3′READS+, a sensitive and accurate method for 3′ end sequencing of polyadenylated RNA

DINGHAI ZHENG, XIAOCHUAN LIU, and BIN TIAN

Department of Microbiology, Biochemistry and Molecular Genetics, Rutgers New Jersey Medical School, Newark, New Jersey 07103, USA

ABSTRACT

Sequencing of the 3′ end of poly(A)$^+$ RNA identifies cleavage and polyadenylation sites (pAs) and measures transcript expression. We previously developed a method, 3′ region extraction and deep sequencing (3′READS), to address mispriming issues that often plague 3′ end sequencing. Here we report a new version, named 3′READS+, which has vastly improved accuracy and sensitivity. Using a special locked nucleic acid oligo to capture poly(A)$^+$ RNA and to remove the bulk of the poly(A) tail, 3′READS+ generates RNA fragments with an optimal number of terminal A's that balance data quality and detection of genuine pAs. With improved RNA ligation steps for efficiency, the method shows much higher sensitivity (over two orders of magnitude) compared to the previous version. Using 3′READS+, we have uncovered a sizable fraction of previously overlooked pAs located next to or within a stretch of adenylate residues in human genes and more accurately assessed the frequency of alternative cleavage and polyadenylation (APA) in HeLa cells (~50%). 3′READS+ will be a useful tool to accurately study APA and to analyze gene expression by 3′ end counting, especially when the amount of input total RNA is limited.

Keywords: alternative polyadenylation; deep sequencing; gene expression; 3′ end counting

INTRODUCTION

Studies in recent years have revealed that most mRNA genes in eukaryotes contain multiple cleavage and polyadenylation sites, or pAs, resulting in alternative cleavage and polyadenylation (APA) isoforms with different coding sequences (CDS) and/or variable 3′ untranslated regions (3′UTRs) (Ozsolak et al. 2010; Jan et al. 2011; Shepard et al. 2011; Wu et al. 2011; Derti et al. 2012; Hoque et al. 2013). Dynamic APA regulation has been reported in different tissue types (Zhang et al. 2005; Wang et al. 2008; Lianoglou et al. 2013), cell proliferation/differentiation and development (Sandberg et al. 2008; Ji et al. 2009; Shepard et al. 2011), cancer cell transformation (Mayr and Bartel 2009; Singh et al. 2009), and response to extracellular stimuli (Flavell et al. 2008). In addition, a sizable fraction of long noncoding RNA genes also display APA (Hoque et al. 2013), whose consequences are yet to be fully appreciated.

While APA can be analyzed with data from microarrays (Sandberg et al. 2008; Ji and Tian 2009), serial analysis of gene expression (SAGE) (Ji et al. 2009) or RNA-seq (Katz et al. 2010; Xia et al. 2014), these techniques were not specifically designed to identify pAs, leading to incomplete analysis. These methods are particularly ineffective when pAs of different isoforms are located close to one another. However, isoforms using different pAs within a short window have been shown to have quite different metabolisms (Geisberg et al. 2014), making it necessary to examine APA isoforms with precise tools. A number of deep sequencing methods have been developed to specifically sequence the 3′ end of transcripts (Fox-Walsh et al. 2011; Fu et al. 2011; Jan et al. 2011; Shepard et al. 2011; Derti et al. 2012; Hoque et al. 2013). These methods can not only identify pAs but also examine gene expression. Most methods use primers containing the oligo(dT) sequence for reverse transcription (RT). While efficient, oligo(dT) can prime at internal A-rich sequences (Nam et al. 2002), leading to false pA identification. This issue is usually addressed computationally by eliminating putative pAs in A-rich regions (Lee et al. 2007). However, this approach not only cannot guarantee full elimination of false positives caused by internal priming, but can also discard bona fide pAs.

Some sequencing methods are not affected by internal priming, including 3P-seq [poly(A)-position profiling by sequencing] (Jan et al. 2011) and 3′READS (3′ region extraction and deep sequencing) (Hoque et al. 2013). Both approaches involve removal of most of the poly(A) tail sequence by RNase H followed by ligation of an adapter to the 3′ end of digested RNA. A short poly(A) sequence

Corresponding author: btian@rutgers.edu

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.057075.116.

unalignable to the genome is used as evidence for the poly(A) tail, which is important for identification of genuine pAs. Since RT is primed at the 3′ adapter region, internal priming at A-rich sequences is avoided. However, both methods require a large amount of input RNA (25 µg RNA typically used by 3′READS and 20–70 µg RNA recommended for 3P-seq). In addition, pAs located in a long stretch of A's cannot be effectively identified by these methods because the short poly(A) tail left after RNase H digestion can be completely aligned to the A-stretch sequence, leaving no additional A's as evidence of the poly(A) tail.

Here we present a much improved version of 3′READS, named 3′READS+. In addition to the capability of addressing the internal priming issue, the method has several key advantages over the previous version and other related methods, including efficient capture of all RNAs with a poly(A) tail ≥10 nt; ability to reliably sequence a small amount of input RNA (as low as 100 ng total RNA) with high reproducibility, and generation of reads with an optimal number of T's (∼13) for accurate identification of genuine pAs located in A-stretch regions. Using 3′READS+, we revealed a sizable fraction of pAs in human genes that are located within a stretch of A's (≥5). With more complete and accurate identification of pAs, we assessed the frequency of APA in HeLa cells.

RESULTS AND DISCUSSION

A locked nucleic acid-containing oligonucleotide improves poly(A)$^+$ RNA capture and poly(A) protection

In our previous 3′READS method (Hoque et al. 2013), we used a DNA/RNA hybrid oligonucleotide (oligo) containing 45 deoxythymidines (T's) and five uridines (U's) to remove the bulk of poly(A) tail by RNase H, leaving behind a few A's that are annealed to the U's and are thus undigested by the enzyme (RNase H is known to digest DNA:RNA but not RNA:RNA duplexes). The terminal A's that are unalignable to the genome are considered as evidence of the poly(A) tail, allowing accurate identification of genuine pAs. In an initial effort to gain more clear evidence of the poly(A) tail by increasing the length of terminal A's, we increased the number of U's in the oligo from five to 15 ($T_{35}U_{15}$, Fig. 1A, top) and tested its ability to protect the poly(A) sequence from RNase H digestion using an in vitro synthesized RNA containing 60 A's, named A60. We found that $T_{35}U_{15}$ indeed protected 14–20 A's from digestion by RNase H, as expected. However, while desirable poly(A) protection was achieved with RNase H at 1/32 units (U)/reaction (Fig. 1A, bottom), variation of its concentration by merely twofold resulted in either over or insufficient digestion (1/16 and 1/64 U/reaction in Fig. 1A, respectively), indicating that U's do not give reliable protection of A's in RNase H digestion.

We reasoned that the observed lack of robustness in protection of A's by U's might be caused by interaction between the 14 and 20 remaining A's after RNase H digestion and the T's in $T_{35}U_{15}$ oligo or indiscriminant digestion of RNA:RNA molecules when the RNase H concentration was high. Since locked nucleic acid (LNA) oligos interact with RNAs with higher affinities than does RNA:RNA, and LNA:RNA duplexes are refractory to RNase H digestion (Kurreck et al. 2002), we designed an LNA/DNA hybrid oligo, $T_{15}(+TT)_5$, which contains 15 T's in the 5′ region and five locked deoxythymidines (denoted as +T) alternating with regular T's at the 3′ end (Fig. 1B, upper). Using an oligo containing 50 T's (T_{50}) as a control, we found that at 0.5 U RNase H/reaction, the highest concentration of RNase H tested, the $T_{15}(+TT)_5$ oligo preserved ∼13 A's, whereas the T_{50} and $T_{35}U_{15}$ oligos led to digestion of 60 A's into 3–5 A's. This result indicated that the $T_{15}(+TT)_5$ oligo is reliable for poly(A) protection in RNase H digestion.

We next tested the ability of $T_{15}(+TT)_5$ to capture RNAs with different poly(A) tail sizes. Using RNAs with 5, 10, or 15 consecutive A's (named A5, A10, and A15, respectively), we tested various washing conditions. In a binding buffer containing 150 mM NaCl, the biotin-$T_{15}(+TT)_5$ oligo immobilized on the streptavidin-coated magnetic beads successfully captured RNAs with ≥10 A's but not those with 5 A's (Fig. 1C,D). Reducing NaCl concentration of the washing buffer from 150 mM to 5 or 1 mM slightly decreased the relative binding of A15 and A10, when compared to A60, while adding 10% formamide to the washing buffer significantly reduced their bindings. We therefore decided to use a washing buffer containing 1 mM NaCl without formamide to capture RNAs with a poly(A) sequence ≥10 nt. Note that this design is important for comprehensive sequencing of poly (A)$^+$ RNAs because recent studies have shown that a substantial fraction of cellular mRNAs have short poly(A) tails (Chang et al. 2014; Subtelny et al. 2014). However, because a very short A-tail (∼5 nt) can also be added during the exosome-mediated degradation (Wlotzka et al. 2011), having the cutoff of 10 nt for poly(A) tail selection can balance sensitivity and specificity.

Efficient ligation steps improve cDNA yield and data quality

In our original 3′READS protocol, the RNA fragments after RNase H digestion were first ligated to a 5′ adenylated 3′ adapter with a truncated RNA ligase II, and then to a 5′ adapter by RNA ligase I in the same reaction tube, an approach often used in small RNA sequencing (Landgraf et al. 2007). Using in vitro synthesized control RNA, we found that this method (protocol A in Fig. 2A) was not efficient. Without polyethylene glycol (PEG), a crowding agent known to increase ligation efficiency, only 10% of RNA could be ligated with both 5′ and 3′ adapters (Fig. 2B). With PEG, the percentage increased to 33% (Fig. 2B). However, when we used PEG in our library preparation, 45% of the resultant reads contained no or short inserts (<23 nt), approximately fourfold higher than the protocol without PEG (Fig. 2C).

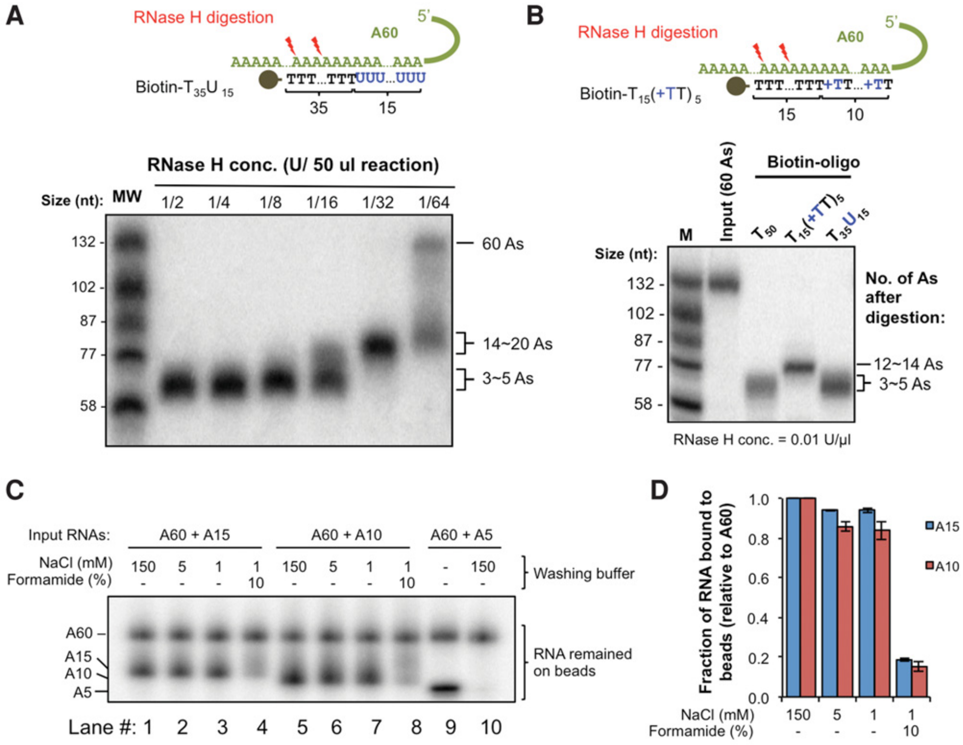


FIGURE 1. Digestion of poly(A) RNA with a LNA oligo. (*A*, *top*) Schematic showing digestion of the poly(A) tail annealed to the $T_{35}U_{15}$ oligo by RNase H. In theory, the A's hybridized to T's are digested by RNase H, whereas those to U's are not. RNase H digestion is indicated by a lightning symbol. The $T_{35}U_{15}$ oligo contains a 5′ biotin group that can bind to streptavidin-coated beads. (*Bottom*) Autoradiography showing digestion products of an RNA molecule containing 60 A's (named A60) by different amounts of RNase H (U/reaction is units per reaction) using the $T_{35}U_{15}$ oligo. MW, molecular weight markers (sizes indicated). Number of remaining A's in digestion products are indicated, which were calculated based on the molecular weight markers. (*B*, *top*) Schematic showing digestion of the poly(A) tail annealed to the T_{15}(+TT)$_5$ oligos. +T, locked deoxythymidine. (*Bottom)* Autoradiography showing digestion products of A60 by 1/2 U/reaction of RNase H with different oligos. Number of remaining A's in the digestion products is indicated. (*C*) Autoradiography showing binding of RNAs with different A's to the biotin-T_{15}(+TT)$_5$ attached to magnetic beads after washing with buffers containing different concentrations of NaCl and formamide. A60, A15, A10, and A5 have different numbers of consecutive A's and are otherwise the same. (*D*) Quantification of the amount of A15 and A10 bound to biotin-T_{15}(+TT)$_5$ relative to A60 in each washing condition based on the data in *C*.

This result indicates that PEG could cause ligation of adapters with short RNA contaminants or ligation between 5′ and 3′ adapters without insert.

In an effort to improve ligations of 5′ and 3′ adapters separately, we found that while PEG could significantly stimulate 3′ adapter ligation efficiency by >10-fold (Fig. 2D), its enhancement of 5′ adapter ligation was limited (Fig. 2E). In fact, PEG is problematic for 5′ adapter ligation because it also caused concatenation of RNA fragments, leading to a lower amount of desirable products (Fig. 2E). In view of these results, we first performed 5′ adapter ligation in the absence of PEG, followed by purification of RNA using oligo(dT) beads to eliminate unused 5′ adapters. Purified RNA was then ligated to the 3′ adapter in the presence of PEG. This new protocol (protocol B in Fig. 2A) resulted in a 5.8-fold increase of the amount of desirable product compared to protocol A without PEG (58% versus 10%) and a 1.8-fold increase compared to protocol A with PEG (Fig. 2B). Importantly, the fraction of reads with insert size <23 was ∼12%, comparable to protocol A without PEG (Fig. 2C).

3′READS+ is sensitive and robust

Based on the optimization experiments described above, we designed a new 3′READS protocol, named 3′READS+ (+ denotes improvement and use of LNA), as illustrated in Figure 3A. Briefly, poly(A)$^+$ RNAs were first selected using oligo (dT)$_{25}$ beads and fragmented by RNase III on the beads. After washing the beads, poly(A)$^+$ RNA fragments were eluted and ligated to a 5′ adapter (without PEG). The ligation products with a poly(A) tail were then purified using biotin-T_{15}(+TT)$_5$ attached to magnetic beads. Unused 5′ adapter was washed away during this step to eliminate ligation between 5′ and 3′ adapters. While the RNAs were on the beads, long poly(A) tails were trimmed to ∼13 nt by RNase H. This was followed by rigorous washing to discard any RNA fragments that cannot bind T_{15}(+TT)$_5$. After elution, RNA fragments were ligated with a 5′ adenylated 3′ adapter in the presence of PEG. The 5′ and 3′ adapters contained several random nucleotides next to the ligation end to mitigate ligation bias (Jayaprakash et al. 2011; Sun et al. 2011). The

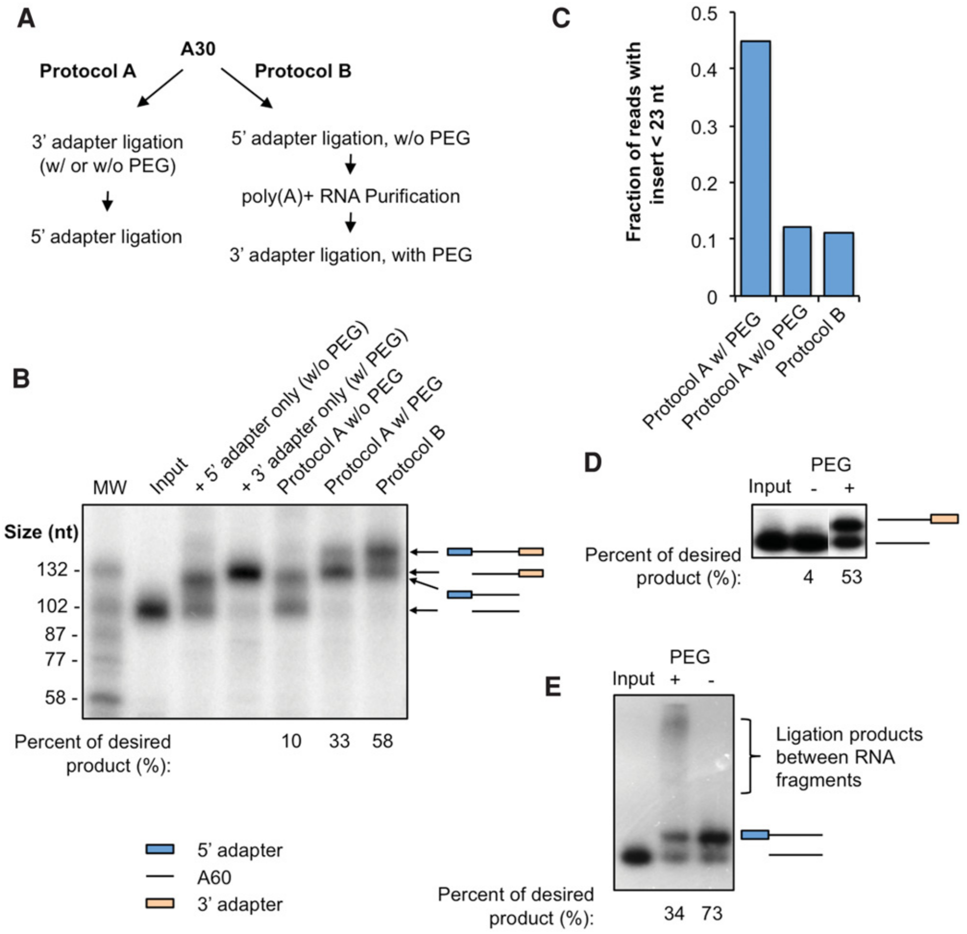
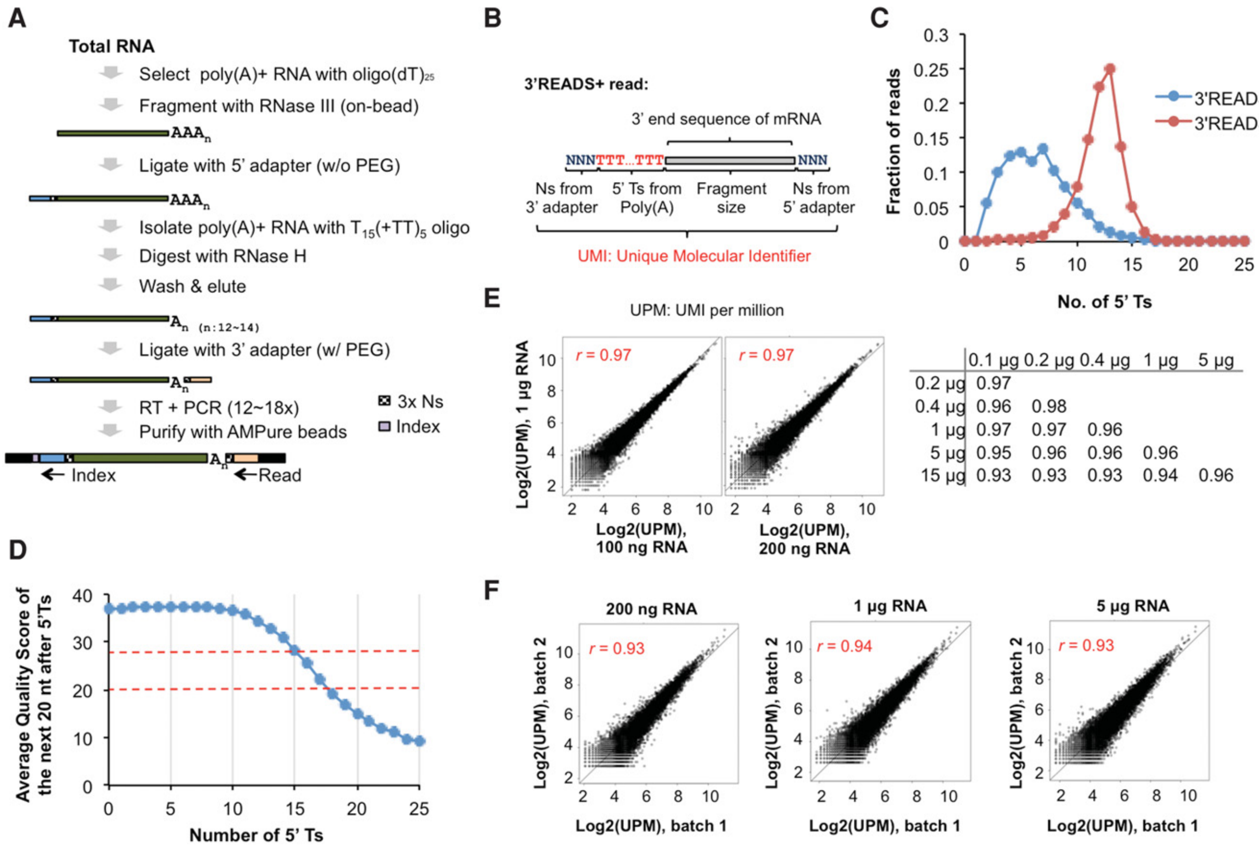


FIGURE 2. Optimization of 5′ and 3′ adapter ligation steps. (*A*) Ligation protocols tested. In protocol A, ligation with 3′ and 5′ adapters was performed sequentially in the same tube. The 5′ adapter is an RNA oligo with hydroxyl groups at both 5′ and 3′ ends, and the 3′ adapter is a 5′-adenylated DNA oligo with a 3′ blocker (ddC). In protocol B, 5′ adapter ligation was performed first without PEG, and the ligation product was purified using oligo(dT)$_{25}$ beads and then ligated to the 3′ adapter in the presence of PEG. (*B*) Autoradiography showing ligation products using different ligation protocols. MW, molecular weight markers (sizes indicated). Schematics of ligation products and their expected sizes are shown on the *right*. The percent of product shown *below* the image is based on the amount of RNA with both 5′ and 3′ adapters relative to that of input RNA. (*C*) Bar plot showing the fractions of raw reads with inserts <23 nt from the 3′READS libraries prepared with ligation protocol A with (*left*) or without (*right*) PEG and with ligation protocol B. (*D*) Autoradiography showing the effect of PEG on 3′ adapter ligation. RNAs corresponding to the bands are indicated. Percent of product shown *below* the image is based on the amount of RNA with ligated 3′ adapter relative to that of input RNA. (*E*) Autoradiography showing the effect of PEG on 5′ adapter ligation. Percent of product shown *below* the image is based on the amount of RNA with ligated 5′ adapter to that of input RNA.

ligation products were then reverse transcribed, PCR-amplified (12–18 cycles) with primers containing an index sequence for multiplexing in sequencing, and size-selected using AMPure beads.

The 3′READS+ libraries were sequenced from the 3′ adapter region (Fig. 3A), yielding reads beginning with several random N's derived from the 3′ adapter (three N's in this study) followed by a run of T's (named 5′T's) corresponding to the poly(A) tail and a reverse complement sequence of the 3′ end region upstream of the poly(A) tail (Fig. 3B). Reads with ≥2 unaligned 5′ T's after mapping to the genome were called poly(A) site-supporting (PASS) reads (Hoque et al. 2013). Using HeLa cell RNA, we found that, consistent with our in vitro result, the number of 5′T's in PASS reads peaked at ~13 nt and were <17 nt for 99% of reads (Fig. 3C), indicating protection of ~13 A's at the 5′-most portion poly(A) tail by the T_{15}(+TT)$_5$ oligo. In contrast, the data from the original 3′READS method showed a peak at ~5 nt (Fig. 3C).

Sequencing homopolymers is challenging for current sequencing platforms (Quail et al. 2012). We thus examined the sequencing quality after the 5′T region, using averaged quality score (QS) over 20 immediately downstream bases. We found that sequencing up to fifteen 5′T's had little effect on the quality of subsequent bases, with the average QS all >28, a value considered to be high quality (Fig. 3D). The QS dropped below 28 but above 20 (a cutoff for poor quality) after sequencing of 16–17 5′T's (Fig. 3D). Eighteen 5′T's led to subsequent bases having QS below 20 (Fig. 3D). This result indicates that using T_{15}(+TT)$_5$ to generate RNA fragments with a peak of ~13 A's and no more than 17 A's is the optimal



	0.1 μg	0.2 μg	0.4 μg	1 μg	5 μg
0.2 μg	0.97				
0.4 μg	0.96	0.98			
1 μg	0.97	0.97	0.96		
5 μg	0.95	0.96	0.96	0.96	
15 μg	0.93	0.93	0.93	0.94	0.96

FIGURE 3. 3′READS+. (*A*) The 3′READS+ protocol incorporating optimized RNase H digestion and ligation steps. AAA$_n$, poly(A) tail; A$_n$, shortened poly(A) tail. 5′ adapter, 3′ adapter, random sequences in the adapters (3× N's), and index region in PCR primer are indicated. (*B*) Schematic showing different parts of a raw read generated by 3′READS+. (*C*) Number of 5′ T's in reads from 3′READS+ and 3′READS. Only the reads mapped to pAs are shown. (*D*) Sequencing quality of the bases after 5′T's. (*Left*) Schematic showing the analyzed region. (*Right*) The average quality score (QS) of the next 20 bases after 5′T's is shown. QS > 28 is usually considered high quality, whereas <20, low quality. (*E*, *left*) Scatter plots comparing log$_2$ (UPM) of transcript between libraries with different amounts of input RNAs. (*Right*) Table summarizing correlations between different samples. UPM, UMI per million. UMI was based on cleavage site location, number of 5′T's, and the three random nucleotides from the 3′ adapter, as shown in *B*. Only transcripts with more than five unique PASS reads were used for the plots. Pearson correlation coefficient (*r*) is indicated in each graph and the table. (*F*) As in *E*, except that samples from different batches were compared.

choice, maximizing the number of A's that can be used for pA identification and yet not compromising sequencing quality in the subsequent region.

We next tested the sensitivity and reproducibility of 3′READS+, using 100 ng, 200 ng, 400 ng, 1 µg, 5 µg, or 15 µg total RNAs from HeLa cells as inputs. Because RNA fragments can be overamplified by PCR, leading to redundant reads, we used the random sequence (3× N's) derived from the 3′ adapter, the number of 5′ T's, and the cleavage site location, collectively called unique molecular identifier (UMI), to identify unique RNA fragments and quantify the expression level of each pA isoform (illustrated in Fig. 3B). In addition, if the 5′ adapter region was reached by sequencing (when RNA fragment was short), the RNA fragment size and the random sequence from the 5′ adapter were also used as part of UMI (Fig. 3B). We calculated UMI per million (UPM) as the quantitative measure of transcript expression. Comparisons between libraries with different amounts of input RNA showed good consistency, with Pearson's correlation coefficients above 0.95 for all comparisons (Fig. 3E), indicating that 3′READS+ has high sensitivity for input RNA as low as 100 ng at least, and high linearity from 100 ng to 5 µg. In addition, we purposely prepared libraries using the same input RNA but at different times to gauge batch differences. As shown in Figure 3F, the Pearson correlation coefficients between different batches were above 0.93, indicating low batch effect. Therefore, 3′READS+ is both sensitive and robust.

3′READS+ identifies pAs in genomic A-stretch sequences

pAs can be located within a stretch of A's in the genome, making them difficult to identify. For simplicity, these pAs are called A-stretch pAs (illustrated in Fig. 4A). They would be discarded from the data generated by oligo(dT)-based 3′ end sequencing, because they could not be distinguished from false sites stemmed from internal priming. Current

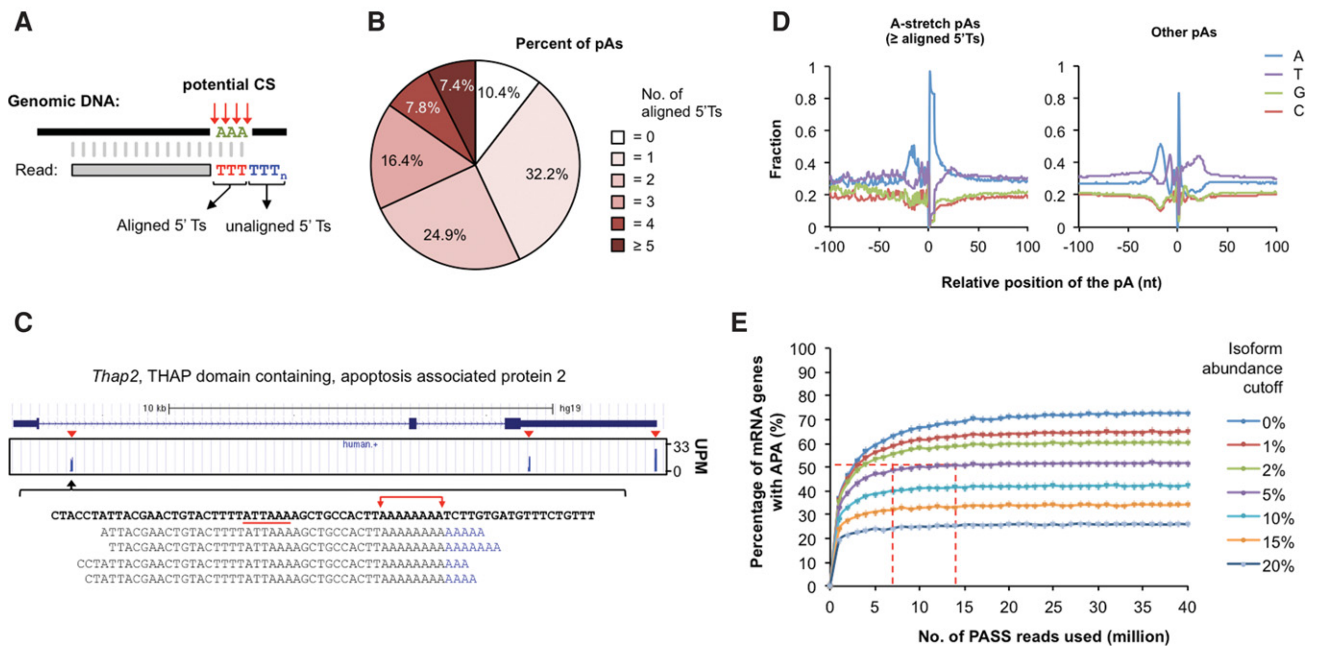


FIGURE 4. 3′READS+ identifies a large number of pAs within A-stretch regions. (*A*) Schematic showing alignment of a PASS read with an A-stretch region. (*B*) Distribution of pAs with different A-stretch region sizes (number of aligned 5′T's) using PASS reads from HeLa cells. (*C*) Nucleotide profiles around the A-stretch and other pAs. (*D*) An example gene (*Thap2*) with an A-stretch pA. (*Top*) Gene structure as shown in the UCSC Genome Browser. (*Middle*) UPM values for pAs of *Thap2*. Three alternative pAs are indicated. (*Bottom*) Sequence surrounding the A-stretch pA. The AUUAAAPAS is underlined, and the A-stretch region is indicated. Several 3′READS+ reads are shown to indicate additional A's used as evidence for the poly(A) tail. (*E*) Assessment of APA rate in HeLa cells using different numbers of PASS reads and different isoform abundance cutoffs. The plateaued value (51% genes with APA) with the 5% isoform abundance cutoff is indicated by a horizontal line, and two vertical lines indicate seven and 14 million reads, which gave rise to 49% and 51% APA rates, respectively.

nonoligo(dT)-based methods generate reads with only short A's/T's as poly(A) tail evidence, making them insufficient to identify pAs located within a long stretch of genomic A's. Failure to identify A-stretch pAs could lead to incomplete mapping of pAs and inaccurate quantification of APA isoforms or gene expression. With the long 5′T sequence (~13 nt) generated by 3′READS+, we set out to address to what extent A-stretch pAs exist in the human genome. Using the HeLa cell data, we found that ~7.4% of pAs detected in HeLa cells were within five or more genomic A's (Fig. 4B). Note that for some A-stretch pAs, not all the constituent cleavage sites were within a stretch of pAs. In these cases, exclusion of A-stretch cleavage sites would lead to partial quantification of pA isoform expression. One example of A-stretch pA is shown in Figure 4C, where an intronic pA of the *Thap2* (THAP domain containing, apoptosis-associated protein 2) gene is within a stretch of eight genomic A's. 3′READS+ reads containing 11–15 5′T's provided crucial evidence for the identification of this pA (Fig. 4C). Note that we in fact do not know the precise cleavage site position within the A-stretch region (illustrated in Fig. 4A), because the genomic A's are indistinguishable from the poly(A) tail sequence. Nucleotide profiles around all A-stretch pAs (≥5 A's) showed upstream A-rich and downstream U-rich peaks similar to those of other pAs (Fig. 4D), suggesting that A-stretch pAs are flanked by *cis* elements similar to other pAs. Taken together, these data indicate that there exists a sizable fraction of pAs in the human genome that are located in A-stretch sequences and thus have hitherto been largely overlooked. More detailed analyses of this group of pAs, such as their functional relevance to APA regulation, await future investigations.

APA in HeLa cells

With a total of 42 million (M) PASS reads generated by 3′READS+ with HeLa cell RNAs during the development of the 3′READS+ method (Supplemental Table S1), we asked what the APA frequency was for genes expressed in a given type of human cell, like HeLa, an important question that had not been addressed so far. Using random sampling, we assessed how the observed APA frequency changes with different sequencing depths and relative abundance cutoff for calling isoforms (Fig. 4E). As expected, more PASS reads allowed us to identify more genes with APA, and increasing the isoform relative abundance cutoff led to lower APA rates. For example, with 40 M PASS reads, 73% and 26% of genes were found to display APA with 0% and 20% cutoffs, respectively (Fig. 4E). After using relative abundance of 5% to filter APA isoforms, a commonly used cutoff value, we found that the percent of genes expressed in HeLa cells displaying APA plateaued at ~51% with 14 M PASS reads. However, there was only a slight drop of the rate to 49% when 7 M PASS reads were used. Thus, we conclude that about half of the genes expressed in HeLa cells display APA and >7 M PASS reads are needed to have a complete assessment of APA with HeLa cell samples. It is notable, however, that these numbers are likely to vary in other cell types when the diversity of transcriptome and APA mechanisms is different.

In summary, we present 3′READS+, a method that can be used to accurately, comprehensively, and quantitatively examine APA with limited amounts of input RNA. It is also conceivable that with complete and accurate mapping of the 3′ ends, the method can be readily used to examine gene expression through 3′ end counting.

MATERIALS AND METHODS

Cells and RNAs used in this study

Human HeLa cells were cultured in high glucose Dulbecco's modification of Eagle's medium (DMEM) with 10% fetal bovine serum (Atlanta Biologicals). Total cellular RNA was extracted using the TRIzol reagent (Life Technologies). RNA concentration was measured with NanoDrop 2000 (Thermo Scientific) and RNA quality was examined on an Agilent Bioanalyzer using the RNA 6000 pico kit.

In vitro synthesized RNAs

Plasmids expressing RNAs containing 15, 30, or 60 terminal A's (A15, A30, or A60, respectively), named pALL-A15, pALL-A30, or pALL-A60, respectively, were obtained from Bioo Scientific Co. and were previously described (Hoque et al. 2014). Plasmids expressing RNAs containing 5 or 10 terminal A's (A5 or A10, respectively) were made by subcloning sequences containing 5 and 10 A's into the pALL-A60 plasmid using EcoRI and PvuII sites. All in vitro transcription products of these plasmids were the same except for the poly(A) length. The template for A0 was prepared by cutting the HindIII site right upstream of the A60 sequence in the pALL-A60 plasmid. Radioactively labeled RNAs were synthesized by in vitro transcription with SP6 RNA polymerase (Promega) and linearized plasmids. α-P32 uridine 5′-triphosphate (PerkinElmer) was used for labeling of RNA. RNAs were purified with Micro Bio-Spin P-30 gel columns (Bio-Rad).

RNase H digestion assay

Radioactive A60 RNA was first denatured by heat, captured by biotin-$T_{35}U_{15}$ (IDT), biotin-T_{50} (IDT), or biotin-T_{15}(+TT)$_5$ (Exiqon) oligos attached to magnetic beads (Dynabeads MyOne Streptavidin C1, Life Technologies) at room temperature for 30 min on a rotator, and digested with different concentrations of RNase H (Epicentre) at 37°C for 30 min. The whole reaction was mixed with an equal volume of 2× RNA loading buffer (95% formamide, 0.02% SDS, 0.02% bromophenol blue, 0.01% xylene cyanol, and 20 mM EDTA), incubated at 70°C for 5 min, and put on a magnetic stand. The supernatant was resolved on an 8% TBE-Urea-polyacrylamide gel. Radioactive signals were analyzed using a phosphor screen (Amersham) and a Typhoon 9400 scanner (Amersham). Image quantification and calculation of molecular weight using molecular size makers were performed with the ImageJ software (Schneider et al. 2012).

RNA-binding assay

The A60 RNA was mixed with A15, A10, or A5 RNAs, followed by heat denaturation and incubation with the biotin-T_{15}(+TT)$_5$ oligo attached to magnetic beads (Dynabeads MyOne Streptavidin C1, Life Technologies) at room temperature for 30 min on a rotator. The beads were then washed three times with buffers containing different concentrations of NaCl and formamide, mixed with 1× RNA loading buffer, heated at 70°C for 5 min, and put on a magnetic stand. RNA in the supernatant was then analyzed by gel electrophoresis and by autoradiography as described above. The A10 and A15 signals were normalized to the A60 signal in the same lane.

Adapter ligation assays

In vitro transcribed radioactive A30 was captured using oligo(dT)$_{25}$ beads, dephosphorylated with calf intestinal alkaline phosphatase (NEB) at 37°C for 45 min, and then phosphorylated with T4 polynucleotide kinase (NEB) at 37°C for 45 min (on a rotator). RNA was then washed to remove free ATP and eluted from the beads with nuclease-free H_2O. Two types of ligation protocols were tested. In protocol A, a 5′ adenylated 3′ adapter made by the 5′ DNA Adenylation Kit (NEB) was ligated to A30 using T4 RNA ligase II (truncated KQ version, NEB) with or without 15% polyethylene glycol (PEG) 8000 (NEB) at 22°C for 1 h. The reaction was then incubated in the same tube with a 5′ adapter, 1 mM ATP and T4 RNA ligase I at 22°C for 1 h. In protocol B, A30 was ligated to the 5′ adapter with T4 RNA ligase I (NEB) at 22°C for 1 h, in the presence of ATP. The RNA was then captured using oligo(dT)$_{25}$ magnetic beads (NEB) and eluted with H_2O at 70°C for 2 min, followed by ligation to the 5′ adenylated 3′ adapter by the T4 RNA ligase I in the presence of 15% PEG 8000. The RNAs in the reactions were then purified by phenol-chloroform extraction, precipitated in ethanol, and examined by gel electrophoresis and by autoradiography as described above.

3′READS+

Poly(A)$^+$ RNA in 0.1–15 µg of total RNA was captured using 12 µL of oligo(dT)$_{25}$ magnetic beads (NEB) in 200 µL 1× binding buffer (10 mM Tris-Cl, pH 7.5, 150 mM NaCl, 1 mM EDTA, and 0.05% TWEEN 20) and fragmented on the beads using 1.5 U of RNase III (NEB) in 30 µL RNase III buffer (10 mM Tris-Cl, pH 8.3, 60 mM NaCl, 10 mM $MgCl_2$, and 1 mM DTT) at 37°C for 15 min. After washing away unbound RNA fragments with binding buffer, poly(A)$^+$ fragments were eluted from the beads with TE buffer (10 mM Tris-Cl, 1 mM EDTA, pH 7.5) and precipitated with ethanol, followed by ligation to 3 pmol of heat-denatured 5′ adapter (5′-CCUUGGCACCCGAGAAUUCCANNNN, Sigma) in the presence of 1 mM ATP, 0.1 µL of SuperaseIn (Life Technologies), and 0.25 µL of T4 RNA ligase 1 (NEB) in a 5 µL reaction at 22°C for 1 h. The ligation products were captured by 10 pmol of biotin-T_{15}-(+TT)$_5$ attached to 12 µL of Dynabeads MyOne Streptavidin C1 (Life Technologies). After washing with washing buffer (10 mM

Tris-Cl, pH 7.5, 1 mM NaCl, 1 mM EDTA, and 0.05% TWEEN 20), RNA fragments on the beads were incubated with 0.01 U/µL of RNase H (Epicentre) at 37°C for 30 min in 30 µL of RNase H buffer (50 mM Tris-Cl, pH 7.5, 5 mM NaCl, 10 mM $MgCl_2$, and 10 mM DTT). After washing with RNase H buffer, RNA fragments were eluted from the beads in elution buffer (1 mM NaCl, 1 mM EDTA, and 0.05% TWEEN 20) at 50°C, precipitated with ethanol, and then ligated to 3 pmol of heat-denatured 5′ adenylated 3′ adapter (5′-rApp/NNNGATCGTCGGACTGTAGAACTCTGAAC/3ddC) with 0.25 µL T4 RNA ligase 2 (truncated KQ version, NEB) at 22°C for 1 h in a 5 µL reaction containing 15% PEG 8000 (NEB) and 0.2 µL of SuperaseIn (Life Technologies). The ligation products were then precipitated and reverse transcribed using M-MLV reverse transcriptase (Promega), followed by PCR amplification using Phusion high-fidelity DNA polymerase (NEB) and bar-coded PCR primers for 12–18 cycles (12 cycles for 15 µg input RNA, 13 cycles for 5 µg input, 15 cycles for 1 µg input, and 18 cycles for inputs <1 µg). Both RT primer (5′-GTTCAGAGTTCTACAGTCCGAC GATC) and PCR primers with indexes were described previously in Hoque et al. (2013). PCR products were size-selected twice with AMPure XP beads (Beckman Coulter), using 0.6 volumes of beads (relative to the PCR reaction volume) to remove large DNA molecules and an additional 0.4 volumes of beads to remove small DNA molecules. The eluted DNA was selected again with 1 volume of AMPure XP beads to further remove small DNA molecules. The size and quantity of the libraries eluted from the AMPure beads were examined using a high sensitivity DNA kit on an Agilent Bioanalyzer (Agilent). The library concentrations were further measured by qPCR using primers corresponding to 5′ and 3′ end regions of cDNAs. Libraries were sequenced on an Illumina HiSeq 2000 machine (1 × 50 bases). Raw read numbers are shown in Supplemental Table S1.

Data analysis

The sequence corresponding to the 5′ adapter was first removed from raw 3′READS+ reads using the cutadapt program (Martin 2011). The 5′ random nucleotides and 5′T's in the reads were trimmed before the reads were mapped to the human (hg19) genome using bowtie2 (global mode) (Langmead and Salzberg 2012). Only reads with a mapping quality score (MAPQ) ≥10 were used for further analysis. The trimmed 5′T's of each read were then compared to the genomic region downstream from the last aligned position of the read to identify aligned 5′T's. The reads with ≥2 nongenomic 5′T's after this process were called polyA site supporting (PASS) reads. Cleavage sites within 24 nt of each other were clustered into pAs as previously described (Hoque et al. 2013). UPM of a transcript with a given pA was calculated with unique PASS reads, based on 5′ random nucleotides, number of 5′T's, and cleavage site location. The 3′READS data were the mouse mixed cell line data we previously published (Hoque et al. 2013). Sequencing quality scores were retrieved using the Biostrings package of Bioconductor (Gentleman et al. 2004). The raw data can be downloaded from NCBI's GEO database (GSE84170).

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank Mainul Hoque and Yu Zhang for technical assistance at the early stage of this project and other members of the B.T. laboratory for helpful discussions. This work was funded by the National Institutes of Health (grant no. GM084089) to B.T.

Received May 3, 2016; accepted June 21, 2016.

REFERENCES

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.

Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22:** 1173–1183.

Flavell SW, Kim TK, Gray JM, Harmin DA, Hemberg M, Hong EJ, Markenscoff-Papadimitriou E, Bear DM, Greenberg ME. 2008. Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. *Neuron* **60:** 1022–1038.

Fox-Walsh K, Davis-Turak J, Zhou Y, Li H, Fu XD. 2011. A multiplex RNA-seq strategy to profile poly(A^+) RNA: application to analysis of transcription response and 3′ end formation. *Genomics* **98:** 266–271.

Fu Y, Sun Y, Li Y, Li J, Rao X, Chen C, 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.

Geisberg JV, Moqtaderi Z, Fan X, Ozsolak F, Struhl K. 2014. Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell* **156:** 812–824.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5:** R80.

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

Hoque M, Li W, Tian B. 2014. Accurate mapping of cleavage and polyadenylation sites by 3′ region extraction and deep sequencing. *Methods Mol Biol* **1125:** 119–129.

Jan CH, Friedman RC, Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. *Nature* **469:** 97–101.

Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R. 2011. Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Res* **39:** e141.

Ji Z, Tian B. 2009. Reprogramming of 3′ untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS One* **4:** e8419.

Ji Z, Lee JY, Pan Z, Jiang B, Tian B. 2009. Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc Natl Acad Sci* **106:** 7028–7033.

Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7:** 1009–1015.

Kurreck J, Wyszko E, Gillen C, Erdmann VA. 2002. Design of antisense oligonucleotides stabilized by locked nucleic acids. *Nucleic Acids Res* **30:** 1911–1918.

Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, et al. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129:** 1401–1414.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359.

Lee JY, Yeh I, Park JY, Tian B. 2007. PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Res* **35:** D165–D168.

Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. 2013. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev* **27:** 2380–2396.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17:** 10–12.

Mayr C, Bartel DP. 2009. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138:** 673–684.

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

Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM. 2010. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143:** 1018–1029.

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

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* **320:** 1643–1647.

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9:** 671–675.

Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17:** 761–772.

Singh P, Alley TL, Wright SM, Kamdar S, Schott W, Wilpan RY, Mills KD, Graber JH. 2009. Global changes in processing of mRNA 3′ untranslated regions characterize clinically distinct cancer subtypes. *Cancer Res* **69:** 9422–9430.

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.

Sun G, Wu X, Wang J, Li H, Li X, Gao H, Rossi J, Yen Y. 2011. A bias-reducing strategy in profiling small RNAs using Solexa. *RNA* **17:** 2256–2262.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476.

Wlotzka W, Kudla G, Granneman S, Tollervey D. 2011. The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. *EMBO J* **30:** 1790–1803.

Wu X, Liu M, Downie B, Liang C, Ji G, Li QQ, Hunt AG. 2011. Genome-wide landscape of polyadenylation in *Arabidopsis* provides evidence for extensive alternative polyadenylation. *Proc Natl Acad Sci* **108:** 12533–12538.

Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, Li W. 2014. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nat Commun* **5:** 5274.

Zhang H, Lee JY, Tian B. 2005. Biased alternative polyadenylation in human tissues. *Genome Biol* **6:** R100.