https://doi.org/10.1093/nargab/lqad051

Transcriptome sequencing suggests that pre-mRNA splicing counteracts widespread intronic cleavage and polyadenylation

Maria Vlasenok, Sergey Margasyuk and Dmitri D. Pervouchine*

Center for Molecular and Cellular Biology, Skolkovo Institute of Science and Technology, Bolshoy Bulvar 30, Moscow 121205, Russia

Received January 17, 2023; Revised May 09, 2023; Editorial Decision May 15, 2023; Accepted May 17, 2023

ABSTRACT

Alternative splicing (AS) and alternative polyadenylation (APA) are two crucial steps in the post-transcriptional regulation of eukaryotic gene expression. Protocols capturing and sequencing RNA 3′-ends have uncovered widespread intronic polyadenylation (IPA) in normal and disease conditions, where it is currently attributed to stochastic variations in the pre-mRNA processing. Here, we took advantage of the massive amount of RNA-seq data generated by the Genotype Tissue Expression project (GTEx) to simultaneously identify and match tissue-specific expression of intronic polyadenylation sites with tissue-specific splicing. A combination of computational methods including the analysis of short reads with non-templated adenines revealed that APA events are more abundant in introns than in exons. While the rate of IPA in composite terminal exons and skipped terminal exons expectedly correlates with splicing, we observed a considerable fraction of IPA events that lack AS support and attributed them to spliced polyadenylated introns (SPI). We hypothesize that SPIs represent transient byproducts of a dynamic coupling between APA and AS, in which the spliceosome removes the intron while it is being cleaved and polyadenylated. These findings indicate that cotranscriptional pre-mRNA splicing could serve as a rescue mechanism to suppress premature transcription termination at intronic polyadenylation sites.

INTRODUCTION

The majority of transcripts that are generated by the eukaryotic RNA Polymerase II undergo endonucleolytic cleavage and polyadenylation (CPA) at specific sites called the polyadenylation sites (PASs) (1). More than half of human genes have multiple PASs resulting in alternative polyadenylation (APA) (2,3). APA modulates gene expression by influencing mRNA stability, translation, nuclear export, subcellular localization, and interactions with microRNAs and RNA binding proteins (RBPs) (4,5). APA is widely implicated in human disease, including hematological, immunological, neurological disorders, and cancer (6,7).

APA can generate transcripts not only with different 3′-untranslated regions (3′-UTR) but also transcripts encoding proteins with different C-termini (8). Recent studies have shown that more than 20% of human genes contain at least one intronic PAS located upstream of the 3′-most exon (9). Intronic polyadenylation (IPA) can lead to important functional changes due to alterations in the protein primary sequence (10). For instance, IPA in *DICER* generates a truncated protein with impaired miRNA cleavage ability that results in decreased endogenous miRNA expression (11,12). Remarkably, the truncated oncosuppressor proteins that are generated by IPA often lack tumor-suppressive functions and contribute significantly to the tumor onset and progression (11).

The interplay between splicing and polyadenylation has long been recognized as being related to cotranscriptional pre-mRNA processing (13). Many splicing factors have dual roles serving both splicing and polyadenylation, including *U2AF* (14), *PTBP1* (15), members of Hu protein family (16) and others (8). The observation that IPA is associated with weaker 5′-splice sites and longer introns, and experiments on mutagenesis of CPA and splicing signals in plants together suggest that splicing and polyadenylation operate in a dynamic competition with each other (9,17). Furthermore, nascent RNA polymerase II transcripts that are susceptible to CPA at cryptic PASs are protected from it by U1 snRNP in a process called telescripting, most remarkably in genes with longer introns (18). These results raise a number of challenging questions about the actual abundance and function of cryptic intronic PASs, mechanisms of their inactivation, and relation to AS.

*To whom correspondence should be addressed. Tel: +7 495 280 14 81; Fax: +7 495 280 14 81;; Email: d.pervouchine@skoltech.ru

A number of experimental protocols have been developed to identify the genomic positions of PASs (19). Many of them use oligo(dT) (3'RNA-seq, PAS-seq, polyA-seq) or similar primers (3'READS) to specifically capture transcript ends (2,20–23). A combination of these protocols yielded a consolidated set of more than 500 000 human PASs (24–26), however many more PASs may be active in tissue- and disease-specific conditions. A number of computational methods also attempt to identify PASs from the standard polyA$^+$ RNA-seq data as genomic loci that exhibit an abrupt decrease in read coverage (27–32). However, since the density of RNA-seq reads is highly non-uniform along the gene length, many of these methods are limited to PASs that are located in the last exon or 3$'$-UTR, thus implicitly focusing on quantifying relative usage of PASs in the gene 3$'$-end rather than on identifying novel intronic PASs.

On the other hand, RNA-seq data contain an admixture of reads that cover the junction between the terminal exon and the beginning of the polyA tail. They align to the reference genome only partially due to a stretch of non-templated adenine residues. Although the fraction of such reads is quite small and normally does not exceed 0.1%, they can potentially be used for *de novo* identification of PASs. Previous studies such as ContextMap2 (30) and KLEAT (29) demonstrated that the analysis of RNA-seq reads containing a part of the polyA tail can offer a powerful alternative to coverage-based methods when analyzing a sufficiently large panel of RNA-seq experiments.

In this work, we took advantage of the massive amount of RNA-seq data generated by the Genotype Tissue Expression Project (GTEx), the largest to-date compendium of human transcriptomes, to simultaneously assess alternative splicing and intronic polyadenylation and match their tissue-specific patterns (33). Unlike previous studies, which extensively characterized the tissue-specific polyadenylation using coverage-based methods (31,34–36), here we focused specifically on intronic PAS by combining the information on polyA reads to identify PAS, split reads to measure the AS rate, and the read coverage to assess the CPA rate. We identified a core set of 318 898 PAS clusters that are stably expressed in GTEx tissues, which is consistent with other published sets, and characterized their attribution to the UTRs, exonic, and intronic regions. After normalizing the number of polyA reads to the background read coverage, we observed that intronic PAS are used more frequently than PASs in regions that are not spliced, i.e., exons. Moreover, in inspecting the concordance between IPA and AS, we unexpectedly found a considerable fraction of unannotated intronic PAS that are inconsistent with previously proposed IPA models (skipped and composite terminal exons). We attributed them to Spliced Polyadenylated Introns (SPI), a term we introduce here to describe transient byproducts of the dynamic coupling between CPA and AS, and conjecture that they are generated by the spliceosome removing the intron while it is being cleaved and polyadenylated.

MATERIALS AND METHODS

Genome assembly and transcript annotation

The February 2009 (hg19) assembly of the human genome and comprehensive GENCODE transcript annotation v34lift37 were downloaded from Genome Reference Consortium (37) and GENCODE websites (38), respectively.

Genome and gene partitions

To partition the genome, we considered genomic regions defined by the intervals annotated in the GENCODE database. A region that was not covered by any annotated gene was classified as intergenic. The remaining regions not covered by any annotated protein-coding gene were classified as non-coding, and those covered by at least one protein-coding gene were referred to as protein-coding. Further, a region was classified as 5$'$-UTR (respectively, 3$'$-UTR) if it belonged to the 5$'$-UTR (respectively, 3$'$-UTR) of at least one annotated protein-coding transcript. The rest of protein-coding regions were classified as ORFs, which were further subdivided into exonic, intronic, and alternative regions. A region was classified as constitutive exonic (respectively, intronic) if it belonged to exonic (respectively, intronic) parts of all annotated transcripts that overlap the region; otherwise, it was classified as alternative exonic. Terminal exons of protein-coding transcripts were excluded from the alternative category.

Identification of PAS from RNA-seq data

GTEx RNA-seq data were downloaded from dbGaP (dbGaP project 15872) in fastq format and aligned to the human genome assembly hg19 using STAR aligner version 2.7.3a in paired-end mode (39). PySAM suite was used to extract uniquely mapped reads (NH:1) (40). To identify polyA reads, we considered all reads containing a soft clipped region of at least 6 nts excluding reads with average sequencing quality below 13, which corresponds to the probability 0.05 of calling a wrong base. We required that the reported nucleotide sequence of the clipped region contain at least 80% T's if the soft clip was in the beginning of the read, and 80% A's if the soft clip was in the end of the read. In fact, the requirement of 80% A's or T's was excessively strict since 87% of soft clip regions consisted entirely of A's or T's. Samples that contained an exceptionally high number of polyA reads were excluded from analysis (Supplementary Figure S1). PolyA reads were pooled by the genomic position of the first non-templated nucleotide, referred to as PAS position, resulting in read counts (f_i) for each value of the overhang (i). Accordingly, each PAS was characterized by the number of aligned polyA reads $f = \sum_i f_i$ and Shannon entropy of the overhang distribution $H = -\sum\limits_i p_i \log_2 p_i$, where $p_i = f_i/f$.

To find optimal cutoffs, we repeated the above steps using an array of thresholds on the minimal overhang length and Shannon entropy threshold H and computed the number of annotated gene ends that are supported by PAS (Supplementary Figure S2). The threshold $H \geq 2$ in combination with the minimum overhang length of 6 nts appeared to be optimal since it captured 85% annotated gene ends and yielded 565 387 PAS, a number that corresponds by the order of magnitude to the size of the PAS set reported in PolyASite 2.0 (24). PASs that were located within 10 nts of each other were merged into clusters (PASCs) using the *GenomicRanges* package (41).

Precision and recall

The list of PASCs obtained from the GTEx RNA-seq data (referred to as GTEx) was validated against two reference sets, the published set of PASCs inferred from the 3′-end sequencing (PolyASite 2.0, referred to as Atlas) and the set of annotated TEs provided by GENCODE consortium (referred to as GENCODE). First, GTEx and Atlas were both compared to GENCODE so that a PASC was considered a true positive if it was located within 100 nts from an annotated TE, as in the previous studies (29,31,32). The precision and recall metrics varied depending on the number of supporting polyA reads (in GTEx) and the average expression (in Atlas) reaching the optimal $F_1 = 2(P^{-1} + R^{-1})^{-1}$ score at $P = 0.57 - 0.58$ and $R = 0.49 - 0.51$ (Supplementary Figure S3, top left). The same metrics for PASCs weighted by polyA read support showed a better performance with the optimal F_1 score at $P = 0.83 - 0.86$ and $R = 0.73 - 0.76$ (Supplementary Figure S3, bottom left). In comparison to Atlas as a reference set by the number of PASCs, GTEx showed a moderate performance with $P = 0.66$ and $R = 0.30$, especially in terms of recall, i.e., a large fraction of PASCs from Atlas were not detected (Supplementary Figure S3, top right). However, when the same comparison was made by weighting PASCs by the number of polyA reads, the precision and recall were 0.92 and 0.97, respectively, indicating that the GTEx primarily misses PASCs with low level of read support (Supplementary Figure S3, bottom right).

Relative position in the gene

For each PASC, which is characterized by the interval [x, y] in the gene [a, b], where x, y, a, and b are genomic coordinates on the plus strand, we defined p, the relative position in the gene as $p = \frac{x-a}{(y-x)-(b-a)+1}$ for genes on the positive strand, and used the value of $1 - p$ for genes on the opposite strand. The values of p outside of the interval [0,1] indicate that the PASC is located outside of the annotated gene boundaries. PASC relative positions with respect to exonic and intronic regions were computed similarly.

Read coverage and fold change

To quantify the extent, to which CPA happen at a specific PASC in a specific tissue, we first calculated the read coverage genomewide for each GTEx sample by considering only uniquely mapped reads (MAPQ = 255 when processed via STAR mapper) with *bamCoverage* utility using flags –binSize 10 –minMappingQuality 255 (42) and averaged the read coverage values between samples within each tissue using *wiggletools mean* utility (43).

Next, we calculated the mean read coverage per nucleotide in 150-nt windows starting 10 nts upstream and downstream of each PASC in each tissue (referred to as wi_1 and wi_2) using *multiBigwigSummary* utility (42). The fold change (wi_1/wi_2) metric was computed using a pseudocount of 10^{-3}. To take into account the variation between samples when assessing PASC expression, we followed the approach described previously (11) by detecting significant differences in read counts between the upstream and downstream windows ($P_{\text{adj}} < 10^{-3}$) using DESeq2 (44), separately in each tissue.

Intronic PASCs (iPASCs) were defined as PASCs located within at least one annotated intron of a protein-coding gene >200bp away from the closest annotated splice site. The read coverage in we_1 and we_2 was computed with respect to the shortest intron containing the iPASC. An iPASC located within 100 nts from an annotated TE of a protein-coding transcript ($n = 3188$) was categorized as an annotated STE (respectively, CTE) if the terminal exon of the transcript fully belonged to the containing intron (respectively, contained the interval from the 5′-splice site to iPASC). This categorization yielded 1136 CTEs and 1948 STEs; 104 PASCs located near multiple TEs were excluded due to the conflicting annotation.

To estimate the mean read coverage in constitutive exons, alternative exons, and introns, the total read coverage values per nucleotide in GTEx samples were averaged between windows located in the respective regions, resulting in the normalization factors of 3.3×10^6, 3.2×10^6 and 8.0×10^4, respectively.

Splicing metrics

To quantify tissue-specific alternative splicing associated with intronic PASCs, we computed split read counts using the IPSA pipeline (33,45). The counts of split reads were pooled within each tissue to compute the $\psi = a/(a + b + c)$ metric, where a, b, and c are the number of split reads supporting the canonical splicing, the number of split reads landing before iPASC, and the number of continuous reads spanning the exon-intron boundary, respectively. The values of ψ with the denominator below 30 were discarded as unreliable.

Cleave-seq 5′-end coverage and 3′-RNA capping and pulldown data

Cleave-seq data in HeLa cells were downloaded from Gene Expression Omnibus under the accession number GSE165742 (samples GSM5566266–GSM5566269) in bigwig format (46). The per-bin Cleave-seq signal was computed around 5′-splice sites using deeptools *computeMatrix* tool with the following parameters *reference-point -a 150 -b 20 -bs 5 –nanAfterEnd –missingDataAsZero –skipZeros* and consequently averaged between replicas and introns for visualization.

The 3′-RNA capping and pulldown (3′-PD) data in U2OS cells (47,48) were downloaded from Gene Expression Omnibus under the accession number GSE84068 including three 3′-PD replicas (GSM2226722–GSM2226724) and three total polyA$^+$ RNA-seq replicas for normalization (GSM2226713–GSM2226715). The per-bin coverage around 5′-splice sites was computed as for Cleave-seq. For visualization, the 3′-PD coverage values were averaged between replicates, normalized to the respective total RNA-seq coverage in each bin of each intron, and averaged between introns.

Statistical analysis

The data were analyzed using R statistics software version 3.6.3. One-sided non-parametric tests were performed using normal approximation with continuity correction. In all figures, the significance levels of 5%, 1% and 0.1% are denoted by *, ** and ***, respectively; whiskers denote standard deviation; log denotes base-10 logarithm.

RESULTS

The identification of PAS

The majority of short reads in the output of polyA$^+$ RNA-seq protocols align perfectly to the genome, but a small fraction map partially due to stretches of non-templated adenines generated by CPA. Since RNA-seq reads with incomplete alignment to the genomic reference tend to map to multiple locations, we took a conservative approach by analyzing only uniquely mapped reads from 9021 GTEx RNA-seq experiments (33) with additional restrictions on sequencing quality (see Methods). We extracted polyA reads, defined as short reads containing a soft clipped region of at least six nucleotides that consists of 80% or more adenines, excluding reads aligning to adenine-rich genomic tracks and omitting samples with exceptionally large numbers of polyA reads (Supplementary Figure S1). Out of ∼356 billion uniquely mapped reads, ∼591 million (0.17%) polyA reads were obtained. At that, the average adenine content in soft clipped regions of polyA reads was 98% despite the original 80% threshold, confirming that the selected short reads indeed contain polyA tails.

The alignment of a polyA read is characterized by the genomic position of the first non-templated nucleotide, which presumably corresponds to a PAS, and the length of the soft clip region, here referred to as overhang (Figure 1A). Consequently, each PAS is characterized by the number of supporting polyA reads, referred to as polyA read support, and the distribution of their overhangs. Our confidence in PAS correlates not only with polyA read support, but also with the diversity of the overhang distribution (33), which is measured by Shannon entropy H. Out of 9.6 million candidate PASs, 2.1 million (22%) had $H \geq 1$ and 565 387 (6%) had $H \geq 2$ (Supplementary Figure S2). PASs located near annotated transcript ends tend to have higher H values compared to other PASs (Supplementary Figure S4A). In further analysis, we chose to use the threshold $H \geq 2$ in order to obtain a list of PASs that matches by the order of magnitude the consolidated atlas of polyadenylation sites from 3′-end sequencing (24) and captures sufficiently many annotated gene ends (Supplementary File 1). Out of 565 387 PASs with $H \geq 2$, 331 563 contained a sequence motif similar to the canonical consensus CPA signal (NAUAAA, ANUAAA, or AAUANA) in the 40-nt upstream region (49,50). The latter PASs will be referred to as PASs with a signal.

To characterize the occurrence of PASs in different genomic regions, we subdivided the human genome into a disjoint union of intervals corresponding to protein-coding genes, non-coding genes, and intergenic regions. In total, 336 045, 49 665 and 179 677 PASs were detected in these respective regions; of these 69%, 61%, and 39% were PASs with a signal, respectively. The level of polyA read support in different genomic regions also varied, e.g. 25.5%, 14% and 7% PASs were supported by 100 or more polyA reads in protein-coding, non-coding, and intergenic regions, respectively (Figure 1B). As expected, protein-coding regions had the largest density of PASs per megabase. However, a large absolute number of PASs in intergenic regions, including PASs without canonical consensus CPA signals, indicates that a substantial number of RNA Pol II transcripts are transcribed from them, in accordance with current hypotheses on pervasive transcription (51–53).

An example of a gene that is highly covered by polyA reads is *RPL5* (Figure 1C). We identified several PASs in the vicinity of its annotated transcript end (TE), some of which were supported by as many as 100 000 polyA reads with more than 20 different overhangs. While instead of a single peak we observed a relatively dispersed cluster of PASs spanning twelve nucleotides, the majority of polyA reads supported CPA at only two closely located positions. The 3′-seq read coverage in *RPL5* locus also followed this pattern (Supplementary Figure S4B). Remarkably, the number of polyA reads decayed with increasing the length of the overhang (Figure 1C, bottom). This decrease could result from the mapping bias, in which a lower fraction of reads with larger soft clip regions can be mapped uniquely, or be a consequence of degradation of the substrates possessing multiple terminal adenines by exonucleases (54).

PAS clusters

The variability of PASs positions in *RPL5* motivated us to explore the distribution of distances from each PAS to its closest annotated TE in protein-coding genes (Figure 2A). Among PASs that were located within 100 nts from an annotated TE, 71% fell within 10 nts, and 78% of PASs with a signal did so. We therefore chose to cluster PASs that were located within 10 nts of each other (Figure 2B). This yielded 318 898 PAS clusters (PASCs), of which 90% had length below or equal to 10 nts, 72% consisted of a unique PAS, and 99% consisted of less than 10 individual PASs (Supplementary File 2). In comparison, PASCs derived from the 3′-end sequencing tend to be wider (Supplementary Figure S4C). In what follows, a PASC will be referred to as PASC with a signal if it contains at least one individual PAS with a signal; the polyA read support of a PASC is defined as the total number of supporting polyA reads of its constituent individual PASs.

We next asked how PASCs identified from GTEx RNA-seq data correspond to those in the consolidated polyadenylation atlas (PolyASite 2.0 (24), in what follows referred to as Atlas) and TEs annotated by the GENCODE consortium (38). To assess this, we surrounded TEs from GENCODE by 100-nt windows and analyzed pairwise intersections of the three respective sets (Figure 2C). The precision of GTEx with respect to GENCODE, i.e., the proportion of PASCs from GTEx that were located within 100 nts of an annotated TE, was higher than that of PolyASite 2.0, while the recall, i.e., the proportion of annotated TEs that are supported by at least one PASC from GTEx within 100 nts, was lower. Conversely, the precision of GTEx with respect to PolyASite 2.0 was lower compared to that of GENCODE, while the recall was higher. A similar interplay between

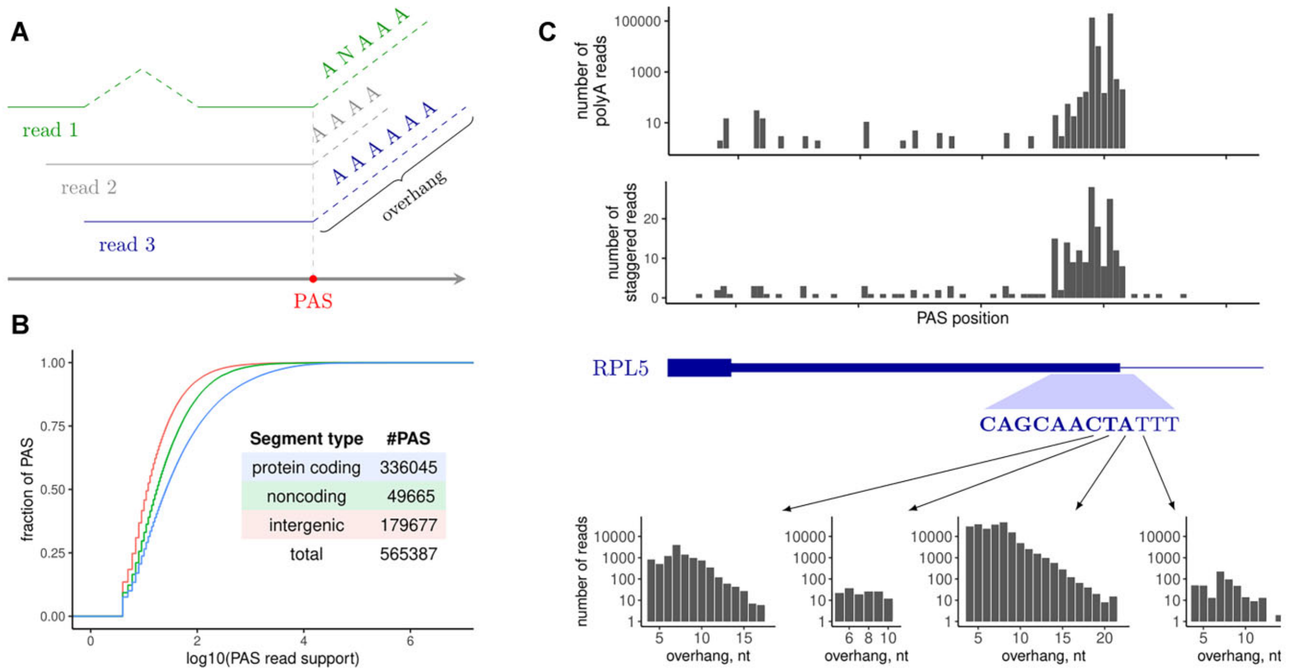

Figure 1. The identification of PAS. (**A**) The alignments of short reads with non-templated adenine-rich ends (polyA reads). The genomic position of the first non-templated nucleotide corresponds to a PAS. The length of the soft clip region is referred to as overhang. (**B**) The polyA read support of PAS in protein-coding genes, non-coding genes, and intergenic regions. The number of PASs in each group is indicated in the inset. (**C**) The 3′-end of the *RPL5* gene is highly covered by polyA reads. Top: the positional distribution of the number of polyA reads (in log scale) and the number of staggered polyA reads (i.e. the number of different overhangs). Bottom: the distribution of overhangs at the indicated positions (in log scale).

precision and recall values was observed when shortening the window around TEs to 50 nts and also for a subset of PASCs located intronically (Supplementary Figure S5). This comparison indicates that GTEx RNA-seq data yields a slightly more conservative set of PASCs than PolyASite 2.0. The benefit of using GTEx PASCs is that RNA-seq provides a snapshot of alternative splicing and polyadenylation assessed in the same conditions. Additional analysis of the relationship between precision and recall for GTEx and PolyASite 2.0 weighted by the polyA read support confirmed that the two sets are largely consistent (Supplementary Figure S3).

Since 85% of newly identified PASCs did not have an annotated TE within 100 nts, we focused on this group of PASCs (referred to as unannotated PASCs) and explored their relative position within the gene length, which is equal to 0% and 100% for the 5′-end and 3′-end of the gene, respectively (Figure 2D). Despite TEs no longer being considered, we observed a considerable increase in PASC density towards the 3′-end for those with and without a signal, and a much weaker, but noticeable increase in the 5′-end. This recapitulates the general tendency of PASCs to occur more frequently towards the 3′-end of the gene, a pattern that is also observed for unannotated PASCs from Atlas (Supplementary Figure S6). Of note, 89% of PASCs documented in Atlas also did not have an annotated TE within 100 nts, thus raising a concern about the biological relevance of these unannotated PASCs and their role in premature transcription termination.

PAS clusters in protein-coding regions

We next focused on a subset of 164 497 PASCs that were located in protein-coding genes and explored their attribution to gene parts, namely to the 5′-untranslated region (5′-UTR), the 3′-untranslated region (3′-UTR), and the coding part (ORF). Each ORF region was further subdivided into intronic, constitutive exonic , and alternative exonic parts (see Methods). Since these regions differ by length, we quantified PASCs not only by absolute number but also by density, i.e., the number of PASCs per nucleotide. Additionally, we quantified the expression of PASCs by taking into account their polyA read support, in which each PASC was weighted by the number of supporting polyA reads (Figure 3).

As expected, PASCs were quite frequent in 3’UTRs and ORF by absolute number, but their density was the highest in 3′-UTRs since ORF regions are also longer than UTRs (Figure 3A). The enrichment in 3′-UTRs was more prominent when taking into account the number of supporting polyA reads. Similarly, PASCs were most frequent in introns by absolute number, but their density was the lowest after normalization (Figure 3B). The positional distribution of PASCs had a pronounced peak in the end of exonic regions and in the beginning of intronic regions (Supplementary Figure S7), and similar peaks were also observed for PolyASite 2.0 (Supplementary Figure S8). However, despite low density, intronic PASCs were still quite frequent by absolute number, and among them there could be PASCs leading to premature CPA.

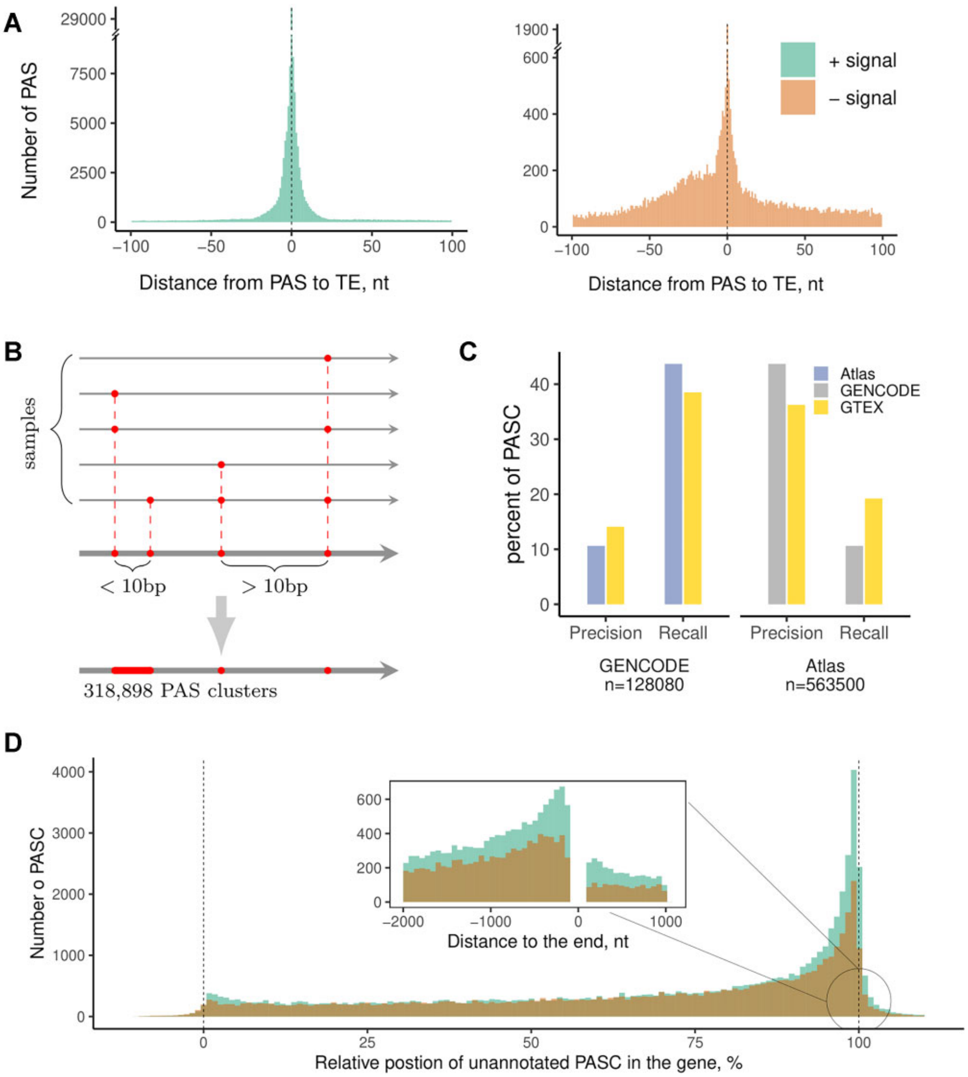

Figure 2. PAS clusters in protein-coding genes. (**A**) The distribution of distances from each PAS to its closest annotated transcript end (TE) for PAS with (n = 122 448) and without a signal (n = 22 361). (**B**) PAS located <10 bp from each other are merged into PAS clusters (PASCs). (**C**) Pairwise comparison of PASs inferred from GTEx, PolyASite 2.0 (Atlas), and GENCODE. Left: the proportion of PASC from GENCODE that are supported by Atlas or GTEx (precision) and the proportion of PASC from Atlas or GTEx that are supported by GENCODE (recall). Right: the proportion of PASC from Atlas that are supported by GENCODE or GTEx (precision) and the proportion of PASC from GENCODE or GTEx that are supported by Atlas (recall). (**D**) The relative positions of unannotated PASCs (i.e., ones not within 100 bp of any annotated TE) along the gene length. 0% and 100% correspond to the 5′-end and 3′-end of the gene, respectively. The inset shows distribution of absolute positions of unannotated PASCs around the gene end.

Current models assume that introns containing PASs cannot undergo splicing after they are cleaved and polyadenylated (1). Here we challenge this assumption by supposing that splicing and CPA machineries can operate on the same pre-mRNA simultaneously, and that the spliceosome, once assembled on the intron, is able to complete intron excision even after CPA has already occurred in it. In this case, some of the intronic CPA events would still be visible in RNA-seq as intronic polyA reads despite intron removal. The extent, to which it happens, may depend on intron debranching and degradation rates as well as on other intron-specific factors such as RNA secondary structure or G-quadruplex formation (55,56).

To estimate the CPA rate, at which it acts on the nascent pre-mRNA, and to take into account the bias arising from intron degradation, we normalized the number of polyA reads to the average read coverage in exons and introns and found that the relative density of polyA reads in introns is substantially larger than that in exons (Figure 3C). Furthermore, we matched introns, constitutive, and alternative exons by the read coverage (Supplementary Figure S9A) and selected a subset of intervals of each type that were covered by approximately the same number of reads (133 ± 6.7 reads per kb per sample). Then, we computed the number of polyA reads in these matched subsets and, again, found a prominent enrichment of polyA reads in introns as compared to exons both in terms of the number of polyA reads (Figure 3D, left) and their density per nt (Figure 3D, right). This enrichment remained significant in other read coverage ranges (Supplementary Figure S9B, C). In sum, this

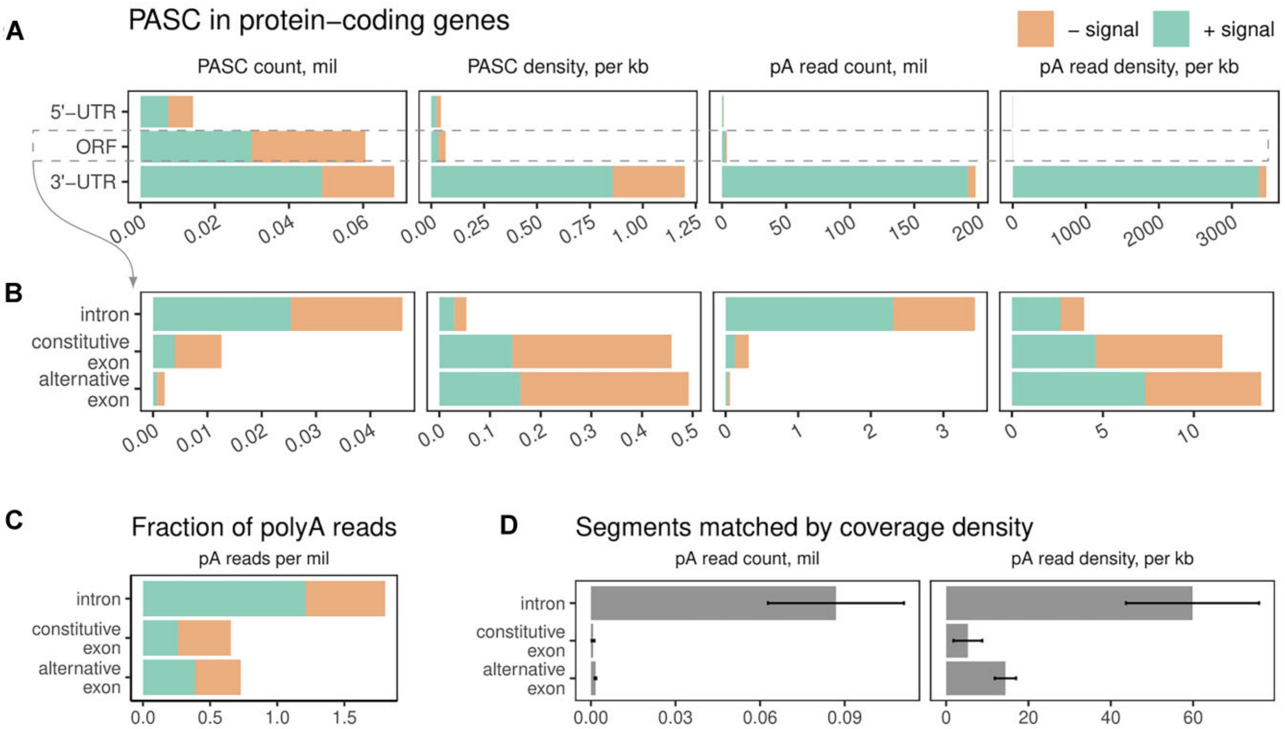

Figure 3. PAS clusters in protein-coding regions. (**A**) The distribution of PASCs in 5′-UTRs, ORF, and 3′-UTRs. Shown are the total number of PASC (PASC count), PASC density per Kb (PASC density), the total number of polyA reads (polyA read count), the total number of polyA reads per kb (polyA read density). (**B**) The distribution of PASCs from ORF in introns, constitutive exons, and alternative exons. PASC located within 2bp of exon borders were excluded. (**C**) The number of polyA reads normalized to the average read coverage in each region (defined as the number of polyA reads per million aligned reads; see Methods for details). (**D**) The number of polyA reads in segments matched by the read coverage density. Whiskers denote standard deviation.

indicates that if introns and exons were equally represented in the RNA-seq data, the frequency of CPA events in introns would have appeared several times larger than that in exons.

Tissue-specific polyadenylation

While PASC positions can be robustly identified by pooling hundreds of millions of polyA reads across the entire GTEx dataset, the rate of their tissue-specific usage cannot be assessed in the same way due to insufficient number of polyA reads in individual samples. Instead, the rate of PASC expression in tissues can be measured by coverage-based methods, as their positions have been already identified. Here, we adapted a simple procedure from (11), in which the average read coverage was measured in 150-nt windows, wi_1 and wi_2, before and after each PASC. To quantify PASC expression, we used $\log_{10}(wi_1/wi_2)$ metric, which captures the magnitude of read coverage drop at a PASC, and a more elaborate method based on DESeq2 (44), which additionally accounts for variation between samples (Figure 4A).

First, we analyzed the set of 164 497 PASCs in protein-coding genes by pooling read coverage profiles across all GTEx samples and excluding PASCs located within 200 nts from splice sites to avoid measuring the read coverage drop at exon-intron boundaries. In the resulting set of 126 310 PASCs (Supplementary File 3), the read density in wi_1 and wi_2 averaged to 8.8 and 3.7 reads per nucleotide per sample, respectively, indicating at least twofold average drop after PASCs. Consistently, the wi_1/wi_2 distribution was skewed towards positive values with a noticeably bigger skewness for PASCs with a signal and PACSs near annotated TEs (Figure 4B). Remarkably, the number of supporting polyA reads was positively correlated with wi_1/wi_2 not only for PASCs near annotated TEs, but also for unannotated PASCs with a signal (Figure 4C).

For each PASC, we computed the average read density in wi_1 and wi_2 separately in each tissue. Out of 126 310 PASCs, on average 18 470 per tissue (15%) had $wi_1/wi_2 > 10$, while DESeq2 analysis has identified a significant difference between read coverage in wi_1 and wi_2 for on average 43 615 (35%) of PASCs per tissue. In each tissue, on average 90% of PASCs with $wi_1/wi_2 > 10$ were also significant according to DESeq results. Since the results of the two methods overlapped, we chose to call a PASC with $wi_1/wi_2 > 10$ as expressed in the corresponding tissue.

We next compared the set of expressed PASCs to a reference set containing 689 346 PASs in 3′-UTRs of human genes that was derived from the GTEx using DaPars (34). Since the exact positions of PASCs in 3′-UTRs may vary, we selected 3′-UTRs that contain at least one expressed PASC and matched them against 3′-UTRs that were called as expressed by DaPars in genes with more than one annotated 3′-UTR. On average 85% of 3′-UTRs containing an expressed PASC were also called as expressed by DaPars, and vice versa 50% of 3′-UTRs called as expressed by DaPars contained at least one expressed PASC. That is, the

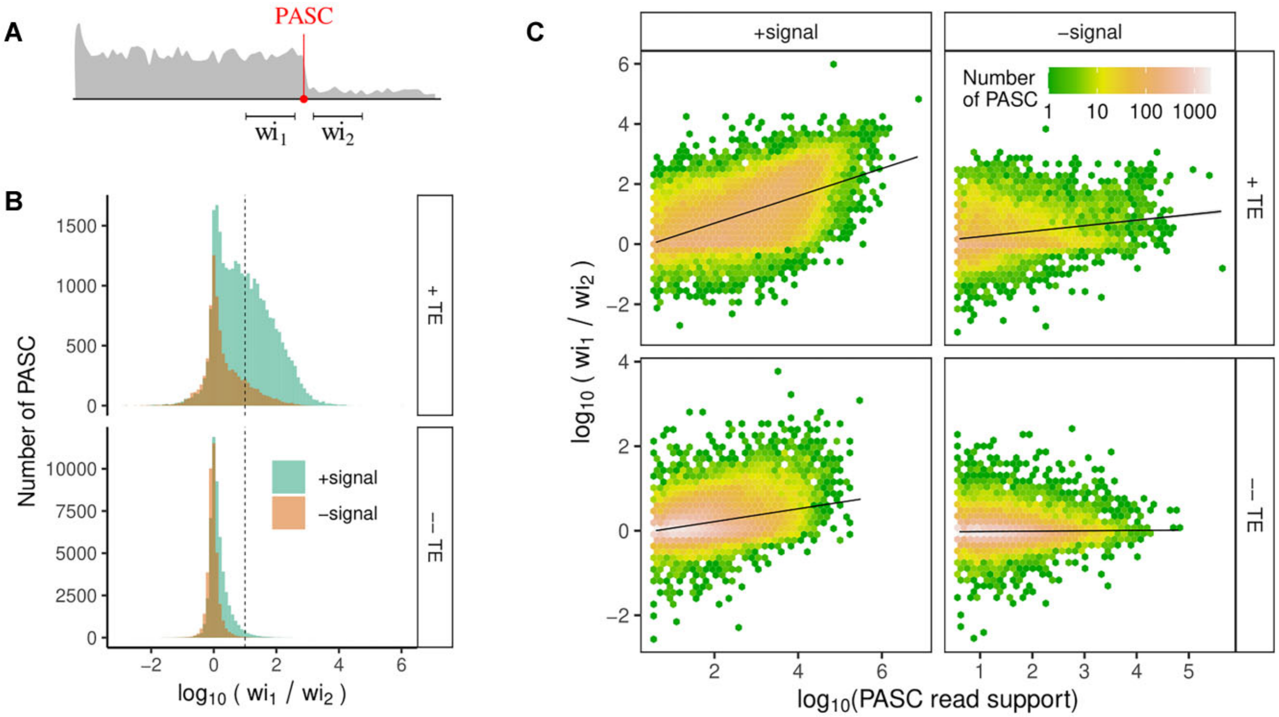

Figure 4. Coverage-based metrics of PASC expression. (**A**) The average read coverage was measured in 150-nt upstream and downstream windows, wi_1 and wi_2, around PASC. (**B**) The distribution of $\log_{10}(wi_1/wi_2)$ metric for annotated ($n = 37\ 194$, top) and unannotated PASCs ($n = 89\ 116$, bottom). A PASC is referred to as annotated if it is within 100 bp of an annotated TE. The dashed line represents the cutoff $wi_1/wi_2 = 10$. (**C**) The $\log_{10}(wi_1/wi_2)$ metric positively correlates with the number of supporting polyA reads not only for annotated, but also for unannotated PASCs with a signal.

expression of PASCs in tissues as measured by the wi_1/wi_2 metric and the results obtained by DaPars are consistent on a subset of PASCs in 3′-UTRs.

Previous studies have extensively characterized tissue-specific polyadenylation in the GTEx dataset using coverage-based methods, however focusing on polyadenylation in 3′-UTRs (31,34–36). Here, we specifically considered intronic PASCs (iPASCs) identified by using polyA reads and examined the relationship between IPA and AS by juxtaposing the information on polyA reads to identify PASC positions, metrics based on split reads to measure AS rate, and the read coverage to assess IPA rate.

Intronic polyadenylation and splicing

According to (9), alternative terminal exons that are generated through IPA can be categorized into two classes, skipped terminal exons (STE), which may be used as terminal exons or excluded, and composite terminal exons (CTE), which result from CPA in a retained intron (Figure 5A, right). To distinguish between these possibilities, we estimated the average read coverage in two additional windows, we_1 and we_2, at the exon-intron boundary (Figure 5A, left). For simplicity, the read coverage values in the four windows will also be denoted by we_1, we_2, wi_1 and wi_2. We expect that, in addition to a large wi_1/wi_2 ratio, STE must be characterized by a large we_1/we_2 ratio, while CTE must be characterized by a small we_1/we_2 ratio.

To quantify the rate of splicing, we computed the number of split reads starting at the intron 5′-end and landing before iPASC (b), after iPASC at the canonical 3′-splice site (a), and the number of continuous reads (c) that span the exon-intron boundary (Figure 5A, left). These metrics were combined into the $\psi = a/(a + b + c)$ ratio, referred to as the rate of canonical splicing, where $\psi \simeq 1$ indicates that the canonical splicing (a) prevails, while $\psi \simeq 0$ indicates the presence of AS events before iPASC. We expect that both STE and CTE are characterized by $\psi \simeq 0$ due to the lack of canonical splicing, with prevailing b in the case of STE and prevailing c in the case of CTE.

In what follows, we confined the analysis to iPASCs with a signal only. The values of we_1, we_2, wi_1, wi_2, and ψ were computed for 1,115,690 iPASC-tissue pairs comprising 35 990 iPASCs in 31 tissues. We observed a significant negative association between ψ and IPA rate measured by polyA read support (Figure 5B) or $\log_{10}(wi_1/wi_2)$ (Supplementary Figure S10A). Of note, ψ is a relative quantity, which is not influenced by the read coverage. This association also manifested itself as a negative skew in the distribution of Pearson correlation coefficients of ψ and IPA rate across tissues as compared to the background distribution, in which the tissue labels were shuffled (Figure 5C, left, and Supplementary Figure S10B). The read coverage at iPASC changed two orders of magnitude when ψ increased from 25% to 100% in some remarkable cases (Figure 5C, right). These observations reconfirm that splicing and CPA naturally counteract each other.

Further, we considered 75 501 iPASC-tissue pairs with a substantial read coverage drop at iPASC ($wi_1/wi_2 > 10$) and a substantially high read coverage in the upstream intronic window ($wi_1 > 0.1we_1$). The bivariate distributions of $\log(we_1)$ and $\log(we_2)$ for 1136 annotated CTEs and

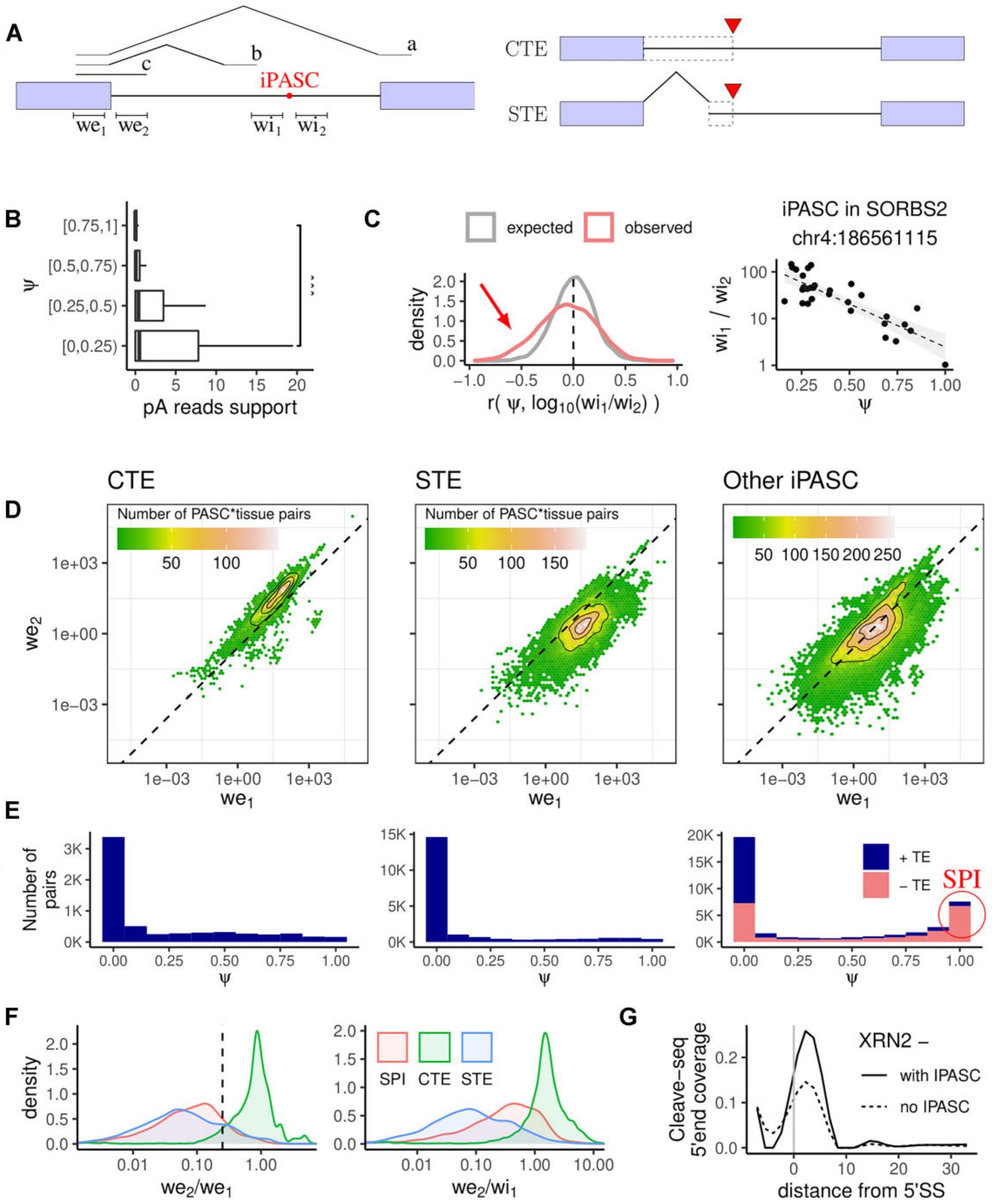

Figure 5. Intronic polyadenylation and splicing. (**A**) Exonic (we_1 and we_2) and intronic (wi_1 and wi_2) 150-nt windows. (**B**) The polyA read support of iPASCs in four ψ quartiles; *** denotes the 0.1% significance level. (**C**) Pearson correlation coefficients of ψ and $\log_{10}(wi_1/wi_2)$ for $n = 12\,261$ iPASCs compared to the label-shuffled control (left). Negative association between ψ and $\log_{10}(wi_1/wi_2)$ in the *SORBS2* gene. (**D**) Bivariate distribution of we_1 versus we_2 in PASC-tissue pairs for CTE ($n = 1136$), STE ($n = 1948$), and other iPASCs ($n = 32\,906$). The dashed line corresponds to $we_2/we_1 = 0.25$. (**E**) The distribution of ψ for CTE, STE, and other iPASCs; +TE (−TE) denote iPASCs within (not within) 100 nts of an annotated TE. (**F**) The distribution of we_2/we_1 (left) and we_2/we_1 (right) values for CTE, STE and SPI. The vertical dashed line denotes $we_2/we_1 = 0.25$. (**G**) The Cleave-seq $5'$-end coverage in introns with ($n = 21\,230$) iPASC and without iPASC ($n = 199\,978$) under *XRN2* knockdown (see Supplementary Figure S10D for the wild type).

1948 annotated STEs were separated by the line $we_2 = 0.25we_1$, with the former expectedly clustering above, and the latter clustering below the line (Figure 5D, left and middle). iPASCs other than CTE or STE formed a mixture of the two distributions (Figure 5D, right). A similar pattern was observed for the bivariate distributions of $\log(wi_1)$ and $\log(we_2)$ (Supplementary Figure S10C). However, while ψ values of CTE and STE were characterized by a single peak at $\psi \simeq 0$ indicating the absence of canonical splicing (Figure 5E, left and middle), the ψ values of iPASC other than CTE or STE had a pronounced second peak at $\psi \simeq 1$ formed mostly by iPASCs without the TE support (Figure 5E, right). This peak is incompatible with CTE and STE models because it implies that IPA coexists with the canonical splicing. To further clarify this, we focused on introns substantially supported by split-reads ($a + b + c \geq 30$) and containing iPASCs with $\psi > 0.9$, termed here as Spliced Polyadenylated Introns (SPI), and compared we_2/we_1 and we_2/wi_1 distributions among STE, CTE, and SPI (Figure 5F).

Similarly to STEs, SPIs were characterized by a low coverage in the intron 5′-end relative to the exon, yet a sufficiently high coverage upstream of iPASC relative to the intron 5′-end. We hypothesized that iPASCs with $\psi \simeq 1$ represent prematurely polyadenylated and spliced introns and hence expected them to have a monophosphate at the 5′-end (5′-p) resulting from the branchpoint (BP) cleavage by RNA debranching enzyme *DBR1* (57,58). Then, the linearized product of CPA at an iPASC upstream of BP would consist of two separate molecules, one corresponding to the intronic RNA upstream of PAS with both 5′-p and polyA tail, and the other corresponding to the intron part downstream of PAS. Consistently with this, the 5′-end coverage of RNAs identified by Cleave-seq, a method designed to capture 3′-polyadenylated RNAs with 5′-p (46), was substantially larger in introns with iPASCs than in introns without iPASCs (Figure 5G) and, among the former, it was the largest in SPI (Supplementary Figure S10D). A similar enrichment in the 5′-end was also observed in 3′-pull down *in vitro* capping experiments (Supplementary Figure S10E–G). Taken together, these results indicate that SPIs undergo both splicing and CPA and are not 3′-ends of distinct Pol II transcripts initiated and terminated within the same intron.

Next, we followed up a few cases of tissue-specific splicing and CPA (Figure 6). The iPASC in the *MEGF8* gene, which encodes a membrane protein associated with Carpenter syndrome (59), is an example of a CTE supported by intronic read coverage in absence of AS before PASC, most remarkably in thyroid tissue (Figure 6A). In the Attractin (*ATRN*) gene, which encodes a transmembrane protein associated with kidney and liver abnormalities in mice (60), an iPASC is expressed in muscle along with the elevation of read coverage in wi_1 and activation of a splice site at its border, likely representing an unannotated STE (Figure 6B). Both these iPASCs are supported by *CSTF2* eCLIP peaks and PolyASite 2.0. In contrast, iPASC in the *ATRX* gene, which encodes a chromatin remodeler linked to a range of diseases (61), exhibits elevated read coverage in wi_1, but it lacks AS events that could support STE, or RNA-seq reads in the beginning of the intron that could support CTE (Figure 6C). The only possible explanation for it would be that the canonical splicing and IPA co-exist and operate concurrently resulting in SPI.

To characterize further the abundance of IPA events, we considered a strict set of iPASC-tissue pairs described above and categorized them as CTE, STE, and SPI according to the following criteria: $\psi \leq 0.9$ and $we_2 > 0.25we_1$ (CTE), $\psi \leq 0.9$ and $we_2 \leq 0.25we_1$ (STE), and $\psi > 0.9$ (SPI), respectively (Supplementary File 4). We categorized an iPASC as CTE, STE, and SPI if it belonged to the respective class for at least one iPASC-tissue pair. This yielded 2846, 2251 and 1482 iPASCs corresponding to STE, CTE and SPI, respectively, with 63% of SPIs also supported by PolyASite 2.0 and >75% of SPIs having more than 200 reads in the ψ denominator. The number of iPASCs attributed to the three classes varied moderately across tissues, presumably reflecting the fact that the bulk of IPA events are not regulated (Supplementary Figure S11A). Accordingly, SPIs exhibit the lowest variation of the relative expression measured by the wi_1/we_1 ratio among the three classes (Supplementary Figure S11B), tend to occur in longer introns, and have a slight preference to the 5′-end of the gene (Supplementary Figure S11C, D). Furthermore, approximately 13% of genes expressing more than one iPASC contain a SPI. We conclude that SPIs represent semi-stable intermediates that can be detected by polyA reads and 3′-end sequencing. They constitute a minor, yet considerable fraction of IPA events and contribute to the observed landscape of intronic polyadenylation.

DISCUSSION

Thousands of recurrent and dynamically changing IPA events have been identified by 3′-end sequencing methods, but the matched data to study the interplay between IPA and AS in the same biological condition are currently in high demand (11). The GTEx dataset represents an ideal resource for studying this interplay because the information on the positions and tissue-specific expression of intronic PASs, which is captured by polyA reads, is complemented by tissue-specific splicing rates inferred from split reads that align to splice junctions.

In this work, we used for the first time the approach based on polyA reads, one that was applied previously to much smaller datasets (29,30), for the identification of PASs at the scale of GTEx project and combined it with a coverage-based method to assess the IPA rate. While RNA-seq is known to have a limited sensitivity when detecting PASs due to the short read length, the magnitude of the GTEx dataset allows for a dramatic improvement, making the results comparable to those of PolyASite 2.0. However, the polyA-read-based approach also has limitations related to the mappability of reads with long soft clip regions. The positional distribution of PASCs in constitutive exons and introns has a pronounced peak in the end of exonic and in the beginning of intronic regions (Supplementary Figure S7) resembling clusters of CAGE tags near internal exons and occurrence of polyA-seq peaks close to exon boundaries (62,63). These anomalies likely arise from erroneous mappings of split reads that contain the polyA tail, e.g. when the adenine-rich part of the read or a short segment between splice junction and the stretch of non-templated adenines are incorrectly attributed to the soft clip region (example in

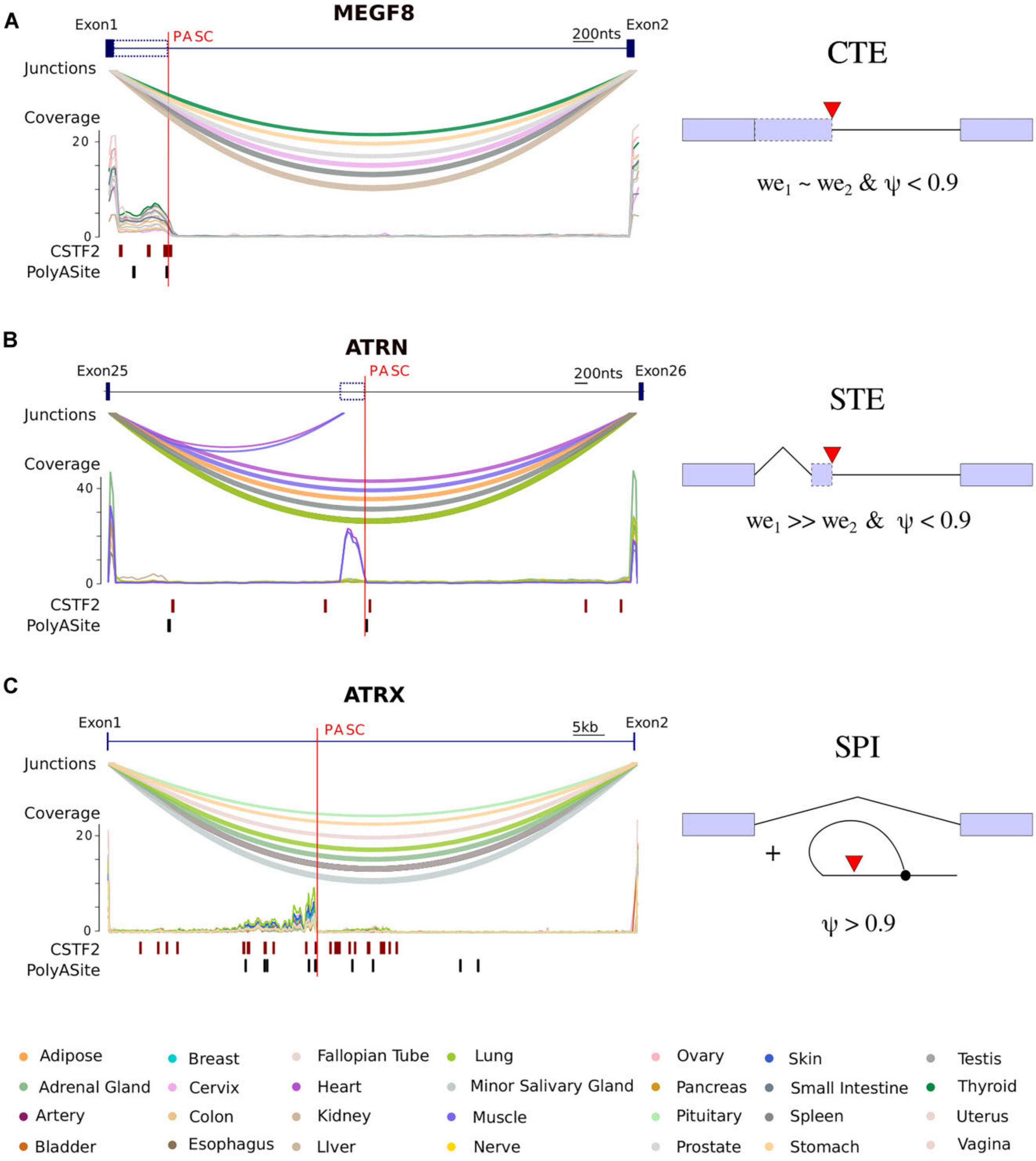

Figure 6. Case studies. (**A**) The iPASC between exons 1 and 2 of *MEGF8* generates a CTE. The eCLIP peaks of *CSTF2* and PASC from PolyAsite 2.0 are indicated in the track below. Arcs represent tissue-specific AS. (**B**) The iPASC between exons 25 and 26 *ATRN* generates a STE with tissue-specific expression in heart and muscle. (**C**) The iPASC between exons 1 and 2 likely generates a SPI because the intron 5′-end is not covered, wi_1 is covered, but there is no evidence of STE.

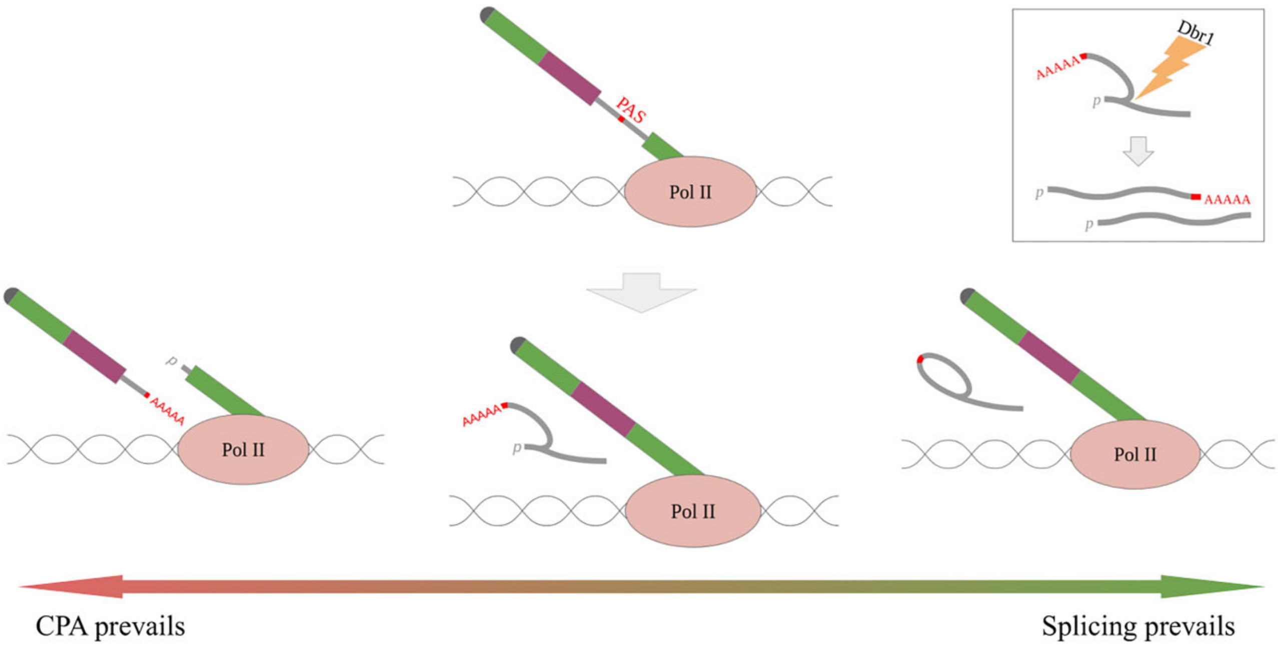

Figure 7. Spliced Polyadenylated Intron (SPI). When the CPA rate exceeds the splicing rate, IPA leads to the generation of a truncated transcript isoform (left). When the splicing rate exceeds the CPA rate, the intron is spliced out and PAS is degraded as a part of a lariat (right). When the CPA and splicing machinery operate at the same rate, the intron is cleaved and polyadenylated while it is being spliced (middle) resulting in SPI. The lariat is debranched producing two separate RNAs (inset) corresponding to intron fragments upstream and downstream of PAS, where the upstream part contains both $5'$-p and polyA tail.

Supplementary Figure S12). However, these details do not invalidate the polyA read strategy since PASCs obtained by other protocols, e.g. in PolyASite 2.0, have similar peaks near exon boundaries (Supplementary Figure S8). The alignment of split reads with a short exonic part appears to be a common problem of all methods.

The widespread nature of IPA has been appreciated recently with the development of $3'$-end sequencing (64). Functionally important IPA cases have been described in specific genes (10,65–69), however most transcripts harboring incomplete reading frames translate into potentially deleterious, truncated proteins that may pose a hazard to the cell (70). In eukaryotes, they usually lack normal termination codons and are rapidly degraded via nonsense mediated decay or nonstop decay pathways (71,72). We found that the majority of polyA reads align to $3'$-UTRs, but a sizable fraction (5–8%) still map to the coding part. Intriguingly, PASs within the coding part appear to be more frequent in introns than in exons, which is partly explained by the higher GC content and stronger evolutionary constraints against generating the canonical AATAAA consensus sequence in exons. However, a remarkably large number of intronic PASs raises concerns about their implication in premature transcription termination (73) and hints at the existence of a mechanism that counteracts their activity. How could it be that 87% of human protein-coding transcripts contain an intronic PAS, but cells are still able to produce full-length transcripts?

Here, we argue that a sizable fraction of intronic PASs observed in polyA read analysis (and also in $3'$-end sequencing) represent SPIs, intermediates that are generated by the spliceosome and the CPA machinery operating concurrently with each other and with the elongating transcription (Figure 7). If CPA occurs first, then it will lead to the generation of a truncated transcript with CTE. If splicing happens first, then the intron containing PAS will be spliced out, and PAS will be degraded as a part of the lariat. However, if PAS-mediated cleavage in the intron starts after the spliceosome has assembled on it and is committed to splicing, then the second catalytic step of the splicing reaction will remove the lariat and all CPA products within it, resulting in SPI (Figure 7 middle). Consequently, SPIs are intronic RNAs spanning from the $5'$-splice site to PAS that contain both $5'$-p due to lariat debranching and the polyA tail. They must be degraded from the $5'$-end by cellular exonucleases, as evidenced in many cases as a characteristic noisy ramp in the read coverage that gradually increases from the $5'$-splice site to PAS (Figure 6C). Nonetheless, a fraction of SPIs are visible through polyA reads due to the presence of the polyA tail. Our conservative estimate is that they constitute almost a quarter of all IPA events, and two thirds of them are supported by PolyASite 2.0. This suggests that $3'$-end sequencing methods may overestimate the rate of IPA, and that their results require careful interpretation.

The enrichment of PASs in introns and the existence of SPIs together suggest that cotranscriptional pre-mRNA splicing may have a possible side function of rescuing eukaryotic transcripts from premature transcription termination. This hypothesis challenges the assumption that when an intronic PAS is used, the surrounding intron can no longer be spliced. The spliceosome that is committed to splicing still can remove the intron that is being cleaved and polyadenylated, thus functioning as rescue. Temporal and spatial interactions of splicing and CPA are orchestrated by a multitude of factors playing dual roles, which recognize signals that are located in the nascent pre-mRNA and bind the same substrates at the same time (14,74,75). It is therefore not impossible that evolution allowed for the generation of dispensable intronic PASs, which are spliced out co-transcriptionally and manifest themselves as SPIs in both RNA-seq and the 3′-end sequencing data. Whether or not SPIs are functional on their own remains a matter of further investigation.

CONCLUSION

Massive amounts of RNA-seq data in the GTEx dataset offered a unique possibility to analyze tissue-specific splicing and polyadenylation. The observed patterns of intronic polyadenylation and splicing reconfirm that splicing and polyadenylation are two inseparable parts of one consolidated pre-mRNA processing machinery, leading to the conjecture that co-transcriptional splicing is a natural mechanism of suppression of premature transcription termination.

DATA AVAILABILITY

The datasets generated during the current study are available online at https://zenodo.org/record/7799648. The source code used for the analysis is available at https://github.com/mashlosenok/RNAseq_PAS_finder (permanent DOI https://doi.org/10.5281/zenodo.7940543).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors thank Vera Rybko, Dmitry Skvortsov, and Olga Donstova for insightful discussions on molecular mechanisms of splicing and polyadenylation.

FUNDING

Russian Science Foundation grant No. [22-14-00330].
Conflict of interest statement. None declared.

REFERENCES

1. Tian,B. and Manley,J.L. (2017) Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.*, **18**, 18–30.
2. 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.
3. 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.
4. Elkon,R., Ugalde,A.P. and Agami,R. (2013) Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.*, **14**, 496–506.
5. Mayr,C. (2019) What are 3’ UTRs doing?. *Cold Spring Harb. Perspect. Biol.*, **11**, a034728.
6. Curinha,A., Oliveira Braz,S., Pereira-Castro,I., Cruz,A. and Moreira,A. (2014) Implications of polyadenylation in health and disease. *Nucleus*, **5**, 508–519.
7. Fang,Z. and Li,S. (2021) Alternative polyadenylation-associated loci interpret human traits and diseases. *Trends Genet.*, **37**, 773–775.
8. Gruber,A.J. and Zavolan,M. (2019) Alternative cleavage and polyadenylation in health and disease. *Nat. Rev. Genet.*, **20**, 599–614.
9. Tian,B., Pan,Z. and Lee,J.Y. (2007) Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res.*, **17**, 156–165.
10. Di Giammartino,D.C., Nishida,K. and Manley,J.L. (2011) Mechanisms and consequences of alternative polyadenylation. *Mol. Cell*, **43**, 853–866.
11. Lee,S.H., Singh,I., Tisdale,S., Abdel-Wahab,O., Leslie,C.S. and Mayr,C. (2018) Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature*, **561**, 127–131.
12. Rakheja,D., Chen,K.S., Liu,Y., Shukla,A.A., Schmid,V., Chang,T.C., Khokhar,S., Wickiser,J.E., Karandikar,N.J., Malter,J.S. *et al.* (2014) Somatic mutations in DROSHA and DICER1 impair microRNA biogenesis through distinct mechanisms in Wilms tumours. *Nat. Commun.*, **2**, 4802.
13. Proudfoot,N.J., Furger,A. and Dye,M.J. (2002) Integrating mRNA processing with transcription. *Cell*, **108**, 501–512.
14. Kyburz,A., Friedlein,A., Langen,H. and Keller,W. (2006) Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3’ end processing and splicing. *Mol. Cell*, **23**, 195–205.
15. Castelo-Branco,P., Furger,A., Wollerton,M., Smith,C., Moreira,A. and Proudfoot,N. (2004) Polypyrimidine tract binding protein modulates efficiency of polyadenylation. *Mol. Cell Biol.*, **24**, 4174–4183.
16. Dai,W., Zhang,G. and Makeyev,E.V. (2012) RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage. *Nucleic Acids Res.*, **40**, 787–800.
17. Li,Q.Q., Liu,Z., Lu,W. and Liu,M. (2017) Interplay between alternative splicing and alternative polyadenylation defines the expression outcome of the plant unique OXIDATIVE TOLERANT-6 gene. *Sci. Rep.*, **7**, 2052.
18. Kaida,D., Berg,M.G., Younis,I., Kasim,M., Singh,L.N., Wan,L. and Dreyfuss,G. (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, **468**, 664–668.
19. Chen,W., Jia,Q., Song,Y., Fu,H., Wei,G. and Ni,T. (2017) Alternative polyadenylation: methods, findings, and impacts. *Genomics Proteomics Bioinformatics*, **15**, 287–300.
20. Yu,F., Zhang,Y., Cheng,C., Wang,W., Zhou,Z., Rang,W., Yu,H., Wei,Y., Wu,Q. and Zhang,Y. (2020) Poly(A)-seq: a method for direct sequencing and analysis of the transcriptomic poly(A)-tails. *PLoS One*, **15**, e0234696.
21. Shepard,P.J., Choi,E.A., Lu,J., Flanagan,L.A., Hertel,K.J. and Shi,Y. (2011) Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, **17**, 761–772.
22. Lianoglou,S., Garg,V., Yang,J.L., Leslie,C.S. and Mayr,C. (2013) Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.*, **27**, 2380–2396.
23. Zheng,D., Liu,X. and Tian,B. (2016) 3’READS+, a sensitive and accurate method for 3’ end sequencing of polyadenylated RNA. *RNA*, **22**, 1631–1639.
24. Herrmann,C.J., Schmidt,R., Kanitz,A., Artimo,P., Gruber,A.J. and Zavolan,M. (2020) PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3’ end sequencing. *Nucleic Acids Res.*, **48**, D174–D179.
25. 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**, D315–D319.
26. You,L., Wu,J., Feng,Y., Fu,Y., Guo,Y., Long,L., Zhang,H., Luan,Y., Tian,P., Chen,L. *et al.* (2015) APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucleic Acids Res.*, **43**, 59–67.
27. 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.
28. Wang,W., Wei,Z. and Li,H. (2014) A change-point model for identifying 3’UTR switching by next-generation RNA sequencing. *Bioinformatics*, **30**, 2162–2170.

29. Birol,I., Raymond,A., Chiu,R., Nip,K.M., Jackman,S.D., Kreitzman,M., Docking,T.R., Ennis,C.A., Robertson,A.G. and Karsan,A. (2015) Kleat: cleavage site analysis of transcriptomes. *Pac. Symp. Biocomput.*, **1**, 347–358.
30. Bonfert,T. and Friedel,C.C. (2017) Prediction of Poly(A) Sites by Poly(A) Read Mapping. *PLoS One*, **12**, e0170914.
31. Cass,A.A. and Xiao,X. (2019) mountainClimber identifies alternative transcription start and polyadenylation sites in RNA-seq. *Cell Syst.*, **9**, 393–400.
32. Zhao,Z., Xu,Q., Wei,R., Wang,W., Ding,D., Yang,Y., Yao,J., Zhang,L., Hu,Y.Q., Wei,G.. *et al.* (2021) Cancer-associated dynamics and potential regulators of intronic polyadenylation revealed by IPAFinder using standard RNA-seq data. *Genome Res.*, **31**, 2095–2106.
33. Melé,M., Ferreira,P.G., Reverter,F., DeLuca,D.S., Monlong,J., Sammeth,M., Young,T.R., Goldmann,J.M., Pervouchine,D.D., Sullivan,T.J. *et al.* (2015) Human genomics. The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.
34. Hong,W., Ruan,H., Zhang,Z., Ye,Y., Liu,Y., Li,S., Jing,Y., Zhang,H., Diao,L., Liang,H. *et al.* (2020) APAatlas: decoding alternative polyadenylation across human tissues. *Nucleic Acids Res.*, **48**, D34–D39.
35. Wang,R. and Tian,B. (2020) APAlyzer: a bioinformatics package for analysis of alternative polyadenylation isoforms. *Bioinformatics*, **36**, 3907–3909.
36. Ha,K. C.H., Blencowe,B.J. and Morris,Q. (2018) QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol.*, **19**, 45.
37. Church,D.M., Schneider,V.A., Graves,T., Auger,K., Cunningham,F., Bouk,N., Chen,H.C., Agarwala,R., McLaren,W.M., Ritchie,G.R. *et al.* (2011) Modernizing reference genome assemblies. *PLoS Biol.*, **9**, e1001091.
38. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
39. 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.
40. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A., Keane,T., McCarthy,S.A., Davies,R.M. *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**, giab008.
41. Lawrence,M., Huber,W., Pagès,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M.T. and Carey,V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.
42. Ramírez,F., Ryan,D.P., Grüning,B., Bhardwaj,V., Kilpert,F., Richter,A.S., Heyne,S., Dündar,F. and Manke,T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–165.
43. Zerbino,D.R., Johnson,N., Juettemann,T., Wilder,S.P. and Flicek,P. (2014) WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics*, **30**, 1008–1009.
44. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
45. Pervouchine,D.D., Knowles,D.G. and Guigó,R. (2013) Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics*, **29**, 273–274.
46. Tang,P., Yang,Y., Li,G., Huang,L., Wen,M., Ruan,W., Guo,X., Zhang,C., Zuo,X., Luo,D. *et al.* (2022) Alternative polyadenylation by sequential activation of distal and proximal PolyA sites. *Nat. Struct. Mol. Biol.*, **29**, 21–31.
47. Malka,Y., Steiman-Shimony,A., Rosenthal,E., Argaman,L., Cohen-Daniel,L., Arbib,E., Margalit,H., Kaplan,T. and Berger,M. (2017) -UTR cleavage of mRNA transcripts generates thousands of stable uncapped autonomous RNA fragments. *Nat. Commun.*, **8**, 2029.
48. Malka,Y., Alkan,F., Ju,S., rner,P.R., Pataskar,A., Shulman,E., Loayza-Puch,F., Champagne,J., Wenzel,C., Faller,W.J. *et al.* (2022) Alternative cleavage and polyadenylation generates downstream uncapped RNA isoforms with translation potential. *Mol. Cell*, **82**, 3840–3855.
49. Sun,Y., Zhang,Y., Hamilton,K., Manley,J.L., Shi,Y., Walz,T. and Tong,L. (2018) Molecular basis for the recognition of the human AAUAAA polyadenylation signal. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, E1419–E1428.
50. Vainberg Slutskin,I., Weinberger,A. and Segal,E. (2019) Sequence determinants of polyadenylation-mediated regulation. *Genome Res.*, **29**, 1635–1647.
51. Jensen,T.H., Jacquier,A. and Libri,D. (2013) Dealing with pervasive transcription. *Mol. Cell*, **52**, 473–484.
52. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
53. Hangauer,M.J., Vaughn,I.W. and McManus,M.T. (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet.*, **9**, e1003569.
54. Zhang,A., Li,S., Apone,L., Sun,X., Chen,L., Ettwiller,L.M., Langhorst,B.W., Noren,C.J. and Xu,M.Q. (2018) Solid-phase enzyme catalysis of DNA end repair and 3' A-tailing reduces GC-bias in next-generation sequencing of human genomic DNA. *Sci. Rep.*, **8**, 15887.
55. Zheng,S., Vuong,B.Q., Vaidyanathan,B., Lin,J.Y., Huang,F.T. and Chaudhuri,J. (2015) Non-coding RNA generated following lariat debranching mediates targeting of AID to DNA. *Cell*, **161**, 762–773.
56. Boutz,P.L., Bhutkar,A. and Sharp,P.A. (2015) Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev.*, **29**, 63–80.
57. Ruskin,B. and Green,M.R. (1985) An RNA processing activity that debranches RNA lariats. *Science*, **229**, 135–140.
58. Mohanta,A. and Chakrabarti,K. (2021) Dbr1 functions in mRNA processing, intron turnover and human diseases. *Biochimie*, **180**, 134–142.
59. Twigg,S.R., Lloyd,D., Jenkins,D., Elçioglu,N.E., Cooper,C.D., Al-Sannaa,N., Annagür,A., Gillessen-Kaesbach,G., Hüning,I., Knight,S.J. *et al.* (2012) Mutations in multidomain protein MEGF8 identify a Carpenter syndrome subtype associated with defective lateralization. *Am. J. Hum. Genet.*, **91**, 897–905.
60. Azouz,A. and Duke-Cohan,J.S. (2020) Post-developmental extracellular proteoglycan maintenance in attractin-deficient mice. *BMC Res. Notes.*, **13**, 301.
61. Nogami,T., Beppu,H., Tokoro,T., Moriguchi,S., Shioda,N., Fukunaga,K., Ohtsuka,T., Ishii,Y., Sasahara,M., Shimada,Y. *et al.* (2011) Reduced expression of the ATRX gene, a chromatin-remodeling factor, causes hippocampal dysfunction in mice. *Hippocampus*, **21**, 678–687.
62. Fejes-Toth,K., Sotirova,V., Sachidanandam,R., Assaf,G., Hannon,G.J., Kapranov,P., Foissac,S., Willingham,A.T., Duttagupta,R., Dumais,E. *et al.* (2009) Post-transcriptional processing generates a diversity of $5'$-modified long and short RNAs. *Nature*, **457**, 1028–1032.
63. Fiszbein,A., McGurk,M., Calvo-Roitberg,E., Kim,G., Burge,C.B. and Pai,A.A. (2022) Widespread occurrence of hybrid internal-terminal exons in human transcriptomes. *Sci. Adv.*, **8**, eabk1752.
64. Sanfilippo,P., Miura,P. and Lai,E.C. (2017) Genome-wide profiling of the 3' ends of polyadenylated RNAs. *Methods*, **126**, 86–94.
65. Early,P., Rogers,J., Davis,M., Calame,K., Bond,M., Wall,R. and Hood,L. (1980) Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell*, **20**, 313–319.
66. Rogers,J., Early,P., Carter,C., Calame,K., Bond,M., Hood,L. and Wall,R. (1980) Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain. *Cell*, **20**, 303–312.
67. 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.
68. Benech,P., Mory,Y., Revel,M. and Chebath,J. (1985) Structure of two forms of the interferon-induced ($2'$-5') oligo A synthetase of human cells based on cDNAs and gene sequences. *EMBO J.*, **4**, 2249–2256.

69. Wang,H., Sartini,B.L., Millette,C.F. and Kilpatrick,D.L. (2006) A developmental switch in transcription factor isoforms during spermatogenesis controlled by alternative messenger RNA 3′-end formation. *Biol. Reprod.*, **75**, 318–323.
70. Wagner,E. and Lykke-Andersen,J. (2002) mRNA surveillance: the perfect persist. *J. Cell Sci.*, **115**, 3033–3038.
71. Lykke-Andersen,S. and Jensen,T.H. (2015) Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.*, **16**, 665–677.
72. Vasudevan,S., Peltz,S.W. and Wilusz,C.J. (2002) Non-stop decay–a new mRNA surveillance pathway. *Bioessays*, **24**, 785–788.
73. Kamieniarz-Gdula,K. and Proudfoot,N.J. (2019) Transcriptional control by premature termination: a forgotten mechanism. *Trends Genet.*, **35**, 553–564.
74. Li,Y., Chen,Z.Y., Wang,W., Baker,C.C. and Krug,R.M. (2001) The 3′-end-processing factor CPSF is required for the splicing of single-intron pre-mRNAs in vivo. *RNA*, **7**, 920–931.
75. Misra,A. and Green,M.R. (2016) From polyadenylation to splicing: dual role for mRNA 3’ end formation factors. *RNA Biol.*, **13**, 259–264.