Molecular Cell
Resource

TAIL-seq: Genome-wide Determination of Poly(A) Tail Length and 3′ End Modifications

Hyeshik Chang,[1,2,3] Jaechul Lim,[1,2,3] Minju Ha,[1,2] and V. Narry Kim[1,2,*]
[1]Center for RNA Research, Institute for Basic Science, Seoul 151-742, Korea
[2]School of Biological Sciences, Seoul National University, Seoul 151-742, Korea
[3]These authors contributed equally to this work
*Correspondence: narrykim@snu.ac.kr
http://dx.doi.org/10.1016/j.molcel.2014.02.007

SUMMARY

Global investigation of the 3′ extremity of mRNA (3′-terminome), despite its importance in gene regulation, has not been feasible due to technical challenges associated with homopolymeric sequences and relative paucity of mRNA. We here develop a method, TAIL-seq, to sequence the very end of mRNA molecules. TAIL-seq allows us to measure poly(A) tail length at the genomic scale. Median poly(A) length is 50–100 nt in HeLa and NIH 3T3 cells. Poly(A) length correlates with mRNA half-life, but not with translational efficiency. Surprisingly, we discover widespread uridylation and guanylation at the downstream of poly(A) tail. The U tails are generally attached to short poly(A) tails (<25 nt), while the G tails are found mainly on longer poly(A) tails (>40 nt), implicating their generic roles in mRNA stability control. TAIL-seq is a potent tool to dissect dynamic control of mRNA turnover and translational control, and to discover unforeseen features of RNA cleavage and tailing.

INTRODUCTION

The 3′ termini of eukaryotic RNAs reflect their regulatory status and play important roles in determining the fates of RNAs. The 3′ ends are generated by endonucleolytic cleavage, tailing (untemplated nucleotidyl transfer), and/or exonucleolytic trimming. In the case of messenger RNA (mRNA), the nascent transcript is cleaved cotranscriptionally by cleavage and polyadenylation specificity factor (CPSF). Soon afterward, the 3′ end of mRNA becomes polyadenylated by canonical poly(A) polymerase (PAP), with an exception of replication-dependent histone mRNAs that lack poly(A) tails (Norbury, 2013). Poly(A) binding proteins (PABPs) not only protect poly(A) tails but also interact with eIF4G bound to the 5′ cap, which is generally thought to facilitate translational initiation (Weill et al., 2012).

Given their central importance, it is surprising that the actual sequences of the 3′ termini remain largely unknown. Current knowledge is limited to a few individual genes investigated by northern- and RT-PCR/Sanger sequencing-based techniques (Norbury, 2013; Sallés et al., 1999). Public sequencing data generated using standard mRNA-seq protocols have a strong bias against mRNA termini and contain little information of the 3′ end sequences of mRNA (Wang et al., 2009). Genome-scale investigation of RNA 3′ end has been difficult for several reasons. First, current deep sequencing technologies cannot determine homopolymeric sequences of longer than ~30 nt, so it is not feasible to read the long poly(A) tail. The most common method currently available is affinity chromatography on oligo(dT) or poly(U) beads followed by differential elution at different temperatures or salt concentrations (Beilharz and Preiss, 2007; Du and Richter, 2005; Meijer et al., 2007). The eluted RNAs are subsequently analyzed by microarray or sequencing. This approach suffers from low resolution (cannot measure small changes in poly[A] tail length) and cannot detect additional modifications at the 3′ end of mRNA. Second, because mRNAs account for only a small fraction of cellular RNA, highly abundant RNAs such as rRNAs and tRNAs dominate cDNA library unless mRNAs are enriched. Oligo(dT)-based affinity purification is often used to enrich mRNA, but it inevitably introduces bias toward mRNAs with long poly(A) tails. Thus, global investigation of RNA 3′ end has been largely limited to the mapping of 3′ UTR and polyadenylation sites that mark the boundary between mRNA body and poly(A) tail (Beck et al., 2010; Derti et al., 2012; Elkon et al., 2013; Fu et al., 2011; Hoque et al., 2013; Jan et al., 2011; Mangone et al., 2010; Martin et al., 2012; Ozsolak et al., 2010; Shepard et al., 2011; Wilkening et al., 2013; Yoon and Brem, 2010), while deep sequencing of the actual end of mRNA has not been feasible.

RESULTS

TAIL-seq: The Challenges and Solutions

To determine directly the 3′ end sequences of transcriptome, we developed a technique termed TAIL-seq (Figure 1A). To briefly list the main features of TAIL-seq, (1) abundant noncoding RNAs are removed by affinity-based depletion (of rRNA) and size fractionation (against tRNA, snRNA, snoRNA, and miRNA). Oligo(dT) is not used in any of the steps. (2) The 3′ adaptor is ligated prior to RNA fragmentation so as to capture the information at the 3′ extremity of RNA. (3) RNase T1 at a low concentration partially digests RNAs while preserving the poly(A) tails as the enzyme cleaves selectively after G residues. (4) The 3′ adaptor has biotin residues that allow purification of the

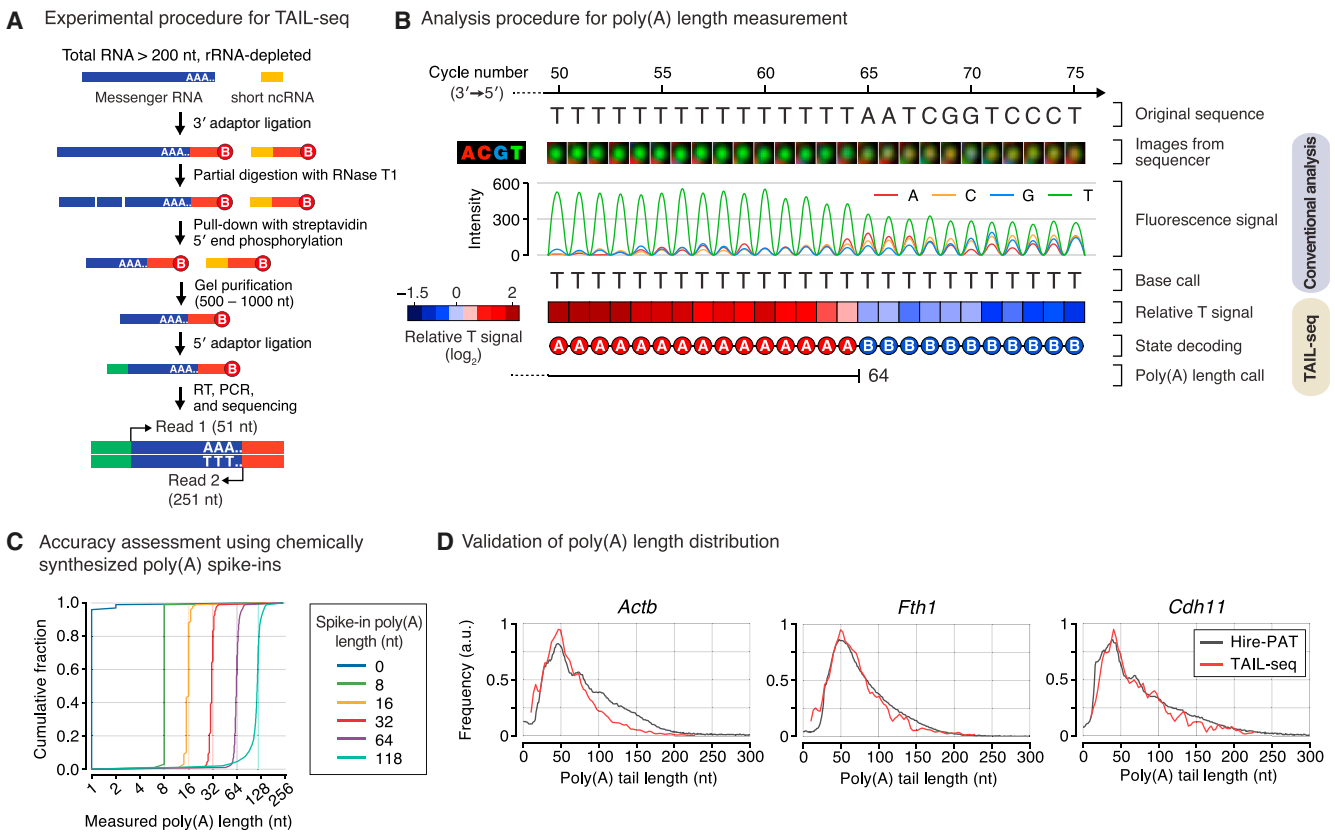

Figure 1. The Procedure and Assessment of TAIL-seq

(A) Schematic description of experimental procedure.

(B) An example of the analysis procedure for poly(A) length measurement. Shown is a spike-in (A_{64}) cluster from cycles corresponding to the 50th to 75th nucleotides from the 3′ end. "Images from sequencer" indicates serial pictures of a cluster taken in each sequencing cycle (red for C, green for T, blue for G; red also reflects A signal due to innate crosstalk between fluorophores). "Fluorescence signal" is the scaled signal intensity measured from the images. "Base call" shows the sequence determined by built-in software (Illumina RTA). "Relative T signal" indicates the T signal divided by the sum of other signals (A, C, and G; see Figure S1A for details), which was then used for machine learning to judge whether or not the cycle is from poly(A) region ("State decoding").

(C) Assessment of accuracy by sequencing synthetic poly(A) spike-ins. A cumulative curve of each spike-in demonstrates the distribution of poly(A) lengths measured by TAIL-seq. Theoretical size of each spike-in is indicated with light vertical lines.

(D) Poly(A) lengths of individual endogenous genes in NIH 3T3 cells measured by Hire-PAT method (Bazzini et al., 2012). Shown here is the representative result from two independent Hire-PAT experiments.

See also Figures S1–S3.

3′-most fragments. (5) The 3′ adaptor contains 15 degenerate bases that improve sequencing performance by diversifying reads from the initial cycles of read 2 and serve as a duplication filter to eliminate uneven PCR amplification artifacts. (6) We chose Illumina HiSeq as the sequencing platform because we needed to sequence highly complex populations of long RNAs containing homopolymers (see below). (7) Paired-end sequencing yields sequences of 51 nt from the 5′ end of the insert (read 1, used for genome mapping to identify the transcript) and 231 nt from the 3′ end (read 2, used for 3′ end sequence determination). (8) In addition to the standard base-calling software, we developed a complementary algorithm that is specialized in detecting signals from long T stretches (Figure 1B and Figure S1A available online, see below).

Although the Illumina sequencing chemistry handles homopolymers relatively well (Bragg et al., 2013), we noticed in our pilot experiments that the sequencing results are far from accurate in the long poly(T) region (which corresponds to poly[A] as sequenced in reverse orientation in read 2) (data not shown). This is due to limited handling of (pre-) phasing in homopolymeric stretches (Ledergerber and Dessimoz, 2011). Moreover, signals from T tend to accumulate over cycle due to incomplete cleavage of fluorophore from thymine (Whiteford et al., 2009) (Figure 1B, see "Fluorescence signal"). So, in read 2, non-poly(T) sequences following poly(T) are often indistinguishable from continuing T stretches, according to standard base-calling algorithm (Wilkening et al., 2013) (Figure 1B, see "Base call"). Thus, when we sequenced a synthetic oligonucleotide containing A_{64} flanked with adaptor sequences, it was often overestimated to have a longer tail (80–100 nt). Likewise, A_{118} was inaccurately measured as an ~155 nt-long A-stretch on average (Figure S1E). Quality score was not informative at all for this problem (data not shown), calling for a different approach.

While looking at the actual cluster images from the sequencer, we made an interesting observation: despite the phasing errors and persisting T signal, the signal intensity for T decreases while non-T signals increase once the cycle reaches non-T sequences (Figure 1B, see "Fluorescence signal"). This transition turned out to be very useful in determining the end of poly(A) stretch. Briefly, we used quantitative fluorescence signals (instead of the discrete values from base calls) to calculate "relative T signal" (T signal intensity divided by the sum of the other signal intensities) (Figures 1B and S1A). We then adopted Gaussian mixture hidden Markov model (GMHMM) to detect the position of transition from poly(A) tail (T stretches) to mRNA body (heterogeneous sequences) (Figures S1A–S1D; see Supplemental Experimental Procedures). The synthetic spike-ins were designed to harbor various lengths of poly(A) tails from 0 to 118 nt flanked by adaptor sequences. The signals from spike-ins (500 reads per each spike-in) were applied for unsupervised learning of GMHMM to model signal outputs from sequencing of poly(A) tails. The poly(A) tail lengths were estimated after decoding hidden states with the model using Viterbi algorithm (Figures 1B and S1A, see "State decoding" and "Poly[A] length call"). Note that A_{118} was used as the longest spike-in because we failed to synthesize longer oligonucleotides of sufficient quality.

This method turned out to be potent, and estimated poly(A) length with unprecedented accuracy and resolution (Figure 1C). Assuming that the spike-in oligonucleotides are synthesized without error, the error rate of poly(A) measurement is estimated to be 14.8% on average of root-mean-square error (RMSE), which is remarkably better than the approaches using standard base calls with or without allowing mismatches (Figures S1A and S1B; see Supplemental Experimental Procedures).

We generated TAIL-seq data from mouse fibroblast cell line NIH 3T3 and human cervical cancer cell line HeLa (29,610,077 and 21,794,337 reads, respectively, after filtering out PCR artifacts and rRNA reads). The tags originate mainly from the 3′ parts of genes, although we also find internal tags that reflect endonucleolytic and exonucleolytic activities (Figure S2A). We could measure the poly(A) length of 4,176 mouse and 4,091 human genes supported by $\geq$30 poly(A)$^+$ tags.

We compared our TAIL-seq data with previous results generated by differential elution from oligo(dT) column which separates mRNAs with short tails (<~30 nt) from those with long tails (Meijer et al., 2007) (>~30 nt) (Figure S2B). Despite the differences between two methods, the long/short tail ratio correlates significantly with our measurements ($p = 0.0024$, Pearson's correlation test; Figure S2B). To validate TAIL-seq data further, we carried out high-resolution poly(A) tail assay (Hire-PAT [Bazzini et al., 2012]) on five spike-ins and ten individual mRNAs. The lengths determined by Hire-PAT assay showed highly similar patterns to those from TAIL-seq for all spike-ins and mRNAs tested (Figures 1D, S2C, and S3A), including the two outliers from Figure S2B (*Hif1a* and *Cdh11*). Furthermore, northern blotting of RNase H cleavage products of *Spp1* mRNA showed a similar poly(A) length distribution to that determined by TAIL-seq, further validating our method (Figure S3B). TAIL-seq measures poly(A) tail length with an unprecedented resolution, accuracy, and scale.

Global Analysis of Poly(A) Tail

Figure 2A presents an example of randomly chosen tags that match to the 3′ end of the *Trp53* mRNA, which encodes the p53 protein. Translation of p53 was previously shown to be regulated by cytoplasmic polyadenylation (Burns and Richter, 2008), so the tail structure of p53 would be particularly interesting to analyze in the context of cellular senescence and malignancy. Read 1 is used to identify the gene, while read 2 is used to sequence the poly(A) tail of heterogeneous lengths. Various types of interesting information can be extracted from the TAIL-seq data, not only at the individual gene level but also at the transcriptome level.

We first examined the global distribution of poly(A) lengths. Overall, the distributions are similar between two cell lines examined (Figure 2B). When the mRNA tags with poly(A) tails of 8–231 nt are plotted, the median lengths are 60 nt and 59 nt in NIH 3T3 and HeLa, respectively. Poly(A) tails over 231 nt could not be counted further due to the limited sequencing cycle, but they account for only ~2% of the total population (see Supplemental Experimental Procedures). Poly(A) tails shorter than 8 nt were excluded from the analysis because the estimation was less accurate with such tags due to the ubiquity of short A stretches in the genome, particularly near polyadenylation sites. Accordingly, poly(A)-free RNAs such as histone mRNAs and decay intermediates were not included in this distribution analysis.

The tags derived from the same gene were clustered to calculate median poly(A) length for each individual gene (4,176 mouse and 4,091 human genes). The distribution of median poly(A) length was consistent over different abundance range of TAIL-seq tags (Figure S3E). As expected, we found that poly(A) lengths vary widely among different genes (mRNA species) (Figure 2C; Table S1). Some mRNA species carry poly(A) tails of ~20 nt, while others have long tails of ~100 nt. Based on these median poly(A) lengths for individual genes, transcriptome-wide median length (median of medians) is estimated to be 61 nt and 60 nt in NIH 3T3 and HeLa cells, respectively. Interestingly, these are significantly shorter than what is generally conceived as a typical length of mammalian poly(A) tail (150–200 nt).

We next asked whether genes with distinct biological function tend to differ in poly(A) length distribution, by gene ontology analysis (Figure 2D; Table S2). Genes associated with regulatory functions such as transcription factors, cell-cycle regulators, embryonic morphogenesis, and protein modification tend to have short tails. On the contrary, those with relatively constitutive functions such as ribosomal subunits, homeostatic genes, and metabolic genes hold relatively longer poly(A) tails, which is similar to ribosomal protein mRNAs in yeasts (Beilharz and Preiss, 2007; Lackner et al., 2007). This result suggests that the poly(A) tail of regulatory genes may be under dynamic control.

To understand which step of gene expression may be influenced by poly(A) tail, we first compared the median poly(A) length of each gene with mRNA half-life that was estimated previously by Schwanhäusser and colleagues (Schwanhäusser et al., 2011). Overall, there is a modest but significant correlation between poly(A) tail length and mRNA half-life ($p = 2.83 \times 10^{-5}$, Pearson's correlation test) (Figure 2E). Thus, deadenylation and/or cytoplasmic polyadenylation may affect mRNA stability, as previously shown (Dreyfus and Régnier, 2002; Norbury,

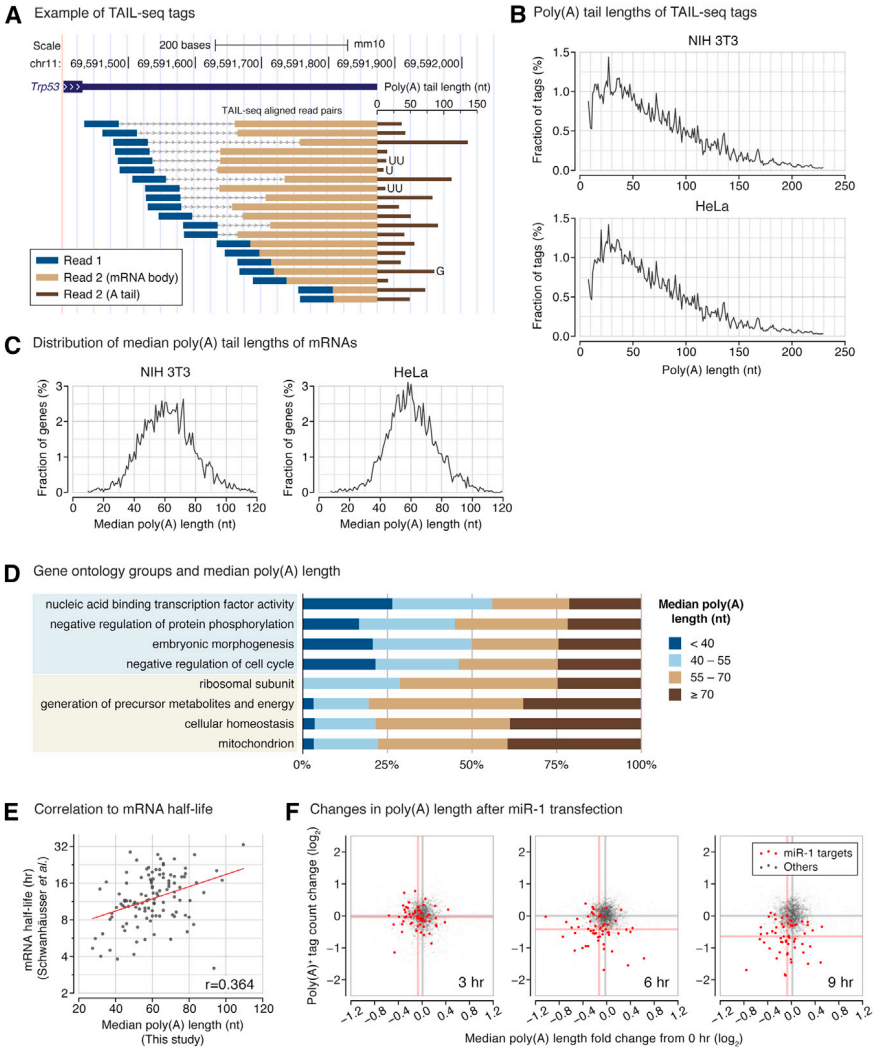

Figure 2. Analyses of Poly(A) Tail

(A) An example of TAIL-seq reads. Blue bar indicates genome-mapped read 1, while the following light brown bar indicates an inferred region of read 2 corresponding to mRNA body, with untemplated adenine residues shown as dark brown bar. Additional modifications are shown on the right.

(B) Global distribution of poly(A) tail lengths of TAIL-seq tags.

(C) Distribution of median poly(A) tail lengths of individual genes.

(D) Functional categorization of genes with their median poly(A) tail lengths. Four categories in the upper panel represent genes with relatively short poly(A) tails, while the lower four categories represent genes with longer tails. See Table S2 for the full list.

(E) A scatterplot showing the correlation between median poly(A) length and mRNA half-life, measured by Schwanhäusser et al. (Schwanhäusser et al., 2011). The r value refers to Pearson correlation coefficient, which is also applied to all the other scatter plots in this manuscript. mRNAs with more than 200 poly(A)$^+$ tags and with total length ranging from 3,000 to 5,000 nt were plotted due to the limited labeling of short RNAs in half-life measurement experiment.

(F) Scatterplots showing the changes of poly(A) tail lengths (x axis) and number of poly(A)$^+$ tags (y axis) after transfection of miR-1. Targets of miR-1 (red dots) are chosen from the list of mRNAs downregulated by more than 30% at 12 hr post-transfection in Guo et al. (2010). Gray dots represent the rest transcripts. Median changes are shown in vertical and horizontal lines. p values from two-sided Mann-Whitney U tests between targets and nontargets are as follows: (3 hr) poly(A) 5.84×10^{-4}, tag count 0.473; (6 hr) poly(A) 1.87×10^{-5}, tag count 1.56×10^{-14}; and (9 hr) poly(A) 6.63×10^{-4}, tag count 2.33×10^{-20}.

See also Figure S4, Table S1, and Table S2.

2013). Of note, poly(A) tail length does not correlate significantly with steady-state mRNA abundance, as was expected because other processes such as transcription also influence mRNA levels (Figure S4A).

It is established that microRNA (miRNA) induces deadenylation, but this model is based on the studies of a few individual genes (Behm-Ansmant et al., 2006; Giraldez et al., 2006; Huntzinger and Izaurralde, 2011; Wu et al., 2006). We tested the model by examining the global effect of miRNA on poly(A) tail of the targets (Figure 2F). Synthetic miR-1 mimic was transfected into HeLa cells, and the poly(A) length was measured by TAIL-seq. Deadenylation of miR-1 targets was evident as early as 3 hr posttransfection without a significant change in mRNA level (Figure 2F, red dots). After 6 or 9 hr, target mRNA level was substantially downregulated. Therefore, although there are some exceptions, our result confirms that, in general, miRNA indeed induces deadenylation. Furthermore, our kinetic global analysis confirms that deadenylation precedes mRNA decay.

We next compared poly(A) length with translation efficiency because it is generally considered that long poly(A) tail is required for effective translation (Kojima et al., 2012; Novoa et al., 2010; Piqué et al., 2008; Udagawa et al., 2012). Unexpectedly, however, poly(A) lengths do not show any meaningful correlation with protein synthesis rates (measured by metabolic labeling and mass spectrometry and divided by mRNA abundance) (Aviner et al., 2013; Schwanhäusser et al., 2011) (Figure S4B; p = 0.893 for NIH 3T3, p = 0.449 for HeLa, Pearson's correlation test). Similarly, when we compared poly(A) length with ribosome density that was determined by ribosomal footprinting (and divided by mRNA abundance) (Guo et al., 2010) (Figure S4C), there was no detectable correlation, further supporting our conclusion. One way of interpreting this result is that poly(A) tail is not a critical element for translation, at least in HeLa and NIH 3T3 cells, and that deadenylation/polyadenylation may not be coupled to translation. But it remains possible that poly(A) tail length at steady state may not faithfully reflect translatability. In future studies, kinetic analyses will be necessary to simultaneously determine the changes of translation and poly(A) tail (by ribosome footprinting and TAIL-seq, respectively) over time following perturbation in poly(A) tail. This study

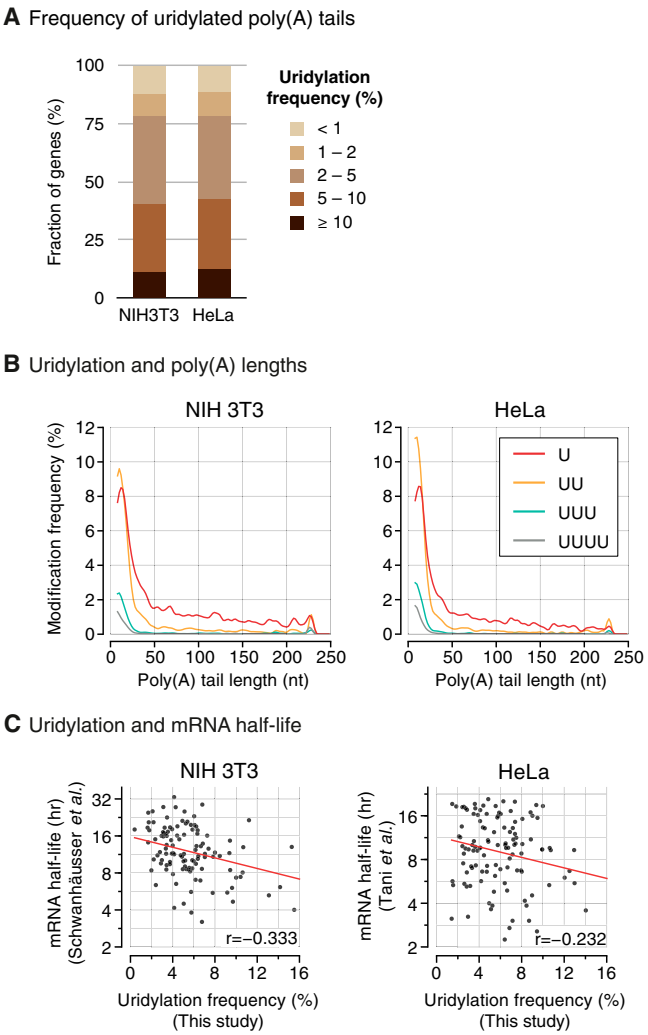

Figure 3. 3′ End Uridylation of Poly(A) Tail

(A) Uridylation frequency of mRNA.

(B) Relationship between uridylation and poly(A) tail length. The density was calculated with 2 nt wide bins, then smoothened with Hanning window (width = 7).

(C) Scatterplots showing the correlation between uridylation frequency and mRNA half-life (Schwanhäusser et al., 2011; Tani et al., 2012).

See also Figure S5.

will be of particular interest as regulation of poly(A) tail may play a determining role under specialized conditions such as in neural synapses and early embryos, where cytoplasmic polyadenylation is known to induce translation of dormant mRNAs with short tails (Besse and Ephrussi, 2008; D'Ambrogio et al., 2013; Mendez and Richter, 2001).

Widespread Uridylation of Mammalian mRNA

One of the unique strengths of TAIL-seq is its ability to determine the sequences of the very end of RNA and to examine if there is any other sequences apart from simple poly(A) stretches. While looking at the 3′ ends of mRNA reads, we were surprised to find widespread uridylation at the downstream of poly(A) tail (Figures 2A and 3A). About half of mRNA species carry U-tails at more than 5% frequency, and ~80% of mRNA species are uridylated at a frequency higher than 2% (Figure 3A). Some mRNAs such as encoding suppressor of glucose autophagy associated 2 (*SOGA2*) and encoding cytoplasmic poly(A) binding protein 4 (*PABPC4*) are frequently uridylated (41% and 24%, respectively), suggesting that at least some of uridylation may have biological importance. We observed a comparable pattern of uridylation in our pilot experiment where we used a different 3′ adaptor and a different RNA fragmentation method (alkaline hydrolysis) (Figure S5A). Uridylation was further validated by Sanger sequencing (Figures S5B and S5C).

Finding of uridylation at the end of poly(A) tail was unexpected because it was thought to occur only in fungi and plants. In mammals, there are only two known cases of uridylation in poly(A)-lacking mRNAs. Poly(A)-lacking histone mRNAs are oligo-uridylated and degraded at the end of S phase (Mullen and Marzluff, 2008; Schmidt et al., 2011). Another example is the 5′ fragment from small RNA-directed cleavage in mammals and plants, which is oligo-uridylated at the cleavage site (Shen and Goodman, 2004). In *S. pombe* and *A. thaliana*, some poly(A)$^+$ mRNAs bear short U tails (1–2 Us), as analyzed by circularized RT-PCR (Rissland et al., 2007; Sement et al., 2013). Uridyl residues were found mainly on decapped mRNAs which represent decay intermediates. When the uridylyl transferase (Cid1 in fission yeast) was mutated, mRNA was stabilized (Rissland and Norbury, 2009). These results collectively suggested that uridylation may be involved in mRNA decay. Our current observation demonstrates that uridylation is much more pervasive than previously anticipated. mRNA uridylation may be an integral part of a generic mRNA turnover pathway that is conserved in all eukaryotes.

It is particularly interesting that uridyl residues are found mainly in mRNAs with short poly(A) tails (<~25 nt) (Figure 3B). This phenomenon is similar to that in *Arabidopsis* where short U tag (1–2 nt) is added to 10–20 nt poly(A) (Sement et al., 2013). It was proposed that uridylation protects the 3′ end against further deadenylation and promotes decapping and 5′-3′ decay (Sement et al., 2013). In filamentous fungus *Aspergillus nidulans*, a mixture of uridyl and cytidyl residues are added to short poly(A) tails (~15 nt) (Morozov et al., 2012).

Consistent with the notion that uridylation may be involved in RNA decay, uridylation frequency shows a modest negative correlation with mRNA half-life in both HeLa and NIH 3T3 cells (Figure 3C), but not with mRNA abundance or translation rate (Figures S5D and S5E). This is intriguing in light of recent reports showing that an oligo-U tail serves as a decay marker by interacting with a 3′-5′ exonuclease Dis3L2 (Chang et al., 2013; Lubas et al., 2013; Malecki et al., 2013) and by recruiting LSM1-7 complex and decapping enzymes (Mullen and Marzluff, 2008; Rissland and Norbury, 2009). In future studies, RNAi of uridylyl transferases and nucleases can be combined with TAIL-seq, so as to elucidate the functional consequence and mechanism of uridylation and decay.

The G Tail

In addition to uridylation, we discover yet another type of modification: guanylation (Figure 4A). About 20% of mRNA species are guanlylated at the downstream of poly(A) tail at a frequency

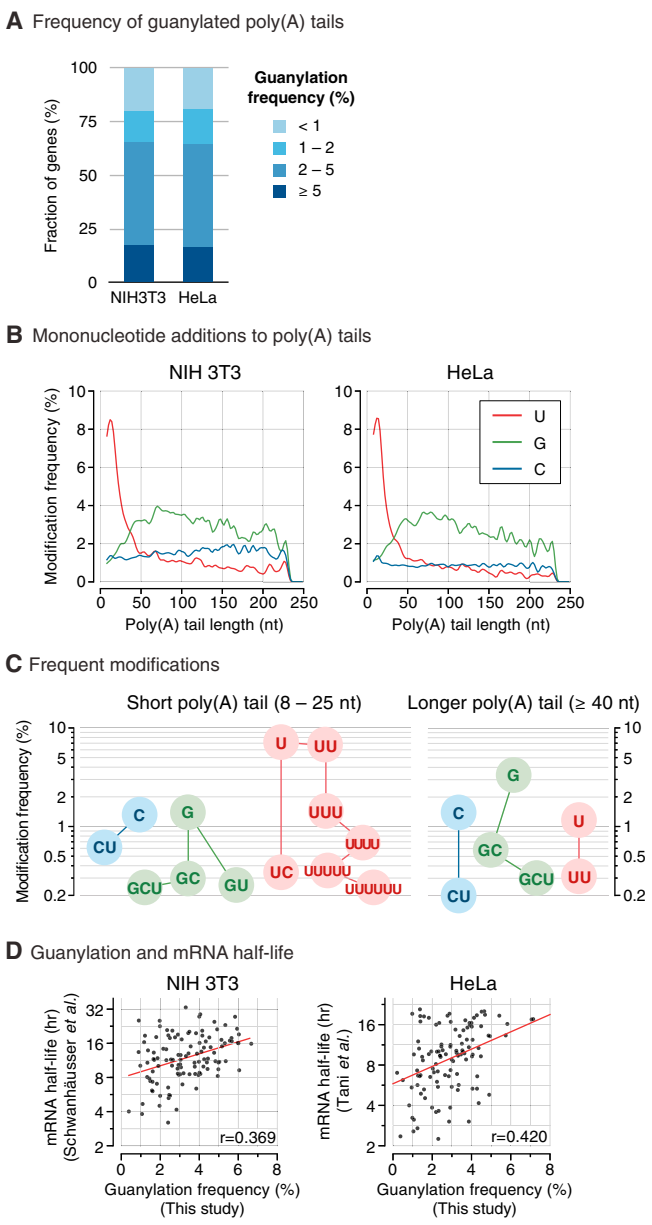

Figure 4. 3′ End Guanylation of Poly(A) Tail

(A) Guanylation frequency of mRNA.

(B) Relationship between guanylation and poly(A) tail length. The density is presented as in Figure 3B.

(C) Additional nucleotides attached to either short poly(A) tails (left panel) or longer poly(A) tails (right panel).

(D) Scatterplots showing the correlation between guanylation frequency and mRNA half-life (Schwanhäusser et al., 2011; Tani et al., 2012).

See also Figure S5.

of higher than 5%; and over 60% of transcripts show G-addition at more than 2% frequency (Figure 4A). Guanylation was detected in our initial experiments using a different 3′ adaptor and alkaline hydrolysis (instead of RNase T1) (Figure S5A), and was confirmed by small-scale Sanger sequencing (Figures S5B and S5C), indicating that the modification is not an artifact of our protocol. To our knowledge, this is the first description of RNA 3′ guanylation, although it was shown previously that some noncanonical PAPs can utilize GTP in vitro (Bai et al., 2011; Heo et al., 2012). In contrast to U tails, terminal G residues are found selectively on longer poly(A) tails (>40 nt) (Figures 4B and 4C). Cytidylation is considerably less frequent and does not show any preference for poly(A) tail size (Figures 4B). Because deadenylases PARN and CCR4 are known to have a preference for terminal diadenosines (Henriksson et al., 2010; Viswanathan et al., 2003) (AA), one can envision that the G addition may slow down deadenylation to protect mRNAs with long poly(A) tail. We indeed detect a modest positive correlation between guanylation frequency and mRNA half-life (Figure 4D), but no correlation between guanylation and mRNA level or translation rate (Figures S5F and S5G). Although it would be too early to draw a conclusion, it is tempting to speculate that guanylation may stabilize mRNAs by antagonizing deadenylation. Not mutually exclusively, it is also plausible that G-tailed mRNAs may represent a specific subcellular location and/or a phase of mRNA life cycle.

Detection of Polyadenylation Sites

Using TAIL-seq data, we could map the poly(A) sites, although this was not our primary goal, and the depth was lower compared to the other specialized tools developed previously (Beck et al., 2010; Derti et al., 2012; Fu et al., 2011; Hoque et al., 2013; Jan et al., 2011; Mangone et al., 2010; Martin et al., 2012; Ozsolak et al., 2010; Shepard et al., 2011; Wilkening et al., 2013; Yoon and Brem, 2010). Nevertheless, when compared with the annotated poly(A) sites in RefSeq, the sites detected from our sequencing are significantly enriched at the annotated sites (Figure 5A). Of note, the 3′ ends detected by TAIL-seq fall predominantly at the upstream of the annotated sites rather than the downstream. The upstream sites may correspond to alternative polyadenylation sites, considering that RefSeq often annotates the most distal sites. The sequences surrounding the detected poly(A) sites show characteristic features of known poly(A) sites (Figure 5B), including the polyadenylation signal (PAS, AAUAAA, and its variants), the U-rich upstream sequence element (USE), and the downstream sequence element (DSE), indicating that TAIL-seq detects poly(A) sites accurately. We could also detect alternative polyadenylation (APA) in some genes (Figure 5C). Notably, certain isoforms differ significantly in their poly(A) length and modification frequency, which is consistent with the notion that APA fundamentally influences mRNA fates (Elkon et al., 2013). For instance, we detected two alternatively processed isoforms from *Bclaf1* gene: one with short 3′ UTR carries a long poly(A) tail and relatively frequent G tags, while another isoform has extended 3′ UTR, a shorter poly(A) tail, and frequent U tails.

DISCUSSION

TAIL-seq is a method that allows global survey of 3′-terminome. Our analysis indicates that poly(A) tails in mammalian cells are substantially shorter (50–100 nt) than generally conceived (150–200 nt). A newly transcribed transcript is known to receive a poly(A) tail of ~230 nt, and they are gradually shortened

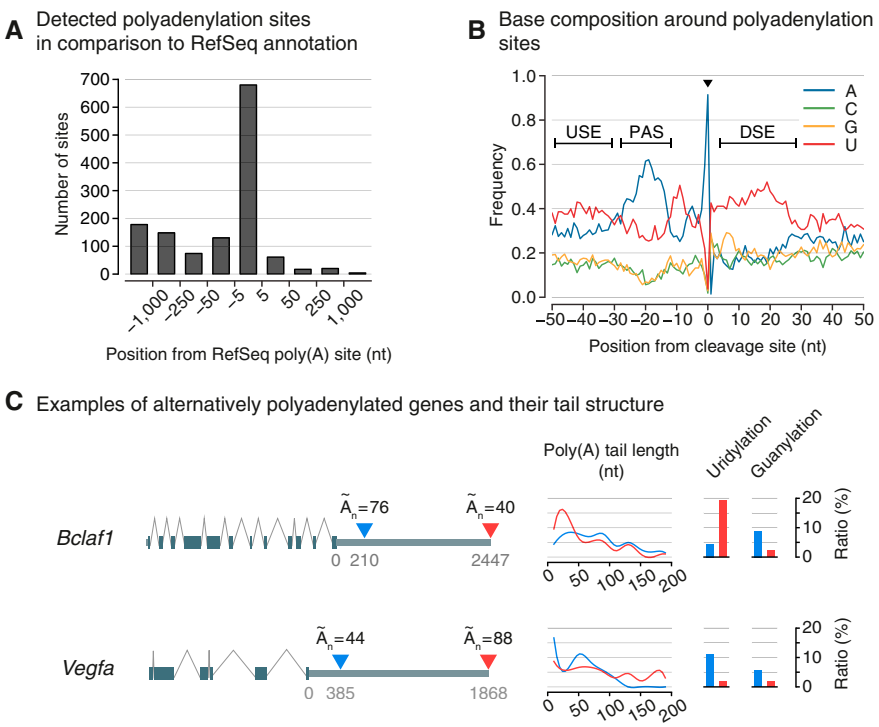

Figure 5. Detection of Polyadenylation Sites by TAIL-seq

(A) Position of the poly(A) site identified by TAIL-seq, against the RefSeq annotation.

(B) Nucleotide composition of genomic sequences near the detected poly(A) sites. Sequence motifs such as PAS (polyadenylation signal), USE (upstream sequence element), and DSE (downstream sequence element) are enriched as shown previously (Derti et al., 2012; Hoque et al., 2013; Jan et al., 2011; Mangone et al., 2010; Martin et al., 2012; Ozsolak et al., 2010).

(C) Simultaneous detection of alternative poly(A) sites and their tail structures. $\tilde{A}_n$ refers to the median length of poly(A) tail. Poly(A) tail length distributions are counted in 20 nt-wide bins, then shown after cubic spline interpolation.

by deadenylases PARN, the PAN2-PAN3 complex, and the CCR4-NOT complex (Garneau et al., 2007). There have been discrepancies over the poly(A) length in earlier studies of bulk poly(A)$^+$ RNA or individual RNAs, which described poly(A) size as ~170 nt in mouse sarcoma polysomes, 100–160 nt in HeLa, and 50–70 nt in rabbit reticulocyte polysomes (Brawerman, 1974). A recent study using oligo(dT) chromatography and microarray suggested that many mammalian mRNAs may have tails of shorter than 30 nt (Meijer et al., 2007). It is noted that we do not rule out a possibility that the current TAIL-seq protocol may underestimate poly(A) length to a certain extent, because the method is based on PCR amplification which generally disfavors homopolymeric sequences. However, several lines of evidence support the notion that mammalian poly(A) tails may indeed be shorter than previous estimations. First, we validated our results with Hire-PAT and northern blotting. Second, TAIL-seq directly sequences the tail so it is free of crosshybridization and low resolution issues. Third, previous methods may have overestimated tail length because they could not detect very short A tails, unlike TAIL-seq.

As the current version of TAIL-seq was designed to look at the 3′-terminome as comprehensively as possible, it allows us to discover many exciting features. Some mRNAs carry unusually short or long poly(A) tails. For instance, *NFKBIA*, *SUZ12*, *PABPC1*, and *EXOSC7* mRNAs have short poly(A) tails (<~40 nt), suggesting that they may rapidly turn over. Furthermore, we made intriguing observations on RNA modifications such as pervasive uridylation and guanylation. The "RNA tailing" may have a fundamental impact on RNA fate determination. Our study raises numerous open questions: which protein factors are involved in each processing and modification, and what are the physiological consequences of the modifications? To this end, TAIL-seq, combined with systematic RNAi, will serve as a potent tool.

TAIL-seq will also be useful to solve various general issues concerning mRNA deadenylation, translation, and decay. It will be particularly interesting to compare the kinetics of deadenylation and translation during miRNA-induced gene silencing, by carrying out TAIL-seq and ribosome footprinting after miRNA transfection. Recent studies indicated that translational suppression may precede deadenylation (Bazzini et al., 2012; Béthune et al., 2012). But because the analyses were done with a small number of target genes and because a modest level of deadenylation was detected at the time of translational suppression (Bazzini et al., 2012; Béthune et al., 2012), it will be necessary to measure the kinetics of deadenylation and translational suppression, at a higher resolution and at the transcriptome level.

Although our analysis of steady-state level of poly(A) tail showed no correlation between poly(A) length and translation rates in nonsynchronous culture of HeLa and NIH 3T3, it is well known that cytoplasmic polyadenylation plays important roles in physiological transitional conditions such as oocyte activation (Mendez and Richter, 2001), cell-cycle progression (Novoa et al., 2010), circadian rhythm (Kojima et al., 2012), neural synapse function (Udagawa et al., 2012), cellular senescence (Burns and Richter, 2008; Groisman et al., 2006), and inflammation (Weill et al., 2012). Cytoplasmic polyadenylation machinery has also been implicated in tumorigenesis (D'Ambrogio et al., 2013). By analyzing such transitional conditions by TAIL-seq and ribosome footprinting, one may identify genes controlled by polyadenylation, and dissect the molecular mechanism of polyadenylation. Thus, TAIL-seq may offer a technical breakthrough in our understanding of cytoplasmic polyadenylation.

Apart from the features described in this paper, TAIL-seq data sets contain rich information that remains to be analyzed. For example, TAIL-seq identifies the 3′ ends of histone mRNAs, DROSHA cleavage sites, numerous putative cleavage sites, and various types of noncoding RNAs, which will be interesting subjects to investigate. For certain types of 3′ ends that are

relatively low in abundance, the technology will need to be modified further to generate more focused libraries. Focused libraries for a subset of RNA termini will increase the depth and reduce the cost of analysis. TAIL-seq is indeed a highly amenable technology that can be modified easily. For instance, one can change the range of size fractionation and/or use RNA extracted from subcellular fractions and immunoprecipitates to enrich for a selective class of RNA. The TAIL-seq protocol can be applied to any species and cell types with minor modifications, which will greatly expand the initial observations made in this study.

EXPERIMENTAL PROCEDURES

NIH 3T3 or HeLa total RNAs were extracted and size fractionated (>200 nt). The rRNAs were depleted by using Epicentre Ribo-Zero kit. The RNAs were ligated to biotinylated 3′ adaptor, partially digested by RNase T1. The fragmented RNAs were pulled down, phosphorylated, and gel purified (500–1000 nt). The purified RNAs were ligated to 5′ adaptor, reverse transcribed, and amplified by PCR. The PCR products were purified again and sequenced on Illumina HiSeq 2500 (51 × 251 bp paired end run) with PhiX control library and the spike-in mixture. The quantified fluorescence signals were transformed to "relative T signal," which is basically the log ratio between T signal and a sum of the others. The transformed signals from spike-ins were used to train a GMHMM to detect poly(A) to mRNA body transitions using Baum-Welch algorithm. Read 2 (3′ end of insert) signals of tags from NIH 3T3 or HeLa were decoded with Viterbi algorithm for the model, then the spans of states 1 and 2 were determined as poly(A) tail length. Read 1 (5′ end of insert) of the tags were aligned against UCSC mm10 or hg19 genome with GSNAP. The tags were classified referring RefSeq, RepeatMasker, miRBase, rfam, and gtRNAdb annotations. Full details are provided in the Supplemental Experimental Procedures.

ACCESSION NUMBERS

Sequenced reads have been deposited in the NCBI Gene Expression Omnibus (GEO) database (accession numbers GSE51299 and GSE54114).

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures, two tables, and Supplemental Experimental Procedures and can be found with this article at http://dx.doi.org/10.1016/j.molcel.2014.02.007.

ACKNOWLEDGMENTS

We are grateful to the members of our laboratory for discussion and technical help. We thank Drs. Jeong-Sun Seo, Gap-Seok Yang, and Sookjin Lee at Macrogen Inc. for the technical help with sequencing. This work was supported by the Research Center Program (EM1302) of IBS (Institute for Basic Science) from the Ministry of Science, ICT and Future Planning of Korea (H.C., J.L., M.H., and V.N.K.), and by the BK21 Research Fellowships from the Ministry of Education of Korea (H.C. and J.L.).

Received: November 19, 2013
Revised: December 23, 2013
Accepted: February 3, 2014
Published: February 27, 2014

REFERENCES

Aviner, R., Geiger, T., and Elroy-Stein, O. (2013). Novel proteomic approach (PUNCH-P) reveals cell cycle-specific fluctuations in mRNA translation. Genes Dev. *27*, 1834–1844.

Bai, Y., Srivastava, S.K., Chang, J.H., Manley, J.L., and Tong, L. (2011). Structural basis for dimerization and activity of human PAPD1, a noncanonical poly(A) polymerase. Mol. Cell *41*, 311–320.

Bazzini, A.A., Lee, M.T., and Giraldez, A.J. (2012). Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science *336*, 233–237.

Beck, A.H., Weng, Z., Witten, D.M., Zhu, S., Foley, J.W., Lacroute, P., Smith, C.L., Tibshirani, R., van de Rijn, M., Sidow, A., and West, R.B. (2010). 3′-end sequencing for expression quantification (3SEQ) from archival tumor samples. PLoS ONE *5*, e8768.

Behm-Ansmant, I., Rehwinkel, J., Doerks, T., Stark, A., Bork, P., and Izaurralde, E. (2006). mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. Genes Dev. *20*, 1885–1898.

Beilharz, T.H., and Preiss, T. (2007). Widespread use of poly(A) tail length control to accentuate expression of the yeast transcriptome. RNA *13*, 982–997.

Besse, F., and Ephrussi, A. (2008). Translational control of localized mRNAs: restricting protein synthesis in space and time. Nat. Rev. Mol. Cell Biol. *9*, 971–980.

Béthune, J., Artus-Revel, C.G., and Filipowicz, W. (2012). Kinetic analysis reveals successive steps leading to miRNA-mediated silencing in mammalian cells. EMBO Rep. *13*, 716–723.

Bragg, L.M., Stone, G., Butler, M.K., Hugenholtz, P., and Tyson, G.W. (2013). Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput. Biol. *9*, e1003031.

Brawerman, G. (1974). Eukaryotic messenger RNA. Annu. Rev. Biochem. *43*, 621–642.

Burns, D.M., and Richter, J.D. (2008). CPEB regulation of human cellular senescence, energy metabolism, and p53 mRNA translation. Genes Dev. *22*, 3449–3460.

Chang, H.M., Triboulet, R., Thornton, J.E., and Gregory, R.I. (2013). A role for the Perlman syndrome exonuclease Dis3l2 in the Lin28-let-7 pathway. Nature *497*, 244–248.

D'Ambrogio, A., Nagaoka, K., and Richter, J.D. (2013). Translational control of cell growth and malignancy by the CPEBs. Nat. Rev. Cancer *13*, 283–290.

Derti, A., Garrett-Engele, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M., and Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. Genome Res. *22*, 1173–1183.

Dreyfus, M., and Régnier, P. (2002). The poly(A) tail of mRNAs: bodyguard in eukaryotes, scavenger in bacteria. Cell *111*, 611–613.

Du, L., and Richter, J.D. (2005). Activity-dependent polyadenylation in neurons. RNA *11*, 1340–1347.

Elkon, R., Ugalde, A.P., and Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. Nat. Rev. Genet. *14*, 496–506.

Fu, Y., Sun, Y., Li, Y., Li, J., Rao, X., Chen, C., and Xu, A. (2011). Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. Genome Res. *21*, 741–747.

Garneau, N.L., Wilusz, J., and Wilusz, C.J. (2007). The highways and byways of mRNA decay. Nat. Rev. Mol. Cell Biol. *8*, 113–126.

Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., van Dongen, S., Inoue, K., Enright, A.J., and Schier, A.F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science *312*, 75–79.

Groisman, I., Ivshina, M., Marin, V., Kennedy, N.J., Davis, R.J., and Richter, J.D. (2006). Control of cellular senescence by CPEB. Genes Dev. *20*, 2701–2712.

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835–840.

Henriksson, N., Nilsson, P., Wu, M., Song, H., and Virtanen, A. (2010). Recognition of adenosine residues by the active site of poly(A)-specific ribonuclease. J. Biol. Chem. *285*, 163–170.

Heo, I., Ha, M., Lim, J., Yoon, M.J., Park, J.E., Kwon, S.C., Chang, H., and Kim, V.N. (2012). Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs. Cell *151*, 521–532.

Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J.Y., Yehia, G., and Tian, B. (2013). Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. Nat. Methods *10*, 133–139.

Huntzinger, E., and Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat. Rev. Genet. *12*, 99–110.

Jan, C.H., Friedman, R.C., Ruby, J.G., and Bartel, D.P. (2011). Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. Nature *469*, 97–101.

Kojima, S., Sher-Chen, E.L., and Green, C.B. (2012). Circadian control of mRNA polyadenylation dynamics regulates rhythmic protein expression. Genes Dev. *26*, 2724–2736.

Lackner, D.H., Beilharz, T.H., Marguerat, S., Mata, J., Watt, S., Schubert, F., Preiss, T., and Bähler, J. (2007). A network of multiple regulatory layers shapes gene expression in fission yeast. Mol. Cell. *26*, 145–155.

Ledergerber, C., and Dessimoz, C. (2011). Base-calling for next-generation sequencing platforms. Brief. Bioinform. *12*, 489–497.

Lubas, M., Damgaard, C.K., Tomecki, R., Cysewski, D., Jensen, T.H., and Dziembowski, A. (2013). Exonuclease hDIS3L2 specifies an exosome-independent 3′-5′ degradation pathway of human cytoplasmic mRNA. EMBO J. *32*, 1855–1868.

Malecki, M., Viegas, S.C., Carneiro, T., Golik, P., Dressaire, C., Ferreira, M.G., and Arraiano, C.M. (2013). The exoribonuclease Dis3L2 defines a novel eukaryotic RNA degradation pathway. EMBO J. *32*, 1842–1854.

Mangone, M., Manoharan, A.P., Thierry-Mieg, D., Thierry-Mieg, J., Han, T., Mackowiak, S.D., Mis, E., Zegar, C., Gutwein, M.R., Khivansara, V., et al. (2010). The landscape of C. elegans 3'UTRs. Science *329*, 432–435.

Martin, G., Gruber, A.R., Keller, W., and Zavolan, M. (2012). Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′UTR length. Cell Rep. *1*, 753–763.

Meijer, H.A., Bushell, M., Hill, K., Gant, T.W., Willis, A.E., Jones, P., and de Moor, C.H. (2007). A novel method for poly(A) fractionation reveals a large population of mRNAs with a short poly(A) tail in mammalian cells. Nucleic Acids Res. *35*, e132.

Mendez, R., and Richter, J.D. (2001). Translational control by CPEB: a means to the end. Nat. Rev. Mol. Cell Biol. *2*, 521–529.

Morozov, I.Y., Jones, M.G., Gould, P.D., Crome, V., Wilson, J.B., Hall, A.J., Rigden, D.J., and Caddick, M.X. (2012). mRNA 3′ tagging is induced by nonsense-mediated decay and promotes ribosome dissociation. Mol. Cell. Biol. *32*, 2585–2595.

Mullen, T.E., and Marzluff, W.F. (2008). Degradation of histone mRNA requires oligouridylation followed by decapping and simultaneous degradation of the mRNA both 5′ to 3′ and 3′ to 5′. Genes Dev. *22*, 50–65.

Norbury, C.J. (2013). Cytoplasmic RNA: a case of the tail wagging the dog. Nat. Rev. Mol. Cell Biol. *14*, 643–653.

Novoa, I., Gallego, J., Ferreira, P.G., and Mendez, R. (2010). Mitotic cell-cycle progression is regulated by CPEB1 and CPEB4-dependent translational control. Nat. Cell Biol. *12*, 447–456.

Ozsolak, F., Kapranov, P., Foissac, S., Kim, S.W., Fishilevich, E., Monaghan, A.P., John, B., and Milos, P.M. (2010). Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. Cell *143*, 1018–1029.

Piqué, M., López, J.M., Foissac, S., Guigó, R., and Méndez, R. (2008). A combinatorial code for CPE-mediated translational control. Cell *132*, 434–448.

Rissland, O.S., and Norbury, C.J. (2009). Decapping is preceded by 3′ uridylation in a novel pathway of bulk mRNA turnover. Nat. Struct. Mol. Biol. *16*, 616–623.

Rissland, O.S., Mikulasova, A., and Norbury, C.J. (2007). Efficient RNA polyuridylation by noncanonical poly(A) polymerases. Mol. Cell. Biol. *27*, 3612–3624.

Sallés, F.J., Richards, W.G., and Strickland, S. (1999). Assaying the polyadenylation state of mRNAs. Methods *17*, 38–45.

Schmidt, M.J., West, S., and Norbury, C.J. (2011). The human cytoplasmic RNA terminal U-transferase ZCCHC11 targets histone mRNAs for degradation. RNA *17*, 39–44.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Sement, F.M., Ferrier, E., Zuber, H., Merret, R., Alioua, M., Deragon, J.M., Bousquet-Antonelli, C., Lange, H., and Gagliardi, D. (2013). Uridylation prevents 3′ trimming of oligoadenylated mRNAs. Nucleic Acids Res. *41*, 7115–7127.

Shen, B., and Goodman, H.M. (2004). Uridine addition after microRNA-directed cleavage. Science *306*, 997.

Shepard, P.J., Choi, E.A., Lu, J., Flanagan, L.A., Hertel, K.J., and Shi, Y. (2011). Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. RNA *17*, 761–772.

Tani, H., Mizutani, R., Salam, K.A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y., and Akimitsu, N. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res. *22*, 947–956.

Udagawa, T., Swanger, S.A., Takeuchi, K., Kim, J.H., Nalavadi, V., Shin, J., Lorenz, L.J., Zukin, R.S., Bassell, G.J., and Richter, J.D. (2012). Bidirectional control of mRNA translation and synaptic plasticity by the cytoplasmic polyadenylation complex. Mol. Cell *47*, 253–266.

Viswanathan, P., Chen, J., Chiang, Y.C., and Denis, C.L. (2003). Identification of multiple RNA features that influence CCR4 deadenylation activity. J. Biol. Chem. *278*, 14949–14955.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Weill, L., Belloc, E., Bava, F.A., and Méndez, R. (2012). Translational control by changes in poly(A) tail length: recycling mRNAs. Nat. Struct. Mol. Biol. *19*, 577–585.

Whiteford, N., Skelly, T., Curtis, C., Ritchie, M.E., Löhr, A., Zaranek, A.W., Abnizova, I., and Brown, C. (2009). Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics *25*, 2194–2199.

Wilkening, S., Pelechano, V., Järvelin, A.I., Tekkedil, M.M., Anders, S., Benes, V., and Steinmetz, L.M. (2013). An efficient method for genome-wide polyadenylation site mapping and RNA quantification. Nucleic Acids Res. *41*, e65.

Wu, L., Fan, J., and Belasco, J.G. (2006). MicroRNAs direct rapid deadenylation of mRNA. Proc. Natl Acad. Sci. USA *103*, 4034–4039.

Yoon, O.K., and Brem, R.B. (2010). Noncanonical transcript forms in yeast and their regulation during environmental stress. RNA *16*, 1256–1267.