Supplementary material for:

An atlas of transposable element induced alternative splicing in cancer

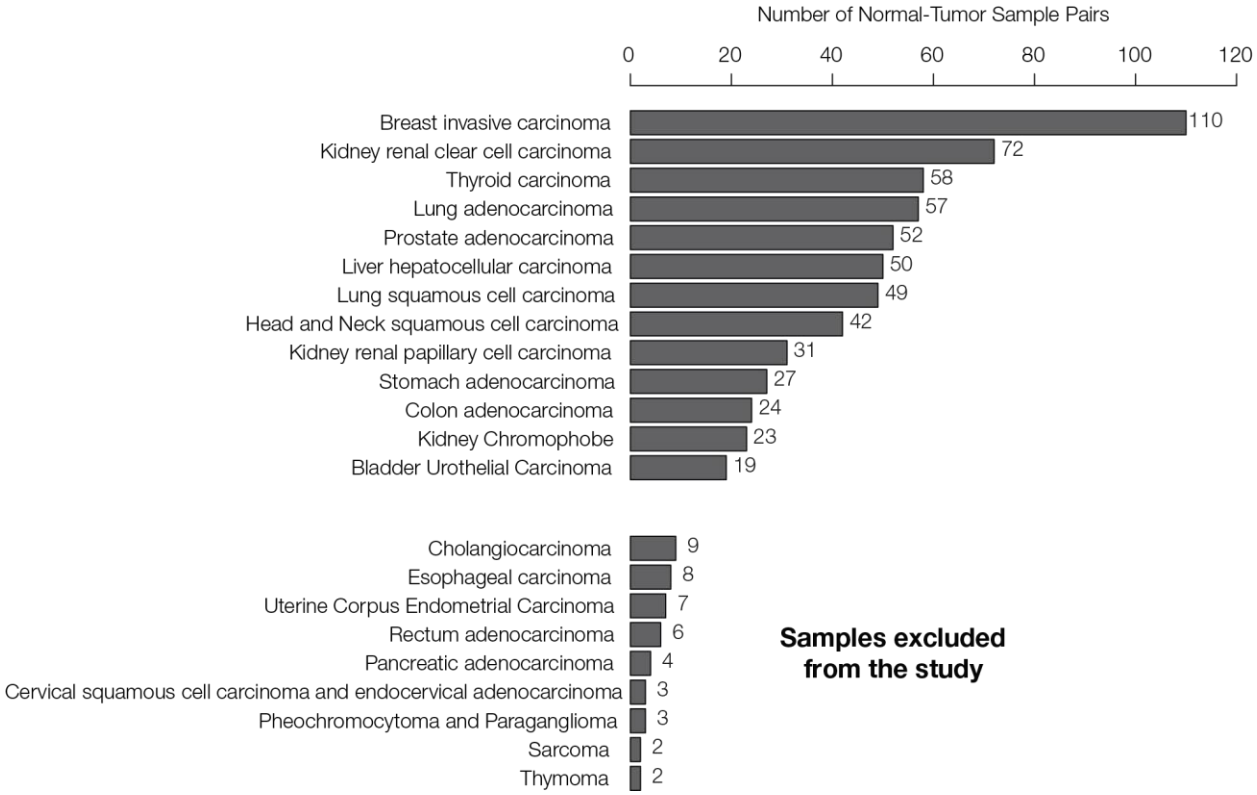
Evan A. Clayton[1], Lavanya Rishishwar[2,3,4], Tzu-Chuan Huang[2], Saurabh Gulati[2], Dongjo Ban[1], John F. McDonald[1] and I. King Jordan[2,3,4*]

[1]Integrated Cancer Research Center, School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA
[2]School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, USA
[3]PanAmerican Bioinformatics Institute, Cali, Colombia
[4]Applied Bioinformatics Laboratory, Atlanta, GA, USA

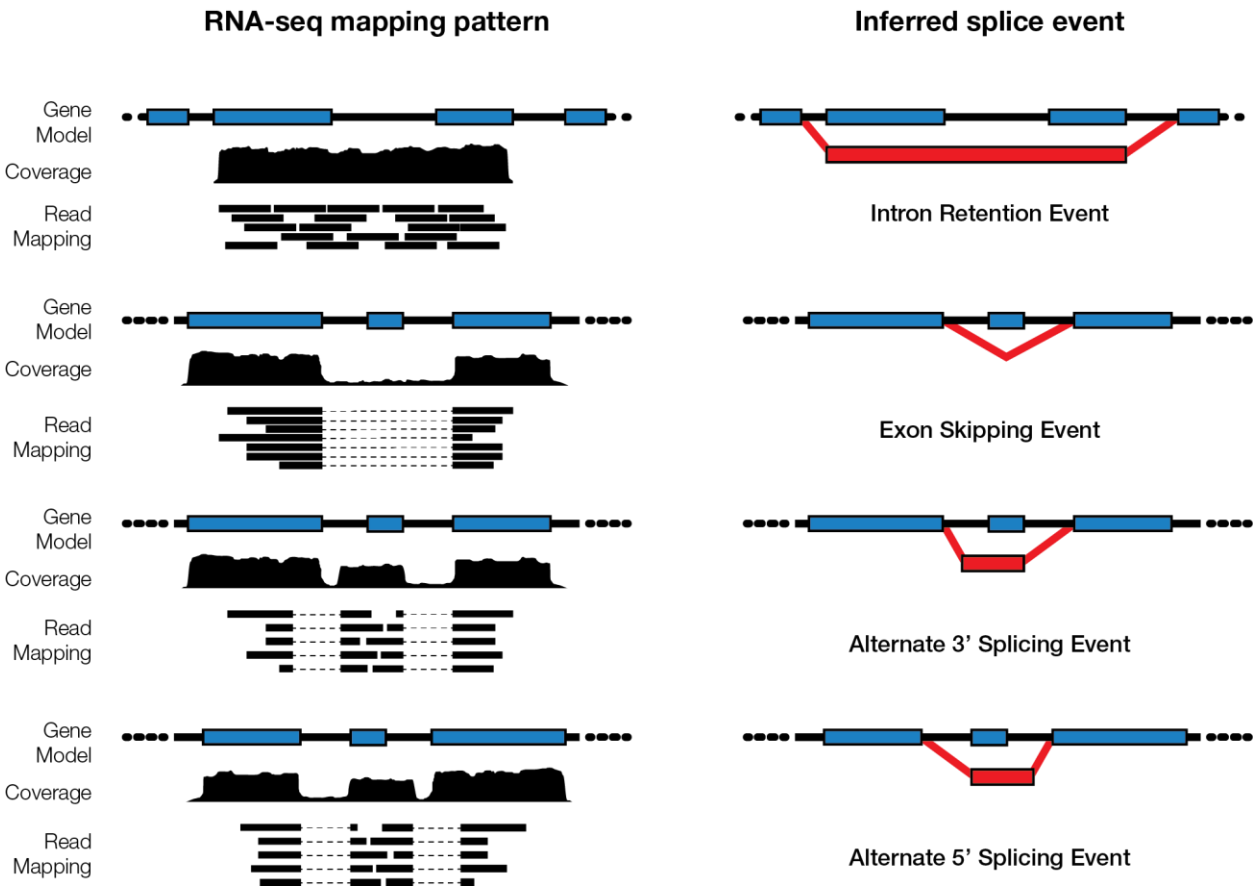
Table of Contents

Supplementary Table 1. **Data sources, programs, and statistical methods used in this study.**

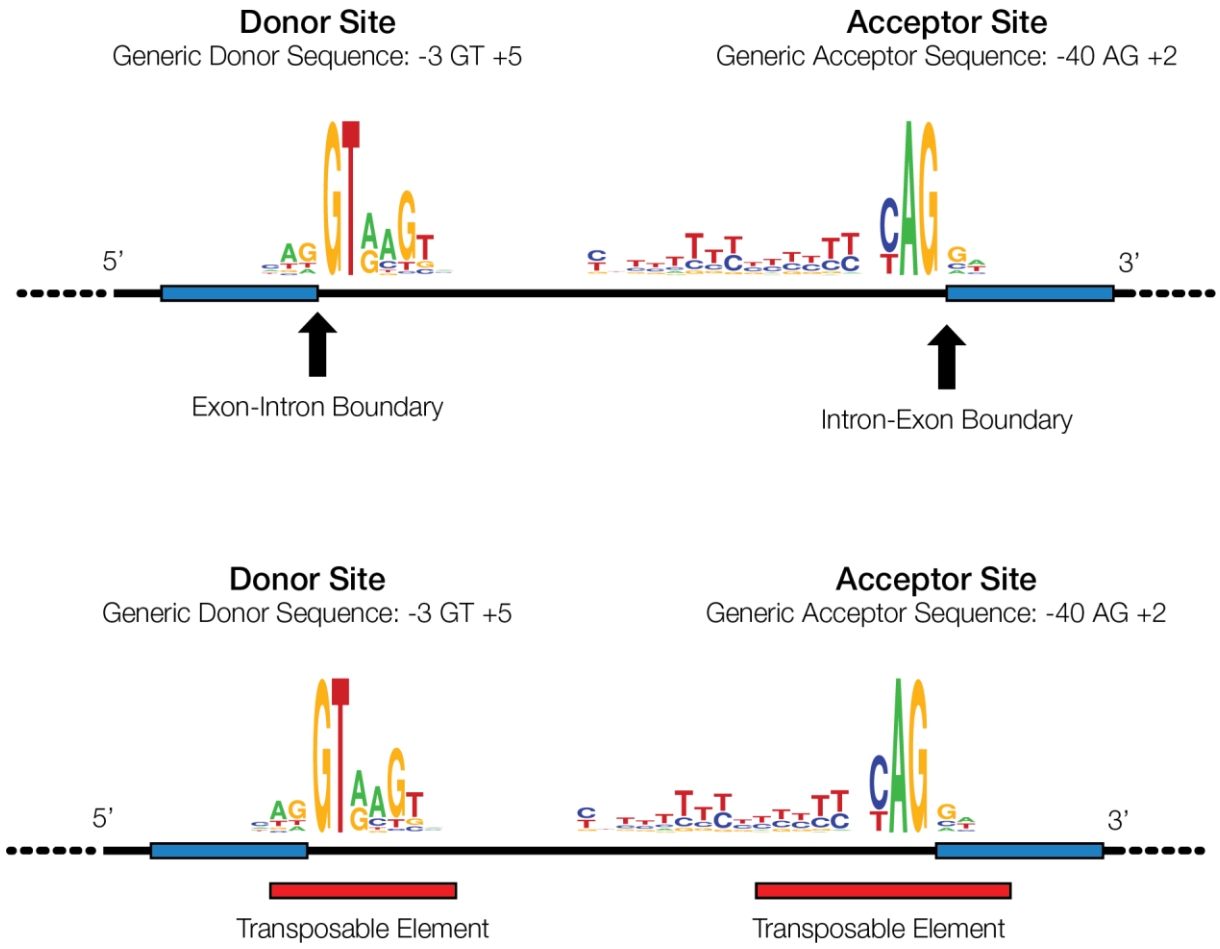
Name	Description	Reference
	Data Sources	
TCGA	RNA-seq data from matched normal-tumor patient samples from 13 cancer types	[1]
COSMIC (v88)	Cancer Gene Census (CGC) annotations for 723 cancer-associated genes	[2]
RepeatMasker (3.2.7)	Genomic coordinates and annotations for human TEs	[3]
NCBI RefSeq (GRCh37.p13 – 2017-04-19)	Genomic coordinates (exon/intron boundaries) for human genes	[4]
GENCODE (v19)	Genomic coordinates (exon/intron boundaries) for human genes	[5]
Genomic Data Commons (GDC)	Coordinates and event counts of alternative splicing events in TCGA samples	[6]
	Programs	
SplAdder (1.2.0)	Detection and quantification of alternative splicing events	[7]
BEDTools (2.28)	Identification of alternative splicing events colocated near TE sequences	[8]
DESeq2 (3.9)	Normalization of alternative splice isoform expression across patients within a cancer type	[9]
UCSC Genome Browser (hg19)	Visualization of putative TE-derived isoforms in cancer associated genes	[10]
	Statistical Methods	
Variance Stabilizing Transformation (VST)	Blind transformation used to remove the experiment-wide trend of variance over mean, normalize alternative splice isoform expression values and produce log2 scale data	[9]
Single linkage clustering	Algorithm to merge overlapping alternative splice isoforms based on $\geq$75% overlap of isoform genomic coordinates in an agglomerative fashion	
Relative expression change (REC)	Quantification of the normalized change in expression levels of TE-derived alternative splice isoforms in tumor versus normal tissue	
G-test	Maximum likelihood statistical significance test for 2x2 isoform expression contingency matrix	



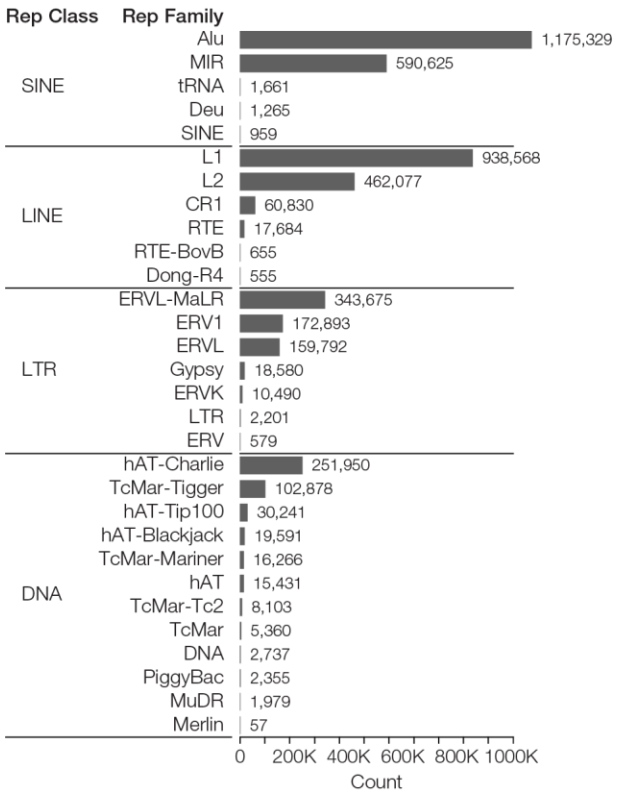
Supplementary Figure 1. **Number of patient samples per cancer type analyzed here.** RNA-seq data for matched normal-tumor sample pairs were taken from The Cancer Genome Atlas (TCGA). Cancers with less than 10 sample pairs were excluded from further analysis.



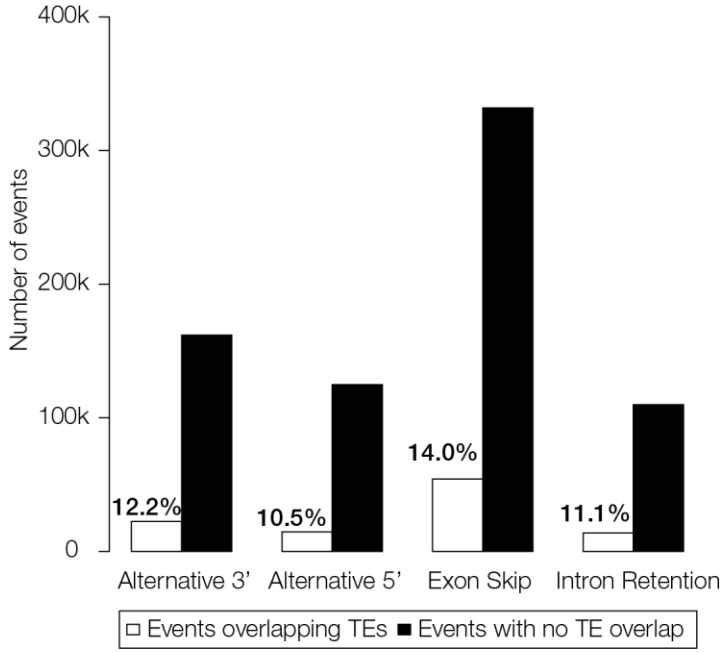
Supplementary Figure 2. **Alternative splicing event types analyzed here**. Four kinds of alternative splice events were analyzed for this study: intron retention, exon skipping, alternate 3' splicing, and alternate 5' splicing. Splicing events were identified and characterized based on the mapping of RNA-seq reads to gene models, using the program SplAdder as previously described [7]. For each type of splicing event, its corresponding RNA-seq read mapping pattern is shown adjacent to a schematic of the inferred splicing event type.



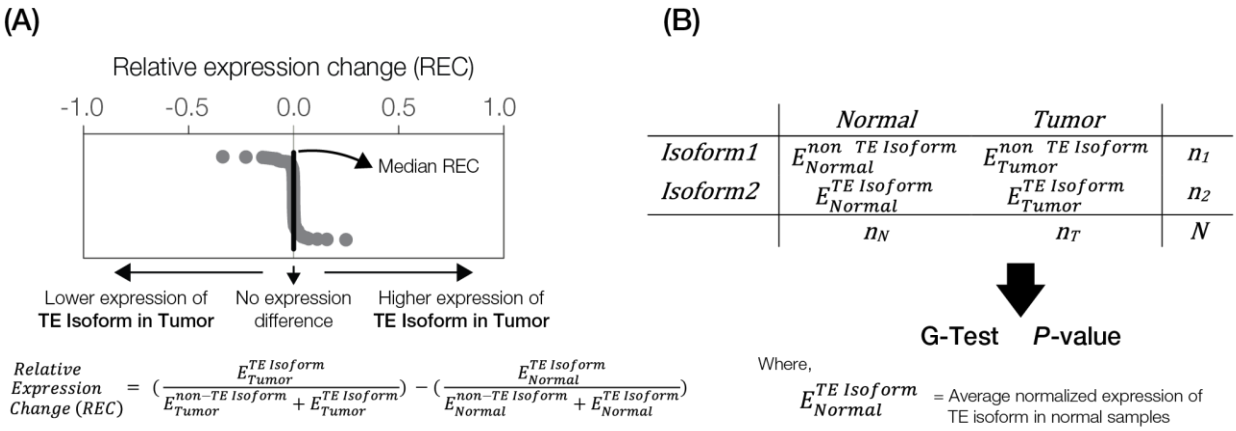
Supplementary Figure 3. **Scheme for the identification TE-derived splice sites.** The top panel shows 3' and 5' exon boundaries along with their canonical splice donor and acceptor site sequence motifs [11]. Potential TE-derived splice donor and acceptor sites were identified where TE sequences were found to overlap the canonical splice site motifs as shown in the bottom panel.



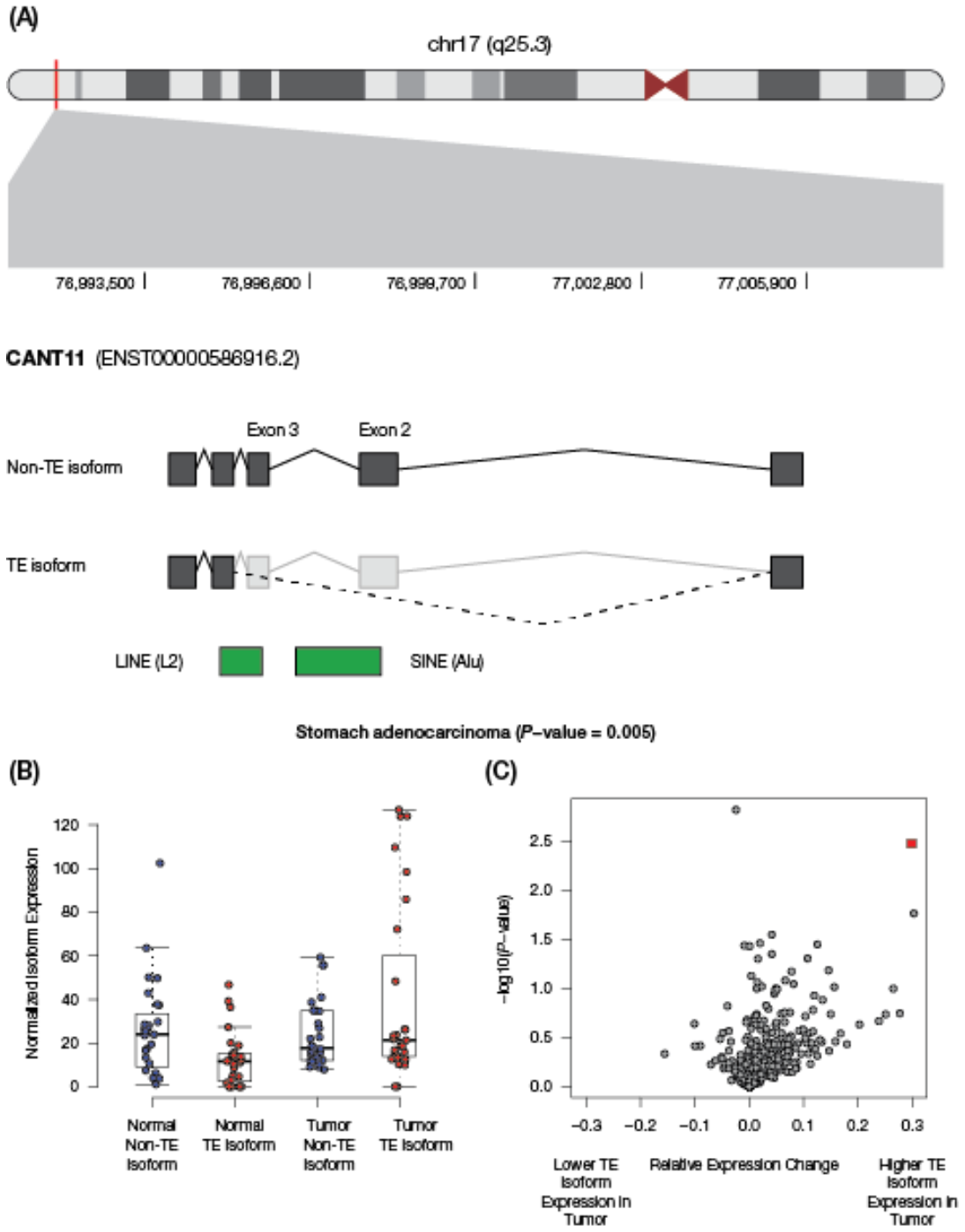
Supplementary Figure 4. **Counts of human transposable element (TE) sequences in the human genome.** TE names and counts are taken from RepeatMakser annotations. TEs are grouped into four major classes, and TE family names are shown for each class. The four major classes are: SINE – short interspersed nuclear element, LINE – long interspersed nuclear element, LTR – long terminal repeat containing element, and DNA – DNA-type element. SINEs, LINEs, and LTRs are retrotransposons that transpose via a copy and paste mechanism catalyzed by reverse transcriptase; DNA-type elements transpose via a cut and paste mechanism catalyzed the transposase enzyme.



Supplementary Figure 5. **Number of alternative splice events seen for human genes.** Counts for the four different alternative splice event types are shown for TE-derived (white) versus non TE-derived isoforms (black). The percentages of TE-derived events are shown.



Supplementary Figure 6. **Quantification and statistical testing for differential expression of TE-derived alternative splice events.** (A) The relative Expression Change (REC) metric quantifies the normalized change in expression levels of TE-derived alternative splice isoforms in tumor versus normal tissue. This metric accounts for the expression of TE and non-TE isoform in both normal and tumor issues. Higher REC values indicate relatively higher expression of TE isoform in tumor tissue and vice versa. Details on the expression counts and formulas can be found in the Methods section. (B) Formulation of the 2x2 contingency matrix used for the G-test of the significance of expression difference.



Supplementary Figure 7. **TE-derived alternative splicing in the *CANT1* gene.** (A) The location of *CANT1* on the long arm of chromosome 17 is shown along with the specific location of its TE-derived alternative splicing event. The presence of LINE and SINE sequences result in an exon skipping event. (B) Distributions of the non-TE (blue) and TE-derived (red) isoforms are shown for matched normal (left) and stomach adenocarcinoma samples (right). (C) Relative expression change (REC) values are plotted against the corresponding G-test *P*-values (see Methods and Supplementary Figure 6) for the matched normal and stomach adenocarcinoma samples. The *CANT1* TE-derived isoform values are shown as a red square.

Supplementary Table 2. **Candidate TE-derived isoform switching in cancer.** Supporting data are shown for the TE-derived isoform events described for KLK2 (Figure 4), MYH11 (Figure 5), WHSC1 (Figure 6), and CANT1 (Supplementary Figure 7).

Cluster[a]	Gene	Cancer Type	%TEi-N[b]	%TEi-T[c]	Event Type	TE coords[d]	Exon coords[e]
467	*MYH11*	Lung squamous cell carcinoma	6.5	51.8	Alt3	chr16:15802811-15803103	chr16:15931764-15932126
412	*CANT1*	Stomach adenocarcinoma	67.3	37.4	Exon (Alu)	chr17:76995725-76996032	chr17:76993921-76994045
					Exon (L2)	chr17:76994098-76994182	chr17:76994229-76994368
132	*WHSC1*	Stomach adenocarcinoma	73.1	42.8	Exon	chr4:1913852-1914626	chr4:1913817-1913921
412	*CANT1*	Breast invasive carcinoma	53.5	32.1	Exon	Refer to cluster 412 above	
154	*KMT2D*	Stomach adenocarcinoma	21.3	31.8	Alt5	chr12:49417694-49417936	chr12:49415825-49415934
397	*POLG*	Stomach adenocarcinoma	34.8	54.6	Alt3	chr15:89861214-89861324	chr15:89862161-89862330
397	*POLG*	Bladder Urothelial Carcinoma	34.8	47.6	Alt3	Refer to cluster 397 above	
261	*PML*	Kidney renal papillary cell carcinoma	57.2	70.3	Intron	chr15:74327328-74327503	chr15:74326818-74326871 chr15:74327512-74328735
261	*PML*	Breast invasive carcinoma	47.0	59.6	Intron	Refer to cluster 261 above	
261	*PML*	Kidney Chromophobe	65.5	75.8	Intron	Refer to cluster 261 above	

[a]Cluster id number corresponding to the distinct TE-derived alternative splicing event.
[b]Relative expression (percentage of total) for the TE-derived isoform in normal tissue.
[c]Relative expression (percentage of total) for the TE-derived isoform in tumor tissue.
[d]Coordinates of the associated TE
[e]Coordinates of the alternatively spliced exon(s)

References

[1] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M. & Network, C. G. A. R. 2013 The cancer genome atlas pan-cancer analysis project. *Nature genetics* **45**, 1113.

[2] Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C. & Ward, S. 2014 COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research* **43**, D805-D811.

[3] Smit, A., Hubley, R. & Green, P. 2015 RepeatMasker Open-4.0. 2013–2015. (

[4] O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. 2016 Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745. (DOI:10.1093/nar/gkv1189).

[5] Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., et al. 2019 GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773. (DOI:10.1093/nar/gky955).

[6] Kahles, A., Lehmann, K.-V., Toussaint, N. C., Hüser, M., Stark, S. G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C. & Caesar-Johnson, S. J. 2018 Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer cell* **34**, 211-224. e216.

[7] Kahles, A., Ong, C. S., Zhong, Y. & Rätsch, G. 2016 SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. *Bioinformatics* **32**, 1840-1847.

[8] Quinlan, A. R. 2014 BEDTools: the Swiss-army tool for genome feature analysis. *Current protocols in bioinformatics* **47**, 11.12. 11-11.12. 34.

[9] Love, M. I., Huber, W. & Anders, S. 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550.

[10] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D. 2002 The human genome browser at UCSC. *Genome research* **12**, 996-1006.

[11] Stephens, R. M. & Schneider, T. D. 1992 Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J Mol Biol* **228**, 1124-1136. (DOI:10.1016/0022-2836(92)90320-j).