Goering et al. BMC Genomics (2021) 22:476
https://doi.org/10.1186/s12864-021-07781-1

BMC Genomics

METHODOLOGY ARTICLE Open Access

# LABRAT reveals association of alternative polyadenylation with transcript localization, RNA binding protein expression, transcription speed, and cancer survival

Raeann Goering[1,2†], Krysta L. Engel[1†], Austin E. Gillen[2,3], Nova Fong[1], David L. Bentley[1,2] and J. Matthew Taliaferro[1,2*]

## Abstract

**Background:** The sequence content of the 3′ UTRs of many mRNA transcripts is regulated through alternative polyadenylation (APA). The study of this process using RNAseq data, though, has been historically challenging.

**Results:** To combat this problem, we developed LABRAT, an APA isoform quantification method. LABRAT takes advantage of newly developed transcriptome quantification techniques to accurately determine relative APA site usage and how it varies across conditions. Using LABRAT, we found consistent relationships between gene-distal APA and subcellular RNA localization in multiple cell types. We also observed connections between transcription speed and APA site choice as well as tumor-specific transcriptome-wide shifts in APA isoform abundance in hundreds of patient-derived tumor samples that were associated with patient prognosis. We investigated the effects of APA on transcript expression and found a weak overall relationship, although many individual genes showed strong correlations between relative APA isoform abundance and overall gene expression. We interrogated the roles of 191 RNA-binding proteins in the regulation of APA isoforms, finding that dozens promote broad, directional shifts in relative APA isoform abundance both in vitro and in patient-derived samples. Finally, we find that APA site shifts in the two classes of APA, tandem UTRs and alternative last exons, are strongly correlated across many contexts, suggesting that they are coregulated.

**Conclusions:** We conclude that LABRAT has the ability to accurately quantify APA isoform ratios from RNAseq data across a variety of sample types. Further, LABRAT is able to derive biologically meaningful insights that connect APA isoform regulation to cellular and molecular phenotypes.

## Background

During the co-transcriptional processing of a pre-mRNA, the 3′ end of the transcript is cleaved and a polyadenine tail is added that promotes the stability and translation of the resulting message [1, 2]. The site where this cleavage occurs determines the sequence content of the 3′ UTR of the transcript. Regulatory *cis*-element sequences can therefore be either included or excluded from the 3′ UTR of the transcript through modulation of where the cleavage and polyadenylation event happens. This regulation of transcript sequence content through alternative polyadenylation (APA) occurs in the majority of genes in yeast, plant, and mammalian genomes [3–6].

* Correspondence: matthew.taliaferro@cuanschutz.edu
†Raeann Goering and Krysta L. Engel contributed equally to this work.
[1]Department of Biochemistry and Molecular Genetics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
[2]RNA Bioscience Initiative, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
Full list of author information is available at the end of the article

The cleavage and polyadenylation reaction is performed by the core CSTF and CPSF complexes and CFIm which associate with RNA polymerase II (Pol II) transcription complexes [7, 8] and together recognize specific sequence elements within 3′ UTRs to determine sites of 3′ end processing [9]. The abundance of these general CPA factors as well as several other RBPs have been found to regulate the relative usage of alternative polyadenylation sites within a transcript [10–15].

Regulation by these factors results in the large variation in 3′ UTR content seen across tissues and developmental stages [16]. Specific tissues, most notably neuronal tissues, are associated with preferential use of gene-distal or downstream APA sites [17]. Similarly, the broad use of gene-proximal or distal APA sites can be developmentally regulated. Undifferentiated, proliferating cells generally display enriched usage of proximal APA sites while more differentiated cells show shifts towards increased usage of distal APA sites [18, 19]. This phenomenon has also been connected to cancer progression where increased usage of proximal APA sites in key oncogenes was associated with elevated cell proliferation and oncogenic transformation [18, 20].

Alternative polyadenylation exists in two structurally distinct forms. The first, which we will refer to as "tandem UTRs", occurs when multiple APA sites are found within the same terminal exon (Fig. 1B**, top**). The second, which we will refer to as "alternative last exons" or "ALEs", occurs when multiple APA sites are found within different terminal exons (Fig. 1B**, bottom**). Regulation of the choice between alternative tandem UTRs can be viewed as a competition between a proximal upstream poly(A) site that is transcribed first with a distal downstream site that is transcribed second. Similarly the choice between ALE's can be viewed as a competition between recognition of a proximal 5′ splice site and a distal poly(A) site [21]. It is not known whether the two forms of APA are subject to common regulatory mechanisms but in this regard it is interesting to note that transcription speed has been reported to influence the competition between alternative splice sites and tandem poly(A) sites [22, 23].

The majority of published high-throughput RNA sequencing data has been produced using libraries in which the entire transcript is represented. Although these libraries are informative for the regulation of processes like alternative splicing, they are not ideally suited for the quantification of APA isoforms. Alternative library preparation strategies that specifically profile the 3′ ends of transcripts, including 3′ end sequencing and 3′ READS [24, 25], likely provide more accurate quantification of polyadenylation site usage. How well whole-transcript RNAseq data compares in its ability to quantify APA isoforms is generally unknown.

The study of APA using high-throughput RNA sequencing has been facilitated through a handful of software packages aimed at quantifying changes in relative APA site usage across conditions [26–29]. However, quantifying APA from transcriptomic alignments can be difficult. Due to their shared isoform structure, different APA isoforms often contain a considerable amount of sequence in common. If the APA isoform quantification software relies on these transcriptomic alignments [27, 28], this can make assigning reads to a specific isoform challenging. Newer transcriptome quantification techniques that assign reads to transcripts by comparing their sequence contents are better equipped to handle this problem [26, 30, 31]..

To take advantage of this advance in isoform quantification and apply it to the analysis of APA isoforms, we developed LABRAT (Lightweight Alignment-Based Reckoning of Alternative Three-prime ends). A particular advantage of our approach is that it permits rapid analysis of large numbers of publically available RNA-seq data sets including patient samples. Here, we applied this approach to tens of thousands of RNAseq samples to study processes and factors that regulate relative APA isoform abundance as well as the consequences of APA site choice on transcript fate.

## Results

### Quantification of alternative polyadenylation with LABRAT

To quantify relative alternative polyadenylation site usage from RNAseq data, LABRAT takes a genome annotation file and first searches the annotation for tags that define transcripts with ill-defined 3′ ends in order to filter and remove them from further analysis (Fig. 1A). Because annotations are isoform-based, they are often rigid in their explicit connection of upstream alternative splicing events to downstream APA sites, even though this connection may not be accurate. Therefore, to exclude spurious contributions of upstream alternative splicing events to APA site quantification, we extracted the final two exons of every transcript and the expression of these transcript "terminal fragments" was quantified using Salmon [30].

For each gene, alternative polyadenylation sites are then defined using terminal fragments. Terminal fragments with 3′ ends within 25 nt of other 3′ ends are grouped together to define a single polyadenylation site, and the sites are ordered from most gene-proximal to most gene-distal. Each APA site within a gene is assigned a value, $m$, which is defined as its position within this proximal-to-distal ordering, beginning with 0. Each gene is assigned a value, $n$, which is defined as the number of distinct APA sites that it contains. The expression (TPM) of every terminal fragment belonging

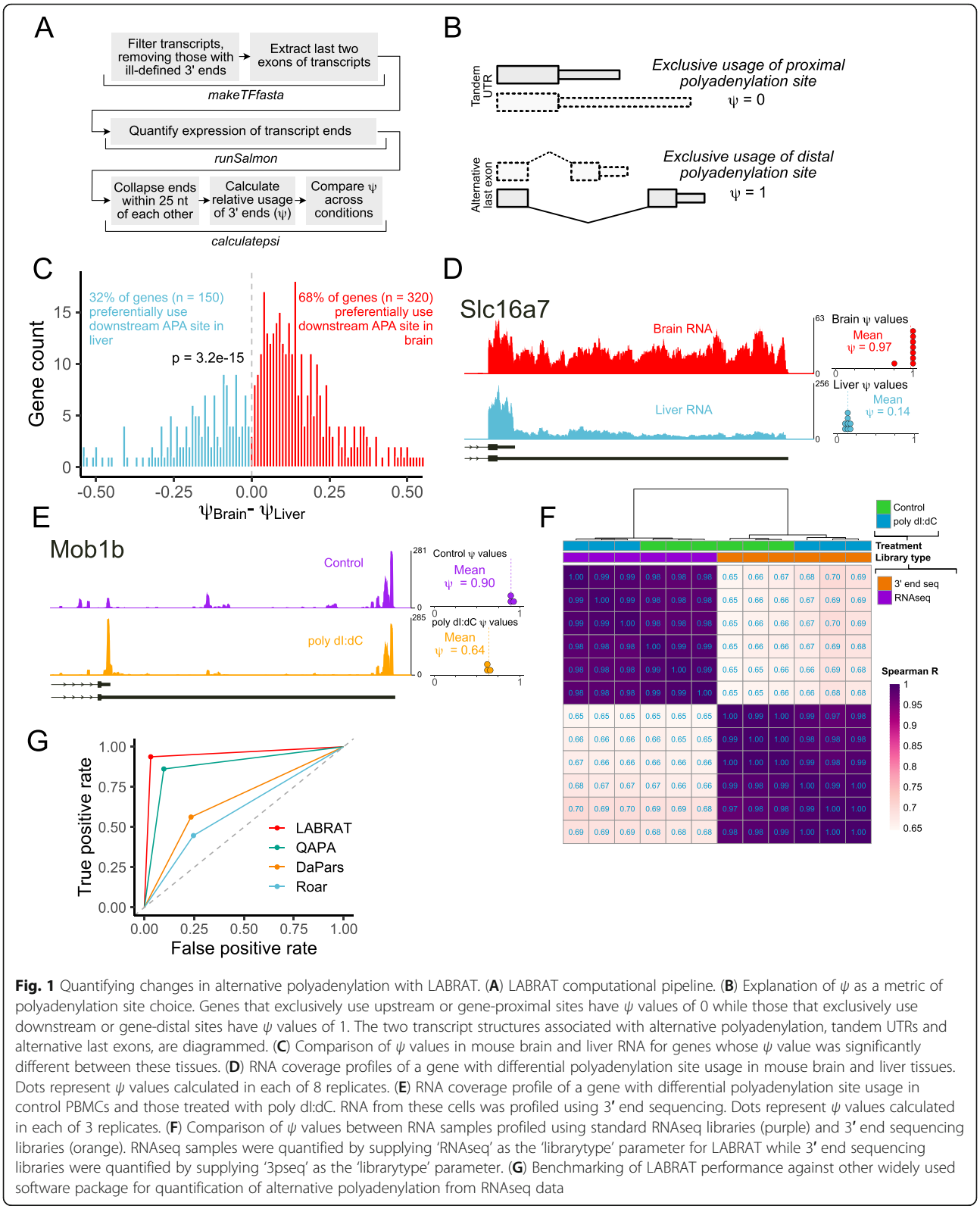


**Fig. 1** Quantifying changes in alternative polyadenylation with LABRAT. (**A**) LABRAT computational pipeline. (**B**) Explanation of $\psi$ as a metric of polyadenylation site choice. Genes that exclusively use upstream or gene-proximal sites have $\psi$ values of 0 while those that exclusively use downstream or gene-distal sites have $\psi$ values of 1. The two transcript structures associated with alternative polyadenylation, tandem UTRs and alternative last exons, are diagrammed. (**C**) Comparison of $\psi$ values in mouse brain and liver RNA for genes whose $\psi$ value was significantly different between these tissues. (**D**) RNA coverage profiles of a gene with differential polyadenylation site usage in mouse brain and liver tissues. Dots represent $\psi$ values calculated in each of 8 replicates. (**E**) RNA coverage profile of a gene with differential polyadenylation site usage in control PBMCs and those treated with poly dI:dC. RNA from these cells was profiled using 3′ end sequencing. Dots represent $\psi$ values calculated in each of 3 replicates. (**F**) Comparison of $\psi$ values between RNA samples profiled using standard RNAseq libraries (purple) and 3′ end sequencing libraries (orange). RNAseq samples were quantified by supplying 'RNAseq' as the 'librarytype' parameter for LABRAT while 3′ end sequencing libraries were quantified by supplying '3pseq' as the 'librarytype' parameter. (**G**) Benchmarking of LABRAT performance against other widely used software package for quantification of alternative polyadenylation from RNAseq data

to a given APA site is then summed to define the expression level of the APA site, and this process is repeated for every APA site within a gene. The expression level of each APA site is then scaled according to the following formula:

$$TPM_{scaled} = TPM_{unscaled}\left(\frac{m}{n-1}\right)$$

To quantify a gene's relative APA site usage, we defined a term, $\psi$. Scaled and unscaled TPM values are summed across all APA sites within a gene, and $\psi$ is defined as the ratio between these summed values:

$$\psi = \frac{\sum TPM_{scaled}}{\sum TPM_{unscaled}}$$

With this strategy, genes that show exclusive usage of the most gene-proximal APA site will be assigned a $\psi$ value of 0, while those that show exclusive usage of the most gene-distal APA site will be assigned a $\psi$ value of 1 (Fig. 1B). Usage of both sites will result in a $\psi$ value between 0 and 1 depending on the relative usage of the sites. Importantly, this strategy also applies to genes with more than 2 APA sites. In these cases, one $\psi$ value is assigned to the entire gene without the need to do multiple pairwise comparisons between APA sites.

After calculating $\psi$ values for genes in all samples, LABRAT compares $\psi$ values of experimental replicates across experimental conditions to identify genes with statistically significantly different $\psi$ values between conditions. This is done using a mixed linear effects model that tests the relationship between $\psi$ values and experimental condition. A null model is also created in which the term denoting the experimental condition has been removed. A likelihood ratio test compares the goodness of fit of these two models to the observed data and assigns a $p$ value for the probability that the real model is a better fit than the null model. In simple comparisons between two conditions, this approach mimics a t-test. However, this technique has the advantage of being able to easily incorporate covariates into significance testing. After performing this test on all genes, the raw $p$ values are corrected for multiple hypothesis testing using a Benjamini-Hochsberg correction [32].

In addition, LABRAT determines whether a gene's APA sites conform to either the tandem UTR or ALE structures (Fig. 1B) and designates the gene accordingly. For genes with more than 2 APA sites, it is possible to contain both tandem UTR and ALE structures. These genes are designated as having a "mixed" APA structure.

### Identifying tissue-specific differences in APA isoform abundance with LABRAT

To demonstrate the ability of LABRAT to identify and quantify differences in APA isoform abundance, we analyzed RNAseq data from mouse brain and liver tissues [33]. Because neuronal tissues are known to be highly enriched for the use of distal APA sites [17], we reasoned that comparison of these two tissues might provide a positive control for LABRAT's ability to identify differential APA isoform abundance.

We found 470 genes that displayed differential relative APA isoform abundance between the tissues (FDR < 0.05) (Fig. 1C). As expected, 68% of these genes showed increased usage of distal APA sites in brain, indicating a significant enrichment for the use of downstream APA sites in this tissue (binomial $p = 3.2e\text{-}15$). To further explore changes in $\psi$ value for specific genes, we plotted read coverages over two genes that showed significantly more downstream APA site usage in brain tissue: *Slc16a7* and *Elavl1* (Fig. 1D**,** **S1A**). For both genes, we observed significantly lower read coverages corresponding to usage of the distal APA site in the liver samples relative to the brain samples. Accordingly, LABRAT assigned these genes to have low $\psi$ values in the liver samples, and high $\psi$ values in the brain samples, indicating that LABRAT can accurately quantify APA.

To perform similar analyses in human samples, we analyzed over 5000 RNAseq samples from over 30 different human tissues produced as part of the Genotype-Tissue Expression (GTEx) project [34]. We quantified APA isoform abundance in these samples and observed relationships between tissue APA using PCA analysis (**Figure S1B**). In this analysis, brain and testis samples were clear outliers. Interestingly, performing the PCA analysis using only tandem UTR (**Figure S1C**) or ALE (**Figure S1D**) genes produced very similar results, suggesting that these two forms of APA are broadly coregulated across many tissues.

To understand more about APA in human brain and testis, we compared their APA profiles to those observed in human liver samples. As expected, we observed that brain samples exhibited a significant bias for the use of downstream APA sites ($p < 2.2e\text{-}16$) (**Figure S1E**). Conversely, testis samples exhibited a similar bias for the use of upstream APA sites ($p < 2.2e\text{-}16$) (**Figure S1F**). The propensity of testis to use upstream APA sites has been previously observed [35–37] and is likely a key feature of spermatogenesis [38]. Overall, these results demonstrate the ability of LABRAT to recapitulate previously reported observations and gave us confidence in its results moving forward.

### Quantifying APA isoform abundance with 3′ end sequencing data using LABRAT

Although the plethora of available RNAseq datasets make it possible to observe APA trends in a variety of

contexts, RNAseq is not perfectly suited to APA quantification. Library preparations that enrich for reads near cleavage and polyadenylation sites provide a more direct, and potentially more accurate, quantification of APA isoforms [24, 25].

To allow LABRAT to quantify APA isoforms in 3′ end sequencing data, we included the 'librarytype' parameter. If this parameter is designated as '3pseq', LABRAT uses the counts of reads assigned to polyA sites for quantification rather than the length-normalized TPM metric. Because 3′ end data is produced using oligo-dT anchors, length normalization is not necessary and if utilized would unfairly penalize long transcripts.

To assess the accuracy of APA isoform quantification with LABRAT from 3′ end data, we used a dataset in which the authors prepared 3′ end libraries from RNA isolated from human peripheral blood mononuclear cells (PBMCs) with and without treatment with poly dI:dC [39]. Salmon, the transcript quantification tool utilized by LABRAT, has been shown to accurately quantify transcript abundances using 3′ end data [39]. We calculated $\psi$ values from this data using LABRAT (Fig. 1E, **S1G**) and compared them to values produced by the more classical approach of simply counting the aligned reads associated with each APA site (see Methods). We found that the $\psi$ values produced by LABRAT were in strong agreement with those produced by alignment-dependent method ($R \sim 0.92$) (**Figure S1H**). It is important to note here that the deviation from perfect agreement is possibly due to the greater ability of Salmon (and therefore LABRAT) to accurately assign reads that could be consistent with multiple polyadenylation sites.

An open question in the field of APA quantification is the extent to which APA isoforms can be accurately quantified using RNAseq as opposed to 3′ end sequencing. To address this question, we took advantage of the fact that in the PBMC study, the authors prepared RNAseq and 3′ end sequencing libraries from the same RNA samples [39]. We used LABRAT to calculate $\psi$ values from the RNAseq and 3′ end data using 'RNAseq' and '3pseq' librarytype parameters, respectively. We found that $\psi$ values from the two library preparation methods were reasonably and reproducibly correlated ($R \sim 0.67$) while $\psi$ values from samples produced using the same highly correlated ($R \sim 0.97$) (Fig. 1F). RNAseq is therefore able to quantify APA isoform abundance with generally acceptable accuracy. Further, both methods accurately segregated samples into treatment and control groups using $\psi$ values (Fig. 1F), giving confidence in the ability of RNAseq libraries to accurately reflect APA status and opening up tens of thousands of RNAseq datasets for quantification.

## Comparison of LABRAT to similar methods of APA isoform quantification

To compare LABRAT with other APA analysis tools, we generated a synthetic RNAseq dataset containing 50 million reads in which 1250 genes displayed increased distal APA site usage, 1250 genes displayed increased proximal APA site usage, and 2500 genes displayed no change in APA site usage [40]. We used the software packages QAPA [26], DaPars [27], and Roar [28] in addition to LABRAT to quantify APA isoforms in these data.

QAPA, like LABRAT, uses lightweight alignments to quantify APA. Reassuringly, we found that $\psi$ values calculated by LABRAT were highly correlated to the analogous metric used by QAPA, PPAU ($R = 0.81$) (**Figure S1I**). In comparing the four methods, LABRAT was the best suited to accurately identify differential APA in the simulated data (Fig. 1G). We further found that the accuracy of LABRAT was not noticeably affected by read depth down to one million reads (**Figure S1J**).

## Alternative polyadenylation isoforms are differentially localized in cell bodies and projections

Multiple studies have found that alternative polyadenylation decisions made during nuclear processing can influence the subcellular localization of the resulting transcript, particularly in neuronal cells [41–43]. However, it has been unclear how widespread this effect is and whether it was driven primarily by tandem UTRs or ALEs. To address this, we used LABRAT to analyze the relative APA status of 26 paired transcriptomic datasets from cell body and projection samples from neuronal cells, NIH 3 T3 cells, and MDA-MB231 cells [41, 42, 44–50].

For all samples, we identified genes whose $\psi$ value was significantly different between subcellular compartments ($FDR < 0.05$), finding between 10 and 740 genes that fit this criterion in each sample (Fig. 2A). Many of these genes were shared across multiple samples (**Figure S2A**). For these genes, we then compared their $\psi$ values across compartments by subtracting the $\psi$ value in the cell body from the $\psi$ in the projection to define $\Delta\psi$. Genes with positive $\Delta\psi$ values therefore had their distal APA isoform enriched in projections while those with negative $\Delta\psi$ values had their proximal APA isoform enriched in projections.

We found that for 19 of these 26 samples, over 50% of significant genes had positive? $\Delta\psi$ values, indicating a broad connection between the use of distal APA sites and localization of the resulting transcript to cell projections (Fig. 2A), consistent with previous reports [41, 42]. Further, we observed a relationship between the amount of time that the projection had been allowed to grow and the fraction of genes with positive $\Delta\psi$ values. Of the samples in which the projections had grown for 2 days

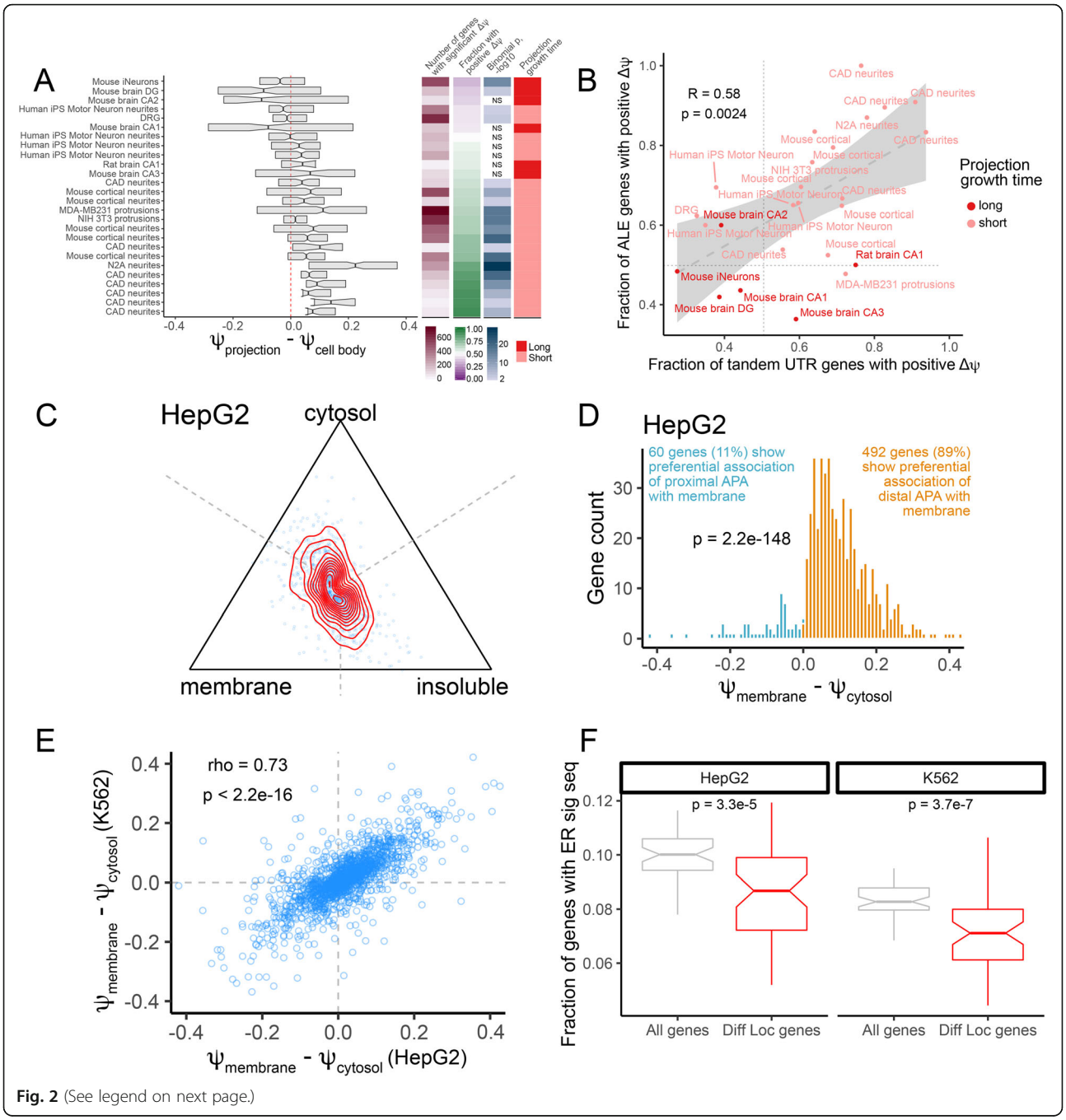


**Fig. 2** (See legend on next page.)

(See figure on previous page.)
**Fig. 2** Alternative polyadenylation is associated with RNA localization in a variety of cell types. (**A**) Comparison of $\psi$ values for RNA isolated from cell projections and cell bodies. $\psi$ values for all genes were calculated using RNA collected from cell projection and cell body compartments, and genes with significantly different $\psi$ values across compartments were identified (FDR < 0.05). $\Delta\psi$ values (cell projection - cell body) for these genes are indicated by boxplots. *P* values in blue represent binomial *p* values for deviations from the expected 50% chance for a gene to have a positive $\Delta\psi$ value. Samples were also separated according to the amount of time that projections were allowed to grow before their RNA content was analyzed. This is represented by the long (at least 6 days) and short (2 days or less) categories colored in red. (**B**) As in A, $\psi$ values for all genes were calculated using RNA collected from cell projection and cell body compartments, and genes with significantly different $\psi$ values across compartments were identified (FDR < 0.05). The fraction of significant tandem UTR and ALE genes with positive $\Delta\psi$ values were plotted on the x and y axes, respectively. (**C**) Simplex plot indicating $\psi$ values calculated from RNA isolated from biochemically defined cytosolic, membrane-associated, and insoluble fractions of HepG2 cells. Genes with equal $\psi$ values in all three fractions are represented by dots equidistant from each vertex (at the intersection of the dotted lines). Genes that displayed higher $\psi$ values in a given fraction than the others are represented by dots placed closer to that fraction's vertex. Red lines indicate the density of dots. (**D**) Comparison of $\psi$ values in HepG2 cytosolic and membrane fractions for genes whose $\psi$ value was significantly different between these compartments (FDR < 0.01). (**E**) Correlation of $\Delta\psi$ values (membrane - cytosol) for all genes expressed in both HepG2 and K562 cells. (**F**) Fraction of genes with nonsignificant $\Delta\psi$ values (membrane vs. cytosol, gray) and those with significant $\Delta\psi$ values (red) that encode peptides that have ER signal sequences as defined by SignalP. Distributions of this fraction were created through bootstrapping in which 40% of the genes were sampled 100 times. *P* values were calculated using a Wilcoxon rank sum test

or less, 15 out of 20 showed a significant bias for the association of distal APA sites with projections. Conversely, of the samples in which the projections had grown for 6 days or more, 0 out of 6 showed a significant bias for the association of distal APA sites with projections. This suggests that distal APA transcripts may play a role in early projection outgrowth but may be less important in mature projections.

Given the conflicting reports about the relative contributions of distal APA produced by tandem UTR and ALEs to the transcriptomes of cell projections [41–43], we analyzed these two classes of APA isoforms separately. Across the 26 subcellular comparisons, we found a strong, significant correlation ($R = 0.58$, $p = 0.0024$) between the fraction of ALE genes with positive $\Delta\psi$ values and the fraction of tandem UTR genes with positive $\Delta\psi$ values (Fig. 2B). This indicates that both classes of genes are preferentially contributing their distal APA isoforms to projections and suggests that these two classes of alternative poly(A) site selection may be regulated by a common mechanism.

## Alternative polyadenylation isoforms are differentially localized in biochemically defined cytosolic and membrane fractions

To further explore connections between APA and RNA localization beyond cell projections, we used LABRAT to analyze RNAseq data from a biochemical fractionation of 3 cell types, *Drosophila* DM-D17-C3 (D17) cells, human HepG2 cells, and human K562 cells [51]. In these data, cells were fractionated into nuclear, cytosolic, membrane-associated and insoluble fractions. RNA was isolated from each of these fractions and prepared for high-throughput sequencing using either polyA-selection-based or ribosomal RNA-depletion-based library preparation. For each fraction, two replicates of each library preparation method were sequenced.

As with the projection data, we compared $\psi$ values for genes across cellular compartments. Hierarchical clustering of samples based on $\psi$ values revealed that samples from the same fraction generally clustered with each other, indicating the high quality of the data (**Figure S2B-D**). To minimize the effect of library preparation on the identification of genes with significantly different $\psi$ values across compartments, we included the library preparation method as a covariate in LABRAT's linear model. This allowed us to pool all of the samples for a given compartment in order to identify genes with significantly different $\psi$ values between compartments regardless of library preparation method.

We first identified genes with significantly different $\psi$ values across any pairwise comparison between cytosolic, membrane-associated, and insoluble fractions (FDR < 0.05). Based on our observations relating distal APA and RNA localization to projections, we then asked if any of these fractions were associated with higher $\psi$ values than the other two. We visualized these comparisons using simplex plots (Fig. 2C). In these plots, each dot represents a gene, and its position is determined by the relative $\psi$ values in each fraction. A gene with a $\psi$ value of 1 in a fraction and $\psi$ values of 0 in the other two would be placed at that fraction's vertex while a gene with equal $\psi$ values in all 3 fractions would be placed equidistant from each vertex at the intersection of the dotted lines. We found that genes tended to have higher $\psi$ values in the membrane fraction (Fig. 2C, **S2E, F**), indicating a preferential association of downstream APA isoforms with that fraction.

Because of the apparent bias toward distal APA site use among membrane-associated transcripts, we next focused on comparing the cytosolic and membrane fractions. When comparing the cytosolic and membrane fractions of HepG2 cells, we identified 552 genes that had significantly different $\psi$ values between the fractions (FDR < 0.01). Of these, 492 (89%) had a $\psi$ value that was

higher in the membrane fraction than the cytosolic fraction, indicating a broad association between transcripts produced using distal APA sites and the membrane fraction (Fig. 2D). We observed highly similar results when comparing the cytosolic and membrane fractions from K562 cells and D17 cells (**Figure S2G, H**).

We then queried whether the same genes had differential APA isoform associations with the cytosolic and membrane fractions in the HepG2 and K562 samples. To test this, we calculated $\Delta\psi$values (membrane - cytosol) for all genes expressed in both cell lines. We observed a strong correlation (R = 0.73) between $\Delta\psi$ values in the two cell lines (Fig. 2E), suggesting that the effects of APA on transcript membrane association are shared between cell lines and are therefore likely transcript-specific with a conserved mechanistic basis.

The ER comprises a large fraction of cellular membranes, and RNA localization to the ER is important for cotranslational access to the secretory pathway. We therefore asked whether transcripts with significant membrane vs. cytosol $\Delta\psi$ values were more or less likely than expected to encode the peptide-based signal sequences required for RNA transport to the ER through cotranslational targeting. We identified ER signal sequences using SignalP [52]. Interestingly, we found that in both the HepG2 and K562 samples, genes that had significant membrane vs. cytosol $\Delta\psi$ values were significantly less likely to contain an ER signal sequence than other genes (Fig. 2F). This observation therefore suggests two alternative modes of RNA localization to the ER: one for transcripts that encode signal peptides and another for those that do not. Specifically, mRNAs that are not cotranslationally targeted by signal peptide recognition appear to be targeted by a mechanism involving distal APA use.

### The transcription speed of RNA polymerase II regulates alternative polyadenylation site choice

The speed of transcription by RNA Polymerase II (Pol II) regulates multiple co-transcriptional processes, including alternative splicing, and termination that is coupled to poly(A) site cleavage [23, 53–56]. To assess how changes in Pol II speed can affect APA, we used LABRAT to analyze RNAseq samples from HEK293 cells that expressed either wildtype or slow Pol II [55]. The slow Pol II mutant used in these studies is a single amino acid substitution in the funnel domain of the Pol II large subunit Rpb1 (R749H).

During transcription, a gene-proximal APA site is necessarily transcribed before a gene-distal APA site. There exists a time, therefore, during which the proximal site is the only APA site that exists on the transcript. Reducing the speed of Pol II transcription would increase this time in which the proximal site is free from competition with the distal site. We hypothesized that this would lead to an increase in usage of the proximal APA site (Fig. 3A). Indeed, we found that for many genes, proximal APA site usage was increased in slow Pol II samples (Fig. 3B), and that overall there was a shift towards increased usage of the proximal site (Fig. 3C).

If the shift in APA was related to the amount of time during which the proximal site was exclusive, then the shift should be most pronounced in genes in which the distance between proximal and distal sites is large. Consistent with this hypothesis, we found that this "inter-polyA distance" for genes that displayed increased proximal APA was significantly longer than expected (Fig. 3D), further suggesting that changes in Pol II kinetics can predictably alter APA.

We found no correlation between a gene's inter-polyA distance and its $\psi$ value from the wildtype Pol II sample alone (R = 0.01). However, in this analysis across genes, other factors that influence APA (e.g. the relative strengths of upstream and downstream polyadenylation sites) can dominate. Comparing within genes but across conditions, as done above using $\Delta\psi$, alleviates this concern.

If alternative polyadenylation of tandem UTRs and ALEs were generally coregulated, then it would be expected that changes in Pol II speed would affect both classes of genes. To test this, we examined the increase in proximal APA site usage caused by slow transcription in the context of tandem UTR and ALE genes separately. We found that proximal APA usage was increased for both tandem UTR and ALE genes (Fig. 3E, F), indicating that the two classes of genes are similarly affected by changes in Pol II speed and consistent with the idea that they are coregulated by a common mechanism.

### Dozens of RNA-binding proteins (RBPs) regulate relative APA isoform abundance across many genes

To investigate the contributions that individual RBPs can have to the regulation of APA isoform abundance, we analyzed the ENCODE RBP knockdown RNAseq datasets with LABRAT [57, 58]. This resource contains 523 shRNA-mediated RBP knockdown RNAseq experiments spread across human HepG2 and K562 cell lines. We compared $\psi$ values for all expressed genes between RBP knockdown and control knockdown samples for 191 RBPs that were expressed in both cell lines. To identify genes that had significantly different $\psi$ values (FDR < 0.05) between RBP knockdown and control knockdown samples, we incorporated the cell line of the experiment as a covariate in LABRAT's linear model.

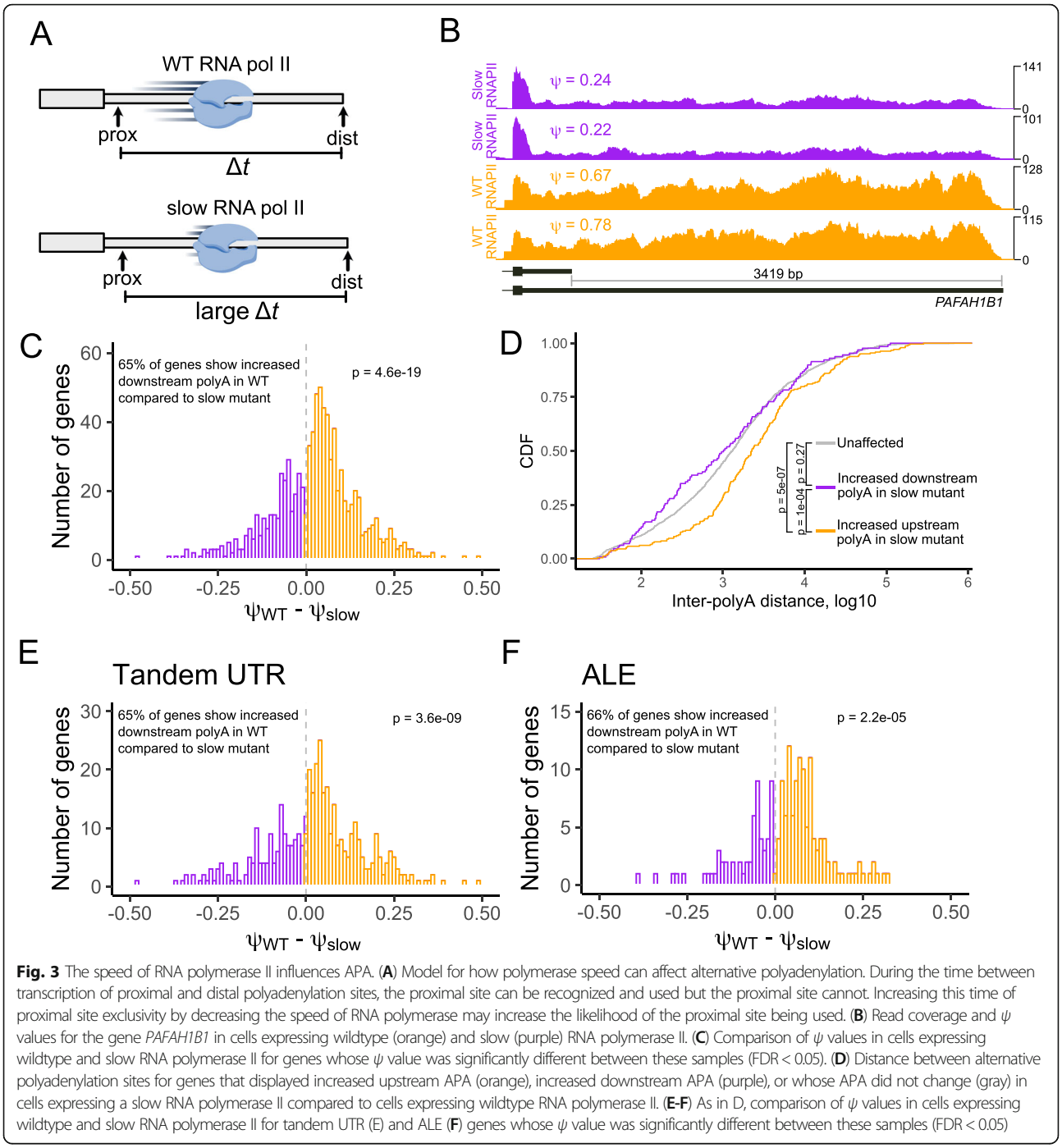


**Fig. 3** The speed of RNA polymerase II influences APA. (**A**) Model for how polymerase speed can affect alternative polyadenylation. During the time between transcription of proximal and distal polyadenylation sites, the proximal site can be recognized and used but the proximal site cannot. Increasing this time of proximal site exclusivity by decreasing the speed of RNA polymerase may increase the likelihood of the proximal site being used. (**B**) Read coverage and $\psi$ values for the gene *PAFAH1B1* in cells expressing wildtype (orange) and slow (purple) RNA polymerase II. (**C**) Comparison of $\psi$ values in cells expressing wildtype and slow RNA polymerase II for genes whose $\psi$ value was significantly different between these samples (FDR < 0.05). (**D**) Distance between alternative polyadenylation sites for genes that displayed increased upstream APA (orange), increased downstream APA (purple), or whose APA did not change (gray) in cells expressing a slow RNA polymerase II compared to cells expressing wildtype RNA polymerase II. (**E-F**) As in D, comparison of $\psi$ values in cells expressing wildtype and slow RNA polymerase II for tandem UTR (E) and ALE (**F**) genes whose $\psi$ value was significantly different between these samples (FDR < 0.05)

We began by assessing the reproducibility of changes in APA isoform abundance upon RBP knockdown between the two cell lines. To do this, we correlated $\Delta\psi$ values (control knockdown - RBP knockdown) for all expressed genes in a given RBP knockdown in HepG2 cells with their $\Delta\psi$ values upon knockdown of the same RBP in K562 cells. We therefore end up with one correlation coefficient per RBP knockdown. As a control, we compared these values to correlations of $\Delta\psi$ values where the RBP that was knocked down was different between the cell lines (Fig. 4A). Reassuringly, we found that correlations between experiments in which the expression of the same RBP was knockdown were significantly higher than those in which the expression different RBPs were knocked down ($p = 1.5e\text{-}19$, Wilcoxon ranksum test). When we restricted the comparison to genes that had significantly different $\psi$ values between RBP and control knockdowns (FDR < 0.05), we

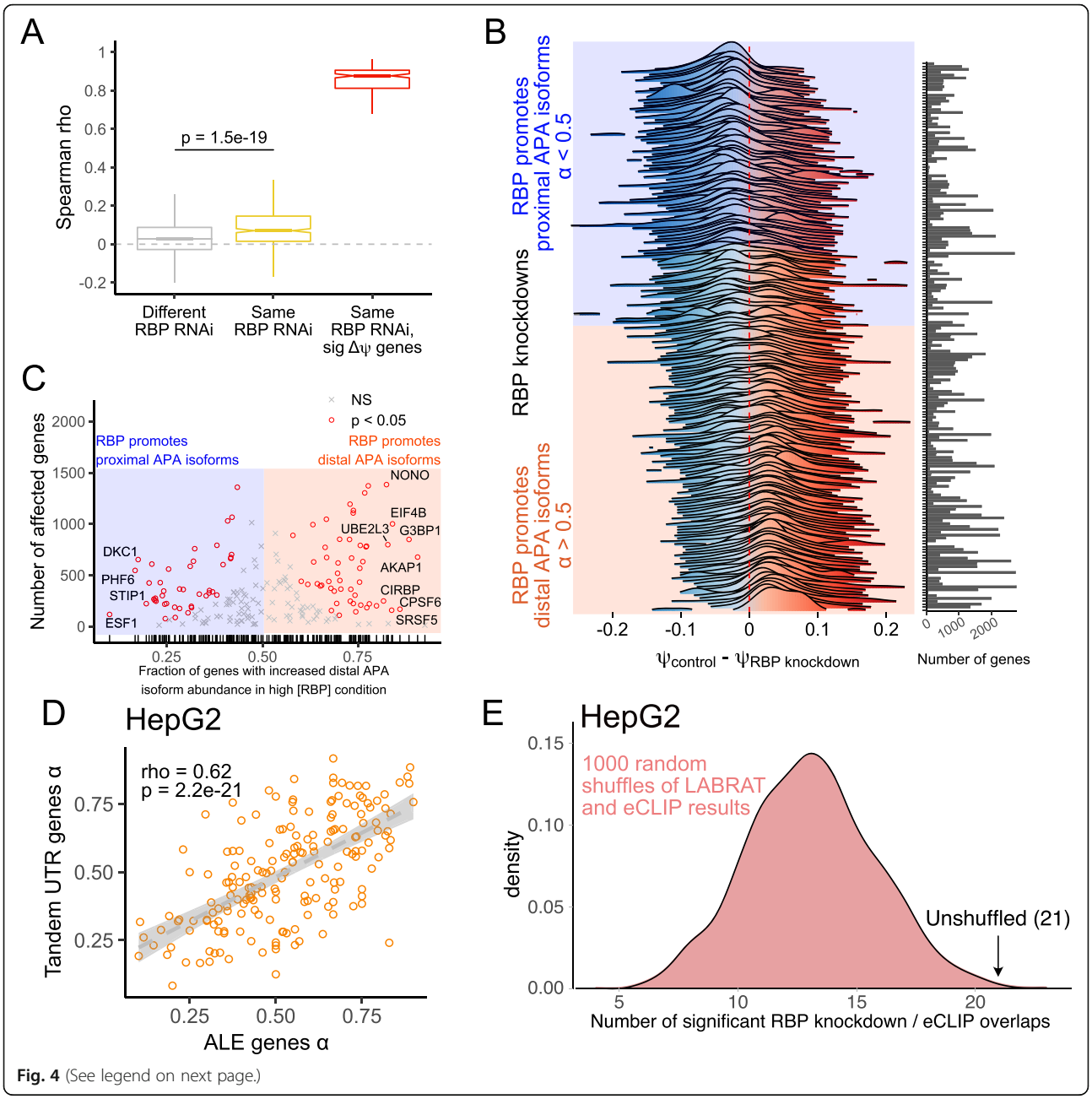


**Fig. 4** (See legend on next page.)

(See figure on previous page.)
**Fig. 4** Many RBPs promote proximal or distal APA isoform abundance in hundreds of genes. (**A**) Correlation of all $\psi$ values across HepG2 and K562 cell lines for all ENCODE RBP-knockdown RNAseq experiments. In gray, correlation coefficients for comparisons of different RBP knockdowns are shown (e.g. RBP X in HepG2 vs. RBP Y in K562). In yellow, correlation coefficients for comparisons of the same RBP knockdown are shown (e.g. RBP X in HepG2 vs RBP X in K562). In red, this comparison is restricted to only those genes whose $\psi$ value significantly differed between the RBP knockdown and control knockdown samples (e.g. RBP X in HepG2 vs RBP X in K562, significant $\Delta\psi$ genes only). In identification of these significant genes, the cell line was included as a covariate. (**B**) Comparison of $\psi$ values in RBP knockdown and control samples for genes whose $\psi$ value was significantly different between these samples (FDR < 0.05). The number of genes with significant $\Delta\psi$ values in each comparison is indicated by the bar graph. A term, $\alpha$, was defined as the fraction of these genes that displayed higher $\psi$ values in the high RBP state (control knockdown) versus the low RBP state (RBP knockdown). (**C**) For each RBP knockdown, the number of genes with significant $\Delta\psi$ values (FDR < 0.05) is indicated on the y axis while the fraction of these genes with positive $\Delta\psi$ values (control knockdown - RBP knockdown) is indicated on the x axis. Knockdowns whose fraction of genes with positive $\Delta\psi$ values significantly differs from the expected 50% are indicated with red circles. (**D**) $\alpha$ values for each RBP knockdown in HepG2 cells were calculated using tandem UTR and ALE genes independently. These were then plotted and correlated. Each dot in this plot represents one RBP knockdown experiment. (**E**) Among 84 RBPs expressed in HepG2 cells, overlaps between the genes whose APA was sensitive to RBP knockdown and the genes whose 3′ UTRs were bound by the RBP in eCLIP experiments were calculated. The significance of this overlap was calculated using a binomial test. 21 RBPs bound the 3′ UTRs of their APA targets more often than expected (binomial $p < 0.05$). To assess whether this was more than the expected number of significant RBPs, relationships between RBPs and their lists of APA and eCLIP targets were shuffled 1000 times, and the analysis was repeated after each shuffle to create a null distribution (pink)

observed a much higher correlation of $\Delta\psi$ values between cell lines (Fig. 4A). These results gave us confidence that we could accurately quantify APA isoform abundance in the ENCODE datasets.

For each RBP knockdown experiment, we then took the genes with significantly different $\psi$ values between RBP and control knockdowns and analyzed the distribution of their $\Delta\psi$ values (control knockdown - RBP knockdown) (Fig. 4B**, Table S1, S2**). We observed that many RBP had distributions of $\Delta\psi$ values that were skewed towards being mostly positive or mostly negative. We defined a term, $\alpha$, as the fraction of these genes with positive $\Delta\psi$ values. RBPs with $\alpha$ values greater than 0.5 therefore were broadly associated with increased distal APA isoform abundance while those with $\alpha$ values less than 0.5 were associated with increased proximal APA isoform abundance. 94 RBPs had $\alpha$ values that were significantly skewed from the expected value of 0.5 (binomial $p < 0.01$), and of these 52 had $\alpha$ values of greater than 0.5 while 42 had $\alpha$ values less than 0.5 (Fig. 4C).

The effects of specific RBPs on APA isoform abundance have been reported for a handful of RBPs. CPSF6 has been found to promote distal APA isoforms [11, 13], and LABRAT analysis of the ENCODE RBP knockdown data agreed with this finding ($\alpha$ = 0.86, $p$ = 5e-20). Fip1 has been found to promote proximal APA isoforms [11]. While Fip1 was not present in the ENCODE data, a related protein, Fip1l1 was present, and LABRAT analysis also found that it promotes proximal APA isoforms ($\alpha$ = 0.39, $p$ = 0.007). Similarly, CSTF2 has been noted to promote proximal APA isoforms [27]. While CSTF2 was not present in the ENCODE data, its related protein CSTF2T was, and LABRAT analysis also found that it promotes proximal isoforms ($\alpha$ = 0.31, $p$ = 0.0001). These LABRAT analyses reflecting prior literature on specific RBPs gave us increased confidence in the ability of LABRAT to probe RBP-specific effects on APA isoform abundance.

For each RBP knockdown experiment we then calculated $\alpha$ values for tandem UTR and ALE genes separately. $\alpha$ values for these two APA types were highly correlated (R = 0.62), further indicating that these two mechanisms of APA regulation are not independent and share elements in common (Fig. 4D**, Figure S3A**).

If changes in APA isoform abundance upon RNAi were directly due to loss of the RBP, then we would expect that the RBP would directly bind the 3′ UTRs of the genes whose APA isoform abundance it regulates. To test this, we analyzed RBP/RNA interactions as measured by the eCLIP experiments performed as part of the ENCODE project [59]. We observed that some RBPs displayed highly promiscuous 3′ UTR binding while others bound very few 3′ UTRs (**Figures S3B, C**).

In HepG2 cells, 84 RBPs had both RNAseq data from RNAi experiments and eCLIP data. For each RBP, we calculated how many of the genes with significant changes in $\psi$ value upon RBP knockdown also contained an eCLIP peak for that RBP in their 3′ UTR. We then calculated whether this overlap of RBP binding and function was statistically significant (binomial $p < 0.05$). For 21 of these RBPs, we observed a significant overlap between the RBPs functional APA isoform regulatory targets and the 3′ UTRs it bound (Fig. 4E). To assess whether this was more or less than the number of expected significant RBPs, we shuffled the relationships between RBPs and their lists of APA targets and bound 3′ UTRs and again calculated the number of RBPs that showed significant overlap between APA and eCLIP data. Repeating this process 1000 times gave us a null distribution of the expected number of RBPs with significant overlaps and indicated that the observed number of overlaps was significant in HepG2 cells ($p$ = 0.006).

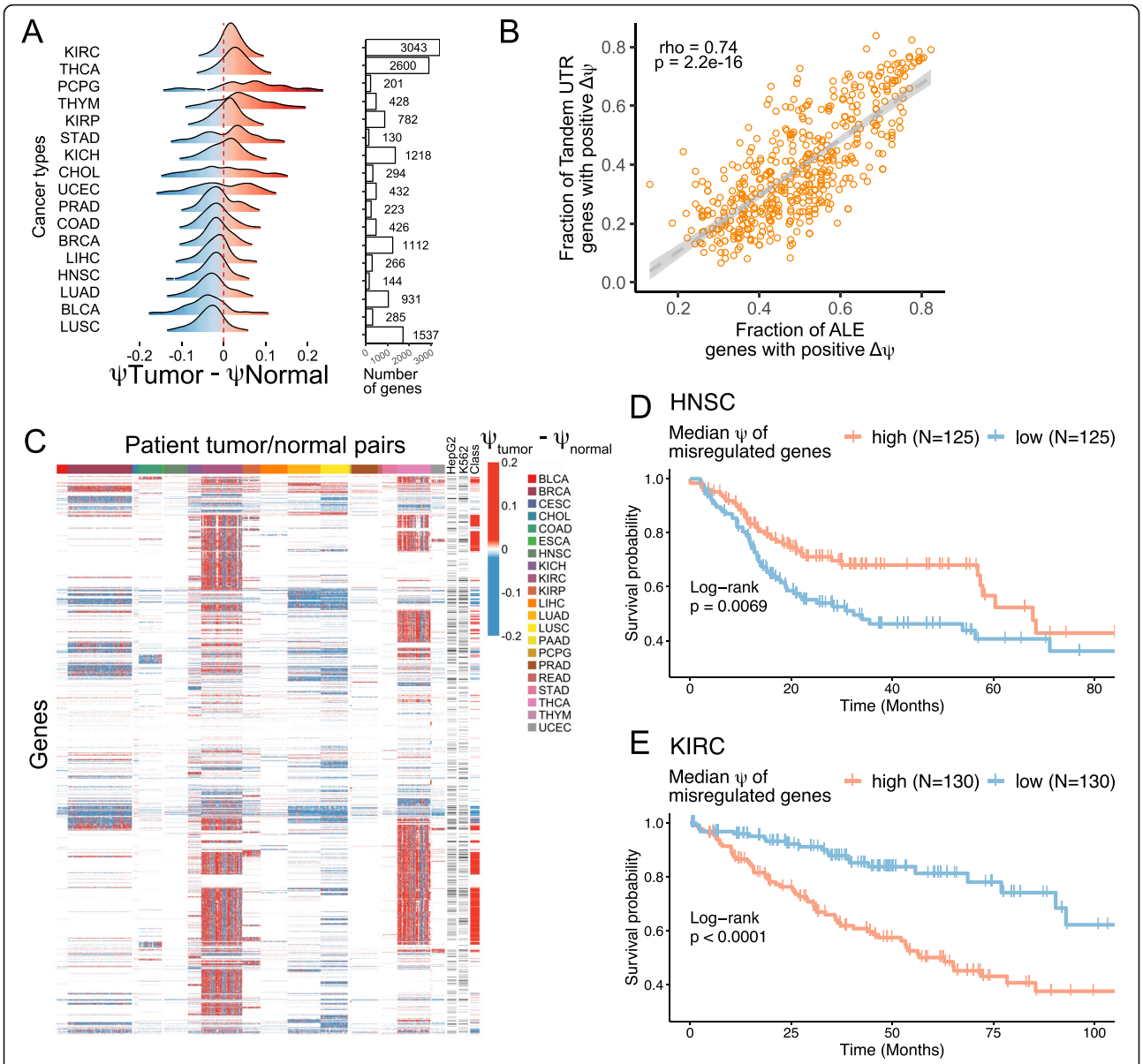


**Fig. 5** Misregulation of alternative polyadenylation in primary tumors. (**A**) Comparison of $\psi$ values in matched patient tumor and control samples for genes whose $\psi$ value was significantly different between these samples (FDR < 0.05). The number of genes with significant $\Delta\psi$ values in each comparison is indicated by the bar graph. (**B**) As in A, $\psi$ values for all genes were calculated in matched patient tumor and normal tissue samples, and genes with significantly different $\psi$ values across samples within a cancer type were identified (FDR < 0.05). The fraction of significant tandem UTR and ALE genes with positive $\Delta\psi$ values were plotted on the x and y axes, respectively. Each dot represents one patient sample pair. (**C**) Genes with significantly different $\psi$ values across samples within a cancer type (FDR < 0.05) are colored according to their $\Delta\psi$ value (tumor - control). Columns represent matched patient samples while rows represent genes. Black ticks (right) represent whether or not the gene displayed a significantly different $\psi$ value (FDR < 0.05) between biochemically defined cytosolic and membrane-associated fractions in HepG2 and K562 cells (Fig. 2D). Genes were further separated into classes of those with increased $\psi$ values in KIRC and THCA tumor samples (red ticks, right) and those with decreased $\psi$ values in BRCA, LUAD and LUSC tumor samples (blue ticks, right). (**D-E**) Survival analysis for APA misregulation in head and neck squamous cell carcinoma and kidney renal clear cell carcinoma, respectively. Patients were grouped into extreme quartiles by ranked median $\psi$ values for misregulated genes as defined in Fig. 5A for the respective tumors. In Fig. 5A, HNSC tumors were associated with decreased $\psi$ values. Here, lower $\psi$ values are associated with poor prognosis. Conversely, in Fig. 5A KIRC tumors were associated with increased $\psi$ values, and here, increased $\psi$ values are associated with poor prognosis

Although we did not observe a similar significant relationship between APA and eCLIP data in K562 cells ($p = 0.4$) (**Figure S3D**), overall, these results indicate that many of the RBPs tested are modulating relative APA isoform abundance through direct interactions. We then repeated the analysis, but looked for eCLIP peak overlaps throughout the gene

bodies of APA targets (**Figure S3E,F**). In this analysis, we did not observe that RBPs were preferentially bound to their APA isoform regulatory targets. Taken together with the results of the 3′ UTR-centric analysis, these results indicate that RBPs that regulate relative APA isoform abundance likely do so through binding the 3′ UTRs of their targets.

### Misregulation of alternative polyadenylation is cancer type specific and correlates with patient survival

Changes APA have long been known to be associated with cancer [60, 61]. Most often, APA is thought to contribute to cancer phenotypes through a general increased usage of proximal APA sites, which are thought to be associated with increased expression of oncogenes and proliferation of cell lines [18, 20]. To further explore this phenomenon, we used LABRAT and data from The Cancer Genome Atlas (TCGA) [62] to examine changes in APA isoform abundance between matched tumor and normal samples from 671 patients across 21 different cancers.

For each cancer, we identified between 130 and 3043 genes that displayed significant differences in $\psi$ values (FDR < 0.05) between tumor and normal samples. We then defined $\Delta\psi$ values (tumor - normal) to ask whether proximal or distal sites showed increased usage in tumor samples. For some cancers, including Lung Squamous Cell Carcinoma (LUSC), Urothelial Bladder Carcinoma (BLCA) and Lung Adenocarcinoma (LUAD), tumors displayed the expected pattern of increased proximal APA in tumors (Fig. 5A). Conversely, Thyroid Cancer (THCA) and Kidney Renal Clear Cell Carcinoma (KIRC) showed strong biases in the opposite direction with increased distal APA in tumors. Mechanisms that drive APA dysregulation are therefore likely specific to different cancer types, and it is not true that increased proximal APA is a general feature of cancer cells.

We then compared $\psi$ values in the TCGA data for tandem UTR genes and ALE genes separately. For each pair of tumor and normal samples, we calculated the fraction of genes with significantly different $\psi$ values across conditions (FDR < 0.05) in which the $\psi$ value was greater in the tumor sample than the normal sample. Put another way, for each patient, we calculated the fraction of significant tandem UTR and ALE genes with positive $\Delta\psi$ (tumor - normal) values (Fig. 5B). The tandem UTR- and ALE-derived fractions were strongly correlated with each other (R = 0.74), again suggesting that these two modes of APA may be coregulated.

We wondered if APA was misregulated in the same genes across many different cancer types or whether the set of genes with misregulated APA was cancer type specific. Although many APA misregulated genes were specific to certain cancers, we did observe that hundreds of genes repeatedly showed misregulation across multiple cancers (Fig. 5C). We defined a set of genes that repeatedly showed increased proximal APA usage in BLCA, LUAD, and LUSC tumors. Using gene ontology analysis, we found that these genes were significantly enriched for those encoding single-stranded RNA binding proteins [63]. Similarly, we defined a set of genes that repeatedly showed increased distal APA usage in THCA and KIRC. These genes were enriched for being involved in programmed cell death and responses to stress.

We enquired whether transcripts we identified whose APA status correlates with membrane association (Fig. 2C, D) are among those subject to misregulation in tumors. Many of these membrane-associated mRNAs showed significantly different $\psi$ values between tumor and normal samples, suggesting that the subcellular localization of these transcripts may be altered in cancerous cells.

To determine if the degree of APA misregulation was related to patient prognosis, we performed survival analyses for patients from the TCGA dataset. In Fig. 5A, we defined genes with tumor-specific APA misregulation by comparing $\psi$ values in tumor and matched normal patient samples. For each tumor, we then calculated a median $\psi$ value across these genes in thousands of tumor RNAseq samples. Using this median $\psi$ of misregulated genes, we ranked patients and separated them into quartiles. The extreme quartiles (patients with the highest and lowest $\psi$ values for misregulated genes) for each cancer were compared. We found that for head and neck squamous cell carcinoma (HNSC), a cancer that typically exhibits increased proximal APA, patients with lower $\psi$ values in misregulated genes had poorer prognoses ($p = 0.0069$) compared to patients with higher $\psi$ values for the same genes (Fig. 5D). Conversely, for kidney renal clear cell carcinoma (KIRC), a cancer that typically exhibits increased distal APA, we found the opposite. Patients with lower $\psi$ values in misregulated genes had better outcomes compared to patients with higher $\psi$ values ($p < 0.0001$) (Fig. 5E). Therefore, the direction of APA misregulation is cancer-specific, and both increased proximal and distal APA are associated with poor patient prognosis, depending on the cancer type.

### Usage of distal APA sites is broadly but weakly associated with decreased RNA expression

Some of the original studies on the relationship between APA and RNA expression reported that distal APA is associated with a decrease in RNA levels [20] while more recent genome-wide studies have reported that the relationship is less clear [64, 65]. To comprehensively examine the relationship between APA and gene expression, we compared changes in $\psi$ and changes in RNA levels across the 191 ENCODE RBP knockdown sample pairs and the 671 TCGA tumor/normal sample pairs. To do

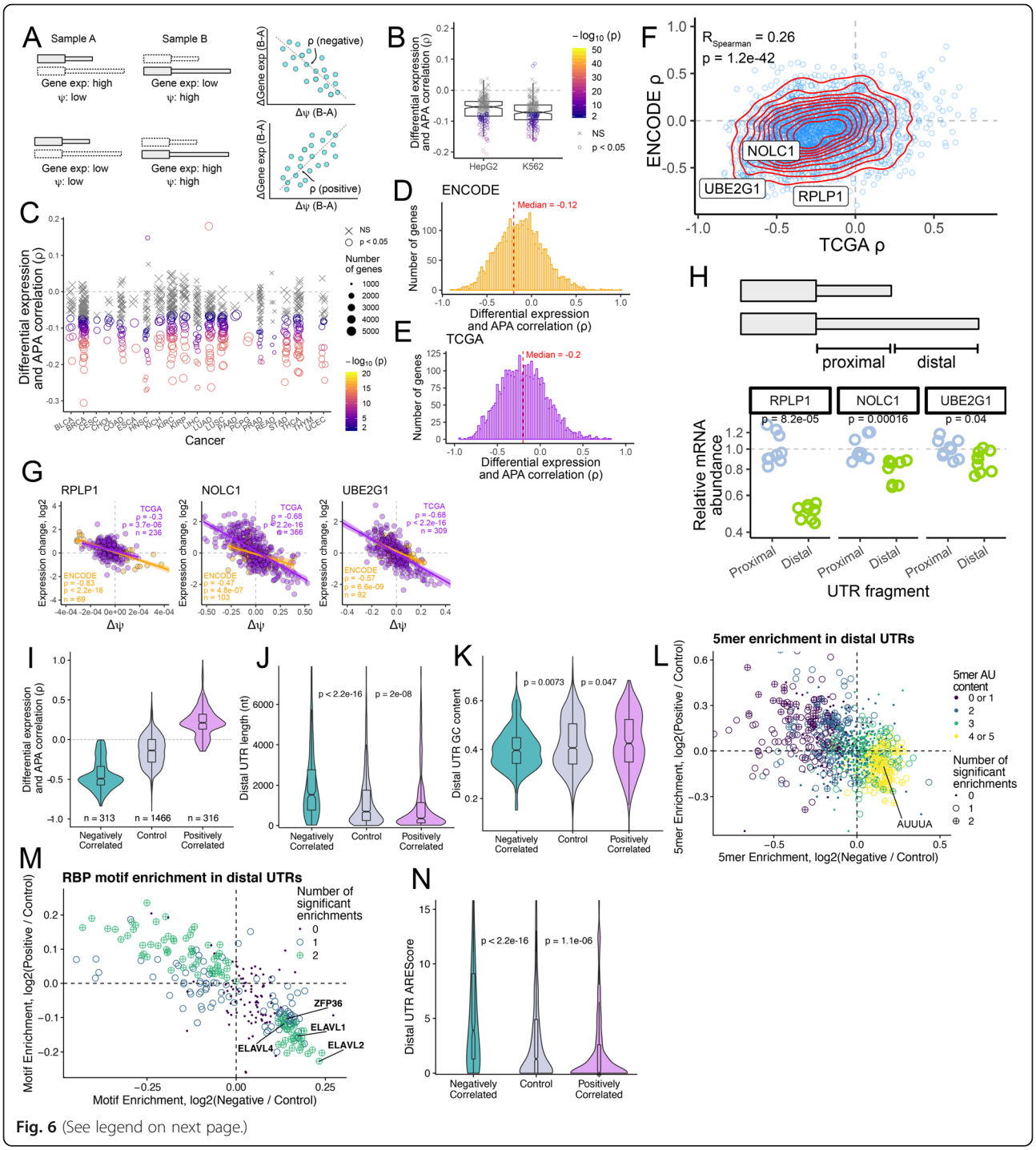


**Fig. 6** (See legend on next page.)

(See figure on previous page.)

**Fig. 6** Comprehensive analyses of connections between alternative polyadenylation and transcript expression. (**A**) Diagram of correlation between APA and transcript expression. Rho ($\rho$) is defined as the correlation between changes in gene expression and changes in $\psi$ value across two conditions. In the scenario described in the top row, the overall RNA expression level for the gene is high in sample A but low in sample B while the gene's $\psi$ value is low in sample A and high in sample B. Changes in gene expression and $\Delta\psi$ are therefore negatively correlated, giving $\rho$ a negative value. Conversely, in the scenario described in the bottom row, changes in gene expression and $\psi$ are positively correlated. (**B**) *P* values across all expressed genes within a comparison for the ENCODE RBP knockdown data. Each dot represents a single comparison (RBP knockdown vs control knockdown). *P* values for the correlation between gene expression and APA are indicated by dot shape and color. (**C**) *P* values across all expressed genes with a comparison for the TCGA paired tumor/control sample data. Each dot represents a single patient's tumor and control samples. P values for the correlation between gene expression and APA are indicated by dot shape and color. (**D**) Gene-level $\rho$ values across all ENCODE RBP knockdown experiments. (**E**) Gene-level $\rho$ values across all TCGA tumor/control sample pairs. (**F**) Correlation of gene-level $\rho$ values derived from the ENCODE and TCGA datasets (**D** and **E**). Red lines indicate the density of points, and the locations of three genes selected for further study are indicated by labels. (**G**) Correlation between gene expression changes and $\Delta\psi$ for three genes. Orange dots represent ENCODE sample pairs (RBP knockdown vs. control knockdown) while purple dots represent TCGA sample pairs (tumor vs. control samples). (**H**) Top: illustration of the UTR fragments fused to the Firefly luciferase gene. Bottom: RT-qPCR-derived relative levels of firefly luciferase mRNA expression when the proximal and distal UTR fragments of the indicated genes were fused. Values indicate ratios between the abundances of Firefly and Renilla luciferase mRNAs with this ratio in the proximal UTR comparison set to 1. P values were calculated using a Wilcoxon ranksum test. (**I**) Correlation between gene expression changes and $\Delta\psi$ was used to define positively correlated, negatively correlated and control genes with two APA isoforms. Correlations are calculated for ENCODE and TCGA separately. (**J**) Distal UTR lengths of each gene set. P values were calculated using a Wilcoxon ranksum test. (K) Distal UTR GC content of each gene set. P values were calculated using a Wilcoxon ranksum test. (L) Five-mer enrichments in the distal 3′ UTRs of positively and negatively correlated gene sets vs control. Five-mers are significantly enriched (BH-adjusted $p < 0.05$, Fisher's exact test) in either both comparisons, one comparison or neither and are represented by a circle plus, open circle or closed dot respectively. Five-mers are colored by their AU content as ranked 0–5. Canonical AU rich element (ARE) "AUUUA" is highlighted as enriched in negatively correlated distal UTRs. (M) RBP motif enrichments in the distal 3′ UTRs of positively and negatively correlated gene sets vs control. RBP motifs are significantly enriched (BH-adjusted $p < 0.05$, Fisher's exact test) in either both comparisons, one comparison or neither and are represented by a green circle plus, blue open circle or purple dot respectively. Canonical ARE binding protein motifs are highlighted as enriched in negatively correlated distal UTRs. (N) Distal UTR ARESCores of each gene set as calculated by ARESscore software. P values were calculated using a Wilcoxon ranksum test

so, we defined a term, rho ($\rho$), as the correlation between changes in $\psi$ and changes in gene expression across two samples (Fig. 6A). Sample comparisons where $\Delta\psi$ and gene expression changes are positively correlated indicate that distal APA and increased RNA levels were associated, and these comparisons will have positive $\rho$ values. Conversely, sample comparisons where $\Delta\psi$ and gene expression changes are negatively correlated indicate that distal APA and decreased RNA levels were associated, and these comparisons will have negative values.

We calculated $\rho$ values across all genes for each RBP knockdown in the ENCODE data. In both the HepG2 and K562 samples, these $\rho$ values overwhelmingly tended to be negative, but weakly so (Fig. 6B). We similarly calculated $\rho$ values across all genes for every patient-derived tumor/normal pair in the TCGA data (Fig. 6C). Again, these $\rho$ values were consistently but weakly negative. These results indicate that although distal APA is generally associated with decreased gene expression, its contribution to changes in RNA levels is modest when comparing all genes in aggregate.

It could be the case, though, that for specific genes, APA and gene expression may be more strongly linked. To explore this, we calculated $\rho$ values for each gene individually across all of the ENCODE and TCGA sample pairs (Fig. 6D, E). The median genes again had weakly negative values (− 0.12 in the ENCODE data, − 0.20 in the TCGA data). ENCODE- and TCGA-derived $\rho$ values for genes were correlated with each other (Fig. 6F, **Table S3**). Tandem UTR genes and ALE genes displayed similar distributions of $\rho$ values, indicating that relationships between gene expression and APA are of approximately equal strength in these two APA classes (**Figure S4A-D**). Relaxing the transcript expression threshold used by LABRAT from 5 TPM to 1 TPM had a minimal effect on these results as the median $\rho$ value remained negative (− 0.10 in the ENCODE data, − 0.14 in the TCGA data).

The tails of the $\rho$ value distributions were long, indicating that there were genes whose changes in $\psi$ value and changes in expression were highly correlated across conditions. We selected three of these, *RPLP1*, *NOLC1*, and *UBE2G1*, for further analysis. Given that each of these genes had strong negative $\rho$ values in both the ENCODE and TCGA data (Fig. 6G), we reasoned that there may be elements in their distal UTRs downstream of the proximal APA site that confer reduced steady-state RNA levels. To test this experimentally, we fused the proximal and distal UTRs of each of these genes to the coding region of Firefly luciferase. Each construct was then site-specifically incorporated into the genome of HeLa cells through Cre-mediated recombination [66]. The Firefly luciferase transcripts were coexpressed from a bidirectional tet-On promoter with unmodified Renilla luciferase. The RNA level of each Firefly-UTR fusion was measured using Taqman qRT-PCR with the Renilla luciferase transcript as a normalizing control. For all three tested genes, fusion of the distal UTR to Firefly luciferase significantly reduced the steady-state level of the

RNA relative to a fusion with the proximal UTR, indicating that sequence elements downstream of the proximal APA sites likely have a role in reducing RNA expression (Fig. 6H). We conclude that by comparing changes in gene expression and APA, we can identify functional elements within 3′ UTRs that regulate mRNA abundance.

### Features enriched in UTRs associated with gene expression changes

To better understand sequence elements downstream of proximal APA sites that may reduce RNA expression, we used the $\rho$ values calculated for individual genes using ENCODE and TCGA sample sets to assign genes to positively correlated, negatively correlated or not correlated (control) gene sets (Fig. 6I, **Figure S4E**). These gene sets behave differently: positively correlated genes are more highly expressed when downstream PAS are used (increased $\psi$) while negatively correlated genes become less expressed as they utilize more downstream PAS.

This analysis was simplified by only considering genes with two APA isoforms such that RNA expression could be explained by proximal or distal UTR usage. The analyzed UTR sequences were unique, meaning that tandem UTRs were separated into proximal and distal UTRs such that distal UTRs lacked their shared 5′ sequence (Fig. 6H). This allowed us to identify sequence characteristics of distal UTRs that explain the differences in RNA expression of the positively correlated and negatively correlated gene sets.

Negatively correlated genes were found to have longer distal UTRs with lower GC content than expected (Fig. 6J, K). This increased 3′ UTR length may make these isoforms more susceptible to NMD, partially explaining their decreased expression [67, 68]. Additionally, negatively correlated genes were generally enriched for AU rich five-mers including the canonical AU rich element (ARE) "AUUUA" (Fig. 6L, **Figure S4F**). Conversely the distal UTRs of positively correlated genes were depleted for AU-rich five-mers (**Figure S4G**). Unsurprisingly given their AU-richness, negatively correlated genes were enriched for ARE binding protein motifs in their distal UTRs and contained more AREs as scored by AREScore [69] (Fig. 6M, N). AREs are destabilizing RNA elements bound by several ARE binding proteins that facilitate RNA degradation. The presence of AREs in distal UTRs of negatively correlated genes is consistent with lower RNA expression when downstream PAS are utilized. It is important to note that the distal UTRs of positively correlated genes are depleted for AREs consistent with their higher expression. These results suggest that APA can regulate gene expression through the inclusion of destabilizing AREs in a transcript's 3′ UTR. Further, given how these results mirror previously observed effects of 3′ UTR AREs [18, 70, 71], they lend further confidence to the ability of LABRAT to accurately quantify relative APA isoform abundance and derive insights regarding its regulation.

### Regulatory effects of RBPs on APA isoform abundance inferred from ENCODE data can be observed in TCGA data

The relation between RBP expression and the widespread misregulation of APA in cancer cells is poorly understood. We investigated this problem by examining expression in patient samples of the 191 RBPs that potentially influence APA isoform abundance revealed by our analysis of ENCODE knockdown RNAseq results (Fig. 4B, C). Based on the ENCODE RBP knockdown data, we defined $\alpha$ values for RBPs where values of greater than 0.5 indicated an RBP that promoted distal APA isoform abundance while values of less than 0.5 indicated an RBP that promoted proximal APA isoform abundance. To compare $\alpha$ values to RBP effects on APA isoform abundance observed in the TCGA data, we defined another term, $\beta$, as the correlation between the change in RNA expression of an RBP between tumor and matched normal TCGA samples and the median $\Delta\psi$ of genes with significantly different APA between the samples (FDR < 0.05) (Fig. 7A). RBPs with positive $\beta$ values are therefore associated with increased distal APA isoform abundance in patient samples while those with negative $\beta$ values are associated with increased proximal APA isoform abundance.

If ENCODE-derived effects of RBPs on APA isoform abundance were recapitulated in the TCGA data, we would expect to see a positive correlation between the $\alpha$ and $\beta$ values for RBPs. We restricted this comparison to the 94 RBPs that had $\alpha$ values significantly different from the expected value of 0.5 ($p < 0.01$, binomial test). For these RBPs, $\alpha$ and $\beta$ values were positively correlated ($R = 0.23$, $p = 0.03$). RBPs with $\alpha$ values greater than 0.5 had significantly higher $\beta$ values than those with $\alpha$ values less than 0.5 (Fig. 7B). Further, when we correlated $\alpha$ and $\beta$ values across all RBPs for all sample pairs within a cancer type, we observed positive correlations in all 12 cancers tested (Fig. 7C). These results further suggest that dozens of RBPs have the ability to regulate relative APA isoform abundance of many genes in a coordinated, directional manner and that the misregulation of APA seen in many cancers may be due to altered expression of specific RBPs.

## Discussion

Alternative polyadenylation is a key step in control of mRNA function, and its misregulation can have large effects on cellular and even organismal phenotype including major effects on the transcriptome of diseased cells including tumors [9, 20, 72–76]. Advances in RNA sequencing and methods of profiling APA from high-throughput data have illuminated the prevalence of APA

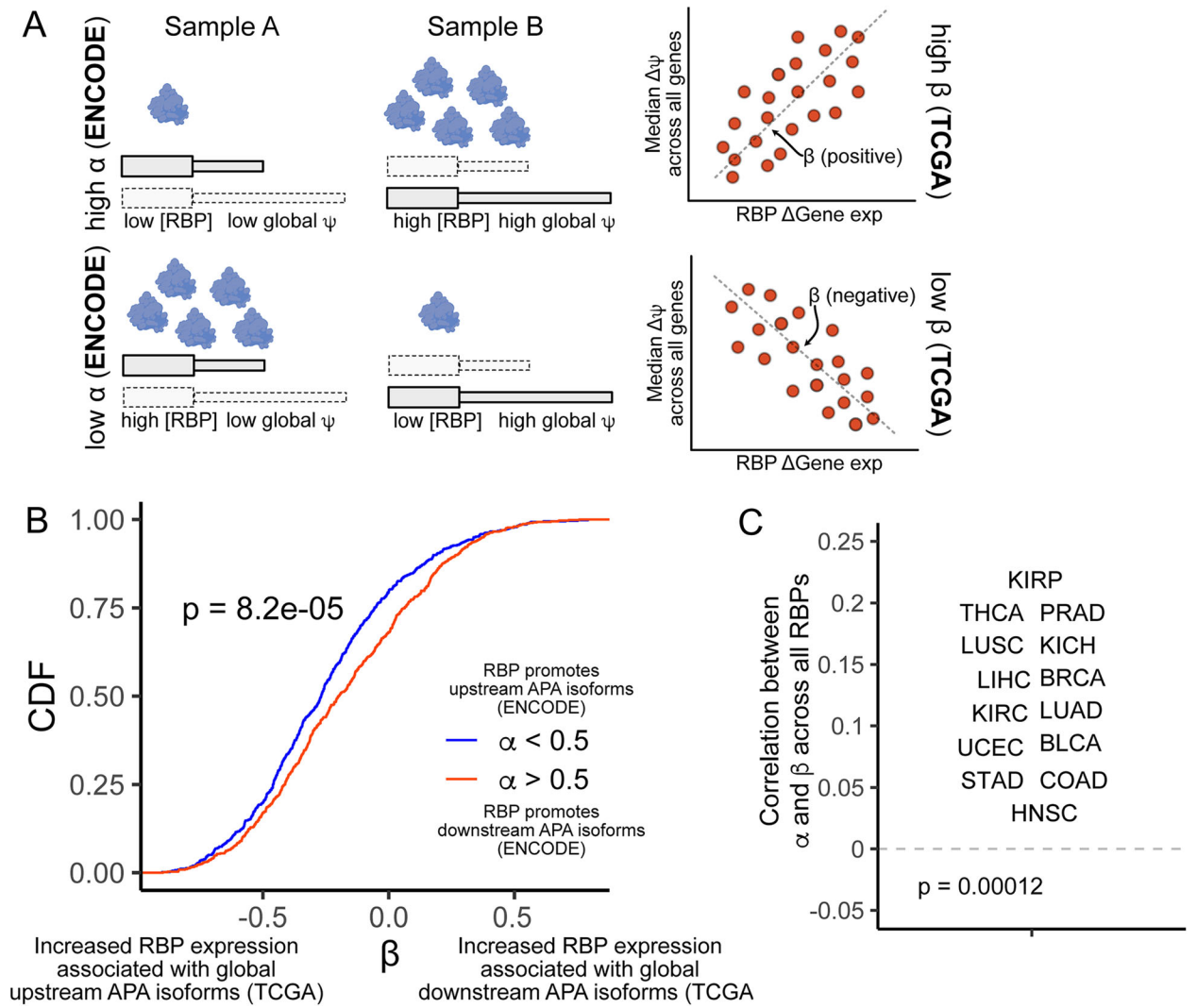


**Fig. 7** APA is regulated by RBP expression in ENCODE and TCGA data. (**A**) Diagram depicting connections between changes in RBP expression between condition and widespread, global in changes in *ψ*. Left: In Fig. 4, RBPs were assigned a value, *α*, based on the effect that their knockdown had on the **Δ***ψ* values for all genes. *α* was defined as the fraction of genes that displayed increased *ψ* values in control knockdown samples compared to RBP knockdown samples. The expression of RBPs with high *α* values was therefore associated with increased *ψ* values transcriptome-wide (top) while expression of RBPs with low *α* values was correlated with decreased *ψ* values transcriptome-wide (bottom). Similar RBP effects were calculated in TCGA data (right) by comparing the change in RBP expression between two matched samples with transcriptome-wide changes in values. A value, *β*, was defined as the correlation between changes in RBP expression and the median **Δ***ψ* across all genes with significant **Δ***ψ* values (FDR < 0.05). *α* and *β* are therefore comparable in relating RBP expression and transcriptome wide changes in with the former designed for ENCODE RBP knockdown data and the latter designed for TCGA paired sample data. (**B**) *β* values for RBPs with low *α* values (*α* < 0.5, blue) and high *α* values (a > 0.5, red). Here, an RBP's *β* value considers the correlation between its expression and global *ψ* across all TCGA sample pairs. The *p* value was calculated using a Wilcoxon ranksum test. (**C**) Correlation between *α* and *β* values across all RBPs for all TCGA sample pairs, separated by cancer type. The p value was calculated using a binomial test for deviation from the expected 0.5 probability that a cancer's correlation between *α* and *β* would be positive

and its regulation across many cell types and physiological conditions [26, 27]. Still, the broad effects of APA on mRNA metabolism, especially beyond changes in mRNA abundance, are not very well understood. Further, the contribution of individual RBPs to the regulation of this process is similarly poorly defined.

To address these challenges, we developed software to accurately quantify alternative polyadenylation and changes in its regulation across conditions from standard RNAseq data. LABRAT builds upon advances in transcriptome quantification using lightweight alignments [30] to determine the relative usage of APA sites within genes. This strategy of using fast, accurate, isoform-level quantification has previously been successfully used to study differential isoform regulation [26, 77]. Here, we have used LABRAT to explore the

regulation and consequences of APA in a variety of contexts using thousands of data sets.

Previous APA quantification software packages, notably DaPars [27], use explicitly mapped read alignments, which may have difficulty in being assigned to a single APA isoform. Further, DaPars does not consider strand information, even when given stranded RNAseq libraries. Neighboring genes on opposite strands can have overlapping 3′ UTRs. Without strand information, reads from this overlapped region can be naively and erroneously assigned.

LABRAT is similar in its approach to a previously developed APA quantification method, QAPA [26]. However, it exceeds the capabilities of QAPA in four areas: (1) LABRAT explicitly classifies genes as having a tandem UTR or ALE structure, allowing comparison of the two groups; (2) LABRAT has a quantification mode dedicated to 3′ end sequencing data; (3) LABRAT employs a statistical test to identify genes whose APA status changes across conditions; and (4) this statistical test can incorporate the use of covariates and complicated experimental designs. Importantly, it must be noted that LABRAT and all currently available APA software quantify the relative abundance of APA isoforms in a sample, not rates of cleavage and polyadenylation.

The subcellular localization of specific transcripts has been known to be regulated by APA. For example, the dendritic localization of *BDNF* mRNA depends on the content of the transcript's 3′UTR as determined by APA [78]. More recent transcriptome-wide studies have shown that this phenomenon is widespread, as hundreds of genes display differential enrichments of APA isoforms across cell body and projection compartments [41–43]. Still, there has been confusion as to the relative contributions of tandem UTR- and ALE-mediated APA to this effect, perhaps due to inefficiencies in studying APA with software that uses genomic alignments. LABRAT is the only currently available APA software that explicitly separates and labels these two classes of genes. We took advantage of this to quantify the distribution of tandem UTR and ALE isoforms across subcellular compartments and found that both classes of APA contribute approximately equally to differences in RNA localization. We further found that differential APA isoform localization is most prevalent in young cellular projections that are less than 3 days old, suggesting that this effect may be important for the initiation of projection outgrowth but less significant for the maintenance of established projections. However, it must be noted that there are differences in cell types among the analyzed samples with young and older projections, potentially confounding the conclusion that projection growth time is related to differential APA isoform localization.

Although RNA localization is most heavily studied in polarized cell types like neurons, transcripts are asymmetrically distributed in essentially all cells. LABRAT identified hundreds of genes with differential APA isoform enrichment between biochemically defined cytosolic and membrane fractions in nonpolarized D17, HepG2, and K562 cells. These results indicate that APA may play a broad role in subcellular localization to membranes in multiple cell types. The consequences of this localization remain unknown, but given that a large fraction of cellular membrane belongs to the ER, modulation of membrane association may be a way to tune the ER association and therefore translation status of a transcript. Further, given the broad misregulation of APA in many cancers, this may mean that the membrane association of many transcripts changes upon transformation. We further found that genes whose APA isoforms are differentially associated with membranes are less likely to encode ER-targeting signal peptides, suggesting that RNA localization to the ER can occur using mechanisms that are independent of the cotranslational targeting. This phenomenon and its misregulation in specific contexts like cancer needs more study.

The abundance of several CPSF and CstF subunits can have important effects on alternative polyA site choice [1, 79–81]. Other RBPs, including CFIm25, have also been shown to strongly directionally regulate APA through activation or repression of specific cleavage events [10, 15]. Using RBP knockdown followed by high-throughput RNA sequencing experiments performed by the ENCODE consortium [57, 58] we interrogated the regulatory effects of 191 RBPs on APA isoform abundance. In this analysis, the knockdown of dozens of RBPs promoted widespread, coordinated directional shifts in relative APA isoform abundance for hundreds to thousands of genes, suggesting that the repertoire of RBPs that can differentially regulate APA isoforms is quite large. It is important to note, though, that many of these RBPs may not be directly regulating APA. For example, many may be differentially regulating stability of 3′ UTR isoforms.

The CPA apparatus processes nascent Pol II transcripts at the ends of genes in the context of complexes with Pol II. According to the "window of opportunity" model [82], the decision between alternative polyA sites can be influenced by the delay between synthesis of upstream and downstream sites which is determined by the speed of transcription. Consistent with this model, we found using LABRAT that slow transcription caused by a mutation in the Pol II large subunit causes a significant shift in favor of upstream polyA sites and that this effect is true for both the ALE and tandem 3′ UTR classes of APA. Moreover, as predicted by the "window of

opportunity" model the mRNAs with the greatest upstream shift in APA correspond to those with the greatest distance between alternative tandem 3'UTR sites (Fig. 3D). In summary, these results show that Pol II speed can significantly modulate alternative polyA site choice. They further suggest the possibility that regulation of transcription elongation could contribute to changes in APA under normal and pathological conditions.

Connections between APA and cancer have been well established [10, 27, 60]. Generally, conclusions regarding this relationship have been focused on the idea of increased proximal APA in cancerous samples [10, 20, 27] with the idea that proximal APA of oncogenic transcripts particular removes repressive regulatory elements in the distal UTR that might otherwise keep the expression of these genes low. However, our results using LABRAT to assess APA changes in 671 paired tumor and normal samples indicate that broad, directional shifts in APA are specific to the type of cancer being studied. Some cancers, including lung cancers and head-neck squamous cell carcinoma, display the canonical increased use of proximal APA sites, while others, including kidney renal clear cell carcinoma and thyroid cancers, show strong shifts in the opposite direction toward distal APA sites. Further, increased proximal and distal APA is associated with poor patient prognosis in head-neck squamous cell carcinoma (HNSC) and kidney renal clear cell carcinoma (KIRC), respectively. Critically, this indicates that increased proximal APA is not a general signature of cancer, but rather that the direction of APA misregulation is cancer-specific.

Relationships between APA and gene expression have also been well documented [18, 20]. Early studies of this connection indicated that distal APA was generally associated with a decrease in gene expression. Later studies, though, indicated that this relationship was less clear [65]. To investigate how APA affects gene expression, we compared changes in $\psi$ values and changes in gene expression for all genes in over 1000 pairs of RNAseq samples. We found that within a sample, correlations between gene expression and APA were weak, but were consistently in the canonical, expected direction where distal APA leads to lower expression. Reorienting the analysis to interrogate the relationship within single genes but across samples again revealed that the average gene has only a very weak connection between APA and gene expression. Still, some genes had remarkable correlations ($R \sim 0.7–0.8$) between these two measurements, indicating that changes in their expression across diverse samples are controlled in large part by modulation of APA site choice.

Across over a thousand pairs of samples, we observed strong correlations between APA changes in genes with tandem UTRs and those with ALEs. If a particular condition promoted increased distal APA in tandem UTR genes, it overwhelmingly also promoted increased distal APA in ALE genes and vice versa. This strongly indicates that the two may be regulated by similar mechanisms, and hints of this connection have been observed before [29]. Tandem UTRs are regulated solely at the level of cleavage/polyadenylation. The simplest interpretation of our results is therefore that the contribution of regulated splicing to ALE control is minor compared to that of regulated cleavage/polyadenylation, perhaps because splicing kinetics are slower. For ALEs, proximal cleavage events obviate potential regulation of the ALE by splicing since the distal ALE is removed from the transcript. If recognition of the proximal APA site by the cleavage and polyadenylation machinery is inhibited, this may provide time for splicing to distal ALEs to occur, and this decision could be affected by the speed of transcription. In this model, splicing acts on ALEs only if given the chance to do so through inhibition of kinetically favored cleavage events.

Overall, the results presented here shed light on the molecular consequences of APA and make predictions about the proteins and mechanisms involved in its regulation. Further experimental studies are needed to fully understand these processes. We envision LABRAT as an important tool in deriving meaningful insights from those experiments.

## Methods

### General LABRAT usage

LABRAT is freely available for download here: https://github.com/TaliaferroLab/LABRAT/. LABRAT searches for specific tags (mRNA_end_NF) in the annotation associated with transcripts with ill-defined 3′ ends. Optionally, LABRAT may also filter out transcripts that are not protein-coding (by looking for the 'protein_coding' tag. This may help remove transcripts that are not fully processed and therefore still nuclear. These tags are present in Gencode (www.gencodegenes.org) gff annotations but may not be present in annotations from other sources. For this reason, we strongly suggest using Gencode annotations for use with LABRAT. For analysis of *Drosophila* data, we modified LABRAT to perform similar filtering on Ensembl annotations for the dm6 *Drosophila* genome build. However, in principle, LABRAT can work with any annotation, including those that make use of 3′ end sequencing data to identify polyadenylation sites, so long as the polyadenylation sites are incorporated into a transcript model. This version of LABRAT is also available at the above GitHub address.

Genes that did not pass an expression filter (TPM ≥ 5) were removed from further analysis. This gene expression was defined as the sum of the expression values for

all valid, filter-passing transcripts for the gene. LABRAT reports these genes as having a $\psi$ value of NA.

Identification of genes with significantly different $\psi$ values across conditions was done using a linear mixed effects model with the Python package statsmodels [83]. For simple comparisons involving two conditions, a simple model relating conditions and $\psi$ values was used ($\psi$ values ~ condition). For analysis of the CeFra and ENCODE data, slightly more complex models were used. In the CeFra data, the method of library preparation, polyA-enrichment or ribosomal RNA depletion, was added as a covariate ($\psi$ values ~ condition + libprep). In the ENCODE data, the cell line, K562 or HepG2, was added as a covariate ($\psi$ values ~ condition + cell line). These models were then compared to null models where the effect of the condition was removed. For simple comparisons, the null models were specified as ($\psi$ values ~ 1). For the CeFra and ENCODE comparisons, these were specified as ($\psi$ values ~ libprep) and ($\psi$ values ~ cell line), respectively. A likelihood ratio test was then used to evaluate the relative fit between the experimental and null models. *P* values were derived from the likelihood ratio test and then corrected for multiple hypothesis testing using a Benjamini-Hochberg correction [32]. $\Delta\psi$ values are defined as differences in mean $\psi$ across conditions.

To define tandem UTR and ALE structures, LABRAT observes the isoform structures at the 3′ end of a gene. If all APA sites are contained within the same exon, then the structure in tandem UTR. If all APA sites are contained within different exons, then the structure is ALE. If a gene has only two APA sites, then its structure must be either tandem UTR or ALE. If a gene has more than two APA sites, it is possible for the gene to fit into neither classification. For example, in a gene with three APA sites, it is possible to have two of them contained within one exon and the third by itself in another exon. In these cases, LABRAT assigns the gene to have a "mixed" structure.

### LABRAT running time

If LABRAT is encountering a gff genome annotation file for the first time, it indexes this file using gffutils (https://github.com/daler/gffutils/). This process can take a few hours, depending on the size of the annotation. However, it only needs to be completed once. All future runs will automatically make use of a database file written after the indexing completes. Importantly, if indexing is interrupted, this file will still be written, and LABRAT will attempt to use this truncated file in the next run. This will cause problems. To prevent this, if indexing is interrupted, be sure to delete the resulting database file. It can be found at the location of the gff annotation, and ends with '.db'.

To test the runtime requirement of LABRAT, we focused on the analysis of RNA polymerase II mutants presented in Fig. 3. This analysis considered two conditions with two replicates per condition. Each sample contained approximately 25 million paired end reads. Using a modern Intel Mac laptop running OSX 10.15 with 12 cores, LABRAT analysis of this data took approximately 25 min. This does not include the time taken to index the genome annotation as described above.

### Comparison of APA in mouse brain and liver tissues

RNAseq data for mouse brain and liver tissues was downloaded from (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA375882) [33]. Each tissue sample contained 8 replicates. Genes with significantly different $\psi$ values were identified as those with an FDR of less than 0.05.

### Analysis of APA in GTEx RNAseq data

RNAseq data from the Genotype-Tissue Expression (GTEx) project (BioProject PRJNA75899) were downloaded from the NCBI Sequence Read Archive (SRA) via dbGaP-authenticated access and quantified using salmon [30] as described elsewhere in this manuscript. $\psi$ values were calculated for each gene in each sample using LABRAT. LABRAT employs an expression level cutoff, returning a $\psi$ value of NA if the sum of expression of all isoforms for a gene is not at least 5 TPM. There were many genes in this analysis of tissue-specific RNAseq that therefore had $\psi$ values of NA in at least one sample. To facilitate PCA analysis, these missing $\psi$ values were imputed using the R package missMDA [84].

The data used for the analyses described in this manuscript were obtained from dbGaP accession number phs000424.vN.pN between 07/16/2020 and 08/31/2020.

### Calculation of $\psi$ values from 3′ end sequencing data

To quantify $\psi$ values from 3′ end sequencing data, we first trimmed 12 nucleotides from the 3′ of the reads as suggested by the authors that produced the data [39]. To calculate $\psi$ values from this data using LABRAT, we added the 'librarytype' parameter. If this value is set to '3pseq', LABRAT will use Salmon-quantified counts for APA abundance estimation instead of length-normalized TPM values.

These 3′ end sequencing libraries were produced using the Quantseq FWD strategy (Lexogen). The single end reads produced by this approach correspond to the end of the fragment opposite of the oligo dT anchor. The beginning of this read is therefore one fragment length away (approximately 200–300 nt) from the polyA tail. Because of this library design, the majority of reads associated with a polyA site in 3′ end sequencing data should lie within 300 nt of the polyA site. Therefore, in

contrast to the LABRAT RNAseq approach where the final two exons of every transcript are used for quantification, the LABRAT 3pseq approach quantifies the last 300 nt of every transcript.

To calculate $\psi$ values from 3′ end sequencing data using counts of *aligned* reads, we first trimmed 12 nucleotides from the 3′ of the reads as suggested by the authors that produced the data [39] and then aligned reads to the human transcriptome (Gencode 28) using STAR [85].

We then wrote a custom Python script (available at https://github.com/TaliaferroLab/LABRAT/blob/master/countfrombam.py) to count the number of aligned reads associated with each polyA site. This was defined as the number of reads in the 300 nt preceding the polyA site. From these counts, $\psi$ values were calculated by combining reads across all of the transcripts with a common polyA site, scaling the counts according to the position of the polyA site within the gene, and computing the ratio of scaled counts to unscaled counts.

### APA analysis of simulated RNAseq data

To compare the performance of LABRAT to QAPA [26], DaPars [27] and Roar [28], we generated a synthetic RNAseq dataset. In this dataset, 5000 genes with only two alternative polyadenylation sites were analyzed. 1250 were randomly assigned to have positive $\Delta\psi$ values, 1250 were assigned to have negative $\Delta\psi$ values, and 2500 were assigned to have no significant change in $\psi$ between conditions. Each gene was then randomly assigned a TPM expression value using a Dirichlet distribution with numpy.random.dirichlet with the alpha parameter set to 1 for every gene.

The simulation was performed by comparing three replicates each from two conditions. For the positive $\Delta\psi$ genes, the minimum $\psi$ from condition B was required to be at least 0.1 greater than the maximum $\psi$ from condition A. Conversely, for the negative $\Delta\psi$ genes, the maximum $\psi$ from condition B was required to be at least 0.1 less than the minimum $\psi$ from condition A. For control genes, the difference between any two $\psi$ values both within and across conditions was required to be less than 0.25. This was performed by randomly sampling $\psi$ values for each gene until the conditions outlined above were met. We found that the varying these $\psi$ value thresholds had minimal effect on the ability of LABRAT to identify differentially regulated genes in the simulated data.

Given a gene's overall expression and its $\psi$ value, TPM values were then relatively split between polyadenylation sites such that the desired $\psi$ value was achieved. TPM values for individual transcripts within polyadenylation sites were then assigned. If a polyadenylation site was only supported by a single transcript, that transcript was given the site's entire TPM value. If a polyadenylation site was supported by multiple transcripts, the site's TPM allotment was randomly distributed among the transcripts.

Given a transcript's assigned TPM value and its length, the desired number of counts for each transcript was then computed by multiplying the TPM value by the length of the transcript. Code for defining expression values, $\psi$ values, and RNAseq count values for the simulation can be found at https://github.com/TaliaferroLab/LABRAT/blob/master/LABRATsimulation.py.

The sequence of each transcript and the desired number of counts were then given to the R package polyester [40] to create the desired number of synthetic, 100 nucleotide, paired-end RNAseq reads.

In analyzing the reads with each package, gene assignments (positive $\Delta\psi$, negative $\Delta\psi$, or control) made by the software were compared to the assignments made during preparation of the synthetic dataset. For analysis of these reads with LABRAT, genes with FDR values of less than 0.05 were called as affected genes (either positive or negative $\Delta\psi$ depending on the reported $\Delta\psi$ value) while those with values of 0.05 or greater were called as control genes. For analysis with QAPA, genes with differences in PPAU values of at least 10 were called as affected genes while those with differences in PPAU values of less than 10 were called as control genes. For analysis with DaPars, genes with adjusted $p$ values of less than 0.05 were called as affected genes while those with adjusted p values of 0.05 or greater were called as control genes. For analysis with Roar, genes with p values less than 0.05 and roar values greater than 1.1 were called as positive $\Delta\psi$ genes, genes with p values less than 0.05 and roar values less than 0.9 were called as negative $\Delta\psi$ genes, while genes with p values of 0.05 or greater were called as control genes.

### Analysis of differential APA isoform enrichment across subcellular compartments

$\psi$ values for each subcellular compartment were quantified using LABRAT, and genes with significant changes in $\psi$ values across compartments were identified using an FDR cutoff of 0.05. The fraction of these significant genes with greater $\psi$ values in the projections than cell bodies was calculated. Binomial $p$ values were calculated for deviations from the expected fraction of 50%. Times of projection growth were manually curated from the methods description of each study.

### Analysis of differential APA isoform enrichment across biochemically defined subcellular fractions

$\psi$ values for each subcellular fraction were quantified using LABRAT, and genes with significant changes in $\psi$ values across compartments were identified using an FDR cutoff of 0.05. FDRs were calculated using a linear

model that incorporated the method of library preparation (polyA-enrichment or ribosomal RNA depletion) as a covariate.

### Quantification of ER signal sequence abundance

For each gene, the translation of its longest CDS sequence was given to the signal sequence prediction program SignalP [52]. For a set of genes, the fraction of genes within the set that contained at least one SignalP-defined ER signal sequence was calculated. For comparing these fractions across sets of genes, a distribution of fractions was created by bootstrapping where 40% of the genes were sampled 100 times.

### Analysis of APA changes induced by changes in RNA polymerase II speed

RNAseq data from HEK293 cells expressing slow (R749H) and wildtype RNA polymerase II [55] were downloaded from the Gene Expression Omnibus (fpol). Using an FDR cutoff of 0.1, genes with significantly different $\psi$ values between wildtype and R749H samples were identified using LABRAT.

### Analysis of ENCODE RBP RNAi knockdown RNAseq samples

In this dataset, each RBP was associated with two RBP RNAi samples and two control RNAi samples. We limited analyses to RBPs that had knockdown samples in both K562 and HepG2 cell lines. $\psi$ values were calculated comparing RBP knockdown and control knockdown samples, and genes with significant $\psi$ differences between RBP RNAi and control RNAi samples were identified using an FDR cutoff of 0.05. FDRs were calculated using a linear model that incorporated the cell line (HepG2 or K562) as a covariate.

For each RBP, the fraction of these significant genes with greater $\psi$ values in the control RNAi than RBP RNAi was calculated. These fractions were defined as a value, $\alpha$, where $\alpha$ ranged from 0 to 1. $\alpha$ values greater than 0.5 were therefore associated with larger $\psi$ values (and therefore more distal APA) in the control RNAi sample. Conversely, $\alpha$ values less than 0.5 were therefore associated with smaller $\psi$ values (and therefore more proximal APA) in the control RNAi sample. Each RBP was therefore assigned one $\alpha$ value from the ENCODE data. Binomial $p$ values were calculated for deviations from the expected fraction of 50%.

### Comparison of ENCODE RBP RNAi knockdowns and eCLIP RBP binding data

The eCLIP narrowpeak bed files for isogenic replicates aligned to GRCh38 for each RBP measured in both HepG2 (103 RBPs) and K562 (120 RBPs) were downloaded from www.encodeproject.org. Analyses were restricted for within each line and not combined. For each individual RBP data set, overlapping peaks were merged using bedtools v2.29.2 [86]. These peaks were then intersected with the longest 3'UTR of genes whose polyA sites were both affected and unaffected by RBP knockdown (as measured by LABRAT described above). RBP occupancy was scored for each 3'UTR as either present or not. The statistical significance of a given RBPs occupancy within the subset of genes whose polyA site choice was affected by knockdown of any RBP was determined using a binomial test.

The number of RBPs that were 'self significant', i.e. the occupancy of a specific RBP was significant for the genes whose polyA site choice was affected by knockdown of that same RBP, was determined for both HepG2 and K562. To determine if that number was greater than what was expected by chance, relationships between RBPs and the genes they bind were shuffled, and the analysis was repeated to identify the number of 'self significant' RBPs. This process was repeated 1000 times to generate a null distribution of the number of 'self significant' RBPs. The number of actual 'self significant' RBPs was then compared to the null distribution and an empirical p value was calculated.

### Analysis of APA in TCGA matched tumor/normal tissue samples

In this dataset, each patient is associated with a pair of samples, one from a tumor and another from matched normal tissue. $\psi$ values were calculated for each sample, and genes with significant $\psi$ differences between all tumor samples and all normal samples within a cancer type were identified using an FDR cutoff of 0.05.

Using the TCGA data, the effect of an RBP's expression on $\psi$ was inferred by correlating changes in the RBP's expression across samples with changes in $\psi$ values of genes that passed the FDR cutoff of 0.05. For each tumor/normal pair, the change in RBP expression was calculated by comparing TPM expression values, and changes in $\psi$ were calculated by finding the median $\Delta\psi$ value across genes with significant changes in $\psi$. The spearman correlation coefficient of this comparison across all tumor/normal pairs was defined as $\beta$. Each RBP was therefore assigned one $\beta$ value from the TCGA data.

### Analysis of survival data in TCGA samples

Using the tumor and matched normal tissue samples from the TCGA dataset, genes with significant $\psi$ differences (FDR < 0.05) were identified for each tumor type as misregulated genes. The median $\psi$ of misregulated genes was then calculated for each patient in samples without matched normal tissue controls. Patients were then ranked by their median $\psi$ of misregulated genes and separated into quartiles. Only

patients within the most extreme quartiles were plotted for each tumor type.

Clinical data for each patient was obtained from cbioportal [87]. Survival analysis and plotting was performed with R packages survival (version = 3.1–8) [88] and survminer (version = 0.4.8) [89]. Log-rank tests for significance were calculated to compare extreme quartiles for each tumor type and were considered significant if less than 0.05.

### Analysis of relationship between APA and RNA expression

For every pair of samples (Control and RBP RNAi in ENCODE and tumor/normal samples in TCGA), the change in RNA expression and $\psi$ value for every gene was calculated. Gene expression filters (TPM ≥ 5) were applied, but FDR cutoffs for $\Delta\psi$ were not. These two values were then compared to each other, and the resulting Spearman correlation coefficient was defined as rho ($\rho$). If distal APA (i.e. increases in $\psi$) was associated with decreases in RNA expression, the resulting $\rho$ value would be negative.

$\rho$ was calculated in two different ways. In the first way, changes in expression and $\psi$ for all genes *within a sample* were correlated. In this comparison, each sample pair ends up with a single $\rho$ value. In the second way, changes in expression and $\psi$ for a single gene *across all sample pairs* were correlated. In this comparison, each gene ends up with a single $\rho$ value in each sample set (ENCODE and TCGA).

The second $\rho$ calculations were used to categorize genes as being either positively or negatively correlated. To achieve similar numbers of genes in each category, a positive $\rho$ in either sample set was considered as positively correlated while a $\rho$ less than − 0.15 in either sample set was considered negatively correlated. Genes behaving inconsistently between sample sets were removed from these categories and placed in the control gene category (25% of positively correlated and 14% or negatively correlated). For simplicity, genes with only two APA isoforms were considered during this categorization resulting in 316 positively correlated genes, 313 negatively correlated genes and 1466 control genes used in UTR sequence analysis.

### Quantifying effects on RNA expression due to UTR content with qRT-PCR

Proximal and distal UTR regions were cloned onto the coding sequence of Firefly luciferase. In this plasmid, Firefly luciferase is driven by a bidirectional tet-On promoter. This promoter also drives Renilla luciferase, which served as a control in these experiments. The resulting plasmids were transfected into HeLa cells using Lipofectamine 2000 (Life Technologies). These cells were engineered to contain a single loxP-flanked cassette within their genome [66]. The plasmid was site-specifically integrated into the genome of the HeLa cells by cotransfecting it with a plasmid expressing Cre recombinase. Recombinants were then selected using 1 μg / mL puromycin for 2 weeks.

The expression of Firefly and Renilla luciferase transcripts was induced by incubating cells with 1 μg / mL doxycycline for 48 h. Total RNA was then isolated using a Quick RNA Isolation Mini Kit (Zymo Research). 1 μg of total RNA was reverse transcribed using iScript Reverse Transcriptase Supermix (BioRad). The relative levels of Firefly and Renilla luciferase transcripts in the sample were then quantified using Taqman qPCR. For each gene, the ratio of Firefly to Renilla luciferase in the case where the proximal UTR was fused to Firefly luciferase was set to 1.

### Identifying features enriched in UTRs associated with gene expression changes

For each gene considered in this analysis (positively correlated, negatively correlated and control genes), proximal and distal UTR sequences were extracted in such a way that they contained unique sequences only. This means that the distal UTRs of genes with tandem UTR models lacked the beginning of their sequence which is unique to the proximal UTR as illustrated in Fig. 6H.

UTR sequence features of either positively or negatively correlated genes were always compared to the control gene set. Enrichment analyses were performed using a custom R package (FeatureReachR) publicly available here: https://github.com/TaliaferroLab/FeatureReachR. This R package utilizes wilcoxon ranksum tests to compare length and GC contents of the three gene sets. Motif and five-mer enrichment significance is calculated with a Fisher's exact test and corrected using the Benjamini & Hochberg method [32]. RBP binding motifs are represented as a sequence match > 80% with position weight matrices sourced from the CISBP-RNA database (http://cisbp-rna.ccbr.utoronto.ca/) [90] or RNA bind-N-seq results [70]. ARESCore [69] was utilized to determine the presence of AU rich elements within the UTRs and compared again using wilcoxon rank-sum tests (http://arescore.dkfz.de/arescore.pl).

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07781-1.

**Additional file 1: Figure S1.** (A) Read coverage plot of *Elavl1* in mouse brain and liver tissues. Dots represent $\psi$ values of 8 replicates. (B) PCA analysis of $\psi$ values calculated from human tissues. Data was produced as part of the GTEx project. (C) As in B, but only using genes that have a tandem UTR APA structure. (D) As in B, but using only genes that have an ALE APA structure. (E) Comparison of $\psi$ values from human brain and liver samples. Delta $\psi$ values for genes with FDR values less than 0.01 are

plotted. (F) Comparison of $\psi$ values from human testis and liver samples. Delta $\psi$ values for genes with FDR values less than 0.01 are plotted. (G) Read coverage plot of *Dab2* in control human PBMCs and those treated with poly dI:dC. RNA from these cells was profiled using 3′ end sequencing. Dots represent $\psi$ values calculated in each of 3 replicates. (H) Comparison of $\psi$ values from 3′ end sequencing data as calculated by LABRAT (orange) and by counting aligned reads (purple, see Methods) (I) Comparison of APA quantifications produced by LABRAT ($\psi$) and QAPA (PPAU). (J) Benchmarking of APA software performance at a range of sequence read depths.

**Additional file 2: Figure S2.** (A) Genes that repeatedly display differential APA isoform localization across repeated neuronal samples. Hierarchical clustering of $\psi$ values from biochemically fractionated *Drosophila* DM-D17-C3 cells (B), HepG2 cells C), and K562 cells (D). (E-F) Simplex plots relating relative $\psi$ values for genes between the cytosolic, membrane-associated, and insoluble fractions of DM-D17-C3 cells (E) and K562 cells (F). A dot that is equidistant from all three vertices had equal $\psi$ values in each fraction while a dot that is closer to one vertex had a higher $\psi$ value in that fraction relative to the other two fractions. (G-H) Comparison of $\psi$ values in K562 (G) and DM-D17-C3 (H) cytosolic and membrane fractions for genes whose $\psi$ value was significantly different between these compartments (FDR < 0.01).

**Additional file 3: Figure S3.** (A) $a$ values for each RBP knockdown in K562 cells were calculated using tandem UTR and ALE genes independently. These were then plotted and correlated. Each dot in this plot represents one RBP knockdown experiment. (B) Binomial $p$ values for overlaps between genes whose APA was sensitive to RBP knockdown and genes whose 3′ UTRs were bound by an RBP in eCLIP experiments. Data taken from ENCODE HepG2 experiments. (C) As in B, but using data from ENCODE K562 experiments. (D) As in Fig. 4E. Among 102 RBPs expressed in K562 cells, overlaps between the genes whose APA was sensitive to RBP knockdown and the genes whose 3′ UTRs were bound by the RBP in eCLIP experiments were calculated. The significance of this overlap was calculated using a binomial test. 14 RBPs bound the 3′ UTRs of their APA targets more often than expected (binomial $p < 0.05$). To assess whether this was more than the expected number of significant RBPs, relationships between RBPs and their lists of APA and eCLIP targets were shuffled 1000 times, and the analysis was repeated after each shuffle to create a null distribution (blue). (E, F) As in Figs. S3D and 4E, but instead of considering eCLIP binding events only in the 3′ UTRs of genes, eCLIP binding events throughout gene bodies were considered.

**Additional file 4: Figure S4.** (A-B) Histogram of gene-wise correlations between changes in $\psi$ and changes in gene expression ($\rho$) derived from TCGA tumor and matched normal samples for tandem UTR (A) genes and ALE (B) genes. (C-D) Histogram of gene-wise correlations between changes in $\psi$ and changes in gene expression ($\rho$) derived from ENCODE RBP knockdown and control samples for tandem UTR (C) genes and ALE (D) genes. (E) Binned scatter plot comparing changes in $\psi$ and changes in gene expression for genes with negative $\rho$ values (blue), positive $\rho$ values (purple) and control genes (gray). (F) Enrichment of 5mers in the distal UTRs of negatively correlated genes compared to the distal UTRs of control genes. (G) Enrichment of 5mers in the distal UTRs of positively correlated genes compared to the distal UTRs of control genes.

**Additional file 5: Table S1.** Delta $\psi$ values (defined as RBP knockdown - control knockdown) for all RBP knockdowns in the ENCODE HepG2 data. Only genes with significant FDR values (less than 0.05) are shown. Delta $\psi$ for genes that did not meet this threshold are indicated as NA.

**Additional file 6: Table S2.** Delta $\psi$ values (defined as RBP knockdown - control knockdown) for all RBP knockdowns in the ENCODE K562 data. Only genes with significant FDR values (less than 0.05) are shown. Delta $\psi$ for genes that did not meet this threshold are indicated as NA.

**Additional file 7: Table S3.** Correlation of expression changes and APA changes in TCGA and ENCODE data. A positive correlation indicates that an increase in gene expression was associated with an increase in $\psi$. Put another way, a positive correlation indicates that increased gene expression was associated with increased distal polyA site usage, while a negative correlation indicates that increased gene expression was associated with increased proximal polyA site usage. Spearman correlation coefficients were calculated across all sample pairs (tumor and matched control in TCGA data, RBP knockdown and control knockdown in ENCODE data) in which the gene was expressed (TPM > 5). If a gene was not expressed in any sample pair, the correlation is noted as NA.

### Acknowledgements
We thank Neel Mukherjee and members of the Taliaferro Lab for helpful discussions regarding analyses. We also thank Rob Patro for helpful discussions regarding using Salmon-based transcript quantification for APA investigation.
The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

### Authors' contributions
RG contributed to manuscript preparation and designed and performed analyses of LABRAT output. KLE contributed to manuscript preparation and designed and performed analyses of LABRAT output. AEG contributed to manuscript preparation and performed LABRAT quantification of TCGA, GTEx, and ENCODE datasets. NF performed experiments related to RNA polymerase II speed mutants. DLB contributed to manuscript preparation and design of analyses. JMT wrote LABRAT software, contributed to manuscript preparation, designed and performed analyses of LABRAT output, and performed experiments related to the expression of reporter constructs. All authors read and approved the final manuscript.

### Funding
This work was funded by the National Institutes of Health (R35-GM133885 to J.M.T. and R35-GM118051 to D.B.) and the RNA Bioscience Initiative at the University of Colorado Anschutz Medical Campus (J.M.T.). It was further supported by a Predoctoral Training Grant in Molecular Biology (NIH-T32-GM008730) (R.G.). Funding sources played no role in the design of the study, interpretation of the data, or writing the manuscript.

### Availability of data and materials
The datasets analyzed during the current study are available in the following repositories:
GTEx tissue expression data: GTEx Portal (https://gtexportal.org/home/).
RNAseq profiling of cells with RNA polymerase II mutants: Gene Expression Omnibus GSE63375 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63375).
RNAseq profiling of cells treated with shRNA against RBPs: ENCODE project (https://www.encodeproject.org/).
RNAseq profiling of matched patient tumor and normal samples: The Cancer Genome Atlas (https://portal.gdc.cancer.gov/).

## Declarations

### Ethics approval and consent to participate
Not applicable.

### Consent for publication
Not applicable.

### Competing interests
None.

### Author details
[1]Department of Biochemistry and Molecular Genetics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [2]RNA Bioscience Initiative, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [3]Division of Hematology, University of Colorado School of Medicine, Aurora, CO, USA.

Received: 14 January 2021 Accepted: 7 June 2021
Published online: 26 June 2021

### References
1. Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates JR 3rd, et al. Molecular architecture of the human pre-mRNA 3′ processing complex. Mol Cell. 2009;33(3):365–76. https://doi.org/10.1016/j.molcel.2008.12.028.

2. Beilharz TH, Preiss T. Widespread use of poly(a) tail length control to accentuate expression of the yeast transcriptome. RNA. 2007;13(7):982–97. https://doi.org/10.1261/rna.569407.
3. Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, et al. A quantitative atlas of polyadenylation in five mammals. Genome Res. 2012; 22(6):1173–83. https://doi.org/10.1101/gr.132563.111.
4. Wu X, Liu M, Downie B, Liang C, Ji G, Li QQ, et al. Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. Proc Natl Acad Sci U S A. 2011;108(30):12533–8. https://doi.org/10.1073/pnas.1019732108.
5. Sherstnev A, Duc C, Cole C, Zacharaki V, Hornyik C, Ozsolak F, et al. Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. Nat Struct Mol Biol. 2012;19(8):845–52. https://doi.org/10.1038/nsmb.2345.
6. Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, et al. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell. 2010;143(6):1018–29. https://doi.org/10.1016/j.cell.2010.11.020.
7. Venkataraman K, Brown KM, Gilmartin GM. Analysis of a noncanonical poly(a) site reveals a tripartite mechanism for vertebrate poly(a) site recognition. Genes Dev. 2005;19(11):1315–27. https://doi.org/10.1101/gad.1298605.
8. Glover-Cutter K, Kim S, Espinosa J, Bentley DL. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. Nat Struct Mol Biol. 2008;15(1):71–8. https://doi.org/10.1038/nsmb1352.
9. Tian B, Manley JL. Alternative polyadenylation of mRNA precursors. Nat Rev Mol Cell Biol. 2017;18(1):18–30. https://doi.org/10.1038/nrm.2016.116.
10. Masamha CP, Xia Z, Yang J, Albrecht TR, Li M, Shyu A-B, et al. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. Nature. 2014;510(7505):412–6. https://doi.org/10.1038/nature13261.
11. Li W, You B, Hoque M, Zheng D, Luo W, Ji Z, et al. Systematic profiling of poly(a)+ transcripts modulated by core 3′ end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. PLoS Genet. 2015;11(4):e1005166. https://doi.org/10.1371/journal.pgen.1005166.
12. Gruber AR, Martin G, Keller W, Zavolan M. Cleavage factor Im is a key regulator of 3′ UTR length. RNA Biol. 2012;9(12):1405–12. https://doi.org/10.4161/rna.22570.
13. Martin G, Gruber AR, Keller W, Zavolan M. Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′ UTR length. Cell Rep. 2012;1(6):753–63. https://doi.org/10.1016/j.celrep.2012.05.003.
14. Takagaki Y, Seipelt RL, Peterson ML, Manley JL. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. Cell. 1996;87(5):941–52. https://doi.org/10.1016/S0092-8674(00)82000-0.
15. Zhu Y, Wang X, Forouzmand E, Jeong J, Qiao F, Sowd GA, et al. Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation. Mol Cell. 2018;69:62–74.e4.
16. Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. Genes Dev. 2013;27(21):2380–96. https://doi.org/10.1101/gad.229328.113.
17. Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC. Widespread and extensive lengthening of 3′ UTRs in the mammalian brain. Genome Res. 2013;23(5):812–25. https://doi.org/10.1101/gr.146886.112.
18. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. Science. 2008;320(5883):1643–7. https://doi.org/10.1126/science.1155390.
19. Ji Z, Lee JY, Pan Z, Jiang B, Tian B. Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc Natl Acad Sci U S A. 2009;106(17): 7028–33. https://doi.org/10.1073/pnas.0900028106.
20. Mayr C, Bartel DP. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell. 2009;138(4): 673–84. https://doi.org/10.1016/j.cell.2009.06.016.
21. Peterson ML, Perry RP. The regulated production of mu m and mu s mRNA is dependent on the relative efficiencies of mu s poly(a) site usage and the c mu 4-to-M1 splice. Mol Cell Biol. 1989;9(2):726–38. https://doi.org/10.1128/mcb.9.2.726-738.1989.
22. Liu X, Freitas J, Zheng D, Oliveira MS, Hoque M, Martins T, et al. Transcription elongation rate has a tissue-specific impact on alternative cleavage and polyadenylation in Drosophila melanogaster. RNA. 2017;23(12): 1807–16. https://doi.org/10.1261/rna.062661.117.
23. de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, et al. A slow RNA polymerase II affects alternative splicing in vivo. Mol Cell. 2003; 12(2):525–32. https://doi.org/10.1016/j.molcel.2003.08.001.
24. Zheng D, Liu X, Tian B. 3′READS+, a sensitive and accurate method for 3′ end sequencing of polyadenylated RNA. RNA. 2016;22(10):1631–9. https://doi.org/10.1261/rna.057075.116.
25. Moll P, Ante M, Seitz A, Reda T. QuantSeq 3′ mRNA sequencing for RNA quantification. Nat Methods. 2014;11 i – iii.
26. Ha KCH, Blencowe BJ, Morris Q. QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. Genome Biol. 2018;19(1):45. https://doi.org/10.1186/s13059-018-1414-4.
27. Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. Nat Commun. 2014;5: ncomms6274.
28. Grassi E, Mariella E, Lembo A, Molineris I, Provero P. Roar: detecting alternative polyadenylation with standard mRNA sequencing libraries. BMC Bioinformatics. 2016;17(1):423. https://doi.org/10.1186/s12859-016-1254-8.
29. Wang R, Tian B. APAlyzer: a bioinformatics package for analysis of alternative polyadenylation isoforms. Bioinformatics. 2020;36(12):3907–9. https://doi.org/10.1093/bioinformatics/btaa266.
30. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017; 14(4):417–9. https://doi.org/10.1038/nmeth.4197.
31. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525–7. https://doi.org/10.1038/nbt.3519.
32. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57:289–300.
33. Li B, Qing T, Zhu J, Wen Z, Yu Y, Fukumura R, et al. A comprehensive mouse transcriptomic BodyMap across 17 tissues by RNA-seq. Sci Rep. 2017;7(1): 4200. https://doi.org/10.1038/s41598-017-04520-z.
34. GTEx Consortium. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45:580–5.
35. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456(7221):470–6. https://doi.org/10.1038/nature07509.
36. Zhang H, Lee JY, Tian B. Biased alternative polyadenylation in human tissues. Genome Biol. 2005;6(12):R100. https://doi.org/10.1186/gb-2005-6-12-r100.
37. Liu D, Brockman JM, Dass B, Hutchins LN, Singh P, McCarrey JR, et al. Systematic variation in mRNA 3′-processing signals during mouse spermatogenesis. Nucleic Acids Res. 2007;35(1):234–46. https://doi.org/10.1093/nar/gkl919.
38. Li W, Park JY, Zheng D, Hoque M, Yehia G, Tian B. Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control. BMC Biol. 2016;14(1):6. https://doi.org/10.1186/s12915-016-0229-6.
39. Corley SM, Troy NM, Bosco A, Wilkins MR. QuantSeq. 3′ sequencing combined with Salmon provides a fast, reliable approach for high throughput RNA expression analysis. Sci rep. Nat Publ Group. 2019;9:1–15.
40. Frazee AC, Jaffe AE, Langmead B, Leek JT. Polyester: simulating RNA-seq datasets with differential transcript expression. Bioinformatics. 2015;31(17): 2778–84. https://doi.org/10.1093/bioinformatics/btv272.
41. Taliaferro JM, Vidaki M, Oliveira R, Olson S, Zhan L, Saxena T, et al. Distal alternative last exons localize mRNAs to neural projections. Mol Cell. 2016; 61(6):821–33. https://doi.org/10.1016/j.molcel.2016.01.020.
42. Tushev G, Glock C, Heumüller M, Biever A, Jovanovic M, Schuman EM. Alternative 3′ UTRs Modify the Localization, Regulatory Potential, Stability, and Plasticity of mRNAs in Neuronal Compartments. Neuron. 2018;98:495–511.e6.
43. Ciolli Mattioli C, Rom A, Franke V, Imami K, Arrey G, Terne M, et al. Alternative 3′ UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. Nucleic Acids Res. 2019;47(5):2560–73. https://doi.org/10.1093/nar/gky1270.
44. Wang T, Hamilla S, Cam M, Aranda-Espinoza H, Mili S. Extracellular matrix stiffness and cell contractility control RNA localization to promote cell migration. Nat Commun. 2017;8(1):896. https://doi.org/10.1038/s41467-017-00884-y.

45. Zappulo A, Van Den Bruck D, Ciolli Mattioli C, Franke V, Imami K, McShane E, et al. RNA localization is a key determinant of neurite-enriched proteome. Nat Commun. 2017;8. Available from:. https://doi.org/10.1038/s41467-017-00690-6.
46. Farris S, Ward JM, Carstens KE, Samadi M, Wang Y, Dudek SM. Hippocampal Subregions Express Distinct Dendritic Transcriptomes that Reveal Differences in Mitochondrial Function in CA2. Cell Rep. 2019;29:522–39.e6.
47. Minis A, Dahary D, Manor O, Leshkowitz D, Pilpel Y, Yaron A. Subcellular transcriptomics-Dissection of the mRNA composition in the axonal compartment of sensory neurons. Dev Neurobiol. 2013; Available from: http://doi.wiley.com/10.1002/dneu.22140.
48. Mardakheh FK, Paul A, Kumper S, Sadok A, Paterson H, McCarthy A, et al. Global analysis of mRNA, translation, and protein localization: local translation is a key regulator of cell protrusions. Dev Cell. 2015;35(3):344–57. https://doi.org/10.1016/j.devcel.2015.10.005.
49. Goering R, Hudish LI, Guzman BB, Raj N, Bassell GJ, Russ HA, et al. FMRP promotes RNA localization to neuronal projections through interactions between its RGG domain and G-quadruplex RNA sequences. bioRxiv. 2019: 784728 [cited 2019 Oct 1]. Available from: https://www.biorxiv.org/content/10.1101/784728v1.
50. Hudish LI, Bubak A, Triolo TM, Niemeyer CS, Sussel L, Nagel M, et al. Modeling Hypoxia-Induced Neuropathies Using a Fast and Scalable Human Motor Neuron Differentiation System. Stem Cell Reports. 2020; Available from. https://doi.org/10.1016/j.stemcr.2020.04.003.
51. Benoit Bouvrette LP, Cody NAL, Bergalet J, Lefebvre FA, Diot C, Wang X, et al. CeFra-seq reveals broad asymmetric mRNA and noncoding RNA distribution profiles in Drosophila and human cells. RNA. 2018;24(1):98–113. https://doi.org/10.1261/rna.063172.117.
52. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019;37(4):420–3. https://doi.org/10.1038/s41587-019-0036-z.
53. Jonkers I, Kwak H, Lis JT. Genome-wide dynamics of pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife. 2014;3:e02407. https://doi.org/10.7554/eLife.02407.
54. Dujardin G, Lafaille C, de la Mata M, Marasco LE, Muñoz MJ, Le Jossic-Corcos C, et al. How slow RNA polymerase II elongation favors alternative exon skipping. Mol Cell. 2014;54(4):683–90. https://doi.org/10.1016/j.molcel.2014.03.044.
55. Fong N, Kim H, Zhou Y, Ji X, Qiu J, Saldi T, et al. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. Genes Dev. 2014;28(23):2663–76. https://doi.org/10.1101/gad.252106.114.
56. Cortazar MA, Sheridan RM, Erickson B, Fong N, Glover-Cutter K, Brannan K, et al. Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a "Sitting Duck Torpedo" Mechanism. Mol Cell. 2019;76:896–908.e4.
57. Consortium, ENCODE Project, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
58. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46(D1):D794–801. https://doi.org/10.1093/nar/gkx1081.
59. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods. 2016;13:508–14. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=27018577&retmode=ref&cmd=prlinks.
60. Masamha CP, Wagner EJ. The contribution of alternative polyadenylation to the cancer phenotype. Carcinogenesis. 2018;39(1):2–10. https://doi.org/10.1093/carcin/bgx096.
61. Yuan F, Hankey W, Wagner EJ, Li W, Wang Q. Alternative polyadenylation of mRNA and its role in cancer. Genes Dis. 2019;8:61–72. Available from: http://www.sciencedirect.com/science/article/pii/S2352304219300984.
62. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, KRM S, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.
63. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10(1):48. https://doi.org/10.1186/1471-2105-10-48.
64. Venkat S, Tisdale AA, Schwarz JR, Alahmari AA, Maurer HC, Olive KP, et al. Alternative polyadenylation drives oncogenic gene expression in pancreatic ductal adenocarcinoma. Genome Res. 2020;30(3):347–60. https://doi.org/10.1101/gr.257550.119.
65. Spies N, Burge CB, Bartel DP. 3′ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. Genome Res. 2013;23(12):2078–90. https://doi.org/10.1101/gr.156919.113.
66. Khandelia P, Yap K, Makeyev EV. Streamlined platform for short hairpin RNA interference and transgenesis in cultured mammalian cells. Proc Natl Acad Sci U S A. 2011;108(31):12799–804. https://doi.org/10.1073/pnas.1103532108.
67. Yepiskoposyan H, Aeschimann F, Nilsson D, Okoniewski M, Mühlemann O. Autoregulation of the nonsense-mediated mRNA decay pathway in human cells. RNA. 2011;17(12):2108–18. https://doi.org/10.1261/rna.030247.111.
68. Hurt JA, Robertson AD, Burge CB. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. Genome Res. 2013;23(10):1636–50. https://doi.org/10.1101/gr.157354.113.
69. Spasic M, Friedel CC, Schott J, Kreth J, Leppek K, Hofmann S, et al. Genome-wide assessment of AU-rich elements by the AREScore algorithm. PLoS Genet. 2012;8(1):e1002433. https://doi.org/10.1371/journal.pgen.1002433.
70. Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, et al. Sequence, structure, and context preferences of human RNA binding proteins. Mol Cell. 2018;70(5):854–67 e9. https://doi.org/10.1016/j.molcel.2018.05.001.
71. Ji Z, Tian B. Reprogramming of 3′ Untranslated Regions of mRNAs by Alternative Polyadenylation in Generation of Pluripotent Stem Cells from Different Cell Types. PLoS One Public Libr Sci. 2009;4:e8419.
72. Shi Y, Manley JL. The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site. Genes Dev. 2015;29(9):889–97. https://doi.org/10.1101/gad.261974.115.
73. Berkovits BD, Mayr C. Alternative 3′ UTRs act as scaffolds to regulate membrane protein localization. Nature. 2015;522(7556):363–7. https://doi.org/10.1038/nature14321.
74. Grassi E, Santoro R, Umbach A, Grosso A, Oliviero S, Neri F, et al. Choice of alternative polyadenylation sites, mediated by the RNA-binding protein Elavl3, Plays a Role in Differentiation of Inhibitory Neuronal Progenitors. Front Cell Neurosci. 2018;12:518.
75. Ulitsky I, Shkumatava A, Jan CH, Subtelny AO, Koppstein D, Bell GW, et al. Extensive alternative polyadenylation during zebrafish development. Genome Res. 2012;22(10):2054–66. https://doi.org/10.1101/gr.139733.112.
76. Zhou X, Zhang Y, Michal JJ, Qu L, Zhang S, Wildung MR, et al. Alternative polyadenylation coordinates embryonic development, sexual dimorphism and longitudinal growth in Xenopus tropicalis. Cell Mol Life Sci. 2019;76(11):2185–98. https://doi.org/10.1007/s00018-019-03036-1.
77. Alamancos GP, Pagès A, Trincado JL, Bellora N, Eyras E. Leveraging transcript quantification for fast computation of alternative splicing profiles. RNA. 2015;21(9):1521–31. https://doi.org/10.1261/rna.051557.115.
78. An JJ, Gharami K, Liao G-Y, Woo NH, Lau AG, Vanevski F, et al. Distinct role of long 3′ UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. Cell. 2008;134(1):175–87. https://doi.org/10.1016/j.cell.2008.05.045.
79. Sun Y, Zhang Y, Hamilton K, Manley JL, Shi Y, Walz T, et al. Molecular basis for the recognition of the human AAUAAA polyadenylation signal. Proc Natl Acad Sci U S A. 2018;115(7):E1419–28. https://doi.org/10.1073/pnas.1718723115.
80. Schönemann L, Kühn U, Martin G, Schäfer P, Gruber AR, Keller W, et al. Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. Genes Dev. 2014;28(21):2381–93. https://doi.org/10.1101/gad.250985.114.
81. Takagaki Y, Manley JL. Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. Mol Cell. 1998;2(6):761–71. https://doi.org/10.1016/S1097-2765(00)80291-9.
82. Bentley DL. Coupling mRNA processing with transcription in time and space. Nat Rev Genet. 2014;15(3):163–75. https://doi.org/10.1038/nrg3662.
83. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. of the 9th Python in Science Conference. 2010 researchgate.net; Available from: https://www.researchgate.net/profile/Josef_Perktold/publication/264891066_Statsmodels_Econometric_and_Statistical_Modeling_with_Python/links/5667ca9308ae34c89a0261a8/Statsmodels-Econometric-and-Statistical-Modeling-with-Python.pdf
84. Josse J, Husson F. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. J Stat Softw [Internet]. Foundation for Open

Access Statistics; 2016 [cited 2020 Sep 25];070. Available from: https://ideas.repec.org/a/jss/jstsof/v070i01.html
85. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.
86. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2. https://doi.org/10.1093/bioinformatics/btq033.
87. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6:l1.
88. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2000.
89. Alboukadel Kassambara MKAPB. survminer: Drawing Survival Curves using "ggplot2". R package version 0.4.8. 2020; Available from: https://CRAN.R-project.org/package=survminer
90. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013;499(7457):172–7. https://doi.org/10.1038/nature12311.

## Publisher's Note