Recurrent alternative splicing isoform switches in tumor samples provide novel signatures of cancer

Endre Sebestyén[1], Michał Zawisza[2], Eduardo Eyras[1,3,4]

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

[2]Universitat Politècnica de Catalunya, Jordi Girona 1-3, Barcelona E08034, Spain

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

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

Abstract

Cancer genomics has been instrumental to determine the genetic alterations that are predictive of various tumor conditions. However, the majority of these alterations occur at low frequencies, motivating the need to expand the catalogue of cancer signatures. Alternative pre-mRNA splicing alterations, which bear major importance for the understanding of cancer, have not been exhaustively studied yet in the context of recent cancer genome projects. In this article we analyze RNA sequencing data for more than 4000 samples from The Cancer Genome Atlas (TCGA) project, including paired normal samples, to detect recurrent alternative splicing isoform switches in 9 different cancer types. We first investigate whether alternative splicing isoform changes are predictive of tumors by applying a rank-based algorithm based on the reversal of the relative expression of transcript isoforms. We find that consistent alternative splicing isoform changes can separate with high accuracy tumor and normal samples, as well as some cancer subtypes. We then searched for those changes that occur in the most abundant isoform, i.e isoform switches, and are therefore more likely to have a functional impact. In total we detected 244 isoform switches, which are associated to functional pathways that are frequently altered in cancer and also separate tumor and normal samples accurately. We further assessed whether these isoform changes are associated to somatic mutations. Surprisingly, only a few cases

appear to have association, including the putative tumor suppressor FBLN2 and the tumor driver MYH11, which show association of an isoform switch to mutations and indels on the alternatively spliced exon. However, the number of observed mutations is in general not sufficient to explain the frequency of the found isoform switches, suggesting that recurrent isoform switching in cancer is mostly independent of somatic mutations. In summary, we present an effective approach to detect novel alternative splicing signatures that are predictive of tumors. Moreover, the same methodology has led to uncover recurrent isoform switches in tumors, which may provide novel prognostic and therapeutic targets.

Software and data are available at: https://bitbucket.org/regulatorygenomicsupf/iso-ktsp and http://dx.doi.org/10.6084/m9.figshare.1061917

Introduction

Cancer genome projects are instrumental to describe the genetic heterogeneity of tumors and to uncover recurrent alterations that may serve as new biomarkers for prognosis and therapeutic targets (TCGA 2012a-d, TCGA 2013). However, known actionable alterations tend to occur at low frequency or are often absent in an individual tumor sample, which hinders the choice of appropriate therapeutic strategies (Vogelstein et al. 2013, Hudson 2013). There is therefore a need to expand the catalogue of molecular signatures in cancer. Alternative splicing alterations, which bear major importance in terms of the understanding and treatment of cancer (Bonomi et al. 2013), have not been exhaustively studied yet in the context of recent cancer genomics efforts. Alternative splicing may confer a selective advantage to the tumor, such as angiogenesis (Amin et al. 2011), proliferation (Bechara et al. 2014), cell invasion (Venables et al. 2013) and avoidance of apoptosis (Izquierdo et al. 2005). Some of these alterations may be caused by somatic mutations (Ward and Cooper 2010), but can also take place as a result of changes in expression, amplifications and deletions in splicing factors (Karni et al. 2007, Furney et al. 2013). This suggests that similar splicing alterations may have different genetic origins and still confer equivalent tumorigenic properties to cells. Accordingly, to uncover potential markers of prognosis and targets of therapy, it is of utmost relevance to describe the patterns of alternative splicing in tumors.

Numerous genome wide surveys have highlighted the role of alternative splicing patterns in tumors. These have mostly been based on the measurement of local patterns of splicing, encoded as events, and studied using microarrays (Thorsen et al. 2008, Lapuk et al. 2010, Misquitta-Ali et al. 2011), RT-PCR platforms (Klinck et al. 2008), or RNA sequencing (Liu et al. 2012). The description of alternative splicing in terms of simple events facilitates the validation using PCR methods and the characterization of regulatory mechanisms using sequence analysis and biochemical approaches. However, alternative splicing takes place through a change in the relative abundance of the transcript isoforms expressed by a gene, and splicing alterations important for tumor progression may involve complex patterns that are not easily described in terms of simple events. Accordingly, to ultimately determine the functional impact of splicing alterations, it is important to describe these in terms of transcript isoforms changes. This has been shown to be relevant for TP53, which produces multiple isoforms with complex variations and with different roles in tumors (Bourdon et al. 2005). Similarly, recent analyses have shown that there are transcript isoforms specific of lung and breast cancers (Kalari et al. 2012, Eswaran et al. 2013), and that transcript analysis can improve expression-based tumor classification (Zhang et al. 2013, Pal et al. 2014). Additionally, transcript isoform changes can be essential to detect resistance to anti-tumor drugs (Mitra et al. 2009, Poulikakos et al. 2011). Thus, the detection of transcript isoform changes characterizing specific tumor types can provide new cancer signatures and could be crucial for the development of diagnostic, prognostic and therapeutic strategies.

With the aim to describe the transcript isoform changes that are characteristic of tumors, we have analyzed more than 4000 RNA sequencing (RNA-Seq) samples available from the Cancer Genome Atlas (TCGA) project. In order to perform an analysis that is robust to the variability between samples from different individuals, we have applied a new rank-based algorithm that searches for consistent reversals of relative isoform expression. This algorithm provides the minimal set of isoform-pairs with relative expression changes that can accurately separate tumor from normal samples. Moreover, the obtained isoform-pairs can accurately classify unseen tumor data. We have applied the same algorithm to breast, lung and colon cancer subtypes to obtain isoform signatures that separate subtypes from each other. In particular, we found a highly significant signature for basal-like breast tumors that distinguish them

from other breast cancer subtypes. We also found that a number of the identified significant isoform changes correspond to transcript isoform switches, for which the relative expression change occurs in the most abundant isoform, and are therefore more likely to have a functional impact. We found a total of 244 isoform switches in all cancer types. These switches can also accurately separate tumor and normal samples, and occur in genes belonging functional pathways frequently altered in cancer. Interestingly, only a few of these switches can be explained by somatic mutations in the gene locus, suggesting that recurrent isoform switching in cancer is mostly independent of somatic mutations. On the other hand, we found that for at least one case the isoform switch is mutually exclusive with mutations affecting the protein-coding region of the transcripts, suggesting that splicing alterations represent an alternative route towards cellular transformation. Our analyses show that recurrent transcript isoform switches represent important novel signatures in cancer that can serve as molecular markers and could lead to the development of new therapeutic targets.

Results

Systematic analysis of splicing isoform changes in cancer

Changes in the relative abundance of the alternative transcripts from a gene translate into a variation in their relative order in the ranking of transcript expression. Accordingly, the problem of finding alternative splicing changes in cancer can be reformulated in terms of the consistency of the reversals of their relative expression of transcript isoforms from the same gene. For this purpose we developed Iso-kTSP, which extends the principle of consistent expression reversals for gene expression (Geman et al. 2004, Tan et al. 2005, Price et al. 2007) to alternative splicing isoforms. The Iso-kTSP algorithm stores the ranking of isoform expression from multiple samples separated into two classes (Figure 1A). All possible isoform pairs from the same gene are sorted according to the sum of frequencies of the two possible relative orders occurring separately in each class, defined as score S_1 (Methods). This score provides an estimate of the probability for the isoforms to change relative order between the two classes. The top scoring isoform pairs are therefore the most consistent changes in isoform relative abundance for a gene between two classes,

tumor and normal, or between two tumor subtypes. Each one of these isoform pairs provides a classification rule based on the relative expression order, with a discrimination power related to the consistency of this reversal across samples.

Using cross-validation, the ranking of isoform-pairs is calculated at each iteration step on a balanced set leaving out one sample from each class, which are used for testing (Figure 1B). The prediction class for a new sample is obtained by evaluating the expression ranking in the new sample against the isoform pair rules. At each iteration step in the cross-validation, the top k-pairs (k=1… k_{max}, with k odd) are evaluated on the test set. Each isoform-pair from the same gene defines a rule, where each gene is only used once. The rule is defined such that for a pair, if the first isoform has lower expression than the second the sample is predicted to be normal, otherwise is predicted to be a tumor. Accordingly, the first isoform is defined as the tumor isoform, and the second as the normal isoform. Significance of the isoform-pair rules is measured by performing 1000 permutations of the sample labels (Figure 1C). At each permutation, the algorithm is run as before keeping only the pair with the highest score S_1. An isoform-pair is defined as significant if its score S_1 and information gain (IG) are larger than the maximum ones obtained from the permutation analysis. The global ranking of isoform-pairs and permutation analysis yields the list of significant isoform changes (Figure 1D). From this set, we derive a minimal classification model with the smallest odd number of isoform-pairs with the highest average performance (Figure 1E). The final classification is based on simple majority voting with an odd number of isoform pairs. On the other hand, some of the isoform-pairs from the list of significant cases are in fact isoform switches, for which the relative expression change occurs in the most abundant isoform of the gene, and which we detect by the anti-correlation of the relative inclusion levels or PSIs of the isoforms (Figure F). Finally, to further assess the accuracy of the minimal classification model, or that of a set of isoform switches, a blind test is carried out on samples that were not used for cross-validation (Figure 1G). On this set we measure the proportion of samples correctly labeled by the classifier, as well as the number of correct votes for each prediction.

TCGA acronym	Cancer type	Paired samples	Unpaired tumor samples	Reference
BRCA	Breast Invasive Carcinoma	107	929	(TCGA 2012c)
COAD	Colon adenocarcinoma	26	236	(TCGA 2012a)
HNSC	Head/neck squamous cell carcinoma	38	384	(Stransky et al. 2011)
KICH	Kidney Chromophobe	21	41	https://tcga-data.nci.nih.gov/
KIRC	Kidney renal clear cell carcinoma	71	434	(TCGA 2013)
LUAD	Lung adenocarcinoma	57	431	(Ding et al. 2008)
LUSC	Lung squamous cell carcinoma	50	433	(TCGA 2012b)
PRAD	Prostate adenocarcinoma	48	247	https://tcga-data.nci.nih.gov/
THCA	Thyroid carcinoma	58	439	https://tcga-data.nci.nih.gov/

Table 1. Number of analyzed paired and unpaired tumor samples from each cancer type: Breast carcinoma (BRCA), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear-cell carcinoma (KIRC) lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), prostate adenocarcinoma (PRAD), and thyroid carcinoma (THCA). All datasets were obtained from https://tcga-data.nci.nih.gov/. For the paired samples we also used the corresponding normal samples from the same patients. For the list of samples used see Supplementary File 1.

Recurrent alternative splicing isoform changes can separate tumor and normal samples

We downloaded the available RNA-Seq datasets for tumor and paired normal samples from TCGA for 9 cancer types (Table 1). Prediction of tissue types from gene expression with URSA (Lee et al. 2013) was used to detect and filter out outliers in the samples (Supplementary Figure 1). The tumor and paired normal samples that were kept show a remarkably similar pattern of tissue expression in BRCA, COAD, KIRC, PRAD and THCA (Supplementary Figure 1), and partly in KICH. On the other hand, LUAD, LUSC, and HNSC show heterogeneous patterns of predicted tissue types, probably due the mixed cellular origins (Lundin et al. 2013). The list of samples used for further analyses are given in Supplementary File 1.

For each cancer type, the Iso-kTSP algorithm was applied to the paired samples to obtain minimal classifiers to separate tumor and normal samples, and the blind test performed on the remaining unpaired tumor samples. This yielded different predictive models for the 9 cancer types (Figure 2A and Supplementary Figure 4A), with PRAD, THCA and KIRC having the lowest average accuracies, and LUSC, LUAD, COAD and KICH having 100% average accuracy. Moreover, the blind tests of these models on the unpaired tumor samples show overall accuracies greater than 84% (Figure 2B and Supplementary Figure 4B). These models provide a minimal set of isoform pairs whose relative expression can separate tumor and normal samples with high accuracy despite the variability of the transcript expression measurement across samples (Supplementary Figures 3-8) (Model files are given in Supplementary File 2). All the isoform-pairs in the derived models are significant according to our permutation analysis (Supplementary Figure 9). This significance also depends in general on the number of samples available and on the heterogeneity of the tumor samples. Permutation analysis for a varying number of input samples indicates that in order to obtain significant isoform changes, more than 13 samples are needed on average (Supplementary Figure 10), which is fulfilled by the cancer types analyzed.

The isoform changes detected include FBLN2, which appears as a single gene model for COAD and is part of the BRCA model (Figure 3A). FBLN2 has been proposed before to be a tumor suppressor (Law et al. 2012) and its cancer-related function seems to be specific of the protein produced in tumor cells (Baird et al. 2013). We

found that FBLN2 undergoes an isoform change related to the skipping of a protein coding exon (Supplementary Figure 11). Additionally, this isoform switch occurs in more than 98% of the unpaired tumor samples in BRCA and COAD (Figure 1A). In the case of LUAD, surprisingly, we found that the most informative isoform change does not occur in NUMB, known to have a splicing switch related to proliferation (Misquitta-Ali et al. 2011, Bechara et al. 2014), but in QKI. In fact, the isoform change in QKI cannot be described in terms of a simple alternative splicing event (Figure 2C). In contrast, LUSC model involves a different set of genes compared to LUAD, including the gene ZNF385A (Figure 2), whose protein product interacts with TP53 thereby promoting growth arrest (Das et al. 2007). The isoform change found is related to the use of an alternative first exon and alternative 3' splice-site (Supplementary Figure 12). Similarly to COAD, THCA and KIRC have single-gene models (Supplementary Figure 2). In particular, for THCA the model involves S100A13, which codes for a calcium binding protein and which has been proposed to be a new marker of angiogenesis in various cancer types (Massi et al. 2010). The isoform change involves an alternative first exon and classifies correctly 84.5% of the unpaired tumor samples (Supplementary Figure 13). Interestingly, S100A13 and another member of the S100 family, S100A16, also have isoform changes in KICH, even though they were not included in the KICH model (Supplementary Figure 13). The single isoform change in KIRC involves the production of an intron-retention transcript, annotated as non-coding in the gene CPAMD8 (Supplementary Figure 14). A similar case occurs in the gene NAGS, which is part of the KICH model (Supplementary Figure 2) and is related to an autosomal recessive urea cycle disorder (Häberle et al. 2003). We predict that a protein coding isoform changes in tumors into an isoform with a retained-intron that is annotated as non-coding (Supplementary Figure 15). Importantly, the loss of the protein coding isoform is predictive of 100% of the KICH tumor samples (Supplementary Figure 15). Other isoform changes are discussed in the supplementary material (Supplementary Figures 16-18). GFF tracks for the isoform-pairs in all derived models can be found in Supplementary File 3.

Changes in alternative splicing isoforms can discriminate tumor subtypes

Cancers are generally classified into subtypes to facilitate patient stratification for more precise prognosis and selection of therapeutic strategy. In particular, breast cancer classification has been recently refined based on molecular information from multiple sources (TCGA 2012c). We thus decided to investigate whether breast cancer subtypes are associated with consistent isoform changes when compared to each other. We separated the BRCA tumor samples into luminal A, luminal B, Her2+ and basal-like as labeled by TCGA (TCGA 2012c) (Supplementary File 1) and run the Iso-kTSP algorithm comparing each subtype against a pool from the rest. In order to maintain balanced sets for the comparison and avoid biases due to sample selection, we subsampled 100 times 45 arbitrary samples for a given subtype and a pool of 15 from each of the other three subtypes together. At each iteration step, we performed permutation analysis of the labels to determine the significance of the detected isoform changes. We found that only basal-like tumors showed isoform changes that were significant in more than 80% of the sampling iterations (Figure 4A and Supplementary Figure 19). Among the most significant cases we found KIF1B, which has been implicated in apoptosis (Schlisio et al. 2008); ATP1A1, proposed to have tumor suppressor activity (Cao et al. 1997); ITGA6, found to be required for the growth and survival of a stem cell like subpopulation of MCF7 cells (Cariati et al. 2008); and CTNND1, whose alternative splicing was previously related to cell invasion and metastasis (Yanagisawa et al. 2008) (Figure 4A). We selected the top 7 isoform-pairs in basal-like that were significant in more than 80% of the iterations. This model classified correctly 93.6% of all the BRCA tumor samples, with 47% of the samples fulfilling all 7 isoform change rules (Figure 4B). Although this cannot be considered a blind test, it provides an estimation of the expected accuracy. For the other BRCA subtypes we found much lower consistency of the isoform changes and none of them were significant for more than 13% of the permutation tests (Supplementary Figure 20).

Four different subtypes have been defined based on mRNA expression for the lung squamous cell carcinomas (LUSC): basal, classical, primitive and secretory; which have different clinical and biological characteristics (Wilkerson et al. 2010). We applied the same approach as above to the four LUSC subtypes using the subtype labeling from TCGA (TCGA 2012b), comparing 24 samples from each subtype

against the pool of three sets of 8 arbitrary samples from the other subtypes. The most relevant isoform change was found for gene GCNT2 in association to the classical subtype in at least 60% of the subsampling iterations, but significant only in 22% of them (Supplementary Figure 21 and 22). Interestingly, GCNT2 overexpression has been linked to breast and lung cancer metastasis (Zhang et al. 2011). All other found changes occurred at lower frequencies and showed significance in no more than 3% of the iterations. Despite of the low recurrence of the isoform changes, tests on the entire set of LUSC tumor samples was able to separate correctly classical from other subtypes for more than 80% of them (Supplementary Figure 21).

Colorectal cancers have been classified into hypermutated and non-hypermutated, where non-hypermutated tumors have generally worse prognosis (TCGA 2012a). Following previous definitions (TCGA 2012a), we labeled COAD samples as hypermutated if they had more than a total of 250 mutations, and as non-hypermutated those with less than 250 mutations. We then compared both subtypes by subsampling 40 samples from each one 100 times. This analysis yielded specific isoform changes between the two types occurring in more than 40% of the iterations (Supplementary Figure 23), including a change in the long non-coding RNA gene antisense of NUTM2A (NUTMA2A-AS1), which appeared in 57% of the models. We tested two different models with the top 5 and 13 isoform-pairs, obtaining an accuracy of more than 80% on the total COAD dataset (Supplementary Figure 18). Models and GFF tracks for all subtype models are provided in Supplementary Files 2 and 3, respectively.

A catalogue of alternative isoform switches in cancer

The alternative isoform changes described above can separate tumor and normal samples, and in some cases specific cancer subtypes. However, these models are optimized to have the minimum number of isoform-pairs and maximum average accuracy, which is convenient for defining biomarkers with potential clinical applications. On the other hand, the frequency of these isoform changes does not imply functional relevance. However, among the recurrent isoform-changes, those more likely to be functionally relevant will be the ones for which the change occurs in the most abundant isoform, i.e. isoform switches (Figure 1F). Accordingly, in order to

obtain all the significant isoform changes with a possible functional relevance in cancer, we decided to calculate all the significant isoform switches between tumor and normal samples. We first retrieved all those genes with significant isoform changes according to our permutation analysis (Figure 1D). This yielded a total 1178 genes for the 9 cancer types. We further filtered these genes by imposing a score $S_I > 0.5$, which corresponds to selecting isoform-pairs with a change in more than 75% of the samples. To select for switches, we kept those cases for which the relative inclusion levels of the isoforms anti-correlate (Figure 1F), as observed for FBLN2, QKI and other genes (Figure 3 and Supplementary Figures 12-18). We thus selected those isoform-pairs having an anti-correlation of PSI values of $R < -0.8$ (Spearman). Finally, we kept only those with average expression per isoform of > 1 TPMs in either tumor or normal samples.

These criteria gave rise to a total of 244 isoform switches, with 59 of them appearing in more than one cancer type (Figure 5) (Supplementary File 4). The most common of the switches is the one described above for FBLN2. From the total 244 switches, 10 occur in known cancer drivers (Figure 5). Moreover, we also found switches in genes whose splicing has been associated before with cancer, like CD44, which has been observed to be relevant in colon cancer initiation (Du et al. 2008), and SLC39A14, whose alternative splicing is regulated by WNT in colon cancer (Thorsen et al. 2011). LUAD, KIRC and LUSC are the cancer types with most switches, with 85, 65 and 54, respectively. LUSC and LUAD have 33 switches in common. In contrast, KIRC and KICH have only 2 switches in common. HNSC and PRAD are the cancer types with the fewest switches, 7 and 2 respectively. Although functional analysis did not yield any significantly enriched Reactome pathways (Croft et al. 2014), isoform switches appear frequently in signal transduction, immune system and metabolism related pathways (Figure 4 and Supplementary Figure 24). On the other hand, Gene-Ontology analysis shows enrichment of multiple categories, including actin activity in relation to cell motility and migration, in categories related to extracellular organization, as well as in response to estrogen and regulation of MAPK activity (Supplementary Figure 25).

We also tested the accuracy of switches as predictive models by performing blind tests on the unpaired tumor samples (Figure 1G) and found accuracies of around 90% and higher (Supplementary Figure 26). These isoform switches are thus good

predictors of tumor samples. All details for the found isoform switches and corresponding GFF tracks are provide in Supplementary Files 4 and 5.

Isoform switches in cancer are not frequently associated with somatic mutations

As splicing changes may be triggered by somatic mutations (Ward and Cooper 2010), we thus investigated whether any of the detected isoform switches may be caused by recurrent somatic mutations in the same genomic locus. To this end, we tested whether in tumor samples there was any association between the presence of the isoform switch and somatic mutations in the region of the transcript isoforms undergoing the switch. Since in addition to intronic mutations, synonymous as well as non-synonymous mutations could alter the splicing of a gene (Sterne-Weiler & Sanford 2014), we considered all mutation types available in TCGA: coding-related (nonsense, missense, frameshift and indel) and non-coding-related (synonymous, splice-site and RNA) mutations. For each isoform-switch and for each cancer type, we calculated the Jaccard index across all samples for the association between the presence of the switch and the presence of somatic mutations (Figure 6A) (see Methods for details). These Jaccard indexes agree with mutual information and do not correlate with the average mRNA length of the switches (Supplementary Figure 27). This analysis shows genes FBLN2, MYH11, FLNA and TNC to have the strongest association between mutations and switches (Figure 6A and Supplementary Figure 20). These four genes are also the ones with the highest frequency of mutated samples (Figure 6A). For FBLN2, we found several mutations in BRCA and COAD samples on the alternative exon and the flanking constitutive exons (Figure 6B). However, there are not enough mutations to explain all the switches observed. We also found frequent mutations in the alternatively spliced region of the oncogene MYH11. In particular, we found recurrent deletions and insertions on the alternative exon in COAD and BRCA tumor samples that coincide with the presence of the switch (Figure 6C). The location of these indels coincides with a region of low conservation and is next to a putative binding site for the splicing factor SRSF1 (Supplementary Figure 28). Nonetheless, the number of found mutations cannot explain either the frequency of the switches observed.

Somatic mutations could also affect the magnitude of the splicing change in specific samples. We therefore tested, for each isoform-switch, whether the presence of mutations is associated with a larger difference of PSI between the tumor and normal isoforms involved in the switch (Methods). Among the four most significant cases, we found FBLN2 and EHBP1 (Supplementary Figure 29). These two cases show some differences in the distributions of samples with and without mutations (Supplementary Figure 29). However, the proportion of mutated samples is very small to make a reliable comparison and after multiple-testing correction, none of the found cases remained significant. This suggests that, except for a limited number of cases, mutations may not be the main cause of the recurrent splicing switches we have found in cancer.

We thus hypothesized that mutations and isoform switches may occur independently as two alternative mechanisms of functional transformation in cancer. To test this possibility, we measured how frequently mutations that affect the protein-coding region occur in tumor samples without the isoform switch in the same gene by defining a mutual-exclusion score based on the number of samples with no switch but with protein-affecting mutations (Methods). We found that in general the mutual exclusion score correlates with the overall proportion of mutated samples (Supplementary Figure 30). However, the number of samples with switch and mutation is generally comparable or higher, except for the gene Tenascin C (TNC), for which we find more samples with a protein-affecting mutation and no switch than with switch and protein-affecting mutation (Supplementary Figure 30). This indicates that there is no strong bias towards this mutual exclusion. We conclude that there are currently not a sufficient number of mutations that can provide an explanation of the described recurrent isoform switches. Nonetheless, there are a few cases for which this association may exist, as described for the genes FBLN2, MYH11 and TNC.

Discussion

We have applied the principle of relative expression reversals (Geman et al. 2009, Tan et al. 2005) to the search of recurrent alternative splicing isoform changes in tumors using available RNA-Seq data from the TCGA project for 9 different cancer types. In our implementation of this algorithm for isoforms (available at:

https://bitbucket.org/regulatorygenomicsupf/iso-ktsp) each classification rule is described in terms of the relative expression of a single pair of isoforms per gene. In this context, the principle of reversals has a natural interpretation as an alternative splicing change between two conditions. This algorithm provides robust classifiers despite of the variability of isoform expression across tumor samples, as the models are not dependent on parameterizations or on any normalization that would maintain the order of the isoform expression. This rank-base method is especially useful for isoform expression from RNA-Seq data, since between-sample normalization methods are not yet fully established. In fact, our approach is applicable to integrate data from heterogeneous platforms, as long as they provide a meaningful ranking of expression. Moreover, the method produces significant isoform changes from a relatively small number of samples, which makes it useful for medium-sized sequencing projects.

We have derived classification rules based on isoform changes that can distinguish tumor from normal samples, and between some tumor subtypes. The predictive models show overall a high accuracy when tested on held-out datasets. Except for the cases of colon, thyroid and kidney carcinomas, individual isoform-pair rules do not show in general a strong predictive power. However, in combination, they accurately classify tumor samples in the blind test. This suggests that splicing alterations are heterogeneous in tumor samples, but in combination they provide characteristic signatures, similarly to the patterns of somatic mutations (Vogelstein et al. 2013). This heterogeneity is further highlighted by the fact that different cancer types share a small fraction of the found isoform changes. Although some of these changes may be explained by the differences in the cell composition of tumors (Venables et al. 2013, Mallinjoud et al. 2014), we observed an homogenous pattern of predicted tissue types in tumor and normal samples for most of the cancer types analyzed, suggesting that the splicing changes are not a consequence of differences in cell type composition.

Comparative analysis between cancer subtypes only yielded a significant model for basal-like breast tumors, which includes genes with known functional relation to cancer. For the LUSC subtypes we did not find significant changes, and only the classical subtype showed frequent isoform changes in the gene GCNT2. Similarly, COAD subtypes, based on the mutations of the samples, did not show significant

isoform changes, but we observed frequent changes in a non-coding RNA gene antisense of NUTMA2A. These results indicate that most of the subtypes considered have similar alternative splicing patterns.

Our initial analysis shows that isoform changes hold sufficient information to separate tumor and normal samples, and specific tumor subtypes. This suggests they could serve as effective molecular markers, as they would only require measuring the expression of two isoforms per gene, for a small number of genes. Furthermore, the application of this method to separate tumors according to clinical information will provide a useful prognostic tool. On the other hand, these signatures are not necessarily related to a biological effect specifically relevant for the tumor. To investigate this aspect, we selected all those significant isoform changes that are also isoform switches, i.e. the change occurs in the most abundant isoform, and therefore more likely to have a functional impact. We found 244 such isoform switches, which can also accurately separate tumor and normal samples, and some of which appear in multiple cancer types. Interestingly, many of the found isoform switches occur in pathways that are often altered in tumors, and some of them occur in known cancer driver genes, including CDKN2C, CTNNB1, ABI1 and MYH11. Additionally, we found that of the splicing switches cannot be described in terms of simple events. This is the case for QKI, for which we predict an isoform switch specific to lung adenocarcinoma. The found isoform switches provide an opportunity to develop experimental strategies based on the detection of specific protein isoforms. In particular, we found genes with isoform switches involved in cell communication pathways, including DST and FLNA, which could be potentially used for diagnostic or prognostic applications, or even for developing tumor-specific therapeutic targets with reduced cross-reactivity to other proteins.

Surprisingly, we did not find strong associations of somatic mutations with the isoform switches. It has been recently proposed that synonymous mutations in known cancer drivers may contribute to the oncogenic process (Supek et al. 2014). However, a direct link was not made between the observed somatic mutations and specific splicing changes measured in the same tumor samples. We found only a handful of genes with significant association between the isoform change and somatic mutations occurring in the same samples. These include the putative tumor suppressor FBLN2 and the cancer driver MYH11, the latter showing a recurrent indel in the alternative

exon in samples where it is skipped. Still, mutations alone cannot explain the splicing changes observed, as 99% of the transcripts analyzed appear mutated in less than 5% of the tumor samples, whereas the majority of switches occur in at least 50% of the samples. This could mean that there are many intronic mutations not represented in the currently available exome-based data, which could explain the observed variations. Alternatively, the recurrent switches could be explained by alterations in splicing factors. Although point mutations and indels on splicing factors do not occur with sufficient frequency to explain the switches (Furney et al. 2013), some splicing factors show frequent amplifications, deletions and changes in expression in tumors (Karni et al. 2007). Another interesting hypothesis is whether alterations in chromatin modifications or DNA methylation may be responsible for the observed changes. These alterations are frequent in cancers (Esteller 2007, Ellis et al. 2008) and can induce changes in splicing (Luco et al. 2010, Maunakea et al. 2013). Interestingly, the gene FBLN2, which presents a switch in various cancers, has been observed frequently methylated in breast and other epithelial tumors (Hill et al. 2010).

Consistent isoform switching in tumors thus seems generally independent of somatic mutations; and moreover, only 10 of the detected 244 switches occur in known cancer drivers. This raised the question of whether these two alterations could be actually mutually exclusive. We found that only TNC, linked to cell invasion in tumors (Hancox et al. 2009), has this mutual exclusion pattern between the isoform switch and somatic mutations affecting the coding regions, suggesting that, albeit to a limited extent, splicing switches may provide an alternative mechanism towards functional transformation in cancer. In summary, we have detected recurrent alternative splicing isoform changes that are predictive of various tumoral conditions, and which may have potential applications for diagnostic and prognostic purposes. The same methodology has allowed us to uncover recurrent isoform switches in tumors, which are likely to have a functional impact, and which may be useful to explore novel therapeutic strategies. Further research will be necessary to determine the functional impact produced by the described isoform changes and how these may actually contribute to the tumor. We hypothesize that the observed recurrent changes in splicing, regardless of their cause, may contribute together with mutations and other alterations to explain tumor formation; hence, providing novel signatures for cancer.

Methods

Available processed RNA-Seq data for tumor and normal samples was downloaded from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) for all cancer types together with the UCSC gene annotation from June 2011 (assembly hg19) and the somatic mutation data. To assess sample quality, the provided estimated read-counts per gene were analyzed using URSA (Lee et al. 2013) and sample pairs that did not cluster with the rest of the samples were removed (Supplementary Figure 1). The list of kept samples can be found in Supplementary File 1.

The abundance of every transcript per sample was calculated in transcripts per million (TPM) (Li et al. 2010) from the transcript-estimated read counts provided by TCGA and the isoform lengths from the UCSC (June 2011) annotation. No further normalization on the TPM values was performed. For each transcript, the relative abundance (or PSI) per sample was calculated by normalizing the TPM by the sum of TPMs for all transcripts in the gene. Genes with one single isoform or no HUGO ID were removed for the Iso-kTSP analysis.

The Iso-kTSP algorithm is based on the kTSP algorithm (Geman et al. 2004, Tan et al. 2005). It stores the isoform expression rankings in samples from two groups, C_m, $m=1,2$. For every pair of isoforms $I_{g,i}$ and $I_{g,j}$ in each gene g, Iso-kTSP calculates a score based on the frequencies of the two possible relative orders in both classes:

$$S_1(I_{g,i}, I_{g,j}) = P(I_{g,i} > I_{g,j} \mid C_1) + P(I_{g,i} < I_{g,j} \mid C_2) - 1$$

Where $P(I_{g,i} > I_{g,j} \mid C_1)$ and $P(I_{g,i} < I_{g,j} \mid C_2)$ are the frequencies at which the isoform $I_{g,i}$ appears later than, or before, $I_{g,j}$ in the expression ranking of classes C_1 or C_2, respectively. To avoid possible ties, a second score S_2 is used, which is based on the average rank difference per class C_m for each isoform pair, as proposed previously (Tan et al. 2005). Defining $R(I_{g,i} \mid S_a, C_m)$ as the rank of isoform $I_{g,i}$ in sample S_a and class C_m, the average rank difference between two isoforms in a given class is calculated as

$$g(I_{g,i}, I_{g,j} \mid C_m) = \frac{1}{|C_m|} \sum_a \left(R(I_{g,i} \mid S_a, C_m) - R(I_{g,j} \mid S_a, C_m) \right)$$

where $|C_m|$ denotes the number of samples in class C_m. The score S_2 for an isoform-pair is then defined as (Tan et al. 2005):

$$S_2(I_{g,i}, I_{g,j}) = \left| g(I_{g,i}, I_{g,j} \mid C_1) - g(I_{g,i}, I_{g,j} \mid C_2) \right|$$

All possible isoform pairs are then sorted by the S_1 score and in the case of a tie, by the S_2 score. Moreover, only pairs of isoforms from the same gene are considered and only a single pair of isoforms per gene is listed in a ranking of isoform-pairs. Classification rules are given in terms of k isoform pairs. The classification of a new sample is performed by evaluating each isoform-pair rule against the ranking of isoform expression of the new sample. For each isoform-pair rule, the classifier selects the class for which the data fulfills the rule. The final decision for classification is established by simple majority voting, by selecting the most voted class from the k rules. In order to avoid ties in the voting, k is always odd. Significance of the computed isoform changes was evaluated by shuffling labels from the two classes 1000 times. For each shuffling step, the Iso-kTSP algorithm was re-run and the top-scoring isoform-pair was selected. An isoform-pair is defined as significant if its Information Gain and Score S_1 are larger than any of the values obtained from the 1000 shufflings of the same cancer type. The Iso-kTSP is implemented in Java. Software and documentation are available at https://bitbucket.org/regulatorygenomicsupf/iso-ktsp.

For the purpose of finding associations, the number of samples for which a gene has a switch was compared with the number of samples for which the transcripts involved overlap mutations. Given the samples with one or more mutations M, and the samples with the isoform switch S, a Jaccard index J for the association of these two variables was calculated:

$$J = \frac{|M \cap S|}{|M \cup S|}$$

A Z-score from the Jaccard index was calculated by comparing each value J for an isoform switch to 100 genes with similar median isoform length. The above analysis was also repeated using only mutations that affect the protein sequence or considering the overlap with genes rather than transcript regions with similar results (see

Supplementary Methods). The mutual information for the association of isoform switches and mutations, and corresponding z-score was also computed (see Supplementary Methods). The distribution of the differences between tumor and normal isoform PSIs was compared between mutated and non-mutated samples using a Mann-Whitney test. A mutual-exclusion between isoform switches and protein-affecting mutations was measured as follows: given the number of samples having an isoform switch and no mutation (n_{10}), and those having a mutation but no isoform switch (n_{01}), a mutual-exclusion score was defined to be:

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

where N is the total number of samples. A z-score was calculated similarly as above (see Supplementary Material). For the Reactome pathway analysis the ReactomePA package from Bioconductor was used. Further details and data are provided as Supplementary Material.

Figure Legends

Figure 1. Methodology for detecting significant transcript isoform changes in cancer. The method is illustrated with data from colon adenocarcinoma (COAD). **(A)** Samples are partitioned into two groups, here tumor (T) and normal (N). **(B)** Cross-validation is performed by iteratively training in all but one pair of tumor-normal samples, and testing on this left-out sample pair. At each step of the cross-validation, the top k=1,3,5, etc. isoform-pairs according to score S_1 are tested on the left-out sample pair. **(C)** Significance is assessed by comparing to the expected values obtained from 1000 permutations of the class labels and by selecting at each permutation the isoform-pair with the highest score S_1. **(D)** The result is a global ranking of isoform-pairs that change relative expression between tumor and normal samples more than expected by change. **(E)** A minimal classification model is obtained by selecting the smallest number of pairs from the global ranking with the largest average accuracy calculated in the cross-validation. The example shows that for colon this is a single pair model in gene FBLN2. **(F)** From the global ranking of significant isoform-pairs, we predict as isoform switches those that anti-correlate

across samples. In the example, CD44 presents a clear switch between two isoforms even though it was not chosen in the minimal classification model. **(G)** The isoform-pairs (either from the minimal classification model or from the set of isoform switches) are tested on a held-out dataset of unpaired tumor samples.

Figure 2: Predictive isoform-pair models. (A) Minimal isoform-pair classifiers for BRCA, PRAD, LUAD and LUSC (models for KICH, KIRC, HNSC and THCA are given in Supplementary Figure 2). Each panel shows the score S_I and information gain (IG) for each isoform-pair in the model. All the isoform-pairs are significant according to the permutation analysis. Next to each cancer label the maximum expected accuracy is given, which is calculated from the cross-validation analysis. Plots of the expression values for each isoform pair are provided in Supplementary Figures 3-8. **(B)** Blind tests of the isoform-pair models on the unpaired samples for each cancer type. The barplots indicate the proportion of samples (y-axis) for each number of possible correct votes (x-axis), i.e., the number of isoform-pair rules from the model fulfilled by the tumor sample. A sample is labeled according to a majority vote from all isoform-pair rules.

Figure 3. Examples of predictive isoform-pairs. (A) Relative inclusion values (PSIs) for the isoform-pair in FBLN2 predicted to separate well tumor from normal in BRCA and COAD. The x-axis represents the inclusion level values (PSI) for the normal isoform and the y-axis the value for the tumor isoform. Tumor samples are shown in red, whereas normal samples are shown in blue. The left panel shows the PSIs in paired samples, whereas the middle and right panels show these for the unpaired samples **(B)** Isoform change for QKI in LUAD samples. The gene locus of QKI is shown, indicating the exon-intron structures of the isoforms predicted to be the tumor and normal isoforms for LUAD. The zoom-in highlights the 3' end region where the splicing variation takes place. The bottom left panel shows the PSI values for the normal (x-axis) and tumor isoforms (y-axis) for the normal (blue) and tumor (red) paired samples.

Figure 4. Isoform-pair rules for the basal-like breast tumors. **(A)** The top 7 recurrent isoform changes found in the 100 comparisons of subsets of basal-like samples against a balanced pool set of the other subtypes (luminal A, luminal B and Her2+). The barplots indicate the frequency of iterations for which the isoform pair was significant according to the permutation analysis performed on the same subsampled sets. All these isoform-pairs are significant according to our permutation analysis. **(B)** Accuracy of the model for the classification of basal-like samples against other subtypes when tested on the entire set of 1036 BRCA tumor samples. The barplots show the proportion of samples (y-axis) with each possible number of correct votes (x-axis), from 0 to the number of genes in the model, and the percentage of all the samples correctly classified.

Figure 5. Catalogue of isoform switches across various cancer types. Heatmap of the 244 isoform switches detected for the 9 cancer types, separated according to whether the genes had an annotated Reactome pathway **(A)** or not **(B)**. The heatmaps show whether the isoform switch occurs in each cancer type, with the color code indicating the information gain (IG) of the switch: from light blue for low IG (0-0.2) to dark purple for high IG (0.8-1). In red we indicate whether the gene with the switch is annotated in COSMIC (Forbes et al. 2011) as a tumor driver. Regarding the mutations, we indicate the Jaccard index and the mutual-exclusion score with light green (0.01-0.02), medium green (0.02-0.03) and dark green (larger than 0.03). We also indicate the presence of a significant difference (p-value < 0.05) of the relative inclusion (PSI) difference between tumor and normal isoforms in mutated and non-mutated tumor samples before multiple-testing correction (brown color). We further show the Reactome Pathway annotation for those genes for which this was available.

Figure 6. Association between somatic mutations and isoform switches. (A) Jaccard index (x-axis) for the association of mutations with switches in tumor samples, and frequency of samples with mutations in the transcripts undergoing the switch (y-axis). **(B)** Example of the tumor suppressor FBLN2. Mutations present in each cancer type are represented in red if the switch is present in the same sample, and blue if that sample does not have the switch. Each mutation is labeled with an identifier of the sample and the type of mutation. **(C)** Example of the oncogene MYH11. The relative

inclusion values (PSI) of the two isoforms in the switch (left panels) separate tumor and normal in COAD and can classify correctly 91.5% of the unpaired tumor samples. Mutations present in each cancer type (right panel) are represented in red if the switch is present in the same sample, and blue if that sample does not have the switch. Each mutation is labeled with an identifier of the sample and the type of mutation.

Acknowledgements

We would like to thank R. Karni, K. Hertel, Q. Morris and R. Castelo, for useful discussions. This work was supported by grants BIO2011-23920 and Consolider RNAREG (CSD2009-00080) from the Ministerio de Ciencia e Innovación of Spain and by the Sandra Ibarra Foundation for Cancer (FSI2013).

References

Amin EM et al (2011) WT1 mutants reveal SRPK1 to be a downstream angiogenesis target by altering VEGF splicing. Cancer Cell 20(6):768–780

Baird BN, Schliekelman MJ, Ahn YH, Chen Y, Roybal JD, Gill BJ, Mishra DK, Erez B, et al. (2013) Fibulin-2 is a driver of malignant progression in lung adenocarcinoma. PLoS One. 8(6):e67054.

Bechara EG, Sebestyén E, Bernardis I, Eyras E, Valcárcel J. RBM5, 6, and 10 Differentially Regulate NUMB Alternative Splicing to Control Cancer Cell Proliferation. Mol Cell. 2013 Dec 12;52(5):720-33.

Bonomi S, Gallo S, Catillo M, Pignataro D, Biamonti G, Ghigna C. Oncogenic alternative splicing switches: role in cancer progression and prospects for therapy. Int J Cell Biol. 2013;2013:962038.

Bourdon JC, Fernandes K, Murray-Zmijewski F, Liu G, Diot A, Xirodimas DP, Saville MK, Lane DP. p53 isoforms can regulate p53 transcriptional activity. Genes Dev. 2005 Sep 15;19(18):2122-37.

Cao, Jiang, et al. "Characterization of colorectal-cancer-related cDNA clones obtained by subtractive hybridization screening." *Journal of cancer research and clinical oncology* 123.8 (1997): 447-451.

Cariati, M., Naderi, A., Brown, J. P., Smalley, M. J., Pinder, S. E., Caldas, C., & Purushotham, A. D. (2008). Alpha-6 integrin is necessary for the tumourigenicity of a stem cell-like subpopulation within the MCF7 breast cancer cell line.*International* Journal of Cancer, 122(2), 298-304.

Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, et al. (2014) The Reactome pathway knowledgebase. Nucleic Acids Research 142(D1):D472-D477

Das S, Raj L, Zhao B, Kimura Y, Bernstein A, Aaronson SA, Lee SW. Hzf Determines cell survival upon genotoxic stress by modulating p53 transactivation. Cell. 2007 Aug 24;130(4):624-37.

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008 Oct 23;455(7216):1069-75.

Du L, Wang H, He L, Zhang J, Ni B, Wang X, Jin H, Cahuzac N, et al. (2008) CD44 is of functional importance for colorectal cancer stem cells. Clin Cancer Res. 14(21):6751-60.

Ellis L, Atadja PW, Johnstone RW. (2008) Epigenetics in cancer: targeting chromatin modifications. Mol Cancer Ther. 8(6):1409-20.

Esteller M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet. 8(4):286-98.

Eswaran J, Horvath A, Godbole S, Reddy SD, Mudvari P, Ohshiro K, Cyanam D, Nair S, Fuqua SA, Polyak K, Florea LD, Kumar R. RNA sequencing of cancer reveals novel splicing alterations. Sci Rep. 2013;3:1689.

Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, et al. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research 39(suppl 1): D945-D950

Furney SJ, Pedersen M, Gentien D, Dumont AG, Rapinat A, Desjardins L, Turajlic S, Piperno-Neumann S, de la Grange P, Roman-Roman S, Stern MH, Marais R. SF3B1 mutations are associated with alternative splicing in uveal melanoma. Cancer Discov. 2013 Oct;3(10):1122-9.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nature Reviews Cancer. 2004 Mar 1;4(3):177–83.

Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. Stat Appl Genet Mol Biol. 2004;3:Article19.

Häberle J, Denecke J, Schmidt E, Koch HG. Diagnosis of N-acetylglutamate synthase deficiency by use of cultured fibroblasts and avoidance of nonsense-mediated mRNA decay. J Inherit Metab Dis. 2003;26(6):601-5.

Hancox RA, Allen MD, Holliday DL, Edwards DR, Pennington CJ, Guttery DS, Shaw JA, Walker RA, et al. (2009) Tumour-associated tenascin-C isoforms promote breast cancer cell invasion and growth by matrix metalloproteinase-dependent and independent mechanisms. Breast Cancer Res. 11(2):R24.

Hill VK, Hesson LB, Dansranjavin T, Dallol A, Bieche I, Vacher S, Tommasi S, Dobbins T, et al. Identification of 5 novel genes methylated in breast and other epithelial cancers. Mol Cancer. 2010 Mar 5;9:51.

Hudson (2013). Genome variation and personalized cancer medicine. J Intern Med. 274(5):440-50.

Izquierdo JM, Majós N, Bonnal S, Martínez C, Castelo R, Guigó R, Bilbao D, Valcárcel J. Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. Mol Cell. 2005 Aug 19;19(4):475-84.

Kalari KR, Rossell D, Necela BM, Asmann YW, Nair A, Baheti S, Kachergus JM, Younkin CS, et al. Deep Sequence Analysis of Non-Small Cell Lung Cancer: Integrated Analysis of Gene Expression, Alternative Splicing, and Single Nucleotide Variations in Lung Adenocarcinomas with and without Oncogenic KRAS Mutations. Front Oncol. 2012 Feb 10;2:12.

Karni R, de Stanchina E, Lowe SW, Sinha R, Mu D, Krainer AR. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nat Struct Mol Biol. 2007 Mar;14(3):185-93.

Klinck R, Bramard A, Inkel L, Dufresne-Martin G, Gervais-Bird J, Madden R, Paquet ER, Koh C, Venables JP, Prinos P, Jilaveanu-Pelmus M, Wellinger R, Rancourt C, Chabot B, Abou Elela S. Multiple alternative splicing markers for ovarian cancer. Cancer Res. 2008 Feb 1;68(3):657-63.

Lapuk A, Marr H, Jakkula L, Pedro H, Bhattacharya S, Purdom E, Hu Z, Simpson K, et al. Exon-level microarray analyses identify alternative splicing programs in breast cancer. Mol Cancer Res. 2010 Jul;8(7):961-74.

Law EW, Cheung AK, Kashuba VI, Pavlova TV, Zabarovsky ER, Lung HL, Cheng Y, Chua D, et al. (2012) Anti-angiogenic and tumor-suppressive roles of candidate tumor-suppressor gene, Fibulin-2, in nasopharyngeal carcinoma. Oncogene 31(6):728-38.

Lee YS, Krishnan A, Zhu Q, Troyanskaya OG. (2013) Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. Bioinformatics 29(23):3036-44.

Li B, Ruotti V, Stewart R, Thomson J, Dewey C. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 26(4):493-500.

Liu J, Lee W, Jiang Z, Chen Z, Jhunjhunwala S, Haverty PM, Gnad F, Guan Y et al. Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events. Genome Res. 2012 Dec;22(12):2315-27.

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. (2010) Regulation of alternative splicing by histone modifications. Science 327(5968):996-1000.

Lundin A, Driscoll B. Lung cancer stem cells: progress and prospects. (2013) Cancer Lett. 338(1):89-93.

Mallinjoud P, Villemin JP, Mortada H, Polay Espinoza M, Desmet FO, Samaan S, Chautard E, Tranchevent LC, Auboeuf D. Endothelial, epithelial, and fibroblast cells exhibit specific splicing programs independently of their tissue of origin. Genome Res. 2014 Mar;24(3):511-21.

Marchionni, L., Afsari, B., Geman, D., & Leek, J. T. (2013). A simple and reproducible breast cancer prognostic test. *BMC genomics*, *14*(1), 336.

Massi D, Landriscina M, Piscazzi A, Cosci E, Kirov A, Paglierani M, Di Serio C, Mourmouras V, et al. (2010) S100A13 is a new angiogenic marker in human melanoma. Mod Pathol. 23(6):804-13.

Maunakea AK, Chepelev I, Cui K, Zhao K. (2013). Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. Cell Res. 23(11):1256-69.

Misquitta-Ali CM, Cheng E, O'Hanlon D, Liu N, McGlade CJ, Tsao MS, Blencowe BJ. Global profiling and molecular characterization of alternative splicing events misregulated in lung cancer. Mol Cell Biol. 2011 Jan;31(1):138-50.

Mitra D, Brumlik MJ, Okamgba SU, Zhu Y, Duplessis TT, Parvani JG, Lesko SM, Brogi E, Jones FE. An oncogenic isoform of HER2 associated with locally disseminated breast cancer and trastuzumab resistance. Mol Cancer Ther. 2009 Aug;8(8):2152-62.

Miura K, Fujibuchi W, Unno M. Splice isoforms as therapeutic targets for colorectal cancer. Carcinogenesis. 2012 Dec;33(12):2311-9.

Pal S, Bi Y, Macyszyn L, Showe LC, O'Rourke DM, Davuluri RV. (2014) Isoform-level gene signature improves prognostic stratification and accurately classifies glioblastoma subtypes. Nucleic Acids Res. 42(8):e64.

Poulikakos PI, Persaud Y, Janakiraman M, Kong X, Ng C, Moriceau G, Shi H, Atefi M, et al. (2011) RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). Nature 480(7377):387-90

Price ND, Trent J, El-Naggar AK, Cogdell D, Taylor E, Hunt KK, Pollock RE, Hood L, Shmulevich I, Zhang W. Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas. Proc Natl Acad Sci U S A. 2007 Feb 27;104(9):3414-9.

Schlisio S, Kenchappa RS, Vredeveld LC, George RE, Stewart R, Greulich H, Shahriari K, Nguyen NV, et al. The kinesin KIF1Bbeta acts downstream from EglN3 to induce apoptosis and is a potential 1p36 tumor suppressor. Genes Dev. 2008 Apr 1;22(7):884-93.

Sterne-Weiler T, Sanford JR. (2014) Exon identity crisis: disease-causing mutations that disrupt the splicing code. Genome Biol. 15(1):201.

Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, Kryukov GV, Lawrence MS, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011 Aug 26;333(6046):1157-60.

Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. Cell. 2014 Mar 13;156(6):1324-35.

Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics. 2005 Oct 15;21(20):3896-904.

Thorsen K, Mansilla F, Schepeler T, Øster B, Rasmussen MH, Dyrskjøt L, Karni R, Akerman M, et al. Alternative splicing of SLC39A14 in colorectal cancer is regulated by the Wnt pathway. Mol Cell Proteomics. 2011 Jan;10(1):M110.002998

Thorsen K, Sørensen KD, Brems-Eskildsen AS, Modin C, Gaustadnes M, Hein AM, Kruhøffer M, Laurberg S, et al. Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. Mol Cell Proteomics. 2008 Jul;7(7):1214-24.

TCGA. The Cancer Genome Atlas Network (2012a). Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012 Jul 18;487(7407):330-7.

TCGA. The Cancer Genome Atlas Research Network (2012b). Comprehensive genomic characterization of squamous cell lung cancers. Nature 489(7417):519-25.

TCGA. The Cancer Genome Atlas Network (2012c) Comprehensive molecular portraits of human breast tumours. *Nature.* 490(7418):61-70

TCGA. The Cancer Genome Atlas Research Network (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature 499(7456):43-9.

Venables JP, Brosseau JP, Gadea G, Klinck R, Prinos P, Beaulieu JF, Lapointe E, Durand M, et al. (2013). RBFOX2 is an important regulator of mesenchymal tissue-specific splicing in both normal and cancer tissues. Mol Cell Biol. 33(2):396-405.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013 Mar 29;339(6127):1546-58.

Wilkerson MD, Yin X, Hoadley KA, Liu Y, Hayward MC, Cabanski CR, Muldrew K, Miller CR, et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. Clin Cancer Res. 2010 Oct 1;16(19):4864-75.

Ward, A. and Cooper,T. (2010) The pathobiology of splicing. J. Pathol., 220, 152–163.

Yanagisawa, M., Huveldt, D., Kreinest, P., Lohse, C. M., Cheville, J. C., Parker, A. S., Anastasiadis, P. Z. (2008). A p120 catenin isoform switch affects Rho activity, induces tumor cell invasion, and predicts metastatic disease. *Journal of biological chemistry*, *283*(26), 18344-18354.

Zhang H, Meng F, Wu S, Kreike B, Sethi S, Chen W, Miller FR, Wu G. (2011) Engagement of I-branching β-1, 6-N-acetylglucosaminyltransferase 2 in breast cancer metastasis and TGF-β signaling. *Cancer research*, *71*(14), 4846-4856.

Zhang Z, Pal S, Bi Y, Tchou J, Davuluri RV. (2013) Isoform level expression profiles provide better cancer signatures than gene level expression profiles. Genome Med. 5(4):33.

Figure 1

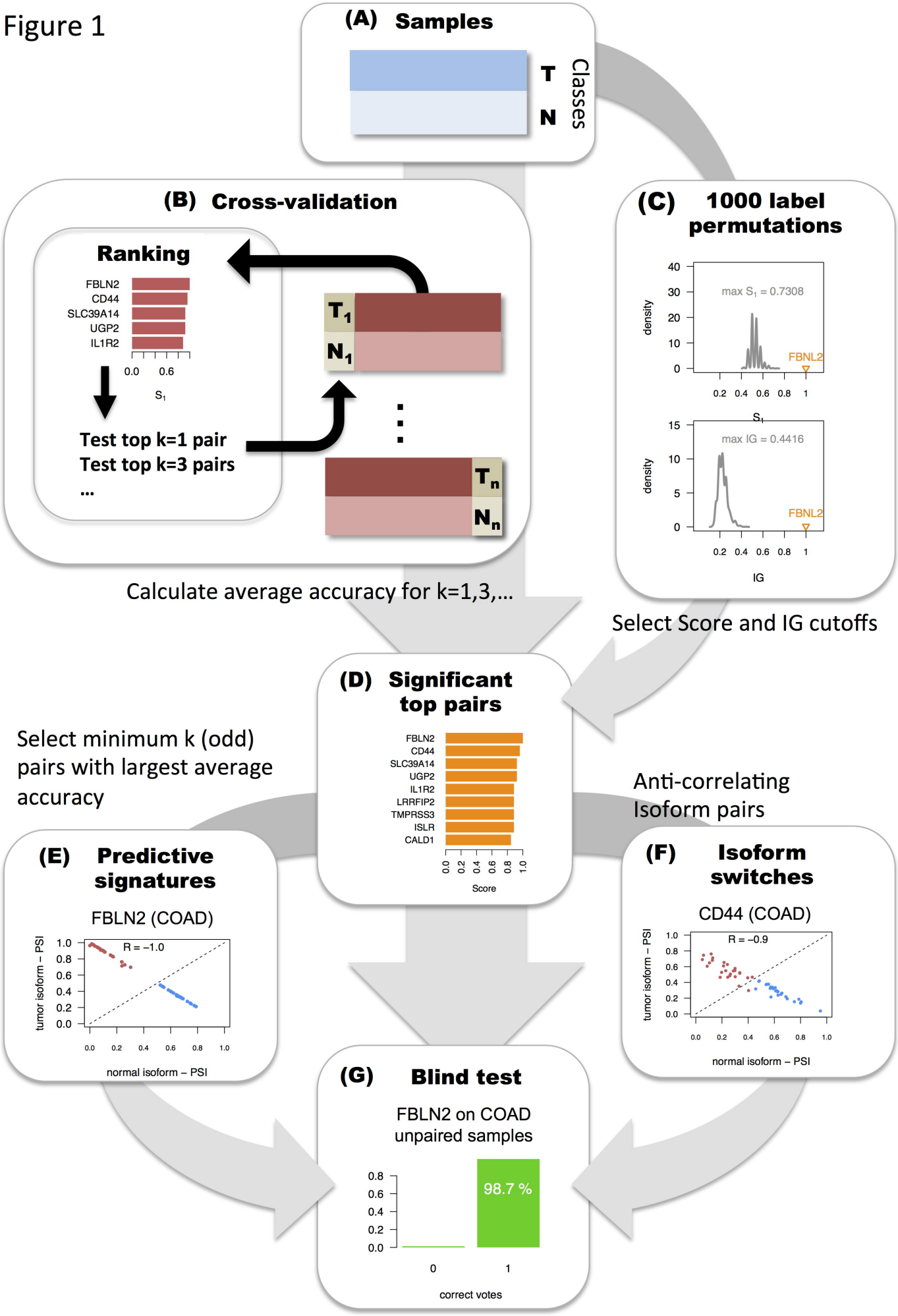
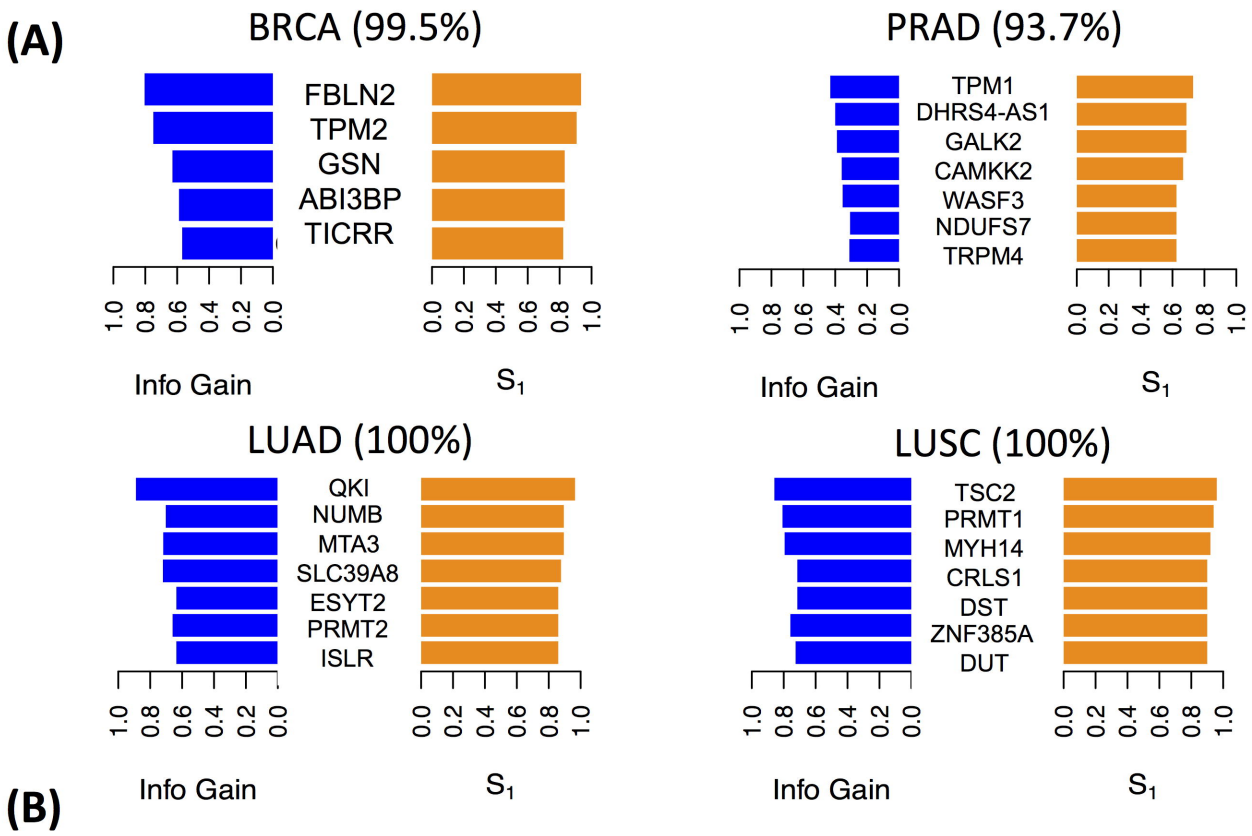

(A) Samples
T
N
Classes
(B) Cross-validation
Ranking
FBLN2
CD44
SLC39A14
UGP2
IL1R2
0.0
0.6
S1
Test top k=1 pair
Test top k=3 pairs
...
T1
N1
Tn
Nn
(C) 1000 label permutations
max S1 = 0.7308
FBNL2
density
max IG = 0.4416
IG
Calculate average accuracy for k=1,3,...
Select Score and IG cutoffs
(D) Significant top pairs
FBLN2
CD44
SLC39A14
UGP2
IL1R2
LRRFIP2
TMPRSS3
ISLR
CALD1
Score
Select minimum k (odd) pairs with largest average accuracy
Anti-correlating Isoform pairs
(E) Predictive signatures
FBLN2 (COAD)
R = −1.0
tumor isoform – PSI
normal isoform – PSI
(F) Isoform switches
CD44 (COAD)
R = −0.9
tumor isoform – PSI
normal isoform – PSI
(G) Blind test
FBLN2 on COAD unpaired samples
98.7 %
correct votes

Figure 2


(A)
BRCA (99.5%)
FBLN2
TPM2
GSN
ABI3BP
TICRR
Info Gain
S1
PRAD (93.7%)
TPM1
DHRS4-AS1
GALK2
CAMKK2
WASF3
NDUFS7
TRPM4
Info Gain
S1
LUAD (100%)
QKI
NUMB
MTA3
SLC39A8
ESYT2
PRMT2
ISLR
Info Gain
S1
LUSC (100%)
TSC2
PRMT1
MYH14
CRLS1
DST
ZNF385A
DUT
Info Gain
S1

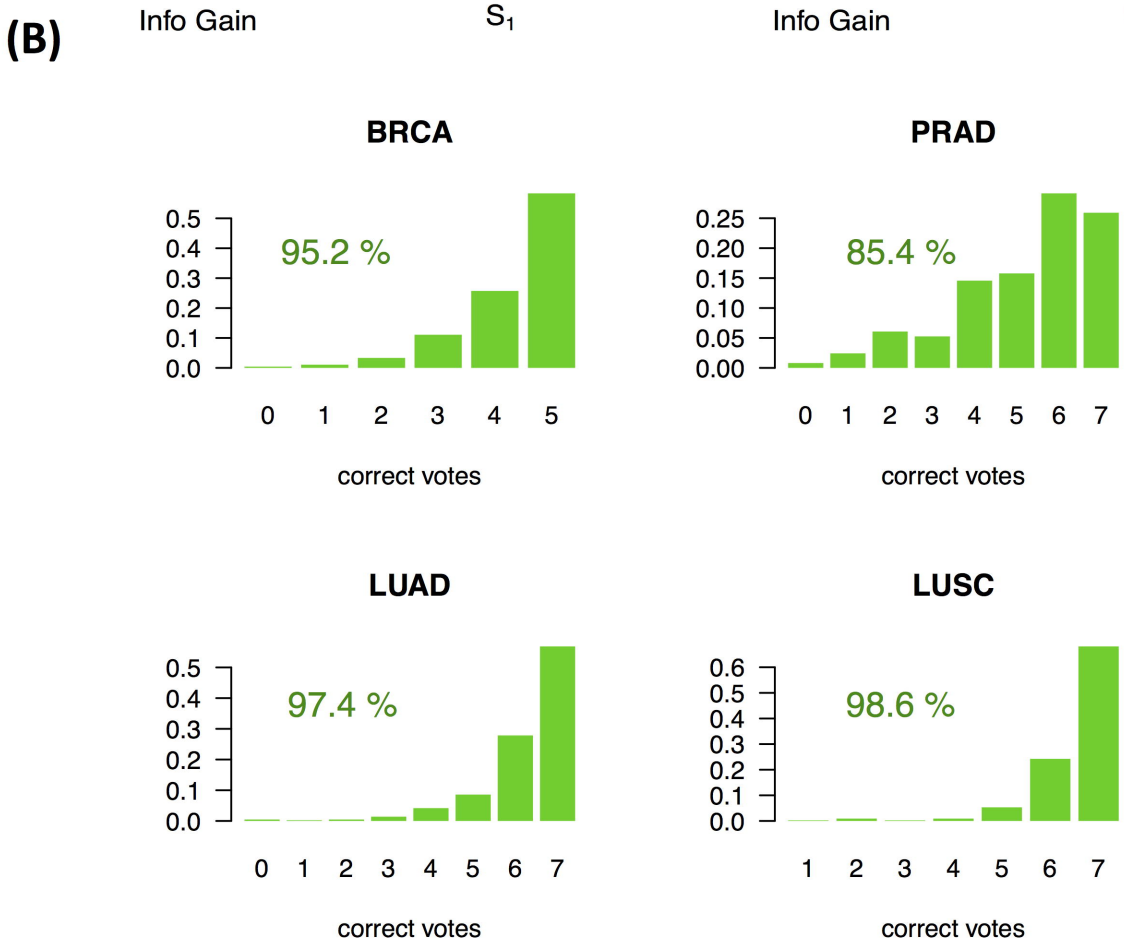

(B)
BRCA
95.2 %
correct votes
PRAD
85.4 %
correct votes
LUAD
97.4 %
correct votes
LUSC
98.6 %
correct votes

Figure 3

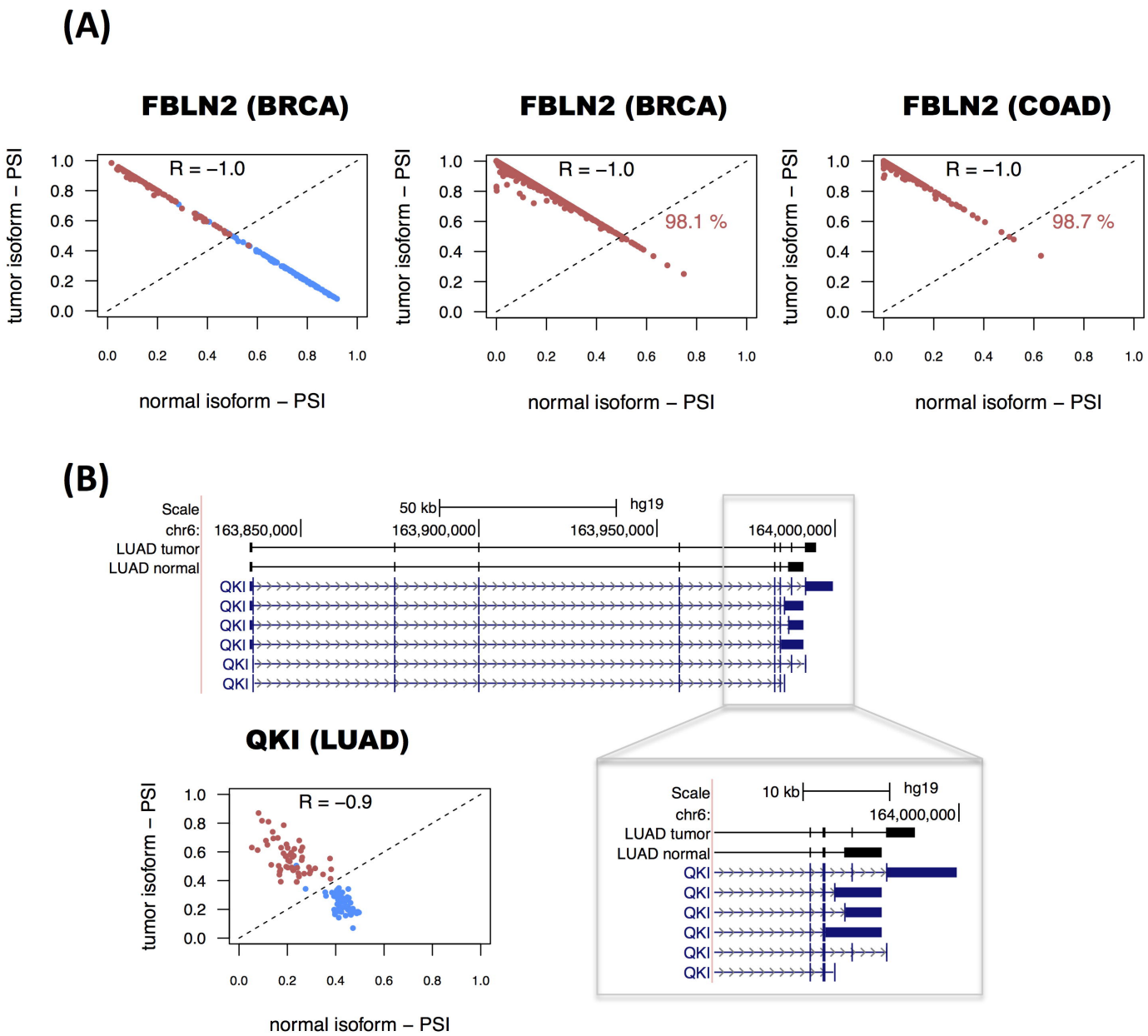

(A)
FBLN2 (BRCA)
R = −1.0
FBLN2 (BRCA)
R = −1.0
98.1 %
FBLN2 (COAD)
R = −1.0
98.7 %
tumor isoform – PSI
normal isoform – PSI
(B)
Scale
chr6:
50 kb
hg19
163,850,000
163,900,000
163,950,000
164,000,000
LUAD tumor
LUAD normal
QKI
QKI (LUAD)
R = −0.9
10 kb
164,000,000

Figure 4

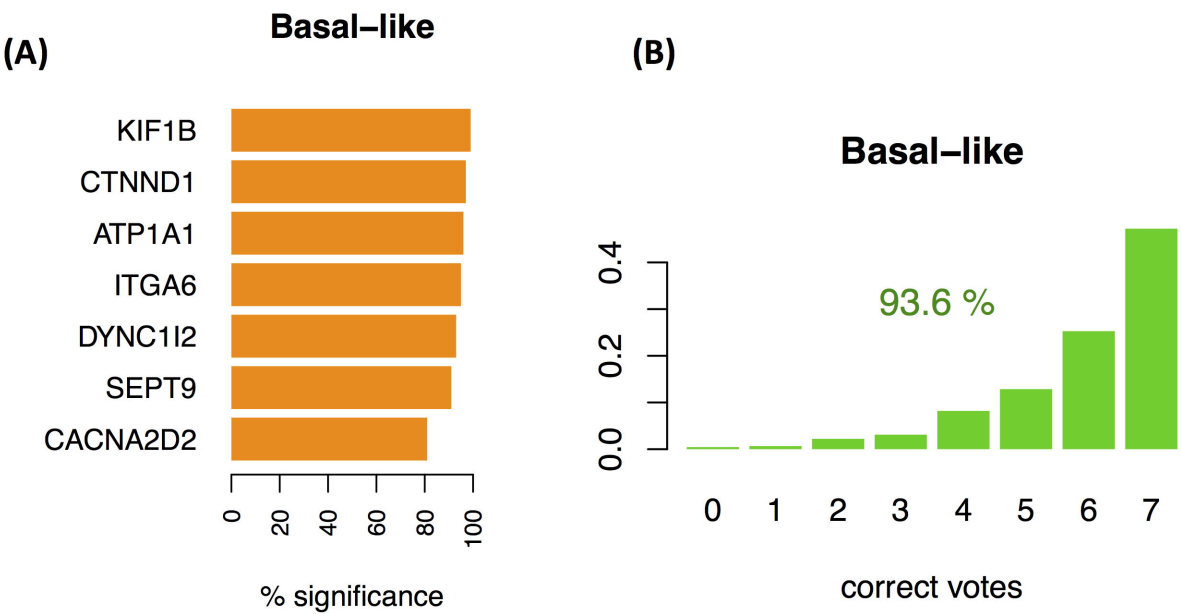
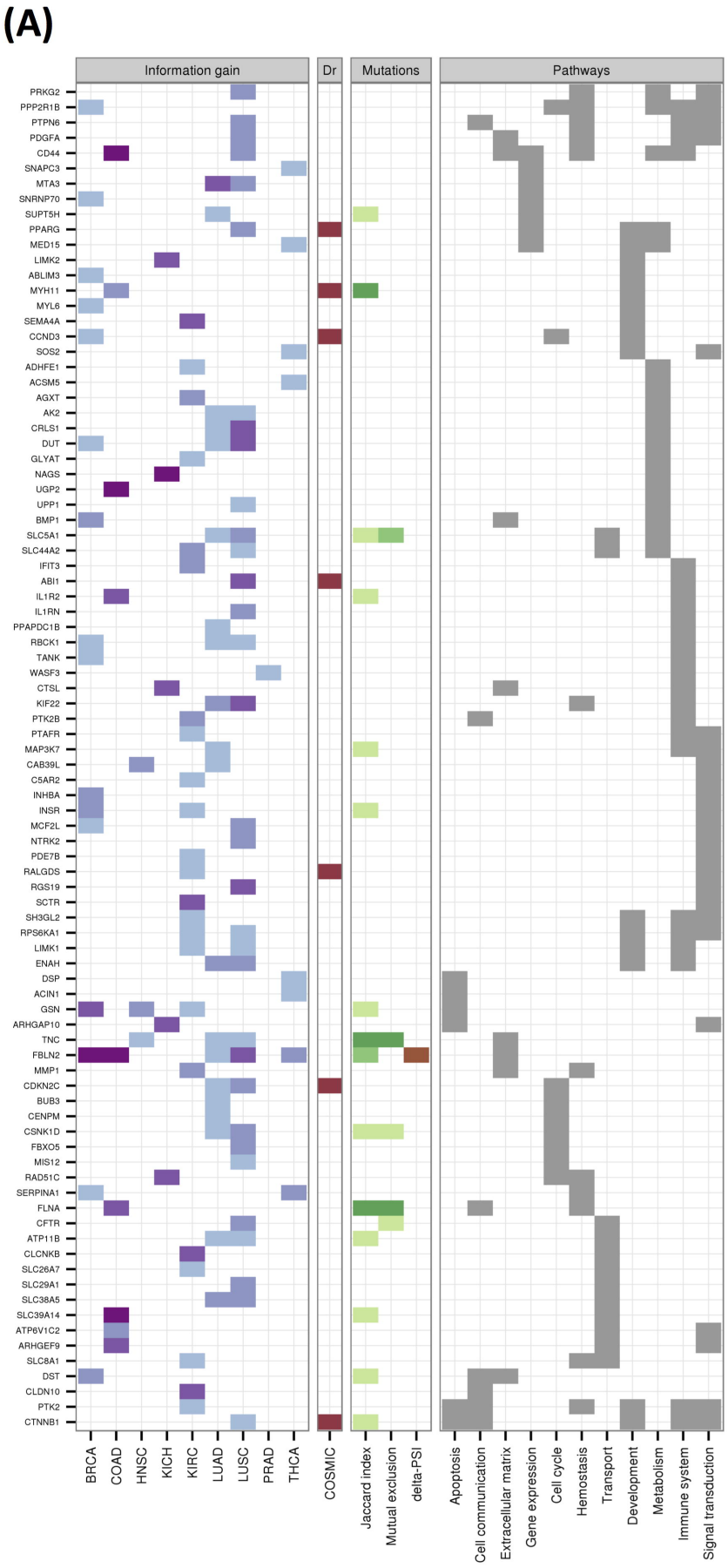

(A)
Basal–like
KIF1B
CTNND1
ATP1A1
ITGA6
DYNC1I2
SEPT9
CACNA2D2
0
20
40
60
80
100
% significance
(B)
Basal–like
93.6 %
0.0
0.2
0.4
0
1
2
3
4
5
6
7
correct votes

Figure 5


(A)
Information gain
Dr
Mutations
Pathways
PRKG2
PPP2R1B
PTPN6
PDGFA
CD44
SNAPC3
MTA3
SNRNP70
SUPT5H
PPARG
MED15
LIMK2
ABLIM3
MYH11
MYL6
SEMA4A
CCND3
SOS2
ADHFE1
ACSM5
AGXT
AK2
CRLS1
DUT
GLYAT
NAGS
UGP2
UPP1
BMP1
SLC5A1
SLC44A2
IFIT3
ABI1
IL1R2
IL1RN
PPAPDC1B
RBCK1
TANK
WASF3
CTSL
KIF22
PTK2B
PTAFR
MAP3K7
CAB39L
C5AR2
INHBA
INSR
MCF2L
NTRK2
PDE7B
RALGDS
RGS19
SCTR
SH3GL2
RPS6KA1
LIMK1
ENAH
DSP
ACIN1
GSN
ARHGAP10
TNC
FBLN2
MMP1
CDKN2C
BUB3
CENPM
CSNK1D
FBXO5
MIS12
RAD51C
SERPINA1
FLNA
CFTR
ATP11B
CLCNKB
SLC26A7
SLC29A1
SLC38A5
SLC39A14
ATP6V1C2
ARHGEF9
SLC8A1
DST
CLDN10
PTK2
CTNNB1
BRCA
COAD
HNSC
KICH
KIRC
LUAD
LUSC
PRAD
THCA
COSMIC
Jaccard index
Mutual exclusion
delta-PSI
Apoptosis
Cell communication
Extracellular matrix
Gene expression
Cell cycle
Hemostasis
Transport
Development
Metabolism
Immune system
Signal transduction

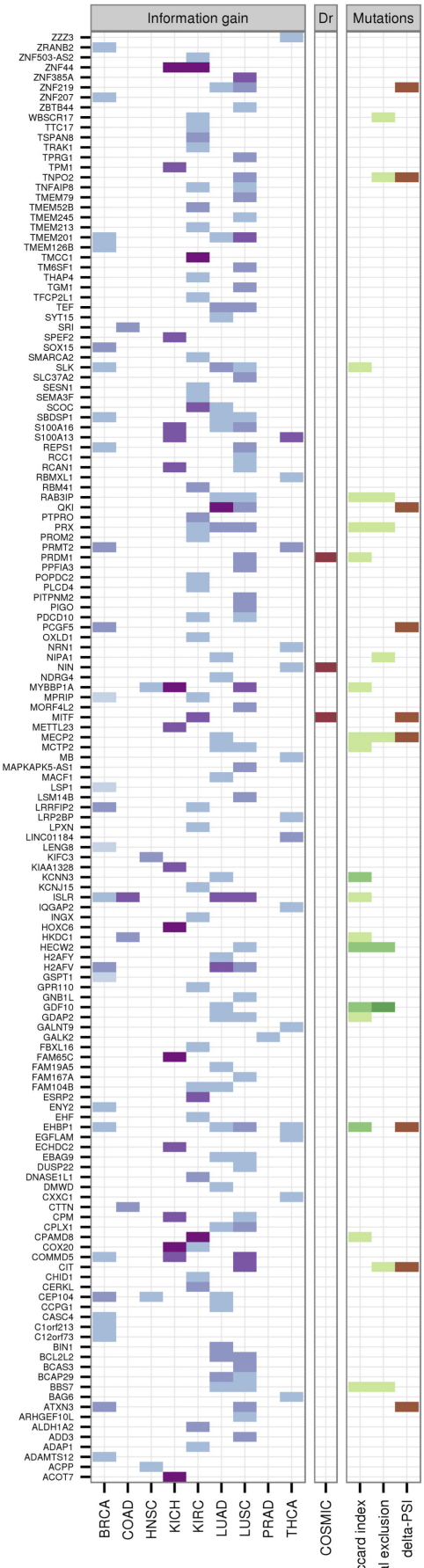
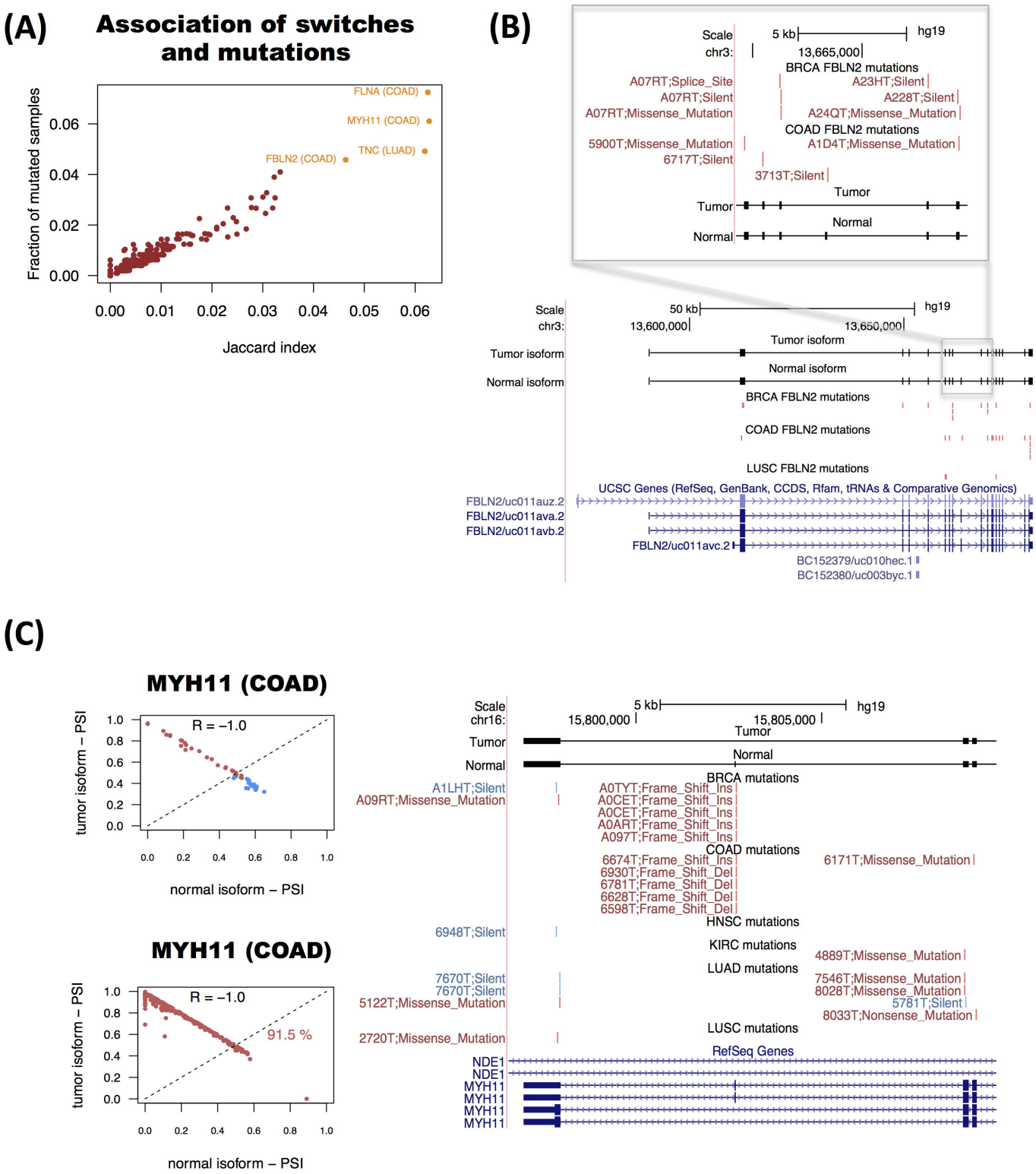

(B)
Information gain
Dr
Mutations
ZZZ3
ZRANB2
ZNF503-AS2
ZNF44
ZNF385A
ZNF219
ZNF207
ZBTB44
WBSCR17
TTC17
TSPAN8
TRAK1
TPRG1
TPM1
TNPO2
TNFAIP8
TMEM79
TMEM52B
TMEM245
TMEM213
TMEM201
TMEM126B
TMCC1
TM6SF1
THAP4
TGM1
TFCP2L1
TEF
SYT15
SRI
SPEF2
SOX15
SMARCA2
SLK
SLC37A2
SESN1
SEMA3F
SCOC
SBDSP1
S100A16
S100A13
REPS1
RCC1
RCAN1
RBMXL1
RBM41
RAB3IP
OKI
PTPRO
PRX
PROM2
PRMT2
PRDM1
PPFIA3
POPDC2
PLCD4
PITPNM2
PIGO
PDCD10
PCGF5
OXLD1
NRN1
NIPA1
NIN
NDRG4
MYBBP1A
MPRIP
MORF4L2
MITF
METTL23
MECP2
MCTP2
MB
MAPKAPK5-AS1
MACF1
LSP1
LSM14B
LRRFIP2
LRP2BP
LPXN
LINC01184
LENG8
KIFC3
KIAA1328
KCNN3
KCNJ15
ISLR
IQGAP2
INGX
HOXC6
HKDC1
HECW2
H2AFY
H2AFV
GSPT1
GPR110
GNB1L
GDF10
GDAP2
GALNT9
GALK2
FBXL16
FAM65C
FAM19A5
FAM167A
FAM104B
ESRP2
ENY2
EHF
EHBP1
EGFLAM
ECHDC2
EBAG9
DUSP22
DNASE1L1
DMWD
CXXC1
CTTN
CPM
CPLX1
CPAMD8
COX20
COMMD5
CIT
CHID1
CERKL
CEP104
CCPG1
CASC4
C1orf213
C12orf73
BIN1
BCL2L2
BCAS3
BCAP29
BBS7
BAG6
ATXN3
ARHGEF10L
ALDH1A2
ADD3
ADAP1
ADAMTS12
ACPP
ACOT7
BRCA
COAD
HNSC
KICH
KIRC
LUAD
LUSC
PRAD
THCA
COSMIC
Jaccard index
Mutual exclusion
delta-PSI

Figure 6


(A)
Association of switches and mutations
Fraction of mutated samples
Jaccard index
FLNA (COAD)
MYH11 (COAD)
TNC (LUAD)
FBLN2 (COAD)
(B)
Scale
chr3:
5 kb
hg19
13,665,000
BRCA FBLN2 mutations
A07RT;Splice_Site
A07RT;Silent
A07RT;Missense_Mutation
A23HT;Silent
A228T;Silent
A24QT;Missense_Mutation
COAD FBLN2 mutations
5900T;Missense_Mutation
6717T;Silent
A1D4T;Missense_Mutation
3713T;Silent
Tumor
Normal
50 kb
13,600,000
13,650,000
Tumor isoform
Normal isoform
BRCA FBLN2 mutations
COAD FBLN2 mutations
LUSC FBLN2 mutations
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
FBLN2/uc011auz.2
FBLN2/uc011ava.2
FBLN2/uc011avb.2
FBLN2/uc011avc.2
BC152379/uc010hec.1
BC152380/uc003byc.1
(C)
MYH11 (COAD)
R = −1.0
tumor isoform − PSI
normal isoform − PSI
MYH11 (COAD)
R = −1.0
91.5 %
chr16:
15,800,000
15,805,000
A1LHT;Silent
A09RT;Missense_Mutation
BRCA mutations
A0TYT;Frame_Shift_Ins
A0CET;Frame_Shift_Ins
A0CET;Frame_Shift_Ins
A0ART;Frame_Shift_Ins
A097T;Frame_Shift_Ins
COAD mutations
6674T;Frame_Shift_Ins
6930T;Frame_Shift_Del
6781T;Frame_Shift_Del
6628T;Frame_Shift_Del
6598T;Frame_Shift_Del
6171T;Missense_Mutation
6948T;Silent
HNSC mutations
KIRC mutations
4889T;Missense_Mutation
LUAD mutations
7670T;Silent
7670T;Silent
5122T;Missense_Mutation
7546T;Missense_Mutation
8028T;Missense_Mutation
5781T;Silent
8033T;Nonsense_Mutation
2720T;Missense_Mutation
LUSC mutations
RefSeq Genes
NDE1
NDE1
MYH11
MYH11
MYH11
MYH11