Combined analysis of genome sequencing and RNA-motifs reveals novel damaging non-coding mutations in human tumors

Babita Singh[1], Juan L. Trincado[1], PJ Tatlow[2], Stephen R. Piccolo[2,3], Eduardo Eyras[1,4,*]

[1]Pompeu Fabra University (UPF), E08003 Barcelona, Spain.

[2]Department of Biology, Brigham Young University, Provo, Utah, USA

[3]Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA

[4]Catalan Institution for Research and Advanced Studies (ICREA). E08010 Barcelona, Spain

*Correspondence to: eduardo.eyras@upf.edu

Abstract

A major challenge in cancer research is to determine the biological and clinical significance of somatic mutations in non-coding regions. This has been studied in terms of recurrence, functional impact, and association to individual regulatory sites, but the combinatorial contribution of mutations at common RNA regulatory motifs has not been explored. We developed a new method, MIRA, to perform the first comprehensive study of significantly mutated regions (SMRs) with overrepresented binding sites for RNA-binding proteins (RBPs) in cancer. We found multiple RBP motifs, including SRSF10, PCBP1 and HNRPLL motifs, as well as a specific subset of 5' and 3' splice-site sequences, enriched in cancer mutations. Gene targets showed association to cancer-related functions, and analysis of RNA sequencing from the same samples identified alterations in RNA processing linked to these mutations. MIRA facilitates the integrative analysis of multiple genome sites that operate collectively through common RBPs and can aid in the interpretation of non-coding variants in cancer. MIRA is available at https://github.com/comprna/mira.

Keywords: cancer, non-coding mutations, RNA-processing, RNA binding proteins

Introduction

Cancer arises from genetic and epigenetic alterations that interfere with essential mechanisms of the normal life cycle of cells such as DNA repair, replication control, and cell death (Hanahan and Weinberg 2011). The search for cancer driver mutations, which confer a selective advantage to cancer cells, has been traditionally performed in terms of how they directly affect protein sequences (Vogelstein et al. 2013). However, systematic studies of cancer genomes have highlighted relevant mutational processes outside of protein-coding regions (Alexandrov et al. 2013b; Weinhold et al. 2014; Juul et al. 2017) and tumorigenic mutations at non-coding regions have been described, like those found in the TERT promoter (Horn et al. 2013; Huang et al. 2013). However, a major challenge remains to more accurately and comprehensively determine the significance and potential pathogenic involvement of somatic variants in regions that do not code for proteins (Piraino and Furney 2016). Current methods to detect potential driver mutations in non-coding regions have been generally based on 1) the recurrence of mutations in predefined regions in combination with measurement of potential functional impacts (Melton et al. 2015; Fredriksson et al. 2014; Mularoni et al. 2016; Weinhold et al. 2014), 2) recurrence in combination with sequence conservation or polymorphism data (Khurana et al. 2016; Piraino and Furney 2017), or 3) the enrichment of mutations with respect to specific mutational backgrounds (Lochovsky et al. 2015; Lanzós et al. 2017; Juul et al. 2017); and some of them have combined multiple such approaches (Juul et al. 2017). However, these methods have so far been restricted to individual genomic positions rather than combining the contributions from multiple functionally equivalent regulatory sites and additionally, have not evaluated the impact on RNA processing measured from the same patient samples.

RNA molecules are bound by multiple RNA binding proteins (RBPs) with specific roles during the different steps of RNA processing, including RNA splicing, stability, localization and translation, and are critical for the proper control of gene expression (Maslon et al. 2014; Rissland 2017). RBPs can act as auxiliary—and sometimes necessary—factors to regulate RNA processing, and in particular splicing, and often to antagonize each other in normal cellular programs and disease states (Fu and Ares 2014). Multiple experimental approaches have established that RBPs generally interact with RNAs through short motifs of 4-7 nucleotides (Ule et al. 2003; Lambert et al. 2014; Ray et al. 2013; Oberstrass et al. 2005). These motifs occur anywhere along the precursor RNA molecule (pre-mRNA), including introns, protein coding regions, untranslated 5' and 3' regions, as well as in short and long non-coding RNAs (Sterne-Weiler and Sanford 2014; Haerty and Ponting 2015; Michlewski et al. 2008).

Mutations on RNA regulatory sequences can impact RNA processing and lead to disease (Soemedi et al. 2017). Studies carried out so far on mutations affecting RNA processing and alternative splicing have mainly focused on a fraction of the motifs associated with the core splicing machinery (Jung et al. 2015), or to protein-coding regions (Supek et al. 2014; Anczuków et al. 2015). Mutations at exon-intron boundaries have been associated with intron retention in tumor suppressors (Jung et al. 2015), whereas mutations on coding exons can trigger oncogenic splicing changes (Supek et al. 2014; Anczuków et al. 2015). *In vitro* screenings of sequence variants in exons has revealed that more than 50% of nucleotide substitutions can induce splicing changes (Ke et al. 2011; Julien et al. 2016), with similar effects for synonymous and non-synonymous sites (Julien et al. 2016). Since RBP binding motifs are widespread along gene loci, and somatic mutations may occur anywhere along the genome, it is possible that mutations in other genic regions could impact RNA processing and contribute to the tumor phenotype. Although mutations and expression alterations in genes coding for RNA binding proteins (RBPs) have an impact on specific cellular programs in cancer, it is not known yet if mutations in the binding sites of RBPs are frequent in cancer, could be damaging to RNA processing, and contribute to oncogenic mechanisms.

To understand the effects of somatic mutations on RNA processing in cancer at a global level we have developed a new approach called MIRA to carry out a comprehensive study of somatic mutation patterns in exons and introns from coding and non-coding genes that operate collectively through interacting with common RBPs. Compared with other existing approaches to detect relevant mutations in non-coding regions, our study provides several novelties and advantages: 1) we searched exhaustively along gene loci, hence increasing the potential to uncover deep intronic pathological mutations; 2) we studied the enrichment of a large compendium of potential RNA regulatory motifs, allowing us to identify potentially novel mechanisms affecting RNA processing in cancer; 3) we showed that multiple mutated genomic loci potentially interact with common RBPs, suggesting novel cancer-related mechanisms; and 4) unlike previous methods, we used RNA sequencing data from the same samples to measure the impact on RNA processing. Our study uncovered multiple sites common to RBPs that impact RNA processing of functions with potential implications in cancer. This study reveals a new layer of insight to aid in the functional interpretation of somatic alterations in cancer, and could also help in interpreting the clinical relevance of non-coding variants in cancer genomes.

Results

Unbiased search for significantly mutated regions (SMRs) along gene loci

RNA binding proteins (RBPs) generally interact with pre-mRNAs at motifs of 4-7 nucleotides, which occur anywhere along the pre-mRNA. We thus performed an exhaustive detection of mutation enrichment along overlapping genomic windows of 7 nucleotides (7-mers) along each gene locus (Fig. 1a) (Methods). Using a dataset of somatic mutations from whole-genome sequencing (WGS) from 505 samples for 14 tumor types (Fredriksson et al. 2014) (PAN505) (Table S1), we performed a double statistical test. First, to account for local variations in mutational processes, we tested each 7-mer window for enrichment in number of mutations by comparing each window against the mutation rate for the entire gene locus, and we selected an enrichment p-value threshold of < 0.05 after correcting for multiple tests (Figure S1a)(Methods). Secondly, to account for nucleotide biases, we compared the mutation count in each 7-mer window with the expected mutation count calculated from the nucleotide sequence of the window and the mutation rate per nucleotide at the same gene locus. With this we defined a nucleotide bias (NB) score per window as the log2-likelihood of the observed versus the expected counts (Methods). Of the 140,704 windows with 3 or more mutations, 93,497 (66%) showed NB-score > 6, whereas of the 45,916,437 7-mer windows with 1 mutation, which we considered to reflect the mutational background, 1,557,310 (3%) had NB-score > 6 (Chi-square test p-value < 2.2e-16) (Additional file 2: Figure S1b). We thus selected NB-score > 6 as a cutoff. After applying these filters (corrected p-value < 0.05 and NB score > 6), our exhaustive analysis produced a total of 78,352 significant 7-mer windows in 8,159 genes.

The functional impact of somatic mutations depends on the specific genic region in which they fall. We thus separated the 7-mer windows according to whether they were in a 5' or 3' untranslated region (5UTR/3UTR), a coding sequence (CDS), an exon in a long non-coding RNA (EXON), an intron (INTRON), or in a 5' or 3' splice site (5SS/3SS) (Fig. 1a) (Methods). These windows were then clustered into significantly mutated regions (SMRs), producing a total of 20,307 SMRs, containing a total of 41,756 substitutions (Figure S1c) (Tables S2 and S3), which we considered for further analysis. Most of the predicted SMRs were 7-15 nucleotides long (Figure S2), and the majority of SMRs were in introns or in exons of non-coding RNAs (EXON) (Table 1).

SMRs	Total	With enriched motifs	With labeled enriched motifs	Impact on RNA-processing
3SS	823	341	335	2
3UTR	294	44	44	7
5SS	1054	546	521	11
5UTR	119	45	45	5
CDS	208	10	10	3
EXON	335	119	114	12
INTRON	17474	3176	1812	427

Table 1. Significant mutated regions (SMRs). For each region type, we indicate the total number of SMRs predicted (Total), SMRs with stranded enriched motifs (with enriched motifs), with stranded enriched motifs that we could label (with labeled enriched motifs), and with a significant association to an RNA-processing change (impact on RNA-processing). In this latter case we count changes in exon-exon junctions and transcript expression changes.

Validation of predicted SMRs

To test our predicted SMRs for possible biases we calculated in each region the DNA replication timing, which is known to correlate with somatic mutations in cancers and can be a source of artefacts in mutational driver predictions (Lawrence et al. 2013; Liu et al. 2013); we observed no association with mutation count (Figure S3). Another potential source of artefacts is the relation between gene expression and the rate of somatic mutations (Lawrence et al. 2013). We used RNA sequencing (RNA-seq) data from the same samples to measure the expression of transcripts whose genomic sequence contained significant regions and observed no association between the mutation count and expression (Figure S4). To further validate our SMRs we used LARVA (Lochovsky et al. 2015) to assess their significance using a statistical model that accounts for over-dispersion of the mutation rate and replication timing (Methods). We observed an overall high similarity between the significance provided by our method and that given by LARVA (Figure S5). In particular, we found a strong agreement for INTRON SMRs (Pearson's R = 0.83), providing support for our intronic predicted regions.

Predicted SMRs recovered known and novel mutational hotspots

We found SMRs in 501 cancer-driver genes out of 889 collected from the literature (Sebestyén et al. 2016) (Fig. 1b) (Figure S7), which is more than expected by chance (Fisher's exact test p-value = 2.2e-16, odds-ratio = 4.6, comparing cancer/non-cancer genes tested with/without SMRs). In particular, we recovered SMRs in 71 genes out of the 108 previously identified with a method that measured the functional impact of mutations (Mularoni et al. 2016). In agreement with prior findings (Mularoni et al. 2016), we observed CDS SMRs in a total of 34 cancer genes, including *BRAF*, *IDH1*, *KRAS*, *PIK3CA* and *PIK3R1* (Fig. 1b). We also found SMRs in *SF3B1*, *CTNNB1*, *TP53* and *KRAS,* which were recovered before with a method based on mutation enrichment and evolutionary conservation (Piraino and Furney 2017). We also found CDS SMRs in cancer genes not found previously, including *NRAS*, *EP300* and *ATM* (Fig. 1c). From the 133 genes with predicted 5UTR SMRs, 17 were identified previously (Mularoni et al. 2016); and we found 5UTR SMRs in 8 cancer genes, including *SPOP* and *EEF1A1*. We found 3UTR SMRs in 3 of the 12 different genes identified before (Mularoni et al. 2016), and found 3UTR SMRs in 28 cancer genes, including *CTNNB1* and *FOXP1*.

From the 519 SMRs in exon of long non-coding RNAs (lncRNAs), which we called EXON SMRs, 18 were located in cancer gene loci. Additionally, of 42 lncRNAs related to cancer (Lanzós et al. 2017), we found only one EXON SMR in *TCL6*. On the other hand, we found INTRON SMRs for 11 of these 42 lncRNAs, indicating that intronic regions in lncRNAs could be more relevant than previously anticipated. From the 13 genes reported as intronic in (Mularoni et al. 2016), we found 6 as INTRON SMRs and 5 as 5SS/3SS SMRs, with the genes *ATG4B*, *NF1* and *TP53* having both types of SMRs. As our analysis was exhaustive along the entire gene loci, we recovered many more intronic SMRs than in previous reports. In particular, we found INTRON SMRs in 317 cancer genes, including *NUMB, ALK*, *EPHB1*, *ARID1A*, *TP73* and *MET* (Fig. 1c). Finally, we found 62 5SS SMRs and 53 3SS SMRs in cancer genes, including *MET, CHEK2*, *BRCA1*, *VEGFA*, *RB1*, *CDKN2A*, as well as *TP53*, *PTEN* and *CHD1*, which were described before to have mutations at splice-sites (Jung et al. 2015). In summary, our SMRs provide a rich resource with potential to understand non-coding alterations in cancer.

Somatic mutations show positional biases on RBP binding motifs

To further understand the properties of our SMRs we calculated the mutation frequencies at trinucleotides considering the strand of the gene in which the SMR was defined. We observed an enrichment of C>T and G>A mutations on SMRs (Fig. 2a, upper panel). However, the stranded triplets showed mutation frequencies similar to their reverse complements, indicating that a considerable proportion of SMRs reflected DNA-related selection processes.

To identify those SMRs that more clearly reflect RNA-related selection processes we studied sequence motifs potentially related to RNA processing regulation. We performed an unbiased *k*-mer enrichment analysis in SMRs, using *k*=6. For each SMR type, we compared the proportion of SMRs in which each 6-mer occurs with the proportion in 100 control region sets (Methods). Further, to keep only those potentially associated with RNA rather than DNA, we reverse-complemented all SMRs and control sequences and repeated the enrichment analysis. Those 6-mers enriched in both calculations for the same region-type were eliminated (Figure S7). We found a total of 357 enriched 6-mers (Table S4) in 3546 SMRs from all region types (Table S4). Enriched 6-mers were AC-rich in CDS regions, GC-rich in 5UTR regions, and T-rich in 3UTR and EXON regions, whereas in 5SS and 3SS regions they were G and A rich, respectively (Figure S7). In contrast, INTRON motifs showed a uniform nucleotide content (Figure S7). Enriched 6-mers showed a different mutational pattern compared to that for all SMRs (Fig. 2a) (Tables S5 and S6). The symmetry between stranded triplets and their reverse complements in SMRs was no longer present in enriched 6-mers, and there was an enrichment of mutations at AGA and TCC triples that is not recapitulated in their reverse complements, indicating that the selected enriched 6-mers reflect RNA-related selection processes.

From the 74 enriched 6-mers found on 5SS SMRs, 46 contained the 5' splice-site (5ss) consensus GT. These enriched 5'ss motifs showed the consensus G|GT(A/G)AG with a strong conservation of G at the +5 intronic position (position 7 in Fig. 2b). The highest density of mutations occurred at two positions on either side of the exon-intron boundary (Fig. 2b), with mostly G>A and G>T substitutions (Fig. 2c). From the 52 enriched 6-mers found on 3SS SMRs, 36 contained the 3' splice site (3'ss) consensus AG, which showed a strong CNCAG|(G/A) motif, with strong conservation of C nucleotides at the -5 position of the intron (position 1 in Fig. 2c) and with positions -1 and -3 (3 and 5 in Fig. 2c) being the most frequently mutated positions, with the most frequent substitution C>T at position -3, and G>A or G>T at position -1 (Fig. 2c). Among the cases found, there was a 5'ss in *NF1* with mutations in skin (SKCM), lung (LUAD) and uterine (UCEC) tumors (Fig. 2d), and a 3'ss in *FGFR3* with mutations in head and neck (HNSC) and bladder (BLCA) tumors (Fig. 2e). Mutations at splice site motifs occurred in higher proportions in lung tumors (LUAD, LUSC) and in uterine tumors (UCEC), 5'ss mutations appeared also more frequent in bladder tumors (BLCA), whereas at 3'ss motifs were more frequent in colorectal tumors (CRC) (Fig. 3a).

To identify RBPs that could potentially bind the enriched 6-mers beyond 5'/3'ss motifs, we used DeepBind (Alipanahi et al. 2015) to score the 6-mers using models for 522 proteins containing KH, RRM and C2H2 domains from human, mouse and Drosophila (Figure S7)

(Methods). Using this procedure we labeled 245 (68.6% of the 357 enriched 6-mers; Table S3). In 5SS and 3SS SMRs, besides the enriched 5'ss or 3'ss consensus motifs, we identified binding sites for multiple RBPs (Figs. 3a and 3b), including motifs for SRSF10, a splicing factor that regulates an alternative splicing response to DNA damage (Chabot and Shkreta 2016). SRSF10 motifs appeared also at CDS and INTRON SMRs (Figs. 3c and 3d) (Figure S8). The observed differences in the consensus motif between region types may reflect differences in binding affinities and nucleotide composition, which have been observed before for other RBPs (Lovci et al. 2013). In 3SS SMRs we also found binding sites for HNRPLL (Fig. 3b), a regulator of T cell activation through alternative splicing (Oberdoerffer et al. 2008), which was also found on EXON and INTRON SMRs (Fig. 3d) (Figures S8).

To evaluate the mutational patterns on these motifs, we studied whether particular positions in the enriched motifs were more frequently mutated than others. For each RBP, we thus grouped the SMRs containing the enriched labelled 6-mers and performed a multiple sequence alignment (MSA) to determine the equivalent positions of the motifs across the SMRs and calculated the density of somatic and germline mutations per position (Methods). For SRSF10 motifs we found an enrichment of A>G mutations at A positions (Fig. 3d), which was recapitulated at INTRON, CDS and 3SS SMRs (Fig. 3c). HNRPLL and PCBP1 motifs showed an enrichment of C>T mutations in EXON SMRs (Figure S9). The most abundant RBP motifs on EXON SMRs were predicted to be for CPEB4, with a T-rich consensus and enriched in T>A mutations (Figure S9). CPEB4 (Ortiz-Zapater et al. 2012) and PCBP1 (Hwang et al. 2017) are known to bind 3'UTR regions of protein coding transcripts. It is possible that they also bind processed long-noncoding RNAs and that this binding is disrupted by cancer-related mutations.

At 5UTR SMRs we found multiple T- and C-rich motifs, which occurred predominantly in melanoma (SKCM) (Fig. 4a) (Table S3). In particular, PCBP3 and PTBP1 motifs at 5UTR SMRs were characterized by an enrichment of C>T substitutions (Fig. 4b) (Figure S10). These motifs might be related to the so-called 5' terminal oligo-pyrimidine tract (5'TOP) motif that is relevant for translational regulation and is bound by multiple RBPs (Sawicka et al. 2008; Pichon et al. 2012); these mutations found could indicate cancer-related alterations of translation. At 5UTR SMRs there were also G-rich motifs (HNRPLL, HNRNPA2B1) with frequent G>A mutations (Figure S10). At 3UTR SMRs, the representation of enriched motifs across tumor types was more variable (Fig. 4c). There were also CT-rich motifs, including PTBP1, HNRNPC, PCBP3, IGF2BP2 and ELAVL1 (HuR) (Figure S11), but these were more frequent in CRC, BLCA and UCEC patients (Fig. 4c). For HNRNPC and other RBPs, although C>T appeared as the most recurrent mutation, there were also frequent T>C and T>G mutations (Fig. 4d) (Figure S11). We also found the ACA-containing IGF2BP2 motif, which was more prominent in UCEC and CRC, and showed enrichment of C>T mutations (Figure S11).

For each enriched motif, we calculated the enrichment of gene sets (GO Biological Function, Pathways and Oncogenic Pathways) in the genes harboring SMRs with the motif (Table S7) (Methods). We found only a few motifs associated with cancer-related functions. Genes harboring SMRs with 5’ss motifs (5SSC) show enrichment in apoptosis and DNA damage response functions, as well as in genes related to NFKB activation and PI3K signal cascade (Figure S12). Genes with SMRs harboring SRSF10 motifs show association to apoptosis and immune response, whereas genes with HNRNPC or HNRNPLL motifs in SMRs are related to metabolic processes (Figure S12). RBM41-motif containing SMRs are also related to genes involved in metabolic processes, as well as in T-cell activation (Figure S12). These results indicate that cancer mutations on RBP motifs may impact to a wide range of functions besides contributing to oncogenic processes.

Somatic mutations in RBP binding motifs impact in RNA expression and splicing

To determine the impact of the described mutations in enriched motifs we decided to test their impact on RNA processing. We first estimated the association of mutations with changes in transcript isoform expression. For each patient with a mutated SMR we considered each transcript overlapping the SMR and compared its abundance with the distribution of abundances of the same transcript in patients from the same tumor type that did not harbor any mutation in the same SMR. From the 20,308 SMRs tested, 148 showed an association with a significant expression change (Fig. 5a) (Table S8) (Methods). Most of the significant changes were associated to INTRON SMRs in skin (SKCM) and colorectal (CRC) tumors (Figures S13 and S14). Motifs with mutations associated with a higher number of significant transcript expression changes included PTBP1, PCBP1, RBM8A and ZNF638 (Fig. 5b) (Figure S14). In particular, mutations in RBM8A motifs were associated with expression changes in transcripts of the histone acetyl transferase gene *KAT6A* (Turner-Ivey et al. 2014) and the pre-B-cell leukemia transcription factor 1 gene *PBX1*, both of which are potential oncogenes (Magnani et al. 2011). In the case of PBX1, two transcript isoforms change expression in opposite directions, indicating an isoform switch (Figure S14).

We also found mutations in an intronic SRSF10 motif associated with expression changes in a transcript from the dystrophin gene (*DMD)* in breast cancer (BRCA), and in two transcripts from the Mitogen-Activated Protein Kinase 10 *MAPK10 (JNK3)* in colorectal cancer

(CRC) (Fig. 5c). In *DMD,* the mutation was associated with increased expression, whereas in *MAPK10* we observed an isoform switch (Fig. 5c). *MAPK10* is a pro-apoptotic gene, whose alternative splicing has been observed in colon tumors (Ying et al. 2006; Sebestyén et al. 2016) Our results suggest that a mutation in an intronic SRSF10 motif conserved in primates is responsible for this splicing change in CRC (Fig. 5c). We also found a CRC mutation on an intronic MBNL1 motif conserved in mammals that is associated with an upregulation of the transcription factor ETV1 (Fig. 5d), which has been linked to prostate cancer (Cai et al. 2007). A mutation at a nearby site in SKCM appears to induce downregulation of a different transcript isoform (Fig. 5d).

In 5UTR SMRs, significant changes were associated only with SKCM mutations (Figure S15). One of these changes corresponds to a mutation on a PCBP3 motif associated with an isoform switch in the galactokinase-2 gene *GALK2* in melanomas (Figure S15). *GALK2* was found to be a regulator of prostate cancer cell growth (Whitworth et al. 2012) and showed extensive, alternative first-exon splicing in multiple tumor types (Sebestyén et al. 2016). Here we showed that these splicing alterations could stem from a mutation at the 5'UTR. Finally, among the changes associated with 3UTR SMRs, we found a mutation in a HuR (ELAVL1) motif in CRC related to the downregulation of the Debrin-like gene *DBNL* (Figure S15).

We also analyzed all possible exon-exon junctions defined from spliced reads mapped to the genome, and for each patient we compared the relative junction inclusion level overlapping a given SMR with the distribution of inclusion levels of the same junction in patients from the same tumor type but without mutations in the same SMR (Methods). Of the SMRs tested, 30 were associated with a significant inclusion change in at least one junction (Fig. 6a) (Table S9). The majority of cases were associated with INTRON SMRs, were more abundant in bladder tumors (BLCA) and head and neck squamous-cell tumors (HNSC). Significant junctions were also commonly associated with 5SS SMRs, mainly in uterine (UCEC) and skin (SKCM) cancers (Fig. 6a). Additionally, there were many significant associations in 5UTRs SMRs, specifically for SKCM. In contrast, we found fewer associations for 3SS SMRs, and no association for CDS SMRs (Fig. 6a).

Significant changes in exon-exon junctions were most commonly associated to the 5' and 3' splice site consensus sequences (5SSC and 3SSC in Fig. 6b) and to IGF2BP2, HNRNPA2B1, HNRNPLL and PCBP1 motifs among others (Fig. 6b). Among the significant changes associated to 5SS SMRs (Fig. 6c), the peptidyl-tRNA hydrolase 2 gene *PTRH2* was associated with a mutation in uterine cancer (UCEC), which would lead to the recognition of an upstream cryptic 5'ss (Fig. 6d). *PTRH2* induces anoikis; its downregulation is linked to metastasis in tumor cells (Karmali et al. 2011) and was proposed as a prognostic marker in ovarian cancer (Hua et al. 2013). The mutation observed in uterine cancer could therefore play a role in cancer progression. We also found a significant splicing change in the Farnesyl Pyrophosphate Synthetase gene *FDPS* in melanoma that would produce an alternative 5'ss and skip part of the Polyprenyl synthetase domain (Figure S16). *FDPS* induces autophagy in cancer cells (Wasko et al. 2011), and an alteration of the autophagy pathway has been related to myeloid neoplasms when *SF3B1* mutations are present (Visconte et al. 2017). It would be interesting to investigate further whether this splicing-induced alteration of *FDPS* could recapitulate a similar phenotype. Among the significant changes associated with 3SS mutations (Figure S16) there was one in the Adenine Phosphoribosyl transferase gene *APRT* associated to G>A mutations at the consensus 3'ss splice site in melanoma (SKCM), which would skip the Phosphoribosyl transferase domain (Figure S16).

We also found mutations in 5UTR motifs associated to splicing changes (Fig. 6e). Among the significant cases, we found one in *C16orf59* associated with mutations in a HNRPLL motif conserved across mammals (Fig. 6f). On the other hand, among the splicing changes associated to EXON mutations, we found one in the *C14orf37* locus associated to LUAD mutations in an IGF2BP2 in an exon of a non-coding transcript (Figure S16). Finally, INTRON SMRs had the most numerous number of junction changes with mutations in multiple RBP motifs, including RBM8A and SRSF10 (Figure S17). We found a significant change in two junctions in the histone gene *HIST1H2AC* related to mutations in the RBM8A motif in HNSC, and in *GMEB1* related to mutations in an SRSF10 motif (Figure S17). BED and GFF tracks representing all the found cases are available as supplementary material (Table S10).

Discussion

We have described a novel method to identify and characterize somatic mutations in coding and non-coding regions in relation to their potential to be involved in common protein-RNA binding. By considering mutational significance in combination with the enrichment of RBP binding motifs, we identified mutations in non-coding regions, and especially in deep intronic regions, with evidence of an impact on RNA processing. Our analysis provides evidence for new potential mechanisms by which somatic mutations impact RNA processing and contribute to tumor phenotypes. RBPs are known to control entire cellular pathways, from epithelial-to-mesenchymal transition (Shapiro et al. 2011) to cellular differentiation (Han et al. 2013). This suggests a general model in which some of the mutations occurring in cancer anywhere along genes could disrupt the function of targets of a given RBP and contribute independently to the disruption of similar pathways (Fig. 7). This provides a new strategy to interpret non-coding mutations and, moreover, indicates that lowly recurrent mutations could still be relevant to the study of cancer, as they can contribute to tumor phenotypes by impacting functions collectively controlled by the same RBP.

Our methodology presents several advantages with respect to previous methods. We can detect deep intronic mutations, whereas other methods have only tested positions on exons or in intronic regions immediately adjacent to exons (Lanzós et al. 2017; Mularoni et al. 2016). Additionally, unlike previous methods (Mularoni et al. 2016; Piraino and Furney 2017; Lanzós et al. 2017), we examined the location of the mutations in the context of regulatory motifs, which was essential to enable the interpretation of the non-coding mutations and to identify regulatory mechanisms that may be altered in cancer. Previous methods attempting to describe recurrent mutations in splice-sites have tested either just a small window around exon-intron boundaries (Jung et al. 2015) or a region too large and undefined to provide substantial mechanistic insight (Mularoni et al. 2016). In our method, the definition of SMRs and motif enrichment is driven by the positions of the mutations, hence providing the precise regions where the mutations are likely to play a role. Furthermore, we did not assume any specific functional impact, like secondary structure or conservation, as not all non-coding regions function through a structure or are necessarily conserved. Other features may determine the processing, stability and function of pre-mRNA or mature RNA molecules, and here we assumed that these are mediated through their ability to interact with proteins. Finally, unlike most previous approaches, we tested the impact on RNA processing and expression using RNA-seq from the same samples. We observed significant changes for a small fraction of SMRs, indicating that the overall impact on RNA is modest. We found multiple intronic SMRs associated to an impacted in RNA processing, which represent new genetic alterations with potential relevance to cancer. Some of these intronic SMRs occurred in non-coding RNAs. Other methods assumed that the function of lncRNAs may be impacted only through exonic mutations (Lanzós et al. 2017). We also found 5UTR SMRs associated to RNA processing changes. Although it has been assumed that RNA structure determines the function in UTR regions (Mularoni et al. 2016), our results suggest alternative mechanisms.

Our approach is subject to several limitations. The analysis may be underpowered due to the relatively small number of patients analyzed. For instance, applying our approach to another data set (Alexandrov et al. 2013a) with 507 patients, we detected no enriched RBPs and limited overlap of the SMRs with the ones described here (data not shown). This indicates that many more samples will be required to detect all possible mutations impacting RNA processing. Another limitation is that to be able to identify functionally analogous regions, we had to choose specific descriptions for the RNA binding motifs. The analysis is thus limited by how accurate these descriptions were. Short nucleotide strings hold sufficient information to describe protein-RNA interactions (Daubner et al. 2013). However, despite computational and technological advances, precise definitions of RBP binding sites at a genome scale remains challenging. Additionally, different RBPs bind very similar sequence motifs; hence the identification of some protein-RNA interactions may be ambiguous. Accordingly, we expect that some of the RBP label assignments could be improved. For instance, the described CT-rich motifs at 5'UTRs and AG-rich motifs in introns could correspond to multiple RBPs.

In summary, our data provides evidence that multiple RNA processing mechanisms may be impaired through cancer mutations. Although non-coding mutations with an impact on RNA only occur in few patients, our results motivate the extension of current methods to analyze non-coding mutations to account for functional analogous sites at different genomic positions, which will allow describing similar phenotypes arising from different alterations.

Methods

Detection of significantly mutated regions

We aimed to identify significantly mutated regions (SMRs) in both coding and non-coding regions of genes, taking into account regional and sequence mutational biases. We used the Gencode gene annotations (v19), excluding pseudogenes. To define gene loci unequivocally, we clustered transcripts that shared a splice site on the same strand, and considered the gene to be the genomic locus and the strand defined by those transcripts. We used somatic mutations from whole genome sequencing for 505 tumor samples from 14 tumor types published previously (Table S1) (Fredriksson et al. 2014): bladder carcinoma (BLCA) (21 samples), breast carcinoma (BRCA) (96 samples), colorectal carcinoma (CRC) (42 samples), glioblastoma multiforme (GBM) (27 samples), head and neck squamous carcinoma (HNSC) (27 samples), kidney chromophobe (KICH) (15 samples), kidney renal carcinoma (KIRC) (29 samples), low grade glioma (LGG) (18 samples), lung adenocarcinoma (LUAD) (46 samples), lung squamous cell carcinoma (LUSC) (45 samples), prostate adenocarcinoma (PRAD) (20 samples), skin carcinoma (SKCM) (38 samples), thyroid carcinoma (THCA) (34 samples), and uterine corpus endometrial carcinoma (UCEC) (47 samples). We only used substitutions, discarding those with a precise allelic match to a germline variant in dbSNP138. We searched for significantly mutated regions (SMRs) using a sliding-window approach, whereby along each gene locus we tested all overlapping windows of fixed length that harbored at least one mutation. As RNA binding proteins (RBPs) interact with pre-mRNAs through short nucleotide stretches, we considered windows of size 7. Using shorter windows increased the number of calculations but the results were similar, whereas with larger windows we lost positional resolution (data not shown). For each 7-mer window, we performed a double statistical test to determine the enrichment and to account for local variations and nucleotide biases in mutation rates. Given a window with *n* mutations in a gene of length *L* and *N* mutations overall, we performed a binomial test using *N/L* as the expected local mutation rate. All tested windows in a gene were adjusted for multiple testing using the Benjamini-Hochberg (BH) method. We kept only 7-mer windows that passed a false discovery rate (corrected p-value) threshold of 0.05. Although our p-values are various orders of magnitude lower than the expected values, the discarded cases showed a trend similar to the expected values (Figure S1).

To account for potential nucleotide biases we performed a second test per 7-mer window: we compared the mutation count in a given window with the expected count according to the distribution of mutations at each nucleotide in the same gene locus as follows: For each base a we calculated the rate of mutations falling in that base along a gene $R(a) = m(a)/n(a)$, where $n(a)$ is the number of a bases in the gene and $m(a)$ is the number of those bases that are mutated. The expected mutation count is then calculated using the nucleotide counts in the window and the mutation rate per nucleotide. For instance, for the 7-mer window AACTGCAG, the expected count was calculated as: $E = 3R(A) + 2R(C) + 2R(G) + R(T)$. This was compared to the actual number of mutations, n, observed in that window to define a nucleotide bias (NB) score: $NB\text{-}score = log2(n / E)$ for each 7-mer window. We discarded 7-mer windows corresponding to single-nucleotide repeats (e.g. AAAAAAA). Further, we compared the NB-scores of windows with only 1 mutation with windows with >= 3 mutations and set the NB-score to be > 6 (Figure S1). For 7-mer windows that overlap any of the three intronic or exonic bases around the exon-intron boundaries, we kept all windows with 1 or more mutations as long as the NB-score was greater than 6.

Significant 7-mer windows were classified according to the genic region in the same strand on which they fell: 5' or 3' untranslated regions (5UTR/3UTR), coding sequence (CDS), exon in short or long non-coding RNA (EXON), 5'/3' splice-site (5SS/3SS), or intron (INTRON). To unambiguously assign each 7-mer window to a region type, we prioritized the assignment as follows: 5SS/3SS > CDS > 5UTR/3UTR > EXON > INTRON. That is, if a window overlapped a splice-site, it was classified as such; else, if it overlapped a CDS, it was classified as CDS; else, if it overlapped an UTR, it was labeled as UTR; else, if it mapped an exon in a non-coding RNA, it was labeled as EXON. All SMRs that could not be matched to an exon (CDS, UTR or EXON) or to a splice site (5SS or 3SS) were classified as intronic. No significant window overlapped start or stop codons. To each SMR we assigned the average NB-score and a corrected p-value using the Simes approach (Lun and Smyth 2014): we ranked the p-values of the n overlapping 7mer windows in increasing order p_i, $i=1,2,...n$ and calculated $p_s = min\{ np_1 / 1, np_2 / 2, np_3 / 3, ..., np_n / n\}$, where p_1 was the lowest and p_n was the highest p-value in the cluster. Each SMR cluster was then assigned the p-value p_s. Code for this analysis is available at https://github.com/comprna/mira.

Comparison to expression, replication timing and LARVA

Data for replication time was obtained from (Lochovsky et al. 2015). Only SMRs for which replication time was available were analyzed. Expression data were analyzed for the samples from the PAN505 cohort (see below). For each SMR in the PAN505 cohort, we considered annotated transcripts whose genomic sequence overlapped with an SMR. We then calculated the total expression in transcripts per million (TPM) units for the overlapping transcripts per patient and averaged them across patients. For each SMR we compared the average expression of the SMR-containing transcripts in the mutated samples with the number of mutations. Additionally, we analyzed all SMRs with LARVA (Lochovsky et al. 2015) using the same mutation dataset. Specifically, we compared the significance of our SMRs with the significance given by LARVA using the model with a beta-binomial distribution and the replication timing correction (p-bbd-cor), which accounts for overdispersion of the mutation rates and regional biases.

Control regions for SMR comparison

For each set of SMRs (5SS, 3SS, CDS, 5UTR, 3UTR, INTRON, EXON), we generated control regions by random sampling non-overlapping regions from the Gencode annotation. For each SMR, we generated 100 control regions of the same length and same type, without mutations, and allowed for a maximum variation of G+C content of 5%. Each of these 100 controls were separated into different sets to generate 100 control sets of the count as the SMRs, each with similar distribution and G+C content distributions. For the 5SS and 3SS SMRs we generated controls by sampling regions with the same length and same relative position from the exon-intron boundary, with no mutations and also controlling for GC content.

Motif analysis

We performed an unbiased search for enriched k-mers (k=6) on the SMRs using MoSEA (https://github.com/comprna/MoSEA) (Sebestyén et al. 2016). For each 6-mer within an SMR, we counted the number of SMRs and control regions in which it appeared. A z-score was computed for each individual 6-mer comparing the observed frequency with the distribution of frequencies in the 100 control regions. The enrichment analysis was repeated, reversing the strand of all SMRs and control regions, and those 6-mers that appeared significantly enriched in the direct and reversed analyses for the same region type were discarded. We considered significantly enriched 6-mers with a z-score > 1.96 and with 5 or more counts on SMRs. This analysis included the GT and AG containing 6-mers at 5SS and 3SS splice sites, respectively. In total we obtained 444 enriched 6-mers (Table S3) in 3456 SMRs (Table S4). From all 20307 SMRs, 749 (3,68%) appeared in both strands due to overlapping genes. However, considering the 3456 SMRs with enriched 6-mers, only 25 (0,72%) overlapped with opposite strands, which is a significant reduction (Fisher's exact test p-value = 2.2e-16, odds-ratio = 5.25). To label enriched 6-mers we used DeepBind (Alipanahi et al. 2015) to score each 6-mer using models for 522 proteins containing KH (24 proteins), RRM (134, 2 in common with KH) and C2H2 (366 proteins) domains from human (413 proteins), mouse (49 proteins) and Drosophila (60 proteins). For each 6-mer, we kept the top three predictions with a score > 0.1. Subsequently, given all 6-mers associated to the same RBP label, we kept only those 6-mers that were at a maximum Levenshtein distance of 2 from the top-scoring 6-mer.

Significant mutations per position of a motif

For each of the RBP labels we located all the associated 6-mers in SMRs and performed a multiple sequence alignment (MSA) with all the 6-mers instances using ClustalW (Thompson et al. 2002). Sequence logos were built from this alignment and somatic mutations were counted per position relative to the MSA. As a control, we shuffled the same number of mutations along the aligned positions to calculate an average per position. Germline mutations from the 1000 genomes project (1000 Genomes Project Consortium et al. 2010) were also considered to calculate the number of germline mutations per position of the MSA.

Gene set enrichment analysis

Annotations for gene sets and pathways were obtained from the Molecular Signatures Database v4.0 (Liberzon et al. 2015). We performed a Fisher's exact test per hallmark set for genes harboring SMRs with labeled enriched motifs in the following way: for each RBP motif and each gene set, we built a contingency matrix with the counts of genes with and without an RBP motif in an SMR, and within or outside each gene set.

RNA-seq data analysis

TCGA RNA-seq data was obtained for the PAN505 samples from the Genomic Data Commons (Grossman et al. 2016) (https://gdc-portal.nci.nih.gov/). We estimated transcript abundances for the Gencode annotations (v19) in TPM units using Salmon (Version 0.8.1) (Patro et al. 2017). For each mutated position in an enriched motif, we calculated whether the mutation was associated to a change in transcript expression using an outlier statistic. For each SMR with an enriched motif and for each transcript whose genomic extension contained the SMR, we compared the transcript $\log_2$(TPM+0.01) for each patient with a mutation on the motif, with the distribution of $\log_2$(TPM+0.01) values for the same transcript in the patients from the same tumor type with no mutations in the motif. We only considered those cases where at least 5 patients lacked a mutation. We kept only those cases with |z-score|>1.96 and with a difference between the observed $\log_2$(TPM+0.01) and the mean of $\log_2$(TPM+0.1) in patients without mutations greater than 0.5 in absolute value. We considered as significant those cases with a p-value < 0.05 after adjusting for multiple testing using the BH approach. RNA-seq reads were also mapped to the human genome (hg19) with STAR (version 2.5.0) (Dobin et al. 2013) and analyzed using Junckey (https://github.com/comprna/Junckey). From the BAM files, we identified all possible exon-exon junctions defined by spliced reads that appeared in any of the samples. All defined junctions were then grouped into junction-clusters. Any two junctions were placed in the same cluster if they shared at least one splice-site. Clusters were built using all junctions present in any patient, but junction read-counts were assigned per patient, i.e. 0 or more. Only clusters that had at least a total of 30 reads in all samples were used. Additionally, we only used junctions <100kbp in length and with a total of >1% of reads from the cluster in all samples. Then, for each patient, the read-count per junction was normalized by the total read count in that cluster to define the junction inclusion level or proportion spliced-in (PSI). A junction not expressed in a sample was given a zero inclusion level. The software for junction analysis is available at https://github.com/comprna/Junckey. For each SMR containing an enriched motif, we compared each patient with a mutation in the motif against all patients for the same tumor type who did not have mutations in the same SMR. We measured a z-score derived from the PSI for each junction overlapping the motif, by comparing with the of PSIs for the same junction in the non-mutated patients, and we kept only those cases with |z-score|>1.96 and |ΔPSI|>0.1. We considered significant those changes with p-value < 0.05 after adjusting for multiple testing using the BH approach.

Supplementary Data and Software

Supplementary Data for this manuscript is available at:
http://comprna.upf.edu/Data/MutationsRBPMotifs/
Code used in this manuscript is available at:
https://github.com/comprna/mira
https://github.com/comprna/MoSEA
https://github.com/comprna/Junckey

Acknowledgements

B.S., J.L.T. and E.E. were supported by the MINECO and FEDER (BIO2014-52566-R), Consolider RNAREG (CSD2009-00080), AGAUR (SGR2014-1121), the European ITN Network RNP-Net (ID:289007) and the Sandra Ibarra Foundation for Cancer (FSI2013). The Computational RNA Biology Group is part of the Research Programme on Biomedical Informatics (GRIB), which is member of ELIXIR-Excelerate of the European Union Horizon 2020 Programme 2014-2020 [No. 676559] and of the Spanish National Bioinformatics Institute (INB), PRB2-ISCIII and is supported by grant PT13/0001/0023, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER. This work was supported by the MINECO and FEDER (BIO2014–52566-R) and AGAUR (SGR2014–1121). BS was funded by an FPI grant from the Spanish Government with reference BES-2012-052683. SRP was supported by startup funds from Brigham Young University. PJT was supported by a fellowship from the Simmons Center for Cancer Research at Brigham Young University. All authors thank The Cancer Genome Atlas project for making their data publicly available.

Authors' contributions

EE proposed and led the study. BS implemented the methods and performed the analyses. JLT developed the method for junction analysis. PJT and SRP performed the mapping of RNA-seq reads and transcript quantification. EE and BS wrote the manuscript with essential inputs from JLT, PJT and SRP.

List of Abbreviations

SMR: significantly mutated region; NB-score: Nucleotide-bias score; RBP: RNA binding protein; RNA-seq: RNA sequencing;

Supplementary files

Table S1: List of patients analyzed

Table S2: Significantly mutated regions (SMRs)

Table S3: All mutations falling on SMRs.

Table S4: All enriched 6-mers found in SMRs with their predicted RBP label (if any)

Table S5: SMRs with enriched k-mers

Table S6: Mutations falling on labelled enriched k-mers

Table S7: Gene signatures and Pathway enrichment analysis

Table S8: SMRs with mutations associated to significant transcript expression changes

Table S9: SMRs with mutations associated to significant differential inclusion of junctions

Table S10: UCSC track for SMRs, enriched motifs, mutations and junctions.

References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–73. http://www.ncbi.nlm.nih.gov/pubmed/20981092.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, et al. 2013a. Signatures of mutational processes in human cancer. *Nature* **500**: 415–21. http://www.ncbi.nlm.nih.gov/pubmed/23945592.

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013b. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**: 246–59.

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and

RNA-binding proteins by deep learning. *Nat Biotechnol* **33**: 831–8. http://www.ncbi.nlm.nih.gov/pubmed/26213851.

Anczuków O, Akerman M, Cléry A, Wu J, Shen C, Shirole NH, Raimer A, Sun S, Jensen MA, Hua Y, et al. 2015. SRSF1-Regulated Alternative Splicing in Breast Cancer. *Mol Cell* **60**: 105–17. http://www.ncbi.nlm.nih.gov/pubmed/26431027.

Cai C, Hsieh C-L, Omwancha J, Zheng Z, Chen S-Y, Baert J-L, Shemshedini L. 2007. ETV1 is a novel androgen receptor-regulated gene that mediates prostate cancer cell invasion. *Mol Endocrinol* **21**: 1835–46. http://www.ncbi.nlm.nih.gov/pubmed/17505060.

Chabot B, Shkreta L. 2016. Defective control of pre-messenger RNA splicing in human disease. *J Cell Biol* **212**: 13–27. http://www.ncbi.nlm.nih.gov/pubmed/26728853.

Daubner GM, Cléry A, Allain FH-T. 2013. RRM-RNA recognition: NMR or crystallography…and new findings. *Curr Opin Struct Biol* **23**: 100–8. http://www.ncbi.nlm.nih.gov/pubmed/23253355.

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

Fredriksson NJ, Ny L, Nilsson JA, Larsson E. 2014. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet* **46**: 1–7. http://dx.doi.org/10.1038/ng.3141.

Fu X-D, Ares M. 2014. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet* **15**: 689–701. http://www.ncbi.nlm.nih.gov/pubmed/25112293.

Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. 2016. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* **375**: 1109–12. http://www.ncbi.nlm.nih.gov/pubmed/27653561.

Haerty W, Ponting CP. 2015. Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* **21**: 333–46. http://www.ncbi.nlm.nih.gov/pubmed/25589248.

Han H, Irimia M, Ross PJ, Sung H-K, Alipanahi B, David L, Golipour A, Gabut M, Michael IP, Nachman EN, et al. 2013. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498**: 241–5. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3933998&tool=pmcentrez&rendertype=abstract.

Hanahan D, Weinberg RA. 2011. Hallmarks of Cancer: The Next Generation. *Cell* **144**: 646–674.

Horn S, Figl A, Rachakonda PS, Fischer C, Sucker A, Gast A, Kadel S, Moll I, Nagore E, Hemminki K, et al. 2013. TERT Promoter Mutations in Familial and Sporadic Melanoma. *Science (80-)* **339**: 959–961.

Hua W, Miao S, Zou W, Yang H, Chen BL. 2013. Pathological implication and function of Bcl2-inhibitor of transcription in ovarian serous papillary adenocarcinomas. *Neoplasma* **60**: 143–50. http://www.ncbi.nlm.nih.gov/pubmed/23259782.

Huang FW, Hodis E, Xu MJ, Kryukov G V., Chin L, Garraway LA. 2013. Highly Recurrent TERT Promoter Mutations in Human Melanoma. *Science (80-)* **339**: 957–959. http://www.ncbi.nlm.nih.gov/pubmed/23348506 (Accessed January 28, 2017).

Hwang CK, Wagley Y, Law P-Y, Wei L-N, Loh HH. 2017. Phosphorylation of poly(rC) binding protein 1 (PCBP1) contributes to stabilization of mu opioid receptor (MOR) mRNA via interaction with AU-rich element RNA-binding protein 1 (AUF1) and poly A binding protein (PABP). *Gene* **598**: 113–130. http://www.ncbi.nlm.nih.gov/pubmed/27836661.

Julien P, Miñana B, Baeza-Centurion P, Valcárcel J, Lehner B. 2016. The complete local genotype-phenotype landscape for the alternative splicing of a human exon. *Nat Commun* **7**: 11558. http://www.ncbi.nlm.nih.gov/pubmed/27161764.

Jung H, Lee D, Lee J, Park D, Kim YJ, Park W-Y, Hong D, Park PJ, Lee E. 2015. Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat Genet* **47**: 1242–1248. http://dx.doi.org/10.1038/ng.3414.

Juul M, Bertl J, Guo Q, Nielsen MM, Świtnicki M, Hornshøj H, Madsen T, Hobolth A, Pedersen JS. 2017. Non-coding cancer driver candidates identified with a sample- and position-specific model of the somatic mutation rate. *Elife* **6**. http://www.ncbi.nlm.nih.gov/pubmed/28362259.

Karmali PP, Brunquell C, Tram H, Ireland SK, Ruoslahti E, Biliran H. 2011. Metastasis of tumor cells is enhanced by downregulation of Bit1. *PLoS One* **6**: e23840. http://www.ncbi.nlm.nih.gov/pubmed/21886829.

Ke S, Shang S, Kalachikov SM, Morozova I, Yu L, Russo JJ, Ju J, Chasin LA. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res* **21**: 1360–74. http://www.ncbi.nlm.nih.gov/pubmed/21659425.

Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M. 2016. Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**: 93–108.

Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB. 2014. RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell* **54**: 887–900. http://www.ncbi.nlm.nih.gov/pubmed/24837674.

Lanzós A, Carlevaro-Fita J, Mularoni L, Reverter F, Palumbo E, Guigó R, Johnson R. 2017. Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. *Sci Rep* **7**: 41544. http://www.ncbi.nlm.nih.gov/pubmed/28128360.

Lawrence MS, Stojanov P, Polak P, Kryukov G V, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**: 214–8. http://www.ncbi.nlm.nih.gov/pubmed/23770567.

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst* **1**: 417–425.

Liu L, De S, Michor F. 2013. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat Commun* **4**: 1502.

http://www.ncbi.nlm.nih.gov/pubmed/23422670.

Lochovsky L, Zhang J, Fu Y, Khurana E, Gerstein M. 2015. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res* **43**: 8123–34. http://www.ncbi.nlm.nih.gov/pubmed/26304545.

Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, Liang TY, Stark TJ, Gehman LT, Hoon S, et al. 2013. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* **20**: 1434–42. http://www.ncbi.nlm.nih.gov/pubmed/24213538.

Lun ATL, Smyth GK. 2014. De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. *Nucleic Acids Res* **42**: e95. http://www.ncbi.nlm.nih.gov/pubmed/24852250.

Magnani L, Ballantyne EB, Zhang X, Lupien M. 2011. PBX1 genomic pioneer function drives ERα signaling underlying progression in breast cancer. *PLoS Genet* **7**: e1002368. http://www.ncbi.nlm.nih.gov/pubmed/22125492.

Maslon MM, Heras SR, Bellora N, Eyras E, Cáceres JF. 2014. The translational landscape of the splicing factor SRSF1 and its role in mitosis. *Elife* **2014**.

Melton C, Reuter JA, Spacek D V, Snyder M. 2015. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat Genet* **47**: 710–6. http://www.ncbi.nlm.nih.gov/pubmed/26053494.

Michlewski G, Guil S, Semple CA, Cáceres JF. 2008. Posttranscriptional regulation of miRNAs harboring conserved terminal loops. *Mol Cell* **32**: 383–93. http://www.ncbi.nlm.nih.gov/pubmed/18995836.

Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. 2016. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol* **17**: 128. http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0994-0.

Oberdoerffer S, Moita LF, Neems D, Freitas RP, Hacohen N, Rao A. 2008. Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL. *Science* **321**: 686–91. http://www.ncbi.nlm.nih.gov/pubmed/18669861.

Oberstrass FC, Auweter SD, Erat M, Hargous Y, Henning A, Wenter P, Reymond L, Amir-Ahmady B, Pitsch S, Black DL, et al. 2005. Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science* **309**: 2054–7. http://www.ncbi.nlm.nih.gov/pubmed/16179478.

Ortiz-Zapater E, Pineda D, Martínez-Bosch N, Fernández-Miranda G, Iglesias M, Alameda F, Moreno M, Eliscovich C, Eyras E, Real FX, et al. 2012. Key contribution of CPEB4mediated translational control to cancer progression. *Nat Med* **18**.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. http://www.ncbi.nlm.nih.gov/pubmed/28263959.

Pichon X, Wilson LA, Stoneley M, Bastide A, King HA, Somers J, Willis AEE. 2012. RNA binding protein/RNA element interactions and the control of translation. *Curr Protein Pept Sci* **13**: 294–304. http://www.ncbi.nlm.nih.gov/pubmed/22708490.

Piraino SW, Furney SJ. 2016. Beyond the exome: the role of non-coding somatic mutations in cancer.

Ann Oncol Off J Eur Soc Med Oncol **27**: 240–8. http://www.ncbi.nlm.nih.gov/pubmed/26598542.

Piraino SW, Furney SJ. 2017. Identification of coding and non-coding mutational hotspots in cancer genomes. *BMC Genomics* **18**: 17. http://www.ncbi.nlm.nih.gov/pubmed/28056774.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**: 172–7. http://www.ncbi.nlm.nih.gov/pubmed/23846655.

Rissland OS. 2017. The organization and regulation of mRNA-protein complexes. *Wiley Interdiscip Rev RNA* **8**. http://www.ncbi.nlm.nih.gov/pubmed/27324829.

Sawicka K, Bushell M, Spriggs KA, Willis AE. 2008. Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem Soc Trans* **36**: 641–7. http://www.ncbi.nlm.nih.gov/pubmed/18631133.

Sebestyén E, Singh B, Miñana B, Pagès A, Mateo F, Pujana MA, Valcárcel J, Eyras E. 2016. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res* **26**: 732–44. http://www.ncbi.nlm.nih.gov/pubmed/27197215.

Shapiro IM, Cheng AW, Flytzanis NC, Balsamo M, Condeelis JS, Oktay MH, Burge CB, Gertler FB. 2011. An emt-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genet* **7**.

Soemedi R, Cygan KJ, Rhine CL, Wang J, Bulacan C, Yang J, Bayrak-Toydemir P, McDonald J, Fairbrother WG. 2017. Pathogenic variants that alter protein code often disrupt splicing. *Nat Genet*. http://www.ncbi.nlm.nih.gov/pubmed/28416821.

Sterne-Weiler T, Sanford JR. 2014. Exon identity crisis: disease-causing mutations that disrupt the splicing code. *Genome Biol* **15**: 201. http://www.ncbi.nlm.nih.gov/pubmed/24456648.

Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. 2014. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**: 1324–35. http://www.ncbi.nlm.nih.gov/pubmed/24630730.

Thompson JD, Gibson TJ, Higgins DG. 2002. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinforma* **Chapter 2**: Unit 2.3. http://www.ncbi.nlm.nih.gov/pubmed/18792934.

Turner-Ivey B, Guest ST, Irish JC, Kappler CS, Garrett-Mayer E, Wilson RC, Ethier SP. 2014. KAT6A, a chromatin modifier from the 8p11-p12 amplicon is a candidate oncogene in luminal breast cancer. *Neoplasia* **16**: 644–55. http://www.ncbi.nlm.nih.gov/pubmed/25220592.

Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. 2003. CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**: 1212–5. http://www.ncbi.nlm.nih.gov/pubmed/14615540.

Visconte V, Przychodzen B, Han Y, Nawrocki ST, Thota S, Kelly KR, Patel BJ, Hirsch C, Advani AS, Carraway HE, et al. 2017. Complete mutational spectrum of the autophagy interactome: a novel class of tumor suppressor genes in myeloid neoplasms. *Leukemia* **31**: 505–510. http://www.ncbi.nlm.nih.gov/pubmed/27773925.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz Jr. LA, Kinzler KW. 2013. Cancer Genome

Landscapes. *Science (80-)* **339**: 1546–1558.

Wasko BM, Dudakovic A, Hohl RJ. 2011. Bisphosphonates induce autophagy by depleting geranylgeranyl diphosphate. *J Pharmacol Exp Ther* **337**: 540–6. http://www.ncbi.nlm.nih.gov/pubmed/21335425.

Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. 2014. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* **46**: 1160–5. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4217527&tool=pmcentrez&rendertype=abstract.

Whitworth H, Bhadel S, Ivey M, Conaway M, Spencer A, Hernan R, Holemon H, Gioeli D. 2012. Identification of kinases regulating prostate cancer cell growth using an RNAi phenotypic screen. *PLoS One* **7**: e38950. http://www.ncbi.nlm.nih.gov/pubmed/22761715.

Ying J, Li H, Cui Y, Wong AHY, Langford C, Tao Q. 2006. Epigenetic disruption of two proapoptotic genes MAPK10/JNK3 and PTPN13/FAP-1 in multiple lymphomas and carcinomas through hypermethylation of a common bidirectional promoter. *Leukemia* **20**: 1173–5. http://www.ncbi.nlm.nih.gov/pubmed/16572203.

Figure Legends

Figure 1. Systematic identification of RNA-related significantly mutated regions (SMRs). **(a)** Short k-mer windows (k=7 in our study) along genes are tested for the enrichment in mutations with respect to the gene mutation rate and the local nucleotide biases. Significant windows are clustered by region type into significantly mutated regions (SMRs). For each SMR we give the NB-score (x axis) and the number of mutations (y axis). **(b)** We show the SMRs detected in CDS regions, introns (INTRON), 5' UTRs (5UTR) and 3'UTRs (3UTR). All SMRs detected are shown, but we only show the gene name for the SMRs with nucleotide-bias (NB) score > 6 and with 5 or more mutations, except for the INTRON SMRs, where we highlight the cases with 12 or more mutations. **(c)** Examples of a CDS SMR in *NRAS* and an INTRON SMR in *MET*. The UCSC screenshots show the SMR (green) and the mutations detected (red).

Figure 2. Enriched splice-site motifs in significantly mutated regions (SMRs). (a) Upper panel: mutation pattern in SMRs. For each nucleotide substitution we give the total count of substitutions observed in SMRs separated according to the nucleotide triplet in which it occurs. SMRs are stranded; hence we give the substitutions according to the SMR strand. Lower panel: mutation patterns in enriched 6-mers in SMRs. For each nucleotide substitution we give the total count observed separated according to the nucleotide triplet in which it occurs. Since the 6-mers

are stranded, give the substitutions according to the 6-mer strand. **(b)** We show the logos for the splice site motifs significantly enriched in 5SS and 3SS SMRs, i.e. they appear more frequently in 5SS or 3SS SMRs than in the corresponding controls. The barplots below show the proportion of all somatic mutations in the 6-mers (y axis) that fall on each position along the motif logo. In orange indicate those somatic mutations that coincide with a germline SNP. **(c)** The plots show for each position the number of splice-sites with each type of substitution indicated with a color code below. **(d)** We give two examples of mutations found at splice sites motifs in the genes *NF1* and *FGFR3*. Above the gene track we show the significantly mutated region (SMR) (green track), the enriched motif found in the SMR (blue track), and the somatic mutations (read track). For each mutation we indicate the patient identifier, the tumor type and the substitution.

Figure 3. Cancer mutations in enriched RBP motifs. We provide the proportion of samples separated by tumor type (y axis) that have a mutated motif in 5SS **(a)**, 3SS **(b)**, CDS **(c)** and INTRON **(d)** SMRs. In each SMR type we show the enriched motifs. For 5SS and 3SS we indicate the consensus 5' or 3' splice site sequences (5SSC/3SSC). The proportions are color coded by tumor type. We show the mutation patterns on SRSF10 motifs in 3SS **(e)**, CDS **(f)** and INTRON **(g)** SMRs. In the upper panel we indicate in dark red the position of the mutations and in light red the positions covered by motif. The barplots below show the proportion of somatic mutations (y axis) that fall on each position along the motif logo. In orange indicate those somatic mutations that coincide with a germline SNP. Below we show for each position the number of motifs with each type of substitution indicated with a color code below.

Figure 4. Cancer mutations in enriched RBP motifs in 5UTR and 3UTR SMRs. (a) For 5UTR **(a)** and 3UTR **(b)** SMRs we provide the proportion of samples in each tumor type (y axis) that have a mutated RNA binding protein (RBP) motif (x axis). The proportions are color-coded by tumor type. The proportion of SMRs with each RBP motif per tumor type is given in Figure S9. We show the positional patterns of mutations on **(c)** PCBP3 motifs in 5UTR SMRs, and on **(d)** HNRNPC3 motifs in 3UTR SMRs. In the upper panels we indicate in red the positions covered by the motif and in dark red the position of the mutations. The barplots below show the proportion of somatic mutations (y axis) that fall on each position along the motif logo. In orange we indicate those somatic mutations that coincide with a germline SNP in position (with a different substitution pattern, as the exact matching substitutions were removed). Below we show for each position, the number of motifs with each type of substitution indicated with a color

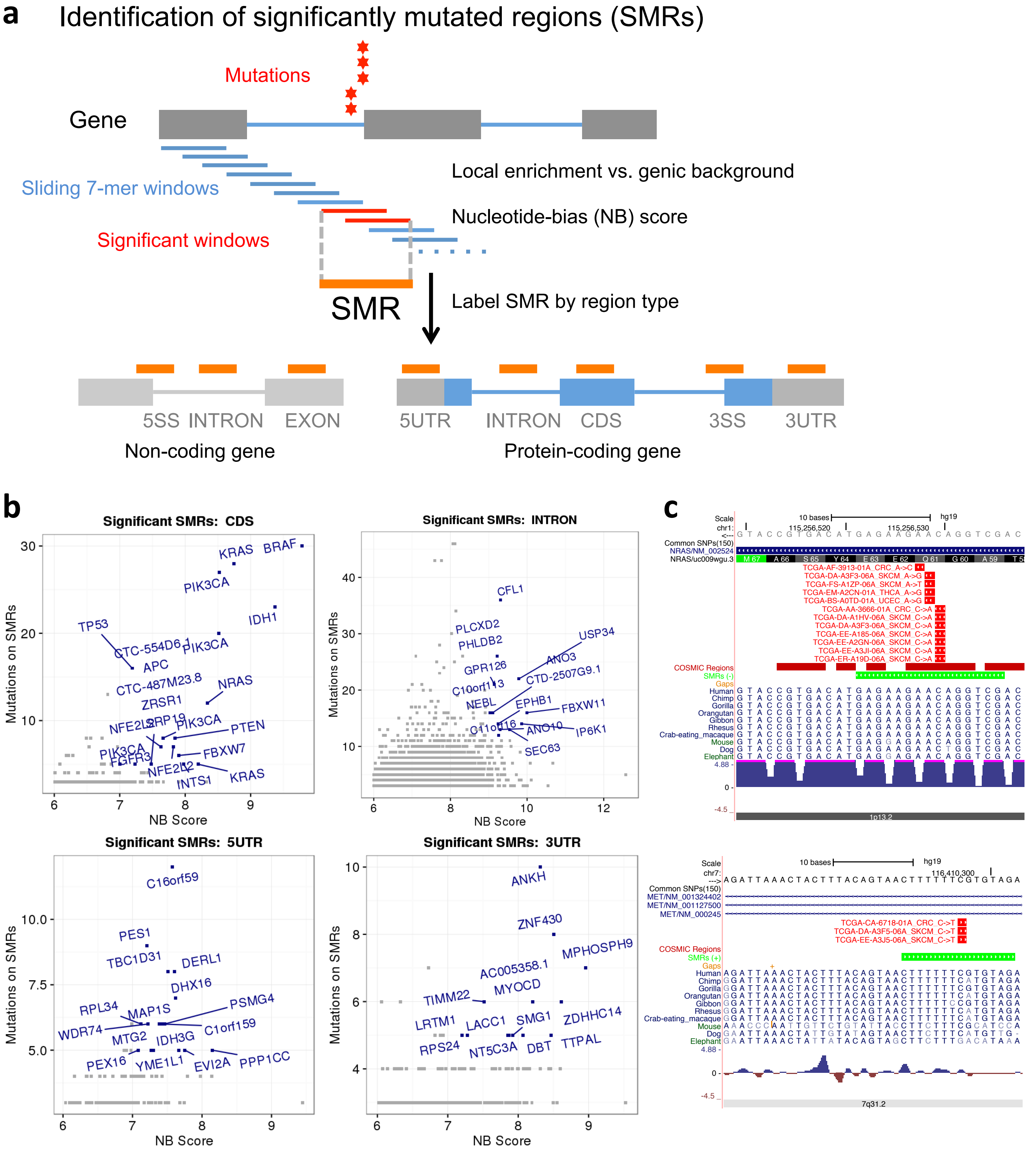
code below.

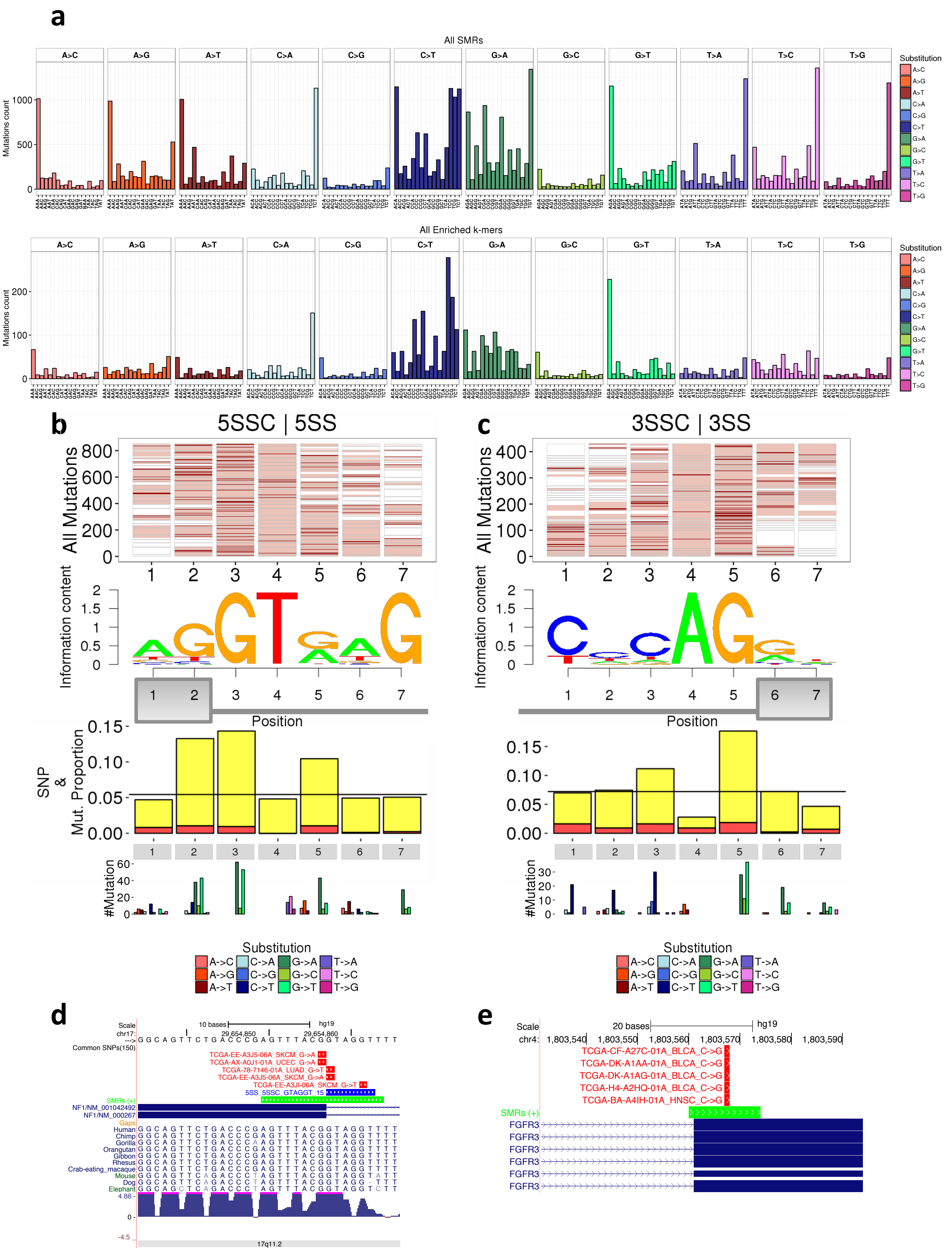
Figure 5. Transcript expression changes associated to mutations in RBP motifs. (a) For each region type (x axis) we give the number of SMRs (y axis) for which we found a significant change in transcript isoform expression associated with somatic mutations in enriched motifs within the SMR. The counts are color-coded by tumor type. **(b)** For each motif (x axis), we give the number of cases for which we found a significant change in transcript expression (y axis) associated with somatic mutations in the motif. The counts are color-coded by tumor type. **(c)** Significant changes in transcript expression associated with mutations in intronic SRSF10 motifs. For each patient (x axis) we show the transcripts (y axis) that have a significant increase (red) or decrease (blue) in expression. We separate them according to tumor type (indicated above). In the UCSC screen shot we indicate the patients and mutations on the CAGAGA motif in the intron of *MAPK10*. Below we show the significant expression changes detected in two different transcripts of *MAPK10* associated to mutations in the SRSF10 motif. **(d)** Significant changes in transcript expression associated with mutations in intronic MBNL1 motifs. For each patient (x axis) we show the transcripts (y axis) that have a significant increase (red) or decrease (blue) in expression. We separate them according to tumor type (indicated above). In the UCSC screen shot we indicate the patients and mutations on the CGCTTT motif in the intron of *ETV1*. Below we show the significant expression changes detected in two different transcripts of *ETV1* associated to mutations in the MBNL1 motif.

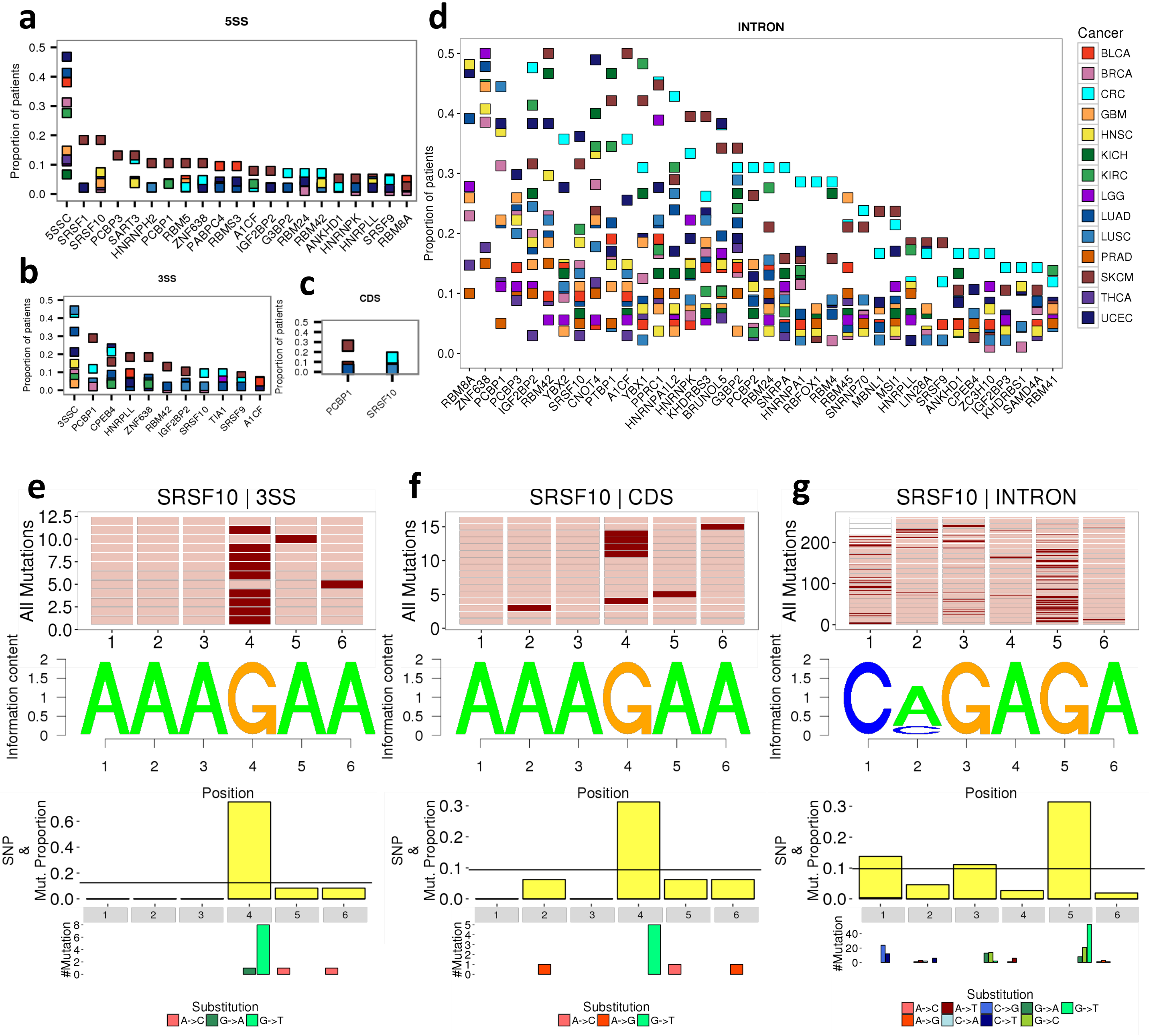
Figure 6. Changes in junction usage associated to mutations in RBP motifs. (a) For each region type (x axis) we give the number of SMRs (y axis) for which we found a significant change in exon-exon junction inclusion associated with somatic mutations in enriched motifs within the SMR. The counts are color-coded by tumor type. **(b)** For each motif (x axis), we give the number of instances (y axis) for which we found a significant change in exon-exon junction inclusion associated with somatic mutations in the motif (x axis). The counts are color-coded by tumor type. **(c)** Significant changes in junction inclusion associated to mutations in 5' splice site (5'ss) motifs. For each patient (x axis) we show the junctions (y axis) that have a significant increase (orange) or decrease (cyan) in inclusion (PSI). We separate them according to tumor type (indicated above). **(d)** Significant junction changes in *PTRH2* in uterine cancer (UCEC). We show the changing junctions in gray. The mutation in the annotated 5'ss induces the usage of an upstream cryptic 5'ss. We show the SMR (green), the enriched motif (blue), and the mutation

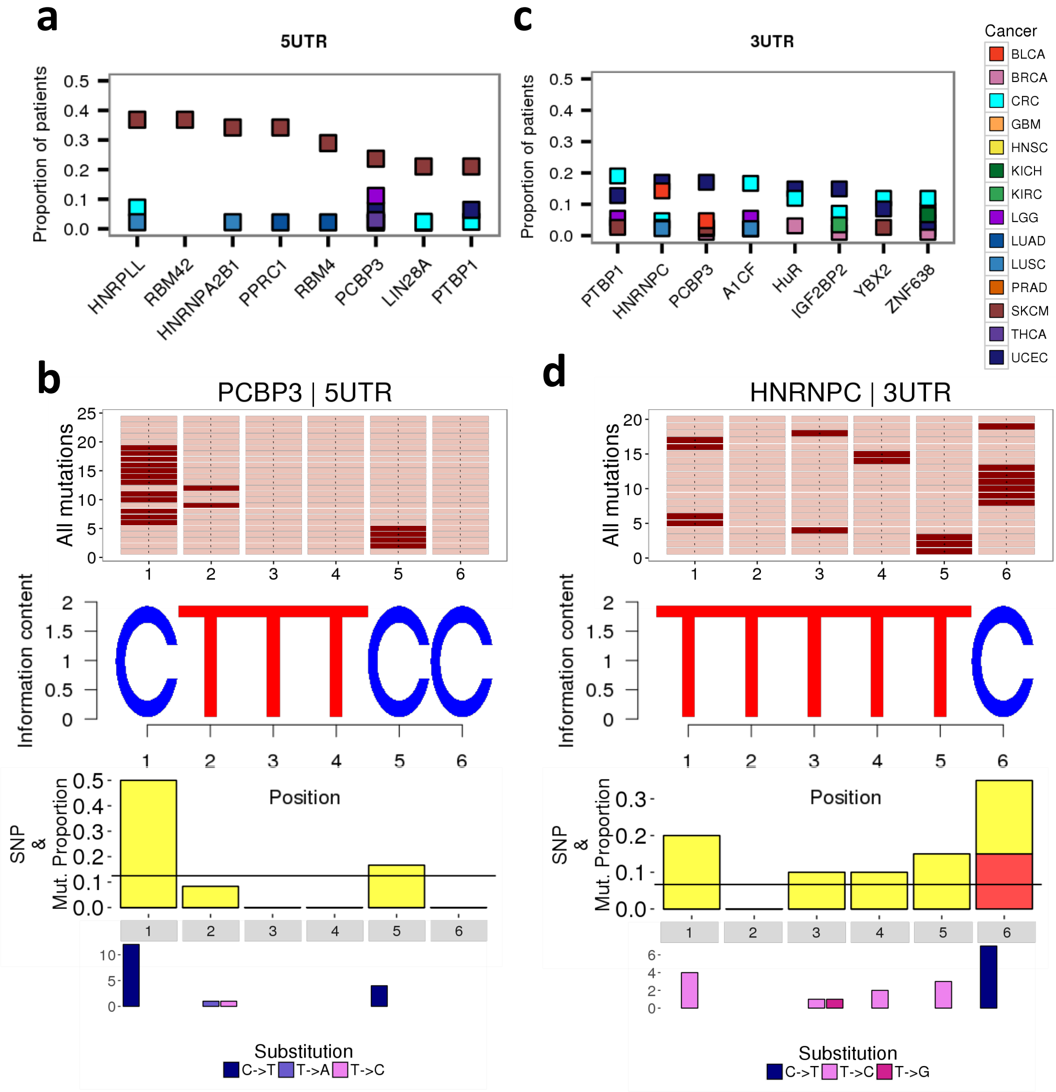
(red). The boxplots blow show the PSI values (y axis) of the two changing junctions separated by samples with mutations and without mutations in this SMR in UCEC, indicated in the x axis. **(e)** Significant changes in junction inclusion associated to mutations in 5UTR SMRs. For each patient (x axis) we show the junctions (y axis) that have a significant increase (orange) or decrease (cyan) in inclusion (PSI). We separate them according to tumor type (indicated above). **(f)** Significant junction changes in *C16orf59* associated to mutations in an HRNPLL motif in melanoma (SKCM). The boxplots show the PSI values (y axis) of the two changing junctions separated by samples with mutations and without mutations, indicated in the x axis. In the screenshot we show the 5UTR SMR (green), the enriched motifs (blue), and the mutations (red), which suggest dinucleotide mutations GG>AA in some patients. The junctions associated to these mutations are downstream of the SMR and do not appear in the genomic range shown in the figure.

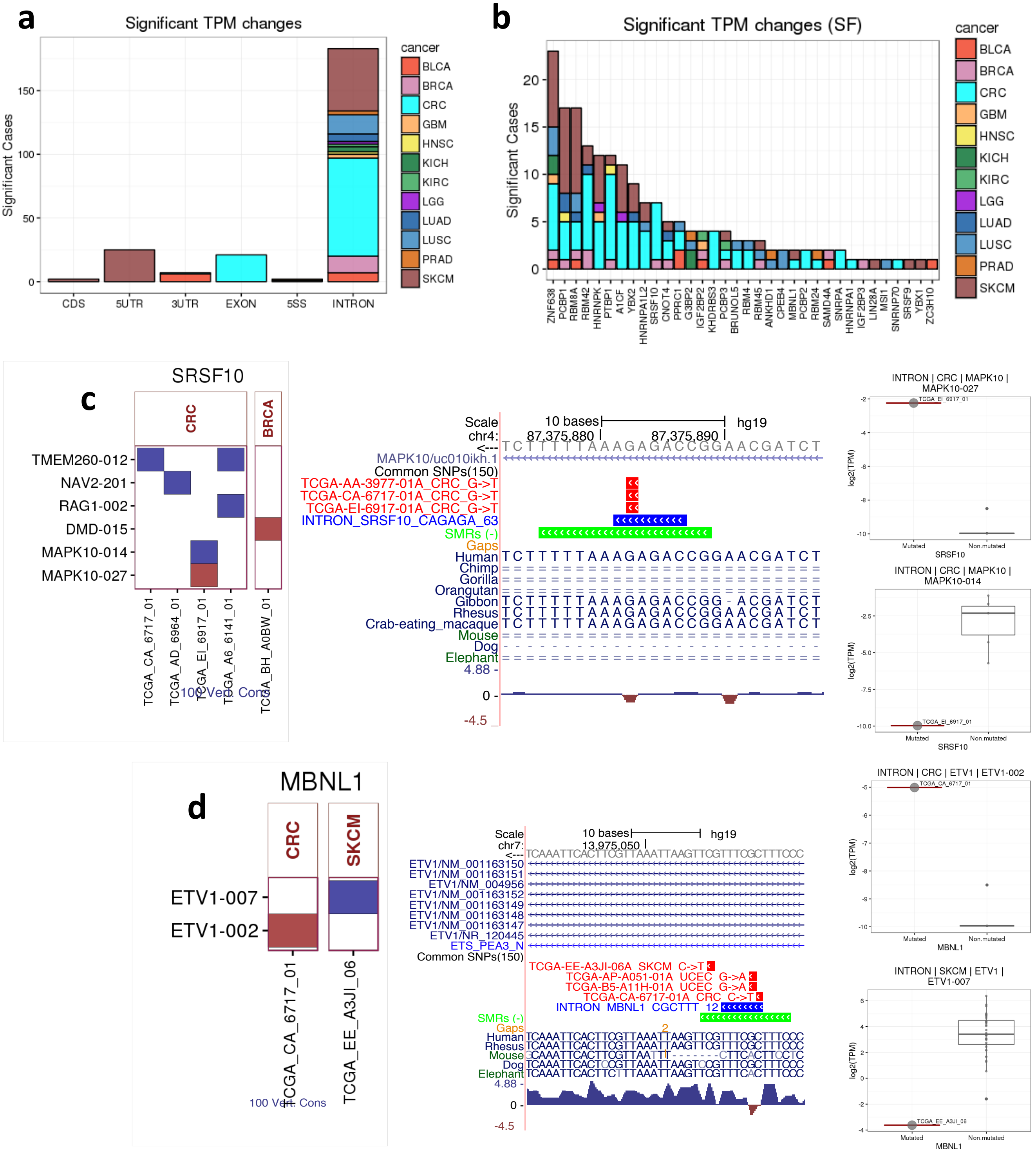
Figure 7. Non-coding mutations in human tumors impact binding sites of RNA binding proteins. Our analysis suggests that many of the mutations (indicated in red) on non-coding regions, predominantly introns, and UTRs, impact binding sites of RNA binding proteins (indicated in orange and green) and affect RNA processing in multiple different genes across patients. In the figure altered RNA processing are indicated as solid gene models. These alterations would contribute to the frequent changes observed in RNA processing in tumors and could indicate novel oncogenic mechanisms.

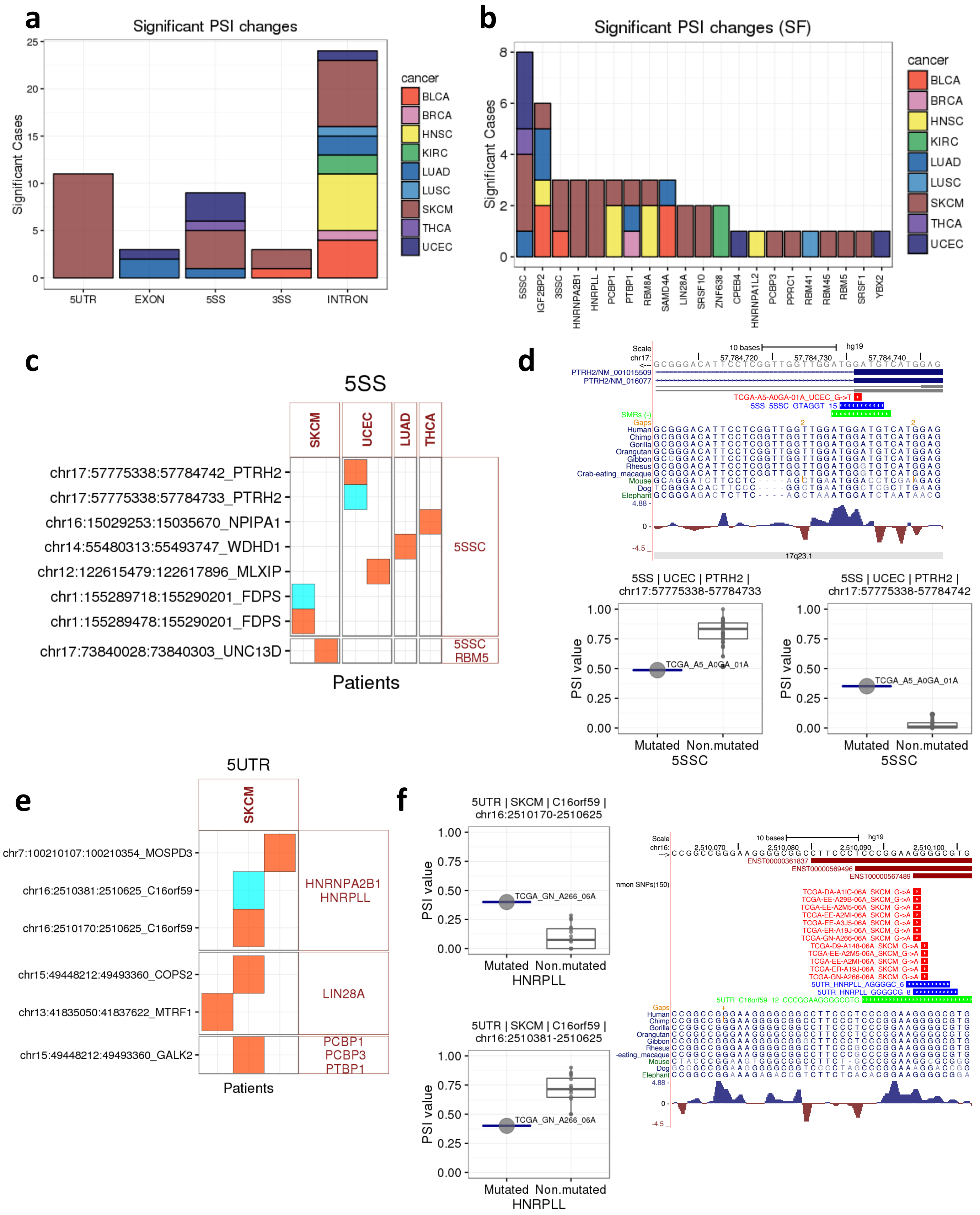

a
Identification of significantly mutated regions (SMRs)
Mutations
Gene
Sliding 7-mer windows
Local enrichment vs. genic background
Nucleotide-bias (NB) score
Significant windows
SMR
Label SMR by region type
5SS
INTRON
EXON
5UTR
INTRON
CDS
3SS
3UTR
Non-coding gene
Protein-coding gene
b
Significant SMRs: CDS
Mutations on SMRs
NB Score
BRAF
KRAS
PIK3CA
IDH1
TP53
CTC-554D6.1
PIK3CA
APC
CTC-487M23.8
NRAS
ZRSR1
NFE2L2
SRP19
PIK3CA
PTEN
PIK3CA
FGFR3
FBXW7
NFE2L2
INTS1
KRAS
Significant SMRs: INTRON
CFL1
PLCXD2
PHLDB2
USP34
GPR126
ANO3
CTD-2507G9.1
C10orf113
NEBL
EPHB1
FBXW11
C1orf116
ANO10
IP6K1
SEC63
Significant SMRs: 5UTR
C16orf59
PES1
TBC1D31
DERL1
DHX16
RPL34
MAP1S
PSMG4
WDR74
C1orf159
MTG2
IDH3G
PEX16
YME1L1
EVI2A
PPP1CC
Significant SMRs: 3UTR
ANKH
ZNF430
MPHOSPH9
AC005358.1
MYOCD
TIMM22
LRTM1
LACC1
SMG1
ZDHHC14
RPS24
NT5C3A
DBT
TTPAL
c
Scale
10 bases
hg19
chr1:
115,256,520
115,256,530
Common SNPs(150)
NRAS/NM_002524
NRAS/uc009wgu.3
M 67
A 66
S 65
Y 64
E 63
E 62
Q 61
G 60
A 59
TCGA-AF-3913-01A_CRC_A->C
TCGA-DA-A3F3-06A_SKCM_A->G
TCGA-FS-A1ZP-06A_SKCM_A->T
TCGA-EM-A2CN-01A_THCA_A->G
TCGA-BS-A0TD-01A_UCEC_A->G
TCGA-AA-3666-01A_CRC_C->A
TCGA-DA-A1HV-06A_SKCM_C->A
TCGA-DA-A3F3-06A_SKCM_C->A
TCGA-EE-A185-06A_SKCM_C->A
TCGA-EE-A2GN-06A_SKCM_C->A
TCGA-EE-A3JI-06A_SKCM_C->A
TCGA-ER-A19D-06A_SKCM_C->A
COSMIC Regions
SMRs (-)
Gaps
Human
Chimp
Gorilla
Orangutan
Gibbon
Rhesus
Crab-eating_macaque
Mouse
Dog
Elephant
4.88
0
-4.5
1p13.2
Scale
10 bases
hg19
chr7:
116,410,300
AGATTAAACTACTTTACAGTAACTTTTTTCGTGTAGA
Common SNPs(150)
MET/NM_001324402
MET/NM_001127500
MET/NM_000245
TCGA-CA-6718-01A_CRC_C->T
TCGA-DA-A3F5-06A_SKCM_C->T
TCGA-EE-A3J5-06A_SKCM_C->T
COSMIC Regions
SMRs (+)
Gaps
Human
Chimp
Gorilla
Orangutan
Gibbon
Rhesus
Crab-eating_macaque
Mouse
Dog
Elephant
4.88
0
-4.5
7q31.2

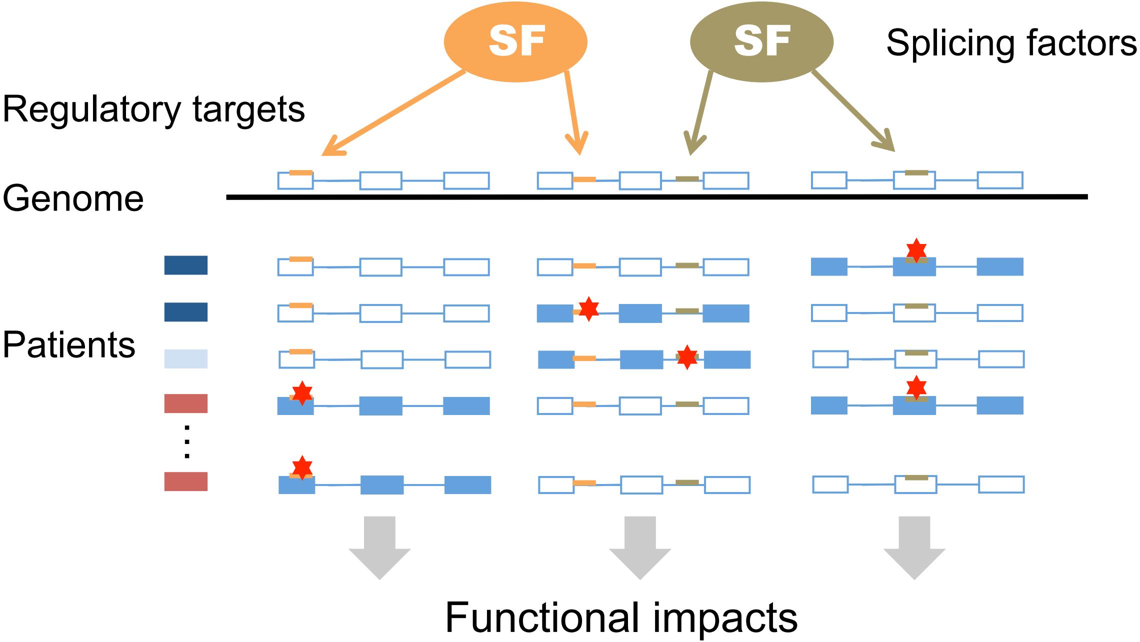

a
All SMRs
A>C
A>G
A>T
C>A
C>G
C>T
G>A
G>C
G>T
T>A
T>C
T>G
Mutations count
1000
500
0
Substitution
A>C
A>G
A>T
C>A
C>G
C>T
G>A
G>C
G>T
T>A
T>C
T>G
All Enriched k-mers
Mutations count
200
100
0
b
5SSC | 5SS
All Mutations
800
600
400
200
0
1 2 3 4 5 6 7
Information content
2
1.5
1
0.5
0
1 2 3 4 5 6 7
Position
SNP & Mut. Proportion
0.15
0.10
0.05
0.00
1 2 3 4 5 6 7
#Mutation
60
40
20
0
Substitution
A->C C->A G->A T->A
A->G C->G G->C T->C
A->T C->T G->T T->G
c
3SSC | 3SS
All Mutations
400
300
200
100
0
1 2 3 4 5 6 7
Information content
2
1.5
1
0.5
0
1 2 3 4 5 6 7
Position
0.15
0.10
0.05
0.00
1 2 3 4 5 6 7
#Mutation
30
20
10
0
Substitution
A->C C->A G->A T->A
A->G C->G G->C T->C
A->T C->T G->T T->G
d
Scale
10 bases
hg19
chr17:
29,654,850
29,654,860
---> G G C A G T T C T G A C C C G A G T T T A C G G T A G G T T T T
Common SNPs(150)
TCGA-EE-A3J5-06A_SKCM_G->A
TCGA-AX-A0J1-01A_UCEC_G->A
TCGA-78-7146-01A_LUAD_G->T
TCGA-EE-A3J5-06A_SKCM_G->A
TCGA-EE-A3JI-06A_SKCM_G->T
5SS_5SSC_GTAGGT_15
SMRs (+)
NF1/NM_001042492
NF1/NM_000267
Gaps
Human G G C A G T T C T G A C C C G A G T T T A C G G T A G G T T T T
Chimp G G C A G T T C T G A C C C G A G T T T A C G G T A G G T T T T
Gorilla G G C A G T T C T G A C C C A A G T T T A C G G T A G G T T T T
Orangutan G G C A G T T C T G A C C C G A G T T T A C G G T A G G T T T T
Gibbon G G C A G T T C T G A C C C G A G T T T A C G G T A G G T T T T
Rhesus G G C A G T T C T G A C C C G A G T T T A C G G T A G G T T T T
Crab-eating_macaque G G C A G T T C T G A C C C G A G T T T A C G G T A G G T T T T
Mouse G G C A G T T C A G A C C C T A G T T T A C G G T A G G T A T T
Dog G G C A G T T C A G A C C C T A G T T T A C G G T A G G - T T T
Elephant G G C A G C T C A G A C C C T A G T T T A C G G T A G G T C T T
4.88 -
0 -
-4.5
17q11.2
e
Scale
20 bases
hg19
chr4:
1,803,540
1,803,550
1,803,560
1,803,570
1,803,580
1,803,590
TCGA-CF-A27C-01A_BLCA_C->G
TCGA-DK-A1AA-01A_BLCA_C->G
TCGA-DK-A1AG-01A_BLCA_C->G
TCGA-H4-A2HQ-01A_BLCA_C->G
TCGA-BA-A4IH-01A_HNSC_C->G
SMRs (+)
FGFR3
FGFR3
FGFR3
FGFR3
FGFR3
FGFR3


a
5SS
Proportion of patients
0.5
0.4
0.3
0.2
0.1
0.0
5SSC
SRSF1
SRSF10
PCBP3
SART3
HNRNPH2
PCBP1
RBM5
ZNF638
PABPC4
RBMS3
A1CF
IGF2BP2
G3BP2
RBM24
RBM42
ANKHD1
HNRNPK
HNRPLL
SRSF9
RBM8A
b
3SS
Proportion of patients
0.5
0.4
0.3
0.2
0.1
0.0
3SSC
PCBP1
CPEB4
HNRPLL
ZNF638
RBM42
IGF2BP2
SRSF10
TIA1
SRSF9
A1CF
c
CDS
Proportion of patients
0.5
0.4
0.3
0.2
0.1
0.0
PCBP1
SRSF10
d
INTRON
Proportion of patients
0.5
0.4
0.3
0.2
0.1
0.0
RBM8A
ZNF638
PCBP1
PCBP3
IGF2BP2
RBM42
YBX2
SRSF10
CNOT4
PTBP1
A1CF
YBX1
PPRC1
HNRNPA1L2
HNRNPK
KHDRBS3
BRUNOL5
G3BP2
PCBP2
RBM24
SNRPA
HNRNPA1
RBFOX1
RBM4
RBM45
SNRNP70
MBNL1
MSI1
HNRPLL
LIN28A
SRSF9
ANKHD1
CPEB4
ZC3H10
IGF2BP3
KHDRBS1
SAMD4A
RBM41
Cancer
BLCA
BRCA
CRC
GBM
HNSC
KICH
KIRC
LGG
LUAD
LUSC
PRAD
SKCM
THCA
UCEC
e
SRSF10 | 3SS
All Mutations
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
6
Information content
2
1.5
1
0.5
0
AAAGAA
Position
SNP & Mut. Proportion
0.6
0.4
0.2
0.0
#Mutation
8
6
4
2
0
Substitution
A->C
G->A
G->T
f
SRSF10 | CDS
All Mutations
15
10
5
0
1
2
3
4
5
6
Information content
2
1.5
1
0.5
0
AAAGAA
Position
SNP & Mut. Proportion
0.3
0.2
0.1
0.0
#Mutation
5
4
3
2
1
0
Substitution
A->C
A->G
G->T
g
SRSF10 | INTRON
All Mutations
200
100
0
1
2
3
4
5
6
Information content
2
1.5
1
0.5
0
CAGAGA
Position
SNP & Mut. Proportion
0.3
0.2
0.1
0.0
#Mutation
40
20
0
Substitution
A->C
A->T
C->G
G->A
G->T
A->G
C->A
C->T
G->C


a
5UTR
Proportion of patients
0.5
0.4
0.3
0.2
0.1
0.0
HNRPLL
RBM42
HNRNPA2B1
PPRC1
RBM4
PCBP3
LIN28A
PTBP1
c
3UTR
Proportion of patients
0.5
0.4
0.3
0.2
0.1
0.0
PTBP1
HNRNPC
PCBP3
A1CF
HuR
IGF2BP2
YBX2
ZNF638
Cancer
BLCA
BRCA
CRC
GBM
HNSC
KICH
KIRC
LGG
LUAD
LUSC
PRAD
SKCM
THCA
UCEC
b
PCBP3 | 5UTR
All mutations
25
20
15
10
5
0
1
2
3
4
5
6
Information content
2
1.5
1
0.5
0
C T T T C C
1
2
3
4
5
6
Position
SNP & Mut. Proportion
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
6
10
5
0
Substitution
C->T
T->A
T->C
d
HNRNPC | 3UTR
All mutations
20
15
10
5
0
1
2
3
4
5
6
Information content
2
1.5
1
0.5
0
T T T T T C
1
2
3
4
5
6
Position
SNP & Mut. Proportion
0.3
0.2
0.1
0.0
1
2
3
4
5
6
6
4
2
0
Substitution
C->T
T->C
T->G


a
Significant TPM changes
Significant Cases
0
50
100
150
CDS
5UTR
3UTR
EXON
5SS
INTRON
cancer
BLCA
BRCA
CRC
GBM
HNSC
KICH
KIRC
LGG
LUAD
LUSC
PRAD
SKCM
b
Significant TPM changes (SF)
Significant Cases
0
5
10
15
20
ZNF638
PCBP1
RBM8A
RBM42
HNRNPK
PTBP1
A1CF
YBX2
HNRNPA1L2
SRSF10
CNOT4
PPRC1
G3BP2
IGF2BP2
KHDRBS3
PCBP3
BRUNOL5
RBM4
RBM45
ANKHD1
CPEB4
MBNL1
PCBP2
RBM24
SAMD4A
SNRPA
HNRNPA1
IGF2BP3
LIN28A
MSI1
SNRNP70
SRSF9
YBX1
ZC3H10
cancer
BLCA
BRCA
CRC
GBM
HNSC
KICH
KIRC
LGG
LUAD
LUSC
PRAD
SKCM
c
SRSF10
CRC
BRCA
TMEM260-012
NAV2-201
RAG1-002
DMD-015
MAPK10-014
MAPK10-027
TCGA_CA_6717_01
TCGA_AD_6964_01
TCGA_EI_6917_01
TCGA_A6_6141_01
TCGA_BH_A0BW_01
Scale
chr4:
10 bases
hg19
87,375,880
87,375,890
<---
TCTTTTTAAAGAGACCGGAACGATCT
MAPK10/uc010ikh.1
Common SNPs(150)
TCGA-AA-3977-01A_CRC_G->T
TCGA-CA-6717-01A_CRC_G->T
TCGA-EI-6917-01A_CRC_G->T
INTRON_SRSF10_CAGAGA_63
SMRs (-)
Gaps
Human TCTTTTTAAAGAGACCGGAACGATCT
Chimp
Gorilla
Orangutan
Gibbon TCTTTTTAAAGAGACCGG-ACGATCT
Rhesus TCTTTTTAAAGAGACCGGAACGATCT
Crab-eating_macaque TCTTTTTAAAGAGACCGGAACGATCT
Mouse
Dog
Elephant
4.88
100 Vert. Cons
0
-4.5
INTRON | CRC | MAPK10 | MAPK10-027
TCGA_EI_6917_01
log2(TPM)
-2
-4
-6
-8
-10
Mutated
Non.mutated
SRSF10
INTRON | CRC | MAPK10 | MAPK10-014
-2.5
-5.0
-7.5
-10.0
TCGA_EI_6917_01
log2(TPM)
Mutated
Non.mutated
SRSF10
d
MBNL1
CRC
SKCM
ETV1-007
ETV1-002
TCGA_CA_6717_01
TCGA_EE_A3JI_06
100 Vert. Cons
Scale
chr7:
10 bases
hg19
13,975,050
<---
TCAAATTCACTTCGTTAAATTAAGTTCGTTTCGCTTTCCC
ETV1/NM_001163150
ETV1/NM_001163151
ETV1/NM_004956
ETV1/NM_001163152
ETV1/NM_001163149
ETV1/NM_001163148
ETV1/NM_001163147
ETV1/NR_120445
ETS_PEA3_N
Common SNPs(150)
TCGA-EE-A3JI-06A_SKCM_C->T
TCGA-AP-A051-01A_UCEC_G->A
TCGA-B5-A11H-01A_UCEC_G->A
TCGA-CA-6717-01A_CRC_C->T
INTRON_MBNL1_CGCTTT_12
SMRs (-)
Gaps
2
Human TCAAATTCACTTCGTTAAATTAAGTTCGTTTCGCTTTCCC
Rhesus TCAAATTCACTTCGTTAAATTAAGTTCGTTTCGCTTTCCC
Mouse GCAAATTCACTTCGTTAATTT-------CTTCACTTCCTC
Dog TCAAATTCACTCCGTTAAATTAAGTCCGTTTCGCTTTCCC
Elephant TCAAATTCACTTCTTTAAATTAAGTTCGTTTCACTTTCCC
4.88
0
-4.5
INTRON | CRC | ETV1 | ETV1-002
TCGA_CA_6717_01
log2(TPM)
-5
-6
-7
-8
-9
-10
Mutated
Non.mutated
MBNL1
INTRON | SKCM | ETV1 | ETV1-007
TCGA_EE_A3JI_06
log2(TPM)
6
4
2
0
-2
-4
Mutated
Non.mutated
MBNL1


a
Significant PSI changes
Significant Cases
0
5
10
15
20
25
5UTR
EXON
5SS
3SS
INTRON
cancer
BLCA
BRCA
HNSC
KIRC
LUAD
LUSC
SKCM
THCA
UCEC
b
Significant PSI changes (SF)
Significant Cases
0
2
4
6
8
5SSC
IGF2BP2
3SSC
HNRNPA2B1
HNRPLL
PCBP1
PTBP1
RBM8A
SAMD4A
LIN28A
SRSF10
ZNF638
CPEB4
HNRNPA1L2
PCBP3
PPRC1
RBM41
RBM45
RBM5
SRSF1
YBX2
cancer
BLCA
BRCA
HNSC
KIRC
LUAD
LUSC
SKCM
THCA
UCEC
c
5SS
SKCM
UCEC
LUAD
THCA
chr17:57775338:57784742_PTRH2
chr17:57775338:57784733_PTRH2
chr16:15029253:15035670_NPIPA1
chr14:55480313:55493747_WDHD1
chr12:122615479:122617896_MLXIP
chr1:155289718:155290201_FDPS
chr1:155289478:155290201_FDPS
chr17:73840028:73840303_UNC13D
5SSC
5SSC
RBM5
Patients
d
Scale
chr17:
10 bases
hg19
57,784,720
57,784,730
57,784,740
<---
GCGGGACATTCCTCGGTTGGTTGGATGGATGTCATGGAG
PTRH2/NM_001015509
PTRH2/NM_016077
TCGA-A5-A0GA-01A_UCEC_G->T
5SS_5SSC_GTAGGT_15
SMRs (-)
Gaps
2
2
Human
Chimp
Gorilla
Orangutan
Gibbon
Rhesus
Crab-eating_macaque
Mouse
Dog
Elephant
4.88
0
-4.5
17q23.1
5SS | UCEC | PTRH2 |
chr17:57775338-57784733
PSI value
0.00
0.25
0.50
0.75
1.00
TCGA_A5_A0GA_01A
Mutated
Non.mutated
5SSC
5SS | UCEC | PTRH2 |
chr17:57775338-57784742
PSI value
0.00
0.25
0.50
0.75
1.00
TCGA_A5_A0GA_01A
Mutated
Non.mutated
5SSC
e
5UTR
SKCM
chr7:100210107:100210354_MOSPD3
chr16:2510381:2510625_C16orf59
chr16:2510170:2510625_C16orf59
chr15:49448212:49493360_COPS2
chr13:41835050:41837622_MTRF1
chr15:49448212:49493360_GALK2
HNRNPA2B1
HNRPLL
LIN28A
PCBP1
PCBP3
PTBP1
Patients
f
5UTR | SKCM | C16orf59 |
chr16:2510170-2510625
PSI value
0.00
0.25
0.50
0.75
1.00
TCGA_GN_A266_06A
Mutated
Non.mutated
HNRPLL
5UTR | SKCM | C16orf59 |
chr16:2510381-2510625
PSI value
0.00
0.25
0.50
0.75
1.00
TCGA_GN_A266_06A
Mutated
Non.mutated
HNRPLL
Scale
chr16:
10 bases
hg19
2,510,070
2,510,080
2,510,090
2,510,100
--->
CCGGCCGGGAAGGGGCGGCCTTCCCTCCCGGAAGGGGCGTG
ENST00000361837
ENST00000569496
ENST00000567489
nmon SNPs(150)
TCGA-DA-A1IC-06A_SKCM_G->A
TCGA-EE-A29B-06A_SKCM_G->A
TCGA-EE-A2M5-06A_SKCM_G->A
TCGA-EE-A2MI-06A_SKCM_G->A
TCGA-EE-A3J5-06A_SKCM_G->A
TCGA-ER-A19J-06A_SKCM_G->A
TCGA-GN-A266-06A_SKCM_G->A
TCGA-D9-A148-06A_SKCM_G->A
TCGA-EE-A2M5-06A_SKCM_G->A
TCGA-EE-A2MI-06A_SKCM_G->A
TCGA-ER-A19J-06A_SKCM_G->A
TCGA-GN-A266-06A_SKCM_G->A
5UTR_HNRPLL_AGGGGC_6
5UTR_HNRPLL_GGGGCG_8
5UTR_C16orf59_12_CCCGGAAGGGGCGTG
Gaps
Human
Chimp
Gorilla
Orangutan
Gibbon
Rhesus
-eating_macaque
Mouse
Dog
Elephant
4.88
0
-4.5


SF
SF
Splicing factors
Regulatory targets
Genome
Patients
Functional impacts