Nucleic Acids Research, 2026, **54**, D247–D254
https://doi.org/10.1093/nar/gkaf1212
Advance access publication date: 29 November 2025
Database issue

PolyA_DB v4: systematic polyA site identification and isoform annotation in human and mouse genomes using 3′ end and long-read sequencing data

Shan Yu[1,†], Wei Chun Chen[1,†], Luyang Wang[1,†], San Jewell[2], Ayna Mammedova[1], Seong Woo Han[2,4], Jayamanna Wickramasinghe[3], Yoseph Barash[2], Bin Tian[1,3,*]

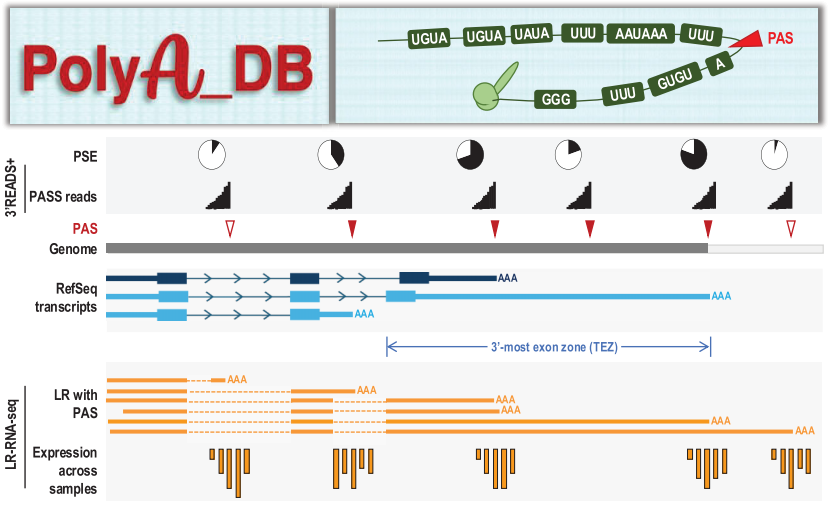
[1]Genome Regulation and Cell Signalling Program, Ellen and Ronald Caplan Cancer Center, The Wistar Institute, Philadelphia, PA 19104, United States
[2]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, United States
[3]Center for Systems and Computational Biology, The Wistar Institute, Philadelphia, PA 19104, United States
[4]Present address: Department of Computer and Information Science, University of Pennsylvania, PA 19104, United States

*To whom correspondence should be addressed. Email: btian@wistar.org
†The first three authors should be regarded as Joint First Authors.

Abstract

The cleavage and polyadenylation site (PAS) defines the 3′ end of almost all protein-coding and long non-coding RNAs in eukaryotes. Most genes harbor multiple PAS, resulting in expression of alternative polyadenylation (APA) isoforms. Here, we present PolyA_DB version 4 (https://exon.apps.wistar.org/polya_db/v4/), an updated database dedicated to PAS in mammalian genomes. By exhaustive mining of human and mouse transcriptomic data sets generated by the 3′ region extraction and deep sequencing plus (3′READS+) method, corresponding to ∼2.3 billion PAS-supporting reads for each species, we identify ∼1.4 million PAS in both human and mouse genomes, increasing PAS coverage over the last database version by 4.9- and 3.5-fold, respectively. Of the full PAS set (named Max collection), 20% of them match the transcript end sites (TES) of public long-read RNA sequencing (LR-RNA-seq) data. Notably, ∼10%–20% of LR-RNA-seq TES do not match our annotated PAS, suggesting 3′ end artifacts derived plausibly from internal A-rich regions of RNA. However, LR-RNA-seq data substantially complement RefSeq-based assignment of PAS to genes and are highly valuable in subtyping APA events in the context of splicing configuration. PolyA_DB v4 also contains PAS conservation and PAS strength information and is linked to UCSC Genome Browser for data visualization.

Graphical abstract



Received: September 16, 2025. **Revised:** October 15, 2025. **Accepted:** October 16, 2025

Introduction

Cleavage and polyadenylation (CPA) is responsible for 3′ end processing of almost all protein-coding RNAs (mRNAs) and long noncoding RNAs (lncRNAs) in eukaryotes [1, 2]. The CPA site, commonly known as the polyA site (PAS), is defined by surrounding motifs (named PAS motifs), which recruit the CPA machinery through its RNA-binding protein (RBP) components for PAS usage. The well-characterized PAS motifs for mammalian PAS include upstream UGUA, U-rich, and A[A/U]UAAA (also known as the PAS hexamer), as well as downstream UGUG, U-rich, and G-rich motifs. These PAS motifs collectively determine the strength of a PAS, or its general frequency of usage. In addition, other RNA motifs and their cognate RBPs can influence the PAS usage, often in a cell type-specific manner [3–5]. CPA is generally coupled with transcriptional termination [6]. Suppression of CPA can lead to transcriptional readthrough and expression of downstream-of-gene (DoG) transcripts, which have been reported to take place under several conditions, such as cellular stress, viral infection, and cancer [7–9]. Moreover, PAS motif mutations have been implicated in a growing number of human diseases [10], highlighting the importance of CPA for gene expression.

Well over half of mammalian genes harbor multiple PAS, resulting in expression of alternative polyadenylation (APA) isoforms [3, 5]. Most of the APA events occur in the 3′-most exon of a gene, leading to isoforms with variable 3′ untranslated regions (3′UTRs); a sizable fraction of genes, e.g. ∼20% in humans, are located in regions upstream of the last exon. These PAS, mostly in regions that can be removed by splicing, could lead to intronic polyadenylation and premature transcriptional termination. APA is increasingly being appreciated as a key mechanism for gene regulation [3, 5], diversifying the transcriptome in different cell types and under various pathological and physiological conditions. Accurate and sensitive analysis of APA isoforms is of great importance for understanding the mechanisms and consequences of APA and can have diagnostic values in clinical settings [11].

PAS can be detected *de novo* by using the poly(A) tail information in the complementary DNA (cDNA) sequence. RNA-seq methods biased to the 3′ end are particularly suitable for PAS identification and expression analysis of APA isoforms, such as QuantSeq [12] and single-cell 3′ sequencing method from 10× Genomics [13]. However, the oligo(dT) used in most methods could prime at internal A-rich sequences or cause template switching, leading to incorrect PAS identification [14–16]. It was estimated that over 20% of cDNA sequences in the public expressed sequence tag database are derived from internal primed transcripts [14]. In contrast, methods that involve direction RNA sequencing [17, 18] and reverse transcription from an adapter sequence ligated to the 3′ end of poly(A)+ RNA provide *bona fide* poly(A) tail information and do not have internal priming or template switch issues [19–21]. Our lab previously developed the 3′ region extraction and deep sequencing (3′READS) method [19] and its updated version, 3′READS+ [21], the latter of which involves poly(A) tail trimming to ∼15 bases by RNase H before 3′ adapter ligation [19, 21]. The ∼15-nt A-stretch region (5′T-stretch in sequence) in the cDNA is often long enough to support the presence of the poly(A) tail but also short enough to avoid the homopolymer sequencing issue on the Illumina platform.

While the importance of PAS regulation and APA isoform expression has been increasingly appreciated in recent years, comprehensive and accurate annotations of PAS and APA isoforms remain severely unfulfilled. Major transcriptomics databases, such as RefSeq [22], Ensembl [23], and GENCODE [24], are neither comprehensive nor accurate for PAS and APA isoform annotations. Efforts have been made to develop databases dedicated to PAS. For example, PolyASite 2.0 [25] annotates PAS based on 12 types of bulk 3′ seq methods, covering a variety of species, from yeast to humans. Its latest version also provides APA isoform expression information from single-cell RNA-seq data [26]. scAPAdb is another database that covers PAS and APA events at the single-cell level for multiple species, including human, mouse, and several plants [27]. More recently, PASSpedia offers a comprehensive resource for PAS annotation and comparison across seven species at single-cell resolution [28]. In addition, databases for tissue-specific and disease-specific APA changes, such as cancers, have been developed to assist in APA research in pathological conditions [29–30]. Moreover, collections of genetic variant-associated APA changes have shed light on APA differences among human populations [31–34]. These databases provide rich information about PAS and APA isoform expression. However, falsely annotated PAS due to internal priming and template switching are not fully addressed, despite that computational approaches are commonly used to remove PAS in A-rich genomic sequences. Another limitation of existing PAS databases is that the short reads used to identify PAS lack information about the full-length transcript, making it difficult to assign a PAS to its corresponding gene with high accuracy, especially those in introns or downstream regions of gene (named orphan PAS in this study). The advent of long-read (LR) RNA-seq technologies, such as PacBio IsoSeq and Oxford Nanopore (ONT) cDNA or direct RNA sequencing, can be instrumental in PAS-to-gene assignment [35, 36] and reveal splicing and transcriptional start site information that are coupled with PAS usage [37].

Here, we present an updated version of PolyA_DB (v4) based on all available data generated by 3′READS+ [21]. In addition to using a consistent data type (3′READS+) for PAS identification, the large number of samples ensure comprehensive PAS coverage and quantitative assessment of APA isoform expression. We incorporate current, public LR-RNA-seq data for thorough annotation of PAS, including their assignment to genes and APA types. LR-RNA-seq data also help validate commonly used PAS and provide expression information across tissues and cell types.

Materials and methods

3′READS+ data processing

We collected all 3′READS+ data in the public domain as well as a few data sets that are soon to be published in research papers by our lab. 3′READS+ data were processed by using the method we previously described [21] with modifications aimed to achieve more stringent identification of PAS. Briefly, the 5′ adapter sequence of each 3′READS+ sequence was first removed using the Cutadapt program [38], followed by trimming of random nucleotides (UMI sequence) and 5′ T-stretch. The remaining sequence, required to be at least 18 bases, was aligned to the reference genome (hg38 and mm10 for human and mouse, respectively) using the Bowtie2 pro-

gram [39]. Reads with mapping quality scores (MAPQ) < 10 were discarded (not including the 5′ T-strech). The last aligned position (LAP) in the genome was identified for each read, and its downstream sequence was compared with the 5′ T-stretch trimmed from the read. If a read had ≥5 bases in its 5′ T-stretch that were not found in the genome, it was considered as a PAS-supporting (PASS) read, and its LAP was considered a cleavage site (CS). We also required each data set to have at least 10% PASS reads of all raw reads to ensure only high-quality samples were used. CS located within 24 nt from one another in the genome were clustered together to address heterogeneous cleavage for the same PAS [40]. The CS with highest overall expression value (reads per million, RPM) in each cluster was used as the representative CS of a PAS.

PAS motif annotation

PAS hexamers were identified within 40 nt upstream of the PAS, include AAUAAA,AUUAAA, and some of their single-nucleotide variants, including AGUAAA, UAUAAA, CAUAAA, GAUAAA, AAUAUA, AAUACA, AAUAGA, AAAAAG, ACUAAA, and AAAAAA (A-rich) [41]. Other known common PAS motifs included UGUA, UAUA/AUAU, and UUUU in the upstream region of PAS, and UGUG/GUGU, UUUU, and GGGG in the downstream region of PAS [41].

PAS strength prediction

Nucleotide sequences flanking each PAS (120 nt upstream and 120 nt downstream) were extracted and used as the input for prediction of PAS strength by using the deep learning model-based method PolyaStrength [42]. We additionally normalized PolyaStrength scores by using percentile-based values. As such, each score reflects its percentile of all PAS (Main collection) in the genome.

PAS annotation by RefSeq

We used one representative RefSeq (Rep. RefSeq) transcript per gene to annoate PAS. The selection of Rep. RefSeq followed the following priority order: Matched Annotation from NCBI and EMBL-EBI (MANE) Select (human only) [43] > transcript with the largest genomic span > transcript type (NM > XM and NR > XR) > transcript with the longest sequence > number of exons. For protein-coding transcripts, 5′UTR, coding sequence (CDS), and 3′UTR were identified. For non-coding transcripts, 5′-most exon, internal exon, 3′-most exon, and intron were used for multi-exon genes. Additionally, the PAS located within genic regions (defined by RefSeq transcripts for each gene) but not covered by the Rep. RefSeq transcript were labeled as the upstream or downstream type, depending upon its location relative to the Rep. RefSeq transcript. If a PAS was assigned to a gene by LR-RNA-seq data (see below) but not by RefSeq data, the PAS was labeled as "LR Only."

Long-read RNA sequencing data processing

Human LR-RNA-seq data include ENCODE4 PacBio IsoSeq and GTEx v9 ONT cDNA datasets [44, 45], and mouse LR-RNA-seq data include PacBio and ONT datasets from ENCODE and GENCODE [44, 46]. Sequence alignments to genomes (hg38 and mm10) were conducted by using Minimap2 (v2.28) [47] with settings adjusted for each sequencing technology to accommodate its specific sequencing error rate: -k 15 -x splice:hg for PacBio reads and -k 14 -x splice:hq for ONT reads [47]. Aligned reads with MAPQ < 10 were discarded by using the SAMtools (v1.15.1) [48].

PAS annotation by LR-RNA-seq data

Filtered BAM files were processed for transcript isoform discovery and quantification by using IsoQuant (v3.5.0) [49]. LR-RNA-seq data from IsoQuant's read_assignments output file were used to annotate human and mouse PAS. Transcripts with isoform assignment types in the following types were removed: ambiguous, noninformative, inconsistent/inconsistent_ambiguous, aligned to unknown chromosome, or not specified for strandedness. The transcript end site (TES) for each LR sequence was identified and compared to PAS in PolyA_DB v4 (Max collection), allowing ±24-nt flexibility. The PAS matched with LR TES constituted the Main collection of PolyA_DB v4.

PAS typing by the 3′-most exon zone

For LR-RNA-seq data, we used a 3′-most exon zone (TEZ)-based PAS typing method, which allows more precise subtyping of APA isoform using the last exon of each LR-RNA-seq sequence. The TEZ of each gene was defined by clustering all last exons of RefSeq transcripts. PAS located upstream or downstream of TEZ are indicated.

PAS expression

To evaluate the expression level of each PAS, we used two metrics, i.e. percent of samples with expression (PSE) and averaged reads per million (AvgRPM) of PASS reads across samples (only those with expression were included for calculation). As such, the total mean expression level for a PAS across all samples can be derived from AvgRPM/PSE.

Database content

PAS collections based on 3′READS+ data

We used 364 human and 451 mouse 3′READS+ data sets in this study, representing all data in the public domain as well as some unpublished 3′READS+ data (Fig. 1A and Supplementary Table S1). 3′READS+ is a 3′ end-based RNA-seq method developed in our lab for PAS identification and quantification [21]. The human samples included primary cells, tissues, and cell lines, and the mouse samples included tissues and cell lines (Supplementary Table S1). Compared with polyA_DB v3.2 [50], polyA_DB 4 has a 3.4-fold increase in human sample number and a 6.4-fold increase in human PAS-supporting (PASS) read number; for the mouse genome, the increase in sample number and PASS read number are 1.8-fold and 3.1-fold, respectively, over PolyA_DB v3.2. In total, we identified 1 429 829 human PAS and 1 346 135 mouse PAS that were supported by at least three samples (Table 1 and Fig. 1B). These PAS display canonical nucleotide profiles in their flanking sequences (Supplementary Fig. 1A and C) [3].

PAS validation and annotation by LR-RNA-seq data

To validate the PAS using LR-RNA-seq data, we matched LR TES with PAS in PolyA_DB, allowing 24-nt flexibility in distance. In total, 20% of the Max collection PAS matched LR TES. Notably, ∼10%–20% of LR TES did not match our PAS, suggesting 3′ end artifacts caused plausibly by internal prim-

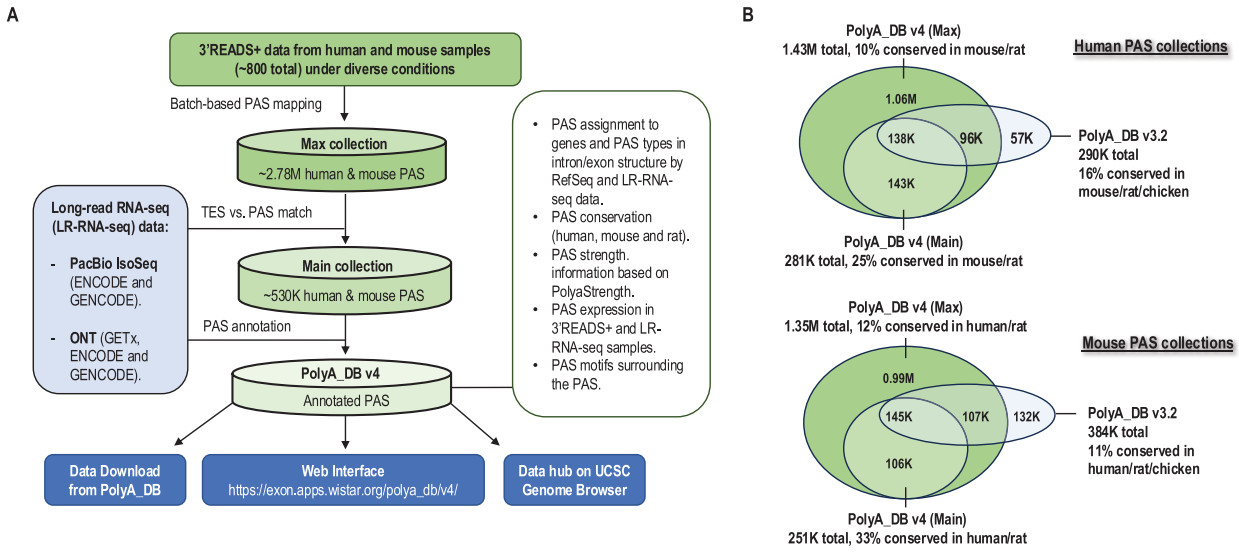


Figure 1. Overview of PolyA_DB v4. (**A**). Schematic of development of PolyA_DB v4 and its data flow. (**B**) Venn diagrams comparing PAS from PolyA_DB v4 (Max and Main collections shown separately) and PolyA_DB v3.2. Both human (top) and mouse (bottom) data are shown. The percentages of conserved PAS across human, mouse, and rat are indicated for different PAS groups.

Table 1. Summary of PolyA_DB v4

Species	Human	Mouse
No. of 3′READS+ samples used	364	451
No. of PAS-supporting 3′READS+ reads used	2 320 997 056	2 322 684 599
No. of LR-RNA-seq samples used	227	165
No. of PAS-supporting long reads used	347 441 037	318 856 685
No. of PAS in the Max collection	1 429 829	1 346 135
No. of conserved PAS in the Max collection	143 303	164 720
No. of PAS in the Main collection	281 091	251 177
No. of conserved PAS (Main collection)	70 139	84 094
No. of PAS assigned to gene (Main collection)	276 956	237 488
No. of protein-coding genes with PAS (Main collection)	17 575	18 091
No. of lncRNA genes with PAS (Main collection)	12 628	6872

ing or template switching (Supplementary Fig. 1B and D). Importantly, the LR-RNA-seq data connected many PAS in introns or in downstream regions of genes (or orphan PAS, ~15% of all PAS) to RefSeq-annotated exons. To assign PAS to genes, we applied two strategies (Figs 1A and 2A). The first method assigned a PAS to a gene through the Rep. RefSeq transcript if the PAS is within the genomic span of the transcript (Fig. 2A). Using this approach, we identified 236 069 gene-linked PAS and 45 022 intergenic PAS in the human genome, including 9220 PAS assigned to more than one gene. In mouse, we obtained 195 647 gene-linked PAS and 55 530 intergenic PAS, of which 2923 PAS were assigned to multiple genes. The second approach was based on all RefSeq transcripts supplemented with LR-RNA-seq data (Fig. 2A). If a PAS was matched with the TES of a LR, the LR sequence was used to find the corresponding gene. Specifically, we used 135 PacBio IsoSeq samples from 18 ENCODE tissue types and 92 Oxford Nanopore samples from 10 GTEx tissue types for humans (Supplementary Tables 2 and 3) and 165 PacBio and ONT samples for mice (Supplementary Tables 4 and 5). This strategy yielded 281 091 human and 251 177 mouse PAS supported by LR-RNA-seq data, which constiute the Main collection of PolyA_DB v4 (Fig. 1B and Table 1). Notably, by using the LR-RNA-seq data, many intergenic PAS defined by the Rep. RefSeq strategy were found to be part of a gene (Fig. 2B) and PAS belonging to genes nested within other genes (host genes) were distinguished from intronic PAS of the host genes.

PAS expression

PolyA_DB v4 catalogs PAS expression in several ways. First, the PSE and AvgRPM (only across samples with expression) were calculated by using all 3′READS+ samples. Second, we calculated PSE and AvgPRM by using LR-RNA-seq datasets. This multi-pronged approach could provide robust expression information on APA isoforms. We found that these expression metrics were well correlated (Fig. 3A and B), supporting the quantitativeness of both 3′READS+ data and LR-RNA-seq data.

PAS strength

To facilitate the evaluation of PAS strength, we used the deep-learning model-based program PolyaStrength [42] to generate a PAS strength score for each PAS and normalized the score using all PAS in the Main collection. Briefly, a PolyaStrength score was first generated for each PAS using its surrounding sequence (±120 nt). We then derived percentile values of all PAS in the Main collection. These percentile values were used to normalize all PAS in the Max collection. Notably, our normalized PAS strength scores were well correlated with PAS expression levels based on PSE (Fig. 3C). In addition, conserved PAS displayed higher normalized strength scores than non-conserved ones (Fig. 3D), indicating that conserved and highly expressed PAS tend to be surrounded by strong PAS motifs in cannonical configurations.

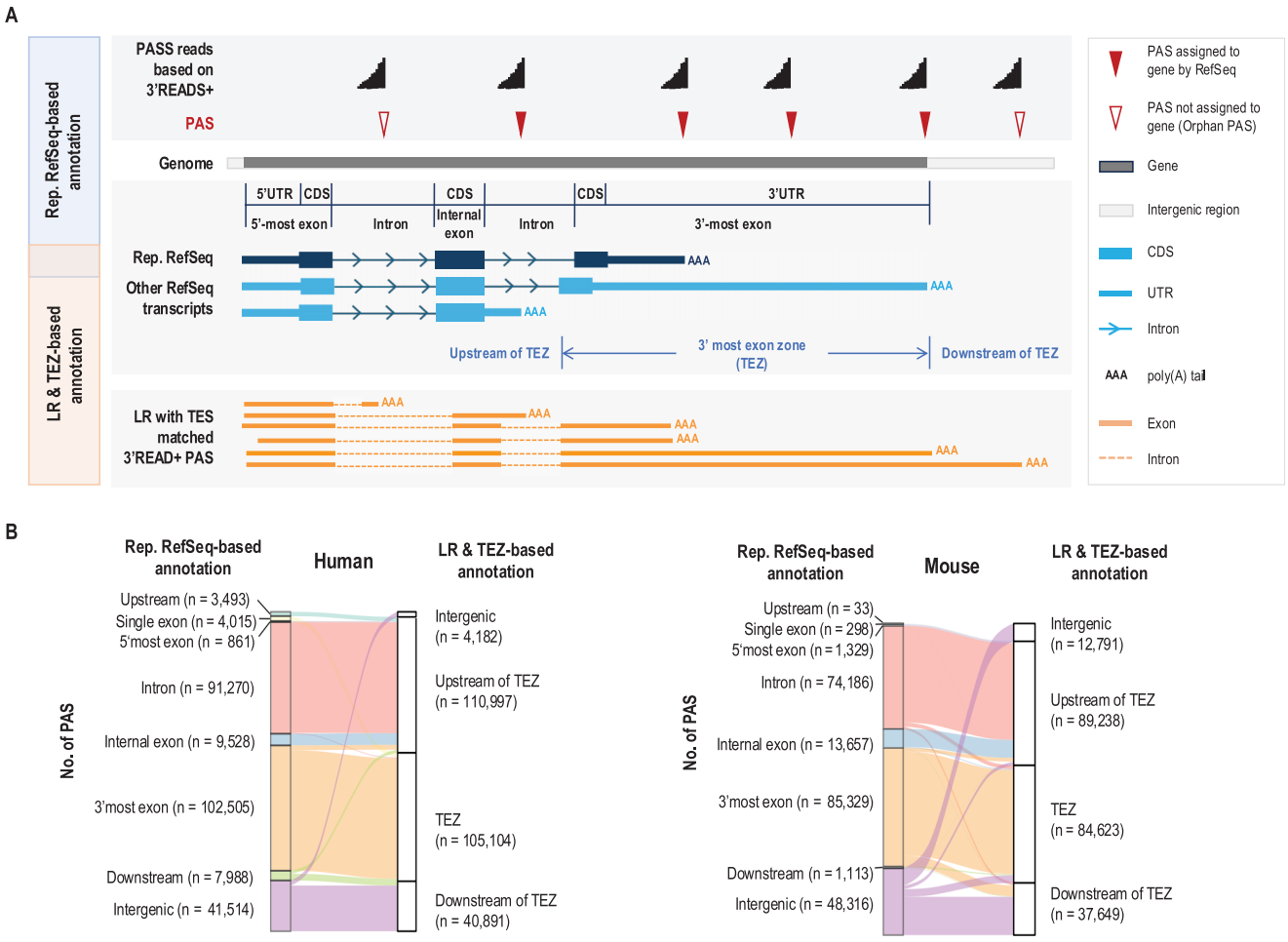


Figure 2. PAS annotation. (**A**) Schematic of PAS annotation. Gene span (gray) is based on all RefSeq transcripts. PAS annotation is based on representative RefSeq (Rep. RefSeq) transcript or the 3′-most exon zone (TEZ) composed of all last exons of RefSeq transcripts. The TEZ is used to annotate the last exon of each LR sequence. Orphan sites are PAS that cannot be connected with exonic regions defined by RefSeq transcripts. (**B**) Sankey plots showing PAS types (Main collection) between Rep. RefSeq-based annotations and those based on LR and TEZ. PAS numbers are indicated.

PAS motifs

We annotated potential, regulatory motifs in surrounding regions of each PAS, including UGUA, UUUU, and PAS hexamers in the upstream region and UGUG/GUGU, UUUU, and GGGG in the downstream region. The identified PAS motifs could provide users information about potentially functional motifs and their configurations around each PAS.

PAS collection comparisons

The Max collection contains all PAS identified by 3′READS+ datasets, whereas the Main collection represents a subset of Max that were validated by LR-RNA-seq. The two datasets differ greatly in sample sources and sequencing depths (Table 1 and Supplementary Tables 1–5). Therefore, whereas the Max collection is comprehensive, the Main collection is more selective. Consistently, compared to the Max collection, the Main collection exhibits more canonical nucleotide profiles in surrounding regions (Supplementary Figs. 1A and C) and higher expression levels (Supplementary Fig. 1E).

We found that ∼80% and ∼47% of human PAS in PolyA_DB v3.2 (identified by using 3′READ datasets) [50] were in the Max and Main collections of PolyA_DB v4, respectively (Fig. 1B). In addition, 42% of human PAS in PolyASite 2.0 [25] were present in PolyA_DB v4. About 130K PAS were shared by PolyA_DB v4 (Main collection) and PolyA_DB v3.2 or PolyASite 2.0, with an additional ∼100K shared between the Max collection and PolyA_DB v3.2 or PolyASite 2.0 (Fig. 1B and Supplementary Fig. 3A and B). Importantly, the PAS common to multiple PAS databases display more canonical nucleotide profiles in surrounding regions (Supplementary Fig. 3C and D) and tended to exhibit higher expression levels (Supplementary Figs 2 and 3E). Therefore, the PAS unique to each of the databases are generally lowly expressed and are less likely to have canonical surrounding motifs.

Data access and website interface

Data in PolyA_DB v4 are stored in a relational database that was implemented with SQLite. The interactive web interface was built with PHP (URL: https://exon.apps.wistar.org/polya_db/v4/). Queries can be made by using a RefSeq gene symbol or ID, or Ensembl gene ID. Four data tables are shown on the web page, as described below.

Gene summary table contains information about gene, including gene symbol, gene ID (RefSeq and Ensembl), gene

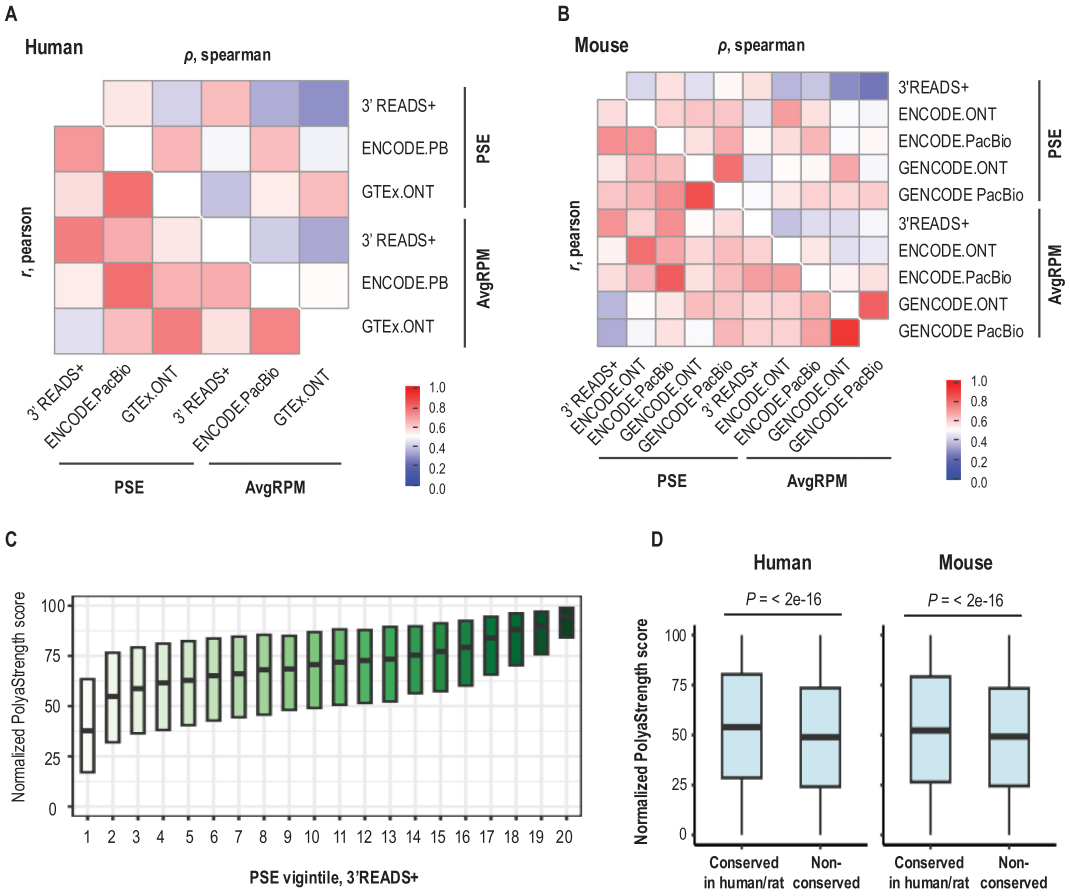


Figure 3. PAS expression, conservation and strength. Correlation of human (**A**) or mouse (**B**) PAS expression values, including PSE and averaged RPM (AvgRPM, $\log_2$) across 3′READS+ and LR-RNA-seq samples (ENCODE PacBio and GTEx ONT). Spearman correlation coefficients (ρ) and Pearson correlation coefficients (r) are shown in the upper and lower triangles, respectively, by using the color scheme indicated in the figure. (**C**) Boxplot showing the relationship between 3′READS+ PSE and normalized PolyaStrength scores. (**D**) Boxplots showing normalized PolyaStrength scores of PAS conserved in mammals and those not conserved.

name, gene type, genomic span based on all RefSeq transcripts, representative (Rep.) RefSeq transcript ID and its genomic span, the 3′ most exon zone (TEZ), last PAS based on RefSeq, and number of PAS identified for the gene. Orthologous genes (among human, moue, and rat) based on the HomoloGene database from NCBI are also shown. A UCSC Genome Browser [51] link is provided for visualization of the gene and its PAS.

PAS summary table lists all PAS assigned to a gene by using RefSeq and/or LR-RNA-seq data. For each PAS, its unique ID (PAS_ID) is based on genomic location, including chromosome, strand, and position. PAS type based on Rep. RefSeq is presented. For protein-coding genes, PAS types include 5′UTR, CDS, 3′UTR, and intron; for lncRNA genes, PAS types include single exon, or 5′-most exon, internal exon, intron, and 3′-most exon for multi-exon genes. PAS types based on both RefSeq and LR-RNA-seq include TEZ (inside of), upstream of TEZ, and downstream of TEZ. Also shown are PAS conservation (between human, mouse, and rat), PAS hexamer sequence, predicted PAS strength, and the number of associated genes. A UCSC Genome Browser [51] link is provided for each PAS.

In addition, PAS motif table displays upstream and downstream sequences for each PAS, respectively, with common motifs highlighted in bold. PAS expression table summarizes PAS expression based on both 3′READS+ and LR-RNA-seq data sets, using PSE and AvgRPM as metrics.

PolyA_DB data are also accessible via the UCSC Genome Browser [51] through a public hub (hyperlinked on PolyA_DB v4 homepage). Both Max and Main PAS collections can be visualized through UCSC Genome Browser [51]. PSE and AvgRPM are shown as peaks, and TEZ information is also displayed. Data download of tab-delimited data is available at https://exon.apps.wistar.org/polya_db/v4/misc/download.php.

Summary and future directions

Here, we present PolyA_DB v4, an updated version of PolyA_DB [52]. The new version is built upon comprehensive and accurate PAS identification by using 3′READS+ data from a large number of samples and thorough PAS annotation by using several public LR-RNA-seq datasets. The database vastly expands PAS collections by 4.9-fold for human and 3.5-fold for mouse over the previous version (v3.2). PASS reads are based solely on the 3′READS+ data, ensuring data consistency and minimizing false positives caused by internal priming and template switching. The database also includes PAS expression information based on both 3′READS+ and LR-RNA-seq data as well as information about PAS conservation across human, mouse, and rat. Putative motifs surrounding the PAS are annotated, and the PAS strength is predicted by using the PolyaStrength program [42]. In sum, PolyA_DB v4

provides a rich resource for PAS and APA isoform information in mammalian genomes.

Future development of the database would include additional species for greater understanding of PAS evolution, APA isoform expression across cells and tissues under various physiological and pathological conditions as well as in different human populations using bulk RNA-seq and scRNA-seq data, and in-depth annotations of APA isoforms in relation to transcriptional start sites and splicing events.

Acknowledgements

We thank Dinghai Zheng and Jihae Shin for generating some of the 3′READS+ data sets used in the database. Computational resources were provided by the Wistar Institute.

Author contributions: Shan Yu (Data curation [lead], Formal analysis [lead], Methodology [lead], Writing—original draft [equal]), Wei Chun Chen (Conceptualization [equal], Formal analysis [lead], Methodology [equal]), Luyang Wang (Conceptualization [supporting], Data curation [supporting], Formal analysis [equal], Methodology [equal]), San Jewell (Methodology [supporting]), Ayna Mammedova (Formal analysis [supporting]), Seong Woo Han (Investigation [supporting], Methodology [supporting]), Jayamanna Wickramasinghe (Methodology [supporting], Resources [supporting]), Yoseph Barash (Project administration [supporting], Resources [supporting]), and Bin Tian (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Funding acquisition [lead], Investigation [lead], Methodology [lead], Project administration [lead], Supervision [lead], Writing—original draft [lead], Writing—review & editing [lead]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the National Institutes of Health (R01GM084089 and R35GM153277 to B.T.; R01LM013437 and R01GM147739 to Y.B.) and National Science Foundation Cooperative Agreement (DBI-2400327 to Y.B.). Funding to pay the Open Access publication charges for this article was provided by National Institutes of Health (R35GM153277).

Data availability

PolyA_DB v4 data are freely available online at https://exon.apps.wistar.org/polya_db/v4/.

References

1. Boreikaitė V, Passmore LA. 3′-End processing of eukaryotic mRNA: machinery, regulation, and impact on gene expression. *Annu Rev Biochem* 2023;92:199–225. https://doi.org/10.1146/annurev-biochem-052521-012445
2. Shi Y, Manley JL. The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site. *Genes Dev* 2015;29:889–97. https://doi.org/10.1101/gad.261974.115
3. Gruber AJ, Zavolan M. Alternative cleavage and polyadenylation in health and disease. *Nat Rev Genet* 2019;20:599–614. https://doi.org/10.1038/s41576-019-0145-z
4. Mitschka S, Mayr C. Context-specific regulation and function of mRNA alternative polyadenylation. *Nat Rev Mol Cell Biol* 2022;23:779–96.
5. Tian B, Manley JL. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* 2017;18:18–30. https://doi.org/10.1038/nrm.2016.116
6. Kamieniarz-Gdula K, Proudfoot NJ. Transcriptional control by premature termination: a forgotten mechanism. *Trends Genet* 2019;35:553–64. https://doi.org/10.1016/j.tig.2019.05.005
7. Vilborg A, Sabath N, Wiesel Y *et al.* Comparative analysis reveals genomic features of stress-induced transcriptional readthrough. *Proc Nat Acad Sci USA* 2017;114:E8362–71.
8. Rodriguez-Molina JB, West S, Passmore LA. Knowing when to stop: transcription termination on protein-coding genes by eukaryotic RNAPII. *Mol Cell* 2023;83:404–15. https://doi.org/10.1016/j.molcel.2022.12.021
9. Morgan M, Shiekhattar R, Shilatifard A *et al.* It's a DoG-eat-DoG world-altered transcriptional mechanisms drive downstream-of-gene (DoG) transcript production. *Mol Cell* 2022;82:1981–91. https://doi.org/10.1016/j.molcel.2022.04.008
10. Grzechnik P, Mischo HE. Fateful decisions of where to cut the line: pathology associated with aberrant 3′ end processing and transcription termination. *J Mol Biol* 2025;437:168802. https://doi.org/10.1016/j.jmb.2024.168802
11. Nourse J, Spada S, Danckwardt S. Emerging Roles of RNA 3′-end Cleavage and Polyadenylation in Pathogenesis, Diagnosis and Therapy of Human Disorders. *Biomolecules* 2020;10:915. https://doi.org/10.3390/biom10060915.
12. Moll P, Ante M, Seitz A *et al.* QuantSeq 3′ mRNA sequencing for RNA quantification. *Nat Methods* 2014;11:i–iii. https://doi.org/10.1038/nmeth.f.376
13. Zheng GX, Terry JM, Belgrader P *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;8:14049. https://doi.org/10.1038/ncomms14049
14. Nam DK, Lee S, Zhou G *et al.* Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc Natl Acad Sci USA* 2002;99:6152–6. https://doi.org/10.1073/pnas.092140899
15. Svoboda M, Frost HR, Bosco G. Internal oligo(dT) priming introduces systematic bias in bulk and single-cell RNA sequencing count data. *NAR Genom Bioinform* 2022;4:lqac035. https://doi.org/10.1093/nargab/lqac035
16. Zhu YY, Machleder EM, Chenchik A *et al.* Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques* 2001;30:892–7. https://doi.org/10.2144/01304pf02
17. Aach J, Bulyk ML, Church GM *et al.* Computational comparison of two draft sequences of the human genome. *Nature* 2001;409:856–9. https://doi.org/10.1038/35057055
18. Workman RE, Tang AD, Tang PS *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods* 2019;16:1297–305. https://doi.org/10.1038/s41592-019-0617-2
19. Hoque M, Ji Z, Zheng D *et al.* Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nat Methods* 2013;10:133–9. https://doi.org/10.1038/nmeth.2288
20. Jan CH, Friedman RC, Ruby JG *et al.* Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. *Nature* 2011;469:97–101. https://doi.org/10.1038/nature09616
21. Zheng D, Liu X, Tian B. 3′READS+, a sensitive and accurate method for 3′ end sequencing of polyadenylated RNA. *RNA* 2016;22:1631–9. https://doi.org/10.1261/rna.057075.116
22. O'Leary NA, Wright MW, Brister JR *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion,

and functional annotation. *Nucleic Acids Res* 2016;44:D733–45. https://doi.org/10.1093/nar/gkv1189
23. Harrison PW, Amode MR, Austine-Orimoloye O *et al.* Ensembl 2024. *Nucleic Acids Res* 2024;52:D891–9. https://doi.org/10.1093/nar/gkad1049
24. Frankish A, Carbonell-Sala S, Diekhans M *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res* 2023;51:D942–9. https://doi.org/10.1093/nar/gkac1071
25. Herrmann CJ, Schmidt R, Kanitz A *et al.* PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Res* 2020;48:D174–9.
26. Moon Y, Herrmann CJ, Mironov A *et al.* PolyASite v3.0: a multi-species atlas of polyadenylation sites inferred from single-cell RNA-sequencing data. *Nucleic Acids Res* 2025;53:D197–204. https://doi.org/10.1093/nar/gkae1043
27. Zhu S, Lian Q, Ye W *et al.* scAPAdb: a comprehensive database of alternative polyadenylation at single-cell resolution. *Nucleic Acids Res* 2022;50:D365. https://doi.org/10.1093/nar/gkab795d370.
28. Zhang PH, Feng H, Ma XK *et al.* PASSpedia: a polyadenylation site database across different species at single cell resolution. *Genomics Proteomics Bioinform* 2025. https://doi.org/10.1093/gpbjnl/qzaf089
29. Hong W, Ruan H, Zhang Z *et al.* APAatlas: decoding alternative polyadenylation across human tissues. *Nucleic Acids Res* 2020;48:D34–9. https://doi.org/10.1093/nar/gkz876
30. Yang X, Tong Y, Liu G *et al.* scAPAatlas: an atlas of alternative polyadenylation across cell types in human and mouse. *Nucleic Acids Res* 2022;50:D356–64. https://doi.org/10.1093/nar/gkab917
31. Yang Y, Zhang Q, Miao YR *et al.* SNP2APA: a database for evaluating effects of genetic variants on alternative polyadenylation in human cancers. *Nucleic Acids Res* 2020;48:D226–32. https://doi.org/10.1093/nar/gkz793
32. Li L, Huang KL, Gao Y *et al.* An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nat Genet* 2021;53:994–1005. https://doi.org/10.1038/s41588-021-00864-5
33. Cui Y, Peng F, Wang D *et al.* 3′aQTL-atlas: an atlas of 3′UTR alternative polyadenylation quantitative trait loci across human normal tissues. *Nucleic Acids Res* 2022;50:D39–45. https://doi.org/10.1093/nar/gkab740
34. Ma X, Cheng S, Ding R *et al.* ipaQTL-atlas: an atlas of intronic polyadenylation quantitative trait loci across human tissues. *Nucleic Acids Res* 2023;51:D1046–52. https://doi.org/10.1093/nar/gkac736
35. Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics Proteomics Bioinform* 2015;13:278–89. https://doi.org/10.1016/j.gpb.2015.08.002
36. Wang Y, Zhao Y, Bollas A *et al.* Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* 2021;39:1348–65. https://doi.org/10.1038/s41587-021-01108-x
37. Alfonso-Gonzalez C, Legnini I, Holec S *et al.* Sites of transcription initiation drive mRNA isoform selection. *Cell* 2023;186:2438–55.e22. https://doi.org/10.1016/j.cell.2023.04.012
38. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* 2011;17:3.
39. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–9.
40. Tian B, Hu J, Zhang H *et al.* A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* 2005;33:201–12. https://doi.org/10.1093/nar/gki158
41. Tian B, Graber JH. Signals for pre-mRNA cleavage and polyadenylation. *WIREs RNA* 2012;3:385–96. https://doi.org/10.1002/wrna.116
42. Stroup EK, Ji Z. Deep learning of human polyadenylation sites at nucleotide resolution reveals molecular determinants of site usage and relevance in disease. *Nat Commun* 2023;14:7378. https://doi.org/10.1038/s41467-023-43266-3
43. Morales J, Pujar S, Loveland JE *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* 2022;604:310–5. https://doi.org/10.1038/s41586-022-04558-8
44. Reese F, Williams B, Balderrama-Gutierrez G *et al.* The ENCODE4 long-read RNA-seq collection reveals distinct classes of transcript structure diversity. bioRxiv, https://doi.org/10.1101/2023.05.15.540865. 16 May 2023, preprint: not peer reviewed.
45. Glinos DA, Garborcauskas G, Hoffman P *et al.* Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* 2022;608:353–9. https://doi.org/10.1038/s41586-022-05035-y
46. Kaur G, Perteghella T, Carbonell-Sala S *et al.* GENCODE: massively expanding the lncRNA catalog through capture long-read RNA sequencing. bioRxiv, https://doi.org/10.1101/2024.10.29.620654, 31 October 2024, preprint: not peer reviewed.
47. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;34:3094–100.
48. Li H, Handsaker B, Wysoker A *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25. 2078–79. https://doi.org/10.1093/bioinformatics/btp352
49. Prjibelski AD, Mikheenko A, Joglekar A *et al.* Accurate isoform discovery with IsoQuant using long reads. *Nat Biotechnol* 2023;41:915–8. https://doi.org/10.1038/s41587-022-01565-y
50. Wang R, Zheng D, Yehia G *et al.* A compendium of conserved cleavage and polyadenylation events in mammalian genes. *Genome Res* 2018;28:1427–41. https://doi.org/10.1101/gr.237826.118
51. Perez G, Barber GP, Benet-Pages A *et al.* The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res* 2025;53:D1243–9. https://doi.org/10.1093/nar/gkae974
52. Zhang H, Hu J, Recce M *et al.* PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res* 2005;33–D116–20. https://doi.org/10.1093/nar/gki055
53. Feng X, Li L, Wagner EJ *et al.* TC3A: the Cancer 3′ UTR Atlas. *Nucleic Acids Res* 2018;46:D1027–30. https://doi.org/10.1093/nar/gkx892
54. Jin W, Zhu Q, Yang Y *et al.* Animal-APAdb: a comprehensive animal alternative polyadenylation database. *Nucleic Acids Res* 2021;49:D47–54. https://doi.org/10.1093/nar/gkaa778
55. Marini F, Scherzinger D, Danckwardt S. TREND-DB-a transcriptome-wide atlas of the dynamic landscape of alternative polyadenylation. *Nucleic Acids Res* 2021;49:D243–53. https://doi.org/10.1093/nar/gkaa722

Received: September 16, 2025. **Revised:** October 15, 2025. **Accepted:** October 16, 2025