nature genetics

ARTICLES

https://doi.org/10.1038/s41588-021-00864-5

An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability

Lei Li[1,7], Kai-Lieh Huang[2,7], Yipeng Gao[3], Ya Cui[1], Gao Wang[4], Nathan D. Elrod[2], Yumei Li[1], Yiling Elaine Chen[5], Ping Ji[2], Fanglue Peng[6], William K. Russell[2], Eric J. Wagner[2] ✉ and Wei Li[1] ✉

Genome-wide association studies have identified thousands of noncoding variants associated with human traits and diseases. However, the functional interpretation of these variants is a major challenge. Here, we constructed a multi-tissue atlas of human 3′UTR alternative polyadenylation (APA) quantitative trait loci (3′aQTLs), containing approximately 0.4 million common genetic variants associated with the APA of target genes, identified in 46 tissues isolated from 467 individuals (Genotype-Tissue Expression Project). Mechanistically, 3′aQTLs can alter poly(A) motifs, RNA secondary structure and RNA-binding protein-binding sites, leading to thousands of APA changes. Our CRISPR-based experiments indicate that such 3′aQTLs can alter APA regulation. Furthermore, we demonstrate that mapping 3′aQTLs can identify APA regulators, such as La-related protein 4. Finally, 3′aQTLs are colocalized with approximately 16.1% of trait-associated variants and are largely distinct from other QTLs, such as expression QTLs. Together, our findings show that 3′aQTLs contribute substantially to the molecular mechanisms underlying human complex traits and diseases.

Genome-wide association studies (GWAS) have identified thousands of genetic variants that have been associated with quantitative traits and common diseases. However, the vast majority of variants occur in noncoding regions, resulting in significant challenges when attempting to elucidate the molecular mechanisms through which these variants contribute to diseases and phenotypes. To provide functional interpretations of GWAS loci, researchers have suggested employing several molecular QTL analyses, including expression QTLs (eQTLs)[1], which are genetic variants associated with the expression of one or more genes. Although these genetic variants can be informative and, in many cases, are thought to impact the transcription of nearby genes, the roles played by a large fraction of trait-associated noncoding variants is unexplained[2].

APA plays an important role during the posttranscriptional regulation of most human genes. By employing different polyadenylation (poly(A)) sites, genes can either shorten or extend 3′UTRs that contain *cis*-regulatory elements, such as microRNAs (miRNA) or RNA-binding protein (RBP) binding sites[3]. Therefore, APA can affect the stability and translation efficiency of target messenger RNA and the cellular localization of proteins[4]. The diverse landscape of poly(A) sites can substantially impact both normal development and the progression of diseases, such as cancer[5]. The broad importance of alternative polyadenylation is well exemplified by the altered expression of NUDT21, a key APA regulator, in diseases such as glioblastoma[6] and idiopathic pulmonary fibrosis[7]. More recently, our work revealed a more nuanced interpretation of APA since 3′UTR shortening in breast cancer represses tumor suppressor genes in *trans* by disrupting competing endogenous RNA crosstalk[8].

In addition to being associated with gene expression, genetic variations have been identified as critical regulatory factors for the APA of individual genes in certain cell lines[9,10]. Moreover, APA-associated genetic changes have been linked to the development of multiple disease states, including cancer[11], α-thalassemia[12], facioscapulohumeral muscular dystrophy[13], bone fragility[14], neonatal diabetes[15] and systemic lupus erythematosus[16,17]. As a prime example of these studies, one SNP (rs10954213) within the 3′UTR of *IRF5* can alter the 3′UTR length and affect mRNA stability[17], which can further contribute to systemic lupus erythematosus susceptibility. Aside from these few isolated examples, the broad implications of genetic determinants impacting APA in various human tissues and their association with phenotypic traits and diseases have not been systematically examined.

Previous studies identified APA-associated SNPs using 3′-end profiling methods, which have not been widely adopted; thus, these methods have only been applied to small sample sizes[9,18]. In contrast, RNA sequencing (RNA-seq) has been extensively used during eQTL studies; however, only a few RNA-seq data have been analyzed in a manner that would systematically identify and quantify APA events[19]. To obtain an insight into the genetic basis of APA regulation in human tissues, we used our dynamic analyses of APA from RNA-seq (DaPars) algorithm[20] to construct an atlas of tissue-specific, human APA events, using 8,277 RNA-seq datasets coupled with whole-genome sequencing genotype data derived from 46 tissues and isolated from 467 individuals by the Genotype-Tissue Expression Project (GTEx)[1]. In total, we identified 403,215 common *cis*-acting genetic variants associated with APA (3′aVariants), which were colocalized with 16.1% of trait-associated

[1]Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA, USA. [2]Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX, USA. [3]Graduate Program in Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX, USA. [4]The Gertrude H. Sergievsky Center and Department of Neurology, Columbia University, New York, NY, USA. [5]Department of Statistics, University of California, Los Angeles, CA, USA. [6]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX, USA. [7]These authors contributed equally: Lei Li, Kai-Lieh Huang. ✉e-mail: ejwagner@utmb.edu; wei.li@uci.edu

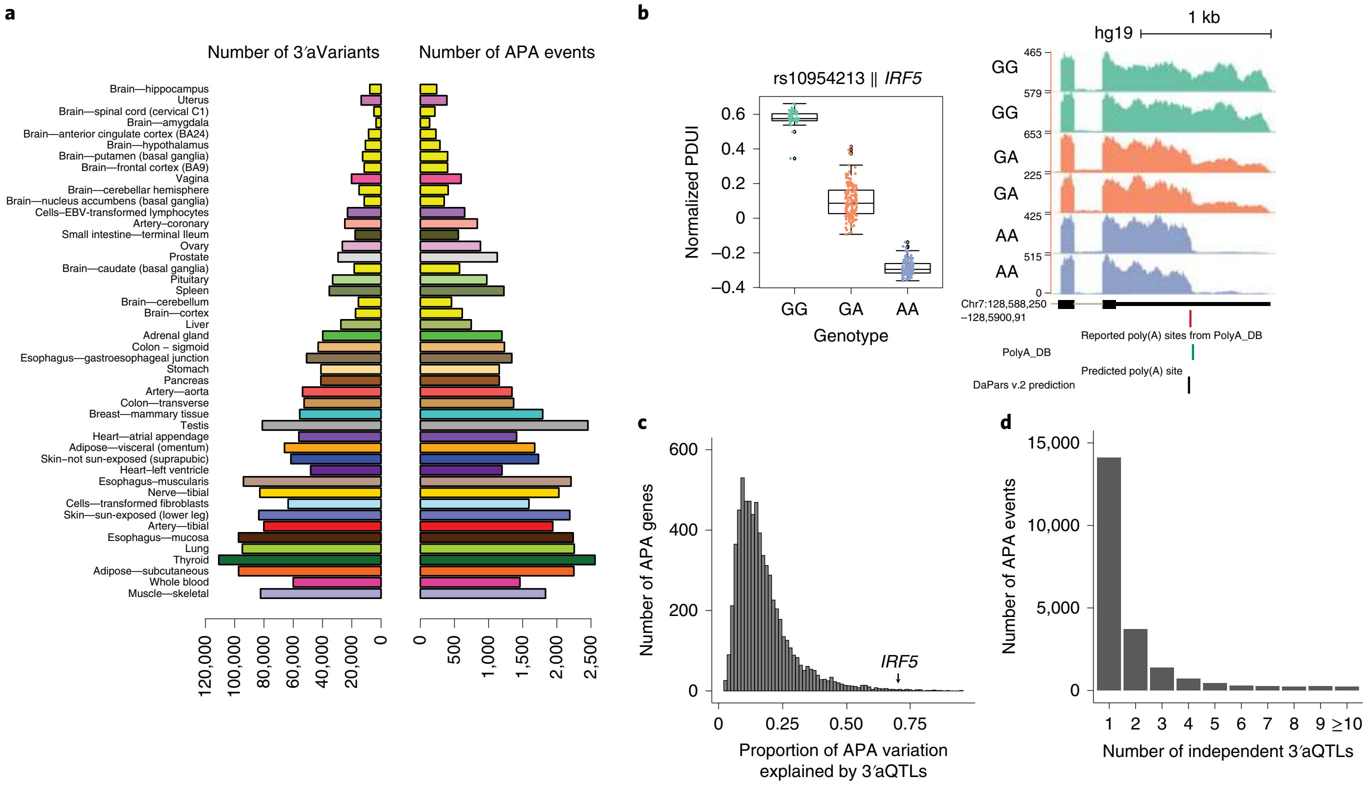


Fig. 1 | Atlas of genetic variations associated with 3′UTR usage across 46 human tissues. a, Distribution of the number of APA events and significant 3′aVariants (FDR ≤ 0.05) for each tissue, sorted by the tissue sample sizes. Each color code indicates a tissue of origin. **b**, Example of a 3′aVariant (rs10954213) that is strongly associated with the *IRF5* 3′UTR usage in whole blood. Left: Distribution of the normalized PDUI for each genotype. Each dot in the box plot represents the normalized PDUI value for one particular sample ($n = 396$). The center horizontal lines within the plot represent the median values and the boxes span from the 25th to the 75th percentile. The whiskers extend to 1.5× interquartile range (IQR) (bottom). Right: RNA-seq coverage track for the *IRF5* 3′UTR. The bottom four tracks show the RefSeq gene structure, 3′aVariant location, reported poly(A) site location and DaPars v.2 prediction. **c**, Average fraction of APA variations that can be explained by 3′aQTLs for each transcript. The *y* axis represents the total transcripts across all human tissues studied. **d**, The distribution of independent 3′aQTLs across tissues.

variants in at least 1 tissue. Collectively, the results of our study indicated that 3′aQTLs reveal the genetic architecture of an emerging molecular phenotype and can be used to interpret a significant portion of the human genetic variants found outside of coding regions.

Results

An atlas of human 3′aQTLs. To detect global APA events in primary human tissues, we used our DaPars v.2.0 algorithm to identify APA events retrospectively and directly using 8,277 standard RNA-seq samples in 46 tissue types from the GTEx v.7 project. The multi-sample DaPars v.2 regression framework calculates a percentage of distal poly(A) site usage index (PDUI) value for each gene in each sample (Supplementary Fig. 1). The PDUI values can then be normalized further after corrections for known covariates including sex, sequencing platform, population structure, RNA integrity number and inferred technical covariates using probabilistic estimation of expression residual (PEER) factors[21]. The inferred PEER factors were strongly associated with several known covariates for each sample and donor (Extended Data Figs. 1–3). We then used Matrix eQTL to identify common genetic variations associated with differential 3′UTR usage (3′aQTLs) in each tissue[22] (Methods). Genes with a 3′aQTL are called 3′aGenes and the corresponding significant variants are called 3′aVariants. Using a false discovery rate (FDR) threshold of 5%, we identified 403,215 3′aVariants associated with 11,613 3′aGenes across 46 tissues, representing approximately 51% of annotated genes (Fig. 1a). Across all tissues, we discovered 56.7% of protein-coding and 26.1% of long noncoding RNA genes detected in at least 1 tissue (Supplementary Fig. 2). The tissues with the highest numbers of 3′aQTLs tended to have larger sample sizes (Supplementary Table 1). This strong association between 3′aQTL number and sample size suggests that additional APA events and 3′aQTLs will continue to be discovered as additional RNA-seq datasets become available. In addition, our global analysis of recent saturation mutagenesis data[23] showed that 3′aQTLs are more enriched in the variants that lead to more notable APA changes (Extended Data Fig. 4).

To evaluate the performance of our 3′aQTL detection method using the current sample size, we compared the detected 3′aQTLs with previously reported SNPs that have been associated with variations in 3′UTR usage. Although previous studies of APA events have been limited to a few cell types, such as lymphoblastoid cells, our approach recaptured many of these 'experimentally validated' 3′aQTLs. For example, the strong association between the SNP rs10954213 and the alternative 3′UTR of *IRF5* (ref. [17]), which encodes a transcription factor involved in multiple immune processes, was replicated in a whole-blood 3′aQTL analysis (Fig. 1b). Interestingly, we also found that this genetic effect on *IRF5* was shared in 22 other tissues, suggesting that the multi-tissue context analysis of this locus could aid further investigations into how *IRF5* variants contribute to autoimmune diseases (Supplementary Fig. 3). Of the 15 previously reported SNP-associated APA genes that were identified in lymphoblastoid cell lines[9,10,24–26], our 3′aQTL analysis was able to recapture 13 (87%) in Epstein–Barr virus (EBV)-transformed lymphocytes (Supplementary Fig. 4). This observation indicated that

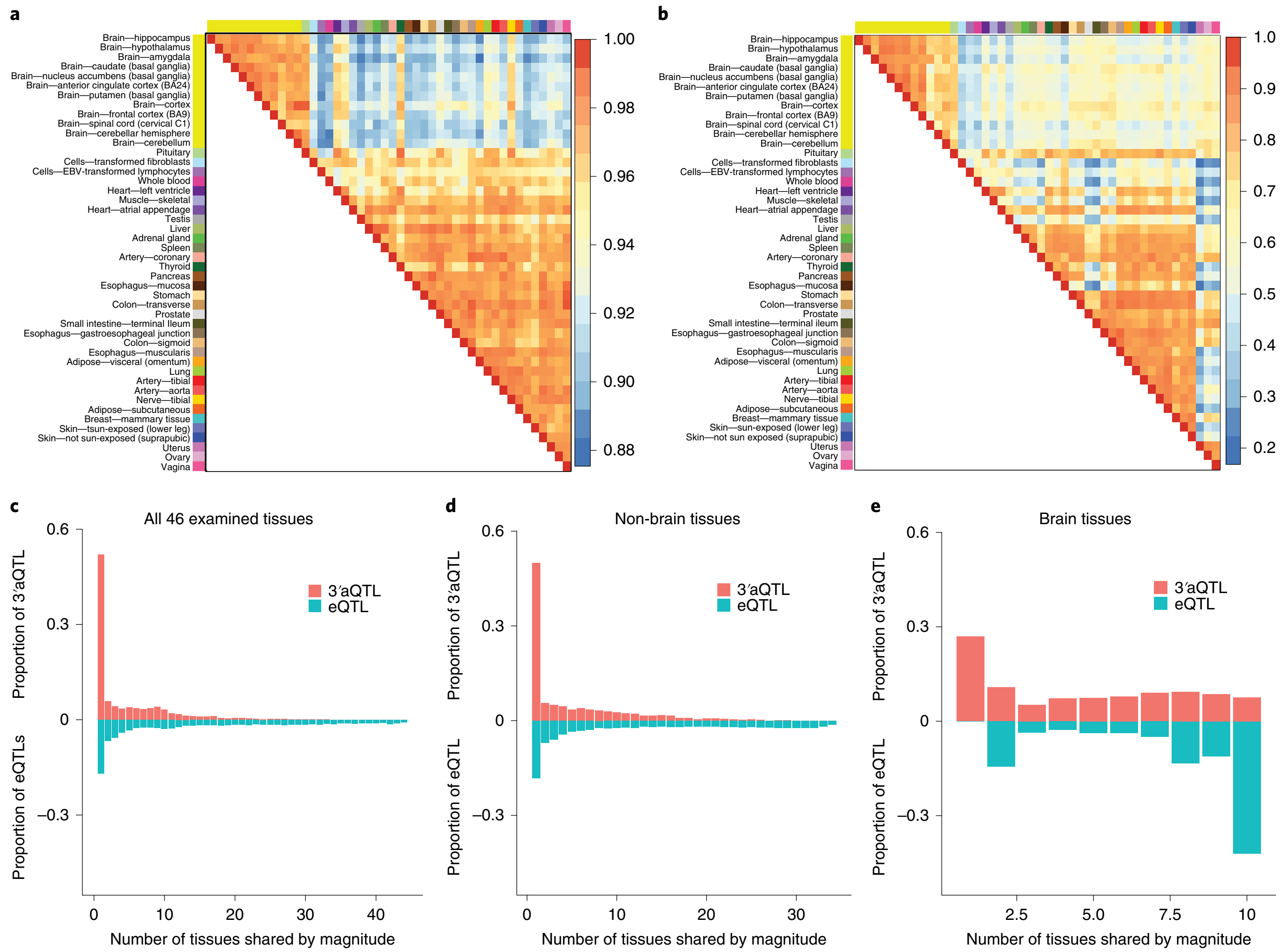


Fig. 2 | Tissue-specific 3′aQTLs. a, Pairwise 3′aQTL sharing by sign among tissues: for each pair of tissues, the proportion of shared lead 3′aQTLs, with the same direction of effect, was calculated. **b**, Pairwise 3′aQTL sharing by magnitude among tissues: for each pair of tissues, the proportion of shared lead 3′aQTLs, with the same direction of effect and within a twofold effect size, was calculated. **c–e**, Histograms showing the estimated proportion of tissues that shared lead 3′aQTLs/eQTLs, by magnitude, with other tissues, among all 46 examined tissues (**c**), among non-brain tissues only (**d**) and among brain tissues only (**e**).

the currently available datasets can be used to capture most of the known APA-associated SNPs in human tissues.

To investigate the global distribution of 3′aQTLs across the human genome, we used Manhattan plots to visualize the locations of 3′aQTLs, with their associated *P* values (Supplementary Fig. 5a). Significant 3′aQTLs were distributed across each chromosome. Importantly, previously reported APA genes were readily detected, including *IRF5* (ref. [17]), *ERAP1* (ref. [10]), *THEM4* (ref. [10]), *EIF2A* and *DIP2B*[9]; however, most of the detected 3′aQTL genes represented, to the best of our knowledge, new events. Several of these new 3′aQTL genes are particularly noteworthy, including *CHURC1* (Supplementary Fig. 5b), which encodes a zinc-finger transcriptional activator that is important during neuronal development[27], and *TPSAB1* (Supplementary Fig. 5c), which encodes α-tryptase and reportedly plays a role in multisystem disorders, such as irritable bowel syndrome, caused by elevated basal serum tryptase levels[28].

We applied heritability estimation and genetic fine-mapping to elucidate the genetic architecture of APA gene variations caused by 3′aQTLs. Specifically, we used a linear mixed model in the genome-wide complex trait analysis genome-based restricted maximum likelihood program[29] to estimate the heritability of the APA variations contributed by all 3′aVariants in each gene, within the 1-megabase (Mb) *cis* region. We observed that 3′aQTLs can explain, on average, 25.2% of APA variations (Fig. 1c). At the individual tissue level, 3′aQTLs can explain between 15.5 and 51.2% of APA variations (Supplementary Table 2). Furthermore, 3′aQTLs could explain >50% of APA variations in 2.2% of APA genes, which are enriched in antigen processing and response to interferon-γ (IFN-γ)-mediated signal pathways (Supplementary Fig. 6). For example, 72.7% of the *IRF5* APA variations can be explained by 3′aQTLs. We also found that 3′aQTLs can explain, on average, 16.2% of APA gene expression changes (Supplementary Fig. 7). To account for correlations among the identified 3′aQTLs, due to linkage disequilibrium (LD), we used sum of single effects (SuSiE) regression[30] to fine-map independent associations (summarized as 95% single-effect credible sets) for each APA transcript in each tissue. SuSiE produces clusters of association signals and each signal is designed to capture exactly one causal SNP independent from those captured by other clusters. SNPs within each signal cluster are highly correlated due to LD. *ALDH16A1* is an APA example where SuSiE revealed two independent 3′aQTL signal clusters (the lead 3′aQTLs are rs1006938 and rs73582462). The maximum R^2 between any 2 SNPs taken separately from the 2 clusters is very small (0.03), suggesting that there are indeed 2 independent signals

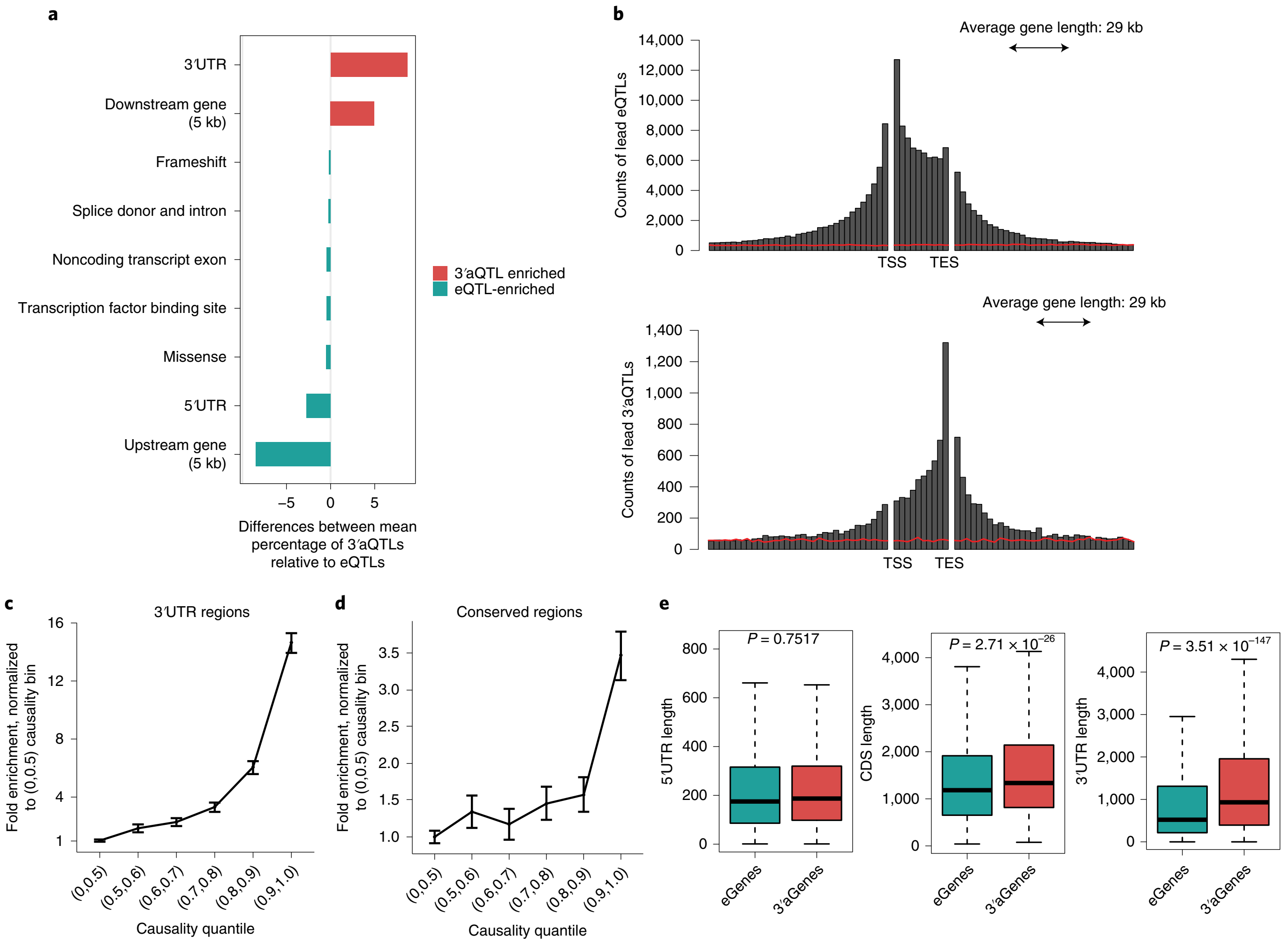


Fig. 3 | 3′aQTL represent a new type of molecular QTL. a, Differences between the mean percentages of lead 3′aQTLs and eQTLs for different annotations. The colors indicate the statistical significance of the differences, with red indicating 3′aQTL-enriched annotations, with an FDR ≤ 0.01, and green indicating eQTL-enriched annotations, with an FDR ≤ 0.01. **b**, Relative distances between eQTLs or 3′aQTLs and their associated genes. TES, transcription end site. The red line represents randomly selected positions within the ±1-Mb window for each gene. **c**, Fold enrichment and 95% confidence intervals (CIs) for 3′aQTLs in each causality bin for the intersection with 3′UTR regions. **d**, Fold enrichment and 95% CIs of 3′aQTLs that intersect with conserved regions, which were defined as regions with UCSC phastCons conservation scores >0.8. **e**, Genomic length was compared between 3′aGenes and eGenes. *P* values were calculated using a two-sided *t*-test. The center horizontal lines of the box plot show the median values and the boxes span from the 25th to the 75th percentile. The whiskers extend to 1.5× IQR (bottom). $n = 46$ tissues examined.

detected. *IRF5* is another APA example where SuSiE detected only one signal cluster (the lead 3′aQTL is rs10954213). In total, 35% of tissue-transcript pairs were associated with more than 1 independent 3′aQTL, which indicated the widespread, allelic heterogeneity of 3′aQTL effects (Fig. 1d). Altogether, the approximately 0.4 million 3′aQTLs we identified provide an extensive display of how common genetic variants are associated with 3′UTR usage across multiple human tissues and expand the number of known 3′aQTLs by several orders of magnitude compared with all previously reported APA-associated SNPs.

Patterns of tissue specificity for 3′aQTLs. To examine how *cis* regulatory elements contribute to APA events in tissue-specific or shared manners (Supplementary Fig. 8), we used multivariate adaptive shrinkage (MASH)[31] to estimate the effect sizes of 3′aQTLs shared across all 46 tissues, while controlling for nongenetic correlations, such as sample overlap. The heterogeneity of cross-tissue effects was evaluated based on the sharing of signs (effects in the same direction) and magnitudes (effects in the same direction and within a twofold effect size change) among 3′aQTLs. This analysis revealed that human tissues cluster into two major groups—brain tissues and non-brain tissues—using hierarchical clustering with complete linkage (Fig. 2a). We also noted that some biologically related tissues grouped within 'non-brain' tissues, such as the uterus/vagina/ovary and colon/stomach groups (Fig. 2b). These patterns revealed developmental and functional similarities between different tissues due to APA regulation. In addition, we found that, although 78.4% of tissues had 3′aQTLs with the same sign, only 13.9% of shared 3′aQTLs displayed similar magnitudes. Compared with eQTLs shared among tissues (85% shared among tissues by sign and 36% shared among tissues by magnitude)[31], 3′aQTLs exhibited similar sign effects (Supplementary Figs. 9 and 10) but a much lower degree of shared-magnitude effects (Fig. 2c–e, Supplementary Fig. 11 and Extended Data Fig. 5). One possible explanation is that APA events are more tissue-specific than gene expression (Supplementary Fig. 12). Considered collectively, these observations suggested that 3′aQTL effect sizes exhibit greater tissue specificity than that of eQTLs.

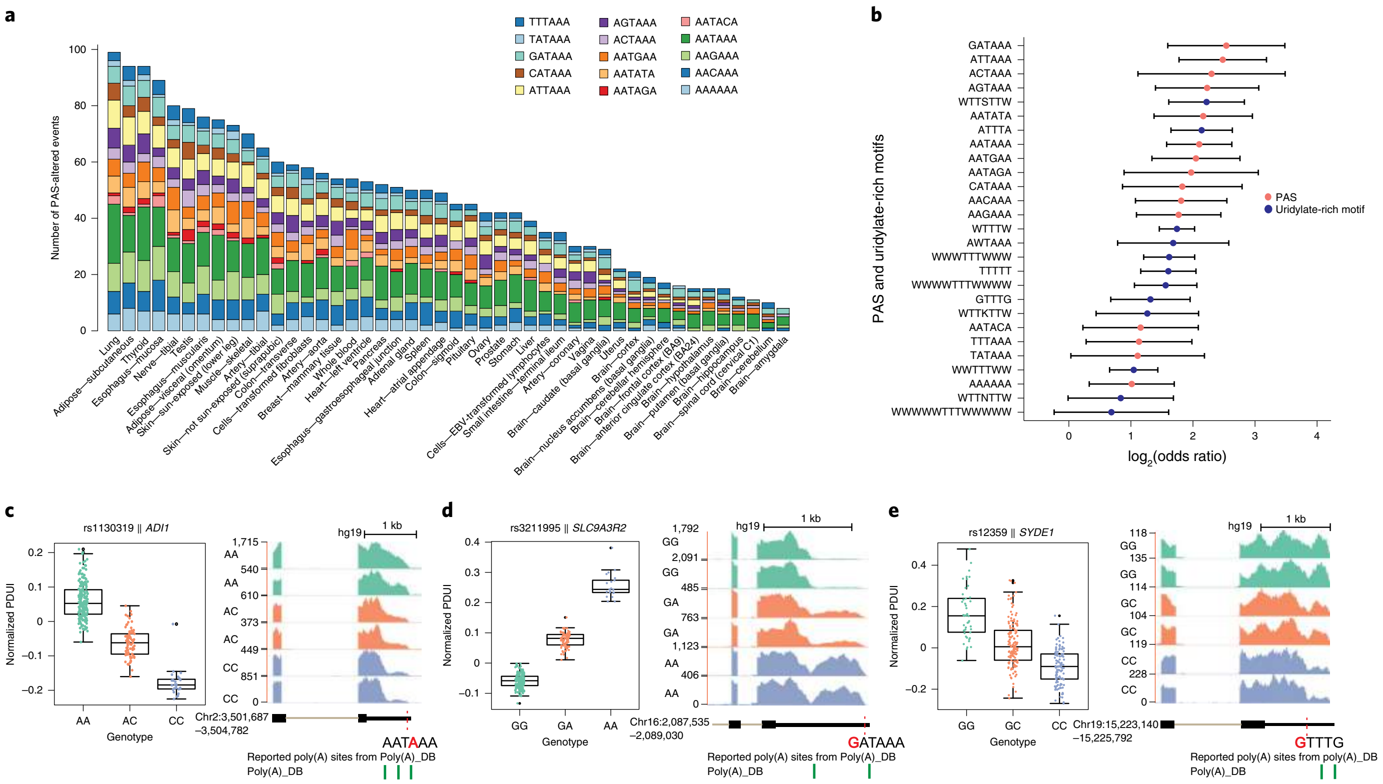


Fig. 4 | 3′aQTLs can alter PAS and uridylate-rich motifs in human tissues. a, Summary of the PAS altered by 3′aVariants across human tissues. The *x* axis shows the tissue names and the *y* axis lists the number of 3′aQTLs that alter the PAS. **b**, Enrichment of 3′aVariants that alter PAS and uridylate-rich motifs and are proximal to poly(A) sites, compared with the rest of the genome. Data are presented as odds ratio and 95% CI. **c**, Box plot showing the significant correlation between the 3′aQTL rs1130319 and *ADI1* APA events for each genotype. Each dot represents a normalized PDUI value from a single sample. The center horizontal lines represent the median values and the boxes span from the 25th to the 75th percentile. The whiskers extend to 1.5× IQR (bottom). The coverage plot illustrates that this SNP could disrupt the canonical PAS. The red dotted line in the RefSeq gene structure indicates the location of the 3′aVariant. The PAS is shown, with the 3′aQTL highlighted in red. **d**, Box plot showing that the 3′aQTL rs3211995 is strongly correlated with the *SLC9A3R2* 3′UTR change for each genotype. The coverage plot illustrates that this SNP could 'create' a canonical PAS. **e**, Box plot showing that the 3′aQTL rs12359, which alters the uridylate-rich motif, is strongly associated with *SYDE1* 3′UTR usage for each genotype.

3′aQTLs have distinct molecular features. To characterize the relationships between different QTLs, we classified lead 3′aQTLs and lead eQTLs across 46 tissue types according to the functional categories defined in SnpEff v.5.0 (ref. [32]). As expected, we found that 3′aQTLs were significantly enriched in 3′UTRs ($P=2.68\times10^{-30}$) or located within 5 kilobases (kb) downstream of genes ($P=9.43\times10^{-08}$), whereas eQTLs were significantly enriched within gene promoters/upstream regions ($P=1.11\times10^{-34}$) or within 5′UTRs ($P=1.42\times10^{-32}$) (Fig. 3a). This observation is consistent with the metagene analysis encompassing the relative position distributions of 3′aQTLs and eQTLs over their associated genes (Fig. 3b). 3′aQTLs are distributed approximately symmetrically around the 3′UTR region and 34% of 3′aQTLs are located in downstream gene regions, likely due to the LD effect[1,33] (Methods). 3′aQTLs also differ markedly from splicing QTLs (sQTLs)[33], which are enriched primarily within gene bodies and splice regions (Extended Data Fig. 6). We also cross-referenced the recent 549 protein QTLs[34] (pQTLs) with lead 3′aQTLs and lead eQTLs. We found that 154 multi-tissue 3′aQTLs are pQTLs for the same gene in 1 or more tissues and 78.5% of pQTL-overlapped 3′aQTLs are not eQTLs. These data suggest that some 3′aQTLs can affect protein expression levels independent of gene expression.

To further determine the genomic context of 3′aQTLs, while also accounting for LD effects, we examined the enrichment of 3′aQTLs according to their posterior causal probabilities. Fine-mapped 3′aQTLs were allocated into six bins based on causality quantiles. We found that 27.4% of 3′aQTLs in the most causal bin (larger than the 90th quantile) were associated with a 14-fold enrichment in 3′UTR regions compared with 3′aQTLs in the least causal bins (less than the 50th quantile) (Fig. 3c). Interestingly, 3′aQTLs are also highly enriched in conserved regions (University of California Santa Cruz (UCSC) phastCons conservation score >0.8) (Fig. 3d) but not in transcription factor binding sites (Supplementary Fig. 13).

Moreover, the structures of 3′aGenes and eQTL-associated genes (eGenes) differed considerably. Compared with eGenes, 3′aGenes harbored comparable 5′UTRs but much longer coding sequences (CDS) ($P=2.71\times10^{-26}$) and 3′UTR lengths ($P=3.51\times10^{-147}$) (Fig. 3e). Furthermore, a significantly higher number of adenylate-uridylate-rich elements proximal to poly(A) sites were observed in 3′aGenes than in eGenes ($P=7.61\times10^{-198}$), suggesting that 3′aGenes harbor more potentially regulatory elements that control APA events (Supplementary Fig. 14). 3′aGenes are also enriched in ontologies related to immune and environmental responses, such as the IFN-γ-mediated signaling pathway. This is in contrast with eGenes, which were underrepresented in genes related to the environmental response[1]. Considered collectively, the results of these analyses suggested that 3′aQTLs and the genes affected by them have different molecular features than other previously defined QTLs and their modulated genes.

Alterations of poly(A) motifs are associated with APA. Next, we investigated the potential mechanisms through which genetic variations contribute to APA events. We hypothesized that some

3′aQTLs alter the motifs important for the 3′-end processing of transcripts. Alterations to the polyadenylation signal (PAS) can produce distinct mRNA isoforms, with 3′UTRs of differing lengths. However, only a few cases have been reported from a limited number of cell lines[9,35]. To systematically examine the prevalence of PAS-altering 3′aQTLs among human populations, we extracted significant 3′aVariants located within 50 base pairs (bp) upstream of annotated poly(A) sites from the Poly(A) database (PolyA_DB)[36], UCSC, Ensembl and RefSeq gene annotations, and performed motif searches based on 15 common PAS motif variants. In total, we identified 2,135 3′aVariants that alter the PAS and generate alternative 3′UTR lengths in their associated genes across 46 human tissues (Fig. 4a and Supplementary Table 3). A total of 991 3′aVariants either disrupted the canonical PAS (AATAAA) or changed other PAS variants to the canonical PAS ($P = 2.827 \times 10^{-10}$) (Fig. 4b). For example, a change in the rs1130319 SNP from the reference A allele to the C allele, which impairs the canonical PAS, AATAAA, correlated with the preferred use of a cryptic poly(A) site in the *ADI1* 3′UTR (Fig. 4c). We validated our finding using recent saturation mutagenesis data[23], where the same 3′aVariant disruption of the *ADI1* canonical poly(A) motif resulted in a 20-fold decrease in the abundance of the long isoform (Extended Data Fig. 7a). In another case, a G>A change in rs3211995 resulted in a strong PAS (AATAAA), instead of the weak noncanonical GATAA motif, at the 3′-end of *SLC9A3R2*, which correlated with a shift to an mRNA isoform with a longer 3′UTR (Fig. 4d). Again, saturation mutagenesis confirmed that this 3′aVariant resulted in a 42.52-fold increase in the abundance of the long isoform (Extended Data Fig. 7b). We also found that 3′aVariants are prone to alter those PAS variants that are proximal to annotated poly(A) sites (Fig. 4b). In addition to the PAS, we also investigated whether 3′aVariants could alter uridylate-rich elements, which are also important for 3′-end processing[4]. Interestingly, adenylate-uridylate, guanylate-uridylate and uridylate-rich motifs were also frequently altered by 3′aQTLs (Fig. 4b and Supplementary Fig. 15). For example, a 3′aVariant at the guanylate-uridylate-rich motif, GTTTG, located near the proximal poly(A) site of the gene *SYDE1*, could lead to significant 3′UTR shortening (Fig. 4e). The uridylate-rich motif variations on APA have been described before[37]. Collectively, these results suggested that a small fraction of detectable APA events are the result of 3′aVariants alterations of PAS or uridylate-rich motifs.

APA-associated RBP binding sites and RNA secondary structure. Alterations in polyadenylation signals can explain only a small percentage of 3′aQTLs, suggesting that most 3′aQTLs affect APA via other mechanisms. To test this hypothesis, we analyzed the extent to which 3′aQTLs interfere with either the transcriptional or posttranscriptional regulation of target genes. First, we used DeepBind v.0.11 (ref. [38]) to evaluate the enrichment of 3′aVariants in 927 binding motifs of 538 DNA-binding proteins and RBPs, in each tissue, using randomly shuffled 3′aVariants as a control group. We identified 125 motifs that were significantly enriched in 3′aVariants, 17 of which were common among at least 20% of the tissues examined (Supplementary Fig. 16). Proteins associated with these 17 common motifs were significantly enriched ($P = 1.06 \times 10^{-5}$; hypergeometric test) with known poly(A) factors, such as PABP[39], CPEB4 (refs. [39,40]), SRSF7 (ref. [41]), RBFOX1 (ref. [42]) and HNRNPC, which was recently described as an APA regulator[43].

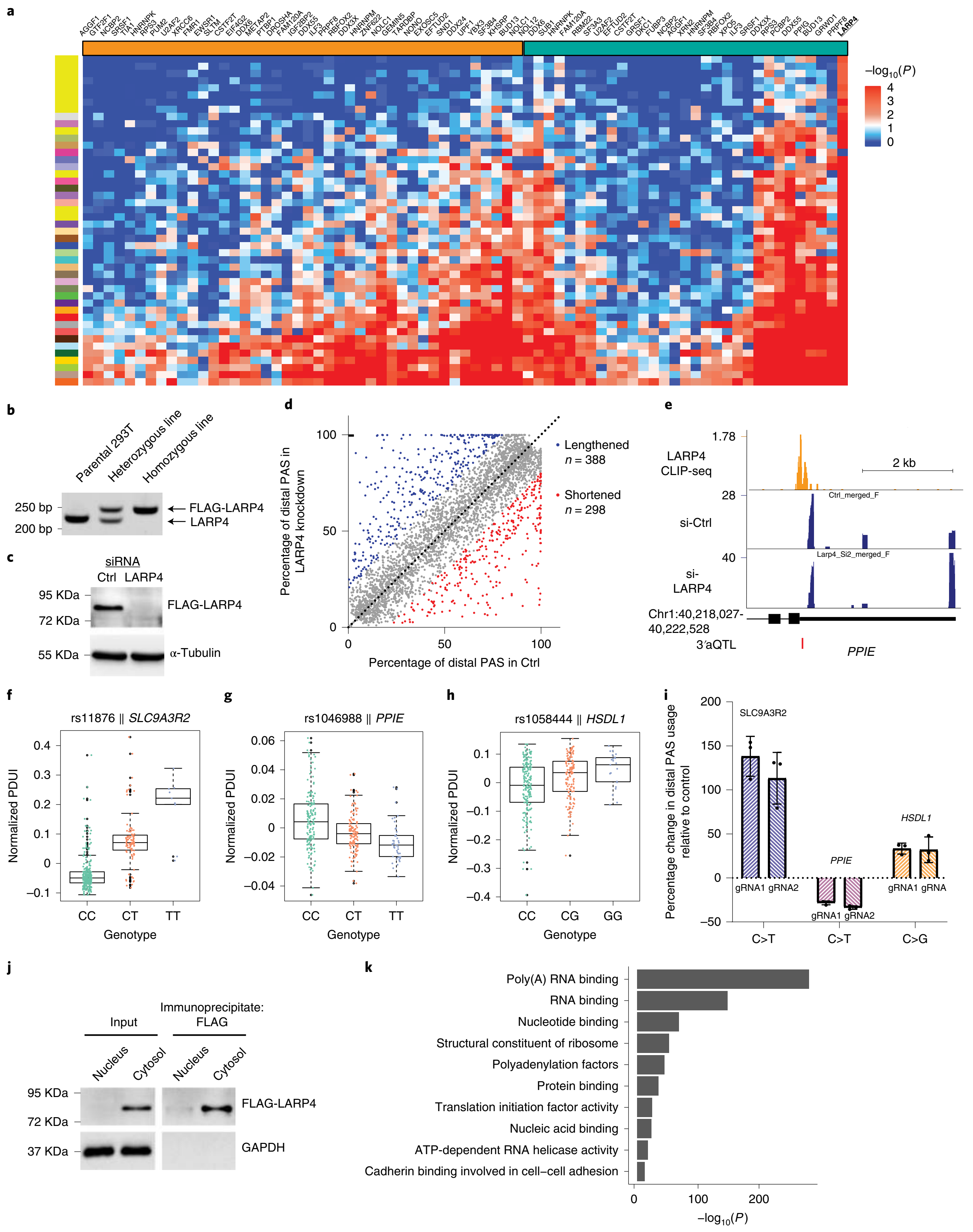
We then analyzed 166 RBP cross-linking immunoprecipitation sequencing (CLIP-seq) datasets, which were available from the Encyclopedia of DNA Elements (ENCODE) project[44]. These datasets are particularly useful because 81.2% of RBPs are not included in the DeepBind resource. We examined whether 3′aQTLs were significantly enriched within the CLIP-seq binding peaks of each RBP compared with a random sequence dataset. We further integrated a new computational strategy to predict the *trans*-regulator of APA (Methods and Extended Data Fig. 8) and identified 73 RBPs that preferentially bound to regions containing 3′aQTLs, including several poly(A) factors, such as CSTF, in addition to many splicing factors (Fig. 5a and Supplementary Table 4). Consistent with a potential functional significance, these splicing factors have previously been linked to alternative 3′UTR usage[40,41].

To evaluate the association between 3′aQTL and RNA structural features, we decided to use the riboSNitch data[45], which are defined as DNA variants affecting RNA secondary structure changes by parallel analysis of RNA structure experiments. We cross-referenced these riboSNitch data with our lead 3′aQTLs. The overlap event was defined as high LD ($R^2 \geq 0.8$) between lead 3′aQTL and riboSNitch for the same transcript. We found that 10.6% of riboSNitch data overlapped with 3′aQTLs (Supplementary Fig. 17), suggesting a strong correlation between 3′aQTLs and RNA secondary structure.

3′aQTL analysis facilitates the identification of APA regulators such as LARP4. Among the 73 3′aQTL-enriched RBPs (Fig. 5a), we found that 1 tumor suppressor, La-related protein 4 (LARP4), with binding sites primarily within 3′UTR regions (Supplementary Fig. 18), was selectively bound to 3′aQTL-containing regions across most tissues. LARP4 is an RBP that binds to the poly(A) tail of mRNA molecules[46] and regulates mRNA translation; however, to our knowledge, its role in APA regulation has not yet been reported. Our observation that LARP4 binding involves regions enriched with 3′aQTLs suggests that LARP4 might be an APA regulator. Importantly, our approach cannot distinguish whether LARP4 APA regulation is mediated through impacting poly(A) site choice in the

Fig. 5 | LARP4 is an APA regulator. a, Heatmap showing the 3′aVariant significance for RBPs identified by ENCODE in each tissue. The left bar shows the color code for each tissue; the top color bar represents the K562 and HepG2 cell lines, separately. Values in the heatmap represent the degree of enrichment for 3′aQTLs in RBP binding peaks compared with the control. **b**, PCR screening gel of clonal 293T lines with homozygous FLAG-LARP4. The primers flanking the integration site of a FLAG epitope tag at the N terminus of the *LARP4* TSS were used. The representing gel of parental 293T cells, a heterozygous targeted line and a homozygous line (*n* = 3 from 12 clonal lines screened) are shown. **c**, Western blot analysis of 293T cells transfected with either control or *LARP4* siRNA to knock down endogenous LARP4 protein. Protein lysates were extracted from whole cells after 72 h of knockdown. The gel represents one of two effective siRNAs tested, as shown in the source data files. **d**, Scatterplot analysis of PAC-seq data comparing distal poly(A) site usage between control and *LARP4* knockdown cells. **e**, Representative genome browser images of the *PPIE* gene, whose poly(A) is regulated by LARP4 and binds with LARP4, as assessed by LARP4 CLIP-seq. **f–h**, Predicted effects of three 3′aQTLs located within the LARP4 binding sites. Each box plot represents the PDUI differences in relation to the SNP genotypes (*n* = 431 for *SLC9A3R2* (**f**), *n* = 396 for *PPIE* (**g**) and *n* = 431 for *HSDL1* (**h**)). The center horizontal lines represent the median values and the boxes span from the 25th to the 75th percentile. The whiskers extend to 1.5× IQR (bottom). **i**, Quantitative RT–qPCR analysis showing the altered APA regulation of three genes in response to CRISPR genome editing to introduce the 3′aQTL that was predicted to alter LARP4 binding. The 3′aQTL of each gene was targeted by two independent gRNAs and each gRNA editing was repeated (*n* = 3, shown by each dot) biologically. Data are presented as the mean ± s.d. **j**, Western blot analysis of nuclear and cytosolic extraction from the homozygous FLAG-LARP4 293T cell line. LARP4 subcellular localization was examined by anti-FLAG M2 antibody. The FLAG immunoprecipitates from each fractionation were subjected to mass spectrometry for orthogonal analysis, which confirmed the results of the western blot through definitive peptide identification. **k**, Functional annotation of the enrichment analysis for LARP4-associated proteins, based on the mass spectrometry results.

nucleus or through regulating differential stability of short/long mRNA isoforms in the cytoplasm. To test the hypothesis that LARP4 regulates APA, we first CRISPR-engineered 293T cells to harbor a single FLAG epitope tag within both copies of the endogenous *LARP4* gene (Fig. 5b). We then transfected these cells with either control small interfering RNA (siRNA) or *LARP4*-targeting siRNA



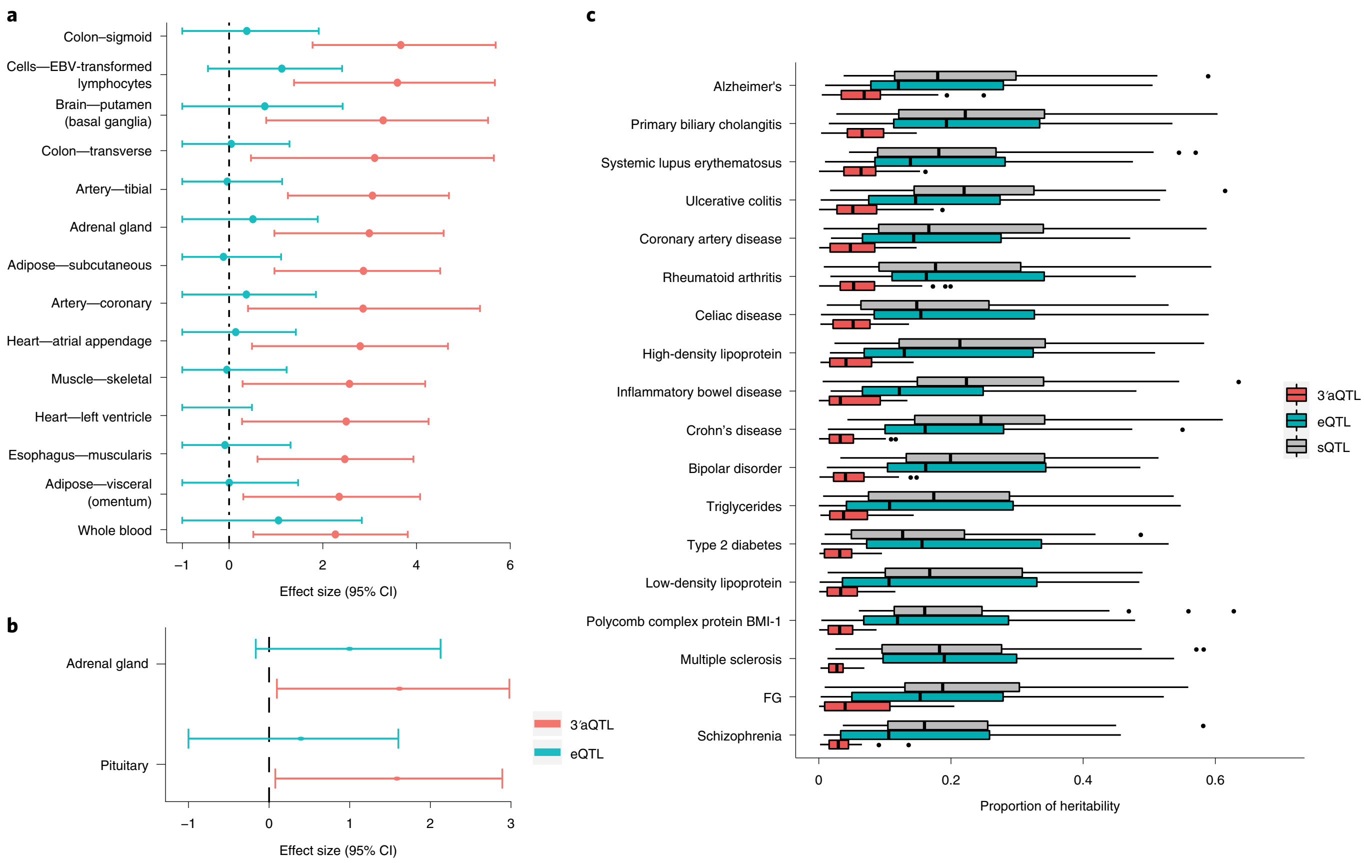


Fig. 6 | Association between 3′aQTLs and human GWAS diseases/traits. a,**b**, Tissues with 3′aVariant enrichment but no eQTL enrichment for Alzheimer's disease (**a**) and rheumatoid arthritis (**b**). The enrichment values (effect size) were calculated using functional genome-wide association analysis, which quantifies the relationships between trait-associated variants and 3′aQTLs/eQTLs. For example, a positive value indicates that variants with stronger association evidence in GWAS are more likely to be 3′aQTLs/eQTLs. The estimated lower and upper bound 95% CIs for the enrichment value are also shown. **c**, Partitioned heritability plot for the percentage of phenotypic variance (*x* axis) that can be explained for 28 traits (*y* axis) by eQTLs, 3′aQTLs and sQTLs in aggregate. FG, fasting glucose. The center lines within the box plot represent the median values and the boxes span from the 25th to the 75th percentile. The whiskers extend to 1.5× IQR (bottom). *n* = 46 tissues examined.

and observed the robust depletion of FLAG-*LARP4* (Fig. 5c). RNA was isolated from both control and knockdown cells and analyzed using 3′-end sequencing (poly(A)-ClickSeq (PAC-seq))[47]. Using PAC-seq, we observed broad changes in poly(A) site usage after knockdown of *LARP4*, which is consistent with a role for *LARP4* in APA regulation (Fig. 5d). Importantly, several of the genes that contain 3′aQTLs that are predicted to alter LARP4 binding were also found to exhibit robust APA in response to *LARP4* knockdown (Fig. 5e and Extended Data Fig. 9). To further test the model that LARP4 can regulate APA, we focused on three genes that exhibit changes in APA after *LARP4* knockdown and contain 3′aQTLs within their LARP4 binding sites, as assessed using the LARP4 CLIP-seq data (Fig. 5f–h). We designed CRISPR-based homologous recombination templates that would allow the introduction of the *LARP4* 3′aQTL into 293T cells (Supplementary Table 5). Cells transfected with Cas9, the homologous recombination template and either of two independent single-guide RNAs (sgRNAs) were selected and APA was assessed using quantitative reverse-transcription PCR (RT–qPCR). In all three cases, we could detect notable changes in the distal poly(A) site selection, which agreed with the predicted effects of 3′aQTLs (Fig. 5f–i), suggesting that the 3′aQTL is sufficient to alter APA regulation. Finally, we generated nuclear and cytoplasmic extracts from FLAG-LARP4 cells, purified LARP4 (using FLAG affinity resin) and analyzed the purified complexes using mass spectrometry (Fig. 5j and Supplementary Table 6). Consistent with previous reports, LARP4 was primarily, but not exclusively, cytoplasmic, and we could robustly detect associated proteins involved in poly(A)-binding. Surprisingly, we also detected numerous components of the cleavage and polyadenylation machinery associated with LARP4, suggesting a potential direct role in APA regulation (Fig. 5k). Altogether, these results support a function of LARP4 in APA regulation and further validate the use of 3′aQTLs as a discovery tool for APA regulators.

3′aQTLs can explain a significant proportion of disease heritability. The GWAS approach has commonly been used to associate genetic variants with complex human traits and diseases. However, explaining how these genetic variations, particularly noncoding variations, contribute to specific phenotypes can be difficult. We hypothesized that 3′aQTLs could be used to interpret GWAS noncoding variants, particularly those located near 3′UTRs (Supplementary Figs. 19 and 20). In this study, we compiled GWAS summary statistics for 23 common human diseases and traits from previously published studies (Supplementary Table 7) and evaluated the enrichment of 3′aVariants within trait-associated GWAS SNPs for each tissue using functional genome-wide association analysis[48]. We identified the enrichment of 3′aVariants within 11.5% of tissue-trait pairs. When further compared with known eVariants that are enriched for these traits, we observed that, overall, eQTLs had larger effects than 3′aQTLs for 26.5% of the tissue-trait pairs

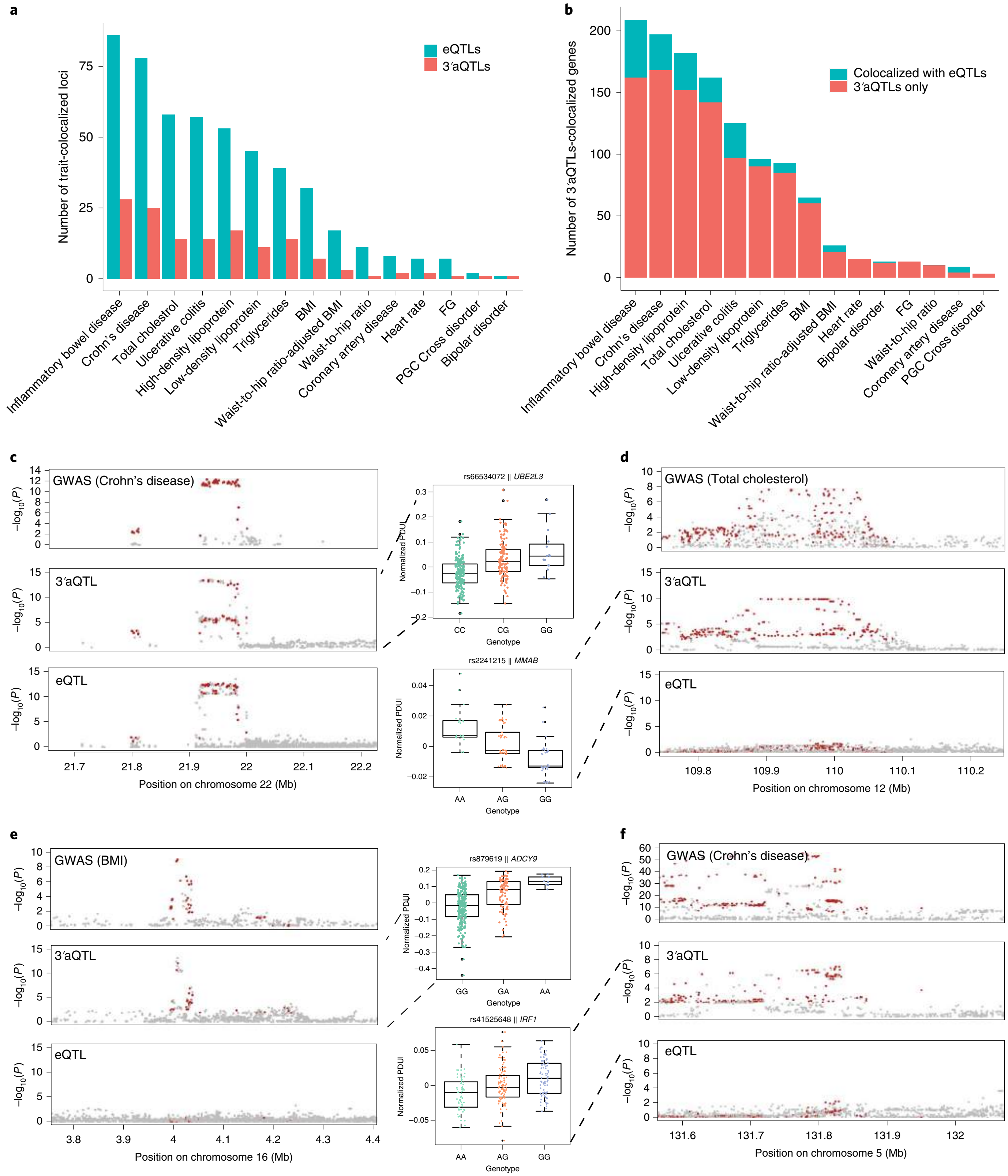


Fig. 7 | Colocalization of 3′aQTLs with complex trait-associated loci. a, Total number of colocalized GWAS signals (*y* axis) for each of 15 traits (*x* axis) across 46 human tissues. FG, fasting glucose. **b**, Number of 3′aQTL-colocalizing genes, colored based on whether the gene also colocalized with eQTLs. Blue represents 3′aQTL-colocalizing genes that are also eQTL colocalizing genes and red represents 3′aQTL-colocalizing genes are not eQTL colocalizing genes. **c**, Colocalization map of Crohn's disease-associated genes with 3′aQTLs in whole blood. The box plot shows that 3′aQTLs are strongly associated with 3′UTR usage in *UBE2L3* ($n=396$). The center horizontal lines represent the median values and the boxes span from the 25th to the 75th percentile. The whiskers extend to 1.5× IQR (bottom). Red indicates 3′aQTLs shared with GWAS SNPs. **d**, Colocalization map of the total cholesterol level trait with 3′aQTLs and eQTLs in liver tissue ($n=135$). **e**, Colocalization map of BMI trait with 3′aQTLs and eQTLs in skeletal muscle tissue ($n=431$). **f**, Colocalization map of Crohn's disease with 3′aQTLs and eQTLs in transformed fibroblasts ($n=291$).

examined. However, in 9.8% of pairs, we found that 3′aQTLs exhibited the increased enrichment of GWAS SNPs compared with eQTLs (Supplementary Table 8), including those associated with Alzheimer's disease and rheumatoid arthritis. Notably, many of the 3′aVariants were enriched in tissues relevant to their respective diseased states, such as the brain putamen (basal ganglia) for Alzheimer's disease and the pituitary gland for rheumatoid arthritis (Fig. 6a,b). Of note, 3′aVariants were also enriched in less biologically relevant tissues, which may represent common 3′aVariants across many tissues or new trait-associated tissues[2].

To quantify the proportion of regulatory variations associated with heritability for each trait, we conducted a partitioned heritability analysis, using LD score regression[49]. Of the traits examined, the median range of SNP heritability that could be explained by 3′aQTLs, sQTLs and eQTLs was 3–7, 13–25 and 10–19% per trait, respectively. Notably, 3′aQTLs were particularly effective for explaining a large proportion of heritability associated with several autoimmune diseases, such as ulcerative colitis, primary biliary cholangitis and Alzheimer's disease. For some diseases, such as multiple sclerosis, 3′aQTLs contributed little to heritability (Fig. 6c and Extended Data Fig. 10). Taken together, although the role of APA in the modulation of these diseases has been studied at the single-gene level, such as for *tau* in Alzheimer's disease[50] and *TCF7L2* in type 2 diabetes[51], our results suggested that 3′aQTLs can explain a significant proportion of disease-associated variants.

Many trait-colocalizing 3′aQTLs are independent of gene expression. The enrichment of 3′aQTLs within disease-associated loci provide disease-specific knowledge about the overall impact of 3′aQTLs but does not necessarily imply a causal relationship. Therefore, we investigated the extent to which 3′aQTLs may function as causal variants for human phenotypes. We used colocalization analysis[52], which identifies 3′aQTLs that share the same putative causal variants with trait-associated signals, to examine 15 complex diseases and traits with known minor allele frequencies (MAFs). Of note, the colocalization model has limited power for the identification of multiple causal variants per gene. In total, 801 trait-associated variants colocalized with either eQTL or 3′aQTL signals. Consistent with previous results[1], 57% of trait-associated variants colocalized with eQTLs in 1 or more tissues. Interestingly, 16.1% of trait-associated variants colocalized with 3′aQTLs in at least 1 tissue (Fig. 7a). Of note, this 3′aQTL colocalization may still be driven by eQTLs or sQTLs (Supplementary Fig. 21). We found that 14 colocalizing 3′aQTLs were overlapped with pQTLs[34]. For example, rs503366 is not only a pQTL for *MTRF1L*, but also the lead 3′aVariant that colocalized with bipolar disorder GWAS variants (the posterior probability of a model with one shared causal variant (PP4) = 0.922). We also found that 83.7% (1,019 out of 1,218) of 3′aQTL-colocalizing genes were not eQTL colocalizing genes (Fig. 7b and Supplementary Table 9). We separated all 3′aGenes into two groups based on whether they overlapped with eQTLs. Within each group, we analyzed the differences of APA usage and gene expression with different 3′aQTL alleles. We observed no APA usage differences between eQTL-overlapped 3′aGenes and non-eQTL-overlapped 3′aGenes ($P = 0.06$; Supplementary Fig. 22a). We further found that eQTL-overlapped 3′aGenes tended to have notable gene expression changes ($P < 2.2 \times 10^{-16}$) (Supplementary Fig. 22b), whereas non-eQTL-overlapped 3′aGenes had almost no associated gene expression changes. To explore the potential regulatory mechanisms, we cross-referenced the 3′UTR regions of 3′aGenes with the TargetScan human v.6.2 (ref. [53]) miRNA binding sites and ENCODE RBP CLIP-seq peaks. We found that eQTL-overlapped 3′aGenes have overall greater miRNA binding site density within the 3′UTR region than non-eQTL-overlapped 3′aGenes ($P = 5.695 \times 10^{-5}$; Supplementary Fig. 22c). We did not find any enrichment of RBP binding sites. These results suggest that eQTL-overlapped 3′aGenes tend to affect gene expression through miRNA-mediated regulation but not through RBP regulation.

UBE2L3 is a representative example of the 16.3% of genes that colocalized with both 3′aQTLs and eQTLs. *UBE2L3* is an E2 ubiquitin-conjugating enzyme that promotes the activation of nuclear factor kappa B signaling during immune responses[54]. The rs66534072 locus in *UBE2L3* has been associated with gene expression levels and confers risk for autoimmune diseases[55]. However, the mechanisms through which these genetic variants affect gene expression are unknown. We determined that *UBE2L3* can be subject to APA and can exhibit dynamic 3′UTR use among different individuals. Moreover, the lead 3′aQTL SNP, rs66534072, was significantly correlated with 3′UTR use in *UBE2L3* (Fig. 7c). Specifically, the C allele was associated with the shortening of the *UBE2L3* mRNA 3′UTR, whereas the G allele was associated with the lengthening of the 3′UTR. We examined the tissues where rs66534072 serves as a 3′aQTL for *UBE2L3* and found that most are known to be affected by autoimmune diseases.

Most 3′aQTL trait-colocalized gene pairs are specific to 3′aQTLs and not eQTLs. For instance, *MMAB* encodes an enzyme involved in adenosylcobalamin formation, which is crucial for cholesterol degradation[56]. A total of 288 3′aQTLs were found to associate with *MMAB* 3′UTR use and were directly correlated with total cholesterol level GWAS loci on chromosome 12 (Fig. 7d). Similarly, variants on chromosome 16 that were associated with body mass index (BMI) also colocalized with 3′aQTLs that regulate 3′UTR length changes in *ADCY9* (Fig. 7e). We also observed a strong colocalization pattern between 3′aQTLs in *IRF1* and the significant GWAS loci for multiple autoimmune diseases, including ulcerative colitis, Crohn's disease and inflammatory bowel disease (Fig. 7f). *IRF1* is induced by IFN-γ signaling and promotes innate and acquired immune responses[57]. In contrast, except in musculoskeletal tissue, no strong association between eQTL and *IRF1* expression was observed. Colocalization analyses of musculoskeletal tissue revealed no colocalization patterns between disease-associated loci and *IRF1* eQTLs. In contrast, colocalization patterns for *IRF1* 3′aQTLs and autoimmune diseases were identified in multiple tissues, including transformed fibroblasts (PP4 = 0.97). These results suggested that *IRF1*-associated 3′aQTLs, more than *IRF1*-associated eQTLs, can explain most of the effects of the *IRF1* variations associated with these diseases. Collectively, our data suggest that many 3′aQTLs contribute to human diseases and traits, independent of their roles in the regulation of gene expression.

Discussion

We defined 3′aQTLs as the genetic basis for an emerging human molecular phenotype that is responsible for alternative 3′UTR usage. By reanalyzing large-scale GTEx data, using our DaPars v.2 algorithm, we identified 11,613 APA genes and approximately 0.4 million 3′aQTLs across 46 human tissues. 3′aQTLs were found to be sufficient to alter APA regulation, as demonstrated by CRISPR-based experiments and saturation mutagenesis data. In contrast with other molecular QTLs, such as eQTLs, 3′aQTLs are highly enriched within 3′UTRs. Mechanistically, 3′aQTLs likely induce changes in 3′UTR usage by either modulating the strength of poly(A) signal motifs, RNA secondary structure or RBP binding sites. 3′aQTLs that reside outside of gene-transcribed regions are likely to involve a more complex mechanistic basis as evidenced by recent work revealing connections between DNA methylation, gene looping and APA regulation[58,59]. eQTLs are important molecular features associated with human phenotypic variations. In this study, we demonstrated that 3′aQTLs represent molecular features that contribute to phenotypic variation in human populations at an unexpectedly similar level as eQTLs. Furthermore, we also validated the use of 3′aQTLs as a discovery tool for identifying APA regulators, such as LARP4.

We found that 3′aQTLs can explain a substantial proportion of trait heritability. Colocalization analyses found that 16.1% of trait-associated loci colocalized with 1 or more 3′aQTLs in human tissues. Furthermore, very few of the 3′aQTL-colocalizing trait-associated loci overlapped with eQTLs, indicating that 3′aQTLs and eQTLs are largely independent. We speculate that eQTL-independent 3′aQTLs regulate the stability, translation or cellular localization of target genes independently of the regulation of gene expression. Collectively, the results of our in-depth analyses of the genetic influence of APA events in 46 human tissues increase the fraction of common noncoding variations that can be associated with molecular phenotypes and suggest interpretations that explain how natural variations can shape human phenotypic diversity and tissue-specific diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00864-5.

Received: 30 January 2020; Accepted: 5 April 2021;
Published online: 13 May 2021

References

1. GTEx Consortium.Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
2. Gamazon, E. R. et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
3. Mayr, C. Regulation by 3′-untranslated regions. *Annu. Rev. Genet.* **51**, 171–194 (2017).
4. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* **18**, 18–30 (2017).
5. Mayr, C. What are 3′ UTRs dDoing? *Cold Spring Harb. Perspect. Biol.* **11**, a034728 (2018).
6. Masamha, C. P. et al. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* **510**, 412–416 (2014).
7. Weng, T. et al. Cleavage factor 25 deregulation contributes to pulmonary fibrosis through alternative polyadenylation. *J. Clin. Invest.* **129**, 1984–1999 (2019).
8. Park, H. J. et al. 3′ UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk. *Nat. Genet.* **50**, 783–789 (2018).
9. Yoon, O. K., Hsu, T. Y., Im, J. H. & Brem, R. B. Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLoS Genet.* **8**, e1002882 (2012).
10. Zhernakova, D. V. et al. DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genet.* **9**, e1003594 (2013).
11. Stacey, S. N. et al. A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat. Genet.* **43**, 1098–1103 (2011).
12. Higgs, D. R. et al. α-Thalassaemia caused by a polyadenylation signal mutation. *Nature* **306**, 398–400 (1983).
13. van der Maarel, S. M., Tawil, R. & Tapscott, S. J. Facioscapulohumeral muscular dystrophy and DUX4: breaking the silence. *Trends Mol. Med.* **17**, 252–258 (2011).
14. Fahiminiya, S. et al. A polyadenylation site variant causes transcript-specific BMP1 deficiency and frequent fractures in children. *Hum. Mol. Genet.* **24**, 516–524 (2015).
15. Garin, I. et al. Recessive mutations in the *INS* gene result in neonatal diabetes through reduced insulin biosynthesis. *Proc. Natl Acad. Sci. USA* **107**, 3105–3110 (2010).
16. Hellquist, A. et al. The human *GIMAP5* gene has a common polyadenylation polymorphism increasing risk to systemic lupus erythematosus. *J. Med. Genet.* **44**, 314–321 (2007).
17. Graham, R. R. et al. Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proc. Natl Acad. Sci. USA* **104**, 6758–6763 (2007).
18. Cannavò, E. et al. Genetic variants regulating expression levels and isoform diversity during embryogenesis. *Nature* **541**, 402–406 (2017).
19. Mariella, E., Marotta, F., Grassi, E., Gilotto, S. & Provero, P. The length of the expressed 3′ UTR is an intermediate molecular phenotype linking genetic variants to complex diseases. *Front. Genet.* **10**, 714 (2019).
20. Xia, Z. et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nat. Commun.* **5**, 5274 (2014).
21. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
22. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
23. Bogard, N., Linder, J., Rosenberg, A. B. & Seelig, G. A deep neural network for predicting and engineering alternative polyadenylation. *Cell* **178**, 91–106.e23 (2019).
24. Kwan, T. et al. Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.* **40**, 225–231 (2008).
25. Hoarau, J.-J., Cesari, M., Caillens, H., Cadet, F. & Pabion, M. *HLA DQA1* genes generate multiple transcripts by alternative splicing and polyadenylation of the 3′ untranslated region. *Tissue Antigens* **63**, 58–71 (2004).
26. Cunninghame Graham, D. S. et al. Association of *IRF5* in UK SLE families identifies a variant involved in polyadenylation. *Hum. Mol. Genet.* **16**, 579–591 (2007).
27. Sheng, G., dos Reis, M. & Stern, C. D. Churchill, a zinc finger transcriptional activator, regulates the transition between gastrulation and neurulation. *Cell* **115**, 603–613 (2003).
28. Lyons, J. J. et al. Elevated basal serum tryptase identifies a multisystem disorder associated with increased *TPSAB1* copy number. *Nat. Genet.* **48**, 1564–1569 (2016).
29. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
30. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).
31. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
32. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
33. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
34. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
35. Thomas, L. F. & Sætrom, P. Single nucleotide polymorphisms can create alternative polyadenylation signals and affect gene expression through loss of microRNA-regulation. *PLoS Comput. Biol.* **8**, e1002621 (2012).
36. Lee, J. Y., Yeh, I., Park, J. Y. & Tian, B. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res.* **35**, D165–D168 (2007).
37. Sun, H. S. et al. A polymorphic 3′UTR element in ATP1B1 regulates alternative polyadenylation and is associated with blood pressure. *PLoS ONE* **8**, e76290 (2013).
38. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
39. Matoulkova, E., Michalova, E., Vojtesek, B. & Hrstka, R. The role of the 3′ untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biol.* **9**, 563–576 (2012).
40. Bava, F.-A. et al. CPEB1 coordinates alternative 3′-UTR formation with translational regulation. *Nature* **495**, 121–125 (2013).
41. Müller-McNicoll, M. et al. SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. *Genes Dev.* **30**, 553–566 (2016).
42. Chen, P.-F., Hsiao, J. S., Sirois, C. L. & Chamberlain, S. J. RBFOX1 and RBFOX2 are dispensable in iPSCs and iPSC-derived neurons and do not contribute to neural-specific paternal *UBE3A* silencing. *Sci. Rep.* **6**, 25368 (2016).
43. Gruber, A. J. et al. A comprehensive analysis of 3′ end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res.* **26**, 1145–1159 (2016).
44. Dominguez, D. et al. Sequence, structure, and context preferences of human RNA binding proteins. *Mol. Cell* **70**, 854–867.e9 (2018).
45. Wan, Y. et al. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706–709 (2014).
46. Yang, R. et al. La-related protein 4 binds poly(A), interacts with the poly(A)-binding protein MLLE domain via a variant PAM2w motif, and can promote mRNA stability. *Mol. Cell. Biol.* **31**, 542–556 (2011).
47. Routh, A. et al. Poly(A)-ClickSeq: click-chemistry for next-generation 3′-end sequencing without RNA enrichment or fragmentation. *Nucleic Acids Res.* **45**, e112 (2017).
48. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).

49. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
50. Dickson, J. R., Kruse, C., Montagna, D. R., Finsen, B. & Wolfe, M. S. Alternative polyadenylation and miR-34 family members regulate tau expression. *J. Neurochem.* **127**, 739–749 (2013).
51. Locke, J. M., Da Silva Xavier, G., Rutter, G. A. & Harries, L. W. An alternative polyadenylation signal in *TCF7L2* generates isoforms that inhibit T cell factor/lymphoid-enhancer factor (TCF/LEF)-dependent target genes. *Diabetologia* **54**, 3078–3082 (2011).
52. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
53. Agarwal, V., Bell, G. W., Nam, J.-W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *eLife* **4**, e05005 (2015).
54. Lewis, M. J. et al. *UBE2L3* polymorphism amplifies NF-κB activation and promotes plasma cell development, linking linear ubiquitination to multiple autoimmune diseases. *Am. J. Hum. Genet.* **96**, 221–234 (2015).
55. Wang, S. et al. A functional haplotype of *UBE2L3* confers risk for systemic lupus erythematosus. *Genes Immun.* **13**, 380–387 (2012).
56. Holleboom, A. G., Vergeer, M., Hovingh, G. K., Kastelein, J. J. & Kuivenhoven, J. A. The value of HDL genetics. *Curr. Opin. Lipidol.* **19**, 385–394 (2008).
57. Kano, S. et al. The contribution of transcription factor IRF1 to the interferon-γ-interleukin 12 signaling axis and T_H1 versus T_H-17 differentiation of $CD4^+$ T cells. *Nat. Immunol.* **9**, 34–41 (2008).
58. Nanavaty, V. et al. DNA methylation regulates alternative polyadenylation via CTCF and the cohesin complex. *Mol. Cell* **78**, 752–764.e6 (2020).
59. Mittleman, B. E. et al. Alternative polyadenylation mediates genetic regulation of gene expression. *eLife* **9**, e57492 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Mapping of GTEx RNA-seq data. Original RNA-seq reads were aligned with the human genome (hg19/GRCh37) using STAR v.2.5.2b[60], with the following alignment parameters: outSAMtype, BAM; SortedByCoordinate; outSAMstrandField, intronMotif; outFilterMultimapNmax, 10; outFilterMultimapScoreRange, 1; alignSJDBoverhangMin, 1; sjdbScore, 2; alignIntronMin, 20; and alignSJoverhangMin, 8. The resulting sorted BAM files were converted into bedGraph formats using BEDTools version 2.17.0 (ref. [61]).

Covariate correction. To account for hidden batch effects and other unobserved covariates in each tissue, we first corrected the sample genotype for population structure. Briefly, we first removed sites marked as 'wasSplit' from the GTEx analysis freeze variant call format (VCF) using BCFtools v.1.3, leaving 39,741,769 biallelic sites. The variants were further filtered with a call rate of >99% and MAF >5%; LD pruning was performed using PLINK v.2.0. The top three principal components from the principal component analysis were consistent with the known three main subpopulations, including White, Black or African American and Asian, in the GTEx samples. We further used PEER[21] with sex, RNA integrity number, top 5 genotype principal components and genotyping platforms as the known covariates to estimate a set of latent covariates for the PDUI values in each tissue. The number of PEER factors was optimized based on suggestions from the GTEx Consortium[1]; for tissue sample sizes <150, 15 PEER factors were chosen. Thirty PEER factors were chosen if the sample size ranged from 150 to 250 and 35 peer factors were chosen for >250 samples. We analyzed the correlation between PEER factors and covariates reported for the GTEx samples and noticed that many of these covariates were strongly associated with PEER factors, such as nucleic acid isolation batch and total ischemic time, which were associated across tissues (Extended Data Fig. 1). We also included three measurements for 3′Bias statistics: (1) 3′ 50-base normalization, which is the ratio between the coverage at the 3′-end and the average coverage of the full transcript, averaged over all transcripts; (2) 5′ 50-base normalization, which is the ratio between the coverage at the 5′-end and the average coverage of the full transcript, averaged over all transcripts; and (3) the number of transcripts that have at least one read at their 5′-end. The inferred PEER factors were highly correlated with the 3′Bias statistics (Extended Data Fig. 1), indicating that most of the 3′Bias effects have been corrected by our PEER analysis.

Furthermore, to comprehensively evaluate the other genotypic covariates, we correlated the PEER factors with donor covariates in each tissue. We observed that our PEER factors were consistently correlated with several donor covariates such as donor death, ischemic time, Hardy scale, EBV immunoglobulin M antibody and age (Extended Data Fig. 2).

3′aQTL mapping for each tissue. A whole-genome sequencing variant file for 635 individuals was obtained from the GTEx database of Genotypes and Phenotypes (dbGaP) website (phs000424.v7.p2), under the name 'GTEx_Analysis_2016-01-15_v7_WholeGenomeSeq_635Ind_PASS_AB02_GQ20_HETX_MISS15_PLINKQC.vcf.gz', from which 17 samples and all the variants that failed to pass the quality control step initially defined by the GTEx Consortium[1] were removed. Any individuals with no RNA-seq data were also removed. 3′aQTL mapping was performed separately for each tissue. Subset VCF data for each tissue were extracted, using BCFtools. VCF files were transformed into an SNP matrix file, including genotyping information, using BioAlcidae v.2.27.1 (ref. [62]). SNPs with a MAF of <0.01 were filtered and at least 10 counts per allele were required. We then tested associations for SNPs within an interval of 1 Mb from the 3′UTR region, with normalized PDUI values, in each tissue, using Matrix eQTL[22], in a linear regression framework.

Permutation analysis was conducted to identify significant 3′aQTL-associated gene pairs. Individual labels were randomly sampled 1,000 times and the minimum P value for each SNP and gene was recorded after each 3′aQTL mapping. These empirical P values were adjusted using the qvalue v.2.0.0 R package[63]. Genes with a $q < 0.05$ were considered to be significant APA genes. All APA gene-associated 3′aQTLs were subsequently identified with the FDR set to 5%.

Fine-mapping of causal variants to 3′aQTLs. We used SuSiE[30] to fine-map 3′aQTL. SuSiE can operate on individual-level data (genotypes and APA phenotypes) and can efficiently analyze loci containing many independent effect variables. We allowed a maximum of 10 independent effects in our analysis. Additionally, we verified our SuSiE results with causal variant identification in associated regions analysis[64], which uses summary statistics (z-scores derived from 3′aQTL association P values) and LD matrices but is limited to the detection of a small number of independent effects per region due to its computational capability constraints.

3′aQTL sharing and specificity analyses among tissues. 3′aQTL sharing and specificity among tissues were analyzed using MASH[31]. Briefly, we converted 3′aQTL association statistics to MASH formats. Lead 3′aQTLs and random SNP sets for each APA gene were extracted from each tissue to calculate MASH priors. A total of 4,470 genes, with no data missing from any tissue, were retained to train the MASH model. Prior covariance matrices were inferred via Empirical Bayes matrix factorization, implemented in factors and loadings by adaptive shrinkage; the multivariate 3′aQTL model was constructed using MASH. Posterior effect sizes were computed by applying the trained model to the lead 3′aQTLs sets. MASH aims to elucidate the heterogeneity of 3′aQTL effect sizes across tissues (Fig. 2). With MASH, we can learn about which 3′aQTLs have tissue-specific effect sizes and which have effect sizes consistent across tissues. This provides interesting insights into the genetic architecture of APA in different tissues. The MASH model was trained on a large random subset of SNPs[31], not the lead SNPs. The trained model was then applied to one lead SNP per gene for posterior inferences, to avoid dealing with LD between SNPs when more than one SNP in a gene was involved. Such 'one effect per region' simplification is widely accepted in a similar context to circumvent LD complications when it comes to evaluating association signals in a small region[48,52,65]. This essentially limits the scope of the investigation to a subset of 3′aQTLs but it is sufficient for our purpose to learn patterns of 3′aQTL sharing across tissues. If a lead SNP is only significant in one tissue and not the others, it will be considered a tissue-specific 3′aQTL; however, if the lead SNP is also significantly associated with APA in other tissues, even though the associations in these tissues are not as strong as the tissue based on which it is selected, it will be considered a shared 3′aQTL among tissues.

To examine whether MASH-estimated magnitudes were affected by read depth, we first downsampled 80% of the raw reads in each sample for the 5 representative tissues and reran the whole analysis. The correlations between the same tissues with different sequencing depths (100 versus 80%) were much stronger than the correlations between different tissues with the same sequencing depths (Supplementary Fig. 9a). We also downsampled the samples in each tissue to match the lowest coverage level, 15 million reads, among the included tissue samples. Still, we observed much stronger correlations between the same tissues with different sequencing depths than between different tissues at the same sequencing depth (Supplementary Fig. 9b).

Prediction of *trans* regulator of APA. For a gene G in a tissue type, all samples were ranked based on the expression levels of gene G. The top 10 most highly expressed samples and bottom 10 least expressed samples were chosen as the two groups. If the mean gene expression fold change between the two groups was >2 with $P < 0.05$, these two groups were treated as control and knockdown groups. Then, the PDUI values between the groups could be compared to identify significant dynamic APA genes between the high and low expression groups of gene G. Using this strategy, we calculated the number of 3′UTR shortening or lengthening effect of each gene, which regulates significant dynamic APA events between the high and low expression groups. The gene will be predicted as a *trans* regulator of APA if $P < 0.05$. We have validated our method in a few known APA regulators, such as CSTF2, which was described as an APA regulator promoting 3′UTR shortening. We observed that there was a marked shift of 3′UTR shortening in individuals with highly expressed CSTF2 (Extended Data Fig. 8a). We also investigated our newly detected APA regulator, LARP4. We often observed many APA events when comparing LARP4$^{\mathrm{high}}$ and LARP4$^{\mathrm{low}}$ individuals (Extended Data Fig. 8b).

Colocalization analyses. We utilized a Bayesian colocalization approach to identify GWAS signals that could exhibit the same genetic effects between eQTLs and 3′aQTLs, using the coloc v.3.2-1 R package[52]. The full summary statistics for 15 GWAS were used when the MAF was available. For each GWAS trait, we extracted the sentinel SNPs, which were defined as GWAS SNPs with $P < 5 \times 10^{-8}$ and located at least 1 Mb away from more significant variants. The colocalized signals were searched for within the 100-kb surrounding region of sentinel SNPs. As defined by the coloc method, five posterior probabilities (PPs) were calculated. PP0 represents the null model of no association. PP1 and PP2 represent the probability that causal genetic variants are either associated with disease signals only or with 3′aQTLs only, respectively. PP3 represents the probability that the genetic effects of disease signals and 3′aQTLs are independent and PP4 represents the probability that disease signals and 3′aQTLs share causal SNPs. The genes were defined as colocalization events if $\mathrm{PP4} \geq 0.75$ and $\mathrm{PP4}/(\mathrm{PP4}+\mathrm{PP3}) \geq 0.9$. Region visualization plots were constructed using LocusZoom v.1.4 (ref. [66]). LDs between reference SNPs and 3′aQTLs were calculated using PLINK[67].

Cell culture and cloning. The HEK 293T cell line (catalog no. CRL-3216; ATCC) was grown in high-glucose DMEM supplemented with 10% FCS and 50 U ml^{-1} penicillin-streptomycin (Thermo Fisher Scientific). The oligonucleotides used for cloning are listed in Supplementary Table 5. pST1374-NLS-flag-linker-Cas9 and pGL3-U6-sgRNA-PGK-puromycin plasmids for CRISPR targeting were a gift from X. Huang (plasmid nos. 44758 and 51133, respectively; Addgene). Each pair of oligonucleotides of sgRNAs was annealed and cloned into a pGL3-U6-sgRNA plasmid. The identities were confirmed by Sanger sequencing. *LARP4* RNA interference experiments were performed using a two-hit strategy, as described previously[68]. Briefly, 60 pmol of *LARP4* siRNA (SASI_Hs01-00187288; Sigma-Aldrich) was diluted in 100 µl of Opti-MEM. For each siRNA, 3 µl of RNAiMAX (Thermo Fisher Scientific) was diluted in 100 µl of Opti-MEM and incubated for 5 min at room temperature. Diluted siRNA and RNAiMAX were mixed and incubated for another 20 min at room temperature. Cells were seeded in 12-well plates at a density of 4×10^5, in 1 ml of regular growth medium, immediately before adding the complexes. Transfected cells were incubated at 37 °C and 5% CO_2 for 24 h. For the second forward transfection, 90 pmol of siRNA and 4.5 µl of RNAiMAX were used to form transfection complexes, as described for the first transfection. The medium was replaced with fresh medium before adding the complexes. After another 24 h, cells were expanded to 6-well plates and grown for a total of 72 h before being collected. To check protein expression, anti-FLAG-HRP M2 (catalog no. A8592; Sigma-Aldrich) in 1:5,000 dilution, anti-alpha Tubulin (catalog no. ab15246; Abcam) in 1:2,000 dilution and anti-GAPDH (catalog no. AM4300; Thermo Fisher Scientific) in 1:4,000 dilution was used for western blotting.

CRISPR genomic editing. CRISPR was used to precisely incorporate the FLAG sequence (gat tac aag gat gac gac gat aag), as described previously[69], into all endogenously expressed LARP4 proteins, at the N terminus. Briefly, a 100-bp genomic sequence surrounding the translation start site (TSS) was input into the CRISPOR program (http://crispor.tefor.net/crispor.py) for guide RNA (gRNA) prediction.

Two gRNAs were selected based on the following: (1) the shortest distance between the Cas9 cutting site (NGG is the protospacer adjacent motif) and the FLAG insertion site; and (2) the specificity score, based on the number of off-target effects. To design the single-strand DNA donor template, a 200-bp genomic sequence (including the 24 bases of the FLAG sequence in the middle) surrounding the TSS was synthesized by Integrated DNA Technologies. The protospacer adjacent motif on the donor template was mutated silently to avoid being attacked by transfected gRNA/Cas9. Equal amounts of gRNA and Cas9 plasmids (720 ng in total) were mixed with 10 pM (approximately 660 ng) of donor template and transfected into 4×10^5 HEK 293T cells in 24-well plates with Lipofectamine 2000. Cells were moved to 6-well plates after overnight incubation; selection (10 µg ml^{-1} of blasticidin and 1 µg ml^{-1} puromycin) was started 24 h after transfection for a total of 48 h. Cells were expanded in regular growth medium, without selection antibiotics. FLAG western blots were performed to determine the signal from pools of cells and confirm the signal from clonal lines. Genomic DNA was extracted from those clonal lines with FLAG signals on western blots. PCR was performed to amplify the approximately 200-bp fragment containing the FLAG sequence; the product was resolved on agarose gels to determine the homogeneity of FLAG insertion on all alleles of the target gene.

For 3′aQTL alterations, double-stranded DNA donor templates (approximately 500 bp from each of three genes) were amplified from HEK 293T genomic DNA. 3′aQTLs were designed to be located approximately two-thirds downstream from the 5′-end for higher CRISPR efficiency[70]. PCR-based mutagenesis was performed to alter the 3′aQTLs. Transfection and selection were performed as described above. RNA and genomic DNA were extracted from a pool of cells for distal PAS usage measurements and Sanger sequencing, respectively.

PAC-seq. To identify alternative polyadenylation sites, PAC-seq[47,71] was adopted to sequence *LARP4* knockdown samples. Briefly, poly(A) mRNA was enriched from 5 µg of total RNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs), as described by the manufacturer's protocol. All enriched mRNA was reverse-transcribed into complementary DNA. First, 2 µl of a 5-mM mixture containing 3′-azido-2′,3′-dideoxyadenosine-5′-triphosphate (N-4007, N-4008 and N-4014; TriLink Biotechnologies) and deoxynucleoside triphosphate, at a ratio of 1 to 5, was added to the RNA sample together with 1 µl of 100 µM 3′Illumina_4N_21T primer. Regular RT–qPCR steps, using SuperScript III, were performed. The sample was treated with 1 µl of ribonuclease H (Thermo Fisher Scientific) for 20 min at 37°C, followed by 10 min at 80°C for inactivation. cDNA was purified using AMPure XP beads, as described by the manufacturer's instructions, and eluted in 12 µl of 50 mM of HEPES, pH 7.4. The click reaction was performed by first adding 23 µl of premixed Click-Adaptor (20 µl of dimethylsulfoxide and 3 µl of 5 µM of Click-Adaptor) to 10 µl of cDNA and then adding 2.4 µl of premixed catalyzer (0.4 µl of 50 mM of vitamin C and 2 µl of 10 mM of Copper(II)-TBTA (Lumiprobe)). After a 30-min incubation at room temperature, 2.4 µl of catalyzer was added to the reaction to boost reaction efficiency. 5′ Clicked cDNA was purified using AMPure XP beads.

PCR amplification was performed using 5′ short universal primer and 3′ indexing primer, which has a unique index for each sample. One*Taq* 2X Master Mix (New England Biolabs) was used to amplify the library under the following conditions: 1 min at 94 °C, 30 s at 55 °C, 10 min at 68 °C and 16 cycles of 30 s at 94 °C, 30 s at 55 °C, 2 min at 68 °C. Finally, the PCR extension was performed at 68 °C for 5 min, followed by 4 °C, indefinitely. The library was purified using AMPure XP beads; size selection was performed on 2% E-Gel EX Agarose Gels (Thermo Fisher Scientific), targeting fragments between 200 and 400 bp. The library was extracted from the gel using ZYMO DNA Clean & Concentrator 5 and quantified by a Qubit 3.0 Fluorometer (Thermo Fisher Scientific) before being sequenced on an Illumina next-generation sequencer. PAC-seq data were analyzed with the differential poly(A) clustering DPAC[72] pipeline using the exon-centric approach, with the --P --M --C --A --B and --D options. The results were filtered such that genes or exons required a minimum of 10 mean reads in each sample, a 1.5-fold change and an adjusted $P < 0.01$ to be considered significantly differentially expressed. Genes with more than one PAS also required a percentage distal PAS usage change of 20% to be considered a change in the length of the 3′UTR.

Nuclear and cytosolic protein extraction. Cells were washed and collected in cold PBS and resuspended in a fivefold cell pellet volume of Buffer A (10 mM of Tris, pH 8, 1.5 mM of $MgCl_2$, 10 mM of KCl, 0.5 mM of dithiothreitol (DTT) and 0.2 mM of phenylmethylsulfonyl fluoride). Cells were allowed to swell during a 15-min rotation at 4 °C, then pelleted at 1,000*g* for 10 min, after which cells were homogenized in twofold the original cell pellet volume Buffer A with a Dounce pestle B for 20 strokes on ice. Nuclear and cytosolic fractions were separated by centrifugation at 2,000*g* for 10 min. For the cytosolic fraction, 10× Buffer B (300 mM of Tris, pH 8, 1.4 M of KCl and 30 mM of $MgCl_2$) was added to the supernatant to a final concentration of 1× Buffer B. Debris was removed by centrifugation at 15,000*g* for 30 min at 4°C. For the nuclear fraction, the pellet was washed once with Buffer A before resuspending the original cell pellet volume of Buffer C (20 mM of Tris, pH 8, 420 mM of NaCl, 1.5 mM of $MgCl_2$, 25% glycerol, 0.2 mM of EDTA, 0.5 mM of phenylmethylsulfonyl fluoride and 0.5 mM of DTT). The sample was homogenized with a Dounce pestle B for 20 strokes on ice and rotated for 30 min at 4 °C before centrifugation at 15,000*g* for 30 min at 4 °C. Supernatants were collected from both fractions and subjected to dialysis in Buffer D (20 mM of HEPES, 100 mM of KCl, 0.2 mM of EDTA, 0.5 mM of DTT and 20% glycerol) overnight at 4°C. Lysates were centrifuged again at 15,000*g* for 3 min at 4 °C to remove any precipitates before downstream applications.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw GTEx RNA-seq and genotype files are available to authorized users through dbGaP release, under accession no. phs000424.v7.p2. A list of 3′aQTLs, lead 3′aQTLs and their associated *APA* genes, isoform usage-controlled 3′aQTLs and expression-controlled 3′aQTLs are freely available at Synapse (accession no. syn22236281; https://doi.org/10.7303/syn22236281). Raw and processed PAC-seq data for the LARP4-depletion experiment have been deposited with the Gene Expression Omnibus under accession no. GSE139548. The proteomics data have been deposited with the MassIVE database under accession no. MSV000087000. A website portal dedicated to trait- and disease-associated 3′aQTLs can be accessed at https://wlcb.oit.uci.edu/3aQTL/index.php. Source data are provided with this paper.

Code availability

The open-source DaPars v.2.0 program is freely available at https://github.com/3UTR/DaPars2.

References

60. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
61. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
62. Lindenbaum, P. & Redon, R. bioalcidae, samjs and vcffilterjs: object-oriented formatters and filters for bioinformatics files. *Bioinformatics* **34**, 1224–1225 (2018).
63. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
64. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
65. Aerts, J. et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
66. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
67. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
68. Wagner, E. J. & Garcia-Blanco, M. A. RNAi-mediated PTB depletion leads to enhanced exon definition. *Mol. Cell* **10**, 943–949 (2002).
69. Baillat, D., Russell, W. K. & Wagner, E. J. CRISPR–Cas9 mediated genetic engineering for the purification of the endogenous integrator complex from mammalian cells. *Protein Expr. Purif.* **128**, 101–108 (2016).
70. Song, F. & Stieger, K. Optimizing the DNA donor template for homology-directed repair of double-strand breaks. *Mol. Ther. Nucleic Acids* **7**, 53–60 (2017).
71. Elrod, N. D., Jaworski, E. A., Ji, P., Wagner, E. J. & Routh, A. Development of Poly(A)-ClickSeq as a tool enabling simultaneous genome-wide poly(A)-site identification and differential expression analysis. *Methods* **155**, 20–29 (2019).
72. Routh, A. DPAC: a tool for differential poly(A)–cluster usage from poly(A)-targeted RNAseq data. *G3 (Bethesda)* **9**, 1825–1830 (2019).

Acknowledgements

We thank L. Hou, Z. Cui and members of the Li laboratory for helpful discussions. This work was supported by a Computational Cancer Biology Training Program fellowship Cancer Prevention Research Institute of Texas (CPRIT) grant no. RP170593 (K.-L.H.), National Institutes of Health grants to W.L. (nos. R01HG007538, R01CA193466, R01CA228140), P.J. (no. R03CA223893), E.J.W. (no. R01GM134539) and a University of California, Irvine (UCI) Cancer Center support grant (no. P30CA062203). E.J.W. acknowledges support from University of Texas Medical Branch (UTMB) Department of Biochemistry and Molecular Biology startup funds and the Welch Foundation (no. H-1889). The UTMB Mass Spectrometry Facility is supported in part by CPRIT grant no. RP190682 (W.K.R.). We also thank I. Toufique and P. Papadopoulos at the UCI Research Cyber Infrastructure Center for high-performance computing support.

Author contributions

L.L. and W.L. conceived and supervised the project. L.L., Y.G., Y.C., Y.E.C. and G.W. performed the data analyses. K.-L.H., N.D.E., W.K.R. and P.J. performed the experiments. L.L., Y.L., Y.C., F.P., E.J.W. and W.L. interpreted the data and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

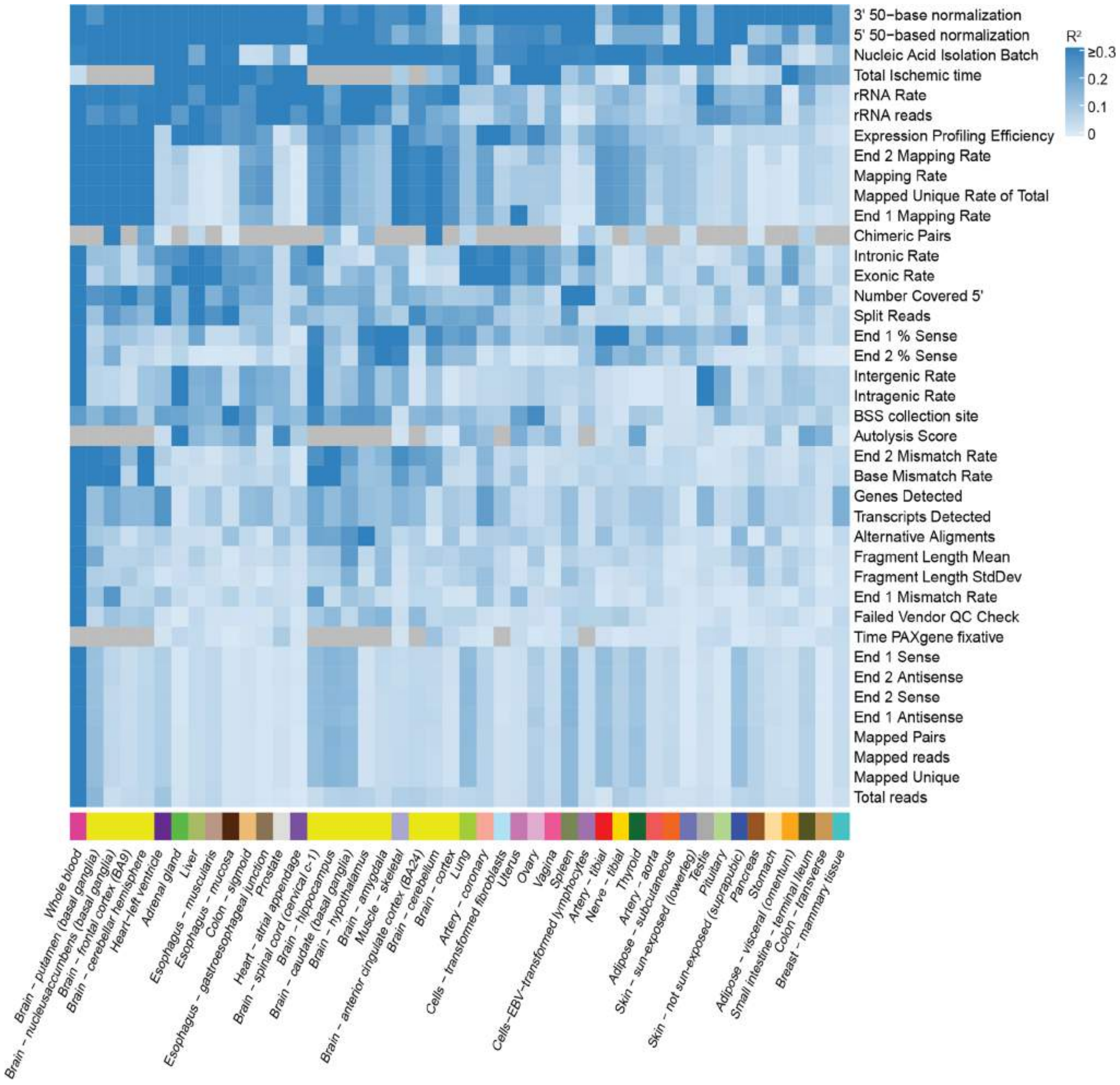
Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00864-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00864-5.

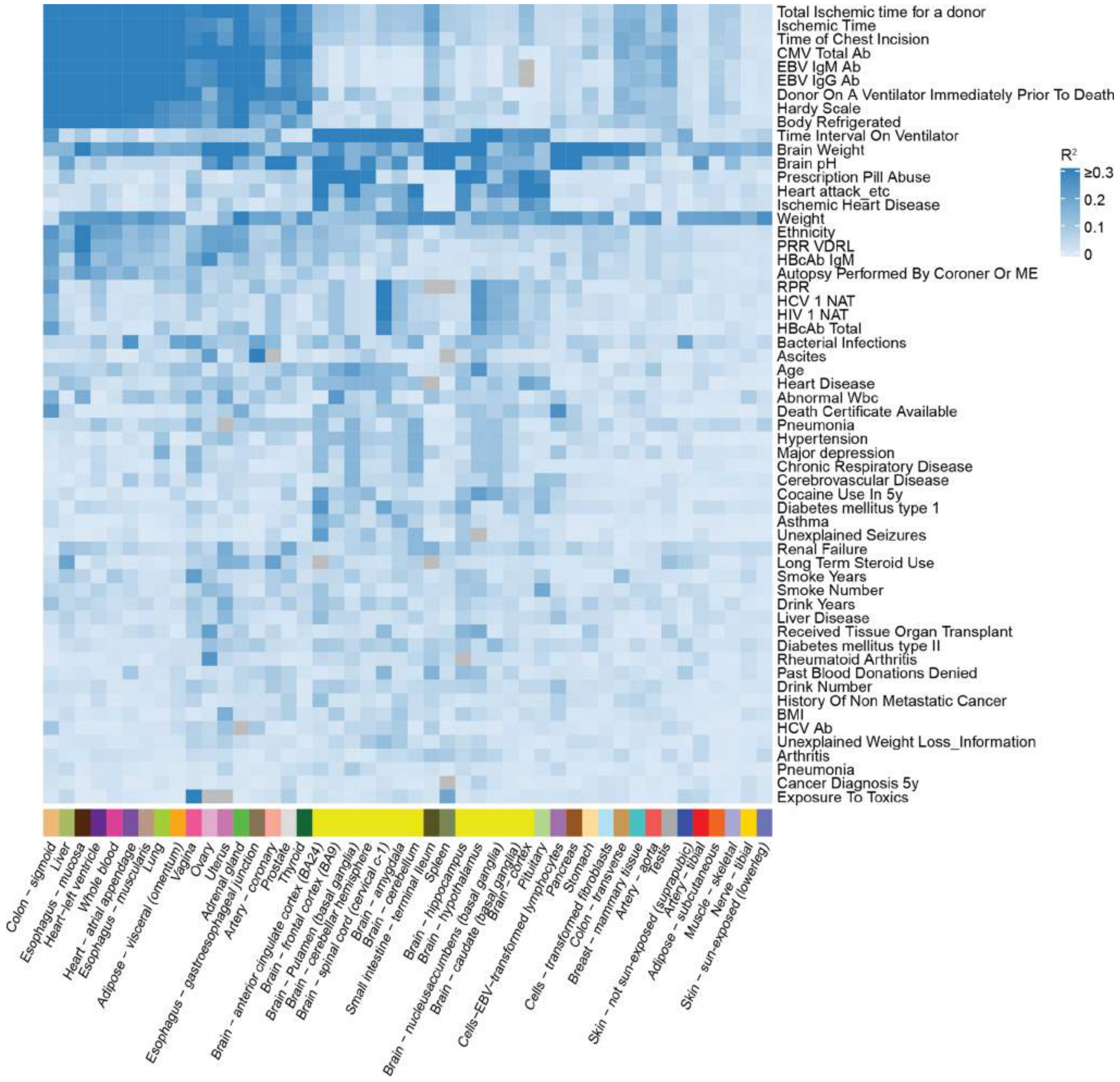
Correspondence and requests for materials should be addressed to E.J.W. or W.L.

Peer review information *Nature Genetics* thanks Stephen Montgomery, Bin Tian and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

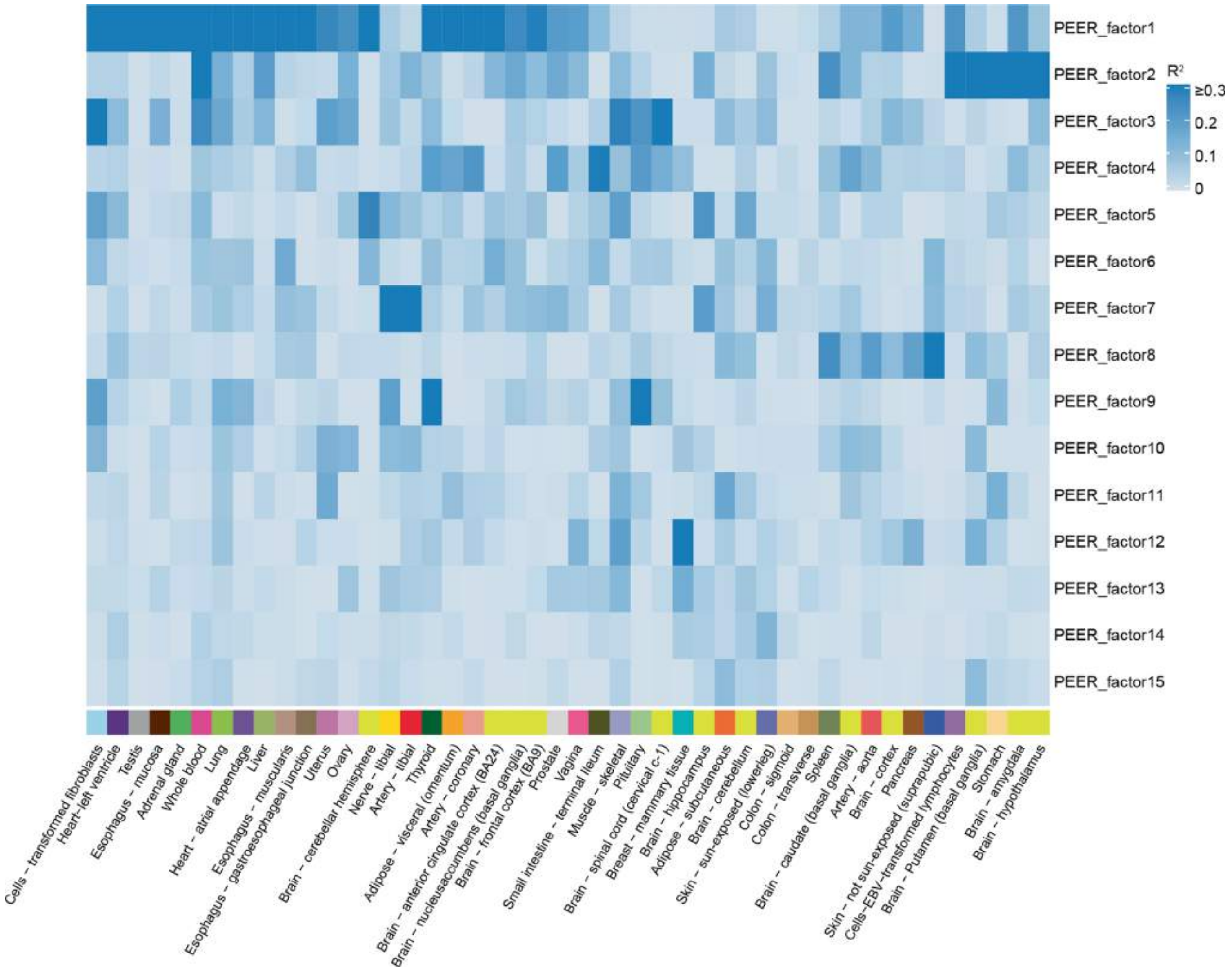
Reprints and permissions information is available at www.nature.com/reprints.



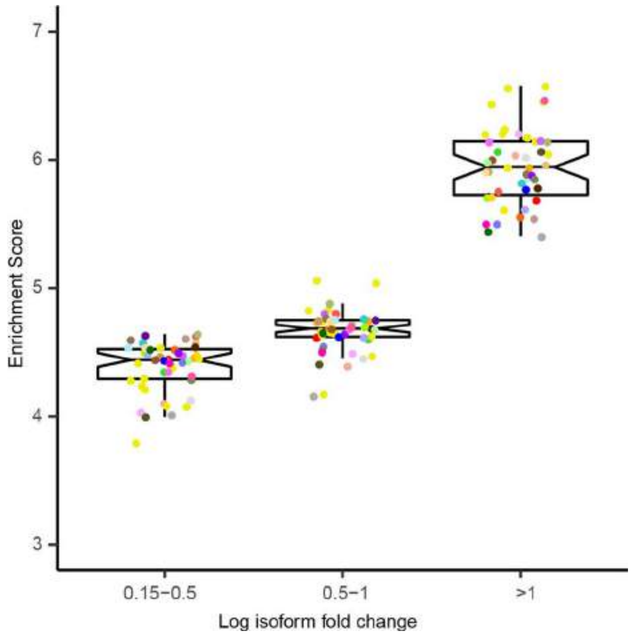
Extended Data Fig. 1 | Known technical covariates associated with inferred PEER factors in each tissue. The R^2 value in each cell represents the percentage of variance explained for each tissue/covariates pair. Only the most relevant sample-specific covariates were used. Gray color represents insufficient data to predict correlations. Each color code below indicates a tissue of origin.



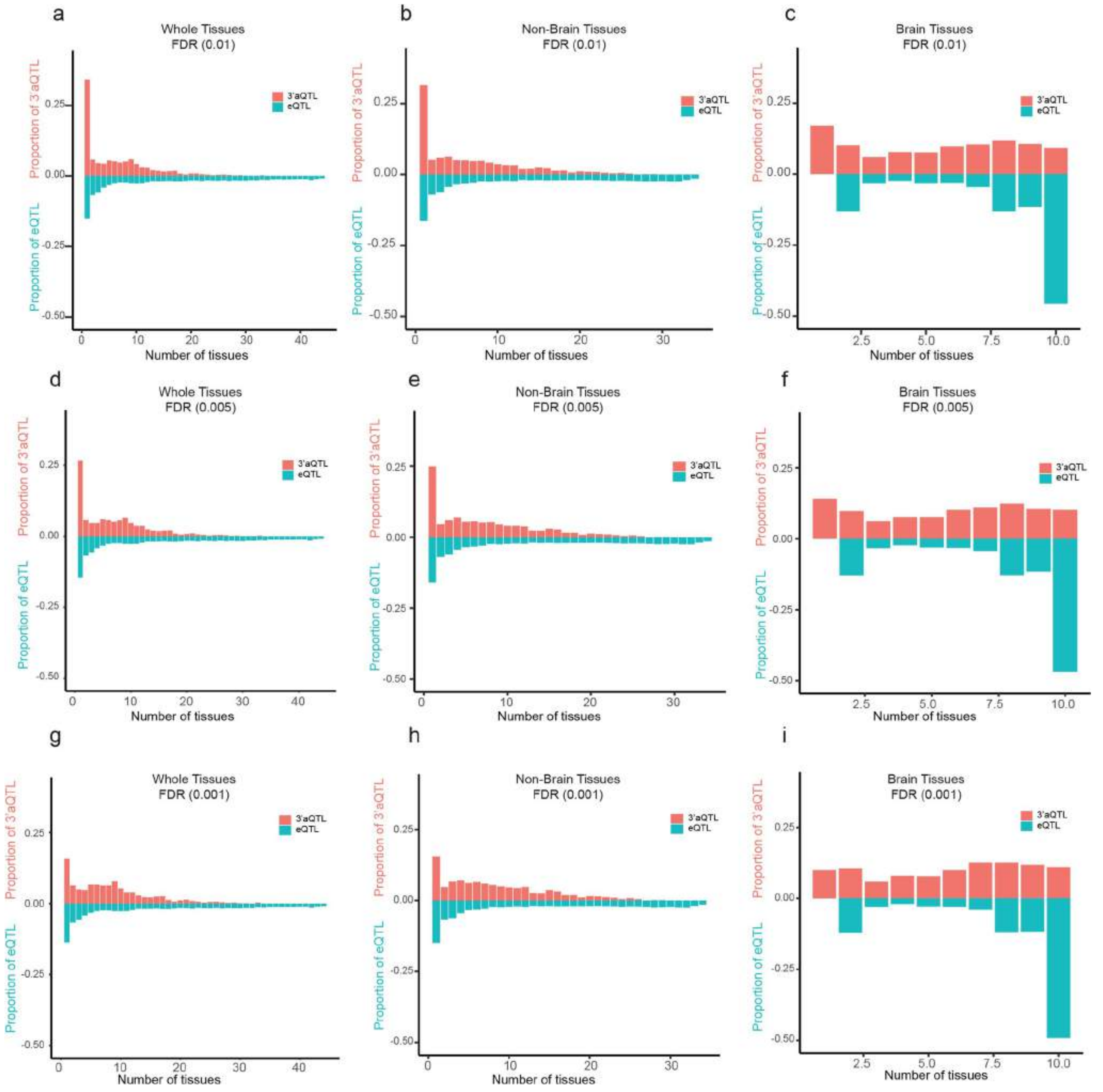
Extended Data Fig. 2 | Known donor covariates associated with inferred PEER factors in each tissue. The R^2 value in each cell represents the percentage of variance explained for each tissue/covariate pair. Only the most relevant donor-specific covariates were used. Gray color represents insufficient data to predict correlations. Each color code below indicates a tissue of origin.



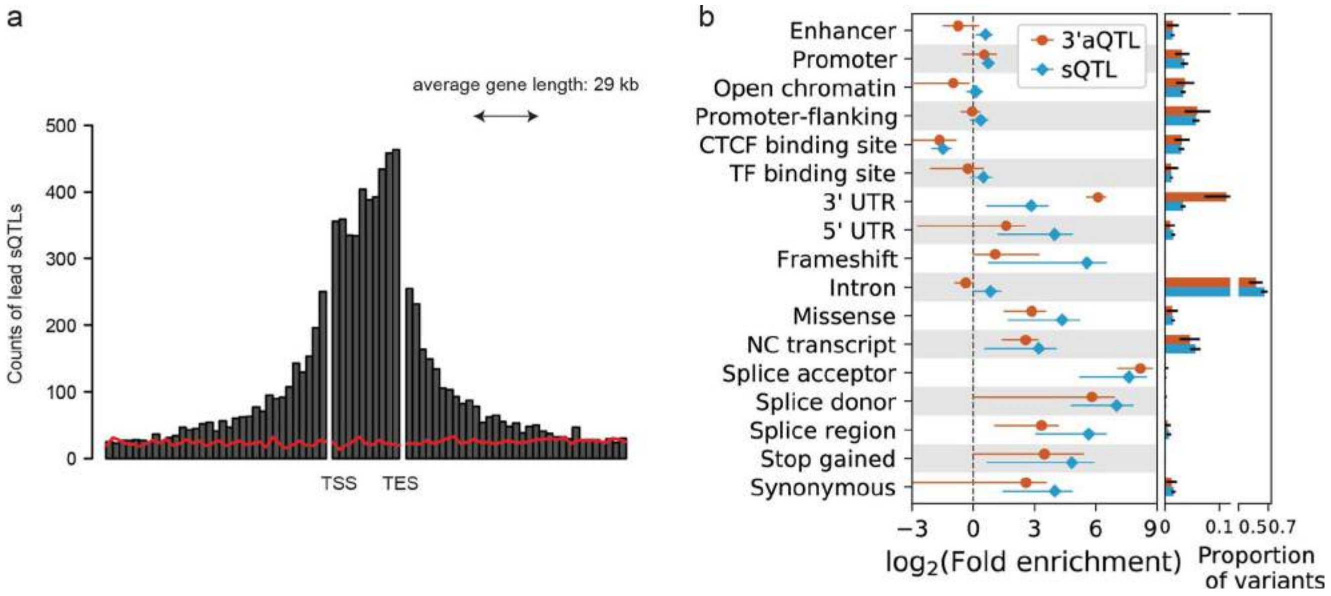
Extended Data Fig. 3 | PEER factors for gene expression associated with PEER factors for PDUI in each tissue. The R^2 value in each cell represents the correlation between the top PEER factors for gene expression (rows) and the most relevant PEER factors for PDUI for each tissue (columns). Each color code below indicates a tissue of origin.



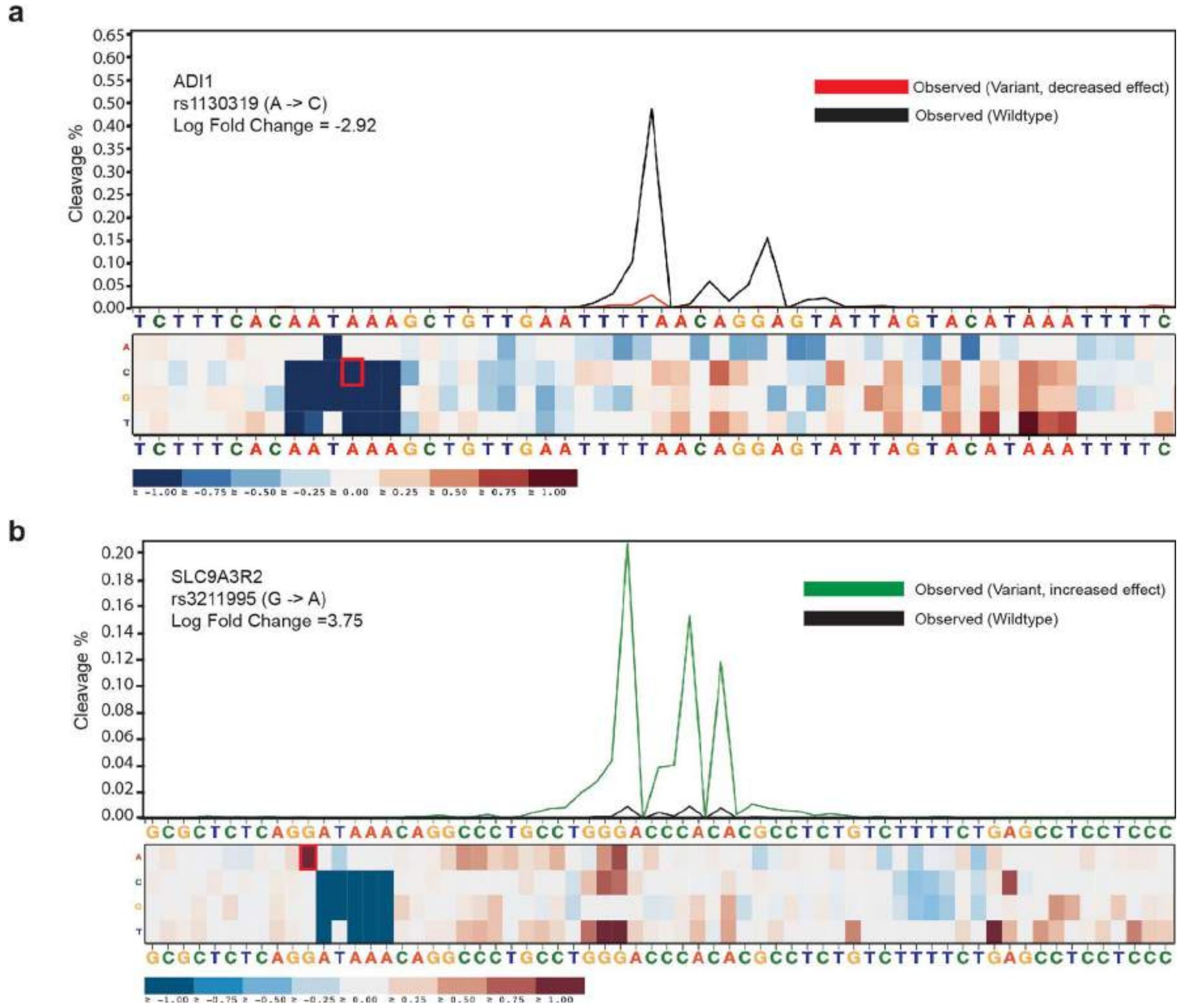
Extended Data Fig. 4 | Enrichment of 3′aQTL in different categories of mutagenesis variants annotations. The enrichment score represents the log odd ratio and accessed by the program Torus. The x-axis represents three categories of variants with different effects in predicting APA isoform log fold change due to the variant. Each color code indicates a tissue of origin. The saturation mutagenesis data with log isoform fold change < 0.15 are not available from Bogard *et al*.



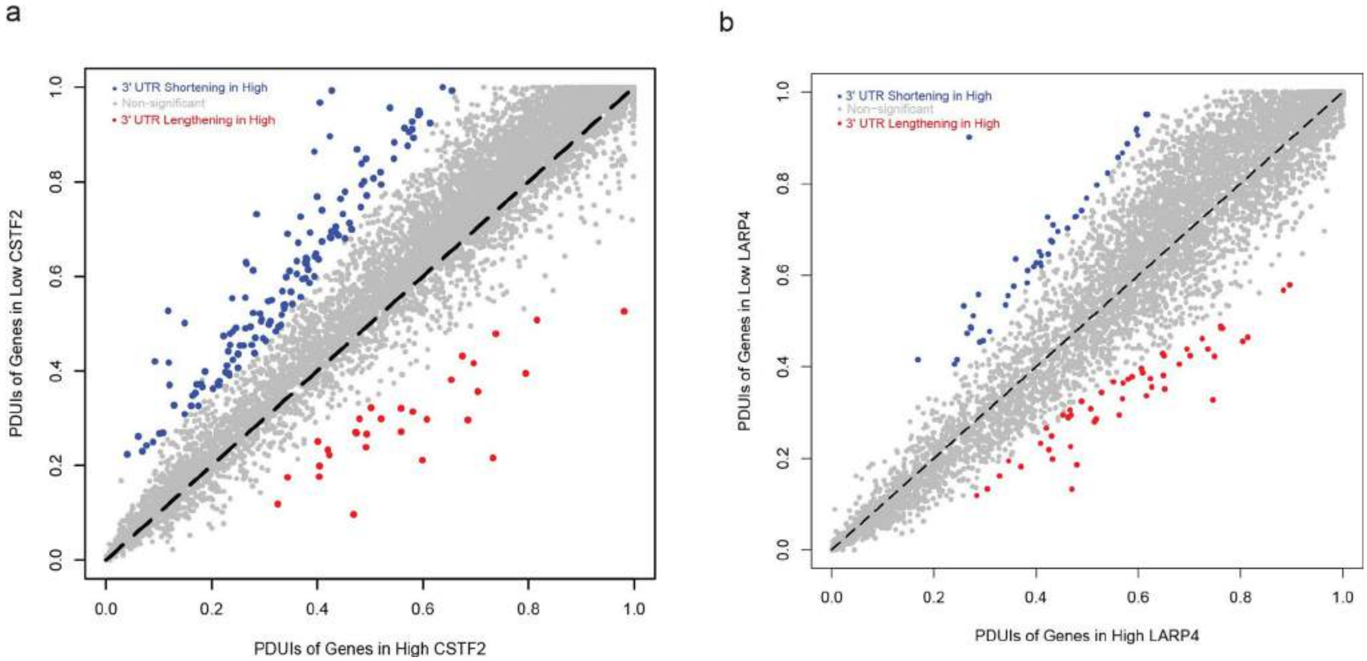
Extended Data Fig. 5 | The sharing magnitude of 3′aQTLs using different FDRs at 0.01, 0.005, 0.001. Histograms showing the estimated proportion of tissues that share lead 3′aQTLs /eQTLs, by magnitude, with other tissues, among all 46 examined tissues, among non-brain tissues only, and among brain tissues only.



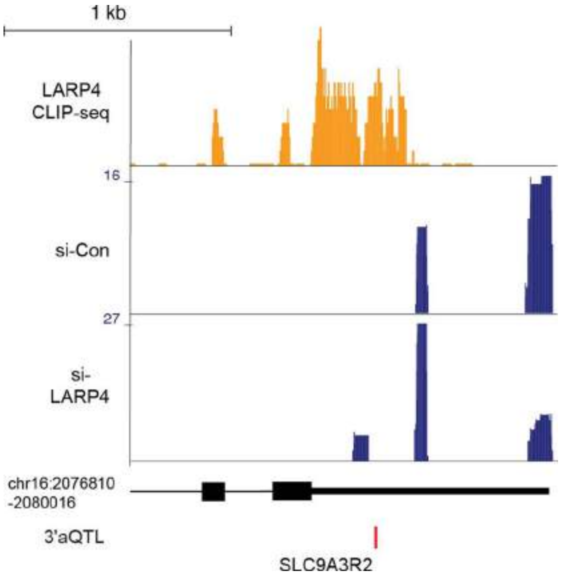
Extended Data Fig. 6 | sQTL have a distinct genomic distribution and functional enrichment compared with 3′aQTL. a, Relative position distance between sQTL and their associated genes. TSS represents the transcription start site; TES represents the transcription end site. Red line represents randomly selected positions within the +/− 1Mb window for each gene. **b**, 3′aQTL and sQTL enrichment in functional annotations. The enrichment is shown as mean with SD across tissues. The proportion of variants was also included for 3′aQTL and sQTL. Data are presented as mean value +/− Standard deviation. n = 46 tissues examined.



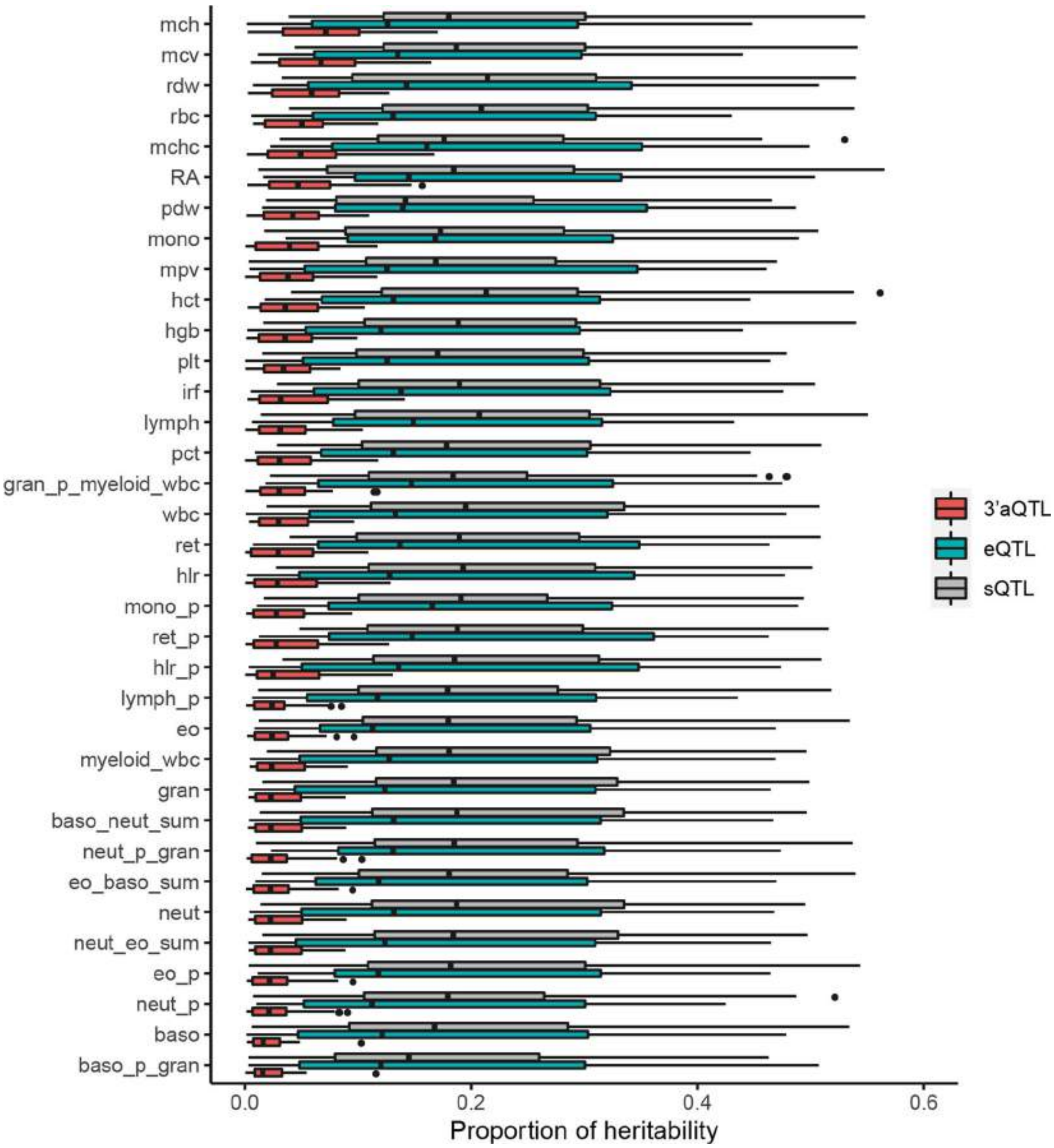
Extended Data Fig. 7 | 3′aQTLs are validated by saturation mutagenesis data. a, Saturation mutagenesis of the *ADI1* PAS. Shown above is the measured wild-type (black) and variant cleavage distribution (red) for the SNP rs1130319. The heatmap below shows the measured isoform fold changes as a result of each SNP. The red box color indicates the SNP rs1130319.



Extended Data Fig. 8 | Trans-regulator APA prediction. a, Scatterplot of the percentage of distal polyA site usage index (PDUI) in CSTF2 over-expressed and low-expressed samples where mRNA significantly shortened (blue) or lengthened (red) are colored. **b**, Scatterplot of PDUI changes for LARP4 over-expressed and low-expressed samples were shown.



Extended Data Fig. 9 | Representative genome browser images of the *SLC9A3R2* gene. *SLC9A3R2* APA is regulated by *LARP4* and binds *LARP4*, as assessed by *LARP4* CLIP-seq.



Extended Data Fig. 10 | A partitioned heritability plot for the percentage of phenotypic variance can be explained, for 35 traits, by 3′aQTLs, eQTLs, and sQTLs in aggregate. The trait/tissue pairs with heritability not significantly greater than 0 are removed. Centre horizontal lines show median values, boxes span from the 25th percentile to the 75th percentile. Whiskers extend to 1.5 × IQR (bottom), where IQR is the interquartile range. n = 46 tissues examined.

nature research

Corresponding author(s):	Eric J Wagner and Wei Li
Last updated by author(s):	3/30/2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ ☒ A description of all covariates tested

☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used
Data analysis	STAR v2.5.2b, APAtrap v1.0, GETUTR v2.0.0, Cufflinks v2.2.1,PEER v1.3, GCTA v1.93, bcftools v1.3, Matrix eQTL v2.1.0, SuSiE (https://github.com/stephenslab/susieR), MEME v5.0.5, SMR v1.0.3, DeepBind v0.11, Fgwas v0.3.6, R 3.4.0, Ldsr v1.0.1, Coloc v3.2-1, bedtools v2.17.0, plink v2.0, The open-source DaPars v2 program is freely available at https://github.com/3UTR/DaPars2.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All the databases/datasets used in the study along with appropriately accessible links in the manuscript under the “Data availability” section as well as in this reporting summary. Raw GTEx RNA-seq and genotype files are available to authorized users through dbGaP release, under accession number phs000424.v7.p2. A list of 3'aQTLs, lead 3'aQTLs, and their associated APA genes, isoform usage-controlled 3'aQTLs, expression-controlled 3'aQTLs are freely available in Synapse (accession number: syn22236281, doi: 10.7303/syn22236281). Raw and processed PAC-seq data for the LARP4-depletion experiment have been deposited to GEO, under the accession number GSE139548. The proteomics data have been deposited to MassIVE database with accession number MSV000087000. A website portal dedicated to trait and disease associated 3'aQTLs can be accessed at https://wlcb.oit.uci.edu/3aQTL/index.php. AREsite2 database can be accessed at http://rna.tbi.univie.ac.at/AREsite2.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined based on the availability of existing GTEx data.
Data exclusions	We removed diseased tissue cells and the leukemia cell line, and seven other tissues, including the cervix endocervix, cervix ectocervix, fallopian tube, bladder, kidney cortex, minor salivary gland, and brain substantia nigra due to small sample sizes.
Replication	The experiments has been performed independently with biological triplicates.
Randomization	The samples have been assigned randomly at the beginning of experiments.
Blinding	The bioinformatics analyses have been corroborated with blinded wet lab experiments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
□	☒ Antibodies
□	☒ Eukaryotic cell lines
☒	□ Palaeontology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Antibodies

Antibodies used	1. Anti-Flag-HRP, clone M2 (Sigma, #A8592); 2. Anti-alpha-tubulin (Abcam, #ab15246); 3. anti-GAPDH, clone 6C5 (ThermoFisher, #AM4300).
Validation	1. ANTI-FLAG M2 monoclonal antibody is useful for detection, identification, and capture of fusion proteins containing a FLAG peptide sequence by common immunological procedures, such as Western blotting and Co-immunoprecipitation. 2. Anti-alpha-tubulin polyclonal antibody is used to detect human microtubule marker by Western blotting. 3. Anti-GAPDH monoclonal antibody is used to detect human GAPDH by Western Blotting. Anti-Flag-HRP M2 monoclonal antibody is registered with ID: AB_439702. It is used for detection of Flag fusion proteins (N-terminal and C-terminal) on Western blots application. The minimum detection range of Flag-fusion protein tested by company is around 8ng shown on Dot blot. Anti-alpha-Tubulin polyclonal antibody is registered with ID: AB_301787. It has been validated by company on Western blot application. This antibody gives a predominant band at expected molecular weight around 55KD after blotting whole cell extracts from mammalian cell lines. Anti-GAPDH monoclonal antibody is registered with ID: AB_437392. It has been validated by company on western blot application. This antibody gives a single band at expected molecular weight around 37KD after blotting whole cell extracts from mammalian cell lines. The citations of each antibody can be bound in the website, The Antibody Registry, by searching its ID.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	293T cell line is purchased from ATCC
Authentication	The 293T cell line was authenticated by STR profiling by ATCC
Mycoplasma contamination	The 293T cell line was tested negative of mycoplasma in our lab using MycoSensor qPCR Assay Kits (#302107, Agilent)
Commonly misidentified lines (See ICLAC register)	There is no commonly misidentified cell line used in the study

Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression

Yipeng Gao[1,2], Lei Li[3], Christopher I. Amos[2], Wei Li[3*]

[1] Graduate Program in Quantitative and Computational Biosciences, Baylor College of Medicine, Houston, TX USA

[2] Department of Medicine, Baylor College of Medicine, Houston, TX USA

[3] Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA USA

* Correspondence: wei.li@uci.edu

Abstract

Alternative polyadenylation (APA) is a major mechanism of post-transcriptional regulation in various cellular processes including cell proliferation and differentiation, but the APA heterogeneity among single cells remains largely unknown. Single-cell RNA sequencing (scRNA-seq) has been extensively used to define cell subpopulations at the transcription level. Yet, most scRNA-seq data have not been analyzed in an "APA-aware" manner. Here, we introduce scDaPars (**D**ynamic Analysis of **A**lternative **P**oly**A**denylation from **S**ingle-**c**ell **R**NA-**s**eq), a bioinformatics algorithm to accurately quantify APA events at both single-cell and single-gene resolution using either 3' end (10x Chromium) or full-length (Smart-seq2) scRNA-seq data. Validations in both real and simulated data indicate that scDaPars can robustly recover missing APA events caused by the low amounts of mRNA sequenced in single cells. When applied to cancer and human endoderm differentiation data, scDaPars not only revealed cell-type-specific APA regulation but also identified cell subpopulations that are otherwise invisible to conventional gene expression analysis. Thus, scDaPars will enable us to understand cellular heterogeneity at the post-transcriptional APA level.

Keywords

Alternative Polyadenylation, Single-cell RNA-sequencing, Single-cell Genomics, Imputation

Introduction

Alternative polyadenylation (APA) is a major mechanism of post-transcriptional regulation under diverse physiological and pathological conditions (Elkon et al. 2013; Tian and Manley 2017). The process of polyadenylation involves endonucleolytic cleavage of the nascent RNA followed by synthesis of a poly(A) tail on the 3' terminus (Tian and Manley 2017). By using different polyadenylation sites (poly(A) sites), which are defined by flanking RNA sequence motifs, APA can generate mRNA isoforms with various 3'-untranslated regions (3' UTRs) in the majority of human genes (Derti et al. 2012; Tian and Manley 2017). While APA in most cases does not alter the protein-coding regions in those mRNA isoforms, it disrupts important *cis*-regulatory elements located in the 3' UTRs, including adenylate-uridylate-rich elements (ARE) and binding sites of miRNAs and RNA-binding proteins, resulting in altered mRNA stability, localization and translation efficiency (Garneau et al. 2007; An et al. 2008; Hoffman et al. 2016).

High-throughput sequencing technologies have revolutionized our understanding of APA over the last decade, illustrating both the pervasiveness of dynamic APA events and complexity of the APA regulatory processes. Recently, multiple studies have shed light on the global regulation of APA in response to changes in cell proliferation and cell differentiation in human diseases including cancer (Tian and Manley 2017; Gruber and Zavolan 2019). Both proliferating cells and transformed cells often express a multitude of alternative mRNA isoforms with shortened 3' UTRs through APA (Sandberg et al. 2008), leading to the activation of several proto-oncogenes such as *CCND1*, by escaping miRNA-mediated repression (Mayr and Bartel 2009). On the other hand, 3' UTR lengthening is more prevalent in cell differentiation (Ji et al. 2009; Ji and Tian 2009). For example, progressive 3' UTR lengthening is observed during mouse embryonic development (Ji et al. 2009), and the generation of induced pluripotent stem cells (iPSCs) (dedifferentiation) is accompanied by global 3' UTR shortening (Ji and Tian 2009). Besides regulating cognate transcripts in *cis*, APA-induced 3' UTR changes can also disrupt competing endogenous RNA (ceRNA) regulation in *trans*, thus repressing several crucial tumor suppressors such as *PTEN* in breast cancer (Park et al. 2018). Although these observations imply a possible cell-state- or cell-type-dependent manner of APA regulation, the variability of APA among individual cells and the utility of APA in revealing novel cell subpopulations remain largely unknown.

Single-cell RNA sequencing (scRNA-seq) has become one of the most widely used technologies in biomedical research by providing an unprecedented opportunity to quantify the abundance of diverse transcript isoforms among individual cells (Shapiro et al. 2013; Saliba et al. 2014). However, methods to quantify relative APA usage across single cells remain underdeveloped. Recently, Velten et al. (Velten et al. 2015) developed an experimental protocol BATseq to quantify various 3'-UTR isoforms at the single-cell resolution. By integrating the standard scRNA-seq protocol and the 3' enriched bulk RNA-seq protocol, Velten et al. found that cell types can be well separated based exclusively on their 3'-UTR isoform usage, indicating that APA is a molecular feature intrinsic to cell states (Velten et al. 2015). While a compelling method, BATseq is hampered by its low sensitivity (~5%) and high procedural complexity (Chen et al. 2017), thereby not being widely adopted in practice. In contrast, standard scRNA-seq data is widely available, yet most of the scRNA-seq data has not been analyzed in an "APA-aware" manner. Since scRNA-seq only captures a small fraction (typically 5%-15%) of the total mRNAs in each cell (Stegle et al. 2015), it can falsely quantify genes, especially lowly expressed ones, as unexpressed; this phenomenon is termed as "dropout". Existing bulk RNA-seq based APA methods such as DaPars (Xia et al. 2014) cannot overcome this vexing challenge when applied directly to scRNA-seq data, as they would lead to a high degree of sparsity in the resulting APA profiles. To address this sparsity, recently published computational approaches such as scDAPA (Ye et al. 2020) and scAPA (Shulman and Elkon 2019) extract and combine reads from cells aggregated based on pre-defined cell types. Alternatively, another study (Kim et al. 2019) aggregates individual genes into “meta-genes” with reference to common functionality. While these strategies cope with the problem of sparsity to some extent, they fail to retain the single-cell or single-gene resolution (Supplemental Table S1).

To fill this knowledge gap, we developed scDaPars (**D**ynamic analysis of **A**lternative **P**oly**A**denylation from **scR**NA-**S**eq), a bioinformatics algorithm for quantifying and recovering APA usage at the single-cell and single-gene resolution using standard scRNA-seq data. Since APA is reported to be regulated in a cell-state- or cell-type-specific manner, scDaPars employs a regression model that enables sharing of APA information across related cells to tackle the sparsity, achieving considerable robustness when applied to noisy scRNA-seq data. In addition, unlike scDAPA and scAPA which are only applicable to 3’ end scRNA-seq datasets, scDaPars can be applied to both 3’ end and full-length scRNA-seq data. To the best of our knowledge, scDaPars is the first single-cell- and single-gene- level APA quantification method for analyzing standard scRNA-seq data.

Results

Overview of the scDaPars algorithm

Figure 1 presents a schematic illustration of the scDaPars algorithm (see “Methods” for detailed definition and computational procedures). Given a scRNA-seq dataset, scDaPars first calculates raw relative APA usage, measured by the percentage of distal poly(A) site usage index (PDUI), based on the two-Poly(A)-site model introduced in DaPars (Xia et al. 2014). scDaPars takes scRNA-seq genome coverage data as input and forms a linear regression model to jointly infer the exact location of proximal poly(A) sites by minimizing the deviation between the observed read density and the expected read density in all single cells. The relative APA usage is then quantified as the proportion of the estimated abundances of transcripts with distal poly(A) sites (longer 3’ UTRs) out of all transcripts (longer and shorter 3’ UTRs), and therefore, genes favoring distal poly(A) site usage (long 3′ UTRs) will have PDUI values near 1, whereas genes favoring proximal poly(A) site usage (short 3′ UTRs) will have PDUI values near 0. This step (step (I)) will generate a PDUI matrix with rows representing genes and columns representing single cells. Of note, the raw PDUI values can only be estimated for genes with sufficient read coverages (default coverage of 5 reads per base), which automatically separates genes into robust genes (genes unaffected by dropout events) and dropout genes for further analysis. Due to the intrinsically low coverage of scRNA-seq data (Brennecke et al. 2013), the resulting PDUI matrix from step (I) is overly sparse with widespread missing data. To further recover the complete PDUI matrix independent of gene expression, we develop a new imputation method by sharing APA information across different cells. For a given cell, scDaPars begins by constructing a nearest neighbor graph based on the sparse PDUI matrix generated in step (I) (Fig.1) to identify a pool of candidate neighboring cells that have similar APA profiles (step (II)). Finally, scDaPars uses a non-negative least square (NNLS) regression model to refine neighboring cells based on robust genes and then borrow APA information in these neighboring cells to impute PDUIs of dropout genes in each cell (step (III)).

Evaluation of the Accuracy and Robustness of scDaPars

To quantitatively evaluate the accuracy of imputed APA usage by scDaPars, we used 384 scRNA-seq libraries of individual human peripheral blood cells (PBMCs) sequenced by Smart-seq2 (Picelli et al. 2013) protocol and a matched bulk RNA-seq library from a benchmark study by Ding et al. (Ding et al. 2020). Since we can estimate poly(A) sites and quantify differential poly(A) sites usage with high sensitivity and specificity in bulk RNA-seq datasets (Xia et al. 2014), we treated the results from the matched bulk sample as pseudo-gold standard for the following evaluation.

First, we showed that scDaPars reliably identified the location of proximal poly(A) sites in single cells. We found that ~84% of poly(A) sites predicted from scRNA-seq data are within 100bp of those predicted in bulk, whereas only ~44% of randomly selected sites from 3' UTR regions are within 100bp of bulk predictions (Fig.2A). We found that ~66.2% of poly(A) sites predicted from scRNA-seq data also overlapped with annotated poly(A) sites complied from RefSeq, Ensembl, UCSC gene models and poly(A)_DB (Wang et al. 2017) within 100bp, and this overlap showed an approximately fivefold enrichment compared with random sites (Fig.2B). In addition, canonical poly(A) signal (PAS) AATAAA was successfully identified by *de novo* motif analysis (Bailey 2011) within the upstream (-100bp) sequence of single-cell predicted poly(A) sites with a p-value ($P = 1.2\times10^{-44}$) similar to that of bulk samples ($P = 5.4\times10^{-48}$) (Fig.2C, Supplemental Fig.S1), supporting the validity of scDaPars's prediction of poly(A) sites.

Next, we showed that scDaPars was able to recover APA usage for genes affected by dropouts in scRNA-seq data. APA is found to be uniquely regulated in distinct immune cell types in PBMCs (Kim et al. 2019). Yet the median Pearson's correlation between APA (PDUI values) of single-cell pairs in the same B cell cluster is only 0.46 when PDUI values were calculated by DaPars (our previous method for bulk RNA-seq) due to dropout effects (Fig. 2D). In contrast, scDaPars successfully recovered PDUI values for most of the affected dropout genes (Supplemental Fig.S2) and increased the median cell-cell correlation by a large margin (0.79) ($P < 2.2\times10^{-16}$) (Fig.2D). We further compared the average APA usage of all single cells with the bulk results. The Pearson's correlation between the average PDUI values of single cells and those of the bulk increased from 0.74 to 0.82 after scDaPars imputation (Fig.2E). Notably, even though the correlation increase was not large, the regression slope increased significantly from 0.59 (DaPars) to 0.8 (scDaPars) ($P = 4.89\times10^{-26}$), indicating

APA usage quantified by scDaPars better represents the linear relationship between the average of single cells and the corresponding bulk.

Finally, we used a simulation study to illustrate scDaPars's ability to identify dynamic APA events (see "Methods") between two cell types. We created a synthetic PDUI matrix of naive and activated CD4 T cells based on bulk RNA-seq data from the DICE project (Schmiedel et al. 2018) (see "Methods"). The naive and activated CD4 T cells are clearly distinguishable using the reference APA profiles estimated from bulk samples (Fig.3A). Additionally, the reference data showed a strong inclination of 3' UTR shortening in activated CD4 T cells ($P = 3.8\times10^{-4}$) (Fig.3D), in line with previous reports that 3' UTR shortening is widely observed upon activation of T cells (Sandberg et al. 2008). However, manually introduced dropout events obscured this differential 3' UTR pattern, in which only ~38% of differential APA genes remained, and the two cell types became less separated by their APA profiles (Fig.3B, E). After we applied the imputation steps of scDaPars, ~79% of differential APA genes are recovered and the clear separation of these two cell types was restored (Fig.3C, F). We further examined the robustness of scDaPars against varying dropout rates. Even though the accuracy of dynamic APA events identified by scDaPars decreased as the dropout rate increased, scDaPars could still achieve > 0.75 area under the receiver operating characteristics (ROC) curve when the proportion of dropout events was as high as 70% (Supplemental Fig.S3).

scDaPars outperforms existing methods by providing single-cell-resolution APA quantification applicable to both 3' end and full-length scRNA-seq data

Several bioinformatics tools have been developed to analyze APA usage using scRNA-seq data (i.e., scDAPA (Ye et al. 2020) and scAPA (Shulman and Elkon 2019)), yet, unlike scDaPars, they were not designed to quantify APA usage at the single-cell resolution. During the preparation of this manuscript, we noticed another method Sierra (Patrick et al. 2020), which detects differential transcript usage in scRNA-seq data, may also be used for quantifying dynamic APA events. To illustrate the superiority of scDaPars over these existing methods, we applied scDaPars, scAPA and Sierra to a benchmark 10x Chromium dataset containing 902 single cells from three lung adenocarcinoma cell lines (Tian et al. 2019) (see "Methods"). scDAPA was excluded from this study since it identifies APA events by pair-wise comparison without quantifying APA usage. scDaPars outperformed both scAPA and Sierra by generating clear and compact cell clusters according to annotated cell lines (UMAP (McInnes et al. 2018) visualization in Supplemental Fig.S4A, B and C). We used silhouette analysis to quantitatively assess the resulting clusters. Compared with scAPA and Sierra, scDaPars showed higher silhouette coefficients which indicated the clustering results from scDaPars are more congruent with the true cell-line labels (Supplemental Fig.S4D, E and F). To further benchmark scDaPars in more complex biological systems, we applied scDaPars, scAPA and Sierra to an immune dataset containing 3362 PBMCs (Ding et al. 2020) (see "Methods"). Again, the APA usage quantified by scDaPars generated compact and accurate immune cell clusters (Fig.4A, D). In contrast, although Sierra outperformed scAPA and was able to separate B cell and CD14+ monocytes (Fig.4B, C), both Sierra and scAPA failed to accurately distinguish the five immune cell types (Fig.4E, F). Besides generating accurate cell clusters, scDaPars also identified 169 dynamic APA genes (genes with differential poly(A) site usage) among the five immune cell types, most of which (96%) were unseen by existing methods. For example, scDaPars identified EIF1 as a dynamic APA gene between B cells and CD14+ monocytes. Both cluster- and single-cell level coverage plots corroborated that EIF1 exhibits 3' UTR lengthening in B cells compared to CD14+ monocytes (Supplemental Fig.S5). Yet, EIF1 was not captured by previous methods (i.e., scAPA), indicating the advantage of scDaPars. More importantly, scDAPA, scAPA and Sierra rely on peak calling using 3' end enriched reads in 10x Chromium to quantify APA usage and thus are not applicable to data generated by full-length sequencing protocols like Smart-seq2 which do not contain enriched peaks in the 3' UTR regions (Picelli et al. 2013).

scDaPars revealed intrinsic tumor APA variations and immune cell subpopulations in primary breast cancer

Global-scale coordinated APA events are commonly observed in cancers (Xia et al. 2014), and APA induced 3' UTR shortening was shown to be associated with tumor aggressiveness and poor survival of cancer patients (Lembo et al. 2012; Xia et al. 2014). However, knowledge of APA regulations in cancer has been largely derived from bulk RNA-seq studies. Therefore, while global APA variations between tumor and normal cells have been well characterized, little is known about the intertumoral APA heterogeneity at the single-cell resolution. To illustrate scDaPars' capacity of characterizing single-cell APA variations in cancers, we applied scDaPars to a Smart-seq2 (Picelli et al. 2013) scRNA-seq dataset containing 563 single cells from 11 breast cancer patients (Chung et al. 2017). In consistent with bulk results, 3' UTRs were shortened in tumor cells compared to normal cells ($P < 2.2\times10^{-16}$) (Fig.5A). Even PDUI values before scDaPars imputation could separate tumor cells from non-tumor cells with effectiveness comparable to that of gene expression values (Supplemental Fig.S6A), suggesting an important role of dynamic APA events in breast cancer progression. As expected, scDaPars imputed APA profiles showed a better separation between tumor and non-tumor groups (Fig.5B, Supplemental Fig.S7).

To further elucidate APA variations among cell subgroups, we analyzed APA profiles of tumor and non-tumor cells separately. On the one hand, contrary to a previous single-cell APA analysis performed on aggregated "meta-genes" in the same breast cancer dataset (Chung et al. 2017), which showed that no differences in APA were associated with cancer subtypes or patients (Kim et al. 2019), we found that tumor cells were not only separated into patient-specific clusters based on scDaPars-imputed APA profiles (Fig.5C), but also further classified into different molecular subtypes (Supplemental Fig.S8), showing evidence of both intertumoral and cancer-subtype-specific APA heterogeneity as well as scDaPars's advantage over existing method. On the other hand, non-tumor cells, which were derived from the same group of patients as tumor cells, were clustered mainly according to their cell types (B cells, Myeloid cells and T cells) instead of patients (Fig.5D, Supplemental Fig.S6B). This result not only reaffirmed that dynamic APA events are cell-type specific characteristics of immune cells, but also indicated that the patient-specific APA profiles observed in tumor cells were unlikely due to batch effects in patient samples but rather reflected true intertumoral variations in APA.

In addition, in consistent with prior knowledge of two B cell subclasses (proliferating and naive/memory B cells) in this dataset, we observed that B cells were classified into two cell subgroups based on scDaPars-imputed APA profiles (Fig.5E) with group 2 B cells showed global 3' UTR shortening compared with group 1 B cells ($P = 2\times10^{-3}$) (Fig.5F). We found that most B cell proliferation signature genes from the literature (Chung et al. 2017) were upregulated in group 2 B cells compared to group 1 B cells (Supplemental Fig.S9, Supplemental Table S2), suggesting that group 2 B cells may represent proliferating B cells. Indeed, the proliferating and naive/memory B cells determined by the expression of B cell proliferating marker genes are highly congruent with scDaPars derived cell subgroups (Supplemental Fig.S10A, B). These results are also in line with previous reports that proliferating cells (i.e., group 2 cells) express more isoforms with shortened 3' UTRs through APA (Sandberg et al. 2008). However, expression analysis of all genes failed to identify these B cell subgroups (Supplemental Fig.S10C), revealing the potential benefits of APA analysis in delineating cell subpopulations. In summary, scDaPars improves the characterization of APA variations and cell subpopulations in single cells.

scDaPars enables identification of novel cell subpopulations invisible to conventional gene expression analysis in endoderm differentiation

As APA patterns appear to be globally regulated in cell differentiation (Ji et al. 2009; Tian and Manley 2017) (i.e., decreased proximal poly(A) site usage in more differentiated states of embryonic development), we hypothesized that they could provide a new aspect to identify cell subpopulations during differentiation. To test this hypothesis, we applied scDaPars to a time-course Smart-seq2 (Picelli et al. 2013) scRNA-seq dataset containing 758 cells sequenced at 0, 12, 24, 36, 72 and 96 h of differentiation during human definitive endoderm (DE) emergence (Chu et al. 2016). scDaPars revealed clear and accurate cell clusters from each time point along the differentiation process (Fig.6A). Dimension 2 of the UMAP projection of raw PDUI values reconstructed single-cell orders matching the true differentiation time points, reflecting the global APA dynamics during cell differentiation (Supplemental Fig.S11).

Next, we investigated whether APA could help delineate novel cell subpopulations invisible to gene expression analysis alone. Imputation based on observed gene expression has been shown to enhance the identification of cell subpopulations (Li and Li 2018). Therefore, to ensure APA is providing additional information beyond expression, we first recovered plausible single-cell gene expression data using scImpute (Li and Li 2018), a state-of-the-art gene expression imputation method. Notably, although the imputed gene expression profile outputs more compact clusters than the raw expression, single cells collected from 72 and 96 h of differentiation were still largely overlapped (Supplemental Fig.S12). To characterize additional cellular heterogeneity, we integrated APA information with imputed gene expression using similarity network fusion (SNF) (Wang et al. 2014). By creating and converging separate similarity networks for APA and gene expression, SNF reduced noisy inter-cluster similarities among cells in 12 and 24 h of differentiation and enhanced intra-cluster similarities observed in one or both similarity networks (Fig.6B). We then quantitatively compared the clustering results by using spectral clustering algorithm (Ng et al. 2002) on different similarity networks with the number of clusters $k = 6$. The clustering results are evaluated by normalized mutual information (NMI) (Witten et al. 2016) where $NMI = 1$ indicates a perfect match between the clustering results and the known differentiation time points. While gene expression imputation increased NMI from 0.76 to 0.85, integration of APA usages with imputed gene expression further increased NMI from 0.85 to 0.89, suggesting the benefits of adding APA information.

Besides unifying the clustering results of APA and gene expression, the fused similarity network also revealed novel and potentially meaningful subpopulations. For example, cells at 96 h of differentiation were divided into two previously unidentified subpopulations (Fig.6B). Through analyzing APA and gene expression between the two subpopulations, we found that APA usage alone can accurately separate the two subpopulations (Fig.6C, Supplemental Fig.S13) and subpopulation 2, which was more distinct from cells in 72 h of differentiation than subpopulation 1, exhibited global 3' UTR lengthening compared to subpopulation 1 ($P = 3.64\times10^{-8}$) (Fig.6D); whereas the imputed gene expression profile alone failed to distinguish the two subpopulations (Fig.6C). The APA profile quantified by DaPars also failed to identify the 2 subgroups (Supplemental Fig.S14), indicating the superiority of scDaPars.

Since subpopulation 2 showed global 3' UTR lengthening, we hypothesized it may represent a more differentiated cell subgroup. To test our hypothesis, we performed differential gene expression analysis between subpopulation 1 and 2 using DESeq2 (Love et al. 2014). As a result, subpopulation 2 was characterized by higher expression of endoderm development marker genes including *GATA6*, *EOMES*, and

SOX17 (Chu et al. 2016) (Fig.6F, Supplemental Table S3). In addition, the transcriptional profile of subpopulation 2 also included significantly upregulated endoderm development related genes like *LHX1*, which is important for renal development (Reidy and Rosenblum 2009), and *HMGA2*, which is required for epithelium differentiation during embryonic lung development (Singh et al. 2014), suggesting subpopulation 2 has a more differentiated phenotype than subpopulation 1. To further elucidate the global biological differences between the two subpopulations, we performed gene oncology (GO) analysis (Luo et al. 2009). We found that several endoderm development related GO terms were highly enriched in the upregulated genes in subpopulation 2 (Fig.6E). Furthermore, using the expression of differential APA genes, we were able to separate the two subpopulations (Supplemental Fig.S15), indicating that some biologically meaningful subpopulations were masked by overall gene expression analysis. Finally, we conducted a trajectory analysis by STREAM (Chen et al. 2019) to independently show the validity of the identified subpopulations. Using cells at 0 h of differentiation as a natural starting point (root), we found that most cells are projected onto the inferred branches according to their corresponding differentiation time points (Supplemental Fig.S16A, B), and the derived pseudotime progression corroborated that cells in subpopulation 2 are more differentiated than those in subpopulation 1 (Fig.6G, Supplemental Fig.S16C). Considered collectively, scDaPars calculated APA usage offered an additional layer of information in characterizing cellular heterogeneity that was otherwise invisible in gene expression analysis.

Discussion

Here, we developed scDaPars, a novel bioinformatics algorithm to *de novo* identify and quantify single-cell dynamic APA events using standard scRNA-seq data. Many methods have been developed to measure the relative APA usages in RNA-seq data from bulk samples (Xia et al. 2014). However, the widespread dropout events in scRNA-seq data impede these bulk-sample based methods to quantify APA usage among single cells (Figs.2D and 2E). To address this technical challenge in scRNA-seq, scDaPars first quantifies raw APA usage based on the two-poly(A)-site model introduced in DaPars (Xia et al. 2014). Since APA exhibits a cell-type specific pattern (Velten et al. 2015; Kim et al. 2019), scDaPars then clusters cells into different cell neighbors based on their calculated raw APA profiles. Next, scDaPars imputes missing APA usage by borrowing APA information of the same gene from neighboring cells. Benchmarking on both real and simulated data show the accuracy of scDaPars in predicting poly(A) sites, the ability in recovering missing APA usages, and the robustness in identifying dynamic APA events across different cell types (Fig.2 and 3).

Previously, methods for analyzing APA usage using scRNA-seq data mostly address the high technical noise in scRNA-seq by creating pseudo-bulk RNA-seq data (i.e. pooled reads from cells that are assigned to the same cell cluster) (Shulman and Elkon 2019; Ye et al. 2020). Unlike scDaPars, even though these methods perform on scRNA-seq data, they do not quantify APA usage at the single-cell resolution but rather measure cell-cluster APA usage, which contradicts the purpose of single-cell sequencing (Supplemental Table S1). Additionally, previous methods are confined by cell cluster assignments determined by conventional gene expression analysis. In contrast, scDaPars quantifies single-cell APA usage independent of gene expression, which provides an additional layer of APA information that helps identify hidden cell states. (Fig.6C).

Finally, unlike existing methods, we expect scDaPars to be widely applicable to any scRNA-seq datasets. While the main analysis presented in this paper builds on scRNA-seq data generated by low-throughput Smart-seq2 (Picelli et al. 2013) protocol and the accuracy of scDaPars decreases as the dropout rate increases (Supplemental Fig.S3), scDaPars can also be applied to datasets generated by high-throughput high-dropout-rate droplet-based methods, e.g. 10x Chromium (Zheng et al. 2017). For example, scDaPars successfully revealed cell-type specific APA patterns in 3362 PBMCs sequenced by 10x Chromium (Ding et al. 2020) (Fig.4A). Together, scDaPars provides an additional layer of APA information that helps identify cell subpopulations invisible to conventional gene expression analysis.

Methods

De novo quantification of dynamic APA events

scDaPars first performs *de novo* identification and quantification of dynamic APA events based on the two-poly(A)-site model introduced in DaPars. The bedGraph files for each single cell were used as input and jointly analyzed to calculate the APA usage measured as the Percentage of Distal poly(A) site Usage Index (PDUI). For each gene, the distal poly(A) site was identified as the end point of the longest 3' UTR among all scRNA-seq samples, and the proximal poly(A) site was inferred by optimizing the following linear regression model:

$$\left(\overline{W_L^{1,2,3,\dots,m}}, \overline{W_S^{1,2,3,\dots,m}}, \bar{P}\right) = \underset{W_L^{1,2,3,\dots,m}, W_S^{1,2,3,\dots,m} \geq 0,\ 1<P<L}{argmin} \sum_{i=1}^{m} \left\| C_i - \left(W_L^i I_L + W_S^i I_P \right) \right\|_2^2 \qquad (1)$$

where W_L^i and W_S^i are the abundances of transcripts with distal and proximal poly(A) sites for cell i, C_i is the read coverage of cell i normalized by total sequencing depth, L is the length of the longest 3' UTR, P is the length of the alternative proximal 3' UTR to be inferred, I_L and I_P are two indicator functions for long and short 3' UTRs such that $I_L = \underset{L}{[1, \cdots, 1]}$ and $I_P = \underset{P, \qquad L-P}{[1, \cdots, 1, 0, \cdots, 0]}$. The optimal proximal poly(A) site is selected by minimizing the deviation between the observed read density C_i and the expected read density $W_L^i I_L + W_S^i I_P$ in all single cells. The APA usage is then quantified as PDUI for each gene in each single cell, with PDUI defined as:

$$PDUI_i = \frac{W_L^{i*}}{} \qquad (2)$$

where W_L^{i*} and W_S^{i*} are the optimal expression levels of transcripts with distal and proximal poly(A) site for cell i. The smaller the PDUI is, the less distal poly(A) site is used, the shorter the 3' UTRs. The final output is a PDUI matrix in which rows represent genes and columns represent cells. Additionally, PDUIs can only be calculated in this step for genes with sufficient read coverage (default coverage of 5 reads per base), which automatically separate genes into robust genes and dropout genes for future analysis. On average, 50% of the genes in a cell are robust genes after quality control and if the dropout rate in the dataset is higher (e.g., in 10x Chromium datasets), the average number of robust genes in the data will decrease. There are overlaps between robust genes of different cells: in the benchmark dataset in Figure 2, the overlap of robust genes between any two cells is ~40%.

Detection of potential neighboring cells and outliers

Since APA exhibits alterations in different cell types and cell states in a global scale, scDaPars recovers missing single-cell level APA usage by borrowing APA information of the same gene from neighboring cells. A critical step here is to determine which cells are from the same cell subpopulation and therefore are neighboring cells. Instead of using observed gene expression, scDaPars uses raw APA usage for this task because (1) APA is a feature intrinsic to cell types or cell states; (2) scDaPars quantifies APA usage independent of gene expression. We first performed a quantitative comparison of clustering using raw APA usage and observed gene expression from the hESC dataset in Figure 6 (Supplemental Fig.S17). We found that clustering of raw APA usage outperformed that of observed gene expression (Supplemental Fig.S17C, D) partly because differentiation is one of the biological processes with the most dramatic APA changes. To further illustrate the benefits of quantifying APA independent of gene expression, we modified our original scDaPars algorithm so that the initial clustering is performed using observed gene expression instead of raw APA usage and re-quantified the APA usage of cells from the hESC dataset in Figure 6. We found that the two subpopulations identified by original scDaPars were obscured by the modified version (Supplemental Fig.S18), indicating the advantage of quantifying APA independent of gene expression.

Due to the technical limitation of scRNA-seq data, it is unlikely to completely cluster cells into true subpopulations based on the sparse PDUI matrix generated in last step. Instead, the goal of this step is to determine a set of potential neighboring cells which scDaPars will fine-tune in the following imputation step.

To increase the robustness and reliability of the clustering results and to find more plausible neighboring cells, scDaPars applies principal component analysis (PCA) to the raw PDUI matrix. While the PDUI matrix is sparse, the modularity of dynamic APA provides redundancy in gene dimensions, which can be exploited. Therefore, scDaPars selects principal components (PCs) that can together explain at least 40% of the variance in the data. Note that the neighboring cells are identified in these PCA dimensions while the imputation is performed on the full PDUI matrix.

$$PDUI_{pca} = pca(PDUI, 0.4) \qquad (3)$$

Next, scDaPars identifies and removes outlier cells from the analysis. The outlier cells may be the result of technical errors or may represent true rare biological variations, in either case, scDaPars will not use these outlier cells to impute missing APA usage in other cells. We calculate the distance matrix $D_{N\times N}$ between cells based on the PCA transformed data $PDUI_{pca}$. For each cell m, we define the Euclidean distance of cell m to its nearest neighbor as d_m, resulting a set $\boldsymbol{d} = \{d_1, \cdots, d_N\}$. We denote the first quantile of $\boldsymbol{d}$ as Q_1 and its third quantile as Q_3 and the distance between Q_1 and Q_3 as interquartile range IQR. The outlier cells are defined as cells which are separated by more than 1.5 IQR to the third quantile Q_3.

$$Outlier = \{m : d_m > Q_3 + 1.5 \times IQR\} \qquad (4)$$

The remaining non-outlier cells $\{1, \cdots, N\} \backslash Outlier$ are then clustered into subpopulations using graph-based community detection algorithm. The single cells are the vertices in the graph, and community detection in graphs will identify groups of vertices with high probability of being connected to each other than to members of other groups. We use R package *RANN* with default parameters to first identify the approximate nearest neighbors and convert neighbor relation matrix into an adjacency matrix. We then use *igraph (Csardi and Nepusz 2006)* to represent the resulting adjacency matrix as a graph and apply *walkstrap (Pons and Latapy 2005) algorithm* to identify communities of vertices (cells) that are densely connected. Suppose scDaPars divides cells into K subpopulations in this step, for each cell m, its potential neighboring cells N_m are the other cells in the same cell subpopulation k.

$$N_m = \{i \in k, i \neq m\} \qquad (5)$$

Imputation of missing APA usage

After potential neighboring cells N_m for each cell are determined, we impute APA usage cell by cell. Recall that PDUIs can only be estimated for genes with sufficient read coverage, scDaPars thereby automatically separates genes into robust genes and dropout genes when calculating the PDUI matrix. Here, we denote the set of robust genes for cell m as R_m and the set of dropout genes that will be imputed in this step as D_m. scDaPars then learns the cells' similarities through the robust gene set $G_{Robust,m}$ and impute the APA usage of D_m by borrowing information from the same gene's APA usage in other neighboring cells learned from R_m . To fine-tune the grouping of neighboring cells from N_m, we use non-negative least squares (NNLS) regression:

$$\overline{\theta_m} = argmin_{\theta_m} \left\| PDUI_{R_m\,,\,m} \; - PDUI_{R_m,\,N_m} \theta_m \right\|_2^2, \; \theta_m > 0 \qquad (6)$$

where N_m represents the indices of cells that are potential neighboring cells of cell m, $PDUI_{Gene_{robust},\,m}$ is a vector of response variables representing R_m rows in the m-th column (cell m) of the original PDUI matrix, $PDUI_{R_m,\,N_m}$ is a sub-matrix of the original PDUI matrix with dimensions $|R_m| \times |N_m|$. The goal is to find the optimal coefficients $\overline{\theta_m}$ of length $|N_m|$ that can minimize the deviation between APA usage of R_m in cell m and those in potential neighboring cells. The advantage of using NNLS is that it has the property of leading to a sparse estimate of θ_m, whose components may have exact zeros, so that true neighboring cells of cell m are conveniently selected from N_m. Once $\overline{\theta_m}$ is computed, we have a vector of weighted neighbors associated with each cell in our data. scDaPars use this coefficient $\overline{\theta_m}$ estimated from the set R_m to impute the APA usage of genes in the set D_m in cell m. All of the above analyses are conducted in *R* (R Core Team 2020).

$$\overline{PDUI_{g,m}} = \begin{cases} PDUI_{g,m}, & \text{if } g \in R_m \\ PDUI_{g,N_m} \cdot \overline{\theta_m}, & \text{if } g \in D_m \end{cases} \qquad (7)$$

Differential percentage of distal APA usage index (PDUI) (Dynamic APA events)

We used the following two criteria to define the significant dynamic APA events: first, given the PDUI values for cells in two cell types, the Benjamini-Hochberg corrected Mann-Whitney *U* p-value between two cell types (FDR) is less than 0.05; second, the absolute difference of mean PDUIs in cell type 1 and cell type 2 is greater than 0.2.

$$\begin{cases} FDR \leq 0.05 \\ \left|PDUI_{cell\ type\ 1} - PDUI_{cell\ type\ 2}\right| \geq 0.2 \end{cases} \qquad (8)$$

Preprocessing of scRNA-seq data

The scRNA-seq datasets used in this manuscript are all publicly available and are summarized in Supplemental Table S4. The 2 single-cell PBMC data are available at the Gene Expression Omnibus (GEO) under accession code GSE132044. The breast cancer data are available at GEO under accession code GSE75688. The time-course definitive endoderm data are available at GEO under accession code GSE75748. The lung adenocarcinoma cell line data are available at GEO under accession code GSE118767. The DICE immune data used to generate synthetic dataset were obtained from dbGaP under study accession code phs001703.v1.p1. For low-throughput datasets generated by Smart-seq2 (Picelli et al. 2013) protocol, we downloaded the publicly available FASTQ files from GEO database and aligned the reads using STAR 2.5.2 (Dobin et al. 2013) with default parameters, generating one BAM file for each single cell. For high-throughput datasets generated by 10x Chromium (Zheng et al. 2017), we downloaded the FASTQ files and aligned the reads using Cell Ranger 3.0.2. We then selected reads with correct unique molecular identifier (UMI) using Drop-seq tools *FilterBAM (Macosko et al. 2015)* and remove reads with duplicated UMIs using UMI-tools *dedup* (Smith et al. 2017). We next merged reads originated from same cells together and generated one BAM file for each single cell. The BAM files are used as inputs for subsequent scDaPars analysis. The average dropout rate (Percentage of missing data) for Smart-seq2 datasets is ~50% in our study. The 10x Chromium dataset in our study has a dropout rate of ~65%.

Generation of synthetic dataset

The synthetic dataset was created based on bulk RNA-seq data generated from 13 immune cell types (Schmiedel et al. 2018). The different immune cell types are isolated so that each sample only contains cells from one cell type. We used DaPars to estimate the APA usage in these bulk samples and generated an APA matrix, in which rows represent genes and columns represent samples. Since widespread dynamic APA events were reported in Naïve and activated CD4 T cells, we selected only samples that belong to these two cell types for the following simulation.

We down-sampled the resulting bulk APA matrix to emulate the APA profiles generated from single-cell data. We first calculated the dropout rate for each gene in the benchmark immune dataset (Ding et al. 2020). Next, for each gene in the bulk APA matrix, the dropout rate is randomly selected from the set of real dropout rates with replacement. Finally, we used Bernoulli distribution with p equals to the selected dropout rate and n equals to the number of samples to introduce dropouts into the synthetic dataset. The final dropout introduced data has a ~50% dropout rate which is similar to the dropout rate of real datasets. Notice that the generation of the synthetic dataset is independent from the models of scDaPars, so that it can be used to evaluate scDaPars in a fair way.

Benchmark comparison of scDaPars

To illustrate the advantage of scDaPars, we applied scDaPars, scAPA and Sierra to two benchmark 10x Chromium datasets. scAPA measures differential usage of poly(A) sites between different cell types by the proximal poly(A) site usage index (proximal PUI). Since we want to test scAPA's ability for quantifying single-cell-level APA usage, we input single-cell coverage into scAPA to generate a cell by transcript proximal PUIs matrix to perform the clustering analysis. The Sierra pipeline does not yield PDUI like measurements. Instead, it generates a peak count matrix in which peak coordinates are annotated according to the genomic features they fall on including UTRs, exons, or introns. In order to calculate APA usage from the peak count matrix, we first selected peaks falling on the 3' UTRs and only kept transcripts with more than one peak. We then transferred the peak count matrix into an APA matrix by calculating the relative usage of the most distal peak. The resulting APA matrix were used for the clustering analysis. Finally, we performed silhouette analysis by *silhouette ()* in R package *cluster* v2.1.0. to quantitatively evaluate the clustering accuracy of the three methods.

Software Availability

The source codes and the R package scDaPars are available as Supplemental Code. scDaPars is also freely available at GitHub (https://github.com/YiPeng-Gao/scDaPars).

Acknowledgements

We thank Yikai Luo, Dr. Joel Neilson at Baylor College of Medicine, members of the Li lab at University of California, Irvine, and Dr. Jingyi Jessica Li at University of California, Los Angeles for insightful discussions. We thank Dr. Chen Chao at Baylor College of Medicine for his suggestions on this manuscript. This work is supported by US National Institutes of Health (NIH) grants R01HG007538, R01CA193466 and R01CA228140 to W.L. and the Cancer Prevention Research Institute of Texas (CPRIT) grant RR170048 to C.A. C.A. is a CPRIT research scholar.

Author Contributions

W. L. conceived and supervised the project. Y.G. performed the data analysis. Y.G., L.L., W.L. interpreted the data. Y.G., L.L., W.L., C.A. wrote the manuscript.

Disclosure Declaration

The authors declare no competing financial interests.

Figure Legends

Figure 1. A schematic illustration of the scDaPars algorithm.
(I) scDaPars predicts both distal and proximal poly(A) sites by joint analysis of all single-cell samples and quantifies the raw relative APA usage by the proportion of estimated abundances of transcripts with distal poly(A) sites (long isoform). (II) scDaPars determines potential neighboring cells by applying community detection methods in APA profiles generated in step(I). (III) scDaPars uses NNLS regression model to refine neighboring cells and impute missing values by borrowing APA information from neighboring cells.

Figure 2. Evaluation of APA detection accuracy of scDaPars using human PBMCs datasets.
(A) Fraction of poly(A) sites predicted in matched bulk RNA-seq data recovered in single cells using scDaPars or random control. Poly(A) sites predicted in scRNA-seq are considered true if they are located within cutoff distance from the bulk results. The cutoffs range from 0 to 100bp with 10bp increment.

(B) Percentage of scDaPars predicted poly(A) sites or random control overlapped with annotated poly(A) sites from RefSeq, Ensembl, UCSC gene models and poly(A)_DB. The confidence interval was derived by taking random sites 10 times.
(C) The top-scoring signal identified by de novo motif analysis (DREME) from the upstream (-100bp) of scDaPars predicted poly(A) sites from single cells.
(D) Boxplot showing Pearson's correlations between PDUI values of B-cell pairs estimated by DaPars and scDaPars (Wilcoxon test $P < 2.2\times10^{-16}$).
(E) Scatter plots of PDUI values between average of all single cells and bulk results estimated by DaPars (left) and scDaPars (right). Red line represents the theoretical linear relationships between bulk and average of all single-cell PDUIs, and blue represents the actual linear relationships estimated from data.

Figure 3. Evaluation of scDaPars in identifying dynamic APA events between two cell types using naive and activated CD4 T cells.
(A) – (C) Scatterplots showing UMAP results of 54 naive CD4 T cells and 31 activated CD4 T cells based on (A) Reference APA profiles or (B) Dropout events introduced APA profiles or (C) scDaPars corrected APA profiles.
(D) – (F) Heatmaps showing APA profiles of 136 differential APA genes (FDR <= 0.05 and PDUI differences >= 0.2) in the (D) reference data (E) dropout events introduced data and (F) scDaPars corrected data. Rows represent differential APA genes and columns represent cells. 88 out of 136 differential APA genes have shorter 3' UTRs in activated CD4 T cells in the reference data.

Figure 4. scDaPars outperforms existing methods by quantifying APA usage in single-cell resolution.
(A) – (C) Scatterplots showing UMAP results of 3362 PBMCs based on (A) scDaPars quantified APA usage or (B) scAPA quantified APA usage or (C) Sierra quantified APA usage.
(D) – (F) Silhouette plots for clustering results from (D) scDaPars, (E) scAPA and (F) Sierra. The x-axis represents cells and y-axis is the corresponding silhouette coefficient Si for each cell. The silhouette coefficient measures how similar a cell is to its own cluster compared to other clusters, therefore a higher silhouette coefficient indicates a better clustering result and a negative coefficient may suggest the cell is assigned to the wrong cluster. The red dashed line is the average Si for all cells.

Figure 5. scDaPars reveals tumor-specific and immune-cell-type specific APA landscape in primary breast cancer.
(A) Scatter plot of PDUI values in Tumor and Normal cells. For each gene, the mean PDUI values in tumor cells (y-axis) are plotted against that in normal cells (x-axis). Genes with shortened or lengthened 3' UTR (FDR <= 0.05 and PDUI difference >= 0.2)

in tumor cells are shown in red and blue. Bar plot shows the number of shortening genes or lengthening genes in tumor cells and p-value is calculated using single-tailed binomial test.

(B) Scatter plot gives UMAP results calculated from scDaPars restored APA profiles. Each dot represents a cell, and cells are labeled based on cell index provided in the original publication.

(C) Scatter plot of UMAP results of tumor cells. Cells are labeled by patient information.

(D) Scatter plot of UMAP results of immune cells. Cells are labeled by cell type information.

(E) Scatter plot of UMAP results of B cells based on scDaPars results.

(F) Scatter plot of PDUI values in group 1 B cells and group 2 B cells. For each gene, the mean PDUI values in group 2 B cells (y-axis) are plotted against that in group 1 B cells (x-axis). Genes with shortened or lengthened 3' UTR (FDR <= 0.05 and PDUI difference >= 0.2) in group 2 B cells are shown in red and blue. Bar plot shows the number of shortening genes or lengthening genes in group 2 cells.

Figure 6. scDaPars helps identify novel cell subpopulations during human embryonic development.

(A) Scatter plot shows UMAP results of single cells based on scDaPars recovered APA profiles. Cells are labeled based on cell differentiation time points given in the original publication.

(B) Cell-by-cell similarities represented by similarity matrices generated by R package SNFtool.

(C) Scatter plots of UMAP results of cells in 96h of differentiation based on scDaPars results (left) and imputed gene expression (right). Cells are labeled by results from (B).

(D) Scatter plot shows mean PDUI values of genes in subpopulation 2 (x-axis) and sub-population 1 (y-axis). Genes with 3' UTR shortening and lengthening (FDR <= 0.05 and PDUI differences >= 0.2) in subpopulation 2 are labeled in blue and red respectively. Bar plot shows the number of genes with shortening or lengthening in subpopulation 2 and p-value is calculated using single-tailed binomial test.

(E) Selected GO terms enriched in the upregulated genes in subpopulation 2.

(F) Example gene expression levels in two subpopulations.

(G) Stream plot from STREAM which shows cell density along different trajectories at a given pseudotime.

References

An JJ, Gharami K, Liao GY, Woo NH, Lau AG, Vanevski F, Torre ER, Jones KR, Feng Y, Lu B et al. 2008. Distinct role of long 3' UTR BDNF mRNA in spine morphology and synaptic plasticity in hippocampal neurons. *Cell* **134**: 175-187.

Bailey TL. 2011. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**: 1653-1659.

Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC. 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods* **10**: 1093.

Chen H, Albergante L, Hsu JY, Lareau CA, Lo Bosco G, Guan J, Zhou S, Gorban AN, Bauer DE, Aryee MJ et al. 2019. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat Commun* **10**: 1903.

Chen W, Jia Q, Song Y, Fu H, Wei G, Ni T. 2017. Alternative Polyadenylation: Methods, Findings, and Impacts. *Genomics Proteomics Bioinformatics* **15**: 287-300.

Chu L-F, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, Choi J, Kendziorski C, Stewart R, Thomson JA. 2016. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology* **17**: 173.

Chung W, Eum HH, Lee H-O, Lee K-M, Lee H-B, Kim K-T, Ryu HS, Kim S, Lee JE, Park YH. 2017. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications* **8**: 15081.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal, complex systems* **1695**: 1-9.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**: 1173-1183.

Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, Hughes TK, Wadsworth MH, Burks T, Nguyen LT. 2020. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature biotechnology* **38**: 737-746.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Elkon R, Ugalde AP, Agami R. 2013. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat Rev Genet* **14**: 496-506.

Garneau NL, Wilusz J, Wilusz CJ. 2007. The highways and byways of mRNA decay. *Nat Rev Mol Cell Biol* **8**: 113-126.

Gruber AJ, Zavolan M. 2019. Alternative cleavage and polyadenylation in health and disease. *Nat Rev Genet* **20**: 599-614.

Hoffman Y, Bublik DR, Ugalde AP, Elkon R, Biniashvili T, Agami R, Oren M, Pilpel Y. 2016. 3’UTR shortening potentiates microRNA-based repression of pro-differentiation genes in proliferating human cells. *PLoS genetics* **12**: e1005879.

Ji Z, Lee JY, Pan Z, Jiang B, Tian B. 2009. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci U S A* **106**: 7028-7033.

Ji Z, Tian B. 2009. Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS One* **4**: e8419.

Kim N, Chung W, Eum HH, Lee H-O, Park W-Y. 2019. Alternative polyadenylation of single cells delineates cell types and serves as a prognostic marker in early stage breast cancer. *PloS one* **14**: e0217196.

Lembo A, Di Cunto F, Provero P. 2012. Shortening of 3′ UTRs correlates with poor prognosis in breast and lung cancer. *PloS one* **7**: e31129.

Li WV, Li JJ. 2018. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature communications* **9**: 997.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**: 550.

Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. 2009. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics* **10**: 161.

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**: 1202-1214.

Mayr C, Bartel DP. 2009. Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**: 673-684.

McInnes L, Healy J, Melville J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:180203426*.

Ng AY, Jordan MI, Weiss Y. 2002. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pp. 849-856.

Park HJ, Ji P, Kim S, Xia Z, Rodriguez B, Li L, Su J, Chen K, Masamha CP, Baillat D et al. 2018. 3' UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk. *Nat Genet* **50**: 783-789.

Patrick R, Humphreys DT, Janbandhu V, Oshlack A, Ho JW, Harvey RP, Lo KK. 2020. Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome biology* **21**: 1-27.

Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* **10**: 1096-1098.

Pons P, Latapy M. 2005. Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, pp. 284-293. Springer.

R Core Team. 2020. R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Reidy KJ, Rosenblum ND. 2009. Cell and molecular biology of kidney development. In *Seminars in nephrology*, Vol 29, pp. 321-337. Elsevier.

Saliba AE, Westermann AJ, Gorski SA, Vogel J. 2014. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* **42**: 8845-8860.

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3'untranslated regions and fewer microRNA target sites. *Science* **320**: 1643-1647.

Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, Ha B, Altay G, Greenbaum JA, McVicker G. 2018. Impact of genetic polymorphisms on human immune cell gene expression. *Cell* **175**: 1701-1715. e1716.

Shapiro E, Biezuner T, Linnarsson S. 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics* **14**: 618-630.

Shulman ED, Elkon R. 2019. Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic acids research* **47**: 10027-10039.

Singh I, Mehta A, Contreras A, Boettger T, Carraro G, Wheeler M, Cabrera-Fuentes HA, Bellusci S, Seeger W, Braun T. 2014. Hmga2 is required for canonical WNT signaling during lung development. *BMC biology* **12**: 21.

Smith T, Heger A, Sudbery I. 2017. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome research* **27**: 491-499.

Stegle O, Teichmann SA, Marioni JC. 2015. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16**: 133-145.

Tian B, Manley JL. 2017. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18**: 18-30.

Tian L, Dong X, Freytag S, Lê Cao K-A, Su S, JalalAbadi A, Amann-Zalcenstein D, Weber TS, Seidi A, Jabbari JS. 2019. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature methods* **16**: 479-487.

Velten L, Anders S, Pekowska A, Jarvelin AI, Huber W, Pelechano V, Steinmetz LM. 2015. Single-cell polyadenylation site mapping reveals 3' isoform choice variability. *Mol Syst Biol* **11**: 812.

Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A. 2014. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods* **11**: 333.

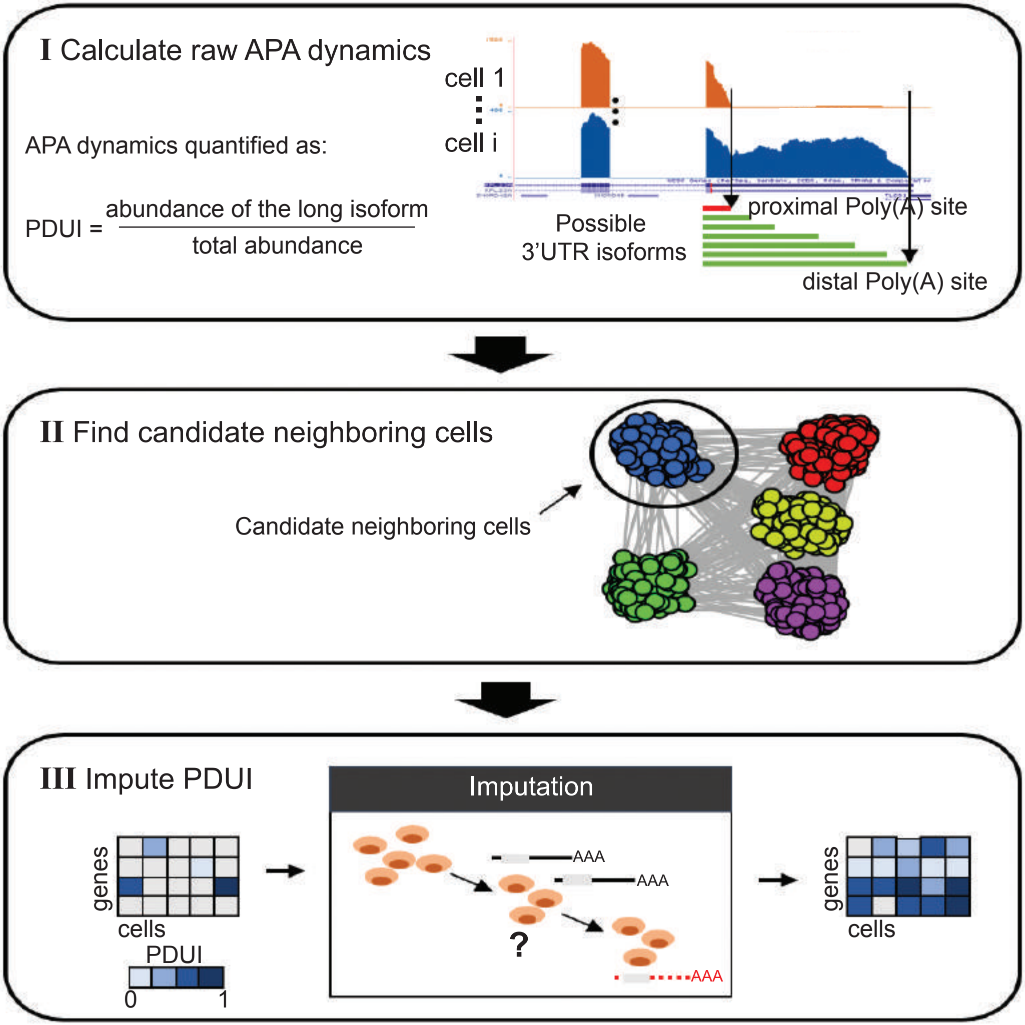
Wang R, Nambiar R, Zheng D, Tian B. 2017. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic acids research* **46**: D315-D319.

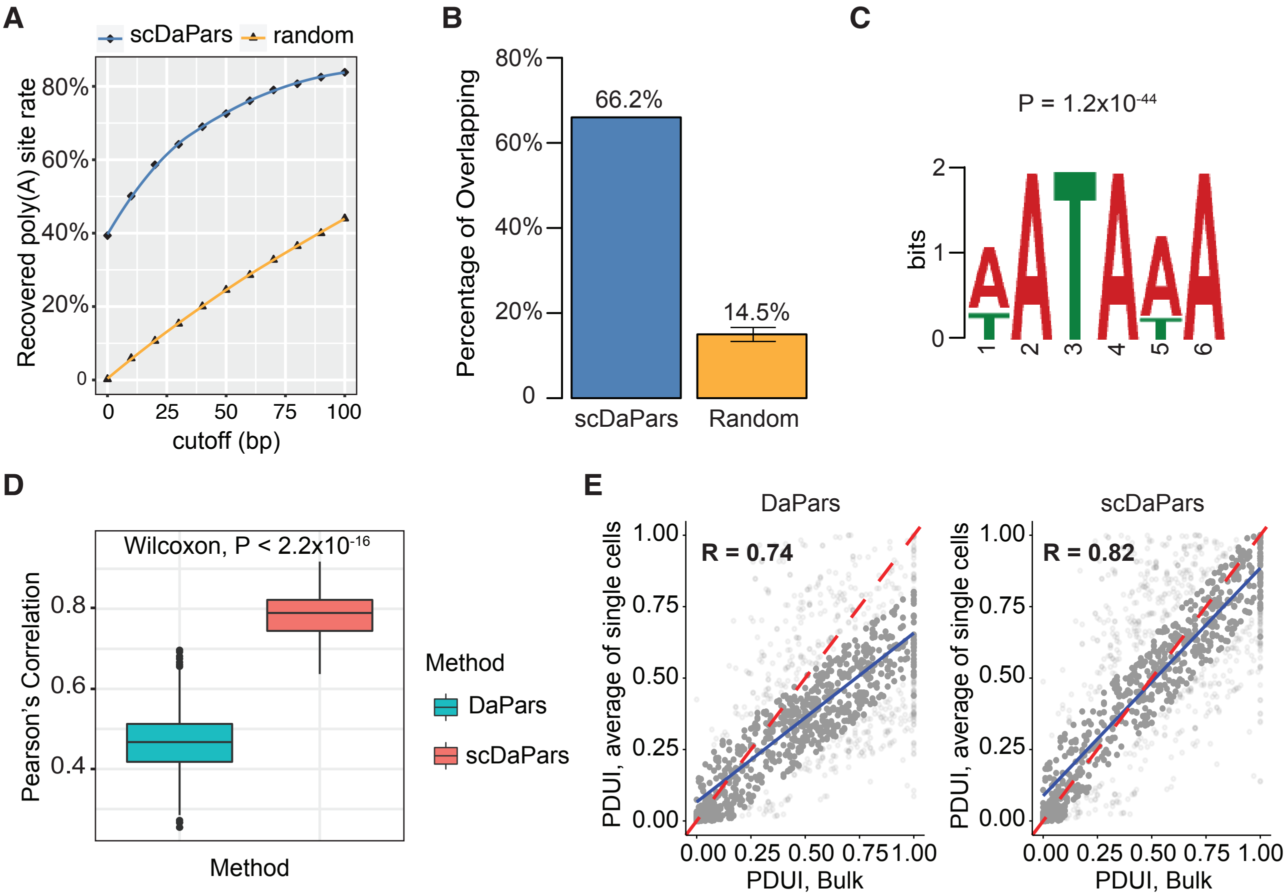
Witten IH, Frank E, Hall MA, Pal CJ. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

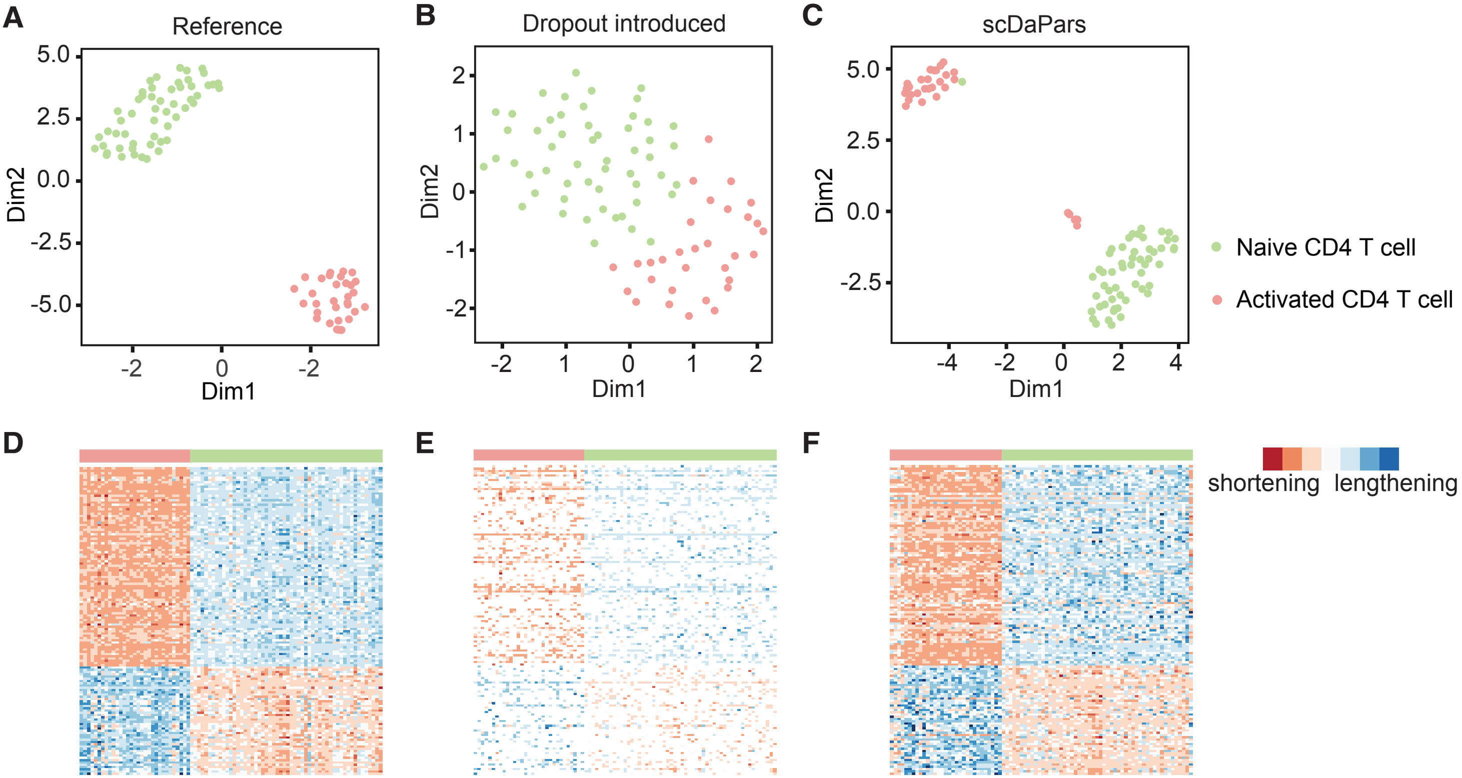
Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, Li W. 2014. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun* **5**: 5274.

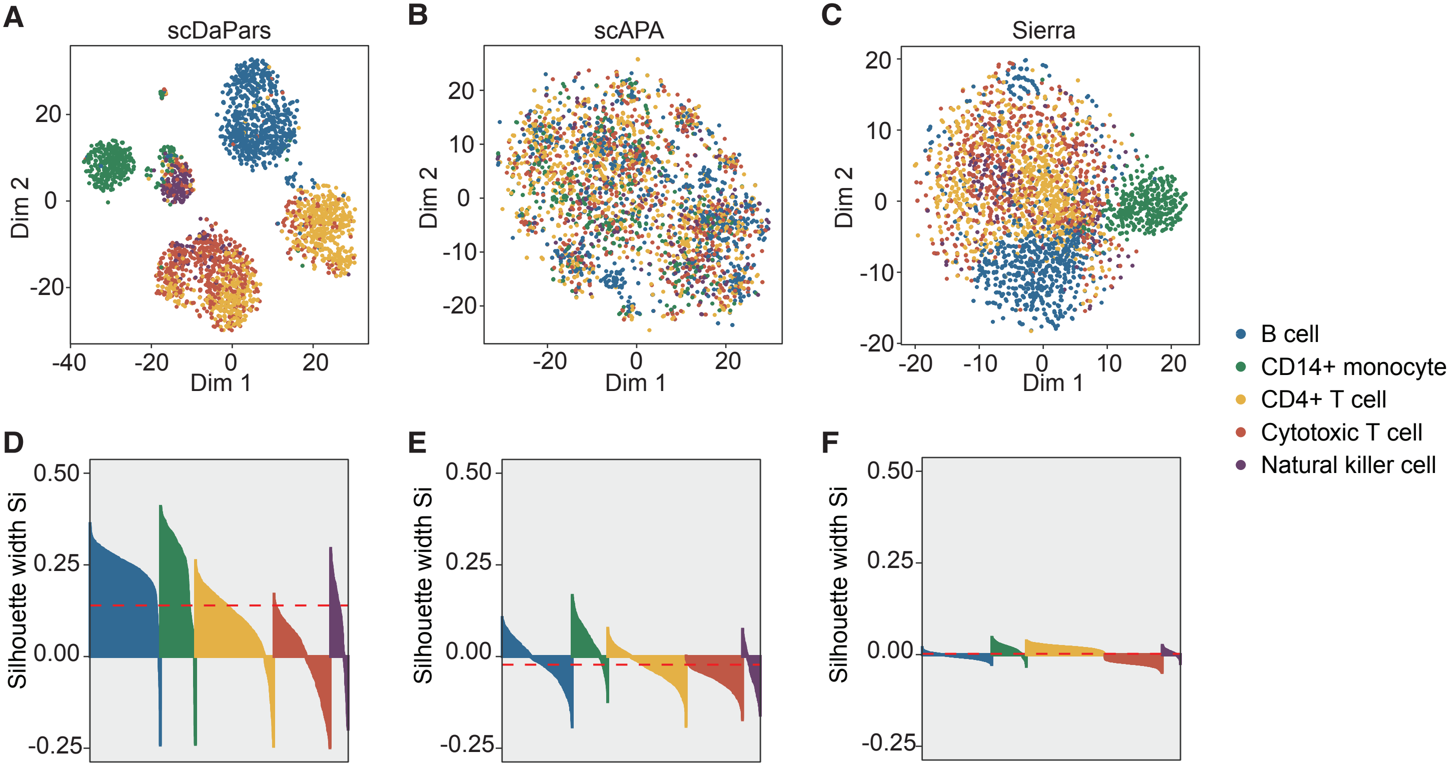
Ye C, Zhou Q, Wu X, Yu C, Ji G, Saban DR, Li QQ. 2020. scDAPA: detection and visualization of dynamic alternative polyadenylation from single cell RNA-seq data. *Bioinformatics* **36**: 1262-1264.

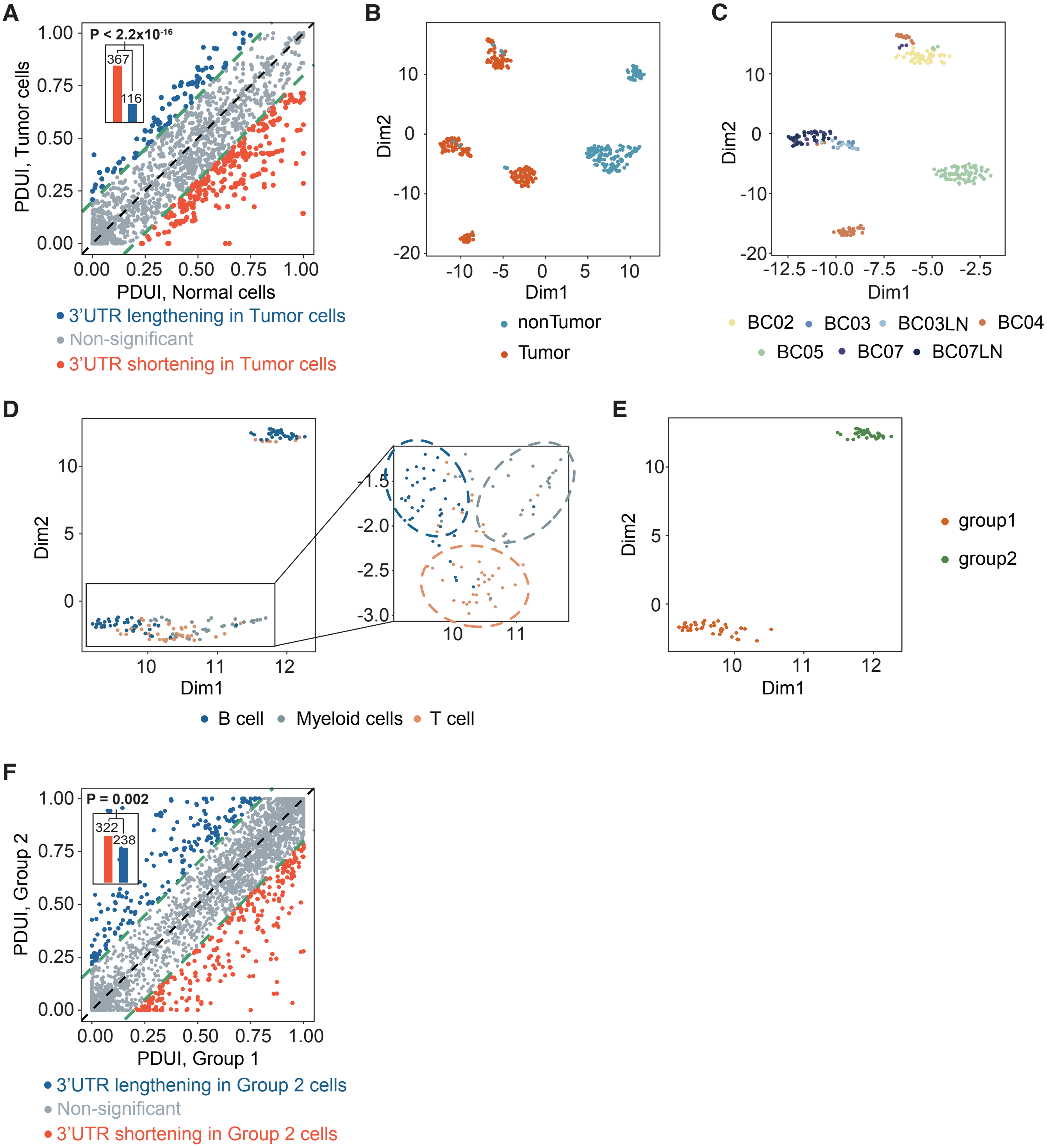
Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J. 2017. Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**: 14049.

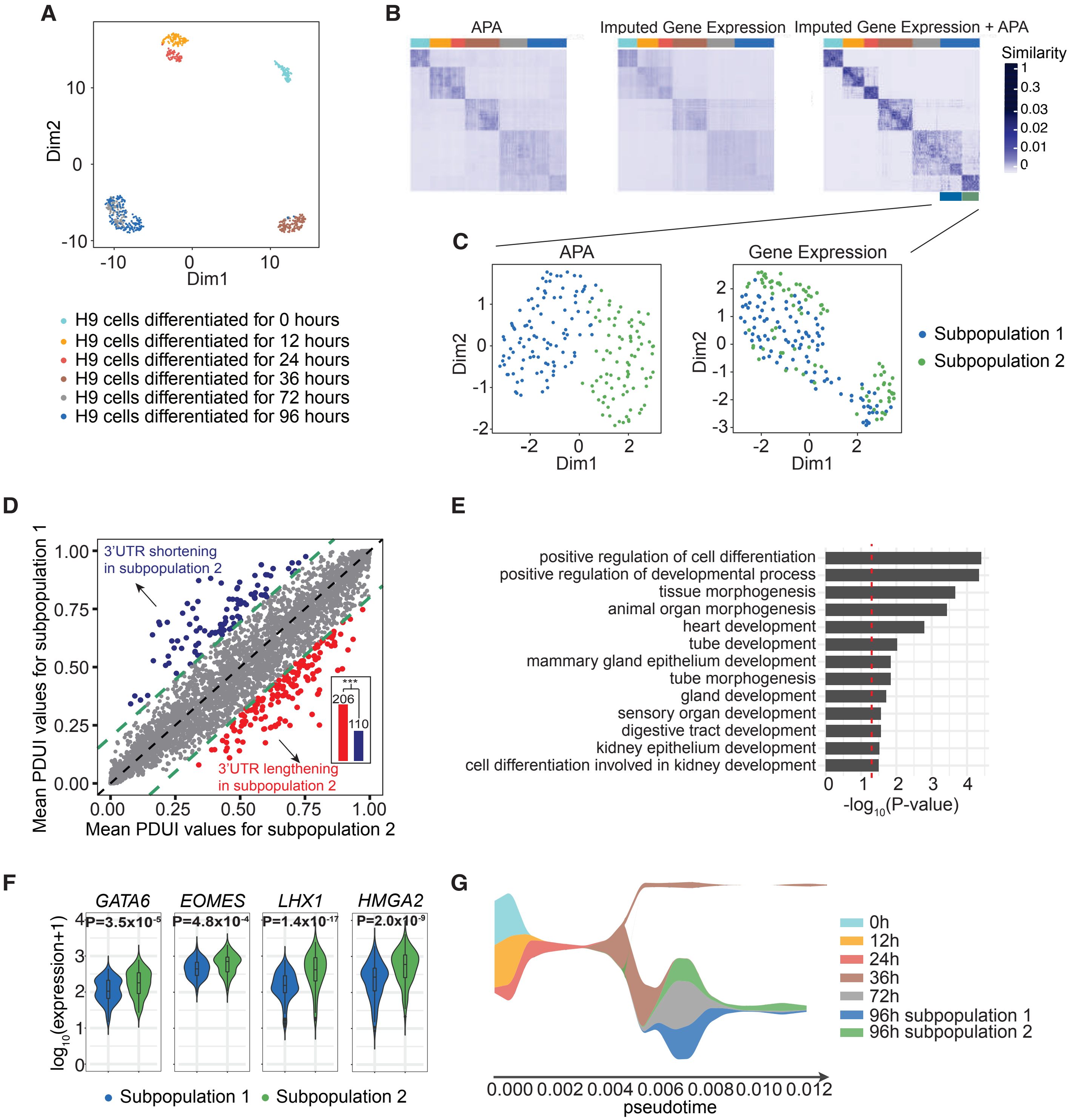

I Calculate raw APA dynamics
cell 1
cell i
APA dynamics quantified as:
PDUI = abundance of the long isoform / total abundance
Possible 3'UTR isoforms
proximal Poly(A) site
distal Poly(A) site
II Find candidate neighboring cells
Candidate neighboring cells
III Impute PDUI
genes
cells
PDUI
0
1
Imputation
AAA
AAA
?
AAA
genes
cells


A
scDaPars
random
Recovered poly(A) site rate
80%
60%
40%
20%
0
0
25
50
75
100
cutoff (bp)
B
Percentage of Overlapping
80%
60%
40%
20%
0
66.2%
14.5%
scDaPars
Random
C
$P = 1.2x10^{-44}$
bits
2
1
0
1
2
3
4
5
6
D
Wilcoxon, $P < 2.2x10^{-16}$
Pearson's Correlation
0.8
0.6
0.4
Method
Method
DaPars
scDaPars
E
DaPars
scDaPars
R = 0.74
R = 0.82
PDUI, average of single cells
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
PDUI, Bulk


A
Reference
5.0
2.5
0.0
-2.5
-5.0
Dim2
-2
0
-2
Dim1
B
Dropout introduced
2
1
0
-1
-2
Dim2
-2
1
0
1
2
Dim1
C
scDaPars
5.0
2.5
0.0
-2.5
Dim2
-4
-2
0
2
4
Dim1
● Naive CD4 T cell
● Activated CD4 T cell
D
E
F
shortening
lengthening


A
scDaPars
B
scAPA
C
Sierra
Dim 2
Dim 1
20
0
-20
-40
10
-10
B cell
CD14+ monocyte
CD4+ T cell
Cytotoxic T cell
Natural killer cell
D
E
F
Silhouette width Si
0.50
0.25
0.00
-0.25


A
P < 2.2x10-16
367
116
1.00
0.75
0.50
0.25
0.00
PDUI, Tumor cells
0.00
0.25
0.50
0.75
1.00
PDUI, Normal cells
● 3'UTR lengthening in Tumor cells
● Non-significant
● 3'UTR shortening in Tumor cells
B
10
0
-10
-20
Dim2
-10
-5
0
5
10
Dim1
● nonTumor
● Tumor
C
10
0
-10
-20
Dim2
-12.5
-10.0
-7.5
-5.0
-2.5
Dim1
● BC02
● BC03
● BC03LN
● BC04
● BC05
● BC07
● BC07LN
D
10
5
0
Dim2
10
11
12
Dim1
-1.5
-2.0
-2.5
-3.0
10
11
● B cell
● Myeloid cells
● T cell
E
10
5
0
Dim2
10
11
12
Dim1
● group1
● group2
F
P = 0.002
322
238
1.00
0.75
0.50
0.25
0.00
PDUI, Group 2
0.00
0.25
0.50
0.75
1.00
PDUI, Group 1
● 3'UTR lengthening in Group 2 cells
● Non-significant
● 3'UTR shortening in Group 2 cells


A
Dim2
10
0
-10
-10
0
10
Dim1
H9 cells differentiated for 0 hours
H9 cells differentiated for 12 hours
H9 cells differentiated for 24 hours
H9 cells differentiated for 36 hours
H9 cells differentiated for 72 hours
H9 cells differentiated for 96 hours
B
APA
Imputed Gene Expression
Imputed Gene Expression + APA
Similarity
1
0.3
0.03
0.02
0.01
0
C
APA
Gene Expression
Dim2
Dim1
1
0
-1
-2
-2
0
2
2
1
0
-1
-2
-3
-2
0
2
Subpopulation 1
Subpopulation 2
D
3'UTR shortening in subpopulation 2
3'UTR lengthening in subpopulation 2
***
206
110
Mean PDUI values for subpopulation 1
Mean PDUI values for subpopulation 2
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
E
positive regulation of cell differentiation
positive regulation of developmental process
tissue morphogenesis
animal organ morphogenesis
heart development
tube development
mammary gland epithelium development
tube morphogenesis
gland development
sensory organ development
digestive tract development
kidney epithelium development
cell differentiation involved in kidney development
0
1
2
3
4
$-\log_{10}$(P-value)
F
GATA6
EOMES
LHX1
HMGA2
$P=3.5\times10^{-5}$
$P=4.8\times10^{-4}$
$P=1.4\times10^{-17}$
$P=2.0\times10^{-9}$
$\log_{10}$(expression+1)
4
3
2
1
0
Subpopulation 1
Subpopulation 2
G
0h
12h
24h
36h
72h
96h subpopulation 1
96h subpopulation 2
0.000
0.002
0.004
0.006
0.008
0.010
0.012
pseudotime

Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression

Yipeng Gao, Lei Li, Christopher Ian Amos, et al.

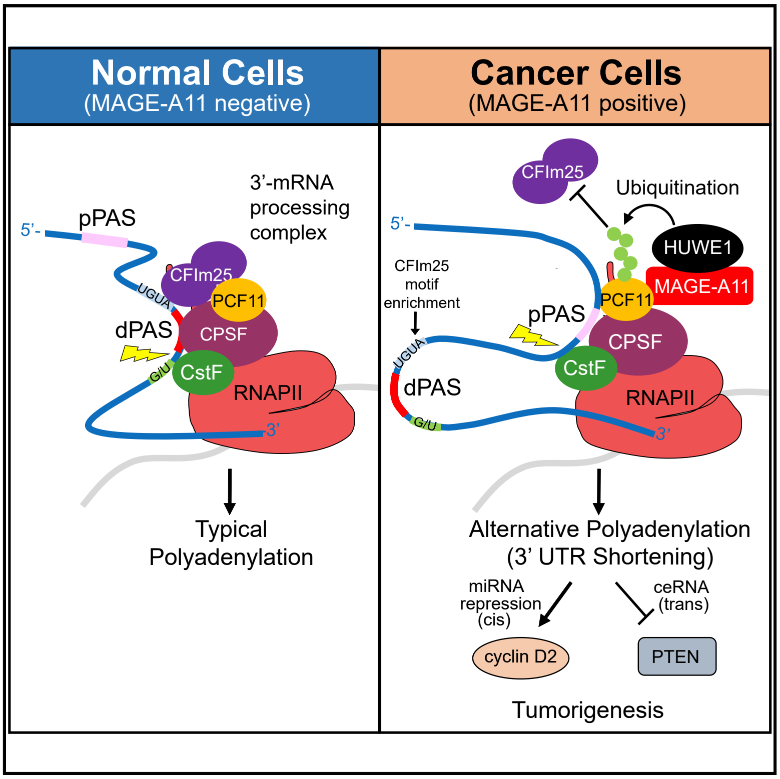
Genome Res. published online May 25, 2021
Access the most recent version at doi:10.1101/gr.271346.120

Article

Molecular Cell

A Cancer-Specific Ubiquitin Ligase Drives mRNA Alternative Polyadenylation by Ubiquitinating the mRNA 3′ End Processing Complex

Graphical Abstract



Highlights

- MAGE-A11 is aberrantly expressed in cancer and is a potent oncogene
- MAGE-A11-HUWE1 ubiquitin ligase promotes ubiquitination and degradation of PCF11
- MAGE-A11 promotes alternative polyadenylation and 3′ UTR shortening in cancer
- MAGE-A11-induced 3′ UTR shortening modulates core oncogenes and tumor suppressors

Authors

Seung Wook Yang, Lei Li,
Jon P. Connelly, ...,
Shondra M. Pruett-Miller, Wei Li,
Patrick Ryan Potts

Correspondence

ryan.potts@stjude.org

In Brief

Yang et al. show that the germ-cell-restricted MAGE-A11 is aberrantly expressed in tumors and drives tumorigenesis through ubiquitination of PCF11, a component of the 3′-mRNA processing complex. PCF11 ubiquitination results in alternative polyadenylation of transcripts leading to 3′ UTR shortening.

Yang et al., 2020, Molecular Cell *77*, 1206–1221
March 19, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.molcel.2019.12.022

CellPress

Molecular Cell
Article

A Cancer-Specific Ubiquitin Ligase Drives mRNA Alternative Polyadenylation by Ubiquitinating the mRNA 3′ End Processing Complex

Seung Wook Yang,[1,6] Lei Li,[2,3,6] Jon P. Connelly,[1] Shaina N. Porter,[1] Kiran Kodali,[4] Haiyun Gan,[1] Jung Mi Park,[1] Klementina Fon Tacer,[1] Heather Tillman,[5] Junmin Peng,[4] Shondra M. Pruett-Miller,[1] Wei Li,[2,3] and Patrick Ryan Potts[1,7,*]

[1]Department of Cell and Molecular Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
[2]Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA
[3]Division of Biostatistics, Dan L. Duncan Cancer Center and Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX 77030, USA
[4]Departments of Structural Biology and Developmental Neurobiology, Center for Proteomics and Metabolomics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
[5]Veterinary Pathology Core, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
[6]These authors contributed equally
[7]Lead Contact
*Correspondence: ryan.potts@stjude.org

https://doi.org/10.1016/j.molcel.2019.12.022

SUMMARY

Alternative polyadenylation (APA) contributes to transcriptome complexity by generating mRNA isoforms with varying 3′ UTR lengths. APA leading to 3′ UTR shortening (3′ US) is a common feature of most cancer cells; however, the molecular mechanisms are not understood. Here, we describe a widespread mechanism promoting 3′ US in cancer through ubiquitination of the mRNA 3′ end processing complex protein, PCF11, by the cancer-specific MAGE-A11–HUWE1 ubiquitin ligase. MAGE-A11 is normally expressed only in the male germline but is frequently re-activated in cancers. MAGE-A11 is necessary for cancer cell viability and is sufficient to drive tumorigenesis. Screening for targets of MAGE-A11 revealed that it ubiquitinates PCF11, resulting in loss of CFIm25 from the mRNA 3′ end processing complex. This leads to APA of many transcripts affecting core oncogenic and tumor suppressors, including cyclin D2 and PTEN. These findings provide insights into the molecular mechanisms driving APA in cancer and suggest therapeutic strategies.

INTRODUCTION

Alternative polyadenylation (APA) of messenger RNA (mRNA) is a widespread phenomenon that frequently occurs in a large proportion of human genes (Elkon et al., 2013; Ji et al., 2009; Mayr and Bartel, 2009; Sandberg et al., 2008). Recent studies have shown that at least 70% of mammalian genes have multiple polyadenylation sites (PASs) in their 3′ untranslated regions (UTRs) (Derti et al., 2012; Hoque et al., 2013). Selection of the PAS is coordinated by recognition of core sequence elements in the mRNA by the mRNA 3′ end processing complex that is composed of several protein complexes, including CPSF, CFI, CFII, and CstF complexes, and single proteins, such as PABPN1, RBBP6, and SYMPK (Elkon et al., 2013; Shi et al., 2009; Tian and Manley, 2017). Modulation of components of these complexes can lead to the use of cryptic PASs, resulting in APA (Martin et al., 2012; Masamha et al., 2014; Yao et al., 2012).

The consequences of APA can be significant, with effects on post-transcriptional gene regulation, including mRNA stability, translation, nuclear export, and cellular localization (reviewed in Tian and Manley, 2017). One well-noted consequence of APA resulting in 3′ UTR shortening (3′ US) is mRNA evasion of microRNA (miRNA)-based repression (Hoffman et al., 2016; Mayr and Bartel, 2009; Sandberg et al., 2008). In addition to regulating cognate transcripts in *cis*, 3′ US can lead to competing-endogenous RNA (ceRNA) regulation in *trans* such that the shortened 3′ UTRs no longer sequester miRNAs and the released miRNAs can be directed to repress their ceRNA partners (Salmena et al., 2011).

APA can be a regulated process that is required for normal physiological functions, including cellular differentiation, neuronal activity, and spermatogenesis (Flavell et al., 2008; Ji and Tian, 2009; Li et al., 2016). For example, APA leading to 3′ UTR lengthening of transcripts in the brain is frequent and results in diverse protein isoforms with differential subcellular localization (Ciolli Mattioli et al., 2019; Miura et al., 2013). Furthermore, 3′ US is associated with T lymphocyte activation and induced proliferation (Sandberg et al., 2008), as well as male germ cell differentiation (MacDonald and Redondo, 2002).

Aberrant APA is often associated with disease, including in cancer, where global 3′ US is a hallmark of most tumors (Fu et al., 2011; Masamha et al., 2014; Mayr and Bartel, 2009; Xia et al., 2014). Pan-cancer analysis revealed that >90% of APA events lead to 3′ US (Xia et al., 2014). Several oncogenes are known to be affected, including the cyclin D1 cell cycle regulator, whose levels are increased due to 3′ US (Mayr and Bartel, 2009). Our recent study also suggests that 3′ US in breast cancer can repress tumor suppressor genes in *trans* by disrupting ceRNA crosstalk (Park et al., 2018). Despite these observations, the mechanisms that promote APA are not well established. Although 3′ US in a subset of glioblastomas can be attributed to CFIm25 downregulation (Masamha et al., 2014), the genetic underpinnings for the vast majority of tumors is largely unknown.

MAGE genes are conserved in all eukaryotes and are defined by a common MAGE homology domain (MHD), which consists of tandem winged helix motifs (Doyle et al., 2010; Lee and Potts, 2017; Newman et al., 2016). A subset of human MAGE proteins is categorized as cancer-testis antigens (CTAs) because they are physiologically restricted to the testis but are aberrantly expressed in cancers (Pineda et al., 2015; Simpson et al., 2005). Recently, MAGE CTAs have gained growing interest as hallmarks of cancers because of their broad expression in aggressive cancers, correlation with poor clinical prognosis, and their oncogenic ability to promote increased tumor growth and metastasis (Pineda et al., 2015; Weon and Potts, 2015). We and others have shown that MAGE proteins function as substrate adaptors through their ability to recruit novel proteins to specific E3 ubiquitin ligases to promote their ubiquitination and often degradation (Doyle et al., 2010; Hao et al., 2013; Pineda et al., 2015). Thus, MAGE proteins may represent a way in which tumors co-opt germ cell functions to rewire key signaling pathways in cancer cells by reprogramming ubiquitin ligases. However, the molecular mechanisms and oncogenic potential of most MAGE CTAs, including MAGE-A11, are unknown.

Here, we show that the normally germ-cell-restricted MAGE-A11 is aberrantly expressed in cancer and acts as a potent oncogene that drives tumorigenesis by promoting APA leading 3′ US of many transcripts. MAGE-A11 acts as a substrate adaptor for the HUWE1 E3 ubiquitin ligase to promote aberrant ubiquitination of the PCF11 subunit of the mRNA 3′ end processing complex in cancer cells. This leads to the loss of CFIm25 from the mRNA 3′ end processing complex and results in 3′ US of transcripts that have enrichment of CFIm25 binding sites upstream of their distal PASs. Importantly, expression of a non-degradable PCF11 mutant suppressed MAGE-A11 oncogenic activity and 3′ US. Analysis of the transcripts affected by MAGE-A11 revealed core oncogenic and tumor suppressor genes and pathways. This includes 3′ US of the cyclin D2 oncogene leading to deregulation of the Rb tumor suppressor pathway. Furthermore, ceRNA partners of 3′ US transcripts included many tumor suppressor genes, such as PTEN that is downregulated by MAGE-A11, resulting in activation of the Akt growth signaling pathway. These findings provide insights into the function of MAGE-A11 and help explain the molecular mechanisms driving APA in cancer.

RESULTS

MAGE-A11 Is Aberrantly Expressed in Cancer and Is Necessary and Sufficient to Drive Tumor Growth

To thoroughly examine the expression pattern of *MAGE-A11*, we analyzed its expression by qRT-PCR in 26 disease-free human tissues and found that it is normally restricted to expression in the testis and placenta (Figure 1A). These findings were confirmed in 51 human tissues from the GTEx project (Figure S1A) and at the protein level by immunohistochemistry, showing expression of MAGE-A11 in germ cells of the testis and syncytiotrophoblasts in placental tissue (Figures S1B and S1C). Like other CTA genes, MAGE-A11 is aberrantly expressed in tumors (Bai et al., 2005; Lian et al., 2012; Su et al., 2013; Xia et al., 2013). Our analysis of The Cancer Genome Atlas (TCGA) transcriptomic data revealed that *MAGE-A11* is frequently expressed in many patient tumors, including lung squamous cell carcinoma (>60%), ovarian carcinoma (>40%), and head and neck squamous cell carcinoma (>40%) (Figure 1B). Furthermore, immunohistochemistry staining of ovarian carcinoma and lung squamous cell carcinoma tumor microarrays confirmed MAGE-A11 protein (Figures S1D and S1E) in 35% of samples (n = 211), regardless of tumor stage or grade (Figure S1F).

To determine whether the aberrant expression of MAGE-A11 in tumor cells is simply a passenger event due to global genomic dysregulation or whether MAGE-A11 has a more active role in driving tumorigenesis, we performed a series of gain- and loss-of-function studies to elucidate the role of MAGE-A11 in driving cancer cell growth. First, we examined whether multiple cancer cells require the expression of MAGE-A11 for viability. Intriguingly, transient knockdown of MAGE-A11 in H520 lung squamous cell carcinoma cells and DAOY medulloblastoma cells resulted in dramatic decrease in cell viability (Figure 1C). Furthermore, knockout of MAGE-A11 decreased the proliferation rate of DAOY and H520 cells, which could be rescued by re-expression of MAGE-A11 (Figures 1D and S1G–S1I). Furthermore, knockout of MAGE-A11 reduced other hallmarks of cancer, such as clonogenic growth and anchorage-independent growth of H520 and DAOY cells (Figures 1E–1G, S1J, and S1K). Re-expression of MAGE-A11 rescued anchorage-independent tumor growth (Figure 1G). Consistent with these findings, knockout of MAGE-A11 slowed xenograft tumor growth, and re-expression of MAGE-A11 rescued tumor growth in mice (Figures 1H and S2A–S2E). Finally, to determine whether overexpression of MAGE-A11 is sufficient to drive tumorigenic phenotypes, we stably expressed MAGE-A11 in A2780 or OV56 ovarian cancer cells that do not naturally express MAGE-A11. Strikingly, expression of MAGE-A11 accelerated anchorage-independent growth of A2780 cells (Figure S2F) and xenograft tumor growth of A2780 and OV56 cells in mice (Figures 1I, 1J, S2G, and S2H). Together, these results suggest that MAGE-A11 is normally restricted to expression in the testis and placenta but is aberrantly expressed in a variety of cancers, where it is necessary and sufficient to drive tumorigenesis.

MAGE-A11 Promotes Ubiquitination and Proteasome-Dependent Degradation of PCF11

To elucidate the molecular mechanisms of MAGE-A11 oncogenic activity, we performed unbiased analysis of MAGE-A11 interacting proteins by tandem affinity purification (TAP) coupled to liquid chromatography-tandem mass spectrometry (LC-MS/MS). Only 4 proteins, PCF11, CLP1, POLR2A, and POLR2B, in addition to the MAGE-A11 bait, were identified repeatedly and specifically in TAP-MAGE-A11 cells compared to TAP-vector controls (Figures 2A and S3A). Remarkably, all four proteins

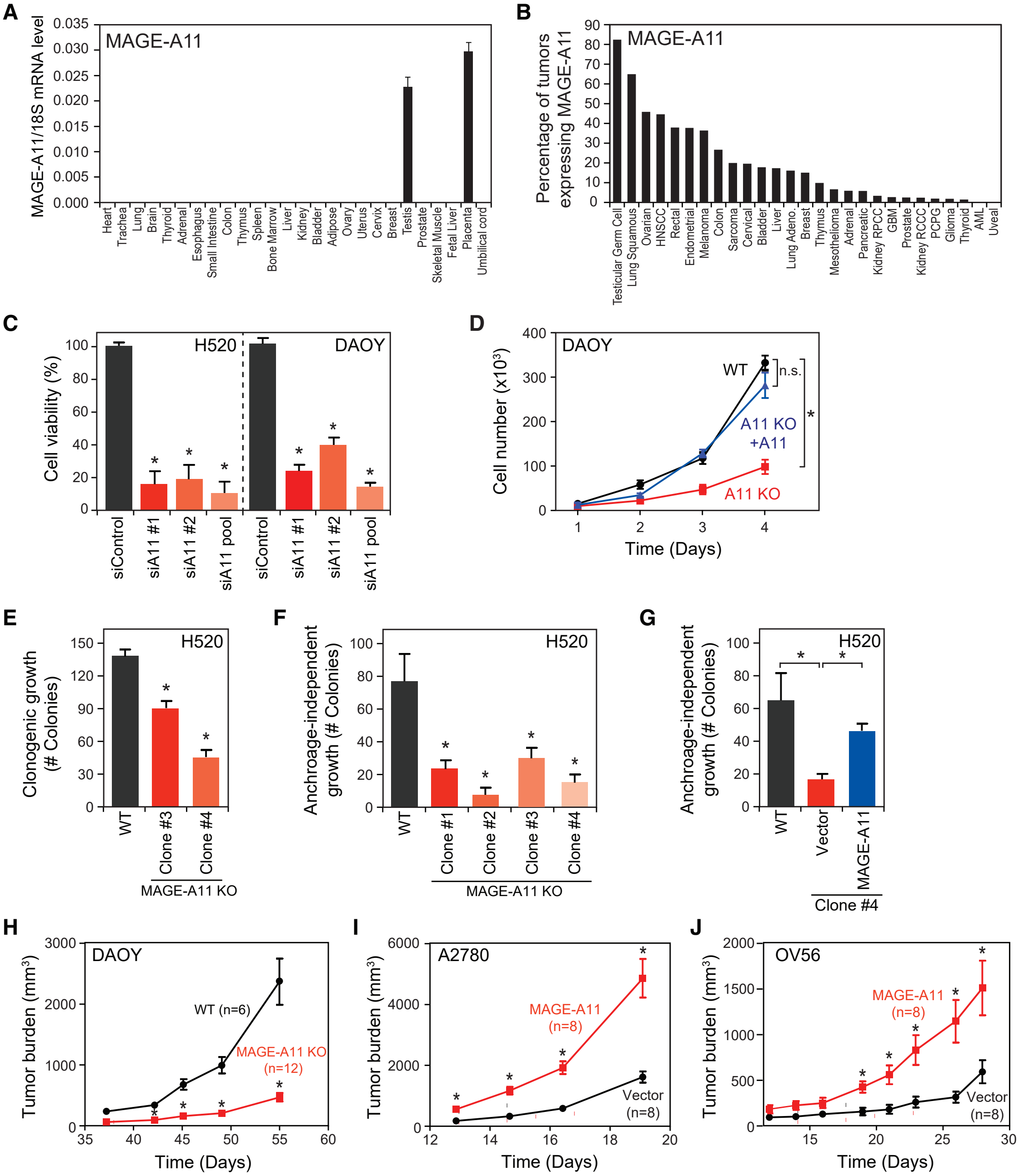


Figure 1. MAGE-A11 Is Aberrantly Expressed in Cancer and Is Necessary and Sufficient to Drive Tumor Growth

(A) qRT-PCR analysis of the normalized expression of human *MAGE-A11* in the indicated tissues (n = 3).

(B) Percentage of patient tumors expressing *MAGE-A11* is shown.

(C) H520 lung squamous cell carcinoma cells and DAOY cerebellar medulloblastoma cells were transfected with control, MAGE-A11 no. 1, MAGE-A11 no. 2, or MAGE-A11 pool siRNAs, and cell viability was measured by alamarBlue assay 72 h later.

(D) MAGE-A11-knockout DAOY cells or those reconstituted with MAGE-A11 were counted for cell proliferation at the indicated time points.

(E and F) Wild-type H520 cells or MAGE-A11-knockout H520 clones were assayed for clonogenic growth (E) and for anchorage-independent growth in soft agar colony formation assays (F).

(G) Re-expression of MAGE-A11 rescues anchorage-independent growth of MAGE-A11-knockout H520 cells.

(H) Knockout of MAGE-A11 in DAOY decreases xenograft tumor growth in mice (n = 6 for wild-type group; n = 12 for MAGE-A11-knockout group).

(I and J) Stable expression of MAGE-A11 in MAGE-A11-negative A2780 (I) and OV56 (J) ovarian cancer cells increases xenograft tumor growth in mice (n = 8 per group). Data are mean ± SD. *p < 0.05.

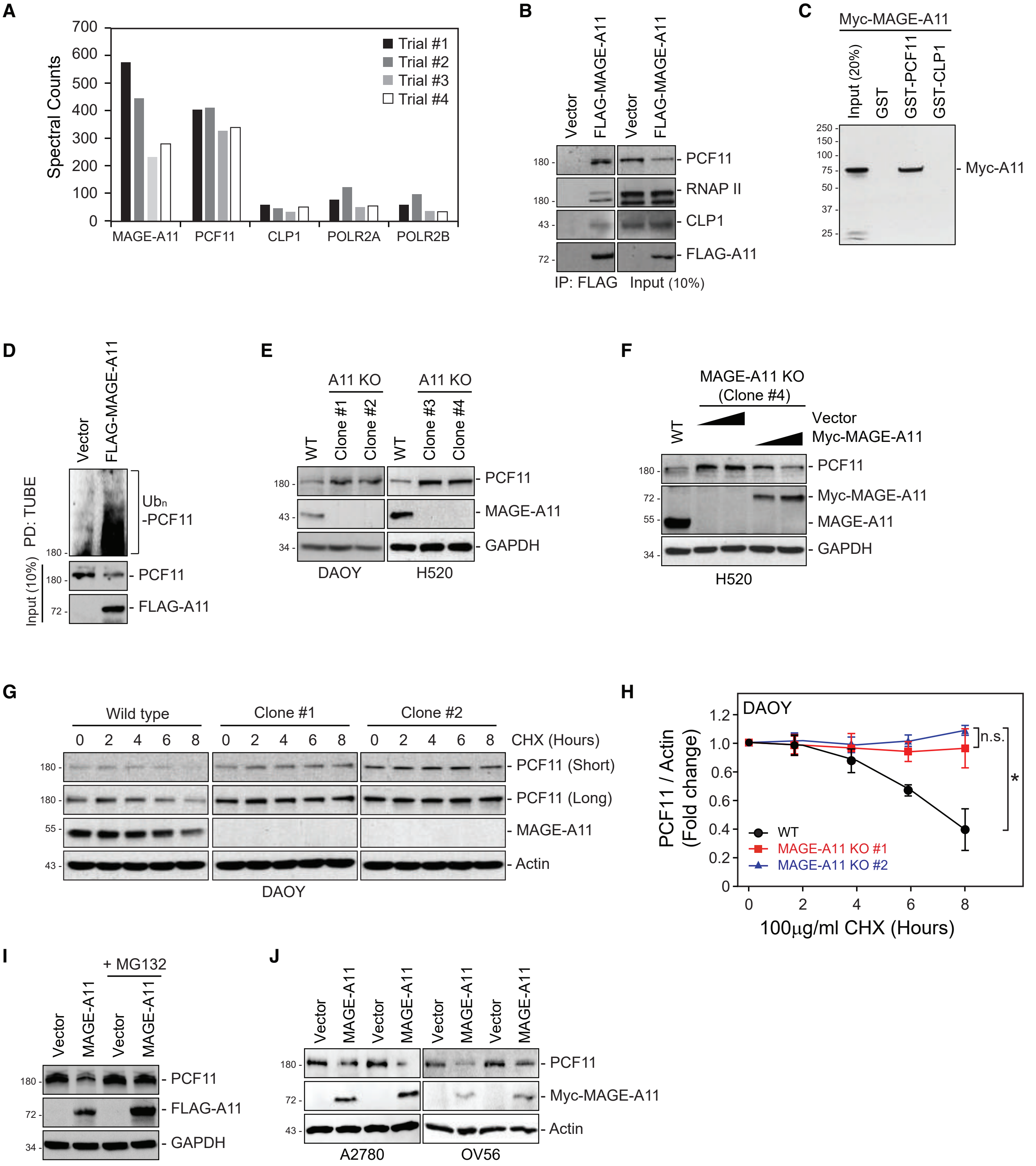


Figure 2. MAGE-A11 Promotes Ubiquitination and Degradation of PCF11

(A) MAGE-A11 interacts with 3′ mRNA processing complex proteins. HEK293 cells stably expressing TAP-vector or TAP-MAGE-A11 were subjected to pull-down followed by SDS-PAGE and LC-MS/MS (n = 4). Note spectral counts for all indicated proteins were 0 in TAP-vector samples.

(B) Interaction between MAGE-A11 and 3′ mRNA processing proteins were validated by immunoprecipitation (IP). HEK293FT cells stably expressing FLAG-vector or FLAG-MAGE-A11 were subjected to pull-down with anti-FLAG followed by SDS-PAGE and immunoblotting for endogenous PCF11, RNAP II, and CLP1.

(C) Recombinant glutathione S-transferase (GST)-PCF11, but not GST-CLP1, binds *in vitro* translated Myc-MAGE-A11.

(legend continued on next page)

are known to interact within the context of the mRNA 3′ end processing complex. PCF11 and CLP1 belong to the cleavage factor II (CFII) subcomplex that directly interacts with RNA polymerase II (RNAPII) via p-S2 residues in the RNAPII CTD-binding PCF11 CID domain (Licatalosi et al., 2002; Meinhart and Cramer, 2004). We confirmed that MAGE-A11 interacts with the CFII complex and RNAPII in cells by co-immunoprecipitation (coIP) (Figure 2B). Further analysis revealed that the MAGE-A11 directly binds PCF11, but not CLP1 (Figure 2C), *in vitro*.

Previously, we have reported that many MAGE proteins bind to specific E3 ubiquitin ligases and modulate ubiquitination of target proteins (Doyle et al., 2010; Hao et al., 2013, 2015; Pineda et al., 2015; Weon et al., 2018). Consistent with this, we found that MAGE-A11 increased PCF11 ubiquitination (Figure 2D). Knockout of MAGE-A11 increased PCF11 protein levels in DAOY and H520 cells (Figure 2E) that could be rescued by re-expression of MAGE-A11 (Figures 2F and S3B). These results were confirmed by protein half-life measurements that showed increased stability of PCF11 in MAGE-A11 knockout cells (Figures 2G and 2H). Furthermore, MAGE-A11 overexpression decreased PCF11 levels in A2780, OV56, and HEK293FT cells in a proteasome-dependent manner (Figures 2I and 2J). Importantly, this effect was specific to PCF11, as MAGE-A11 expression did not alter levels of CPSF, CstF, and CFI complexes or PCF11-interacting proteins (Figure S3C).

MAGE-A11 Recruits PCF11 to the HUWE1 E3 Ubiquitin Ligase for Ubiquitination and Degradation

Next, we utilized the previously described ubiquitin-activated interaction trap (UBAIT) approach (O’Connor et al., 2015) to identify which E3 ubiquitin ligase partners with MAGE-A11 to promote PCF11 ubiquitination and degradation. Follow-up analysis of the candidate E3 ligases revealed that HUWE1 is required for the MAGE-A11-mediated ubiquitination and degradation of PCF11. We confirmed that MAGE-A11 interacted with HUWE1 (Figure 3A) and recruited PCF11 to the HUWE1 ligase (Figure 3B), consistent with the function of MAGEs as substrate adapters. Depletion of MAGE-A11 or HUWE1 decreased ubiquitination of PCF11 (Figure 3C) and increased PCF11 protein levels (Figure 3D). Furthermore, MAGE-A11 induced PCF11 degradation in a HUWE1-dependent manner (Figure 3E). These results were confirmed by protein half-life measurements that showed increased stability of PCF11 upon HUWE1 knockdown in DAOY cells naturally expressing MAGE-A11 (Figure 3F). Together, these results suggest that MAGE-A11 targets PCF11 for ubiquitination and degradation by the HUWE1 E3 ubiquitin ligase.

MAGE-A11 Promotes Alternative Polyadenylation Leading to 3′ US in Tumors

Because PCF11 is one of the polyA cleavage factors responsible for mRNA 3′ end processing, we examined whether MAGE-A11 regulation of PCF11 would alter PAS choice, leading to APA and changes in 3′ UTR length. We performed high-depth (2.5×10^8 reads) RNA sequencing (RNA-seq) and applied our previously described bioinformatics algorithm DaPars (dynamic analysis of alternative polyadenylation from RNA-seq) (Masamha et al., 2014; Xia et al., 2014) to identify 3′ UTR alterations between control and MAGE-A11-expressing HEK293FT cells. The difference in 3′ UTR length between samples was quantified as a change in percentage of distal PAS usage index (PDUI). MAGE-A11 expression resulted in 268 APA events, with the majority, 213, being 3′ US events in which the proximal PAS (pPAS) was preferentially used (Figures 4A and 4B). Similar results were also obtained using the APAtrap algorithm (Ye et al., 2018) with a large number of 3′ US transcripts identified by both approaches (χ^2 $p < 0.00001$). In contrast to MAGE-A11 expression, knockout of MAGE-A11 in DAOY cells resulted in significantly more transcripts with 3′ UTR lengthening ($p = 0.008254$; Figure S4A). These results suggest that MAGE-A11 promotes 3′ US of transcripts.

Next, to examine whether MAGE-A11 induces 3′ US in tumors, we analyzed APA events in A2780 and OV56 xenograft tumors from mice. We identified 531 and 275 significant APA events driven by MAGE-A11 in OV56 and A2780 tumors, respectively (Figures 4C–4E). These APA events were almost exclusively 3′ US (95% and 84% of APA events in OV56 and A2780 tumors, respectively; Figures 4C and 4D). This included a statistically enriched ($p = 1.01^{-8}$) core set of common 3′ US transcripts altered in each tumor type, with a large number of cell-type-specific 3′ US transcripts. Furthermore, analysis of TCGA transcriptomics datasets from human ovarian carcinoma and lung squamous carcinoma patient tumors for 3′ UTR usage revealed a significant number of transcripts (106 [85% of APA events] and 151 [87% of APA events], respectively) with 3′ US in MAGE-A11-expressing tumors compared to MAGE-A11-negative control tumors (Figures 4F–4H). Notably, many of the transcripts with APA had altered mRNA levels, consistent with disruption of *cis*-regulatory elements in the 3′ UTR of these transcripts (Figures S4B–S4F). Together, these results suggest that MAGE-A11 regulation of PCF11 drives APA leading to 3′ US in tumors.

MAGE-A11-Induced PCF11 Ubiquitination Dissociates CFIm25 from RNAPII

Previous studies have shown that changes in the levels of specific components of the mRNA 3′ end processing complex can

(D) Expression of MAGE-A11 promotes PCF11 ubiquitination. Ubiquitinated proteins from FLAG-vector or FLAG-MAGE-A11 stably expressing HEK293FT cells were isolated with tandem ubiquitin binding entity (TUBE)-agarose followed by SDS-PAGE and immunoblotting for endogenous PCF11.

(E) Knockout of MAGE-A11 increases PCF11 protein levels. Wild-type or MAGE-A11 knockout DAOY or H520 cells were blotted for the indicated proteins.

(F) Re-expression of MAGE-A11 decreases PCF11 protein levels in MAGE-A11 knockout H520 cells. Increasing amounts of MAGE-A11 were stably expressed in MAGE-A11 knockout-H520 cells.

(G and H) Knockout of MAGE-A11 increases PCF11 protein stability in DAOY cells. MAGE-A11 wild-type or knockout DAOY cells were treated with 100 μg/mL cycloheximide for the indicated times. Cell lysates were immunoblotted (G) and quantitated (H; n = 3). Data are mean ± SD. $^{*}p < 0.05$.

(I) MAGE-A11 promotes proteasome-dependent PCF11 degradation. HEK293FT cells stably expressing FLAG-vector or FLAG-MAGE-A11 were treated with 10 μM MG132 for 4 h before immunoblotting.

(J) Stable expression of MAGE-A11 decreases PCF11 protein levels in A2780 and OV56 cells. Cell lysates were blotted for the indicated proteins.

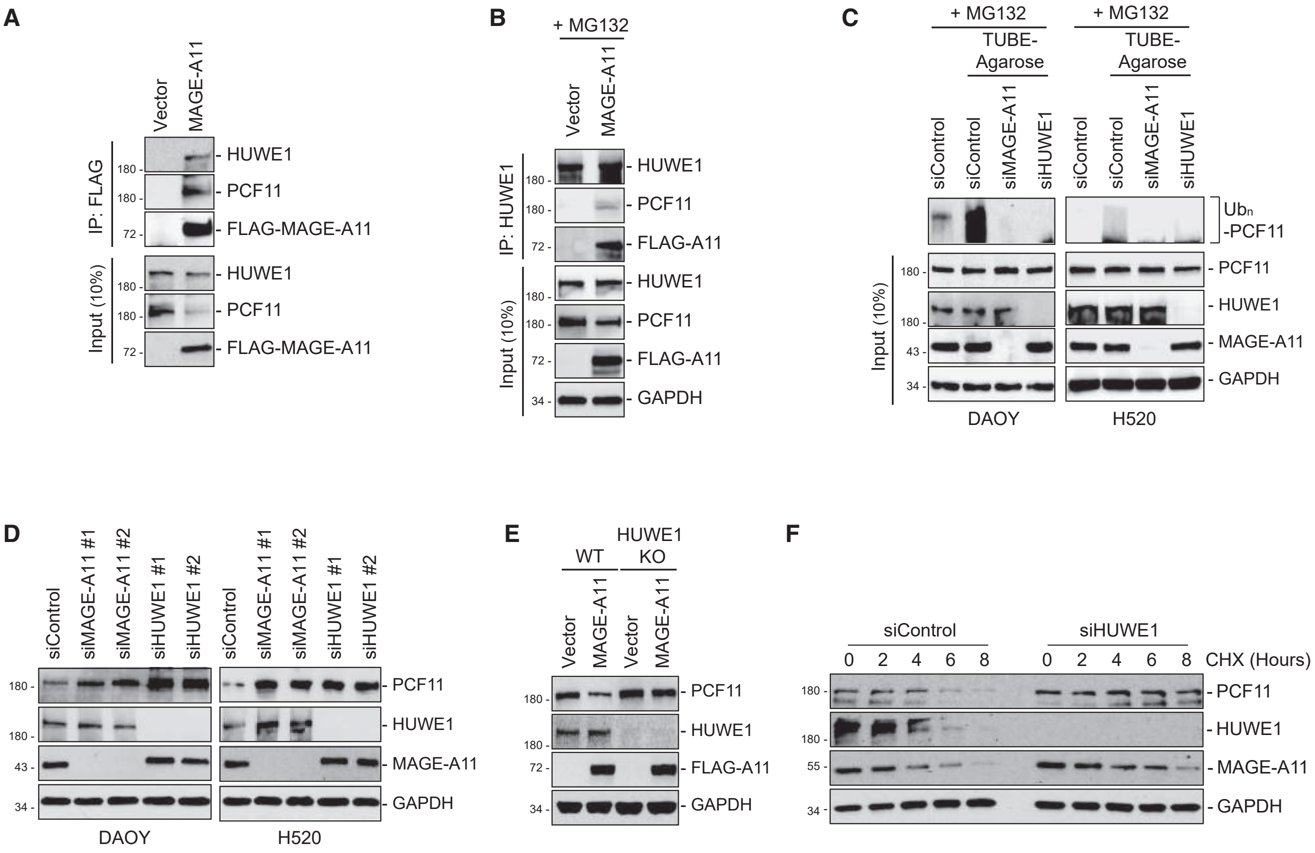


Figure 3. MAGE-A11 Recruits PCF11 to the HUWE1 E3 Ubiquitin Ligase for Ubiquitination and Degradation

(A) MAGE-A11 interacts with PCF11 and HUWE1. HEK293FT cells stably expressing FLAG-vector or FLAG-MAGE-A11 were subjected to pull-down with anti-FLAG followed by SDS-PAGE and immunoblotting for anti-HUWE1 and anti-PCF11.

(B) MAGE-A11 promotes PCF11 binding to HUWE1 E3 ubiquitin ligase. HEK293FT cells stably expressing FLAG-vector or FLAG-MAGE-A11 were treated with 10 μM MG132 for 4 h followed by IP with anti-HUWE1, SDS-PAGE, and immunoblotting for the indicated proteins.

(C) MAGE-A11-induced ubiquitination of PCF11 depends on HUWE1 E3 ligase. DAOY or H520 cells were transfected with the indicated siRNAs for 72 h and treated with 10 μM MG132 for 4 h followed by pull-down with TUBE-agarose, SDS-PAGE, and immunoblotting for PCF11.

(D) Depletion of MAGE-A11 or HUWE1 increases PCF11 protein levels. DAOY or H520 cells were transfected with the indicated siRNAs for 72 h and blotted for the indicated proteins.

(E) MAGE-A11-induced PCF11 degradation is dependent on HUWE1. Wild-type or HUWE1 knockout HEK293T cells stably expressing FLAG-vector or FLAG-MAGE-A11 were immunoblotted for the indicated proteins.

(F) Knockdown of HUWE1 increases PCF11 protein stability in DAOY cells that express MAGE-A11. siControl or siHUWE1 DAOY cells were treated with 100 μg/mL cycloheximide for the indicated times. Cell lysates were immunoblotted for the indicated proteins.

lead to APA. Although depletion of CFIm25 by small interfering RNA (siRNA)-mediated knockdown led to 3′ US, depletion of PCF11 produced 3′ UTR lengthening (Baejen et al., 2017; Kamieniarz-Gdula et al., 2019; Li et al., 2015; Masamha et al., 2014; Ogorodnikov et al., 2018). In contrast, our data suggest that MAGE-A11-induced PCF11 ubiquitination leads to 3′ US (Figure 4). Importantly, there was very little overlap (10 transcripts) in MAGE-A11-indued 3′ US transcripts (213 transcripts) and PCF11 siRNA-induced 3′ UTR lengthened transcripts (545 transcripts), suggesting that dynamic MAGE-A11-induced ubiquitination of PCF11 has distinct outcomes compared to static siRNA-mediated knockdown of PCF11. To explore this further, we examined whether MAGE-A11 ubiquitination of PCF11 could alter the architecture of the mRNA 3′ end processing complex. We found that MAGE-A11 expression resulted in significant reduction in CFIm25 association with RNAPII by coIP (Figures 5A, 5B, and S5A). Moreover, this effect was more pronounced in comparison to siRNA-mediated knockdown of PCF11 (Figures 5A, 5B, and S5A). Consistent with these findings, there is significant overlap (42%; $p = 7.8^{-55}$) in the 3′ US transcripts upon CFIm25 knockdown in HeLa cells (Masamha et al., 2014) and MAGE-A11 overexpression in HEK293FT cells (Figure S5B). Next, we determined whether ubiquitination and/or degradation of PCF11 are required for MAGE-A11-induced CFIm25 dissociation from RNAPII. CFIm25 dissociation from RNAPII by MAGE-A11 is HUWE1 dependent, confirming the importance of ubiquitination (Figure 5C). However, this effect was independent of PCF11 degradation, as rescue of PCF11 levels in MAGE-A11-expressing cells by MG132 led to stabilization of PCF11 but failed to rescue CFIm25 association with RNAPII

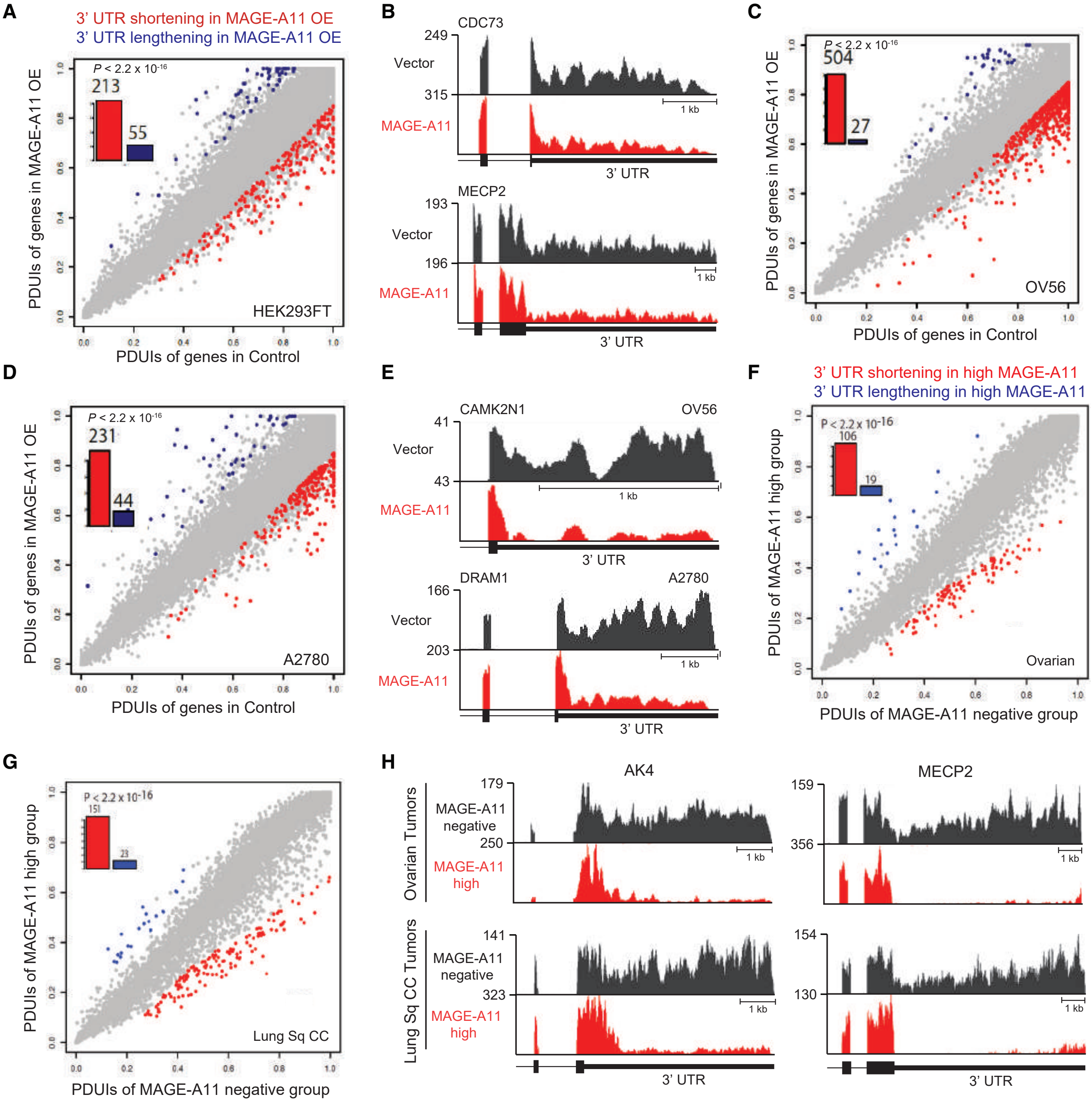


Figure 4. MAGE-A11 Promotes Alternative Polyadenylation Leading to 3′ UTR Shortening in Tumors

(A) Transcriptome analysis of HEK293FT cells stably expressing FLAG-vector or FLAG-MAGE-A11 reveals that MAGE-A11 promotes 3′ US. Scatterplot of percentage of distal polyA site usage index (PDUI) in control and MAGE-A11-overexpressing cells shows shortened 3′ UTRs (n = 213) or lengthened 3′ UTRs (n = 55) in genes by overexpression of MAGE-A11. False discovery rate (FDR) ≤ 0.05, ΔPDUIs ≥ 0.2 and 2-fold change of PDUIs between control and MAGE-A11 overexpression are colored. The shifting toward pPAS is significant ($p < 2.2 \times 10^{-16}$; binomial test).

(B) Representative RNA-seq density plots for genes with 3′ UTR shortening are shown. Numbers on y axis indicate RNA-seq read coverage.

(C and D) Scatterplot of PDUIs from both datasets of mouse xenografts in Figures 1I and 1J using the same cutoffs as in (A). Data from OV56 and A2780 tumors are shown in (C) and (D), respectively. The shifting toward pPAS is significant ($p < 2.2 \times 10^{-16}$; binomial test).

(E) Representative examples of genes with 3′ UTR shortening from datasets shown in (C) and (D) are shown.

(F and G) Global analysis of 3′ UTR changes in ovarian cancer (F) or lung squamous cell carcinoma (G) patient samples with either negative or high levels of MAGE-A11. Scatterplot of PDUIs from both datasets of patient samples is shown. The shifting toward pPAS is significant ($p < 2.2 \times 10^{-16}$; binomial test).

(H) Representative examples of genes show 3′ UTR shortening in patient samples with negative (black) or high MAGE-A11 expression levels (red).

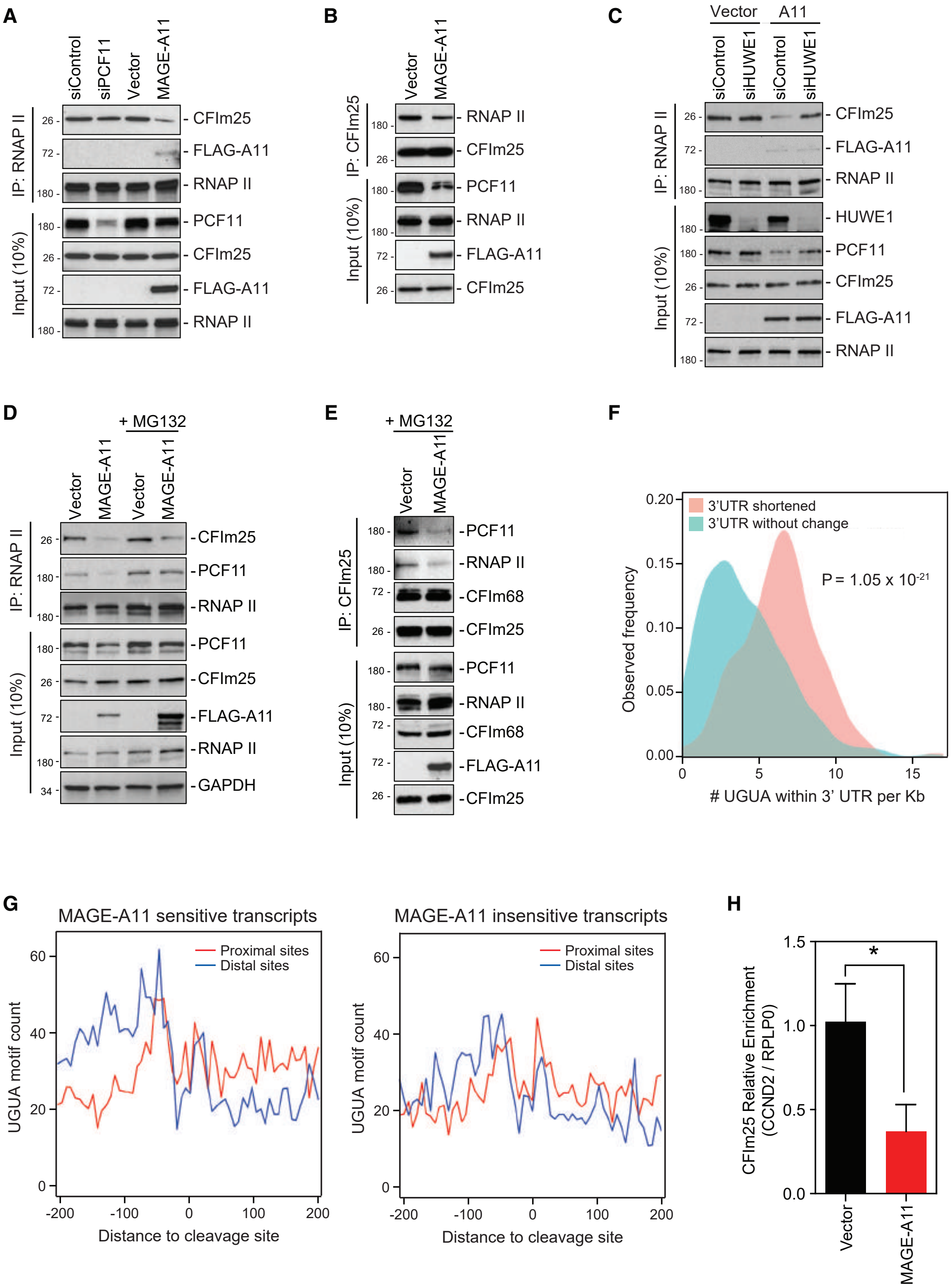


Figure 5. MAGE-A11-Induced PCF11 Ubiquitination Dissociates CFIm25 from RNAPII

(A and B) Overexpression of MAGE-A11 induces dissociation of CFIm25 from RNAPII compared to knockdown of PCF11. HEK293FT cells were transfected with the indicated siRNAs for 72 h or stably expressing FLAG-vector or FLAG-MAGE-A11 were followed by IP with anti-RNAPII (A) and IP with anti-CFIm25 (B), SDS-PAGE, and immunoblotting for the indicated proteins.

(legend continued on next page)

(Figures 5D and 5E). Notably, PCF11 interaction with RNAPII was not altered by MAGE-A11 in MG132-treated cells (Figure 5D). These results suggest that MAGE-A11-induced PCF11 ubiquitination, but not degradation, causes remodeling of the mRNA 3′ end processing complex that leads to dissociation of CFIm25. Moreover, simple steady-state depletion of PCF11 by siRNA does not mimic the effect of MAGE-A11-induced ubiquitination of PCF11.

Consistent with the dissociation of CFIm25 from RNAPII playing an important role in MAGE-A11-induced 3′ US, sequence analysis of MAGE-A11-sensitive transcripts revealed significantly more CFIm25 binding motifs (UGUA) compared to unaffected transcripts (Figure 5F). Furthermore, analysis of UGUA motif distribution near distal and proximal PASs, as described previously (Zhu et al., 2018), showed motif enrichment upstream of distal PASs in MAGE-A11-sensitive transcripts, but not proximal PASs or transcripts unaffected by MAGE-A11 (Figures 5G and S5C–S5F). This was not the case for transcripts lengthened by PCF11 siRNA-mediated knockdown (Figures S5G and S5H). Collectively, these findings provide insights into how PCF11 ubiquitination affects the mRNA 3′ end processing complex through loss of CFIm25 that leads to 3′ US of transcripts with enriched UGUA motifs upstream distal PASs. To further test this model, we performed crosslinking immunoprecipitation and qPCR (CLIP-qPCR) to determine the abundance of CFIm25 associated with a transcript, *CCND2*, which undergoes 3′ US upon MAGE-A11 expression. Expression of MAGE-A11 significantly reduced the abundance of CFIm25 associated with the *CCND2* transcript in relation to a non-MAGE-A11-regulated transcript, RPLP0 (Figure 5H).

Regulation of PCF11 Is Essential for MAGE-A11-Induced Tumorigenesis and APA

To determine whether regulation of PCF11 is required for MAGE-A11 oncogenic activity, we identified a non-degradable PCF11 mutant. The degron motif in PCF11 required for MAGE-A11 binding was mapped to amino acids 653–702 (Figures 6A, 6B, and S6A–S6D). Mutation of conserved residues in PCF11 (Figure S6E) identified I689A mutant that abolished PCF11 interaction with MAGE-A11 (Figures 6C and S6F) and disrupted ubiquitination and degradation by MAGE-A11 (Figure 6D). Importantly, introduction of PCF11 I689A into A2780 cells, by a transgene or homozygous mutation using CRISPR/Cas9, rescued PCF11 protein levels (Figures 6E and 6G) and completely or partially (depending on the clone) blocked MAGE-A11-induced xenograft tumor growth in mice (Figures 6F, 6H, S6G, and S6H). Importantly, MAGE-A11-driven APA was dependent on its regulation of PCF11, as expression of the non-degradable PCF11 I689A mutant by transgene or CRISPR/Cas9 homozygous knockin prevented MAGE-A11-induced APA in A2780 cells (Figures 6I, 6J, and S6I). These results suggest that the ability of MAGE-A11 to regulate PCF11 is critical for its oncogenic activity.

MAGE-A11-Induced 3′ US Modulates Core Oncogenic and Tumor Suppressor Pathways

To identify those 3′ US events that impact levels of their encoded proteins, we performed unbiased, quantitative proteomics using tandem mass tagging (TMT)-LC/LC-MS/MS (Niu et al., 2017) in isogenic DAOY cells with or without MAGE-A11 expression (Table S3). Consistent with previous results, PCF11 was downregulated upon MAGE-A11 expression (Figure 7A). More importantly, we found several 3′ US transcripts with altered protein levels, including the *CCND2* (cyclin D2) oncogene that was upregulated upon MAGE-A11 expression (Figure 7A). We validated these results by expressing MAGE-A11 in an independent cell line, HEK293FT, and again saw 3′ US of the *CCND2* transcript (Figures 7B and S7A) and increased protein levels (Figures 7C and 7D). As a member of the D-type cyclins, cyclin D2 has been widely implicated in cell cycle transition, differentiation, and cellular transformation (Evron et al., 2001; Sherr, 1995), and its overexpression is highly correlated with poor prognosis in various cancers (Mermelshtein et al., 2005; Sicinski et al., 1996; Takano et al., 1999, 2000). Cyclin D-Cdk4/6 inactivates retinoblastoma (Rb) tumor suppressor by progressive multiphosphorylation to release transcription factors, such as E2F (Narasimha et al., 2014; Sherr, 1995). MAGE-A11 increased phospho-Rb (S807 and S811) in HEK293FT cells and MAGE-A11 expression in ovarian and lung squamous cell carcinoma patient tumor samples correlated with increased phospho-Rb (S807/811; Figures 7E, 7F, and S7B). To determine whether cyclin D2 upregulation upon MAGE-A11 expression contributes to MAGE-A11-driven proliferation, cyclin D2 was knocked down in DAOY cells with or without MAGE-A11 expression and proliferation rates were determined. Knockdown of cyclin D2 decreased proliferation of MAGE-A11 expressing DAOY, but not MAGE-A11 knockout DAOY (Figure 7G), thus implicating upregulation of cyclin D2 by MAGE-A11 as an important contributor to MAGE-A11-mediated cellular proliferation.

To better understand how *CCND2* 3′ US may upregulate cyclin D2 protein levels, we determined whether inhibitory factors, such

(C) HUWE1 is required for MAGE-A11-induced dissociation of CFIm25 from RNAPII. HEK293FT FLAG-vector or FLAG-MAGE-A11 stable cell lines were transfected with the indicated siRNAs for 72 h followed by IP with anti-RNAPII, SDS-PAGE, and immunoblotting for the indicated proteins.

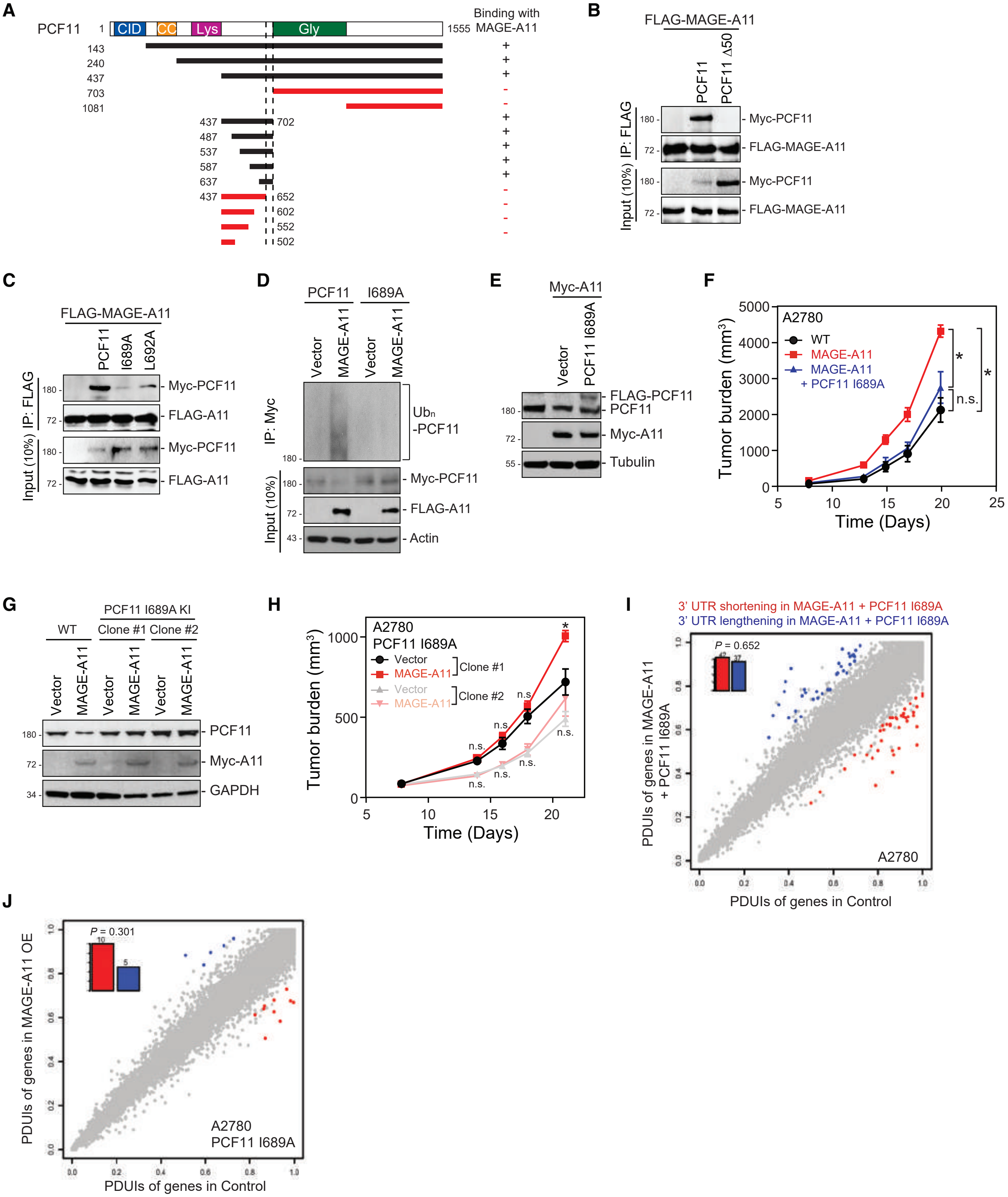
(D) PCF11 ubiquitination, but not degradation, promotes CFIm25 dissociation from RNAPII. HEK293 FLAG-vector or FLAG-MAGE-A11 stable cell lines were treated with or without 10 μM MG132 for 4 h prior to collection, anti-RNAPII IP, SDS-PAGE, and immunoblotting.

(E) MAGE-A11 dissociates CFIm25 from PCF11. Cells were treated as described in (D) before IP with anti-CFIm25, SDS-PAGE, and immunoblotting for the indicated proteins.

(F) The number of UGUA motifs within 3′ US or unaffected transcripts in MAGE-A11 overexpressing HEK293FT cells. Equal numbers of transcripts with no 3′ UTR changes were randomly selected.

(G) The UGUA motif frequency within MAGE-A11-sensitive transcripts is highly enriched upstream of distal PAS compared to proximal PAS (ΔPDUI value ≤ 0.05; $p > 0.5$) for 3′ US transcripts, but not MAGE-A11-insensitive transcripts.

(H) MAGE-A11 reduces CFIm25 associated with 3′ US transcript *CCND2*. CLIP-qPCR analysis was performed from HEK293FT cells using control immunoglobulin G (IgG) or CFIm25 antibodies. Abundance of CCND2 or control RPLP0 was determined by qRT-PCR. Normalized (CFIm25/IgG) ratios of CCND2 and RPLP0 are shown. Data (n = 3) are mean ± SD. $^{*}p < 0.05$.



(legend on next page)

as miRNAs, may repress cyclin D2 expression in MAGE-A11-negative HEK293FT cells. We used the approach pioneered by others to overexpress the 3′ UTR of *CCND2* to act as a “sponge” for potential miRNAs and other factors binding to the endogenous *CCND2* transcript (Mallon and Macklin, 2002; Matoulkova et al., 2012; Rutnam and Yang, 2012). We found that expression of the *CCND2* 3′ UTR upregulated cyclin D2 protein levels in MAGE-A11-negative, but not MAGE-A11-positive, cells (Figures 7H and S7C). In order to determine which particular miRNA(s) might mediate cyclin D2 repression, we analyzed the predicted miRNA binding sites (TargetScan; Agarwal et al., 2015) lost upon *CCND2* 3′ US and correlated these to miRNA expression datasets (miRmine; Panwar et al., 2017) to identify relevant miRNAs. Using this approach, we identified miR-191-5p, a previously reported miRNA targeting *CCND2* (Di Leva et al., 2013), as a likely candidate. We found that the miR-191-5p mimic downregulated cyclin D2 protein levels and miR-191-5p antagomiR increased cyclin D2 protein levels in MAGE-A11-negative cells, but not in MAGE-A11 expressing cells (Figures 7I and 7J). These results suggest that MAGE-A11-mediated 3′ US of *CCND2* leads to increased cyclin D2 protein levels, in part through loss of miR-191-5p repression.

In addition to 3′ US of oncogenes leading to their activation in *cis* through evading miRNA-mediated repression, we and others have also shown that these now-liberated miRNAs can downregulate competing endogenous mRNAs (ceRNAs) in *trans* (Park et al., 2018). Using our previously established computational approach to predict the *trans* effect of 3′ US to their ceRNA partners (Park et al., 2018), we found that many 3′ US ceRNA partners in ovarian cancer or lung squamous cell carcinoma patient samples with high MAGE-A11 levels are tumor suppressors (Figure 7K). Notably, the top ceRNA identified in MAGE-A11 lung squamous cell carcinoma was the tumor suppressor PTEN (Table S4). Consistently, MAGE-A11 expression markedly downregulated PTEN levels and increased downstream phospho-AKT (T308) in HEK293FT cells (Figures 7L and S7D) and ovarian carcinoma patient tumor samples (Figure 7M). To determine whether this effect depends on miRNA targeting of the *PTEN* 3′ UTR, we utilized a luciferase reporter plasmid containing the *PTEN* 3′ UTR. MAGE-A11 expression repressed *PTEN* 3′ UTR luciferase activity (Figure S7E). These results suggest that MAGE-A11-induced 3′ US has both *cis* and *trans* effects on oncogenes (cyclin D2) and tumor suppressors (PTEN), respectively, to alter key cell growth and signaling pathways.

DISCUSSION

The eukaryotic mRNA 3′ end processing complex plays an essential role in defining the transcriptome. This molecular machine interacts with the transcription machinery to define mRNA termination through cleavage of pre-mRNA and polyA tail addition. Recent transcriptomic studies have shown that a majority of mammalian genes have multiple PASs and APA contributes to the complexity of the transcriptome by generating mRNA isoforms with varying 3′ UTR lengths (Derti et al., 2012; Mayr and Bartel, 2009; Sandberg et al., 2008). Interestingly, widespread shortening of mRNA by APA is found in many types of cancers, but the molecular mechanisms contributing to APA in cancer have been unclear. Our findings elucidate a previously undefined molecular mechanism contributing to the widespread 3′ US in tumors.

The regulation of PCF11 ubiquitination by the cancer-specific E3 ubiquitin ligase MAGE-A11–HUWE1 led to changes in the mRNA 3′ end processing complex and increased the number of 3′ US transcripts in cancers. Interestingly, PCF11 is a sub-stoichiometric component of the mRNA 3′ end processing complex in many human cells and tissues (Kamieniarz-Gdula et al., 2019). Thus, even small fluctuations in PCF11 may impact mRNA 3′ end processing and the dynamics of PCF11 association with the mRNA 3′ end processing complex may be important. This is consistent with our findings that MAGE-A11-induced ubiquitination of PCF11 confers unique phenotypes compared to steady-state siRNA knockdown of PCF11. Furthermore, PCF11 couples mRNA 3′ end processing with mRNA export (Johnson et al., 2009) and phosphorylation of PCF11 CID by WNK1 is critical for release of transcripts from chromatin-associated RNAPII (Volanakis et al., 2017). Therefore, nuclear export of mature transcripts in tumor cells could potentially be regulated by MAGE-A11-mediated PCF11 ubiquitination.

Analysis of the transcripts affected by MAGE-A11-induced ubiquitination of PCF11 revealed many oncogenes and tumor suppressors. First, we and others have shown that 3′ US of oncogenes results in their increased production through evading miRNA-mediated repression. Indeed, the alternative isoforms, especially shorter transcripts of genes encoding cyclin D1, cyclin

Figure 6. Regulation of PCF11 Is Essential for MAGE-A11-Induced Tumorigenesis and APA

(A) Summary of *in vitro* binding assays from Figures S6A–S6C mapping the degron region of PCF11 recognized by MAGE-A11.

(B) HEK293FT cells stably expressing FLAG-MAGE-A11 were transfected with PCF11 wild-type or PCF11 653–702 deletion construct for 48 h before IP with anti-FLAG followed by SDS-PAGE and immunoblotting for anti-Myc.

(C) PCF11 I689A or L692A mutants have diminished interaction with MAGE-A11. HEK293FT cells stably expressing FLAG-MAGE-A11 were transfected with the indicated constructs for 48 h before IP with anti-FLAG followed by SDS-PAGE and immunoblotting for anti-Myc.

(D) PCF11 I689A mutant fails to be ubiquitinated and degraded by MAGE-A11. HEK293FT cells stably expressing FLAG-vector or FLAG-MAGE-A11 were transfected with indicated constructs for 48 h before IP with anti-Myc followed by SDS-PAGE and immunoblotting for anti-His.

(E and F) Non-degradable PCF11 I689A was stably expressed in MAGE-A11-expressing A2780 (E), and xenograft tumor growth was determined (F). Data are mean ± SD (n = 6 per group). $^*p < 0.05$.

(G) CRISPR-Cas9-mediated knockin of non-degradable PCF11 I689A mutant into A2780 prevents degradation of PCF11 by MAGE-A11.

(H) Stable expression of MAGE-A11 in PCF11 I689A knockin A2780 clones does not increase xenograft tumor growth in mice (n = 6 per group). Data are mean ± SD. $^*p < 0.05$.

(I and J) Expression (I) or knockin (J) of non-degradable PCF11 I689A rescues 3′ US in A2780 MAGE-A11-expressing tumors. Scatterplot of PDUIs (as described in Figure 4A) from mouse xenografts shown in (F) or (H) is shown. The pPAS is not significant (p = 0.652, I; p = 0.301, J; binomial test).

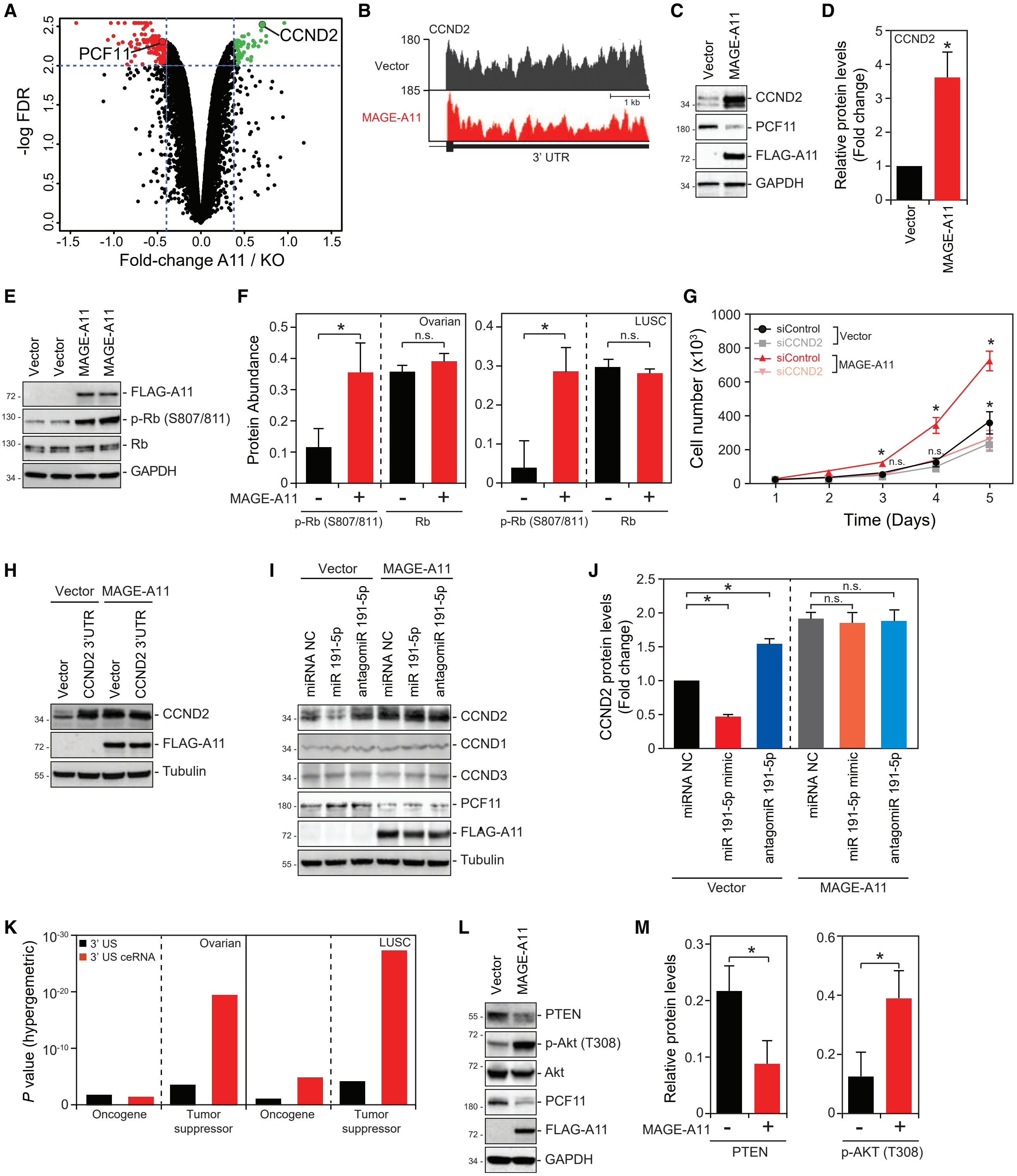


Figure 7. MAGE-A11-Induced 3′ UTR Shortening Modulates Oncogenes and Tumor Suppressors

(A) Quantitative whole-cell proteomics using TMT labeling (n = 5) revealed upregulation of CCND2 (cyclin D2) oncogene upon MAGE-A11 expression in DAOY MAGE-A11 KO cells.

(B) RNA-seq tracks for CCND2 showing reduced 3′ UTR reads in MAGE-A11-expressing cells.

(legend continued on next page)

D2, and FGF2, are more prominently detected in cancers compared to normal tissues (Mayr and Bartel, 2009). Furthermore, 3′ US of cyclin D1 in lymphomas correlates with increased cyclin D1 expression and proliferation of the lymphoma cells (Rosenwald et al., 2003). Interestingly, we found that MAGE-A11 induced 3′ US of cyclin D2, but not cyclin D1 or cyclin D3, resulting in increased protein products (Figures 7C, 7D, and 7I). These results suggest that MAGE-A11 may selectively regulate gene expression through modulation of APA leading to 3′ US in cancers. Second, we report that 3′ US possesses a significant role as ceRNAs for tumor-suppressor genes in *trans* (Park et al., 2018). Intriguingly, the ceRNA partners of 3′ US genes upon expression of MAGE-A11 are strongly enriched for tumor suppressors in lung squamous cell carcinoma ($p = 1.93^{-26}$) and ovarian cancer ($p = 7.71^{-21}$). Remarkably, these are notable tumor suppressors, such as PTEN, whose downregulation resulted in upregulation of Akt pro-survival signaling. These findings indicate MAGE-A11 may orchestrate gene expression changes in *cis* and in *trans* by modulating APA that results in reprogramming critical cellular signaling pathways, such as cell cycle and Akt signaling, to drive tumorigenesis. These findings may have important implications on therapeutic strategies for treating cancer, as MAGE-A11 expression status may confer predictive power to the response of cells against therapies, such as CDK4/6 inhibitors and AKT pathway inhibitors.

APA is known to be differentially regulated across tissue types and developmental stages such that an APA signature, ratio of distal versus proximal PAS choice, can be found. For example, compared to mammalian somatic cells, male germ cells have remarkable APA leading to 3′ US of many transcripts. In particular, PAS choice in male germ cells is often unique compared to somatic cells and results in testis-specific transcripts (Li et al., 2016; MacDonald, 2019; MacDonald and Redondo, 2002). It is not fully appreciated what leads to the widespread alternative PAS usage in germ cells leading to 3′ US but has been suggested to involve changes in the composition of the polyadenylation machinery, including CFIm (Edwalds-Gilbert et al., 1997; McMahon et al., 2006; Sartini et al., 2008; Takagaki and Manley, 1998). Our findings suggest that MAGE-A11–HUWE1 may be important factors in promoting APA in male germ cells. Consistently, HUWE1 has been shown to be important for spermatogonial differentiation and entry into meiosis (Bose et al., 2017). Furthermore, we suggest that the ability of MAGE-A11 to induce APA in cancer cells is not a neomorphic activity but rather is a conserved function of MAGE-A11 in cancer and germ cells.

Overall, our results suggest that dynamic regulation of the mRNA 3′ end processing machinery by ubiquitination can serve as a mechanism to control APA in various biological and pathological states.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Animals
  - ○ Cell lines
  - ○ Microbe strains
- METHOD DETAILS
  - ○ Cell culture transfections
  - ○ Generation of stable overexpression cell lines
  - ○ siRNA and miRNAs
  - ○ Tandem affinity purification
  - ○ RNA preparation and quantitative reverse transcription PCR Analysis (qRT-PCR)
  - ○ Clonogenic growth and anchorage-independent growth soft agar assays
  - ○ Immunoprecipitation and immunoblotting
  - ○ Recombinant protein purification and *in vitro* binding assay
  - ○ Tandem ubiquitin binding entity (TUBE) ubiquitination assay
  - ○ Cell viability assay
  - ○ Xenograft tumor growth assays
  - ○ LightSwitch luciferase reporter assay
  - ○ RNA-seq
  - ○ CRISPR/Cas9 genome editing
  - ○ TCGA 3′-UTR analysis

(C and D) MAGE-A11 overexpression increases cyclin D2 protein levels. HEK293FT cells stably expressing FLAG-vector or FLAG-MAGE-A11 were immunoblotted for the indicated proteins (C) and quantitated (D). Data are mean ± SD (n = 3). $^*p < 0.05$.

(E and F) MAGE-A11 induces phosphorylation of retinoblastoma (Rb). HEK293FT cells stably expressing FLAG-vector or FLAG-MAGE-A11 were immunoblotted for the indicated proteins (E), and ovarian cancer or lung squamous cell carcinoma patient samples with either low or high levels of MAGE-A11 were analyzed for phospho-Rb (S807/811) and total Rb protein levels (F). Data are mean ± SE of tumors indicated. $^*p < 0.01$.

(G) Depletion of CCND2 decreases the proliferation rate of MAGE-A11-re-expressing DAOY cells. MAGE-A11-knockout DAOY cells or those reconstituted with MAGE-A11 were transfected with the indicated siRNAs for 48 h counted for cell proliferation at the indicated time points. Data are mean ± SD. $^*p < 0.05$.

(H) *CCND2* 3′ UTR upregulates CCND2 in HEK293FT control cells, but not those stably expressing MAGE-A11. The indicated HEK293FT cells were transfected with vector control or *CCND2* 3′ UTR for 48 h and blotted for the indicated proteins.

(I and J) miR191-5p decreases CCND2 expression. HEK293FT cells stably expressing FLAG-vector or FLAG-MAGE-A11 were transfected with indicated miRNAs for 72 h, blotted for the indicated proteins (I), and quantitated (J). Data are mean ± SD (n = 3). $^*p < 0.05$.

(K) Oncogene or tumor suppressor gene enrichment of 3′ US mRNAs and 3′ US competing endogenous RNAs (ceRNAs) in ovarian cancer or lung squamous cell carcinoma patient samples with high MAGE-A11 expression levels. Top 10 tumors as high MAGE-A11 expression levels in two cancer types were analyzed; averaged p values with SD are plotted.

(L and M) MAGE-A11 represses PTEN protein levels through 3′ US in *trans*. HEK293FT cells stably expressing FLAG-vector or FLAG-MAGE-A11 were immunoblotted for the indicated proteins (L), and patient samples with either low (n = 101) or high levels of MAGE-A11 (n = 75) were analyzed for PTEN and phospho-Akt (T308) protein levels (M). Data are mean ± SE of tumors indicated. $^*p < 0.01$.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.molcel.2019.12.022.

ACKNOWLEDGMENTS

We thank members of the Potts lab for helpful discussions and critical reading of the manuscript. We would also like to thank Dr. George-Lucian Moldovan, Pennsylvania State University College of Medicine, for generously providing HEK293T HUWE1 knockout cells. This work was partially supported by a Worldwide Cancer Research grant 15-0177 (P.R.P.), American Cancer Society Research Scholar Award 181691010 (P.R.P.), and NIH R01 CA193466 (W.L.).

AUTHOR CONTRIBUTIONS

P.R.P. and S.W.Y. conceptualized the study and designed experiments. L.L. and W.L. designed the computational analyses. L.L., W.L., and H.G. analyzed RNA-seq data. K.K. and J.P. performed proteomics. K.F.T. and H.T. performed tissue gene expression and IHC analysis. S.N.P., J.P.C., and S.M.P.-M. performed gene editing. S.W.Y., J.M.P., and P.R.P. performed experiments and analyzed data. S.W.Y. and P.R.P. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 13, 2019
Revised: December 2, 2019
Accepted: December 23, 2019
Published: January 21, 2020

REFERENCES

Agarwal, V., Bell, G.W., Nam, J.W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. eLife *4*, e05005.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Baejen, C., Andreani, J., Torkler, P., Battaglia, S., Schwalb, B., Lidschreiber, M., Maier, K.C., Boltendahl, A., Rus, P., Esslinger, S., et al. (2017). Genome-wide analysis of RNA polymerase II termination at protein-coding genes. Mol. Cell *66*, 38–49.e6.

Bai, S., He, B., and Wilson, E.M. (2005). Melanoma antigen gene protein MAGE-11 regulates androgen receptor function by modulating the interdomain interaction. Mol. Cell. Biol. *25*, 1238–1257.

Bai, B., Tan, H., Pagala, V.R., High, A.A., Ichhaporia, V.P., Hendershot, L., and Peng, J. (2017). Deep profiling of proteome and phosphoproteome by isobaric labeling, extensive liquid chromatography, and mass spectrometry. Methods Enzymol. *585*, 377–395.

Bose, R., Sheng, K., Moawad, A.R., Manku, G., O'Flaherty, C., Taketo, T., Culty, M., Fok, K.L., and Wing, S.S. (2017). Ubiquitin ligase Huwe1 modulates spermatogenesis by regulating spermatogonial differentiation and entry into meiosis. Sci. Rep. *7*, 17759.

Choe, K.N., Nicolae, C.M., Constantin, D., Imamura Kawasawa, Y., Delgado-Diaz, M.R., De, S., Freire, R., Smits, V.A., and Moldovan, G.L. (2016). HUWE1 interacts with PCNA to alleviate replication stress. EMBO Rep. *17*, 874–886.

Ciolli Mattioli, C., Rom, A., Franke, V., Imami, K., Arrey, G., Terne, M., Woehler, A., Akalin, A., Ulitsky, I., and Chekulaeva, M. (2019). Alternative 3′ UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. Nucleic Acids Res. *47*, 2560–2573.

Connelly, J.P., and Pruett-Miller, S.M. (2019). CRIS.py: a versatile and high-throughput analysis program for CRISPR-based genome editing. Sci. Rep. *9*, 4194.

Davoli, T., Xu, A.W., Mengwasser, K.E., Sack, L.M., Yoon, J.C., Park, P.J., and Elledge, S.J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell *155*, 948–962.

Derti, A., Garrett-Engele, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M., and Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. Genome Res. *22*, 1173–1183.

Di Leva, G., Piovan, C., Gasparini, P., Ngankeu, A., Taccioli, C., Briskin, D., Cheung, D.G., Bolon, B., Anderlucci, L., Alder, H., et al. (2013). Estrogen mediated-activation of miR-191/425 cluster modulates tumorigenicity of breast cancer cells depending on estrogen receptor status. PLoS Genet. *9*, e1003311.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Doyle, J.M., Gao, J., Wang, J., Yang, M., and Potts, P.R. (2010). MAGE-RING protein complexes comprise a family of E3 ubiquitin ligases. Mol. Cell *39*, 963–974.

Edwalds-Gilbert, G., Veraldi, K.L., and Milcarek, C. (1997). Alternative poly(A) site selection in complex transcription units: means to an end? Nucleic Acids Res. *25*, 2547–2561.

Elkon, R., Ugalde, A.P., and Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. Nat. Rev. Genet. *14*, 496–506.

Evron, E., Umbricht, C.B., Korz, D., Raman, V., Loeb, D.M., Niranjan, B., Buluwela, L., Weitzman, S.A., Marks, J., and Sukumar, S. (2001). Loss of cyclin D2 expression in the majority of breast cancers is associated with promoter hypermethylation. Cancer Res. *61*, 2782–2787.

Feng, X., Li, L., Wagner, E.J., and Li, W. (2018). TC3A: The Cancer 3′ UTR Atlas. Nucleic Acids Res. *46* (D1), D1027–D1030.

Flavell, S.W., Kim, T.K., Gray, J.M., Harmin, D.A., Hemberg, M., Hong, E.J., Markenscoff-Papadimitriou, E., Bear, D.M., and Greenberg, M.E. (2008). Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. Neuron *60*, 1022–1038.

Fu, Y., Sun, Y., Li, Y., Li, J., Rao, X., Chen, C., and Xu, A. (2011). Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome Res. *21*, 741–747.

Goldman, M., Craft, B., Swatloski, T., Cline, M., Morozova, O., Diekhans, M., Haussler, D., and Zhu, J. (2015). The UCSC Cancer Genomics Browser: update 2015. Nucleic Acids Res. *43*, D812–D817.

Hao, Y.H., Doyle, J.M., Ramanathan, S., Gomez, T.S., Jia, D., Xu, M., Chen, Z.J., Billadeau, D.D., Rosen, M.K., and Potts, P.R. (2013). Regulation of WASH-dependent actin polymerization and protein trafficking by ubiquitination. Cell *152*, 1051–1064.

Hao, Y.H., Fountain, M.D., Jr., Fon Tacer, K., Xia, F., Bi, W., Kang, S.H., Patel, A., Rosenfeld, J.A., Le Caignec, C., Isidor, B., et al. (2015). USP7 acts as a molecular rheostat to promote WASH-dependent endosomal protein recycling and is mutated in a human neurodevelopmental disorder. Mol. Cell *59*, 956–969.

Hoffman, Y., Bublik, D.R., Ugalde, A.P., Elkon, R., Biniashvili, T., Agami, R., Oren, M., and Pilpel, Y. (2016). 3'UTR shortening potentiates microRNA-based repression of pro-differentiation genes in proliferating human cells. PLoS Genet. *12*, e1005879.

Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J.Y., Yehia, G., and Tian, B. (2013). Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. Nat. Methods *10*, 133–139.

Ji, Z., and Tian, B. (2009). Reprogramming of 3′ untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. PLoS ONE *4*, e8419.

Ji, Z., Lee, J.Y., Pan, Z., Jiang, B., and Tian, B. (2009). Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. Proc. Natl. Acad. Sci. USA *106*, 7028–7033.

Johnson, S.A., Cubberley, G., and Bentley, D.L. (2009). Cotranscriptional recruitment of the mRNA export factor Yra1 by direct interaction with the 3′ end processing factor Pcf11. Mol. Cell *33*, 215–226.

Kamieniarz-Gdula, K., Gdula, M.R., Panser, K., Nojima, T., Monks, J., Wiśniewski, J.R., Riepsaame, J., Brockdorff, N., Pauli, A., and Proudfoot, N.J. (2019). Selective roles of vertebrate PCF11 in premature and full-length transcript termination. Mol. Cell *74*, 158–172.e9.

Lee, A.K., and Potts, P.R. (2017). A comprehensive guide to the MAGE family of ubiquitin ligases. J. Mol. Biol. *429*, 1114–1142.

Li, W., You, B., Hoque, M., Zheng, D., Luo, W., Ji, Z., Park, J.Y., Gunderson, S.I., Kalsotra, A., Manley, J.L., and Tian, B. (2015). 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. *11*, e1005166.

Li, W., Park, J.Y., Zheng, D., Hoque, M., Yehia, G., and Tian, B. (2016). Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control. BMC Biol. *14*, 6.

Lian, Y., Sang, M., Ding, C., Zhou, X., Fan, X., Xu, Y., Lü, W., and Shan, B. (2012). Expressions of MAGE-A10 and MAGE-A11 in breast cancers and their prognostic significance: a retrospective clinical study. J. Cancer Res. Clin. Oncol. *138*, 519–527.

Licatalosi, D.D., Geiger, G., Minet, M., Schroeder, S., Cilli, K., McNeil, J.B., and Bentley, D.L. (2002). Functional interaction of yeast pre-mRNA 3′ end processing factors with RNA polymerase II. Mol. Cell *9*, 1101–1111.

MacDonald, C.C. (2019). Tissue-specific mechanisms of alternative polyadenylation: Testis, brain, and beyond (2018 update). Wiley Interdiscip. Rev. RNA *10*, e1526.

MacDonald, C.C., and Redondo, J.L. (2002). Reexamining the polyadenylation signal: were we wrong about AAUAAA? Mol. Cell. Endocrinol. *190*, 1–8.

Mallon, B.S., and Macklin, W.B. (2002). Overexpression of the 3′-untranslated region of myelin proteolipid protein mRNA leads to reduced expression of endogenous proteolipid mRNA. Neurochem. Res. *27*, 1349–1360.

Martin, G., Gruber, A.R., Keller, W., and Zavolan, M. (2012). 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. *1*, 753–763.

Masamha, C.P., Xia, Z., Yang, J., Albrecht, T.R., Li, M., Shyu, A.B., Li, W., and Wagner, E.J. (2014). CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. Nature *510*, 412–416.

Matoulkova, E., Michalova, E., Vojtesek, B., and Hrstka, R. (2012). The role of the 3′ untranslated region in post-transcriptional regulation of protein expression in mammalian cells. RNA Biol. *9*, 563–576.

Mayr, C., and Bartel, D.P. (2009). Widespread shortening of 3’UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell *138*, 673–684.

McMahon, K.W., Hirsch, B.A., and MacDonald, C.C. (2006). Differences in polyadenylation site choice between somatic and male germ cells. BMC Mol. Biol. *7*, 35.

Meinhart, A., and Cramer, P. (2004). Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors. Nature *430*, 223–226.

Mermelshtein, A., Gerson, A., Walfisch, S., Delgado, B., Shechter-Maor, G., Delgado, J., Fich, A., and Gheber, L. (2005). Expression of D-type cyclins in colon cancer and in cell lines from colon carcinomas. Br. J. Cancer *93*, 338–345.

Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J.O., and Lai, E.C. (2013). Widespread and extensive lengthening of 3′ UTRs in the mammalian brain. Genome Res. *23*, 812–825.

Narasimha, A.M., Kaulich, M., Shapiro, G.S., Choi, Y.J., Sicinski, P., and Dowdy, S.F. (2014). Cyclin D activates the Rb tumor suppressor by monophosphorylation. eLife *3*, e02872.

Newman, J.A., Cooper, C.D., Roos, A.K., Aitkenhead, H., Oppermann, U.C., Cho, H.J., Osman, R., and Gileadi, O. (2016). Structures of two melanoma-associated antigens suggest allosteric regulation of effector binding. PLoS ONE *11*, e0148762.

Niu, M., Cho, J.H., Kodali, K., Pagala, V., High, A.A., Wang, H., Wu, Z., Li, Y., Bi, W., Zhang, H., et al. (2017). Extensive peptide fractionation and y_1 ion-based interference detection method for enabling accurate quantification by isobaric labeling and mass spectrometry. Anal. Chem. *89*, 2956–2963.

O’Connor, H.F., Lyon, N., Leung, J.W., Agarwal, P., Swaim, C.D., Miller, K.M., and Huibregtse, J.M. (2015). Ubiquitin-activated interaction traps (UBAITs) identify E3 ligase binding partners. EMBO Rep. *16*, 1699–1712.

Ogorodnikov, A., Levin, M., Tattikota, S., Tokalov, S., Hoque, M., Scherzinger, D., Marini, F., Poetsch, A., Binder, H., Macher-Göppinger, S., et al. (2018). Transcriptome 3’end organization by PCF11 links alternative polyadenylation to formation and neuronal differentiation of neuroblastoma. Nat. Commun. *9*, 5331.

Pagala, V.R., High, A.A., Wang, X., Tan, H., Kodali, K., Mishra, A., Kavdia, K., Xu, Y., Wu, Z., and Peng, J. (2015). Quantitative protein analysis by mass spectrometry. Methods Mol. Biol. *1278*, 281–305.

Panwar, B., Omenn, G.S., and Guan, Y. (2017). miRmine: a database of human miRNA expression profiles. Bioinformatics *33*, 1554–1560.

Park, H.J., Ji, P., Kim, S., Xia, Z., Rodriguez, B., Li, L., Su, J., Chen, K., Masamha, C.P., Baillat, D., et al. (2018). 3′ UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk. Nat. Genet. *50*, 783–789.

Pineda, C.T., Ramanathan, S., Fon Tacer, K., Weon, J.L., Potts, M.B., Ou, Y.H., White, M.A., and Potts, P.R. (2015). Degradation of AMPK by a cancer-specific ubiquitin ligase. Cell *160*, 715–728.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rosenwald, A., Wright, G., Wiestner, A., Chan, W.C., Connors, J.M., Campo, E., Gascoyne, R.D., Grogan, T.M., Muller-Hermelink, H.K., Smeland, E.B., et al. (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. Cancer Cell *3*, 185–197.

Rutnam, Z.J., and Yang, B.B. (2012). The non-coding 3′ UTR of CD44 induces metastasis by regulating extracellular matrix functions. J. Cell Sci. *125*, 2075–2085.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell *146*, 353–358.

Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A., and Burge, C.B. (2008). Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. Science *320*, 1643–1647.

Sartini, B.L., Wang, H., Wang, W., Millette, C.F., and Kilpatrick, D.L. (2008). Pre-messenger RNA cleavage factor I (CFIm): potential role in alternative polyadenylation during spermatogenesis. Biol. Reprod. *78*, 472–482.

Sherr, C.J. (1995). D-type cyclins. Trends Biochem. Sci. *20*, 187–190.

Shi, Y., Di Giammartino, D.C., Taylor, D., Sarkeshik, A., Rice, W.J., Yates, J.R., 3rd, Frank, J., and Manley, J.L. (2009). Molecular architecture of the human pre-mRNA 3′ processing complex. Mol. Cell *33*, 365–376.

Sicinski, P., Donaher, J.L., Geng, Y., Parker, S.B., Gardner, H., Park, M.Y., Robker, R.L., Richards, J.S., McGinnis, L.K., Biggers, J.D., et al. (1996). Cyclin D2 is an FSH-responsive gene involved in gonadal cell proliferation and oncogenesis. Nature *384*, 470–474.

Simpson, A.J., Caballero, O.L., Jungbluth, A., Chen, Y.T., and Old, L.J. (2005). Cancer/testis antigens, gametogenesis and cancer. Nat. Rev. Cancer *5*, 615–625.

Su, S., Minges, J.T., Grossman, G., Blackwelder, A.J., Mohler, J.L., and Wilson, E.M. (2013). Proto-oncogene activity of melanoma antigen-A11 (MAGE-A11) regulates retinoblastoma-related p107 and E2F1 proteins. J. Biol. Chem. *288*, 24809–24824.

Takagaki, Y., and Manley, J.L. (1998). Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. Mol. Cell *2*, 761–771.

Takano, Y., Kato, Y., Masuda, M., Ohshima, Y., and Okayasu, I. (1999). Cyclin D2, but not cyclin D1, overexpression closely correlates with gastric cancer progression and prognosis. J. Pathol. *189*, 194–200.

Takano, Y., Kato, Y., van Diest, P.J., Masuda, M., Mitomi, H., and Okayasu, I. (2000). Cyclin D2 overexpression and lack of p27 correlate positively and cyclin E inversely with a poor prognosis in gastric cancer cases. Am. J. Pathol. *156*, 585–594.

Tian, B., and Manley, J.L. (2017). Alternative polyadenylation of mRNA precursors. Nat. Rev. Mol. Cell Biol. *18*, 18–30.

Volanakis, A., Kamieniarz-Gdula, K., Schlackow, M., and Proudfoot, N.J. (2017). WNK1 kinase and the termination factor PCF11 connect nuclear mRNA export with transcription. Genes Dev. *31*, 2175–2185.

Wang, X., Li, Y., Wu, Z., Wang, H., Tan, H., and Peng, J. (2014). JUMP: a tag-based database search tool for peptide identification with high sensitivity and accuracy. Mol. Cell. Proteomics *13*, 3663–3673.

Wang, R., Nambiar, R., Zheng, D., and Tian, B. (2018). PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. Nucleic Acids Res. *46* (D1), D315–D319.

Weon, J.L., and Potts, P.R. (2015). The MAGE protein family and cancer. Curr. Opin. Cell Biol. *37*, 1–8.

Weon, J.L., Yang, S.W., and Potts, P.R. (2018). Cytosolic iron-sulfur assembly is evolutionarily tuned by a cancer-amplified ubiquitin ligase. Mol. Cell *69*, 113–125.e6.

Xia, L.P., Xu, M., Chen, Y., and Shao, W.W. (2013). Expression of MAGE-A11 in breast cancer tissues and its effects on the proliferation of breast cancer cells. Mol. Med. Rep. *7*, 254–258.

Xia, Z., Donehower, L.A., Cooper, T.A., Neilson, J.R., Wheeler, D.A., Wagner, E.J., and Li, W. (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. Nat. Commun. *5*, 5274.

Yao, C., Biesinger, J., Wan, J., Weng, L., Xing, Y., Xie, X., and Shi, Y. (2012). Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. Proc. Natl. Acad. Sci. USA *109*, 18773–18778.

Ye, C., Long, Y., Ji, G., Li, Q.Q., and Wu, X. (2018). APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. Bioinformatics *34*, 1841–1849.

Yoon, J.H., and Gorospe, M. (2016). Cross-linking immunoprecipitation and qPCR (CLIP-qPCR) analysis to map interactions between long noncoding RNAs and RNA-binding proteins. Methods Mol. Biol. *1402*, 11–17.

Zhu, Y., Wang, X., Forouzmand, E., Jeong, J., Qiao, F., Sowd, G.A., Engelman, A.N., Xie, X., Hertel, K.J., and Shi, Y. (2018). Molecular mechanisms for CFIm-mediated regulation of mRNA alternative polyadenylation. Mol. Cell *69*, 62–74.e4.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit polyclonal anti-MAGE-A11	This paper	N/A
Rabbit polyclonal anti-PCF11	Bethyl Laboratories	Cat# A303-706A; RRID:AB_11204946
Rabbit polyclonal anti-RNAP II	Cell Signaling Technology	Cat# 14958; RRID:AB_2687876
Rabbit monoclonal anti- anti-CLP1	Abcam	N/A
Rabbit monoclonal anti- anti-HUWE1	Novus Biologicals	Cat# NB100-652; RRID:AB_2264587
Rabbit monoclonal anti-GAPDH	Cell Signaling Technology	Cat# 2118; RRID:AB_561053
Mouse monoclonal anti-FLAG	Sigma-Aldrich	Cat# F3165; RRID:AB_259529
Rabbit polyclonal anti-Myc	Sigma-Aldrich	Cat# C3956; RRID:AB_439680
Mouse monoclonal anti-Actin	Abcam	Cat# ab6276; RRID:AB_2223210
Mouse monoclonal anti-Tubulin	Sigma-Aldrich	Cat# T5168; RRID:AB_477579
Rabbit polyclonal anti-CCND1	ABclonal	Cat# A11022; RRID:AB_2758370
Rabbit polyclonal anti-CCND2	ABclonal	Cat# A1773; RRID:AB_2763815
Rabbit polyclonal anti-CCND3	ABclonal	Cat# A0746; RRID:AB_2757375
Rabbit polyclonal anti-phospho-Rb S807/811	Cell Signaling Technology	Cat# 9308; RRID:AB_331472
Mouse monoclonal anti-Rb	Cell Signaling Technology	Cat# 9309; RRID:AB_823629
Rabbit polyclonal anti-PTEN	Bethyl Laboratories	Cat# A300-701A; RRID:AB_2174186
Rabbit monoclonal anti-phospho-Akt T308	Cell Signaling Technology	Cat# 4056; RRID:AB_331163
Rabbit monoclonal anti-Akt	Cell Signaling Technology	Cat# 4691; RRID:AB_915783
Rabbit polyclonal anti-CPSF160	Bethyl Laboratories	Cat# A301-580A; RRID:AB_1078859
Rabbit polyclonal anti-CPSF100	Thermo Fisher Scientific	Cat# A301-581A; RRID:AB_1078861
Rabbit polyclonal anti-CPSF73	Bethyl Laboratories	Cat# A301-091A; RRID:AB_2084528
Rabbit polyclonal anti-CSTF64	Bethyl Laboratories	Cat# A301-092A; RRID:AB_873014
Rabbit anti-TauCSTF64	Bethyl Laboratories	Cat# A301-487A; RRID:AB_999545
Rabbit polyclonal anti-CPSF68	Bethyl Laboratories	Cat# A301-356A; RRID:AB_937781
Rabbit polyclonal anti-CPSF59	Bethyl Laboratories	Cat# A301-359A; RRID:AB_937869
Mouse monoclonal anti-NUDT21	Proteintech	Cat# 66335-1-Ig; RRID: N/A
Rabbit anti-XRN2	Bethyl Laboratories	Cat# A301-103A; RRID:AB_2218876
Mouse IgG control	Santa Cruz Biotechnology	Cat# sc-2025; RRID:AB_737182
Donkey Anti-Rabbit IgG, HRP Conjugated	GE Healthcare	Cat# NA934; RRID:AB_772206
Sheep Anti-Mouse IgG, HRP Conjugated	GE Healthcare	Cat# NA931; RRID:AB_772210
Bacterial and Virus Strains		
DH5a Competent Cells	Thermo Fisher Scientific	Cat# 18265017
XL1-blue Competent Cells	Agilent Technologies	Cat# 200130
One shot Stbl3	Life Technologies	Cat# C7373-03
BL21-codon plus(DE3)-RILP	Agilent Technologies	Cat# 230280
Chemicals, Peptides, and Recombinant Proteins		
ECL detection reagent	GE Healthcare	Cat# RPN2209
ECL prime detection reagent	GE Healthcare	Cat# RPN2236
Protein A beads	Bio-Rad	Cat# 1560005
Protein G Agarose	Thermo Fisher Scientific	Cat# 20389
Protein A/G PLUS agarose	Santa Cruz Biotechnology	Cat# sc-2003
Anti-FLAG M2 Beads	Sigma	Cat# A2220

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
TUBE2-Agarose	LifeSensors	Cat# UM402
Effectene transfection reagent	QIAGEN	Cat# 301425
Lipofectamine RNAiMAX	Thermo Fisher Scientific	Cat# 13778030
Lipofectamine 2000	Invitrogen	Cat# 11668027
RNAStat60	TelTest	Cat# Cs-112
RNeasy Kit	QIAGEN	Cat# 74104
TEV Protease	Sigma-Aldrich	Cat# T4455
Calmodulin-Sepharose 4B	GE Healthcare	Cat# 17-0529-01
Glutathione Sepharose 4B	GE Healthcare	Cat# 17-0756-05
Critical Commercial Assays		
High Capacity cDNA Reverse Transcription kit	Thermo Fisher Scientific	Cat# 4368813
TNT SP6 Quick *In Vitro* TNT Kit	Promega	Cat# L2080
BCA protein assay kit	Thermo Fisher Scientific	Cat# 23227
AlamarBlue® Cell Viability Reagent	Thermo Fisher Scientific	Cat# DAL1100
Experimental Models: Cell Lines		
HEK293	ATCC	Cat# CRL-1573
HEK293T	Choe et al., 2016	N/A
HEK293T HUWE1 KO	Choe et al., 2016	N/A
HEK293FT	Thermo Fisher Scientific	Cat# R70007
A2780	A gift from Michael White, UT Southwestern	N/A
DAOY	ATCC	Cat# HTB-186
H520	A gift from John Minna UT Southwestern	N/A
OV56	A gift from Michael White, UT Southwestern	N/A
DAOY MAGE-A11 KO	This paper	N/A
H520 MAGE-A11 KO	This paper	N/A
Experimental Models: Organisms/Strains		
NOD.Cg-Prkdcscid Il2rgtm1Wjl/SzJ: NOD scid gamma mice	The Jackson Laboratory	005557; RRID: IMSR_JAX:005557
Oligonucleotides		
See Table S5	This paper	N/A
Software and Algorithms		
ImageJ software	ImageJ	https://imagej.nih.gov/ij
GraphPad Prism 7	GraphPad	https://www.graphpad.com
STAR version 2.5.2b	Dobin et al., 2013	https://github.com/alexdobin/STAR
DaPars	Xia et al., 2014	https://github.com/ZhengXia/dapars
MAT3UTR	Park et al., 2018	https://github.com/thejustpark/MAT3UTR
DESeq2	Anders and Huber, 2010	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
HTSeq	Anders et al., 2015	https://htseq.readthedocs.io/en/release_0.11.1/count.html
Deposited Data		
Sequencing data	NCBI Gene Expression Omnibus	GEO: GSE134898
Proteomics data	MassIVE UCSD	MSV000084123

LEAD CONTACT AND MATERIALS AVAILABILITY

All materials generated in this study are available through request to Lead Contact Patrick Ryan Potts (ryan.potts@stjude.org).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Animals

6-8 week old male NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ (NOD scid gamma) mice from Jackson Labs were used for xenograft growth assays. Animals were group housed under standard conditions. All studies were approved by the St. Jude Children's Research Hospital institutional review committee on animal safety.

Cell lines

HEK293FT, HEK293T, HEK293, A2780, and DAOY cells were grown in DMEM supplemented with 10% (v/v) FBS, 2 mM L-glutamine, 100 units/mL penicillin, 100 units/mL streptomycin, and 0.25 mg/mL amphotericin B. H520 cells were grown in RPMI supplemented with 5% (v/v) heat inactivated serum. OV56 cells were grown in DMEM:HAMS F12 (1:1) supplemented with 5% (v/v) FBS, 2 mM L-glutamine, 0.5 μg/mL hydrocortisone, and 10 μg/mL insulin. HEK293T HUWE1 knockout cells were described previously (Choe et al., 2016). Cell lines were authenticated by STR analysis. Sex of cells used: Female, HEK293FT, HEK293T, HEK293, A2780, OV56; Male, DAOY, H520. All cells were maintained at 37°C in 5% CO_2.

Microbe strains

DH5a (Thermo Fisher Scientific, #18265017) and XL1-blue (Agilent #200130) competent cells were used for standard molecular cloning and plasmid amplification. One shot Stbl3 competent cells (Life Technologies #C7373-03) were used for lentiviral plasmid cloning and plasmid amplification. BL21-codon plus (DE3)-RILP competent cells (Agilent Technologies, #230280) were used for recombinant protein production and purification. Bacteria were cultured under standard conditions at 37°C, 225 rpm.

METHOD DETAILS

Cell culture transfections

siRNA and plasmid transfections were performed using Lipofectamine RNAiMAX, Lipofectamine 2000 (Invitrogen) and Effectene (QIAGEN) according to the manufacturers' protocol.

Generation of stable overexpression cell lines

HEK293FT and DAOY cells were transfected with either HA-FLAG-vector or HA-FLAG-MAGE-A11 using Effectene according to the manufacturer's protocol in 6 cm^2 plates. After 48 hours, cells were selected with 1 μg/mL of puromycin over 2 weeks. HEK293 cells were transfected with either tandem affinity purification (TAP)-vector or TAP-MAGE-A11 using Effectene in 6 cm^2 plates. After 48 hours, cells were selected with 1 μg/mL of puromycin over 2 weeks. HEK293FT cells were transfected with TAP-MAGE-A11-UBAIT or TAP-MAGE-A11-UBAIT GG deletion using Effectene in 6 cm^2 plates. After 48 hours, cells were selected with 1 μg/mL of puromycin over 2 weeks. A2780, H520 and OV56 cells were transduced with Myc-vector or Myc-MAGE-A11 lentivirus using polybrene in 6-well plates. Two days after lentiviral transduction, cells were selected over 2 weeks using 2.5 μg/mL of blasticidin (GIBCO).

siRNA and miRNAs

siRNA transfections were performed using Lipofectamine RNAiMAX according to the manufacturer's protocol. All siRNAs were purchased from Sigma. siRNA targeting sequences; siControl, 5′-ACUACAUCGUGAUUCAAACUU; siMAGE-A11 #1, 5′- CAAGAUAAUUGAUUUGGUU; siMAGE-A11 #2, 5′-CUGAUAGACCCUGAGUCCU; siPCF11 #1, 5′-GUACCUUAUGGAUUCUAUU; siPCF11 #2, 5′- GUAUCUCACUGCCUUUACU; siPCF11 #3, 5′- CAACGUUAUGACAGUGUUA; siPCF11 #4, 5′-CAAUUGUUCCUGAUAUACA. siCCND2 #1, 5′-CUCAUGACUUCAUUGAGCA; siCCND2 #2, 5′-CUGUGUGCCACCGACUUUA; siCCND2 #3, 5′-GAGGAAGUGAGCUCGCUCA; siHUWE1 #1, 5′- CAUGAGACAUCAGCCCACCCUUAAAA; siHUWE1 #2, and 5′- CACACCAGCAAUGGCUGCCAGAAUU. miRNA mimetics were purchased from Sigma. miRNA antagomirs were purchased from Applied Biological Materials Inc.

Antibodies

MAGE-A11 rabbit polyclonal antibody was generated against bacterially expressed MAGE-A11 (amino acids 197-429) (Cocalico Biologicals, Inc). Commercial antibodies used: anti-PCF11 (Bethyl Laboratories, A303-706A), anti-RNAP II (Cell Signaling Technology, 14958S), anti-CLP1 (Abcam, ab133669), anti-HUWE1 (Novus Biologicals, NB100-652), anti-GAPDH (Cell Signaling Technology, 2118S), anti-FLAG (Sigma, F3165), anti-Myc (Roche, 11666606001), anti-Actin (Abcam, ab6276), anti-Tubulin (Sigma, T5168), anti-CCND1 (ABclonal, A11022), anti-CCND2 (ABclonal, A1773), anti-CCND3 (ABclonal, A0746), anti-phospho-Rb S807/811 (Cell Signaling Technology, 9308T), anti-Rb (Cell Signaling Technology, 9309T), anti-PTEN (Bethyl Laboratories, A300-701A), anti-phospho-Akt T308 (Cell Signaling Technology, 4056S), anti-Akt (Cell Signaling Technology, 4691S), anti-CPSF160 (Bethyl Laboratories, A301-580A), anti-CPSF100 (Bethyl Laboratories, A301-581A), anti-CPSF73 (Bethyl Laboratories, A301-091A), anti-CstF64 (Bethyl Laboratories, A301-092A), anti-TauCstF64 (Bethyl Laboratories, A301-487A), anti-CFIm68 (Bethyl Laboratories, A301-356A), anti-CFIm59 (Bethyl Laboratories, A301-359A), anti-CFIm25 (Proteintech, 66335-1-Ig), anti-XRN2 (Bethyl Laboratories, A301-103A), Donkey anti-Rabbit IgG (GE, NA934V), and Sheep anti-Mouse IgG (GE, NA931V).

Tandem affinity purification

Ten 15 cm^2 plates of HEK293 cells stably expressing TAP-vector or TAP-MAGE-A11 were lysed with TAP lysis buffer (10% (v/v) glycerol, 50 mM HEPES-KOH pH 7.5, 100 mM KCl, 2 mM EDTA, 0.1% (v/v) NP-40, 10 mM NaF, 0.25 mM NA_3VO_4, 50 mM β-glyerolphosphate, 2 mM DTT, and 1X protease inhibitor cocktail) and cleared supernatants were bound to IgG-Sepharose beads (GE Amersham) and then washed in lysis buffer and TEV buffer (10 mM HEPES-KOH pH 8.0, 150 mM NaCl, 0.1% NP-40, 0.5 mM EDTA, 1 mM DTT, and 1X protease inhibitor cocktail). Protein complexes were cleaved off the beads by TEV protease and incubated with calmodulin-Sepharose beads (GE Amersham) in calmodulin binding buffer (10 mM HEPES-KOH pH 8.0, 150 mM NaCl, 1 mM Mg acetate, 1 mM imidazole, 0.1% NP-40, 6 mM $CaCl_2$, 10 mM 2-mercaptoethanol) and then washed, eluted with SDS sample buffer, subjected to SDS-PAGE, and stained with GelCode Blue stain (Thermo Fisher Scientific) before protein identification by LC-MS/MS.

RNA preparation and quantitative reverse transcription PCR Analysis (qRT-PCR)

RNA was extracted from cultured cells using RNAStat60 (TelTest) according to manufacturer's instructions. Total RNA was treated with DNase I (Roche) and converted to cDNA using High Capacity Reverse Transcription kit (Life Technologies). cDNA and appropriate primers were plated in triplicate in a 384-well plate and gene expression levels were measured using SYBR green master mix (Applied Biosystems). Oligonucleotides used for qRT-PCR: *MAGE-A11* forward, 5′-GAGGATCACTGGAGGAGAACA; reverse, 5′-TCTTTGCTCAAGAGGCATGAT; *CCND2* forward, 5′-TTCCCTCTGGCCATGAATTAC; reverse, 5′-GGGCTGGTCTCTTTGAGTTT; *CCND2* 3′-UTR forward #1, 5′-CTTCTGGTATCTGGCGTTCTT; reverse #1, 5′-CAGGCTTGTCTGAGGAATGT; *CCND2* 3′-UTR forward #2, 5′-GGACACCTTGTGTTTAGGATCA; reverse #2, 5′-GGGAGAAGGAAGCACCATAAA; *CCND2* 3′-UTR forward #3, 5′-CAAGCCTACCCGACTCTATTTAC; reverse #3, 5′-CCCAAGGATGGGAAAGAGAAA; *CCND2* 3′-UTR forward #4, 5′-TACTGGGT CATCCTTGGTCTAT; reverse #4, 5′-TTGTCTTCTCCTCTGGCTTTG.

Clonogenic growth and anchorage-independent growth soft agar assays

For clonogenic growth assays on plates, wild-type or knockout cells were plated in 6-well plates in triplicate. After 2-3 weeks, cells were fixed and stained with 0.05% (w/v) crystal violet and counted. For anchorage-independent growth soft agar assays, cells were suspended in 0.375% Noble agar (Difco) supplemented with regular growth medium and overlaid on 0.5% Noble agar. Cells were incubated for 2-4 weeks before colonies $\geq$ 100 μm in size were counted.

Immunoprecipitation and immunoblotting

HEK293FT cells were plated in 6 cm^2 plates and transfected 24 hours later with Effectene (QIAGEN) according to the manufacturer's protocol. After 48 hours, cells were washed and scraped in cold PBS, spun down, and resuspended in lysis buffer (50 mM Tris pH 7.4, 150 mM NaCl, 1 mM DTT, 0.1% (v/v) Triton X-100, 10 mM N-Ethylmaleimide (NEM), and 1X protease inhibitor cocktail). Cell lysates were incubated with appropriate antibodies overnight at 4°C and then with protein A beads for 2 hours at 4°C. Beads were then washed with lysis buffer three times and eluted with 2X SDS sample buffer. For immunoblotting, samples in SDS sample buffer were resolved on SDS-PAGE gels and then transferred to nitrocellulose membranes prior to blocking in TBST with 5% (w/v) milk powder or 3% (w/v) bovine serum albumin and probing with primary and secondary antibodies (GE Healthcare). Protein signal was visualized after addition of ECL detection reagent (GE Healthcare) according to manufacturer's instructions.

Recombinant protein purification and *in vitro* binding assay

GST-PCF11, GST-CLP1 or GST tag alone were induced in BL21 (DE3) cells at 16°C with isopropyl β-D-1-thiogalactopyranoside (IPTG). GST-tagged proteins were purified from bacterial lysates in lysis buffer (50 mM Tris pH 7.7, 150 mM KCl, 0.1% (v/v) Triton X-100, 1 mM DTT, 1 mg/mL lysozyme) with glutathione Sepharose (GE Amersham) and eluted with 10 mM glutathione. For *in vitro* binding assay, Myc-tagged proteins were *in vitro* translated using the SP6-TNT Quick rabbit reticulocyte lysate system (Promega) according to manufacturer's instructions. Purified GST-tagged proteins were bound to glutathione Sepharose beads for 1 hour in binding buffer (25 mM Tris pH 8.0, 2.7 mM KCl, 137 mM NaCl, 0.05% (v/v) Tween-20, and 10 mM 2-mercaptoethanol) and then were blocked for 1 hr in binding buffer containing 5% (w/v) milk powder. *In vitro* translated proteins were then incubated with the bound beads for 1 hour, washed, and the proteins were eluted in SDS-sample buffer, boiled, and subjected to SDS-PAGE and immunoblotting.

Tandem ubiquitin binding entity (TUBE) ubiquitination assay

HEK293FT, DAOY or H520 cells (1-2 10 cm^2 plates) were lysed with TUBE lysis buffer (50 mM Tris pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% (v/v) NP-40, 10% (v/v) glycerol, 20 mM N-Ethylmaleimide (NEM), and 1X protease inhibitor cocktail), and the lysates were bound to TUBE-agarose (LifeSensors) overnight at 4°C. Beads were subsequently washed three times in wash buffer (20 mM Tris pH 8.0, 150 mM NaCl, 0.1% (v/v) Tween 20) and then the ubiquitinated proteins were eluted in SDS-sample buffer, boiled, and subjected to SDS-PAGE and immunoblotting.

Cell viability assay

To assess cell viability after siRNA knockdown, 1 X 10^4 cells/mL were transfected with 50 nM siRNA using Lipofectamine RNAiMAX according to the manufacturer's protocol and incubated for 72-96 hours prior to changing the media and adding alamarBlue (Thermo

Fisher Scientific) and incubating for 4 hours at 37°C. Plates were read by measuring the fluorescence with excitation wavelength at 540 nm and emission wavelength at 590 nm on an Enspire plate reader.

Xenograft tumor growth assays

3 X 10^6 DAOY wild-type and MAGE-A11 knockout cells were mixed with matrigel (Corning) before injection into the flank of NOD scid gamma mice (Jackson Lab) (n = 6 for wild-type group, n = 12 for MAGE-A11-knockout group). MAGE-A11 negative A2780 and OV56 ovarian cancer cells were made to stably express Myc-vector or Myc-MAGE-A11 before injection of 3 X 10^6 cells in PBS with matrigel into NOD scid gamma mice (n = 8 per group). Tumor size was measured 2-3 times a week during the duration of the experiment.

LightSwitch luciferase reporter assay

HEK293FT cells were seeded in a 96-well white plate (Corning Costar) in triplicate. After 24 hours, cells were transfected with 100 ng of LightSwitch luciferase reporter construct with PTEN 3′-UTR (SWTICHGEAR genomics) or 100 ng Renilla luciferase reporter using Lipofectamine 2000 according to the manufacturer's instructions and incubated for 24 hours. The luciferase assays were performed according to the manufacturer's protocol (SWTICHGEAR genomics).

RNA-seq

Total RNA was extracted from cultured cells or xenograft tumors using RNeasy kit (QIAGEN) according to manufacturer's instructions. RNA quality was assessed by 2100 Bioanalyzer RNA 6000 Nano assay (Agilent). Libraries were prepared using TruSeq Stranded mRNA kits (Illumina) and subjected to 100 cycle paired-end sequencing on the Illumina HiSeq platform.

CRISPR/Cas9 genome editing

Genetically modified cell lines were generated using CRISPR-Cas9 technology. Briefly, MAGEA11 KO DAOY cells were created by transiently co-transfecting 400,000 cells with 500 ng of each gRNA expression plasmid (cloned into Addgene plasmid #43860), 1 µg Cas9 expression plasmid (Addgene plasmid #43945), and 200 ng of pMaxGFP via nucleofection (Lonza, 4D-Nucleofector X-unit) using solution P3 and program EN-158 in small (20 µl) cuvettes according to the manufacturer's recommended protocol. MAGEA11 KO H520 cells were created by transiently co-transfecting 400,000 cells with 500 ng of each gRNA expression plasmid (cloned into Addgene plasmid #43860), 1 µg Cas9 expression plasmid (Addgene plasmid #43945), and 200ng of pMaxGFP via nucleofection (Lonza, 4D-Nucleofector X-unit) using solution P3 and program EH-100 in small (20 µl) cuvettes according to the manufacturer's recommended protocol. PCF11 I689A cells were created by transiently co-transfecting 400,000 cells with 500 ng of gRNA expression plasmid (cloned into Addgene plasmid #43860), 1 µg Cas9 expression plasmid (Addgene plasmid #43945), 2.1 µg of ssODN, and 200 ng of pMaxGFP via nucleofection (Lonza, 4D-Nucleofector X-unit) using solution P3 and program CA-137 in small (20 µl) cuvettes according to the manufacturer's recommended protocol. Five days post-nucleofection, cells were single-cell sorted by FACS to enrich for GFP+ (transfected) cells, clonally expanded and verified for the desired targeted modification via targeted deep sequencing followed by analysis using CRIS.py (Connelly and Pruett-Miller, 2019). Clones were identified for each modification and assessed in relevant assays. The sequences for genome editing reagents and applicable primers are listed below.

Name	Sequence (5′ to 3′)
hMAGEA11 KO reagents	
hMAGEA11.sgRNA.g1	GACGGCGGGACUAUGGGGGG
hMAGEA11.sgRNA.g2	UGUGGCCCUGAAGCAUGCAU
hMAGEA11.NGS.F partial Illumina adaptors (upper case)	CACTCTTTCCCTACACGACGCTCTTCCGATCTagcaaggctccctctctgctgtcag
hMAGEA11.NGS.R partial Illumina adaptors (upper case)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTtctccaagtcacccgatggaagaga
hPCF11 I689A reagents	
hPCF11 sgrna	CAGCUAUUUCAGUAUCAAGA
hPCF11 I689A ssODN I689A and blocking modifications (upper case)	agtgaacgtttagcatctggtgaaattacacaggatgacttccttgttgttgtgcatcaaGCtcgacagctatttcaAtaCcaGgaaggtaaacatagatgcaatgtacgggatagtcctacagaagaaaataaaggtggatta
hPCF11.NGS.F partial Illumina adaptors (upper case)	CACTCTTTCCCTACACGACGCTCTTCCGATCTcccctatacagacgagtgaacg
hPCF11.NGS.R partial Illumina adaptors (upper case)	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTtgaatgctgaacctgtgtcct

TCGA 3′-UTR analysis

The original TCGA RNA-seq gene expression data were obtained from the UCSC Cancer Genomics Hub (CGHub). All of the patients in a tumor type were ranked based on the CPM (count per million) values of *MAGE-A11* gene. The top 10 most highly MAGE-A11-expressed patients and bottom 10 least *MAGE-A11* expressed patients were chosen as two groups. The significant dynamics 3′-UTR usage genes between these two groups will be identified if the mean percentage of distal polyA usage (PDUI) change between these two groups is larger than 0.2 and mean fold change is larger than 1.5, also the p value calculated from Student's t test is less than 0.05. Finally, we observed *MAGE-A11* promotes strong 3′-UTR shortening in two tumor types including ovarian cancer (OV) and lung squamous cell carcinoma (LUSC).

RNA-seq data analysis

RNA from cells with MAGE-A11 knocked out or overexpressed, and PCF mutant-expressing cells and controls were sequenced by HiSeq. The raw paired-end RNA-seq reads were filtering out low-quality reads using Trim Galore, and then aligned to the human genome (hg19/GRCh37) using STAR version 2.5.2b (Dobin et al., 2013) using the following alignment parameters: –outSAMtype BAM SortedByCoordinate –outSAMstrandField intronMotif –outFilterMultimapNmax 10 –outFilterMultimapScoreRange 1 –alignSJDBoverhangMin 1 –sjdbScore 2 –alignIntronMin 20 –alignSJoverhangMin 8. The resulted BAM files were converted into bedgraph format using bedtools version 2.17.0 (Quinlan and Hall, 2010). For each gene, the read count were calculated by HTSeq (Anders et al., 2015), and then CPM values based on read count were used. The read coverage was visualized at UCSC Genome Browsers (Goldman et al., 2015). Differential gene analysis was performed using DESeq2 (Anders and Huber, 2010).

DaPars analysis

DaPars (Feng et al., 2018; Xia et al., 2014) was used to identify the most significant APA events between two conditions. We require that significant APA events should meet three criteria. First, the adjusted p value of PDUI differences was controlled at 5%. Second, the absolute mean difference of PDUI must be no less than 0.2. Third, the mean PDUI fold change must be no less than 1.5.

CFIm25 motif analysis

MAGE-A11 sensitive transcripts were defined as those transcripts with significant 3′-US upon MAGE-A11 overexpression, while MAGE-A11 insensitive transcripts were an equal number of randomly selected unaffected transcripts (PDUI differences less than 0.05; p value larger than 0.5). For each DaPars predicted PAS, the nearest annotated PAS was defined as the true PAS. The annotated PASs were compiled from multiple domains including Refseq, ENSEMBL, UCSC gene models and PolyA_DB version 3 (Wang et al., 2018) databases. The sequences of 200 nucleotides upstream and downstream of the PASs were used for motif analysis. The CFIm25 motif density was calculated by counting the number of UGUA motif (smoothed over 7 nucleotide) along these specified annotation features, which included proximal and distal PAS.

CLIP-qPCR

Cross-linking immunoprecipitation and QPCR (CLIP-QPCR) was carried out as previously described (Yoon and Gorospe, 2016). Briefly, HEK293FT/HA-FLAG-Vector or HEK293FT/HA-FLAG-MAGE-A11 cells (10 15 cm^2 dishes) were washed in ice-cold, magnesium-free PBS and irradiated on ice with 150 mJ/cm^2 of UVC (254 nm) in a Stratalinker 2400 (Agilent). Cells were collected in ice-cold PBS, pelleted, lysed in NP-40 lysis buffer (50 mM Tris-HCl pH 7.5, 150 mM KCl, 0.5% (v/v) NP-40), and centrifuged for 15 min at 10,000 xg at 4°C. Supernatants were collected and subjected to immunoprecipitation. Cell lysates were incubated with 20 μL pre-coupled antibody-protein A/G PLUS-agarose beads (Santa Cruz Biotechnology) for 3 hr at 4°C rotating. Antibodies (10 μg) used were as follows: normal mouse IgG control (Santa Cruz, sc-2025) and anti-CFIm25 (Proteintech, 66335-1-Ig). Beads were then washed three times in NP-40 lysis buffer, treated with 20 units of RNase-free DNase I for 15 min at 37°C, and proteins degraded by treatment with 0.5 mg/mL proteinase K (Invitrogen) in 0.5% SDS at 55°C for 15 min. RNA was then separated by phenol: chloroform extraction, followed by ethanol precipitation. RNA was then converted to cDNA using the High Capacity cDNA reverse transcriptase kit (Invitrogen) according to manufacturer's instructions. qPCR analysis was performed on cDNA using PowerUp SYBR Green master mix (Applied Biosystems) according to manufacturer's instructions using the following primers: *CCND2* forward, 5′-TTCCCTCTGGCCATGAATTAC; reverse, 5′-GGGCTGGTCTCTTTGAGTTT and *RPLP0* forward 5′-TCTACAACCCTGAAGTGCTT GAT-3′, *RPLP0* reverse 5′-CAATCTGCAGACAGACACTGG-3′. Data were analyzed by ΔΔCt method normalizing to *RPLP0* and control normal IgG pulldowns.

Trans-effect analysis of 3′-US

We used MAT3UTR (Park et al., 2018) for the detection of trans-effect of MAGEA11-induced 3′-UTR shortening in ceRNA in two tumor types OV and LUSC. The Briefly, MAT3UTR can predict ceRNA partner expression changes by using its 3′-UTR shortening gene expression, 3′-UTR shortening gene level, microRNA binding sites and miRNA expression. The miRNA binding sites were compiled from a collection of TarBase, miRecords, miRTarBase and predicted miRNA-binding sites from TargetScanHuman version 6.2. Exon and CDS annotation for TCGA and miRNA expression were downloaded from Xena UCSC Genome browsers. The enrichment of ceRNA partner genes with tumor suppressor gene (TSG) and oncogene (OG) was calculated by fisher exact test. The annotation of TSG and OG were from TUSON prediction (Davoli et al., 2013) with top 500 genes ($p < 0.01$) selected.

Mass spectrometry analysis

Protein samples were digested and the resulting peptides were analyzed by an optimized LC-MS/MS platform (Pagala et al., 2015). For quantitative TMT analysis, the digested peptides were labeled with individual TMT reagents, equally pooled, and fractioned by basic pH reversed phase LC chromatography. Each fraction was then analyzed using acidic pH reverse phase nanoscale LC-MS/MS

(Bai et al., 2017). The collected MS data were processed for protein identification and quantification by database search using the JUMP software suite (Wang et al., 2014).

DATA AND CODE AVAILABILITY

Proteomics data are available at MassIVE MSV000084123. RNA-seq data are available at NCBI GEO: GSE134898.

nature genetics

LETTERS
https://doi.org/10.1038/s41588-018-0118-8

3′ UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk

Hyun Jung Park[1,2,3,9], Ping Ji[4,9], Soyeon Kim[5], Zheng Xia[1,2], Benjamin Rodriguez[1,2], Lei Li[1,2], Jianzhong Su[1,2], Kaifu Chen[1,2], Chioniso P. Masamha[6], David Baillat[4], Camila R. Fontes-Garfias[4], Ann-Bin Shyu[7], Joel R. Neilson[8], Eric J. Wagner[4,10]* and Wei Li[1,2,10]*

Widespread mRNA 3′ UTR shortening through alternative polyadenylation[1] promotes tumor growth in vivo[2]. A prevailing hypothesis is that it induces proto-oncogene expression in cis through escaping microRNA-mediated repression. Here we report a surprising enrichment of 3′ UTR shortening among transcripts that are predicted to act as competing-endogenous RNAs (ceRNAs) for tumor-suppressor genes. Our model-based analysis of the trans effect of 3′ UTR shortening (MAT3UTR) reveals a significant role in altering ceRNA expression. MAT3UTR predicts many trans-targets of 3′ UTR shortening, including *PTEN*, a crucial tumor-suppressor gene[3] involved in ceRNA crosstalk[4] with nine 3′ UTR-shortening genes, including *EPS15* and *NFIA*. Knockdown of *NUDT21*, a master 3′ UTR-shortening regulator[2], represses tumor-suppressor genes such as *PHF6* and *LARP1* in trans in a miRNA-dependent manner. Together, the results of our analysis suggest a major role of 3′ UTR shortening in repressing tumor-suppressor genes in trans by disrupting ceRNA crosstalk, rather than inducing proto-oncogenes in cis.

Widespread 3′ UTR shortening (3′US) through alternative polyadenylation (APA) occurs during enhanced cellular proliferation and transformation[1,5–8]. Recently, we reported that *NUDT21*-mediated 3′US promotes glioblastoma growth, further underscoring its significance to tumorigenesis[2]. A prevailing hypothesis is that a shortened 3′ UTR results in activation of proto-oncogenes in cis through escaping microRNA (miRNA)-mediated repression. Indeed, several well-characterized oncogenes, such as *CCND1*, have been shown to use 3′US to increase their protein levels, but mostly in cell lines[5]. However, in recent PolyA sequencing[7] and our TCGA RNA sequencing (RNA-Seq) APA analysis[1] (5 and 358 tumor/normal pairs, respectively), most oncogenes with 3′US previously identified in vitro[5] displayed almost no changes in their 3′ UTR lengths in tumors (Fig. 1a). For example, we identified 1,346 recurrent (occurrence rate >20%) 3′US genes in 358 tumor/normal pairs[1]. However, *CCND1* is not on that list as its 3′US occurred in only a very small portion (8 out of 358; 2.2%) of tumors (Fig. 1b). Furthermore, similar to random genes, 3′US genes from all 5 previous APA studies have little overlap with the top 500 ($P<0.01$) high-confidence oncogenes as defined on the basis of distinct somatic mutational patterns of >8,200 tumor/normal pairs[9] (Fig. 1c). These results challenge the previous hypothesis and suggest a different function of 3′US for tumorigenesis.

Aside from regulating its cognate transcript in cis, the 3′ UTR has also been implicated in competing-endogenous RNA (ceRNA) regulation in trans[10]. Although the scope is not fully understood, ceRNA is generally thought to form global regulatory networks (ceRNETs) controlling important biological processes[11]. For example, the tumor suppressor *PTEN*'s ceRNAs, *CNOT6L* and *VAPA*, have been shown to regulate *PTEN* and phenocopy its tumor-suppressive properties[12]. As the ceRNA's regulatory axis is mostly based on miRNA-binding sites on 3′ UTRs, we hypothesize that when genes with shortened 3′ UTRs no longer sequester miRNAs, the released miRNAs would then be directed to repress their ceRNA partners, such as tumor-suppressor genes, in trans, thereby contributing to tumorigenesis.

To test this hypothesis, we first used well-established strategies to reconstruct two ceRNETs from 97 TCGA breast tumors and their matched normal tissues, respectively, based on miRNA-binding-site overlap and co-expression[13,14] between genes of active ceRNA regulation (Methods). In general, transcripts are less correlated between each other in tumors than in normal tissues, partially due to tumor heterogeneity[15] and global reduction of miRNA expression in tumors[16] (Fig. 2a). As expected, the loss of co-expression results in a much smaller (tenfold reduced) ceRNET for tumors than for normal tissues (Fig. 2b).

To investigate the role of 3′US in ceRNET disruption, we focused on estrogen-receptor-positive (ER$^+$) breast tumors, which comprise the majority (68/97) of TCGA breast tumor samples. We built normal and tumor ceRNETs using the same procedure as above. Using the DaPars algorithm[1], we identified 427 3′US genes recurring in >20% of tumors. Close inspection indicates that 3′US is associated with ceRNET disruption. For example, we identified *PTEN* and *EPS15* as a ceRNA pair in normal ceRNET (4 miRNA-binding-site overlap and $\rho=0.63$ co-expression). However, since *EPS15* underwent 3′US in 23 (33.8%) out of 68 tumors, thereby losing its capability to compete with *PTEN* for miRNAs, it lost ($\rho=0.32$) the co-expression (and ceRNA) relationship with *PTEN* in tumors (Fig. 2c). Globally, the top 100 ceRNAs with the most significant 3′US genes all lost their interactions in tumors, while 12 out of 100 ceRNAs lacking 3′US retained ($P=0.0002$) their interactions.

[1]Division of Biostatistics, Dan L Duncan Cancer Center, Baylor College of Medicine, Houston, TX, USA. [2]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX, USA. [3]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA. [4]Department of Biochemistry & Molecular Biology, University of Texas Medical Branch, Galveston, TX, USA. [5]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA. [6]Department of Pharmaceutical Sciences, College of Pharmacy and Health Sciences, Butler University, Indianapolis, IN, USA. [7]Department of Biochemistry and Molecular Biology, University of Texas, McGovern Medical School, Houston, TX, USA. [8]Department of Molecular Physiology and Biophysics, Baylor College of Medicine, Houston, TX, USA. [9]These authors contributed equally: Hyun Jung Park, Ping Ji. [10]These authors jointly supervised this work: Eric J. Wagner, Wei Li. *e-mail: ejwagner@utmb.edu; WL1@bcm.edu

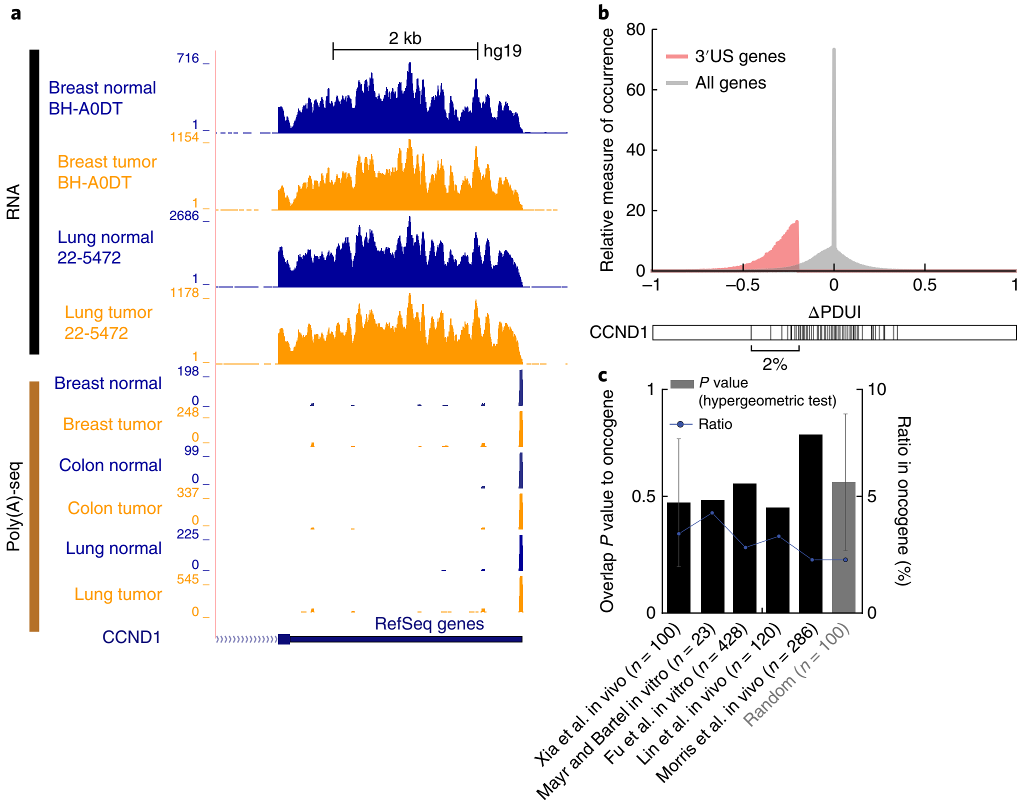


Fig. 1 | 3′US genes are not strongly associated with oncogenes. a, TCGA RNA-Seq data for *CCND1* demonstrates no change in 3′UTR usage between tumors (yellow) and matched normal samples (blue). A similar pattern was also observed in PolyA-seq[7] of *CCND1*. **b**, ΔPDUI values for 3′US genes (red) and all genes (gray) in 358 TCGA tumor/normal pairs[1] (upper panel). A negative ΔPDUI represents 3′UTR shortening. The lower panel shows ΔPDUI values for CCND1 across 358 tumor/normal pairs[1]. Significant *CCND1* 3′ UTR shortening occurred only in a very small portion (8 out of 358; 2.2%) of tumors. **c**, Overlap *P* values and the ratios between previously identified 3′US genes and oncogenes. 'Random ($n = 100$)' represents the averaged *P* value from 100 random sampling of 100 RefSeq genes. The error bar represents standard variation values of *P* values from 100 random trials.

Furthermore, in separate ceRNETs from 30 tumor/normal pairs with the least and most amount of 3′US (upper panel in Fig. 2d), more 3′US is clearly associated with more ceRNET loss (38.6 versus 16.4 in fold decrease, $P < 1 \times 10^{-16}$, lower panel in Fig. 2d). From these findings, we conclude that 3′US is strongly associated with ceRNA network disruption in tumors.

To understand the function of 3′US-mediated ceRNET disruption, we selected 381 3′US genes and 2,131 of their ceRNA partner genes (3′US ceRNAs), including 591 3′US ceRNA hub and 1,540 3′US ceRNA non-hub genes, in the normal ceRNET (Supplementary Table 1, Methods). We hypothesized that 3′US genes released their miRNAs to repress their ceRNA partners in trans. Consistent with our hypothesis, expression changes of 2,131 3′US ceRNA genes in tumors are anti-correlated ($r = -0.21$; $P = 5 \times 10^{-24}$) with the degree of 3′US of the associated 3′US genes (Supplementary Fig. 1a). As a result, among 976 genes in normal ceRNET downregulated in tumors, 816 (83.6%) are ceRNAs of 3′US genes. Surprisingly, 3′US ceRNA hub genes are enriched in tumor-suppressor genes ($P \sim 1 \times 10^{-20}$) but not in oncogenes (Fig. 3a), suggesting that the 3′US represses tumor suppressors in trans. For example, 3′US of *EPS15* would contribute to downregulating its ceRNA partner *PTEN* in tumors (Fig. 2c). Globally, 160 expressed tumor-suppressor genes from 3′US ceRNAs are more likely downregulated than 226 control tumor-suppressor genes not in ceRNET ($P = 8 \times 10^{-3}$, Fig. 3b), indicating a significant association between 3′US and tumor-suppressor gene repression.

Additional analyses on sequence features partially explain why 3′US genes, but not tumor suppressors in their ceRNA partners, are likely to have alternative proximal polyadenylation sites, leading to 3′US (Supplementary Note). We have also analyzed TCGA 450K methylation array data and found that the 3′US-mediated ceRNA repression is independent of promoter hypermethylation (Supplementary Note).

To better quantify the trans effects of 3′US, we developed a mathematical model (MAT3UTR) based on its 3′US gene(s) expression, 3′US level, miRNA-binding site(s) and miRNA expression(s) (Methods). In 1,548 differentially expressed 3′US ceRNAs, MAT3UTR can explain 47.6% of variation in gene expression (Supplementary Fig. 3c). In contrast, the MAT3UTR-control model, which considers miRNA expression but not 3′US, explains only 27.2% of variation (Supplementary Fig. 3d), consistent with previous reports[17] that miRNA alone has a weak role in regulating gene expression. The results suggest that the trans effects of 3′US plays a major role in regulating ceRNA gene expression.

MAT3UTR predicts many trans-target genes of 3′US, including *PTEN*, in ceRNA crosstalk[11–13] (top 1% MAT3UTR score, Supplementary Table 2). In normal ceRNET, *PTEN* is predicted to be a ceRNA of nine 3′US genes (Fig. 3c). When we ranked 97 breast tumor/normal pairs by the amount of 3′US across these nine genes (upper panel in Fig. 3d), tumors with more 3′US showed more downregulation of *PTEN* ($P = 0.03$, lower panel in Fig. 3d). Furthermore, MAT3UTR can explain 86.9% of the variation in *PTEN*'s expression across tumors (Supplementary Fig. 3g), suggesting that the trans effects of 3′US play a major role in downregulating *PTEN*.

To empirically test the hypothesis that 3′US can downregulate *PTEN* in trans, we focused on *EPS15* among the nine 3′US genes

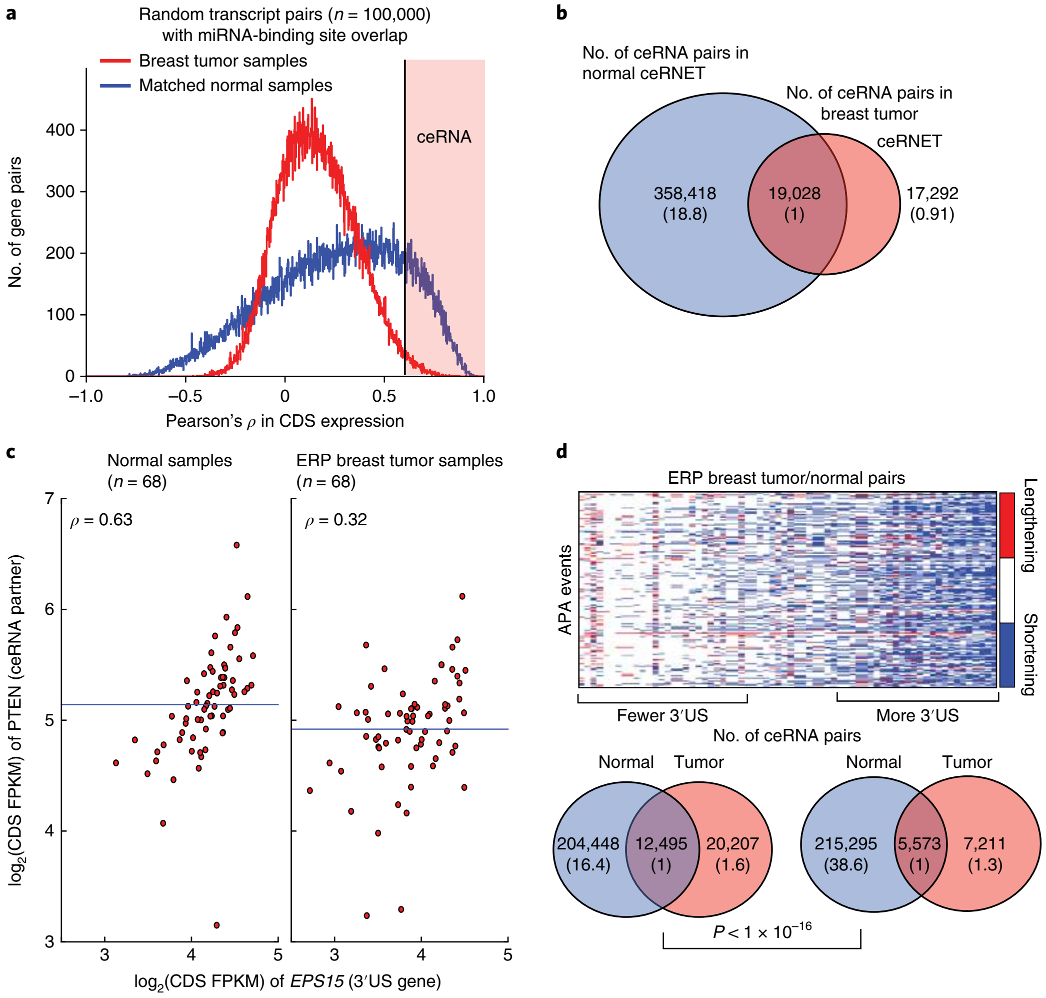


Fig. 2 | 3′UTR shortening contributes to ceRNET disruption. a, Pearson's correlation coefficients of 100,000 randomly selected transcript pairs with significant miRNA-binding-site overlap in breast tumors and matched normal tissues. **b**, The number of ceRNA pairs in breast tumor and the matched normal ceRNETs. The numbers in parentheses are normalized to the number of edges shared between tumor and normal. **c**, Gene expression of *EPS15* (3′US gene) and *PTEN* (ceRNA partner) on 68 estrogen-receptor-positive (ERP) breast tumors and matched normal samples. The horizontal lines represent the mean expression values of *PTEN*, which is decreased in tumors (FDR = 2.1×10^{-10}). **d**, The upper heatmap exhibits significant APA events (rows) across 68 ERP tumor/normal pairs (columns), ranked by the number of 3′US genes. The Venn diagrams show the number of ceRNA pairs in the normal and tumor ceRNET. The numbers in parentheses are normalized to the number of edges shared between tumor and normal tissues. The *P* value was calculated from a one-tailed Pearson's chi-squared test.

(Methods). We observed that depletion of *EPS15* by siRNA in MCF7 cells reduces *PTEN* expression (Fig. 3e). To ascertain whether this effect depends on miRNA-based targeting of the *PTEN* 3′ UTR, we used a luciferase reporter vector with the *PTEN* 3′ UTR (pLightSwitch-PTEN 3′ UTR) to test the effect of *EPS15* knockdown on its expression. We observed that reduction of *EPS15* reduces *PTEN* 3′ UTR luciferase activity (Fig. 3f). To further understand whether the crosstalk is miRNA-dependent, we depleted *DICER1* to abolish miRNA biogenesis and found that loss of *DICER1* can relieve the influence of *EPS15* knockdown on *PTEN* 3′ UTR expression (Fig. 3g). Finally, overexpression of the *EPS15* 3′ UTR increased the number of PTEN-positive cells (Fig. 3h,i). Thus, *EPS15* 3′US may impact *PTEN* expression.

To gain insights into the global cause-and-effect relationship between 3′US and the repression of tumor-suppressor genes, we revisited our previous data from *NUDT21*-knockdown HeLa cells, since NUDT21 is one of the master regulators of 3′US [2]. We identified 1,168 3′US ceRNAs in *NUDT21*-knockdown cells solely on the basis of significant miRNA-binding-site overlap with 1,450 3′US genes, since co-expression cannot be effectively estimated from two replicates of our experiments. With 9,914 expressed RefSeq genes with no significant miRNA-binding-site overlap with 3′US genes as random controls, the tumor-suppressor genes remain strongly enriched in 3′US ceRNAs ($P \sim 1 \times 10^{-38}$, Fig. 4a). Among 57 tumor-suppressor genes in 3′US ceRNAs, 33 (57.9%) showed repression in *NUDT21*-knockdown samples; whereas a smaller portion (44.5%) of 339 control tumor-suppressor genes showed repression ($P \sim 0.03$, Fig. 4b), suggesting that *NUDT21*-mediated 3′US represses tumor-suppressor genes in trans. In spite of potentially higher false positives due to lack of co-expression in ceRNA identification, these results are highly consistent with our observations in TCGA breast cancer. On the basis of these results, we posit that repression of tumor-suppressor ceRNAs would correlate with increased occupancy of AGO2 in the RISC complex. To formally test this hypothesis, we isolated cytoplasmic fractions from control or *NUDT21*-knockdown cells and conducted RNA immunoprecipitation (RIP) using anti-AGO2 antibodies. On average, we observed ~200-fold enrichment of ceRNAs in Ago2

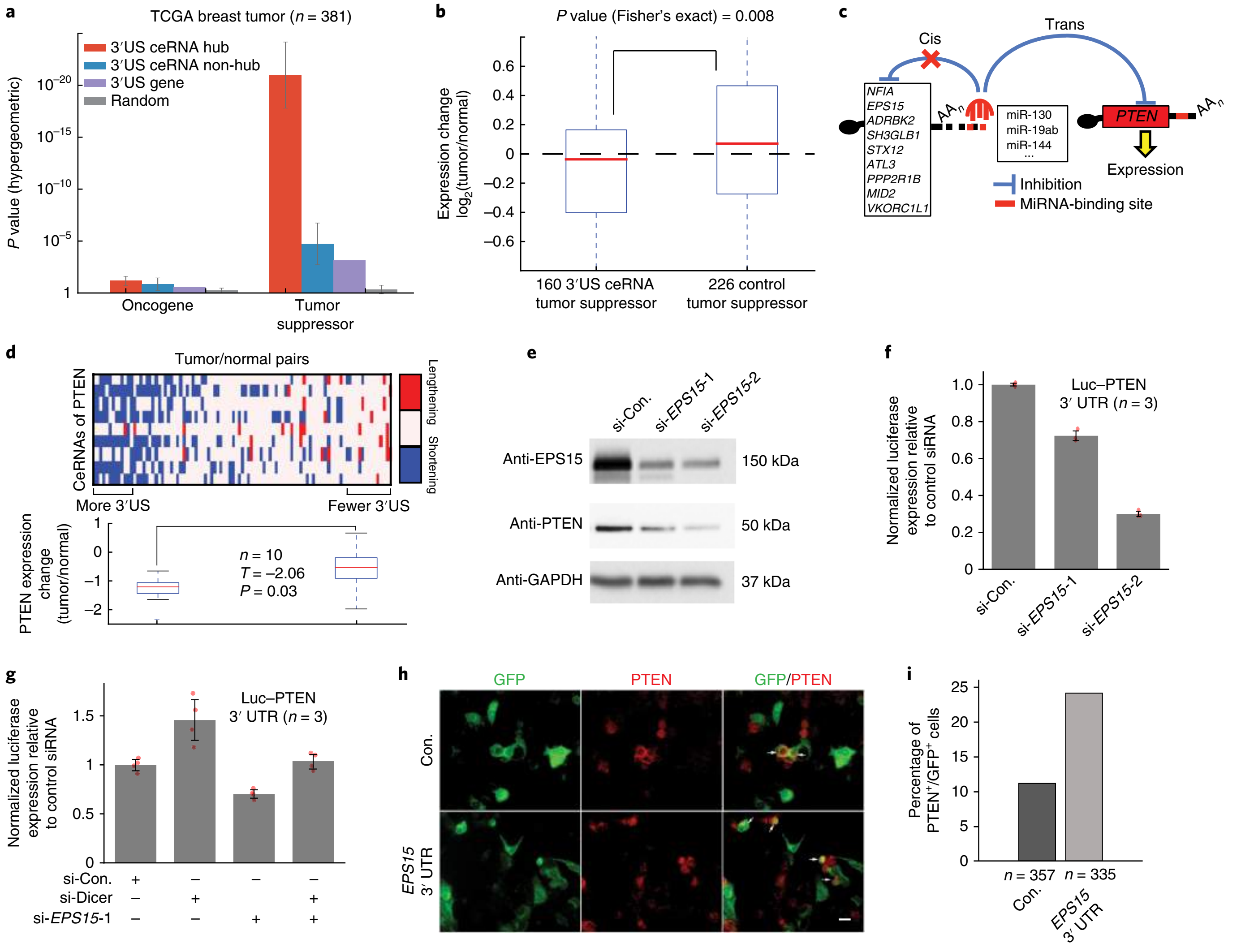


Fig. 3 | 3′UTR shortening represses tumor-suppressor genes in TCGA breast cancer. a, Functional enrichment of 3′US ceRNA hub genes (red), 3′US ceRNA non-hub genes (blue), 3′US genes (purple) and random RefSeq genes (gray). We randomly sampled each gene category to the same number (381) 100 times; averaged *P* values with standard deviation are plotted. **b**, Relative expression (tumor/normal) of tumor-suppressor genes that are 3′US ceRNAs ($n = 160$, left box) is lower than for those that are not in ceRNET ($n = 226$, right box) ($P = 8\times10^{-3}$). **c**, 3′US genes (left) might repress their ceRNA partner *PTEN* in trans through miRNAs (middle) commonly released through 3′ UTR shortening. **d**, A heatmap in the top panel showing APA events for the nine 3′US genes (rows). The boxplots in the bottom panel show *PTEN* expression levels in 10 tumors with the most (left) or least (right) 3′ UTR shortening. **e**, Western blot analysis of lysates from MCF7 cells treated with control (Con.) or *EPS15*-targeting siRNAs. The image is representative of three independent experiments. **f**, Quantification of luciferase activity from cell lysates derived from MCF7 cells transfected with a luciferase reporter containing the *PTEN* 3′ UTR and *EPS15*-targeting siRNAS. Data are the average luciferase activity ± standard deviation from three independent experiments ($P = 0.011$ and $P < 0.001$, two-sided *t*-test). **g**, *PTEN* 3′ UTR luciferase reporter activity in MCF7 cells transfected with *EPS15*- and *DICER*-targeting siRNAS. Data are the average luciferase activity ± standard deviation from three independent experiments ($P = 0.045$, $P = 0.003$ and $P = 0.645$, two-sided *t*-test). **h**, Indirect immunofluorescence of MCF7 cells transfected with either a heterologous reporter containing a vector-derived 3′ UTR (Con.) or the *EPS15* 3′ UTR together with a *GFP* construct. PTEN was detected by anti-PTEN antibody conjugated with Alexa Fluor-594. The arrows highlight $PTEN^+$ transfected cells. A representative image is shown from three independent experiments. Scale bar, 20 μM. **i**, The number of PTEN-positive cells in the transfected cells with either the *EPS15* 3′ UTR ($n = 335$) or the control 3′ UTR ($n = 357$) from three images.

RIP complexes relative to control IgG (Supplementary Fig. 4b). Reduced expression of *NUDT21* does not impact AGO2/DICER1 expression and *GAPDH* messenger RNA binding to AGO2 (Fig. 4c,d and Supplementary Fig. 4b). Furthermore, we sequenced miRNAs from control and *NUDT21*-knockdown cells, and found that miRNAs are equally likely to be upregulated or downregulated (Supplementary Fig. 4d), ruling out a general effect on miRNA biogenesis. Importantly, we could detect increased association of multiple tumor-suppressor ceRNAs with AGO2 following *NUDT21* depletion that ranged from 1.5-fold to nearly 7-fold (Fig. 4d). These results demonstrate that 3′US can lead to reduction of tumor-suppressor genes through their increased association with repressive AGO2 complexes.

To further validate the miRNA-dependent, repressive trans effects of 3′US, we monitored expression of the tumor-suppressor genes *PHF6* and *LARP1* and their ceRNA partners, *YOD1* and *LAMC1* (Supplementary Table 3). We consistently observed that *PHF6* and *LARP1* expression levels were decreased in

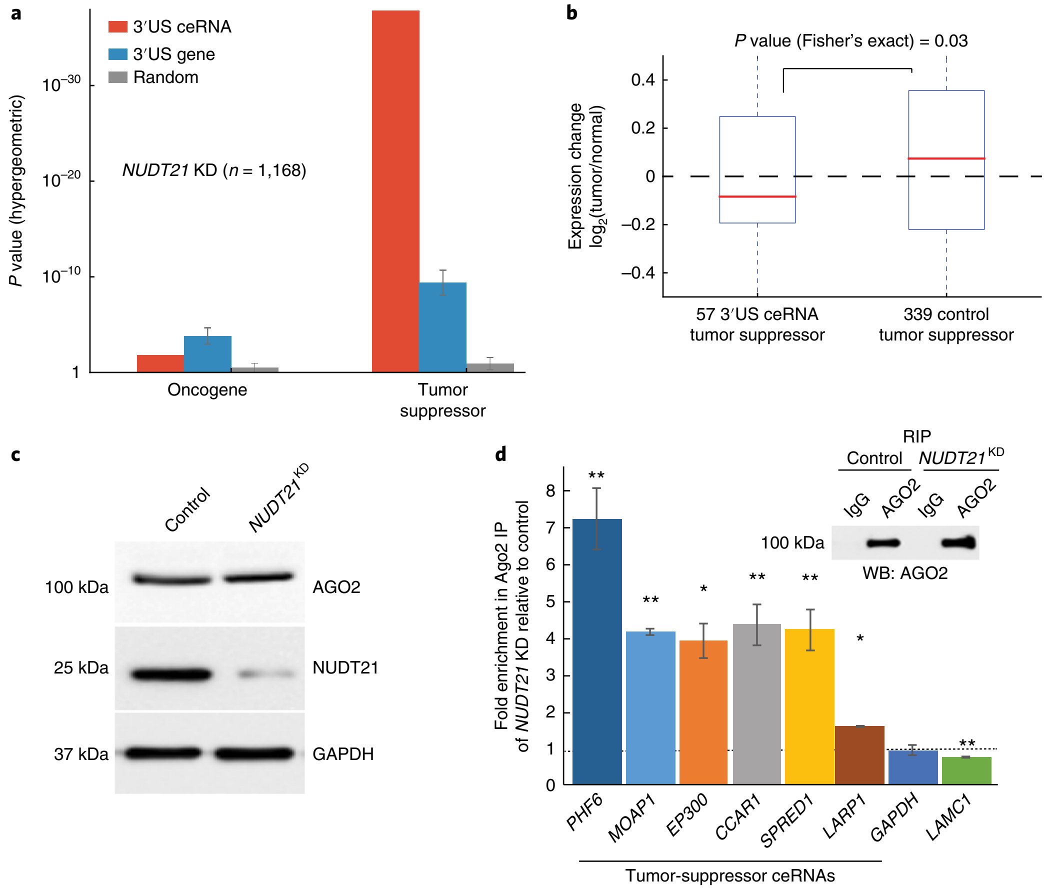


Fig. 4 | *NUDT21*-mediated 3′ UTR shortening causes tumor-suppressor repression in trans. a, Oncogene or tumor-suppressor gene enrichment of 3′US ceRNAs (red), 3′US genes (blue) and RefSeq genes (gray), in the *NUDT21*-knockdown (KD) experiment. We randomly sampled each gene category to the same number ($n = 1{,}168$) 100 times, and reported the averaged P values with standard deviation (error bar). **b**, Expression change of tumor-suppressor genes that are 3′US ceRNAs ($n = 57$, left box) or that are not connected to 3′US genes ($n = 339$, right box). 3′US ceRNA tumor-suppressor genes showed lower expression in *NUDT21*-knockdown samples ($P = 0.03$). **c**, Knockdown of *NUDT21* in HeLa cells using CRISPR/Cas9 and reduced NUDT21 was detected by western blot analysis in three independent experiments. **d**, RIP was performed with anti-AGO2 antibody; normal mouse IgG served as a control. The RIP complexes were detected by western blot with a distinct AGO2 antibody from rat (inset). The indicated ceRNAs associated with AGO2 enrichment in *NUDT21*-knockdown cytoplasmic lysates versus the control are shown with average fold change ± standard deviation from three independent assays ($P = 0.0002$, $P = 5.2 \times 10^{-6}$, $P = 0.0004$, $P = 0.0005$, $P = 0.0006$, $P = 0.01$ and $P = 5.47 \times 10^{-7}$, two-sided t-test, $^{**}P < 0.001$, $^{*}P < 0.01$).

NUDT21-knockdown cells while both *YOD1* and *LAMC1* expression levels were increased (Fig. 5a). To determine whether the 3′ UTR mediated this effect, we transfected luciferase reporters containing the 3′ UTR of either *PHF6* or *LARP1* into control or *NUDT21*-knockdown cells and measured luciferase activity. We found that both reporters were downregulated after *NUDT21* knockdown (Fig. 5b). Both *PHF6* and *LARP1* have been shown as tumor-suppressor genes[9,18,19] and downregulation of *PHF6* or *LARP1* in HeLa cells increases cell growth, confirming their tumor suppressive activity (Supplementary Fig. 5).

To further investigate the mechanism of tumor-suppressor ceRNA downregulation, we chose *PHF6* on the basis of MAT3UTR analysis and experimental results (Methods). We selected two miRNAs targeting *PHF6* (Fig. 5c), which were released by 3′US of *YOD1* (miR-3187-3p as the highest and miR-549 as the sixth highest in terms of β_{miR_j}in equation (3); Methods and Supplementary Table 4). Neither of these miRNAs was found to change its expression following *NUDT21* knockdown (Supplementary Fig. 4d). However, *PHF6* expression was partially rescued by an antagomir blocking the activity of miR-549 and completely rescued by an antagomir targeting miR-3187-3p (Fig. 5d). Moreover, *PHF6* 3′ UTR-mediated luciferase activity was partially rescued by the miR-3187-3p antagomir or *YOD1* siRNA (Fig. 5e). To understand whether reduced expression of *PHF6* depends on *YOD1* levels, we transfected *YOD1* cDNA into cells depleted of *YOD1* and found that re-expression of *YOD1* could not restore either the expression of endogenous *PHF6* (Fig. 5f) or the expression of the *PHF6* 3′ UTR-mediated luciferase (Fig. 5g), suggesting that the trans effect on *PHF6* is due to the 3′ UTR of *YOD1*. Finally, to determine whether the crosstalk between *PHF6* and *YOD1* is miRNA-dependent, we also showed that depletion of *DICER1* abolishes *PHF6* and *YOD1* crosstalk (Fig. 5h). Collectively, the data strongly suggest that *NUDT21*-mediated 3′US causes tumor-suppressor gene repression in trans in a miRNA-dependent manner.

Although analyzing ceRNA crosstalk in light of 3′US has been briefly suggested[20–22], our MAT3UTR analysis of 97 breast cancer RNA-Seq data followed by functional validation suggests a widespread causal role of 3′US in repressing tumor-suppressor genes in trans. While the trans effect further emphasizes the importance of APA in tumor progression, it also provides an additional layer of gene regulation and underscores the need for further investigation into other potential mechanisms[23,24] that could perturb ceRNA crosstalk, such as RNA editing and competition with RNA-binding proteins.

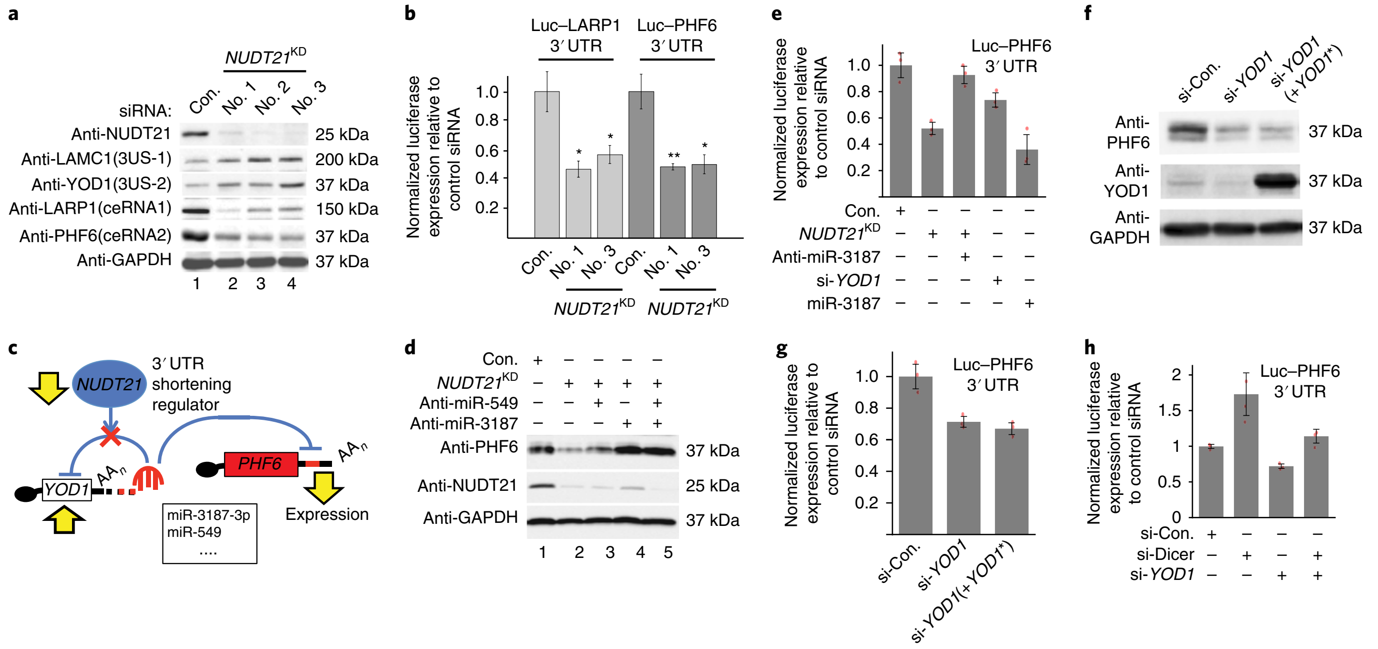


Fig. 5 | *NUDT21*-mediated 3′UTR shortening represses the tumor-suppressor genes *PHF6* and *LARP1*. **a**, Western blot of 3′US ceRNA tumor-suppressor genes (*PHF6*/*LARP1*) and 3′US genes (*LAMC1*/*YOD1*) in *NUDT21*-knockdown cells. A representative image is shown from three independent experiments. **b**, Activity of the *PHF6* 3′ UTR and *LARP1* 3′ UTR luciferase reporter constructs in *NUDT21*-knockdown cells relative to control siRNA-transfected cells. The data are the average of luciferase activity ± standard deviation from three independent experiments ($P=0.037$ and 0.05; $P=0.016$ and 0.025, two-sided *t*-test). **c**, *NUDT21* knockdown induces 3′ UTR shortening and upregulation of *YOD1*, allowing miR-3187-3p and miR-549 to repress *PHF6*. **d**, Western blot analysis using the indicated antibodies on lysates from HeLa cells transfected with siRNA for *NUDT21* (si-*NUDT21*-4) and two antagomirs that block miR-549 and miR-3187-3p. The image is representative of two independent experiments. **e**, Activity of the *PHF6* 3′ UTR luciferase reporter construct in HeLa cells with the indicated siRNAs, miRNAs or antagomirs. The data are the average of luciferase activity ± standard deviation from three independent experiments ($P=0.004$, $P=0.90$, $P=0.018$ and $P=0.015$, two-sided *t*-test). **f**, Western blot analysis of cell lysates from cells transfected with either control siRNA or *YOD1* siRNA. In the third lane, the cells were transfected with *YOD1* siRNA and then transfected with *YOD1* cDNA. The data are representative of three independent experiments. **g**, Activity of the *PHF6* 3′ UTR luciferase reporter in cells treated with the same experimental design as in **f**. The data are the average of luciferase activity ± standard deviation from three independent experiments ($P=0.016$ and $P=0.01$, two-sided *t*-test). **h**, Activity of the *PHF6* luciferase reporter in cells transfected with the indicated siRNAs. The data are the average of luciferase activity ± standard deviation from three independent experiments ($P=0.025$, $P=0.009$ and $P=0.99$, two-sided *t*-test).

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0118-8.

Received: 20 June 2016; Accepted: 22 March 2018;
Published online: 21 May 2018

References

1. Xia, Z. et al. Dynamic analyses of alternative polyadenylation from RNA-Seq reveal landscape of 3′ UTR usage across 7 tumor types. *Nat. Commun.* **5**, 5274, (2014).
2. Masamha, C. P. et al. CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* **510**, 412–416 (2014).
3. Zhang, S. & Yu, D. PI(3)king apart PTEN's role in cancer. *Clin. Cancer Res.* **16**, 4325–4330 (2010).
4. Poliseno, L. et al. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033–1038 (2010).
5. Mayr, C. & Bartel, D. P. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673–684 (2009).
6. Fu, Y. et al. Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.* **21**, 741–747 (2011).
7. Lin, Y. et al. An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res.* **40**, 8460–8471 (2012).
8. Morris, A. R. et al. Alternative cleavage and polyadenylation during colorectal cancer development. *Clin. Cancer* **18**, 5256–5266 (2012).
9. Davoli, T. et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
10. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146**, 353–358 (2011).
11. Thomson, D. W. & Dinger, M. E. Endogenous microRNA sponges: evidence and controversy. *Nat. Rev. Genet.* **17**, 272–283 (2016).
12. Tay, Y. et al. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell* **147**, 344–357 (2011).
13. Sumazin, P. et al. An extensive microRNA-mediated network of RNA–RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* **147**, 370–381 (2011).
14. Ala, U. et al. Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. *Proc. Natl Acad. Sci. USA* **110**, 7154–7159 (2013).
15. Swanton, C. Intratumor heterogeneity: evolution through space and time. *Cancer Res.* **72**, 4875–4882 (2012).
16. Lu, J. et al. MicroRNA expression profiles classify human cancers. *Nature* **435**, 834–838 (2005).
17. Hausser, J. & Zavolan, M. Identification and consequences of miRNA–target interactions — beyond repression of gene expression. *Nat. Rev. Genet.* **15**, 599–612 (2014).
18. Mets, E. et al. MicroRNA-128-3p is a novel oncomiR targeting PHF6 in T-cell acute lymphoblastic leukemia. *Haematologica* **99**, 1326–1333 (2014).
19. Selcuklu, S. D. et al. MicroRNA-9 inhibition of cell proliferation and identification of novel miR-9 targets by transcriptome profiling in breast cancer cells. *J. Biol. Chem.* **287**, 29516–29528 (2012).
20. Tian, B. & Manley, J. L. Alternative cleavage and polyadenylation: the long and short of it. *Trends Biochem. Sci.* **38**, 312–320 (2013).

21. Mueller, A. A., Cheung, T. H. & Rando, T. A. All's well that ends well: alternative polyadenylation and its implications for stem cell biology. *Curr. Opin. Cell Biol.* **25**, 222–232 (2013).
22. Li, L. et al. 3′UTR shortening identifies high-risk cancers with targeted dysregulation of the ceRNA network. *Sci. Rep.* **4**, 5406 (2014).
23. Tay, Y., Rinn, J. & Pandolfi, P. P. The multilayered complexity of ceRNA crosstalk and competition. *Nature* **505**, 344–352 (2014).
24. Wang, Y. et al. The emerging function and mechanism of ceRNAs in cancer. *Trends Genet.* **32**, 211–224 (2016).

Acknowledgements

This work was supported by US National Institutes of Health (NIH) grants R01HG007538, R01CA193466 and U54CA217297, Cancer Prevention Research Institute of Texas (CPRIT) grant RP150292 to W.L., CPRIT RP100107 to E.J.W. and A.-B.S., CPRIT RP140800 and Welch Foundation AU-1889 to E.J.W., and NIH RO1GM046454 and the Houston Endowment, Inc. to A.-B.S.

Author contributions

H.J.P. and W.L. conceived the project, designed the experiments and performed the data analysis. S.K. performed the regression analysis. Z.X. performed the APA analysis. L.L., J.S. and K.C. helped with data analysis. C.P.M., E.J.W. and A.-B.S. obtained the miRNA-Seq data. P.J., C.R.F.-G. and D.B. performed the *NUDT21*-knockdown experiments. H.J.P., P.J., E.J.W. and W.L. wrote the manuscript with input from B.R., A.-B.S., C.P.M. and J.R.N.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0118-8.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to E.J.W. or W.L.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Tumor-suppressor genes and oncogenes. The tumor-suppressor genes and oncogenes used in this study were defined by the TUSON algorithm from genome sequencing of >8,200 tumor/normal pairs[9], namely residue-specific activating mutations for oncogenes and discrete inactivating mutations for tumor-suppressor genes. TUSON is a computational method that analyzes patterns of mutation in tumors and predicts the likelihood that any individual gene functions as a tumor-suppressor gene or oncogene. We ranked genes by their TUSON prediction P values from the most to the least significant and used the top 500 genes ($P < 0.01$) as the reference tumor-suppressor genes or oncogenes. After removing 30 genes in common, 470 tumor-suppressor genes and oncogenes were used for the enrichment analysis. Note that there were very few breast tumor-specific tumor-suppressor genes and oncogenes (36 and 3 with breast q-value ≤ 0.5, respectively) and 90% of them were found in the top 500 pan-cancer predictions.

Previously identified 3′US genes in cancers. Xia et al. identified 1,187 3′US genes across 7 TCGA cancer types[1]. Mayr and Bartel selected 23 3′US genes from 27 cancer cell lines[5]. Fu et al. identified 428 3′US genes in human breast cancer cell lines[6]. Lin et al. reported 120 3′US genes in major cancers and tumor cell lines[7]. Morris et al. found 286 3′US genes in human colorectal tumor samples[8]. The 3′US genes of Xia et al. were randomly sampled to 100 genes for a fair comparison.

Selection of miRNA-binding sites. Predicted miRNA-binding sites were obtained from TargetScanHuman version 6.2[25]. Only those with a preferentially conserved targeting score (Pct) more than 0 were used[1]. Experimentally validated miRNA-binding sites were obtained from TarBase version 5.0[26], miRecords version 4[27] and miRTarBase version 4.5[28]. The binding sites found in indirect studies such as microarray experiments and high-throughput proteomics measurements were filtered out[29]. Another source is the microRNA target atlas composed of public AGO-CLIP data[30] with significant binding sites (q-value <0.05). The predicted and validated binding site information was then combined to use in this study.

TCGA breast tumor RNA-Seq and miRNA-Seq data. Quantified gene expression files (RNASeqV1) for primary breast tumors (TCGA sample code 01) and their matching solid normal samples (TCGA sample code 11) were downloaded from the TCGA Data Portal[31]. We used 97 breast tumor samples that have matched normal tissues. A total of 10,868 expressed RefSeq genes (fragments per kilobase of transcript per million mapped reads (FPKM) ≥ 1 in >80% of all samples) were selected for downstream analyses. To better quantify gene expression in the presence of 3′US, we used only coding regions (CDS) to quantify mRNA expression. Exon and CDS annotation for TCGA data and miRNA expressions (syn1445790) were downloaded from Sage Bionetworks' Synapse database.

CeRNA identification in TCGA breast tumors. CeRNAs were identified by miRNA-binding-site overlap and expression correlation[13,14]. Only microRNAs with intermediate expression (between 0.01 and 100 in averaged fragments per million mapped fragments (FPM)) were used to capture dynamic interactions[14]. After removing genes with fewer than six such miRNA-binding sites, gene pairs with significant miRNA-binding-site overlap (<0.05 in Benjamini–Hochberg-corrected P value) were selected. Among them, pairs correlated (>0.6 in Pearson's correlation coefficient) ($P < 1 \times 10^{-10}$) in gene expression were defined as ceRNAs. To account for mRNAs with variable 3′ UTRs, we used only CDS to quantify mRNA expression. Genes that are connected with >500 ceRNAs were defined as hub genes.

Model-based analysis of trans effect of 3′US (MAT3UTR). Suppose transcript x has a constitutive proximal 3′ UTR (pUTR) and a distal 3′ UTR that might be shortened in tumors (dUTR) (Supplementary Fig. 3a). We define $\mathrm{MiRs}(x,\ \mathrm{miR}_j)$ as the amount of binding sites for miRNA miR_j in x.

$$\begin{aligned}\mathrm{MiRs}(x, \mathrm{miR}_j) = \; & (\mathrm{pUTR}(x,\ \mathrm{miR}_j) + \mathrm{dUTR}(x,\ \mathrm{miR}_j) \\ & \times \mathrm{PDUI}(x)) \times \mathrm{FPKM}(x)\end{aligned} \tag{1}$$

where $\mathrm{pUTR}(x,\ \mathrm{miR}_j)$ and $\mathrm{dUTR}(x,\ \mathrm{miR}_j)$ are the numbers of miR_j binding sites in pUTR and dUTR of x, and $FPKM(x)$ is expression of x. PDUI indicates the percentage of dUTR usage index[1]. Note that equation (1) can also estimate for genes with no distal 3′ UTR by setting $\mathrm{PDUI} = 1$.

To estimate the trans effect of 3′US on gene y', we define X to be a set of 3′US genes that are ceRNA partners of y' (Supplementary Fig. 3b) and Y to be a set of ceRNA partners to $x \in X$, including y'. Only moderately expressed miRNAs are considered, since they are likely to bind all possible binding sites. Thus, we can roughly use the amount of miRNA-binding sites to represent the miRNA function. The miR_j-binding effect on each copy of y' can be defined as follows:

$$\mathrm{TransE}(y',\ \mathrm{miR}_j) = \frac{\mathrm{FPM}(\mathrm{miR}_j)}{\sum_{x \in X} \mathrm{MiRs}(x, \mathrm{miR}_j) + \sum_{y \in Y} \mathrm{MiRs}(y, \mathrm{miR}_j)} \tag{2}$$

where $\mathrm{FPM}(\mathrm{miR}_j)$ is the miR_j expression level. Since miRNA can bind to any binding sites in the genes connected by the ceRNA relationship ($X \cup Y$), both X and Y need to be considered.

The high-dimensional MAT3UTR input data are often highly correlated with each other (for example, 588 miRNAs in equation (2)). Therefore, MAT3UTR employs the ridge regression that is known to address the dimensionality and collinearity[32,33] in biological data. Indeed, the ridge regression yields a remarkably higher prediction power than classical linear regression. For example, MAT3UTR has a much smaller mean square error (0.38) than classical linear regression (mean square error = 10.84) (Supplementary Fig. 3f).

$$\mathrm{MAT3UTR}(y') = \sum_{\mathrm{miR}_j \in 3'\mathrm{UTR}(y')} \beta_{\mathrm{miR}_j} \times \log \frac{\mathrm{transE}(y', \mathrm{miR}_j)_{\mathrm{tumor}}}{\mathrm{transE}(y', \mathrm{miR}_j)_{\mathrm{normal}}} + \epsilon_{y'} \tag{3}$$

subject to $\sum_{\mathrm{miR}_j \in 3'\mathrm{UTR}(y')} \beta_{\mathrm{miR}_j} \leq t$, the ridge regression penalty. $MAT3UTR(y')$ is the trans effect of 3′US; β_{miR_j} is the regression coefficient of miR_j; $\epsilon_{y'}$ is the gene-specific error term. We use R^2 to show how much variation in gene expression can be explained by the $MAT3UTR$ model. We also used 10-fold cross-validation (CV) to choose the optimal regularization parameter t with 75% of data for training and the remaining 25% for testing. CV error is measured by mean-squared error. Then, to estimate β, we fit the ridge regression with the entire data set using the selected regularization parameter as chosen by CV.

As a result, y' would be more repressed following 3′US, if: y' contains more miRNA-binding sites in its 3′ UTR; X and Y contain fewer miRNA-binding sites; and more transcripts in X undergo 3′US. The MAT3UTR-control model, which considers miRNA expression but not 3′US, is defined as:

$$\mathrm{MAT3UTR\text{-}control}(y') = \sum_{\mathrm{miR}_j \in 3'\mathrm{UTR}(y')} \beta_{\mathrm{miR}_j} \times \log \frac{\mathrm{FPM}(\mathrm{miR}_j)_{\mathrm{tumor}}}{\mathrm{FPM}(\mathrm{miR}_j)_{\mathrm{normal}}} + \epsilon_{y'} \tag{4}$$

where $\mathrm{FPM}(\mathrm{miR}_j)$ is the miR_j expression level. For model comparison between MAT3UTR and MAT3UTR-control, we randomly selected 75% of data for training and the remaining 25% for testing. We perform random division 100 times to evaluate the performance of the MAT3UTR and MAT3UTR-control models, where 10-fold CV also confirms that MAT3UTR has a 2-fold higher prediction power on gene expression variation than the MAT3UTR-control model (Supplementary Fig. 3e).

Selecting genes for experimental validations. To test the trans repressive effect of 3′US on PTEN, we chose EPS15 on three grounds. First, its expression is easily detected in MCF-7 cells; second, analysis of RNA-Seq from MCF-7 cells[34] indicates distal polyA site usage of the *EPS15* transcript; third, the *EPS15* 3′ UTR contains four microRNA target sites that compete with the *PTEN* 3′ UTR.

To investigate the tumor-suppressor ceRNA downregulation mechanism, we chose *PHF6*, because among 57 tumor-suppressor genes in 3′US ceRNAs, *PHF6* was predicted as a strong (sixth highest in MAT3UTR score, Supplementary Table 3) trans-target of 3′US, was significantly downregulated (second highest in gene expression) and was the most enriched in AGO2 RIP complexes of the ceRNA tested (Fig. 4d).

Statistical analyses. Differential expression analyses were carried out by edgeR (version 3.8.6)[35] (tumor samples versus normal samples) with false discovery rate (FDR) control at 0.05. The significance of observed values for a particular class compared to its control is calculated from one-tailed Pearson's χ^2 test. Each variable follows either a binomial or multinomial distribution and each case consists of at least five counts, which meets the assumption of Pearson's χ^2 test. To test whether there is a significant enrichment of tumor-suppressor genes or oncogenes among a gene list of our interest, we conducted hypergeometric tests with normalized overlap counts, since assessing overlap between sets meets all criteria to use hypergeometric tests, including trials without replacement. To compare means of two groups that have different variances, we used Welch's t-test, which does not assume equal population variance. To check the normality assumption for the t-test, we conducted a Shapiro-Wilk normality test for small samples ($n < 50$). All statistical computations were performed in the Python scipy stats package (version 0.15.1) or R (version 3.1.1).

RNA-Seq for *NUDT21* depletion experiment. We previously sequenced two control and two *NUDT21* depletion samples of HeLa cells by HiSeq 2000 (LC Sciences)[2]. After trimming adaptors using Trim Galore (version 0.4.1), paired-end RNA-Seq reads of 101 base pairs in each end were used to reconstruct the transcriptome in the Tuxedo protocol[36] (TopHat 2.0.6 and Cufflinks 2.1.1). The resulting FPKM values were normalized for comparison using Cuffdiff 2.2.0. Further analyses are based on 10,681 expressed (FPKM ≥ 1 in >3 samples) RefSeq genes. We sequenced miRNAs from control and *NUDT21*-knockdown cells to utilize only miRNAs with intermediate expression in ceRNA identification.

CeRNA identification in the *NUDT21*-knockdown experiment in the HeLa cell line. Due to the small sample size (two for each condition wild-type and *NUDT21* knockdown), ceRNAs were identified solely on the basis of miRNA-binding-site overlap. We considered only binding sites for miRNAs with intermediate expression (between 0.01 and 100 in averaged FPM). A total of 1,450 3′US genes identified by DaPars had significant miRNA-binding-site overlap with 1,168 ceRNA genes (3′US ceRNA partners).

MiRNA-Seq for the *NUDT21* depletion experiment. HeLa cells were transfected with control or *NUDT21* siRNA. *NUDT21* depletion was validated as previously described[2]. Small RNA libraries were generated from one control and one *NUDT21* depletion sample using the Illumina Truseq Small RNA Preparation kit, and sequenced on Illumina GAIIx. Raw sequencing reads (40 nucleotides) were obtained using Illumina's Sequencing Control Studio software following image analysis and base-calling by Illumina's Real-Time Analysis (v 1.8.70). Then a script ACGT101-miR v 4.2 (LC Sciences) was used for data analysis, where reads are mapped to the reference database (miRbase). The script also normalizes the counts by a library size parameter for comparison.

CeRNA tumor-suppressor repression in HeLa cells with *NUDT21* knockdown. Parental HeLa cells were purchased from ATCC (cat. no. CCL-2) and maintained in Eagle's minimum essential medium (Lonza, cat. no. 12-604F) with 10% fetal bovine serum. The cells were made mycoplasma free by incubating with Plasmocin (InvivoGen, cat. no. ant-MPT) for two weeks before transfection with three different siRNAs for *NUDT21* (Sigma Aldrich, ID: SASI_Hs01_00146875~77) and negative control siRNA (Sigma Aldrich, ID:SIC002) using previously established approaches[2]. Western blotting was also performed as described in our previous work[2] using antibodies raised against: PHF6 (Santa Cruz, cat. no. sc-271767), YOD1 (abcam, ab178979), NUDT21 (Proteintechlab, cat. no. 10322-1-AP) and GAPDH (Sigma, G9545). To block miRNA function, we selected two miRNAs with a strong trans effect targeting *PHF6* (miR-3187-3p and miR-549) and HeLa cells were co-transfected with siRNA for *NUDT21* and the two antagomirs, to block the two predicted miRNAs, miR-549 and miR-3187-3p in the *PHF6* 3′ UTR. The two antagomirs were designed[37] and synthesized from Sigma-Genosys: Antagomir-3187-3p: 5′-[mU]s[mU]s[mG]mG][mC][mC][mA][mU][mG][mG][mG][mG][mC][mU][mG] [mC][mG]s[mC]s[mG]s[mG]s-chol-3′; and Antagomir-549: 5′-[mU]s[mG]s[mA][mC] [mA][mA][mC][mU][mA][mU][mG][mG][mA][mU][mG][mA][mG][mC]s[mU]s[mC]s[mU]s-chol-3′. PHF6 and YOD1 expression were detected by western blotting and quantified by Image Lab software (version 5.2.1) from Bio-Rad.

Detection of ceRNA tumor-suppressor gene enrichment by RIP with quantitative PCR. HeLa cells were seeded in a 6-well plate at 4×10^5 cells per well and transfected with a *Cas9* and single-guide RNA (sgRNA) plasmid targeting *NUDT21* or with *Cas9* and *GFP* as a control. sgRNAs for *NUDT21* (top, ccggccgcccaatcgctcgcagac; bottom, aaacgtctgcgagcgattg ggcgg) were synthesized (Sigma), and the annealing double-stranded DNA was cloned into pGL3-U6-sgRNA-PGK-puromycin. The transfected cells from three wells were combined and then selected with 10 µg ml^{-1} blasticidin for three days. *NUDT21*-knockdown efficiency was determined by western blot with NUDT21 antibody. RIP was performed with anti-AGO2 antibody, and AGO2-associated RNAs were purified and measured by quantitative real-time PCR[38]. Briefly, the cells were harvested and lysed with 100 µl polysome lysis buffer (100 mM KCl, 5 mM $MgCl_2$, 10 mM Hepes pH 7.0, 0.5% NP50, 1 mM DTT and 1×PI cocktail). The cell lysate was centrifuged at 10,000*g* for 15 min and added to magnetic beads (A+G) with 5 µg anti-Ago2 antibody or normal mouse IgG suspended in 900 µl of NET2 buffer (50 mM Tris-Cl pH 7.4, 150 mM NaCl, 1 mM $MgCl_2$, 0.05% NP-40, 17.5 mM EDTA pH 8.0, 1 mM DTT and 100 units ml^{-1} RNaseOUT). The beads were washed six times with NT2 buffer (50 mM Tris-Cl pH 7.4, 150 mM NaCl, 1 mM $MgCl_2$, 0.05% NP-40). Beads were resuspended in 150 µl proteinase K buffer (50 mM Tris-Cl pH 7.4, 150 mM NaCl, 1 mM $MgCl_2$, 0.05% NP-40 and 1% SDS) with 9 µl proteinase K. Samples were incubated at 55 °C for 30 min and isolate total RNAs with 150 µl phenol–chloroform. The total RNA was reverse transcribed and the candidate ceRNAs were determined by quantitative real-time PCR using primers described in Supplementary Table 5 (Bio-Rad real-time PCR system).

LightSwitch luciferase reporter assay with *PTEN*, *PHF6* and *LARP1* 3′ UTR. LightSwitch luciferase reporter constructs with *PTEN*, *PHF6* and *LARP1* 3′ UTR were purchased from SWITCHGEAR genomics. Briefly, HeLa cells were seeded in a 96-well white TC plate in 100 µl total volume to yield ≥80% confluence at the time of transfection. For each transfection, the following reagents were combined: 50 nM siRNA and/or miRNAs and/or antagomir RNA, individual GoClone reporter (30 ng $\mu\mathrm{l}^{-1}$) 3.33 µl and 1 ng Rluc reporter. Lipofectamine 2000 was diluted in OPTI-MEM medium at 1:10 and incubated at room temperature for 5 min and then added to each tube. Following a 20-min incubation at room temperature, 80 µl of pre-warmed (37 °C) OPTI-MEM medium per replicate was added for a total of 100 µl per replicate transfection. All 100 µl of the transfection mixture was added to each well and incubated overnight. The luciferase reporter assays were performed according to the manufacturer's protocol (Invitrogen).

Immunofluorescence staining for PTEN in MCF7 cells with *EPS* 3′ UTR. pLightSwitch-*EPS15* 3′ UTR construct was purchased from SWITCHGEAR genomics and transfected into MCF7 cells. PTEN expression was detected by immunofluorescence staining with anti-PTEN antibody from Cell Signaling. Briefly, 1×10^5 MCF7 cells were seeded in 4-well chamber slides overnight, and transfected with pLightSwitch-*EPS15* 3′UTR/GFP constructs at 10:1 or pLightSwitch-3′ UTR/GFP constructs as a control. One day after transfection, the cells were fixed with 90% cold methanol at −20 °C overnight. The next day, 0.5% Triton X-100 in PBS was added and incubated at room temperature for 30 min. Samples were blocked in 3% BSA in PBS at room temperature for 1 h. *PTEN* antibody was used at 1:200 dilution in 3% BSA/PBS and 200 µl per well was added to the chamber slides and incubated for 1 h at room temperature. After washing three times, the cells were incubated with Alexa-594-conjugated secondary antibody in 3% BSA/PBS for 1 h at room temperature, in the dark. The cells were rinsed three times with PBS, with the third wash containing DAPI. The coverslips were mounted in anti-fade mounting medium and detected by immunofluorescence microscopy. Both *PTEN*- and *GFP*-positive cells were counted in *EPS15* 3′ UTR/GFP cells and pLightSwitch-3′ UTR/GFP control cells.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The open source MAT3UTR program (version 0.9.2) is freely available at https://github.com/thejustpark/MAT3UTR with necessary example data for this analysis.

Data availability. Raw and processed miRNA-Seq data for the *NUDT21*-depletion experiment have been deposited to GEO under the accession number GSE78198.

References

25. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).
26. Papadopoulos, G. L., Reczko, M., Simossis, V. A., Sethupathy, P. & Hatzigeorgiou, A. G. The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.* **37**, D155–D158 (2009).
27. Xiao, F. et al. miRecords: an integrated resource for microRNA–target interactions. *Nucleic Acids Res.* **37**, D105–D110 (2009).
28. Hsu, S.-D. et al. miRTarBase update 2014: an information resource for experimentally validated miRNA–target interactions. *Nucleic Acids Res.* **42**, D78–D85 (2014).
29. Dvinge, H. et al. The shaping and functional consequences of the microRNA landscape in breast cancer. *Nature* **497**, 378–382 (2013).
30. Hamilton, M. P. et al. Identification of a pan-cancer oncogenic microRNA superfamily anchored by a central core seed motif. *Nat. Commun.* **4**, 2730 (2013).
31. Goldman, M. et al. The UCSC Cancer Genomics Browser: update 2013. *Nucleic Acids Res.* **41**, D949–D954 (2013).
32. Friedman, J. M., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
33. Kim, S., Baladandayuthapani, V. & Lee, J. J. Prediction-oriented marker selection (PROMISE): with application to high-dimensional regression. *Stat. Biosci* **9**, 217–245 (2016).
34. Bayerlová, M. et al. Newly constructed network models of different WNT signaling cascades applied to breast cancer expression data. *PLoS ONE* **10**, 1–19 (2015).
35. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
36. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
37. Krutzfeldt, J. et al. Silencing of microRNAs in vivo with 'antagomirs'. *Nature* **438**, 685–689 (2005).
38. Tenenbaum, S. A., Lager, P. J., Carson, C. C. & Keene, J. D. Ribonomics: identifying mRNA subsets in mRNP complexes using antibodies to RNA-binding proteins and genomic arrays. *Methods* **26**, 191–198 (2002).

nature research

Corresponding Author: Eric Wagner and Wei Li

Date: Jan 19, 2018

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.	computational analyses used all 97 breast tumor samples that have matched normal breast samples in the database.

2. Data exclusions

Describe any data exclusions.	no data were excluded

3. Replication

Describe whether the experimental findings were reliably reproduced.	all attempts at replication were successful

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.	We used all 97 breast tumor samples that have matched normal breast samples in the database. Since there's no sampling, there's no randomization.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.	There was no group allocation, because we used all 97 samples of normal and tumor conditions. Since there's no group allocation (or sampling), blinding is not relevant.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. p values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

All statistical computations were performed in python scipy stats package (version 0.15.1) or R (version 3.1.1). Gene differential expression analysis was done by edgeR (version 3.8.6).

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No restrictions on availability of unique materials.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

All the antibodies used are commercial available and validated indicated in company website and their references.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

HeLa and MCF7 cell lines are purchased from ATCC.

b. Describe the method of cell line authentication used.

The cell lines are authenticated by the provider and store in our lab with low passage.

c. Report whether the cell lines were tested for mycoplasma contamination.

They were tested and showed mycoplasma free.

d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

None of the cell lines is listed in the database of commonly misidentified cell lines maintained by ICLAC.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

no animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

the study did not involve human research participants.

doi:10.1038/nature13261

CFIm25 links alternative polyadenylation to glioblastoma tumour suppression

Chioniso P. Masamha[1]*, Zheng Xia[2]*, Jingxuan Yang[3], Todd R. Albrecht[1], Min Li[3], Ann-Bin Shyu[1], Wei Li[2] & Eric J. Wagner[1]

The global shortening of messenger RNAs through alternative polyadenylation (APA) that occurs during enhanced cellular proliferation represents an important, yet poorly understood mechanism of regulated gene expression[1,2]. The 3′ untranslated region (UTR) truncation of growth-promoting mRNA transcripts that relieves intrinsic microRNA- and AU-rich-element-mediated repression has been observed to correlate with cellular transformation[3]; however, the importance to tumorigenicity of RNA 3′-end-processing factors that potentially govern APA is unknown. Here we identify CFIm25 as a broad repressor of proximal poly(A) site usage that, when depleted, increases cell proliferation. Applying a regression model on standard RNA-sequencing data for novel APA events, we identified at least 1,450 genes with shortened 3′ UTRs after CFIm25 knockdown, representing 11% of significantly expressed mRNAs in human cells. Marked increases in the expression of several known oncogenes, including cyclin D1, are observed as a consequence of CFIm25 depletion. Importantly, we identified a subset of CFIm25-regulated APA genes with shortened 3′ UTRs in glioblastoma tumours that have reduced CFIm25 expression. Downregulation of CFIm25 expression in glioblastoma cells enhances their tumorigenic properties and increases tumour size, whereas CFIm25 overexpression reduces these properties and inhibits tumour growth. These findings identify a pivotal role of CFIm25 in governing APA and reveal a previously unknown connection between CFIm25 and glioblastoma tumorigenicity.

Recently, it has become increasingly clear that mRNA 3′-end formation is subject to dynamic regulation under diverse physiological conditions[2–5]. Over 50% of human genes have multiple polyadenylation signals, thereby increasing the potential diversity in mRNA transcript length[6]. The formation of mRNA transcripts using these distinct poly(A) sites (PASs) is carried out by APA, with the most common form involving differential use of alternative PASs located within the same terminal exon (reviewed in ref. 7). Processing at the PAS most proximal to the stop codon (pPAS) removes negative regulatory elements that reduce mRNA stability or impair translation efficiency, such as AU-rich elements (AREs)[8] and microRNA (miRNA) targeting sites[9,10]. It has been reported that both rapidly proliferating cells[1,2] and transformed cells[3,11] preferentially express mRNAs with shortened 3′ UTRs. Despite these observations, the mechanisms that control the extensive distal-to-proximal PAS switch observed in proliferative and/or transformed cells, the relationship between cause and effect, and the critical target genes subject to this regulation, are not well characterized.

To measure relative changes in endogenous APA events, we devised a quantitative polymerase chain reaction after reverse transcription (qRT–PCR) assay to monitor the transcript-specific use of the distal PAS (dPAS) while normalizing for total mRNA levels for three test transcripts, cyclin D1 (*CCND1*), *DICER1* and *TIMP2*, known to undergo APA[3,12]. Using this approach, we readily detected appreciable usage of dPASs for all three genes in HeLa cells (Extended Data Fig. 1). This was somewhat surprising given their highly transformed state, but is consistent with previous reports that not all transformed cells tested exhibit appreciable 3′ UTR shortening[1,3]. Previous studies implicate multiple members of the cleavage and polyadenylation (CPA) machinery as potentially regulating poly(A) site selection[12–15]. To test the relative contribution of these factors to the APA of the three test genes, we used systematic RNA interference (RNAi) (Fig. 1a–c). We observed only small changes in the relative use of the dPAS after knockdown of members of the cleavage and polyadenylation specificity factor (CPSF), cleavage stimulation factor (CSTF) and cleavage factor IIm (CFIIm) complexes (Fig. 1d–f). By contrast, we detected significant reduction in dPAS usage after knockdown of the members of the CFIm complex. These results are consistent with a recent report that CFIm68 depletion decreases 3′ UTR length[14]; however, the most notable PAS switching was found to occur after knockdown of CFIm25. We therefore focused all further analyses on CFIm25.

Traditional methods of global PAS profiling use mRNA partitioning and digestion to sequence poly(A) junctions within messages[1,16,17]. To identify global targets of CFIm25 with a more streamlined approach requiring less sample manipulation, we performed high-depth ($>3 \times 10^8$ reads) RNA sequencing (RNA-seq) after knocking down CFIm25 in parallel with a control knockdown. We determined that 23% of RNA-seq reads can be uniquely mapped to 3′ UTRs of expressed genes leading to approximately 200-fold sequence coverage (Extended Data Fig. 2a, b). We first analysed the three test genes and observed markedly reduced read density within the 3′ UTRs in response to CFIm25 depletion (Fig. 2a). These results not only confirm our qRT–PCR findings that HeLa cells robustly use the dPAS for all three test genes under basal conditions but also demonstrate that considerable 3′ UTR shortening induced by CFIm25 knockdown is readily visualized by analysing the read density of RNA-seq data.

On the basis of this promising observation, we applied a novel bioinformatics algorithm termed ‘dynamic analysis of alternative polyadenylation from RNA-seq’ (DaPars; see Methods) for the *de novo* identification of all instances of 3′ UTR alterations between control and CFIm25 knockdown cells, regardless of a pre-annotated pPAS within each RefSeq transcript. DaPars uses a linear regression model to identify the exact location of this novel proximal 3′ UTR as the optimal fitting point (Fig. 2b, red point) as well as the abundance of both novel and annotated UTRs. The degree of difference of 3′ UTR usage between the samples was then quantified as a change in percentage dPAS usage index (ΔPDUI), which is capable of identifying lengthening (positive index) or shortening (negative index) within the 3′ UTR. When applied to the 12,273 RefSeq transcripts whose average terminal exon sequence coverage is more than 30-fold, DaPars identified 1,453 transcripts possessing a significant, reproducible shift in 3′ UTR usage in response to CFIm25 depletion (Fig. 2c and Extended Data Fig. 2c, d). Notably, among this group of transcripts, 1,450 are shifted to pPAS usage in CFIm25 knockdown cells. We found a significant enrichment of the CFIm25 UGUA binding motif and previously reported CFIm25 iCLIP sequence tags[14] within 3′ UTRs that shortened after CFIm25 knockdown relative to transcripts exhibiting no length change (Extended Data Fig. 3).

[1]Department of Biochemistry and Molecular Biology, The University of Texas Medical School at Houston, Houston, Texas 77030, USA. [2]Division of Biostatistics, Dan L Duncan Cancer Center and Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, 77030 Texas, USA. [3]The Vivian L. Smith Department of Neurosurgery, The University of Texas Medical School at Houston, Houston, Texas 77030, USA.

*These authors contributed equally to this work.

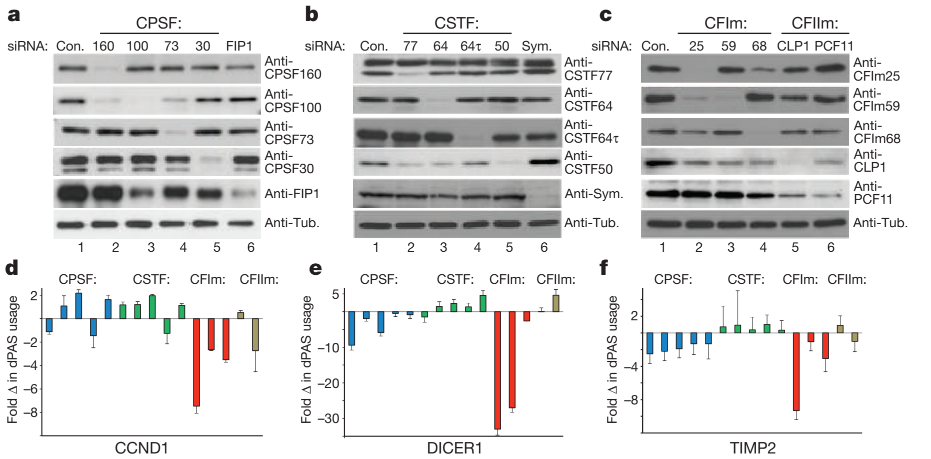


Figure 1 | CFIm25 depletion leads to consistent and robust 3′ UTR shortening of test genes. **a**–**c**, Western blot analysis of HeLa cell lysates treated with control siRNA (Con.) and siRNAs individually targeting each of members of the CPA machinery and Symplekin (Sym.). In all cases, tubulin (Tub.) was used as a loading control. **d**–**f**, Quantified results of three biologically independent qRT–PCR experiments on RNA isolated from cells represented in panels **a**–**c** with the factors presented in the same order as shown in western blots **a**–**c**. See Methods for quantification details.

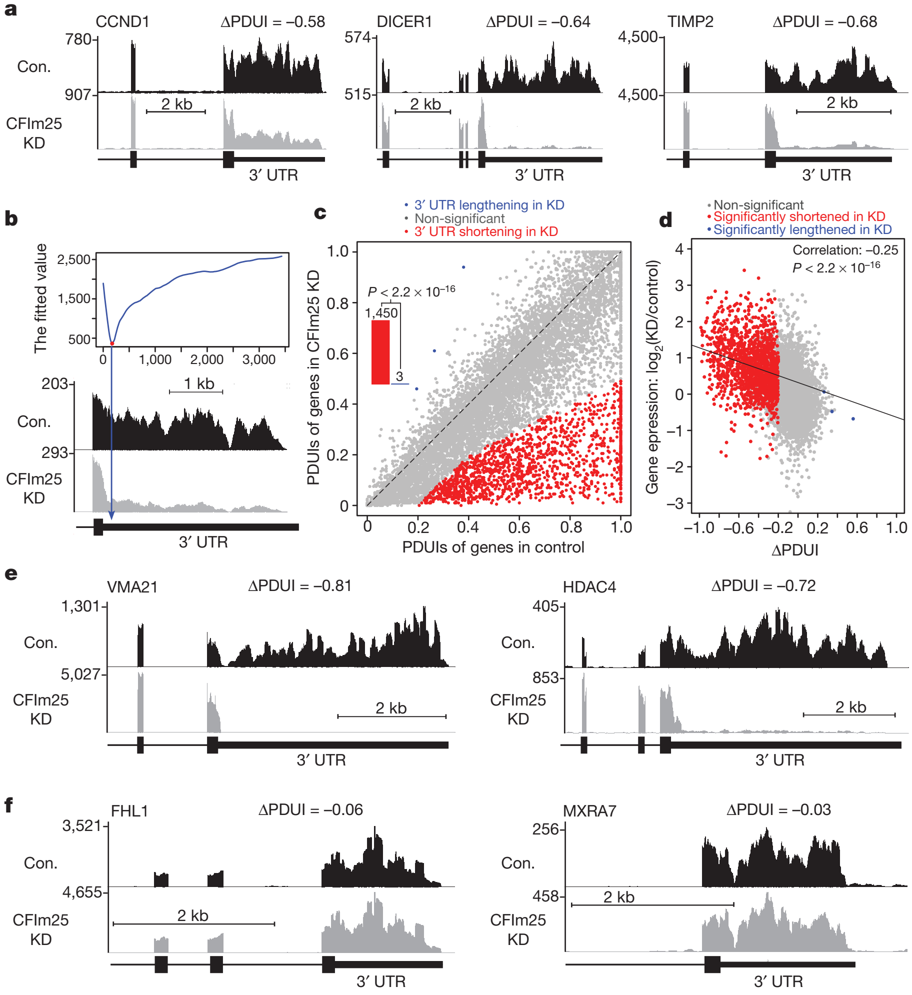


Figure 2 | The DaPars algorithm identifies broad targets of CFIm25 in standard RNA-seq data. **a**, RNA-seq read density for 3′ UTR, terminal exon and upstream exon(s) after the control (Con.) siRNA treatment and CFIm25 knockdown (KD) in HeLa cells. Numbers on *y*-axis indicate RNA-seq read coverage. **b**, Diagram depicts how the differential alternative 3′ UTR usage was identified based on DaPars. The *y*-axis shows the fitting value of the DaPars regression model and the locus with minimum fitting value (red point) is the predicted alternative pPAS for the RNA-seq data (bottom). **c**, Scatterplot of PDUIs in control and CFIm25 knockdown cells where mRNAs significantly shortened ($n = 1{,}450$) or lengthened ($n = 3$) after CFIm25 knockdown (false discovery rate (FDR) ≤ 0.05, absolute ΔPDUI ≥ 0.2 and at least twofold change of PDUIs between CFIm25 knockdown and control cells) are coloured. The shifting towards pPAS is significant ($P < 2.2 \times 10^{-16}$, binomial test). **d**, Correlation between dPAS site usage and gene expression levels of control and CFIm25 knockdown cells. The *x*-axis shows ΔPDUI; a negative value indicates that pPAS is prone to be used in CFIm25 knockdown cells. The *y*-axis shows the logarithm of the expression level of genes from the CFIm25 knockdown relative to the control sample. **e**, Representative RNA-seq density plots along with ΔPDUI values for genes whose 3′ UTR is shortened in response to CFIm25 knockdown. Numbers on *y*-axis indicate RNA-seq read coverage. **f**, Representative RNA-seq density plots along with ΔPDUI values of genes whose 3′ UTR is unchanged by CFIm25 knockdown.

Moreover, we determined that 70% of transcripts whose 3′ UTR is shortened after CFIm25 knockdown use a pPAS within the first one-third of their 3′ UTR. By contrast, only 29% of multi-PAS transcripts that did not alter 3′ UTR length in response to CFIm25 have an annotated pPAS in the first third of their 3′ UTR. This demonstrates that CFIm25 APA targets are enriched with pPASs positioned close to the stop codon to maximize their degree of 3′ UTR shortening. Collectively, these results clearly indicate that the function of CFIm25 is to broadly repress proximal poly(A) site choice, and consequently, the shortening of 3′ UTR length is considerable for the majority of CFIm25-regulated transcripts upon its depletion.

One potential consequence of 3′ UTR shortening in CFIm25 knockdown is the loss of miRNA-binding sites and/or AREs, resulting in truncated mRNA transcripts that evade negative regulation. Although the correlation between transcript expression change and ΔPDUI was modest (Pearson correlation = −0.25), it does reveal that transcripts with shorter 3′ UTR in CFIm25 knockdown cells have overall higher expression levels (Fig. 2d). We observed that 64% of transcripts with shortened 3′ UTRs exhibited significantly increased steady-state levels, 34% were unchanged, and only 2% were significantly reduced (Extended Data Fig. 4). We have also organized the list of CFIm25-regulated genes with respect to their ΔPDUI score, change in relative levels of transcript, and predicted numbers of ARE motifs and miRNA target sites lost after APA (Supplementary Table 1) and observed that gene expression positively correlates with the number of lost ARE motifs and miRNA target sites (Extended Data Fig. 5). Several examples of novel genes whose APA is regulated by CFIm25 are shown in Fig. 2e and it is important to note that not all long 3′ UTRs were observed to shorten in response to CFIm25 knockdown, indicating that the CFIm complex regulates many, but not all genes capable of APA (Fig. 2f). Collectively, these data demonstrate the power and ease of the DaPars algorithm to identify APA within standard RNA-seq, and indicate that the major form of CFIm25 regulation is to repress pPAS choice at a global level.

To validate the ΔPDUI results, we created qRT–PCR amplicons to monitor dPAS usage of six genes whose 3′ UTRs were found to be shortened after CFIm25 knockdown and two that were not altered. Using these amplicons, we analysed RNA isolated from cells effectively depleted of CFIm25 using two independent short interfering RNAs (siRNAs) (Fig. 3a, inset), and observed high congruence between qRT–PCR results and those obtained using RNA-seq and ΔPDUI (Fig. 3a, graph). To test formally for the presence of de-repressed protein expression from mRNAs with shortened 3′ UTRs, we measured their levels in lysates from knockdown cells (Fig. 3b). We observed considerable increases in protein levels of CFIm25 target genes, including several that have a well-documented role in tumour growth, such as cyclin D1, glutaminase and methyl-CpG-binding protein 2 (MECP2)[18–22]. It is worth noting that the 3′ UTR of each of these genes has been shown to be subject to miRNA-mediated inhibition[23–25]. Consistent with this observation, we also noted enhanced cellular proliferation in response to knockdown of CFIm25 relative to control knockdown in HeLa cells (Fig. 3c). Finally, to determine whether the 3′ UTR is sufficient to elicit translational de-repression of a heterologous protein in response to CFIm25 knockdown, we used reporters with the SMOC1 3′ UTR cloned downstream of luciferase or the GAPDH 3′ UTR, which was not found to alter its poly(A) site usage. We observed that only the luciferase activity specifically resulting from the luciferase–SMOC1 reporter was increased in response to knockdown of CFIm25 (Fig. 3d), supporting the idea that the increased expression of endogenous SMOC1 protein when CFIm25 is depleted is mediated through its 3′ UTR.

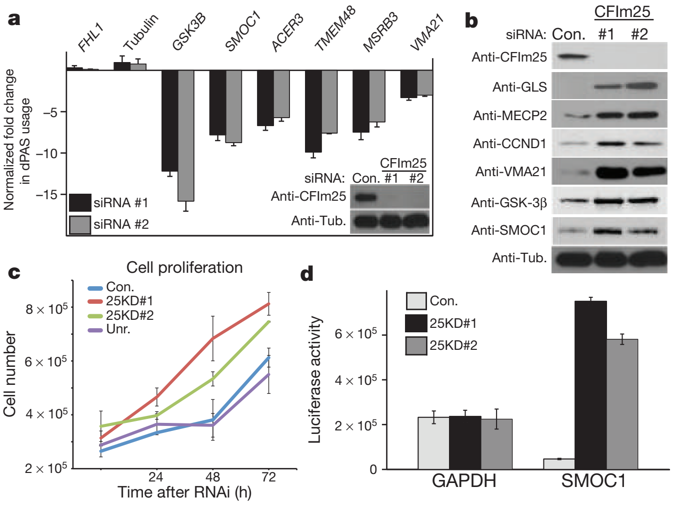


Figure 3 | Increased pPAS usage after CFIm25 depletion results in increased protein translation and enhanced cell proliferation. **a**, qRT–PCR results of select genes shown as fold change in dPAS usage after CFIm25 depletion. Experiments were performed in triplicate with data shown as mean + standard deviation from the mean (s.d.). The inset shows western blot analysis demonstrating effective knockdown of CFIm25 using two distinct siRNAs. Tub., tubulin. **b**, Results of western blot analysis of cell lysates after knockdown of CFIm25 using siRNA. **c**, Growth of HeLa cells was measured after RNAi of CFIm25 compared with cells transfected with control siRNA or the siRNA to the CFIIm complex subunit PCF11 (Unr.). Results shown are mean ± standard deviation (s.d.) ($n = 3$). **d**, Graph representing luciferase activity from cells transfected with a luciferase reporter containing the 3′ UTR of either GAPDH or of SMOC1 after being transfected with either control or CFIm25 siRNA. Data are the average of three independent experiments and error bars show s.d.

The collective observations that CFIm25 depletion leads to broad 3′ UTR shortening, enhanced expression of growth promoting genes and increased cell proliferation support the hypothesis that CFIm25 is a novel anti-proliferative gene whose levels may be reduced in human cancers. We focused our analysis on glioblastoma, as recent reports indicate that brain tissue possesses the longest 3′ UTRs[26,27]. We reasoned that tumours derived from these cells might be more sensitive to changes in CFIm25 levels than other cancers. To test this prediction, we downloaded archived patient RNA-seq data from The Cancer Genome Atlas (TCGA), stratified it according to CFIm25 expression, and analysed it using DaPars. Indeed, following the same cut-offs in our HeLa RNA-seq 3′ UTR analysis, we identified 60 genes with altered 3′ UTRs, with 59 of those experiencing shortening in glioblastoma expressing lower levels of CFIm25 (Fig. 4a and Supplementary Table 2). Among those genes, a significant number of events (24 genes; $P = 2.2 \times 10^{-12}$ by hypergeometric testing) were also shortened in CFIm25 knockdown HeLa cells and this percentage of overlap increased markedly to 86% as the ΔPDUI cut-off increased from 0.2 to 0.4 (Extended Data Fig. 6). Two representative examples of genes, FOS-related antigen 2 (*FRA2*; also known as *FOSL2*) and *MECP2*, with shortened 3′ UTRs in low CFIm25-expressing glioblastoma tumours is shown in Fig. 4b, demonstrating a compelling similarity between the patient samples and HeLa cells before and after CFIm25 knockdown. Overexpression of either of these genes has been shown to enhance cell proliferation[18,28].

To test formally whether altering CFIm25 expression can modulate glioblastoma tumorigenic properties, we screened a panel of glioblastoma cell lines and observed that U251 cells naturally express lower levels of CFIm25 compared with LN229 cells (Fig. 4c). To raise CFIm25 levels in U251 cells, we created cell lines stably expressing either Myc-tagged CFIm25 or green fluorescent protein (GFP) as a control. In parallel, we used RNAi to reduce CFIm25 levels in LN229 cells (Fig. 4c). We observed a significant reduction in anchorage-dependent growth and cellular invasion in U251 cells overexpressing CFIm25 compared with the GFP control, whereas reducing CFIm25 in LN229 cells caused an increase in both of these properties (Extended Data Fig. 7). To determine if the altered *in vitro* properties of glioblastoma cells affected tumour growth kinetics *in vivo*, we used a subcutaneous xenograft model. Increased expression of CFIm25 in U251 cells resulted in a marked reduction in tumour growth and decreased tumour cell proliferation (Fig. 4d and Extended Data Fig. 8). By contrast, depletion of CFIm25 in LN229 cells caused a profound increase in tumour size (Fig. 4e and Extended

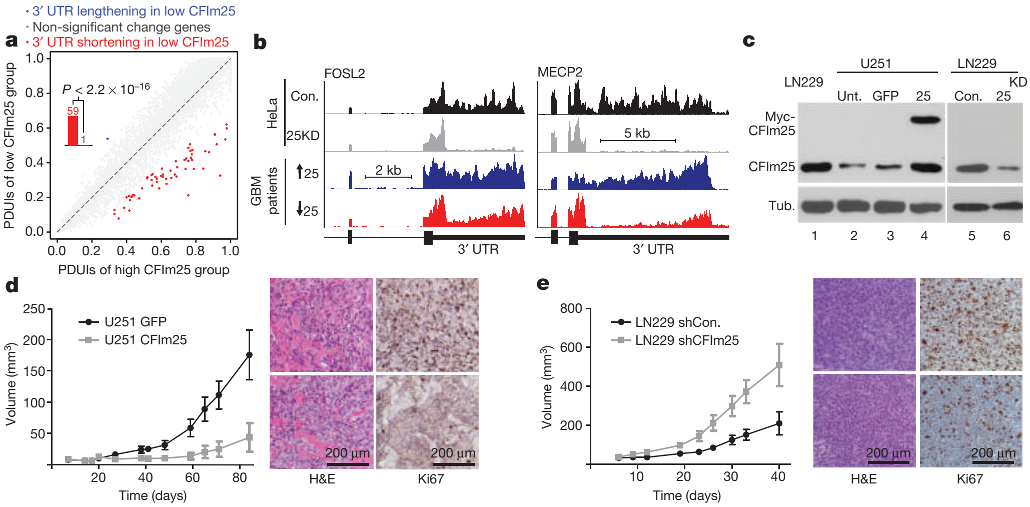


Figure 4 | Altered expression of CFIm25 modulates glioblastoma tumour growth. a, The global analysis of 3′ UTR changes in glioblastoma (GBM) patient samples with either high or low levels of CFIm25. Scatterplot of PDUIs from both data sets using the same cut-offs as in Fig. 2c. The shifting to pPAS in the low CFIm25 group is significant ($P < 2.2 \times 10^{-16}$; binomial test). **b**, Representative UCSC Genome Browser images of RNA-seq data, demonstrating 3′ UTR shortening after CFIm25 knockdown in HeLa cells and in glioblastoma patient samples with high (blue) or low CFIm25 expression (red). KD, knockdown. **c**, Western blot analysis of lysates from two glioblastoma cell lines. Note that the overexpressed Myc–CFIm25 also increases endogenous CFIm25 levels in U251 cells. Tub., tubulin; Unt., untreated. **d**, Growth comparison of U251 tumours overexpressing either GFP (control) or CFIm25. Data represent the average of ten mice per group. Right panel shows representative haematoxylin and eosin (H&E) and Ki67 staining of U251 GFP tumours (top) or U251 CFIm25 tumours (bottom). Scale bars, 200 μm. **e**, Growth comparison of LN229 tumours derived from cells transduced with lentiviruses expressing a scrambled short hairpin RNA (shRNA) (control) or with lentiviruses expressing shRNA targeting CFIm25. Data represent the average of ten mice per group. Right panel shows representative H&E and Ki67 staining of LN229 tumours expressing shRNA targeting CFIm25 (top) or LN229 tumours expressing scrambled shRNA (bottom). Scale bars, 200 μm.

Data Fig. 9). Collectively, these results uncover a tumour suppressive property of CFIm25 in glioblastoma that is probably mediated through its broad repression of APA-dependent mRNA 3′ UTR shortening.

We identified CFIm25 among 15 cleavage and polyadenylation factors as a key factor that broadly regulates APA. Importantly, the data presented here also extend our understanding of APA in regulated gene expression through the demonstration that extensive shortening of 3′ UTRs causally leads to enhanced cellular proliferation and tumorigenicity, probably through the upregulation of growth promoting factors, such as cyclin D1. These results indicate the importance of 3′ UTR usage in cell growth control and underscore the need for further research into the mechanism and regulation of APA and its potential links to other human diseases.

METHODS SUMMARY

Human cell lines used were cultured using standard techniques. RNAi and western blot experiments were conducted as described previously[29]. For luciferase experiments, one day after the second siRNA hit, cells were transfected with 3′ UTR *Renilla* luciferase plasmids and activity was assayed after 24 h. Total RNA for pRT–PCR was reverse transcribed using MMLV-RT (Invitrogen). qRT–PCR reactions were performed using SYBRGREEN (Fermentas). Duplicate control and CFIm25 knockdown samples were sequenced by HiSeq 2000. RNA-seq reads were aligned (hg19) using TopHat 2.0.10[30]. All the TCGA glioblastoma RNA-seq BAM files were downloaded from the UCSC Cancer Genomics Hub (https://cghub.ucsc.edu/). DaPars was used to identify differential 3′ UTR usage from RNA-seq (Z.X. *et al.*, unpublished observations; https://code.google.com/p/dapars). For tumour xenografts, U251 cells were stably transfected with GFP or CFIm25 plasmids. LN229 cells were transfected with lentivirus expressing CFIm25 shRNA. After subcutaneous injection of cell lines into nude mice, glioblastoma tumour size was monitored and tumours were removed and histologically analysed.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 January 2013; accepted 13 March 2014.

Published online 11 May 2014.

1. Elkon, R. *et al.* E2F mediates enhanced alternative polyadenylation in proliferation. *Genome Biol.* **13,** R59 (2012).
2. Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A. & Burge, C. B. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* **320,** 1643–1647 (2008).
3. Mayr, C. & Bartel, D. P. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138,** 673–684 (2009).
4. Ji, Z., Lee, J. Y., 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. USA* **106,** 7028–7033 (2009).
5. Mangone, M. *et al.* The landscape of *C. elegans* 3′UTRs. *Science* **329,** 432–435 (2010).
6. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33,** 201–212 (2005).
7. Di Giammartino, D. C., Nishida, K. & Manley, J. L. Mechanisms and consequences of alternative polyadenylation. *Mol. Cell* **43,** 853–866 (2011).
8. Chen, C.-Y. A. & Shyu, A.-B. AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem. Sci.* **20,** 465–470 (1995).
9. Farh, K. K. *et al.* The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310,** 1817–1821 (2005).
10. Wu, L. & Belasco, J. G. Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. *Mol. Cell* **29,** 1–7 (2008).
11. Singh, P. *et al.* Global changes in processing of mRNA 3′ untranslated regions characterize clinically distinct cancer subtypes. *Cancer Res.* **69,** 9422–9430 (2009).
12. Kubo, T., Wada, T., Yamaguchi, Y., Shimizu, A. & Handa, H. Knock-down of 25 kDa subunit of cleavage factor Im in Hela cells alters alternative polyadenylation within 3′-UTRs. *Nucleic Acids Res.* **34,** 6264–6271 (2006).
13. Yao, C. *et al.* Transcriptome-wide analyses of CstF64–RNA interactions in global regulation of mRNA alternative polyadenylation. *Proc. Natl Acad. Sci. USA* **109,** 18773–18778 (2012).
14. Martin, G., Gruber, A. R., 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 Reports* **1,** 753–763 (2012).
15. Thomas, P. E. *et al.* Genome-wide control of polyadenylation site choice by CPSF30 in *Arabidopsis. Plant Cell* **24,** 4376–4388 (2012).
16. Jan, C. H., Friedman, R. C., Ruby, J. G. & Bartel, D. P. Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. *Nature* **469,** 97–101 (2011).
17. Shepard, P. J. *et al.* Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17,** 761–772 (2011).
18. Bernard, D. *et al.* The methyl-CpG-binding protein MECP2 is required for prostate cancer cell growth. *Oncogene* **25,** 1358–1366 (2006).
19. Sicinski, P. *et al.* Cyclin D1 provides a link between development and oncogenesis in the retina and breast. *Cell* **82,** 621–630 (1995).
20. Weinstat-Saslow, D. *et al.* Overexpression of cyclin D1 mRNA distinguishes invasive and *in situ* breast carcinomas from non-malignant lesions. *Nature Med.* **1,** 1257–1260 (1995).

21. Liu, W. *et al.* Reprogramming of proline and glutamine metabolism contributes to the proliferative and metabolic responses regulated by oncogenic transcription factor c-MYC. *Proc. Natl Acad. Sci. USA* **109,** 8983–8988 (2012).
22. Wang, J. B. *et al.* Targeting mitochondrial glutaminase activity inhibits oncogenic transformation. *Cancer Cell* **18,** 207–219 (2010).
23. Klein, M. E. *et al.* Homeostatic regulation of MeCP2 expression by a CREB-induced microRNA. *Nature Neurosci.* **10,** 1513–1514 (2007).
24. Deshpande, A. *et al.* 3′UTR mediated regulation of the cyclin D1 proto-oncogene. *Cell Cycle* **8,** 3592–3600 (2009).
25. Gao, P. *et al.* c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism. *Nature* **458,** 762–765 (2009).
26. Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J. O. & Lai, E. C. Widespread and extensive lengthening of 3′ UTRs in the mammalian brain. *Genome Res.* **23,** 812–825 (2013).
27. Ulitsky, I. *et al.* Extensive alternative polyadenylation during zebrafish development. *Genome Res.* **22,** 2054–2066 (2012).
28. Nakayama, T. *et al.* Aberrant expression of Fra-2 promotes CCR4 expression and cell proliferation in adult T-cell leukemia. *Oncogene* **27,** 3221–3232 (2008).
29. Wagner, E. J. & Garcia-Blanco, M. A. RNAi-mediated PTB depletion leads to enhanced exon definition. *Mol. Cell* **10,** 943–949 (2002).
30. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank members of the E.J.W., A.-B.S. and W.L. laboratories for helpful discussions, P. Carpenter for reviewing the manuscript, and Q. Zhu and T. Shan of LC Sciences for their efforts on the RNA-seq. This work was supported by a CPRIT grant to E.J.W. and A.-B.S. (RP100107), and in part by a Department of Defense grant to E.J.W. (W81XWH-11-1-0304), National Institutes of Health (NIH) grants to A.-B.S. (GM046454) and E.J.W. (CA167752 and CA166274), and an endowment from Houston Endowment (to A.-B.S.). Work in the W.L. laboratory is funded by grants from the Department of Defense (W81XWH-10-1-0501), CPRIT (RP110471-C3) and NIH (R01HG007538). Work in the M.L. laboratory is funded by grants from the Dr Marnie Rose Foundation and the William and Ella Owens Medical Research Foundation. C.P.M. acknowledges a Department of Defense Postdoctoral Visionary Award (W81XWH-12-1-0218).

Author Contributions A.-B.S., W.L. and E.J.W. designed the study. C.P.M., T.R.A. and J.Y. performed the described experiments with conceptual advice from M.L. W.L. and Z.X. conducted bioinformatic analyses and developed the DaPars algorithm. C.P.M., Z.X., W.L., A.-B.S. and E.J.W. wrote the manuscript.

Author Information Raw sequence data has been deposited in the Gene Expression Omnibus under accession number GSE42420. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.J.W. (Eric.J.Wagner@uth.tmc.edu), A.-B.S. (Ann-Bin.Shyu@uth.tmc.edu) or W.L. (WL1@bcm.edu).

METHODS

RNA-seq. We used whole transcriptome RNA-seq to investigate alternative PAS usage in a genome-wide fashion. Two control and two CFIm25 knockdown samples were sequenced by HiSeq 2000 (LC Sciences). Paired-end RNA-seq reads with 101 bp in each end were aligned to the human genome (hg19) using TopHat 2.0.10[30]. RefSeq gene expressions were quantified by RSEM[31]. A statistical summary of read alignments and average gene expressions can be found in Extended Data Fig. 2. More than 12,000 (~50%) human RefSeq genes can be detected through RNA-seq with expression levels more than 1 fragments per kilobase of transcript sequence per million mapped paired-end reads (FPKM)[32]. More importantly, the average of 23% of RNA-seq reads can be uniquely mapped to 3′ UTRs of expressed genes that renders around 200× coverage on UTRs. All the TCGA glioblastoma RNA-seq BAM files were downloaded from the UCSC Cancer Genomics Hub (CGHub; https://cghub.ucsc.edu/).

Analysis of APA from RNA-seq. We used a novel bioinformatics algorithm DaPars (Z.X. *et al.*, unpublished observations; https://code.google.com/p/dapars) for the *de novo* identification of APA from RNA-seq. The observed sequence coverage was represented as a linear combination of novel and annotated 3′ UTRs. For each RefSeq transcript with annotated PAS, we used a regression model to infer the end point of alternative novel PAS within this 3′ UTR at single nucleotide resolution, by minimizing the deviation between the observed read coverage and the expected read coverage based on a two-PAS model, in both control and CFIm25 knockdown samples simultaneously.

To quantify the relative PAS usage, we defined the percentage of dPAS usage for each sample as PDUI index. The greater the PDUI is, the more the dPAS of a transcript is used and vice versa.

ΔPDUI. We used the following three criteria to detect the most significant shifted 3′ UTR events: First, given the expression levels of short and long 3′ UTRs in two samples in each condition, we compute the significance of the difference of mean PDUIs using Fisher's exact test, which is further adjusted by Benjamini–Hochberg (BH) procedure to control the FDR at a level of 5%. Second, the absolute difference of mean PDUIs must be no less than 0.2. Third, the absolute $\log_2$ ratio (fold change) of mean PDUIs must be no less than 1. To avoid false positive estimation on low coverage transcripts, we required that there be more than 30-fold coverage on the 3′ UTR region of both samples. For genes with multiple annotated PASs, we only kept the one with the greatest absolute ΔPDUI value. Last, we identified 1,453 transcripts possessing a significant shift in 3′ UTR usage in response to CFIm25 knockdown, the vast majority of which have shortened 3′ UTRs in CFIm25 knockdown.

Bioinformatic analyses of 3′ UTR shortening. As miRNA binding sites and other regulatory sequences such as AREs reside in 3′ UTRs[33,34], APA has an important role in mRNA stability, translation and translocation. Indeed, it has been reported that shorter 3′ UTRs produce higher levels of protein[3]. To elucidate the consequences of 3′ UTR shortening, we provided the numbers of lost ARE motifs and miRNA binding sites due to the 3′ UTR shortening for the transcripts shifting to proximal 3′ UTR usage in CFIm25 knockdown cells (Supplementary Table 1). The ARE is one of the most prominent *cis*-acting regulatory elements found in 3′ UTRs to target mRNAs for rapid degradation[35]. The eight different consensus ARE motifs, including the plain AUUUA pentamer, were retrieved from the ARE site database[35]. miRNA–mRNA binding information was based on miRNA target prediction database TargetScanHuman version 6.2[36–38]. To limit the miRNA to high-confidence sites, we required the probability of the preferentially conserved targeting (PCT) score to be more than 0 for all highly conserved miRNA families[38].

Differentially expressed gene expression analysis. With two replicates in each group, we used edgeR[39] to call differentially expressed genes with FDR < 0.05. To better quantify gene expression with shorter 3′ UTRs, we counted reads based on the coding regions of each transcript.

Cell culture and cell counts. All the cell lines used (HeLa, U251 and LN229) were cultured in DMEM supplemented with 10% FBS (+1% penicillin and streptomycin) in a 5% CO_2 incubator at 37 °C. Cell counts were done using a standard hemacytometer.

siRNA and western blot assays. Both siRNA transfection and western blot analysis were performed as previously described[29]. The siRNA was purchased from Sigma and all the siRNAs used are shown below. After transfection, cells were harvested for mRNA extraction, western blotting or Matrigel assay. To detect 3′-end-processing factors by western blotting, the following primary antibodies from Bethyl Laboratories were used: CPSF160, CPSF100, CPSF73, CPSF30, FIP1, CSTF77, CSTF64τ, CSTF50, CFIm68 and CFIm59. Other antibodies used include CFIm25 (PTGlabs), CSTF64 and CFIIm PCF11, and Symplekin (Sigma and CFIIm CLP1 (Epitomics). Additional antibodies include VMA21, GLS, ACER3 and GSK-3β (PTGlabs); cyclin D1 (Cell Signaling); and SMOC1 and tubulin (Abcam).

siRNA sequences. We used the following siRNA sequences. CPSF160 si1: 5′-GCUUUAAGAAGGUCCCUCA; si2: 5′-CUUACCACGUGGAGUCUAA; CPSF100 si1: 5′-CUCAACUUCUUGAUCAGAU; si2: 5′-GGAUAGAUGGUGUCUUAGA; CPSF73 si1: 5′-CCAUAUACUGGUCCCUUUA; si2: 5′-GAUAUUGGAAGUUCAGUCA; CPSF30 si1: 5′-GUGCCUAUAUCUGUGAUUU; si2: 5′-CCUAUAUCUGUGAUUUGAA; FIP1 si1: 5′-CGAAUGGGACUUGAAGUUA; si2: 5′-GACAAGUACUGCCUCCAGA; CSTF77 si1: 5′-GAAGACUUAUGAACGCCUU; si2 5′-CACAGAAUCAACCUAUAGA; CSTF64 si1: 5′-GGCUUUAGUCCCGGGCAGA; si2: 5′-GGUUAUGGCUUCUGUGAAU; CSTF64τ si1: 5′-GUCUUAGAGACACGUGUAA; si2: 5′-CUAAUGUUCUGCUGAACCA; CSTF50 si1: 5′-GUCGUAAGUCCGUGCACCA; si2: 5′-CUACUCUUCGCCUUUAUGA; Symplekin si1: 5′-CAGUUCAACUCGGGCCUGA; si2: 5′-GAGACAUUGAGUUGCUGCU; CFIm25 si1: 5′-CCUCUUACCAAUUAUACUU; si2: 5′-GCUAUAUACAGUGUAGAAU; CFIm59 si1: 5′-CUCAUCUGCUCGUGUGGAU; si2: 5′-GCAAUUUCCAGCAGUGCCA; CFIm68 si1: 5′-CUGCAAUUUCUUUAAUUAA; si2: 5′-GGAUCAAGACGUGAACGAU; CFIIm CLP1 si1: 5′-GCUUAUGUCUCCAAGGACA; si2: 5′-CAGUUCAGUUGGAGUUGUU; CFIIm PCF11 si1: 5′-GUACCUUAUGGAUUCUAUU; si2: 5′-GUAUCUCACUGCCUUUACU) and the control siRNA used was described elsewhere[29].

qRT–PCR. After appropriate transfections, total RNA was extracted using TRIzol Reagent (Life Technologies) using the manufacturer's protocol. For qRT–PCR the mRNA was reverse transcribed using MMLV-RT (Invitrogen) using the manufacturer's protocol to generate cDNA. The qRT–PCR reactions were performed using Stratagene MxPro3000P (Agilent Technologies) and SYBRGREEN (Fermentas). Common primers were designed to target the open reading frame and normalize for total transcript. The distal primers were designed to target sequences just before the dPAS and detect long transcripts that use the dPAS. All primers used are shown below. Data were calculated using a modified version of the $2^{-\Delta\Delta CT}$ method to show changes in dPAS usage, where CT is the threshold cycle. First, the CT values for the common and distal amplicons were normalized to the levels of 7SK, where ΔCT (common or distal) $= CT_{\text{common or distal}} - CT_{7SK}$. Then $\Delta\Delta CT = \Delta CT_{\text{distal}} - \Delta CT_{\text{common}}$ (note that we applied the correction factor for difference in amplification efficiency calculated in Extended Data Fig. 1). To show fold changes normalized to the control siRNA-transfected samples the following equation was used: normalized ΔΔΔCT $= \Delta\Delta CT_{\text{average target siRNA}} - \Delta\Delta CT_{\text{average of control siRNA}}$. Then the decrease (−) or increase (+) in dPAS usage was calculated as $\pm 2^{\text{normalized }\Delta\Delta\Delta CT}$.

Oligonucleotides used for qRT–PCR. Cyclin D1 common forward, 5′-CTGCCAGGAGCAGATCGAAG; reverse, 5′-AATGCTCCGGAGAGGAGGGACT; distal forward, 5′-ATCGAGAGGCCAAAGGCT; reverse, 5′-CGTCTTTTTGTCTTCTGCTGGA; DICER1 common forward, 5′-CTCATTATGACTTGCTATGTCGCCTTG; reverse, 5′-CACAATCTCACATGGCTGAGAAG; distal forward 5′-TGCTTTCCGCAGTCCTAACTATG; reverse, 5′-AATGCCACAGACAAAAATGACC; TIMP2 common forward, 5′-CAACCCTATCAAGAGGATCCAGTAT; reverse, 5′-GATGTCGAGAAACTCCTGCTTG; distal forward, 5′-GACATCAGCTGTAATCATTCCTGTG; reverse, 5′-CGATGCCAAATGGAGAGC; FHL1 common forward, 5′-CTGGCACAAAGACTGCTTCACCTGT; reverse 5′-GATTGTCCTTCATAGGCCACCACACTGG; distal forward, 5′-GCCAGGGCTGTCATCAACATGGATA; reverse 5′-TGCATTTCAGGTAAGCGGTAGGTGGA; tubulin common forward, 5′- GAAGGCCTCATCCTCCACTTTGGAAAG; reverse, 5′-TGCTAGCAGTGTCTCATGCTCG; distal forward, 5′-GCATCAGTAGCTGAGTGCACTCCTGGT; reverse, 5′-GTAGAGGGTATGAAGGGCAAGAACTCT; VMA21 common forward, 5′- GATAAGGCGGCGCTGAACGCACTGC; reverse, 5′-TGAGCCTTCATTCCAGGCCACATACACA; distal forward, 5′-CATCTGCACAGCACCTTACAGTTTGC; reverse, 5′-GAAATGCAGCACATCCAAATCCTCCC; GSK-3β common forward, 5′-CTGGTCCGAGGAGAACCCAATGTTTCG; reverse, 5′-CAGCCAACACACAGCCAGCAGACCATAC; distal forward, 5′-GAGCTGAGCCCATGGTTGTGTGTAAC; reverse, 5′-GGTTCACTTCAGCAGGCAGGACAACTC; SMOC1 common forward, 5′-CTCTGATGGCAGGTCCTACGAGTCCA; reverse, 5′-GTATGGCACTGCACCTGGGTAAAGGAG; distal forward, 5′-GAGTCCTGCAATTGTACTGCGGACTCCA; reverse 5′-CATGGGATCTGGACTCCCTTCCTCTC; ACER3 common forward, 5′-CACGCTGGACTGGTGCGAGGAGAACT; reverse, 5′-GTGGAAGCACCAGGATCCCATTCCTACC; distal forward, 5′-CTGTTCAAGCTAATACAGCATTTCCT; reverse, 5′-GTGAATAAGCAGACTGAGATTACCTG; TMEM48 common forward, 5′-CATTCATCCTCAGCAACTCATGCACTC; reverse 5′-CTGTTAGTACCAGTGCAGGGAACCAC; distal forward, 5′-GTGCTGTGTACTAAATACAGGCCACATAGTG; reverse 5′-CCTGGTTCCAACAGATGGTGTGTAGA; MSRB3 common forward, 5′-CTCTGGGAAGTGCGCAGTCCGGGT; reverse, 5′-GTCCCTTTCTCCTGAGTGACATGG; distal forward, 5′-GCAGGATATGGAGTGCAATGAACTGAG; reverse, 5′-ACAGTAAGAGCTGGAGTGCAGAGA; 7SK forward, 5′-GACATCTGTCACCCCATTGATC; reverse, 5′-TCTGCAGTCTTGGAAGCTTGAC.

Luciferase assays. One day after a second hit with siRNA (as described earlier), HeLa cells were transfected with 0.25 μg of gene-specific 3′ UTR *Renilla* luciferase plasmids (SMOC1 and GAPDH from Switchgear Genomics) using Lipofectamine

2000 (Invitrogen). *Renilla* luciferase activity was assayed 24 h after plasmid transfection using Stop and Glo reagent (Promega).

Generation of stable cell lines. LN229 cells were transfected with CFIm25-specific shRNA or control shRNA using polybrene in 6-well plates. Two days after lentiviral transfection cells were transfected with a second hit of lentivirus. Selection was done using 1 μg ml^{-1} of puromycin over 2 weeks. U251 cells were transfected with either GFP or CFIm25 expressing pcDNA3 plasmids using Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol. Selection was performed over 1–2 weeks using 2.5 mg ml^{-1} of G418.

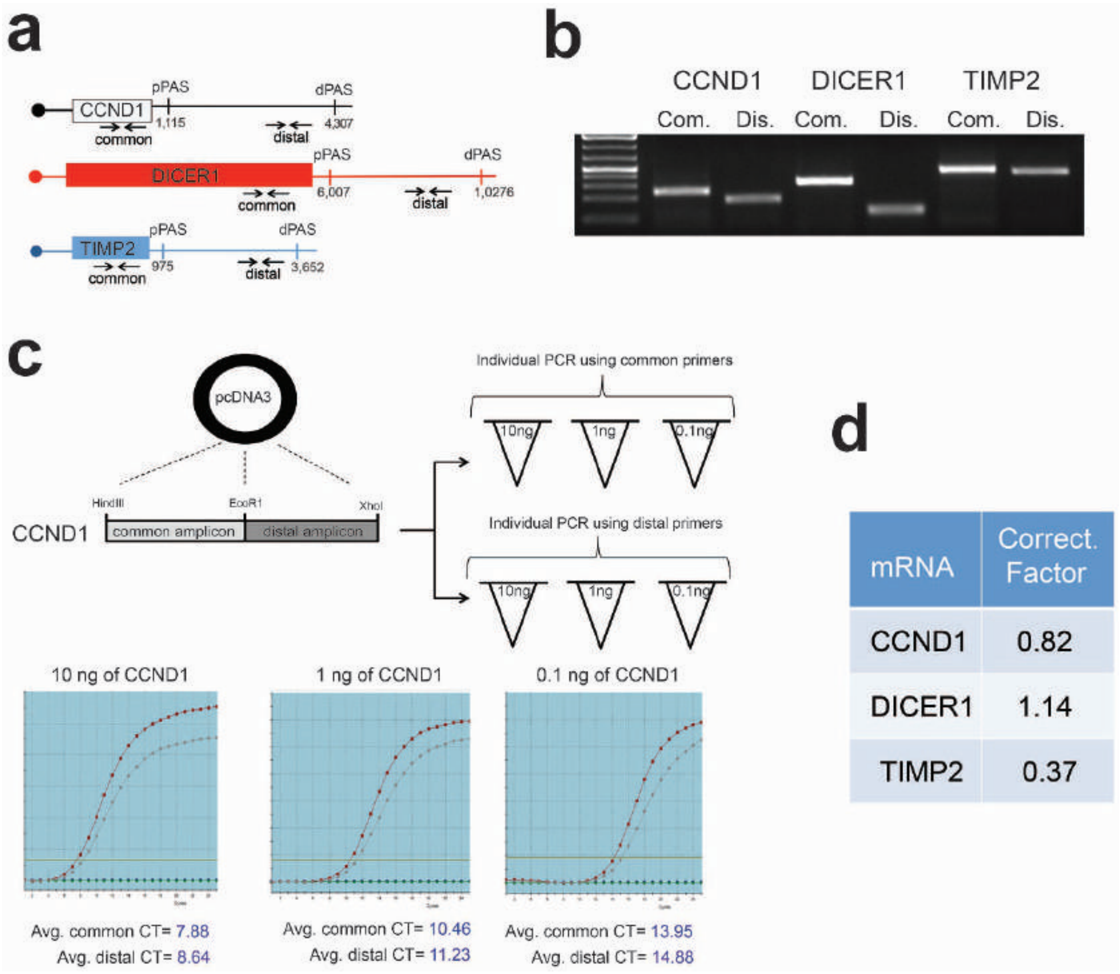
Soft agar assay. Soft agar assays were used to determine anchorage-dependent growth. For the base layer, 1% of UltraPure low melting point agarose (Invitrogen) was mixed 1:1 with 2× DMEM media and plated in 6-well plates giving a 1.5 ml bottom layer of 0.5% agar. Then 3×10^4 cells of LN229 shRNA stably transfected cells were titrated into 2× DMEM and mixed with an equal volume of 0.6% agar to give a 0.3% layer and 1.5 ml was dispensed into each well. The agar was covered with 1 ml of 1× DMEM and incubated in a humidified incubator at 37 °C (5% CO_2). Fresh media was added once a week. After 2 weeks, colonies formed were stained with 0.01% crystal violet, photographed and counted. For U251 plasmid transfected cells the same protocol was followed except that a third (0.3%) layer of agar was plated on top of the layer containing the cell suspension.

Matrigel invasion assay. The Matrigel invasion assay was performed following the manufacturer's protocol. Briefly, the 6-well BioCoat Matrigel Invasion Chamber (Becton Dickinson) was rehydrated with FBS free DMEM. The Matrigel trans-well inserts were then transferred to 6-well plates containing 10% FBS on the bottom. U251 siRNA-transfected or LN229 shRNA-transfected cells were plated (5×10^5 cells per well) in triplicate wells of the upper chamber in serum-free media. After 24 h, cells were stained with 0.01% crystal violet, and the number of invading cells was counted at ×20 magnification in 10 fields for each well.

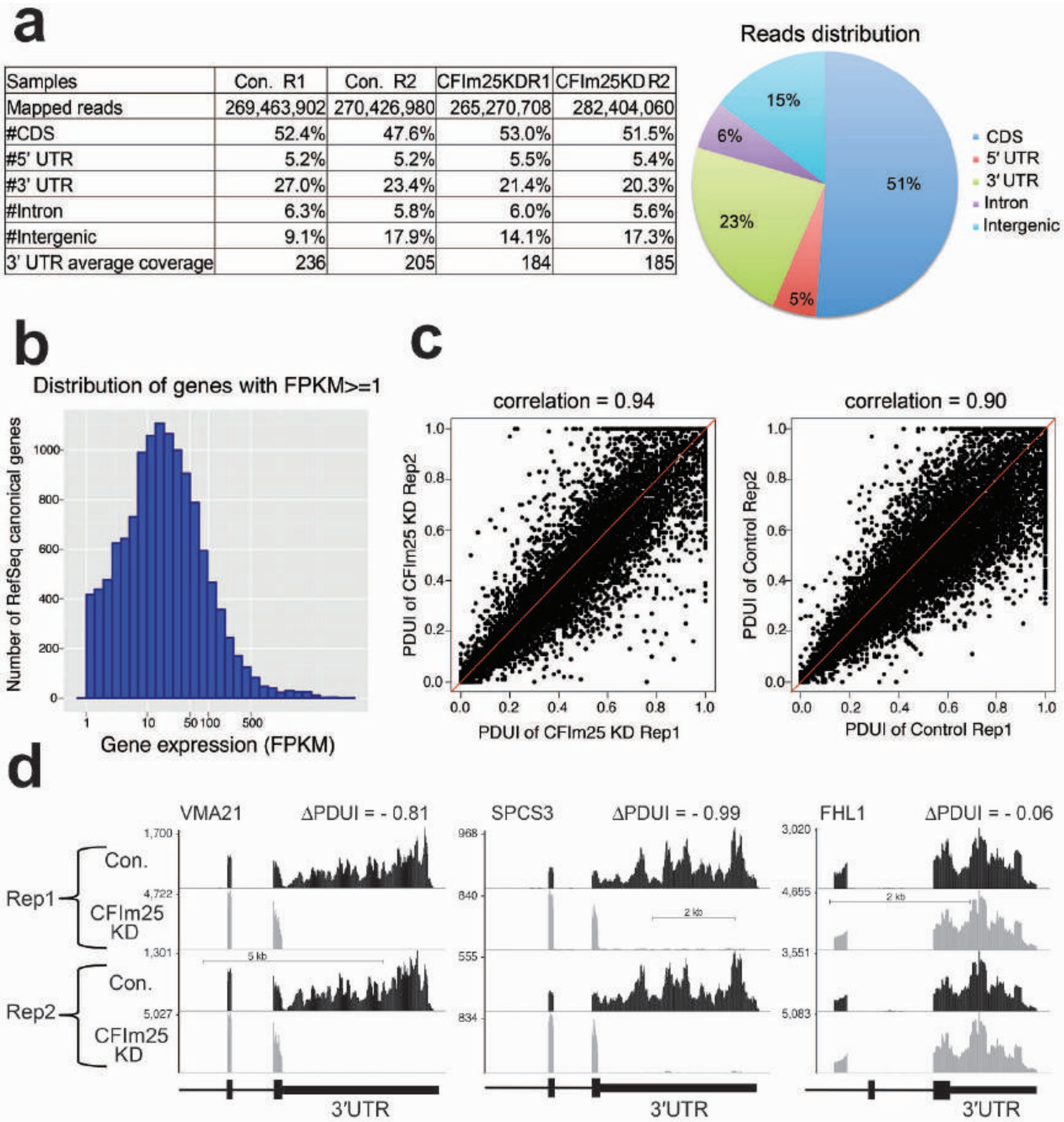
Statistical tests. Unless otherwise specified, experiments were done using three biological replicates and data are shown as average ± s.d., and statistical analysis was done using a two-tailed student *t*-test.

Subcutaneous xenograft tumour model. Hsd:Athymic Nude-Foxn1nu nude mice at age 5–6 weeks were used. For each cell line (LN229 or U251), 20 male nude mice were randomly assigned into two groups ($n = 10$). Stably transfected LN229 and U251 cells were resuspended in pure culture medium with the concentration of 3×10^7 cells ml^{-1}. One-hundred-microlitre cell suspensions (3×10^6 cells) were inoculated subcutaneously into the lower right flank of the mice using a 27-gauge needle. Tumour diameters are measured with digital callipers, and the tumour volume in mm^3 is calculated by the formula: volume = (width)^2 × length/2. The tumour size data were collected and processed blindly. The animal experiments were performed under the Institutional Review Board approved animal protocol AWC-13-115.

31. Ward, A. & Dutton, J. R. Regulation of the Wilms' tumour suppressor (*WT1*) gene by an antisense RNA: a link with genomic imprinting? *J. Pathol.* **185,** 342–344 (1998).
32. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28,** 511–515 (2010).
33. Kaplan, P. J., Mohan, S., Cohen, P., Foster, B. A. & Greenberg, N. M. The insulin-like growth factor axis and prostate cancer: lessons from the transgenic adenocarcinoma of mouse prostate (TRAMP) model. *Cancer Res.* **59,** 2203–2209 (1999).
34. Fabian, M. R., Sonenberg, N. & Filipowicz, W. Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* **79,** 351–379 (2010).
35. Braulke, T., Dittmer, F., Gotz, W. & von Figura, K. Alteration in pancreatic immunoreactivity of insulin-like growth factor (IGF)-binding protein (IGFBP)-6 and in intracellular degradation of IGFBP-3 in fibroblasts of IGF-II receptor/IGF-II-deficient mice. *Horm. Metab. Res.* **31,** 235–241 (1999).
36. Hu, J. F. *et al.* Lack of reciprocal genomic imprinting of sense and antisense RNA of mouse insulin-like growth factor II receptor in the central nervous system. *Biochem. Biophys. Res. Commun.* **257,** 604–608 (1999).
37. Ellis, M. J. *et al.* Insulin-like growth factors in human breast cancer. *Breast Cancer Res. Treat.* **52,** 175–184 (1998).
38. Friedman, R. C., Farh, K. K.-H., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19,** 92–105 (2009).
39. Gómez-Angelats, M., Teeguarden, J. G., Dragan, Y. P. & Pitot, H. C. Mutational analysis of three tumor suppressor genes in two models of rat hepatocarcinogenesis. *Mol. Carcinog.* **25,** 157–163 (1999).

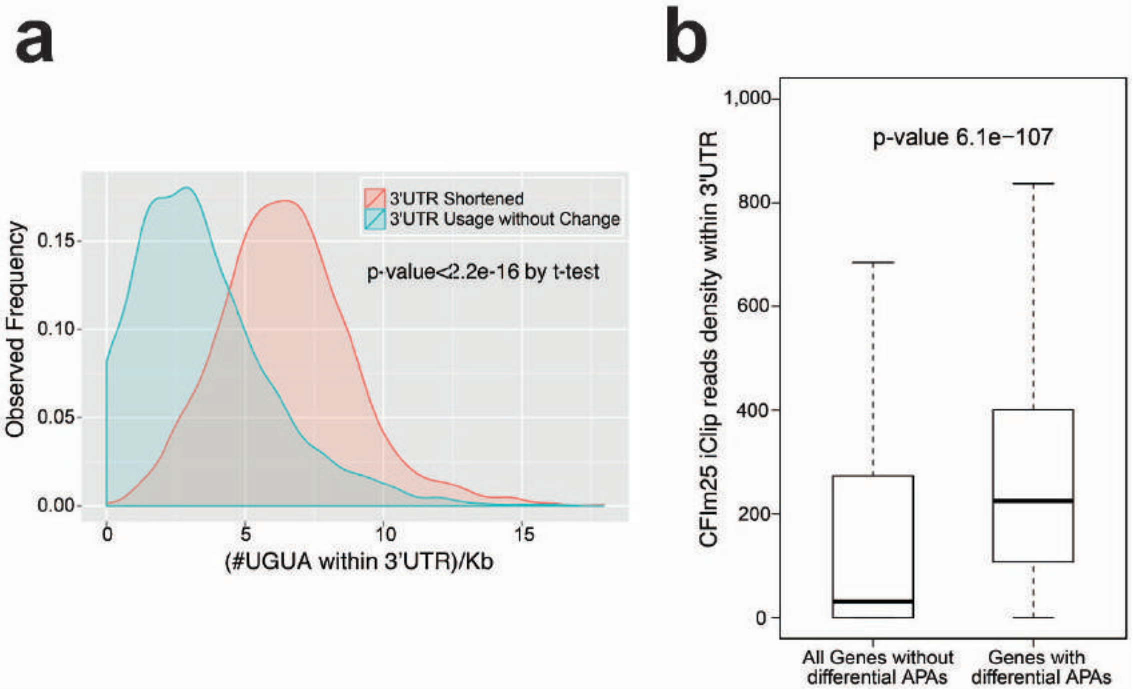


Extended Data Figure 1 | Design and optimization of the qRT–PCR assay to monitor APA of three test genes. a, Schematic denotes the relative location of the common and distal primer annealing sites in each test gene and the approximate locations of the annotated proximal and distal poly(A) sites, depicted as pPAS and dPAS, respectively. The numbers demarcate where the 3′ UTR starts and ends according to ENSEMBL. **b**, Ethidium-stained agarose gel of RT–PCR products of equal cycle number from the different amplicons using HeLa cell mRNA. **c**, Both the common and distal cyclin D1 amplicons were cloned into the same pcDNA3 plasmid in tandem. Three dilutions of each plasmid were made and amplified individually with each amplicon in triplicate. The two lines on the graph depict the amplification curve for the common and distal amplicons. The expectation is that identical cycle threshold (CT) values should be attained for each, given that the PCR reactions were conducted using identical amounts of starting material. The average of three individual experiments is shown for each dilution and the average CT deviation of either amplicon at all of the dilutions was calculated as a correction factor. **d**, The experiment shown in **c** was repeated for DICER1 and TIMP2 to determine their respective correction factors, which was then applied to experiments shown in Fig. 1.

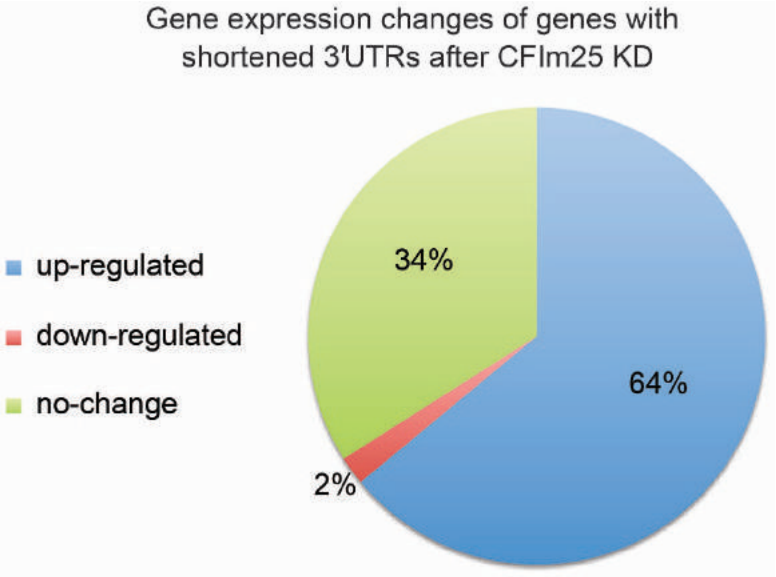


Samples	Con. R1	Con. R2	CFIm25KDR1	CFIm25KD R2
Mapped reads	269,463,902	270,426,980	265,270,708	282,404,060
#CDS	52.4%	47.6%	53.0%	51.5%
#5' UTR	5.2%	5.2%	5.5%	5.4%
#3' UTR	27.0%	23.4%	21.4%	20.3%
#Intron	6.3%	5.8%	6.0%	5.6%
#Intergenic	9.1%	17.9%	14.1%	17.3%
3' UTR average coverage	236	205	184	185

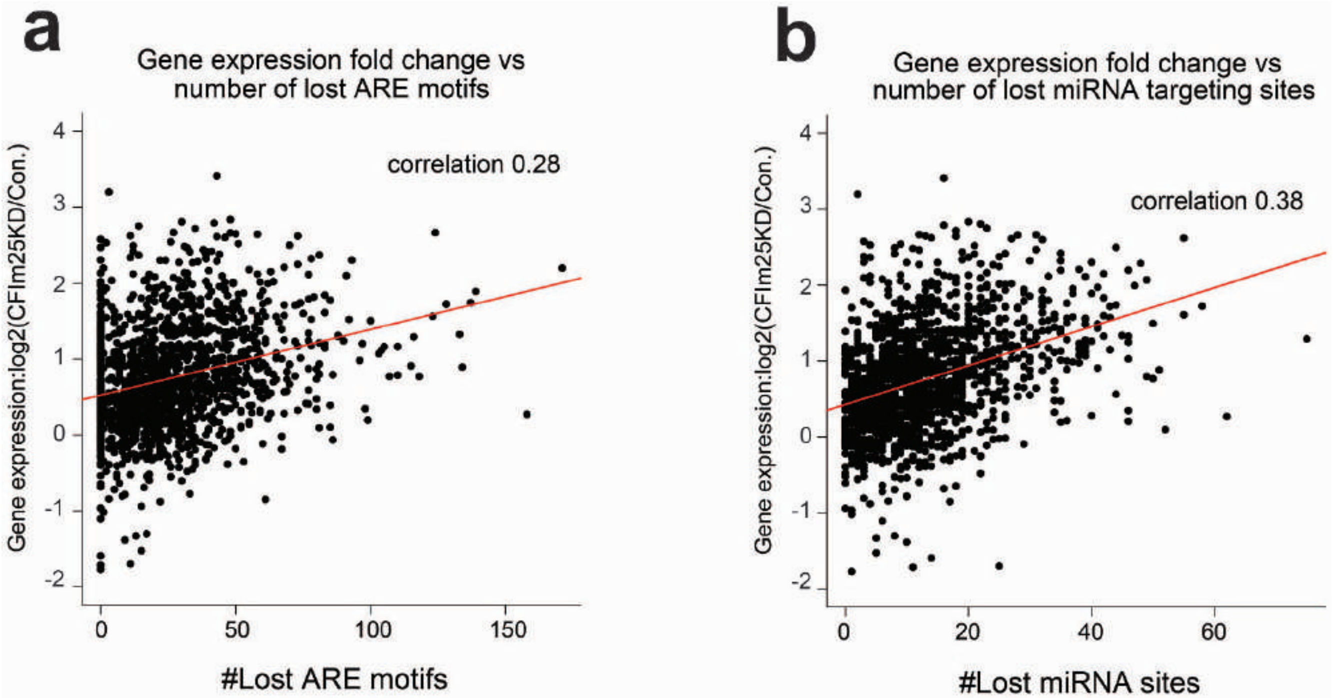
Extended Data Figure 2 | Summary of RNA-seq alignment and reproducibility of PDUI and CFIm25-depletion-induced 3′ UTR shortening. a, RNA-seq read statistics of the four biologically independent experiments where HeLa cells were treated with either control siRNA (Control) or CFIm25 siRNA (CFIm25kD). Pie chart on the right represents genomic distribution of reads that were mapped to human genome hg19. The percentage was calculated by averaging all samples. CDS, coding region. **b**, Histogram of gene expression of RefSeq genes with fragments per kilobase of transcript sequence per million mapped paired-end reads (FPKM) no less than 1. **c**, Scatterplot of the two biological replicates for each condition with high Pearson correlation ($r \geq 0.9$) demonstrating a high level of reproducibility between sample PDUI scores. Each dot represents the PDUI of a transcript. **d**, Genome browser screen images from four independent RNA-seq experiments. Each represents an independent biological sample where HeLa cells were transfected with either the control siRNA (Con.) or an siRNA that knocked down CFIm25. Both VMA21 and SPCS3 were found to undergo 3′ UTR shortening after CFIm25 knockdown whereas FHL1 was found not to change.



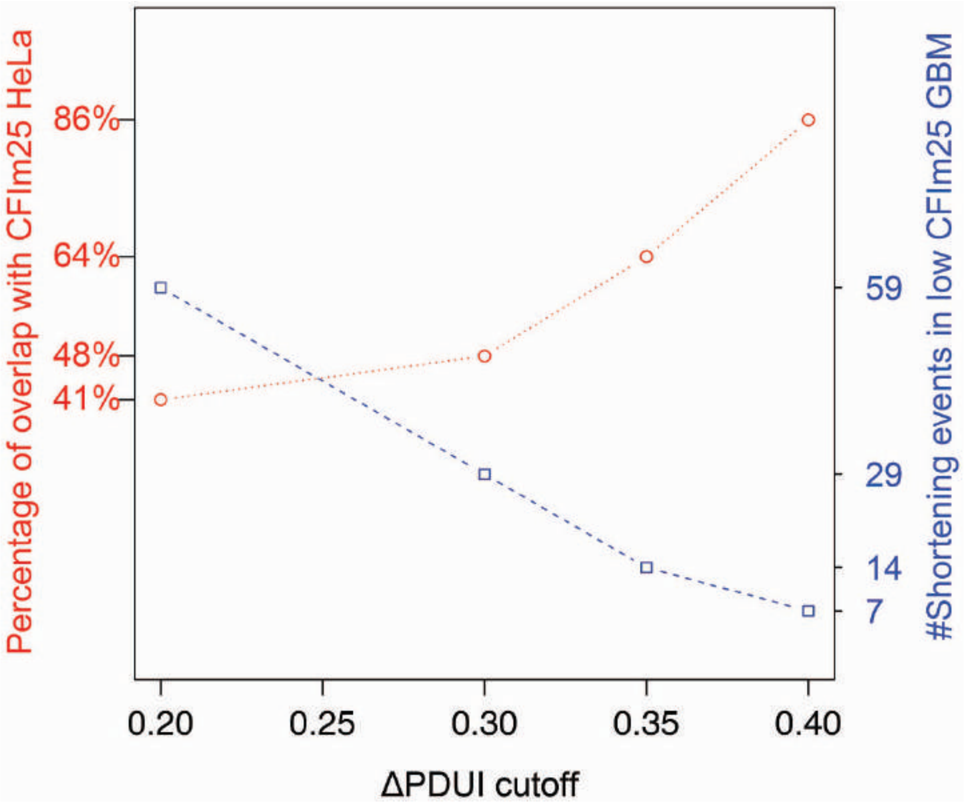
Extended Data Figure 3 | Shortened transcripts have more UGUA CFIm25-binding motifs than unaltered transcripts. a, CFIm25 is known to bind to the UGUA motif. The number of UGUA motifs within the 3′ UTRs of genes with 3′ UTR shortening after CFIm25 knockdown relative to genes with unaltered 3′ UTRs was calculated and compared. Here we selected the genes without 3′ UTR change according to them having a ΔPDUI value ≤ 0.05. **b**, iCLIP tags from ref. 14 (Gene Expression Omnibus accession number GSE37398) were superimposed onto data derived from PDUI analysis of CFIm25 knockdown cells. The box plot demonstrates the enrichment of CFIm25 binding within 3′ UTRs that are altered after CFIm25 knockdown ($P = 6.1 \times 10^{-107}$, *t*-test).



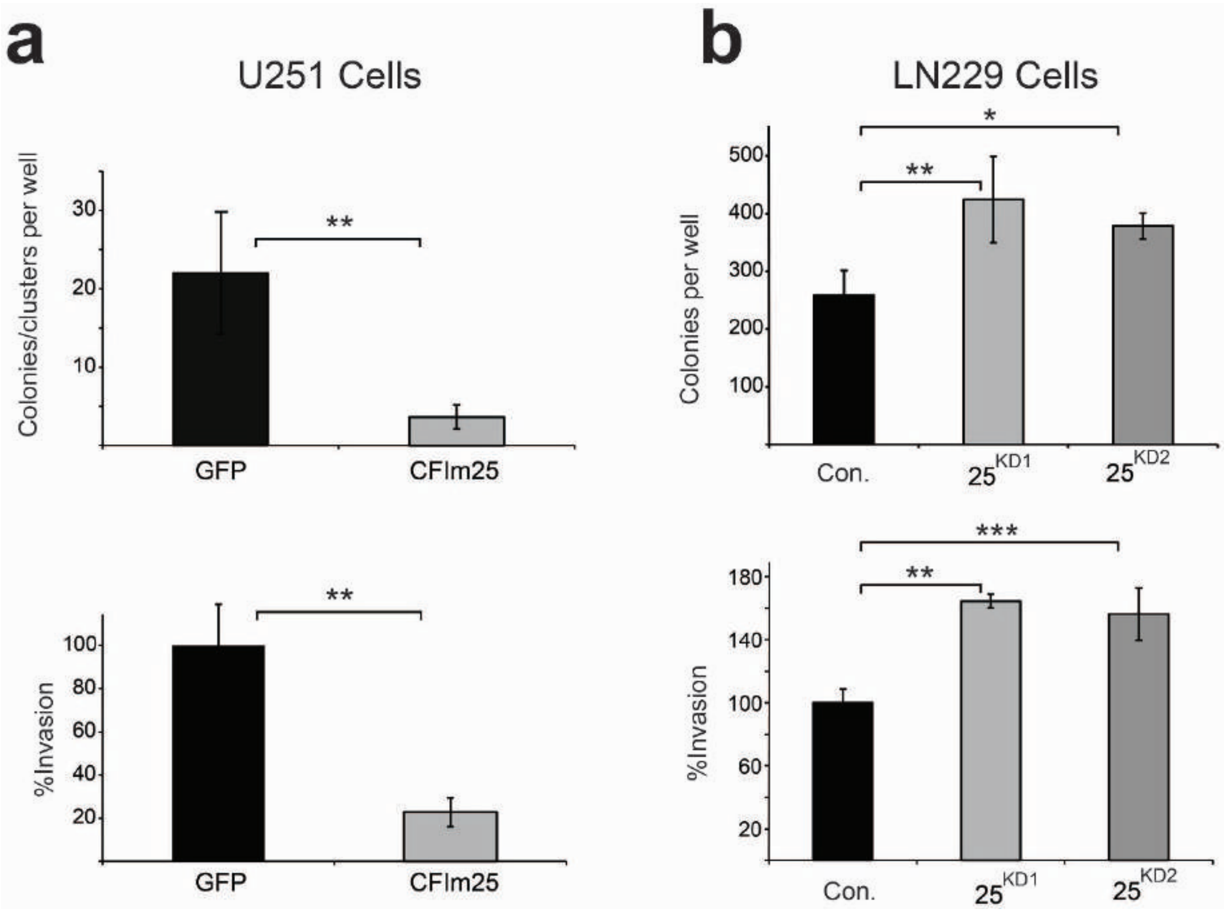
Extended Data Figure 4 | Gene expression changes of genes with shortened 3′ UTRs. Pie chart was calculated from the list of 1,450 genes exhibiting shortened 3′ UTRs due to CFIm25 knockdown (dn, down). Differentially expressed gene analysis was performed using edgeR with FDR ≤ 0.05 (see Methods).



Extended Data Figure 5 | The Pearson correlation between gene expression fold change and the number of lost negative regulatory elements. Left, the number of lost AREs (AU-rich elements) due to 3′ UTR shortening was calculated using the ARE database and plotted against change in gene expression levels after CFIm25 knockdown (KD). Right, similar to the left except the number of lost patented miRNA target sites (Targetscan 6.2) was plotted.



Extended Data Figure 6 | Overlap between shortening events in glioblastoma with low CFIm25 and shortening events in HeLa cells after CFIm25 knockdown. Left, *y*-axis (red) represents the percentage of shortening events in low CFIm25 glioblastoma that are also shortened in HeLa cells after CFIm25 knockdown. Right, *y*-axis (blue) shows the number of shortening events in low CFIm25 glioblastoma (GBM) against different ΔPDUI cut-offs.



Extended Data Figure 7 | Overexpression of CFIm25 reduces invasion and colony formation whereas CFIm25 depletion increases invasion and colony formation. **a**, U251 cells were transfected with either GFP or CFIm25. Top left, Cells were replated in soft agar and the number of colonies/clusters formed were determined. Bottom left, Matrigel invasion assay for cells overexpressing CFIm25 or GFP. **b**, Top right, LN229 cells were transfected with either control or two different lentiviral plasmids targeting CFIm25 (KD1 and KD2). Stably transfected cells were plated on soft agar and the resulting colonies were counted for KD1 and KD2, respectively. Bottom right, LN229 cells were transfected with either control or two different siRNAs (KD1 and KD2) directed against CFIm25 and were replated for a Matrigel invasion assay. All the experiments were done in biological triplicates and shown is the mean ± s.d. All *P* values were from the two-tailed student *t*-test of the control versus sample. $^*P < 0.1$, $^{**}P < 0.01$, $^{***}P < 0.001$.

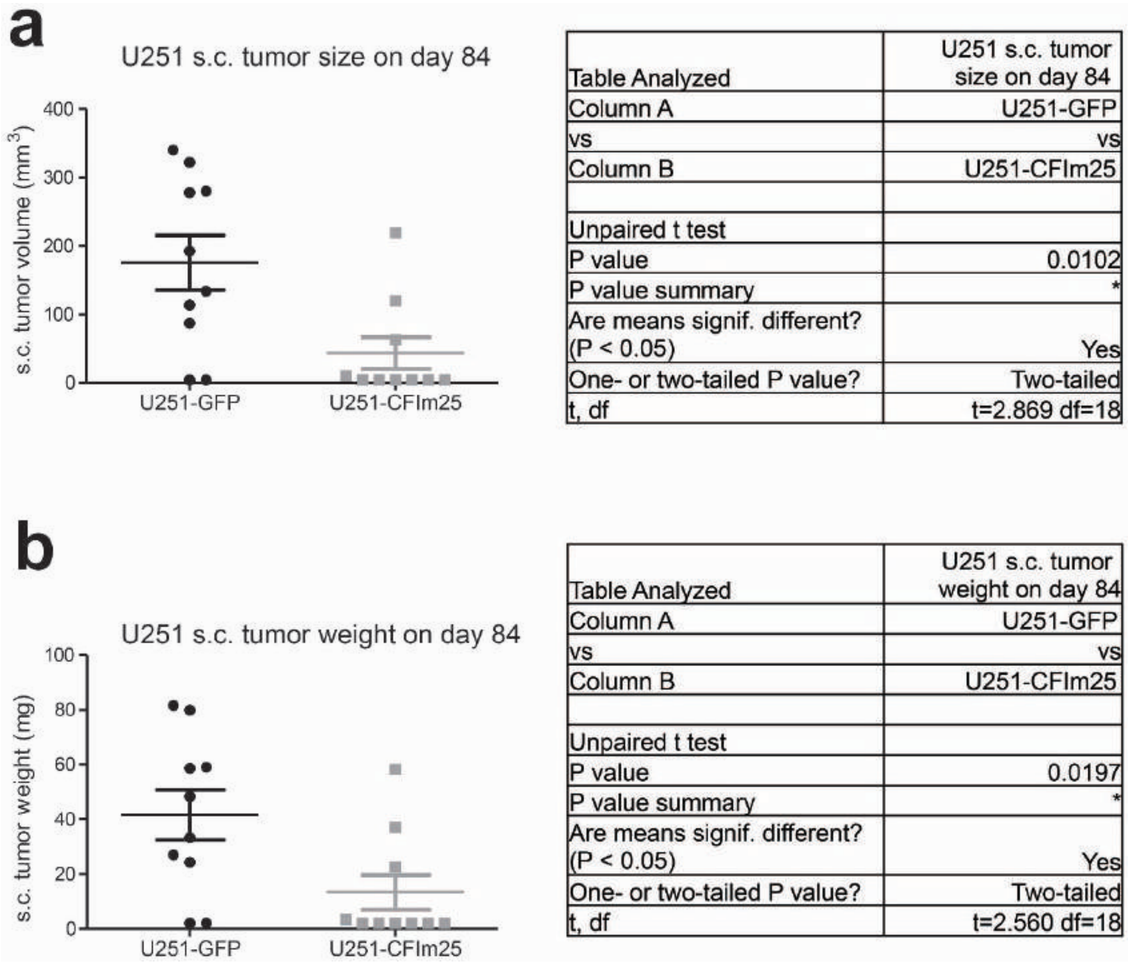
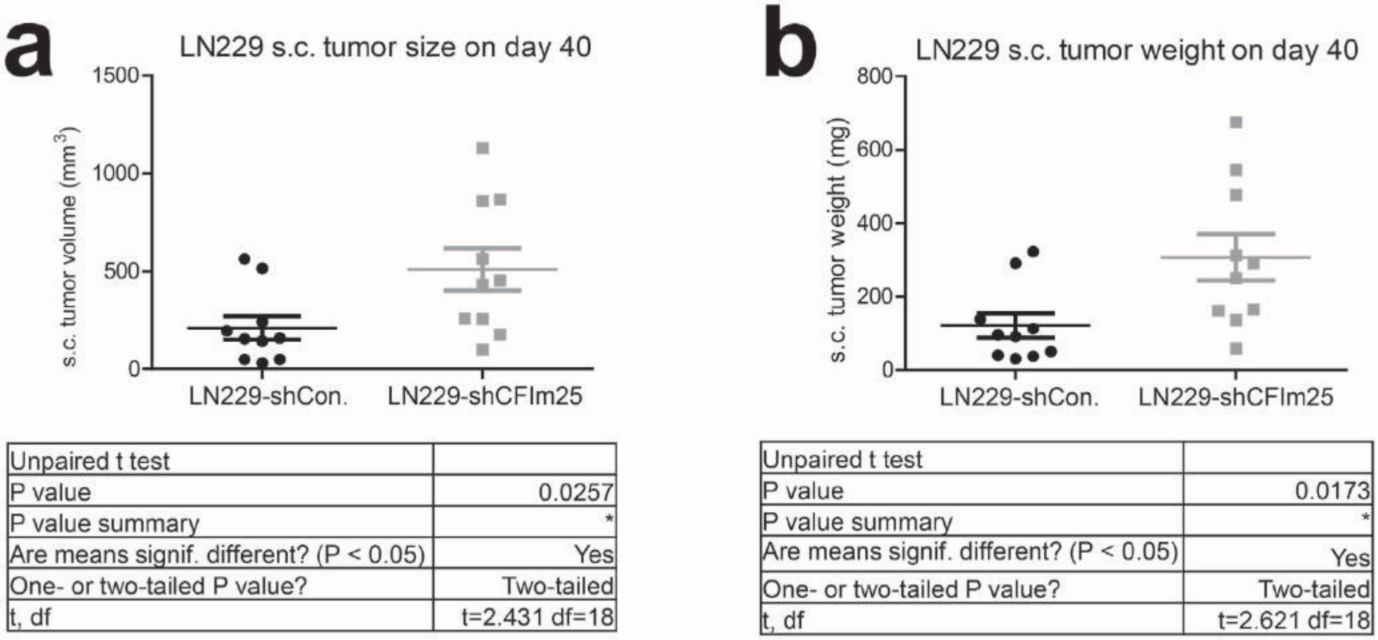


Table Analyzed	U251 s.c. tumor size on day 84
Column A	U251-GFP
vs	vs
Column B	U251-CFIm25
Unpaired t test	
P value	0.0102
P value summary	*
Are means signif. different? (P < 0.05)	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=2.869 df=18

Table Analyzed	U251 s.c. tumor weight on day 84
Column A	U251-GFP
vs	vs
Column B	U251-CFIm25
Unpaired t test	
P value	0.0197
P value summary	*
Are means signif. different? (P < 0.05)	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=2.560 df=18

Extended Data Figure 8 | Overexpression of CFIm25 in U251 tumours reduces their size and weight. **a**, **b**, U251 subcutaneous (s.c.) xenograft tumours were isolated from nude mice on day 84 after implantation and measured for volume (**a**) and weight (**b**) (*n* = 10). U251-GFP indicates control U251 cells expressing GFP and U251-CFIm25 indicates cells transduced with a lentivirus that overexpresses CFIm25.



Unpaired t test	
P value	0.0257
P value summary	*
Are means signif. different? (P < 0.05)	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=2.431 df=18

Unpaired t test	
P value	0.0173
P value summary	*
Are means signif. different? (P < 0.05)	Yes
One- or two-tailed P value?	Two-tailed
t, df	t=2.621 df=18

Extended Data Figure 9 | Reduction in CFIm25 expression levels enhances LN229 tumour size and weight. **a**, **b**, LN229 subcutaneous (s.c.) xenograft tumours were isolated from nude mice on day 40 after implantation and measured for volume (**a**) and weight (**b**) ($n = 10$). LN229-shCon. indicates control lentiviral transduced cells and LN229-shCFIm25 indicates cells transduced with a lentivirus that expresses shRNA targeting CFIm25.

nature COMMUNICATIONS

ARTICLE

Received 11 Jul 2014 | Accepted 15 Sep 2014 | Published 20 Nov 2014

DOI: 10.1038/ncomms6274

Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types

Zheng Xia[1,2], Lawrence A. Donehower[3,4], Thomas A. Cooper[2,5,6], Joel R. Neilson[6], David A. Wheeler[4,7], Eric J. Wagner[8] & Wei Li[1,2]

Alternative polyadenylation (APA) is a pervasive mechanism in the regulation of most human genes, and its implication in diseases including cancer is only beginning to be appreciated. Since conventional APA profiling has not been widely adopted, global cancer APA studies are very limited. Here we develop a novel bioinformatics algorithm (DaPars) for the *de novo* identification of dynamic APAs from standard RNA-seq. When applied to 358 TCGA Pan-Cancer tumour/normal pairs across seven tumour types, DaPars reveals 1,346 genes with recurrent and tumour-specific APAs. Most APA genes (91%) have shorter 3′-untranslated regions (3′ UTRs) in tumours that can avoid microRNA-mediated repression, including glutaminase (*GLS*), a key metabolic enzyme for tumour proliferation. Interestingly, selected APA events add strong prognostic power beyond common clinical and molecular variables, suggesting their potential as novel prognostic biomarkers. Finally, our results implicate *CstF64*, an essential polyadenylation factor, as a master regulator of 3′-UTR shortening across multiple tumour types.

[1] Division of Biostatistics, Dan L Duncan Cancer Center, Baylor College of Medicine, Houston, Texas 77030, USA. [2] Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas 77030, USA. [3] Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, Texas 77030, USA. [4] Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. [5] Department of Pathology and Immunology, Baylor College of Medicine, Houston, Texas 77030, USA. [6] Department of Molecular Physiology and Biophysics, Baylor College of Medicine, Houston, Texas 77030, USA. [7] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA. [8] Department of Biochemistry and Molecular Biology, The University of Texas Medical School at Houston, Houston, Texas 77030, USA. Correspondence and requests for materials should be addressed to W.L. (email: WL1@bcm.edu).

The dynamic usage of messenger RNA (mRNA) 3′-untranslated region (3′ UTR), mediated through alternative polyadenylation (APA), plays an important role in post-transcriptional regulation under diverse physiological and pathological conditions[1,2]. Approximately 70% of human genes[3] are characterized by multiple polyA sites that produce distinct transcript isoforms with variable 3′-UTR length and content, thereby significantly contributing to transcriptome diversity[4]. The majority of APA examples utilize alternative polyA sites located within the terminal exon proximally downstream of the stop codon (tandem APA). As a result, while the protein-coding sequence is unaltered, the regulatory elements in the distal 3′ UTR that might reduce mRNA stability or impair translation efficiency can be removed, including AU-rich elements[5] and microRNA (miRNA)-binding sites[6]. A small percentage of APA sites can be located within internal introns/exons (splicing APA) and are coupled with alternative splicing to produce mRNA isoforms encoding distinct proteins. A well-documented example occurs during B-cell differentiation, where IgM switches from a membrane-bound form to a secreted form using a proximal polyA site instead of a distal one[7]. More recent studies[8] have shed light on the importance of APA in human diseases such as cancer, but its clinical significance to tumorigenesis is only beginning to be appreciated. Both proliferating cells[2,9] and transformed cells[10] have been shown to favour expression of shortened 3′ UTRs through APA, leading to activation of several proto-oncogenes, such as cyclin D1 (ref. 8). Collectively, these observations imply that truncation of the 3′ UTRs may serve as prognostic biomarkers[10,11]. While compelling, these studies were highly limited to either a limited number of genes or a small sample size. It remains to be determined to what extent APA occurs in cancer patients, what level of clinical utility APA may have and the molecular mechanisms and functional consequences of APA during tumorigenesis across multiple tumour types.

RNA-seq has become a routine protocol for gene expression analysis; however, methods to quantify relative APA usage are still under development. Previous global APA studies use microarrays[2,12], which are limited by the dependence on annotated polyA databases as well as inherent technical biases such as cross-hybridization. Recent APA protocols use polyA junction sites enrichment followed by high-throughput sequencing (PolyA-seq)[13,14]. These PolyA-seq protocols, although powerful in providing the precise locations of polyA sites, are hampered by technical issues, such as internal priming artefacts, and thus have not been widely adopted by the cancer community. In contrast, RNA-seq has been widely employed in almost every large-scale genomics project, including The Cancer Genome Atlas (TCGA). However, very few RNA-seq reads contain polyA tails, challenging our ability to identify APA events. For example, an ultra-deep sequencing study[15] only identified ~40 thousand putative polyA reads (~0.003%) from 1.2 billion total RNA-seq reads. Moreover, although the popular RNA-seq tool MISO[16] can detect annotated alternative tandem 3′ UTRs, it cannot identify any novel APA events beyond polyA databases. Finally, the short 3′ UTRs are often embedded within the long ones, and thus the isoforms with short 3′ UTRs are commonly overlooked by transcript assembly tools, such as Cufflinks[17]. Despite these inherent limitations, we hypothesize that any major changes in APA usage between different conditions will result in localized changes in RNA-seq density near the 3′-end of mRNA. And this localized RNA-density change can be readily detected through single-nucleotide resolution RNA-seq analysis. We therefore developed a novel bioinformatics algorithm, Dynamic analyses of Alternative PolyAdenylation from RNA-Seq (DaPars), to directly infer dynamic APA events through the comparison of standard RNA-seq data between different conditions.

TCGA has characterized a comprehensive list of genomic, epigenomic and transcriptomic features in thousands of tumour samples; however, it lacks a PolyA-seq platform for APA analysis. To fill this knowledge gap, we used DaPars to retrospectively analyse the existing RNA-seq data of tumours and matched normal tissues derived from 358 patients across 7 tumour types. We discover 1,346 genes with highly recurrent tumour-specific dynamic APA events, demonstrate the additional prognostic power of APA beyond common clinical and molecular variables and expand our knowledge of the mechanisms and consequences of APA regulation during tumorigenesis.

Results

DaPars identifies dynamic APA events. DaPars performs *de novo* identification and quantification of dynamic APA events between tumour and matched normal tissues, regardless of any prior APA annotation. For a given transcript, DaPars first identifies the *de novo* distal polyA site based on a continuous RNA-seq signal independent of the gene model (Fig. 1a; Supplementary Fig. 1a,b). Assuming there is an alternative *de novo* proximal polyA site, DaPars models the normalized single-nucleotide resolution RNA-seq-read densities of both tumour and normal as a linear combination of both proximal and distal polyA sites. DaPars then uses a linear regression model to identify the location of the *de novo* proximal polyA site as an optimal fitting point (vertical arrow in Fig. 1a) that can best explain the localized read-density change. Furthermore, this regression model is extended towards internal exons, so that splicing-coupled APA events can also be detected. Finally, the degree of difference in APA usage between tumour and normal can be quantified as a change in Percentage of Distal polyA site Usage Index (ΔPDUI), which is capable of identifying lengthening (positive index) or shortening (negative index) of 3′ UTRs. The dynamic APA events with statistically significant ΔPDUI between tumour and normal will be reported. The DaPars algorithm is described in further detail in the Methods. One example of an identified dynamic APA event is given for the *TMEM237* gene (Fig. 1b), where the shorter 3′ UTR predominates in both breast (breast invasive carcinoma (BRCA)) and lung (lung squamous cell carcinoma (LUSC)) tumours compared with matched normal tissues. Another example is *LRRFIP1* (Fig. 1c), where the distal 3′ UTR is nearly absent in both breast and lung tumours.

DaPars evaluation using simulated and experimental APA data. To assess the performance of DaPars, we conducted a series of proof-of-principle experiments. First, we used simulated RNA-seq data with predefined APA events to evaluate DaPars as a function of sequencing coverage. We simulated 1,000 genes in tumour and normal at different levels of sequencing coverage (reads per base gene model). For each gene, we simulated two isoforms with long and short 3′ UTRs (3,000 and 1,500 bp), respectively. The relative proportion of these two isoforms is randomly generated, so that the ΔPDUI between tumour and normal for each gene is a random number ranging from −1 to 1. According to these gene models and expression levels, we used Flux Simulator[18] to generate 50-bp paired-end RNA-seq reads with a 150-bp fragment length, taking into account typical technical biases observed in RNA-seq. The simulated RNA-seq reads were used as the input for DaPars analysis, while the short/long isoforms and the ΔPDUI values were hidden variables to be determined by DaPars. As a criterion for accuracy, the DaPars dynamic APA prediction is considered to be correct if the predicted *de novo* APA is within 50-bp distance of the *bona fide*

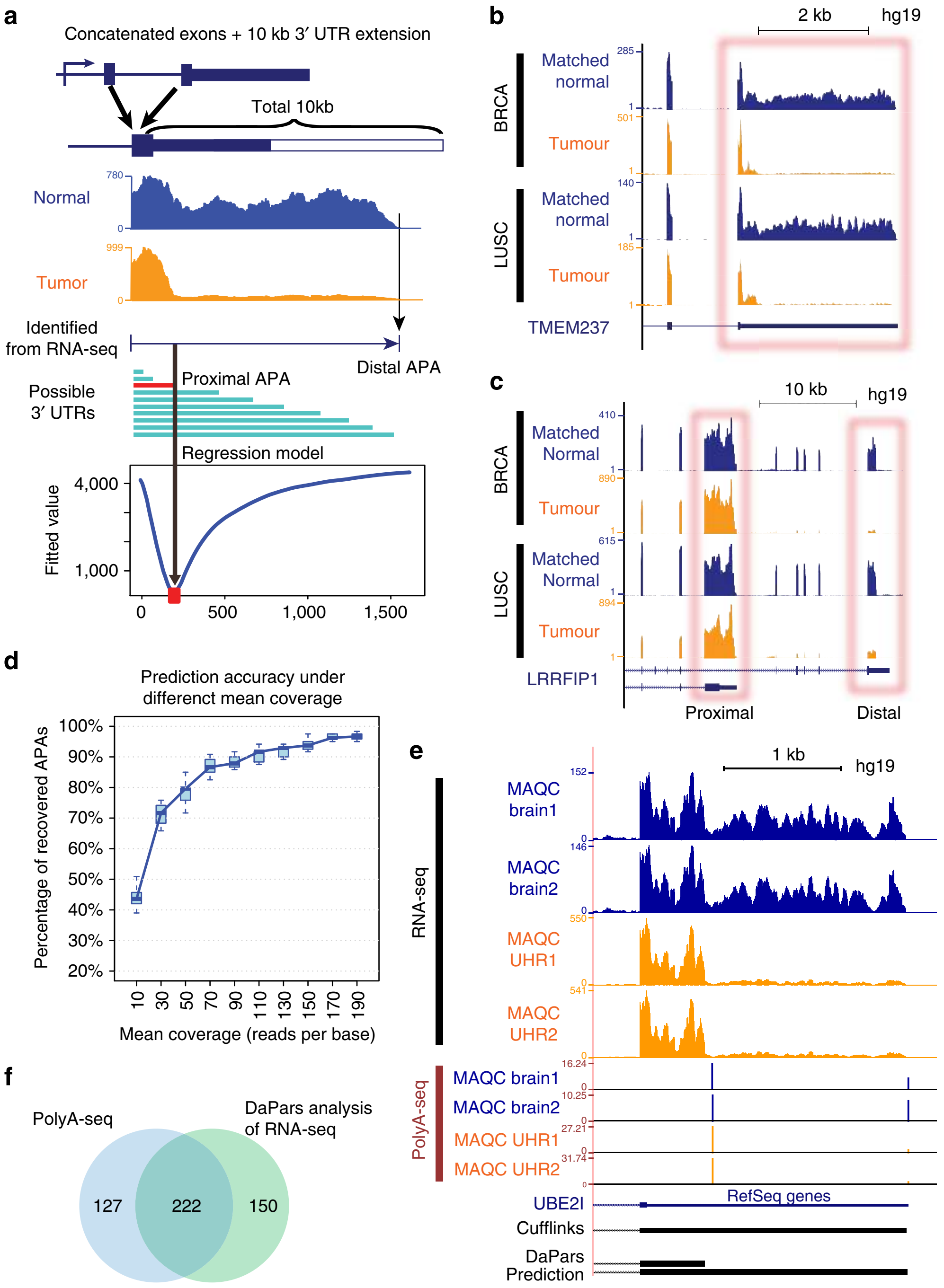


Figure 1 | Overview of the DaPars algorithm and its performance evaluation. (**a**) Diagram depicts the DaPars algorithm for the identification of dynamic APA between tumour and normal samples. The top panel shows RNA-seq coverage on exons with 10 kb extension without any prior knowledge of APA sites. The distal APA site is inferred directly from the combined RNA-seq data of tumour and normal tissues (middle panels). The *y* axis of the bottom panel is the fitted value of our regression model and the locus with the minimum fitted value (red point below vertical arrow) corresponds to the predicted proximal APA site (red horizontal bar). (**b**) An example of DaPars identified dynamic APA from the TCGA RNA-seq data. The shorter 3′ UTR of *TMEM237* is preferred in BRCA and LUSC tumours. (**c**) Another example of dynamic APA, here the distal APA of *LRRFIP1* is nearly absent in both BRCA and LUSC tumours while the proximal APA is unchanged. (**d**) A simulation study to demonstrate DaPars performance. The percentage of recovered APA events is plotted against different sequencing coverage. The quantile box shows the variation of DaPars prediction based on 1,000 simulated events. The black line in each box is the median recovery rate. (**e**) An example of dynamic APA between MAQC UHR and brain detected by both DaPars analysis of RNA-seq and PolyA-seq. The three bottom tracks are the RefSeq gene structure, Cufflinks prediction and DaPars prediction. (**f**) Venn diagram comparison between PolyA-seq and DaPars analysis of RNA-seq based on the same MAQC UHR and brain samples.

polyA site, and the predicted ΔPDUI is within 0.05 from the predetermined ΔPDUI. The final prediction accuracy (percentage of recovered APAs) is plotted as a function of the different coverage levels (Fig. 1d). Using genes with a single isoform as negative controls, we also reported receiver-operating characteristic curves at different coverage levels with areas under receiver-operating characteristic curves (AUCs) ranging from 0.762 to 0.985 (Supplementary Fig. 2). Our results indicate that dynamic APA events can be readily identified across a very broad range of coverage levels. Importantly, we determined that a sequencing coverage of 30-fold can achieve $>70\%$ accuracy and close to 0.9 AUC in dynamic APA detection. Therefore, we filtered out genes with <30-fold coverage for all further analysis.

As an additional proof-of-principle, we directly compared APA events detected by DaPars with that of PolyA-seq. To achieve this, we used the RNA-seq data[19] and PolyA-seq data[3] based on the same Human Brain Reference and the Universal Human Reference (UHR) MAQC samples[20]. For PolyA-seq, the differentially altered 3′-UTR usage was identified as described in Methods. From the comparison between brain and UHR, we found that ~60% (*P* value $<2.2\mathrm{e}-16$; Fisher's exact test) of 372 DaPars predicted APA events could be strongly supported by PolyA-seq (Fig. 1e,f). Both PolyA-seq and DaPars reported longer 3′ UTRs in brain than in UHR in $>94\%$ dynamic APA events, which is consistent with recent reports that brain tissues normally have the longest 3′ UTRs[21,22]. Close inspection of the raw data indicates that the non-overlapping dynamic APA events can be partially explained by the individual assay limitations. For example, PolyA-seq is designed to amplify polyA tags; therefore, some dynamic APA events reported by PolyA-seq may have a small magnitude of changes that are not readily detectable by RNA-seq (Supplementary Fig. 1c). Meanwhile, probably due to additional steps such as fractionation, PolyA-seq may also fail to detect dynamic APAs that are clearly observed by RNA-seq (Supplementary Fig. 1d). Together, we conclude that DaPars can reliably detect dynamic APA events between different conditions using standard RNA-seq.

Broad and recurrent shortening of 3′ UTRs across tumour types. Since TCGA lacks a PolyA-seq platform for APA analysis, we sought to fill this knowledge gap through DaPars retrospective analysis of existing TCGA RNA-seq data, which were originally sequenced for gene expression. We focused our analysis on seven tumour types that have >10 tumour/normal pairs, including bladder urothelial carcinoma (BLCA), head and neck squamous cell carcinoma, LUSC, lung adenocarcinoma (LUAD), BRCA, kidney renal clear cell carcinoma (KIRC) and uterine corpus endometrioid carcinoma (UCEC) (Supplementary Table 1). TCGA RNA-seq data are of high quality with a mean coverage of around 50-fold, which corresponds to 80% accuracy for DaPars APA analysis based on our simulation study (Fig. 1d). For each tumour type, we identified 224–744 genes with statistically significant and recurrent (occurrence rate $>20\%$) dynamic APA events during tumorigenesis, leading to a total of 1,346 non-redundant events across 7 tumour types (Fig. 2a; Supplementary Fig. 3a; Supplementary Data 1). As a negative control, we did not observe any recurrent APA events between different batches of normal tissues of the same tumour type, indicating that the 1,346 DaPars reported tumour-specific APA events are not likely due to technical artefacts, such as sequencing bias or batch effect. Overall, lung (LUSC and LUAD), uterine (UCEC), breast (BRCA) and bladder (BLCA) cancers possess the highest amount dynamic APA events than the other tumour types (Fig. 2a; Supplementary Fig. 3a,b). Furthermore, 55% of the 1,346 dynamic APA events occur in at least 2 tumour types (Supplementary Fig. 3c), indicating potential concerted mechanisms in APA regulation across tumour types. Strikingly, the majority (61–98%) of APA events have shorter 3′ UTRs in tumours (Fig. 2a; Supplementary Fig. 3a), which is consistent with previous reports that transformed cells preferentially express mRNAs with shortened 3′ UTRs[8].

Multiple lines of evidence indicate that DaPars reported *de novo* APA events are indeed regulated through APA. First, 51% of DaPars predictions are within 50 bp of the annotated APAs compiled from Refseq, ENSEMBL, UCSC gene models and polyA database[23]. There is an approximately sixfold enrichment of annotated APAs in our DaPars predictions compared with random controls (Fig. 2b). Second, in the upstream (-50 nt) of our *de nov* APA sites, canonical polyA signal AATAAA can be successfully identified by MEME motif enrichment analysis[24] (Fig. 2c). In addition, AATAAA and ATTAAA are the most prevalent motifs among variants[25] of polyA signals (Supplementary Fig. 4)[4]. By comparing ± 50 bp flanking sequences of the distal and proximal polyA sites of the 3′-UTR shortening events, DREME[26] discriminative motif discovery algorithm reported that AATAAA motif is significantly stronger in distal polyA sites (Supplementary Fig. 5), suggesting the molecular basis for differential polyA site selection[27]. Furthermore, the canonical polyA signal can also be identified (Supplementary Figs 6 and 7) on those *de novo* APA sites that do not coincide with previous annotation. As expected, the *de novo* DaPars analysis enables us to detect novel APAs that are not annotated in any database. For example, we found a potential novel proximal APA site in *AGPS* that is significantly upregulated in LUSC tumour (Fig. 2d). Together, we conclude that DaPars reliably identified a comprehensive list of novel and existing APA target genes across seven TCGA tumour types, and the preferential shortening of 3′ UTR is a major layer of transcriptomic dynamics during tumorigenesis.

APA events remain far from complete. To explore to what extent the discovered 1,346 APA events have reached saturation, we performed 'down-sampling' saturation analysis. We repeated DaPars analysis (occurrence rate $>20\%$) on random subsets of samples of various smaller sizes. Saturation is expected to occur when increasing sample size fails to discover additional APA events. The results indicate that the number of APA events increases steadily with increasing sample size in total (Fig. 2e), sample size per tumour type (Supplementary Fig. 3d) and the number of tumour types studied (Fig. 2f). This suggests that APA events derived from 358 samples across 7 tumour types remain far from complete. DaPars analysis on a larger sample size or more tumour types is likely to reveal many more novel APA events. This prediction is consistent with a recent report demonstrating that cancer genome sequencing normally requires thousands of samples per tumour type to approach saturation[28]. This observation also highlights the need for *de novo* discovery of APA, since any *prior* annotation-based detection methods are likely to miss a significant portion of novel APA events from tumour samples.

Genes with shorter 3′ UTRs are prone to be upregulated. The current model predicts that 3′-UTR shortening through APA during tumorigenesis may upregulate its parental gene by escaping miRNA repression. To test this hypothesis, we calculated the numbers of miRNA-binding sites lost due to 3′-UTR shortening in tumours (Fig. 3a). Using this approach, we determined that ~67% genes with shorter 3′ UTRs in tumours have lost at least 1 predicted miRNA-binding site (Fig. 3a). Furthermore, when compared with all the genes of sufficient sequencing

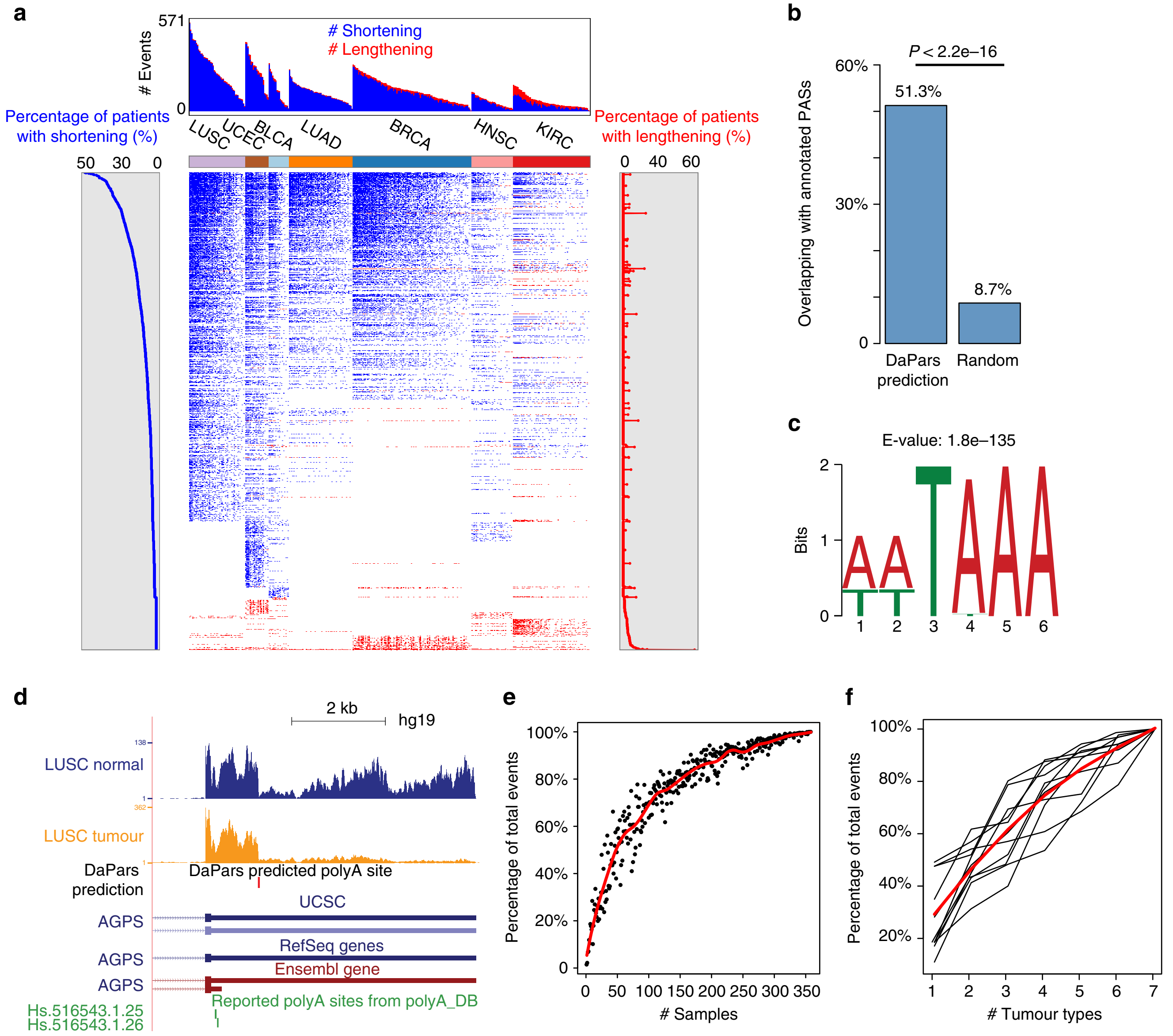


Figure 2 | Broad shortening of 3′ UTRs across seven TCGA tumour types. (**a**) The central heatmap shows genes (rows) undergoing 3′-UTR shortening (blue) or lengthening (red) in each of the 358 tumours (columns) compared with matched normal tissues across seven tumour types. The upper histogram shows the number of APA events per tumour. The side histograms show the percentage of tumours with 3′-UTR shortening (left) or lengthening (right) for each APA gene. (**b**) Bar plots show the percentages of DaPars-predicted APAs and randomly selected APAs from 3′-UTR regions overlapping with annotated APAs from four databases (Refseq, UCSC, ENSEMBL and PolyA_DB). The *P* value was calculated by *t*-test using 50 × bootstrapping of data. (**c**) MEME identifies the canonical polyA motif AATAAA with a very significant E-value (1.8e – 135) from the upstream (– 50 bp) of the proximal polyA sites predicted by DaPars. (**d**) An example of DaPars-predicted novel polyA site (red bar) in a LUSC tumour that is far away from any annotated polyA sites. (**e**) Saturation analysis of APA events by adding more samples. Each point is a random subset of samples of various smaller sizes. All the points were fitted by a smoothed read line. (**f**) Saturation analysis by adding more tumour types. Each grey line represents a random ordering of seven tumour types and the red curve is the fitting line. The percentage of dynamic APA events increased with the number of tumour types.

coverage, those genes with shorter 3′ UTRs in tumours have overall greater miRNA-binding site density in their gene models (*P* value = 1.8e – 11, *t*-test; Fig. 3b). These data imply that APA regulation tends to maximize the miRNA-binding loss through preferentially shortening those 3′ UTRs already heavily regulated by miRNA. To examine the consequences of 3′ UTR and miRNA-binding loss, we compared the gene expression between tumours and matched normal tissues. As expected, those genes with shorter 3′ UTRs in tumours tend to be more upregulated in tumours (Fig. 3c). In conclusion, our data strongly support the hypothesis that many genes are upregulated during tumorigenesis by shortening their 3′ UTRs to escape post-transcriptional miRNA repression.

APA events add prognostic power beyond common covariates. Very little is known of the clinical implications of the dynamic 3′ UTRs in cancer patients. To address this issue, we used a standard Cox proportional hazards model[29] for the correlation between patient overall survival and multiple clinical and molecular covariates. Here we only used BRCA, LUSC and KIRC due to high mortality rate and large sample size. We first used common

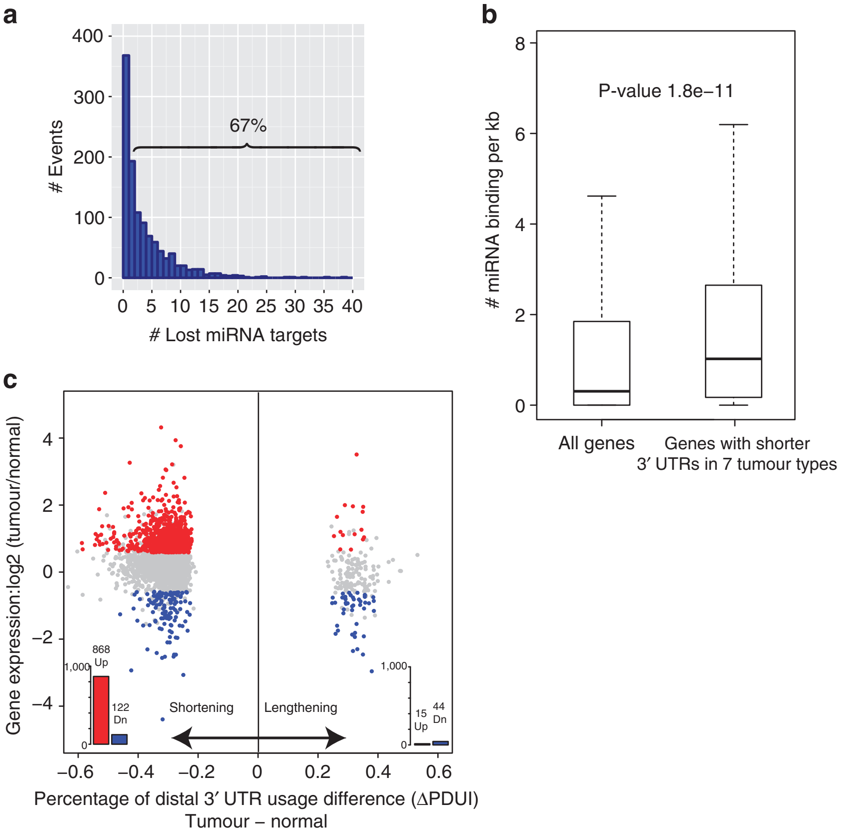


Figure 3 | Genes with shorter 3′ UTRs in tumours are prone to be upregulated. (**a**) Number of genes losing miRNA-binding sites due to the shortening of their 3′ UTRs. Here we selected miRNA-bindings sites predicted by both TargetScanHuman V6.2 (refs 52,53) and miRanda[54], as a more conservative list of miRNA targets. Number in the bracket represents the percentage of genes losing at least 1 miRNA-binding site. (**b**) Genes with shorter 3′ UTRs in tumours have greater miRNA-binding-site density in the 3′-UTR region than all RefSeq genes. We used RefSeq gene models for all the calculations regardless of the APA detection. The *y* axis is the number of miRNA-binding sites normalized by 3′-UTR length (per kb). The *P* value was calculated by a *t*-test. (**c**) For genes with shorter 3′ UTRs in tumours, their fold-change expression between tumours and normal tissues are plotted against their ΔPDUI values. All isoforms of the same gene were combined for the expression measurement. The genes significantly up- or downregulated in tumours are shown in red and blue, respectively, which were identified by paired *t*-test with Benjamini–Hochberg (BH) false-discovery rate at 5%. Accordingly, the red and blue bar plots indicate the number of up- and downregulated genes, respectively.

clinical covariates including only tumour stage, age, gender (excluding breast cancer) and smoking status (lung cancer only) to generate low- and high-risk groups, which are visualized by Kaplan–Meier plots and compared by the log-rank test (Fig. 4a). We next used the same Cox regression model integrated with LASSO to select the APA (ΔPDUI) events besides clinical covariates that can best separate risk groups. With clinical covariates always included, we used leave-one-out cross-validation (CV) to select the optimal 1–3 APA events (Supplementary Table 2) to constitute new APA-clinical Cox regression models (Fig. 4d), which have much more significant *P* values in the risk group comparison. To quantify the added prognostic power of APA events, we used a likelihood-ratio test (LRT) to compare the new APA-clinical models with the clinical only models. The LRT results (Fig. 4e) clearly demonstrate a strong additional prognostic power of APA events beyond clinical covariates. Among these six APA covariates, significant worse survival is associated with 3′-UTR shortening of three genes (*SYNCRIP* in BRCA; *TMCO7* and *PLXDC2* in KIRC) and 3′-UTR lengthening of two genes (*ATP5S* in BRCA; *RAB23* in LUSC) (Supplementary Table 2). This result strongly suggests that, depending on the tumour types or genes studied, either lengthening or shortening of 3′ UTRs may be associated with poor clinical outcome. Since our CV procedure only selects the optimal APA events, it is highly likely that even more APA events can be associated with patient survival. Furthermore, we combined clinical covariates with tumour mRNA expression (mRNA-clinical) and tumour-vs-normal gene expression fold-change (mRNA-FC-clinical model) of the same APA genes (Supplementary Table 2) as two additional Cox regression models and repeated the same analyses. Compared with the APA-clinical model, both mRNA-clinical and mRNA-FC-clinical models provide much less additional prognostic power (Fig. 4e), less-significant log-rank *P* values in risk group comparison (Fig. 4b–d). Finally, we show that the separated high- and low-risk groups by APA-clinical models have no correlation with the TCGA Pancan12 significantly mutated gene (doi:10.7303/syn1750331) (Fig. 4f). Together, APA events provide additional power in survival prediction beyond clinical covariates, and independent of commonly used molecular data such as gene expression and somatic mutations.

Cancer metabolism gene *GLS* is regulated through APA. Ingenuity IPA and literature searches were used to characterize the pathways enriched in 1,346 dynamic APA events (Fig. 5a; Supplementary Data 2). The vast majority of enriched pathways are cancer related, such as ERK/MAP signalling and glutamine metabolism. The metabolism gene glutaminase (*GLS*) is of particular interest. It is well known that tumours are considerably

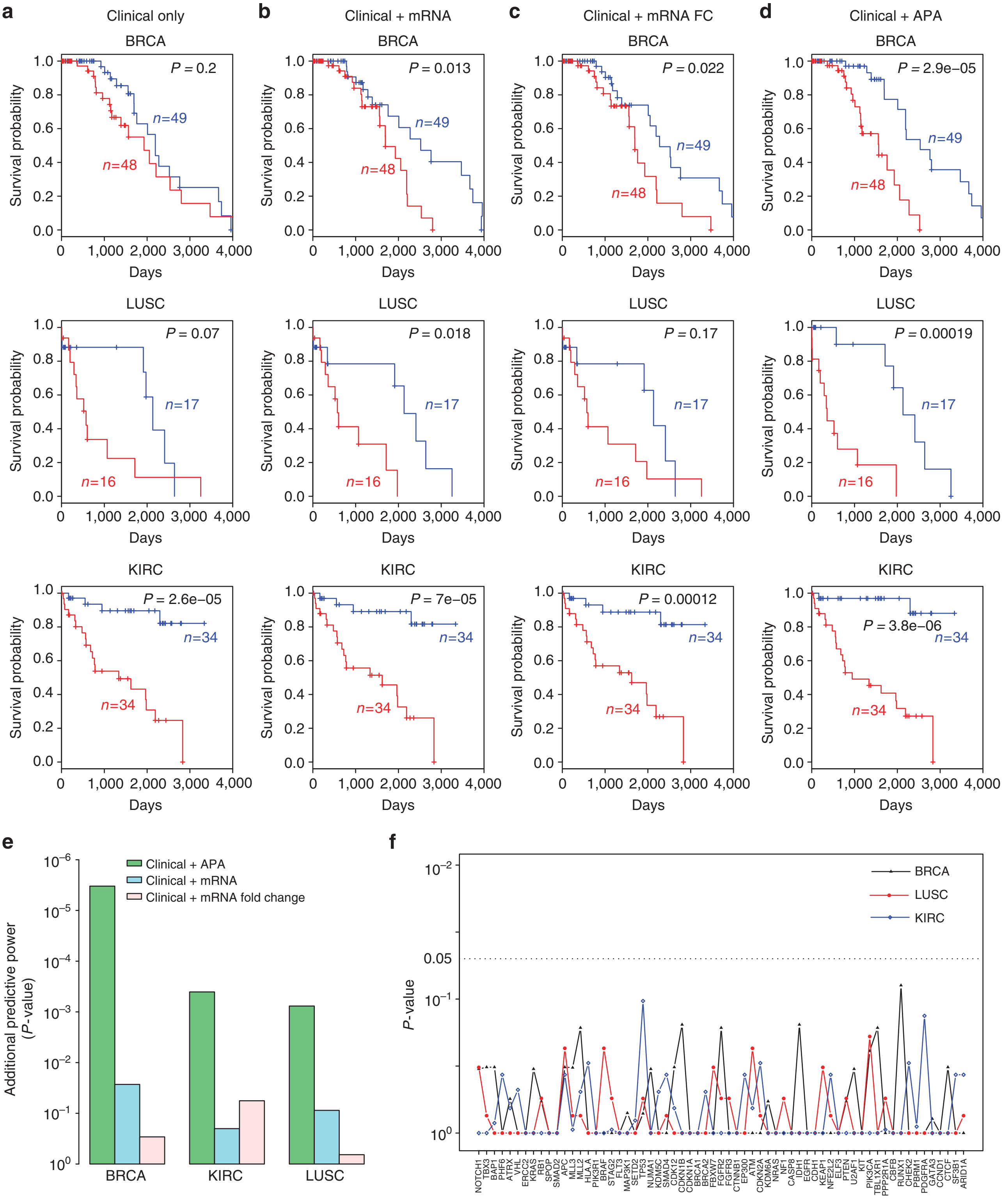


Figure 4 | Prognostic power of dynamic APA events. (**a**–**d**) Kaplan–Meier survival plots for high- (red line) and low (blue line)-risk groups separated by clinical only (**a**), clinical with mRNA expression (**b**), clinical with tumour-vs-normal mRNA expressions fold-change (**c**) and clinical with dynamic APAs (**d**). *P* value was calculated using the log-rank test. (**e**) Additional prognostic power of APA, mRNA expression and mRNA tumour-vs-normal expression fold-change beyond clinical variables. The *P* value is calculated by the likelihood-ratio test. (**f**) No correlation between risk groups separated by APA-clinical models and mutation profiles of significantly mutated genes (SMGs). The dotted vertical line represents the *P* value (Mann–Whitney test) cutoff of 0.05. All SMG *P* values are below this cutoff and thus are not significant.

more dependent on the glycolytic pathway, regardless of oxygen availability, to supply a great deal of their energetic and biosynthetic demand for cell division. This phenomenon, termed the Warburg effect, is a hallmark of cancer[30]. *GLS* is a key enzyme in glutaminolysis and its high expression is essential to support the cancer metabolic phenotype[31]. There are two major *GLS* isoforms

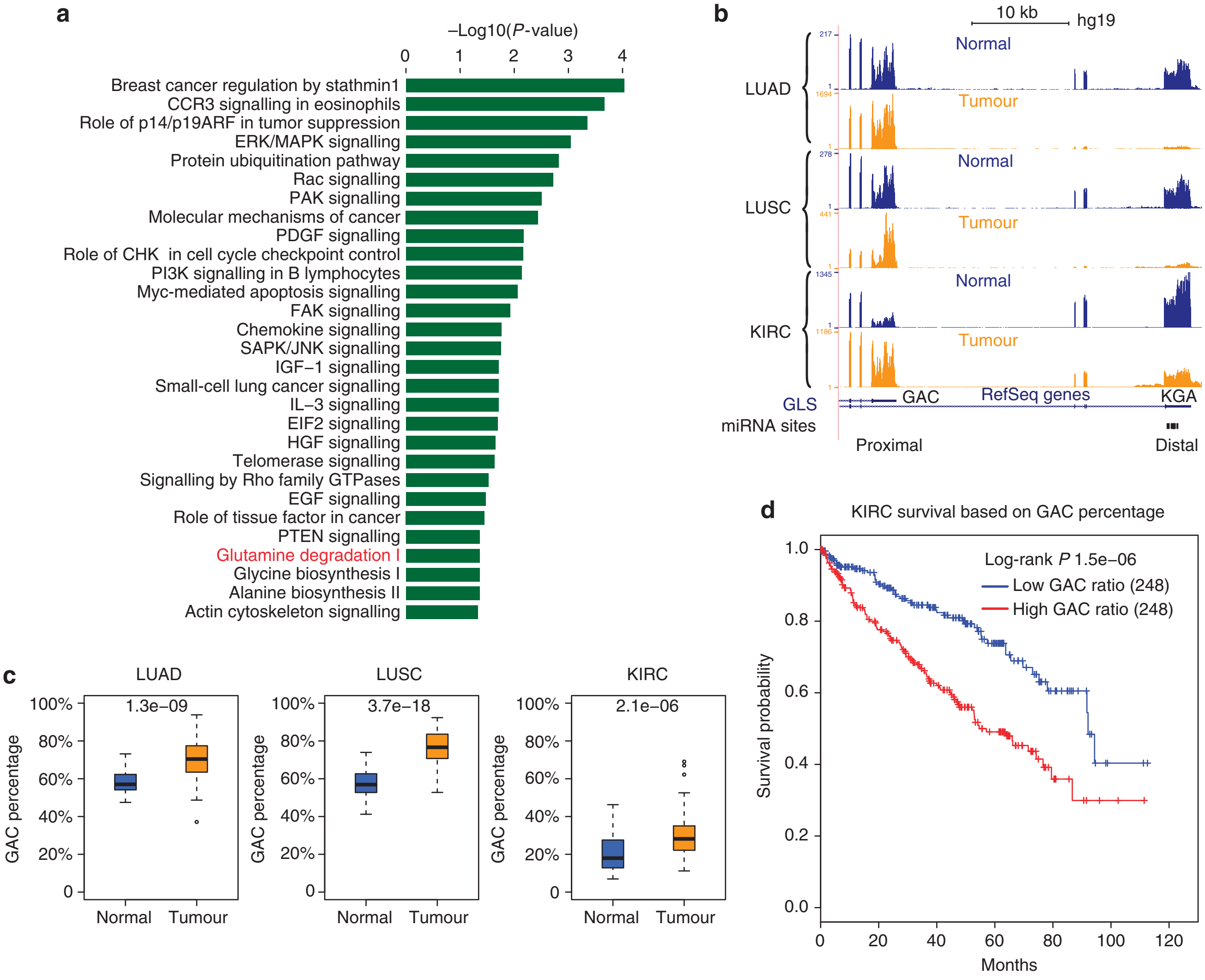


Figure 5 | Pathway analysis. (**a**) Significantly enriched (*P* value <0.05; Fisher's exact test) Ingenuity canonical pathways in the 1,346 dynamic APA events. (**b**) *GLS* has a significant 3′-UTR shift from *KGA* long isoform in normal to *GAC* short isoform in tumour. (**c**) *GAC* percentages are significantly higher in LUAD, LUSC and KIRC tumours. The *P* value in each box was calculated using a paired *t*-test. (**d**) Kaplan–Meier survival plot of two KIRC tumour groups stratified by the GAC ratios with equal patient number in each group. *P* value was calculated using the log-rank test.

termed distal Kidney-type (*KGA*) and proximal Glutaminase C (*GAC*), which have distinct 3′ UTRs and slightly different C termini[32–34] (Fig. 5b). *KGA* has a number of miRNA-binding sites within its 3′ UTR, whereas *GAC* surprisingly is not predicted to have any (Fig. 5b). Furthermore, it has been shown that *miR-23* represses *KGA* in most cells. However, in myc-transformed cells, *MYC* overexpression de-represses *GLS* through down-regulation of *miR-23*, resulting in glutamine-dependent growth characteristics[35]. Interestingly, we found a strong alternative-splicing-coupled 3′-UTR shift from *KGA* in normal to *GAC* in tumour, leading to a significantly increased percentage of *GAC* in LUAD, LUSC and KIRC (Fig. 5b,c). This is consistent with previous report that *GAC* is key to the mitochondrial glutaminase metabolism of cancer cells[31]. The implication of the 3′ UTR switch to *GAC* is that the expression of *GLS* is no longer regulated by *miR-23* or *MYC*. Consistently, we did not observe any significant expression changes of *miR-23* between tumours and normal tissues, although *MYC* is upregulated in LUSC and KIRC tumours (Supplementary Fig. 8a), suggesting that *GLS* potentially utilizes 3′-UTR switch, rather than *MYC* to escape *miR-23*-mediated repression.

To investigate the potential clinical utility of the APA-mediated *GLS* isoform switch, we examined the correlation between *GAC* percentage and clinical survival information for KIRC tumours, using the Cox proportional hazards model with age and gender as covariates. We found that higher *GAC* percentage is highly correlated with worse survival ($P=3.2\mathrm{e}-13$, hazard ratio $=50$, 95% confidence interval: 17–141; Fig. 5d), which is consistent with previous studies indicating that *GAC* is essential for cancer cell growth[36]. Overall, patients with high *GAC* ratios in KIRC have a median survival of ~55 months, compared with >92 months for those with low GAC ratios. We did not find a statistically significant correlation between *GAC* percentage and survival outcome in LUSC and LUAD possibly because the *GAC* percentages ((0.5, 0.97) and (0.59,0.98), respectively) (Supplementary Fig. 8b) have very limited dynamic range in these two tumour types, and thus may not have enough power to stratify patients. In contrast, *GAC* percentage ranges from 0.05 to 0.96 in KIRC (Supplementary Fig. 8b), allowing patient stratification based on a full range of *GAC* levels. Together, the *GLS* APA regulation suggests a novel and potentially *MYC*-independent and miRNA-independent mechanism of glutaminase regulation in tumours, and *GLS* APA can be used to predict patient survival in KIRC.

Potential mechanisms for APA regulation during tumorigenesis. We sought to investigate the potential mechanisms governing APA dynamics in tumorigenesis. Although many details remain poorly understood, APA is thought to be regulated in *cis* through genetic aberrations[37,38] of the underlying nascent mRNA (derived from DNA), and in *trans* by regulatory proteins in responding to dynamic environmental changes[39]. These *cis*-elements include canonical polyA signal AAUAAA and other auxiliary sequences such as U/GU-rich downstream elements[40]. The core polyadenylation *trans*-factors involve four multi-subunit protein complexes, CPSF (cleavage and polyadenylation specificity factor), CstF (cleavage stimulation factor), CFI and CFII (cleavage factors I and II). The chemical cleavage of pre-mRNA process mainly employs CPSF to recognize the canonical polyA signal upstream of the cleavage site, and utilizes CstF to bind downstream U/GU-rich elements[40] mainly through the *CstF64* subunit[39].

To examine the role of genetic aberrations in the regulation of APA, we compared our 1,346 recurrent APA events with 64 Pancan12 Significantly Mutated Genes (doi:10.7303/syn1750331). Surprisingly, there are only five genes in common (Fig. 6a; *P* value 0.48 by Fisher's exact test). This result indicates that most of the dynamic APA events are probably not due to aberrations of underlying *cis*-elements but may be the result of aberrant expression of polyA *trans*-factors. To address this possibility, we investigated the gene expression of 22 important polyA *trans*-factors[41] based on the TCGA RNA-seq data. The significantly up- and downregulated factors between tumours and matched normal tissues are indicated by yellow and blue, respectively (Fig. 6b). In general, we observed global upregulation of most polyA factors in five tumour types (LUSC, LUAD, UCEC, BLCA and BRCA), which also have more 3′-UTR-shortening events. Therefore, we conclude that most core polyadenylation factors are expressed at higher levels in tumour types where proximal APAs

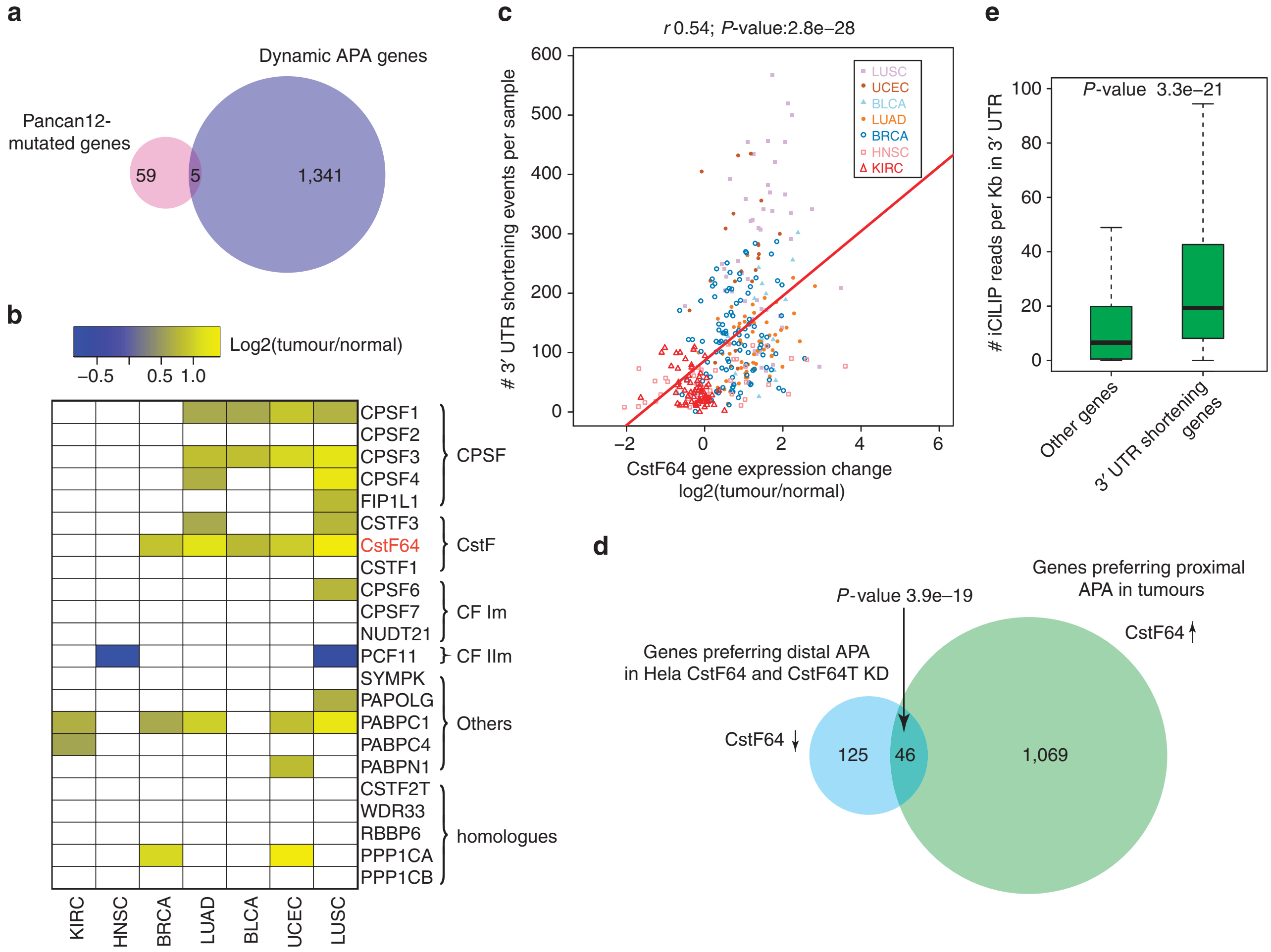


Figure 6 | Potential mechanisms for APA regulation during tumorigenesis. (**a**) Only five genes are in common between genes undergoing dynamic APA and genes significantly mutated in Pancan12 tumour types. (**b**) Heatmap of gene expression fold-change of known polyadenylation factors. Each rectangle represents the mean log2 fold-change between tumour and matched normal tissues of one factor in one tumour type. A factor is considered differentially expressed if the false-discovery rate from edgeR[55] is <0.05 and the mean absolute fold-change is >1.5. Yellow and blue boxes indicate the significantly upregulated and downregulated genes, respectively. White boxes are non-significant genes. (**c**) Correlation between *CstF64* expression fold-change and number of 3′-UTR-shortening events per sample. Each point represents a patient sample across seven tumour types. The *x* axis is the *CstF64* log2 fold-change between tumours and matched normal tissues. The *y* axis is the number of shortening events per sample. Spearman's correlation coefficient (0.54) and *P* value (2.8e − 28) are indicated on the top. (**d**) Venn diagram comparison between genes preferring proximal APAs in tumour with higher expression of *CstF64* and genes preferring distal APA in Hela cells with knockdown of *CstF64* and *CstF64T*. (**e**) Genes with 3′-UTR shortening in tumours have more *CstF64* iCLIP data derived from HeLa cells than background (*P* value 3.3e − 21 by *t*-test).

are favoured. Our results are consistent with previous studies showing that 3′-UTR shortening in proliferating cells is also accompanied by an increased expression of polyadenylation factors[9,12,27].

We further investigated the correlation between gene expression and 3′-UTR shortening for four polyadenylation factors (*CPSF1*, *CPSF3*, *CstF64* and *PABPC1*), which are differentially expressed between tumour and normal in at least three cancer types (Fig. 6b). Among them, *CstF64* has the greatest correlation between gene expression fold-change and the number of shortening events per patient in tumours (Spearman's correlation 0.54 with *P* value 2.8e − 28, Fig. 6c), followed by *CPSF3*. In contrast, *CPSF1* and *PABPC1* have weak correlations (Supplementary Fig. 9). This result is consistent with a recent iClip-seq study, suggesting that *CstF64* is one of the top three most important factors for polyA site selection[42]. Also, a recent study indicated that CPSF plays an important role in recruiting CstF64 to RNAs[43]. Furthermore, a recent global study in HeLa cells suggests that *CstF64* induces the usage of proximal APAs[43]. They reported 171 genes with lengthening in 3′ UTRs upon knockdown of *CstF64*, among which 46 genes from our analysis have shortened 3′ UTRs in tumours where *CstF64* is upregulated (Fig. 6d; *P* value = 3.9e − 19 using Fisher's exact test; Supplementary Fig. 10). This significant overlap indicates that a subset of 3′-UTR-shortening events we observed in tumours can indeed be explained by the expression level of *CstF64*. Finally, using *CstF64* iCLIP-seq in HeLa cells[43], we showed that those 1,346 genes have more *CstF64* bindings in their 3′ UTRs than other genes (Fig. 6e). Together, our study provides strong evidence that key polyA *trans*-factors, such as *CstF64*, are upregulated in tumorigenesis, leading to preferential 3′-UTR shortening in tumours.

Discussion

We have developed a novel bioinformatics algorithm, termed DaPars, dedicated to the *de novo* identification and quantification of dynamic APA events using standard RNA-seq. The accuracy of DaPars is evidenced by the fact that our *de novo* predicted APAs are enriched for the canonical polyA signal AATAAA and have a high degree of overlap with annotated polyA sites (Fig. 2b,c). Our extensive DaPars analysis of TCGA data sets convincingly demonstrate that any investigator(s) conducting standard RNA-Seq is now capable of identifying the majority of functionally important APA events in most biological systems. DaPars is not just yet another APA assay; instead, its key methodology innovation is the inference of *de novo* APA events from existing RNA-seq data without relying on any additional wet-bench experiments. For example, our current APA analysis was based on RNA-seq of 358 tumour/normal pairs across 7 cancer types. An analysis of this scale would be prohibitively cumbersome using any previous method, such as microarrays, EST and PolyA-seq, but was made possible now with our DaPars method.

While our paper was under review, Wang *et al.*[44] reported a change-point model to detect 3′-UTR switching using RNA-seq. The model by Wang *et al.* relies on the annotated distal polyA sites to infer the proximal ones, only supports genes with two polyA sites and only supports pair-wise comparison. In contrast, our DaPars method is fully *de novo*, can handle multiple (>2) polyA sites and multiple (>2) samples and thus is much more powerful and flexible than the model by Wang *et al.*. Most of our analyses based on hundreds of TCGA patient samples would not be possible using the model by Wang *et al.*

It has been reported that shorter 3′ UTRs are preferentially used by several oncogenes in cancer cell lines[8], but what was not clear from this work is how pervasive and recurrent APA is in clinical samples. Lin *et al.*[45] reported 126 3′-UTR-shortening genes in 5 tumour/normal pairs but unfortunately did not provide a supplementary table for those genes. To directly compare our results with Lin *et al.*[45], we repeated the same analysis as described in their paper and detected a total of 120 genes with 3′-UTR shortening and upregulation of the short isoform. Among them, 53% were also found in our 1,201 shortening APA genes (Supplementary Fig. 11; *P* value 2e − 43 by Fisher's exact test), including *POLR2K*, the main APA gene reported by Lin *et al.* Two examples of consistence and inconsistence between TCGA RNA-seq and PolyA-seq from Lin *et al.*[45] are shown in Supplementary Figs 12 and 13, respectively. In this study, we have substantially increased the number of dynamic APA events based on 358 tumour/normal pairs. To our knowledge, this is the largest global APA analysis to date, leading to a 71-fold increase in sample size compared with Lin *et al.*[45]

Several novel and significant biological and clinical insights are noticeable from our large-scale APA analysis. First, dynamic APA events are highly tumour type and patient specific. We observe that lung, uterus, breast and bladder cancers have significantly more APAs than head/neck and kidney cancers. Moreover, similar to other caner genomic data, there is considerable APA heterogeneity among patients within the same tumour type. Second, our saturation analysis indicates that APA events derived from 358 samples across 7 tumour types remain far from complete, highlighting the need for *de novo* discovery of APA, and the need for expanding DaPars analysis to more tumour types and samples when they become available. Third, selected APA events provide a surprisingly strong additional prognostic power beyond common clinical covariates and conventional molecular data, such as somatic mutation and gene expression. A recent study[46] also indicated that conventional molecular data had poor prognostic power beyond clinical data. Although the exact cause is unknown, we speculate that it may reflect a role for APA as a driver of tumour progression. Fourth, our study reveals a novel link between altered 3′-UTR usage and cancer metabolism. We observed that the *GAC* isoform of the glutaminase gene (*GLS*), which lacks any predicted miRNA binding, is predominantly expressed in LUSC, LUAD and KIRC tumours. Therefore, this APA event would abrogate the need to attenuate miR-23 expression through *MYC* upregulation and result in increased Glutaminase expression and altered glutamine metabolism. Fifth, our observation of correlating *CstF64* levels with increased 3′-UTR shortening suggests that this factor is a potential master activator of proximal APA usage in tumorigenesis. This hypothesis predicts that tumour cells increase *CstF64* levels to promote the 3′-end processing at the proximal and weaker polyA sites thereby preventing the usage of the distal polyA sites[39,43]. Finally, APA is likely to be regulated by many factors in a tissue-specific manner. For example, we recently reported *CFIm25* (ref. 47) as a global repressor of proximal APA in brain tumour. *CFIm25* has opposite function of *CstF64,* since its decreased expression correlates with increased 3′-UTR shortening. However, *CFIm25* is not a master APA regulator in the cancer types we studied here, because it is not differentially expressed between tumour and normal (*NUDT21* in Fig. 6b).

Our DaPars analysis of RNA-seq reveals a comprehensive list of previously unobserved, highly recurrent and functionally important 3′-UTR somatic 'RNA aberrations'. These RNA aberrations represent an illustrative case of genomic 'dark matter' beyond coding regions, and thus may also provide new directions for tumour gene discovery[48]. Although there is a lack in observed genetic aberrations within 3′ UTRs of most genes undergoing APA, caution should be taken as current TCGA mutation analyses utilize primarily exome sequencing, which excludes 3′ UTR. We will revisit this issue in the future when more whole-genome sequencing data are available. Finally, although focused on cancer genomics in this study, our novel DaPars framework will open the door for APA analysis in numerous biological and pathologic systems. It also underscores the power of innovative bioinformatics analyses that can derive novel biological insights from existing sequence data.

Methods

Data sets. All the RNA-seq BAM files were downloaded from the UCSC Cancer Genomics Hub (CGHub, https://cghub.ucsc.edu/). Here we only processed BRCA, BLCA, LUSC, LUAD, head and neck squamous cell carcinoma, UCEC and KIRC cancers that have >10 tumour–normal pairs (Supplementary Table 1). Gene expression and miRNA expression were downloaded from The Cancer Genome Atlas data portal (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm). MAQC brain and UHR RNA-seq reads were obtained from Sequence Read Archive with accessions ERP000016 and ERP000400, respectively. For MAQC PolyA-seq, the filtered polyA sites with normalized read counts were downloaded from the UCSC browser[3].

DaPars algorithm. DaPars performs *de novo* identification and quantification of dynamic APA events between two conditions, regardless of any *prior* APA annotation. DaPars identifies a distal polyA site based on RNA-seq data, uses a regression model to infer the exact location of the proximal APA site after correcting the potential RNA-seq non-uniformity bias along gene body, detects statistically significant dynamic APAs and has the potential to detect >2 dynamic APA events.

Distal polyA site identification from RNA-seq. Given two or more RNA-seq samples, distal polyA site refers to the end point of the longest 3′ UTR among all the samples, which will be used in the next step to identify the proximal polyA within this longest 3′-UTR region. To identify possible distal polyA site that may locate outside of gene annotation, we extend the annotated gene 3′ end by up to10 kb before reaching a neighbouring gene. RNA-seq data from all input samples will be merged to have a combined coverage along the extended gene model. To address possible uneven and discontinuous issues, we applied a 50-bp window to smooth this combined coverage. We then scan the extended 3′ UTR from 5′ to 3′ to find the distal polyA site whose coverage is significantly lower (that is, < predefined cutoff at 5%) than the coverage at the start of the preceding exon. A similar strategy has been successfully used to detect lengthening of 3′ UTRs in the mammalian brain[21]. The *de novo* distal APA estimated directly from RNA-seq, which may not be included in gene model, will benefit the downstream proximal APA identification (Supplementary Fig. 1a).

Since most current RNA-seq data sets are not strand-specific, potential overlapping of 3′ UTRs from two neighbouring 'tail-to-tail' genes from different strands may give false-positive distal polyA. So after previous distal APA analysis, if 3′ UTRs of two neighbouring genes overlap, we will gradually increase the cutoffs until the two 3′ UTRs are separated. In this way, we can recover the proper distal polyA, which may be overlooked by other methods such as Cufflinks (Supplementary Fig. 1b). The distal polyA site identification method implemented in DaPars has very good performance. For all the predicted distal polyA sites from TCGA RNA-seq, on average 81% are within 50 bp of the annotated polyA sites.

Regression model in DaPars. For each RefSeq transcript with a distal APA estimated from previous step, we use a regression model to infer the exact location of a *de novo* proximal polyA site at single-nucleotide resolution, by minimizing the deviation between the observed read density and the expected read density based on the two-polyA-site model, in both tumour and matched normal samples simultaneously. This regression model solves the following optimization problem:

$$(w_L^{1*}, w_L^{2*}, w_S^{1*}, w_S^{2*}, P^*) = \underset{w_L^1, w_L^2, w_S^1, w_S^2 \geq 0, 1<P<L}{\arg\min} \sum_{i=1}^{2} \|\ \mathbf{C}_i - (w_L^i \mathbf{I}_L + w_S^i \mathbf{I}_P)\ \|_2^2 \qquad (1)$$

where w_L^i and w_S^i are the abundances of transcripts with distal and proximal polyA sites for sample i, respectively, $\mathbf{C}_i = [\mathbf{C}_{i1}, \cdots, \mathbf{C}_{ij}, \cdots, \mathbf{C}_{iL}]^T$ is the read coverage of sample i at single-nucleotide resolution normalized by total sequencing depth, L is the length of the longest 3′ UTR from previous step, P is the length of alternative proximal 3′ UTR to be estimated, $\mathbf{I}_L$ and $\mathbf{I}_P$ are indicator functions such that $\mathbf{I}_L = [\underbrace{1, \cdots, 1}_{L}]$ and $\mathbf{I}_P = [\underbrace{1, \cdots, 1}_{P}, \underbrace{0, \cdots, 0}_{L-P}]$.

For each given $1 < P < L$, the expression levels of two transcripts with distal and proximal polyA sites in both tumour and normal tissues can be estimated by optimizing this linear regression model using quadratic programming[49]. The optimal *de novo* proximal polyA site P^* is the one with the minimal objective function value, as demonstrated by the vertical arrow in Fig. 1a. To quantify the relative polyA site usage, we define the PDUI for sample i as the following:

$$\mathrm{PDUI} = \frac{w_L^{i*}}{w_L^{i*} + w_S^{i*}} \qquad (2)$$

where w_L^{i*} and w_S^{i*} are the estimated expression levels of transcripts with distal and proximal polyA sites for sample i. The greater the PDUI is, the more distal polyA site of a transcript is used and vice versa. Finally, the regression model is extended towards the internal exons, so that splicing-coupled APA events can also be detected.

Non-uniformity correction. It has been reported that RNA-seq reads are not uniformly distributed along the gene body. DaPars provides an option to address the issue of non-uniformity by statistical modelling[50]. Since it is technically difficult to distinguish non-uniform distribution from dynamic APA, we decide to train our statistical model based on a subset of genes with no APA change, that is, with only one 3′ UTR. We first run DaPars to select those genes with no APA change and divide their RNA-seq gene body coverage into 100 bins, yielding an observed gene body-sequencing profile (Supplementary Fig. 1e). In the conventional DaPars, the elements of I_L and I_P in equation (1) are unweighted and all 1s on 3′-UTR regions. We will infer the weighted I_L and I_P based on the observed gene body-sequencing profile, then re-run DaPars with the weighted I_L and I_P to correct the non-uniformity in RNA-seq (Supplementary Fig. 1e).

Differential percentage of distal APA usage index. We used the following three criteria to detect the most significant APA events:

first, given long 3′-UTR expression level w_L^i and short 3′-UTR expression level w_S^i estimated from (equation (1)), we used Fisher's exact test to determine the P value of PDUI difference between tumour and matched normal tissue of the same patient, which is further adjusted by the Benjamini–Hochberg procedure to control the false-discovery rate (FDR) at 5%. Second, the absolute mean difference of PDUIs of all the patients in the same tumour type must be no less than 0.2. Third, the mean fold-change of PDUIs of all the patients in the same tumour type must be no less than 1.5.

$$\begin{cases} \mathrm{FDR} \leq 0.05 \\ |\Delta\mathrm{PDUI}| = |\mathrm{PDUI}_{\mathrm{tumor}} - \mathrm{PDUI}_{\mathrm{normal}}| \geq 0.2 \\ \left|\log_2\left(\frac{\mathrm{PDUI}_{\mathrm{tumor}}}{\mathrm{PDUI}_{\mathrm{normal}}}\right)\right| \geq 0.59 \end{cases}. \qquad (3)$$

To avoid false-positive estimation on lowly expressed genes, we only included genes with >30-fold mean coverage (reads per base gene model).

More than 2 dynamic APAs. Our DaPars framework can be easily extended to address >2 dynamic APAs. We formulated the multiple APA analysis in the following matrix format,

$$\begin{bmatrix} c_{11} & c_{21} \\ c_{12} & c_{22} \\ \vdots & \vdots \\ c_{1(m-1)} & c_{2(m-1)} \\ c_{1m} & c_{2m} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 0 & 1 & \cdots & 1 & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{m\times m} \begin{bmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ \vdots & \vdots \\ w_{1(m-1)} & w_{2(m-1)} \\ w_{1m} & w_{2m} \end{bmatrix}_{m\times 2} \qquad (4)$$

where m is the length of the longest 3′ UTR of a transcript. w_{ij} is the expression level of one possible 3′ UTR j on sample i. The number of non-zero w_{ij} determines how many polyA sites will be derived from RNA-seq. In most cases, there are only a few w_{ij} will be non-zero. So we can solve this equation using a positive Lasso optimization method as reformulated in the following form:

$$\underset{\mathbf{W}}{\arg\min} \frac{1}{2} \|\mathbf{C} - \mathbf{MW}\|_2^2 + \lambda \|\mathbf{W}\|_1 \qquad (5)$$

where **C**, **M** and **W** are corresponding to the left, middle and right matrix in equation (4), respectively. In practice, we only consider no more than 4 APAs in a real data set to reduce the complexity of model selection and avoid over-fitting issues. In Supplementary Fig. 1f, we showed that our DaPars can also identify >2 APAs from RNA-seq and the predictions are highly consistent with the annotation. Although many genes have >2 annotated APAs, the majority of dynamic APAs only involve 2 polyA sites[1]. Therefore in the current large-scale TCGA RNA-seq analysis, we only focus on 2 APAs in the dynamic APA detection.

PolyA-seq processing. We downloaded the processed polyA sites with normalized read counts of MAQC brain and UHR PolyA-seq data sets (two replicates for each tissue) from the UCSC Genome Browser[3]. We calculated the signal intensity of a given polyA site based on all the same-strand PolyA-seq reads within 50 bases of the polyA site. We then used Fisher's exact test to detect the statistically significant differential APAs between brain and UHR with a Benjamini–Hochberg-adjusted FDR cutoff of 0.1 and read count difference of >10%. For a fair comparison, we also used FDR of 0.1 and 10% ΔPDUI for DaPars analysis of MAQC RNA-seq data derived from the same brain and UHR samples.

Survival analysis using Cox proportional hazards model. A standard Cox proportional hazards model[29] implemented in the R package 'survival' was used for patient survival and Kaplan–Meier plotting. Hazard ratios exceeding 1 indicate poor prognosis for patients possessing shorter 3′ UTR, whereas those below 1 are associated with better outcome. The high-risk group and low-risk group were generated based on the prognostic index (PI). The PI is the linear component of the Cox model, $\mathrm{PI} = \sum_{i=1}^{m} \beta_i x_i$ where x_i is the value of covariate i and its risk coefficient, β_i was estimated from the Cox fitting. The high-risk and low-risk groups were generated for survival plot by splitting the ordered PI (higher values for higher risk) with equal number of samples in each group.

Survival analysis using Cox model and LASSO feature selection. We combined tumour-vs-normal shortening/lengthening events of APA genes (ΔPDUI values) with clinical covariates, such as age, gender, stage and smoking status (lung cancer), in survival analysis. We used a Cox regression model with LASSO feature selection to determine the contributions of APAs in survival prediction using the R package 'glmnet'[51]. We chose the optimal APA genes based on the leave-one-out CV. Here the clinical covariates are not penalized and always selected. Finally, we used a LRT to estimate the additional prediction power of the new APA-clinical models over the clinical-only models.

Software availability. The open source DaPars program is freely available at https://code.google.com/p/dapars/. We will update this website periodically with new versions.

References

1. Elkon, R., Ugalde, A. P. & Agami, R. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.* **14,** 496–506 (2013).
2. Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A. & Burge, C. B. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* **320,** 1643–1647 (2008).
3. Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res.* **22,** 1173–1183 (2012).
4. Tian, B., Hu, J., Zhang, H. & Lutz, C. S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33,** 201–212 (2005).
5. Barreau, C., Paillard, L. & Osborne, H. B. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Res.* **33,** 7138–7150 (2005).
6. Jacobsen, A. *et al.* Analysis of microRNA-target interactions across diverse cancer types. *Nat. Struct. Mol. Biol.* **20,** 1325–1332 (2013).
7. Takagaki, Y. & Manley, J. L. Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. *Mol. Cell* **2,** 761–771 (1998).
8. Mayr, C. & Bartel, D. P. Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138,** 673–684 (2009).
9. Elkon, R. *et al.* E2F mediates enhanced alternative polyadenylation in proliferation. *Genome. Biol.* **13,** R59 (2012).
10. Singh, P. *et al.* Global changes in processing of mRNA 3′ untranslated regions characterize clinically distinct cancer subtypes. *Cancer Res.* **69,** 9422–9430 (2009).
11. Morris, A. R. *et al.* Alternative cleavage and polyadenylation during colorectal cancer development. *Clin. Cancer Res.* **18,** 5256–5266 (2012).
12. Ji, Z., Lee, J. Y., 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. USA* **106,** 7028–7033 (2009).
13. Hoque, M. *et al.* Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nat. Methods* **10,** 133–139 (2013).
14. Sun, Y., Fu, Y., Li, Y. & Xu, A. Genome-wide alternative polyadenylation in animals: insights from high-throughput technologies. *J. Mol. Cell Biol.* **4,** 352–361 (2012).
15. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).
16. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7,** 1009–1015 (2010).
17. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
18. Griebel, T. *et al.* Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.* **40,** 10073–10083 (2012).
19. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11,** 94 (2010).
20. Consortium, M. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24,** 1151–1161 (2006).
21. Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J. O. & Lai, E. C. Widespread and extensive lengthening of 3′ UTRs in the mammalian brain. *Genome Res.* **23,** 812–825 (2013).
22. Ulitsky, I. *et al.* Extensive alternative polyadenylation during zebrafish development. *Genome Res.* **22,** 2054–2066 (2012).
23. Zhang, H., Hu, J., Recce, M. & Tian, B. PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res.* **33,** D116–D120 (2005).
24. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37,** W202–W208 (2009).
25. Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* **10,** 1001–1010 (2000).
26. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27,** 1653–1659 (2011).
27. Shepard, P. J. *et al.* Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17,** 761–772 (2011).
28. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).
29. Andersen, P. K. & Gill, R. D. Cox's regression model for counting processes: a large sample study. *Ann. Statist.* 1100–1120 (1982).
30. Hsu, P. P. & Sabatini, D. M. Cancer cell metabolism: Warburg and beyond. *Cell* **134,** 703–707 (2008).
31. Cassago, A. *et al.* Mitochondrial localization and structure-based phosphate activation mechanism of Glutaminase C with implications for cancer metabolism. *Proc. Natl Acad. Sci. USA* **109,** 1092–1097 (2012).
32. Aledo, J. C., Gomez-Fabre, P. M., Olalla, L. & Marquez, J. Identification of two human glutaminase loci and tissue-specific expression of the two related genes. *Mamm. Genome* **11,** 1107–1110 (2000).
33. de la Rosa, V. *et al.* A novel glutaminase isoform in mammalian tissues. *Neurochem. Int.* **55,** 76–84 (2009).
34. Elgadi, K. M., Meguid, R. A., Qian, M., Souba, W. W. & Abcouwer, S. F. Cloning and analysis of unique human glutaminase isoforms generated by tissue-specific alternative splicing. *Physiol. Genomics* **1,** 51–62 (1999).
35. Gao, P. *et al.* c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism. *Nature* **458,** 762–765 (2009).
36. van den Heuvel, A. P., Jing, J., Wooster, R. F. & Bachman, K. E. Analysis of glutamine dependency in non-small cell lung cancer: GLS1 splice variant GAC is essential for cancer cell growth. *Cancer Biol. Ther.* **13,** 1185–1194 (2012).
37. Stacey, S. N. *et al.* A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat. Genet.* **43,** 1098–1103 (2011).
38. Wiestner, A. *et al.* Point mutations and genomic deletions in CCND1 create stable truncated cyclin D1 mRNAs that are associated with increased proliferation rate and shorter survival. *Blood* **109,** 4599–4606 (2007).
39. Di Giammartino, D. C., Nishida, K. & Manley, J. L. Mechanisms and consequences of alternative polyadenylation. *Mol. Cell* **43,** 853–866 (2011).
40. Shi, Y. Alternative polyadenylation: new insights from global analyses. *RNA* **18,** 2105–2117 (2012).
41. Shi, Y. *et al.* Molecular architecture of the human pre-mRNA 3′ processing complex. *Mol. Cell* **33,** 365–376 (2009).
42. Martin, G., Gruber, A. R., 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.* **1,** 753–763 (2012).
43. Yao, C. *et al.* Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. *Proc. Natl Acad. Sci. USA* **109,** 18773–18778 (2012).
44. Wang, W., Wei, Z. & Li, H. A change-point model for identifying 3′ UTR switching by next-generation RNA sequencing. *Bioinformatics* **30,** 2162–2170 (2014).
45. Lin, Y. *et al.* An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res.* **40,** 8460–8471 (2012).
46. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10,** 1108–1115 (2013).
47. Masamha, C. P. *et al.* CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* **509,** 412–416 (2014).
48. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).
49. Bohnert, R. & Ratsch, G. rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res.* **38,** W348–W351 (2010).
50. Li, J., Jiang, H. & Wong, W. H. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol.* **11,** R50 (2010).
51. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33,** 1–22 (2010).
52. Grimson, A. *et al.* MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell* **27,** 91–105 (2007).
53. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120,** 15–20 (2005).
54. John, B. *et al.* Human MicroRNA targets. *PLoS Biol.* **2,** e363 (2004).
55. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

Acknowledgements

We would like to thank Benjamin Rodriguez for critical reading of this manuscript and members of the Li laboratory for helpful discussions. This work was partially supported by NIH R01HG007538, CPRIT RP110471 and DOD W81XWH-10-1-0501 (to W.L.), NIH grants CA167752 and CA166274 (to E.J.W.). We gratefully acknowledge the contributions from TCGA Research Network and its TCGA Pan-Cancer Analysis Working Group.

Author contributions

Z.X. and W.L. conceived the project, performed the analysis and wrote the manuscript. Z.X. developed the DaPars algorithms. Z.X., L.A.D., T.A.C., J.R.N., D.A.W., E.J.W. and W.L. interpreted the results and edited the manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Xia, Z. *et al.* Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nat. Commun.* 5:5274 doi: 10.1038/ncomms6274 (2014).