Computational methods to identify and characterize the functionality of polyadenylation isoforms

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Dominik Burri

2023

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Erstbetreuerin: Prof. Dr. Mihaela Zavolan

Zweitbetreuer: Prof. Dr. Erik van Nimwegen

Externer Experte: Prof. Dr. Sebastian Leidel

Basel, den 23.05.2023

Prof. Dr. Marcel Mayor
Dekan

Acknowledgments

I want to thank all the former and current Zavolan group members. They supported me with general advice, technical assistance, and elucidations of mathematical and biological questions. They gave me an appreciation of community across disciplines that I did not have before. They motivated me with their curiosity and passion for research. And this made my journey a pleasure.

I express my sincere gratitude to my supervisor Mihaela, who made it possible that I carried out my PhD in her group. I am very thankful for all the support and mentorship I was given for the whole duration of the training. I am tremendously grateful that I was able to experiment with new approaches and to immediately apply what I learnt to practice.

I extend my gratitude towards my committee members Erik and Sebastian that provided me with thorough advice and valuable insights during the highly fruitful committee meeting discussions.

I was extremely pleased to have been a member of the Biozentrum PhD Fellowship program. This not only gave me the chance to connect with fellows early on. It also gave me the liberty and independence to decide on a group to pursue my PhD.

Finally, I want to do credit to my family, friends, and partner. They always supported and encouraged me throughout this extended venture. I would not be here without you.

Abstract

Polyadenylation is the process by which poly(A) tails are added at the 3’ end of an RNA molecule. Alternative polyadenylation (APA) can occur when multiple poly(A) sites exist. This gives rise to transcript isoforms which can have different 3’ untranslated regions (3’UTRs) or can lead to different protein products.

APA has been shown to influence the stability, localization and translation of the mRNA molecules. APA is known to be involved in health and disease and is tissue- and cell type-specific. Although a lot is known about the processing machinery, less is known about the context-specific selection of a specific poly(A) site. With the appearance of single-cell transcriptomics (scRNA-seq), it got possible to study APA on the level of individual cells.

I developed SCUREL, a computational method that detects 3’UTR changes between two sets of cells. We found SCUREL to be more sensitive compared to a similar method. Applying SCUREL to lung tumor, we found that the global 3’UTR shortening in tumor tissues cannot be explained by the proliferative cancer cells alone, but by a combination of most cell types composing the tumor tissue. Additionally, the proteins targeted by 3’UTR shortening are mostly implicated in protein metabolism and localization processes.

Since we noticed that the 3’UTR annotations were incomplete, we developed a computational method that identifies novel poly(A) sites from scRNA-seq and termed it SCINPAS. It extracts poly(A) containing reads and identifies poly(A) sites irrespective of genome annotation. We assessed the performance of SCINPAS on systems with known effects and against a competing method. We demonstrated its usability and its ability to detect novel poly(A) sites in genic and non-genic regions.

I have been involved in large collaborative projects, such as APAeval, where I was a main co-organizer. The APAeval hackathon was a community-driven effort to evaluate tools related to APA analysis based on conventional RNA-seq data. We built a benchmark suite for the reliable and reproducible assessment of the tool’s performance against ground truth data. Furthermore, I co-developed ZARP, an RNA-seq workflow, which performs the basic steps of an RNA-seq analysis in an automatic and reproducible manner. ZARP follows best programming practices.

Table of contents

Table of figures

1 Introduction

1.1. Cell types and cell states

Each cell in a mammalian organism resides in a particular organ, resp. tissue, and has a function. For example, the human organism consists of over 200 cell types, ranging from stem, immune, bone, muscle to epithelial cells, as shown in the Human Cell Atlas (Regev et al., 2017). And each type comes with a specific morphology, function, and molecular composition. Yet all cells stem from the same zygote and contain the same genetic information stored as DNA. This diversity is reached by the organized and structured way of "reading" the DNA.

1.2. Gene expression

The DNA is organized into chromosomes that carry sequences of bases. Chromosomes contain genes that are transcribed into messenger RNAs (mRNAs) by the RNA polymerase. mRNAs are translated into proteins by the ribosome. The set of all proteins in a cell, called the proteome, carries out the main functions of the cells. Proteins are also post-translationally modified, for example by phosphorylation and ubiquitination. These modifications can activate and/or repress proteins, thereby affecting many cellular processes, such as apoptosis, cell cycle, DNA transcription or immune responses (Blom et al., 2004; Caragea et al., 2007; Haltiwanger & Lowe, 2004; Karve & Cheema, 2011; Mann & Jensen, 2003; Ohtsubo & Marth, 2006; Ramazi & Zahiri, 2021; Strumillo & Beltrao, 2015; M. Wang et al., 2015; Wei et al., 2019; Y. Xu & Chou, 2015).

In the past decades, lots of effort was put into measuring especially the transcriptomes of various cell types, with the idea that the mRNA level reflects the activity and functions of cells. However, the correlation between transcriptome and proteome is low (Dhingra et al., 2005; Rogers et al., 2008). This may have to do with the measurement technologies, but could also be due to the many layers of regulation between mRNA and protein. Interestingly, a new technology has emerged, that of ribosome footprinting (Ingolia, 2010, 2016). It produces more comprehensive data on protein outputs and also with higher correlation with the protein level than the mRNA level measurements (Brar & Weissman, 2015; Eastman et al., 2018; Riba et al., 2019).

1.3. Transcriptome variation in single cells

Up until recently, the transcriptome could only be measured from bulk samples consisting of many cells, which masked the individual types and their functions. With the advent and progression of single-cell transcriptomic technologies (Proserpio, 2019; F. Tang et al., 2009), the gene expression state of individual cells can now be measured, which enables the discovery of new functions and interplay between cell types (Elmentaite et al., 2022; Gulati et al., 2020; Hermann, 2018; Lambrechts et al., 2018; Laughney et al., 2020; Leader, 2021; Lukassen et al., 2018b; The Tabula Muris Consortium et al., 2018, 2020; The Tabula Sapiens Consortium & Quake, 2021; Travaglini et al., 2020).

For example, tumor tissue is highly heterogeneous and consists not only of malignant cells, but also of transformed epithelial cells and invading immune cells, among others. The tissue has its own microenvironment with surrounding blood vessels, signaling molecules and extracellular matrix (Alfarouk et al., 2011; Spill et al., 2016).

In development, the stem cells in the developing organism proliferate and produce cells that differentiate and form the different organs and tissues with their own functions.

For my Master thesis, I mathematically studied the build-up and development of mutations in healthy skin. This tissue is heterogeneous and consists of many different types including stem cells that renew and regenerate the skin. The growth of the tissue is constrained by its architecture. Therefore, the mutations that arise within a stem-cell niche cannot easily outcompete neighboring cells, decreasing the chance of selective sweeps. In contrast to that stands fluid tissue such as the blood, in which mutations conferring growth benefits are more easily and readily reaching saturation (Noble et al., 2021).

1.4. Regulation of gene expression

The production of proteins at the right location at right time is crucial for the functioning of the organism. Thus, the process of protein production starting from the DNA template (known as gene expression) is highly regulated to achieve the fine-tuned protein function and high diversity observed in higher animals. For example, not producing a protein in a tissue can have severe effects and cause diseases such as Albinism (Oetting & Adams, 2018) or Phenylketonuria (Blau, 2016).

The genes in the genome are organized into chromatin, a complex of DNA and histones. The chromatin exists as euchromatin that allows the access of transcription machinery and heterochromatin that is condensed and restricts the access of the transcription machinery. Post-translational modification of histones as well as changes in the DNA such as methylation, lead to chromatin remodeling and thus regulate wide reaching genome regions, such as whole chromosome sections. Individual genes are regulated by the binding of transcription factors to promoters, enhancers and silencers.

1.5. Co- and Post-transcriptional modifications

During and after transcription of a gene into pre-mRNA, multiple processes modify the nascent transcript to produce the mature RNA molecule (Figure 1A). The pre-mRNA is capped at the 5’ end by adding 7-methylguanosines. The spliceosome carries out the splicing and also regulates the splicing pattern (e.g. exon exclusion). At the 3’ end, the pre-mRNA is cleaved by the 3’ end processing complex and then a tail of non-templated adenosines is added by the poly(A) polymerase (Figure 1B). Almost all eukaryotic mRNA and many non-coding RNA, in particular long non-coding RNA, are polyadenylated. Various RBPs attach to the RNA and regulate the fate of the molecule.

The 3’ end processing complex, aka CPA machinery, carrying out the cleavage and polyadenylation consists of four subcomplexes, the cleavage and polyadenylation factor (CPSF), cleavage stimulation factor (CSTF), cleavage factor I and II (CFI and CFII), and other proteins including polyadenylate polymerase (PAP). The CPSF subcomplex recognizes the poly(A) signal, a hexamer with the canonical sequence AAUAAA, and cleaves approximately 21 nucleotides downstream. The poly(A) signal is conserved between mouse and human and is present in the majority of poly(A) sites, though it has a number of functional variants (A. J. Gruber et al., 2016). Downstream of the cleavage site is also a U/GU rich region which is bound by a trimeric CSTF. CFIm 25, part of mammalian CFI (CFIm), binds on the UGUA motif, around 40 nucleotides upstream of the cleavage site. Less is known about the role of the subcomplex CFII.

The PAP adds a poly(A) tail to the 5’ cleavage product (extending the 3’ end of the transcript) and the poly(A) binding protein 1 (PABN1) stops the growth of the poly(A) tail once it reached around 250 nucleotides (Eckmann et al., 2011; Keller et al., 2000; Kühn et al., 2009; Park et al., 2016). The transcript is exported from the nucleus to the cytoplasm by binding the poly(A)-binding protein (PABP). This can also recruit proteins that affect translation, one of which is eukaryotic initiation factor 4G (eIF4G), which is in the eIF4F complex that recruits the 40S ribosomal subunit. The ribosome assembles the 40S and the 60S subunits and starts translation (Gorgoni, 2004). For a schematic description see Figure 1C.
The poly(A) tail is itself regulated and varies in length. During the mRNA’s lifetime, the initially 250bp long poly(A) tail decays. The poly(A) tail is considered a key factor in regulating the stability and translational efficiency of the mRNA (Bilska et al., 2022; Eckmann et al., 2011; Eisen et al., 2022; Legnini et al., 2019; Park et al., 2016; Subtelny et al., 2014). Deadenylation is the first step and typically controls mRNA decay. It is carried out by deadenylases that progressively remove adenosines from the 3’ end (Figure 1C). Once the poly(A) tail is removed, the mRNA can be degraded by the exosome (chapter 23 in (Rorbach & Bobrowicz, 2014)).

While a lot has been unraveled, the regulation of the various co- and post-transcriptional modifications is not understood in complete mechanistic detail. Here I focus on the 3’ end processing complex that carries out the cleavage and polyadenylation of the mRNA molecule.

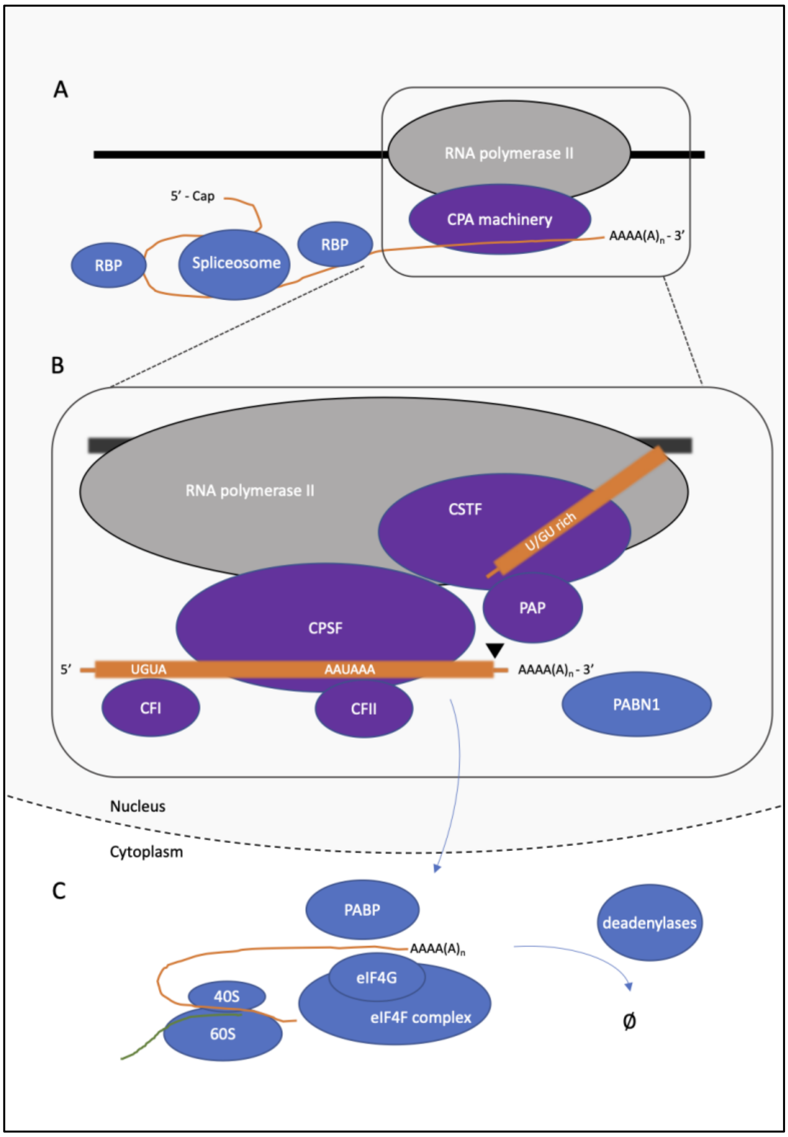


Figure 1: Co- and post-transcriptional modifications of an mRNA molecule.

A) The pre-mRNA (orange) is transcribed from the DNA (black) by the RNA polymerase II. It undergoes multiple maturation processes in the nucleus, including capping at 5' end, intron removal by the spliceosome and cleavage and polyadenylation by the CPA machinery. Various RBPs interact with the nascent RNA molecule. Adapted from (A. J. Gruber & Zavolan, 2019). **B)** The CPA machinery in more detail. The CPSF subcomplex recognizes the polyA(A) signal, AAUAAA, and cleaves 21 nucleotides downstream at the PAS (black triangle). The UGUA motif 40 nucleotides upstream of the PAS is bound by CFI. CSTF binds U/GU-rich region downstream of the PAS. The role of CFII is not yet fully understood. The PAP adds numerous adenosines at the 3' end of the RNA molecule and its length is regulated by PABN1. Adapted from (A. J. Gruber & Zavolan, 2019). **C)** Export of the matured RNA molecule into the cytoplasm. Through recruitment of PABP and the eIF4F complex containing eiF4G, the ribosome (40S and 60S subunits) is recruited and translation into the polypeptide (green) can be started. Translation is countered by mRNA decay. The mRNA decay pathway starts with the poly(A) tail degradation by deadenylases and ends with the mRNA being broken down in the exosome. The translation part is adapted from (Gorgoni, 2004). RBP: RNA-binding protein, CPA: cleavage and polyadenylation, CSTF: cleavage stimulation factor, CFI and II: cleavage factor I and II, PAP: polyadenylate polymerase, PABN1: poly(A) binding protein 1, PABP: poly(A) binding protein (aka PABP1), eIF4G and eiF4F: eukaryotic initiation factor 4G, eiF4F: eukaryotic initiation factor 4F complex, 40S: small subunit and 60S: large subunit or the ribosome.

1.6. Alternative polyadenylation

Most of the human genes have multiple poly(A) sites (PAS) (Derti et al., 2012; Elkon et al., 2013; Shepard et al., 2011; Tian, 2005), and thus, they typically generate multiple isoforms, depending on the cell type and condition. This process is called alternative polyadenylation (APA). The CPA machinery recognizes, dependent on the context, a more proximal or more distal PAS. Each of these PAS is recognized by its own motifs, the poly(A) signal upstream, the U/GU rich region downstream, and probably other signals that are recognized by various RBPs. How exactly the decision for the use of a specific PAS is made is still unclear. The CPA machinery itself as well as other regulators are involved (A. J. Gruber & Zavolan, 2019; Mitschka & Mayr, 2022). The effects of perturbing the expression of CPA machinery factors are known, but the mechanistic details of 3' end processing complex recruitment are unknown.

RBPs can regulate poly(A) site choice, either directly as part of the of the CPA machinery or through binding adjacent regions (Yeo, 2014). The main isoforms resulting from APA are those which are polyadenylated in introns and those that differ in 3'UTR length (Mitschka & Mayr, 2022). Alternative 3'UTR isoforms have been shown to play a role in the regulation of protein abundance, local translation and the formation of protein complexes (Fu et al., 2018; Mayr & Bartel, 2009; Sandberg et al., 2008).

Alternative polyadenylation was studied in multiple human tissues and found to be tissue-specific (Lianoglou et al., 2013; MacDonald & McMahon, 2010; E. T. Wang et al., 2008; H. Zhang et al., 2005). That is, the expression of APA isoforms of a given gene is unique to the tissue.

1.6.1. Types of polyadenylation isoforms

As mentioned above the alternative usage of poly(A) sites leads to transcript isoforms. Alternative PAS can be located on the same 3'UTR, leading to 3'UTR isoforms. It can also be that the PAS is located in an intron, leading to a composite terminal exon, essentially an extended version of an exon. The third type is using an alternative exon by which alternative splicing and polyadenylation give rise to a cassette terminal exon (A. J. Gruber & Zavolan, 2019). Here we will concentrate on 3'UTR isoforms, that only differ in the length of 3'UTRs (Figure 2).

The 3'UTR of an mRNA can contain various cis-regulatory sequence motifs and the long 3'UTR contains more. Most discussed among the regulatory factors are micro-RNAs (miRNAs) that bind to the long 3'UTR isoform to down-regulate the translation and stability of the mRNA. If the short 3'UTR isoform is expressed, the miRNA cannot bind and therefore not regulate the nascent transcript. The miRNA-dependent regulation of 3'UTR isoforms has been implicated in cancers (Mayr & Bartel, 2009; S. Xu et al., 2021; S. Yang et al., 2021).

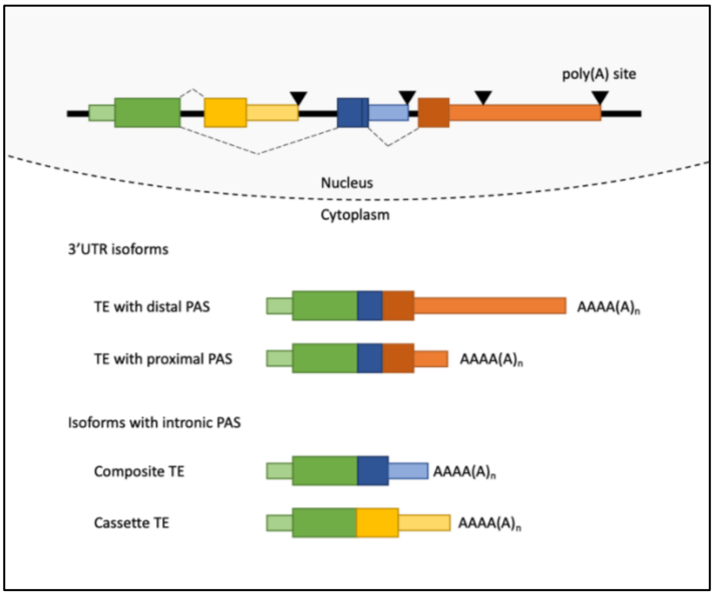


Figure 2: Types of polyadenylation isoforms.

The RNA (black line) can be composed of multiple terminal exons (TEs) (colored bars) and with multiple poly(A) sites (PAS) (triangle). The terminal exons consist of coding (tall) and non-coding (small) sequences. The CPA machinery recognizes, dependent on the context the different poly(A) sites and creates transcript isoforms. Terminal exons that are either processed at the distal or proximal PAS belong to the 3'UTR isoforms. The isoform at the distal PAS has a longer 3'UTR which contains additional cis-regulatory sequence motifs. Composite TEs result from processing of intronic PAS downstream of exons, essentially being extended version of an exon. Cassette TEs follows from the interplay between alternative splicing and polyadenylation. Selection of intronic PAS can lead to isoforms with truncated of the coding sequences. Adapted from (A. J. Gruber & Zavolan, 2019).

1.6.2. Intronic polyadenylation

Selection of an intronic poly(A) site can lead to the truncation of the coding sequences and therefore to a shorter protein product or a protein with a different C-terminal sequence.

(A. J. Gruber, Gypas, et al., 2018) developed the *TECtool* that identifies terminal exons ending at intronic poly(A) sites by combining RNA-seq data with poly(A) site annotations. They showed that such sites have high prevalence in immune and germ cells. Similar results were obtained by (Singh et al., 2018), who analyzed 3'-seq and RNA-seq datasets from human tissues, immune cells, and multiple myeloma samples. They constructed an atlas of intronic polyadenylation events and found that such isoforms are often expressed in immune cells. In contrast, multiple myeloma cells have fewer intronic polyadenylation isoforms.

(Lee et al., 2018) observed intronic polyadenylation in chronic lymphocytic leukemia. The truncated mRNAs lead to truncated proteins that lack tumor-suppressive functions, compared to the protein translated from the full-length mRNA.

(R. Wang et al., 2019) reported that the protein PCF11, a component of the CPA machinery, regulates gene expression via intronic polyadenylation, depending on the length of introns. The downregulation of PCF11 during cell differentiation leads to an upregulation of genes with long introns (which tend to be related to cell morphology, adhesion, and migration).

These studies show that alternative polyadenylation on intronic sites leads to alternative protein products and that it is tightly regulated in physiological conditions and dysregulated in disease conditions.

1.6.3. APA and its consequences in various cell types

Alternative polyadenylation is an integral part of gene expression programs that are associated with physiological changes. During differentiation, the 3'UTRs tend to become longer, increasing the number of cis-regulatory elements that are available for recognition by RBPs. During proliferation, the 3'UTRs tend to become shorter, therefore decreasing the number of cis-regulatory elements (A. J. Gruber & Zavolan, 2019; Sandberg et al., 2008).

(Z. Ji et al., 2009) reported that during mouse embryonic development many genes tend to express mRNAs with longer 3'UTR. Also (Shepard et al., 2011) showed with PAS-seq, a method for the specific capture and sequencing of mRNA 3' ends, that many mRNAs acquire longer 3'UTR during differentiation from embryonic stem cells to neurons.

In contrast, (Sommerkamp et al., 2020) showed that in hematopoietic stem cells a global 3'UTR shortening occurs during differentiation and the transition from quiescent to proliferating cells. The authors used RNA-seq datasets of hematopoietic and progenitor cells (HSPCs) to study APA. While this appears to contradict previous studies that relied on other methods such as EST libraries (Z. Ji et al., 2009) or PAS-seq (Shepard et al., 2011), it is noteworthy that the system involves not only a differentiation process but also an increase in proliferation, thus processes that have been shown to have antagonistic effects on 3'UTR length.

During B cell activation, the membrane-bound IgM is switched to the secreted form. This occurs when the concentration of CSTF2 increases, leading to proximal alternative polyA sites (Edwalds-Gilbert, 1997; Takagaki et al., 1996; Yao et al., 2012).

In T cells, (Chuvpilo et al., 1999) has shown that the transcription factor NF-ATc switches from the distal PAS in naïve T cells to the proximal PAS in T effector cells, but more generally, activation of murine naïve T cells leads to global shortening of 3'UTRs (Sandberg et al., 2008). (Peattie et al., 1994) showed that the FKBP12 gene (the protein binds the immunosuppressants FK506 and rapamycin) encodes three transcripts containing the same open reading frame varying in the 3'UTR. The transcripts are generated by the processing of different splice junctions and multiple poly(A) sites. Upon *in vitro* activation of T cell populations, the transcripts with longer 3'UTRs increase in abundance and/or stability, suggesting that T cell activation requires more FKBP12 protein.

The T cell activation model has been used in many studies of APA, including some that focused on the development of methods to infer poly(A) sites. The data from (Pace et al., 2018) was especially useful in demonstrating that single-cell sequencing data obtained with the 10x Genomics technology reveals poly(A) site with very high resolution. For this reason, this data is described in detail here. (Pace et al., 2018) studied the role of the histone methyltransferase Suv39h1 in silencing gene expression in murine CD8$^+$ T cells. They infected CD8+ T cells with OVA-expressing Listeria monocytogenes (LM-OVA) bacteria and triggered their activation into CD8+ T effector cells. Suv39h1-defective CD8+ T cells were found to survive for longer and have a higher capacity for memory reprogramming. Single-cell RNA sequencing was done to analyze the heterogeneity of wild-type and Suv39h1 knockout LM-OVA infected CD8+ T cells. The authors purified these cells by fluorescence-activated cell sorting (FACS) 7 days after LM-OVA infection and processed them for scRNA-seq. They sequenced around 1000 cells per mouse, 1 mouse for naïve, and 3 for infected cells (2 technical replicates and one biological replicate), in total around 4'000 cells.

Another system broadly used for method development in the APA field is mouse spermatogenesis. During this differentiation process progressive 3'UTR shortening occurs during the maturation of germ cells to sperm (Bao et al., 2016; W. Li, 2016; D. Liu et al., 2007). Spermatogonia stem cells differentiate (mitotic division) into primary and secondary spermatocytes (meiotic division I), which form elongating, condensing and round spermatids (meiosis II), and finally spermatozoa (spermiogenesis). There are a number of scRNA-seq experiments that are useful for the study of APA.

(Lukassen et al., 2018b) sequenced over 2'500 cells from the mouse testis to comprehensively characterize the mouse transcriptome during spermatogenesis. The study specifically described rare cell populations. The authors prepared cell suspensions for two 8-week-old C57BL/6J mice and obtained approximately 1'250 cells for each mouse. The cell type annotation was performed by clustering the cells and checking the expression of over 200 published spermatogenesis stage markers.

(Hermann, 2018) gathered the single cell transcriptome of over 62'000 cells from immature and adult male mice and adult men. Similar to (Lukassen et al., 2018b), cell types and subtypes were identified with known spermatogenic cell type specific marker genes. Besides this, the cell types were also compared against the results of sorted cell types from FACS or gravity sedimentation.

APA is also a prominent mechanism for regulating gene expression in neurons. In these cells, the transcript isoforms using a distal terminal exon or proximal PAS are preferentially localized in the neurites instead of the soma. Such isoforms are induced during differentiation (Taliaferro et al., 2016). (Guvenek & Tian, 2018) showed that neurons have the longest 3'UTRs among the cell types of the brain and that 3'UTRs lengthen during neurogenesis.

1.6.4. APA in cancer

Alternative polyadenylation is also reported in diseases, in particular in the majority of cancers (A. J. Gruber, Schmidt, et al., 2018; Z. Xia et al., 2014). Generally, cancer cells tend to express mRNAs with shortened 3'UTRs, consistent with the reported 3'UTR shortening of proliferative cells. The causes likely are a combination of genetic alterations, global upregulation of 3' end processing factors and other regulators. Genetic alterations can lead to the loss of poly(A) sites which, in turn, leads to reduced 3' end processing and decreases mRNA expression. The gain of poly(A) sites by mutation has the opposite effects. Mutations in genes encoding the CPA machinery are likely causing a global perturbation of the poly(A) site usage. But these two processes do not quantitatively explain the changes in poly(A) site usage (A. J. Gruber & Zavolan, 2019). So, changes in other genes can impact the regulation as well. RNA-binding proteins play a major role both in the formation and the function of mRNAs with alternative 3'UTRs. RBPs involved in splicing are often found to regulate APA too, and since polyadenylation is co-transcriptional, also transcriptional processes influence PAS choice (Tian & Manley, 2017).

In the context of cancer, it is interesting to mention the 3'UTR-dependent regulation of the anti-apoptotic protein CD47. CD47 can translocate from the endoplasmic reticulum to the plasma membrane depending on the 3'UTR isoform from which it is expressed. Binding of the HuR RBP to the long 3'UTR isoform recruits SET to the translation site, enabling the joining of RAC1, which leads to the translocation of CD47 to the plasma membrane. This mechanism, called 3'UTR-dependent protein localization (UDPL) appears to apply to a number of proteins (Berkovits & Mayr, 2015).

Lung cancer appears to be the cancer with the most prevalent and drastic shortening of 3'UTRs (A. J. Gruber, Schmidt, et al., 2018; Mayr & Bartel, 2009; Z. Xia et al., 2014). As single-cell sequencing data sets are generated at an astounding pace, it became possible to study the polyadenylation landscape of cancers in great detail.

(Lambrechts et al., 2018) studied the tumor microenvironment (TME). They obtained the single cell transcriptomes of over 92'000 cells from human lung tumor and matching non-malignant lung samples. They identify 52 stromal cell subtypes with new subpopulations in cell types previously considered homogeneous and validated them on selected markers with immune-histochemistry.

(Laughney et al., 2020) studied the emergence of regenerative cell types in human primary lung adenocarcinomas (a type of non-small cell lung cancer). For this they obtained single-cell transcriptomes of over 40'000 cells from 17 human tissue samples from primary and metastatic lung adenocarcinoma including matching non-malignant tissue samples. They find a high level of immune cell infiltration in their cohort, with cancer cell fraction ranging from 7 to 32% per sample.

To gain more insight, whether only the cancer cells are perturbed and show the association to proliferative cells, or whether perturbations in other cell types can explain the global 3'UTR shortening in transcripts, we make use of the single-cell transcriptomics data of lung cancer with matched normal tissue and develop the method SCUREL to detect 3'UTR changes between two conditions (see section 2).

1.7. Approaches to the inference of APA

The process of cleavage and polyadenylation has been studied for some time, e.g. (Niwa & Berget, 1991; Sheets et al., 1994). A variety of methods have been developed to take advantage of all types of sequence information available at a given time, starting from the rather low coverage and low resolution expressed sequence tag (EST) and microarray data.

The first global studies of APA - ESTs and microarrays EST databases made it for the first time possible to study APA on multiple genes (Elkon et al., 2013). Various groups catalogued APA and performed motif enrichment searches to uncover sequence elements involved in the recognition of poly(A) sites (Ara et al., 2006; Beaudoing, 2000; Legendre, Matthieu et al., 2006; H. Zhang et al., 2005). The first method for global quantification of gene expression was based on cDNA microarrays. In this approach, cDNAs are used for the in vitro transcription of biotin-containing RNAs, which are then hybridized to a chip containing a vast number of gene-specific probes and then quantified. When the probes come from regions of transcripts that identify specific APA isoforms (e.g. a probe comes from the coding region of the mRNA and another from the long form of the 3'UTR), the microarray data can, in principle, reveal the relative usage of proximal and distal poly(A) sites in a 3'UTR.

(Sandberg et al., 2008) took advantage of microarray data to carry out the first global analysis of APA in resting and activated T cells.

(Hu et al., 2014) performed a meta-analysis for APA events from public mouse microarray data. They found that global differential APA affects the biological processes development, differentiation, and immune responses, and observed differential APA in RBP-encoding genes such as Rbm3, Eif4e2 and Elavl1. Since RBPs regulate APA, the authors further analyzed crosslinking and immunoprecipitation (CLIP) data for selected RBPs, concluding that Nova2 represses and Mbnl1 promotes the usage of proximal PAS.

Lembo & Provero in chapter 12 of (Rorbach & Bobrowicz, 2014) describe their computational method to analyze alternative 3’UTR isoforms from Affymetrix 3’ IVT microarray data, which tends to capture the 3’ ends of the probes. They used public data, mainly from Gene Expression Omnibus (Barrett et al., 2012), to study retrospectively APA in cancer.

Thus, microarray-based analysis provided the first insights into the prevalence of APA as a mechanism of gene regulation, particularly in cancer. However, microarrays are rather limited by their design, that is the choice of probes and therefore which transcripts can be measured and make it very difficult to study anything in APA beyond the switch in usage between two APA isoforms (Z. Ji & Tian, 2009; Sandberg et al., 2008).

1.7.1. Inference of PAS usage with dedicated 3’ end RNA-seq protocols

The prevalence of APA revealed by the above-mentioned studies prompted the development of experimental methods that use the poly(A) tail to capture and enrich mRNAs. In contrast to standard RNA-seq, these methods yield reads from around the poly(A) site and do not have a uniform read distribution along transcripts. Poly(A) sites are then inferred from peaks in the read coverage along the genome. Compared to previous EST libraries or microarray databases, the use of RNA-seq made the approach truly transcriptomic, as each expressed transcript in a sample could be measured without prior selection of genes of interest. For a recent overview of 3’ end tailored methods see Table 1 in (A. J. Gruber & Zavolan, 2019). Multiple groups developed 3’ end sequencing approaches.

Using this kind of data various atlases have been built for mammals including human and mouse (Herrmann et al., 2019; Muller et al., 2014; R. Wang, Nambiar, et al., 2018; You et al., 2015). For example, the polyAsite atlas (Herrmann et al., 2019) contains over 569’000 poly(A) sites for Homo sapiens (GRCh38.96), inferred from 221 samples from ten different protocols with a total of over 1 billion reads. The atlas can be added to the UCSC genome browser or be downloaded as BED formatted file for own use.

These atlases are great for studying APA and a good reference, but the samples used do not cover all organs and tissues in enough depth of the organism in question.

1.7.2. Inference of PAS usage with bulk RNA-seq data

3’ end sequencing is not nearly as commonly used as bulk RNA-seq. Thus, the vast availability of short-read RNA-seq prompted the development of computational methods that can infer poly(A) site usage from bulk RNA-seq data. The principle is that while the read coverage profile along a gene is uniform, drops in the coverage occur when a 3’UTR isoform terminates. Although many tools have been proposed, in our group (A. J. Gruber, Schmidt, et al., 2018) developed PAQR that quantifies PAS usage from RNA-seq data. It uses the read coverage profile to subdivide 3’UTRs, respectively terminal exons which include the 3’UTR, based on known poly(A) sites from the polyAsite atlas (Herrmann et al., 2019). It finds the PAS used in the sample by optimally parsing the coverage of 3’UTRs by reads, given the available poly(A) sites, calculating the mean squared error (MSE) between up- and downstream regions and comparing it against the overall MSE around the candidate PAS. Finally, after the PAS are identified, the normalized expression and relative usage within a terminal exon is calculated. The relative PAS usage of a sample can be summarized by calculating the average terminal exon length over all transcripts by summing the relative frequency and the length of the terminal exon in bases and normalized by the maximum length. This yields a measure in

percent that equals to 0 when the proximal site is exclusively used and 100 when the most distal site is exclusively used.
Various other methods have been developed to identify and quantify PAS and ultimately detect differential APA from RNA-seq data. (Chen et al., 2020) reviewed these methods. They classify them into methods requiring a-priori annotations of poly(A) sites (MISO, Roar, QAPA, PAQR), transcript reconstruction (PASA, Scripture, Cufflinks, 3USS, ExUTR), poly(A)-capped reads (KLEAT, ContextMap2), or based on read coverage profile fluctuations (PHMM, GETUTR, ChangePoint, EBChangePoint, IsoSCM, DaPars, APAtrap, TAPAS). The authors benchmark the methods on RNA-seq and real PAS data sets from human, mouse, and Arabidopsis and on simulated data. They match the predicted PAS to real ones with some flexibility by allowing a particular distance (e.g. 50bp) to an annotated PAS. If the prediction is within such a distance, it is considered a true positive, else a false positive. They calculate various performance metrics, including sensitivity, precision, and Receiver Operating Characteristic curves. They conclude that TAPAS has generally the best PAS prediction performance, although it overestimates the number of APA sites and the genes with differential APA. They also note that the overall prediction of all methods studied is only mediocre and the overlap between methods is small.

(Shah et al., 2021) benchmarked TAPAS, QAPA, DaPars2, GETUTR and APAtrap against 3'-Seq, a 3' end-based RNA-seq protocol, and Iso-Seq, a single-molecule full-length RNA-seq method. They first showed that all methods can define poly(A) sites with some reliability, like having the poly(A) signal in their vicinity or being in an annotated 3'UTR, but that 3'-Seq and Iso-Seq are performing better. Next, they benchmarked the methods using RNA-seq and 3'-Seq against Iso-Seq and found that a maximum of 75% of Iso-Seq PASs can be identified by those methods. Also, the similarity in number and distribution of PAS is bigger between 3'-Seq and Iso-Seq compared to the RNA-seq based methods. The PAS identification and quantification is more variable for the RNA-seq methods. Estimating isoform abundance from RNA-seq is difficult, as only short snippets of transcripts are sampled and alternative transcripts can overlap each-other. These authors suggested that it is not always wise to create specialized datasets for studying APA, also given the plentiful public RNA-seq datasets. Although, combining small, specialized data with large amount of RNA-seq data can be a good balance for the near future.

Benchmarking studies that were done so far provide a good overview of the different methods available, but they did not thoroughly evaluate the scope of the methods for analyzing APA. For example, some methods are designed to identify PAS, others to detect 3'UTR shortening or lengthening, the latter also expressed in different metrics. On the technical side, these studies did not provide an easy integration of additional datasets or computational methods.
We therefore sought to tackle these limitations in a benchmarking effort called *APAeval* that led to the comprehensive evaluation of computational methods that use RNA-seq to study APA, in a reproducible and open-source environment. APAeval also brought together researchers from the experimental and the computational side in a collaborative manner (see section 4).

1.7.3. Inference of PAS usage with single cell sequencing methods

Bulk sequencing methods of course mask cell-type specific APA effects. The tissue normally consists of multiple cell types, such as connective, epithelial, immune, or specialized cells (Regev et al., 2017; The Tabula Muris Consortium et al., 2018). Cell isolation from solid tissue may be challenging. This is not a problem for suspended cells such as those in the blood where the cell types can be separated by various means like centrifugation. In-vitro studies can circumvent this issue as well, as cell lines can be used. However, for clinical studies, only tissue sections can be obtained. For example, lung cancer patients provide their cancerous but also adjacent healthy tissue (Lambrechts et al., 2018; Laughney et al., 2020).

The first method to measure the transcriptome of single cells has already been described a decade ago (F. Tang et al., 2009). There exist various single-cell sequencing (scRNA-seq) methods, each with its own focus and goal. The main purpose is to characterize heterogeneous tissues on the level of individual cells and their transcriptional states. This enables the detection of rare subpopulations, which would be masked by bulk sequencing methods. This is a more unbiased way compared to FACS, which relies on enrichment of cells with surface protein expression. The scRNA-seq methods vary by their throughput, sensitivity, and scalability.

While most studies concentrate on profiling gene expression, the technique specifically captures the 3' end of transcripts. Thus, scRNA-seq can be used to interrogate the cleavage sites and study APA with very high resolution, of individual cell types and, to an increasing extent, of single cells (W. Ye et al., 2022).

The experimental protocol BATSeq by (Velten et al., 2015) is used to quantify various 3'UTR isoforms at single cell resolution. It integrates unique molecular identifiers (UMIs) and a PAS mapping protocol to develop barcoded, 3' specific sequencing method (BATSeq). These authors used BATSeq to sequence and retain 107 mouse embryonic and neural stem cells. With Bayesian modeling they found variability in isoform choice across single cells in consistent populations, and that cell types can be distinguished by their 3'UTR isoform usage (Y. Gao & Li, 2021).

Since BATSeq is tailored to study APA in single cells, it is very specialized and only few public datasets are available. The more general scRNA-seq methods focus on gene expression, which is a more common field of study, and are therefore more readily available. This trend is similar to bulk RNA-seq and dedicated 3' end protocols.

1.7.3.1. Droplet-based methods

Droplet-based scRNA-seq methods such as 10x Genomics and Drop-seq are among the most popular choices in large scale studies, as they offer simultaneous measurement of thousands of cells and are therefore considered high-throughput. They are in general more cost-effective and cut the library preparation cost to one tenth compared to FACS approaches (like CEL-Seq2 or Smart-seq2). The technical variation with 10x Genomics has decreased compared to bulk 3' end sequencing. This enabled the study of APA at cell type resolution (Mitschka & Mayr, 2022). However, as generally only the 3'-end of the transcript is measured, information such as internal exon isoform is lost. This additional information can be obtained by full-length transcript approaches such as Smart-seq2 (Proserpio, 2019).

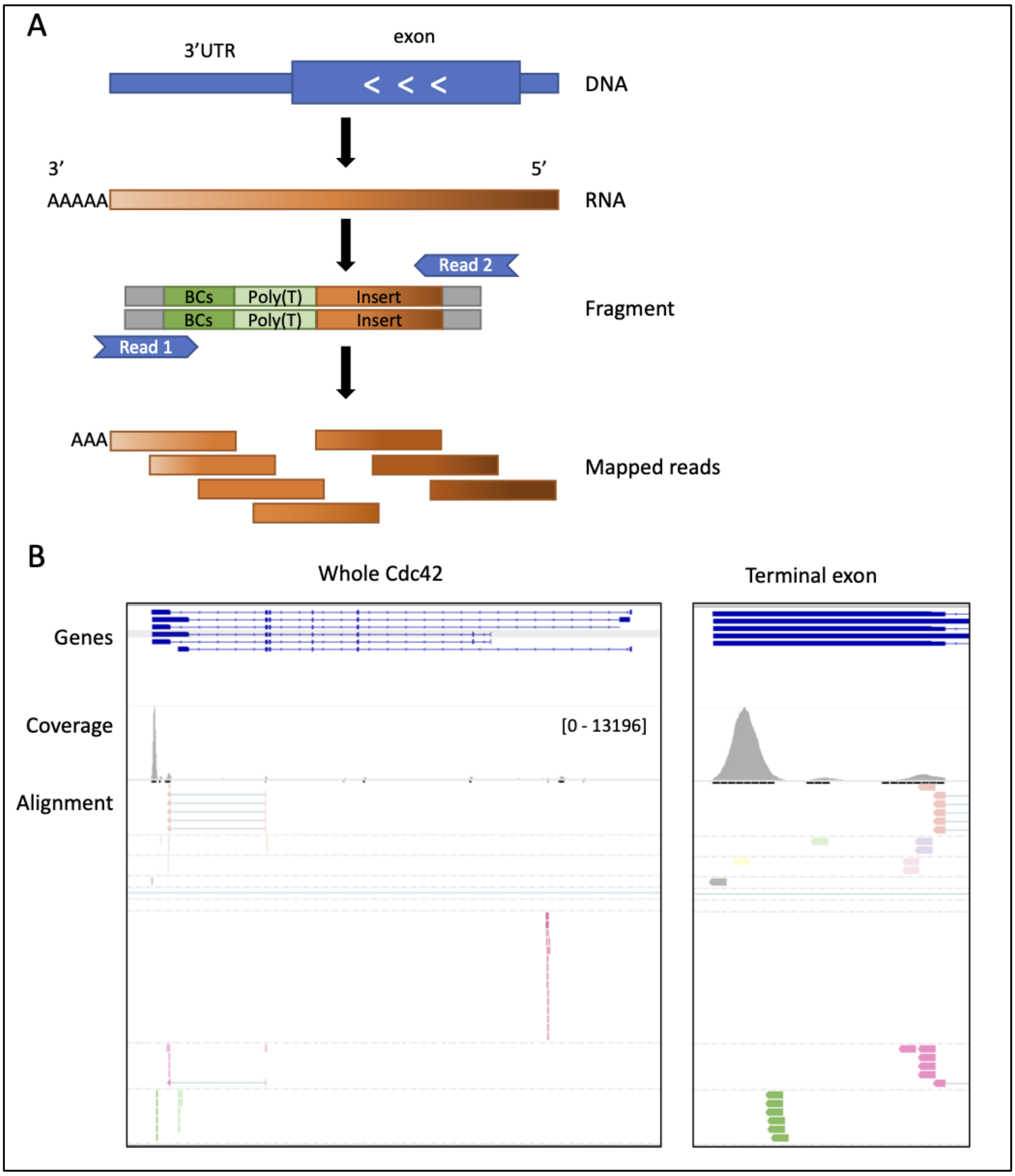


Figure 3: Library generation with Chromium Single Cell 3' v2.

A) Schematic of read generation with 10x Chromium Single Cell 3’ v2 library. The genomic sequence (blue) is transcribed into RNA (shaded orange). The library preparation involves reverse transcription, template switching, enzymatic fragmentation and two rounds of PCR amplification steps. This yields a double stranded fragment with cell and unique molecule identifiers (BCs - green), poly(T) inclusion (light green) and the insert of the RNA molecule. The standard stranded sequencing of read 1 gives the barcodes and read 2 the insert. Mapping of read 2 from the same molecule yields a slightly different position because of the fragmentation (different shading). Some read insertions were short enough such that the poly(T) section of the fragment was sequenced, which produces non-templated adenosines. **B)** Example of a raw coverage profile obtained by CellRanger, from a 10x library. The reads are heavily enriched around the 3’ end of the transcript. The whole Cdc42 region is shown on the left and the terminal exon on the right. The alignments are sorted by cell barcode and colored by the UMI barcode. The PCR duplicated reads are visible as reads belonging to the same cell and having the same UMI. Data from (Pace et al., 2018).

The main principle of droplet-based methods is to capture single cells in droplets, where they are prepared with reagents for sequencing. Afterwards, the cells are pooled together and sequenced as in conventional short-read RNA-seq.

The most prominent and widely used droplet-based scRNA-seq protocol is from Chromium 10x Genomics, a public company designing and manufacturing sequencing technologies for research. The library generation of Chromium Single Cell 3' v2 libraries is depicted in Figure 3A. In short, the cell within a droplet (termed Gel Bead in Emulsion, GEM) is lysed and together with the other components (oligos and Master Mix) full-length, barcoded cDNAs are generated. The mRNA transcripts with poly(A) tails are captured by oligo(dT) primers, the reverse transcriptase extends the poly(A) tail and the template switch oligo is added via a template switch reaction to the 5' end of the transcript, yielding a single-stranded, barcoded cDNA molecule. These molecules from single cells are pooled after breaking the GEMs. A bulk PCR-amplification is performed to obtain enough double-stranded cDNAs. Enzymatic fragmentation creates fragmented cDNAs that are size-selected for optimal insert size for library construction. Read 2 is added by adapter ligation, Illumina P5 and P7 sequences and the sample index sequences are added during the sample index PCR. The final library fragments are made up of the P5, P7, read 1 and read 2 sequences, used for Illumina paired-end sequencing (commercial kits, see https://www.illumina.com). Read 1 contains the cell barcode (CB), which is specific for each droplet, and the UMI, which is used to track the transcripts and delineate from PCR duplicates. Read 2 contains the mRNA-derived fragment. The reads can be conveniently mapped on a transcriptome with the 10x Genomics-provided software called *CellRanger* (Zheng et al., 2017). Interestingly, even though this sequencing method is 3' end based, it normally does not capture the cleavage and polyadenylation sites of transcripts. This is because the sequencing is done from the 5' end of terminal fragments, and may not reach into the poly(A) tail. However, it can happen that the fragment contains parts of the polyA tail, being apparent as adenosines at the 3' end not mapped to the genome. An example coverage profile of a gene and its terminal exon is shown in Figure 3B. We take advantage of such reads in the SCINPAS method that we developed to identify experimentally supported PAS from 3' end based scRNA-seq data (section 3).

(Macosko et al., 2015) developed Drop-seq, a droplet-based technique to measure the transcriptome of thousands of cells. A single cell suspension is prepared, each cell is captured into a DNA-barcoded bead with a custom microfluidics device. Once in the droplet, the cells are lysed, and the available mRNA is captured onto a microparticle with oligo(dT) primers. These loaded microparticles are reverse-transcribed with template switching forming beads called STAMPs. These barcoded STAMPs are pooled, amplified with PCR, and sequenced with high-throughput RNA-seq. Like 10x Genomics, paired read 1 contains the barcodes and paired read 2 the transcript sequences (typically 50bp of length). This sequence is aligned to the genome and the gene count matrix can be constructed.

1.7.3.2. FACS-based methods

CEL-Seq (Hashimshony et al., 2012) was the first to use in vitro transcription for linear RNA amplification, in contrast to PCR amplification by other methods. This also eliminates the need for template switching increasing the sensitivity. CEL-Seq uses an oligo(dT) primer with a cell barcode and selects the 3' ends of the transcripts. The library construction is based on bulk RNA-seq. The cells can be pooled early on, and the library is sequenced as paired-end short reads, where read 1 contains the barcodes and read 2 the transcript-derived sequence.

Its successor CEL-Seq2 (Hashimshony et al., 2016) uses UMIs which reduces amplification biases further. CEL-Seq works on manually selected or FACS sorted cells and is a plate-based protocol (typically on 96 or 384 well plates). Automation and microfluidic devices can be combined to increase the throughput. The original protocol was modified into MARS-seq (Jaitin et al., 2014) and APA-seq (Levin et al., 2020). APA-seq uses by default both reads to capture the 3’ ends, read 1 for the exact cleavage site and read 2 for locating the gene. The authors note that APA-seq is in principle applicable to any poly(A) anchored RNA-seq method, including 10x. But this would require that the read 2 containing the barcodes needs to be sequenced into the poly(A) tail and into the 3’UTR, which seems unfeasible.

Detection and separation of cells with FACS (Ibrahim & van den Engh, 2007; Julius et al., 1972) is not always possible. Maybe the sample is too small and not enough material can be obtained. Or the cell types of interest are not known or described in enough detail, like a rare subpopulation, and thus markers for FACS separation are not available. Currently, FACS can simultaneously measure around 10 fluorescence markers (Autissier et al., 2010; Chattopadhyay et al., 2008; Chattopadhyay & Roederer, 2012; Maes et al., 2020; Perfetto et al., 2004), which is a limitation. Furthermore, differentiating cells are often defined by their lack of surface protein expression. Thus, FACS makes sense for settings where the cell type of interest is known and expresses specific surface proteins that can be detected with fluorescence markers.

1.7.3.3. Full-length scRNA-seq

Smart-seq2 (Picelli et al., 2014) is a method that captures full-length transcripts and therefore enables isoform analysis, including APA isoforms. It relies on the SMART technology (switching mechanism at the 5’ end of the RNA transcript) and is based on reverse transcription and template switching. Single cells are lysed, and mRNA transcripts captured with oligo(dT) primers. Reverse transcription is carried out and template switching introduces 2-5 untemplated C nucleotides at the 5’ end. The template switching oligonucleotide (TSO) adds helper oligonucleotides for stable annealing. Then the cDNA strand can be synthesized, which enables the amplification of the entire transcriptome in a single PCR reaction. These full-length fragments need to be fragmented because the method relies on Illumina short-read sequencing. The most popular choice for this is to use the tagmentation reaction, a neologism from Illumina for tagging and fragmentation of the double-stranded DNA with a prokaryotic Tn5 transposase. A second PCR amplification enriches the sequences and adds the Illumina sequences (P5, P7 and sample index) required for the library. The library can be sequenced in single- or paired-end mode (Proserpio, 2019).

(Ziegenhain et al., 2017) evaluated six scRNA-seq methods and found Smart-Seq2 to have the highest sensitivity. Smart-Seq3 increases the sensitivity by detecting more transcripts per cell (Hagemann-Jensen et al., 2020). Smart-Seq3 is 5’ end-based and cannot be used for profiling APA.

Caveats of the method are the comparatively low throughput because the cells can only be pooled after tagmentation, just before library preparation. The cells are collected manually or by FACS and deposited in single tubes or in 96-, 384-well plates. Also, the method is not strand-specific and therefore unique assignment of reads mapping to overlapping genes is impossible. And common to the other methods, Smart-seq2 cannot detect RNAs other than polyadenylated RNAs, like microRNAs, non-polyadenylated long non-coding RNA, or PIWI-interacting RNAs.

1.7.3.4. Microarray-based methods

(Han et al., 2018) developed Microwell-seq, a high-throughput and low-cost scRNA-seq method. The cells are loaded onto an agarose microarray and the mRNAs captured with magnetic beads. Each bead contains 10^7 to 10^8 oligonucleotides, consisting of a cell barcode, a UMI, a poly(T) tail, and a primer sequence. The cells are lysed and beads with mRNAs are retrieved with a magnet. The beads are collected in tubes to perform reverse transcription and template switching using the Smart-seq2 protocol (Picelli et al., 2014). The amplified cDNAs are fragmented with a customized transposase, during PCR the 3' ends of the transcripts are enriched and sequenced with Illumina Hiseq platform.

These authors used their Microwell-seq to create a mouse cell atlas with all major cell types. The preprint by (Fansler et al., 2021) used this atlas to quantify 3'UTR isoforms with the open-access pipeline *scUTRquant*. They use the fact that Microwell-seq reads cover the transcripts cleavage site and contain poly(A) bases.

1.8. Computational methods for studying APA from single cell transcriptomics

The detection of PAS and the quantification of APA isoform changes from single-cell RNA-seq data is non-trivial due to the complex technical and biological issues associated with the data acquisition.

Various computational methods were developed, dealing with the challenges in various ways. Some rely solely or primarily on the genome annotation, others use scRNA-seq data and PAS databases, some predict PAS *de novo* using the 3' end nature of the scRNA-seq protocols. The computational methods using scRNA-seq data can broadly be categorized into peak calling and density-based methods. (C. Ye, Lin, et al., 2020) note that the peak calling methods have troubles with the sparsity of the 3'end based scRNA-seq datasets and the density-based methods cannot quantify PAS usage. They state a potential way to improve it is to include PAS databases.

1.8.1. Genome sequence-based methods

(Leung et al., 2018) developed a convolutional neural network (CNN) called *Conv-Net* that predicts tissue-specific strength of PAS from genomic sequence alone. The model can discover sequence motifs irrespective of location and prior knowledge. The model can also predict which PAS is more likely to be selected in genes with multiple sites. The *Conv-Net* model is an artificial neural network and requires training for the regression task. The training data consists of the sequence of a pair of PAS from a gene and the regression target is their relative read counts. The authors compare *Conv*-Net with a set of hand-crafted features including the poly(A) signal, called *Feature-Net*. The AUC for PAS selection performance between competing sites across different tissues is pretty much the same between *Feature-Net* and *Conv-Net*, suggesting that *Conv-Net* learns the already known motifs.

(X. Gao et al., 2018) developed *DeepPolyA* that uses a deep CNN to classify plant Arabidopsis thaliana gene sequences into PAS or not PAS, that is into the binary classes positive or negative. For training the CNN model the authors used 13'427 positive sequences and 13'427 negative sequences. The positive sequences were obtained from a dataset and the negative ones by randomly sampling sequences from the Arabidopsis Information Resources database. *DeepPolyA* visualizes the first convolutional layer as sequence logo, and they show that it is able to learn poly(A) signal motifs without prior knowledge. It outperforms other machine learning and deep learning methods on metrics such as area under the receiver-operating

characteristic curve and Matthew's correlation coefficient (or mean square contingency coefficient; measure of association for two binary variables).
(Z. Xia et al., 2019) developed *DeeReCT-polyA*, a CNN for poly(A) signal identification (in contrast to the poly(A) site). The DNA sequence is one-hot encoded, as in the other CNN-based models, for representing the four nucleotides and is classified on whether a poly(A) signal is found. The authors state that the 16 filters of the first layer are basically sequence motif indicators. The training involves discriminating true poly(A) signals from pseudo signals and they visualize the convolutional filters as sequence logos. The performance is evaluated with the error rate, which they define as 1 minus the accuracy. They show that *DeeReCT-polyA* has the lowest average error rate on the Dragon (Kalkatawi et al., 2013) and Omni (Magana-Mora et al., 2017) human poly(A) data compared to three other methods.
(Arefeen et al., 2019) developed *DeepPASTA*, a tool to predict PAS from both sequence and RNA secondary structure data. The RNA secondary structure is predicted by RNAshapes (Steffen et al., 2006). The contribution of the RNA secondary structure on the prediction performance of PAS is not too big in regards of AUC (area under receiver operating characteristic) and AUPRC (area under precision-recall curve) values, but still consistently outperforms a similar model without the secondary structure. The authors extended the method to predict tissue-specific PAS. Also, it can predict the most dominant PAS of a gene in a specific tissue and relative dominance when two PAS of the same gene are given. The training data consists of read counts and is taken from the polyA-Seq data in (Derti et al., 2012), with a similar procedure as in (Leung et al., 2018). This experimental method is used to globally map poly(A) sites in 24 matched tissues in human, mouse and other organisms. These authors also compared their method against (X. Gao et al., 2018; Leung et al., 2018; Z. Xia et al., 2019) and two other methods.
(Bogard et al., 2019) proposed *APARENT*, a model that predicts the proximal-to-distal APA isoform ratio from DNA sequence alone. The model also revealed the cis-regulatory code for APA, visualized as sequence logos, with known motifs but also unknown sequence motifs of 3' end processing. The model was also developed to engineer poly(A) signals computationally and some predictions were validated experimentally.

These methods and models make use of the genomic sequence around the poly(A) site, to predict novel poly(A) sites and in some instances their relative usage within a gene. However, it seems they retrieved mainly the previously known sequence motifs, such as the canonical poly(A) signal. Also, these methods disregard gene expression profiles (W. Ye et al., 2022). As these methods rely on DNA sequence alone, they are not able to infer and study cell-type and tissue-specific effects.

1.8.2. Methods that use the genome annotation

These methods do not identify novel PAS but rather study APA. They make use of the genome annotation, which can be considered previous knowledge, especially in comparison to methods relying on genomic sequence alone.
MAAPER (W. V. Li et al., 2021) uses a likelihood model to predict PAS from 3' end-based reads. First, the method learns the distance of reads to PAS from genes with one PAS only. These genes are obtained from the PAS database PolyA_DB (v3) (R. Wang, Nambiar, et al., 2018) . Second, PAS are predicted and quantified by a likelihood model. The model uses the annotated PAS in genes by the same PAS database and estimates their proportions. They

extend the procedure to select PAS based on the statistical significance such that MAAPER uses as few annotated PAS as possible.

Sierra (Patrick et al., 2020) uses splice-aware peak calling to identify potential PAS, which are used to build an annotated UMI count matrix for each gene. The peak is called by fitting a Gaussian distribution to the read count using NLS (non-linear least squares). A drawback is that it does not infer PAS usage in single samples but can only compare cell types in a pairwise fashion.

scAPA (Shulman & Elkon, 2019) does peak identification with Homer's function findPeaks (Heinz et al., 2010). Peaks overlapping 3'UTRs (from the GENCODE annotation) were used as PAS and reads within the peaks counted. A Gaussian finite mixture model was used to split nearby peaks. This was the first method to explore APA on single cell resolution by calculating the mean proximal peak usage index. The authors found a strong correlation of APA status with cell type, though an open question is whether this is to some extent due to the sparsity of coverage in scRNA-seq. These results provided a proof-of-principle that scRNA-seq can be used for analysis of APA regulation.

scMAPA (Y. Bai et al., 2022), single-cell multi-group identification of APA, is a computational change-point algorithm and a statistical model. It uses 3'UTR annotations to estimate the abundance of long, resp. short 3'UTR isoforms. The method identifies APA genes across multiple cell types with quadratic programming by extending *DaPars* v.2.0 (L. Li et al., 2021). The authors compared their *scMAPA* against *scAPA* and *Sierra*. *scMAPA* is more sensitive and has similar specificity against *scAPA* using simulated data. Performance evaluation with PBMC data (from the 10x Genomics website) on the proportion of annotated PAS from polyA site atlas (Herrmann et al., 2019) showed that *scMAPA* outperforms *scAPA* and *Sierra*. *scDAPA* was not used, as no PAS or similar are returned.

scDAPA (C. Ye, Zhou, et al., 2020) is a histogram-based method to detect APA. The 3' ends of reads in gene regions are divided into distinct bins of equal width. A site distribution difference (SDD) index is calculated to quantify the APA difference between two conditions. Significance is assessed by Wilcoxon rank-sum test, adjusted for multiple testing (BH) and genes with SDD > cut-off and p-value < cut-off are considered as APA genes. This tool can only compare cell types in a pairwise fashion.

scDaPars (Y. Gao et al., 2021) is an application of *DaPars* (Z. Xia et al., 2014) to single cells to identify and quantify APA events. For each gene and cell, the APA usage is measured by the Percentage of Distal poly(A) site Usage Index. The higher the index, the longer the 3'UTR. This index can only be estimated for genes with enough coverage. To recover more genes and indices, the authors impute indices by a non-negative least square regression model on neighboring cells based on the APA profile. The authors applied *scDaPars* on primary breast cancer and endoderm differentiation datasets and were able to characterize APA variations and cell subpopulations in single cells.

1.8.3. De novo identification of polyA sites

SAPAS (Y. Yang et al., 2021) is a method for identifying poly(A) sites from 3' tag-based scRNA-seq based on poly(A)-containing reads. The trimmed poly(A) and the other non-poly(A) reads are mapped with HISAT2 to the ENSEMBL annotation. Poly(A)-containing reads are filtered for internal priming (more than five consecutive adenosines 20 base-pairs downstream of reads 3' end). The 3' ends of the reads falling into 3'UTR regions of GENCODE annotation are clustered (20 bp distance) and filtered based on expression. The peak/mode of the cluster region is assigned as poly(A) site. This means that *SAPAS* does *de novo* PAS identification, but only in known 3'UTR regions, and thus no novel PAS is reported in intergenic (also within 1kb of 3'UTR regions), exonic or intronic regions.

SCAPE (R. Zhou et al., 2022) is a probabilistic mixture model for identification and quantification of poly(A) sites in single cells by utilizing insert (read: fragment) size information. The data is modelled as a mixture of K isoform components and one noise component. Each component has three parameters: mean, standard deviation and weight. For paired-end sequencing, the insert size can be estimated. The poly(A) length distribution can be estimated from poly(A) site covering reads. The number of components is selected using the Bayesian Information Criterion. Expectation-Maximization is used to infer the parameters. The authors benchmarked *SCAPE* on simulated data for precision and recall against *Sierra*, *scAPAtrap*, *scAPA*, *SCAPTURE* and *MAAPER*. *SCAPE* had the highest recall and 2nd highest precision and in total the highest F-score.

scPolyA-pipe (J. Wang et al., 2022) uses scPolyA-seq, which is based on Smart-seq2, so full-length transcript reads are used. It uses reads with poly(A) tails (consecutive A's within or at the end of the read) for the identification of poly(A) sites. Only these reads are mapped and each read's 3' end considered as poly(A) site position by merging nearby sites. Internal priming events are defined as more than five consecutive adenosines or more than 14 adenosines in the 20 bp downstream region of the PAS.

scAPAtrap (Wu et al., 2020) performs detection poly(A) sites based on peaks, searched in the genome. An additional module (findTails) identifies poly(A) sites based on A/T stretches (start or end of the read unaligned to the reference genome). The authors also reason that poly(A)-containing reads are not stemming from internal priming and therefore the anchoring step would remove possible such events.

scUTRquant (Fansler et al., 2021) is a method for single cell 3'UTR isoform quantification, which uses Microwell-seq to construct an atlas of poly(A) sites. Cleavage sites were directly read out from the read alignments to genome. Those that overlapped with annotated genes or with cleavage sites from the polyA site atlas (Herrmann et al., 2019) were kept. The remaining cleavage sites were filtered for possible internal priming events and whether they mapped within 5'000 bp downstream of known transcripts sites. This approach led to the addition of approximately 10'000 cleavage sites to the GENCODE annotation, which were used for APA quantification. However, the procedure does not allow to include novel intronic or intergenic cleavage sites.

SCAPTURE (G.-W. Li et al., 2021) uses the 3' end-based nature of 10x Genomics. PAS are identified by peak calling using Homer (Heinz et al., 2010). High confidence PAS are obtained by a deep learning method which classifies sequences into PAS or no-PAS based on the genomic region around the putative PAS in question. The DL method (DeepPASS) was trained on PAS databases.

1.9. Internal priming

An important challenge in calling poly(A) sites coming from oligo(dT)-based methods is that the primer can align to internal A-rich region of a transcript. This is especially a problem in scRNA-seq, because many of the transcripts are pre-mRNAs, containing still unprocessed introns. For the annotation of poly(A) sites this is of particular interest as internal priming leads to a region as falsely marked as a poly(A) site. Most computational methods using scRNA-seq data deal in one way or another with possible internal priming artefacts. In most cases, these artefacts are identified based on a number of adenosines downstream of the possible cleavage sites (Agarwal et al., 2021; Y. Bai et al., 2022; Shulman & Elkon, 2019; J.

Wang et al., 2022). Some methods check the downstream region instead (G.-W. Li et al., 2021; Y. Yang et al., 2021). In early EST-based studies, the frequency of internal priming during the reverse transcription step was estimated to be up to 12% (Nam et al., 2002).

Methods based on bulk RNA-seq that rely on oligo(dT) primers deal in one way or another with internal priming. Putative internal priming events are removed by considering the reads, their mapping positions, and the genomic sequence (G. Ji et al., 2015).

For the construction of the quantitative atlas of poly(A) sites, (Derti et al., 2012) dealt with internal priming by considering the genomic sequence downstream of the reads. They built an empirical model to give a probability of any given site being a true polyadenylation site.

(Sheppard et al., 2013) used a naïve Bayes classifier to identify internal priming events. The classifier was trained on RNA-seq and PAS-Seq data with true positive and true negative poly(A) sites. Their naïve Bayes classifier outperformed the heuristic filters, such as 8 As downstream within 10 nucleotides window and 8 adenosines downstream plus no poly(A) signal. In this way they were able to identify novel poly(A) sites.

(L. Wang et al., 2013) used a combination of motifs upstream and downstream from the defined cleavage site and assigned 3Seq read peaks into four classes. The peaks within the class without a canonical poly(A) signal but with an A-rich region mostly stemmed from internal priming. They validated the results by demonstrating that the expected position-dependent nucleotide bias and PAS-associated sequence motifs in their vicinity. The study showed that real poly(A) sites could be reasonably well distinguished from sequencing artefacts.

The deep learning neural network *DeepPASS* (G.-W. Li et al., 2021) successfully identifies internal priming events. The classification model was trained on sequences around poly(A) sites and assigns probabilities of each peak being a poly(A) site.

(Svoboda et al., 2022) dedicatedly developed *polyAfilter*, an algorithm to filter out internal alignments based on stretches of adenosines in genomic sites. The tool can be applied to both single-cell and bulk RNA-seq data.

There are also methods that read out cleavage sites directly from reads that still contain poly(A) tails (Wu et al., 2020). These tails do not map to the genome, so they appear as soft-masked regions of the reads flagged by some of the mapping programs.

We deal with internal priming in SCINPAS (section 3) by considering poly(A) containing reads only. Reads that have soft-clipped (i.e. not mapped to genome) bases at the 3’ end and whose soft-clipped regions are sufficiently A-rich are called polyA reads. Also, most computational methods focus on predicting PAS in 3’UTRs, but with dedicated 3’ end sequencing protocols intronic and exonic PAS are found (W. Ye et al., 2022). Since in SCINPAS we consider poly(A) containing reads, we can predict PAS outside of 3’UTRs.

2 Development of a computational method to detect changes in 3’UTRs

My first project was stimulated by Mihaela and had the aim to quantify the 3’UTR changes in tumor tissue compared to matched normal tissue on the level of cell types. The motivation thereof was that the tumor tissues were known to exhibit global 3’UTR shortening compared to healthy counterpart. It is also known that tumor tissue is made up of a variety of cell types. Furthermore, previous studies have shown that proliferating cells and activated immune cells (specifically T cells) exhibit widespread 3’UTR shortening. It was unclear, however, whether only perturbations in cancer cells or a combination of cell types can explain the 3’UTR changes in the tumor tissue. With the appearance of single cell transcriptomics, it was now possible to study changes in 3’UTR length on the level of individual cell types.
At that time, no computational method was available to perform this task, so I developed SCUREL, a method to quantify 3’UTR length changes. SCUREL compares two sets of cells in annotated 3’UTRs on the read density and computes a summary statistic. With the use of a statistical background distribution, we are able to detect genes with significant changes in 3'UTR length. We assessed the performance on two datasets with known effects, namely T cell activation and spermatogenesis. By then, a competing method was available. We compared the performance of SCUREL against this competitor and found that SCUREL was more sensitive and detected more genes with 3’UTR length changes. Next, we carried out the analyses from two studies of the same lung cancer type with matching non-tumor tissues. The multi-study approach gave a more robust assessment. We found a myeloid to lymphoid switch in lung tumor, but the RNAs coming from immune cells were not enough to explain the observed 3’UTR shortening pattern in in the pseudo-bulk profile. We analyzed the cancer cells against their putative origin, the alveolar cells and found genes involved in the protein metabolism to be targeted by 3’UTR shortening. The SCUREL analysis on individual cell types, contrasting tumorous and matching non-tumor tissues, revealed conserved targets of 3’UTR shortening. Taken together, we found that most cell types within tumor tissues are involved in 3’UTR shortening, and the proteins targeted are enriched in protein metabolism and organization of subcellular structure.
SCUREL and its results are published at the *RNA* journal (Burri & Zavolan, 2021) and are available in Appendix A. It includes the correction we recently introduced regarding the wording of spermatogenesis. Supplemental table 1 is not provided in the appendix, please refer to the supplementary material of the *RNA* journal publication.

Since the publication SCUREL has been cited by several publications. For example, (W. Ye et al., 2022) mention it in their survey of methods predicting poly(A) sites as a method for APA analysis rather than poly(A) site prediction from scRNA-seq data. (Mitschka & Mayr, 2022) cite it for the global 3’UTR shortening in spermatogenesis.

I was able to apply SCUREL on a large aging mouse bone marrow dataset generated in our group, see the section 2.1. below. This project involved the identification and annotation of major cell types in multiple conditions and the subsequent differential gene expression analysis between those conditions. The SCUREL analysis would reveal the extent and direction of 3’UTR changes between the conditions.

2.1. Application of SCUREL on aging bone marrow in mice

2.1.1. Introduction

Hematopoiesis is the process by which the blood components are formed. Daily, the bone marrow produces numerous cells to maintain steady state levels in the peripheral circulation. A pool of hematopoietic stem cells (HSCs) resides in the bone marrow, around two to five HSCs per 10^5 total bone marrow cells (Geiger et al., 2013). The pool of HSCs is sustained, as they divide asymmetrically, and some daughter cells remain HSCs. The other daughter cells differentiate into common lymphoid or myeloid progenitor cells. The lymphoid progenitors give rise to T and B lymphocytes and natural killer cells. The common myeloid progenitors further differentiate into granulocytes (neutrophils, basophils, and eosinophils), macrophages, thrombocytes, and erythrocytes (red blood cells).

One well-known hallmark of aging in human physiology is cellular senescence, the reduction in the renewal capacity of various systems (Ferrucci et al., 2020; Geiger et al., 2013; Y. Liu et al., 2022). Alternative splicing and polyadenylation have been found to play a role in cellular senescence (Deschênes & Chabot, 2017; H. Li et al., 2017; Shen et al., 2019; L. Wang et al., 2020).

The number of phenotypic HSCs increases with aging, probably because of increased self-renewal activity of aged HSCs. Also, in aged mice, the common myeloid progenitors increase, but the common lymphoid progenitors decrease in numbers, causing a myeloid skewing (Andersson & Florian, 2022; Geiger et al., 2013).

(Sommerkamp et al., 2020) demonstrated that the APA regulator Pabpn1 is required for HSC function and that global 3'UTR shortening during differentiation into effector cells happens. APA regulates an isoform switch in the glutamine metabolism which is necessary for proper HSC self-renewal and stress response.

In mice, various life-span prolonging treatments exist. One of which is caloric restriction (CR), in which the nutrient intake is limited normally to approximately 70% of the normal feeding. CR has systemic effects and impacts many tissues. (Swindell, 2008) find in a comprehensive microarray data study that stress-response pathway is a shared response to CR. CR has been shown to reduce the risk of chronic diseases (Ryu et al., 2022).

CR also affects the hematopoietic system. (M. Bai et al., 2022) find that intermittent CR mice have more reticulocytes (immature red blood cells), and continuous CR mice have more red blood cells and hemoglobin. They conclude that short-term intermittent, but not continuous CR, has a profound effect on hematopoiesis, which can improve a form of acute anemia in mice.

The chemical compound rapamycin has anti-aging effects (J. Li et al., 2014; Y. Zhang et al., 2021). Rapamycin, aka Sirolimus and marketed as *Rapamune* is an immunosuppressant that is delivered after organ transplantations, among other treatments (European Medicines Agency, 2001). Rapamycin acts by inhibiting the kinase mammalian target of rapamycin (mTOR). mTOR exists in two complexes, mTORC1, which main role is to activate protein translation and the subsequent outcome is cell growth, proliferation and building the actin cytoskeleton (Lipton & Sahin, 2014). mTORC2, the second complex, is rapamycin-insensitive, has similar functions to mTORC1, but also promotes activation of insulin receptors and insulin-growth factor 1 receptors (Yin et al., 2016). Both treatments have similar but distinct life-prolonging effects in aging skeletal muscles (Ham et al., 2022).

Taken together, hematopoiesis is changing in aging and some anti-aging treatments have an impact on this system. We are wondering whether the treatments could slow-down or revert some of the aging effects and how, to better understand the molecular underpinnings. Furthermore, APA regulates HSC development and could therefore also be involved in the aging bone marrow.

2.1.2. Data and Methods

To study the effects of the treatments on the aging bone marrow, adult mice (10M) were treated with CR, rapamycin, or kept as control and fed ad-libitum. The four conditions were measured in triplicates, giving a total of 12 mice (Figure 4A). Among other tissues, the bone marrow was collected and the single cell transcriptome of around 80'000 cells was obtained.

The single cells of the scRNA-seq dataset were annotated with the *haemopedia* RNA-seq database (Choi et al., 2019). The database consists of over 120 mouse samples of the bone marrow and each sample is characterized by the cell type, immunophenotype (surface protein markers) and the cell lineage. The cell type is used as label and occurs in two or more samples.

The gene expression profile of an individual cell is compared pairwise against the gene expression profile of each label from the reference dataset with *singleR* (Aran et al., 2019). The classifier is trained on the known labels, detecting marker genes between labels. The Spearman's rank correlation is computed for each cell-label pair and the label with the highest score is assigned. An additional fine-tuning step increases resolution and resolves similarly high scores by re-computing marker genes. Low scores are pruned, leaving some cells unassigned because they do not resemble any label in the reference dataset. The cells are visualized in two dimensions by selecting the 10% most variable genes, performing PCA, selecting the first 8 PCs for dimensionality reduction with UMAP (McInnes et al., 2018) (Figure 4B).

2.1.3. Results

We observed a similar lymphoid to myeloid skewing in age. The fraction of cell lineages changes in age, particular the B, T cell, erythrocyte, macrophage, neutrophil and restricted potential progenitor lineages. The CR treated mice exaggerate the aging effects. Rapamycin treated mice seem to revert this change and look more like young mice again. For example, the fraction of cells from the erythrocyte lineage increases in age, further increase after CR treatment, but revert to the fraction of the young mice in rapamycin treated aged mice (Figure 4C).

We find less cells from spleen and peripheral blood in aged mice. The erythrocyte lineage consists mostly of the reticulocytes, the most differentiated cell type of the lineage. In line with previous findings, the hemoglobin expression increases along the differentiation.

Differential gene expression (DGE) was performed for selected cell types between the conditions. The cells of a type and sample were pooled into a pseudo-bulk sample, summing the counts per gene. DGE was performed with the function *pseudoBulkDGE* from the R package *scran* (Lun et al., 2016) using contrasts. The method wraps *edgeR*'s (Robinson et al., 2010) quasi-likelihood method, meaning that cell types of sample triplicates are compared among the contrasts old versus young, CR vs old and rapamycin vs old.

The results show that differentially expressed genes exist between the conditions for cell types of the erythrocyte, T cell, B cell, neutrophil, and restricted progenitor lineage. The results for two cell types of the erythrocyte lineage are shown in Figure 4D. Generally, the genes are not enriched in specific pathways or GO terms. The gene Uba52 is consistently expressed in young, but not in old or treated mice. It has a role in development. Cdc42 regulates cell cycle and is pretty much exclusively expressed in young mice.

We checked the changes in 3'UTR length with the same cell types and condition comparisons. For this we applied SCUREL (Burri & Zavolan, 2021) on the mapped reads. We find during aging the majority (approximately 80%) of 3'UTR events are lengthening events. For the CR and rapamycin treatments compared against the age mice, we find approximately equal 3'UTR shortening and lengthening events (Figure 4E). Pathway enrichment analysis of affected genes does not yield much, as only few 3'UTRs show significant changes.

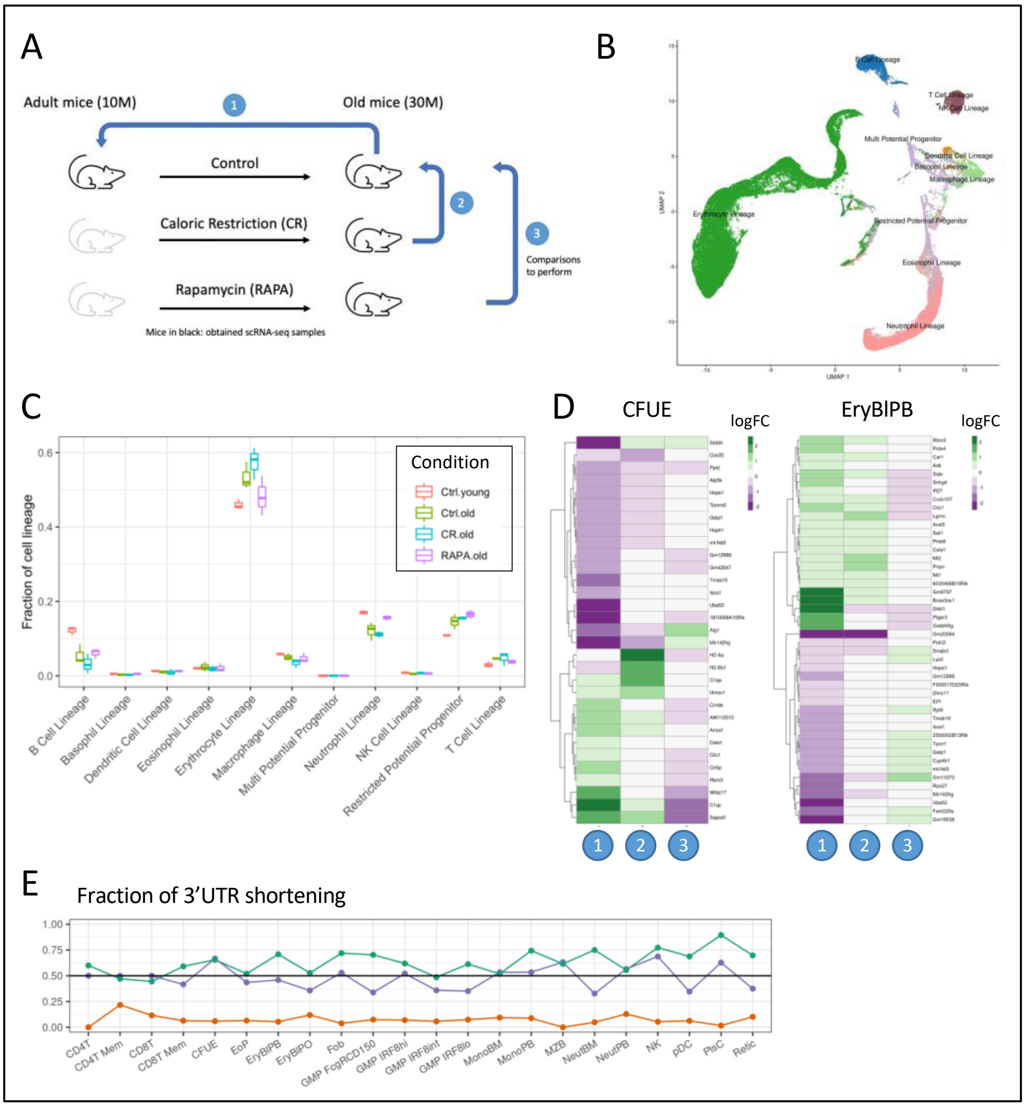


Figure 4: Aging bone marrow analyses.

A) Study set-up of aging bone marrow in mice. The single-cell transcriptome of four conditions in triplicate were measured with 10x Genomics. The comparisons performed are 1: old versus adult, 2: caloric restriction (CR) versus old, 3: Rapamycin (RAPA) versus old. **B)** UMAP representation of the 12 mice. Each point represents an individual cell, 87'704 cells in total. Colored by cell lineages for better visualization. See main text for information on the annotation procedure. **C)** Fraction of cell lineages per sample, separated by condition (color). Boxplots combine the triplicates. **D)** Differential gene expression analysis in subset of erythrocyte lineage. At least one comparison in each row is statistically significant. Each column corresponds to a comparison and is numbered as in A. **E)** SCUREL analysis. Fraction of 3'UTR shortening per cell type in the old versus adult (orange), CR versus old (green) and RAPA versus old (purple) comparisons.

Panels A-D adapted from a Poster I presented during the PhD retreat 2022. Cell types from *haemopedia* mouse RNA-seq database (Choi et al., 2019). CD4/8T: Total CD4/8+ T cell, CFUE: Erythroid Colony Forming Unit, EoP: Eosinophil Progenitor, EryBlPB: Polychromatic erythroblasts (some Basophillic erythroblasts), EryBlPO: Polychromatic erythroblasts (some Orthochromatic erythroblasts), Fob: Follicular B cells, GMP FcgRCD150: FcgammaR+ CD150+ Progenitor, GMP IRF8hi/int: IRF8 High Granulocyte Macrophage Progenitor, GMP_IRF8lo: IRF8 Low Granulocyte Macrophage Progenitor, MonoBM: Total Monocyte Bone Marrow, MonoPB: Total Monocyte Peripheral Blood, MZB: Marginal zone B cells, NeutBM: Bone Marrow Neutrophil, NeutPB: Peripheral Blood Neutrophil, NK: Mature Natural Killer Cell, pDC: Plasmacytoid dendritic cell, PlsC: Bone marrow plasma cells, Retic: Reticulocyte.

2.1.4. Discussion

Taken together, we find that the aging effects are nuanced and systemic, having small changes in many genes. This is not surprising, as aging affecting the whole organism and thus has systemic and far-reaching effects on many processes and genes. APA seems to play a role, but it’s unclear what exactly.

To elucidate further, one could analyze the differential gene expression changes and APA not by the strict and rigid fold change-based methods, but rather with gene set enrichment analysis. The GSEA is better able to find small but consistent effects.

3 Identification of de novo polyadenylation sites with SCINPAS

Mihaela and I experienced during the SCUREL analyses, that the 3'UTR annotations in the genome seemed incomplete. We first thought that we cannot infer the cleavage sites directly from the sequenced reads. But we noticed that reads with non-templated 3' ends were available and wondered whether those could be indicative of a cleavage site. We performed a preliminary check which revealed that a steady small percentage of reads do contain non-templated adenosines. This motivated us to start to investigate this in more detail and identify de novo poly(A) sites directly from the reads.

At that time a Master student joined the group and we started to develop a workflow extracting poly(A) containing reads from publicly available scRNA-seq datasets based on 10x Genomics. We worked on the workflow together and I supervised him throughout and after his Master thesis.

We termed the workflow SCINPAS, an acronym for single cell identification of novel poly(A) sites. SCINPAS takes as input mapped reads and performs a custom deduplication step, favoring distal reads. This increased the fraction of reads with non-templated 3' ends compared to the standard approach. Since the mapping procedure was not always accurate, we included a correction step that mapped-back soft-clipped, i.e. non-templated, nucleotides at the 3'end if they matched the genomic sequence. This increased the mapped part of the read. The cleavage site of such reads was defined as the 3' end-most nucleotide that reached a pre-defined number of mismatches to the genomic sequence. Poly(A) containing reads were identified as reads with soft-clipped 3' ends that were sufficient in length and consisted mostly of adenosines. To decrease variability, we clustered cleavage sites as in (A. J. Gruber et al., 2016) and annotated the most-frequently used genome position as poly(A) site.

We ensured the justifiability of the obtained poly(A) containing reads by comparing them to known and trustworthy poly(A) sites from terminal exons and the polyAsite atlas by (Herrmann et al., 2019). I assessed the performance of SCINPAS to similar methods that also perform poly(A) site identification from scRNA-seq data. Because a method already carried out a benchmark and it favored reasonably well, I evaluated the performance of SCINPAS against this one method. We found that SCINPAS compared well, as it not only identified poly(A) sites in terminal exons but also in intronic and intergenic regions. We further assessed the capability of SCINPAS to identify novel poly(A) sites in genic and non-genic regions by checking the poly(A) signal motif distributions in various annotation classes. Finally, we applied SCINPAS on a large scRNA-seq dataset to demonstrate its generality and usefulness. We measured the reliability of inferred poly(A) sites based on the poly(A) signal motifs. Taken together, SCINPAS is able to identify poly(A) sites from scRNA-seq data, in an autonomous and robust procedure.

The manuscript is published in *NAR Genomics and Bioinformatics* (Moon et al., 2023) and available in Appendix B.

4 Benchmark of computational tools that study polyadenylation

Alternative polyadenylation can be studied with the wealth of data generated by conventional RNA-seq. Various computational tools have been developed, with diverse purposes and assumptions. It was continuously more difficult to keep track of their practicality and limitations. The need for a test suite to benchmark and compare the tools got more urgent. The *APAeval* challenge set out to establish a benchmark for tools for the identification and quantification of poly(A) sites from RNA-seq data. The *APAeval* challenge is conducted by an international community of researchers ranging from RNA biologists to bioinformaticians. It started as an online hackathon, a collaborative event with several people to write computer programs, during the RNA Society meeting in 2021.

We reviewed 17 tools and benchmarked 8 for their performance. The tools participated in the identification and/or absolute quantification and/or relative quantification challenges. For example, some tools would only perform one task such as identification of poly(A) sites, so it would only participate in the identification challenge. We treated 3'-end sequencing data as ground truth and therefore selected a range of such data with matching RNA-seq data. We selected datasets from human and mouse organism as well as synthetic data. We used a range of metrics to assess the performance. We used and developed a suite of workflows, packaged into containers, discrete environments that contain only the operating system and the application to run. The workflows were used for the consistent and reproducible execution of the tools and computation of the metrics. We found varying performance of the tools across factors such as tissue. We believe that our APAeval benchmarking suite is a valuable resource for researchers to pick an appropriate tool for their task at hand. The manuscript is published in the *RNA* journal (Bryce-Smith et al., 2023) and can be found in Appendix C.

I was a main co-organizer of the APAeval challenge during the RNA Society meeting in 2021. This involved the preparation of pilot workflows, that exemplify their usage and that can be used as a template for productive workflows. I structured the work by dividing it into work packages and monitored their timely completion during the hackathon. We quickly realized that the resources would not be enough for a timely finish during the RNA Society meeting. We therefore presented the progress and decided to continue afterwards. During this longer period, I was planning and organizing regular meetings. I was involved in the programming of workflows for the tool executions and their integration into the *OpenEBench* online platform. In the later stages, I was more and more involved in the selection and visualization of identification and absolute quantification metrics, in addition to drafting and writing the manuscript.

5 Development of an automated RNA-seq pipeline

Our group performed RNA-seq experiments on a regular basis. But the analysis was specific to the experimentalist or bioinformatician. We therefore sought to harmonize and automate the initial steps in an RNA-seq analysis. Another main objective was to work collaboratively and practice good programming principles. We developed ZARP, Zavolan-Lab Automated RNA-seq Pipeline, a general-purpose RNA-seq analysis workflow that executes the basic steps of short-read sequencing libraries. ZARP makes use of publicly available tools and packages them into an easy-to-use workflow. The input can either be either the nucleotide sequence in a fastq-formatted file or a public accession number for automated download. The user needs to provide metadata in a sample table and configure the workflow. The workflow can be run locally or on a high-performance cluster. One of the main outputs is an interactive report displaying all available quality metrics of the samples executed. The mapped reads are reported in standard files. Gene and transcript expression levels are also reported. We believe ZARP will help in the autonomous and reproducible analysis of RNA-seq samples, both for experimentalists and bioinformaticians alike. The manuscript is published as pre-print (Katsantoni et al., 2021) and available in Appendix D.

I was mainly involved in the development and software engineering parts of the project. The project had a participatory and democratic spirit, but other group members were in charge and made decisions about the design and direction of the project. I incorporated the annotation of reads into genomic classes. We decided to use an open-source method and I designed tests to ensure the proper execution covering the different library types. I also developed a sub-workflow for the automatic download of fastq files from public accession numbers.
I executed ZARP on several in-house data sets, including one from another research group. This enabled me to give a quick overview of the sample quality and proceed with downstream analyses if needed. A group from another university executed ZARP and reached out to us because they had some difficulties. I gave guidance and helped to troubleshoot their problem. In both settings, this gave me the chance to collect and report on problems and improvements, not only on the technical but also conceptual side.

We continued to improve and extend ZARP by simplifying and reducing the amount of metadata needed for execution. I was mainly involved in the initial set-up of the new command line interface, which is implemented as an object-oriented python class. Furthermore, we tested ZARP on multiple organisms to showcase its strength. The updated version will soon be submitted to a peer-reviewed journal.

6 Current and future prospects

6.1. State of my work

I developed a sensitive computational method to detect 3'UTR changes and named it SCUREL, short for single cell 3' untranslated region lengths. Given that the annotation of PAS in genomes is incomplete, we developed SCUREL to be agnostic to PAS annotation and robust with respect to the sparsity of the scRNA-seq data. I showed that SCUREL works by applying it to ground truth datasets, where the dynamics of PAS usage changes is knowns. SCUREL recapitulates the global 3'UTR shortening upon T cell activation and spermatogenesis. Furthermore, I compared the results against the competitor method available at the time and showed that SCUREL is generally more sensitive, detecting even small changes in PAS usage given limitations in the 3'UTR annotations. Application of SCUREL to single cell sequencing data from lung cancers revealed widespread 3'UTR shortening in cancer cells compared to the non-malignant counterpart, the alveolar epithelial cells. The proteins encoded by the genes that exhibited 3'UTR shortening were enriched in protein metabolism and traffic. I applied SCUREL analysis on all major cell types of the lung tissue to find out whether the global 3'UTR shortening occurs in all cells in the tumor microenvironment. I found that most cell types experience a trend towards 3'UTR shortening in tumor samples. Targets of 3'UTR shortening in T and myeloid cells were enriched in cellular components such as membranes, vesicles, and granules. Lastly, I analyzed the variability in 3'UTR shortening between patients. I found that two patients were highly similar with a clear trend of 3'UTR shortening, whereas the third patient displayed rather 3'UTR lengthening across cell types. Enrichment analyses of targeted genes of 3'UTR shortening strengthened the notion that transport processes are affected. Taken together, SCUREL makes use of widely employed scRNA-seq data and enables the PAS-agnostic analysis of 3'UTR changes between two sets of cells.

Given that the annotation of 3'UTRs and, consequently, PAS are incomplete, we sought to fill this knowledge gap by developing a computational tool to infer PAS from scRNA-seq data. We developed SCINPAS, short for single cell identification of novel poly(A) sites, that extracts reads with non-templated poly(A) tails, which provide experimental support for the poly(A) site. PAS clusters corresponding to individual processing sites and taking into account known imprecision in 3' end processing, were recovered by a modification of the conventional process of de-duplicating reads with the same unique molecular identifier. SCINPAS favors the 3'-most reads, and includes a custom procedure that corrects the read-to-genome alignment to more reliably identify the location of the last mapped nucleotide of a transcript. Closely spaced PAS were gathered into clusters and reported as PAS clusters, with the most frequent position reported as a representative PAS. We demonstrated that most of the retrieved PAS clusters from SCINPAS fall within few nucleotides of annotated terminal exons. Furthermore, we showed that the PAS clusters show the expected enrichment of poly(A) signal variants around 20 nucleotides upstream. We evaluated SCINPAS's ability to identify non-canonical PAS on two systems with such dynamics, namely T cell activation and spermatogenesis. SCINPAS was also able to identify novel PAS in genic and non-genic regions, as the enrichment of poly(A) signals upstream of these PAS demonstrated. Furthermore, application of SCINPAS to a subset of the *Tabula Muris Senis* data (The Tabula Muris Consortium et al., 2020) demonstrated the robustness of our tool to retrieve PAS from various samples. We assessed the performance of SCINPAS against the main current competitor and found that our tool retrieves more de novo PAS in non-canonical locations, such as intronic, exonic and intergenic regions. Moreover, due to its model-free nature, SCINPAS has much higher resolution in pinpointing the location of the 3' end cleavage compared to its competitor. In summary, SCINPAS makes use of the 3' end-biased nature of scRNA-seq data to identify novel poly(A) sites again without relying on genome annotation.

I have also been involved in large collaborative projects, such as APAeval, where I was a main co-organizer. The *APAeval* hackathon was a community-driven effort to evaluate tools related to APA analysis based on RNA-seq data. The hackathon was held online during the RNA Society meeting in 2021, was continued afterwards and culminated in a state-of-the-art benchmarking effort whose results we are about to submit for publication.
The objective was to assess the performance of various open-source computational tools that use the vast collection of conventional RNA-seq datasets data to infer poly(A) sites and/or quantify their usage, and to do this as a collaborative team of researchers from RNA biology, bioinformatics, and software development. The numerous tools were assessed on their assumptions and unique limitations and classified into the tasks of identification, absolute and relative quantification. Some tools could not be integrated in the benchmarking study as they made too specific assumptions, did not generate clearly interpretable metrics or were too brittle to install and run. We developed and used benchmarking workflows to ensure the uniform, comparable, reusable, and reproducible execution of tools and the computation of performance metrics. To aid the analyses, the tools and the benchmarking workflows were packaged into *Docker* containers (Merkel, 2014). The performance metrics were computed against 3' end sequencing data obtained from the same cellular systems as the RNA-seq data, which we treated as ground truth. We selected high quality RNA-seq datasets with matching 3' end sequencing data from human and mouse and simulated data, based on isoform expression distributions and read coverage profiles observed in real data sets. We observed varying performance of the tools across organisms, tissues and dependent on the availability of poly(A) site databases. This benchmarking resource is available for biologists and bioinformaticians alike and will assist to select an appropriate tool, based on the needs of any user's study.

Our group regularly performed RNA-seq experiments, but no standard approach to assess the quality of obtained the sequences existed. The bioinformaticians in our group therefore sought to develop an automated workflow to process RNA-seq data and provide an initial assessment of sample composition. Beyond this, another main purpose was to work collaboratively and to learn best practices in software engineering in a largely remote working environment, during the COVID pandemic.
We developed ZARP (Zavolan-Lab's Automated RNA-seq Pipeline) that uses state-of-the-art bioinformatics tools bundled into a *Snakemake* (Koster & Rahmann, 2012) workflow (see section 5). ZARP handles bulk, stranded, single or paired-end RNA-seq data, can run on a local computer or on a high-performance cluster. A final report provides a detailed summary of the numerous sample and quality statistics. Following best practices, ZARP makes use of continuous integration with the automatic execution of integration tests upon code changes. We demonstrated ZARP's capabilities on several datasets from different organisms. ZARP requires a sample table with metadata such as library type, read orientation, or organism. To further facilitate its use, we developed *zarp-cli*, the command line interface to ZARP. It utilizes the *HTSinfer* package, also developed in our group, to infer the missing metadata. It also

enables the download of samples from the Sequence Read Archive. After the inference, *zarp-cli* executes ZARP automatically. With ZARP, we developed a flexible, versatile and multi-purpose RNA-seq pipeline that will simplify the first steps in RNA-seq analyses.

6.2. Biological implications

With single cell transcriptomic data, I was able to study alternative polyadenylation and 3'UTR isoform expression in more depth than ever before. The link between APA and cell states seems clearer. The regulation of APA and the impact of other regulatory mechanisms are explored in more detail. The progress in spatial transcriptomics will enable an even more fine-grained view of APA.

6.2.1. APA as a manifestation of cell states

In my work, I found that APA is cell type-specific and that changes mirror cell state transitions. That APA is cell type-specific was already known from previous studies with other technologies (e.g. microarray and bulk RNA-seq). However, my SCUREL analysis of single cell RNA-seq data provided a greatly increased resolution, added another viewpoint and reinforced the notion of cell type-specificity of APA. We showed that in human lung cancer APA is widespread, that most major cell types are affected, and we identified differences between the tumors and matched normal tissues. We found common biological processes affected by APA, but also individual cell type-specific effects. Also, we showed that APA is to some extent patient-specific, which could be a reflection of the heterogeneity of the tumor makeup (i.e. the differing driving mutations), even though the tumors were classified as lung adenocarcinomas, the most frequent type of non-small cell lung cancer. Differences in sample collection procedure could also play a role.

Analyses of datasets from different tissues, including T cells from spleen and liver and spermatocytes from testis, revealed and reinforced the tissue-specificity of APA. We observed a clear separation between proximal and distal usage in spermatogenesis, in contrast to a more nuanced and slight changes in T cell activation and in cancer. Also, the set of genes that are affected by APA is different, signifying the tissue-specific nature of APA.

These insights hinge primarily on the cell type annotation, which is constrained by the sparse transcriptomic measurements of the single cell RNA-seq technology. For a detailed discussion, see section 6.3. . In carrying out the analyses, we realized that the annotation of PAS and 3'UTRs is still highly incomplete, as many 3' end clusters mapped to regions that did not have PAS annotations. This is why I have worked along two lines: developing a method to estimate 3'UTR lengths that does not rely on a specific annotation in SCUREL and developing an approach to improve the coverage of PAS annotation in SCINPAS.

6.2.2. Regulation of APA

The 3' end processing complex carries out the pre-mRNA cleavage and polyadenylation, but little is still known about how the choice of poly(A) site is made. Even though APA appears cell type and tissue specific there seems to be common pathways that are affected throughout. In particular, we find in the SCUREL study that in human lung tumors the targets of APA have roles in protein metabolism, secretion and localization. A recent review by (Mitschka & Mayr, 2022) found that APA is a major regulator of mRNA (subcellular) localization, spatial organization of protein synthesis and protein abundance (by their half-time). This suggests that APA regulates specific processes by which it is itself regulated.

Many processes are involved in the regulation of APA. The DNA and RNA sequence can attract and interact with transcription factors or RNA-binding proteins to regulate many steps of RNA processing, including APA. Interestingly, various studies have noted that the distal PAS contain more of the canonical signals for 3’ end processing (A. J. Gruber & Zavolan, 2019; Mitschka & Mayr, 2022; P. Tang et al., 2022; Tian & Manley, 2017). This indicates that weak polyA signals at the proximal sites are overlooked in normal conditions and require specific regulators to promote their usage in other conditions.

Previous studies indicated that APA directly impacts translation and therefore gene expression (Jackson et al., 2010; Sonenberg & Hinnebusch, 2009; Weill et al., 2012). However, recent studies could not establish this link and did not find a correlation between APA and protein abundance (Fansler et al., 2021; A. R. Gruber et al., 2014; Lianoglou et al., 2013; Spies et al., 2013; R. Wang, Zheng, et al., 2018). We investigated this in our SCUREL study as well and could not find a correlation either. This indicates that alternative 3’UTRs have other functions than setting the protein level. The 3’UTR-dependent protein localization described for the CD47 protein represents one such role.

APA is observed in cell proliferation and differentiation, both of which require cell growth and division. These processes are part of the cell cycle, and we found this process to be enriched in erythrocyte progenitors (polychromatic erythroblasts) in the aging mouse bone marrow (see section 2.1.). The cell cycle is tightly controlled at several cell cycle checkpoints. It could be that these checkpoints are the upstream processes that cause the CPA machinery to change in level or localization leading to APA changes.

(Mitra et al., 2018) reported that quiescent mouse fibroblasts express longer transcripts from distal poly(A) sites compared to proliferative fibroblasts. This change is accompanied by a decrease in the level of the cleavage and polyadenylation factor CstF-64. Fibroblasts are often induced to proliferate and migrate, and these data indicate that changes in CstF-64 levels link these two processes.

6.2.3. Impact of other regulatory mechanisms

I concentrated on one type of alternative polyadenylation isoform, those that only differ in the 3’UTR length. But RNA processing can also generate composite or cassette terminal exons, that differ in that they have PAS in intronic regions or stem from alternative exons respectively. The co- and post-transcriptional modifications are regulated and intertwined.

6.2.3.1. Alternative splicing

One type of isoform generated by APA is the cassette terminal exon, which is a product of alternative splicing and polyadenylation. These isoforms contain exons that are spliced out in other transcript isoforms. During RNA maturation, various RNA-binding proteins interact with the spliceosome and the 3’ end processing complex, and they can have activating or repressing functions (A. J. Gruber & Zavolan, 2019). For example, the splicing factor U2AF and the 3’ end processing complex are a component in the designation of the terminal exon (Kyburz et al., 2006; Millevoi et al., 2006; Niwa et al., 1990).

This intertwining means that even when a putative regulator is found, one has to ensure the direct interaction with the 3’ end processing complex, and not indirect via another mechanism, e.g. the splicing machinery.

6.2.3.2. poly(A) tail length

The 3’ end processing complex adds adenosines at the cleaved 3’ end of the transcript. This happens in the nucleus and is performed by canonical poly(A) polymerases. The poly(A) tail is necessary for the translocation of the transcript to the cytoplasm. It is also important for mRNA stability, as deadenylation is the first step in the mRNA decay pathway (Meyer et al., 2004; Parker & Song, 2004). The poly(A) tail can already be shortened in the nucleus (Alles et al., 2023).

Both the UTRs and the poly(A) tail impact the stability and translation rate of the mRNA transcript, while the UTRs also impact the subcellular localization. Disentangling the effects of the two sources, i.e. UTRs and poly(A) tails, might not be an easy task, as the same or similar set of proteins are involved in their regulation.

The complexity and links between RNA processing steps make it challenging to computationally predict regulators of individual steps. More fine-grained measurements of RNA abundance in different cellular compartments are needed to resolve this issue.

6.2.4. Spatial transcriptomics

Spatial transcriptomic methods are able to spatially resolve tissue sections by measuring transcriptomes from spots (Lebrigand et al., 2023; McKellar et al., 2022; Ståhl et al., 2016) or measure subcellular RNA localization by image-based transcriptomics (Eng et al., 2019; C. Xia et al., 2019). A recent review by (Moffitt et al., 2022) illustrates the methodological progress in spatial transcriptomics and the possibility of image-based proteome profiling.

These methods enable the localization of gene expression patterns and computational methods were developed to explore such trends (Edsgärd et al., 2018; S. Sun et al., 2020; Svensson et al., 2018; Zhu et al., 2021). Often, the analysis is restricted to gene expression, but transcript isoforms could also be measured.

A study using long-read single-cell sequencing showed that brain regions display differential isoform expression patterns (Joglekar et al., 2021), highlighting the importance of spatial transcriptomics.

(G. Ji et al., 2023) developed a method to study APA from spatial transcriptomics datasets. The method called *stAPAminer* uses spatially barcoded data and the computational method *scAPAtrap* (Wu et al., 2020) for poly(A) site identification, with the modification of using single spots instead of single cells. Analysis of mouse olfactory bulb data revealed genes with APA enriched in gene ontology terms associated with olfactory bulb development.

Given the progress in spatial transcriptomics, including its overcoming of technical challenges such as high error rates in long-read sequencing and the limited capture of mRNA molecules, the technique will further increase to single cell resolution and the ability of transcript isoform switch analysis. It will likely not take long to study APA in tissue sections in a cell type specific manner.

6.3. Technical challenges

Single cell transcriptomic sequencing led the way for cell type and tissue-specific study of APA. However, the obtained data is sparse and has difficulties in cell type annotation and transcript isoform discovery. This poses challenges that could be overcome with new technical advances.

6.3.1. Dealing with sparsity in scRNA-seq

We are still not at the point being able to capture all mRNAs present in a cell at the time of sequencing. Along with the bursty nature of transcription, this means that even for similar cells a different set of genes is measured which leads to complications comparing them.

(Lähnemann et al., 2020) reports that handling sparsity in single-cell RNA sequencing is performed in two broad approaches, by either modelling the sparsity or by imputation of observed zeros. They note the open problem is the circularity that arises from imputing values from internal information only. This can lead to inflated correlation between genes or cells. The imputation methods need to account for unwanted technical and expected biological variation.

Some methods try to overcome the sparsity and impute expression levels (Linderman et al., 2022; Ran et al., 2020; J. Wang et al., 2019). For example, (J. Wang et al., 2019) developed *SAVER-X*, a deep autoencoder coupled with a Bayesian model, to obtain gene to gene relationships to denoise data sets. As the model is using transfer learning, it can help to automate data integration across studies.

The droplet-based based single-cell RNA-seq methods do not capture the full transcriptome of individual cells. They rely on PCR amplification to obtain more material for sequencing, but there are limits to it. The library preparation and PCR amplification steps are a sampling process in which certain transcripts are under- and others over-represented.

Detection of over-represented transcripts are for sequencing methods using UMIs not a big problem. The UMI denotes the original transcript and can be used to remove duplicate reads. The real downside is that the complexity of the library is reduced because other original transcripts were not sequenced.

The under-representation of transcripts poses a bigger challenge and can ultimately lead to the falsely zero expression of a gene, a so-called drop-out. Other technical factors, such as gene length, GC-content, or sequencing depth, introduce biases in the raw read counts as well. The normalization of the scRNA-seq data takes care of such biases. Various methods were developed to model the gene expression in a more accurate and robust manner to tackle these biases, including drop-out events (Breda et al., 2021; Huang et al., 2018; W. V. Li & Li, 2018; van Dijk et al., 2018). For example, (Breda et al., 2021) developed *Sanity*, a Bayesian normalization method from first principles. It models the mRNA counts as a sampling process with Poisson noise coming on top of the multiplicative noise of the PCR amplification. The authors compare the normalization to a selection of popular methods and find that *Sanity* removes Poisson sampling fluctuations and outperforms in downstream analysis tasks such as clustering.

The sparse gene expression states of single cells make their annotation and detection of functional relations difficult. Modelling gene expression to normalize the transcriptome of a cell is achievable. However, methods incorporating the transcript isoforms, like alternative splicing or alternative polyadenylation isoforms, are still lacking. The study of single-cell APA would greatly benefit from normalization in that the expression levels would become comparable and could be used for downstream differential usage analyses.

6.3.2. Overcoming the problem of cell type annotation with additional protein measurements

Often, the annotation is performed manually with the help of marker genes. These are often surface proteins known from experimental studies. However, these markers are not always expressed, even though the cells are known or thought to express them. For example, CD4+ T cells should express CD4, but the mRNA level may be low or even absent.

The lack of correlation between the gene and surface protein expression could have biological reasons. Maybe the half-life of the mRNA transcript and the protein are so different, that the mRNA is already degraded when the protein is still stably expressed in the cytoplasm or cell membrane.

With the progression of sequencing methods that enable the simultaneous measurement of gene expression and selected surface protein markers, it is now also possible to impute and correlate gene expression with surface markers (Javaid & Frost, 2022; Linderman et al., 2022; van Dijk et al., 2018; Z. Zhou et al., 2020).

For example, (Z. Zhou et al., 2020) developed *cTP-net* that learned the mapping between gene expression and surface protein expression for PBMC data sets (mainly immune cells). They used available multi-omic data from CITE-seq (Stoeckius et al., 2017) and REAP-seq (Peterson et al., 2017) as training data. They apply the trained *cTP-net* (on PBMCs, CBMCs and BMMCs) to perform imputation on *Human Cell Atlas* CBMC and BMMC data sets. They state that labelling subtypes (e.g. CD8 senescent T) is easier compared to pure RNA expression.

The progressive use and availability of such datasets is quite promising, as it gives an additional perspective on the correlation between gene and protein expression. It can also help in the annotation of cells, as the trusted surface protein markers can be used. It also makes the study of APA easier, mostly by making the cell type specific analyses more confident.

6.3.3. Novel transcript isoforms with long-read sequencing

Often, the short read scRNA-seq data is used for studying gene expression patterns. Some efforts have been made for transcript-based analyses (Patrick et al., 2020; Tekath & Dugas, 2021). For example, (Tekath & Dugas, 2021) developed *DTUrtle* which enables differential transcript usage analysis from bulk and single-cell RNA-seq. *DTUrtle* performs a two-stage statistical procedure *stageR* to detect genes with DTU and specific transcripts within those genes.

So far, the PAS could be assigned based on the genome with the additional transcript level annotation. However, it would be interesting to combine the two types of information and perform transcript-based analysis for the identification of PAS and studying APA. It would enable a more-fine grained and nuanced view at the complex process of APA. However, the current methods lack power to perform differential transcript usage, even more so for APA isoforms that increase the number of transcript isoforms again.

Long-read sequencing enables identification of novel isoforms (R. Li et al., 2020; Wright et al., 2022) and study of differential transcript usage, including APA isoforms (Chang et al., 2022).

The combination of long-read and single-cell sequencing seems promising. Considerable progress has been made in the last couple years showing that it is possible (Y. H. Sun et al., 2021; Q. Wang et al., 2021). Commercial sequencing kits also became available, for example PacBio's MAS-seq for 10x Single Cell 3' kit (Al'Khafaji et al., 2021; PacBio, 2021). However, availability of such data is still sparse. Technical challenges, like high error rates and increased sparsity, would need to be overcome or at least dealt by in some way, either by improving the sequencing technologies and/or computational efforts.

6.4. Directions of future development

Beyond the scope of this dissertation await further studies. The most straightforward and natural extension is the application of SCINPAS on the wealth of scRNA-seq data to extend the poly(A) site atlas. With the abundance of sequencing data, it would be interesting to identify additional sequence motifs that regulate PAS choice or to identify regulatory networks from miRNA activity.

6.4.1. Extending the poly(A) site atlas

We made considerable progress in the identification of PAS from scRNA-seq data with the development of SCINPAS. This paves the way to extend the polyAsite atlas (Herrmann et al., 2019) with a new data type. Furthermore, it is possible to increase the PAS expression resolution from sample to tissue and even cell type. This would give a more complete picture of PAS events and would be a great resource to study APA.

In SCINPAS, we identify PAS from scRNA-seq by using the fact that some reads provide direct evidence for the cleavage site, which is apparent from the stretch of non-templated adenosines. This contrasts with the other methods currently used in construction of the polyAsite atlas, which are dedicated 3’ end protocols based on conventional RNA-seq. The polyAsite atlas could be considerably extended with the wealth of scRNA-seq datasets publicly available, such as (Regev et al., 2017; The Tabula Muris Consortium et al., 2018, 2020; Travaglini et al., 2020). Furthermore, since these single cell atlases contain the cell type and tissue information, the polyA site atlas could be extended to make use of it and annotate PAS by the expression in particular cell types and tissues. Previously annotated PAS could be enhanced with expression information across all cell types represented in scRNA-seq data sets. This would give a more comprehensive picture of the PAS and their cell type- and tissue-specific usage. For example, during our analyses on the spermatogenesis dataset (Lukassen et al., 2018a) we discovered that spermatocytes and elongating spermatids exclusively express the distal or proximal PAS. This is in contrast to other tissues and cell types we studied, where the expression of either PAS is partial.

Such a comprehensive polyAsite atlas would make it easier to retrieve and validate new PAS from another study. It would also make it possible to study the regulation of tissue- and cell type-specific APA in a more systematic manner.

6.4.2. Identification of additional sequence motifs regulating PAS choice

The choice of the polyadenylation site might be determined by the combination of the poly(A) signal, DNA elements around the signal and the expression levels of binding factors (Tian, 2005). Therefore, computational methods were developed to identify PAS and sequence motifs regulating PAS choice.

Most, if not all of the machine learning and deep learning-based models concentrated on the hexameric poly(A) signal as the main sequence motif for polyadenylation events. (Xie et al., 2013) developed a model that combined a hidden Markov model with a support vector machine to classify genomic sequences into PAS or not. Later, (Magana-Mora et al., 2017) built on top a method called *Omni-PolyA* that combined decision trees, neural networks and random forests to recognize true PAS. Some methods used convolutional neural networks to predict PAS (Arefeen et al., 2019; Bogard et al., 2019; Z. Xia et al., 2019). The models used about 200 nucleotides of genomic sequence as input, which covers the U-rich region upstream and the U/GU-rich region downstream around the cleavage site (A. J. Gruber & Zavolan, 2019).

It is unknown whether more sequence motifs contribute to the PAS choice, especially for cell type-specific APA. The models could be modified to use longer stretches of genomic sequence as input, to include such putative binding sites or motifs in the analyses. However, an increase of the genomic sequence length would lead to a more complex model and the use of more model parameters (and possibly additional layers in deep learning models), which would require considerably more training data. And it is questionable whether enough annotated data is available. Furthermore, most of those models are sequence position sensitive, meaning that the motif has to occur in the same position along the input sequence. Additionally, these models are trained to classify a genomic sequence into PAS and are not tailored to identify sequence motifs. They are therefore also called black box models. Some of these models however enable the visualization of the learnt weights as sequence logos, e.g. (Bogard et al., 2019; Z. Xia et al., 2019).

Given these limitations, I would find it interesting to use models that incorporate more prior knowledge, like sequential, structural, statistical, or evolutionary properties. And since the methods mentioned above use genomic sequence only, it would be meaningful to incorporate expression data to correlate differential usage with sequence context. For example, the binary classification task could be extended to multi-class classification or even to regression analysis. Moreover, methods from natural language processing could be used to create models that use the genomic sequence as a document and learn the relationship between the individual words (could be the sequence motifs). These approaches could reveal conditional binding and help to elucidate the regulation of PAS choice.

6.4.3. Identification of regulatory networks from miRNA activity

We could use a computational method to infer the transcription factors and miRNAs driving the expression changes between conditions. For example, we could check the changes in cell types in lung cancer compared to matched normal tissue and study whether different or similar transcription factors and/or miRNAs participate across cell types. In a next step we could correlate the miRNA changes with the 3'UTR changes identified by SCUREL or other methods. This would give more insights on how the 3'UTR changes would impact the miRNA-dependent regulation. For example, assuming that miRNA X binds on the 3'UTRs of gene A and gene B, would it be possible, when the 3'UTR of gene A is shortened and the binding site of miRNA X is lost, that the miRNA X would increase or decrease activity towards gene B and lead to more or less repression of B (see Figure 5).

Several methods exist that use scRNA-seq data to predict regulatory networks with transcription factors (Aibar et al., 2017; Behjati Ardakani et al., 2020). But I found no method that attempts to infer regulatory networks from miRNAs.

One method that models the regulatory network by transcription factors and miRNAs is the Integrated System for Motif Activity Response Analysis (ISMARA) (Balwierz et al., 2014). ISMARA identifies key transcription factors and miRNAs spurring the observed expression state changes. In short, ISMARA uses a collection of promoters and precalculated predicted transcription factor binding sites in proximity to promoters and miRNA target sites in annotated 3'UTRs of promoter associated transcripts. For gene expression data, based on the read densities across the transcripts, an expression signal is calculated for each promoter and sample. The linear MARA model is used to explain the signals and to infer the unknown motif activity profiles. The motifs are sorted by significance and a list of predicted target promoters, respectively associated genes, for each motif is provided. ISMARA uses estimates of expression driven by specific promoters to evaluate the effect of transcription factors binding to the promoters. Although miRNAs binding sites are also included in the ISMARA model, the 3'UTR and miRNA binding site annotations are much less refined than the annotation of transcription factor binding site. Thus, this is another direction in which the APA analyses can help. It would be interesting to combine scRNA-seq and the Bayesian modelling of ISMARA to model and infer miRNA activity and to gain more insights into the regulation of APA.

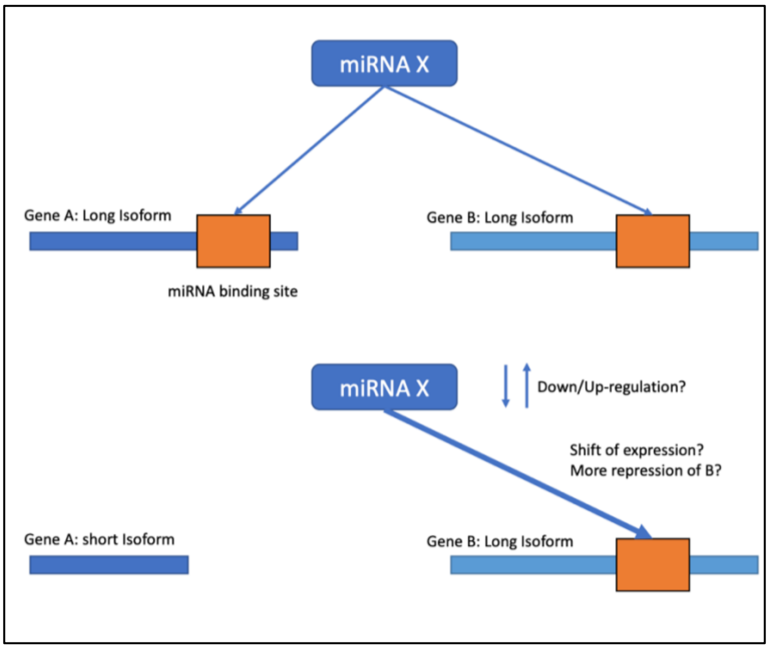


Figure 5: Sketch of miRNA regulation.
Upon alternative polyadenylation and the generation of short 3'UTR isoforms of gene A, the effects of a putative miRNA X, which cannot bind to gene A anymore, are unknown.

7 References

Agarwal, V., Lopez-Darwin, S., Kelley, D. R., & Shendure, J. (2021). The landscape of alternative polyadenylation in single cells of the developing mouse embryo. *Nature Communications*, *12*(1), 5101. https://doi.org/10.1038/s41467-021-25388-8

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., & Aerts, S. (2017). SCENIC: Single-cell regulatory network inference and clustering. *Nature Methods*, *14*(11), 1083–1086. https://doi.org/10.1038/nmeth.4463

Alfarouk, K. O., Muddathir, A. K., & Shayoub, M. E. A. (2011). Tumor Acidity as Evolutionary Spite. *Cancers*, *3*(1), 408–414. https://doi.org/10.3390/cancers3010408

Al'Khafaji, A. M., Smith, J. T., Garimella, K. V., Babadi, M., Sade-Feldman, M., Gatzen, M., Sarkizova, S., Schwartz, M. A., Popic, V., Blaum, E. M., Day, A., Costello, M., Bowers, T., Gabriel, S., Banks, E., Philippakis, A. A., Boland, G. M., Blainey, P. C., & Hacohen, N. (2021). *High-throughput RNA isoform sequencing using programmable cDNA concatenation* [Preprint]. Genetics. https://doi.org/10.1101/2021.10.01.462818

Alles, J., Legnini, I., Pacelli, M., & Rajewsky, N. (2023). Rapid nuclear deadenylation of mammalian messenger RNA. *iScience*, *26*(1), 105878. https://doi.org/10.1016/j.isci.2022.105878

Andersson, R., & Florian, M. C. (2022). Living a longer life: Unique lessons from the naked mole-rat blood system. *The EMBO Journal*. https://doi.org/10.15252/embj.2022111759

Ara, T., Lopez, F., Ritchie, W., Benech, P., & Gautheret, D. (2006). Conservation of alternative polyadenylation patterns in mammalian genes. *BMC Genomics*, *7*(1), 189. https://doi.org/10.1186/1471-2164-7-189

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., Butte, A. J., & Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, *20*(2), 163–172. https://doi.org/10.1038/s41590-018-0276-y

Arefeen, A., Xiao, X., & Jiang, T. (2019). DeepPASTA: Deep neural network based polyadenylation site analysis. *Bioinformatics*, *35*(22), 4577–4585. https://doi.org/10.1093/bioinformatics/btz283

Autissier, P., Soulas, C., Burdo, T. H., & Williams, K. C. (2010). Evaluation of a 12-color flow cytometry panel to study lymphocyte, monocyte, and dendritic cell subsets in humans. *Cytometry Part A*, *9999A*, NA-NA. https://doi.org/10.1002/cyto.a.20859

Bai, M., Cao, P., Lin, Y., Yu, P., Song, S., Chen, L., Wang, L., & Chen, Y. (2022). Intermittent Caloric Restriction Promotes Erythroid Development and Ameliorates Phenylhydrazine-Induced Anemia in Mice. *Frontiers in Nutrition*, *9*, 892435. https://doi.org/10.3389/fnut.2022.892435

Bai, Y., Qin, Y., Fan, Z., Morrison, R. M., Nam, K., Zarour, H. M., Koldamova, R., Padiath, Q. S., Kim, S.,

& Park, H. J. (2022). scMAPA: Identification of cell-type–specific alternative polyadenylation in complex tissues. *GigaScience*, *11*, giac033. https://doi.org/10.1093/gigascience/giac033

Balwierz, P. J., Pachkov, M., Arnold, P., Gruber, A. J., Zavolan, M., & van Nimwegen, E. (2014). ISMARA: Automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Research*, *24*(5), 869–884. https://doi.org/10.1101/gr.169508.113

Bao, J., Vitting-Seerup, K., Waage, J., Tang, C., Ge, Y., Porse, B. T., & Yan, W. (2016). UPF2-Dependent Nonsense-Mediated mRNA Decay Pathway Is Essential for Spermatogenesis by Selectively Eliminating Longer 3'UTR Transcripts. *PLOS Genetics*, *12*(5), e1005863. https://doi.org/10.1371/journal.pgen.1005863

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2012). NCBI GEO: Archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*(D1), D991–D995. https://doi.org/10.1093/nar/gks1193

Beaudoing, E. (2000). Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome Research*, *10*(7), 1001–1010. https://doi.org/10.1101/gr.10.7.1001

Behjati Ardakani, F., Kattler, K., Heinen, T., Schmidt, F., Feuerborn, D., Gasparoni, G., Lepikhov, K., Nell, P., Hengstler, J., Walter, J., & Schulz, M. H. (2020). Prediction of single-cell gene expression for transcription factor analysis. *GigaScience*, *9*(11), giaa113. https://doi.org/10.1093/gigascience/giaa113

Berkovits, B. D., & Mayr, C. (2015). Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature*, *522*(7556), 363–367. https://doi.org/10.1038/nature14321

Bilska, A., Krawczyk, P. S., Dziembowski, A., & Mroczek, S. (2022). Measuring the tail: Methods for poly(A) tail profiling. *WIREs RNA*. https://doi.org/10.1002/wrna.1737

Blau, N. (2016). Genetics of Phenylketonuria: Then and Now. *Human Mutation*, *37*(6), 508–515. https://doi.org/10.1002/humu.22980

Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., & Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *PROTEOMICS*, *4*(6), 1633–1649. https://doi.org/10.1002/pmic.200300771

Bogard, N., Linder, J., Rosenberg, A. B., & Seelig, G. (2019). A Deep Neural Network for Predicting and Engineering Alternative Polyadenylation. *Cell*, *178*(1), 91-106.e23. https://doi.org/10.1016/j.cell.2019.04.046

Brar, G. A., & Weissman, J. S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature Reviews Molecular Cell Biology*, *16*(11), 651–664. https://doi.org/10.1038/nrm4069

Breda, J., Zavolan, M., & van Nimwegen, E. (2021). Bayesian inference of gene expression states from single-cell RNA-seq data. *Nature Biotechnology*. https://doi.org/10.1038/s41587-021-00875-x

Bryce-Smith, S., Burri, D., Gazzara, M. R., Herrmann, C. J., Danecka, W., Fitzsimmons, C. M., Wan, Y. K., Zhuang, F., Fansler, M. M., Fernández, J. M., Ferret, M., Gonzalez-Uriarte, A., Haynes, S., Herdman, C., Kanitz, A., Katsantoni, M., Marini, F., McDonnel, E., Nicolet, B. P., … Zavolan, M. (2023). Extensible benchmarking of methods that identify and quantify polyadenylation sites from RNA-seq data. *RNA*, rna.079849.123. https://doi.org/10.1261/rna.079849.123

Burri, D., & Zavolan, M. (2021). Shortening of 3′ UTRs in most cell types composing tumor tissues implicates alternative polyadenylation in protein metabolism. *RNA*, *27*, 13. https://doi.org/10.1261/rna.078886.121

Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D., & Honavar, V. (2007). Glycosylation site prediction using ensembles of Support Vector Machine classifiers. *BMC Bioinformatics*, *8*(1), 438. https://doi.org/10.1186/1471-2105-8-438

Chang, J. J.-Y., Gleeson, J., Rawlinson, D., De Paoli-Iseppi, R., Zhou, C., Mordant, F. L., Londrigan, S. L., Clark, M. B., Subbarao, K., Stinear, T. P., Coin, L. J. M., & Pitt, M. E. (2022). Long-Read RNA Sequencing Identifies Polyadenylation Elongation and Differential Transcript Usage of Host Transcripts During SARS-CoV-2 In Vitro Infection. *Frontiers in Immunology*, *13*, 832223. https://doi.org/10.3389/fimmu.2022.832223

Chattopadhyay, P. K., Hogerkorp, C.-M., & Roederer, M. (2008). A chromatic explosion: The development and future of multiparameter flow cytometry. *Immunology*, *125*(4), 441–449. https://doi.org/10.1111/j.1365-2567.2008.02989.x

Chattopadhyay, P. K., & Roederer, M. (2012). Cytometry: Today's technology and tomorrow's horizons. *Methods*, *57*(3), 251–258. https://doi.org/10.1016/j.ymeth.2012.02.009

Chen, M., Ji, G., Fu, H., Lin, Q., Ye, C., Ye, W., Su, Y., & Wu, X. (2020). A survey on identification and quantification of alternative polyadenylation sites from RNA-seq data. *Briefings in Bioinformatics*, *21*(4), 1261–1276. https://doi.org/10.1093/bib/bbz068

Choi, J., Baldwin, T. M., Wong, M., Bolden, J. E., Fairfax, K. A., Lucas, E. C., Cole, R., Biben, C., Morgan, C., Ramsay, K. A., Ng, A. P., Kauppi, M., Corcoran, L. M., Shi, W., Wilson, N., Wilson, M. J., Alexander, W. S., Hilton, D. J., & de Graaf, C. A. (2019). Haemopedia RNA-seq: A database of gene expression during haematopoiesis in mice and humans. *Nucleic Acids Research*, *47*(D1), D780–D785. https://doi.org/10.1093/nar/gky1020

Chuvpilo, S., Zimmer, M., Kerstan, A., Glöckner, J., Avots, A., Escher, C., Fischer, C., Inashkina, I., Jankevics, E., Berberich-Siebelt, F., Schmitt, E., & Serfling, E. (1999). Alternative Polyadenylation Events Contribute to the Induction of NF-ATc in Effector T Cells. *Immunity*, *10*(2), 261–269. https://doi.org/10.1016/S1074-7613(00)80026-6

Derti, A., Garrett-Engele, P., MacIsaac, K. D., Stevens, R. C., Sriram, S., Chen, R., Rohl, C. A., Johnson, J. M., & Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. *Genome Research*, *22*(6), 1173–1183. https://doi.org/10.1101/gr.132563.111

Deschênes, M., & Chabot, B. (2017). The emerging role of alternative splicing in senescence and aging. *Aging Cell*, *16*(5), 918–933. https://doi.org/10.1111/acel.12646

Dhingra, V., Gupta, M., Andacht, T., & Fu, Z. F. (2005). New frontiers in proteomics research: A perspective. *International Journal of Pharmaceutics*, *299*(1–2), 1–18. https://doi.org/10.1016/j.ijpharm.2005.04.010

Eastman, G., Smircich, P., & Sotelo-Silveira, J. R. (2018). Following Ribosome Footprints to Understand Translation at a Genome Wide Level. *Computational and Structural Biotechnology Journal*, *16*, 167–176. https://doi.org/10.1016/j.csbj.2018.04.001

Eckmann, C. R., Rammelt, C., & Wahle, E. (2011). Control of poly(A) tail length: Control of poly(A) tail length. *Wiley Interdisciplinary Reviews: RNA*, *2*(3), 348–361. https://doi.org/10.1002/wrna.56

Edsgärd, D., Johnsson, P., & Sandberg, R. (2018). Identification of spatial expression trends in single-cell gene expression data. *Nature Methods*, *15*(5), 339–342. https://doi.org/10.1038/nmeth.4634

Edwalds-Gilbert, G. (1997). Alternative poly(A) site selection in complex transcription units: Means to an end? *Nucleic Acids Research*, *25*(13), 2547–2561. https://doi.org/10.1093/nar/25.13.2547

Eisen, T. J., Li, J. J., & Bartel, D. P. (2022). *The interplay between translational efficiency, poly(A) tails, microRNAs, and neuronal activation*. 25.

Elkon, R., Ugalde, A. P., & Agami, R. (2013). Alternative cleavage and polyadenylation: Extent, regulation and function. *Nature Reviews Genetics*, *14*(7), 496–506. https://doi.org/10.1038/nrg3482

Elmentaite, R., Domínguez Conde, C., Yang, L., & Teichmann, S. A. (2022). Single-cell atlases: Shared and tissue-specific cell types across human organs. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-022-00449-w

Eng, C.-H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-C., & Cai, L. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, *568*(7751), 235–239. https://doi.org/10.1038/s41586-019-1049-y

European Medicines Agency. (2001). *Rapamune medicine at EMA* [Official website of the European Union]. https://www.ema.europa.eu/en/medicines/human/EPAR/rapamune

Fansler, M. M., Zhen, G., & Mayr, C. (2021). *Quantification of alternative 3'UTR isoforms from single cell RNA-seq data with scUTRquant* [Preprint]. Bioinformatics. https://doi.org/10.1101/2021.11.22.469635

Ferrucci, L., Gonzalez-Freire, M., Fabbri, E., Simonsick, E., Tanaka, T., Moore, Z., Salimi, S., Sierra, F., & Cabo, R. (2020). Measuring biological aging in humans: A quest. *Aging Cell*, *19*(2). https://doi.org/10.1111/acel.13080

Fu, Y., Chen, L., Chen, C., Ge, Y., Kang, M., Song, Z., Li, J., Feng, Y., Huo, Z., He, G., Hou, M., Chen, S., & Xu, A. (2018). Crosstalk between alternative polyadenylation and miRNAs in the regulation of protein translational efficiency. *Genome Research*, *28*(11), 1656–1663. https://doi.org/10.1101/gr.231506.117

Gao, X., Zhang, J., Wei, Z., & Hakonarson, H. (2018). DeepPolyA: A Convolutional Neural Network Approach for Polyadenylation Site Prediction. *IEEE Access*, *6*, 24340–24349. https://doi.org/10.1109/ACCESS.2018.2825996

Gao, Y., Li, L., Amos, C. I., & Li, W. (2021). Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression. *Genome Research*, gr.271346.120. https://doi.org/10.1101/gr.271346.120

Gao, Y., & Li, W. (2021). Computational analysis of alternative polyadenylation from standard RNA-seq and single-cell RNA-seq data. In *Methods in Enzymology* (Vol. 655, pp. 225–243). Elsevier. https://doi.org/10.1016/bs.mie.2021.03.015

Geiger, H., de Haan, G., & Florian, M. C. (2013). The ageing haematopoietic stem cell compartment. *Nature Reviews Immunology*, *13*(5), 376–389. https://doi.org/10.1038/nri3433

Gorgoni, B. (2004). The roles of cytoplasmic poly(A)-binding proteins in regulating gene expression: A developmental perspective. *Briefings in Functional Genomics and Proteomics*, *3*(2), 125–141. https://doi.org/10.1093/bfgp/3.2.125

Gruber, A. J., Gypas, F., Riba, A., Schmidt, R., & Zavolan, M. (2018). Terminal exon characterization with TECtool reveals an abundance of cell-specific isoforms. *Nature Methods*, *15*(10), 832–836. https://doi.org/10.1038/s41592-018-0114-z

Gruber, A. J., Schmidt, R., Ghosh, S., Martin, G., Gruber, A. R., van Nimwegen, E., & Zavolan, M. (2018). Discovery of physiological and cancer-related regulators of 3′ UTR processing with KAPAC. *Genome Biology*, *19*(1), 44. https://doi.org/10.1186/s13059-018-1415-3

Gruber, A. J., Schmidt, R., Gruber, A. R., Martin, G., Ghosh, S., Belmadani, M., Keller, W., & Zavolan, M. (2016). 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. https://doi.org/10.1101/gr.202432.115

Gruber, A. J., & Zavolan, M. (2019). Alternative cleavage and polyadenylation in health and disease. *Nature Reviews Genetics*, *20*(10), 599–614. https://doi.org/10.1038/s41576-019-0145-z

Gruber, A. R., Martin, G., Müller, P., Schmidt, A., Gruber, A. J., Gumienny, R., Mittal, N., Jayachandran, R., Pieters, J., Keller, W., van Nimwegen, E., & Zavolan, M. (2014). Global 3′ UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nature Communications*, *5*(1), 5465. https://doi.org/10.1038/ncomms6465

Gulati, G. S., Sikandar, S. S., Wesche, D. J., Manjunath, A., Bharadwaj, A., Berger, M. J., Ilagan, F., Kuo, A. H., Hsieh, R. W., Cai, S., Zabala, M., Scheeren, F. A., Lobo, N. A., Qian, D., Yu, F. B., Dirbas, F. M., Clarke, M. F., & Newman, A. M. (2020). *Single-cell transcriptional diversity is a hallmark of developmental potential*. 8.

Guvenek, A., & Tian, B. (2018). Analysis of alternative cleavage and polyadenylation in mature and differentiating neurons using RNA-seq data. *Quantitative Biology*, *6*(3), 253–266.

https://doi.org/10.1007/s40484-018-0148-3

Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.-J., Larsson, A. J. M., Faridani, O. R., & Sandberg, R. (2020). Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nature Biotechnology*, *38*(6), 708–714. https://doi.org/10.1038/s41587-020-0497-0

Haltiwanger, R. S., & Lowe, J. B. (2004). Role of Glycosylation in Development. *Annual Review of Biochemistry*, *73*(1), 491–537. https://doi.org/10.1146/annurev.biochem.73.011303.074043

Ham, D. J., Börsch, A., Chojnowska, K., Lin, S., Leuchtmann, A. B., Ham, A. S., Thürkauf, M., Delezie, J., Furrer, R., Burri, D., Sinnreich, M., Handschin, C., Tintignac, L. A., Zavolan, M., Mittal, N., & Rüegg, M. A. (2022). Distinct and additive effects of calorie restriction and rapamycin in aging skeletal muscle. *Nature Communications*, *13*(1), 2025. https://doi.org/10.1038/s41467-022-29714-6

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., … Guo, G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, *172*(5), 1091-1107.e17. https://doi.org/10.1016/j.cell.2018.02.001

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K. J., Rozenblatt-Rosen, O., Dor, Y., Regev, A., & Yanai, I. (2016). CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, *17*(1), 77. https://doi.org/10.1186/s13059-016-0938-8

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, *2*(3), 666–673. https://doi.org/10.1016/j.celrep.2012.08.003

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, *38*(4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Hermann, B. P. (2018). *The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids*. 27. https://doi.org/10.1016/j.celrep.2018.10.026

Herrmann, C. J., Schmidt, R., Kanitz, A., Artimo, P., Gruber, A. J., & Zavolan, M. (2019). PolyASite 2.0: A consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Research*, gkz918. https://doi.org/10.1093/nar/gkz918

Hu, W., Liu, Y., & Yan, J. (2014). Microarray Meta-Analysis of RNA-Binding Protein Functions in Alternative Polyadenylation. *PLoS ONE*, *9*(3), e90774. https://doi.org/10.1371/journal.pone.0090774

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., & Zhang, N. R. (2018). SAVER: Gene expression recovery for single-cell RNA sequencing. *Nature Methods*, *15*(7), 539–542. https://doi.org/10.1038/s41592-018-0033-z

Ibrahim, S. F., & van den Engh, G. (2007). Flow Cytometry and Cell Sorting. In A. Kumar, I. Y. Galaev, & B. Mattiasson (Eds.), *Cell Separation* (Vol. 106, pp. 19–39). Springer Berlin Heidelberg. https://doi.org/10.1007/10_2007_073

Ingolia, N. T. (2010). Genome-Wide Translational Profiling by Ribosome Footprinting. In *Methods in Enzymology* (Vol. 470, pp. 119–142). Elsevier. https://doi.org/10.1016/S0076-6879(10)70006-9

Ingolia, N. T. (2016). Ribosome Footprint Profiling of Translation throughout the Genome. *Cell*, *165*(1), 22–33. https://doi.org/10.1016/j.cell.2016.02.066

Jackson, R. J., Hellen, C. U. T., & Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular Cell Biology*, *11*(2), 113–127. https://doi.org/10.1038/nrm2838

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., & Amit, I. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*, *343*(6172), 776–779. https://doi.org/10.1126/science.1247651

Javaid, A., & Frost, H. R. (2022). *SPECK: An Unsupervised Learning Approach for Cell Surface Receptor Abundance Estimation for Single Cell RNA-Sequencing Data* [Preprint]. Bioinformatics. https://doi.org/10.1101/2022.10.08.511197

Ji, G., Guan, J., Zeng, Y., Li, Q. Q., & Wu, X. (2015). Genome-wide identification and predictive modeling of polyadenylation sites in eukaryotes. *Briefings in Bioinformatics*, *16*(2), 304–313. https://doi.org/10.1093/bib/bbu011

Ji, G., Tang, Q., Zhu, S., Zhu, J., Ye, P., Xia, S., & Wu, X. (2023). stAPAminer: Mining Spatial Patterns of Alternative Polyadenylation for Spatially Resolved Transcriptomic Studies. *Genomics, Proteomics & Bioinformatics*, S1672022923000037. https://doi.org/10.1016/j.gpb.2023.01.003

Ji, Z., Lee, J. Y., Pan, Z., Jiang, B., & Tian, B. (2009). Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences*, *106*(17), 7028–7033. https://doi.org/10.1073/pnas.0900028106

Ji, Z., & Tian, B. (2009). Reprogramming of 3′ Untranslated Regions of mRNAs by Alternative Polyadenylation in Generation of Pluripotent Stem Cells from Different Cell Types. *PLoS ONE*, *4*(12), e8419. https://doi.org/10.1371/journal.pone.0008419

Joglekar, A., Prjibelski, A., Mahfouz, A., Collier, P., Lin, S., Schlusche, A. K., Marrocco, J., Williams, S. R., Haase, B., Hayes, A., Chew, J. G., Weisenfeld, N. I., Wong, M. Y., Stein, A. N., Hardwick, S. A., Hunt, T., Wang, Q., Dieterich, C., Bent, Z., … Tilgner, H. U. (2021). A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. *Nature Communications*, *12*(1), 463. https://doi.org/10.1038/s41467-020-20343-5

Julius, M. H., Masuda, T., & Herzenberg, L. A. (1972). Demonstration That Antigen-Binding Cells Are Precursors of Antibody-Producing Cells After Purification with a Fluorescence-Activated Cell

Sorter. *Proceedings of the National Academy of Sciences*, *69*(7), 1934–1938. https://doi.org/10.1073/pnas.69.7.1934

Kalkatawi, M., Rangkuti, F., Schramm, M., Jankovic, B. R., Kamau, A., Chowdhary, R., Archer, J. A. C., & Bajic, V. B. (2013). Dragon PolyA Spotter: Predictor of poly(A) motifs within human genomic DNA sequences. *Bioinformatics*, *29*(11), 1484–1484. https://doi.org/10.1093/bioinformatics/btt161

Karve, T. M., & Cheema, A. K. (2011). Small Changes Huge Impact: The Role of Protein Posttranslational Modifications in Cellular Homeostasis and Disease. *Journal of Amino Acids*, *2011*, 1–13. https://doi.org/10.4061/2011/207691

Katsantoni, M., Gypas, F., Herrmann, C. J., Burri, D., Bak, M., Iborra, P., Agarwal, K., Ataman, M., Börsch, A., Zavolan, M., & Kanitz, A. (2021). *ZARP: An automated workflow for processing of RNA-seq data* [Preprint]. Bioinformatics. https://doi.org/10.1101/2021.11.18.469017

Keller, R. W., Kühn, U., Aragón, M., Bornikova, L., Wahle, E., & Bear, D. G. (2000). The nuclear poly(A) binding protein, PABP2, forms an oligomeric particle covering the length of the poly(A) tail. *Journal of Molecular Biology*, *297*(3), 569–583. https://doi.org/10.1006/jmbi.2000.3572

Koster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kühn, U., Gündel, M., Knoth, A., Kerwitz, Y., Rüdel, S., & Wahle, E. (2009). Poly(A) Tail Length Is Controlled by the Nuclear Poly(A)-binding Protein Regulating the Interaction between Poly(A) Polymerase and the Cleavage and Polyadenylation Specificity Factor. *Journal of Biological Chemistry*, *284*(34), 22803–22814. https://doi.org/10.1074/jbc.M109.018226

Kyburz, A., Friedlein, A., Langen, H., & Keller, W. (2006). Direct Interactions between Subunits of CPSF and the U2 snRNP Contribute to the Coupling of Pre-mRNA 3′ End Processing and Splicing. *Molecular Cell*, *23*(2), 195–205. https://doi.org/10.1016/j.molcel.2006.05.037

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. de, Cappuccio, A., … Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, *21*(1), 31. https://doi.org/10.1186/s13059-020-1926-6

Lambrechts, D., Wauters, E., Boeckx, B., Aibar, S., Nittner, D., Burton, O., Bassez, A., Decaluwé, H., Pircher, A., Van den Eynde, K., Weynand, B., Verbeken, E., De Leyn, P., Liston, A., Vansteenkiste, J., Carmeliet, P., Aerts, S., & Thienpont, B. (2018). Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature Medicine*, *24*(8), 1277–1289. https://doi.org/10.1038/s41591-018-0096-5

Laughney, A. M., Hu, J., Campbell, N. R., Bakhoum, S. F., Setty, M., Lavallée, V.-P., Xie, Y., Masilionis, I., Carr, A. J., Kottapalli, S., Allaj, V., Mattar, M., Rekhtman, N., Xavier, J. B., Mazutis, L., Poirier, J. T., Rudin, C. M., Pe'er, D., & Massagué, J. (2020). Regenerative lineages and immune-mediated pruning in lung cancer metastasis. *Nature Medicine*, *26*(2), 259–269. https://doi.org/10.1038/s41591-019-0750-6

Leader, A. M. (2021). *Single-cell analysis of human non-small cell lung cancer lesions refines tumor classification and patient stratification*. 29. https://doi.org/10.1016/j.ccell.2021.10.009

Lebrigand, K., Bergenstra, J., Meletis, K., Barbry, P., Waldmann, R., & Lundeberg, J. (2023). *The spatial landscape of gene expression isoforms in tissue sections*.

Lee, S.-H., Singh, I., Tisdale, S., Abdel-Wahab, O., Leslie, C. S., & Mayr, C. (2018). Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature*, *561*(7721), 127–131. https://doi.org/10.1038/s41586-018-0465-8

Legendre, Matthieu, Ritchie, William, Lopez, Fabrice, & Gautheret, Daniel. (2006). *Differential Repression of Alternative Transcripts: A Screen for miRNA Targets*. https://doi.org/10.1371/journal.pcbi.0020043

Legnini, I., Alles, J., Karaiskos, N., Ayoub, S., & Rajewsky, N. (2019). FLAM-seq: Full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nature Methods*, *16*(9), 879–886. https://doi.org/10.1038/s41592-019-0503-y

Leung, M. K. K., Delong, A., & Frey, B. J. (2018). Inference of the human polyadenylation code. *Bioinformatics*, *34*(17), 2889–2898. https://doi.org/10.1093/bioinformatics/bty211

Levin, M., Zalts, H., Mostov, N., Hashimshony, T., & Yanai, I. (2020). Gene expression dynamics are a proxy for selective pressures on alternatively polyadenylated isoforms. *Nucleic Acids Research*, *48*(11), 5926–5938. https://doi.org/10.1093/nar/gkaa359

Li, G.-W., Nan, F., Yuan, G.-H., Liu, C.-X., Liu, X., Chen, L.-L., Tian, B., & Yang, L. (2021). SCAPTURE: A deep learning-embedded pipeline that captures polyadenylation information from 3′ tag-based RNA-seq of single cells. *Genome Biology*, *22*(1), 221. https://doi.org/10.1186/s13059-021-02437-5

Li, H., Wang, Z., Ma, T., Wei, G., & Ni, T. (2017). Alternative splicing in aging and age-related diseases. *Translational Medicine of Aging*, *1*, 32–40. https://doi.org/10.1016/j.tma.2017.09.005

Li, J., Kim, S. G., & Blenis, J. (2014). Rapamycin: One Drug, Many Effects. *Cell Metabolism*, *19*(3), 373–379. https://doi.org/10.1016/j.cmet.2014.01.001

Li, L., Huang, K.-L., Gao, Y., Cui, Y., Wang, G., Elrod, N. D., Li, Y., Chen, Y. E., Ji, P., Peng, F., Russell, W. K., Wagner, E. J., & Li, W. (2021). An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nature Genetics*, *53*(7), 994–1005. https://doi.org/10.1038/s41588-021-00864-5

Li, R., Ren, X., Ding, Q., Bi, Y., Xie, D., & Zhao, Z. (2020). Direct full-length RNA sequencing reveals unexpected transcriptome complexity during Caenorhabditis elegans development. *Genome Research*, *30*(2), 287–298. https://doi.org/10.1101/gr.251512.119

Li, W. (2016). *Alternative cleavage and polyadenylation in spermatogenesis connects chromatin regulation with post-transcriptional control*. 17.

Li, W. V., & Li, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature Communications*, *9*(1), 997. https://doi.org/10.1038/s41467-018-03405-7

Li, W. V., Zheng, D., Wang, R., & Tian, B. (2021). MAAPER: Model-based analysis of alternative polyadenylation using 3′ end-linked reads. *Genome Biology*, *22*(1), 222. https://doi.org/10.1186/s13059-021-02429-5

Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S., & Mayr, C. (2013). Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & Development*, *27*(21), 2380–2396. https://doi.org/10.1101/gad.229328.113

Linderman, G. C., Zhao, J., Roulis, M., Bielecki, P., Flavell, R. A., Nadler, B., & Kluger, Y. (2022). Zero-preserving imputation of single-cell RNA-seq data. *Nature Communications*, *13*(1), 192. https://doi.org/10.1038/s41467-021-27729-z

Lipton, J. O., & Sahin, M. (2014). The Neurology of mTOR. *Neuron*, *84*(2), 275–291. https://doi.org/10.1016/j.neuron.2014.09.034

Liu, D., Brockman, J. M., Dass, B., Hutchins, L. N., Singh, P., McCarrey, J. R., MacDonald, C. C., & Graber, J. H. (2007). Systematic variation in mRNA 3′-processing signals during mouse spermatogenesis. *Nucleic Acids Research*, *35*(1), 234–246. https://doi.org/10.1093/nar/gkl919

Liu, Y., Schwam, J., & Chen, Q. (2022). Senescence-Associated Cell Transition and Interaction (SACTAI): A Proposed Mechanism for Tissue Aging, Repair, and Degeneration. *Cells*, *11*(7), 1089. https://doi.org/10.3390/cells11071089

Lukassen, S., Bosch, E., Ekici, A. B., & Winterpacht, A. (2018a). Characterization of germ cell differentiation in the male mouse through single-cell RNA sequencing. *Scientific Reports*, *8*(1), 6521. https://doi.org/10.1038/s41598-018-24725-0

Lukassen, S., Bosch, E., Ekici, A. B., & Winterpacht, A. (2018b). Single-cell RNA sequencing of adult mouse testes. *Scientific Data*, *5*(1), 180192. https://doi.org/10.1038/sdata.2018.192

Lun, A. T. L., McCarthy, D. J., & Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data. *F1000Research*, *5*, 2122. https://doi.org/10.12688/f1000research.9501.1

MacDonald, C. C., & McMahon, K. W. (2010). Tissue-specific mechanisms of alternative polyadenylation: Testis, brain, and beyond: Tissue-specific mechanisms of alternative polyadenylation. *Wiley Interdisciplinary Reviews: RNA*, *1*(3), 494–501. https://doi.org/10.1002/wrna.29

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Maes, E., Cools, N., Willems, H., & Baggerman, G. (2020). FACS-Based Proteomics Enables Profiling of Proteins in Rare Cell Populations. *International Journal of Molecular Sciences*, *21*(18), 6557.

https://doi.org/10.3390/ijms21186557

Magana-Mora, A., Kalkatawi, M., & Bajic, V. B. (2017). Omni-PolyA: A method and tool for accurate recognition of Poly(A) signals in human genomic DNA. *BMC Genomics*, *18*(1), 620. https://doi.org/10.1186/s12864-017-4033-7

Mann, M., & Jensen, O. N. (2003). Proteomic analysis of post-translational modifications. *Nature Biotechnology*, *21*(3), 255–261. https://doi.org/10.1038/nbt0303-255

Mayr, C., & Bartel, D. P. (2009). Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell*, *138*(4), 673–684. https://doi.org/10.1016/j.cell.2009.06.016

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. https://doi.org/10.48550/ARXIV.1802.03426

McKellar, D. W., Mantri, M., Hinchman, M. M., Parker, J. S. L., Sethupathy, P., Cosgrove, B. D., & De Vlaminck, I. (2022). Spatial mapping of the total transcriptome by in situ polyadenylation. *Nature Biotechnology*. https://doi.org/10.1038/s41587-022-01517-6

Merkel, D. (2014). *Docker: Lightweight Linux Containers for Consistent Development and Deployment*. https://dlnext.acm.org/doi/10.5555/2600239.2600241

Meyer, S., Temme, C., & Wahle, E. (2004). Messenger RNA Turnover in Eukaryotes: Pathways and Enzymes. *Critical Reviews in Biochemistry and Molecular Biology*, *39*(4), 197–216. https://doi.org/10.1080/10409230490513991

Millevoi, S., Loulergue, C., Dettwiler, S., Karaa, S. Z., Keller, W., Antoniou, M., & Vagner, S. (2006). An interaction between U2AF 65 and CF Im links the splicing and 3' end processing machineries. *The EMBO Journal*, *25*(20), 4854–4864. https://doi.org/10.1038/sj.emboj.7601331

Mitra, M., Johnson, E. L., Swamy, V. S., Nersesian, L. E., Corney, D. C., Robinson, D. G., Taylor, D. G., Ambrus, A. M., Jelinek, D., Wang, W., Batista, S. L., & Coller, H. A. (2018). Alternative polyadenylation factors link cell cycle to migration. *Genome Biology*, *19*(1), 176. https://doi.org/10.1186/s13059-018-1551-9

Mitschka, S., & Mayr, C. (2022). Context-specific regulation and function of mRNA alternative polyadenylation. *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/s41580-022-00507-5

Moffitt, J. R., Lundberg, E., & Heyn, H. (2022). The emerging landscape of spatial profiling technologies. *Nature Reviews Genetics*, *23*(12), 741–759. https://doi.org/10.1038/s41576-022-00515-3

Moon, Y., Burri, D., & Zavolan, M. (2023). Identification of experimentally-supported poly(A) sites in single-cell RNA-seq data with SCINPAS. *NAR Genomics and Bioinformatics*, *5*(3), lqad079. https://doi.org/10.1093/nargab/lqad079

Muller, S., Rycak, L., Afonso-Grunz, F., Winter, P., Zawada, A. M., Damrath, E., Scheider, J., Schmah, J., Koch, I., Kahl, G., & Rotter, B. (2014). APADB: A database for alternative polyadenylation and microRNA regulation events. *Database*, *2014*(0), bau076–bau076. https://doi.org/10.1093/database/bau076

Nam, D. K., Lee, S., Zhou, G., Cao, X., Wang, C., Clark, T., Chen, J., Rowley, J. D., & Wang, S. M. (2002). Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proceedings of the National Academy of Sciences*, *99*(9), 6152–6156. https://doi.org/10.1073/pnas.092140899

Niwa, M., & Berget, S. M. (1991). Mutation of the AAUAAA polyadenylation signal depresses in vitro splicing of proximal but not distal introns. *Genes & Development*, *5*(11), 2086–2095. https://doi.org/10.1101/gad.5.11.2086

Niwa, M., Rose, S. D., & Berget, S. M. (1990). In vitro polyadenylation is stimulated by the presence of an upstream intron. *Genes & Development*, *4*(9), 1552–1559. https://doi.org/10.1101/gad.4.9.1552

Noble, R., Burri, D., Le Sueur, C., Lemant, J., Viossat, Y., Kather, J. N., & Beerenwinkel, N. (2021). Spatial structure governs the mode of tumour evolution. *Nature Ecology & Evolution*, *6*(2), 207–217. https://doi.org/10.1038/s41559-021-01615-9

Oetting, W. S., & Adams, D. (2018). Albinism: Genetics. In John Wiley & Sons, Ltd (Ed.), *eLS* (1st ed., pp. 1–8). Wiley. https://doi.org/10.1002/9780470015902.a0006081.pub3

Ohtsubo, K., & Marth, J. D. (2006). Glycosylation in Cellular Mechanisms of Health and Disease. *Cell*, *126*(5), 855–867. https://doi.org/10.1016/j.cell.2006.08.019

PacBio. (2021). MAS-seq for 10X Single Cell 3' kit. *Single-Cell RNA Sequencing. Full-Length Isoform Information for Your Single-Cell Transcriptome Studies*. https://www.pacb.com/products-and-services/applications/rna-sequencing/single-cell-rna-sequencing/

Pace, L., Goudot, C., Zueva, E., Gueguen, P., Burgdorf, N., Waterfall, J. J., Quivy, J.-P., Almouzni, G., & Amigorena, S. (2018). The epigenetic control of stemness in CD8+ T cell fate commitment. *Science*, *359*(6372), 177–186. https://doi.org/10.1126/science.aah6499

Park, J.-E., Yi, H., Kim, Y., Chang, H., & Kim, V. N. (2016). Regulation of Poly(A) Tail and Translation during the Somatic Cell Cycle. *Molecular Cell*, *62*(3), 462–471. https://doi.org/10.1016/j.molcel.2016.04.007

Parker, R., & Song, H. (2004). The enzymes and control of eukaryotic mRNA turnover. *Nature Structural & Molecular Biology*, *11*(2), 121–127. https://doi.org/10.1038/nsmb724

Patrick, R., Humphreys, D. T., Janbandhu, V., Oshlack, A., Ho, J. W. K., Harvey, R. P., & Lo, K. K. (2020). Sierra: Discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome Biology*, *21*(1), 167. https://doi.org/10.1186/s13059-020-02071-7

Peattie, D. A., Hsiao, K., Benasutti, M., & Lippke, J. A. (1994). Three distinct messenger RNAs can encode the human immunosuppressant-binding protein FKBP12. *Gene*, *150*(2), 251–257. https://doi.org/10.1016/0378-1119(94)90434-0

Perfetto, S. P., Chattopadhyay, P. K., & Roederer, M. (2004). Seventeen-colour flow cytometry: Unravelling the immune system. *Nature Reviews Immunology*, *4*(8), 648–655. https://doi.org/10.1038/nri1416

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., Moore, R., McClanahan, T. K., Sadekova, S., & Klappenbach, J. A. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, *35*(10), 936–939. https://doi.org/10.1038/nbt.3973

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, *9*(1), 171–181. https://doi.org/10.1038/nprot.2014.006

Proserpio, V. (Ed.). (2019). *Single Cell Methods: Sequencing and Proteomics* (Vol. 1979). Springer New York. https://doi.org/10.1007/978-1-4939-9240-9

Ramazi, S., & Zahiri, J. (2021). Post-translational modifications in proteins: Resources, tools and prediction methods. *Database*, *2021*, baab012. https://doi.org/10.1093/database/baab012

Ran, D., Zhang, S., Lytal, N., & An, L. (2020). scDoc: Correcting drop-out events in single-cell RNA-seq data. *Bioinformatics*, *36*(15), 4233–4239. https://doi.org/10.1093/bioinformatics/btaa283

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., … Human Cell Atlas Meeting Participants. (2017). The Human Cell Atlas. *eLife*, *6*, e27041. https://doi.org/10.7554/eLife.27041

Riba, A., Di Nanni, N., Mittal, N., Arhné, E., Schmidt, A., & Zavolan, M. (2019). Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. *Proceedings of the National Academy of Sciences*, *116*(30), 15023–15032. https://doi.org/10.1073/pnas.1817299116

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Rogers, S., Girolami, M., Kolch, W., Waters, K. M., Liu, T., Thrall, B., & Wiley, H. S. (2008). Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics*, *24*(24), 2894–2900. https://doi.org/10.1093/bioinformatics/btn553

Rorbach, J., & Bobrowicz, A. J. (Eds.). (2014). *Polyadenylation: Methods and Protocols* (Vol. 1125). Humana Press. https://doi.org/10.1007/978-1-62703-971-0

Ryu, S., Sidorov, S., Ravussin, E., Artyomov, M., Iwasaki, A., Wang, A., & Dixit, V. D. (2022). The matricellular protein SPARC induces inflammatory interferon-response in macrophages during aging. *Immunity*, *55*(9), 1609-1626.e7. https://doi.org/10.1016/j.immuni.2022.07.007

Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., & Burge, C. B. (2008). Proliferating Cells Express mRNAs with Shortened 3’ Untranslated Regions and Fewer MicroRNA Target Sites. *Science*, *320*(5883), 1643–1647. https://doi.org/10.1126/science.1155390

Shah, A., Mittleman, B. E., Gilad, Y., & Li, Y. I. (2021). Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation. *Genome Biology*, *22*(1), 291. https://doi.org/10.1186/s13059-021-02502-z

Sheets, M. D., Fox, C. A., Hunt, T., Vande Woude, G., & Wickens, M. (1994). The 3’-untranslated regions of c-mos and cyclin mRNAs stimulate translation by regulating cytoplasmic polyadenylation. *Genes & Development*, *8*(8), 926–938. https://doi.org/10.1101/gad.8.8.926

Shen, T., Li, H., Song, Y., Li, L., Lin, J., Wei, G., & Ni, T. (2019). Alternative polyadenylation dependent function of splicing factor SRSF3 contributes to cellular senescence. *Aging*, *11*(5), 1356–1388. https://doi.org/10.18632/aging.101836

Shepard, P. J., Choi, E.-A., Lu, J., Flanagan, L. A., Hertel, K. J., & Shi, Y. (2011). Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, *17*(4), 761–772. https://doi.org/10.1261/rna.2581711

Sheppard, S., Lawson, N. D., & Zhu, L. J. (2013). Accurate identification of polyadenylation sites from 3′ end deep sequencing using a naïve Bayes classifier. *Bioinformatics*, *29*(20), 2564–2571. https://doi.org/10.1093/bioinformatics/btt446

Shulman, E. D., & Elkon, R. (2019). Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic Acids Research*, *47*(19), 10027–10039. https://doi.org/10.1093/nar/gkz781

Singh, I., Lee, S.-H., Sperling, A. S., Samur, M. K., Tai, Y.-T., Fulciniti, M., Munshi, N. C., Mayr, C., & Leslie, C. S. (2018). Widespread intronic polyadenylation diversifies immune cell transcriptomes. *Nature Communications*, *9*(1), 1716. https://doi.org/10.1038/s41467-018-04112-z

Sommerkamp, P., Altamura, S., Renders, S., Narr, A., Ladel, L., Zeisberger, P., Eiben, P. L., Fawaz, M., Rieger, M. A., Cabezas-Wallscheid, N., & Trumpp, A. (2020). Differential Alternative Polyadenylation Landscapes Mediate Hematopoietic Stem Cell Activation and Regulate Glutamine Metabolism. *Cell Stem Cell*, *26*(5), 722-738.e7. https://doi.org/10.1016/j.stem.2020.03.003

Sonenberg, N., & Hinnebusch, A. G. (2009). Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell*, *136*(4), 731–745. https://doi.org/10.1016/j.cell.2009.01.042

Spies, N., Burge, C. B., & Bartel, D. P. (2013). 3′ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Research*, *23*(12), 2078–2090. https://doi.org/10.1101/gr.156919.113

Spill, F., Reynolds, D. S., Kamm, R. D., & Zaman, M. H. (2016). Impact of the physical microenvironment on tumor progression and metastasis. *Current Opinion in Biotechnology*, *40*, 41–48. https://doi.org/10.1016/j.copbio.2016.02.007

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., … Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, *353*(6294), 78–82. https://doi.org/10.1126/science.aaf2403

Steffen, P., Voß, B., Rehmsmeier, M., Reeder, J., & Giegerich, R. (2006). RNAshapes: An integrated RNA analysis package based on abstract shapes. *Bioinformatics*, *22*(4), 500–503. https://doi.org/10.1093/bioinformatics/btk010

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., & Smibert, P. (2017). *Large-scale simultaneous measurement of epitopes and transcriptomes in single cells* [Preprint]. Genomics. https://doi.org/10.1101/113068

Strumillo, M., & Beltrao, P. (2015). Towards the computational design of protein post-translational regulation. *Bioorganic & Medicinal Chemistry*, *23*(12), 2877–2882. https://doi.org/10.1016/j.bmc.2015.04.056

Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H., & Bartel, D. P. (2014). Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature*, *508*(7494), 66–71. https://doi.org/10.1038/nature13007

Sun, S., Zhu, J., & Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods*, *17*(2), 193–200. https://doi.org/10.1038/s41592-019-0701-7

Sun, Y. H., Wang, A., Song, C., Shankar, G., Srivastava, R. K., Au, K. F., & Li, X. Z. (2021). Single-molecule long-read sequencing reveals a conserved intact long RNA profile in sperm. *Nature Communications*, *12*(1), 1361. https://doi.org/10.1038/s41467-021-21524-6

Svensson, V., Teichmann, S. A., & Stegle, O. (2018). SpatialDE: Identification of spatially variable genes. *Nature Methods*, *15*(5), 343–346. https://doi.org/10.1038/nmeth.4636

Svoboda, M., Frost, H. R., & Bosco, G. (2022). Internal oligo(dT) priming introduces systematic bias in bulk and single-cell RNA sequencing count data. *NAR Genomics and Bioinformatics*, *4*(2), lqac035. https://doi.org/10.1093/nargab/lqac035

Swindell, W. R. (2008). Comparative analysis of microarray data identifies common responses to caloric restriction among mouse tissues. *Mechanisms of Ageing and Development*, *129*(3), 138–153. https://doi.org/10.1016/j.mad.2007.11.003

Takagaki, Y., Seipelt, R. L., Peterson, M. L., & Manley, J. L. (1996). The Polyadenylation Factor CstF-64 Regulates Alternative Processing of IgM Heavy Chain Pre-mRNA during B Cell Differentiation. *Cell*, *87*(5), 941–952. https://doi.org/10.1016/S0092-8674(00)82000-0

Taliaferro, J. M., Vidaki, M., Oliveira, R., Olson, S., Zhan, L., Saxena, T., Wang, E. T., Graveley, B. R., Gertler, F. B., Swanson, M. S., & Burge, C. B. (2016). Distal Alternative Last Exons Localize mRNAs to Neural Projections. *Molecular Cell*, *61*(6), 821–833.

https://doi.org/10.1016/j.molcel.2016.01.020

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382. https://doi.org/10.1038/nmeth.1315

Tang, P., Yang, Y., Li, G., Huang, L., Wen, M., Ruan, W., Guo, X., Zhang, C., Zuo, X., Luo, D., Xu, Y., Fu, X.-D., & Zhou, Y. (2022). Alternative polyadenylation by sequential activation of distal and proximal PolyA sites. *Nature Structural & Molecular Biology*. https://doi.org/10.1038/s41594-021-00709-z

Tekath, T., & Dugas, M. (2021). Differential transcript usage analysis of bulk and single-cell RNA-seq data with DTUrtle. *Bioinformatics*, *37*(21), 3781–3787. https://doi.org/10.1093/bioinformatics/btab629

The Tabula Muris Consortium, Almanzar, N., Antony, J., Baghel, A. S., Bakerman, I., Bansal, I., Barres, B. A., Beachy, P. A., Berdnik, D., Bilen, B., Brownfield, D., Cain, C., Chan, C. K. F., Chen, M. B., Clarke, M. F., Conley, S. D., Darmanis, S., Demers, A., Demir, K., … Zou, J. (2020). A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, *583*(7817), 590–595. https://doi.org/10.1038/s41586-020-2496-1

The Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, & Principal investigators. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, *562*(7727), 367–372. https://doi.org/10.1038/s41586-018-0590-4

The Tabula Sapiens Consortium, & Quake, S. R. (2021). *The Tabula Sapiens: A multiple organ single cell transcriptomic atlas of humans* [Preprint]. Cell Biology. https://doi.org/10.1101/2021.07.19.452956

Tian, B. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, *33*(1), 201–212. https://doi.org/10.1093/nar/gki158

Tian, B., & Manley, J. L. (2017). Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology*, *18*(1), 18–30. https://doi.org/10.1038/nrm.2016.116

Travaglini, K. J., Nabhan, A. N., Penland, L., Sinha, R., Gillich, A., Sit, R. V., Chang, S., Conley, S. D., Mori, Y., Seita, J., Berry, G. J., Shrager, J. B., Metzger, R. J., Kuo, C. S., Neff, N., Weissman, I. L., Quake, S. R., & Krasnow, M. A. (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature*, *587*(7835), 619–625. https://doi.org/10.1038/s41586-020-2922-4

van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., Bierie, B., Mazutis, L., Wolf, G., Krishnaswamy, S., & Pe'er, D. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, *174*(3), 716-729.e27. https://doi.org/10.1016/j.cell.2018.05.061

Velten, L., Anders, S., Pekowska, A., Järvelin, A. I., Huber, W., Pelechano, V., & Steinmetz, L. M. (2015).

Single-cell polyadenylation site mapping reveals 3′ isoform choice variability. *Molecular Systems Biology*, *11*(6), 812. https://doi.org/10.15252/msb.20156198

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., & Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. https://doi.org/10.1038/nature07509

Wang, J., Agarwal, D., Huang, M., Hu, G., Zhou, Z., Ye, C., & Zhang, N. R. (2019). Data denoising with transfer learning in single-cell transcriptomics. *Nature Methods*, *16*(9), 875–878. https://doi.org/10.1038/s41592-019-0537-1

Wang, J., Chen, W., Yue, W., Hou, W., Rao, F., Zhong, H., Qi, Y., Hong, N., Ni, T., & Jin, W. (2022). Comprehensive mapping of alternative polyadenylation site usage and its dynamics at single-cell resolution. *Proceedings of the National Academy of Sciences*, *119*(49), e2113504119. https://doi.org/10.1073/pnas.2113504119

Wang, L., Chen, M., Fu, H., Ni, T., & Wei, G. (2020). Tempo-spatial alternative polyadenylation analysis reveals that 3′ UTR lengthening of Mdm2 regulates p53 expression and cellular senescence in aged rat testis. *Biochemical and Biophysical Research Communications*, *523*(4), 1046–1052. https://doi.org/10.1016/j.bbrc.2020.01.061

Wang, L., Dowell, R. D., & Yi, R. (2013). Genome-wide maps of polyadenylation reveal dynamic mRNA 3′-end formation in mammalian cell lineages. *RNA*, *19*(3), 413–425. https://doi.org/10.1261/rna.035360.112

Wang, M., Jiang, Y., & Xu, X. (2015). A novel method for predicting post-translational modifications on serine and threonine sites by using site-modification network profiles. *Molecular BioSystems*, *11*(11), 3092–3100. https://doi.org/10.1039/C5MB00384A

Wang, Q., Bönigk, S., Böhm, V., Gehring, N., Altmüller, J., & Dieterich, C. (2021). *Single cell transcriptome sequencing on the Nanopore platform with ScNapBar*. 22.

Wang, R., Nambiar, R., Zheng, D., & Tian, B. (2018). PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Research*, *46*(D1), D315–D319. https://doi.org/10.1093/nar/gkx1000

Wang, R., Zheng, D., Wei, L., Ding, Q., & Tian, B. (2019). Regulation of Intronic Polyadenylation by PCF11 Impacts mRNA Expression of Long Genes. *Cell Reports*, *26*(10), 2766-2778.e6. https://doi.org/10.1016/j.celrep.2019.02.049

Wang, R., Zheng, D., Yehia, G., & Tian, B. (2018). A compendium of conserved cleavage and polyadenylation events in mammalian genes. *Genome Research*, *28*(10), 1427–1441. https://doi.org/10.1101/gr.237826.118

Wei, L., Xing, P., Shi, G., Ji, Z., & Zou, Q. (2019). Fast Prediction of Protein Methylation Sites Using a Sequence-Based Feature Selection Technique. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *16*(4), 1264–1273. https://doi.org/10.1109/TCBB.2017.2670558

Weill, L., Belloc, E., Bava, F.-A., & Méndez, R. (2012). Translational control by changes in poly(A) tail length: Recycling mRNAs. *Nature Structural & Molecular Biology*, *19*(6), 577–585. https://doi.org/10.1038/nsmb.2311

Wright, D. J., Hall, N. A. L., Irish, N., Man, A. L., Glynn, W., Mould, A., Angeles, A. D. L., Angiolini, E., Swarbreck, D., Gharbi, K., Tunbridge, E. M., & Haerty, W. (2022). Long read sequencing reveals novel isoforms and insights into splicing regulation during cell state changes. *BMC Genomics*, *23*(1), 42. https://doi.org/10.1186/s12864-021-08261-2

Wu, X., Liu, T., Ye, C., Ye, W., & Ji, G. (2020). scAPAtrap: Identification and quantification of alternative polyadenylation sites from single-cell RNA-seq data. *Briefings in Bioinformatics*, bbaa273. https://doi.org/10.1093/bib/bbaa273

Xia, C., Fan, J., Emanuel, G., Hao, J., & Zhuang, X. (2019). Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proceedings of the National Academy of Sciences*, *116*(39), 19490–19499. https://doi.org/10.1073/pnas.1912459116

Xia, Z., Donehower, L. A., Cooper, T. A., Neilson, J. R., Wheeler, D. A., Wagner, E. J., & Li, W. (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nature Communications*, *5*(1), 5274. https://doi.org/10.1038/ncomms6274

Xia, Z., Li, Y., Zhang, B., Li, Z., Hu, Y., Chen, W., & Gao, X. (2019). DeeReCT-PolyA: A robust and generic deep learning method for PAS identification. *Bioinformatics*, *35*(14), 2371–2379. https://doi.org/10.1093/bioinformatics/bty991

Xie, B., Jankovic, B. R., Bajic, V. B., Song, L., & Gao, X. (2013). Poly(A) motif prediction using spectral latent features from human DNA sequences. *Bioinformatics*, *29*(13), i316–i325. https://doi.org/10.1093/bioinformatics/btt218

Xu, S., Tang, L., Dai, G., Luo, C., & Liu, Z. (2021). Immune-related genes with APA in microenvironment indicate risk stratification and clinical prognosis in grade II/III gliomas. *Molecular Therapy - Nucleic Acids*, *23*, 1229–1242. https://doi.org/10.1016/j.omtn.2021.01.033

Xu, Y., & Chou, K.-C. (2015). Recent Progress in Predicting Posttranslational Modification Sites in Proteins. *Current Topics in Medicinal Chemistry*, *16*(6), 591–603. https://doi.org/10.2174/1568026615666150819110421

Yang, S., Jiang, W., Yang, W., Yang, C., Yang, X., Chen, K., Hu, Y., Shen, G., Lu, L., Cheng, F., Zhang, F., Rao, J., & Wang, X. (2021). Epigenetically modulated miR-1224 suppresses the proliferation of HCC through CREB-mediated activation of YAP signaling pathway. *Molecular Therapy - Nucleic Acids*, *23*, 944–958. https://doi.org/10.1016/j.omtn.2021.01.008

Yang, Y., Paul, A., Bach, T. N., Huang, Z. J., & Zhang, M. Q. (2021). Single-cell alternative polyadenylation analysis delineates GABAergic neuron types. *BMC Biology*, *19*(1), 144. https://doi.org/10.1186/s12915-021-01076-3

Yao, C., Biesinger, J., Wan, J., Weng, L., Xing, Y., Xie, X., & Shi, Y. (2012). Transcriptome-wide analyses of CstF64–RNA interactions in global regulation of mRNA alternative polyadenylation. *Proceedings of the National Academy of Sciences*, *109*(46), 18773–18778. https://doi.org/10.1073/pnas.1211101109

Ye, C., Lin, J., & Li, Q. Q. (2020). Discovery of alternative polyadenylation dynamics from single cell types. *Computational and Structural Biotechnology Journal*, *18*, 1012–1019. https://doi.org/10.1016/j.csbj.2020.04.009

Ye, C., Zhou, Q., Wu, X., Yu, C., Ji, G., Saban, D. R., & Li, Q. Q. (2020). scDAPA: Detection and visualization of dynamic alternative polyadenylation from single cell RNA-seq data. *Bioinformatics*, *36*(4), 1262–1264. https://doi.org/10.1093/bioinformatics/btz701

Ye, W., Lian, Q., Ye, C., & Wu, X. (2022). A Survey on Methods for Predicting Polyadenylation Sites from DNA Sequences, Bulk RNA-seq, and Single-cell RNA-seq. *Genomics, Proteomics & Bioinformatics*, S1672022922001218. https://doi.org/10.1016/j.gpb.2022.09.005

Yeo, G. W. (Ed.). (2014). *Systems Biology of RNA Binding Proteins* (Vol. 825). Springer New York. https://doi.org/10.1007/978-1-4939-1221-6

Yin, Y., Hua, H., Li, M., Liu, S., Kong, Q., Shao, T., Wang, J., Luo, Y., Wang, Q., Luo, T., & Jiang, Y. (2016). mTORC2 promotes type I insulin-like growth factor receptor and insulin receptor activation through the tyrosine kinase activity of mTOR. *Cell Research*, *26*(1), 46–65. https://doi.org/10.1038/cr.2015.133

You, L., Wu, J., Feng, Y., Fu, Y., Guo, Y., Long, L., Zhang, H., Luan, Y., Tian, P., Chen, L., Huang, G., Huang, S., Li, Y., Li, J., Chen, C., Zhang, Y., Chen, S., & Xu, A. (2015). 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. https://doi.org/10.1093/nar/gku1076

Zhang, H., Lee, J., & Tian, B. (2005). Biased alternative polyadenylation in human tissues. *Genome Biology*, *6*(12), R100. https://doi.org/10.1186/gb-2005-6-12-r100

Zhang, Y., Zhang, J., & Wang, S. (2021). The Role of Rapamycin in Healthspan Extension via the Delay of Organ Aging. *Ageing Research Reviews*, *70*, 101376. https://doi.org/10.1016/j.arr.2021.101376

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*(1), 14049. https://doi.org/10.1038/ncomms14049

Zhou, R., Xiao, X., He, P., Zhao, Y., Xu, M., Zheng, X., Yang, R., Chen, S., Zhou, L., Zhang, D., Yang, Q., Song, J., Tang, C., Zhang, Y., Lin, J., Cheng, L., & Chen, L. (2022). SCAPE: A mixture model revealing single-cell polyadenylation diversity and cellular dynamics during cell differentiation and reprogramming. *Nucleic Acids Research*, gkac167. https://doi.org/10.1093/nar/gkac167

Zhou, Z., Ye, C., Wang, J., & Zhang, N. R. (2020). Surface protein imputation from single cell transcriptomes by deep neural networks. *Nature Communications*, *11*(1), 651. https://doi.org/10.1038/s41467-020-14391-0

Zhu, J., Sun, S., & Zhou, X. (2021). SPARK-X: Non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology*, *22*(1), 184. https://doi.org/10.1186/s13059-021-02404-0

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., & Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, *65*(4), 631-643.e4. https://doi.org/10.1016/j.molcel.2017.01.023

Appendix

Appendix A

Shortening of 3’ UTRs in most cell types composing tumor tissues implicates alternative polyadenylation in protein metabolism

Dominik Burri[#], Mihaela Zavolan[#]

[#] Computational and Systems Biology, Biozentrum, University of Basel

Klingelbergstrasse 50-70, Basel, CH-4056, Switzerland

Keywords

Alternative polyadenylation, single cell transcriptomics, lung cancer, bioinformatics, computational method

A. 1. Abstract

During pre-mRNA maturation 3’ end processing can occur at different polyadenylation sites in the 3’ untranslated region (3’ UTR) to give rise to transcript isoforms that differ in the length of their 3’UTRs. Longer 3’ UTRs contain additional cis-regulatory elements that impact the fate of the transcript and/or of the resulting protein.

Extensive alternative polyadenylation (APA) has been observed in cancers, but the mechanisms and roles remain elusive. In particular, it is unclear whether the APA occurs in the malignant cells or in other cell types that infiltrate the tumor. To resolve this, we developed a computational method, called SCUREL, that quantifies changes in 3’UTR length between groups of cells, including cells of the same type originating from tumor and control tissue. We used this method to study APA in human lung adenocarcinoma (LUAD).

SCUREL relies solely on annotated 3’UTRs and on control systems, such as T cell activation and spermatogenesis gives qualitatively similar results at much greater sensitivity compared to the previously published scAPA method.

In the LUAD samples, we find a general trend towards 3’UTR shortening not only in cancer cells compared to the cell type of origin, but also when comparing other cell types from the tumor vs. the control tissue environment. However, we also find high variability in the individual targets between patients. The findings help to understand the extent and impact of APA in LUAD, which may support improvements in diagnosis and treatment.

A. 2. Introduction

The processing of most human pre-mRNAs involves 3′ end cleavage and addition of a polyadenosine (poly(A)) tail. Typically, there are multiple cleavage and polyadenylation sites within a gene, and alternative polyadenylation (APA) has emerged as a major source of transcriptome diversity (Reyes & Huber, 2018). A prevalent type of APA isoforms are those that differ only in the length of their 3′ untranslated regions (3′ UTRs). 3′ UTRs become shorter upon T cell activation (A. R. Gruber et al., 2014; Sandberg et al., 2008), in cancer cells (Mayr & Bartel, 2009; Xia et al., 2014) and upon induction of reprogramming in somatic cells (Ji & Tian, 2009). Although the responsible regulators are still to be determined, core 3′ end processing factors under the transcriptional control of cell cycle-related transcription factors have been implicated, at least in the context of cell proliferation (Elkon et al., 2012). Various RNA-binding proteins (RBPs) are also involved in specific cellular systems (A. J. Gruber, Schmidt, et al., 2018; Lee et al., 2021; Martin et al., 2012; Masuda et al., 2020; So et al., 2019).

While APA-dependent 3′ UTR shortening has been observed in many cancers (Schmidt et al., 2018; Xia et al., 2014), it is presently unclear whether it is a manifestation of the change in cell composition of the tissue or of functional changes in all cell types within the tumor environment. As single cell RNA sequencing (scRNA-seq) technologies specifically capture mRNA 3′ ends, and datasets of tumor and matched control tissue samples have started to become available, this question can now be addressed, provided a few challenges are overcome. First, the number of transcripts that can be reliably quantified is still low (Breda et al., 2021), because the total number of reads obtained from individual cells is in the 10^3-10^4 range. Thus, quantifying gene expression at the isoform level is still very challenging. This issue can be partially circumvented by pooling the reads from cells of the same type. Second, while 3′ biased, scRNA-seq reads do not always reach the PAS and may also result from internal priming. Thus, identifying which reads correspond to the same 3′ end is also not trivial. This problem can be mitigated by associating scRNA-seq reads with already-annotated transcript 3′ ends. However, the current annotation is still far from complete (A. J. Gruber, Gypas, et al., 2018), leading to PAS usage quantification that is imprecise and incomplete. For this reason we developed a PAS-agnostic approach for quantifying changes in 3′ UTR length between samples, based on the entire 3′ end read distribution along the 3′ UTR. Applying the method to single cell sequencing data from human lung adenocarcinoma (LUAD), we found that 3′ UTR shortening is not specific to a cell type but rather occurs in most cell types that compose the tumor. Furthermore, our analysis revealed that the targeted transcripts encode proteins that are involved in various steps of protein metabolism, including synthesis at the endoplasmic reticulum (ER), transport between ER and the Golgi network and finally secretion of proteins. Our data thus implicates APA in the remodeling of protein metabolism in tumors.

A. 3. Results

A. 3. 1. A myeloid to lymphoid switch in lung tumors

While analyses of bulk RNA-seq data revealed the shortening of 3′ UTRs in virtually all studied cancers with respect to matched control tissue, the shortening is especially pronounced in lung tumors (A. J. Gruber, Schmidt, et al., 2018). Thus, to better understand the mechanism and function of APA in cancers, we identified two studies in which single cell sequencing of lung adenocarcinoma (LUAD) and matched control tissue from multiple patients was carried out on the same platform, 10x Genomics (Lambrechts et al., 2018; Laughney et al., 2020). These data enable us to not only identify 3′ UTR changes in specific cell types, but also to assess their generality between studies and patients. We followed the procedure described in (Lambrechts et al., 2018) to annotate the type of individual cells. Briefly, we integrated the data with the *harmony* package (see Methods, Suppl. Fig. 1), clustered the normalized gene expression vectors of all cells (Fig. 1A) with the *Seurat* package (Butler et al., 2018), and annotated the type of 38’156 cells from 12 samples of the (Lambrechts et al., 2018) study (samples 3a-d, 4a-d, 6a-d, representing 3 tumor samples and a matched control for each of three patients) and 18’543 cells of the (Laughney et al., 2020) study (3 pairs of tumor-matched control samples) based on known markers. We used the markers proposed in the (Lambrechts et al., 2018) study, but also added a few markers for mast cell (*TPSAB1*, *TPSB2* and *CPA3*; (Dwyer et al., 2016) Table 1) (Fig. 1B). As described in the initial study (Lambrechts et al., 2018), the most abundant cell types in the tumor samples were T cells, myeloid and B cells, while the matched control samples were dominated by myeloid and alveolar cells (Fig.1C). We further identified a small cluster of mast cells, annotated as B cells in the initial study that did not consider mast cell markers. We observed a similar myeloid to T cell switch between control and cancer samples from the (Laughney et al., 2020) study (Fig. 1D). In addition, the matched control samples from this latter study had a more homogenous cell type composition compared to those from the (Lambrechts et al., 2018) study, consisting almost exclusively of lymphocytes and myeloid cells (Fig. 1D).

Given that T cells are the most numerous cell type in tumor samples and that T cell activation leads to 3’ UTR shortening (A. R. Gruber et al., 2014; Sandberg et al., 2008) we wondered whether the pattern of 3’ UTR usage that was previously inferred from ‘bulk’ samples can be attributed to the infiltration of the tumor with activated T cells. To investigate this possibility, we first determined the distribution of RNA molecules (unique molecular identifiers, UMI) per cell in various cell types in the two studies (Suppl. Fig. 2A) and the total number of UMIs obtained from each cell type in each data set (Suppl. Fig. 2B). While T cells were the most numerous cell type in tumors, their relatively small RNA content per cell led to a smaller overall contribution to the total RNA pool compared to the less numerous myeloid cells, which have substantially more RNA molecules per cell (Suppl. Fig. 2A). Thus, the ‘bulk’ RNA obtained from tumor samples is not dominated by RNA originating from T cells, suggesting that other cell types also contribute to the 3’ UTR shortening that was previously described in tumors. We therefore carried out a cell type-specific analysis of 3’ UTR usage in tumors relative to matched controls.

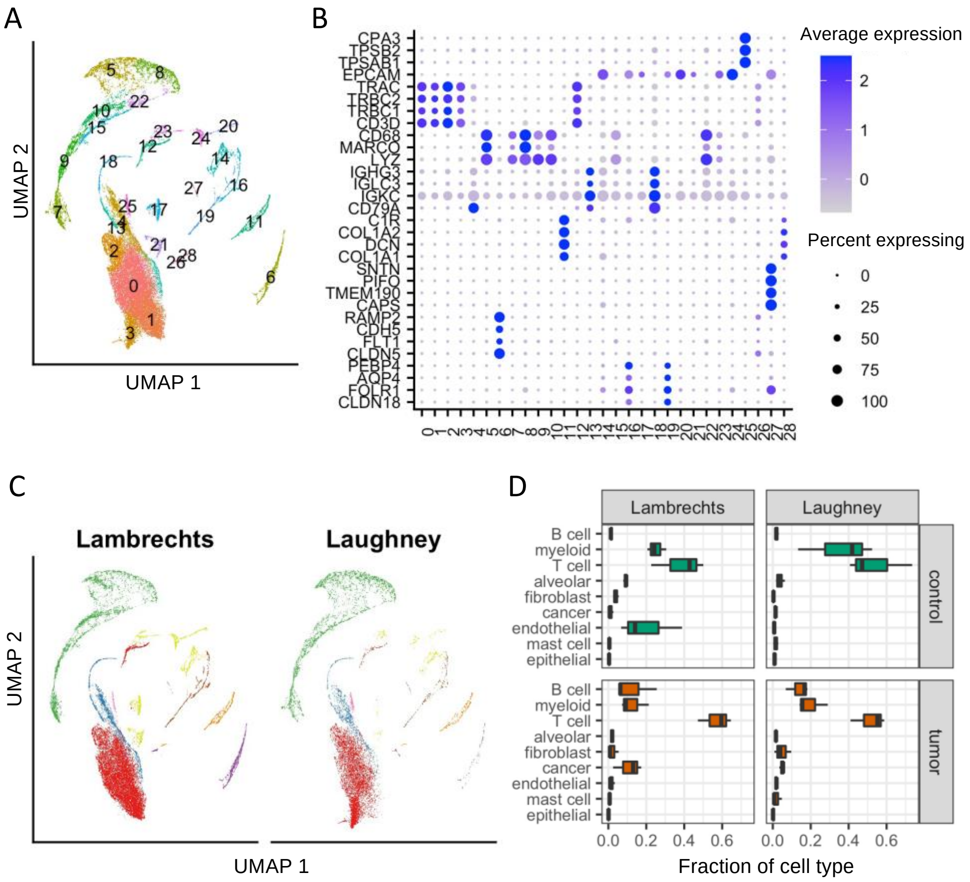


Figure 1. Cell type composition of lung adenocarcinoma and matched control samples.
A. 2-dimensional projection (Uniform Manifold Approximation and Projection, UMAP) of gene expression vectors. The projections were obtained with the *RunUMAP* function from *Seurat* v3.2.3 (Butler et al., 2018), based on the first 10 principal components. The two datasets were integrated with *harmony*. Cell clustering was done on the shared nearest neighbour (SNN) graph (see Methods). **B.** Dot plot of marker gene expression across the clusters shown in panel A. Shown is the average expression and percent of expressing cells per cluster for the markers used in (Lambrechts et al., 2018) (see also Table 1). The dot plot was created with *Seurat*. **C.** 2-dimensional projection (created with *Seurat*) of gene expression vectors as in **A**, but highlighting only cells from one study in each panel. **D.** Box plot of relative proportion of each cell type in control (green) and tumor (red) samples from individual patients from the Lambrechts and Laughney datasets.

A. 3. 2. A PAS-agnostic approach to quantify 3' UTR shortening and APA events

A few approaches have been proposed for assessing APA in scRNA-seq data sets (Patrick et al., 2020; Shulman & Elkon, 2019; Wu et al., 2020). However, their robustness with respect to the sparsity of the data and the incompleteness of PAS annotation has not been checked (Ye et al., 2020). Thus, we developed a novel approach (single cell analysis of 3' untranslated region lengths, SCUREL) (Figure 2A), specifically designed to circumvent these issues and implemented in a Snakemake (Koster & Rahmann, 2012) workflow. SCUREL enables two different comparisons of 3' UTR length: between two different cell types in a data set ("cell type" mode), or for the same cell type between two different conditions (e.g. tumor and matched control tissue, "condition" mode). We frame the detection of changes in 3' UTR length between two groups of cells as a problem of identifying the cell group from which the reads originated by inspecting the positions where the reads map in the terminal exons (TEs).

That is, read 3’ ends are tabulated and the cumulative coverage along individual TEs is calculated and normalized (Fig. 2B). Then, analyzing each TE individually, we record the fraction of reads from the two cell groups that map within an extending window of the TE starting from the 3’ end (Fig. 2C). This yields a curve in the plane defined by the proportions of reads in the two cell groups, which is similar to a receiver operating characteristic (ROC). The area under this curve (AUC) indicates the similarity of TE length between the compared cell groups. The curve is anchored at coordinates (0,0), corresponding to the end of the TE, where no reads have been observed yet, and (1,1), corresponding to the start of the TE, where all reads from the TE have been accounted for. If the coverage of a TE by read 3’ ends were similar between the two groups of cells and thus the cell group cannot be identified from the position of the reads, the curve would trace the diagonal line. Deviations above the diagonal indicate higher coverage of the distal region of the TE in the cell group represented on the y-axis, while deviations below the diagonal line indicate higher coverage of the distal TE region in the cell group represented on the x-axis. When the number of read mapping to a given TE is small, the curve will show discrete jumps of 1/*n* step size (where *n* is the number of reads mapping to the TE), as individual reads are encountered along the TE. This could lead to AUC values that deviate strongly from the 0.5 value expected under the assumption of similar coverage in the two cell groups. To avoid false positives that are caused by these finite sampling effects, we constructed a background coverage data set by randomizing the labels indicating the cell group from which each read originated. This preserves the depth of coverage of each TE in each group of cells while randomizing the location of each read, thus allowing us to determine changes in 3’ UTR length that cannot be explained by the sparsity of the data. For considerations of efficiency, we carried out the randomization once, and used the information from TEs with similar average coverage to detect significant AUC values. That is, the distribution of AUC values being wider for TEs with low coverage (in counts per million, CPM) compared to TEs with high coverage (Fig. 2D), we binned TEs by the average coverage in the two cell groups (in log(mean CPM)) and within each of the 20 bins, we used the 1% quantile of the randomized read data as the threshold for significant AUC values. Finally, noting that in some cases the difference in TE exon was small and unlikely to be due to APA, we selected only those TEs for which the read 3’ ends span a sufficiently large distance. That is, we calculated the interquartile range (IQR) of read 3’ end positions and, if the union of these intervals for the two cell clusters that were analyzed was larger than 200 nucleotides, we considered the range of 3’ end variation sufficient to be indicative of APA (Fig. 2D).

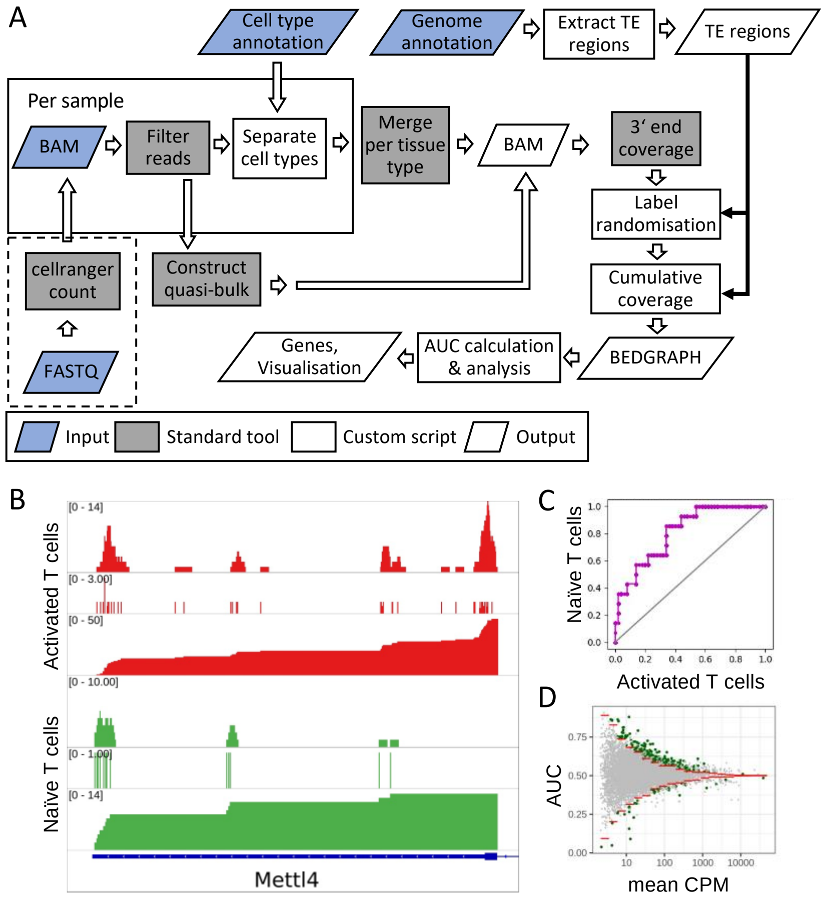


Figure 2. Overview of SCUREL.

A. Schematic representation of the workflow for detecting significant changes in 3′ UTR length between two cell populations. Input data (blue) consist of mapped reads from *cellranger count* and a table of annotated cell barcodes. The genome annotation is used to extract TEs, their cumulative 3′ end coverage in the two cell groups yielding the AUC value, which we used as a measure of APA. Dashed box: Alternative start of the workflow, from scRNA-seq reads in FASTQ format. The cell type annotation is done semi-automatically, based on marker gene expression (see Methods). **B.** Cumulative 3′ end coverage of the TE of mouse *Mettl4* gene in activated (red) and naive (green) T cells from the (Pace et al., 2018) study. For each cell type, the first track shows the read coverage along the TE, the second track the location of read 3′ ends and the third track the reverse cumulative of the 3′ end coverage. The gene is on the negative strand of the chromosome. **C.** Summary of the cumulative 3′ end read distribution along the TE of *Mettl4* in activated versus naive T cells, from the 3′ (at 0,0) to the 5′ (at 1,1) end. Points correspond to individual nucleotides of the TE where 3′ end reads are observed. The upwards deviation of the curve relative to the diagonal line indicates higher coverage of the distal region of the TE in naive T cells, quantified by the AUC value of 0.582. **D.** Distribution of AUC values as a function of log10(mean CPM) per TE in the mouse T cell activation data set (Pace et al., 2018). 9′099 TEs are represented, 218 showing significant shortening and 43 TEs significant lengthening (green points) attributed to APA.

A. 3. 3. SCUREL detects 3’ UTR length changes in previously characterized systems

To validate our approach, we analyzed the dynamics of 3’ UTR length in two well-characterized cellular systems, namely T cell activation, where 3’ UTRs become shorter, and sperm cell development, where the 3’ UTRs are known to become longer. Furthermore, we compared our results with those generated on these data sets by the previously published scAPA method (Shulman & Elkon, 2019).

We annotated the mouse T cell scRNA-seq data (Pace et al., 2018) with *Seurat,* obtaining 1605 activated and 1535 naïve T cells (Figure 3A), with 5.8 and 1.8 million reads mapped to TEs, respectively. Applying SCUREL, we identified 261 TEs whose length changed significantly upon T cell activation, of which 218 (84%) became shorter (Figure 3B). These results recapitulate those obtained from bulk RNA sequencing in a similar system (A. R. Gruber et al., 2014). Applying the previously published scAPA method (Shulman & Elkon, 2019) (see Methods) we only obtained 14 TEs with a significant length change, 12 of which (85%) became shorter (Figure 3C). ⅔ of the scAPA-identified targets (8 of 12 TEs) were also identified by our method, while the 4 cases missed by SCUREL involved either very small TE length changes (3 cases) or a difference in the annotation of the TE, because scAPA also quantifies PAS downstream of annotated TEs. In contrast, inspection of 9 randomly chosen TEs identified only by SCUREL indicated that they correspond to genes with relatively low expression, which are overlooked by scAPA (Suppl. Fig. 3). Examples of TEs from each of these categories are shown in Fig. 3G.

We carried out a similar analysis on a mouse spermatogenesis data set (Lukassen et al. 2018), as it is well known that 3'UTRs become progressively shorter during maturation of germ cells (spermatogonia) to spermatocytes, spermatids and finally spermatozoa. We used the markers described in the original publication (Lukassen et al., 2018) to annotate 386 elongating spermatids (ES) and 667 spermatocytes (SC), with 8 and 12 million reads in the TE regions, respectively (Figure 3D). Applying SCUREL, we found 2060 TEs whose length changed significantly from SCs to ES, almost all of which (1992, 97%) became shorter (Fig. 3E). scAPA yielded a similar proportion of shortened TEs (but fewer in absolute number), 96% (165 of 171 significant APA events, Figure 3F). As in the case of T cells, most of the scAPA-identified TEs were also found by our method (146 of 165 TEs), while TE annotation and small changes in PAS usage accounted for the cases that were unique to scAPA. Inspection of 9 randomly chosen TEs identified only by SCUREL indicated that they correspond to genes with relatively low expression or exclusively express one PAS or the other (Suppl. Fig. 4).

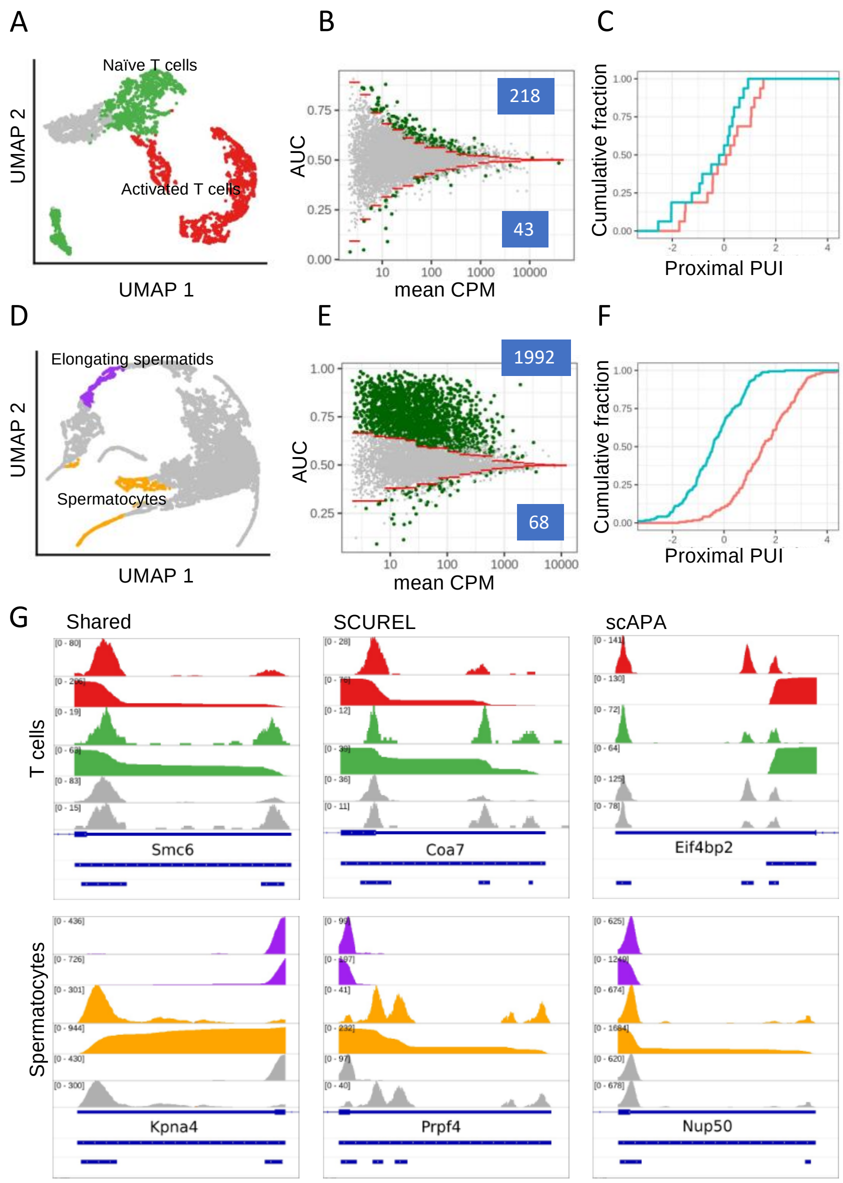
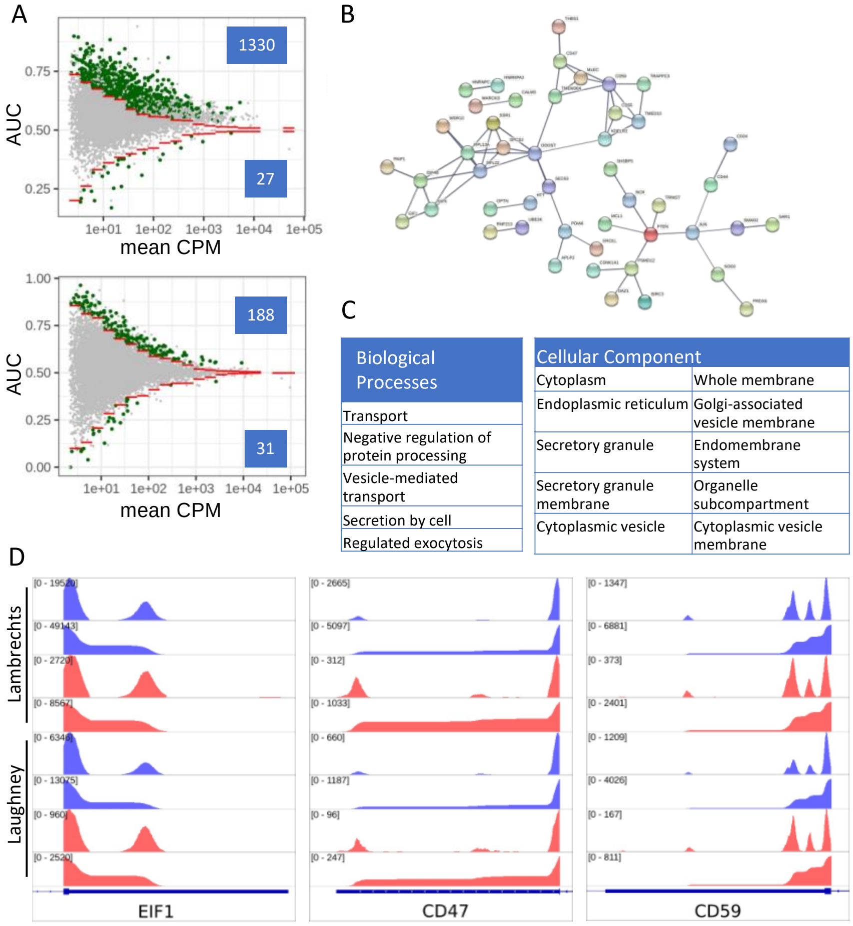


Figure 3: Analysis of APA in T cell activation and spermatogenesis.

A. UMAP projection of the T cell activation dataset (Pace et al., 2018) showing activated (red), naive (green) and unassigned (grey) T cells. **B.** Scatter plot of AUC in function of log10(mean CPM) for 9'099 TEs. The 1% quantiles (red lines) of the distributions obtained from the randomized dataset were used to identify TEs whose length changed significantly. AUC values > 0.5 indicate shorter 3' UTRs in activated T cells. TEs whose length changes were attributed to APA based on the span of the read 3' ends (see Methods) are shown in green. **C.** Cumulative distribution of proximal peak usage index (proximal PUI) for genes deemed by scAPA to undergo significant 3' UTR length changes. Activated T cells (red) generally have higher proximal PUI compared to naive T cells (blue), indicating 3'UTR shortening in activated T cells. **D.** UMAP projection of the spermatogenesis dataset (Lukassen et al., 2018), with highlighted elongating spermatids (purple) and spermatocytes (orange). **E.** Scatter plot of AUC in function of log10(mean CPM) for 7'875 TEs (see panel B for details). AUC values > 0.5 indicate longer 3' UTRs in spermatocytes. **F.** As in C, but comparing elongating spermatids (red) with spermatocytes (blue). **G.** Examples of genes deemed to exhibit significant change in 3' UTR length by both methods (left), by SCUREL only (middle) or by scAPA only (right). For each example, the tracks are: read coverage and cumulative distribution in the two conditions (activated - red - and resting - green - T cells for T cell examples, elongating spermatids - purple - and spermatocytes - orange - for the spermatogenesis examples, followed by coverage tracks from scAPA for the same two conditions in grey. The three blue tracks on the bottom denote in order, the Refseq annotation of the gene, the TE region analyzed in SCUREL and the peaks identified by scAPA.

A. 3. 4. Genes involved in protein metabolism are targets of 3' UTR shortening in lung cancer cells

Having established that our method reproduces previously reported patterns of 3' UTR length change in physiological settings, we then turned to the question of whether 3' UTRs are also different in lung cancer cells compared to their non-malignant counterpart, the alveolar epithelial cells. We identified 1'330 TEs that were shorter in the 3'607 cancer compared to the 851 alveolar cells in the Lambrechts dataset (with 22 and 3.7 million reads in TEs respectively), representing 98% of 1'357 significant events (Figure 4A, top). Similarly, we identified 188 shortened TEs from the Laughney dataset of 489 cancer and 292 alveolar cells (with 6 and 1.3 million reads in TEs respectively), representing 85% of 219 significant events (Figure 4A, bottom). While much fewer events were found in the Laughney data set, the majority (105 of 188 TEs, 56%) were shared with the Lambrechts dataset. To determine whether specific biological processes are subject to APA-dependent regulation in cancer cells, we submitted the set of 105 shared genes to functional analysis via the STRING web server (Szklarczyk et al., 2019). This revealed that the corresponding proteins are associated with membranes, vesicles and granules (Figure 4B,C). Interestingly, these APA targets cover the entire lifecycle of membrane and secreted proteins, from synthesis (i.e. translation initiation factors and ribosomal proteins), to traffic into the ER (e.g. *SSR1*, *SPCS3*, *SEC63*) and Golgi (e.g. *TRAPPC3*, *KDELR2*), to proteasome-mediated degradation (*PSMD12*). Some of the APA targets are surface receptors with well-known involvement in cancers (*CD44*, *CD47* and *CD59*). These results indicate that APA contributes to the orchestration of protein metabolism and traffic in cancer cells. Examples of TEs from Figure 4B are shown in Figure 4D.



Biological Processes
Transport
Negative regulation of protein processing
Vesicle-mediated transport
Secretion by cell
Regulated exocytosis

Cellular Component	
Cytoplasm	Whole membrane
Endoplasmic reticulum	Golgi-associated vesicle membrane
Secretory granule	Endomembrane system
Secretory granule membrane	Organelle subcompartment
Cytoplasmic vesicle	Cytoplasmic vesicle membrane

Figure 4: APA in lung adenocarcinoma cells.

A. Scatter plot of AUC in function of log10(mean CPM) for cancer and alveolar cells in the Lambrechts (top) and Laughney (bottom) datasets. TEs with significant APA-induced length changes are highlighted in green (numbers shown in insets). **B.** The interaction network (from the STRING web server) of proteins whose transcripts undergo 3'UTR shortening in both datasets. **C.** Functional enrichment analysis for genes whose TEs undergo shortening in cancer cells. Shown are the top 10 GO biological process terms (sorted by the false discovery rate, FDR). Analysis was performed with STRING web server, using as background the set of genes found to be expressed in the lung samples. **D.** Read coverage along TEs for a few example genes from panel B (*EIF1*, *CD44* and *CD59*). Each panel shows four tracks per data set, blue: cancer cells, red: alveolar cells, coverage of the TE by reads (top track) and the cumulative coverage of the TE by read 3' ends (bottom track). In all cases, the 3' UTRs are shorter in cancer compared to alveolar cells.

A. 3. 5. Conserved targets of 3' UTR shortening in individual cell types

The next question we wanted to answer is whether 3' UTR shortening affects all cells in the tumor environment, or it is rather restricted to specific cell types. We thus carried out the SCUREL analysis for each individual cell type for which we had at least ~20 cells in each data set, comparing TE lengths between cells of the same type, from the tumor sample and matched control sample. We found many more TEs becoming significantly shorter than longer (Fig. 5A-B), across almost all cell types and in both data sets. This is summarized in Fig. 5C, which shows that the proportion of shortened among significantly changed TEs is almost always greater than 0.5. By grouping all reads from the tumors and from matched control samples, respectively, we also recapitulated the result of previous 'bulk' RNA-seq data analyses (Fig. 5D). Thus, 3’ UTR shortening is not restricted to a specific cell type, but seems to generally take place in all cell types, associated with the tumor environment.

Moreover, in spite of the differences between the studies, there was a highly significant overlap between the targets of TE shortening in individual cell types (Fig. 5E-F). To gain further insight into the processes that may be regulated by APA, we submitted the intersection sets of genes exhibiting TE shortening in T lymphocytes and myeloid cells in these studies to functional enrichment analysis. We found significant enrichments especially in cellular components such as membranes, vesicles and granules (Fig. 5G-H), similar to what we observed in cancer cells.

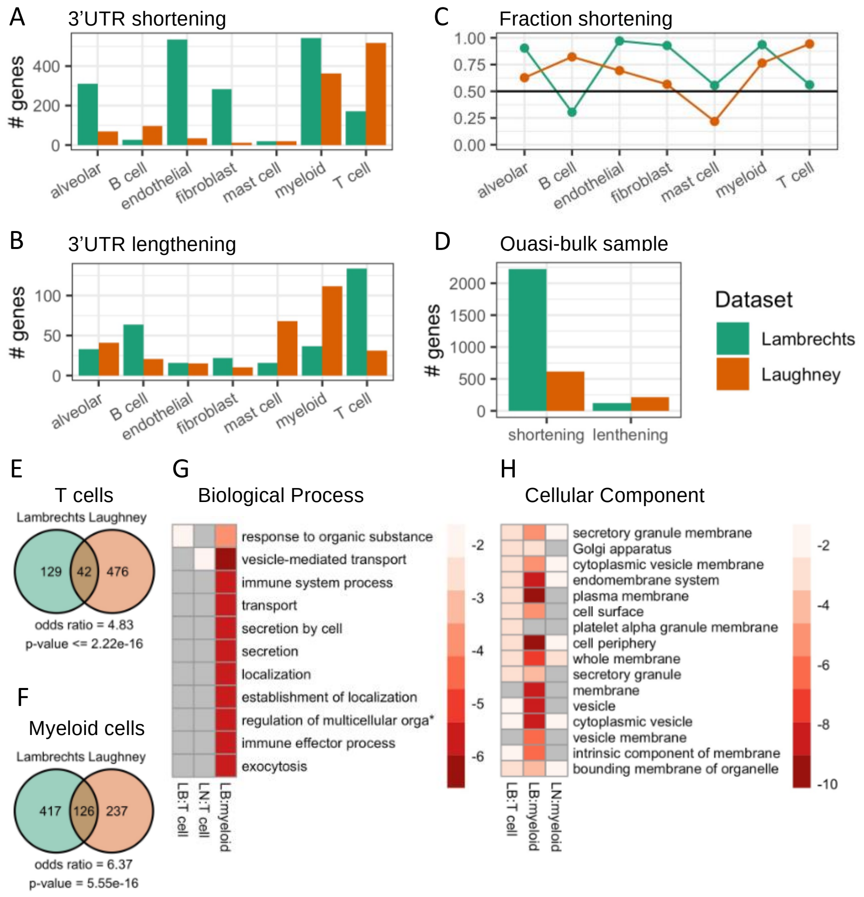


Figure 5. APA events in individual cell types.

A. Number of genes with APA-associated 3’UTR shortening in the Lambrechts (green) and Laughney (orange) data sets. **B.** Number of genes with APA-associated 3’UTR lengthening, same colors as A. **C.** Fraction of 3’UTR shortening events in individual cell types, among all significant events. **D.** Number of genes whose TEs undergo significant length change in quasi-bulk samples, shortening and lengthening events being shown separately. **E.** Venn diagram of TE shortening events in T cells from the two studies. Calculation of odds ratio and p-value of overlap with hypergeometric distribution (see Methods). **F.** Similar for myeloid cells. **G.** Biological process enrichment for TEs undergoing significant shortening in T cells and myeloid cells from the Lambrechts (LB) and Laughney (LN) studies. No process was specifically enriched in myeloid cells from the Laughney dataset. Plot generated with *pheatmap* (v 1.0.12). **H** Cellular component enrichment for TEs undergoing significant shortening in T cells and myeloid cells from the two studies. No component was specifically enriched in T cells from the Laughney dataset. Plot generated as in G.

A. 3. 6. Variability in 3’ UTR shortening among individuals

Finally, we asked to what extent are the targets of 3’ UTR shortening similar across patients. To answer this question, we analyzed individually the cells obtained from three patients in the Lambrechts study. Interestingly, in spite of the similar histopathological classification of the samples, one of the three samples was markedly different from the others, not exhibiting any tendency towards 3’ UTR shortening (Fig. 6A-D). The other two samples showed highly significant overlaps between shortened 3’ UTRs in different cell types (Fig. 6E). Analysis of biological process enrichment in individual cell types based on the genes targeted in both of these patients reinforced the concept that transport processes are affected in multiple cell types (Fig. 6F). It also provided further granularity. For example, leukocyte activation and secretion are terms enriched in the myeloid cell data, whereas metabolic processes are enriched in T cells, interaction with immune cells in endothelial cells and interaction with endothelial cells and angiogenesis in fibroblasts. Altogether these data demonstrate the power of SCUREL identifying changes in APA-related changes in 3’ UTR length, revealing common functional themes, in spite of substantial variability between samples. A complete table of genes with significant 3’UTR shortening across all LUAD comparisons we conducted is available in Suppl. Table 1. The data further indicate that protein transport processes and intercellular communication are preferential targets of APA across multiple cell types.

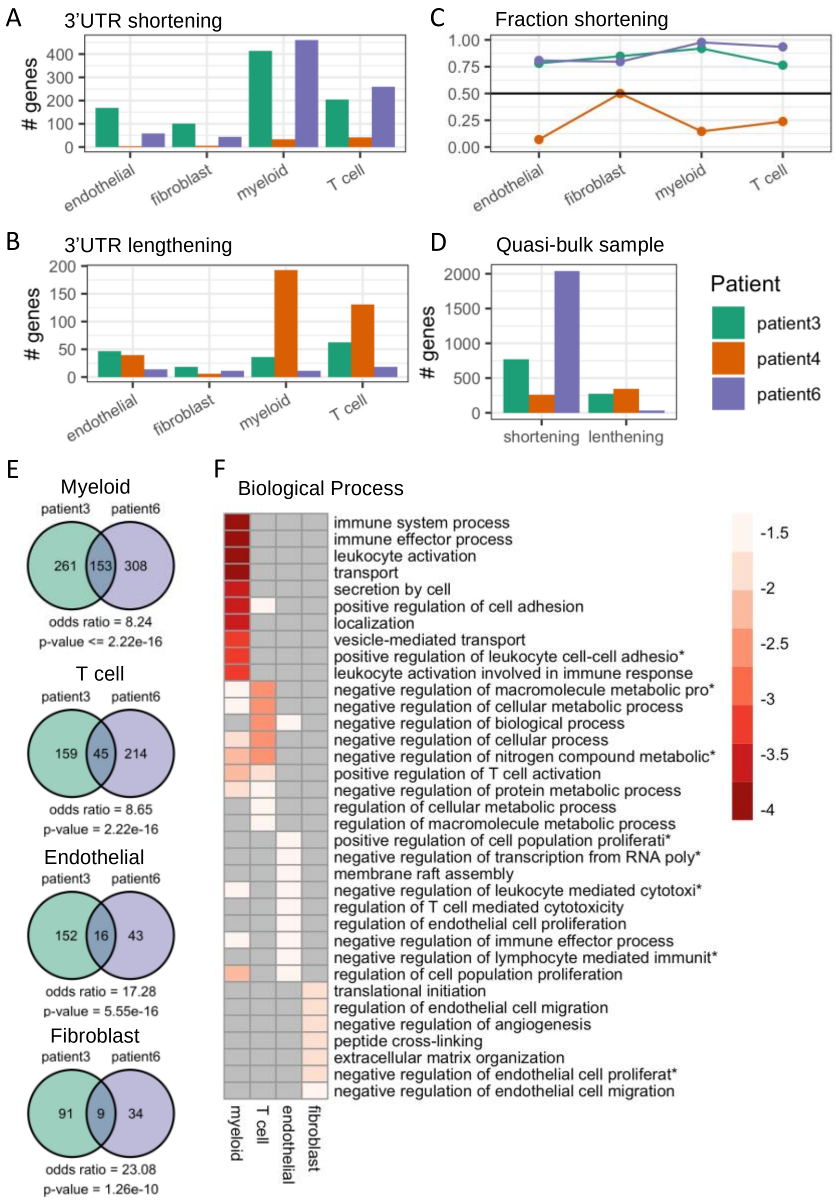


Figure 6. APA events in individual cell types from individual patients.

A. Number of genes with 3'UTR shortening inferred from patient 3 (green), patient 4 (orange) and patient 6 (purple) samples from the Lambrechts dataset. **B.** Number of genes with 3'UTR lengthening, same colors as A. **C.** Fraction of 3'UTR shortening events in individual cell types, among all significant events. **D.** Number of genes whose TEs undergo significant length change in quasi-bulk samples, shortening and lengthening events being shown separately. **E.** Venn diagrams of significantly shortened TEs in myeloid, T, endothelial and fibroblast cells from tumor relative to matched control samples from distinct patients. Calculation of odds ratio and p-value of overlap with hypergeometric distribution (see Methods). **F.** Biological process enrichment for TEs found to be shorter in cancer compared to matched control cells of individual cell types, from patient 3 and patient 6. Plot generated with *pheatmap* (v 1.0.12).

A. 4. Discussion

The remodeling of gene expression in cancers involves, among other processes, alternative polyadenylation. A tendency toward 3' UTR shortening has been generally observed, though to different extents, in virtually all studied cancers (Schmidt et al., 2018; Xia et al., 2014). Whether this is the result of changes in the cell type composition of the tissue or to cancer-related changes in functionality in all cell types has not been investigated so far. We set out to answer this question, taking advantage of single cell sequencing data sets obtained from human lung adenocarcinoma. As the sparsity of the scRNA-seq data poses some challenges (Lähnemann et al., 2020) we sought two distinct studies that used the same sequencing platform, to identify shared patterns of variation. Furthermore, we developed an approach that controls for both imperfect annotation of transcript isoforms and low read coverage in scRNA-seq.

Comparing data from cells of the same type, but originating either from tumor samples or from matched control tissue, we found similar tendencies towards 3' UTR shortening in the tumor environment for most cell types. Furthermore, the proteins encoded by the transcripts that are affected in various cell types cluster into specific functional classes, specifically the synthesis, traffic, secretion and degradation of proteins. This implicates APA in the regulation of protein metabolism and the organization of subcellular structure.

Initial studies that described the phenomenon of 3' UTR shortening in T cells and cancer cells proposed a role in the regulation of protein levels, as short 3' UTR isoforms are more stable than those with long 3' UTRs (Mayr & Bartel, 2009; Sandberg et al., 2008). However, when the decay rates of 3' UTR isoforms were measured, they turned out to be rather similar (A. R. Gruber et al., 2014; Spies et al., 2013), leaving open the question of functional differences between 3' UTR isoforms (Mayr, 2018). More recent work uncovered additional layers of 3' UTR-mediated regulation. For example, a role of 3' UTRs in the localization of the translated protein (UDPL) has been described for a number of membrane proteins, including the immunoglobulin family member CD47, whose localization to the cell membrane protects host cells from phagocytosis by macrophages (Berkovits & Mayr, 2015). Interestingly, *CD47* is a conserved APA target in both LUAD datasets that we analyzed here, its 3' UTR becoming shorter in cancer cells compared to lung alveolar cancer cells. This would predict decreased localization of CD47 to the surface of cancer cells, making them more susceptible to apoptosis compared to normal alveolar cells. This may explain why increased levels of CD47 are associated with increased cancer-free survival of patients with lung cancers (kmplot.com, (Nagy et al., 2021)) . It will be very interesting to apply methods for simultaneous profiling of protein and mRNA expression in single cells (Stoeckius et al., 2017) to better understand the interplay between APA, gene expression, and membrane localization of CD47 in cancers.

The concept that 3' UTR shortening is associated with proliferative states was challenged in a recent study that instead demonstrated its association with the secretion of proteins, both in trophoblast and in plasma cells (Cheng et al., 2020). Our data fully support this notion, extending the data to cancer cells as well as T lymphocytes and myeloid cells. As the protein production apparatus is present in all cells, APA is a well-suited mechanism for fine-tuning the expression of various components in a cell type and cell state-dependent manner (Lianoglou et al., 2013). Associating APA with protein metabolism rather than cell proliferation makes the question of its upstream regulation ever more puzzling because the shortening of 3' UTRs in proliferating cells has been attributed to an increased expression of 3' end processing factors mediated by cell cycle-associated E2F transcription factors (Elkon et al., 2012). It will be interesting to revisit this issue in a system where the increased protein production and secretion can be decoupled from cell proliferation, as the B cell maturation system (Cheng et al., 2020).

In conclusion, among the many applications of scRNA-seq, analysis of cell type-dependent polyadenylation reveals the relevance of APA as a general mechanism for regulating the metabolism and traffic of proteins within cells. With SCUREL we provide a robust method for detecting changes in 3' UTR length for even low-expression genes between cell types, in a manner that does not rely on accurate PAS annotation.

A. 5. Materials and Methods

A. 5. 1. Datasets

Lung cancer samples

Lung adenocarcinoma (LUAD) and matched control samples were downloaded from the GEO database (Barrett et al., 2013), based on the accession numbers in the original publications. Specifically, from the (Lambrechts et al., 2018; Szklarczyk et al., 2019) data set we used the LUAD samples listed in Table 1 of the original publication (corresponding to patients 3, 4 and 6, 3 tumor samples and one matched control sample for each patient). scRNA-seq data (ArrayExpress (Athar et al., 2019) accession numbers E-MTAB-6149 and E-MTAB-6653) were generated in this study with the 10x Genomics Single Cell 3' V2 protocol. From the (Laughney et al., 2020) study we also used LUAD and matched control samples, which originated from 3 donors. These samples were also generated with the 10x Genomics Single Cell 3' V2 protocol (accession number GSE123904).

Mouse testis samples

scRNA-seq data from the testes of two 8-week old C57BL/6J mice (Lukassen et al., 2018) were downloaded from the GEO database (accession number GSE104556).

Mouse T cell samples

scRNA-seq data of FACS sorted T cells from the lymph nodes and spleen of C57BL/6J mice, three infected with OVA-expressing *Lysteria monocytogenes* and one naive (Pace et al., 2018) were downloaded from the GEO database (accession number GSE106268).

A. 5. 2. Execution of scAPA

scAPA (Shulman & Elkon, 2019) was downloaded from the github repository and executed with the same genome sequence that was used throughout the study. For compatibility, the “chr” prefix in the chromosome names was removed. The lengths of the chromosomes were obtained with *samtools faidx*. The *homer* software (v4.11.1) required by the scAPA package was manually downloaded from http://homer.ucsd.edu/homer/. We collected all other requirements specified on scAPA github page in a conda environment. The removal of duplicate reads was done by adjusting the existing *umi_tools dedup* command in *scAPA.shell.script.R* for 10X Genomics, using the following options " --per-gene", " --gene-tag=GX", " --per-cell ". This was necessary because according to the protocol, one RNA fragment could result in reads that do not map at identical positions.

A. 5. 3. Extraction of terminal exons

Terminal exons were obtained from the RefSeq genome annotations (gff), GRCm38.p6 for mouse and GRCh38.p13 for human, with a custom script, as follows. Chromosome names from the RefSeq assembly were converted to ENSEMBL-type names based on the accompanying ‘assembly_report.txt’ file. Only autosomes, allosomes and mitochondrial DNA were retained. Based on the genome annotation file, protein-coding and long non-coding transcripts were retained, while model transcripts (‘Gnomon’ prediction; accession prefixes XM_, XR_, and XP_) were discarded. From this transcript set, the 3’-most exons (i.e. terminal exons, TEs) were retrieved. Overlapping TEs on the same chromosome strand were clustered with *intervaltree* (v3.0.2; python package) and from each cluster, the longest exon was kept. The resulting set of TEs was sorted by chromosome and start position and saved to a BED-formatted file. TE IDs were converted to gene names with *biomaRt* (v 2.46.3) using the ensembl BioMart database. Duplicate gene names were discarded.

A. 5. 4. Processing of scRNA-seq reads

The workflow can start from mapped reads in *cellranger*-compatible format, a file with cell barcode-to-cell type annotation and a genome annotation file. Alternatively, the *cellranger count* function can be used to map reads from FASTQ input data. Reads from the FASTQ files were mapped with the function *count* from the *cellranger* (v5.0.0) package to the reference human genome GRCh38-3.0.0 sequence obtained directly from *10X genomics* website. This genome is a modified version of the GRCh38 genome, compatible with the cellranger analysis pipeline. Reads are also aligned to the transcriptome. In this step, cell barcodes and UMIs correction also takes place. Aligned reads (BAM) with mapping quality (MAPQ) scores > 30 were selected with *samtools* (v1.12, (Li et al., 2009)). Reads without a cell barcode “CB” tag were removed with *samtools view*, as were duplicated reads using *umi_tools dedup* (v1.1.1, (Smith et al., 2017)). The mapped reads are filtered, deduplicated and grouped by cell type in the “cell type” mode or by cell type and tissue of origin in the “condition” mode. In the latter case, quasi-bulk samples are also constructed from the filtered reads that come from individual conditions.

A. 5. 5. Cell type annotation

The annotation of cell types in all datasets was carried out with the approach described in (Lambrechts et al., 2018). Filtered data (so as to remove artifacts such as empty droplets) consisting of cellular barcodes and count matrices from individual data sets were loaded in R (v4.0.3) with *Read10X* (from *Seurat* v3.2.3 (Butler et al., 2018)), and Seurat objects were created with *CreateSeuratObject*. For the lung cancer datasets, cells with < 201 Unique Molecular Identifiers (UMIs), with < 101 or > 6000 genes or with > 10% UMIs from mitochondrial genes (which may indicate apoptotic or damaged cells) were removed. For all datasets, genes with zero variance across all cells (i.e. sum = 0) were discarded. The gene expression counts for each cell were log-normalised with *NormalizeData* with a default scale factor of 10’000. In Seurat, 2’000-2’500 most variable genes are used to cluster the cells. Here we used the 2’192 variable most variable genes, as in (Lambrechts et al., 2018). These were selected with *FindVariableFeatures*, with normalised expression between 0.125 and 3, and a quantile-normalised variance exceeding 0.5 for lung cancer and mouse T cell samples, and normalised expression between 0.1 and 8 for mouse testis samples. Gene expression levels were then centered and scaled across all cells. After Principal Component Analysis (PCA) on the most variable genes, the number of relevant dimensions *n* for each data set was determined by assessing the variance explained by individual Principal Components (PC) with *ElbowPlot* from Seurat. UMAP (McInnes et al., 2018) was used to visualize the data projected on the *n* dimensions. For T cell activation and LUAD samples, batch correction and data integration were performed with *harmony* (v1.0) (Korsunsky et al., 2019). *Harmony* was run on the first 30 PCs and set to group by dataset. The transformed data set was used for downstream analysis (i.e. clustering of cells, visualization in 2D).

Various Seurat functions were used to identify the cell type of individual cells. Cells were clustered using the Shared Nearest Neighbor (SNN) algorithm, which aims to optimize modularity. First, *FindNeighbors* was executed using the first *n* dimensions from PCA or *harmony* and with otherwise default settings (k = 20). Then, *FindClusters* with resolution parameter 0.6 for LUAD, 0.2 for T cells and 0.3 for spermatocytes was run, so as to retrieve a number of clusters similar to those in the original publications. The expression of cell type markers in each cluster was assessed with *FindAllMarkers*. This function finds genes that are differentially expressed between cells from one cluster and all other cells, by applying a Wilcoxon Rank Sum test on the log-normalized expression. Individual clusters were downsampled to the number of cells in the smallest cluster or to at least 100 cells. Only genes expressed in a minimum of 10% of the cells in either population and with a log (base *e*) -fold-change of at least 0.25 (default values in *Seurat*) were tested. Markers with adjusted p-value < 0.01 were considered significant and those with higher expression in the selected cluster were considered as potential markers for that cell cluster. For each cluster we counted the number of significant markers that matched known cell type markers (Table 1) and assigned the cell type to be the one for which a proportion of > 0.6 of known markers were specifically expressed in the cell cluster. Generally, this assignment was unambiguous, and when it was not, the cell type assignment was done manually, taking into account the adjusted p-value and average log-fold-change of all considered marker genes as well as the cell type annotation from the Suppl. Table 3 of (Lambrechts et al., 2018), which contains additional cell type markers. At least 3 marker genes were required to assign a cluster to the corresponding cell type, except for cancer cells that were annotated only based on the expression of EPCAM.

Table 1: **Marker genes for cell type annotation.** Based on (Lambrechts et al., 2018).

Cell type	**marker genes**	**Cell type**	**marker genes**
alveolar	*CLDN18*	fibroblast	*C1R*
alveolar	*FOLR1*	B cell	*CD79A*
alveolar	*AQP4*	B cell	*IGKC*
alveolar	*PEBP4*	B cell	*IGLC3*
endothelial	*CLDN5*	B cell	*IGHG3*
endothelial	*FLT1*	myeloid	*LYZ*
endothelial	*CDH5*	myeloid	*MARCO*
endothelial	*RAMP2*	myeloid	*CD68*
epithelial	*CAPS*	myeloid	*FSGR3A*
epithelial	*TMEM190*	T cell	*CD3D*

epithelial	*PIFO*	T cell	*TRBC1*
epithelial	*SNTN*	T cell	*TRBC2*
fibroblast	*COL1A1*	T cell	*TRAC*
fibroblast	*DCN*	cancer	*EPCAM*
fibroblast	*COL1A2*		

A. 5. 6. Assessing 3’ UTR length differences with the AUC measure

To assess changes in 3’ UTR length between groups of cells we used the following approach. For simplicity, the analysis is carried out for terminal exons (TEs) rather than 3’ UTRs, as 3’ UTRs are generally contained in TEs, covering almost the entire length of the TEs. We started from the BAM files of mapped reads from two groups of cells. We computed the 3’ end coverage of individual TEs per strand with *bedtools genomecov* and parameter “-bga”. The BED file with read 3’ end positions was used to obtain the normalized reverse cumulative coverage of individual TEs, i.e. starting at the TE 3’ end and ending at the 5’ most nucleotide. Individual TEs were traversed from the end to the beginning, recording the reverse cumulative coverage in the two groups of cells as a function of position. The area under the resulting curve (AUC) was then calculated. An AUC of 0.5 corresponds to identical position-dependent coverage of the TE by 3’ end reads in the two groups of cells, i.e. no difference in TE length. An AUC value of 1 corresponds to all the 3’ end reads from the group of cells indicated on the y-axis being clustered at the end of the TE, before any reads from the other group are observed, thus the TEs are longest in this group of cells. Vice versa, an AUC value of 0 corresponds to all the 3’ end reads from the group indicated on x-axis are observed before any reads of the other group, thus the 3’UTRs are longest in this group of cells.

If the read coverage of a TE is very sparse, the curve representing the coverage in the two cell groups will not be smooth, but rather change in steps of 1/*n* where *n* is the number of reads mapping to the TE; deviations from the diagonal line of identical coverage in the two groups will be common, due to the stochastic sampling of the reads. To mitigate this effect and identify TEs whose coverage cannot be explained by stochastic sampling of low-expression genes we generated a background dataset, in which we randomized the cell group label of the reads. This procedure preserves the number of reads obtained in each TE in each group, but randomizes their position in the TE.

Finally, we identified TEs with AUCs indicating significant shifts in PAS usage. For this, we extracted TEs with a normalized read count (CPM) >= 2 in both cell groups, roughly corresponding to TEs with at least one count in each of the groups. As AUC values depend on the overall expression of the TE, we used an expression-dependent AUC cutoff to identify the TEs significantly changing length. This corresponded to the two-tailed 1% quantile of the background distribution in each of the 20 equal-sized log(mean expression between cell groups) bins, smoothened using the median over a running window of 5 values. Finally, to ensure that the change in read coverage was due to APA, we only retained significantly changed TEs for which the union of the interquartile range of TE positions that were covered by 3’ end reads in the two samples spanned at least 200 nucleotides.

A. 5. 7. Analysis of overlaps between data sets

We used a sample-specific background for the calculation of the probability of overlap of genes and for the pathway enrichment analysis carried out on the STRING web server. All TEs considered in the AUC analysis, i.e. TEs with CPM >= 2, in each sample were combined and the unique set of TEs was used as background. In particular, for the cell type analysis of the Lambrechts dataset, we used the cell type-specific union of TEs from patients 3, 4 and 6 and obtained 10'966 genes for myeloid cells, 10'473 for T cells, 11'269 for endothelial cells and 11'857 for fibroblasts. For the cell type analysis of lung cancer datasets, the union of TEs consisted of 10'177 genes in T cells and 9'970 genes in myeloid cells. We used the hypergeometric distribution to calculate the odds ratio and associated p-value of the overlap between gene sets.

A. 5. 8. Pathway analysis

The gene symbols for TEs with significant APA events were analyzed via the STRING web server, which provides enriched Gene ontology (GO) terms, KEGG and reactome pathways. As a background gene set for the enrichment analysis we provided the dataset-specific list of expressed genes (CPM >= 2).

A. 5. 9. Workflow execution

SCUREL was packaged in Snakemake and can be obtained from https://github.com/zavolanlab/SCUREL.

A. 6. Acknowledgements

We thank the tissue donors and the authors of the single cell studies (Lambrechts et al., 2018; Laughney et al., 2020) for providing the data. We thank Erik van Nimwegen for the extensive and fruitful discussions on measures of 3' UTR shortening. We also thank members of the Zavolan lab and Dr. Sebastian Leidel for constructive comments. We especially acknowledge the help of Arka Banerjee in the cell marker identification and Dr. Jeremie Breda for helpful discussions regarding the AUC measure. We are grateful to the sciCORE team for the maintenance of the computing infrastructure on which the computations were carried out. This work was supported by the Swiss National Science Foundation grant #310030_189063 to M.Z. D.B. is a recipient of a Fellowship for Excellence from the Biozentrum, University of Basel.

A. 7. References

Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N. A., Petryszak, R., Papatheodorou, I., Sarkans, U., & Brazma, A. (2019). ArrayExpress update — from bulk to single-cell expression data. *Nucleic Acids Research*, *47*(D1), D711–D715.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*(Database issue), D991–D995.

Berkovits, B. D., & Mayr, C. (2015). Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature*, *522*(7556), 363–367.

Breda, J., Zavolan, M., & van Nimwegen, E. (2021). Bayesian inference of gene expression states from single-cell RNA-seq data. *Nature Biotechnology*. https://doi.org/10.1038/s41587-021-00875-x

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420.

Cheng, L. C., Zheng, D., Baljinnyam, E., Sun, F., Ogami, K., Yeung, P. L., Hoque, M., Lu, C.-W., Manley, J. L., & Tian, B. (2020). Widespread transcript shortening through alternative polyadenylation in secretory cell differentiation. *Nature Communications*, *11*(1), 3182.

Dwyer, D. F., Barrett, N. A., Austen, K. F., & Immunological Genome Project Consortium. (2016). Expression profiling of constitutive mast cells reveals a unique identity within the immune system. *Nature Immunology*, *17*(7), 878–887.

Elkon, R., Drost, J., van Haaften, G., Jenal, M., Schrier, M., Vrielink, J. A. O., & Agami, R. (2012). E2F mediates enhanced alternative polyadenylation in proliferation. *Genome Biology*, *13*(7), R59.

Gruber, A. J., Gypas, F., Riba, A., Schmidt, R., & Zavolan, M. (2018). Terminal exon characterization with TECtool reveals an abundance of cell-specific isoforms. *Nature Methods*, *15*(10), 832–836.

Gruber, A. J., Schmidt, R., Ghosh, S., Martin, G., Gruber, A. R., van Nimwegen, E., & Zavolan, M. (2018). Discovery of physiological and cancer-related regulators of 3' UTR processing with KAPAC. *Genome Biology*, *19*(1), 44.

Gruber, A. R., Martin, G., Müller, P., Schmidt, A., Gruber, A. J., Gumienny, R., Mittal, N., Jayachandran, R., Pieters, J., Keller, W., van Nimwegen, E., & Zavolan, M. (2014). Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nature Communications*, *5*, 5465.

Ji, Z., & Tian, B. (2009). Reprogramming of 3' untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PloS One*, *4*(12), e8419.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wie, K., Baglaenko, Y., Brenner, M., Loh, P.-R., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, *16*(12), 1289–1296.

Koster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. In *Bioinformatics* (Vol. 28, Issue 19, pp. 2520–2522). https://doi.org/10.1093/bioinformatics/bts480

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. de, Cappuccio, A., … Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, *21*(1), 31.

Lambrechts, D., Wauters, E., Boeckx, B., Aibar, S., Nittner, D., Burton, O., Bassez, A., Decaluwé, H., Pircher, A., Van den Eynde, K., Weynand, B., Verbeken, E., De Leyn, P., Liston, A., Vansteenkiste, J., Carmeliet, P., Aerts, S., & Thienpont, B. (2018). Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature Medicine*, *24*(8), 1277–1289.

Laughney, A. M., Hu, J., Campbell, N. R., Bakhoum, S. F., Setty, M., Lavallée, V.-P., Xie, Y., Masilionis, I., Carr, A. J., Kottapalli, S., Allaj, V., Mattar, M., Rekhtman, N., Xavier, J. B., Mazutis, L., Poirier, J. T., Rudin, C. M., Pe'er, D., & Massagué, J. (2020). Regenerative lineages and immune-mediated pruning in lung cancer metastasis. *Nature Medicine*, *26*(2), 259–269.

Lee, S., Wei, L., Zhang, B., Goering, R., Majumdar, S., Wen, J., Taliaferro, J. M., & Lai, E. C. (2021). ELAV/Hu RNA binding proteins determine multiple programs of neural alternative splicing. *PloS Genetics*, *17*(4), e1009439.

Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S., & Mayr, C. (2013). Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & Development*, *27*(21), 2380–2396.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , *25*(16), 2078–2079.

Lukassen, S., Bosch, E., Ekici, A. B., & Winterpacht, A. (2018). Characterization of germ cell differentiation in the male mouse through single-cell RNA sequencing. *Scientific Reports*, *8*(1), 6521.

Martin, G., Gruber, A. R., Keller, W., & Zavolan, M. (2012). Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Reports*, *1*(6), 753–763.

Masuda, A., Kawachi, T., Takeda, J.-I., Ohkawara, B., Ito, M., & Ohno, K. (2020). tRIP-seq reveals repression of premature polyadenylation by co-transcriptional FUS-U1 snRNP assembly. *EMBO Reports*, *21*(5), e49890.

Mayr, C. (2018). What Are 3' UTRs Doing? *Cold Spring Harbor Perspectives in Biology*. https://doi.org/10.1101/cshperspect.a034728

Mayr, C., & Bartel, D. P. (2009). Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell*, *138*(4), 673–684.

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. In *Journal of Open Source Software* (Vol. 3, Issue 29, p. 861). https://doi.org/10.21105/joss.00861

Nagy, Á., Munkácsy, G., & Győrffy, B. (2021). Pancancer survival analysis of cancer hallmark genes. *Scientific Reports*, *11*(1), 6047.

Pace, L., Goudot, C., Zueva, E., Gueguen, P., Burgdorf, N., Waterfall, J. J., Quivy, J.-P., Almouzni, G., & Amigorena, S. (2018). The epigenetic control of stemness in CD8 T cell fate commitment. *Science*, *359*(6372), 177–186.

Patrick, R., Humphreys, D. T., Janbandhu, V., Oshlack, A., Ho, J. W. K., Harvey, R. P., & Lo, K. K. (2020). Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome Biology*, *21*(1), 167.

Reyes, A., & Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Research*, *46*(2), 582–592.

Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., & Burge, C. B. (2008). Proliferating Cells Express mRNAs with Shortened 3′ Untranslated Regions and Fewer MicroRNA Target Sites. *Science*, *320*(5883), 1643–1647.

Schmidt, R., Ghosh, S., & Zavolan, M. (2018). The 3′ UTR Landscape in Cancer. In *eLS* (pp. 1–9). https://doi.org/10.1002/9780470015902.a0027958

Shulman, E. D., & Elkon, R. (2019). Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic Acids Research*, *47*(19), 10027–10039.

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, *27*(3), 491–499.

So, B. R., Di, C., Cai, Z., Venters, C. C., Guo, J., Oh, J.-M., Arai, C., & Dreyfuss, G. (2019). 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.

Spies, N., Burge, C. B., & Bartel, D. P. (2013). 3′ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Research*. https://doi.org/10.1101/gr.156919.113

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., & Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, *14*(9), 865–868.

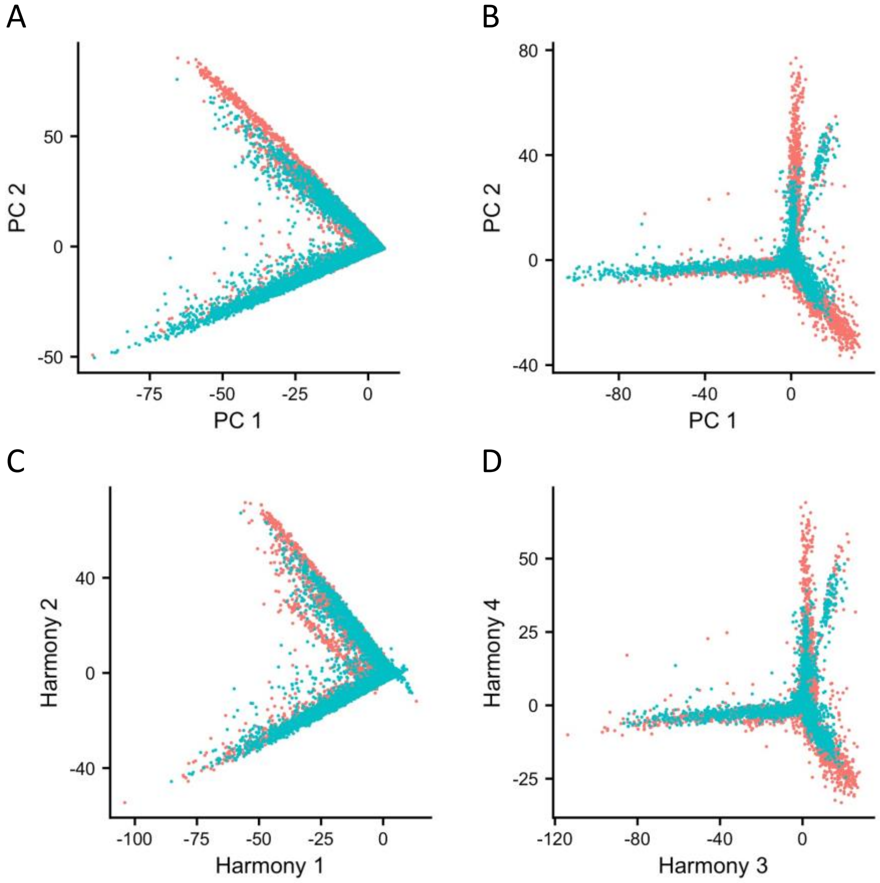
Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Mering, C. von. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613.

Wu, X., Liu, T., Ye, C., Ye, W., & Ji, G. (2020). scAPAtrap: identification and quantification of alternative polyadenylation sites from single-cell RNA-seq data. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbaa273

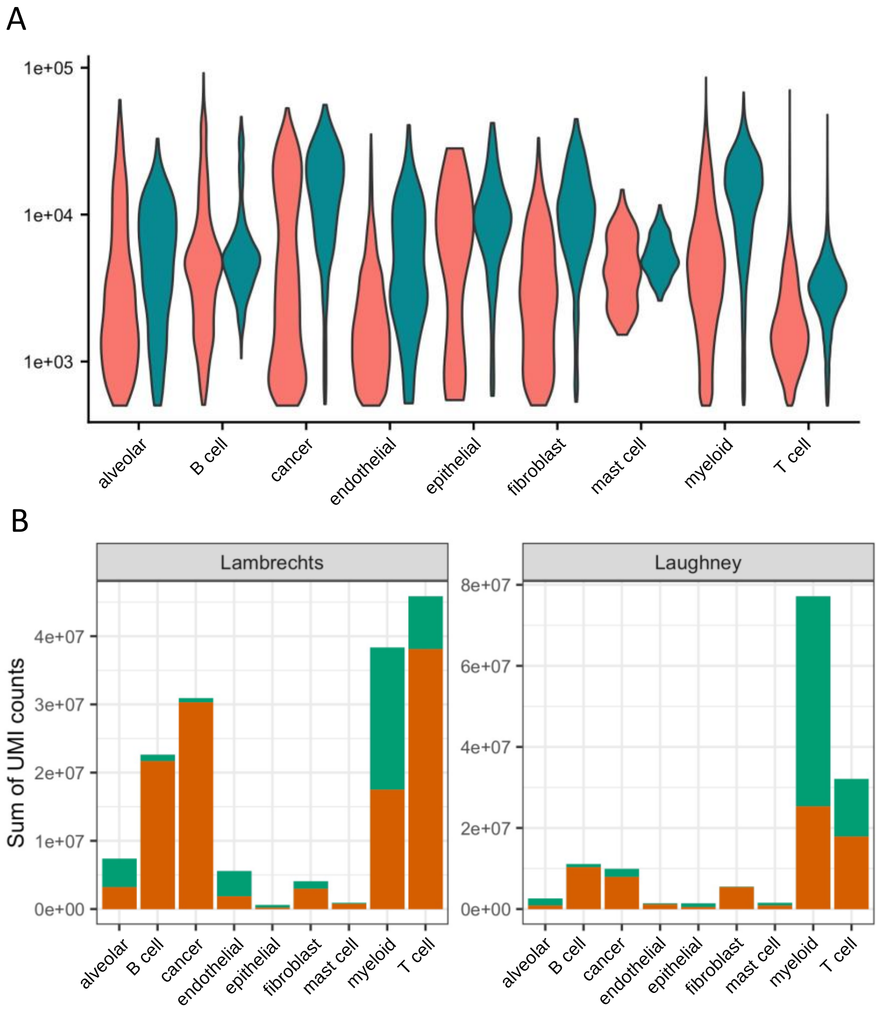
Xia, Z., Donehower, L. A., Cooper, T. A., Neilson, J. R., Wheeler, D. A., Wagner, E. J., & Li, W. (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nature Communications*, *5*, 5274.

Ye, C., Zhou, Q., Wu, X., Yu, C., Ji, G., Saban, D. R., & Li, Q. Q. (2020). scDAPA: detection and visualization of dynamic alternative polyadenylation from single cell RNA-seq data. *Bioinformatics*, *36*(4), 1262–1264.

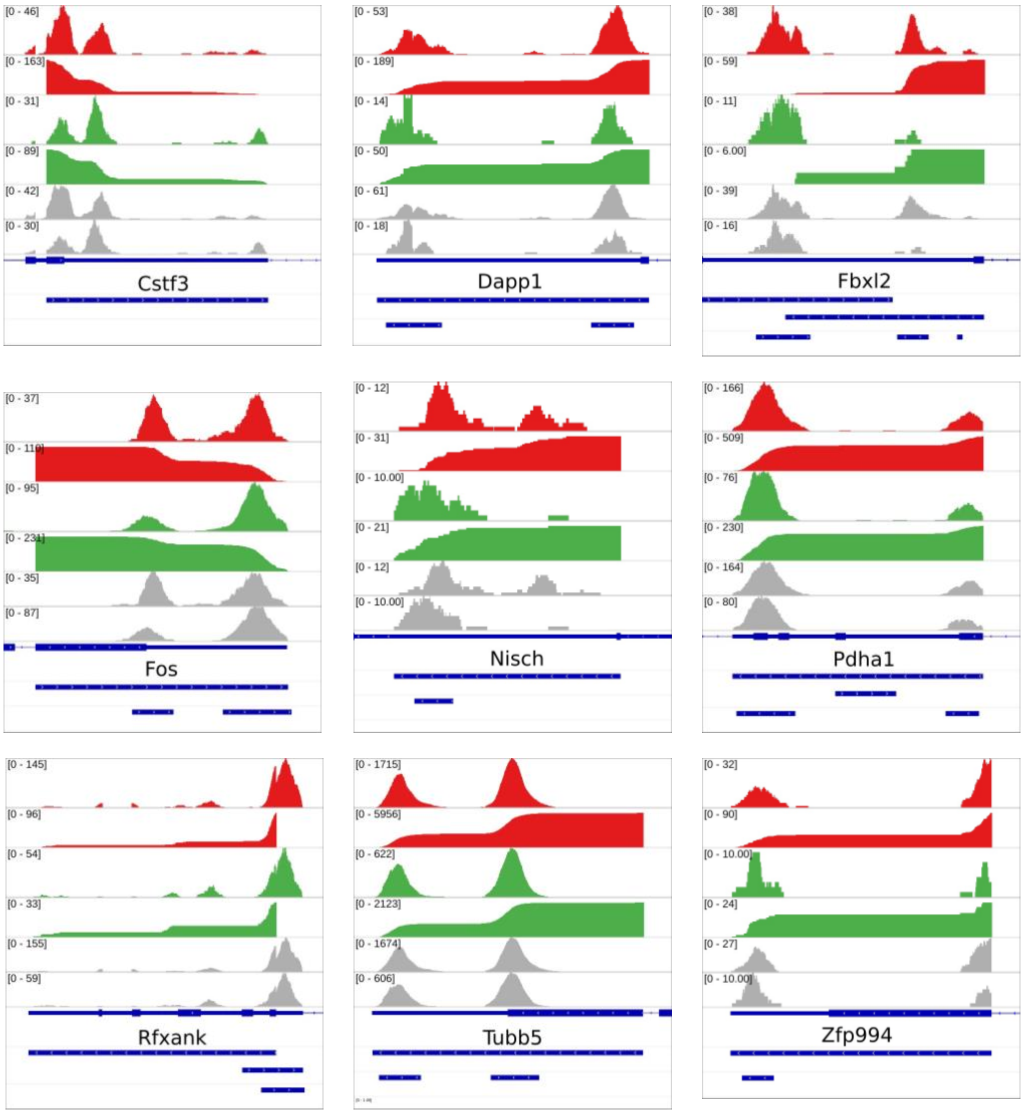
A. 8. Supplementary Figures



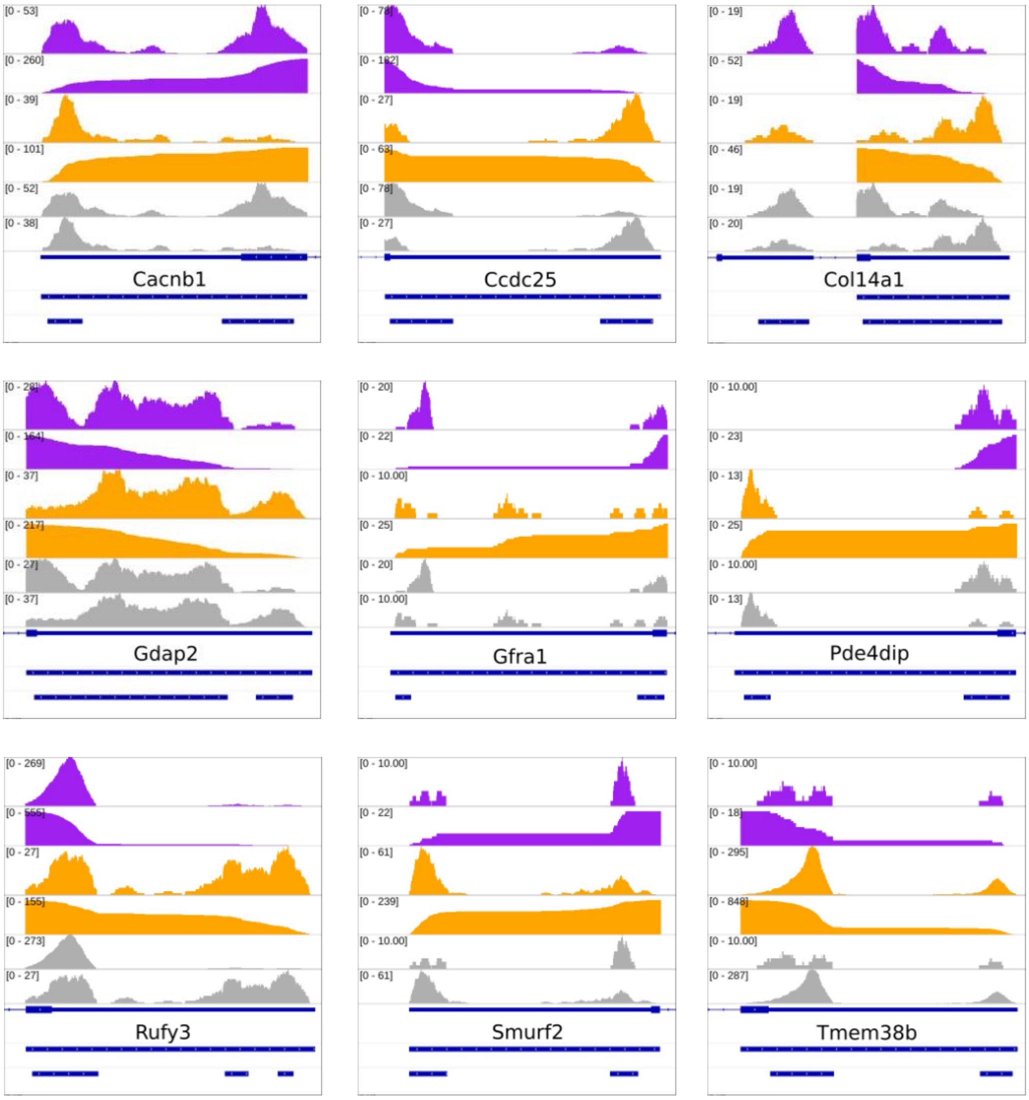
Supplementary Figure 1. Principal component analysis of the combined datasets from the (Lambrechts et al. 20218; Laughney et al. 2020) studies. **A, B.** Projection of the gene expression data from the 56'699 cells from the Lambrechts (red) and Laughney (blue) data sets, on the first two principal components, PC 1 and 2 (A), and on PC 3 and 4 (B). **C, D.** Projection of the gene expression data as above, but after dataset integration with *harmony*, for PC 1 and 2 (C) and PC 3 and 4 (D).



Supplementary Figure 2. Relative contributions of individual cell types to the total pool of mRNA reads. **A** Distributions of the number UMI counts per cell for all annotated cell types in the Lambrechts (red) and Laughney (blue) datasets. The graphs were generated with *Seurat*. **B** Total number of UMI counts contributed by individual cell types to tumor (red) and matched control (green) samples from the Lambrechts and Laughney data.



Supplementary Figure 3. Nine random examples of SCUREL (and not scAPA)-detected 3’UTR shortening events in activated T cells (red) versus naive T cells (green). Only the TE of each gene is shown. The strand on which genes are encoded is indicated by arrows in the TE track (> indicates the Watson and < the Crick strands).



Supplementary Figure 4. Nine random examples of SCUREL detecting 3’UTR shortening in elongating spermatids (purple) versus spermatocytes (orange). Only the TE of each gene is shown. The strand on which genes are encoded is indicated by arrows in the TE track (> indicates the Watson and < the Crick strands).

Appendix B

Identification of experimentally-supported poly(A) sites in single-cell RNA-seq data with SCINPAS

Youngbin Moon[1,2,†], Dominik Burri[1,2,†], Mihaela Zavolan[1,2,*]

[1]Computational and Systems Biology, Biozentrum University of Basel, Spitalstrasse 41, CH-4056 Basel, Switzerland

[2]Swiss Institute of Bioinformatics, Basel, Switzerland

[†]These authors contributed equally to this work

[*]Corresponding author mihaela.zavolan@unibas.ch

B. 1. Abstract

Alternative polyadenylation is a main driver of transcriptome diversity in mammals, generating transcript isoforms with different 3’ ends via cleavage and polyadenylation at distinct polyadenylation (poly(A)) sites. The regulation of cell type-specific poly(A) site choice is not completely resolved, and requires quantitative poly(A) site usage data across cell types. 3’ end-based single-cell RNA-seq can now be broadly used to obtain such data, enabling the identification and quantification of poly(A) sites with direct experimental support. We propose SCINPAS, a computational method to identify poly(A) sites from scRNA-seq datasets. SCINPAS modifies the read deduplication step to favor the selection of distal reads and extract those with non-templated poly(A) tails. This approach improves the resolution of poly(A) site recovery relative to standard software. SCINPAS identifies poly(A) sites in genic and non-genic regions, providing complementary information relative to other tools. The workflow is modular, and the key read deduplication step is general, enabling the use of SCINPAS in other typical analyses of single cell gene expression. Taken together, we show that SCINPAS is able to identify experimentally-supported, known and novel poly(A) sites from 3’ end-based single-cell RNA sequencing data.

B. 2. Introduction

The majority of genes in the human genome have multiple isoforms, most of which come from the use of alternative transcription start or polyadenylation sites (1). While the regulation of transcription initiation by transcription factors has been extensively studied, much less is known about the regulation of poly(A) site (PAS) choice (2, 3). Comprehensive and quantitative PAS usage data across cell types is essential for studying the PAS choice, and a variety of methods have been developed to obtain such data by specifically sequencing mRNA 3’ ends (2, 4). With the introduction of single-cell RNA sequencing (scRNA-seq) the scale and resolution of PAS choice analyses can be dramatically expanded, because the broadly used 10x Genomics technology targets the 3’-terminal fragments of mRNAs. Consequently, various studies have emerged, describing the polyadenylation landscape of various cell types (5–10). However, as the scRNA-seq reads are generated from the 5’ ends of terminal mRNA fragments, they do not typically reach into the poly(A) tails to directly define the PAS. These are inferred computationally by associating peaks in read coverage with putative PAS, which can and does lead to a loss of resolution in PAS identification. Moreover, analyses of PAS usage in scRNA-seq data invariably start from genome-mapped reads, once the pre-processing and the "deduplication" of the reads based on their unique molecular identifiers (UMIs) have been performed with standard tools like CellRanger (11) and UMI-tools (12). These tools were not developed with the specific intent of detecting and quantifying the usage of PAS, and therefore, they do not attempt to extract the reads that are most relevant for PAS analyses. To fill this gap, we have developed SCINPAS, a tool that modifies the pre-processing of scRNA-seq data to improve the extraction of reads that carry non-templated poly(A) tails and thus provide direct evidence for PAS usage. SCINPAS should be applicable to any dataset generated with a 3'-biased approach to increase the recovery of PAS from individual cells and cell types, and thus improve the understanding of PAS usage and 3' untranslated region (UTR) dynamics across cell types.

B. 3. Methods

B. 3. 1. Analyzed datasets

Dataset	Accession number	Sample	BAM File Size (GB)	Tissue
Tabula Muris Senis	NA	10X_P4_2	21.8	Liver
	NA	10X_P4_7	23.8	Spleen
	NA	10X_P7_4	17.8	Heart and Aorta
	NA	10X_P7_11	22.2	Thymus
	NA	10X_P7_14	18.8	Limb muscle
	NA	10X_P7_15	15.0	Limb muscle
T cell activation dataset	SRR6228889	10X_naive_1	5.7	Blood
	SRR6228891	10X_infected_1	7.0	Blood
	SRR6228892	10X_infected_2	6.8	Blood
	SRR6228895	10X_infected_3	5.8	Blood
Sperm cell development dataset	SRR6129050	10X_mouse_1	16.6	Germ line
	SRR6129051	10X_mouse_2	16.1	Germ line

Table 1. scRNA-seq datasets used in the study.

SRR: Sequence read archive run identifier, BAM: Binary Alignment Map, GB: Gigabyte, NA: Not Available. Single-cell RNA sequencing data of the *Tabula Muris Senis* dataset were downloaded from czb-tabula-muris-senis S3 Public Bucket (13). Single-cell RNA sequencing data of mouse CD8+ T cells - naïve and from Listeria monocytogenes infection (14) - as well as from mouse germ cells (15) were downloaded from the NCBI's GEO database (accession numbers GSE106264 and GSE104556 respectively). Table 1 provides further information on these samples.

B. 3. 2. Mapping reads to the genome

Alignments of reads-to-genome were obtained with the CellRanger software, as it provided only the primary, highest-accuracy alignments and did not discard reads that mapped to non-exonic regions.

For the *Tabula Muris Senis* datasets, we used the alignments already available at czb-tabula-muris-senis S3 Public Bucket (13), generated with CellRanger version 2.0.1 (11), using the GENCODE GRCm38 vM19 annotation (available from the same S3 Public Bucket).

For the T cell activation and sperm cell development datasets, we used the 10x Genomics CellRanger software version 5.0.0 (11) to map the reads to the CellRanger-provided genome assembly, which is a modification of the GENCODE GRCm38 vM23 assembly version of the mouse genome.

B. 3. 3. Read deduplication

A key step in the scRNA-seq data analysis is the identification and "deduplication" of reads that come from the PCR copies of the same initial mRNA. This is done based on the UMIs that are added during the cDNA synthesis step and then sequenced as part of read 2, while the 5' end of the mRNA fragment is captured in read 1, in a paired-end sequencing approach. In principle, reads carrying the same cell identifier and the same UMI should come from PCR copies of one mRNA molecule. However, mutations may be introduced in the UMIs during sample preparation and sequencing, so that distinct UMIs do not always imply distinct initial mRNAs. CellRanger corrects apparent sequencing errors in the molecule identifiers (UR tag), providing read barcodes (UB). Moreover, as the UMIs are very short, there is a small chance that two distinct mRNAs end up with the same UMI. The standard approach for read deduplication with the UMI-tools software uses the genome annotation, to collapse the reads that have the same UMI only if they fall inside one gene. This of course makes sense, since the reads should be derived from a unique initial mRNA, but it also means that reads that fall outside of annotated regions are not considered. Furthermore, UMI-tools is not optimized to extract the most distal and thus most likely to contain a poly(A) tail from among reads with the same UMI. As our goal is to identify PAS in as comprehensive a manner as possible, including those outside of annotated genes or exonic regions, we do not use the gene annotation for deduplication, but implemented a different pre-processing approach.

Determination of read spans

First, we investigated the span of the genome covered by reads that originated in the same cell (same cell barcode - CB tag, provided by CellRanger) and the same molecular identifier, not trying to correct errors in the molecular identifier (UR tag). We calculated the span of a set of reads as follows:

$$span\ of\ read\ set(CB, UR) = max(read\ end|CB, UR)\ - min\ (read\ start|CB, UR)$$

The start and end coordinates refer to the genomic coordinates of reads within the set with a specific (CB, UR) combination. For reads that spanned splice junctions (coming from

adjacent exons of spliced mRNA), only the most distal part of the mapped read was used to compute the span.
The distribution of spans had two distinct peaks, one at 100-1000 and the other at 10-100 million nucleotides. Only the first one corresponds to terminal fragment sizes that are generated in the experiments, while the second peak may correspond to cases where two distinct mRNAs ended up with the same UMI.

Read clustering
Based on these results, we restricted the deduplication to reads with the same (CB, UR) tag combination that covered a maximum span of 100’000 nucleotides. That is, we traversed the genome, adding reads to the 3’ end of a cluster for as long as the maximum cluster span was not reached. Once this happened, we initiated a new subcluster with a new subcluster tag (YB tag, Table 2). In the very unlikely case that reads originating in the same mRNA will be split into multiple clusters by this procedure, the identification of PAS will not be impacted, because only the distal cluster will contain reads with poly(A) tails.

UMI correction
Similar to CellRanger, we then corrected errors in the molecular identifiers, by merging clusters whose span overlapped, and whose UR tags differed in one nucleotide. The majority UR tag in a merged cluster was then taken as the UMI of all reads in the cluster.

Read selection
Finally, we chose the most distal read from each cluster, as this should come closest to the PAS, possibly covering part of the poly(A) tail. If a cluster contained reads mapping to both strands of the chromosome (as well as having the same CB and UMI tags), we applied deduplication only to reads corresponding to the majority strand. In case of an equal number of reads mapping to the positive and negative strands we chose arbitrarily those from the negative strand.

B. 3. 4. Alignment correction

Inspection of read-to-genome alignments indicated that there were some cases where the alignment program did not fully extend the mappable parts of the reads into regions of low nucleotide complexity. This resulted in unmapped (i.e. “soft-clipped”) regions of the reads that in fact matched the genome. As we rely on soft-clipping to identify the PAS, it is important that the alignment is correct, extending over the entire alignable part of each read. We therefore implemented an additional step following the read-to-genome alignment, extending the mapped region of a soft-clipped read for as long as the number of mismatches between the soft-clipped region and reference genome remained under a threshold, which was

$$threshold = max(length\ of\ soft\ clipped\ region/10, 2)$$

That is, we extended the alignment for as long as the number of errors in the extended alignment stayed under 10%, or, for short extensions, until the number of errors remained less than 2. Once this point was reached, we backtracked to the 3’-most position in the alignment where the read and the genome matched over 3 consecutive bases. The corrected cleavage site was set to the nucleotide after the last of these 3 positions. For further processing, we defined two additional tags associated with the extended read alignments, XO and XF (Table 2), corresponding to the old cleavage site implied by the initial alignment, and the new cleavage site, after the alignment extension.

B. 3. 5. Extraction of poly(A) tail-containing reads (PATR)

Many reads have a few soft-clipped nucleotides at their 3' end that cannot be aligned to the genome. In the dataset that we used for developing the method, *Tabula Muris Senis* sample 10X_P7_14, the distribution of soft-clipped region length decreased abruptly up to 4-5 nucleotides, and slower beyond this point, consistent with two processes generating these soft-clipped regions. The longer soft-clipped regions were also very A-rich (not shown), indicating that they represent poly(A) tails. Thus, we extracted as poly(A) tail-containing reads (PATR) those reads that, after the alignment extension and cleavage site correction, had at least 5 soft-clipped nucleotides at the 3' ends, with more than 80% A's.

B. 3. 6. Standard approach to read deduplication

To illustrate the utility of our tool in extracting experimentally-supported PAS we compared the extracted reads with those obtained with the standard workflow for scRNA-seq analysis. That is, we carried out the read deduplication with the UMI-tools (12) software (version 1.1.1). Throughout we used one sample from the *Tabula Muris Senis* dataset, 10X_P7_14 for these benchmarks. UMI-tools `dedup` was used with parameters extract-umi-method=tag, umi-tag=UB, cell-tag=CB, gene-tag=GX, method=unique, per-gene and per-cell.

We sorted and indexed the alignments with samtools (16) and the set of reads was then processed as the set extracted by SCINPAS, starting with the identification of PATR.

B. 3. 7. Clustering of read 3' ends into PAS clusters

It has been observed before (e.g. (17)) that poly(A) sites are not processed with single-nucleotide precision, but rather mRNAs ending a few nucleotides upstream or downstream of a dominant PAS are typically observed in large scale datasets. For analyses such as of regulatory motifs, it is important to identify these dominant sites, which we refer to simply as PAS, and their respective clusters of secondary cleavage sites. To retrieve these PAS, BAM files containing alignments of PATR were used to construct BED files where the end positions were set to the corrected cleavage sites implied by the reads, the start positions were those preceding the end (i.e. corrected cleavage site -1) and the score was the number of reads with identical corrected cleavage site. We clustered individual cleavage sites as done before (17): in each iteration, we started from the cleavage site with the highest score, which became a new PAS, and associated with it all corrected cleavage sites within 25 nucleotides upstream or downstream. The score of the PAS cluster (PAS score) was computed as the total number of reads supporting the PAS cluster (Fig. 1). We then removed all the cleavage sites associated with the cluster, and moved to the next most frequent cleavage site not yet considered. We repeated the procedure until all cleavage sites were examined (17). For the various controls, we started with the appropriate set of reads (depending on the analysis, reads without poly(A) tails, i.e. non-PATR, or reads deduplicated by standard tools) and applied the same clustering procedure described above.

B. 3. 8. Classification of PAS clusters

To evaluate the SCINPAS-identified PAS we annotated the clusters it produced by intersecting them with non-overlapping features annotated on the genome, i.e. intergenic regions (IG), intronic regions (I), non-terminal exons (NTE) or terminal exons (TE). We used the CellRanger-provided GTF annotation mm10-2020-A_build, which is a modified version of the GRCm38 mouse genome assembly from GENCODE. We extracted entries corresponding to lncRNA and protein-coding mRNAs, and then intersected the locations of PAS clusters with these annotation features. For example, intronic clusters were those that intersected gene loci but not exons (Fig. 1). The intersections were done with the BEDTools (software version 2.27.1) 'window' function with w = 1 (18), to allow for the ambiguity in assigning by different tools of the A nucleotide that frequently occurs after the cleavage position to either the transcript or to the poly(A) tail. For clusters annotated to TE we further distinguished those whose PAS was less than 100 nucleotides from the annotated TE end (annotated in terminal exon, ATE) and those whose PAS was farther away (unannotated in terminal exon, UTE).

B. 3. 9. Classification of PATR

Tag name	Description	Value
XO	Cleavage site implied by initial alignment	Integer
XF	Corrected cleavage site implied by the extended alignment	Integer
YB	Cluster of reads with same unique molecular identifier (UR)	String (URID-subcluster #)
ZI	PAS cluster annotation	class_chromosome:start:end:strand:clusterID[a]
ZS	PAS score	Integer
ZD	Tag indicating whether a read maps to the boundary between the 2 clusters	Integer (0/1)
Zi	PAS sub-cluster id	String (ATE/UTE)
Zd	Tag indicating whether a read maps to the boundary between the 2 sub-clusters	Integer (0/1)

Table 2. Tags added for deduplication and classification of read 3' ends and PAS clusters.

[a]clusterID consists of chromosome, cluster representative, corrected cleavage site and strand separated by ':'.

We also annotated individual reads within the clusters, by propagating the cluster annotation to individual reads. This was achieved by identifying the cluster in which each read belonged and assigning it the annotation of the cluster (ZI tag) and the PAS score (ZS tag). If a read mapped to the boundary between 2 clusters, we assigned it to the cluster with the highest score, and we noted the potential ambiguity by setting another tag, ZD=1. If a read belonged to exactly 1 cluster, the ZD tag value was set to 0. Finally, we used another tag, 'Zi' to denote the ATE or UTE annotation (and a corresponding 'Zd' tag to indicate whether the read overlapped two PAS clusters in the same terminal exon (Table 2). Tag names are in accordance with SAM format specification (https://github.com/samtools/hts-specs).

B. 3. 10. Computation of summary statistics

Number of reads associated with various categories of PAS

The BAM files enhanced with the tags indicating the annotation of the reads were used as input to the 'pysam' python package (version 0.18.0) (16, 19) to count all types of reads (i.e. raw reads, deduplicated, soft-clipped, non-PATR, PATR, TE, ATE, UTE, NTE, I, IG).

Number of covered genes

We considered as annotated those genes for which a transcript with support level (TSL) <= 3 is annotated in the GTF file. TSL 3 signifies that there is at least one sequenced expressed sequence tag providing evidence for a transcript. We counted the number of annotated genes in the GTF file. We then computed the number of expressed genes in a sample as the number of unique gene IDs (GX tag) in the deduplicated BAM file for which there were at least 2 reads mapping to one of the gene's annotated exons. Similarly, we computed the number of genes covered with identified PATR.

Position-dependent nucleotide frequencies around PAS

To determine whether different categories of PAS had the expected nucleotide composition in their vicinity, PAS clusters of specific types were identified in BED files and the PAS, i.e. the cleavage site with the highest read support (found in the ZI tag, see Table 2) was used to extract 101 nucleotides-long genomic sequences centered on these PAS. The relative frequencies of the four nucleotides were computed and visualized for each PAS category.

Position-dependent frequency of polyadenylation signals

The most conserved signal for polyadenylation, i.e. the poly(A) signal, has the consensus sequence AAUAAA, but 12 variants (AAUAAA, AUUAAA, UAUAAA, AGUAAA, AAUACA, CAUAAA, AAUAUA, GAUAAA, AAUGAA, AAGAAA, ACUAAA, AAUAGA) have been found conserved between human and mouse (17), and we refer to them as "canonical". We determined the position-dependent frequency distribution of these canonical poly(A) signals around PAS of various categories as done before (17). Specifically, we extracted the sequence centered on each of the PAS and stored all these sequences into a dataframe. For each sequence we recorded which of the 12 canonical poly(A) signals (17) occurred in it, as a 0 or 1 value in the column corresponding to each poly(A) signal. A column sum then gives the frequency of PAS where each of the poly(A) signals occurs. We then traversed the dataframe iteratively, recording the highest frequency motif, constructing the position-dependent distribution of its occurrence in the sequences that contained it, then removing all these sequences from the dataframe and repeating the process for the next-most frequent poly(A) signal. If a motif occurred more than once in a sequence, its contribution towards each of the positions where it occurred was weighted by 1/number of occurrences, so that each sequence contributed with equal weight to the motif frequency distribution. The analysis was done for entire PAS datasets as well as for subsets of PAS with particular annotations. Running averages (5 nucleotides to the left and right of a given position) were plotted.

Position-dependent frequency of polyadenylation signals in PAPERCLIP-identified PAS

To determine whether the position-dependent frequency of polyadenylation signals depends on the method by which the PAS were inferred, we also analyzed data generated with the PAPERCLIP method (20), in which mRNA termini are identified by crosslinking and immunoprecipitation of the poly(A)-binding protein. We extracted PAPERCLIP-identified PAS from the polyAsite atlas (21), which contains pre-analyzed data for 28 samples mapped to the mouse genome assembly version GRCm38.96. Any PAS with TPM expression > 0 across all

PAPERCLIP samples was written out to a BED file and further used to construct position-dependent frequency of occurrence of poly(A) signals, following the procedure described in the previous section.

Position-dependent frequency distribution of AAUAAA around SCINPAS- and SCAPE-identified PAS

We applied the procedure described in the previous two sections to compare the position-dependent frequency distribution of the main polyadenylation motif, AAUAAA, relative to PAS identified with either SCINPAS or SCAPE.

Consistency of poly(A) signal distribution at PAS and annotated mRNA 3’ ends

To determine whether novel PAS located in various genomic regions are characterized by the same poly(A) signals as annotated PAS we used the following procedure. First, we constructed reference distributions of poly(A) signals upstream of the 3’ ends of annotated mRNAs, as described in the above paragraph. Then, for each of the 12 canonical poly(A) signals, we determined the location of its peak around the 3’ ends of mRNAs and recorded the interval around the peak where the frequency was >= 90% of the peak value. This interval was considered the expected location of the poly(A) signal at true poly(A) sites. Then, for each category of PAS in a dataset we constructed the position-dependent frequency of each canonical poly(A) signal and we determined whether the peak position of each poly(A) signal fell within the interval expected from the true PAS. Finally, we counted for how many poly(A) signals this condition held and we defined this count to be the motif score for each category of PAS in a given dataset. Hence, the minimum motif score of a dataset is 0 and maximum motif score is 12. As negative control, we started from reads without poly(A) tails (non-PATR reads) and applied the same procedure, i.e. clustering, identifying the position with most read support in each cluster, and finally determining the motif scores for these clusters.

Number of PAS in a given category

To compare the performance of SCINPAS with that of other tools that identify PAS from scRNA-seq, we extracted PAS with specific annotations from the relevant BED files (see section Classification of PAS clusters) and counted the number of clusters supported by at least 2 PATR, thus requiring a minimum of 2 reads to support a PAS.

Comparison of PAS usage between 2 different cell types

To compare the pattern of PAS usage in previously analyzed datasets, we used the metadata provided in the respective studies to identify cell types and merge the reads (aligned and deduplicated) from individual cell types. The merged BAM files were further processed to get the PAS of individual cell types. We then intersected the set of PAS identified by SCINPAS with terminal exons of annotated transcripts, and for each terminal exon, we calculated the length implied by the location of PAS within this terminal exon. That is, given the PAS score (number of supporting reads) s_i of a PAS i located at distance d_i from the start of the terminal exon, the average length l_i of the terminal exon in the respective sample is given by $(\Sigma_i\, d_i s_i\,)/(\Sigma_i\, s_i)$. If a cluster overlapped multiple terminal exons, the PAS score was uniformly divided between these terminal exons. We then calculated the ratio of average lengths of each terminal exon between two cell types and the distribution of log-values of this ratio.

B. 3. 11. Comparison with SCAPE

Execution of SCAPE

We downloaded SCAPE from https://github.com/LuChenLab/SCAPE, tested it with the provided example data and executed it with default parameters (see below). By default, SCAPE requires stranded data to infer the insert size. For the widely used 10x Genomics data, the second read contains only the barcodes and the insert size is approximated from paired-end datasets. The number of PAS to search for in a specific terminal exon has to be provided. We used the parameter values suggested by the authors, namely maximum number of PAS = 5, minimum number of PAS = 1, the mean and standard deviation of insert size of the library = 300 and 50 bases, respectively, the length of the poly(A) tail = Uniform(20,150) nucleotides, minimum distance between two PAS = 100 nucleotides, and maximum length of UTR = 6000 nucleotides. SCAPE performs the optimisation step-wise `theta_step=9` and fixes the maximum standard deviation `max_beta=70`. This explains the discrete spans of the regions centered on poly(A) sites.

Obtaining classes of PAS clusters

For the comparison with other tools/resources, we created an additional annotation class, namely of regions of size 1kb downstream of annotated genes and termed it '1kb downstream genes'. We created these regions with BEDTools 'flank' function, then removed regions that overlapped with other genes on the same strand.

For SCINPAS PAS clusters, this is an additional intergenic class, which was obtained with the BEDTools 'window' function using the '1kb downstream genes' regions and the intergenic PAS clusters (parameter -u and one base pair added up- and downstream of the PAS clusters (parameter -w 1)). The remaining intergenic PAS clusters were also obtained with BEDTools 'window' applied to the '1kb downstream genes', but with parameters -w 1 and -v, to report to complement of the previously identified class, i.e. PAS that were initially classified as intergenic, and were further located outside of the 1kb downstream of annotated genes. For both cases only overlaps on the same strand are reported (parameter -sm).

The main output of SCAPE, the 'pasite.csv.gz' file, contains the count for each cell barcode and PAS. These values were summed and saved into a standard BED file. The start and end coordinates of each PAS was computed as floor(mean - beta/2) and floor(mean + beta/2), where mean and beta were the parameters of the fitted Normal distribution from SCAPE.

The SCAPE PAS were classified with BEDTools 'intersect', similar to the classification of SCINPAS PAS. Exonic PAS were obtained from the intersection with exons but not terminal exons, intronic PAS from the intersection with genes but not exons, and intergenic PAS were those that did not intersect with genes or with '1kb downstream genes'. The annotation '1kb downstream genes' was obtained when PAS did not intersect genes but overlapped completely (-f=1) with the class '1kb downstream genes'. Lastly, terminal exon PAS were obtained from the intersection of terminal exons only.

Analysis and graphics

In general, we used SCINPAS-extracted PAS clusters with at least 2 supporting PATRs. This was also the case when we compared SCINPAS to SCAPE. To plot the number of PAS clusters, the individual classes (TE, exons, introns, intergenic and '1kb downstream genes') were extracted and plotted as stacked bar charts with `geom_col`.

The distance between PAS and the closest PAS cluster downstream was computed as follows. For each chromosome and strand, PAS clusters were sorted by start and end positions. Then for each but the last cluster we obtained the distance from its end position to the start position of the following cluster. The distance distribution plot was created with `geom_freqpoly` using density estimates.

The scatter of the number of supporting reads associated with SCINPAS and SCAPE-identified PAS in individual genes were generated as follows. For each gene, overlaps between gene (g) and PAS cluster (p) were found by requiring the same chromosome and strand and (g_s <= p_e) & (g_e >= p_s) for SCAPE and (g_s <= p_e+1) & (g_e >= p_s-1) for SCINPAS, where (s) and (e) are start and end coordinates respectively. This allows for partial overlaps, which is also the default behavior of BEDTools intersect and window functions. The found overlaps were counted and the individual PAS scores (i.e. number of reads supporting the PAS cluster) were summed. The log(read count+1) values were plotted as a scatter. Density estimates were created with `geom_density2d` using 200 grid points in each direction. The Spearman rank correlation rho and associated p-value was computed with `cor.test(method="spearman")` on the PAS score at the gene level.

For all PAS clusters, irrespective of annotation, the span was computed as the distance between the end and start coordinates (from the BED file coordinates).

All plots were generated with ggplot2 (22).

Examples of PAS and read coverage of gene loci were visualized with IGV v2.11.9 (23).

B. 3. 12. Overlap of SCINPAS-inferred PAS from the *Tabula Muris Senis* samples with the polyAsite atlas

We used the 6 *Tabula Muris Senis* samples from Table 1 to infer PAS, requiring a minimum of 2 reads support. We then determined whether a SCINPAS PAS cluster (x) overlapped a PAS cluster (y) from the polyAsite atlas (21), located on the same chromosome and strand if the start (s) and end (e) coordinates of the clusters satisfied the condition (y_s <= x_e + 1) & (y_e >= x_s - 1). This condition, which allows for clusters to be immediately adjacent to each other rather than overlapping, accounts for the possibility that tools may differ in whether they assign an A nucleotide that frequently occurs in the genome immediately downstream of the cleavage, to the templated part of the transcript or to the poly(A) tail. We then counted the fraction of SCINPAS clusters that overlapped a PAS cluster, for various numbers of SCINPAS clusters, sorted by their read support (i.e. top 100, 500, 1000, etc.).

B. 3. 13. Expression levels of RNAs with or without the AAGAAA PAS motif

We first filtered representative cleavage sites that overlap with terminal exons and grouped them by the gene name of the terminal exons with which they overlapped. Definition of overlap is the one in the paragraph above. If multiple terminal exons overlapped with a given representative cleavage site, the terminal exon whose end was closest to the representative cleavage site was associated with the respective cleavage site. The selected terminal exons were then divided in two sets, depending on whether or not any of the PAS within them had the AAGAAA motif in the region -40bp to +20bp. The distribution of transcript expression levels (number of reads in the PAS clusters of the terminal exon) was then calculated for the two categories of TEs in the three datasets used for benchmarking: *Tabula Muris Senis* sample 10X_P7_14, T cell activation dataset (union of sites in all samples) and sperm cell development dataset (also union of all sites in these samples).

B. 3. 14. Distance of PAS to terminal exon ends

To determine how precise different methods are in identifying TE ends, we first filtered representative cleavage sites that overlap with terminal exons. The definition of overlap is the same as the two paragraphs above. If a given representative cleavage site overlapped multiple TEs, the TE whose end was closest to the representative cleavage site was associated with the respective cleavage site. Then the distance was computed as

$$distance = abs(end\ of\ terminal\ exon\ -\ representative\ cleavage\ site)$$

The distances were computed for both samples and control to generate a cumulative frequency plot. For the control, UMI-tools deduplicated 10X_P7_14 was used.

B. 4. Results

B. 4. 1. scRNA-seq reads provide direct evidence of polyadenylation sites

Increasingly many studies have started to investigate APA from scRNA-seq datasets that are generated with the 10x Genomics technology, which captures 3’ fragments of mRNAs (5, 6, 8, 10). Invariably, these studies start from “deduplicated” reads mapped to the reference genome with the CellRanger software (11). While a unique molecular identifier (UMI) is attached to the 3’ end of an mRNA, PCR copies of the mRNA are fragmented and 3’-terminal fragments are sequenced in the 5’-to-3’ direction, yielding distinct reads associated with the same UMI. For quantifying gene expression it is not crucial which of the reads with the same UMI is selected for quantification during the read deduplication process. However, reads that map most distally in the gene locus are more likely to reach the 3’ end of the mRNA. Thus, for the purpose of identifying reads that contain poly(A) tails and thus provide experimental evidence of the PAS, it is important to select these distal reads from among those with identical UMIs. To demonstrate this, we determined the number of reads with unmapped (soft-clipped) nucleotides at the 3’ end that were extracted either with standard software (CellRanger followed by UMI-tools) or by our software. On a randomly chosen sample from the *Tabula Muris Senis* dataset (ID:10X_P7_14), we found that 0.44% of the reads that were extracted with the standard software had soft-clipped nucleotides at their 3’ end, while this proportion was ~3-fold higher, 1.12%, when selecting distal reads. Similar results were obtained on other datasets (not shown). This result emphasized the need for a tool to pre-process scRNA-seq reads so as to maximize the recovery of poly(A) tail-containing reads and thereby polyadenylation sites with experimental support.

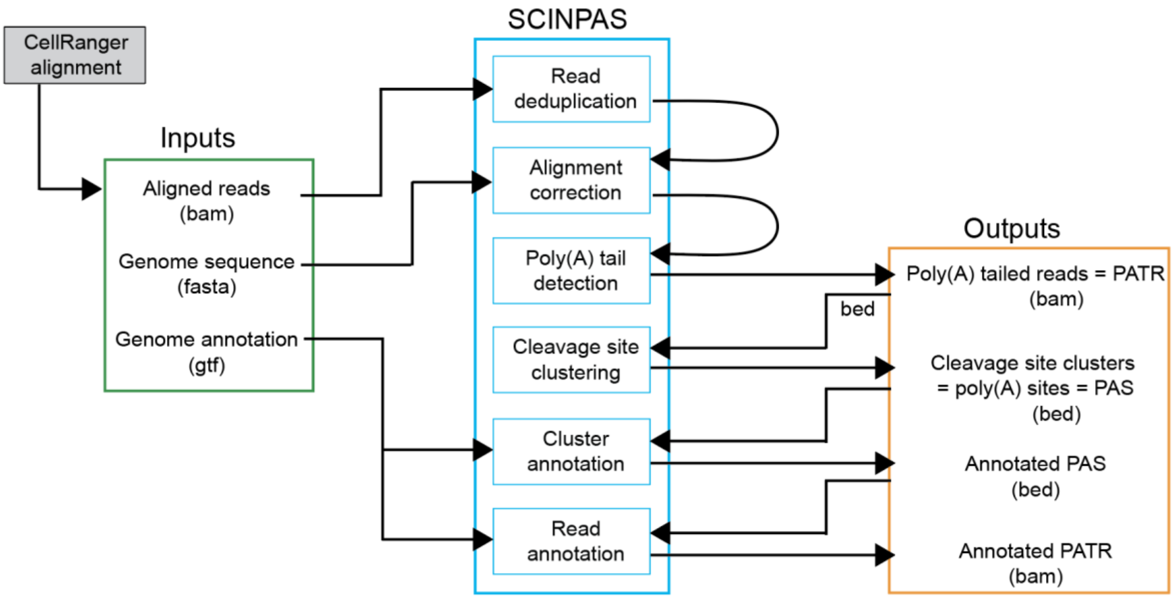


Figure 1. Scheme of SCINPAS workflow.

The inputs to SCINPAS are indicated in the green box. Alignments of reads from primary samples are generated with CellRanger. The SCINPAS processing steps are shown in the cyan boxes and the outputs of the workflow are indicated in the orange box. File formats for inputs and outputs are indicated in parentheses.

A scheme of the SCINPAS - short for **sc**RNA-seq-based **i**dentification of **n**ovel **p**oly(**A**) **s**ites - workflow is shown in Fig. 1. SCINPAS is written in the nextflow language (24) and its key features are the following. First, in contrast to UMI-tools, which uses the genome annotation to collapse reads that have the same UMI and map to the same gene, SCINPAS does not assume a specific genome annotation but rather is able to identify PAS that are located outside of the currently annotated exonic/genic regions. To demonstrate this, we first clustered the reads that came from the same cell and had the same unique molecular identifier. Most clusters spanned less than 10 kilobases (Fig. 2A), as expected when reads come from terminal fragments of mRNAs, terminal exons being generally kilobases-long (25). However, some clusters had a much larger span. This could occur when the sequenced fragments span splice junctions, or perhaps from rare cases when distinct mRNAs were tagged with the same UMI. In SCINPAS, we collapse all the reads with the same CB and UMI, but only within some maximum cluster span. That is, we traverse the genome in the 5'-to-3' direction to construct clusters of such reads, ending a cluster when a predefined threshold (100'000 nucleotides) in length is reached. The selection of the distal read is done separately for each such cluster (Fig. 2B). As only reads with poly(A) tails contribute to PAS identification, if reads with the same UMI end up erroneously in multiple clusters, the reads originating from the upstream clusters would not have poly(A) tails and thus spurious PAS will not be generated, despite the error in read clustering. On the other hand, if the initially large cluster span was really due to the same UMI being attached to multiple isoforms, then the upstream clusters should also contain reads with poly(A) tails, and they will be kept for further analysis. The second key step is to identify the reads containing poly(A) tails. For this, we extracted all the reads whose 3' ends could not be mapped to the genome, i.e. those with soft-clipped nucleotides at the 3' end. In the sample that we arbitrarily picked for the tool development, the 10X_P7_14 sample from the *Tabula Muris Senis* dataset, the soft-clipped part of the reads was generally very short, 1-3 nucleotides in 58.2% of the cases (Fig. 2C). However, many reads still had longer soft-clipped regions, up to ~30 nucleotides. Inspection of read-to-genome alignments revealed some cases in which the alignment (generated by the STAR software (26)) could be further extended into the soft-clipped region, without a decrease in the alignment quality (Fig. 2D). Thus, we implemented an additional step of refining the alignment by extending the mapped regions of soft-clipped reads until the number of mismatches between the soft-clipped region and reference genome reached a maximum threshold and then correcting the cleavage site implied by the read (see Methods). Finally, we selected the reads that contained non-templated poly(A) tails of at least 5 nucleotides and over 80% A's, and we clustered them as described previously (17), to remove the small variability in cleavage sites. We consider the most frequently used cleavage site in a cluster (cluster representative) to be the poly(A) site (simply PAS). In the 10X_P7_14 sample we found that 1.6% of the deduplicated reads contained poly(A) tails. The clusters and individual cleavage site positions (including corrected positions) within them were then saved in BED and BAM files, respectively, and then finally, annotated (Fig. 2E).

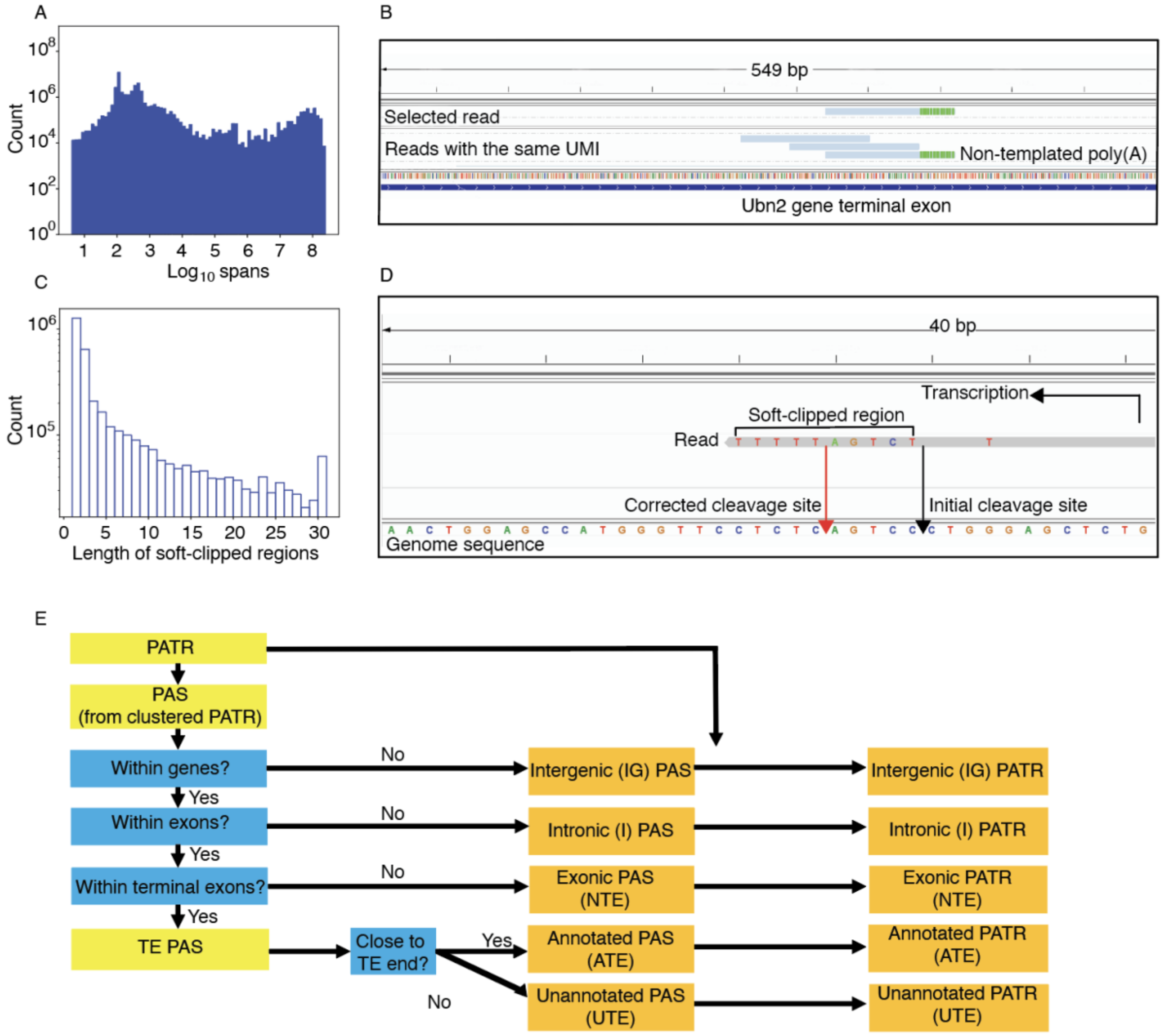


Figure 2. Key steps in SCINPAS.

A. Distribution of genomic spans of reads with the same cell and molecular identifier (CB and UR tag, $\log_{10}$) constructed from the sample 10X_P7_14 of *Tabula Muris Senis*. **B.** Illustration of distal read selection, from among the reads with the same CB and UMI. In this case, only the most 3′ read has 3′ non-templated A nucleotides (indicated by the green color). **C.** Distribution of soft-clipped region length in reads from the same sample, as given by the STAR software. **D.** Illustration of a read-to-genome alignment that could be extended further over the region marked as soft-clipped in the initial alignment. The read maps to the negative strand of the genome. The start of the soft-clipped region marks the “Initial cleavage site” implied by the alignment. The “Corrected cleavage site” (red arrow) results from the extension of the alignment over the mappable part of the soft-clipped region. **E.** Scheme of SCINPAS annotation of PAS and PATR.

B. 4. 2. SCINPAS improves the recovery of poly(A) sites relative to standard software

To compare PAS recovered from reads extracted by either SCINPAS or the standard software, we investigated a few properties previously found to characterize true PAS. First, the mouse genome being quite extensively annotated, we expect that most well-expressed isoforms are already represented in this annotation, and are recovered by an accurate PAS identification tool. Of the 652’288 poly(A) tail-containing reads (PATR) extracted by SCINPAS from the *Tabula Muris Senis* 10X_P7_14 sample, 415’299 mapped to annotated terminal exons (TE - 63.7 %), 2’329 to other exons (NTE - 0.4 %), 34’484 to introns (I - 5.3 %) and 200’176 to intergenic regions (IG - 30.7 %) (Fig. 3A). In contrast, only 133’536 PATR were extracted after applying the UMI-tools software, 126’958 from terminal (95.1 %), 566 (0.4%) from other types of exons, 952 (0.7 %) from introns and 5’060 (3.8 %) from intergenic regions. The main difference is that SCINPAS identifies PATR in intergenic regions. When these are not considered, as in the standard analysis, the proportion of PATR in terminal exons compared to other genic regions is indeed very high, 91.9 %. The small number of reads that end up with intergenic and intronic annotation after the application of UMI-tools deduplication come from regions that were considered genic in the older mouse genome annotation that was used by the *Tabula Muris Senis* project for mapping the reads to the genome, but not in the newer annotation that we used in SCINPAS for read and PAS classification. Thus, SCINPAS identifies many more polyadenylated reads, the majority of which come from terminal exons, but also some that come from intergenic regions.

We also asked whether the transcript ends implied by the inferred PAS indeed correspond to the ends of annotated terminal exons. To answer this, we calculated the distances between PAS, weighted by the number of supporting reads, and the annotated ends of the terminal exons in which the PAS are located. The cumulative density function of $\log_{10}$ values of the distance, shown in Fig. 3B, confirms that the vast majority of SCINPAS-extracted PATR are located within 10 nucleotides from the annotated terminal exons, while UMI-tools-extracted reads end hundreds of nucleotides away from the terminal exon end. For better resolution of PAS annotation (Fig. 2E), we distinguished between PAS located at most 100 nucleotides upstream of the terminal exon end (we called these “annotated” TE PAS, or ATE) and those that were located further upstream in terminal exons (UTE PAS).

The sequence composition around PAS, determined in many previous studies (17, 27, 28), is strongly enriched in A nucleotides at ~20 nucleotides upstream of the PAS, where the poly(A) signal is located, while the region downstream of the PAS is enriched in U nucleotides. To test this, we first clustered cleavage sites implied by the PATR into clusters of closely-spaced sites, and took the most frequently used position in a cluster (the “cluster representative”) as the actual poly(A) site (see Methods). Computing the nucleotide frequencies around these PAS, we obtained the expected pattern (Fig. 3C). This was not the case when the reads used to infer cleavage sites came from the UMI-tools deduplication and were not constrained to contain poly(A) tails (Fig. 3D). Furthermore, different categories of PAS individually exhibited a similarly biased nucleotide composition (Fig. S1).

We also specifically checked for the presence of the poly(A) signal, which has the AAUAAA consensus and is located at ~20 nucleotides upstream of the cleavage site (17, 29, 30). There are 12 variants of the consensus that are conserved between human and mouse (17), and almost all showed the expected peak at ~20 nucleotides upstream of the PAS (Fig. 3E). In contrast, no such pattern was exhibited by the negative control data set, constructed from reads without poly(A) tails obtained with the standard UMI-tools-based deduplication (Fig. 3F). Altogether, these results demonstrate that our tool improves the recovery of bona fide PAS from scRNA-seq data relative to the standard workflows.

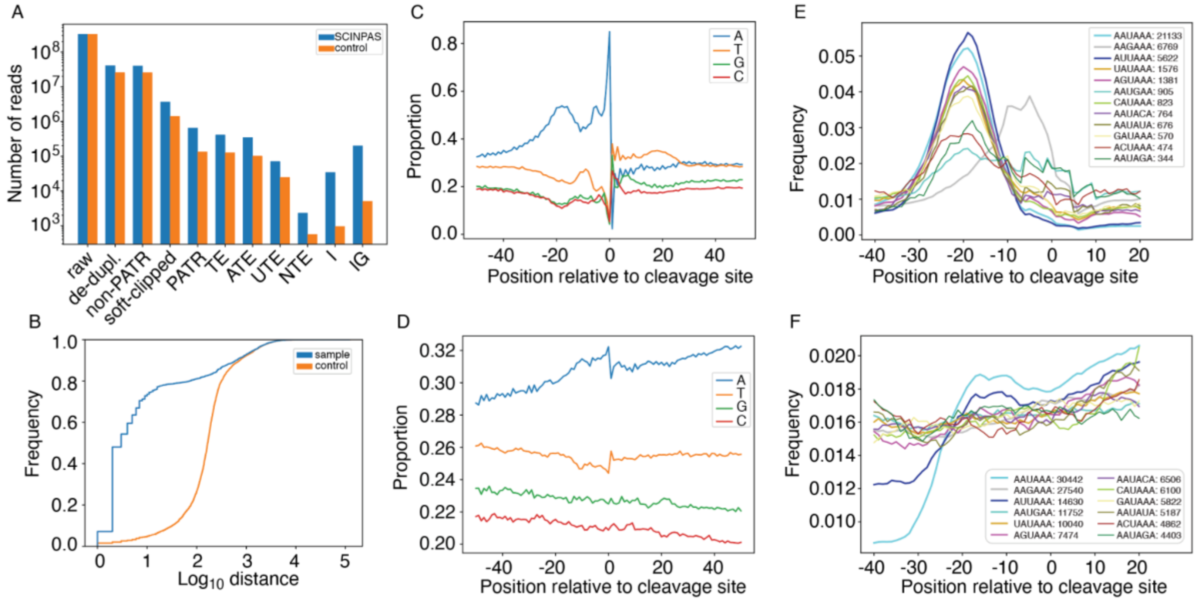


Figure 3. PAS extracted by SCINPAS contain the expected poly(A) signals.

A. The number of reads from the *Tabula Muris Senis* sample 10X_P7_14, at different steps of the processing pipeline, when the processing is done with SCINPAS (blue) or the standard UMI-tools-based workflow (orange). **B.** Distribution of $\log_{10}$ distances between inferred and annotated terminal exons, when processing is done with SCINPAS (blue) or the standard workflow (orange). **C.** Position-dependent nucleotide frequencies in PAS constructed from SCINPAS-extracted reads. PAS are anchored at position 0, and the genomic sequence upstream and downstream (from -50 to +50 nucleotides) was used to calculate nucleotide frequencies. **D.** Similar for negative control sites. **E.** Position-dependent occurrence of poly(A) signals. The genomic sequence from -40 to +20 around PAS was extracted, poly(A) signals were identified and tabulated, and the frequency of poly(A) signal occurrence across all examined sequences was calculated. **F.** Similar for negative control sites.

B. 4. 3. SCINPAS identifies PAS in genic and non-genic regions

Given that the majority of PATR and PAS correspond to terminal exon ends, we wondered whether PAS that SCINPAS identified in other types of genomic regions also carry the expected signals for 3’ end processing and polyadenylation. Thus, we constructed position-dependent distributions of occurrence of canonical poly(A) signals around putative PAS of different annotation categories. As negative control, we compared these distributions with those obtained for a similarly analyzed dataset, where the reads were deduplicated with UMI-tools and did not contain soft-masked nucleotides. Indeed, all but the smallest category of SCINPAS-extracted PAS had the expected enrichment of almost all poly(A) signals at ~20 nucleotides upstream of the PAS (Fig. 4A-E). The few PAS identified in non-terminal exons had the expected enrichment of the main poly(A) signal, AAUAAA, while for the other signals the number of occurrences was low and the positioning relative to PAS less clear. These results indicate that reads with poly(A) tails selected by our tool identify bona fide PAS across all types of genomic regions. The results also suggest that position-specific patterns of occurrence of poly(A) signals are very reliable and can be used to flag datasets from which PAS are not accurately identified.

One of the 12 conserved signals, AAGAAA showed a different positional pattern than the other motifs, peaking not at ~ -20 nucleotides of the PAS, but in the region -10 to 0. We also checked this motif's frequency around the ends of the annotated TEs in our genome annotation and found it to peak at ~ +10 nucleotides, i.e. downstream of the TE end (Fig. 4F). To exclude the possibility that priming on internal poly(A) stretches underlies the differences in motif occurrence around SCINPAS PAS compared to annotated TEs, we further determined the position-dependent frequency of the motif occurrence in the vicinity of PAS that were determined with an orthogonal experimental method, PAPERCLIP (20), which uses crosslinking and immunoprecipitation of the poly(A) binding protein rather than priming with oligo(dT) to detect poly(A) tails. We extracted the PAPERCLIP-identified sites from the polyAsite atlas (21) and constructed the position-dependent motif distribution as done for all other categories of sites. The results show that in this data set as well, the AAGAAA motif peaks at ~10 nucleotides upstream of the PAS, similar to SCINPAS-identified PAS, and not to annotated TEs (Fig. 4G).

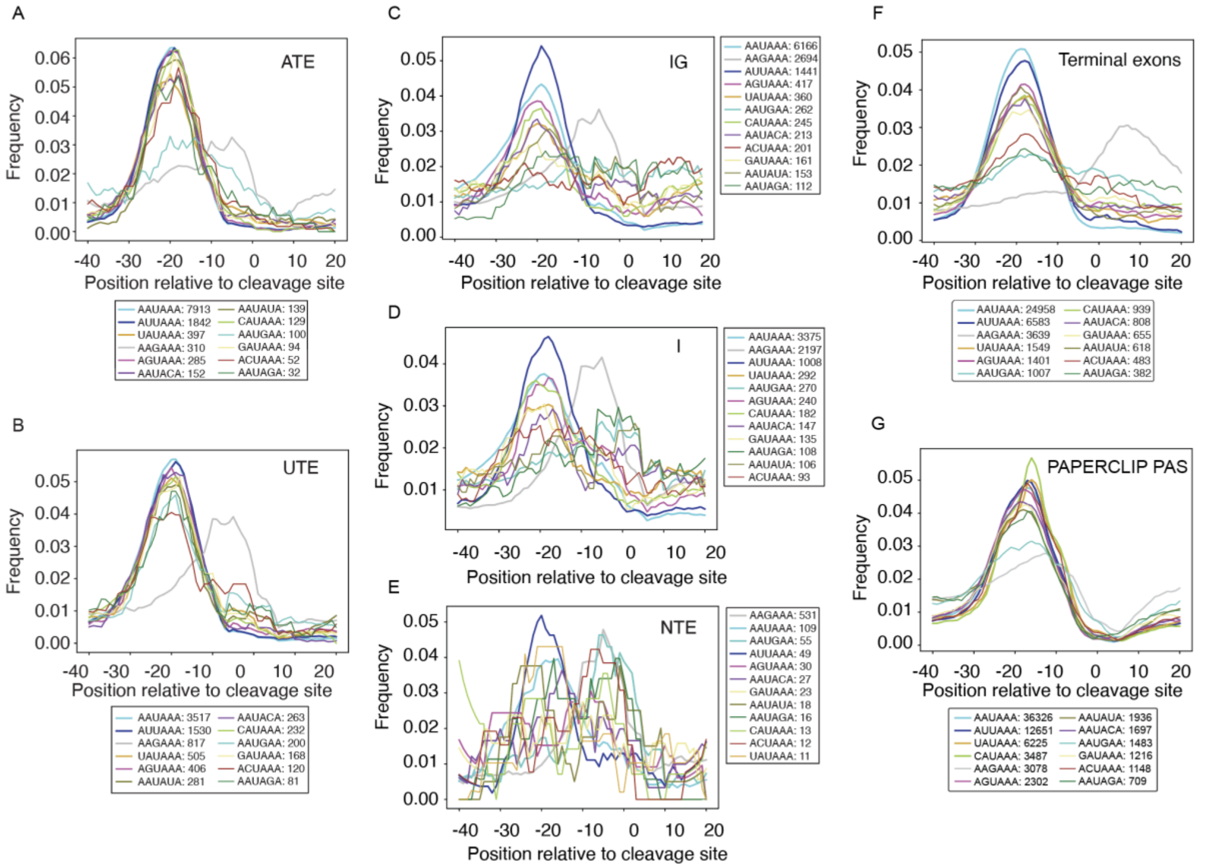


Figure 4. Position-dependent frequency of occurrence of poly(A) signals at different types of PAS.

A-E. PAS were extracted and annotated with SCINPAS from the *Tabula Muris Senis* sample 10X_P7_14. ATE - PAS within 100 nucleotides of annotated TE ends; UTE - PAS in TEs but >100 nucleotides from the annotated TE ends; IG - intergenic; I - intronic; NTE - PAS in exons that are not TE. **F.** Motif distributions around TE ends from the annotation of the GRCm38 mouse genome assembly. **G.** Similar, for PAS identified by the PAPERCLIP (20) method for experimental identification of PAS.

B. 4. 4. PAS identified by SCINPAS exhibit the expected dynamics during cell differentiation

To further test the ability of SCINPAS to identify non-canonical PAS, we applied it to two systems in which the abundance and dynamics of such sites has been reported before, T cell activation and sperm cell development, systems in which the usage of intronic and/or coding-region-proximal PAS is activated (31, 32). Applying SCINPAS to the T cell activation dataset (14) we found that intronic PAS are more frequent, 13.9% of all annotated PAS, in activated T cells compared to the naive T cells, where 10.3% of PAS were annotated as intronic. The average terminal exon length as implied by the PAS inferred from the respective samples, remained largely unchanged, as we observed similar numbers of terminal exons that became shorter or longer by at least a factor of 2 upon T cell activation (3.3% vs. 2.8%, Fig. 5A). We carried out a similar analysis for a sperm cell development dataset (15), comparing PAS usage in elongating spermatids and spermatocytes. The proportion of intronic PAS in this dataset was more similar between the two differentiation stages 10.8% vs 9.3% in spermatocytes and elongating spermatids, respectively, but many more terminal exons (13.4%) became at least 2-fold shorter upon spermatocyte differentiation into elongating spermatids than becoming longer by the same factor (1.6%, Fig. 5B). As with other analyzed datasets, the intronic PAS inferred from activated T cells (Fig. 5C) and elongating spermatids (Fig. 5D) had the expected peak poly(A) signals at ~20 nucleotides upstream of the inferred cleavage site (Fig. 5C-D). An example of intronic PAS usage in the sperm cell differentiation system is shown in Fig. 5E.

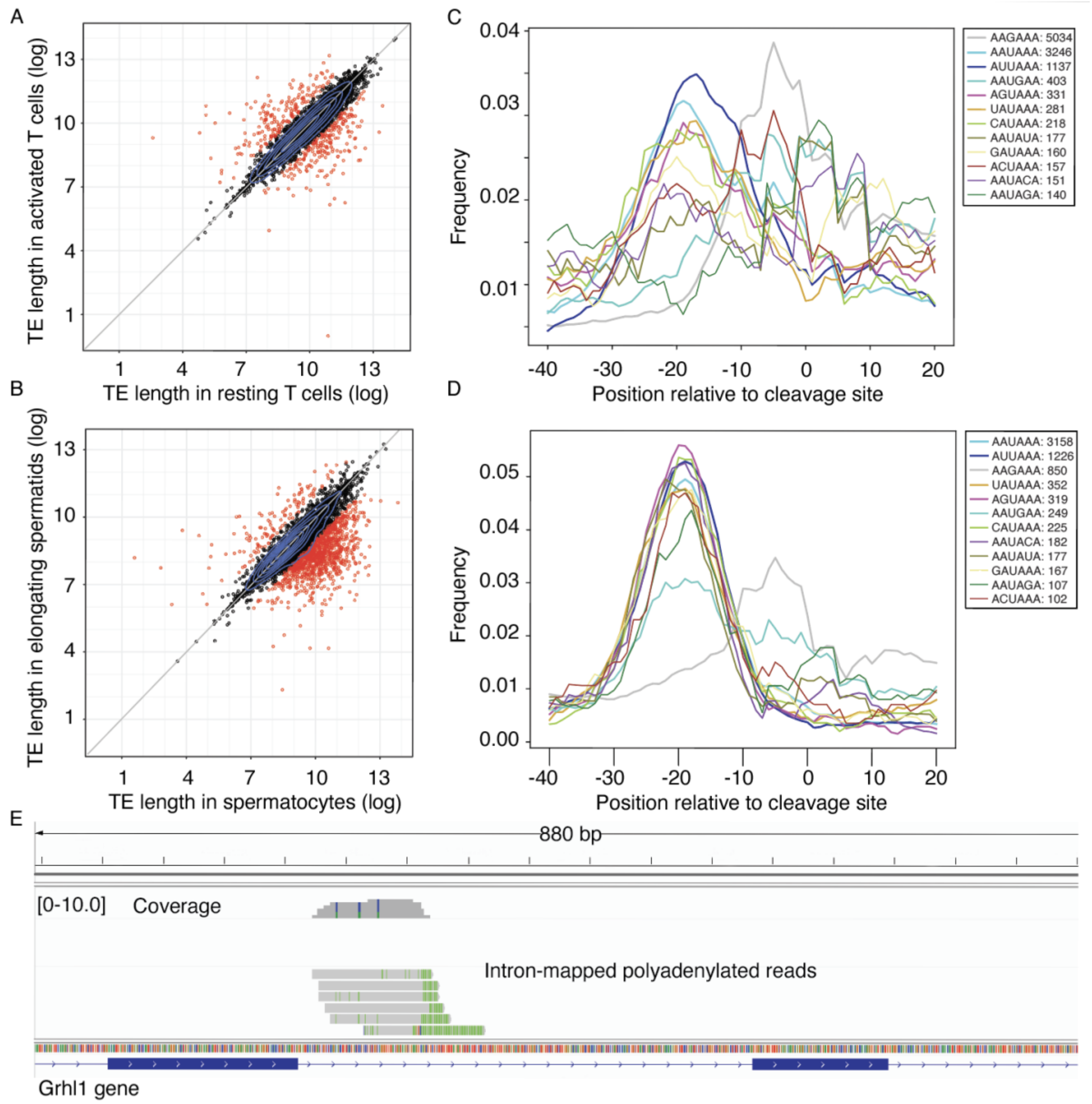


Figure 5. SCINPAS-recovered sites reproduce APA patterns in previously characterized systems.

A. Scatter plot of average terminal exon (TE) length (log2 values) computed from the location and relative abundance of PATR mapping to individual terminal exons. Highlighted in red are TEs whose length changes (increases or decreases) by more than a factor of 2 in activated compared to resting T cells. **B.** Similar to A but comparing elongating spermatids with spermatocytes. **C.** Position-dependent frequency distribution of canonical poly(A) signals at intronic PAS identified in activated T cells. **D.** Similar to C, for intronic PAS of elongating spermatids. **E.** Example of an intronic PAS identified from sperm cell development dataset. Top track shows the coverage of the region by reads, individual reads with poly(A) tails are shown in subsequent tracks ('A' nucleotides are shown with green color) and the gene annotation is shown in the bottom track.

B. 4. 5. SCINPAS provides complementary information relative to other tools

As already mentioned, a number of tools have been developed for extracting PAS from scRNA-seq data, though they do not focus on PATR. A very recently-published and benchmarked tool, called SCAPE (6), uses PATR in the estimation of insert length in paired-end sequencing datasets, so that peaks in read coverage corresponding to PAS can be appropriately positioned on the genome. SCAPE was also found to perform favorably with respect to the other tools developed to date, namely scAPA (33), Sierra (34), scAPAtrap (8), SCAPTURE (10) and MAAPER (35).

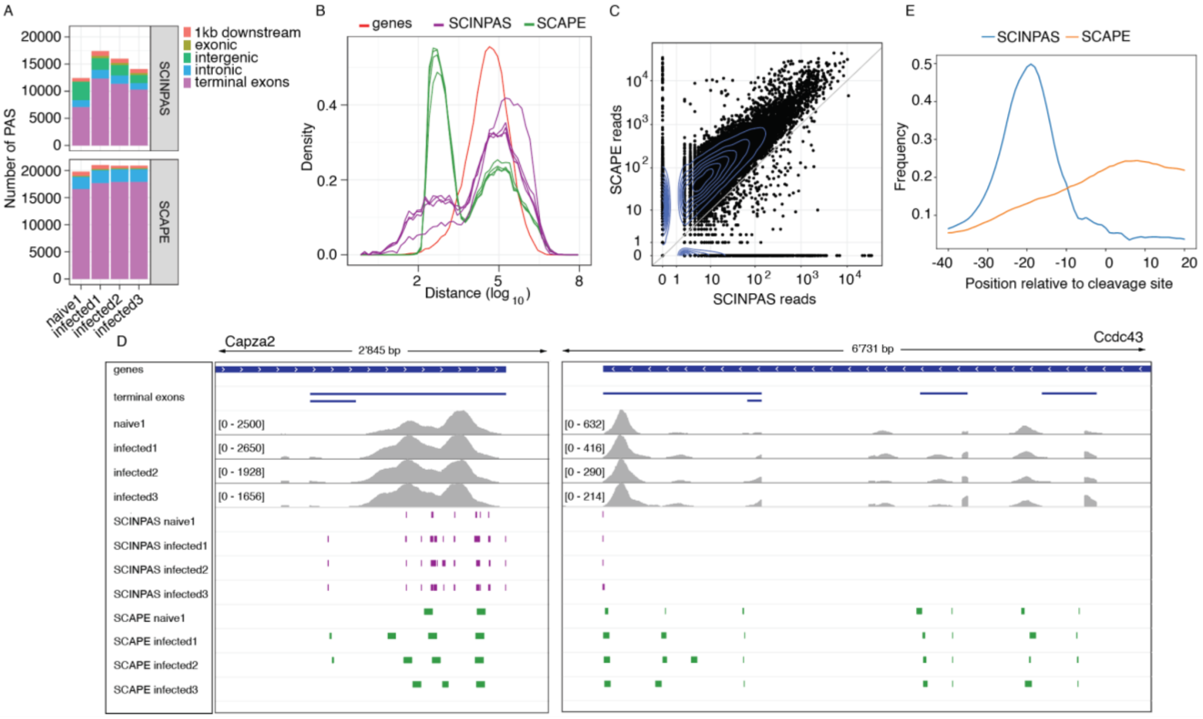


Figure 6. PAS recovery by SCINPAS and SCAPE.

A. Number of PAS inferred by SCINPAS (top) and SCAPE (bottom) from the T cell activation data. The colors indicate different classes of PAS (see legend). **B.** Distribution of the distances from each PAS to the closest PAS downstream for SCINPAS (purple) and SCAPE (green). For comparison, the distribution of 3'end-to-5'end distances between genes is shown in red. **C.** Scatter of the total number of PAS-associated reads within a gene for SCINPAS (x-axis) and SCAPE (y-axis). Spearman correlation coefficient was 0.68 (p-value < 2.2e-16). The diagonal of equal read counts is shown in gray. 2D kernel density estimates obtained with the geom_density_2d(n=200) function of ggplot2 are shown as blue contours. **D.** Examples of PAS recovered by SCINPAS (purple) and SCAPE (green) in the Capza2 (left) and Ccdc43 (right) genes, from the T cell activation dataset. Genes and terminal exons are shown in the IGV browser (23) in blue, and the coverage tracks in gray. **E.** Position-dependent distribution of the canonical polyadenylation signal AAUAAA around SCINPAS- and SCAPE-identified PAS.

First, we determined the number of PAS clusters identified by SCAPE and SCINPAS in each sample in the T cell activation dataset. As shown in Figs. 6A and S3A, while the number of PAS from terminal exons does not show a consistent difference between SCINPAS and SCAPE, SCINPAS identifies many more PAS in intronic and intergenic regions that are not analyzed by SCAPE. The number of PAS identified per sample is more variable for SCINPAS, probably because SCINPAS only uses PATR, which represent only a few percent of the deduplicated reads in a library (Fig. 3A). To better understand what the two methods extract from the data it is insightful to examine the distance from each PAS to the closest PAS downstream. The distributions constructed from each sample in the T cell activation dataset are shown in Fig. 6B and in both cases they have a prominent peak located at approximately 50'000 nucleotides, roughly corresponding to end-to-end distances between genes (red line), as expected. The left sides of the distributions, however, are very different. SCAPE identifies PAS that are ~500 nucleotides apart, likely reflecting choices in the SCAPE model (Gaussian shape of the peaks with mean insert size of 300 and standard deviation of 50 nucleotides). In contrast, the distances between SCINPAS clusters have a broad distribution between ~100 and ~10’000 nucleotides, with no preferred distance, as may be expected if the PAS occurred randomly within terminal exons. SCINPAS clusters are either composed of single cleavage sites, or have a relatively small span (peak at 5 nucleotides), indicating that the cleavage sites are well-defined, but also that the supporting data is sparse. In contrast the span of SCAPE clusters shows a periodicity of 9 nucleotides, again likely indicating parameter choices of the method (Fig. S2). We also compared the number of supporting reads associated with PAS in individual genes. While the SCINPAS counts were ~10-fold lower, as expected from the fact that it only uses PATR and not all deduplicated reads, the Spearman correlation coefficient of SCINPAS and SCAPE counts was relatively high, 0.68 (p-value < 2.2e-16, Fig. 6C). In some instances, SCINPAS detected more PAS clusters per gene compared to SCAPE (Fig. 6D, left panel), though examples where the opposite was the case also occurred (Fig. 6D, right panel). We performed the same analysis as above on the sperm cell development dataset and found similar trends (Fig. S3). Overall, SCINPAS detected fewer PAS clusters per gene in the T cell activation dataset (Fig. S2B) but more PAS clusters in the sperm cell development dataset (Fig. S2D). The increased positional resolution of SCINPAS-identified sites is also emphasized by the position-dependent distribution of the canonical polyadenylation motif, which has a sharper peak for the SCINPAS-identified sites compared to those identified in SCAPE (Fig. 6E, S3E).

Finally, we asked how reproducible the PAS identified by the two methods were between replicate samples, by calculating the Jaccard statistic with BEDTools (18). As indicated in Table 3, the Jaccard statistics were higher for SCINPAS than for SCAPE when comparing replicates, and lower when comparing PAS obtained from naive and activated T cells.

	SCINPAS	SCAPE
Mouse 1 vs 2	0.3887	0.3088
Infected 1 vs 2	0.3903	0.3012
Infected 2 vs 3	0.3879	0.3834
Infected 1 vs 3	0.38478	0.2852
Naive 1 vs infected 1	0.2230	0.2479

Table 3: Jaccard statistics.

Pairwise comparison of SCINPAS (left) and SCAPE (right) predicted PAS in individual samples from the sperm cell development (mouse 1 and 2) and T cell (naive 1 and infected 1-3) activation datasets.

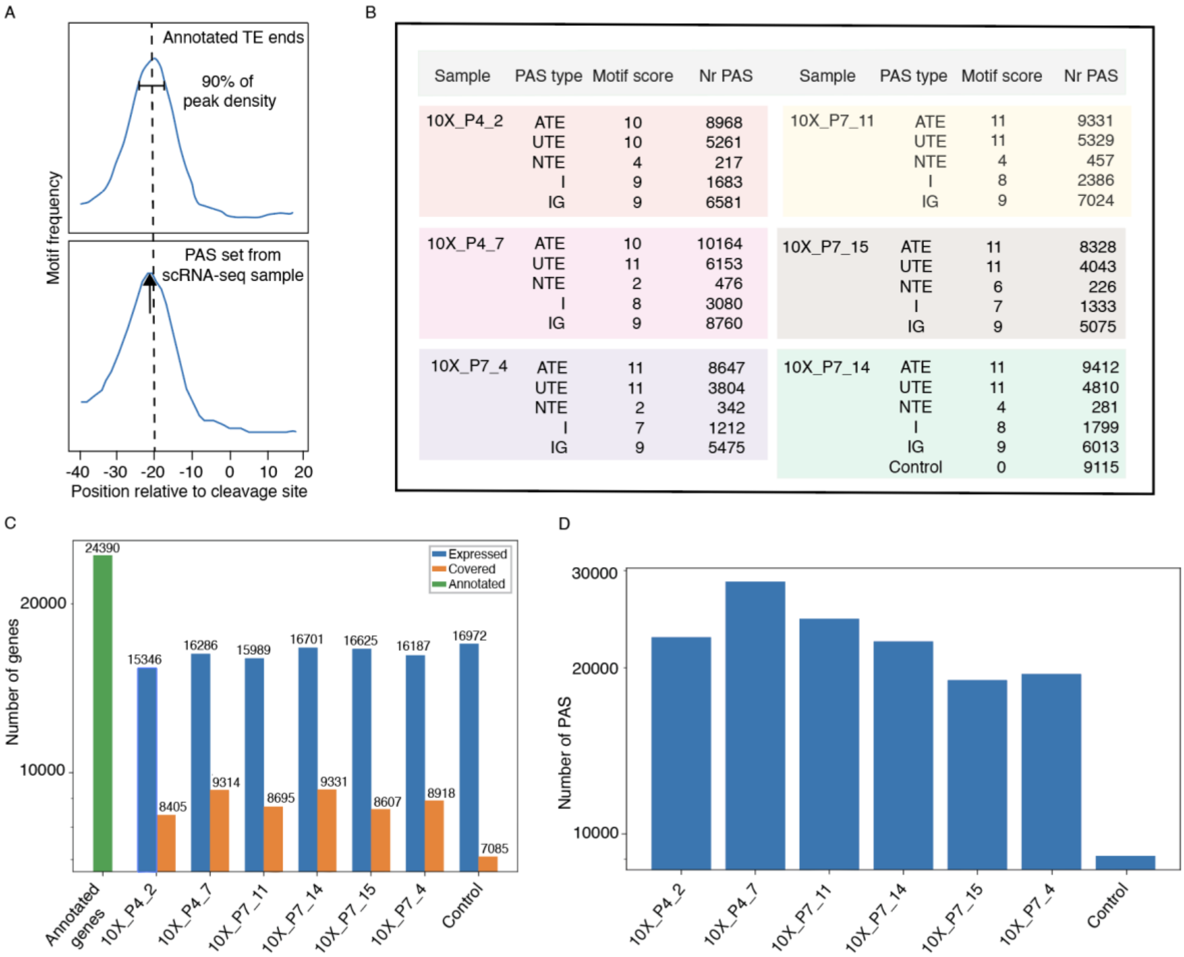
Taken together, SCINPAS compared well with the most up-to-date method available, identifying not only sites in terminal exons, but also in intronic and intergenic regions. The method is efficient, as it uses a much smaller fraction of the sequenced reads than SCAPE, and gives more reproducible PAS when applied to closely-related samples.

B. 4. 6. SCINPAS-based annotation of PAS from the *Tabula Muris Senis* dataset

Finally, to illustrate the generality and utility of SCINPAS we applied it to a large dataset of mouse scRNA-seq, *Tabula Muris Senis* (13), which was generated with a view of building an atlas of gene expression in the mouse. The run time of SCINPAS ranged from 1.5 to 8 hours for all samples in an individual dataset (Table 1). To roughly assess the reliability of PAS inferred from a given sample, we used a measure based on the poly(A) signal distribution around the PAS. Namely, we determined the number of canonical poly(A) signals that peaked at the same position in the SCINPAS-inferred PAS as in annotated terminal exon ends. We considered a peak to occur at the expected position if it was located within the 90% peak frequency window inferred from annotated terminal exon ends (Fig. 7A). As shown in Fig. 7B, in all datasets, all but the NTE PAS categories had the known poly(A) signal peaking at the expected position upstream of the PAS. This was not the case for the negative control which was constructed based on non-PATR reads from the UMI-tools-deduplicated 10X_P7_14 sample. The PAS located in non-terminal exons (NTE) generally represented a small proportion of all the inferred PAS in each dataset (0.96% - 1.86%, depending on the dataset), and for these, only the main poly(A) signals, AAUAAA and AUUAAA occurred in sufficient frequency to yield stable profiles (Fig. 7B).

To evaluate the sensitivity of our method we determined the proportion of expressed genes (supported by at least 2 reads) for which a PAS with a minimum support of 2 PATR was found. The results in Fig. 7C show that SCINPAS identified a PAS for approximately 52-57 % of expressed genes, whereas only 42% were covered by PAS inferred when starting from UMI-tools-deduplicated reads. The total number of PAS we identified in each of the samples is shown in Fig. 7D.

We further compared the PAS that we obtained here with the polyAsite atlas (21), which contains a curated collection of ~300'000 PAS identified in the mouse genome by bulk 3' end sequencing. By taking the union of PAS from the *Tabula Muris Senis samples* (13) defined in Table 1, we obtained a total of 67'829 PAS. 35'741 of these are represented in the polyAsite atlas, while 32'088 can be considered novel. The overlap with polyAsite atlas is larger when considering only the most supported PAS (Fig. S4), as may be expected. These results demonstrate the utility of our tool in the mining of scRNA-seq data to obtain a comprehensive coverage of PAS in a given species.



Sample	PAS type	Motif score	Nr PAS	Sample	PAS type	Motif score	Nr PAS
10X_P4_2	ATE	10	8968	10X_P7_11	ATE	11	9331
	UTE	10	5261		UTE	11	5329
	NTE	4	217		NTE	4	457
	I	9	1683		I	8	2386
	IG	9	6581		IG	9	7024
10X_P4_7	ATE	10	10164	10X_P7_15	ATE	11	8328
	UTE	11	6153		UTE	11	4043
	NTE	2	476		NTE	6	226
	I	8	3080		I	7	1333
	IG	9	8760		IG	9	5075
10X_P7_4	ATE	11	8647	10X_P7_14	ATE	11	9412
	UTE	11	3804		UTE	11	4810
	NTE	2	342		NTE	4	281
	I	7	1212		I	8	1799
	IG	9	5475		IG	9	6013
					Control	0	9115

Figure 7. Application of SCINPAS on *Tabula Muris Senis* samples.

A. Illustration of the motif score calculation. The positional preference of polyadenylation motifs relative to 3′ ends of annotated terminal exons was first determined. Then, the motif frequency in PAS from individual samples and classes was calculated and was deemed consistent with the annotation when the peak fell within the 90% interval around the maximum frequency for annotated terminal exons. The motif score was the number of motifs found to be consistent in a given sample and PAS class. **B.** Statistics of PAS classes in the analyzed *Tabula Muris Senis* samples. **C.** Number of genes (y-axis, $\log_{10}$ scale) with an annotated PAS ("covered" genes) from among the expressed genes in each of the analyzed *Tabula Muris Senis* samples. The control was obtained with the same processing workflow as the PAS, but starting from UMI-tools-based deduplicated reads from the *Tabula Muris Senis* sample 10X_P7_14. A minimum of 2 reads support was required for both considering a gene expressed and for considering a PAS. **D.** Total number of SCINPAS-identified PAS (y-axis, $\log_{10}$ scale) with at least 2 PATR support in each of the samples.

B. 5. Discussion

APA is one of the main mechanisms of isoform diversification in humans (1), with a wide range of consequences for cell signaling and gene expression (reviewed in (3)). In the past decade, dedicated 3′ end sequencing methods have been developed to map the relative usage of PAS across tissues and conditions, and the resulting data have been consolidated in specialized repositories (36). However, as it has become clear from various types of single cell analyses, much remains to learn about the processes that give each cell its identity and alternative polyadenylation seems to play an important role (37). scRNA-seq has opened new possibilities for studying the polyadenylation landscape of individual cell types because available technologies target mRNA 3’ ends. Yet the field has not fully exploited scRNA-seq data to extract reads that provide direct evidence for the usage of specific PAS by virtue of containing part of the poly(A) tail. While this property has been used before for PAS identification from bulk sequencing datasets (e.g. (38)), the volume of the data and the breadth of coverage of cell types afforded by scRNA-seq, especially using the technology from 10x Genomics, is unmatched.

A number of methods have already been proposed for analyzing the polyadenylation landscape from scRNA-seq data (5, 7, 8, 10, 35). However, none of these methods address the very first step in the processing pipeline, which is read deduplication. This is the focus of SCINPAS, which improves the recovery of reads containing poly(A) tails several fold. The reads without poly(A) tails are also extracted, which means that previously developed models for interpreting the entire dataset can also be used. We also implemented a procedure for identifying PAS, clustering data from closely-spaced reads, and compared the PAS that we recovered with those recovered by a recently developed method, SCAPE (6). We show that SCINPAS provides complementary information (e.g. recovering PAS in non-exonic regions) and also, much higher resolution in PAS identification. SCINPAS enables studies of cleavage site microheterogeneity, as well as detection of alternative PAS in 3’ UTRs without specific assumptions about their relative distance. A small fraction of the PAS that we classify as intergenic are located within a relatively short distance (< 1kb) downstream of terminal exon ends (Fig. 6 and S3). Small variations in the position of cleavage sites can occur for multiple reasons, including the imprecision of the processing machinery, observed in many previous studies, as well as the ambiguity of assigning terminal A nucleotides when the cleavage occurs immediately upstream of a genome-encoded A nucleotide. However, in these cases the variation is much smaller than 1kb. Further analysis of SCINPAS-identified sites along with long read data should clarify the transcription units to which these PAS belong.

The most conserved poly(A) signal that guides the 3’ end processing of pre-mRNAs is the AAUAAA hexamer, bound by the WDR33 and CPSF30 components of the 3’ end processing complex (30, 39). 12 variants of this sequence have been previously found to have a similar pattern of position-dependent enrichment upstream of the PAS (17, 29) and also to promote polyadenylation *in vitro* (40). Here we found that the peak of the AAGAAA variant was located at ~10 nucleotides upstream of the SCINPAS-identified PAS, but ~10 nucleotides *downstream* of annotated TE ends (Fig. 4, S5). To resolve this discrepancy, we also analyzed the position-dependent frequency of the motif at PAS obtained with PAPERCLIP, an orthogonal method for PAS identification that uses crosslinking and immunoprecipitation of the poly(A)-binding protein to identify *bona fide* poly(A) tails (20). In PAPERCLIP-identified PAS, AAGAAA peaked also at ~10 nucleotides upstream of PAS (Fig. 4). PAS that are located in non-terminal exons, introns and intergenic regions are more likely to contain this motif, and genes with AAGAAA-containing PAS have higher expression levels than genes that do not contain such PAS (Fig. S5). These results suggest that AAGAAA-containing PAS are non-canonical PAS that can only be observed under normal conditions when the gene expression level is high (Fig. S5). Whether they are functionally relevant in specific conditions or cell types remains to be determined in future studies. Interestingly, while AAGAAA was found to promote the polyadenylation of a substrate *in vitro* (40), it has also been observed associated with a specific class of genes; these genes have multiple PAS in both introns and exons, and they couple polyadenylation with splicing to generate long or short transcripts (41). An example studied in detail is that of the immunoglobulin E-encoding gene (42), which generates either a short, secreted form of the protein by the usage of an intronic AAUAAA PAS, or a long, membrane-bound form that depends on the usage of multiple PAS, including one containing the AAGAAA poly(A) signal. Also noted before is that AAGAAA is a splice enhancer (41, 43), and thus, the position-dependent enrichment of this signal may vary depending on the location of analyzed PAS within genes. For the other signals, the position-dependent enrichment was similar between annotated 3' ends and the PAS identified by SCINPAS, in terminal exons or elsewhere, supporting the accuracy of the method.

Altogether, these results indicate that SCINPAS is an accurate method for extracting experimentally-supported PAS from scRNA-seq data. Running SCINPAS on typical datasets as we used here takes 1~8 hours, allowing SCINPAS to be applied to the many datasets available in the public domain. While SCINPAS focuses on the extraction of PATR, it also carries out deduplication of all reads, and thus can be used in general workflows for scRNA-seq data analysis. Moreover, non-polyadenylated reads may be further taken into consideration when quantifying PAS usage starting from the experimentally-supported PAS in the system of interest. The vast volume of scRNA-seq data makes it possible to substantially improve the coverage of PAS in public repositories, to thus reach an improved understanding of PAS usage in individual cell types. This is an exciting research direction for the future. SCINPAS is available from https://github.com/zavolanlab/SCINPAS.

B. 6. Availability of data and material

SCINPAS is packaged into a nextflow workflow (24). The code and subsequent analysis is available at: https://github.com/zavolanlab/SCINPAS.

The data and additional scripts used for the plots and graphs are deposited in the zenodo repository, with the DOI 10.5281/zenodo.7868155.

B. 7. Conflict of interest statement

None declared.

B. 8. Acknowledgements

We are grateful to the sciCORE team for the maintenance of the computing infrastructure on which the computations were carried out. We thank the Zavolan group members for feedback and discussions. D.B. was a recipient of a Fellowship for Excellence from the Biozentrum, University of Basel. This project was supported in part by the Swiss National Science grant #310030_189063 to M.Z.

B. 9. Author contributions

D.B. and M.Z. conceived the study. Y.M. wrote the code. Y.M. and D.B. performed the analyses and generated the figures. Y.M., D.B. and M.Z. wrote the manuscript. The authors read and approved the final manuscript.

B. 10. References

1. Reyes,A. and Huber,W. (2018) Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.*, **46**, 582–592.

2. Gruber,A.J. and Zavolan,M. (2019) Alternative cleavage and polyadenylation in health and disease. *Nat. Rev. Genet.*, 10.1038/s41576-019-0145-z.

3. Mitschka,S. and Mayr,C. (2022) Context-specific regulation and function of mRNA alternative polyadenylation. *Nat. Rev. Mol. Cell Biol.*, 10.1038/s41580-022-00507-5.

4. Ji,G., Guan,J., Zeng,Y., Li,Q.Q. and Wu,X. (2015) Genome-wide identification and predictive modeling of polyadenylation sites in eukaryotes. *Briefings in Bioinformatics*, **16**, 304–313.

5. Yang,Y., Paul,A., Bach,T.N., Huang,Z.J. and Zhang,M.Q. (2021) Single-cell alternative polyadenylation analysis delineates GABAergic neuron types. *BMC Biol.*, **19**, 144.

6. Zhou,R., Xiao,X., He,P., Zhao,Y., Xu,M., Zheng,X., Yang,R., Chen,S., Zhou,L., Zhang,D., *et al.* (2022) SCAPE: a mixture model revealing single-cell polyadenylation diversity and cellular dynamics during cell differentiation and reprogramming. *Nucleic Acids Res.*, **50**, e66.

7. Wang,J., Chen,W., Yue,W., Hou,W., Rao,F., Zhong,H., Qi,Y., Hong,N., Ni,T. and Jin,W. (2022) Comprehensive mapping of alternative polyadenylation site usage and its dynamics at single-cell resolution. *Proc. Natl. Acad. Sci. U. S. A.*, **119**, e2113504119.

8. Wu,X., Liu,T., Ye,C., Ye,W. and Ji,G. (2021) scAPAtrap: identification and quantification of alternative polyadenylation sites from single-cell RNA-seq data. *Brief. Bioinform.*, **22**.

9. Gao,Y., Li,L., Amos,C.I. and Li,W. (2021) Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression. *Genome Res.*, **31**, 1856–1866.

10. Li,G.-W., Nan,F., Yuan,G.-H., Liu,C.-X., Liu,X., Chen,L.-L., Tian,B. and Yang,L. (2021) SCAPTURE: a deep learning-embedded pipeline that captures polyadenylation information from 3' tag-based RNA-seq of single cells. *Genome Biol.*, **22**, 221.

11. Zheng,G.X.Y., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J., *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.

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

13. Tabula Muris Consortium (2020) A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, **583**, 590–595.

14. Pace,L., Goudot,C., Zueva,E., Gueguen,P., Burgdorf,N., Waterfall,J.J., Quivy,J.-P., Almouzni,G. and Amigorena,S. (2018) The epigenetic control of stemness in CD8+ T cell fate commitment. *Science*, **359**, 177–186.

15. Lukassen,S., Bosch,E., Ekici,A.B. and Winterpacht,A. (2018) Characterization of germ cell differentiation in the male mouse through single-cell RNA sequencing. *Sci. Rep.*, **8**, 6521.

16. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

17. Gruber,A.J., Schmidt,R., Gruber,A.R., Martin,G., Ghosh,S., Belmadani,M., Keller,W. and Zavolan,M. (2016) A comprehensive analysis of 3’ end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res.*, **26**, 1145–1159.

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

19. Bonfield,J.K., Marshall,J., Danecek,P., Li,H., Ohan,V., Whitwham,A., Keane,T. and Davies,R.M. (2021) HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience*, **10**.

20. Hwang,H.-W., Park,C.Y., Goodarzi,H., Fak,J.J., Mele,A., Moore,M.J., Saito,Y. and Darnell,R.B. (2016) PAPERCLIP Identifies MicroRNA Targets and a Role of CstF64/64tau in Promoting Non-canonical poly(A) Site Usage. *Cell Rep.*, **15**, 423–435.

21. Herrmann,C.J., Schmidt,R., Kanitz,A., Artimo,P., Gruber,A.J. and Zavolan,M. (2020) PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3’ end sequencing. *Nucleic Acids Res.*, **48**, D174–D179.

22. Wilkinson,L. (2011) Ggplot2: Elegant graphics for data analysis by WICKHAM, H. *Biometrics*, **67**, 678–679.

23. Thorvaldsdóttir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.

24. Di Tommaso,P., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.

25. Mayr,C. (2016) Evolution and Biological Roles of Alternative 3’UTRs. *Trends Cell Biol.*, **26**, 227–237.

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

27. Legendre,M. and Gautheret,D. (2003) Sequence determinants in human polyadenylation site selection. *BMC Genomics*, **4**, 7.

28. Ozsolak,F., Kapranov,P., Foissac,S., Kim,S.W., Fishilevich,E., Monaghan,A.P., John,B. and Milos,P.M. (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*, **143**, 1018–1029.

29. Beaudoing,E., Freier,S., Wyatt,J.R., Claverie,J.M. and Gautheret,D. (2000) Patterns of variant polyadenylation signal usage in human genes. *Genome Res.*, **10**, 1001–1010.

30. Schönemann,L., Kühn,U., Martin,G., Schäfer,P., Gruber,A.R., Keller,W., Zavolan,M. and Wahle,E. (2014) Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Genes Dev.*, **28**, 2381–2393.

31. Sandberg,R., Neilson,J.R., Sarma,A., Sharp,P.A. and Burge,C.B. (2008) Proliferating Cells Express mRNAs with Shortened 3’ Untranslated Regions and Fewer MicroRNA Target Sites. *Science*, **320**,

1643–1647.

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

33. Shulman,E.D. and Elkon,R. (2019) Cell-type-specific analysis of alternative polyadenylation using single-cell transcriptomics data. *Nucleic Acids Res.*, **47**, 10027–10039.

34. Patrick,R., Humphreys,D.T., Janbandhu,V., Oshlack,A., Ho,J.W.K., Harvey,R.P. and Lo,K.K. (2020) Sierra: discovery of differential transcript usage from polyA-captured single-cell RNA-seq data. *Genome Biol.*, **21**, 167.

35. Li,W.V., Zheng,D., Wang,R. and Tian,B. (2021) MAAPER: model-based analysis of alternative polyadenylation using 3′ end-linked reads. *Genome Biol.*, **22**, 222.

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

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

38. Smibert,P., Miura,P., Westholm,J.O., Shenker,S., May,G., Duff,M.O., Zhang,D., Eads,B.D., Carlson,J., Brown,J.B., *et al.* (2012) Global patterns of tissue-specific alternative polyadenylation in Drosophila. *Cell Rep.*, **1**, 277–289.

39. Chan,S.L., Huppertz,I., Yao,C., Weng,L., Moresco,J.J., Yates,J.R., Ule,J., Manley,J.L. and Shi,Y. (2014) CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3′ processing. *Genes Dev.*, **28**, 2370–2380.

40. Sheets,M.D., Ogg,S.C. and Wickens,M.P. (1990) Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res.*, **18**, 5799–5805.

41. Tian,B., Hu,J., Zhang,H. and Lutz,C.S. (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.*, **33**, 201–212.

42. Karnowski,A., Achatz-Straussberger,G., Klockenbusch,C., Achatz,G. and Lamers,M.C. (2006) Inefficient processing of mRNA for the membrane form of IgE is a genetic mechanism to limit recruitment of IgE-secreting cells. *Eur. J. Immunol.*, **36**, 1917–1925.

43. Fairbrother,W.G., Yeo,G.W., Yeh,R., Goldstein,P., Mawson,M., Sharp,P.A. and Burge,C.B. (2004) RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res.*, **32**, W187–90.

B. 11. Supplementary Figures

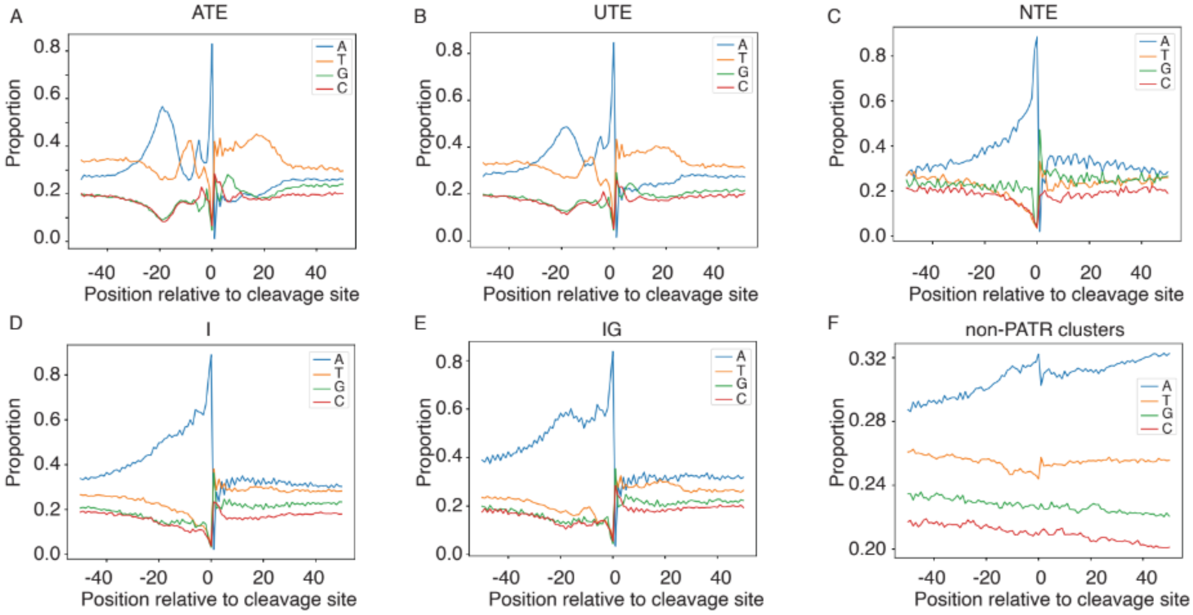


Figure S1. Position-dependent nucleotide frequencies in PAS constructed from SCINPAS-extracted reads. Data was from *Tabula Muris Senis* sample 10X_P7_14. PAS were anchored at position 0, and the genomic sequence upstream and downstream (from -50 to +50 nucleotides) was used to calculate nucleotide frequencies. ATE: annotated terminal exon, UTE: unannotated terminal exon, NTE: non-terminal exons, I: intronic, IG: intergenic regions, PATR: poly(A) containing reads.

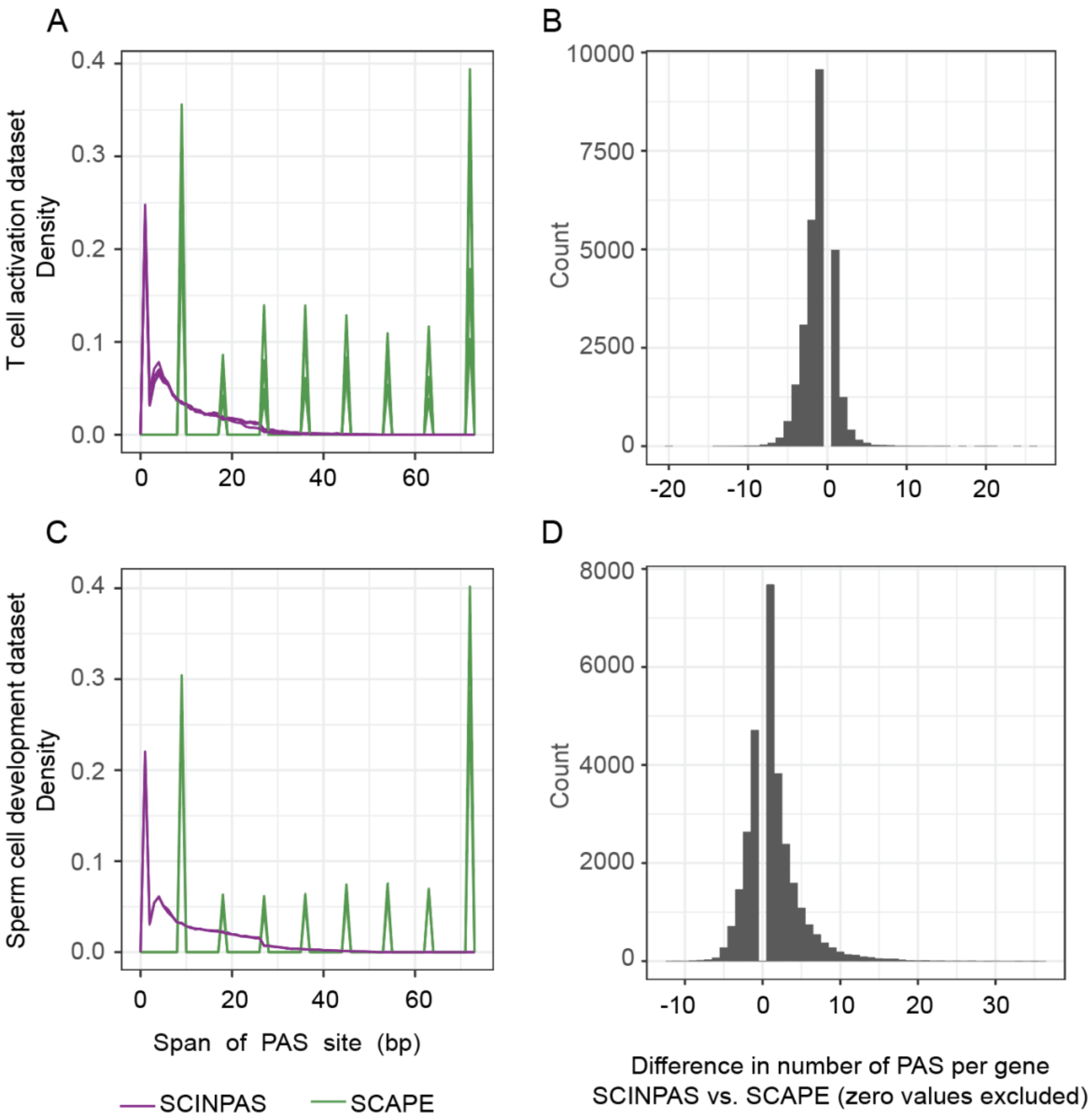


Figure S2. PAS cluster statistics.

A. Density distribution of span of PAS clusters, for each sample of T cell activation dataset separately. SCAPE displays discrete PAS cluster spans. **B.** Histogram of the difference in PAS site count identified within individual genes by SCINPAS and SCAPE. Zero difference excluded. A value > 0 means that SCINPAS identified more PAS clusters in this gene compared to SCAPE. **C.** Same as **A**, but on samples of sperm cell development dataset. **D.** Same as **B**, but on samples of sperm cell development dataset.

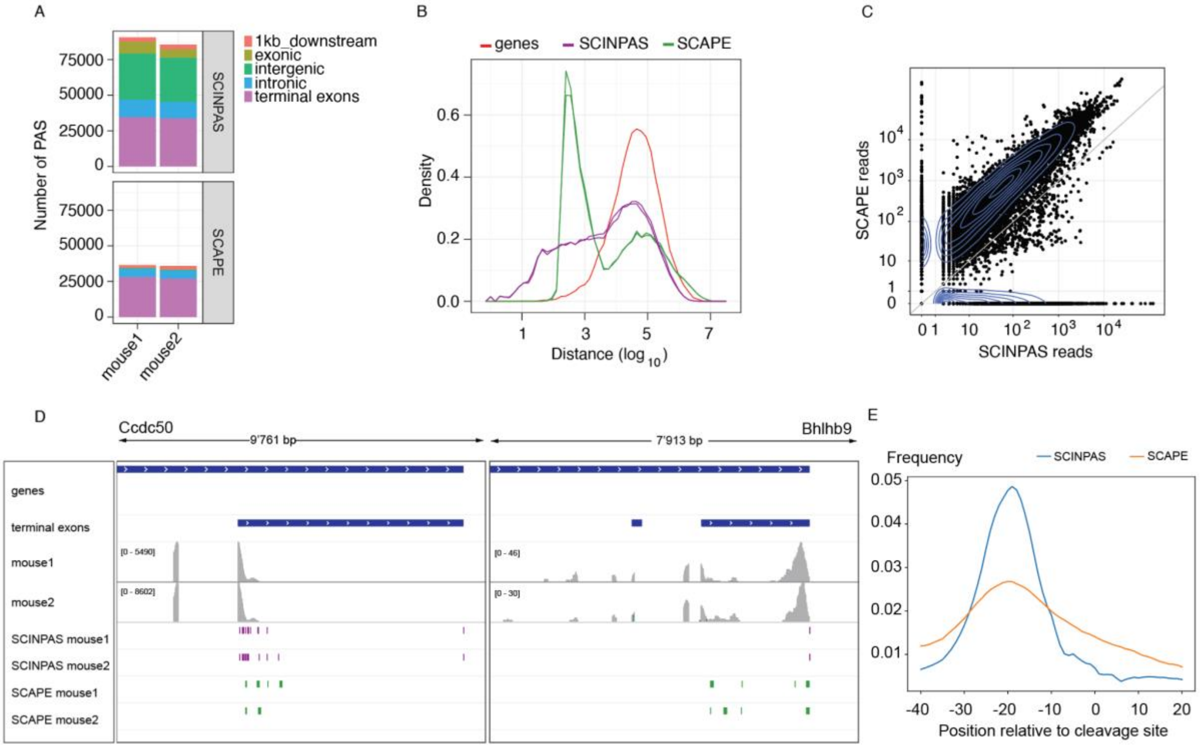


Figure S3. PAS recovery by SCINPAS and SCAPE from the sperm cell development dataset.

A. Number of PAS inferred by SCINPAS (top) and SCAPE (bottom) from the sperm cell development datasets. The colors indicate different classes of PAS (see legend). **B.** Distribution of the distances from each PAS to the closest PAS downstream for SCINPAS (purple) SCAPE (green). For comparison, the distribution of 3'end-to-5'end distances between genes is shown in red. **C.** Scatter of the total number of PAS-associated reads within a gene for SCINPAS (x-axis) and SCAPE (y-axis). Spearman correlation coefficient was 0.6 (p-value < 2.2e-16). The diagonal of equal read counts is shown in gray. 2D kernel density estimates obtained with the geom_density_2d(n=200) function of ggplot2 are shown as blue contours. **D.** Examples of PAS recovered by SCINPAS (purple) and SCAPE (green) in the Ccdc50 (left) and Bhlhb9 (right) genes, from the sperm cell development dataset. Genes and terminal exons are shown in the IGV browser (23) in blue, the coverage tracks shown in gray. **E.** Position-dependent distribution of the canonical polyadenylation signal AAUAAA around SCINPAS- and SCAPE-identified PAS.

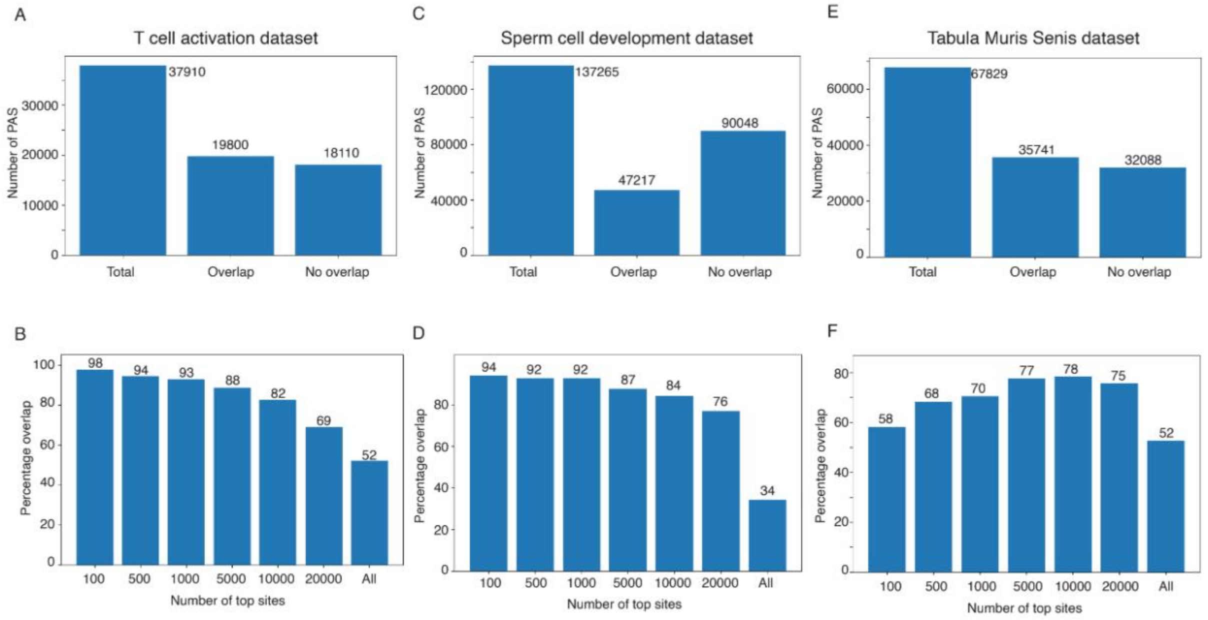


Figure S4. Overlap of PAS sites with the polyAsite atlas.

A. Number of PAS identified by SCINPAS in the T cell activation dataset and the overlap with the polyAsite atlas. **B.** Proportion of top sites (ranked by abundance) from the T cell activation dataset that overlap with the polyAsite atlas. **C.** Same as **A** for the sperm cell development dataset. **D.** Same as **B** for the sperm cell development dataset. **E.** Same as **A** but for the *Tabula Muris Senis* dataset. **F.** Same as **B** but for the *Tabula Muris Senis* dataset.

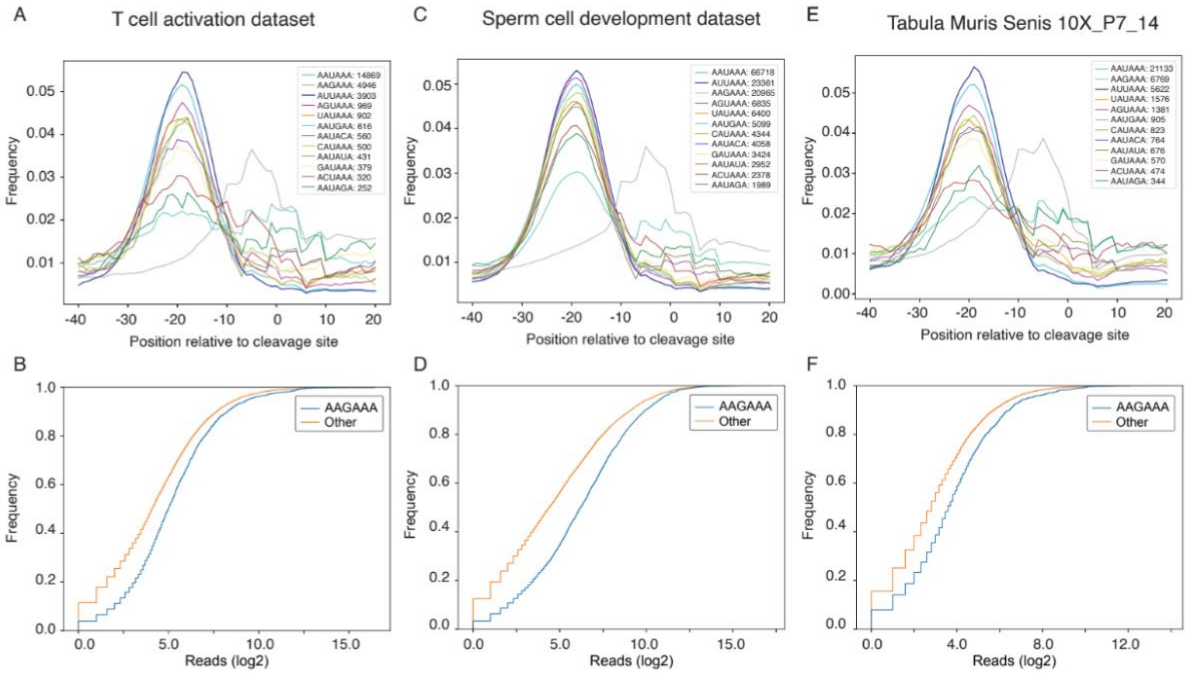


Figure S5. Poly(A) signal frequency in different data sets.

A,C,E: Position-dependent frequency of poly(A) signals (color) across datasets: T cell activation (**A**), sperm cell development (**C**) and *Tabula muris Senis* 10X_P7_14 (**E**). The hexamer AAGAAA (grey) displays a distinct pattern in comparison to the other poly(A) signals. **B,D,F:** cumulative distribution of gene expression levels (total number of reads in terminal exons), for genes that have at least one PAS containing the AAGAAA motif within -40 to +20 nucleotides around the PAS (blue) compared to all other expressed genes (orange). The data sets are in the same order as in the top row panels.

Appendix C

Extensible benchmarking of methods that identify and quantify polyadenylation sites from RNA-seq data

Sam Bryce-Smith[1]*, Dominik Burri[2,3]*, Matthew R. Gazzara[4]*, Christina J. Herrmann[2,3]*, Weronika Danecka[5]^, Christina M. Fitzsimmons[6]^, Yuk Kei Wan[7,8]^, Farica Zhuang[9]^, Mervin M. Fansler[10,11], José M. Fernández[12,13], Meritxell Ferret[12,13], Asier Gonzalez-Uriarte[12,13], Samuel Haynes[5], Chelsea Herdman[14], Alexander Kanitz[2,3], Maria Katsantoni[2,3], Federico Marini[15], Euan McDonnel[16], Ben Nicolet[17], Chi-Lam Poon[18], Gregor Rot[3,19], Leonard Schärfen[20], Pin-Jou Wu[21], Yoseop Yoon[22], Yoseph Barash[4,9]#, Mihaela Zavolan[2,3]#

Within each respective category, authors are listed alphabetically.
*These authors contributed equally
^These authors contributed equally
#To whom correspondence should be addressed:
yosephb@upenn.edu and mihaela.zavolan@unibas.ch

C. 1. Author Affiliations

A UCL Queen Square Motor Neuron Disease Centre, Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, UCL, London, UK.

B Biozentrum, University of Basel, Basel, Switzerland

C Swiss Institute of Bioinformatics, Lausanne, Switzerland

D Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

E Institute for Cell Biology, School of Biological Sciences, The University of Edinburgh, Edinburgh, United Kingdom

F Laboratory of Cell Biology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA.

G Genome Institute of Singapore, Buona Vista, Singapore

H National University of Singapore, Kent Ridge, Singapore

I Department of Computer and Information Science, School of Engineering, University of Pennsylvania, Philadelphia, USA

J Tri-Institutional Program in Computational Biology and Medicine, Weill Cornell Graduate Studies, New York, NY, USA.

K Cancer Biology and Genetics, Sloan-Kettering Institute, MSKCC, New York, NY, USA.

L Barcelona Supercomputing Center, Barcelona, Spain

M Spanish National Bioinformatics Institute (INB/ELIXIR-ES)

N Department of Neurobiology, University of Utah, Utah, USA

O Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI) - University Medical Center of the Johannes Gutenberg, University Mainz, Germany

P Leeds Institute for Data Analytics, School of Molecular and Cellular Biology, University of Leeds, United Kingdom

Q Department of Hematopoiesis, Sanquin Research, Landsteiner Laboratory, Amsterdam UMC, University of Amsterdam, and Oncode Institute, Amsterdam, The Netherlands.
R Weill Cornell Medicine, New York, NY, USA
S Institute of Molecular Life Sciences, Zurich, Switzerland
T Department of Molecular Biophysics & Biochemistry, Yale University, New Haven CT, USA
U Center for Plant Molecular Biology (ZMBP), University of Tübingen, Germany
V Department of Microbiology and Molecular Genetics, School of Medicine, University of California Irvine, Irvine, California, USA

Short Title:
APAeval: Benchmarking methods for APA from RNA-seq

Keywords:
Benchmarking, (alternative) polyadenylation, bioinformatics, RNA-seq, community initiative

C. 2. ABSTRACT:

The tremendous rate with which data is generated and analysis methods emerge makes it increasingly difficult to keep track of their domain of applicability, assumptions, limitations and consequently, of the efficacy and precision with which they solve specific tasks. Therefore, there is an increasing need for benchmarks, and for the provision of infrastructure for continuous method evaluation. APAeval is an international community effort, organized by the RNA Society in 2021, to benchmark tools for the identification and quantification of the usage of alternative polyadenylation (APA) sites from short-read, bulk RNA-sequencing (RNA-seq) data. Here, we reviewed 17 tools and benchmarked eight on their ability to perform APA identification and quantification, using a comprehensive set of RNA-seq experiments comprising real, synthetic, and matched 3’-end sequencing data. To support continuous benchmarking, we have incorporated the results into the OpenEBench online platform, which allows for continuous extension of the set of methods, metrics, and challenges. We envisage that our analyses will assist researchers in selecting the appropriate tools for their studies, while the containers and reproducible workflows could easily be deployed and extended to evaluate new methods or datasets.

C. 3. INTRODUCTION:

Alternative polyadenylation (APA) is a mechanism of RNA processing that generates distinct 3′ termini, allowing the expression of multiple transcript isoforms from a single genomic locus as been observed in nearly all eukaryotes, and in humans (Tian & Manley, 2017). The choice of different polyadenylation sites (PAS) can give rise to protein isoforms with distinct C-termini and thereby distinct functions. Even without changing the coded protein, APA-derived changes to 3′ untranslated regions (UTRs) can impact gene expression by influencing mRNAs' nuclear export, interactions with miRNA or RNA binding proteins, stability, and translational efficiency (Elkon et al., 2013). It is estimated that over 70 percent of all genes produce alternatively polyadenylated mRNAs (Derti et al., 2012). A number of studies have reported a critical role for APA-mediated gene regulation during development (Ji et al., 2009a; Lianoglou et al., 2013; Sommerkamp et al., 2021; Yoon et al., 2019) or disease (Goering et al., 2021; Morris et al., 2012; reviewed in Gruber & Zavolan, 2019)

Given the importance of APA, identifying and quantifying the usage of polyadenylation sites on a transcriptome-wide scale is critical for understanding both the underlying mechanisms and functional implications of APA-mediated gene regulation. Early studies of APA used microarray platforms and discovered widespread changes in PAS usage (Flavell et al., 2008; Ji et al., 2009b). While these studies laid the groundwork for large scale study of APA-mediated gene regulation, they were limited by the dependence of the microarray design on previously annotated transcript isoforms. With the advancement of high throughput sequencing (HTS) technologies, scientists have developed a number of targeted 3′-end sequencing methods for global profiling of PAS usage. Most methods utilize oligo(dT)-based reverse transcription to enrich reads derived from mRNA 3′ ends (Derti et al., 2012; Lianoglou et al., 2013; Martin et al., 2012; Sanfilippo et al., 2017; Shepard et al., 2011; Yoon et al., 2021; Zhou et al., 2016) while other methods developed oligo(dT)-independent approaches to avoid the issue of internal priming (Hoque et al., 2013; Hwang et al., 2016; Jan et al., 2011; Ogorodnikov & Danckwardt, 2021; Zheng et al., 2016). These studies identified a large number of previously unknown sites and also demonstrated cell- and tissue-specific regulation of APA. However, the number of datasets generated by targeted 3′-end sequencing remains limited compared to the enormous amount of publicly available RNA-seq datasets.

Unlike the aforementioned methods, standard RNA-seq does not target the 3′ end of transcripts. Instead, reads are sampled from the entire length of any expressed isoform. Computational methods to detect and quantify APA usage from such data generally rely on the pattern of coverage of a genomic locus by reads, which is a superposition of the sampled isoforms. However, the large fluctuations in coverage even along loci expressing single isoforms make it challenging to identify drops in read coverage that correspond to 3′ ends. As a consequence, many computational methods have been developed by various labs in the context of specific projects to answer often related, but not identical questions. The reliance of many researchers within the RNA community on these computational tools for data analysis points to a clear need for their comprehensive evaluation. A few previous reports have endeavored to benchmark computational methods for APA analysis from RNA-seq data. Notably, (Chen et al., 2020) described their efforts to review 11 methods, using RNA-seq data sets from human, mouse, and *Arabidopsis* in their analysis. While the study provided assessments of the precision and sensitivity of PAS site inference, the RNA-seq data sets were pre-processed using different tools, making the results difficult to interpret and compare.

Moreover, the code was not presented in a manner that made it easily reproducible and extendable. Other benchmarking efforts have similar shortcomings (Chen et al., 2020; Shah et al., 2021; W. Ye et al., 2022).

To specifically address the issues of reproducible and continuous benchmarking, we organized a hackathon focused on software that detects and quantifies poly(A) sites from RNA-seq data. This international effort in the RNA community had several goals: (1) To bring together RNA biologists, bioinformaticians, and developers within the RNA Society to foster the dialog between RNA researchers of different backgrounds; (2) to provide informative benchmarking results for current methods for APA detection and quantification; and (3) to develop a framework for reproducible, cloud-based benchmarking for bioinformatic tools. This benchmarking infrastructure was designed to be modular, extendable, and standardized at all levels, with the idea that additional tools or metrics could be added in the future, or the infrastructure applied to the benchmarking of other types of tools.

Overall, we benchmarked eight methods on five RNA-seq data sets and compared the results to five "ground truth" datasets of known APA sites. This work constitutes the most extensive reproducible evaluation of APA detection and quantification methods to date. In addition to being described here, selected benchmarks and results are made available on the ELIXIR benchmarking platform OpenEBench (https://openebench.bsc.es/benchmarking/OEBC007)(Capella-Gutierrez et al., 2017).

C. 4. RESULTS:

C. 4. 1. Methods selected for benchmarking

From an algorithmic point of view, methods for identifying or quantifying APA from conventional short-read RNA-seq data can broadly be grouped into two categories. In the first group are methods that utilize annotated poly(A) sites, including QAPA (Ha et al., 2018) and PAQR (Gruber et al., 2018). While they can be used to estimate (differential) poly(A) site usage, these methods have the limitation that they cannot identify novel sites, beyond those listed in poly(A) site databases or implied in the genome annotation. In the second group are methods that can perform *de novo* identification of sites based on changes in the read coverage along the mRNAs, as this is expected to drop at mRNA 3′ ends. This category includes DaPars, DaPars2, GETUTR, TAPAS, IsoSCM, and APAtrap (Arefeen et al., 2018; Feng et al., 2018; Kim et al., 2015; Shenker et al., 2015; Xia et al., 2014; C. Ye et al., 2018).

One of the main goals of APAeval was to make all benchmarked methods accessible for the larger RNA community. Thus, beyond the performance of the method, we assessed additional factors that included the ease of installation, tool accessibility for new users, the breadth of use of the method (a metric biased towards older methods), the responsiveness of the authors to email questions or GitHub issues, and whether or not the method is currently being maintained (a metric biased towards recent methods). Based on these criteria, we evaluated 17 methods for their ability to perform our benchmarking challenges (Table 1). This list was narrowed to eight methods that we were able to install and run on the selected benchmarking datasets (see below), that performed robustly, and whose output was compatible with our evaluation metrics.

Table 1. List of methods evaluated in APAeval.

We evaluated 17 methods for possible inclusion in APAeval. Green background: benchmarked in APAeval. **Reasons for exclusion from APAeval benchmarking (columns 6 & 7):** [1] Incompatibility with APAeval input; [2] Incompatibility with APAeval metrics; [3] Reported bugs not fixed by authors; [4] Other (unable to install/run etc.). **Remarks on method workflows created in APAeval (column 7):** [5] workflow has very high time or memory consumption, [6] workflow only tested on small test files, [7] workflow does not include building of machine-learning model ; uses authors' published model instead, [8] workflow contains steps to build custom annotation, defaults are hardcoded, and parameters for pseudoalignment cannot be changed; benchmarking was run also with the annotation used by the authors in the original publication, [9] differential usage functionality of the method is not implemented in the APAeval workflow. **Other remarks:** [10] features present according to publication/manual but not tested by APAeval.

Note that if a tool can perform absolute quantification it qualifies for our APAeval relative quantification event, even if it doesn't produce a dedicated relative quantification output. The benchmarking of differential poly(A) site usage is not discussed in the current publication.

Method	PAS Identification	Absolute PAS Quantification	Relative PAS Quantification	Differential PAS usage	Benchmarked in APAeval?	APAeval method workflow	Citation
APA-Scan	Yes	No	Yes	Yes	No[1,3,4]	No[3,4]	Fahmi et al. 2022
APAlyzer	No	No	Yes	Yes	No[2]	Snakemake[2]	Wang et al. 2020

APAtrap	Yes	Yes	Yes	Yes	Yes	Nextflow[5]	Ye et al. 2018
Aptardi	Yes	No	No	No	No[4]	Nextflow[5,6,7]	Lusk et al. 2021
CSI-UTR	No	No	No	Yes	No[1,3]	Nextflow[6]	Harrison et al. 2019
DaPars	Yes	No	Yes	Yes	Yes	Nextflow	Xia et al. 2014
DaPars2	Yes	No	Yes	No	Yes	Snakemake	Feng et al. 2018
diffUTR	No	No	No	Yes	No[2]	Nextflow[6]	Gerber et al. 2021
GETUTR	Yes	No	Yes	Yes	Yes	Nextflow	Kim et al. 2015
IsoSCM	Yes	No	Yes	Yes	Yes	Nextflow	Shenker et al. 2015
LABRAT	No	No	Yes	Yes	No[2]	Nextflow[2,6]	Goering et al. 2021
MISO	No	Yes	Yes	Yes	No[1,4]	No[1,4]	Katz et al. 2010
mountain Climber	Yes[10]	Yes[10]	Yes[10]	Yes[10]	No[3,4]	No[3,4]	Cass et al. 2019
PAQR	No	Yes	Yes	Yes	Yes	Snakemake	Gruber et al. 2018
QAPA	No	Yes	Yes	No	Yes	Nextflow[8]	Ha et al. 2018
Roar	No	No	Yes	Yes	No[1]	Snakemake[2,6]	Grassi et al. 2016
TAPAS	Yes	No	Yes	Yes	Yes	Nextflow[9]	Arefeen et al. 2018

Experimental data

To be able to evaluate the inferences made by a computational method, a ground truth, independent (orthogonal) dataset is necessary. In our case, the ideal dataset would have RNA-seq data for samples where the precise abundance of transcript isoforms is known and can be directly compared with the abundance inferred by the method for APA site inference. Quantifying transcript abundances genome-wide with a method different than RNA-seq is not generally done. However, there are studies in which both RNA-seq and 3′ end sequencing data have been obtained from the same experimental system. Therefore, we used two types of "ground truth" estimates of PAS usage: 1) from targeted sequencing of mRNA 3′ ends; and 2) from "realistic" simulated data, where reads were sampled so as to match transcript isoform expression levels estimated by the RSEM algorithm (Li & Dewey, 2011) from specific RNA-seq samples in the GTEx compendium of human tissue data (The GTEx Consortium, 2020).

Based on the above ground truth definitions, we searched for publicly available RNA-seq datasets with matching 3′-end sequencing data. To represent diverse biological conditions, we selected both human and mouse datasets, originating both from cell lines (HEK293 or P19 cells) and primary tissues (cortex and immune cell populations). Moreover, to avoid potential biases caused by technical characteristics of datasets, we chose datasets with varying sequencing depths (30 to 200 million reads) and from both paired and single-end sequencing. The ground truth data was obtained with a broad range of techniques for 3′ end sequencing, namely 3′-seq (Lianoglou et al., 2013), MACE-seq (Zawada et al., 2014), A-seq2 (Martin et al., 2017) and PAPERCLIP (Hwang et al., 2016). Where possible, we selected datasets where the RNA-seq and orthogonal 3′-end sequencing data were generated by the same lab. The GTEx-based "realistic" simulated data was derived from a recent study (Vaquero-Garcia et al., 2023). For comparability and reproducibility, we processed all raw RNA-seq data with the same workflow, described in the Materials and Methods section. RNA-seq/ground truth pairs are summarized in Supplemental Table 1. RNA-seq dataset characteristics are shown in Supplemental Fig. 1, and the distribution of mRNA abundances in the ground truth data is shown in Supplemental Fig. 2A.

C. 4. 2. The APAeval benchmarking workflow

APAeval benchmarking was divided into three "events", according to the different tasks the evaluated methods perform: 1) identification of *de novo* poly(A) sites; 2) quantification of the usage of individual poly(A) sites within the transcriptome; and 3) quantification of the relative usage of poly(A) sites compared to other sites within the same terminal exon (TE). For each of these events, specific metrics were defined (see below) and then computed for each dataset pair separately.

The benchmarking infrastructure that we developed contains two types of modules: workflows to execute individual methods ("method workflows") and workflows to compute the benchmarking metrics for all evaluated methods ("benchmarking workflows"). Detailed input and output specifications are available in the APAeval Github repository (https://github.com/iRNA-COSI/APAeval/) and additional details can be found in the Material and Methods section. We used standard data formats, BAM files for aligned reads, FASTA files for nucleotide sequences, GTF files for gene/transcript model annotations and BED files for reference poly(A) sites. We included all the code necessary to make the inputs compatible with a method as well as to generate the standardized outputs in the respective method workflow. To ensure reproducibility and extensibility of our work, whenever Docker or Singularity containers for a method were not readily available, we custom-built them. As some methods were developed for a specific genome annotation, we carried out the analysis both with the preferred annotation of the tool and the corresponding GENCODE annotations

(Release 38 for real human data, release 26 for simulated data, and release M18 for mouse) (Frankish et al., 2021). A graphic summary of the entire workflow is shown in Figure 1.

Identification of poly(A) sites based on the profile of genome coverage by RNA-seq reads is challenging to achieve with nucleotide-level precision. For this reason we tested associating poly(A) sites identified from RNA-seq data by the benchmarked method (PD - prediction) to poly(A) sites detected in an orthogonal 3′-end sequencing dataset (GT - ground truth) with various degrees of tolerance in calling corresponding sites. Specifically, the coordinates of ground truth sites were extended by *n* nucleotides (*n* = 10, 25, 50 or 100 nucleotides) in both directions (Supplemental Fig. 3) and the BEDTools window (Quinlan & Hall, 2010) tool was used to find PD sites that intersected these windows. Note that we use the term “predicted” to denote outputs of computational methods applied to RNA-seq data, to distinguish them from the corresponding values in the ground truth datasets derived from simulation or obtained experimentally by 3’ end sequencing.

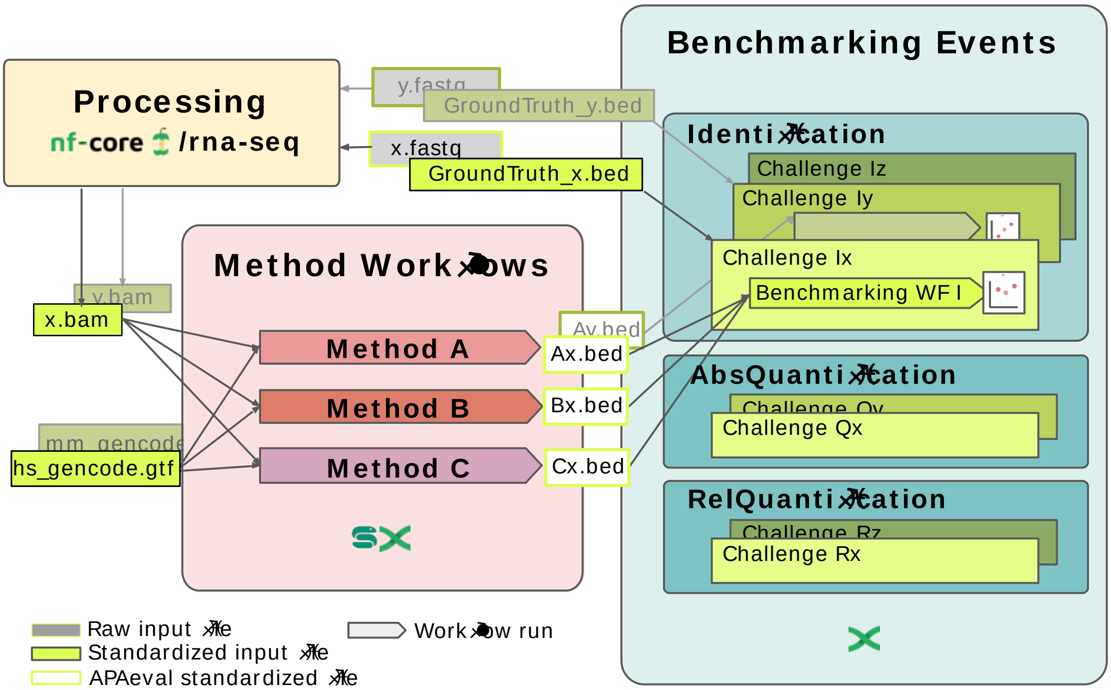


Figure 1: Overview of APAeval benchmarking strategy.
RNA-seq data (“x.fastq”) was processed with the nf-core RNA-seq pipeline (nf-core/rna-seq) for quality control and mapping. The matching ground truth data (“GroundTruth_x.bed”) was retrieved from the respective publications in bed format. The processed input data (“x.bam”), as well as a genome annotation (“hs_gencode.gtf”), and if required a reference PAS atlas in BED format (not shown), were provided to the benchmarked methods. For running the methods, a reusable “method workflow” was written for each tool in either Snakemake or Nextflow. Each method workflow contains all necessary pre- and post-processing steps needed to process data from the input formats provided by APAeval, to the format required for the benchmarking workflows (“Ax.bed”, “Bx.bed”, etc.). For each benchmarking event (“Identification”, ”AbsoluteQuantification”, “RelativeQuantification”), a reusable “benchmarking workflow” was written to compute a defined set of metrics from the comparison of outputs of method workflows with the corresponding ground truth data. Finally, the metrics for all methods for all datasets were compared within each event.

C. 4. 3. PAS identification

The first metric we evaluated was the ability of various tools to perform identification of *de novo* poly(A) sites. We tested the tools both on our reference annotation (GENCODE, Supplemental Fig. 4) as well as using the tools' preferred annotation (Fig. 2). The sensitivity, specificity, and Jaccard Index (see details in the Methods) with a window size of 50 bases indicate comparable performance of most methods across datasets, with GETUTR having relatively poor performance and TAPAS and IsoSCM relatively good performance as defined by sensitivity and precision. (Fig. 2, Supplemental Fig. 3, Supplemental Fig. 4). DaPars2 performed similarly to its predecessor, DaPars. On simulated data, the precision was much higher than on real data, at the cost of lower sensitivity. This difference from real data is likely due to the fact that the simulated data does not capture all of the sources of variation in real RNA-Seq coverage at UTR ends well. This leads to higher precision values but, by design, assigns definite "truth" even to lowly covered genes (Supplemental Fig. 2A, compare the left panel (simulated) to other panels (real data)) where APA methods struggle, resulting in lower sensitivity. The Jaccard indices calculated for different methods and samples are in the range of 0.1-0.2, with the same methods as above having higher or lower values. With the exception of GETUTR, the values obtained on the simulation data are at the high end for each of the methods. This suggests that the experimental variability affects site identification for PD, GT, or both, reducing their overlap. However, we did not detect any obvious dependence of the methods' performance on quality metrics of the RNA-seq data (Supplemental Fig. 5A). Finally, we find that the number of identified PAS matched the number of sites found in the GT data for approximately 25% of the genes. This proportion is typically lower in the simulation data, which is again an indication that the simulation data contains isoforms with very low abundance (Supplemental Fig. 2A, compare the left panel (simulated) to other panels (real data)) that are not detected by any of the tools, leading to an underestimation of used PAS. Altogether these results indicate that even when calling only PAS of highly expressed transcript isoforms, the non-uniform coverage of mRNAs by RNA-seq reads makes it difficult to reliably detect drops in coverage for PAS identification. Nevertheless, TAPAS showed consistently better performance in the PAS identification task compared to other methods.

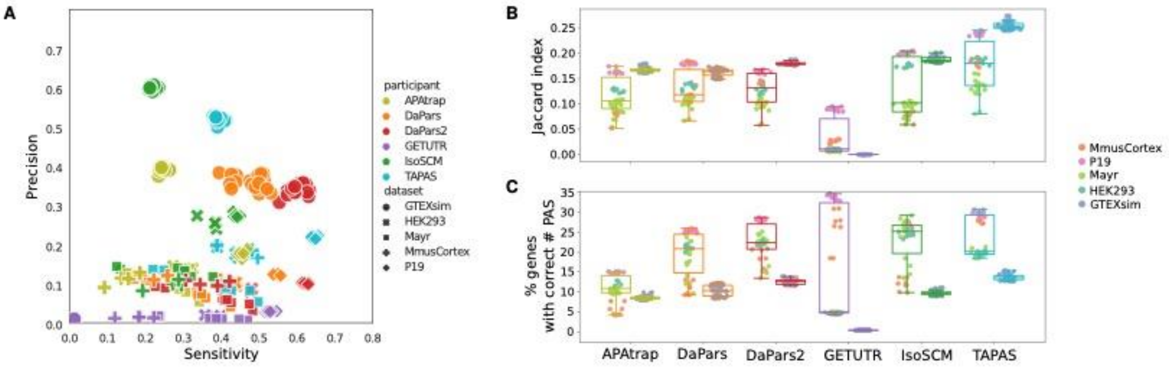


Figure 2: Results of the PAS identification event.

Predicted site locations were extended by 50 nucleotides in both directions before the intersection with GT sites and each tool was given their preferred annotation (if specified by the developers) to identify the PAS. Results using GENCODE annotation are given in Supplemental Fig. 4. (**A**) Scatter plot of precision versus sensitivity. Each symbol corresponds to a sample-tool pair, with the shape of the symbol indicating the sample set and its color indicating the tool. (**B**) Box plots of Jaccard indices indicating the overlap of predicted and ground truth sites, with predicted sites being extended symmetrically by 50 nucleotides. The tools used to predict the sites are shown on the x-axis, each with two associated box plots, one for the real data (left) and another for simulated data (right). Each point is labeled according to the code given in the legend. **(C)** Percentage of genes for which the number of PD sites was the same as the number of GT sites. Color scheme and organization as in B.

C. 4. 4. Absolute Quantification

The next task was to assess how accurate was the estimation of PAS usage (quantified in Transcripts Per Million (TPM) provided by a given tool. For simplicity of terminology, we call this "absolute quantification". Three tools provided such values and were thus included in this analysis: PAQR, APAtrap, and QAPA. QAPA and APAtrap were each tested with two different annotations: QAPA with GENCODE (Frankish et al., 2021) and a custom annotation provided by the authors, and APAtrap with GENCODE and RefSeq (O'Leary et al., 2016). PAQR was only tested with GENCODE and the results are duplicated in Figure 3A.

Given that the site identification from RNA-seq data does not generally have nucleotide-level precision, if one PD site matched multiple GT sites its score (expression level as TPM) was split between the GT sites into shares inversely proportional to the distance of the PD site to the respective GT sites. If multiple PD sites matched one GT site they were merged and their scores were summed up. By default, we included all sites estimated to have expression level > 0, but we also explored the impact of filtering the data, using in the analysis only predicted and ground truth sites with TPM > 1. Additionally, we determined the fraction of the gene expression coming from unmatched sites ("pct-FP"), to evaluate whether the accuracy of quantification is associated with the ability of a method to correctly identify used PAS.

First, as for the identification task, we determined the precision and sensitivity with which these methods identify the used PAS. Compared to APAtrap, which identifies PAS *de novo* and quantifies their usage, both PAQR and QAPA, methods that quantify the usage of annotated sites, achieve better performance, as they focus on PAS that have been found to be functional in prior data (Fig. 3A). PAQR is more conservative, achieving relatively high precision at low sensitivity, while QAPA has a much higher sensitivity at the cost of lower precision. As with the other methods presented above, precision is high on simulated data at the cost of sensitivity. Next, we determined the Pearson correlation coefficient of PD expression with the expression in the GT data (Fig. 3B). To assess how the method's performance is due to mis-allocation of reads, we simultaneously determined the fraction of expression that is assigned to false positive sites, which are not present in the GT set. The correlations between measures were good, especially for PAQR and QAPA, two methods that assign most of the expression in a given sample to sites that are also present in the GT. Interestingly, these methods give lower F1-scores on the simulated data than on the real data (Fig. 3C). This is likely another reflection of a discrepancy between the abundance of PAS isoforms in the simulation and the real data, which leads to low-abundance sites from the simulation data not being quantified by the computational tools. Indeed, the distribution of expression values (TPM) had a much more prominent peak in the simulated data set compared to the real datasets (Supplemental Fig. 2). Consequently, filtering for TPM > 1 only improved performance of the methods on the simulated data, albeit to a negligible extent (Increase of mean Pearson R by approximately 0.01; data not shown). As with the identification task, the performance of the methods was generally slightly better when using their preferred annotation as opposed to an independently chosen standard (GENCODE). An exception was QAPA on simulated data

(Supplemental Fig. 6). Again, we did not detect any significant dependence of a method's performance on quality metrics of the RNA-seq data, though QAPA tends to perform better with higher-quality (i.e. deeper sequencing, longer reads, more replicates) RNA-seq input data (Supplemental Fig. 5B).

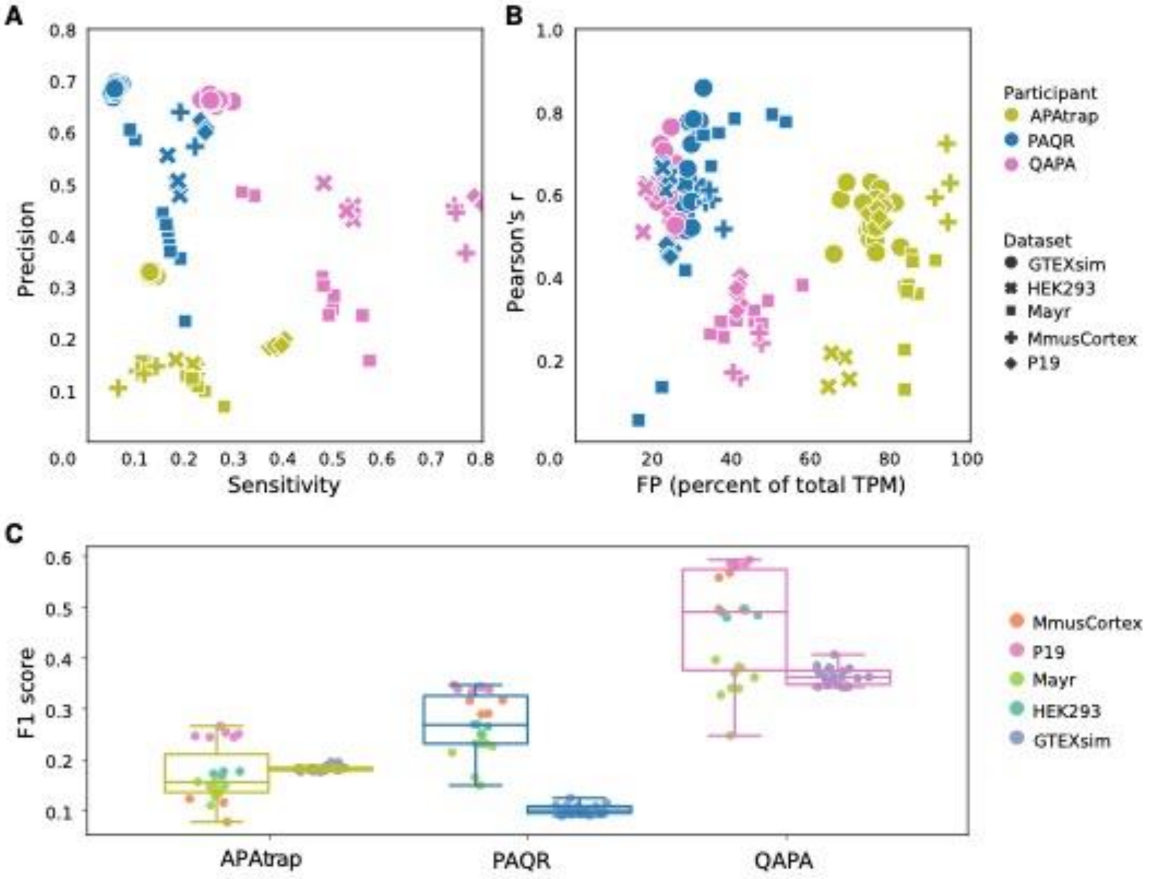


Figure 3: Results of PAS isoform quantification.

(A) Scatter plot of precision vs. sensitivity. Each sample and tool combination is represented as a symbol, with shape and color defined in the legend. **(B)** Pearson correlation of PD and GT site expression. The correlation coefficient for each sample is plotted against the percentage of total TPM that a method attributes to PAS that are not expressed in the ground truth (pct-FP). **(C)** Box plots of F1 scores. Box plots are drawn separately for real (left, see color scheme in the legend) and simulation (right) samples.

C. 4. 5. Relative quantification

When quantifying PAS site usage it is often useful to report the relative usage of each PAS within a 3'UTR or a gene, rather than the absolute expression in TPM, particularly when one is interested in the usage of PAS in different conditions in which gene expression may change. Because of this, many tools, including highly cited tools such as DaPars or DaPars2, exclusively report relative PAS usage in some form, which made them ineligible for our definition of absolute quantification benchmarking. Additionally, all tools which we benchmarked for absolute quantification also report some form of relative PAS usage. Therefore, we sought to extend our quantification benchmark to include relative quantification as well. We found a total of eight tools with suitable outputs that could be matched to the ground truth datasets by the PAS window approach described above (Table 1).

To assess the relative quantification accuracy, we first defined a set of high-confidence and clear-cut APA sites within TEs based on the orthogonal 3'-end sequencing and simulation ground-truth datasets (see Materials and Methods). Briefly, this approach first collapses overlapping TEs to create non-overlapping, composite TEs for every gene based on the appropriate annotation (RefSeq or GENCODE) for each dataset. Next, we required the composite TE to contain at least 2 quantified PAS and the total expression of these PAS to be >= 1 TPM. We further required that the top two PAS overlapping with a composite TE represented at least 80% of the total expression of that TE and also that the second most utilized PAS had at least 5% relative Poly(A) Usage (PAU). Finally, because our previous results for PAS matching found an overlap window of 50 nt to behave well, we removed TEs where the top 2 PAS were less than 50 nt apart. Figure 4 illustrates how a single gene with multiple transcript annotations can lead to multiple TEs and PAS that are considered and then filtered out or retained for downstream benchmarking analysis. We applied this strategy to all five ground truth datasets using both RefSeq and GENCODE annotations, resulting in a range of GT-TEs counts and relative usage levels of the distal PAS (Supplemental Fig. 7).

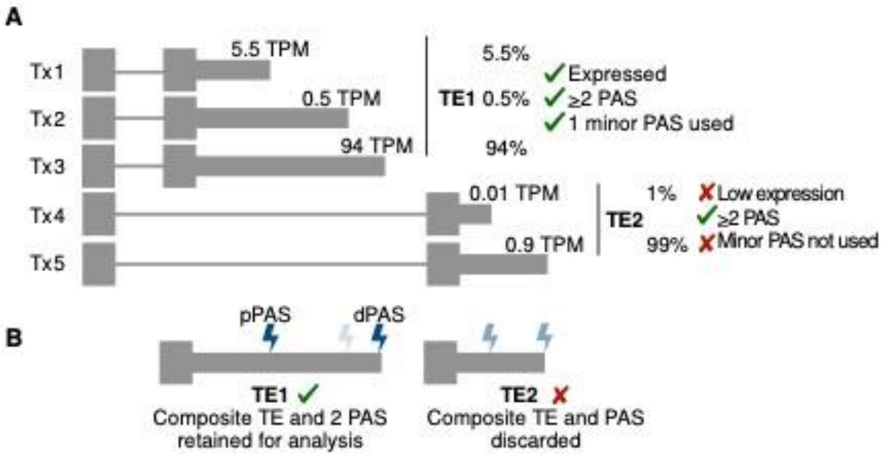


Figure 4: Ground-truth (GT) terminal exon (TE) and polyadenylation site (PAS) filtering for high-confidence, alternative polyadenylation (APA) sites.

(A) Cartoon example of heuristics applied to composite TEs based on transcript (Tx) annotations and overlapping GT-PAS based on expression (transcripts per million, TPM) and relative usage of each GT-PAS within each composite TE. Percentages represent the polyadenylation usage (PAU) for each PAS relative to other PAS in the same TE. **(B)** Final GT TE and PAS retained for downstream comparison to tool predictions.

With the above set of filtered orthogonal/ground-truth terminal exons (GT-TE) and PAS (GT-PAS), there are a number of parameters to consider for benchmarking. These include the window size for an allowable match between GT-PAS and predicted PAS (PD-PAS) as well as which transcriptome annotation to use when running each tool (e.g., RefSeq, GENCODE, custom). Additionally, upon matching GT-PAS and PD-PAS, one must consider how to handle multiple matches and which GT-PAS values are most relevant for comparisons: values from the proximal PAS (GT-pPAS), the distal PAS (GT-dPAS), or all PAS considered together (GT-allPAS)) (Supplemental Fig. 8). We describe GT-pPAS, GT-dPAS, and GT-allPAS in more detail below.

Given a specific combination of the above parameter settings, we assessed how well the RNA-seq-based estimated PAS of a given algorithm matched the high-confidence set of PAS using the following statistics: The number of ground truth APA TEs and PAS captured; the correlation coefficient between a tool's quantified values and ground truth; and by plotting the distribution of absolute differences between inferred and ground truth relative quantification values.

Given the results of the identification and absolute quantification benchmarks and the fact that a window size of 50 nt seemed to perform similarly well for most tools (Supplemental Fig. 9), we focused our analysis on results derived using the 50 nt window for matching computationally-inferred and ground truth PAS. We next asked what fraction of the filtered high-confidence ground truth PAS arising from terminal exons with APA based on the RefSeq annotation were reported on by each tool (Figure 5A). Overall, GETUTR and PAQR reported the smallest fractions of GT-PAS matched to predictions, QAPA reported the most, and the remaining algorithms reported similar, intermediate numbers (Figure 5A). This was also true when considering the fraction of APA GT-TEs with any PAS matched to an algorithm-inferred PAS (Supplemental Fig. 10A). Strikingly, with the exception of QAPA, most tools reported a relatively small fraction (around 20%) of APA GT-TEs where both the distal and proximal PAS matched a computationally-inferred PAS (Figure 5B). In the case of APAtrap, DaPars, DaPars2, IsoSCM, and TAPAS, the small number of GT-TEs with both PAS matched to predictions was likely driven by the fact that these tools report more GT-TE distal PAS matches (Supplemental Fig. 10B) compared to proximal PAS matches (Supplemental Fig. 10C).

Next, we assessed how well the relative PAS usage inferred by the algorithms correlated to that of matched PAS within GT APA TEs. Importantly, different tools report different types of relative usage. Some, like DaPars and DaPars2, report on each version of each TE in the annotation that was expressed, which can lead to the same PAS coordinate having multiple quantification values. Other methods, like QAPA, only report on the most distal, composite TE in a gene and report all annotated PAS regardless of expression level. Other tools we also considered, like LABRAT and APAlyzer, quantify PAS usage per transcript, but do not provide specific coordinates of the potentially collapsed PAS clusters or inferred PAS and therefore we did not include these in the benchmarking. These disparate approaches to produce PAS quantifications can lead to multiple valid matches between computationally-inferred and ground-truth PAS quantifications for one or more PAS within each TE. See, for example, Supplemental Fig. 8. This example also highlights how annotation can be influential in which PAS are considered and quantified by certain tools (with the exception of IsoSCM which is annotation-agnostic).

To overcome this, we performed the benchmarking based on all GT- to PD-PAS matches for a window of 50 nt (all-PD) and also on only the best possible GT- to PD-PAS match that minimizes the absolute difference between PD and GT quantification (best-PD) (Supplemental Fig. 8). For a tool that outputs multiple quantifications for a single PAS, it is difficult for a user to prioritize the TE or PAS from the outputs, therefore we focus on the results for all PAS quantified by a tool that matched GT-PAS (all-PD, Fig. 5C and Supplemental Fig. 11). The results for the best possible PAS quantification of a tool matched to GT-PAS (best-PD) are shown in Supplemental Fig. 12.

The effect of choosing to correlate all-PD matches compared to only the best-PD matches was apparent at different window sizes, where most tools had higher correlations when a window size of 100 nt was used and only the best-PD match was considered, but similar or worse correlations when all-PD matches at this window size were considered (e.g., APAtrap, DaPars/DaPars2, GETUTR, IsoSCM, and QAPA). PAQR and TAPAS performed similarly for all window sizes when considering PD-all or only PD-best matches (Supplemental Fig. 9).

Due to partial, incomplete, or redundant matches between a tool's PAS quantifications and the high confidence set of GT-PAS, choosing which set of GT-PAS value(s) to correlate with the computationally-inferred values can also have an effect. We considered all GT-PAS together (GT-allPAS, Fig. 5C and Supplemental Fig. 11A), or separated out distal GT-PAS values (GT-dPAS, Supplemental Fig. 11B) from proximal GT-PAS values (GT-pPAS, Supplemental Fig. 11C) and correlated those individually to all-PD (Fig. 5, Supplemental Fig. 11) or best-PD PAS matches (Supplemental Fig. 12). In most cases, methods performed better when estimating the usage of the distal PAS site on both real and simulated RNA-seq datasets (compare panels B (dPAS) and C (pPAS) in Supplemental Fig. 11 and 12). This may be due, in part, to the fact that the distal site is typically the "canonical" one, containing the typical polyadenylation signals, and thereby having higher usage/expression, as observed in the ground truth samples (Supplemental Fig. 7B). Similar results were seen when we evaluated the absolute value of the difference in relative usage between PD-PAS and GT-PAS for each dataset (see Fig 6, comparing GT-pPAS, GT-dPAS, and GT-allPAS). Given this, we consider as most informative the metrics obtained when matching PD sites to all GT-PAS (GT-allPAS).

We also considered how the annotation may influence method performance. We found that the more conservative annotation, RefSeq, which was often mentioned in many tools' documentations (e.g., APAtrap, DaPars, DaPars2, GETUTR, and TAPAS), led to better correlations for many of these tools, particularly on the simulated dataset (Compare Supplemental Fig. 11 and Supplemental Fig. 13). We note that IsoSCM does not use a reference annotation and QAPA recommends a custom annotation resulting from a number of filtering steps. PAQR was only tested with GENCODE, as its main principle is to quantify the usage of PAS from the PolyASite database (Herrmann et al., 2020) and does not have a "preferred" TE annotation.

Given the above results, we chose to visualize the distributions of absolute differences between all-PD matches to different GT-PAS values using each tool's preferred annotation with an overlap window of 50 nt (Fig. 6 for representative datasets, Supplemental Fig. 14 for all datasets). As with the correlation results, absolute differences between all-PD to GT-pPAS values were worse than those observed when using GT-allPAS values or GT-dPAS values. This trend was particularly pronounced for tools like DaPars and DaPars2 across all datasets. These same methods that performed much worse on pPAS relative quantification also found much fewer GT-TEs with pPAS matches compared to dPAS matches (Fig. 6, Supplemental Figs. 14 and 15, inset barcharts).

PAQR was the method that generally achieved the highest performance on the correlation metrics described above for most datasets (Fig. 5C and 6, Supplemental Fig. 11). However, PAQR also consistently reported fewer PD to GT-PAS matches than other methods (Fig. 5A and B, Fig. 6, insets). Also of note, the dataset with lowest correlation between RNA-seq-based quantifications and GT was MmusCortex. This was the only dataset that was not obtained based on oligo(dT) selection and, due to low coverage, the GT was obtained by pooling two replicates (Supplemental Fig. 1).

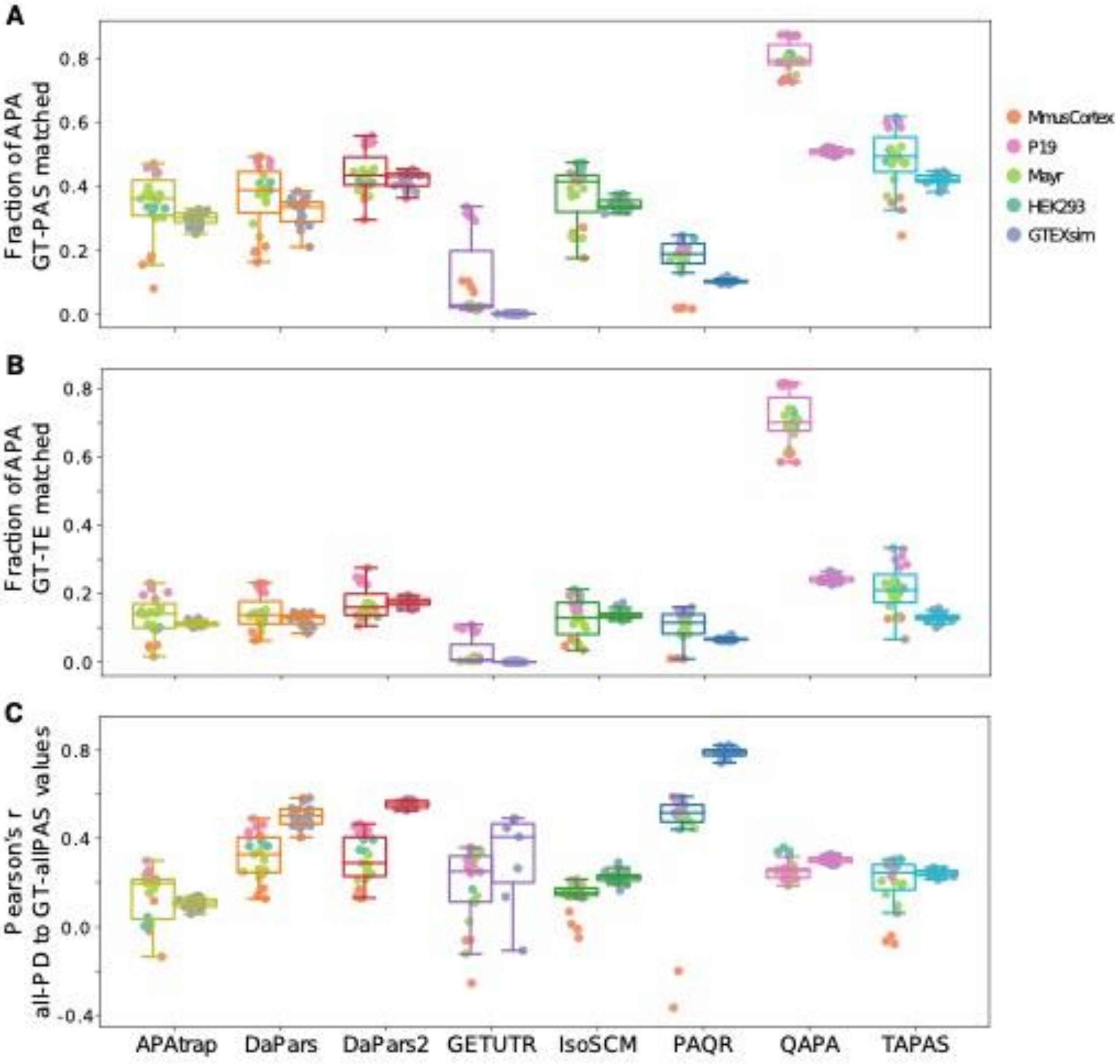


Figure 5: Overlap and correlations of algorithm predicted PAS to ground truth PAS within APA TEs.

(A) Distribution across datasets of the fraction of ground truth polyadenylation sites (GT-PAS) from terminal exons (TEs) with alternative polyadenylation (APA) that were matched to a tool predicted polyadenylation site (PD-PAS) based on RefSeq/preferred annotations within a window of 50 nucleotides. Left boxes are from real RNA-seq datasets while right boxes are for simulated datasets with points colored according to experimental groups. Each point is labeled according to dataset grouping given in the legend. (B) Fraction of ground truth (GT) TEs with APA that had PD-PAS matches to both distal and proximal GT-PAS within a window of 50 nucleotides. (C) Pearson correlation coefficient when considering all-PD predicted values that match GT-allPAS values (both distal and proximal) using each tool's preferred annotation and a match window of 50 nt. Left boxes for each method represent real RNA-seq data and right boxes are simulated RNA-seq data. Each dataset needed a minimum of 20 matched values to be plotted. Correlations between predictions either ground truth proximal PAS (GT-pPAS) or ground truth distal PAS (GT-dPAS) considered separately are plotted in Supplemental Figure 11.

C. 5. DISCUSSION:

The landscape of bioinformatics software is broad, and often there are many tools available to complete the same task. Therefore, it can be a challenge for researchers to make informed decisions on which tool to select. Recognizing the need for continuous and independent benchmarking of computational methods, the APAeval hackathon was held during the 2021 RNA Society Meeting. Bringing together a community of RNA biologists, bioinformaticians, and software developers, our goal was to benchmark various open-source computational tools for the identification and quantification of poly(A) sites from RNA-seq data, as well as set up an extendable framework for carrying out continuous benchmarking. We strove to package tools for easy installation and use, as well as developed workflows that could be applied to new datasets or used for incorporating new tools into the benchmark. We also made an effort to identify appropriate ground truth datasets and preprocess them for use not only in our study, but also in future studies of computational methods for APA analysis. By benchmarking many of the currently available tools, we provide researchers with the basis for making informed decisions on which methods best fit their application(s).

From the 17 surveyed tools, we were able to benchmark eight tools across five distinct ground-truth datasets. Our reasons for not being able to include all tools varied, from not being able to install or run them, to the tools generating outputs that could not be transformed into the basic metrics for the respective task, i.e., PAS identification or quantification (Table 1). That is, although all the aforementioned tools analyze PAS usage from RNA-seq data, they use different types of information as input, compute distinct measures, and solve somewhat different tasks: PAS identification, PAS quantification (both per PAS isoform and relative to all PAS within a gene), differential PAS usage between samples, and detection of changes in TE length. As each type of task required considerable effort to implement, we limited ourselves to the first two. We refer the reader to (Chen et al., 2020; Shah et al., 2021; W. Ye et al., 2022) for additional description and other types of benchmarking not covered here. Further increasing the complexity of the effort was the fact that a non-trivial number of parameters needed to be defined and tested for each task, which also emphasizes the importance of reproducible and parameterizable workflows that can be tested and used by any researcher.

The results show that, in general, the methods strike different balances between sensitivity and specificity. The highly non-uniform coverage of genes by RNA-seq reads makes it challenging to reliably identify the drops in coverage that reveal the PAS. Thus, methods that do not identify the PAS *ab initio*, but rather use pre-defined PAS to assess their usage (e.g., QAPA and PAQR), generally have better performance than methods that do not use such information. However, the advantage of considering only known APA sites clearly comes at a price, especially for species that are not as well annotated, or when researchers study polyadenylation in contexts when novel sites are likely to play a role. We found that TAPAS has the highest accuracy for PAS identification, while for quantification, the performance metrics vary widely between methods. The correlation of PAQR-inferred PAS usage with the ground truth is consistently higher compared to other methods such as DaPars(2), although PAQR consistently quantifies fewer sites. Thus, if researchers are interested in a higher accuracy set of PAS PAQR quantifications may be preferable, but if a broader coverage of PAS is needed for downstream analyses, this could be obtained with methods such as QAPA, TAPAS and DaPars. Also, users should take into account that in general, proximal PAS are much less accurately quantified and far fewer are properly identified compared to distal PAS, particularly for methods that infer proximal sites *de novo*, from the RNA-seq read coverage. These results are summarized in Fig. 7, which provides an overview of methods' performance and tradeoffs that the users can easily assess when deciding what approach to use in analyzing their data.

The above results are generally in line with other recent efforts to compare methods for calling PAS from RNA-Seq. Specifically, the closest benchmarking effort (Chen et al., 2020) included assessment of PAS identification and differential usage. The authors used four datasets to test PAS identification and a single dataset for differential usage. They also found TAPAS to be the top performer in site identification and observed similar low accuracy when calling PAS from RNA-seq alone. Chen et al. also note similar differences when comparing synthetic to real data, though the simulated data was limited to a thousand genes with 1-4 PAS and high coverage, resulting in higher accuracy than we observe here with larger datasets. Notably, Chen et al. included extensive testing of each method's parameter setting and the effect of the read coverage depth. However, PAQR and DaPars2 were not part of that evaluation, and the authors did not test the quantification accuracy, neither absolute nor relative.

Importantly, even though we made a strong effort to identify appropriate ground truth data, having diverse "gold standard" data remains a challenge. Any orthogonal method for measuring 3′ end usage will exhibit a different bias compared to RNA-seq, and therefore, the overlap between RNA-seq-based inferences and ground truth will be inherently limited. Furthermore, even when comparing replicates of the ground truth datasets, the Jaccard index of the identified sites was roughly between 0.5 and 1, and the Pearson correlation coefficient roughly between 0.75 and 1 (data not shown). This indicates that the methods for experimental identification of PAS can also be substantially improved, and also sets the upper bound on the performance of the computational methods.

Given these limitations of the methods for quantifying APA from RNA-seq data, the question arises as to their utility overall. Ideally, studies of APA would use direct measurements of APA isoform abundance, obtained on platforms such as PacBio or Oxford Nanopore, which are designed for sequencing full-length cDNAs. However, substantial efforts have been put in the past decades into the profiling of samples from a wide range of conditions, including tumors, by short read RNA sequencing. Such datasets are much more extensive than datasets where 3′ end sequencing has been applied. Thus, the interest in mining RNA-seq data to uncover APA remains, until perhaps 3′-biased single-cell sequencing will have generated comparable coverage of human cell types and diseases. Our study should facilitate the choice and application of available methods to these RNA-seq data sets.

C. 6. MATERIALS AND METHODS:

C. 6. 1. Data processing

Dataset preprocessing of the RNA-seq data was performed using the nf-core/rnaseq v3.8.1 RNA-seq pipeline (Patel et al., 2022) with options:

```
--profile docker --aligner star_salmon --save_reference --
gencode --save_trimmed --skip_markduplicates --skip_stringtie -
-save_unaligned --skip_bbsplit
```

For ground truths from A-seq2 and 3′-seq protocols, the raw data was downloaded from SRA (for identifiers see Table S1) and processed as in (Herrmann et al., 2020). Here, sites were discarded as likely internal priming events if the 10 nt genomic region downstream of the putative site contained 6 consecutive As, or 7 As in total. For MACE-seq, processed PAS files were obtained from the authors (Schwich et al., 2021) and used as ground truth. According to the publication, internal priming filtering was performed by mapping a sequence of 10 As to the genome, allowing two mismatches, and discarding PAS if they were located adjacent to the genomic coordinates of a mapped A stretch. For PAPERCLIP, bed files containing PAS with associated read counts pooled across the duplicate experiments were downloaded from SRA (accessions see Table S1). The counts for each site were multiplied by 10^6 and divided by the total read count in the sample to obtain the expression level as TPM. All ground truth files were converted to a BED6 format for benchmarking. If applicable, PAS clusters were collapsed to their representative site to obtain single nucleotide PAS.

Simulated RNA-seq data based on transcript level expression quantification from GTEx v8 were used from (Vaquero-Garcia et al., 2023). We selected ten cerebellum and ten skeletal muscle samples from this study at random. Ground truth poly(A) site expression levels for these samples were extracted from GTEx v8 transcript quantifications (GTEx_Analysis_2017-06-05_v8_RSEMv1.3.0_transcript_tpm.gct.gz downloaded from the GTEx portal) where the last nucleotide of each quantified transcript was defined as the poly(A) site.

The APAeval Benchmarking Workflow

For each of the APAeval benchmarking events, specific metrics were defined and then computed for each dataset pair (RNA-seq-based inferences (predictions) - ground truth data) separately.

The benchmarking infrastructure was broadly separated into two modules: workflows to execute individual methods (“method workflows”) and workflows to compute the benchmarking metrics for all evaluated methods (“benchmarking workflows”). To ensure compatibility between those two modules, the output of the method workflows was required to adhere to a previously defined standardized format, namely the well-known BED format. Thus, 1) the genomic location of poly(A) sites had to be reported as single-nucleotide; 2) the absolute quantification of PAS was done as TPM (Transcripts Per Million); or 3) the relative quantification of PAS was done as fractional usage compared to one or more additional PAS within the same TE. Detailed input and output specifications are available in the APAeval Github repository (https://github.com/iRNA-COSI/APAeval/).

A method workflow was developed for each participating method using either the Snakemake (Mölder et al., 2021) or Nextflow (Di Tommaso et al., 2017) workflow management systems. The execution of each individual step was isolated in a Docker or Singularity container. When possible, we selected publicly available containers (e.g., from the Biocontainers project ((Da Veiga Leprevost et al., 2017)). When necessary, we custom-built Docker images. Input file formats were restricted to maintain consistency across method workflows while also allowing some flexibility in individual method execution requirements. As input file formats, we selected BAM files for aligned reads, FASTA files for nucleotide sequences, GTF files for gene/transcript model annotations and BED files for reference poly(A) sites. Any steps necessary to make those inputs compatible with a method were included in the method workflow. While we generally used the same annotation files for all methods, some required a specific annotation (e.g., QAPA). In such cases we carried out the analysis both with the preferred annotation of the tool and with the standard one. All configurable parameters of a tool were specified via a configuration file and test data was provided for each workflow. For computational tasks common to several method workflows we created a Python package that could be imported in the method workflows to avoid variability in implementation and code duplication. All workflows were reviewed for clarity and accuracy by two independent members of the APAeval team.

Benchmarking workflows were created in Nextflow, based on previously published guidelines from the OEB initiative (https://github.com/inab/TCGA_benchmarking_workflow). In the workflows three distinct containerized steps were executed; validation of the file formats created by the method workflows, computation of the metrics defined for the respective benchmarking event, and consolidation of results into OEB compatible JSON files. Those files were used for uploading APAeval results to the OEB database, and the metrics were extracted to create the figures presented in the manuscript. All the code is available on GitHub (https://github.com/iRNA-COSI/APAeval/).

C. 6. 2. PAS matching strategy

Any evaluation of detection or quantification relies on first matching poly(A) sites identified from RNA-seq data by the benchmarked method (PD - prediction) to poly(A) sites detected in an orthogonal 3'end sequencing dataset (GT - ground truth). To achieve this we used BEDTools window (Quinlan & Hall, 2010). Matching was performed with different window sizes (n = 10, 25, 50 or 100 nucleotides) , i.e., the coordinates of ground truth sites were extended by n nucleotides in both directions to allow for variation in poly(A) site identification. For absolute quantification, if one PD site matched multiple GT sites its score (expression level as TPM) was split between the GT sites into shares inversely proportional to the distance of the PD site to the respective GT sites. If multiple PD sites matched one GT site they were merged and their scores were summed up. For identification, merging of PD sites was not performed. For relative quantification, merging of multiple or redundant PD site matches was not performed and either all PD site matches were considered for correlations (all-PD) or the single best PD match was considered (best-PD). See Supplemental Fig. 8 for more details and an example.

C. 6. 3. Identification metrics

In the identification metrics, we define true positives (TP) as predicted sites that fall within windows of specified size (see above) around GT sites. In contrast, false negatives (FN) are GT sites that do not have a matching prediction, and false positives (FP) are predicted sites without a GT match. Accordingly, we calculated the precision (TP / (TP + FP)), sensitivity (TP / (TP + FN)) and Jaccard index (TP / (TP + FP + FN)) for a range of window sizes.

Finally, a metric evaluating the prediction per gene was calculated, called "Percentage of genes with correct number of PAS". For this, we obtained the number of PAS per gene from the GT. Genes with no PAS in the GT were not considered further. Similarly, the number of PAS per gene from unique PD sites was calculated. Finally, the number of genes with the same number of PAS in GT and PD was calculated and divided by the number of genes with at least one PAS and multiplied by 100 to obtain the percentage of genes with the correct number of PAS.

C. 6. 4. Absolute quantification metrics

Given the matching procedure described above, the main quantification accuracy metric we used is the Pearson or Spearman rank correlation of the normalized prediction and the ground truth expression (TPM) values for *matched* sites. To assess a method's ability to focus on relevant sites we also determined the fraction of the gene expression coming from unmatched sites ("pct-FP"). For that, the percentage of TPM expression of non-matched predicted sites (FP) was calculated by summing the TPM expression of all predicted sites without a GT match and dividing by the sum of expression values of all predicted PAS (FP + TP). Although the absolute quantification event aims to assess the performance of methods in the "true" quantification of PAS expression, we provide the sensitivity and precision metrics as defined above, as well as the F1-score (TP / (TP + 0.5*(FP + FN))) to shed light on the sets of PAS each method considers.

C. 6. 5. Relative quantification metrics

For relative quantification benchmarking, we first defined sets of high confidence TEs that exhibited alternative polyadenylation (APA) to be detected by the various tools. The procedure is outlined in Figure 4. Given a transcriptome annotation, the summary workflow first defines composite TEs by collapsing TEs from different transcripts of the same gene that overlapped in their genome coordinates. Next, given the coordinates and expression values (as TPMs) of ground truth PAS (GT-PAS), we retained all GT-PAS that overlapped with a composite TE. Relative Poly(A) Usage (PAU) was calculated for each retained GT-PAS by dividing each GT-PAS TPM by the sum of TPMs of all GT-PAS associated with the same composite TE. Finally, a number of filtering steps were applied to define high confidence TEs that exhibited APA to be retained for benchmarking. These APA TEs needed to have a sum of GT-PAS TPMs greater than or equal to 1 and contain at least two GT-PAS with relative usage of at least 5%. Any GT-PAS with less than 5% PAU was filtered out. We also filtered out potential overly complex TEs and their associated GT-PAS by requiring the top two GT-PAS within a composite TE to represent at least 80% of the relative PAS usage. Finally, we filtered out the remaining composite TEs and their associated GT-PAS if the retained PAS were closer than 50 nt together. The GT-PAS closest to the start of the TE was defined as the proximal PAS (pPAS) and all other PAS were defined as distal PAS (dPAS).

We applied the PAS window matching strategy described above, to match the filtered ground truth (GT-PAS) and the tool predictions (PD-PAS) and then calculated metrics over these matched sites. We calculated Pearson correlation between matched GT and PD relative usage values and plotted the empirical cumulative distribution functions (eCDFs) of absolute differences between GT-PAS and PD-PAS PAUs. Importantly, for these metrics we also considered different subsets of matches between GT and PD to identify potential shortcomings in the tool's ability to quantify either the pPAS or the dPAS accurately. Therefore, we calculated these metrics for all GT to PD matches (GT-allPAS), just the matches to GT-pPAS, and just the matches to GT-dPAS. Some methods provide duplicate quantifications for the same PAS coordinate because they consider each transcript-defined TE separately, leading to some multi-matched sites (see, for example, Supplemental Figure 8). Because a user does not know *a priori* which TE version or PAS quantification is correct we calculated metrics based on all matched PD-PAS usage values together, including duplicates (all-PD matches), or by only considering the best PD-PAS usage value that minimized the difference to GT-PAS usage (best-PD matches). Finally, to shed light on the number of TEs with APA and the types of PAS (distal or proximal) each tool detects and quantifies in this relevant subset, we calculated the fraction of GT-TEs with APA that had matches to any PD-PAS as well as the fraction that had a PD match to the GT-pPAS site, the GT-dPAS site, or to both.

C. 7. Data availability

All code and metrics for the project are publicly available at https://github.com/iRNA-COSI/APAeval. Unless otherwise stated, publicly available containers (e.g., from the Biocontainers project (Da Veiga Leprevost et al., 2017)) were utilized for execution workflows. In a few cases, we generated custom-built docker images. These are hosted at https://hub.docker.com/u/apaeval. The most informative metrics for all challenges of the identification and absolute quantification events are stored in the OpenEBench database (https://openebench.bsc.es/benchmarking/OEBC007/events), where the results can be visualized and further explored. All datasets used in this study are publicly available. For SRA accessions and download links see Table S1. Annotation files, input ground truth and RNA-seq files, code for figure plotting, and other data to reproduce or supplement this analysis have been deposited in a Zenodo repository available at https://doi.org/10.5281/zenodo.8290348

C. 8. ACKNOWLEDGMENTS

We thank the RNA Society (https://www.rnasociety.org/) for making the hackathon possible and bringing together people from various disciplines. We thank the OpenEBench platform (https://openebench.bsc.es/) for hosting the APAeval benchmark. We thank the OEB members Anna Redondo, Laura Portell, Laura Rodriguez-Navas and Salvador Capella for help and contributions. We thank Seqera Labs (https://seqera.io/) for kindly hosting a workshop regarding the use of Nextflow and Nextflow Tower during the RNA Society 2021 meeting, and Amazon Web Services (https://aws.amazon.com/) for providing us with credits for the use of their cloud platform. We are grateful to host the source code on the iRNA COSI (https://irnacosi.org/) GitHub account.

C. 9. Author contributions

Author contributions are listed following the CREDIT taxonomy system and are as follows:
Conceptualization: YB, AK, DB, MRG, CJH, and MZ
Data Curation: YB, SBS, DB, WD, M Fansler, MRG, CJH, AK, FM, BN, LS, YKW, MZ, and FZ
Formal Analysis: YB, SBS, DB, WD, MMF, MF, MRG, PLG, AGU, CJH, ZK, FM, CLP, GR, LS, YKW, MZ, and FZ
Funding acquisition: YB and MZ

Investigation: YB, SBS, DB, WD, MMF, MF, CMF, MRG, AGU, SH, CH, CJH, AK, MK, FM, EM, BN, CLP, GR, LS, DS, YKW, PJW, MZ, and FZ
Methodology: YB, SBS, DB, WD, MMF, MRG, AGU, CJH, AK, GR, MZ, and FZ
Project Administration: YB, SBS, DB, WD, MMF, CMF, MRG, AGU, CH, CJH, AK, BN, YKW, MZ, and FZ
Software: SBS, DB, WD, MMF, MF, MRG, AGU, SH, CJH, EM, BN, CLP, GR, LS, DS, YKW, PJW, and FZ
Supervision: YB and MZ
Validation: SBS, DB, WD, MF, MRG, CJH, AK, GR, YKW, and FZ
Visualization: SBS, DB, CMF, MRG, CH, CJH, AK and YKW
Writing–Original Draft: YB, SBS, DB, WD, CMF, MRG, CH, CJH, YY, MZ, and FZ
Writing–Review and Editing: YB, SBS, DB, CMF, MRG, CH, CJH, AK, FM, YKW, MZ, and FZ

CREDIT Author Taxonomy. See also the Contributors section of https://github.com/iRNA-COSI/APAeval.

C. 10. Funding information

SBS was supported by a UK Motor Neurone Disease Association and Masonic Charitable Foundation PhD Studentship (893-792). CMF was partially supported by a postdoctoral fellowship from the American Cancer Society (PF-19-157-01-RMC). MRG was supported by the Blavatnik Family Fellowship in Biomedical Research and by NHLBI of the NIH under award number F31 HL162546. AGU was supported by the The Emerging Human Brain Cluster (Clúster Emergent del Cervell Humà – CECH) project, number 001-P-001682, cofunded in a 50% by the European Regional Development Fund of the European Union (Programa Operatiu FEDER Catalunya 2014-2020) with the support of the Catalan Government. YKW was supported by the Singapore International Graduate Award. OpenEBench is partly funded by the Horizon 2020 ELIXIR-CONVERGE programme, grant agreement id 871075. YB work was supported by R01 GM128096, R01 GM147739 , R01 LM013437. DB and CJH were partially supported by the Swiss National Science Foundation grant #310030_189063 to MZ.

C. 11. REFERENCES

Arefeen, A., Liu, J., Xiao, X., & Jiang, T. (2018). TAPAS: Tool for alternative polyadenylation site analysis. *Bioinformatics*, *34*(15), 2521–2529. https://doi.org/10.1093/bioinformatics/bty110

Capella-Gutierrez, S., Iglesia, D. de la, Haas, J., Lourenco, A., Fernández, J. M., Repchevsky, D., Dessimoz, C., Schwede, T., Notredame, C., Gelpi, J. L., & Valencia, A. (2017). *Lessons Learned: Recommendations for Establishing Critical Periodic Scientific Benchmarking* (p. 181677). bioRxiv. https://doi.org/10.1101/181677

Cass, A. A., & Xiao, X. (2019). MountainClimber Identifies Alternative Transcription Start and Polyadenylation Sites in RNA-Seq. *Cell Systems*, *9*(4), 393-400.e6. https://doi.org/10.1016/j.cels.2019.07.011

Chen, M., Ji, G., Fu, H., Lin, Q., Ye, C., Ye, W., Su, Y., & Wu, X. (2020). A survey on identification and quantification of alternative polyadenylation sites from RNA-seq data. *Briefings in Bioinformatics*, *21*(4), 1261–1276. https://doi.org/10.1093/bib/bbz068

Da Veiga Leprevost, F., Grüning, B. A., Alves Aflitos, S., Röst, H. L., Uszkoreit, J., Barsnes, H., Vaudel, M., Moreno, P., Gatto, L., Weber, J., Bai, M., Jimenez, R. C., Sachsenberg, T., Pfeuffer, J., Vera Alvarez, R., Griss, J., Nesvizhskii, A. I., & Perez-Riverol, Y. (2017). BioContainers: An open-source and community-driven framework for software standardization. *Bioinformatics*, *33*(16), 2580–2582. https://doi.org/10.1093/bioinformatics/btx192

Derti, A., Garrett-Engele, P., Macisaac, K. D., Stevens, R. C., Sriram, S., Chen, R., Rohl, C. A., Johnson, J. M., & Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. *Genome Research*, *22*(6), 1173–1183. https://doi.org/10.1101/gr.132563.111

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Elkon, R., Ugalde, A. P., & Agami, R. (2013). Alternative cleavage and polyadenylation: Extent, regulation and function. *Nature Reviews. Genetics*, *14*(7), 496–506. https://doi.org/10.1038/nrg3482

Fahmi, N. A., Ahmed, K. T., Chang, J.-W., Nassereddeen, H., Fan, D., Yong, J., & Zhang, W. (2022). APA-Scan: Detection and visualization of 3′-UTR alternative polyadenylation with RNA-seq and 3′-end-seq data. *BMC Bioinformatics*, *23*(3), 396. https://doi.org/10.1186/s12859-022-04939-w

Feng, X., Li, L., Wagner, E. J., & Li, W. (2018). TC3A: The Cancer 3′ UTR Atlas. *Nucleic Acids Research*, *46*(D1), D1027–D1030. https://doi.org/10.1093/nar/gkx892

Flavell, S. W., Kim, T.-K., Gray, J. M., Harmin, D. A., Hemberg, M., Hong, E. J., Markenscoff-Papadimitriou, E., Bear, D. M., & Greenberg, M. E. (2008). Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. *Neuron*, *60*(6), 1022–1038. https://doi.org/10.1016/j.neuron.2008.11.029

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., … Flicek, P. (2021). GENCODE 2021. *Nucleic Acids Research*, *49*(D1), D916–D923. https://doi.org/10.1093/nar/gkaa1087

Gerber, S., Schratt, G., & Germain, P.-L. (2021). Streamlining differential exon and 3′ UTR usage with diffUTR. *BMC Bioinformatics*, *22*(1), 189. https://doi.org/10.1186/s12859-021-04114-7

Goering, R., Engel, K. L., Gillen, A. E., Fong, N., Bentley, D. L., & Taliaferro, J. M. (2021). LABRAT reveals association of alternative polyadenylation with transcript localization, RNA binding protein expression, transcription speed, and cancer survival. *BMC Genomics*, *22*(1), 476. https://doi.org/10.1186/s12864-021-07781-1

Grassi, E., Mariella, E., Lembo, A., Molineris, I., & Provero, P. (2016). Roar: Detecting alternative polyadenylation with standard mRNA sequencing libraries. *BMC Bioinformatics*, *17*(1), 423. https://doi.org/10.1186/s12859-016-1254-8

Gruber, A. J., Schmidt, R., Ghosh, S., Martin, G., Gruber, A. R., van Nimwegen, E., & Zavolan, M. (2018). Discovery of physiological and cancer-related regulators of 3′ UTR processing with KAPAC. *Genome Biology*, *19*(1), 44. https://doi.org/10.1186/s13059-018-1415-3

Gruber, A. J., & Zavolan, M. (2019). Alternative cleavage and polyadenylation in health and disease. *Nature Reviews. Genetics*, *20*(10), 599–614. https://doi.org/10.1038/s41576-019-0145-z

Ha, K. C. H., Blencowe, B. J., & Morris, Q. (2018). QAPA: A new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biology*, *19*(1), 45. https://doi.org/10.1186/s13059-018-1414-4

Harrison, B. J., Park, J. W., Gomes, C., Petruska, J. C., Sapio, M. R., Iadarola, M. J., Chariker, J. H., & Rouchka, E. C. (2019). Detection of Differentially Expressed Cleavage Site Intervals Within 3′ Untranslated Regions Using CSI-UTR Reveals Regulated Interaction Motifs. *Frontiers in Genetics*, *10*, 182. https://doi.org/10.3389/fgene.2019.00182

Herrmann, C. J., Schmidt, R., Kanitz, A., Artimo, P., Gruber, A. J., & Zavolan, M. (2020). PolyASite 2.0: A consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Research*, *48*(D1), D174–D179. https://doi.org/10.1093/nar/gkz918

Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J. Y., Yehia, G., & Tian, B. (2013). Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nature Methods*, *10*(2), 133–139. https://doi.org/10.1038/nmeth.2288

Hwang, H.-W., Park, C. Y., Goodarzi, H., Fak, J. J., Mele, A., Moore, M. J., Saito, Y., & Darnell, R. B. (2016). PAPERCLIP Identifies MicroRNA Targets and a Role of CstF64/64tau in Promoting Non-canonical poly(A) Site Usage. *Cell Reports*, *15*(2), 423–435. https://doi.org/10.1016/j.celrep.2016.03.023

Jan, C. H., Friedman, R. C., Ruby, J. G., & Bartel, D. P. (2011). Formation, regulation and evolution of Caenorhabditis elegans 3′UTRs. *Nature*, *469*(7328), 97–101. https://doi.org/10.1038/nature09616

Ji, Z., Lee, J. Y., Pan, Z., Jiang, B., & Tian, B. (2009a). Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences*, *106*(17), 7028–7033. https://doi.org/10.1073/pnas.0900028106

Ji, Z., Lee, J. Y., Pan, Z., Jiang, B., & Tian, B. (2009b). Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences*, *106*(17), 7028–7033. https://doi.org/10.1073/pnas.0900028106

Katz, Y., Wang, E. T., Airoldi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, *7*(12), 1009–1015. https://doi.org/10.1038/nmeth.1528

Kim, M., You, B.-H., & Nam, J.-W. (2015). Global estimation of the 3′ untranslated region landscape using RNA sequencing. *Methods*, *83*, 111–117. https://doi.org/10.1016/j.ymeth.2015.04.011

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12(323), https://doi.org/10.1186/1471-2105-12-323

Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S., & Mayr, C. (2013). Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & Development*, *27*(21), 2380–2396. https://doi.org/10.1101/gad.229328.113

Lusk, R., Stene, E., Banaei-Kashani, F., Tabakoff, B., Kechris, K., & Saba, L. M. (2021). Aptardi predicts polyadenylation sites in sample-specific transcriptomes using high-throughput RNA sequencing and DNA sequence. *Nature Communications*, 12(1), 1652. https://doi.org/10.1038/s41467-021-21894-x

Martin, G., Gruber, A. R., Keller, W., & Zavolan, M. (2012). Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′ UTR length. *Cell Reports*, *1*(6), 753–763. https://doi.org/10.1016/j.celrep.2012.05.003

Martin, G., Schmidt, R., Gruber, A. J., Ghosh, S., Keller, W., & Zavolan, M. (2017). 3′ End Sequencing Library Preparation with A-seq2. *JoVE (Journal of Visualized Experiments)*, *128*, e56129. https://doi.org/10.3791/56129

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. *F1000Research*, *10*, 33. https://doi.org/10.12688/f1000research.29032.2

Morris, A. R., Bos, A., Diosdado, B., Rooijers, K., Elkon, R., Bolijn, A. S., Carvalho, B., Meijer, G. A., & Agami, R. (2012). Alternative Cleavage and Polyadenylation during Colorectal Cancer Development. *Clinical Cancer Research*, *18*(19), 5256–5266. https://doi.org/10.1158/1078-0432.CCR-12-0543

Ogorodnikov, A., & Danckwardt, S. (2021). TRENDseq—A highly multiplexed high throughput RNA 3′ end sequencing for mapping alternative polyadenylation. In *Methods in Enzymology* (Vol. 655, pp. 37–72). Elsevier. https://doi.org/10.1016/bs.mie.2021.03.022

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Patel, H., Ewels, P., Peltzer, A., Hammarén, R., Botvinnik, O., Sturm, G., Moreno, D., Vemuri, P., silviamorins, Pantano, L., Binzer-Panchal, M., BABS-STP1, bot, nf-core, FriederikeHanssen, Garcia, M. U., Yates, J. A. F., Cheshire, C., rfenouil, Espinosa-Carrasco, J., … Hall, G. (2022). *nf-core/rnaseq: Nf-core/rnaseq v3.8.1 - Plastered Magnesium Mongoose*. Zenodo. https://doi.org/10.5281/zenodo.6587789

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Sanfilippo, P., Miura, P., & Lai, E. C. (2017). Genome-wide profiling of the 3′ ends of polyadenylated RNAs. *Methods (San Diego, Calif.)*, *126*, 86–94. https://doi.org/10.1016/j.ymeth.2017.06.003

Schwich, O. D., Blümel, N., Keller, M., Wegener, M., Setty, S. T., Brunstein, M. E., Poser, I., Mozos, I. R. D. L., Suess, B., Münch, C., McNicoll, F., Zarnack, K., & Müller-McNicoll, M. (2021). SRSF3 and SRSF7 modulate 3′UTR length through suppression or activation of proximal polyadenylation sites and regulation of CFIm levels. *Genome Biology*, *22*(1), 82. https://doi.org/10.1186/s13059-021-02298-y

Shah, A., Mittleman, B. E., Gilad, Y., & Li, Y. I. (2021). Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation. *Genome Biology*, *22*(1), 291. https://doi.org/10.1186/s13059-021-02502-z

Shenker, S., Miura, P., Sanfilippo, P., & Lai, E. C. (2015). IsoSCM: Improved and alternative 3′ UTR annotation using multiple change-point inference. *RNA*, *21*(1), 14–27. https://doi.org/10.1261/rna.046037.114

Shepard, P. J., Choi, E.-A., Lu, J., Flanagan, L. A., Hertel, K. J., & Shi, Y. (2011). Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA (New York, N.Y.)*, *17*(4), 761–772. https://doi.org/10.1261/rna.2581711

Sommerkamp, P., Cabezas-Wallscheid, N., & Trumpp, A. (2021). Alternative Polyadenylation in Stem Cell Self-Renewal and Differentiation. *Trends in Molecular Medicine*, *27*(7), 660–672. https://doi.org/10.1016/j.molmed.2021.04.006

The GTex Consortium. (2020). *The GTEx Consortium atlas of genetic regulatory effects across human tissues*.

Tian, B., & Manley, J. L. (2017). Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology*, *18*(1), 18–30. https://doi.org/10.1038/nrm.2016.116

Vaquero-Garcia, J., Aicher, J. K., Jewell, S., Gazzara, M. R., Radens, C. M., Jha, A., Norton, S. S., Lahens, N. F., Grant, G. R., & Barash, Y. (2023). RNA splicing analysis using heterogeneous and large RNA-seq datasets. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-36585-y

Wang, R., & Tian, B. (2020). APAlyzer: A bioinformatics package for analysis of alternative polyadenylation isoforms. *Bioinformatics,* *36*(12), 3907–3909. https://doi.org/10.1093/bioinformatics/btaa266

Xia, Z., Donehower, L. A., Cooper, T. A., Neilson, J. R., Wheeler, D. A., Wagner, E. J., & Li, W. (2014). Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nature Communications*, *5*(1), 5274. https://doi.org/10.1038/ncomms6274

Ye, C., Long, Y., Ji, G., Li, Q. Q., & Wu, X. (2018). APAtrap: Identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics*, *34*(11), 1841–1849. https://doi.org/10.1093/bioinformatics/bty029

Ye, W., Lian, Q., Ye, C., & Wu, X. (2022). A Survey on Methods for Predicting Polyadenylation Sites from DNA Sequences, Bulk RNA-seq, and Single-cell RNA-seq. *Genomics, Proteomics & Bioinformatics*. https://doi.org/10.1016/j.gpb.2022.09.005

Yoon, Y., Klomp, J., Martin-Martin, I., Criscione, F., Calvo, E., Ribeiro, J., & Schmidt-Ott, U. (2019). Embryo polarity in moth flies and mosquitoes relies on distinct old genes with localized transcript isoforms. *ELife*, *8*, e46711. https://doi.org/10.7554/eLife.46711

Yoon, Y., Soles, L. V., & Shi, Y. (2021). PAS-seq 2: A fast and sensitive method for global profiling of polyadenylated RNAs. *Methods in Enzymology*, *655*, 25–35. https://doi.org/10.1016/bs.mie.2021.03.013

Zawada, A. M., Rogacev, K. S., Müller, S., Rotter, B., Winter, P., Fliser, D., & Heine, G. H. (2014). Massive analysis of cDNA Ends (MACE) and miRNA expression profiling identifies proatherogenic pathways in chronic kidney disease. *Epigenetics*, *9*(1), 161–172. https://doi.org/10.4161/epi.26931

Zheng, D., Liu, X., & Tian, B. (2016). 3’READS+, a sensitive and accurate method for 3’ end sequencing of polyadenylated RNA. *RNA (New York, N.Y.)*, *22*(10), 1631–1639. https://doi.org/10.1261/rna.057075.116

Zhou, X., Li, R., Michal, J. J., Wu, X.-L., Liu, Z., Zhao, H., Xia, Y., Du, W., Wildung, M. R., Pouchnik, D. J., Harland, R. M., & Jiang, Z. (2016). Accurate Profiling of Gene Expression and Alternative Polyadenylation with Whole Transcriptome Termini Site Sequencing (WTTS-Seq). *Genetics*, *203*(2), 683–697. https://doi.org/10.1534/genetics.116.188508

C. 12. Glossary

TERM	DEFINITION
APA	Alternative polyadenylation
APAeval	Community effort for benchmarking bioinformatics methods that identify and quantify APA from RNA-seq data
Benchmarking	Comparison of performance of methods designed for a specific task
Benchmarking Workflow [1]	Computational workflow created to calculate various performance metrics
Challenge [1]	Specific task solved by the benchmarked methods
Event [1]	Set of challenges and metrics for evaluating method functionality; in APAeval, one of “Identification”, “Absolute quantification” or “Relative quantification”
Ground Truth	Information known to be true, used to evaluate the accuracy of predictions made by computational methods; in APAeval - 3’end seq data from the same sample/cell type as the RNA-seq data consumed by the methods in a particular challenge; aka “orthogonal data”
Method/Tool	Published bioinformatics software for analyzing APA from RNA-seq data;
Method Workflow	Computational workflow created by APAeval to reproducibly apply a method on all challenges
Metrics	Distinct performance indicators; in APAeval these are computed based on the predictions of each method and the ground truth data
OpenEBench [2]	The ELIXIR (European Life Science Infrastructure) [3] platform for community benchmarking and software monitoring
PAS	Poly(A) site, where the 3’ end cleavage of mRNAs occurs
UTR	Untranslated regions of protein-coding mRNAs

[1] Adapted from https://openebench.readthedocs.io/en/latest/glossary/glossary.html

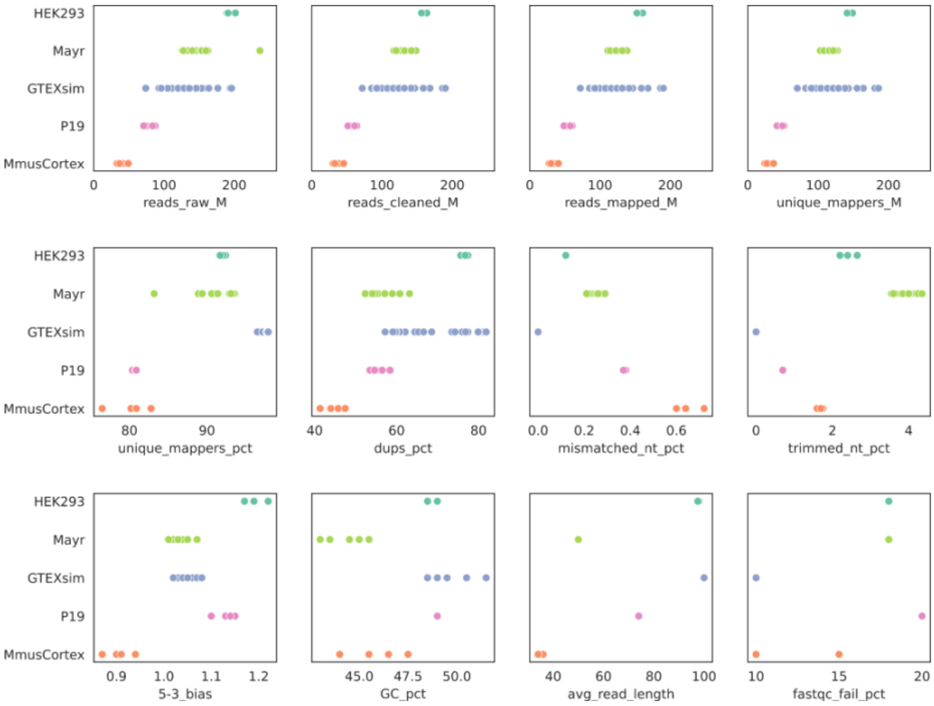
[2] https://openebench.bsc.es

[3] https://elixir-europe.org

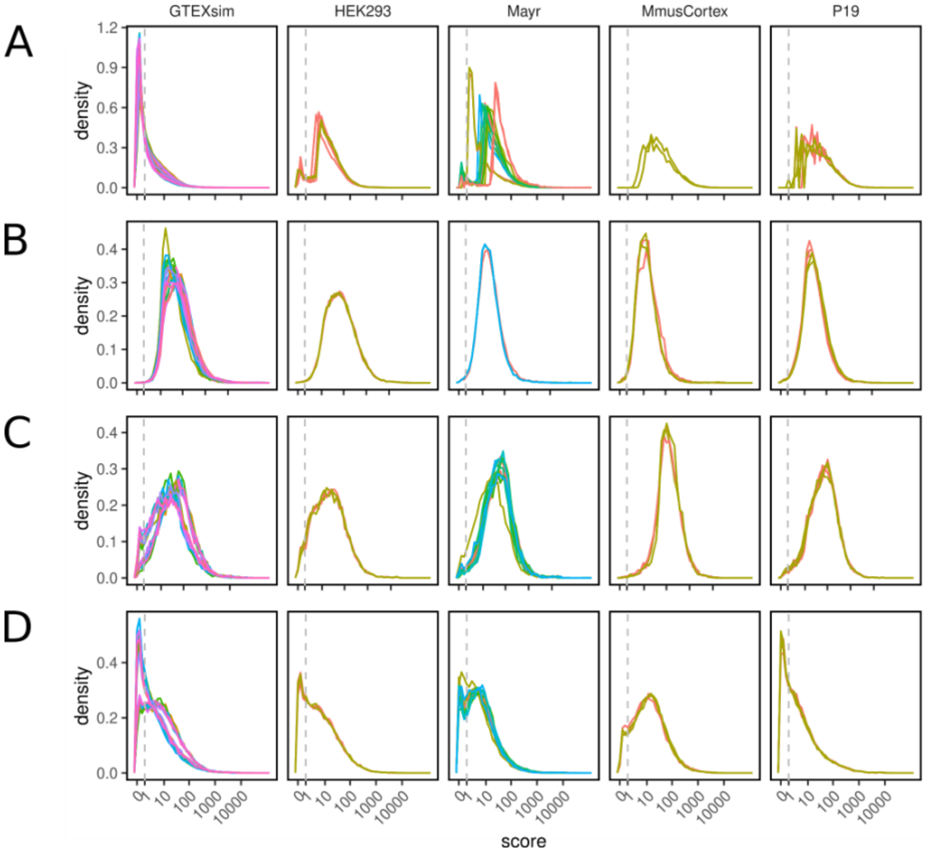
C. 13. Supplementary Material

Supplemental Table 1: RNA-seq datasets with corresponding orthogonal datasets used for benchmarking

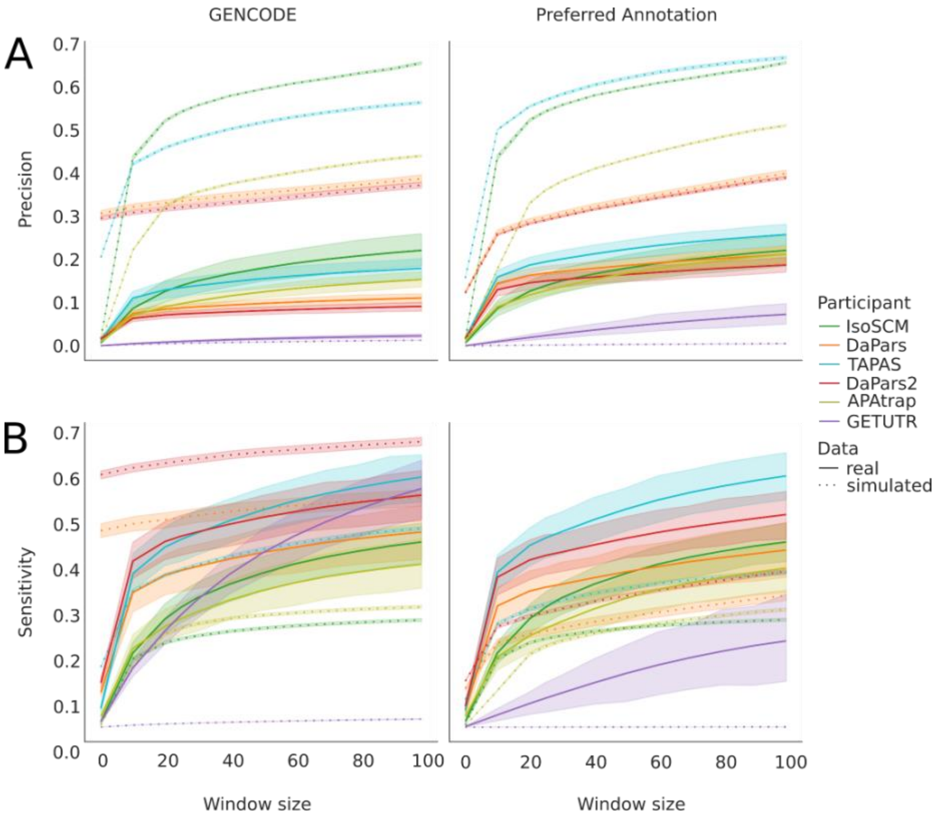
RNA- seq data						**matching orthogonal data**	
SRA accession	**sample_name**	**strandedness**	**layout**	**organism**	**length**	**SRA accession**	**sequencing method**
SRR1573494	HEK293_siControl_R1	reverse	paired	Hsap	100	SRR2922409	A-seq2
SRR1573495	HEK293_siControl_R2	reverse	paired	Hsap	100	SRR2922448	A-seq2
SRR1573496	HEK293_siHNRNPC_R1	reverse	paired	Hsap	100	SRR2922419	A-seq2
SRR1573497	HEK293_siHNRNPC_R2	reverse	paired	Hsap	100	SRR2922449	A-seq2
SRR6795718	Mayr_CD5B_R3	reverse	paired	Hsap	51	SRR6795684	3'-seq
SRR6795719	Mayr_CD5B_R4	reverse	paired	Hsap	51	SRR6795685	3'-seq
SRR6795720	Mayr_NB_R1	reverse	paired	Hsap	51	SRR1005606	3'-seq
SRR6795721	Mayr_NB_R2	reverse	paired	Hsap	51	SRR1005607	3'-seq
SRR6795723	Mayr_NB_R3	reverse	paired	Hsap	51	SRR6795688	3'-seq
SRR6795724	Mayr_NB_R4	reverse	paired	Hsap	51	SRR6795689	3'-seq
SRR6795726	Mayr_M_R2	reverse	paired	Hsap	51	SRR6795691	3'-seq
SRR6795713	Mayr_GC_R2	reverse	paired	Hsap	51	SRR6795693	3'-seq
SRR6795715	Mayr_GC_R1	reverse	paired	Hsap	51	SRR6795692	3'-seq
SRR11918577	P19_siControl_R1	unstranded	single	Mmus	75	SRR11918617	MACEseq
SRR11918578	P19_siControl_R2	unstranded	single	Mmus	75	SRR11918618	MACEseq
SRR11918579	P19_siSrsf3_R1	unstranded	single	Mmus	75	SRR11918619	MACEseq
SRR11918580	P19_siSrsf3_R2	unstranded	single	Mmus	75	SRR11918620	MACEseq
SRR11918581	P19_siSrsf7_R1	unstranded	single	Mmus	75	SRR11918621	MACEseq
SRR11918582	P19_siSrsf7_R2	unstranded	single	Mmus	75	SRR11918622	MACEseq
SRR1811005	MmusCortex_adult_R1	forward	paired	Mmus	37	GSM1614167	PAPERCLIP
SRR3067958	MmusCortex_adult_R2	forward	paired	Mmus	35	GSM1614167	PAPERCLIP
SRR3067957	MmusCortex_embryonic_R1	forward	paired	Mmus	37	GSM1614169	PAPERCLIP
SRR3067959	MmusCortex_embryonic_R2	forward	paired	Mmus	35	GSM1614169	PAPERCLIP
SRR22955576	GTEXsim_cerebellum_R1	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955574	GTEXsim_cerebellum_R2	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955639	GTEXsim_cerebellum_R3	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955510	GTEXsim_cerebellum_R4	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955630	GTEXsim_cerebellum_R5	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955420	GTEXsim_cerebellum_R6	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955571	GTEXsim_cerebellum_R7	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955570	GTEXsim_cerebellum_R8	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955441	GTEXsim_cerebellum_R9	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955647	GTEXsim_cerebellum_R10	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955539	GTEXsim_muscle_R1	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955532	GTEXsim_muscle_R2	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955403	GTEXsim_muscle_R3	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955603	GTEXsim_muscle_R4	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955459	GTEXsim_muscle_R5	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955398	GTEXsim_muscle_R6	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955458	GTEXsim_muscle_R7	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955611	GTEXsim_muscle_R8	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955449	GTEXsim_muscle_R9	forward	paired	Hsap	100	simulatedPAS	simulated
SRR22955444	GTEXsim_muscle_R10	forward	paired	Hsap	100	simulatedPAS	simulated



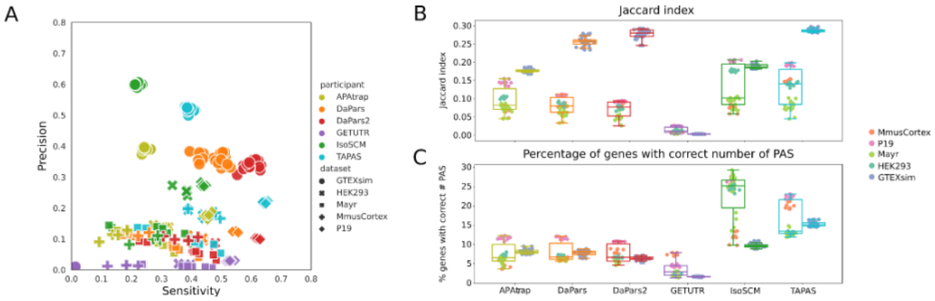
Supplemental Figure 1: Dataset characteristics. Quality characteristics of the RNA-seq datasets used for benchmarking. reads_raw_M: number of raw reads (Mio); reads_cleaned_M: number of reads after adapter trimming and minimum length selection (Mio); reads_mapped_M: number of reads successfully mapped to the genome (Mio); unique_mappers_M: number of reads mapped to a unique position in the genome (Mio); unique_mappers_pct: fraction of total reads mapped to a unique position in the genome; dups_pct: percentage of duplicate reads; mismatched_nt_pct: average percentage of mismatched nucleotides within a read; trimmed_nt_pct: average percentage of nucleotides trimmed from a read; 5-3_bias: ratio between 5' and 3' bias, where those biases are the ratio between mean coverage at the 5' region and 3' region, respectively, and the whole transcript; GC_pct: average percentage of GC nucleotides per read; avg_read_length: average read length in nucleotides; fast_qc_fail_pct: Percentage of tests failed in FastQC report of nf-core/rnaseq .



Supplemental Figure 2. TPM distributions for (A) ground truth datasets (GT), (B) APAtrap, (C) PAQR and (D) QAPA predictions (from top to bottom). The TPM scores are displayed in log-space. Sample replicates colored for better differentiation. Only all annotations (for GT) and preferred annotation (for predictions) shown.

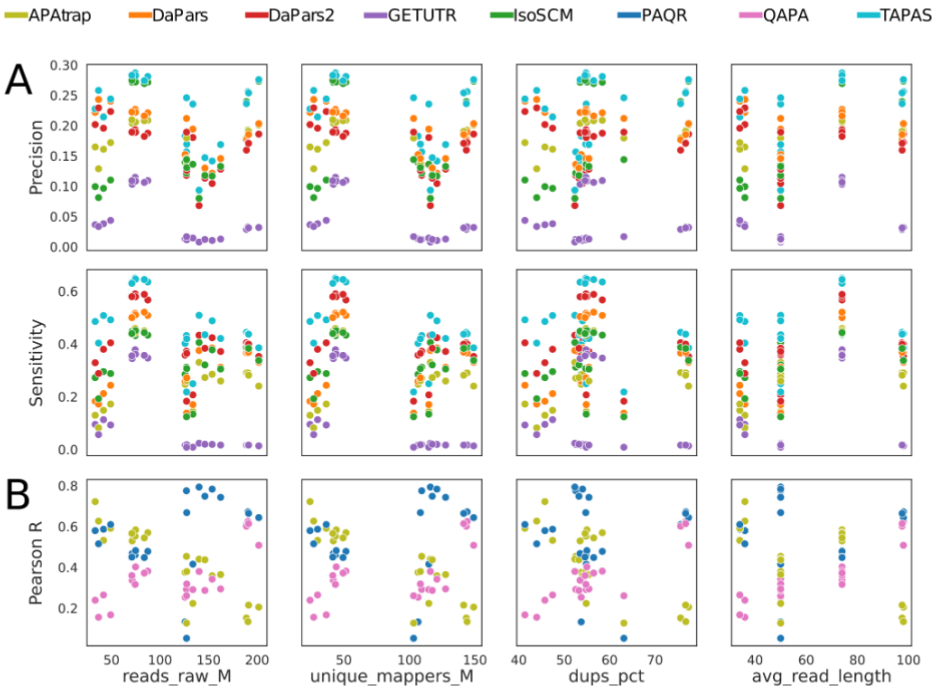


Supplemental Figure 3: Dependence of method performance in identification event on window size. Windows between 0 and 100 nt in steps of 10nt have been tested. Performance on all samples from all datasets has been combined. Shaded areas around the lines depict the 95% confidence interval. A) Precision; left column: GENCODE annotation, right column: preferred annotation. B) Sensitivity; columns as above.

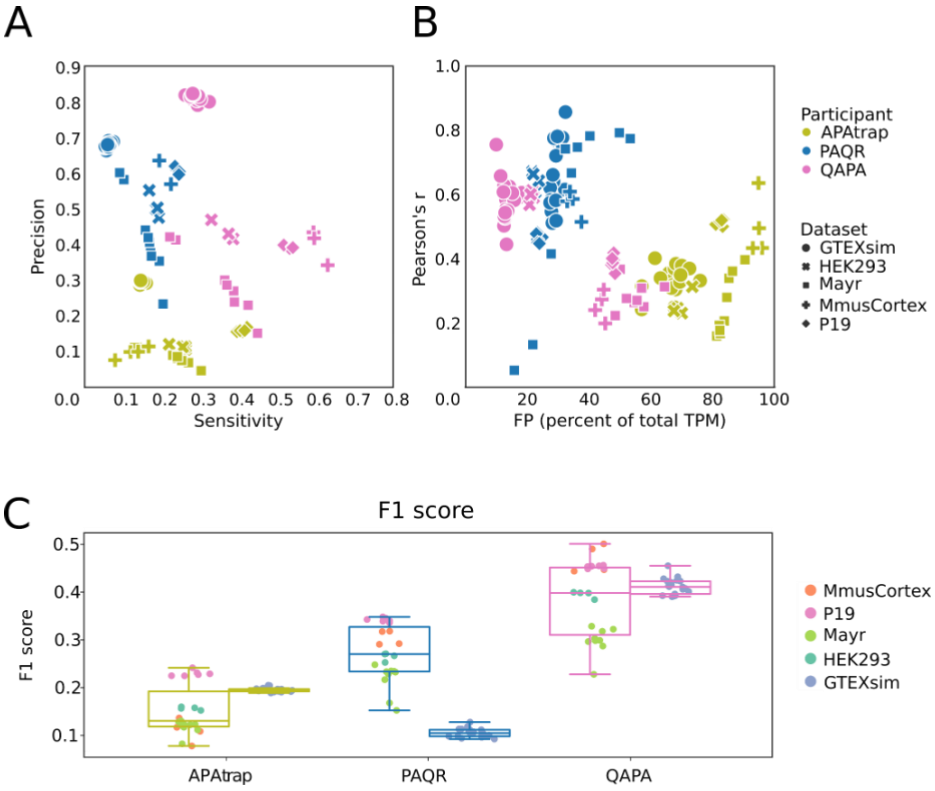


Supplemental Figure 4: Results of the PAS identification event (like Figure 2 but GENCODE instead of preferred annotation)

Box plots of Jaccard indices indicating the overlap of predicted and ground truth sites, with predicted sites being extended symmetrically by 50 nucleotides. The tools used to predict the sites are shown on the x-axis, each with two associated box plots, one for the real data (left) and another for simulated data (right). Each point is labeled according to the code given in the legend.

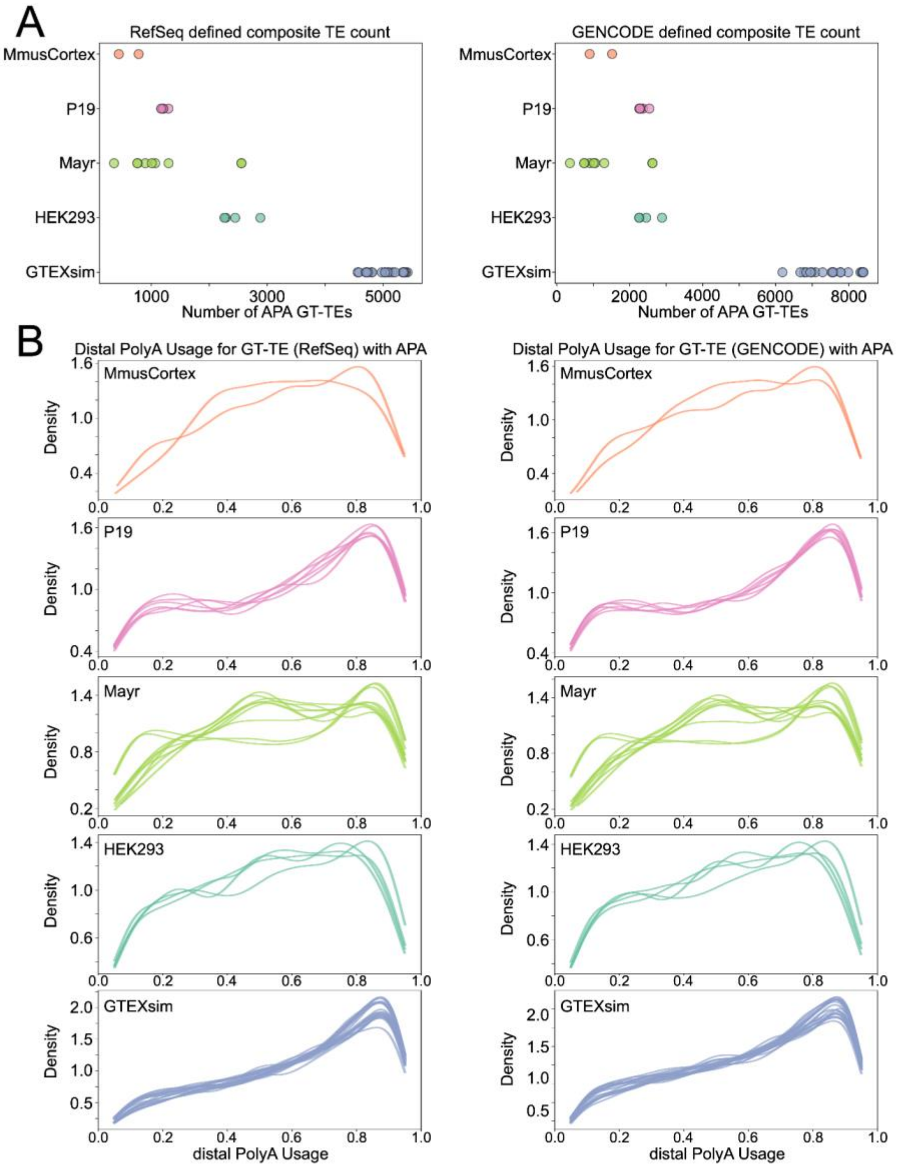


Supplemental Figure 5: Methods' performance in relation to selected dataset characteristics of real data. A) Precision and Sensitivity from the identification event. B) Pearson R from the absolute Quantification event. Each dot represents one sample.

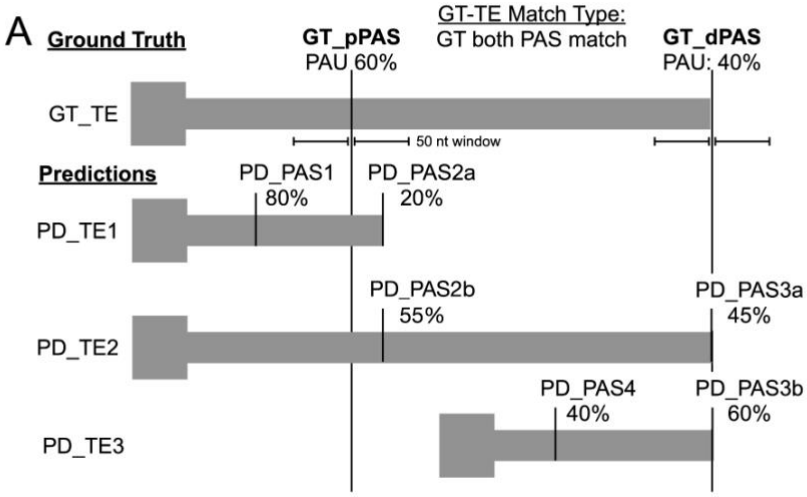


Supplemental Figure 6: Results of PAS isoform quantification (like Figure 3 but GENCODE instead of preferred annotation).

A) Scatter plot of Precision vs. Sensitivity. Each sample and tool combination is represented as a symbol, with shape and color defined in the legend. B) Pearson correlation of PD and GT site expression. The correlation coefficient for each sample is plotted against the percentage of total TPM that an algorithm attributes to PAS that are not expressed in the ground truth (false positives). C) Box plots of F1 scores. Box plots are drawn separately for real (left, see color scheme in the legend) and simulation (right) samples.



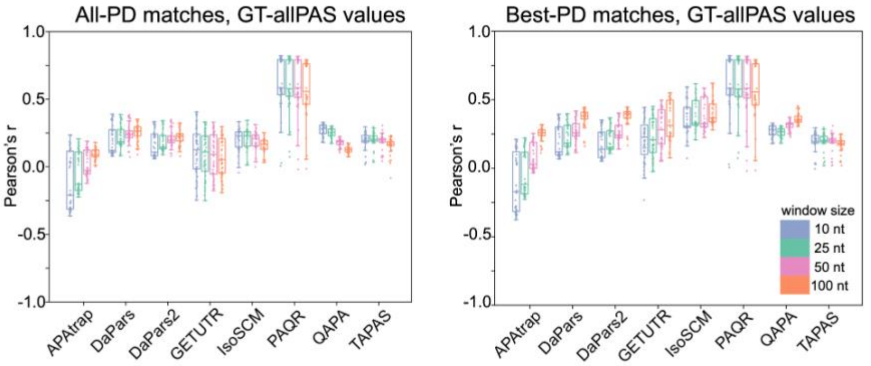
Supplemental Figure 7. Characteristics of high confidence, ground-truth (GT) terminal exon (TE) with alternative polyadenylation (APA) for relative quantification benchmarking. (A) Number of composite GT-TEs with APA defined as in Figure 4 using either RefSeq (left) or GENCODE (right) annotations. Each dot represents a ground truth experimental/simulation sample. (B) Distribution of distal polyA Usage (dPAU) for GT-TEs with APA for each GT experiment separated by experimental group based on composite TEs from RefSeq (left) or GENCODE (right).



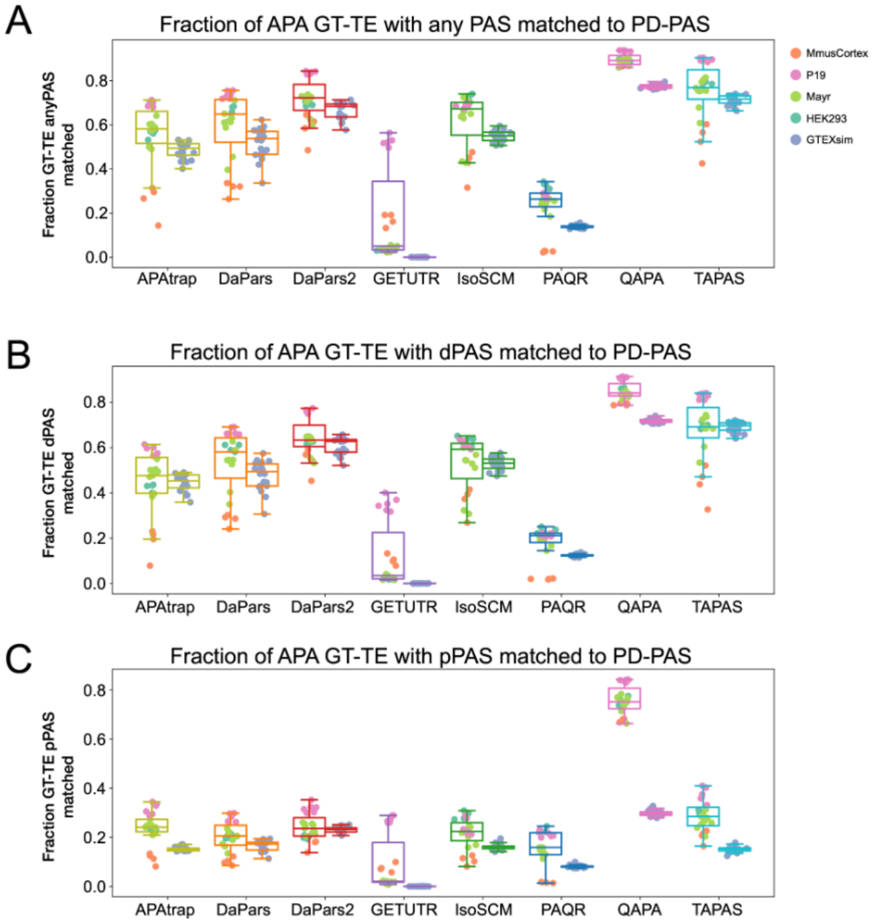
B

	GT-allPAS match?	GT-pPAS match?	GT-dPAS match?	best-PD correlation?	all-PD correlation?
PD_PAS1					
PD_PAS2a	Yes	Yes			Yes (GT_pPAS)
PD_PAS2b	Yes	Yes		Yes (GT_pPAS)	Yes (GT_pPAS)
PD_PAS3a	Yes		Yes	Yes (GT_dPAS)	Yes (GT_dPAS)
PD_PAS3b	Yes		Yes		Yes (GT_pPAS)
PD_PAS4					

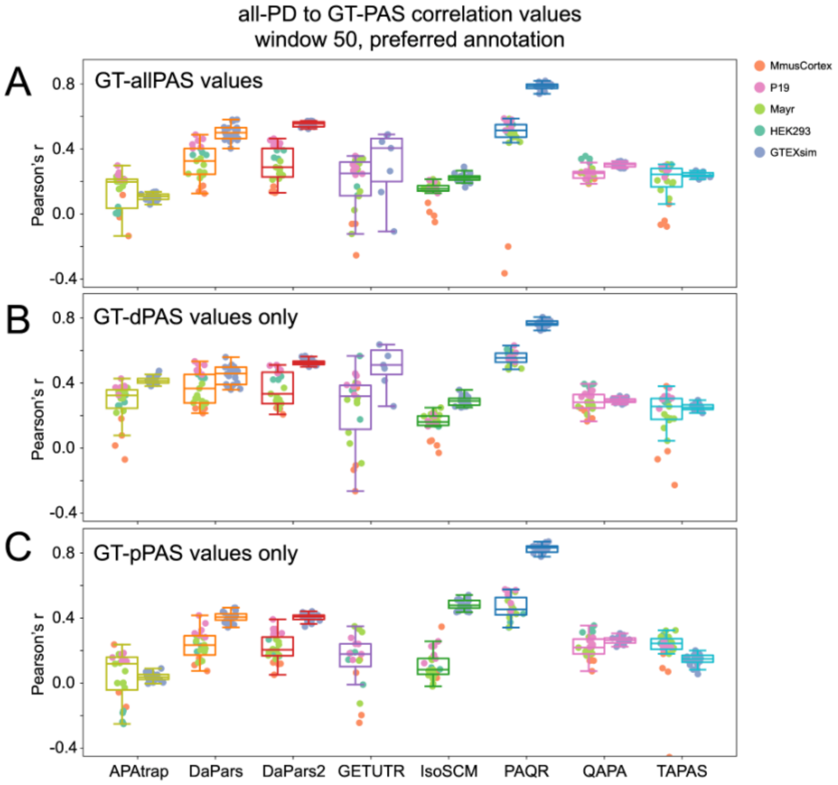
Supplemental Figure 8. Multi-match examples for certain algorithm PD to GT-PAS matches. (A) Top shows an example of a theoretical high-confidence, alternative polyadenylation containing ground-truth (GT) terminal exon (TE) and proximal PAS (GT_pPAS) and distal PAS (GT_dPAS) with polyadenylation usage (PAU) calculated by the filtering algorithm described in Methods. Bottom shows an example of an algorithm like DaPars which outputs a number TEs (PD-TEs) based on each transcript with predicted PD-PAS, some of which overlap and can have different relative quantification values (e.g. PD_PAS2a versus PD_PAS2b). (B) Table showing the different match types for each PD-PAS. “best-PD” column shows only the best match between unique GT-PAS and PD-PAS which minimizes the absolute difference between the two while “all-PD” column shows all PD-PAS to GT-PAS matches which can be used for downstream benchmarking.



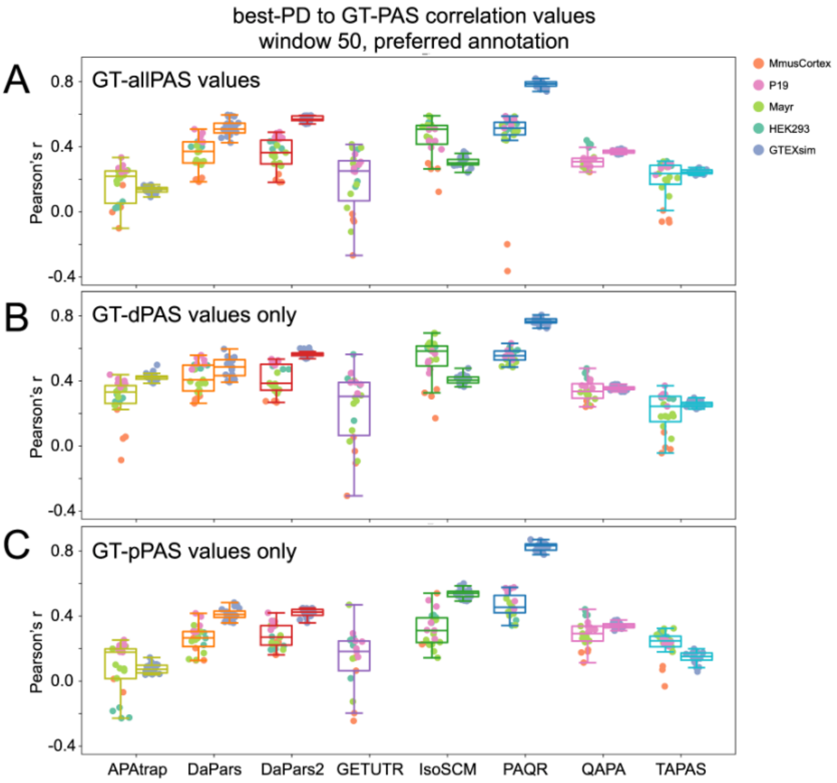
Supplemental Figure 9. Boxplots of Pearson correlation coefficients for all datasets using given window sizes and considering all PD- to GT-PAS quantification matches (left) or only the best possible PD- to GT-PAS match (right).



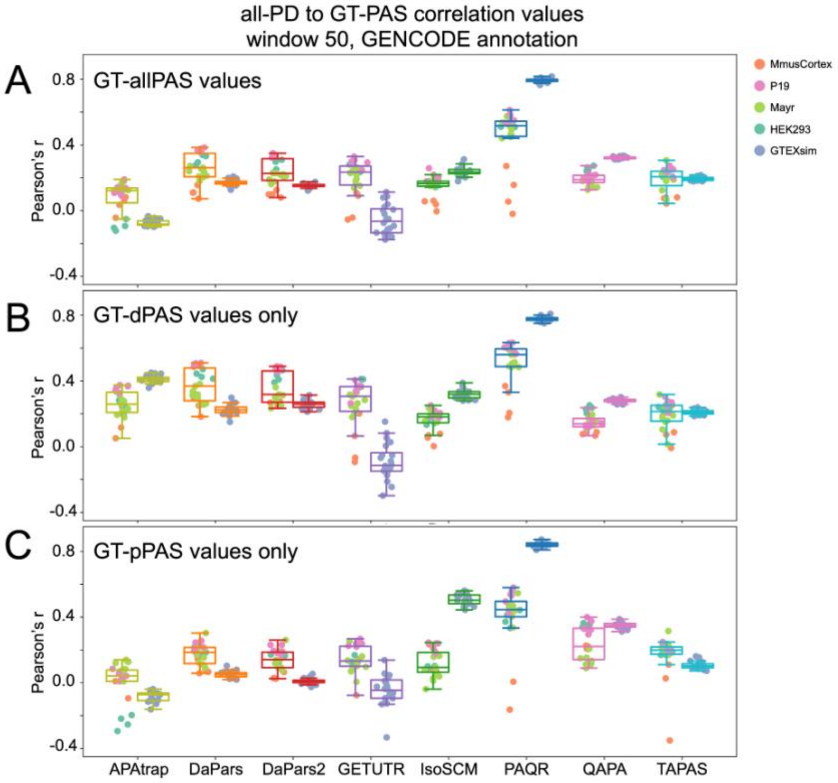
Supplemental Figure 10. (A) Fraction of unique ground truth terminal exons with alternative polyadenylation (APA GT-TE, based on RefSeq annotations) that matched to any algorithm predicted PAS (PD-PAS) using a window of 50 nucleotides. (B) Fraction of unique APA GT-TEs that had algorithm prediction PAS matched at least the distal PAS (GT-dPAS). (C) Fraction of unique terminal exons (TEs) from the ground-truth (GT) filtering that had algorithm prediction PAS matched at least the proximal PAS (GT-pPAS).



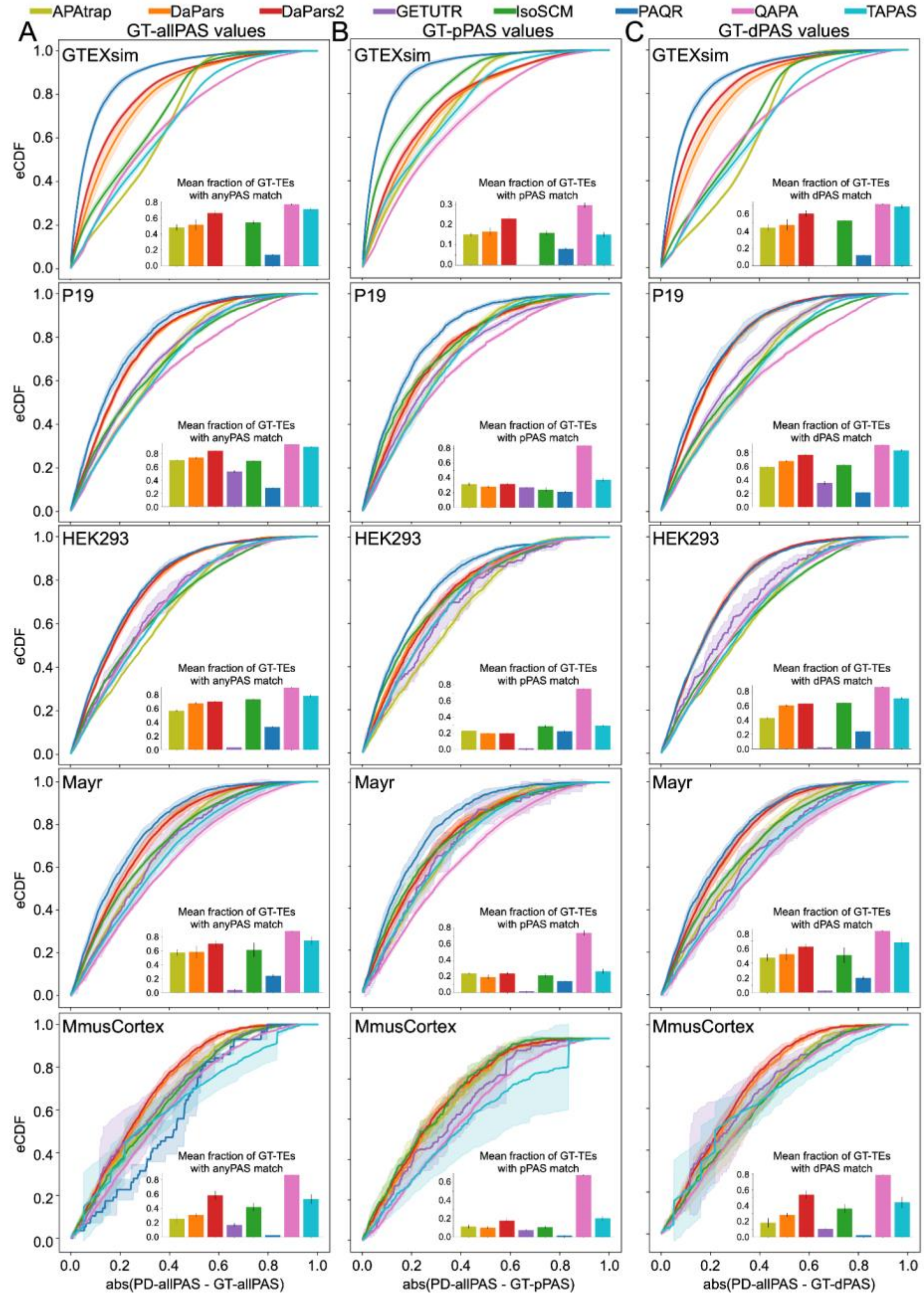
Supplemental Figure 11: The effect of GT-PAS type choice on correlation with predictions (all-PD, preferred annotation): (A) Repeat of Figure 5C shown for comparison. Pearson correlation coefficient when considering all-PD predicted values that match GT-allPAS values (both distal and proximal) using each algorithm's preferred annotation and a match window of 50 nt. Left boxes for each algorithm represent real RNA-seq data and right boxes are simulated RNA-seq data. Each point is labeled according to dataset grouping given in the legend. (B) As in (A), but using all-PD PAS matches to distal GT-PAS (GT-dPAS) values only. (C) As in (A), but using all-PD PAS matches to proximal GT-PAS (GT-pPAS) values only.



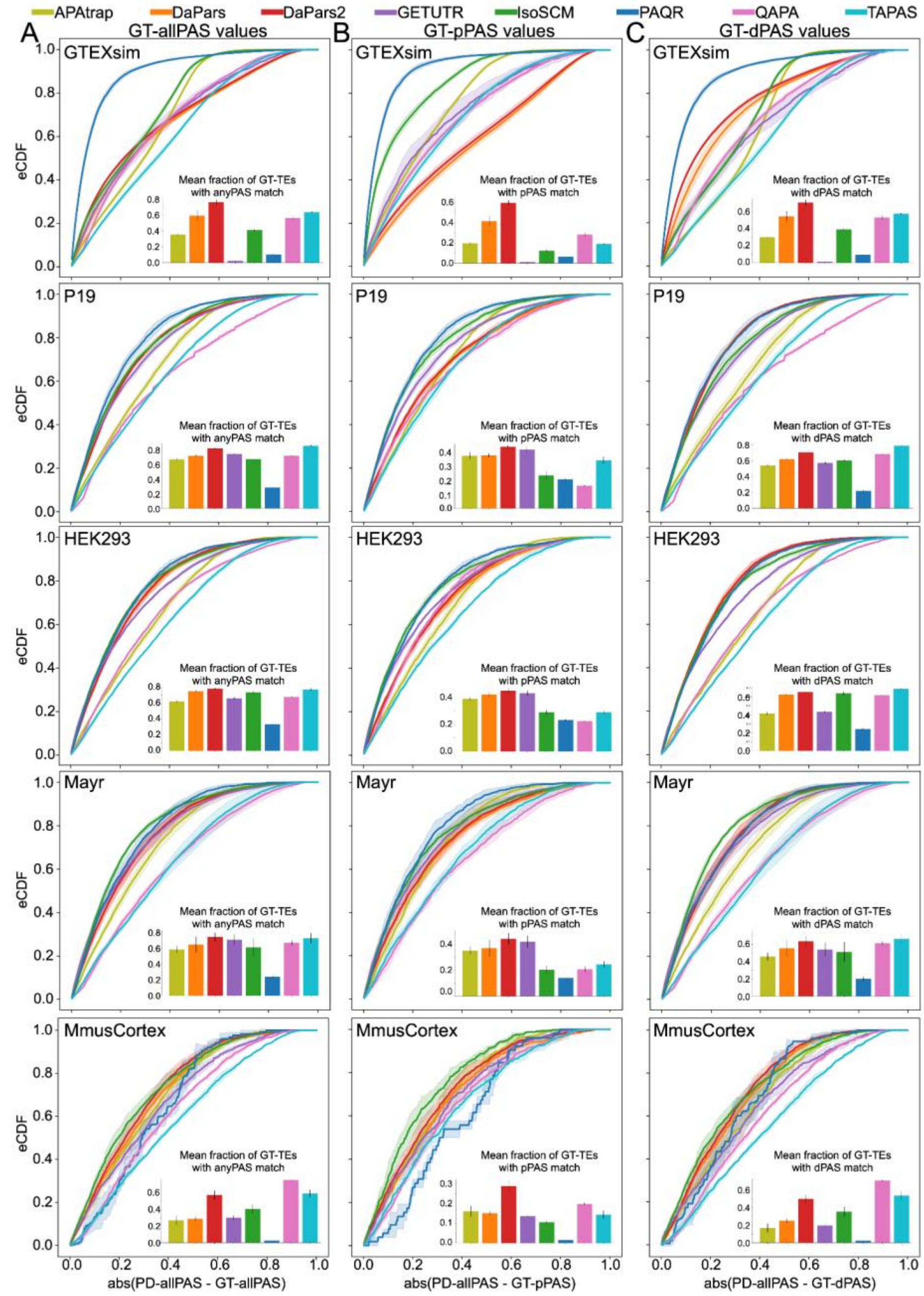
Supplemental Figure 12: The effect of GT-PAS type choice on correlation with predictions (best-PD, preferred annotation): (A) Pearson correlation coefficient when considering the single best-PD predicted values that match GT-allPAS values (both distal and proximal) using each algorithm's preferred annotation and a match window of 50 nt. Left boxes for each algorithm represent real RNA-seq data and right boxes are simulated RNA-seq data. Each point is labeled according to dataset grouping given in the legend. (B) As in (A), but using best-PD PAS matches to distal GT-PAS (GT-dPAS) values only. (C) As in (A), but using best-PD PAS matches to proximal GT-PAS (GT-pPAS) values only.



Supplemental Figure 13: The effect of GT-PAS type choice on correlation with predictions (all-PD, GENCODE): (A) Pearson correlation coefficient when considering all-PD predicted values that match GT-allPAS values (both distal and proximal) using GENCODE annotation and a match window of 50 nt. Left boxes for each algorithm represent real RNA-seq data and right boxes are simulated RNA-seq data. Each point is labeled according to dataset grouping given in the legend. (B) As in (A), but using all-PD PAS matches to distal GT-PAS (GT-dPAS) values only. (C) As in (A), but using all-PD PAS matches to proximal GT-PAS (GT-pPAS) values only.



Supplemental Figure 14. Distribution of absolute differences between ground truth and all prediction values using preferred annotations for all datasets. (A) Average eCDF for the absolute difference between all-PD matches to GT-allPAS values for each algorithm's preferred annotation for the given datasets. Lines represent the mean of all experiments in the group and shaded regions represent plus/minus one SD. Inset barchart shows the mean fraction of unique, ground-truth terminal exons with APA (defined in Figure 4, based on RefSeq annotation) represented by all-PD matches. Error bar shows one SD. Each dataset needed a minimum of 20 matched values to be plotted. (B) Same as (A), but only for matches to proximal GT-PAS (GT-pPAS) values. Inset barchart shows mean fraction of unique GT terminal exons with a pPAS matched to the algorithm predictions. Error bar shows one SD. (C) Same as (A), but only for matches to distal GT-PAS (GT-dPAS) values. Inset barchart shows mean fraction of unique GT terminal exons with a dPAS matched to the algorithm. Error bar shows one SD.



Supplemental Figure 15. Distribution of absolute differences between ground truth and all prediction values using GENCODE annotations for all datasets. (A) Average eCDF for the absolute difference between all-PD matches to GT-allPAS values for each algorithm using GENCODE annotation for the given datasets. Lines represent the mean of all experiments in the group and shaded regions represent plus/minus one SD. Inset barchart shows the mean fraction of unique, ground-truth terminal exons with APA (defined in Figure 4, based on GENCODE annotation) represented by all-PD matches. Error bar shows one SD. Each dataset needed a minimum of 20 matched values to be plotted. (B) Same as (A), but only for matches to proximal GT-PAS (GT-pPAS) values. Inset barchart shows mean fraction of unique GT terminal exons with a pPAS matched to the algorithm predictions. Error bar shows one SD. (C) Same as (A), but only for matches to distal GT-PAS (GT-dPAS) values. Inset barchart shows mean fraction of unique GT terminal exons with a dPAS matched to the algorithm. Error bar shows one SD.

Appendix D

ZARP: An automated workflow for processing of RNA-seq data

Maria Katsantoni [1,2], Foivos Gypas [3], Christina J. Herrmann [1,2], Dominik Burri [1,2], Maciej Bak [1,2], Paula Iborra [1,2], Krish Agarwal [1], Meric Ataman [1,2], Anastasiya Börsch [1,2], Mihaela Zavolan [1,2,*] & Alexander Kanitz [1,2,*]

A Biozentrum, University of Basel, Basel, 4056, Switzerland

B Swiss Institute of Bioinformatics, Lausanne, 1015, Switzerland

C Friedrich Miescher Institute for Biomedical Research, Basel, 4058, Switzerland.

* Corresponding author

D. 1. Abstract

RNA sequencing (RNA-seq) is a crucial technique for many scientific studies and multiple models, and software packages have been developed for the processing and analysis of such data. Given the plethora of available tools, choosing the most appropriate ones is a time-consuming process that requires an in-depth understanding of the data, as well as of the principles and parameters of each tool. In addition, packages designed for individual tasks are developed in different programming languages and have dependencies of various degrees of complexity, which renders their installation and execution challenging for users with limited computational expertise. The use of workflow languages and execution engines with support for virtualization and encapsulation options such as containers and Conda environments facilitates these tasks considerably. Computational workflows defined in those languages can be reliably shared with the scientific community, enhancing reusability, while improving reproducibility of results by making individual analysis steps more transparent.

Here we present ZARP, a general purpose RNA-seq analysis workflow which builds on state-of-the-art software in the field to facilitate the analysis of RNA-seq data sets. ZARP is developed in the Snakemake workflow language using best software development practices. It can run locally or in a cluster environment, generating extensive reports not only of the data but also of the options utilized. It is built using modern technologies with the ultimate goal to reduce the hands-on time for bioinformaticians and non-expert users. ZARP is available under a permissive Open Source license and open to contributions by the scientific community.

Contact: mihaela.zavolan@unibas.ch, alexander.kanitz@unibas.ch

Keywords

Computational workflow, pipeline, RNA-seq, high-throughput, reproducible research, FAIR, transcriptomics, bioinformatics

D. 2. Introduction

Recent years have seen an exponential growth in bioinformatics tools [1], a large proportion of which are dedicated to High Throughput Sequencing (HTS) data analysis. For example, for transcript-level analyses there are tools to quantify the expression level of transcripts and genes from RNA-seq data [2], identify RNA-binding protein (RBP) binding sites from crosslinking and immunoprecipitation (CLIP) data [3,4], improve transcript annotation with the help of RNA 3'end-sequencing data [5,6], estimate gene expression at the single cell level [7] or improve the annotation of transcripts and quantification of splicing events based on long read sequencing (e.g., on the Oxford Nanopore platform) [8,9]. Such tools are written in different programming languages (e.g., Python, R, C, Rust) and have distinct library requirements and dependencies. In most cases, the tools expect the input to be in one of the widely accepted file formats (e.g., FASTQ [10], BAM [11]), but custom formats are also frequently used. In addition, the variations in protocols or instruments across experiments may make it necessary to use different parameterization for every sample, rendering a joint analysis of samples from multiple studies challenging. Combining tools into an analysis protocol is a time-consuming and error-prone process. As these tasks have become so common, and as the data sets and analyses continue to increase in size and complexity, there is an urgent need for expertly curated, well-tested, maintained and easy-to-use reusable computational workflows.

A number of feature-rich, modern workflow specification languages and corresponding management systems [12,13] like Snakemake [14,15], Nextflow [16] and CWL [17] are now gaining widespread popularity in life sciences, as they make it easier for such workflows to be developed, tested, shared and executed. This leads to more reusable code and reproducible results, while fostering scientific collaborations and Open-Source Software along the way. In addition, to facilitate the installation and execution of these workflows across different hardware architectures and host operating systems, modern workflow management systems make use of virtualization and encapsulation techniques relying on containers (e.g., Docker [18] and Singularity [19]) and/or package managers (e.g., Conda [20] and Bioconda [21]). An added advantage of using workflows is the metadata stored along with the expected results. These can be invaluable for re-analyzing the data but may also provide additional insights into the results and cost analyses (e.g., runtimes, resources usage).

The aim of the presented work is the development of a flexible, easy-to use workflow for bulk RNA-seq data processing. The inclusion of the most widely used and best performing tools for the various processing steps minimizes time spent by users on making tool choices. Use of a workflow language for the development ensures the reproducibility and reliable execution of each analysis and it facilitates (meta)data management and reporting.

D. 3. Methods/Results

ZARP (**Z**avolan-Lab **A**utomated **R**NA-seq **P**ipeline) is a general purpose RNA-seq analysis workflow that allows users to carry out the most general steps in the analysis of Illumina short-read sequencing libraries with minimum effort. The workflow is developed in Snakemake [14,15], a widely used workflow language [12]. It relies on publicly available bioinformatics tools that follow best practices in the field [22], and handles bulk, stranded RNA-seq data, single or paired-end.

D. 3. 1. Workflow inputs

ZARP requires two distinct input files: (1) A tab-delimited file with sample-specific information, such as paths to the sequencing data (FASTQ format), reference genome sequence (FASTA format), transcriptome annotation (GTF format) and additional experiment protocol- and library-preparation specifications like adapter sequences or fragment size. (2) A configuration file in YAML format containing workflow-related parameters, such as results and log directory paths and user-related information. Advanced users can take advantage of ZARP's flexible design to provide tool-specific configuration parameters via an optional third input file, which allows adjusting the behaviour of the workflow to their specific needs. More information on the input files can be found in ZARP's documentation [23].

D. 3. 2. Analysis steps

A general schema of the workflow in its current version (0.3.0) is presented in Figure 1 (see Supplementary Figure 1 for a more technical representation of the entire workflow, including all of its steps). Table 1 below lists the main tools/functionalities of ZARP:

Table 1. Core tools/functionalities included in ZARP.
See main text for more information on use cases for each tool and why we chose those tools to be included in ZARP.

Tool	Description	Reference
FastQC	Generates various quality control metrics based on raw FASTQ data.	[24]
Cutadapt	Trims sequence fragments of non-biological origin or low information content.	[25]
STAR	Aligns reads to reference genome.	[26]
tin-score-calculation	Calculates a Transcript INtegrity score (TIN) on aligned reads that reflects the state of RNA degradation of a sample.	[27]
ALFA	Annotates read alignments based on gene/transcript annotations.	[28]
kallisto	Estimates gene/transcript expression levels.	[29]
Salmon	Estimates gene/transcript expression levels.	[30]
zpca	Performs principal component analyses of gene/transcript expression level estimates across samples in a given workflow run.	[31]
MultiQC	Aggregates tool results and generates interactive reports.	[32]

Calculation of per-sample quality statistics by applying **FastQC** [24] directly on the input files (FASTQ) provides a quick assessment of the overall quality of the samples. These consist of a considerable range of metrics, including, for example, GC content, overrepresented sequences and adapter content. An excessive bias in GC content may affect downstream analyses and may have to be corrected for [33]. Overrepresented sequences may be the result of PCR duplication, which, if excessive, may skew expression estimates and other downstream analyses. Information about adapter content may be used to cross-check whether it matches with whatever the user has selected to trim. For more information on the metrics that FastQC reports and how they can be interpreted, please refer to [24].

Trimming of any 5′ and/or 3′ adapters as well as poly(A/T) stretches using **Cutadapt** [25] ensures a more reliable alignment as well as removal of contaminant adapter sequences. The adapters and poly(A/T) stretches to be removed are indicated by the user.

Alignment against a given set of genome resources (either only the genome or the genome and a set of corresponding gene annotations) is the step where each read is assigned to the genomic region from which it originated. Even though there are many available aligners, **STAR** has been chosen [26][34]. BAM-formatted files that are then sorted (based on coordinates) and indexed using SAMtools [11]. These steps enable faster random access and visualisation by tools such as genome browsers. The sorted, indexed BAM files are further converted into the BigWig (BedGraphtoBigWig from UCSC tools [35]) format, which allows for library normalisation, and is thus convenient for visualising or comparing coverages across multiple samples.
The aligned reads are also used to calculate per-transcript Transcript Integrity Numbers (TIN scores) [36], a metric to assess the degree of RNA degradation in the sample. This is done with **tin-score-calculation** [27], which is based on a script originally included in the RSeQC package [37] but modified by us to enable multiprocessing for increased performance.

To provide a high-level topographical/functional annotation of which gene segments (e.g., CDS, 3’UTR, intergenic) and biotypes (e.g., protein coding genes, rRNA) are represented by the reads in a given sample, ZARP includes **ALFA** [28].

Salmon [30] and **kallisto** [29] along with a transcriptome are used to infer transcript and gene expression estimates. Since both of these tools have been shown to be equally fast, memory efficient and accurate [38], they are both included in ZARP. The main output metrics provided by either tool are estimates of normalized gene/transcript expression, in Transcripts Per Million (TPM) [39], as well as raw read counts per gene/transcript.

Within ZARP, TPM estimates are essential for performing principal component analyses (PCA) [40] with the help of **zpca** [31], a tool created by us for the use in ZARP, but packaged separately so that it can be easily used on its own or as part of other workflows. PCAs on gene/transcript expression levels can help users understand whether differences in gene/transcript expression levels across different sample groups are sufficiently high that meaningful results in downstream analyses may be expected.

TPM and raw count estimates can be further used in downstream analyses, e.g., for differential gene/transcript expression, differential transcript usage or gene set enrichment analyses. Given that such analyses require an experiment design table and are difficult to configure generically for a wide range of experiments, we chose not to include these in ZARP. However, to facilitate downstream analyses, gene/transcript estimates are aggregated for all samples with the aid of Salmon and merge_kallisto [41], which generate summary tables that can be plugged into a variety of available tools.

ZARP produces two user-friendly, web-based, interactive reports: one with a summary of sample-related information generated by **MultiQC** [32], the other with estimates of utilized computational resources generated by Snakemake itself. Note that both for tin-score-calculation and ALFA, we have created plugins so that the respective results can be explored interactively through MultiQC.

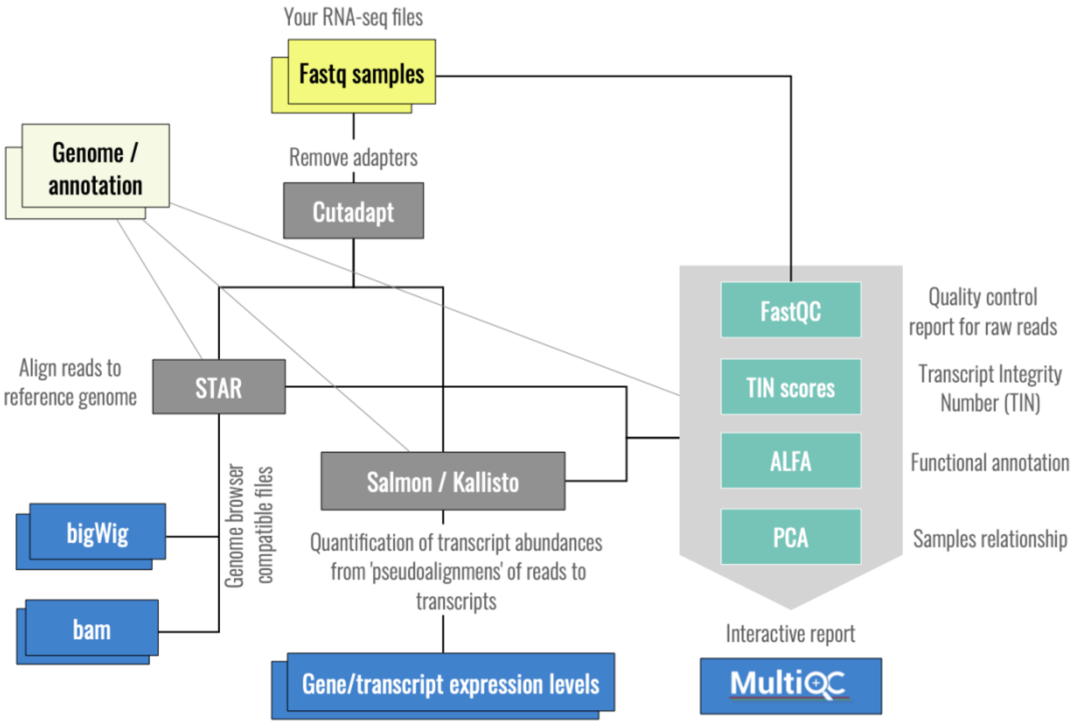

Your RNA-seq files
Fastq samples
Genome / annotation
Remove adapters
Cutadapt
FastQC
Quality control report for raw reads
Align reads to reference genome
STAR
TIN scores
Transcript Integrity Number (TIN)
ALFA
Functional annotation
Salmon / Kallisto
PCA
Samples relationship
bigWig
Genome browser compatible files
Quantification of transcript abundances from 'pseudoalignmens' of reads to transcripts
bam
Interactive report
Gene/transcript expression levels
MultiQC

Figure 1. Schematic overview of the ZARP workflow.

D. 3. 3. Reproducibility and reusability

To enhance reproducibility of results and reusability of the workflow, each step (referred to as "rule" in Snakemake) of the workflow definition relies either on Conda environments mostly hosted in the Bioconda channel [21] or on Docker images. The latter are converted by Snakemake to Singularity images [19] on the fly where needed, enabling seamless execution of the workflow in environments with limited privileges (e.g., HPC clusters). Users can choose between Conda- and container-based execution by selecting or preparing an appropriate profile when/before running a workflow. At the moment, we include profiles for the Slurm job scheduler and we plan to add new profiles over time. For that, we encourage users to feed their own profiles back to the original ZARP repository so that the entire community can benefit.

D. 3. 4. Output and documentation

In addition to the transcript/gene expression tables, ZARP collects log files and metadata for downstream analyses. Intermediate files can be optionally cleaned up by ZARP to minimize disk space usage. The workflow is hosted in its own GitHub repository, and each ZARP version released is accompanied by an up-to-date workflow-oriented description.

D. 3. 5. Continuous Integration and Testing

To facilitate collaborative development of the workflow and associated software and to reduce the chance of the codebase regressing with ongoing changes, ZARP is making use of a GitHub Actions-based workflow for Continuous Integration and Delivery (CI/CD). Each modification to the remote repository triggers a variety of integration tests (Conda environments test, Snakemake graph test, dry run, minimal-example based run) to guarantee ZARP's correct execution throughout the development cycle as the source code is refactored and new features are added.

D. 4. Use Cases

Apart from quickly gaining insights into individual samples or smaller sets of samples, ZARP is very well suited to analyze large RNA-Seq experiments or even run meta-analyses across multiple different experiments.

To demonstrate how ZARP can be used to gain meaningful insights into typical RNA-seq experiments, we tested it on an RNA-seq dataset that was generated by Ham et al. (GEO [42] accession number GSE139213) while analyzing the role of mTORC1 signalling in the age-related loss of muscle mass and function in mice [43]. The dataset consists of 20 single-ended RNA-seq libraries (read length: 101 nt, gzipped FASTQ file sizes ranging from 0.8 to 3.2 Gb, library sizes ranging from 18.5 to 75.3 reads), corresponding to four cohorts of 3-months old mice (with five biological replicates per cohort): (1) wild-type, (2) rapamycin-treated, (3) tuberous sclerosis complex 1 (TSC1) knockout and (4) rapamycin-treated TSC1 knockout. The samples were mapped against ENSEMBL's [44] GRCm38 genome primary assembly and corresponding gene annotations (release: 99) for standard human chromosomes. Other parameters for populating ZARP's samples table were obtained from the GEO accession entries of the respective samples. Sample tables and results for the test run are publicly available [45].

In Figure 2, we are presenting a subset of the outputs that ZARP generated for this dataset. We can see that the GC content of reads (Figure 2A) is slightly skewed towards being more AU-rich, yet all samples pass the FastQC-defined threshold for GC bias. Moreover, GC content does not exhibit a strong bias across samples. There is no evidence of extensive sequencing of residual adapters ("adapter contamination") (Figure 2B; black), as less than 1% of reads have been discarded in each sample because of insufficient length after adapter trimming. Transcript integrity across samples is also uniform and high (Figure 2C), with the highest density of expressed transcripts at TIN scores of 75 to 85. Similarly, alignment statistics as reported by STAR are also consistently high (Figure 2D), with rates of reads mapped uniquely against the mouse genome of more than 72% across all samples (<4% unmapped), irrespective of sequencing depth. As expected, ALFA analysis of transcript categories shows that uniquely mapped reads overwhelmingly originated from protein coding genes (over 86% for all samples) (Figure 2E). Taken together, these metrics indicate that all samples are of sufficiently high quality for downstream analyses.

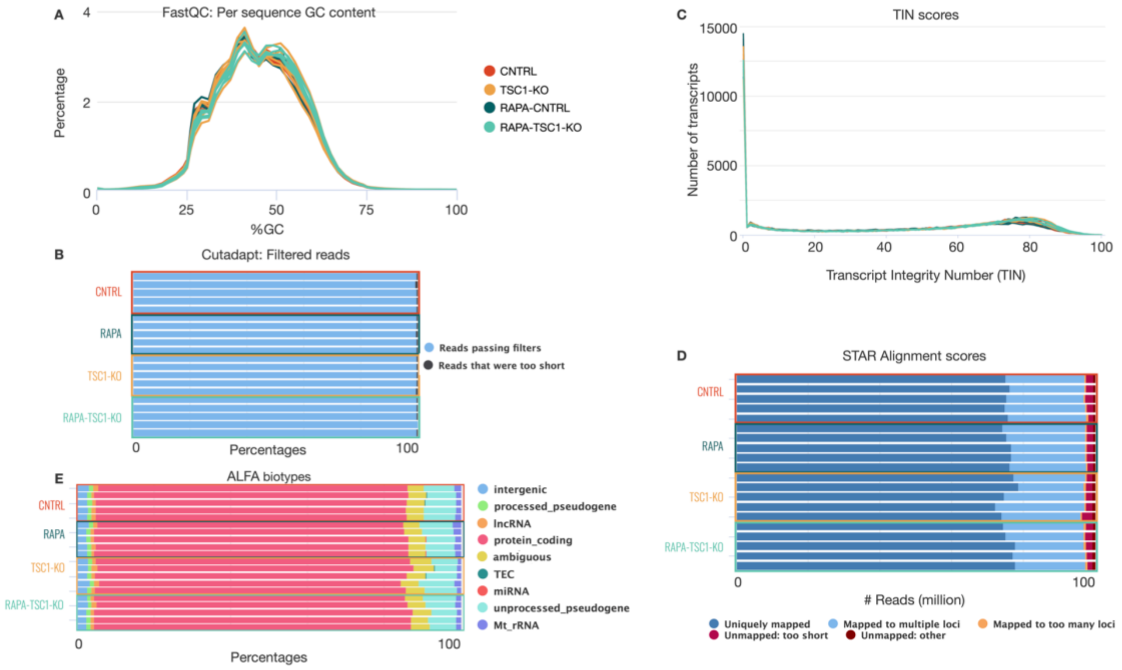


Figure 2. Selection of metrics reported by ZARP.

Shown are (A) GC content, (B) adapter removal report, (C) Transcript Integrity Number (TIN) score, (D) STAR alignment statistics, and (E) ALFA biotypes for the test run described in the main text. Figures have been edited for visibility purposes in order to group samples according to cohorts. Additionally, some biotypes have been omitted from (E) as they are not meaningfully represented. Note that in (C), transcripts that are not expressed are assigned a TIN score of 0. The complete raw html report can be found at [45].

In addition to sample-specific metrics, ZARP also provides tooling to compute principal component analyses across samples (Figure 3). For the test run, the distribution of samples in the space of the first two principal components shows a clustering by condition, with a clear separation between knockout and wild type, as well as between the untreated and rapamycin-treated TSC1 knockout mice. This separation is more pronounced at the gene expression level (Figure 3A), but is also present at the transcript level (Figure 3B). This shows that the differences across conditions are more pronounced than any replicate biases (multiplicative noise, sequencing errors), i.e., the signal-to-noise ratio is favorable, which strongly increases the likelihood that any subsequent analyses (e.g., differential gene/transcript expression analysis) will provide targets of biological importance.

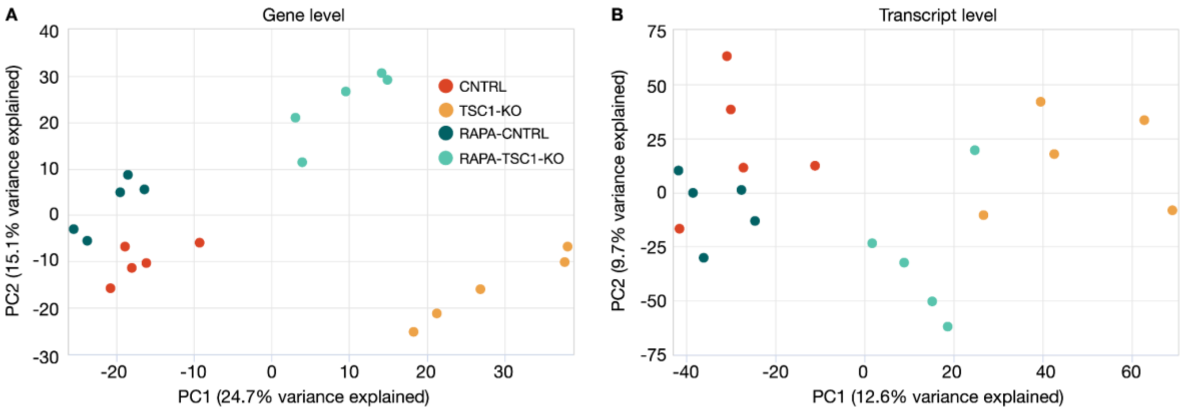


Figure 3. Principal component analysis.

Principal component analysis (PCA) at the (A) gene and (B) transcript level. PC1 and PC2 correspond to the first and second principal components, respectively. Variances explained by each of them are stated in the parentheses of the corresponding axes labels. Expression levels used in this figure are those reported by kallisto, but ZARP also generates corresponding PCA plots for Salmon-based quantifications.

The total wall clock time to execute the entire test run was just over one hour (1.01h) for all 20 samples on our Slurm-managed HPC cluster [46], where we could make heavy use of ZARP's parallelization capabilities. This translates to a total CPU time of 68.79 h, out of which 6.68h were run-specific, i.e., jobs that had to be executed only once for all samples. The accumulated sample-specific CPU time used for each sample varied between 2.75h and 8.44h. While the actual runtime may differ considerably across different compute environments, we project that most users would be able to run even large-scale analyses with dozens to hundreds of samples in less than a day on an HPC cluster, with very little hands-on time. Maximum memory usage for any of the steps and across all samples was <32 Gb (for STAR indexing and mapping of/against the human genome), indicating that ZARP is suitable for execution on state of the art computers, albeit at considerably higher runtimes due to limited parallelization capabilities, particularly for large sample groups. None of the jobs took longer than ~20 min (wall clock time) for any of the samples (Figure 4). Among the most time-consuming steps are the creation of indices (STAR, Salmon, kallisto), which however have to be performed only once per set of genome resources. Among the sample-specific steps, the calculation of the Transcript INtegrity (TIN) score was the most time-consuming. However, we had already considerably reduced its runtime by adding parallelization capabilities to the original script (see subsection "Analysis steps" in section "Methods/Results" for details).

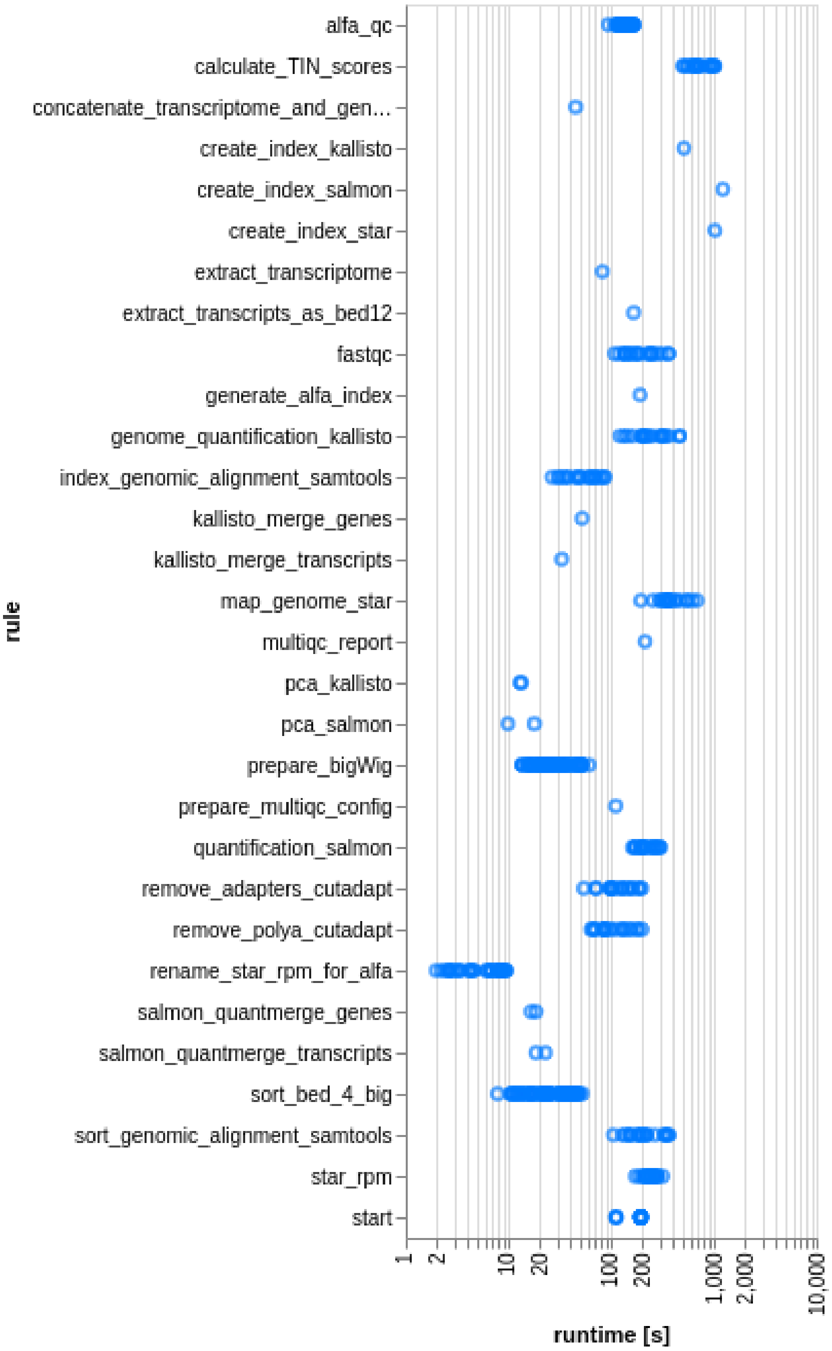


Figure 4. Runtime statistics.

Runtime (in seconds; wall clock time) of the different steps (“rules”) of the workflow run are depicted for each sample. The workflow was executed in an HPC cluster managed by the Slurm job scheduler, so the reported runtimes include the time that jobs spent queuing. Additional variation in runtimes may result from individual jobs being executed on cluster nodes with different specifications.

In summary, our test case demonstrates how ZARP can be used to quickly gain informative insights (Figures 2 & 3) into a non-trivial real-world RNA-seq analysis in a reasonable timeframe (Figure 4).

D. 5. Discussion/Conclusions

ZARP is a general purpose, easy-to-use, reliable and efficient RNA-seq processing workflow that can be used by molecular biologists with minimal programming experience. Scientists with access to a UNIX-based computer (ideally a Linux machine with enough memory to align sequencing reads) or a computing cluster can run the workflow to get an initial view of their data on a relatively short time scale. ZARP has been specifically fine-tuned to process bulk RNA-seq datasets, allowing users to run it out of the box with default parameters. At the same time, ZARP allows advanced users to customize workflow behavior, thereby making it a helpful and flexible tool for edge cases, where a more generic analysis with default settings is unsuitable. The outputs that ZARP provides can serve as entry points for other project-specific analyses, such as differential gene and transcript expression analyses. ZARP is publicly available and open source (Apache License, Version 2.0), and contributions from the bioinformatics community are welcome. Please address all development-related inquiries as issues at the official GitHub repository [47].

D. 6. Data and Software Availability

D. 6. 1. Data

Raw data analysed in section “Use Cases” are publicly available for anyone to download from the NCBI:GEO server, accession number GSE139213.

D. 6. 2. Software

The ZARP code is available on GitHub at [47] and is published under Apache License, Version 2.0. A snapshot of the ZARP version described in this manuscript (0.3.0) has been additionally uploaded to Zenodo for long-term storage [23]. Both services are public and allow anyone to download the software without prior registration.

D. 6. 3. Results

Analysis results presented in section “Use Cases” are publicly available for anyone to download from Zenodo.

D. 7. Author Contributions

MK, FG, MZ, AK conceived the project. MK, FG, CJH, DB, MB, PI, AK developed the method. MK, FG, CJH, DB, MB, PI, KA, MZ, AK developed custom tools used in the study. MA, AB tested the method with real datasets. MK, FG, CJH, DB, MB, PI, MA, MZ, AK wrote the manuscript. MK, MZ, AK supervised the study. MK, MB, AK managed the software repository. All of the authors approved the manuscript.

D. 8. Competing Interests

None declared.

D. 9. Grant Information

M.K., M.B., D.B. were supported by Biozentrum Basel International Ph.D. Program Fellowships for Excellence.

D. 10. Acknowledgements

This work would not have been possible without the sciCORE Team [46] at the University of Basel. We are thankful for their support regarding the computational infrastructure as well as dedicated time and effort to aid us in this project. We would also like to express our deepest gratitude towards all members of the Zavolan Lab who contributed to this work with numerous pieces of advice, during the initial development as well as by testing the workflow in later stages. We would also like to thank the bioconda community for helping us package and distribute some of the custom tools we developed.

D. 11. References

1. Clément L, Emeric D, J GB, Laurent M, David L, Eivind H, et al. A data-supported history of bioinformatics tools [Internet]. arXiv [cs.DL]. 2018. Available from: http://arxiv.org/abs/1807.06808

2. Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol. 2015;16:150.

3. Khorshid M, Rodak C, Zavolan M. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. Nucleic Acids Res. 2011;39:D245–52.

4. Hafner M, Katsantoni M, Köster T, Marks J, Mukherjee J, Staiger D, et al. CLIP and complementary methods. Nature Reviews Methods Primers. Nature Publishing Group; 2021;1:1–23.

5. Gruber AJ, Gypas F, Riba A, Schmidt R, Zavolan M. Terminal exon characterization with TECtool reveals an abundance of cell-specific isoforms. Nat Methods [Internet]. 2018; Available from: http://dx.doi.org/10.1038/s41592-018-0114-z

6. Herrmann CJ, Schmidt R, Kanitz A, Artimo P, Gruber AJ, Zavolan M. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. Nucleic Acids Res. Oxford Academic; 2019;48:D174–9.

7. Breda J, Zavolan M, van Nimwegen E. Bayesian inference of gene expression states from single-cell RNA-seq data. Nat Biotechnol [Internet]. 2021; Available from: http://dx.doi.org/10.1038/s41587-021-00875-x

8. Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD, Hussain S. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. Nat Commun. 2019;10:3359.

9. Karousis ED, Gypas F, Zavolan M, Mühlemann O. Nanopore sequencing reveals endogenous NMD-targeted isoforms in human cells. bioRxiv [Internet]. biorxiv.org; 2021; Available from: https://www.biorxiv.org/content/10.1101/2021.04.30.442116v1.abstract

10. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2009;38:1767–71.

11. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

12. Perkel JM. Workflow systems turn raw data into scientific knowledge. Nature. 2019;573:149–50.

13. Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods. Nature Publishing Group; 2021;1–8.

14. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics [Internet]. academic.oup.com; 2012; Available from: https://academic.oup.com/bioinformatics/article-abstract/28/19/2520/290322

15. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021;10:33.

16. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.

17. Amstutz P, Crusoe MR, Tijanić N, Chapman B, Chilton J, Heuer M, et al. Common Workflow Language, v1.0 [Internet]. Figshare; 2016. Available from: http://dx.doi.org/10.6084/M9.FIGSHARE.3115156.V2

18. Merkel D, Others. Docker: lightweight linux containers for consistent development and deployment. Linux J. 2014;2014:2.

19. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One. 2017;12:e0177459.

20. Anaconda Documentation — Anaconda documentation [Internet]. [cited 2021 Aug 23]. Available from: https://docs.anaconda.com

21. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018;15:475–6.

22. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.

23. Katsantoni M, Gypas F, Herrmann CJ, Burri D, Bak M, Iborra P, et al. ZARP: An automated workflow for processing of RNA-seq data [Internet]. Zenodo; 2021. Available from: https://zenodo.org/record/5703358

24. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010.

25. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–2.

26. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

27. tin-score-calculation: Given a set of BAM files and a gene annotation BED file, calculates the Transcript Integrity Number (TIN) for each transcript [Internet]. Github; [cited 2021 Aug 23]. Available from: https://github.com/zavolanlab/tin-score-calculation

28. Bahin M, Noël BF, Murigneux V, Bernard C, Bastianelli L, Le Hir H, et al. ALFA: annotation landscape for aligned reads. BMC Genomics. 2019;20:250.

29. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.

30. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods [Internet]. 2017; Available from: http://dx.doi.org/10.1038/nmeth.4197

31. zpca: PCA analysis [Internet]. Github; [cited 2021 Aug 23]. Available from: https://github.com/zavolanlab/zpca

32. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–8.

33. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40:e72.

34. Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Methods. Nature Publishing Group; 2016;14:135–9.

35. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2013;14:144–61.

36. Wang L, Nie J, Sicotte H, Li Y, Eckel-Passow JE, Dasari S, et al. Measure transcript integrity using RNA-seq data. BMC Bioinformatics. 2016;17:58.

37. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28:2184–5.

38. Teng M, Love MI, Davis CA, Djebali S, Dobin A, Graveley BR, et al. A benchmark for RNA-seq quantification pipelines. Genome Biol. 2016;17:74.

39. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131:281–5.

40. Jolliffe I. Principal Component Analysis [Internet]. Encyclopedia of Statistics in Behavioral Science. Chichester, UK: John Wiley & Sons, Ltd; 2005. Available from: https://onlinelibrary.wiley.com/doi/10.1002/0470013192.bsa501

41. merge_kallisto: Merge kallisto results from multiple runs [Internet]. Github; [cited 2021 Aug 23]. Available from: https://github.com/zavolanlab/merge_kallisto

42. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41:D991–5.

43. Ham DJ, Börsch A, Lin S, Thürkauf M, Weihrauch M, Reinhard JR, et al. The neuromuscular junction is a focal point of mTORC1 signaling in sarcopenia. Nat Commun. 2020;11:4510.

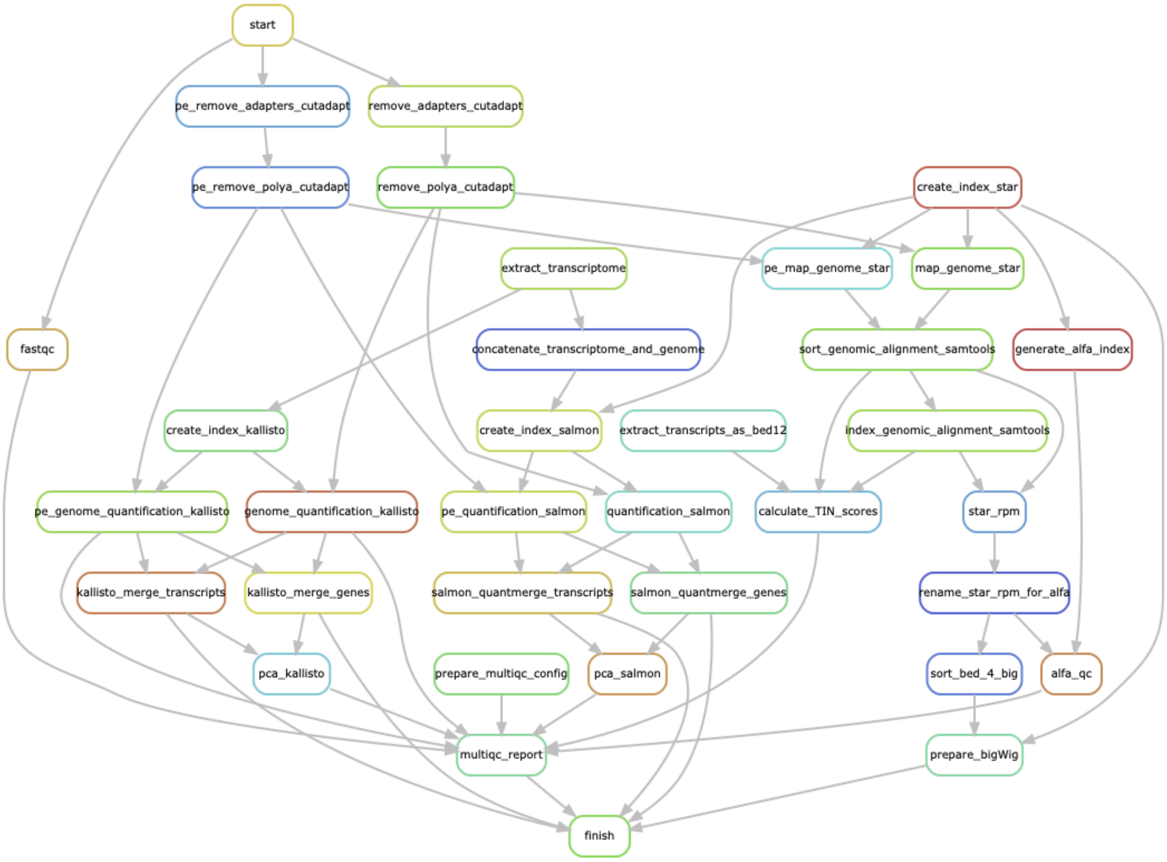
44. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Res. 2021;49:D884–91.

45. Ataman M, Börsch A, Bak M. ZARP: Supplementary Materials [Internet]. Zenodo; 2021. Available from: https://zenodo.org/record/5683524

46. sciCORE [Internet]. [cited 2021 Nov 15]. Available from: http://scicore.unibas.ch/

47. zavolanlab. GitHub - zavolanlab/zarp: Zavolan-Lab Automated RNA-Seq Pipeline [Internet]. [cited 2021 Nov 15]. Available from: https://github.com/zavolanlab/zarp

D. 12. Supplementary material



Supplementary Figure 1. ZARP workflow schema.
Graph-based representation of ZARP v0.3.0, including all of its steps ("rules"), as produced by running Snakemake with the `--rulegraph` option. Steps for both the single and the paired end workflows are shown.