Skolkovo Institute of Science and Technology

TRANSCRIPTOMIC ANALYSIS OF THE INTERACTION BETWEEN PRE-MRNA SPLICING AND INTRONIC POLYADENYLATION

Doctoral Thesis
by

Mariia Vlasenok

Doctoral Program in Life Sciences

Supervisor
Associate Professor, Dmitri D. Pervouchine

Moscow - 2023

I hereby declare that the work presented in this thesis was carried out by myself at Skolkovo Institute of Science and Technology, Moscow, except where due acknowledgement is made, and has not been submitted for any other degree.

Mariia Vlasenok (Candidate)

Associate Professor Dmitri D. Pervouchine (Supervisor)

Abstract

Alternative splicing (AS) and alternative polyadenylation (APA) are two crucial steps in the post-transcriptional regulation of eukaryotic gene expression. APA can generate transcripts with different C-termini as a result of intronic polyadenylation (IPA). Protocols capturing and sequencing RNA 3'-ends have uncovered widespread IPA in normal and disease conditions, where it is currently attributed to stochastic variations in the pre-mRNA processing. The approach presented in this dissertation is based on the analysis of short reads with non-templated adenines, which provides a powerful alternative to the coverage-based methods when applied to a sufficiently large panel of RNA-seq experiments. Simultaneous assessment of tissue-specific patterns of AS and IPA based on an extensive panel of RNA-seq datasets from the Genotype Tissue Expression project (GTEx) revealed that APA events are more frequent in introns than in exons. While the rate of IPA in the so-called composite terminal exons and skipped terminal exons expectedly correlates with splicing, a considerable fraction of IPA events are not associated with AS events. These IPA events are attributed to the spliced polyadenylated introns (SPI), a term introduced in this dissertation to refer to transient byproducts of the dynamic coupling between APA and AS, in which the spliceosome removes the intron while it is being cleaved and polyadenylated. The results obtained in this Thesis contribute to the understanding of the mechanisms of pre-mRNA processing, providing new evidence for dynamic competition between splicing and polyadenylation within introns and suggesting a potential role for splicing as a safeguard against premature transcription termination at intronic polyadenylation sites.

Publications

Main author

1. Maria Vlasenok, Sergey Margasyuk, and Dmitri D. Pervouchine. Transcriptome sequencing suggests that pre-mRNA splicing counteracts widespread intronic cleavage and polyadenylation. *NAR Genomics and Bioinformatics*, 5(2):lqad051, mar 2023. doi:10.1093/nargab/lqad051.

Co-author

1. Sergey D. Margasyuk, Maria A. Vlasenok, Gaofeng Li, Ch Cao, and Dmitri D. Pervouchine. RNAcontacts: A Pipeline for Predicting Contacts from RNA Proximity Ligation Assays. *Acta Naturae*, 15(1-56):51–57, 2023. doi:10.32607/actanaturae.11893.

2. Alexei Mironov, Marina Petrova, Sergey Margasyuk, Maria Vlasenok, Andrey A Mironov, Dmitry Skvortsov, and Dmitri D Pervouchine. Tissue-specific regulation of gene expression via unproductive splicing. *Nucleic Acids Research*, 51(7):3055–3066, apr 2023. doi:10.1093/nar/gkad161.

Conference presentations

1. D. D. Pervouchine, M. Vlasenok. Transcriptome sequencing suggests lariat polyadenylation. Computational Approaches to RNA Structure and Function 2022. Benasque, Spain.

2. D. D. Pervouchine, M. Vlasenok. Transcriptome sequencing suggests that pre-mRNA splicing counteracts widespread intronic cleavage and polyadenylation. ISMB/ECCB 2023, Lyon, France.

Acknowledgments

I would like to express my gratitude to my supervisor, Dr. Dmitri Pervouchine, for his endless patience, invaluable guidance, and encouragement through the course of this research. I also wish to thank all my teachers and mentors, whose insights, expertise and wisdom inspire me in my academic journey.

Contents

List of Figures

List of Tables

Glossary

APA alternative polyadenylation. 21–23, 26, 33

AS alternative splicing. 38, 83, 87, 97, 101, 133

ATE alternative terminal exon. 22, 101

CPA cleavage and polyadenylation. 23, 26, 39, 51, 60, 61, 63, 64, 67, 76, 83, 89, 90, 96, 97, 99, 101

CPSF Cleavage and Polyadenylation Specificity factor. 17, 20, 28, 40, 102

CSTF Cleavage stimulation Factor. 17, 20, 25, 26, 28, 40

CTD C-terminal domain. 20, 28, 39

CTE Composite Terminal Exon. 22, 83, 84, 87–89, 91, 94, 102, 132, 133

DSE dowstream sequence element. 18–20, 24, 100

GTEx Genotype Tissue Expression project. 46, 50, 51, 58, 69, 78, 97, 101

IPA intronic polyadenylation. 22, 27, 39, 83, 84, 97, 99, 101, 103

iPASC intronic PAS cluster. 52, 83, 84, 88, 93, 95, 132, 133

IQR interquartile range. 49, 62, 66

PAS polyadenylation site. 11, 46, 48, 54, 56, 58–60, 63, 65, 66, 97, 102, 137

PASC PAS cluster. 11, 51, 68, 69, 78, 90, 99, 137

SPI Spliced Polyadenylated Intron. 88–90, 92, 94–96, 102, 105, 132, 133

STE Skipped Terminal Exon. 22, 83, 84, 87–90, 92, 94, 104, 132, 133

TE transcript end. 60–62, 66, 69, 74, 80, 82, 87, 98, 100, 133

Chapter 1

Introduction

The majority of transcripts that are generated by the eukaryotic RNA polymerase II (Pol II) undergo a series of modifications including 5'-end capping, splicing, and 3'-end processing. During splicing, introns are removed, and the remaining exonic regions are ligated. The 3'-end processing involves endonucleolytic cleavage and subsequent attachment of the polyadenylic tail at specific sites called the polyadenylation sites (PASs). The 3'-end processing is therefore commonly referred to as cleavage and polyadenylation (CPA).

Splicing and the 3'-end processing do not always occur at the same sites on the pre-mRNA, leading to alternative splicing (AS) and alternative polyadenylation (APA). AS and APA act in the majority of human genes and are believed to be crucial mechanisms for increasing transcriptomic and proteomic diversity in eukaryotes [121, 12]. APA affects mRNA stability, translation and localization, which makes it implicated in numerous diseases, ranging from cancers to neurological disorders [52]. APA can also happen in introns. Over 20% of human genes undergo intronic polyadenylation (IPA) and many more possess cryptic intronic PASs [159]. An important feature of IPA is that it results in protein isoforms with different C-termini and can lead to major functional alterations through protein domain loss. However, current estimates of the actual IPA rate in the human genome remain unclear, and the identification of cryptic intronic PASs and the mechanisms of their suppression is an outstanding and challenging question.

Experimental protocols to identify PASs often employ oligo(dT)-based primers

to specifically capture transcript ends. Although these techniques have identified over half a million human PASs, many remain undiscovered due to tissue- and disease-specific variations [58]. However, there are not many such experiments that are available for public use, and they are too costly to conduct on a large scale. Instead, computational methods could also be employed to identify PASs from more prevalent experimental datasets such as RNA-seq.

In RNA-seq, the existing approaches primarily detect PASs based on the abrupt decrease in read coverage. However, since the density of RNA-seq reads is highly non-uniform along the gene length, most of these methods are focused on PASs within the terminal exons [52]. However, polyA(+) RNA-seq data also include a small fraction of the so-called polyA reads, which cover the junction between the terminal exon and the start of the polyA tail and contain detectable non-templated adenine residues. Studies demonstrated that the analysis of polyA reads could serve as a specific, albeit less sensitive, alternative to the coverage-based methods, potentially effective when analyzing large panels of RNA-seq experiments [184, 180]. Such large RNA-seq experiment panels, however, have not yet been analyzed using polyA reads.

The Genotype Tissue Expression (GTEx) project released an extensive compendium of RNA-seq datasets that enables systematic analysis of human tissue transcriptomes. This dissertation uses the GTEx dataset to systematically examine the interplay between pre-mRNA splicing and intronic polyadenylation by analyzing IPA through polyA reads. Previous works used coverage-based approaches and were primarily focused on tissue-specific polyadenylation. In contrast, this study combines the analysis of polyA reads for PAS identification, split reads to quantify alternative splicing rates, and read coverage to evaluate cleavage and polyadenylation activity, thus matching tissue-specific patterns of AS and IPA. This approach allowed estimation of the true IPA rate not confounded by low read coverage in introns, leading to a striking observation that intronic PASs are utilized more frequently than their exonic counterparts. In inspecting the concordance between IPA and AS, a considerable number of novel IPA events were found, ones that are inconsistent with the established models and can be attributed to byproducts of

the dynamic coupling between CPA and AS. These IPA events are termed here as spliced polyadenylated introns (SPI). The main hypothesis behind SPIs is that they are generated by the spliceosome removing the intron while it is being cleaved and polyadenylated.

The structure of the dissertation unfolds as follows: After this introductory chapter, Chapter 2 provides a thorough review of the literature and sets the background necessary for the understanding of the subsequent sections. Chapter 3 delineates the specific objectives of the thesis articulating the research questions this work aims to answer. Chapter 4 describes the materials and methods. Chapter 5 constitutes the main body of the dissertation and presents the results. It is divided into two sections; the first one focuses on identifying PASs from RNA-seq data and highlights how the increased dataset size enhances the sensitivity of the polyA reads-based approach. The second section characterizes the identified PAS clusters expressed in GTEx tissues and presents a simultaneous tissue-specific analysis of alternative splicing and intronic polyadenylation, which reveals events inconsistent with accepted IPA models. The section ends with additional corroborative evidence from other datasets. The dissertation concludes with Chapters 6 and 7, which are devoted to the discussion of results and conclusions.

Chapter 2

Background

Most eukaryotic precursor mRNAs (pre-mRNAs) undergo a series of critical RNA processing events that dictate their function and fate. These events include 5'-end capping, where the pre-mRNA receives a 7-methylguanosine cap, splicing, which removes introns while ligating the remaining exons, and 3'-end processing, which involves endonucleolytic cleavage and polyadenylation (CPA) at specific sites [158]. While these processing steps have historically been investigated as isolated phenomena, emerging evidence underscores their interdependence and reciprocal regulation [71].

This dissertation focuses on the interplay between intronic CPA and splicing. Accordingly, the literature review starts with a comprehensive description of the machinery and mechanisms governing cleavage and polyadenylation. It is followed by a discussion on the prevalence, outcomes, and regulation of APA. The section concludes with an overview of the current methodologies used for identifying the polyadenylation sites. The second part of the review briefly describes the molecular mechanism of splicing, processing of the excised intron lariat, and types of alternative splicing events. The review concludes with a focus on existing evidence and proposed models of the reciprocal regulation between splicing and 3'-end processing.

2.1 Cleavage and polyadenylation

The majority of transcripts that are generated by the eukaryotic RNA Polymerase II undergo endonucleolytic cleavage and subsequent polyadenylation at specific sites called polyadenylation sites (PASs). The PASs are defined by surrounding RNA sequence elements, that are generally conserved across metazoans but exhibit notable variations in yeast and plants [158]. Nonetheless, PAS can vary significantly across species: only about 10% of human sites are conserved in mammals [171]. Strikingly, metazoan histone pre-mRNAs are subjected to 3'-end cleavage without polyadenylation and, thus, require specialized 3'-end processing machinery.

While the core protein factors for polyadenylation are conserved across eukaryotes, significant differences in protein composition and subcomplex organization exist between mammalian and yeast systems. Notably, despite reconstituted active machinery for human histone and yeast canonical systems, the structure of active mammalian CPA machinery remains unclear, which is a critical gap in our understanding of human CPA mechanisms [151].

2.1.1 The molecular mechanism of cleavage and polyadenylation

In mammals, CPA is normally performed by several complexes including Cleavage and Polyadenylation Specificity factor (CPSF), Cleavage stimulation Factor (CSTF), cleavage factor I (CFI) and cleavage factor II (CFII) (Figure 2-1). Studies have revealed that the core factor CPSF consists of two distinct components: the mammalian polyadenylation specificity factor (mPSF) and the mammalian cleavage factor (mCF). The mPSF is responsible for recognising the polyadenylation signal and attracting the poly(A) polymerase (PAP) to execute polyadenylation. It consists of four subunits: CPSF160 (also known as CPSF1), CPSF30 (CPSF4), WDR33, and FIP1 (factor interacting with PAP). mCF catalyzes the cleavage process and comprises three subunits: CPSF73 (CPSF3), CPSF100 (CPSF2), and Symplekin. CPSF73 serves as the endonuclease for the cleavage process. A conserved region of CPSF100 interacts with WDR33 and CPSF160 and holds mCF and mPSF to-

gether, the region is referred to as PIM for "PSF Interaction Motif". Interestingly, mCF is functionally analogous to the histone pre-mRNA cleavage complex (HCC), meaning that both systems have similar cleavage components [151, 158]. This fact is currently being used to predict the structure of the activated mCF since the latter has not been resolved yet.

Other structural data suggest that the combination of CPSF160 and WDR33 serves as the central element of the machinery [191]. This core duo is essential for interactions of CPSF30 identifying the polyadenylation signal, mCF excising the pre-mRNA and CSTF recognizing another *cis*-regulatory element. Subsequently, CPSF30 attracts FIP1, which interacts with PAP to carry out polyadenylation of the mRNA.

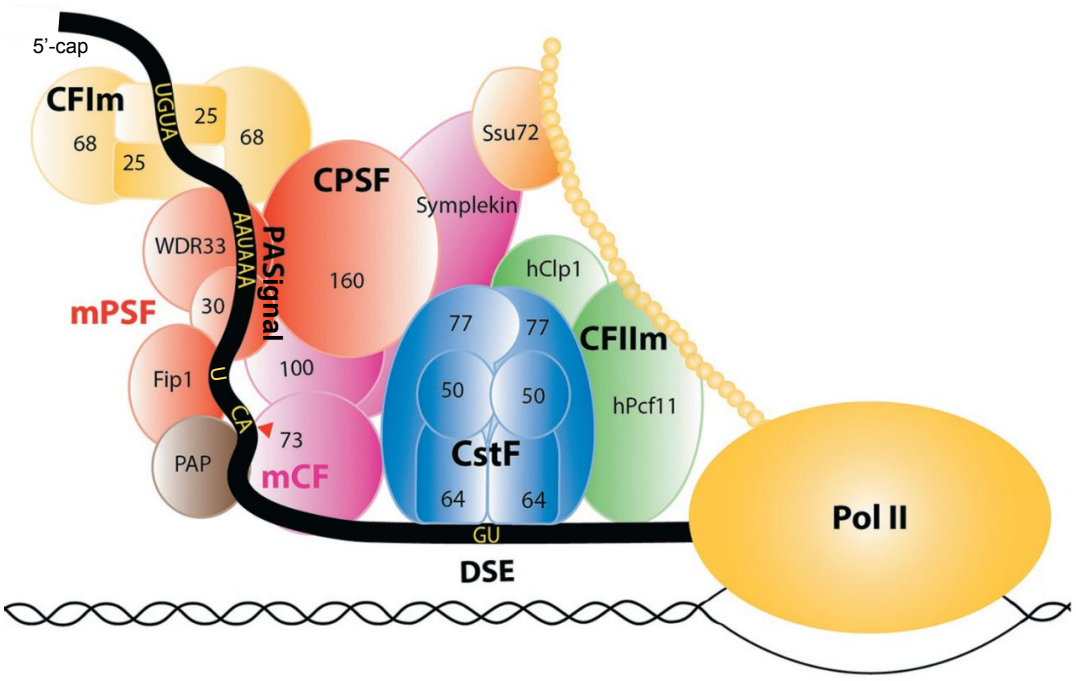


Figure 2-1: **Cleavage and polyadenylation complex**. Schematic of Human Pre-mRNA 3'-End Processing Machinery. The diagram depicts CPA factors attached to pre-mRNA and the Pol II C-terminal domain during transcription. Various CPA protein complexes are colour-coded: mPSF (red), mCF (magenta), CSTF (blue), CFI (yellow), and CFII (green). The beaded yellow chain represents the CTD of Pol II, the red triangle denotes the cleavage site, DSE stands for "downstream sequence element". Figure from [151].

In metazoan genes, PASs are primarily defined by the hexameric consensus motif AAUAAA also known as the canonical polyadenylation signal. According to recent

structural studies, WDR33 and CPSF30 recognise the signal. WDR33 interacts with the U3 and A6, which appear to be the most conserved residues in the motif [152]. In humans and mice, the motif is positioned on average 21 nucleotides upstream of the cleavage site (Figure 2-2B). About 70% of the PASs are preceded by the canonical motif, 14% by the AUUAAA motif, and ∼15% contain single-nucleotide variants of one of the two hexamers [151]. Notably, the signal may be located far from the cleavage site as secondary-structure elements in the pre-mRNA can bring them together [175]. The genome-wide frequency of signal variants correlates with *in vitro* CPA efficiency at their adjacent PAS [144]. Accordingly, the canonical signal is enriched at stronger distal PAS compared to the weaker proximal ones. These observations led to a hypothesis that signal sequences of individual genes have been selected in evolution to allow CPA to proceed with specific kinetics or efficiency [52]. On the other hand, a study about selection acting on human polyadenylation signals found no evidence of negative selection against mutations improving the CPA-stimulating efficiency of non-consensus polyadenylation signals [73].

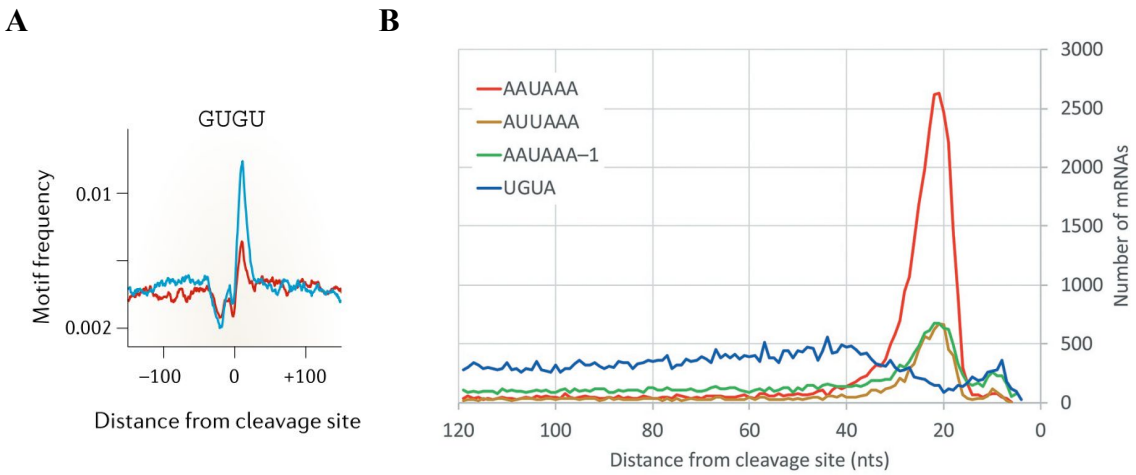


Figure 2-2: **Distribution of sequence elements at 3'-ends of human pre-mRNAs**. **(A)** Frequencies of the GUGU motif, often referred to as DSE, near the proximal (red) and distal (blue) PAS located in the same terminal exon. Figure from [52]. **(B)** The number of known human mRNAs with AAUAAA (red), AUUAAA (yellow), single-nucleotide variants of AAUAAA (green, labelled as AAUAAA-1), and UGUA (blue) motifs at specified positions. Figure from [151].

Other complexes in the canonical CPA machinery are CSTF, CFI and CFII. CFI

consists of CFIM25 (also CPSF5, encoded by *NUDT21*) and CFIM68 or CFIM59 (CPSF7). This complex has a particularly strong impact on the PAS choice. CFIM25 binds the upstream UGUA motif (Figures 2-1, 2-2B), which is substantially more enriched near distal PASs compared to the proximal ones. Thus, CFI associates preferentially with distal PASs in terminal exons and enhances their usage [13]. Some studies also suggest that CFIM68 interacts with FIP1 [197].

CSTF is required for cleavage but is dispensable for polyadenylation. It is composed of four types of subunits: CSTF50, CSTF77, CSTF64 and its paralogue τCSTF64. Both CSTF64 and τCSTF64 recognize U- or GU-rich dowstream sequence element in the pre-mRNA (Figure 2-2A), and their upregulation is associated with increased usage of weaker PAS [176, 91].

CFII is the least studied factor of the canonical machinery. It consists of CLP1 and PCF11. The latter has a C-terminal segment essential for cleavage and the RNA Polymerase II CTD interaction domain. In yeast, factors homologous to CSTF and CFII are required for strong endonuclease activity of the cleavage and polyadenylation core, but not for its specificity [60].

In addition to the polyadenylation signal, other cis-regulatory sequence elements can enhance 3'-end processing. These elements are recognized by various components of the CPA machinery. Such sequence elements include the previously mentioned UGUA motif and the GU- or U-rich DSE. Notably, the DSE can consist of either GU- or U-repeats, both of which are recognized by the CSTF complex [187, 157]. The cleavage sites are typically followed by an A nucleotide (about 80% of PAS) and preceded by a C, U, or G nucleotide, with a slight preference towards C [151] (Figure 2-1). This dinucleotide is preceded by a U-rich region, which interacts with the FIP1 subunit of the CPSF complex [77].

The current model of the CPA process in mammals is based on studies of the yeast cleavage and polyadenylation factor (CPF) complex. It states that the factors CPSF, CSTF, CFI, and CFII, recognise and bind their respective RNA sequences independently. Then, they assemble into a cleavage-competent complex after structural rearrangements of the RNA [60]. However, the particularities of the CPA process in humans, such as the mechanism of the CPSF complex activation, are not

fully understood.

2.1.2 Alternative polyadenylation

About 70% of human genes have multiple PASs resulting in alternative polyadenylation (APA) [63, 29]. Although APA expands the transcriptome diversity, its evolutionary benefit is debatable. Some authors believe it to be generally deleterious [178]. Still, about 40% of human genes have APA events conserved in mammals [171]. APA can generate transcripts with different 3'-untranslated regions (3'-UTR) and transcripts encoding proteins with different C-termini [114] (Figure 2-3). The latter isoforms are a result of intronic polyadenylation upstream of the 3'-most exon. Studies have shown that more than 20% of human genes contain at least one intronic PAS [159].

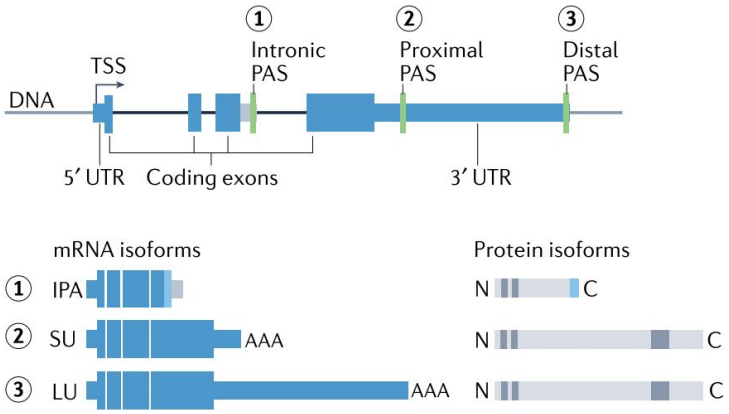


Figure 2-3: **CPA positions in a pre-mRNA and the corresponding isoforms**. SU stands for short 3' UTRs and LU - for long 3' UTRs. The light blue colour shows alternative coding regions. The grey boxes in the protein symbols represent protein domains. Figure from [114].

3'-UTR APA modulates gene expression by influencing mRNA stability, translation, nuclear export, subcellular localization, and interactions with microRNAs and RNA binding proteins (RBPs) [52, 106]. Moreover, some processes are associated with widespread shifts in 3'-UTR length in genes with tandem PASs in the same terminal exon. For example, the induction of pluripotency in somatic cells is associated with the shortening of 3'-UTRs, whereas differentiated cells tend to have longer ones [149]. Also, in mice and humans, the length of 3'-UTRs was shown to

vary across tissues [52]: brain and blood tissues are characterized by long and short 3'-UTRs, respectively [169]. In neurons, isoforms with long and short 3'-UTRs are over-represented in the soma and neurites, respectively [105]. Other studies showed that genes with tandem PASs adjust their isoform ratios to tissue-specifically regulate protein expression [92, 28].

2.1.3 Intronic polyadenylation

CPA events in an intron upstream of the stop codon, also termed intronic polyadenylation (IPA), are less studied than tandem PASs. One reason for this might be that intronic PASs tend to have more tissue-specific distribution and, thus, require an analysis of many diverse tissues and conditions [53]. Also, IPA events are less conserved among species compared to other APA events [159, 171]. About 40% of genes in the mouse genome are already known to undergo IPA [63]. In a recent study, more than 100 thousand novel PAS were identified in human introns [50, 53]. These observations speak to the immensity of the IPA landscape, which is still being discovered.

IPA events generate truncated transcripts with alternative terminal exons (ATEs). These ATEs can partially match one of the internal exons of the original isoform (Composite Terminal Exon (CTE)) or be an omitted exon expressed only in the truncated transcript (Skipped Terminal Exon (STE)) (Figure 2-4). The subtype depends on the class of the alternative splicing event associated with IPA: CTE is an example of retained intron, while STE represents a cassette exon (2.2).

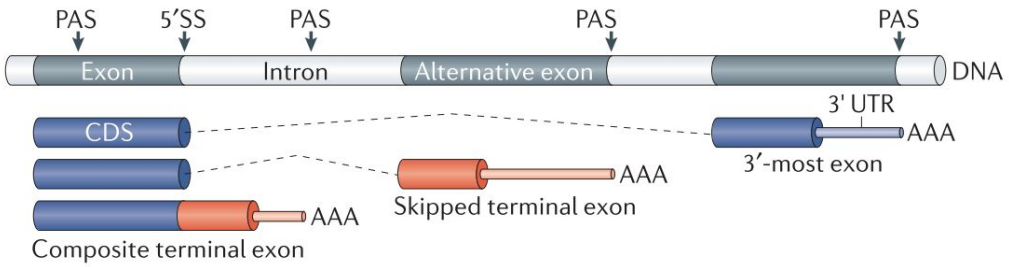


Figure 2-4: **Types of IPA events**. IPA generates isoforms with alternative terminal exons. The use of the PAS in the cassette exon generates a transcript containing an STE (middle). Cleavage at the PAS within a retained intron results in a transcript with a CTE (bottom). Figure from [158].

IPA can lead to important functional changes due to alterations in the protein amino acid sequence [31]. For example, IPA is a mechanism of switching from a membrane-bound form of the protein to the soluble one, which was primarily observed decades ago on the immunoglobulin M (*IgM*) gene during the activation of B cells [137]. The same mechanism was recently shown to work for the vascular endothelial growth factor receptor 2 (*VEGFR2*) gene, and more than 350 other genes were predicted to be regulated in a similar way [27, 164]. Another recent example is IPA in the *DICER* gene that generates a truncated protein with impaired miRNA cleavage ability and results in decreased endogenous miRNA expression [88]. Additionally, IPA generates truncated isoforms of oncosuppressor proteins, such as MGA and NUP98, that often lack tumour-suppressive functions or even contribute to the tumour onset and progression [88, 80]. Overall, all types of APA are widely implicated in human disease, including haematological, immunological, neurological disorders, and cancer [52, 23, 38].

2.1.4 The role of alternative cleavage and polyadenylation in disease

Defects in CPA *cis*-regulatory elements, particularly polyadenylation signals, have been implicated in various diseases through their role in APA [122]. For example, SNP in the canonical polyadenylation signal of *TP53* gene leads to its reduced expression and has been linked to increased susceptibility to prostate cancer, glioma, and colorectal adenoma [150]. In the genetic disorder α-thalassemia, a mutation in the canonical signal of the haemoglobin subunit alpha 2 (*HBA2*) gene weakens the polyadenylation signal and results in elongated transcripts that are often degraded (Figure 2-5). Similar processes are observed in disorders like β-thalassemia, bone fragility disorder and IPEX syndrome [122].

Mutations in non-canonical signals also can lead to disease-specific outcomes, as seen in Fabry disease where mutations in the α-galactosidase A (*GLA*) gene alter protein localization [182]. Other conditions like systemic lupus erythematosus and syndromic microphthalmia experience similar shifts in transcript isoform ra-

tios due to mutations in the polyadenylation signal [122, 69]. Mutations can also create new cryptic PAS, as seen in mantle cell lymphoma, Huntington's disease, and FSHD [122]. In amyotrophic lateral sclerosis, RNA editing introduces sequence variations that affect APA [42].

Alterations in other *cis*-regulatory elements can also affect PAS strength. Notable examples include thrombosis and thrombophilia, linked to mutations in the cleavage site and DSE, respectively, of the coagulation factor II (*F2*) (Figure 2-5). This gene-specific regulation mechanism stems from an unusual gene architecture, in which an upstream sequence element recognised by CFI compensates for the weak activity of the DSE and the noncanonical cleavage site (CG instead of CA) [52, 24].

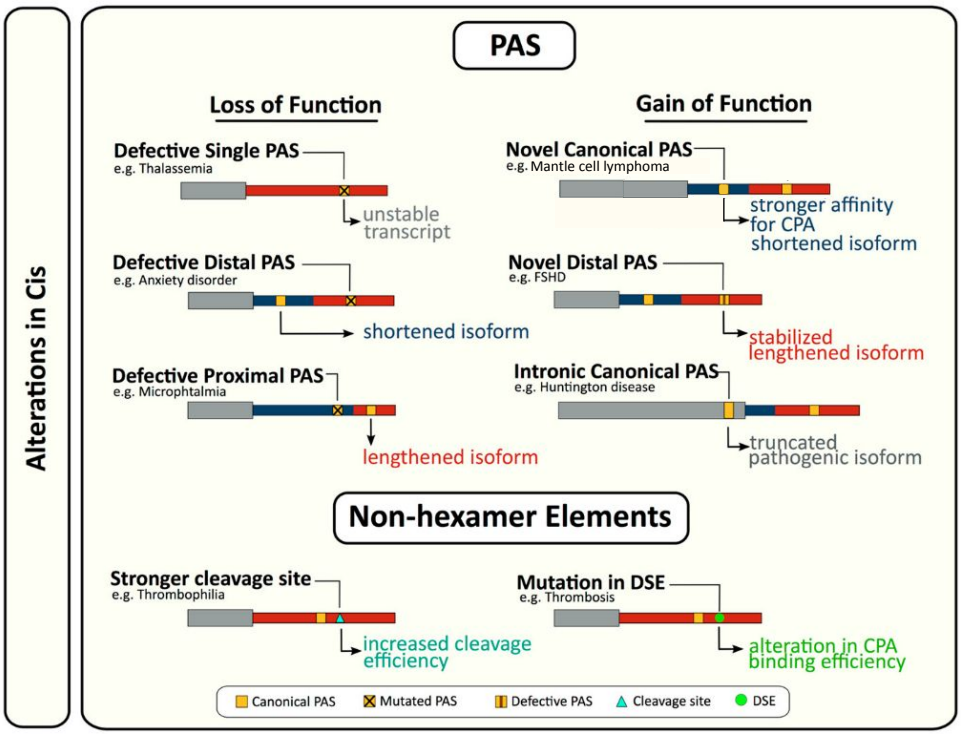


Figure 2-5: **Examples of alterations in *cis*-regulatory elements resulting in disease**. Loss-of-function alterations destroy or weaken a PAS. Gain-of-function mutations introduce new CPA sites or increase CPA efficiency at an existing PAS. Note: In this figure, "PAS" refers to the polyadenylation signal, not the polyadenylation site. Figure from [122]

Overall, genome-wide association studies have identified multiple variants in the 3'-UTRs associated with changes in APA ("apaQTL"), and up to 19% of them were associated with known human traits or diseases [89]. Surprisingly, these apaQTLs

were found to rarely overlap with expression QTLs, indicating that they contribute to diseases without significantly altering the mRNA abundance [89, 114].

Alterations in genes of core CPA machinery can impact PAS selection and lead to diseases (2.1.5). For instance, in oculopharyngeal muscular dystrophy (OPMD), an extended (GCG)-repeat in the *PABPN1* gene creates trePABPN1, which disrupts normal PABPN1 function and causes unregulated cleavage at proximal PASs [52, 15] (Figure 2-6). Neuropsychiatric diseases are often associated with elevated CFIM25 expression, which reduces levels of *MECP2* transcription factor [122](Figure 2-6). In cardiac hypertrophy, CSTF upregulation shortens 3'-UTRs [124]. Moreover, mutations in lesser-known CPA factors, like the CLP1 subunit of CFII, can disrupt tRNA biogenesis and potentially lead to neurodegenerative disorders [75, 142].

A broad remodelling of 3'-UTRs was observed in cancer cell lines. Numerous tumour types, including colorectal, gastric, and breast cancers, as well as neuroendocrine and hepatocellular carcinomas, neuroblastoma, and glioblastoma, exhibit a noticeable trend towards the usage of proximal PASs [176, 122, 51, 107, 180]. These data agree with the observations of 3'-UTR lengthening with cell differentiation. Moreover, genome-wide shortening of transcript isoforms has been linked to adverse outcomes in conditions like breast and lung cancers, as well as pancreatic ductal adenocarcinoma and neuroblastoma [52, 147, 114].

These patterns can partially be attributed to alterations in the expression levels of 3'-end processing factors. For instance, the upregulation of CSTF64 and PCF11 as well as a decrease in CFIM25 or PABPN1 levels have been linked to proximal PAS usage in cancer data [122, 52, 176, 114] (Figure 2-6). It is hypothesized that by shortening the 3'-UTRs, cancer cells can circumvent the inhibitory effects of microRNAs and RBPs. This is particularly significant since more than 70% of genes contain conserved microRNA target sites or destabilizing AU-rich elements within their 3'-UTRs [122].

More recent large-scale studies contradicted the established trends and highlighted the complexity of the landscape of 3'-UTR alterations in cancer [180]. Moreover, APA changes are often highly specific to both the type of tumour and individual patient characteristics. In chronic lymphocytic leukaemia, for example,

widespread intronic CPA results in numerous truncated proteins [88]. In contrast to most other studied cancers, chronic myelocytic leukaemia displays unusually high levels of CFIM25. Depletion of this factor in K562 leukaemia cells has been shown to inhibit growth and proliferation [190]. In sum, the role of APA in cancer encompasses not only global CPA shifts but also gene-specific mutations. Both types of changes can be pivotal in driving tumour development and influencing clinical outcomes.

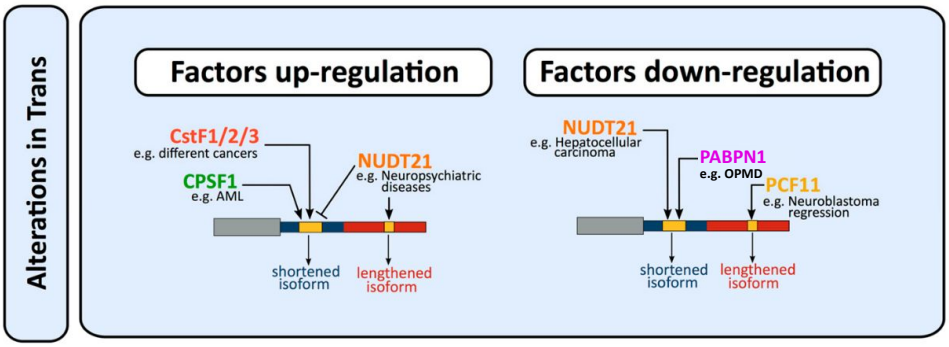


Figure 2-6: **Global alterations in the CPA factors as a significant feature of diseases.** The concentration of CPA factors as APA determinant. The left (right) panel shows the consequences of factor upregulation (downregulation, respectively). For example, NUDT21 (CFIM25) upregulation is associated with enhanced distal PAS usage. In neuroblastoma samples with low PCF11 levels, transcript lengthening and tumour regression are observed. Note: in the main text, CSTF2 is mostly referred to as CSTF64. Figure adapted from [122].

2.1.5 Regulation of alternative polyadenylation

Both gene-specific and global features can influence polyadenylation patterns across the genome. Specifically, the deregulation of CPA factors induces genome-wide shifts in APA. In proliferating cells, transcript shortening is linked with elevated expression of CSTF, CFII factors, and PAP [122]. Consistent with this, the CFII subunit PCF11 promotes cleavage at proximal PAS [91], and the simultaneous depletion of CSTF64 and τCSTF64 enhances the selection of distal PAS in almost 500 genes in HeLa and LN229 cells [66](Figure 2-6). The downregulation of the CFI subunits CFIM25 (encoded by *NUDT21* gene) and CFIM68 results in the shortening of 3'-UTRs in numerous genes [66]. Decreased CFIM25 levels are common in solid tumours like glioblastoma, hepatocellular carcinoma, and bladder cancer, contributing

to the aforementioned 3'-UTR shortening in cancers [51, 13, 104]. The factor FIP1 globally promotes cleavage at proximal PAS [91]. And the Poly(A) Binding Protein N1 (PABPN1), controlling the length of the polyA tails, interacts directly with pre-mRNAs near intronic or TSS-proximal PASs to block their cleavage [91, 67].

The concentration of many subunits of the canonical machinery during proliferation is, in turn, regulated by E2F transcription factors [37]. Remarkably, the IPA rate determines the expression of the CPA factors PCF11 and CSTF77 through autoregulation: the rate of cleavage at PASs in the 3rd intron of CSTF77 and the 1st intron of PCF11 is controlled by CPA factors levels [99, 170].

Several RNA-binding proteins recognizing sequence elements in the gene body are also involved in the regulation of APA. Among them there are splicing factors SRSF3 and SRSF7, NOVA1, PCBP1, members of the hnRNP family and ESRP1 [195, 52] (Figure 2-4). Polypyrimidine tract-binding proteins (PTBPs) are known to compete with canonical splicing factors to repress the splicing of target exons. However, they can also promote CPA at an intronic PAS located upstream of their binding site [97]. In contrast, when PTBP1 binds up to 75 nts downstream of a distal PAS, it can mask the site from cleavage [51] (Figure 2-4). One proposed mechanism is that PTBP1 blocks the downstream binding of CPA factors such as CSTF or CFII. A similar mechanism has been demonstrated for Hu family proteins, which impact CPA by competing with CSTF for access to the downstream sequence element [52]. Many other splicing factors including members of the hnRNP family (HNRNP H1, HNRNP H2, HNRNP L, HNRNPC, etc.), neuro-oncological ventral antigen (NOVA) and epithelial splicing regulatory protein 1 (ESRP) are known to alter CPA positions. However, their effect appears to be more complex, and the underlying regulatory mechanisms remain elusive [195, 114]. This suggests that their influence might be indirect and operate through splicing alterations.

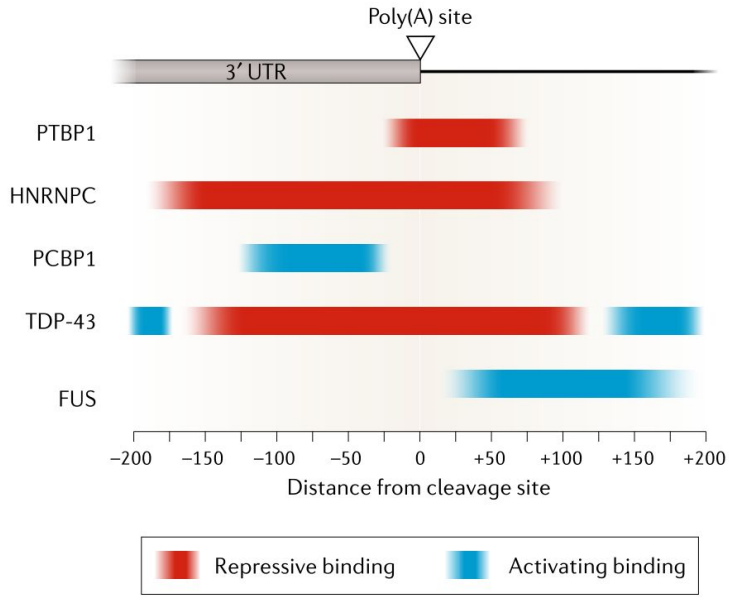


Figure 2-7: **Impact of RBPs on pre-mRNA CPA**. Figure from [52].

The dynamics of RNA Polymerase II elongation also play a significant role in controlling APA. In genes with multiple PASs, a slower transcription rate or Pol II pausing promotes the use of the upstream cleavage site by extending the period during which it is the only available option. As the CPA machinery interacts with Pol II CTD, pausing may enhance its contact with the pre-mRNA and promote the cleavage. Accordingly, studies in yeast and *Drosophila m.* showed that a substantial decrease in the transcriptional elongation rate leads to a moderate increase in proximal, weaker PASs usage, both within 3'-UTRs and introns [94, 45]. Also, a localized slowing of Pol II downstream of a PAS is associated with CPA at the site. Polymerase pausing can be induced by specific sequences, transient DNA/RNA structures (such as G-quadruplexes) and chromatin environment; all these factors have been shown to affect APA [49, 6, 114, 119]. However, the direction of the regulation can be ambiguous: CPSF and CSTF were shown to promote Pol II pausing [126].

Global regulation by CPSF subunit FIP1 presents an interesting example here: it depends on the distance between the two PASs. If the sites are far from each other, and there is a significant lag between the times when they are transcribed, higher levels of FIP1 promote the recognition of the weaker proximal PASs. However, in the case of closely located PAS, in addition to the smaller time window advantage,

FIP1 binding upstream of the distal PAS blocks access to the proximal one, thus obstructing its usage [85].

CTD-associated elongation factors and the CTD phosphorylation status regulate Pol II with respect to its relative location in the gene and influence Pol II speed. These factors include several cyclin-dependent kinases (CDK12, CDK9 and CDK13), phosphatases (SSU72, FCP1, PP2A), and the PAF1C complex [114, 181]. For instance, the inhibition of CDK12 leads to an increased probability of using cryptic intronic PAS, due to the slowing of productive Pol II elongation [80, 35]. CDK9 affects PAS usage through deregulation of transcription termination [83]. SR-related CTD-associated factors SCAF4 and SCAF8 suppress cleavage at intronic PASs [48]. SSU72, a Pol II CTD phosphatase, interacts with Symplekin and PCF11 during CPA [82, 177]. Notably, its roles in CPA and Pol II regulation are independent [81, 64].

In conclusion, beyond the canonical CPA machinery, PAS determination is controlled by many splicing factors, chromatin remodelers and Pol II CTD-associated elongation and termination factors. This highlights the tight coordination between CPA and both transcription and co-transcriptional processes.

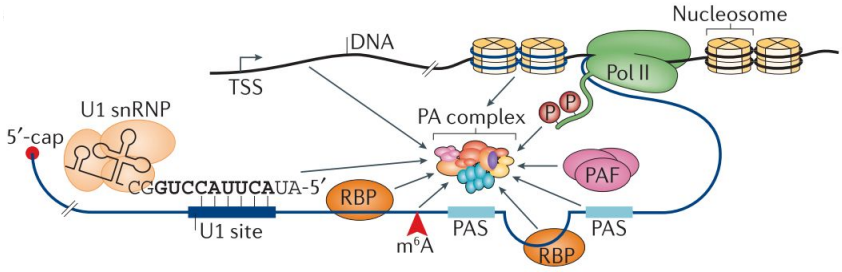


Figure 2-8: **Multifaceted regulation of APA**. PAS selection is shaped by factors such as the transcription start site, recruitment of CPA factors and other RBPs, nucleosome density around the PAS, Pol II elongation rate, RNA methylation, and 'telescripting' involving U1 snRNP. Figure from [158].

Splicing adds a layer of regulation to APA events. Studies have shown that the splicing machinery, particularly spliceosome subunits like U1 and U2 snRNP, can both suppress and facilitate specific CPA events. The proposed mechanisms behind this will be further explored in a specific section 2.3.

Finally, recent studies discovered numerous examples of post-transcriptional 3'-UTR cleavage and sequential CPA in the nucleus, thus, revealing novel types of the APA control [101, 102, 155].

In summary, APA is regulated globally by levels of the CPA factors, and gene-specifically by the *trans*-acting factors involved in splicing, transcription initiation, elongation and termination. The PAS choice depends on both local and global Pol II elongation speeds, chromatin environment, and DNA looping and can be altered post-transcriptionally (Figure 2-9). This complexity results in the specific PAS sets for each cell state, cell type and organism and calls for precise and accessible methods for PAS identification.

2.1.6 Polyadenylation sites identification and quantification

Various experimental methods have been developed to pinpoint the genomic locations of PASs [19, 52]. Techniques like 3'RNA-seq, PAS-seq, polyA-seq, and 3'READS commonly utilize oligo(dT) or analogous primers to selectively isolate transcript ends [186, 146, 29, 92, 194, 63]. For example, the 3'-seq method, introduced by the group of Dr. C. Mayr in 2013, was designed to quantitatively profile 3'-UTR isoforms [92]. This approach uses an oligo(dT) primer with uridine and a VN-anker to capture the junction between the terminal exon and the poly(A) tail. A subsequent nick at the uridine, followed by a 50-75 nucleotide shift, ensures capturing a sufficiently long mRNA fragment upstream of the CPA site (Figure 2-10).

Several databases, such as `APADB`, `APASdb`, `PolyA_DB` and `PolyASite`, collate PAS from various species, identified through the 3'-end sequencing techniques [29, 118, 58, 185, 87, 168]. Some databases are under regular updates, with the 2021 version of PolyASite hosting over 500,000 human PAS. Nevertheless, many more PASs could be active in specific tissues, cell states and abnormal conditions. Concurrently, massive polyA(+) RNA-seq datasets for different cellular contexts are becoming increasingly available, thus, accelerating the development of methods to identify and evaluate PASs from RNA-seq data [184, 14, 18].

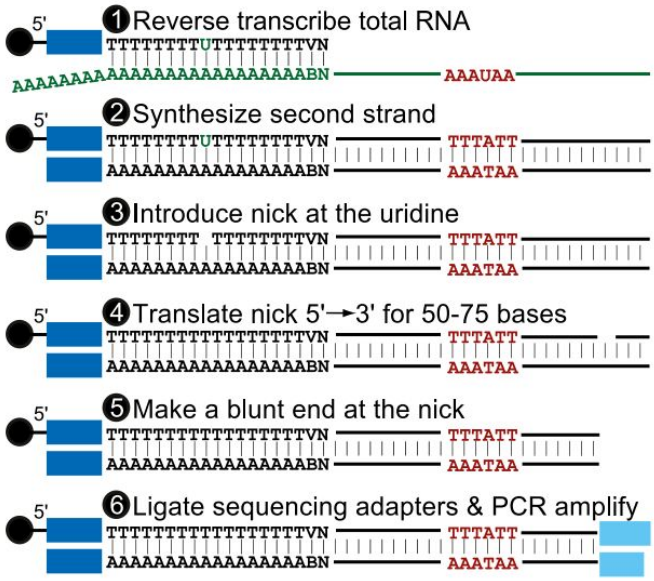


Figure 2-9: **3'-Seq protocol**. Total RNA is reverse-transcribed with an oligo(dT) primer containing one uridine, the primer is ligated to a sequencing adapter bound to a magnetic bead. After the second strand is synthesized, a nick is introduced at the uridine and then shifted at least 50nt away from the 3'-end. A blunt end is created at the new position of the nick. The ligation of the second sequencing adapter is followed by PCR, gel purification and sequencing. Figure from [92].

A PAS can be identified in standard polyA(+) RNA-seq data as a genomic locus exhibiting an abrupt decrease in read coverage (Figure 2-10). Most published tools, including `DaPars`, `GETUTR`, `IsoSCM`, `APAtrap`, and `TAPAS`, employ this approach [176, 172, 79, 145, 183, 3]. Since the density of the RNA-seq reads is noisy and highly non-uniform along the gene length, most of these tools are limited to the PASs in the terminal exon and can not detect novel intronic sites. Three more recent programs, `mountinClimber`, `IPAFinder`, and `Aptardi`, integrated RNA-seq read coverage with the information about splice junctions, either annotated or identified *de novo*. This integration enabled the prediction of the intronic PAS [16, 193, 100].

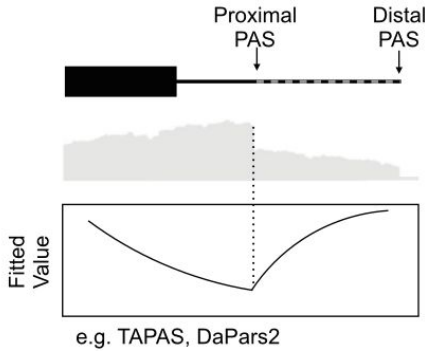


Figure 2-10: **A generalized algorithm of the coverage-based *de novo* PAS identification from RNA-seq data**. The method finds a genomic locus where the read coverage abruptly decreases. The middle panel depicts the RNA-seq coverage across the terminal exon extended by a variable length specific to the tool. The bottom panel's y-axis represents the fitted value of the regression model, and the point of the minimum value corresponds to the predicted PAS location. Figure from [143].

In 2023, the APAeval community reviewed and benchmarked the coverage-based methods using RNA-seq datasets with matching 3'-end sequencing data from two human and two mouse cell types. Each of the nine selected tools was applied to the RNA-seq datasets. The PAS predictions obtained were then compared against the annotated transcript ends from GENCODE. Subsequently, the tools were ranked based on precision and recall metrics. `TAPAS` demonstrated the best performance with ∼25% precision and ∼65% recall values on a mouse dataset containing about 480 million raw reads [14]. Again, `TAPAS` performed best and maintained these levels of precision and recall when the PAS from 3'-end sequencing data were used as the ground truth. Unfortunately, the authors reported critical technical problems in the IPA-centered tools `mountinClimber` and `Aptardi`; `IPAFinder` was not evaluated in the study [14, 18].

Additionally, 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 and are often referred to as "polyA reads". 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 [153, 184]. Moreover, the polyA read-based approaches have the advantage of determining the precise PAS locations. A couple of tools

have applied this method: an RNA-seq aligner `ContextMap2` [10] and `KLEAT` [9]. The latter employs polyA reads to improve transcript assembly and, consequently, identify the PASs. It is worth mentioning that any transcriptome assembler such as `Cufflinks` [161] or `Trans-ABySS` [135] predicts transcript ends and, thus, the PASs.

The authors of `ContextMap2` evaluated both polyA-read-based tools using six RNA-seq samples, each containing about 250M raw reads [10]. `KLEAT` and `ContextMap2` predicted around 15,000 and 7,000 PASs, respectively. These sites were then compared with PAS sets obtained from the same samples via RNA-PET, a protocol designed to capture the 3' and 5' ends of transcripts. The reported recall values for both tools were less than 12%, and the precision values ranged from 65% to 81% for `KLEAT` and from 75% to 94% for `ContextMap2`. An independent benchmark analysis by the authors of `KLEAT` yielded similar results [180]. Remarkably, they also evaluated `DaPars`, which showed considerably lower precision compared to the other two tools. The low sensitivity of the polyA-read-based approach initially led to the conclusion that it is ineffective for the PAS identification [18]. However, this method is characterized by relatively good precision [180, 14] and can offer a powerful alternative to coverage-based methods, particularly when analyzing a sufficiently large panel of RNA-seq experiments.

Based on	**Intronic PAS identification**	**Name and publication year**	**Ref.**
Read coverage drop	No	DaPars (2014, 2018)	[176, 40]
		ChangePoint (2014)	[172]
		GETUTR (2015)	[79]
		IsoSCM (2015)	[145]
		APAtrap (2018)	[183]
		TAPAS (2018)	[3]
Read coverage drop	Yes	mountinClimber (2019)	[16]
		IPAFinder (2021)	[193]
		Aptardi (2021)	[100]
polyA reads	Yes	KLEAT (2015)	[9]
		ContextMap2 (2017)	[10]

Table 2.1: **Tools for *de novo* PAS identification from RNA-seq data.** Methods successfully evaluated by APAeval community are underlined [14].

For studies of APA, it is crucial to link PAS locations with the expression levels

of the respective transcript isoforms. Therefore, there is a growing emphasis on tools that quantify the usage of the individual PASs within the transcriptome from RNA-seq data. Due to the scarcity of polyA reads, all approaches utilize the read coverage in some form. Several methods integrate *de novo* PAS identification with the isoform quantification [16, 40, 79, 3, 183, 193, 100]. Other tools like `Roar` [47], `QAPA` [55], `PAQR` [51], `APAlyzer` [169], and `MISO` [76], focus on quantifying PAS usage and often are applied to PAS sets from 3'-end sequencing databases described above. These tools usually depend on the annotation, requiring not just the PAS but also the genomic coordinates of the terminal exon, which can be identified by `TECtool` [50]. Previous studies extensively characterized tissue-specific polyadenylation patterns using the coverage-based approach [16, 62, 169, 55, 40].

Long-read and whole-transcript sequencing data, commonly used for transcriptome assembly, are also applicable for identifying polyadenylation sites as transcript ends [1]. However, such datasets are not typically used for *de novo* PAS identification. A recent preprint introduced the first tool for PAS identification from long-read RNA-seq data [17]. The study compared several datasets of similar size generated with various long-read and 3'-seq protocols and showed that the 3'-seq protocol is more effective at detecting PAS, particularly in genes with low expression levels. 3'-seq-derived PAS set contained about 97% of all PAS identified by the other protocols. Notably, another comparison between PacBio Iso-Seq and 3'-seq found 3'-seq to be more proficient in identifying PAS in introns [143]. However, long-read sequencing has lower mapping error rates and offers a broader range of applications. It is instrumental in transcript isoform identification, estimating polyA tail lengths [95, 65], and analyzing the coordination between splicing and polyadenylation, as evidenced by several studies [192, 131].

In summary, methods for PAS identification and quantification have made significant progress with the development of numerous experimental techniques and computational tools. However, challenges remain, particularly in the accurate identification of intronic PASs and PASs specific to particular conditions.

2.2 Splicing

One of the cotranscriptional processes tightly connected to CPA is splicing. During splicing some parts of the pre-mRNA (introns) are removed and the remaining regions (exons) are ligated.

Splicing is catalysed by the spliceosome, a complex ribonucleoprotein machine comprising five small nuclear RNAs (snRNAs; U1, U2, U4, U5 and U6 snRNA) and approximately 200 proteins. The spliceosome forms stepwise on every newly synthesised intron, its components are recruited, in part, through base-pairing interactions between the snRNAs and short sequences in the pre-mRNA. The start of the intron, or 5'-splice site (5'SS), interacts with U1 snRNA at the very beginning of splicing. The 3'-splice site (3'SS) at the other end of the intron is preceded by the polypyrimidine tract and the branch point (BP) sequence that pairs with U2 snRNA (Figure 2-11). While 3'SS and 5'SS are represented by highly conserved GU and AG dinucleotides, other sequence elements are relatively loosely defined [30]. An exception is the BP adenosine, located 18-37 nts upstream of the 3'SS; it initiates a nucleophilic attack on the 5'SS during the first step of splicing [110]. In addition to snRNAs, the spliceosome contains proteins that associate with each snRNA to form snRNPs and a set of non-snRNP proteins.

In the course of splicing, the spliceosome catalyzes two transesterification reactions i.e. two replacements of one phosphodiester bond in the pre-mRNA for another (Figure 2-11). In the first reaction (also called the branching reaction), the 2'-hydroxyl group of the BP adenosine carries out a nucleophilic attack on the phosphate group in the phosphodiester bond in the 5'SS; the products are an intron lariat-3' exon intermediate and a loose 5' exon. During the second reaction (the exon-ligation reaction) the exposed 3'-hydroxyl of the freed upstream exon attacks the phosphodiester bond within the 3'SS, releasing the intron lariat and ligating the two exons [173, 30].

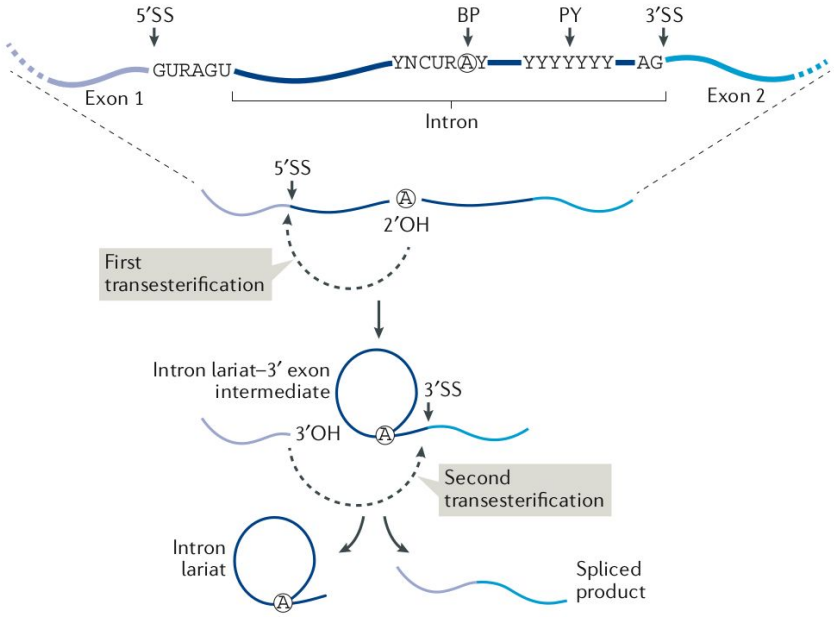


Figure 2-11: **The transesterification reactions in pre-mRNA splicing**. The scheme describes the two transesterification reactions, catalyzed by the spliceosome. In the figure N represents any nucleotide, R represents purine and Y represents pyrimidine. Figure from [30].

Splicing begins when U1 snRNP and three non-snRNP proteins bind the pre-mRNA. U1 snRNP interacts with 5'SS, factors SF1, U2AF1 and U2AF2 bind, respectively, to the BP sequence, the 3'SS and the polypyrimidine tract (Figure 2-12). Next, U2 snRNP displaces SF1 and binds the BP, interaction between U1 and U2 snRNPs loops the intron [22, 30] (Figure 2-12:1). The binding of U2 snRNP to the pre-mRNA triggers recruitment of the preformed U4/U6.U5 tri-snRNP (Figure 2-12:2). The spliceosome is then rearranged into its catalytically active state; in the process of transformation U1 and U4 snRNPs are released, and U6 snRNA and U2 snRNA form the catalytic centre of the machinery [173]. The BP adenosine, docked into the active site, performs the branching reaction (Figure 2-12:3). Then, while the 3'-end of the upstream exon remains in the active site, the 3'SS site substitutes the BP adenosine in the active site for the exon-ligation reaction. U5 snRNP aligns the two exons for the ligation [30]. When the splicing reaction is complete (Figure 2-12:4), the pre-mRNA and the remaining U2, U5 and U6 snRNPs are released from the intron lariat. Subsequently, the excised intron is degraded.

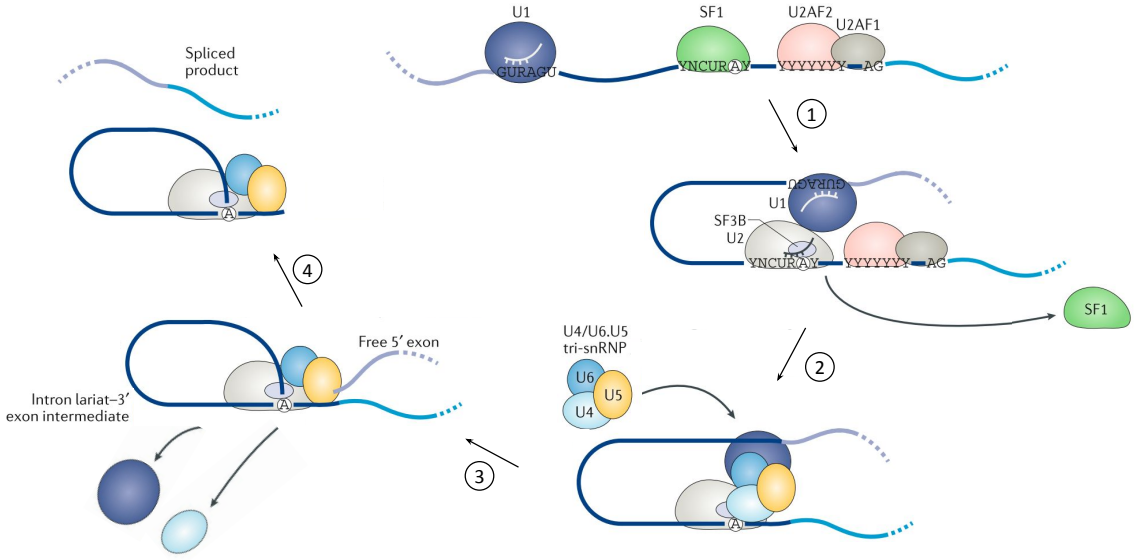


Figure 2-12: **The mechanism of the pre-mRNA splicing.** The spliceosome is assembled in an orderly manner, activated to form the active site and remodelled extensively to perform the branching and exon-ligation reactions. Figure adapted from [30].

The intron lariat degradation is also a stepwise process. Following excision, the 3'-tails of the lariats are shortened by 3'-exonucleases up to the lariat BP. Further exonucleolytic degradation requires cleavage of the 2'-5' bond formed by the 5'SS guanine and the BP adenosine. It is performed by the RNA debranching enzyme DBR1, a phosphodiesterase found in all eukaryotes [116]. Generally, the intron lariats are degraded within minutes of splicing and are low abundance RNAs in sequencing datasets [20, 120]. Surprisingly, a genome-wide screen for intronic lariats resulted in the identification of stable intronic RNAs in the oocyte nucleus of *X. tropicalis*. Later, additional studies identified cytoplasmic lariats in human, mouse, chicken, and zebrafish cells [154]. These RNAs were relatively short (up to 500 nucleotides) and contained an unusual cytosine branchpoint, an unfavourable target for the DBR1 recognition. Some of these lariats are actively exported from the nucleus [154]. Sno-RNAs and some miRNAs are also processed from introns excised by the spliceosome [120]. These examples indicate that while most intron lariats are instantly degraded, some have a different fate.

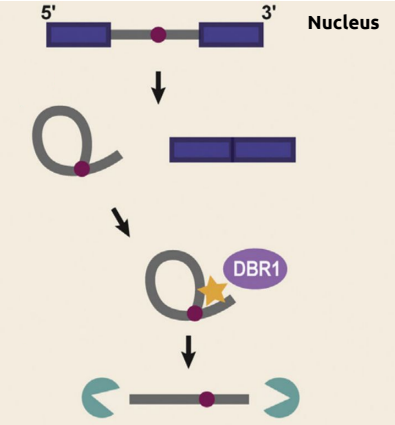


Figure 2-13: **Schematic representation of an intron lariat processing.** An intronic lariat is typically debranched by DBR1 (light purple) and processed for degradation in the nucleus by exonucleases (green). Figure from [120].

2.2.1 Alternative splicing

Alternative splicing AS is a mechanism that allows pre-RNA to be processed into alternative transcript isoforms via the excision of different introns. Recent studies suggest that AS impacts more than 90% of all genes [132]. The repertoire of AS events is broad, with exon skipping, alternative 3' splice sites, alternative 5' splice sites, and intron retention being the most prevalent types [43] (Figure 2-14). Splicing generally occurs co-transcriptionally [30]. Therefore, similarly to CPA, AS is affected not only by the level and activity of splice factors but also by the Pol II elongation rate, chromatin structure and other epigenetic factors [44].

Gene architecture complexity differs across phyla, which requires various splice site identification mechanisms. Unlike the well-conserved splice sites in yeast, metazoan splice sites are less conserved and, thus, demand specific strategies for accurate identification [59]. In metazoa, splice site selection relies on short, conserved sequences called splicing regulatory elements (SREs) within introns or exons. Regulatory proteins like SR proteins and hnRNPs bind to the SREs and enhance or suppress splice site recognition and spliceosome assembly [5]. In vertebrates, intron length varies from hundreds to thousands of nucleotides, and the median length of an internal exon is about 137 nucleotides. In this context, an "exon definition" mechanism commits the upstream intron to splicing [8, 59]. This is achieved by pairing

the 3'SS of the upstream intron with the 5'SS of the downstream intron across the exon. Conversely, in lower eukaryotes with substantially shorter introns, an "intron definition" mechanism triggers splicing by pairing the 5'SS with the downstream 3'SS within the same intron.

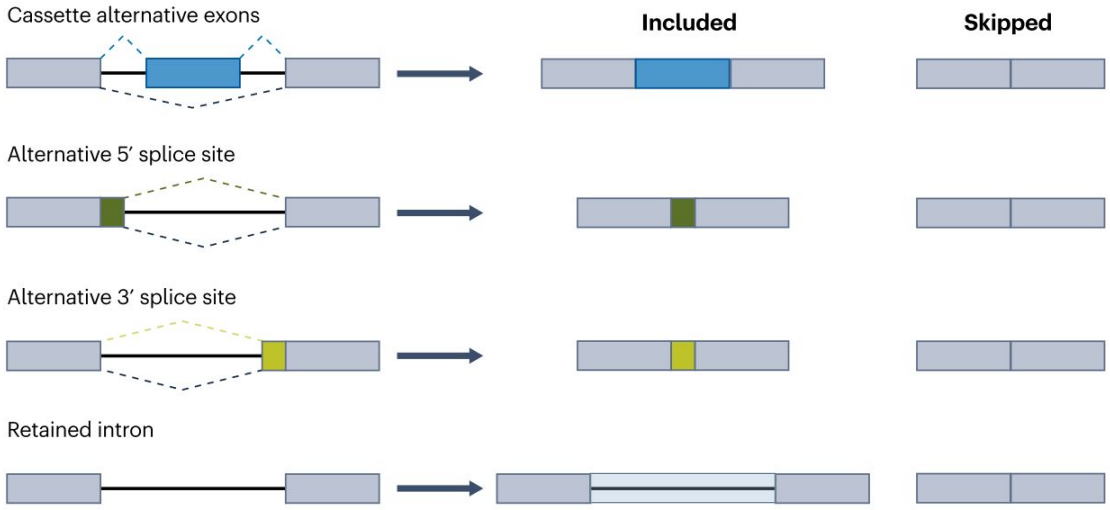


Figure 2-14: **Alternative splicing events are classified into cassette exon expression, alternative 5'SS or 3'SS usage, and intron retention.** Figure adapted from [12].

2.3 The interplay between cleavage and polyadenylation and splicing

The interplay between splicing and CPA is a crucial component of the cotranscriptional pre-mRNA processing [127, 159, 166]. Extensive evidence supports the functional coupling between these two mechanisms. For instance, numerous splicing factors have dual roles and serve in both splicing and polyadenylation, including U2AF [84], PTBP1 [121], members of the Hu protein family [121], and others [52] (see section 2.1.5). Furthermore, given that splicing and CPA factors both physically interact with the Pol II CTD, they may compete for access to the elongating polymerase and counteract each other [114, 64]. Finally, empirical findings, such as the association of IPA with weaker 5'SS and longer introns, along with mutagenesis studies on polyadenylation and splicing signals in plants, together suggest a competitive interplay between splicing and IPA [159, 90].

A decade ago, it was demonstrated that the spliceosome subunit U1 snRNP co-transcriptionally suppresses premature CPA at cryptic intronic PAS in metazoan cells, a phenomenon termed "telescripting" [72]. More recent studies have revealed that U1 snRNP, when bound to the pre-mRNA, can form an inactive complex with key CPA factors CFI, CPSF, and CSTF. In HeLa cells, these complexes are bound to the transcripts around the cryptic PAS in approximately 1,500 genes, preferentially those with longer introns. The inactivity of the complex is hypothesized to stem from the lack of PABPN1 and CFIM68 factors [148].

Two general models have been proposed to explain the inhibition of cryptic intronic PAS by splicing. The first, exemplified by the telescripting phenomenon, is known as the "antitermination model" [160]. It suggests that splicing and CPA factors competitively bind the Pol II CTD or the *cis*-regulatory elements of the pre-mRNA [133, 72]. According to this model, cleavage at the intronic PAS does not occur when splicing prevails over CPA (see Figure 2-15A). In the "kinetic model," the PAS is cleaved, but the splicing reaction initiates before the transcription termination. Thus, the cleavage occurs, but it does not affect the resulting mRNA because the lariat containing the PAS is excised [160] (see Figure 2-15B). Both of these models describe a dynamic counteraction between splicing and intronic CPA, which is dependent on factors such as the polymerase velocity, intron length, and the strengths of the splice sites and the polyadenylation signal.

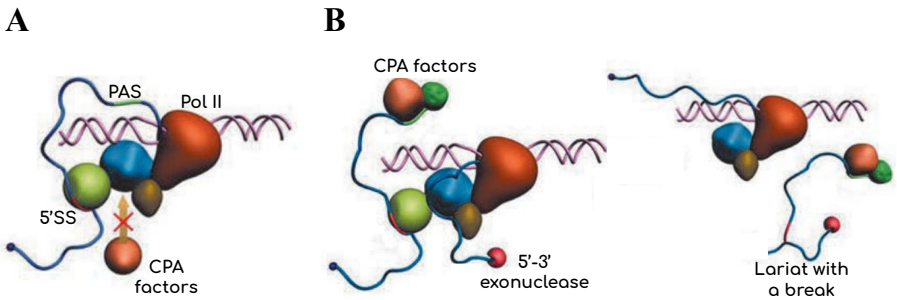


Figure 2-15: **Two models for the inhibition of cryptic intronic PAS by splicing. (A)**. The "antitermination model": PAS is inaccessible to the CPA factors due to the binding of splicing factors to the elongation complex. **(B)**. The "kinetic model": PAS is accessible, but Pol II synthesizes the 3'SS quickly enough to initiate the splicing reaction (left). The introduced break remains in the cut-out lariat (right). Figure adapted from [160]

Multiple studies have shown that the splicing of the terminal intron is linked to the CPA of the pre-mRNA in metazoa [133, 134, 71, 26, 117, 34, 131]. On one hand, elimination of the PAS suppresses splicing of the last intron [134]. This phenomenon can be explained by the exon definition model of the splice site recognition, which is prevalent in organisms with long introns (see 2.2). In this model, the splicing of the last intron depends on the definition of the terminal exon, including the recognition of the PAS [8]. On the other hand, mutations in the last 3'SS inhibit the 3'-end processing and transcription termination [134]. Moreover, splicing factor binding to the 3'SS, U2AF65, interacts with CFI and stimulates CPA *in vitro* [71, 111]. Thus, 3'SS recognition by the spliceosome is required for the 3'-end processing. These and other observations suggest that the machineries for splicing and 3'-end processing function as recruitment platforms for each other. Notably, their catalytic activity is not required for the regulation of the complementary process [59, 134, 26].

In summary, the current body of research offers compelling evidence for the functional interconnection between pre-mRNA splicing and CPA in metazoan cells. These processes not only share regulatory factors but also exhibit a cooperative relationship around the terminal exon and enhance each other's efficiency. However, within individual introns, splicing and CPA are competitive. Both the competitive and cooperative interactions are influenced by numerous factors, such as intron length, splice site strength, and the presence of other *cis*-regulatory elements.

Chapter 3

Thesis Objectives

The main goal of this dissertation is to systematically examine the interplay between pre-mRNA splicing and intronic polyadenylation across human tissues by leveraging a large-scale RNA-seq dataset from GTEx.
Specific aims of the study are

- To identify stably expressed PAS on a genome-wide scale across human tissues;
- To quantify the rate of tissue-specific cleavage and polyadenylation for the identified PASs;
- To characterize the association between tissue-specific rates of intronic polyadenylation and alternative splicing;
- To predict classes of tissue-specific alternative splicing events associated with novel intronic PASs;
- To characterize the abundance and tissue-specificity of skipped and composite alternative terminal exons;
- To examine the hypothesis on the kinetic counteraction between intronic polyadenylation and splicing.

Chapter 4

Materials and methods

4.1 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 [21] and GENCODE websites [57], respectively.

4.2 Genome and gene partitions

To partition the genome, I considered genomic regions defined by the intervals annotated in the GENCODE database. A region that was not covered by any annotated transcript was classified as intergenic. 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 the 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.

4.3 Matched RNA-seq and 3'-seq data

To validate the PAS identification procedure, I used a dataset containing matched 3'-seq and RNA-seq samples of malignant B cells from chronic lymphocytic leukaemia [88]. Since the samples were matched, one could expect that the same PASs were expressed in both 3'-seq and RNA-seq experiments. Thus, I considered the PAS set identified from 3'-seq data as one of the reference sets for validation. The RNA-seq data were downloaded from the Gene Expression Omnibus website (GSE111793) in fastq format and aligned to the reference genome using `STAR v2.7` with the following parameters:

`-outSAMstrandField intronMotif -outFilterType BySJout`
`-outFilterMultimapNmax 20 -outFilterMismatchNmax 999`
`-outFilterMismatchNoverReadLmax 0.04 -alignIntronMin 20`
`-alignIntronMax 1000000 -alignMatesGapMax 1000000`
`-alignSJoverhangMin 8 -alignSJDBoverhangMin 1` [33].

Since the adenine runs at the 3'-ends of short reads from the 3'-seq dataset were already trimmed, I mapped the reads to the same reference genome using `HISAT2` mapper with the default parameters [78].

Gene expression was quantified from RNA-seq data with `featureCounts` program based on exon coverage of protein-coding transcripts of the genes [93].

4.3.1 The identification of PAS from 3'-seq data

Reads generated by the 3'-sequencing technique correspond to adenine stretch starting at the polyadenylation site and ∼50 preceding nucleotides of the mRNA. The location of PAS can be identified as the start of the polyA run.

In the fastq files available in GEO, the adenine runs at the 3'-ends of short reads were already trimmed. Consequently, the genomic position of a PAS corresponded to the very end of each read. Additional information was required to establish, at which end of the read the site was located. Precisely, if the mRNA was complementary to the plus strand (the gene is on the minus strand), then PAS was at the leftmost position of the aligned region and *vice versa*. Therefore, if a read was mapped to

an annotated gene, this gene's coding strand was used to determine the location of the site. Practically, I selected all annotated genes that did not intersect with a gene from another strand (14,922 out of 20,089 protein-coding genes). Then, the `bedtools intersect` utility was used to identify reads that map to the selected genes, and only these reads were considered in the downstream analysis [129]. The coverage of the reads 3'-ends was computed using the `bedtools` utility with the parameter `genomecov -3`, the coding strand of the corresponding gene was used instead of the mapping/coding strand of the read. In the resulting bedgraph file, all regions with coverage lower than ten reads were filtered out leaving only 8% of positions, similar to the initial 3'-seq work where peaks with fewer than five supporting reads were removed [92]. All adjacent intervals were then merged by `bedtools merge` into clusters; the total number of read ends covering each resulting interval characterised the usage of the PAS cluster.

4.3.2 The identification of PAS from RNA-seq data

To identify polyA reads, all reads containing a soft clipped region of at least 6 nts were considered. I 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. PolyA reads were pooled by the genomic position of the first non-templated nucleotide, referred to as PAS position; thus, each PAS was characterized by the number of corresponding polyA reads i.e. polyA read support (Figure 5-1). The position of the soft-clipped region within the read can be used to filter out duplicate reads that were generated by PCR during the library preparation. However, due to the small dataset size, this approach did not significantly improve the quality of PAS identification. PolyA reads originating from genomic adenine-rich regions could contain non-templated adenines due to PCR-induced errors and sequencing artefacts. Thus, PAS located within adenine stretches were excluded from the analysis. Specifically, I located all genomic loci with at least ten consecutive As or ten consecutive Ts (n=1,003,583) and excluded the candidate PAS within these regions or adjacent to them.

Since multiple samples were used in the analysis, the read counts were pooled

across them to increase the coverage of individual PASs. Pre-mRNA cleavage is not completely deterministic but occurs with high frequency at the PAS and with lower frequency at neighbouring positions [157]. Thus, adjacent individual PAS are often clustered together. The appropriate distance threshold in the previous studies varied between 8 and 12 nts [53, 157], so initially 10 nts were picked in this study. The downstream analysis of PAS distribution around annotated TEs showed that this number was adequate. `PySAM` suite was used to process the bam files [25].

To validate the pipeline, I assessed the accuracy of PASs predicted from RNA-seq by comparing them with other data sources. The obvious way to quantify the overlap between the two site sets is to count the number of overlapping PAS, however, this approach does not account for the different read support (level of confidence) of the predicted PAS. Thus, another metric was also applied: each PAS was assigned a weight equal to its read support, and the weighted sums were calculated to estimate the precision and recall values (5.1.1). The precision-recall analysis was restricted to PAS clusters located in genes containing at least one RNA-seq-derived PAS and at least one 3'seq-derived PAS. First, the 3'seq-based PAS cluster set and RNA-seq-based PAS cluster set were both compared against annotated GENCODE TEs. Next, the RNA-seq-based clusters were validated against the 3'-seq-based clusters. The PAS cluster was considered a true positive if it was located within 10nts of a PAS from the reference set.

4.4 GTEx dataset

4.4.1 The 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 v2.7.3a` in paired-end mode [33].

Initially, the same pipeline for identification of PAS from RNA-seq data was applied as in 4.3.2. As previously, the soft clip length threshold was 6 nts and its A/T content threshold was 80%. 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. 220

samples that contained an exceptionally high number of polyA reads were excluded from the analysis (Figure 4-1). Interestingly, they all corresponded to the Whole Blood tissue type. PolyA reads were again pooled by the genomic position of the first non-templated nucleotide, PAS position, which resulted in read counts (f_i) for each value of the overhang (i). Each PAS was characterized by the number of aligned polyA reads $f = \sum_i f_i$ and Shannon entropy of the overhang distribution $H = -\sum_i p_i \log_2 p_i$, where $p_i = f_i/f$. Filtering the PAS by the Shannon entropy values ($H > 2$) reduces the effect of PCR duplicates.

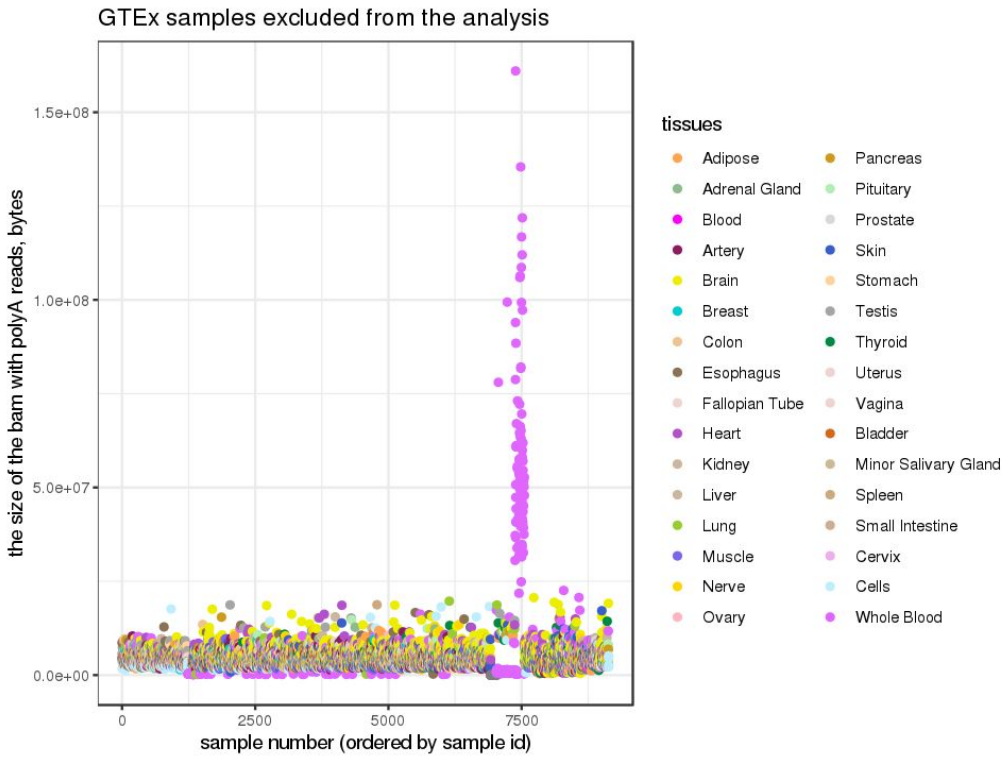


Figure 4-1: **Samples with an exceptionally large number of polyA reads.** A diagram of the number of polyA reads (measured as BAM file size) of 9,135 GTEx samples, of which 220 samples had an exceptionally large number of polyA reads and were excluded from further analysis. The excluded samples are from the Whole Blood tissue type.

After assessment of the initially identified PAS set two additional constraints of the polyA reads were introduced. First, only uniquely mapped reads (NH:1) were considered. Secondly, I excluded reads with average sequencing quality below 13, which corresponds to the probability 0.05 of calling a wrong base.

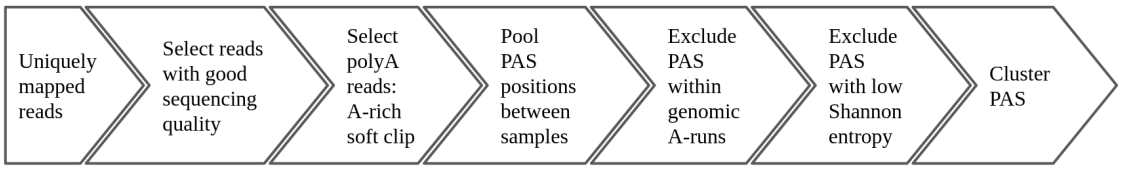


Figure 4-2: **The final pipeline for PAS identification from GTEx RNA-seq data.**

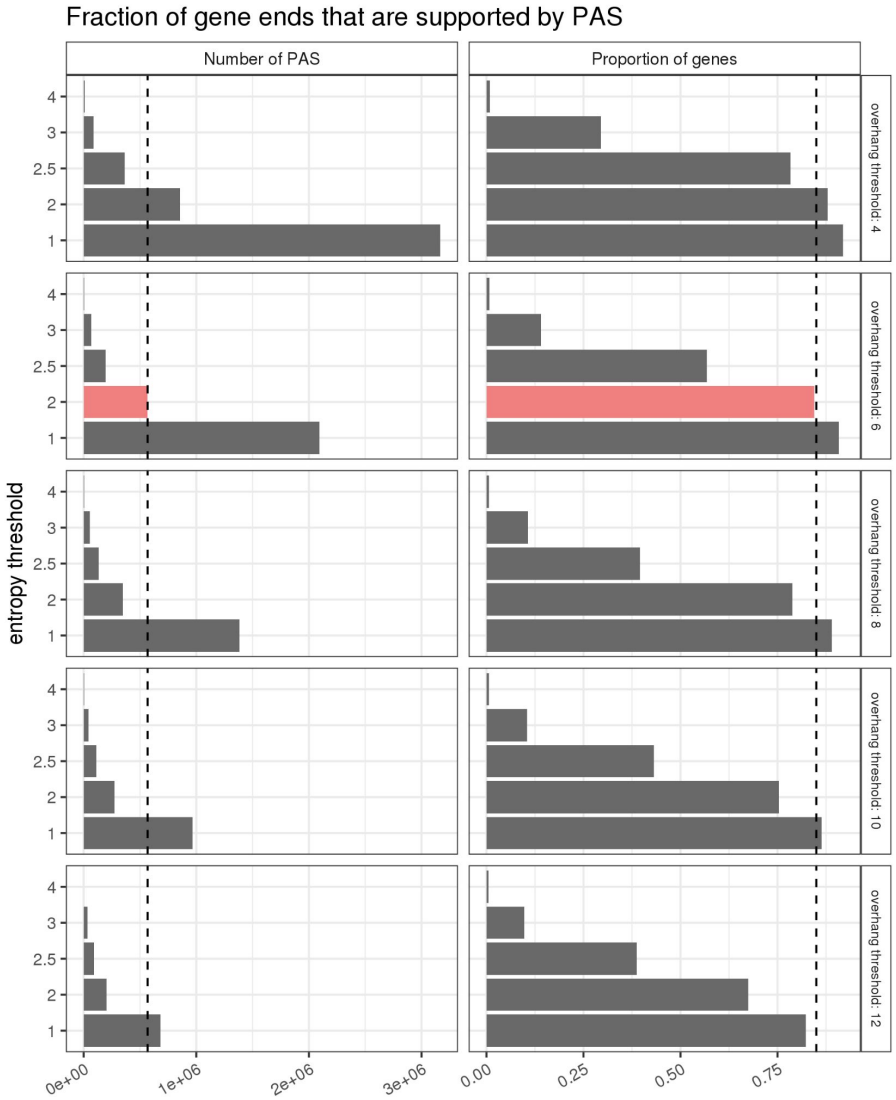


Figure 4-3: **The choice of thresholds for Shannon entropy and minimal overhang length**. The PAS identification procedure was carried out using an array of thresholds on the minimal overhang length and Shannon entropy (H). Shown are the number of PAS (left) and the proportion of genes, in which an annotated transcript end was identified within 100 bp of a PAS. The dashed line represents the number of PAS reported in PolyASite 2.0. The condition $H \geq 2$ in combination with the minimum overhang length of 6 nts (red) gives the optimal cutoff.

To find optimal cutoffs, I 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 4-3). 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 [58]. PASs that were located within 10 nts of each other were merged into clusters (PASCs) using the `GenomicRanges` package [86].

4.4.2 Estimation of CPA precision

The precision of cleavage and polyadenylation position for annotated TEs was estimated based on the spread of the identified PASs around the TE, specifically the interquartile range (IQR) of the PASs positions. For weighted IQR (wIQR) each PAS is weighted by its polyA read support i.e. in wIQR calculations each PAS was repeated n times, where n - polyA read support of the PAS.

$$wIQR = IQR(\{PAS_1, ..., \underbrace{PAS_i, ..., PAS_i}_{s_i \text{ times}}, ..., PAS_N\})$$

where $s_i = PAS_i$ polyA read support, N = number of PAS within 100nts of the TE. All analyses with IQR calculation did not include TEs that had only one PAS within 100 nts.

To estimate the stability of potential RNA structures around TEs, the difference of Gibbs energies for folded and unstructured forms (dG) were predicted using `RNAfold` from the ViennaRNA with `-noPS -noDP -p0` parameters [96], 100nt regions upstream and downstream of the TEs were considered. To eliminate the influence of the GC content on the energy values, for each region, its counterpart with the same nucleotide content but a shuffled sequence was generated. Then, the difference between dG and dG_{shuf} was computed for each TE, $dG_{structure} = dG - dG_{shuf}$.

4.4.3 Saturation analysis

For saturation analysis, downsampling was performed 10 times. For each subsample, all candidate PAS within 10 nts were clustered and the number of obtained clusters was recorded (Figure 4-4). The number of obtained PASCs monotonously increased with the number of samples and, accordingly, with the total number of RNA-seq reads subjected to the analysis. I also estimated the number of PASCs with polyA read support higher than 1, 10, 50 or 100 for each subsample, however, these restrictions did not alter the shape of the curve.

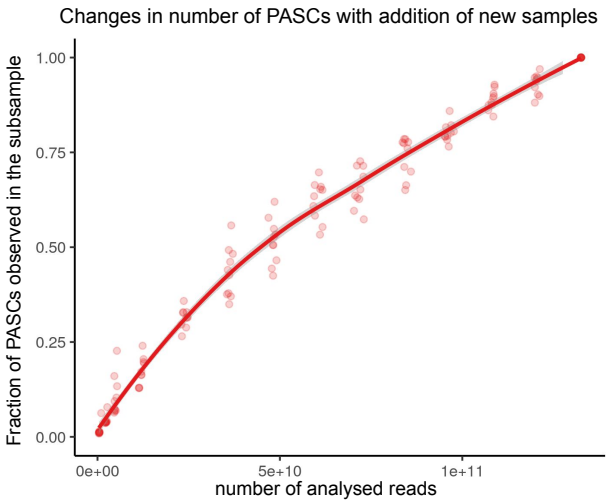


Figure 4-4: **Saturation of PAS clusters.** Number of PAS clusters identified from random subsets of samples from the GTEx RNA-seq dataset. PASCs were formed from all candidate PAS in the corresponding subsample; the total number of clusters was ∼5mil.

To analyze PASCs newly identified in each subset a different approach was used. Only candidate sites with substantial per-sample entropy value $H_{sample} \geq 2$ were clustered. The number of obtained clusters and their positions were recorded for each subsample. This information was used to analyse the distribution of newly identified PASCs among different regions of protein-coding genes (Figure 5-18, 5.1.5).

4.4.4 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 [9, 16, 193]. The precision and recall metrics varied depending on the number of supporting polyA reads (in GTEx) and the 3'-end sequencing read coverage (in Atlas). For each comparison the point with maximum $F_1 = 2(P^{-1} + R^{-1})^{-1}$ metric was identified.

4.4.5 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, I 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.

4.4.6 Read coverage and fold change

To quantify the extent, to which CPA happen at a specific PASC in a specific tissue, I 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` [130] and averaged the read coverage values between samples within each tissue using `wiggletools mean` utility [188].

Next, I 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 [130]. 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 usage, I followed the approach described

previously [88] by detecting significant differences in read counts between the upstream and downstream windows ($p_{adj} < 10^{-3}$) using `DESeq2` [98], 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. An iPASC located within 100 nts from an annotated TE of a protein-coding transcript ($n = 3,188$) 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 1,136 CTEs and 1,948 STEs; 104 PASCs located near multiple TEs were excluded due to the conflicting annotation.

In 5.2.1 I evaluated the CPA rate, at which it acts on the nascent pre-mRNA. To account for the bias arising from intron degradation, I normalized the polyA read count to the average read coverage in exons and introns. To estimate the read coverage in constitutive exons, alternative exons, and introns, the total read coverage values per nucleotide in GTEx samples were averaged between windows (*wi* and *we*) located in the respective regions, resulting in the normalization factors of $3.3 \cdot 10^6$, $3.2 \cdot 10^6$, and $8.0 \cdot 10^4$, respectively.

4.4.7 Splicing metrics

To quantify tissue-specific alternative splicing associated with intronic PASCs, I computed split read counts using the IPSA pipeline [125, 109]. 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 in the intron before iPASC, and the number of continuous reads spanning the upstream exon-intron boundary, respectively. The values of ψ with the denominator below 30 were discarded as unreliable.

4.5 Cleave-seq and 3'-RNA capping and pulldown data

The results of Cleave-seq experiments in HeLa cells were downloaded from Gene Expression Omnibus (GEO) under the accession number GSE165742 (downloaded samples GSM5566266 – GSM5566269) in bigwig format [155]. The per-bin Cleave-seq signal was computed around 5'-splice sites using the `computeMatrix` from `deeptools` suit 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 [102, 101] were downloaded from GEO 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.

Chapter 5

Results

5.1 De novo PAS identification from RNA-seq data

A fraction of the reads from the traditional polyA(+) RNA-seq protocol map only partially to the genome because they contain stretches of non-templated adenines from the polyA tails. These reads are referred to as polyA reads. The mapping positions of these reads can be used to identify the PASs.

The developed pipeline processes BAM files and outputs a set of predicted PAS (see 4.3.2 and 4.4.1). The first script isolates polyA reads, defined as reads with a soft-clipped region of at least six nucleotides, of which 80% or more are adenines. Subsequently, the genomic position of the first non-templated nucleotide in each polyA read is determined (see Figure 5-1). These positions constitute the preliminary PAS set.

In the following steps, this preliminary set undergoes filtration to minimize false positives. By definition, a polyA read's alignment contains a soft-clipped region, termed here as the overhang (Figure 5-1). Thus, each predicted PAS is characterized by both the number of supporting polyA reads (read support) and the distribution of the lengths of their overhangs. Our confidence in a PAS correlates with the variability of this distribution, quantified by the Shannon entropy (denoted as H, see 4.4.1), as well as the read support. Therefore, in the second step of the pipeline, PAS with low entropy values are excluded. Entropy-based filtering also eliminates PAS with insufficient read support and reduces the influence of the duplicate reads

generated by PCR during library preparation.

Since the sequencing of mononucleotide repeats is prone to high error rates, PAS in genomic adenine-rich regions are often falsely identified. This happens because unmapped adenines at the ends of the polyA reads stem from sequencing errors rather than the polyA tail. Thus, any predicted PAS located in these adenine-rich regions are also removed.

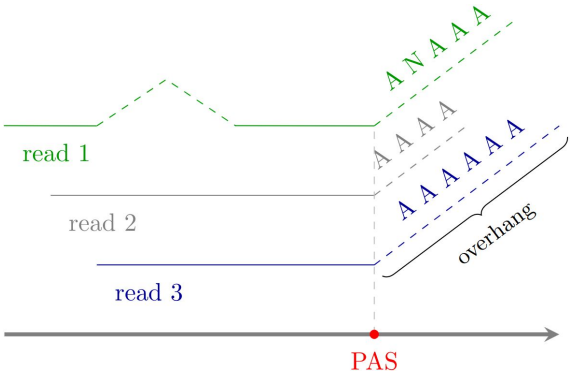


Figure 5-1: **The identification of PAS.** The alignments of short reads with non-templated adenine-rich ends (polyA reads). The genomic position of the first non-templated nucleotide corresponds to a PAS. The length of the soft clip region is referred to as the overhang.

Finally, the predicted PASs located within 10 nts of each other are merged, and the result is returned as a separate PAS cluster (PASC) set. A detailed description of the pipeline is provided in Chapter 4, specifically in Sections 4.3.2 and 4.4.1 for the analysis of the small and large datasets, respectively.

The reads with non-templated adenine stretches could also result from misalignment, reverse transcription, or sequencing artifacts in genomic adenine- or thymine-rich regions. Thus, I primarily assessed the accuracy of the PASs predicted from RNA-seq by comparing them to PAS from other data sources.

5.1.1 Matched RNA-seq and 3'-seq dataset

First, I analyzed a dataset containing matched 3'-seq and RNA-seq data [88]. The developed pipeline for polyA read analysis was applied to the 13 RNA-seq samples, which had a total of approximately 0.7 billion reads (see 4.3.2). This processing yielded roughly 400,000 reads with non-templated adenines, 202,216 predicted PAS,

and 186,249 PAS clusters with varying levels of polyA read support. Consistent with previous reports, only 16% of the predicted PAS were backed by multiple polyA reads. The 3'-seq data was processed as described by the authors of the protocol [92] (4.3.1), resulting in 194,488 3'-seq-based PAS clusters with a median coverage density of around 20 reads per nt.

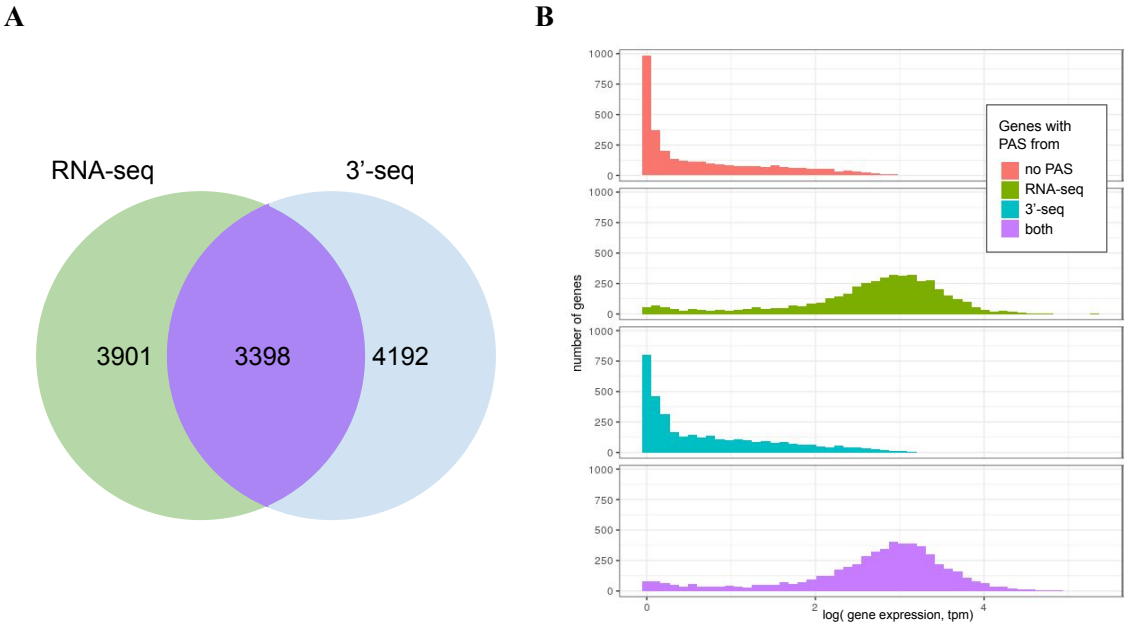


Figure 5-2: **Genes with PAS in RNA-seq and 3'-seq data. (A)** Venn diagram of genes containing at least one PAS from RNA-seq (blue) and 3'seq (green) data. **(B)** Histograms of levels of gene expression for the genes grouped by the presence of PAS from RNA-seq and/or 3'-seq dataset. Gene sets are coloured in the same way as in A: green - genes containing only PAS from RNA-seq based set, blue - only from the 3'-seq-based set, violet - genes with PAS from both sets, red - no PAS were identified in the gene.

Since the samples were matched, it was expected that the 3'-seq and RNA-seq datasets would contain reads from the same genes. However, PAS in only about 4,000 genes were identified by both methods, while each method individually predicted PAS in over 7,000 genes (Figure 5-2A). Notably, 3'-seq often did not detect PAS in many highly expressed genes (Figure 5-2B). This discrepancy partially stems from the fact that genes that intersect with a gene from another strand were excluded from the 3-'seq data analysis (see 4.3.2). The primary aim of this study was to compare the PAS positions identified by the two methods. Therefore, to minimize the effect of discrepancies between the gene sets, subsequent analyses were limited

to genes harbouring both 3'-seq- and RNA-seq-derived PAS (n=3398).

I clustered adjacent candidate sites and evaluated the precision and recall of the PASs in relation to the PAS set derived from 3'-seq data at varying polyA read support levels (Figure 5-3, bottom right). Precision was defined as the proportion of predicted PASs located within 10 nts of a PAS derived from 3'-seq data, and recall was defined as the proportion of 3'-seq-derived PASs that had a predicted PAS within 10 nts. This method of quantifying the overlap between the two sets did not account for the different confidence levels of the predicted PAS. Thus, another metric was also applied: each PAS was assigned a weight equal to its read support, and the weighted sums were calculated to estimate the precision and recall (Figure 5-3, top right). At the highest F_1 measure (defined in 4.4.4), nearly 70% of the PAS from 3'-seq were captured by RNA-seq, with a precision of 63%. As anticipated, precision consistently decreased with diminishing read support for the PAS.

The 3'-seq protocol has its limitations. It's susceptible to various sequencing artefacts typical of homopolymers, and as previously mentioned, its efficiency can vary across genes. Additionally, oligo(dT) primers used in the 3'-seq protocol can anneal to genomic A-rich regions, yielding a sequence of adenines that is indiscernible from the poly(A) tail [194, 138]. This led me to utilize the set of annotated transcript ends as a more conservative reference set (Figure 5-3, top left).

The precision of PAS identified from RNA-seq surpassed that of PAS derived from 3'-seq in a narrow region corresponding to highly covered transcripts. However, beyond this region, RNA-seq identifies far fewer annotated PAS compared to 3'-seq. Furthermore, when I analyzed just the number of overlapping PAS without considering read support, the performance deteriorated (Figure 5-3, bottom). This emphasizes that only well-supported sites are reliable. Such findings suggest that the coverage of non-templated adenines provided by the 13 RNA-seq samples isn't sufficient to pinpoint genuine annotated PAS. Consequently, I explored the efficacy of the developed approach on a more extensive RNA-seq dataset.

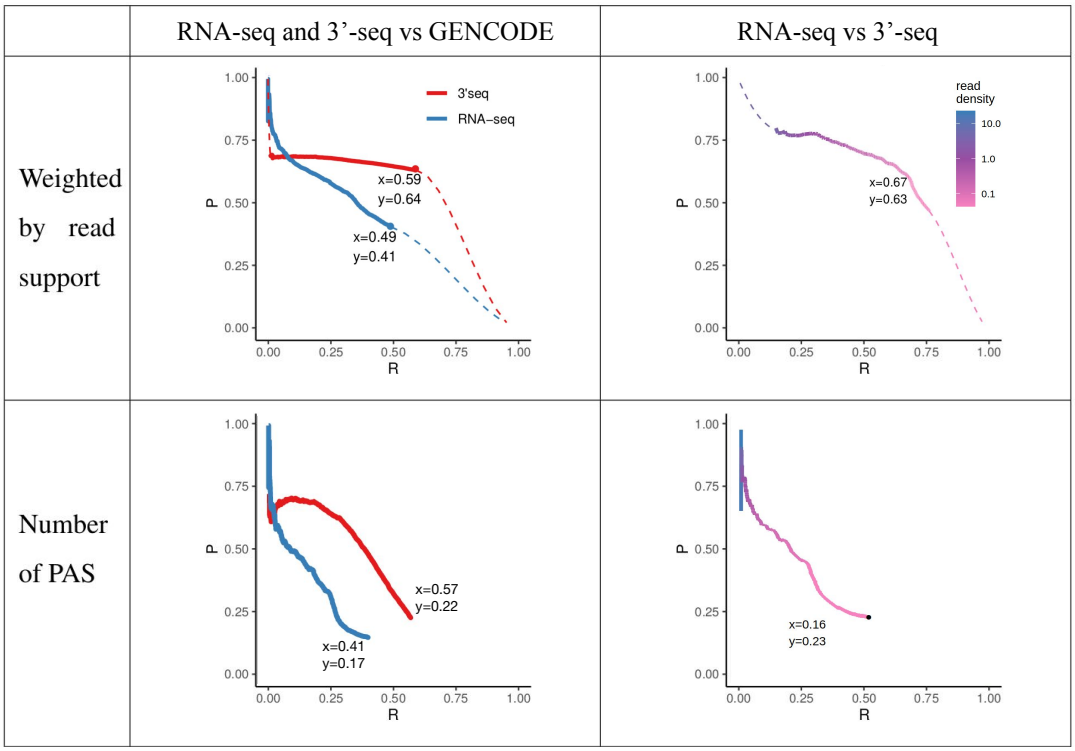


Figure 5-3: **Precision-recall curves for PASs validation against 3'seq-based set and GENCODE**. The comparison of PAS identified from RNA-seq and 3'-seq vs. GENCODE (left) and RNA-seq vs. 3'seq (right) in terms of the polyA reads (top) and the number of PAS clusters (bottom) using precision (P) and recall (R) metrics (4.3.2). The curves are obtained by varying the threshold on polyA read support. The values of precision and recall with the largest F_1 metric are shown.

5.1.2 GTEx – a large-scale dataset

As a larger RNA-seq panel, I utilized the Genotype Tissue Expression project (GTEx) dataset, which encompasses over 10,000 RNA-seq samples from 31 human tissues [109]. Since 3'-seq data for these samples were not available, I opted to use the consolidated atlas of PAS derived from 3'-end sequencing and similar protocols (PolyASite 2.0 [58], henceforth referred to as "Atlas") as the primary reference.

The developed pipeline was applied to 9,021 GTEx RNA-seq experiments, excluding samples with anomalously high numbers of polyA reads (4.4.1, Figure 4-1). Out of approximately 554 billion input reads, around 562 million (0.1%) polyA reads were obtained, which led to 37,300,049 potential PAS. Approximately 1,5% of them were located within adenine-rich genomic tracks.

As discussed in 5.1, confidence in a PAS correlates not only with the read support

but also with the variability in the overhang distribution, measured by the Shannon entropy. Only 20% of potential PAS were supported by polyA reads with a variety of overhang lengths (i.e., possessed non-zero entropy). Initially, I set a threshold value proximate to the 0.95 percentile: PAS with entropy greater than 2 were included in further analysis (n=1,446,521). To illustrate, an entropy of 2 characterises a PAS supported by four polyA reads, with each read featuring a unique overhang length.

To classify the distribution of PAS across genomic regions, I partitioned the human genome into distinct intervals corresponding to protein-coding genes, non-coding genes, and intergenic areas. Altogether, 548,478; 126,977; and 771,066 PAS were located in these respective sections. The magnitude of polyA read support diverged across the genomic sections. For instance, 20%, 10%, and 4% of PAS were backed by 100 or more polyA reads in protein-coding, non-coding, and intergenic areas respectively (Figure 5-4A). All in all, a greater number of candidate PAS were identified in intergenic regions than in protein-coding ones. Given that many intergenic PAS might be false positives, I imposed more rigorous selection criteria for polyA reads.

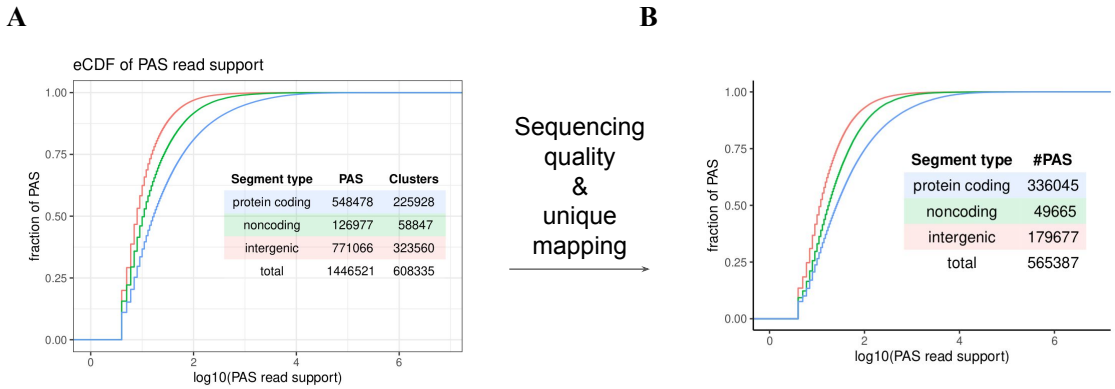


Figure 5-4: **PAS in genomic regions. (A)** The polyA read support of PAS in protein-coding genes, non-coding genes, and intergenic regions. The number of PASs in each group is indicated in the inset. **(B)** Same but after filtering the polyA reads by sequencing quality and number of mapping hits (NH1). The fraction of PAS located in protein-coding regions increased substantially.

RNA-seq reads with incomplete alignment to the genomic reference map only partially, which elevates the relative importance of sequencing quality for each mapped nucleotide. Therefore, for PAS identification, I only considered polyA reads

with an average sequencing quality of 13 or higher, which corresponds to the probability 0.05 of calling a wrong base. This filtration considerably diminished the count of polyA reads from 562 million to 468 million and resulted in 10,782,482 candidate PAS.

RNA-seq reads with partial alignment to the genomic reference tend to map to multiple locations. Hence, I adopted a conservative approach and considered solely the uniquely mapped reads (see 4.4.1). Out of approximately 356 billion uniquely mapped reads, around 413 million (0.11%) polyA reads were obtained. Approximately 9.6 million candidate PASs were identified from these uniquely mapped polyA reads with good average sequencing quality. Broadly speaking, rigorous supervision of sequencing and mapping quality for the polyA reads reduced the candidate sites count by 65%. The mean adenine content in soft clipped portions of the chosen polyA reads stood at 98%, surpassing the initial 80% threshold, indicating that the selected short reads indeed contain polyA tails.

Once again, I evaluated the diversity of polyA reads supporting the PASs. Out of the 9.6 million prospective PASs, 2.1 million (22%) possessed $H \geq 1$ and 565,387 (6%) had $H \geq 2$. Furthermore, the distribution of PASs across distinct genomic regions changed substantially. 336,045; 49,665; and 179,677 PASs were identified within protein-coding genes, non-coding genes, and intergenic areas respectively (Figure 5-4B). It is worth noting that PAS from protein-coding segments dominated the set. Fluctuations in polyA read support across different genomic regions remained the same.

In further analysis, I persisted with the threshold $H \geq 2$ as it yielded a list of PASs that matched by the order of magnitude the consolidated atlas of polyadenylation sites from 3'-end sequencing [58] and encompassed a substantial number of annotated gene termini (4.4.1, Figure 4-3). A detailed comparison of the PAS collection from the GTEx RNA-seq data versus the polyadenylation atlas and TEs annotated by the GENCODE consortium is discussed in a subsequent section (5.1.4).

Out of the 565,387 PASs with $H \geq 2$, 331,563 contained a sequence motif similar to the canonical consensus CPA signal (NAUAAA, ANUAAA, or AAUANA) within the preceding 40-nt sequence [152, 162]. Such PASs will henceforth be referred to

	protein-coding genes	**noncoding genes**	**intergenic regions**
Number of PAS	336,045	49,665	179,677
length, Mb	1,374.032	583.840	4,233.482
Density per Mb	244,5	85,1	42,4

Table 5.1: PAS in genomic regions.

as PASs with a signal. In the other two genomic regions, 61% and 39% were PASs with a signal, respectively.

As expected, protein-coding regions had the largest density of PASs per megabase (Table 5.1). However, a 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 [68, 32, 56].

5.1.3 Estimation of CPA precision from individual PAS positions

An example of a gene that is highly covered by polyA reads is *RPL5*, which encodes a component of the 60S subunit of the Ribosome (Figure 5-5). I 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. In line with previous studies, instead of a single peak, I observed a relatively dispersed cluster of PASs spanning twelve nucleotides [157]. The manual inspection confirmed that the RNA-seq read alignments ending at all these positions indeed were followed by non-templated polyA tracks, thus indicating that the observed pattern was due to biological stochasticity and not mapping artefacts. The 3'-seq read coverage in the *RPL5* locus also followed this pattern (Figure A-1). Remarkably, the number of polyA reads decayed with the increase in the length of the overhang (Figure 5-5, 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 [189].

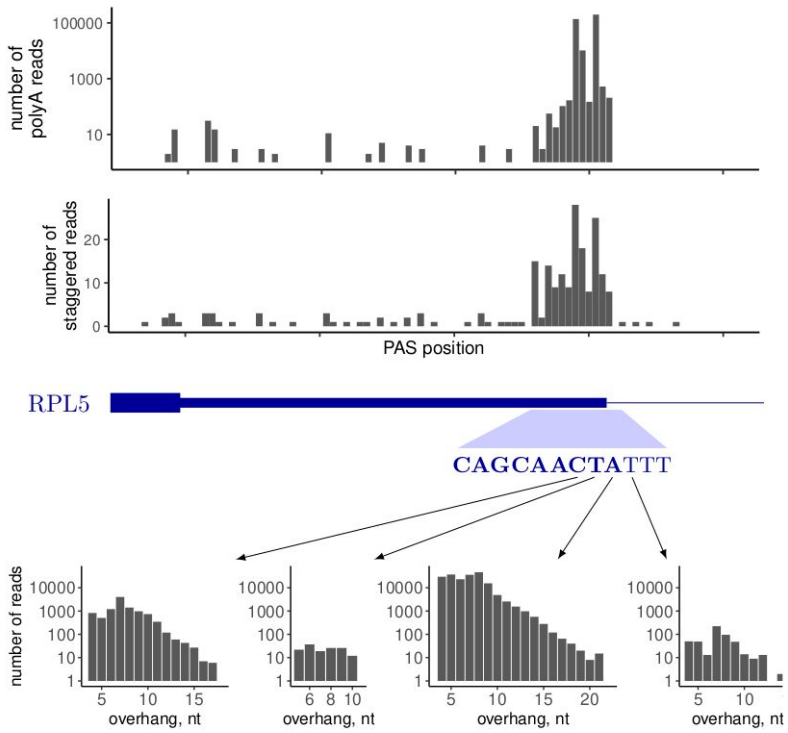


Figure 5-5: **The width of PAS groups. Example.** The 3'-end of the *RPL5* gene is highly covered by polyA reads. Top: the positional distribution of the number of polyA reads (in log scale) and the number of staggered polyA reads (i.e., the number of different overhangs). Bottom: the distribution of overhangs at the indicated positions (in log scale).

Large variability of PASs positions in *RPL5* motivated me to explore the distribution of distances from each PAS to its closest annotated TE in protein-coding genes (Figure 5-6A). Among PASs that were located within 100 nts from an annotated TE, the median distance to the TE was 5 nts, 71% fell within 10 nts, and 78% of PASs with a signal did so. Additionally, for each annotated TE, I computed the interquartile range (IQR) of the distances to all PASs located within 100 nts, excluding TEs with a single PAS (Figure 5-6B). Approximately 83% of TEs had an IQR below 10 nts, and 87% of TEs did so when considering only PASs with a signal.

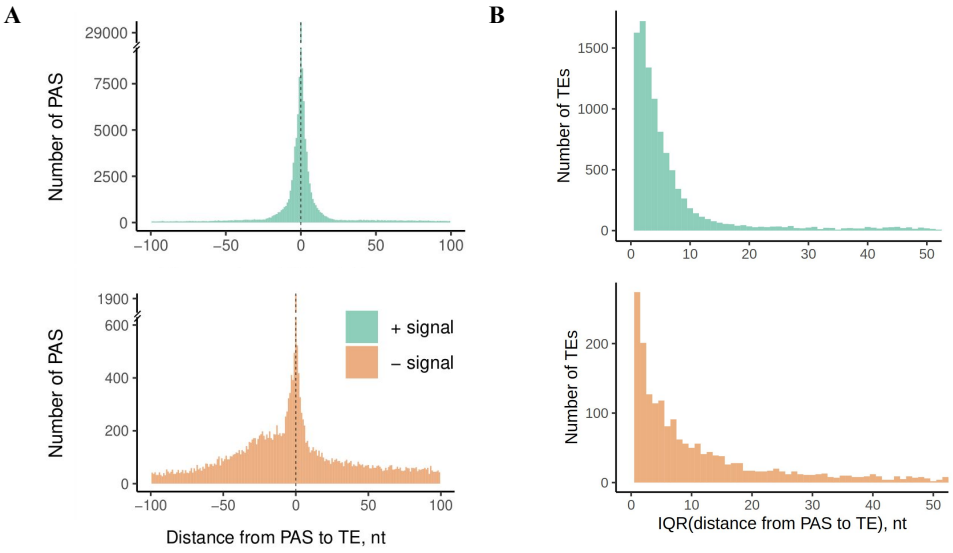


Figure 5-6: **PAS around annotated TEs.** (**A**) The distribution of distances from each PAS to its closest annotated transcript end (TE) for PAS with (green, $n = 122,448$) and without a signal (orange, $n = 22,361$). (**B**) The distribution of IQRs of distances from neighbouring (within 100bp) PAS to each annotated TE with (green) and without a signal (orange).

To explore the frequency of CPA events at each PAS around TEs, I analyzed the distribution of the polyA reads and computed the IQRs of PAS positions weighted by their polyA read support (or "weighted IQR"). These tend to be lower than the unweighted IQRs (Figure 5-7, bottom panel, details in 4.4.2). This observation suggests that PASs with higher read support are located close to each other. It also implies that CPA occurs with higher frequency at "strong" 3'-end processing sites and with lower frequency at the surrounding positions.

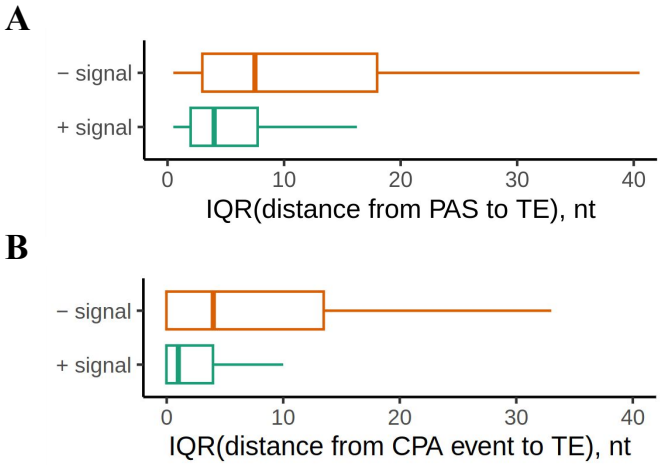


Figure 5-7: **TEs with signal tend to be surrounded by narrow PAS groups. (A)** IQRs of distances from neighbouring PAS to each annotated TE with (green) and without a signal (orange). **(B)** IQRs of distances from neighbouring PAS to each annotated TE weighted by the PAS read support, can be a proxy for the IQR of the CPA events around the TE. Details about wIQR in 4.4.2.

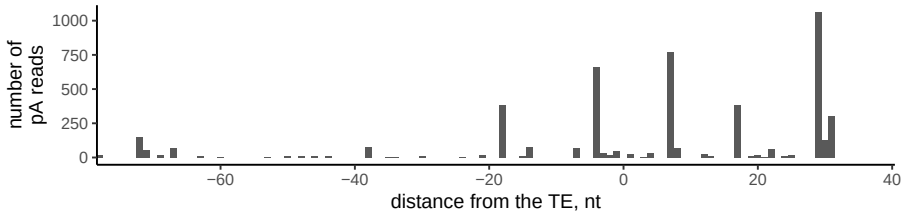


Figure 5-8: **An example of a transcript end with an imprecise cleavage point.** The 3'-end of the ENST00000442760 transcript in the *TPT1* gene is surrounded by multiple PASs. The gene plays a role in carcinogenesis and encodes a regulator of cellular growth and proliferation. The plot shows the positional distribution of the polyA reads.

The polyA reads provide a snapshot of CPA at single nucleotide resolution, which reveals that PASs form peaks of varying widths around the TEs (Figures 5-5, 5-8). This observation indicates that the precision of the CPA machinery is highly variable, producing either narrow clusters of closely spaced PASs or broader regions with imprecise cleavage points. These findings prompt a question about the determinants of CPA precision across different PAS classes, which I sought to address. The first noted feature was the presence of the polyadenylation signal, associated with narrower PAS groups (Figure 5-6B, 5-7; one-sided Wilcoxon rank sum test p-value for smaller weighted IQRs in the "+ signal" group was $< 10^{-16}$).

Moreover, this effect was most pronounced for TEs with a polyadenylation signal located 18 to 15 nts upstream of the TE (Figure 5-9). As another known CPA motif is the GU/U-rich region 20 nts downstream of the PAS, I also analyzed the influence of GU-content in the downstream region [162] but did not observe any significant effect (Figure 5-10).

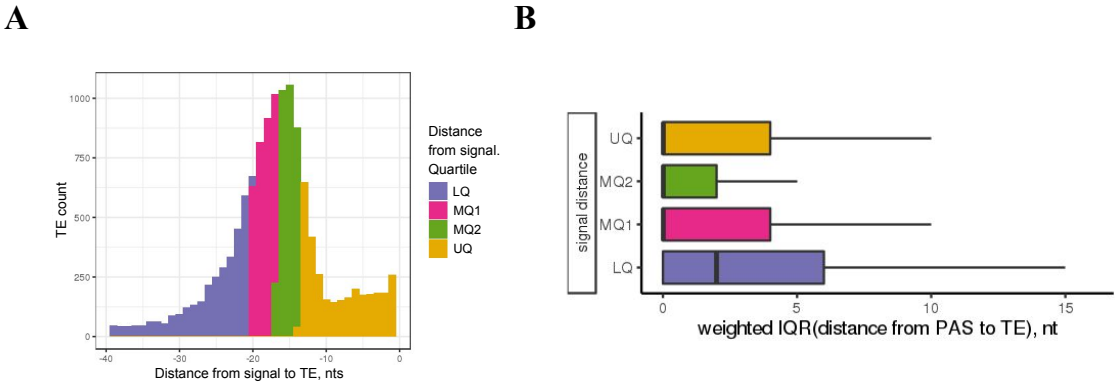


Figure 5-9: **TEs with signal tended to be surrounded by narrow PAS groups.** **(A)** Distribution of the polyadenylation signal positions relative to the TE. Color-coded by the quartiles. **(B)** Boxplots representing the IQRs of distances from neighbouring PAS to each annotated TE, weighted by the PAS read support. TEs were grouped by the quartile of the distance from the signal to the TE.

Figure 5-10: **GU content did not affect the width of the PAS groups.** Boxplots representing the IQRs of distances from neighbouring PAS to each annotated TE, weighted by the PAS read support. TEs are grouped based on the high and low G, GU or U content in the 20 nts region downstream of the TE. Only TEs with content values from the upper and lower quartiles are shown. Plots are color-coded by the quartile: high (yellow) and low content (violet). Asterisks represent the p-values of the one-sided Wilcoxon rank sum test ($* \leq 0.05$, $** \leq 0.01$, $*** \leq 0.001$).

Next, I examined potential determinants related to the genomic sequence, specifically the frequency of all potential bases within 50 nts of the TE (Figure 5-11). The adenine content in the downstream region exhibited the most substantial effect on the IQR value, with broader PAS groups correlating with higher A frequency. G-content around the TE, both upstream and downstream, also significantly influenced the IQR of the PASs distribution. The CFII subunit PCF11, which binds RNA downstream of the PAS, has a preference for G-rich sequence elements [141]. This preference aligns with thinner PAS clusters observed for TEs with high downstream G-content. Interestingly, a similar significant association was observed for downstream C content.

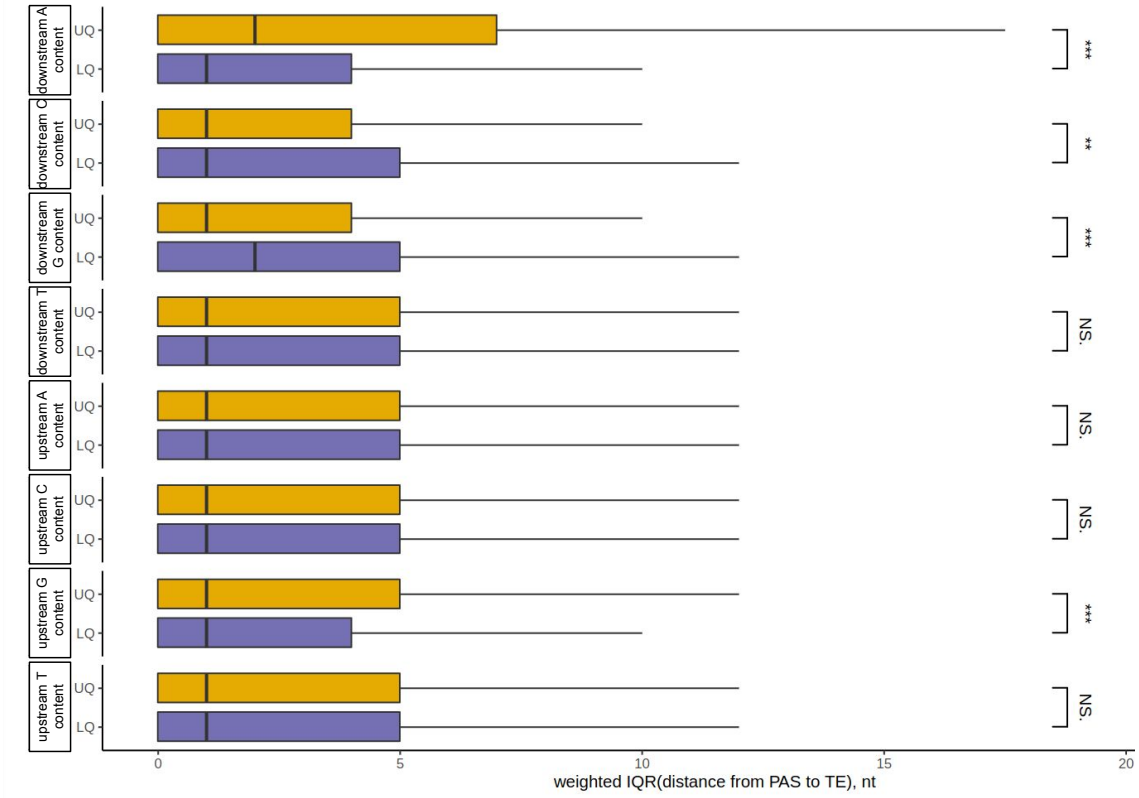


Figure 5-11: **Other determinants of the PAS cluster widths related to the genomic sequence.** Boxplots representing the IQRs of distances from neighbouring PASs to each annotated TE, weighted by PAS read support. TEs are grouped based on the high and low A, T, G, or C genomic content within the 50-nt region either upstream or downstream of the TE. Only TEs with content values from the upper and lower quartiles are shown. Boxplots are color-coded by the quartile: high content (yellow) and low (violet).

Given that high GC content is linked to a greater likelihood of secondary struc-

ture formation, I hypothesised that the distribution of CPA events might be influenced by RNA secondary structures. To investigate this, I assessed the stability of RNA secondary structures formed in the regions 100 nts upstream and downstream and found that a higher absolute difference in Gibbs energy in the downstream region corresponded with narrower PAS groups (Figure 5-12). However, after adjusting for the effect of nucleotide content in the sequence (4.4.2), this association was no longer evident (Figure 5-12, bottom plot). Since the stability of these structures is also influenced by A-content, I deduced that the observed impact of nucleotide content on CPA precision might arise from artefacts linked to the A-richness of the downstream region. While PAS at the boundaries of adenine genomic runs were excluded (5.1.1), regions with high A content slightly below the threshold could still produce spurious PASs.

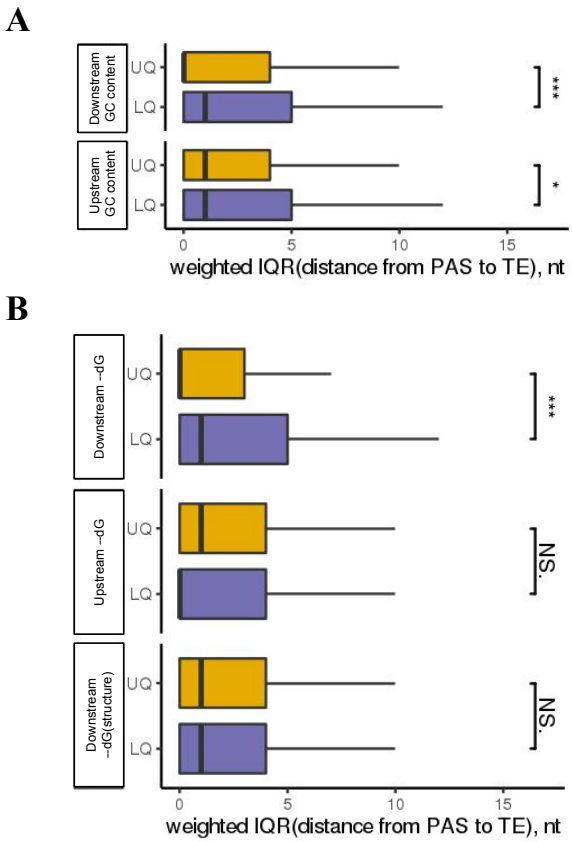


Figure 5-12: **Stability of RNA secondary structures forming in the upstream and downstream regions as a possible cluster width determinant.** Only TEs with energies from the upper and lower quartiles are shown. Boxplots are color-coded by the quartile: high content (yellow) and low (violet). The caption continues onto the following page.

Figure 5-12: (Previous page.)**(A)** Boxplots representing the IQRs of distances from neighbouring PASs to each annotated TE, weighted by the PAS read support. TEs are grouped based on the GC content in the 100nts region downstream (top) or upstream (bottom) of the TE. **(B)** Boxplots representing the IQRs of distances from neighbouring PASs to each annotated TE, weighted by the PAS read support. TEs were grouped by the high and low Gibbs energies of the corresponding upstream/-downstream structure (upper and middle plots). Gibbs energies were calculated via RNAfold [96]. For the bottom boxplot TEs were grouped by the Gibbs energy corresponding to the secondary structure formation (not the AT/GC content of the sequence). The $dG_{structure}$ was estimated by subtracting the Gibbs energies of shuffled sequences (see 4).

In the distribution of distances from each PAS to its closest annotated TE in protein-coding genes (Figure 5-6), approximately 83% of TEs had an IQR below 10 nts. When considering only PASs with a signal, this fraction increased to 87%. This variability in PAS positions is consistent with findings from a massively parallel reporter assay [162] and the seminal work by B. Tian et al [157]. As a result, I chose to combine PASs located within 10 nts of one another (Figure 5-13), which yielded 318,898 PAS clusters (PASCs). Of these, 90% were 10 nts or shorter, 72% contained a unique PAS, and 99% comprised fewer than ten individual PASs (Figure 5-14). In the following analyses, a PASC will be termed a "PASC with a signal" if it includes at least one individual PAS with a signal. The polyA read support for a PASC is determined by the cumulative number of supporting polyA reads across its individual PASs.

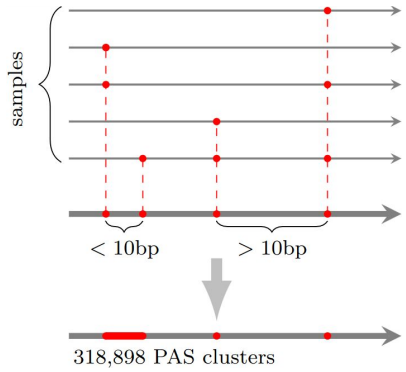


Figure 5-13: **PAS clusters.** PASs located <10 bp from each other are merged into a PAS cluster (PASC).

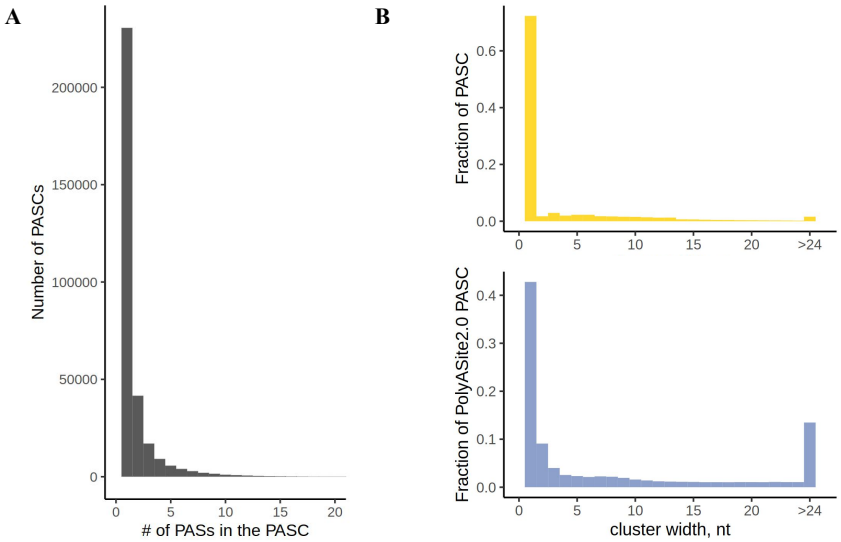


Figure 5-14: **PASC characteristics.** (**A**) The distribution of the number of individual PAS in the cluster for GTEx PASCs. (**B**) The distribution of cluster widths for GTEx PASCs and Atlas clusters.

5.1.4 The PASC set validation

I next examined how PASCs identified from GTEx RNA-seq data compared to those in the consolidated polyadenylation atlas (PolyASite 2.0 [58]) and TEs annotated by the GENCODE consortium [57]. To evaluate this, I enclosed TEs from GENCODE within 100-nt windows and assessed pairwise intersections among the three datasets (Figure 5-15). The precision of GTEx in relation to GENCODE, i.e., the proportion of PASCs from GTEx situated within 100 nts of an annotated TE, surpassed that of PolyASite 2.0. However, its recall, i.e., the proportion of annotated TEs supported by at least one PASC from GTEx within 100 nts, was less. In contrast, the precision of GTEx with respect to PolyASite 2.0 was lower compared to that of GENCODE, while the recall was higher. This trade-off between precision and recall persisted when reducing the window around TEs to 50 nts and also when focusing on intronic PASCs (Figure A-3). This analysis suggests that GTEx RNA-seq data provides 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.

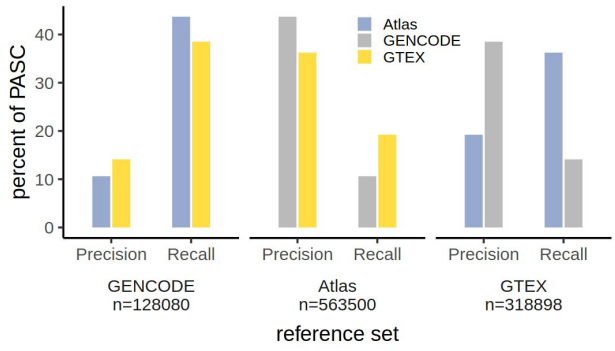


Figure 5-15: **PASCs validation.** Pairwise comparison of PASs inferred from GTEx, PolyASite 2.0 (Atlas), and GENCODE. Left: the proportion of PASCs from Atlas or GTEx that are supported by GENCODE (precision) and the proportion of PASCs from GENCODE that are supported by Atlas or GTEx (recall). Middle: the proportion of PASC from GENCODE or GTEx that are supported by Atlas (precision) and the proportion of PASC from Atlas that are supported by GENCODE or GTEx (recall). Right: the proportion of PASC from Atlas or GENCODE that are supported by GTEx (precision) and the proportion of PASC from GTEx that are supported by Atlas or GENCODE (recall).

Similarly to section 5.1.1, I probed the relationship between precision and recall for GTEx and PolyASite 2.0, with weighting the PAS by their polyA read support (Figure 5-16). Specifically, I selected PASCs in protein-coding genes that were expressed in both datasets and estimated the precision and recall in relation to TEs from GENCODE across various polyA read support levels. Importantly, out of approximately 20,000 protein-coding genes, 18,271 contained at least one PASC from both datasets.

The precision and recall metrics varied depending on the PAS support and reached the optimal F_1 (4.4.4) score when $P = 0.57 - 0.58$ and $R = 0.49 - 0.51$ (Figure 5-16, top left). For PASCs weighted by polyA read support, these metrics demonstrated superior performance reaching an optimal F_1 score with $P = 0.83 - 0.86$ and $R = 0.73 - 0.76$ (Figure 5-16, bottom left). In relation to Atlas GTEx exhibited moderate performance with $P = 0.66$ and $R = 0.30$. Notably, a significant fraction of PASCs from Atlas were undetected (Figure 5-16, top right). Yet, when weighting PASCs by the number of polyA reads, precision and recall improved to 0.92 and 0.97, respectively. This suggests that GTEx primarily overlooks PASCs with minimal read support (Figure 5-16, bottom right).

Remarkably, when compared to the annotated TEs, the two datasets showed nearly identical performance but often identified distinct low-expression sites (Figure 5-16, top row). This suggests that they can be combined to obtain a set of high-confidence PAS. In summary, further analysis verified the substantial consistency between the two PAS sets. The larger dataset substantially enhances the quality of polyA-read-based *de novo* PAS identification from RNA-seq data.

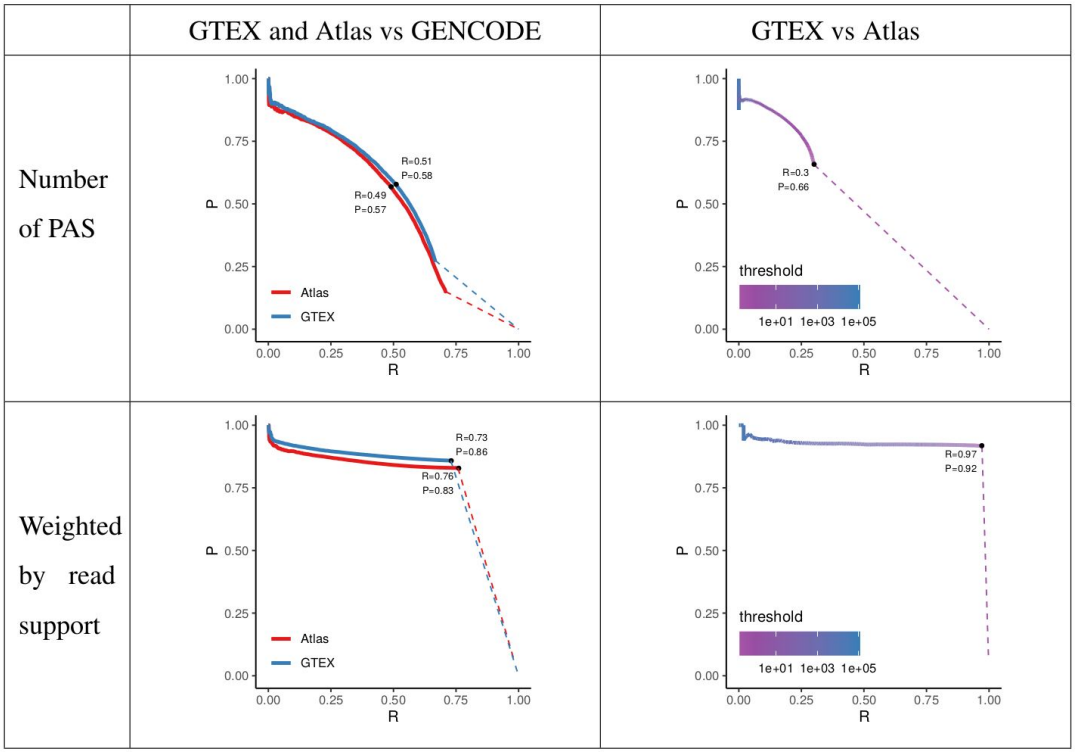


Figure 5-16: **Precision-recall curves for PASCs validation against Atlas and GENCODE**. The comparison of PASCs from GTEX and Atlas vs. GENCODE (left) and GTEX vs. Atlas (right) in terms of the number of PASCs (top) and the number of polyA reads (bottom) using precision (P) and recall (R) metrics. The curves are obtained by varying the threshold on polyA read support. The values of precision and recall with the largest F_1 metric are shown.

Given that 85% of the newly identified PASCs lacked an annotated TE within 100 nts, my focus shifted to this cohort (denoted as unannotated PASCs). I examined their relative position within the gene, which is equal to 0% and 100% for the 5'-end and 3'-end of the gene, respectively (Figure 5-17A). Even when TEs were excluded from consideration, a pronounced rise in PASC density towards the 3'-end was evident for both signal-present and signal-absent groups. A subtler increase was

discernible at the 5'-end. This reinforces the prevalent trend of PASCs appearing more frequently towards a gene's 3'-end, a pattern mirrored by the unannotated PASCs from Atlas (Figure 5-17B). Notably, 89% of the PASCs listed in Atlas also lacked 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.

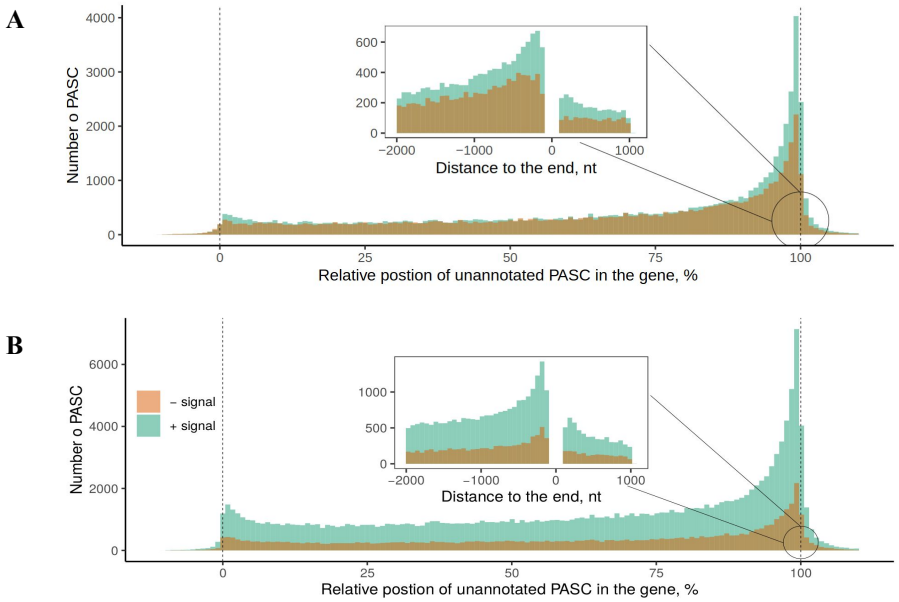


Figure 5-17: **Unannotated PASCs in protein-coding genes.** The relative positions of unannotated PASCs (i.e., ones not within 100 bp of any annotated TE) along the gene length. 0% and 100% correspond to the 5'-end and 3'-end of the gene, respectively. The inset shows the distribution of absolute positions of unannotated PASCs around the gene end. **(A)** PASCs from GTEx, **(B)** clusters from Atlas.

5.1.5 Saturation analysis

Analysis of the GTEx RNA-seq dataset yielded 150 times more predicted PAS than the processing of the small 13-sample dataset. To determine how much the number of identified PAS might increase with a further increase in total sequencing depth, I conducted a down-sampling saturation analysis (4.4.3). The number of PAS clusters observed consistently rose as more RNA-seq reads were incorporated into the analysis (Figure 4-4).

Subsequently, the PAS identification pipeline was applied to expanding subsets of the GTEx RNA-seq dataset. I slightly adjusted the pipeline and filtered the PASCs

using a per-sample entropy criterion of $H_{sample} \geq 2$. Nonetheless, the count of PASCs grew steadily with the sample size and, consequently, with the aggregate number of RNA-seq reads analyzed (Figure 5-18A). Notably, the rate of newly detected PASCs per sample decelerated as the total read depth approached its maximum. This continuing growth implies that the PAS set may not yet be exhaustive for the 31 human tissues represented, or there may be a need for additional measures to minimize false positives. Intriguingly, while the initial thousands of PASCs primarily localized within 3'-UTRs, after accumulating more data, roughly one-third of the newly identified PASCs emerged in non-UTR regions (Figure 5-18B).

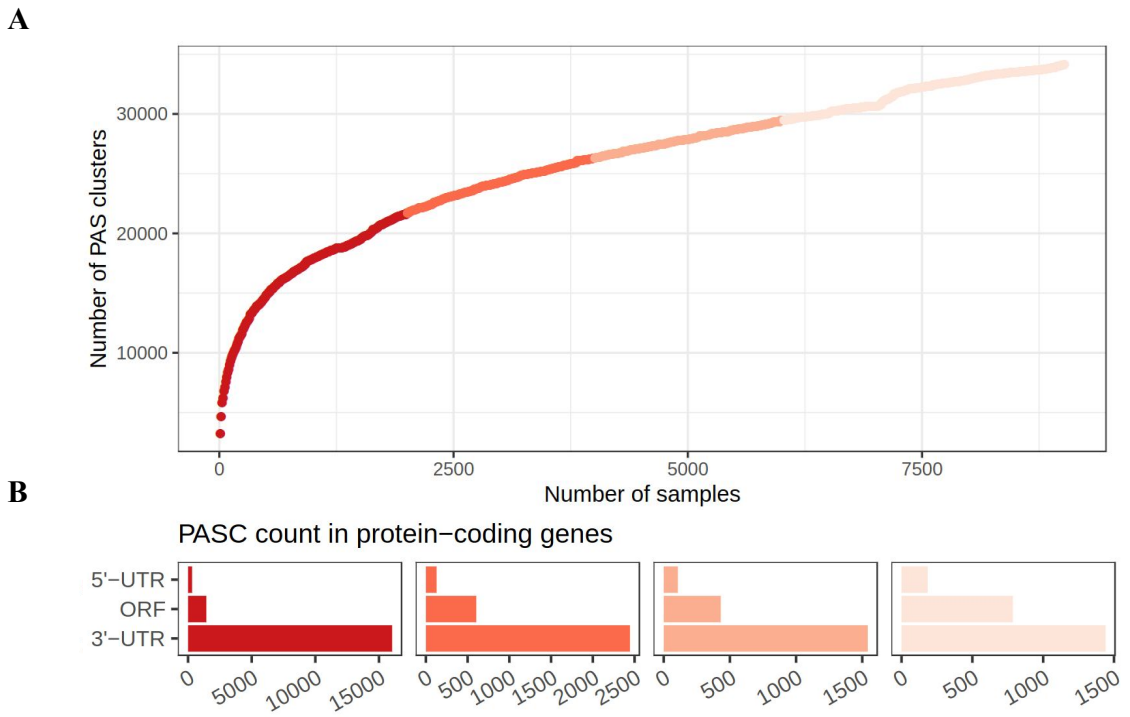


Figure 5-18: **PAS clusters in protein-coding regions at various saturation stages.** (**A**) Count of PAS clusters derived from increasing subsets of samples within the GTEx RNA-seq dataset. Sample permutations did not influence the general trajectory of the curve.(**B**) The distribution of PASCs newly identified in the respective samples across 5'-UTRs, ORFs, and 3'-UTRs. The color signifies the sample set: red represents the 1st to the 2000th sample, dark orange denotes the 2001st to the 4000th, and so forth.

5.2 Intronic polyadenylation and splicing

5.2.1 PAS clusters in protein-coding regions

I next focused on a subset of 164,497 PASCs situated within protein-coding genes and examined their distribution across different gene regions, namely the 5'-untranslated region (5'-UTR), the 3'-untranslated region (3'-UTR), and the open reading frame (ORF). By definition, the ORF did not encompass any annotated TEs, so the PASCs located there were novel. Each ORF was further dissected into intronic, constitutive exonic, and alternative exonic sections (see 4.2). Given that these regions vary in length, I assessed PASCs based on both their absolute count and their density, defined as the number of PASCs per nucleotide. Additionally, I evaluated the frequency of PASCs usage by accounting for their polyA read support (Figure 5-19, Table A.1).

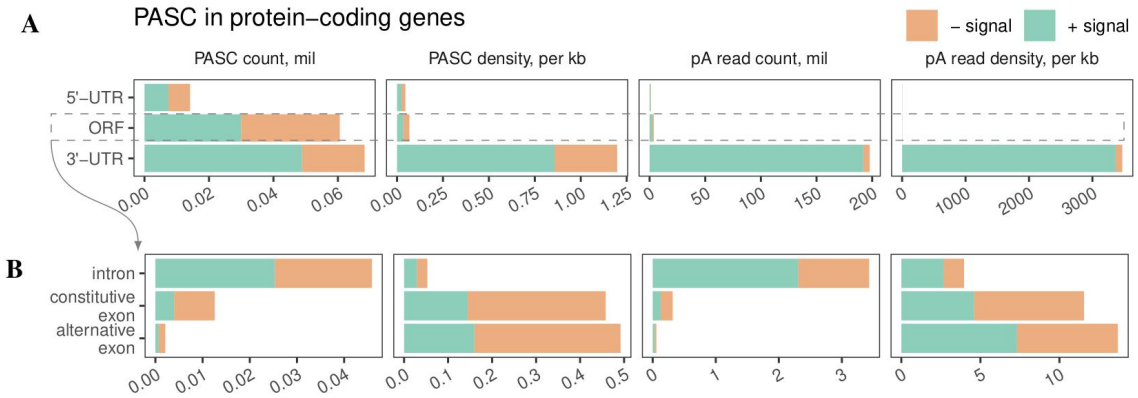


Figure 5-19: **PASCs in protein-coding regions. (A)** The distribution of PASCs in 5'-UTRs, ORF, and 3'-UTRs. Shown are the total number of PASC (PASC count), PASC density per Kb (PASC density), the total number of polyA reads (polyA read count), and the total number of polyA reads per kb (polyA read density). **(B)** The distribution of PASCs from ORF in introns, constitutive exons, and alternative exons. PASC located within 2bp of exon borders were excluded.

As anticipated, 3'-UTRs and ORFs contained a considerable number of PASCs. However, the highest density of PASCs was observed in 3'-UTRs because ORFs are typically longer (Figure 5-19A). This concentration in 3'-UTRs became even more evident when considering the number of polyA reads. While introns held the majority of PASCs in absolute terms, their density ranked the lowest after normal-

ization (Figure 5-19B). Sequence analysis revealed that every spliced protein-coding transcript harboured a cryptic intronic PAS, i.e., it possessed a motif resembling the canonical consensus CPA signal within an intron. Hence, over 82% of these transcripts included at least one intronic PASC from either GTEx or PolyASite2.0 (68,060 out of 82,506).

The positional distribution of PASC showed a distinct peak towards the end of exonic regions and the beginning of intronic ones (Figure 5-20). Specifically, PASC density was significantly elevated in the final nucleotides of exons (and the initial ones of introns). Similar peaks were observed for PolyASite 2.0 (Figure 5-21). Although both PAS sets displayed these spikes, a detailed examination of individual events indicated they might be artifacts from the mapping procedure (discussed in chapter 6).

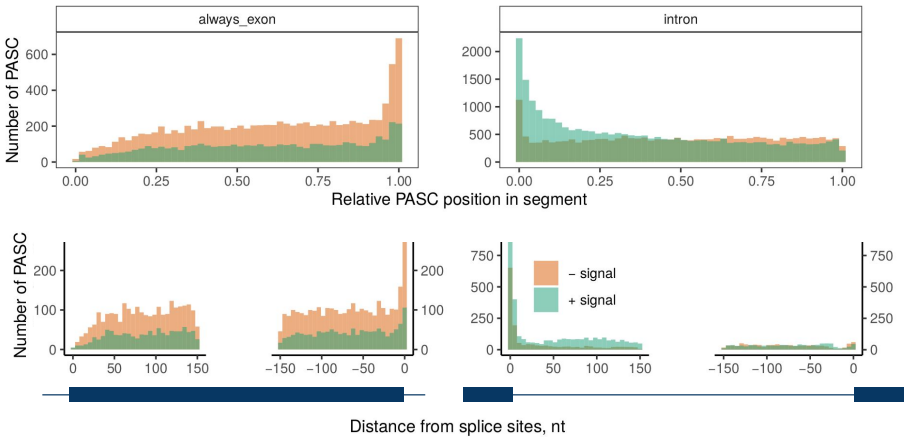


Figure 5-20: **Positional distribution of PASCs from GTEx in exons and introns.** The relative (top) and absolute (bottom) positions of PASCs from GTEx in constitutive exons and introns.

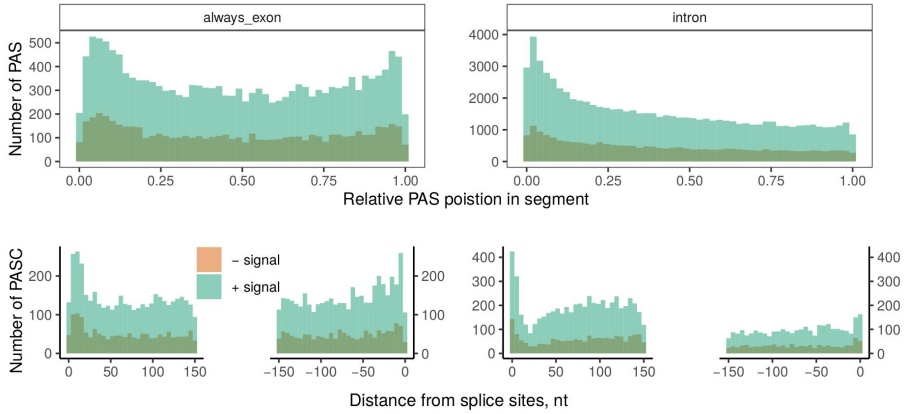


Figure 5-21: **Positional distribution of PolyASite 2.0 clusters in exons and introns**. The relative (top) and absolute (bottom) positions of PASCs from Atlas in constitutive exons and introns.

Despite the low density, intronic PASCs were quite frequent by absolute number, and among them, there could be PASCs leading to premature CPA. Current models assume that introns containing PASs cannot undergo splicing after they are cleaved and polyadenylated [158]. In this work I challenged 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, can complete intron excision even after CPA has already occurred in it. In this case, the co-occurrence of the two processes would result in short polyadenylated RNAs that could be present in the cell for some time before they are degraded by 5'exonucleses. Thus, products of some of the intronic CPA events would still be visible in RNA-seq as 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 RNA G-quadruplex formation [196, 11].

To estimate the CPA rate, at which it acts on the nascent pre-mRNA, and to account for the bias arising from intron degradation, I normalized the polyA read count to the average read coverage in exons and introns. I discovered that the relative density of polyA reads within introns is substantially larger than within exons (Figure 5-22, Table A.2).

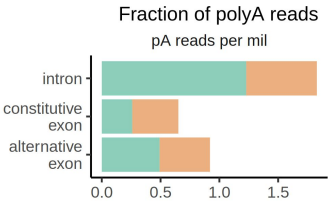


Figure 5-22: **Fraction of polyA reads.** The number of polyA reads normalized to the average read coverage in each region (defined as the number of polyA reads per million aligned reads; see 4.4.6 for details).

Furthermore, I matched introns and constitutive exons by the read coverage (Figure 5-23) 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, I 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 5-24, left) and their density per nucleotide (Figure 5-24, middle). This enrichment remained significant in other read coverage ranges (Figure 5-24B, C). In sum, this 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.

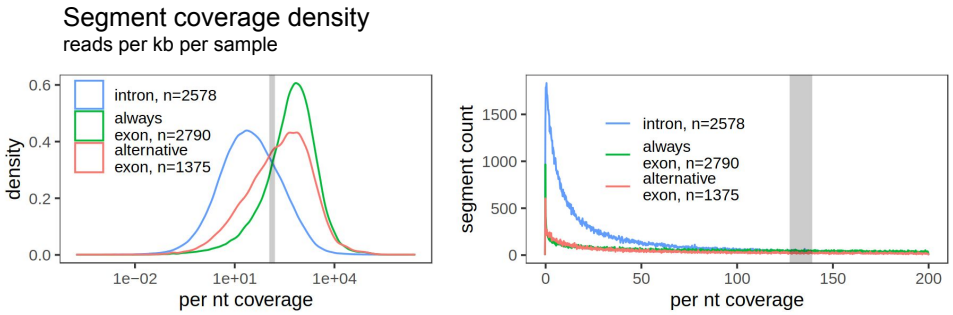


Figure 5-23: **Matching genomic intervals by the read coverage density.** The distribution of read coverage density values in three types of genomic segments. For each segment, the mean read coverage was computed as the number of reads per kb per sample. The segments with average read coverage of 133 ± 6.7, which corresponds to the intersection of density curves (highlighted by grey area), were selected for the analysis in Figure 5-24A.

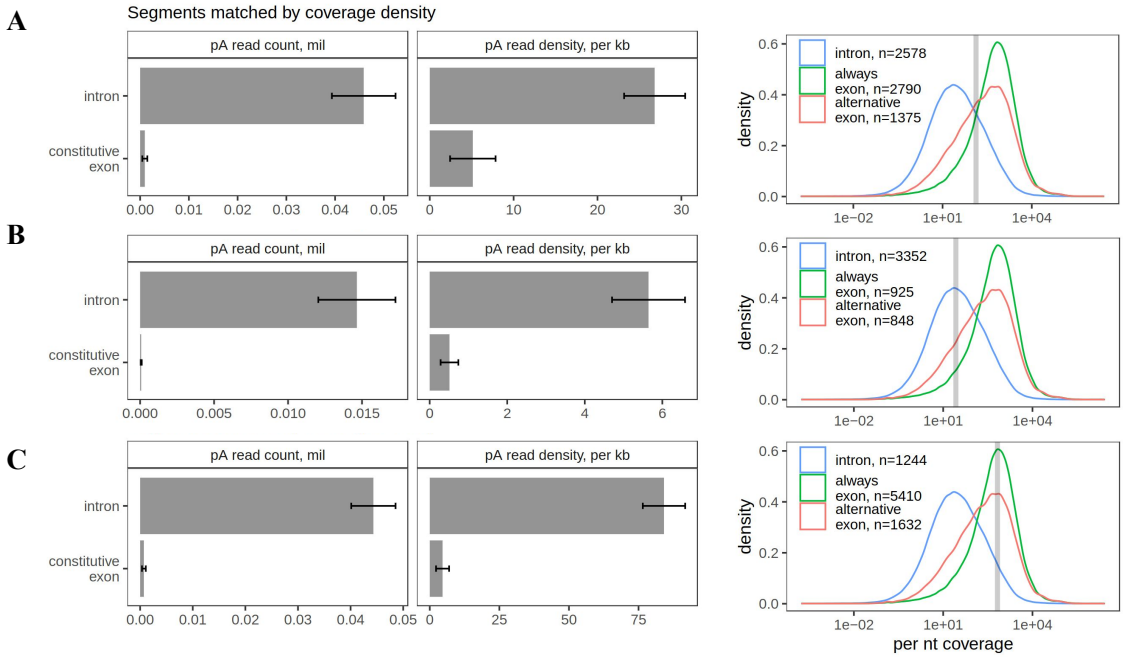


Figure 5-24: **The number of polyA reads in segments matched by the read coverage density.** (**A**) The number and density of polyA reads in segments matched by the read coverage density for coverage values highlighted in Figure 5-23. (**B**) the same procedure was repeated for the average read coverage interval 27 ± 5. (**C**) the same procedure was repeated for the average read coverage interval 666 ± 5.

5.2.2 Coverage-based metrics of PASC usage

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. Only 16.5% of PASCs are supported by ten or more polyA reads in more than one tissue.

On the other hand, the rate of PASC usage in tissues can be measured by coverage-based methods, as their positions have been already identified. Here, I adapted a simple procedure from [88], in which the average read coverage was measured in 150-nt windows, wi_1 and wi_2, before and after each PASC. To quantify the frequency of cleavage at a specific PASC, I used the $FC = wi_1/wi_2$ metric, which captures the magnitude of read coverage drop at the PASC location (Figure 5-25A).

There are several strategies to classify a PASC as "used in a tissue" based on this Fold Change metric. While it is possible to compare FC against a predetermined threshold, such an approach does not take sample variation into consideration. As a result, I also evaluated the significance of FC using the Wilcoxon signed-rank test and another method based on `DESeq2` [98] (see 4.4.6).

First, I analyzed the set of 164,497 PASCs located in protein-coding genes. This was done by pooling read coverage profiles from all GTEx samples and excluding PASCs situated within 200 nts of splice sites to avoid potential interference from read coverage drops at exon-intron junctions. In the resulting set of 126,310 PASCs, the median read densities in wi_1 and wi_2 were approximately 5.9 and 2.4 reads per nucleotide per sample, respectively, indicating at least twofold average drop after PASCs.

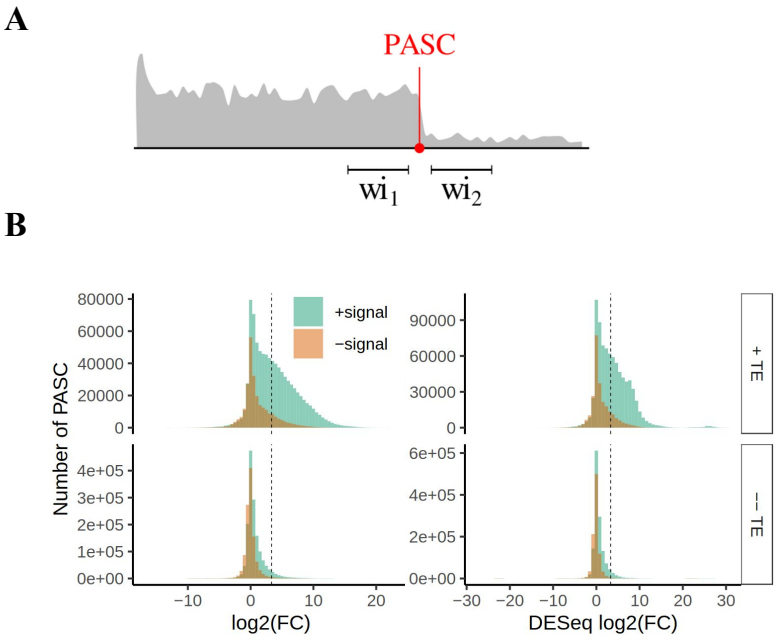


Figure 5-25: **Coverage-based metrics of PASC usage. (A)** The average read coverage was measured in 150-nt upstream and downstream windows, wi_1 and wi_2, around PASC. **(B)** The distribution of $\log_{10}(wi_1/wi_2)$ metric for annotated ($n = 37,194$, top) and unannotated PASCs ($n = 89,116$, bottom). A PASC is referred to as annotated if it is within 100 bp of an annotated TE. The dashed line represents the cutoff $wi_1/wi_2 = 10$.

Subsequently, I computed the average read density in wi_1 and wi_2 for every PASC in each tissue. As anticipated, the ratio wi_1/wi_2 was skewed to the right, and this

bias was more pronounced for PASCs with a signal and those adjacent to annotated TEs (see Figure 5-25B). The distribution of fold changes did not exhibit significant variations among tissues (as shown in Figure A-4). The only standout was Whole Blood comprising atypical samples with exceedingly high counts of polyA reads. These were previously ruled out from the PAS identification pipeline (see 5.1.2 and Figure 4-1). Consequently, I also omitted Whole Blood from subsequent analyses.

For certain PASCs, the fold change values reported by `DESeq2` were smaller than manually computed ones due to sample-specific normalization factors (see Figure 5-25B and 4). Nonetheless, categorization based on a predetermined threshold appeared more stringent. Of the 126,310 PASCs, an average of 18,470 per tissue (or 15%) exhibited a wi_1/wi_2 ratio greater than 10. Meanwhile, the `DESeq2` analysis identified a significant difference in read coverage between wi_1 and wi_2 for an average of 43,615 (or 35%) PASCs. For every tissue, around 90% of PASCs with a wi_1/wi_2 ratio exceeding 10 were also reported in the `DESeq2` findings. Considering PASC-tissue combinations, both methodologies discerned a significant coverage reduction in 504,469 pairs out of 3,789,300 (illustrated in Figure 5-26). Given the considerable overlap between the results of both techniques and the more stringent nature of the threshold-based method, I decided to designate a PASC with $wi_1/wi_2 > 10$ as "used in the respective tissue".

I subsequently compared the set of used PASCs to a reference set that included 689,346 PASs in 3'-UTRs of human genes, which were derived from GTEx using the `DaPars` algorithm [62, 176]. Since the exact positions of PASs in 3'-UTRs can vary, I selected 3'-UTRs with at least one used PASC based on the $FC > 10$ criteria. These were then contrasted with 3'-UTRs that `DaPars` identified as expressed in genes with multiple annotated 3'-UTRs. On average, 81% of 3'-UTRs containing a PASC with $FC > 10$ were also designated as expressed by `DaPars`. Conversely, 51% of the 3'-UTRs labelled as expressed by `DaPars` contained at least one PASC with $FC > 10$. This confirms that the usage of PASCs in tissues, as measured by the FC metric, and the `DaPars` findings are consistent.

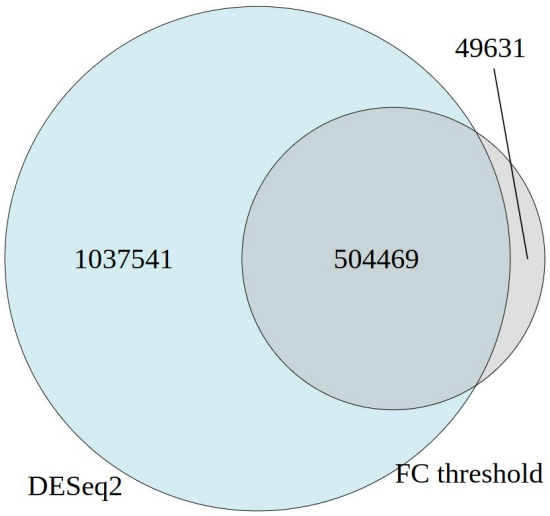


Figure 5-26: **The overlap of used PASC sets identified by DESeq2 and threshold-based rule.** Venn diagram illustrating the overlap between used PASC sets identified by DESeq2 (left) and the simple $wi_1/wi_2 > 10$ rule (right). Both methods detected a significant coverage drop in 504,469 of the 3,789,300 PASC-tissue pairs evaluated.

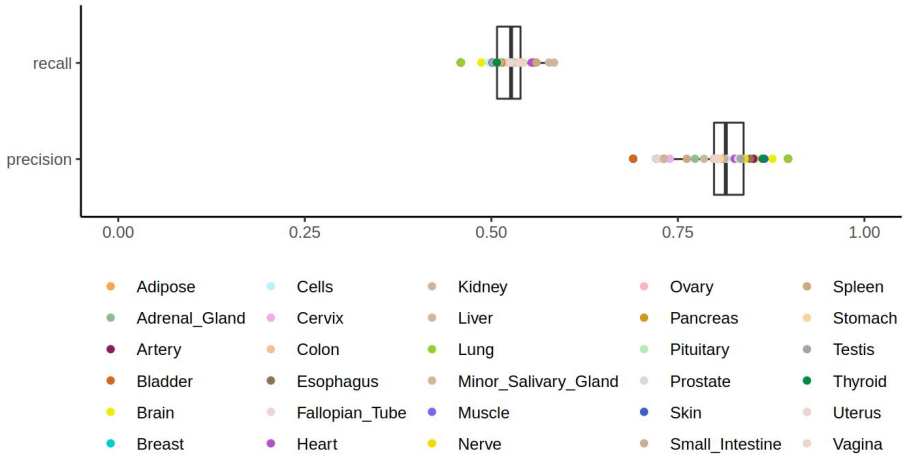


Figure 5-27: **Comparison of expressed 3'-UTRs identified by DaPars and the threshold-based rule.** This compares 3'-UTRs containing at least one used PASC based on the $FC > 10$ criterion with those identified as expressed by DaPars in each tissue. The recall represents the fraction of 3'-UTRs in the DaPars set that also appears in our dataset for a given tissue. Precision denotes the reverse: the percentage of 3'-UTRs with a PASC defined by $FC > 10$ that DaPars also recognizes as expressed in the same tissue. Each dot indicates a value for a specific tissue.

As mentioned earlier, the density of polyA reads was not adequate to analyze tissue-specific polyadenylation. However, there was a positive correlation between

the number of reads supporting a PASC and the wi_1/wi_2 ratio. This held true not only for PASCs proximate to annotated TEs but also for unannotated PASCs with a signal (see Figure 5-28). It once again highlights the ubiquity and biological relevance of these unannotated PASCs.

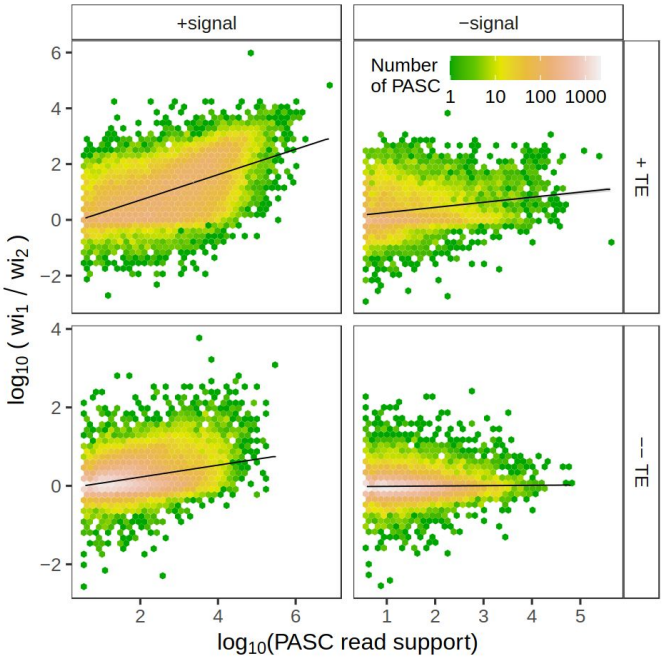


Figure 5-28: **Coverage drop correlates with polyA read support.** The $\log_{10}(wi_1/wi_2)$ metric positively correlates with the number of supporting polyA reads not only for annotated PASCs but also for unannotated PASCs with a signal.

Based on the *FC* metric, I identified approximately a hundred unannotated PASCs with pronounced tissue-specific usage (see Figure A-5). Among all tissues, the testis and brain exhibited the most distinct PAS usage profiles. One notable PAS is located within the 3'-UTR of the Synaptopodin-2 (*SYNPO2*) gene, which is linked to Nephrotic Syndromes and several cancer types. Synaptopodin-2 is an actin-binding protein [70]. Interestingly, this PAS is predominantly used in the muscles and heart. Cleavage at the PAS location eliminates a binding site for the microRNA miR-760, which has been demonstrated to regulate skeletal muscle proliferation [156]. This suggests that usage of this PAS might serve as a strategy to bypass regulation by miR-760.

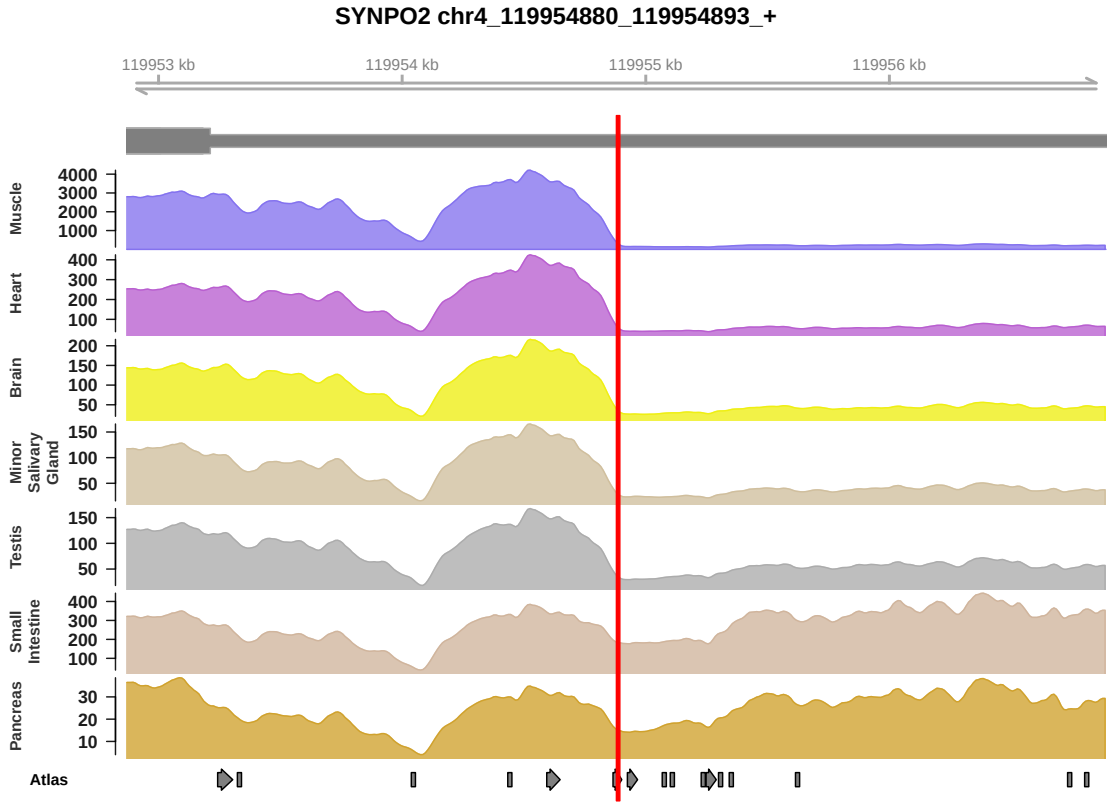


Figure 5-29: **An example of a tissue-specific PAS.** The figure depicts read coverage surrounding an unannotated PASC in the *SYNPO2* gene across six tissues. Cleavage at this PASC is more frequent in the Muscle, Heart and Brain than in Pancreas and Small intestine. The PASC is situated within the 3'-UTR of the ENST00000429713 transcript. A red vertical line marks the location of the PASC. The bottom panel displays PASs from the Atlas in proximity to the PASC.

The previous studies have extensively characterized tissue-specific polyadenylation in the GTEx dataset using coverage-based methods and focusing on polyadenylation in 3'-UTRs. Thus, I shifted my focus away from PASC tissue-specificity, delved into intronic PASCs (iPASCs) and examined the interplay between intronic polyadenylation (IPA) and alternative splicing (AS).

5.2.3 Intronic polyadenylation and alternative splicing

According to [159], an alternative terminal exon generated through can be attributed to one of the two classes: Skipped Terminal Exon (STE), which is a result of cleavage and polyadenylation in a cassette exon, and Composite Terminal Exon (CTE), which arises from cleavage and polyadenylation in a retained intron (Fig-

ure 5-30). To discern between these scenarios and identify the alternative splicing event associated with each IPA event, I calculated the average read coverage in two windows, we_1 and we_2, at the exon-intron boundary. For the sake of clarity, the read coverage values in the four windows will be denoted as we_1, we_2, wi_1, and wi_2. I anticipated that, besides a large wi_1/wi_2 ratio, an STE would exhibit a substantial we_1/we_2 ratio, whereas a CTE would show a minimal we_1/we_2 ratio (Figure 5-30). Essentially, the cassette exon and retained intron can be differentiated by the relative coverage immediately following the 5'-splice site; it should be minuscule in the former scenario and proportionate to the rate of alternative splicing in the latter.

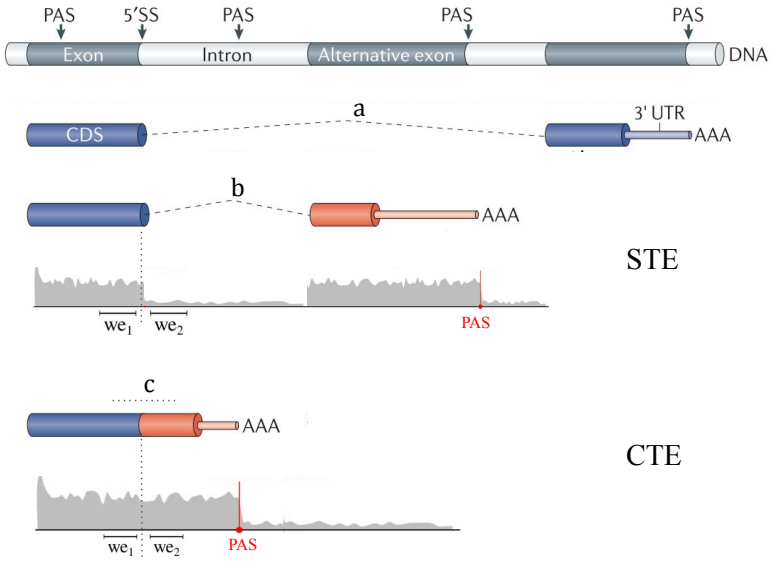


Figure 5-30: **IPA and associated alternative splicing events.** The figure presents three isoforms and their corresponding terminal exon types. The isoform with the skipped terminal exon (STE) is in the middle, and the isoform with the composite terminal exon (CTE) is at the bottom. Corresponding expected coverage profiles are shown beneath the transcript schematics. Exonic (we_1 and we_2) and intronic (wi_1 and wi_2) 150-nt windows were used for the classification. Here, a denotes split reads supporting the canonical splice junction; b signifies splice junctions starting at the 5'SS and landing before the iPASC; and c corresponds to the continuous reads spanning the 5'SS. Dashed lines indicate splicing events. Parts of the figure were adapted from a review by B. Tian and J.L. Manley [158].

To quantify the rate of splicing, I computed the number of split reads starting at the intron 5'-end and landing before the iPASC (b), after the iPASC at the canonical 3'-splice site (a), and the number of continuous reads (c) that span the exon-intron boundary (Figure 5-30). The canonical 5'SS and 3'SS were defined as the borders

of the shortest annotated intron containing the iPASC. These three 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 $\psi \simeq 0$ indicated the presence of alternative splicing events before the iPASC. Of note, ψ is a relative quantity, which is not influenced by the read coverage. I expected that both Skipped and continuous terminal exons would be 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. Thus, quantification of the split and continuous reads spanning 5' splice sites represented the second approach to terminal exon classification.

PASCs that lacked the canonical polyadenylation signal did not exhibit a positive correlation between the wi_1/wi_2 rates and the number of reads supporting the site. Additionally, they exhibited a smaller coverage drop compared to PASCs with a signal (see 5.2.2). Thus, in what follows, I confined the analysis to conservative iPASCs set with a signal. The values of we_1, we_2, wi_1, wi_2, and ψ were computed for a total of 1,115,690 iPASC-tissue combinations, encompassing 35,990 iPASCs across 31 tissues.

As discussed, the current paradigm is that an intron that contains a PAS cannot be spliced out after CPA machinery cleaves the pre-mRNA at this PAS. Thus, a negative association between canonical splicing rate and CPA rate in the corresponding intron was expected. Indeed, I observed the anticorrelation between ψ and iPASC usage measured by polyA read support or wi_1/wi_2 (Figure 5-31, left). 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 5-31, right). The read coverage at iPASC changed two orders of magnitude when ψ increased from 25% to 100% in some remarkable cases (Figure 5-32). These observations indicate that general trends in the data are in agreement with the current paradigm and once again reconfirm that splicing and CPA naturally counteract each other.

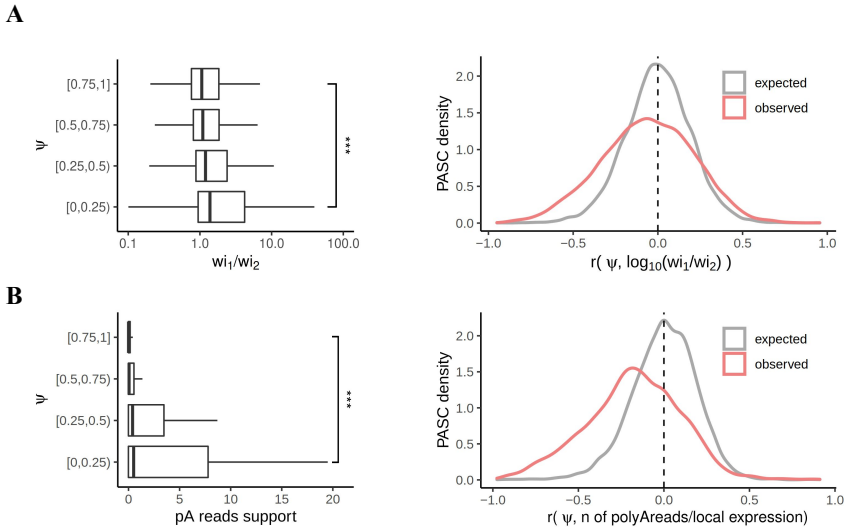


Figure 5-31: **Negative association between canonical splicing rate and CPA rate. (A)** The wi_1/wi_2 for iPASCs in four ψ quartiles (left); *** denotes the 0.1% significance level in Wilcoxon rank-sum test. Pearson correlation coefficients of ψ and $\log_{10}(wi_1/wi_2)$ for $n = 12,261$ iPASCs compared to the label-shuffled control (right). **(B)** The relative polyA read support of iPASC in four ψ quartiles (left). Pearson correlation coefficients of ψ and the relative polyA read support for iPASCs compared to the label-shuffled control (right). The relative polyA read support was computed as the ratio between the number of polyA reads supporting the iPASC and the ψ denominator (also referred to as local expression).

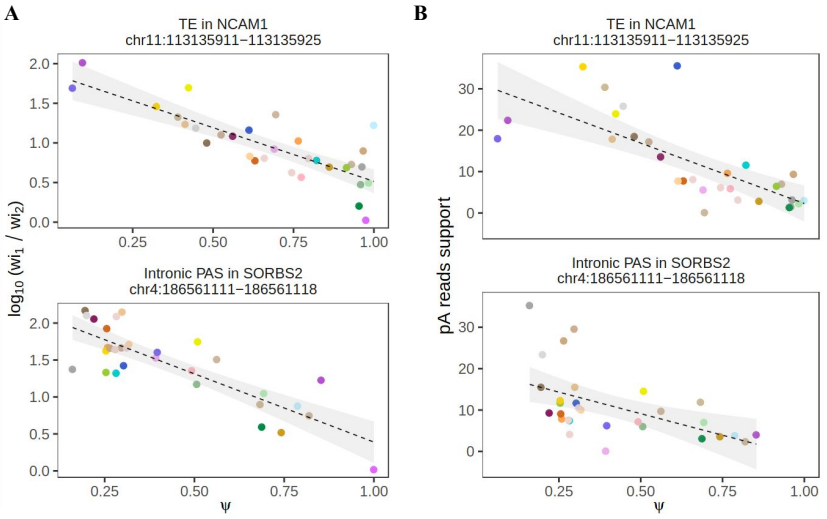


Figure 5-32: **Negative association between canonical splicing rate and CPA rate. Examples. (A)** Negative association between ψ and $\log_{10}(wi_1/wi_2)$ in the *NCAM1* and *SORBS2* genes. **(B)** Negative association between ψ and the relative polyA read support in the *NCAM1* and *SORBS2* genes. Tissue colours are the same as in the figure 5-27.

I utilized the coverage data to identify the tissues in which each iPASC was used. A total of 75,501 iPASC-tissue pairs exhibited a pronounced read coverage drop at the iPASC ($wi_1/wi_2 > 10$) and a notably high read coverage in the upstream intronic window ($wi_1 > 0.1we_1$). Overall, 7,427 iPASCs were used in at least one tissue. The class of an associated terminal exon (STE or CTE) could be deduced from the GENCODE transcript annotation for 1,506 of these iPASCs (see 4.4.6). The bivariate distributions of $\log(we_1)$ and $\log(we_2)$ for the 439 annotated CTEs and 1,067 annotated STEs were separated by the line $we_2 = 0.25we_1$. As expected, the former clustered above this line, while the latter were below it (Figure 5-33, left and middle). iPASCs of unidentified alternative terminal exon type displayed a composite of these two distributions (Figure 5-33, right). A similar pattern was observed in the bivariate distributions of $\log(wi_1)$ and $\log(we_2)$.

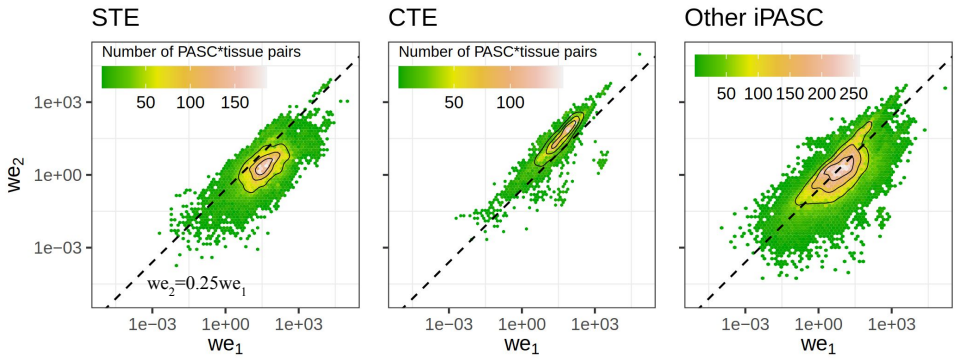


Figure 5-33: **Bivariate distribution of we_1 vs. we_2** in PASC-tissue pairs for expressed CTE ($n = 1,136$ pairs), STE ($n = 1,948$ pairs), and other used iPASCs ($n = 32,906$ pairs). The dashed line corresponds to $we_2/we_1 = 0.25$.

Moreover, ψ values of expressed annotated CTE and STE were characterized by a single peak at $\psi \simeq 0$ indicating the absence of canonical splicing (Figure 5-34, left and middle). In STEs, reads that support AS were primarily represented by the split-reads (b), whereas in CTEs, the continuous reads spanning 5'SS (c) dominated (see Figure 5-35). Unexpectedly, the ψ values of other used iPASCs had a pronounced second peak at $\psi \simeq 1$ formed mostly by iPASCs without the TE support (Figure 5-34, right). The absence of annotated transcript ends around these iPASCs suggests that they are not endpoints of short non-coding transcripts situated within the intron. This peak is incompatible with the current paradigm

and CTE and STE models because it implies that intronic polyadenylation coexists with canonical splicing. Accordingly, in what follows these events will be referred to as Spliced Polyadenylated Introns (SPIs).

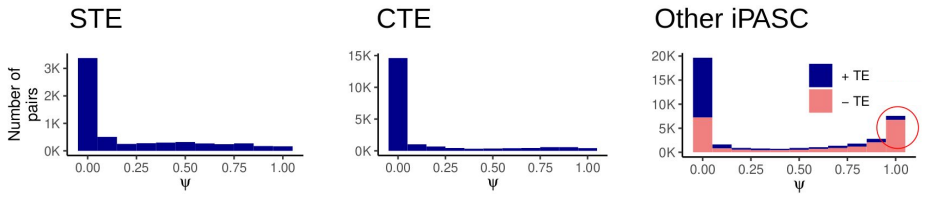


Figure 5-34: **The distribution of** ψ for CTE, STE, and other iPASCs; +TE (−TE) denote iPASCs within (not within) 100 nts of an annotated TE.

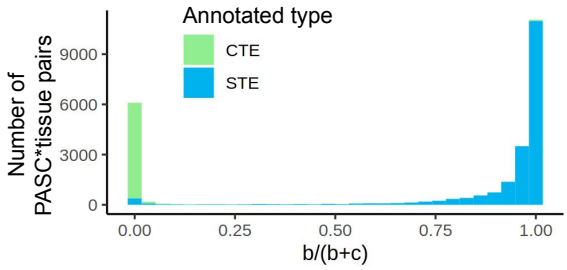


Figure 5-35: **The distribution of reads supporting cassette exon and retained intron AS events in annotated CTE and STE iPASCs.** The fraction of split reads supporting cassette exon events (b) among all reads supporting AS events upstream of the iPASC for annotated CTE and STE iPASCs i.e. the distribution of $b/(b+c)$.

To discern the origin of SPIs, I focused on events with $\psi > 0.9$ substantially supported by 5'SS continuous and split-reads (local expression or $a+b+c \geq 30$, n=7960), and compared we_2/we_1 and we_2/wi_1 distributions among STEs, CTEs, and SPIs (Figure 5-36). Similarly to skipped terminal exons, SPIs were characterized by a low coverage in the intron 5'-end relative to the exon (same as high we_1/we_2, Figure 5-36B). However, when normalizing the coverage upstream of the PAS by the coverage at the intron 5'-end, SPI values exceeded those of CTEs but were less than the typical values for STEs (wi_1/we_2 in Figure 5-36B). At this point I preliminary concluded that SPIs were associated with a substantial drop of coverage at the 5'SS followed by its partial restoration upstream of the iPASC. However, an examination of specific examples was required to form a detailed understanding of the coverage profiles.

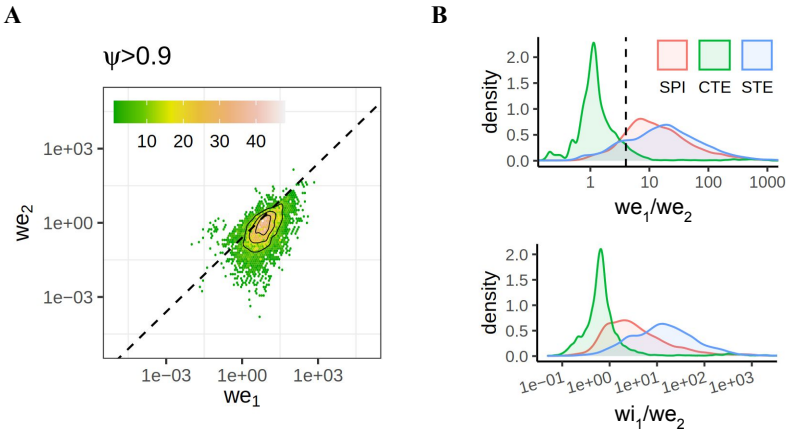


Figure 5-36: **Coverage profile distributions in STE, CTE, and SPI iPASCs.** **(A)** Bivariate distribution of we_1 vs. we_2 in PASC-tissue pairs for used iPASCs with $\psi > 0.9$ and local expression ≥ 30 ($a + b + c \geq 30$, n=7960 pairs). **(B)** The distribution of we_1/we_2 (top) and wi_1/we_2 (bottom) values for CTE, STE, and SPI. The vertical dashed line denotes $we_2/we_1 = 0.25$.

Thus, I followed up on a few cases of tissue-specific splicing and CPA. In the Attractin (*ATRN*) gene, which encodes a transmembrane protein linked to kidney and liver abnormalities in mice [4], an intronic polyadenylation site is used in muscle and heart (Figure 5-37). Substantial coverage drop at the iPASC along with the elevation of read coverage upstream of it and simultaneous activation of an adjacent acceptor splice site, suggested that the PAS is associated with an unannotated tissue-specific skipped terminal exon.

The iPASC in the *TRIP11* gene, which is related to skeletal dysplasia [108], is an example of CTE. It is supported by intronic read coverage, with no evidence of AS events preceding the PAS, most remarkably in pituitary tissue (Figure 5-38A). Cleavage at this iPASC would result in a nonfunctional protein since the PAS is in the first intron. Another example of CTE is an intronic polyadenylation site in the *TMEM38A* gene. This gene encodes a transmembrane protein that represses several myogenic genes by relocating them to the nuclear periphery [136] (Figure 5-38B). A transcript ending at the iPASC would lack two coding exons and, thus, possess a truncated TRIC (trimeric intracellular cation) domain. All three listed iPASCs are supported by CSTF2 eCLIP peaks and sites from PolyASite 2.0.

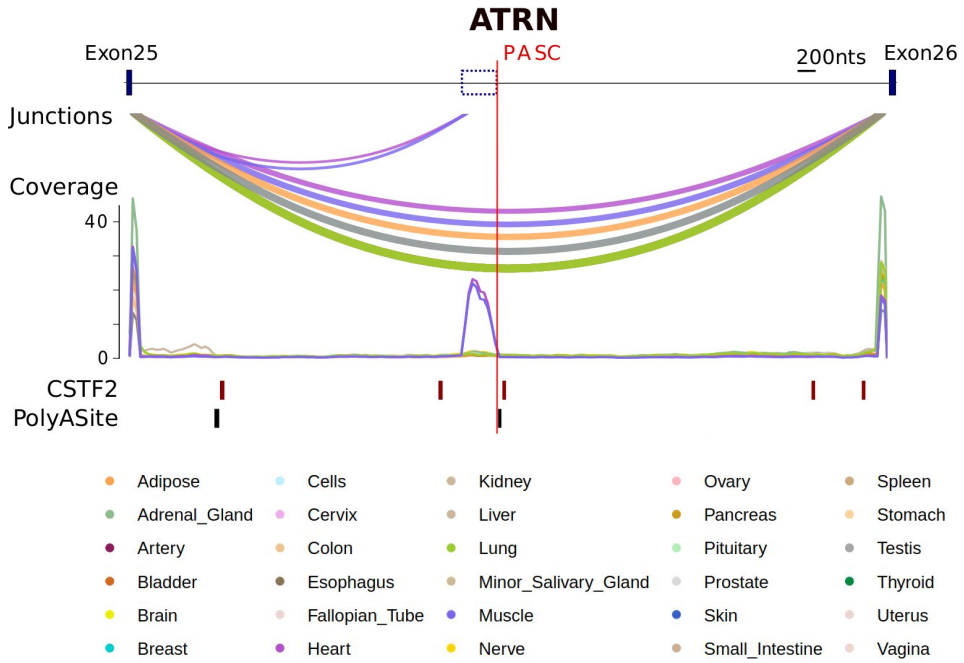


Figure 5-37: **STE case studies.** The iPASC between exons 25 and 26 in *ATRN* gene generates a STE with tissue-specific usage in heart and muscle. Arcs represent tissue-specific AS. Smoothed per nt coverage in sequenced nucleotides is shown under the arcs for each analysed tissue. The eCLIP peaks of CSTF2 and PASC from PolyAsite 2.0 are indicated in the track below.

In contrast, iPASC in the autism risk factor gene *ASH1L*, which encodes a histone methyltransferase [128], exhibits elevated read coverage in wi_1, but it lacks AS events that could support STE, or RNA-seq reads at the beginning of the intron that would validate CTE (Figure 5-39A). The iPASC is surrounded by eCLIP peaks of three known CPA factors: CSTF2, CSTF2T and CPSF6. The only plausible interpretation is that canonical splicing and IPA co-exist and operate concurrently resulting in an SPI.

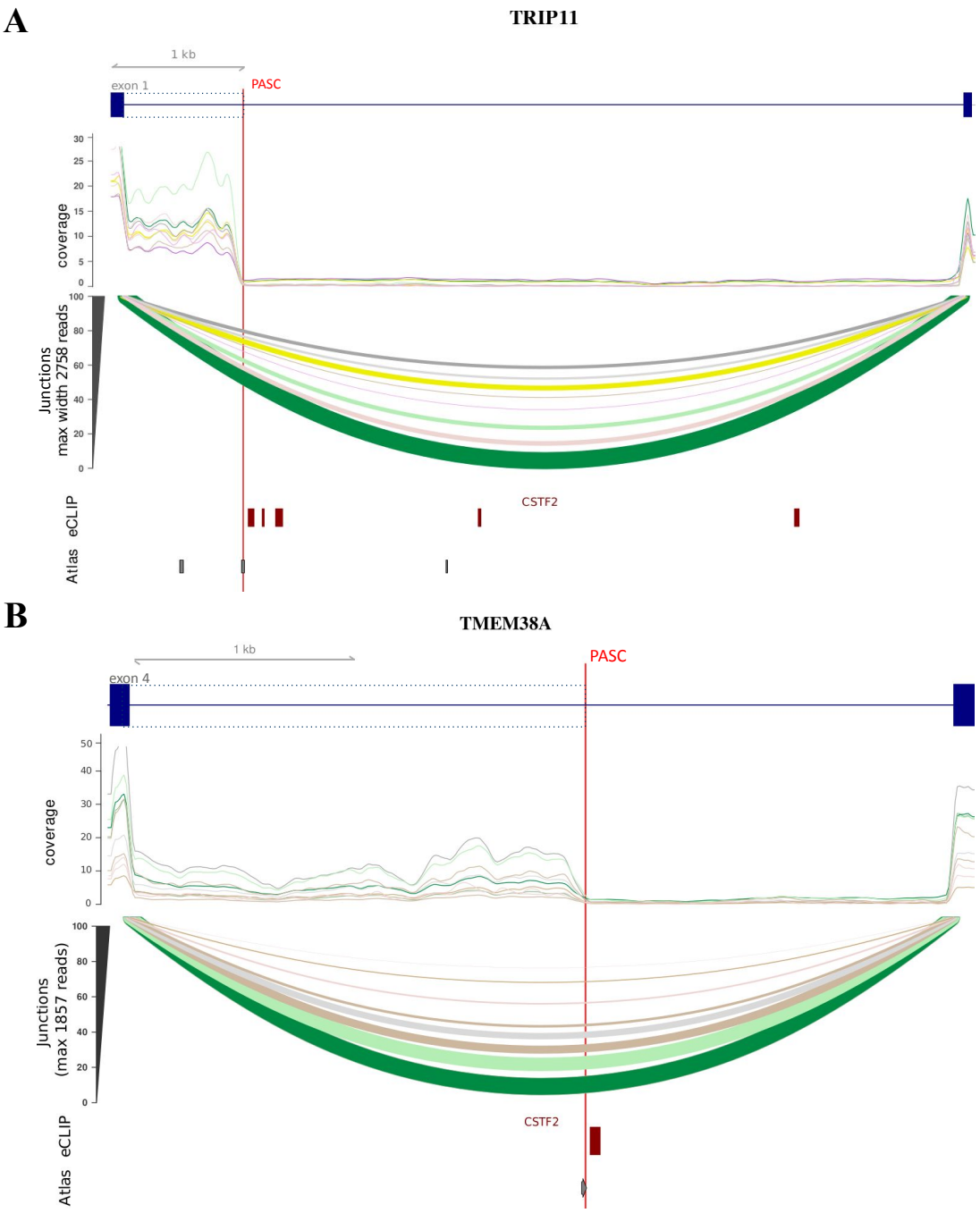


Figure 5-38: **CTE case studies. (A)** The iPASC between exons 1 and 2 of *TRIP11* generates a CTE. iPASC (position chr14:92504996-92505000 on − strand) is shown by the vertical red line. **(B)** The iPASC between exons 4 and 5 of *TMEM38A* generates a CTE. iPASC position is chr19:16795410-16795417 on +. Tissue colours are the same as in the previous figure. Widths of the arcs correlate with the number of supporting split reads, the number of reads corresponding to 100% is shown in the y-axis title.

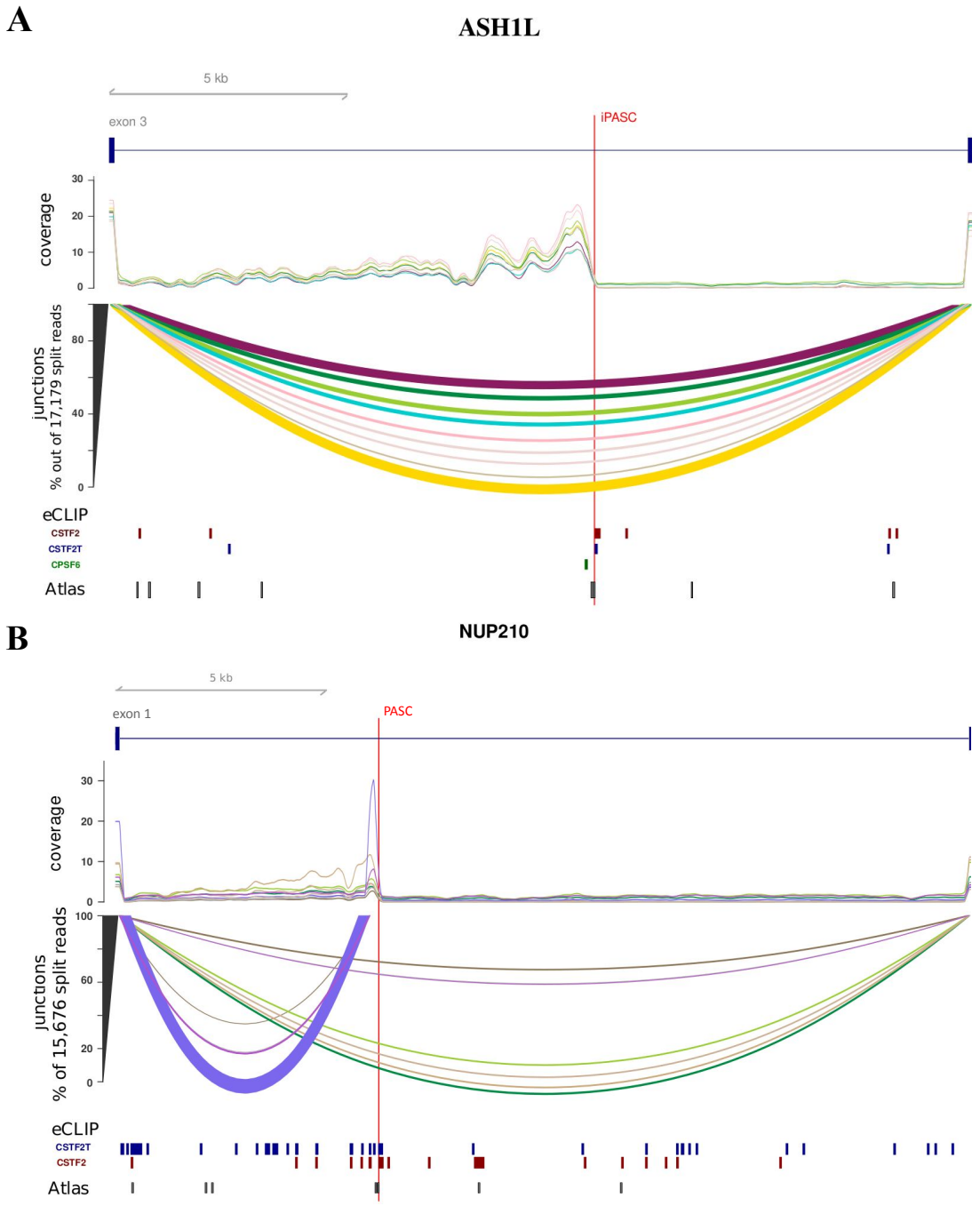


Figure 5-39: **SPI case studies. (A)** The iPASC between exons 3 and 4 of *ASH1L* likely generates a SPI because the intron 5'-end is not covered, wi_1 is covered, but there is no evidence of STE. iPASC position is chr1:155437554-155437561 on − strand. **(B)** The iPASC between exons 1 and 2 *NUP210* likely generates a STE with tissue-specific usage in heart and muscle and a SPI in several other tissues such as the spleen (beige). iPASC position is chr3:13455398-13455407 on −. Tissue colours are the same as in the two previous figures.

I also observed an intriguing case of tissue-specific shift between STE and SPI

types in *NUP210* gene, encoding nuclear pore membrane glycoprotein, which is associated with increased metastasis in breast cancer and controls muscle differentiation [2, 46]. In muscle (blue line) sharp elevation of read coverage upstream of the PAS was accompanied by activation of an alternative acceptor splice site, a similar less prominent effect was also observed in heart and esophagus tissues (purple and dark brown, respectively, in Figure 5-39B). Moreover, in these three tissues, the coverage of the second exon was much lower than the coverage of the first one. On the other hand, in the spleen (beige) the increase of the intronic coverage was gradual, there were no split reads supporting alternative splicing and the coverage was similar in the two exons. All these observations suggest that in muscle, heart and esophagus tissues the shorter transcript with a skipped terminal exon is expressed, while in the spleen the canonical splicing prevails and a spliced polyadenylated intron is generated.

5.2.4 Abundance and tissue-specificity of alternative terminal exons and spliced polyadenylated introns

To characterize further the abundance of each type of events, I considered a strict set of iPASC-tissue pairs described above and categorized them as continuous terminal exons, skipped terminal exons or spliced polyadenylated introns 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 (Table A.7).

In the previous section, I described a case where an iPASC behaved like an STE in a couple of tissues and like an SPI in other ones (Figure 5-39, bottom). Thus, one polyadenylation site could be attributed to different types in different tissues. I categorized a polyadenylation site as CTE, STE, and SPI if it belonged to the respective class in at least one iPASC-tissue pair. This yielded 2,846, 2,251 and 1,482 sites 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 (Figure 5-40A, A-7C). Notably, this classification approach correctly labeled 91% of annotated expressed continuous terminal exons, and 88% of skipped

type	Number of PASC-tissue pairs	Number of unique PASCs	Number of PASCs with annotated type	Recall among annotated PASCs
CTE	19573	2235	406	0.91
STE	31230	2843	1024	0.88
SPI	7194	1481	-	-
-	17286	4347	-	-

Table 5.2: **Classification of PAS-tissue pairs and individual PASCs.** The last column shows the fraction of PASCs annotated as STE(CTE) classified as STE(CTE) in at least one tissue. "-" type are used iPASCs that could not be classified as CTE, STE or SPI due to low local expression.

terminal exons (Table 5.2).

The number of iPASCs attributed to the three classes varied moderately across tissues, presumably reflecting the fact that the bulk of intronic polyadenylation events is not regulated tissue-specifically (Figure 5-40B). While introns with STE PAS were expectedly longer than those with CTE [193], introns with SPIs were even larger (Figure A-6A). Also, SPIs had a slight preference for the 5'-end of the gene (Figure A-6B). I concluded that spliced polyadenylated introns represent semi-stable intermediates that can be detected by polyA reads and 3'-end sequencing.

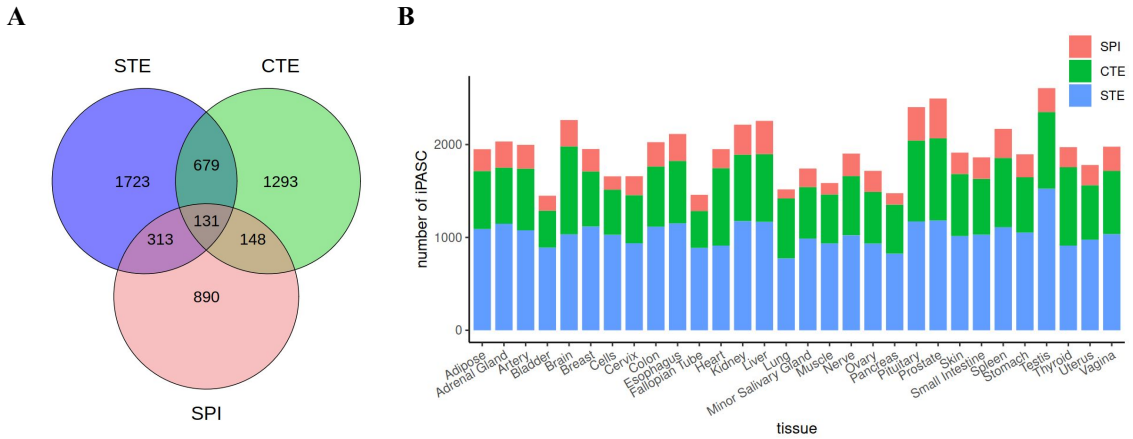


Figure 5-40: **Expression of STE, CTE, and SPI across tissues.** The number of iPASCs attributed to STE, CTE, and SPI across tissues. (**A**) Venn diagram showing the amount of iPASCs attributed to different types in different tissues. (**B**) Barplot showing numbers of iPASCs of each type in each tissue. An iPASC was categorized as CTE, STE, and SPI if it belonged to the respective class in at least one iPASC-tissue pair.

5.2.5 Linearized SPIs captured by Cleave-seq

The processing of intron lariats involves cleavage at the branch point by the debranching enzyme DBR1, followed by degradation by exonucleases [139, 115] (see 2.2). Since SPIs represent prematurely polyadenylated introns that are removed by the spliceosome, they should also possess a 2'-5' bond formed by the 5'SS guanine and the BP adenosine. Thus, I anticipated that DBR1 would also linearize SPIs. This linearization would yield two distinct molecules: one representing the intronic RNA upstream of the PAS with both a 5'-monophosphate (5'-p) and a polyA tail, and the other representing the portion of the intron downstream of the PAS.

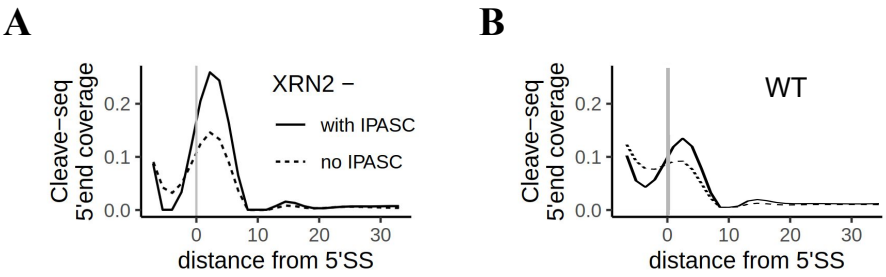


Figure 5-41: **Intron coverage in Cleave-seq data.** The Cleave-seq 5'-end coverage in introns with ($n = 21,230$) iPASC and without iPASC ($n = 199,978$) under *XRN2* knockdown **(A)** and in the wild type **(B)**.

To test this hypothesis I reanalysed a dataset generated through Cleave-seq, a protocol tailored for capturing polyadenylated RNAs with a 5'-p [155]. In line with my expectations, introns with iPASCs showed a higher 5'-end coverage by Cleave-seq reads compared to introns without iPASCs (see Figure 5-41). The linear RNA is degraded slower in cells with knockdown of *XRN2* exonuclease, thus the observed difference was much more pronounced in *XRN2* KD samples. Among the introns with iPASCs, the coverage was the highest in those containing an SPI (Figure 5-42). A similar increase in 5'-end coverage was observed in 3'-pull down *in vitro* capping experiments, another protocol designed to detect RNA molecules with the same attributes (Figure 5-43, see 4.5).

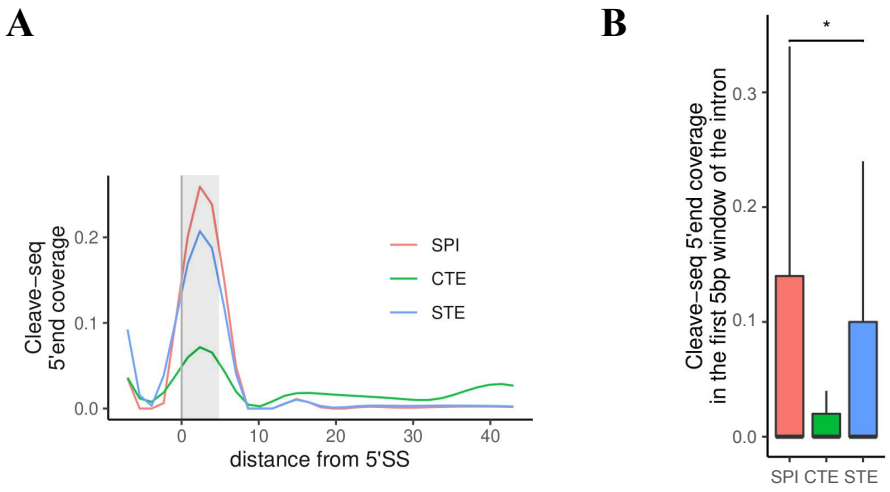


Figure 5-42: **Introns with different PAS types in Cleave-seq data.** (**A**) The Cleave-seq 5'-end coverage in introns with STEs, CTEs, and SPIs in HeLa cells following *XRN2* knockdown. Introns were categorized as containing CTE ($n = 729$), STE ($n = 957$), or SPI ($n = 533$) if they contained an unambiguous iPASC of the respective type, while introns with iPASCs of different types were discarded. The shaded region illustrates the 5bp window featured in panel (B). (**B**) The Cleave-seq 5'-end coverage in the first 5bp-window of introns with STE, CTE, and SPI iPASCs. The set of introns is the same as in A. Introns with SPIs have significantly higher coverage than introns with iPASCs of the other types. The * denotes the 5% significance level in the Wilcoxon rank-sum test.

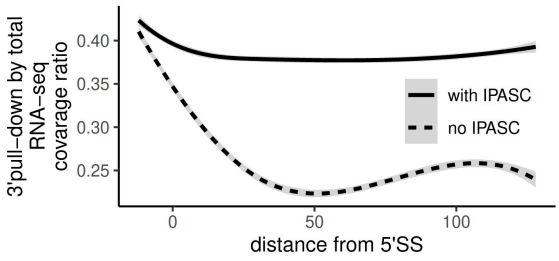


Figure 5-43: **Intron coverage in 3'-pull down *in vitro* capping data.** Normalized coverage in 3'-pull down *in vitro* capping experiments as a function of the distance from the 5'-splice site, in introns with and without iPASCs. Relative coverage is higher in introns containing a PAS. Set of introns is the same as in figure 5-41.

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. SPIs constitute a minor, yet considerable fraction of IPA events and contribute to the observed landscape of intronic polyadenylation.

Chapter 6

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 [88]. The GTEx dataset represents an ideal resource for studying this interplay because the information on the positions and tissue-specific usage of intronic PASs, which is captured by polyA reads and changes in RNA-seq reads coverage, is complemented by tissue-specific splicing rates inferred from split reads that align to splice junctions.

6.1 PAS identification

The polyA-read-based approach used in this study was previously limited to smaller datasets. In 2017, T. Bonfert and C.C. Friedel evaluated existing tools, `ContextMap2` and `KLEAT`, on three dual-replicate RNA-seq datasets. They reported recall rates under 12%, with ambiguous precision rates ranging from 39% to 95% [10, 9]. Since the main problem of the method was the scarcity of polyA reads, I hypothesized that a comprehensive RNA-seq panel could substantially enhance the quality of the resulting PAS set. Thus I used this approach for the identification of PASs at the scale of the GTEx project, which has not been attempted previously. The obtained PAS set was then combined with coverage-based methods to estimate the CPA rate.

To validate the developed pipeline for polyA read analysis, I primarily applied it to 13 RNA-seq samples derived from a matched 3'-seq/RNA-seq dataset [88].

At the peak F1 measure, merely 16% of 3'-seq PAS were recovered by RNA-seq with a precision of 23% (Figure 5-3). Notably, when the PAS from both datasets were weighted by their read support, the performance improved significantly and achieved 70% recall and 63% precision. This indicates that highly covered PAS were much more likely to coincide. A comparison of the two sets against a conservative reference set (TEs catalogued by GENCODE) in both precision and recall terms clearly favoured 3'-seq (see 5.1.1).

In a similar analysis, PASs derived from the large GTEx RNA-seq dataset were compared to PolyASite2.0 atlas, which encompasses hundreds of 3'-end sequencing libraries (see 5.1.4). The precision and recall values of the predicted PAS set and the atlas were nearly identical. This indicates that the increase in dataset size substantially enhances the quality of polyA-read-based *de novo* PAS identification from RNA-seq data.

Notably, the 3'-seq analysis failed to identify any sites in several highly expressed genes, while it successfully identified many PASs in genes with considerably lower expression levels (Figure 5-2B). This observed discrepancy between the 3'-seq- and RNA-seq-derived PAS sets can be attributed to several factors. A primary factor arises from the 3'-seq dataset analysis. As detailed in 4.3.2, I inferred strand information for 3'-seq-derived PASs based on the gene harboring the PAS. This method led to the exclusion of genes overlapping with those from a different strand. These excluded genes could not contain any 3'-seq-derived PAS. Additionally, patient variability may have influenced the results, as cells from only 9 out of 15 patients participating in the study were used for both RNA-seq and 3'-seq libraries. While differences in protocols might also contribute to the discrepancy in gene sets, previous research has shown that mRNA expression, calculated as the aggregate of 3'-seq reads supporting different PASs of a gene, is consistent with mRNA coverage in the corresponding RNA-seq data [92].

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 led to a dramatic improvement. Moreover, I attempted to minimise the rate of false positive predictions by removing non-unique and low-quality mappings, excluding A-rich genomic re-

gions and using filters on the diversity of polyA read distribution (Figure 4-3, A-2). Nevertheless, in down-sampling saturation analysis, the number of obtained PAS clusters monotonously increased with the total read depth, even toward the curve's end (Figure 5-18A). This implies either that the GTEx-based PAS set was not exhaustive for the encompassed tissues, or that stricter false-positive control was required. The former reason is supported by the shift of newly identified PASCs from 3'UTR to non-UTR parts of protein-coding genes upon the data accumulation (Figure 5-18B). The abundance of these sample-specific sites agrees with the known tissue- and context-specificity of many IPA events [114] (2.1.2).

Another possible limitation of the polyA-based method is 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 at the end of exonic and the beginning of intronic regions (Figure 5-20) resembling clusters of CAGE tags near internal exons and occurrence of polyA-seq peaks close to exon boundaries [39, 41]. 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 (for example in Figure A-8). 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 (Figure 5-21). The alignment of split reads with short exonic parts appears to be a common problem of all such methods.

One feature of the method is that polyA reads provide a snapshot of cleavage and polyadenylation at single nucleotide resolution, which revealed that PASs form clusters of varying sizes (see 5.1.3). This indicated that the precision of CPA machinery is highly variable, providing narrow clusters of closely spaced PASs in some cases and broad regions with imprecise cleavage points in others [162, 157] (Figure 5-6). Notably, the PASs distribution pattern around annotated TEs was almost identical to the reported distribution based on polyA-seq, including a characteristic skew to the left [29] (Figure 5-6A). 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 [61]. The functional relevance of stochastic variations in CPA events is currently not well understood. In this work, I explored the possible sequence determinants of CPA precision in annotated transcript ends.

Indeed, sites accompanied by the canonical polyadenylation signal had narrow PAS groups, which was most pronounced if the signal was located between 18 and 15 nts upstream of the TE (Figure 5-9). This distance scope includes the median TE and corresponds to the positional preference of the canonical upstream motif known from the literature (∼20 nts) [162, 157]. Supposedly, that is the most convenient location of the signal for its recognition by WDR33 and CPSF-30 CPA factors [152], making these TEs strong CPA sites with high polyA read support and narrow peak. Indeed, mutations in these positions were also shown to affect the PAS usage the most [162]. Moreover, the hexamer signal is the most defined CPA regulating *cis*-element both in terms of sequence and position [114] (2.1.1), which implies that it primarily determines the position of the PAS. Accordingly, my observations show that the signal position affects not only the PAS strength but also the CPA precision.

Conversely, another known GU/U-rich downstream motif did not show any association with the PAS cluster width (Figure 5-10). Previous studies suggest that this motif is crucial to PAS with weaker noncanonical hexamer signals and serves to enhance usage frequency (2.1.1). Results presented here agree with the literature and imply that GU/U-rich DSE does not influence the precision of the CPA reaction. Analysis of other potential determinants showed that only adenine content in the downstream region had a significant effect, with broader PAS clusters associated with higher adenine frequency (5.1.3). Even though PASs at the boundaries of genomic adenine stretches were excluded, regions with an adenine content just below the set threshold might still give rise to spurious PASs. In sum, I observed that CPA precision is predominantly controlled by the polyadenylation signal, and not upstream or downstream sequence elements. However, any influence from the DSEs might have been masked by mapping errors related to the high adenine content.

6.2 Intronic polyadenylation and splicing

As mentioned above, in this work I used the GTEx dataset to study the interplay between CPA and splicing because the information about the positions and tissue-specific usage of intronic PASs, captured by polyA reads and coverage-based methods, could be complemented by tissue-specific splicing rates inferred from split reads that align to splice junctions. Thus, after obtaining the PAS set I focused on the intronic sites.

In the established model, any IPA event is associated with an AS event: intron retention or a cassette exon inclusion [159]. To detect the type of AS event associated with each intronic PAS, I analysed split reads aligning to exon junctions and local read coverages around the 5'SS and the PAS. Both `TECtool` and `IPAFinder` programs address the same question. However, these tools adopt a complex approach and primarily identify the exact location of the alternative terminal exon. In this study, I showed that a simpler method, which does not determine the ATE start, accurately predicted the type for 91% of annotated used CTE iPAS and 88% of STEs. Since AS events were categorized tissue-specifically, I was able to observe switching between CTE and STE iPAS types (Figure 5-40A). While such a tissue-specific switch in iPAS type requires further validation and investigation, it highlights the dynamic nature of IPA that has been observed repeatedly in the past studies [193, 104, 146].

The abundance of IPA has been appreciated recently with the development of 3'-end sequencing [140]. Functionally important IPA cases have been described in specific genes [137, 36, 31, 7, 167, 88], however, most transcripts harbouring incomplete reading frames translate into potentially deleterious, truncated proteins that may pose a hazard to the cell [165] (2.1.3). I found that the majority of polyA reads align to 3'-UTRs, but a sizable fraction (about 5%) still maps to the coding part.

Within the coding part, the polyA read density in introns is lower than in exons. However, since the intronic regions are removed from the transcripts and degraded, the total RNA-seq read coverage in introns is also substantially lower. To compare

the frequency of CPA events in intronic and exonic regions, we normalized the polyA read density by the total read coverage. Intriguingly, we observed that the normalized polyA read density is substantially higher in the intronic regions. Thus, CPA within the coding part appears to be more frequent in introns than in exons. This disbalance can be partly explained by the higher GC content and stronger evolutionary constraints against generating the canonical AATAAA consensus sequence in exons.

The exonic polyadenylation signal hexamers are under stronger selection than intronic ones [73]. Additionally, sequence analysis showed that each spliced protein-coding transcript contains an intronic polyadenylation signal hexamer, which can be specifically recognized and bound by CPSF, one of the key CPA factors (2.1.1). This remarkably large number of intronic PASs raises concerns about their implication in premature transcription termination [74]. All these observations hint at the existence of a mechanism that counteracts activity of the intronic PAS. How could it be that 82% of human protein-coding transcripts contain an intronic PAS, but cells are still able to produce full-length transcripts?

Here, I 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. If CPA occurs first, then it will lead to the generation of a truncated transcript with CTE (Figure 6-1, left). If splicing happens first, then the intron-containing PAS will be spliced out, and PAS will be degraded as a part of the lariat (Figure 6-1, right). 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 a Spliced Polyadenylated Intron (Figure 6-1, middle).

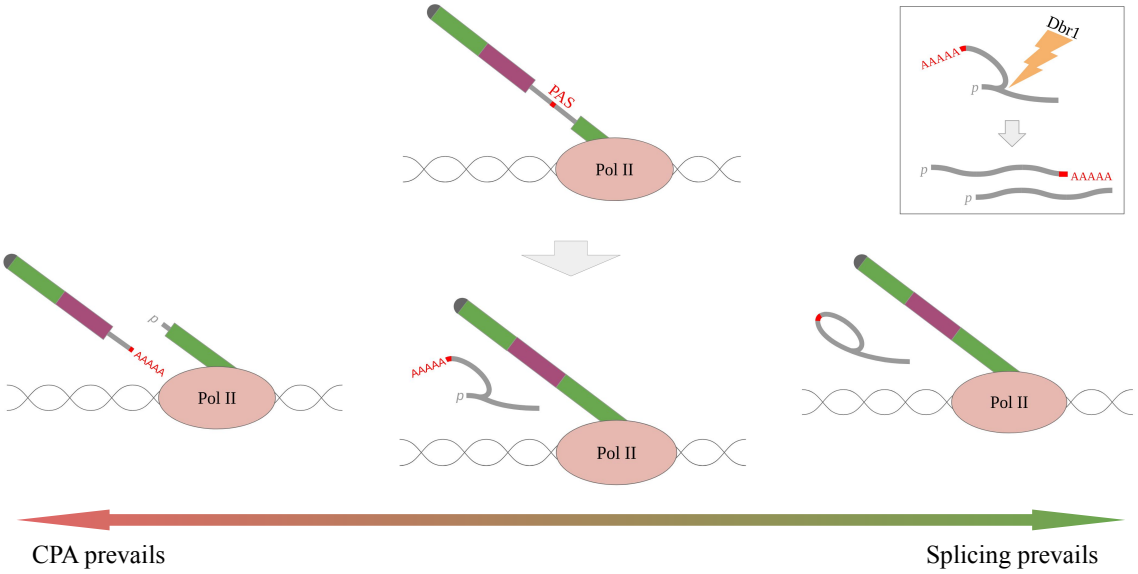


Figure 6-1: **Spliced Polyadenylated Intron (SPI).** When the CPA rate exceeds the splicing rate, IPA leads to the generation of a truncated transcript isoform (left). When the splicing rate exceeds the CPA rate, the intron is spliced out and PAS is degraded as a part of a lariat (right). When the CPA 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.

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 (Figure 6-1, inset). 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 5-39). Nonetheless, a fraction of SPIs are visible through polyA reads due to the presence of the polyA tail. A conservative estimate is that they constitute almost 15% 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.

In short, SPI is a result of intronic CPA not accompanied by an AS event. In this work, an intronic PAS was classified as an SPI in the tissue if RNA-seq read coverage was substantial upstream of the site and dropped significantly right after it, but no reads supported an AS event that could explain the observed intronic

coverage in the upstream region. One possible concern is that SPIs are a result of STEs misclassification in genes missing the spliced reads due to low expression level. It is additionally supported by the similarity of we_2/we_1 and we_2/wi_1 distributions between STEs and SPIs. To avoid such misclassification due to low coverage, I introduced a threshold of 30 reads on the ψ denominator, which is a proxy for the local expression around the intron. In the resulting SPI set $\geq 75\%$ had a ψ denominator bigger than 200. Furthermore, the local expression value distributions for CTEs and SPIs are very similar, confirming that SPIs are not STE or CTE events in weakly covered genes (Figure A-7).

After CPA concludes, the polyadenylated pre-mRNA dissociates from the transcribing Pol II. Since successful intron splicing would require both the polyadenylated upstream pre-mRNA and the Pol II-associated downstream RNA, this dissociation could hinder the process. U2 snRNP bound to BP and U1 snRNP at 5'SS interact as early as A complex stage in the spliceosome formation (2.2). After that, the upstream pre-mRNA and the Pol-II-associated RNA would not dissociate even upon the CPA completion. Moreover, this interaction loops out the majority of the intron from 5'SS to the BP, allowing the CPA reaction to occur without steric hindrances. This model agrees with the observation that SPIs are located in longer introns (Figure A-6). U1 telescripting is also known to repress intronic PAS in long introns, presumably because they stochastically contain more polyadenylation signals [123] (2.3). However, this mechanism relies on U1 snRNP binding sites upstream of each cryptic PAS. Thus, spliceosome-mediated excision of an intron, in which CPA has already occurred, could be a less robust but more universal mechanism to suppress premature transcription termination.

While SPIs can be a valuable indicator of the splicing and CPA interplay, this study does not suggest that they, as a class of short RNAs, have any other specific functions. On the other hand, some SPIs could have obtained an independent role in the course of evolution similar to known functional circular RNAs [179, 54]. To sum up, whether or not SPIs have a biological purpose on their own is a matter of further investigation.

One of the most remarkable observations made here is the enrichment of polyA

read density in introns, which can be detected after proper normalization. This difference suggests that cotranscriptional pre-mRNA splicing may have an important side role in rescuing transcripts from premature transcription termination. This hypothesis challenges the assumption that when an intronic PAS is used, the surrounding intron is not spliced. The spliceosome that is committed to splicing still can remove the intron that was cleaved and polyadenylated, thus functioning as a 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 pre-mRNA substrate at the same time [174, 84, 113] (2.3). 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.

Chapter 7

Conclusion

This dissertation offers a comprehensive examination of the interplay between pre-mRNA splicing and intronic polyadenylation across human tissues, utilizing the extensive compendium of RNA-seq datasets provided by the GTEx project.
It can be concluded that

- The increased scale of the dataset significantly increases the sensitivity of the polyA-read-based method of PAS detection, thus achieving performance comparable to that of meta-analyses of hundreds of 3'-seq libraries;
- The accuracy of the cleavage and polyadenylation depends on the relative position of the polyadenylation signal with respect to PAS;
- After proper normalization, the relative frequency of polyA reads in introns is larger than that in nonterminal exons, reflecting intrinsically higher cleavage and polyadenylation rate in introns;
- A substantial fraction of intronic polyadenylation events are unaccompanied by alternative splicing and can be attributed to byproducts of the dynamic interaction between CPA and splicing mechanisms, here called spliced polyadenylated introns (SPI);
- Splicing and polyadenylation are two inseparable parts of one consolidated pre-mRNA processing machinery, and co-transcriptional splicing may be a natural mechanism of suppression of premature transcription termination.

Bibliography

[1] Shanika L. Amarasinghe, Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. Opportunities and challenges in long-read sequencing data analysis, feb 2020.

[2] Ruhul Amin, Anjali Shukla, Jacqueline Jufen Zhu, Sohyoung Kim, Ping Wang, Simon Zhongyuan Tian, Andy D. Tran, Debasish Paul, Steven D. Cappell, Sandra Burkett, Huaitian Liu, Maxwell P. Lee, Michael J. Kruhlak, Jennifer E. Dwyer, R. Mark Simpson, Gordon L. Hager, Yijun Ruan, and Kent W. Hunter. Nuclear pore protein NUP210 depletion suppresses metastasis through heterochromatin-mediated disruption of tumor cell mechanical response. *Nature Communications*, 12(1), dec 2021.

[3] Ashraful Arefeen, Juntao Liu, Xinshu Xiao, and Tao Jiang. TAPAS: Tool for alternative polyadenylation site analysis. *Bioinformatics*, 34(15):2521–2529, aug 2018.

[4] Abdallah Azouz and Jonathan S. Duke-Cohan. Post-developmental extracellular proteoglycan maintenance in attractin-deficient mice. *BMC Research Notes*, 13(1), jun 2020.

[5] Yoseph Barash, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe, and Brendan J. Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, may 2010.

[6] Jean Denis Beaudoin and Jean Pierre Perreault. Exploring mRNA 3'-UTR G-quadruplexes: Evidence of roles in both alternative polyadenylation and mRNA shortening. *Nucleic Acids Research*, 41(11):5898–5911, jun 2013.

[7] P. Benech, Y. Mory, M. Revel, and J. Chebath. Structure of two forms of the interferon-induced (2'-5') oligo A synthetase of human cells based on cDNAs and gene sequences. *The EMBO journal*, 4(9):2249–2256, 1985.

[8] Susan M. Berget. Exon recognition in vertebrate splicing. *Journal of Biological Chemistry*, 270(6):2411–2414, feb 1995.

[9] Inanc Birol, Anthony Raymond, Readman Chiu, Ka Ming Nip, Shaun D Jackman, Maayan Kreitzman, T Roderick Docking, Catherine A Ennis, A Gordon Robertson, and Aly Karsan. KLEAT: cleavage site analysis of transcriptomes. In *Biocomputing 2015*, pages 347–358. World Scientific, nov 2014.

[10] Thomas Bonfert and Caroline C. Friedel. Prediction of Poly(A) Sites by Poly(A) Read Mapping. *PLOS ONE*, 12(1):e0170914, jan 2017.

[11] Paul L. Boutz, Arjun Bhutkar, and Phillip A. Sharp. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes and Development*, 29(1):63–80, jan 2015.

[12] Robert K. Bradley and Olga Anczuków. RNA splicing dysregulation and the hallmarks of cancer. *Nature reviews. Cancer*, pages 1–21, jan 2023.

[13] Justin Brumbaugh, Bruno Di Stefano, Xiuye Wang, Marti Borkent, Elmira Forouzmand, Katie J. Clowers, Fei Ji, Benjamin A. Schwarz, Marian Kalocsay, Stephen J. Elledge, Yue Chen, Ruslan I. Sadreyev, Steven P. Gygi, Guang Hu, Yongsheng Shi, and Konrad Hochedlinger. Nudt21 Controls Cell Fate by Connecting Alternative Polyadenylation to Chromatin Signaling. *Cell*, 172(1-2):106–120.e21, jan 2018.

[14] Sam Bryce-Smith, Dominik Burri, Matthew R. Gazzara, Christina J. Herrmann, Weronika Danecka, Christina M. Fitzsimmons, Yuk Kei Wan, Farica Zhuang, Mervin M. Fansler, José M. Fernández, Meritxell Ferret, Asier Gonzalez-Uriarte, Samuel Haynes, Chelsea Herdman, Alexander Kanitz, Maria Katsantoni, Federico Marini, Euan McDonnel, Ben Nicolet, Chi-Lam Poon, Gregor Rot, Leonard Schärfen, Pin-Jou Wu, Yoseop Yoon, Yoseph Barash, and Mihaela Zavolan. Extensible benchmarking of methods that identify and quantify polyadenylation sites from RNA-seq data. *bioRxiv*, page 2023.06.23.546284, jun 2023.

[15] Angelo Calado, Ulrike Kutay, Uwe Kühn, Elmar Wahle, and Maria Carmo-Fonseca. Deciphering the cellular pathway for transport of poly(A)-binding protein II. *RNA*, 6(2):245–256, feb 2000.

[16] Ashley A. Cass and Xinshu Xiao. mountainClimber Identifies Alternative Transcription Start and Polyadenylation Sites in RNA-Seq. *Cell Systems*, 9(4):393–400.e6, 2019.

[17] Muhammed Hasan Çelik and Ali Mortazavi. Analysis of alternative polyadenylation from long-read or short-read RNA-seq with LAPA. pages 1–17, 2022.

[18] Moliang Chen, Guoli Ji, Hongjuan Fu, Qianmin Lin, Congting Ye, Wenbin Ye, Yaru Su, and Xiaohui Wu. A survey on identification and quantification of alternative polyadenylation sites from RNA-seq data, 2019.

[19] Wei Chen, Qi Jia, Yifan Song, Haihui Fu, Gang Wei, and Ting Ni. Alternative Polyadenylation: Methods, Findings, and Impacts. *Genomics, Proteomics and Bioinformatics*, 15(5):287–300, 2017.

[20] Zhi Cheng and Thomas M. Menees. RNA splicing and debranching viewed through analysis of RNA lariats. *Molecular Genetics and Genomics*, 286(5-6):395–410, 2011.

[21] Deanna M. Church, Valerie A. Schneider, Tina Graves, Katherine Auger, Fiona Cunningham, Nathan Bouk, Hsiu Chuan Chen, Richa Agarwala, William M. McLaren, Graham R.S. Ritchie, Derek Albracht, Milinn Kremitzki, Susan Rock, Holland Kotkiewicz, Colin Kremitzki, Aye Wollam, Lee Trani, Lucinda Fulton, Robert Fulton, Lucy Matthews, Siobhan Whitehead, Will Chow, James Torrance, Matthew Dunn, Glenn Harden, Glen Threadgold, Jonathan Wood, Joanna Collins, Paul Heath, Guy Griffiths, Sarah Pelan, Darren Grafham, Evan E. Eichler, George Weinstock, Elaine R. Mardis, Richard K. Wilson, Kerstin Howe, Paul Flicek, and Tim Hubbard. Modernizing reference genome assemblies. *PLoS Biology*, 9(7), jul 2011.

[22] Marie claire Daugeron, Jamal Tazi, Philippe Jeanteur, Claude Brunel, and Guy Cathola. U1-U2 snRNPs interaction induced by an RNA complementary to the 5' end sequence of U1 snRNA. *Nucleic Acids Research*, 20(14):3625–3630, jul 1992.

[23] Ana Curinha, Sandra Oliveira Braz, Isabel Pereira-Castro, Andrea Cruz, and Alexandra Moreira. Implications of polyadenylation in health and disease. *Nucleus*, 5(6):508–519, oct 2014.

[24] Sven. Danckwardt, K. Hartmann, B. Katz, M. W. Hentze, Y. Levy, R. Eichele, V. Deutsch, A. E. Kulozik, and Oded. Ben-Tal. The prothrombin 20209 C-T mutation in jewish-moroccan caucasians: Molecular analysis of gain-of-function of 3' end processing. *Journal of Tissue Engineering and Regenerative Medicine*, 4(5):1078–1085, may 2006.

[25] Petr Danecek, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, Thomas Keane, Shane A. McCarthy, and Robert M. Davies. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), feb 2021.

[26] Lee Davidson and Steven West. Splicing-coupled 3' end formation requires a terminal splice acceptor site, but not intron excision. *Nucleic Acids Research*, 41(14):7101–7114, aug 2013.

[27] Melissa J. Davis, Kelly A. Hanson, Francis Clark, J. Lynn Fink, Fasheng Zhang, Takeya Kasukawa, Chikatoshi Kai, Jun Kawai, Piero Carninci, Yoshihide Hayashizaki, and Rohan D. Teasdale. Differential use of signal peptides and membrane domains is a common occurrence in the protein output of transcriptional units. *PLoS Genetics*, 2(4):554–563, apr 2006.

[28] Nicola de Prisco, Caitlin Ford, Nathan D Elrod, Winston Lee, Lauren C Tang, Kai-Lieh Lieh Huang, Ai Lin, Ping Ji, Venkata S Jonnakuti, Lia Boyle, Maximilian Cabaj, Salvatore Botta, Katrin Õunap, Karit Reinson, Monica H Wojcik, Jill A Rosenfeld, Weimin Bi, Kristian Tveten, Trine Prescott, Thorsten Gerstner, Audrey Schroeder, Chin-To To Fong, Jaya K George-Abraham, Catherine A Buchanan, Andrea Hanson-Khan, Jonathan A Bernstein, Aikaterini A Nella, Wendy K Chung, Vicky Brandt, Marko Jovanovic, Kimara L Targoff,

Hari Krishna Yalamanchili, Eric J Wagner, and Vincenzo A Gennarino. Alternative polyadenylation alters protein dosage by switching between intronic and 3'UTR sites. *Science advances*, 9(7):eade4814, feb 2023.

[29] Adnan Derti, Philip Garrett-Engele, Kenzie D Macisaac, Richard C Stevens, Shreedharan Sriram, Ronghua Chen, Carol A Rohl, Jason M Johnson, and Tomas Babak. A quantitative atlas of polyadenylation in five mammals. *Genome research*, 22(6):1173–83, jun 2012.

[30] Joana Desterro, Pedro Bak-Gordon, and Maria Carmo-Fonseca. Targeting mRNA processing as an anticancer strategy. *Nature Reviews Drug Discovery*, 19(2):112–129, feb 2020.

[31] Dafne Campigli Di Giammartino, Kensei Nishida, and James L. Manley. Mechanisms and Consequences of Alternative Polyadenylation. *Molecular Cell*, 43(6):853–866, sep 2011.

[32] Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Nadav S. Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cedric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J. Luo, Eddie Park, Kimberly Persaud, Jonathan B. Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guig, and Thomas R. Gingeras. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, sep 2012.

[33] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, jan 2013.

[34] Heather L. Drexler, Karine Choquet, and L. Stirling Churchman. Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Molecular Cell*, 77(5):985–998.e8, mar 2020.

[35] Sara J. Dubbury, Paul L. Boutz, and Phillip A. Sharp. CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. *Nature*, 564(7734):141–145, dec 2018.

[36] Gretchen Edwalds-Gilbert, Kristen L. Veraldi, and Christine Milcarek. Alternative poly(A) site selection in complex transcription units: Means to an end? *Nucleic Acids Research*, 25(13):2547–2561, jul 1997.

[37] Ran Elkon, Jarno Drost, Gijs van Haaften, Mathias Jenal, Mariette Schrier, Joachim A.O. Vrielink, and Reuven Agami. E2F mediates enhanced alternative polyadenylation in proliferation. *Genome Biology*, 13(7):1–15, jul 2012.

[38] Zhixiao Fang and Shengli Li. Alternative polyadenylation-associated loci interpret human traits and diseases. *Trends in Genetics*, 37(9):773–775, sep 2021.

[39] Katalin Fejes-Toth, Vihra Sotirova, Ravi Sachidanandam, Gordon Assaf, Gregory J. Hannon, Philipp Kapranov, Sylvain Foissac, Aarron T. Willingham, Radha Duttagupta, Erica Dumais, and Thomas R. Gingeras. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature*, 457(7232):1028–1032, feb 2009.

[40] Xin Feng, Lei Li, Eric J Wagner, and Wei Li. TC3A: The Cancer 3' UTR Atlas. *Nucleic Acids Research*, 46(D1):D1027–D1030, jan 2018.

[41] Ana Fiszbein, Michael McGurk, Ezequiel Calvo-Roitberg, Gyeung Yun Kim, Christopher B. Burge, and Athma A. Pai. Widespread occurrence of hybrid internal-terminal exons in human transcriptomes. *Science Advances*, 8(3):1752, jan 2022.

[42] Rachel Flomen and Andrew Makoff. Increased RNA editing in EAAT2 pre-mRNA from amyotrophic lateral sclerosis patients: Involvement of a cryptic polyadenylation site. *Neuroscience Letters*, 497(2):139–143, jun 2011.

[43] Eric R. Gamazon and Barbara E. Stranger. Genomics of alternative splicing: Evolution, development and pathophysiology. *Human Genetics*, 133(6):679–687, jan 2014.

[44] Niels H. Gehring and Jean Yves Roignant. Anything but Ordinary – Emerging Splicing Mechanisms in Eukaryotic Gene Regulation, apr 2021.

[45] Joseph V. Geisberg, Zarmik Moqtaderi, and Kevin Struhl. The transcriptional elongation rate regulates alternative polyadenylation in yeast. *eLife*, 9:1–55, aug 2020.

[46] J. Sebastian Gomez-Cavazos and Martin W. Hetzer. The nucleoporin gp210/Nup210 controls muscle differentiation by regulating nuclear envelope/ER homeostasis. *Journal of Cell Biology*, 208(6):671–681, 2015.

[47] Elena Grassi, Elisa Mariella, Antonio Lembo, Ivan Molineris, and Paolo Provero. Roar: detecting alternative polyadenylation with standard mRNA sequencing libraries. *BMC Bioinformatics*, 17(1):423, dec 2016.

[48] Lea H. Gregersen, Richard Mitter, Alejandro P. Ugalde, Takayuki Nojima, Nick J. Proudfoot, Reuven Agami, Aengus Stewart, and Jesper Q. Svejstrup. SCAF4 and SCAF8, mRNA Anti-Terminator Proteins. *Cell*, 177(7):1797–1813.e18, jun 2019.

[49] Natalia Gromak, Steven West, and Nick J. Proudfoot. Pause Sites Promote Transcriptional Termination of Mammalian RNA Polymerase II. *Molecular and Cellular Biology*, 26(10):3986–3996, may 2006.

[50] Andreas J. Gruber, Foivos Gypas, Andrea Riba, Ralf Schmidt, and Mihaela Zavolan. Terminal exon characterization with TECtool reveals an abundance of cell-specific isoforms. *Nature Methods*, 15(10):832–836, sep 2018.

[51] Andreas J. Gruber, Ralf Schmidt, Souvik Ghosh, Georges Martin, Andreas R. Gruber, Erik van Nimwegen, and Mihaela Zavolan. Discovery of physiological and cancer-related regulators of 3' UTR processing with KAPAC. *Genome Biology*, 19(1):44, dec 2018.

[52] Andreas J. Gruber and Mihaela Zavolan. Alternative cleavage and polyadenylation in health and disease. *Nature Reviews Genetics*, pages 1–16, jul 2019.

[53] Andreas R. J. Gruber, Ralf Schmidt, Andreas R. J. Gruber, Georges Martin, Souvik Ghosh, Manuel Belmadani, Walter Keller, and Mihaela Zavolan. A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Research*, 26(8):1145–1159, aug 2016.

[54] Franziska Gruhl, Peggy Janich, Henrik Kaessmann, and David Gatfield. Circular RNA repertoires are associated with evolutionarily young transposable elements. *eLife*, 10, 2021.

[55] Kevin C. H. Ha, Benjamin J. Blencowe, and Quaid Morris. QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biology*, 19(1):45, dec 2018.

[56] Matthew J. Hangauer, Ian W. Vaughn, and Michael T. McManus. Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genetics*, 9(6), jun 2013.

[57] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel

Rodriguez, Iakes Ezkurdia, Jeltje Van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, 22(9):1760–1774, sep 2012.

[58] Christina J. Herrmann, Ralf Schmidt, Alexander Kanitz, Panu Artimo, Andreas J. Gruber, and Mihaela Zavolan. PolyASite 2.0: A consolidated atlas of polyadenylation sites from 3' end sequencing. *Nucleic Acids Research*, 48(D1):D174–D179, 2020.

[59] Lydia Herzel, Diana S.M. Ottoz, Tara Alpert, and Karla M. Neugebauer. Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function. *Nature Reviews Molecular Cell Biology*, 18(10):637–650, 2017.

[60] Chris H. Hill, Vytautė Boreikaitė, Ananthanarayanan Kumar, Ana Casañal, Peter Kubík, Gianluca Degliesposti, Sarah Maslen, Angelica Mariani, Ottilie von Loeffelholz, Mathias Girbig, Mark Skehel, and Lori A. Passmore. Activation of the Endonuclease that Defines mRNA 3' Ends Requires Incorporation into an 8-Subunit Core Cleavage and Polyadenylation Factor Complex. *Molecular Cell*, 73(6):1217–1231.e11, mar 2019.

[61] Chung Chau Hon, Christian Weber, Odile Sismeiro, Caroline Proux, Mikael Koutero, Marc Deloger, Sarbashis Das, Mridula Agrahari, Marie Agnes Dillies, Bernd Jagla, Jean Yves Coppee, Alok Bhattacharya, and Nancy Guillen. Quantification of stochastic noise of splicing and polyadenylation in Entamoeba histolytica. *Nucleic Acids Research*, 41(3):1936–1952, feb 2013.

[62] Wei Hong, Hang Ruan, Zhao Zhang, Youqiong Ye, Yaoming Liu, Shengli Li, Ying Jing, Huiwen Zhang, Lixia Diao, Han Liang, and Leng Han. APAatlas: Decoding alternative polyadenylation across human tissues. *Nucleic Acids Research*, 48(D1):D34–D39, jan 2020.

[63] Mainul Hoque, Zhe Ji, Dinghai Zheng, Wenting Luo, Wencheng Li, Bei You, Ji Yeon Park, Ghassan Yehia, and Bin Tian. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nature Methods*, 10(2):133–139, 2013.

[64] Jing-Ping Hsin and James L. Manley. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes & Development*, 26(19):2119–2137, oct 2012.

[65] Li Huang, Guangnan Li, Chen Du, Yu Jia, Jiayi Yang, Weiliang Fan, Yong-Zhen Xu, Hong Cheng, and Yu Zhou. The polyA tail facilitates splicing of last introns with weak 3' splice sites via PABPN1. *EMBO reports*, page e57128, sep 2023.

[66] Hun Way Hwang, Christopher Y. Park, Hani Goodarzi, John J. Fak, Aldo Mele, Michael J. Moore, Yuhki Saito, and Robert B. Darnell. PAPERCLIP

Identifies MicroRNA Targets and a Role of CstF64/64tau in Promoting Non-canonical poly(A) Site Usage. *Cell Reports*, 15(2):423–435, apr 2016.

[67] Mathias Jenal, Ran Elkon, Fabricio Loayza-Puch, Gijs Van Haaften, Uwe Kühn, Fiona M. Menzies, Joachim A.F.Oude Vrielink, Arnold J. Bos, Jarno Drost, Koos Rooijers, David C. Rubinsztein, and Reuven Agami. The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell*, 149(3):538–553, apr 2012.

[68] Torben Heick Jensen, Alain Jacquier, and Domenico Libri. Dealing with Pervasive Transcription. *Molecular Cell*, 52(4):473–484, nov 2013.

[69] Jennifer J. Johnston, Kathleen A. Williamson, Christopher M. Chou, Julie C. Sapp, Morad Ansari, Heather M. Chapman, David N. Cooper, Tabib Dabir, Jeffrey N. Dudley, Richard J. Holt, Nicola K. Ragge, Alejandro A. Schäffer, Shurjo K. Sen, Anne M. Slavotinek, David R. Fitzpatrick, Thomas M. Glaser, Fiona Stewart, Graeme C.M. Black, and Leslie G. Biesecker. NAA10 polyadenylation signal variants cause syndromic microphthalmia. *Journal of Medical Genetics*, 56(7):444–452, jul 2019.

[70] Fui Boon Kai, James P. Fawcett, and Roy Duncan. Synaptopodin-2 induces assembly of peripheral actin bundles and immature focal adhesions to promote lamellipodia formation and prostate cancer cell migration. *Oncotarget*, 6(13):11162–11174, may 2015.

[71] Daisuke Kaida. The reciprocal regulation between splicing and 3'-end processing. *WIREs RNA*, 7(4):499–511, jul 2016.

[72] Daisuke Kaida, Michael G. Berg, Ihab Younis, Mumtaz Kasim, Larry N. Singh, Lili Wan, and Gideon Dreyfuss. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, 468(7324):664–668, dec 2010.

[73] Yaroslav A. Kainov, Vasily N. Aushev, Sergey A. Naumenko, Elena M. Tchevkina, and Georgii A. Bazykin. Complex selection on human polyadenylation signals revealed by polymorphism and divergence data. *Genome Biology and Evolution*, 8(6):1971–1979, jun 2016.

[74] Kinga Kamieniarz-Gdula and Nick J. Proudfoot. Transcriptional Control by Premature Termination: A Forgotten Mechanism. *Trends in Genetics*, 35(8):553–564, aug 2019.

[75] Ender Karaca, Stefan Weitzer, Davut Pehlivan, Hiroshi Shiraishi, Tasos Gogakos, Toshikatsu Hanada, Shalini N. Jhangiani, Wojciech Wiszniewski, Marjorie Withers, Ian M. Campbell, Serkan Erdin, Sedat Isikay, Luis M. Franco, Claudia Gonzaga-Jauregui, Tomasz Gambin, Violet Gelowani, Jill V. Hunter, Gozde Yesil, Erkan Koparir, Sarenur Yilmaz, Miguel Brown, Daniel Briskin, Markus Hafner, Pavel Morozov, Thalia A. Farazi, Christian Bernreuther, Markus Glatzel, Siegfried Trattnig, Joachim Friske, Claudia Kronnerwetter, Matthew N. Bainbridge, Alper Gezdirici, Mehmet Seven, Donna M.

Muzny, Eric Boerwinkle, Mustafa Ozen, Tim Clausen, Thomas Tuschl, Adnan Yuksel, Andreas Hess, Richard A. Gibbs, Javier Martinez, Josef M. Penninger, and James R. Lupski. Human CLP1 mutations alter tRNA biogenesis, Affecting both peripheral and central nervous system function. *Cell*, 157(3):636–650, apr 2014.

[76] Yarden Katz, Eric T. Wang, Edoardo M. Airoldi, and Christopher B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7(12):1009–1015, dec 2010.

[77] Isabelle Kaufmann, Georges Martin, Arno Friedlein, Hanno Langen, and Walter Keller. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO Journal*, 23(3):616–626, feb 2004.

[78] Daehwan Kim, Ben Langmead, and Steven L Salzberg. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, apr 2015.

[79] MinHyeok Kim, Bo-Hyun You, and Jin-Wu Nam. Global estimation of the 3' untranslated region landscape using RNA sequencing. *Methods*, 83:111–117, jul 2015.

[80] Malgorzata Krajewska, Ruben Dries, Andrew V. Grassetti, Sofia Dust, Yang Gao, Hao Huang, Bandana Sharma, Daniel S. Day, Nicholas Kwiatkowski, Monica Pomaville, Oliver Dodd, Edmond Chipumuro, Tinghu Zhang, Arno L. Greenleaf, Guo-Cheng Yuan, Nathanael S. Gray, Richard A. Young, Matthias Geyer, Scott A. Gerber, and Rani E. George. CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation. *Nature Communications*, 10(1):1757, dec 2019.

[81] Shankarling Krishnamurthy, Xiaoyuan He, Mariela Reyes-Reyes, Claire Moore, and Michael Hampsey. Ssu72 is an RNA polymerase II CTD phosphatase. *Molecular Cell*, 14(3):387–394, may 2004.

[82] Jason N. Kuehner, Erika L. Pearson, and Claire Moore. Unravelling the means to an end: RNA polymerase II transcription termination. *Nature Reviews Molecular Cell Biology*, 12(5):283–294, may 2011.

[83] Buki Kwon, Mervin M. Fansler, Neil D. Patel, Jihye Lee, Weirui Ma, and Christine Mayr. Enhancers regulate 3' end processing activity to control expression of alternative 3'UTR isoforms. *Nature Communications*, 13(1):1–14, may 2022.

[84] Andrea Kyburz, Arno Friedlein, Hanno Langen, and Walter Keller. Direct Interactions between Subunits of CPSF and the U2 snRNP Contribute to the Coupling of Pre-mRNA 3' End Processing and Splicing. *Molecular Cell*, 23(2):195–205, jul 2006.

[85] Brad Lackford, Chengguo Yao, Georgette M. Charles, Lingjie Weng, Xiaofeng Zheng, Eun A. Choi, Xiaohui Xie, Ji Wan, Yi Xing, Johannes M. Freudenberg, Pengyi Yang, Raja Jothi, Guang Hu, and Yongsheng Shi. Fip1 regulates mRNA alternative polyadenylation to promote stem cell self-renewal. *EMBO Journal*, 33(8):878–889, apr 2014.

[86] Michael S. Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, 9(8):e1003118, aug 2013.

[87] Ju Youn Lee, Ijen Yeh, Ji Yeon Park, and Bin Tian. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Research*, 35(SUPPL. 1):D165, jan 2007.

[88] Shih Han Lee, Irtisha Singh, Sarah Tisdale, Omar Abdel-Wahab, Christina S. Leslie, and Christine Mayr. Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature*, 561(7721):127–131, 2018.

[89] Lei Li, Kai Lieh Huang, Yipeng Gao, Ya Cui, Gao Wang, Nathan D. Elrod, Yumei Li, Yiling Elaine Chen, Ping Ji, Fanglue Peng, William K. Russell, Eric J. Wagner, and Wei Li. An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nature Genetics*, pages 1–12, may 2021.

[90] Qingshun Q. Li, Zhaoyang Liu, Wenjia Lu, and Man Liu. Interplay between Alternative Splicing and Alternative Polyadenylation Defines the Expression Outcome of the Plant Unique OXIDATIVE TOLERANT-6 Gene. *Scientific Reports*, 7(1):1–9, may 2017.

[91] Wencheng Li, Bei You, Mainul Hoque, Dinghai Zheng, Wenting Luo, Zhe Ji, Ji Yeon Park, Samuel I. Gunderson, Auinash Kalsotra, James L. Manley, and Bin Tian. Systematic Profiling of Poly(A)+ Transcripts Modulated by Core 3' End Processing and Splicing Factors Reveals Regulatory Rules of Alternative Cleavage and Polyadenylation. *PLoS Genetics*, 11(4):e1005166, apr 2015.

[92] Steve Lianoglou, Vidur Garg, JL Yang, Christina S. Leslie, and Christine Mayr. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes*, pages 2380–2396, 2013.

[93] Yang Liao, Gordon K. Smyth, and Wei Shi. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, apr 2014.

[94] Xiaochuan Liu, Jaime Freitas, Dinghai Zheng, Marta S. Oliveira, Mainul Hoque, Torcato Martins, Telmo Henriques, Bin Tian, and Alexandra Moreira. Transcription elongation rate has a tissue-specific impact on alternative cleavage and polyadenylation in Drosophila melanogaster. *Rna*, 23(12):1807–1816, 2017.

[95] Yusheng Liu, Hu Nie, Hongxiang Liu, and Falong Lu. Poly(A) inclusive RNA isoform sequencing (PAIsoseq) reveals wide-spread non-adenosine residues within RNA poly(A) tails. *Nature Communications*, 10(1):5292, dec 2019.

[96] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1), nov 2011.

[97] Hua Lou, Robert F. Gagel, and Susan M. Berget. An intron enhancer recognized by splicing factors activates polyadenylation. *Genes and Development*, 10(2):208–219, jan 1996.

[98] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, dec 2014.

[99] Wenting Luo, Zhe Ji, Zhenhua Pan, Bei You, Mainul Hoque, Wencheng Li, Samuel I. Gunderson, and Bin Tian. The Conserved Intronic Cleavage and Polyadenylation Site of CstF-77 Gene Imparts Control of 3' End Processing Activity through Feedback Autoregulation and by U1 snRNP. *PLoS Genetics*, 9(7):1–14, 2013.

[100] Ryan Lusk, Evan Stene, Farnoush Banaei-Kashani, Boris Tabakoff, Katerina Kechris, and Laura M Saba. Aptardi predicts polyadenylation sites in sample-specific transcriptomes using high-throughput RNA sequencing and DNA sequence. *Nature Communications*, 12(1), 2021.

[101] Yuval Malka, Ferhat Alkan, Shinyeong Ju, Pierre-Rene Körner, Abhijeet Pataskar, Eldad Shulman, Fabricio Loayza-Puch, Julien Champagne, Casper Wenzel, William James Faller, Ran Elkon, Cheolju Lee, and Reuven Agami. Alternative cleavage and polyadenylation generates downstream uncapped RNA isoforms with translation potential. *Molecular Cell*, 82(20):3840–3855.e8, oct 2022.

[102] Yuval Malka, Avital Steiman-Shimony, Eran Rosenthal, Liron Argaman, Leonor Cohen-Daniel, Eliran Arbib, Hanah Margalit, Tommy Kaplan, and Michael Berger. Post-transcriptional 3´-UTR cleavage of mRNA transcripts generates thousands of stable uncapped autonomous RNA fragments. *Nature Communications*, 8(1):1–11, dec 2017.

[103] Sergey D. Margasyuk, Maria A. Vlasenok, Gaofeng Li, Ch Cao, and Dmitri D. Pervouchine. RNAcontacts: A Pipeline for Predicting Contacts from RNA Proximity Ligation Assays. *Acta Naturae*, 15(1-56):51–57, 2023.

[104] Federico Marini, Denise Scherzinger, and Sven Danckwardt. TREND-DB - A transcriptome-wide atlas of the dynamic landscape of alternative polyadenylation. *Nucleic Acids Research*, 49(D1):D243–D253, jan 2021.

[105] Camilla Ciolli Mattioli, Aviv Rom, Vedran Franke, Koshi Imami, Gerard Arrey, Mandy Terne, Andrew Woehler, Altuna Akalin, Igor Ulitsky, and Marina Chekulaeva. Alternative 3 UTRs direct localization of functionally diverse protein isoforms in neuronal compartments. *Nucleic Acids Research*, 47(5):2560–2573, mar 2019.

[106] Christine Mayr. What are 3' utrs doing? *Cold Spring Harbor Perspectives in Biology*, 11(10):a034728, oct 2019.

[107] Christine Mayr and David P. Bartel. Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell*, 138(4):673–684, aug 2009.

[108] Cristina T.N. Medina, Renata Sandoval, Gabriela Oliveira, Karina da Costa Silveira, Denise P. Cavalcanti, and Robert Pogue. Pathogenic variants in the TRIP11 gene cause a skeletal dysplasia spectrum from odontochondrodysplasia to achondrogenesis 1A. *American Journal of Medical Genetics, Part A*, 182(4):681–688, apr 2020.

[109] Marta Melé, Pedro G. Ferreira, Ferran Reverter, David S. DeLuca, Jean Monlong, Michael Sammeth, Taylor R. Young, Jakob M. Goldmann, Dmitri D. Pervouchine, Timothy J. Sullivan, Rory Johnson, Ayellet V. Segrè, Sarah Djebali, Anastasia Niarchou, Fred A. Wright, Tuuli Lappalainen, Miquel Calvo, Gad Getz, Emmanouil T. Dermitzakis, Kristin G. Ardlie, and Roderic Guigó. The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665, may 2015.

[110] Tim R. Mercer, Michael B. Clark, Stacey B. Andersen, Marion E. Brunck, Wilfried Haerty, Joanna Crawford, Ryan J. Taft, Lars K. Nielsen, Marcel E. Dinger, and John S. Mattick. Genome-wide discovery of human splicing branchpoints. *Genome Research*, 25(2):290–303, feb 2015.

[111] Stefania Millevoi, Clarisse Loulergue, Sabine Dettwiler, Sarah Zeïneb Karaa, Walter Keller, Michael Antoniou, and Stéphan Vagner. An interaction between U2AF 65 and CF Im links the splicing and 3' end processing machineries. *EMBO Journal*, 25(20):4854–4864, oct 2006.

[112] Alexei Mironov, Marina Petrova, Sergey Margasyuk, Maria Vlasenok, Andrey A Mironov, Dmitry Skvortsov, and Dmitri D Pervouchine. Tissue-specific regulation of gene expression via unproductive splicing. *Nucleic Acids Research*, 51(7):3055–3066, apr 2023.

[113] Ashish Misra and Michael R. Green. From polyadenylation to splicing: Dual role for mRNA 3' end formation factors. *RNA Biology*, 13(3):259–264, mar 2016.

[114] Sibylle Mitschka and Christine Mayr. Context-specific regulation and function of mRNA alternative polyadenylation. *Nature reviews. Molecular cell biology*, pages 1–18, jul 2022.

[115] Arundhati Mohanta and Kausik Chakrabarti. Dbr1 functions in mRNA processing, intron turnover and human diseases. *Biochimie*, 180:134–142, jan 2021.

[116] Eric J. Montemayor, Adam Katolik, Nathaniel E. Clark, Alexander B. Taylor, Jonathan P. Schuermann, D. Joshua Combs, Richard Johnsson, Stephen P. Holloway, Scott W. Stevens, Masad J. Damha, and P. John Hart. Structural basis of lariat RNA recognition by the intron debranching enzyme Dbr1. *Nucleic Acids Research*, 42(16):10845–10855, sep 2014.

[117] Maliheh Movassat, Tara L. Crabb, Anke Busch, Chengguo Yao, Derrick J. Reynolds, Yongsheng Shi, and Klemens J. Hertel. Coupling between alternative polyadenylation and alternative splicing is limited to terminal introns. *RNA Biology*, 13(7):646–655, jul 2016.

[118] Sören Müller, Lukas Rycak, Fabian Afonso-Grunz, Peter Winter, Adam M. Zawada, Ewa Damrath, Jessica Scheider, Juliane Schmäh, Ina Koch, Günter Kahl, and Björn Rotter. APADB: a database for alternative polyadenylation and microRNA regulation events. *Database : the journal of biological databases and curation*, 2014, 2014.

[119] Vishal Nanavaty, Elizabeth W. Abrash, Changjin Hong, Sunho Park, Emily E. Fink, Zhuangyue Li, Thomas J. Sweet, Jeffrey M. Bhasin, Srinidhi Singuri, Byron H. Lee, Tae Hyun Hwang, and Angela H. Ting. DNA Methylation Regulates Alternative Polyadenylation via CTCF and the Cohesin Complex. *Molecular Cell*, 78(4):752–764.e6, may 2020.

[120] Christopher R. Neil and William G. Fairbrother. Intronic RNA: Ad'junk' mediator of post-transcriptional gene regulation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1862(11-12):194439, nov 2019.

[121] Jonathan Neve, Radhika Patel, Zhiqiao Wang, Alastair Louey, and André Martin Furger. Cleavage and polyadenylation: Ending the message expands gene regulation. *RNA Biology*, 14(7):865–890, jul 2017.

[122] Jamie Nourse, Stefano Spada, and Sven Danckwardt. Emerging Roles of RNA 3'-end Cleavage and Polyadenylation in Pathogenesis, Diagnosis and Therapy of Human Disorders. *Biomolecules*, 10(6):915, jun 2020.

[123] Jung Min Oh, Chao Di, Christopher C Venters, Jiannan Guo, Chie Arai, Byung Ran So, Anna Maria Pinto, Zhenxi Zhang, Lili Wan, Ihab Younis, and Gideon Dreyfuss. U1 snRNP telescripting regulates a size-function-stratified human genome. *Nature Structural and Molecular Biology*, 24(11):993–999, nov 2017.

[124] Ji Yeon Park, Wencheng Li, Dinghai Zheng, Peiyong Zhai, Yun Zhao, Takahisa Matsuda, Stephen F. Vatner, Junichi Sadoshima, and Bin Tian. Comparative analysis of mRNA isoform expression in Cardiac hypertrophy and development reveals multiple Post-Transcriptional regulatory modules. *PLoS ONE*, 6(7):e22391, 2011.

[125] Dmitri D. Pervouchine, David G. Knowles, and Roderic Guigó. Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics*, 29(2):273–274, jan 2013.

[126] Nick J. Proudfoot. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science*, 352(6291), 2016.

[127] Nick J. Proudfoot, Andre Furger, and Michael J. Dye. Integrating mRNA processing with transcription. *Cell*, 108(4):501–512, feb 2002.

[128] Luye Qin, Jamal B. Williams, Tao Tan, Tiaotiao Liu, Qing Cao, Kaijie Ma, and Zhen Yan. Deficiency of autism risk factor ASH1L in prefrontal cortex induces epigenetic aberrations and seizures. *Nature Communications*, 12(1), dec 2021.

[129] Aaron R. Quinlan and Ira M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, mar 2010.

[130] Fidel Ramírez, Devon P. Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1):W160–W165, jul 2016.

[131] Kirsten A. Reimer, Claudia A. Mimoso, Karen Adelman, and Karla M. Neugebauer. Co-transcriptional splicing regulates 3' end cleavage during mammalian erythropoiesis. *Molecular Cell*, 81(5):998–1012.e7, mar 2021.

[132] Marina Reixachs-Solé and Eduardo Eyras. Uncovering the impacts of alternative splicing on the proteome with current omics techniques, 2022.

[133] Frank Rigo and Harold G. Martinson. Functional coupling of last-intron splicing and 3'-end processing to transcription in vitro: the poly(A) signal couples to splicing before committing to cleavage. *Molecular and cellular biology*, 28(2):849–62, jan 2008.

[134] Frank Rigo and Harold G. Martinson. Polyadenylation releases mRNA from RNA polymerase II in a process that is licensed by splicing. *RNA*, 15(5):823–836, may 2009.

[135] Gordon Robertson, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D. Jackman, Karen Mungall, Sam Lee, Hisanaga Mark Okada, Jenny Q. Qian, Malachi Griffith, Anthony Raymond, Nina Thiessen, Timothee Cezard, Yaron S. Butterfield, Richard Newsome, Simon K. Chan, Rong She, Richard Varhol, Baljit Kamoh, Anna Liisa Prabhu, Angela Tam, Yongjun Zhao, Richard A. Moore, Martin Hirst, Marco A. Marra, Steven J.M. Jones, Pamela A. Hoodless, and Inanc Birol. De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7(11):909–912, oct 2010.

[136] Michael I. Robson, Jose I. de las Heras, Rafal Czapiewski, Phú Lê Thành, Daniel G. Booth, David A. Kelly, Shaun Webb, Alastair R.W. Kerr, and

Eric C. Schirmer. Tissue-Specific Gene Repositioning by Muscle Nuclear Membrane Proteins Enhances Repression of Critical Developmental Genes during Myogenesis. *Molecular Cell*, 62(6):834–847, jun 2016.

[137] J. Rogers, P. Early, C. Carter, K. Calame, M. Bond, L. Hood, and R. Wall. Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin μ chain. *Cell*, 20(2):303–312, 1980.

[138] Kevin R. Roy and Guillaume F. Chanfreau. Robust mapping of polyadenylated and non-polyadenylated RNA 3' ends at nucleotide resolution by 3'-end sequencing. *Methods*, 176:4–13, apr 2019.

[139] Barbara Ruskin and Michael R. Green. An RNA processing activity that debranches RNA lariats. *Science*, 229(4709):135–140, 1985.

[140] Piero Sanfilippo, Pedro Miura, and Eric C. Lai. Genome-wide profiling of the 3' ends of polyadenylated RNAs. *Methods*, 126:86–94, aug 2017.

[141] Peter Schäfer, Christian Tüting, Lars Schönemann, Uwe Kühn, Thomas Treiber, Nora Treiber, Christian Ihling, Anne Graber, Walter Keller, Gunter Meister, Andrea Sinz, and Elmar Wahle. Reconstitution of mammalian cleavage factor II involved in 3' processing of mRNA precursors. *RNA*, 24(12):1721–1737, dec 2018.

[142] Ashleigh E. Schaffer, Veerle R.C. Eggens, Ahmet Okay Caglayan, Miriam S. Reuter, Eric Scott, Nicole G. Coufal, Jennifer L. Silhavy, Yuanchao Xue, Hulya Kayserili, Katsuhito Yasuno, Rasim Ozgur Rosti, Mostafa Abdellateef, Caner Caglar, Paul R. Kasher, J. Leonie Cazemier, Marian A. Weterman, Vincent Cantagrel, Na Cai, Christiane Zweier, Umut Altunoglu, N. Bilge Satkin, Fesih Aktar, Beyhan Tuysuz, Cengiz Yalcinkaya, Huseyin Caksen, Kaya Bilguvar, Xiang Dong Fu, Christopher R. Trotta, Stacey Gabriel, André Reis, Murat Gunel, Frank Baas, and Joseph G. Gleeson. CLP1 founder mutation links tRNA splicing and maturation to cerebellar development and neurodegeneration. *Cell*, 157(3):651–663, apr 2014.

[143] Ankeeta Shah, Briana E. Mittleman, Yoav Gilad, and Yang I. Li. Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation. *Genome Biology*, 22(1):1–21, dec 2021.

[144] Michael D. Sheets, Stephen C. Ogg, and Marvin P. Wickens. Point mutations in AAUAAA and the poly (A) addition site: Effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Research*, 18(19):5799–5805, oct 1990.

[145] Sol Shenker, Pedro Miura, Piero Sanfilippo, and Eric C Lai. IsoSCM: Improved and alternative 3' UTR annotation using multiple change-point inference. *RNA*, 21(1):14–27, 2015.

[146] Peter J. Shepard, Eun A. Choi, Jente Lu, Lisa A. Flanagan, Klemens J. Hertel, and Yongsheng Shi. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, 17(4):761–772, apr 2011.

[147] Priyam Singh, Travis L. Alley, Sarah M. Wright, Sonya Kamdar, William Schott, Robert Y Wilpan, Kevin D. Mills, and Joel H. Graber. Global changes in processing of mRNA 3' untranslated regions characterize clinically distinct cancer subtypes. *Cancer Research*, 69(24):9422–9430, dec 2009.

[148] Byung Ran So, Chao Di, Zhiqiang Cai, Christopher C. Venters, Jiannan Guo, Jung-Min Oh, Chie Arai, and Gideon Dreyfuss. A Complex of U1 snRNP with Cleavage and Polyadenylation Factors Controls Telescripting, Regulating mRNA Transcription in Human Cells. *Molecular Cell*, 76(4):590–599.e4, nov 2019.

[149] Pia Sommerkamp, Nina Cabezas-Wallscheid, and Andreas Trumpp. Alternative Polyadenylation in Stem Cell Self-Renewal and Differentiation. *Trends in Molecular Medicine*, 27(7):660–672, 2021.

[150] Simon N. Stacey, Patrick Sulem, Aslaug Jonasdottir, Gisli Masson, Julius Gudmundsson, Daniel F. Gudbjartsson, Olafur T. Magnusson, Sigurjon A. Gudjonsson, Bardur Sigurgeirsson, Kristin Thorisdottir, Rafn Ragnarsson, Kristrun R. Benediktsdottir, Bjørn A. Nexø, Anne Tjønneland, Kim Overvad, Peter Rudnai, Eugene Gurzau, Kvetoslava Koppova, Kari Hemminki, Cristina Corredera, Victoria Fuentelsaz, Pilar Grasa, Sebastian Navarrete, Fernando Fuertes, Maria D. García-Prats, Enrique Sanambrosio, Angeles Panadero, Ana De Juan, Almudena Garcia, Fernando Rivera, Dolores Planelles, Virtudes Soriano, Celia Requena, Katja K. Aben, Michelle M. Van Rossum, Ruben G.H.M. Cremers, Inge M. Van Oort, Dick Johan Van Spronsen, Jack A. Schalken, Wilbert H.M. Peters, Brian T. Helfand, Jenny L. Donovan, Freddie C. Hamdy, Daniel Badescu, Ovidiu Codreanu, Mariana Jinga, Irma E. Csiki, Vali Constantinescu, Paula Badea, Ioan N. Mates, Daniela E. Dinu, Adrian Constantin, Dana Mates, Sjofn Kristjansdottir, Bjarni A. Agnarsson, Eirikur Jonsson, Rosa B. Barkardottir, Gudmundur V. Einarsson, Fridbjorn Sigurdsson, Pall H. Moller, Tryggvi Stefansson, Trausti Valdimarsson, Oskar T. Johannsson, Helgi Sigurdsson, Thorvaldur Jonsson, Jon G. Jonasson, Laufey Tryggvadottir, Terri Rice, Helen M. Hansen, Yuanyuan Xiao, Daniel H. Lachance, Brian Patrick O'Neill, Matthew L. Kosel, Paul A. Decker, Gudmar Thorleifsson, Hrefna Johannsdottir, Hafdis T. Helgadottir, Asgeir Sigurdsson, Valgerdur Steinthorsdottir, Annika Lindblom, Robert S. Sandler, Temitope O. Keku, Karina Banasik, Torben Jørgensen, Daniel R. Witte, Torben Hansen, Oluf Pedersen, Viorel Jinga, David E. Neal, William J. Catalona, Margaret Wrensch, John Wiencke, Robert B. Jenkins, Eduardo Nagore, Ulla Vogel, Lambertus A. Kiemeney, Rajiv Kumar, José I. Mayordomo, Jon H. Olafsson, Augustine Kong, Unnur Thorsteinsdottir, Thorunn Rafnar, and Kari Stefansson. A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nature Genetics*, 43(11):1098–1103, nov 2011.

[151] Yadong Sun, Keith Hamilton, and Liang Tong. Recent molecular insights into canonical pre-mRNA 3'-end processing. *Transcription*, 11(2):83–96, mar 2020.

[152] Yadong Sun, Yixiao Zhang, Keith Hamilton, James L. Manley, Yongsheng Shi, Thomas Walz, and Liang Tong. Molecular basis for the recognition of the human AAUAAA polyadenylation signal. *Proceedings of the National Academy of Sciences of the United States of America*, 115(7):E1419–E1428, feb 2018.

[153] Krzysztof Szkop and Irene Nobeli. Untranslated parts of genes interpreted: making heads or tails of high-throughput transcriptomic data via computational methods. *Bioessays*, 39(12):71–84, 2019.

[154] Gaëlle J.S. Talhouarne and Joseph G. Gall. Lariat intronic RNAs in the cytoplasm of vertebrate cells. *Proceedings of the National Academy of Sciences of the United States of America*, 115(34):E7970–E7977, aug 2018.

[155] Peng Tang, Yang Yang, Guangnan Li, Li Huang, Miaomiao Wen, Wen Ruan, Xiaolong Guo, Chen Zhang, Xinxin Zuo, Daji Luo, Yongzhen Xu, Xiang Dong Fu, and Yu Zhou. Alternative polyadenylation by sequential activation of distal and proximal PolyA sites. *Nature Structural and Molecular Biology*, 29(1):21–31, 2022.

[156] Xujun Tang, Jiuxia Wang, Shuhong Zhou, Jing Zhou, Guyou Jia, Han Wang, Chunlei Xin, Guoning Fu, and Jiahong Zhang. MiR-760 regulates skeletal muscle proliferation in rheumatoid arthritis by targeting Myo18b. *Molecular Medicine Reports*, 20(6):4843–4854, 2019.

[157] Bin Tian, Jun Hu, Haibo Zhang, and Carol S. Lutz. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, 33(1):201–212, jan 2005.

[158] Bin Tian and James L. Manley. Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology*, 18(1):18–30, 2016.

[159] Bin Tian, Zhenhua Pan, and Ju Youn Lee. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome research*, 17(2):156–65, feb 2007.

[160] Maksim Tikhonov, Pavel Georgiev, and Oksana Maksimenko. Competition within Introns: Splicing Wins over Polyadenylation via a General Mechanism. *Acta Naturae*, 5(4):52, 2013.

[161] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. Van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, may 2010.

[162] Ilya Vainberg Slutskin, Adina Weinberger, and Eran Segal. Sequence determinants of polyadenylation-mediated regulation. *Genome Research*, 29(10):1635–1647, oct 2019.

[163] Maria Vlasenok, Sergey Margasyuk, and Dmitri D. Pervouchine. Transcriptome sequencing suggests that pre-mRNA splicing counteracts widespread intronic cleavage and polyadenylation. *NAR Genomics and Bioinformatics*, 5(2):lqad051, mar 2023.

[164] Sandra Vorlová, Gina Rocco, Clare V. LeFave, Francine M. Jodelka, Ken Hess, Michelle L. Hastings, Erik Henke, and Luca Cartegni. Induction of Antagonistic Soluble Decoy Receptor Tyrosine Kinases by Intronic PolyA Activation. *Molecular Cell*, 43(6):927–939, sep 2011.

[165] Eileen Wagner and Jens Lykke-Andersen. mRNA surveillance: The perfect persist. *Journal of Cell Science*, 115(15):3033–3038, aug 2002.

[166] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, nov 2008.

[167] Hang Wang, Becky L. Sartini, Clarke F. Millette, and Daniel L. Kilpatrick. A developmental switch in transcription factor isoforms during spermatogenesis controlled by alternative messenger RNA 3'-end formation. *Biology of Reproduction*, 75(3):318–323, sep 2006.

[168] Ruijia Wang, Ram Nambiar, Dinghai Zheng, and Bin Tian. PolyA-DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Research*, 46(D1):D315–D319, jan 2018.

[169] Ruijia Wang and Bin Tian. APAlyzer: A bioinformatics package for analysis of alternative polyadenylation isoforms. *Bioinformatics*, 36(12):3907–3909, jun 2020.

[170] Ruijia Wang, Dinghai Zheng, Lu Wei, Qingbao Ding, and Bin Tian. Regulation of Intronic Polyadenylation by PCF11 Impacts mRNA Expression of Long Genes. *Cell Reports*, 26(10):2766–2778.e6, mar 2019.

[171] Ruijia Wang, Dinghai Zheng, Ghassan Yehia, and Bin Tian. A compendium of conserved cleavage and polyadenylation events in mammalian genes. *Genome Research*, 28(10):1427–1441, 2018.

[172] Wei Wang, Zhi Wei, and Hongzhe Li. A change-point model for identifying 3'UTR switching by next-generation RNA sequencing. *Bioinformatics*, 30(15):2162–2170, aug 2014.

[173] Max E. Wilkinson, Clément Charenton, and Kiyoshi Nagai. RNA Splicing by the Spliceosome. *Annual review of biochemistry*, 89(1):359–388, jun 2020.

[174] Steven Wormsley, Dmitry A. Samarsky, Maurille J. Fournier, and Susan J. Baserga. The 3'-end-processing factor CPSF is required for the splicing of single-intron pre-mRNAS in vivo. *RNA*, 7(6):920–931, 2001.

[175] Xuebing Wu and David P. Bartel. Widespread Influence of 3'-End Structures on Mammalian mRNA Processing and Stability. *Cell*, 169(5):905–917.e11, may 2017.

[176] Zheng Xia, Lawrence A. Donehower, Thomas A. Cooper, Joel R. Neilson, David A. Wheeler, Eric J. Wagner, and Wei Li. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nature Communications*, 5(1):5274, dec 2014.

[177] Kehui Xiang, Takashi Nagaike, Song Xiang, Turgay Kilic, Maia M. Beh, James L. Manley, and Liang Tong. Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex. *Nature*, 467(7316):729–733, oct 2010.

[178] Chuan Xu and Jianzhi Zhang. Alternative Polyadenylation of Mammalian Transcripts Is Generally Deleterious, Not Adaptive. *Cell Systems*, 6(6):734–742.e4, jun 2018.

[179] Chuan Xu and Jianzhi Zhang. Mammalian circular RNAs result largely from splicing errors. *Cell Reports*, 36(4), jul 2021.

[180] Zhuyi Xue, René L. Warren, Ewan A. Gibb, Daniel MacMillan, Johnathan Wong, Readman Chiu, S. Austin Hammond, Chen Yang, Ka Ming Nip, Catherine A. Ennis, Abigail Hahn, Sheila Reynolds, and Inanc Birol. Recurrent tumor-specific regulation of alternative polyadenylation of cancer-related genes. *BMC Genomics*, 19(1):536, jul 2018.

[181] Yan Yang, Wencheng Li, Mainul Hoque, Liming Hou, Steven Shen, Bin Tian, and Brian D. Dynlacht. PAF Complex Plays Novel Subunit-Specific Roles in Alternative Cleavage and Polyadenylation. *PLoS Genetics*, 12(1):e1005794, 2016.

[182] Makiko Yasuda, Junaid Shabbeer, Makiko Osawa, and Robert J. Desnick. Fabry disease: Novel α-galactosidase A 3'-terminal mutations result in multiple transcripts due to aberrant 3'-end formation. *American Journal of Human Genetics*, 73(1):162–173, jul 2003.

[183] Congting Ye, Yuqi Long, Guoli Ji, Qingshun Quinn Li, and Xiaohui Wu. APAtrap: Identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics*, 34(11):1841–1849, jun 2018.

[184] Wenbin Ye, Qiwei Lian, Congting Ye, and Xiaohui Wu. A Survey on Methods for Predicting Polyadenylation Sites from DNA Sequences, Bulk RNA-seq, and Single-cell RNA-seq. *Genomics, Proteomics & Bioinformatics*, sep 2022.

[185] Leiming You, Jiexin Wu, Yuchao Feng, Yonggui Fu, Yanan Guo, Liyuan Long, Hui Zhang, Yijie Luan, Peng Tian, Liangfu Chen, Guangrui Huang, Shengfeng

Huang, Yuxin Li, Jie Li, Chengyong Chen, Yaqing Zhang, Shangwu Chen, and Anlong Xu. APASdb: A database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucleic Acids Research*, 43(D1):D59–D67, jan 2015.

[186] Fengyun Yu, Yu Zhang, Chao Cheng, Wenqing Wang, Zisong Zhou, Wenliang Rang, Han Yu, Yaxun Wei, Qijia Wu, and Yi Zhang. Poly(A)-seq: A method for direct sequencing and analysis of the transcriptomic poly(A)-tails. *PLoS ONE*, 15(6), jun 2020.

[187] Margarita I. Zarudnaya, Iryna M. Kolomiets, Andriy L. Potyahaylo, and Dmytro M. Hovorun. Downstream elements of mammalian pre-mRNA polyadenylation signals: Primary, secondary and higher-order structures. *Nucleic Acids Research*, 31(5):1375–1386, mar 2003.

[188] Daniel R. Zerbino, Nathan Johnson, Thomas Juettemann, Steven P. Wilder, and Paul Flicek. WiggleTools: Parallel processing of large collections of genome-wide datasets for visualization and statistical analysis. *Bioinformatics*, 30(7):1008–1009, apr 2014.

[189] Aihua Zhang, Shaohua Li, Lynne Apone, Xiaoli Sun, Lixin Chen, Laurence M. Ettwiller, Bradley W. Langhorst, Christopher J. Noren, and Ming Qun Xu. Solid-phase enzyme catalysis of DNA end repair and 3' A-tailing reduces GC-bias in next-generation sequencing of human genomic DNA. *Scientific Reports*, 8(1):15887, dec 2018.

[190] Lan Zhang and Weihua Zhang. Knockdown of NUDT21 inhibits proliferation and promotes apoptosis of human K562 leukemia cells through ERK pathway. *Cancer Management and Research*, 10:4311–4323, 2018.

[191] Yixiao Zhang, Yadong Sun, Yongsheng Shi, Thomas Walz, and Liang Tong. Structural Insights into the Human Pre-mRNA 3'-End Processing Machinery. *Molecular Cell*, 77(4):800–809.e6, feb 2020.

[192] Zhiping Zhang, Bongmin Bae, Winston H. Cuddleston, and Pedro Miura. Coordination of alternative splicing and alternative polyadenylation revealed by targeted long read sequencing. *Nature Communications*, 14(1):1–14, sep 2023.

[193] Zhaozhao Zhao, Qiushi Xu, Ran Wei, Weixu Wang, Dong Ding, Yu Yang, Jun Yao, Liye Zhang, Yue Qing Hu, Gang Wei, and Ting Ni. Cancer-associated dynamics and potential regulators of intronic polyadenylation revealed by IPAFinder using standard RNA-seq data. *Genome Research*, 31(11):2095–2106, sep 2021.

[194] Dinghai Zheng, Xiaochuan Liu, and Bin Tian. 3'READS+, a sensitive and accurate method for 3' end sequencing of polyadenylated RNA. *RNA*, 22(10):1631–1639, oct 2016.

[195] Dinghai Zheng and Bin Tian. RNA-binding proteins in regulation of alternative cleavage and polyadenylation. In Gene W. Yeo, editor, *Systems Biology of RNA Binding Proteins*, pages 97–127. Springer New York, New York, NY, 2014.

[196] Simin Zheng, Bao Q. Vuong, Bharat Vaidyanathan, Jia Yu Lin, Feng Ting Huang, and Jayanta Chaudhuri. Non-coding RNA Generated following Lariat Debranching Mediates Targeting of AID to DNA. *Cell*, 161(4):762–773, may 2015.

[197] Yong Zhu, Xiuye Wang, Elmira Forouzmand, Joshua Jeong, Feng Qiao, Gregory A. Sowd, Alan N. Engelman, Xiaohui Xie, Klemens J. Hertel, and Yongsheng Shi. Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation. *Molecular Cell*, 69(1):62–74.e4, jan 2018.

Appendix A

Additional Resources

The source code for the PAS identification pipeline is available at https://github.com/mashlosenok/RNAseq_PAS_finder.

A.1 Supplementary figures

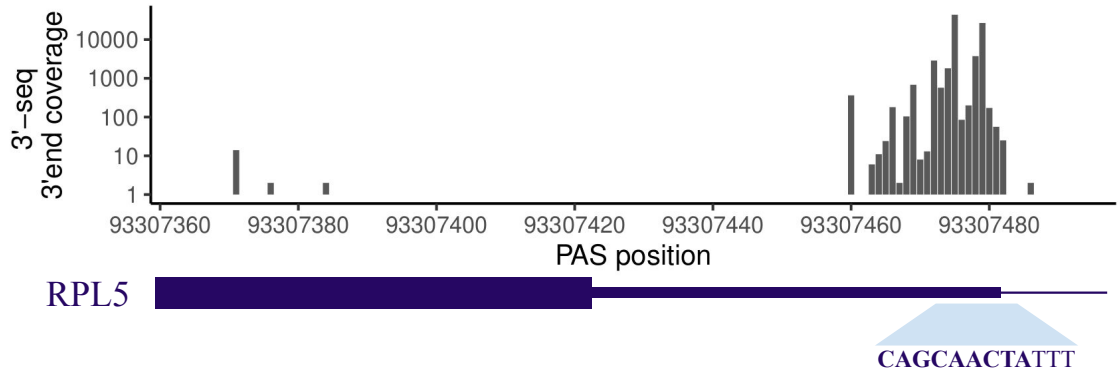


Figure A-1: **3'-seq reads 5'-end peaks width. Example.** The distribution of 3'-seq reads from GSE111793 near *RPL5* transcript end.

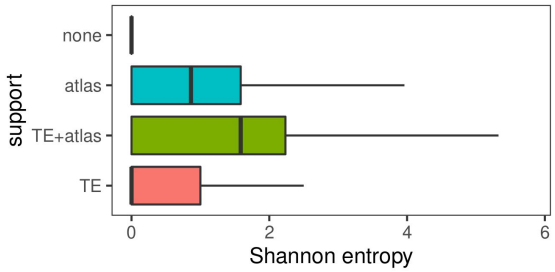


Figure A-2: **Shannon entropy of soft clip length is informative for true PAS.** Shannon entropy distribution of PASs supported by GENCODE transcript ends (red), PolyASite2.0 clusters (blue), both (green) and neither (violet).

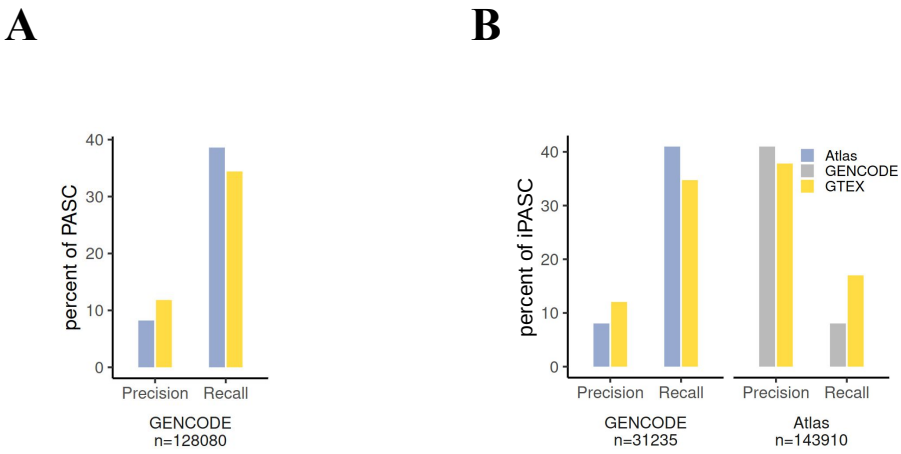


Figure A-3: **Pairwise comparison of PASs inferred from GTEx, PolyA-Site 2.0 (Atlas), and GENCODE** for the window of 50 nts around the annotated transcript end (A) and for a set of intronic PASCs (B). (**A**) The 50-nt window was used in the validation of DaPars [176]. (**B**) Same as Figure 5-15, but restricted to the subset of intronic PASCs. The rest of the legend is the same as in Figure 5-15.

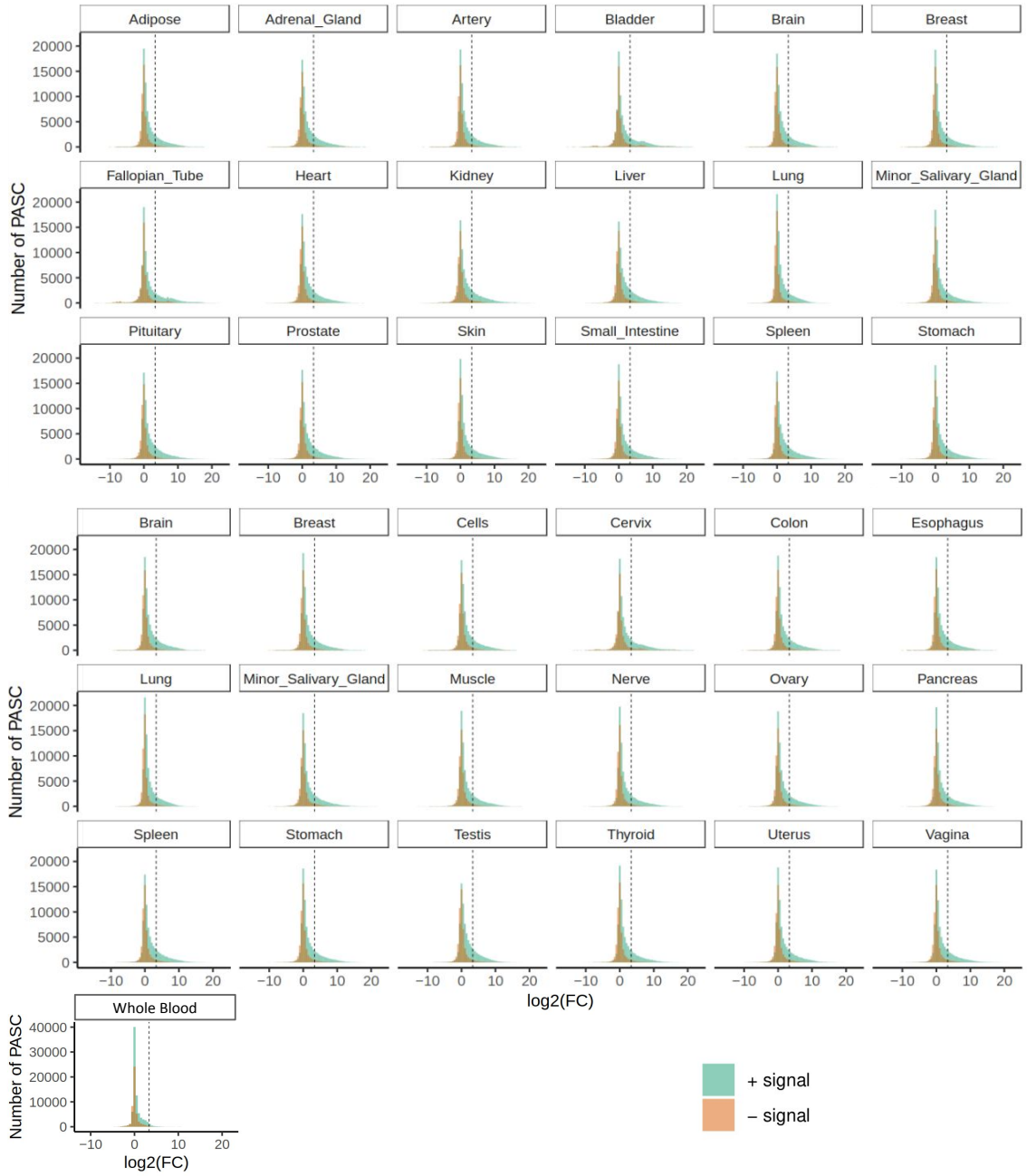


Figure A-4: **Coverage-based metrics of PASC usage. Per tissue.** The distribution of $\log_2(wi_1/wi_2)$ metric for each PASC plotted separately for each tissue. The dashed line represents the cutoff $wi_1/wi_2 = 10$.

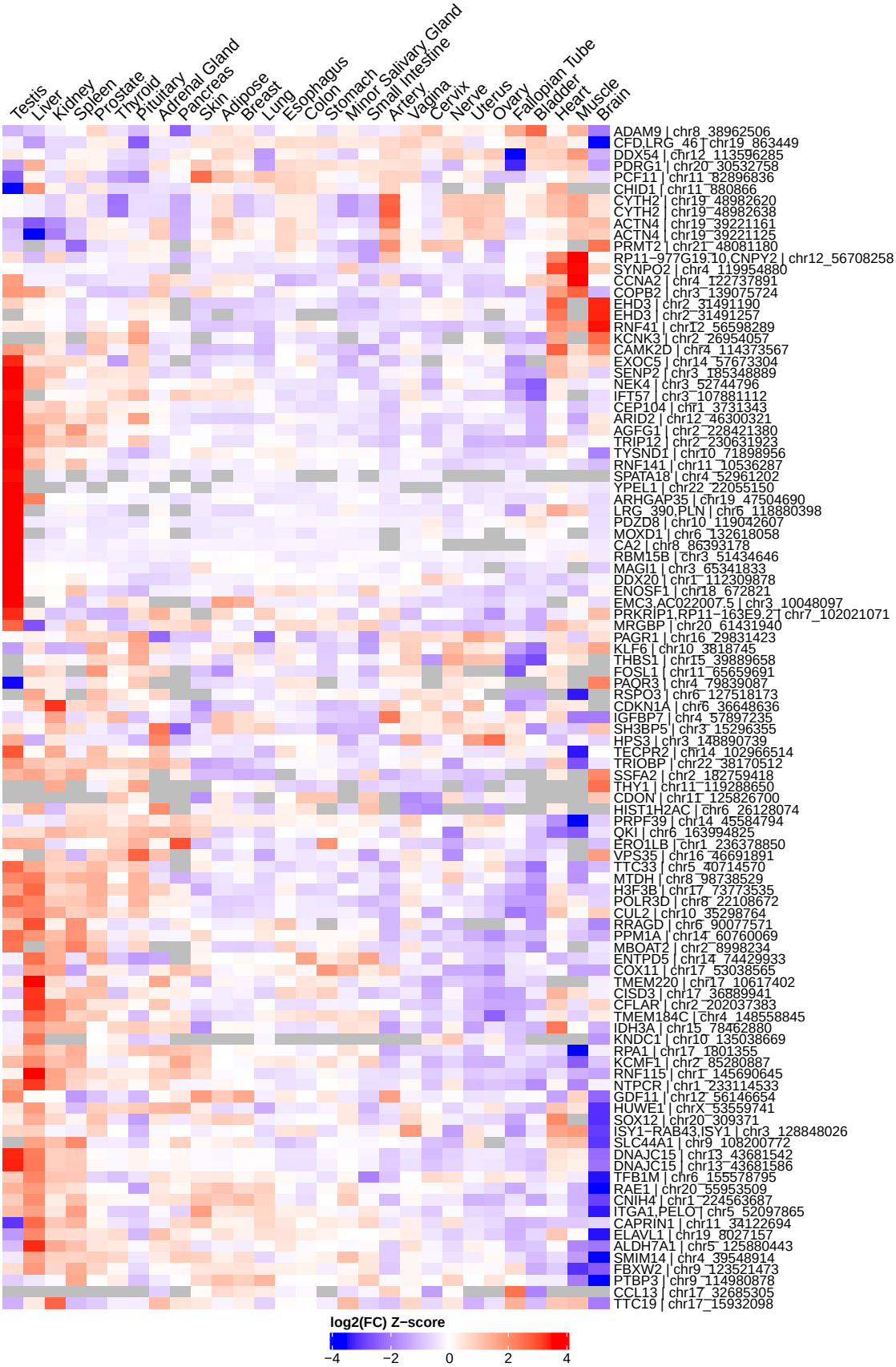


Figure A-5: **Tissue-specific PAS.**

Figure A-5: (Previous page.) The heatmap depicts the Z-scores of coverage fold changes (FC) at PAS for 100 selected sites showing substantial variation among tissues. Read coverage at each of these sites experiences a minimum 10-fold decrease in at least one tissue and less than a 2-fold decrease in another. Grey squares represent tissues where the coverage upstream of the PAS was lower than 10.

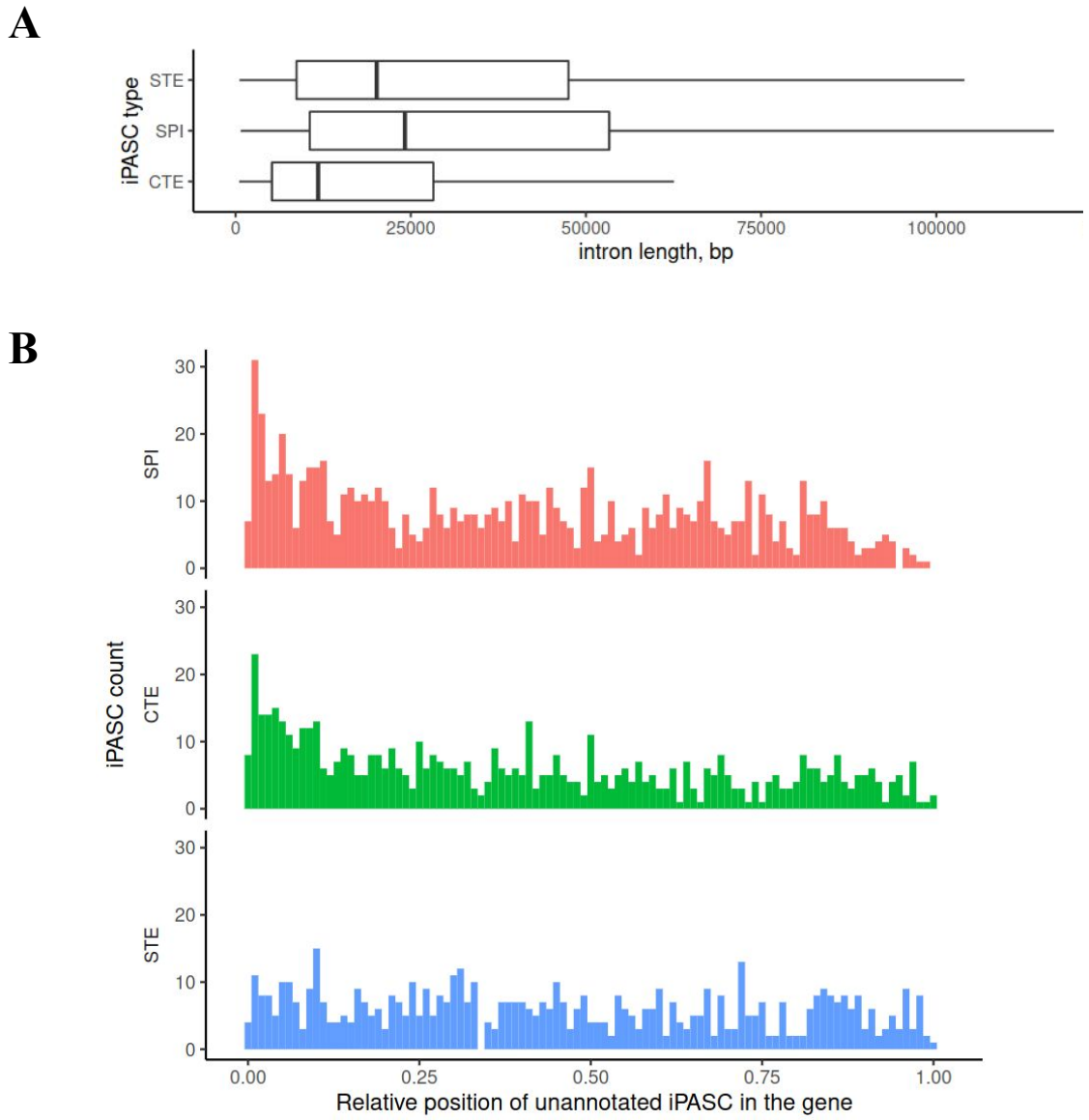


Figure A-6: **STEs, CTEs, and SPIs in the gene.**(**A**) The distribution of lengths of the introns containing the iPASCs. For each iPASC only the shortest containing it intron was considered. (**B**) The relative position in the gene of unannotated iPASCs (same as in Figure 5-17). An iPASC was categorized as CTE, STE, and SPI if it belonged to the respective class for at least one iPASC-tissue pair.

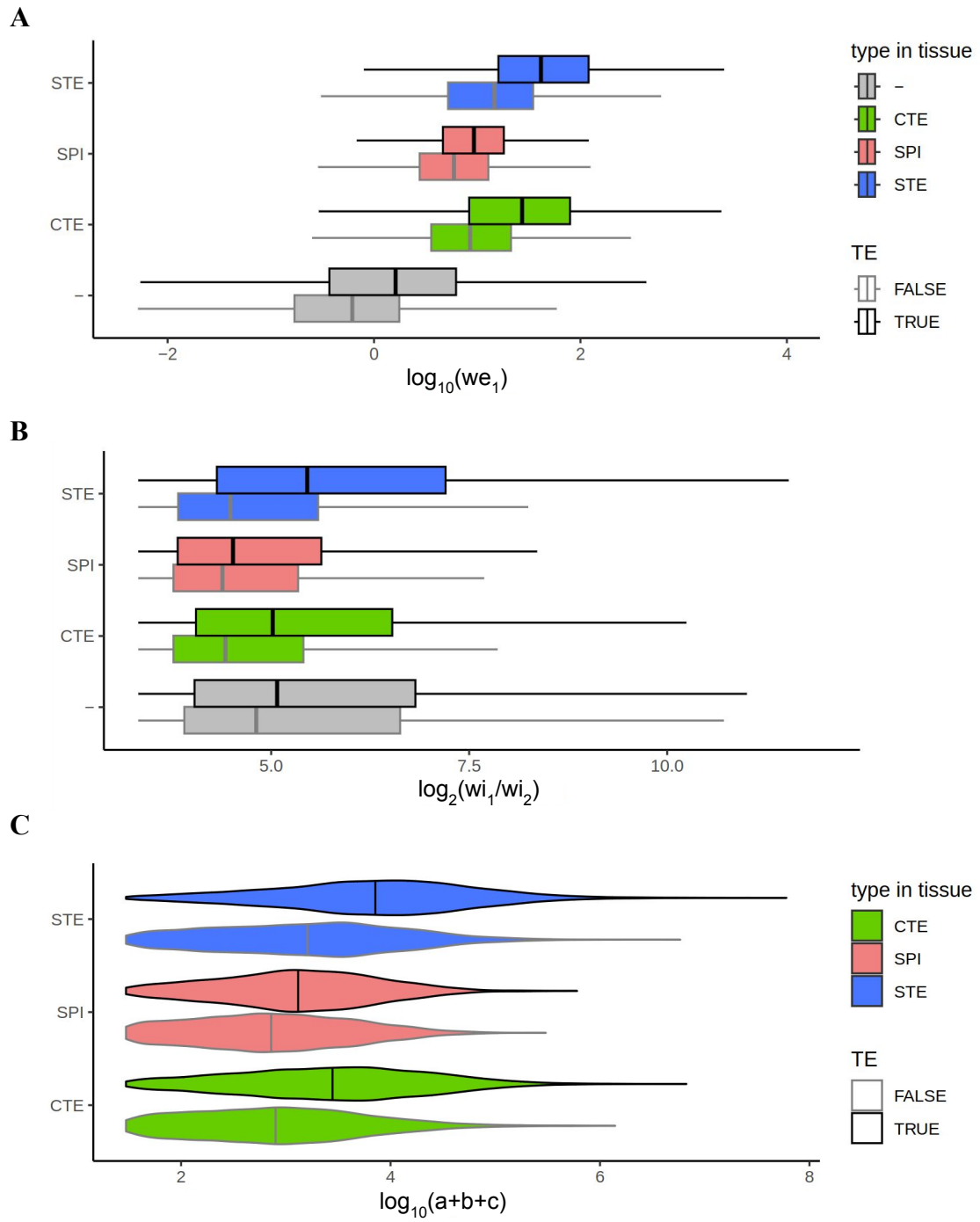


Figure A-7: **Coverage and PAS usage of STE, CTE, and SPI iPASC types in tissues.** (**A**) The distribution of coverages of the preceding exon's end (we_1) for the three iPASC-tissue pair types. (**B**) Coverage drop at PAS for the three iPASC-tissue pair types. (**C**) Local expression levels ($= a+b+c$) for the three iPASC-tissue pair types. Each iPASC was categorized as CTE, STE, SPI,"-" or "not expressed" in each tissue based on tissue-specific coverage in the four windows we_1, we_2, wi_1, wi_2 and number of split or continuous reads supporting AS events in the intron (5.2.4). Annotated (within 100 nt of an annotated TE) and unannotated iPASCs are shown separately (grey and black outlines, respectively). "-" type are used iPASCs that could not be classified as CTE, STE or SPI due to low local expression. They are omitted in the C plot because they are defined by a low $a + b + c$ value.

A

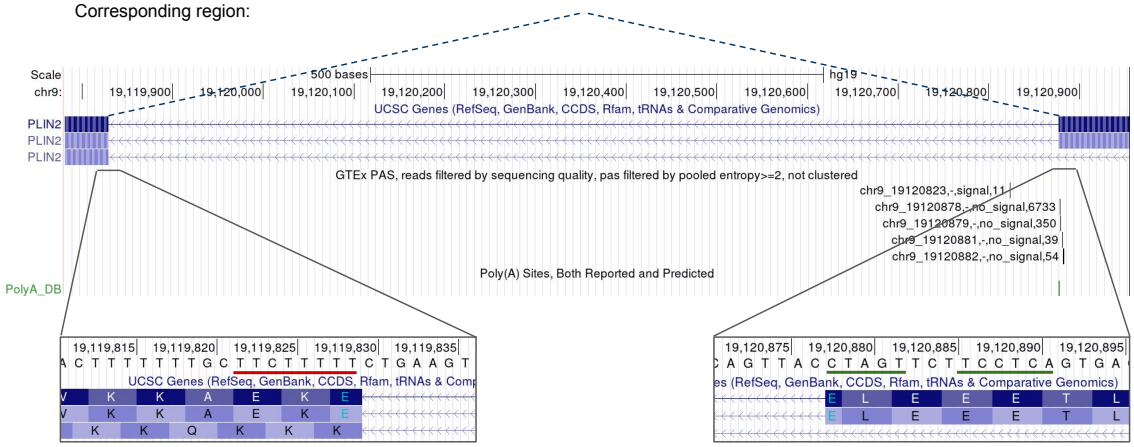
polyA reads:

SRR1400910.12641453 83 9 19120877 255 8S68M 9 19119788 68
TTCTTTTTCTAGTAATTCCTCAATGAGAGGGAGGTACTGTTTTACCAACAGCAATGATTTGGTGAGTGCATTTACT

SRR1400910.12645416 163 9 19120877 255 6S70M 9 19120948 70
CTTTTTCTAGTCCTTCCTCAGTGAGAGGGAGGTACTGTTCTACCAACAGCTCTGATTTGGTGAGTGCATTTTCTAC

SRR1400910.12652323 163 9 19120877 255 9S67M 9 19123575 67
TTTTTTTTTCTAGTTCGTCCTCAGTGAGAGGGAGGTACTGTTCTACCAACAGCTCTGATTTGGTGAGTGCATTTTC

Corresponding region:



B

polyA reads:

SRR1402414.19165097 163 14 23851636 255 8S68M 14 23851740 68
ATTTTTTICTTGGCACCAATGTCACGGCTCTTGGCTCGAAGCTTGTTGACCTGGGACTCATCGATGTCCGCCCGCT

SRR1402414.19165230 147 14 23851636 255 7S69M 14 23851140 69
TTTTTTGCTTGGCCCCAATGTCACGGCTCTTGGCTCGAAGCTTGTTGACCTGGGACTCAGCGATGTCCGCCCGCTC

SRR1402414.19180339 163 14 23851636 255 6S69M1S 14 23851721 69
TTTTTGCTTGICACCAATGTCACGGCTCTTGGCTCGAAGCTTGTTGACCTGGGACTCAGCGATGTCCGCCCGCTCN

Corresponding region:

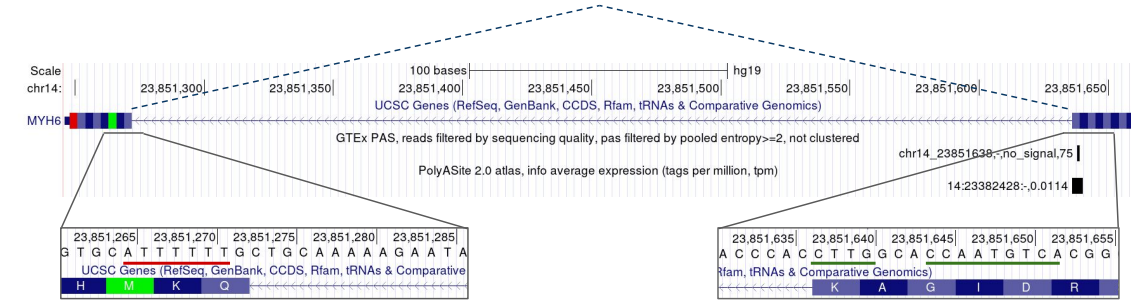


Figure A-8: Two examples (A and B) of mapping artifacts, where a spurious PAS is incorrectly placed near the exon boundary due to the A-rich regions at the beginning of the next exon. The A/T-rich ends of the reads (red) are incorrectly soft-clipped by the mapper. Short-read alignments are shown on the positive strand as they appear in the BAM file.

A.2 Supplementary tables

Online materials are available at https://zenodo.org/record/7799648.

Table A.1: **PASCs distribution among different regions of protein-coding genes.**

spliced/ unspliced	**translated/ untranslated**	**length kb**	**PASC count, 1000s**	**pA read count, million**	**PASC density, per kb**	**pA read density, per kb**
always exon	3'-UTR	31135	58.0	186.87	1.86	6001.92
always exon	5'-UTR	4976.5	0.9	0.09	0.17	17.86
always exon	ORF	27456	12.7	0.35	0.46	12.68
intron	3'-UTR	18203	2.2	1.21	0.12	66.63
intron	5'-UTR	297960	12.7	1.49	0.04	4.99
intron	ORF	866167	47.0	11.40	0.05	13.16
alternative exon	3'-UTR	7657	9.5	19.80	1.24	2585.53
alternative exon	5'-UTR	5830.5	0.7	0.16	0.12	28.27
alternative exon	ORF	4276	2.1	0.07	0.50	17.19

Table A.2: **Fraction of polyA reads in non-UTR regions of protein-coding genes.** The polyA read density normalized to the average read coverage in exonic and intronic regions.

spliced/ unspliced	**length, kb**	**pA read density, per Mb**	**total read density, per bp**	**pA reads per mil reads**
always exon	27456	12682	17686	0.717
intron	866167	13161	2199	5.986
alternative exon	4276	17193	18831	0.913

Table A.3: **iPASC types in tissues**. Median values of coverage characteristics for different iPASC types in tissues. Window coverage values we_1 and wi_1 are in sequenced nts per bp.

type in tissue	**annotated**	we_1/we_2	wi_1/we_2	wi_1/wi_2	we_1	wi_1	**local expression**
-	FALSE	3.54	2.11	27.62	0.93	0.52	3
-	TRUE	2.71	1.74	33.46	1.55	0.93	3
CTE	FALSE	1.86	0.84	21.45	8.49	4.75	779.5
CTE	TRUE	1.38	0.85	32.45	27.11	23.55	2800
SPI	FALSE	12.58	3.00	20.89	5.95	1.46	698.5
SPI	TRUE	12.19	2.96	22.92	9.28	2.13	1267
STE	FALSE	11.70	5.98	22.40	14.65	5.36	1578
STE	TRUE	22.73	18.20	43.88	41.45	32.97	7212
no expres.	FALSE	8.03	0.73	5.40	6.69	0.57	517
no expres.	TRUE	6.31	1.10	5.97	14.93	2.51	1229

Table A.4: The list of 565,387 human polyadenylation site (PAS) from GTEx with entropy$\geq$ 2 and minimum overhang of 6 nucleotides.

See online materials.

Table A.5: The list of 318,898 PAS cluster (PASC)s in human protein-coding genes.

See online materials.

Table A.6: The list of 126,310 PASCs located $>$ 200 nts away from exon boundaries with the coverage fold change at the PASCs in GTEx tissues.

See online materials.

Table A.7: The list of 67,075 intronic PASCs in 31 tissues, the coverage fold change at the PASCs in GTEx tissues, annotation status and categorization as CTE, STE, or SPI.

See online materials.