Published online 28 September 2020 *Nucleic Acids Research, 2021, Vol. 49, Database issue* ***D47–D54***
doi: 10.1093/nar/gkaa778

Animal-APAdb: a comprehensive animal alternative polyadenylation database

Weiwei Jin[1,†], Qizhao Zhu[2,†], Yanbo Yang[1], Wenqian Yang[1], Dongyang Wang[1], Jiajun Yang[1], Xiaohui Niu[1], Debing Yu[2,*] and Jing Gong[1,3,*]

[1]Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, P.R. China, [2]College of Animal Science and Technology, Nanjing Agricultural University, Nanjing 210095, P.R. China and [3]College of Biomedicine and Health, Huazhong Agricultural University, Wuhan 430070, P.R. China

Received August 01, 2020; Revised August 27, 2020; Editorial Decision September 05, 2020; Accepted September 08, 2020

ABSTRACT

Alternative polyadenylation (APA) is an important post-transcriptional regulatory mechanism that recognizes different polyadenylation signals on transcripts, resulting in transcripts with different lengths of 3′ untranslated regions and thereby influencing a series of biological processes. Recent studies have highlighted the important roles of APA in human. However, APA profiles in other animals have not been fully recognized, and there is no database that provides comprehensive APA information for other animals except human. Here, by using the RNA sequencing data collected from public databases, we systematically characterized the APA profiles in 9244 samples of 18 species. In total, we identified 342 952 APA events with a median of 17 020 per species using the DaPars2 algorithm, and 315 691 APA events with a median of 17 953 per species using the QAPA algorithm in these 18 species, respectively. In addition, we predicted the polyadenylation sites (PAS) and motifs near PAS of these species. We further developed Animal-APAdb, a user-friendly database (http://gong_lab.hzau.edu.cn/Animal-APAdb/) for data searching, browsing and downloading. With comprehensive information of APA events in different tissues of different species, Animal-APAdb may greatly facilitate the exploration of animal APA patterns and novel mechanisms, gene expression regulation and APA evolution across tissues and species.

INTRODUCTION

Alternative polyadenylation (APA) is a widespread mechanism that contributes to the generation of transcript isoforms with different lengths of 3′ untranslated regions (3′UTR) by recognizing different polyadenylation signals (1), which may cause the alteration of some important regulatory elements, such as miRNA binding sites and RNA protein binding sites, thus affecting mRNA stability, localization and translation (2,3). It has been revealed that approximately 70% of eukaryotic genes possess multiple functional polyadenylation sites (PAS) (3–6) and nearly half of genes in fruitfly (7), worm (8) and zebrafish (9) undergo APA. APA-mediated gene regulation functions in a tissue-specific (3,10), and cell-specific manner (11,12). For example, brain and neuronal cells tend to have longer 3′UTRs than testis and ovary cells (13,14). Global 3′UTR shortening has been found in proliferating cells, cancer cells and tumor samples (13,15–17), whereas 3′UTR lengthening is associated with embryonic differentiation (16) and animal neurogenesis (18). Recent studies have highlighted the important roles of APA in human. Several APA dysregulations have been identified in human diseases (6–9), such as diabetic nephropathy, systemic lupus erythematosus and muscular dystrophy (19). However, the scope for gene regulation at the level of cleavage and polyadenylation in other animals except human has not been well recognized.

Several methods have been developed to identify PAS and quantify APA events (1,20–23). Compared with early APA identification methods based on complementary DNAs, expressed sequence tags and 3′-sequencing data, which can only detect limited APA events, RNA sequencing (RNA-seq) has become an alternative technology for detecting APA events at the genome level (24–26). Accordingly, several algorithms have been developed for the identification of APA events from RNA-seq data, either based on de novo identification algorithms including IsoSCM (27), DaPars (15,28), APAtrap (29) and TAPAS (30) or annotation-based algorithms such as MISO (31), roar (32) and QAPA (33). In human, TC3A (34) and APAatlas (24) databases systematically characterize APA events in different tissues using a large amount of RNA-seq data from The Cancer

*To whom correspondence should be addressed. Tel: +86 25 8439 5036; Fax: +86 25 8439 5036; Email: yudebing@njau.edu.cn
Correspondence may also be addressed to Jing Gong. Tel: +86 27 8728 5085; Fax: +86 27 8728 5085; Email: gong.jing@mail.hzau.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Genome Atlas and Genotype-Tissue Expression project, respectively. However, there is no database that provides comprehensive APA information for other animals except human in a large number of tissues.

In this study, we systematically characterized APA profiles in 9244 samples of 18 species using the RNA-seq data collected from public databases. These species include baboon, chicken, chimp, clawed frog, cow, crab-eating macaque, dog, fruitfly, green monkey, horse, mouse, pig, rabbit, rat, rhesus, sheep, worm and zebrafish. In addition, we predicted the PAS and motifs near PAS (APA motifs) of these species. Finally, we further developed Animal-APAdb (http://gong_lab.hzau.edu.cn/Animal-APAdb/), a user-friendly database for the browsing, searching and downloading of APA-related information.

MATERIALS AND METHODS

Collection and processing of RNA-seq data

To obtain a complete list of RNA-seq data of other animals except human, we conducted a comprehensive search from the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) (35,36) of the National Center for Biotechnology Information. The following search terms were used for SRA searching on 10 December 2019: ('cdna'[Selection]) AND 'transcriptomic'[Source]) AND ('rna seq'[Strategy]) NOT ('human'[Organism]) NOT ('single cell'[Text Word]), and a total of 443 318 records were obtained. Because certain files are required in the quantification of APA events, including gene_bed.file, gene_symbol_file, ensembl_identifiers.txt, gencode.basic.txt, genome.fa and genome.annotation.gtf from UCSC (https://genome.ucsc.edu/) (37) and Ensembl (http://www.ensembl.org) (38), the candidate species were screened based on the availability of the required files and the ranking of their sample sizes. As a result, a total of 18 species were selected for further study. Then, we extracted all BioProjects from the sample list and checked each description manually. Finally, 106 BioProjects of normal tissues were retained. The non-repetitive raw RNA-seq data from these BioProjects were downloaded, converted into standard fastq, subjected to quality control using FastQC (version: v0.11.8), cleaned with Trim Galore (version: 0.6.4_dev) and then aligned to the corresponding reference genome using HISAT2 (39). Subsequently, samples and BioProjects with low mapping rates were discarded, and finally 9244 samples of 97 BioProjects were retained (Figure 1A).

Identification of PAS and PAS cluster

Recent studies have demonstrated the possibility of using *de novo* algorithms to identify novel PAS based on RNA-seq data (15,28). Here, we used the well-established de novo algorithm DaPars2 (15) to identify the alternative proximal PAS within each sample. Based on the two-PAS model, DaPars2 applies a linear regression model to infer the location of the APA site within the 3′UTR region. Considering that the position of PAS predicted by DaPars2 might be inconsistent among different samples, the sites were grouped into a cluster based on the principle of the site position distance $\leq$24 nt (Figure 1B) (40,41). For a gene, the median position of a PAS cluster is usually the most representative site among samples, so the median site was defined as the PAS.

Identification of alternative polyadenylation

In this study, we utilized two popular algorithms, DaPars2 and QAPA, to quantify APA events from standard RNA-seq data. DaPars2 only predicts single proximal site, and the end of 3′UTR was taken as the distal site by default, so we used the percentage of the distal poly(A) site usage index (PDUI) to quantify APA events. PDUI value was a novel, intuitive ratio for quantifying APA events based on RNA-seq data (28), which was calculated by the expression level of isoform with the distal poly(A) site, divided by the total expression level of isoforms with both distal and proximal poly(A) sites. To reduce false positives, we discarded the PDUIs of certain transcripts for which the coverage of the last exon $<30\times$ or the percentage of samples supporting this PAS cluster (SampleP) $< 5\%$ (Figure 1B) (24,28). For QAPA, which is based on transcript-level abundance, it can calculate the relative proportion of each isoform in a gene using the PAS annotation files from GENCODE basic poly(A) annotation track, PolyASite (42) and/or custom file, so we used Poly(A) Usage (PAU) to quantify APA events. Due to the lack of PAS annotation files for most animals, we first created PAS annotation files based on the PAS extracted from DaPars2 results with the SampleP $\geq$ 5% (Figure 1B). Since mouse and worm have PAS annotation files in PolyASite database, these PAS annotation files were merged with our PAS annotation files for QAPA calculation.

Identification of APA motifs

Polyadenylation is the result of an RNA processing reaction. In the polyadenylation process, a multiprotein complex assembles on specific sequences of the pre-mRNAs, which are called the cleavage and polyadenylation signals (pA signals) (43). pA signals are composed of sequences that flank either side of where the pre-mRNA is endonucleolytically cleaved and subsequently polyadenylated (43). The classic pA signal is a bipartite sequence element that usually consists of a PAS hexamer, as well as upstream and downstream motifs of the cleavage site. In this study, we scanned the 50 nt (1,26) upstream sequence of the PAS to find PAS hexamers by DREME (44). In addition, for each PAS, motifs respectively at 200 nt upstream and downstream (1) from the PAS were obtained using MEME (45). Motifs were further filtered based on the following conditions: the statistical significance of the motif (*E*-value) $>$ 0.05, the percentage of sites contributing to the construction of the motif (CountP) $< 5\%$ for MEME, or the percentage of sequences matching the motif (CountP) $< 5\%$ for DREME (Figure 1B).

IMPLEMENTATION

Animal-APAdb (http://gong_lab.hzau.edu.cn/Animal-APAdb/) was built based on the THINKPHP (version 5.0.24) framework and Bootstrap 4, and runs on the

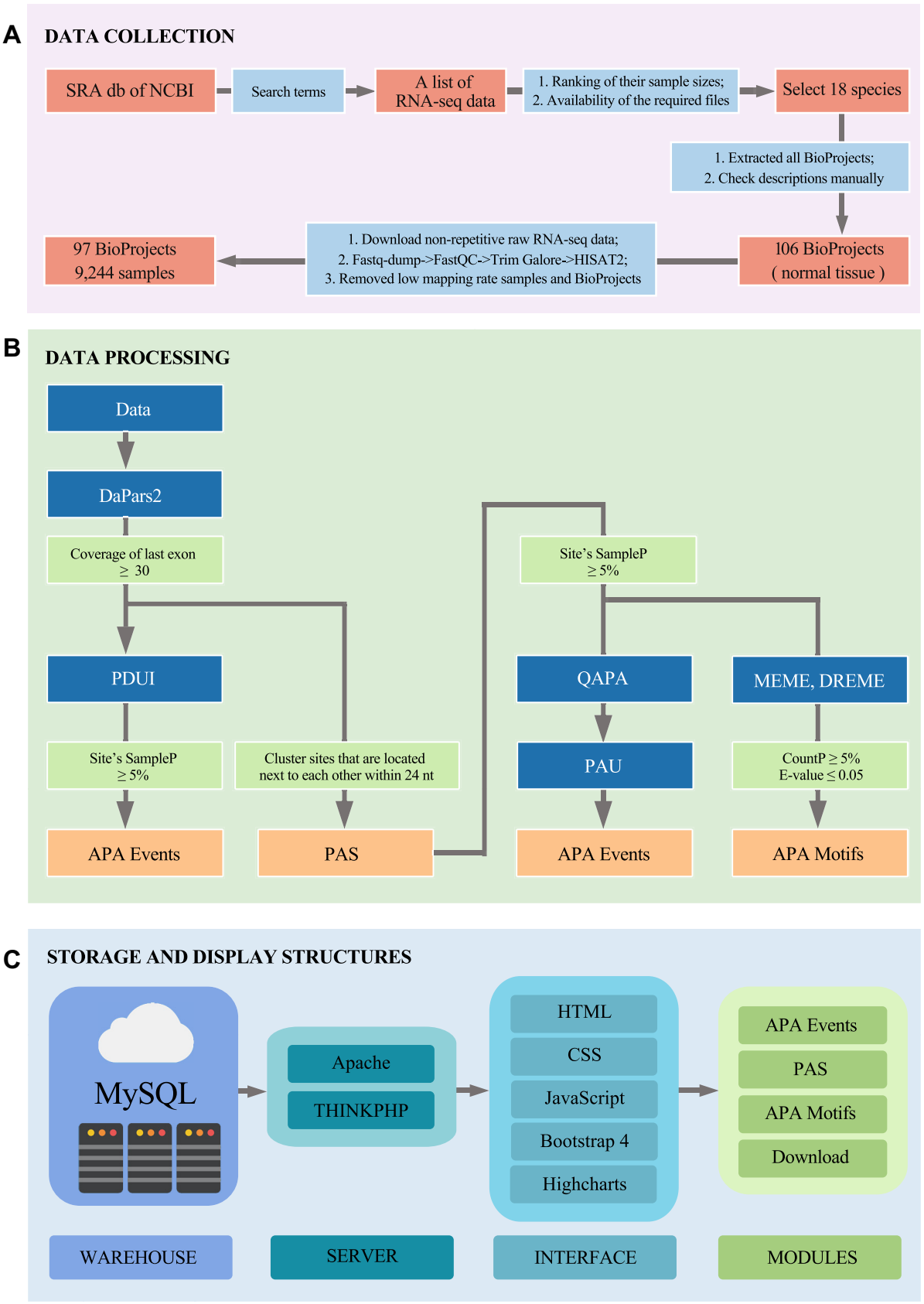


Figure 1. Flow chart of Animal-APAdb. (**A**) Data collection. (**B**) Data processing. (**C**) Storage and display structures of Animal-APAdb.

Table 1. Data summary in Animal-APAdb

Species	No. of samples	APA events identified by DaPars2	APA events identified by QAPA	Identified PAS	Genes with multiple PAS (%)
Papio anubis (Baboon)	766	2657	2401	11 694	21.73
Gallus gallus (Chicken)	656	25 600	20 680	72 508	60.06
Pan troglodytes (Chimp)	262	11 524	14 447	41 063	31.66
Xenopus tropicalis (Clawed frog)	284	19 782	18 164	49 648	59.07
Bos taurus (Cow)	838	17 203	17 741	68 535	55.41
Macaca fascicularis (Crab-eating macaque)	87	29 269	26 956	60 990	54.83
Canis lupus familiaris (Dog)	292	16 837	14 066	55 969	47.21
Drosophila melanogaster (Fruitfly)	774	7332	8572	34 261	52.12
Chlorocebus sabaeus (Green monkey)	327	13 922	14 645	43 972	41.18
Equus caballus (Horse)	160	11 149	7186	36 641	44.59
Mus musculus (Mouse)	1235	54 448	53 710	166 132	43.26
Sus scrofa (Pig)	819	36 280	24 441	160 005	61.41
Oryctolagus cuniculus (Rabbit)	338	7687	7442	22 165	40.88
Rattus norvegicus (Rat)	901	19 605	20 378	74 092	47.71
Macaca mulatta (Rhesus)	257	29 138	20 625	82 352	39.65
Ovis aries (Sheep)	730	7029	4189	26 652	32.27
Caenorhabditis elegans (Worm)	319	17 218	18 459	28 830	24.79
Danio rerio (Zebrafish)	199	16 272	21 589	46 991	40.31
Sum	9244	342 952	315 691	1 082 500	-
Max	1235	54 448	53 710	166 132	61.41
Min	87	2657	2401	11 694	21.73
Median	333	17 020	17 953	48 320	43.93

Apache 2 web server with MySQL (version 5.7.29) as its database engine and Highcharts for graph drawing (Figure 1C). Animal-APAdb is available online without registration and optimized for Chrome (recommended), Internet Explorer, Opera, Firefox, Windows Edge and macOS Safari.

DATABASE CONTENT AND USAGE

Samples of 18 species in Animal-APAdb

In total, 9244 samples of 18 species were analyzed in Animal-APAdb, ranging from 87 samples in Crab-eating macaque to 1235 samples in mouse (Table 1). The detailed information, including the number of samples per species, reference genome versions and the number of APA events, is available on the 'Document' page. The sample information of each species is presented in the 'BioProjects of each species' module on the 'Document' page, including species, the ID of BioProject, library layout, sample size and breed.

APA events in Animal-APAdb

Considering that *de novo* identification may introduce some false positives, part of the results was filtered as aforementioned. Finally, we identified a total of 342 952 APA events (median: 17 020 per species) using the DaPars2 algorithm, and 315 691 APA events (median: 17 953 per species) using the QAPA algorithm in these 18 species, respectively. The summary of these APA events is shown in 'APA event summary' module on the 'Document' page and Table 1.

PAS in Animal-APAdb

By using DaPars2, we identified a total of 1 082 500 PAS in these species, ranging from 11 694 in baboon to 166 132 in mouse. About 44% genes have multiple PAS, ranging from 22% in baboon to 61% in pig. We found that the 3′UTR length of genes (median: 773 nt) with multiple PAS is obviously longer than that of genes (median: 149 nt) with single PAS among all species (Figure 2A). We then calculated the number of occurrences of classic polyadenylation signal AATAAA and its 1 nt variants at upstream 50 nt from PAS (1,26,46), and found that about 18% PAS having these signals, which is similar to the percentage of 15% reported in another study (26).

APA motifs in Animal-APAdb

By using the MEME, DREME tool and the threshold value mentioned above, we obtained a total of 336 valid motifs, including 154 PAS hexamers, 90 motifs at 200 nt upstream, 92 motifs at 200 nt downstream. Among these PAS hexamers, the most frequent motifs are GGAGGA and TGTAAA, which are presented in 11 species, followed by GGAAGA, TGTATA and AGAAGA. It is actually difficult to determine the differences or similarities between motifs generated from MEME tool due to their different lengths. However, some similar short sequences could still be found from motifs in different species here, such as GAGGAAGA, CTGCTG and their variants at upstream 200 nt (Figure 2B), and A-rich sequence, CTGCAG and their variants at downstream 200 nt (Figure 2C).

Web interface

Animal-APAdb provides a user-friendly web interface. Four main modules, including 'APA Events', 'PAS', 'APA Motifs' and 'Download' (Figure 3A), are provided for the users to query APA events of genes in the tissues of certain species, retrieve PAS in the gene/genomic region of interests, browse probable APA motifs and download corresponding datasets.

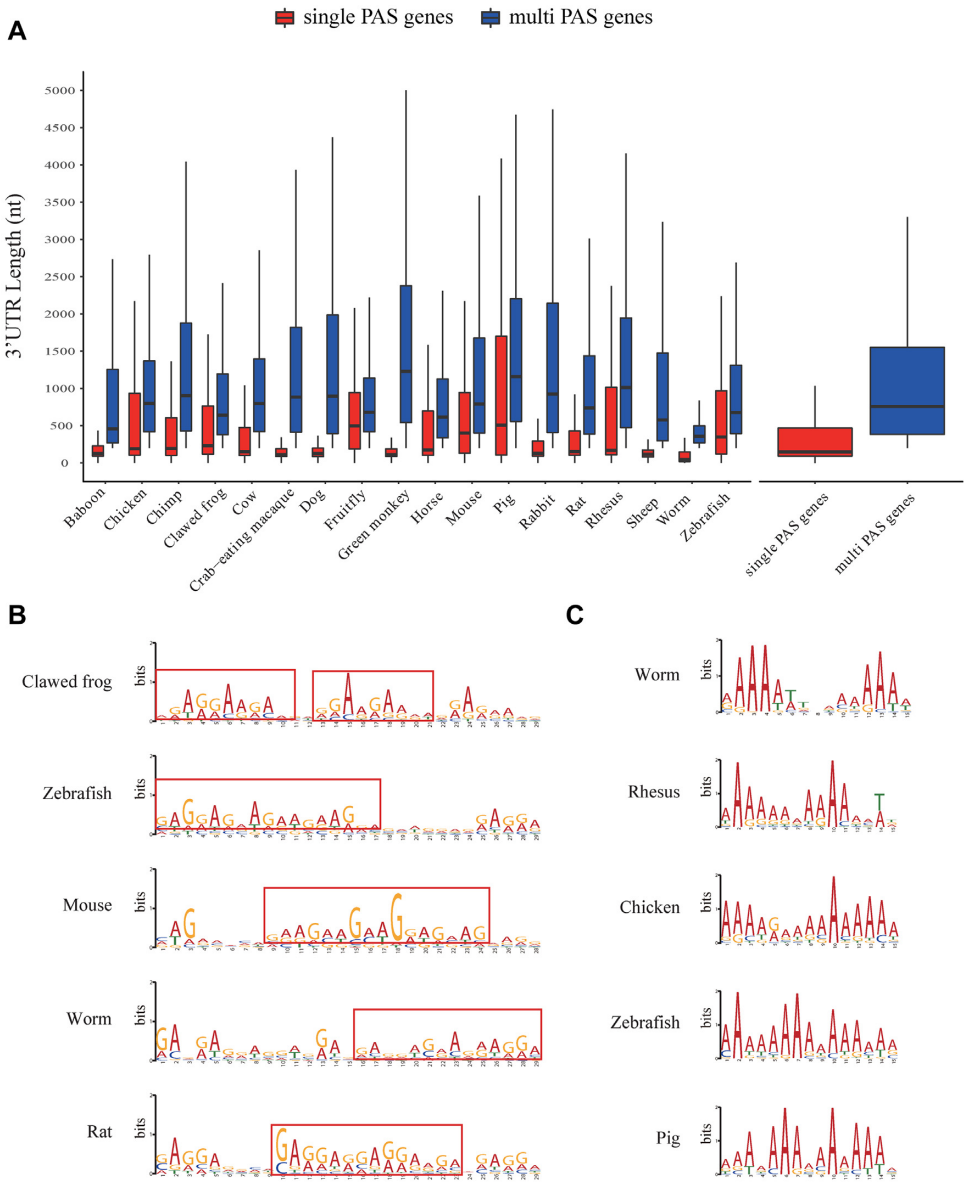


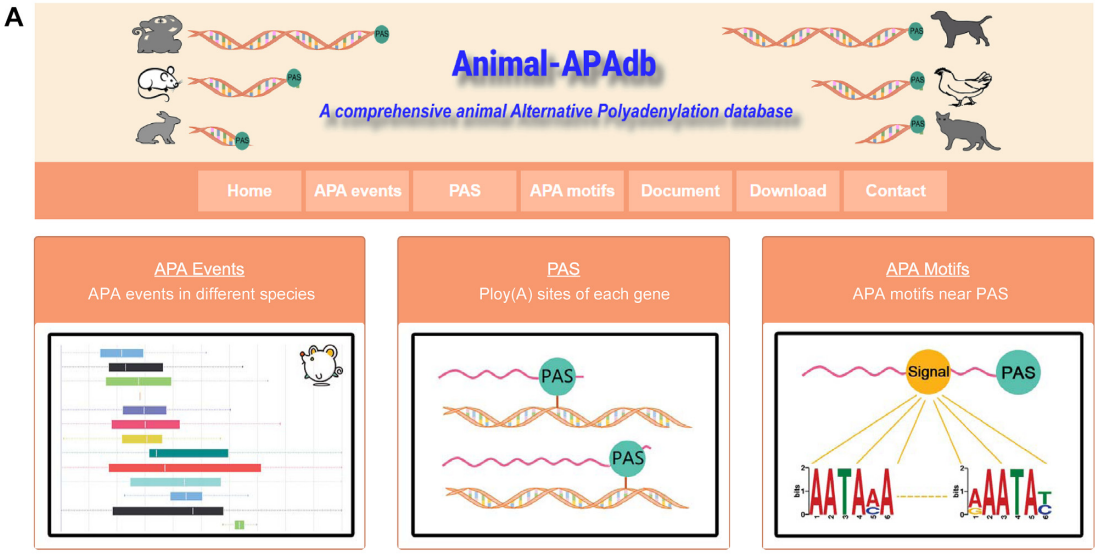
Figure 2. Some results of PAS and APA motifs. (**A**) 3′UTR length differences between single PAS genes and multi PAS genes. (**B**) A case of motifs at upstream 200 nt. (**C**) A case of motifs at downstream 200 nt.

On the 'APA Events' page, the users can query APA events by selecting an algorithm, species and tissue and typing a gene symbol or Ensembl gene ID in the search box. A table with the species, tissue, gene symbol, Ensembl ID and Ensembl Transcript ID of the queried APA events will be shown (Figure 3B). Then, by clicking the 'Plot' button, the users can view the position graph (Figure 3C) including the range of 3′UTR of the gene, the position of PAS and a box-plot graph of APA events (Figure 3D). It is worth noting that QAPA can calculate the usage of multiple sites (the distal site may be different from the end of 3′UTR) by PAS annotation file. Hence the users need to click the point on the position graph to retrieve the box-plot graph if they selected QAPA algorithm.

On the 'PAS' page, the users can select a species and input a genomic region (e.g. chr1:1–2000000:+), gene symbol or Ensembl ID to query the PAS clusters. Then, a table will be presented to provide details of the cluster with gene symbol, Ensembl ID, site ID, 3′UTR, PAS cluster, all PAS in the cluster (PAS ClusterS), PAS, the percentage (SampleP) and number (SampleS) of samples that support this PAS cluster and signals (Figure 3E). The users can click the 'Download' button to download the queried data, or click the '?' button for more information.

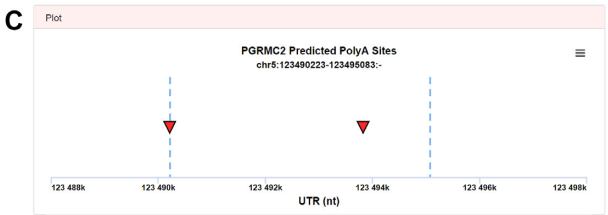
On the 'APA Motifs' page, when the users select the species and motif location, a table with species, motif location, motif, CountP and *E*-value will be provided, and more detailed reports can be obtained by clicking the 'More Detail' button (Figure 3F).

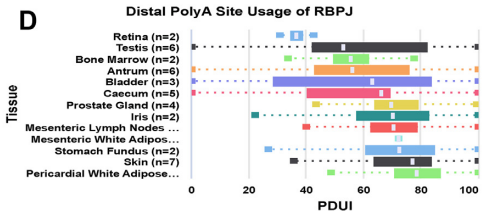
In Animal-APAdb, the main datasets of tissues for each species can be freely available from the 'Download' page. The 'Document' page provides the sample information, ref-



B

Species	Tissue	Gene Symbol	Ensembl Id	Ensembl Trans Id	Plot	Data
Baboon	AllTissue	PGRMC2	ENSPANG00000017275	ENSPANT00000000557		





E

Gene Symbol	Ensembl ID	Site ID	3'UTR	PAS Cluster	PAS ClusterS	PAS	SampleP	SampleS	Signals
TTLL10	ENSPANG00000012551	chr1|162956|+	162805- 163220	162956-162956	details	162956	0.13	1	NA
TTLL10	ENSPANG00000012551	chr1|163256|+	162805- 163403	163256-163256	details	163256	0.13	1	NA
CPTP	ENSPANG00000000592	chr1|288715|+	288564- 289086	288715-288715	details	288715	1.44	11	NA
VWA1	ENSPANG00000006170	chr1|402374|+	402223- 404848	402374-402374	details	402374	0.26	2	NA
VWA1	ENSPANG00000006170	chr1|402490|+	402223- 404848	402480-402497	details	402490	0.52	4	NA
VWA1	ENSPANG00000006170	chr1|403888|+	402223- 404848	403882-403897	details	403888	0.91	7	NA
MIB2	ENSPANG00000018509	chr1|578429|+	578278- 578516	578429-578429	details	578429	0.52	4	NA
GABRD	ENSPANG00000025625	chr1|949406|+	949255- 949554	949406-949406	details	949406	2.87	22	NA
GABRD	ENSPANG00000025625	chr1|949406|+	949255- 949478	949406-949406	details	949406	2.87	22	NA
GABRD	ENSPANG00000025625	chr1|949921|+	949255- 952230	949897-949957	details	949921	3.13	24	NA

F

Species	Motif Location	Motif	CountP	E-value
Baboon	PAS hexamer		15.08	7.0E-8
Baboon	PAS hexamer		8.46	0.00088
Baboon	PAS hexamer		7.44	0.00089
Baboon	PAS hexamer		6.66	0.00019

Figure 3. Overview of the Animal-APAdb. (**A**) The main functions in Animal-APAdb, including 'APA Events', 'PAS', 'APA Motifs' and 'Download' modules. (**B**) A table with species, tissue, gene symbol, Ensembl ID and Ensembl Trans ID of queried APA events. (**C**) The PAS graph of the queried gene. (**D**) The box-plot graph of APA events of the queried gene. (**E**) An example of search results in the 'PAS' module. (**F**) A case in the 'APA Motifs' module.

erence genome versions, APA event summary, pipeline of database construction and some other information. Besides, Animal-APAdb welcomes any feedback with email address provided on the 'Contact' page.

SUMMARY AND FUTURE DIRECTIONS

Great progress has been achieved in animal genome research in recent decades. Several animal-related databases, such as AnimalQTLdb (47) and Animal-ImputeDB (48), have been widely used by researchers. However, there are still big gaps in the research on the mechanisms and functions of APA in other animals except human. In this study, we developed the Animal-APAdb by collecting public available data, and provided comprehensive APA information of different tissues in 18 species. To the best of our knowledge, Animal-APAdb is the largest and most comprehensive animal APA database to date. In this version of Animal-APAdb, by using the data of 9244 samples, numerous PAS in multiple species are provided, and large amounts of APA events in different tissues and probable APA motifs are identified. In the future, we will integrate more samples and species into Animal-APAdb and continue to update the database. With comprehensive APA information in various tissues of different species, we believe that Animal-APAdb will be useful for uncovering animal APA patterns and novel mechanisms, gene expression regulation and APA evolution across tissues and species.

FUNDING

National Natural Science Foundation of China [31970644 to J.G.]; Huazhong Agricultural University Scientific & Technological Self-innovation Foundation [11041810351 to J.G.]; Jiangsu Agricultural Science and Technology Independent Innovation Fund [CX (17) 3014 to D.B.Y.]; Fundamental Research Funds for the Central University (Huazhong Agricultural University) [2662017JC048 to X.H.N.]. Funding for open access charge: Jiangsu Agricultural Science and Technology Independent Innovation Fund [CX (17) 3014 to D.B.Y.].
Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Elkon,R., Ugalde,A.P. and Agami,R. (2013) Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.*, **14**, 496–506.
3. Tian,B. and Manley,J.L. (2017) Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.*, **18**, 18–30.
4. Hoque,M., Ji,Z., Zheng,D., Luo,W., Li,W., You,B., Park,J.Y., Yehia,G. and Tian,B. (2013) Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nat. Methods*, **10**, 133–139.
5. Wu,X. and Bartel,D.P. (2017) Widespread influence of 3′-end structures on mammalian mRNA processing and stability. *Cell*, **169**, 905–917.
6. Mayr,C. (2016) Evolution and biological roles of alternative 3′UTRs. *Trends Cell Biol.*, **26**, 227–237.
7. 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.
8. Jan,C.H., Friedman,R.C., Ruby,J.G. and Bartel,D.P. (2011) Formation, regulation and evolution of Caenorhabditis elegans 3′UTRs. *Nature*, **469**, 97–101.
9. Ulitsky,I., Shkumatava,A., Jan,C.H., Subtelny,A.O., Koppstein,D., Bell,G.W., Sive,H. and Bartel,D.P. (2012) Extensive alternative polyadenylation during zebrafish development. *Genome Res.*, **22**, 2054–2066.
10. 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.
11. MacDonald,C.C. (2019) Tissue-specific mechanisms of alternative polyadenylation: testis, brain, and beyond (2018 update). *Wiley Interdiscip. Rev. RNA*, **10**, e1526.
12. Di Giammartino,D.C., Nishida,K. and Manley,J.L. (2011) Mechanisms and consequences of alternative polyadenylation. *Mol. Cell*, **43**, 853–866.
13. 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.
14. Guvenek,A. and Tian,B. (2018) Analysis of alternative cleavage and polyadenylation in mature and differentiating neurons using RNA-seq data. *Quant. Biol.*, **6**, 253–266.
15. Xiang,Y., Ye,Y., Lou,Y., Yang,Y., Cai,C., Zhang,Z., Mills,T., Chen,N.Y., Kim,Y., Muge Ozguc,F. *et al.* (2018) Comprehensive characterization of alternative polyadenylation in human cancer. *J. Natl. Cancer Inst.*, **110**, 379–389.
16. Ji,Z., Lee,J.Y., Pan,Z., Jiang,B. and Tian,B. (2009) Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 7028–7033.
17. Mayr,C. and Bartel,D.P. (2009) Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, **138**, 673–684.
18. Miura,P., Shenker,S., Andreu-Agullo,C., Westholm,J.O. and Lai,E.C. (2013) Widespread and extensive lengthening of 3′ UTRs in the mammalian brain. *Genome Res.*, **23**, 812–825.
19. Chang,J.W., Yeh,H.S. and Yong,J. (2017) Alternative polyadenylation in human diseases. *Endocrinol. Metab.*, **32**, 413–421.
20. 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.
21. You,L., Wu,J., Feng,Y., Fu,Y., Guo,Y., Long,L., Zhang,H., Luan,Y., Tian,P., Chen,L. *et al.* (2015) APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucleic Acids Res.*, **43**, D59–D67.
22. Zhang,H., Hu,J., Recce,M. and Tian,B. (2005) PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res.*, **33**, D116–D120.
23. Lee,J.Y., Yeh,I., Park,J.Y. and Tian,B. (2007) PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res.*, **35**, D165–D168.
24. Hong,W., Ruan,H., Zhang,Z., Ye,Y., Liu,Y., Li,S., Jing,Y., Zhang,H., Diao,L., Liang,H. *et al.* (2020) APAatlas: decoding alternative polyadenylation across human tissues. *Nucleic Acids Res.*, **48**, D34–D39.
25. Bonfert,T. and Friedel,C.C. (2017) Prediction of Poly(A) sites by Poly(A) read mapping. *PLoS One*, **12**, e0170914.
26. Chen,M., Ji,G., Fu,H., Lin,Q., Ye,C., Ye,W., Su,Y. and Wu,X. (2019) A survey on identification and quantification of alternative polyadenylation sites from RNA-seq data. *Brief. Bioinform.*, **21**, 1261–1276.
27. Shenker,S., Miura,P., Sanfilippo,P. and Lai,E.C. (2015) IsoSCM: improved and alternative 3′ UTR annotation using multiple change-point inference. *RNA*, **21**, 14–27.
28. Xia,Z., Donehower,L.A., Cooper,T.A., Neilson,J.R., Wheeler,D.A., Wagner,E.J. and Li,W. (2014) Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nat. Commun.*, **5**, 5274.

29. Ye,C., Long,Y., Ji,G., Li,Q.Q. and Wu,X. (2018) APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics*, **34**, 1841–1849.
30. Arefeen,A., Liu,J., Xiao,X. and Jiang,T. (2018) TAPAS: tool for alternative polyadenylation site analysis. *Bioinformatics*, **34**, 2521–2529.
31. Katz,Y., Wang,E.T., Airoldi,E.M. and Burge,C.B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**, 1009–1015.
32. Grassi,E., Mariella,E., Lembo,A., Molineris,I. and Provero,P. (2016) Roar: detecting alternative polyadenylation with standard mRNA sequencing libraries. *BMC Bioinformatics*, **17**, 423.
33. Ha,K.C.H., Blencowe,B.J. and Morris,Q. (2018) QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol.*, **19**, 45.
34. Feng,X., Li,L., Wagner,E.J. and Li,W. (2018) TC3A: the Cancer 3′ UTR Atlas. *Nucleic Acids Res.*, **46**, D1027–D1030.
35. Kodama,Y., Shumway,M., Leinonen,R. and International Nucleotide Sequence Database, C. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
36. Sayers,E.W., Beck,J., Brister,J.R., Bolton,E.E., Canese,K., Comeau,D.C., Funk,K., Ketter,A., Kim,S., Kimchi,A. *et al.* (2020) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **48**, D9–D16.
37. Lee,C.M., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Gonzalez,J.N., Hinrichs,A.S., Lee,B.T., Nassar,L.R., Powell,C.C. *et al.* (2020) UCSC Genome Browser enters 20th year. *Nucleic Acids Res.*, **48**, D756–D761.
38. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.
39. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
40. Wu,X., Liu,M., Downie,B., Liang,C., Ji,G., Li,Q.Q. and Hunt,A.G. (2011) Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 12533–12538.
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. 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.
43. Neve,J., Patel,R., Wang,Z., Louey,A. and Furger,A.M. (2017) Cleavage and polyadenylation: ending the message expands gene regulation. *RNA Biol.*, **14**, 865–890.
44. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, **27**, 1653–1659.
45. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.
46. 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.
47. Hu,Z.L., Fritz,E.R. and Reecy,J.M. (2007) AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyon3d. *Nucleic Acids Res.*, **35**, D604–D609.
48. Yang,W., Yang,Y., Zhao,C., Yang,K., Wang,D., Yang,J., Niu,X. and Gong,J. (2020) Animal-ImputeDB: a comprehensive database with multiple animal reference panels for genotype imputation. *Nucleic Acids Res.*, **48**, D659–D667.