CLT-seq as a universal homopolymer-sequencing concept reveals poly(A)-tail-tuned ncRNA regulation

Qiang Su[1], Yi Long[2], Jun Wang[1] and Deming Gou*[1]

[1] Shenzhen Key Laboratory of Microbial Genetic Engineering, Vascular Disease Research Center, College of Life Sciences and Oceanography, Guangdong Provincial Key Laboratory of Regional Immunity and Disease, Shenzhen University, 1066 Xueyuan Street, Nanshan District, Shenzhen, 518055, Guangdong, China.

[2] The School of Medicine, Life and Health Sciences (MED | LHS), Chinese University of Hong Kong, Shenzhen, No. 2001 Longxiang Blvd., Longgang District, Shenzhen, 518172, Guangdong, China.

* To whom correspondence should be addressed. Email: dmgou@szu.edu.cn

Running title: Poly(A) tail profiling using CLT-seq

ABSTRACT

Dynamic poly(A)-tail tuning is a central mechanism to control eukaryotic mRNA translation and stability. A more comprehensive understanding of poly(A)-tail-tuned regulation requires an unbiased abundance quantification of the poly(A)-tail length-sorted transcripts on the common-used high-throughput sequencing platforms. However, current methods have limitations due to each method-specified complicated setups and elaborate library-preparing plan-introduced biases. Here we describe the Central Limit Theorem (CLT)-managed RNA-seq (CLT-seq) as a simple homopolymer-sequencing concept. Since the anchor-free oligo(dT) equiprobably bound to and rapidly unbound from any low-affinity binding complementary segment of the poly(A) tail till it was primed by a coincident-loading transcriptase, the CLT mechanism enables the CLT-shortened poly(T) lengths that correspond to poly(A) tail to distribute normally. Based on the well-fitted pseudoguassion-derived poly(A)-poly(T) conversion model, the actual poly(A)-tail profile is reconstructed from the acquired poly(T)-length profile through matrix operations. The CLT-seq followed a simple procedure without pre-treatment, enrichment, and selection of the total RNA and the CLT-shortened poly(T) stretches are more compatible with existing sequencing platforms. This conceptual approach facilitates the straightforward homopolymer base-calling and features the unbiased RNA-seq. Therefore, the CLT-seq provides unbiased, robust, and cost-efficient transcriptome-wide poly(A)-tail profiling. We demonstrate that the CLT-seq on the most common Illumina platform delivers the high-quality poly(A)-tail profiling at a transcriptome-wide scale in human cellular contexts. We find that the poly(A)-tail-tuned ncRNA regulation underwent a dynamic complex process as the mRNA.

INTRODUCTION

Dynamic poly(A)-tail tuning in all eukaryotic cells is a central mechanism to control both mRNA translation and stability.(Garneau et al. 2007; Passmore and Coller 2022) Likewise, polyadenylation also occurs in non-coding RNAs (ncRNAs).(Doma and Parker 2007; Schaffer et al. 2019) Yet despite the importance of poly(A) tail in biological process, there has been only a partial mechanism of poly(A)-tail-regulated mRNA expression unveiled to

date.(Nicholson and Pasquinelli 2019) Furthermore, poly(A)-tail-tunned ncRNA regulation remains largely unknown. Such limited information is primarily due to the difficulty in base-calling the homopolymeric sequences.(Quail et al. 2012b)

Some creative approaches have been developed to address the homopolymeric sequencing problem. Two pioneering methods, TAIL-seq(Chang et al. 2014) and PAL-seq(Subtelny et al. 2014), open an era of transcriptome-wide poly(A)-tail profiling using high-throughput sequencing. A machine-learning Tailseeker program was applied in TAIL-seq for base-calling A-residue in homopolymer row.(Chang et al. 2014; Lim et al. 2016; Lima et al. 2017) An internal standard and calibration curve method in PAL-seq was utilized to estimate poly(A)-tail length.(Subtelny et al. 2014; Eisen et al. 2020) Further advancements in long-read sequencing by Nanopore(Garalde et al. 2018; Nicholson-Shaw et al. 2022) and PacBio(Legnini et al. 2019; Liu et al. 2019; Mattijssen et al. 2020; Long et al. 2021) also handled homopolymeric sequences through repeatedly correcting reading-errors. All these methods have indeed expanded our understanding of the dynamic nature of the mRNA poly(A) tail. However, the method-specified complicated setups limit the applicability of each method to various scenarios. For instance, the Tailseeker 1/2 programs are only applicable to the Illumina 4-channel system, instead of a 2-channel system even with more efficient sequencing. And the Nanopore and PAL-seq methods provide only the length of the poly(A) tail rather than its specific sequence.

Here, we describe the Central Limit Theorem-managed RNA-seq (CLT-seq) as a universal and simple homopolymer sequencing concept that facilitated the straightforward poly(A)-tail base-calling and featured the unbiased RNA-seq. Since the anchor-free oligo(dT) equiprobably bound to and rapidly unbound from any low-affinity binding complementary segment of the poly(A) tail till it was primed by a coincident-loading transcriptase, the CLT mechanism arranges the poly(A)-tail template-directed poly(T)-synthesis length to distribute normally, upon which the CLT-shortened poly(T) stretches can be sequenced straightforwardly and reliably. Based on the well-fitted pseudoguassion-derived

poly(A)-poly(T) conversion model, the poly(A)-tail profile is reconstructed from the acquired poly(T)-length profile through matrix operations. Moreover, the CLT-seq enables to mitigate artificial and reverse-transcription biases for transcript-abundance estimation. We validated the CLT-seq with simulation-measurement comparisons, Ligation-Mediated Poly(A) Tail (LM-PAT) Assays, and statistical analyses. By applying the CLT-seq approach on the most common Illumina system, we comprehensively profiled the poly(A) tail of mitochondrial transcriptome as a sample from the full sets of whole transcriptomes and found that poly(A)-tail-tuned ncRNA regulation underwent a dynamic process as complex as that of mRNA.

RESULTS

Principle and feasibility of CLT-seq

To profile poly(A) tail in multiple cellular contexts, we developed CLT-seq following a proposed simple procedure without pre-treatment, enrichment, and selection (Fig. 1, a). First, an adaptor-flanked anchor-free oligo(dT) randomly anneals to poly(A) tail for the 1st cDNA synthesis. Then, after excess primer depletion, the resulting cDNA couples with the second adaptor through a splint-ligation approach. Finally, the fully equipped cDNA serves as a template for PCR-based library amplification. Since the anchor-free oligo(dT) binds equiprobably to and rapidly unbind from any complementary segment of the poly(A) tail till it is primed by a coincident-loading transcriptase, the CLT mechanism enables the randomly primed poly(T) lengths to distribute normally. When considering the variable sample size that determines the number of the randomly selected oligo(dT)-binding events, a pseudoguassion distribution function (Function in Method) can convert the measured poly(A) length into the CLT-shortened poly(T) length for modelling purposes. Upon this well-fitted poly(A)-poly(T) conversion model, the actual poly(A)-tail profile is reconstructed through mathematical operations from the acquired poly(T)-length distribution.

To model this pseudoguassion-based conversion, a set of poly(A) tails is generated by poly(A) polymerase-catalysed spike-in RNA polyadenylation under time-controlled incubation. We measured the extended poly(A) lengths precisely with Agilent 2100 bioanalyzer (Fig. 1, b).

Further, we characterized the poly(A)-tailed spike-ins and their corresponding 1st cDNAs primed from non-anchored oligo(dT_{18}) and anchored oligo(dT_{18}VNN) by electrophoresis (Supplemental Fig. S1, a). Compared with the anchored oligo(dT) priming at the fixed position on the poly(A)-tailed spike-in template, the non-anchored oligo(dT) primed randomly along the poly(A) tail, leading to diffuse 1st-cDNA components. Through PCR amplification, libraries based on oligo(dT_{12}) (Fig. 1, c), oligo(dT_{18}), and oligo(dT_{30}) (Supplemental Fig. S1, b) displayed a poly(A)-tail-dependent library-shifting tendency but with different diffusion for the long poly(A) tails, attesting CLT's controlled mechanism. To assess the outcome of CLT-seq, the replicated experiments for a series of spike-ins gave a good correlation of poly(T) profile in the Wilcoxon rank-sum test (Fig. 1, d). As the poly(A)-tail length increased, the poly(T) profile shifted to the right along with curve flattening in the poly(T)-length spectrum, which the Kruskal-Wallis ANOVA had confirmed ($p = 3.9 \times 10^{-87}$ (df=9)).

In the homopolymer-length determination, the acquired sequencing lengths followed an empirical distribution where the symmetric length-distribution profile was flattened as the measured homopolymeric length increased.(Chang et al. 2014) The mean length in the sequencing-length distribution should be equal to the actual homopolymer size, theoretically. However, in practice, the mean-length determination needs to consider the effect of the acquired-length dispersion (standard deviation). Since the Illumina sequencing chemistry could handle homopolymers relatively well,(Bragg et al. 2013) we applied the two-channel NovaSeq instrument that has a lower error rate and less variation than most of Illumina platforms to determine the poly(A)-tail length.(Ma et al. 2019; Stoler and Nekrutenko 2021) It is known that the read quality in homopolymer regions decreased rapidly on the Illumina platform because of the stochastic (pre-) phasing (that usually occurs per ~15 sequencing cycles).(Whiteford et al. 2009; Ledergerber and Dessimoz 2011) The quality scores here indicate only the homopolymer length-determining accuracy instead of the canonical base calling-error rate. The stochastic (pre-) phasing recurs throughout the sequencing, where the regular lagging and decaying A-base signal leads to a predictable sequencing-length

overestimation. The standard base-calling algorithm can properly correct this error at an early stage of sequencing. Although the Tailseeker 1/2 base-calling algorithm can significantly improve the homopolymer-sequencing accuracy, a few length-overestimating stretches are still observed in the 64-nt homopolymer sequencing, and the overestimation is more noticeable in the 118-nt-based length determination.(Chang et al. 2014) It is quite reasonable to suppose that the overestimating-occurrence rate rises with increasing the homopolymer length, even with this advanced base-calling algorithm. Nevertheless, there are very reliable sequencing measurements for poly(A) tail even longer than 250 nt. Likewise, the reliable measurements based on the standard base-calling algorithm for the CLT-shortened homopolymer stretches are also achieved in our biological replicates (Fig. 1, d and Supplemental Fig. S2).

The CLT mechanism arranges the CLT-shortened homopolymer lengths to distribute normally. Accordingly, the sequencing-poly(T)-length distribution of a set of poly(A) tail from A_{30} to A_{149} is shown in detail in Supplemental Fig. S2. In the longest A_{149} tail-templated CLT poly(T)-length distribution, an 80-nt (73 and 87 nt) mean length determined by the pseudoguassion function fitting is close to the 75 nt theoretical length. Although the cumulative overestimation could skew the sequencing-length distribution and flatten the length-distribution profile further,(Chang et al. 2014) the proposed pseudoguassion function allows us to account for the skewness and kurtosis effects for modelling purposes. Based on the standard base-calling algorithm, the Illumina homopolymer-sequencing accuracy is insufficient to measure all various lengths of mammalian poly(A) tail,(Nicholson and Pasquinelli 2019) but it is sufficient to profile the CLT-shortened homopolymeric stretches for characterizing.

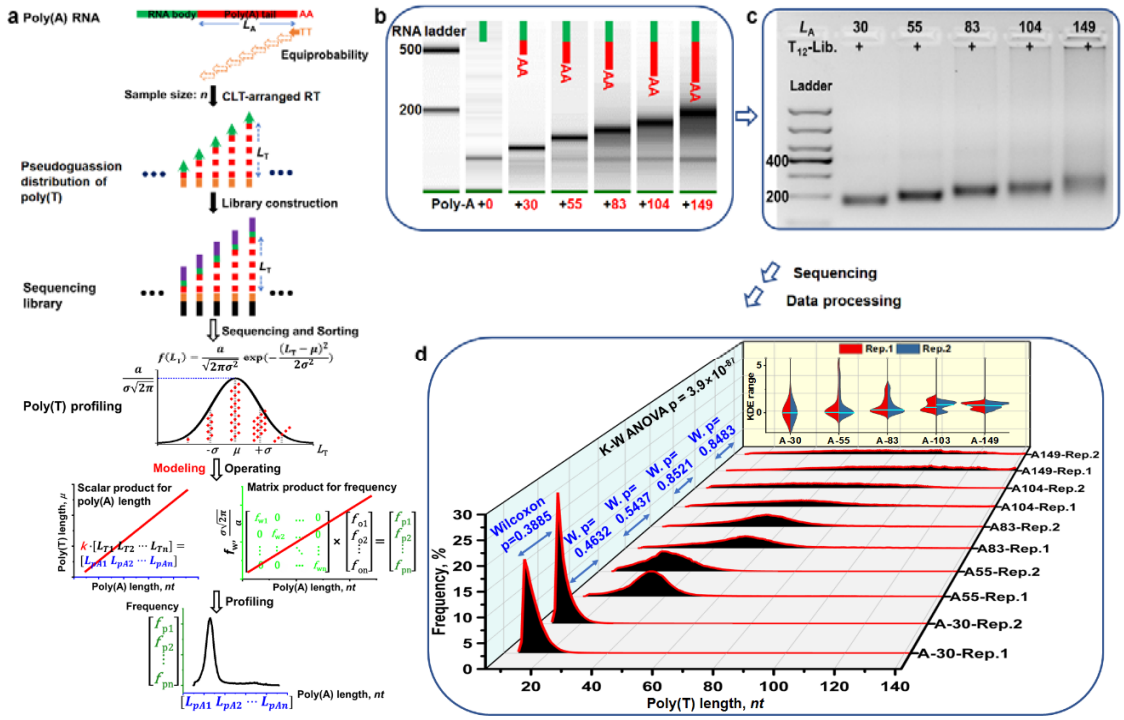

Fig. 1. The principle and feasibility of CLT-seq on the Illumina platform. **a**) Experimental and computational outline of CLT-seq. **b**) Home-made spike-ins with varying poly(A)-tail lengths (red bar). The lengths are measured by Agilent 2100 bioanalyzer with RNA 6000 Pico Kit. **c**) Agarose-gel (3%) electrophoresis of the PCR-amplified libraries. **d**) 3-D plot CLT-arranged poly(T) length against counting frequency for a series of spike-ins with two replicates. Wilcoxon between A-30, A-55, A-83, A-104, and A-149 replicates p is 0.3885 (z=0.8624), 0.4632 (z=-0.7335), 0.5437 (z=0.6073), 0.8521 (z=0.1865), and 0.8483 (z=0.1913). Kruskal-Wallis ANOVA p is 3.9×10^{-87} (df=9). Splint violin with KDE range is displayed (median, cyan line).

Mathematical consideration for CLT-seq

An only-thymine-containing short oligo(dT) annealing to its complementary poly(A) segment exhibits a very low binding affinity, particularly at 37 ℃ for reverse transcription (RT).(Gutiérrez Sánchez et al. 2015) The poly(A) tail consists of a fixed number of binding sites (segments) for oligo(dT) (Fig. 2, a). The forming weak oligo(dT)-poly(A) monovalent complex that stems from an individual binding event could unbind instantly if without

transcriptase-loading assistance or become progressively more stable as the annealed-oligo(dT) priming by the coincident-loading transcriptase. In other word, all binding sites on each poly(A) tail are subjected equiprobably to bind to and unbind from by the oligo(dT) before the reverse transcription succeeded. Each binding event assigns a potential start priming position along the poly(A) template to the transcription. We assume that the binding events are randomly packed to define the sample size underlying the CLT mechanism (Function in Methods), upon which the packed numbers denoting binding positions on the poly(A) tail are averaged to determine the coincidentally oligo(dT)-priming position (Fig. 2, a, bottom). The poly(A)-tail template directs the poly(T) stretch synthesizing from this starting position. In probability theory, the synthesized-poly(T) length-distribution approximates a normal distribution as the sample size becomes larger. Accordingly, the distribution of poly(T) length at T_{12} and A_{55} RT conditions was simulated as the sample size ranged from 1 to 30 (Fig. 2, b), on which an experimental conducting profile from the T_{12} primed-spike-in-A_{55} CLT-seq was also overlapped. The conducting experimental profile approximates the computational CLT-simulating profile of a sample size of 5. A two-sample Kolmogorov-Smirnov test with $p > 0.7522$ by cumulative distribution function-based probability-probability (PP) plot shows that these two profiles come from the same distribution (Fig. 2, c). We validated the CLT-arranging length-distribution assumption by high-fidelity mathematical simulation, which lays a solid theoretical foundation for CLT-seq.

The T_{12}-, T_{18}-, and T_{30}-primed poly(T) profiles corresponding to a series of the spike-ins are superimposed (Fig. 2, d) and fitted by the pseudoguassion function (Supplemental Fig. S2 and S3), which confirms a linear relationship between the input poly(A) length and the output mean poly(T) length or standard deviation (Fig. 2, e). By comparing standard least square-fitting results in T_{12}, T_{18}, and T_{30}-based datasets (Fig. 5S), we find that the T_{12}-based CLT-seq offers the best accuracy for modelling poly(A)-poly(T) conversion (Fig. 2, f and g; Supplemental Fig. S4). As the poly(A) tail lengthened, the observed frequency amplitude decreased hyperbolically (Fig. 2, h). According to the hyperbolical equation (Equation in

Methods) derived from the pseudogaussian function, a predicted frequency amplitude is deduced from the fitted standard deviation in theory. The measuring amplitude and the predicted value tally very well (Fig. 2, i), which attests to the fully qualified pseudogaussian function in the CLT-mechanism interpretation. Moreover, the T_{12}-based CLT-seq can mitigate the poly(A)-tail-length introduced bias (Fig. 2, j) due to the T_{12}-featured weak binding affinity. The poly(A)-tail length effect in the T_{12}-primed PCR amplification efficiency is likely to be overlooked only with around a 2-fold gain in yield even over a length of 129 nt; however, there is a significant increase in T_{18} and T_{30}-primed amplification efficiency with rising tail length. This bias should be widely existed in the anchor-free oligo(dT)-based sequencing that is currently one of the most common protocols for whole transcriptome profiling(Macosko et al. 2015), however, to our best knowledge, it has not been reported.

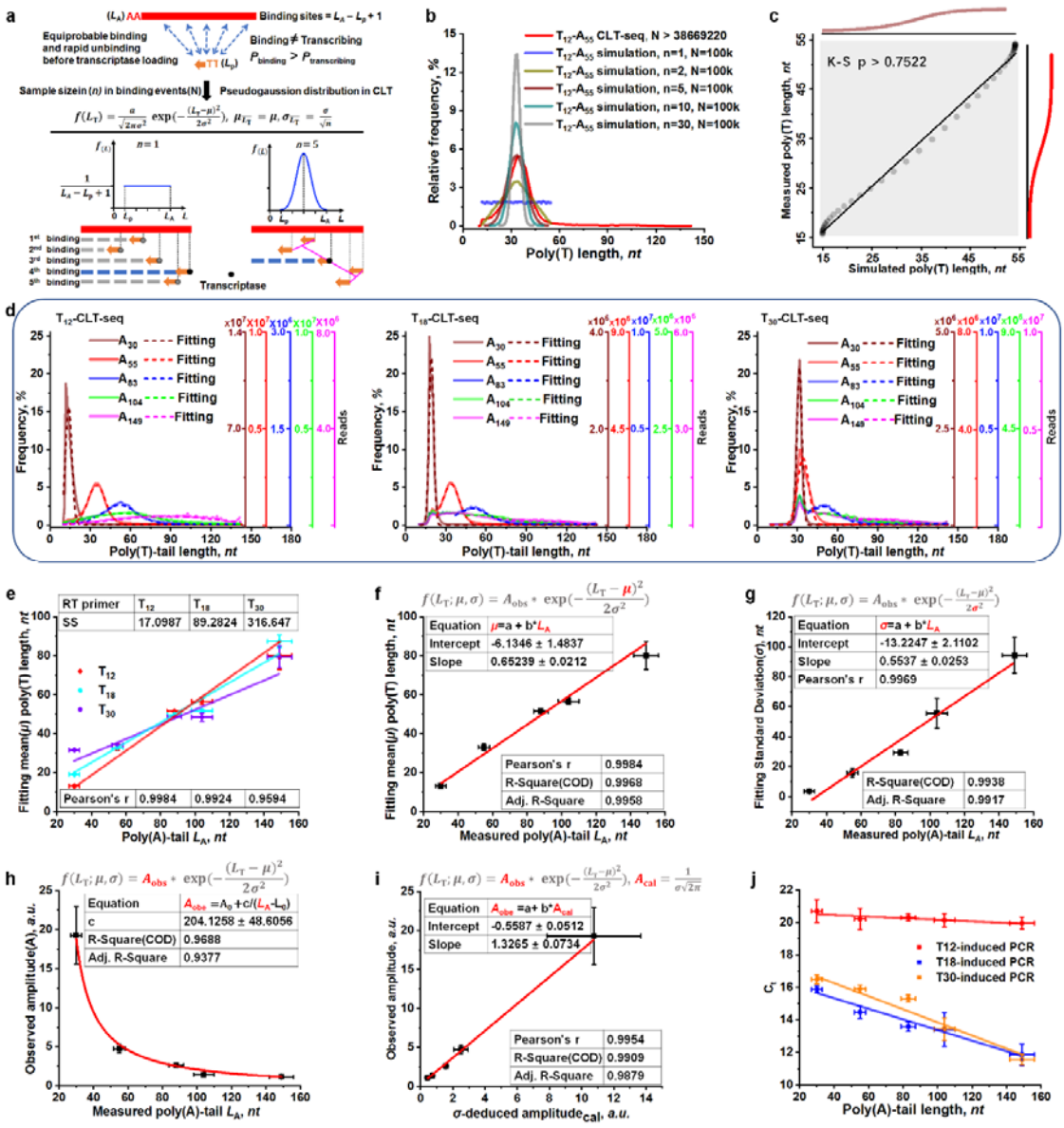

Fig. 2. Mathematical consideration of CLT-seq. **a**) Theoretical interpretation of CLT-arranged poly(T)-length distribution as sampling size increases. The poly(A)-oligo(dT) complexes are more than the poly(A)-oligo(dT)-transcriptase complex. The oligo(dT) primer equiprobably bound to and instantly unbound from the poly(A) segment till it was primed by the coincident-loading transcriptase. The sample size (n) determines the number of the poly(T)-length-assigned binding events (first to fifth, for instance) that are randomly packed. The mean length corresponding to the *n* independent binding events (*n*=5, for instance) is close to the deserved length of the middle oligo(dT) (3rd). **b**) CLT-seq simulation under T_{12} and A_{55}-fixed conditions and an experimental T_{12}-primed-A_{55}-template profile. **c**) The PP plot

of poly(T) length ranging from 12 to 55 nt and Kolmogorov-Smirnov test (K-S test) for simulation-measurement comparison. K-S p is 0.7523 (z=0.6794). **d**) The overlapped poly(T)-length profiles generated from T_{12}, T_{18}, and T_{30}-based CLT-seq. Counting frequency and reads are plotted versus the poly(T) length. **e**) T_{12}, T_{18}, and T_{30}-based regression lines of poly(A) tail length versus fitting poly(T) mean length and fitted standard deviation. **f-g**) The correlation analysis between the measured poly(A)-tail length and output mean poly(T) length or standard deviation. Data are represented as mean ± s.e.m. for two biological replicates. **h**) The regression curve of poly(A) tail length (L_A) versus frequency amplitude (A) from the T_{12}-primed dataset. **i**) The correlation analysis between the standard deviation (σ)-deduced amplitude versus the observed amplitude in experimental poly(T)-length profiles in the T_{12}-primed dataset. **j**) qPCR assessment of poly(A) tail-length preference corresponding to T_{12}, T_{18}, and T_{30} primers.

Practical utility of CLT-seq

Since the poly(A) tail is profiled by assigning the sequencing-determined relative abundance to the sorted lengths, the CLT-seq may suffer from the PCR bias. To investigate the PCR bias in the CLT-seq, we tracked the pattern of insert-length (Supplemental Fig. S5) and poly(T)-length distribution (Fig. 3, a) at the different phases of PCR amplification. A spike-in-A_{83} is subjected to CLT-seq with T_{12} primer, and the resulting cDNA is directly amplified with real-time ΔR_n output by qPCR. It suggests that the PCR amplification at different phases almost does not change the poly(T)/insert-length proportion within a wide GC-content range. This performance is due primarily to the KAPA HiFi polymerase's high processivity and exceptional fidelity.(Quail et al. 2012a) Library construction requires an appropriate PCR amplification, especially for rare cells or clinical samples, but must avoid an over-amplification. Since the CLT mechanism operates in multiple cellular contexts, we also investigated the oligo(dT)-primer/transcript-object ratio influence on the poly(T)-length distribution (Fig. 3, b). A spike-in-A_{55} is subjected to transcription at a T_{12} primer-transcript ratio from 0.2:1 to 200:1 with ten-fold serial dilutions, and then the resulting cDNA is

quantified by qPCR. Even over a 3-order-of-magnitude ratio range, two bases shift at most in the poly(T)-distribution spectrum. This tiny shift is entirely consistent with a previous study where increasing the primer/transcript ratio in the RT system has almost no effect on the poly(T) profiling.(Nam et al. 2002)

To evaluate the reliability and utility of the CLT-seq under a complex environment, we performed T_{12}, T_{18}, and T_{30}-based CLT-seq for the total RNA of HEK293T according to the gene-specific PCR-amplified procedure (Fig. 3, c and d). Here the CLT-generated 1st cDNA is directly amplified by the gene-specific primers of the human GAPDH and DNMT1. A desired GAPDH and DNMT1 libraries are identified by gel electrophoresis (Fig. 3, e). The minimum size of the PCR amplicon in the GAPDH library based on T_{12}, T_{18}, and T_{30} is 439, 445 (439+6), and 457 (439+18) nt, respectively. The minimum size of the amplicon in the DNMT1 library based on T_{12}, T_{18}, and T_{30} is 326, 332 (326+6), and 344 (326+18) nt, respectively. A more diffuse library generates from DNMT1 than GAPDH, implying that the mean poly(A)-tail length of DNMT1 is longer than that of GAPDH based on the diffusion ranging function (Function in Methods). Furthermore, the comparison of the CLT-seq sequencing poly(T)-length distribution shows a similar finding (Fig. 3, f). And a significant difference between the DNMT1 and GAPDH is confirmed further by the Wilcoxon-Mann-Whitney test $p = 3.2 \times 10^{-5}$ ($z=-4.16$) (Fig. 3, g), exhibiting good reliability and utility of the CLT-mechanism operation in multiple cellular contexts.

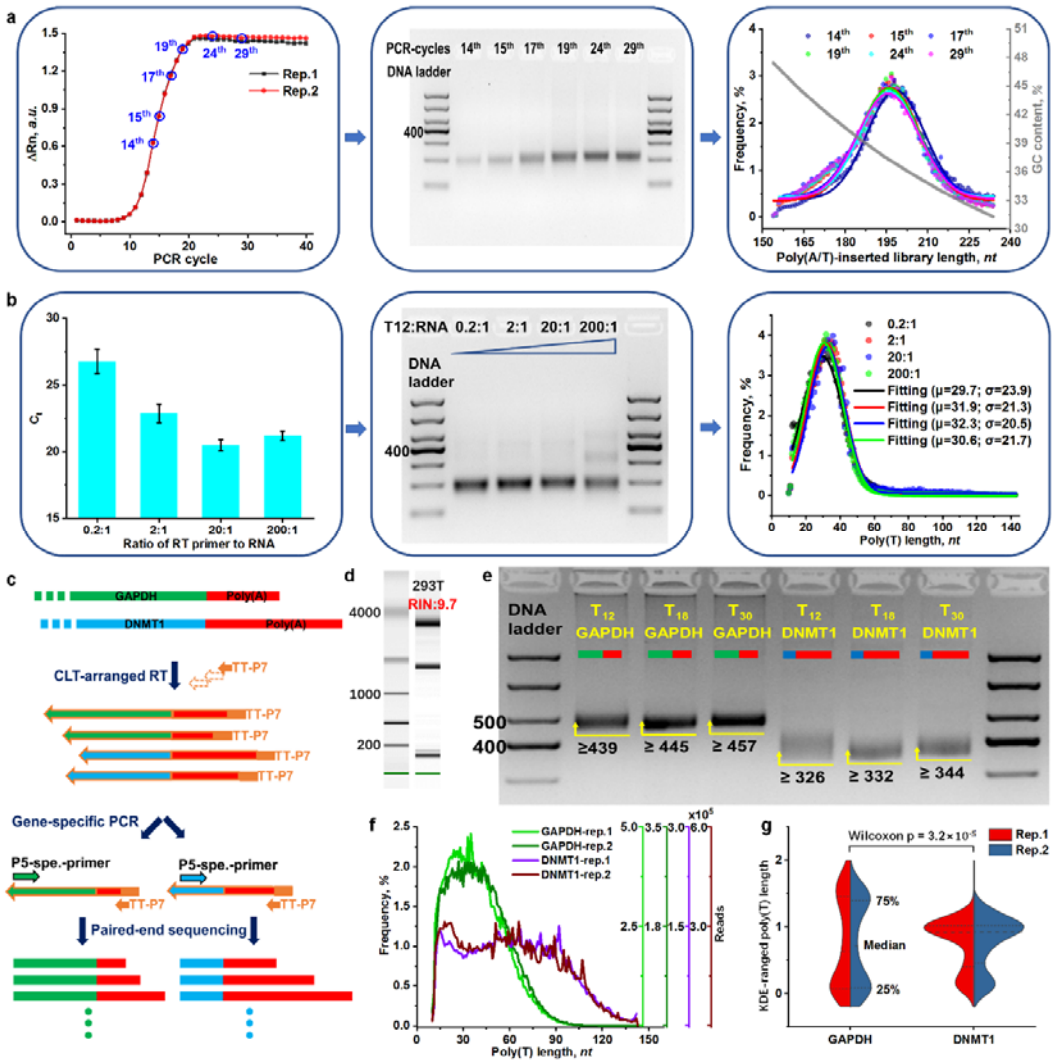

Fig. 3. Practical utility of CLT-seq. **a**) Poly(T) profiling at the different phases of PCR amplification. The PCR cycle number of 14th, 15th, 17th, 19th, 24th, and 29th in two replicates are marked by circles (left), and the cleaned-up libraries are identified by electrophoresis with 3% agarose gel (middle). The poly(T)-length profile of each library is overplotted versus counting frequency and GC content (right). Arrows represent the stepwise procedure. **b**) Primer-transcript ratio in poly(T) profiling. The different ratio-induced 1st cDNA is quantified by qPCR with relative C_t value (n=3, triplicates) (left), and the cleaned-up libraries are characterized by electrophoresis with 3% agarose gel (middle). The pseudogaussian fitted poly(T) profiles are overlapped with a mean (μ) and standard deviation (σ) (right). **c**) The stepwise procedure of gene-specific poly(T) length determination by CLT-seq. **d**) An integer

total RNA isolated from HEK293T. **e**) Characterizing the libraries of T_{12}, T_{18}, and T_{30}-based CLT-seq by electrophoresis. **f**) The poly(T) length of GAPDH and DNMT1 plotted versus the counting frequency and sequencing reads (n=2, two replicates). **g**) A split violin based on Kernel density estimation (n=2, two replicates). The dash and dot lines are median and quantiles, respectively.

Transcriptome-wide profiling poly(A) tail by CLT-seq

To determine directly the 3' end sequences of transcriptome, we performed CLT-seq to profile multiple-transcripts poly(A) tails at a transcriptome-wide scale according to a standard procedure (Fig. 4, a). Considering the reading-length limit and read-through overlap in the Illumina paired-end sequencing, the poly(T) length reading from the paired-end R_2 should not be shorter than the poly(A) length reading from the paired R_1 (Fig. 4, a, bottom library). As a result, the R_2-outputted poly(T)-length distribution exhibits a right-shift profile than the R_1-outputted poly(A)-length distribution in the GAPDH and ZNF544 (Supplemental Fig. S6). It indicates that the paired-end R_1 and R_2 ought to be in their good position to precisely map genes and output poly(A) sequence, respectively. We performed RNA-seq with two types of oligo(dT) primers to assess transcriptome-wide mapping performance. By comparison to the typical representatives of the anchored (Ramsköld et al. 2012; Jaitin et al. 2014) and non-anchored (Kolodziejczyk et al. 2015; Macosko et al. 2015) long-oligo(dT) primers, an anchor-free and short T_{12} offered a comparable pattern of read-mapping distribution (Fig. 4, b). This over-genome distribution feature highlights a pronounced oligo(dT)-specific poly(A)-dependent reverse transcription. The mapping coverage over each transcript is illustrated, where the R_1-based mapping result across six primer-based datasets confirms the 3'-end preference of transcripts.(Moll et al. 2014) Compared to the R_1, the paired R_2 carrying more poly(T) sequence and less mRNA information gives a poor gene-specific identification, thereby it is only applicable to outputting poly(T) sequence.

Based on the well-fitted conversion, the acquired poly(T) length was initially rescaled to the poly(A) length using the scalar-product formula (Fig. 4, c; Formula in Methods). Since the

pseudogaussian-shaped poly(T) profile flattened as rising poly(A)-tail length (Fig. 2, d), the reduced frequency (amplitude) should be offset by a weighted frequency where the length-overestimation effect has been taken into account (Function in Methods). Accordingly, the poly(A)-tail-length frequency is derived as the matrix product of the length-weighting factor and the poly(T)-length-sorted frequency (Fig. 4, c; Formula in Methods). Moreover, the mean poly(A)-tail length is also computed as the matrix product of the poly(A)-tail length and the weighted frequency (Formula in Methods).

The global poly(A)-tail profiles are restored from the T_{12}, T_{18}, and T_{30}-primed poly(T) profile (Fig. 4, d, up). A similar poly(A)-tail-length distribution exists, especially between T_{12} and T_{18} datasets, where two prominent frequency peaks emerge at ~25-nt and ~88-nt poly(A) lengths. This long highlighting length entirely agrees with the estimate of mean global poly(A) length (88.4 nt) in HEK293T by PAL-seq.(Subtelny et al. 2014) Since the CLT-seq followed an unbiased procedure without pre-treatment, enrichment, and selection of the total RNA, the naturally high abundance(Lima et al. 2017) of the short-tailed transcripts is likely to be profiled by the CLT-seq. We thought that the short-tail and/or low-abundance transcripts are easily filtered out in traditional library-preparing procedures, for example, the oligo(dT)-bead based capturing. To examine the reliability of CLT-seq in multi-transcript poly(A)-tail profiling, a good correlation is achieved between the biological replicates of GAPDH measurement (Fig. 4, d, down). Nevertheless, a relatively large fluctuation is observed in the frequency. We found that the degree of frequency fluctuation depends on the sequencing depth (the total number of gene-identified reads) (Supplemental Fig. S7). The number has grown to over the ~100 for obtaining a clear and smooth poly(A)-tail profile. Moreover, we conducted a ligation-mediated poly(A) tail assay (LM-PAT in Methods) with two replicates to profile the poly(A) tail of GAPDH. The LM-PAT(Sallés and Strickland 1999) followed an enrichment- and pre-treatment-free procedure as simple as we proposed in CLT-seq. There is substantial agreement between the CLT-seq profile and PCR-amplicon electrophoretic profile (Fig. 4, e).

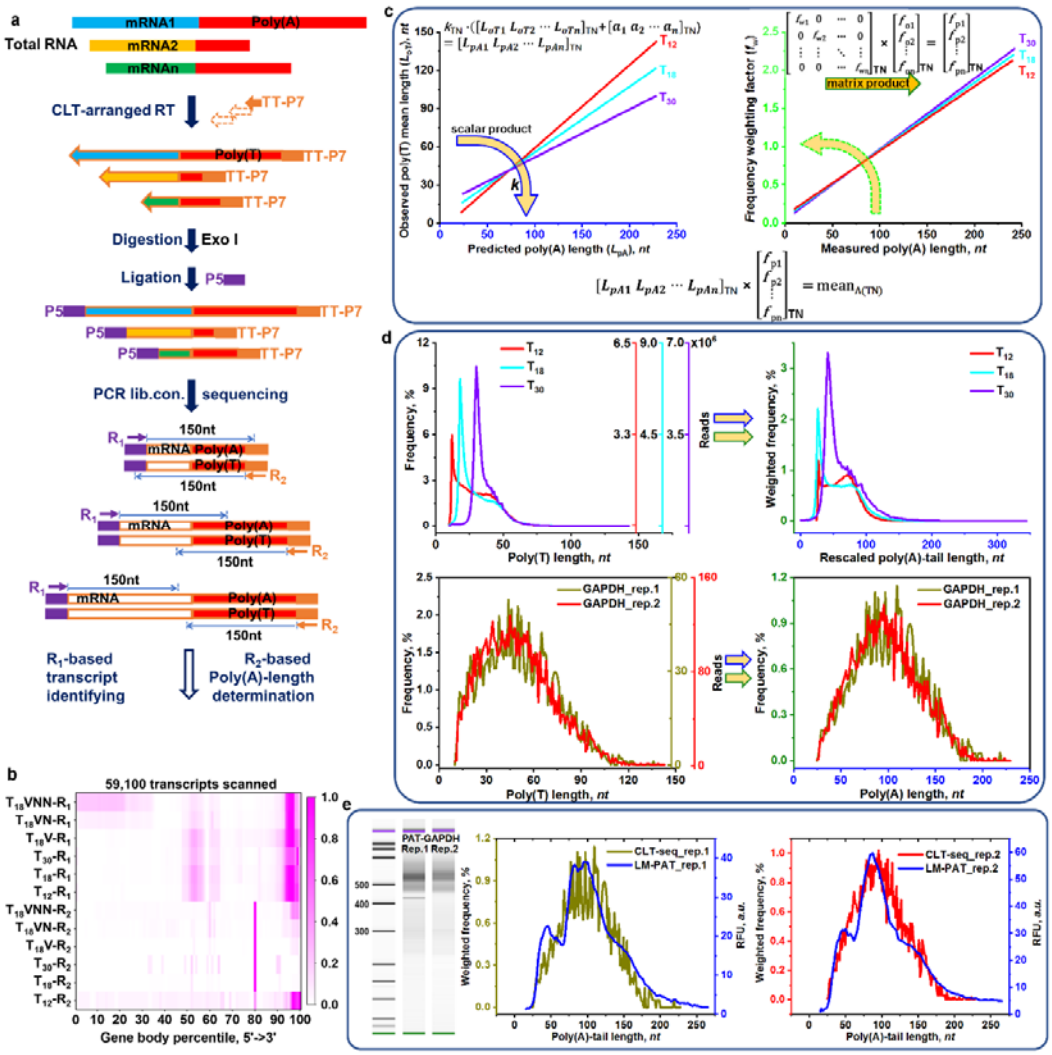

Fig. 4. Transcriptome-wide poly(A)-tail length determination and validation. **a**) The stepwise procedure of transcriptome-wide tail-length determination by CLT-seq. **b**) Mapping read coverage over 59 000 transcripts across six oligo(dT) primers based on paired-end R_1 and R_2 reads. **c**) Scalar and matrix operation for poly(A) rescaling, frequency weighting, and mean poly(A)-tail length determining. T_{12} (red), T_{18} (cyan), or T_{30} (violet) primer takes a different set of parameters. The arrow means the operating process. **d**) Global profile of poly(A) tail in HEK293T with T_{12} (red), T_{18} (cyan), or T_{30} (violet)-based CLT-seq. GAPDH-specific poly(A) tail profile from T_{12}-based CLT-seq, n=2 (two replicates). **e**) Poly(A) length validation by LM-PAT assay for GAPDH with two replicates. PAT assay product identified by Agilent 2100 bioanalyzer with DNA chip in left. PAT assay profile (blue) with RFU (*a.u.*) overlapped the

poly(A) length distribution (red) with weighted frequency (%) restored from CLT-seq data in the right panel.

Poly(A)-tail length control in ncRNA regulation

The mitochondrial (MT) transcriptome affords a great diversity of the poly(A)-tailed mRNAs and ncRNAs to sample the full sets of cellular transcriptomes.(Mercer et al. 2011) Since the poly(A)-tail length is associated explicitly with translation efficiency in early embryonic development or during the cell cycle,(Subtelny et al. 2014; Park et al. 2016) the embryonic HEK293T cell and lung cancer A549 cell are adopted in the comprehensive poly(A)-tail study by using CLT-seq (Supplemental Fig. S1, c). We identified all mitochondrial transcripts (37 RNA symbols and D-Loop) by aligning the R_1-decoded sequence to human sapiens mitochondrion (NC_012920.1) and restored the poly(A)-tail profile based on the paired R_2 reads (Fig. 5, a). A different pattern of mitochondrial-transcript abundance is observed between HEK293T and A549 (Fig. 5, b). And the mean tail length of literature-reported mitochondrial-mRNAs(Legnini et al. 2019) agrees with the results measured by CLT-seq (data not shown).

Because the mean or median length in terms of ‘short’ or ‘long’ cannot thoroughly reflect the total length spectrum of poly(A) tail, we profiled poly(A) tail by assigning the sequencing-determined relative abundance to the sorted lengths (Fig. 5, c). Every mitochondrial poly(A)-tail profile in HEK293T or A549 is unique, even though the mean lengths are very close (ND3 versus CYTB and ND6 versus RNR2). We also find that many profiles exhibit a double-humped tail-length distribution, which usually arises from poly(A)-tail pruning in the different phases of mRNA in gene expression regulation.(Castellano and Bazzini 2017) A similar pruning mechanism seemingly also occurs in ncRNAs. Interestingly, a positively skewed distribution is first observed in the D-Loop transcript, demonstrating the poly(A)-tail-promoted degradation of the nonfunctional ncRNAs.(Houseley and Tollervey 2009) To confirm these new findings, T_{12}, T_{18}, and T_{30}-primed profiles overlapped for each mitochondrial transcript, exhibiting a similar tail-length pattern (Supplemental Fig. S8). We

used the KDE method to normalize the poly(A)-tail profile with a Gaussian-shaped *Kernel* (Fig. 5, d). There is a significant difference between all profiles with Kruskal-Wallis ANOVA $p < 3.6 \times 10^{-105}$, and there also exists a pairwise comparison across all transcripts with Dunn's test. By comparing HEK293T and A549, an apparent discrepancy presents in many ncRNA profiles. We find an overall increase of poly(A)-tail length in A549 tRNAs that may couple with the specific upregulation of tRNA-derived small RNA (tsRNA) in lung cancer.(Pekarsky et al. 2016; Slack and Chinnaiyan 2019) CLT-seq here helps us unravel the biogenesis and regulation mechanism of tsRNAs.

To assess the reliability in discriminating subtle differences between profiles, we performed CLT-seq with two biological replicates for HEK293T (Supplemental Fig. S9) and A549 (Supplemental Fig. S10). There exists a good correlation between the two replicates across all mitochondrial transcripts. By comparison, the correlation coefficient in ncRNAs with a short tail was higher than that of mRNAs with a relatively long poly(A) tail, most likely due to a larger variation in the long tail measurement. Nevertheless, the long-length variation effect in CLT-seq only shifted the profile to some extent but did not change the fundamental pattern underlying the poly(A)-tail profile.

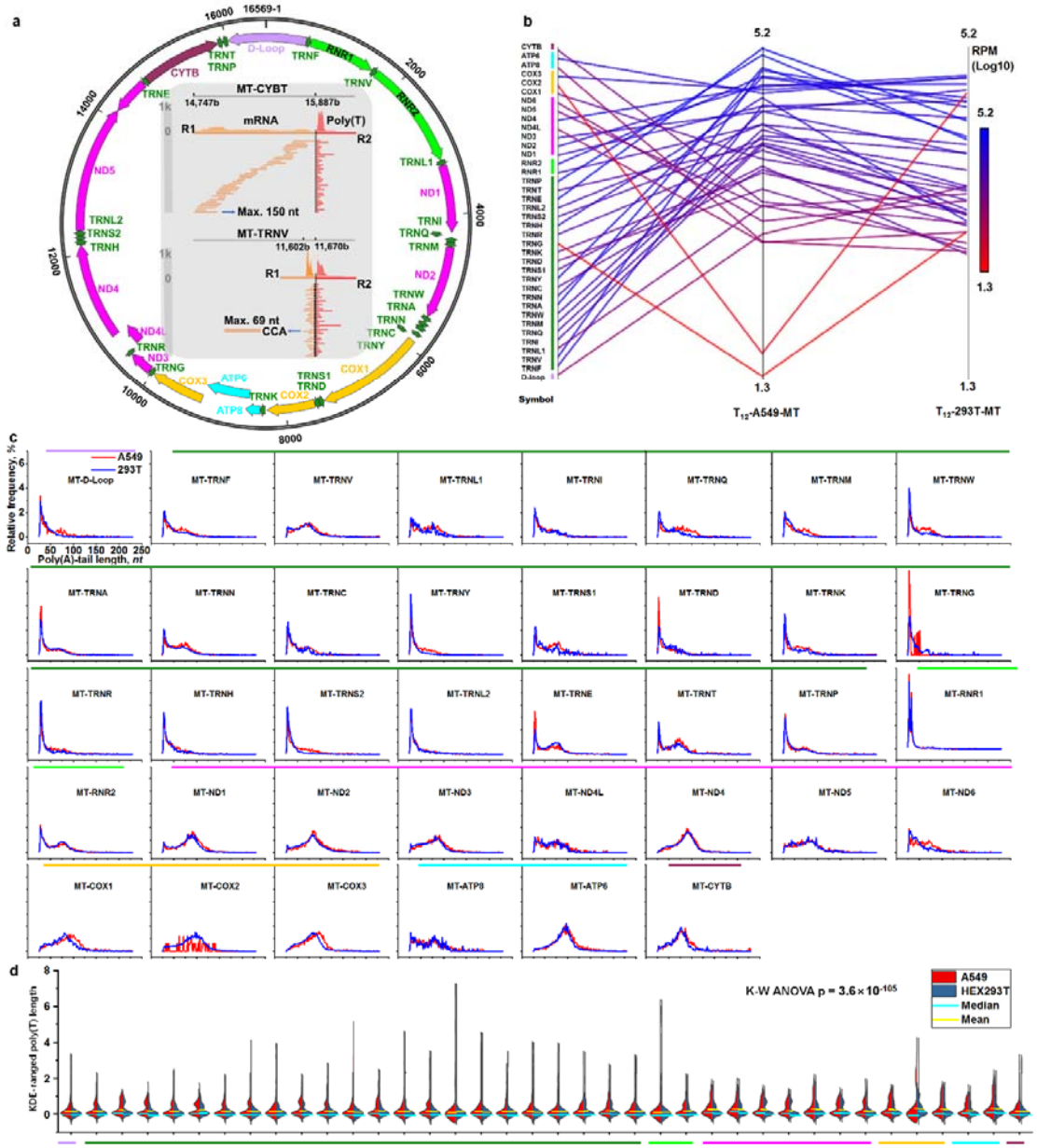

Fig. 5. Poly(A)-tail length analysis of whole mitochondrial transcripts. **a**) Map of the human mitochondrial transcriptome with annotated gene features regarding NC_012920.1 (outer tracker) and two poly(A)-tail profiling strategies for individual transcript (inner). In an example of a lengthy transcript (mitochondrial-CYTB), the sequence with a maximum length of 150 nt from paired-end R_1 is mapping onto 1141-nt CYTB mRNA reference, and the corresponding poly(T) length is outputting. For short transcript (mitochondrial-TRNV), the sequence from R_1 directly aligns to the ncRNA-poly(A) tail junction for the poly(T) tail profiling. **b**) Transcript expression abundance. All transcripts are in type and subclass order with different colors.

The abundance of RPM is plotted parallelly with the transcript symbol in the HEK293T and A549 dataset. **c**) The restored Poly(A)-tail profile by CLT-seq. The overlapped profiles of HEK293T and A549 tally counting frequency for each position with single-based resolution. **d**) The split violin plot of all mitochondrial transcripts with Kruskal-Wallis ANOVA analysis.

Characterizing poly(A)-tail profile at transcriptome-wide scale

Although the rank-based estimation in Kruskal-Wallis ANOVA analysis confirmed the difference of mitochondrial poly(A)-tail profile in HEK293T or A549, it is vague and imprecise in many pairwise comparisons. Therefore, we performed a statistical analysis using the Pearson correlation method for HEK293T (Fig. 6, a) and A549 (Fig. 6, b) datasets. A more exact Pearson's correlation-coefficient value is assigned to each pairwise comparison. This two-dimensional correlation pattern offers substantially quantitative interpretation between multiple endogenous poly(A) tails, even across two or more heterogeneous cell lines.

It is worthwhile noting that the poly(A) tail mostly added to the 3'CCA nucleotides following the 3' end of the acceptor's arm across all mitochondrial-tRNAs in the T_{12}-based (Fig. 6, c), T_{18}-base (Supplemental Fig. S11, a), and T_{30}-based dataset (Supplemental Fig. S11, b). The proportion of CCA-matured mitochondrial-tRNA concerning total mitochondrial-tRNA reached around 80% on average (Supplemental Fig. S11, c). The T_{12}, T_{18}, and T_{30}-based CLT-seq were conducted parallelly to confirm this high percentage, achieving consistent accounting-for results in all mitochondrial-tRNAs. Strikingly, about 0.3% mitochondrial-tRNAs tagged with two CCA repeats (5'-NCCACCA-3') at the 3' end of the acceptor's arm were identified (data not shown), which yielded strong evidence and stringent validation for the cytosolic tRNA nucleotidyltransferase 1 (TRNT1)-mediated tRNA degradation.(Wilusz et al. 2011) It was the first time directly observing 3'-end CCACCA (two repeats) sequences at mitochondrial tRNA, showing that the cytosolic TRNT1-based tRNA clearing system was not only applicable to the nuclear-coded tRNAs but also to the mitochondrial-coded tRNAs.

We computed the mean tail length of transcripts with a relatively low abundance in T_{12}-based data from HEK293T and A549 (Fig. 6, d). Mean tail length does not correlate with

RNA abundance. In contrast, it positively correlates with transcript size (Fig. 6, e), which is in accord with the relationship between the median tail length and 3'-UTR length.(Legnini et al. 2019) In addition, CLT-seq also enables to characterize the poly(A) tail of all kinds of APA (alternative polyadenylation)-spliced isoforms. An isoform-specific poly(A)-tail profile was plotted at each APA site on the same 3'-UTR across different genes (Supplemental Fig. S12), where a unique isoform was identified by the assembly of specific 3'-end of UTR and poly(A) tail. Moreover, the cross-intron spliced isoform's poly(A) tail in an individual gene can also be profiled (Supplemental Fig. S13). The skipped region distinguishes these isoforms with the same APA site on the 3'-UTR.

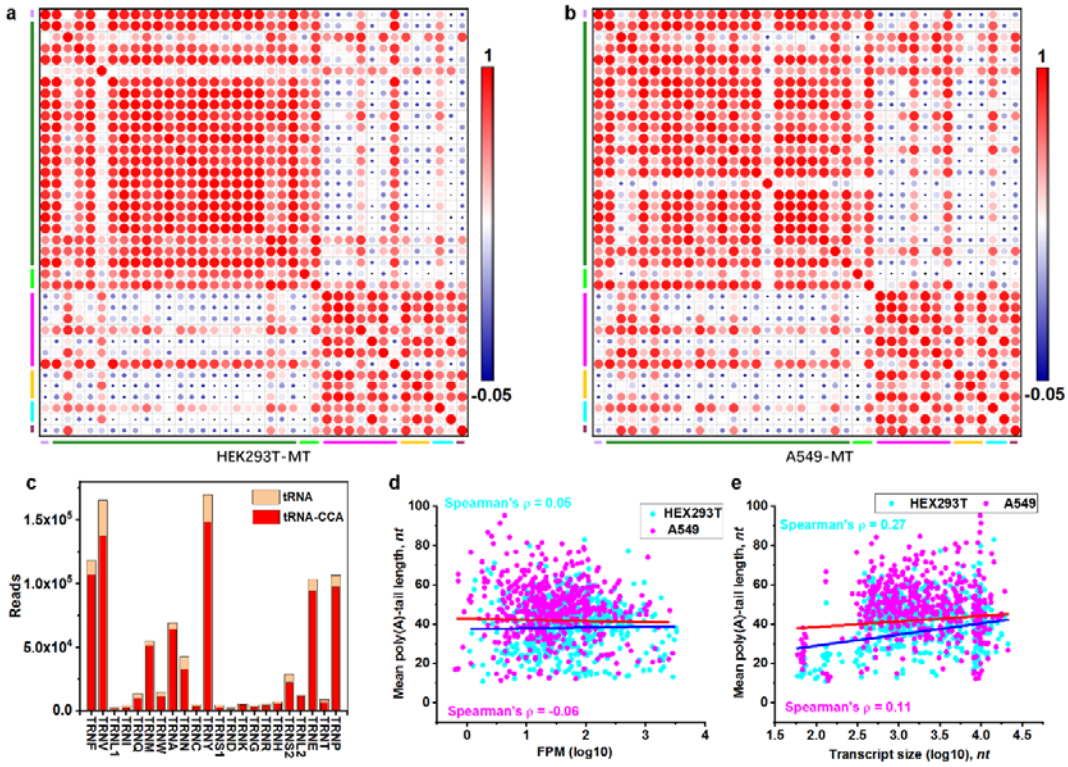

Fig. 6. The featured poly(A)-tail. **a)** and **b**) Correlation scatter plot of pairwise comparison across all mitochondrial transcripts. Pearson correlation ecoefficiency (r) is related to the color and diameter of each spot in HEK293T and A549 datasets. **c**) The read count of CCA marked tRNA and all identified tRNA across all mitochondrial tRNAs in HEK293T. **d)** and **e**) Correlation of mean tail length with transcript abundance and size in HEK293T (cyan) or

A549 (magenta) cells (n=1, Spearman correlation). Comparison of poly(A)-tailed ncRNA transcript abundance determined by T_{12}-based CLT-seq between HEK293T and A549.

DISCUSSION

The poly(A) tail is the main participant in eukaryotic RNA quality control.(Houseley et al. 2006) Even so, the transcriptome-wide analyses of poly(A)-tail length started only when the TAIL-seq(Chang et al. 2014) and PAL-seq(Subtelny et al. 2014) methods broke through the homopolymer base-calling technical limits. In recent years the Nanopore technology(Garalde et al. 2018) and PacBio system(Legnini et al. 2019; Liu et al. 2019; Mattijssen et al. 2020) also proved to overcome homopolymer-reading barriers. However, all such methods for the high-quality homopolymer reading must rely on an elaborate sequencer-dependent library-preparing plan that inevitably introduces biases in sequencing. As a result, the current methods make it hard to print the actual poly(A)-tail profile, especially for the rare transcripts, at a transcriptome-wide scale. Moreover, there still exist applicability issues in such homopolymer-reading solutions. Therefore, we developed CLT-seq based on a simple workflow and a universal homopolymer-sequencing concept. It enables to bypass the enriching and/or selecting steps to avoid common artificial biases. In addition, the non-anchored oligo(dT) priming mechanism circumvents the GC bias, and a short oligo(dT) mitigates tail-length bias. Moreover, the randomized splint format adopted in CLT-seq could significantly reduce bias during the introduction of the second adaptor.(Maguire et al. 2020) Since CLT-arranged transcription only requires low-input total RNA (≤ 500 ng), the buffer contains divalent cations strong enough to shear the long transcripts into proper fragments for short-read Illumina sequencing.(Wery et al. 2013) Finally, an unbiased RNA-seq at transcriptome-wide scale is accomplished through such concerted efforts.

The short oligo(dT) binds equiprobably to and unbinds instantly from its target segment of the poly(A) tail till it is primed by the coincidently-loading transcriptase, upon which the CLT mechanism arranges the poly(A)-template-directed poly(T)-synthesis length to distribute normally. Theoretically, the size of the poly(A) tail shrinks by around half, which facilitates the

high-quality homopolymer sequencing. Furthermore, the well-fitted pseudoguassion-derived model underlying the weighted KDE method interconverts between the length of the poly(A) tail and poly(T) stretch. Since the overall median poly(A)-tail length of the most human mRNAs concentrates in a range of 50 and 100 nt(Subtelny et al. 2014), the CLT-shortened mean poly(T) length should range from 25 to 50 nt that can be measured by Illumina sequencing. Moreover, this shortened homopolymer can be read out easily by the advanced Tailseeker, PacBio, and Nanopore. Therefore, the CLT model could serve as a universal concept for poly(A) tail profiling, especially on the Illumina platform with the widespread availability and well-tested bioinformatic pipeline.

By applying CLT-seq on the Illumina platform, we found two peaks at ~25 nt and ~88 nt in the global poly(A)-tail profile in HEK293T. The ~88 nt poly(A) length is basically consistent with the global tail length measured by the PAL-seq.(Subtelny et al. 2014) However, the ~25 nt poly(A) length is scarcely reported because the short-tail transcripts are easily filtered out in traditional library-preparing procedures and principally unfavoured in ordinary oligo(dT) primer-based RT. There likely exists a relatively high abundance of the short-tailed transcripts coming from the normal metabolism of long poly(A) tail, conserved ~25-30 nt footprint tail of highly expressed genes, and random RNA degradation.(Lima et al. 2017) Still, it probably cannot be captured and quantified by the current sequencing methods. The CLT-seq enables the transcripts with short poly(A) tail to be sequenced unbiasedly. We provided the first comprehensive analysis of poly(A)-tail length in the human mitochondrial transcriptome. The mean tail length of the most mRNAs is longer than that of ncRNAs, supporting the notion that a relatively long tail is responsible for mRNA dynamic regulation and a short poly(A) tail extension is required only for nuclease loading.(Doma and Parker 2007) Even though the mitochondrial-ncRNA is tagged with a short tail, the poly(A)-tail profile does not follow a consistent length-distribution pattern. It indicates that poly(A)-tail-stimulated ncRNA degradation is a complex process under highly coordinated control. Moreover, compared to the poly(A)-tail profile of mitochondrial ncRNAs in a normal cell, a double-bumped profile

emerged in many mitochondrial tRNAs in a lung cancer cell. The altered pattern based on the poly(A)-tail-profile correlation enables the reliable tracking of gene-expression dynamics. It helps us sample the full sets of total RNA and characterize some rare events.

METHODS

Data analysis:

1. *Data pre-processing*. The published-software pipeline processes the pair-end raw data: FastQC (v.0.11.8) (Babraham Bioinformatics), cutadapt (v.2.10)(Kechin et al. 2017), trimmomatic (v.0.39)(Bolger et al. 2014), hisat2 (v.2.2.1)(Kim et al. 2019), samtools (v.1.7)(Li et al. 2009), and RSeQC package(Wang et al. 2012; Wang et al. 2016).

2. *Poly(T) profiling*. For paired-end R_1, poly(A) calling is preceded by RNA body, while for R_2, poly(T) calling precedes RNA-body. R_1 as forward read primarily calling RNA body and R_2 as reverse read primarily calling poly(T) stretch are deployed to identify gene-specific transcript and output its corresponding poly(T) length, respectively. We firstly identified genes from the trimmed R1 reads aligned to the reference hg19_RefSeq.bed using the FPKM_count.py script. The sequence at the end of 3'-UTR from the identified gene packs the gene-specific R1-read ID from massive pair-end reads, and then the R2-read with the same index-ID was pooled together for poly(T)-length outputting.

3. *The code for poly(T) profiling is as following*:

The **code** for gene-specific grouping from R1:

```
grep -B1 'UTR_sequence' trimmed_input_R1.fastq | sed
's/1:N:0:ATCTCGTA+ATAGAGAG(barcode)//g' | awk NR%3==1 > R1_ID.txt
```

The **code** for R2-read pooling with the outputted R1_ID:

```
for i in $(cat R1_ID.txt)
do
        grep -A1 $i trimmed_input_R2.fastq >> R2_reads_indexed_ID.txt
done
```

The **code** for gene-specific grouping:

```
for ((i=10; i<144; i++))
do
egrep -c T{$i}> R2_reads_indexed_ID.txt >> abundance_assigned_sorted_tail-length.txt
done
```

4. *Read-mapping distribution*. The raw data is treated following pipeline to generate .bam file that is processed by read_distribution.py script with reference hg19_RefSeq.bed. The acquired tags per Kb are normalized by sequencing depth.

5. *Mapping coverage*. The .bam file is treated by geneBody_coverage.py script with reference hg19_RefSeq.bed.

6. *Nucleotide frequency*. The .bam file is processed by read_NVC.py script with reference hg19_RefSeq.bed.

7. *Expression abundance quantification*. The .bam file is processed by FPKM_count.py script with reference hg19_RefSeq.bed.

Mathematical operation:

1. *Weighted KDE method*. To transcriptome-widely obtain the poly(A)-tail profile, the acquired poly(T)-length distribution is processed *via* matrix operation based on this verified model using the kernel density estimation (KDE) method. In statistics, KDE with a gaussian kernel is often used to estimate the probability density function. Given CLT-based pseudogaussian kernel, a weighted KDE with function:

$$\hat{f}(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} f_{wi} \cdot \emptyset(L - L_{Ai})$$

is applied to restore poly(A)-length profile (probability density function) where $\hat{f}(x)$ is the estimating probability density function, f_{wi} indicates the weighting factor for poly(A) length of i nt, and $\emptyset(L - L_{Ai})$ is pseudogaussian function for i nt poly(A) length.

2. *Scalar product operation*. The sorted poly(T) length is rescaled to the poly(A) tail length by a scalar product with function (Fig. 4, c, left):

$$[L_{pA1}\ L_{pA2} \cdots L_{pAn}]_{TN} = k_{TN} \cdot ([L_{oT1}\ L_{oT2} \cdots L_{oTn}]_{TN} + [a_1\ a_2 \cdots a_n]_{TN})$$

where L_{oTn} and L_{pAn} are the observed poly(T) length and the resulting poly(A) length in length order, respectively, $_{TN}$ denotes the oligo(dT) primer (T_{12}, T_{18} and T_{30}), k is a scale factor, and a_n is a constant for all length n ($a_1=a_2=\cdots=a_n$).

3. *Matrix product operation*. The counting frequency (density) of poly(A)-tail length (variable) is the matrix product of the indexed weighting factor and the observed frequency of the poly(T) length with the following function (Fig. 4, c, right):

$$\begin{bmatrix} f_{w1} & 0 & \cdots & 0 \\ 0 & f_{w2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f_{wn} \end{bmatrix}_{TN} \times \begin{bmatrix} f_{o1} \\ f_{o2} \\ \vdots \\ f_{on} \end{bmatrix}_{TN} = \begin{bmatrix} f_{p1} \\ f_{p2} \\ \vdots \\ f_{pn} \end{bmatrix}_{TN}$$

where f_{wn} in n×n matrix and f_{on} in n×1 matrix is the weighting factor and the observed frequency of poly(T) length of n nt, respectively. The weighted factors are listed orderly in a square diagonal matrix with poly(A)-length-dependent diagonal entries. This a weighted frequency should offset length-reduced frequency to reflect the transcript abundance-determined poly(A)-tail profile. Therefore, the weighted KDE method enables the CLT-based poly(A)-tail profile to approximate the actual tail-length distribution.

4. *Mean length (expected value).* The mean poly(A)-tail length is the matrix product of the poly(A) length and the weighted frequency. It follows that

$$\left[L_{pA1} L_{pA2} \cdots L_{pAn}\right]_{TN} \times \begin{bmatrix} f_{p1} \\ f_{p2} \\ \vdots \\ f_{pn} \end{bmatrix}_{TN} = \sum_{i=1}^{n} \left(L_{pAn} \cdot f_{pn}\right)_{TN} = \mathrm{mean}_{\mathrm{A(TN)}},$$

where L_{pAn} and f_{pn} represents the rescaled poly(A) length and the corresponding weighted frequency, respectively; $\mathrm{mean}_{\mathrm{A(TN)}}$ denotes the mean poly(A) tail length (expected value) (Fig. 4, c, bottom).

Pseudoguassion function:

A pseudoguassion distribution function describes bell-like curve with

$f(L_T) = \frac{a}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(L_T-\mu)^2}{2\sigma^2}), \mu_{\overline{L_T}} = \mu,\ \sigma_{\overline{L_T}} = \frac{\sigma}{\sqrt{n}}$, where L_T is the length of poly(T) in 1st cDNA,

$f(L_T)$ denotes the counting frequency (density), μ and σ is the population mean and standard deviation, respectively, a is constant, $\mu_{\overline{L_T}}$ and $\sigma_{\overline{L_T}}$ indicates the sampling mean and sampling standard deviation, respectively, and n is the sample size.

Frequency amplitude-standard deviation Equation:

According to the pseudogaussian function, a bell-like curve of CLT-based poly(T) length distribution flattened as rising poly(A)-tail length (Fig. 2d), leading to a reduced amplitude (counting frequency). The frequency amplitude is inversely proportional to the standard deviation (σ) as expressed by equation: $A_{obs} = \frac{a}{\sqrt{2\pi\sigma_{LA}{}^2}}$.

Sequencing library diffusion range function:

The library diffusion range (length-distribution range, R) is well-defined by an equation: R = L_A-L_p+1. The mean poly(A)-tail length of DNMT1 with strong diffusion is accordingly longer than that of GAPDH ($R_{DNMT1} > R_{GAPDH}$ leading to $L_{A\text{-}DNMT1} > L_{A\text{-}GAPDH}$ with fix-length primer (L_p)).

Cell culture:

HEK293T cells were cultured in DMEM high glucose (catalog no. SH30022.01, HyClone) supplemented with 10% FBS (catalog no. 10100147, Thermo Fisher). In addition, A549 cells are cultured in F-12K (Gibco/Invitrogen) supplemented with 10% FBS. The cell lines grow at 37 ℃ and 5% CO_2 with saturating humidity.

RNA isolation:

All total RNAs were extracted from cells following the worksheet in RNAiso Plus (catalog no. 9109, TaKaRa Biotechnology) and resuspended at the appropriate concentration(s) in RNase-free water. The RNase-free water was used throughout the RNA-related assays. Total RNA was checked on 2100 Bioanalyzer RNA picochip and then aliquoted into 5 μg for long-term storing at -80 ℃.

Spike-in RNA and polyadenylation:

0.5 μL of 20 μM spike-in RNA was added in 20 μL polyadenylation buffer with 0.4 μL of 5 U/μL Poly(A) Polymerase (part no. P7460L, QIAGEN) and 2 μL of 10 mM ATP for 0, 1-, 3-, 5-, and 10-min incubation at 37℃:

Spike-in:

5'-CACCUCUCUAUACACUCUUUCCCUACACGACGCUCUUCCGAUCUGAUCGTAAUGGCUUAG-3'

We noted the polyadenylated reaction started instantly (0 min incubating) after blending all reagents together.

Library preparation:

500 ng of total RNA was subjected to 5 μl RT reaction with 0.5 μL of 1 μM primer T_{12}, T_{18}, T_{30}, $T_{18}V$, $T_{12}VN$, and $T_{18}VNN$:

T_{12}:

5'-TACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGCGTAGCTTTTTTTTTTTT-3'

T_{18}:

5'-TACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGCGTAGCTTTTTTTTTTTTTTTTTT-3'

T_{30}:

5'-TACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGCGTAGCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT-3'

$T_{18}V$:

5'-TACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGCGTAGCTTTTTTTTTTTTTTTTTTV-3'

$T_{12}VN$:

5'-TACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGCGTAGCTTTTTTTTTTTTTTTTTTVN-3'

T_{18}VNN:

5'-TACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGCGTAGCTTTTTTTT TTTTTTTTTTVNN-3'

1 μl RT buffer mix (5×), 1 μl 200 U/μl of Super MMLV Reverse Transcriptase (catalog no. MD028, FAPON), 0.5 μl of 10 mM dNTP (each), and 3.3 mM Mg^{2+} at 37℃ for 30 min. And then, 1 μL of 20 U/μl of Exonuclease I (catalog no. M0293S, New England Biolabs) was added directly to the RT mix for 30 min at 37°C and 20 min at 80°C. The resulting 1st cDNA was mixed with 1 μL of 10 μM splint-second adaptor that was incubated at 95 °C for 5 min, slowly cooled to 25 °C:

5'-CACCTCTCTATACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNN-invT-3'

5'-p-AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTATAGAGAGGTG-invT-3'

in 20 μl ligation with 1 μl of 400 U/μl T4 DNA ligase (catalog no. M0202L, New England Biolabs) and 2 μl DNA ligase Buffer (10×). After 2 hours at 16°C, the final mix was then amplified by PCR with HIFI KAPA master mix (2×) according to the following mixture: 25 μl 2× HIFI KAPA master mix, 10 μl cDNA, 13 μl H_2O and 1 μl one of the two universal forward and reverse primers (10 μM):

universal-forward-primer:

5'-AATGATACGGCGACCACCGAGATCTACACCTCTCTATACACTCT*T-3'

(*:phosphorothioate)

universal-reverse-primer:

5'-CAAGCAGAAGACGGCATACGAGATGTGACTGGAGT*T-3' (*:phosphorothioate)

The PCR reaction was then performed in a thermocycler as follows: 95 °C for 5 min, 23× (95 °C for 15 s and 60 °C for 30 s). The amplified library was then purified with 1.8× Ampure XP DNA Beads and resuspended in 20 μl H_2O. Finally, one microliter of the library was checked by Qubit 4 Fluorometer with dsDNA HS (high sensitivity) Assay Kit (Invitrogen).

Since there was not any clean-up step in the workflow, the library preparation in negative control yields a variable number of 4-types primers-introduced dimers or concatemers, which

is minimized by the deserved target cDNA competition in the mRNA sample. Moreover, the non-specific amplicon without flow cell binding sites and sequencing primer binding sites are not likely to be sequenced.

Illumina sequencing:

The purified PCR libraries were submitted to the Hoplox for Illumina sequencing (Novaseq 6000, two channel). The data output was according to a yield set.

RNA standards and Spike-ins:

The poly(A) lengths were precisely measured by Agilent 2100 bioanalyzer as standard spike-ins carrying poly(A) tail of 30±2, 55±3, 83±4, 104±4, and 149±6 nt, respectively.

GAPDH and DNMT1 electrophoresis assay:

The 500 ng of total RNA was subjected to a 5 μl RT reaction with 0.5 μL of 1 μM T_{12}, T_{18}, or T_{30}, 1 μl RT buffer mix (5×), 0.5 μl of 10 mM dNTP (each), 1 μl of 200 U/μl Super MMLV Reverse Transcriptase (catalog no. MD028, FAPON), 3.3 mM Mg^{2+} at 37℃ for 30 min. And then, 0.5 μL of 20 U/μl Exonuclease I (catalog no. M0293S, NEB) was added directly into the RT mix for 30 min at 37℃ and 20 min at 80℃. Subsequently, the mix was then amplified by PCR with 25 μl HIFI KAPA master mix (2×), 1 μl universal reverse primer (10 μM) and 1 μl forward primer GAPDH-tag or DNMT1-tag (10 μM),

GAPDH-tag:

5'-CACCTCTCTATACACTCTTTCCCTACACGACGCTCTTCCGATCTCCTCAACGACCACTTTGTCAAG-3'

DNMT1-tag:

5'-CACCTCTCTATACACTCTTTCCCTACACGACGCTCTTCCGATCTCTGTTGTGTGAGGTTCGCTTATCAAC-3'

The PCR reaction was then performed in a thermocycler as follows: 95 °C for 5 min, 20× (95 °C for 15 s and 65 °C for 30 s).

LM-PAT assay:

The 500 ng of total RNA was subjected to a 5 μl RT reaction with 0.5 μL of 100 nM the phosphorylated T_{12-18} and 1 μl RT buffer mix (5×) at 65 °C for 5 min, quickly cooled to 37 °C. 15 μl of the following prewarmed mixture is added and incubated at 37 °C for 30 min. For each reaction add: 4 μl RT buffer mix (5×), 2 μl DNA ligase Buffer (10×), 0.5 μl of 10 mM dNTP (each), 0.5 μl of 10 mM ATP, 1 μl of 200 U/μl Super MMLV Reverse Transcriptase (catalog no. MD028, FAPON), 1 μl of 400 U/μl T4 DNA ligase (catalog no. M0202L, New England Biolabs). At the end of the incubation, while still at 37 °C, 0.5 μL of 1 μM T_{12} was added and incubated at 16 °C for 2 hours. Subsequently, 1 μl of 200 U/μl Super MMLV Reverse Transcriptase (catalog no. MD028, FAPON) was added and incubated at 37 °C for 30 min and 65 °C for 30 min.

Finally, the cDNA was pre-amplified by PCR with 2 μl universal reverse primer (10 μM) and 2 μl forward primer GAPDH-1-tag (10 μM). The PCR reaction was performed in a thermocycler as follows: 95 °C for 5 min, 15× (95 °C for 15 s and 60 °C for 30 s). The pre-amplified product was further amplified with two universal primers for the 2100 Bioanalyzer and Illumina sequencer. The PCR reaction was performed in a thermocycler as follows: 95 °C for 5 min, 20× (95 °C for 15 s and 60 °C for 30 s).

AVAILABILITY

The code used for data analysis is available at https://github.com/QiangSu/polyA_profiling. The code for poly(T) length determination is also available in methods.

ACCESSION NUMBERS

All sequencing raw data and processing data have been deposited on NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE202256. Any other relevant data are available from the authors upon request.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We thank Ms Qinyi Xia for providing HEK293T cells and Dr Kang Kang and Mr Haoyang Yin for providing A549 cells, Ms Shasha Zhang, and Mr Yongjie Liang for sequencing library preparation and preprocessing of sequencing data. We also thank all the members of the Prof Deming Gou laboratory for critical and useful discussions.

FUNDING

We acknowledge financial support from the National Natural Science Foundation of China (82170070, 91739109, 81970053), Guangdong Provincial Key Laboratory of Regional Immunity and Diseases (2019B030301009), Shenzhen Municipal Research Grant (JCYJ20190808123219295, JCYJ20210324120206017) and Shenzhen-Hong Kong Collaborative Innovation Research (SGDX20201103095404019).

CONFLICT OF INTEREST

The authors declare no financial interests.

REFERENCES

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* **30**: 2114-2120.

Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. 2013. Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS computational biology* **9**: e1003031.

Castellano LA, Bazzini AA. 2017. Poly(A) tails: longer is not always better. *Nat Struct Mol Biol* **24**: 1010-1011.

Chang H, Lim J, Ha M, Kim VN. 2014. TAIL-seq: Genome-wide Determination of Poly(A) Tail Length and 3' End Modifications. *Mol Cell* **53**: 1044-1052.

Doma MK, Parker R. 2007. RNA Quality Control in Eukaryotes. *Cell* **131**: 660-668.

Eisen TJ, Eichhorn SW, Subtelny AO, Lin KS, McGeary SE, Gupta S, Bartel DP. 2020. The Dynamics of Cytoplasmic mRNA Metabolism. *Molecular cell* **77**: 786-799.e710.

Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A et al. 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**: 201-206.

Garneau NL, Wilusz J, Wilusz CJ. 2007. The highways and byways of mRNA decay. *Nat Rev Mol Cell Biol* **8**: 113-126.

Gutiérrez Sánchez C, Su Q, Schönherr H, Grininger M, Nöll G. 2015. Multi-Ligand-Binding Flavoprotein Dodecin as a Key Element for Reversible Surface Modification in Nano-biotechnology. *ACS Nano* **9**: 3491-3500.

Houseley J, LaCava J, Tollervey D. 2006. RNA-quality control by the exosome. *Nat Rev Mol Cell Biol* **7**: 529-539.

Houseley J, Tollervey D. 2009. The many pathways of RNA degradation. *Cell* **136**: 763-776.

Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A et al. 2014. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science* **343**: 776-779.

Kechin A, Boyarskikh U, Kel A, Filipenko M. 2017. cutPrimers: A New Tool for Accurate Cutting of Primers from Reads of Targeted Next Generation Sequencing. *Journal of computational biology : a journal of computational molecular cell biology* **24**: 1138-1143.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**: 907-915.

Kolodziejczyk Aleksandra A, Kim JK, Svensson V, Marioni John C, Teichmann Sarah A. 2015. The Technology and Biology of Single-Cell RNA Sequencing. *Mol Cell* **58**: 610-620.

Ledergerber C, Dessimoz C. 2011. Base-calling for next-generation sequencing platforms. *Briefings in bioinformatics* **12**: 489-497.

Legnini I, Alles J, Karaiskos N, Ayoub S, Rajewsky N. 2019. FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control. *Nat Methods* **16**: 879-886.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**: 2078-2079.

Lim J, Lee M, Son A, Chang H, Kim VNJG, development. 2016. mTAIL-seq reveals dynamic poly (A) tail regulation in oocyte-to-embryo development. *Genes Dev* **30**: 1671-1682.

Lima SA, Chipman LB, Nicholson AL, Chen Y-H, Yee BA, Yeo GW, Coller J, Pasquinelli AE. 2017. Short poly(A) tails are a conserved feature of highly expressed genes. *Nat Struct Mol Biol* **24**: 1057-1063.

Liu Y, Nie H, Liu H, Lu F. 2019. Poly(A) inclusive RNA isoform sequencing (PAIso−seq) reveals wide-spread non-adenosine residues within RNA poly(A) tails. *Nat Commun* **10**: 5292.

Long Y, Jia J, Mo W, Jin X, Zhai J. 2021. 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. *Nature Protocols* **16**: 4355-4381.

Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, Liu Y, Chen X, Newman S, Nakitandwe J et al. 2019. Analysis of error profiles in deep next-generation sequencing data. *Genome biology* **20**: 50.

Macosko Evan Z, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas Allison R, Kamitaki N, Martersteck Emily M et al. 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**: 1202-1214.

Maguire S, Lohman GJS, Guan S. 2020. A low-bias and sensitive small RNA library preparation method using randomized splint ligation. *Nucleic acids research* **48**: e80.

Mattijssen S, Iben JR, Li T, Coon SL, Maraia RJ. 2020. Single molecule poly(A) tail-seq shows LARP4 opposes deadenylation throughout mRNA lifespan with most impact on short tails. *eLife* **9**.

Mercer TR, Neph S, Dinger ME, Crawford J, Smith MA, Shearwood AM, Haugen E, Bracken CP, Rackham O, Stamatoyannopoulos JA et al. 2011. The human mitochondrial transcriptome. *Cell* **146**: 645-658.

Moll P, Ante M, Seitz A, Reda T. 2014. QuantSeq 3′ mRNA sequencing for RNA quantification. *Nature methods* **11**: i-iii.

Nam DK, Lee S, Zhou G, Cao X, Wang C, Clark T, Chen J, Rowley JD, Wang SM. 2002. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(A) priming during reverse transcription. *Proc Natl Acad Sci USA* **99**: 6152-6156.

Nicholson-Shaw AL, Kofman ER, Yeo GW, Pasquinelli AE. 2022. Nuclear and cytoplasmic poly(A) binding proteins (PABPs) favor distinct transcripts and isoforms. *Nucleic acids research* **50**: 4685-4702.

Nicholson AL, Pasquinelli AE. 2019. Tales of Detailed Poly(A) Tails. *Trends Cell Biol* **29**: 191-200.

Park JE, Yi H, Kim Y, Chang H, Kim VN. 2016. Regulation of Poly(A) Tail and Translation during the Somatic Cell Cycle. *Molecular cell* **62**: 462-471.

Passmore LA, Coller J. 2022. Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. *Nat Rev Mol Cell Biol* **23**: 93-106.

Pekarsky Y, Balatti V, Palamarchuk A, Rizzotto L, Veneziano D, Nigita G, Rassenti LZ, Pass HI, Kipps TJ, Liu C-g et al. 2016. Dysregulation of a family of short noncoding RNAs, tsRNAs, in human cancer. *Proc Natl Acad Sci USA* **113**: 5071.

Quail MA, Otto TD, Gu Y, Harris SR, Skelly TF, McQuillan JA, Swerdlow HP, Oyola SO. 2012a. Optimal enzymes for amplifying sequencing libraries. *Nat Methods* **9**: 10-11.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012b. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genom* **13**: 341.

Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC et al. 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**: 777-782.

Sallés FJ, Strickland S. 1999. Analysis of poly(A) tail lengths by PCR: the PAT assay. *Methods Mol Biol* **118**: 441-448.

Schaffer AE, Pinkard O, Coller JM. 2019. tRNA Metabolism and Neurodevelopmental Disorders. *Annu Rev Genomics Hum Genet* **20**: 359-387.

Slack FJ, Chinnaiyan AM. 2019. The Role of Non-coding RNAs in Oncology. *Cell* **179**: 1033-1055.

Stoler N, Nekrutenko A. 2021. Sequencing error profiles of Illumina sequencing instruments. *NAR genomics and bioinformatics* **3**: lqab019.

Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. 2014. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**: 66-71.

Wang L, Nie J, Sicotte H, Li Y, Eckel-Passow JE, Dasari S, Vedell PT, Barman P, Wang L, Weinshiboum R et al. 2016. Measure transcript integrity using RNA-seq data. *BMC bioinformatics* **17**: 58.

Wang L, Wang S, Li W. 2012. RSeQC: quality control of RNA-seq experiments. *Bioinformatics (Oxford, England)* **28**: 2184-2185.

Wery M, Descrimes M, Thermes C, Gautheret D, Morillon A. 2013. Zinc-mediated RNA fragmentation allows robust transcript reassembly upon whole transcriptome RNA-Seq. *Methods (San Diego, Calif)* **63**: 25-31.

Whiteford N, Skelly T, Curtis C, Ritchie ME, Löhr A, Zaranek AW, Abnizova I, Brown C. 2009. Swift: primary data analysis for the Illumina Solexa sequencing platform. *Bioinformatics (Oxford, England)* **25**: 2194-2199.

Wilusz JE, Whipple JM, Phizicky EM, Sharp PA. 2011. tRNAs marked with CCACCA are targeted for degradation. *Science* **334**: 817-821.