circTAIL-seq, a targeted method for deep analysis of RNA 3′ tails, reveals transcript-specific differences by multiple metrics

VAHID H. GAZESTANI,[1] MARSHALL HAMPTON,[2] JUAN E. ABRAHANTE,[3] REZA SALAVATI,[1] and SARA L. ZIMMER[4]

[1]Institute of Parasitology, McGill University, Québec H9X 3V9, Canada
[2]Department of Mathematics, University of Minnesota Duluth, Duluth, Minnesota 55812, USA
[3]University of Minnesota Informatics Institute, University of Minnesota, Minneapolis, Minnesota 55455, USA
[4]Department of Biomedical Sciences, University of Minnesota Medical School, Duluth, Minnesota 55812, USA

ABSTRACT

Post-transcriptionally added RNA 3′ nucleotide extensions, or tails, impose numerous regulatory effects on RNAs, including effects on RNA turnover and translation. However, efficient methods for in-depth tail profiling of a transcript of interest are still lacking, hindering available knowledge particularly of tail populations that are highly heterogeneous. Here, we developed a targeted approach, termed circTAIL-seq, to quantify both major and subtle differences of heterogeneous tail populations. As proof-of-principle, we show that circTAIL-seq quantifies the differences in tail qualities between two selected *Trypanosoma brucei* mitochondrial transcripts. The results demonstrate the power of the developed method in identification, discrimination, and quantification of different tail states that the population of one transcript can possess. We further show that circTAIL-seq can detect the tail characteristics for variants of transcripts that are not easily detectable by conventional approaches, such as degradation intermediates. Our findings are not only well supported by previous knowledge, but they also expand this knowledge and provide experimental evidence for previous hypotheses. In the future, this approach can be used to determine changes in tail qualities in response to environmental or internal stimuli, or upon silencing of genes of interest in mRNA-processing pathways. In summary, circTAIL-seq is an effective tool for comparing nonencoded RNA tails, especially when the tails are extremely variable or transcript of interest is low abundance.

Keywords: Illumina; mitochondrion; polyadenylation; trypanosome; uridylation

INTRODUCTION

The majority of eukaryotic transcripts acquire a nontemplated nucleotide addition or "tail" on their 3′ termini after or nearly simultaneously with transcription. Although tail addition appears ubiquitous, tail length and composition are finely regulated in various cell compartments (e.g., nucleus, cytoplasm, mitochondria, and chloroplast) with implications for the mRNAs stability, transport, and translation initiation (Zhang et al. 2010; Norbury 2013). Interestingly, the regulatory roles of these end structures can differ based on cellular needs. For example, while extension of tails in the early embryonic stages of zebrafish and xenopus enhances the translation rate of cognate transcripts, experimental data do not support the functionality of this regulatory process in the later life stages (Subtelny et al. 2014). Moreover, tails may have different states, each with a distinct regulatory role and biological impact on transcripts; e.g., while one tail state regulates the stability of a transcript, the other state with possibly differing tail length and/or composition regulates the translational rate of the transcript (Aphasizheva et al. 2011).

The average tail length of cytoplasmic transcripts varies between different organisms from relatively short lengths of 20–30 nt in yeast to less than 100 for some metazoan organisms (Chang et al. 2014; Subtelny et al. 2014). The situation for the organellar mRNAs is also varied. Yeast mitochondrial transcripts (mtRNAs) entirely lack tails, adenine (A)-rich and uridine (U)-rich oligomers are part of the chloroplast and mitochondrial mRNA decay pathways in plants and algae, and multiple roles of poly(A) and other tails on human mtRNAs appear transcript-specific (Schuster and Stern

Corresponding authors: szimmer3@d.umn.edu, reza.salavati@mcgill.ca
Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.054494.115.

2009; Zimmer et al. 2009; Chang and Tong 2012; Rorbach and Minczuk 2012). However, the biological implications of tail variations within and across organisms are largely unknown.

The *Trypanosoma brucei* mitochondrion is an interesting system to study tailing mechanisms and their regulatory impacts on the transcripts, where an apparently convoluted tailing process is intertwined with other post-transcriptional events. Expression of *T. brucei* mitochondrial genes starts with their constitutive polycistronic transcription, followed by processing that is coupled with addition of fairly ubiquitous relatively short tails. These are termed here as (in)itial tails or "in-tails" and are thought to mediate the transcript stability (Ryan et al. 2003; Kao and Read 2005; Etheridge et al. 2008; Aphasizheva and Aphasizhev 2010). Tail addition and modification are also linked to the unique type of editing that 12 trypanosome mtRNAs must undergo prior to translation (for review, see Stuart et al. 2005; Aphasizhev and Aphasizheva 2011, 2014; Hashimi et al. 2013). Finally, transcripts are potentially marked for translation by extensions appended to a subset of in-tails on translatable mtRNAs only. Extensions contain both A and U and are described as having a 7:3 A/U ratio (Etheridge et al. 2008), with a fairly consistent frequency of switching of addition from A to U and back. We are naming these latter, presumably translation-associated tails that possess these described extensions "ex-tails."

During the past decades, biochemical and other approaches have been developed for the identification of tail characteristics of populations of individual genes (Temperley et al. 2003; Beilharz and Preiss 2011; Slomovic and Schuster 2013). Some of these approaches provide only qualitative information on the tail length. Others such as circular RT-PCR or cloning of end-adapted 3′ ends are labor intensive and thus examine characteristics of a relatively small sample of tails from the transcript of interest. These limitations hamper their application to: (i) quantitative comparison of tail characteristics of the transcript in different biological conditions; (ii) accurate description of multiple tail states in the tail population of the transcript; and (iii) identification of tail characteristics for rare, but biologically interesting transcripts or degradation intermediates.

High-throughput sequencing approaches to genome-wide tail inference of transcripts are now available and have profoundly expanded our understanding of the functions and regulatory potentials of tails (Chang et al. 2014; Slevin et al. 2014; Subtelny et al. 2014; Welch et al. 2015). However, due to the genome-wide design of these approaches, the number of sampled tails for most transcripts is still below 100 and can be even significantly less than that (or zero) for low-abundance transcripts, a limitation that might explain, in some extent, the observed discrepancies between the results of some of these approaches (Lee et al. 2014; Zheng and Tian 2014). Therefore, genome-wide approaches are not universally suitable for in-depth analysis of tail characteristics for a focused subset of transcripts of interest.

In this work, we developed an approach to characterize tails on transcript populations and quantitatively compared tail populations of transcripts. For in depth and high-resolution analysis of tail population for a transcript of interest, we coupled conventional circular reverse transcription—polymerase chain reaction (cRT-PCR) to next-generation sequencing techniques and termed this "circTAIL-seq." As proof-of-principle, we applied the circTAIL-seq approach to mtRNAs of *T. brucei*. The depth of circTAIL-seq allowed us to accurately detect and quantify different tail states and subtle differences of tail populations. Furthermore, we also captured tail characteristics of rare but biologically intriguing variants of mtRNA ends that would likely be missed with other approaches. The diversity of tail lengths and compositions on trypanosome mtRNAs proved a tremendous asset for circTAIL-seq development. We describe a methodology that addresses both experimental and computational aspects to identify and characterize the tail population of target transcripts.

RESULTS

Capturing tail census of a transcript by circTAIL-seq

The circTAIL-seq approach is composed of three major steps: library generation, next-generation sequencing, and the informatics workflow to extract tail information from the raw sequencing output (Fig. 1A). Library generation parallels the conventional 3′ tail analysis of cRT-PCR used to investigate RNA 5′ and 3′ ends. (Perrin et al. 2004a,b; Slomovic and Schuster 2008, 2013; Aphasizheva and Aphasizhev 2010; Aphasizheva et al. 2011; Zimmer et al. 2012). Total RNA is first circularized using RNA ligase. Next, carefully positioned gene-specific primers containing adaptor sequence are used in reverse transcription and PCR to generate tail-containing amplicons bridging the 3′–5′ junction that can be directly used as Illumina sequencing libraries. As detailed in Materials and Methods, the obtained reads are then preprocessed and subsequently aligned to the reference sequence for the gene of interest to identify the embedded tails and associated 3′ and 5′ termini sites (Fig. 1B).

We initially performed pilot study sequencing of two *T. brucei* mitochondrial transcript tail populations acquired as described above. Results clearly indicated that sequential method optimization for circTAIL-seq was required. Table 1 describes read usability between the pilot experiment and two subsequent trials (total of three separate library preparation and sequencing trials). The initial trial was performed at the smallest possible sequencing scale on two single transcript amplicons only. With it, we confirmed that amplicons could be subjected to deep sequencing, and also obtained reads with which to develop an informatics workflow. Studies of 5′ and 3′ RNA ends often use end-ligation of adaptors to RNA, eliminating the need for gene-specific reverse primers. In this first trial, we also sequenced amplicons of 3′ end-adapted rather than circularized RNAs generated as

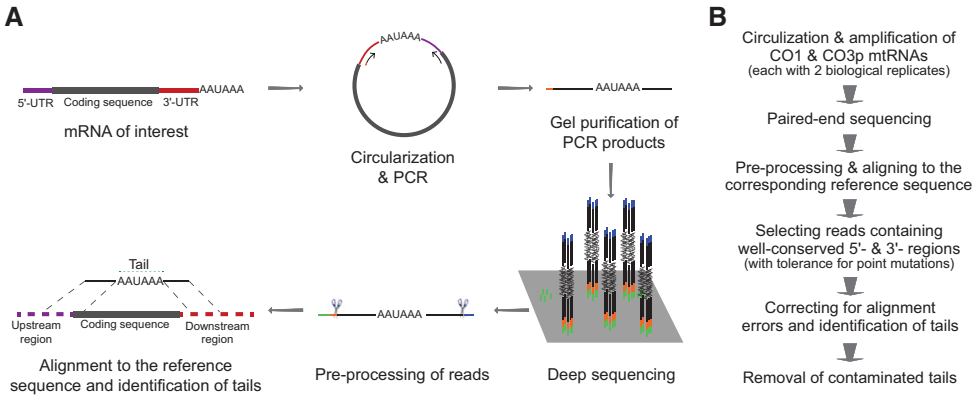

FIGURE 1. (*A*) Schematic illustrating steps of circTAIL-seq. Thick gray line indicates a coding region of a generic mRNA that is not amplified in the process. The generic RNAs 5′ UTR and 3′ UTRs are represented in violet and red, respectively. "AAUAAA" represents all potential nonencoded tails on the ends of the RNAs. The orange region on PCR products is the bar code introduced during PCR. (*B*) Informatics workflow for circTAIL-seq data analysis.

described for miRNA end-sequencing (Diebel et al. 2010) facilitated by preadenylation of the adaptor (Hafner et al. 2008). As we obtained only a few hundred gene-specific reads within the entire Illumina read file from end-adapted amplicons (not shown), we proceeded with optimization of circTAIL-seq.

The second trial was performed on a larger set of individual libraries that more realistically mirrored the number of samples we anticipate when circTAIL-seq is experimentally applied. We included six transcripts in biological replicate (12 total) in order to gain insight into reproducibility. Reads for the first two trial experiments confirmed that

TABLE 1. Quality of sequencing output during method development

Trial	MiSeq scale	Multiplex	Read pairing method	RNA	Replicate	Lowest MQS	Bar-coded reads	% Reads with primer sequence
1	nano	2	fastq-join	CO1		28	89,284	92.5%
				mRNA B		33	144,914	44.0%
2	nano	6	fastq-join	CO1	r1	22	38,187	83.5%
					r2	28	10,041	81.0%
				mRNA A	r1	32	17,717	55.3%
					r2	28	24,354	86.7%
				mRNA B	r1	31	40,902	65.8%
					r2	31	9998	51.4%
				mRNA C	r1	27	3118	55.0%
					r2	27	18,522	59.2%
				CO3p	r1	21	206,445	artifacts
					r2	18	80,422	artifacts
				mRNA D	r1	18	73,668	artifacts
					r2	19	160,920	artifacts
3[a]	V2	6	PEAR	CO1[a]	r1	32	456,410	90.0%
					r2	32	165,529	98.8%
				mRNA A	r1	32	730,684	90.2%
					r2	33	867,025	99.9%
				mRNA B	r1	35	151,966	94.4%
					r2	34	179,089	93.9%
				mRNA C	r1	35	223,581	95.6%
					r2	33	581,992	95.3%
				CO3p[a]	r1	33	227,924	93.5%
					r2	34	235,995	93.8%
				mRNA D	r1	33	614,614	94.9%
					r2	31	1,421,938	94.8%

Lowest mean quality score is across all positions using FastQC quality control analysis. Analyzed for quality score are the single R1 read for Experiments 1 and 2, and the paired reads for Experiment 3. (MQS) Mean quality score. Note: Sample replicates are designated by lower-case "r" and read direction with upper-case "R." (Artifacts) Indicates that the vast majority of reads for this sample were PCR artifacts.
[a]Transcripts selected for analysis of tail characteristics in this study.

primers annealed in locations resulting in generated amplicons containing enough transcript 5′ and 3′ end sequence for the subsequent tail extraction program (Table 2). Unfortunately, very few tails that met the traditional definition of an ex-tail (Etheridge et al. 2008; Aphasizheva et al. 2011; Zimmer et al. 2012) were observed in either of the first two trials; we discovered instead that most of the longest reads were artifacts resulting from aberrant PCR amplification. Thus, optimization of the amplicon-generating PCR reaction was necessary to reduce or eliminate artifacts. We developed a PCR optimization protocol (see Supplemental Method) to optimize PCR reactions for each transcript. In addition, we changed our thermostable polymerase to one requiring short annealing and extension times that seems to reduce artifact abundance (data not shown). In conclusion, our first two optimization experiments suggest that for any circTAIL-seq experiments, especially those involving transcripts of unknown 3′ and 5′ UTR length, a trial amplicon sequencing at nano scale to verify identity of generated amplicons and good primer selection location is prudent.

The third sequencing trial performed on the same targets as trial 2 with amplicons generated using optimized PCR provided us ample reads and very few artifacts. There were no clear influences of read yield on analysis (Figs. 2, 3; Table 3) in this trial where analyzable tails varied from ∼150,000 to ∼1.5 million per sample. Nor was there an influence attributable to having a particular barcode in the primer sequence. High proportions of the returned reads contained primer annealing sequence and were deemed usable (Table 1). Encouragingly, some of the sample files contained reads in which ex-tails were observed by manual perusal. Output from the circTAIL-seq tails extraction workflow for all three transcripts can be found in Table 3 for this trial, which is what we used for subsequent analysis. Importantly, the fraction of reads appearing only once in each sequencing run (ranging from 4.5% to 38%) indicated that abundance of circularized mitochondrial templates in the initial RNA pool is not limiting, suggesting that even by devoting high-throughput sequencing to a prespecified transcript, we may not fully capture the tail diversity of the transcript (Supplemental Fig. S1). Finally, we note that sample PCR product abundances are quantitated individually and multiplexed with a goal of equal proportions of each sample the multiplexed sequencing reaction. Therefore, differences between tail populations to be compared will typically be less than an order of magnitude, sometimes differing no more than twofold (Table 3). In conclusion, we have developed amplicon generation and sequencing protocols for circTAIL-seq that provide adequate sample sizes, high percentages of usable reads, and the potential for capturing longer tail sequences.

TABLE 2. Length of UTRs based on the most common terminus shown in Figure 4 for transcripts for which tails were analyzed in our proof-of-concept experiments

RNA	3′ Approx. UTR	Distance 3′ primer to approx. 3′ end	5′ Approx. UTR	Distance 5′ primer to approx. 5′ end	Tail length possible to capture
CO1	26 nt	74 nt	32 nt	62 nt	144 nt
CO3p	39 nt	20 nt	30 nt	34 nt	217 nt

Distance of primer to average end of RNA includes length of primer sequence. Maximum tail length assumes 150 bp paired-end reads and a minimum overlap of 10 nt for pairing reads.

Two *T. brucei* mitochondrial transcripts were selected to demonstrate the analytical potential of circTAIL-seq

We next developed methodologies that could eventually be used to compare mtRNA tails between transcripts and changes in tails in response to internal and/or environmental stimuli. An analysis of our entire third sequencing trial data set and the resulting implications for trypanosome biology are too extensive to be reported here, and will be presented elsewhere (VH Gazestani et al., in prep.). Instead, we report here methodology development and proof-of-principle for circTAIL-seq using selected replicate tail populations of two transcripts only, profiling tail population for the mitochondrial transcripts of CO1 (cytochrome oxidase subunit I) and pre-edited CO3 (CO3p; encoding cytochrome oxidase subunit III). The tailing process of *T. brucei* mitochondrial transcripts is regulated and is coupled with the RNA editing status of transcripts, i.e., although most mitochondrial transcripts acquire in-tails, only the transcripts with correct open reading frames (ORFs) can undergo the ex-tailing process that mark them for translation. CO1 gene encodes the transcripts with correct ORFs, so it is never edited and it can be translated upon cleavage to the monocistronic form. It is known from conventional RNA blots that a sub-population of CO1 is ex-tailed, although details of the length differential are not possible to garner with that method. In contrast, CO3p transcript does not possess a translatable ORF. Thus it should not be associated with the ribosome and we would not expect it to be ex-tailed (Aphasizheva et al. 2011). Additionally, limited published tail sequences suggest that the typical length and composition of CO1 and CO3 in-tails would likely be different (Decker and Sollner-Webb 1990; Kao and Read 2007). Because of these differences, tail populations of CO1 and CO3p transcripts were used to illustrate the functionality of circTAIL-seq results.

Analysis of circTAIL-seq data demonstrates complexity in tail populations that may not be captured in low-resolution settings

Large-scale analysis of tail populations has been performed on tails of cytosolic mRNAs (Chang et al. 2014; Lim et al. 2014; Subtelny et al. 2014), histone mRNA degradation products (Slevin et al. 2014; Welch et al. 2015), and miRNAs

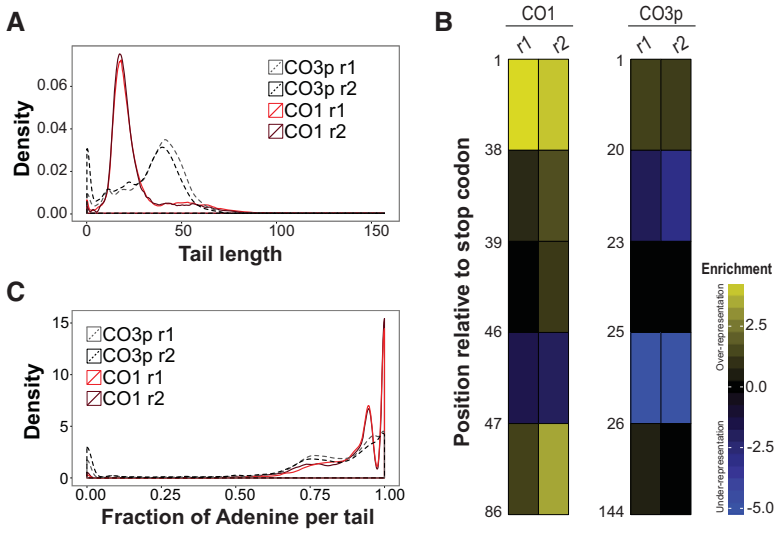

FIGURE 2. (*A*) Density curves comparing lengths of tail populations from CO1 or CO3p transcripts. CO1 tails >100 nt are present but are not abundant enough to be observed on a chart of this scale. (*B*) Enrichment analysis of tail-less reads in five roughly equally populated bins. The figure is pseudo-colored, showing only significant enrichments (*P*-values <0.01 and fold-enrichment >1.5), with blue and yellow colors indicating underrepresentation and overrepresentation of tail-less reads, respectively. (*C*) Density curves comparing fraction nucleotides that are "A" in each tail [0 = tails that are oligo(U), 1 = tails that are oligo(A)] from CO1 and CO3p transcripts. Distributions were inferred using R statistical package. Area *under* each curve is 1. "r1" and "r2" are biological replicates.

(Wyman et al. 2011), yet trypanosome tails are far more complex than these. Analyzing complementary aspect of tails with multiple metrics proved useful because it provided multiple lines of evidence from which to draw conclusions. Moreover, to ascertain that our reported results are not overwhelmingly affected by potential PCR amplification bias favoring short length sequences, we verified that reported results are supported by the analysis of unique reads (not shown) as well as total reads. This was possible as a benefit of the extremely high heterogeneity of read sequences.

Tail length

The simplest tail characteristic to describe is its length. We constructed probability density functions from the collected tails in each sample, with the area under each density curve set to 1. Profiles for replicate samples for CO1 and CO3p are displayed on single graphs in Figure 2A, allowing us to analyze reproducibility and transcript-specific differences. The density curves demonstrated overall good reproducibility of tail length data. Comparison of density curves indicated that CO3p transcripts have significantly longer tails than CO1 (P-value $< 2.2 \times 10^{-16}$, Wilcoxon–Mann–Whitney rank-sum test with pooling the replicates). Additionally, CO3p tail length distribution showed a wide

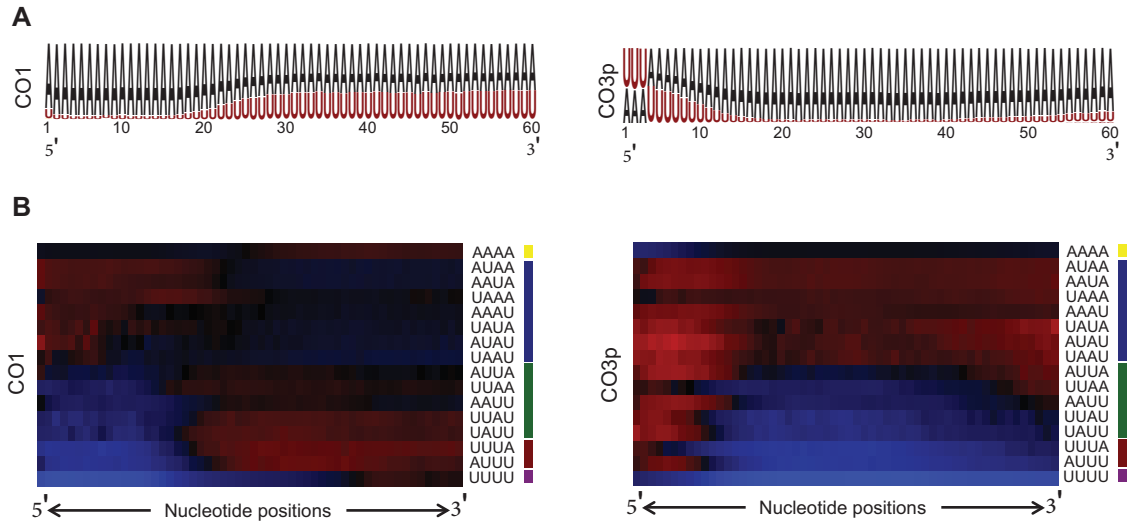

FIGURE 3. (*A*) Relative abundance of each nucleotide at each position from tail positions 1–60. All tails possessing a nucleotide at the analyzed position in the total population were considered. (*B*) Heat map describing the occurrence of the indicated tetramer (nucleotide position 1 of the tetramer at the position indicated at the *bottom*), relative to the likelihood of that tetramer given the average nucleotide compositions at those positions under an independently distributed model. "Nucleotide positions" are tail positions 1–60 from 5′ to 3′ as in *A*. Both plots are colored on the same scale; the lowest (most intense red) value was 0.0217 (occurring 1/46th as often as expected under an independently distributed model), and the highest (most intense blue) value was 3356 (occurring 3356 times as often as expected under an independently distributed model). At each position, the entire population of tails possessing a full tetramer starting at that position is analyzed.

TABLE 3. Summary of circTAIL-seq read analysis

RNA	Replicate	Number reads input	Number reads analyzed	Number of different tails	Occurrence of most frequent tail	Tails occurring one time only in file	Reads with untailed 3′ ends
CO1[a]	r1	410,922	376,011 (91.5%)	39,580	14,833	22,810 (6.0%)	3161 (0.84%)
	r2	163,552	140,534 (85.9%)	18,266	5806	10,609 (7.5%)	642 (0.46%)
mRNA A	r1	659,230	594,279 (90.1%)	49,582	12,783	27,177 (4.5%)	1175 (0.20%)
	r2	862,249	779,547 (90.4%)	71,989	16,319	39,865 (5.1%)	3384 (0.43%)
mRNA B	r1	143,505	101,048 (70.4%)	49,238	3423	38,377 (38.0%)	3423 (3.39%)
	r2	168,156	124,288 (73.9%)	47,608	6307	34,458 (27.8%)	6307 (5.07%)
mRNA C	r1	213,730	163,559 (76.5%)	76,787	5520	58,086 (35.5%)	5520 (3.38%)
	r2	554,785	474,264 (85.5%)	216,792	10,608	158,682 (33.5%)	10,608 (2.24%)
CO3p[a]	r1	212,992	168,268 (79.0%)	52,027	3601	35,528 (21.1%)	3601 (2.14%)
	r2	221,001	187,542 (84.9%)	50,541	16,548	34,709 (18.5%)	16,548 (8.82%)
mRNA D	r1	582,922	434,990 (74.6%)	111,743	6193	71,981 (16.5%)	6193 (1.42%)
	r2	1,347,417	1,003,267 (74.5%)	227,514	13,685	148,494 (14.8%)	13,685 (1.36%)

Number reads input are reads containing gene-specific primer annealing region of reverse amplification primer for the 5′ end at 80% similarity or greater. These are the sequences that were input (column 3) for the circTAIL-seq Analyzer program. Number reads analyzed are reads with acceptable sequence conservation to the 5′ and 3′ regions of the reference sequence for each transcript, from which tails were extracted. "Un-tailed" was considered a single type of tail. Note: Sample replicates are designated by lower-case "r" as opposed to read direction with upper-case "R."
[a]Transcripts selected for analysis of tail characteristics in this study.

variation with two major peaks, while the majority of CO1 tails were narrowly distributed in length with a second minor population in a longer length category of 50 or more nucleotides. Thus, circTAIL-seq allows for deep analysis of even basic characteristics such as length.

Tail-less reads

Tail length density curves also demonstrated the presence of tail-less reads in circTAIL-seq results (Fig. 2A; Table 3). The percentage of CO1 tails lacking reads ranged from 0.02% to 0.84%, barely detectable, while CO3p contained higher percentages of tail-less reads ranging from 1.4% to 8.8% (total number of tails per sample appears in Table 3). We found that tail-less reads reproducibly have significantly shorter 3′ untranslated regions (UTRs) than the tailed reads in all biological replicates of both transcripts (P-value $< 2.2 \times 10^{-10}$, Wilcoxon–Mann–Whitney test). Shorter 3′ ends may in fact be decay intermediates, suggesting the circTAIL-seq approach has been able to capture this difficult-to-detect population. Another interesting observation was the moderate enrichment of tail-less reads with unusually long embedded 3′ UTRs (Fig. 2B). The polycistronic nature of mitochondrial transcription suggests that these reads might be RNAs currently undergoing 3′ exonucleolytic cleavage to generate the mature, typical 3′ end from a larger precursor. Evidence such as this would be supportive of the idea that 3′ exoribonucleases act during polycistronic mtRNA processing (Mattiacio and Read 2008; Aphasizhev and Aphasizheva 2011). Hence, circTAIL-seq analysis of all mtRNAs could yield valuable insights into the process of polycistron processing, of which little is known.

Tail composition

The other important characteristic of 3′ RNA tails is their overall nucleotide composition. Figure 2C presents density curves for the percentage of A in total composition from none (value of 0) to 100% (value of 1) *per tail* for each sample's population. We found that CO3p incorporates far more Us than CO1. This was surprising as CO3p transcripts are expected to be exclusively in an in-tailed state that reports commonly described as primarily poly(A) (Bhat et al. 1992; Etheridge et al. 2008; Aphasizhev and Aphasizheva 2011; Aphasizheva et al. 2011). To probe this distinction, we examined the nucleotide composition in each tail population as a function of nucleotide position in the tail (Fig. 3A) by first merging the two biological replicates of each transcript. The analysis was performed out to 60 nt, after which no additional differences were observed (not shown).

Results for tails on the two transcript populations analyzed indicated A is the most common nucleotide in all positions but the first three nucleotides of CO3p tails. However, U abundance not only differed between transcripts, but also showed transcript-specific variability in positioning. In CO1 tails, Us are rare until about position 17, where they increase in frequency to reach to the relative frequency of about 1/3 in position ~28, and remain near this frequency afterwards. Assuming ex-tail sequence extensions to start somewhere between tail position 20 and 40 nt (Aphasizheva et al. 2011; Zimmer et al. 2012), this observation on the CO1 transcript is consistent with a transition from an in-tail (mainly As) to an ex-tail (extension of the in-tail with A/U composition of approximately 7:3 ratio). In contrast, Us on CO3p tails are at positions consistent with belonging to in-tails, demonstrating that U can be a common

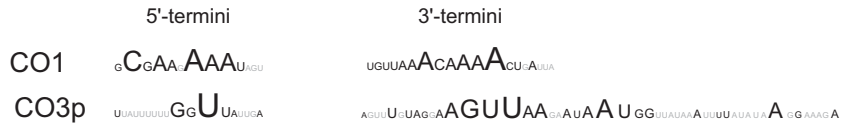

FIGURE 4. 3′ and 5′ UTR termini derived from populations of circularized molecules. Replicate experiments have been pooled by first normalizing for tail counts, so average density for each nucleotide is shown. Black represents the termini with occurrence probability >0.01%, with the size of the nucleotide corresponding to its frequency as a terminus. The gray nucleotides represent positions with probabilities between 0.1% and zero that exist between nucleotide positions that are more frequently termini. Other nucleotides that are termini with a probability <0.1% are not shown.

component of in-tails for some mitochondrial transcripts. These observations indicate that analyzing tail composition by position (possible only with high numbers of tail sequences) can provide important insights not possible by analyzing overall composition alone. To summarize our entire analysis of tail composition, CO1 and CO3p appear to have overall nucleotide compositional differences (Fig. 2C) as well as positional signatures (Fig. 3A).

Other differences in nucleotide patterns

While the above metrics are highly informative, further details can be elucidated from the tails data that shed light on potential differences in in-tail and ex-tail A/U addition patterns. For instance, a tail comprised of 50% A and 50% U can consist of alternations of single As and Us, or of alternating homopolymer stretches of As and Us. The initial regions of tails in Figure 3A represent the sum of the entire population, and therefore could consist of sub-populations of A and U homopolymers, or a single population that is fairly heteropolymeric. In Figure 3B we present deviation from expected probability of each listed nucleotide tetramer at every position along the first 60 nt of all tails. These heat maps are arranged so that poly(A) tetramer is the top listing, combinations of single Us within A stretches are next, then alternating A/U combinations, next U polymer stretches interspersed with A, and finally a U tetramer. Therefore, vertical locations in the heat map are more or less associated with multiple versus single additions of a certain nucleotide.

Figure 3B demonstrates that with tail numbers obtained from circTAIL-seq, we can observe variable tetramer probability frequencies, with different tail nucleotide heat map patterns observed for CO1 and CO3p. U polymers of two nucleotides or more were overrepresented in CO3p tails along the entire analyzed region of 60 nt with the exception of the first 10 nt, where only U tetramers were overrepresented. In contrast, single Us within the sequence were either underrepresented or occurred approximately as expected. As CO3p is a pre-edited transcript that should not associate with the ribosome, we do not expect to see a transition to ex-tails with frequent A/U alterations. This result is consistent with such an expectation.

While U polymer-containing tetramers are also overrepresented in CO1 tails, this overrepresentation only occurs within the first 20–30 nt. This likely relates to the fact that for CO1, nucleotide addition beyond the first 30 or 40 is in ex-tail state and thus switching of A/U addition is far more frequent. Analyzing the deviation from expected probability of these different tetramers provided important observations, especially when we concentrate on differences specifically between tetramers consisting of U stretches of two or more (homopolymer) compared to tetramers containing single Us. Overall, in tails from a transcript where we not expect to find ex-tails (CO3p), and initial tail regions of tail populations from a transcript where we expect to find ex-tails (CO1), U homopolymers are overrepresented. In contrast, we see tetramers containing single Us become overrepresented in the 3′ end of CO1 where we expect to find sequence consistent with ex-tail additions. Until now, demonstrating the existence of an ex-tail among tail sequences required publishing the tail sequence and indicating the site of transition to frequent nucleotide switching compared to the beginning of the tail (Etheridge et al. 2008; Aphasizheva et al. 2011). This tetramer analysis therefore represents a way to demonstrate an increase or change in frequency of nucleotide switching in an entire population. We note, however, that although our sequencing strategy detects both types of tails, because of possible PCR shorter-tail amplification bias, the relative abundance of ex-tails to in-tails may not be a true representative of the underlying populations. Because of this, transitions in deviations from expected tetramer frequencies may be even more dramatic than suggested by our plots.

circTAIL-seq can capture differences in both 5′ and 3′ UTR lengths

Aligning reads to a reference sequence also provided 3′ and 5′ termini information for each transcript population. Sequenced amplicons can thus reveal UTR lengths and degree of termini homogeneity. For instance, analysis of the 5′ termini derived from CO3p reads is shown in Figure 4, and the most frequently encountered 5′ terminus (the largest "U" in 5′ UTR sequence) was the terminus previously identified by primer extension (A Estevez and L Simpson, unpubl.). Therefore, transcript termini can also be defined by circTAIL-seq.

This is particularly useful when UTRs are heterogeneous in length, such as the 3′ UTR of CO3p, which has much higher length heterogeneity than what we observed for CO1. We hypothesize that transcripts with high UTR length heterogeneity undergo more exonuclease processing after a downstream cleavage than transcripts such as CO1. In contrast, CO1 possesses a 3′ UTR that could have been generated by a tightly

controlled endonuclease cleavage event, although proving this is well beyond this study's scope.

DISCUSSION

Here we have developed experimental and computational methodologies that allow coupling of circular RT-PCR tail analysis with high-throughput sequencing technology. Development required optimization of tail-containing amplicon generation, Illumina sequencing modifications to adjust for amplicons containing problematic regions of identical sequence, and development of a stringent informatics workflow to separate true tails from contaminating sequences. Using this rigorous process we obtain 100,000–1,000,000 tails per transcript amplicon.

circTAIL-seq has three major advantages over conventional circular RT-PCR approach. First, it eliminates the cloning step that is the most labor intensive, and thus limiting step of the circular RT-PCR approach. Second, the bias of circularized RT-PCR approach toward collecting a population's shortest tails, introduced in both the PCR and cloning steps, is reduced (albeit not removed) by eliminating the cloning step. Benefiting from the extremely high heterogeneity of read sequences that stem from variations in tail length and composition as well as 5′ and 3′ termini, we were able to show that the circTAIL-seq data are minimally affected by PCR bias as considering either total reads or unique reads for the analysis led mainly to the same results. For future experiments, a spike-in addition of RNAs of known length and frequency could be added to remove the bias completely if it is deemed necessary.

Third and most importantly, although augmentation of tails from low-abundance cytosolic reads is possible (Welch et al. 2015), circTAIL-seq, to the best of our knowledge, provides highest depth of tail analysis compared to other techniques developed thus far. It provides a high-resolution picture on tail population for specific transcripts of interest, and is possibly the only current method that is efficient for analyzing 5′ as well as 3′ ends of organellar RNAs. This critical characteristic empowers the user to go beyond overall average tail characteristics to examine interesting sub-populations of tails within a data set and identify specific nucleotide patterns that appear in populations. Applications of this approach include determining changes in tail qualities in response to environmental or internal stimuli, or upon silencing of genes of interest in mRNA-processing pathways. Especially when transcript tail populations that are an investigative focus prove low abundance relative to other tails in the sequenced population, and/or populations are highly heterogeneous such as on decay intermediates in *Chlamydomonas* chloroplasts and mitochondria (Zimmer et al. 2009), circTAIL-seq will prove invaluable.

We have demonstrated the range of characteristics that can be compared in large tail data sets using this methodology, and provided evidence that previously unknown differences between tail populations exist on trypanosome mtRNAs, many of which may not be captured in lower resolution settings. We were able to define transcript-specific differences in tail populations and concurrently define population-wide 3′ and 5′ termini of the transcripts. These sequencing and analysis methods can potentially be used to describe and compare 3′ and 5′ ends of any sort of RNA when specific transcripts are the study's focus.

Finally, the high-throughput sequencing setting of circTAIL-seq approach can be adjusted based on the expected tail characteristics for the transcript of interest. Here, the employed setting, 150 bp paired-end, was selected based on the previous biological knowledge on tail characteristics of *T. brucei* mtRNAs which are highly heterogeneous (composed of As and Us) with almost no A homopolymer stretches of longer than 30 nt. Therefore, inaccurate quantitation of length of homopolymers problematic in Illumina sequencing of poly(T) stretches (Chang et al. 2014) is not a problem in our study, but could be addressed in a context where circTAIL-seq was used to analyze highly homopolymeric tail populations. In summary, there are multiple possibilities for adaptation of circTAIL-seq to answer outstanding questions about the roles of nontemplated tails on RNAs.

MATERIALS AND METHODS

Cell culture

Trypanosoma brucei 29–13 cell line was grown in CO_2 incubators at 27°C in SDM-79 supplemented with G418 and hygromycin and harvested in late log stage (1.5×10^7 cells/mL).

Generation of circTAIL amplicon libraries

Five micrograms of two DNase-treated RNA samples used in the qRT-PCR analysis were circularized in a 200 µL total reaction volume overnight at room temperature with 8 µL T4 RNA ligase (Epicentre) and 2 µL RNase inhibitor (Applied Biosystems) in 33 mM Tris acetate, 66 mM ammonium acetate, 10 mM magnesium acetate, 0.5 mM DTT, and a final ATP concentration of 25 µM. Samples were then extracted with phenol:chloroform, pH 5.2, and precipitated overnight. Pellets were washed and resuspended in 20 µL H_2O. One microgram of circular RNA was reverse transcribed in a 20 µL reaction using Episcript Reverse Transcriptase (Epicentre) with all gene-specific primers together (2 pmol each) that anneal to each RNA 3′ to the 5′ reverse PCR primer (provided in Supplemental Table S1). PCR was performed with KAPA2G Robust polymerase (Kapa Biosystems) with manufacturer-provided buffer and weighted dNTPs (8 mM dGTP, dCTP, 12 mM dTTP, dATP), using HPLC-purified primers adapted to generate Illumina-sequencable amplicons (Supplemental Table S1). Twenty microliters of PCR reactions were performed to optimize the PCR step of the protocol according to the Supplemental Protocol. One hundred microliters of PCR reactions were used to generate adequate product for sequencing. Optimized conditions for the two evaluated transcripts were as

follows. 0.25 µL cDNA per reaction for all three primer sets. After incubation at 95°C for 3:00, reactions were cycled; 33×: 95°C 0:15, 62°C 0:15, 72°C 0:15 (CO1); 32×: 95°C 0:15, 62°C 0:15, 72°C 0:15, (CO3p). Reactions were electrophoresed on a 1.5 mm × 20 cm long 6% polyacrylamide-TBE gel. Gel slabs containing products just under minimum expected size to minimum expected size plus 150 bp were excised. DNA was eluted from the gel slabs for each sample according to the protocol established elsewhere (Riley et al. 2014), using 400 rather than 200 µL elution volume and two spin columns per sample, ethanol precipitating, and resuspending each sample in 15 µL H_2O.

Sequencing

Generation of amplicons above utilized primers containing bar codes for each of six RNAs in biological replicate (12 bar-coded samples total). University of Minnesota Genomics Center performed quality control on all samples by examining quality and quantity on a Bioanalyzer, performed KapaQC to confirm its ability to be amplified, and sequenced the samples. Equal Bioanalyzer-determined quantities of six samples were multiplexed into a run on an Illumina MiSeq using the MiSeq V2 kit, acquiring 150 bp paired-end reads (two runs total). Runs were under clustered and spiked with a PhiX diverse library to improve *Q* scores at cycles where amplicon diversity is low.

Read processing

Raw reads (deposited in Sequence Read Archive SRP064265) were sorted by barcode and Illumina primer ends removed. Downstream read processing was performed on a Galaxy platform maintained by the Minnesota Supercomputing Institute. Variable sequences (4, 5, or 6 nt long) that are part of the PCR primers positioned between the Illumina primer sequence and gene-specific primer sequence were removed from both R1 and R2 reads using Trimmomatic HEADCROP task (Bolger et al. 2014), specifying a number of nucleotides to remove. R1 and R2 reads were then merged into single consensus reads using PEAR (Zhang et al. 2014; default settings). After conversion by FASTQ groomer, consensus reads were reverse complemented so reads would possess the proper directionality. This file was verified for per base sequence quality of 30 or better. The sequence was then subjected to a search for gene-specific primer annealing region of reverse amplification primer for the 5′ end at 80% similarity. Reads fulfilling this criterion were selected for the follow-up analysis.

We developed a software package written in C#, called circTAIL-seq Analyzer (available at http://trypsNetDB.org/circTAILseq Analyzer.zip), to systematically extract and analyze tail sequences from the preprocessed reads generated by circTAIL-seq deep sequencing. The circTAIL-seq Analyzer first aligns the reads to the reference sequence for the transcript of interest (including DNA sequence downstream and upstream of CDS) using Needleman–Wunsch pairwise global alignment algorithm (Needleman and Wunsch 1970). To have reliable identification of tail sequences, circTAIL-seq Analyzer only considers reads as valid that contain the well conserved 5′ and 3′ regions of the reference sequence, excluding the binding regions of the primers ("well conserved" reference sequence regions were defined as those regions where 90% or more reads aligned). The program permits a limited number of point mutations in the conserved region (max two mutations, different settings for each gene based on the observed diversity and the length of the conserved region) to account for diversity present in the population. The well-conserved regions and allowable number of point mutations can be selected by the program as default values or adjusted based on the users' needs. In the case of CO3p, primers were specifically designed to cover the initial editing region of the mtRNA immediately 5′ to the final editing site, thus permitting for selection of tails from RNAs we can demonstrate have not initiated editing as judged by alignment.

circTAIL-seq Analyzer next infers the embedded tails in the selected reads based on the alignment results. However, visual inspection of results demonstrated that small fractions of tails (<0.2% of tails in each sample) are contaminated with genomic/transcriptomic sequences (mostly rRNA) that can arise due to fragment incorporation during circularization. Therefore, the program filters out those tails that match (using NCBI BLAST, *e*-value <0.001) to a masked version of *T. brucei* reference genome in which interspersed repeats and low-complexity parts of the genome were masked out by dustMasker program (Morgulis et al. 2006). The program reports back primary alignment results, reads lacking the well-conserved 5′ and 3′ regions, tails contaminated with other genomic/transcriptomic sequences as judged by BLAST, the inferred tails for the reads that passed filtration criteria with inferred 3′ and 5′ termini, overall tail counts, tail length distribution, and nucleotide composition distributions.

Positional probabilities

Single position frequency plots shown in Figure 3A were produced with WebLogo (Crooks et al. 2004).

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank Aaron Becker and the staff at University of Minnesota Genomics Center for their assistance in troubleshooting early experiments. We also acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota (http://www.msi.umn.edu) for providing resources that contributed to the research results reported in this paper.

Received September 28, 2015; accepted December 9, 2015.

REFERENCES

Aphasizheva I, Aphasizhev R. 2010. RET1-catalyzed uridylylation shapes the mitochondrial transcriptome in *Trypanosoma brucei*. *Mol Cell Biol* **30:** 1555–1567.

Aphasizhev R, Aphasizheva I. 2011. Mitochondrial RNA processing in trypanosomes. *Res Microbiol* **162:** 655–663.

Aphasizhev R, Aphasizheva I. 2014. Mitochondrial RNA editing in trypanosomes: small RNAs in control. *Biochimie* **100:** 125–131.

Aphasizheva I, Maslov D, Wang X, Huang L, Aphasizhev R. 2011. Pentatricopeptide repeat proteins stimulate mRNA adenylation/uridylation to activate mitochondrial translation in trypanosomes. *Mol Cell* **42:** 106–117.

Beilharz TH, Preiss T. 2011. Polyadenylation state microarray (PASTA) analysis. *Methods Mol Biol* **759:** 133–148.
Bhat GJ, Souza AE, Feagin JE, Stuart K. 1992. Transcript-specific developmental regulation of polyadenylation in *Trypanosoma brucei* mitochondria. *Mol Biochem Parasitol* **52:** 231–240.
Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114–2120.
Chang JH, Tong L. 2012. Mitochondrial poly(A) polymerase and polyadenylation. *Biochim Biophys Acta* **1819:** 992–997.
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.
Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14:** 1188–1190.
Decker CJ, Sollner-Webb B. 1990. RNA editing involves indiscriminate U changes throughout precisely defined editing domains. *Cell* **61:** 1001–1011.
Diebel KW, Smith AL, van Dyk LF. 2010. Mature and functional viral miRNAs transcribed from novel RNA polymerase III promoters. *RNA* **16:** 170–185.
Etheridge RD, Aphasizheva I, Gershon PD, Aphasizhev R. 2008. 3′ adenylation determines mRNA abundance and monitors completion of RNA editing in *T. brucei mitochondria*. *EMBO J* **27:** 1596–1608.
Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, Lin C, Holoch D, Lim C, Tuschl T. 2008. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* **44:** 3–12.
Hashimi H, Zimmer SL, Ammerman ML, Read LK, Lukeš J. 2013. Dual core processing: MRB1 is an emerging kinetoplast RNA editing complex. *Trends Parasitol* **29:** 91–99.
Kao CY, Read LK. 2005. Opposing effects of polyadenylation on the stability of edited and unedited mitochondrial RNAs in *Trypanosoma brucei*. *Mol Cell Biol* **25:** 1634–1644.
Kao CY, Read LK. 2007. Targeted depletion of a mitochondrial nucleotidyltransferase suggests the presence of multiple enzymes that polymerize mRNA 3′ tails in *Trypanosoma brucei* mitochondria. *Mol Biochem Parasitol* **154:** 158–169.
Lee M, Kim B, Kim VN. 2014. Emerging roles of RNA modification: m^6A and U-tail. *Cell* **158:** 980–987.
Lim J, Ha M, Chang H, Kwon SC, Simanshu DK, Patel DJ, Kim VN. 2014. Uridylation by TUT4 and TUT7 marks mRNA for degradation. *Cell* **159:** 1365–1376.
Mattiacio JL, Read LK. 2008. Roles for TbDSS-1 in RNA surveillance and decay of maturation by-products from the 12S rRNA locus. *Nucleic Acids Res* **36:** 319–329.
Morgulis A, Gertz EM, Schäffer AA, Agarwala R. 2006. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol* **13:** 1028–1040.
Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48:** 443–453.
Norbury CJ. 2013. Cytoplasmic RNA: a case of the tail wagging the dog. *Nat Rev Mol Cell Biol* **14:** 643–653.
Perrin R, Lange H, Grienenberger JM, Gagliardi D. 2004a. AtmtPNPase is required for multiple aspects of the 18S rRNA metabolism in *Arabidopsis thaliana* mitochondria. *Nucleic Acids Res* **32:** 5174–5182.
Perrin R, Meyer EH, Zaepfel M, Kim YJ, Mache R, Grienenberger JM, Gualberto JM, Gagliardi D. 2004b. Two exoribonucleases act sequentially to process mature 3′-ends of atp9 mRNAs in *Arabidopsis* mitochondria. *J Biol Chem* **279:** 25440–25446.
Riley TR, Slattery M, Abe N, Rastogi C, Liu D, Mann RS, Bussemaker HJ. 2014. SELEX-seq: a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. *Methods Mol Biol* **1196:** 255–278.
Rorbach J, Minczuk M. 2012. The post-transcriptional life of mammalian mitochondrial RNA. *Biochem J* **444:** 357–373.
Ryan CM, Militello KT, Read LK. 2003. Polyadenylation regulates the stability of *Trypanosoma brucei* mitochondrial RNAs. *J Biol Chem* **278:** 32753–32762.
Schuster G, Stern D. 2009. RNA polyadenylation and decay in mitochondria and chloroplasts. *Prog Mol Biol Transl Sci* **85:** 393–422.
Slevin MK, Meaux S, Welch JD, Bigler R, Miliani de Marval PL, Su W, Rhoads RE, Prins JF, Marzluff WF. 2014. Deep sequencing shows multiple oligouridylations are required for 3′ to 5′ degradation of histone mRNAs on polyribosomes. *Mol Cell* **53:** 1020–1030.
Slomovic S, Schuster G. 2008. Stable PNPase RNAi silencing: its effect on the processing and adenylation of human mitochondrial RNA. *RNA* **14:** 310–323.
Slomovic S, Schuster G. 2013. Circularized RT-PCR (cRT-PCR): analysis of the 5′ ends, 3′ ends, and poly(A) tails of RNA. *Methods Enzymol* **530:** 227–251.
Stuart KD, Schnaufer A, Ernst NL, Panigrahi AK. 2005. Complex management: RNA editing in trypanosomes. *Trends Biochem Sci* **30:** 97–105.
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.
Temperley RJ, Seneca SH, Tonska K, Bartnik E, Bindoff LA, Lightowlers RN, Chrzanowska-Lightowlers ZM. 2003. Investigation of a pathogenic mtDNA microdeletion reveals a translation-dependent deadenylation decay pathway in human mitochondria. *Hum Mol Genet* **12:** 2341–2348.
Welch JD, Slevin MK, Tatomer DC, Duronio RJ, Prins JF, Marzluff WF. 2015. EnD-Seq and AppEnD: sequencing 3′ ends to identify nontemplated tails and degradation intermediates. *RNA* **21:** 1375–1389.
Wyman SK, Knouf EC, Parkin RK, Fritz BR, Lin DW, Dennis LM, Krouse MA, Webster PJ, Tewari M. 2011. Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res* **21:** 1450–1461.
Zhang X, Virtanen A, Kleiman FE. 2010. To polyadenylate or to deadenylate: that is the question. *Cell Cycle* **9:** 4437–4449.
Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30:** 614–620.
Zheng D, Tian B. 2014. Sizing up the poly(A) tail: insights from deep sequencing. *Trends Biochem Sci* **39:** 255–257.
Zimmer SL, Schein A, Zipor G, Stern DB, Schuster G. 2009. Polyadenylation in *Arabidopsis* and *Chlamydomonas* organelles: the input of nucleotidyltransferases, poly(A) polymerases and polynucleotide phosphorylase. *Plant J* **59:** 88–99.
Zimmer SL, McEvoy SM, Menon S, Read LK. 2012. Additive and transcript-specific effects of KPAP1 and TbRND activities on 3′ non-encoded tail characteristics and mRNA stability in *Trypanosoma brucei*. *PLoS One* **7:** e37639.

circTAIL-seq, a targeted method for deep analysis of RNA 3′ tails, reveals transcript-specific differences by multiple metrics

Vahid H. Gazestani, Marshall Hampton, Juan E. Abrahante, et al.

RNA 2016 22: 477-486 originally published online January 12, 2016
Access the most recent version at doi:10.1261/rna.054494.115