Table S1: Summary of different types of 3′ seq techniques and data characteristics

3′ seq techniques			*Example data*						
Category	**3′ seq technique**	**Reference**	**Accession No. of example data**	**A or T stretch**	**Species**	**Biological sample**	**Read length (nt)**	**5' end**	**3' end**
Oligo(dT) priming-based	PAS-seq (II)	(Conrad, et al., 2021)	SRR1183738	>10 nt As	Human	293 cells	100	6nt UMI	Adapter
	PolyA-Tag seq	(Zhu, et al., 2020)	SRR5055884	>10 nt Ts	*Arabidopsis thaliana*	Leaf	100	8nt UMI	Adapter
	QuantSeq 3′mRNA-seq	(Kamieniarz-Gdula, et al., 2019)	SRR8383324	>10 nt Ts	Human	Hela cells	20-50	-	-
	PAS-seq (III)	(Routh, 2019)	PRJNA436720	>10 nt As	Mouse	mTOR activation cellular	100	4nt UMI	Adapter
	WTTS-seq	(Zhang, et al., 2018)	SRR6312600	>6 nt Ts	Mouse	Muscle	25-721	-	-
	2P-seq	(Zhou, et al., 2018)	PRJNA419320	>10 nt As	*Neurospora crassa*	Wild-type	151	-	Poly(G)
	PAC-seq (Poly(A)-ClickSeq)	(Routh, et al., 2017)	SRR5440716	>10 nt As	Human	Hela cells	151	6nt UMI	Adapter
	WTTS-seq	(Zhou, et al., 2016)	GSM1937549	>10 nt Ts	*Xenopus tropicalis*	Whole body	8-389	-	-
	SAPAS	(You, et al., 2015)	SRR065455	>10 nt Ts	Human	Breast cancer cells	76	-	-
	A-seq2	(Gruber, et al., 2014)	SRR1168402	3 nt Ts	Human	T cells	51	4nt UMI	-
	3′-seq	(Lianoglou, et al., 2013)	SRR6830250	>10 nt As	Human	B cells	22-51	-	-
	2P-seq	(Spies, et al., 2013)	SRR766745	>10 nt Ts	Mouse	3T3 cells	40	-	-
	PolyA-seq	(Derti, et al., 2012)	SRR299116	-	Human	UHR	76	-	Adapter
	PAS-seq (I)	(Shepard, et al., 2011)	GSM624868	>10 nt Ts	Human	Hela cells	40	-	-
	PolyA-Tag seq	-	SRR14512171	>10 nt Ts	Human	Neural differentiation in response to AgSCs	110	8nt UMI	Adapter
	3′READS (I)	(Wang, et al., 2018)	SRR16867064	>10 nt Ts	Human	Hela cells	76	6nt UMI	Adapter
	cTag-PAPERCLIP	(Jereb, et al., 2018)	GSM2901339	>10 nt As	Mouse	Cerebellar granule cells	64	4nt UMI	-

RNA manipulation-based	PolyA-Test seq	(Tucey, et al., 2018)	SRR5666331	>10 nt As	Mouse	Macrophages	151	4nt UMI	-
	PAPERCLIP	(Hwang, et al., 2016)	SRR1810989	>10 nt As	Human	Hela cells	87	-	-
	PolyA-Test seq	(Harrison, et al., 2015)	GSM1293999	>10 nt As	*Saccharomyces cerevisiae*	Budding yeast cells	101	-	Adapter
	3P-seq	(Spies, et al., 2013)	SRR766743	>2 nt Ts	Mouse	3T3 cells	36	-	-
	3′READS (II)	(Hoque, et al., 2013)	GSM1039195	>10 nt As	Mouse	Cell line mix	80	-	-
DRS-based	DRS	(Majerciak, et al., 2013)	ERR294004	-	*Arabidopsis thaliana*	Seed	25-57	-	-

Table S2: Existing studies involving identifying polyA sites from data generated by different 3′ seq techniques

Id	*Study*	*Toolkit*	*Species*	*Applicable 3′ seq techniques*	*pA annotation* [a]	*pA clustering strategy* [b]	*Internal priming strategy* [c]
1	(Wang, et al., 2023)	×	*Spartina alterniflora*	PAT-seq	x	24nt	A-rich
2	(Yalamanchili, et al., 2020)	PolyA-miner;	Animals	Several 3′-seq techniques	BED file, only consider 3′UTR and downstream 16Kb	15nt	A-rich and known pAs
3	(Herrmann, et al., 2020)	PolyASite 2.0	Animals	Several 3′-seq techniques	x	24nt	A-rich
4	(Zhu, et al., 2020)	PlantAPAdb	Plants	Several 3′-seq techniques	GenomicFeatures (R)	24nt	A-rich
5	(Routh, 2019)	DPAC	Animals	Several 3′-seq techniques with As/Ts > 10nt	BED file of UCSC	25nt	A-rich
6	(Zhou, et al., 2019)	×	*Xenopus tropicalis*	WTTS-seq	Cuffcompare	24nt	A-rich
7	(Wang, et al., 2018)	PolyA_DB 3	Animals	3′READS	RefSeq	24nt	-
8	(West, et al., 2018)	×	*Caenorhabditis elegans*	LITE-Seq	HTSeq	12nt	A-rich
9	(Jereb, et al., 2018)	×	Mouse	cTag-PAPERCLIP	BED-tools	Custom scripts	A-rich
10	(Guvenek and Tian, 2018)	×	Mouse	3′READS	GenomicFeatures (R)	24nt	-
11	(Chen, et al., 2018)	×	Mouse	PA-seq	RefSeq	F-seq	A-rich
12	(Sanfilippo, et al., 2017)	×	Drosophila	3′-seq	FeatureCounts	25nt	A-rich
13	(Routh, et al., 2017)	Python scripts	Drosophila	Poly(A)-ClickSeq	UCSC genome browser	10nt	A-rich
14	(Liu, et al., 2017)	×	Yeast	3′READS	Ensembl	7nt	-
15	(Jia, et al., 2017)	×	Human, mouse	IVT-SAPAS	√	24nt	A-rich
16	(Hong, et al., 2017)	×	Arabidopsis	PAT-seq	Perl scripts	24nt	A-rich
17	(Fu, et al., 2016; Zhou, et al., 2019)	Perl scripts	Rice	PAT-seq	Perl scripts	24nt	A-rich
18	(Bell, et al., 2016; de Lorenzo, et al., 2017)	×	Arabidopsis, *Chlamydomonas*	PAT-seq	BEDTools	Custom scripts	A-rich
19	(Gruber, et al.,	PolyASite	Human, mouse	3′-seq	√	24nt	A-rich, tetramers

	2016)						
20	(Li, et al., 2015)	×	Mouse	3′READS	RefSeq	24nt	-
21	(Blazie, et al., 2015)	×	*Caenorhabditis elegans*	PAT-Seq	√	√	A-rich
22	(Wilkening, et al., 2013)	x	Yeast	3′T-fill	√	√	A-rich
23	(Hoque, et al., 2013)	×	Mouse	3′READS	RefSeq	24nt	-
24	(Schlackow, et al., 2013)	×	Yeast	RNA-seq	√	6nt	A-rich
25	(Ni, et al., 2013)	×	Human	PA-seq	RefSeq	F-seq	A-rich
26	(Mata, 2013)	Perl scripts	Yeast	3PC	√	25nt	A-rich
27	(Lianoglou, et al., 2013)	×	Human	3′-seq	√	peak	A-rich
28	(Ulitsky, et al., 2012)	×	Zebrafish	3P-seq	Custome script	10nt	×
29	(Smibert, et al., 2012)	×	Drosophila	Poly(A) Enrichment	√	10nt	A-rich
30	(Sherstnev, et al., 2012)	×	Arabidopsis	DRS	√	peak	-
31	(Derti, et al., 2012)	×	Mammals	PolyA-seq	√	peak	robust filtering
32	(Fu, et al., 2011; Li, et al., 2012)	×	Zebrafish, Human	SAPAS	√	24nt	A-rich
33	(Thomas, et al., 2012; Wu, et al., 2011)	Perl scripts	Arabidopsis	PAT-seq	Perl scripts	24nt	A-rich
34	(Shepard, et al., 2011)	×	Human, mouse	PAS-Seq	√	40nt	A-rich
35	(Jan, et al., 2011)	×	*Caenorhabditis elegans*	3P-Seq	√	21nt	-
36	(Ozsolak, et al., 2010)	×	Human, yeast	DRS	SeqSolve	×	-
37	(Mangone, et al., 2010)	×	*Caenorhabditis elegans*	polyA capture	√	Clustering	A-rich

[a] pA annotation: “x” means annotation was not provided in the respective study; “√” means annotation was provided but the method is unknown.

[b] pA clustering strategy: “N nt” means that adjacent cleavage sites within X nt were grouped into polyA site clusters; “√” means cleavage sites were grouped but the method is unknown; “peak” means polyA site clusters were obtained by peak calling methods.

[c] internal priming strategy: “A-rich” means internal priming artifacts were eliminated by heuristic filtering based on A-rich sequences around alignment locations; “-” means internal priming removing was not mentioned in the study.

Table S3: Existing pipelines or tools for identifying polyA sites from 3′ seq data

Pipeline / tool	*Code*	*Preprocessing of raw sequencing data*	*Read aligner*	*Strategy to remove internal priming*	*Strategy to cluster polyA reads*	*Strategy to filter polyA sites*	*Applicable 3' seq techniques*
PlantAPAdb (Zhu, et al., 2020)	R, in-house Perl scripts	Adapter trimming, polyA tail trimming	STAR	Heuristic filtering: six consecutive As or more than six As within the -10 to +10 nt window of polyA site.	24 nt grouping	Normalized read count > 1 in two or more experiments	many
DPAC (Routh, 2019)	Shell scripts, Python scripts	Adapter trimming, polyA tail trimming. Reads containing polyA tracts shorter than 10 As are discarded.	HISAT2	Heuristic filtering: 12 or more As within 20 nt downstream.	25 nt grouping	-	3'READS, PAS-Seq, Poly(A)-Click-Seq
PolyASite 2.0 (Herrmann, et al., 2020)	Perl, Python, R scripts	Protocol-specific pre-processing	Segemehl	Heuristic filtering: six consecutive As or more than six As within the -10 to +10 nt window of polyA site.	25 nt grouping	PolyA signal in the region of -60 to +10 nt in >=90% reads in a cluster.	many
PolyA-miner (Yalamanchili, et al., 2020)	Python scripts	The first six nucleotides and adapter contamination is trimmed and reads <40 bp are filtered out.	bowtie2	Sites with >65% of As in a sliding window of 20 bp are filtered out. Sites within 50 bp of an annotated cleavage site (Zhang, et al., 2005) are considered accurate.	Peaks with base coverage, merged within a given distance.	Map novel sites to genes if they are within 16 Kb from respective transcriptional end site.	PAC-seq
(Sheppard, et al., 2013)	In-house Perl scripts	Reads were trimmed 30 terminal nt	Tophat	Bayes classifier	5 nt grouping	-	PAS-Seq, 3P-seq
RECTIFY (Roy and Chanfreau, 2020)	In-house Python scripts, R scripts	3'-trimmed for quality score < 20 and adapters.	BWA	Heuristic filtering: downstream of polyA sites with 6 or more bases meeting certain criteria are flagged as uncertain.	R package VariantAnnotation	-	QuantSeq 3' mRNA seq
poly-A-seq (Yu, et al., 2020)	Shell scripts, Perl scripts	Poly(G) trimming, polyA tail trimming;	Tophat	Heuristic filtering: six consecutive As or more than six As within the -10 to +10 nt window from the cleavage site	24 nt grouping	-	3'READS, 2P-Seq

Table S4: Applicability of different polyA site identification pipelines or tools on various 3′ seq datasets

3'seq technique	*Reference*	*PolyA-miner*	*DPAC*	*PolyASite*	*Poly-A-seq*
3′READS	(Wang, et al., 2018)	√	√	√	√
PAS-seq	(Shepard, et al., 2011)	√	√	√	×[d]
PAS-seq	(Conrad, et al., 2021)	√	√	×[c]	×[d]
PAC-seq	(Routh, et al., 2017)	√	√	×[c]	×[d]
PolyA-Tag seq	-	√	√	×[c]	×[d]
3′-seq	(Lianoglou, et al., 2013)	√	√	√	×[d]
PAPERCLIP	(Hwang, et al., 2016)	√	√	√	×[d]
QuantSeq 3'mRNA-seq	(Kamieniarz-Gdula, et al., 2019)	√	√	√	×[d]
SAPAS	(You, et al., 2015)	√	√	√	×[d]
A-seq2	(Gruber, et al., 2014)	√	×[b]	×[c]	×[d]
PolyA-seq	(Derti, et al., 2012)	√	×[b]	√	×[d]
2P-seq	(Spies, et al., 2013)	√	√	√	×[d]
PolyA-Test seq	(Tucey, et al., 2018)	√	√	×[c]	×[d]
WTTS-seq	(Zhang, et al., 2018)	√	√	×[c]	×[d]
3P-seq	(Spies, et al., 2013)	√	√	√	×[d]
3′READS	(Hoque, et al., 2013)	√	√	√	√
PAS-seq	(Routh, 2019)	√	√	×[c]	×[d]
cTag-PAPERCLIP	(Jereb, et al., 2018)	√	√	√	×[d]
PolyA-Tag seq	(Zhu, et al., 2020)	√	√	×[c]	×[d]
PolyA-Test seq	(Harrison, et al., 2015)	×[a]	√	×[c]	×[d]
2P-seq	(Zhou, et al., 2018)	×[a]	√	×[c]	√
WTTS-seq	(Zhou, et al., 2016)	×[a]	√	×[c]	×[d]
DRS	(Majerciak, et al., 2013)	√	×[b]	√	×[d]

[a] the dataset cannot be analyzed by the tool because of the lack of polyA site annotation for the species (Arabidopsis) of this dataset.

[b] the dataset cannot be analyzed by the tool because that the polyA stretch of this dataset is shorter than 10 nt.

[c] the dataset cannot be analyzed by the tool because that the processing script is not applicable for this dataset.

[d] the dataset cannot be analyzed by the tool because no script is applicable for this dataset.

Table S5: 3′ seq datasets used in this study

Accession No.	*3′ seq technique*	*Reference*	*3′ seq category*	*A/T stretch*	*Species*	*Biological sample*	*Read Length (nt)*	*5′ end*	*3′ end*
SRR1183738	PAS-seq	(Conrad, et al., 2021)	Oligo(dT) based	>10 nt As	Human	293 cells	100	6nt UMI	adapter
SRR8383324	QuantSeq 3′mRNA-seq	(Kamieniarz-Gdula, et al., 2019)	Oligo(dT) based	>10 nt Ts	Human	Hela cells	20-50	-	-
SRR16867064	3'READS	(Wang, et al., 2018)	RNA manipulation-based	>10 nt Ts	Human	Hela cells	76	6nt UMI	adapter
SRR5440716	PAC-seq	(Routh, et al., 2017)	Oligo(dT) based	>10 nt As	Human	Hela cells	151	6nt UMI	adapter
SRR1810989	PAPERCLIP	(Hwang, et al., 2016)	RNA manipulation-based	>10 nt As	Human	Hela cells	87	-	-
SRR065455	SAPAS	(You, et al., 2015)	Oligo(dT) based	>10 nt Ts	Human	breast cancer cell	76	-	-
SRR1168402	A-seq2	(Gruber, et al., 2014)	Oligo(dT) based	3 nt Ts	Human	T cell	51	4nt UMI	-
SRR6830250	3′-seq	(Lianoglou, et al., 2013)	Oligo(dT) based	>10 nt As	Human	B cells	22-51	-	-
SRR299116	PolyA-seq	(Derti, et al., 2012)	Oligo(dT) based	-	Human	UHR	76	-	adapter
GSM2901339	cTag-PAPERCLIP	(Jereb, et al., 2018)	RNA manipulation-based	>10 nt As	Mouse	Cerebellar granule cells	64	4nt UMI	-
SRR5666331	PolyA-Test seq	(Tucey, et al., 2018)	RNA manipulation-based	>10 nt As	Mouse	macrophages	151	4nt UMI	-
SRR6312600	WTTS-seq	(Zhang, et al., 2018)	Oligo(dT) priming-based	>6 nt Ts	Mouse	muscle	25-721	-	-
SRR766743	3P-seq	(Spies, et al., 2013)	RNA manipulation-based	>2 nt Ts	Mouse	3T3 cells	36	-	-
SRR766745	2P-seq	(Spies, et al., 2013)	Oligo(dT) based	>10 nt Ts	Mouse	3T3 cells	40	-	-
SRR5055884	PolyA-Tag seq	(Zhu, et al., 2020)	Oligo(dT) based	>10 nt Ts	Arabidopsis	Leaf	100	8nt UMI	adapter
ERR294004	DRS	(Majerciak, et al., 2013)	DRS-based	-	Arabidopsis	Seed	25-57	-	-
SRR5276077	3′READS	(Yu, et al., 2014)	RNA manipulation-based	>10 nt Ts	Budding yeast	wild type	51	-	adapter
SRR5276080	3′READS	(Yu, et al., 2014)	RNA manipulation-based	>10 nt Ts	Fission yeast	wild type	51	-	adapter

Table S6: Performance of the DeepIP model for different species

Species-specific evaluation: the model was trained and tested on the same species.

Trained species	Precision	Recall	F1
Arabidopsis	0.9973	0.9504	0.9733
Chlamy	1	0.9886	0.9943
Human	0.9732	0.9674	0.9703
Mouse	0.9421	0.964	0.9529
Budding yeast	0.9927	0.965	0.9786
Fission yeast	0.9834	0.9757	0.9796
Average score	0.9815	0.9685	0.9748

Cross-species evaluation: the model was trained on one species and test on the other species.

Trained species	Test species	Precision	Recall	F1 score	Trained species	Test species	Precision	Recall	F1 score
Arabidopsis	Chlamy	0.9878	0.9221	0.9538	Mouse	Arabidopsis	0.9311	0.9412	0.9361
Arabidopsis	Human	0.9986	0.8762	0.9334	Mouse	Chlamy	0.75	0.9964	0.8558
Arabidopsis	Mouse	0.9994	0.847	0.9169	Mouse	Human	0.9515	0.978	0.9646
Arabidopsis	Budding yeast	1	0.9393	0.9687	Mouse	Budding yeast	0.9214	0.9543	0.9375
Arabidopsis	Fission yeast	1	0.9514	0.9751	Mouse	Fission yeast	0.9413	0.9729	0.9568
Chlamy	Arabidopsis	0.9841	0.9258	0.954	Budding yeast	Arabidopsis	0.9661	0.9358	0.9507
Chlamy	Human	0.9559	0.9008	0.9275	Budding yeast	Chlamy	0.9993	0.98	0.9895
Chlamy	Mouse	0.9439	0.8708	0.9059	Budding yeast	Human	0.9226	0.9134	0.918
Chlamy	Budding yeast	0.9873	0.9407	0.9634	Budding yeast	Mouse	0.9245	0.8848	0.9042
Chlamy	Fission yeast	0.9845	0.9557	0.9699	Budding yeast	Fission yeast	0.9898	0.9729	0.9813
Human	Arabidopsis	0.9639	0.9352	0.9493	Fission yeast	Arabidopsis	0.9733	0.9404	0.9566
Human	Chlamy	0.7181	0.9971	0.8349	Fission yeast	Chlamy	0.9978	0.9857	0.9917
Human	Mouse	0.9637	0.9498	0.9567	Fission yeast	Human	0.9553	0.9014	0.9276
Human	Budding yeast	0.9601	0.9443	0.9521	Fission yeast	Mouse	0.9772	0.8775	0.9247
Human	Fission yeast	0.972	0.9671	0.9696	Fission yeast	Budding yeast	0.9875	0.9629	0.975
	Average score	0.9536	0.9374	0.9434					

Table S7: Comparison of the steps of the pipeline between PolyAseqTrap and other 3′ seq tools / pipelines.

Pipeline / tool	*Code*	*Preprocessing of raw sequencing data*	*Strategy to remove internal priming*	*Strategy to cluster polyA reads*	*Strategy to annotate polyA sites*
PolyAseqTrap (this study)	R package	Using a polyA read prioritization strategy with fine post-inspection.	Using a transferrable cross-species deep learning model.	Using a weighted density peak clustering method.	Allowing the use of various genome annotation files; generating well formatted reports from three levels – polyA read, cleavage site, and PAC.
DPAC (Routh, 2019)	Shell scripts, Python scripts	Adapter trimming, polyA tail trimming. Reads containing polyA tracts shorter than 10 As are discarded.	12 or more As within 20 nt downstream.	25 nt grouping.	Using genome annotation in BED format.
	☹ Flaw	☹ Only suitable for certain 3′ seq data.	☹ The threshold is arbitrarily set and there is no unified standard.	☹ The cutoff is subjective and arbitrary, tending to mix adjacent pA sites.	☹ No flexible annotation or visualization; no rich reports.
PolyA-miner (Yalamanchili, et al., 2020)	Python scripts	The first six nucleotides and adapter contamination is trimmed and reads <40 bp are filtered out.	Sites with >65% of As in a sliding window of 20 bp; sites not within 50 bp of annotated sites (Zhang, et al., 2005).	Peaks with base coverage, merged within a given distance.	Map novel sites to genes if they are within 16 Kb from respective transcriptional end site.
	☹ Flaw	☹ Only suitable for certain 3′ seq data.	☹ The threshold is arbitrarily set and known pA sites are needed.	☹ The cutoff is subjective and arbitrary, tending to mix adjacent pA sites.	☹ Missing pA sites in intergenic regions; no flexible annotation or visualization; no rich reports.
PolyASite 2.0 (Herrmann, et al., 2020)	Perl, Python, R scripts	Protocol-specific pre-processing.	Six consecutive As or more than six As within the -10 to +10 nt window of polyA site.	25 nt grouping.	Only polyA position list.
	☹ Flaw	☹ Lack flexibility in coping with new types of 3′seq data.	☹ The threshold is arbitrarily set and there is no unified standard.	☹ The cutoff is subjective and arbitrary, tending to mix adjacent pA sites.	☹ No flexible annotation or visualization; no rich reports.

Table S8: Key functions and adjustable parameters in PolyAseqTrap

Parameter	Description	Default value	Variable type
FindPTA: function used to identify and quantify polyA sites from BAM files			
Bam	Input BAM file to process	—	File Path
yieldSize	Maximum number of records read at a time using Rsamtools::scanBam	2.60E+07	Integer
reverse	Reverse complement reads mapped to the minus strand (BAM files store them as if on the plus strand)	FALSE	Logical (TRUE/FALSE)
bsgenome	A BSgenome object for reference genome (e.g., BSgenome.Hsapiens.UCSC.hg38)	—	BSgenome object
d	Distance (nt) to merge nearby polyA sites into a polyA cluster (PAC)	24	Integer
poly	Type of polyA tail at 3’ end. Must be 'T' or 'A'	T	Character ('T'/'A')
adjust.chr	Add 'chr' prefix if BAM chromosome names lack it but genome includes it	FALSE	TRUE/FALSE
threeUTRregion	A GRanges object for 3′UTR regions; used when reads lack polyA tails	—	GRanges object
ext3UTRlen	Length (nt) to extend 3′UTR regions for polyA site analysis	1000	Integer
cutoffCount	Minimum number of supporting reads required for C3–V8 category PACs to be considered as valid polyA sites	5	Integer
isDRS	Specify TRUE if input is Direct RNA Sequencing (DRS) data	FALSE	Logical (TRUE/FALSE)
run.quantify	If TRUE, then perform quantification of polyA sites	TRUE	Logical (TRUE/FALSE)
detect.repeat	Check whether polyA sites originate from A-rich repeat regions	FALSE	Logical (TRUE/FALSE)
trainDeepIP: function used to train DeepIP model for given input sequences			
condaEnv	The Python conda environment	—	Conda environment path
inTrainSeq	Input sequences (200 nt each), including positive and negative polyA sites	—	File Path
outTrainedModel	Output file to save the trained model (HDF5 format)	—	File Path
epoch	Number of training iterations	100	Integer
seqLable	Sequence labels for inTrainSeq, *e.g.*, 0/1 or 0:1	—	Character (e.g.,0/1, 0:1)

testDeepIP: function used to test given input sequences using given trained model			
inTestSeq	Input sequences (FASTA file), 100 nt upstream and downstream of examined polyA sites	—	File Path
inTrainedModel	File of the pre-trained DeepIP model (HDF5 format)	—	File Path
outTestCsv	Output CSV file to store prediction results	—	File Path
seqLabel	Labels for sequences in inTestSeq, e.g., 0/1 or 0:1	—	Character (e.g.,0/1, 0:1)
resut.PA: function used to regroup nearby polyA sites into PACs and quantify their expression			
pa.data	A list containing polyA site information, typically the output from findTailAT	—	An R object
d	Maximum distance (nt) between adjacent polyA sites to be grouped into a single PAC	24	Integer
split_pac: function used to refine polyA site clusters to reduce microheterogeneity			
pa.data	A list containing polyA site information, typically the output from findTailAT	—	An R object
d	Threshold distance (nt) for cluster refinement: initial PACs with a range ≥ d are further processed using density peak clustering to detect potential sub-clusters	24	Integer
mc.cores	Number of processors/cores to use for parallel clustering	4	Integer

Table S9: PolyA site datasets used for DeepIP model training and test

Species	*PolyA site category*	*PolyA site collections*
Human	High-quality	PolyA_DB 3 (Wang, et al., 2018)
	Mixed-quality	PolyASite 2.0 (Herrmann, et al., 2020)
Mouse	High-quality	PolyA_DB 3 (Wang, et al., 2018)
	Mixed-quality	PolyASite 2.0 (Herrmann, et al., 2020)
Arabidopsis	High-quality	PolyA sites from DRS data (Sherstnev, et al., 2012)
	Mixed-quality	"Arabidopsis_thaliana.high_confidence.PAC.bed" downloaded from PlantAPAdb (Zhu, et al., 2020)
Budding yeast	High-quality	PolyA sites with level V1 obtained by PolyAseqTrap from 3′READS data (SRR5276077 (Yu, et al., 2014))
	Mixed-quality	PolyA sites with other levels obtained by PolyAseqTrap (SRR5276077 (Yu, et al., 2014))
Fission yeast	High-quality	PolyA sites with level V1 obtained from PolyAseqTrap for 3′READS data (SRR5276080 (Yu, et al., 2014))
	Mixed-quality	PolyA sites with other levels obtained by PolyAseqTrap (SRR5276080 (Yu, et al., 2014))
Chlamy	High-quality	"chlamydomonas_reinhardtii.high_confidence.PAC.bed" downloaded from PlantAPAdb (Zhu, et al., 2020)
	Mixed-quality	-

Note: The polyA collections for a same species may be based on different versions of genome annotation. In our analyses, we aligned the coordinates from different genome versions to the same version using tools like LiftOver. We have ensured that the versions of the genome annotation and genome assembly match.

References

Bell, S.A., *et al.* Experimental Genome-Wide Determination of RNA Polyadenylation in Chlamydomonas reinhardtii. *PLoS One* 2016;11(1):e0146107.

Blazie, S.M., *et al.* Comparative RNA-Seq analysis reveals pervasive tissue-specific alternative polyadenylation in Caenorhabditis elegans intestine and muscles. *BMC Biol.* 2015;13(1):4.

Chen, M., *et al.* 3' UTR lengthening as a novel mechanism in regulating cellular senescence. *Genome Res.* 2018.

Conrad, N.K., *et al.* Mechanism and consequences of herpes simplex virus 1-mediated regulation of host mRNA alternative polyadenylation. *PLos Genet.* 2021;17(3):e1009263.

de Lorenzo, L., *et al.* Noncanonical Alternative Polyadenylation Contributes to Gene Regulation in Response to Hypoxia. *The Plant Cell* 2017;29(6):1262-1277.

Derti, A., *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Res.* 2012;22(6):1173-1183.

Fu, H., *et al.* Genome-wide dynamics of alternative polyadenylation in rice. *Genome Res.* 2016;26(12):1753-1760.

Fu, Y., *et al.* Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.* 2011;21(5):741-747.

Gruber, A.J., *et al.* 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.* 2016;26(8):1145-1159.

Gruber, A.R., *et al.* Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nat. Commun.* 2014;5(1):5465.

Guvenek, A. and Tian, B. Analysis of alternative cleavage and polyadenylation in mature and differentiating neurons using RNA-seq data. *Quant Biol* 2018;6(3):253-266.

Harrison, P.F., *et al.* PAT-seq: a method to study the integration of 3′-UTR dynamics with gene expression in the eukaryotic transcriptome. *RNA* 2015;21(8):1502-1510.

Herrmann, C.J., *et al.* PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing. *Nucleic Acids Res* 2020;48(D1):D174-d179.

Herrmann, C.J., *et al.* PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing. *Nucleic Acids Res.* 2020;48(D1):D174-D179.

Hong, L., *et al.* Alternative polyadenylation is involved in auxin-based plant growth and development. *The Plant Journal* 2017;93(2):246-258.

Hoque, M., *et al.* Analysis of alternative cleavage and polyadenylation by 3 ' region extraction and deep sequencing. *Nat. Methods* 2013;10(2):133-139.

Hwang, H.W., *et al.* PAPERCLIP Identifies MicroRNA Targets and a Role of CstF64/64tau in Promoting Non-canonical poly(A) Site Usage. *Cell Rep.* 2016;15(2):423-435.

Jan, C.H., *et al.* Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature* 2011;469(7328):97-101.

Jereb, S., *et al.* Differential 3' Processing of Specific Transcripts Expands Regulatory and Protein Diversity Across Neuronal Cell Types. *Elife* 2018;7.

Jereb, S., *et al.* Differential 3' Processing of Specific Transcripts Expands Regulatory and Protein Diversity Across Neuronal Cell Types. *Elife* 2018;7.

Jia, X., *et al.* The role of alternative polyadenylation in the antiviral innate immune response. *Nat. Commun.* 2017;8:14605.

Kamieniarz-Gdula, K., *et al.* Selective roles of vertebrate PCF11 in premature and full-length transcript termination. *Mol. Cell* 2019;74(1):158-172 e159.

Li, W., *et al.* Systematic Profiling of Poly(A)+ Transcripts Modulated by Core 3’ End Processing and Splicing Factors Reveals Regulatory Rules of Alternative Cleavage and Polyadenylation. *PLos Genet.* 2015;11(4):e1005166.

Li, Y., *et al.* Dynamic landscape of tandem 3 ' UTRs during zebrafish development. *Genome Res.* 2012;22(10):1899-1906.

Lianoglou, S., *et al.* Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.* 2013;27(21):2380-2396.

Liu, X., *et al.* Comparative analysis of alternative polyadenylation in S. cerevisiae and S. pombe. *Genome Res.* 2017;27(10):1685-1695.

Majerciak, V., *et al.* A viral genome landscape of RNA polyadenylation from KSHV latent to lytic infection. *PLoS Pathog.* 2013;9(11):e1003749.

Mangone, M., *et al.* The landscape of C. elegans 3'UTRs. *Science* 2010;329(5990):432-435.

Mata, J. Genome-wide mapping of polyadenylation sites in fission yeast reveals widespread alternative polyadenylation. *RNA Biol.* 2013;10(8):1407-1414.

Ni, T., *et al.* Distinct polyadenylation landscapes of diverse human tissues revealed by a modified PA-seq strategy. *BMC Genomics* 2013;14(1):615.

Ozsolak, F., *et al.* Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* 2010;143(6):1018-1029.

Routh, A. DPAC: A Tool for Differential Poly(A) Cluster Usage from Poly(A)-Targeted RNAseq Data. *g3 genes genomes genetics* 2019;9(6):1825-1830.
Routh, A., *et al.* Poly(A)-ClickSeq: click-chemistry for next-generation 3-end sequencing without RNA enrichment or fragmentation. *Nucleic Acids Res.* 2017;45(12):e112.
Routh, A., *et al.* Poly(A)-ClickSeq: click-chemistry for next-generation 3′-end sequencing without RNA enrichment or fragmentation. *Nucleic Acids Res.* 2017;45(12):e112.
Roy, K.R. and Chanfreau, G.F. Robust mapping of polyadenylated and non-polyadenylated RNA 3′ ends at nucleotide resolution by 3′-end sequencing. *Methods* 2020;176:4-13.
Sanfilippo, P., Wen, J. and Lai, E.C. Landscape and evolution of tissue-specific alternative polyadenylation across Drosophila species. *Genome Biol.* 2017;18(1):229.
Schlackow, M., *et al.* Genome-wide analysis of poly(A) site selection in Schizosaccharomyces pombe. *RNA* 2013;19(12):1617-1631.
Shepard, P.J., *et al.* Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* 2011;17(4):761-772.
Sheppard, S., Lawson, N.D. and Zhu, L.J. Accurate identification of polyadenylation sites from 3′ end deep sequencing using a naïve Bayes classifier. *Bioinformatics* 2013;29(20):2564-2571.
Sherstnev, A., *et al.* Direct sequencing of Arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. *Nat. Struct. Mol. Biol.* 2012;19(8):845-852.
Smibert, P., *et al.* Global Patterns of Tissue-Specific Alternative Polyadenylation in Drosophila. *Cell Rep.* 2012;1(3):277-289.
Spies, N., Burge, C.B. and Bartel, D.P. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* 2013;23(12):2078-2090.
Thomas, P.E., *et al.* Genome-Wide Control of Polyadenylation Site Choice by CPSF30 in Arabidopsis. *Plant Cell* 2012;24(11):4376-4388.
Tucey, T.M., *et al.* Glucose Homeostasis Is Important for Immune Cell Viability during Candida Challenge and Host Survival of Systemic Fungal Infection. *Cell Metab.* 2018;27(5):988-1006 e1007.
Ulitsky, I., *et al.* Extensive alternative polyadenylation during zebrafish development. *Genome Res.* 2012;22(10):2054-2066.
Wang, R., *et al.* PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res.* 2018;46(D1):D315-d319.
Wang, R., *et al.* A compendium of conserved cleavage and polyadenylation events in mammalian genes. *Genome Res.* 2018;28(10):1427-1441.
Wang, T., *et al.* Alternative 3′-untranslated regions regulate high-salt tolerance of Spartina alterniflora. *Plant Physiol.* 2023;191(4):2570-2587.
West, S.M., *et al.* Developmental dynamics of gene expression and alternative polyadenylation in the Caenorhabditis elegans germline. *Genome Biol.* 2018;19(1):8.
Wilkening, S., *et al.* An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic Acids Res.* 2013;41(5):e65-e65.
Wu, X., *et al.* Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation. *Proc. Natl. Acad. Sci. USA* 2011;108(30):12533-12538.
Yalamanchili, H.K., *et al.* PolyA-miner: accurate assessment of differential alternative polyadenylation from 3'Seq data using vector projections and non-negative matrix factorization. *Nucleic Acids Res.* 2020;48(12):e69.
You, L., *et al.* APASdb: a database describing alternative poly(A) sites and selection of heterogeneous cleavage sites downstream of poly(A) signals. *Nucleic Acids Res.* 2015;43(Database issue):D59-67.
Yu, F., *et al.* Poly(A)-seq: A method for direct sequencing and analysis of the transcriptomic poly(A)-tails. *PLoS One* 2020;15(6):e0234696.
Yu, R., *et al.* Determinants of heterochromatic siRNA biogenesis and function. *Mol. Cell* 2014;53(2):262-276.
Zhang, H., *et al.* PolyA_DB: a database for mammalian mRNA polyadenylation. *Nucleic Acids Res.* 2005;33:D116-120.
Zhang, S., *et al.* Alternative polyadenylation drives genome-to-phenome information detours in the AMPKalpha1 and AMPKalpha2 knockout mice. *Sci. Rep.* 2018;8(1):6462.
Zhou, Q., *et al.* Differential alternative polyadenylation contributes to the developmental divergence between two rice subspecies Japonica and Indica. *The Plant Journal* 2019;98(2):260-276.
Zhou, X., *et al.* Accurate Profiling of Gene Expression and Alternative Polyadenylation with Whole Transcriptome Termini Site Sequencing (WTTS-Seq). *Genetics* 2016;203(2):683-697.
Zhou, X., *et al.* Alternative polyadenylation coordinates embryonic development, sexual dimorphism and longitudinal growth in Xenopus tropicalis. *Cell. Mol. Life Sci.* 2019;76(11):2185-2198.
Zhou, Z., *et al.* Codon usage biases co-evolve with transcription termination machinery to suppress premature cleavage and polyadenylation. *eLife* 2018;7:e33569.
Zhu, S., *et al.* PlantAPAdb: a comprehensive database for alternative polyadenylation sites in plants. *Plant Physiol.* 2020;182(1):228-242.