Title: The Landscape of Isoform Switches in Human Cancers

Authors: Kristoffer Vitting-Seerup[1,*], Albin Sandelin[1,*]

* Corresponding authors

Affiliations: [1]The Bioinformatics Centre, Department of Biology and Biotech Research & Innovation Centre, University of Copenhagen, Denmark.

Running title: Isoform Switches in Cancer

Keywords: RNA isoform switches, pan-cancer analysis, Molecular diagnosis and prognosis, Mechanisms of transcription, bioinformatics tools

Corresponding authors

Kristoffer Vitting-Seerup:
k.vitting.seerup@gmail.com
Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark
Phone: +45 35330235

Albin Sandelin:
albin@binf.ku.dk
Ole Maaløes Vej 5, DK-2200 Copenhagen N, Denmark
+45 35321285

Conflict of interest disclosure statement:
The authors declare no potential conflicts of interest

Financial support:
This study was supported by grants from the Novo Nordisk Foundation, the Lundbeck Foundation, and Elixir Denmark to Albin Sandelin.

Manuscript statistics:

Word count (excl. abstract): 6,282
Number figures: 6
Number tables: 1
Supplementary data: 6 figures, 5 tables, 3 texts, 1 file.

ABSTRACT

Alternative usage of transcript isoforms from the same gene has been hypothesized as an important feature in cancers. However, differential usage of gene transcripts between conditions - isoform switching - has not been comprehensively characterized in and across cancer types.

To this end, we developed methods for identification and visualization of isoform switches with predicted functional consequences. Using these methods, we characterized isoform switching in RNA-seq data from >5500 cancer patients covering 12 solid cancer types. Isoform switches with potential functional consequences were common, affecting ~19% of multiple-transcript genes. Amongst these, isoform switches leading to loss of DNA sequence encoding protein domains were more frequent than expected, particularly in pan-cancer switches. We identified several isoform switches as powerful biomarkers: 31 switches were highly predictive of patient survival independent of cancer types.

Our data constitute an important resource for cancer researchers, available through interactive web tools. Moreover, our methods, available as an R package, enable systematic analysis of isoform switches from other RNA-seq data sets.

Implications

This study indicates that isoform switches with predicted functional consequences are common and important in dysfunctional cells, which in turn means that gene expression should be analyzed on the isoform level.

INTRODUCTION

The ability to produce different transcripts – gene isoforms – through alternative splicing (AS), alternative transcription start sites (aTSS) and alternative transcription termination sites (aTTS) is a major determinant of the increased complexity of higher vertebrates (1). The large majority of human genes uses alternative isoforms: ~95% of multi-exon genes show evidence of AS (2) and ~60% of genes have at least one aTSS (3). Recently, the ENCODE project estimated that, on average, each gene has 6.3 isoforms (3.9 different protein-coding isoforms) (4). It is therefore no surprise that gene isoform usage has an important role in many biological processes, including development, homeostasis, pluripotency and apoptosis (5–9). Moreover, isoforms are often tissue-specific and may alter the function, cellular localization and stability of the corresponding RNA or protein (10,11).

Differential usage of isoforms in different conditions - often referred to as isoform switching (Fig 1A) - can have substantial biological impact, caused by the difference in the functional potential of the two isoforms. Isoform switches are implicated in many diseases (11) and are especially prominent in cancer (12). A well-described example is the isoform switch in the *ALK* gene, occurring in 11% of melanoma patients, caused by the differential usage of aTSSs. The switch results in the production of a truncated protein lacking extracellular domains (13), which in turn promotes cell proliferation and drives tumorigenesis in vitro (13). Many such examples exist in the literature and include genes central to all eight cancer hallmarks (reviewed in (14,15)). This has resulted in an increasing interest in targeting both general and specific splicing events for therapeutic purposes (16,17).

The combination of mature methods for genome-wide RNA-sequencing (RNA-seq) and availability of advanced tools for the analysis of the resulting short DNA reads (e.g. Cufflinks (18) or Kallisto (19)) has enabled the quantification of transcriptomes with isoform resolution. This has enabled the genome-wide analysis of isoform usage and thereby the identification of isoform switches (20). Relative isoform usage can be defined as the fraction

of gene expression originating from each associated isoform. Here we define these fractions as Isoform Fractions (IFs) (Fig. 1B, top panel). It follows that changes in the usage of individual isoforms can be quantified as the difference in IF values between conditions (dIF values, Fig. 1B, bottom panel), and by extracting isoforms with opposite dIF values (one going up, one going down), isoform switches may be identified. The validity of such switches can be evaluated with statistical tests across samples (Fig. 1C). For simplicity, we will here refer to isoforms with increased or decreased dIF values respectively as upregulated or downregulated.

RNA-seq and related genome-wide RNA profiling methods have been cornerstone for consortia projects profiling the transcriptional states of tissues, cells and disease states (4,7,21,22). Of particular interest in this context is The Cancer Genome Atlas (TCGA)(7). TCGA has produced RNA-seq data from thousands of cancer patients covering a wide range of cancer types, quantified at both gene and isoform resolution, with the ultimate goal of profiling transcriptional states of different cancer types.

The availability of data sets of this type has spurred the development of statistical tools and methods for finding genes with isoform switches, such as Cufflinks2/Cuffdiff2, IUTA, rSeqNP and DRIMSeq (23–26). In particular, isoform switches in cancer have been analyzed by several groups (27–29). Of special interest is the study by Sebestyen *et al* (27), which analyzed TCGA data sets from 9 cancer types and identified 244 genes with isoform switches, of which 59 were found in multiple cancer types (pan-cancer switches). However, this and other studies have only analyzed isoform switching descriptively, and not estimated the potential impact of switches. Functional consequences of switches such as gain or loss of protein domains or signal peptides, loss of protein coding sequence and changes in propensity for Nonsense Mediated Decay (NMD), may all be computationally predicted once the specific isoforms involved in a switch are detected (30–32).

Analyses of RNA-seq data utilizing isoform level information remain rare, even though RNA-seq is now a standard technique and new RNA-seq data sets are

produced on a daily basis. To illustrate this, we examined 100 randomly selected articles from the first 5 months of 2016 analyzing RNA-seq data sets: only 11% of these articles performed analysis at the isoform. Thus, RNA-seq data is clearly under-utilized. We believe this is partially due to the lack of dedicated computational tools for the post-analysis of RNA-seq data at isoform resolution. Specifically, there are to our knowledge, no tools enabling: i) Statistical identification of the specific isoform involved in an isoform switch. ii) Prediction of the potential effects of an isoform switch. iii) Straightforward visualization of isoform switches and their potential consequences.

Here, we describe a set of methods for the identification of isoform switches and the subsequent integration of multiple predictors of isoform function, in order to identify isoform switches with predicted functional consequences. We applied these methods to TCGA RNA-seq data sets from more than 5,500 cancer patients covering 12 solid cancer types, making the most comprehensive analysis of isoform switches in cancer to date. We found that isoform switches with predicted functional consequences were common, affecting ~19% (*N*=2,352) of multi-isoform genes. Amongst these, switches leading to the loss of sequence encoding protein domains were frequent, particularly in pan-cancer isoform switches. We also identified a set of 31 pan-cancer switches as highly predictive of patient survival, independent of cancer types. Our data is can be easily explored through interactive online visualization tools and our methods are implemented as an easy-to-use R package, available at [Bioconductor link].

MATERIAL AND METHODS

All analysis was performed with R > 3.1.1 except if specifically stated. *P* <0.05 were considered significant. False Discovery Rate (*FDR*) was used to correct for multiple testing.

TCGA data acquisition

TCGA data was analyzed using the hg19 assembly. Isoform count matrices (RNASeqV2, quantified with RSEM (33) by TCGA), and patient meta-data for all available TCGA RNA-seq data sets (all 33 cancer types) were downloaded from https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.html on November 3rd 2015, specifying no cell line controls. The GAF file containing annotation of the quantified isoforms was downloaded from https://tcga-data.nci.nih.gov/docs/GAF/GAF.hg19.June2011.bundle/outputs/TCGA.hg19.June2011.gaf. Clinical data was obtained from https://genome-cancer.ucsc.edu/proj/site/hgHeatmap 24th August 2016. COSMIC coding point mutations from targeted and genome wide screens ("CosmicMutantExport") and copy number alterations (CNA, "CosmicCompleteCNA") were downloaded from https://cancer.sanger.ac.uk/cosmic 22nd February 2017.

Isoform Annotation

We manually constructed strand-specific gene ids defined as the most up-and downstream genomic coordinates of overlapping isoforms all having the same gene name.

Library filtering

We discarded libraries with <20 million reads as well as multiple libraries from the same patient (only keeping one). Only cancers types with >24 paired samples after filtering were considered.

RNA-seq data processing

For each cancer type isoform expression was obtained from the downloaded count matrix. To take transcript length into account we calculated Relative Log

Expression (RLE) normalized FPKM (Fragments Per Kilobase of transcript per Million fragments mapped) values. Gene expression was calculated as the sum of the FPKM values of all isoforms associated with that gene id. Isoform Fractions were calculated by dividing the isoform expression with the parent gene expression ($FPKM_{iso}$ / $FPKM_{gene}$, Fig 1B).

Analysis of differential isoform usage in TCGA data

For each cancer type we only considered genes containing multiple isoforms, requiring that the lower boundary of the 95% confidence interval (CI) of the mean expression of a given isoform >1 FPKM in at least one condition from genes where the lower boundary of the 95% CI of the mean gene expression >1 FPKM in both control and tumor samples. The significance of the change in isoform usage for the paired analysis and analysis of all samples was respectively done with a paired and an unpaired Whitney-Mann *U* test (Fig 2A). *dIF* values for paired analyses were calculated as the mean of all paired *dIF* values. *dIF* values from the analysis of all samples were calculated as $mean_{IF2} - mean_{IF1}$. The statistical test was only performed if both conditions had at least 25 defined IF values. We considered a gene as having an isoform switch if it contained at least one isoform where the |*dIF*| value >0.1 and *FDR*<0.05 in both paired analysis and the analysis of all libraries.

IsoformSwitchAnalyzeR

To systematically identify and analyze the potential consequences of isoform switches we developed the R package IsoformSwitchAnalyzeR [Bioconductor link]. See supplementary text 1-2 for a detailed introduction.

Analysis of TCGA data with IsoformSwitchAnalyzeR

Isoform switch analysis was performed with IsoformSwitchAnalyzeR v0.98.0. Except for the isoform switch identification a standard IsoformSwitchAnalyzeR workflow was performed on the TCGA data (see supplementary text 1).

As described above, we identified isoforms with a significant change in isoform usage. For each gene, pairs of isoforms having opposite change in usage (requiring a minimum change of |*dIF*|>0.1) were identified. Next we

annotated isoforms involved in a switch using computational predictions. We annotated pre-mature stop-codons (PTCs) using the 50nt rule (34). Intron retentions were annotated by comparing each isoform to the hypothetical pre-RNA generated when combining all exons of the parent gene, done by the spliceR (32). Using the known exon structure of each isoform, we obtained the corresponding spliced nucleotide sequence and corresponding coding sequence using annotated ORF positions. These sequences were used for sequence analysis of protein domains (via Pfam (30) only using the pfamA database and by specifying –"as"), signal peptides (via SignalP (35) specifying "-f summary") and coding potential (via CPAT). The cutoff for distinguishing coding and non-coding isoforms based on the CPAT analysis was 0.364 as in (36).

We predicted functional consequences of an isoform pair by comparing the isoforms that were used more (*dIF*>0.1) to the isoforms that were used less (*dIF*<-0.1), and identified differences in the obtained annotation via the analyzeSwitchConsequences() function. We focused on six annotation types: gain/loss of protein domains, signal peptides, introns, or coding potential, changes in NMD sensitivity and large changes in amino acid sequence. The latter was defined as a Jacard similarity (length of aligned region w.o gaps) / ((length of sequence a) + (length of sequence b) - (length of aligned region w.o gaps)) of the pairwise alignment of the sequences <0.9. All plots of isoforms and their annotations were made with the switchPlot() function. The global analysis of changes in isoform features was performed with the IsoformSwitchAnalyzeR's extractGenomeWideAnalysis().

Analysis of potential 3'end bias

BRCA samples were divide based on their subtype classification into 'Basal-like', 'Luminal A' and 'Luminal B' and all pairwise comparisons were considered, as well as all control samples vs. all cancer samples. Isoform switches were identified via IsoformSwitchAnalyzeR (supplementary text 2). Predictions of functional consequences were done as described above. The analysis of which transcription start sites (aTSS), alternative splicing (AS) and alternative transcription termination (aTTS) caused the consequences were

analyzed via analyzeSwitchConsequences() by specifying consequencesToAnalyze =c('tss','intron_structure','tts').

Analysis of mutation frequency

The COSMIC database (37) was reduced to cancer types analyzed and the genes tested for isoform switches. For each mutation type the number of times a gene was annotated as mutated was counted and genes were divided into those with and without isoform switches with predicted functional consequences. Only deleterious mutation types found at least 25 genes in each category were considered.

Enrichment tests

All enrichment analysis was done using a Fishers exact test except if especially stated. Expressed multi-isoform genes were used as background.

Gene set enrichment analysis

Gene Ontology (GO) gene-sets were downloaded (6th Jan 2016) from EBI. GO-terms from the 5th level were used. MSigDB's gene sets c2, c6 and H was downloaded (6th Jan 2016) from http://bioinf.wehi.edu.au/software/MSigDB/ (38). Gene names were converted to Ensembl gene ids. Only gene-sets with >9 genes were considered.

The functional class scoring analysis (Fig. S1D), using enrichment *P*-values, was done for each of each gene set as follows: 1) For each cancer type a Fishers Exact test was performed only testing enrichment. In each cancer type the only expressed multi-isoform genes were used as background. 2) The average –log10(*P*) across all 12 cancer types was calculated. 3) A bootstrap p-value was calculated by, from the cancer-type specific expressed multi-isoform genes, sampling 1000 random gene sets of equal size as the gene set in question and the average –log10(*P*) was calculated.

Significant gene sets across cancer types were required to have: 1) bootstrap $P<0.001$. 2) Average $-\log_{10}(P)>-\log10(0.05)$. 3) Median odds ratio>2. 4) Average genes found in >5 gene-set.

Protein network analysis

The protein network analysis was done via the STRING(39) webserver http://string-db.org/ only considering “high confidence” interactions (interaction score > 0.7).

Literature searches

All literature searches were performed on PubMed (http://www.ncbi.nlm.nih.gov/pubmed) using the RISmed R package. We used two queries, one for finding RNA-seq articles (“RNA-seq search”) and for finding RNA-seq articles analyzing isoforms (“isoform search). Both queries were limited to 1) articles published in the first 5 months of 2016. 2) article type=’Journal Article’. 3) English articles. 4) Articles about human or mouse data. The RNA-seq search was performed searching for ”RNA-seq or ”RNA-sequencing” specifically excluding single cell studies, resulting in 518 articles. The isoform search was performed using the search terms ’isoform’, ’isoforms’, "Alternative Splicing"[Mesh] ’transcript’ or ’transcripts’ resulting in 1000 articles; 51 articles was present in both groups.

We defined an ”RNA-seq only” set, containing the RNA-seq search articles not overlapping with the isoform search, and a ”RNA-seq and isoform” set containing the articles found in both searches. We randomly selected 50 articles from each set, which were read and categorized as analyzing RNA-seq data with isoform resolution or not; 30 articles were discarded because they were either analyzing non-standard RNA-seq data, describing tools for analysis or were not available online. From this categorization we estimated fraction of articles expected to be analyzing RNA-seq data with isoform resolution.

To identify known human isoform switches in cancer we constructed a comprehensive set of phrases for describing isoform switches in cancer. We identified 26 different phrases used in the literature we already knew. These were deconstructed into 9 prefixes (e.g. ’differential’), 14 feature references (e.g. ’isoform’) and 4 suffixes (e.g. ’usage’). All possible combinations of a

prefix, a feature reference and a suffix were created. Combinations were used to perform a PubMed search (13th Jan 2016) in combination with variations of 'human' and 'cancer'. All relevant articles were read and genes containing isoform switches with a described functional consequence in cancer were extracted. Only isoform switches in the 12 cancers analyzed here (Table1) were considered.

Analysis of patient survival

We identified isoform switches in individual patients as described in the main text. We used a cox proportional hazard regression model (the coxph() function from the survival R package (40)). To test for differences in the overall survival rates and for differences in hazard ratios two separate tests were needed since the *P*-values for the survival test were not uniformly distributed when covariates were included. We therefore tested each isoform switch both with and without age, gender and parent gene expression as covariates.

To label an isoform switch as significantly associated with poor patient survival we required:

1) The switch had to be identified in at least 5 diseased patients and not be identified in at least other 5 patients.
2) Hazard ratio *FDR* (with covariates) < 0.05.
3) Log-likelihood test *FDR* (without covariates) < 0.05.
4) Median survival rate < 1.
5) Hazard ratio > 1.

The pan-cancer analysis was stratified by cancer type via the strata() function and gender was omitted as co-factor for PRAD due to no females in that cohort.

RESULTS

Identification of isoform switches with predicted consequences from TCGA data

In order to comprehensively characterize isoform switches in cancer we utilized the extensive TCGA RNA-seq (7). After filtering, our dataset contained 6194 RNA-seq libraries from 5,562 cancer patients covering 12 solid cancer types (Table 1). For each cancer type, RNA-seq libraries from tumors and corresponding healthy tissue were available, and for a substantial subset, tumor and healthy tissues were paired (originating from the same patient; Table 1). The availability of paired samples made it possible to identify switches using either a paired approach, comparing RNA-seq profiles from tumor and healthy tissue from the same patient, or an unpaired approach also including tumor and healthy tissue samples that were from different patients (Fig 2A, top panel). The advantage of the paired approach is increased statistical power, as potential artifacts such as batch effects and inter-patient variance are omitted, while on the other hand the unpaired analysis includes more samples and thereby allows for better generalization.

To exploit the advantages of both approaches, we tested each isoform for differential usage by comparing the IF values of tumor and healthy tissue samples using a paired Mann-Whitney *U* test for the paired samples and a standard Mann-Whitney U test for the unpaired analysis. For each such test, we required: i) at least 25 patients in each group (healthy and tumor), ii) a difference in average isoform usage larger than 10% |*dIF* |>0.1 and iii) an FDR <0.05. Isoform switches were called when both tests passed the outlined requirements and where we also found opposite usage |*dIF* |>0.1 of another isoform from the same gene (Fig. 2A, bottom panel).

Many isoform switches identified this way may be inconsequential to cells, since the isoforms might have identical biological functions. One example is closely located alternative TSSs in the 5' UTR, where the effect of an isoform switch will be to slightly change the 5' UTR length: such a change may in some cases not have a measurable biological effect.

To focus on switches with potential biological impact, we only analyzed isoform switches where we could predict such consequences. To predict consequences, we first collected a set of annotations for each isoform involved in a switch utilizing its annotated open reading frame (ORF). The collected annotation included i) protein domains predicted by Pfam (30), ii) signal peptides identified via SignalP (35), iii) protein coding potential predicted by CPAT (36), iv) large changes in amino acid sequence (see material and methods), v) intron retention identified by spliceR (32) and vi) NMD sensitivity, analyzed by the 50-nt rule (see methods)(34). For each gene with a significant isoform switch, we then in a pairwise manner compared the annotations for the isoform used more in a given state (*dIF* > 0.1) to the isoform used less (*dIF* < -0.1). Cases where we found differences in the annotations between the two isoforms were considered as isoform switches with predicted functional consequences (workflow illustrated in Fig. 2B, dashed line).

We assessed whether the prediction of isoform switches could be confounded by systematic bias in the data set due to either differences in GC content or changes in RNA quality between cancer and control and found that such biases have either no or minor effect on our results (Supplementary Text 3).

Extent and verification of isoform switching with potential functional consequences

We found isoform switches with predicted functional consequences to be common in cancers: on average each cancer type had 357 genes (3.5% of expressed multi-isoform genes) containing at least one isoform switch predicted to result in functionally different RNAs/proteins (Fig. 3A, Fig. S1A-C, Table S1-2). Across all 12 cancer-types analyzed, we found 2,352 genes (18.96% of all expressed multi-isoform genes) that had such a switch (Fig. 3B). These genes were generally associated with signal transduction and stemness, and were part of known cancer gene signatures (Fig. S1D). Interestingly, these genes were also more frequently mutated in somatic cancers, as annotated in the COSMIC database (37), compared to genes

without such switches (Fig 3C). We reasoned that the high occurrence of isoform switches might be the result of the disruption of the spliceosome. If this was the case one would expect to see global changes in isoform usage (41). To investigate this, we compared the genome-wide distribution of control and cancer isoform usage for isoforms with a particular feature (e.g. ORF, protein domain etc.). With a single exception in KIRP, we found no indications of global disruptions (Fig. S1E).

To analyze the quality of the identified isoform switches, we performed a systematic literature review and identified 47 genes with known isoform switches that all were described as having functional consequences in their respective studies (Table S3). These literature-curated switches were highly enriched amongst the set of genes where we identified isoform switches with predicted consequences compared to all genes we tested for isoform switching (when only considering the parent gene: odds ratio >4.2, P < 3.8e-12, Fisher's Exact test, when also considering the cancer type in which the isoform switch is described: odds ratio >16.33, P < 3.4e-35, Fisher's Exact test). Because we systematically analyzed 12 cancer types, we could in many cases extend the number of cancer types in which a switch was observed, compared to the cancers where the switch was previously described (Fig 3D, Table S3). For example, *ZAK* (also known as MLTKα), a gene with known isoform-specific oncogenic properties, has an isoform switch which was previously described in three cancer types (BRCA, COAD and STAD)(29). We identified this isoform switch in three additional cancer types (KIRC, KIRP and PRAD) (Fig. 3D) and found that the isoform switch causes the inclusion of exons encoding the SAM_1 protein domain, which typically facilitates protein-protein interactions (Fig 3E), illustrating the usefulness of our methods for hypothesis generation.

Consequences of isoform switching

Because we applied the same analysis across 12 cancer types, we were able to systematically investigate the occurrence of different predicted consequences across different cancer types. The most commonly predicted

consequences of isoform switching were large changes in amino acid sequence (corresponding to loss and/or gain of amino acid coding exon(s)), change of protein domains and change of signal peptides (Fig. 4A, Fig. S2A). Surprisingly, the distribution of protein domain gain *vs.* domain loss was highly skewed towards domain loss for isoforms upregulated in cancer (Fig. 4A, 1st column). To investigate what gain/loss ratio would be expected if isoform switching occurred randomly (i.e. if it were not regulated) we calculated, for each cancer-type, the expected distribution of gain/loss ratios by randomly sampling the same number of isoform pairs from non-switching protein coding genes 1000 times. Compared to this background, loss of protein domains was significantly more frequent than domain gain in all 12 cancer-types ($P < 0.001$, by sampling) (Fig. 4B, Fig. S2B). Importantly, the observed prevalence of protein domain loss was only partially explained by the commonly observed isoform switches resulting in protein-coding to non-coding change (ORF loss) (Fig 4A, 3rd column, Fig. S2C).

Switches resulting in ORF loss occurred significantly more often than expected in all 12 cancer types (*FDR* < 5.82e-10, Fisher Exact test)(Fig 4C, black dots). Using a more conservative ORF loss threshold (requiring ORF loss as well as loss of coding potential, as calculated by CPAT (36)), this held true for 10 of the 12 cancer types (*FDR* < 0.05, Fisher Exact test) (Fig 4C grey dots, Fig. S2D). Considering the overrepresentation of ORF loss, and that upregulated isoforms in cancer typically were shorter (Fig. S2E) it was particularly interesting that isoform switching resulting in gain of signal peptides occurred significantly more often than expected in 9 of the 12 cancer types (*FDR* < 0.05, Fisher Exact test) (Fig 4D).

The regulatory mechanisms controlling the production of alternative isoforms are often assumed to originate from AS. Thus, the contribution from aTSS or aTTS has not been comprehensively investigated. To analyze the contribution of the individual mechanisms we decomposed the isoforms involved in isoform switches with predicted consequences into those originating from AS, aTSS, aTTS and combinations thereof. AS and aTSS were the main contributors: 78.4% and 70.3% of isoform switches involved AS and aTSS

respectively, while 52% of isoform switches involved both (Fig. 4E-F). For switches where only one mechanism was used, aTSS and AS accounted for 40.4% and 51.5%, respectively. In comparison, aTTS contributed much less: globally only 17.3% of switches utilize aTTS, and aTTS could by itself only explain 8.1% of the single mechanism switches (Fig. 4E-F).

Pan-cancer isoform switches with predicted functional consequences

We reasoned that isoform switches, with predicted functional consequences, which were observed across two or more cancer types, referred to as pan-cancer switches, were of particular interest. We identified 945 genes containing such switches (Fig 5A, Table S4), which were enriched for genes involved in signal transduction and stemness, and were over-represented in known cancer signature gene lists (Fig 5B). In these pan-cancer isoform switches, the predicted protein domain gain/loss ratio was particularly skewed towards domain loss (Fig 5C) (78.26 % of events, $P < 0.001$, bootstrap test), a ratio higher than the domain gain/loss ratio for 10 of the 12 individual cancer types analyzed (Fig. S3A). In the pan-cancer isoform switches we identified 11 protein domains that, compared to single cancer isoform switches, were lost from significantly more genes than expected by chance (*FDR* < 0.05, Fisher Exact test) (Fig 5D, left). These domains were linked to signal transduction (e.g. protein kinase domains), extracellular space (e.g. immunoglobulin domains) and protein-protein interaction (e.g. PH domains) (Fig 5D, right).

We hypothesized that genes containing pan-cancer switches may be functionally connected. To assess this we used the STRING database (39), a resource summarizing protein-protein interaction evidence from multiple sources. Using genes containing pan-cancer switches observed in at least four cancer types, we identified a high-confidence protein interaction network (7.45x more interactions than expected, $P < 5.8e-04$, STRING online analysis) (Fig. 5E, Fig. S3B). The protein interaction network, which mainly consisted of genes involved in cell signaling, contained *VEGFA* and *AKT1* (also known as

PKB1 or PKBα) as central nodes (Fig 5E). The previously uncharacterized isoform switch in *AKT1* is of particular interest, since *AKT1* is a known tumor suppressor that inhibits tumor invasion (42). We found this isoform switch in five different cancer types (COAD, KICH, KIRC, KIRP and THCA) and our detailed analysis revealed that the upregulated isoform lacked the PH domain present in the normal isoform (Fig. 5F). Since PH domains typically facilitate protein-protein interactions, this could indicate that the ability of the cancer form of *AKT1* to co-localize with its normal signal transduction partners is decreased, perhaps decreasing its suppressor effect. However, the causality and functional importance of this switch remains to be tested.

Isoform switches associated with worse pan-cancer patient survival

Since analyses of patient survival rates using exon level expression data have previously been shown to complement gene-level expression (43), we hypothesized that isoform switches could be predictors of patient survival. To investigate this we analyzed each cancer patient for the presence of any of the identified 2,792 isoform-pairs involved in switches with predicted consequences as follows. First, for each cancer type, we analyzed the average isoform fractions of the isoform pair in all control samples. Next, for each of the 5,232 cancer patients for whom survival data was available (Table 1), we compared these isoform fractions to the distributions from corresponding healthy samples. If the isoform fraction in a particular cancer patient was different from healthy samples (defined as a $|Z\text{-score}| > 3$ and a combined *dIF* score of the two isoforms > 0.2 compared to the average of healthy samples), we classified the individual cancer patient as having the isoform switch. We then compared the survival rates of patients with and without the isoform switch within each cancer type and across all cancers (a pan-cancer analysis) taking age at diagnosis, cancer type, gender and the expression of the parent gene into account (see materials and methods). This resulted in the identification of 1195 isoform switches significantly associated with worse patient survival in at least one cancer type. Of these we found 111

isoform switches that were significantly associated with worse patient survival in the pan-cancer analysis (Table S5).

Encouraged by these findings, and the large number of pan-cancer isoform switches identified we hypothesized that for some isoform switches the associations with worse survival rates might be independent of cancer type. In other words, there may be cases where an isoform switch is associated with worse survival rates regardless of which cancer type it was identified in.

We reasoned that if the confidence (as estimated by *P*-value) of the association between an isoform switch and worse survival rates was larger in the pan-cancer analysis (smaller *P*-value) compared to any of the corresponding single cancer analysis, the association was independent of cancer types. Importantly, such a result would not be explained by differences in the number of patients analyzed since our pan-cancer analysis is stratified by cancer type and thereby only reflects cancer types where patients with the switch are identified (Fig. S4).

To identify such isoform switches, we compared the *P*-value of the pan-cancer analysis to the *P*-value of the most significant single cancer analysis. This resulted in the identification of 31 isoform switches which were significantly associated with worse survival rates independent of cancer types (Fig. 6A)(Table S5).

Since an isoform switch is not a binary event we hypothesized that the magnitude of isoform switches could be an important parameter for survival analysis. To analyze this we divided cancer patients based on the magnitude of the isoform switch for each of the 31 pan-cancer predictors: isoform switches with a combined *dIF* score of the two analyzed isoforms >0.5 were defined as “large” switches and remaining cases as “normal”. Based on this stratification we reanalyzed the survival rates and obtained significant results for the individual analysis of both normal and large switches for 13 out of 30 pan-cancer predictors (details in material and methods). As hypothesized, 9

out of 13 (69%) cases showed clear differences in median survival rates with worse outcome for patients with larger switches (Table S5).

To illustrate the results of our survival analysis, we here provide an in depth description of four isoform switches with potential functional consequences where each switch was predictive of patient survival independent of cancer type.

The *ZNF12* gene is a transcriptional repressor that, through its KRAB domain, suppresses AP-1 and SRE mediated transcriptional activity (44). An isoform switch in *ZNF12* (identified in 5 cancer types, Table S5) resulted in the loss of the transcriptional inhibitory KRAB domain (Fig. 6B, top). This switch and in the pan-cancer analysis this isoform switch was associated with worse survival rates than any of the single-cancer analysis (Hazard ratio: 1.58, Hazard ratio $P < 0.0013$) (Fig. 6B, bottom).

Another isoform switch resulting in the loss of a protein domain was found in *ERCC1*, a gene with a central role in the nucleotide excision repair system, where mutations have previously been associated with worse survival rates (45). We identified this isoform switch in 1,091 patients (corresponding to an average 23.5% patients in each of the 12 individual cancer-types, Table S5). This isoform switch was predicted to result in a protein lacking the HHH domain in cancer states (Fig. 6C, top). Since the HHH domain is a sequence-nonspecific DNA binding element this switch could potentially compromise the function of *ERCC1*. Patients with this isoform switch consistently had lower survival rates than patients without it, in both the analysis of HNSC (the single cancer type where the survival difference was the most significant) and the pan-cancer analysis (Hazard ratio: 1.28, Hazard ratio $P < 0.001$) (Fig. 6C, bottom).

FAM36A, which encodes an assembly factor important for the last step of the electron transport chain in oxidative phosphorylation, had an isoform switch which was found in more than 1000 patients across all 12 cancer-types analyzed (Table S5). This isoform switch, which constitutes a change from a

non-coding to a coding isoform (Fig. 6D top), was associated with worse survival rates across cancer types (Hazard ratio: 1.4, Hazard ratio *P* < 1.63e-06)(Figure 6D bottom), supporting recent findings that oxidative phosphorylation might play an important role in cancer progression (46). Importantly, the size of the isoform switch seems to be associated with worse patient outcome (Fig. 6E).

One of the most extreme pan-cancer predictors identified was the isoform switch in the *SHC1* gene. This isoform switch was found in 11 cancer types (Table S5) and associated a median survival rate of 0.58 (Fig. 6A). The isoform switch was predicted to result in increased usage of the longer isoform, which is often referred to as p66Shc (Fig. 6F, top). This long isoform included an additional 110 N-terminal amino acid region, and is known to play a central role in prostate cancer metastasis as well as many other cellular functions (47). In agreement with these results we found the isoform switch to be associated with poor patient survival across cancer types (Hazard ratio: 1.58, Hazard ratio *P*-value < 4.70e-06) (Fig. 6F, Bottom) a trend worse for patients with large isoform switches (Fig. 6G).

Resources and software for mining isoform switches in cancer

To facilitate the analysis of isoform switches with potentially functional consequences in cancer we provide access to the data presented here in several ways, aimed at different types of users. For easy and fast exploration of isoform switches in cancer we generated three interactive online web-services, which produce isoform switch plots (similar to Fig. 3D or 5F), where the genes can be selected with specific focuses, either: i) gene-oriented, ii) cancer-type oriented, iii) pan-cancer oriented. These tools are available at http://www.binf.ku.dk/services/#switch_cancer. We also provide all the results presented here as supplementary data for power users (Table S1-5). In order to streamline analysis of isoform switches with predicted consequences (Fig. 2B) and for others to be able to apply our methods to other RNA-seq data sets, we implemented our methods as an R package, IsoformSwitchAnalyzeR, which enables statistical identification of isoform

switches with predicted functional consequences from RNA-seq data. IsoformSwitchAnalyzeR is available through Bioconductor [Bioconductor link] and is to our knowledge the first tool enabling statistical identification of the specific isoforms involved in a switch (Fig. S5A).

Details about IsoformSwitchAnalyzeR can found in the vignette (Text S1) and Text S2. Lastly two SwitchAnalyzeRlist objects, which can be explored via IsoformSwitchAnalyzeR, and contains the full switch analysis of TCGA on which all of the above analysis is made are available via figshare https://figshare.com/s/804a08cff94ba9909073.

DISCUSSION

Despite the large amount of RNA-seq data and computational methods available, isoform-based expression analysis is rare. This means that the potential of existing RNA-seq data is untapped, and as a consequence, our general understanding of differential isoform usage is poor. The few efforts at analyzing individual isoform switches have typically dealt with isoforms by describing their frequent occurrence rather than trying to systematically predict their consequence. Overall, this is unsatisfying, as isoform usage is important in disease and especially cancer, where many individual isoform switches have been described.

Here we present methods for the statistical identification and analysis of isoform switches with predicted functional consequences. We utilized these methods to make the most comprehensive analysis of isoform switches in cancer to date, analyzing data for more than 5,500 cancer patients. According to our analysis, isoform switches with predicted functional consequences are common, occur in genes important for cancer and are often shared between cancer types. Such isoform switches often lead to protein domain loss and/or ORF loss. Conversely, signal peptides are more often gained than lost in cancer-upregulated isoforms. Many of these switches have previously been described as functionally important, and in many instances causal for cancer development and/or progression. Regardless of causality, a subset of such

isoform switches was proven to predict patient survival independent of cancer type. As the survival rates were associated with the magnitude of switches, these switches may be useful as biomarkers.

Although we found that isoform switching with predicted functional consequences is common, we believe our results are a conservative estimate for several reasons: First, the analysis presented here is based on UCSC KnownGenes isoform models released in 2009, while the most recent GENCODE annotation have >2.5x more transcripts annotated and contains ~50% more multi-isoform genes. Similarly, it is likely that many novel isoforms, which are not in the current annotation, are used in cancer patients. Consequently, we may not be analyzing the full spectrum of isoforms, which limits our ability to detect isoform switches. Second, in the analysis presented here we did not analyze cancer sub-types because we prioritized a high number of paired samples to account for inter-patient and batch effects. This means we only identify isoform switches shared by the majority of patients in each cancer-type. Thus, we will not identify isoform switches found in smaller sub-types. One such example is the isoform switch in *ERCC1,* which we first globally identified in LUAD and KICH, but deeper analysis showed that the isoform switch existed in subsets of patients from all 12 cancer types analyzed. Finally, we have focused on isoform switches with predicted functional consequences. Such predictions are limited, since we cannot predict all functional features of genes. For example we identified an isoform switch from a short to a long isoform in the insulin receptor (*INSR*) in eight cancer-types (Fig. S5B), an isoform switch suggested to have a role both in fetal growth and cancer biology (48). However, because we did not predict a functional consequence of this switch, it was not analyzed further.

Overall our results strongly indicate that isoform switches with predicted functional consequences are both common and important in dysfunctional cells, illustrating the potential of augmenting gene-level analysis with isoform-level analysis. This idea is especially appealing in the light of recent findings showing that utilization of isoform-level expression data also might lead to more reliable gene-level expression estimates (49). While there is ample room

for improvements in algorithms for isoform reconstruction and quantification, our results suggest that current methods are still adequate for isoform-level-based analysis.

We expect our data sets will be a significant resource for cancer biology research and that the methods described here will be useful for analyzing other RNA-seq sets, enabling analysis of isoform switches with predicted functional consequences in other diseases or physiological states.

ACKNOWLEDGEMENTS

We thank Sarah Rennie for comments on the manuscript. This study was supported by grants from the Novo Nordisk Foundation and the Lundbeck Foundation to AS.

CONTRIBUTIONS

KVS and AS designed the study. KVS developed and implemented the algorithms and analyzed the data supervised by AS. KVS and AS wrote the manuscript.

REFERENCES

1. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012;338:1587–93.
2. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40:1413–5.
3. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, et al. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet. 2006;38:626–35.
4. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
5. Kornblihtt AR, Schor IE, Alló M, Dujardin G, Petrillo E, Muñoz MJ. Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat Rev Mol Cell Biol. Nature. 2013;14:153–65.
6. Schwerk C, Schulze-Osthoff K. Regulation of apoptosis by alternative pre-mRNA splicing. Mol Cell. 2005;19:1–13.
7. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. Nature Publishing Group; 2013;45:1113–20.
8. Kalsotra A, Cooper TA. Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet. Nature Publishing Group; 2011;12:715–29.
9. Ye J, Blelloch R. Regulation of pluripotency by RNA binding proteins. Cell Stem Cell. 2014;15:271–80.
10. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456:470–6.
11. Davuluri R V., Suzuki Y, Sugano S, Plass C, Huang THM. The functional consequences of alternative promoter use in mammalian genomes. Trends Genet. 2008;24:167–77.
12. Rajan P, Elliott DJ, Robson CN, Leung HY. Alternative splicing and

biological heterogeneity in prostate cancer. Nat Rev Urol [Internet]. Nature Publishing Group; 2009;6:454–60.

13. Wiesner T, Lee W, Obenauf AC, Ran L, Murali R, Zhang QF, et al. Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. Nature. 2015;526:453–7.
14. Sveen a, Kilpinen S, Ruusulehto A, Lothe R a, Skotheim RI. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene. Nature Publishing Group; 2015;1–15.
15. Venables JP. Aberrant and alternative splicing in cancer. Cancer Res. 2004;64:7647–54.
16. Lee SC-W, Abdel-Wahab O. Therapeutic targeting of splicing in cancer. Nat Med. 2016;22:976–86.
17. Bonnal S, Vigevani L, Valcárcel J. The spliceosome as a target of novel antitumour drugs. Nat Rev Drug Discov. 2012;11:847–59.
18. Trapnell C, Williams BB a, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.
19. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
20. Bao J, Vitting-Seerup K, Waage J, Tang C, Ge Y, Porse BT, et al. UPF2-Dependent Nonsense-Mediated mRNA Decay Pathway Is Essential for Spermatogenesis by Selectively Eliminating Longer 3'UTR Transcripts. PLOS Genet. 2016;12:e1005863.
21. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. Science . 2015;348:660–5.
22. Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Lassmann T, et al. A promoter-level mammalian expression atlas. Nature. 2014;507:462–70.
23. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31:46–53.
24. Niu L, Huang W, Umbach DM, Li L. IUTA: a tool for effectively detecting

differential isoform usage from RNA-Seq data. BMC Genomics. 2014;15:862.

25. Shi Y, Chinnaiyan a. M, Jiang H. rSeqNP: a non-parametric approach for detecting differential expression and splicing from RNA-Seq data. Bioinformatics. 2015;31:2222–4.
26. Nowicka M, Robinson MD. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. F1000Research. 2016;5:1356.
27. Sebestyen E, Zawisza M, Eyras E. Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. Nucleic Acids Res. 2015;43:1345–56.
28. Zhao W, Hoadley KA, Parker JS, Perou CM. Identification of mRNA isoform switching in breast cancer. BMC Genomics. BMC Genomics; 2016;17:181.
29. Liu J, McCleland M, Stawiski EW, Gnad F, Mayba O, Haverty PM, et al. Integrated exome and transcriptome sequencing reveals ZAK isoform usage in gastric cancer. Nat Commun. Nature Publishing Group; 2014;5:1–8.
30. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012;40:D290-301.
31. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. 2007;2:953–71.
32. Vitting-Seerup K, Porse BT, Sandelin A, Waage J. spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. BMC Bioinformatics. BMC Bioinformatics; 2014;15:81.
33. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
34. Weischenfeldt J, Waage J, Tian G, Zhao J, Damgaard I, Jakobsen JS, et al. Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. Genome Biol. 2012;13:R35.

35. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785–6.
36. Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 2013;41:e74.
37. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45:D777–83.
38. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27:1739–40.
39. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43:D447–52.
40. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York, NY: Springer New York; 2000.
41. Oka Y, Varmark H, Vitting-Seerup K, Beli P, Waage J, Hakobyan A, et al. UBL5 is essential for pre-mRNA splicing and sister chromatid cohesion in human cells. EMBO Rep. 2014;15:956–64.
42. Chin YR, Toker A. Akt isoform-specific signaling in breast cancer: uncovering an anti-migratory role for palladin. Cell Adh Migr. 2011;5:211–4.
43. Shen S, Wang Y, Wang C, Wu YN, Xing Y. SURVIV for survival analysis of mRNA isoform variation. Nat Commun. 2016;7:11548.
44. Zhao Y, Zhou L, Liu B, Deng Y, Wang Y, Wang Y, et al. ZNF325, a novel human zinc finger protein with a RBaK-like RB-binding domain, inhibits AP-1- and SRE-mediated transcriptional activity. Biochem Biophys Res Commun. 2006;346:1191–9.
45. Takenaka T, Yano T, Kiyohara C, Miura N, Kouso H, Ohba T, et al. Effects of excision repair cross-complementation group 1 (ERCC1) single nucleotide polymorphisms on the prognosis of non-small cell lung cancer patients. Lung Cancer. 2010;67:101–7.
46. Tan AS, Baty JW, Dong LF, Bezawork-Geleta A, Endaya B, Goodwin J,

et al. Mitochondrial genome acquisition restores respiratory function and tumorigenic potential of cancer cells without mitochondrial DNA. Cell Metab. 2015;21:81–94.

47. Rajendran M, Thomes P, Zhang L, Veeramani S, Lin M. p66Shc - a longevity redox protein in human prostate cancer progression and metastasis : p66Shc in cancer progression and metastasis. Cancer Metastasis Rev. 2010;29:207–22.
48. Frasca F, Pandini G, Scalia P, Sciacca L, Mineo R, Costantino A, et al. Insulin receptor isoform A, a newly recognized, high-affinity insulin-like growth factor II receptor in fetal and cancer cells. Mol Cell Biol. 1999;19:3278–88.
49. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2015;4:1521.

TABLES

Table 1: Overview of RNA-seq samples analyzed

Cancer Type		Number of RNA-seq Libraries Analyzed						Number Patients with Survival Data		
Abbreviation	Full Name	Unpaired cancer	Paired	Unpaired healthy	Cancer total	Healthy total	Combined	Alive	Dead	Combined
BRCA	Breast invasive carcinoma	981	113	0	1094	113	1207	911	130	1041
COAD	Colon adenocarcinoma	255	26	15	281	41	322	208	58	266
HNSC	Head and Neck squamous cell carcinoma	477	43	1	520	44	564	316	190	506
KICH	Kidney Chromophobe	41	25	0	66	25	91	57	8	65
KIRC	Kidney renal clear cell carcinoma	459	72	0	531	72	603	357	165	522
KIRP	Kidney renal papillary cell carcinoma	256	32	0	288	32	320	216	30	246
LIHC	Liver hepatocellular carcinoma	321	50	0	371	50	421	222	105	327
LUAD	Lung adenocarcinoma	456	55	3	511	58	569	336	135	471
LUSC	Lung squamous cell carcinoma	444	51	0	495	51	546	287	181	468
PRAD	Prostate adenocarcinoma	442	52	0	494	52	546	447	7	454
STAD	Stomach adenocarcinoma	375	32	3	407	35	442	240	126	366
THCA	Thyroid carcinoma	445	59	0	504	59	563	486	14	500
Pan-cancer	Across all 12 cancer types analyzed	4952	610	22	5562	632	6194	4083	1149	5232

FIGURE LEGENDS

Figure 1: Illustration and definitions of isoform switching.

A) Schematic overview of an isoform switch. Top panel shows two isoforms originating from the same gene. Bottom pane shows the expression of both isoforms in two conditions (healthy and tumor). A switch is defined as cases where the relative contribution of the isoforms to the parent gene expression changes significantly between conditions.

B) Conceptual examples of calculation of isoform fraction values (*IF*) and difference between isoforms (*dIF*). The two top tables show the expression of two isoforms in two conditions, measured in FPKM. The isoform fraction is calculated by the ratio of each FPKM value to the total gene expression. Calculation of the difference in IF values (*dIF*) is defined as the difference between IF values of a given isoform in the two conditions, as shown in the bottom panel (*IF2 – IF1*).

C) Statistical testing of isoform fraction values across samples. Boxplots show the distribution of IF values across multiple individuals for tumor/healthy

states and two isoforms. A statistical tests can be made for each isoform by comparing the IF value distributions between states.

Figure 2: Data and data analysis.

A) Strategy for detecting of isoform switching from RNA-seq data in a single cancer cell type. Within each cancer type, the TCGA RNA-seq libraries can be divided by sample type, as illustrated in this classification tree. At the first level from the top, samples are divided depending on whether they are from healthy tissue or from tumors, as indicated by grey boxes. Such samples do not have to be paired (originating from the same patient). Isoform switch analysis between healthy and tumor samples on this level will be referred to as “unpaired” analysis. At the second level, we focus on those tumors and healthy tissue samples originating from the same patients, forming pairs. Isoform switch analysis between such paired samples will be referred to as “paired” analysis. Only isoform switches detected in both the paired and the unpaired analysis are considered for functional consequence prediction (panel B).

B) Overview of computational pipeline for isoform detection and functional prediction. The pipeline allows analysis starting from full-length isoform quantification data from RNA-seq experiments (e.g. the output of Cufflinks(18), Kallisto(19) or similar). The pipeline allows the identification of isoform switches, followed by computational annotation of isoforms and the prediction of functional consequences. Because of the specific samples setup in the TCGA data (panel A), we used only the parts of the pipeline for annotation concatenation and prediction of functional consequences for the analysis of the TCGA data (indicated by dotted lines), only considering the set of predicted consequences listed in the grey area. The pipeline is implemented as an R package ‘IsoformSwitchAnalyzeR’ (Methods, Supplementary text 1-2) available though Bioconductor.

Figure 3: Isoform Switches with Predicted Consequences

A) Extent of isoform switching across cancer types. Left panel shows the number of genes having at least one isoform switch between healthy and tumor samples predicted to have a functional consequence in each cancer

type. Right panel shows corresponding fractions of analyzed genes having at least one such isoform switch.

B) Extent of isoform switch usage in any cancer type. Bar plots show the number and percentage of tested genes and isoforms involved in isoform switches with predicted functional consequences in at least one cancer-type.

C) Mutation frequency in genes with isoform switches.

Y-axis shows the mutation frequency of genes taken from the COSMIC database as boxplots. Genes are split by whether they contain isoform switches with predicted consequences or not (X-axis). Subplot shows different mutation types. Significance was assessed with a Mann-Whitney *U* test.

D) Visualization of analysis of literature-derived isoform switches with functional consequences. Rows indicate cancer types analyzed. Columns indicate genes with literature curated isoform switches found in at least two cancer types after extension with our analysis. Light grey color indicates isoform switches identified both by our analysis and in previous studies in the indicated cancer type. Grey indicates switch/cancer type combinations only found in literature, while dark grey indicates switch/cancer type combinations only found in this study.

E) Example of isoform switch in the ZAK gene in KIRP cancer type, identified in panel F.

Top panel shows two analyzed isoforms for the ZAK gene. Genomic features are rescaled to the square root of their original size in bp. PFAM domains are indicated in color. UCSC gene transcript IDs are shown for each isoform. Left bottom panel shows the average gene expression in healthy tissues and KIRP tumors. Middle bottom pale shows the average isoform expression in healthy tissues and KIRP tumors. Bottom right pale shows the average isoform fraction in healthy tissues and KIRP tumors. ‘ns’ indicates 'not significant' and *** indicates a FDR < 0.001 (EdgeR for expression tests and Mann-Whitney two sided (as described in main text) for isoform fraction analysis). Error bars indicate 95% confidence interval.

Figure 4: Consequences of Isoform Switches

A) Extent of predicted functional consequences. Each panel shows boxplots indicating the number of genes (Y-axis) with at least one isoform switch with predicted consequences across cancer types (indicated by dots). The X-axis indicates the feature of the upregulated isoform divided into category of consequence as indicated at the top of panels. A feature "switch" indicates that both a gain and a loss in that category occurred.

B) Loss vs. gain of protein domain analysis. Only genes that are subject to a switch in which a protein domain is lost or gained were analyzed. X-axis shows the fraction of such genes that lost a domain. Y-axis shows the cancer types analyzed. Error bars indicate the 95% confidence intervals of the true fraction. The expected fraction by random sampling (0.5) is indicated by a dotted line. Significance levels were estimated by sampling (see Fig S2B and main text). Triple asterisks (***) indicates $P < 0.001$.

C) Coding to non-coding switch analysis. X-axis shows the enrichment (as $\log_2$ odds ratio) of switches resulting in a non-coding isoform compared to the background frequency of non-coding isoforms in all transcripts tested for isoform switches. Y-axis shows the cancer types analyzed. Two definitions of coding sequence loss were analyzed: ORF loss (black) and ORF loss combined with a loss of predicted coding potential, as calculated by CPAT (grey; also see Suppl. Fig 8)(grey). Dashed line indicates no enrichment. Error bars indicate the 95% confidence interval of the odds ratio. Significance levels were estimated by Fishers exact tests. Triple asterisks (***) indicates *FDR* < 0.001, and single asterisks (*) indicates *FDR* < 0.05.

D) **Loss vs. gain of signal peptide analysis.** X-axis shows the enrichment (as $\log_2$ odds ratio) of switches resulting in a signal peptide gain compared to the background frequency of signal peptides in all transcripts tested for isoform switches. Y-axis shows the cancer types analyzed. Dashed line indicates no enrichment. Error bars indicate the 95% confidence interval of the odds ratio. Significance levels were estimated by Fishers exact tests. Triple asterisks (***) indicates *FDR* << 0.001, and single asterisks (*) indicates FDR corrected P < 0.05.

E-F) Mechanisms underlying observed isoform switches. Venn diagrams show the number (E) and percentages (F) of isoform switches with predicted consequences, where the changes in the upregulated isoform, compared to

the downregulated isoform, are caused by alternative splicing, alternative transcription start sites, alternative transcription termination sites or combinations thereof. Percentages in parentheses are calculated from isoform switches that only utilize a single mechanism.

Figure 5: Pan cancer isoform switches with predicted consequences.

A) Reoccurrence of isoform switches across cancer types. X-axis shows the number of isoform switches with predicted consequences that are identified as reoccurring in the numbers of cancer types indicated on the Y-axis.

B) Gene ontology enrichment in genes subject to pan-cancer isoform switching. Left panel: Y-axis shows GO-terms or gene set tested for enrichment. X-axis shows the enrichment of Y-axis set or terms in genes subject to pan-cancer switching vs. all genes tested for isoform switches. Significance levels were estimated by two-sided Fishers exact tests. Single asterisks (*) indicates FDR corrected $P < 0.05$. Right panel: simplified classification of the gene sets in respective enrichment test set.

C) Loss of protein domains in pan-cancer isoform switches. Y-axis shows the number of genes subject to a pan-cancer isoform switch resulting in a protein domain gain, loss or switch as indicated on by the X-axis. “Switch” indicates genes where both gain and loss occurred. Significance levels were estimated by a bootstrapping approach (see main text). Triple asterisks (***) indicates $P < 0.001$.

D) Specific protein domains are lost in pan-cancer isoform switches. Left side: X-axis shows the over-representation of the loss of a specific protein domain in the genes with pan-cancer isoform switches compared to the frequency of loss in all genes with isoform switches, as a log2 odds ratio. Y-axis shows the protein domains lost significantly more often than expected. Error bars indicate the 95% confidence interval of the odds ratio. The significance was estimated via a Fishers exact test. Triple asterisks (***) indicates $FDR < 0.001$, single asterisks (*) indicates $FDR < 0.05$. Right panel: simplified classification of the protein domains.

E) Interaction network of genes involved in pan-cancer isoform switches. The network is created by genes containing an isoform switch with

predicted consequences in at least four cancer types using the STRING database of protein-protein interactions. Only networks of genes with minimum 5 interactions are shown. Asterisks (*) indicate previously described isoform switches.

F) Example of switch identified from network analysis.

Top panel shows two analyzed isoforms for the *AKT1* gene: Genomic features are rescaled to the square root of their original size in bp. PFAM domains are indicated by greyscale. UCSC gene transcript IDs are shown for each isoform. Left bottom panel shows the average expression in healthy tissues and KIRP tumors on gene level. Middle bottom panel shows the average expression in healthy tissues and KIRP tumors on isoform level. Bottom right panel shows the average isoform fraction in healthy tissues and KIRP tumors. 'ns' indicates 'not significant' and *** indicates a *FDR* < 0.001 (EdgeR for expression tests and Mann-Whitney two sided test (as described in main text) for isoform fraction analysis). Error bars indicate 95% confidence intervals.

Figure 6: Isoform switches as predictors of patient survival

A) Isoform switches predictive of pan-cancer survival. Y-axis shows the median survival rate (top plot) and hazard ratio (bottom plot) as a barplot for each of 31 genes (X-axis) containing isoform switches predictive of pan-cancer survival. Genes in bold are shown as examples in panels B-G.

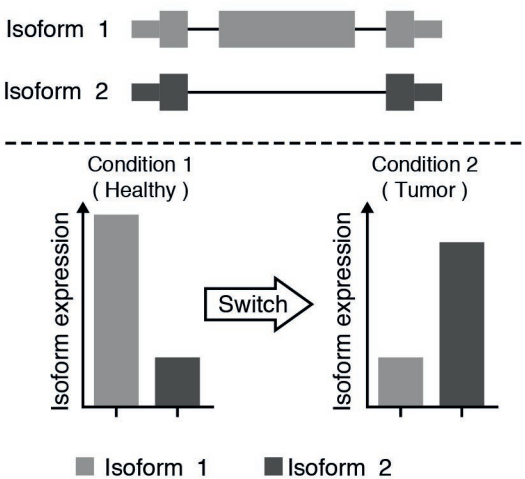
B,C,D,F) Examples of isoform switches as predictors of pan-cancer survival. Isoform switches in ZNF12 (B), ERCC1 (C), FAM36A (D) and SH1 (F) are shown. Top panel shows the two isoforms switching for respective gene. Genomic features are rescaled to the square root of their original size in bp. PFAM domains are indicated in greyscale. UCSC gene transcript IDs are shown for each isoform as well as an indication of which isoforms are used more/less. Bottom panel: A Kaplan-Meier plot showing the survival probability (Y-axis) as a function of time since diagnosis in years (X-axis). Grey lines indicates patients with an isoform switch, black lines indicate patients without it. Full lines show the results from pan-cancer analysis; dashed lines show the result from the single cancer where the isoform was the most predictive of cancer survival. Inset shows the number of events identified. *P*-values indicating the significance of the difference in survival rates (log likelihood

test) between patients with and without the isoform switch are shown. Note the *P*-values in the main text and figures are different since they measure different properties.

E,G) **Example of pan-cancer survival prediction as a function of isoform switch magnitude.** Kaplan-Meier plot (as in B) for the isoform switches in FAM36A (E) and SH1 (G) (see panel D and F) stratified by switch magnitude size as indicated by linetype. *P*-values indicating the significance of the difference in survival rates (log likelihood test) are shown. Note the *P*-values in the main text and figures are different since they measure different properties.

Figure 1

A **Illustration of isoform switch**

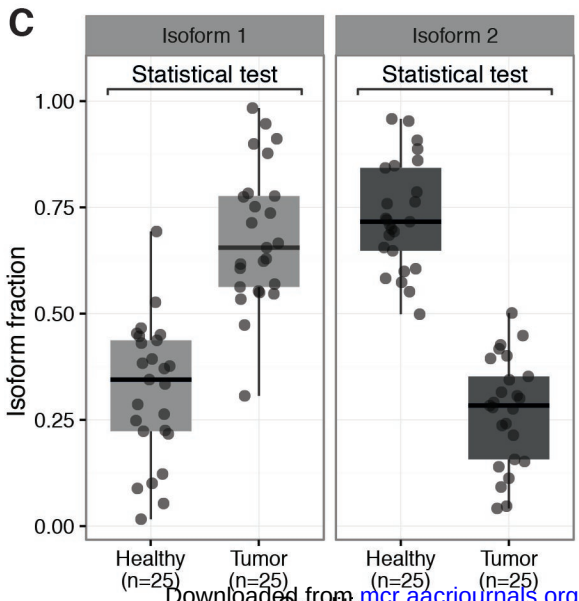


B **Example of measuring change in isoform usage**

		Expression (FPKM)	Isoform fraction(IF)
Condition 1 (Healthy)	Isoform 1	30	30/150=0.2
	Isoform 2	120	120/150=0.8
	Gene(total)	150	
Condition 2 (Tumor)	Isoform 1	120	120/150=0.8
	Isoform 2	30	30/150=0.2
	Gene(total)	150	

	IF condition 1	IF condition 2	dIF
Isoform 1	0.2	0.8	+0.6
Isoform 2	0.8	0.2	-0.6

C



Author Manuscript Published OnlineFirst on June 5, 2017; DOI: 10.1158/1541-7786.MCR-16-0459
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

Figure 2

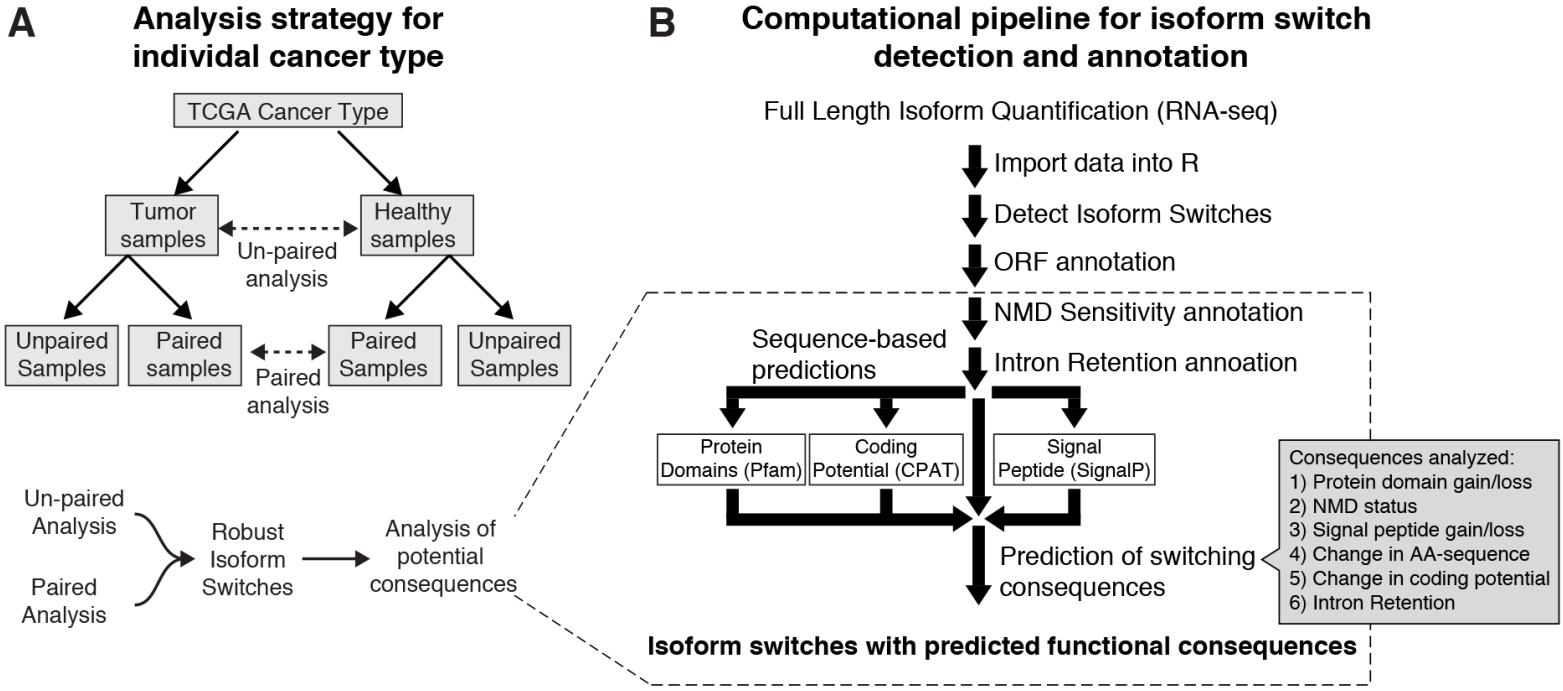


Figure 3

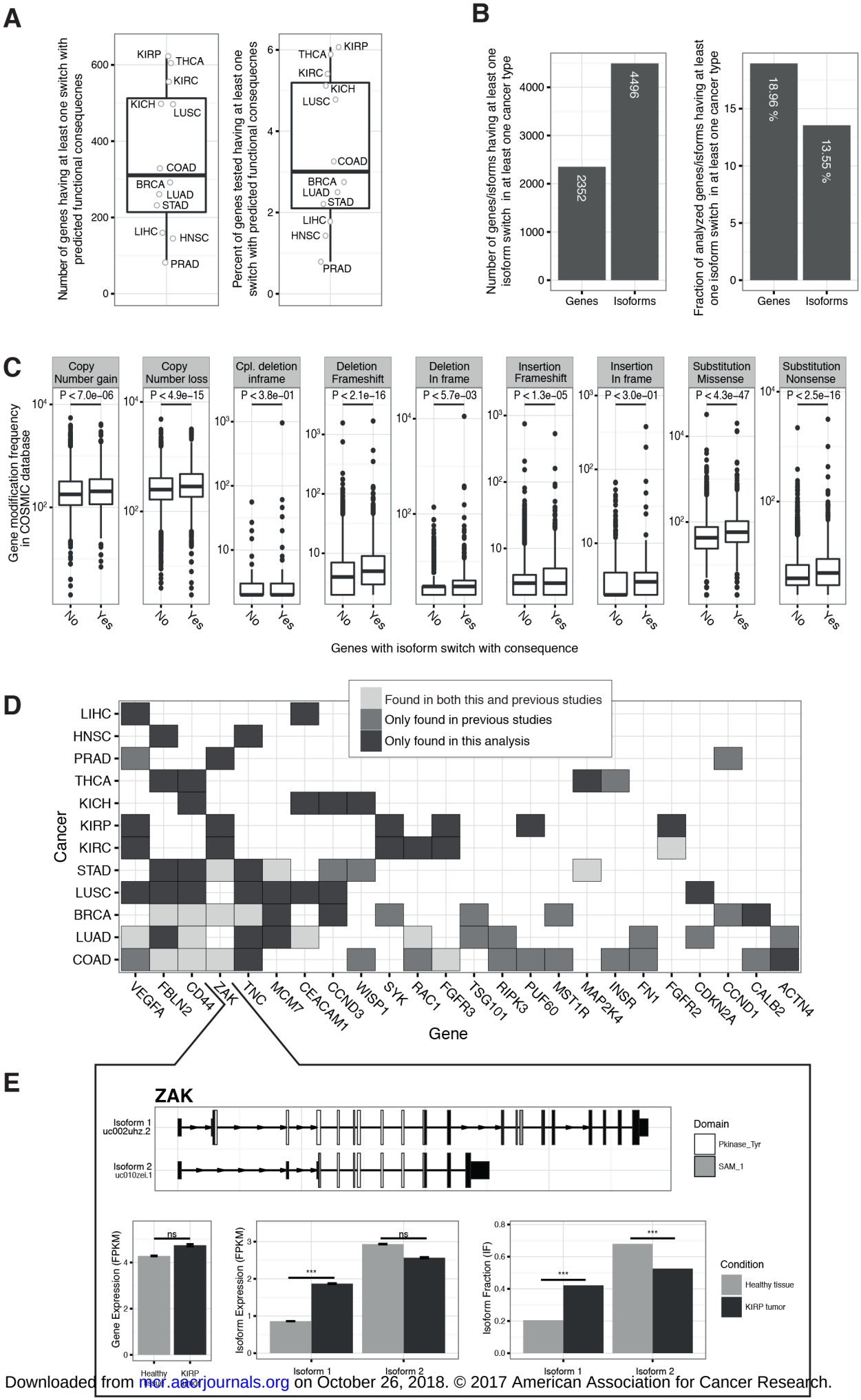
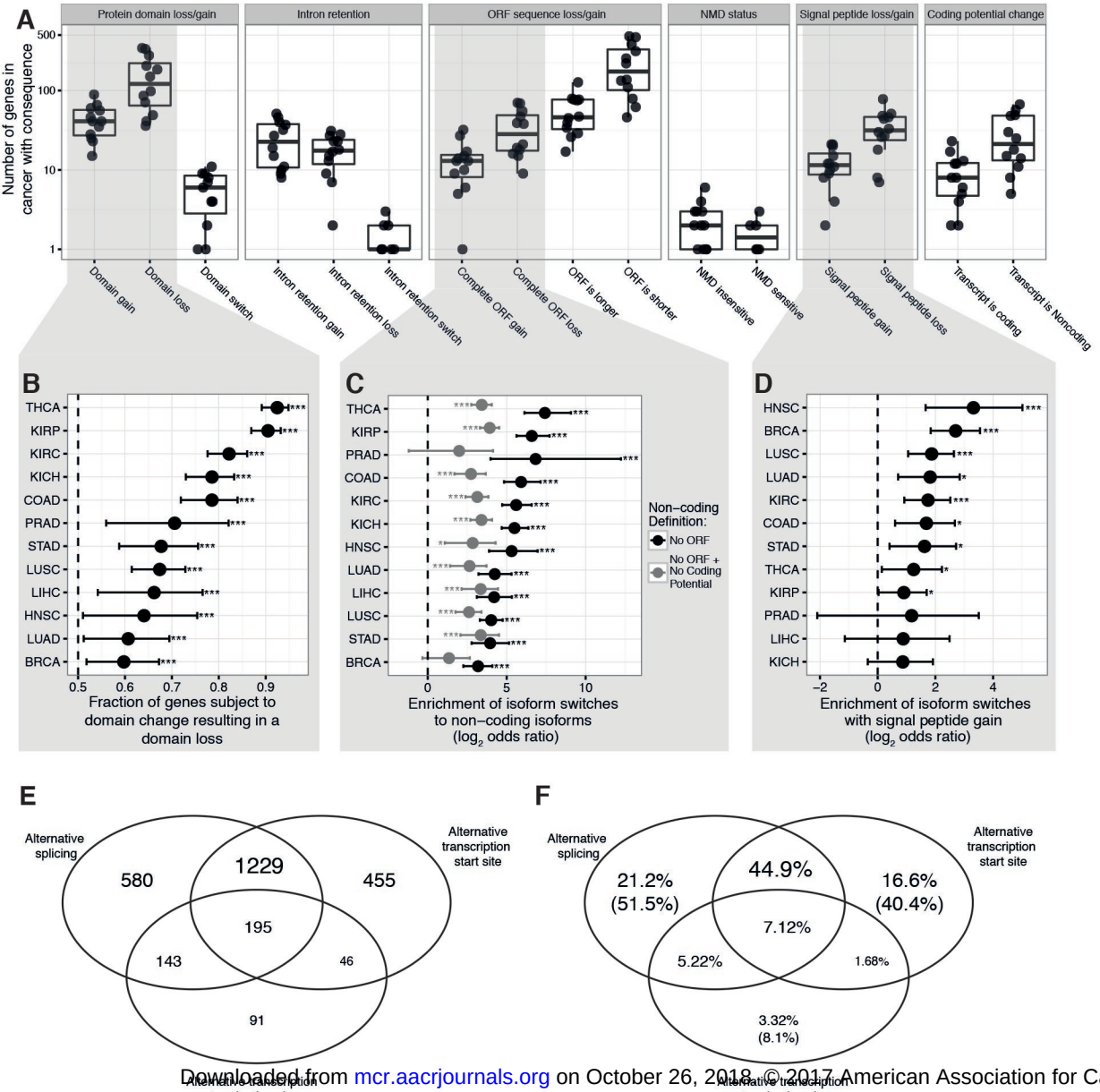


Figure 4



Author Manuscript Published OnlineFirst on June 5, 2017; DOI: 10.1158/1541-7786.MCR-16-0459
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

Figure 5

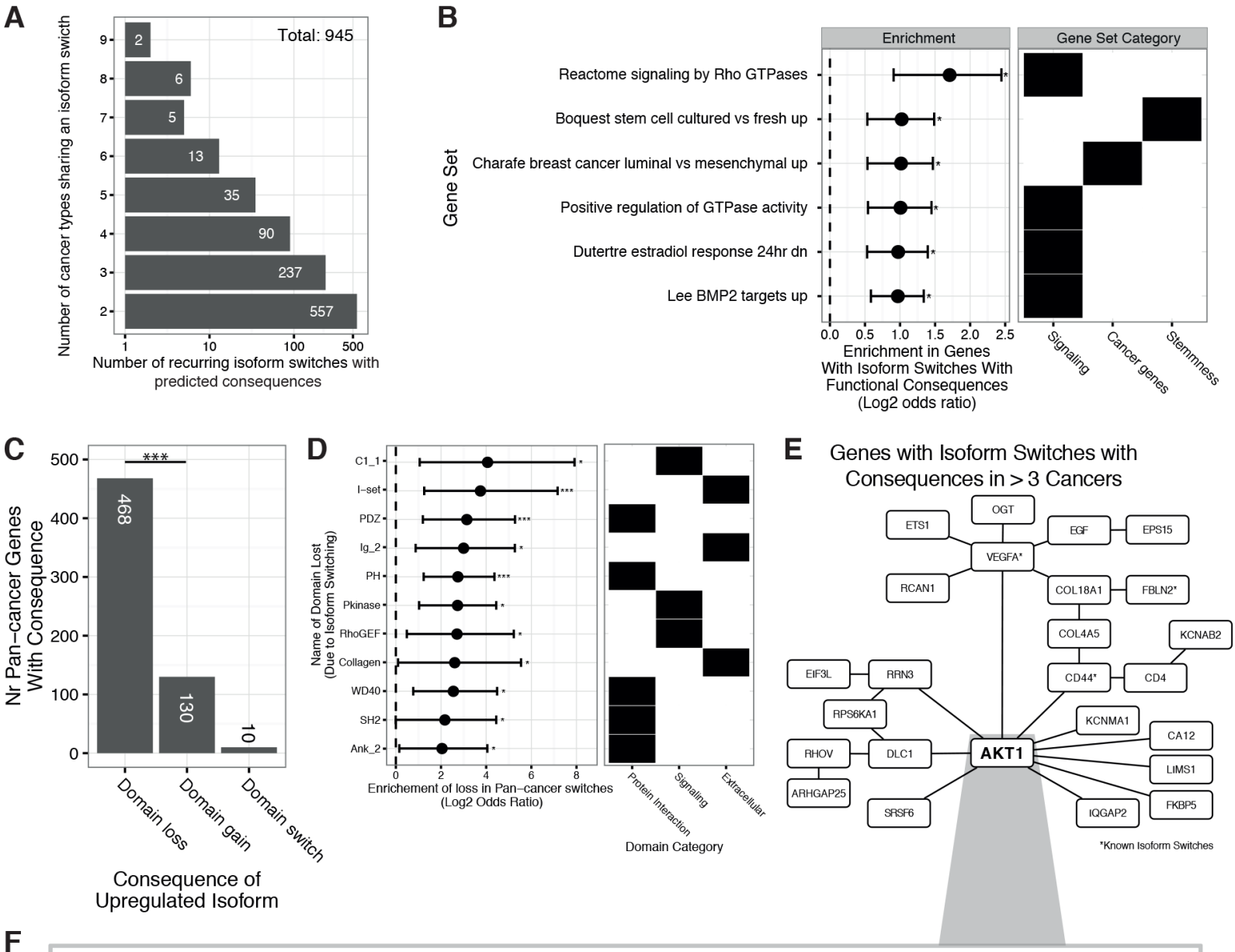
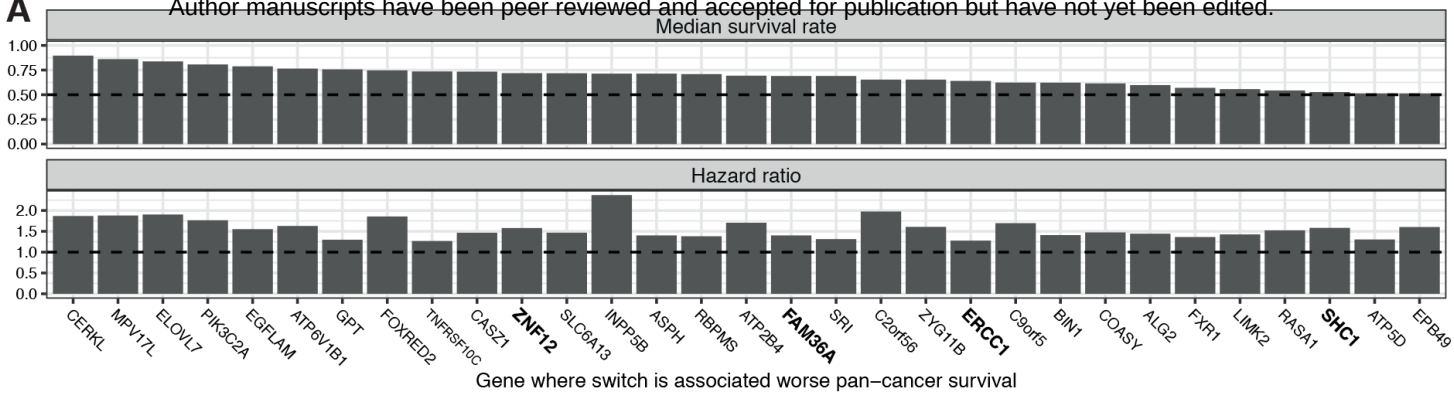
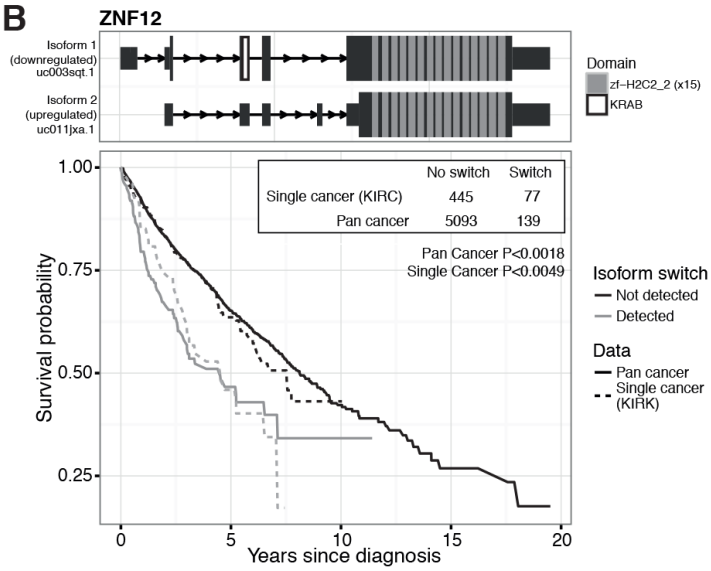
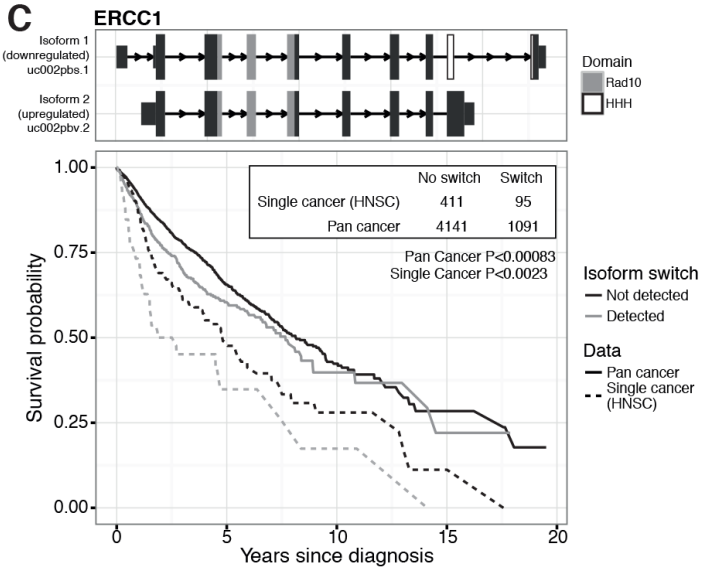
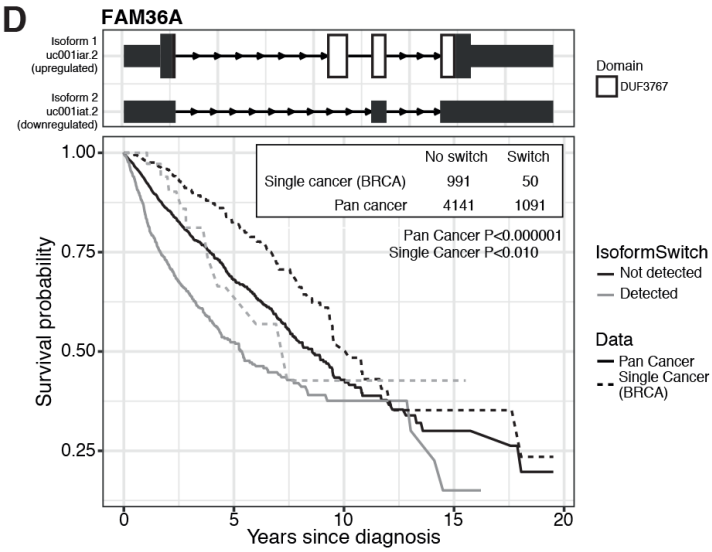


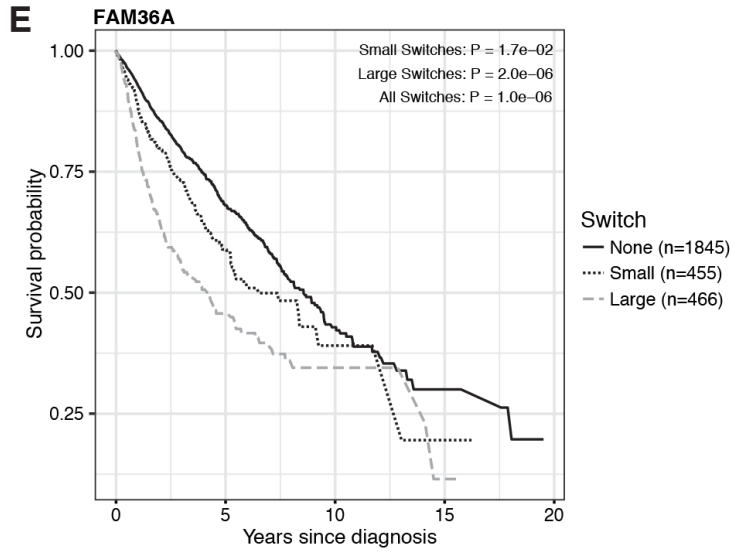
Figure 6

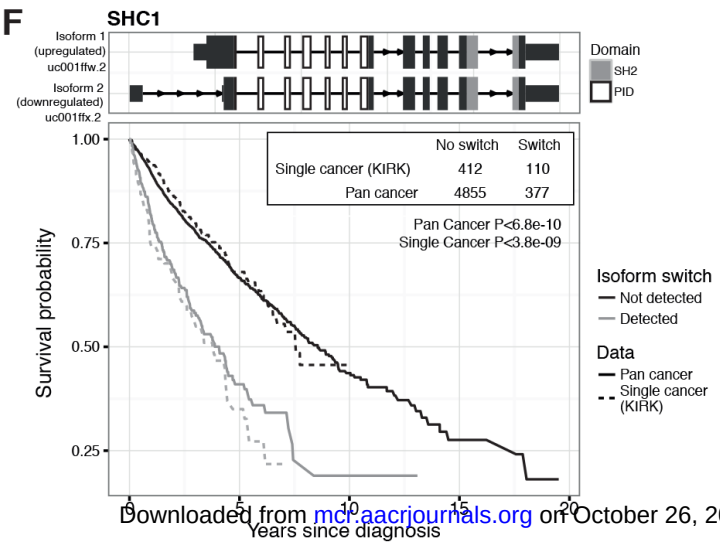


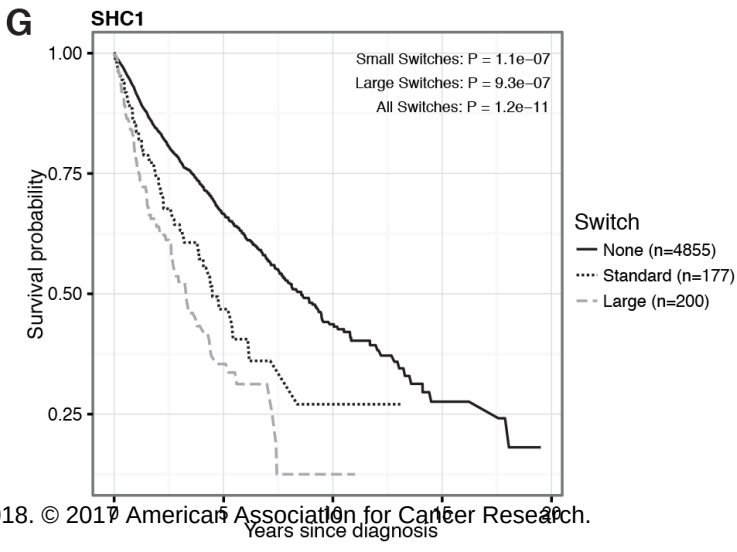












AACR American Association for Cancer Research

Molecular Cancer Research

The Landscape of Isoform Switches in Human Cancers

Kristoffer Vitting-Seerup and Albin Sandelin

Mol Cancer Res Published OnlineFirst June 5, 2017.