclusion score, $mx$, with values between 0 and 1, was defined as $mx = 2 \cdot \min(n_{10}, n_{01})/N$, where $N$ is the total number of samples with both mutation and expression data available. For samples with 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. We required that the full locus of the RBP fall within a copy number region. We defined focal copy number alterations as those smaller than 5Mb. The frequency of focal or all CNV gains and losses across all samples with available data was calculated per gene and tumor type and the top 5% of genes were considered for further analyses. The robust $Z$-score was used to calculate the Jaccard index between up regulation ($Z$-score $> 1.96$) or down regulation ($Z$-score $< -1.96$) with CNV gains or losses, respectively, for all changes focal or not.

**Alternative splicing events**

A total of 30820 alternative splicing events were calculated from the gene annotation using SUPPA (Alamancos et al. 2015): 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. Percent spliced-in (PSI) values per event and sample were calculated with SUPPA from the transcript quantifications in transcript per million (TPM) units. Only cases for which the transcripts defining the event had a total TPM$>0.1$ were considered. Differentially spliced events were obtained by comparing the PSI value distributions between 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 with the Benjamini-Hochberg method.

Events were considered differentially spliced if the difference between the tumor and normal median PSIs > 0.1 in absolute value and a corrected p-value <0.05. Non-regulated events were defined as having an absolute difference of median PSI ≤ 0.01 and corrected p-value > 0.05. To calculate differentially spliced events associated with mutations in RBPs, tumor samples were separated according to whether they had or not mutations in an RBP, and the two groups were compared using a Wilcoxon signed rank test, correcting for multiple testing. Tests were performed using protein-affecting mutations. Only RBPs with mutations in at least 10 samples were tested. An enrichment Z-score was calculated per RBP and tumor type by comparing the number of events changing significantly with the median value obtained using all RBPs tested.

Tumor type specific alternative splicing events were calculated by comparing their PSI values pairs of tumor types. An equal number of samples from each tumor pair were subsampled 100 times and compared using information gain (IG). For each pair, an average IG per event was calculated from the 100 iterations, keeping only events differentially spliced in at least one of the two tumor types, and the top 1% were kept as the most discriminating. All events obtained this way were finally combined into a non-redundant set. The list of differentially spliced events changing between embryonic stem cells (ESCs) and differentiated cells and tissues (Han et al. 2013), or in single knockdown or overexpression experiments (Brosseau et al. 2014) (Shen et al. 2012) (Vanharanta et al. 2014) were matched to the tumor events. Additionally, we analysed the ENCODE data corresponding to the RNA-Seq i in 2 biological replicates for the knockdown and controls in two cell lines, HepG2 and K562 (https://www.encodeproject.org/), for the RBP genes DAZAP1, FMRI, FUS, HNRNPA1, HNRNPA2B1, HNRNPA8, HNRNPC, HNRNPK, HNRNPL, KHDRBS1, MATR3, SART3, SRSF7, SRSF9, TIA1 and U2AF2. For each RBP we considered the tumor types where they are up or down regulated and compared the differentially spliced events of that tumor type with the events from the cell-line experiment that were either reported (MBNL1, QKI, RBFOX2, PTBP1, RBM47 and ESRP1) or calculated with SUPPA (ENCODE RBPs) with |ΔPSI| > 0.1. From the common events, we calculated the number of events that change in the expected direction according to the direction of regulation of the RBP in the tumor, i.e. correlating if the
RBP is down regulated or anti-correlating if the RBP is up regulated, and how many of each set have the RBP motif. Results are reported in Supplemental Table S8.

**Gene sets**

Annotations for 50 cancer hallmarks were obtained from the Molecular Signatures Database v4.0 (Liberzon et al. 2015) 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 cancer was collected from the literature (Supplemental Table S5). Additionally, a set of 889 cancer drivers based on mutations and CNVs (34 in common with previous set) (Supplemental Table S6) was obtained from a combination of COSMIC (October 2014) (Forbes et al. 2015), (Jelinic et al. 2014), 291 high-confidence tumor drivers (HCD) (Tamborero et al. 2013) and 260 drivers from the set Cancer5000 (Lawrence et al. 2014). This resulted in a total of 937 genes, collectively called drivers, which were labeled as oncogenes or tumor suppressors based on the annotations from COSMIC, (Vogelstein et al. 2013), and the TSGene database (Zhao et al. 2013). Unlabeled cases were predicted with OncodriveROLE (Schroeder et al. 2014) using cutoffs 0.3 (loss-of-function class) and 0.7 (activating class).

**RNA binding motif enrichment**

RNAcompete motifs (Ray et al. 2013) were used for 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 (Supplemental Table S2). Motifs from different species were used after confirming the conservation of the RNA binding domain in human: RBM47 (chicken), SF1 (Drosophila), SRP54 (Drosophila), TRA2 (Drosophila), and PCBP3 (mouse) (Supplemental Table S2). The tool fimo (Bailey and Elkan 1994) was used to scan the motifs in the event regions using p-value < 0.001. Motif enrichment analysis was performed by comparing the frequency of regions in differentially spliced events with a motif with 100 random subsamples of the same size from equivalent regions in non-differentially spliced events controlling for similar G+C content. An enrichment z-score per motif, region and direction of change (ΔPSI > 0.1, ΔPSI < -0.1) was
calculated by normalizing the observed frequency with the mean and standard deviation in the 100 random control sets. We considered a differentially spliced event to be a potential target of a differentially expressed RBP if the correlation between the event PSI value and the gene expression robust Z-score was $|R| > 0.5$ (Spearman) and the event contained the RBP binding motif. To assess significance, the same number of differentially spliced events in a tumor type was randomly selected from all events 100 times, and events associated to the RBPs calculated each time as described previously. A z-score was calculated from the mean and standard deviation. 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 (Mahony and Benos 2007) 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 events involved.

Networks of RBPs and events were built using the correlations between RBPs through events. For each pair of RBP a correlation 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 these correlations using the glasso algorithm (Friedman et al. 2008) and then searching for dense, highly connected sub-graphs with a greedy algorithm (Clauset et al. 2004). 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 2.5mM glutamine, 15mM HEPES, 10% FBS (Life Technologies, 10270), Pen-Strep (Life Technologies, 15070063), human insulin solution at 10ng/ml concentration (SIGMA, I9278), Hydrocortisone at 0.50µg/ml
(Merck, 386698) and human recombinant EGF from *E. coli* at 25ng/ml (Merck, 324831). 250,000 MCF10A cells were plated in 6-well plates and transfected in triplicate using 2µl Lipofectamine RNAiMax (Life Technologies, 13778150) per 1ml 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 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, 1.5mM MgCl2, 20mM TrisHCl pH7.5, 150 mM NaCl, 1% NP40) with 1x Complete protease inhibitor (ROCHE, 11697498001). MBNL1 siRNA (Life Technologies, s8553 and s8555), QKI siRNA (Life Technologies, s18084 and s18085) and RBM42 siRNA (Life Technologies, s35670) were used at 20, 60, 100nM concentrations, as well as Silencer® Select Negative Control No. 1 siRNA (Life Technologies, 4390843) at 20 and 100nM concentrations.

**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 0.45µM nitrocellulose membrane (Protran BA85 10401196, Whatman). MBNL1 monoclonal antibody (M02), clone 3E7 (ABNOVA, H00004154-M02), rabbit anti-QKI antibody (ABCAM, ab126742), rabbit anti-RBM42 (Abcam, ab171136), 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 were detected by Western Lightning Plus ECL chemiluminescence reagent (PERKIN-ELMER, NEL105001EA) and exposed to Kodak BioMax MR film (SIGMA, Z353949). For RBM42 Anti-RBM42 antibody - C-terminal (Abcam ab171136) was used.

**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); 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 Supplemental Table S23.

**Antisense oligonucleotides treatment**

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

NUMA1_ex16_5’ss: 5’- ggcacauacCUGCUUUAGUUGC-3’
NUMA1_ex16_3’ss: 5’- CCUCUAGCUGCUCCACcugu-3’
RANDOM 2’-O-Methyl RNA oligo: 5’- GCAAGGCGUCAAGUGUGUGUCG-3’

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

**Cell proliferation/viability assay**

2500 MCF10A cells/well were seeded the night before treatment in 96-well plates (NUNC, 167008) in 100µl 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 37°C 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.

**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 cold methanol for 10 minutes and washed in phosphate-buffered saline. The secondary antibody was Alexa Fluor 488 (Molecular Probes, Life Technologies) 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 (Supplemental Table S21). The chromosome instability signature CIN25 (Carter et al. 2006) 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.

**References**


Supplemental Figures
Figure S1. Up (orange) and downregulation (blue) of the expression of RNA binding proteins (RBPs) in different tumor types compared to normal samples. The intensity of the color indicates the log2-fold change (log2 FC). The barplot above indicates the frequency of tumor types with up (red) or down (blue) regulation for each RBP.
**Figure S2.** Dendogram (top) and heatmap (bottom) corresponding to 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. Samples are color-labeled according to tumor type. The plot includes the mutations (upper heatmap), copy number variations (CNVs) (middle heatmap). The mutation panel shows only factors with mutations in 7% or more of the tumor samples in a tumor type. CNVs are separated according to amplifications (red) and deletions (blue) and are only shown for genes with strongest associations between amplification and upregulation or deletion and downregulation. Samples with no CNV data available are indicated in gray. Bottom panel heatmap shows the expression Z-scores with orange and blue showing Z > 0 and Z < 0, respectively (Methods). LUAD and LUSC samples separate into two large subgroups, with the majority of LUSC samples showing frequent upregulation of TRA2B. KIRC and KIRP samples cluster together and separately from KICH. The prostate adenocarcinoma (PRAD) and HNSC samples cluster closely and show a general pattern of low expression variation.
Figure S3. Dendrograms for the unsupervised clustering of the 4442 tumor samples normalized using the robust z-score expression values (Methods) for 1348 genes encoding known and putative RNA binding proteins (RBPs) (a) and 1426 genes encoding transcription factors (TFs) (b). Samples are color-labeled according to tumor type. Heatmaps of the unsupervised hierarchical clusterings for 4442 tumor samples for RBP genes (c) and TF genes (d). Genes are given along the y-axis whereas samples are along the x-axis. Cells show the robust z-score expression relative to normal samples (Methods) with orange and blue showing $Z > 0$ and $Z < 0$, respectively.
Figure S4. Dendrogram (top) and heatmap (bottom) for the clustering of 1036 tumor samples from BRCA and the 262 tumor samples from COAD according to expression Z-score. Samples are color-labeled according to subtypes (Methods). Mutations and CNVs are indicated as in Supp. Fig. 2. 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 (Yae et al. 2012).
Figure S5. Analogous to Figure 1a, but for 1036 tumor samples from breast invasive carcinoma (BRCA) and the 262 tumor samples from colon adenocarcinoma (COAD) separated by subtype (Methods): breast luminal A (BRCA-LumA) and B (BRCA-LumB), basal (BRCA-Basal), HER2+ (BRCA-HER2), colon hypermutated (COAD-H+), and colon non-hypermutated (COAD-H−).
Figure S6. Barplots with the number of events (y-axis) that appear in a given number of tumor types (x-axis) for putative cancer drivers (a) and non-drivers (b). The event count is separated according by type (legend). Putative cancer drivers (c) and non-drivers (d) that have regulated events in 4 and 6 or more cancer types, respectively. The color indicates whether the regulated event has ΔPSI > 0.1 (orange) or ΔPSI < 0.1 (blue).
Figure S7. Enrichment of cancer hallmarks in genes differentially expressed in each tumor type. The plot indicates whether specific sets of genes defining a hallmark (x-axis) are enriched (Fisher test p-value < 0.05) in differentially expressed genes in each tumor type (y-axis). The color indicates the odds ratio of the enrichment.
Figure S5. (a) Proportion of samples with mutations in each tumor type for RBPs that have protein-affecting mutations in at least 10 samples and have > 0 associated significantly differentially spliced events. (b) Number of differentially spliced events related to mutations of RBPs that are significant after correcting for multiple testing (Methods).
Figure S9. (a-e) Patterns of protein-affecting mutations in the RBPs with the largest number of associated differentially spliced events, RBM10, TCF20, U2AF1, MACF1, plus SF3B1 for comparison. (f-j) Enrichment or depletion of specific event types in association to mutations in these RBPs compared to the overall event proportions. Cases with Fisher's test p-value < 0.05 are labeled in red. Contingency tables are provided as Supplementary Table.
Figure S10. For each pair of tumor types (x and y axes) and for each splicing factor (SF) (indicated in each legend) that is differentially expressed in both tumor types, we plot the ΔPSI values for events that have a correlation of |R| > 0.5 (Spearman) with these SFs in both tumor types. These include all regulated and non-regulated events. Only SFs with more than 50 events with |R| > 0 in both tumor types are considered. We only show those cases for which the correlation of events if |R| > 0.8 from Figure 3. We also included the BRCA and COAD comparison as it includes MBNL1 and QKI, which we test experimentally breast cancer cells. The correlations for the individual factors are given in Supplementary Table S17.
**Figure S11. Tumor type specific events.** Dendrogram (a) and heatmap (b) for the clustering of the 380 events according to PSI values that are in the top 1% separating at least two tumor types based on their PSI values (Methods). Clustering is performed using Gower distance and Ward method. (b) and (c) PSI value distributions in normal (blue) and tumor (red) samples in each cancer type for the events in the genes NAGS and KALRN found to be specific of KICH and KIRC, respectively.
Figure S12. (a). Regions used for motif sequence analysis. For the skipping exons (SE) and retained introns (RI) events, the full sequence of the regulated exons and intron, respectively, is considered. (b) List of the RBPs for which we have motifs that show either differential expression (DE) and/or motif enrichment (EN) in differentially spliced events in each tumor type. (c) Proportion of the differential expressed (DE) RBPs that have motif enrichment (EN) in the differentially spliced events.
Figure S13. The plot shows the similarities between the motifs selected from RNAcompete (Supplementary Table S2) using STAMP (Mahoney et al. 2007). The motifs were clustered using Euclidean distance and Ward method. The version of STAMP used is a modified one where the motifs are not reverse-complemented when compared, which is available at https://github.com/shaunmahony/stamp
Figure S14. Motifs associated to differentially expressed RNA binding proteins that are enriched in alternative 3′ splice site (a) and 5′ splice site (b) events differentially spliced between tumor and normal (see Methods for details). Only motifs with (median-based) z-score > 1.96 are shown. The motif enrichment is indicated in red or blue if the corresponding RBP is up- or downregulated in the tumor type. The intensity of the color is proportional to the z-score. RBPs are separated into groups according to Figure 1A.
Figure S15. Motifs associated to differentially expressed genes encoding for RNA binding proteins that are enriched in mutual exclusion events that are differentially spliced between tumor and normal. Only motifs with (median-based) z-score $> 1.96$ are shown. The motif enrichment is indicated in red or blue if the corresponding RBP is up- or downregulated in the tumor. The intensity of the color is proportional to the motif z-score.
Figure S16. Motifs associated to differentially expressed RNA binding proteins that are enriched in intron retention events differentially spliced between tumor and normal (see Methods for details). Only motifs with (median-based) z-score > 1.96 are shown. The motif enrichment is indicated in red or blue if the corresponding RBP is up- or downregulated in the tumor type. The intensity of the color is proportional to the motif z-score.
Figure S17. Enriched motifs in regulated skipping exon (a) and retained intron (b) events in luminal A and B breast tumors, separated by inclusion (upper panels) or skipping (lower panels); and upstream (left), exonic (middle) or downstream (right) regions. Only enriched motifs for splicing factors that are differentially expressed in each tumor type are indicated. Splicing factors up- and downregulated are indicated in red and blue, respectively. The color gradient indicates the Z-score of the motif enrichment.
Figure S18. Motifs associated to differentially expressed RNA binding proteins that are enriched in A3 (q), A5 (b), and MX (c) events in luminal A and B BRCA tumors. The motif enrichment calculation is performed as explained in Methods. The motif enrichment is indicated in red or blue if the corresponding splicing factor is up- or downregulated in the tumor set. The intensity of the color is proportional to the z-score of the motif enrichment.
Figure S19. For each differentially expressed splicing factor (y-axis) we calculate the proportion (x-axis) of cancer drivers with differentially spliced events whose PSI correlate with the splicing factor expression (Spearman $|R| > 0.5$) and contains the splicing factor binding motif. We only show the top 10 splicing factors in each tumor type.
Figure S20. Correlation of ΔPSI values for the differentially spliced events in each tumor type with those in common with the joint knockdown of MBNL1 and MBNL2 in 293T (a) and HeLa (b) cells (Han et al., Nature 2013). We only show tumors with a correlation |R| > 0.5 (Spearman). Tumors where MBNL1 is differentially expressed are inside the blue boxes. Events with a predicted RNA binding motif for the same factor are indicated with blue dots.
Figure S21. Correlation of ΔPSI values for the differentially spliced events in BRCA Luminal A and Luminal B tumors that are common to those found upon knockdown of MBNL1+MBNL2 (Han et al., Nature 2013), QKI (Brosseau et al., RNA 2014), PTBPT1 (Han et al., Nature 2013), RBFOX2 and PTBP1 (Brosseau et al., RNA 2014). Events with a predicted RNA binding motif for the same factor are indicated in blue. Only cases with | R | > 0.5 (Spearman) are shown.
Figure S22. Correlation of ΔPSI values for the differentially spliced events in tumors that are common to those found upon the knockdown of QKI (Brosseau et al., RNA 2014), PTBPT1 (Han et al., Nature 2013), RBFOX2 (Brosseau et al., RNA 2014) and RBM47 (Vanharanta et al., eLife 2014), and ESRP1 overexpression (Shen et al., Nucleic Acids Res 2012). We only show tumors with a correlation $|R| > 0.5$ (Spearman). Tumors where the factor is differentially expressed are inside blue boxes. Events with a predicted RNA binding motif for the same factor are indicated with blue dots. PTBP1 dependent events show correlation in KIRP, LIHC and PRAD, even though PTBP1 is not differentially expressed in these tumors.
Figure S23. Correlation of ΔPSI values for the differentially spliced events with events from (Han et al. 2013) comparing with the ΔPSI values obtained from stem cells (ESCs) vs. differentiated cells (CL) (a), and from ESCs vs. differentiated tissues (Diff) (b). Events with a predicted MBNL motif are indicated in blue. We only show tumors with $|R| > 0.4$ (Spearman).
**Figure S24.** Modules of alternative splicing regulation according to cancer hallmarks in different tumor types. For each cluster of splicing factors (x-axis) we indicate in gray the total number of genes associated to these factors in each of the hallmarks found enriched in each tumor type (y-axis). 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.
Figure S25. Modules of alternative splicing regulation according to cancer hallmarks in different tumor types. For each cluster of splicing factors (x-axis), we indicate in gray the total number of genes associated to these factors in each hallmark. For each cluster of splicing factors (y-axis), 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.
Figure 526: Potential regulators of cancer drivers differentially spliced in BRCA luminal A (a) and B (b) tumors. The plot shows the cancer drivers with differentially spliced events (y-axis) and the set of splicing regulators that are differentially expressed, correlated with those events. If the |R| Spearman > 0.5 and whose motifs are present in the event (x-axis), we highlight in red those events and factors that appear frequently associated and that include the MBNL genes.
Figure S27. (a) Predicted regulators of NUMA1 exon 16 alternative splicing. The cartoon shows the splicing factors that are differentially expressed in breast tumors, have a high correlation (Spearman |R| > 0.5) with NUMA1 exon skipping PSI value, and their motif is found in the event. In red or blue we indicate those factors that are up- or downregulated. Their upstream and downstream position indicate the position of the predicted motif relative to the alternative exon. (b) Sequence context of NUMA1 exon 16, indicating the predicted MBNL1 motif and its conservation in other species.
Figure S28. (a) Western blots of MBNL1 and QKI proteins in MCF10A, MCF7, MDA-MB-231, BT474, MDA-MB-436, MDA-MB-468 and A459 cell lines. (b) Left panel: western blots of MBNL1 in cells where MBNL1 is depleted at various siRNA concentrations (lanes 2,3,4) compared to a control (lane 1). Middle panel: western blots of QKI in cells where QKI is depleted at various siRNA concentrations (lanes 2,3,4) compared to a control (lane 1). Right panel: in duplicate experiments, western blots of MBNL1 in cells where either MBNL1 or QKI are depleted compared to control (c) RT-PCR validation of NUMA1 alternative splicing in MCF10A cells transfected with siMBNL1 (lanes 3 and 4) or siQKI (lanes 5 and 6) compared to cell lines MCF10A (lane 7), MCF7 (lane 8), MDA-MB-231 (lane 9), A549 (lane 10) and HCT-116 (lane 11) and to scrambled siRNAs in MCF10 cells (lanes 1 and 2). We indicate below each lane the percent skipping of the NUMA1 exon 16. (d) RT-PCR validation of NUMA1 alternative splicing in MCF10A (upper panel) and MCF7 (lower panel) cells transfected with siRBM42 (lanes 7-12) compared to scrambled siRNAs (lanes 1-6). We indicate below each lane the percent skipping of the NUMA1 exon 16.
Figure S29. (a) Antisense RNA oligonucleotides (AON) (2'-O-Methyl RNA oligos) targeting both splice-sites of exon 16 in NUMA1. The image shows the sequence of the AONs, and the sequence of the regulated exon and part of the flanking introns. Below we show the sequence context of the exon 16, indicating a conserved MBNL1 binding motif in the downstream intron. (b) Effects of AONs targeting the 3' splice-site (lanes 3, 4), 5' splice-site (lanes 5, 6) or both (lanes 7, 8) of NUMA1 cassette exon 16, compared with the knockdowns (siRNAs) of MBNL1 and QKI (lanes 14-17). Lanes 1 and 2 show the controls with randomized AONs. Lane 10 shows the non-template control (NTC) (PCR without any nucleic acids) and lane 11 shows the molecular weight marker (MWM). Lanes 12 and 13 show the knockdown controls with randomized siRNAs. The quantification of the lanes are indicated below.
Figure S30. (a) Correlation of NUMA1 exon skipping PSI (x-axis) with the CIN25 signature of aneuploidy (y-axis) (Methods) across the tumor (red) and normal (blue) samples for the different tumor types. The Spearman correlation R is given for each case. (b) Representation of the region of the NUMA1 gene and protein that changes upon the exon skipping described, and representation of GPS results with the phosphorylation sites predicted in both isoforms. The plot shows the GPS scores, putative kinases targeting a certain site, and the grey area shows the region lost during exon skipping.