bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Pathogenic impact of isoform switches in 1209 cancer samples covering 27 cancer types using an isoform-specific interaction network

Abdullah Kahraman[1,2,3] and Christian von Mering[1,3]

[1] University of Zurich, Institute of Molecular Life Sciences (Zurich, Switzerland)

[2] University Hospital Zurich, Department of Pathology and Molecular Pathology, Molecular Tumor Profiling lab (Zurich, Switzerland)

[3] Swiss Institute of Bioinformatics

Abstract

Under normal conditions, cells of almost all tissues types express the same predominant canonical transcript isoform at each gene locus. In cancer, however, splicing regulation is often disturbed, leading to cancer-specific switches in the most dominant transcripts (MDT). But what is the pathogenic impact of these switches and how are they driving oncogenesis? To address these questions, we have developed CanIsoNet, a novel isoform-specific protein-protein interaction network that identifies binding domain losses and interaction disruptions in known alternatively spliced isoforms. We applied CanIsoNet on 1209 cancer samples covering 27 different cancer types from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project of the International Cancer Genomics Consortium (ICGC). Our study revealed large variations in the number of cancer-specific MDTs (cMDT) between cancer types. While carcinomas of the head and neck, and brain had none or few cMDT switches, cancers of the female reproduction organs showed the highest number of cMDTs.

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Interestingly, in contrast to the mutational load the cMDT load was tissue-specific, i.e. cancers arising from the same primary tissue had a similar number of cMDTs. Some cMDT switches were found in 100% of all samples in a cancer type, making them candidates for diagnostic biomarkers. cMDTs showed a tendency to fall at densely populated network regions where they disrupted protein interactions in the proximity of pathogenic cancer genes. A gene ontology enrichment analysis showed that these disruptions occurred mostly in enzyme signaling, protein translation, and RNA splicing pathways. Interestingly, no significant correlation between the number of cMDT switches and the number of coding and non-coding mutations was found. However, for some transcripts, we show that their expression correlates with mutations in non-coding splice-site and promoter regions of their genes. This work demonstrates for the first time the large extent of cancer-specific alterations in alternative splicing for 27 different cancer types, highlights distinct and common patterns of cMDT switches and suggests novel pathogenic transcripts and markers that induce large network disruption in cancers.

Keywords: Alternative splicing, Pan-cancer, Most-dominant transcripts, Protein-protein interaction network, Whole-genome sequencing, Pathogenicity

Introduction

Cells express on average four alternatively spliced transcripts per gene (see Ensembl database v97). The expression values follow an extreme value distribution (Hu *et al.*, 2017), where a single or a few transcripts show significantly higher expression than the remaining alternative transcripts. In most of the cases, the Most Dominant Transcript (MDT) of a gene is shared between different tissue types (Gonzàlez-Porta *et al.*, 2013; Ezkurdia *et al.*, 2015). In cancer, however, splicing regulation is often disturbed with alternative transcripts being more dominantly expressed than in normal tissues (Sebestyén *et al.*, 2015). The resulting MDT switches are known to contribute to tumor progression, metastasis, therapy resistance, and other oncogenic processes that are part of cancer hallmarks (Oltean and Bates, 2014). Exon skipping events, intron retention, or alternative exon usage can produce transcripts and proteins whose transactivation or binding domains, localization signals, active sites, stop codons or untranslated regions (UTR) are altered (Sveen *et al.*, 2015; Kelemen *et al.*, 2013). Other transcripts can even be marked for nonsense-mediated decay (Popp and Maquat,

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2018). For example, in gliomas, prostate and ovarian cancers a short Epidermal Growth Factor Receptor (EGFR) splice variant has been described to lack exon 4. The exclusion of exon 4 removes 45 amino acids from the extracellular domain of EGFR causing elevated levels of cell proliferation by ligand-independent activation and constitutive downstream signaling (H. Wang *et al.*, 2011). Alternative splicing of the BCL-X gene generates two isoforms, where the shorter isoform BCL-XS is a tumor suppressor and downregulated in prostate cancer, while the longer isoform BCL-XL is an oncogene blocking apoptosis (Lapuk *et al.*, 2014). Furthermore, melanoma tumors often develop drug resistance to BRAF(V600E) inhibitors by expressing a shorter isoform of mutated BRAF that lacks the RAS binding domain and allows BRAF(V600E) proteins to dimerize and signal in a RAS independent manner (Poulikakos *et al.*, 2011; Samatar and Poulikakos, 2014). Other splice variants are used as prognostic biomarkers in the clinics such as the Variant 7 of the Androgen Receptor (AR)-V7, which when overexpressed in hormone-refractory prostate cancer patients, correlates with poor patient survival and higher recurrence rate (B.-D. Wang *et al.*, 2017).

Fundamentally, these phenotypes arise through alterations in interaction networks (Vidal *et al.*, 2011) in which alternative splicing changes the interaction capabilities of gene products by disrupting protein binding domains or protein availability (Corominas *et al.*, 2014). The interaction landscape of alternatively spliced isoforms is often distinct from the canonical isoform, allowing cells to widely expand their protein interaction capabilities (Yang *et al.*, 2016). Earlier studies showed that tissue-specific exons were often found in unstructured protein regions, peptide-binding motifs and phosphorylation sites (Buljan *et al.*, 2012). At the same time, such exons were often part of hub genes in interaction networks, whose differential expression disrupted and promoted new protein interactions (Ellis *et al.*, 2012). The Eyras lab discovered in a recent study in over 4,500 cancer samples from 11 cancer types from The Cancer Genome Atlas (TCGA) (Network *et al.*, 2013), significant alterations in alternative splicing and MDT switches (Climente-González *et al.*, 2017). In their analysis, they were able to show an association between recurrent functional switches in MDT and the loss of protein functions while those gaining functional capabilities were mostly found in oncogenes. Furthermore, they observed that genes often mutated in various cancers were also those frequently altered in their alternative splicing patterns but often in a mutually exclusive manner. By mapping the isoform switches onto protein-protein interaction modules, they were able to show that disruptions of protein interactions mostly occurred in apoptosis-, ubiquitin-, signaling-, splicesome- and ribosome-related

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

pathways. In a similar analysis of over 5,500 TCGA samples from 12 cancer types, Vitting-Seerup et al. discovered that 19% of multiple transcript genes were affected by some functional loss due to isoform switching (Vitting-Seerup and Sandelin, 2017). They identified 31 switches that had prognostic biomarker qualities, predicting patient survival in all cancer types.

Cancer-specific MDTs are believed to be fundamentally caused by genomic mutations. Splicing Quantitative Trait Loci (sQTL) calculations in which exon expression is linearly correlated with mutations in nearby cis-regions or distant trans locations are supporting this hypothesis. For example, over half a million sQTLs were measured in whole blood samples of which 90% were located in intergenic and intronic regions (Xiaoling Zhang *et al.*, 2015). Over 520 sQTLs were associated with disease phenotypes from previous Genome-Wide Association Studies (GWAS). Interestingly, 395 GWAS associated SNPs overlapped with cis-sQTLs whose genes were not differentially expressed, giving additional insights into the functional mechanism of GWAS results. An independent Pan-Cancer Analysis of Whole Genomes (PCAWG) analysis group that focused on cancer transcriptomes found over 1800 splicing alterations, which correlated with nearby mutations in intronic regions (PCAWG Transcriptome Core Group *et al.*). They identified over 5,200 mutations mostly located in or near splice-sites which had a major impact on the expression of cassette exons. Interestingly, only 4% of these mutations increased splicing efficiency, while the large majority had negative effects on splicing.

Building further on these studies, we are presenting here the most comprehensive functional assessment of alternative spliced transcripts to date, covering the transcriptome and matched whole-genome sequences of 1209 cancer samples from 27 different cancer types from the PCAWG project of the International Cancer Genomics Consortium (ICGC) (Figure 1). Our study extends the aforementioned studies by an additional 16 cancer types from 10 primary tissues most notably cancers of the brain, blood, female reproductive organs, and melanoma. Brain cancers had particularly little deregulation. Similarly, melanoma showed little changes in the alternative splicing despite having the highest mutational burden. In contrast uterus, ovary and cervix cancers had the highest number of splicing alterations. We based our study on the hypothesis that alternative transcripts are pathogenic when they disrupt protein interactions and pathways. To test this, we focused on cancer-specific switches in MDTs and assessed the extent to which these rewire protein-

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

protein interactions - using a novel isoform-specific protein interaction network that we termed *CanIsoNet*. Our analyses revealed a large diversity in the number of cMDTs between cancer types, most of which were tissue specific. Some cMDT were found in all samples of a cancer type but not in any sample of a matched normal cohort, which makes them ideal candidates for diagnostic biomarkers. We show large scale disruptions of protein-protein interactions that are enriched in enzyme signaling, protein translation, and splicing pathways. We provide evidence that some cMDTs were likely pathogenic, given their proximity to cancer-related genes and their location in densely populated PPI network regions. Finally, we present correlation data between somatic mutations and transcript expressions.

Methods

Scripts, input files and a step-by-step instruction to reproduce the presented analyses can be found at https://github.com/abxka/CanIsoNet.

Accessing pan-cancer analysis of whole genomes data

Transcript isoform-specific expression levels for 1393 Pan-Cancer Analysis of Whole Genomes (PCAWG) samples (syn7536587) and 3249 Genotype-Tissue Expression (GTEx) (Lonsdale *et al.*, 2013) samples (syn7596599) were provided by the PCAWG project for download from a dedicated Synapse database (https://www.synapse.org). Expression levels were given in Transcript Per Million (TPM) counts computed for all known transcripts in Ensembl version 75 using Kallisto (v.0.42.1) (Bray *et al.*, 2016) with default parameters (see Method section in PCAWG Transcriptome Core Group paper (PCAWG Transcriptome Core Group *et al.*), for more details). 1209 PCAWG samples remained after selecting those labeled as whitelisted and as a tumor in the RNAseq metadata file (syn7416381). 2232 GTEx samples remained after selecting those matching primary tumor tissues from PCAWG RNAseq cancer samples using metadata on GTEx samples (syn7596611) (see Table S1).

Coding and non-coding mutation calls from the independent PCAWG working group “Novel somatic mutation calling methods” were downloaded from Synapse (syn7118450). Only mutations located in functional regions (i.e. promoter core, promoter domain, 5'UTR, coding sequence, splice site, 3'UTR)

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

were taken into consideration. Information on the genomic location of functional regions was also downloaded from Synapse (syn7345646).

Constructing an isoform-specific protein-protein interaction network

The implementation of an isoform-specific protein-protein interaction network is primarily based on the integration of functional interaction from the STRING database (Szklarczyk *et al.*, 2015) with physical domain-domain interactions from the 3did database (Stein, 2004) and known alternatively spliced protein isoforms from the Ensembl database (Hubbard *et al.*, 2002) (see Figure 1). For the integration, we downloaded all human functional protein-protein interactions and the FASTA sequences of all canonical isoforms from the STRING database (version 10.0)(Szklarczyk *et al.*, 2015). To identify physical interactions, we downloaded the 3did database (version 2018_04) (Mosca *et al.*, 2014), which lists pairs of PFAM domains (Punta *et al.*, 2012) that are physically interacting with each other in the Protein structure Data Bank (PDB) (Berman *et al.*, 2007). For integrating the STRING database with 3did, we received PFAM domain annotations for each STRING protein from the STRING developer team, which we extended with PFAM information for human from the PFAM database itself (version 32.0). To guarantee that no PFAM assignment was missed, we additionally ran the pfam_scan.pl script (available on the PFAM FTP server) on all Ensembl protein isoforms. STRING proteins whose sequences were not identical to their sequence in Ensembl were discarded. STRING interactions with proteins having interacting PFAM domains in 3did were considered to be of physical nature. Only high-confidence interactions with a STRING combined score of ≥ 900 were used in this study. For the next step of testing the remaining alternative isoforms for the existence of the PFAM domain in their protein sequence, the sequence of the PFAM domain in the canonical isoform was extracted from the STRING FASTA file. If the same sequence existed in an alternative isoform, the interaction was assumed to persist, otherwise, we assumed the interaction to be lost due to alternative splicing. A table representing a database of human isoform-specific protein-protein interaction with information which interactions are lost, and which persist for alternatively spliced isoforms and transcripts, can be found in Table S2.

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Identifying most dominant transcripts

For assessing the impact of disrupted alternative splicing in cancer, we chose to focus on the most extreme alteration events, namely on those cases in which the identity of the Most Dominant Transcript (MDT) in a PCAWG sample is unique and not known to exist in matched cohorts of GTEx normal samples. We call these most dominant transcripts cancer-specific MDT (cMDT). Note, that in this study the term transcript and protein isoform is interchangeable. We worked only with transcripts that had a protein ID (ENSP) in the Ensembl database (see Constructing an isoform-specific protein-protein interaction network).

To identify MDTs in any PCAWG and GTEx sample, Kallisto counts in Transcripts per Million (TPM) were extracted for each Ensembl transcript from files provided by the PCAWG Transcriptome Core group (see Accessing pan-cancer analysis of whole genomes data). For each gene, all transcripts were ordered by their TPM counts. The transcript with the highest TPM count was designated as MDT if its expression was at least twice as high as the TPM count of the 2nd ranked transcript. Transcripts having an NA value were assigned a TPM of 0. MDTs were required to have a minimum TPM value ≥2, which corresponded to the maximum expression value of 99% of olfactory receptor proteins. As the expression of olfactory transcripts is known to be limited to nasal tissue only, we used their expression as a threshold for separating background noise in the PCAWG and GTEx RNAseq data (Ezkurdia *et al.*, 2014).

Identifying cancer-specific most dominant transcripts

Once all MDTs in each PCAWG and GTEx sample were determined, we next checked whether the MDTs in the cancer samples were unique and specific to PCAWG, in which case we designated them as cancer-specific MDT (cMDT). To qualify as a cMDT, an MDT must

1.) be found in PCAWG
2.) not be an MDT in any of the samples of the matched GTEx cohort (see Table S1)
3.) have an alternative transcript that is an MDT in any of the samples of the matched GTEx cohort
4.) derive from a gene that has an MDT in at least 50% of samples from the matched GTEx cohort.

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

5.) have a significantly different relative expression than in GTEx

Note that for point 5, we used a sign-test to assess the significance. For the sign test, we counted the number of times the relative expression of an MDT in a cancer sample was higher or lower to all relative expression values in the samples from the matched GTEx cohort. The positive and negative counts were subsequently put in a two-sided binomial test and the p-value was calculated. After the p-values for all MDTs in a cancer type were determined, they were subjected to a Benjamini-Hochberg FDR correction. An MDT that fulfilled all the 5 criteria above and had a q-value of < 0.01 qualified as a cancer-specific MDT (cMDT).

Predicting the pathogenic impact of cancer-specific MDT (cMDT)

To predict the pathogenic impact of cMDTs, we assessed their proximity to 723 genes from the COSMIC gene census list (version 89) in the STRING interaction network and checked whether they were located in densely populated network regions, following the idea that cMDT might interact with known cancer genes or their interaction partners and effect numerous network interactions. For the cancer gene proximity calculations, we computed the shortest path in the STRING interaction network between a cMDT and all known genes in the COSMIC Cancer Gene Census (CGC) using a breadth-first-search algorithm (Kahraman *et al.*, 2011).

For assessing the interaction density at each node A of the STRING network, we computed a Network Density Score (NDS) using the following equation:

$$NDS(a) = int(a) + \sum_{s=1}^{3} \sum_{b=1}^{B} \left(\frac{1}{2^s} * int(b_s) * score(b_s, b_{s-1}) \right)$$

where *a* is the protein of interest, *b* is an interactor being *s* interaction nodes apart, *int()* is the number of interactors of *b* and *score()* is the STRING combined interaction score between *b* and its interaction partners. *B* is the maximum number of interactors of *a*. To put a meaning to raw NDS values, we ordered all STRING proteins by their NDS value and assigned each value a relative rank position (0 - 1.0) within the ordered list. The highest density had the guanine nucleotide binding protein GNB1 while the lowest density was observed amongst other for the ankyrin repeat protein ANKS6.

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

STRING gene ontology enrichment analysis

A hypergeometric test was used to determine the enrichment of disrupted interactions in biological processes from Gene Ontology (GO) (Ashburner *et al.*, 2000). The statistical test was performed using the STRINGdb R package (version 1.22), with a score threshold of 0, STRING database version 10.0, species identifier 9606, the category GO biological processes, FDR multiple testing correction and the parameter Inferred from Electronic Annotations set to true. The test was performed for each subnetwork of disrupted interactions of a PCAWG cancer sample. A subnetwork contained proteins of disrupted interactions that were overlapping with one or both interaction partners. Only the most significant biological process was selected for each subnetwork. The remaining processes were ignored.

Correlation between somatic mutations and transcript expression

Similar to expressed quantitative trait loci calculations, in which the expression of a gene is correlated with mutations in the proximity or distance, we performed a correlation analysis between transcript expression and mutations located within the associated gene. The correlation analysis was performed only within a cancer-type, to reduce biases from confounders. The expression of a transcript was assigned to the group *Mutated* if its gene was found to carry a mutation in the promoter, 5'UTR, coding sequence, 3'UTR and splice sites. Otherwise, the expression of the transcript was added to the group *Wildtype*. A non-parametric Wilcox-rank sum test was performed to test the difference in the expression values between both groups. Once the difference was tested for all transcripts with expression values in both groups, the p-values were corrected using the Benjamini-Hochberg FDR method.

Results

The goal of this study was to identify common patterns in the choices of "Most Dominant Transcripts (MDT)" of 27 different cancer types while testing for their pathogenicity and disruptive nature in protein-protein interaction networks (Figure 1). On a median average, 37 samples per cancer type were available with Kidney Renal Cell Carcinomas (Kidney-RCC) having most samples (117x) and Cervix adenocarcinoma (Cervix-AdenoCA) and undifferentiable Lymphoma (Lymph-NOS) having only

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2 samples (see Figure 2). The latter two cancer types were discarded for in-depth analysis due to their small cohort size. A detailed data file listing all detected cancer-specific MDTs, the disrupted interactions and a rich set of annotations is available in Table S3.

PCAWG samples with cancer-specific most dominant transcript switches

In each of the 1209 samples, the MDT of each gene was determined and subsequently its expression compared to the expression of the transcript in related GTEx samples (GTEx samples originating from the same primary organ as the PCAWG sample) (see Figure 2A and Table S1). An MDT switch was called when the MDT was unique to PCAWG and its expression significantly different (higher) than the median expressions in related GTEx samples. We called these MDT switches cancer-specific MDTs (cMDT). Following these rules, we detected 11,040 unique cMDTs from 7,143 genes that underwent a total of 122,051 cMDT switches in all 1209 PCAWG samples, with a median average of 58 cMDTs per sample. The highest number of cMDTs was detected in cancers of female reproductive organs. The mean average number of cMDTs per sample was 322 for uterus adenocarcinoma (Uterus-AdenoCA), 101 cMDTs for cervix squamous-cell carcinoma (Cervix-SCC) and 129 cMDTs for ovarian adenocarcinoma (Ovary-AdenoCA) (see Figure 2B). On the other side, none of the 42 Head and Neck Squamous Cell Carcinoma samples (Head-SCC) showed any cMDTs, while B-cell Non-Hodgkin Lymphomas (Lymph-BNHL) had only 6 cMDT switches per case. Interestingly, melanoma samples had only a few more cMDTs with a median of 10, despite having the highest median mutational burden (see C). In contrast, Pancreas-AdenoCA had only a few mutations, while having the second highest number of cMDTs in the dataset. Another observation we made is that the cMDT load, i.e. the number of cMDTs in a cancer sample, was tissue specific. Cancer types originating from the same primary tissue, e.g. CNS-GBM and CNS-Oligo, Lymph-BNHL and Lymph-CLL, Liver-HCC and Biliary-AdenoCA, tended to have the same cMDT load. Interestingly, the tissue specificity existed only for cMDTs and not for the mutational load (compare Figure 2B and 2C). This is not surprising as gene expression programs especially those defining tissue-identity persist through neoplastic progressions in cancer cells (Bradner *et al.*, 2017). Furthermore, in about 50% of cases, it is the same cMDT that is overexpressed in different cancer types (data not shown). For those cMDTs affecting known cancer genes, most were found in tumor suppressor genes followed by fusion and oncogenes,

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

which follows the expectation as tumor suppressor genes are also the most frequent genes in the COSMIC Cancer Gene Census.

In summary, these results highlight the large variation in the cMDT load in different cancer types, which is tissue-specific and difficult to predict using genomics data only.

Most dominant transcript switches as diagnostic biomarkers

22 transcripts from 19 genes were found to have an cMDT switch in 100% of the samples of a cancer type, while the large majority with over 75% showed an cMDT switch in up to 10% of samples of a cancer type (see Table S4). Due to the omnipresence of the 22 transcripts, they are likely playing an important role in the seven cancer types in which they occurred and could serve as a diagnostic biomarker (Danan-Gotthold *et al.*, 2015; Vitting-Seerup and Sandelin, 2017). Among the 22 transcripts, the one with the highest number of cMDTs was the kinetochore protein named Nuclear Division Cycle (NDC)80 or HEC1 (Highly Expressed in Cancer). 100% of Bone-Leiomyo and Breast-LobularCA cases, 98% of Breast-AdenoCA, 96% of CNS-GBM and Bladder-TCC cases, and 95% of Ovary-AdenoCA expressed mostly the long transcript ENST00000261597 of the NDC80/HEC1 gene (see Figure 3A). In the associated normal tissues from the GTEx project, the short transcript ENST00000576274 that lacks the coiled-coiled protein domains for protein complex and mitotic spindle formation (Valverde *et al.*, 2016), was found to be mostly expressed. In hepatocellular and colorectal carcinomas, NDC80/HEC1 was recently shown to be highly overexpressed while its knockdown in cancer cell lines led to apoptosis and cell cycle arrest (Ju *et al.*, 2017; Yan *et al.*, 2018). Our results indicate that the function of the *NDC80* gene is not only regulated by its total gene expression but also by the distinct distribution of its alternative transcripts, with non-functional transcripts expressed primarily in normal cells while fully functional transcripts mostly expressed in cancer cells. The overexpression of the long transcript could serve as a biomarker for the aforementioned cancer types.

Similarly, the full-length transcript of chromosome-associated kinesin KIF4A was expressed in 100% of both Breast cancers, 96 % in Bladder cancers and in 93% Ovarian-AdenoCA cases where its expression indicated mitotic activity. In the associated GTEx normal tissues a shorter transcript that lacked the last three exons encoding a C-terminal cysteine-rich motif was mostly expressed. The

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

cysteine-rich motif is critical for chromatin binding and functioning of KIF4A in mitosis (G. Wu and Chen, 2008; Almeida and Maiato, 2018).

Also, the shorter transcript ENST00000300403 (747 Amino Acids (AA) long) of the microtubule-associated TPX2 gene was found to be the cMDT in 100% of samples in Breast-AdenoCA and Breast-LobularCA, while its longer sister transcript ENST00000340513 (783 AA) was the primary isoform in 67% of GTEx samples of the breast. Similar ratios were found for Uterus-AdenoCA (98%) and Bladder-TCC where the shorter transcript was the cMDT in at least 91% of the cancer samples (see Figure 3B). The longer isoform in GTEx samples has an additional 36 AA after position 351, which due to its proximity to the N-terminus of the TPX2_importin domain (361-489 AA) might interfere with its function of binding its inhibitor importin-alpha (see Figure 3B). In complex with importing-alpha, TPX2 moves from the nucleus to the cytoplasm to regulates mitotic spindle formation (Pérez de Castro and Malumbres, 2012). Thus, besides the expression of TPX2, the transcription of the longer isoform might be another regulatory principle to inhibit TPX2 functionality in the cell.

Another candidate cMDT biomarker which was also one of the most frequent cMDTs in our study, was the transcript ENST00000366999 of the gene *Never in mitosis A-related kinase 2* (NEK2). The transcript also known as NEK2A was detected as the cMDT in 100% of Cervix-SCC samples, 91% of Bladder-TCC samples and in 90% of Ovary-AdenoCA samples (see Figure 3C). In GTEx normal samples the MDT was mostly the shorter transcript ENST00000366998 (NEK2B). Both isoforms differ in their C-terminal domain, with NEK2A having an additional coiled-coil region to bind the E3-ubiquitin ligase Anaphase Promoting Complex (APC/C), which induces its degradation in the M-phase of mitosis (Xia *et al.*, 2015). Reassuringly, NEK2A was shown to be overexpressed in many cancer types including breast, lung, colon, and as here, in ovarian cancers (Fang and Xiongwen Zhang, 2016).

In summary, our analysis highlights the existence of cancer-specific transcripts whose dominant expression could serve as a diagnostic biomarker for clinical applications.

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Cancer-specific most dominant transcripts disrupting protein-protein interactions

Of the aforementioned 7,143 genes, 2,638 genes had not a single high-quality Protein-Protein Interaction (PPI, STRING combined score ≥ 0.9). But for the remaining genes with 122,051 cMDT switches, we found 461,437 high-quality PPI of which 28% (129,496) were found to be disrupted due to cancer-specific MDT switches. The 129,496 interactions corresponded to 12,885 unique PPIs originating from 1410 different proteins in 990 different samples. Among the unique PPI interactions, only 14% of cMDTs disrupted one or more PPI of the canonical isoform. However, in 88% of these cases the cMDTs were predicted to disrupt all known PPIs of their associated genes (see Table 1). The high percentage of total PPI losses can be explained by the fact that proteins often interact via the same binding domain (Keskin *et al.*, 2016).

The most frequent cMDTs disrupting interactions in over 90% of samples of a cancer type were those from the genes USP46, WDR74, RPS19, BOLA2B, NDUFA9, and LAMA3 (see Table S5). Three of the genes whose protein products have a PDB structure available with their interaction partner are shown in Figure 4. Most of the disrupted interactions were found in Uterus-AdenoCA, which had on average 86.2 disrupted interactions per sample, followed by Eso-AdenoCa and Cervix-SCC with 45.1 and 44.1 disrupted interaction, respectively. In contrast, cancers of the central nervous system and Lymph-BNHL had on average less than a single disrupted interaction per sample (see Table 2).

For the Ubiquitin carboxyl-terminal hydrolase 46 (USP46) that plays a role in neurotransmission, histone deubiquitination and tumor suppression (Li *et al.*, 2013), the most frequent transcript in cancer cells was ENST00000451218, which lacked the second exon of the GTEx transcript ENST00000441222. The exon skipping event removes a beta-strand from an N-terminal two-strand beta-sheet in the palm motif of USP46 (see Figure 4A), which has dramatic effects on the protein conformation (Birzele *et al.*, 2008). Besides, the spliced-out beta-strand is part of the interaction interface with the Polyubiquitin-B (UBB) protein. UBB stabilizes the finger motif of USP46, which is used by USP46 to interact with its allosteric activator WD repeat-containing protein 48 (WDR48) (Yin *et al.*, 2015) and other proteins (see Figure 4A). Thus, the cancer-specific expression of the transcript ENST00000451218 is disabling the tumor suppressor function of USP46.

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In another example, an alternative promotor region in the ribosomal protein S19 (RPS19) gene induced the expression of the cancer-specific 71 AA short MDT ENST00000221975 in 100% of Panc-AdenoCA, 98% of Uterus-AdenoCA and 93% of Stomach-AdenoCA (see Figure 4B). A longer 145 AA transcript ENST00000593863 was mainly expressed in the matched GTEx tissues. The alternative promoter caused an elongation of the 5'UTR region, which led to the removal of the first 74 N-terminal AAs in the cancer-specific isoforms. The N-terminal region of RPS19 holds the entirety of the interaction interface with RPS16. Thus, the interaction between RPS19 and RPS16 as well as to RPS5, RPS18, MRPS7, and MRPS9 was lost in the effected 240 cancer samples in PCAWG. A fully functioning RPS19, however, is required for the E-site release of tRNA and the maturation of 40S ribosomal subunits (Flygare *et al.*, 2007). Truncating mutations in ribosomal proteins are known to cause cancer (Goudarzi and LIindström, 2016) or syndromes like the autosomal inherited Diamond-Blackfan anemia (Flygare *et al.*, 2007).

The mitochondrial NADH dehydrogenase (ubiquinone) alpha subcomplex 9 (NDUFA9) was also mainly expressed via an alternative promoter in 93% Uterus-AdenoCA samples and 35% of ColoRect-AdenoCA samples. As a result, the cancer-specific transcript ENST00000540688 had a length of only 136 AA and lacked the first 235 AA of the longer GTEx-specific transcript ENST00000266544. But not only was ENST00000540688 shorter, the first 57 AA were also encoded by an alternative exon, making the N-terminus of the cancer-specific transcript distinct from the MDT in GTEx. As a result, most of the canonical protein sequence including the binding site sequence for NDUFS7 was missing in the cancer-specific transcript of NDUFA9 (see Figure 4C). Thus, in the 66 PCAWG cancer samples, NDUFA9 is not able to interact with NDUFS7, which will destabilize the structure and function of the Respiratory complex I, impacting the electron transfer from NADH to ubiquinone. Various germline mutations in NDUFA9 are known to cause severe neurological disorders (Baertling *et al.*, 2018). In breast cancer cell-lines, dysregulation of the NAD+/NADH balance was found to correlate with enhanced cancer progression (Santidrian *et al.*, 2013). Thus, we postulate that the short cancer-specific NDUFA9 transcript causes mitochondrial respiratory defects, which could promote aerobic glycolysis in the effected cancer cells leading to cancer progression (Srinivasan *et al.*, 2016).

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

An enrichment analysis on the disrupted interactions using Gene Ontology biological processes revealed that 9% of disrupted interactions were mostly impacting “Enzyme linked receptor protein signaling” pathways, followed by “Translational termination” with 5%, “Transmembrane receptor protein tyrosine kinase signaling” pathways with 4% and “RNA splicing” with 2%. Most of the disruptions were due to losses of Protein-kinase domains, WD40 repeat domains and Pleckstrin homology domains, which were also the most frequent domains in our STRING-3did interaction network. Translational initiation, but in particular also translational termination, were impacted in most cancer types, while Enzyme linked receptor protein signaling pathways were disrupted mainly in Uterus-AdenoCA, Ovary-AdenoCA and Kidney-RCC, which set them apart from the rest (see Figure 4D). In contrast, Prost-AdenoCA, Skin-Melanoma, Lymph-BNHL, Breast-LobularCA, and CNS-Oligo had a disruption in less than 4 pathways within the top 50 disrupted pathways in the PCAWG dataset (see Figure 4D). The ribosomal proteins RPS19, RPLP0, and RPL13 were among the topmost frequent proteins whose cMDT switch disrupted interaction in 181 to 240 different samples, most often in Uterus-AdenoCA (82x), Panc-AdenoCA (75x), Kidney-RCC (69x), and ColoRect-AdenoCA (58x) (see Table S6). The high frequency was also evident from the relatively high number of the disrupted ribosomal protein domains Ribosomal_S19e and Ribosomal_60s (top 8 disrupted domains), which are usually found beyond the top 470 domains in the STRING-3did database.

In summary, our results highlight extensive PPI network disruptions by cMDTs mainly impacting signaling, translational and RNA splicing pathways.

Pathogenic disruptions of protein interactions due to alternative splicing

An indication that the cMDTs in the PCAWG dataset were pathogenic to various degrees, came from an analysis where we assessed the edgetic distances of cMDTs to the closest COSMIC Cancer Gene Census (CGC) gene in the STRING interaction network. As the STRING interaction network is built on canonical isoforms, we used the associated canonical isoform of cMDTs gene in the PCAWG dataset. The distance distribution was compared to a random distribution that was generated by selecting a random protein from all expressed proteins in a cancer type. Figure 5A shows a clear preference of cMDTs to be located close to CGC genes. 58% of cMDTs were CGC gene themselves or direct interaction partners. The preference even increased for cMDTs that disrupted protein interactions and found its maximum with cMDTs that are located at densely populated regions of the

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

STRING interaction network. The last observation is expected as PPI networks are biased towards disease-associated genes that are generally more studied than non-disease causing genes (Schaefer *et al.*, 2015).

Nonetheless, cMDT switches that lead to the disruption of many protein-protein interactions are likely more pathogenic than cMDT switches that disrupt few interactions. We have therefore computed a network density score (NDS), which estimates the number of interactions of a protein and its neighborhood in a protein interaction network. Plotting the ranked NDS values of proteins and their interactors from the COSMIC Cancer Gene Census (CGC) versus the remaining non-CGC proteins, showed the aforementioned tendency of CGC proteins to be located at denser network regions than non-CGC genes (see Figure 5B). Of the 1,276 genes that had the highest 30% NDS, 124 were CGC genes, which corresponds to 53% of all CGC genes for which we detected an cMDT switch in the PCAWG dataset. The remaining 1152 genes were non-CGC genes, from a pool of a total of 4270 non-CGC genes with an cMDT switch in the PCAWG dataset. The large majority of the 1152 non-CGC genes were direct interactors of CGC, with only 79 having a distance > 1 in the STRING interaction network.

For the following analysis, we selected interesting examples from the top 30% NDS. One of the most disrupted interactions in the PCAWG dataset was between the regulatory and scaffolding subunit of the PP2A complex. This interaction is located within the 16% of densest network regions in the STRING interaction network. In 75% of Uterus-AdenoCA, 39% of Cervix-AdenoCA, 24% in Colon-AdenoCA and 17% of Ovary-AdenoCA a shorter isoform of regulatory PP2A subunit PPP2R5D (ENST00000230402) was expressed that lacked the first N-terminal 80 AA and 18 AA within the B56 binding domain of the GTEx-specific isoform (ENST00000485511) (see Figure 5C). The B56 binding domain, however, is central to the interaction of the regulatory subunit with the scaffolding subunit PPP2R1A and the catalytic subunit PPP2CA. The B56 binding domain is build up by ankyrin repeats; a common protein-protein binding motif in nature (Jernigan and Bordenstein, 2015). The deletion of an ankyrin repeat segment is likely destabilizing the domain (Tripp and Barrick, 2004), which disrupts the structure and binding capability of PPP2R5D. The disruption has likely an oncogenic effect given that PP2A is known as a tumor suppressor and any disruptions in the function of PP2A can lead to cell motility, invasiveness, and loss of cell polarity (Seshacharyulu *et al.*, 2013).

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The most frequent cMDT switches among the COSMIC cancer genes were observed for the E3 ubiquitin-protein ligases FBXW7 and MDM2, and the Cyclin Dependent Kinase CDK4. The F-box/WD repeat-containing protein 7 (FBXW7) is part of the Skp, Cullin, F-box (SCF) complex and is known to be a tumor suppressor. It ranks in the top 25% of the densest STRING network regions. In 37% of Panc-AdenoCA, we found a short isoform (ENST00000604872) mostly expressed that only consisted of the N-terminal region of the canonical isoform, lacking the F-box and WD40 repeat domains. The SCF complex without a functioning FBXW7 protein is unable to degrade cyclin E, which causes sustained proliferation and genome instability (Senft *et al.*, 2018).

The human homolog of Murine Double Minute-2 (MDM2) resides in the top 11% of densest network regions in the STRING database and was found to have cMDT switches in 33% of Cervix-SCC, 25% of Uterus-AdenoCA and 13% in Bladder-TCC with the transcript ENST00000428863 being mostly expressed. Compared to the GTEx normal isoform ENST00000462284, ENST00000428863 lacks the N-terminal domain which contains the SWIB domain that is essential for binding tumor suppressors like TP53, the ubiquitin proteins like RPS27A, UBA52, UBB, UBC, the ribosomal protein RPL11 and MDM4 (J. Zheng *et al.*, 2015) (see Figure 5D). Thus, these cMDTs lose the ability to bind and ubiquitinate p53. Interestingly, 11 of the affected 24 samples carry besides the MDM splice variant various TP53 mutations. The cMDT of MDM2 could enhance in these cases the gain-of-function effect of mutated TP53 genes (Oren and Rotter, 2010). MDM2 isoforms lacking the SWIB domain but containing the C-terminal zinc-finger RING domain *zf-C3HC4_3* can dimerize and withdraw full-length canonical MDM2 from interacting with p53 (T. Zheng *et al.*, 2013). The same domain drives also the interaction between MDM2 and the ubiquitin-conjugating enzymes UBE2A, UBE2D1, UBE2D2, UBE2D3 and to the SUMO associated enzyme UBE2I (see Figure 5D). Thus, the cancer-specific MDM2 likely induces a gain-of-function effect on TP53 by breaking the negative-feedback loop between wildtype MDM2 and TP53.

In the case of CDK4, the 111 amino acid short cMDT (ENST00000312990), which lacks the entire C-lobe of the kinase domain from the GTEx-specific transcript (ENST00000257904) was expressed in 36% of Uterus-AdenoCA and 14% in Eso-AdenoCA. The loss of the C-lobe disrupted the kinase activity of CDK4. In CDK4 knock-out mice, the loss of CDK4 function has mild effects on cell cycle progression, due to CDK6 compensating for CDK4 loss (Berthet and Kaldis, 2007). However, in 11 of

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

24 samples this compensation effect is most likely absent due to mutated CDK6, which hints towards a strong functional impact of CDK4 cMDT in these tumors. CDK4 lies in the top 11% of densest network regions in STRING.

In summary, our analysis shows that many cMDTs are located in the direct neighborhood of known cancer relevant genes within densely populated PPI network regions.

Discovering novel pathogenic genes via cancer-specific most dominant transcripts

All genes and cMDTs discussed above were known to have a role in cancer. To discover new cancer-associated genes driving neoplasm via cancer-specific alternative splicing, we searched for cMDT switches in the top 30% of the densest regions in the STRING database that were not interactors of COSMIC CGC genes.

The cMDT switch with the highest number of disrupted interactions located in the top 15% of densest network regions was the Natriuretic Peptide receptor 2, NPR2. In 57% of Uterus-AdenoCA, 19% Ovary-AdenoCA, 15% Bone-Leiomyo, and 14% ColoRect-AdenoCA cases, NPR2 expressed an cMDT switch (ENST00000448821), which undergoes nonsense-mediated decay (see Ensembl entry of transcript). The loss of NPR2 disrupts interactions of the canonical transcript ENST00000342694 with the hormone Natriuretic peptide type A, B and C (NPPA, NPPB and NPPC). Disrupted interactions between NPR2 and NPPC have been shown to cause disorganized chromosomes in mouse oocytes (Kiyosu *et al.*, 2012). Given that the chromosome structure is often altered in cancer, the cMDT switch in NPR2 could hint towards a role of NPR2 in cancer. Interestingly, we find potential deleterious mutations along the NPR2 gene in 113 PCAWG samples that supports this hypothesis.

Furthermore, the Charged multivesicular body protein 7 (CHMP7) was found to have an cMDT switch (ENST00000517325) in 50% of Uterus-AdenoCA, 17% of Breast-LobularCA and 13% of Breast-AdenoCA. It is part of the densest 22% of network regions in STRING. According to the Ensembl database also this cMDT is also predicted to undergo nonsense-mediated decay. As a result, the interaction between CHMP7 and other CHMP family members (2A, 2B, 3, 4A, 4B, 4C, 5, 6) and the Vacuolar protein sorting-associated homolog protein VTA1 is deleted. CHMP7 is known to play an important role in repairing envelope raptures after cancer cell migration (Denais *et al.*, 2016). Lacking

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

functional CHMP7 proteins in cancer cells and a fully functional nuclear envelope can induce extensive double-strand breaks and damage to nuclear DNA (Willan *et al.*, 2019). Thus, the NMD driven loss of CHMP7 could play an important role in the cancer hallmark describing genome instability and mutations (Hanahan and Weinberg, 2011).

In 41% of Uterus-AdenoCA, 17% of Bladder-TCC and one of Eso-AdenoCA PCAWG samples, we also identified switches in the cMDT for Tubulin beta-6 chain (TUBB6), where a short transcript ENST00000591909 was mainly expressed. The short transcript lacked most of the central and C-terminal sequence of the canonical isoform (ENST00000317702), which contains both of the tubulin domains. The canonical isoform is located in the top 29% of densest STRING network regions. Thus, the cMDT switch will disrupt interactions not only to other Tubulin family members (1, 1A, 1B, 1C, 2A, 2B, 3C, 3E, 4A, 4B) and the dynein 1 and 2 heavy chains but also have far-reaching impact beyond the direct interaction partners. The loss of specific Tubulin functions was associated with more aggressive forms of cancer tumors and resistance formation upon tubulin-binding chemotherapy agents (Parker *et al.*, 2017).

In summary, through our analysis on cMDTs we provide evidence for previously non-cancer associated genes to play an important role in tumor progression.

Non-coding mutations associated with cMDT switches

Plotting the sum of all single (SNVs) and multi-nucleotide variants (MNVs; joining of adjacent SNVs), and insertion and deletions (indels) against the number of cMDTs in the PCAWG dataset, revealed for the entirety of the dataset no correlation, R= -0.03 (Spearman's rank correlation) (see Figure 6A). Interestingly, the points of various cancer types clustered together at different regions in log-space, indicating that the number of mutations and the number of cMDTs were more similar within cancer types than between different cancer types. Separating the correlation analysis by cancer type showed in contrary to the global correlation analysis, a wide range of positive and negative correlations for the various cancer types. ColoRect-AdenoCA had the lowest negative correlation with R= -0.23 in the PCAWG dataset. In contrast, Breast-AdenoCA, CNS-Oligo, Prost-AdenoCA, and Breast-LobularCA samples had the highest positive correlation between their number of mutations and alternative splicing disruption with R values between 0.25 and 0.49. Note however that the correlation analysis

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

was mostly inconclusive due to high p-values (see Table 3). This somewhat contradicts the results of Eduardo and co-workers, who found a significant inverse correlation between protein-affecting mutations and functional MDTs (Climente-González *et al.*, 2017).

Next, we compared the mutations in the PCAWG dataset with the expression values of the transcripts to identify potential causative mutations for the cMDT switches in this study. To minimize confounder effects, we compared the expression values between mutated and wildtype transcripts from the same cancer type only. GTEx samples were not taken into account. In total, we were able to identify an association between mutations in cis and the expression value of a transcript for 20 cases, of which none was an cMDT described in this study. It seems that the dramatic alternations of the cMDT switches are not caused by mutations in cis-regions but rather by other alternative mechanisms (see Discussion).

The transcripts whose expression was most significantly correlated with mutations within the gene structure were those of the apoptosis regulator Bcl-2 in Lymph-BNHL (see Figure 6C). In total 44 of 103 samples had mutations either in the promoter region, 5 and 3'UTR, splice-site, intronic or exonic region that correlated with transcript expression in the gene. Interestingly, the expression of three out of four transcripts (ENST00000333681 (FDR corrected Wilcox test = 2.4e-08), ENST00000589955 (FDR corrected Wilcox test = 1.6e-07), ENST00000398117 (FDR corrected Wilcox test = 3.6e-05)) of Bcl-2 showed a high correlation with the mutations (see Figure 6C and Figure S1), hinting towards a general upregulation of the gene due to the detected mutations. Additional transcripts in Lymph-BNHL whose over-expression significantly correlated with mutations in cis were those of MYC (see Figure 6D and Figure S1) with mutations in 5'UTR, promoter, splice-site in particular exonic mutations (ENST00000377970 (FDR corrected Wilcox test = 4.0e-05), ENST00000524013 (FDR corrected Wilcox test = 5.3e-05)) and those of Serum/glucocorticoid-regulated kinase 1 SGK1 with various mutations in the promotor, UTR regions, splice site and coding sequence (ENST00000460769 (FDR corrected Wilcox test = 0.005), ENST00000367858 (FDR corrected Wilcox test = 0.008)) (see Figure 6E and Figure S1).

Over the 55 Panc-AdenoCA samples, we found a significant correlation between expressions of the CDKN2A transcripts ENST00000479692 and ENST00000497750 and 22 mutations covering the entire coding sequence and a single splice site mutation in the second last exon (ENST00000479692

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(FDR corrected Wilcox test = 0.004, ENST00000497750 (FDR corrected Wilcox test = 0.008)) (see Figure 6F and Figure S1).

Additional correlations between expression and mutations were identified for the canonical transcript of TERT (ENST00000310581) whose expression significantly correlated with mutations in the promoter region of 11/47 Thy-AdenoCA samples (FDR corrected Wilcox test = 5.5e-06) (see Figure 6B). The expression of the transcripts ENST00000547379 and ENST00000367714 from the Monocarboxylate transporter gene SLC16A7 and Sodium/hydrogen exchanger gene SLC9C2, respectively, were significantly correlated with various mutations in 10 and 8 ColoRect-AdenoCA samples, respectively (ENST00000547379 (FDR corrected Wilcox test = 0.002), ENST00000367714 (FDR corrected Wilcox test = 0.005)) (see Figure 6G and 6H). However, the median expression of these transcripts and the TERT transcript was generally below 2 TPM, which was the threshold for transcripts to be included in our study. Thus, these transcripts were not considered for the cMDT analysis.

In summary, these results indicate that mutations in cis change the expression of transcripts but are not driving the large-scale changes observed with cMDT switches.

Discussion

We have performed (as of today) the most comprehensive analysis on the pathogenic consequences of alternative splicing alterations in 27 different cancer-types. To perform the analysis, we introduced the concept of cancer-specific most dominant transcripts (cMDT) and have developed a novel isoform-specific protein-protein interaction network to assess their functional and pathogenic impact. We demonstrated large variations in the number of cMDTs but also showed that the cMDT load is tissue-specific, in contrast to the mutational load in the same samples. We identified some cMDT as candidate diagnostic biomarkers which were found in 100% of cancer samples but not in any sample of the matched normal cohort. 28% of protein-protein interactions were disrupted due to cMDTs which were mostly related to enzyme signaling, protein translation, and RNA splicing. When disruptive, cMDTs destroyed all known interactions of a given protein. Most cMDTs were interaction partners of cancer-associated genes. Based on the density of local network regions, we predicted CHMP7, NPR2, and TUBB6 as novel pathogenic genes whose splice variants impact the interaction network

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

similarly as splice variants of cancer-associated genes. And finally, we didn't find evidence of genomic alterations explaining the large extent of cMDT switches but identified transcripts whose expression correlated with various somatic mutations in cis.

Despite the large extent of functional and pathogenic consequences that were detected and predicted for all the different cancer types, two main problems remain with our assessments. Firstly, the RNAseq data on which we based our MDT measurements were collected with short-read sequencing technologies, which have an intrinsic limitation to detect and quantify long transcripts (Steijger *et al.*, 2013). But several benchmarks have shown that alignment-free transcript quantification methods like Kallisto are among the most accurate quantification tools for known transcripts (D. C. Wu *et al.*, 2018; Chi Zhang *et al.*, 2017). Nevertheless, as Kallisto only quantifies known transcripts, we might have underestimated the impact of altered alternative splicing by not considering novel transcripts. Long-read sequencing (Tilgner *et al.*, 2015) in bulk or on single cells (Gupta *et al.*, 2018; X. Liu *et al.*, 2017) are ideal methods to overcome these problems. Their application on large cohorts like PCAWG will certainly advance our understanding of the true extent of cMDT switches in cancer.

Secondly, there remains the possibility that the detected cMDTs are not translated into proteins, in which case all predicted consequences on the interaction networks would be invalid. However, there is currently no technology for measuring protein isoform expression on a proteome wide scale. Mass-spectrometry (MS) based methods which are most widely used to probe the proteome of cancer cells suffer from similar limitations as short-read sequencing technologies. In MS-based methods often only a single or a few peptides are identified to quantify proteins. In most cases, however, these peptides are shared between different transcripts. In a recent study by the Aebersold lab, only 65 peptides which were unique to an isoform were identified from a whole proteome measurement (Y. Liu *et al.*, 2017). Thus, the small number of MS detectable isoform-specific peptides make current proteome-wide judgments on the translation of cMDTs unfeasible.

We also noticed that the identification of cMDTs is somewhat dependent on method parameters, which forced us to be conservative with our choices of fold-change thresholds, interaction score confidences, p-values and the exclusion of any normal cohort matches. Despite the restrictive parameters, we failed to identify any causative mutations in cis that could explain the observed overexpression of the cMDTs. There are multiple reasons why this might be the case. Firstly, even

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

though we had over 1209 samples available for our study, on a cancer-type level we had only 37 cases on a median average. Also, most mutations were unique and found at various locations within a gene's structure. To counteract the data sparsity we had to combine the different mutations which further reduces the power of our correlation analysis (PCAWG Transcriptome Core Group *et al.*). Secondly, the causative mutations could lie outside the cMDT genes like in splicing factors. And indeed, 100 of the 1209 samples have a mutation in at least one of the splicing factors SF3B1, SRSF2, U2AF1, ZRSR2 that are often mutated in cancers (Dvinge *et al.*, 2016). 97 samples have mutations in one of the RNA Polymerase II proteins (RBP1-12), which can also lead to aberrantly spliced products (Oesterreich *et al.*, 2016; Saldi *et al.*, 2016). Thirdly, epigenetic regulations by histone modifications and DNA methylations could have led to some of the observed deregulations in alternative splicing (Luco *et al.*, 2011; Zhu *et al.*, 2018). And finally, more recently a connection between glucose metabolism and splicing efficacy was demonstrated (Biamonti *et al.*, 2018), which could also have contributed to aberrant splicing in our cancer samples. Further in-depth analyses are required to fully understand the genomic causes behind the observed cMDT patterns.

The functional and pathogenic impact that we demonstrated for cMDT switches emphasizes the importance of alternative splicing in tumorigenesis and cancer progression. Future work will show which of the presented findings can be observed on the proteome level and whether these findings can be translated as diagnostic biomarkers for precision oncology into the clinics.

Acknowledgments

We would like to acknowledge first of foremost Dr. Damian Szklarczyk for providing us data and annotations from the STRING database. We also would like to thank Dr. Nuno A. Fonseca for the insightful discussions on the concept of most dominant transcripts and QTL analyses. Furthermore, we would also like to thank Prof. Dr. Juri Reimand, Prof. Dr. Mark D. Robinson. Dr. Kjong Lehmann and Dr. Andre Kahles for in-depth discussions on various aspects of the project. Special thanks go to the PCAWG community and especially the PCAWG-5 group working on "Consequences of somatic mutations on pathway and network activity" led by Prof. Dr. Ben Raphael and Prof. Dr. Josh Stuart for their constant support throughout this project.

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figures

Figure 1: Overview of methodology to assess the impact of cancer-specific Most Dominant Transcript (cMDT) switches using CanIsoNet, an isoform-specific interaction network. Top of the figure with a grey background shows the steps and filters for cMDT switch detection based on the relative expression value analysis of transcripts in PCAWG and GTEx. Bottom of the figure with a violet background describes the methods and databases we used to develop CanIsoNet which combines functional interactions from STRING, with physical domain-domain interactions from 3did on all known transcripts from Ensembl. For the assessment of the functional and pathogenic impact, CanIsoNet counts the relative number of disrupted interactions, collects network density information from the STRING database and proximity information to genes to the COSMIC Cancer Gene Census within the STRING database. The middle section with white background depicts the combination of cMDT switch information with data from CanIsoNet to assess the functional impact of alternatively spliced isoforms.

Figure 2: Overview of the PCAWG dataset. A) Mapping table between code and cancer type name in PCAWG. The upper column plot shows the number of samples with RNAseq and WGS data in PCAWG per cancer type colored according to PCAWG specifications. GTEx tissue type names are followed by PCAWG cancer type codes. The bottom column plot displays the number of samples per GTEx tissue type. B) Number of cancer-specific Most Dominant Transcript (cMDT) per sample, grouped by cancer type and ordered according to the median number of cMDT per cancer type. The top axis displays the sum of all cMDTs per cancer type. Red lines with numbers in each cancer type point towards the median number of cMDT per cancer type. C) Number of Short Variances, i.e. single/multi nucleotide variances and indels per cancer type in log-scale. The top axis represents the sum of all mutations per cancer type. Red lines show the median number of mutations per cancer type.

Figure 3: Structure and frequency of selected most dominant transcripts in the PCAWG dataset. A) NDC80/HEC1 and its Most Dominant Transcripts (MDT) in cancer and matched normal tissues. Note the high relative frequency of the expression of the long transcript in Bone-Leiomyo, Breast-LobularCA, Breast-AdenoCA, CNS-GBM, Bladder-TCC, and Ovary-AdenoCA. B) TPX2 and its MDT in cancer and matched normal tissues. The transcript ENST00000300403 has an extra exon 5' to the central TPX2_importin domain highlighted with the green lines, which inhibit TPX2 translocation to the cytoplasm and hinder its binding to the mitotic spindle. C) NEK2 and its MDT in various cancers and matched normal tissues. In various cancers, the MDT contains an additional C-terminal coiled-coiled region which allows the NEK2A isoform to form a complex with APC/C and be degraded in mitosis.

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 4: STRING interactions with 3did protein domain information and structural representation of the most frequently disrupted Protein-Protein Interactions (PPI) due to cancer-specific Most Dominant Transcript (cMDT) switches. Disrupted interactions are highlighted with black-colored lines. A) Ubiquitin carboxyl-terminal hydrolase 46 (USP46) (shown in lavender color) whose interaction with WD repeat-containing protein 48 (WDR48) (shown in green color) and Polyubiquitin-B (UBB)(shown in orange color) is likely disrupted due to the cMDT of USP46 lacking an N-terminal exon (red-colored segment), which encodes part of a beta-sheet. The loss of the beta-strand has likely major impact on the structure of USP46, disrupting its interaction with UBB (shown as sphere) and the finger motive that interacts directly with WDR48 (Structure from Protein Data Bank (PDB) ID: 5cvn, USP46: chain B, WDR48: chain A, UBB: chain D). B) Complex of 40S ribosomal protein S19 (RPS19) and S16 (RPS16) extracted from the electron microscopy 40S ribosome structure (PDB ID: 5flx, RPS19: chain T, RPS16: chain Q). The cMDT of RPS19 lacks a large portion of the N-terminus which usually forms an interface with RPS16 (shown as spheres). C) The mitochondrial NADH dehydrogenases NDUFA9 and NDUFS7 are shown in complex. The coordinates were extracted from the electron microscopy structure of the human respiratory complex PDB ID: 5xtb (NDUFA9: chain J, NDUFS7: chain C). The interface between NDUFA9 and NDUFS7 is highlighted with spheres. Spliced exons are shown in red color.

Figure 5: Predicting the Pathogenicity of cancer-specific Most Dominant Transcript (cMDT). A) cMDTs and their shortest distance to a COSMIC CGC gene in the STRING interaction network. Relative frequencies of all cMDTs are shown in red, while cMDTs disrupting protein interactions are shown in dark red. Frequencies of randomly selected and expressed proteins are shown in grey. The significance of cancer and random frequency differences was p-value < 2.2e-16 (Wilcox-Rank sum test). B) A protein Network Density Score (NDS) was computed for all genes in the PCAWG dataset based on the number of interactions of a gene and its neighborhood. The histogram shows the distribution of NDS for genes from the COSMIC Cancer Gene Census (CGC) and their interaction partners vs. remaining genes. C-F) Shown are exemplary structures of protein complexes whose integrity is lost due to cMDTs lacking important residues of the binding interface (shown in sphere representations). The spliced-out regions in the cMDTs are shown in red color. Next to the protein structures, are the STRING interactions shown for the cMDT with all interaction partners that could be identified in our STRING-3did database. Interactions that are lost due to an cMDT switch are highlighted with thick black lines. C) Trimeric Protein Phosphatase 2A (PP2A) - Shugoshin 1 (SGO1) complex, with an 18 AA long segment in the ankyrin repeat domain that is spliced out in various cancer types. This short segment is not directly involved in the interaction with the other PP2A subunits. However, its removal by alternative splicing is likely distorting the structure of PPP2R5D and its interactions. The 80 AA long N-terminus of PPP2R5D, which is also spliced out, has no structural coordinates, why the atoms of the first N-terminal amino acid in the structure (Phe92) are shown as spheres to indicate the location of the N-terminus of PPP2R5D. The inset figure shows the

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

same complex rotated horizontally by -90º. The structure of PPP2R5D is a homology model mapped on the PP2A complex of the Protein Data Bank entry PDB-ID: 3fga (Herzog et al., 2012). The STRING interaction map indicates that all known interactions in the STRING-3did network would be lost due to this cMDT. D) X-ray crystal structure (PDB-ID 1ycr) of a small section from the MDM2 – TP53 complex that shows the interface between MDM2 and TP53. The entirety of the MDM2 segment was lost due to alternative splicing. Nevertheless, not all known interactions in the STRING-3did network were affected. The interactions to the ubiquitin-conjugating enzymes likely remain despite the cMDT switch. E) Structure showing the cryo-Electron Microscopy (EM) image of a dimeric microtubule element assembled from human TUBA1A and TUBB6. TUBB6 is a homology model from SWISS-MODEL (Biasini et al., 2014) mapped on the location of TUBB3, which was the original protein in the cryo-EM complex. All known interactions in the STRING-3did database are lost in 23 PCAWG samples expressing the TUBB6 cMDT.

Figure 6: Integrating PCAWG Whole Genome Sequencing data with the Most Dominant Transcript (MDT) information. A) Comparing for each sample the number of single- and multi-nucleotide mutations including insertion and deletions (indels) with the number of cancer-specific MDTs. While globally no correlation can be identified, on a single cancer type basis, positive and negative correlations can be observed (see also Table 3). Please see to assess the functional impact of alternatively spliced isoforms.

Figure 2 for the color code. B-H) Transcripts whose expression most significantly correlates with mutations in the gene structure. q-values are FDR corrected p-values from Wilcox-Rank sum tests between PCAWG expression values from mutated samples vs. non-mutated samples. All transcripts having any expression were taken into consideration. Note, in this context the little expression of the SLC9C2 transcript, which could hint towards a false-positive correlation detection for the transcript. Results are shown only for transcripts for which ≥ 6 samples exist in cancer types which have a mutation in the associated gene. The significance threshold for q-values was < 0.01. Significant correlations were found for additional transcripts of BCL2, MYC, SGK1, and CDKN2A (see Figure S1).

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Tables

Table 1: Cancer-specific Most Dominant Transcripts (cMDT) can disrupt Protein-Protein Interactions (PPI) in cases in which they are not encoding the canonical isoform and are lacking protein domains important for the interaction. For each gene in the dataset, we measured the percentage of interactions that were predicted to be lost due to the expression of a non-canonical cMDT and counted their frequency in the dataset. The total number of interactions in the dataset was 461,437 of which 129,496 PPI were predicted to be lost due to cMDT expression.

Percentage of a Protein's PPI disrupted due to cMDT	Number of cMDTs disrupting a Percentage of a Protein's PPI	Relative frequency in entire dataset
0%	104,700	0.858
1% – 10%	275	0.002
11% – 20%	213	0.002
21% – 30%	98	0.001
31% – 40%	305	0.002
41% – 50%	257	0.002
51% – 60%	30	0.000
61% – 70%	111	0.001
71% – 80%	200	0.002
81% – 90%	403	0.003
91% – 99%	126	0.001
100%	15,333	0.126

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2: Number of disrupted Protein-Protein-Interactions (PPI) due to cancer-specific Most Dominant Transcript (cMDT) switches per cancer type.

Cancer type	Total number of disrupted interactions due to cMDT switch	Total number of samples in cancer type	Mean number of PPI disrupted due to cMDT switch per sample
Uterus/Uterus-AdenoCA	3792	44	86%
Esophagus/Eso-AdenoCA	316	7	45%
CervixUteri/Cervix-SCC	794	18	44%
Colon/ColoRect-AdenoCA	1651	51	32%
Bladder/Bladder-TCC	637	23	28%
Ovary/Ovary-AdenoCA	2788	110	25%
Kidney/Kidney-RCC	2328	117	20%
Pancreas/Panc-AdenoCA	1059	75	14%
Liver/Biliary-AdenoCA	243	18	14%
Muscle/Bone-Leiomyo	415	34	12%
Thyroid/Thy-AdenoCA	482	47	10%
Liver/Liver-HCC	841	100	8%
Breast/Breast-AdenoCA	626	85	7%
Stomach/Stomach-AdenoCA	169	29	6%
Blood/Lymph-CLL	360	68	5%
Lung/Lung-AdenoCA	190	37	5%
Breast/Breast-LobularCA	30	6	5%
Kidney/Kidney-ChRCC	194	43	5%
Lung/Lung-SCC	206	47	4%
Prostate/Prost-AdenoCA	71	19	4%
Skin/Skin-Melanoma	74	36	2%
Blood/Lymph-BNHL	74	103	1%
Brain/CNS-Oligo	5	18	0.3%
Brain/CNS-GBM	6	28	0.2%
SalivaryGland/Head-SCC	0	42	0%

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 3: Spearman correlation coefficients R and p-value between number of short variants and number of cancer-specific Most Dominant Transcript (cMDT) switches for all cancer types. The scatter plot of all points is shown in Figure 5A. A correlation coefficient for SalivaryGland/Head-SCC could not be computed, as it lacked any detectable cMDT switches.

Cancer type	R	P-value
Colon/ColoRect-AdenoCA	-0.23	0.109
Liver/Biliary-AdenoCA	-0.10	0.683
Bladder/Bladder-TCC	-0.09	0.673
Kidney/Kidney-RCC	-0.07	0.486
Uterus/Uterus-AdenoCA	-0.06	0.716
Liver/Liver-HCC	-0.02	0.861
Thyroid/Thy-AdenoCA	-0.02	0.914
Esophagus/Eso-AdenoCA	0.00	1.000
Pancreas/Panc-AdenoCA	0.01	0.933
Brain/CNS-GBM	0.02	0.923
Blood/Lymph-BNHL	0.04	0.708
Muscle/Bone-Leiomyo	0.05	0.784
Ovary/Ovary-AdenoCA	0.10	0.314
Stomach/Stomach-AdenoCA	0.10	0.620
Blood/Lymph-CLL	0.11	0.389
Kidney/Kidney-ChRCC	0.13	0.422
CervixUteri/Cervix-SCC	0.15	0.545
Lung/Lung-AdenoCA	0.16	0.340
Skin/Skin-Melanoma	0.18	0.295
Lung/Lung-SCC	0.19	0.212
Breast/Breast-AdenoCA	0.25	0.019
Brain/CNS-Oligo	0.28	0.256
Prostate/Prost-AdenoCA	0.28	0.247
Breast/Breast-LobularCA	0.49	0.356
SalivaryGland/Head-SCC	NA	NA

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary material

Table S1: Mapping table between PCAWG code, PCAWG cancer type name and matched GTEx tissue cohort.

Table S2: Isoform-specific interaction network with information on which interactions are lost and which remain for each alternatively spliced isoform/transcript. Please read the comment in the file's header for more information on the file format.

Table S3: List of all detected cancer specific Most Dominant Transcript switches (cMDT) and the protein interactions they disrupt with a rich set of various annotations.

Table S4: Most dominant transcripts found in the PCAWG dataset. The Ensembl Gene ID, as well as the gene name, are listed. The number of samples in which the cMDT was observed in the cancer type is given in the Frequency column. The total number of samples per cancer type is listed 6th column, followed by the percentage of samples expressing the transcript as cMDT.

Table S5: Disrupted protein interactions due to cancer-specific Most Dominant Transcript (cMDT) switches in the PCAWG dataset.

Table S6: Most significant biological processes from GeneOntology, which were found enriched for cancer-specific Most Dominant Transcripts (cMDT) that are disrupting protein interactions. Enrichment analysis with FDR corrected p-values were computed on the STRING interaction network using STRINGdb R-package (Franceschini et al., 2013).

Figure S1: Transcripts whose expression significantly correlated with mutations in their associated gene structure. These transcripts are part of multiple transcripts from the same gene that all show a significant correlation between expression and gene mutation (compare to Figure 6).

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

Almeida,A.C. and Maiato,H. (2018) Chromokinesins. *Curr. Biol.*, **28**, R1131–R1135.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25–29.

Baertling,F. *et al.* (2018) NDUFA9 point mutations cause a variable mitochondrial complex I assembly defect. *Clin. Genet.*, **93**, 111–118.

Berman,H. *et al.* (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res*, **35**, D301–3.

Berthet,C. and Kaldis,P. (2007) Cell-specific responses to loss of cyclin-dependent kinases. *Oncogene*, **26**, 4469–4477.

Biamonti,G. *et al.* (2018) The Krebs Cycle Connection: Reciprocal Influence Between Alternative Splicing Programs and Cell Metabolism. *Front Oncol*, **8**, 1029.

Biasini,M. *et al.* (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.*, **42**, W252–8.

Birzele,F. *et al.* (2008) Alternative splicing and protein structure evolution. *Nucleic Acids Res*, **36**, 550–558.

Bradner,J.E. *et al.* (2017) Transcriptional Addiction in Cancer. *Cell*, **168**, 629–643.

Bray,N.L. *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*, **34**, 525–527.

Buljan,M. *et al.* (2012) Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell*, **46**, 871–883.

Climente-González,H. *et al.* (2017) The Functional Impact of Alternative Splicing in Cancer. *Cell Rep*, **20**, 2215–2226.

Corominas,R. *et al.* (2014) Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nat Commun*, **5**, 3650.

Danan-Gotthold,M. *et al.* (2015) Identification of recurrent regulated alternative splicing events across human solid tumors. *Nucleic Acids Res.*, **43**, 5130–5144.

Denais,C.M. *et al.* (2016) Nuclear envelope rupture and repair during cancer cell migration. *Science*, **352**, 353–358.

Dvinge,H. *et al.* (2016) RNA splicing factors as oncoproteins and tumour suppressors. *Nature Reviews Cancer*, **16**, 413–430.

Ellis,J.D. *et al.* (2012) Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol Cell*, **46**, 884–892.

Ezkurdia,I. *et al.* (2014) Analyzing the First Drafts of the Human Proteome. *J Proteome Res*, **13**, 3854–3855.

Ezkurdia,I. *et al.* (2015) Most highly expressed protein-coding genes have a single dominant isoform. *J Proteome Res*, **14**, 1880–1887.

Fang,Y. and Zhang,Xiongwen (2016) Targeting NEK2 as a promising therapeutic approach for cancer treatment. *Cell Cycle*, **15**, 895–907.

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Flygare,J. *et al.* (2007) Human RPS19, the gene mutated in Diamond-Blackfan anemia, encodes a ribosomal protein required for the maturation of 40S ribosomal subunits. *Blood*, **109**, 980–986.

Franceschini,A. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*, **41**, D808–15.

Gonzàlez-Porta,M. *et al.* (2013) Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biol*, **14**, R70.

Goudarzi,K.M. and Llindström,M.S. (2016) Role of ribosomal protein mutations in tumor development (Review). *Int. J. Oncol.*, **48**, 1313–1324.

Gupta,I. *et al.* (2018) Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol*.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of Cancer: The Next Generation. *Cell*, **144**, 646–674.

Herzog,F. *et al.* (2012) Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. *Science*, **337**, 1348–1352.

Hu,J. *et al.* (2017) Stochastic principles governing alternative splicing of RNA. *PLoS Comput Biol*, **13**, e1005761.

Hubbard,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

Jernigan,K.K. and Bordenstein,S.R. (2015) Tandem-repeat protein domains across the tree of life. *PeerJ*, **3**, e732.

Ju,L.-L. *et al.* (2017) Effect of NDC80 in human hepatocellular carcinoma. *World J. Gastroenterol.*, **23**, 3675–3683.

Kahraman,A. *et al.* (2011) Xwalk: computing and visualizing distances in cross-linking experiments. *Bioinformatics*, **27**, 2163–2164.

Kelemen,O. *et al.* (2013) Function of alternative splicing. *Gene*, **514**, 1–30.

Keskin,O. *et al.* (2016) Predicting Protein–Protein Interactions from the Molecular to the Proteome Level. *Chem Rev*, **116**, 4884–4909.

Kiyosu,C. *et al.* (2012) NPPC/NPR2 signaling is essential for oocyte meiotic arrest and cumulus oophorus formation during follicular development in the mouse ovary. *Reproduction*, **144**, 187–193.

Lapuk,A.V. *et al.* (2014) The role of mRNA splicing in prostate cancer. *Asian J. Androl.*, **16**, 515–521.

Li,X. *et al.* (2013) The deubiquitination enzyme USP46 functions as a tumor suppressor by controlling PHLPP-dependent attenuation of Akt signaling in colon cancer. *Oncogene*, **32**, 471–478.

Liu,X. *et al.* (2017) Detecting alternatively spliced transcript isoforms from single-molecule long-read sequences without a reference genome. *Molecular Ecology Resources*, **17**, 1243–1256.

Liu,Y. *et al.* (2017) Impact of Alternative Splicing on the Human Proteome. *Cell Rep*, **20**, 1229–1241.

Lonsdale,J. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, **45**, 580–585.

Luco,R.F. *et al.* (2011) Epigenetics in alternative pre-mRNA splicing. *Cell*, **144**, 16–26.

Mosca,R. *et al.* (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res*, **42**, D374–9.

Network,T.C.G.A.R. *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, **45**, 1113–1120.

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Oesterreich,F.C. *et al.* (2016) Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell*, **165**, 372–381.

Oltean,S. and Bates,D.O. (2014) Hallmarks of alternative splicing in cancer. *Oncogene*, **33**, 5311–5318.

Oren,M. and Rotter,V. (2010) Mutant p53 gain-of-function in cancer. *Cold Spring Harbor Perspectives in Biology*, **2**, a001107–a001107.

Parker,A.L. *et al.* (2017) An Emerging Role for Tubulin Isotypes in Modulating Cancer Biology and Chemotherapy Resistance. *Int J Mol Sci*, **18**, 1434.

PCAWG Transcriptome Core Group *et al.* Genomic basis for RNA alterations revealed by whole-genome analyses of 27 cancer types. **63**, 2665.

Pérez de Castro,I. and Malumbres,M. (2012) Mitotic Stress and Chromosomal Instability in Cancer: The Case for TPX2. *Genes & Cancer*, **3**, 721–730.

Popp,M.W. and Maquat,L.E. (2018) Nonsense-mediated mRNA Decay and Cancer. *Curr. Opin. Genet. Dev.*, **48**, 44–50.

Poulikakos,P.I. *et al.* (2011) RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). *Nature*, **480**, 387–390.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Saldi,T. *et al.* (2016) Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J Mol Biol*, **428**, 2623–2635.

Samatar,A.A. and Poulikakos,P.I. (2014) Targeting RAS-ERK signalling in cancer: promises and challenges. *Nature Reviews Drug Discovery*, **13**, 928–942.

Santidrian,A.F. *et al.* (2013) Mitochondrial complex I activity and NAD+/NADH balance regulate breast cancer progression. *J. Clin. Invest.*, **123**, 1068–1081.

Schaefer,M.H. *et al.* (2015) Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front Genet*, **6**, 260.

Sebestyén,E. *et al.* (2015) Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res.*, **43**, 1345–1356.

Senft,D. *et al.* (2018) Ubiquitin ligases in oncogenic transformation and cancer therapy. *Nature Reviews Cancer*, **18**, 69–88.

Seshacharyulu,P. *et al.* (2013) Phosphatase: PP2A structural importance, regulation and its aberrant expression in cancer. *Cancer Lett.*, **335**, 9–18.

Srinivasan,S. *et al.* (2016) Mitochondrial respiratory defects promote the Warburg effect and cancer progression. *Mol Cell Oncol*, **3**, e1085120.

Steijger,T. *et al.* (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods*, **10**, 1177–1184.

Stein,A. (2004) 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.*, **33**, D413–D417.

Sveen,A. *et al.* (2015) Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene*, **35**, 2413–2427.

Szklarczyk,D. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, **43**, D447–D452.

bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Tilgner,H. *et al.* (2015) Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat Biotechnol*, **33**, 736–742.

Tripp,K.W. and Barrick,D. (2004) The tolerance of a modular protein to duplication and deletion of internal repeats. *J Mol Biol*, **344**, 169–178.

Valverde,R. *et al.* (2016) Conserved Tetramer Junction in the Kinetochore Ndc80 Complex. *Cell Rep*, **17**, 1915–1922.

Vidal,M. *et al.* (2011) Interactome networks and human disease. *Cell*, **144**, 986–998.

Vitting-Seerup,K. and Sandelin,A. (2017) The Landscape of Isoform Switches in Human Cancers. *Mol. Cancer Res.*, **15**, 1206–1220.

Wang,B.-D. *et al.* (2017) Alternative splicing promotes tumour aggressiveness and drug resistance in African American prostate cancer. *Nat Commun*, **8**, 15921.

Wang,H. *et al.* (2011) Identification of an exon 4-deletion variant of epidermal growth factor receptor with increased metastasis-promoting capacity. *Neoplasia*, **13**, 461–471.

Willan,J. *et al.* (2019) ESCRT-III is necessary for the integrity of the nuclear envelope in micronuclei but is aberrant at ruptured micronuclear envelopes generating damage. *Oncogenesis*, **8**, 29.

Wu,D.C. *et al.* (2018) Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics*, **19**, 510.

Wu,G. and Chen,P.-L. (2008) Structural requirements of chromokinesin Kif4A for its proper function in mitosis. *Biochem Biophys Res Commun*, **372**, 454–458.

Xia,J. *et al.* (2015) Role of NEK2A in human cancer and its therapeutic potentials. *Biomed Res Int*, **2015**, 862461.

Yan,X. *et al.* (2018) Nuclear division cycle 80 promotes malignant progression and predicts clinical outcome in colorectal cancer. *Cancer Med*, **7**, 420–432.

Yang,X. *et al.* (2016) Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell*, **164**, 805–817.

Yin,J. *et al.* (2015) Structural Insights into WD-Repeat 48 Activation of Ubiquitin-Specific Protease 46. *Structure*, **23**, 2043–2054.

Zhang,Chi *et al.* (2017) Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics*, **18**, 1–11.

Zhang,Xiaoling *et al.* (2015) Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nat Genet*, **47**, 345–352.

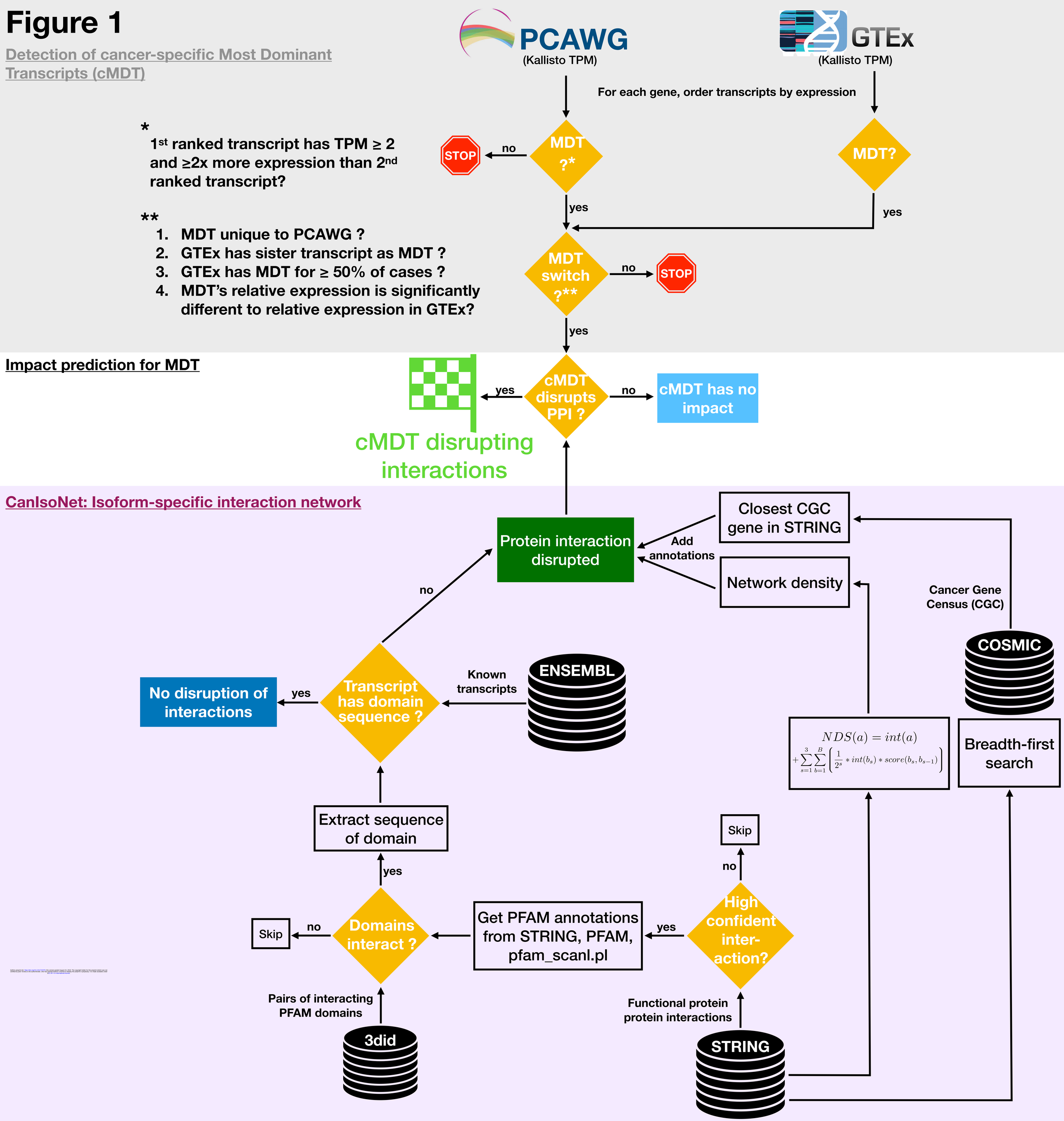
Zheng,J. *et al.* (2015) Structure of human MDM2 complexed with RPL11 reveals the molecular basis of p53 activation. *Gene Dev*, **29**, 1524–1534.

Zheng,T. *et al.* (2013) Spliced MDM2 isoforms promote mutant p53 accumulation and gain-of-function in tumorigenesis. *Nat Commun*, **4**, 2996.

Zhu,L.-Y. *et al.* (2018) Epigenetic regulation of alternative splicing. *American Journal of Cancer Research*, **8**, 2346–2358.

Figure 1

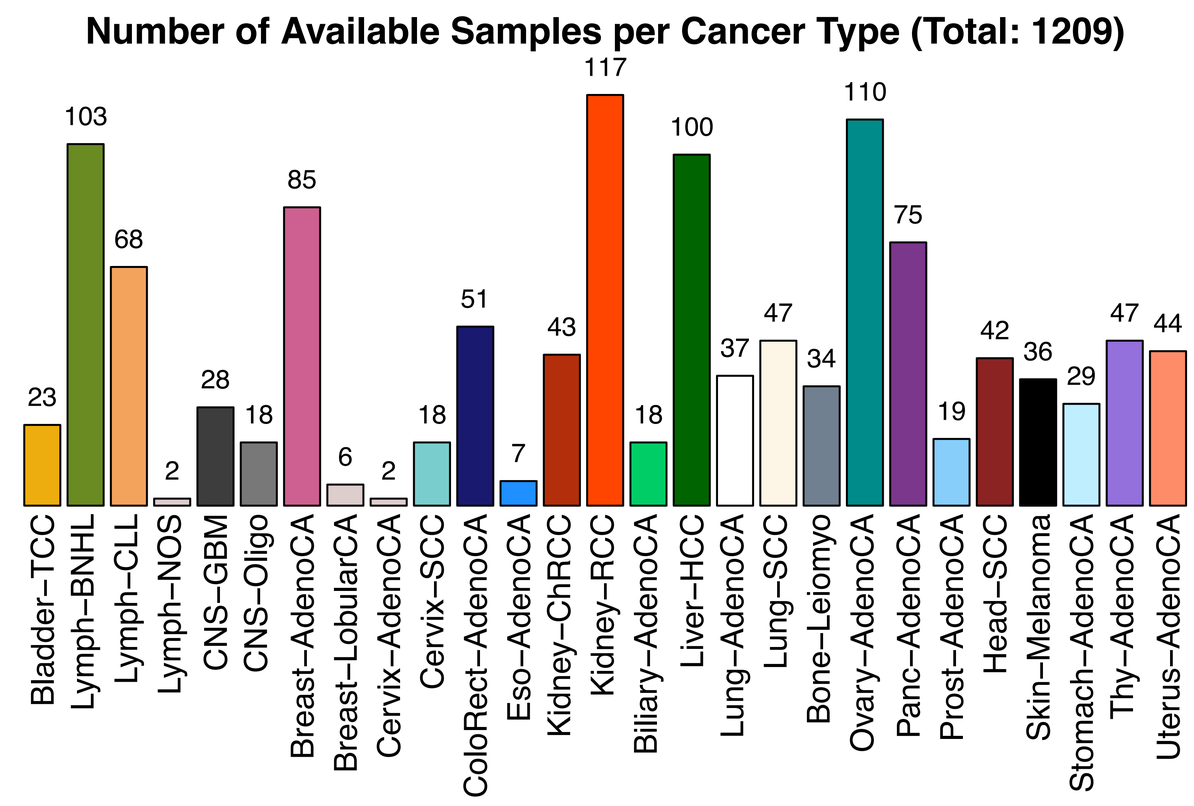
Detection of cancer-specific Most Dominant Transcripts (cMDT)

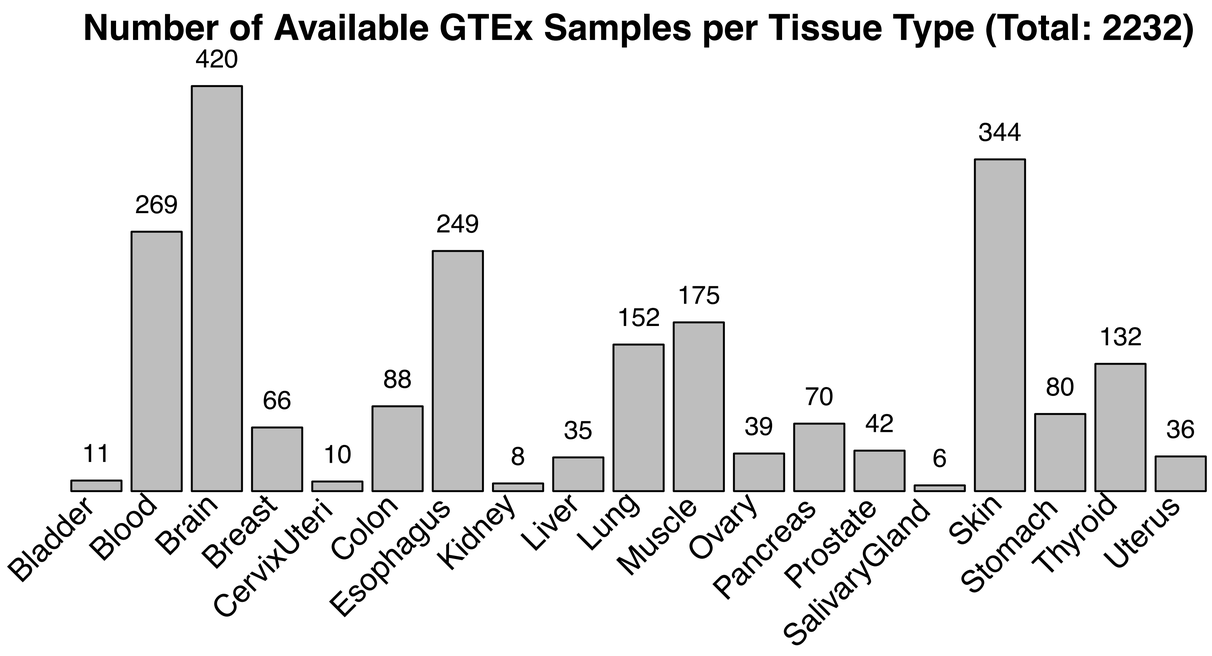


bioRxiv preprint doi: https://doi.org/10.1101/742270; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under CC-BY 4.0 International license.

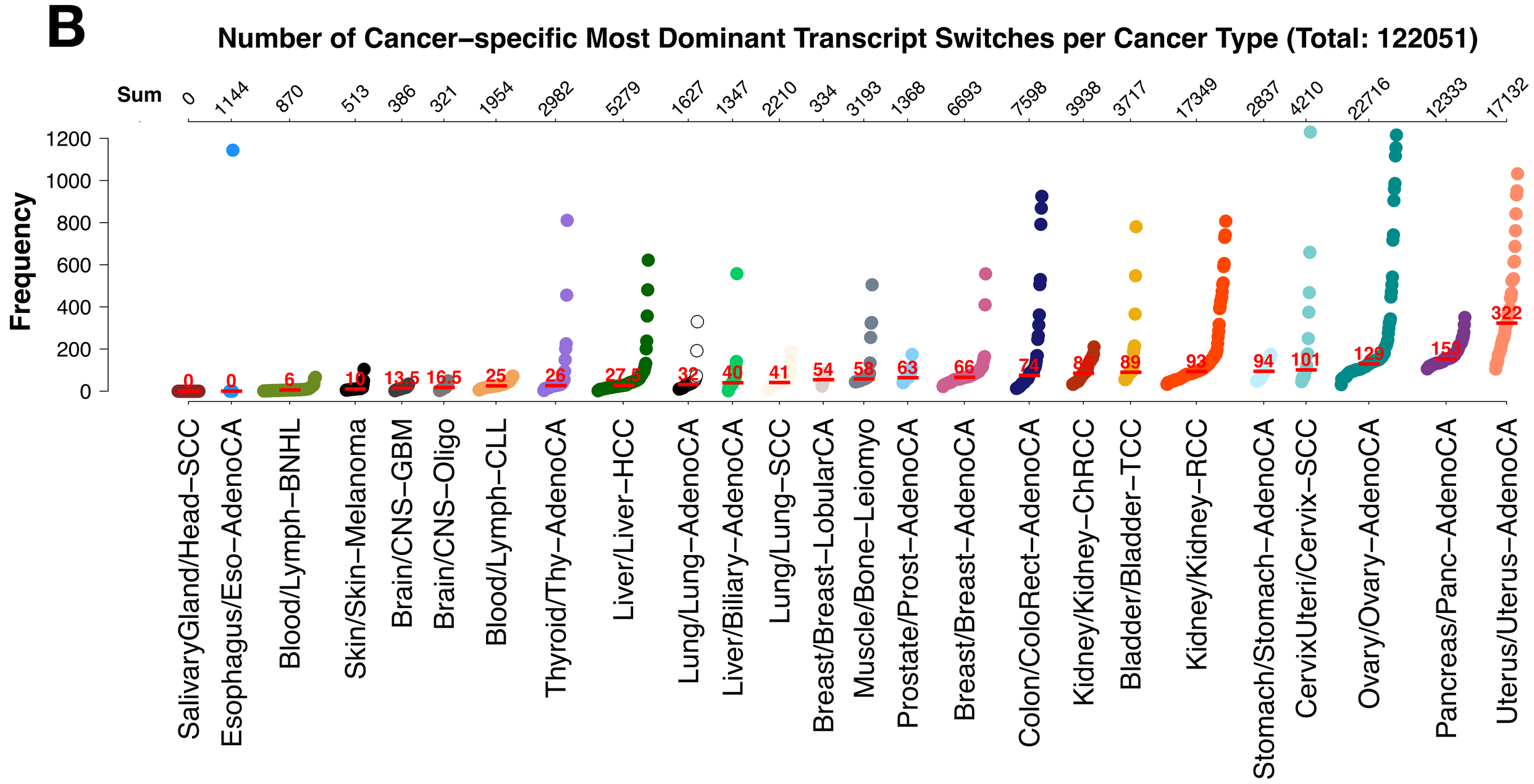
Figure 2 A

PCAWG Code	Cancer Type Name
Biliary-AdenoCA	Liver Cancer
Bladder-TCC	Bladder Urothelial Cancer
Bone-Leiomyo	Sarcoma
Breast-AdenoCA	Breast Cancer
Breast-LobularCA	Breast Invasive Lobular Carcinoma
Cervix-AdenoCA	Cervical Squamous Cell Carcinoma
Cervix-SCC	Cervical Squamous Cell Carcinoma
CNS-GBM	Brain Glioblastoma Multiforme
CNS-Oligo	Brain Lower Grade Gliona
ColoRect-AdenoCA	Colorectal Adenocarcinoma
Eso-AdenoCA	Esophageal Adenocarcinoma
Head-SCC	Head and Neck Squamous Cell Carcinoma
Kidney-ChRCC	Kidney Chromophobe Renal Cell Carcinoma
Kidney-RCC	Kidney Renal Papillary Cell Carcinoma
Liver-HCC	Liver Hepato-Cellular Carcinoma
Lung-AdenoCA	Lung Adenocarcinoma
Lung-SCC	Lung Squamous Cell Carcinoma
Lymph-BNHL	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
Lymph-CLL	Chronic Lymphocytic Leukemia
Lymph-NOS	Malignant Lymphoma
Ovary-AdenoCA	Ovarian Cancer
Panc-AdenoCA	Pancreatic Cancer
Prost-AdenoCA	Prostate Adenocarcinoma
Skin-Melanoma	Skin Cutaneous Melanoma
Stomach-AdenoCA	Gastric Adenocarcinoma
Thy-AdenoCA	Head and Neck Thyroid Carcinoma
Uterus-AdenoCA	Uterine Corpus Endometrial Carcinoma

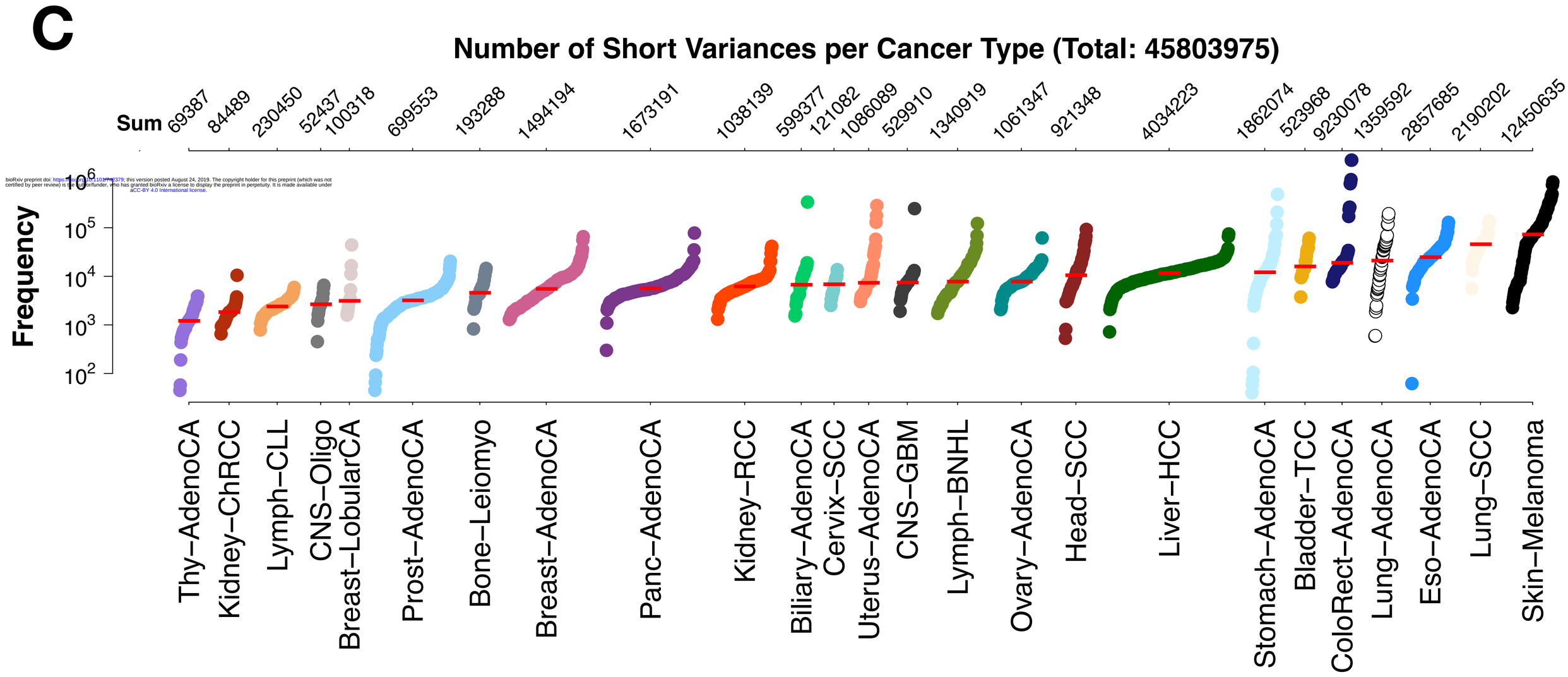




B



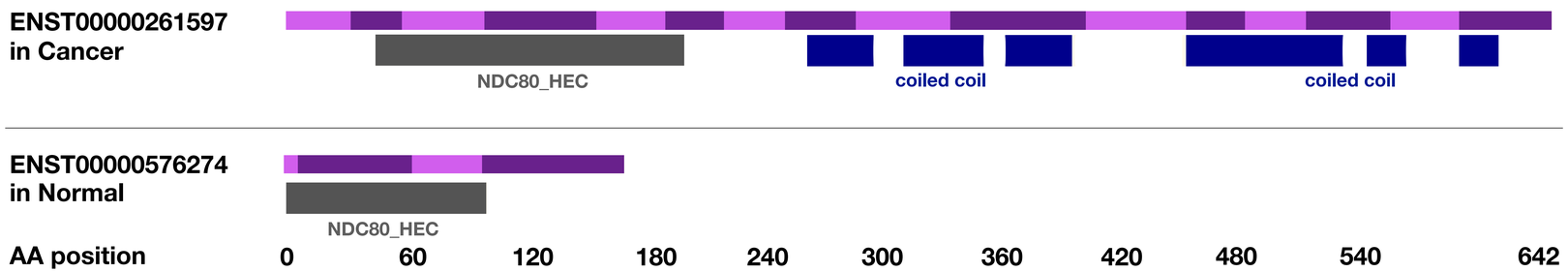
C



bioRxiv preprint doi: https://doi.org/10.1101/746379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

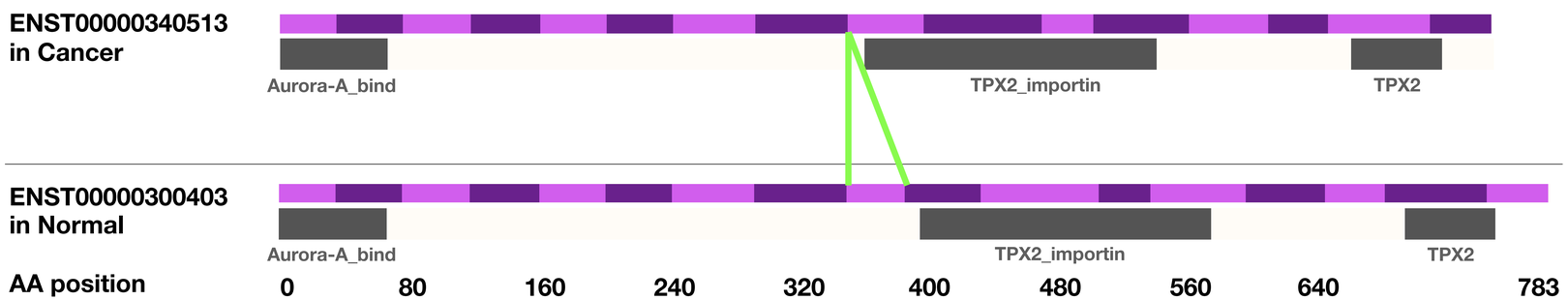
Figure 3

A NDC80/HEC1



GTEx tissue/PCAWG cancer type	# of samples with cMDT ENST00000261597	Total # of samples	Relative frequency
Muscle/Bone-Leiomyo	34	34	1.00
Breast/Breast-LobularCA	6	6	1.00
Breast/Breast-AdenoCA	83	85	0.98
Brain/CNS-GBM	27	28	0.96
Bladder/Bladder-TCC	22	23	0.96
Ovary/Ovary-AdenoCA	104	110	0.95
Liver/Liver-HCC	58	100	0.58
Pancreas/Panc-AdenoCA	38	75	0.51
Liver/Biliary-AdenoCA	9	18	0.50
Kidney/Kidney-RCC	41	117	0.35
Thyroid/Thy-AdenoCA	8	47	0.17
Brain/CNS-Oligo	3	18	0.17
Kidney/Kidney-ChRCC	5	43	0.12

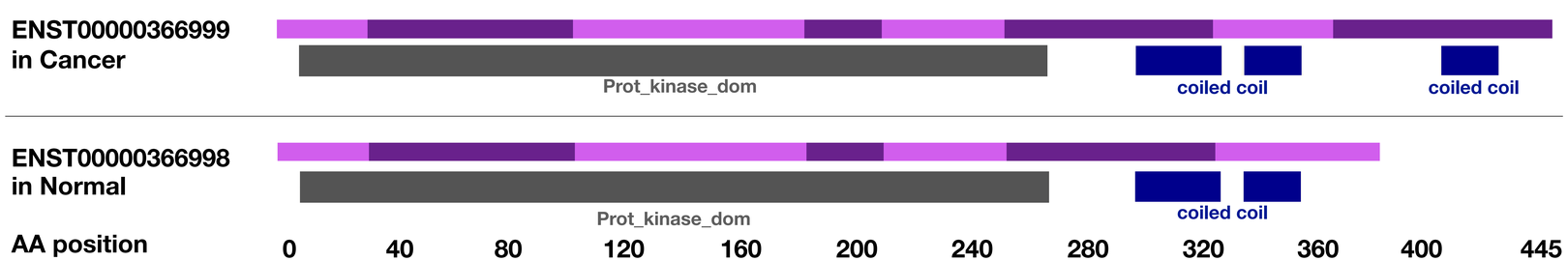
B TPX2



GTEx tissue/PCAWG cancer type	# of samples with cMDT ENST00000340513	Total # of samples	Relative frequency
Breast/Breast-LobularCA	6	6	1.00
Breast/Breast-AdenoCA	85	85	1.00
Uterus/Uterus-AdenoCA	43	44	0.98
Bladder/Bladder-TCC	21	23	0.91

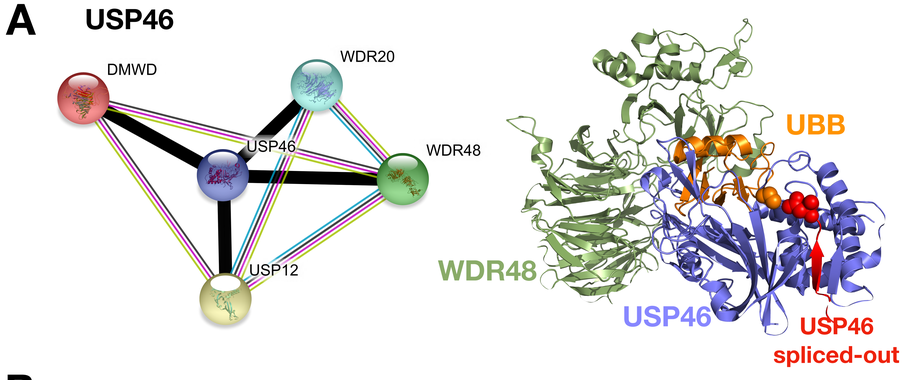
bioRxiv preprint doi: https://doi.org/10.1101/742379; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

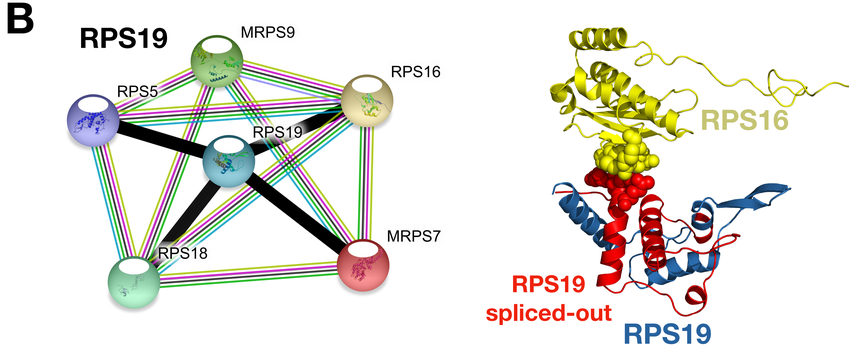
C NEK2

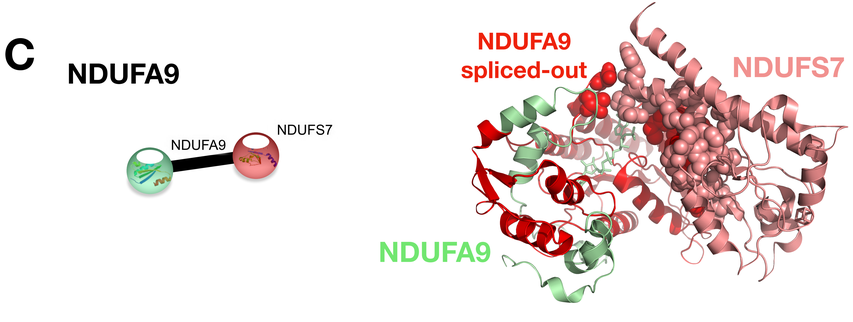


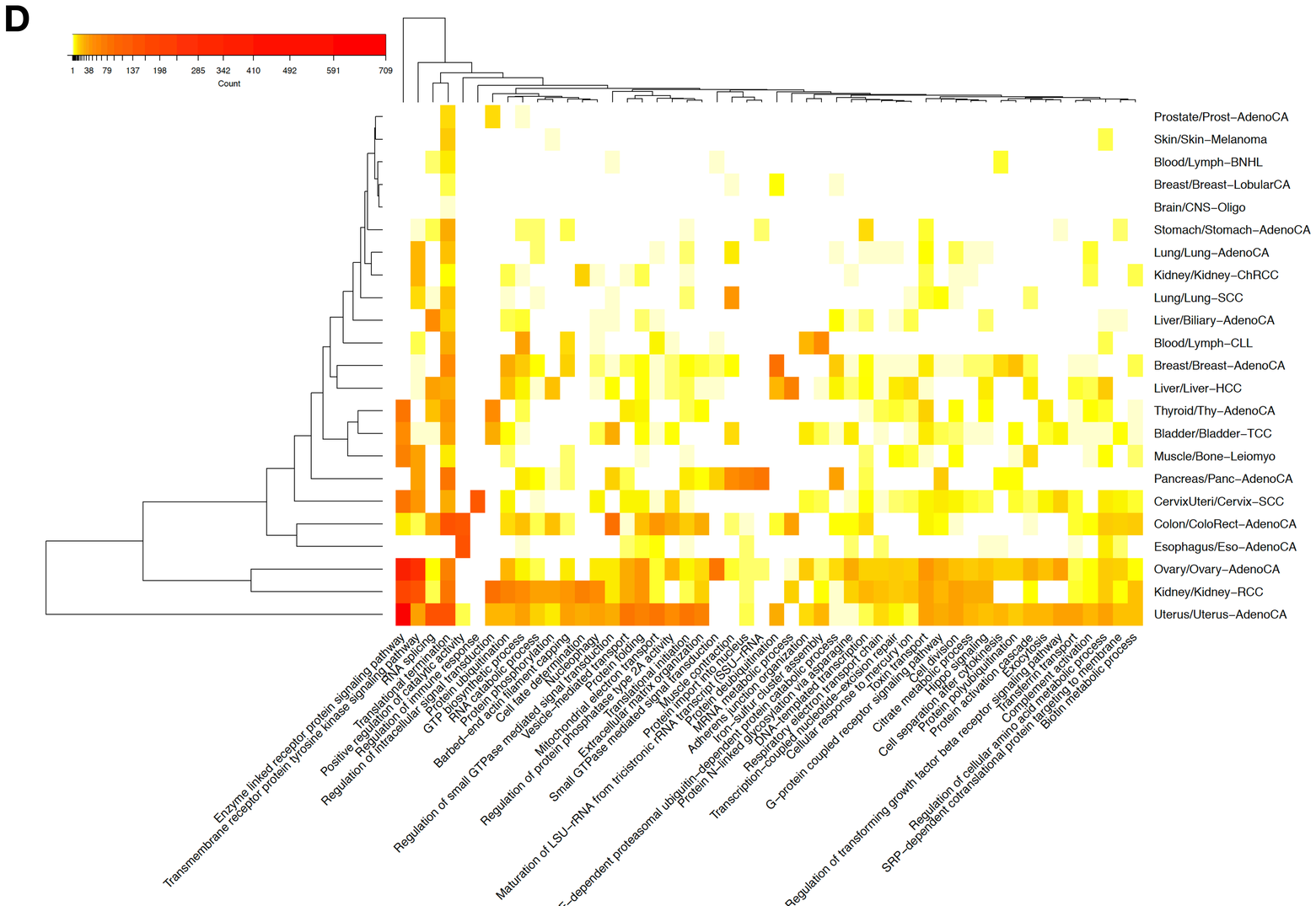
GTEx tissue/PCAWG cancer type	# of samples with cMDT ENST00000366999	Total # of samples	Relative frequency
CervixUteri/Cervix-SCC	18	18	1.00
Bladder/Bladder-TCC	21	23	0.91
Ovary/Ovary-AdenoCA	99	110	0.90
Muscle/Bone-Leiomyo	26	34	0.76
Uterus/Uterus-AdenoCA	33	44	0.75
Pancreas/Panc-AdenoCA	44	75	0.59
Kidney/Kidney-RCC	13	117	0.11
Kidney/Kidney-ChRCC	2	43	0.05
Thyroid/Thy-AdenoCA	1	47	0.02

Figure 4









bioRxiv preprint doi: https://doi.org/10.1101/763276; this version posted August 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 5

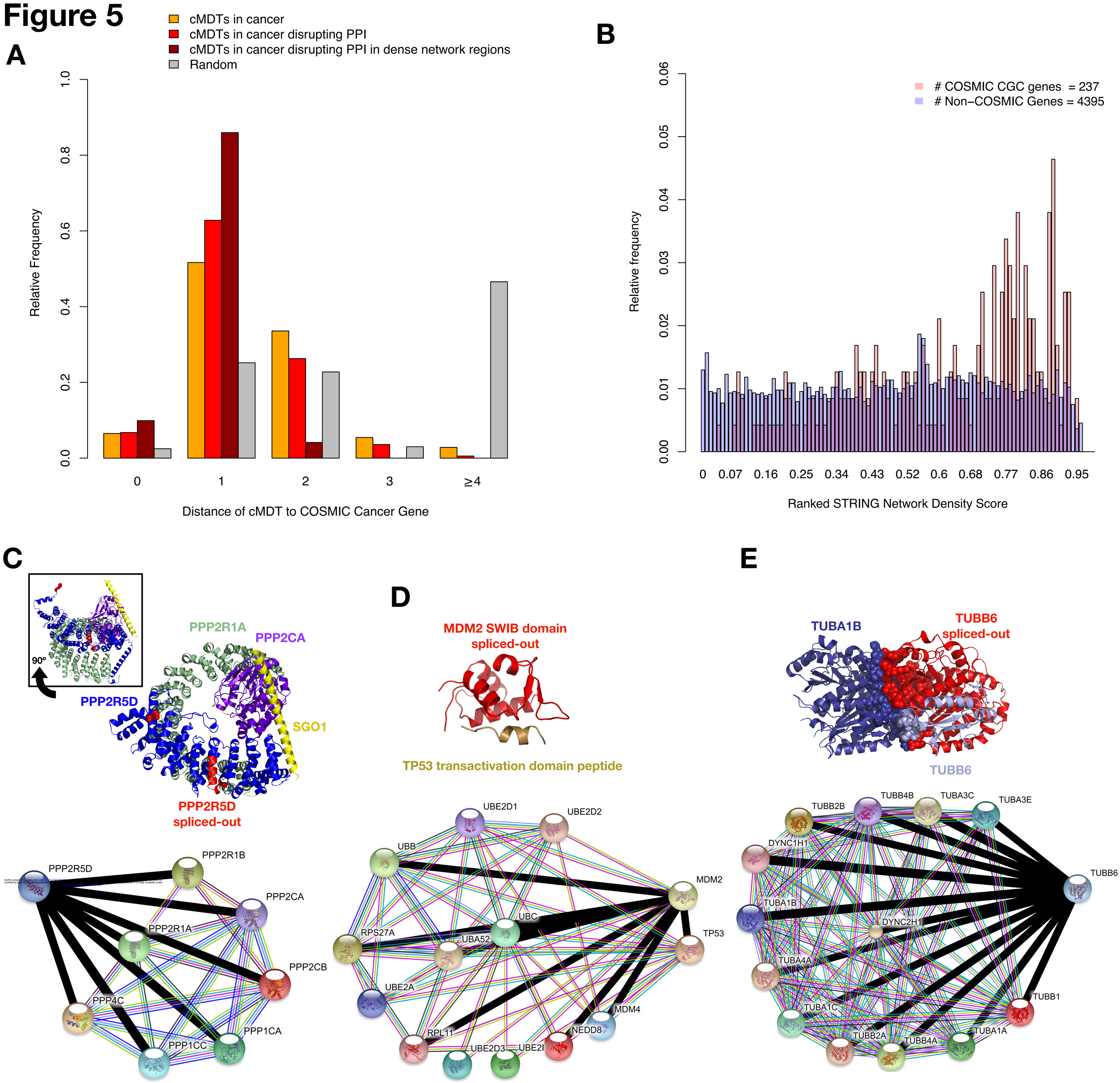
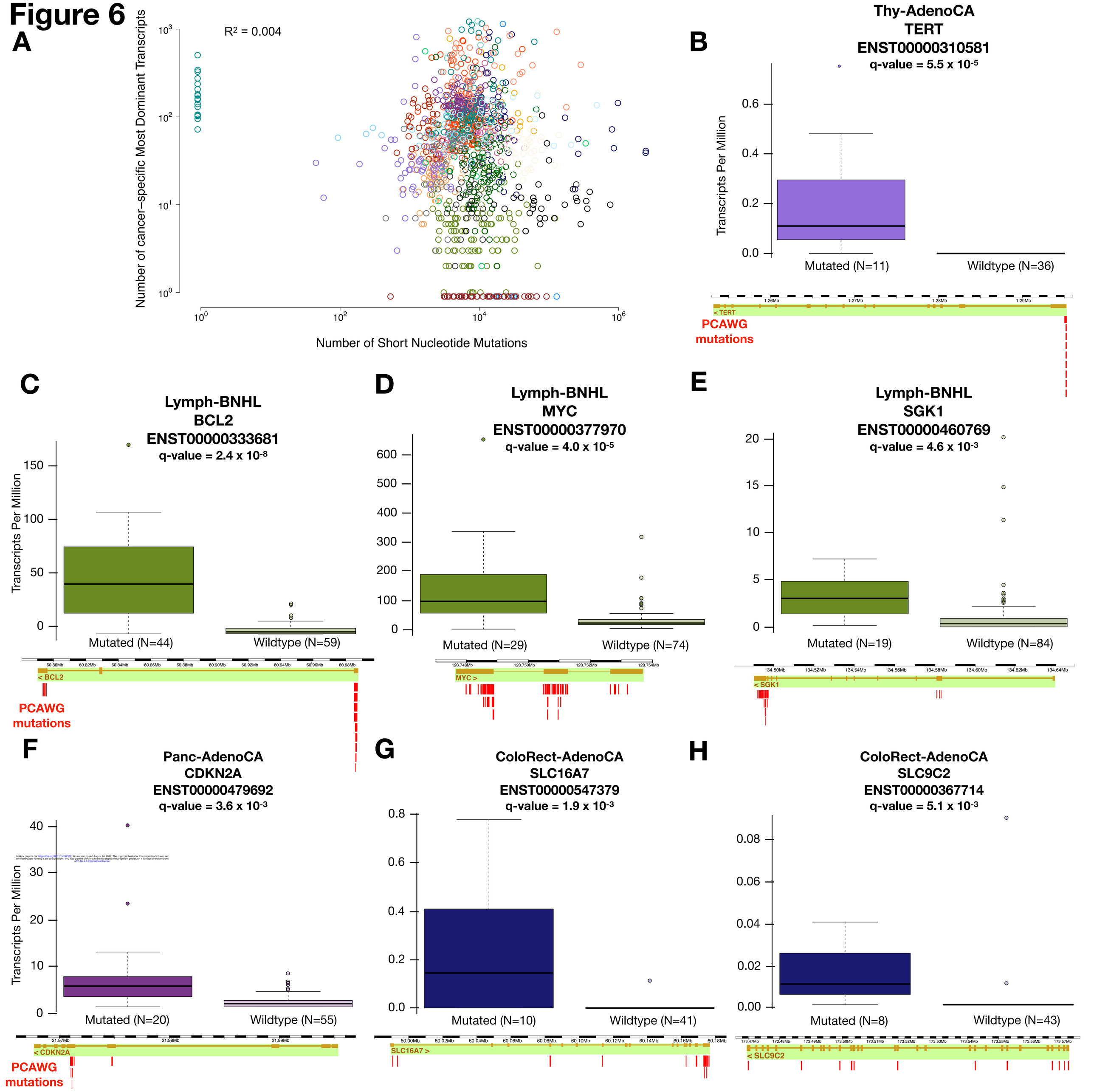

A
cMDTs in cancer
cMDTs in cancer disrupting PPI
cMDTs in cancer disrupting PPI in dense network regions
Random
Relative Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
≥4
Distance of cMDT to COSMIC Cancer Gene
B
COSMIC CGC genes = 237
Non−COSMIC Genes = 4395
Relative frequency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0
0.07
0.16
0.25
0.34
0.43
0.52
0.6
0.68
0.77
0.86
0.95
Ranked STRING Network Density Score
C
90°
PPP2R1A
PPP2CA
PPP2R5D
SGO1
PPP2R5D
spliced-out
PPP2R5D
PPP2R1B
PPP2CA
PPP2R1A
PPP2CB
PPP4C
PPP1CA
PPP1CC
D
MDM2 SWIB domain
spliced-out
TP53 transactivation domain peptide
UBE2D1
UBE2D2
UBB
MDM2
UBC
RPS27A
UBA52
TP53
UBE2A
MDM4
RPL11
NEDD8
UBE2D3
UBE2I
E
TUBA1B
TUBB6
spliced-out
TUBB6
TUBB2B
TUBB4B
TUBA3C
TUBA3E
DYNC1H1
TUBB6
TUBA1B
DYNC2H1
TUBA4A
TUBB1
TUBA1C
TUBA1A
TUBB2A
TUBB4A

Figure 6


A
R² = 0.004
Number of cancer–specific Most Dominant Transcripts
Number of Short Nucleotide Mutations
B
Thy-AdenoCA
TERT
ENST00000310581
q-value = 5.5 x 10-5
Transcripts Per Million
Mutated (N=11)
Wildtype (N=36)
< TERT
PCAWG mutations
C
Lymph-BNHL
BCL2
ENST00000333681
q-value = 2.4 x 10-8
Transcripts Per Million
Mutated (N=44)
Wildtype (N=59)
< BCL2
PCAWG mutations
D
Lymph-BNHL
MYC
ENST00000377970
q-value = 4.0 x 10-5
Mutated (N=29)
Wildtype (N=74)
MYC >
E
Lymph-BNHL
SGK1
ENST00000460769
q-value = 4.6 x 10-3
Mutated (N=19)
Wildtype (N=84)
< SGK1
F
Panc-AdenoCA
CDKN2A
ENST00000479692
q-value = 3.6 x 10-3
Transcripts Per Million
Mutated (N=20)
Wildtype (N=55)
< CDKN2A
PCAWG mutations
G
ColoRect-AdenoCA
SLC16A7
ENST00000547379
q-value = 1.9 x 10-3
Mutated (N=10)
Wildtype (N=41)
SLC16A7 >
H
ColoRect-AdenoCA
SLC9C2
ENST00000367714
q-value = 5.1 x 10-3
Mutated (N=8)
Wildtype (N=43)
< SLC9C2