Transcriptome sequencing suggests that pre-mRNA splicing counteracts premature intronic polyadenylation

Mariia Vlasenok[1], Sergei Margasyuk[1], and Dmitri D. Pervouchine[1,*]

[1]Skolkovo Institute of Science and Technology, Moscow 143026, Russia

[*]e-mail: d.pervouchine@skoltech.ru

May 27, 2022

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 physiological and disease conditions, where it is currently attributed to stochastic variations in 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 usage of intronic polyadenylation sites with tissue-specific splicing. A combination of computational methods including the analysis of short reads with non-templated adenines confirmed highly abundant IPA events. Among them, composite terminal exons and skipped terminal exons expectedly correlate with splicing, however we also observed a considerable fraction of IPA events that lack AS support and can be attributed to lariat polyadenylation (LPA). We hypothesize that LPA originates from a dynamic coupling between APA and AS, in which the spliceosome removes an

intron after CPA have already occurred in it. Taken together, these results suggest that co-transcriptional pre-mRNA splicing could serve as a natural mechanism of suppression of premature transcription termination.

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 or 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, resulting in intronic polyadenylation (IPA) (*9*). While alternative 3'-UTRs contain cis-regulatory elements that impact the stability, localization and translation rate of the mRNAs (*5*), the alteration of the protein primary sequence can lead to important functional changes (*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 the tumor-suppressive functions and contribute significantly to tumor onset and progression (*11*). Thousands of recurrent and dynamically changing IPA events have been identified in transcriptomic studies, indicating that current knowledge on IPA is largely incomplete (*13*).

The interplay between splicing and polyadenylation has long been recognized as being related to cotranscriptional pre-mRNA processing (*14*). Many splicing factors have dual roles serving both splicing and polyadenylation, including *U2AF* (*15*), *PTBP1* (*16*), members of Hu protein family (*17*), and others (*8*). The observation that IPA is associated with weaker 5'-splice sites and longer introns (*9*), and experiments on mutagenesis of CPA and splicing signals in

plants (*18*) together suggest that splicing and polyadenylation operate in a dynamic competition with each other. Furthermore, nascent RNA polymerase II transcripts, which are susceptible to CPA at cryptic PASs, are protected from premature transcription termination by U1 snRNP in a process called telescripting, most remarkably in genes with longer introns (*19*). This raises a number of challenging questions about the abundance of cryptic intronic PASs, mechanisms of their inactivation, and relation to alternative splicing.

Various experimental protocols have been developed to identify the genomic positions of PASs (*20*). Many of them use oligo(dT) (3'RNA-seq, PAS-seq, polyA-seq) or similar primers (3'READS) to specifically capture transcript ends (*21–24*). A combination of these protocols yielded a consolidated set of more than 500,000 human PASs (*25–27, 27*), 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 (*13, 28, 29, 29–32*). 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 focusing on quantifying relative usage of PASs with known genomic positions rather than identifying novel 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 poly(A) 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 implementations of this approach, ContextMap2 (*32*) and KLEAT (*31*), demonstrated that the analysis of RNA-seq reads containing a part of the poly(A) tail 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 (*33*), to simultaneously assess alternative splicing and intronic polyadenylation and match their tissue-specific patterns. We found a remarkable variability of PAS positions around annotated transcript ends and identified a core set of 318,898 PAS clusters that are expressed in

GTEx tissues, which is consistent with other published sets. We further characterized the distribution of PAS clusters in the untranslated, exonic, and intronic regions of protein-coding genes and described the relationship between tissue-specific IPA and AS. In inspecting the concordance between IPA and AS patterns, we unexpectedly found a considerable fraction of unannotated intronic PAS clusters lacking splicing support and attributed them to lariat polyadenylation (LPA), a term we introduce here to describe the dynamic coupling between CPA and AS.

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 only 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 9,021 GTEx RNA-seq experiments (*33*) with additional restrictions on sequencing quality (see Methods). We extracted polyA reads, defined as reads containing a soft clipped region of at least six nucleotides that consists of 80% or more adenines, excluding short reads aligning to adenine-rich genomic tracks and omitting samples with exceptionally large numbers of polyA reads (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, indicating 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 read support, and the distribution of their overhangs. Our confidence in PAS correlates not only with the read support, but also with the diversity of the overhang distribution, 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 $H \geq 2$ (Fig-

ure S2). 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 (*25*) 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 (*34, 35*). 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, large absolute number of PASs in intergenic regions, including PASs without canonical consensus CPA signals, points at a remarkable number of RNA Pol II transcripts that are transcribed from them consistently with the current knowledge on pervasive transcription (*36–38*).

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. Unexpectedly, instead of a single peak, we observed a relatively dispersed cluster of PASs spanning twelve nucleotides. Manual inspection confirmed that the RNA-seq read alignments ending in all these positions indeed were followed by non-templated polyA tracks, thus indicating that the observed pattern was due to biological stochasticity and not to mapping artifacts. 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 (*39*).

PAS clusters

Large 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. Similarly, for each annotated TE, we computed the interquartile range (IQR) of the distances to all PASs located within 100 nts, excluding TEs with a single PAS (Figure 2B). Approximately 83% of TEs had IQR below 10 nts, and 87% of TEs did so when considering only PASs with a signal. A similar variability of PAS positions was observed in a massively parallel reporter assay (*35*). We therefore chose to merge PASs that were located within 10 nts of each other (Figure 2C). 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 ten individual PASs (Supplementary File 2). 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 the consolidated atlas of PASs derived from 3'-end sequencing (PolyASite 2.0 (*25*), in what follows referred to as Atlas) and TEs annotated within the GENCODE consortium (*40*). 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. 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 consistent

(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 2E). Despite TEs no longer being considered, we observed a considerable increase in the density of PASCs 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 (Figure S4).

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 distribution within gene parts, namely in the 5'-untranslated region (5'-UTR), the 3'-untranslated region (3'-UTR), and the coding part (CDS). Each CDS region was further subdivided into intronic, always 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 the read support, in which each PASC was weighted by the number of supporting polyA reads (Figure 3).

As expected, PASCs were quite frequent in CDS by absolute number, but their density was the highest in 3'-UTRs since CDS 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 (Figure S5), and similar peaks were also observed for PolyASite 2.0 (Figure S6). However, despite low density, intronic PASCs were still quite frequent in number, and among them there could be PASCs leading to premature CPA.

One obvious reason for low intronic density of PASCs is the undercoverage bias of the

polyA RNA-seq protocol, in which introns become invisible for RNA-seq after they are removed by the spliceosome and degraded. This indicates that the number of supporting polyA reads could strongly underestimate the actual abundance of IPA events. We therefore normalized the number of polyA reads to the average read coverage in the respective CDS parts and found that polyA reads would have been most frequent in intronic regions if the coverages were the same (Figure 3C).

Complementary to this, we estimated the relative frequency of single nucleotide substitutions that give rise to the canonical polyA signal (AATAAA) from a pre-consensus sequence in the common ancestor of human and macaque (see Methods). The number of substitutions creating the AATAAA signal relative to the number of substitutions creating other hexamers was significantly higher in intronic as compared to other regions (Figure 3D). In sum, these findings indicate that intronic PASCs could be more active than exonic PASCs both in terms of expression and evolutionary dynamics.

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. We adapted the 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 $\mathrm{logFC} = \log_{10}(wi_1/wi_2)$ metric, which captures the magnitude of read coverage drop at PASC, and DESeq2 (*41*), 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, excluding PASCs located within 200 nts from exon boundaries 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 drop af-

ter PASCs. Consistently with this decrease, the $\log\mathrm{FC}$ distribution was skewed towards positive values with a noticeably bigger skewness for PASCs with a signal and PACSs near annotated TEs (Figure 4B). The number of supporting polyA reads was positively correlated with $\log\mathrm{FC}$ 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 (15%) had $\log\mathrm{FC} > 1$ per tissue, 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 $\log\mathrm{FC} > 1$ were also significant according to DESeq results. Since the results of the two methods overlapped, we chose to call a PASC with $\log\mathrm{FC} > 1$ as expressed.

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 algorithm (*42*). Since the exact positions of PASCs in 3'-UTRs may vary, we selected 3'-UTRs that contain at least one expressed PASC according to $\log\mathrm{FC} > 1$ condition and compared them to 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 a PASC with $\log\mathrm{FC} > 1$ were also called as expressed by DaPars, and vice versa 50% of 3'-UTRs called as expressed by DaPars contained at least one PASC with $\log\mathrm{FC} > 1$, thus confirming that the expression of PASCs in tissues as measured by the $\log\mathrm{FC}$ metric and the results obtained by DaPars are consistent.

Since the analysis of tissue-specific polyadenylation *per se* falls beyond the scope of this report, we next focused on intronic PASCs and examined the relationship between IPA and AS.

Intronic polyadenylation and splicing

Alternative terminal exons that are generated through IPA can be divided 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 (*9*). To distinguish between these possibilities, we estimated the average read coverage in two additional windows, we_1 and we_2, and computed the number of split reads starting at the intron 5'-end and landing before (b)

and after (a) each intronic PASC (iPASC) in each tissue (Figure 5A). We expect that, in addition to large wi_1/wi_2 ratio, STE are characterized by large we_1/we_2 ratio and presence of split reads landing before PASC, while CTE are characterized by small we_1/we_2 ratio and absence of such split reads. For simplicity, the values of the read coverage in the four windows are also denoted by we_1, we_2, wi_1, and wi_2. To quantify the rate of splicing, we used $\psi = a/(a+b)$ ratio. Large ψ values ($\psi \simeq 1$) indicate that the intron is spliced canonically, while low ψ values ($\psi \simeq 0$) indicate the presence of unannotated AS events before iPASC.

As a result, we obtained 2,079,325 iPASC-tissue pairs comprising 67,075 iPASCs in 31 tissues and evaluated Pearson correlation coefficient r between ψ and logFC for each iPASC with a sufficiently large ψ range ($IQR > 0.03$) across tissues. As expected, the distribution of r was significantly skewed towards negative values as compared to the distribution, in which tissue labels were shuffled (Figure 5B). We manually followed specific examples in *NCAM1* (Neural Cell Adhesion Molecule 1) and *SORBS2* (Sorbin And SH3 Domain Containing 2) genes and, indeed, observed a substantial negative association between CPA and splicing, i.e., the larger the splicing rate, the lower the CPA rate (Figure 5C).

Next, we considered 87,622 iPASC-tissue pairs with a substantial read coverage drop at PASC ($\mathrm{logFC} > 1$) and a sufficiently high read coverage in the intronic window before PASC ($wi_1 > 0.1we_1$). The bivariate distribution of $\log(we_1)$ and $\log(we_2)$ (Figure 5D, left) revealed two groups of PASCs separated by the line $we_2 = 0.3we_1$, one with comparable values of we_1 and we_2 (above the line) and the other, in which we_2 was substantially lower than we_1 (below the line). Our expectation was that these two groups, $we_1 \simeq we_2$ and $we_1 \gg we_2$, correspond to CTE and STE, respectively. Indeed, when considering 968 annotated CTEs and 1880 annotated STEs, we found that the former clustered above the separating line (Figure 5D, middle), while the latter clustered below (Figure 5D, right). In accordance with this observation, the distributions of ψ values of the annotated CTE were characterized by a pronounced peak at $\psi \simeq 1$ indicating the absence of splicing events in the intron before PASC, while STE had a peak at $\psi \simeq 0$ indicating that a splice site upstream of PASC was used (Figure 5E, middle and right).

However, in inspecting the distribution of all 87,622 iPASC-tissue pairs, of which approximately 35% (respectively, 65%) were located above (respectively, below) the separating line,

we unexpectedly found a bimodal distribution of ψ values in the latter group, which should presumably correspond to STE (Figure 5E, left). As expected for STE, the peak at $\psi = 0$ indicates the activation of a splice site in the intron upstream of PASC, while the peak at $\psi = 1$ corresponds to a group of iPASCs that are characterized by a large drop from we_1 to we_2 in the absence of splicing between we_2 and wi_1, which is incompatible with the STE model.

We next followed tissue-specific splicing and CPA patterns in a few cases (Figure 6). The intronic PASC in the *MEGF8* gene, which encodes a membrane protein associated with Carpenter syndrome (*43*), is an example of a CTE supported by intronic reads in absence of splicing events 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 (*44*), PASC is expressed in muscle tissue along with the elevation of read coverage in the upstream region and activation of a splice site at its border, thus likely representing an unannotated STE (Figure 6B). These PASCs are supported by *CSTF2* eCLIP footprints and PolyASite 2.0 (*25*).

In contrast, PASC in the *ATRX* gene, which encodes a chromatin remodeler linked to a range of diseases (*45*), exhibits elevated read coverage upstream of PASC, however it lacks splice junctions 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 these findings would be that canonical splicing and IPA co-exist and operate concurrently, a possibility that we named lariat polyadenylation (LPA). Our hypothesis is that LPA originates from a dynamic coupling between APA and AS, in which the spliceosome can remove an intron after CPA has already occurred in it.

To estimate the abundance of LPA events, we considered a strict set of iPASC-tissue pairs described above and categorized them as CTE, STE, and LPA according to the following criteria: $we_2 > 0.3we_1$ (CTE), $we_2 \leq 0.3we_1$ and $\psi \leq 0.9$ (STE), and $we_2 \leq 0.3we_1$ and $\psi < 0.9$ (LPA) (Supplementary File 4). This yielded 4,435, 2,863, and 2,821 CTE, STE, and LPA cases, respectively, indicating that the latter are almost as abundant as STE and, therefore, must contribute greatly to the observed landscape of intronic polyadenylation.

Discussion

The GTEx dataset represents an ideal resource for studying the interaction between IPA and AS because the information on the positions and tissue-specific expression of intronic PAS is complemented by tissue-specific splicing rates inferred from split reads aligning to splice junctions. Such matched data are currently in high demand (*11*). In this work, we applied for the first time the approach based on short reads containing a part of the poly(A) tail, one that was used previously on much smaller datasets (*31, 32*), to identify PASs from RNA-seq data at the sequencing scale when it becomes efficient. The method can be combined with coverage-based methods to detect tissue-specific usage of PASs, remarkably not only in the untranslated, but also in the coding regions.

PolyA reads provide a snapshot of CPA at single nucleotide resolution, which reveals that PASs form clusters of different sizes. This indicates that the precision of CPA machinery is highly variable, in some cases providing narrow clusters of closely spaced PASs, and broad regions with imprecise cleavage points in the others. Other steps of pre-mRNA processing such as splicing are more restricted to producing error-free mRNAs due to protein-coding constraints, however they are also prone to stochastic variations (*46*). The functional relevance of stochastic variations in CPA events is currently not well understood. Our results raise a valid concern about the determinants of CPA precision in different PAS classes, thus opening new avenues to be explored in future studies.

The approach based on polyA reads has limitations related to mappability of reads with long soft clip regions. The frequency distribution of polyA reads decays with increasing the length of the overhang, likely due to mapping bias. Positional distribution of PASCs in constitutive exons and introns has a pronounced peak in the end of exonic regions and in the beginning of intronic regions, with a particular enrichment for PASCs without a signal (Figure 3C). Yet, PASCs with a signal are also enriched in the 50–150 nt region downstream of the donor splice site. This pattern resembles the correlation between CAGE tags and internal exons of annotated transcripts and widespread occurrence of polyA-seq peaks near exon boundaries (*47, 48*), but it could also result from erroneous mappings of split reads, e.g. when adenine-rich part of the read or a short segment between splice junction and the stretch of non-templated adenines

is incorrectly attributed to a soft clip region by the mapper (example in Figure S7). Of note, mapping split reads with short exonic parts appears to be a common problem of all methods since the positional distribution of PASCs obtained by other protocols, e.g., in PolyASite 2.0 data, has similar peaks near exon boundaries (Figure S6).

While the majority of polyA reads align to 3'-UTRs, a small fraction (5–8%) still map to the coding part raising important concerns about their implication in premature transcription termination (*49*). Transcripts harboring incomplete reading frames translate into potentially deleterious truncated proteins that may pose a hazard to the cell (*50*). In eukaryotes, early transcription termination is tightly linked with CPA, which occurs at cryptic PAS typically located in introns, although a small fraction of PAS-mediated cleavage may also occur within internal exons. IPA typically generates transcripts that harbor a premature termination codon (PTC) or transcripts without a stop codon, which are unstable and get rapidly degraded via nonsense mediated decay (*51*) and nonstop decay pathways (*52*). A number of functionally important IPA cases have been described in specific genes (*10, 53–57*), but the widespread nature of IPA has been appreciated only recently with the development of 3'-end sequencing methods (*58*).

Strikingly, despite low density, PASCs within the coding part are quite abundant in number and, after proper normalization of the read coverage, they appear to be much more frequent in introns than in exons. Higher abundance of PASCs in introns is complemented by weaker evolutionary pressure on generating the canonical AATAAA consensus from pre-consensus sequences in introns, which on one hand may reflect the constraints on maintaining the amino acid sequence, but, on the other hand, also hints at the existence of a mechanism that counteracts the activity of cryptic intronic PASs. A remarkably large number of intronic PASs that are listed in current catalogs brings an outstanding question of how could it be that virtually every intron of every human gene contains a cryptic intronic PAS, but cells are still able to produce full-length transcripts?

Our hypothesis is that a considerable fraction of the observed IPA cases could be attributed to LPA, a situation in which splicing and polyadenylation co-exist and operate concurrently along with the elongating transcription. The spliceosome and the CPA machinery both recognize signals that are located in the nascent pre-mRNA and bind the same pre-mRNA substrate

at the same time. These processes operate at their intrinsic rates subordinate to the transcription elongation speed and, depending on tissue-specific conditions, one of them could operate faster than the other. Particularly, if the spliceosome has already assembled on an intron when CPA started PAS-mediated cleavage, the second catalytic step of the splicing reaction would remove the polyadenylated part, thus leading to LPA. If CPA machinery has operated faster than the spliceosome could excise the intron, then the outcome would be IPA. Recently, we proposed a related mechanism to explain RNA structure-mediated suppression of premature CPA (*59*). However, besides RNA structure, a multitude of tissue-specific factors, all which are impossible to list here, are responsible for correct temporal and spatial interactions of splicing and CPA machineries. In this light, it appears plausible that an important side function of co-transcriptional splicing might be to prevent premature transcription termination by counteracting the activity of cryptic intronic PASs through LPA.

Conclusion

Massive amounts of RNA-seq data in the GTEx dataset open a unique possibility to jointly analyze tissue-specific splicing and polyadenylation. Patterns of intronic polyadenylation and splicing again demonstrate that splicing and polyadenylation are two inseparable parts of one consolidated pre-mRNA processing machinery, leading to a conjecture that co-transcriptional splicing could be a natural mechanism of suppression of premature transcription termination.

Acknowledgments

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

Funding

All authors acknowledge Russian Science Foundation grant 21-64-00006.

Availability of data and materials

The datasets generated during the current study are available online at `https://zenodo.org/record/6587186`. The source code used for the analysis is available at `https://github.com/mashlosenok/RNAseq_PAS_finder`.

Authors' contributions

DP designed and supervised the study; MV and SM performed data analysis; DP and MV wrote the first draft of the manuscript. All authors edited the final version of the manuscript.

Methods

Genome assembly and transcript annotation

February 2009 (hg19) assembly of the human genome and GENCODE transcript annotation v34lift37 were downloaded from Genome Reference Consortium (*60*) and GENCODE website (*40*), respectively. Transcript annotations were parsed by custom scripts to extract the coordinates of transcript ends, exons and introns. The attribution of PAS to protein-coding, non-coding, and intergenic segments was done on the basis of their occurrence in the corresponding gene types.

Partition of protein-coding genes

To partition protein-coding genes into segments, we parsed the annotation of protein-coding transcripts from GENCODE and extracted 5'-UTRs, 3'-UTRs and CDS of all transcripts as follows. Genomic regions that were not covered by any transcript were classified as intergenic. A gene part was classified as 5'-UTR (respectively, 3'-UTR) if it belonged to the 5'-UTR (respectively, 3'-UTR) of at least one annotated transcript of the gene; the rest of the gene sequence was classified as CDS. We next considered exons and introns of all annotated protein-coding transcripts and used them to further subdivide CDS regions into exonic, intronic, and alternative

regions. A genomic region was classified as always exonic (respectively, intronic) if it belonged to exonic (respectively, intronic) parts of all annotated transcripts that overlap the region; otherwise, it was classified as an alternative exonic region.

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 (*61*). PySAM suite was used to extract uniquely mapped reads (NH:1) (*62*). 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, which always corresponds to the positive strand according to BAM format, contained 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 (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

$$H = \frac{\sum\limits_i f_i \log_2 f_i}{f} - \log_2 f.$$

In order to select a reasonable number of PAS, 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 (Figure S2). The threshold $H \geq 2$ in combination with the minimum overhang length of 6 nts appears to be optimal since it captures 85% annotated gene ends and yields 565,387 PAS, a number that corresponds by

the order of magnitude with the size of the PAS set reported in PolyASite 2.0 (*25*). PASs that were located within 10 nts of each other were merged into clusters (PASCs) using the GenomicRanges package (*63*).

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 (*25*) (referred to as Atlas) and the set of annotated TEs provided by GENCODE consortium (*40*) (referred to as GENCODE). In each comparison, we calculated the precision and recall metrics of GTEx with respect to the reference set by imposing variable thresholds on PASC support level. 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. 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 $F_1 = (P^{-1} + R^{-1})^{-1}$ score at $P = 0.57 - 0.58$ and $R = 0.49 - 0.51$ (Figure S3, top left). The same scores, in which each PASC was weighted by the read support, showed a similar performance with the optimal F_1 score of $P = 0.83 - 0.86$ and $R = 0.73 - 0.76$ (Figure S3, bottom left). In comparison to Atlas as a reference set by the number of PASC, 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 (Figure S3, top right). However, when the same comparison was made by the number of transcripts, i.e., by weighting PASCs by the read support, the precision and recall were 0.92 and 0.97, respectively, indicating that the GTEx primarily misses PASCs with low level of read support (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. In the same way, PASC relative positions were computed in exonic and intronic regions.

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 (*64*) and averaged the read coverage values between samples within each tissue using `wiggletools mean` utility (*65*).

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` (*64*) and computed the log-fold-change metric ($\log\mathrm{FC}$) as the logarithm of the ratio of the mean read coverage in the upstream and downstream windows, respectively, with 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 ($p_{adj} < 10^{-3}$) using DESeq2 (*41*), separately in each tissue.

Intronic PASCs were defined as PASCs located within at least one annotated intron of a protein-coding gene >200bp away from the closest annotated splice site ($n = 67,075$). The shortest intron containing a PASC was chosen, and the average read coverage was computed not only in wi_1 and wi_2, but also in 150-nt windows starting 10 nts upstream and downstream of the intron 5'-end (we_1 and we_2, Figure 5A). An intronic PASC located within 100 nts from an annotated TE of a protein-coding transcript ($n = 2,921$) was categorized as STE (respectively, CTE) if the terminal exon of the transcript fully belonged to the containing intron (respectively, overlapped the interval from 5'-splice site to PASC). This categorization yielded 968 CTEs and 1880 STEs, while 73 PASCs were located near TEs of several transcripts resulting in conflicting annotation.

To estimate the mean read coverage in constitutive exons, alternative exons, and introns, the total read coverage values per nucleotide in all GTEx samples were averaged between windows

located in the respective regions to obtain normalization factors ($3.3 \cdot 10^6$, $3.2 \cdot 10^6$, and $8.0 \cdot 10^4$, respectively). The latter were used to normalize the fraction of polyA reads in the respective regions (Figure 3C) relative to the average read coverage.

Splicing metrics

To quantify tissue-specific alternative splicing associated with intronic PASCs, we computed split read counts using IPSA pipeline as explained earlier (*33, 66*). The counts of split reads were pooled within each tissue to compute the $\psi = a/(a+b)$ metric (Figure 5A), defined here as the number of split reads supporting splicing of the shortest annotated intron that contains PASC (a) as a fraction of the number of split reads supporting splicing of the shortest annotated intron and the number of split reads supporting splicing from the donor site to any acceptor site located before PASC (b). The latter split reads are referred to as "landing before PASC".

Evolutionary dynamics of consensus sequences

In order to quantify the number of single nucleotide substitutions that convert a pre-consensus polyadenylation signal (defined as any sequence that differs by 1 nt from the canonical AATAAA sequence) to the canonical polyadenylation signal AATAAA, we downloaded multiple sequence alignment of 45 vertebrate genomes with the human genome (GRCh37) from the UCSC Genome Browser in MAF format (*67*). The alignments with *M. mulatta* (rhesus) and *C. jacchus* (marmoset) genomes were extracted from MAF, and the alignment blocks were concatenated. The genomic sequence in the common ancestor (CA) of human and rhesus with marmoset as an outgroup was reconstructed by parsimony. We identified all positions of pre-consensus hexamers in the CA and computed the number of single nucleotide substitutions on the human branch that led to the canonical AATAAA signal as a fraction of single nucleotide substitutions on the human branch that led to any change in the pre-consensus, separately in always exonic, alternative exonic, and intronic regions. In total, 16,408,153 such substitutions were analyzed.

Figures

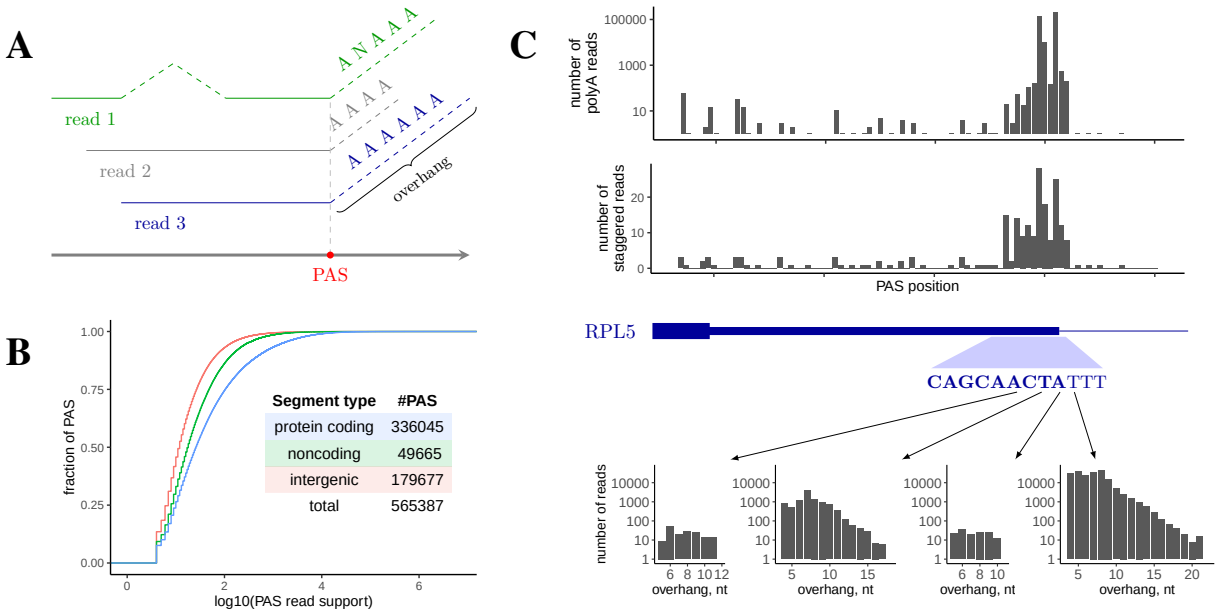

Figure 1: The identification of PAS. (A) The alignments of short reads with non-templated adenine-rich ends (polyA reads). The genomic position of the first non-templated nucleotide corresponds to a PAS. The length of the soft clip region is referred to as overhang. **(B)** PolyA read support of PAS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

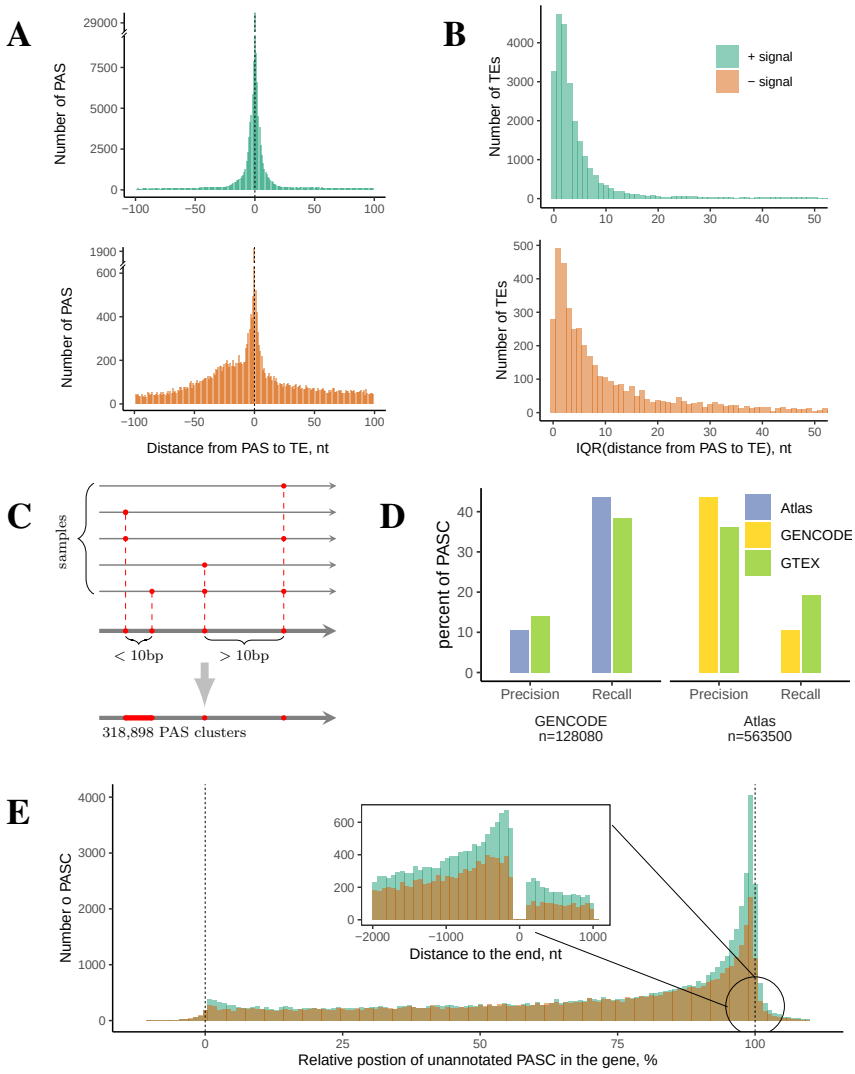

Figure 2: PAS clusters in protein-coding genes. **(A)** The distribution of distances from each PAS to its closest annotated transcript end (TE) for PAS with ($n = 122,448$) and without a signal ($n = 22,361$). **(B)** The variability of PAS positions around TEs, measured as the interquartile range (IQR) of distances from the TE to all PASs within 100 nts. **(C)** PAS located <10 bp from each other are merged into PAS clusters (PASCs). **(D)** Pairwise comparison of PASs inferred from GTEx, PolyASite 2.0 (*25*) (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). **(E)** 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.

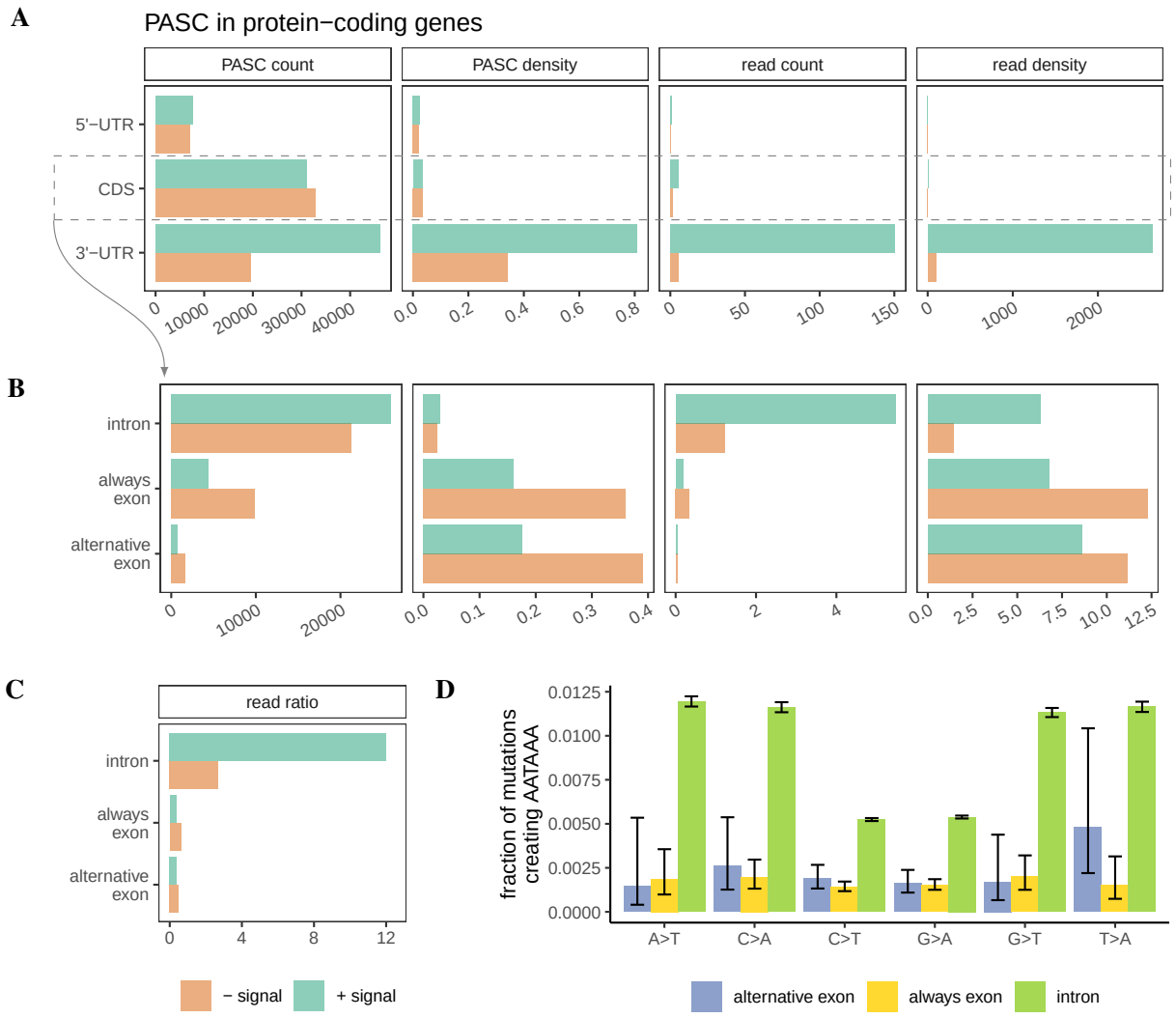

Figure 3: PAS clusters in protein-coding regions. **(A)** The distribution of PASCs in 5'-UTRs, CDS, and 3'-UTRs. Shown are the total number of PASC (PASC count), PASC density per nt (PASC density), the total number of polyA reads (read count), the total number of polyA reads per nt (read density). **(B)** The distribution of PASCs from CDS in introns, constitutive exons (always exon), and alternative exons. **(C)** The proportion of polyA reads (reads ratio) 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 relative frequency of single nucleotide substitutions in pre-consensus sequences that give rise to the canonical polyA signal (AATAAA) in the human lineage.

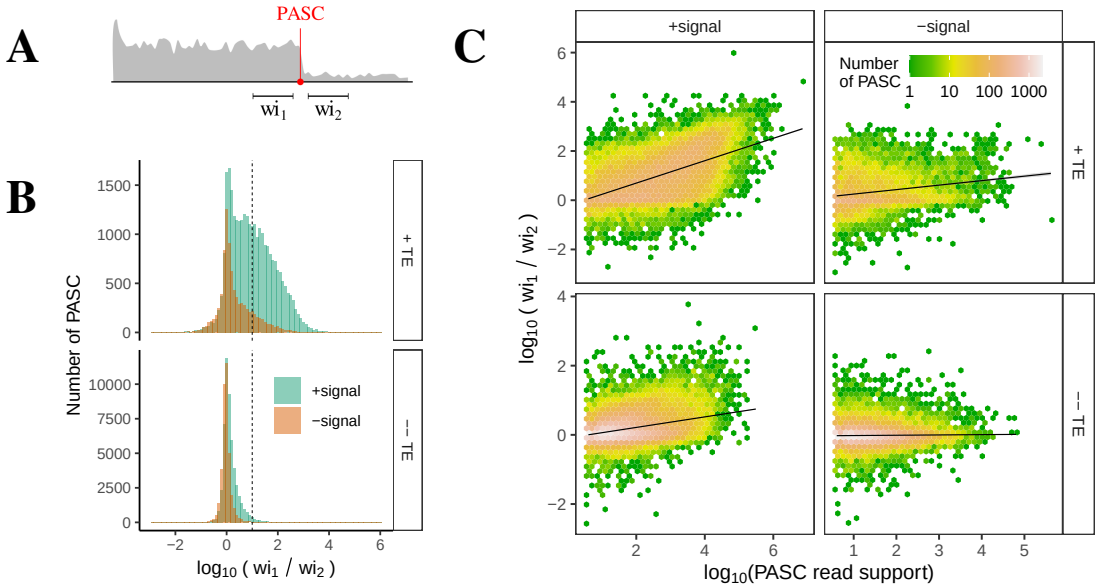

Figure 4: Coverage-based metrics of PASC expression. **(A)** The average read coverage was measured in 150-nt upstream and downstream windows, wi_1 and wi_2, around PASC. **(B)** The distribution of $\text{logFC} = \log_{10}(wi_1/wi_2)$ metric for annotated ($n = 37,194$, top) and unannotated PASCs ($n = 89,116$, bottom). A PASC is referred to as annotated if it is within 100 bp of an annotated TE. The dashed line represents the cutoff $\text{logFC} = 1$. **(C)** The $\text{logFC} = \log_{10}(wi_1/wi_2)$ metric positively correlates with the number of supporting polyA reads not only for annotated, but also for unannotated PASCs with a signal.

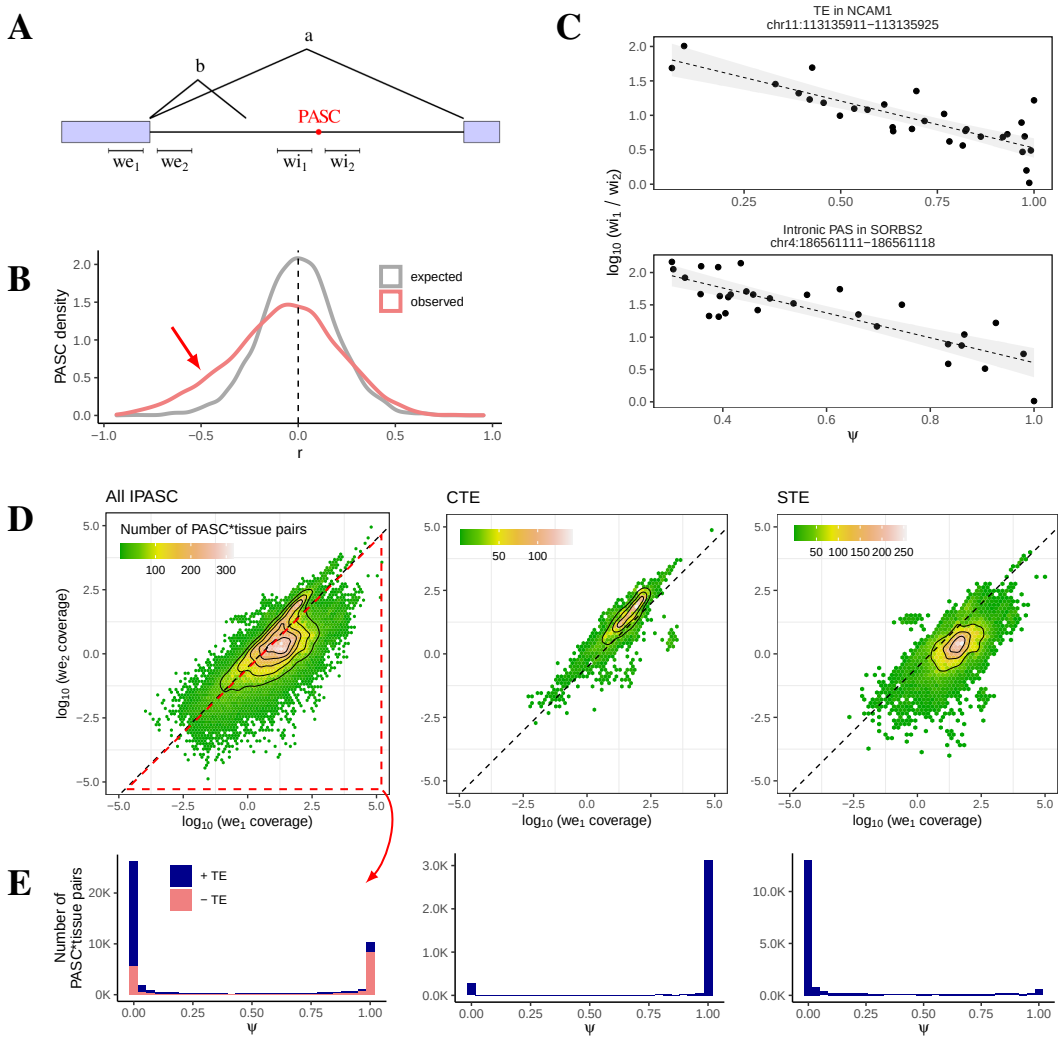

Figure 5: Intronic polyadenylation and splicing. **(A)** Exonic (we_1 and we_2) and intronic (wi_1 and wi_2) 150-nt windows used to assess PASC expression and splicing; a denotes the number of split reads supporting the annotated intron. b denotes the number of split reads landing before PASC. **(B)** The distribution of Pearson correlation coefficients of ψ and $\log_{10}(wi_1/wi_2)$ for $n = 5,081$ PASCs, as compared to shuffled control. The bias towards negative values is indicated by an arrow. **(C)** Case studies of negative association between ψ and $\log_{10}(wi_1/wi_2)$ in *NCAM1* and *SORBS2* genes. The genomic coordinates of PASC are in GRCh37 assembly. **(D)** Bivariate distribution of we_1 vs. we_2 in PASC-tissue pairs for all PASCs ($n = 67,075$, left), annotated CTE ($n = 968$, middle), and STE ($n = 1,880$, right). The dotted line in log coordinates corresponds to $we_2/we_1 = 0.3$. To further analyze unannotated STEs (red triangle), only tissues where iPASC was expressed ($\text{logFC} > 1$) and where the intron coverage was at least 10% of the exon coverage ($wi_1 > 0.1we_1$) were considered. **(E)** ψ distribution for PASCs from the red triangle in panel D (left), CTE (middle), and STE (right); $+$TE ($-$TE) denote PASCs within (not within) 100 nts of annotated TE. The peak at $\psi \simeq 1$ represents putative STEs without evidence of splicing between the upstream exon and PASC, attributed here to lariat polyadenylation (LPA).

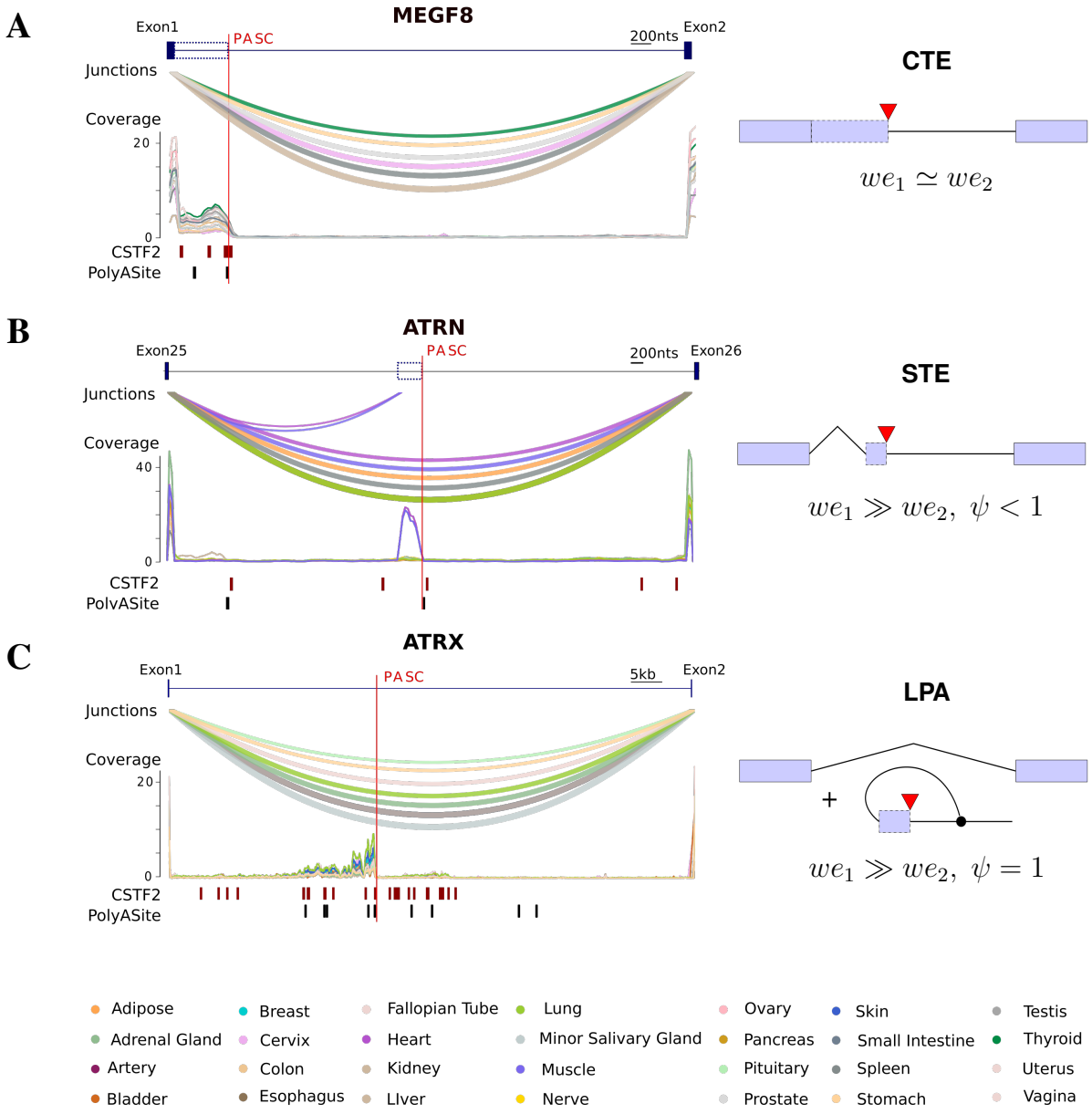

Figure 6: Case studies of IPA. (A) The iPASC between exons 1 and 2 of *MEGF8* represents a CTE, as evidenced by high read coverage in we_2 and the absence of other splicing events $\psi \simeq 1$. The eCLIP peaks of *CSTF2* and PASC from PolyAsite 2.0 are indicated below. Arcs represent tissue-specific splice junctions. **(B)** The iPASC between exons 25 and 26 *ATRN* represents a STE with tissue-specific expression in heart and muscle tissues, as evidenced by splice junctions and the read coverage. **(C)** The iPASC between exons 1 and 2 likely represents a LPA case because the read coverage is low at the 5'-end of the intron and detectable directly upstream of iPASC, but there is no evidence of STE by splice junctions.

References

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

2. Hoque, M. *et al.* Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat Methods* **10**, 133–139 (2013).

3. Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22**, 1173–1183 (2012).

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

5. Mayr, C. What Are 3' UTRs Doing? *Cold Spring Harb Perspect Biol* **11** (2019).

6. Curinha, A., Oliveira Braz, S., Pereira-Castro, I., Cruz, A. & Moreira, A. Implications of polyadenylation in health and disease. *Nucleus* **5**, 508–519 (2014).

7. Fang, Z. & Li, S. Alternative polyadenylation-associated loci interpret human traits and diseases. *Trends Genet* **37**, 773–775 (2021).

8. Gruber, A. J. & Zavolan, M. Alternative cleavage and polyadenylation in health and disease. *Nat Rev Genet* **20**, 599–614 (2019).

9. Tian, B., Pan, Z. & Lee, J. Y. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res* **17**, 156–165 (2007).

10. Di Giammartino, D. C., Nishida, K. & Manley, J. L. Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**, 853–866 (2011).

11. Lee, S. H. *et al.* Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature* **561**, 127–131 (2018).

12. Rakheja, D. *et al.* Somatic mutations in DROSHA and DICER1 impair microRNA biogenesis through distinct mechanisms in Wilms tumours. *Nat Commun* **2**, 4802 (2014).

13. Zhao, Z. *et al.* Cancer-associated dynamics and potential regulators of intronic polyadenylation revealed by IPAFinder using standard RNA-seq data. *Genome Res* **31**, 2095–2106 (2021).

14. Proudfoot, N. J., Furger, A. & Dye, M. J. Integrating mRNA processing with transcription. *Cell* **108**, 501–512 (2002).

15. Kyburz, A., Friedlein, A., Langen, H. & Keller, W. 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 (2006).

16. Castelo-Branco, P. *et al.* Polypyrimidine tract binding protein modulates efficiency of polyadenylation. *Mol Cell Biol* **24**, 4174–4183 (2004).

17. Dai, W., Zhang, G. & Makeyev, E. V. RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage. *Nucleic Acids Res* **40**, 787–800 (2012).

18. Li, Q. Q., Liu, Z., Lu, W. & Liu, M. Interplay between Alternative Splicing and Alternative Polyadenylation Defines the Expression Outcome of the Plant Unique OXIDATIVE TOLERANT-6 Gene. *Sci Rep* **7**, 2052 (2017).

19. Kaida, D. *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664–668 (2010).

20. Chen, W. *et al.* Alternative Polyadenylation: Methods, Findings, and Impacts. *Genomics Proteomics Bioinformatics* **15**, 287–300 (2017).

21. Yu, F. *et al.* Poly(A)-seq: A method for direct sequencing and analysis of the transcriptomic poly(A)-tails. *PLoS One* **15**, e0234696 (2020).

22. Shepard, P. J. *et al.* Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**, 761–772 (2011).

23. Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S. & Mayr, C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27**, 2380–2396 (2013).

24. Zheng, D., Liu, X. & Tian, B. 3'READS+, a sensitive and accurate method for 3' end sequencing of polyadenylated RNA. *RNA* **22**, 1631–1639 (2016).

25. Herrmann, C. J. *et al.* PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing. *Nucleic Acids Res* **48**, D174–D179 (2020).

26. 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* **46**, D315–D319 (2018).

27. You, L. *et al.* 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 (2015).

28. Xia, Z. *et al.* Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun* **5**, 5274 (2014).

29. Cass, A. A. & Xiao, X. mountainClimber Identifies Alternative Transcription Start and Polyadenylation Sites in RNA-Seq. *Cell Syst* **9**, 393–400 (2019).

30. Wang, W., Wei, Z. & Li, H. A change-point model for identifying 3'UTR switching by next-generation RNA sequencing. *Bioinformatics* **30**, 2162–2170 (2014).

31. Birol, I. *et al.* Kleat: cleavage site analysis of transcriptomes. *Pac Symp Biocomput* 347–358 (2015).

32. Bonfert, T. & Friedel, C. C. Prediction of Poly(A) Sites by Poly(A) Read Mapping. *PLoS One* **12**, e0170914 (2017).

33. Melé, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).

34. Sun, Y. *et al.* Molecular basis for the recognition of the human AAUAAA polyadenylation signal. *Proc Natl Acad Sci U S A* **115**, E1419–E1428 (2018).

35. Vainberg Slutskin, I., Weinberger, A. & Segal, E. Sequence determinants of polyadenylation-mediated regulation. *Genome Res* **29**, 1635–1647 (2019).

36. Jensen, T. H., Jacquier, A. & Libri, D. Dealing with pervasive transcription. *Mol Cell* **52**, 473–484 (2013).

37. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

38. Hangauer, M. J., Vaughn, I. W. & McManus, M. T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet* **9**, e1003569 (2013).

39. Zhang, A. *et al.* 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 (2018).

40. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760–1774 (2012).

41. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

42. Hong, W. *et al.* APAatlas: decoding alternative polyadenylation across human tissues. *Nucleic Acids Res* **48**, D34–D39 (2020).

43. Twigg, S. R. *et al.* Mutations in multidomain protein MEGF8 identify a Carpenter syndrome subtype associated with defective lateralization. *Am J Hum Genet* **91**, 897–905 (2012).

44. Azouz, A. & Duke-Cohan, J. S. Post-developmental extracellular proteoglycan maintenance in attractin-deficient mice. *BMC Res Notes* **13**, 301 (2020).

45. Nogami, T. *et al.* Reduced expression of the ATRX gene, a chromatin-remodeling factor, causes hippocampal dysfunction in mice. *Hippocampus* **21**, 678–687 (2011).

46. Hon, C. C. *et al.* Quantification of stochastic noise of splicing and polyadenylation in Entamoeba histolytica. *Nucleic Acids Res* **41**, 1936–1952 (2013).

47. Fejes-Toth, K. *et al.* Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).

48. Fiszbein, A. *et al.* Widespread occurrence of hybrid internal-terminal exons in human transcriptomes. *Sci Adv* **8**, eabk1752 (2022).

49. Kamieniarz-Gdula, K. & Proudfoot, N. J. Transcriptional Control by Premature Termination: A Forgotten Mechanism. *Trends Genet* **35**, 553–564 (2019).

50. Wagner, E. & Lykke-Andersen, J. mRNA surveillance: the perfect persist. *J Cell Sci* **115**, 3033–3038 (2002).

51. Lykke-Andersen, S. & Jensen, T. H. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol* **16**, 665–677 (2015).

52. Vasudevan, S., Peltz, S. W. & Wilusz, C. J. Non-stop decay–a new mRNA surveillance pathway. *Bioessays* **24**, 785–788 (2002).

53. Early, P. *et al.* Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20**, 313–319 (1980).

54. Rogers, J. *et al.* Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain. *Cell* **20**, 303–312 (1980).

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

56. Benech, P., Mory, Y., Revel, M. & Chebath, J. 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 (1985).

57. Wang, H., Sartini, B. L., Millette, C. F. & Kilpatrick, D. L. A developmental switch in transcription factor isoforms during spermatogenesis controlled by alternative messenger RNA 3'-end formation. *Biol Reprod* **75**, 318–323 (2006).

58. Singh, I. *et al.* Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nat Commun* **9**, 1716 (2018).

59. Kalmykova, S. *et al.* Conserved long-range base pairings are associated with pre-mRNA processing of human genes. *Nat Commun* **12**, 2300 (2021).

60. Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol* **9**, e1001091 (2011).

61. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

62. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10** (2021).

63. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**, e1003118 (2013).

64. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160–165 (2016).

65. Zerbino, D. R., Johnson, N., Juettemann, T., Wilder, S. P. & Flicek, P. WiggleTools: parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics* **30**, 1008–1009 (2014).

66. Pervouchine, D. D., Knowles, D. G. & Guigó, R. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* **29**, 273–274 (2013).

67. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res* **47**, D853–D858 (2019).