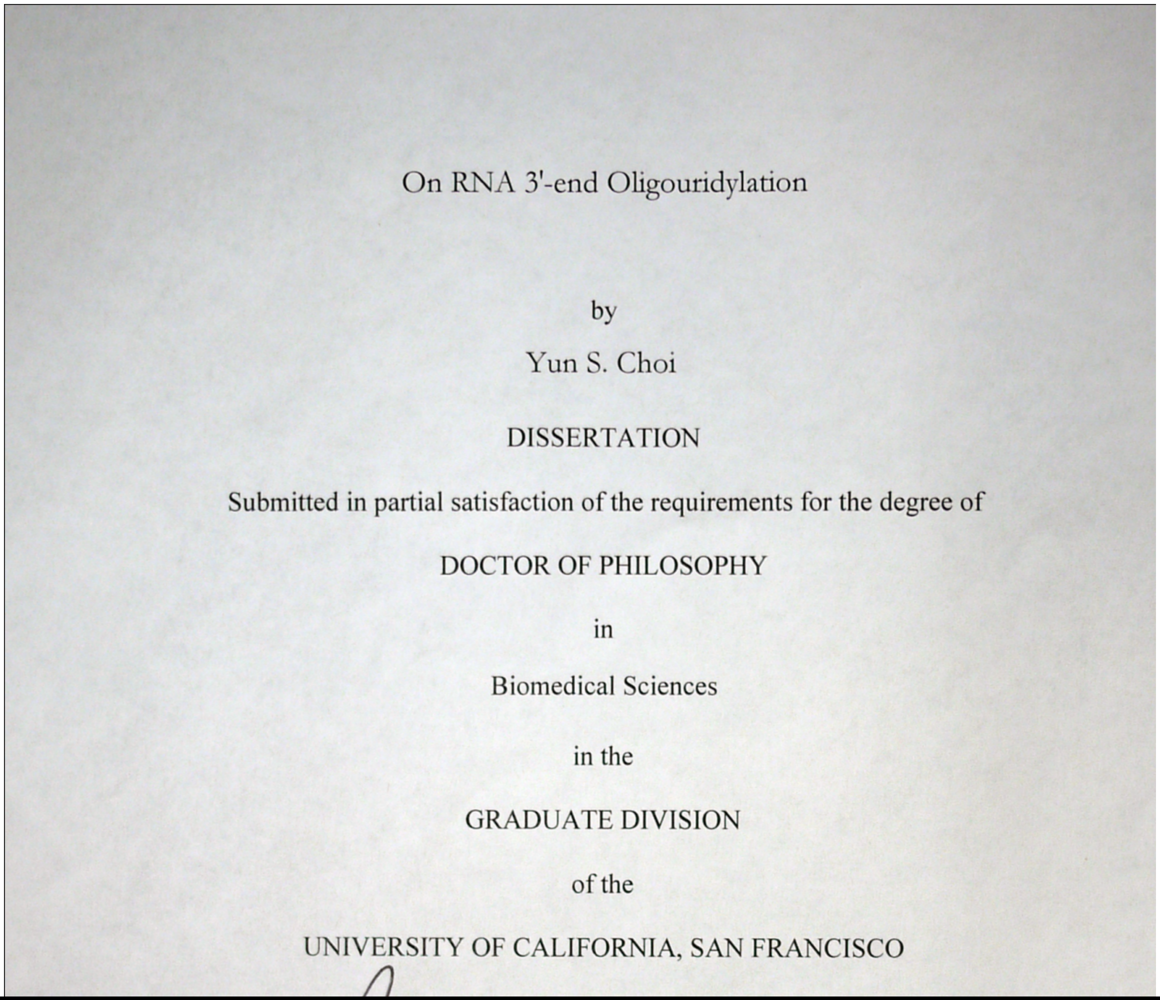


On RNA 3'-end Oligouridylation

by

Yun S. Choi

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Dedication

In memory of those who made this possible but have passed on:

To my grandmother who raised a most unruly grandson,

To Lisa who showed me the way,

And, Yuki who showed love and loyalty are not bounded by size or humanity.

Acknowledgements

Even for my optimistic nature, I could not have wished for a more supportive community for grad school. From the start, the program administrators took care of any and all red tape and moreover were an abundant source of free food. At the end, it took the patience, guidance and encouragement of my committee and mentor to see this project succeed. Words cannot do justice to the lengths they went to so that I could reach the goal.

UCSF is an amazing place for science where brilliant people are at every turn. More importantly, the atmosphere is one of openness that encourages collaboration. I was fortunate to share the expertise of the Leavitt and German labs. They're good people.

I am embarrassed at how naïve I was when I joined Michael's lab. It is due to him that I could try any experiment (almost) and practice and develop the skills needed to be a scientist. Michael also fostered an atmosphere where everyone was respectful and supportive of each other. He brought together a wonderful group of people, many of whom are dear friends even after they have all moved on to new things.

Trinna was there from the very beginning (I haven't forgotten about the van and the garage) where her infectious enthusiasm and passion sustained me through an abysmal period. At the end there was Nikki, Hunter and Ruby to share in life and when death visited again. In between, there were many talented post-docs and technicians who set a high standard. I was also fortunate to have equally bright and supportive classmates. We had more than just a common interested in science; they were there to share in celebrations and struggles as we figured out just what we had signed up for.

It is because of all of them that I was able to conquer mountains, literal and figurative.

On RNA 3'-end Oligouridylation

by Yun S. Choi

Abstract

Non-templated 3'-end oligouridylation is a poorly characterized RNA modification found in many species including humans. Recent studies have reported non-templated 3'-end oligouridylation of small non-coding RNAs and long mRNAs. To determine the prevalence of oligouridylation in higher eukaryotes, we used strand-specific paired-end RNA-Seq to deeply survey the population of small RNAs in human cells. We used custom sequence analysis programs to identify non-templated addition to RNA. Our data revealed widespread non-templated nucleotide addition to the 3'-ends of many classes of RNA, with short stretches of uridine being the most frequently added nucleotide.

To expand the search for RNAs subject to non-templated oligouridylation beyond those identified in the small RNA population, several changes to the RNA-Seq library preparation method were tested. A new RNA-Seq library was prepared based on these optimizations using total RNA purified from human embryonic stem cells. Analysis shows that RNA generated from repetitive elements are subject to non-templated oligouridylation.

Table of Contents

List of Tables

List of Figures

Chapter 1: General Introduction

In the beginning was RNA, or so sayeth the RNA world hypothesis (Yarus 2011). Evidence from all domains of life speaks to the ancient origins of the diverse and essential roles played by RNA in all facets of cellular activity. For example, the process of protein synthesis is carried out using the same mechanism in all organisms in all domains. Messenger RNA (mRNA) carrying information in a universal genetic code is translated into a polypeptide in a ribozyme complex with the help of adaptor RNAs (Cavicchioli 2011). But RNA is more than a transient copy of information stored in DNA; RNA is involved in regulating gene expression at every step along the path from transcription to eventual turnover of the message-bearing RNA. Many of these processes are so fundamental that they have been conserved from prokaryotes to complex eukaryotes.

Post-transcriptional processing of RNA genes

It is easy to appreciate the importance of synthesizing mRNAs to express proteins but non-coding RNA genes are just as essential for life. The maturation process for many classes of small RNA genes have been conserved over the course of evolutionary history. For example, ribosomal RNAs (rRNA) (Kaczanowska and Ryden-Aulin 2007; Henras et al. 2008) and transfer RNAs (tRNA) (Hartmann et al. 2009; Heinemann et al. 2010; Phizicky and Hopper 2010) are made by similar biogenesis pathway in all domains of life using related enzymes and accessory RNAs. Both of these classes of RNA require endonucleases to cleave their respective precursor transcripts that are subsequently trimmed by exonucleases to their mature sequences. In eukaryotes, other classes of small RNAs such as microRNAs (miRNAs) (Yang and Lai 2011) and small nucleolar RNAs (snoRNAs) (Dieci et al. 2009) are made in a similar fashion of processing by endo- and exonucleases.

Yet another type of processing in the generating of some mature RNA genes is the post-transcriptional addition of nucleotides to an initial transcript. Nearly all tRNA genes require the non-templated post-transcriptional addition of –CCA to their 3'-ends (Hou 2010). U6 small nuclear RNA (snRNA) undergoes 3'-trimming followed by non-templated 3'-oligouidylation in the course of its maturation (Terns et al. 1992). The initial transcript of 5S rRNA also undergoes trimming before the addition of non-templated uridines at the 3'-end (van Hoof et al. 2000; Perumal and Reddy 2002).

Non-templated nucleotide addition and RNA decay

RNA turnover is also essential for life with the importance of RNA degradation seen in the numerous RNases present in all organisms, some that are conserved in all domains (Yang 2011). The first enzyme identified as an RNA polymerase was polynucleotide phosphorylase (PNPase) which ironically happens to have a key role in RNA degradation (Xu and Cohen 1995). PNPase is an ancient enzyme present in bacteria, eukaryotic organelles and is related to the eukaryotic and archaeal exosomes (Schmid and Jensen 2008). The apparent paradox of an RNA degrading enzyme with RNA polymerase activity is reconciled by the activity of RNases being determined by RNA structure. RNA exonucleases have difficulty on structured RNA substrates and the addition of non-templated bases provides a handle to grab on to the RNA to be degraded, either directly or by recruiting other RNA processing complexes (Yang 2011). PNPase is able to polymerize a heterogeneous tail, predominantly composed of adenosines, to the 3'-end of RNA which is then degraded 3' → 5' by PNPase or other exonucleases (Mohanty and Kushner 2000). The enzyme responsible for polyadenylation in prokaryotes adds on a longer tail than PNPase but long polyadenylation in prokaryotes also acts to destabilize RNA (O'Hara et al. 1995).

In eukaryotes, 3'-end polyadenylation is associated with mRNA stability and increased gene expression. However, the ancient function of polyadenylation as a destabilizing mark is still maintained in certain contexts. Polyadenylation acts to destabilize RNA in chloroplasts and plant mitochondria (Zimmer et al. 2009). Yeast mitochondrial transcripts are destabilized with heterogeneous nucleotide addition composed of mostly adenines and uracil (Butow et al. 1989). There is conflicting data in human mitochondria where some mitochondrial RNAs seem to be destabilized by polyadenylation but other mitochondrial transcripts requires the post-transcriptional addition of a polyA tail to make the stop codon and a translatable mRNA (Borowski et al. 2010) . In the nucleus, the TRAMP complex (Trf4/Air2/Mtr4p) adds a short poly-A tail to improperly processed nuclear RNAs to promote their degradation by the nuclear exosome (LaCava et al. 2005; Vanacova et al. 2005).

Eukaryotic mRNA processing

Eukaryotes have developed several modifications to the basic process of translation first evolved in prokaryotes that help to stabilize RNA and distinguish cellular messages from viral intruders. One such modification is the addition of a modified guanidine nucleotide to the initial 5'-triphosphate of newly transcribed RNA (Moore and Proudfoot 2009). The capping complex is recruited to the C-terminal domain of RNAPol-II and transcripts are quickly capped once transcription begins (Peterlin and Price 2006). Capping is not limited to protein coding mRNAs as snRNAs and a few snoRNAs also posses this modification (Jacobson and Pederson 1998). U6 snRNA is transcribed by RNAPol-III but it too is modified on its 5'-end with the unusual monomethyl addition to the γ-phosphate (Singh and Reddy 1989). In addition to protecting the 5'-end from exonucleases, the cap is important for the localization and function of capped snoRNAs with the subcellular localization of capped snoRNAs changing as the methylation state of the cap changes (Jacobson and Pederson 1998).

The development of capping required the development of specialized enzymes to remove that protective group. Multiple decapping enzymes and regulatory factors are present in eukaryotes to control decapping in a tissue specific and RNA specific manner (Coller and Parker 2004; Jiao et al. 2006). After removal of the 5'-cap, RNA can be degraded 5' → 3' by Xrn1 or Xrn2 (Coller and Parker 2004). Loss of Xrn1 in *D. melanogaster* results in defects in gastrulation and thorax closure (Grima et al. 2008). Loss of the *C. elegans* homologue of Xrn1 shows a related phenotype where the worms have defects in embryonic ventral closure (Newbury and Woollard 2004) again illustrating the conserved and essential functions of regulating RNA abundance.

Gene expression in eukaryotes is further complicated by the presence of introns in the initial pre-mRNA transcript. A large complex of RNAs and proteins called the spliceosome must precisely recognize the boundaries between intron and exon (Rino and Carmo-Fonseca 2009). Sequential transesterification reaction releases the intron from the pre-mRNA in the form of a lariat with the 5'-end of the intron joined to itself at the branch point. The lariat form protects the 5'-end from exonucleases. A debranching enzyme is needed to hydrolyze the 5'-end linkage from the branch point to allow introns to be degraded (Chapman and Boeke 1991). Most introns are degraded quickly but introns can harbor non-coding RNA genes such as snoRNAs and miRNAs (Chiang et al. ; Dieci et al. 2009). In rare cases of short introns, the debranched intron can adopt a stem-loop structure and be processed directly as a miRNA precursor (Okamura et al. 2007; Ruby et al. 2007).

The secondary structure of snoRNAs and miRNAs and their assembly into protein-RNA complexes may block exonucleases to allow the expression of these genes (Dragon et al. 2001).

Non-templated oligouridylation

Oligouridylation is another type of RNA post-transcriptional modification has been recently identified in eukaryotes. Similar to the effects of non-templated RNA polyadenylation, many groups have reported the destabilizing effects of non-templated RNA oligouridylation. Examples of this modification have been observed in many species for both mRNA and non-coding RNAs. Histone mRNAs are among the rare RNAPol-II transcripts that lack poly-A tails. The 3'-ends of the mRNA of these ancient genes are similar to prokaryotic mRNAs that are Rho-independently terminated and form a 3'-end stem-loop that is bound and protected by a protein complex (Krieg and Melton 1984). After DNA replication is complete, the 3'-ends of histone mRNAs are oligouridylated which promotes swift degradation of these mRNAs (Mullen and Marzluff 2008). Other studies support a general role for oligouridylation in stimulating RNA decapping and in turning over polyadenylated mRNA (Song and Kiledjian 2007; Rissland and Norbury 2009).

The large and ancient family of miRNA genes is also regulated by non-templated addition of uridines. These small RNAs are protected from degradation in plants by modification of their 3'-ends by 2'-O-methylation of the terminal nucleotide (Yu et al. 2005). Loss of HEN1, the RNA methyltransferase that acts on miRNAs in *A. thaliana*, results in oligouridylation of miRNAs and a reduction in miRNA abundance (Li et al. 2005). Mammalian miRNAs are not methylated which allows miRNAs in mammals to be frequently modified post-transcriptionally, usually with additional adenines and uridines (Landgraf et al.

2007). Nucleotide addition to miRNAs may not always be destabilizing. In the case of mir-122, the addition of one non-templated adenine seems to be required for expression of that miRNA (Katoh et al. 2009). Another class of small RNAs called piwi-interacting RNAs (piRNAs) in metazoans is methylated on the 2'-OH of the terminal nucleotide by the homologue of the plant HEN1 gene (Horwich et al. 2007; Kirino and Mourelatos 2007a; Saito et al. 2007). Loss of 2'-O-methylation results in a reduction in piRNAs *in vitro* suggesting a conserved protective function of RNA-methylation (Kirino and Mourelatos 2007b). Non-templated nucleotide addition is an ancient regulatory motif extending beyond plants and metazoans to unicellular algae. Work in the *C. reinhardtii* model system shows an increase in abundance of miRNAs and other small interfering RNAs (siRNAs) when the terminal nucleotidyltransferase MUT68 is deleted (Ibrahim et al. 2010).

Expression of the let-7 miRNA is regulated in embryonic stem cells by non-templated oligouridylation of pre-let-7 (Heo et al. 2009). Lin-28 recognizes a conserved motif in the stem-loop of pre-let-7 and recruits the terminal uridylyltransferase (TUTase) Zcchc11 which then polymerizes an oligouridine tail on pre-let7 (Heo et al. 2009) preventing miRNA expression by stimulating degradation and blocking processing by Dicer. Lin-28 is a gene identified in worms because mutations in this gene results in defects in developmental timing or lineage (Moss et al. 1997). Despite the great separation between worms and mammals, misregulation of Lin-28 results in changes in developmental timing in mouse and humans (Ong et al. 2009; Zhu et al. 2010).

Non-templated oligouridylation is associated with another aspect of small RNA biology. The RNA-Induced Silencing Complex (RISC) uses small RNAs as a guide to selectively target RNAs for destruction (Martinez et al. 2002). When Ago2 finds extensive complementary base-paring between the small RNA guide and the target RNA, it slices the target RNA between the 10th and 11th nucleotide measured from the 5'-end of the small RNA guide (Elbashir et al. 2001). Slicing of the target generates a new 3'-OH which is subsequently oligouridylated. The addition of non-templated uridines to the sliced RNA is seen in plants and mammals and seems to promote 5' → 3' degradation (Shen and Goodman 2004).

Presented here is the work undertaken to study this recently identified post-transcriptional RNA modification. This dissertation describes the development of methods to deeply sequence native RNA 3'-ends; the development of a method to analyze RNA-Seq data to identify non-templated nucleotide addition and classify RNA-Seq reads and; the development of an RNA-Seq method to selectively enrich for and sequence RNAs that end in 3'-uridines.

Chapter 2: Widespread RNA 3'-end Oligouridylation in Mammals

This chapter includes material from an accepted manuscript (Choi *et al.* RNA 2011).

Abstract

Non-templated 3'-end oligouridylation of RNA occurs in many species, including humans. Unlike the familiar phenomenon of polyadenylation, non-templated addition of uridines to RNA is poorly characterized in higher eukaryotes. Recent studies have reported non-templated 3'-end oligouridylation of small RNAs and mRNAs. Oligouridylation is involved in many aspects of microRNA biology from biogenesis to turnover of the mature species and it may also mark long mRNAs for degradation by promoting decapping of the protective 5'-cap structure. To determine the prevalence of oligouridylation in higher eukaryotes, we used next generation sequencing technology to deeply examine the population of small RNAs in human cells. Our data revealed widespread non-templated nucleotide addition to the 3'-ends of many classes of RNA, with short stretches of uridine being the most frequently added nucleotide.

Introduction

RNA can undergo several modifications between transcription and eventual degradation. Well known examples of post-transcriptional RNA processing such as 5'-cap addition and polyadenylation of mRNAs and 3'-CCA addition to tRNAs illustrate the critical importance of RNA modifications in gene expression (Martin and Keller 2007; Moore and Proudfoot 2009). Another form of processing is the generation of a mature RNA from a longer precursor transcript. Many pre-mRNAs require removal of introns and primary transcripts of rRNAs and tRNAs are processed by endo- and exonucleases to release the mature RNA (Granneman and Baserga 2005; Phizicky and Hopper 2010). Maturation of other non-coding RNAs such as miRNAs and snoRNAs also requires processing by endo and exonucleases (Filipowicz and Pogacic 2002; Olena and Patton 2010).

Altering RNA 3'-ends can help stabilize or degrade RNA. The 3'-end of U6 snRNA is stabilized after the addition of non-templated uridines by 2'-3' cyclization of the final nucleotide (Lund and Dahlberg 1992). The destabilizing effect of uridine addition to mRNAs has been observed in yeast and mammalian cells and has been proposed to function by stimulating removal of the 5'-cap and degradation of mRNA (Song and Kiledjian 2007; Rissland and Norbury 2009). After DNA replication is complete, degradation of histone mRNA is initiated by non-templated oligouridylation to coordinate histone expression with DNA abundance (Mullen and Marzluff 2008).

Oligouridylation is associated with miRNA biogenesis, function and turnover. The addition of uridines to *let-7* precursors has been reported to regulate the expression of *let-7* in embryonic stem cells (Heo et al. 2009). Mature miRNAs are often sequenced with non-templated uridine or adenines (Landgraf et al. 2007). These non-templated bases may mark miRNAs to promote their decay (Ameres et al. 2010). Furthermore, small RNAs with extensive sequence complementarity to a target RNA can guide Ago2 to cleave, or slice, the target RNA (Liu et al. 2004; Meister et al. 2004). Sequencing the 5'-fragment of the sliced RNA often shows non-templated uridines added to the cleavage site (Shen and Goodman 2004).

In vitro studies suggest that uridylyltransferases may act on many RNAs in human cells (Sinha et al. 1998). To determine the extent of non-templated 3'-end oligouridylation in mammals, we used next generation sequencing with strand-specific RNA linker-ligation to directly examine the 3'-ends of small RNAs (<200 nt) in human cells. Examples of previously identified RNA processing steps and non-templated 3'-end nucleotide addition are present. We also find many novel cases of non-templated 3'-oligouridylation on several classes of RNA including transcriptional start site-associated RNAs (TSSa-RNAs) and spliced introns suggesting that oligouridylation may play a larger role in RNA metabolism in mammals.

Results

Identification of non-templated 3'-end nucleotide addition to RNA

Because of the growing body of literature showing that non-templated 3'-end oligouridylation can regulate RNA metabolism in mammals, we used next generation sequencing to better define the scope of this RNA modification. Small RNAs (<200 nt) of the optimal size for deep sequencing were purified without shearing to maintain native 3'-ends. While this method excludes long RNAs, it also avoids ribosomal RNAs, the major fraction of total RNA. Sequential ligation of adaptors maintains RNA strand information and is specific for a subset of the total RNA population: small RNAs with 5'-phosphate and 3'-OH, the functional groups that result from many endo- and exonucleases.

Illumina paired-end sequencing was used to directly examine the 3'-end of RNA. Because the paired-end 2 (PE2) primer begins sequencing 3' → 5' relative to the sense orientation, this dataset provides high quality reads starting at the 3'-end of captured RNA. Post-transcriptional processing of RNA poses a challenge for aligning RNA-Seq data because non-templated nucleotides are likely to disagree with the genomic sequence at the aligned loci. Mismatches between the read and reference genome are penalized with some alignment programs having a maximum cutoff for number of mismatches allowed before rejecting a read. To compensate for homopolymeric, non-templated nucleotide addition to RNA 3'-ends, all initial homopolymer clusters for all four nucleotides, regardless of length, were removed from the start of sequences in the PE2 dataset. These initial clusters represent the 3'-end terminal homopolymer (THP). Alignment after removal of the THP resulted in more aligned reads and more usable reads that aligned to unique loci (Table 1). The mismatch rate in the THP was more than 10-fold greater than the mismatch rate in the following 20 nucleotides (Table 2). Stripping the THP allowed read mapping using bases that were more likely to match the reference genome.

A broad overview shows that out of 16.3 million reads that aligned to unique genomic loci, nearly 20% (3.2 million) had at least one mismatch in their 3'-end (Figure 1A). Partitioning reads by gene class found non-templated nucleotides on many classes of RNA, often at high frequency (Figure 1B-F). The gene family that made up the largest fraction of reads (73%) with non-templated 3'-end nucleotides was miRNAs (Figure 1A). The mir-302-367 cluster produces an embryonic stem cell-specific family of miRNAs identified by traditional and next generation sequencing of small RNAs (Suh et al. 2004; Morin et al. 2008). One member of this cluster, *hsa-mir-302c,* accounted for the vast majority of miRNA reads with non-templated 3'-ends. Sequencing H9 genomic DNA of the mir-302-367 cluster confirmed that the cell line agrees with the reference genome and that the high frequency and abundance of *mi*R-*302c* +U was not encoded in the DNA (data not shown). Examination of other miRNAs showed that our data is consistent with previous studies that have reported frequent uridine or adenine addition to miRNAs (Landgraf et al. 2007) (Figure 1B).

Uridines and adenines accounted for nearly all cases of non-templated nucleotides for most gene classes (Figure 1C). We identified known examples of non-templated RNA 3'-end addition such as the nearly universal post-transcriptional 3'-CCA addition to tRNAs (Figure 1D). Less than 10% of reads aligned to tRNA genes, much less than their contribution to the small RNA mass in cells. This underrepresentation may be due to aminoacylation on either 3' or 2' hydroxyl which would inhibit adapter ligation by T4 RNL2truc (Munafo and Robb 2010). Essentially all reads with unambiguous, non-templated

3'-CCA or 3'-CA aligned to tRNA genes (Table 3). Also noted were rare cases of non-templated uridine or adenine addition to mitochondrial tRNAs and U or UU addition to nuclear encoded tRNAs following the 3'-CCA modification. Reads with more extensive non-templated adenylation (>10 nt) were usually found on mitochondrial tRNAs with much fewer cases of polyadenylated nuclear-encoded tRNAs (Supplementary figure 1).

Another example of previously identified RNA 3'-end nucleotide addition found in our dataset was the oligouridylation of U6 snRNA. The low read counts for U6 and for snRNAs as a group (Figure 1E) likely demonstrates the selection of the cloning protocol. Before snRNAs can be captured, they must be processed to remove their 5'-cap or 5'-tri-phosphate. U6 needs to be doubly processed to remove the 2'-3' cyclic UMP found on the mature form of U6. By contrast, snoRNA genes had the most overall reads but were least likely to have non-templated nucleotide addition (Figure 1F). Unlike snRNAs, mature snoRNAs have been processed by nucleases (Filipowicz and Pogacic 2002) and already possess the functional groups selected for by the cloning protocol.

Transcriptional start-site associated RNAs are oligouridylated

After confirming the presence of known cases of non-templated 3'-end nucleotide addition in our dataset, we searched for new examples of this modification, particularly from RNAs derived from long transcripts. Although the RNA purification protocol enriched for small RNAs, approximately 400,000 total reads mapped to long coding and non-coding genes. This relatively small population contributed significantly to the diversity of unique sequences with non-templated nucleotides. The overall population of 3.2 million reads with non-templated RNA 3'-ends (Figure 1A) collapses to 82,176 unique sequences with 24,079 derived from long RNA transcripts. The cumulative frequency of 3'-end THP lengths shows non-templated uridines were the most frequent base for tails up to five nucleotides (Figure 2). Because different types of RNA processing occurs on RNAPol-II transcripts, reads that aligned to the UCSC refGene database were grouped on the basis of where the 3'-end aligned with respect to the gene annotation (Table 3).

Recent studies in mouse and human embryonic stem cells have described TSSa-RNAs, short RNAs that map near transcriptional start-sites (Guenther et al. 2007; Seila et al. 2008). These short RNAs derived from transcripts that have aborted or paused after initiation have been postulated to result from regulation of RNAPol-II elongation and escape from the promoter. TSSa-RNAs are present at low copy numbers per cell per gene (Guenther et al. 2007; Seila et al. 2008); 84,298 sense and 35,284 antisense TSSa-RNAs (Table 1) were identified in our dataset. The read length distribution of TSSa-RNAs in our RNA-Seq library made from H9 human embryonic stem cells shows a broad peak from 22 – 50 nt with 6.9% of sense and 5.4% of antisense TSSa-RNAs extending the full 76 cycles of sequencing (Figure 3A). The distribution of reads with respects to their distance from the TSS is similar to earlier reports describing this class of RNA (Figure 3B). Most of the sense TSSa-RNAs that aligned upstream of the TSS are due to incomplete annotations that lack the 5'-UTR. Non-templated nucleotides were found on 20% of TSSa-RNAs with uridines being six-fold more abundant than adenine homopolymers for both sense (Figure 3C) and antisense orientations (Figure 3D). The position of reads with non-templated uridines is similar to the distribution for all sense TSSa-RNAs (Figure 3B). It should be noted that not all TSSa-RNAs would be detected by the cloning method used to prepare our RNA-Seq library. The initial 5'-tri-phosphate or added 5'-cap would have to be removed, an activity which is promoted by the addition of non-templated uridines (Song and Kiledjian 2007).

Spliced introns are oligouridylated

Of all categories of sequences that aligned to long genes in the UCSC refGene database, splice site sequences were the RNA most likely to have non-templated 3'-end nucleotides (Table 1). They were also the rarest population with 5,600 total reads. Most genes had fewer than five reads with only two intron splice sites covered by more than 100 reads (Supplemental Figure 1). By contrast, the much larger population of other intronic reads had the lowest frequency of non-templated 3'-end nucleotide addition (Table 1). After intron splicing, the lariat must be processed before it can captured for sequencing. Analysis of the lengths of splice site reads reveals a peak for short reads, ~25% were ≤26 nt, and a broad population that extend past the branch point with 23% of all splice site reads and 38% of splice site reads with non-templated THPs spanning the full 76 cycles of PE2 sequencing (Figure 4A). Uridines were the most frequently found non-templated nucleotide in this population of processed RNA, 3 to 15-fold more abundant than non-templated adenines depending on the length of the homopolymer (Figure 4B).

The fraction of spliced introns with non-templated nucleotides may be larger than what our analysis reports. Sequence analysis can only identify mismatches between the read and the reference genome. It cannot distinguish between templated transcription and non-templated nucleotide addition after RNA processing if the non-templated base happens to match the genome. Alignment of sequences to splice sites with relatively high read coverage illustrates this problem. Ninety-six reads that mapped to the 3'-splice site of intron 7 of *TATDN1* included an extra adenine that could be templated and the result of mis-splicing or it could be due to non-templated nucleotide addition after correct splicing (Figure 5). Similarly, the extra uridine found on nine reads of an intron of *HSPG2* could either be due to templated transcription or added following splicing. Non-templated RNA 3'-end modification would be underestimated on other classes of RNAs as well by the requirement for unambiguous mismatches in the alignment to the reference genome.

Discussion

Deep sequencing of small RNA from human embryonic stem cells captured many previously identified examples of post-transcriptional RNA 3'-end processing. We also found many new examples of non-templated nucleotide addition with oligouridylation being the most frequently observed modification. The abundance of non-templated uridines in our dataset cannot be simply explained by the enzymes used in preparing the RNA-Seq library. The non-templated activity of Taq DNA polymerase adds to the end of the DNA template and there are no reports of this enzyme being prone to insert multiple nucleotides in the middle of a template. The error rate of DNA polymerase also does not account for the clearly higher rate of mismatches to the reference genome in the THP compared to the adjacent 20 nucleotides even after excluding the first PE2 cluster (Supplementary Table 2). T4 RNA ligase 2-truncated, the enzyme used to add the 3'-adaptor, is catalytically impaired and requires an activated, pre-adenylated DNA or RNA oligomer as a substrate (Ho et al. 2004). It is unlikely to act as a polymerase to add free nucleotides to RNA. Extra uridines are not likely introduced by the 3'-adaptor which was PAGE purified and begins: 5'-AGATCG. RNA ligases are known to have different ligation efficiencies depending on the sequence of the donor and acceptor molecules (England and Uhlenbeck 1978). Even if this enzyme was particularly efficient at ligating RNA with 3'-uridines, that would not explain why the uridines are so often mismatched to the reference genome.

Our RNA-Seq preparation and analysis method may explain the high rate of 3'-end mismatches we observed in our dataset and why other methods may exclude this RNA modification. The common practice of excluding reads with mismatches to the reference genome, especially when aligning short sequences, would be blind to non-templated nucleotides. Previous studies of TSSa-RNAs used DNA microarrays (Guenther et al. 2007), which do not detect sequence variants of the probe. Others used short sequencing in the sense orientation (Seila et al. 2008) that excluded reads with mismatches and would not have reached the 3'-ends of most TSSa-RNAs (Figure 3A). RNA-Seq studies of long coding transcripts often use polyA selection to enrich for mRNAs from total RNA thereby excluding spliced introns.

The RNA-Seq library preparation method that we used has limitations that must be considered when interpreting the data. The low total count of reads that aligned to 3'-splice sites may be caused in part by the size fractionation of RNA during purification. Half of splice site reads aligned to short introns <300 nt with the other half originating from introns up to 129 kb long. The scarcity of reads is also seen in the low read coverage of most introns (<5 reads). It was rare for more than one intron of a single transcript to be detected, usually in genes with several short introns. While it would be premature to draw conclusions about specific splice sites with such low read coverage, the scarcity of splice site reads does not address why these processed RNAs as a whole were so often sequenced with non-templated uridines.

Though we highlighted the non-templated oligouridylation of TSSa-RNAs and splice site-derived RNAs, it should be noted that sequences derived from longer mRNA or pre-mRNA transcripts were in the “other” category of refGene reads (Table 1). These reads included fragments of 5'-UTR, exon or exon with upstream intron. As a group, these reads had a similar rate of non-templated nucleotide addition as TSSa-RNAs which may allude to prior nuclease processing of these small RNA fragments necessary to be in our library.

The low rate of non-templated nucleotide addition to snoRNAs may provide a clue as to the relative abundance of modified RNAs compared to the mature form. Unlike RNAs that are protected at the 5'-end, mature snoRNAs can be captured without further processing which could explain why modified snoRNAs were rarely seen against the background of mature transcripts. The largest fraction of reads that aligned to long RNA genes was derived from intronic sequences more than 10 nt from a splice sites. These reads had a low rate of 3'-end non-templated nucleotide addition (Table 4). Two loci made up more than 50% of this category of reads (Figure 6). These reads sequenced with high frequency could simply be RNA that formed nuclease-resistant secondary structures. However, three of these abundantly sequenced genes have many characteristic of snoRNAs. They mapped to a portion of intron with high species conservation; the mapped region was found in other EST databases; the RNA could be folded into stable hairpins; and known snoRNAs with sequence similarity were in other introns of the same host gene. This suggests that many of our intronic refGene sequences could be derived from processing of unannotated snoRNAs or snoRNA pseudogenes (Figure 7).

Though most human snoRNAs are processed from introns and have 5'-monophosphates (Dieci et al. 2009), there are a few snoRNAs that are known to be 5'-capped. Many studies on U3, U8 and U14 have examined the changes in methylation status of their caps and subcellular localization of these RNAs (Reddy et al. 1985; Tyc and Steitz 1989; Terns et al. 1995; Jacobson and Pederson 1998; Speckmann et al. 2000). The low read counts for U3 (18) and U14 (2,270 U14A and 13 U14B) are consistent with the requirement for RNAs to have 5'-phosphates in ordered to be captured by the RNA-Seq cloning method used to prepare our library. However, U8 is the most frequently sequenced gene in our library with 2,768,660 reads aligned to that locus. Reads to U8 also had a low rate of non-templated nucleotide addition (0.93%). Because almost all reads to U8 (98.4%) in the PE2 library extended the full 76 cycles of sequencing we searched the paired-end 1 sequencing primer dataset to identify the 5'-ends of U8. Over 80% of reads had the annotated 5'-end sequence of U8 and were not cleaved fragments indicating that there was a significant population of U8 snoRNAs with 5'-monophosphates. This poses several questions about the processing of U8 in human embryonic stem cells. Examination of the U8 locus shows that this snoRNA is not intronic but it overlaps the 3'-UTR of *TMEM107*. One possibility is that in some cases, the cap may be added on after processing the snoRNA out of the 3'-UTR of *TMEM107* and not due to independent transcription of U8. There were 47 reads to *TMEM107* in our dataset: three TSSa-RNAs, one mRNA fragment including part of intron 2 + exon 3, with all other reads aligned to the 3'-UTR downstream of U8. Other questions about the rates of capping and decapping of snoRNAs are beyond the scope of this study but we have made available the PE1 sequencing primer dataset for those interested in human small RNAs.

While we must be cautious about inferring the *in vivo* abundance of an RNA from sequencing read counts (Linsen et al. 2009), the different rates at which non-templated uridines are observed on different classes of RNAs may hint at a broader functional role for non-templated uridine addition in RNA metabolism in mammals. There is precedence for the addition of non-templated nucleotides to mark RNA for degradation. In contrast to the stabilizing effect of long polyadenylation of eukaryotic mRNA, polyadenylation in prokaryotes destabilizes RNA (Arraiano et al. 2010). Polyadenylation can destabilize eukaryotic RNA as well (West et al. 2006). The TRAMP complex (Trf4/Air2/Mtr4p) adds poly-A tails to improperly processed nuclear RNAs to promote their degradation by the nuclear exosome (LaCava et al. 2005; Vanacova et al. 2005). Studies have identified and characterized a conserved family of nucleotidyl transferases present in many eukaryotes including yeast, plants and humans (Kwak and Wickens 2007; Rissland et al. 2007). These enzymes, called poly(U) polymerases (PUPs) because they have a strong preference *in vitro* to incorporate uridines on the 3'-ends of RNA, are related to the Trf4 and Trf5 components of the TRAMP complex. It is has already been shown that oligouridylation can regulate diverse RNA species. Studies in plant and animal models show that small noncoding RNAs can be regulated by oligouridylation (Li et al. 2005; Heo et al. 2009; Ameres et al. 2010). Longer mRNAs can also be marked for degradation by the addition of uridines (Mullen and Marzluff 2008). We find it interesting that the highest frequency of non-templated 3'-end oligouridylation is found on RNAs that required processing by nucleases to be captured by our cloning method. Our analysis shows post-transcriptional oligouridylation of RNA is widespread and found on many classes of RNA genes. These observations suggest that oligouridylation may be involved in several aspects of RNA metabolism and opens new questions for further study. Further refinements of RNA-Seq methods may be required to identify more examples of RNA3'-end oligouridylation.

Table 1. Alignment with and without terminal homopolymer

Input 31,824,185 sequences	Complete sequence	-Terminal homopolymer
Aligned	19,612,301	21,352,356
Unique alignment	14,817,548	16,527,417
Gapped Alignment	1,343,686	1,003,231
Quality Filter	7,335,647	350,7262
Homopolymer Filter	29,085	869
Alignments to unique loci –N		16,429,038
Alignments to unique loci –N, –mm9		16,304,574
Collapsed to unique sequences		742,711

Table 1. Alignment with and without terminal homopolymer.

The number of reads that aligned to the reference genome using the complete sequence of each read compared to the number that aligned after removal of the THP as reported by novoalign. The remaining number of reads with unique alignments after excluding clusters with ambiguous base calls (-N) and after excluding potential mouse RNA contamination (-N, -mm9).

Table 2. Mismatch rate in THP vs. next 20 bases (Bases/Mismatch)

	All initial PE2 clusters	Excluding first PE2 cluster
THP mismatch rate	59.04	68.56
Next 20 bases	711.96	632.28

Table 2. Mismatch rate in terminal homopolymer vs. next 20 bases.

The rate at which the alignment did not agree with the reference genome was calculated for the THP and for the next 20 bases following the THP. In case the THP was longer than one nucleotide, the mismatch rate was calculated by dividing the total number of nucleotides in THPs by the total number of mismatched bases in THPs. The calculation was repeated after excluding the first cycle of PE2 sequencing**.**

Table 3. Genes with non-templated –CCA addition

	-CCA	-CA	Total Reads
tRNAs	68,700	105,757	1,398,306
Mitochondrial	68	111	279,818
No Annotation	342	347	212,992
Pseudogene	10	19	729,045
U75	8	12	124,949
U5F	7	7	29,932
U26	7	7	119,963
U8	4	4	2,793,760
U2	2	20	18,344
U30		25	124,049
U31		20	41,409
U5Ds		3	317,435
SUPT4H1		3	13
DCAKD		2	11
AKIRIN2		2	10

Table 3. Genes with non-templated –CCA addition.

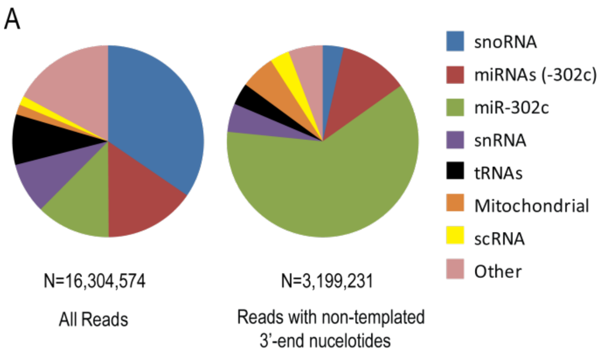
Non-templated 3'-CCA or –CA addition tabulated by gene class. Listed are the read counts for non-tRNA genes found with these modifications. Mitochondrial reads are non-tRNA genes. Also shown are the total number of reads in the library for each category or gene.

Table 4. Alignment of refGene annotated genes.

	TSS	3'-UTR	Splice Site	Intron	All Antisense	Antisense <1000 bp from TSS	Other
All reads	84298	12616	5600	177158	97865	35284	15882
All unique reads	38256	4739	2460	37382	36076	16614	7412
Genes	9075	2218	1175	7476	10189	6129	2768
Reads with MM	14614	2299	2080	14092	15847	5127	2964
Unique MM Reads	7503	1030	994	6016	6981	2943	1555
Genes	2897	629	464	2621	3314	1756	846
MM Reads (%)	17.3	18.2	37.1	8.0	16.2	14.5	18.7
MM Unique Reads (%)	19.6	21.7	40.4	16.1	19.4	17.7	21.0

Table 4. Alignment of non-small RNA refGene reads.

Reads that mapped to the UCSC refGene database that did not align to small RNA genes were categorized according to the position of the 3'-end of the read in the gene. Sense reads that were no more than 200 nt downstream or antisense reads less than 1000 bp from a TSS were considered TSSa-RNAs. 3'-UTR reads were those that ended in a 3'-UTR. A special category of intron reads were those that mapped within 10 nt of an annotated intron 3'-end which were called “splice site”. “Other” reads were mostly derived from exons and 5'-UTRs. If there were more than one annotation of a gene overlapping the read position, the annotation with the most exons was used to categorize the read. Unique reads are counts after removing duplicate sequences.



	% of All Reads	# of MM Reads	Non-templated Nucleotide				Gene Name
			A	C	G	U	
B	96.11	1965434	0.00	0.04	0.00	0.96	mir-302c
	14.83	371347	0.44	0.08	0.03	0.46	miRNAs (-302c)
	5.44	33907	0.47	0.03	0.11	0.39	hsa-mir-302a
	25.86	118036	0.28	0.02	0.01	0.70	hsa-mir-302b
	16.57	80878	0.45	0.22	0.03	0.30	hsa-mir-302d
	18.92	31078	0.82	0.01	0.01	0.17	hsa-mir-367
	19.22	16576	0.70	0.02	0.01	0.27	hsa-mir-20a
	86.66	6702	0.00	0.00	0.00	0.99	hsa-mir-301a
	15.75	10788	0.86	0.11	0.01	0.02	hsa-mir-130a
	15.48	9683	0.84	0.03	0.06	0.07	hsa-mir-21
C	66.92	186312	0.90	0.07	0.02	0.01	Mitochondrial
	10.95	153583	0.10	0.16	0.07	0.67	snRNA
	8.51	118038	0.63	0.08	0.03	0.25	tRNAs
	2.02	113621	0.37	0.30	0.14	0.20	snoRNA
	42.20	104871	0.60	0.01	0.02	0.37	scRNA
	12.33	26152	0.23	0.08	0.04	0.65	No Annotation
	17.04	19699	0.23	0.06	0.03	0.68	Antisense
D	92.47	93304	0.90	0.09	0.02	0.00	Mitochondrial tRNA-Pro
	21.87	8950	0.88	0.06	0.03	0.03	Mitochondrial tRNA-Leu
	4.64	17553	0.22	0.09	0.03	0.66	tRNALys-TTT
	6.55	8411	0.53	0.08	0.03	0.36	tRNAGly-CCC
	10.79	7184	0.79	0.11	0.03	0.07	tRNAAsp-GTC
	39.07	25567	0.96	0.01	0.02	0.01	tRNAGlu-TTC
	19.94	9697	0.90	0.03	0.01	0.06	tRNALys-CTT
E	44.57	82	0.44	0.12	0.05	0.39	U1
	30.07	5435	0.63	0.03	0.02	0.32	U2
	13.24	1802	0.28	0.04	0.06	0.62	U4
	23.23	73102	0.01	0.09	0.04	0.85	U5Ds
	2.92	9053	0.32	0.33	0.10	0.25	U5E
	35.54	145	0.02	0.02	0.03	0.93	U6
	6.69	275	0.24	0.05	0.03	0.68	U4atac
	24.32	45	0.31	0.11	0.02	0.56	RNU6ATAC
	9.34	45053	0.05	0.26	0.13	0.57	U7
	4.60	7136	0.53	0.23	0.12	0.12	U11
	12.06	1120	0.14	0.10	0.04	0.73	U12
	48.60	519	0.67	0.03	0.01	0.29	7SK_RNA
F	0.93	25879	0.05	0.79	0.12	0.03	U8
	10.19	15449	0.92	0.02	0.01	0.05	U74
	2.95	11862	0.06	0.43	0.41	0.10	SNORD2
	9.72	8461	0.13	0.01	0.09	0.77	U13
	18.05	6212	0.89	0.00	0.08	0.03	U101
	3.88	1515	0.05	0.48	0.40	0.06	SNORD18B
	3.18	3845	0.05	0.47	0.40	0.08	U26
	3.95	4704	0.36	0.04	0.03	0.57	U44
	2.47	3825	0.02	0.61	0.33	0.03	SNORD100

0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

Figure 1. The non-templated 3'-ends of small RNAs in human ES cells. (A) The distribution of paired-end-2 sequencing primer reads by RNA gene class. The distribution for all uniquely aligned reads (left) and those reads with non-templated 3'-end THP (right) are shown. (B-F) The frequency at which each nucleotide was found as the non-templated THP was calculated for each RNA gene class. The first column is the percent of reads for a specific gene or gene class that had non-templated 3'-ends. The numbers in the (MM) column are the number of reads with mismatches between the THP and the reference genome. (C) The overall frequency of non-templated 3'-ends for several classes of RNA genes. Shown are the most highly sequenced genes of miRNAs (B), tRNA anti-codons (D), snRNAs (E) and snoRNAs (F)

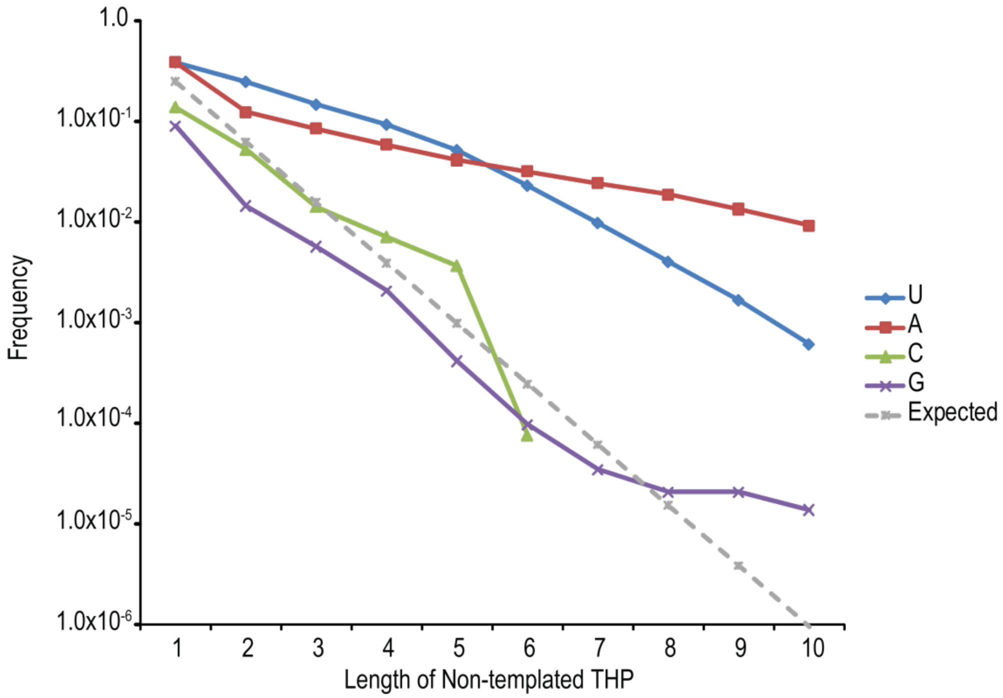


Figure 2. Distribution of non-templated THP lengths.

The THP lengths were counted for reads with non-templated 3'-end nucleotides. 3.2 million reads covering 11,616 genes were collapsed to 82,176 unique sequences. The cumulative frequency of THP lengths of this collapsed dataset is shown, that is if an RNA had a THP of five uridines, it was included in the totals for lengths of one through five. The expected line is the probability of finding a homopolymer of a given length at random based on equal representation of all nucleotides.

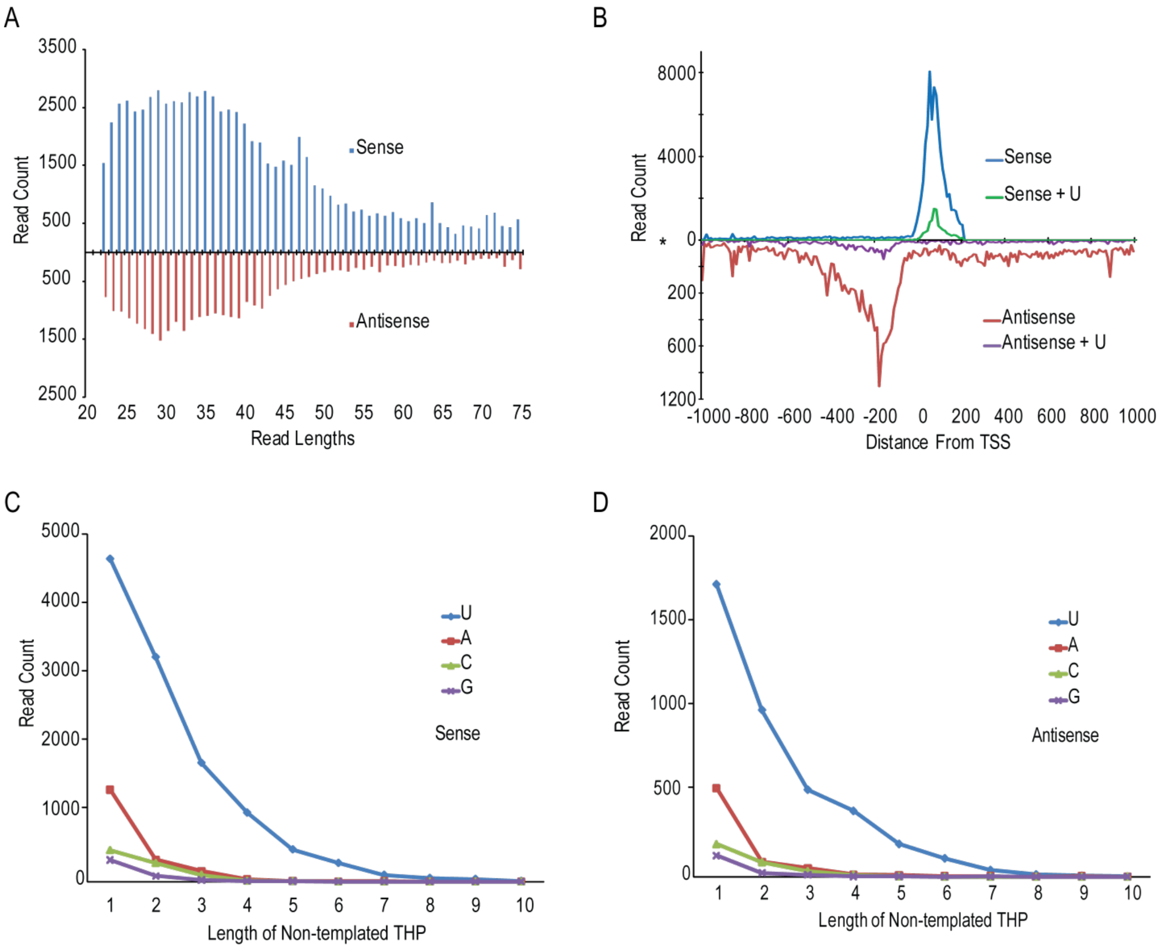


Figure 3. Characteristics of TSSa-RNA.

(A) The length distribution of TSSa-RNAs. (B) The distribution of distances of TSSa-RNAs measured from the TSS to the 3'-end of the read for all TSSa-RNAs and TSSa-RNAs with non-templated uridine addition. * Note the different scales for sense and antisense. The distribution of non-templated 3'-end THP lengths for sense (C) and antisense (D) TSSa-RNAs.

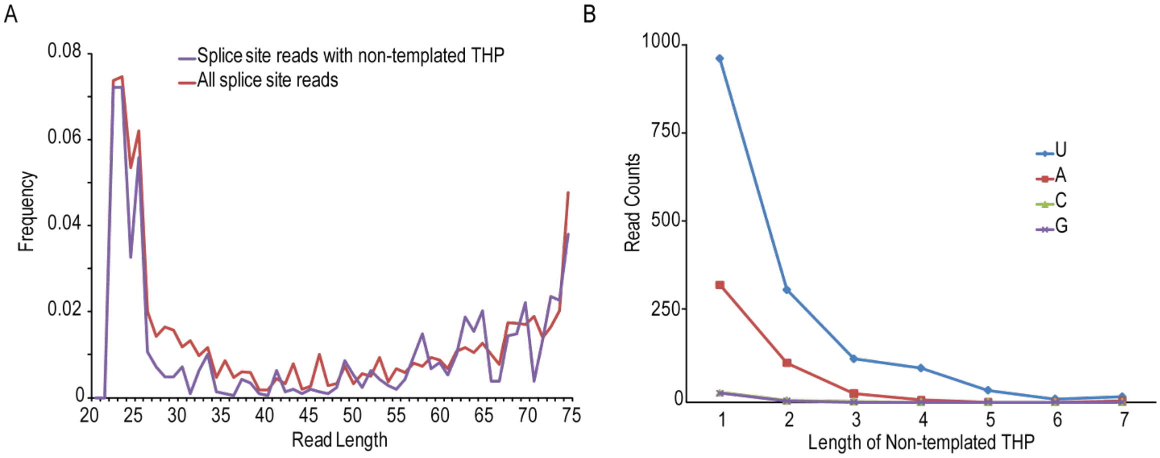


Figure 4. Characteristics of spliced introns.

(A) The distribution of lengths of all splice site reads and splice site reads with non-templated 3'-end THPs plotted as the fraction of their respective populations. (B) The distribution of non-templated 3'-end THP lengths sequenced on splice site reads.

```
                                                                              RPL34 (171)
…taagtatttctgaaaattttaagtatatattgtcatttactctacaaaatgctgacctactgactgtttcactttctagGTCCC…
                                                        CCUACUGACUGUUUCACUUUCUAGA       (47)
                                                       ACCUACUGACUGUUUCACUUUCUAGU        (3)
                                                       ACCUACUGACUGUUUCACUUUCUAGA        (7)
                                                AAUGCUGACCUACUGACUGUUUCACUUUCUAGA        (5)
                                                        CCUACUGACUGUUUCACUUUCUAG        (54)
                                                       ACCUACUGACUGUUUCACUUUCUAG        (48)


                                                                            TATDN1 (126)
…aaccaaagtctcattcagaacttaagaattttgtagaaatcaagctatttgctaaaagttctttgttttaattcacagACATA…
                                                           UUCUUUGUUUUAAUUCACAGU        (20)
                                                           UUCUUUGUUUUAAUUCACAGA        (96)
                                                           UUCUUUGUUUUAAUUCACAGAA        (1)
                                                          GUUCUUUGUUUUAAUUCACAG          (6)
   CCAAAGUCUCAUUCAGAACUUAAGAAUUUUGUAGAAAUCAAGCUAUUUGCUAAAAGUUCUUUGUUUUAAUUCACA           (2)


                                                                              HSPG2 (23)
…cagggagtggggatggaggaagtctgtcttctggggtctctgagccccacccatgaccccctttcgccttgccctgcagTCCC…
     GAGUGGGGAUGGAGGAAGUCUGUCUUCUGGGGUCUCUGAGCCCCACCCAUGACCCCCUUUCGCCUUGCCCUGCAGU        (9)
     GAGUGGGGAUGGAGGAAGUCUGUCUUCUGGGGUCUCUGAGCCCCACCCAUGACCCCCUUUCGCCUUGCCCUGCAGUUU      (5)
```

Figure 5. Alignments of splice site reads.

Examples of splice site reads that aligned to introns with high coverage. The genomic sequence is above the grey line with the intron in lower case and the exon marked by the box and upper case. The number of total reads to the gene is in parenthesis as is the read count for unique sequences. Sequences were converted to sense orientation for easier comparison to the reference genome. Unambiguous mismatches are in red.

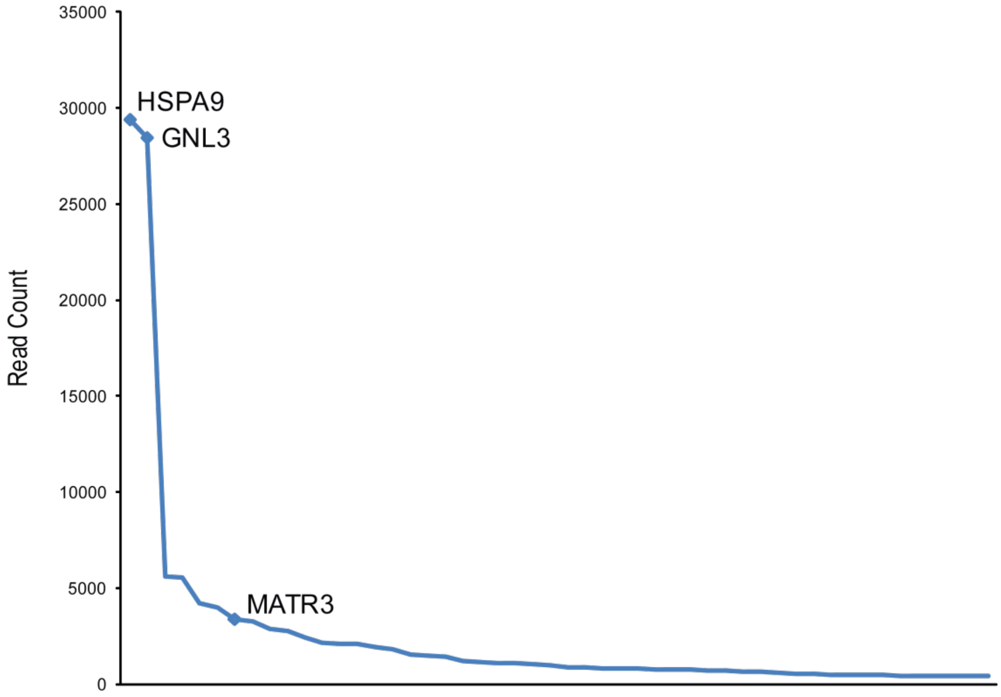


Figure 6. Distribution of intronic reads by gene.

The distribution of refGene reads that aligned to introns more than 10 nt from a splice site plotted by rank order of total reads per gene.

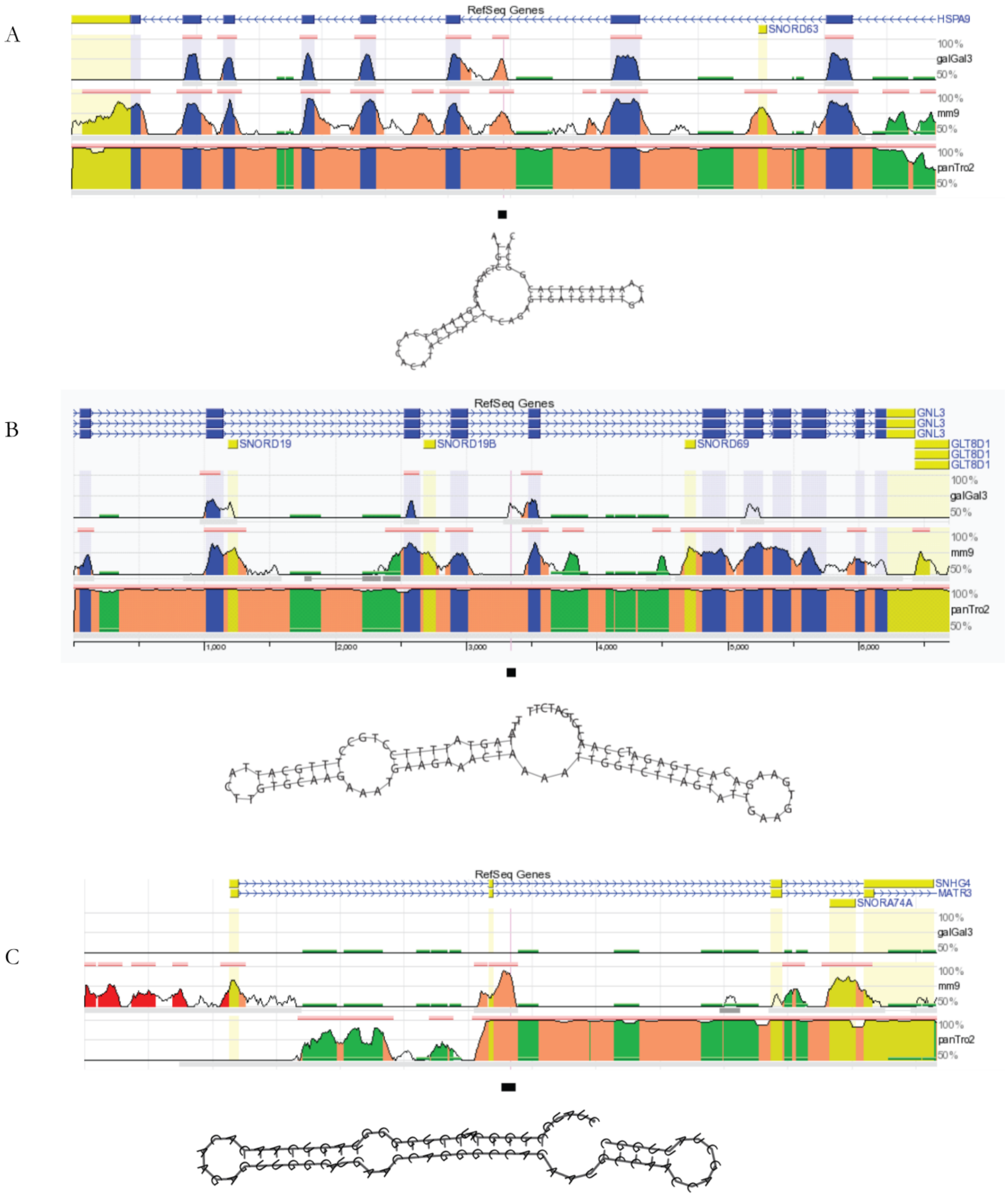
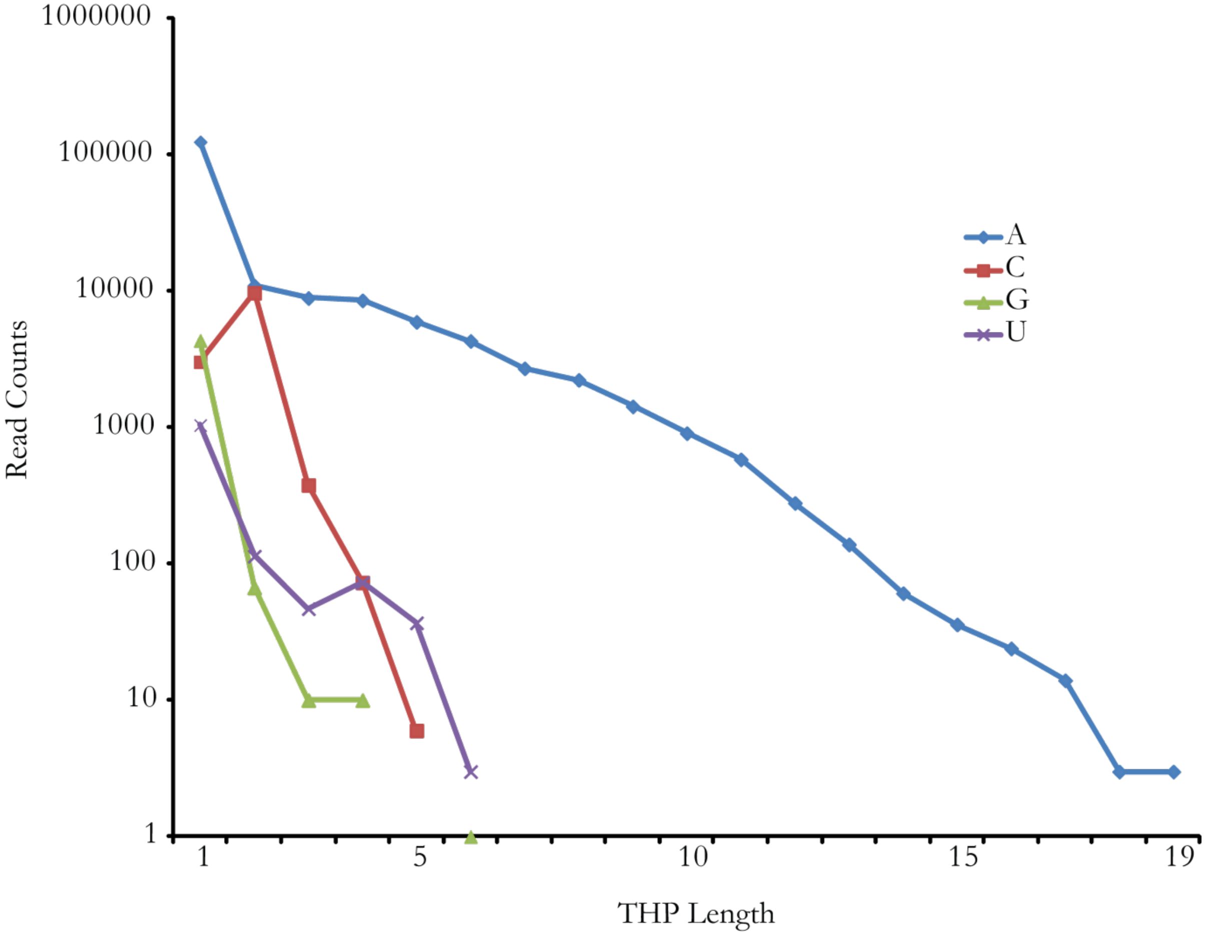

A
RefSeq Genes
HSPA9
SNORD63
100%
galGal3
50%
100%
mm9
50%
100%
panTro2
50%
B
RefSeq Genes
GNL3
GNL3
GNL3
SNORD19
SNORD19B
SNORD69
GLT8D1
GLT8D1
GLT8D1
100%
galGal3
50%
100%
mm9
50%
100%
panTro2
50%
1,000
2,000
3,000
4,000
5,000
6,000
C
RefSeq Genes
SNHG4
MATR3
SNORA74A
100%
galGal3
50%
100%
mm9
50%
100%
panTro2
50%

Figure 7. Identification of potential new snoRNA or snoRNA pseudogenes.

Alignment positions are indicated by the black box. Species conservation to chimpanzee, mouse and chicken genomes are from the Evolutionary Conserved Regions browers (http://ecrbrowser.dcode.org/). Structures from the RNAfold webserver (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) are shown for reads of (A) *HSPA9*, (B) *GNL3* and (C) *MATR3*.



Supplemental Figure 1. Non-templated nucleotide addition in mitochondria.

The non-templated THP length distribution of mitochondrial transcripts.

Materials And Methods

Cell culture

The hESC line H9 (NIH designation, WA09), obtained from the University of Wisconsin Alumni Research Foundation, was cultured as described (Gaur et al. 2006). In brief, H9 cells were maintained as undifferentiated colonies by co-culture on irradiated CF1 MEFs (Bodnar et al. 2004) plated on 0.1% gelatin-coated 6-well dishes in DMEM high glucose supplemented with 2 mM L-glutamine, 1X penicillin/streptomycin (Gibco BRL), and 10% fetal bovine serum (Hyclone). One day after plating, the media was removed, and the MEFs were washed once with phosphate-buffered saline without calcium or magnesium (PBS-CMF; Gibco BRL) immediately prior to plating the hESCs. The hESCs were grown on MEFs in KSR media: Knockout-DMEM medium containing 20% Knockout Serum Replacement, 2 mM Lglutamine, 0.1 mM (1X) non-essential amino acids, 4 ng/ml human basic fibroblast growth factor (Gibco BRL), and 0.1 mM β-mercaptoethanol (Sigma). To selectively release hESC colonies from the MEFS, the cells were incubated with a combination of collagenase (Type IV)/dispase (Gibco BRL) for 20 min at 37°C. Released hESC colonies were gently transferred to a 15ml conical tube and allowed to settle by gravity for 5 minutes. Supernatant was removed and hESC colonies were gently washed once with KSR media and once with PBS. Clumps were again allowed to gravity settle, supernatant was removed, and clumps disrupted by P-1000 pipet. Cells were >99% free of MEFs and were used for RNA isolation.

RNA and DNA purification

Small RNAs (<200 nt) were purified from ~3 million cells using the mirVana miRNA Isolation Kit (Ambion) according to manufacturer's directions. DNA from H9 cells were purified using the PureLink Genomic DNA Mini Kit (Invitrogen).

Strand-specific RNA linkering and cDNA library preparation

DNA and RNA oligos for RNA sequencing were synthesized by IDT. The 3'-linker (5'-rApp-AGA TCG GAA GAG CGG TTC AGC AGG AAT GCC GAG/3InvdT/) was adenylated and purified as described in (Pfeffer et al. 2005). RNA linker ligation was done as previously described (Neilson et al. 2007) but briefly, pre-adenylated 3'-linker was ligated to small RNAs using T4 RNL2truc (NEB) without ATP in the reaction buffer. After ethanol precipitation, the reaction products were used for 5'-linker ligation (/5Cy3/rArCrA rCrUrC rUrUrU rCrCrC rUrArC rArCrG rArCrG rCrUrC rUrUrC rCrGrA rUrCrU) without gel purification with T4 RNL1 (NEB) with ATP. RNA was reverse transcribed (5'-GCT GAA CCG CTC TTC CGA TCT) with SuperScript III (Invitrogen) and amplified with 25 cycles of PCR with Platinum Taq polymerase (Invitrogen) and Illumina PE PCR primers (Forward primer: AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T, Reverse primer: CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG ATC T). PCR products greater than 140 bp were agarose gel purified (Qiagen QIAquick Gel Extraction Kit), ethanol precipitated and resuspended in elution buffer. Taking into account the length of Illumina PE linkers, this corresponds to RNA inserts > ~21 nt. Library fragment size and concentration were measured with the Bioanalyzer High Sensitivity DNA chip assay (Agilent).

Paired-end next generation sequencing and analysis

The prepared library was submitted for one lane of Illumina GAIIx pair-end sequencing (76 x 76 cycles) at the Michael Smith Genome Sciences Centre, Vancouver, Canada. The paired-end 2 (PE2) sequencing primer library was used to analyze the 3'-end of RNAs. Using custom Python programs, reads were trimmed of the initial homopolymer, regardless of length, for each nucleotide. Trimmed nucleotides were noted and the remaining sequence was aligned to the full hg18 genome (UCSC genome browser) with Novoalign (version 2.07 Novocraft Technologies). Only reads that aligned to unique loci were further analyzed. Reads were also compared to the mm9 mouse genome. Reads that had fewer mismatches when mapped to mm9 were considered contamination from mouse feeder cells and removed from further analysis. The previously removed homopolymer portion of each read was compared to the expected genomic sequence at the aligned locus to identify RNAs with mismatches in the 3'-end.

Read annotation

Reads were annotated by checking for any overlap of coordinates on the same strand when searched for in annotation databases in the following order: mirBase 16, UCSC (hg18) rnaGene.txt, UCSC tRNAs.txt, UCSC refGene.txt and UCSC knownGene.txt. For reads that did not match, the search was repeated after adding 50 bp to the coordinates of mirBase, rnaGene and tRNA genes and adding 1000 bp to refGene and knownGene coordinates. Reads matching these relaxed coordinates were labeled "Gene Name – flank". For remaining reads that failed to match the expanded coordinates, the search was repeated to check for anti-sense strand origin and reads matching those conditions were labeled "Gene Name – anti" or "Gene Name – anti – flank". Sequences that included "SNORD" or "SCARN" in the annotation to refGene were removed and included as snoRNAs genes.

Sequencing data

The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (Edgar et al. 2002)and are accessible through GEO Series accession number GSE31051 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31051).

Acknowledgements

We would like to thank Shaohui Wang for her help with ES cell culture. This work was supported by a National Science Foundation Graduate Research Fellowship (YSC), a grant from the California Institute for Regenerative Medicine RC1-00347 (ADL, MTM), and a National Institutes of Health grant 5R01DA026065 (MTM). We are grateful to members of the McManus lab, especially Matt Hangauer, Greg Ku and Nikki Shariat for their help in drafting this manuscript.

Chapter 3: Development of a U-tailed RNA-Seq Enrichment Method

Abstract

The complexities of eukaryotic RNA transcription and processing pose significant challenges to accurately surveying the cellular RNA repertoire. Several variations in RNA-Seq library preparation methods are used to enrich for different subsets of the total cellular RNA pool. The specific choices made in RNA purification, post-purification processing, adaptor ligation, cDNA synthesis, library amplification and sequencing technology platform can all affect enrichment of specific RNAs of interest but can also introduce inherent biases in the resulting libraries. To expand the search for RNAs subject to non-templated oligouridylation beyond those identified in the small RNA population, several changes to the RNA-Seq library preparation method were tested. A new RNA-Seq library was prepared based on these optimizations using total RNA purified from human embryonic stem cells. Analysis shows a new class of RNAs that may be subject to non-templated oligouridylation.

Introduction

Advances in DNA sequencing technology has made it routine to study gene expression by preparing cDNA libraries and using frequency of read counts as a measure of RNA transcript abundance. The enormous amount data generated from these deep sequencing methods are capable of detecting rare transcripts that would defy capture in shallow surveys using Sanger sequencing (Wang et al. 2009). However, even with these new technologies, total cellular RNA cannot be used directly because of the huge dynamic range of RNA transcript abundance *in vivo* that spans more than seven orders of magnitude (von der Haar 2008). By mass, rRNAs and tRNAs can make up more than 95% of the total RNA in cells (Warner 1999). Different methods have been devised to selectively remove these sequences from RNA-Seq libraries (Davis et al. 2007; Armour et al. 2009).

Size fractionation is a commonly used method for purifying specific RNA genes and is especially useful in studies of small RNAs because it removes the bulk portion of rRNA and tRNA. Due to the long polyA tail found on most eukaryotic mRNAs, oligo-dT primers are often used to purify RNAPol-II transcripts for gene expression studies (Wang et al. 2009). Selective degradation of undesired RNA by hybridization of DNA oligomers complementary to target RNAs followed by RNase H digestion is used in studies in prokaryotes because long polyA tails are not found on microbial mRNA (He et al. 2010).

Next-generation sequencing technologies still require the addition of known sequences, or adaptors, to the 5' and 3' ends of RNA. The known sequence could be a simple non-templated polyA tail added on by polyA-polymerase (Ingolia et al. 2009). This type of pre-processing can be useful because it creates a uniform sequence on the 3'-ends of a small RNA population, thus removing the biases created by RNA ligases that are used to add a known adaptor to the pool of purified RNA (Ingolia et al. 2009). However, this approach alters the native 3'-end and cannot be used to study post-transcriptional processing of RNA 3'-ends. Ligases derived from T4 bacteriophage are often used to add defined adaptors to RNA. A modified form of T4 RNA Ligase 2 (T4 RNL2) that is truncated to only the first 124 amino acid residues is used to ligate adaptors to small RNAs. This truncated form of the enzyme, T4 RNL2trnc, is catalytically impaired and requires an activated, pre-adenylated RNA or DNA adaptor (Ho et al. 2004). This is useful because a fully competent RNA ligase, such as T4 RNL1, when presented with an RNA with a 5'-monophosphate will complete the ligation reaction on the nearest available 3'-OH which in the case of small RNAs results in self-circularization (Pfeffer et al. 2005). Studies have characterized the biases of T4 RNL1 which shows different rates of ligation depending on the sequence of the donor and acceptor molecules (England and Uhlenbeck 1978). T4 RNL2 and its catalytically impaired derivatives have not been systematically studied but they show different rates of activity depending on the sequence of the molecules involved in the reaction. These biases can be large; more than three orders of magnitude difference in read counts has been observed from sequencing a defined, equimolar RNA pool (Hafner et al. 2011). Modifications to the RNA 3'- or 2'-OH will also influence the efficiency of adaptor ligation (Munafo and Robb 2010) and are unavoidable limitation of cloning nucleic acids.

There are further complications to consider when adding an adaptor to the 5'-end of RNA because there are several possible functional groups that can be present in eukaryotic RNAs. A 5'-monophosphate can be directly used as a substrate by T4 RNL1. However, newly transcribed RNA will have a 5'-triphosphate or a 5'-cap (Peterlin and Price 2006). RNA that has spontaneously degraded and some endonucleases may leave a 5'-OH

(Okorokov et al. 1997). Methods using nucleic acid modifying enzymes are used to selectively capture capped RNAs to identify transcriptional start sites and to identify the 5'-ends of transcripts (Ozsolak and Milos 2011).

Alternative RNA cloning methods do not require adaptor ligation. In the case of polyadenylated RNA, first and second strand cDNA synthesis can proceed with oligo-dT and random primers. While this method removes the strong 3'-end coverage bias of reverse transcriptases (RT) and can lead to more uniform coverage of long mRNAs, it has the drawback of losing RNA strand information (Mortazavi et al. 2008). An alternative strategy to avoid the complications of 5'-end structure is to use T4 RNL1 to ligate an adaptor to the first strand cDNA (Pak and Fire 2007). This method is independent of the 5'-end structure of the RNA and maintains strand information however; the reaction is impeded by having a DNA substrate.

Other variations on cDNA library preparation using properties of reverse transcriptases do not require 5'-linker ligation. Reverse transcriptases are not as processive or accurate as replicative DNA polymerases and are prone to adding non-templated cytidines to the end of a template (Kulpa et al. 1997) so that a primer with 3'-guanosines can be used for second strand cDNA synthesis. An undesired consequence of the ability of reverse transcriptase to leave one template and jump to another is the formation of unintended, chimeric sequencing artifacts composed of segments of different RNAs (Brakenhoff et al. 1991).

The Illumina sequencing platform requires PCR amplification of the cDNA library prior to sequencing. As with any PCR strategy, there are options beyond the standard primers that are complementary to adaptor sequences. A common addition in RNA-Seq methodologies is a barcode sequence to the standard primers (Smith et al. 2010). More creative primer designs are permitted when amplifying genomic DNA such as multiplexed primer pools that target specific genes of interest (Tang et al. 2006; Li et al. 2009).

The initial survey of a subset of the small RNA population in H9 embryonic stem cells showed that many classes of RNA are subject to 3'-end non-templated nucleotide addition. To extend upon our earlier work and to broaden the search for oligouridylated RNAs, several changes to our previous the RNA-Seq library preparation method were tested. A modified PE2 adaptor was designed that was more efficient at ligating to RNAs with 3'-end uridines. A new library was made from rRNA-depleted H9 total RNA to increase the chance of sequencing long RNAs. A modified PE2 reverse PCR primer was used to selectively amplify cDNAs made from RNAs that ended in 3'-uridines. Most of the reads in the new library aligned to loci that were previously sequenced in the small RNA dataset. Examination of reads that originated from new loci suggests that RNA transcribed from repetitive elements may also be oligouridylated.

Results

Development of an RNA-Seq protocol that enriches for oligouridylated RNA

Many possible modifications to the RNA-Seq cloning were considered to expand the search for RNAs subject to post-transcriptional 3'-end oligouridylation beyond those identified in the survey of small RNAs. Some aspects of the cloning strategy could not be altered such as using sequential ligation of adaptors to maintain RNA strand information. Also, RNA could not be sheared prior to adaptor ligation to maintain the native 3'-ends. However, the RNA purification method and the sequence of the PE2 adaptor were changed.

Total cellular RNA without size selection was used to prepare an RNA-Seq library. This introduced the complication of the great cellular abundance of rRNA. Intact rRNA was removed using a pool of lock nucleic acid (LNA) probes against highly conserved sequences in eukaryotic rRNA. There was no selection against tRNAs as that population produced fewer reads than expected in our previous study (Figure 1).

Characterization of T4 RNL1 showed that ligation efficiency of the enzyme depended on the sequence of both acceptor and donor molecules (England and Uhlenbeck 1978). RNAs with 5'-cytidines in the linker were the most efficient and 5'-adenine the least efficient. While the 3'-end sequence is independent in our purified library, it is interesting to note that the study of T4 RNL1 showed that adenines made the most efficient substrates and RNAs that end with uridines had the slowest ligation rate. T4 RNL2trunc has not been analyzed in the same fashion but it too shows biases when ligating small RNAs (Hafner et al. 2011). The sequence of the PE2 adaptor oligomer was modified to introduce 5'-CTGCTG, the recognition sequence for EcoP15I, a type III DNA restriction endonucleases that cuts 25 bases away from the recognition site. This site was introduced in case it would be necessary to compensate for the biases against long RNAs in amplification by RT-PCR and in cluster generation on the Illumina flow cell. Limited testing of T4 RNL2trunc shows that the addition of the EcoP15I site may increases the efficiency of ligating to some RNAs that end in uridines (data not shown).

Enzymatic selection of 3'-uridylated RNA

Enzymatic selection for RNA with terminal uridines was tested using reverse transcriptase. These enzymes derived from retroviral sources require a primer that is complementary to a target sequence to begin cDNA synthesis. A primer extension assay was used to determine if RT could selectively synthesize cDNA from RNAs with 3'-uridines. The ability of RT to distinguish between a primer that exactly matched or mismatched a template was tested under different primer concentrations and dNTP concentration. At the higher dNTP concentration that is typically used for cDNA synthesis, a considerable signal from mismatched primer extension was seen (Figure 8, lane 4). Reducing the dNTP concentration ten-fold reduced the rate of mispriming (Figure 8, lane 8) but there was a reduction in overall cDNA synthesis that required more cycles of PCR amplification (data not shown).

The ability of *Taq* DNA polymerase to discriminate between a matched and mismatched template was tested. PCR depends on the complimentary annealing of a DNA oligomer to a template, especially at the 3'-end of the primer. This essential requirement is the basis of the single nucleotide primer extension assay (SNuPE) assay (Singer-Sam et al. 1992; Singer-Sam 1994) and is the reason why in site-directed PCR mutagenesis, the mutation in the PCR primer must be distant from the 3'-end of the primer. A reverse primer for PCR that was complementary to the modified PE2 adaptor but was extended by five additional 3'-adenosines was able to selectively amplify the matching template present at

1,000-fold lower concentration than a competing, mismatched template (Figure 9). Because of the stringency of *Taq* DNA polymerase in amplifying correctly annealed primers, we used a PCR-primer based strategy to selectively amplify cDNAs made from RNAs with 3'-uridines (Figure 10).

Identification of new U-tailed RNAs by PCR-enriched paired-end sequencing

The sequence analysis method used in our earlier study had to be modified because of the addition of the EcoP15I site to the PE2 adaptor and because the PCR reverse primer masks the terminal five nucleotides of the ligated RNA. The first eleven bases of the PE2 sequencing primer library, being introduced by the adaptor and by the PCR primer, were removed from the start of all reads in the dataset. Because adenines were selected for by the PCR reverse primer, only further adenines were trimmed after removing the initial sequence. These additional trimmed adenines would correspond to a uridine THP longer than five nucleotides. Any further trimmed adenine nucleotides were noted and then alignment and annotation proceeded as was done for the small RNA dataset.

To identify new RNAs that may be subject to post-transcriptional oligouridylation, we compared the genome alignment position of the 3'-end of all reads with unique alignments in both libraries. If the 3'-end of a read in the PCR-enriched library was within ten nucleotides of any uniquely aligned sequence in the small RNA library, then that read was considered as a match to a previously sequenced RNA. RNAs with potential non-templated uridines were sorted by the number of mismatches in the AAAAA extension on the PCR reverse primer. RNAs in the PCR-enriched library should have terminal uridines either genomically encoded or post-transcriptionally added. As in our previous work, only those nucleotides that disagreed with the reference genome were attributed to post-transcriptional addition.

The PCR-enriched library contains few new RNAs

The number of reads and the number of uniquely aligned reads in the PCR-enriched library were comparable to the small RNA library (Table 5). However, the composition of the libraries differed greatly. Overall, the PCR-enriched library was dominated by sequences derived from 28S rRNA (Figure 11). Small RNA genes accounted for nearly all reads in the dataset with only 5.97% of reads aligning to long RNA genes in the refGene annotation database. As most reads aligned to small RNA genes, there were few reads in the PCR-enriched library that had not previously been identified (Table 6). The small fraction of reads that were newly observed were mostly derived from long RNA genes (Table 6, refGene) and from loci that did not have a known, annotated gene (Table 6, other).

Reads were sorted by the number of mismatches in the AAAAA PCR primer binding site to identify candidate RNAs with non-templated oligouridylation (Table 7). Overall, most reads appeared to have templated 3'-uridines with only 2.1% of reads having four or five mismatches in the PCR primer binding site. In contrast, RNAs that were not previously sequenced were eight-fold more likely to have more than three mismatches in the PCR primer binding site. Sequences from known genes were categorized according to their alignment position. In our previous work using a small RNA library, we observed frequent oligouridylation of TSSa-RNAs. In the PCR-enriched library, no new TSSa-RNAs or anti-sense TSSa-RNAs were in the high-mismatch reads group (Table 8). Most reads were initially assigned to intronic regions distant from splice sites or from fragments of exons.

Splicing complicates mismatch detection

Most of the reads with more than three mismatches in the PCR primer binding site were generated from single locations in six genes (Table 9). The majority of refGene reads were from exon fragments except for *TRIM28* reads which aligned to a splice site and *PODXL* reads that aligned in the 3'-UTR. Detailed examination of the sequences revealed that the some of the mismatches in the PCR primer binding site could be resolved by compensating for mRNA splicing. For example, the reads that aligned to *HNRNPD* would match the RNA sequence of a known alternative splice variant that skips exon 2. The alternatively spliced alignment would have only one mismatch in the PCR primer binding site (Figure 12A).

Because this paired-end library was not designed to study mRNA sheared to an expected size, the alignment was done using parameters for aligning short sequences. For the reads that aligned to *UTP14C/ALG11*, Novoalign calculated a higher score for a continuous alignment with three mismatches (Figure 12B) even though a gapped alignment that spanned an intron would create an exact match (Figure 12C) to *UTP14A*. However, in either UTP14 alignments, there are four mismatches in the expected PCR primer binding site.

Repetitive elements may be oligouridylated

After the six most frequently sequenced genes, regions with no known gene annotation generated the most new RNAs that may be oligouridylated (Table 9). Closer examination of the approximately 6.8% of reads with four or five mismatches in the PCR primer binding site revealed that many of these sequences originated from repetitive elements. Of the reads that had five mismatches to the PCR primer, 392 reads (237 unique sequences) did not have a known gene at the alignment position but overlapped a known repetitive element (UCSC Genome Browser RepeatMasker track) at the alignment loci. L1

LINE (Long Interspersed Element), tRNA-derived pseudogenes, L2 LINE and 28S rRNA were the most commonly found repetitive elements in this population.

The RNAs with four mismatches in the PCR primer binding site had a much less complex sequence composition; 1,607 reads came from 79 unique sequences. Nearly all of the reads in this group (1,546) aligned to a region 1.3 kb upstream of the transcriptional start site of *ZNF 496*, too distant to be classified as a TSSa-RNA. There were no annotated repetitive elements annotated upstream of ZNF 496 but there was a spliced EST that spanned 2.7 kb upstream of the TSS. Other reads in this group with four mismatches, aligned to Alu SINEs (Short Interspersed Element) and L1 and L2 LINEs. Sequences that aligned to known genes could have sequence similarity to repetitive DNA sequences as well. The reads that aligned to the 3'-UTR of *PODXL* overlapped the position of a MIR3 SINE. Most of the remaining reads in the high-mismatch group were reads anti-sense to known genes and were counted once or twice.

Because the PCR reverse primer masks the 3'-end sequence of the RNA in the process of PCR-enrichment and amplification, any candidate RNA that may be oligouridylated must be sequenced directly by an alternative method to determine the sequence of the 3'-THP. To directly examine the 3'-end sequences of RNA that did not align to known genes, we analyzed the intergenic reads in the small RNA library prepared without PCR selection. There were 206,150 reads that aligned to intergenic regions of the nuclear genome, 12.6% of these had non-templated 3'-end nucleotides. This population was divided according to the number of 3'-end mismatches and sorted into categories of repetitive elements in the UCSC Genome Browser RepeatMasker track (Table 10). Almost all sequences of in this population aligned to a region of the genome with a repetitive element.

These reads were collapsed to unique sequences and their THP length distribution was calculated (Figure 13). There is an over-representation of 3'-end oligouridines even in the reads without mismatches which is consistent with most SINEs being transcribed by RNA-Pol-III (Kriegs et al. 2007) and the templated uridines being part of the RNAPol-III terminator (Bogenhagen and Brown 1981). As seen in other classes of RNA, uridines and adenines were the most frequently added non-templated nucleotides (Figure 1) with oligouridine addition dominating for short THPs and polyadenylation more frequent at longer lengths.

Discussion

Because there were several changes made in the modified Illumina paired-end RNA-Seq protocol, it is difficult to make direct comparisons between the PCR-enriched library and the small RNA library. The EcoP15I site added to the PE2 adaptor changes the ligation bias and thus could changes the population of RNA captured for sequencing. The PCR-enriched library also used a different starting pool of RNA that included long RNAs. The difference in the starting RNA populations is difficult to calculate as the amount of long RNAs after rRNA depletion was not quantifiable. Also, it is not the mass of RNA but the molar concentration of 5'-monophosphates and 3'-OH that is relevant in terms of ligation.

The AAAAA extension on the PE2 reverse PCR primer was designed to create a bias specific for cDNAs synthesized from RNAs with 3'-uridines. There is no simple way to measure if the resulting library was enriched for 3'-U RNAs because there were different selection bias at RNA ligation and PCR amplification. A crude measure of enrichment is to compare the THP distribution of the sequences in the small RNA library that were found in the PCR-enriched library to the THPs of the small RNA sequences that were not cloned in the PCR-enriched library. 8,343,939 reads in the small RNA library were also in the PCR-enriched library while 6,229,137 sequences failed to align to a sequence in the PCR-enriched library. If enrichment worked, then uridine THPs, either genomically encoded or created post-transcriptionally, should be more abundant in the small RNAs that were also sequenced in the PCR-enriched library. Because post-transcriptional processing shifts the position of the 3'-end, a buffer of ten nucleotides was used to allow for a processed RNA to count as a match to a nearby read. This imprecise position matching permits oligouridylated and non-oligouridylated sequences to count as being in both RNA-Seq libraries. Another caveat of this measure of "U-ness" is overall uridine content is not counted for short homopolymers of one or two nucleotides. For example, if the end of an RNA was UUAUU-3', it would be read as a two nucleotide uridine THP and any uridines preceding the adenine hidden by this counting method. However, even with these limitations, there was a noticeably higher frequency of oligouridine THPs in set of small RNAs that were also sequenced in the PCR-enriched library compared to the small RNAs that were not present in the enriched library (Figure 14).

It is not surprising that the composition of the libraries were so different given the differences in the preparation method. The biggest change was in the fraction of reads that aligned to rRNA (Figure 11). The electropherogram trace from the Bioanalyzer RNA Nano assay showed effective removal of intact rRNAs but even with no visible signal, 28S rRNA related sequences had the largest read count of all genes. It is possible that the LNA probes did not bind to truncated or degraded rRNA transcripts so there could have been a significant population of rRNA-derived transcripts still in the RNA pool used for library preparation. Polyadenylated rRNA has been reported by others (Slomovic et al. 2010) and non-templated nucleotide 3'-end addition may be a signal of quality surveillance used to mark rRNAs for degradation (Wang and Pestov 2011). Uridine addition may have been missed by shallow sequencing used in these studies and would require independent validation using deeper, targeted cloning.

After 28S rRNA, mitochondrial tRNA-Pro was the most frequently sequenced gene (2.7 million reads) in the PCR-enriched library (Figure 11). This gene was sequenced 100,920 times in the small RNA library, 93% of with non-templated nucleotide addition. Most of the reads with non-templated nucleotides were due to 3'-CCA addition (80.0%) or polyadenylation. The implied sequence of the reads that aligned to this tRNA in the PCR-enriched library requires loss of or no 3'-CCA addition and further trimming or endonucleases cleavage by eight nucleotides. Only two reads matching that 3'-end were in the small RNA library. Given the scarcity of the truncated form of this tRNA and the low rate of non-templated oligouridylation of mitochondrial transcripts (Supplemental figure 1), an alternative explanation for the great abundance of reads to this gene is unintended PCR primer binding. Instead of binding to the PE2 linker at the 3'-end of a captured RNA, the long reverse primer could have primed internally with 10 out of the last 11 bases of the reverse primer annealing to the target (Figure 15). This type of internal mispriming could happen on other templates so new RNAs in this dataset must be verified by independent cloning of the 3'-end to confirm the presence of non-templated uridines.

A change in the composition in the PCR-enriched library was expected because of the selection bias of the reverse primer. There appears to be a favorable bias for RNAs with short oligouridine THPs but there was a lower frequency of RNAs with long uridine THPs in the PCR-enriched library (Figure 14) and a lower diversity of sequences overall (Table 5). Most reads were to short RNA genes with few long RNAs captured. We used PCR conditions to allow long amplicon extension but there could still be bias and loss of diversity due to shorter products being favored during PCR and during cluster generation on the Illumina flowcell.

However, the new sequences that were observed in the PCR-enriched library were mostly derived from long RNA transcripts which may be due to including long RNAs in the starting pool of purified RNA. The reads that aligned to long RNA genes were usually mRNA fragments or parts of long introns. Only 40 new TSS-RNAs were found in the entire

PCR-enriched dataset. Long, intact mRNAs are not likely to be captured or analyzed by our method. Studies in yeast using circularized PCR of targeted genes have reported uridine addition to intact polyA tails (Rissland et al. 2007; Rissland and Norbury 2009). This type of post-transcriptional oligouridylation would not be detected by our sequence analysis method. The typical polyA tail is ~200 nts and those reads would be rejected by the homopolymer filter and removed from further analysis.

Using long RNAs will require changes to the sequence analysis method as short read alignments are not optimized to detect spliced RNA. Reads that aligned to long RNA genes will need to be processed separately by an alignment program that is better at detecting spliced mRNA. This will improve accuracy in alignment and mismatch calculation to identify with more confidence transcripts that may be oligouridylated.

The requirement for having a unique alignment to the genome may skew the relative abundances of RNA genes that are analyzed. A unique alignment is defined as a parameter in the alignment scoring algorithm. The calculation used in Novoalign to determine a unique alignment was to compare the scores given to the highest scoring alignments. The threshold requirement was that the best alignment score for a read needed to be three-fold greater than the next best alignment to count as a unique mapping to the reference genome. Many genes, especially small RNA genes, have multiple copies in the genome as functioning genes and as pseudogenes relics. Along with the enormous copy number load of repetitive elements, the unique read threshold requirement removed 6.7 million clusters from further analysis (Table 5) and a smaller fraction of reads in the small RNA dataset (Table 1). This may account for the low number of 5S and 5.8S rRNA sequences in the small RNA dataset even though those RNAs were within the size range captured in the small RNA fraction. Reads aligned to small rRNA genes along with many snRNAs would be filtered by having multiple alignment loci.

Many of the repetitive elements in the genome have diverged over time so that unique alignments are possible given a sufficiently long sequence (Lerat 2010). The new RNAs in the PCR-enriched library that aligned to repetitive elements pointed to the possibility that this class of RNAs may be subject to non-templated oligouridylation. This was confirmed by analysis of sequences in the small RNA library that did not align to known genes. The small RNA library directly sequenced the 3'-end of RNAs from repetitive elements and independently confirmed that non-templated uridines and adenines were frequently found on this class of RNA. The THP length distribution was similar to non-templated nucleotides added to other classes of RNA.

The modified RNA-Seq method seemed to enrich for U-tailed RNA and new RNAs derived from long RNA genes that may be oligouridylated were sequenced. However, because PCR-primer selection masks the terminal sequence of the RNA, these new candidate oligouridylated RNAs need to be confirmed by methods that do not introduce artifactual bases at the ligation site. Modification of the PE2 adaptor by the addition of the EcoP15I site created a separate complication. This restriction enzyme recognition sequence is the most over-represented trimer repeat in the human genome (Han et al. 1994). Having this GC-rich repeat next to the A_5 extention to the PCR primer may have allowed for mispriming and generated sequences upstream of the actual RNA 3'-end. The few new RNA sequences in the enriched library may be an indicator of the scarcity of oligouridylated RNAs. Further modifications to RNA-Seq methods are necessary to identify new cases of RNA oligouridylation. Sequencing RNA after blocking RNA degradation may be necessary to improve the chances of capturing these rare RNAs and to test the functional consequence of this post-transcriptional modification.

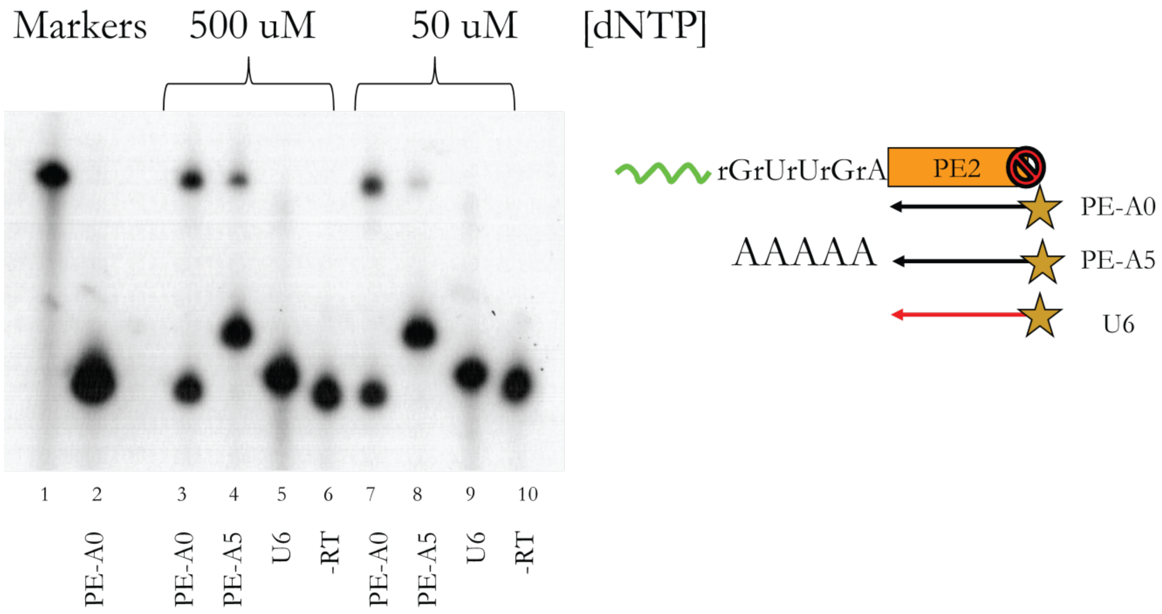


Figure 8.RT primer selection assay.

A 5'-end radiolabeled primer as indicated was incubated with a template with different concentration of dNTPs. Lane 1 is a radiolabeled DNA oligo of the length of a fully extended primer. Lane 2 is the PE-A0 primer. Reaction products were resolved by denaturing 15% PAGE and imaged by autoradiography on film. –RT lanes included the template and primer but no enzyme. U6 is a DNA probe against U6 and is not complementary to the PE2 adaptor.

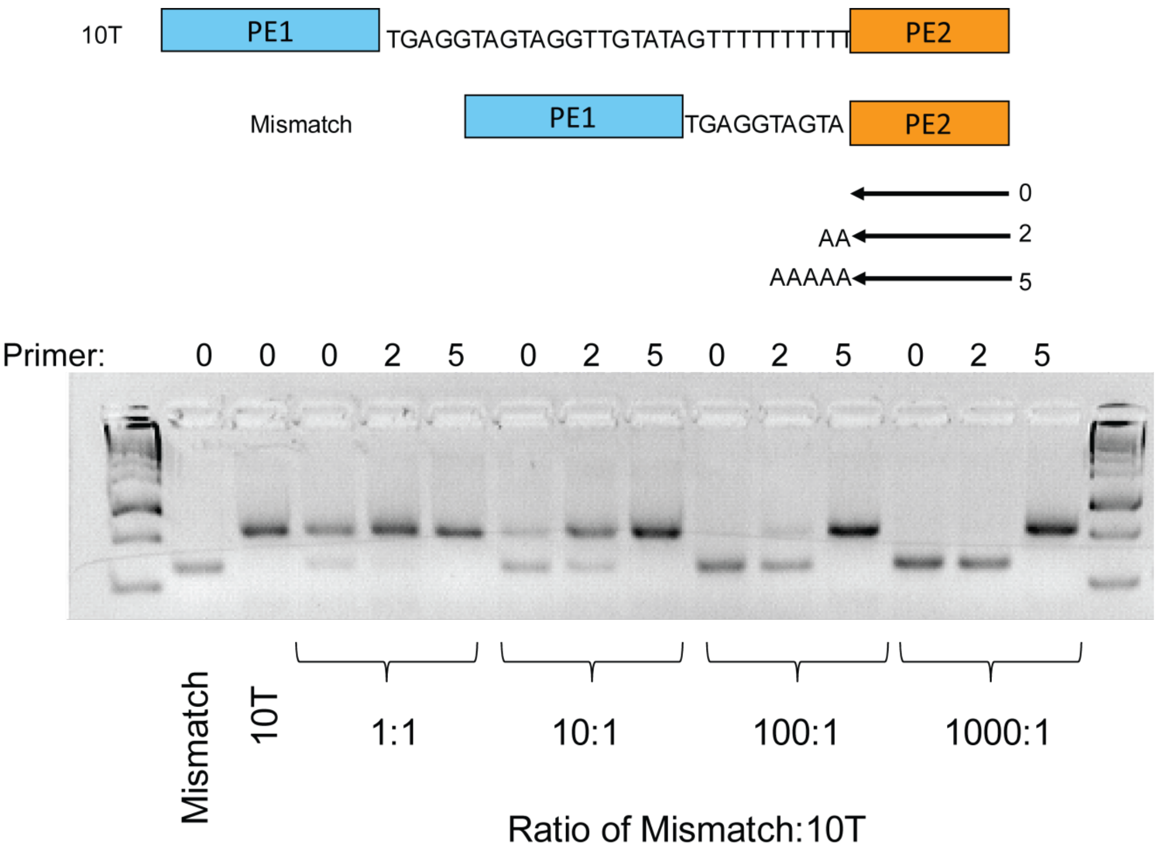


Figure 9. PCR primer selection assay.

Synthetic DNA templates were PCR amplified with the indicated reverse primer for different molar ratios of Mismatch or 10T templates. The concentration of Mismatch template was kept constant at [20 µM] and the 10T template was serially diluted to the indicated ratio. The lanes on the left are reactions with a single template to mark the position of each amplicon. Reactions were resolved by 5% agarose native gel electrophoresis in 1X TBE buffer and visualized by Sybr Safe dye.

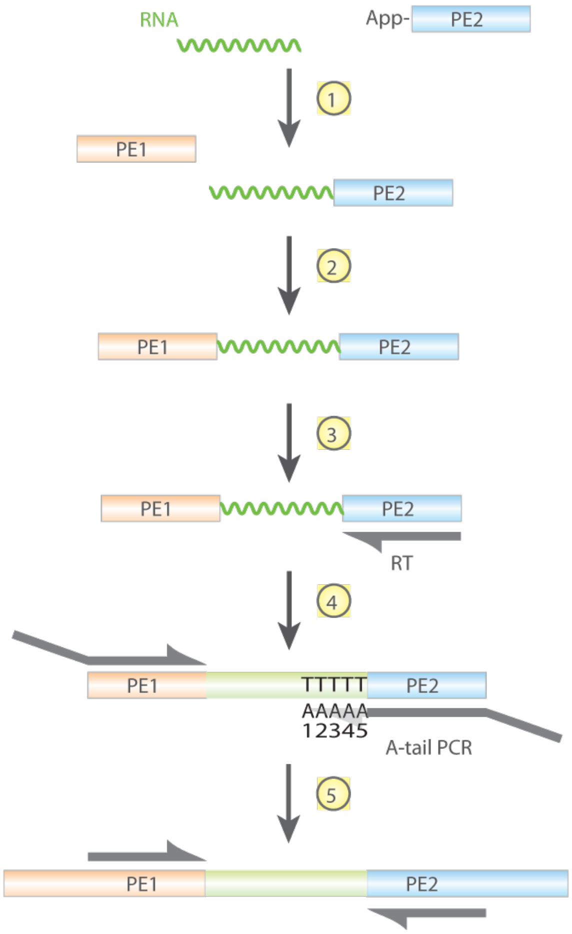


Figure 10. 3'-U-RNA PCR enrichment cloning protocol.

The modified RNA-Seq cloning method used to enrich for RNAs with 3' -uridine tails. (1) 3'-RNA linker ligation and (2) 5'-linker ligation is followed by (3) reverse transcription and (4) PCR to generate full length Illumina paired-end adaptors. For PCR-enrichment, a reverse primer with a 3'-extension of AAAAA was used to select for cDNAs with TTTTT at the ligation site. (5) The PE1 sequencing primer reads in the sense orientation 5' to 3' and the PE2 sequencing primer reads antisense 3' to 5'.

Table 5. Alignment of reads in PCR-enriched RNA-Seq library.

Input 35,986,119 cluster	Number of sequence
Aligned	24,893,272
Unique alignment	18,206,773
Gapped Alignment	1,594,461
Quality Filter	9,664,957
Homopolymer Filter	275
Alignments to unique loci –N, -mm9	17,338,808
Collapsed to unique sequences	672,855

Table 5. Alignment of reads in PCR-enriched library.

The alignment results of the PCR-enriched library prepared from H9 total RNA as reported by Novoalign. The PE2 sequencing primer library was filtered of reads with ambiguous base calls (-N) and better alignments to the mm9 mouse genome (-mm9).

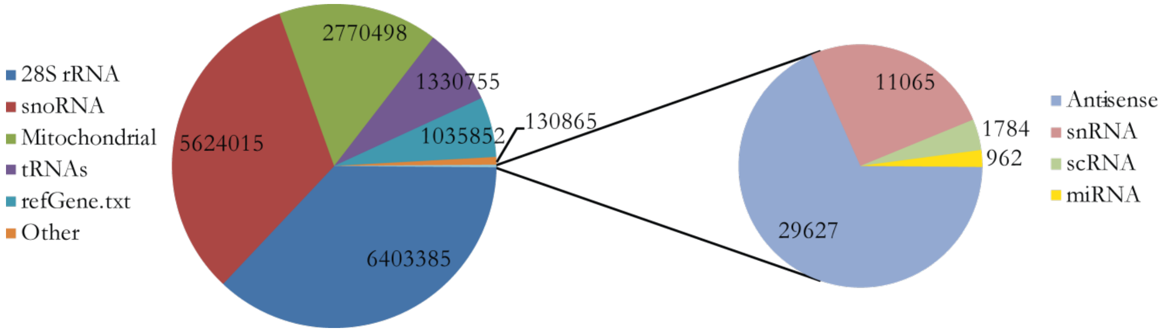


Figure 11. The distribution of reads in the PCR-enriched PE2 sequencing primer library.

The reads in the PCR-enriched RNA-Seq library were sorted into gene categories. Mitochondrial includes all mitochondrial gene including tRNAs. "Other" includes all other categories of reads including intergenic loci without known genes and flanking sequences to known genes.

Table 6. Annotation of reads from PCR-Enriched Library

Gene Class	All Reads	Reads in H9 Small RNA	New read not in H9 Small RNA library
28S rRNA	6,403,385	6,402,619	766
snoRNAs	5,624,015	5,624,004	11
Mitochondrial	2,770,498	2,770,497	1
tRNAs	1,330,755	1,330,682	73
refGene	1,035,852	872,932	163,920
Other	130,865	71,279	59,586
Anti-sense	29,627	9,875	19,752
snRNA	11,065	10,978	87
scRNA	1,784	1,778	6
miRNA	962	950	12
Total	17,338,808	17,095,594	243,214

Table 6. Annotation of PCR-Enriched Library.

Reads in the PCR-enriched library were annotated and categorized into gene classes and were filtered for sequences that were not previously sequenced in the H9 small RNA library made without PCR-enrichment.

Table 7. Distribution of mismatches in the A_5 PE2 reverse primer binding site

# Mismatches	All Reads	Reads in H9 Small RNA	Reads Not in H9Small RNA
0	2,783,450 (16.1%)	2,685,178 (15.7%)	98,272 (40.4%)
1	3,430,234 (19.8%)	3,401,521 (19.9%)	28,713 (11.8 %)
2	10,728,746 (61.9 %)	10,659,832 (62.4%)	68,914 (28.3%)
3	290,683 (1.7%)	272,825 (1.6%)	17,858 (7.3%)
4	78,772 (0.5%)	52,693 (0.3%)	26,079 (10.7%)
5	26,923 (0.2%)	23,545 (0.1%)	3,378 (1.4%)
Total	17,338,808	17,095,594	243,214

Table 7. Distribution of mismatches in the A_5 PE2 reverse primer binding site.

Reads in the PCR-enriched library were partitioned into sequences that were or were not previously observed in the H9 small RNA-Seq library (Chapter 2). Reads were sorted by number of mismatches in the PCR primer binding site by comparing the AAAAA sequence in the primer to the expected sequence in the hg18 reference genome. Shown in parenthesis are the percentages of reads of the entire library or in the subgroup with the indicated number of mismatches in the primer binding site.

Table 8. Types of refGene reads with mismatches in the A_5 PE2 reverse primer binding site

Position of Read	0 – 3 MM	4 MM	5 MM
TSSa-RNA	40	0	0
3'-UTR	34,051	5	67
Splice site	3,151	3	55
Intronic	79,531	19,024	36
Antisense	18,699	8	2
Other (exon)	22,416	4,967	2,812
Total	157,888	24,007	2,972

Table 8. Types of refGene reads with mismatches in the A_5 PE2 reverse primer binding site.

New sequences that aligned to long RNA genes in the refGene annotation database were categorized in the same was as refGene annotated reads in the small RNA dataset (Table 4). Reads were grouped according to the number of mismatches in the PCR primer binding site.

Table 9. New RNAs that may be oligouridylated.

4 Mismatches	
Gene name	**Read count**
HNRNPD	19,020
SPTBN2	4,964
No known gene	1,607
UTP14C/ALG11	438
5 Mismatches	
EXOC2	2,812
No known gene	392
PODXL	66
TRIM28	55
HNRNPD	33

Table 9. New RNAs that may be oligouridylated.

Listed are the most frequently sequenced genes that had reads with four or five mismatches in the PCR primer binding site.

Table 10. Intergenic read in the small RNA library.

Category	Exact	Non-templated THP
ERV	14,822	463
MIR SINE	13,326	640
tRNA-repeat	9,747	909
Alu SINE	3,332	260
rRNA	1,015	51
No known gene or repeat	292	496
LINE	261	479
Alt-TSS *POU2F3*	239	23
snRNA		40
scRNA		12
Total	43,034	3,373

Table 10. Intergenic read in the small RNA library.

Reads that aligned to intergenic regions of the nuclear genome were divided based on the mismatches in the 3'-THP. The sequences with no mismatches in the THP (Exact) counted more than 100 times and sequences with non-templated 3'-end nucleotides counted more than 10 times were sorted into classes of repetitive DNA sequences. Most reads overlapped a known repetitive element. Endogenous retrovirus (ERV), mammalian-wide interspersed repeat (MIR), Short INterspersed Elements (SINE), Long INterspersed Elements, Alternative Transcriptional Start Site (Alt-TSS)

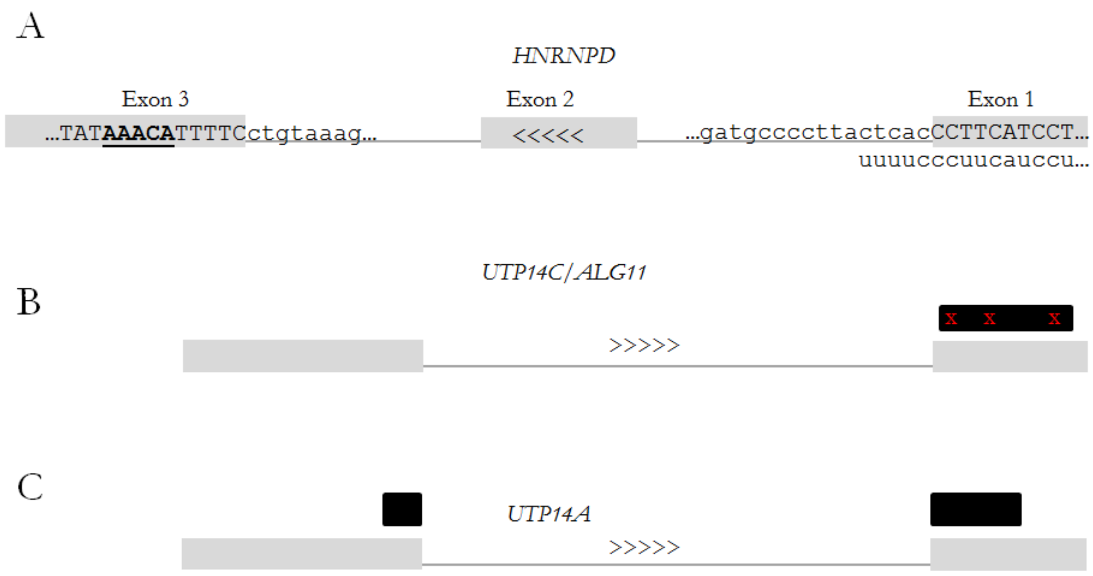


Figure 12. Splicing complicates mismatch detection.

(A) The gene structure of HNRNPD is shown with exons as grey boxes and introns as lines. Sequences of exons are in uppercase and introns in lowercase. The gene is on the reverse strand with the sequence aligned to the end exon 1 in lower case. A known splice variant is skipping of exon 2. The expected PCR primer binding site is underlined. (B) The best alignment scored by Novoalign for reads to *UTP14C/ALG11* (black box) is a continuous alignment with three mismatches. (C) A lower scoring alignment across spliced exon 4 and exon5 has no internal mismatches. Both alignments have four mismatches in the PCR primer binding site.

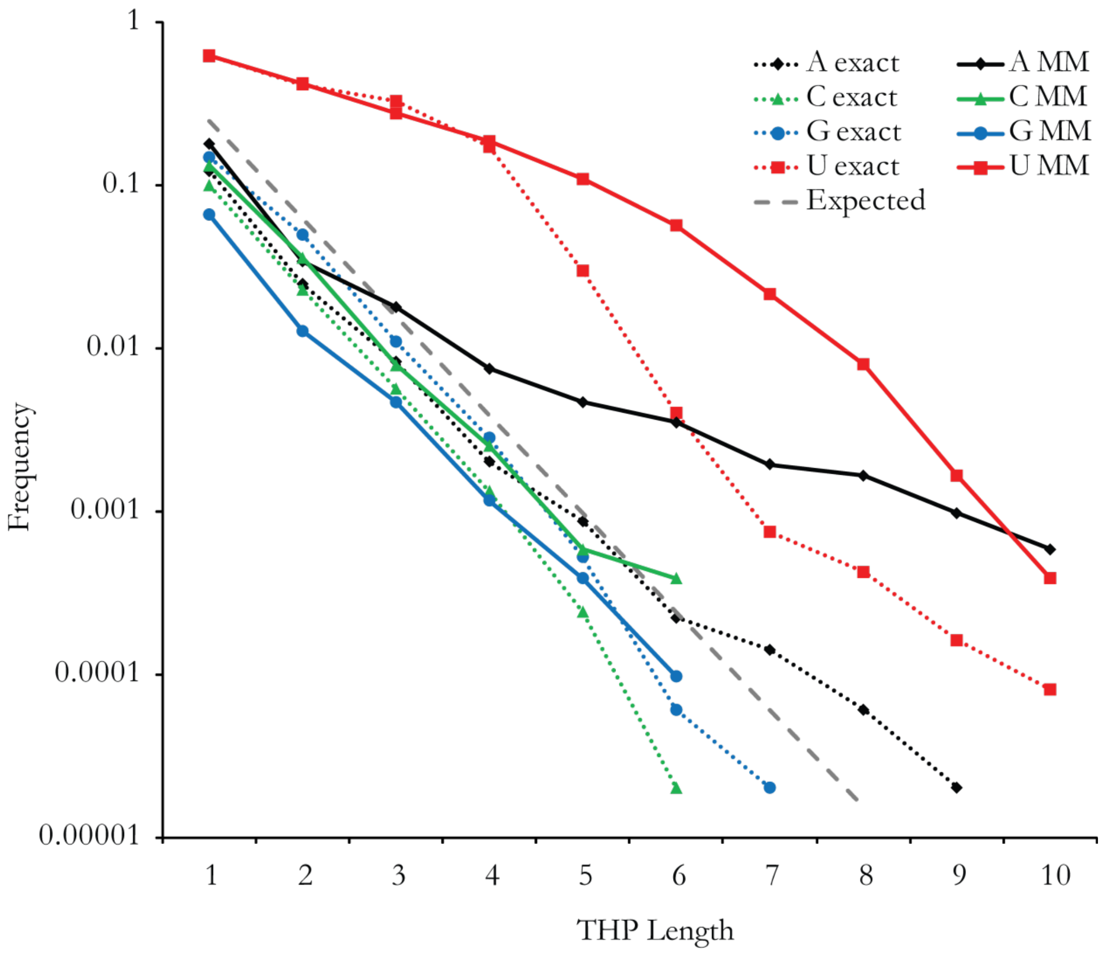


Figure 13. The distribution of THP lengths on intergenic RNA.

The THP lengths were counted for reads with no mismatches (exact) or reads with non-templated 3'-end nucleotides (MM). 185,933 exact reads collapsed to 49,332 to unique sequences and 26,152 mismatched reads collapsed to 10,258 unique sequences. The cumulative frequency of THP lengths of this collapsed dataset is shown, that is if an RNA had a THP of five uridines, it was included in the totals for lengths of one through five. The expected line is the probability of finding a homopolymer of a given length at random based on equal representation of all nucleotides.

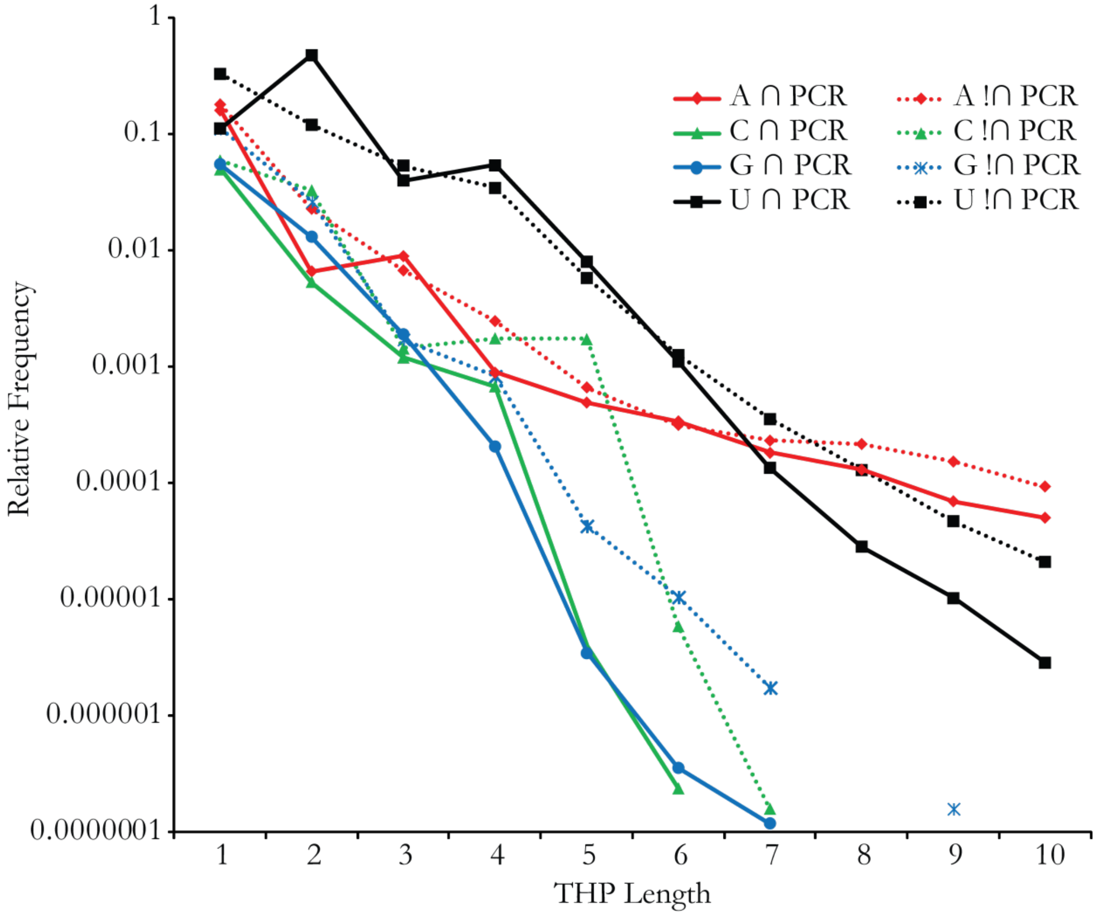


Figure 14. Effect of PCR-selection on RNA 3'-end THPs.

The THP length distribution of 8,343,939 reads from the small RNA library that were also in the PCR-enriched library (∩) were compared to the 6,229,137 small RNAs that were not observed (!∩) in the PCR-enriched library. A read from the small RNA library was considered to intersect (∩) with the PCR-enriched library if the 3'-end alignment position was within ten nucleotides of the 3'-end of an RNA in the PCR-enriched dataset.

```
cagagaauaguuuaaauuagaaucuuagcuuugggugcuaaugguggaguuaaagacuuuuuuCTGCTGA
cagagaauaguuuaaauuagaaucuuagcuuugggugcuaaugguggaguuaaagacuuuuuCTGCTGA
cagagaauaguuuaaauuagaaucuuagcuuugggugcuaaugguggaguuaaagacuuuuuuuCTGCTGA
CAGAGAATAGTTTAAATTAGAATCTTAGCTTTGGGTGCTAATGGTGGAGTTAAAGACTTTTTCTCTGAT
cagagaauaguuuaaauuagaaucuuagcuuugggugcuaaugguggaguuaaagacuuuuucucugacca Small RNA
                                                         AAAAAGA-GAC…
                                                                C
```

Figure 15. Potential mispriming in PCR.

The reference sequence of mitochondrial tRNA-Pro is in upper case in the grey box. The major form of the reads that aligned to this gene in the small RNA library are below with non-templated 3'–CCA addition underlined. The most frequently sequenced variants of reads in the PCR-enriched library are above. In red is the 3'-end sequence of the reverse primer showing how 10 of 11 bases would anneal if one nucleotide was displaced.

Materials And Methods

Primer extension assay

Primer extensions were done with 2.0 pmol of a ^{32}P-labeled primer mixed with 5.0 pmol of template, annealed at 65°C for 5 min before incubation at 55°C for 60 min with 10 U SuperScript III (Invitrogen). Reactions were heat inactivated at 70°C for 15 min. Reactions were resolved by denaturing PAGE on 15% gels (National Diagnostics) and imaged by autoradiography on film.

Cell culture

The hESC line H9 (NIH designation, WA09), obtained from the University of Wisconsin Alumni Research Foundation, was cultured as described (Gaur et al. 2006). In brief, H9 cells were maintained as undifferentiated colonies by co-culture on irradiated CF1 MEFs (Bodnar et al. 2004) plated on 0.1% gelatin-coated 6-well dishes in DMEM high glucose supplemented with 2 mM L-glutamine, 1X penicillin/streptomycin (Gibco BRL), and 10% fetal bovine serum (Hyclone). One day after plating, the media was removed, and the MEFs were washed once with phosphate-buffered saline without calcium or magnesium (PBS-CMF; Gibco BRL) immediately prior to plating the hESCs. The hESCs were grown on MEFs in KSR media: Knockout-DMEM medium containing 20% Knockout Serum Replacement, 2 mM Lglutamine, 0.1 mM (1X) non-essential amino acids, 4 ng/ml human basic fibroblast growth factor (Gibco BRL), and 0.1 mM β-mercaptoethanol (Sigma). To selectively release hESC colonies from the MEFS, the cells were incubated with a combination of collagenase (Type IV)/dispase (Gibco BRL) for 20 min at 37°C. Released hESC colonies were gently transferred to a 15ml conical tube and allowed to settle by gravity for 5 minutes. Supernatant was removed and hESC colonies were gently washed once with

KSR media and once with PBS. Clumps were again allowed to gravity settle, supernatant was removed, and clumps disrupted by P-1000 pipet. Cells were >99% free of MEFs and were used for RNA isolation.

RNA purification.

Total RNA from H9 cells was purified with Trizol Reagent (Invitrogen) following manufacturer's protocol for adherent cells grown in monolayer. Forty µg of total RNA was depleted of rRNA with the RiboMinus Eukaryote Kit (Invitrogen) following manufacturer's protocol. Total RNA quality and rRNA depletion was measured with the Bioanalyzer RNA 6000 Nano chip assay (Agilent) according to manufacturer's protocol.

Strand-specific RNA linkering and cDNA library preparation

DNA and RNA oligos for RNA sequencing were synthesized by IDT. The 3'-linker consisting of an EcoP15I site added to the standard Illumina PE2 linker sequence (/5'-rApp/rCrUrG rCrUrG rArGrA TCG GAA GAG CGG TTC AGC AGG AAT /3InvdT/) was adenylated and purified as described in (Pfeffer et al. 2005). RNA linker ligation was done as previously described (Neilson et al. 2007) but briefly, pre-adenylated 3'-linker was ligated to small RNAs using T4 RNL2truc (NEB) without ATP in the reaction buffer. After ethanol precipitation, the reaction products were used for 5'-linker ligation (/5Cy3/rArCrA rCrUrC rUrUrU rCrCrC rUrArC rArCrG rArCrG rCrUrC rUrUrC rCrGrA rUrCrU) without gel purification with T4 RNL1 (NEB) with ATP. RNA was reverse transcribed (5'-ATT CCT GCT GAA CCG CTC TTC CGA TCT CAG CAG) with SuperScript III (Invitrogen).

PCR selection of U-tailed RNAs

A PCR-enrichment strategy was used to selectively amplify cDNAs with 3'-TTTTT at the PE2 ligation site. The cDNA library prepared from rRNA-depleted H9 total RNA was amplified with 30 cycles of PCR with Platinum Taq polymerase (Invitrogen) and modified Illumina PE PCR primers (Forward primer: 5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T, Reverse primer: 5'-CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG ATC TCA GCA GAA AAA). PCR products greater than 150 bp were agarose gel purified (Qiagen QIAquick Gel Extraction Kit), ethanol precipitated and resuspended in elution buffer. Taking into account the length of modified Illumina PE linkers, this corresponds to RNA inserts > ~20 nt. Library fragment size and concentration were measured with the Bioanalyzer High Sensitivity DNA chip assay (Agilent).

Paired-end sequencing and analysis

The prepared library was submitted for one lane of Illumina GAIIx pair-end sequencing (85 x 85 cycles) at the University of California, Davis Genome Center. The paired-end 2 (PE2) sequencing primer library was used to analyze the 3'-end of RNAs. Because the standard PE2 sequencing primer does not include the EcoP15I site and the five adenines in the PE2 reverse primer, custom Python programs were used to trim the initial 11 bases from all sequences in the library. If there were additional adenines revealed after initial trimming, they were removed as well as it could represent a uridine tail longer than five nucleotides in the RNA. Trimmed nucleotides were noted and the remaining sequence was aligned to the full hg18 genome (UCSC genome browser) with Novoalign (version 2.07 Novocraft Technologies). Only reads that aligned to unique loci were further analyzed. Reads were also compared to the mm9 mouse genome. Reads that had fewer mismatches when mapped to mm9 were considered contamination from mouse feeder cells and removed from further analysis. The previously removed homopolymer portion of each read was compared to the expected genomic sequence at the aligned locus to identify RNAs with mismatches in the 3'-end.

Read annotation

Reads were annotated by checking for any overlap of coordinates on the same strand when searched for in annotation databases in the following order: mirBase 16, UCSC (hg18) rnaGene.txt, UCSC tRNAs.txt, UCSC refGene.txt and UCSC knownGene.txt. For reads that did not match, the search was repeated after adding 50 bp to the coordinates of mirBase, rnaGene and tRNA genes and adding 1000 bp to refGene and knownGene coordinates. Reads matching these relaxed coordinates were labeled “Gene Name – flank”. For remaining reads that failed to match the expanded coordinates, the search was repeated to check for anti-sense strand origin and reads matching those conditions were labeled “Gene Name – anti” or “Gene Name – anti – flank”. Sequences that included “SNORD” or “SCARN” in the annotation to refGene were removed and included as snoRNAs genes.

Identification of new RNAs with 3'-oligouridine tails

Reads from the PCR-selected library were checked to the reads in the H9 Small RNA library prepared without PCR enrichment (Chapter 2) to identify new RNAs with 3'-uridine tails. A custom python program compared the alignment position of the 3'-end of reads in the libraries. If a sequence in the PCR-selected library was within 10 nts of any read in the small RNA library, it was considered a match to an already sequenced RNA. Those reads that were > 10 nts distant from a known sequence were considered a potentially new oligouridylated RNA.

The number of mismatches in the AAAAA reverse primer extension was counted to identify potential cases of non-templated nucleotide addition. Reads with four or five mismatches were considered as optimal candidates for oligouridylated RNA.

Acknowledgements

We would like to thank Shaohui Wang for her help with ES cell culture. This work was supported by a National Science Foundation Graduate Research Fellowship (YSC), a grant from the California Institute for Regenerative Medicine RC1-00347 (ADL, MTM), and a National Institutes of Health grant 5R01DA026065 (MTM).

Chapter 4: General Summary

The work presented here describes the efforts to better understand the scope and function of non-templated oligouridylation, a modification that has been recently reported to occur on many coding and non-coding RNAs. We also tested and optimized a modified RNA-Seq method to specifically enrich for what may be rare and short-lived RNAs. Using paired-end sequencing and custom sequence analysis programs, we were able to deeply profile non-templated nucleotide addition in RNA purified from human embryonic stem cells. We identified examples of already discovered non-templated nucleotide addition on many classes of RNA. We also identified and reported for the first time the post-transcriptional uridylation of several types of RNA. These findings point to a possible function for this RNA modification and suggest future experiments to address the many implications and questions raised by our data and analysis.

Oligouridylation on processed RNAs

Extensive sequencing of RNA from human embryonic stem cells revealed widespread non-templated 3'-end nucleotide addition. Non-templated uridine and adenine accounted for nearly all cases of mismatched nucleotides and were commonly found on RNAs that required processing by nucleases in order to be captured and sequenced by our cloning method. Oligouridylation was especially abundant and was found on mRNAs at all stages of biogenesis from newly transcribed TSSa-RNAs, RNAs generated during maturation such as spliced introns and on degraded mRNA fragments. Many lines of circumstantial evidence point to oligouridylation being linked to RNA degradation. With so many types of RNA found with non-templated uridines, there are numerous targets to pursue to test for functional consequence of this modification and to determine the kinetics of non-templated oligouridylation.

Experiments are in progress to study the functional consequences of oligouridylation and RNA stability. In addition to the seven known TUTases, siRNAs were picked to target RNA processing factors such as decapping enzymes and components of the nuclear and cytoplasmic exosome. In the absence of conclusive data, we can only speculate on the necessity of this type of RNA modification or on the interactions between different enzymes and pathways involved in RNA processing and speculate about the evolutionary path taken to arrive at this activity.

Oligouridylation and decapping

There is evidence from *in vitro* studies that shows that oligouridylation can stimulate decapping and that maximal increase in decapping rate is reached with five uridines (Song and Kiledjian 2007). Other studies have shown that there are accessory factors that regulate decapping in a tissue or gene-specific manner (Ghosh et al. 2004; Fenger-Gron et al. 2005). These findings point to the complications created by eukaryotic modifications to mRNA. The protective 5'-cap must be removed to allow for efficient 5' →3' turnover of mRNA, the dominant decay mode in yeast. Decapping enzymes must be regulated so that they do not act prematurely on nascent transcripts or on mRNAs stored for later translation upon activation. A mature mRNA is protected at both 5'- and 3'-ends by cap binding proteins and poly(A) binding proteins (PABP) (Kuhn and Wahle 2004). Histone mRNAs 3'-ends and snRNAs, the rare RNAPol-II transcripts that lack poly(A) tails, have their 3'-ends associated with protein complexes to protect those ends from nucleases (Krieg and Melton 1984). Oligouridylation of histone mRNA is able to overcome the protection of the 3'-end complex and signal for decapping and turnover of these messages after DNA replication is completed (Mullen and Marzluff 2008). The addition of uridines may be an effective way to mark capped RNA such as TSSa-RNAs and mRNA fragment by adding nucleotides that are not normally found on intact, properly processed RNAPol-II transcripts. Two PUPs (Mullen and Marzluff 2008) have been linked to histone mRNA oligouridylation and it would be interesting to test if these enzymes act on other capped transcripts and if there are factors that bind to 3'-oligouridines that interact with RNA processing complexes.

Oligouridylation and splicing

Another complication of eukaryotic mRNA processing is the removal of introns from pre-mRNAs. Introns generated by splicing are in the form of an RNA lariat that requires an essential debranching enzyme (DBR1) to hydrolyze the branch point linkage (Chapman and Boeke 1991). Most introns have short half-lives and are quickly degraded but the details of intron degradation are unknown. While debranching followed by degradataion by exonucleases is thought to account for turnover of most introns, there are other mechanisms of intron decay independent of debranching such as processing by RNase III (Danin-Kreiselman et al. 2003).

The high frequency of oligouridylation on RNA aligned to spliced site may be a clue to dissect the molecular details of intron decay. However, the example of the TRAMP complex gives reason to be cautious about drawing conclusions about non-templated nucleotide addition and RNA decay. Work done in yeast showed that the TRAMP complex enhances the turnover of a small population of introns and that this decay is not dependent on the polyadenylation activity of Trf4p or Trf5p (San Paolo et al. 2009). There is reason to question the importance of polyadenylation by the TRAMP complex and RNA decay as the lethal *trf4* and *trf5* double knock out mutant can be rescued with a catalytically inactive Trf4p (Wyers et al. 2005). RNA decay may be enhanced by other mechanisms such as the Mtr4p helicase component of the TRAMP complex which could disrupt strong RNA secondary structure ubiquitous on nuclear RNAs. Polyadenylation may be a nonessential accessory to enhance decay. Further study of the complexes bound to introns is needed to determine the enzyme or enzymes that uridylates spliced introns to test for functional consequences of uridine addition.

Non-templated uridines were found on both short (<26 nt) and longer spliced intron reads (Figure 4A). The short fragments could be generated by endonucleases that cleave near the branch point or it could the accumulation of the remainders of 5'→3' degradation following debranching. Targeting DBR1 and XRN1 to prevent debranching and 5'→ 3' exonucleases degradation may help to show if there is a major decay pathway for intron degradation in human cells and show if debranching must precede oligouridylation.

Oligouridylation kinetics

Sequencing of many types of RNA generated from disparate processes shows that oligouridylation is short, typically five nucleotides and limited to less than 10 nucleotides *in vivo*. This is the case for replication dependent histone mRNAs, RNAs sliced by Ago2, and miRNA precursors (Shen and Goodman 2004; Mullen and Marzluff 2008; Newman et al. 2011). There was a similar pattern overall in our RNA-Seq libraries with non-templated uridines being abundant on for THP lengths up to five nucleotides (Figure 2). There is an inflection point at five uridines where the slope changes from uridines being more abundant than expected by chance to a distribution that more closely follows the slope of random nucleotide incorporation. The reduced probability of non-templated oligouridylation beyond five nucleotides is more easily seen in the population of reads that derived from repetitive elements (Figure 13 and 16). Some of the unexpected high frequency of uridine THP may be due to 3'-end formation by RNAPol-III terminator sequence (TTTTT) but usually only two uridines transcribed before termination (Bogenhagen and Brown 1981). Also, short oligouridylation is seen on RNAs not transcribed by RNAPol-III.

Studies of the TRAMP complex show that the activity of a polymerase can be modulated by other members of the complex and that the rate of polymerization can change with the length of the polymerized RNA tail (Jia et al. 2011). Measurement of polyA tails *in vivo* shows that the TRAMP complex has fast polymerase activity until the tail reaches ~4-5 nt (Wlotzka et al. 2011). PUPs are related to the Trf4 and Trf5 components of the TRAMP complex. These enzymes show markedly higher processivity *in vitro* than the short tails that are observed *in vivo* (Kwak and Wickens 2007). Identification of PUP containing complexes and interacting partners using affinity purification could help to illuminate the factors regulating oligouridylation *in vivo* and how these marked RNAs are channeled to the next step in their processing.

Oligouridylation and repetitive elements

Analysis of candidate RNAs sequenced in the PCR-enriched library suggested that transcripts generated by repetitive elements may be oligouridylated. The presence of non-templated uridines was confirmed by analysis of reads to this class of RNA in the small RNA library (Figure 13). Repetitive elements such as L1 and Alu elements have been shown to be active in human embryonic stem cell lines (Macia et al. 2011) and in the germline (O'Donnell and Boeke 2007). Transcriptional and post-transcriptional mechanisms are employed to limit the expression of these extremely abundant elements (Esnault et al. 2005; Zilberman 2008).

Further study is needed to determine if non-templated oligouridylation is part of another, independent means of controlling retrotransposons or if this modification is linked to other genome defense mechanisms. We must be cautious about the extent of non-templated nucleotide addition in this class of genes as it has been reported that one member of this class can be extensively edited in undifferentiated stem cells (Osenberg et al. 2010) making accurate alignment and identification of non-templated bases even more challenging.

Why oligouridylation?

I would like to close this work on a philosophical note and speculate as to why oligouridylation may be an ideal way to mark RNA for turnover. Uridine has the lowest energy cost of the common nucleotides to synthesize (Berg 2002) and is a distinguishing feature of RNA. Other than transcription termination by RNAPol-III, oligouridine tracts are not found on properly processed, mature coding or non-coding RNAs. The 3'-ends of small RNA genes transcribed by RNAPol-III are removed during maturation. tRNAs are cleaved to remove the 3'-trailer and then receive the canonical CCA addition. 5S rRNA is trimmed at the 3'-end by exonucleases and is then associated with ribosomal proteins. U6 snRNA undergoes trimming and uridylation as part of its maturation but its 3'-end is protected by 2'-3' cyclization.

New RNAPol-III transcripts are bound by La, a protein with high affinity for RNAs with 3'-uridines (Teplova et al. 2006). La protects RNA from 3'→ 5' degradation until they form mature RNA-protein complexes where the 3'-end can be hidden and protected such as in the case of 7SL, 5s rRNA, and U6 snRNA (Wolin and Cedervall 2002). This suggests another method to enrich for RNA with 3'-uridines—use La to pull down RNA for cloning and sequencing.

As mRNA processing evolved from prokaryotes to eukaryotes, the initial 5'-triphosphate on new RNA transcripts transformed from being a functional group resistant to exonucleases to a danger signal that provokes an immune response (Schlee et al. 2009; Uzri and Gehrke 2009). Perhaps the presence of an oligouridine tail can act as a clear signal to mark an aberrant RNA for degradation.

The key to understanding the function of RNA oligouridylation is to identify the enzyme(s) responsible for this activity on the disparate classes of RNA that are subject to this modification. Many groups have attempted to characterize TUTases with limited success. Only one of the seven putative human TUTases has a clear function ascribed to it (Trippe et al. 2006). Knockdown of two TUTases showed they may be involved with histone mRNA regulation (Mullen and Marzluff 2008) and there appears to be changes to A or U addition to different miRNAs upon knockdown of different TUTases (Wyman et al. 2011). The task of characterizing this family of enzymes with possible functional redundancy may be challenging, but it could reveal the evolutionary history of an ancient RNA regulatory mechanism that still persists within us.

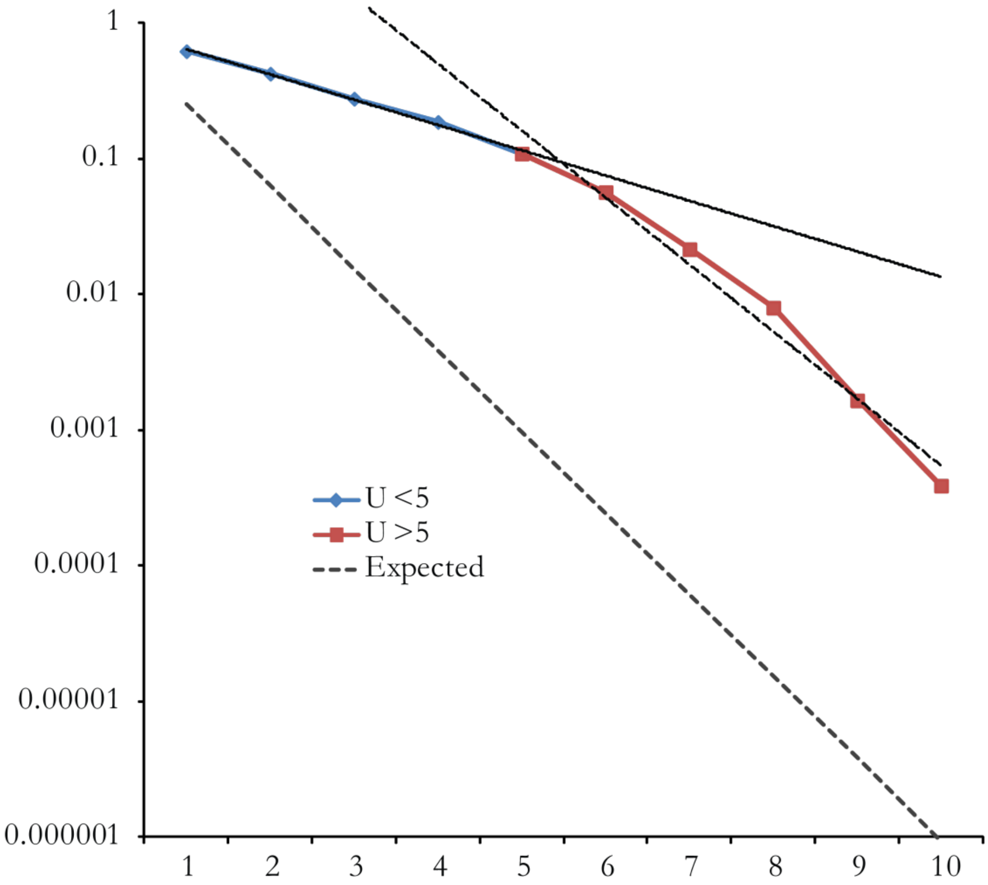


Figure 16. Rate of oligouridylation on intergenic RNA.

The cumulative frequency of non-templated oligouridine THP lengths of unique sequences is shown, that is if an RNA had a THP of five uridines, it was included in the totals for lengths of one through five. The expected line is the probability of finding a homopolymer of a given length at random based on equal representation of all nucleotides. Overlaid are the best fit exponential regression lines for THPs shorter or longer than five nucleotides.

References

Ameres, S.L., Horwich, M.D., Hung, J.H., Xu, J., Ghildiyal, M., Weng, Z., and Zamore, P.D. 2010. Target RNA-directed trimming and tailing of small silencing RNAs. *Science* **328**(5985): 1534-1539.

Armour, C.D., Castle, J.C., Chen, R., Babak, T., Loerch, P., Jackson, S., Shah, J.K., Dey, J., Rohl, C.A., Johnson, J.M. et al. 2009. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat Methods* **6**(9): 647-649.

Arraiano, C.M., Andrade, J.M., Domingues, S., Guinote, I.B., Malecki, M., Matos, R.G., Moreira, R.N., Pobre, V., Reis, F.P., Saramago, M. et al. 2010. The critical role of RNA processing and degradation in the control of gene expression. *FEMS Microbiol Rev* **34**(5): 883-923.

Berg, J.M. 2002. Chapter 25. Nucleotide Biosynthesis. In *Biochemistry 5th Edition*

Bodnar, M.S., Meneses, J.J., Rodriguez, R.T., and Firpo, M.T. 2004. Propagation and maintenance of undifferentiated human embryonic stem cells. *Stem Cells Dev* **13**(3): 243-253.

Bogenhagen, D.F. and Brown, D.D. 1981. Nucleotide sequences in Xenopus 5S DNA required for transcription termination. *Cell* **24**(1): 261-270.

Borowski, L.S., Szczesny, R.J., Brzezniak, L.K., and Stepien, P.P. 2010. RNA turnover in human mitochondria: more questions than answers? *Biochim Biophys Acta* **1797**(6-7): 1066-1070.

Brakenhoff, R.H., Schoenmakers, J.G., and Lubsen, N.H. 1991. Chimeric cDNA clones: a novel PCR artifact. *Nucleic Acids Res* **19**(8): 1949.

Butow, R.A., Zhu, H., Perlman, P., and Conrad-Webb, H. 1989. The role of a conserved dodecamer sequence in yeast mitochondrial gene expression. *Genome* **31**(2): 757-760.

Cavicchioli, R. 2011. Archaea--timeline of the third domain. *Nat Rev Microbiol* **9**(1): 51-61.

Chapman, K.B. and Boeke, J.D. 1991. Isolation and characterization of the gene encoding yeast debranching enzyme. *Cell* **65**(3): 483-492.

Chiang, H.R., Schoenfeld, L.W., Ruby, J.G., Auyeung, V.C., Spies, N., Baek, D., Johnston, W.K., Russ, C., Luo, S., Babiarz, J.E. et al. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev* **24**(10): 992-1009.

Coller, J. and Parker, R. 2004. Eukaryotic mRNA decapping. *Annu Rev Biochem* **73**: 861-890.

Danin-Kreiselman, M., Lee, C.Y., and Chanfreau, G. 2003. RNAse III-mediated degradation of unspliced pre-mRNAs and lariat introns. *Mol Cell* **11**(5): 1279-1289.

Davis, C., Barvish, Z., and Gitelman, I. 2007. A method for the construction of equalized directional cDNA libraries from hydrolyzed total RNA. *BMC Genomics* **8**: 363.

Dieci, G., Preti, M., and Montanini, B. 2009. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics* **94**(2): 83-88.

Dragon, F., Lemay, V., and Trahan, C. 2001. snoRNAs: Biogenesis, Structure and Function. In *eLS*. John Wiley & Sons, Ltd.

Edgar, R., Domrachev, M., and Lash, A.E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**(1): 207-210.

Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W., and Tuschl, T. 2001. Functional anatomy of siRNAs for mediating efficient RNAi in Drosophila melanogaster embryo lysate. *EMBO J* **20**(23): 6877-6888.

England, T.E. and Uhlenbeck, O.C. 1978. Enzymatic oligoribonucleotide synthesis with T4 RNA ligase. *Biochemistry* **17**(11): 2069-2076.

Esnault, C., Heidmann, O., Delebecque, F., Dewannieux, M., Ribet, D., Hance, A.J., Heidmann, T., and Schwartz, O. 2005. APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses. *Nature* **433**(7024): 430-433.

Fenger-Gron, M., Fillman, C., Norrild, B., and Lykke-Andersen, J. 2005. Multiple processing body factors and the ARE binding protein TTP activate mRNA decapping. *Mol Cell* **20**(6): 905-915.

Filipowicz, W. and Pogacic, V. 2002. Biogenesis of small nucleolar ribonucleoproteins. *Curr Opin Cell Biol* **14**(3): 319-327.

Gaur, M., Kamata, T., Wang, S., Moran, B., Shattil, S.J., and Leavitt, A.D. 2006. Megakaryocytes derived from human embryonic stem cells: a genetically tractable system to study megakaryocytopoiesis and integrin function. *J Thromb Haemost* **4**(2): 436-442.

Ghosh, T., Peterson, B., Tomasevic, N., and Peculis, B.A. 2004. Xenopus U8 snoRNA binding protein is a conserved nuclear decapping enzyme. *Mol Cell* **13**(6): 817-828.

Granneman, S. and Baserga, S.J. 2005. Crosstalk in gene expression: coupling and co-regulation of rDNA transcription, pre-ribosome assembly and pre-rRNA processing. *Curr Opin Cell Biol* **17**(3): 281-286.

Grima, D.P., Sullivan, M., Zabolotskaya, M.V., Browne, C., Seago, J., Wan, K.C., Okada, Y., and Newbury, S.F. 2008. The 5'-3' exoribonuclease pacman is required for epithelial sheet sealing in Drosophila and genetically interacts with the phosphatase puckered. *Biol Cell* **100**(12): 687-701.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**(1): 77-88.

Hafner, M., Renwick, N., Brown, M., Mihailovic, A., Holoch, D., Lin, C., Pena, J.T., Nusbaum, J.D., Morozov, P., Ludwig, J. et al. 2011. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* **17**(9): 1697-1712.

Han, J., Hsu, C., Zhu, Z., Longshore, J.W., and Finley, W.H. 1994. Over-representation of the disease associated (CAG) and (CGG) repeats in the human genome. *Nucleic Acids Res* **22**(9): 1735-1740.

Hartmann, R.K., Gossringer, M., Spath, B., Fischer, S., and Marchfelder, A. 2009. The making of tRNAs and more - RNase P and tRNase Z. *Prog Mol Biol Transl Sci* **85**: 319-368.

He, S., Wurtzel, O., Singh, K., Froula, J.L., Yilmaz, S., Tringe, S.G., Wang, Z., Chen, F., Lindquist, E.A., Sorek, R. et al. 2010. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nat Methods* **7**(10): 807-812.

Heinemann, I.U., Soll, D., and Randau, L. 2010. Transfer RNA processing in archaea: unusual pathways and enzymes. *FEBS Lett* **584**(2): 303-309.

Henras, A.K., Soudet, J., Gerus, M., Lebaron, S., Caizergues-Ferrer, M., Mougin, A., and Henry, Y. 2008. The post-transcriptional steps of eukaryotic ribosome biogenesis. *Cell Mol Life Sci* **65**(15): 2334-2359.

Heo, I., Joo, C., Kim, Y.K., Ha, M., Yoon, M.J., Cho, J., Yeom, K.H., Han, J., and Kim, V.N. 2009. TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. *Cell* **138**(4): 696-708.

Ho, C.K., Wang, L.K., Lima, C.D., and Shuman, S. 2004. Structure and mechanism of RNA ligase. *Structure* **12**(2): 327-339.

Horwich, M.D., Li, C., Matranga, C., Vagin, V., Farley, G., Wang, P., and Zamore, P.D. 2007. The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr Biol* **17**(14): 1265-1272.

Hou, Y.M. 2010. CCA addition to tRNA: implications for tRNA quality control. *IUBMB Life* **62**(4): 251-260.

Ibrahim, F., Rymarquis, L.A., Kim, E.J., Becker, J., Balassa, E., Green, P.J., and Cerutti, H. 2010. Uridylation of mature miRNAs and siRNAs by the MUT68 nucleotidyltransferase promotes their degradation in Chlamydomonas. *Proc Natl Acad Sci U S A* **107**(8): 3906-3911.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**(5924): 218-223.

Jacobson, M.R. and Pederson, T. 1998. A 7-methylguanosine cap commits U3 and U8 small nuclear RNAs to the nucleolar localization pathway. *Nucleic Acids Res* **26**(3): 756-760.

Jia, H., Wang, X., Liu, F., Guenther, U.P., Srinivasan, S., Anderson, J.T., and Jankowsky, E. 2011. The RNA helicase Mtr4p modulates polyadenylation in the TRAMP complex. *Cell* **145**(6): 890-901.

Jiao, X., Wang, Z., and Kiledjian, M. 2006. Identification of an mRNA-decapping regulator implicated in X-linked mental retardation. *Mol Cell* **24**(5): 713-722.

Kaczanowska, M. and Ryden-Aulin, M. 2007. Ribosome biogenesis and the translation process in Escherichia coli. *Microbiol Mol Biol Rev* **71**(3): 477-494.

Katoh, T., Sakaguchi, Y., Miyauchi, K., Suzuki, T., Kashiwabara, S., and Baba, T. 2009. Selective stabilization of mammalian microRNAs by 3' adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2. *Genes Dev* **23**(4): 433-438.

Kirino, Y. and Mourelatos, Z. 2007a. The mouse homolog of HEN1 is a potential methylase for Piwi-interacting RNAs. *RNA* **13**(9): 1397-1401.

-. 2007b. Mouse Piwi-interacting RNAs are 2'-O-methylated at their 3' termini. *Nat Struct Mol Biol* **14**(4): 347-348.

Krieg, P.A. and Melton, D.A. 1984. Formation of the 3' end of histone mRNA by post-transcriptional processing. *Nature* **308**(5955): 203-206.

Kriegs, J.O., Churakov, G., Jurka, J., Brosius, J., and Schmitz, J. 2007. Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet* **23**(4): 158-161.

Kuhn, U. and Wahle, E. 2004. Structure and function of poly(A) binding proteins. *Biochim Biophys Acta* **1678**(2-3): 67-84.

Kulpa, D., Topping, R., and Telesnitsky, A. 1997. Determination of the site of first strand transfer during Moloney murine leukemia virus reverse transcription and identification of strand transfer-associated reverse transcriptase errors. *EMBO J* **16**(4): 856-865.

Kwak, J.E. and Wickens, M. 2007. A family of poly(U) polymerases. *RNA* **13**(6): 860-867.

LaCava, J., Houseley, J., Saveanu, C., Petfalski, E., Thompson, E., Jacquier, A., and Tollervey, D. 2005. RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell* **121**(5): 713-724.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M. et al. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**(7): 1401-1414.

Lerat, E. 2010. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* **104**(6): 520-533.

Li, J., Yang, Z., Yu, B., Liu, J., and Chen, X. 2005. Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in Arabidopsis. *Curr Biol* **15**(16): 1501-1507.

Li, J.B., Levanon, E.Y., Yoon, J.K., Aach, J., Xie, B., Leproust, E., Zhang, K., Gao, Y., and Church, G.M. 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**(5931): 1210-1213.

Linsen, S.E., de Wit, E., Janssens, G., Heater, S., Chapman, L., Parkin, R.K., Fritz, B., Wyman, S.K., de Bruijn, E., Voest, E.E. et al. 2009. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* **6**(7): 474-476.

Liu, J., Carmell, M.A., Rivas, F.V., Marsden, C.G., Thomson, J.M., Song, J.J., Hammond, S.M., Joshua-Tor, L., and Hannon, G.J. 2004. Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305**(5689): 1437-1441.

Lund, E. and Dahlberg, J.E. 1992. Cyclic 2',3'-phosphates and nontemplated nucleotides at the 3' end of spliceosomal U6 small nuclear RNA's. *Science* **255**(5042): 327-330.

Macia, A., Munoz-Lopez, M., Cortes, J.L., Hastings, R.K., Morell, S., Lucena-Aguilar, G., Marchal, J.A., Badge, R.M., and Garcia-Perez, J.L. 2011. Epigenetic control of retrotransposon expression in human embryonic stem cells. *Mol Cell Biol* **31**(2): 300-316.

Martin, G. and Keller, W. 2007. RNA-specific ribonucleotidyl transferases. *RNA (New York, NY* **13**(11): 1834-1849.

Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., and Tuschl, T. 2002. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* **110**(5): 563-574.

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl, T. 2004. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell* **15**(2): 185-197.

Mohanty, B.K. and Kushner, S.R. 2000. Polynucleotide phosphorylase functions both as a 3' right-arrow 5' exonuclease and a poly(A) polymerase in Escherichia coli. *Proc Natl Acad Sci U S A* **97**(22): 11966-11971.

Moore, M.J. and Proudfoot, N.J. 2009. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136**(4): 688-700.

Morin, R.D., O'Connor, M.D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M. et al. 2008. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res* **18**(4): 610-621.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**(7): 621-628.

Moss, E.G., Lee, R.C., and Ambros, V. 1997. The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. *Cell* **88**(5): 637-646.

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**(1): 50-65.

Munafo, D.B. and Robb, G.B. 2010. Optimization of enzymatic reaction conditions for generating representative pools of cDNA from small RNA. *RNA* **16**(12): 2537-2552.

Neilson, J.R., Zheng, G.X., Burge, C.B., and Sharp, P.A. 2007. Dynamic regulation of miRNA expression in ordered stages of cellular development. *Genes Dev* **21**(5): 578-589.

Newbury, S. and Woollard, A. 2004. The 5'-3' exoribonuclease xrn-1 is essential for ventral epithelial enclosure during C. elegans embryogenesis. *RNA* **10**(1): 59-65.

Newman, M.A., Mani, V., and Hammond, S.M. 2011. Deep sequencing of microRNA precursors reveals extensive 3' end modification. *RNA* **17**(10): 1795-1803.

O'Donnell, K.A. and Boeke, J.D. 2007. Mighty Piwis defend the germline against genome intruders. *Cell* **129**(1): 37-44.

O'Hara, E.B., Chekanova, J.A., Ingle, C.A., Kushner, Z.R., Peters, E., and Kushner, S.R. 1995. Polyadenylylation helps regulate mRNA decay in Escherichia coli. *Proc Natl Acad Sci U S A* **92**(6): 1807-1811.

Okamura, K., Hagen, J.W., Duan, H., Tyler, D.M., and Lai, E.C. 2007. The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell* **130**(1): 89-100.

Okorokov, A.L., Panov, K.I., Offen, W.A., Mukhortov, V.G., Antson, A.A., Karpeisky, M., Wilkinson, A.J., and Dodson, G.G. 1997. RNA cleavage without hydrolysis. Splitting the catalytic activities of binase with Asn101 and Thr101 mutations. *Protein Eng* **10**(3): 273-278.

Olena, A.F. and Patton, J.G. 2010. Genomic organization of microRNAs. *J Cell Physiol* **222**(3): 540-545.

Ong, K.K., Elks, C.E., Li, S., Zhao, J.H., Luan, J., Andersen, L.B., Bingham, S.A., Brage, S., Smith, G.D., Ekelund, U. et al. 2009. Genetic variation in LIN28B is associated with the timing of puberty. *Nat Genet* **41**(6): 729-733.

Osenberg, S., Paz Yaacov, N., Safran, M., Moshkovitz, S., Shtrichman, R., Sherf, O., Jacob-Hirsch, J., Keshet, G., Amariglio, N., Itskovitz-Eldor, J. et al. 2010. Alu sequences in undifferentiated human embryonic stem cells display high levels of A-to-I RNA editing. *PLoS One* **5**(6): e11173.

Ozsolak, F. and Milos, P.M. 2011. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**(2): 87-98.

Pak, J. and Fire, A. 2007. Distinct populations of primary and secondary effectors during RNAi in C. elegans. *Science* **315**(5809): 241-244.

Perumal, K. and Reddy, R. 2002. The 3' end formation in small RNAs. *Gene Expr* **10**(1-2): 59-78.

Peterlin, B.M. and Price, D.H. 2006. Controlling the elongation phase of transcription with P-TEFb. *Mol Cell* **23**(3): 297-305.

Pfeffer, S., Lagos-Quintana, M., and Tuschl, T. 2005. Cloning of small RNA molecules. *Curr Protoc Mol Biol* **Chapter 26**: Unit 26 24.

Phizicky, E.M. and Hopper, A.K. 2010. tRNA biology charges to the front. *Genes Dev* **24**(17): 1832-1860.

Reddy, R., Henning, D., and Busch, H. 1985. Primary and secondary structure of U8 small nuclear RNA. *J Biol Chem* **260**(20): 10930-10935.

Rino, J. and Carmo-Fonseca, M. 2009. The spliceosome: a self-organized macromolecular machine in the nucleus? *Trends Cell Biol* **19**(8): 375-384.

Rissland, O.S., Mikulasova, A., and Norbury, C.J. 2007. Efficient RNA polyuridylation by noncanonical poly(A) polymerases. *Mol Cell Biol* **27**(10): 3612-3624.

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**(6): 616-623.

Ruby, J.G., Jan, C.H., and Bartel, D.P. 2007. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**(7149): 83-86.

Saito, K., Sakaguchi, Y., Suzuki, T., Siomi, H., and Siomi, M.C. 2007. Pimet, the Drosophila homolog of HEN1, mediates 2'-O-methylation of Piwi- interacting RNAs at their 3' ends. *Genes Dev* **21**(13): 1603-1608.

San Paolo, S., Vanacova, S., Schenk, L., Scherrer, T., Blank, D., Keller, W., and Gerber, A.P. 2009. Distinct roles of non-canonical poly(A) polymerases in RNA metabolism. *PLoS Genet* **5**(7): e1000555.

Schlee, M., Hartmann, E., Coch, C., Wimmenauer, V., Janke, M., Barchet, W., and Hartmann, G. 2009. Approaching the RNA ligand for RIG-I? *Immunol Rev* **227**(1): 66-74.

Schmid, M. and Jensen, T.H. 2008. The exosome: a multipurpose RNA-decay machine. *Trends Biochem Sci* **33**(10): 501-510.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. 2008. Divergent transcription from active promoters. *Science* **322**(5909): 1849-1851.

Shen, B. and Goodman, H.M. 2004. Uridine addition after microRNA-directed cleavage. *Science* **306**(5698): 997.

Singer-Sam, J. 1994. Quantitation of specific transcripts by RT-PCR SNuPE assay. *PCR Methods Appl* **3**(4): S48-50.

Singer-Sam, J., LeBon, J.M., Dai, A., and Riggs, A.D. 1992. A sensitive, quantitative assay for measurement of allele-specific transcripts differing by a single nucleotide. *PCR Methods Appl* **1**(3): 160-163.

Singh, R. and Reddy, R. 1989. Gamma-monomethyl phosphate: a cap structure in spliceosomal U6 small nuclear RNA. *Proc Natl Acad Sci U S A* **86**(21): 8280-8283.

Sinha, K.M., Gu, J., Chen, Y., and Reddy, R. 1998. Adenylation of small RNAs in human cells. Development of a cell-free system for accurate adenylation on the 3'-end of human signal recognition particle RNA. *J Biol Chem* **273**(12): 6853-6859.

Slomovic, S., Fremder, E., Staals, R.H., Pruijn, G.J., and Schuster, G. 2010. Addition of poly(A) and poly(A)-rich tails during RNA degradation in the cytoplasm of human cells. *Proc Natl Acad Sci U S A* **107**(16): 7407-7412.

Smith, A.M., Heisler, L.E., St Onge, R.P., Farias-Hesson, E., Wallace, I.M., Bodeau, J., Harris, A.N., Perry, K.M., Giaever, G., Pourmand, N. et al. 2010. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res* **38**(13): e142.

Song, M.G. and Kiledjian, M. 2007. 3' Terminal oligo U-tract-mediated stimulation of decapping. *Rna* **13**(12): 2356-2365.

Speckmann, W.A., Terns, R.M., and Terns, M.P. 2000. The box C/D motif directs snoRNA 5'-cap hypermethylation. *Nucleic Acids Res* **28**(22): 4467-4473.

Suh, M.R., Lee, Y., Kim, J.Y., Kim, S.K., Moon, S.H., Lee, J.Y., Cha, K.Y., Chung, H.M., Yoon, H.S., Moon, S.Y. et al. 2004. Human embryonic stem cells express a unique set of microRNAs. *Dev Biol* **270**(2): 488-498.

Tang, F., Hajkova, P., Barton, S.C., O'Carroll, D., Lee, C., Lao, K., and Surani, M.A. 2006. 220-plex microRNA expression profile of a single cell. *Nat Protoc* **1**(3): 1154-1159.

Teplova, M., Yuan, Y.R., Phan, A.T., Malinina, L., Ilin, S., Teplov, A., and Patel, D.J. 2006. Structural basis for recognition and sequestration of UUU(OH) 3' temini of nascent RNA polymerase III transcripts by La, a rheumatic disease autoantigen. *Mol Cell* **21**(1): 75-85.

Terns, M.P., Grimm, C., Lund, E., and Dahlberg, J.E. 1995. A common maturation pathway for small nucleolar RNAs. *EMBO J* **14**(19): 4860-4871.

Terns, M.P., Lund, E., and Dahlberg, J.E. 1992. 3'-end-dependent formation of U6 small nuclear ribonucleoprotein particles in Xenopus laevis oocyte nuclei. *Mol Cell Biol* **12**(7): 3032-3040.

Trippe, R., Guschina, E., Hossbach, M., Urlaub, H., Luhrmann, R., and Benecke, B.J. 2006. Identification, cloning, and functional analysis of the human U6 snRNA-specific terminal uridylyl transferase. *RNA* **12**(8): 1494-1504.

Tyc, K. and Steitz, J.A. 1989. U3, U8 and U13 comprise a new class of mammalian snRNPs localized in the cell nucleolus. *EMBO J* **8**(10): 3113-3119.

Uzri, D. and Gehrke, L. 2009. Nucleotide sequences and modifications that determine RIG-I/RNA binding and signaling activities. *J Virol* **83**(9): 4174-4184.

van Hoof, A., Lennertz, P., and Parker, R. 2000. Three conserved members of the RNase D family have unique and overlapping functions in the processing of 5S, 5.8S, U4, U5, RNase MRP and RNase P RNAs in yeast. *EMBO J* **19**(6): 1357-1365.

Vanacova, S., Wolf, J., Martin, G., Blank, D., Dettwiler, S., Friedlein, A., Langen, H., Keith, G., and Keller, W. 2005. A new yeast poly(A) polymerase complex involved in RNA quality control. *PLoS Biol* **3**(6): e189.

von der Haar, T. 2008. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Syst Biol* **2**: 87.

Wang, M. and Pestov, D.G. 2011. 5'-end surveillance by Xrn2 acts as a shared mechanism for mammalian pre-rRNA maturation and decay. *Nucleic Acids Res* **39**(5): 1811-1822.

Wang, Z., Gerstein, M., and Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**(1): 57-63.

Warner, J.R. 1999. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci* **24**(11): 437-440.

West, S., Gromak, N., Norbury, C.J., and Proudfoot, N.J. 2006. Adenylation and exosome-mediated degradation of cotranscriptionally cleaved pre-messenger RNA in human cells. *Mol Cell* **21**(3): 437-443.

Wlotzka, W., Kudla, G., Granneman, S., and Tollervey, D. 2011. The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. *EMBO J* **30**(9): 1790-1803.

Wolin, S.L. and Cedervall, T. 2002. The La protein. *Annu Rev Biochem* **71**: 375-403.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.C., Dufour, M.E., Boulay, J., Regnault, B., Devaux, F., Namane, A., Seraphin, B. et al. 2005. Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121**(5): 725-737.

Wyman, S.K., Knouf, E.C., Parkin, R.K., Fritz, B.R., Lin, D.W., Dennis, L.M., Krouse, M.A., Webster, P.J., and Tewari, M. 2011. Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res* **21**(9): 1450-1461.

Xu, F. and Cohen, S.N. 1995. RNA degradation in Escherichia coli regulated by 3' adenylation and 5' phosphorylation. *Nature* **374**(6518): 180-183.

Yang, J.S. and Lai, E.C. 2011. Alternative miRNA Biogenesis Pathways and the Interpretation of Core miRNA Pathway Mutants. *Mol Cell* **43**(6): 892-903.

Yang, W. 2011. Nucleases: diversity of structure, function and mechanism. *Q Rev Biophys* **44**(1): 1-93.

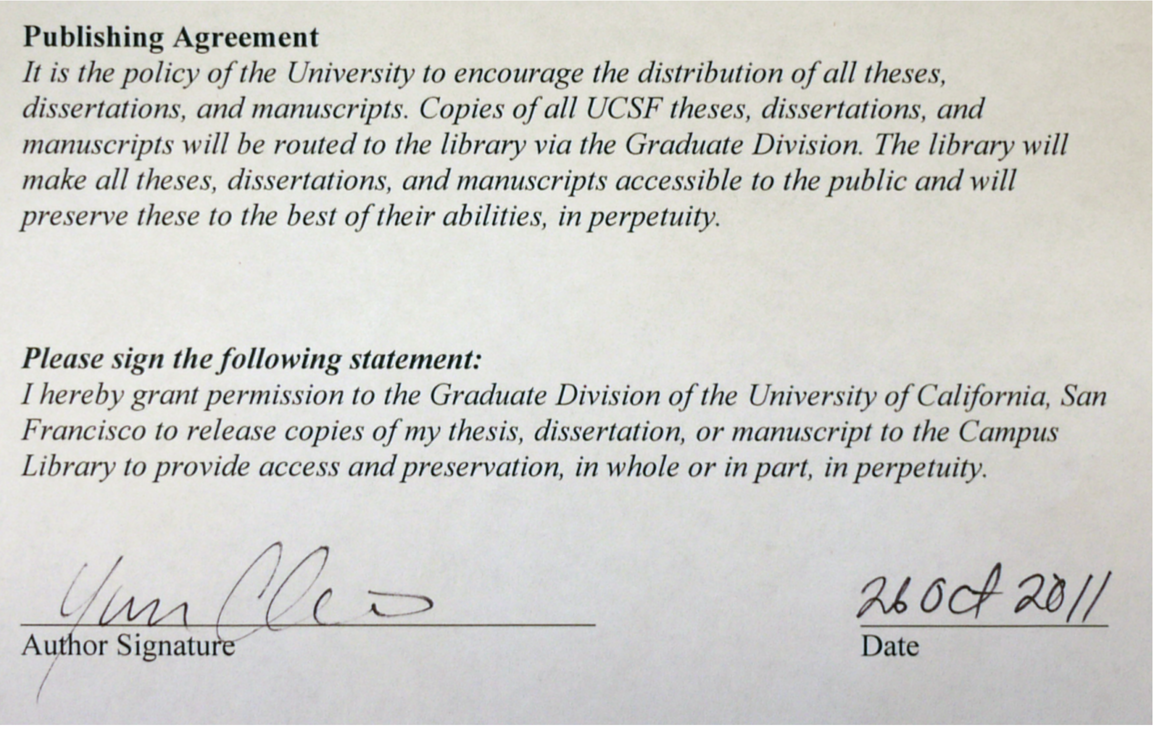
Yarus, M. 2011. Getting past the RNA world: the initial Darwinian ancestor. *Cold Spring Harb Perspect Biol* **3**(4).

Yu, B., Yang, Z., Li, J., Minakhina, S., Yang, M., Padgett, R.W., Steward, R., and Chen, X. 2005. Methylation as a crucial step in plant microRNA biogenesis. *Science* **307**(5711): 932-935.

Zhu, H., Shah, S., Shyh-Chang, N., Shinoda, G., Einhorn, W.S., Viswanathan, S.R., Takeuchi, A., Grasemann, C., Rinn, J.L., Lopez, M.F. et al. 2010. Lin28a transgenic mice manifest size and puberty phenotypes identified in human genetic association studies. *Nat Genet* **42**(7): 626-630.

Zilberman, D. 2008. The evolving functions of DNA methylation. *Curr Opin Plant Biol* **11**(5): 554-559.

Zimmer, S.L., Schein, A., Zipor, G., Stern, D.B., and Schuster, G. 2009. Polyadenylation in Arabidopsis and Chlamydomonas organelles: the input of nucleotidyltransferases, poly(A) polymerases and polynucleotide phosphorylase. *Plant J* **59**(1): 88-99.



Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Author Signature

26 Oct 2011

Date