nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-022-00704-8

Check for updates

Transcriptome-wide measurement of poly(A) tail length and composition at subnanogram total RNA sensitivity by PAIso-seq

Yusheng Liu[1]✉, Yiwei Zhang[2], Jiaqiang Wang[2]✉ and Falong Lu[1,3]✉

Poly(A) tails are added to the 3′ ends of most mRNAs in a non-templated manner and play essential roles in post-transcriptional regulation, including mRNA export, stability and translation. Measuring poly(A) tails is critical for understanding their regulatory roles in almost every aspect of biological and medical studies. Previous methods for analyzing poly(A) tails require large amounts of input RNA (microgram-level total RNA), which limits their application. We recently developed a poly(A) inclusive full-length RNA isoform-sequencing method (PAIso-seq) at single-oocyte-level sensitivity (a single mammalian oocyte contains ~0.5 ng of total RNA) based on PacBio sequencing that enabled accurate measurement of the poly(A) tail length and non-A residues within the body of poly(A) tails along with the full-length cDNA, providing the opportunity to study precious in vivo samples with very limited input material. Here, we describe a detailed protocol for PAIso-seq library preparation from single mouse oocytes or bulk oocyte samples. In addition, we provide a complete bioinformatic pipeline to perform the analysis from the raw data to downstream analysis. The minimum time required is ~14.5 h for PAIso-seq double-stranded cDNA preparation, 2 d for PacBio sequencing in HiFi mode and 8 h for the initial data analysis.

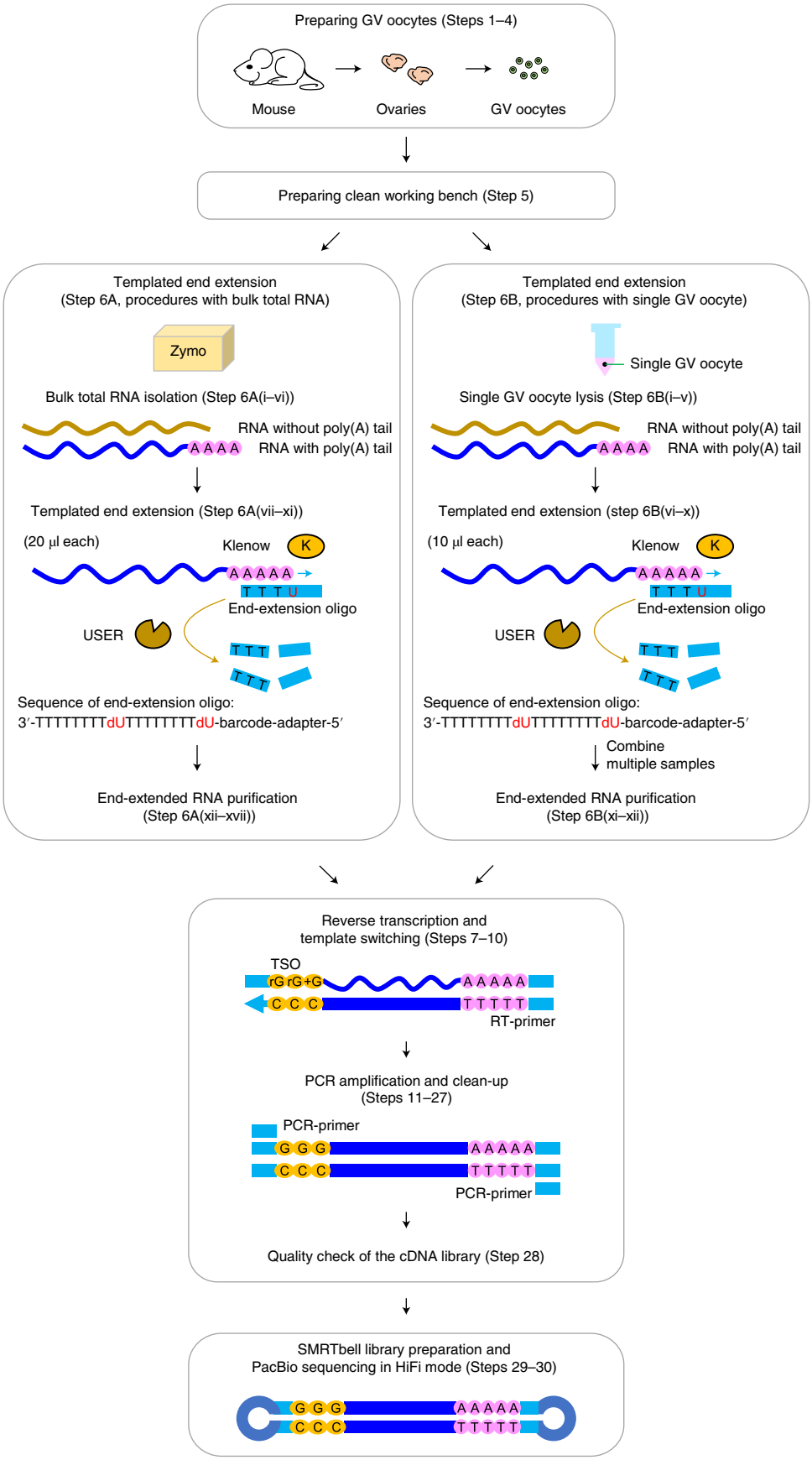
Introduction

The poly(A) tail is an important post-transcriptional modification for almost all mRNAs and long noncoding RNAs and has been known to play key roles in RNA stability, nuclear export and translational efficiency[1–7]. In oocytes and early embryos, the length of poly(A) tails has been demonstrated to positively regulate the translational efficiency of mRNA, which plays critical roles in oocyte-to-embryo transition[8–12]. Localized translational regulation at the neuronal synapses can also be achieved through cytoplasmic re-polyadenylation upon neuronal stimulation[13–16]. Recent studies reveal that non-A residues can be added to the 3′ ends of poly(A) tails, with 3′-end U residues found to have roles in promoting mRNA degradation and 3′-end G residues being important in stabilizing mRNA[2,3,17]. More recently, non-A residues have also been shown to be wide-spread in the body of RNA poly(A) tails, the function of which remain to be explored[18–20].

Comparison with alternative methods

Currently, there are several methods to measure RNA poly(A) tails. The methods based on the Illumina platform include poly(A)-tail length profiling by sequencing (PAL-seq)[1], PAL-seq-v2[21,22], a method for sequencing the very end of mRNA molecules (TAIL-seq)[17], mRNA TAIL-seq (mTAIL-seq)[4], PAT-seq[23], TED-seq[24] and poly(A)-seq[25]. Sequencing quality on the Illumina platform quickly degenerates on homopolymers, which makes it unable to sequence homopolymeric sequences longer than ~30 nt[17]. Among these methods, PAL-seq, PAL-seq-v2, TAIL-seq and mTAIL-seq achieve relatively good quantification of the length of poly(A) tails. In addition, TAIL-seq can also measure the non-A residues at the 3′ end of poly(A) tails. However, the PAL-seq, PAL-seq-v2, TAIL-seq and mTAIL-seq methods have to customize the sequencing recipes and base-calling algorithm, which makes them harder to adopt for most laboratories. Nanopore sequencing can be used to quantify the length of poly(A) tails but is not able to measure non-A residues within poly(A) tails[26–29]. Full-length poly(A) and mRNA sequencing (FLAM-seq) on the PacBio platform can accurately quantify the

[1]State Key Laboratory of Molecular Developmental Biology, Institute of Genetics and Developmental Biology, Innovative Academy of Seed Design, Chinese Academy of Sciences, Beijing, China. [2]College of Life Science, Northeast Agricultural University, Harbin, China. [3]University of Chinese Academy of Sciences, Beijing, China. ✉e-mail: liuys1126@foxmail.com; wangjiaqiang@neau.edu.cn; fllu@genetics.ac.cn


Preparing GV oocytes (Steps 1–4)
Mouse
Ovaries
GV oocytes
Preparing clean working bench (Step 5)
Templated end extension
(Step 6A, procedures with bulk total RNA)
Zymo
Bulk total RNA isolation (Step 6A(i–vi))
RNA without poly(A) tail
RNA with poly(A) tail
Templated end extension (Step 6A(vii–xi))
(20 µl each)
Klenow
End-extension oligo
USER
Sequence of end-extension oligo:
3′-TTTTTTTTdUTTTTTTTTdU-barcode-adapter-5′
End-extended RNA purification
(Step 6A(xii–xvii))
Templated end extension
(Step 6B, procedures with single GV oocyte)
Single GV oocyte
Single GV oocyte lysis (Step 6B(i–v))
RNA without poly(A) tail
RNA with poly(A) tail
Templated end extension (step 6B(vi–x))
(10 µl each)
Klenow
End-extension oligo
USER
Sequence of end-extension oligo:
3′-TTTTTTTTdUTTTTTTTTdU-barcode-adapter-5′
Combine
multiple samples
End-extended RNA purification
(Step 6B(xi–xii))
Reverse transcription and
template switching (Steps 7–10)
TSO
RT-primer
PCR amplification and clean-up
(Steps 11–27)
PCR-primer
PCR-primer
Quality check of the cDNA library (Step 28)
SMRTbell library preparation and
PacBio sequencing in HiFi mode (Steps 29–30)

◀ **Fig. 1 | Flowchart for the PAIso-seq protocol.** The input sample can be either purified total RNA (bulk mouse GV oocyte as an example) or direct cell lysis for low-input samples (single-mouse GV oocytes as an example). For the PAIso-seq double-stranded cDNA preparation, poly(A) tail preservation is performed with templated end extension following either bulk procedures or single-cell procedures. Then, full-length cDNA is generated by performing RT together with template switching, followed by PCR amplification. The resulting PAIso-seq double-stranded cDNA is made into an SMRTbell library and sequenced on a PacBio sequencer in HiFi mode. Procedure step numbers are shown in parentheses. GV, germinal vesicle; USER, uracil-specific excision reagent.

length of poly(A) tails and measure the non-A residues in poly(A) tails[19]. In addition, full-length elongating and polyadenylated RNA sequencing (FLEP-seq) can also be used for poly(A) tail length analysis on both PacBio and Nanopore platforms[30,31], although its performance measuring non-A residues within poly(A) tails has not been evaluated. In addition, all the above methods need a microgram level of total RNA with which to start[1,4,17,19,21,22,26–28,30].

We developed a method that could read transcriptome-wide poly(A) tail inclusive full-length RNA isoforms accurately and sensitively from the total RNA without the need for poly(A)$^+$ RNA enrichment, which we named 'poly(A) inclusive RNA isoform-sequencing' (PAIso-seq)[18]. Compared with other existing methods in analyzing poly(A) tails, PAIso-seq is more sensitive, being able to analyze RNA from a single mammalian oocyte[18]. Therefore, PAIso-seq can be applied to almost all kind of samples, especially as the best choice for in vivo samples with a very limited number of cells, such as human oocytes and pre-implantation embryos that are extremely rare for research. PAIso-seq uses an oligo dT/dU end-extension template to anneal to the poly(A) tails, which provides the template for end extension of the poly(A)$^+$ RNA (Fig. 1). Then, the templated end-extended RNA can be reverse-transcribed and amplified by using the template-switching approach to achieve the poly(A) tail inclusive full-length cDNA, which includes the full body of the mRNA sequence and the complete poly(A) tail sequence (Fig. 1). Then, the PAIso-seq double-stranded cDNA can be made into a SMRTbell library and sequenced on the PacBio platform in HiFi mode (Fig. 1). An accompanying bioinformatic pipeline was also developed for extraction of the transcriptome-wide poly(A) tail information from the PAIso-seq data (Fig. 2).

Compared to the Illumina and Nanopore platforms, PacBio sequencing in HiFi mode has the advantage of accurately measuring the homopolymeric poly(A) tail sequence[32–35], including the length and the nucleotide composition. Because PAIso-seq can sequence the full-length cDNA isoform together with the poly(A) tail, PAIso-seq data provide a good opportunity to study the interplay between alternative splicing, alternative polyadenylation, poly(A) tail length and non-A residues within poly(A) tails[18]. The power of the PAIso-seq method in analyzing RNA poly(A) tails has started to be recognized[36–40]. In the future, PAIso-seq has the potential to be further improved for poly(A) tail measurement from single somatic cells to dissect poly(A) tail heterogeneity among populations of cells, providing a new way to dissect the unique features of single cells, although further work would be needed in this case to improve the sensitivity of the method.

Applications of the method

We have previously successfully applied PAIso-seq to analyze transcriptome-wide poly(A) tails in mouse germinal vesicle (GV) oocytes[18]. First, it can accurately measure the RNA poly(A) tail length for each of the sequenced transcripts transcriptome-wide (Fig. 3a). The poly(A) tail length analysis can then be applied to each individual gene. In mouse GV oocytes, the median length of poly(A) tails for individual genes is ~58 nt (Fig. 3b). Because PAIso-seq can also measure full-length cDNAs together with their poly(A) tails, it provides the opportunity to look into the RNA isoform-specific features of RNA poly(A) tails. For example, mRNAs can have different lengths of poly(A) tails for alternative polyadenylation isoforms for the same gene (Fig. 3c), which can also be seen for alternative splicing isoforms[18].

Moreover, PAIso-seq, as well as the FLAM-seq method, have revealed widespread presentation of non-A residues in the body of poly(A) tails and serve as good tools for analyzing these non-A residues within RNA poly(A) tails[18,19]. For example, U, C and G residues can be identified in a good proportion of mRNA poly(A) tails in mouse GV oocytes (Fig. 3d and Table 1). The much higher level of non-A residues in the poly(A) tails detected in mouse GV oocytes over poly(A) spike-ins proves the confidence of PAIso-seq in measuring non-A residues within poly(A) tails (Fig. 3d). Moreover, the level of non-A residues can be different among RNA isoforms of the same gene (Fig. 3e)[18]. Finally,

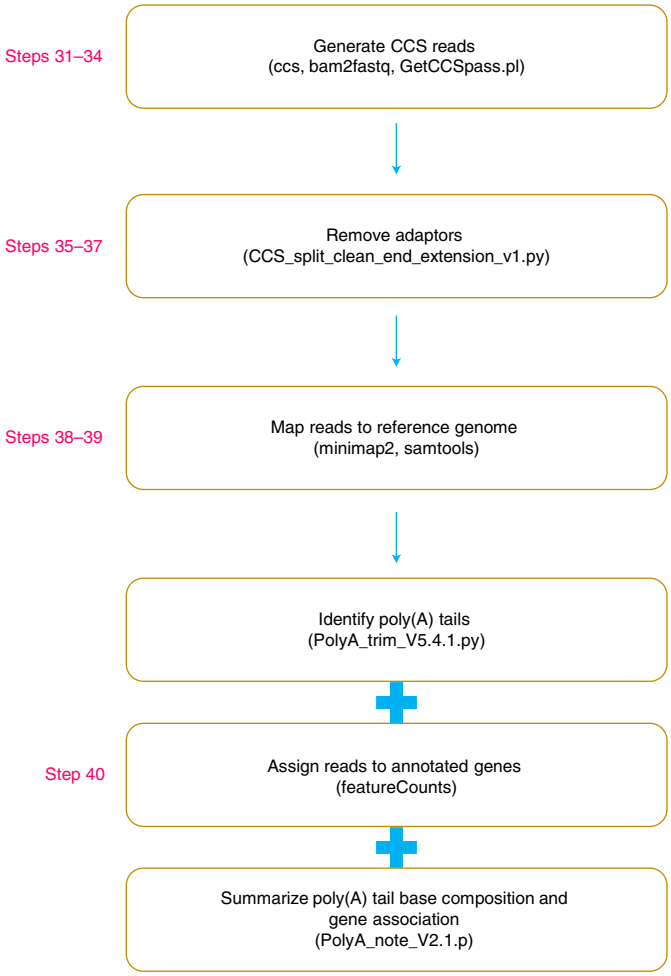


Fig. 2 | Flowchart for the bioinformatic pipeline. The data processing scheme of PAIso-seq data analysis. For each step, the main software and scripts used are shown in the brackets. The corresponding procedure steps are shown in magenta on the left. CCS, circular consensus sequencing.

PAIso-seq has been demonstrated to be powerful enough to analyze single GV oocytes, which contain ~0.5 ng of total RNA per cell[18].

Although mouse GV oocytes are used for method demonstration here, we expect that PAIso-seq could be applied to any species, including mammals such as mice, humans, rats, pigs, cows, dogs, sheep and horses; model organisms such as zebrafish, *Drosophila* and *Caenorhabditis elegans*; plants such as *Arabidopsis*, rice, wheat and maize; and even viruses.

Limitations

First, PAIso-seq uss an 18-nt dT/dU sequence to anneal to target RNA molecules for templated end extension to add a 3′ handle RT and PCR amplification, so it cannot capture transcripts without poly(A) tails and is inefficient in measuring short poly(A) tails (less than ~18 nt). Second, pairing of the very end A nucleotides with the dT/dU template is required for efficient end extension; therefore, RNA molecules with non-A residues at their 3′ ends cannot be efficiently captured. Third, because the last A residue of the RNA poly(A) tail does not always exactly match the junction of oligo(dT/dU) and adapter sequence of the templated end extension oligos, the sequenced poly(A) tail may have a minimal number of additional A residues. Fourth, incomplete RT may happen to RNA with a complex secondary structure, which can lead to truncated transcripts. Fifth, because RNA chemical modification information is lost during the RT and PCR amplification steps, further development is needed for measuring poly(A) tails and assessing their chemical modifications concurrently.

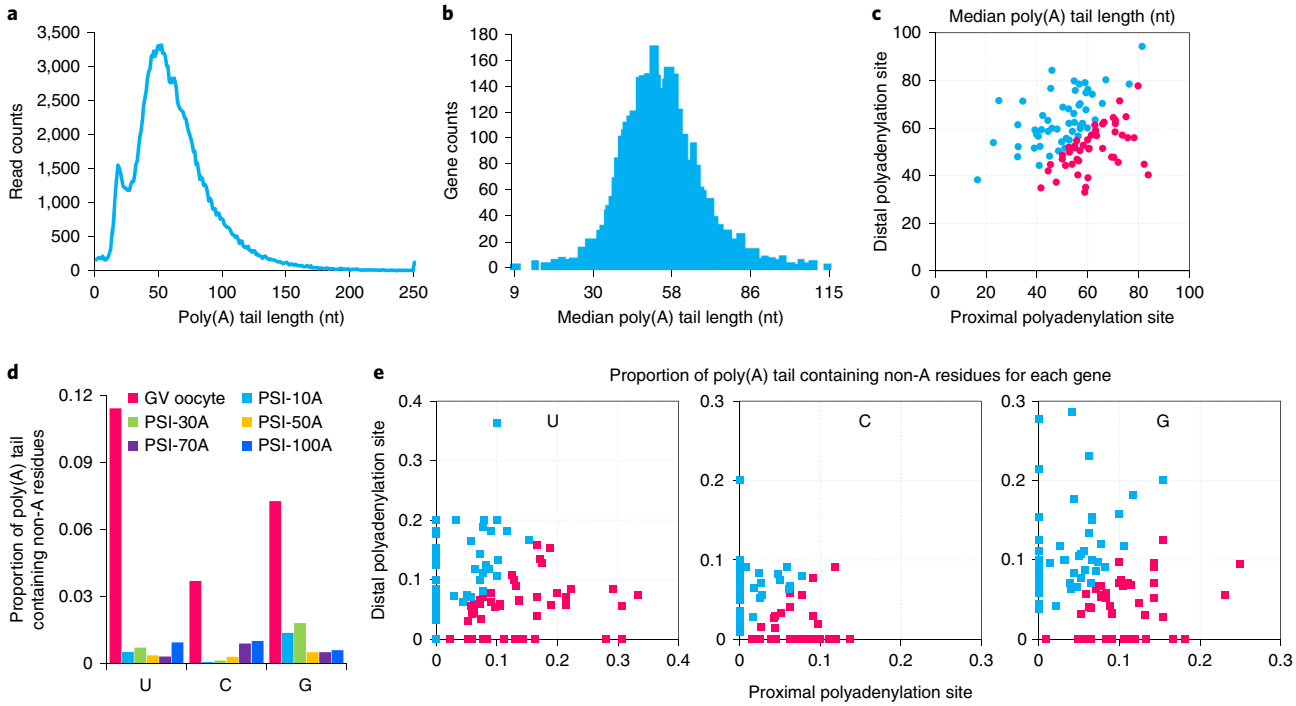


Fig. 3 | Application of PAIso-seq. a, Global distribution of poly(A) tail lengths of all transcripts (CCSs) in GV oocytes (all data combined as one replicate). Read counts (y axis) is the number of CCS reads with poly(A) tails at given length. Reads with pass number ≥10 and poly(A) tail length ≥1 nt are included. The bin size of the histogram is 1 nt. Reads with poly(A) tail length >250 nt are assigned to the 250-nt bin. **b**, The distribution of median poly(A) tail lengths of genes with ≥10 detected poly(A) tails (all data combined as one replicate). Gene counts (y axis) is the number of genes with poly(A) tails at given length (geometric mean of all the detected transcripts with pass number ≥10, transcript number ≥10 and poly(A) tail length ≥1 associated with each of the genes). The bin size of the histogram is 1 nt. **c**, Scatter plot for poly(A) tail length of proximal and distal isoforms of genes with different alternative polyadenylation sites. Genes (n = 49) with longer poly(A) tails in proximal isoforms are shown in magenta, while genes (n = 53) with longer poly(A) tails in distal isoforms are shown in cyan. The poly(A) tail length is the geometric mean of all the detected transcripts associated with the given polyadenylation site. Genes with ≥10 detected poly(A) tails for both proximal and distal polyadenylation sites are analyzed here. **d**, The proportion of poly(A) tails with non-A residues (U, C or G) in GV oocyte and poly(A) spike-in data (all data combined as one replicate). The proportion (y axis) shows the number of poly(A) tails containing the non-A residues divided by the total number of poly(A) tails for each sample. PSI-10A, PSI-30A, PSI-50A, PSI-70A and PSI-100A are poly(A) spike-ins with a poly(A) tail sequence of 10, 30, 50, 70 and 100 nt, respectively. **e**, Scatter plot for the proportion of poly(A) tails with U, C or G residues for proximal and distal isoforms of genes with different alternative polyadenylation sites. Genes (n = 50 for U, n = 36 for C and n = 49 for G) with a higher proportion of poly(A) tails with non-A residue in proximal isoforms are shown in magenta, and genes (n = 50 for U, n = 34 for C and n = 46 for G) with a higher proportion of poly(A) tails with non-A residue in distal isoforms are shown in cyan. Genes with ≥10 detected poly(A) tails are analyzed here. Figure 3a,b adapted with permission from ref. [18]. The data presented here are combined data of the GV_rep2 and SCGV dataset, which are available at the Genome Sequence Archive hosted by the National Genomic Data Center (https://ngdc.cncb.ac.cn/gsa/) under accession number CRA005547.

Experimental design

Choice between starting with total RNA purification or a direct cell lysate

Purified total RNA is the default input material to start with for PAIso-seq. PAIso-seq has sub-nanogram sensitivity, which makes it suitable for precious samples with very limited input materials. However, if a total RNA purification step is used for the low-input samples, it will cause substantial sample loss. For these samples, we prefer to use a direct cell lysate as the starting material. Therefore, for the samples with up to ~10 ng of total RNA (up to 1,000 mammalian somatic cells and 10 mammalian oocytes), we recommend starting PAIso-seq with a direct cell lysate. However, if the input sample is too much, impurities in the cell lysate will inhibit the enzymatic reactions and cause RNA degradation. Thus, for samples with larger amounts of input RNA, we recommend starting PAIso-seq with purified total RNA. Here, we use mouse GV oocytes as an example for PAIso-seq, using both total RNA extracted from bulk oocytes and direct single oocyte lysis. In addition, we expect that RNA prepared from other fractionation or purification methods, such as cytoplasmic and nuclear RNA fractionation[41,42], will also fit the PAIso-seq protocol well.

Capture of the poly(A) tail

PAIso-seq uses templated end extension with templated end-extension oligos, which contain 16 dT and 2 dU at the 3′ end and can pair with the 3′ end of RNA poly(A) sequences to serve as a template

Table 1 | Examples of the poly(A) tail sequences from the mouse GV oocyte sample[18]

Gene	Poly(A) sequence (5′→3′)	CCS ID
Btg4	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAAAAAAAAAAAAA	16777603
Btg4	UUUAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAA	66847194
Btg4	AAAAAAAAAAAAAAUUUAAAAAAAAAAAAAAAAAAAAAAAA	4391575
Btg4	UUUUAAAAAAAAAAAAAAA	51773986
Btg4	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	4522945
Btg4	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	5309206
Ooep	AAAAAAAAAAAAAAAAAAAAAAUUAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	4588230
Ooep	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAAAAAAAA	24314598
Ooep	AAAAAAAAAUUUUAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	44958205
Ooep	AAAAAAAAUUUUAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAA	71435079
Ooep	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	5505833
Ooep	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAA	5898906
Tle6	AAAAAAAAAUAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	9372506
Tle6	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	17433001
Tle6	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAAAAA	47842031
Tle6	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	45940936
Tle6	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAA	10682467
Tle6	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	11338669
Tle6	AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	14484244

Some non-A residues (U, C and G) in the tails are visible. The genes of the mRNAs associated with these tails are shown in column 1, and the CCS IDs for the corresponding reads from the PacBio HiFi sequencing are shown in column 3.

for end extension to add a 3′-end handle after the RNA poly(A) tail for subsequent RT and PCR amplification (Fig. 1). Therefore, the complete poly(A) tail sequence is preserved. This reaction takes place only on poly(A)$^+$ RNAs, avoiding the need for poly(A)$^+$ RNA enrichment. Uracil-specific excision reagent (USER enzyme) is then used to digest the templated end-extension oligos at the dU sites, because the templated end-extension oligos can impair RT and PCR amplification if not removed. The templated end-extension oligo sequences are included in Table 2 and consist of three main parts. The 5′-end part is a 25-nt sequence identical to the 4–28 nt in the template-switching oligo (TSO) described in the next section, which will be used later in RT to pair with the RT primer (RT primer in Table 2) and in PCR amplification to pair with a single PCR primer for both ends of cDNA (IS PCR primer in Table 2). The middle part contains a 16-nt barcode sequence that can be used to assign reads to different samples from multiplexed sequencing. The barcode sequences suggested by PacBio (https://github.com/PacificBiosciences/Bioinformatics-Training/blob/master/barcoding/pacbio_384_barcodes.fasta) are used here. The 3′-end part is the 2 dU and 16 dT nucleotides mentioned above to pair with the end of RNA poly(A) tails.

cDNA 5′ handle addition by template switching

RT can add a few non-template nucleotides (mostly C) to the 3′ end of cDNA when reaching the 5′ end of the RNA template. Therefore, a TSO with several G residues at its 3′-end can be used to pair with the non-templated C residues at the end of cDNA, which can direct the reverse transcriptase to switch templates and continue cDNA synthesis to the end of the oligo[43]. This process called 'template

Table 2 | Sequences of oligos used in this protocol

Name	Sequence	Purpose
TSO	5′-iCiGiCAAGCAGTGGTATCAACGCAGAGTACATrGrG+G/NH_2-C3-3′	To facilitate template switching in RT (used in Steps 8 and 9)
PAT_end-extend_2dU-BC1	5′- AAGCAGTGGTATCAACGCAGAGTAC**TCAGACGATGCGTCAT**dUTTTTTTTTdUTTTTTTTT -3′	Templated end-extension oligos used in templated end extension (used in Steps 6A(vii), 6A(viii) and 6B(vii))
PAT_end-extend_2dU-BC2	5′- AAGCAGTGGTATCAACGCAGAGTAC**CTATACATGACTCTGC**dUTTTTTTTTdUTTTTTTTT-3′	
PAT_end-extend_2dU-BC3	5′- AAGCAGTGGTATCAACGCAGAGTAC**TACTAGAGTAGCACTC**dUTTTTTTTTdUTTTTTTTT-3′	
PAT_end-extend_2dU-BC4	5′- AAGCAGTGGTATCAACGCAGAGTAC**TGTGTATCAGTACATG**dUTTTTTTTTdUTTTTTTTT-3′	
PAT_end-extend_2dU-BC5	5′-AAGCAGTGGTATCAACGCAGAGTAC**GATCTCTACTATATGC**dUTTTTTTTTdUTTTTTTTT-3′	
PAT_end-extend_2dU-BC6	5′-AAGCAGTGGTATCAACGCAGAGTAC**ACAGTCTATACTGCTG**dUTTTTTTTTdUTTTTTTTT-3′	
PAT_end-extend_2dU-BC7	5′-AAGCAGTGGTATCAACGCAGAGTAC**ATGATGTGCTACATCT**dUTTTTTTTTdUTTTTTTTT-3′	
PAT_end-extend_2dU-BC8	5′-AAGCAGTGGTATCAACGCAGAGTAC**CTGCGTGCTCTACGAC**dUTTTTTTTTdUTTTTTTTT-3′	
PAT_end-extend_2dU-BC9	5′-AAGCAGTGGTATCAACGCAGAGTAC**GCGCGATACGATGACT**dUTTTTTTTTdUTTTTTTTT-3′	
PAT_end-extend_2dU-BC10	5′-AAGCAGTGGTATCAACGCAGAGTAC**CGCGCTCAGCTGATCG**dUTTTTTTTTdUTTTTTTTT-3′	
PAT_end-extend_2dU-BC11	5′-AAGCAGTGGTATCAACGCAGAGTAC**GCGCACGCACTACAGA**dUTTTTTTTTdUTTTTTTTT-3′	
PAT_end-extend_2dU-BC12	5′-AAGCAGTGGTATCAACGCAGAGTAC**ACACTGACGTCGCGAC**dUTTTTTTTTdUTTTTTTTT-3′	
PAT_end-extend_2dU-BC13	5′-AAGCAGTGGTATCAACGCAGAGTAC**CGTCTATATACGTATA**dUTTTTTTTTdUTTTTTTTT-3′	
PAT_end-extend_2dU-BC14	5′-AAGCAGTGGTATCAACGCAGAGTAC**ATAGAGACTCAGAGCT**dUTTTTTTTTdUTTTTTTTT-3′	
PAT_end-extend_2dU-BC15	5′-AAGCAGTGGTATCAACGCAGAGTAC**TAGATGCGAGAGTAGA**dUTTTTTTTTdUTTTTTTTT-3′	
RT primer	5′-AAGCAGTGGTATCAACGCAGAGTAC-3′	RT (used in Step 7)
IS PCR primer	5′-AAGCAGTGGTATCAACGCAGAG-3′	PCR preamplification and large-scale PCR amplification (used in Steps 11 and 24)
PSI-0A-F	5′- AAGCAGTGGTATCAACGCAGAGTAC**GCACATACACGCTCAC**ATGGTGAGCAAGGGCGAGGAGGATAAC-3′	PCR primers used in poly(A) spike-in preparation (used in Step 1 in Box 1)
PSI-0A-R	5′- AAGCAGTGGTATCAACGCAGAGTACTCACTTGTACAGCTCGTCCATGCCGCCGGTG-3′	
PSI-10A-F	5′- AAGCAGTGGTATCAACGCAGAGTAC**GCTCGTCGCGCGCACA**ATGGTGAGCAAGGGCGAGGAGGATAAC-3′	
PSI-10A-R	5′-AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTCACTTGTACAGCTCGTCCATGCCGCCGGTG-3′	
PSI-30A-F	5′- AAGCAGTGGTATCAACGCAGAGTAC**ACAGTGCGCTGTCTAT**ATGGTGAGCAAGGGCGAGGAGGATAAC-3′	
PSI-30A-R	5′- AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCACTTGTACAGCTCGTCCATGCCGCCGGTG-3′	
PSI-50A-F	5′- AAGCAGTGGTATCAACGCAGAGTAC**TCACACTCTAGAGCGA**ATGGTGAGCAAGGGCGAGGAGGATAAC-3′	
PSI-50A-R	5′- AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCACTTGTACAGCTCGTCCATGCCGCCGGTG-3′	
PSI-70A-F	5′- AAGCAGTGGTATCAACGCAGAGTAC**TCACATATGTATACAT**ATGGTGAGCAAGGGCGAGGAGGATAAC-3′	
PSI-70A-R	5′- AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCACTTGTACAGCTCGTCCATGC-3′	
PSI-100A-F	5′- AAGCAGTGGTATCAACGCAGAGTAC**CGCTGCGAGAGACAGT**ATGGTGAGCAAGGGCGAGGAGGATAAC-3′	
PSI-100A-R	5′- AAGCAGTGGTATCAACGCAGAGTACTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTCACTTGTACAGCTCGTCCATGCC-3′	

Barcode sequences in oligos are shown in bold letters. +G, LNA-modified guanosine; /NH_2-C3-3′, 3′-aminated.

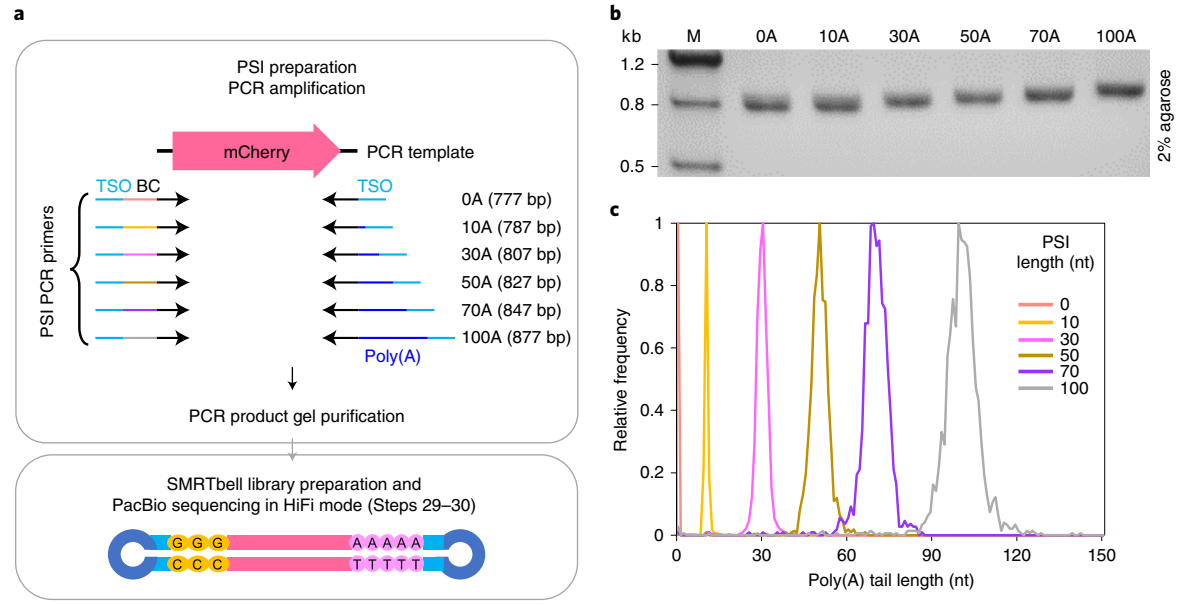


Fig. 4 | Accurate measurement of poly(A) tail length validated by PSIs. a, The PSIs can be prepared by direct PCR amplification with distinct 5′ primers containing barcodes (BC) for read identification and 3′ primers containing a different length of poly(A) sequence (PSI-F and PSI-R PCR primers, Table 2). After PCR and gel purification, the PSIs can be made into a SMRTbell library so that they are ready for spiking into the PAIso-seq runs for PacBio HiFi sequencing. **b**, The PSIs checked in 2% (wt/vol) agarose gel. This gel image has been published in Fig. 1b in ref. [18]. TSO refers to the following sequence: AAGCAGTGGTATCAACGCAGAGTAC. The template for PCR amplification is cDNA encoding mCherry described in ref. [18]. **c**, Histogram of the sequenced poly(A) tail length for the PSIs. Figure adapted with permission from ref. [18].

switching' has been widely used for efficient full-length cDNA synthesis and amplification[44–47] and is used to add the 5′ handle to the cDNA in PAIso-seq.

The TSO (Table 2) carries two iso-deoxycytosine (iC) and one iso-deoxyguanosine (iG) residues at the 5′ end, two riboguanosines (rG) and one locked nucleic acid–modified guanosine (+G) at the 3′ end. The rGrG+G is included at the 3′ end to increase the affinity of the TSO toward the non-templated CCC residues added to the ends of RT products, which will increase the success rate of template switching[47]. The iCiGiC at the 5′ end of the TSO can inhibit the reverse transcriptase from extending the cDNA beyond the TSO, which avoids the formation of concatamers of TSOs and increases cDNA yield[48]. Moreover, the oligo is 3′-aminated (/NH_2-C3-3′) to block it from working as an RT primer in RT. The current version of PAIso-seq does not include a unique molecular identifier sequence, which would be a potentially beneficial addition for molecular counting but would need additional tests in future PAIso-seq updates.

Synthetic poly(A) spike-in to evaluate the performance of PAIso-seq (optional)

To evaluate the accuracy of PacBio sequencing in measuring poly(A) sequences, poly(A) spike-ins (PSIs) consisting of a set of DNAs with different-length poly(A) tails can be prepared (Fig. 4a and Box 1). These PSIs are synthesized by PCR amplification of the same template DNA with different sets of PSI primers (PSI-*A-F and PSI-*A-R in Table 2). The '*' in the PSI primer name refers to the number of A residues for the PSI. The forward primer PSI-*A-F contains a 16-nt unique barcode sequence suggested by PacBio for identification of the reads after multiplexed sequence. The reverse primer PSI-*A-F contains the desired number of A residues added to the PSIs. We expect that RNA poly(A) tail spike-ins can be prepared by chemical synthesis and made into PAIso-seq double-stranded cDNA by following standard PAIso-seq double-stranded cDNA preparation.

Choice among long-read sequencers

PAIso-seq double-stranded cDNA can be made into SMRTbell libraries and sequenced on a PacBio Sequel or Sequel II system under HiFi mode, which can accurately handle long homopolymers[32–35] and enables highly accurate generation of circular consensus sequencing (CCS) reads. One Sequel or Sequel II SMRT cell can yield ~0.3 million or 3 million CCS reads, respectively. The choice between Sequel and Sequel II depends on the amount of output data needed. The Sequel II system delivers

Box 1 | (Optional) PSI preparation ● Timing 2 h

Procedure (Fig. 4a)

▲ **CRITICAL** The PSIs are synthesized by direct PCR amplification by using primers containing the desired number of A residues.

PCR amplification ● Timing 1 h 15 min

1 Prepare six PCR mixes with PSI PCR primers (PSI-0A-F and PSI-0A-R, PSI-10A-F and PSI-10A-R, PSI-30A-F and PSI-30A-R, PSI-50A-F and PSI-50A-R, PSI-70A-F and PSI-70A-R and PSI-100A-F and PSI-100A-R; Table 2) by combining and mixing the following components:

Component of PCR mix	Volume (μl)	Final concentration
KOD-Plus-Neo (1 U/μl)	1	0.02 U/μl
Buffer for KOD-Plus-Neo (10×)	5	1×
dNTPs (2 mM each)	5	0.2 mM each
$MgSO_4$ (25 mM)	3	1.5 mM
Forward primer (10 μM)	1.5	0.3 μM
Reverse primer (10 μM)	1.5	0.3 μM
Template DNA	1	1 ng/μl
Nuclease-free water	32	-
Total volume	50	-

2 Perform the PCR with the following program:

Cycle number	Denature	Anneal	Extend
1	98 °C, 5 min	-	-
2–19	98 °C, 20 s	68 °C (reduced by 1 °C each cycle), 15 s	72 °C, 70 s
20–40	98 °C, 20 s	61 °C, 15 s	72 °C, 70 s
41	-	-	72 °C, 10 min

Gel recovery of the above PCR product ● Timing 45 min

▲ **CRITICAL** Use Zymoclean gel DNA recovery kit for gel recovery.

3 Mix 25 μl of the PCR product with 5 μl of gel loading dye (6×) and run on a 1% (wt/vol) agarose gel at 7 V/cm at room temperature for 20 min.

4 Excise the DNA fragment from the agarose gel by using a blade, transfer it into a 1.5-ml microcentrifuge tube, add 3 volumes of agarose dissolving buffer (ADB) to each volume of the gel and incubate at 55 °C for 5–10 minutes until the gel slice is completely dissolved.

5 Transfer the melted agarose solution to a Zymo-Spin column in a collection tube, centrifuge at 10,000*g* for 30 s at room temperature and discard the flow-through.

6 Add 200 μl of DNA Wash Buffer to the column and centrifuge at 10,000*g* for 30 s at room temperature. Discard the flow-through. Repeat the wash step once.

7 Add 20 μl of nuclease-free water directly to the column matrix. Place the column into a 1.5-ml tube and centrifuge at 10,000*g* for 30 s at room temperature to elute DNA.

8 Quantify the concentration with a spectrophotometer and the size with agarose gel (Fig. 4b). Mix an equal amount of each differently sized PCR product together as the PSIs.

▲ **CRITICAL** The PSIs can be used the same as PAIso-seq double-stranded cDNA for subsequent SMRTbell library preparation and PacBio sequencing as shown in Steps 29 and 30.

▲ **CRITICAL** The PSIs can be mixed with the PAIso-seq double-stranded cDNA to be sequenced at ~1% (wt/wt) of the amount for the subsequent SMRTbell library preparation.

■ **PAUSE POINT** The PSIs can be stored at −20 °C for ≥12 months.

~8–10 times the number of output CCS reads at around three times the total cost, which means about three times the total cost per cell while one-third the cost per read for Sequel II compared to Sequel. The preferred number of CCS reads is ≥1 million for a bulk sample and ≥50,000 for a single mammalian oocyte. For other low-input samples, the number of CCS reads to be sequenced needs to be tested before investing a large amount of sequencing power. Other third-generation sequencers such as Oxford Nanopore can also be used to estimate poly(A) tail length, although not as accurately as PacBio CCS reads. However, the information about non-A residues within the body of poly(A) tails will be lost if sequenced on a Nanopore platform.

Data processing and analysis

The raw PacBio HiFi data from the sequencer are sequences of subreads in bam format. CCS reads are produced by combining multiple subreads (the number of subreads for a CCS read is called the

'pass number') of the same SMRTbell molecule by using the *ccs* software provided by PacBio to achieve an accurate consensus sequence, which is key for the accuracy of PacBio HiFi reads. For accurate quantification of poly(A) tail length and base composition, we use a high threshold requiring ≥10 passes for the CCS reads, which means that the single molecular sequence has been sequenced ≥10 times for the consensus sequence calling, to ensure the accuracy of poly(A) tails[49,50].

When preparing the SMRTbell library, hairpin loop adapters are added by ligation, resulting in a small fraction of the library with multiple PAIso-seq double-stranded cDNAs to be ligated into a single SMRTbell circular molecule. Therefore, one CCS read may contain multiple PAIso-seq full-length cDNAs, which need to be split on the basis of the adapters added to the 5′ ends and the 3′ ends to achieve clean reads. The above considerations are integrated in our bioinformatic analysis pipeline with the pass number information included in the output file for each of the reads (Fig. 2).

Materials

Biological materials

Seven- to eight-week-old female CD1 (ICR) mice were purchased from Beijing Vital River Laboratory Animal Technology and housed in individually ventilated cage systems with no more than five mice per cage in specific pathogen-free rooms **! CAUTION** All mouse procedures are performed in compliance with the guidelines of the Animal Care and Use Committee of the Institute of Genetics and Development Biology, Chinese Academy of Sciences.

Reagents

- Pregnant mare serum gonadotropin (PMSG; ProSpec, cat. no. HOR-272)
- M2 medium (Sigma, cat. no. M7167)
- RNaseZap (Ambion, cat. no. AM9780)
- DNA-OFF (TaKaRa, cat. no. 9036)
- PBS, pH 7.4, RNase-free (10×; Invitrogen, cat. no. AM9625)
- BSA (Sigma-Aldrich, cat. no. A1933)
- Triton X-100 (Sigma-Aldrich, cat. no. T9284) **! CAUTION** Triton X-100 is harmful if swallowed. It causes serious eye damage. When handling this reagent, wear protective gloves, eye protection, face protection and a laboratory coat.
- TRIzol reagent (Invitrogen, cat. no. 15596026) **! CAUTION** TRIzol is toxic if swallowed, toxic in contact with skin and toxic if inhaled. It causes severe skin burns and eye damage, is suspected of causing genetic defects, may cause respiratory irritation and may cause damage to organs through prolonged or repeated exposure. When handling this reagent, wear protective gloves, eye protection, face protection and a laboratory coat. In addition, avoid breathing fumes of this reagent.
- Recombinant RNase inhibitor (TaKaRa, cat. no. 2313A)
- Ethanol, 200 proof, molecular biology grade (Sigma-Aldrich, cat. no. E7023) **! CAUTION** Ethanol is highly flammable. When handling this reagent, keep away from heat, sparks, open flames and hot surfaces.
- Nuclease-free water (Invitrogen, cat. no. AM9938)
- dNTP mix (10 mM each; NEB, cat. no. N0447L)
- Klenow fragment 3′→5′ exo$^-$ (NEB, cat. no. M0212L)
- USER enzyme (NEB, cat. no. M5505L)
- RNA Clean & Concentrator-5 kit (Zymo Research, cat. no. R1016)
- $MgCl_2$ (1 M; Invitrogen, cat. no. AM9530G)
- SuperScript II reverse transcriptase (Invitrogen, cat. no. 18064-014)
- SuperScript II first-strand buffer (5×; 250 mM Tris-HCl, pH 8.3 at room temperature, which is between 20 and 28 °C; 375 mM KCl; 15 mM $MgCl_2$; Invitrogen, cat. no. 18064-014)
- DTT (100 mM; Invitrogen, cat. no. 18064-014) **! CAUTION** DTT is harmful if swallowed. It causes skin irritation and serious eye irritation. It may cause respiratory irritation. When handling this reagent, wear protective gloves, eye protection, face protection and a laboratory coat. Avoid breathing fumes.
- Betaine (Sigma-Aldrich, cat. no. 61962)
- KAPA HiFi HotStart ReadyMix (2×; KAPA Biosystems, cat. no. KK2601) **▲ CRITICAL** KAPA HiFi HotStart DNA polymerase is a robust hot start polymerase to achieve efficient amplification with minimal background amplification.

- SPRIselect beads (Beckman Coulter, cat. no. B23318)
- KOD-Plus-Neo DNA polymerase (Toyobo, cat. no. KOD-401)
- Buffer for KOD-Plus-Neo (10×; Toyobo, cat. no. KOD-401)
- $MgSO_4$ (25 mM; Toyobo, cat. no. KOD-401)
- dNTP mix (2 mM each; Toyobo, cat. no. KOD-401)
- Zymoclean gel DNA recovery kit (Zymo Research, cat. no. D4001)
- Direct-zol RNA MicroPrep (Zymo Research, cat. no. R2060)
- DNA ladder (Tiangen, cat. no. MD103)
- Agarose (Vetec, cat. no. V900510)
- Gel loading dye, purple (6×; NEB, cat. no. B7024S)
- GelRed nucleic acid gel stains (Biotium, cat. no. 41003)
- Agilent RNA 6000 Pico kit (Agilent Technologies, cat. no. 5067-1513)
- Agilent Genomic DNA ScreenTape (Agilent Technologies, cat. no. 5067-5365)
- Agilent Genomic DNA reagents (Agilent Technologies, cat. no. 5067-5366)
- SMRTbell template prep kit 1.0-SPv3 (PacBio, cat. no. 100–991-900)
- DNA oligos and modified oligos (see Reagent setup and Table 2). All oligos were ordered from TaKaRa (https://www.takarabiomed.com.cn/) with HPLC purification.

Equipment

- Disposable plastic syringes (Aladdin, cat. no. A2292-01-1000EA)
- Surgical scissors (Aladdin, cat. no. I2942-01-1EA)
- Surgical forceps (Aladdin, cat. no. D3056-01-1EA)
- Needle, lavender, 30-gauge 0.5 inch (Thomas Scientific, cat. no. JG3005X)
- Microcentrifuge tube, 1.5 ml, MaxyClear SnapLock polypropylene, clear, nonsterile (Axygen, cat. no. MCT-150-C)
- Thin-walled PCR tubes with flat cap, 0.2 ml, clear, nonsterile (Axygen, cat. no. PCR-02-C)
- Falcon polystyrene conical tube (50 ml; BD Biosciences, cat. no. 352095)
- Syringe filter, 0.22 μm (Millipore, cat. no. SLGP033RB)
- Mini vortexer (VWR, cat. no. 82019-170)
- Thermal cycler (Eppendorf, Mastercycler nexus X2; another thermal cycler with a heated lid can also be used)
- DynaMag Spin magnet (Invitrogen, cat. no. 12320D)
- DynaMag-96 Side magnet (Invitrogen, cat. no. 12331D)
- Filter tips: 10, 20, 200 and 1,000 μl (Axygen, cat. nos. TF-10-R-S, TF-20-R-S, TF-200-R-S and TF-1000-B-R-S)
- Horizontal electrophoresis gel box (Dingguo, cat. no. DG-31DN)
- Gel imaging system (Tanon, cat. no. Tanon 1600)
- Agilent 2100 Bioanalyzer (Agilent Technologies, cat. no. G2938C)
- Agilent 4200 TapeStation (Agilent Technologies, cat. no. G291BA)
- Fluorometer and spectrophotometer (DeNovix, cat. no. DS-11 FX+)
- A compatible PacBio DNA-sequencing instrument or PacBio sequencing service
- Computational hardware (Computing server or workstation running Linux)

Software

- ccs software (https://github.com/PacificBiosciences/ccs, version 5.0.0)
- pbbam software (https://github.com/PacificBiosciences/pbbam, version 1.0.6)
- subread software (http://subread.sourceforge.net/)
- minimap2 software (https://github.com/lh3/minimap2, version v2.15)
- SAMtools (http://samtools.sourceforge.net, version 1.9)
- Parafly (https://github.com/ParaFly/ParaFly)
- Python 3.7 or above, and the following packages: pysam (https://github.com/pysam-developers/pysam), regex (https://bitbucket.org/mrabarnett/mrab-regex), parasail (https://github.com/jeffdaily/parasail-python), itertools (https://github.com/more-itertools/more-itertools), numpy (https://numpy.org/) and statistics (https://pypi.python.org/pypi/statistics).

Reagent setup

▲ CRITICAL Avoid contamination from RNase and previous amplified cDNA when handling the following reagents.

PBS with 0.1% (wt/vol) BSA

Weigh 50 mg of BSA, dissolve in 5 ml of PBS (10×) and 45 ml of nuclease-free water. Filter the buffer through a 0.22-μm syringe filter. Divide the buffer into 1-ml aliquots and store at −20 °C for ≤6 months.

Cell lysis buffer

Cell lysis buffer contains 0.2% (vol/vol) Triton X-100. For 10 μl of cell lysis buffer, add 0.5 μl of 40-U/μl RNase inhibitor before use. This buffer can be stored at 4 °C for 6 months. Add RNase inhibitor just before use.

TSO

Dissolve the TSO in nuclease-free water to 100 μM and store it in aliquots at −80 °C for ≤6 months. Keep the freeze-thaw cycles to no more than five for each aliquot.

Templated end-extension oligos (PAT_end-extend_2dU-BC*)

These oligos anneal to the 3′ end of RNAs containing a poly(A) tail. 'BC' in the name of each oligo refers to barcode. The '*' in the title of this paragraph refers to any number shown in Table 2. Each oligonucleotide contains two dU residues, which can be cleaved by a USER enzyme (Table 2). Dissolve each oligonucleotide in nuclease-free water to a final concentration of 100 μM. These oligonucleotides can be stored in aliquots at −20 °C for ≤12 months. Keep the freeze-thaw cycles to no more than five for each aliquot.

RT primer

This oligonucleotide acts as an RT primer (Table 2). Dissolve the oligonucleotide in nuclease-free water to a final concentration of 100 μM. This oligonucleotide can be stored in aliquots at −20 °C for ≤12 months.

IS PCR primer

This oligonucleotide serves as a PCR primer (Table 2) in the preamplification and large-scale PCR steps after RT and template-switching reactions. Dissolve the oligonucleotide in nuclease-free water to a final concentration of 10 μM. This oligonucleotide can be stored in aliquots at −20 °C for ≤12 months.

PSI PCR primers (PSI-*A-F and PSI-*A-R)

These oligos act as PCR primers (Table 2) in the PCR amplification step of PSI preparation (Box 1). Dissolve these oligonucleotides in nuclease-free water to a final concentration of 10 μM. These oligonucleotides can be stored at −20 °C for ≤12 months.

Commands for sequencing data processing

The commands used in the protocol should all be run from the Linux shell prompt within a terminal window connected to a Linux server. We encourage the user to create a separate directory (e.g., '*mouse_GV_polyA_analysis/*') to store all the example data and files created by the analysis. All commands are described under the assumption that the user is working in this directory. The commands are to be executed in the Linux shell environment and are prefixed with a '$' character in the following parts.

Downloading and organizing example data

In this protocol, we took data from mouse GV oocytes as an example to illustrate the sequencing data processing. We used raw subread data from the PacBio sequencer (one bulk GV dataset (*GV_rep2.subreads.bam*) and one single GV oocyte dataset (*SCGV.subreads.bam*)) as examples that can be downloaded from the Genome Sequence Archive (GSA) hosted by the National Genomic Data Center (https://ngdc.cncb.ac.cn/gsa/) under the accession number CRA005547. CCS reads containing PSIs (*polyA_spike-ins-1.ccs.fastq.gz* and *polyA_spike-ins-2.ccs.fastq.gz*) were included for optional PSI analysis, which can be downloaded from GSA under the accession number CRA005706, and the

accompanying pass number files (*polyA_spike-ins-1.ccs.pass.txt* and *polyA_spike-ins-2.ccs.pass.txt*) can be downloaded from GitHub (https://github.com/Lulab-IGDB/PAIso-seq_scripts/blob/main/polyA_spike-in_pass_file/).

The user can download these six files by using the following commands:

```
$ wget -c ftp://download.big.ac.cn/gsa/CRA005547/CRR352158/CRR352158.bam
-O SCGV.subreads.bam

$ wget -c ftp://download.big.ac.cn/gsa/CRA005547/CRR352157/CRR352157.bam
-O GV_rep2.subreads.bam

$ wget -c ftp://download.big.ac.cn/gsa/CRA005706/CRR356911/CRR356911.fastq.gz
-O polyA_spike-ins-1.ccs.fastq.gz

$ wget -c ftp://download.big.ac.cn/gsa/CRA005706/CRR356912/CRR356912.fastq.gz
-O polyA_spike-ins-2.ccs.fastq.gz

$ wget -c https://github.com/Lulab-IGDB/PAIso-seq_scripts/blob/main/polyA_spike-in_pass_file/
polyA_spike-ins-1.ccs.pass.txt

$ wget -c https://github.com/Lulab-IGDB/PAIso-seq_scripts/blob/main/polyA_spike-in_pass_file/
polyA_spike-ins-2.ccs.pass.txt
```

The user should get these six files in the current path:
- *SCGV.subreads.bam*
- *GV_rep2.subreads.bam*
- *polyA_spike-ins-1.ccs.fastq.gz*
- *polyA_spike-ins-2.ccs.fastq.gz*
- *polyA_spike-ins-1.ccs.pass.txt*
- *polyA_spike-ins-2.ccs.pass.txt*

▲ **CRITICAL** The user should always get the **.subreads.bam* files from the PacBio sequencing instrument from the sequencing provider. The md5 value is provided from the GSA or the sequencing provider. The md5 value is usually stored in a txt file called 'md5.txt'.

The user should check the integrity of the data files by using the following command:

```
$ md5sum -c md5.txt
```

▲ **CRITICAL** It will give an 'OK' result if the data file is intact. Otherwise, contact the data provider to transfer the correct data again immediately.

Downloading the reference files

The user should download two reference files from GENCODE (https://www.gencodegenes.org/mouse/release_M25.html) into the *mouse_GV_polyA_analysis/genome/* folder by using the following command:

```
$ mkdir genome/

$ wget -c ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M25/gencode.vM25.
primary_assembly.annotation.gtf.gz -P genome/

$ wget -c ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M25/GRCm38.prima
ry_assembly.genome.fa.gz -P genome/
```

Decompress the *gencode.vM25.primary_assembly.annotation.gtf.gz* file by using the following command:

```
$ gunzip genome/gencode.vM25.primary_assembly.annotation.gtf.gz
```

Downloading and installing software

Install all the executable software used in this protocol by using conda (if none already exists):

```
$ conda install -c bioconda pbccs
$ conda install -c bioconda pbbam
$ conda install -c bioconda minimap2
$ conda install -c bioconda samtools=1.9
$ conda install -c bioconda subread
$ conda install -c bioconda parafly
```

Download and unzip fastq-splitter.pl from the developer's website.

```
$ wget -c http://kirill-kryukov.com/study/tools/fastq-splitter/fastq-splitter-0.1.2.zip
$ unzip fastq-splitter-0.1.2.zip
```

Download the customized python scripts and perl scripts used in this protocol from https://github.com/Lulab-IGDB/PAIso-seq_scripts, which include four python scripts and one perl script:

```
$ wget -c https://github.com/Lulab-IGDB/PAIso-seq_scripts/blob/main/CCS_split_clean_end_extension_v1.py
$ wget -c https://github.com/Lulab-IGDB/PAIso-seq_scripts/blob/main/PolyA_trim_V5.4.1.py
$ wget -c https://github.com/Lulab-IGDB/PAIso-seq_scripts/blob/main/PolyA_note_V2.1.py
$ wget -c https://github.com/Lulab-IGDB/PAIso-seq_scripts/blob/main/GetCCSpass.pl
$ wget -c https://github.com/Lulab-IGDB/PAIso-seq_scripts/blob/main/DNA_spikein_extract_2019_NC_V1.3.py
```

The user should get these five scripts in the current path:

- CCS_split_clean_end_extension_v1.py
- PolyA_trim_V5.4.1.py
- PolyA_note_V2.1.py
- GetCCSpass.pl
- DNA_spikein_extract_2019_NC_V1.3.py

Procedure

Preparing mouse GV oocytes ● Timing 48 h 30 min (30 min of hands-on time)

! CAUTION All mouse procedures are performed in compliance with the guidelines of the Animal Care and Use Committee of the Institute of Genetics and Development Biology, Chinese Academy of Sciences. House the mice in individually ventilated cage systems with no more than five mice per cage in specific pathogen-free rooms. The number of mice used per experiment is according to the number of oocytes needed. Around 20 GV oocytes can be expected from one female mouse[51].

1 Intraperitoneally inject 10 U of PMSG to each of the seven- to eight-week-old CD1 (ICR) female mice with disposable plastic syringes.

2 Euthanize mice by carbon dioxide followed by cervical dislocation 48 h after the PMSG injection and collect ovaries with surgical scissors and forceps.

3 Release the GV oocytes from the ovaries with a 30-gauge needle.

4 Wash the GV oocytes three times with M2 medium and three times with PBS with 0.1% (wt/vol) BSA.

▲ CRITICAL STEP The relatively long time for preparing this sample is specific for mouse oocytes, which need superovulation to achieve a relatively large number of cells per animal (~25 GV oocytes from each female ICR mouse). The time for input sample preparation highly depends on the sample type to be analyzed, which can vary from minutes to days.

■ PAUSE POINT For bulk samples, the GV oocytes can either be used immediately after preparation or be stored in Trizol reagent at −80 °C for ≥4 weeks. For single-oocyte analysis, the single oocyte can either be used immediately after preparation or can be stored individually in ~0.5 µl of PBS with 0.1% (wt/vol) BSA and 2 U/µl recombinant RNase inhibitor in 0.2-ml thin-walled PCR tubes at −80 °C for ≤2 weeks. RNA fragmentation will happen if stored longer.

Preparing a clean working bench ● Timing 15 min

5 Clean the bench with DNA-OFF reagent first and then with RNaseZap reagent. Spray the surface of pipettes with RNaseZap.

▲ CRITICAL STEP All the experiments up to the PCR amplification step (Step 12) must be performed on a clean bench free of DNA contamination from previously amplified cDNA and free of RNase to prevent potential unwanted degradation of RNA in the samples. Locating the clean bench in a separate room from handling amplified product is preferred.

Templated end extension ● Timing 2 h 30 min

6 Templated end extension can be achieved by following option A for total RNA (including total RNA isolation from bulk GV oocytes followed by templated end extension and purification of templated end-extended RNA) or option B for single GV oocytes (including cell lysis followed by templated end extension and purification of templated end-extended RNA).

(A) **Total RNA ● Timing 2 h 30 min**

Total RNA isolation from bulk GV oocytes ● Timing 15 min

▲ CRITICAL Use a Direct-zol RNA MicroPrep kit for total RNA isolation. Other methods for total RNA purification that preserve the integrity of the RNA can also be used.

▲ CRITICAL We used GV oocytes as an example; however, the procedures are suitable for other bulk samples.

(i) Count and transfer GV oocytes from Step 4 by using a microcapillary pipette at the lowest possible volume into a 1.5-ml tube containing 50 µl of TRIzol reagent and mix thoroughly.

! CAUTION Perform all mouse procedures in accordance with relevant guidelines and regulations.

▲ CRITICAL STEP Oocytes can be collected in batches. Oocytes in TRIzol can be stored at −80 °C after each collection. Oocytes collected in different batches can be combined together and used for RNA extraction. About 50–100 ng of total RNA can be extracted from 200 oocytes.

(ii) Add additional TRIzol reagent to achieve the volume of 500 µl. Add 500 µl of absolute ethanol and mix thoroughly. Transfer the mixture into a Zymo-Spin IC column in a collection tube, centrifuge at 10,000*g* for 30 s at 4 °C and discard the flow-through.

(iii) Add 400 µl of RNA Wash PreBuffer to the column and centrifuge at 10,000*g* for 30 s at 4 °C. Discard the flow-through.

(iv) Add 700 µl of RNA Wash Buffer to the column and centrifuge at 10,000*g* for 30 s at 4 °C. Discard the flow-through. Repeat this step once.

(v) Add 12 µl of nuclease-free water directly to the column matrix. Place the column into a 1.5-ml tube and centrifuge at 10,000*g* for 60 s at 4 °C to elute RNA.

(vi) Check the quality of the extracted total RNA with an Agilent 2100 Bioanalyzer to measure the RNA integrity number (RIN). With a spectrophotometer, measure the total RNA for concentration based on the A_{260} absorption and for purity based on the A_{260}/A_{280} ratio.

▲ CRITICAL STEP It is good to proceed by using RNA with a RIN >8 and an A_{260}/A_{280} ratio in the range of 1.8–2.0. The RIN value was 9.10 for the bulk GV oocyte sample used in the example here. 50–100 ng of total RNA is expected from ~200 mouse GV oocytes.

■ PAUSE POINT The purified RNA can either be used immediately or stored at −80 °C for ≤1 month.

Templated end extension for bulk samples ● Timing 2 h

▲ CRITICAL Perform all temperature-controlled incubations in a thermal cycler with a heated lid set to 105 °C throughout this protocol.

(vii) For each sample, one templated end-extension oligo is used. Dilute each templated end-extension oligo (Table 2) to 50 µM by adding 10 µl of 100 µM templated end-extension oligo and 10 µl of nuclease-free water to a tube and mix well.

(viii) Add 1 µl of a different templated end-extension oligo to 50 ng of each total RNA sample from Step 6A(v) and then add nuclease-free water to a total volume of 11 µl. Vortex the tubes briefly to mix and then centrifuge at 2,000*g* for 5 s at room temperature to collect the solution at the bottom of the tubes. Incubate the reaction at 80 °C for 5 min and then at 37 °C for 10 min in a thermal cycler with a heated lid. Put the tubes immediately back on ice. The templated end-extension oligo is now annealed to the 3′ end of the poly(A) tail of poly(A)$^+$ RNA molecules.

▲ **CRITICAL STEP** Record the barcode of the templated end-extension oligo for each sample.

(ix) Prepare the following mix for templated end extension reaction by mixing the reagents listed below for the number of samples plus one additional reaction. This table indicates the volume for one sample. Calculate the required volume according to the sample number and mix the components thoroughly.

Component of templated end-extension mix	Volume (μl)	Final concentration
SuperScript II first-strand buffer (5×)	4	1×
DTT (100 mM)	1	5 mM
dNTP mix (10 mM each)	1	0.5 mM each
Recombinant RNase inhibitor (40 U/μl)	1	2 U/μl
Klenow fragment, 3′→5′ exo$^-$ (5 U/μl)	2	0.5 U/μl
Total volume	9	-

(x) Add 9 μl of the templated end-extension mix to each of the samples in Step 6A(viii) to obtain a final reaction volume of 20 μl. Vortex the tubes briefly to mix and then centrifuge at 2,000*g* for 5 s at room temperature to collect the solution at the bottom of the tubes. Incubate the reactions at 37 °C for 1 h and then at 80 °C for 10 min in a thermal cycler with a heated lid. Put the sample tubes back on ice immediately after the completion of the reaction.

(xi) Add 1 μl of USER enzyme and 30 μl of nuclease-free water to each tube. Vortex the tubes briefly to mix and then centrifuge at 2,000*g* for 5 s at room temperature to collect the solution at the bottom of the tubes. Incubate the reaction at 37 °C for 30 min in a thermal cycler with a heated lid. Put the sample tubes back on ice immediately after the completion of the reaction.

▲ **CRITICAL STEP** The USER enzyme (an enzyme mix) is used to digest the templated end-extension oligos at the two dU sites by excision of uracil bases with uracil DNA glycosylase followed by phosphodiester backbone breaking with endonuclease VIII. Inefficient digestion of templated end-extension oligos by the USER enzyme would lead to false RT from the undigested oligos starting from the internal part of poly(A) tails, leading to truncated poly(A) tails. Therefore, ensure that the USER enzyme is fully active; multiple freeze-thaw cycles should be avoided.

■ **PAUSE POINT** The templated end-extended RNA can be stored at −80 °C for ≤1 month.

Purification of templated end-extended RNA from bulk sample ● Timing 15 min

▲ **CRITICAL** Use RNA Clean & Concentrator-5 kit for RNA purification.

(xii) Add 2 volumes of RNA Binding Buffer to each tube. Vortex the tubes briefly to mix and then centrifuge at 2,000*g* for 5 s at room temperature to collect the solution at the bottom of the tubes.

(xiii) Add an equal volume of absolute ethanol. Vortex the tubes briefly to mix and then centrifuge at 2,000*g* for 5 s at room temperature to collect the solution at the bottom of the tubes.

(xiv) Transfer the mixture to the Zymo-Spin IC column in a collection tube and centrifuge at 10,000*g* for 30 s at 4 °C. Discard the flow-through.

(xv) Add 400 μl of RNA Prep Buffer to the column and centrifuge at 10,000*g* for 30 s at 4 °C. Discard the flow-through.

(xvi) Add 700 μl of RNA Wash Buffer to the column and centrifuge at 10,000*g* for 30 s at 4 °C. Discard the flow-through. Repeat this step again.

(xvii) Add 7 μl of nuclease-free water directly to the column matrix and centrifuge at 10,000*g* for 60 s at 4 °C.

■ **PAUSE POINT** The purified templated end-extended RNA can be stored at −80 °C for ≤1 month.

(B) **Single GV oocytes ● Timing 2 h 30 min**

Single oocyte lysis ● Timing 15 min

▲ **CRITICAL** We used single-mouse GV oocytes as an example; however, the procedures are suitable for most other samples with very limited input materials, including but not limited to mammalian oocytes, early embryos or other samples with a small number of cells.

▲ **CRITICAL** Perform all temperature-controlled incubations in a thermal cycler with a heated lid set to 105 °C throughout this protocol.

(i) Prepare cell lysis buffer by adding 4 µl of 2% (vol/vol) Triton X-100 and 2 µl of recombinant RNase inhibitor to 34 µl of nuclease-free water and mix thoroughly.

(ii) Add 2.5 µl of cell lysis buffer into each of the 0.2-ml thin-walled PCR tubes.

(iii) Transfer a single GV oocyte from Step 4 by using a microcapillary pipette at the lowest possible volume (~0.5 µl) into each of the 0.2-ml thin-walled PCR tubes containing 2.5 µl of cell lysis buffer from Step 6B(ii), to a final volume of ~3 µl. For single GV oocytes already frozen in 0.2-ml thin-walled PCR tubes from Step 4, add 2.5 µl of cell lysis buffer into each of the tubes.

! CAUTION Perform all mouse procedures in accordance with relevant guidelines and regulations.

(iv) Vortex the tubes briefly to mix and then centrifuge at 2,000*g* for 5 s at room temperature to collect the solution at the bottom of the tubes.

(v) Incubate the tubes at 85 °C for 5 min, followed by holding at 4 °C. Place the tubes on ice immediately after the temperature reaches 4 °C.

▲ **CRITICAL STEP** The single-oocyte lysate (or other very limited input materials) can be stored in the −80 °C refrigerator after each collection.

■ **PAUSE POINT** The single-oocyte lysate (or lysates from very limited input materials) can either be used immediately or stored at −80 °C for ≤2 weeks.

Templated end extension for single-oocyte samples ● Timing 2 h

(vi) Dilute each templated end-extension oligo as detailed in Step 6A(vii).

(vii) Add 1 µl of templated end-extension oligo (50 µM) and 1.5 µl of nuclease-free water to each sample from Step 6B(v) (3 µl) to a total volume of 5.5 µl. Vortex the tubes briefly to mix and then centrifuge at 2,000*g* for 5 s at room temperature to collect the solution at the bottom of the tubes. Incubate the reaction at 80 °C for 5 min and then at 37 °C for 10 min in a thermal cycler with a heated lid. Put the tubes immediately back on ice. The templated end-extension oligo is now annealed to the 3′ end of the poly(A) tail of poly(A)$^+$ RNA molecules.

▲ **CRITICAL STEP** Record the barcode of the templated end-extension oligo for each sample.

(viii) Prepare the following mix for templated end-extension reaction by mixing the reagents listed below for the number of samples plus one additional reaction. This table indicates the volume for one sample. Calculate the required volume according to the sample number and mix the components thoroughly.

Component of templated end extension mix	Volume (µl)	Final concentration
SuperScript II first-strand buffer (5×)	2	1×
DTT (100 mM)	0.5	5 mM
dNTP mix (10 mM each)	0.5	0.5 mM each
Recombinant RNase inhibitor (40 U/µl)	0.5	2 U/µl
Klenow fragment, 3′→5′ exo$^-$ (5 U/µl)	1	0.5 U/µl
Total volume	4.5	-

(ix) Add 4.5 µl of the templated end-extension mix to each tube from Step 6B(vii) to obtain a final reaction volume of 10 µl. Vortex the tubes briefly to mix and then centrifuge at 2,000*g* for 5 s at room temperature to collect the solution at the bottom of the tubes. Incubate the reactions at 37 °C for 1 h and then at 80 °C for 10 min in a thermal cycler with a heated lid. Put the sample tubes back on ice immediately after the completion of the reaction.

(x) Add 1 µl of USER enzyme and 15 µl of nuclease-free water to each of the tubes. Vortex the tubes briefly to mix and then centrifuge at 2,000*g* for 5 s at room temperature to collect the solution at the bottom of the tubes. Incubate the reactions at 37 °C for 30 min in a thermal cycler with a heated lid. Put the sample tubes back on ice immediately after the completion of the reaction.

■ **PAUSE POINT** The templated end-extended samples can be stored at −80 °C for ≤1 month.

Purification of end-extended RNA from single-oocyte samples ● Timing **15 min**

▲ **CRITICAL** Use RNA Clean & Concentrator-5 kit for RNA purification.

(xi) Mix different single-oocyte samples from Step 6B(x) with different barcodes together into one tube.

▲ **CRITICAL STEP** We mix the templated end-extended product from 15 individual single-mouse GV oocytes, each bearing a different barcode. The users can change the number of samples mixed according to their own experimental design. Make sure that the samples to be mixed together use different barcodes; otherwise, you will not be able to distinguish them in the future.

(xii) Purify the end-extended RNA as described in Step 6A(xii–xvii).

■ **PAUSE POINT** The purified templated end-extended RNA can be stored at −80 °C for ≤1 month.

Reverse transcription ● Timing 3 h

▲ **CRITICAL** From this step, total RNA or single-oocyte samples both follow the same steps.

▲ **CRITICAL** Perform all temperature-controlled incubations in a thermal cycler with a heated lid set to 105 °C throughout this protocol.

7 Add 0.4 µl of RT primer (100 µM; Table 2) and 1 µl of dNTP mix (10 mM each) to the samples from Step 6A(xvii) or Step 6B(xii). Vortex the tubes briefly to mix and then centrifuge at 2,000*g* for 5 s at room temperature to collect the solution at the bottom of the tubes. Incubate the tubes at 72 °C for 3 min in a thermal cycler with a heated lid. Put the tubes immediately back on ice. The RT primer is now annealed to the 3′-end adapter of RNA molecules.

8 Dilute the TSO (Table 2) to 20 µM by adding 10 µl of 100 µM TSO and 40 µl of nuclease-free water to a tube and mix well.

9 Prepare the following mix for the RT reaction by mixing the reagents listed below for the number of samples plus one additional reaction. This table indicates the volume for one sample. Calculate the required volume according to the sample number and mix the components thoroughly.

▲ **CRITICAL STEP** Thaw all the reagents before performing RNA denaturation and prepare the mix during the time of denaturation (Step 7).

Component of RT mix	Volume (µl)	Final concentration
SuperScript II reverse transcriptase (200 U/µl)	1	10 U/µl
SuperScript II first-strand buffer (5×)	4	1×
Recombinant RNase inhibitor (40 U/µl)	0.5	1 U/µl
DTT (100 mM)	1	5 mM
Betaine (5 M)	4	1 M
$MgCl_2$ (1 M)	0.12	6 mM
TSO (20 µM, Step 8)	0.98	0.98 µM
Total volume	11.6	-

10 Add 11.6 µl of the RT mix to each tube to obtain a final reaction volume of 20 µl. Vortex the tubes briefly to mix and then centrifuge at 2,000*g* for 5 s at room temperature to collect the solution at the bottom of the tubes. Incubate the tubes in a thermal cycler with a heated lid following the program below.

Cycle	Temperature (°C)	Time	Purpose
1	42	90 min	RT and template switching
2–11	50	2 min	Unfold the RNA secondary structures
	42	2 min	Continue RT and template switching
12	70	15 min	Inactivate the enzyme
13	4	Hold	Safe storage

▲ **CRITICAL STEP** An additional 10 cycles between 50 °C and 42 °C after the initial incubation at 42 °C can give a slight increase in final cDNA yield[47].

■ **PAUSE POINT** The RT products can be safely stored at −80 °C for ≥2 weeks.

PCR preamplification ● Timing 3 h

▲ **CRITICAL** Perform all PCR in a thermal cycler with a heated lid set to 105 °C throughout this protocol.

11 Prepare the following mix for PCR reaction by mixing the reagents listed below for the number of samples plus one additional reaction. This table indicates the volume for one sample. Calculate the required volume according to the sample number and mix the components thoroughly.

Component of PCR mix	Volume (µl)	Final concentration
KAPA HiFi HotStart ReadyMix (2×)	25	1×
IS PCR primer (10 µM) (Table 2)	5	1 µM
Total volume	30	-

12 Add 30 µl of the above mix to each tube containing 20 µl of first-strand reaction from Step 10. Vortex the tubes briefly to mix and then centrifuge at 2,000*g* for 5 s at room temperature to collect the solution at the bottom of the tubes.

13 Perform the PCR in a thermal cycler with a heated lid by following the program below:

Cycle number	Denature	Anneal	Extend
1	98 °C, 3 min	-	-
2–16	98 °C, 20 s	67 °C, 15 s	72 °C, 6 min
17	-	-	72 °C, 10 min

▲ **CRITICAL STEP** The reaction buffer in the KAPA HiFi HotStart ReadyMix contains a higher salt concentration than regular PCR buffers, which affects DNA melting. Therefore, it is good to perform the denaturation step at 98 °C and use a higher annealing temperature of 67 °C for the IS PCR primer[47]. 15 cycles are normally performed for samples with input material of no more than 50 ng of total RNA, including those single-oocyte samples. For samples with input material more than 50 ng, decrease the cycle number by 1 if the input material amount doubles compared to 50 ng.

■ **PAUSE POINT** PCR products can be safely stored at −20 °C for ≥12 months.

Clean-up and quantification of preamplification product ● Timing 1 h

14 Vortex SPRIselect beads thoroughly to achieve even suspension.

15 Add 0.8× volume (40 µl) of SPRIselect beads and mix by pipetting up and down until the mixture looks homogeneous.

▲ **CRITICAL STEP** Keep the volume of beads in this purification step at 0.8× ratio. This ratio is suitable for the purification of poly(A) inclusive full-length cDNA to avoid very short fragments.

16 Incubate the mixture undisturbed for 8 min at room temperature to bind the DNA to the beads.

17 Place the tubes on a magnetic stand (DynaMag-96 Side magnet) for 5 min. The solution will become clear, with the beads collected at one corner of the tube. Keep the tubes on the magnetic stand and carefully remove most of the liquid without taking any beads by pipetting.

18 Wash the beads with 150 µl of 80% (vol/vol) ethanol solution with the tubes on the magnetic stand. Incubate for 30 s without disturbing the beads and then remove the ethanol. Repeat this step twice.

▲ **CRITICAL STEP** Prepare the 80% (vol/vol) ethanol solution freshly every time.

19 Remove any trace amount of liquid by pipetting and let the beads air-dry at room temperature on the magnetic stand for 5 min or until small cracks are visible on the surface of the beads.

▲ **CRITICAL STEP** Do not overdry the beads. The overdried beads are difficult to resuspend, which leads to reduced yield.

20 Add 20 µl of nuclease-free water to each of the tubes on the magnetic stand. Take the tube off the magnetic stand. Mix 10 times by pipetting to completely resuspend the beads. Incubate the tubes on a regular stand for 2 min to elute the DNA.

21 Place the tubes on the magnetic stand and leave them until the solution becomes clear, with beads collected in a corner of each of the tubes.

22 Collect the liquid, which contains the purified preamplification product, without taking any beads by pipetting and transfer it to a fresh 1.5-ml tube.
▲ **CRITICAL STEP** Do not take all the liquid; leave ~1 µl of the solution in the tube to minimize bead carryover.

23 Measure the DNA concentration of the purified preamplification product by using a fluorometer-based method, such as DeNovix DS-11 FX+ or Qubit 3.0.
▲ **CRITICAL STEP** Preamplification product of ~25 ng can be expected from the combination of 15 single-mouse GV oocytes, while ~200 ng can be expected from 50 ng of total RNA of mouse GV oocytes.
■ **PAUSE POINT** The purified preamplification product can be safely stored at −20 °C for ≥12 months.
? TROUBLESHOOTING

Large-scale PCR ● Timing 3 h

▲ **CRITICAL STEP** Perform all PCR in a thermal cycler with a heated lid set to 105 °C throughout this protocol.

24 Set up 800 µl of large-scale PCR reactions with 20 ng of purified PCR pre-amplification product (Step 23) by combining and mixing the following components thoroughly. This table indicates the volume for one sample.

Component of PCR mix	Volume (µl)	Final concentration
KAPA HiFi HotStart ReadyMix (2×)	400	1×
PCR pre-amplification product	20	-
IS PCR primer (10 µM)	80	1 µM
Nuclease-free water	300	-
Total volume	800	-

▲ **CRITICAL STEP** Use 20 ng of purified preamplification product as input for large-scale PCR; add nuclease-free water if the input volume is <20 µl. If the total amount of purified preamplification product is <20 ng, use all of it as input for large-scale PCR.

25 Vortex the tubes briefly to mix and then centrifuge at 2,000*g* for 5 s at room temperature to collect the solution at the bottom of the tubes. Distribute 50-µl aliquots of this PCR mix into sixteen 0.2-ml thin-walled PCR tubes.

26 Perform the PCR with the following program:

Cycle number	Denature	Anneal	Extend
1	98 °C, 3 min	-	-
2–11	98 °C, 20 s	67 °C, 15 s	72 °C, 6 min
12	-	-	72 °C, 10 min

▲ **CRITICAL STEP** Typically, 10 cycles are used for large-scale PCR for ~20 ng of purified preamplification product. If the input amount is <20 ng, increase the number of cycles for large-scale PCR depending on the amount of input.
■ **PAUSE POINT** PCR products can be safely stored at −20 °C for ≥12 months.

Clean-up of large-scale PAIso-seq double-stranded cDNA ● Timing 45 min

27 Perform the clean-up of large-scale PAIso-seq double-stranded cDNA as shown in Steps 14–22 with 0.8× volume (640 µl) of SPRIselect beads, wash with 900 µl of 80% (vol/vol) ethanol and elute in 100 µl of nuclease-free water.
▲ **CRITICAL STEP** Keep the volume of beads in this purification step at 0.8× ratio. This ratio is suitable for the purification of poly(A) inclusive full-length cDNA to avoid very short fragments. Use a DynaMag Spin magnet for 1.5-ml tubes.
■ **PAUSE POINT** The purified PAIso-seq double-stranded cDNA can be safely stored at −20 °C for ≥12 months.

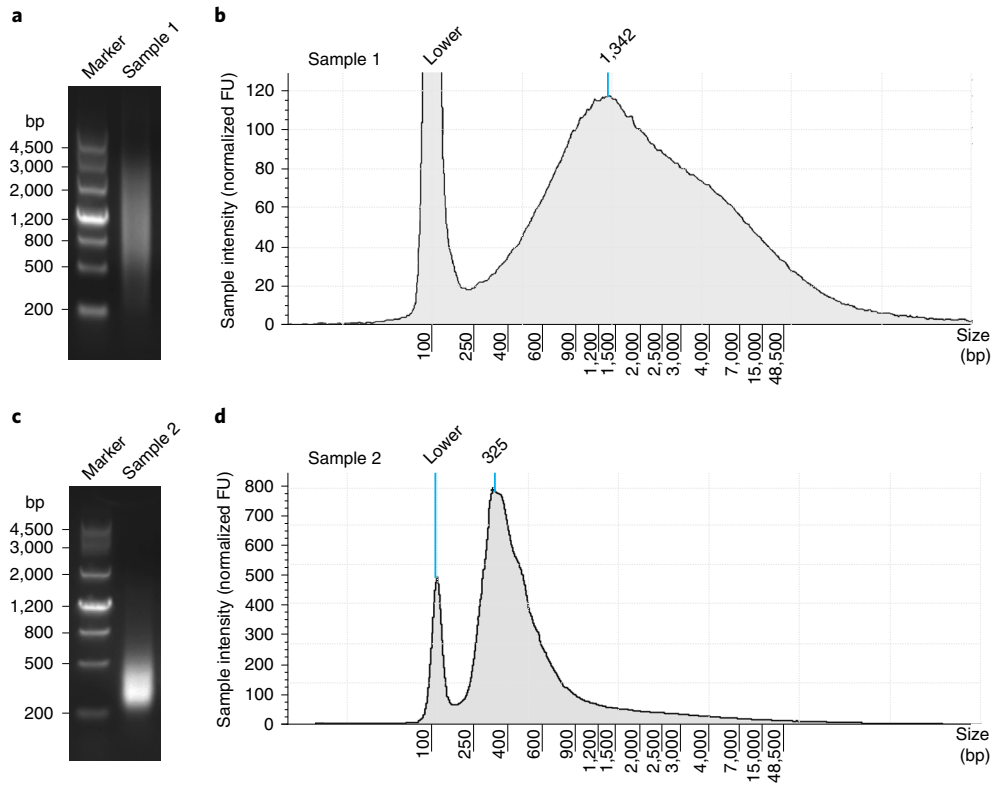


Fig. 5 | Size distribution of PAIso-seq double-stranded cDNA. The PAIso-seq double-stranded cDNA can be checked by either agarose gel electrophoresis (**a** and **c**) or Agilent 4200 TapeStation (**b** and **d**). Sample 1 represents a bulk PAIso-seq double-stranded cDNA sample showing one of the most common profiles, with a peak around 1–2 kb (1,342 bp). Sample 2 represents a single oocyte PAIso-seq double-stranded cDNA from a sample with partially fragmented RNA with a peak of short length (325 bp). 'Lower' indicates the lower marker (100 bp) of the Agilent genomic DNA reagents for Agilent 4200 TapeStation.

Quality check of the PAIso-seq double-stranded cDNA ● Timing 1 h

28 Measure the DNA concentration and quality of the purified cDNA by using a fluorometer-based method, such as DeNovix DS-11 FX+ or Qubit 3.0. The expected yield should be ~1–10 μg. The size distribution of the PAIso-seq double-stranded cDNA can be measured by agarose gel electrophoresis, Agilent TapeStation or other similar tools for analyzing DNA fragment size (Fig. 5). Typical PAIso-seq double-stranded cDNA shows a peak around 1–2 kb (Fig. 5a,b). However, for some precious in vivo samples in which RNA fragmentation cannot be avoided, PAIso-seq can still be performed with the ability to analyze the poly(A) tails while sacrificing its ability to measure full-length RNA isoforms. The size of this type of PAIso-seq double-stranded cDNA is generally shorter (Fig. 5c,d).

■ PAUSE POINT PAIso-seq double-stranded cDNA can be safely stored at −20 °C for ≥12 months.

? TROUBLESHOOTING

SMRTbell library preparation and long-read sequencing on the PacBio platform ● Timing 2 d

29 Perform SMRTbell library construction by using the PAIso-seq double-stranded cDNA with or without PSIs according to the instructions of the SMRTbell template prep kit 1.0-SPv3 with 160–500 ng of input DNA.

30 Sequence the SMRTbell library on a PacBio Sequel or Sequel II system according to the standard PacBio Iso-Seq procedures.

? TROUBLESHOOTING

Data analysis ● Timing 8 h

▲ CRITICAL All the command examples use SCGV in this protocol. For other samples, the GV_rep2 sample for example, change the name of the input and output files accordingly.

Generate CCS reads from subreads ● Timing 2 h

31 Generate highly accurate single-molecule consensus reads by using *ccs* software, as follows:

```
$ ccs SCGV.subreads.bam SCGV.ccs.bam -j 30 &> SCGV.ccs.log
```

▲ **CRITICAL STEP** Converting subreads to CCS is a computation-intensive and time-consuming step. Therefore, it is better to do this computation on a multi-core server in parallel. The number of threads (-j option here) depends on the available CPU cores on the computing server, which we set to 30 here. For the data from one cell of a Sequel HiFi run (the single-cell GV oocyte (SCGV) sample here, for example), it took ~47 CPU h (<2 h when executed in parallel by using 30 CPU cores) in total to generate CCS reads (249,163 successful CCS reads from an input of 438,861 zero-mode waveguides (ZMWs)) on an Intel Xeon 4110 CPU at 2.1 GHz. For the data from one cell of a Sequel II HiFi run, it will take ~10 times the CPU hours compared to that from a Sequel HiFi run. In this case, performing CCS conversion with more computing threads in parallel is recommend. Alternatively, the subreads.bam files can be split into multiple chunks to be submitted as a batch job in a computing cluster to save time.

32 Check the success of the CCS read conversion by using the following command:

```
$ cat SCGV.ccs.log
```

▲ **CRITICAL STEP** It is important to check that the conversion is successful. The SCGV.ccs.log will be empty if the CCS read conversion is successful.

33 Convert the ccs.bam file to a fastq.gz file by using the following command:

```
$ bam2fastq -o SCGV.ccs SCGV.ccs.bam > SCGV.ccs.bam2fastq.log
```

▲ **CRITICAL STEP** Ensure that the input SCGV.ccs.bam.pbi file that is generated in Step 31 together with SCGV.ccs.bam is in the current path; otherwise, it will report an error. If the SCGV.ccs.bam2fastq.log file is empty, it means that no error has happened during the step. An SCGV.ccs.fastq.gz file is expected.

34 Get the number of passes for each of the CCS reads by using the following command:

```
$ perl GetCCSpass.pl SCGV.ccs.bam > SCGV.ccs.pass.txt
```

Extract the clean reads containing the poly(A) inclusive full-length cDNA from the CCS reads on the basis of the barcodes for each sample ● Timing 2 h

35 Prepare files for splitting ccs and cleaning the reads by using the following commands:

```
$ gunzip SCGV.ccs.fastq.gz
$ fastq-splitter.pl --n-parts 100 SCGV.ccs.fastq
$ vim barcode.fa
>BC1
ATGACGCATCGTCTGAGTACTCTGCGTTGATACCACTGCTT
>BC2
GCAGAGTCATGTATAGGTACTCTGCGTTGATACCACTGCTT
>BC3
GAGTGCTACTCTAGTAGTACTCTGCGTTGATACCACTGCTT
>BC4
CATGTACTGATACACAGTACTCTGCGTTGATACCACTGCTT
>BC5
GCATATAGTAGAGATCGTACTCTGCGTTGATACCACTGCTT
>BC6
CAGCAGTATAGACTGTGTACTCTGCGTTGATACCACTGCTT
>BC7
AGATGTAGCACATCATGTACTCTGCGTTGATACCACTGCTT
>BC8
GTCGTAGAGCACGCAGGTACTCTGCGTTGATACCACTGCTT
>BC9
```

```
AGTCATCGTATCGCGCGTACTCTGCGTTGATACCACTGCTT
>BC10
CGATCAGCTGAGCGCGGTACTCTGCGTTGATACCACTGCTT
>BC11
TCTGTAGTGCGTGCGCGTACTCTGCGTTGATACCACTGCTT
>BC12
GTCGCGACGTCAGTGTGTACTCTGCGTTGATACCACTGCTT
>BC13
TATACGTATATAGACGGTACTCTGCGTTGATACCACTGCTT
>BC14
AGCTCTGAGTCTCTATGTACTCTGCGTTGATACCACTGCTT
>BC15
TCTACTCTCGCATCTAGTACTCTGCGTTGATACCACTGCTT
```

▲ **CRITICAL STEP** Extracting the clean reads by using the commands in Step 36 is time consuming. Therefore, we separate the fastq file into multiple (100 here) files for execution in parallel. This will create SCGV.ccs.part-001.fastq, SCGV.ccs.part-002.fastq…and SCGV.ccs.part-100.fastq. The number of parts can be increased or lowered according to the computing resource.

▲ **CRITICAL STEP** The barcode.fa file contains the barcode information for replicates of the SCGV sample in fasta format. The user should create the barcode.fa file according to the actual barcode sequences used for the samples.

36 Extract clean reads from the CCS reads by using the following commands:

```
$ for i in {001..100}; do echo 'python CCS_split_clean_end_extension_v1.py SCGV.ccs.part-'${i}'.fastq SCGV.ccs.pass.txt barcode.fa 2 1> SCGV. '${i}'.out.txt 2> SCGV.'${i}'.err.txt'; done > split_clean.sh
$ ParaFly -c split_clean.sh -CPU 25
```

▲ **CRITICAL STEP** This step extracts full-length cDNA sequences on the basis of the 3′-end barcode sequence and the 5′-end TSO sequence and outputs the poly(A) inclusive full-length cDNA sequence in the 5′-end-to-3′-end orientation with both 5′ adapter and 3′ adapter removed. The parameter 2 means the number of mismatches allowed for matching barcodes in the barcode.fa file with reads, for which we recommend using no more than 4. The *.out.txt files contain clean reads, while the *.err.txt files contain reads with incomplete structure or ccs containing no barcodes, which are kept for potential troubleshooting. If the sequencing data contain reads of the PSIs, follow the procedure in Box 2 to get clean poly(A) sequences of PSIs.

37 Convert the clean reads into fastq format by using the following commands:

```
$ cat *.out.txt > SCGV.all.out.txt
$ cat *.err.txt > SCGV.all.err.txt
$ awk '{print "@"$2"\n"$6"\n+\n"$7}' SCGV.all.out.txt | gzip -nc > SCGV.clean.fastq.gz
$ rm SCGV.*.err.txt SCGV.*.out.txt *fastq.log.txt SCGV.ccs.part-*.fastq
```

? TROUBLESHOOTING

Align the clean-reads to the genome ● Timing 2 h

38 Generate reference files for genome alignment by using the following commands:

```
$ paftools.js gff2bed genome/gencode.vM25.primary_assembly.annotation.gtf > gencode.vM25.primary_assembly.annotation.bed
$ minimap2 -x splice -t 20 -d gencode.vM25.mmi genome/GRCm38.primary_assembly.genome.fa.gz
```

▲ **CRITICAL STEP** The paftools.js is a script provided by the *minimap2* software that converts the gtf format annotation file into the bed format needed for read mapping using *minimap2*.

▲ **CRITICAL STEP** This step generates the index for *minimap2* from the reference genome sequence.

Box 2 | (Optional) Extract clean poly(A) sequences of PSIs from the CCS reads

Procedure

1 Prepare a barcode file for PSI extraction by using the following command:

```
$ vim PSI-barcode.fa
>PSI-0A
GCACATACACGCTCAC
> PSI-10A
GCTCGTCGCGCGCACA
> PSI-30A
ACAGTGCGCTGTCTAT
> PSI-50A
TCACACTCTAGAGCGA
> PSI-70A
TCACATATGTATACAT
> PSI-100A
CGCTGCGAGAGACAGT
```

2 Extract clean poly(A) sequences of PSIs from the CCS reads by using the following command:

```
$ python DNA_spikein_extract_2019_NC_V1.3.py polyA_spike-ins-1.ccs.fastq
polyA_spike-ins-1.ccs.pass.txt PSI-barcode.fa 2 1 1> PSI.out.txt 2> PSI.err.txt
```

▲ **CRITICAL STEP** This step extracts PSI sequences on the basis of the barcode sequence and the PSI-specific sequence from the CCS reads and outputs the poly(A) sequence in the 5′-end-to-3′-end orientation. The input CCS reads are converted from raw subreads the same way as described in Steps 31–33. The polyA_spike-ins-2 can be processed the same way by changing the names of the input and output files. The parameter 2 means the number of mismatches allowed for matching barcodes in the PSI-barcode.fa file with reads, for which we recommend using no more than 4. The PSI.out.txt files contain clean reads, while the PSI.err.txt files contain reads with incomplete structure or CCS containing no barcodes, which are kept for potential troubleshooting.

3 The length of the poly(A) tails of these spike-ins can then be plotted, which shows sharp peaks around the expected size (Fig. 4c).

39 Map the clean reads to the genome by using the following command:

```
$ minimap2 -ax splice -uf --secondary=no -t 40 -L --MD --cs --junc-bed
gencode.vM25.primary_assembly.annotation.bed gencode.vM25.mmi SCGV.clean.
fastq.gz 2> align.log | samtools view -F 3844 -bS > SCGV.clean.filter.bam
```

▲ **CRITICAL STEP** This step maps the clean reads to the reference genome, allowing splicing to get the mapped clean reads in bam format.

Extract and annotate poly(A) tails from mapped clean-reads ● Timing 2 h

40 Extract and annotate poly(A) tails from clean reads by using the following command:.

```
$ python PolyA_trim_V5.4.1.py SCGV.clean.filter.bam > SCGV.polyA_trim.
out.txt
$ featureCounts -L -g gene_id -t exon -s 1 -R CORE -a genome/gencode.
vM25.primary_assembly.annotation.gtf -o SCGV.featureCounts SCGV.clean.
filter.bam &> featureCounts.log
$ python PolyA_note_V2.1.py SCGV.polyA_trim.out.txt SCGV.clean.filter.
bam.featureCounts 1> SCGV.polyA_note.txt 2> SCGV.polyA_note.err.txt
```

▲ **CRITICAL STEP** The 3′ terminal clipped sequence of the mapped clean reads in the bam file is used as a candidate poly(A) tail sequence. These candidate poly(A) tails are filtered by using 'PolyA_trim_V5.4.1.py'. The mapped clean reads are assigned to each annotated gene by using the *featureCounts* software. Then, the poly(A) tail information for each mapped clean read that can be uniquely assigned to annotated genes is summarized by using 'PolyA_note_V2.1.py'.

▲ **CRITICAL STEP** The *featureCounts* software is provided by the *subread* package. The user will get three files in this step: SCGV.featureCounts, SCGV.featureCounts.summary and SCGV.clean.filter.bam.featureCounts.

▲ **CRITICAL STEP** If the SCGV.polyA_note.err.txt file is empty, it means that no error has happened during the step; otherwise, an error message will be reported.

▲ **CRITICAL STEP** SCGV.polyA_note.txt contains all the poly(A) tail information for the reads assigned a poly(A) tail (including reads without tails that give tail size as 0 nt and the sequence as 'No'), which will be the main starting file for further downstream analysis. Each line in this file contains the following 13 columns of information for one read: barcode, read_id, ensembl_id, pass_number, '1', number_of_residue_A, number_of_residue_T, number_of_residue_C, number_of_residue_G, number_of_residue_T+C+G, '0', poly(A)_tail_sequence and average_quality_value_of_poly(A)_tail_bases.

Troubleshooting

Troubleshooting advice can be found in Table 3.

Table 3 | Troubleshooting table

Step	Problem	Possible reason	Solution
23	The yield of preamplification is much lower than expected	The input RNA was too low	Increase the amount of starting material or increase the PCR cycles for preamplification
		Cells were dead at the time of collection	Use cells in good condition
		Failed templated end extension or RT	Change new aliquots of reagents for templated end extension and RT
28	The overall double-stranded cDNA size is short	RNA fragmentation	Use RNase-free aliquots of reagents and consumables
			Use fresh samples or samples stored at −80 °C for ≤2 weeks. For some rare samples in which RNA fragmentation cannot be avoided due to the nature of sample collection, the cDNA can still be sequenced to get the poly(A) tail information while sacrificing the full-length RNA isoform information
30	The sequencing output is too low	Failed PacBio HiFi run because of a defective SMRT cell	Run the remaining SMRTbell library on a new SMRT cell under HiFi mode
37	Some barcodes for single-cell samples with no or very low number of reads	Cells were dead or damaged	Use cells in good condition and analyze more single cells if possible

Timing

Preparing mouse GV oocytes

Steps 1–4, preparing mouse GV oocytes: 48 h 30 min

PAIso-seq double-stranded cDNA preparation

Step 5, preparing a clean working bench: 15 min
Step 6, templated end extension: 2 h 30 min
Steps 7–10, RT: 3 h
Steps 11–13, PCR preamplification: 3 h
Steps 14–23, clean-up and quantification of preamplification product: 1 h
Steps 24–26, large-scale PCR: 3 h
Steps 27–28, PAIso-seq double-stranded cDNA clean-up and quality control: 1 h 45 min

PSI preparation (optional)

Box 1, poly(A) spike-in preparation: 2 h

Sequencing on a PacBio platform

Steps 29–30, SMRTbell library construction and sequencing: 2 d

Sequencing data processing

Steps 31–34, convert subreads to ccs: 2 h
Steps 35–37, split ccs and clean the reads: 2 h
Steps 38–39, align the clean reads to the genome: 2 h
Step 40, extract and annotate poly(A) tails from clean reads: 2 h

Anticipated results

Step 6A(vi)

The concentration and quality of total RNA extracted from bulk samples can be assessed with an Agilent 2100 Bioanalyzer to measure the RIN value and a spectrophotometer to measure the concentration and the A_{260}/A_{280} ratio. Typically, the user can expect a yield of 50–100 ng of total RNA from ~200 mouse GV oocytes, with the RIN value >8 and the A_{260}/A_{280} ratio in the range of 1.8–2.0.

Step 23

After PCR preamplification and purification, the user can measure the quantity of cDNA product. This is normally performed with a fluorometer-based method, such as DeNovix DS-11 FX+ or Qubit 3.0. Typically, 15 cycles of PCR preamplification can yield ~25 ng of cDNA from 15 single-mouse GV oocytes and ~200 ng of cDNA from 50 ng of total RNA, respectively. Several known common problems with the possible reasons and solutions are included in the Troubleshooting section (Table 3).

Step 28

After large-scale PCR amplification and purification, the user can determine the overall quality of the PAIso-seq double-stranded cDNA. This is normally performed with an Agilent TapeStation. Typically, after 10 cycles of large-scale PCR from ~20 ng of preamplification product, the user can expect a yield of ~1–10 μg of final PAIso-seq double-stranded cDNA. The size of the peak of the transcripts in mouse GV oocytes should be ~1–2 kb as evaluated by agarose gel or Agilent 4200 TapeStation (Fig. 5a,b). Several known common problems with the possible reasons and solutions are included in the Troubleshooting section (Table 3).

Steps 31–40

A typical PAIso-seq sequencing run composed of 15 individual single-mouse GV oocytes sequenced on a PacBio Sequel platform, which we have published previously[18], yielded ~249,000 raw CCS reads, from which ~173,000 clean reads were extracted. Among these clean reads, ~133,000 reads were mapped to the mouse genome, and 131,000 reads were assigned with poly(A) tail information, of which 93,000 belonged to nuclear protein coding genes covering 8,732 nuclear protein–coding genes. Currently, the sequencing cost per CCS read is about one-fifth compared to the per-read cost in our previous published study due to the introduction of the new PacBio sequel II system[18]. Therefore, more comprehensive transcriptome-wide poly(A) tail information can be achieved at a reasonable cost now.

The user can get comprehensive information about transcriptome-wide poly(A) tail information for each sequenced read, including the pass number of the CCS read, the full poly(A) tail sequence, an accurate length of the poly(A) tail, the number of non-A residues in the poly(A) tail and the gene that the read is assigned to.

Data availability

The data used in this study for bioinformatic analysis are from our published dataset[18]. The PAIso-seq CCS reads in fastq format we previously deposited[18] are available in the NCBI Sequence Read Archive under the accession number PRJNA529588. For this protocol, we uploaded the raw subread data for the GV_rep2 and SCGV datasets, which had not been deposited before, to the GSA hosted by the National Genomic Data Center (https://ngdc.cncb.ac.cn/gsa/) under the accession number CRA005547. The CCS read data containing the PSIs in fastq format are available in GSA under the accession number CRA005706, and the accompanying pass number files are available in GitHub (https://github.com/Lulab-IGDB/PAIso-seq_scripts/blob/main/polyA_spike-in_pass_file/).

Code availability

Custom scripts used for data analysis are available at GitHub: https://github.com/Lulab-IGDB/PAIso-seq_scripts.

References

1. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**, 66–71 (2014).
2. Lim, J. et al. Uridylation by TUT4 and TUT7 marks mRNA for degradation. *Cell* **159**, 1365–1376 (2014).

3. Lim, J. et al. Mixed tailing by TENT4A and TENT4B shields mRNA from rapid deadenylation. *Science* **361**, 701–704 (2018).
4. Lim, J., Lee, M., Son, A., Chang, H. & Kim, V. N. mTAIL-seq reveals dynamic poly(A) tail regulation in oocyte-to-embryo development. *Genes Dev.* **30**, 1671–1682 (2016).
5. Ma, J., Fukuda, Y. & Schultz, R. M. Mobilization of dormant Cnot7 mRNA promotes deadenylation of maternal transcripts during mouse oocyte maturation. *Biol. Reprod.* **93**, 48 (2015).
6. Kumar, A., Clerici, M., Muckenfuss, L. M., Passmore, L. A. & Jinek, M. Mechanistic insights into mRNA 3′-end processing. *Curr. Opin. Struct. Biol.* **59**, 143–150 (2019).
7. Eckmann, C. R., Rammelt, C. & Wahle, E. Control of poly(A) tail length. *Wiley Interdiscip. Rev. RNA* **2**, 348–361 (2011).
8. Charlesworth, A., Meijer, H. A. & de Moor, C. H. Specificity factors in cytoplasmic polyadenylation. *Wiley Interdiscip. Rev. RNA* **4**, 437–461 (2013).
9. Sha, Q. Q. et al. CNOT6L couples the selective degradation of maternal transcripts to meiotic cell cycle progression in mouse oocyte. *EMBO J.* **37**, e99333 (2018).
10. Yu, C. et al. BTG4 is a meiotic cell cycle-coupled maternal-zygotic-transition licensing factor in oocytes. *Nat. Struct. Mol. Biol.* **23**, 387–394 (2016).
11. Pasternak, M., Pfender, S., Santhanam, B. & Schuh, M. The BTG4 and CAF1 complex prevents the spontaneous activation of eggs by deadenylating maternal mRNAs. *Open Biol.* **6**, 160184 (2016).
12. Liu, Y. et al. BTG4 is a key regulator for maternal mRNA clearance during mouse early embryogenesis. *J. Mol. Cell Biol.* **8**, 366–368 (2016).
13. Costa-Mattioli, M., Sossin, W. S., Klann, E. & Sonenberg, N. Translational control of long-lasting synaptic plasticity and memory. *Neuron* **61**, 10–26 (2009).
14. Huang, Y. S., Jung, M. Y., Sarkissian, M. & Richter, J. D. *N*-methyl-D-aspartate receptor signaling results in Aurora kinase-catalyzed CPEB phosphorylation and alpha CaMKII mRNA polyadenylation at synapses. *EMBO J.* **21**, 2139–2148 (2002).
15. Wu, L. et al. CPEB-mediated cytoplasmic polyadenylation and the regulation of experience-dependent translation of alpha-CaMKII mRNA at synapses. *Neuron* **21**, 1129–1139 (1998).
16. Alarcon, J. M. et al. Selective modulation of some forms of schaffer collateral-CA1 synaptic plasticity in mice with a disruption of the CPEB-1 gene. *Learn. Mem.* **11**, 318–327 (2004).
17. Chang, H., Lim, J., Ha, M. & Kim, V. N. TAIL-seq: genome-wide determination of poly(A) tail length and 3′ end modifications. *Mol. Cell* **53**, 1044–1052 (2014).
18. Liu, Y., Nie, H., Liu, H. & Lu, F. Poly(A) inclusive RNA isoform sequencing (PAIso-seq) reveals wide-spread non-adenosine residues within RNA poly(A) tails. *Nat. Commun.* **10**, 5292 (2019).
19. Legnini, I., Alles, J., Karaiskos, N., Ayoub, S. & Rajewsky, N. FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nat. Methods* **16**, 879–886 (2019).
20. Zhao, T. et al. Impact of poly(A)-tail G-content on *Arabidopsis* PAB binding and their role in enhancing translational efficiency. *Genome Biol.* **20**, 189 (2019).
21. Eisen, T. J. et al. The dynamics of cytoplasmic mRNA metabolism. *Mol. Cell* **77**, 786–799.e10 (2020).
22. Eisen, T. J., Eichhorn, S. W., Subtelny, A. O. & Bartel, D. P. MicroRNAs cause accelerated decay of short-tailed target mRNAs. *Mol. Cell* **77**, 775–785.e8 (2020).
23. 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* **21**, 1502–1510 (2015).
24. Woo, Y. M. et al. TED-seq identifies the dynamics of poly(A) length during ER stress. *Cell Rep.* **24**, 3630–3641.e7 (2018).
25. Yu, F. et al. Poly(A)-seq: a method for direct sequencing and analysis of the transcriptomic poly(A)-tails. *PLoS One* **15**, e0234696 (2020).
26. Parker, M. T. et al. Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m^6A modification. *eLife* **9**, e49658 (2020).
27. Workman, R. E. et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).
28. Roach, N. P. et al. The full-length transcriptome of *C. elegans* using direct RNA sequencing. *Genome Res.* **30**, 299–312 (2020).
29. Kim, D. et al. The architecture of SARS-CoV-2 transcriptome. *Cell* **181**, 914–921.e10 (2020).
30. Long, Y., Jia, J., Mo, W., Jin, X. & Zhai, J. FLEP-seq: simultaneous detection of RNA polymerase II position, splicing status, polyadenylation site and poly(A) tail length at genome-wide scale by single-molecule nascent RNA sequencing. *Nat. Protoc.* **16**, 4355–4381 (2021).
31. Jia, J. et al. Post-transcriptional splicing of nascent RNA contributes to widespread intron retention in plants. *Nat. Plants* **6**, 780–788 (2020).
32. Laehnemann, D., Borkhardt, A. & McHardy, A. C. Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Brief. Bioinform.* **17**, 154–179 (2016).
33. Ross, M. G. et al. Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51 (2013).
34. Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* **12**, R112 (2011).
35. Hebert, P. D. N. et al. A Sequel to Sanger: amplicon sequencing that scales. *BMC Genomics* **19**, 219 (2018).
36. Zhang, Y. et al. Alternative polyadenylation: methods, mechanism, function, and role in cancer. *J. Exp. Clin. Cancer Res.* **40**, 51 (2021).

37. Morgan, M., Kumar, L., Li, Y. & Baptissart, M. Post-transcriptional regulation in spermatogenesis: all RNA pathways lead to healthy sperm. *Cell. Mol. Life Sci.* **78**, 8049–8071 (2021).
38. Liudkovska, V. & Dziembowski, A. Functions and mechanisms of RNA tailing by metazoan terminal nucleotidyltransferases. *Wiley Interdiscip. Rev. RNA* **12**, e1622 (2021).
39. Kandhari, N., Kraupner-Taylor, C. A., Harrison, P. F., Powell, D. R. & Beilharz, T. H. The detection and bioinformatic analysis of alternative 3′ UTR isoforms as potential cancer biomarkers. *Int. J. Mol. Sci.* **22**, 5322 (2021).
40. Yu, S. & Kim, V. N. A tale of non-canonical tails: gene regulation by post-transcriptional RNA tailing. *Nat. Rev. Mol. Cell Biol.* **21**, 542–556 (2020).
41. Hu, S. B. et al. Protein arginine methyltransferase CARM1 attenuates the paraspeckle-mediated nuclear retention of mRNAs containing IRAlus. *Genes Dev.* **29**, 630–645 (2015).
42. Wang, Y. et al. Genome-wide screening of NEAT1 regulators reveals cross-regulation between paraspeckles and mitochondria. *Nat. Cell Biol.* **20**, 1145–1158 (2018).
43. Matz, M. et al. Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Res.* **27**, 1558–1560 (1999).
44. Ramskold, D. et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
45. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
46. Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. & Siebert, P. D. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* **30**, 892–897 (2001).
47. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
48. Kapteyn, J., He, R., McDowell, E. T. & Gang, D. R. Incorporation of non-natural nucleotides into template-switching oligonucleotides reduces background and improves cDNA synthesis from very small RNA samples. *BMC Genomics* **11**, 413 (2010).
49. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
50. van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The third revolution in sequencing technology. *Trends Genet.* **34**, 666–681 (2018).
51. Luo, C. et al. Superovulation strategies for 6 commonly used mouse strains. *J. Am. Assoc. Lab. Anim. Sci.* **50**, 471–478 (2011).

Acknowledgements

We thank Hu Nie for help in the bioinformatic analysis. This work was supported by the National Key Research and Development Program of China (2018YFA0107001), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA24020203), the National Natural Science Foundation of China (31970588, 32170606 and 81891001), the Natural Science Foundation of Heilongjiang province (YQ2020C003), the China Postdoctoral Science Foundation (2020M670516 and 2020T130687) and the State Key Laboratory of Molecular Developmental Biology.

Author contributions

Y.L., J.W. and F.L. conceived and designed the study. Y.L. developed the method and performed the experiments. Y.Z. performed part of the bioinformatic analysis. Y.L., J.W. and F.L. designed the computational pipeline, analyzed the data and wrote the manuscript.

Competing interests

Y.L. and F.L. are named inventors on a patent (number: 201910837492.2) filed by the Institute of Genetics and Developmental Biology covering the PAIso-seq method. The other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Yusheng Liu, Jiaqiang Wang, Falong Lu.

Peer review information *Nature Protocols* thanks Qingshun Quinn Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Additional initial assessment was performed by informal referee Andrzej Dziembowski.

Received: 29 December 2021; Accepted: 23 March 2022;
Published online: 13 July 2022

Related links
Key reference using this protocol
Liu, Y. et al. *Nat. Commun.* **10**, 5292 (2019): https://doi.org/10.1038/s41467-019-13228-9