Supplementary Figures and Methods

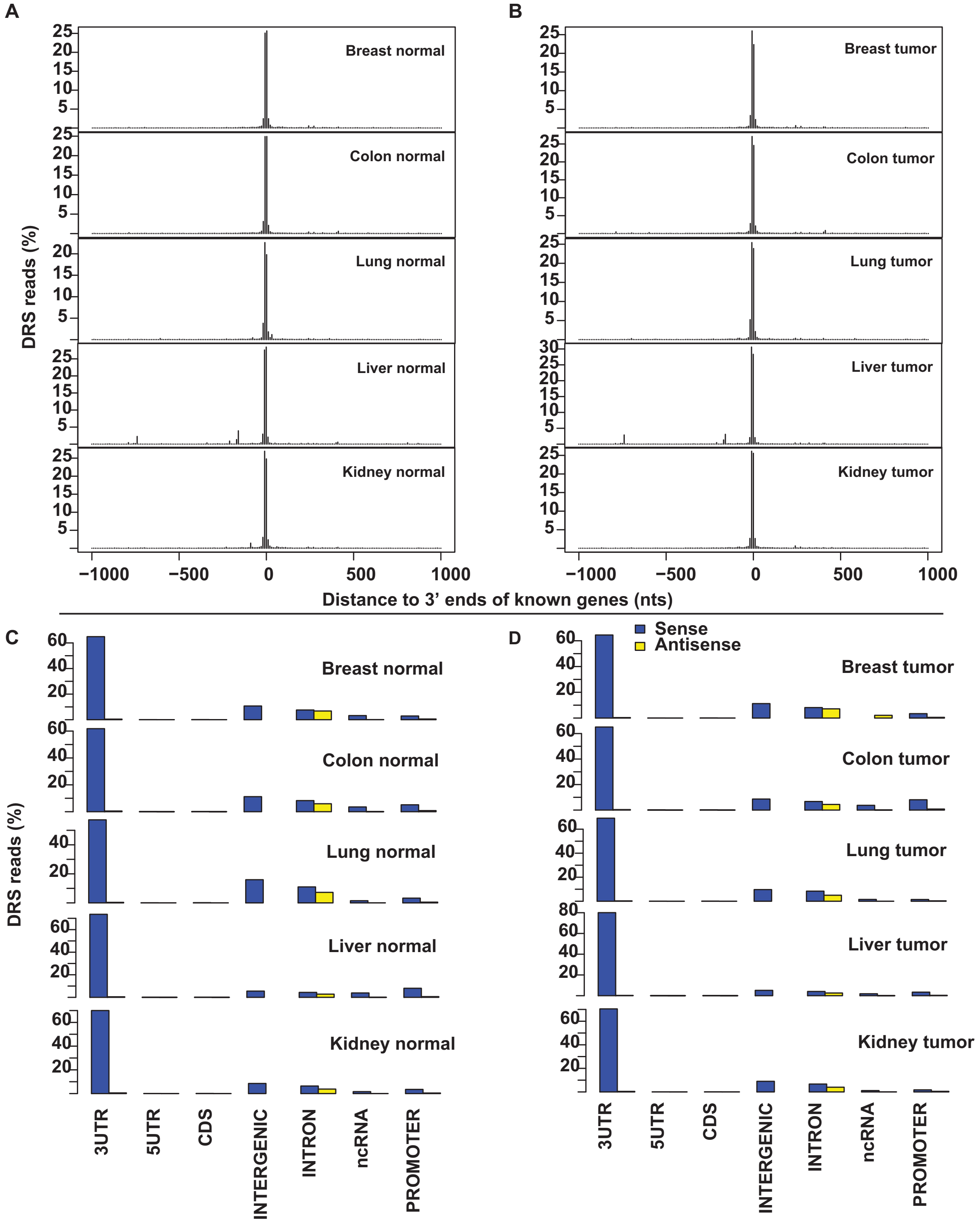

Figure S1: Polyadenylation sites manifest similar overall characteristics across all tissues A-B) DRS reads predominantly match (bin size=10 nts) to annotated 3' ends of known genes across all tissue types in both normal **(A)** and tumor **(B). C-D)** Majority of DRS reads match to sense strand of transcriptionally active regions.

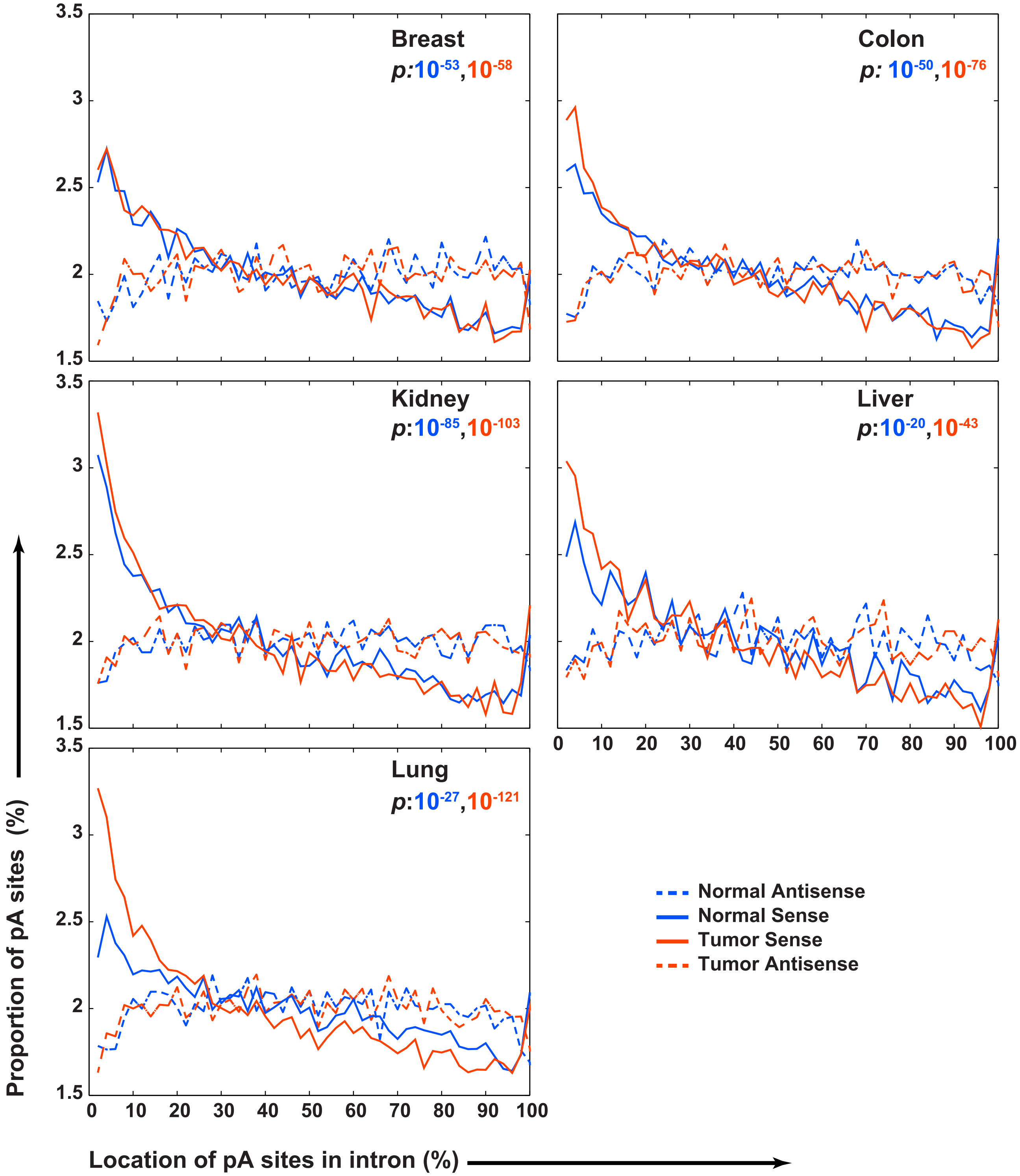

Figure S2: Distribution of the locations of polyadenylation sites in sense and antisense intronic transcripts. The full-range of each intron is standardized to 0-100%, and the polyadenylation site location is with respect to the start of the intron (sense strand; bin size=2%). For example, in normal breast, ~3% of pA sites that occur in sense introns are located within the first 2% of the full-length of intron. For each pair of sense and antisense distributions in either normal or tumor tissue, one-sided two-sample Kolmogorov-Smirnov test was performed to analyze if the observed preference of sense intronic transcripts to occur towards intron start is statistically significant in comparison to that of antisense intronic pA sites.

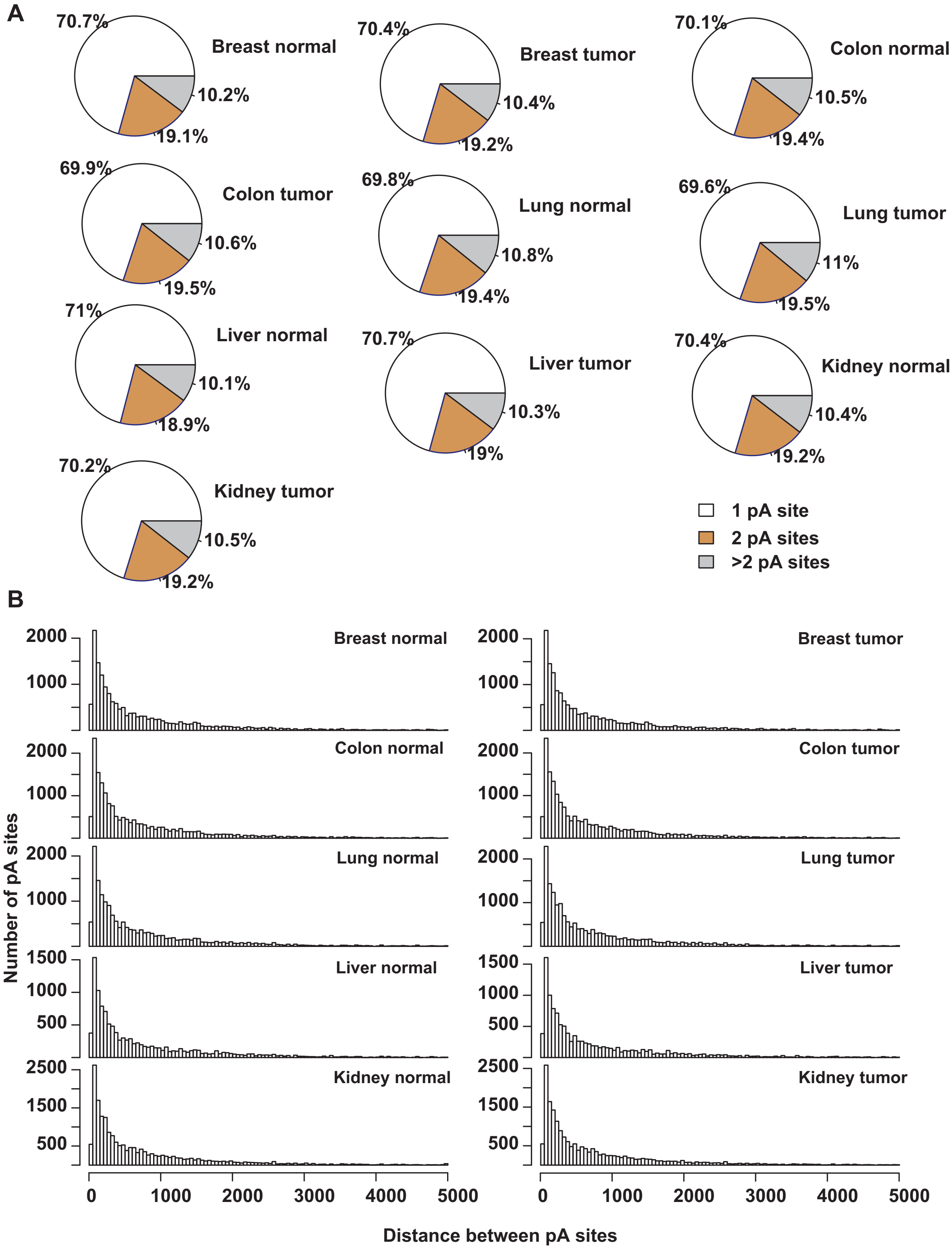

Figure S3: Overall properties of polyadenylation sites are consistent across tissues. **A)** Approximately 30% of genes contain tandem APA sites. **B)** Distribution of distances between adjacent tandem polyadenylation sites in the 3' UTR of genes expressed (bin size=50 nts). Adjacent pA sites that are separated by at least 5000 nts correspond to a small fraction (0.9-1.1%).

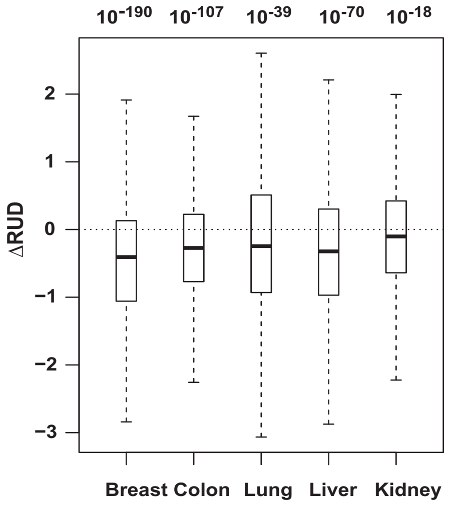

Figure S4: Change in relative abundance of long and short isoforms between normal and tumor samples. The median of ΔRUD is consistently below zero in all five cancer samples confirming that the relative expression ratio of long and short isoforms is lower in the patient samples tested.

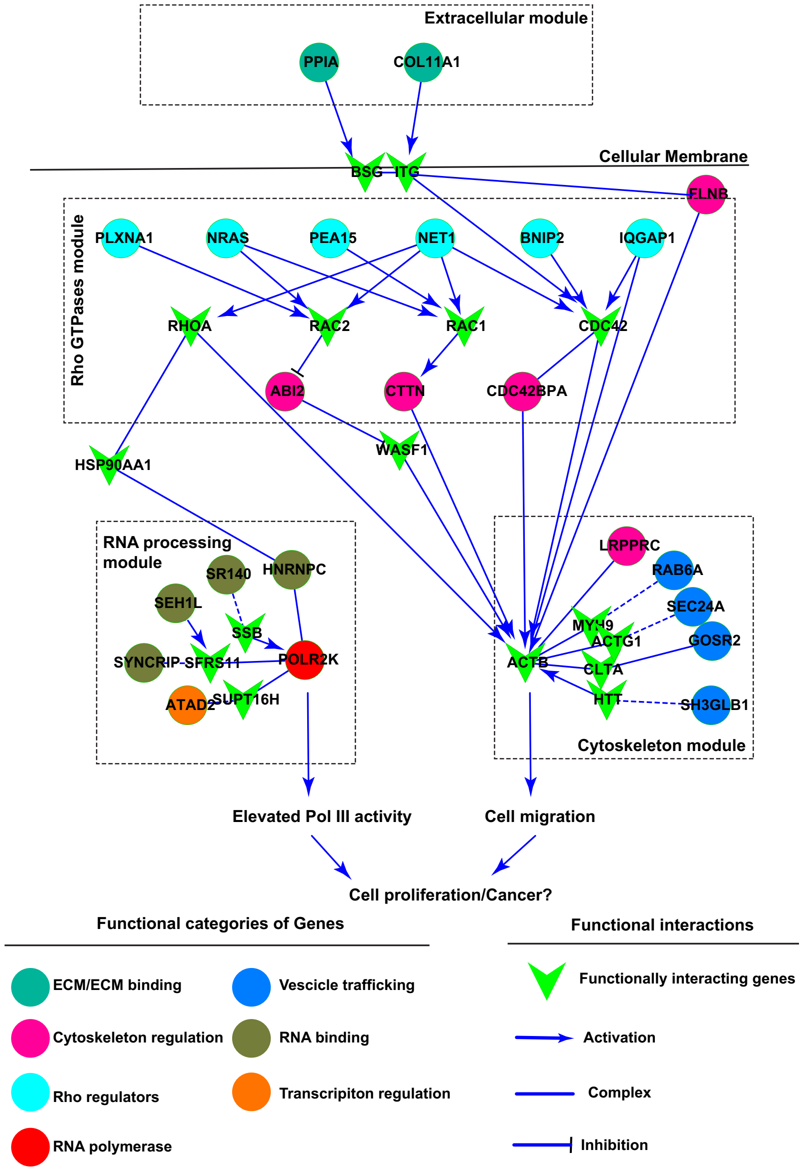

Figure S5: Functional network genes affected by the preferential up-regulation of short isoforms. Network consists of genes that manifest preferential up-regulation of their short isoforms (circles) and other genes (V-shaped nodes) that physically or functionally interact with them. Functional interactions correspond to activation and inactivation, while physical interaction corresponds to co-occurrence within a molecular complex. Note that activation is used in the broad sense: activation may include up-regulation, catalytic activation, or stabilization. The network can be divided into four interlinked modules (dotted lines). The activation of the modules can elevate transcription of Pol III and re-organize actin, both of which are prominent events in cancer progression.

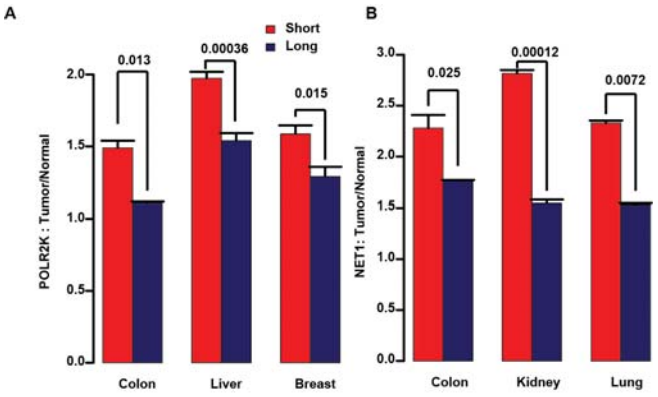

Figure S6: Additional validation for POLR2K and NET1. As noted POLR2K and NET1 short isoforms were found to be upregulated by DRS in at least three tissues**.** Real-time RT-PCR results for two genes, POLR2K (**A**) and NET1 (**B**) across the respective three tissues. Fold changes between the expression levels of each isoform in normal and tumor are illustrated (p-values on top). Error bars represent standard error (n=3).

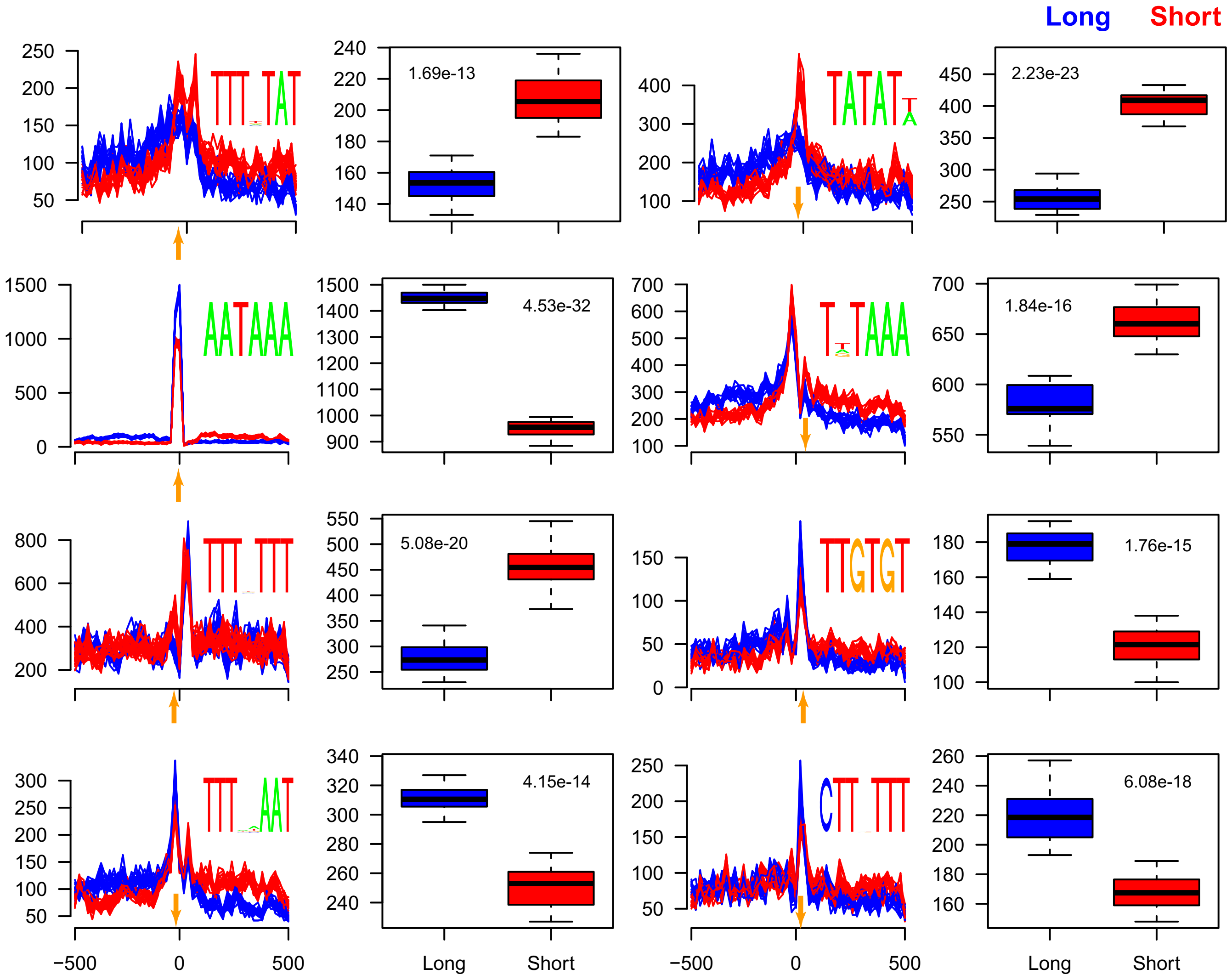

Figure S7: Bootstrapping-based analysis to quantify the preference of each motif towards a given isoform. Ten different bootstrapping samples for both short (red) and long (blue) isoforms, and the average number of occurrences at each location (arrow) was calculated and then compared between the two isoforms using the student t-test.

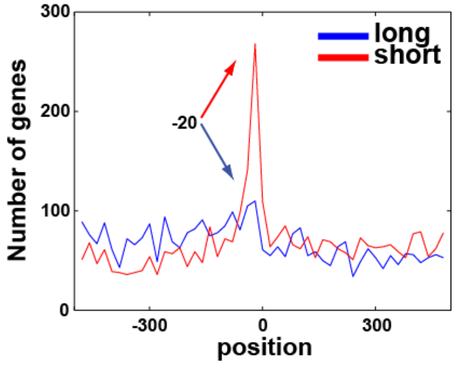

Figure S8: Strong positional preference of ATATAT motif. The motif occurs 20 nts upstream of the polyadenylation site, exclusive to short isoforms. We note that a considerable proportion (~12% or ~400) of genes containing 3' UTR isoforms used for the analysis seem affected by this motif.

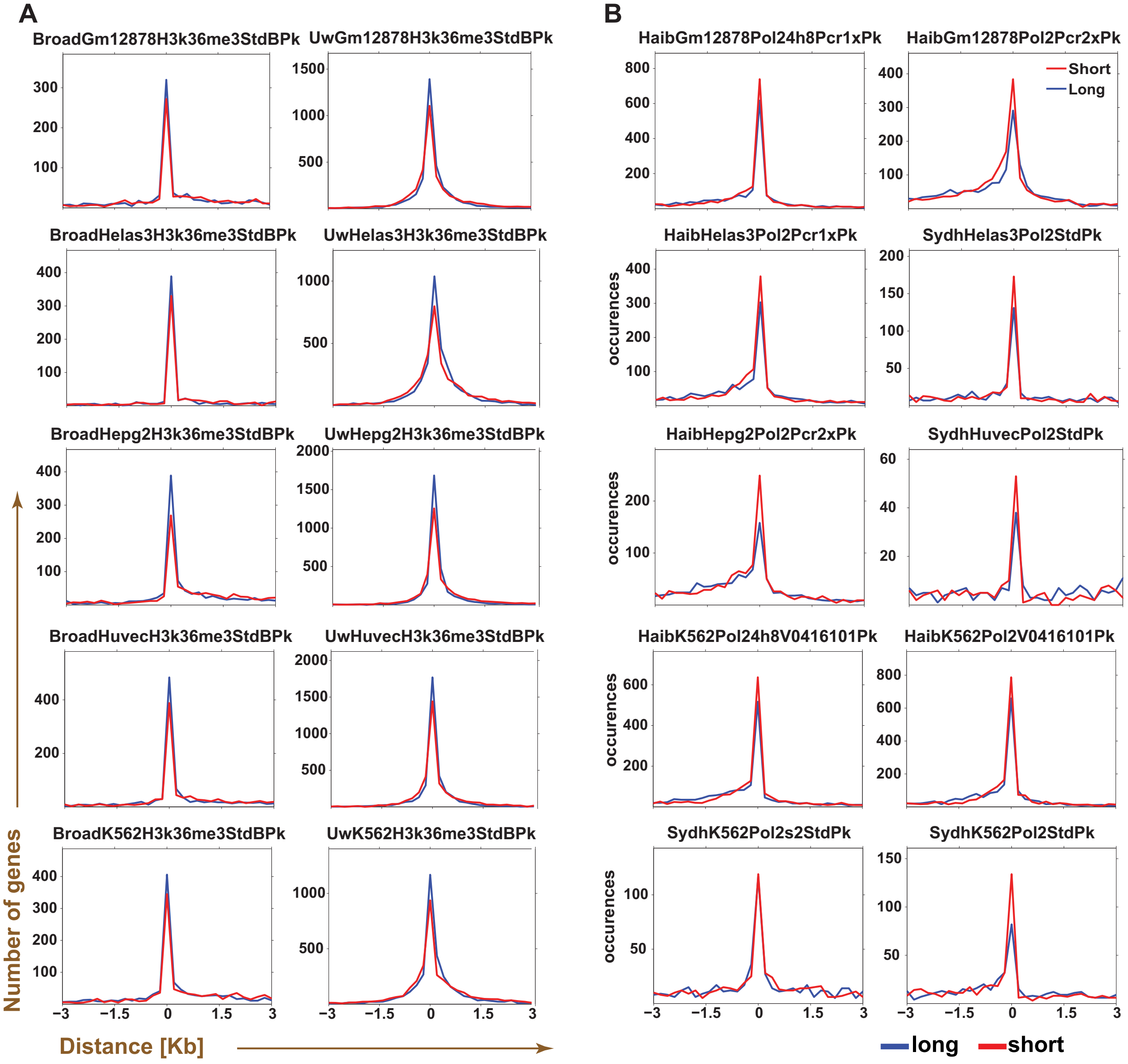

Figure S9: Distance distribution of the H3K36Me3 and Pol2 marks next to short and long isoform polyadenylation sites in multiple cell lines.

Filtering of Reads

To help retain only high quality reads, in addition to Helisphere software, we used in house tools to remove additional reads that occur due to imperfections in surface that results in reads that resemble the sequence of sequencing reactions, resulting in perfect or imperfect repeats of AGCT. Although Helisphere software already eliminates such reads, we have observed using other data sets that the use of the method outlined below can be helpful. For each read, we determined the extent of divergence from the repeating sequence of AGCT; Kullback–Leibler (KL) divergence measure (MATLAB) was used based on the two different frequency distributions ($\boldsymbol{P},\boldsymbol{Q}$) of tri-nucleotides of the read ($\boldsymbol{P}$) and that of the repeat sequence of AGCT [$\boldsymbol{Q}(i)$ =0.25 for i= AGC/GCT/CTA/TAG; $\boldsymbol{Q}(i)$ =0 otherwise] KL score was defined as $\sum_i \boldsymbol{P}(\boldsymbol{i}).log\,\frac{\boldsymbol{P}(i)}{\boldsymbol{Q}(i)}$, where $\boldsymbol{i}$ represents one of the 64 possible tri-nucleotides. Only reads above a KL divergence threshold of 40 was considered for further filtering; threshold was determined by tests on randomly generated sequence datasets, which on average retained ~90% of reads at the threshold of 40. The reads were then further filtered using genome mapping, as described in text.

Determination of usage of short and long 3' UTR isoforms

For genes with more than one tandem polyadenylation site, *Short* and *Long* represents the total number of reads (quantile normalized) from the proximal and distal polyadenylation sites, respectively. To compare the relative usage of the distal polyadenylation site between normal and tumor tissues, we used the previously reported RUD (70) indices (RUD= log2[*Long*/*Short*]; ∆RUD= $\mathrm{RUD}_{\mathrm{tumor}} - \mathrm{RUD}_{\mathrm{normal}}$). RUD is used in the strict sense of its definition; manipulations to emulate mixed signals from short isoforms in parent studies that used RUD were avoided for sake of simplicity, particularly since our conclusions were consistent with previous studies. For example, ∆RUD ≥ 1, corresponds to at least two-fold increase in *Long*/*Short* ratio in tumor, and hence represents those genes that favor the expression of their long isoforms in tumor. To determine differences in expression levels of either short (∆*Short*) or long (∆*Long*) isoform levels between normal and tumor tissues, we computed the log2-transformed fold-change ratio (∆Log*Short*=log[$Short_{\mathrm{tumor}}/Short_{\mathrm{normal}}$]; ∆Log*Long*=log[$Long_{\mathrm{tumor}}/\ Long_{\mathrm{normal}}$]). For network analysis, we selected all short isoforms that are preferentially more up-regulated in tumors than their long forms ($2^{\Delta RUD}$ ≥ 1.5) and are at least 1.5 fold up-regulated in their absolute abundance ($Short^{\mathrm{Tumor}}/Short^{Normal}$ ≥ 1.5). These genes were used to perform functional annotation and pathway enrichment analysis using Cytoscape (71) with the Functional Interaction API from Reactome (72). Only those statistically significant pathways (binomial test p-value ≤ 0.05) were considered as relevant pathways. The complete network was manually curated by close inspection of the sub-networks and published reports of the underlying interactions.

Real-time RT-PCR

For validation of novel gene regions and their differential expression in cancer, we used total RNA from human breast and colon (Biochain, Newark, CA). The colon sample used is identical to that used for DRS sequencing, while the breast RNA is obtained from an independent patient due to limits in the availability of RNA from breast samples. Total RNA (1µg) was reverse-transcribed with random primers using the iScript Select cDNA Synthesis Kit (BIO-RAD, CA). The primers that were designed using Real-Time PCR Assay Design Tool from Integrated DNA Technologies were: TGTACCTTCACAGCCACTCCATCA (forward), GTTAATGATGCCCACAGGATCCAC (reverse) for pA site at Chr9: 29,257,640; TCT ATG GCA TTC CAG CGG AT (forward), TCAGGCGTTGCTGAATACTGTCCA (reverse) for pA site at Chr7:41,724,750. qPCR amplification was monitored using Maxima SYBR Green qPCR Master Mix with ROX Reference Dye (Fermentas, MD). The reactions were performed in 96-well plates (ABI, CA) using a Stratagene MX3000P. The PCR conditions were 10 min at 95°C, followed by 40 cycles of 95°C for 30 sec, 60°C for 1 min and 72°C for 1 min. The data were normalized to ACTB and relative changes in gene expression were quantified using the ΔΔCT method. Since the fold changes were high, additional housekeeping genes were not selected for normalization.

To re-confirm the observed preferential up-regulation of short isoforms for candidate genes, we designed primers immediately before the proximal and distal poly(A) sites on the genome. The primer sequences used are as follows. POLR2K: TTT GAT GCT CGA TGA ATG CTG GGA A (short isoform forward); ACA ACA ATC AGA AAT GGG AAG AGA GCA A (short isoform reverse); ACA CAA TAA AGG GTA AAG TTG CTT CCC CA (long isoform forward); TGC ATT TTT AAC AAG GAC TGC AGA TGA TTC (long isoform reverse). NET1: CGG CCA TTG CCC CCT TCC AG (short isoform forward); CGG ATG CCG GGC TGT GTC TG (short isoform reverse); TGT AGA AGG CTC GAG GGG ACG T (long isoform forward); CGG GCT TTT ACT CCT CCT CCC CC (long isoform reverse). A preliminary study was first performed to identify the least invariant (normal *vs.* tumor) housekeeping genes in each tissue type. The selected housekeeping genes were: GUSB (colon, liver, and lung), HPRT (kidney), and PRLPO (breast). The primer sequences for the housekeeping genes consisted of, GUSB: TCA GCA CCA CTC TTG TGC GCA GGT (forward), TGG ACA TGC CAG TTC CCT CCA GCT T (reverse); HPRT: TGT GAT GAA GGA GAT GGG AGG CCA (forward), CAG TGC TTT GAT GTA ATC CAG CAG GTC AGC (reverse); RPLPO: TGG CCT TGC GCA TCA TGG TGT TCT (forward), (reverse) AGC AGA CAA TGT GGG CTC CAA GCA. The amount of RNA, and conditions used for PCR was the identical to previous set up (above). Genomic DNA was used as the reference sample to construct a standard curve for each primer pair. After converting Ct values to copy numbers based on standard curves, the copy numbers for long isoforms were subtracted from those for short isoforms to approximate the copy numbers for short isoforms. As noted earlier, similar to microarrays, since PCR primers that targets short isoforms also amplify the long isoforms in a primer-dependent manner, the absolute quantification of short isoforms by PCR will likely contain artifacts. To further reduce artifacts, the estimated copy numbers were normalized between tumor and normal tissues to determine the fold change in expression between the tissues. The copy numbers for short or long isoforms were normalized against housekeeping genes before determining fold changes.

References

70. Ji, Z., Lee, J.Y., Pan, Z., Jiang, B. and Tian, B. (2009) Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci U S A*, **106**, 7028-7033.
71. Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc*, **2**, 2366-2382.
72. Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*, **37**, D619-622.