

RESEARCH

Open Access



# APC mutations dysregulate alternative polyadenylation in cancer

Austin M. Gabel<sup>1,2,3,4</sup>, Andrea E. Belleville<sup>1,2,3,5</sup>, James D. Thomas<sup>1,2</sup>, Jose Mario Bello Pineda<sup>1,2,3,4</sup> and Robert K. Bradley<sup>1,2,3\*</sup> 

\*Correspondence:  
rbradley@fredhutch.org

<sup>1</sup> Computational Biology Program, Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA

<sup>2</sup> Basic Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA

<sup>3</sup> Department of Genome Sciences, University of Washington, Seattle, WA, USA

<sup>4</sup> Medical Scientist Training Program, University of Washington, Seattle, WA, USA

<sup>5</sup> Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA

## Abstract

**Background:** Alternative polyadenylation (APA) affects most human genes and is recurrently dysregulated in all studied cancers. However, the mechanistic origins of this dysregulation are incompletely understood.

**Results:** We describe an unbiased analysis of molecular regulators of poly(A) site selection across The Cancer Genome Atlas and identify that colorectal adenocarcinoma is an outlier relative to all other cancer subtypes. This distinction arises from the frequent presence of loss-of-function *APC* mutations in colorectal adenocarcinoma, which are strongly associated with long 3' UTR expression relative to tumors lacking *APC* mutations. *APC* knockout similarly dysregulates APA in human colon organoids. By mining previously published *APC* eCLIP data, we show that *APC* preferentially binds G- and C-rich motifs just upstream of proximal poly(A) sites. Lastly, we find that reduced *APC* expression is associated with APA dysregulation in tumor types lacking recurrent *APC* mutations.

**Conclusions:** As *APC* has been previously identified as an RNA-binding protein that preferentially binds 3' UTRs during mouse neurogenesis, our results suggest that *APC* promotes proximal poly(A) site use and that *APC* loss and altered expression contribute to pervasive APA dysregulation in cancers.

## Background

Alternative cleavage and polyadenylation (APA) occurs in a majority of human genes and is known to vary across cell types and biological contexts, including cancer [1–6]. Mechanistic study of distinct APA events demonstrates that changes in poly(A) site selection can alter mRNA stability, translation kinetics, and localization [7–9]. APA is globally dysregulated in all studied human cancers [10, 11], and alterations in APA correlate with clinically relevant cancer phenotypes such as patient prognosis, immune infiltration, and others [10, 12–14].

Although canonical protein regulators of APA and 3' end processing of mRNAs have been well studied and demonstrated to contribute to cancer-associated dysregulation



© The Author(s) 2024, corrected publication 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

of APA, the mechanistic origins of the widespread APA dysregulation that characterizes almost all cancers remains incompletely understood. Previous analyses of tumor and matched normal control samples have identified *CSTF2* (*CstF64*) as a potential pivotal regulator of 3' UTR shortening across a subset of human cancers [11]. Interestingly, expression of other core components of the 3' end processing complex, including *NUDT21* and *PCF11*, which both have compelling evidence for a key role in poly(A) site selection from targeted knockdown experiments [15, 16] did not display strong associations with global changes in APA across tumor types [11]. Other studies have assessed correlations between gene expression of putative regulators of poly(A) site selection and APA, focusing on genes previously associated with the 3' end processing complex [10, 11, 16]. The potential contributions of other genes, including those that are not historically thought of as core components of the APA machinery, to modulation of poly(A) site selection in both healthy and cancerous cells likely have not yet been fully elucidated.

## Results

### Expression of canonical poly(A) regulators correlates with global polyadenylation site selection in all cancer subtypes except colorectal adenocarcinoma

To identify potential regulators of APA in human cancer, we assessed the correlation between gene expression and a summary statistic of 3' UTR across RNA-seq datasets of 29 distinct cancer subtypes from The Cancer Genome Atlas (TCGA). Our approach builds off of previous studies which focused on smaller subsets of tumors for which peritumoral, healthy control tissues are available (Fig. 1A) [10, 11]. For each RNA-seq sample within a given dataset, we imputed the 3' UTR length per gene using existing computational tools, in this case, DaPars [11]. We then took the median of all imputed gene level 3' UTR measurements to generate a median 3' UTR per sample that serves as a summary statistic of global poly(A) site selection [17]. A larger value indicates a given sample uses, on average, more distal poly(A) sites, and thus the 3' UTR of any given gene tends to be longer. We next calculated gene expression of all coding genes present in our lab annotation ( $N=15,332$ ) represented as transcripts per million (TPM). We then correlated the median 3' UTR and gene expression (TPM) across a given dataset and obtained a Pearson correlation coefficient per gene within each cancer type (Fig. 1B). A positive Pearson correlation coefficient indicates higher expression of that gene is associated with, on average, longer 3' UTRs in that cancer type, and a negative–positive Pearson correlation coefficient indicates the opposite.

We first sought to validate this approach for genes known to control poly(A) site selection globally. This includes *PABPN1*, where high levels are known to reduce proximal poly(A) site selection [16, 18], and the paralogs *CSTF2* and *CSTF2T*, where high expression correlates with increased use of proximal poly(A) site selection [10, 11, 19], and thus should have either a positive or negative correlation using our approach. We applied this method to RNA-seq data from tumors across 30 cancer subtypes in the Cancer Genome Atlas (TCGA) database and quantified gene expression—median 3' UTR correlations for each cancer subtype.

As expected, we observed *PABPN1* and *CSTF2T* expression are positively and negatively correlated with median 3' UTR length, respectively (Fig. 1C–D; Additional file 1: Fig. S1A–C). In addition, we identified strong correlations with other genes previously



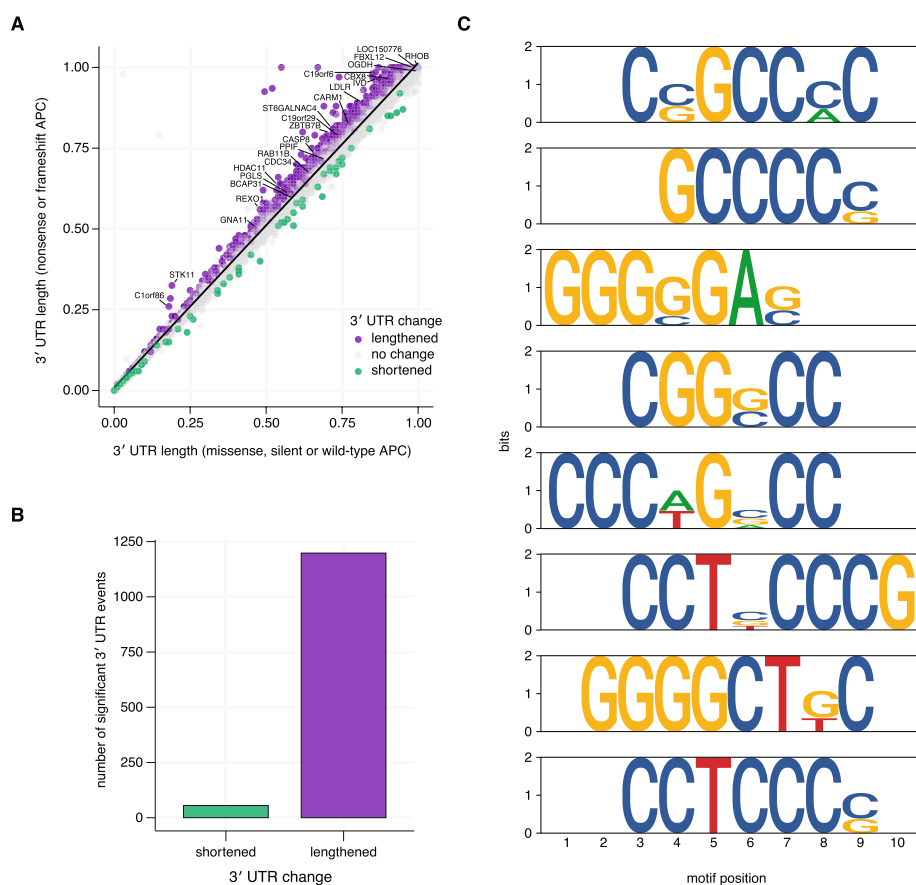
identified as core components of the mRNA 3' end processing machinery including *CSTF3*, *CPSF2*, and *NUDT21* (Additional file 1: Fig. S1D; Additional file 2–3: Table S1–S2). We next sought to understand how similar correlations of gene expression and 3' UTR length were between distinct cancer subtypes. Pairwise comparisons of subtype correlations of gene expression and median 3' UTR length were remarkably similar across most TCGA cancer subtypes analyzed, with a median pairwise Pearson correlation coefficient of 0.699 (Fig. 1E–F). Despite the concordance across most datasets, colorectal adenocarcinoma stood out as a striking outlier (median pairwise Pearson correlation coefficient of  $-0.14$ ), and known regulators of poly(A) site selection such as *PABPN1* and *CSTF2T* showed no correlation with median 3' UTR length, unlike in all other analyzed datasets (Fig. 1E–F; Additional file 1: Fig. S1A–D). This suggested there may be some unique feature of colorectal adenocarcinoma that obscures the expected correlations with known regulators of poly(A) site selection, perhaps by altering poly(A) site selection in a manner that is distinct from all other analyzed cancer subtypes.

#### **APC nonsense and frameshift mutations are associated with enhanced distal poly(A) site selection in colorectal adenocarcinoma**

Colorectal adenocarcinomas are largely driven by mutations in the gene adenomatous polyposis coli (*APC*) [20], and inherited loss-of-function mutations in *APC* cause familial adenomatous polyposis, which carries nearly a 100% risk of colorectal adenocarcinoma development at some point in life [21–24]. Most studies of *APC* in colorectal adenocarcinoma focus on the protein's role in regulating WNT signaling or cytoskeletal organization; however, several studies focused on murine neurogenesis have previously identified *APC* as an RNA-binding protein [25–27]. Crosslinking immunoprecipitation sequencing (CLIP-seq) of *APC*–RNA binding revealed more than 90% of identified *APC* binding sites are in the 3' UTR of the target mRNA [26]. This identified role of *APC* as an RNA-binding protein has been largely unexplored in the context of cancer, and a link between *APC* and global regulation of 3' UTR length is unreported in any context.

We first reasoned that frameshift or nonsense mutations in *APC* were more likely to disrupt RNA binding capacity, as opposed to missense mutations, which may have more heterogeneous effects on protein function. We stratified colorectal adenocarcinoma samples into two groups that either harbored at least one nonsense or frameshift mutation ( $N=342$ ) or those without a nonsense or frameshift mutation ( $N=282$ ). Samples with a frameshift or nonsense mutation displayed significant, global 3' UTR lengthening compared to samples without a frameshift or nonsense mutation (Fig. 2A–B; Additional file 4: Table S3). We validated this result with two distinct algorithms used to quantify 3' UTR length—DaPars and APalyzer—which use distinct computational approaches to quantify 3' UTR length (Additional file 1: Fig. S2A–E) [11, 28]. While this approach has limitations, specifically that some missense mutations may also result in protein loss-of-function, we validated the stratification by completing differential gene expression analysis, which revealed that samples with at least one nonsense or frameshift mutation expressed significantly higher levels of canonical WNT signaling genes as expected [20] (Additional file 1: Fig. S3A–C).

We then completed de novo motif enrichment of lengthened 3' UTRs to computationally predict potential *APC* binding sites [29] and found that the top 8 motifs were largely G- and C-rich (Fig. 2C). This is concordant with results from a prior study of mouse



**Fig. 2** APC nonsense and frameshift mutations are associated with enhanced distal poly(A) site use in colorectal adenocarcinoma. **A** Scatter plot of the median imputed 3' UTR length using DaPars for samples without APC loss-of-function mutations versus samples with a nonsense or frameshift APC mutation. All 3' UTR measurements per group were compared using a two-sided Wilcoxon rank-sum test. Comparisons with a Benjamin-Hochberg corrected  $p$  value  $< 0.05$  were called as significantly altered. Points labeled in green indicate significant shortening in APC nonsense or frameshift samples, points labeled in purple indicate significant lengthening in APC nonsense or frameshift samples, and points labeled in gray indicate no significant difference. **B** Bar plot of significantly altered lengthening (purple,  $n = 1198$ ) or shortening events (green,  $n = 55$ ) in APC nonsense or frameshift samples. **C** Sequence logo plot of de novo motif enrichment of 3' UTRs exhibiting significant lengthening in APC nonsense or frameshift colorectal adenocarcinoma samples compared to all other UTRs analyzed in panel A that exhibit no significant change in 3' UTR length. Top 8 significantly enriched motifs displayed

APC CLIP-seq data, which identified three consensus motifs: a G-rich motif, a C-rich motif, and CUGU [26]. These data indicate APC may function directly or indirectly to promote proximal poly(A) site use in a sequence-specific manner.

#### Targeted APC knockout in human *colon* organoids alters poly(A) site selection

As RNA-seq obtained from cancer samples are often extremely heterogeneous in terms of cell populations, genetic background, and sample quality, we wanted to assess alterations in APA in a more controlled experimental setting. We analyzed previously published data from human *colon* organoids at baseline or 24 h after targeted knockout of *APC* using CRISPR/Cas9 [30]. We computed differentially polyadenylated 3' UTRs and identified 196 shortened and 207 lengthened 3' UTRs 24 h after *APC* knockout (Fig. 3A; Additional file 5: Table S4).

We screened all analyzed 3' UTRs in the organoid data for presence of at least one of the top five enriched motifs from COAD (CSGGCCMC, GCCCCS, GGGGGAS, CGGSCC, CCCWGSCC) and found that lengthened 3' UTRs, but not shortened 3' UTRs, were significantly more likely to contain at least one of these motifs (73.8%) compared to unchanged 3' UTRs (61.4%) (Fig. 3B). We validated 3' UTR lengthening events by visualizing the RNA-seq data as overlaid BAM coverage plots and comparing the quantified 3' UTR length per condition for a number of genes (Fig. 3C–H). These data demonstrate that APC loss-of-function likely perturbs poly(A) site selection leading to 3' UTR lengthening of a subset of 3' UTRs enriched in G- and C-rich motifs.

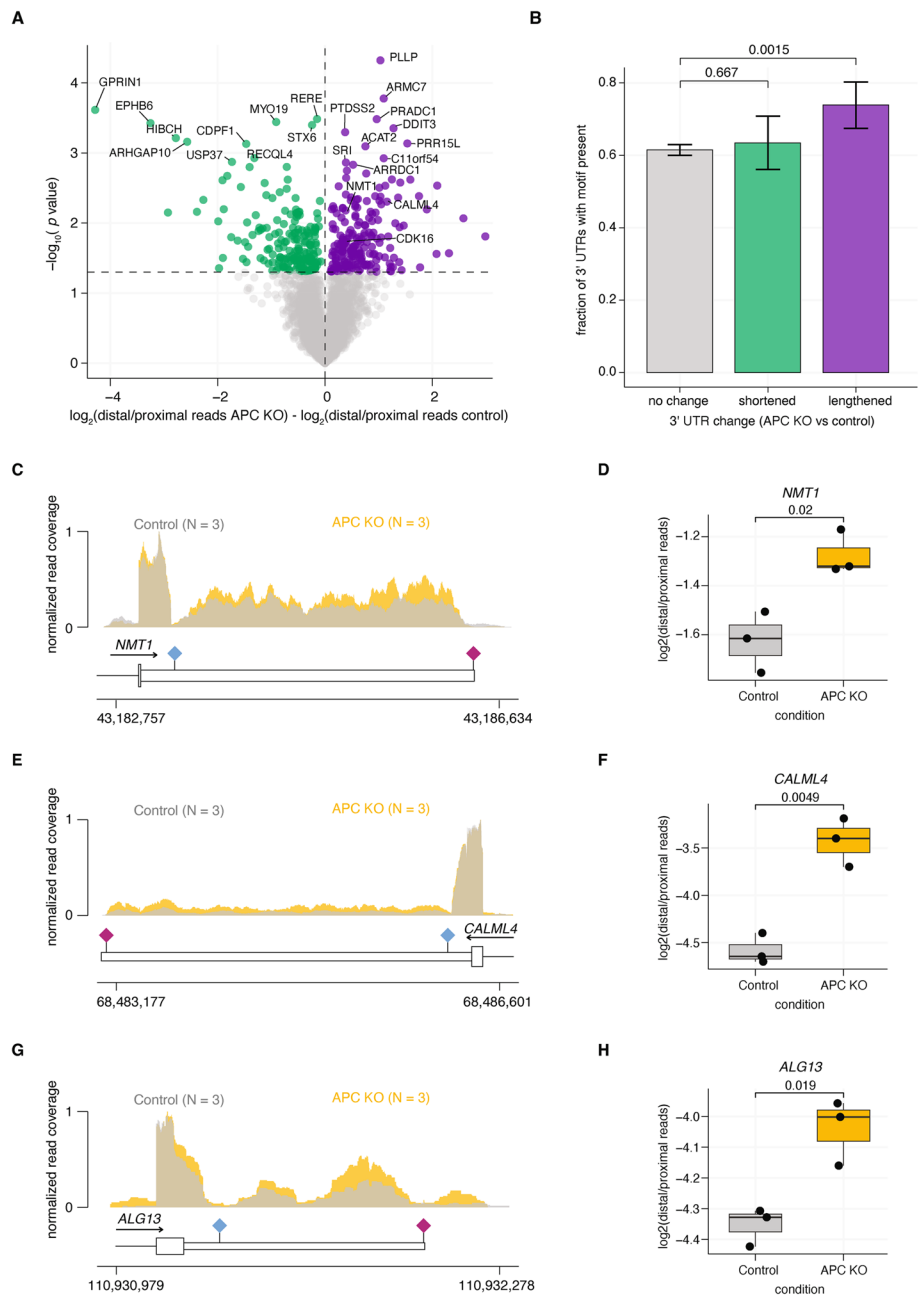
#### **APC binds mRNAs proximal to canonical poly(A) sites**

To gain further insight into the relationship between APC RNA binding and poly(A) site selection, we reanalyzed previously published APC HITS-CLIP data (Fig. 4A) [26]. We identified enriched APC binding sites in the CLIP data and searched a 1000-bp window around those sites to evaluate the distance to canonical poly(A) signal sequences (AAT AAA, ATTAAA) [3, 31]. We found that APC binding sites exhibited enrichment of canonical poly(A) sequences proximal to APC binding sites, with a maximum density of poly(A) signal sequences ~62 bp downstream of APC binding sites (Fig. 4B; Additional file 1: Fig. S4A; Additional file 6: Table S5). We then augmented those analyses by testing for association between APC binding and APA. We assessed where APC binds RNAs relative to proximal versus distal poly(A) sites within the 3' UTR. We used a published database of annotated poly(A) signal sequences [32] and measured the distance between APC binding sites and annotated poly(A) signal sequences to find that APC binds significantly closer to proximal and intermediate poly(A) signal sequences than it does to distal poly(A) signals (Fig. 4C–D). These analyses indicate that APC typically binds mRNAs immediately upstream of more proximal canonical poly(A) sequences within 3' UTRs.

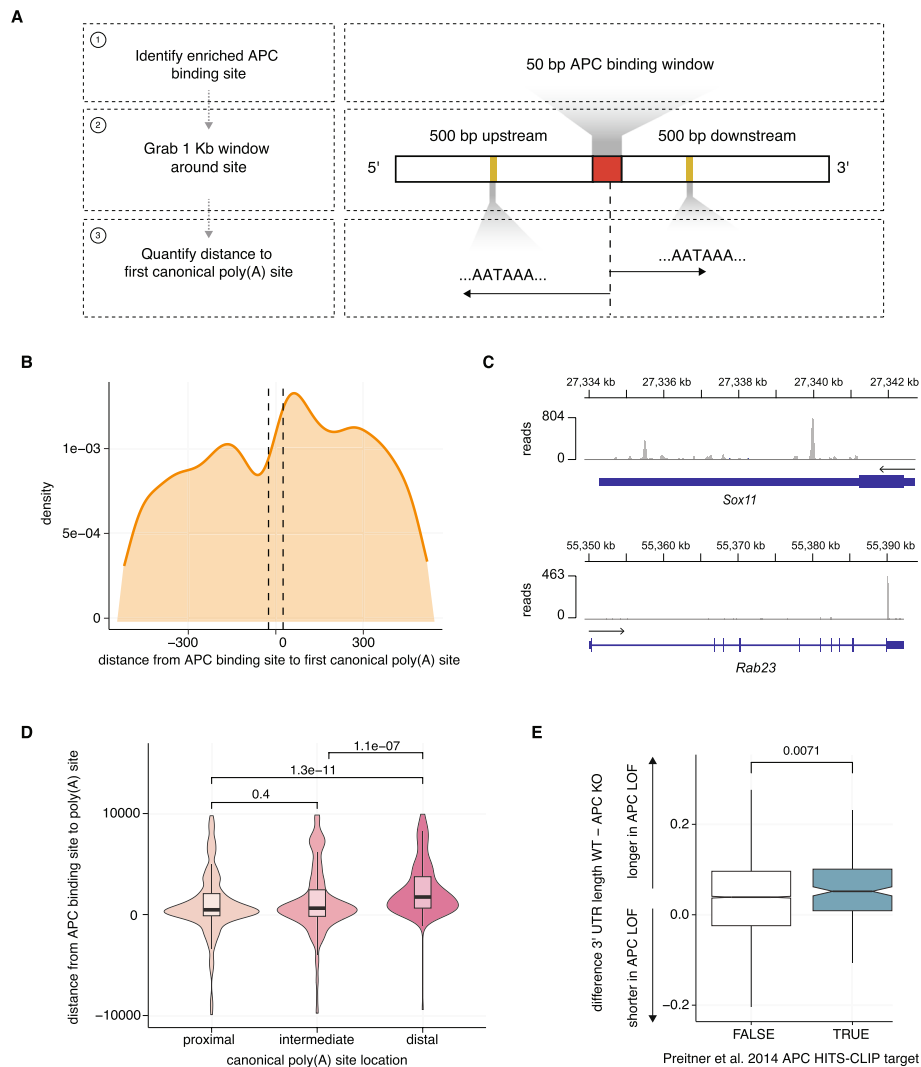
Given the similarity in identified motifs between our computational analyses of human transcriptomes and experimental measurements in mice (Fig. 2C) [26], we reasoned the identified RNA targets of APC binding may demonstrate more significant phenotypes than do other genes in our analysis. We assessed the degree of 3' UTR lengthening among APC-bound genes identified from HITS-CLIP data and found that those genes, on average, displayed significantly more lengthening than did other genes included in our analysis not identified as APC targets in Preitner et al. (Fig. 4E). When assessing the fraction of the APC binding targets displaying significant lengthening, we found that the observed fraction of genes displaying significant lengthening (17.6%) was significantly higher than what we would expect to observe based on random chance (Additional file 1: Fig. S4B–E). Taken together, these data demonstrate that APC binds RNA proximal to canonical poly(A) signal sequences and suggest that the RNA-binding targets identified in mice are potentially conserved in humans.

#### **APC expression correlates with 3' UTR length and expression levels of genes involved in growth factor signaling in low-grade glioma**

To evaluate if these analyses were relevant in other contexts, we assessed global correlations of poly(A) site selection and *APC* gene expression in other cancer subtypes. Brain tissues generally demonstrate the highest expression of *APC* [33] and, concordantly,



**Fig. 3** Acute induction of APC loss in colon organoids drives global dysregulation of poly(A) site selection. **A** Scatter plot of the difference in gene level 3' UTR lengths in human colon organoids with APC gene knockout using CRISPR/Cas9 comparing 0- versus 24-h timepoints. 3' UTRs that show significant shortening are indicated in green lengthening in purple. **B** Fraction of 3' UTRs identified in colon organoids as lengthened, shortened, or no change (panel **A**) that contain at least one of the five most significantly enriched motifs (CSGGCCMC, GCCCCS, GGGGGAS, CGGSCC, CCCWGSCC) identified from colorectal adenocarcinoma de novo motif enrichment analysis (Fig. 2D). *P* values were calculated using a Chi-squared test compared to the 3' UTRs in the no change group; error bars reflect 95% confidence interval using bootstrapping. **C**, **E**, **G** Bam coverage plots of RNA-seq data for individual 3' UTRs (*NMT1*, *CALML4*, *ALG13*) that display lengthening 24 h after APC knock out. Plots are overlaid for time 0 (gray, *n* = 3) or 24 h (yellow, *n* = 3). **D**, **F**, **H** Boxplots of 3' UTR length quantified per gene using APALyzer, values reflect  $\log_2(\text{distal/proximal reads})$ . *P* value reflective of a two-sided Student's *T*-test



**Fig. 4** APC binds mRNAs proximal to poly(A) signal sequences. **A** Overview of computational pipeline, including mapping of HITS-CLIP data and detection of enriched binding sites using DEWseq, definition of a 1000-bp window surrounding identified binding sites, and quantification of distance to the nearest canonical poly(A) signal sequences (AATAAA or ATTTAA). **B** Density plot of nearest canonical poly(A) sequence from identified APC binding sites using HITS-CLIP. Dotted lines indicate a 50-bp window of identified HITS-CLIP binding. Maximal signal occurring at +62.19 bp from the center of APC binding. **C** Representative read coverage plots of APC HITS-CLIP binding data. Plots illustrate two of the most-enriched genes, *Sox11* and *Rab23*, identified from reanalysis of the APC HITS-CLIP binding data from Preitner et al. (2014) [26]. **D** Violin plots of distance from APC binding site to high-confidence poly(A) sites from PolyADB v3 [32]. Poly(A) sites are binned into either proximal (most 5' poly(A) signal within the 3' UTR), distal (most 3' poly(A) signal within the 3' UTR), or intermediate (all other poly(A) sites). *P* values from a two-sided Wilcoxon rank-sum test. **E** Boxplot comparing the differences in change of 3' UTR length between colorectal adenocarcinoma samples with or without APC loss-of-function mutations for genes identified as direct binding targets of APC from Preitner et al. (2014) (blue) or genes not identified in their experiments (white). A more positive value indicates, on average, a longer 3' UTR in the APC loss-of-function colorectal adenocarcinoma samples. *P* value from a two-sided Wilcoxon rank-sum test



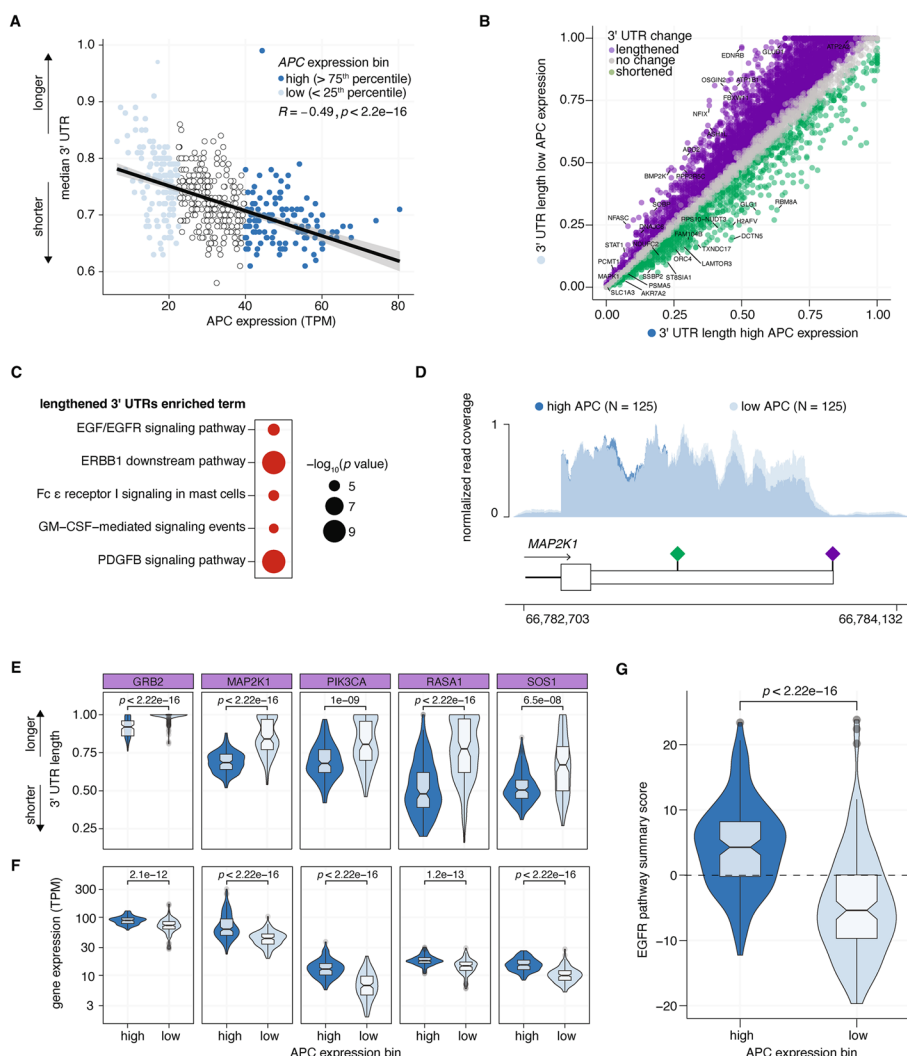
low-grade gliomas display the highest expression of *APC* among all TCGA cancer subtypes (Additional file 1: Fig. S5A). We first analyzed the correlation between *APC* expression and global median 3' UTR length and found that higher *APC* expression was associated with significantly shortened 3' UTRs globally (Fig. 5A). We next stratified patients into high (>75th percentile) and low (<25th percentile) *APC* expression to identify differentially polyadenylated genes, which revealed 5517 lengthening and 1158 shortening events in *APC* low expression samples (Fig. 5B; Additional file 7: Table S6). Gene ontology (GO) analysis of lengthened 3' UTRs revealed significant enrichment for genes involved in a number of growth factor signaling pathways, including EGFR and PDGF signaling cascades (Fig. 5C). EGFR signaling is clinically relevant and therapeutically targetable in low-grade gliomas as well as other brain and non-brain malignancies [34], and so we sought to further characterize APA of genes involved in EGFR signaling pathways.

We confirmed 3' UTR lengthening of several genes involved in EGFR signaling cascades that corresponded with significant reductions in gene expression that clustered by *APC* expression (Fig. 5D–F). Of 28 genes identified as hallmark EGFR signaling cascade genes, 21 demonstrated significant 3' UTR lengthening, 7 displayed no change and 1 displayed significant shortening in *APC* low expression low-grade glioma samples (Additional file 1: Fig. S5B–E). We next devised an EGFR summary score, defined as the sum of *Z* scores of gene expression across 28 genes per sample, and found that *APC* low expressing samples demonstrated significantly reduced EGFR signaling scores compared to *APC* high expression samples (Fig. 5G). These data demonstrate that *APC* expression levels correlate with global alterations in poly(A) site selection, which are associated with significant alterations in gene expression of a clinically actionable signaling pathway.

## Discussion

Here, we build on previous insights that *APC* is an RNA-binding protein that binds overwhelmingly in the 3' UTR during mouse neurogenesis [26]. We extend those findings to human cancers and show that *APC* loss-of-function mutations as well as reduced expression are strongly correlated with 3' UTR lengthening for many RNA targets. Among these targets are WNT signaling genes, as demonstrated by changes in 3' UTR length in low-grade glioma (Fig. 5C–G). Differential expression of *APC* and corresponding altered poly(A) site selection could thereby influence gene expression of key growth-promoting pathways, including WNT signaling genes, that are implicated in tumorigenesis. Such a hypothesis is concordant with observations from colorectal adenocarcinoma, where homozygous deletions of *APC* are extremely rare and thought to arise from a need to optimize  $\beta$ -catenin activity [35]. The functional and potentially pro-tumorigenic consequences of the 3' UTR lengthening associated with *APC* loss-of-function mutations or reduced expression remains unclear—in particular, it is unknown which specific 3' UTR lengthening events are relevant to cancer progression—but this important question could be experimentally answered by functionally interrogating poly(A) site selection in a multiplexed fashion in relevant biological models [17].

Our study motivates several avenues of investigation for future work. First, targeted knockdown of *APC* coupled with sequencing methods developed to unambiguously



**Fig. 5** Low APC expression correlates with 3' UTR lengthening in low-grade glioma. **A** Scatter plot of median 3' UTR versus APC gene expression (TPM) per sample in the TCGA low-grade glioma cohort ( $N = 500$ ). Samples are colored by expression bin.  $P$  value and  $R$  from Pearson correlation. **B** Differential poly(A) site use in high versus low APC expressing low-grade glioma samples. Genes with 3' UTR lengthening are indicated in purple, no change in gray, and shortening in green. **C** Top 5 enriched pathways of lengthened 3' UTRs in APC low expressing low-grade glioma samples. **D** BAM coverage plot of RNA-seq of MAP2K1 3' UTR from low-grade glioma samples stratified by APC expression bin ( $N = 125$  each bin). Schematic of MAP2K1 3' UTR indicates the proximal poly(A) sequence with a green diamond and the distal poly(A) sequence with a purple diamond. **E** Violin plot of 3' UTR length of several EGFR signaling genes in high versus low APC expressing stratified low-grade glioma samples.  $P$  value from two-sided Wilcoxon rank-sum test. **F** Violin plot of gene expression of several EGFR signaling genes in high versus low APC expressing stratified low-grade glioma samples.  $P$  value from two-sided Wilcoxon rank-sum test. **G** Violin plots of EGFR signaling pathway summary score (summed  $Z$  score of gene expression across 28 genes) per sample for low-grade glioma samples stratified by APC expression bin.  $P$  value from two-sided Wilcoxon rank-sum test

map alterations in poly(A) site selection such as Poly(A)-seq could refine and focus subsequent functional study of individual genes exhibiting significantly altered 3' UTRs. Second, further study of protein–protein and protein-RNA interactions with techniques like eCLIP or proximity labeling is needed to strengthen our analysis indicating that APC frequently binds RNA proximal to canonical poly(A) sequences

(Fig. 4B). Such studies could validate this result in human cells and yield specific insights into the molecular mechanisms by which APC shapes poly(A) site selection. For example, immunoprecipitation mass spectrometry data has identified components of the U4/5/6 tri-snRNP complex as direct interactors with APC [36]. While these components are not classically linked to poly(A) site selection, direct binding of APC to mRNA as well as components of the canonical poly(A) processing or splicing machineries could directly or indirectly promote cleavage at proximal poly(A) sites. Finally, focused investigation of how the altered poly(A) site selection driven by APC loss influences tumorigenesis and tumor phenotypes may yield new insights into the process of cellular transformation. For example, rapid cell division is strongly associated with global 3' UTR shortening [37], raising interesting questions about cellular growth kinetics following the acquisition of *APC* loss-of-function mutations.

There are several limitations to our work that are important to note. First, the association between *APC* knockdown and 3' UTR lengthening in organoids is clearly evident for transcripts containing enriched motifs identified via motif enrichment analyses in primary human cancer samples but not when this restriction is removed. We hypothesize that this may be due to the relatively short interval (24 h) between *APC* knockdown and collection for RNA-seq but cannot directly test this hypothesis without additional experimental evidence. Second, we have not directly demonstrated that APC is directly (mechanistically) responsible for changes in poly(A) site selection in human cells. Although previous studies in mouse cells demonstrated direct binding of APC to 3' UTRs [26], we cannot exclude the hypothesis that the global shifts in APA site selection that we observe occur downstream of changes in gene expression induced by low levels of functional APC. Given this, we cannot say with certainty that *APC* mutations are solely or even primarily responsible for the distinct global patterns of APA that we observed in colorectal adenocarcinoma samples (Fig. 1E–F). Third, although we present exploratory analyses indicating that *APC* expression is correlated with alterations in APA as well as with expression of genes in pathways critical for tumor growth and survival, additional work is needed to assess the potential functional relevance of those changes to tumorigenesis. We hope that this work motivates future study to mechanistically characterize how *APC* shapes poly(A) site selection in cancer as well as functional study of specific APA events that contribute to tumorigenesis.

## Methods

### RNA-seq mapping and analysis

RNA-seq was analyzed as previously described [38]. RNA-seq reads were mapped to an annotated transcriptome created using Ensembl 71 [39] UCSC knownGene [40] and Misov2.0 [41] annotations using RSEM version 1.2.4 [42] (modified to call Bowtie [43] with option “-v 2”). Unaligned reads were then mapped to the human genome (hg19/GRCh37 assembly) and a database containing all possible pairings of 5' and 3' splice sites per gene in our merged transcriptome annotation using TopHat version 20.8b [44]. All mapped reads were merged and then input into MISO v2.0. For TCGA studies, we analyzed 9045 available samples across 30 cancer types.

### 3' UTR analyses

For TCGA data, gene level 3' UTR measurements were downloaded from a previously published database [45] or BAM files were analyzed using the APALyzer package in R [28] and the gene level  $\log_2(\text{distal reads/proximal reads})$  was computed per sample. For the Schwank 2020 organoid data, RNA-seq data was mapped as stated above and BAM files were analyzed using the APALyzer package in R [28] and the gene level  $\log_2(\text{distal reads/proximal reads})$  was computed per sample for each respective strand.

### Median 3' UTR analyses and Pearson correlation

For each TCGA sample, we computed the median 3' UTR of all assayed genes, referred to as the median 3' UTR. We then assessed the Pearson correlation between gene expression (transcripts per million) computed as described above and the median 3' UTR among each TCGA cancer subtype for every gene in our lab annotation ( $N=15,332$ ). We then assessed the Pearson correlation coefficient  $R$  and the  $P$  value for each gene expression and median 3' UTR correlation with Bonferroni correction for multiple hypothesis testing.

We then completed pairwise Pearson correlation analysis of all the distinct Pearson correlations obtained for each TCGA cancer subtype. We calculated the median Pairwise Pearson correlation between the datasets by assessing every calculated Pearson correlation between each TCGA subtype and then taking the median value.

All statistical analyses were performed in R with Bioconductor, and the tables and plots were generated using dplyr [46] and ggplot2 [47] packages.

### APC mutation calling

Mutation Annotation Files (MAF) files for the TCGA colorectal adenocarcinoma dataset were downloaded from the NCI Genomic Data Consortium (GDC). For each sample, we identified if a patient harbored at least one nonsense, frameshift insertion, or frameshift deletion mutation in APC which were already annotated per GDC.

### De novo motif enrichment

For all sets of 3' UTRs, we downloaded available sequences [48]. We then utilized the R package memes version 5.5.0 [29], input the longest annotated 3' UTR per gene, and ran MEME analysis with default settings for ungapped de novo motif identification and enrichment analyses.

### Generation of an empirical distribution of expected rates of 3' UTR lengthening

Genes identified as direct APC binders were downloaded from supplementary data published from previous mouse neurogenesis HITS-CLIP data [26]. Of the identified genes, we identified the 210 human orthologs present in our colorectal adenocarcinoma APA analyses. We then assess the fraction of those genes that displayed significant shortening, lengthening, or no change in APC loss-of-function versus no loss-of-function mutations. We then randomly subsampled 10,000 random groupings of 210 genes from the differential APA analysis to generate an empirical distribution of the expected fraction of genes that would be expected to be lengthened, unchanged, or shortened by random chance.

### HITS-CLIP data analysis

Previously published HITS-CLIP data was downloaded from NCBI SRA (SRP042131) [26]. Data was then processed and analyzed in accordance with the DEWseq data preprocessing designed for eCLIP data [49]. In brief, FASTQ files were mapped and aligned to an annotated mouse genome (mm10) using the default setting in HISAT2 [50]. Cross-link sites were then extracted using HTSeq-CLIP using default settings with a sliding window approach with default window size and step settings [51]. Count matrices were then loaded into R and the reads per sliding scale window were pooled across all 4 available replicates to maximize statistical power. We then restricted analyses to sites containing at least 50 mapped reads. We then obtained the sequences of 500 bp up and downstream of the sliding window utilizing Biostrings [52] and then screened the up and downstream sequences for the first instance of a canonical poly(A) site relative to the HITS-CLIP binding site. For each HITS-CLIP target, we identified the nearest up and downstream canonical poly(A) signal sequences, AATAAA or ATTAAA [3, 31].

Annotated poly(A) sites were downloaded from a publicly available resource, PolyADB v3 [32]. We measured the distance from annotated APC binding sites to each annotated poly(A) site in a given 3' UTR. Poly(A) signal sequences were then binned as either proximal (most 5' within a given 3' UTR), distal (most 3' within a given 3' UTR), and intermediate (all other poly(A) signals within the 3' UTR).

### EGFR pathway expression and summary scoring

The canonical EGFR signaling pathway genes were identified from the Gene Set Enrichment Analysis (GSEA) database [53, 54]. For each sample, we assessed gene expression as described above (transcripts per million). For each individual sample, we also assessed the overall level of APC expression (transcripts per million) per sample and divided the TCGA low-grade glioma cohort into quartiles based on APC expression. We then calculated the individual Z score per gene within the EGFR signaling pathway. We then summed the Z scores of all canonical EGFR genes as a summary score of average EGFR signaling for each sample, where a higher score reflects on average higher expression of canonical EGFR genes and a lower score reflects on average lower expression of canonical EGFR genes.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-024-03406-4>.

Additional file 1: Figures S1-S5.

Additional file 2: Table S1. Pan-cancer genome wide Pearson correlation of gene expression and median 3' UTR. Individual Pearson correlation coefficients ( $R$ ) and  $P$  values associated between gene expression and sample median 3' UTR length. For each gene the ensemble gene ID and gene name are listed, along with the  $R$  and  $p$  value per TCGA cancer cohort identified with the standard TCGA subtype acronyms.

Additional file 3: Table S2. Summary Pearson correlation between gene expression and median 3' UTR length. Summary data generated for Additional file 1: Figure S1D reflecting the mean and standard deviation of Pearson correlation coefficients generated for all 29 TCGA cancer cohorts analyzed.

Additional file 4: Table S3. Differentially polyadenylated transcripts in colorectal adenocarcinoma samples harboring APC nonsense or frameshift mutations. Quantification of the average 3' UTR length per gene in samples with or without nonsense or frameshift mutations in APC. For each gene, the median 3' UTR length per group is identified which was quantified using DaPars [11].

Additional file 5: Table S4. Differentially polyadenylated transcripts in human colon organoids with APC gene knockout. APALyzer analysis output [28] indicating the difference in the  $\log_2(\text{distal/proximal reads})$  between human colon organoids with or without APC knockout.

Additional file 6: Table S5. APC HITS-CLIP reanalysis target enrichment. Reanalysis of APC HITS-CLIP data from Preitner et al. 2014 [26]. Enriched RNA targets with specific window, number of reads quantified per replicate, and specific gene annotation if enriched within a genic region. Data shown for all targets with more than 50 reads per target window total across all four replicates.

Additional file 7: Table S6. Differentially polyadenylated transcripts in low grade glioma samples with high or low APC gene expression. Quantification of the average 3' UTR length per gene in samples with either high or low APC gene expression. For each gene, the median 3' UTR length per group is identified which was quantified using DaPars [11].

Additional file 8: Review history.

### Acknowledgements

The results in this publication are based in part on data from The Cancer Genome Atlas Research Network (<http://cancergenome.nih.gov>).

### Peer review information

Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

### Review history

The review history is available as Additional file 8.

### Authors' contributions

A.G. and R.K.B. designed the study. A.G., A.E.B., J.D.T., and J.M.B.P. analyzed data and performed genomic analyses. A.G. and R.K.B. wrote the manuscript, with input from all authors.

### Authors' X handles

X handles: @a\_gabel2 (Austin M. Gabel).

### Funding

A.M.G. is an ARCS Foundation scholar. R.K.B. was supported in part by the NIH/NCI (R01 CA251138), NIH/NHLBI (R01 HL128239, R01 HL151651), and the Blood Cancer Discoveries Grant program through the Leukemia & Lymphoma Society, Mark Foundation for Cancer Research, and Paul G. Allen Frontiers Group (8023–20). R.K.B. is a Scholar of The Leukemia & Lymphoma Society (1344–18) and holds the McIlwain Family Endowed Chair in Data Science.

### Availability of data and materials

RNA-seq data generated by TCGA were downloaded from the Genomic Data Commons (GDC) and the Cancer Genomics Hub (CGHub). Human colon organoid RNA-seq data generated by Ringel et al., 2020 were downloaded from the Gene Expression Omnibus (GSE145185) [30, 55]. HITS-CLIP data from Preitner et al., 2014 were downloaded from the Sequence Read Archive (SRP042131) [26, 56]. Data from all analyses, including gene expression and 3' UTR level measurements and correlations generated for this paper, are included in the supplementary materials as follows. Pearson correlation coefficients from gene expression and median 3' UTR per dataset and the associated *P* value are included in Additional file 2: Table S1. Mean Pearson correlation coefficients (*R*) across all analyzed datasets are included in Additional file 3: Table S2. Median 3' UTR length per colorectal adenocarcinoma generated using the DaPars algorithm are included in Additional file 4: Table S3, and the imputed values for that same dataset but using APALyzer are included in Additional file 5: Table S4. Reanalysis of HITS-CLIP data from Preitner et al. 2014 [26] are included in Additional file 6: Table S5. Differential APA analysis using DaPars of TCGA LGG samples with high or low APC expression are in Additional file 7: Table S6. No other scripts or software were utilized other than those mentioned in the Methods section.

### Declarations

#### Ethics approval and consent to participate

This was not applicable to this study.

#### Competing interests

R.K.B. is a founder and scientific advisor of Codify Therapeutics and Synthesize Bio and holds equity in both companies. R.K.B. has received research funding from Codify Therapeutics unrelated to the current work. The remaining authors declare no competing interests.

Received: 27 November 2023 Accepted: 29 September 2024

Published: 7 October 2024

### References

1. Agarwal V, Lopez-Darwin S, Kelley DR, Shendure J. The landscape of alternative polyadenylation in single cells of the developing mouse embryo. *Nat Commun.* 2021;12(1). <https://doi.org/10.1038/s41467-021-25388-8>.

2. Cheng LC, Zheng D, Baljinnyam E, Sun F, Ogami K, Yeung PL, Hoque M, Lu CW, Manley JL, Tian B. Widespread transcript shortening through alternative polyadenylation in secretory cell differentiation. *Nat Commun.* 2020;11(1):1–14. <https://doi.org/10.1038/s41467-020-16959-2>.
3. Derti A, Garrett-Engle P, MacIsaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. A quantitative atlas of polyadenylation in five mammals. *Genome Res.* 2012;22(6):1173–83. <https://doi.org/10.1101/gr.132563.111>.
4. Gruber AJ, Zavolan M. Alternative cleavage and polyadenylation in health and disease. *Nat Rev Genet.* 2019;20(10):599–614. <https://doi.org/10.1038/s41576-019-0145-z>.
5. Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* 2013;27(21):2380–96. <https://doi.org/10.1101/gad.229328.113>.
6. Zhang H, Lee JY, Tian B. Biased alternative polyadenylation in human tissues. *Genome Biol.* 2005;6(12):1–13. <https://doi.org/10.1186/gb-2005-6-12-r100>.
7. Berkovits BD, Mayr C. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature.* 2015;522(7556):363–7. <https://doi.org/10.1038/nature14321>.
8. Ji Z, Lee JY, Pan Z, Jiang B, Tian B. Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci U S A.* 2009;106(17):7028–33. <https://doi.org/10.1073/pnas.0900028106>.
9. Spies N, Burge CB, Bartel DP. 3' UTR-Isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* 2013;23(12):2078–90. <https://doi.org/10.1101/gr.156919.113>.
10. Goering R, Engel KL, Gillen AE, Fong N, David L, Taliaferro JM. LABRAT reveals association of alternative polyadenylation with transcript localization, RNA binding protein expression, transcription speed, and cancer survival.
11. Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, Li W. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'2-UTR landscape across seven tumour types. *Nat Commun.* 2014;5. <https://doi.org/10.1038/ncomms6274>.
12. Hu C, Liu C, Li J, Yu T, Dong J, Chen B, Du Y, Tang X, Xi Y. Construction of Two Alternative Polyadenylation Signatures to Predict the Prognosis of Sarcoma Patients. *Front Cell Dev Biol.* 2021;9. <https://doi.org/10.3389/fcell.2021.595331>.
13. Zhang Y, Wang Y, Li C, Jiang T. Systemic Analysis of the Prognosis-Associated Alternative Polyadenylation Events in Breast Cancer. *Front Genet.* 2020;11. <https://doi.org/10.3389/fgene.2020.590770>.
14. Zhang Y, Xu Y, Wang Y. Alternative polyadenylation associated with prognosis and therapy in colorectal cancer. *Sci Rep.* 2022;12(1). <https://doi.org/10.1038/s41598-022-11089-9>.
15. Brumbaugh J, Di Stefano B, Wang X, Borkent M, Forouzmand E, Clowers KJ, Ji F, Schwarz BA, Kalocsay M, Elledge SJ, Chen Y, Sadreyev RI, Gygi SP, Hu G, Shi Y, Hochedlinger K. Nudt21 Controls Cell Fate by Connecting Alternative Polyadenylation to Chromatin Signaling. *Cell.* 2018;172(1–2):106–20.e21. <https://doi.org/10.1016/j.cell.2017.11.023>.
16. Li W, You B, Hoque M, Zheng D, Luo W, Ji Z, Park JY, Gunderson SI, Kalsotra A, Manley JL, Tian B. Systematic Profiling of Poly(A)+ Transcripts Modulated by Core 3' End Processing and Splicing Factors Reveals Regulatory Rules of Alternative Cleavage and Polyadenylation. *PLoS Genet.* 2015;11(4). <https://doi.org/10.1371/journal.pgen.1005166>.
17. Gabel AM, Belleville AE, Thomas JD, McKellar SA, Nicholas TR, Banjo T, Crosse EI, Bradley RK. Multiplexed screening reveals how cancer-specific alternative polyadenylation shapes tumor growth in vivo. *Nat Commun.* 2024;15(1):959. <https://doi.org/10.1038/s41467-024-44931-x>.
18. Jenal M, Elkon R, Loayza-Puch F, Van Haaften G, Kühn U, Menzies FM, Vrieling JAFO, Bos AJ, Drost J, Rooijers K, Rubinsztein DC, Agami R. The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell.* 2012;149(3):538–53. <https://doi.org/10.1016/j.cell.2012.03.022>.
19. Yao C, Choi EA, Weng L, Xie X, Wan JI, Xing YI, Moresco JJ, Tu PG, Yates JR, Shi Y. Overlapping and distinct functions of CstfF64 and CstfF64t in mammalian mRNA 3' processing. *RNA.* 2013;19(12):1781–90. <https://doi.org/10.1261/rna.042317.113>.
20. Fearnhead NS, Britton MP, Bodmer WF. The ABC of APC. *Hum Mol Genet.* 2001;10(7):721–33.
21. Groden J, Thliveris A, Samowitz W, Carlson M, Gelbert L, Albertsen H, Joslyn G, Stevens J, Spirio L, Robertson M. Identification and characterization of the familial adenomatous polyposis coli gene. *Cell.* 1991;66(3):589–600.
22. Joslyn G, Carlson M, Thliveris A, Albertsen H, Gelbert L, Samowitz W, Groden J, Stevens J, Spirio L, Robertson M. Identification of deletion mutations and three new genes at the familial polyposis locus. *Cell.* 1991;66(3):601–13. [https://doi.org/10.1016/0092-8674\(81\)90022-2](https://doi.org/10.1016/0092-8674(81)90022-2).
23. Kinzler KW, Nilbert MC, Su LK, Vogelstein B, Bryan TM, Levy DB, Smith KJ, Preisinger AC, Hedge P, McKechnie D. Identification of FAP locus genes from chromosome 5q21. *Science.* 1991;253(5020):661–5. <https://doi.org/10.1126/science.1651562>.
24. Nishisho I, Nakamura Y, Miyoshi Y, Miki Y, Ando H, Horii A, Koyama K, Utsunomiya J, Baba S, Hedge P. Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science.* 1991;253(5020):665–9. <https://doi.org/10.1126/science.1651563>.
25. Baumann S, Komissarov A, Gili M, Ruprecht V, Wieser S, Maurer SP. A reconstituted mammalian APC-kinesin complex selectively transports defined packages of axonal mRNAs. *Sci Adv.* 2020;6. <https://doi.org/10.1126/sciadv.abc3580>.
26. Preitner N, Quan J, Nowakowski DW, Hancock ML, Shi J, Tcherkezian J, Young-Pearse TL, Flanagan JG. APC is an RNA-binding protein, and its interactome provides a link to neural development and microtubule assembly. *Cell.* 2014;158(2):368–82. <https://doi.org/10.1016/j.cell.2014.05.042>.
27. Wang T, Hamilla S, Cam M, Aranda-Espinoza H, Mili S. Extracellular matrix stiffness and cell contractility control RNA localization to promote cell migration. *Nat Commun.* 2017;8(1). <https://doi.org/10.1038/s41467-017-00884-y>.
28. Wang R, Tian B. APALyzer: A bioinformatics package for analysis of alternative polyadenylation isoforms. *Bioinformatics.* 2020;36(12):3907–9. <https://doi.org/10.1093/bioinformatics/btaa266>.
29. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* 2009;37(SUPPL. 2). <https://doi.org/10.1093/nar/gkp335>.
30. Ringel T, Frey N, Ringnald F, Janjuha S, Cherkaoui S, Butz S, Srivatsa S, Pirkel M, Russo G, Villiger L, Rogler G, Clevers H, Beerwinkel N, Zamboni N, Baubec T, Schwank G. Genome-Scale CRISPR Screening in Human Intestinal Organoids Identifies Drivers of TGF- $\beta$  Resistance. *Cell Stem Cell.* 2020;26(3):431–40.e8. <https://doi.org/10.1016/j.stem.2020.02.007>.

31. Proudfoot NJ, Brownlee GG. 3' non-coding region sequences in eukaryotic messenger RNA. *Nature*. 1976;263(5574):211–4. <https://doi.org/10.1038/263211a0>.
32. Wang R, Nambiar R, Zheng D, Tian B. PolyA\_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res*. 2018 Jan 4;46(D1). <https://doi.org/10.1093/nar/gkx1000>.
33. Bhat RV, Baraban JM, Johnson RC, Eipper BA, Mains RE. High levels of expression of the tumor suppressor gene APC during development of the rat central nervous system. *J Neurosci*. 1994;14(5 Pt 2):3059–71. <https://doi.org/10.1523/JNEUROSCI.14-05-03059.1994>.
34. Hatanpaa KJ, Burma S, Zhao D, Habib AA. Epidermal growth factor receptor in glioma: signal transduction, neuropathology, imaging, and radioresistance. *Neoplasia*. 2010;12(9):675–84. <https://doi.org/10.1593/neo.10688>.
35. Albuquerque C, Breukel C, van der Luijt R, Fidalgo P, Lage P, Slors FJ, Leitão CN, Fodde R, Smits R. The 'just-right' signaling model: APC somatic mutations are selected based on a specific level of activation of the beta-catenin signaling cascade. *Hum Mol Genet*. 2002;11(13):1549–60.
36. Popow O, Paulo JA, Tatham MH, Volk MS, Rojas-Fernandez A, Loyer N, Newton IP, Januschke J, Haigis KM, Näthke I. Identification of Endogenous Adenomatous Polyposis Coli Interaction Partners and  $\beta$ -Catenin-Independent Targets by Proteomics. *Mol Cancer Res*. 2019;17(9):1828–41. <https://doi.org/10.1158/1541-7786.MCR-18-1154>.
37. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science*. 2008;320(5883):1643–7. <https://doi.org/10.1126/science.1155390>.
38. Dvinge H, Ries RE, Ilagan JO, Stirewalt DL, Meshinchi S, Bradley RK. Sample processing obscures cancer-specific alterations in leukemic transcriptomes. *Proc Natl Acad Sci U S A*. 2014;111(47):16802–7. <https://doi.org/10.1073/pnas.1413374111>.
39. Kersey PJ, Allen JE, Christensen M, Davis P, Falin LJ, Grabmueller C, Hughes DS, Humphrey J, Kerhornou A, Khobova J, Langridge N, McDowall MD, Maheswari U, Maslen G, Nuhn M, Ong CK, Paulini M, Pedro H, Toneva I, Tuli MA, Walts B, Williams G, Wilson D, Youens-Clark K, Monaco MK, Stein J, Wei X, Ware D, Bolser DM, Howe KL, Kulesha E, Lawson D, Staines DM. Ensembl 2013. *Nucleic Acids Res*. 2013;41(D1). <https://doi.org/10.1093/nar/gks1236>.
40. Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., Sloan, C. A., Rosenbloom, K. R., Roe, G., Rhead, B., Raney, B. J., Pohl, A., Malladi, V. S., Li, C. H., Lee, B. T., Learned, K., Kirkup, V., Hsu, F., Heitner, S., Harte, R. A., ... Kent, W. J. The UCSC Genome Browser database: Extensions and updates 2013. *Nucleic Acids Res*. 2013;41(D1). <https://doi.org/10.1093/nar/gks1048>.
41. Katz Y, Wang ET, Airolidi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*. 2010;7(12):1009–15. <https://doi.org/10.1038/nmeth.1528>.
42. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12. <https://doi.org/10.1186/1471-2105-12-323>.
43. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009;10(3). <https://doi.org/10.1186/gb-2009-10-3-r25>.
44. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105–11. <https://doi.org/10.1093/bioinformatics/btp120>.
45. Feng X, Li L, Wagner EJ, Li W. TC3A: The Cancer 3' UTR Atlas. *Nucleic Acids Res*. 2018;46(D1)–30. <https://doi.org/10.1093/nar/gkx892>.
46. Wickham H, François R, Henry L, Müller K, Vaughan D. dplyr: A Grammar of Data Manipulation. R package version 1.1.4. 2023. <https://CRAN.R-project.org/package=dplyr>.
47. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer; 2009.
48. O'Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badredin A, Bao Y, Blinkova O, Brover V, Chetverin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1). <https://doi.org/10.1093/nar/gkv1189>.
49. Schwarzl T, Sahadevan S, Lang B, Miladi M, Backofen R, Huber W, Hentze MW, Tartaglia GG. Improved discovery of RNA-binding protein binding sites in eCLIP data using DEWSeq. *Nucleic Acids Res*. 2024;52(1). <https://doi.org/10.1093/nar/gkad998>.
50. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019;37(8):907–15. <https://doi.org/10.1038/s41587-019-0201-4>.
51. Sahadevan S, Sekaran T, Ashaf N, Fritz M, Hentze MW, Huber W, Schwarzl T. htseq-clip: a toolset for the preprocessing of eCLIP/iCLIP datasets. *Bioinformatics*. 2023;39(1). <https://doi.org/10.1093/bioinformatics/btac747>.
52. Pagès H, Aboyou P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings. R package version 2.66.0. 2022.
53. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1 $\alpha$ -responsive genes involved in oxidative phosphorylation are coordinately down-regulated in human diabetes. *Nat Genet*. 2003;34(3):267–73. <https://doi.org/10.1038/ng1180>.
54. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–50. <https://doi.org/10.1073/pnas.0506580102>.
55. Ringel T, Frey N, Ringnalda F, Janjuha S, Cherkaoui S, Butz S, Srivatsa S, Pirkl M, Russo G, Villiger L, Rogler G, Clevvers H, Beerenwinkel N, Zamboni N, Baubec T, Schwank G. Pooled CRISPR screens in human intestinal organoids. GSE145185. *Gene Expression Omnibus*. 2020. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145185>.
56. Preitner N, Quan J, Nowakowski DW, Hancock ML, Shi J, Tcherkezian J, Young-Pearse TL, Flanagan JG. Adenomatous Polyposis Coli (APC) HITS-CLIP from E14 mouse brain. PRJNA248117. Sequence Read Archive. <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA248117> (2014).

## Publisher's Note

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