Published online 11 September 2023

NAR Genomics and Bioinformatics, 2023, Vol. 5, No. 3 **1**
https://doi.org/10.1093/nargab/lqad079

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

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

[1]Computational and Systems Biology, Biozentrum University of Basel, Spitalstrasse 41, CH-4056 Basel, Switzerland and [2]Swiss Institute of Bioinformatics, Basel, Switzerland

Received April 27, 2023; Revised August 15, 2023; Editorial Decision August 21, 2023; Accepted August 23, 2023

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.

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.

MATERIALS AND METHODS

Analyzed datasets

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

*To whom correspondence should be addressed. Tel: +41 61 207 15 77; Email: mihaela.zavolan@unibas.ch
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

(15) were downloaded from the NCBI's GEO database (accession numbers GSE106264 and GSE104556 respectively). Table 1 provides further information on these samples.

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.

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.

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

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

Table 1. scRNA-seq datasets used in the study

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

SRR: sequence read archive run identifier, BAM: binary alignment map, GB: gigabyte, NA: not available.

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

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)

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

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 <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.

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 >80% A's.

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.

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 (Figure 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.

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 (Figure 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 <100 nucleotides from the annotated TE end (annotated in terminal exon, ATE) and those whose PAS was farther away (unannotated in terminal exon, UTE).

Classification of PATR

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).

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 containing the respective poly(A) signal. We then traversed the data frame 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 data frame 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

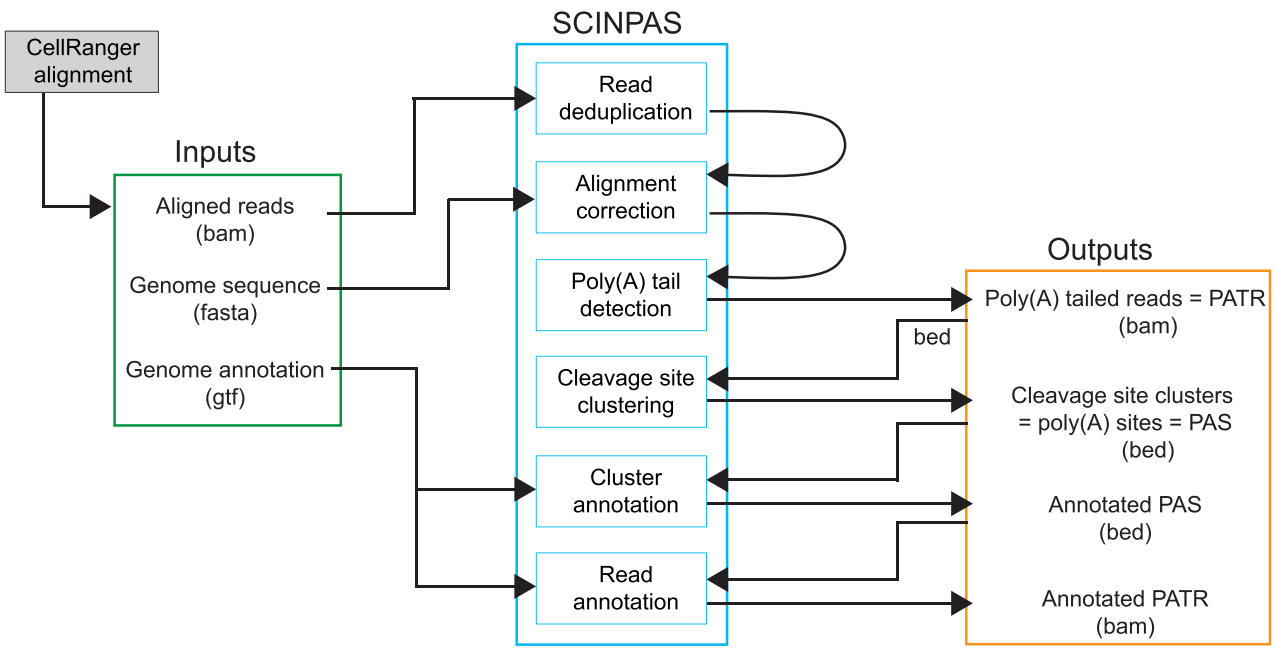


Figure 1. Scheme of SCINPAS workflow**.** The inputs to SCINPAS are indicated in the green box. Alignments of reads from primary samples are generated with CellRanger. The SCINPAS processing steps are shown in the cyan boxes and the outputs of the workflow are indicated in the orange box. File formats for inputs and outputs are indicated in parentheses.

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.

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 '1 kb 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 PATR. This was also the case when we compared SCINPAS to SCAPE. To visualize the number of PAS clusters, the individual classes (TE, exons, introns, intergenic and '1 kb downstream genes') were extracted and plotted as stacked bar charts with `geom_col`.

The distance between a PAS cluster 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 ($g_s \leq p_e$) & ($g_e \geq p_s$) for SCAPE and ($g_s \leq p_e + 1$) & ($g_e \geq p_s - 1$) 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).

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 ($y_s \leq x_e + 1$) & ($y_e \geq x_s - 1$). 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.).

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 −40 bp to +20 bp. 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).

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.

RESULTS

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.

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 Figure 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 <10 kb (Figure 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 (Figure 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 (Figure 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 (Figure 2D). Thus, we implemented an additional step of refining the alignment by extending

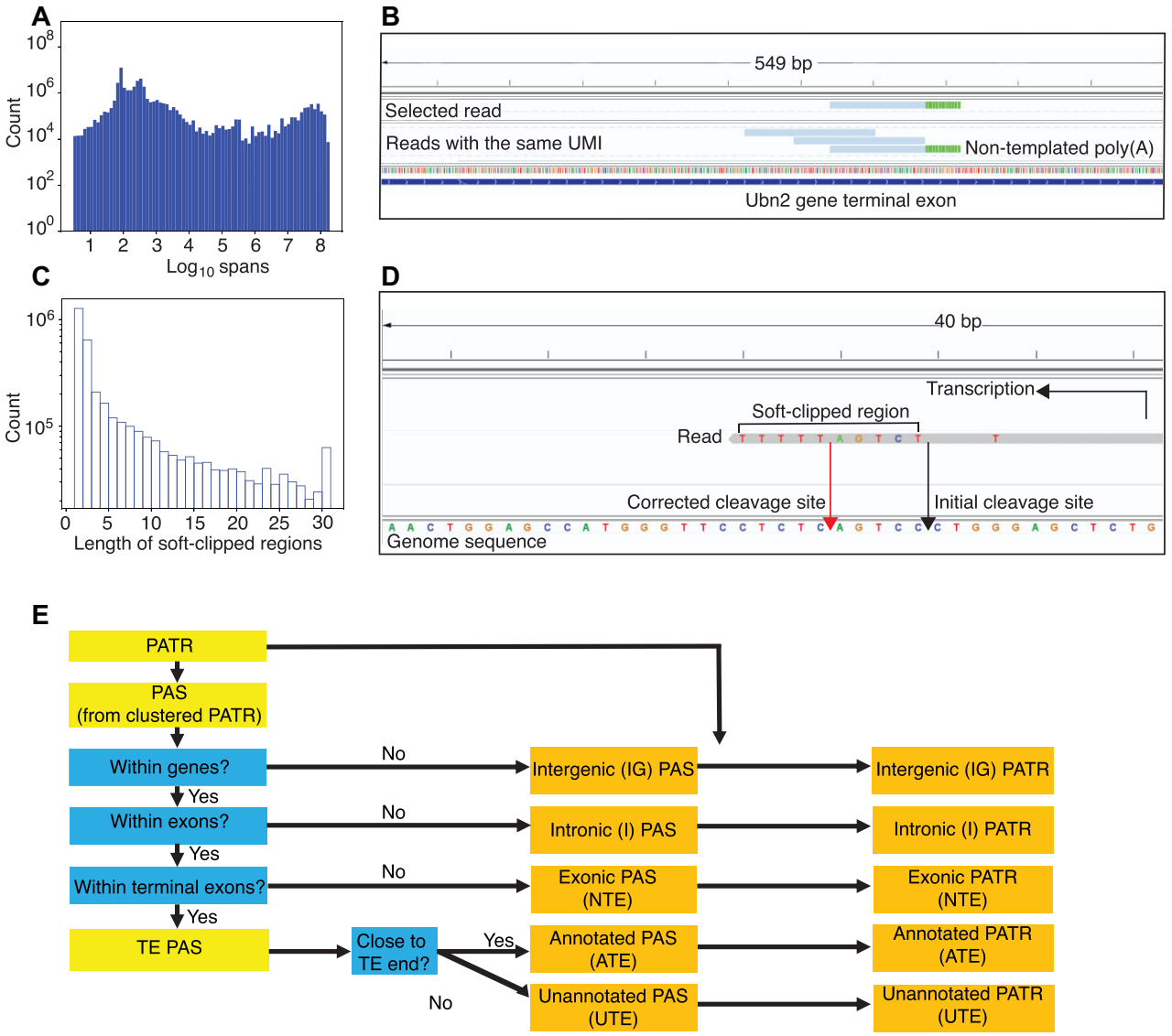


Figure 2. Key steps in SCINPAS. (**A**) Distribution of genomic spans of reads with the same cell and molecular identifier (CB and UR tag, $\log_{10}$) constructed from the sample 10X_P7_14 of *Tabula Muris Senis*. (**B**) Illustration of distal read selection, from among the reads with the same CB and 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.

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 Materials and Methods). Finally, we selected the reads that contained non-templated poly(A) tails of at least five 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 (Figure 2E).

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 %), 2329 to other exons (NTE – 0.4 %), 34 484 to introns (I – 5.3 %) and 200 176 to intergenic regions (IG – 30.7 %) (Figure 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 5060 (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 Figure 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 (Figure 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 (Figure 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 (Figure 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 (Figure 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 (Figure 3F). Altogether, these results demonstrate that our tool improves the recovery of bona fide PAS from scRNA-seq data relative to the standard workflows.

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 (Figure 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 (Figure 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 (Figure 4G).

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

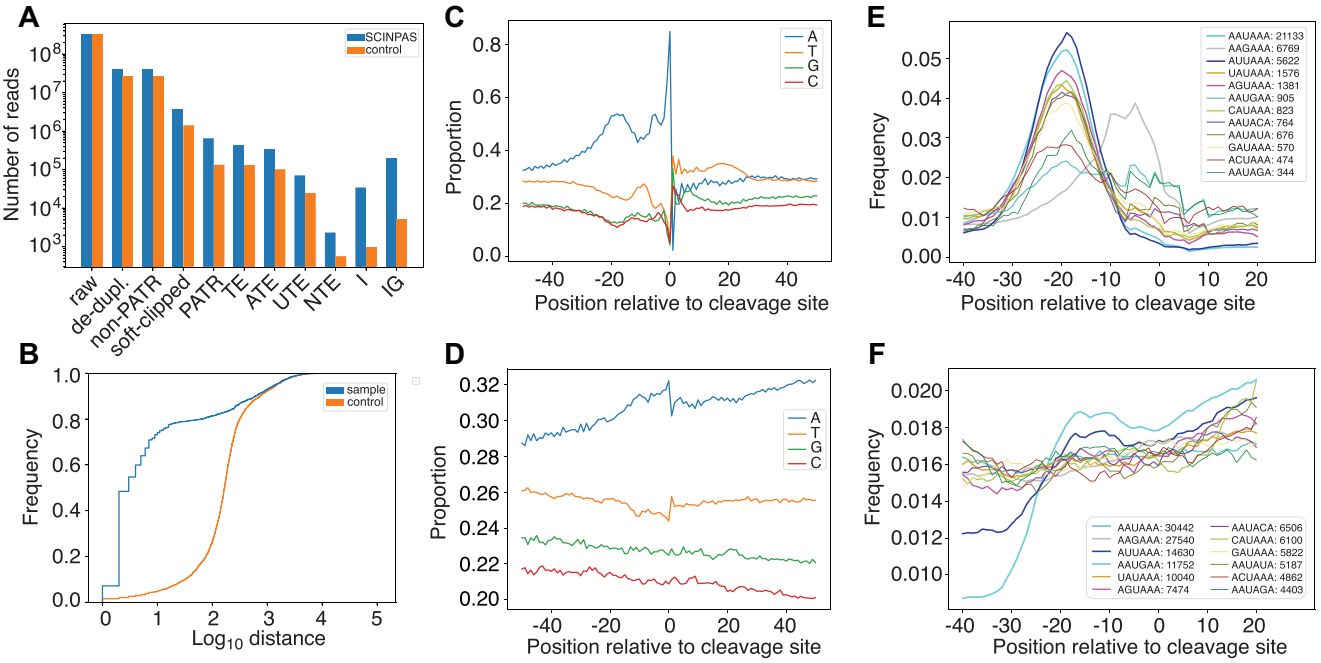


Figure 3. PAS extracted by SCINPAS contain the expected poly(A) signals. (**A**) The number of reads from the *Tabula Muris Senis* sample 10X_P7_14, at different steps of the processing pipeline, when the processing is done with SCINPAS (blue) or the standard UMI-tools-based workflow (orange). (**B**) 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.

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%, Figure 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% versus 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%, Figure 5B). As with other analyzed datasets, the intronic PAS inferred from activated T cells (Figure 5C) and elongating spermatids (Figure 5D) had the expected peak poly(A) signals at ∼20 nucleotides upstream of the inferred cleavage site (Figure 5C, D). An example of intronic PAS usage in the sperm cell differentiation system is shown in Figure 5E.

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).

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 Figures 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 (Figure 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 Figure 6B and in both cases they have a prominent peak located at approximately 50’000

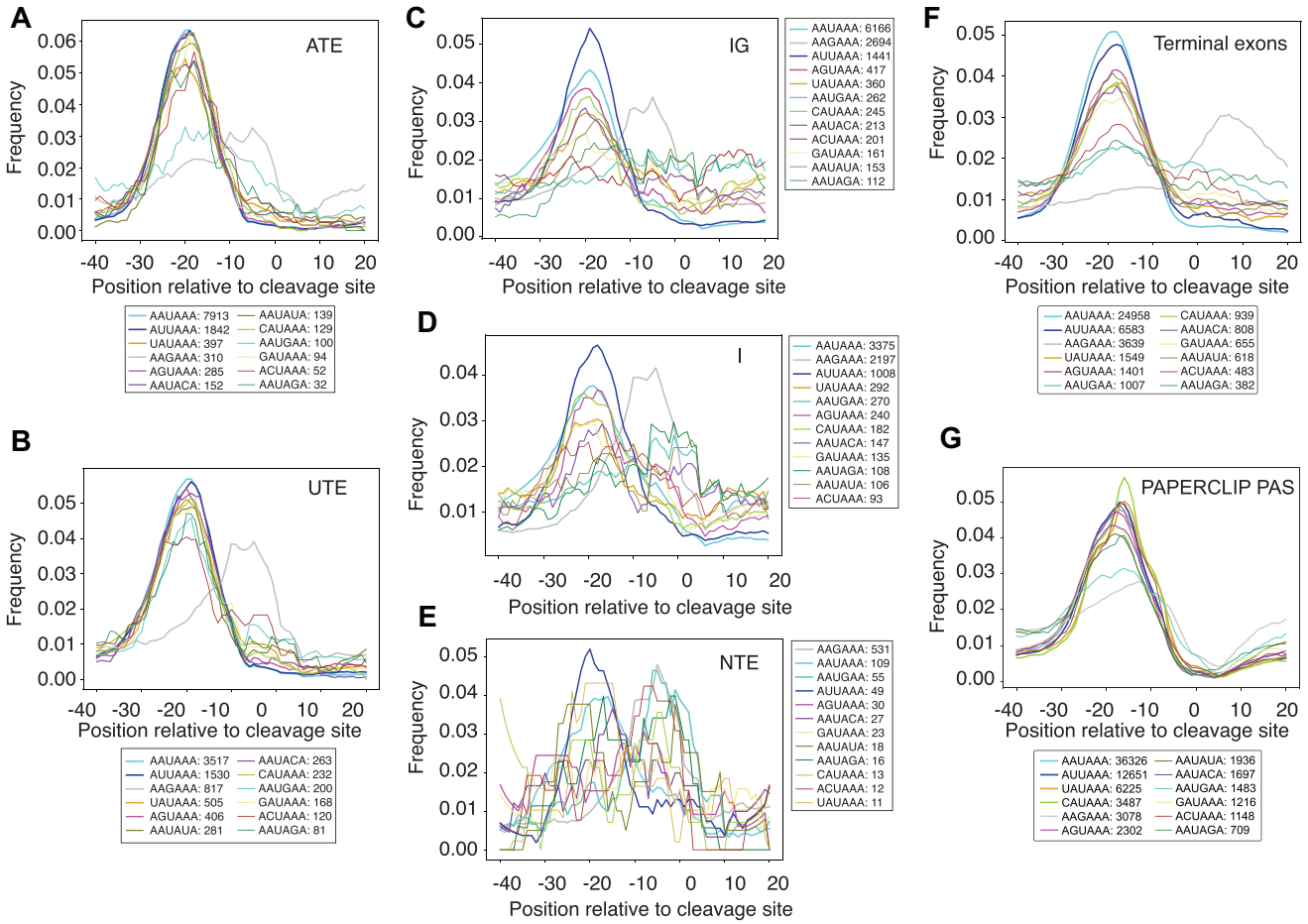


Figure 4. Position-dependent frequency of occurrence of poly(A) signals at different types of PAS. (**A–E**) PAS were extracted and annotated with SCINPAS from the *Tabula Muris Senis* sample 10X_P7_14. ATE – PAS within 100 nucleotides of annotated TE ends; UTE – PAS in TEs but >100 nucleotides from 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.

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 (Figure 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, Figure 6C). In some instances, SCINPAS detected more PAS clusters per gene compared to SCAPE (Figure 6D, left panel), though examples where the opposite was the case also occurred (Figure 6D, right panel). We performed the same analysis as above on the sperm cell development dataset and found similar trends (Figure 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 (Figures 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.

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.

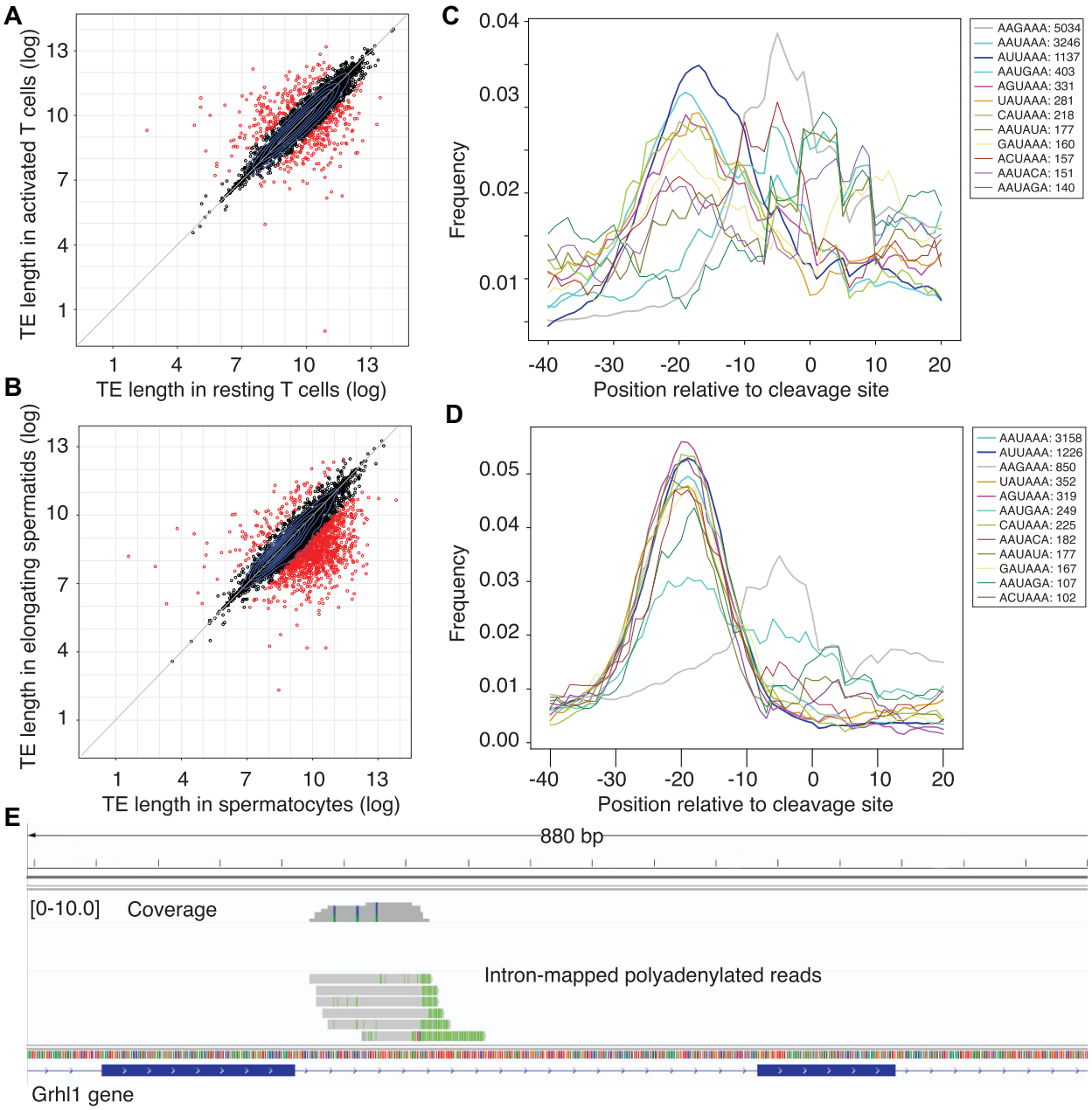


Figure 5. SCINPAS-recovered sites reproduce APA patterns in previously characterized systems. (**A**) Scatter plot of average terminal exon (TE) length (log2 values) computed from the location and relative abundance of PATR mapping to individual terminal exons. Highlighted in red are TEs whose length 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.

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.

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 h 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

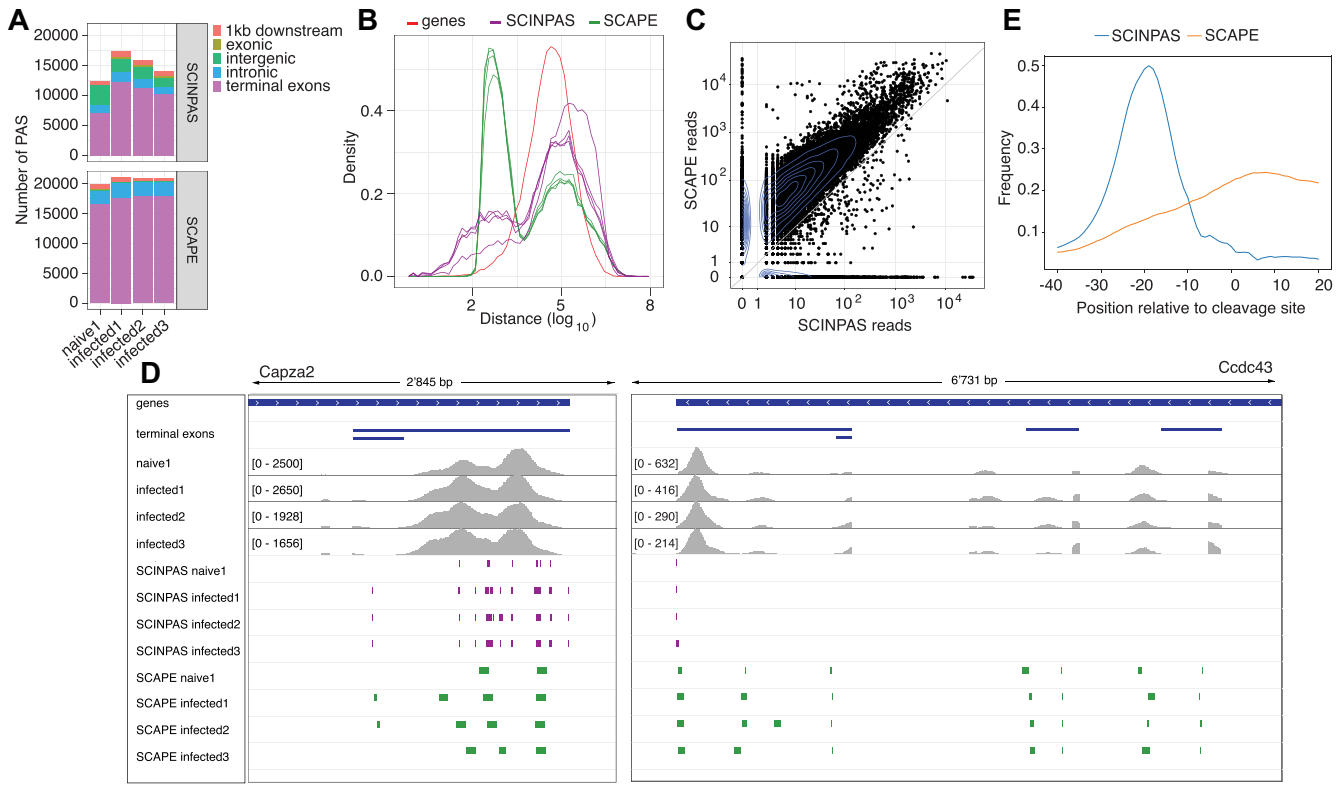


Figure 6. PAS recovery by SCINPAS and SCAPE. (**A**) Number of PAS inferred by SCINPAS (top) and SCAPE (bottom) from the T cell activation data. The colors indicate different classes of PAS (see legend). (**B**) Distribution of the distances from each PAS to the closest PAS downstream for SCINPAS (purple) 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.

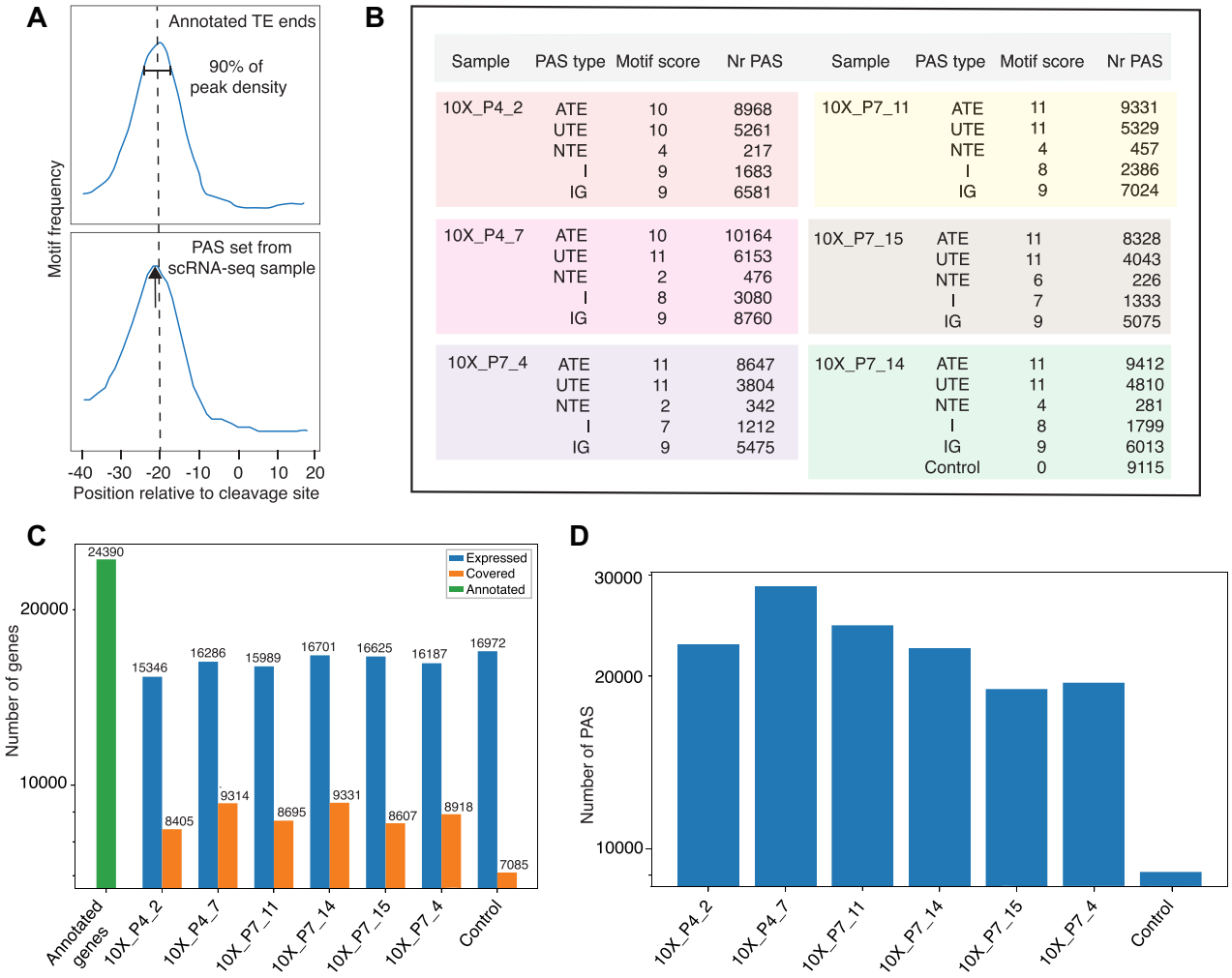
Table 3. Jaccard statistics. Pairwise comparison of SCINPAS (left) and SCAPE (right) predicted PAS in individual samples from the sperm cell development (mouse 1 and 2) and T cell (naive 1 and infected 1–3) activation datasets

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

to occur at the expected position if it was located within the 90% peak frequency window inferred from annotated terminal exon ends (Figure 7A). As shown in Figure 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 (Figure 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 Figure 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 Figure 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 (Figure 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
10X_P4_2	ATE	10	8968
	UTE	10	5261
	NTE	4	217
	I	9	1683
	IG	9	6581
10X_P4_7	ATE	10	10164
	UTE	11	6153
	NTE	2	476
	I	8	3080
	IG	9	8760
10X_P7_4	ATE	11	8647
	UTE	11	3804
	NTE	2	342
	I	7	1212
	IG	9	5475
10X_P7_11	ATE	11	9331
	UTE	11	5329
	NTE	4	457
	I	8	2386
	IG	9	7024
10X_P7_15	ATE	11	8328
	UTE	11	4043
	NTE	6	226
	I	7	1333
	IG	9	5075
10X_P7_14	ATE	11	9412
	UTE	11	4810
	NTE	4	281
	I	8	1799
	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.

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 addresses 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 (<1 kb) downstream of terminal exon ends (Figures 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). Twelve 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 (Figures 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 (Figure 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 (Figure 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 h, 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.

DATA AVAILABILITY

SCINPAS is packaged into a nextflow workflow (24). The code and analysis are available from: https://github.com/zavolanlab/SCINPAS (permanent DOI: https://doi.org/10.5281/zenodo.8272892).

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

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

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.

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.

FUNDING

Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung [310030_189063].

Conflict of interest statement. None declared.

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.*, **20**, 599–615.
3. Mitschka,S. and Mayr,C. (2022) Context-specific regulation and function of mRNA alternative polyadenylation. *Nat. Rev. Mol. Cell Biol.*, **23**, 779–796.
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 Bioinf.*, **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**, bbaa273.
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. Consortium,T.M. (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. 1000 Genome Project Data Processing Subgroup, Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R.(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**, giab007.
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 [Electronic Resource]*, **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–W190.