2015-10-06

Optimizing RNA Library Preparation to Redefine the Translational Status of 80S Monosomes: A Dissertation

Erin E. Heyer
University of Massachusetts Medical School

Optimizing RNA Library Preparation
to
Redefine the Translational Status of 80S Monosomes

A Dissertation Presented

By

Erin E. Heyer

Submitted to the Faculty of the
University of Massachusetts
Graduate School of Biomedical Sciences, Worcester
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

October 6, 2015

BIOCHEMISTRY

Optimizing RNA Library Preparation
to
Redefine the Translational Status of 80S Monosomes

A Dissertation Presented

By

Erin E. Heyer

The signatures of the Dissertation Defense Committee signify completion and approval as to style and content of the Dissertation

Melissa J. Moore, Ph.D., Thesis Advisor

Sean P. Ryder, Ph.D., Member of Committee

Phillip D. Zamore, Ph.D., Member of Committee

Christopher V. Nicchitta, Ph.D., Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee

Allan Jacobson, Ph.D., Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the school.

Anthony Carruthers, Ph.D.

Dean of the Graduate School of Biomedical Sciences

Biochemistry and Molecular Pharmacology

October 6, 2015

For Mem

and

For Uncle Dave

Forever missed, forever loved.

Acknowledgements

My interest in science started early, and I was lucky enough to have a fantastic high school biology teacher who encouraged my endless questioning and would hunt down the answers once I'd stumped him. Mr. William Warren, thank you for inspiring my interest in biology.

During my summer breaks from college, I worked at the University of Vermont in Kenneth Mann's lab, under the supervision of Matthew Whelihan and Professors Kathleen Brummel-Ziedins and Thomas Orfeo. Thank you for a great introduction to research science, and for teaching me how to pipette correctly - your instructions have served me well.

As an undergraduate at Harvey Mudd College, I was incredibly fortunate to work in David Asai's lab (the Asailum!). Learning confocal microscopy as a sophmore in college and using it to studying micronuclear division in *Tetrahymena thermophila* got me hooked on research, and I stayed in the lab for my thesis work. A rarity at an undergraduate-only college, we had two wonderful postdocs in the lab - David Wilkes and Vidyalakshmi Rajagopalan - who kept everything running and made the Asailum a great place to be. David, David and Vidya, thank you for all the opportunities and for everything you taught me.

Coming to graduate school directly from a small college, I had no idea what I wanted to study. After hearing Melissa speak at the Wood's Hole retreat, I still didn't know what I wanted to study, but I knew it would be in her lab. Melissa, thank you for the opportunity to learn so much from you.

Many thanks to my committee - Allan Jacobson, Oliver Rando, Sean Ryder and Phillip Zamore - for your guidance through graduate school. You pushed me to think outside my comfort zone and it helped immensely.

When first rotating in Melissa's lab, I was paired with Guramrit Singh. An outstanding postdoc, Amrit became a benchmate, educator, mentor and friend throughout our years together in the lab. Amrit, I wouldn't be the scientist I am today without your thoughtful teaching and guidance; thank you.

Hakan and Emiliano, thanks for your help with OmniPrep; it was fun! Luis and Francois, it was great working with you on the Huntington's project.

Academics aside, the Moore lab is filled with truly wonderful individuals who make it fun to be in the lab. Special thanks to Akiko, Alicia, Ami, Andrew, Blandine, Carrie, and Emiliano for always being up for a chat and a good laugh. Chris and Eric, I couldn't imagine 2 greater guys to go through grad school with - thanks for everything.

My graduate work came to a bit of a standstill when I needed to learn bioinformatics in order to analyze my monosome footprinting data. It would have taken me much longer to develop the necessary analytical skills without a very generous offer from John Mattick to learn bioinformatics from his lab at the Garvan Institute in Sydney, Australia. Many thanks to John and his research group, especially Martin Smith, for hosting me and for everything you taught me.

On a personal note, I'm incredibly lucky to have the never-ending support of my truly awesome family. Mom, Papa, Kathleen, Dave and Aunt Cin; thanks for always being there. To Anna and Bert: life wouldn't be the same without you.

Finally, and most importantly, thank you to my husband Bryan for your love and unending support. Thank you for your understanding; thank you for your encouragement; thank you for believing in me.

UNIVERSITY OF MASSACHUSETTS MEDICAL SCHOOL

Abstract

BIOCHEMISTRY AND MOLECULAR PHARMACOLOGY

Doctor of Philosophy

Optimizing RNA Library Preparation
to
Redefine the Translational Status of 80S Monosomes

by Erin E. Heyer

Deep sequencing of strand-specific cDNA libraries is now a ubiquitous tool for identifying and quantifying RNAs in diverse sample types. The accuracy of conclusions drawn from these analyses depends on precise and quantitative conversion of the RNA sample into a DNA library suitable for sequencing. Here, we describe an optimized method of preparing strand-specific RNA deep sequencing libraries from small RNAs and variably sized RNA fragments obtained from ribonucleoprotein particle footprinting experiments or fragmentation of long RNAs. Because all enzymatic reactions were optimized and driven to apparent completion, sequence diversity and species abundance in the input sample are well preserved.

This optimized method was used in an adapted ribosome-profiling approach to sequence mRNA footprints protected either by 80S monosomes or polysomes in *S. cerevisiae*. Contrary to popular belief, we show that 80S monosomes are translationally active as demonstrated by strong three-nucleotide phasing of monosome footprints across open reading frames. Most mRNAs exhibit some degree of monosome occupancy, with monosomes predominating on upstream ORFs, canonical ORFs shorter than ~590 nucleotides and any ORF for which the total time required to complete elongation is substantially shorter than the time required for initiation. Additionally, endogenous NMD targets tend to be monosome-enriched. Thus, rather than being inactive, 80S monosomes are significant contributors to overall cellular translation.

Contents

List of Figures

List of Tables

bp	Base pair, usually of DNA
CDS	Coding sequence
DNA	Deoxyribonucleic acid
ddNTP	Dideoxynucleoside triphosphate
dNTP	Deoxynucleoside triphosphate
ER	Endoplasmic reticulum
HTS	High-throughput sequencing (see also NGS)
kb	kilo-base of RNA or DNA (in nt)
miRNA	microRNA, a small non-coding RNA
mRNA	Messenger RNA
mRNP	Messenger ribonucleoprotein particle
NGS	Next-generation sequencing
NMD	Nonsense-mediated mRNA decay
nt	A nucleotide of either DNA or RNA
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
PIC	Pre-initiation complex
RNA	Ribonucleic acid
RT	Reverse transcription
rRNA	Ribosomal RNA
ssDNA	Single-stranded DNA
ssRNA	Single-stranded RNA
tRNA	Transfer RNA
uORF	Upstream open reading frame

A-site Acceptor site: the site on the 80S ribosomal subunit that is mostly occupied by aminoacyl-tRNA waiting to accept the elongating polypeptide

barcode A nucleotide sequence uniquely identifying the sample each read originated from

codon A set of 3 nucleotides encoding a single amino acid

coverage A measure of the number of times each nt of a genome is sequenced

E-site Exit site: the site on the 80S ribosomal subunit where a tRNA, now deacylated, is held before release from the ribosome.

footprint The fragment of RNA protected from RNase digestion by a bound protein or protein complex

insert The RNA molecule captured between two adaptor sequences in high-throughput sequencing library construction

monosome A single 80S ribosome, either bound to a mRNA or free

P-site Peptidyl site: the site on the 80S ribosomal subunit that is mostly occupied by the tRNA linked to the growing polypeptide

paired-end A deep sequencing run, identifying the DNA sequence from either side of a DNA template

polysome Also known as a polyribosome, this structure consists of a single mRNA bound by multiple ribosomes

polysome profiling The process of sedimenting cellular lysate through a sucrose gradient to separate mRNAs based on the number of bound ribosomes

read The sequence of nucleotides produced from each spot on a high-throughput sequencing machine

read depth The number of reads in each library obtained from high-throughput sequencing analysis

read length The number of nucleotides in each read; often corresponds to the number of cycles in the sequencing reaction

RNA-Seq A technology wherein RNA is fragmented, converted to DNA, and analyzed on a high-throughput sequencing instrument

single-end sequencing A deep sequencing run, originating only from one side of a DNA template

transcriptome The complete set of cellular transcripts

uORF An ORF located within the 5' transcript leader, upstream of the canonical start codon

Preface

Some work reported in this dissertation has been published elsewhere:

Chapter II has been accepted for publication at Cell.

Chapter III has been published previously as:

> *Heyer, E. E., Özadam, H., Ricci, E. P., Cenik, C., and Moore, M. J. (2015). An optimized kit-free method for making strand-specific deep sequencing libraries from RNA fragments.* Nucleic Acids Research*, 43(1):e2*

Additional work studying the translation of huntingtin mRNA was performed during thesis studies but will not be presented in this thesis. The majority of this work involved method optimization for (a) polysome profiling from mouse brain and (b) quantitative PCR detection of WT and mutant huntingtin mRNA. The result of this work is a functional method for future experiments, but the system in which the work was performed did not allow for biologically relevant conclusions, and therefore the work will not be discussed here.

Chapter I

Introduction

Translation

Gene expression is the cellular process by which genetic information is transformed into a functional product. Genetic information - DNA - is transcribed into RNA molecules, which are either a functional molecule on their own (ribosomal RNA, transfer RNA, micro RNA, etc.) or an intermediary to a final protein product (messenger RNA). In translation, each mRNA molecule is decoded by a ribosome to assemble amino acids into a polypeptide chain. As the polypeptide emerges from the ribosome, it folds into a functional protein. Through this process, the ribosome is responsible for the creation of the proteome, the collection of all proteins expressed in a cell. Translation is a highly regulated process, as cell and tissue growth depend on protein synthesis. Our understanding of gene expression, especially the translation pathway, began to blossom with the discovery of DNA.

Discovery of Translation Machinery and Polysomes

After the discovery of the DNA double helix in 1953 (Watson and Crick, 1953), a considerable amount of scientific research focused on identifying the relationship between genes and proteins, since DNA’s ability to direct protein synthesis was not immediately apparent from its structure. By the end of the decade, a great deal of evidence pointed to RNAs playing major roles in this process, and the "central dogma" of molecular biology was published by Francis Crick in 1958 (Crick, 1958). This sequence hypothesis accurately described the flow of genetic information between the 3 major classes of polymeric biomolecules (RNA and DNA are polynucleotides; protein is a polypeptide), stating that "once 'information' has passed into protein *it cannot get out again*." Impressively, this hypothesis proved to be quite accurate, including the prediction of reverse transcription.

The first observations of ribosomes as dense granules in electron microscopy images were made by George Emil Palade (Palade, 1955), work for which he would share the 1974 Nobel Prize in Physiology or Medicine. He found these particles both in the cytoplasm and attached to endoplasmic reticulum membranes and identified their substance as ribonucleoprotein (Palade, 1955). Subsequent work isolated both cytoplasmic and membrane-bound ribosomes to identify these structures as the site of protein synthesis (Kirsch et al., 1960). Other components of the translational machinery, such as transfer RNA (tRNA), had been discovered a few years earlier. Then called soluble RNA or sRNA, tRNA was shown to be the link between the 3-nt nucleic acid code and the amino-acid building blocks of protein (Crick, 1958, Hoagland et al., 1958). The discovery of messenger RNA (mRNA) as the template for protein synthesis (Brenner et al., 1961, Gros et al., 1961) was the final piece of the puzzle, a finding that came a few years after an RNA template was proposed by Francis Crick (Crick, 1958).

However, the mechanisms controlling how these components came together to synthesize proteins - and their relative proportions - were still unknown. Work from James Watson's lab (Risebrough et al., 1962) described ribosomes sedimenting more rapidly than the canonical *Escherichia coli* 70S ribosome, suggesting that the addition of mRNA to the complex caused this increased sedimentation rate. However, the salt conditions used in these preparations also drove the assembly of 100S ribosome complexes. Based on this data, it was proposed that these 100S ribosomes might be the "principal sites of protein synthesis" (Risebrough et al., 1962). Work in Alexander Rich's lab by Jonathan Warner and Paul Knopf (and at the same time by the labs of Alfred Gierer and

Hans Noll) used a recently published technique of separating large molecules through sucrose gradients (Britten and Roberts, 1960) to analyze translation on nascent hemoglobin mRNAs in rabbit reticulocyte lysate by incorporating radioactively-labelled amino acid into the growing polypeptide (Warner et al., 1963). They found that the vast majority of incorporated radioactive signal corresponded to rapidly sedimenting peaks indicative of multiple ribosomes. The name 'polyribosome' - or polysome for short - was assigned to these structures. Additional experiments showed that polysomes are connected by RNA, as the addition of RNase, but not DNase, moved the signal from heavy peaks to light peaks (Warner et al., 1963). Experiments in Hans Noll's lab specifically identified the connector RNA as mRNA by isolating polysomes from the livers of rats injected with actinomycin, an antibiotic that inhibits mRNA synthesis (Staehelin et al., 1963). In a time- and concentration- dependent manner, ribosome aggregates were converted into 80S monomers due to the lack of mRNA template molecules. Electron micrographs confirmed the polyribosomal structure as "extended arrays of ribosomes connected by a strand approximately 10 A in width" (Slayter et al., 1963).

The discovery that multiple ribosomes can assemble on a single mRNA molecule was a huge step forward for the field of translation, but the relative ratios of tRNA and polypeptide per ribosome remained unclear. It was known that the polypeptide chain was covalently linked to a tRNA, and the first proposed translation mechanism suggested that only a single tRNA was bound by the ribosome, the polypeptide chain remaining with the ribosome through transference from one tRNA to the next (Cannon et al., 1963). However, experimental evidence disproved this, demonstrating that each actively translating ribosome in a polysome binds two tRNAs (Warner and Rich, 1964). This was measured by radioactively-labeling tRNAs and quantifying the tRNA signal in polysomes relative to the known number of ribosomes in the polysome. Consequently, Warner and Rich concluded that each ribosome had two tRNA-binding sites. This was ultimately disproven by work from the labs of Hans Knoll (Wettstein and Noll, 1965) and Knud Nierhaus (Rheinberger et al., 1981) demonstrating that each ribosome has three tRNA binding sites.

In a relatively short amount of time, the individual components of the translation machinery and the general manner in which they all worked together were discovered. Building on this body of knowledge, many of the proteins involved in translation initiation were discovered in the 1970s, using classical biochemical techniques to isolate and reconstitute *in vitro* the initiation process (Fraser, 2015). Since then, technical and methodological developments have led to many discoveries which together inform our current, high resolution structural understanding of the ribosome and its many accessory factors.

Translation Cycle

The translation cycle comprises four stages: initiation, elongation, termination and ribosome recycling. The canonical translational mechanism is cap-dependent and will be described in detail below.

The first translational stage is initiation, in which a ribosome assembles on the first codon to be translated. The accuracy of initiation site selection is highly regulated and crucially important, as this determines which reading frame will be used to translate the mRNA. Before any interaction with the mRNA, however, initiation begins with the formation of a 43S pre-initiation complex (PIC). This structure is composed of a variety of initiation factors (eIF1, eIF1A, multisubunit eIF3, eIF5, and the ternary complex eIF2-GTP-tRNA$_{\mathrm{met}}$), which mutually stabilize their binding to the 40S small ribosomal subunit (Fraser, 2015, Hinnebusch, 2014, Jackson et al., 2010). These interactions slightly alter the conformation of the 40S subunit to encourage an open-state PIC, where tRNA$_{\mathrm{met}}$ is not fully seated in the P-site, and the subunit can scan along an mRNA (Llácer et al., 2015). The PIC then binds to an mRNA near the 5'-7-methylguanosine (m^{7}G) cap with the assistance of cap-binding complex eIF4F (cap-binding protein eIF4E, scaffold protein eIF4G, and RNA helicase eIF4A), eIF4B, and poly(A)-binding protein. Once bound, the PIC scans down the 5' UTR until the start codon (initiator AUG) is recognized. As this recognition occurs, the 40S subunit adopts a closed conformation, blocking its migration along the mRNA, and the $\mathrm{tRNA}_{\mathrm{met}}$ engages fully in the P-site, forming a 48S-initiation complex. As initiation continues, several initiation factors are released and the large 60S subunit docks, catalyzed by eIF5B (Hinnebusch and Lorsch, 2012). This forms a complete 80S initiation complex, containing a mRNA, initiator $\mathrm{tRNA}^{\mathrm{met}}$ base-paired to AUG in the P-site, and an empty A site containing the 2nd codon in the ORF (Figure 1.1). With the recent development of ribosome profiling (see I for details), the precise mRNA location of a ribosome can be mapped. This has demonstrated that at initiation, the 5' and 3' edges of the *S. cerevisiae* ribosome protect 12 nts upstream and 12-13 nts downstream of the start codon.

This 80S structure is now poised for elongation, the second phase of translation, in which the remainder of the codons are decoded (Figure 1.2). First, an aminoacylated elongator-tRNA is delivered to the A-site in complex with eEF1A. If the tRNA anticodon is able to stably base-pair with the mRNA codon in the A-site, eEF1A is released through GTP hydrolysis and the tRNA becomes fully seated in the A-site (Dever and Green, 2012, Voorhees and Ramakrishnan, 2013). To extend the peptide chain, the methionine delivered by the initiation complex must form a peptide bond with the newly delivered amino acid. This bond forms through a nucleophilic attack on peptidyl tRNA by aminoacyl tRNA, catalyzed by the ribosomal peptidyl transferase center bringing these reactive groups into close proximity.

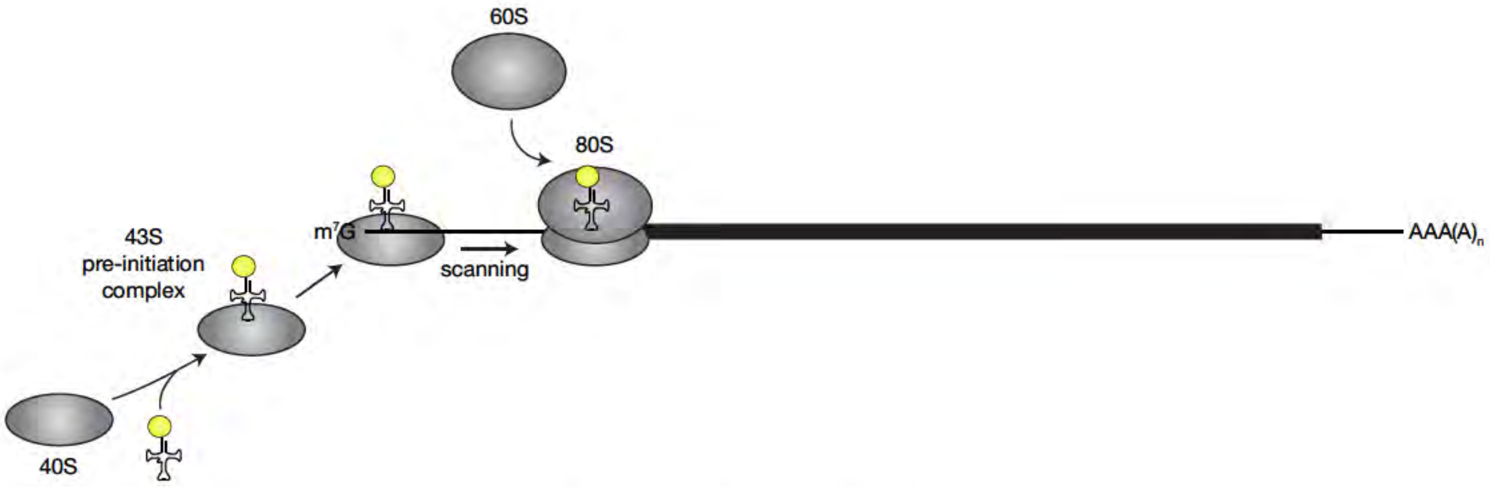


FIGURE 1.1: Translation Initiation

A 43S pre-initiation complex (PIC) forms to load a 40S ribosomal subunit with initiator tRNA$_{met}$. The PIC then binds to an mRNA near the 5'-7-methylguanosine (m^{7}G) cap and scans down the 5' UTR until the start codon (initiator AUG) is recognized. As this recognition occurs, the 40S subunit adopts a closed conformation to prevent migration along the mRNA, and the tRNA$_{met}$ engages fully in the P-site, forming a 48S-initiation complex. As initiation continues, the 60S subunit docks and forms a complete 80S initiation complex, containing a mRNA, initiator tRNA^{met} base-paired to AUG in the P-site, and an empty A site containing the 2nd codon in the ORF.

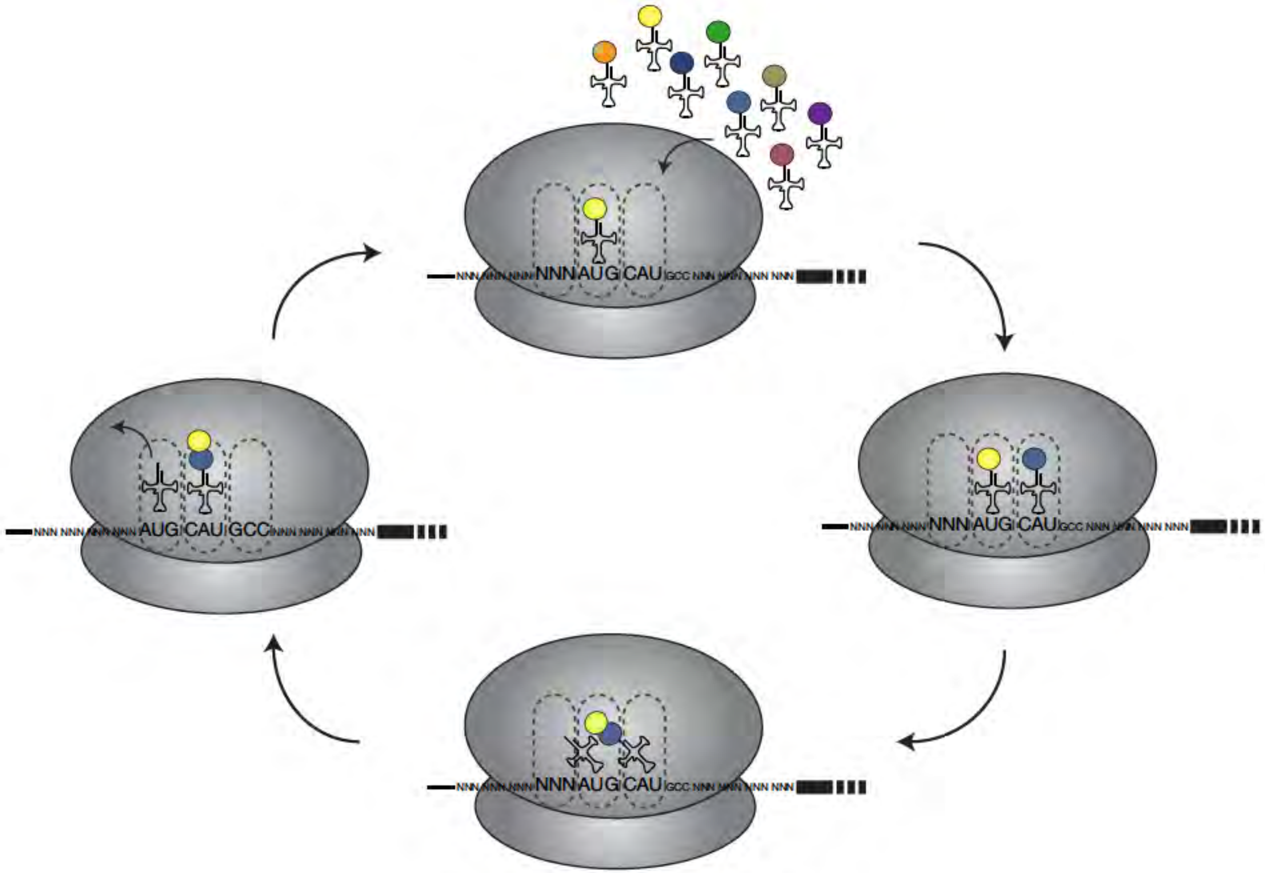

NNN NNN NNN NNN AUG CAU GCC NNN NNN NNN NNN
NNN NNN NNN NNN AUG CAU GCC NNN NNN NNN NNN
NNN NNN NNN NNN AUG CAU GCC NNN NNN NNN NNN
NNN NNN NNN NNN AUG CAU GCC NNN NNN NNN NNN

FIGURE 1.2: Elongation Cycle

An aminoacylated tRNA is delivered to the A-site and base-pairs with the A-site anticodon. The acceptor end of the A-site tRNA moves into the P-site and receives the P-site amino acid. Once this peptide bond forms, a large-scale conformational rearrangement results in (a) the transference of the polypeptide-carrying tRNA to the P-site and the deacylated tRNA to the E-site and (b) an exact 3-codon movement of the mRNA to bring the next codon into the A-site. After the E-site tRNA dissociates, another round of peptide-bond formation can occur.

Once the peptide has been transferred to the A-site tRNA, the ribosome undergoes a large-scale conformational change where the individual subunits rotate relative to each other in a ratchet motion (Agirrezabala et al., 2008, Frank and Agrawal, 2000). This structural rearrangement results in (a) the transference of the polypeptide-carrying tRNA to the P-site and the deacylated tRNA to the E-site and (b) an exact 3-codon movement of the mRNA to bring the next codon into the A-site. This rearrangement occurs via a hybrid state, where the anticodon ends of the bound tRNAs remains in the P- and A- sites, but the acceptor ends move to the E- and P- sites, respectively (Moazed and Noller, 1989, Noeske and Cate, 2012). Elongation factor 2 catalyzes translocation by stabilizing this rotated conformation, and the tRNAs move into the canonical E- and P- sites. The ribosome now exists in a posttranslocation state, which recent work has found to deviate slightly from the classic pretranslation state of the ribosome, suggesting that additional structural rearrangements must be made before another elongation cycle can occur (Budkevich et al., 2014). After the E-site tRNA spontaneously dissociates and the ribosomal subunits "roll" back to their pretranslation state, another round of peptide-bond formation can occur.

Elongation occurs over and over until the ribosome encounters a stop codon (UAA, UGA, or UAG) in the A-site. At this point, the third phase of translation - termination - occurs. Two release factors (eRF1 and eRF3) work together to catalyze termination (Dever and Green, 2012, Skabkin et al., 2013). A class

I factor, eRF1 is a tRNA-shaped protein that recognizes the stop codon and activates release of the synthesized protein via peptide hydrolysis (Brown et al., 2015). Additionally, eRF1 interacts with eRF3, a class II release factor that accelerates polypeptide release.

The final step in translation is ribosome recycling, which connects translation termination and initiation. Following polypeptide release, the mRNA and deacylated tRNA remain bound to the 80S ribosome; all components must be freed from this complex to restore the individual components required for subsequent rounds of translation (Dever and Green, 2012, Nürenberg and Tampé, 2013). The ATP-binding cassette ABCE1 dissociates posttermination complexes into free 60S and 40S subunits, the latter still associated with mRNA and tRNA (Pisarev et al., 2010). The manner in which mRNA and tRNA dissociate from the 40S subunit remains unclear (Nürenberg and Tampé, 2013), though their release is accelerated by Ligatin (also known as eIF2D), potentially by stabilizing an open conformation of the 40S subunit which allows rapid tRNA dissociation (Dever and Green, 2012).

Translational Control

Regulation of Initiation

The final step of gene expression is translation of mRNA into protein, described in detail in the previous section. This is a highly regulated process, as cells must produce and maintain proper protein levels to function in their environment and respond to environmental cues. Translational regulation occurs mostly through the modulation of translation initiation, as this allows for a quicker proteomic response to stimuli than induction of transcription. Control of initiation has become more complex throughout evolution, illustrated by humans having an order of magnitude more initiation factor mass than bacteria (Fraser, 2015). In general, there are two cellular mechanisms regulating translation initiation: (a) global control, where the overall level of translation within the cell is altered, and (b) mRNA-specific control, where the translation of specific messages is modulated (Gebauer and Hentze, 2004, Jackson et al., 2010).

Global translational control predominantly relies on inhibition through modulation of initiation factor phosphorylation. In one pathway, initiation factor eIF2α is phosphorylated, reducing the cellular amount of active initiation complex. Typically, eIF2α leaves the 40S subunit after translation initiation and regenerates through interaction with eIF2B. Phosphorylated eIF2α, however, binds very tightly to eIF2B, an interaction which sequesters eIF2α and prevents its function in translation initiation (Van Der Kelen et al., 2009). A second pathway prevents initiation by hindering recognition of the mRNA 5' cap, a step that depends on the interaction between eIF4E and eIF4G. A family of eIF4E-binding proteins (eIF4E-BPs) interact with eIF4E using the same binding site as eIF4G, resulting in competition between eIF4E-BPs and eIF4G for eIF4E. This competition strongly favors eIF4E-BPs once they have been phosphorylated (Sonenberg and Hinnebusch, 2009).

Translation initiation on individual mRNAs is primarily controlled by regulatory protein complexes binding to unique sequence elements, normally in mRNA 5' or 3' UTRs (Gebauer and Hentze, 2004). Often reversible, this binding will inhibit translation by altering the conformation of the mRNP and preventing eIF4F from accessing the mRNA 5' cap (Abaza and Gebauer, 2008, Jackson et al., 2010). In addition to sequence elements, recent work in human cells had shown that translation initiation can also be controlled by base modifications within the RNA. A newly discovered protein, YTHDF1, interacts with N^6-methyladenosine modifications within an mRNA to recruit the translation initiation machinery, resulting in increased translational output of the message (Wang et al., 2015).

Upstream ORFs

One specific class of mRNA-specific translational control elements are upstream ORFs (uORFs). Located within transcript leaders (5' UTRs), uORFs precede the initiation codon of the canonical ORF and are defined by an initiation codon with a downstream, in-frame stop codon. Typically, uORFs suppress translation of the downstream ORF (Barbosa et al., 2013, Calvo et al., 2009, Morris and Geballe, 2000). In cap-dependent translation initiation (see Translation Cycle section), the 43S pre-initiation complex scans along the 5' UTR until it encounters a start codon. As the uORF is first, it monopolizes the ribosome, thus reducing the efficiency of translation initiation on the downstream start codon. Upstream ORFs also reduce protein expression levels by triggering mRNA decay (Barbosa et al., 2013). However, uORFs can increase protein expression in response to cellular stress (Hinnebusch, 2005) or cell cycle stage (Brar et al., 2012). In 2009, it was predicted that roughly half of human and mouse transcripts contained a uORF (Calvo et al., 2009). With the development of ribosome profiling, the ease of uORF identification has grown enormously, which will enhance discoveries of uORFs and their unique form of translational regulation.

Nonsense-mediated mRNA Decay

Another class of *cis*-acting sequence elements regulating gene expression is premature termination codons, which are recognized through a translation-dependent process called nonsense-mediated mRNA decay (NMD). NMD mainly functions to prevent the production of erroneous proteins by rapidly degrading mRNAs containing premature termination codons, though it also regulates the abundance of many endogenous transcripts (Peccarelli and Kebaara, 2014). Discovered nearly 40 years ago in *S. cerevisiae* (Losson and Lacroute, 1979) and human cells (Chang and Kan, 1979, Maquat et al., 1981), it has since been shown to be an active mechanism of translational control in all eukaryotes.

Activation of NMD is dependent on three proteins: Upf1 (UP-Frameshift 1), Upf2, and Upf3. Upf1 is recruited when the ribosome encounters a premature termination codon, which differs from a normal termination codon due to its location relative to other cis-elements (e.g., a downstream element in yeast or an exon-junction complex in humans; Baker and Parker, 2004). Decay of the ribosome-bound mRNA occurs once Upf1 interacts with both Upf2 and Upf3 (Chang et al., 2007), promoting rapid degradation of the mRNA through the recruitment of decay enzymes (Kervestin and Jacobson, 2012). Because NMD is translation-dependent, it has been proposed that decay occurs the first time a ribosome encounters that codon, during the initial, or pioneer, round of translation (Gao et al., 2005, Ishigaki et al., 2001). However, recent evidence in *S. cerevisiae* has demonstrated that decay can occur on polysomes (Hu et al., 2010) or at any point during the translation cycle (Maderazo et al., 2003).

Localization

A third type of mRNA-sequence driven regulation is the control of translation through localization of the mRNA. By specifying the subcellular localization of an mRNA and inhibiting translation of this message until it reaches its destination, the cell is able to control the location of the encoded protein. Thus, the encoded protein only functions within a discrete subcellular region. To accomplish this targeting, a *cis*-acting localization sequence within the mRNA ("zipcode" or "address") is bound by specific RNA-binding proteins. Through interaction with other complexes, these proteins typically function in both localization and translational repression. Localization is achieved through interactions with molecular motors which drive the mRNP to its target location (Czaplinski and Singer, 2006).

Translational regulation through mRNA transport is a wide-spread phenomenon, with examples of localized mRNAs found across a wide range of eukaryotic species. In *S. cerevisiae* undergoing mitosis, *ASH1* mRNA is localized to the bud tip to inhibit mating type switching in the daughter cell (Paquin and Chartrand, 2008). In *Drosophila*, an impressive study demonstrated that 71% of mRNAs expressed during embryogenesis are subcellularly localized (Lécuyer et al., 2007). Gene expression in embryogenesis is also temporally controlled, adding yet another layer to gene regulation. In animals, mRNA localization is used in neuronal dendrites and axons to respond to environmental cues (see section entitled Neurons: An Alternate Experimental System).

One specific type of mRNA localization directs mRNAs bound by ribosomes to the endoplasmic reticulum (ER) as part of the secretory pathway. Messages are directed to this pathway by a *cis*-element called the signal sequence, which is typically located at the 5' end of the mRNA. Briefly, the encoded signal peptide emerges first from the ribosomal exit tunnel and is recognized by signal recognition particle (SRP), which inhibits further translation. This ribosome-SRP complex then localizes to the ER membrane with the help of SRP receptor and docks with the Sec61 complex. Docking releases the ribosome back into elongation, with the nascent chain passing through the Sec61 tunnel into the lumen of the ER, where chaperones help the protein fold into its proper conformation (Akopian et al., 2013). As ~30% of the eukaryotic proteome traffics through the ER, this is an incredibly common and highly regulated class of mRNA localization (Akopian et al., 2013).

Techniques to Study Translation

Low Throughput Techniques: Polysome Profiling, Toeprinting and Ribosome Density Mapping

Although transcriptome analysis can act as a proxy for gene expression, mRNA levels are often uncorrelated to the abundance of the proteins they encode (Maier et al., 2009, Vogel and Marcotte, 2012), hindering accurate quantifications. Other than protein degradation, this lack of correlation is likely due to translational regulation, as two transcripts present at equimolar concentrations can be translated at different rates. Therefore, it is necessary to measure the mRNAs undergoing active translation to truly understand gene expression.

To assess the translational activity of an mRNA, the metric most commonly used is the degree of its association with ribosomes, often measured by polyribosome profiling (Figure 1.3). This technique relies on velocity sedimentation of a sample through a gradient of increasing sucrose concentrations (Britten and Roberts, 1960). Sucrose gradients are fairly stable, withstanding both the addition of sample to the top of the gradient and centrifugation at high speed. During centrifugation, the components of the sample will separate based on their individual sedimentation rates, which are a function of the mass, density and shape of each molecule. Because the ribosome is such a large RNP, mRNA sedimentation rate is driven by the number of associated ribosomes. Sample components of similar sedimentation rates will form bands within the gradient, which are then collected and analyzed. It is very common for RNA to be extracted from the gradient fractions and then analyzed by northern blotting or quantitative PCR (or high-throughput techniques; see Microarrays and Deep Sequencing section).

Polysome profiling reports on the average number of ribosomes associated with a given transcript, but it offers no information as to where the ribosomes sit on the mRNA. The first technique to identify the location of ribosomes along an mRNA was a primer extension inhibition assay known as toeprinting (Hartz et al., 1988, Kozak, 1998). In the eukaryotic version of this assay, *in vitro* translation of an mRNA is inhibited by the addition of cycloheximide, and the mRNA is reverse transcribed into cDNA using a radiolabeled primer. A ribosome stably bound to an mRNA will block the movement of reverse transcriptase along this mRNA, producing a short cDNA product. The mRNA location where reverse transcription stopped corresponds to the 3' edge of the ribosome. The throughput for this method is low, as only a single mRNA can be analyzed in each reaction. A modification to this approach uses fluorescent primers and instrumentation to automate the detection of RT fragments (Gould et al., 2005, Shirokikh et al., 2010), simplifying the toeprinting process and improving its accuracy.

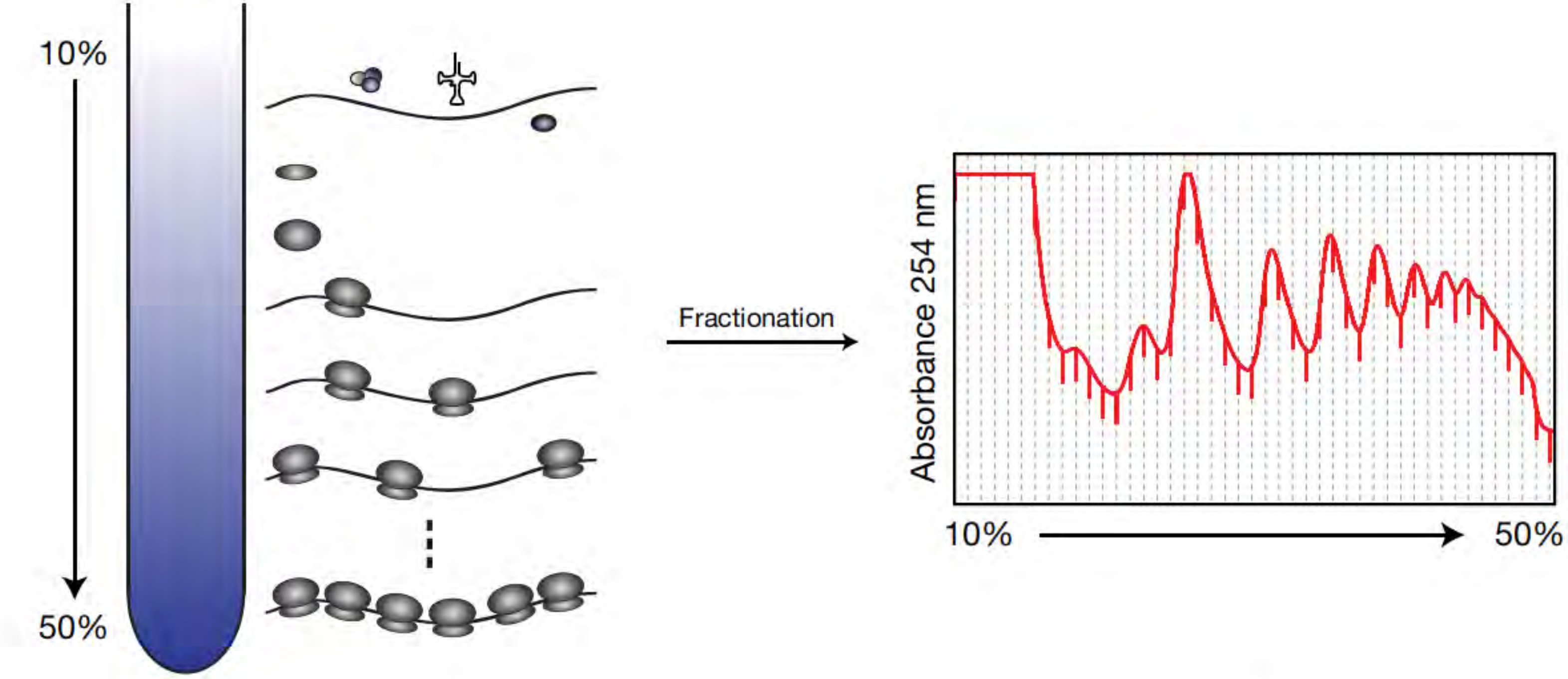


FIGURE 1.3: Polysome Profiling by Sucrose Density Gradient Centrifugation

Cellular lysate is separated across a 10-50% sucrose gradient. The sedimentation rate of each mRNP is dependent on the number of ribosomes, so mRNAs bound by the same number of ribosomes will sediment together. Free RNAs and proteins do not move far into the gradient, while mRNAs bound by many ribosomes sediment to the bottom of the gradient. The gradient is fractionated and collected while continuously monitoring the absorbance at 254 nm to detect RNA peaks and identify the number of ribosomes bound to each mRNA.

While toeprinting reports the specific location of a ribosome, the total number of ribosomes bound to that mRNA is unknown. A technique used to identify the general location of multiple ribosomes along an mRNA is Ribosome Density Mapping (RDM; Arava et al., 2005). First, polysome profiling is used to isolate a pool of polysomal mRNAs associated with a specific number of ribosomes. Then, the target mRNA is site-specifically cleaved by addition of an antisense oligonucleotide and RNase H. This digested sample is separated on a second sucrose gradient, and northern blotting is used to determine the number of ribosomes associated with each fragment of the cleaved mRNA. Thus, RDM results in slightly increased location sensitivity by identifying the number of ribosomes associated with a specific portion of an mRNA. However, this process is very low-throughput and labor intensive.

Microarrays and Deep Sequencing

Several technological advancements enabled the study of the translational status of all mRNAs expressed in an organism, one of which was the development of microarrays. A microarray is a slide with thousands of DNA probes arranged across its surface, enough to probe for all known genes in a genome. To identify RNAs in a sample, reverse transcriptase synthesizes a cDNA library, incorporating fluorescent molecules. Then, the fluorescently-labeled cDNA library is hybridized to the microarray. The appearance of fluorescent signal in a specific location identifies a specific gene sequence, and the intensity of the signal indicates the amount of that specific RNA in the original sample. For the study of translation, microarray analysis is usually paired with polysome profiling to generate genome-wide maps of ribosome occupancies and densities (for examples, see Arava et al., 2003, Thomas and Johannes, 2007). To do this, RNA is extracted from sucrose gradient fractions and analyzed on a microarray, inferring translational status from the number of ribosomes with which it cosedimented during polysome profiling.

While microarrays presented a major step forward in the ability to interrogate all genes in a genome, the scope of analysis was limited to known genes. This shortcoming was overcome by the development of deep sequencing (discussed in detail in the section entitled Deep Sequencing). With this technology, the identity of an RNA cosedimenting with ribosomes is revealed during analysis; there is no requirement for *a priori* knowledge of mRNA sequence. Thus, deep sequencing enables the characterization of novel translation substrates.

Ribosome Profiling

The advent of deep sequencing technology paved the way for a major advancement in the study of translation: the development of ribosome profiling (Ingolia et al., 2009). Developed by Nicholas Ingolia and Jonathan Weissman, ribosome profiling captures a "snapshot" of cellular translation at sub-codon resolution. This method draws on work isolating mRNA fragments protected from nuclease digestion by ribosomal association to identify the bacteriophage mRNA sequence bound by initiating ribosomes (Steitz, 1969). This footprinting by nuclease digestion was exploited by Ingolia and Weissman, who digested *S. cerevisiae* cellular lysate with RNase I, an endonuclease which cleaves after all 4 bases. The digested lysate was separated through a sucrose gradient, collecting the 80S monosome fractions afterwards. RNA was extracted from these fractions and loaded onto denaturing PAGE gels to size select ribosome footprints (~28 nts in *S. cerevisiae*). Once these mRNA footprints were extracted from the gel, they were constructed into deep sequencing libraries (see global footprinting, Figure 2.1).

Through deep sequencing of these footprints, gene expression can be measured at the translational level. Typically, an RNA-Seq library is prepared from the same starting material, enabling normalization of footprint abundance to mRNA abundance. The real power of ribosome profiling, however, is that ribosomal locations throughout the transcriptome at the time of lysis can be identified. This position information can identify the reading frame being translated and the codons sitting in the A and P sites. For mRNAs with multiple ORFs, ribosome profiling data enables the identification of which ORF(s) is being translated, enhanced by the 3-nt periodicity of ribosome footprints.

This ability was demonstrated in the original ribosome profiling paper, where non-AUG uORF translation was found to increase sixfold during starvation (Ingolia et al., 2009). Further work from the Weissman lab demonstrated a significant increase in uORF translation during meiosis as well (Brar et al., 2012). Significant uORF translation has also been found in mammalian cells (Fritsch et al., 2012, Ingolia et al., 2011, Lee et al., 2012). Additionally, novel short ORFs (other than uORFs) have been identified in *S. cerevisiae* (Smith et al., 2014) and *Drosophila* S2 cells (Aspden et al., 2014) using ribosome profiling.

In addition to informing ORF annotation, ribosome profiling data has greatly expanded our understanding of translational mechanisms. Ribosome profiling in *S. cerevisiae* strains deficient for a ribosome recycling factor revealed abundant fragments of different sizes, with short footprints (15-24 nts) protected by a ribosome stalled at the 3' end of a truncated RNA and long footprints (40-80 nts) protected by 2 closely stacked ribosomes that accumulate after the 1st ribosome has stalled (Guydosh and Green, 2014). In the absence of cycloheximide, *S. cerevisiae* ribosomes were found to protect ~21 nt fragments in addition to the canonical 28 nt fragments, potentially protected by an alternative conformation of the ribosome (Lareau et al., 2014).

In the original manuscript describing ribosome profiling, footprint signal aggregated across hundreds of highly-expressed genes showed a 3-fold increase over the first 30-40 codons (Ingolia et al., 2009). This signal was interpreted as increased ribosome occupancy, supporting the idea of a "translational ramp" where ribosomes move more slowly through this initial region to prevent downstream traffic jams (Shah et al., 2013, Tuller et al., 2010). A genome-wide enrichment in low efficiency (rare) codons in this same region suggested that footprint signal increased because of longer decoding time at rare codons (Tuller et al., 2010). The translational ramp theory will be specifically addressed in Chapter II in the section entitled The First Round of Translation and Translational Ramps.

By identifying the exact mRNA locations bound by ribosomes, ribosome profiling has fundamentally changed the way translation is studied. With this technique, gene expression can be measured, mechanisms of translational control can be studied, and the rate of protein synthesis and the abundance of the encoded protein can be predicted. In the ~6 years since the ribosome profiling method was published, this technique has revealed many features of translation that were previously unrecognized, and will certainly continue to do so in the future.

Monosome Translational Status

Density gradient fractionation of cytoplasmic lysates reveals two populations of fully assembled ribosomes: polysomes and monosomes. Polysome fractions contain mRNAs occupied by two or more ribosomes, whereas the monosome fraction contains mRNAs bound by a single ribosome plus "vacant couples", wherein the large and small ribosomal subunits stably associate with no bound mRNA (Noll et al., 1973). As initially demonstrated by Warner, Knopf and Rich in their seminal 1963 paper (Warner et al., 1963), polysomes are sites of active protein synthesis. Compelling evidence for this was that radioactive amino acids incorporated into nascent peptides by rabbit reticulocyte lysate cosedimented with a RNase-sensitive complex of comparable size to ~5 individual ribosomes (i.e., polysomes). In the same gradient, however, almost no radioactivity cosedimented with monosomes. This led to their reasonable conclusion that "protein synthesis in the reticulocyte occurs only on [a polyribosome] and not on a single ribosomal unit" (Warner et al., 1963). Further evidence for translationally-active polysomes and translationally-inactive monosomes came from a follow-up paper demonstrating, again in reticulocyte lysate, that 2 tRNA molecules cosedimented with each polysomal ribosome, whereas <1 tRNA molecule cosedimented with each monosome (Warner and Rich, 1964). This suggested that monosome-associated mRNAs in reticulocyte extracts are predominated by ribosomes in the process of initiation (i.e., sitting at the start codon with a stably-bound initiator tRNA^{met} in the P site, but no tRNA yet in the A site).

The primary role of reticulocytes *in vivo* is to produce hemoglobin for oxygen transport once the cell becomes a mature erythrocyte. Hemoglobin open reading frames (ORFs) are long enough to accommodate 5-6 ribosomes each, and α- and β- hemoglobin mRNAs account for the vast majority of all mRNA molecules in this cell type (Bonafoux et al., 2004). Thus reticulocyte translation is predominated by just two mRNAs evolutionarily tuned to produce massive quantities of protein. In such a system, it would be expected that the vast majority of radioactive amino acid incorporation would occur on polysomes containing ~5 ribosomes each (exactly as observed over 50 years ago). However, other cell types expressing a more diverse mRNA population with different ORF lengths and translation efficiencies might exhibit quite different profiles with regard to polysome and monosome translational activity. In particular, for proteins undergoing localized translation whose optimal levels are just a few molecules per cell or subcellular region, many ribosomes per mRNA molecule seems unlikely. Nonetheless, the reticulocyte experiments - and others like it (Gierer, 1963, Wettstein et al., 1963) - are the apparent basis of the widespread notion that active translation is limited to polysome-associated mRNA in all cell types. The corresponding belief that monosomes are translationally inactive is also quite common, with statements to that effect evident in numerous publications [for examples, see (Aspden et al., 2014, Cosgrove et al., 1982, Irier et al., 2009, Van Der Kelen et al., 2009).

Over the years, some anecdotal evidence has demonstrated that monosomes can be translationally active. For example, RPL41A and B are primarily translated by monosomes in *S. cerevisiae* (Yu and Warner, 2001). Due to their extremely short length (78 nts), these ORFs can likely only accommodate a single ribosome. Because the majority of canonical ORFs are much longer than 78 nts, these data have done little to change the opinion of the field regarding monosome translational status. The body of work presented in Chapter II challenges the hypothesis that monosomes are translationally inactive and adapts ribosome profiling to determine the translational status of 80S monosomes in *S. cerevisiae*. This work demonstrates monosomes are translationally active and translate specific subsets of mRNAs.

Nucleic Acid Sequencing

Ribosome profiling is dependent upon deep sequencing, which is the sequencing of millions of different nucleic acid sequences at one time. Deep sequencing, also called high-throughput sequencing (HTS) or next generation sequencing (NGS), is a relatively new technology that has only been on the market for approximately

10 years. The following sections will discuss the history of sequencing, the technology behind deep sequencing, and the challenges of RNA sequencing.

DNA Sequencing

Before 1953, the structure of DNA was unknown. Specific components had been identified; it was known to contain phosphates, sugars, and different amounts of two purine (adenine and guanine) and two pyrimidine (cytosine and thymine) bases - but not how all the building blocks fit together. Structural models had been proposed by Linus Pauling and Robert Corey (Pauling and Corey, 1953) and by Bruce Fraser (unpublished), but were ultimately disproven. In 1953, James Watson and Francis Crick accurately predicted the structure of DNA (Watson and Crick, 1953), a discovery that won them the Nobel Prize in 1962.

However, it was not until 1977 that technological development enabled the specific identification of nucleotide order within a DNA sequence. Though Maxam-Gilbert sequencing was the first published method (Maxam and GILBERT, 1977), its technical complexity led to its minimal use once Sanger sequencing was published later the same year (Sanger et al., 1977). Developed by Frederick Sanger, this method was based upon the incorporation of dideoxynucleotides (ddNTPs) during DNA replication. These modified nucleotides were chain terminators, preventing subsequent base incorporation by DNA polymerase. Four sequencing reactions were performed at once, each with the same sequencing primer and mixture. A minor (~1%) amount of a single ddNTP (ddATP, ddCTP, ddGTP or ddTTP) was added to each reaction, such that full-length product could be synthesized, but fragments would be produced whenever the ddNTP was incorporated. Analyzing these four reactions together (originally by denaturing gel electrophoresis of radiolabeled DNA) allowed for sequence identification based upon the relative position of the sequence fragments.

Technological advances, including the development of fluorescently-labeled ddNTPs (Smith et al., 1986), eventually automated the collection of sequencing information. Though these developments eased the technical difficulties and increased the accuracy of Sanger sequencing, it remained the most widely used sequencing method for ~25 years. Recently, the frequency of Sanger sequencing has dropped precipitously, as whole gene and genome sequencing is now typically done with deep sequencing (see section below entitled Deep Sequencing). However, Sanger sequencing still has its place in the lab, and will likely continue to for years to come.

Deep Sequencing

Deep sequencing, as briefly mentioned above, refers to the massive parallelization of a sequencing reaction similar to Sanger sequencing. At this scale, the sequence of millions of nucleic acids can be determine at once, allowing a broader look at genomic identity and gene expression patterns with single nucleotide resolution. In general, sequencing machines rely on some form of immobilization to hold the DNA sequence in one place, which allows for imaging after each reaction cycle to identify the incorporated nucleotide. The manner in which this is done varies between different platforms and will be discussed below. However, one preparation step used by all common sequencing platforms is colony amplification. This step usually occurs immediately before or after the sample is applied to the sequencing machine and is necessary for signal acquisition. Most detection methods used on deep sequencing machines require a strong signal, as they are unable to detect single molecules. Therefore, it is necessary to generate these colonies of identical sequence (clonal colonies) which will incorporate new nucleotides en masse, amplifying the signal.

The first successful NGS approach was 454 sequencing, which relied on pyrosequencing technology to achieve a throughput approximately 100X over Sanger sequencing (Margulies et al., 2005). DNA templates were amplified inside water droplets to form a clonal colony on a primer-coated bead, and then each bead was deposited into a tiny well on a slide where the sequencing reaction occurred. Base incorporation was monitored by detection of light generated via a luciferase reaction. This reaction is dependent on the pyrophosphate released from a dNTP molecule when it is incorporated into the growing DNA chain. Because the readout is the same for every base, individual solutions of A, C, T and G nucleotides are sequentially added to and removed from the sequencing reaction.

Solexa technology was the next major player in the sequencing market, and relied on reversible dye-terminator technology (Bentley et al., 2008). Complementary adaptor sequences drive hybridization of the DNA sample to a slide, after which colony amplification generates clusters containing ~500-1000 copies of each DNA sequence. In each sequencing cycle, polymerase incorporates fluorescently-labeled nucleotides with a chain terminating blocking agent, limiting nucleotide incorporation to one base per sequencing cycle. An image of the slide is recorded after every sequencing cycle, after which the chain terminator is removed so the sequencing reaction can proceed. Each sequence is tracked by location, with the color after each sequencing cycle identifying the incorporated base (see Figure 1.4).

Illumina purchased Solexa and their technology in early 2007, just after the launch of the first Solexa sequencer, and has dominated the sequencing market ever since. In 2014, approximately 90% of the DNA sequenced was generated on Illumina platforms (Regalado, 2014). New sequencing technologies face the difficult task of breaking into this market, as they need to compete with the standards of low sequencing cost/base and low error rates set by Illumina. Though a few other platforms have made it to market, they will not be discussed in detail, as the sequencing data in this thesis (Chapters II and III) were collected using Illumina technology.

Challenges of RNA Sequencing

Though first used for genome sequencing, researchers quickly realized the utility of applying deep sequencing technology to the study of gene expression. RNA sequencing (RNA-Seq) allows the detection and quantification of all expressed transcripts in a sample, with applications to differential expression, novel transcript identification and alternative splicing. Researchers in all RNA fields have benefited enormously from this technology, but it is rife with challenges, both pre- and post- sequencing. In terms of sample preparation, all deep sequencing platforms commonly used today only sequence DNA, so RNA must be captured and converted to cDNA prior to sequencing. This multi-step conversion process - called "library construction" or "library preparation" in this thesis - append specific adaptor sequences to the cDNA and amplify the sample to make it viable for sequencing. These adaptor sequences facilitate attachment to a surface [either a bead (454) or a slide (Illumina)] via hybridization to complementary sequences bound to the sequencing surface. Second, they provide primer sites for clonal amplification, which as discussed above is required to amplify the detected signal from each sequencing reaction. Finally, adaptor sequences allow for the hybridization of a sequencing primer, which is necessary to initialize the sequencing reaction. The nucleotide sequences of these adaptors are platform-specific and vary in length. The manner in which these adaptor sequences can be appended to either the RNA or cDNA will be discussed in detail in Chapter III.

One challenge during library construction is ensuring that the RNA remains intact. As mentioned above, there are many types of RNA sequencing, and often the length of the RNA molecule is an important piece of information. Therefore, mRNA degradation during library construction is a major concern since RNA, unlike DNA, is chemically unstable. Even though RNA, in many ways, is very similar to DNA - the components are the same, both form helical structures, and the sugar-phosphate backbones connecting individual nucleotides are virtually identical - the presence of a $2'$ hydroxyl (OH) instead of a single hydrogen atom (H) gives RNA the ability to hydrolyze itself into fragments. This 2' OH breaks a phosphodiester bond in the sugar-phosphate backbone through a nucleophilic attack on a neighboring phosphorus, resulting in ester cleavage of the backbone.

RNA auto-hydrolysis makes it challenging to work with, and special care must be taken to maintain the integrity of the RNA sample during construction of an RNA deep sequencing library. While this seems straightforward and obvious for RNA researchers, it is nevertheless an important point to note when discussing a technique where RNA length can be crucially important.

RNAs are expressed across a wide range of cellular concentrations (at least 5 orders of magnitude in eukaryotic cells; Mortazavi et al., 2008) and may only be expressed in a specific cellular location or at a particular time (for examples, see Lécuyer et al., 2007, Madabhushi et al., 2015). Consequently, sequencing coverage varies considerably across transcripts. Therefore, the number of sequencing reads required by any individual experiment will be determined by the least abundant RNA of interest. However, the amount of signal coming from low expression genes as a percentage of the total amount of RNA is minuscule, due to the incredibly high expression levels of noncoding RNAs. In *S. cerevisiae*, ribosomal RNA (rRNA) accounts for ~80% of the total transcriptome, and transfer RNA accounts for ~15%, leaving only ~5% of the transcriptome to mRNA and other noncoding RNAs (von der Haar, 2008, Warner, 1999). Thus, to avoid wasting time and money repeatedly sequencing rRNA and tRNA, these sequences are often avoided during library construction. One common approach to avoid rRNA is to prepare libraries from polyA-selected RNA, while others have attempted to deplete rRNA either before (Benes et al., 2011) or after (Archer et al., 2014, Zhulidov et al., 2004) library construction (see Chapter III). As tRNAs have a specific size, it is possible to remove them from a sample by size selecting RNAs either much smaller or larger; therefore, tRNA contamination is a relatively minor problem for most library preparations.

After removal of rRNAs and tRNAs, sequencing is used to define the transcriptome of each sample. In order to specifically identify which genes are turned on, however, RNA strand information must be maintained. Some of the first library construction methods developed for RNA captured each molecule in a random orientation, resulting in the loss of strand information (Mortazavi et al., 2008). While these first methods enabled amazing transcriptomic discoveries, it quickly became clear that strand information would be crucial for complete transcriptome identification. In addition to yielding more information about transcription, the maintenance of strand information has recently been shown to result in more accurate estimates of gene expression compared to unstranded libraries (Zhao et al., 2015). This has facilitated the fascinating discoveries of the massive amount of transcription that happens all over the human genome. Recent work has demonstrated that antisense transcription is a ubiquitous phenomenon (He et al., 2008, Katayama et al., 2005), though both the amount and the full extent of its function remains to be discovered. Initial experiments have shown that it can influence gene expression either through the act of transcription itself, or by the noncoding RNA that is produced (Pelechano and

Steinmetz, 2013).

It would be negligent to discuss the challenges of RNA sequencing without briefly mentioning all the analytical work subsequent to data collection. Minimally, sequencing reads need to be (a) trimmed and filtered for quality, (b) mapped to the genome or built into a transcriptome to identify the set of genes expressed in a given sample, and (c) quantified to measure expression level. However, the required level of data analysis will depend on the organism being studied and the biological questions being asked. This bioinformatic analysis is computationally intensive and requires a set of coding skills not commonly found among biologists, so it is not uncommon for data to sit unanalyzed for months - or years - before someone with the correct set of skills is able to analyze it.

Sequence Read Length

Initial versions of 454 deep sequencing technology generated ~250,000 reads with a 100 nt read length, enough to sequence ancient DNA samples from wooly mammoths (Miller et al., 2008) and from Neanderthals (Green et al., 2008). This technology was also used to sequence James Watson's genome, which was the first complete genome sequencing of an individual (Wheeler et al., 2008). Technological developments have increased the read length of both 454 and Illumina platforms, with 454's current read length around 1 kb and Illumina's maximum read length of 300 bps (though if sequencing from both ends of the cDNA, as with paired-end sequencing, 600 bps can be sequenced). Both read lengths fall short of the average mammalian mRNA length (~2,000 nts; Ravasi et al., 2006). Therefore, any attempt to sequence intact mRNAs would result in reads limited to the ends of transcripts. Thus, to completely sequence across most mRNAs, they must be fragmented prior to library construction.

Once the fragmented RNA library has been sequenced, computational analysis determines the identity and abundance of transcripts in the original sample. This is often accomplished by identifying the genomic location where the RNAs were transcribed (genome-mapping); the accuracy of this mapping depends on the length of the sequencing read, as the probability of genomic mapping is proportional to $1/4^n$, where n is the read length (Vivancos et al., 2010). Therefore, longer read lengths are preferred. Besides genome-mapping, it is often desirable to bioinformatically piece the RNA fragments back together to identify the specific transcriptome of the sample. Longer read lengths are also preferred for transcriptome assembly as they are more likely to contain an exon-exon junction. This information is crucially important when studying systems with alternatively splicing (Wang et al., 2008, Xing and Lee, 2006), where multiple RNAs are encoded by the same gene with variability in the inclusion of certain exons.

While sequence read length typically has less of an impact on libraries prepared from small RNAs or short mRNA fragments, it can still affect the downstream conclusions. For most small RNA libraries, read length must be longer than the original RNA to ensure accurate identification of the boundaries.

Sources of Bias in RNA Sequencing

As mentioned previously, RNA-Seq has redefined transcriptomic studies through its incredibly sensitive detection of RNAs with single-nucleotide resolution. However, any conclusions regarding the abundance of RNAs is dependent on the sequencing data read count accurately reflecting the abundance of these RNAs in the starting pool. Unfortunately, library construction often introduces significant bias by altering the nucleotide content and abundances of RNA libraries (for a review, see (Linsen et al., 2009) and Chapter III).

Most of the biases in library construction are due to enzymatic preferences. Though there are many approaches to library construction (discussed in detail in Protocol Design section), they all rely on enzymatic catalysis which can have various effects on the resulting library. In brief, ligation enzymes (acting on RNA or cDNA) have nucleotide biases which result in more efficient ligation of certain sequence species, altering the base composition of the library. Reverse transcriptases might incorporate the incorrect base, or have low processivity, producing less than full-length cDNA (see Figure 3.4). The latter would be exacerbated by secondary structures, such as stem loops, which could impede reverse transcription of specific sequences. PCR enzymes have biases as well, preferentially amplifying certain sequences over others.

In addition to enzymatic bias, secondary structure can introduce bias. Sorefan et al. (2012) demonstrated that miRNAs discovered on the Illumina platform will fold *in silico* more strongly with multiple versions of Illumina adaptor sequences than with 454 adaptor sequences. The converse was true as well; miRNAs identified on the 454 platform fold more strongly with 454 sequencing adaptors than Illumina adaptors. When performing this analysis, Sorefan et al. (2012) utilized all miRNAs which had been sequenced at least once on the 454 and Illumina platforms, but ignored read number. Thus, if a miRNA had been sequenced on both platforms it was folded with adaptors for both platforms. This suggests that sequence complementarity to deep sequencing adaptors has played a role in miRNA identification, biasing miRNA discovery against the miRNAs which do not fold with adaptors. Other studies have observed similar structure preferences, demonstrating that ligation efficiency is enhanced by RNA base-pairing with the adaptor sequence (Fuchs et al., 2015, Zhuang et al., 2012). Additionally, secondary structure within a miRNA can reduce its ligation efficiency (Hafner et al., 2011).

Consequences of Bias in RNA Sequencing

That the relative levels of different sequences within the same library will be under- or over- represented due to preparation bias has been known for quite some time (Linsen et al., 2009), but the impacts of these effects are still being determined. A few case studies, however, have suggested that these effects can be quite significant. For example, Illumina sequencing data ranks miR-29b as the 29th most abundant miRNA in DLD-1 cells. However, alternative libraries constructed to minimize bias and quantitative northern blotting both rank miR-29b as the most abundant miRNA (Sorefan et al., 2012). Similarly, in one extreme case measuring the effect of adaptor 5' end sequence on capture efficiency, miR-106b was captured well (>4-fold more signal) only by a single 5' terminal sequence of the 12 sequences tested (Jayaprakash et al., 2011).

What effect can these biases have on the biological conclusions resulting from deep sequencing data? A disfavored miRNA sequence could cause researchers to overlook interesting miRNAs because they do not seem abundant in a sample. miRNAs display a significant amount of tissue-specific expression, with only ~25% expressed at a given time (Baer et al., 2013). They have become biomarkers for many diseases; the variable expression profile of 217 human miRNAs across cancerous tumor samples better defined the type of cancer than the expression profile of ~16,000 mRNAs (Lu et al., 2005). Given this, their potential use as prognostic or diagnostic biomarkers is quite high (Rosenfeld et al., 2008), but it will be critical to quantify them correctly. Adaptor sequence bias can also affect the magnitude of processing error attributed to a specific processing pathway mutation. For example, a distinguishing feature of primary piRNAs is the bias for U at position 1 (Zhang et al., 2015). Changes in this 5' bias are indicative of piRNA processing defects, so accurately capturing the extent of 5' U bias is important to piRNA studies. However, libraries prepared from cells with mutations in the processing pathway showed a range of 5' U bias (57 to >80%) depending on the sequence identity of the adaptor used in library preparation (Jayaprakash et al., 2011).

To a lesser extent, bias can also affect libraries prepared from long RNAs. Low complexity (often a result of excessive PCR amplification) libraries limit the conclusions that can be made from the data, as read redundancy severely decreases coverage depth. Variations in nucleotide composition can result in highly non-uniform read coverage across transcripts (first noted by Mortazavi et al. 2008). This uneven coverage hinders transcriptome assembly, and may affect the quantification of transcript abundances (Li et al., 2013).

Biases in RNA deep sequencing libraries (whether large or small RNA) undermine the potential sensitivity and accuracy of this technology. Therefore, it is important to construct RNA libraries in a way that introduces as little bias as possible. Chapter III, entitled **An optimized kit-free method for making strand-specific deep sequencing libraries from RNA fragments**, discusses my work developing an optimized protocol for constructing RNA libraries. This protocol works for all types of RNAs while minimizing bias at each enzymatic step in the construction process.

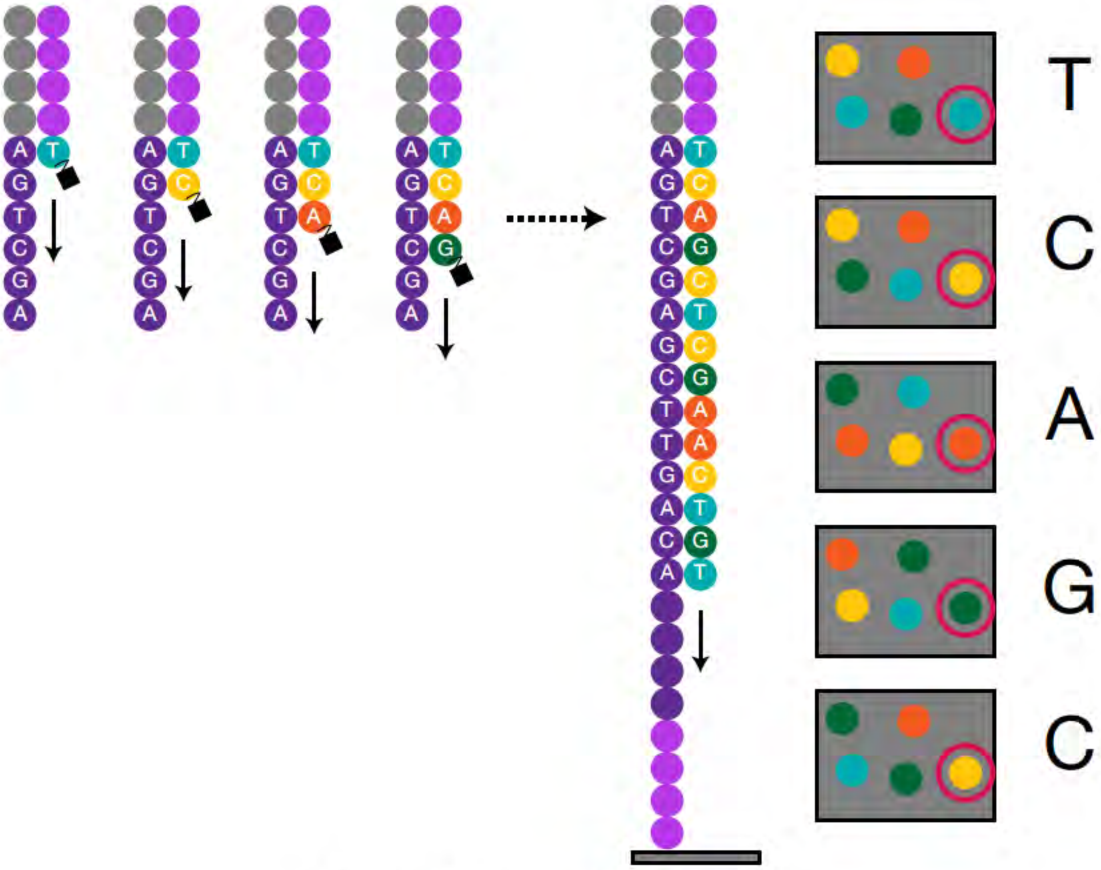


FIGURE 1.4: Sequencing by Synthesis

Adaptors (shown in pink and grey) allow the DNA template to attach to the flow cell and enable primer hybridization. During each sequencing cycle, nucleotides reversibly-labeled with a fluorescent terminator moiety are washed onto the flow cell and incorporated into the growing DNA chain by polymerases. Unincorporated nucleotides are washed away, and an image is captured of the flow cell (right). Finally, the terminator moiety is removed in preparation for another round of sequencing. The order in which fluorescence appears at a specific location on the slide indicates the DNA sequence.

Chapter II

Redefining the translational status of 80S monosomes

Preface

The contents of this Chapter appear as they were accepted for publication at Cell, December 2015.

Supplemental Table 3 is not provided in this thesis due to size. Please refer to the online supplemental material.

Introduction

The cytoplasm contains two populations of ribosomes: polysomes and monosomes. Polysomes consist of mRNAs occupied by two or more ribosomes, whereas monosomes are mix of mRNAs bound by a single ribosome plus "vacant couples" wherein the large and small ribosomal subunits stably associate in the absence of mRNA (Noll et al., 1973). Ample evidence from radioactive amino acid incorporation studies indicates that the vast majority of new peptide bonds are formed on polysomes (Noll, 2008, Warner and Knopf, 2002). Thus, polysomes are generally equated with the translationally-active mRNA pool, with monosomes often presumed to be newly assembled at the start codon and therefore translationally inactive (for examples, see Aspden et al., 2014, Van Der Kelen et al., 2009). Nonetheless, some fraction of monosomes must be translationally active. For example, the first or "pioneer" round of translation on any newly transcribed mRNA necessarily involves translation by a single ribosome until it has moved far enough to allow a second ribosome to assemble at the start codon. Further, the average distance between elongating ribosomes in *Saccharomyces cerevisiae* has been estimated to be >100 nucleotides (Arava et al., 2003, Shah et al., 2013). With such spacing, some ORFs (e.g., RPL41A and RPL41B, each 78 codons) are so short that they should be occupied by just one ribosome (Yu and Warner, 2001). Consistent with this, *S. cerevisiae* mRNAs with very short ORFs cosediment predominantly with 80S monosomes (Arava et al., 2003). Notably, that study also revealed that several longer ORF mRNAs known to be translationally-regulated (e.g., GCN4, CPA1, and ICY2) are primarily monosome-associated.

Ribosome profiling enables precise mapping of ribosome position on mRNAs undergoing active translation (Ingolia et al., 2009). Here, we adapted this protocol to specifically examine the translational status of 80S monosomes in *S. cerevisiae*. We provide definitive evidence that the vast majority of monosomes are in the act of elongation, not initiation. As expected, monosomes predominate on nonsense-mediated decay (NMD) targets, including unspliced transcripts. Transcriptome-wide, relative polysome and monosome occupancy is a function of initiation versus total elongation time. That is, if initiation is faster than elongation, an mRNA will be predominantly polysome-associated. Conversely, if initiation is much slower than elongation, an mRNA will be predominantly monosome-associated. A high initiation:elongation ratio can be driven either by ORF length or by slow initiation rate, often indicative of translation regulation. Thus, in addition synthesizing extremely short proteins, monosomes also translate key regulatory proteins such as transcription factors, kinases and phosphatases. Such regulatory factors are often transiently expressed (i.e., have short mRNA and protein half-lives) at very low levels. Therefore, relative monosome:polysome association may prove a useful metric for identifying and studying mRNAs subject to negative translation regulation.

Results

Monosome, Polysome and Global Footprinting

We took multiple precautions to ensure that the ribosome footprints we isolated accurately reflected intracellular conditions (Figure 2.1). To minimize ribosome movement during sample workup, we briefly incubated log phase cultures with 100 µg/ml cycloheximide prior to rapid collection by vacuum filtration and immediate resuspension in ice-cold lysis buffer, and performed all subsequent steps at 4°C (Ingolia et al., 2009). To preserve polysome integrity, we lysed cells by vortexing with glass beads rather than ball milling, as ball milling tends to shear polysomes and thereby artificially increase the monosome fraction. We also excluded detergent (e.g., TritonX-100) as it led to excessive foaming and poor cell lysis. Because efficient extraction of membrane-bound polysomes requires detergent (Potter and Nicchitta, 2002), we expected cytoplasmic species to predominate in our lysates. Although Mg^{2+} concentrations up to 30 mM are often used when preparing yeast extract for polysome profiling (Bhattacharya et al., 2010), $[Mg^{2+}]$ in excess of 8 mM can drive vacant couple formation (Favaudon and Pochon, 1976); therefore, we limited total $[Mg^{2+}]$ to 5 mM in all experiments.

For global ribosome profiling (Figure 2.1, left), cell lysates are digested with

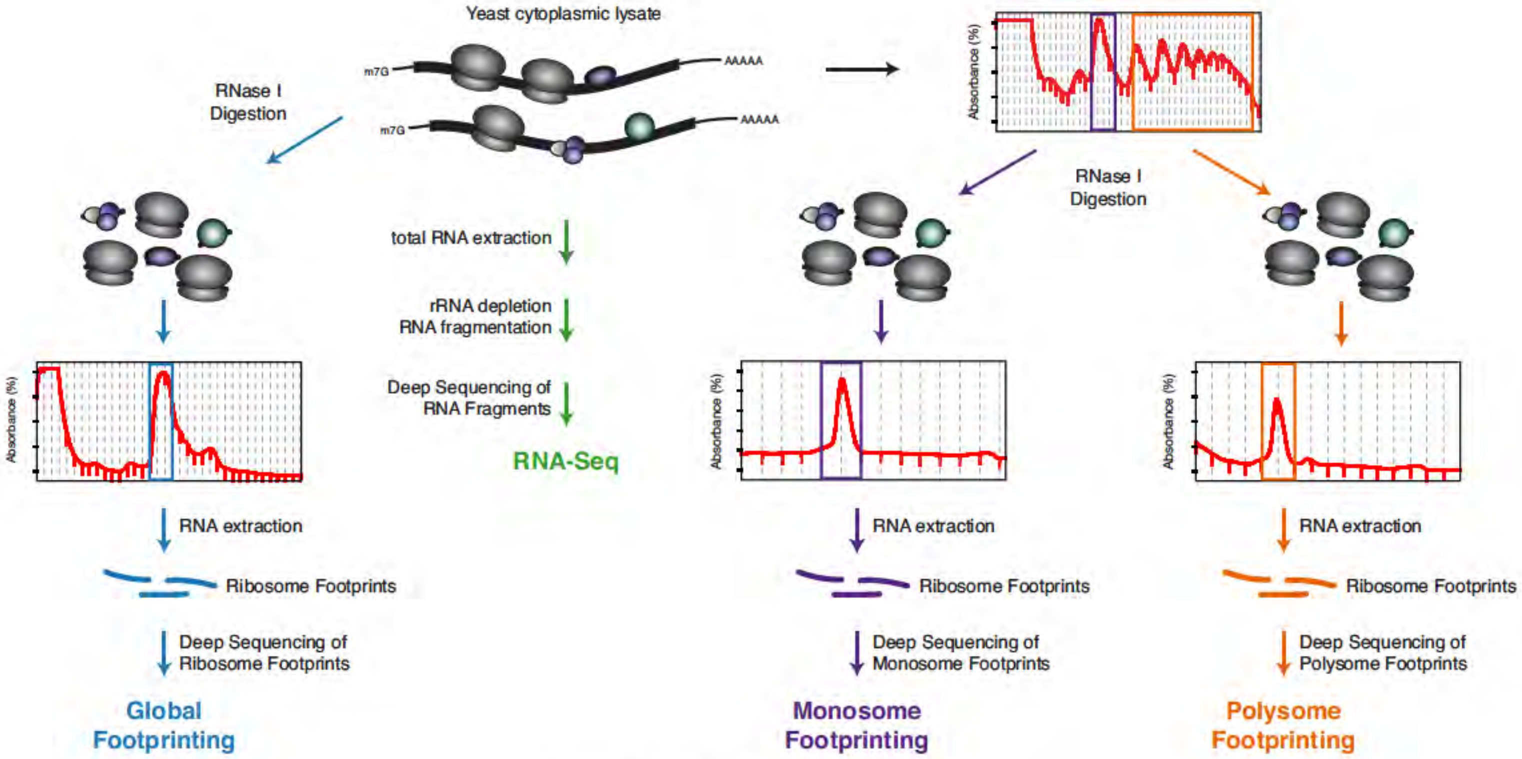


FIGURE 2.1: Experimental Scheme

RNase I prior to ribosome isolation (Ingolia et al., 2009) - consequently, the entire ribosome population is sampled without regard to monosome or polysome status. To generate monosome- and polysome-specific footprints, we first separated the two populations on a 6-38% (w/v) sucrose gradient. These gradients were centrifuged for a sufficient time to sediment the 80S monosome peak to the middle fractions, allowing both monosomes and polysomes to be collected with optimal separation between the two pools. These conditions also allowed for clean separation between monosomes and 48S initiation complexes that, although unlikely, could leave mRNA footprints upon RNase I digestion (Aspden et al., 2014, Ingolia et al., 2014). Pooled fractions were then separately digested with RNase I prior to loading on a second sucrose gradient (Figure 2.1, middle and right), ensuring that any mRNA fragments obtained were bona fide ribosome footprints and not similarly sized protected fragments originating from non-ribosome-bound positions on the intact monosome or polysome-bound mRNAs isolated from the first gradient. We also generated RNA-Seq libraries from total lysate RNA extracted prior to gradient fractionation.

For both biological replicates of isolated monosomes and polysomes, ~80-90% of uniquely mapping reads post-ncRNA removal aligned to the sacCer3 genome (Supplemental Table 2.1), with 5,045 of 6,692 annotated ORFs having at least 10 reads in all four libraries. Scatter plots comparing either total ribosome occupancy per ORF (reads per million mapped; RPM) or ribosome density per

ORF (reads per kilobase per million mapped; RPKM) revealed high correlations between biological replicates (Pearson coefficient >0.99) (Figure 2.2). Thus our monosome- and polysome-specific ribosome profiling data were highly reproducible and covered the vast majority (75%) of annotated ORFs.

Ribosome Position Analysis

When aggregated across all coding sequence (CDS) genes, ribosome footprints tend to be highly enriched at ORF 5' ends and then sharply decrease before reaching a plateau that persists throughout the remainder of the ORF (Ingolia et al., 2009). Both metagene and aggregation plots of our global ribosome footprints replicated these features (Figure 2.3A and B, left panels). Monosome and polysome plots, however, were distinctly different (Figure 2.3A and B, middle and right panels). Whereas both exhibited ORF 5' end enrichment, this enrichment was much more pronounced for monosomes and much less pronounced for polysomes than the global pattern. Further, the plateau was lower for monosomes and higher for polysomes than the global plateau. Therefore, monosomes and polysomes make distinct contributions to the global ribosome footprint pattern.

High monosome occupancy at ORF 5' ends might suggest that a large fraction is in the process of initiation with tRNA_{met} in the P-site. Because our double

TABLE 2.1: Statistics for RiboSeq and RNASeq Libraries

				sacCer3 Genome Mapping								**Transcriptome Mapping**	
Library	**Replicate**	**Barcode**	**# of reads**	**Total mapped reads**		**ncRNA**		**other genomic**		**length filtered*, other genomic unique reads**		**length filtered*, other genomic unique reads**	
				#	**%**	**#**	**%**	**#**	**%**	**#**	**%**	**#**	**%**
Monosome	1	CGGGACC	20,314,216	19,201,314	95%	13,532,186	70%	5,669,128	30%	4,889,880	86%	4,509,934	80%
	2	ACAAGCC	23,827,254	22,586,545	95%	14,887,835	66%	7,698,710	34%	6,701,367	87%	6,153,087	80%
Polysome	1	TAGCCCC	24,526,964	23,314,582	95%	12,244,519	53%	11,070,063	47%	8,777,740	79%	8,109,829	73%
	2	ATTCACC	24,322,817	23,239,971	96%	13,993,192	60%	9,246,779	40%	7,578,966	82%	7,013,751	76%
Global	1	TGACTCC	26,007,920	24,949,684	96%	24,002,025	96%	947,659	4%	781,401	82%	707,759	75%
	2	TCTACCC	24,556,489	23,533,151	96%	22,782,120	97%	751,031	3%	588,116	78%	522,845	70%
RNASeq	1	ATTCACC	13,732,341	12,946,709	94%	11,268,252	87%	1,678,457	13%	1,196,896	71%	–	–
	2	TCTACCC	14,400,141	13,115,357	91%	11,285,179	86%	1,830,178	14%	980,574	54%	–	–

***** See Extended Experimental Procedures

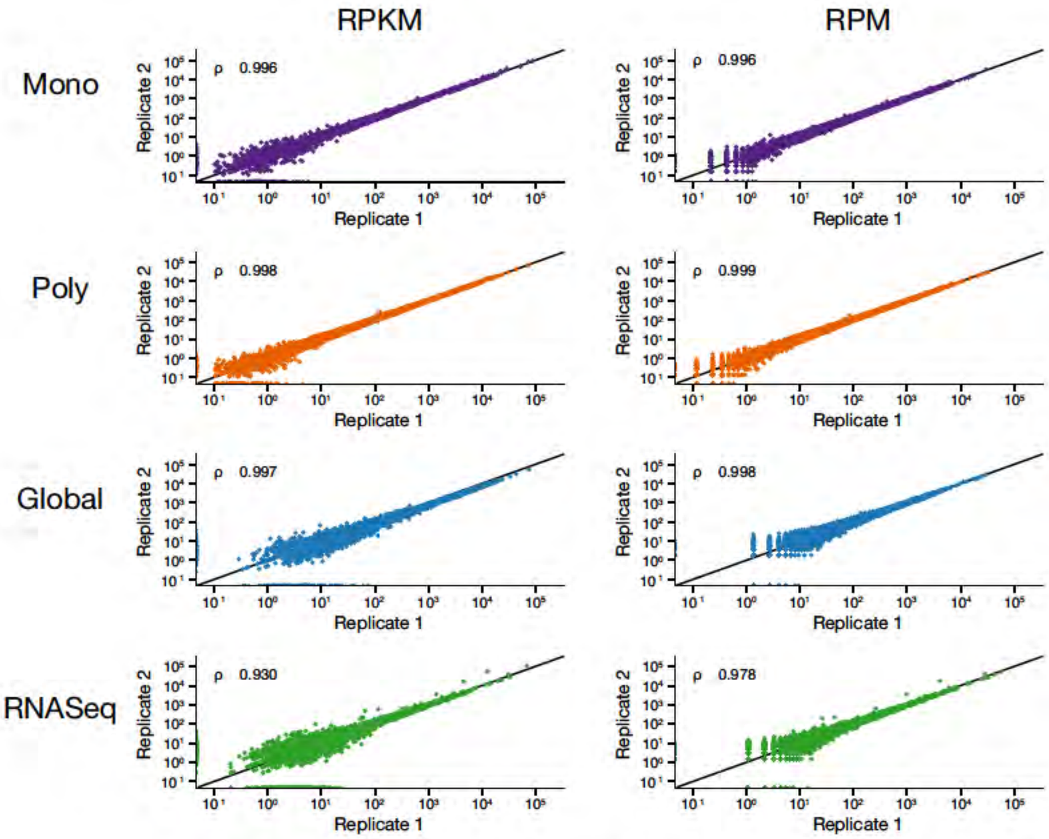


FIGURE 2.2: Correlation between Biological Replicates

Correlation (ρ = Pearson coefficient) between reads per kilobase per million uniquely mapped reads (RPKM; left) and reads per million uniquely mapped reads (RPM; right) for individual genes (dots) in biological replicates. Black line is y=x.

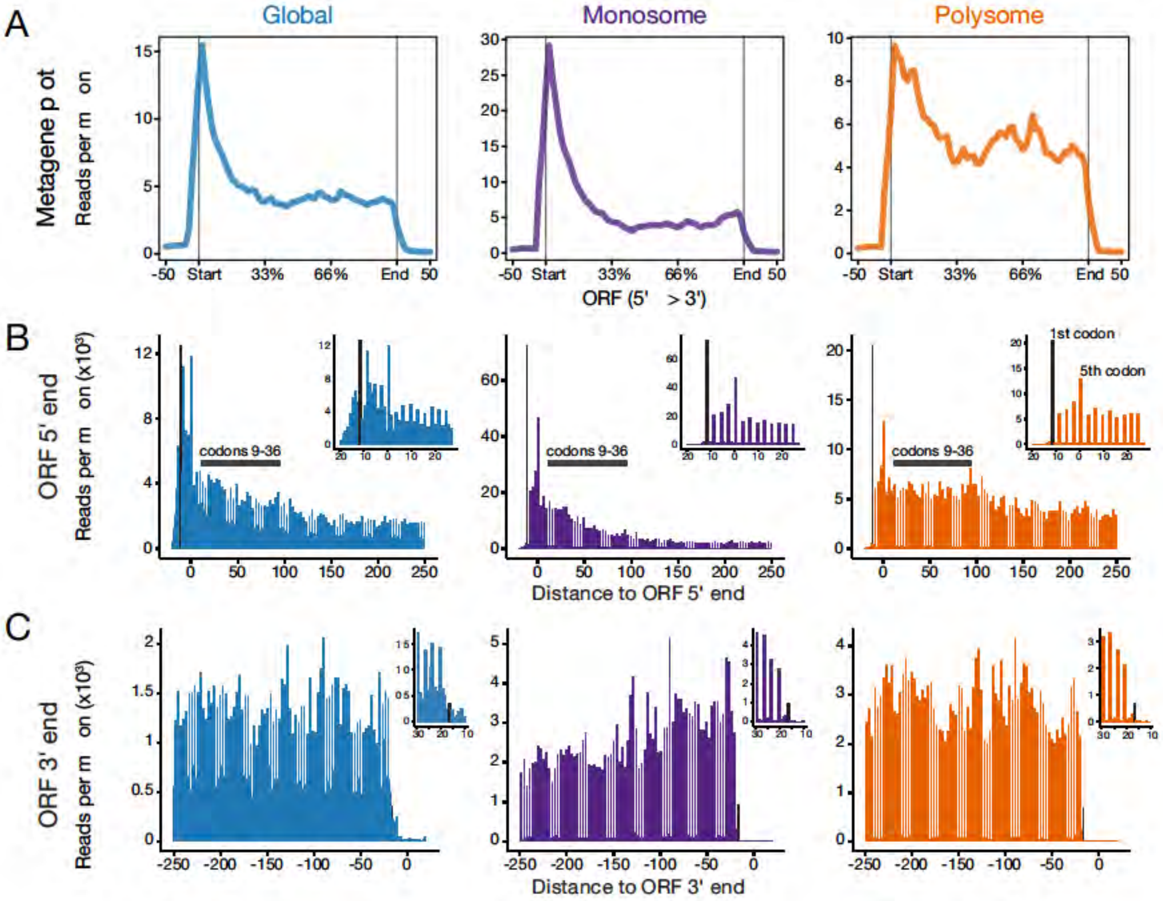


FIGURE 2.3: Global, Monosome and Polysome Footprint Read Coverage (continued on next page)

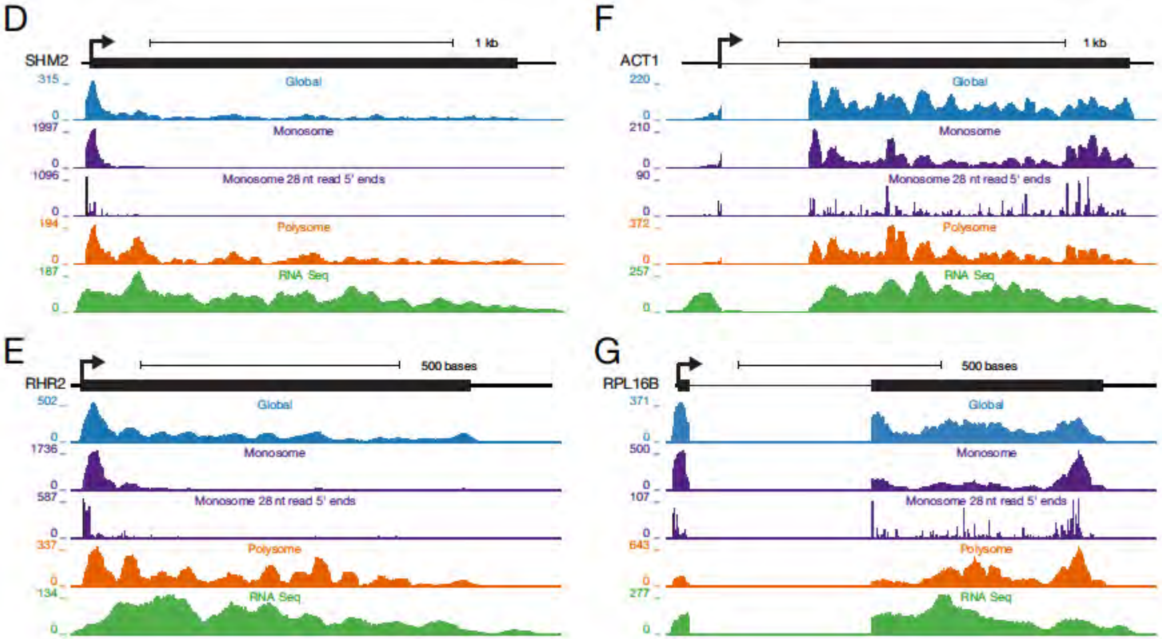


FIGURE 2.3: Global, Monosome and Polysome Footprint Read Coverage

A) Metagene plots using all ≥25 nt reads mapping to annotated genes. B) Aggregation plots using only 28 nt reads for all ORFs >300 nts. X-axis: Distance from ORF 5' end (i.e., first nt of start codon) to read 5' end; Y-axis: reads per million (RPM). Grey bar indicates codons 9 to 36. Insets show first 13 codons (nts -20 to +27), with peak at codon 1 indicated in black. C) Aggregation plots, as in B, except distances are from ORF 3' end (i.e., third nt of stop codon) to read 5' end. Insets show last 5 codons (nts -30 to -10), with 4 nt offset peak indicated in black. D-G) Distribution of ≥25 nt reads from indicated libraries (solid plots) or 28 nt monosome 5' read ends relative to the start codon (middle track) across individual genes. Whereas monosome reads on some genes predominate at the start codon (e.g., SHM2) or near the ORF 5' end (e.g., RHR2), monosome reads on other genes (e.g., ACT1 and RPL16B) are spread throughout the entire ORF in a pattern similar to polysome and global footprints. All plots (A-G) were constructed from biological replicate 1 using only uniquely mapping reads. See also Figure 2.4.

sedimentation strategy strongly disfavored free tRNA contamination in our footprinting libraries, any tRNA fragments in our libraries likely originated from stably-bound tRNAs. In both monosome and polysome libraries, very similar fractions of tRNA-mapping reads (0.76-1.03% Mono; 0.62-1.08% Poly) corresponded to tRNA_{met} (Table 2.2), indicating there was no major difference in tRNA_{met} association between monosomes and polysomes. Further, only 7% of monosome 28 nt reads were positioned over canonical ORF start codons, compared to 2% for polysomes. All other 28 nt reads mapping to canonical ORFs (93% and 98%, respectively) mapped to internal codon positions, demonstrating that most monosomes are in the process of elongation.

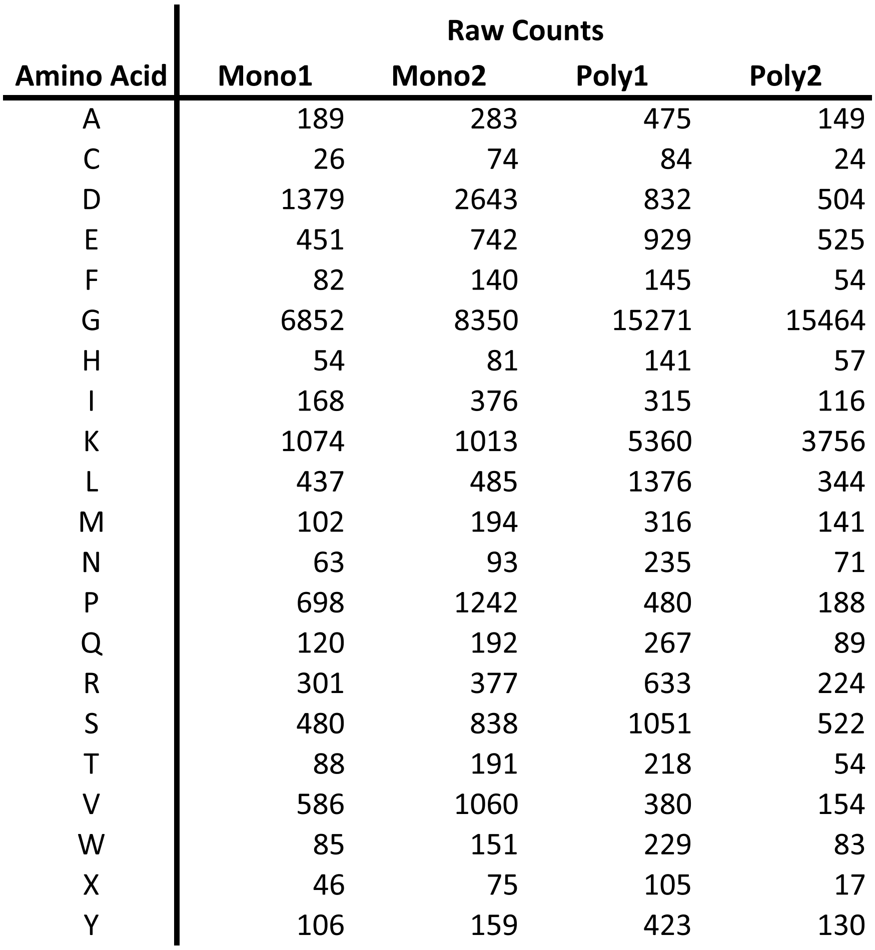
Further evidence for elongating monosomes came from aggregating 28 nt read 5' end positions relative to the start codon. Both monosome and polysome footprints exhibited the 3 nt phasing characteristic of elongating ribosomes (Figure 2.3B and inset), and for both populations, this strong phasing continued all the way to the stop codon (Figure 2.3C and inset). Among all 6,692 annotated CDS genes, 5,029 (75%) had more than five 28 nt monosome footprints in each biological replicate indicative of elongation (i.e., P-site inside the ORF). Thus, most monosomes (93%) were in the process of elongation, with most genes (75%) having multiple 28 nt monosome footprints within the ORF. We conclude that the preponderance of 80S monosomes in our samples were translationally active, and that at least a fraction of elongation events on most mRNAs occurs while the mRNA is occupied by a single ribosome.

A prominent difference between the monosome and polysome aggregation plots occurs across codons 9-36 (Figure 2.3B). Whereas monosome read coverage decreased ~3-fold over this region (5' ends at nt positions +12 to +93), polysome read coverage remained relatively even. These same patterns were observed when aggregation plots were limited to cytoplasmic mRNAs (see below; Supplemental Figure 2.4C-F). Likely explanations for these different patterns are discussed below. Another difference between the aggregation plots occurred at ORF 3' ends, where monosome reads exhibited a slight uptick (~1.5 fold) over the last 100 nts, while polysome reads did not (Figure 2.3C). This uptick in monosome reads was even more pronounced in aggregation plots limited to mRNAs that were otherwise polysome-enriched (Supplemental Figure 2.4G and H). This signal could originate from the final ribosome on an otherwise polysome-associated mRNA as that ribosome completes translation prior to or coincident with mRNA degradation (Pelechano et al., 2015).

Features common to both monosome and polysome aggregation plots were: (1) strong peaks at codons 1 and 5 (Figure 2.3B, labeled in right inset); and (2) a four- rather than three-nucleotide gap between the last coding position peak and final peak over the stop codon (Figure 2.3C, inset). Because the +5 codon peak did not disappear when aggregation plots were normalized so

TABLE 2.2: tRNA Mapping Statistics

Number of tRNA reads, binned by encoded amino acid, in each monosome and polysome biological replicate library.



	Raw Counts			
Amino Acid	**Mono1**	**Mono2**	**Poly1**	**Poly2**
A	189	283	475	149
C	26	74	84	24
D	1379	2643	832	504
E	451	742	929	525
F	82	140	145	54
G	6852	8350	15271	15464
H	54	81	141	57
I	168	376	315	116
K	1074	1013	5360	3756
L	437	485	1376	344
M	102	194	316	141
N	63	93	235	71
P	698	1242	480	188
Q	120	192	267	89
R	301	377	633	224
S	480	838	1051	522
T	88	191	218	54
V	586	1060	380	154
W	85	151	229	83
X	46	75	105	17
Y	106	159	423	130

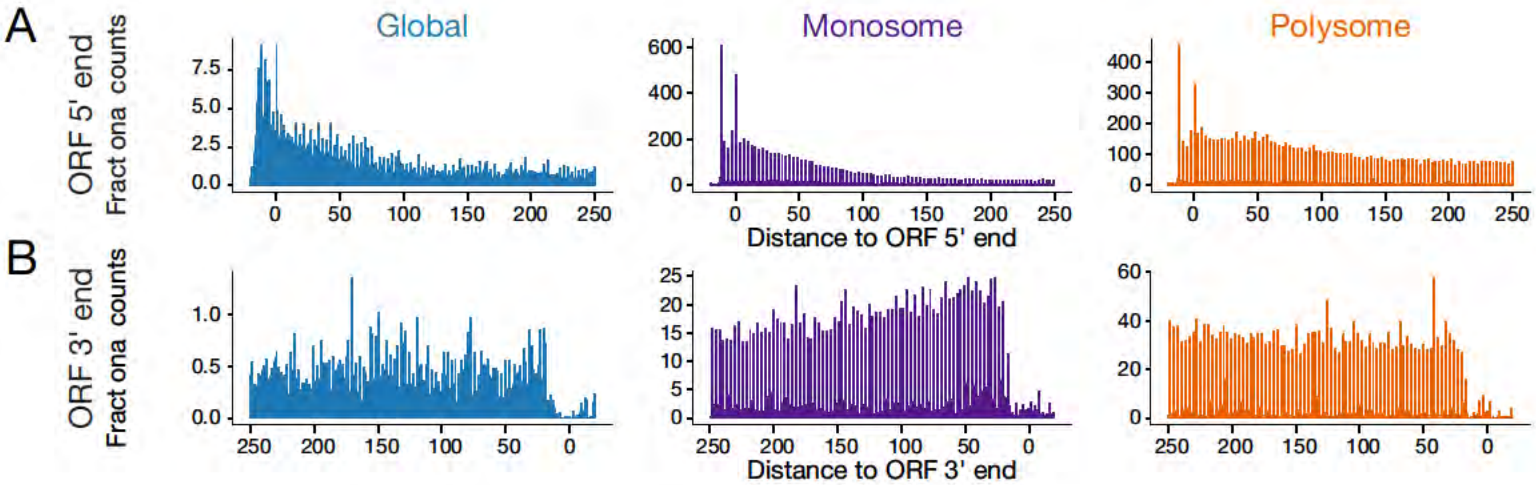


FIGURE 2.4: Aggregation Plots of ORF 5' and 3' Ends (continued on next page)

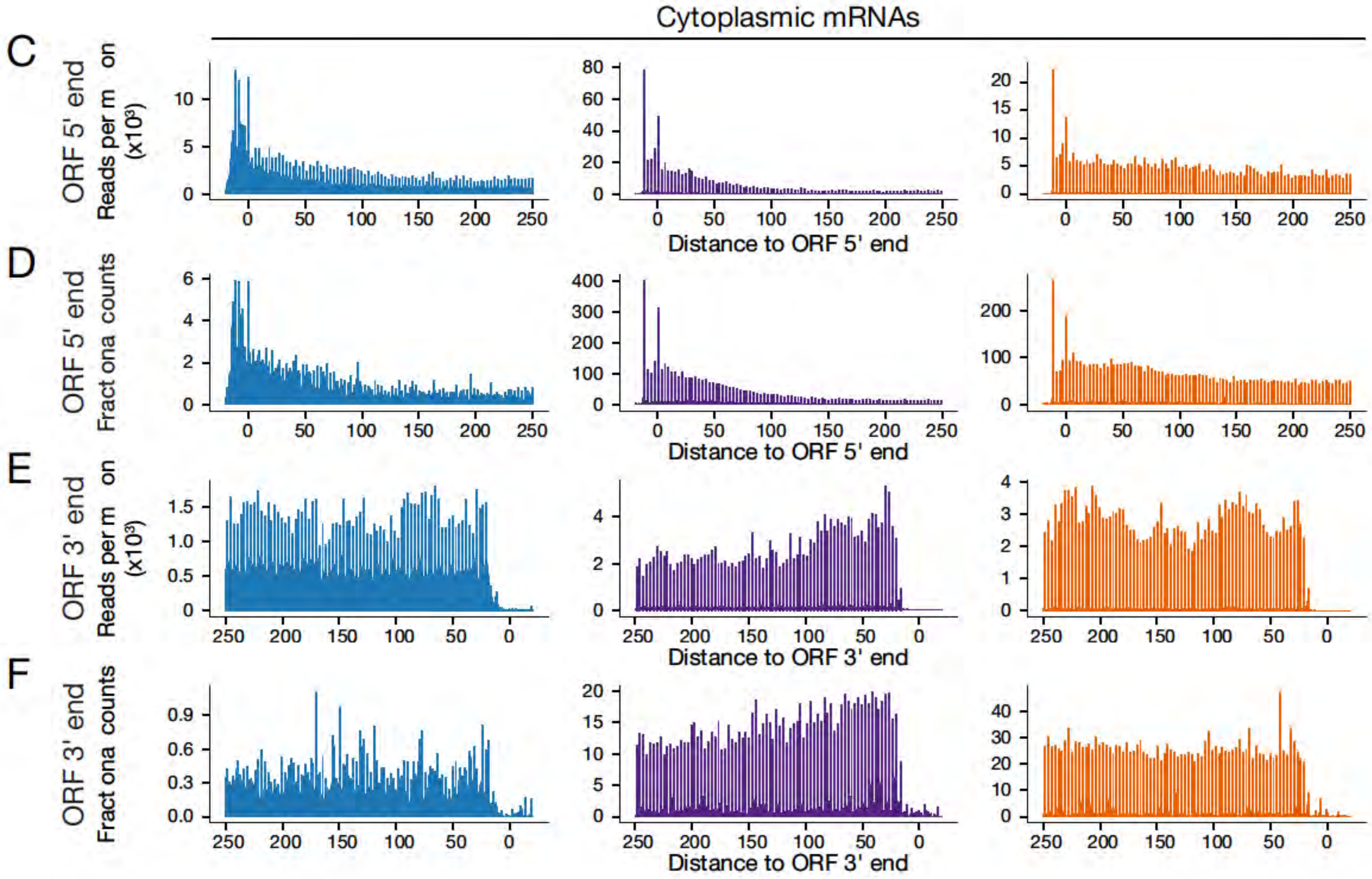


FIGURE 2.4: Aggregation Plots of ORF 5' and 3' Ends (continued on next page)

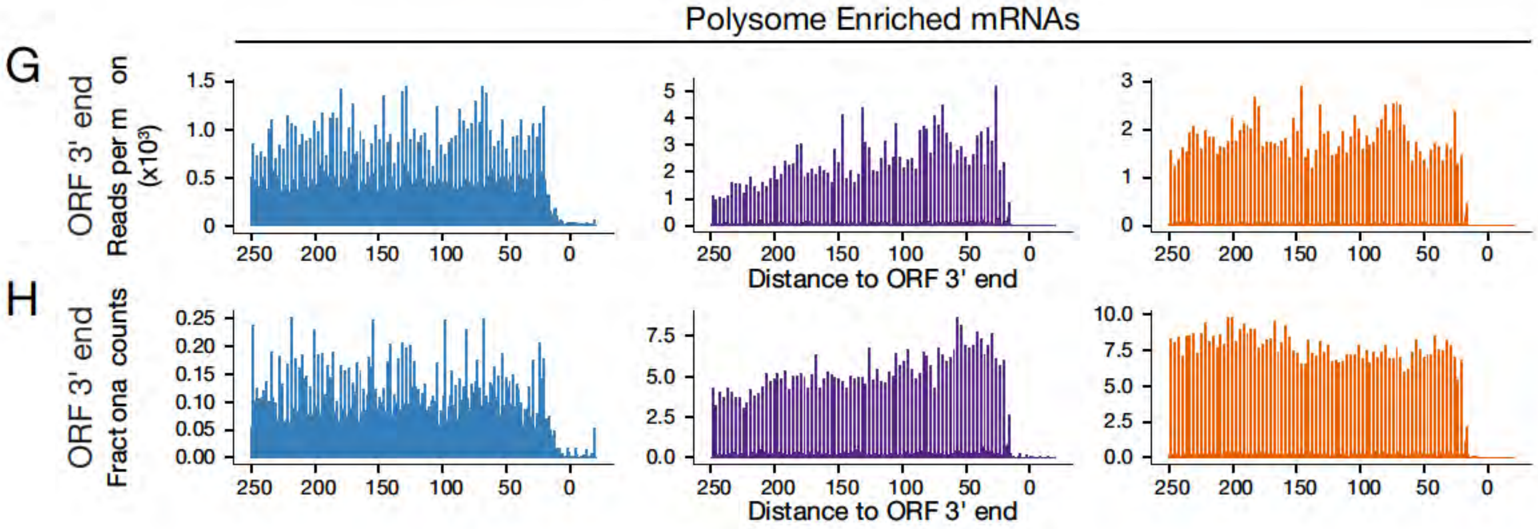


FIGURE 2.4: Aggregation Plots of ORF 5' and 3' Ends

A, B) Aggregation plots of 28 nt read 5' ends as in Figure 2.3B and C, respectively, with counts normalized such that each gene contributes equally (fractional counts). C-F) Aggregation plots limited to cytoplasmic mRNAs (membrane and secreted proteins removed) with ORFs >300 nts. C-D, ORF 5' end signal; E-F, ORF 3' end signal. In C and E, counts were normalized only for library size. In D and F, fractional counts as in A and B. G,H) Aggregation plots of ORF 3' end signal limited to polysome enriched mRNAs defined in Figure 3A. Counts were normalized only for library size (G) or are shown as fractional counts. All plots in this figure were constructed from biological replicate 1 using only uniquely mapping reads.

that each gene contributed equally (Supplemental Figure 2.4A and D), it was a general feature of our libraries and not due to any single gene or small gene subset. Thus +5 pausing may be a general post-initiation feature in both yeast and mammals, where it was recently attributed to exit tunnel geometry (Han et al., 2014). The offset peak at ORF 3' ends was also apparent in normalized aggregation plots (Supplemental Figure 2.4B and F), indicating that its generality. This previously observed spacing (Guydosh and Green, 2014) may be due to mRNA compaction during termination (Brown et al., 2015). Thus, with regard to these previously characterized features at the 5' and 3' ends of ORFs, we could detect no differences between the monosome and polysome populations.

To determine if the transcriptome-wide patterns accurately represented footprint patterns across single genes, we next examined individual ORFs. As expected from the metagene and aggregation plots, monosome footprints on many ORFs predominated at and immediately downstream of the start codon, with polysome footprints exhibiting much higher coverage across the entire ORF (e.g. SHM2 and RHR2, Figure 2.3D and E). Other genes, however, exhibited very similar patterns between the monosome, polysome, and global libraries. Some such genes encoded long and abundant proteins (e.g., actin/ACT1 and RPL16B, Figure 2.3F and G). Hence, some mRNA molecules encoding even highly abundant housekeeping genes are apparently translated by monosomes.

Features of Monosome- and Polysome-Enriched mRNAs

We next used the differential expression package DESeq2 (Love et al., 2014) to compute monosome versus polysome fold-enrichment for each mRNA (monosome:polysome score). For cytoplasmic mRNAs, our data paralleled the previous microarray estimate of ribosome number per mRNA (Arava et al., 2003), with monosome:polysome scores increasing as estimated ribosome number decreased (Figure 2.5A). This relationship did not exist, however, for membrane-associated mRNAs, which were highly skewed toward monosome occupancy (Figure 2.5B). For these mRNAs, monosome footprints accumulated over and immediately downstream of predicted signal sequences (Figure 2.5C-D). This fits the long-standing model of ER protein import (Figure 2.5E) where the signal sequence is first translated by a single cytoplasmic ribosome prior to signal recognition particle (SRP) recruitment and membrane engagement. Because membrane-associated polysomes were likely under-sampled due to lack of detergent in our cell lysis procedure, we limited all subsequent analyses to the 4,342 mRNAs for which no evidence exists of membrane association (i.e., cytoplasmic mRNAs).

A major determinant of ribosome number per mRNA is ORF length (Arava et al., 2003). Consistent with this, a scatter plot of monosome:polysome score versus ORF length revealed a strong inverse relationship, with shorter ORFs being more monosome-associated than longer ORFs (Figure 2.6A). Mean and median monosome:polysome scores of ordered bins each containing 50 genes revealed this relationship to be particularly strong and nearly linear for ORFs $\leq$590 nts (Figure 2.6B and 2.5F). Thus the shortest canonical ORFs in CDS genes tend to be occupied by a single ribosome.

Besides short canonical ORFs, two other classes of short ORFs are sORFs [short <300 nt ORFs in transcripts not originally annotated as protein-coding genes; Smith et al., 2014] and uORFs [ORFs upstream of canonical ORFs; Ingolia et al., 2009]. Both classes were strongly biased toward monosome occupancy (Figure 2.6C). Like short canonical ORFs, sORFs exhibited a negative correlation between ORF length and monosome:polysome read ratio; accordingly, sORFs are likely bona fide protein-coding genes translated predominantly by monosomes. Strikingly, a different behavior was observed for uORFs. Consistent with all currently annotated uORFs in non-membrane genes being <250 nts, the population as a whole was strongly biased toward monosome occupancy (Figure 2.6C). However, no relationship was detectable between monosome enrichment and uORF length. A likely explanation is that uORFs, by definition, are contained within multi-cistronic mRNAs whose cosedimentation with monosomes or polysomes is determined by the combined ribosome occupancy on all ORFs. Consequently, any simultaneous ribosome occupancy on multiple ORFs (even if each ORF is only occupied by a single

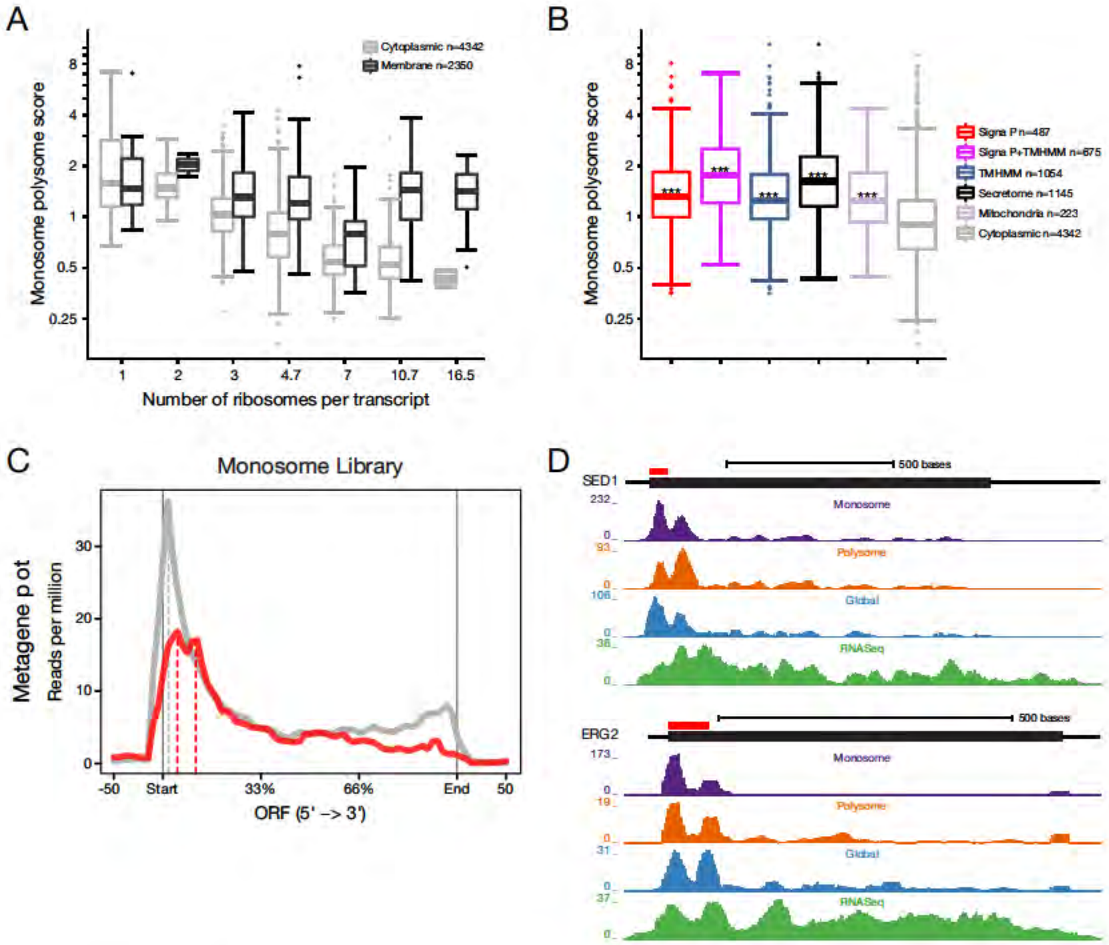


FIGURE 2.5: Comparison of Monosome:Polysome Score to Various Features

(continued on next page)

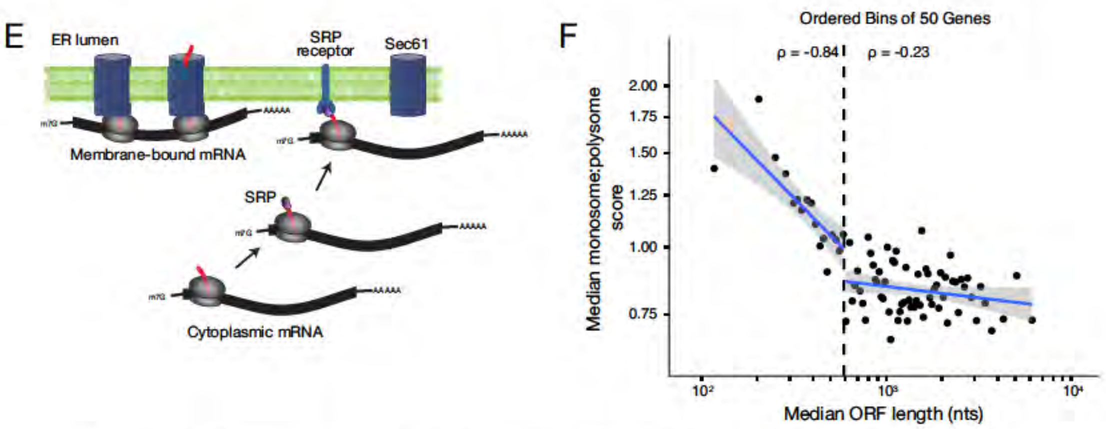


FIGURE 2.5: Comparison of Monosome:Polysome Score to Various Features

A) Boxplots comparing monosome:polysome score for 'Cytoplasmic' and 'Membrane' mRNAs to the number of ribosomes typically bound by each mRNA as calculated from gradient fraction microarray analysis (Arava et al., 2003). 'Membrane' genes combine all co-translationally targeted mRNA subsets indicated in (B); 'Cytoplasmic' genes comprise the remainder. B) Boxplot comparing monosome:polysome score for various mRNA subsets for which evidence exists of co-translational targeting to a cellular organelle. $***$ p < 2.2 x 10^{-16}, Wilcoxon rank sum test for each class compared to 'Cytoplasmic'. SignalP and TMHMM predictions are from SGD; secretome annotation from Jan et al., 2014; mitochondria from Williams et al., 2014. See extended experimental procedures for further details on gene classification. C) Metagene plots using all $\geq$25 nt reads mapping to annotated signal-sequence containing (red) or cytoplasmic (grey) mRNAs, as defined in (B). D) Distribution of $\geq$25 nt reads from indicated libraries across individual signal-sequence containing genes. Plots were constructed using only uniquely mapping reads from biological replicate 1. E) Model of SRP-dependent co-translational stalling and localization to the ER. Because all mRNAs molecules stably associated with membranes (e.g., monosomes and polysomes engaged in protein translocation via the Sec61 channel) were depleted during lysate preparation, lysates were enriched for membrane mRNA molecules in the earliest stages of translation (e.g., prior to SPR receptor engagement by the first ribosome). F) Median monosome:polysome enrichment for ordered bins each containing 50 genes. Gray shading: 0.95 confidence interval; ρ is Pearson correlation coefficient.

ribosome) will cause the entire mRNA to cosediment with polysomes. Nonetheless, the strong bias of uORF ribosome footprints toward monosome cosedimentation indicates that when a ribosome is engaged on a uORF, all other ORFs in the mRNA tend to be unoccupied.

Because sORFs and uORFs are predominantly monosome-associated, these regions had much higher occupancy in the monosome libraries than in either the polysome or global libraries. This enhanced detection suggested that monosome footprinting might prove more effective for identifying new sORFs than global ribosome footprinting, especially sORFs in monocistronic transcripts. To find new translationally-active ORFs, we combined the two monosome libraries, removed reads associated with previously annotated ORFs (canonical, sORFs and uORFs) and then identified clusters of overlapping or adjacent genome-mapping reads in the remainder. Examination of high coverage clusters revealed that most occurred either within annotated 5' UTRs or the region immediately upstream. One example is a uORF upstream of PCL5 (Figure 2.6D). Previously published 5' transcript leader sequencing data (Arribere and Gilbert, 2013) suggests the existence of alternate transcription start sites (TSSs) for PCL5. Therefore, this uORF is likely an alternatively included element regulating PCL5 translation (Pelechano et al., 2013).

To identify features other than ORF length that affect the monosome:polysome

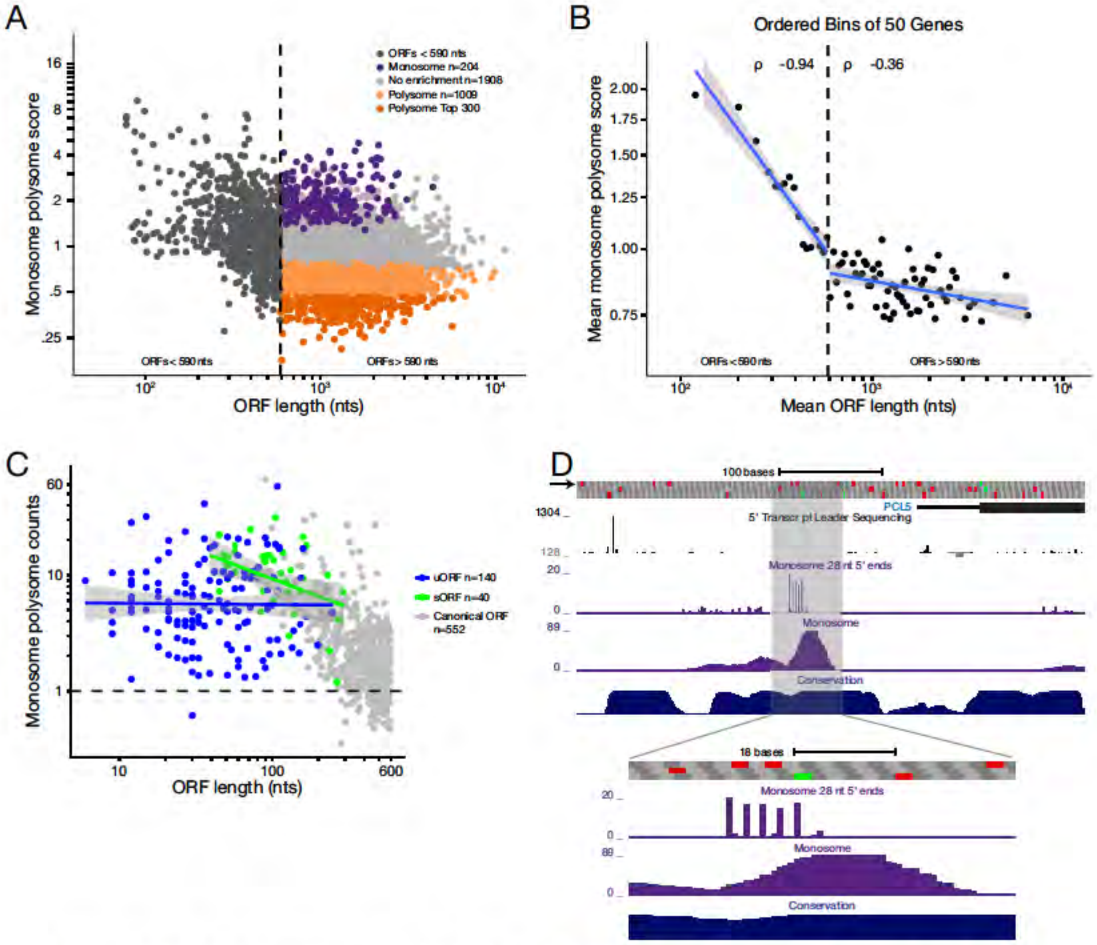


FIGURE 2.6: Relationship Between ORF Length and Monosome versus Polysome Enrichment

A) Scatterplot of monosome:polysome score versus ORF length, with monosome (purple dots), polysome (orange dots) and no enrichment (light gray dots) sets indicated. B) Mean monosome:polysome score versus ORF length for ordered bins each containing 50 genes. Gray shading: 0.95 confidence interval; ρ: Pearson correlation coefficient. C) Scatterplot of monosome:polysome count ratio versus ORF length for canonical ORFs <590 nts (Gene; gray dots), sORFs (Smith et al., 2014; green dots) and uORFs (SGD-curated list from Ingolia et al., 2009; blue dots). D) Genome browser screenshot showing monosome footprint coverage and 3-nucleotide phasing of 28 nt monosome footprints over a novel 6-codon sORF upstream of PCL5. Arrow indicates exon reading frame; green and red boxes indicate start and stop codons, respectively. Conservation track represents evolutionary nucleotide conservation across 7 *Saccharomyces* species. See also Figure 2.5.

score, we next considered only the 3,121 CDS genes with a canonical ORF >590 nts (Figure 2.6A). Within this set, DESeq2 identified 204 monosome-enriched (p-adj $\leq$ 0.001; Figure 2.6A, purple dots) and 1009 polysome-enriched (p-adj $\leq$ 0.001; Figure 2.6A, orange dots) mRNAs (Table S3). To define the most extreme set of polysome-enriched mRNAs, we also picked the 300 mRNAs exhibiting the smallest monosome:polysome score (Figure 2.6A, dark orange dots). The remaining 1908 mRNAs not meeting the above cutoffs formed the ’no enrichment’ set (p-adj > 0.001; Figure 2.6A, grey dots). We then compared various features of these four gene sets (Figure 2.7; data from this paper or previously published). Polysome-enriched genes have higher median mRNA and protein abundances than the no-enrichment and monosome-enriched sets (Figure 2.7A and F). Further, as also expected for highly expressed genes, the polysome-enriched sets exhibit higher mRNA synthesis rates and longer mRNA half-lives than either the no-enrichment or monosome-enriched set (Figure 2.7B and C). Finally, consistent with an evolutionary pressure toward more efficient translation, polysome-enriched mRNAs tend to have shorter 5' UTRs (Figure 2.7D) and a higher optimal codon frequency (Figure 2.7E). In short, polysome-enriched mRNAs tend to be highly transcribed, have long mRNA half-lives, short 5' UTRs, high codon optimality, and encode highly abundant proteins.

Monosomes, mRNA Half-Life and NMD

In the above analysis, the monosome-enriched gene set had a lower median mRNA half-life than the no-enrichment set (Figure 2.7C; data from Presnyak et al., 2015. One class of mRNAs expected to be monosome-enriched with shorter than average half-lives are those subject to nonsense-mediated mRNA decay (NMD). NMD preferentially eliminates transcripts wherein the stop codon exists in a suboptimal context for termination (Amrani et al., 2004), and NMD targeting is thought to occur when the first or 'pioneer' ribosome encounters this suboptimal stop codon (Gao et al., 2005). Two previous studies independently identified two classes of genes in *S. cerevisiae* whose mRNA levels depend on NMD: (1) those with strong evidence of being 'direct' NMD targets, and (2) those for which the evidence might indicate 'indirect' regulation by NMD (Guan et al., 2006, Johansson et al., 2007). To facilitate our own comparative analysis, we parsed these partially overlapping gene sets according to whether both studies concurred on direct NMD target status (our Class A; 33 non-membrane genes with ORF length >590 nts), a single study indicated direct NMD target status (Class B; 144 genes), and any other gene indicated as an indirect target by either study (Class C; 166 genes) (Supplemental Table 3). All other genes were placed into a 'non-NMD target' bin. Consistent with expectation, global ribosome footprint RPKM boxplots revealed that all three NMD target sets exhibited

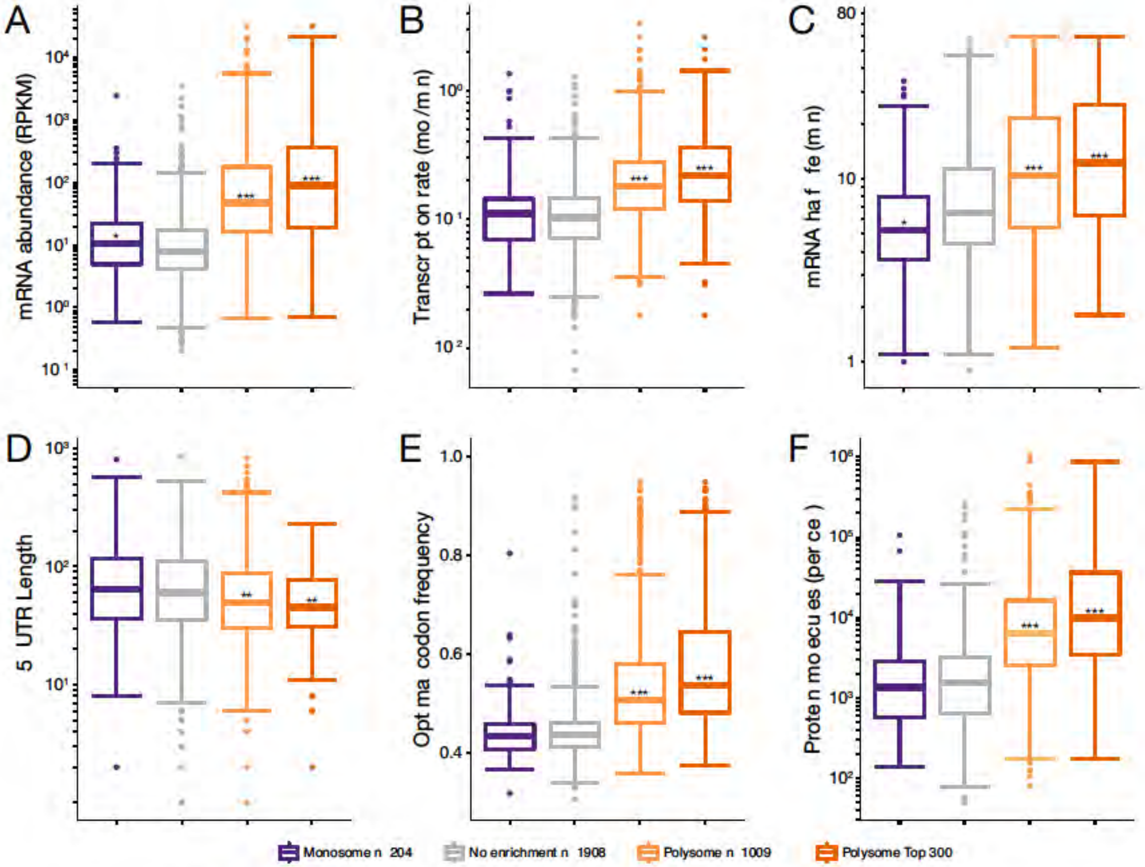


FIGURE 2.7: Features of Monosome and Polysome Enriched Genes

Boxplots comparing mRNA and protein features for the gene sets defined in Figure 2.6A. mRNA abundances were calculated from our own RNA-Seq libraries. Transcription rates (Pelechano et al., 2010), mRNA half life (Presnyak et al., 2015), 5' UTR length (Nagalakshmi et al., 2008), protein abundance (Ghaemmaghami et al., 2003), and optimal codon frequency (SGD) numbers were from indicated references. $***$ p < 2.2 x 10^{-16}, $**$ p $\leq$ 1.5 x 10^{-5}; $*$ p $\leq$ 0.0041; all others p > 0.05; Wilcoxon rank sum test compared to 'no enrichment' set.

significantly lower overall ribosome occupancy than the non-NMD set (Figure 2.8A). This tendency toward lower ribosome occupancy was also readily apparent in both a scatter plot of monosome versus polysome footprint RPKM (Figure 2.8C; quantified in Figure 2.9A and 2.9B) and boxplots of monosome:polysome scores (Figure 2.8C, inset). Unexpectedly, however, none of the three NMD target sets was statistically different from the non-NMD set with regard to mRNA half-life (Figure 2.8B). Thus, while NMD targets tend toward monosome occupancy, enrichment for NMD targets does not explain the lower median half-life of monosome-enriched mRNAs (Figure 2.7C). Consistent with this, monosome-enriched mRNA median half-life remained significantly lower than the no enrichment set even when all known NMD targets were removed (Figure 2.9D). Thus, some feature other than NMD must be driving the lower stability of monosome-enriched mRNAs.

Ribosome Occupancy on Introns

Another set of known NMD targets are intron-containing transcripts that escape the nucleus without having been spliced (Sayani et al., 2008). Of 222 non-membrane genes harboring an intron inside the canonical ORF, 130 had ≥10 total monosome footprints that either partially or completely overlapped the intron. For these 130 introns, comparing the canonical ORF monosome:polysome score to the intron monosome:polysome count ratio revealed that ribosome

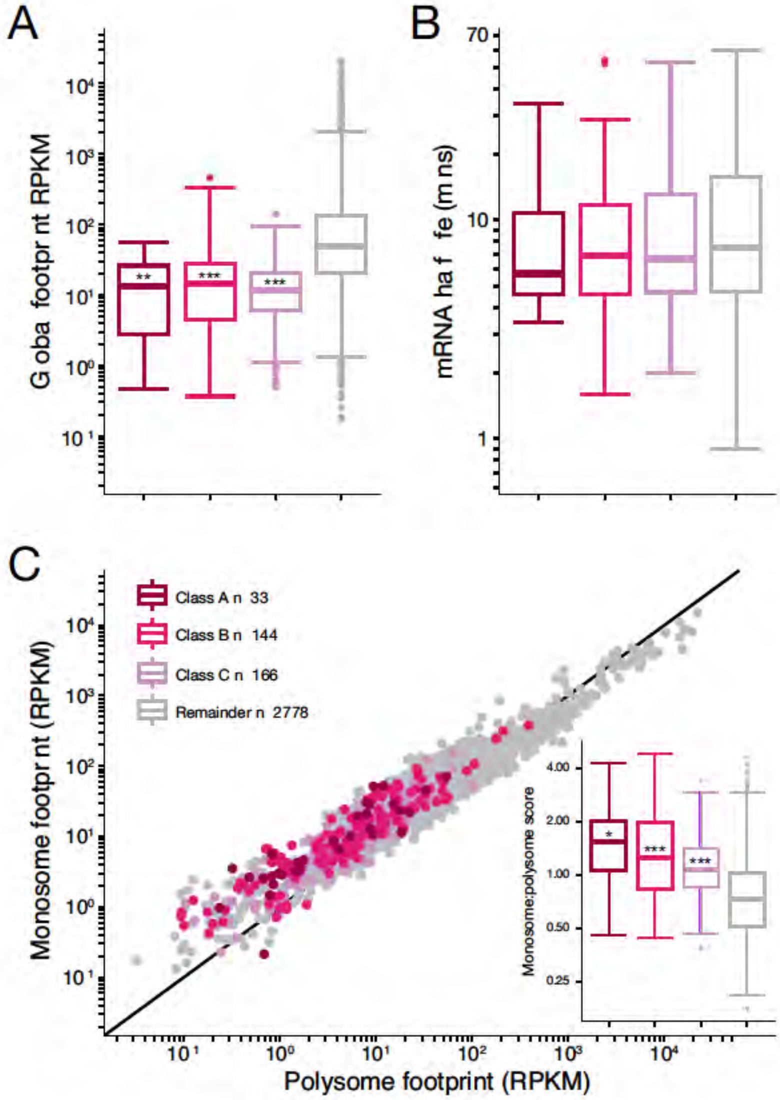


FIGURE 2.8: Characteristics of NMD-Regulated Genes

A, B) Boxplots comparing mean global footprint RPKM (A) and mRNA half-life (B; Presnyak et al., 2015) for classes of direct and indirect NMD targets (Guan et al., 2006, Johansson et al., 2007) as indicated in (C). C) Scatterplot comparing mean monosome and polysome footprint RPKM for indicated gene sets. Inset, boxplots of monosome:polysome scores for the same gene sets. $***$ $p < 2.2 \times 10^{-16}$; $**$ $p = 6.1 \times 10^{-11}$; $*$ $p = 2.0 \times 10^{-10}$; all others $p > 0.05$; Wilcoxon rank sum test compared to 'Remainder'. See also 2.9.

footprints in introns were highly skewed toward monosomes, regardless of ORF monosome:polysome score (Figure 2.10A). Aggregation of 28 nt monosome footprint 5' ends revealed strong exonic reading frame maintenance across the first 20 intronic codons (Figure 2.10B,C and 2.11A). The sharp decrease in aggregated counts after the first few intron positions was due to the presence of an in-frame stop codon near the 5' end of most introns. For genes with no early in-frame stop codon, phased monosome reads were readily observable within the intron (ex. Figure 2.10E). In other introns, phased reads were also apparent on one or more internal sORFs on which ribosomes likely reinitiated after encountering the first stop codon (ex. Figure 2.10F). The presence of such sORFs led to a decrease in phasing when all intron positions were taken into account (Figure 2.10D and 2.11B). Taken together, these data strongly support the idea that targeting of unspliced transcripts for NMD occurs predominantly on monosomes. They further demonstrate that introns can harbor translationally active sORFs, so may represent a previously unrecognized source of short peptide translation products. As with other sORFs, ribosome occupancy on intronic sORFs was only readily detectable in the monosome and not polysome or global libraries. This again highlights the usefulness of monosome footprinting for detecting low-density ribosome occupancy.

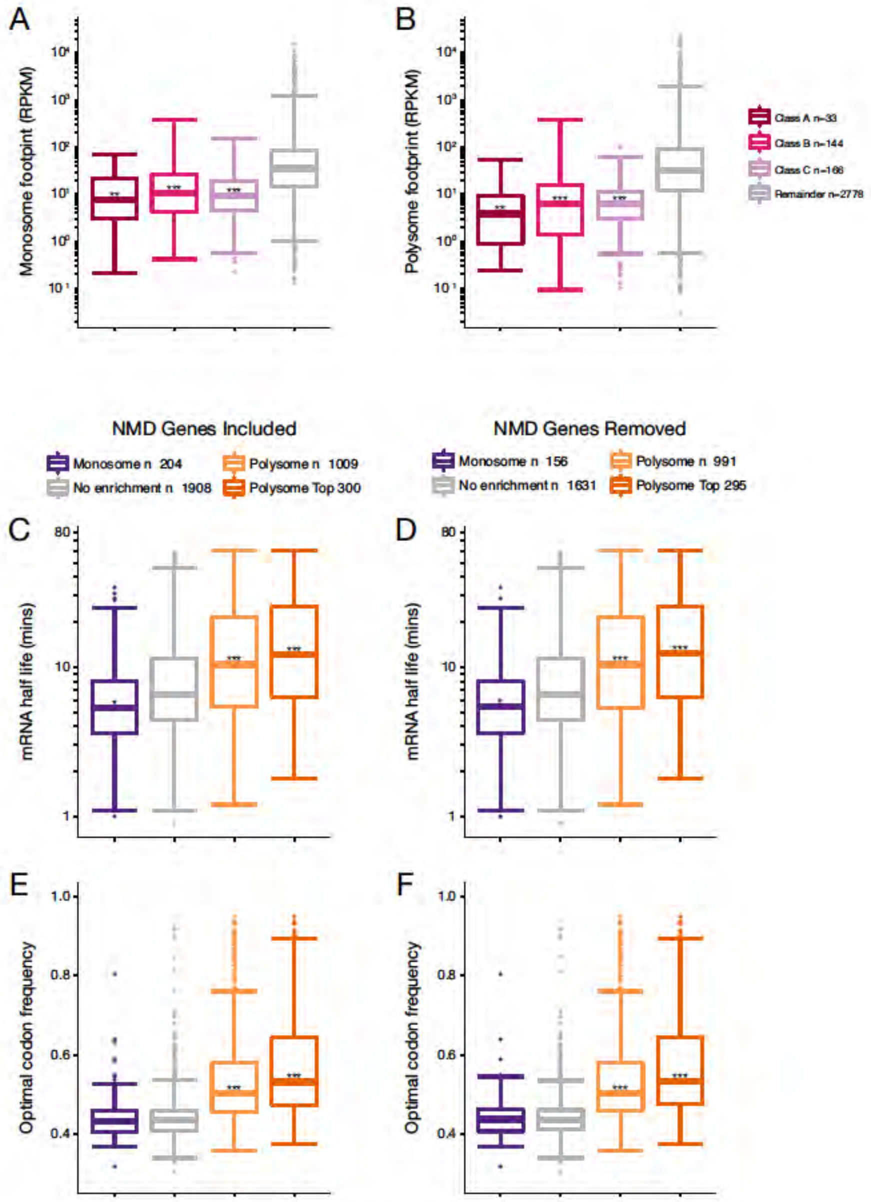


FIGURE 2.9: Features of NMD-Regulated Genes

A, B) Boxplots comparing mean ribosome footprint RPKM for NMD targets in monosome and polysome libraries, respectively. C, D) Boxplots comparing mRNA half-life for monosome- and polysome-enriched genes sets either with (C) or without (D) NMD-regulated genes. E, F) Boxplot comparing optimal codon frequency as in (C) and (D). $***$ p < 2.2 x 10^{-16}; $**$ p $\leq$ 3.2 x 10^{-9}; $*$ p $\leq$ 0.0016; all others p > 0.05; Wilcoxon rank sum test compared to either 'Remainder' or 'No enrichment'.

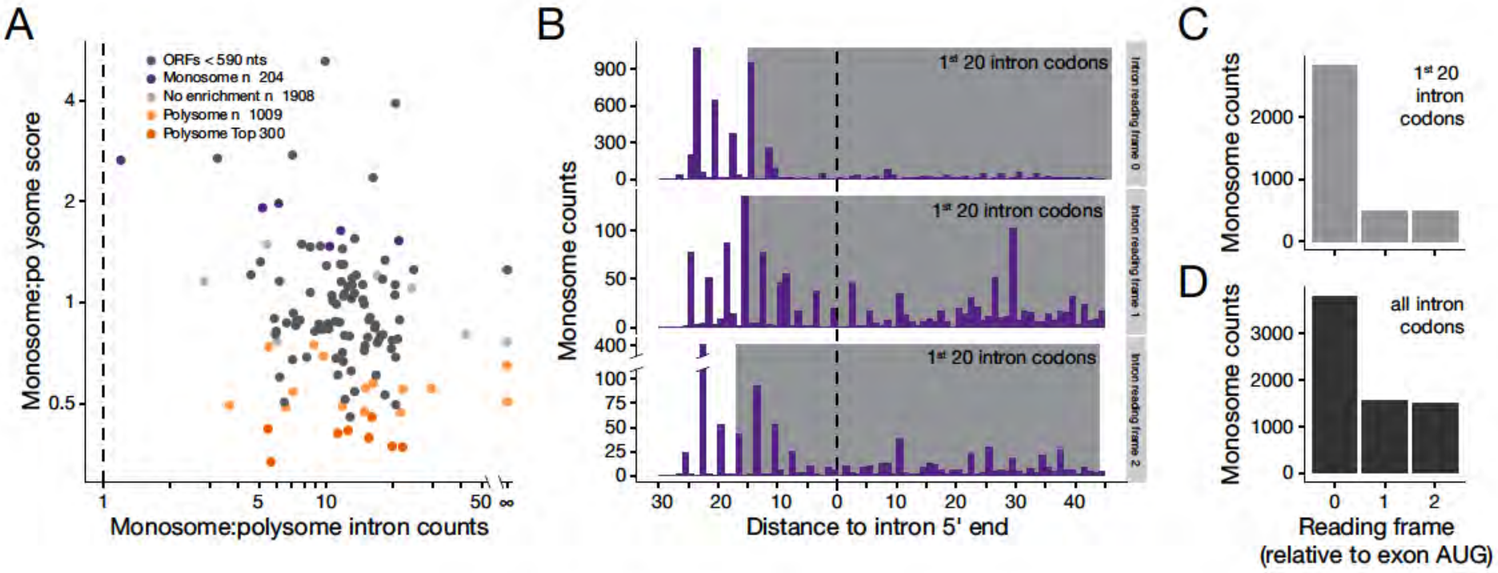


FIGURE 2.10: Monosome Footprints on Introns

(continued on next page)

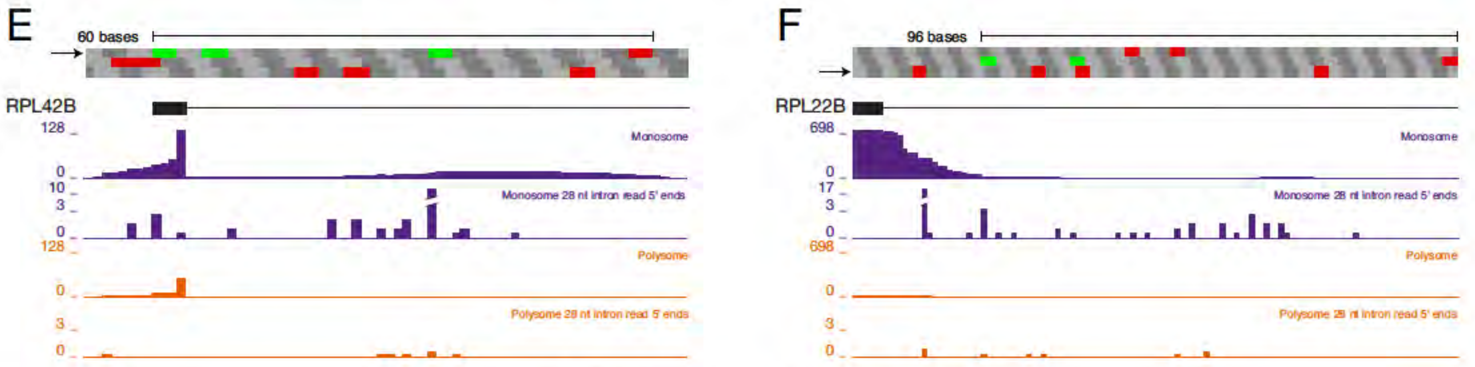


FIGURE 2.10: Monosome Footprints on Introns

A) Scatterplot showing ORF monosome:polysome score versus intronic monosome:polysome footprint ratio for intron-containing members of the gene sets defined in Figure 2.6A. B) Aggregation plots combining both monosome biological replicates of 28 nt reads overlapping the 1st intron by $\geq$ 1 nt. X-axis: Intron 5' end to read 5' end distance; Y-axis: read count; Grey boxes: 5' read end positions indicative of ribosomal A-site occupancy by intronic codons 1-20. C-D) Bar charts quantifying reading frame use across intronic codons 1-20 (C; region highlighted by grey boxes in B) or across the entire intron (D). E-F) Distribution of $\geq$25 nt reads from indicated libraries (solid plots) or 5' ends of 28 nt intronic reads across individual genes. Plots were constructed from biological replicate 1 using only uniquely mapping reads. Arrow indicates exon reading frame; green and red boxes indicate start and stop codons, respectively. See also 2.11.

Initiation and Elongation Times Determine Monosome Association

For genes with canonical ORFs long enough to accommodate more than one ribosome, why are some still predominantly monosome associated? Compared to all 3,121 cytoplasmic CDS genes with canonical ORFs >590 nts (i.e., our background population), mRNAs encoding kinases and transcription regulators were overrepresented in the 204-member monosome-enriched set (Table 2.3). Such regulatory proteins tend to be required in low copy numbers per cell. Notably, both overall ribosome density and calculated protein output were substantially lower for our monosome-enriched genes than all other gene sets (Figure 2.12A and B). Many mRNAs encoding regulatory proteins are also subject to negative translation regulation (e.g., by uORFs; see below). Consistent with this, canonical ORFs downstream of a uORF exhibited greater monosome enrichment than canonical ORFs not preceded by any annotated uORF (Figure 2.12D). Thus long ORF mRNAs typically occupied by single ribosomes tend to encode low-abundance proteins and be subject to negative translation regulation.

By integrating fixed parameters such as average cell size and ribosome abundance with numerous transcriptome-wide datasets (e.g., RNA-Seq, RiboSeq, mRNA half-life, and tRNA decoding specificity), Siwiak et al. (2010) recently established a quantitative, computational model of translation in *S.*

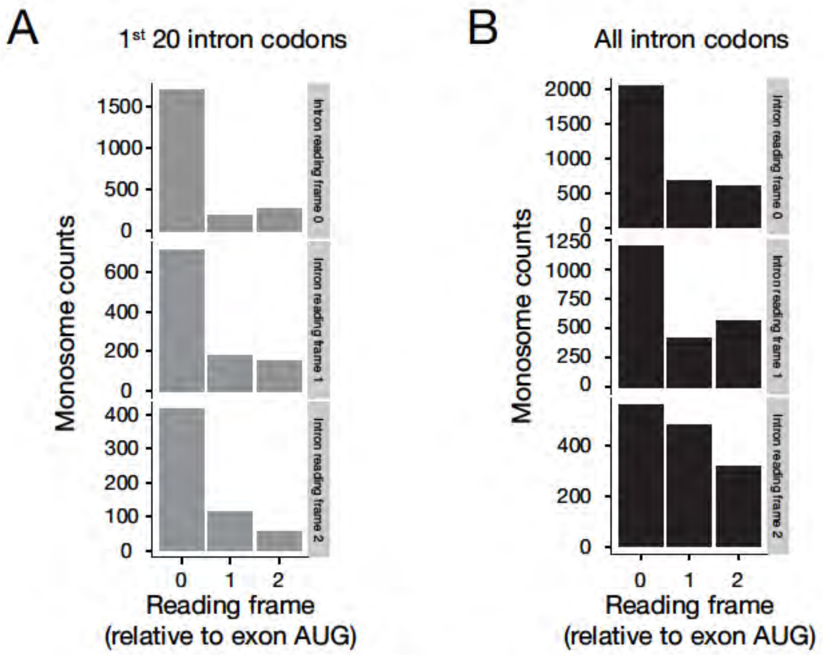
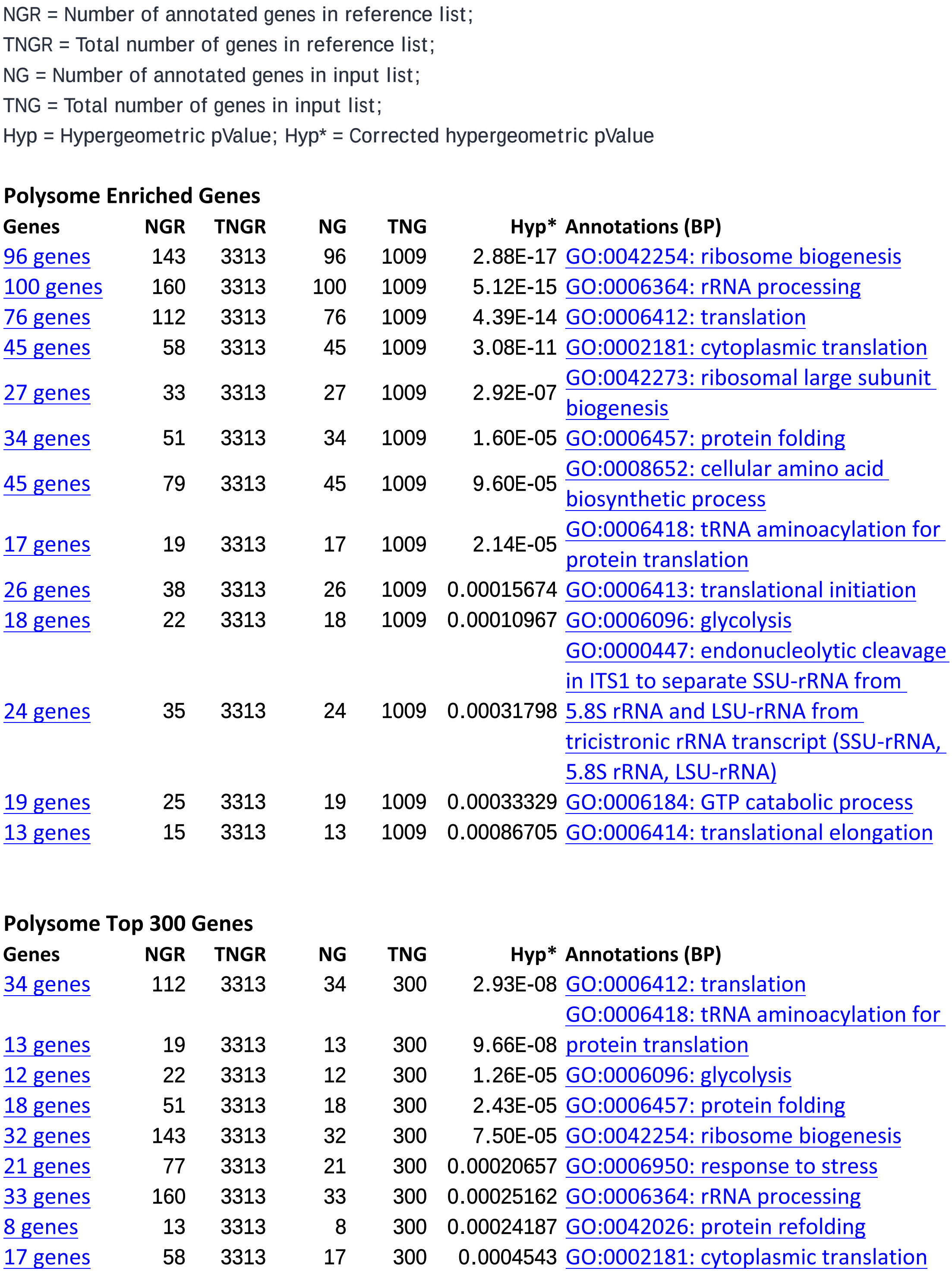


FIGURE 2.11: Intron Reading Frame

Bar charts quantifying reading frame use, separated by intron reading frame, across either the first 20 intronic codons (A; highlighted in grey boxes in Figure 2.10B) or the entire intron (B).

TABLE 2.3: Gene Ontology Analysis (continued on next page)



NGR = Number of annotated genes in reference list;
TNGR = Total number of genes in reference list;
NG = Number of annotated genes in input list;
TNG = Total number of genes in input list;
Hyp = Hypergeometric pValue; Hyp* = Corrected hypergeometric pValue

Polysome Enriched Genes

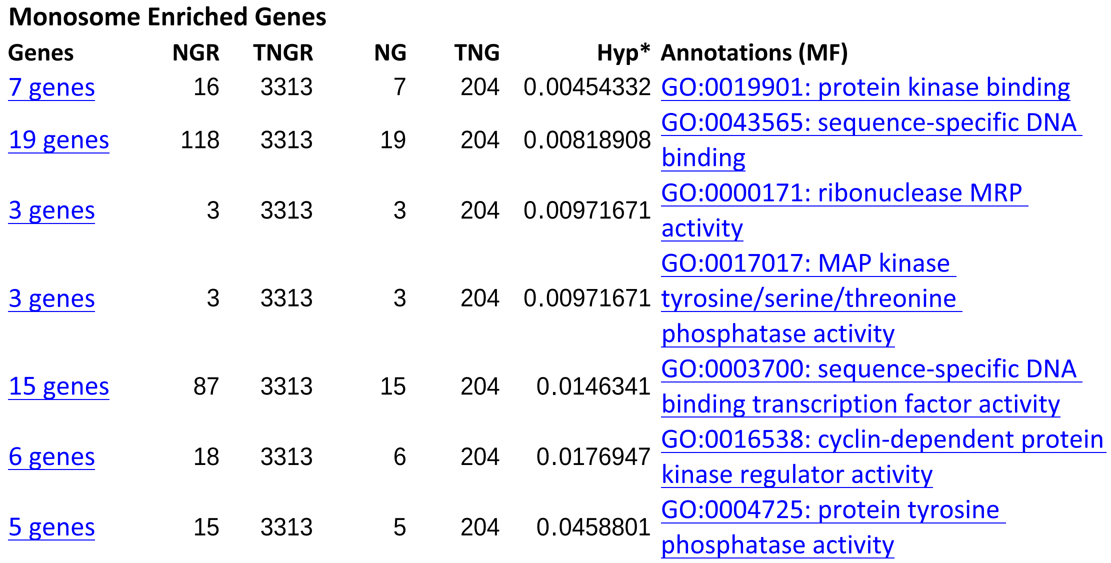
Genes	NGR	TNGR	NG	TNG	Hyp*	Annotations (BP)
96 genes	143	3313	96	1009	2.88E-17	GO:0042254: ribosome biogenesis
100 genes	160	3313	100	1009	5.12E-15	GO:0006364: rRNA processing
76 genes	112	3313	76	1009	4.39E-14	GO:0006412: translation
45 genes	58	3313	45	1009	3.08E-11	GO:0002181: cytoplasmic translation
27 genes	33	3313	27	1009	2.92E-07	GO:0042273: ribosomal large subunit biogenesis
34 genes	51	3313	34	1009	1.60E-05	GO:0006457: protein folding
45 genes	79	3313	45	1009	9.60E-05	GO:0008652: cellular amino acid biosynthetic process
17 genes	19	3313	17	1009	2.14E-05	GO:0006418: tRNA aminoacylation for protein translation
26 genes	38	3313	26	1009	0.00015674	GO:0006413: translational initiation
18 genes	22	3313	18	1009	0.00010967	GO:0006096: glycolysis
24 genes	35	3313	24	1009	0.00031798	GO:0000447: endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
19 genes	25	3313	19	1009	0.00033329	GO:0006184: GTP catabolic process
13 genes	15	3313	13	1009	0.00086705	GO:0006414: translational elongation

Polysome Top 300 Genes

Genes	NGR	TNGR	NG	TNG	Hyp*	Annotations (BP)
34 genes	112	3313	34	300	2.93E-08	GO:0006412: translation
13 genes	19	3313	13	300	9.66E-08	GO:0006418: tRNA aminoacylation for protein translation
12 genes	22	3313	12	300	1.26E-05	GO:0006096: glycolysis
18 genes	51	3313	18	300	2.43E-05	GO:0006457: protein folding
32 genes	143	3313	32	300	7.50E-05	GO:0042254: ribosome biogenesis
21 genes	77	3313	21	300	0.00020657	GO:0006950: response to stress
33 genes	160	3313	33	300	0.00025162	GO:0006364: rRNA processing
8 genes	13	3313	8	300	0.00024187	GO:0042026: protein refolding
17 genes	58	3313	17	300	0.0004543	GO:0002181: cytoplasmic translation

TABLE 2.3: Gene Ontology Analysis

Statistically significant gene ontology terms for monosome-enriched, polysome-enriched, and polysome top 300 gene sets compared to the background population of cytoplasmic mRNAs with ORFs > 590 nts.



Monosome Enriched Genes

Genes	NGR	TNGR	NG	TNG	Hyp*	Annotations (MF)
7 genes	16	3313	7	204	0.00454332	GO:0019901: protein kinase binding
19 genes	118	3313	19	204	0.00818908	GO:0043565: sequence-specific DNA binding
3 genes	3	3313	3	204	0.00971671	GO:0000171: ribonuclease MRP activity
3 genes	3	3313	3	204	0.00971671	GO:0017017: MAP kinase tyrosine/serine/threonine phosphatase activity
15 genes	87	3313	15	204	0.0146341	GO:0003700: sequence-specific DNA binding transcription factor activity
6 genes	18	3313	6	204	0.0176947	GO:0016538: cyclin-dependent protein kinase regulator activity
5 genes	15	3313	5	204	0.0458801	GO:0004725: protein tyrosine phosphatase activity

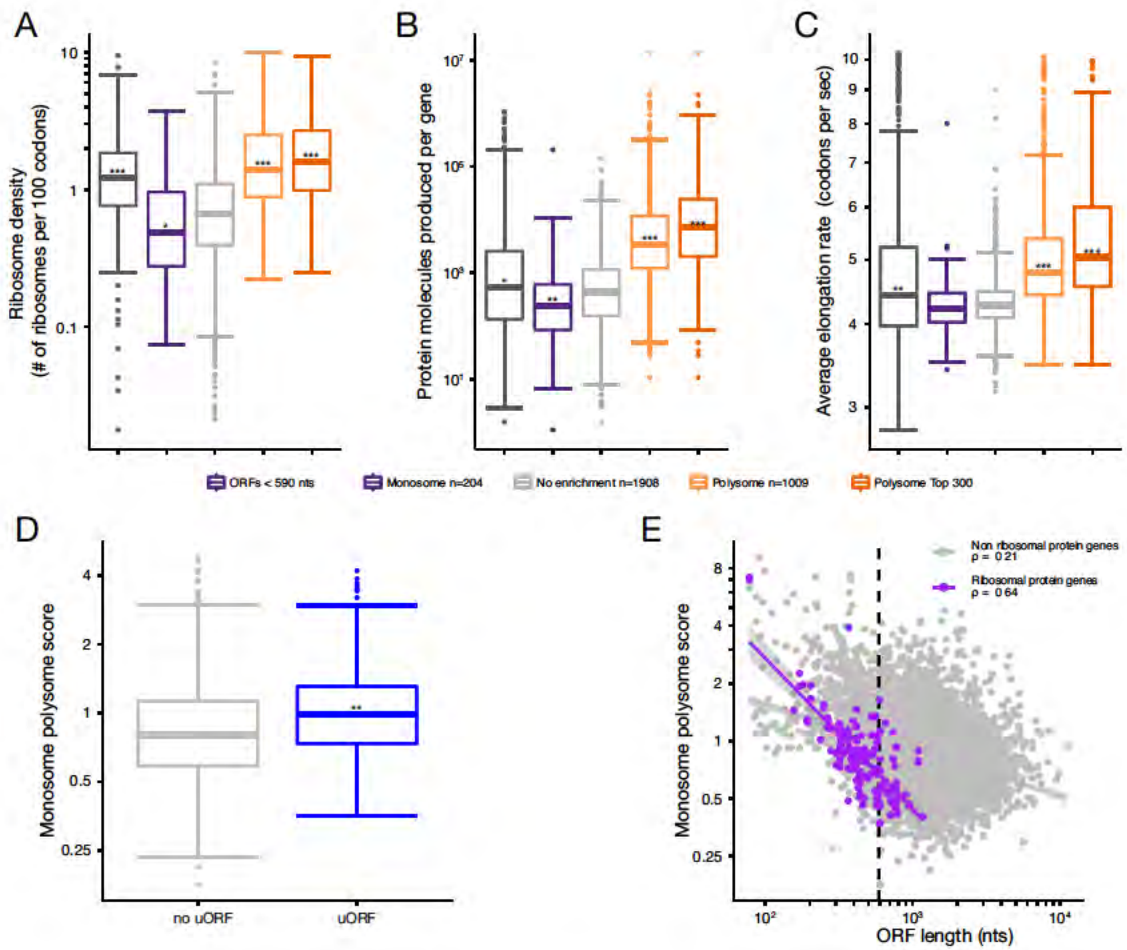


FIGURE 2.12: Translation Measurements for Defined Gene Sets

(continued on next page)

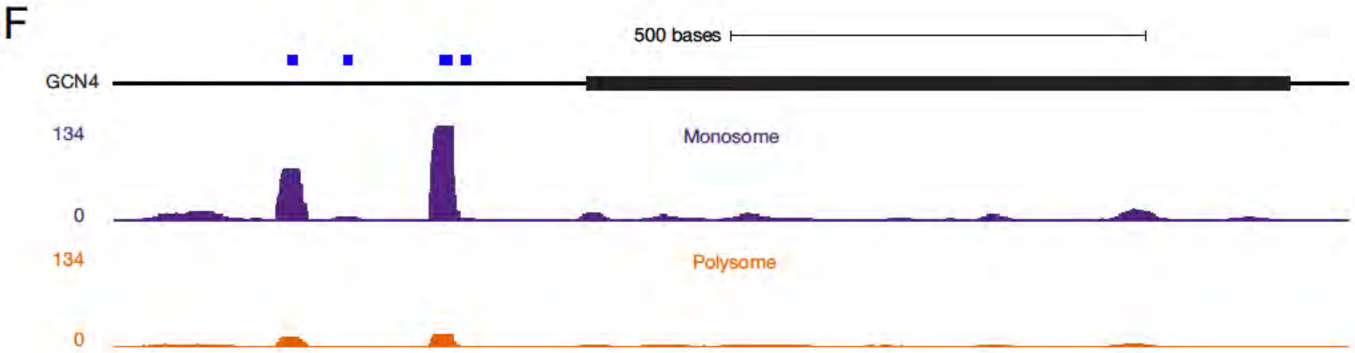


FIGURE 2.12: Translation Measurements for Defined Gene Sets

A-C) Boxplots comparing calculated ribosome density (A), protein molecules produced per gene (B), and average elongation rate (C) for the gene sets defined in Figure 2.6A; data from Siwiak and Zielenkiewicz, 2010. $\ast\ast\ast$ p < 2.2 x 10^{-16}; $\ast\ast$ p $\leq$ 1.6 x 10^{-6}; $\ast$ p < 0.0015; all other p > 0.02; Wilcoxon rank sum test compared to 'No enrichment'. D) Boxplot comparing monosome:polysome score for canonical ORFs with or without a uORF; $\ast\ast$ p = 1.1 x 10^{-11} Wilcoxon rank sum test compared to no uORF gene set. E) Scatterplot of monosome:polysome score versus ORF length, with ribosomal protein genes (magenta dots) indicated. Dotted line indicates 590 nt length cutoff. Gray shading: 0.95 confidence interval; ρ: Pearson correlation coefficient. F) Distribution of $\geq$25 nt reads from indicated libraries across entire GCN4 transcript. Blue bars represent uORFs. Plots were constructed using only uniquely mapping reads from biological replicate 1.

cerevisiae from which various statistics (e.g., initiation time = time required to form a new 80S ribosome at the start codon; total elongation time = time required for a ribosome to elongate through the entire ORF) were calculated for each mRNA. In theory, the number of ribosomes occupying an mRNA should be a function of initiation time versus total elongation time. If initiation time is considerably longer than total elongation time, then an mRNA should on average be occupied by zero or one ribosome, but rarely by two or more. Consistent with this, the ratio of initiation time to total elongation time was highest for our two monosome-enriched populations (the ORF <590 nt and ORF >590 nt monosome-enriched sets) compared to all other gene sets (Figure 2.13C). For the ORF <590 nt set, the major driver of this ratio was short total elongation time (Figure 2.13B) due to ORF length. Conversely, for the >590 nt monosome-enriched set, the major driver was initiation time (Figure 2.13A). Thus many long ORFs tend toward monosome occupancy due to slow initiation rates.

The relationship between initiation and elongation times also leads to different monosome and polysome footprint patterns across individual genes (Figure 2.13D and F). For mRNAs where initiation time is substantially slower than total elongation time, any ribosome occupying that mRNA will generally be in the process of elongation. It follows that mRNAs with extremely slow initiation rates should be predominantly monosome-associated, with ribosome footprints

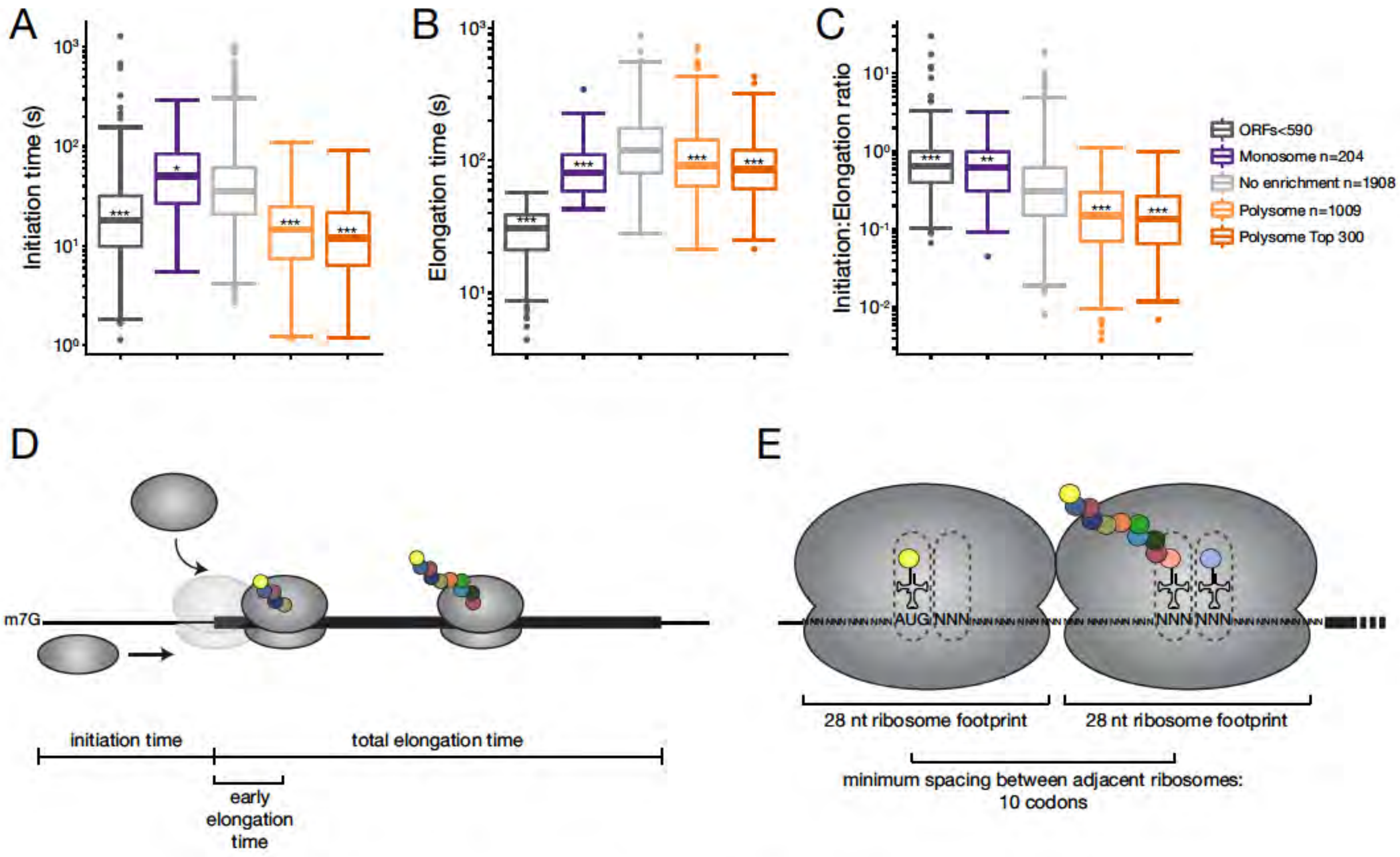


FIGURE 2.13: Relationship Between Initiation and Elongation Times Determines Degree of Monosome versus Polysome Association (continued on next page)

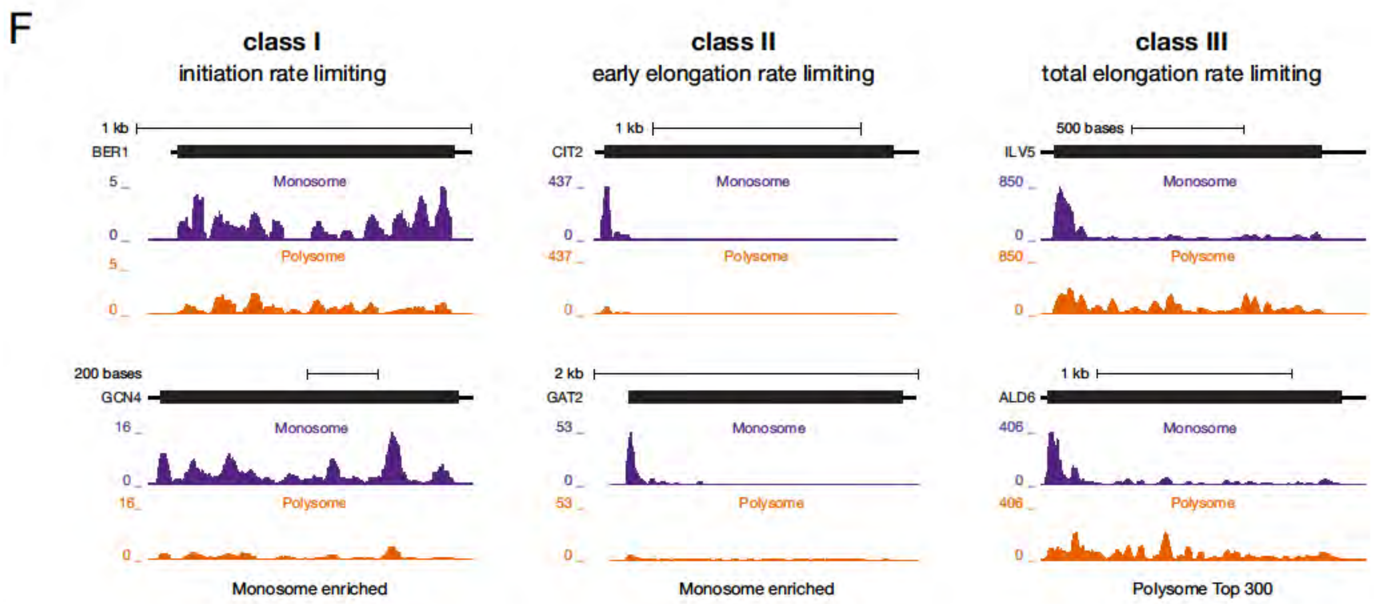


FIGURE 2.13: Relationship Between Initiation and Elongation Times Determines Degree of Monosome versus Polysome Association

A-C) Boxplots comparing initiation time (A), elongation time (B), or the ratio of initiation:elongation time (C) for the gene sets defined in Figure 3A; data from Siwiak and Zielenkiewicz, 2010. $***$ $p < 2.2 \times 10^{-16}$; $**$ $p = 7.5 \times 10^{-15}$; $*$ $p = 5.2 \times 10^{-5}$; all others $p > 0.05$; Wilcoxon rank sum test compared to 'No enrichment'. D) Schematic showing the relative position of ribosomes and individual subunits during initiation, early elongation and total elongation. Opaque ribosome over start codon indicates block to 80S formation while another ribosome occupies the region immediately downstream of the start codon. E) Schematic showing minimum spacing between adjacent ribosomes. F) Distribution of $\geq$25 nt reads from monosome and polysome libraries across genes representative of classes discussed in text. See also Figure 2.12.

distributed across the entire ORF (Figure 2.13F, class I). One example is BER1, a regulator of microtubule stability involved in proper kinetochore function (Fiechter et al., 2008). Another is GCN4, a highly studied transcription factor required for up-regulation of amino acid biosynthesis upon starvation. Initiation on the canonical GCN4 ORF is regulated by four uORFs; translation across a uORF generally decreases downstream re-initiation efficiency (Hinnebusch, 2005). The very high enrichment of all four GCN4 uORFs in the monosome libraries (monosome:polysome counts = 9.0, 8.3, 10.5 and 6.1, respectively; Figure 2.12F) indicates a strong preference for only one GCN4 ORF (one uORF or the canonical ORF) to be occupied at a time. Consistent with this, the canonical ORF exhibited strong monosome enrichment (monosome:polysome counts = 3.7), with footprints distributed throughout its entire length (Figure 2.13F, class I). Therefore, the rate-limiting step for GCN4 translation during logarithmic growth in rich media is initiation on the canonical ORF.

Other mRNAs primarily occupied by monosomes are those on which a newly initiated 80S lingers for an extended time either at the start codon (i.e., the transition from initiation into elongation is extremely slow) or immediately downstream (i.e., elongation through codons 2-9 is extremely slow) (Figure 2.3D and 2.13F, class II). While the possibility of new initiation events occurring during sample workup always warrants cautious interpretation of reads at ORF 5' ends (Gerashchenko and Gladyshev, 2014), any ribosome occupancy at the beginning of the ORF necessarily prevents a second ribosome from assembling over the start codon due to steric hindrance (see Figure 2.13E schematic). The footprint pattern expected for class II genes is high monosome signal limited to the very beginning of the ORF, combined with low polysome signal wherein the footprint distribution exhibits a strong peak akin to the monosome peak at the beginning of the ORF and low but even coverage across the remainder.

Examples of class II genes were readily apparent in the 204-member monosome-enriched set. For SHM2, most of the monosome reads occurred immediately over the start codon, with the remainder of the ORF only occupied in polysomes (Figure 2.3D). The reason for AUG stalling on SHM2 mRNA may be the highly suboptimal CCU proline codon (Artieri and Fraser, 2014, Gardin et al., 2014) at position 2. Regardless of the cause, the transition from initiation to elongation is clearly rate limiting for SHM2 translation in logarithmically growing yeast. CIT2 and GAT2 are paradigmatic examples of mRNAs for which elongation through codons 2-9 is rate limiting for overall translation (Figure 2.13F, class II). For both, monosome footprints were confined to the beginning of the ORF, with the rest of the ORF only exhibiting low ribosome occupancy in the polysome libraries. Slow transit at the beginning of an ORF might be due to highly suboptimal codons in this region. It should be noted, however, that suboptimal codons tend to be enriched at ORF 5' ends transcriptome-wide (Tuller et al., 2010), and when we calculated optimal codon frequency and codon adaptation index across codons

2-9, we found no statistically significant difference between class II genes and any other gene set (data not shown). Consequently, we currently have no clear explanation for slow ribosome transit across codons 2-9 in class II genes, though codon arrangement could certainly play a role (Ciandrini et al., 2013).

At the opposite end of the spectrum with regard to initiation rate are genes encoding high abundance proteins (Figure 2.13F, class III). Both the 1009-member and top-300 polysome-enriched sets with ORFs >590 nts (Figure 2.6A) are highly enriched in mRNAs encoding proteins involved in translation, RNP biogenesis and general metabolism (e.g., amino acid biosynthesis, glycolysis) (Table 2.3). Because each mRNA molecule must turn out massive amounts of protein during logarithmic growth (Figure 2.12B), these genes have the shortest initiation times (Figure 2.13A) and the highest elongation rates (i.e., codons per second; Figure 2.12C). Their translation is limited only by the time required to complete elongation (Figure 2.13B). Therefore, the paradigmatic footprint pattern for this class is high and uniform density across the ORF in the polysome libraries, with monosome reads predominating at ORF 5' ends (Figure 2.13F, class III). For such highly translated genes (as exemplified by mRNAs encoding ribosomal proteins; Figure 2.12E), their relative association with monosomes or polysomes is almost entirely a function of ORF length.

In summary, the above results clearly demonstrate that the ratio of total time required to complete initiation and liberate the start codon for occupancy by another ribosome versus total time required to complete elongation is a major factor determining polysome versus monosome association. Whereas mRNAs with long ORFs and high initiation rates tend to be translated primarily on polysomes, mRNAs with short ORFs and/or slow initiation rates are predominately occupied by monosomes.

Discussion

In this paper, we examined the translational status of 80S monosomes in *S. cerevisiae*. Countering the widespread notion that translationally-active mRNAs are limited to polysomes, we found ample evidence for translation elongation by monosomes. Strong 3-nt phasing of monosome footprints at both 5' and 3' ends of ORFs in transcriptome-wide aggregation plots indicates that monosomes can both initiate and complete elongation. Indeed, the vast majority of monosome footprints were located downstream of the start codon, and 75% of canonical ORFs exhibited internal monosome occupancy. Thus, for most mRNA species, some fraction of molecules is occupied by a single, translationally-active ribosome. For species with short ORFs or slow initiation rates, the majority of mRNA molecules are monosome-associated. Monosomes also predominate on NMD targets, unspliced pre-mRNAs and mRNAs encoding low abundance regulatory proteins. Therefore, the 80S monosome fraction should no longer be viewed as translationally inactive. Rather, monosomes are key contributors to the overall cellular translatome.

The First Round of Translation and Translational Ramps

When aggregated across all genes, global ribosome footprints tend to peak at ORF 5' ends, sharply decrease across the first 30-40 codons and then gradually reach a plateau that persists throughout the remainder of the ORF (Ingolia et al., 2009). This pattern has been proposed to reflect an evolutionarily-conserved 'translational ramp', a region containing suboptimal codons through which newly initiated ribosomes elongate slowly before speeding up to maximal efficiency within the ORF body, presumably to minimize ribosome traffic jams and thereby the energetic cost of protein synthesis (Tuller et al., 2010). Our data do not support the translational ramp hypothesis. If all newly initiated ribosomes first proceed slowly, the same sharp footprint drop-off at ORF 5' ends should occur regardless of whether or not the mRNA is concurrently occupied by other ribosomes. However, our polysome and monosome libraries displayed quite different profiles, with polysomes almost completely lacking a ramp and monosomes having an even more pronounced ramp than the global libraries (Figure 2.3B, 2.4A, 2.4C-D). We conclude that the observed global ramp is almost entirely due to the monosome component.

Why do monosome aggregation plots display such a steep ramp at the beginning of the ORF? When any mRNA transitions from the free mRNP pool to the translationally-active pool, the first several codons must necessarily be translated by a monosome. This is because a second ribosome cannot form at the start codon until the first has moved sufficiently far into the ORF that it no longer sterically blocks a second from assembling (Figure 2.13E). Based on the known lengths of ribosome footprints, 10 codons is the minimum spacing between the first elongating ribosome and a second at the AUG (i.e., codon 11 in the P site of the first ribosome and the AUG in the P site of the second ribosome). Upon assembly of the second ribosome, the mRNA becomes a polysome, so is thereby removed from the monosome pool. The dramatic decrease starting after codon 9 in the transcriptome-wide monosome aggregation plots fully supports this minimum spacing (Figure 2.3B, 2.4A, 2.4C-D). This pattern was also clearly present on individual mRNAs (Figure 2.3E and 2.13F, class II). We therefore conclude that the apparent 'translational ramp' in global footprint aggregation plots is simply due to steric constraints on assembly of multiple ribosomes at the 5' ends of ORFs.

Monosomes, NMD Targets and mRNA Half-Lives

Nonsense-mediated decay (NMD) is a cellular process that degrades both aberrant mRNAs containing premature termination codons and regulates a subset of wild-type mRNAs (He et al., 2003). It has been proposed that NMD occurs predominantly as a consequence of the first round of translation (Culbertson and Neeno-Eckwall, 2005, Gao et al., 2005). If so, NMD substrates should exhibit lower than average overall ribosome occupancy in global footprinting experiments and be predominantly monosome-associated. Both expectations proved valid for three non-overlapping sets of previously identified NMD targets (Figure 2.8B and C, inset). Surprisingly, however, no NMD target set was statistically different from the non-NMD set with regard to mRNA half-life (Figure 2.8A). Consequently, a predominance of NMD targets cannot explain the lower median mRNA half-life of the 204-member monosome-enriched gene set compared to the no-enrichment set (Figure 2.6C). Consistent with this, removal of all NMD targets only negligibly affected median half-lives of the monosome-enriched, no-enrichment and polysome-enriched gene sets (Figure 2.9C and D), with the difference between the monosome-enriched and no-enrichment median half-lives still being highly significant ($p = 0.001$). So why do monosome-enriched mRNAs have shorter half-lives? Overall codon optimality was recently shown to be the major determinant of mRNA half-life in *S. cerevisiae* (Presnyak et al., 2015). However, even with the NMD targets removed, we found no statistically significant difference between the monosome-enriched and no-enrichment sets with regard to codon optimality (Figure 2.9E and F). In the end, while our data do indicate a relationship between mRNA half-life and monosome occupancy, we have yet to find a mechanistic explanation.

Monosomes, sORFs and Biologically-Acive Peptides

Recent work over diverse organisms has discovered the existence of thousands of biologically-active peptides synthesized directly from sORFs rather than being proteolytically cleaved from larger precursors (for reviews, see Chu et al., 2015, Landry et al., 2015, Storz et al., 2014). If we are to understand the complete repertoire of such peptides, new methods for identifying translationally-active sORFs are required. Two recent studies expressly sought to accomplish this by ribosome profiling (Aspden et al., 2014, Smith et al., 2014). Because both analyses were limited to transcripts cosedimenting with polysomes, however, only those sORFs long enough to accommodate two or more ribosomes could be interrogated. Not surprisingly, our datasets reveal a very strong relationship between ORF length and monosome:polysome score, with the shortest canonical ORFs being highly monosome-enriched (Figure 2.6A and B). This same trend exists for the *S. cerevisiae* sORFs identified above (Smith et al., 2014; Figure 2.6C), suggesting that identification of new sORFs is better accomplished by monosome profiling. Our own datasets have already revealed a conserved uORF upstream of the canonical PCL5 ORF (Figure 2.6D), as well as several translationally-active sORFs in introns (ex. Figure 2.10F). Thus, monosome profiling is a highly effective method for expanding the universe of sORFs that either serve as translational regulators (e.g., uORFs) and/or sources of new biologically-active peptides.

Monosome Association: A Function of Initiation Versus Elongation

In addition to transcripts with sORFs and uORFs, scores of mRNAs with canonical ORFs were highly enriched in the monosome fraction. Some encode high abundance species such as ribosomal proteins (RPs). Because RPs are constantly required to produce new ribosomes during logarithmic growth, RP mRNAs are among the most efficiently translated of all mRNAs. The two shortest RP ORFs are both 75 nts and encode RPL41A and B. Based on the number of ribosomes per cell and *S. cerevisiae* doubling time in rich media, Warner estimated that initiation and completion of RPL41A/B translation requires only ~2 secs. If the combined rates of elongation and termination are faster than initiation, then, at steady state, the preponderance of short, highly-translated mRNAs should be associated with just a single ribosome. Consistently, both RPL41 mRNAs were previously shown to be predominantly monosome-associated (Yu and Warner, 2001) and they were among the most highly monosome-enriched canonical ORF transcripts in our datasets (Figure 2.6A).

RPL41A and B illustrate the general principle that any mRNA will be predominantly monosome-associated if the combined time required for elongation and termination is much shorter than the time required for initiation. Because total elongation time strongly depends on ORF length (Siwiak and Zielenkiewicz, 2010), monosome:polysome score should be a function of ORF length up the point where the elongation phase is of similar duration to the initiation phase. This likely explains the steep slope and tight correlation between mean/median monosome:polysome score and ORF length up to 590 nts (Figure 2.6B and 2.5F). Beyond this inflection point, most mRNAs are predominantly polysome-associated because the elongation phase is now longer than the initiation phase. Nonetheless, even among mRNAs with ORFs >590 nts, many remained predominantly monosome-associated (our 204-member monosome gene set; Figure 2.6A). This set has significantly longer initiation times than all other gene sets (Figure 2.13A), driving the initiation:elongation ratio to be comparable to the ORF <590 nts gene set (Figure 2.13C). Therefore, mRNAs with very long initiation times are predominantly monosome-associated. Included among mRNAs with long initiation times are those subject to negative translation regulation and those encoding low-abundance regulatory proteins.

Perspective

The long-standing assumption that all translation occurs on polysomes and, therefore that 80S monosomes are translationally inactive, has had important ramifications in multiple biological systems. For instance, studies that focus solely on mRNAs co-sedimenting with polysomes (e.g, Aspden et al., 2014, Krishnan et al., 2014, Reboll and Nourbakhsh, 2014, Smith et al., 2014) will severely underestimate translational flux for mRNAs on which initiation is significantly slower than elongation and termination. These include mRNAs with sORFs, mRNAs with long and highly structured 5' UTRs, and those on which translation initiation is subject to negative regulation under the particular cellular conditions are being examined. Pre-selection of polysomes also eliminates the possibility of identifying sORFs that are only long enough to accommodate a single ribosome.

The 'polysome-only' assumption has also served as a strong and longstanding argument against localized translation in mature mammalian axons, where visible polysomes are generally lacking (Holt and Schuman, 2013, Steward and Schuman, 2003). Even in dendrites where localized translation is well established, the polysome-only assumption leads to large discrepancies between biochemical measurements of translation and the amount of translation theoretically possible based only on visible polysome numbers per dendritic spine (Ostroff et al., 2002). Our finding that mRNAs encoding key regulatory factors and other low-abundance proteins are predominantly translated by monosomes in *S. cerevisiae* opens the possibility that monosomes are also active in neuronal processes, where many polypeptides required for modulating synaptic strength are required at low stoichiometries per synapse (Sheng and Hoogenraad, 2007, Sheng and Kim, 2011). At least one of these synaptic modulators, Arc/Arg3.1, is a natural NMD target (Bicknell et al., 2012, Giorgi et al., 2007), and so may be preferentially monosome-associated as are *S. cerevisiae* NMD targets (Figure 2.8C, inset).

Accession Numbers

All sequencing data have been deposited in NCBI's GEO database under accession number GSE76117.

Acknowledgements

We thank John Mattick, Martin Smith and other Mattick lab members for bioinformatic training and guidance, and Nicholas Ingolia, Christopher Nicchitta, Elisabet Mandon, Hakan Ozadam and other Moore lab members for helpful discussions and critical manuscript review. This research was funded by HHMI.

Experimental Procedures

Extended Experimental Procedures are provided as Supplemental Information.

Ribosome Footprinting

BY4741 yeast were grown in YEPD, harvested at OD_{600} 0.6 after a 2 min cycloheximide treatment and lysed by vortexing with glass beads. For global ribosome footprinting, 50 A_{260} units of clarified lysate was digested with RNase I and separated through a 35 ml 6-38% sucrose gradient. 80S monosome fractions were collected and RNA extracted. For monosome or polysome footprinting, clarified lysate was separated through a 35 ml 6-38% gradient. Fractions corresponding to either monosomes or polysomes were collected separately; each was diluted in an equal volume of gradient buffer and concentrated. Post-concentration, 2 A_{260} units of each sample were digested with RNase I and separated through a 10.5 ml 10-50% sucrose gradient. 80S fractions were collected and the RNA extracted as for global footprints.

RNA-Seq and Library Preparation

5 µg of total RNA from clarified lysate was depleted of rRNA prior to fragmentation and size selection. RNA fragments (~20-45 nts) were isolated by denaturing-PAGE for RNA-Seq; ribosome footprints (27-31 nts) were isolated in a similar manner. All RNA fragments were converted into deep sequencing libraries using

a modified version of our standard laboratory protocol (Heyer et al., 2015). Briefly, a preadenylated adaptor is ligated to RNA 3' ends, after which the ligated RNAs are reverse transcribed. The RT product is gel purified, circularized, and PCR amplified prior to sequencing.

Mapping and Analysis of Deep Sequencing Data

Barcoded libraries were pooled and sequenced on either an Illumina HiSeq2000 (ribosome footprints) or MiSeq (RNA-Seq). Reads were parsed into appropriate libraries by 5' barcode, and then adaptor sequences removed. Trimmed reads were filtered for non-coding RNAs, and the remaining reads were mapped to both the sacCer3 genome (in a splice-aware fashion) and transcriptome, with the former being viewed on the UCSC genome browser. Uniquely mapping reads $\geq$25 nts (ribosome footprints) or $\geq$22 nts (RNA-Seq) were used for all analyses unless otherwise indicated. Data analyses were performed using the R software package.

Extended Experimental Procedures

Yeast Lysate Preparation

BY4741 yeast were grown under standard conditions in YEPD until OD_{600} ~0.6. Cycloheximide was added to a final concentration of 100 µg/ml, followed by 2 min of additional shaking at 30°C. Cells were collected via vacuum filtration over a 0.45 µm nitrocellulose filter. Collected cells were immediately scraped off the filter and resuspended in 4 ml ice-cold lysis buffer (20 mM Tris-Cl pH 8.0, 140 mM KCl, 5 mM $MgCl_2$, supplemented with 1 mg/ml heparin, 100 µg/ml cycloheximide, 500 µM DTT, 1X protease inhibitor cocktail set V, EDTA-free (Calbiochem #539137) and 1 mM PMSF). This resuspension was split equally between 2 chilled 15 ml corex tubes each containing 1.2 ml of 425-600 µm glass beads. Cells were lysed by vortexing on high 8x for 15 sec each, with 30-120 sec pause on ice between each cycle. Lysates were clarified by two rounds of centrifugation: one at 13,000 x g for 10 min at 4°C, followed by another in eppendorf tubes at 20,000 x g for 10 min at 4°C. Clarified lysates were then recombined into a single tube.

Monosome and Polysome Ribosome Footprinting

Sucrose gradient solutions were prepared w/v in gradient buffer (20 mM Tris-Cl pH 8.0, 140 mM KCl, 5 mM $MgCl_2$, supplemented with 100 µg/µl cycloheximide and 500 µM DTT). Gradients were poured using a Gradient Master (Biocomp). 2 ml of clarified lysate was loaded onto a 35 ml 6-38% gradient and spun for 3 hr 30 min at 27,000 rpm at 4°C. Gradient fractions were collected using a Density Gradient Fractionation System (Brandel #BR-188). Fractions corresponding to either the monosome peak or polysome peaks were pooled, resulting in ~3 ml of monosomes and ~15 ml of polysomes. To dilute the sucrose, an equal volume of gradient buffer was added to each pool. Samples were then concentrated on

Amicon-Ultra 100K columns (Millipore #UFC910024 and #UFC810024) by spinning at 5,000xg for either 10 mins (monosome fractions) or 20 mins (polysome fractions). 2 A_{260} units of concentrated monosome or polysome fractions (270-370 µl total) were digested with 60 µl RNase I (3,000 U RNase I per A_{260} unit RNA; Ambion #AM2294) at room temperature for 30 min with gentle shaking at 400 rpm. Digested fractions were loaded onto a second gradient (10.5 ml 10-50% w/v; prepared as above) and centrifuged at 35,000 rpm for 3 hr 12 min at 4°C. Gradient fractions were collected as above, and the monosome fractions were pooled.

Global Ribosome Footprinting

50 A_{260} units of clarified lysate were digested with 7.5 µl RNase I, rotating at RT for 1 hr. Digested lysate was loaded on a 35 ml 6-38% gradient and spun for 3 hr 30 min at 27,000 rpm at 4°C. Fractions corresponding to the 80S monosome peak were collected and pooled as above.

Ribosome Footprint Isolation

To isolate ribosome footprints, 800 µl of pooled RNase I-digested monosome fraction was first mixed with 1200 µl 8M guanidine-HCl. Following addition of 600 µl isopropanol, samples were incubated overnight at -20°C. Following centrifugation, precipitated RNA was washed 1X with 75% ethanol and resuspended in 400 µl H_2O. RNA was further purified by phenol-chloroform extraction, followed by ethanol precipitation and resuspension in 100 µl H_2O. To remove any remaining heparin, 100 µl of 5M LiCl; 33 mM EDTA was added to each sample before incubating overnight at -20°C overnight. RNA was collected by centrifugation, washed 2X with 75% ethanol, and dried briefly at 37°C before resuspension in H_2O (monosome and polysome samples, 20 µl; global samples, 100 µl).

Size selection of 26-32 nt RNA fragments was carried out by electrophoresis through a 15% Urea-PAGE gel (prepared using AccuGel reagents; National Diagnostics). RNA was eluted from gel fragments in RNA Elution Buffer (300 mM sodium acetate pH 5.2, 1 mM EDTA) plus a small amount (<100 µl) of phenol pH 4.5 to prevent any spurious RNase-catalyzed degradation. After an overnight incubation with constant rotation at RT, the eluate was phenol-chloroform extracted and ethanol precipitated. The RNA pellet was resuspended in 10 µl H20, briefly heat denatured, and dephosphorylated in MES buffer (100 mM MES-NaOH pH 5.5, 600 mM NaCl, 10 mM $MgCl_2$, 20 mM β-mercaptoethanol, 12.5 U T4 PNK (NEB #M0201)) at 37°C for 3 hrs. Dephosphorylated RNA was ethanol precipitated and resuspended in 8.25 µl H_2O prior to library construction.

mRNA Isolation for RNA-Seq

Total RNA was phenol extracted from the remaining undigested yeast lysate, ethanol precipitated, and resuspended in 200 µl H_2O. To remove any remaining heparin, the sample was reprecipitated by adding 100 µl of 5M LiCl; 33 mM EDTA to each sample before incubating overnight at -20°C overnight. RNA was collected by centrifugation, washed 2X with 75% ethanol, and dried briefly at 37°C before resuspension in 200 µl H_2O.

Ribosomal RNAs were depleted from 5 µg total RNA using the Ribo-Zero Magnetic Gold Yeast kit (Epicentre #MRZY13). Remaining RNAs were fragmented using RNA Fragmentation Reagent (Ambion #AM8740). RNAs were size selected for 20-45 nts and dephosphorylated as above.

Library Construction

Post-dephosphorylation, RNA fragments were prepared into deep sequencing libraries as described in (Heyer et al., 2015) with the following modifications. 3' adaptor ligation was carried out at 16°C overnight in a 10 µl reaction with 15% PEG8000. Ligated RNAs were reverse transcribed with 20 pmol of RT primer in a 30 µl reaction. Gel-purified RT product was circularized without betaine and PCR-amplified for 8 cycles (RNA-Seq), 11 (global footprinting), or 12 cycles (monosome or polysome footprinting).

Library Sequencing and Genome Alignment

Footprinting libraries were sequenced on an Illumina HiSeq2000 using a single-end, 50 bp run. RNA-Seq libraries were sequenced on an Illumina MiSeq using similar run parameters. Data were parsed by 5' barcode (Table 2.1) using cutadapt v1.7.1 (Martin, 2011) and an error rate of 0.2. The 3' adaptor sequence (5'-TGGAATTCTCGGGTGCCAAGG-3') was removed using the same method.

Footprinting reads were filtered for non-coding RNAs by mapping to a file containing sacCer3 sequences for rRNAs, tRNAs, snRNAs and snoRNAs using Bowtie 1 (Langmead et al., 2009) with parameters "-v 3 -k 1". After this filtering, remaining reads from each library were mapped to the sacCer3 genome using TopHat (Trapnell et al., 2009) with arguments "-N 2 -g 20 – segment-length 12 –coverage-search –library-type fr-secondstrand –bowtie1 –max-intron- length 1500 –max-segment-intron 1500 –max-coverage-intron 1500" and providing a GTF file to define known ORF boundaries. Only reads $\geq$25 nts were retained for further analysis; size selection was performed with NGSUtils (Breese and Liu, 2013). In addition, only uniquely mapping reads were used, filtered for a mapping score $\geq$10 using SAMtools (Li et al., 2009).

RNA-Seq reads were mapped as above, with a few modifications. After removing the 3' adaptor sequence, the terminal (3' end of the read) 3 nts were removed from all reads where a 3' adaptor sequence was not trimmed. Libraries were filtered for ncRNAs as above, but a significant number of tRNA mapping reads remained in the library. These reads were removed by mapping to a tRNA-only .fasta file with an additional CCA added to the end of every tRNA sequence. After genome-mapping, reads were limited to those $\geq$ 22 nts.

Transcriptome Alignment

A sacCer3 "transcriptome" was downloaded from the UCSC Table Browser (Karolchik et al., 2004), adding 20 nts upstream and 48 nts downstream of annotated start and stop codons, respectively, for 5' and 3' UTRs, and eliminating intron sequences. All reads post ncRNA removal were mapped to this transcriptome using the TopHat parameters "-N 2 -g 20 –no-coverage-search –library-type fr-secondstrand –bowtie1". Uniquely mapping reads were selected by eliminating those with a flag of 16 or 272, and then selecting those with a mapping score $\geq$ 25. Only reads $\geq$ 25 nts were retained for further analysis.

Genome Counts and Monosome:Polysome Score

Counts per gene were calculated from genome-mapping reads using HTSeq 0.6.1 (Anders et al., 2014) with parameters "–stranded=yes –type=exon –idattr=gene_id –mode=union". Resulting monosome and polysome counts were fed into DESeq2 (Love et al., 2014) for quantification of enrichment in either library. The assigned monosome:polysome score was the $\log_2$ fold change calculated by DESeq2.

tRNA Counts

Monosome and polysome tRNA mapping reads were isolated from files containing all previously removed noncoding RNA reads. The number of reads corresponding to each amino acid were counted.

Defining Cytoplasmic mRNAs

To limit our analysis to mRNAs translated in the cytoplasm, we removed genes for which there is evidence of translation on a membrane (either ER or mitochondrial). Lists of signal-sequence containing genes predicted by SignalP and transmembrane-domain containing genes predicted by TMHMM were downloaded from the Saccharomyces Genome Database (SGD; http://www.yeastgenome.org/), and these genes were excluded from further analysis. In addition, transcripts recently found to be translated on the mitochondria (no cycloheximide gene set from Williams et al., 2014) were also removed. Though not specifically removed from analysis (because they were already identified by SignalP or TMHMM), the secretome annotation came from Jan et al., 2014.

Determining Length Cutoff

To limit our analysis to mRNAs where a factor other than ORF length determined monosome association, we analyzed ORFs longer than 590 nts. To determine the 590 nt ORF length cutoff, we ordered all cytoplasmic mRNAs by ORF length, separated them into bins of 50 genes each, and calculated the mean and median monosome:polysome score of each bin. Above 590 nts, the difference in mean and median from bin to bin was much larger than the <590 bins, and the Pearson correlation coefficient between mean monosome:polysome score and ORF length was less than 0.95.

Outside Datasets

Other than those mentioned above, several published datasets were included in this analysis. Data were downloaded from supplemental material and merged by gene name with data generated in this paper. Transcription rate came from Table S1 in Pelechano et al., 2010. Data on mRNA half-lives were downloaded from Table S1 in Presnyak et al., 2015, using the total half life value. Protein molecules per cell were gathered from supplemental material from Ghaemmaghami et al., 2003. The average number of ribosomes associated with each mRNA species was downloaded from the Data Summary for 2,128 high confidence genes from Arava et al., 2003. Frequency of optimal codons came from SGD (http://downloads.yeastgenome.org/curation/calculated_protein_info/protein_properties.tab). Lists of NMD-regulated genes were taken from He et al., 2003 and Table S2 of Guan et al., 2006. Translation initiation and elongation times, as well as the calculated ribosome density from the same model, came from Table S1 in Siwiak and Zielenkiewicz, 2010.

Metagene Plots

Metagene plots were created using the ngsplot package (Shen et al., 2014) with genome-mapping reads and parameters "-R genebody –FL 30 –SE 0 –L 100

–SS same". The flanking regions were then trimmed to represent only 50 nts.

Aggregation Plots

Transcriptome-mapping reads were filtered with NGSUtils to only include 28 nt reads. The alignment file for these reads was converted to a BED file using BEDTools (Quinlan and Hall, 2010), and the 5' end position of each read determined. To calculate the distance to the start codon, 20 nts was subtracted from the 5' end position of each read to account for the 20 nts added onto the annotation of each ORF (see transcriptome alignment above). Read distance from the 3' end was calculated by subtracting the length of the ORF from the 5' end position.

To create 5' end or 3' end aggregation plots, all 28 nt reads for each distance were summed across all genes. For the RPM aggregation plots, the number of reads at each position was normalized to the total number of 28 nt reads in each library. For aggregation plots normalized such that each gene contributed equally, the number of reads per transcript at each position was divided by the total number of reads from that transcript. These fractions were then summed across all ORFs.

The decrease over codons 9-36 in aggregation plots were calculated by averaging the signal at positions corresponding to either codons 9 and 10 (nt positions 12 and 15 in transcriptome mapping) or codons 35 and 36 (nt positions 90 and 93).

Gene Ontology Analysis

GeneCodis (http://genecodis.cnb.csic.es/) gene ontology tool was used to identify enriched genes in the monosome and polysome gene sets (Carmona-Saez et al., 2007, Nogales-Cadenas et al., 2009, Tabas-Madrid et al., 2012). All cytoplasmic ORFs > 590 nts were used as the background population.

sORF and uORF Counts

sORF annotations were taken from Table S3 in Smith et al., 2014 and converted into a BED file. uORF annotations were downloaded from SGD (http://downloads.yeastgenome.org/published_datasets/Ingolia_2009_PMID_19213877/track_files/Ingolia_2009_canonical_uORFs_V64.bed and http://downloads.yeastgenome.org/published_datasets/Ingolia_2009_PMID_19213877/track_files/Ingolia_2009_noncanonical_translated_uORFs_V64.bed) and combined. In a strand-specific manner, these coordinates were adjusted 12 nts upstream to allow capture of read 5' ends. To count the number of reads mapping to these regions, the number of mapped read 5' ends (from genome mapping above) falling within each region was counted using BEDTools intersectBed with the "-s" option.

Intron Analysis

A BED file containing the genomic coordinates of sacCer3 annotated introns was downloaded from the UCSC Table Browser. Uniquely mapping reads (either $\geq$25 nts or only 28 nts) overlapping introns by $\geq$ 1 nt were identified from the genome mapping reads (see Library sequencing and genome alignment above) by a 2-step process. First, all reads overlapping each intron were isolated using BEDTools intersectBed with the "-s -wo -bed" options. Second, these intron-overlapping reads were limited to BED entries with a single block count (i.e. exon-junction mapping reads spanning the intron are 2 block entries, and were discarded). Genome-browser tracks were created as discussed below.

Intron reading frame was assigned as the remainder when the distance between the ORF start and intron start (i.e., 5' splice site) was divided by 3. Thus, ’intron reading frame’ was assigned based on the reading frame of the first intronic nucleotide relative to the ORF start codon in the upstream exon. To combine all ’intron reading frames’ into a single ’read reading frame’ term, the 5' position of each read was adjusted to the reading frame of the intron it overlapped.

Graphics

Genome browser tracks were generated from genome mapping reads using SAMtools and BEDTools, normalizing for library size and splitting the reads by strand. Images are screenshots from the UCSC Genome Browser (Kent et al., 2002). All plots were generated in RStudio (R Core Team, 2015) using ggplot2 (Wickham, 2009) and cowplot (Wilke, 2015) packages.

Chapter III

An optimized kit-free method for making strand-specific deep sequencing libraries from RNA fragments

Preface

The contents of this Chapter have been published previously as:

> *Heyer, E. E., Özadam, H., Ricci, E. P., Cenik, C., and Moore, M. J. (2015). An optimized kit-free method for making strand-specific deep sequencing libraries from RNA fragments.* Nucleic Acids Research*, 43(1):e2*

For information not contained in this chapter (i.e. supplemental tables), please refer to the following locations:

NCBI: http://www.ncbi.nlm.nih.gov/pubmed/25505164

Nucleic Acids Research: http://nar.oxfordjournals.org/content/43/1/e2/suppl/DC1

Additional protocol information can also be found on the Moore Lab website: http://umassmed.edu/moorelab/resources/protocols/

Introduction

In cells, all RNA molecules interact with RNA binding proteins (RBPs) to form ribonucleoprotein particles (RNPs). An ever-increasing number of methodologies employ deep sequencing to map these protein-RNA interaction sites transcriptome-wide. Such techniques include ultraviolet crosslinking methods (e.g. CLIP, PAR-CLIP; Hafner et al., 2010, Ule et al., 2003) to map the ribonucleotides directly in contact with an individual RBP and RNP footprinting (e.g. Ribo-Seq, RIPiT-Seq; Ingolia et al., 2009, Singh et al., 2012) to map the occupancy sites of larger complexes. Many projects in our laboratory are focused on transcriptome-wide RNP footprint analysis (Chen et al., 2014, Ricci et al., 2014, Singh et al., 2014). Depending on the complex being examined and the RNA fragmentation method utilized (e.g. RNase or sonication), bound RNA fragments can range from 10 to 200 nucleotides (nts). Therefore, we require a strand-specific library generation method that works for diverse RNA lengths, faithfully preserves their relative abundances in the original sample and excludes any contaminating DNA fragments.

Multiple commercial kits currently exist for strand-specific library preparation, but most are intended to capture either long RNAs (e.g. RNA-Seq) or short RNAs (e.g. miRNA-Seq), but not both. Further, commercial kits are regularly updated with new preparation methods. Because preparation method is the primary source of variability between deep sequencing libraries (Linsen et al., 2009), quantitative comparisons are best done between identically generated libraries (i.e. with a single commercial kit version). However, the expense of commercial kits (and remaking libraries as new kits appear and older versions are phased out) is cost prohibitive for many academic laboratories. We therefore set out to develop an optimized, strand-specific RNA library preparation protocol that utilizes commonly available reagents and works over a wide range of input amounts. We also wanted an approach that can be used to capture full-length RNP footprints as well as map sites of reverse transcriptase stalling (e.g. sites of RNA-protein crosslinking from CLIP experiments or abasic/alkylated sites).

All current library preparation methods utilize enzymes to capture nucleic acid fragments by appending 5' and 3' adaptor sequences. Enzymes have inherent substrate preferences that are most significant at low substrate concentrations (k_{cat}/K_m conditions) and at short reaction times (Fersht, 1985). For ligation reactions, low temperatures can favor capture of sequences capable of base pairing with the adaptor (Sorefan et al., 2012). Low temperatures can also disfavor capture of sequences containing internal secondary structures. Many published library preparation protocols are suboptimal for one or more of these factors, resulting in differential capture of small RNAs (e.g. miRNA-Seq; Bissels et al., 2009, Hafner et al., 2011, Sorefan et al., 2012) and highly non-uniform ("peaky") coverage of long RNAs (e.g. RNA-Seq of RNA Pol II transcripts; Levin et al., 2010). For these reasons, we decided to re-examine 5' and 3' end capture conditions, with the goal of driving every reaction to completion.

Here, we present the detailed protocol for strand-specific RNA library preparation currently in use in our laboratory, as well as the titration and time course data we used to optimize each step. Also presented are deep sequencing data on (i) the effects of time and temperature on initial 3' end capture and (ii) capture uniformity analysis for an equimolar pool of 29 miRNAs. Taken together, these data show that our method faithfully preserves fragment diversity and abundance in complex starting mixtures and is minimally affected by fragment sequence or folding potential.

Results

Protocol Design

To generate strand-specific deep sequencing libraries, both ends of the captured RNA must be appended to fixed sequences (adaptors) to enable primer hybridization for amplification and sequencing. These adaptors generally correspond to the forward and reverse primer sequences used for clonal cluster amplification on the desired sequencing platform. All strand-specific RNA-Seq and small RNA library preparations published to date capture the 3' end in one of the following ways: (i) RT of full length or fragmented RNAs with oligo-dT and/or random hexamers, or a longer DNA primer containing a 3' randomized region (Armour et al., 2009, Cloonan et al., 2008, Kwok et al., 2013, Langevin et al., 2013, Zhang et al., 2012); (ii) polyA tailing of RNA fragments followed by RT with an anchored oligo-dT 3' end sequence (Ingolia et al., 2009, Linsen et al., 2009); or (iii) direct 3' end adaptor ligation (Elbashir et al., 2001, Lau et al., 2001, Pan and Uhlenbeck, 1992). Disadvantages of random hexamer RT include the introduction of mutations at the point of primer hybridization plus capture biases resulting from differential hybridization efficiencies on different sequences (Hansen et al., 2010). Random hexamer RT is also not an option for small RNAs. In our hands, polyA tailing of fragmented RNA samples proved inconsistent (data not shown). Therefore, we decided to adopt a 3' end adaptor ligation approach widely used in the small RNA field (Lau et al., 2001) - direct ligation of a preadenylated DNA adaptor to the 3' end of RNA fragments using RNA ligase (Figure 3.1, Step 1). We chose to use a truncated and mutant form of T4 RNA Ligase 2 (RNL2 Tr. K227Q) because published reports indicated it has less substrate bias and produces fewer side products than the full-length wild-type enzyme (Bissels et al., 2009, Viollet et al., 2011), and RNL2 is known to be less affected by nt identity at the ligation site than T4 RNA Ligase 1 (Zhuang et al., 2012). Following 3' adaptor ligation, a highly efficient method for appending the 5' adaptor is to reverse transcribe the RNA from the 3' adaptor with an RT primer containing the 5' adaptor sequence at the other end and then circularize the resulting single-stranded cDNA using CircLigase (Ingolia et al., 2009) (Figure 3.1, Steps 2 and 4). A long flexible linker (Spacer 18, an 18-atom hexa-ethyleneglycol spacer) is placed between the fixed adaptor sequences to minimize structural constraints for circularization and preclude the possibility of rolling circle PCR (Ingolia, 2010).

A common strategy for reducing deep sequencing costs is to "barcode" individual libraries so that they can be mixed together and sequenced in a single lane. Barcodes consist of 2–10 unique nts appended either 5' or 3' to the captured sequences (Parameswaran et al., 2007), and ideally differ by more than 2 nts so as to minimize incorrect library identification due to sequencing errors. Barcodes can be placed in one of the adaptors (Alon et al., 2011, Hafner et al., 2012) or in the reverse PCR primer (Alon et al., 2011), or they can be ligated to the double-stranded library post-PCR amplification (Van Nieuwerburgh et al., 2011). Barcode incorporation immediately downstream of the forward sequencing primer hybridization site allows both the barcode and the adjacent captured fragment to be decoded in one single-end sequencing reaction. In theory, barcodes can be appended to either end of the captured fragment. However, RNL2 ligation efficiency is significantly affected by the 3' adaptor sequence - therefore, placement of the barcode at the 5' end of the 3' adaptor can result in significant and different sequence biases dependent on the barcode (Hafner et al., 2011, Jayaprakash et al., 2011). Because we were able to find conditions under which cDNA circularization is quantitative (see below), we chose to place our barcodes at the 3' end of the 5' adaptor (i.e. between the forward primer sequence and the captured sequences). Nonetheless, to minimize any confounding effects of varying the nt composition at the site of circularization, we introduced two guanine residues at the 5' end of each RT primer so that the nts interacting with CircLigase would be the same regardless of barcode.

A final consideration for making strand-specific cDNA libraries is the quantity of starting material required. Major factors leading to material loss during library preparation are the number of gel purification steps and the number of different surfaces (i.e. tips and tubes) with which the sample comes in contact. Thus, we

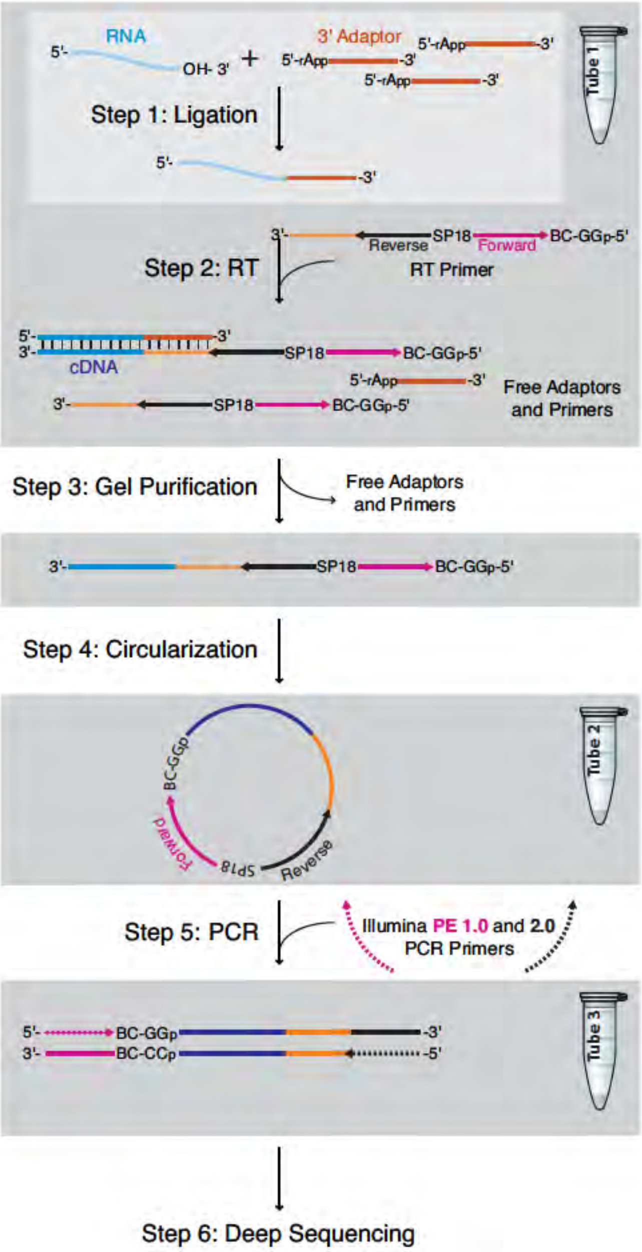


FIGURE 3.1: Method Overview

Step 1: Ligation. RNA, shown in blue, is ligated to a preadenylated DNA adaptor to form a RNA:DNA hybrid. In the same tube, RT is performed (Step 2). The RT primer contains both the reverse and forward priming sequences for Illumina sequencing, as well as a barcode to uniquely identify the sample. Step 3: The RT product is gel purified, removing unligated adaptors and unextended RT primers from the sample. Step 4: The gel purified RT product is circularized, forming a template for PCR (Step 5). The PCR product is then purified and used for deep sequencing (Step 6).

opted for a protocol wherein the ligation (Step 1) and RT (Step 2) were carried out in a single tube without any cleanup or buffer exchange in between, and the sample is only subjected to a single gel purification (Step 3) after RT.

Protocol Optimization

For optimization of each step, we used a pool of randomized RNA 24mers (N24) to mimic the diversity of sequences in a biological sample. Ligation reactions were visualized using 5' end ^{32}P-labeled RNAs. RT products were visualized by including α-^{32}P-dCTP in the RT reaction. Circularization reactions were visualized using either body-labeled or 5' end-labeled RT products.

Step 1 - Preadenylated 3' Adaptor Ligation

When we initiated this project, the manufacturer's (NEB) suggested conditions for RNL2 Tr. K227Q ligation reactions were 500 nM single-stranded RNA, 1 µM 3' adaptor, 10 U/µl enzyme and 15% w/v PEG8000 in 1x reaction buffer at 16°C overnight. As our goal was to create a robust protocol that could be successfully employed over a wide range of RNA input concentrations, we set out to explore the limits of these parameters (Figure 3.2). For all experiments below, we pre-mixed the RNA and 3'-adaptor in water and incubated this mixture at 65°C for 10 min prior to enzyme addition.

Ligation efficiency depends on successful collision of multiple components. Such collisions can be increased by molecular crowding agents (e.g. PEG) and/or dehydrating co-solutes (e.g. dimethyl sulfoxide; DMSO), and published 3' adaptor ligation protocols vary with regard to PEG8000 and DMSO inclusion (Eminaga et al., 2013, Mamanova and Turner, 2011, Munafo and Robb, 2010, Pfeffer et al., 2005, Vivancos et al., 2010). Consistent with a recent report that 25% PEG8000 enhances ligation efficiency (see Figure 4B in Munafo and Robb, 2010), we found that 25% PEG8000 resulted in near complete N24 ligation at 16°C O/N (Figure 3.2A). However, increasing DMSO had no effect, regardless of PEG8000 absence or presence (Figure 3.2B). Thus, all subsequent ligation reactions included 25% PEG8000 but no DMSO.

We next titrated preadenylated 3' adaptor, N24 and enzyme concentrations. Using two different N24 concentrations, near complete ligation was observed at all adaptor concentrations above 130 nM (Figure 3.2C). At 470 nM adaptor, ligation was highly efficient with N24 concentrations above 50 nM (Figure 3.2D) and enzyme concentrations above 6 U/µl (Figure 3.2E). A greater dependence of ligation efficiency on enzyme concentration at 10 nM N24 does suggest, however, that additional enzyme will increase yields for very dilute RNA samples (Sterling et al., 2015).

Published reports using T4 RNA ligases for library preparation employ a wide

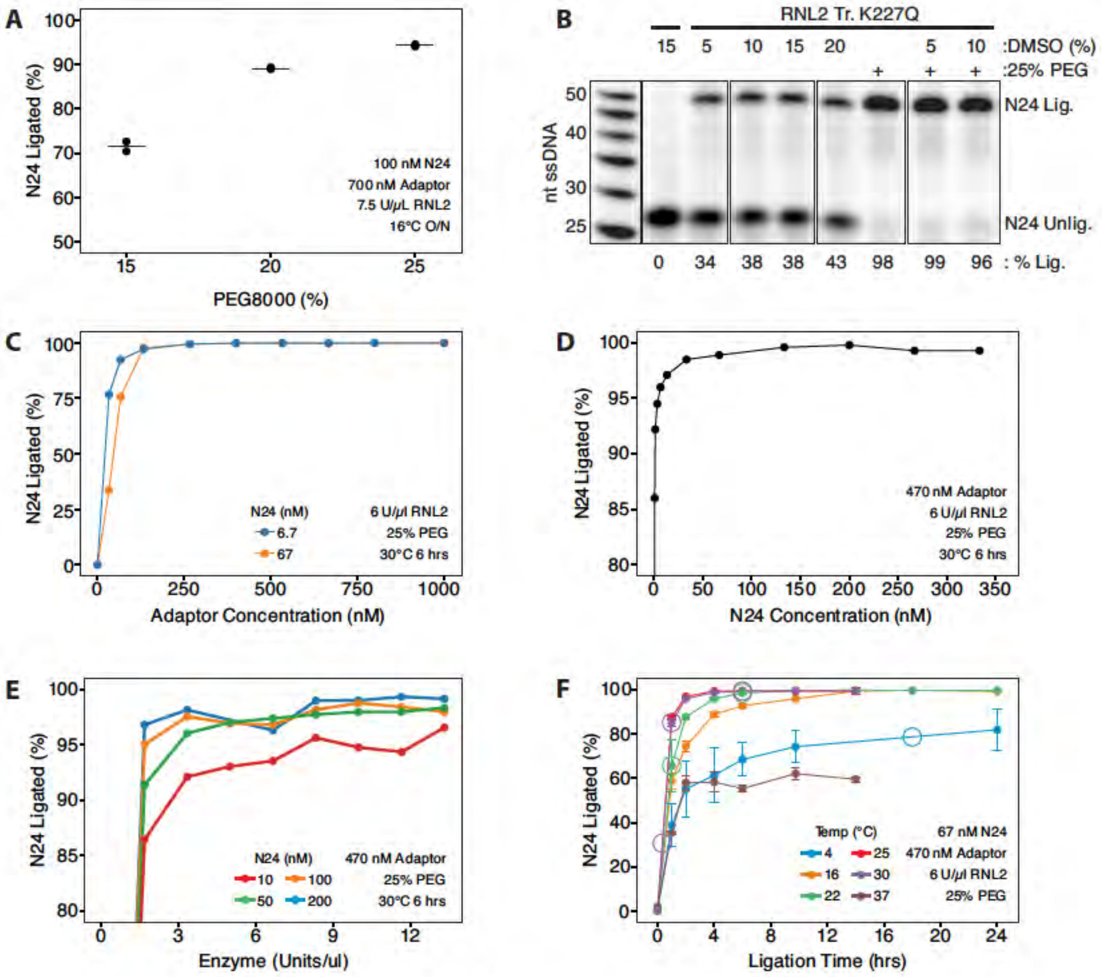


FIGURE 3.2: 3' Adaptor Ligation Optimization

(A) Ligation efficiency versus % PEG8000 (w/v) (n = 2; black line, mean). (B) Comparison of DMSO and PEG as ligation enhancers. Absence or presence of indicated species are indicated by – and +; ligation efficiencies are indicated below each lane. N24 RNA was 5' end labeled with ^{32}P-γ-ATP. (C) Ligation efficiency versus 3' adaptor concentration (n = 1). (D) Ligation efficiency versus N24 concentration (n = 1). (E) Ligation efficiency versus RNL2 concentration at four different N24 RNA concentrations (n = 1). (F) Ligation efficiency versus time and temperature (n = 3; error bars, standard deviation). Circles indicate ligation conditions for N24 libraries. In all panels, data were generated by quantification of denaturing polyacrylamide gels similar to that shown in panel B; ligation efficiency = (ligated RNA:DNA product)/(unligated RNA + ligated RNA:DNA product) in each lane.

range of reaction times (1 h to overnight) and temperatures (5°C–37°C) (Ingolia et al., 2013, Lau et al., 2001, Lee and Ambros, 2001, Lui et al., 2007, Morin et al., 2008, Pfeffer et al., 2005, Ule et al., 2003, Vivancos et al., 2010). However, colder temperatures should stabilize both intra- and inter-molecular secondary structures, potentially biasing ligations against internally structured RNAs and toward RNA sequences that partially base pair with the 3' adaptor (Hafner et al., 2011, Sorefan et al., 2012, Zhuang et al., 2012). Higher temperatures should alleviate these issues, but could decrease enzyme stability and increase RNA degradation. Using our N24 pool, we assessed ligation efficiencies across a range of incubation times and temperatures (Figure 3.2F). Both 4°C and 37°C yielded poor ligation efficiencies at all incubation times. Using radioactively labeled RNA, we determined that the lower yields at 37°C were not due to increased RNA degradation (data not shown); rather, the plateau reached after 2 h suggests that enzyme is unstable at 37°C. All reactions incubated between 16°C and 30°C ultimately resulted in near complete ligation. However, the 16°C and 22°C reactions took longer to reach completion (10–14 h) than did the 25°C and 30°C reactions (4–6 h).

Based on all of the above data, we adopted the following as our standard ligation reaction conditions: 470 nM adaptor, 50–330 nM RNA, ≥6 U/µl RNL2 K227Q, 1X RNL2 reaction buffer (from NEB: 50 mM Tris-HCl, pH 7.5 at 25°C, 10 mM MgCl2, 1 mM DTT) plus an additional 1 mM DTT to ensure a reducing environment, incubated for 6 h at 30°C and then 20 min at 65°C (to heat inactivate the enzyme). These conditions yield efficient ligation over the wide range of RNA fragment lengths (Supplemental Figure 3.3) we generally obtain when footprinting endogenous RNP complexes (Ricci et al., 2014, Singh et al., 2012, 2014).

Step 2 - Reverse Transcription

A number of high fidelity reverse transcriptases are commercially available. For our purposes, we wanted an enzyme that produced a high yield of full-length product with minimal side products when added directly to the heat-inactivated/diluted 3' adaptor ligation reaction from Step 1. We tested Accuscript (Agilent), AMV RT (Finnzymes), Superscript III (Invitrogen) and Transcriptor (Roche) (Figure 3.4A). In all cases, ligation reactions were diluted and supplemented with either (i) the appropriate amount of manufacturer-supplied 5X or 10X RT buffer or (ii) the same buffer minus $MgCl_2$ (as the Step 1 reaction already contains $MgCl_2$, and concentrations of $MgCl_2$ above 3 mM can inhibit RT (Gerard et al., 1997)). For all four enzymes (tested at the manufacturer's recommended concentration), we observed more full-length RT product when no Mg^{2+} was added beyond that supplied by the diluted ligation reaction. As SuperScript III gave the highest RT product yield, we chose it for subsequent optimization. By varying the amount of the heat-inactivated

Step 1 reaction in the Step 2 reaction, we determined that maximal RT product yield was obtained when the ligation reaction constituted one-third of the final volume of the RT reaction (data not shown). This resulted in a final $MgCl_2$ concentration of 3.3 mM. At this 3-fold dilution, we found no inhibitory effect on RT by the PEG8000 present in the Step 1 reaction; rather, Step 1 reactions containing 25% PEG8000 gave the highest Step 2 yields (Figure 3.4B).

We next varied RT primer, enzyme and RNA input amounts. To maximize RT product yield, it is important that the RT primer concentration be greater than the 3' adaptor concentration but not excessively so, as this would favor empty circle formation in the subsequent circularization reaction (Step 4). We observed no advantage for RT yield when the RT primer:3' adaptor ratio was significantly higher than 1.3:1 (Figure 3.4C). Further, all SuperScript III concentrations above 3 U/µl gave comparable product yields (Figure 3.4D). Varying the temperature (50°C, 55°C and 60°C) and time (30 min and 1 h) of the RT reactions revealed 55°C for 30 min to be optimal (data not shown). When the input RNA was varied between 3.3 and 133 nM, the yield of RT product increased linearly across this range (Figure 3.4E and F). Thus, like the ligation reaction, the RT reaction proved highly robust and amenable to library construction over a wide range of input amounts.

Based on the above data, we adopted the following as our standard Step 2

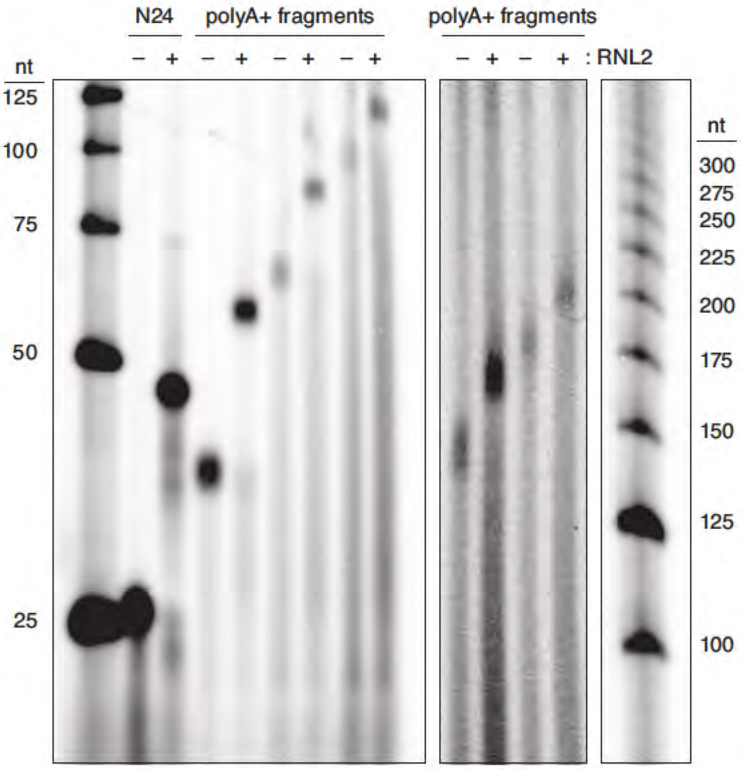


FIGURE 3.3: 3' Adaptor Ligation to Different Length RNAs

reaction conditions: 3-fold dilution of the heat-denatured ligation reaction from Step 1, supplemented with 333 nM RT primer, 5.33 U/µl SuperScript III (to ensure consistent results and allow for some variability in nucleic acid concentration determination and enzyme activity), 50 mM Tris-HCl (pH 8.3 at room temperature), 75 mM KCl and 5 mM DTT. This mixture is incubated at 55°C for 30 min followed by heat inactivation at 75°C for 15 min.

Step 3 - Gel Purification

See Experimental Procedures section.

Step 4 - Circularization

There are currently two commercially available enzymes for ssDNA circularization: CircLigase I and II (Epicentre). We tested both at 50 nM input ssDNA and found that CircLigase I gave much higher circularization efficiencies (98–99%) than CircLigase II (45–61%) (Figure 3.5A). Betaine, a compound commonly used in PCR reactions to eliminate the energy difference between A-T and G-C base pairs, is recommended by Epicentre for use with CircLigase II. However, as no amount of betaine improved CircLigase II efficiency to that obtained with CircLigase I, we decided to proceed with CircLigase I.

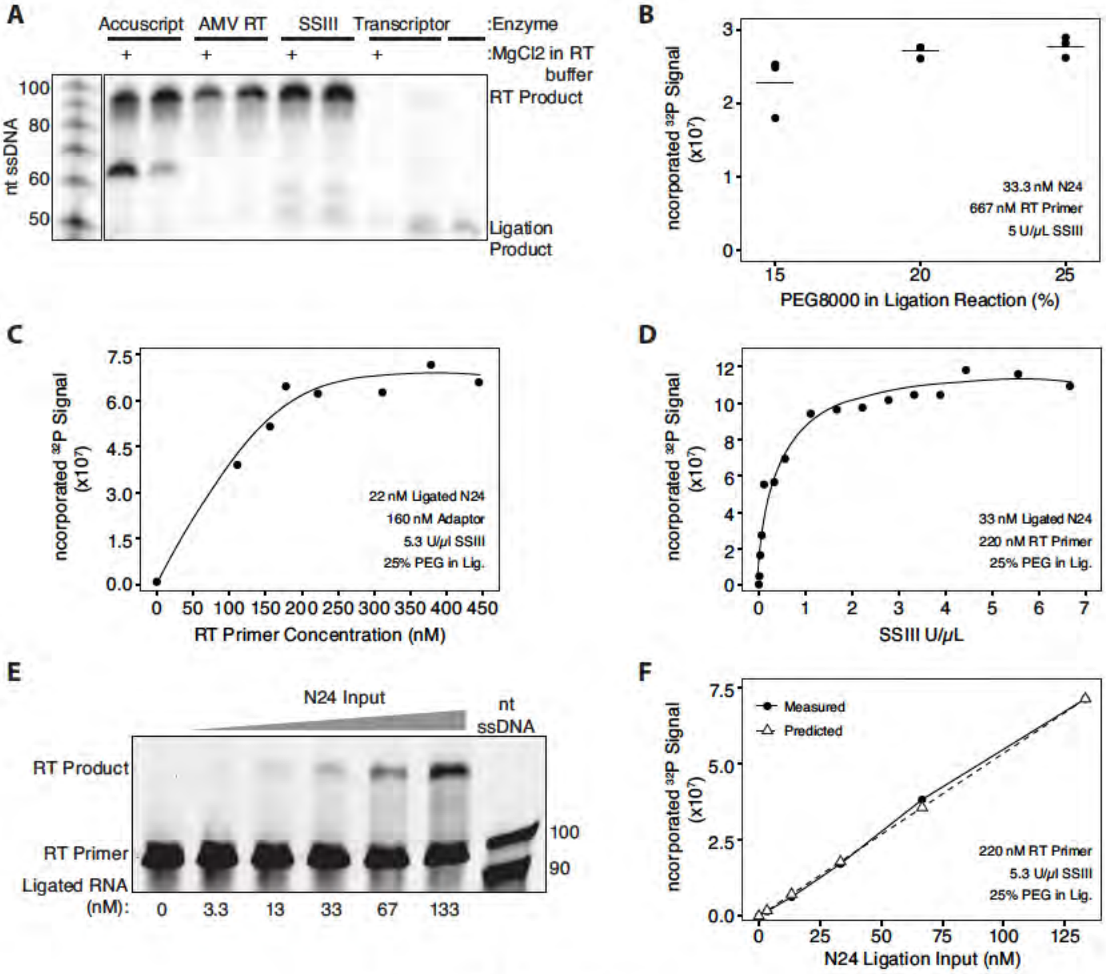


FIGURE 3.4: RT Optimization

(A) Comparison of high-fidelity reverse transcriptases for the amount of RT product generated ± $MgCl_2$ in the RT buffer. Absence or presence of indicated species are indicated by – and +. (B) RT product signal versus % PEG8000 (w/v) in the ligation reaction (n = 3; black line, mean). (C) RT product signal versus RT primer concentration (n = 1). (D) RT product signal versus SSIII concentration (n = 1). (E) RT product signal varies with RNA input concentration, ranging from 3.3 nM (lane 2) to 133 nM (lane 6). (F) RT product signal versus RNA input concentration (n = 1). Replicate of panel E, incorporating ^{32}P in the RT for quantification. In panels B-D and F, data were generated by quantification of denaturing polyacrylamide gels similar to panels A and E.

To explore the limits of CircLigase I performance, we tested a range of conditions. Changing the enzyme concentration and doubling or reducing by half the reaction volume had no significant effect on circularization efficiency (data not shown), so we continued to use the manufacturer's suggested conditions. A timecourse revealed that complete circularization with 5 U/µl enzyme and 50 nM input N24 RT product required at least 2 h at 60°C (Figure 3.5B). Titration of the N24 RT product indicated that ligation efficiencies dropped off precipitously below 25 nM ssDNA (Figure 3.5C). This dropoff was unaffected by either increasing or decreasing the enzyme concentration (data not shown), but was substantially rescued by the inclusion of 1 M betaine in the circularization reaction (Figure 3.5D). In this case, as circularization of <5 nM N24 RT product could not be detected by direct observation of the ^{32}P-labeled substrate and product on a gel, relative PCR product yields served as a proxy for circularization yields, with cycle number adjusted for RNA input amount. In order to exclude the possibility of betaine stimulating the yield of the PCR reaction instead of the circularization reaction, we added betaine subsequent to heat inactivation of CircLigase I; under these conditions, no betaine-dependent increase in PCR signal was observed (data not shown).

Based on the above data, we adopted the following as our standard Step 4 reaction conditions: 1X CircLigase buffer (Epicentre), 1 M betaine, 50 µM adenosine triphosphate, 2.5 mM MnCl2 and 5 U/µl CircLigase I in 20 µl containing all of the ssDNA isolated in Step 3. This mixture is incubated at 60°C for 3 h followed by heat inactivation at 80°C for 10 min.

Step 5 - PCR

To eliminate another gel purification step, we decided to use a portion of the completed and inactivated circularization reaction as direct input to PCR amplification. Adding 1.5 µl of a heat-inactivated circularization reaction containing ~88 nM input RT product directly to a 25 µl (final volume) PCR reaction, we tested the following high fidelity polymerases, each using their respective manufacturer's supplied buffer and recommended cycling conditions (i.e. times and temperatures) for 8 cycles: PfuUltraII (Stratagene), Herculase II (Stratagene), Phusion (Finnzymes), KAPA HiFi (Kapa Biosystems), Advantage HD (Clontech), PrimeSTAR Max (Clontech) and AccuPrime Pfx (Invitrogen). Addition of DMSO, a PCR enhancing agent, did not significantly increase PCR amplification with any enzyme, perhaps with the exception of PfuUltra II (Figure 3.6A and B). PfuUltraII, Herculase II, Phusion, PrimeSTAR Max and KAPA HiFi all gave comparable product yields, but KAPA HiFi generated the least amount of slower migrating side products (indicated by $*$) just above the desired product (Figure 3.6A and B). Because of this and an independent report demonstrating its robustness with regard to GC content (Quail et al., 2012), we decided to proceed with KAPA HiFi.

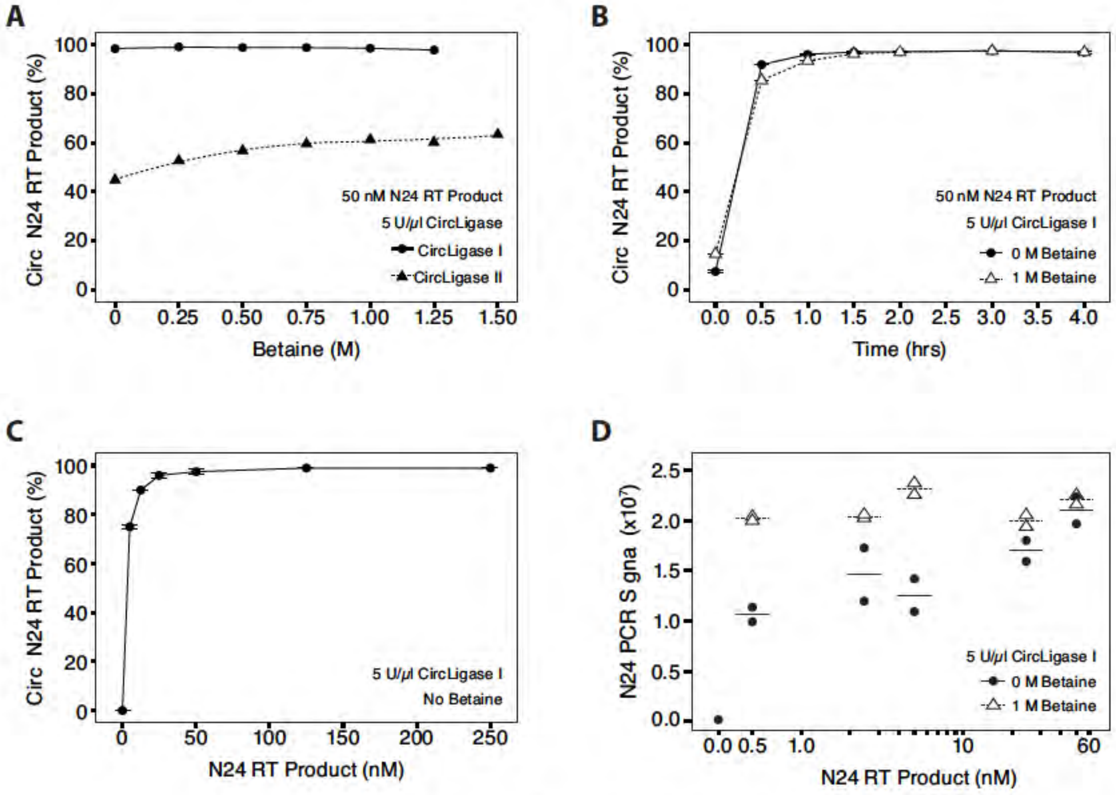


FIGURE 3.5: Circularization Optimization

(A)Circularization efficiency versus betaine concentration (n=1) for CircLigase I and II (n=1). (B) Circularization efficiency versus time and betaine concentration (n = 1). (C) Circularization efficiency versus N24 RT product concentration (n = 2). (D) N24 PCR signal versus N24 RT product concentration prior to circularization (n = 2; line, mean) at 0M and 1M betaine. In all panels, data were generated by quantification of polyacrylamide gels (denaturing, panels A–C; non-denaturing, panel D). Circularization efficiency = (circularized RT product)/(linear RT product + circularized RT product) in each lane. N24 PCR signal = intensity of N24 PCR product band.

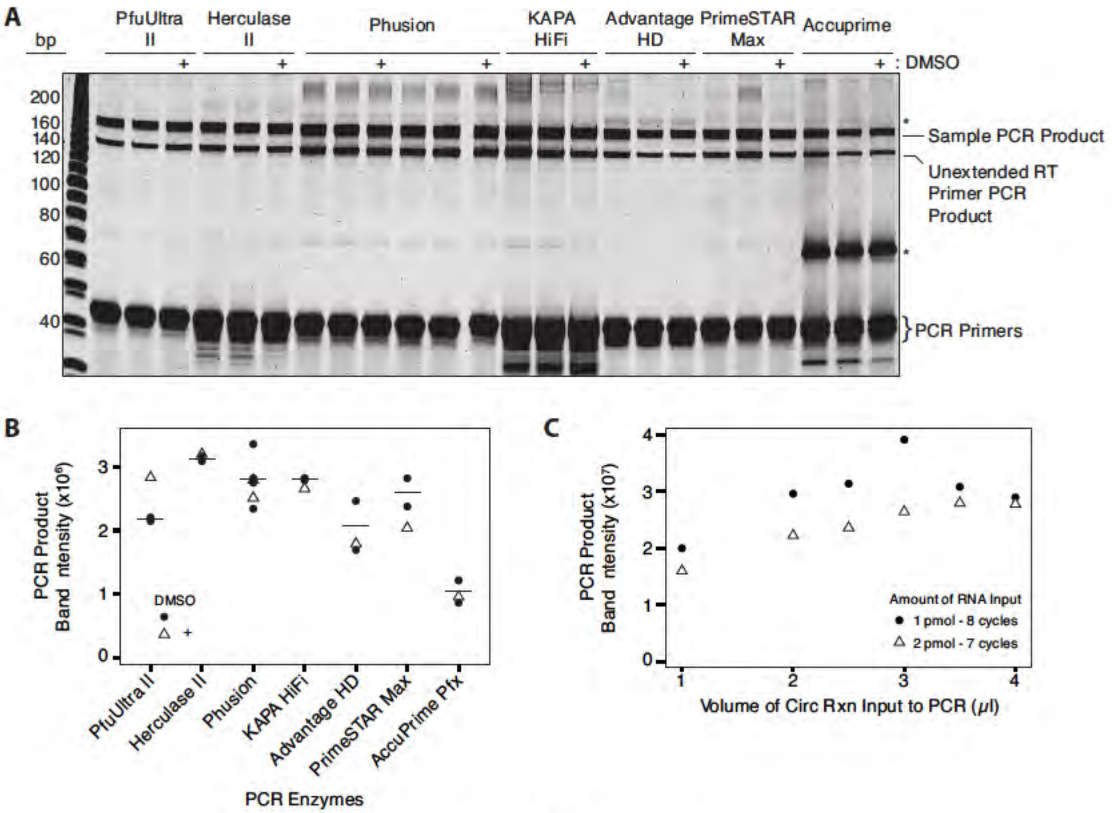


FIGURE 3.6: PCR Optimization

(A) Comparison of proofreading PCR enzymes for the amount of sample PCR product ± DMSO. $*$, PCR by-products. (B) PCR product signal versus enzyme; quantification of panel A; black line, mean. (C) PCR product signal versus circularization reaction input volume (n = 1) for 1 pmol and 2 pmol RNA starting material. In panels B and C, data were generated by quantifying sample PCR product band on non-denaturing polyacrylamide gels.

When preparing deep sequencing libraries, higher amounts of input DNA and low cycle numbers are desirable to amplify the greatest number of unique species. However, as with the RT reaction (Step 2), we were concerned that the diluted circularization buffer might affect PCR efficiency. Therefore, we titrated the volume of CircLigase reaction included in each PCR reaction. When this volume was varied from 0.5 to 3.5 µl in a 15 µl PCR reaction, the PCR band intensity increased with increasing input, but not to scale (i.e. a 2-fold increase in input from 1 to 2 µl produced only a 1.5-fold increase in output; Figure 3.6C), likely indicating some inhibitory effect of the CircLigase reaction on PCR efficiency. We therefore limit the amount of added CircLigase reaction to one-fifth of the total PCR reaction volume.

Consequences of Incomplete 3' Adaptor Ligation

Having optimized each step in the protocol (Supplementary Table 3.1), we next wanted to assess the quality of libraries it generates. Because many published protocols use lower 3' adaptor ligation temperatures and/or shorter incubation times than our optimized conditions (Figure 3.2F), we also wanted to test the effects of these variables. Therefore, we prepared seven different libraries using our synthetic N24 pool. All libraries were prepared identically except for the 3' adaptor ligation step, for which the conditions are shown in Figure 3.2F and Supplementary Figure 3.7A. In one library, we also included four randomized nts at the 5' end of the 3' adaptor (N4 adaptor) to assess whether this would reduce 3' end capture bias, as has been previously suggested (Jayaprakash et al., 2011, Sorefan et al., 2012, Zhang et al., 2013). To eliminate possible sequencing variability, all libraries were barcoded, mixed together and sequenced to similar depth within a single Illumina HiSeq 2000 lane (Supplementary Figure 3.7A). Also included in this lane was a library of random ~500 nt fragments generated from the PhiX174 genome (~15% of total sequences); PhiX inclusion increases the nt diversity at every position, thereby increasing the base calling accuracy (Illumina, 2013).

To address the concern that long incubation times at higher temperatures could lead to significant RNA hydrolysis, we first examined the lengths of the captured sequences (Figure 3.8A). In all libraries, the majority of captured sequences were 24 nts. As expected, however, incubation at 22°C or 30°C for 6 h did result in a small decrease (<7%) in the fraction of full-length species compared to the 20 min and 1 h incubation times (Figure 3.8A, inset I). Also as expected, this effect was somewhat less apparent at 4°C. Nonetheless, the impact of this material loss must be weighed against the higher capture variability introduced by shorter ligation times and lower temperatures (see below).

For further analysis we focused solely on full-length (24 nt) reads. Because the number of possible sequences in a 24-nt random oligo ($>10^{14}$) so vastly

TABLE 3.1: Quickguide to Method

	Initial Sample Mix	Heat Denature	Additional Materials	Incubation	Heat Inactivation
1. Ligation	1 μl 7 μM 3' Adaptor x μl RNA Water to 4.8 μl	65°C 10 min; 4°C hold	1.5 μl 10X RNL2 Buffer 7.5 μl 50% PEG8000 0.45 μl 33 mM DTT 0.75 μl T4 RNL2 Tr. K227Q	30°C 6 hr	65°C 20 min
2. RT	15 μl Ligation Rxn 1 μl 10 μM RT Primer 2.25 μl 10 mM dNTP mix 14.3 μl Water	65°C 5 min; 4°C hold	9 μl 5X FS Buffer w/o $MgCl_2$ 2.25 μl 100 mM DTT 1.2 μl SSIII	55°C 30 min	70°C 15 min
3. Gel Purification	Gel percentages and run times vary with insert length				
4. Circularization	10 μl RT Product (all) 2 μl CircLigase I Buffer 1 μl 1 mM ATP 1 μl 50 mM $MnCl_2$ 4 μl 5M Betaine 1 μl CircLigase I			60°C 3 4 hr	80°C 10 min
5. PCR: Test PCR	7.5 μl KAPA 2X HiFi 0.75 μl 10 μM PE 1.0 0.75 μl 10 μM PE 2.0 ≤ 3 μl Circularization Rxn to 15 μl Water		Denaturation Cycling	98°C 45 sec 98°C 15 sec 65°C 30 sec 72°C 30 sec	
Large Scale PCR	25 μl KAPA 2X HiFi 2.5 μl 10 μM PE 1.0 2.5 μl 10 μM PE 2.0 3.3*Circ. Rxn volume used in test to 50 μl Water		Final Extension	72°C 1 min	

A

Ligation Conditions		Barcode	3' Adaptor Sequence	Library Size	Filtered N24 Reads	Species Count
4°C	18 hr	ATCTG	Fixed	19,793,346	13,681,230 (69.1%)	13,625,394 (99.6%)
22°C	1 hr	AGCTA	Fixed	21,013,751	14,997,628 (71.4%)	14,939,326 (99.6%)
	6 hr	TGACT	Fixed	17,191,147	11,431,878 (66.5%)	11,405,124 (99.8%)
30°C	20 min	CGGGA	Fixed	17,780,191	12,997,144 (73.1%)	12,936,600 (99.5%)
	1 hr	ACAAG	Fixed	17,768,912	12,874,223 (72.5%)	12,824,685 (99.6%)
	6 hr	ATTCA	Fixed	13,714,279	8,996,410 (65.6%)	8,971,183 (99.7%)
	6 hr	TCTAC	5'NNNN	17,915,319	12,278,968 (68.5%)	12,235,347 (99.6%)
			Combined Data:	125,176,945	87,257,481 (69.7%)	86,935,208 (99.6%)

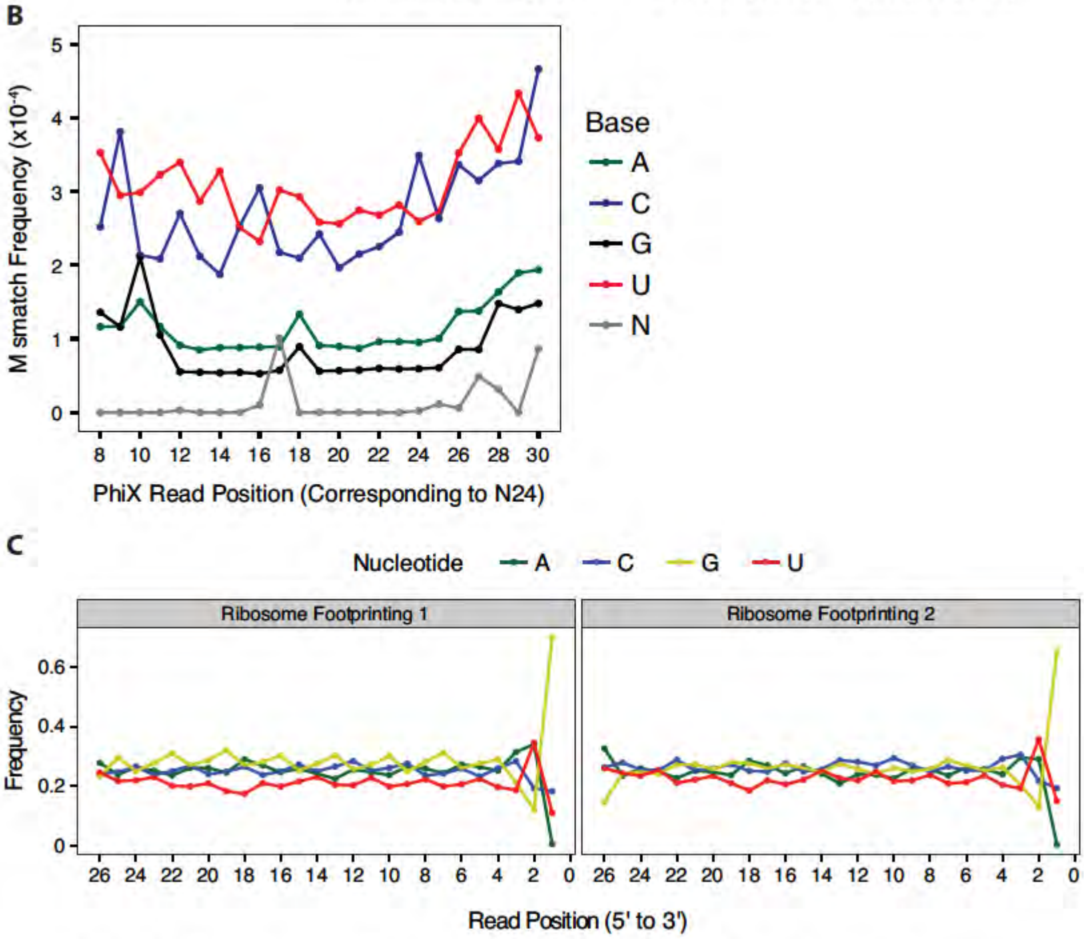


FIGURE 3.7: N24 Library Details, Error Frequency and Bias in Ribo-Seq Libraries

(A) Table of ligation conditions for N24 libraries. (B) Sequencing mismatch frequency as a function of PhiX read position. (C) Nucleotide frequency as a function of ribosome footprinting read position. Two biological replicates from HEK-293 cells are shown (Ricci et al., 2014). The nt frequency at the 3' end (position 1) is due to footprint isolation by digestion with RNases A and T1.

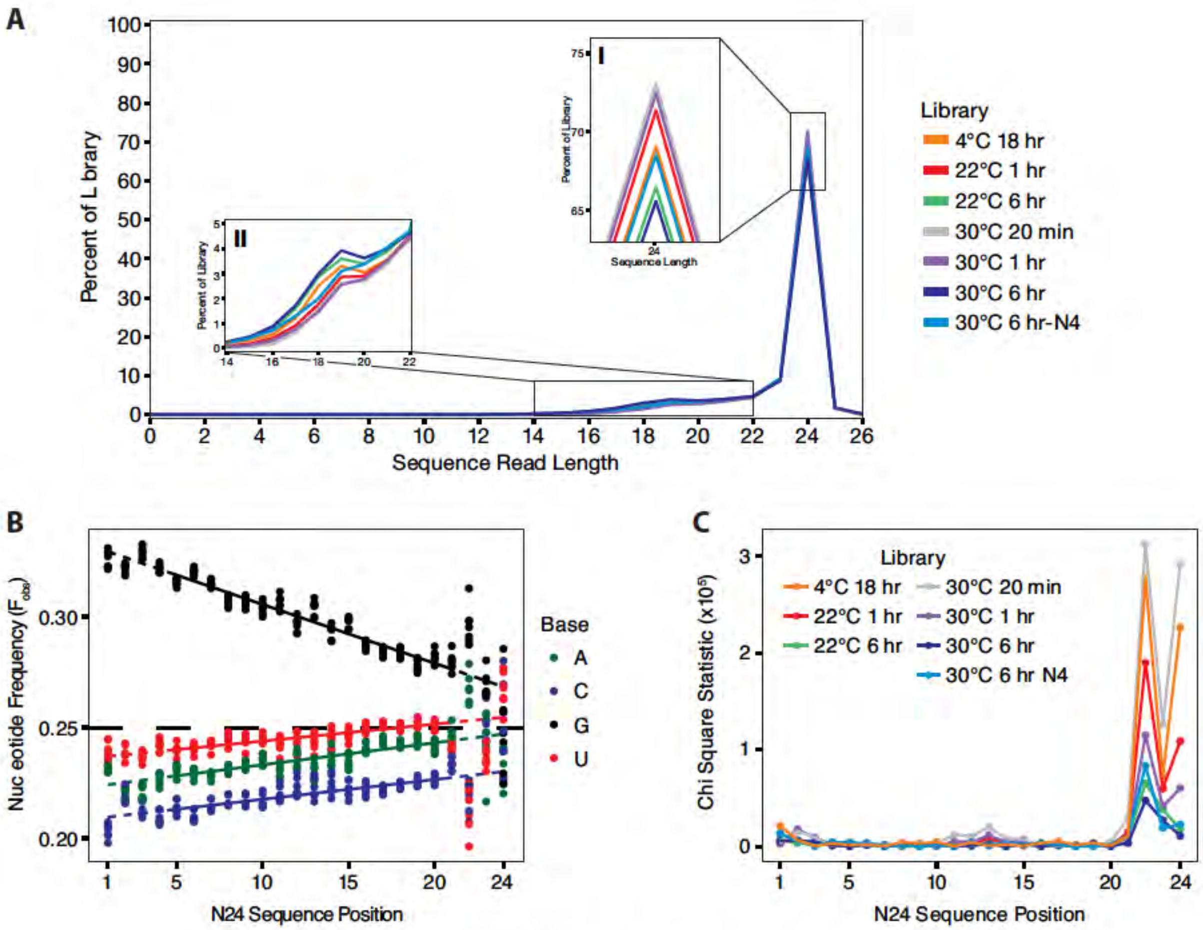


FIGURE 3.8: N24 Length and Bias Analysis

(A) Distribution of read lengths, shown as a percent of the total sequences. (B) Nt frequency versus N24 sequence position. Dashed line indicates ideal 25% incorporation and capture of all four nts. (C) Total bias at each N24 sequence position.

outnumbers the reads obtained per library (~10^7), unique species constituted >99.5% of each library and >99.6% of the entire pooled data set (Supplementary Figure 3.7A). Because each library captured a unique sequence set, it was not possible to calculate the capture frequency for individual species. Therefore, to assess capture bias driven by nt identity, we measured nt frequency at each position in our captured fragments (Figure 3.8B). Across all libraries, there was a notable enrichment in G that decreased linearly in the 5' → 3' direction. To determine the extent to which this might be due to base misincorporation/miscalling at the sequencing level, we determined the mismatch frequency in the PhiX fragments sequenced alongside our N24 libraries (Supplementary Figure 3.7B). Across all positions corresponding to our N24 inserts, the PhiX mismatch frequency was no greater than 0.00049 for any of the 4 nts, with G being the least frequently miscalled base (<0.00021). Additionally, when analyzing the nt frequency per position in ribosome footprinting libraries made with our optimized ligation conditions, we see no 3' -5' trend toward G enrichment (Supplementary Figure 3.7C). Thus, the most likely explanation for the overabundance of G in the N24 libraries was guanosine phosphoramidite overincorporation during oligonucleotide synthesis (Bartel and Szostak, 1993).

Examination of Figure 3.8B reveals that the majority of interlibrary variance occurred at the 3' termini of captured RNAs (positions 21–24). To estimate expected nt frequencies (F_{exp}) at these terminal positions, we used the observed frequency (F_{obs}) data from all libraries to generate four best-fit lines (one for each nt) through positions 5–20 (Figure 3.8B), as these internal positions should be least affected by enzyme preference during 3' adaptor ligation and circularization. We then used these best-fit lines to calculate expected nt counts at every nt position for each library. Calculating the chi-square statistic allowed us to quantify the deviation in observed nt count from expected nt count (Figure 3.8C). This analysis revealed that the chi-square statistic at positions 21–24 decreased in the following order: 30°C–20 min > 4°C–18 h > 22°C–1 h > 30°C–1 h > (30°C–6 h ~ 30°C–6 h-N4 ~ 22°C–6 h). That is, the libraries exhibiting the greatest deviation from expected were those wherein 3' adaptor ligation was only ~30–85% complete (Figure 3.2F), either because of insufficient incubation time or a suboptimal ligation temperature. For reactions that did proceed to apparent completion (the three 6-h libraries), inclusion of four randomized nts at the 5' end of the 3' adaptor (5' N4) had no additional benefit in reducing position 21–24 deviation compared to the fixed-sequence 3' adaptor (although see miRNA data below).

Unexpectedly, position 22 exhibited equal or greater deviation than position 24 in all seven libraries. When comparing F_{obs}–F_{exp} for each nt, another feature readily observable in the 30°C–20 min library, and to a lesser extent in the 30°C–1 h library, is a tendency toward higher GC content at positions 11–15

(Supplementary Figure 3.9). Currently, we have no clear explanations for either of these effects (see Discussion section), but both strengthen the point that uneven capture is accentuated by short ligation times.

Method Validation

To assess how our optimized protocol performs on a known RNA sample, we made libraries from 50 fmol or 1 pmol of an equimolar 29 miRNA pool previously used to benchmark small RNA library preparation (SRR899527 and SRR899530; Zhang et al., 2013). Barcoded libraries were generated using either the fixed or N4 preadenylated 3' adaptor, then pooled and sequenced on a single MiSeq lane (Table 3.2). Plotting F_{obs} versus F_{exp} (where F_{exp} = 1/29 = 0.0345) revealed no recurring over- or underrepresentation pattern for any individual miRNA across our four libraries (Figure 3.10A). Importantly, all four of our libraries exhibited less variability than both the previous benchmark (Zhang et al., 2013) (Figure 3.10B) and a new library preparation protocol for capturing scarce miRNAs (Sterling et al., 2015). In our libraries, the lowest CV in F_{obs} were obtained with the fixed adaptor at 1 pmol input and the N4 adaptor at 50 fmol and 1 pmol input. At 50 fmol input, however, the fixed adaptor did result in somewhat higher variability. Therefore, the N4 adaptor may be preferable when using our protocol to construct libraries from very low input RNA.

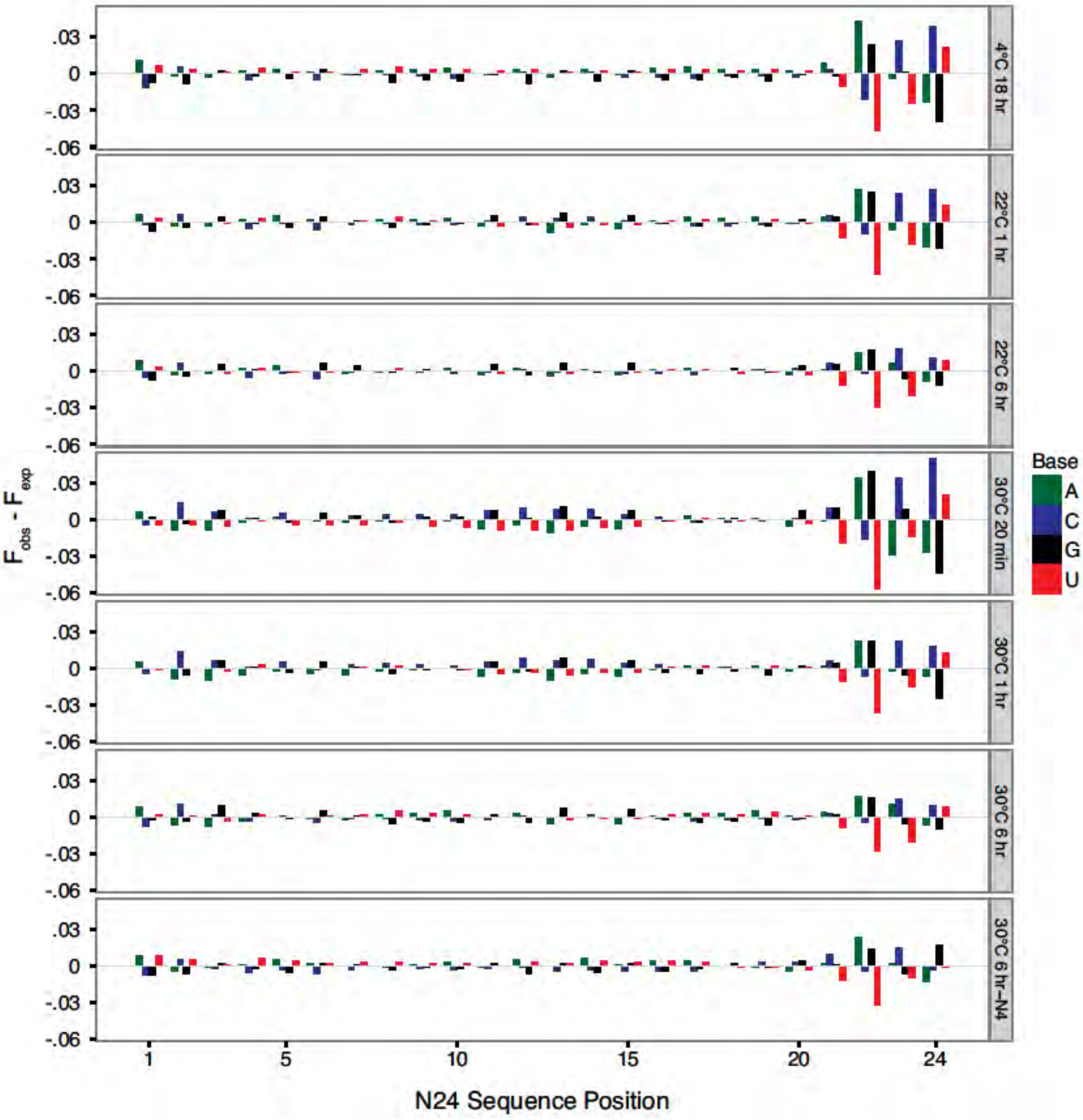


FIGURE 3.9: Nucleotide Bias

Nucleotide bias across all N24 positions in each library.

TABLE 3.2: miRNA Libraries

Input	Adaptor	Sequencing Platform	Mapped Reads
1 pmol	Fixed	MiSeq	1,044,234
	N4		1,393,238
50 fmol	Fixed		1,389,911
	N4		676,609
SRR899527		HiSeq 2000	715,728
SRR899530			1,424,004

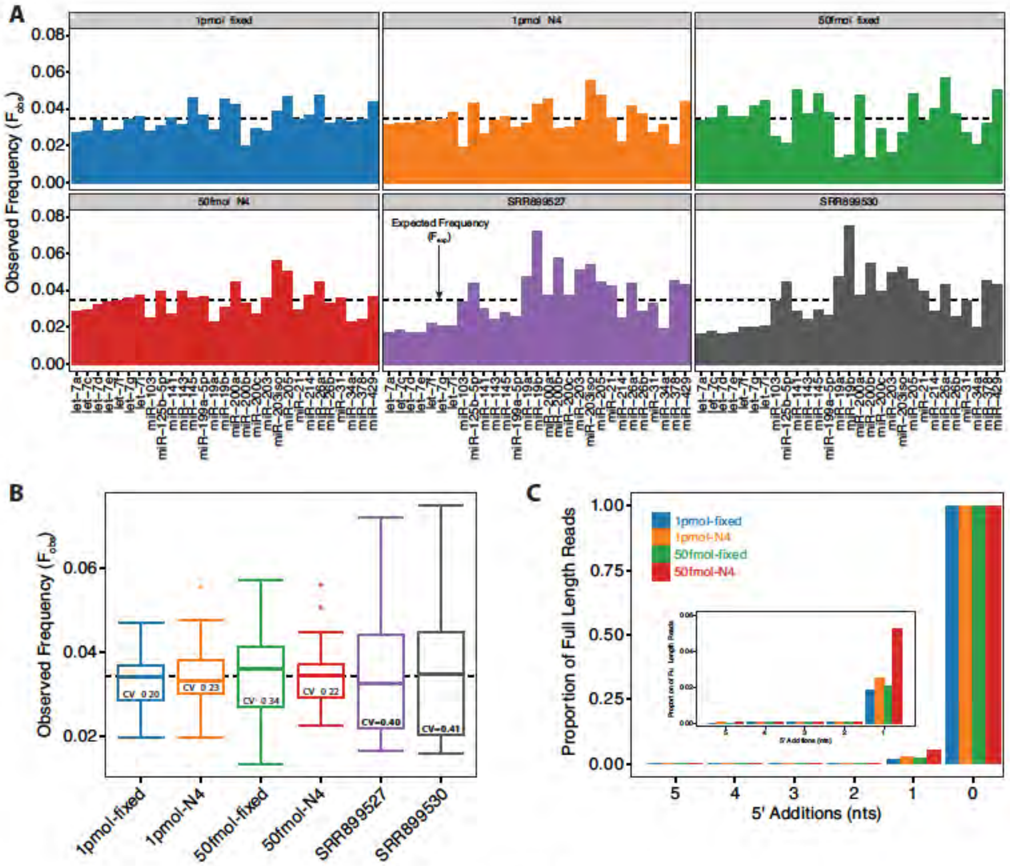


FIGURE 3.10: miRNA Pool Libraries

(A) Proportion of each miRNA in each library. Line represents perfectly even capture with each miRNA representing 1/29th of the reads. (B) Boxplot showing the distribution of proportions. CV = standard deviation (miRNA counts)/mean (miRNA counts). (C) Terminal transferase activity. Barchart showing percent of 5' additions and subtractions as a percentage of full-length reads.

It has previously been noted that both secondary structure internal to individual miRNAs and the ability of individual miRNAs to hybridize to the 3' adaptor can affect capture efficiency (Sorefan et al., 2012, Zhuang et al., 2012). To address this possibility, we made scatter plots of read frequency versus individual miRNA features and calculated both slope and ρ-value for the line best fitting the data (Supplementary Figure 3.11). (We note that a slope other than 0 is potentially indicative of bias, with the magnitude of the slope indicating the strength of the bias dependent on the particular feature being plotted. The ρ-value indicates only how well the line fits the data.) These plots revealed no correlation with a $|\rho$-value$| > 0.5$ between F_{obs} and GC-content, or between F_{obs} and the calculated folding energies (ΔG) for each miRNA alone or each miRNA co-folded with the adaptor in any of our four libraries. We could also detect no apparent folding energy effects in the previous benchmark libraries. With the latter samples, however, there were readily observable trends with regard to nt composition, the most significant being a negative correlation (mean slope m = -0.058; mean ρ = -0.72) between F_{obs} and the number of U's in the last 10 nts of each miRNA (Supplementary Figure 3.12). This is consistent with our N24 data showing an increased bias against U's in the last few nts when ligation reactions conditions are suboptimal (Supplementary Figure 3.9). The absence of the same trend in our miRNA libraries highlights the more even coverage provided by our optimized ligation conditions.

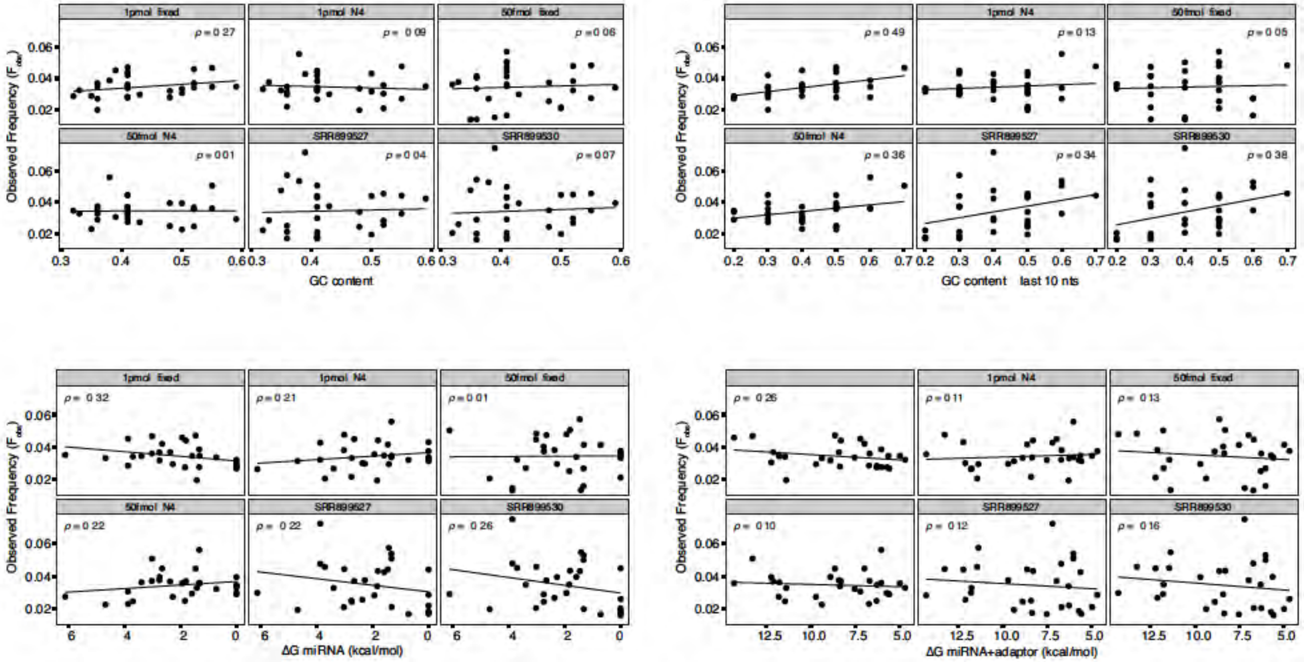


FIGURE 3.11: Effect of GC Content and ΔG on miRNA Capture Frequency

Scatterplots showing the relationship between miRNA capture frequency and GC content and folding energies. ρ is the pearson correlation coefficient; m is the slope.

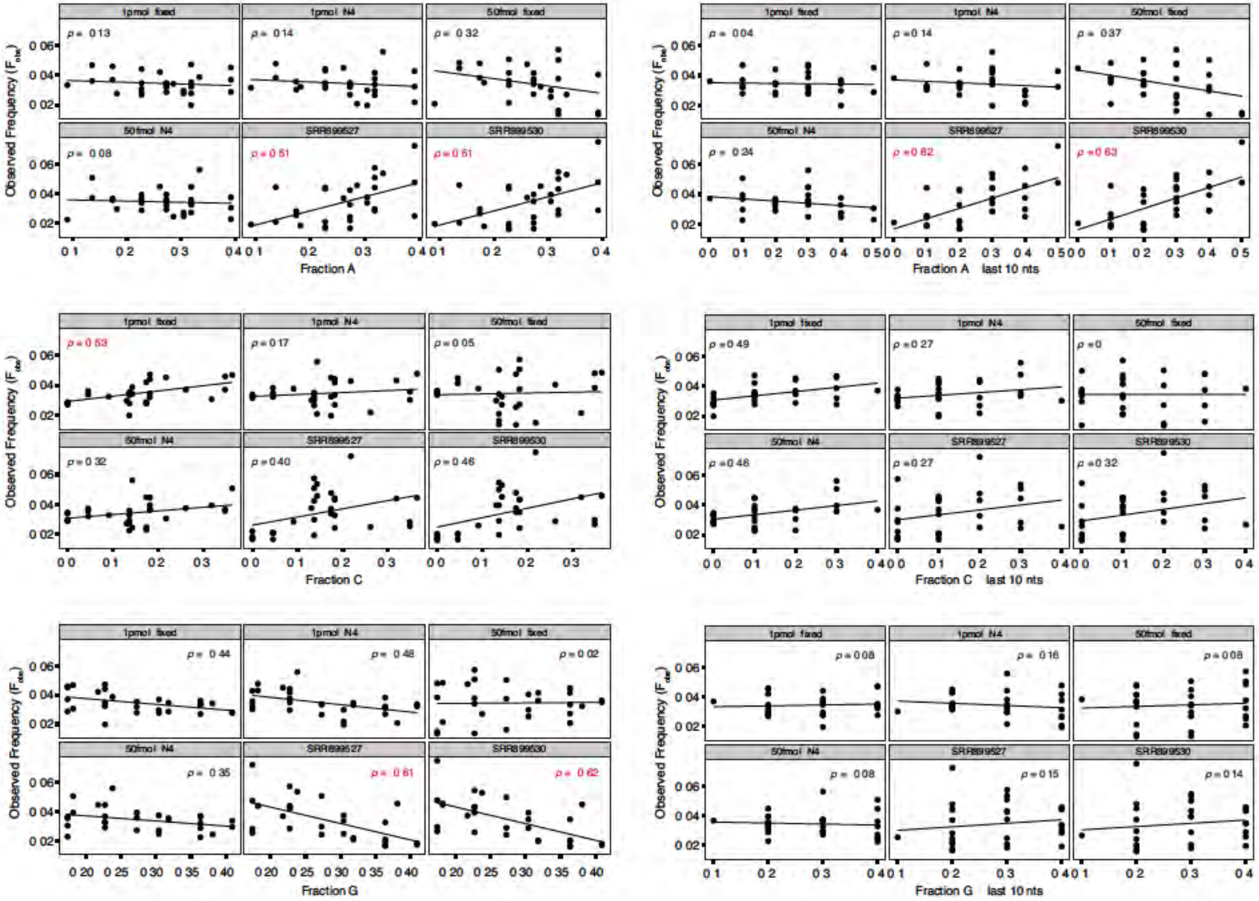


FIGURE 3.12: Effect of Nt Content on miRNA Capture Frequency (continued on next page)

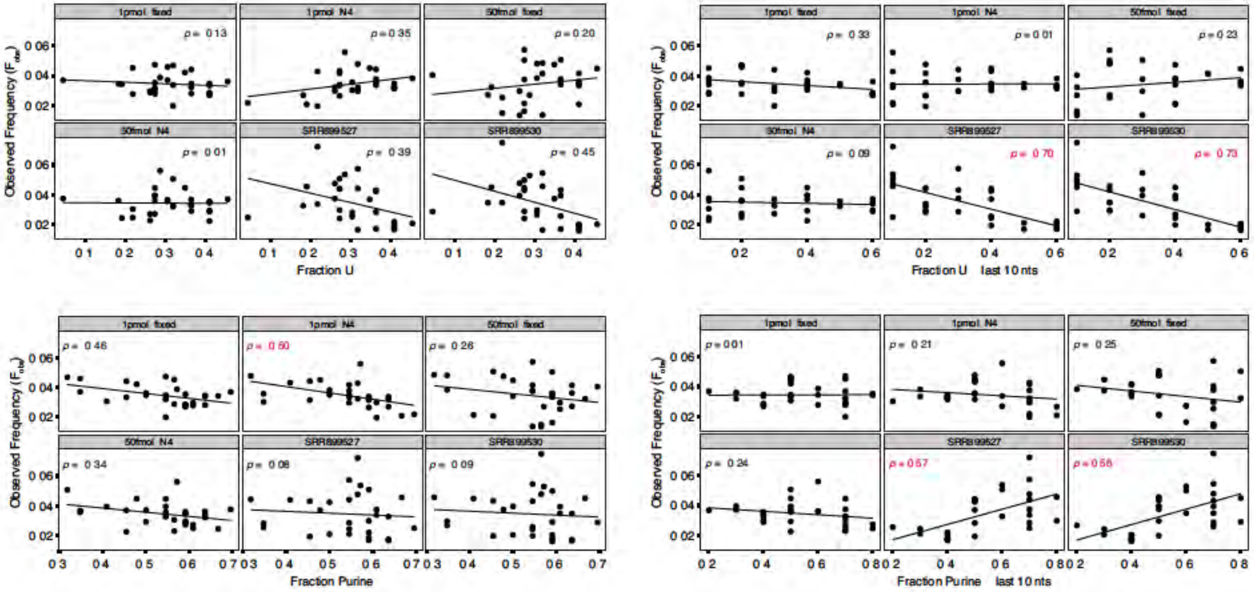


FIGURE 3.12: Effect of Nt Content on miRNA Capture Frequency

Scatterplots showing the relationship between miRNA capture frequency and nt content. ρ is the pearson correlation coefficient; m is the slope.

Under some conditions, reverse transcriptases can exhibit terminal transferase (TdT) activity, resulting in non-templated nt addition to cDNA 3' ends (Chen and Patton, 2001). Examination of our miRNA libraries revealed that, while some untemplated addition did occur, extensions were generally limited to a single nt and these extended species were 20- to 50-fold less abundant than full-length species (Figure 3.10C). During preparation, these samples were immediately gel purified after RT (Supplementary Table 3.1). With one set of libraries, we observed more extensive TdT activity when the RT reaction was maintained at 4°C overnight following the heat inactivation step (data not shown). This suggests that Superscript III is not completely inactivated by the manufacturer's suggested heat inactivation regimen and will continue to add untemplated nts during long, low temperature incubations.

Discussion

In this study, we set out to develop a method that yields robust strand-specific deep sequencing libraries from diverse RNA inputs. Our method involves 3' ligation of a preadenylated adaptor followed by RT, circularization and PCR. This approach combines features of several previously published protocols (Ingolia et al., 2009, Lau et al., 2001, Lui et al., 2007), with modifications to enhance capture efficiency and minimize sample loss. Our method works across a range of input amounts, is easy to follow, and produces a library in 2–3 days at relatively low reagent cost (<$25 per sample), all while giving the user complete control over every step. Because the input to our method is generic single-stranded RNA with a 3' hydroxyl, it can be used to capture many different sized RNA footprints. Our approach can also be used to map sites of RNA-protein crosslinking (e.g. from CLIP experiments) and other base modifications that cause reverse transcriptase to either stall (e.g. abasic or alykylated sites) or incorporate the wrong base (e.g. PAR-CLIP). To date, various members of our laboratory have used this method to generate multiple footprinting libraries for Ribo-Seq and other RNA-protein complexes, as well as RNA-Seq libraries (Ricci et al., 2014 and unpublished results). Input fragment sizes have ranged from 20 to 200 nts, input amounts have ranged from 400 pg to 200 ng RNA and all resulted in highly complex libraries. Our method is highly reproducible, with both read counts and RPKM for Ribo-Seq and RNA-Seq biological replicates having correlation coefficients of 0.93–0.99 (Ricci et al., 2014 and unpublished results).

One of our major goals in developing this protocol was to minimize capture biases. We did so by identifying conditions wherein both the RNL2 and CircLigase reactions were driven to apparent completion, thereby minimizing ligase sequence preferences and any intra- and inter-molecular secondary structure effects. Our analysis of the effects of time and temperature on

3' adaptor ligation clearly indicates that incomplete ligation exacerbates capture bias (Figures 3.2, 3.8B and 3.8C and Supplementary Figure 3.9). Nonetheless, even under conditions where the ligation reaction appeared to proceed to completion, apparent 3' end biases were not fully eliminated (Figure 3.8C). Three recent papers reported that 3' end capture bias can be reduced by including a short (2–4 nt) randomized region at the 5' end of the 3' adaptor (Jayaprakash et al., 2011, Sorefan et al., 2012, Zhang et al., 2013). Inclusion of degenerate nts in the adaptor also allows for identification of species that are preferentially amplified during the PCR reaction (König et al., 2010). Although we observed no advantage of the N4 adaptor over our fixed sequence adaptor with 1–2 pmol N24 or miRNA pool input (Figures 3.8C, 3.10A, 3.10B and Supplementary Figure 3.9), the N4 adaptor was clearly superior when the miRNA pool input was lowered to 50 fmol (Figure 3.10A and B). Therefore, using a 5' randomized adaptor is recommended.

Contrary to expectation (Sorefan et al., 2012, Zhuang et al., 2012), we could detect no effects on N24 or miRNA capture efficiency that could be attributed to either internal secondary structure forming propensity or the ability of captured sequences to hybridize with the adaptor (Supplementary Figure 3.11 and data not shown). In our N24 data, however, we did detect an unexpected nt identity bias at the -3 position relative to the 3' adaptor ligation site (Figure 3.8C). This is consistent with a previous report demonstrating -3 substrate bias by both

RNL1 and RNL2 (Zhuang et al., 2012). Currently, there is no clear explanation for this effect, as a crystal structure of RNL2 bound to substrate suggests that RNL2 substrate specificity is dictated solely by the nts at positions -1 and -2 (Nandakumar et al., 2006). Nonetheless, our N24 data highlight the importance of driving the 3' ligation reaction as close to completion as possible.

Following ligation, RT of the captured RNA attaches a sequence tag to the 3' end of the RNA, allowing for PCR amplification and deep sequencing. Although the adaptor sequences used here are for sequencing on Illumina platforms, libraries can be prepared for any deep sequencing platform by simply modifying the 5' and 3' adaptor sequences. Our method employs a variety of RT primers that differ only by their 5' barcode, allowing multiple samples to be sequenced on the same flow cell lane. Barcoding the samples during the RT step minimizes opportunities for accidental mixing or cross-contamination of samples. We currently use a set of twelve 5-nt barcodes (see Experimental Procedures section) that were chosen such that the first position is balanced (to increase initial base calling accuracy by Illumina platforms) and there is no possibility for barcode misidentification, even with two sequencing errors. After circularization, the barcode is positioned 5' to the captured cDNA sequence, allowing for barcode identification and fragment sequencing all in one single-end sequencing run.

Following circularization, one must determine the optimal number of PCR cycles for each sample. Cycle number is highly dependent on the original RNA input amount. Our current approach is to empirically determine the correct number of PCR cycles by gel analysis; too few cycles will result in product yield below the sequencing input requirement; too many cycles will result in PCR jackpots that can overwhelm the library and introduce significant bias. A recently published qPCR approach for identifying the correct number of cycles can easily be applied to our method (Langevin et al., 2013).

Two similar protocols for making strand-specific libraries were recently published (Epicentre Technologies Corporation, 2012, Ingolia et al., 2012), speaking to the overall strength of this strategy. Nonetheless, the modifications we describe here (i.e. inclusion of 25% PEG in the 3' adaptor ligation reaction; no additional MgCl2 in the RT reaction; a single gel purification step; inclusion of 1M betaine in the CircLigase I reaction; and optimized times and temperatures to ensure completion of all reactions) offer significant improvements over similar methods. To assist the reader in implementing our protocol, we have included a short summary of the conditions (Supplementary Table 3.1) and placed a detailed protocol at http://www.umassmed.edu/moorelab/resources/protocols/.

Accession Numbers

High-throughput sequencing data have been deposited in the GEO database under accession number GSE63606.

Acknowledgements

We thank Catherine Sterling, Victor Ambros and Phillip Zamore for helpful discussions related to method development. Rui Yi kindly provided the 29 miRNA mix. Fred Hyde and Hank Daum at Epicentre aided in troubleshooting the circularization reaction.

Experimental Procedures

Gel Analysis

All acrylamide gels were prepared using AccuGel reagents (National Diagnostics). Ligation samples were prepared in an equal volume of 2X denaturing load buffer (12% Ficoll Type 400-DL, 7 M Urea, 1X TBE, 0.02% Bromophenol Blue, 0.02% Xylene Cyanol), denatured for 5 min at 95°C and cooled on ice prior to loading on denaturing 15% polyacrylamide (19:1)-8 M Urea-1X TBE gels. Reverse transcription (RT) samples were diluted in one-third volume of 3X denaturing load buffer (18% Ficoll Type 400-DL, 10.5 M Urea, 1.5X

TBE, 0.02% Bromophenol Blue, 0.02% Xylene Cyanol), denatured for 5 min at 95 °C, and analyzed on 10% denaturing polyacrylamide gel electrophoresis (PAGE) gels. Circularization reactions were prepared similarly to ligation reactions and analyzed on 10% denaturing PAGE gels. Polymerase chain reaction (PCR) products for gel analysis were mixed with 5X non-denaturing load buffer (15% Ficoll Type 400-DL, 1X TBE, 0.02% Bromophenol Blue, 0.02% Xylene Cyanol) before separation on native 8% PAGE gels. PCR products to be sequenced were similarly prepared and analyzed on the Double Wide Mini-Vertical system (C.B.S. Scientific) to limit the amount of heat denaturation. Gels were either exposed to a phosphorimager screen (Amersham Biosciences) or stained with SYBR Gold (Invitrogen) prior to visualization on a Typhoon Trio (Amersham Biosciences). Quantifications were performed with ImageQuant (GE Healthcare).

3' Adaptor Ligation

Indicated amounts of either 5' ^{32}P-labeled N24 RNA oligonucleotide (Dharmacon) or 28-mer oligonucleotide (5'-AUGUACACGGAGUCGAC CCGCAACGCGA-3'; IDT) were ligated to preadenylated adaptor mirCat-33 (5'-rAppTGGAATTCTCGGGTGCCAAGGddC-3'; IDT) or EH-preaden (5'-rAppNNNNTGGAATTCTCGGGTGCCAAGGddC-3'; IDT) using T4 RNL2 Tr. K227Q (NEB) with the conditions described in this paper. Due to the high viscosity of 50% PEG8000, we found that low retention filter tips aided consistent pipetting while simultaneously preventing sample cross-contamination. Ligation efficiencies were calculated by dividing the quantified pixel signal of ligated RNA by the total amount of RNA signal (bands corresponding to both ligated and unligated RNA) in each lane, and multiplying by 100.

Reverse Transcription

RT was performed with gel purified RT primers 5'-pGG-**B**-AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-SP18-CTCGGCATTCC TGCTGAACCGCTCTTCCGATCT-CCTTGGCACCCGAGAATTCCA-3', where **B** indicates a 5-nt barcode of sequence ATCAC, CGATG, TAGCT, GCTCC, ACAGT, CAGAT, TCCCG, GGCTA, AGTCA, CTTGT, TGAAT or GTAGA. RT products were detected by incorporating α-^{32}P-dCTP in the reaction. RT products intended for circularization were gel purified. For the data in Figures 3.5 and 3.6, we eluted the cDNA from crushed gel pieces in 300 mM NaCl, 1 mM ethylenediaminetetraacetic acid (EDTA) during an overnight incubation at room temperature with constant rotation; eluted material was ethanol precipitated before circularization. We have since modified our approach to increase elution yield by eluting in TE (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0) and incubating at 37°C overnight with constant rotation. With this buffer, we can concentrate the eluate (either by butanol extraction or SpeedVac) before precipitating the sample in a single tube.

Circularization Efficiency and PCR Amplification

Circularization reactions were performed on gel-purified RT product as described in the text. The single-stranded DNA input was either body-labeled with α-^{32}P-dCTP in the RT reaction or end-labeled in an exchange reaction with ^{32}P-γ-ATP. Circularized RT product was separated from non-reactive, linear RT product on 10% denaturing PAGE gels, and the gels were exposed and quantified as described. The amount of circularization was determined by quantifying the pixel signal corresponding to the circularized product and dividing that value by the total pixel signal corresponding to the circularized product plus the remaining linear input, and multiplying by 100.

PCR amplification from the circularized RT product was performed with KAPA HiFi Library Amplification Kit (Kapa Biosystems) according to manufacturer's instructions, except where otherwise noted. All PCR products were analyzed on native 8% PAGE gels and quantified as described above. Samples to be sequenced were excised and gel extracted as described for RT products, precipitated and quantified by gel analysis before sample submission.

N24 Library Construction and Analysis

N24 libraries were constructed from 2 pmol of N24 RNA oligo using the optimized conditions shown in Supplementary Table 3.1, except for the described variations in 3' ligation conditions. In one case (22°C 6 hr library), a minute amount of

28-mer oligo was added. All libraries were amplified with 7 PCR cycles and gel purified prior to sequencing on a single Illumina HiSeq2000 lane (Genewiz).

Deep sequencing data were analyzed with custom scripts unless otherwise noted. Data were parsed into individual libraries by 5' barcode, allowing 1 mismatch. The 3' adaptor sequence was removed from all libraries allowing 3 mismatches. Once individual sequence reads were identified, read lengths were calculated. All subsequent analysis utilized only 24 nt reads. For each library, we calculated the observed nt frequencies at each of the 24 positions. To determine expected values, we used the data across positions 5–20 from all libraries and fitted least squares lines to the frequency pattern for each nt. The equations for the line-fits yielded the expected nt frequencies at all 24 positions. The chi-square statistic was calculated for each library by summing [(observed nt count - expected nt count)^2/(expected nt count)] across all four nts at each N24 position.

PhiX reads were identified if they mapped to the PhiX174 genome with a maximum of 6 errors within the 51 sequenced nts. Mismatches were identified and counted if the sequenced nt was different than the PhiX174 genome sequence. Mismatch frequencies were calculated by dividing the mismatch counts at each position by the total number of PhiX reads. For analysis of nt distribution across ribosome footprints (Ricci et al., 2014), all 26–30 nt reads were selected and aligned by their 3' ends; nt frequencies were calculated by dividing the observed nt count at each position by the total number of reads.

miRNA Library Construction and Analysis

Libraries were constructed from either 1 pmol or 50 fmol of an equimolar mix of 29 miRNAs (Zhang et al., 2013) according to the optimized conditions shown in Supplementary Table 3.1. For each input amount, the ligation was performed with either the fixed or N4 preadenylated 3' adaptor. Libraries were pooled and sequenced on a single MiSeq lane. Deep sequencing data were parsed into individual libraries by 5' barcode using cutadapt version 1.3 (Martin, 2011), allowing 1 mismatch. Reads were mapped to reference sequences using a custom script which (i) required that the 3' adaptor be present in the read and (ii) only counted reads mapping to reference miRNA sequences with 0 mismatches. Additionally, we counted the reads with 5 or fewer non-templated 5' terminal additions and 5 or fewer 5' terminal deletions. Observed miRNA frequencies (F_{obs}) were calculated using the total number of reads for each miRNA (including 5' terminal additions and subtractions). The expected frequency (F_{exp}) for each miRNA is 1/29 or 0.0345. Coefficients of variation (CV) were calculated by dividing standard deviation (miRNA counts) by the mean (miRNA counts). Terminal transferase activity was assessed by dividing total miRNA reads in each 5' addition bin by the total full-length miRNA reads in each library. Free energy values from in silico folding were calculated using the Vienna RNA Package v. 2.1.7 using the -T 30 parameter to obtain structure predictions at

30°C (Lorenz and Bernhart, 2011).

Chapter IV

Discussion

Monosome Translation

The work presented in Chapter II overturns a long-existing dogma that monosomes are inactive for translation. By combining monosome enrichment analysis with multiple genome-wide datasets, we found that monosomes predominantly translate mRNAs with short ORFs, endogenous NMD targets and mRNAs for which the time to initiate or translate the first portion of the ORF is much greater than the time to elongate across the remainder of the ORF.

Implications of the Inactive Monosome Assumption

Ever since the discovery of polysomes (Warner et al., 1963), it has commonly been assumed that monosomes are translationally inactive. This assumption is based on two pieces of data: (1) No incorporation of radioactive amino acids was detected in monosomes (Warner et al., 1963), and (2) in experiments measuring the number of tRNAs per ribosome, only 1 tRNA cosedimented with 80S ribosomes (Warner and Rich, 1964). These experiments were both performed in rabbit reticulocyte lysate, which is a very unique system in that the *in vivo* role of reticulocytes is to produce hemoglobin; therefore, the vast majority of mRNAs in reticulocyte lysate code for hemoglobin proteins. Warner et al. were quite careful in the assessment of their data demonstrating translation by polysomes, stating that "our experiments indicate that protein synthesis in the reticulocyte occurs only on this structure and not on a single ribosomal unit" (1963). However, this finding has been indiscriminately applied to translation across all organisms, resulting in an oversimplification of the translation system: an mRNA's association with 0 or 1 ribosome indicates no translational activity, while association with 2 or more ribosomes indicates translational activity.

This assumption has both limited experimental findings and resulted in potentially inaccurate conclusions. Take, for example, the recent ribosome profiling experiments identifying short ORFs in both *S. cerevisiae* and *Drosophila* S2 cells (Aspden et al., 2014, Smith et al., 2014). In both cases, the assumption that translationally active mRNAs must be associated with polysomes resulted in experimental analysis being limited to those mRNAs which cosediment with polysomes. Short ORFs which cosediment exclusively with monosomes, either due to extremely long initiation times or short ORF lengths, were excluded from collection and data analysis. Similarly, the inactive monosome assumption has affected conclusions drawn from comparative polyribosome profiling. Consider an experimental setup where mRNAs are monitored for the number of ribosomes with which they cosediment - either by northern blotting or qPCR of RNA extracted from polysome profile fractions - before and after a specific treatment, or across cell cycle stages. An mRNA shift from polysomes to monosomes has typically been interpreted as a lack of translation due to cell treatment or stage. However, we now know that this likely reflects a decrease in translation and not a complete shutoff; that is, the translation of this mRNA has been downregulated.

Often, we think about identifying the set of mRNAs translated during each experimental condition or treatment, which can lead to thinking of translation as a binary system, with the 2 possible states of translational status for each mRNA being either on (expressed) or off (not expressed). However, we need to think of translation as more fluid, with varying states of expression levels that change depending on cell cycle or environment. A shift from polysomes to monosomes indicates a decrease in protein expression but not a complete halt; nuances in gene expression regulation may have been overlooked because of these assumptions. As an example, a recent experiment examined translation across the temporal scale of *S. cerevisiae* meiosis, identifying sets of genes with increased or decreased expression throughout various meiotic stages (Brar et al., 2012). Given the sensitivity of deep sequencing and ribosome profiling, we now have the tools to measure translation across a wide range of expression levels.

Future Monosome Footprinting Experiments

When preparing *S. cerevisiae* lysate for polysome profiling, I excluded detergent from the lysis buffer for reasons mentioned previously. However, if repeated, I would add an additional step to fully recover ribosomes from the ER membrane via post-lysis incubation with detergent (Seiser and Nicchitta, 2000). As a result, both membrane-bound and cytoplasmic ribosome footprints would be captured, generating a more accurate snapshot of total cellular translation. In addition to adding information on the ~2400 genes *S. cerevisiae* genes which are predominantly translated on membrane-bound ribosomes, quantifications for cytoplasmic mRNAs might change as well, as work from the Nicchitta lab has demonstrated that mRNAs encoding soluble proteins can be translated by ER-bound ribosomes (Reid and Nicchitta, 2015). To specifically examine the role of monosome translation on the ER, it would be interesting to perform compartment-specific monosome- and polysome- ribosome profiling in *S. cerevisiae*. Previous analysis in human cell lines showed little difference between cytoplasmic- and ER-ribosome footprint profiles of predominantly cytoplasmic mRNAs (Jagannathan et al., 2014), so we might expect a similar finding in yeast. However, by teasing apart the relative contributions of monosome- and polysome- footprinting, we might detect a change in footprint patterns across ORFs that would depend on the subcellular site of translation.

As part of our analysis of monosome translation, we calculated that 93% of monosomes are translationally active. However, due to the nature of the experiment, we cannot draw any conclusions about the percentage of monosomes associated with mRNA (Bhattacharya et al., 2010). We expect that some portion of the signal in the 80S monosome peak represents empty couples, though the 5 mM $MgCl_2$ in the lysis buffer should minimize ribosomal subunit interaction in the absence of mRNA (see Chapter II, section entitled Monosome, Polysome and Global Footprinting). Given the long-standing debate regarding monosome translational status, it would be interesting to actually quantify the percentage of monosomes that are associated with mRNA. PolyA+ selection from monosome fractions could be used to separate mRNA-bound ribosomes from free ribosomes, followed by ribosomal RNA quantification by northern blot or qPCR. Either an excess of oligo-dT or multiple rounds of polyA+ pulldown would be necessary to ensure that all mRNAs were removed from the sample. This experiment may underestimate the amount of mRNA-bound monosomes, however, since polyA+ selection will not capture mRNAs with a short polyA tail or mRNA fragments.

Monosome Footprinting to Identify Novel Short ORFs

As demonstrated in Figures 2.6A-C and 2.5D, monosome-associated mRNAs are enriched for sORFs, a finding not entirely surprising since monosomes are primarily responsible for the translation of RPL41A and B, the shortest ORFs in *S. cerevisiae* (Yu and Warner, 2001). This enrichment included both short upstream ORFs (uORFs) and novel sORFs. Since we believe this monosome enrichment is dependent on the relationship between translation initiation time and elongation time, we would expect monosome footprints to be enriched for sORFs in other organisms as well.

Recently, there has been a dramatic increase in the publication of sORF papers; seven reviews (Chu et al., 2015, Crappé et al., 2014, Eguen et al., 2015, Kemp and Cymer, 2014, Landry et al., 2015, Ramamurthi and Storz, 2014, Storz et al., 2014) on sORFs have been published in the last 2 years alone! Short ORFs have been found as alternative ORFs in reference genes, though most often are identified within previously annotated non-coding RNAs (Anderson et al., 2015, Lauressergues et al., 2015, Ruiz-Orera et al., 2014). Thus far, the shortest functional ORFs to be discovered are 11 amino acids long (Galindo et al., 2007). Short ORFs are thought to play an important role in biology, as many small molecules and small proteins often have signaling or regulatory roles. However, most of our current knowledge about small proteins comes from short peptides which are enzymatically cleaved from longer precursor proteins (Crappé et al., 2014). There is much to be learned about the function of these small peptides!

The characterization of sORFs and their encoded peptides will both expand our knowledge of the proteome and provide critical insights into cellular biology. However, mapping of entire proteomes is currently biased against small proteins, due to a combination of experimental limitations and historical assumptions. Mass-spectrometry has been used to identify protein sequences genome-wide, but the lower limit of detection means that short proteins are underrepresented compared to longer proteins (Ruiz-Orera et al., 2014). Additionally, historical identification of protein-coding sequences utilized a lower limit of 100 codons, as it is unlikely that by chance, a 300-nt genomic region will lack any stop codons (Harrison et al., 2002). Therefore, the overall abundance of sORFs is likely underestimated, and the assumptions used to identify ORFs need to be revisited. Initial computational experiments have demonstrated the usefulness of combining global ribosome footprinting data with evolutionary codon conservation to identify novel sORFs in yeast, flies, zebrafish, mice, humans and a plant (Bazzini et al., 2014, Crappé et al., 2013, Ruiz-Orera et al., 2014). As monosome footprints are diluted in polysome footprints when following the global ribosome profiling assay, sORF monosome footprints could be overlooked due to low count numbers. Therefore, sORF identification would only be improved with monosome footprinting data.

Potential Applications of Polysome Footprinting

This work has focused on elucidating the translational status of monosomes; consequently, the polysome footprinting data was largely ignored except to provide a comparison for the monosome footprinting libraries. Our limited analysis revealed that polysome-enriched mRNAs have higher mRNA synthesis rates, longer half-lives, and encode highly abundant proteins, results which were hardly surprising. There are, however, many interesting things to be learned from polysome footprinting.

Using an approach similar to the one presented here, polysomes of various sizes could be pooled together and analyzed (ex. 2-4 ribosomes, 4-6 ribosomes, 7+ ribosomes). Split-polysome footprinting has previously been used to identify sORFs (Aspden et al., 2014), likely overlooking some sORFs for the reasons presented above. Several interesting questions could be answered by dividing the polysome pool into multiple groups. For example, would the pattern of footprints change between a small polysome and a large polysome? From our data, we concluded that occupancy on either monosomes or polysomes was predominantly driven by the ratio of initiation rate to elongation rate (see Figure 2.13). What if elongation rate changed depending on the number of ribosomes occupying an ORF? For example, a region with significant secondary structure could slow a ribosome as it unwinds this structure, but a steady flow of ribosomes might prevent the structure from reforming, and as a result footprint abundance would decrease. However, if only a few ribosomes were translating the mRNA, the secondary structure might reform after every ribosome transits this region, resulting in slower moving ribosomes that would create more footprints in this region.

When calculating monosome- versus polysome-enrichment, we found a strong relationship between monosome occupancy and ORF length for short ORFs (Figures 2.6B and 2.5D). This changed drastically around 590 nts, but a slight trend towards increased polysome-enrichment with longer ORFs was still visible. Would small polysomes contain, on average, shorter ORFs than are found in larger polysomes? To some extent this is expected, as longer ORFs can physically accommodate more ribosomes. However, a slow initiation rate on a long ORF could drive its association with small polysomes, while a fast initiation rate could drive a short ORF to large polysome status.

Additionally, the difference between preferential association with small or large polysomes might depend on cellular environment. As the cell responds to some external queue, it could quickly ramp up translation of the proteins required to mount a response by shifting an mRNA from a small to a large polysome. If merely a quick response was needed, a time-course study could show a immediate transition into large polysomes, then medium sized polysomes, then small polysomes or even a lack of footprinting signal if translation was completely inhibited post-response. In contrast, an mRNA could stay in the same polysomal pool but simply transition from inhibited ribosomes to actively elongating ribosomes (see section entitled Comments on Technical Aspects of Ribosome Profiling).

Neurons: An Alternate Experimental System

The work presented above was performed in *S. cerevisiae* for several reasons. First and foremost, this work began soon after the original ribosome profiling paper was published (Ingolia et al., 2009), so our experiments were based on the methods described in that paper. Second, since *S. cerevisiae* is a well-studied model organism, there is a wealth of knowledge upon which to base our potential conclusions regarding monosome translation. Finally, *S. cerevisiae* is easy to grow, so availability and abundance of sample would never be an issue. However, our original plan for monosome footprinting was to perform the experiments in *S. cerevisiae* only as a proof of concept, and move on to a higher organism to study a "more interesting" biological system. As it turns out, the *S. cerevisiae* datasets I generated held a wealth of novel findings, and I did not apply monosome footprinting to another system. However, when first developing the hypothesis of translating monosomes, we were attracted by the hypothesis that monosome translation could be particularly abundant in subcellular compartments where only a very small amount of protein would be necessary to execute its specific function. As neurons are a highly polarized cell, we reasoned that monosome translation may play a role in axon and dendrite local translation. In dendrites, local translation plays a role in maintaining synaptic function and plasticity (Perry and Fainzilber, 2014, Steward and Schuman, 2003, Sutton and Schuman, 2006, Wang et al., 2009). Similarly, local translation in axons enables the response to extrinsic signals (Jung et al., 2011).

Although local translation within both dendrites and axons is now widely accepted, the apparent lack of ribosomes in early studies on mature vertebrate axons lent support to the assumption that axons cannot synthesize proteins. Though there were a few reports of polysomes in axons (for review, see Twiss and Fainzilber, 2009), they were so infrequent compared to reports of dendritic translation that they were largely ignored. Significant evidence regarding the number of dendritic polysomes near synapses came from electron micrographs, where polysomes were identified based on their distinct structural morphology (Ostroff et al., 2002). Using this data, Sutton and Schuman (2006) calculated that, on average, each dendritic spine could only have a single polyribosome, which would introduce serious constraints to the variety of locally synthesized proteins. However, they noted the difficulty of identifying free ribosomes in electron micrographs. So what if a significant amount of dendritic translation happens on monosomes? As little is known about the abundance of mRNAs or proteins in neuronal projections

(Holt and Schuman, 2013), it is possible that a few protein molecules synthesized by a single ribosome would effectively execute its function. Though the overall number of protein molecules may be low, the relative concentration in such a small subcellular region could be quite high.

In recent years, significant work has been devoted to identifying the transcriptome of these neuronal projections (Cajigas et al., 2012, Zivraj et al., 2010). One intriguing finding from these analyses was the abundance (13%) of transmembrane and secretory protein-encoding mRNAs. Can these proteins be translated in neurites? This could be addressed with ribosome profiling to identify ribosome positioning across these mRNAs. The issue of secretory pathway mRNAs in neurites aside, ribosome profiling analysis would add a wealth of knowledge to the local translation field. However, I think it would be much more interesting to perform monosome- and polysome- ribosome profiling in these cellular projections (a) to measure the amount of overall translation and (b) to determine if monosomal translation contributes significantly to the overall amount of translation in axons and dendrites.

Such a system immediately lends itself to several types of experiments. Where would monosome footprints be found in resting neurites, and would that profile change upon pharmacological or chemical stimulation? Would the identity of ribosome-associated mRNA change depending on the stimulant? One possibility is that the entire repertoire of translated mRNAs would change upon stimulation, whether the translation is happening on monosomes or polysomes. On the other hand, monosomes could be pre-loaded on certain mRNAs and merely redistribute into the ORF once translation elongation began.

It would be fascinating to prepare monosome- and polysome- footprinting libraries from different timepoints post-stimulation to determine how quickly translation of individual mRNAs began and ended. If paired with RNA-Seq data, some of the subtle interdependencies between translational control and mRNA levels could be teased apart. If footprints for a specific message were no longer seen after a certain amount of time, ribosome profiling data alone would be unable to address whether ribosomes stopped translating that message or if the mRNA was degraded. This information, however, could be provided by RNA-Seq libraries prepared in tandem with the ribosome profiling libraries.

Comments on Technical Aspects of Ribosome Profiling

In the first iteration of the ribosome profiling protocol, footprints were isolated by digestion with RNase I, an endoribonuclease that cleaves after every base. Thus, the exact footprint of the ribosome could be identified. As this technique has been applied to systems other than *S. cerevisiae*, however, it is not always feasible to use RNase I for footprint isolation. For example, MNase was used to isolate *E. coli* ribosome footprints because RNase I activity is inhibited by bacterial 30S ribosomes (Oh et al., 2011). In the Moore lab, RNase I proved ineffective for preparing ribosome footprints from HEK293 cells, so a combination of RNases A and T1 were used instead (Ricci et al., 2014).

These discoveries of ineffective RNase I digestion would not have been possible without polysome profiling of the digested lysate. The absorbance trace collected during sucrose gradient fractionation reports on the abundance of each ribosome peak and can be used to qualitatively measure the extent of RNase digestion and degradation. Recent work in *C. elegans* demonstrated that RNase I overdigestion causes significant loss of signal in the monosome peak, presumably due to ribosome degradation (Aeschimann et al., 2015). Thus, it is important to optimize the RNase digestion step using polysome profiling when designing ribosome profiling experiments in new systems. Technical developments for the isolation of digested monosomes, such as the use of a sucrose cushion (Ingolia et al., 2012) or a size-exclusion column (TruSeq Ribosome Profiling Kit; Epicentre), are making ribosome profiling experiments accessible to research groups without the capability to perform polysome profiling. While very attractive to the novice user, as they are quicker and more straightforward than polysome profiling, blindly digesting with a fixed concentration of RNase without any optimization is unwise. However, after digestion optimization, these approaches would likely yield similar results to monosome isolation on a sucrose gradient.

The conclusions drawn from ribosome profiling are usually based on the assumption that each ribosome footprint belonged to an actively elongating ribosome. This is similar to the common assumption that the larger the polysome (i.e. the more ribosomes bound to each mRNA), the greater the protein output. However, there is no guarantee that all ribosomes captured in ribosome profiling are actively elongating. In fact, several groups have shown that multiple ribosomes can remain bound to an translationally-inhibited mRNA, indicating that polysomal mRNAs can be translationally inactive (Braat et al., 2004, Nottrott et al., 2006, Olsen and Ambros, 1999, Petersen et al., 2006). For a ribosome leaving a footprint at any location besides the start codon, we can definitively conclude that it was actively elongating at one point, but not necessarily at the time of sample collection.

Consequently, what percentage of footprints originate from stalled ribosomes? This is currently unknown, but could be measured by comparing normal ribosome profiling data to ribosomal run-off footprinting data. In the latter experiments, footprints are likely to be enriched at - and upstream of - stall sites, as any ribosome upstream of the stalled ribosome is likely to accumulate behind it (Guydosh and Green, 2014). Could there be a difference in stuck ribosomes between monosomes and polysomes? It seems likely that stuck ribosomes would predominantly be found in polysomes, as another ribosome could initiate and stack behind the stalled ribosome. However, this would depend on initiation rate and whether the stuck ribosome sterically blocked another 80S from forming.

Future Library Construction Methodological Development

The RNA library construction method described here (Chapter III) was born out of a desire to design a library construction method that minimized bias while allowing us complete control over every step in the process. The Moore lab was beginning to focus heavily on deep sequencing experiments and had grown frustrated with the meager information about sample manipulations provided by commercial library preparation kits. Therefore, we set out to design our own method for preparing libraries from all types of RNAs, thus the internal name "OmniPrep." We minimized the bias introduced at each construction step through careful enzymatic choice and titration experiments to find optimal conditions for each enzyme. In addition to measuring the bias of our optimized procedure, we were able to demonstrate that our method introduced less bias (see Figure 3.10) than a previously published bias-minimizing protocol. However, as with any biologically-related process, there are still aspects of this method that could be improved, which will be discussed below.

Optimizing Removal of Unextended RT Primer

One methodological step to be further optimized is the isolation of full-length RT product away from unextended primer, as the latter will circularize and contaminate PCR amplification. Currently, this step utilizes gel excision and elution following size-selection on a denaturing-PAGE gel. However, this isolation step is imperfect and results in carryover of unextended RT primer, which can be more abundant than the desired RT product for very low RNA input amounts.

Our method was published back-to-back with a method (Sterling et al., 2015) developed by Catherine Sterling - a postdoc in Victor Ambros' lab - as we had collaborated after discovering our that our work overlapped. Catherine's focus was slightly different, as her project depended on constructing miRNA libraries from minute amounts of RNA. To increase RT product capture sensitivity, she incorporated biotinylated dNTPs during reverse transcription and recovered her product via incubation with streptavidin beads. Catherine was not able to detect any bias introduced by use of biotinylated dNTPs, likely because the small biotin molecule did not interfere with reverse transcription. Incorporating this approach into OmniPrep could negate the gel purification of RT product, but would only be an improvement if the ratio of extended RT product:unextended RT primer was greater than with OmniPrep. Currently, Catherine gel purifies the circularized RT product because of unextended RT primer contamination post-streptavidin pull-down (personal communication, Catherine Sterling). Therefore, in order to maximize the potential of biotinylated dNTP incorporation for RT product isolation, the amount of contaminating unextended RT primer would need to be significantly reduced.

Another approach to isolating full-length RT product could utilize size-exclusion chromatography. This chromatographic method separates molecules by size, where larger molecules elute first due to their exclusion from pores in the column material, and has many advantages over gel purification. First, there should be greater recovery of sample material compared to gel purification (Kurien and Scofield, 2002). Second, molecules of a similar size tend to fractionate in narrow bands, enabling good sensitivity and minimal contamination with any similarly sized RT by-products. Typically, size exclusion chromatography requires a 10% difference in size to achieve good resolution, though single-residue resolution can be achieved (Ellington and Pollard, 2001). This 10% size difference would require the addition of $\geq$10 nts during reverse transcription; as read length >10 nts is needed to uniquely map to the genome, this would not introduce any extra limitations to OmniPrep. Third, the RT product would stay in solution the whole time. Finally, this type of chromatography can be automated, which would reduce the amount of hands-on time that OmniPrep currently requires. One potential drawback is the lack of denaturation, as secondary structure can alter RT product elution time. The extent of this effect could vary depending upon the cDNA sequence in the RT product, which could alter the relative abundances within the sample. In summary, size-exclusion chromatography seems very promising as a method of purifying full-length RT product, but some optimization may be required.

Approaching the problem a different way, RT product isolation would be unnecessary if the unextended RT primer could be specifically removed after reverse-transcription. The Preiss group has published a protocol using Exonuclease I (ExoI) enzyme immediately after reverse transcription to degrade unextended RT primer (Archer et al., 2014). This enzyme specifically degrades single-stranded DNA, so this approach requires that the newly-synthesized RT product remain hybridized to its RNA template. I briefly tested ExoI treatment with OmniPrep and detected a slight reduction in the amount of full-length RT product. However, these experiments were far from exhaustive and may warrant revisiting.

Addition of Randomer Sequences to Detect PCR Jackpots

When optimizing the PCR step, we specifically chose an enzyme (KAPA HiFi) which introduces minimal bias. The use of KAPA HiFi will reduce the chances of PCR jackpots - individual sequences which amplify much more efficiently than the overall sample - as this enzyme has little preference for nucleotide content. However, we could improve OmniPrep by adding randomer DNA tags which enable the tracking of PCR jackpots (Kivioja et al., 2011). These sequence tags can be added to either the RNA, cDNA, or dsDNA prior to PCR amplification. The goal is to label each molecule with a unique identifying sequence so the amount of PCR amplification can be measured. Post-PCR amplification, if a read appears multiple times with the same randomer tag sequence, all reads likely originated from the same molecule and the read count can be adjusted. However, if a read appears multiple times with different randomer tag sequences, those reads likely originated from individual capture events.

Error Correcting Barcodes

As discussed in Chapter III, the barcodes at the 5' end of the PCR-amplified cDNA allow for pooled samples of multiple libraries to be separated into their component libraries after sequencing. The barcode sequences used in OmniPrep are based on barcodes used by Illumina in their kits and were chosen to be base-balanced at each position across the barcode. At the time of development, sequence read length was limiting, so the length of this in-line barcode was limited to 5 nts to avoid negatively impacting the potential mappability of the cDNA fragment. However, the number of libraries that can be identified by a 5-nt barcode with a high degree of confidence - while maintaining base balance - is low. Current high-quality read lengths have increased to the point where a longer in-line barcode can be used without compromising the amount of information obtained from the sample. The number of reads per lane has increased along with read length; therefore, it is desirable to combine more samples into each sequencing lane. In order to maintain a high level of confidence when de-multiplexing these pooled libraries, the length and complexity of the barcode must increase as well.

Based on combinatorics, a simple increase in the length of the barcode will increase the number of potential barcodes. However, it is important that the barcodes can be identified even when they contain sequencing errors. A simple solution is to choose barcodes that cannot be converted into another barcode even with multiple sequencing errors. A more elegant and comprehensive approach is to utilize error-correcting barcodes, which enable the possibility of detecting and correcting sequencing errors. The most common error-correcting barcodes are based on either Hamming codes (Bystrykh, 2012, Hamady et al., 2008) or Levenshtein codes (Buschmann and Bystrykh, 2013). The former is a binary code made of data bits that are interrupted by parity bits, which are used for a checksum function to identify substitution errors (Bystrykh, 2012). As substitution errors are the most common errors produced by Illumina sequencing machines (Nakamura et al., 2011), Hamming codes have become a popular source of sequencing barcodes. However, one limitation of a Hamming code approach is the inability to detect insertions and deletions in the linear sequence code. This can be overcome by utilizing barcodes based on Levenshtein codes, which are capable of correcting substitution, insertion and deletion errors (Buschmann and Bystrykh, 2013). Either of these types of barcodes could easily be incorporated into OmniPrep, depending on the needs of the user and the sequencing platform being used. A simple substitution of error-correcting barcodes for the current barcodes in the RT primers would result in a set of libraries that could be parsed into their component samples with extremely high confidence.

Paired-end Sequencing Compatibility

The current OmniPrep protocol was designed with the sequences of Illumina's paired-end adaptors. These adaptor sequences, as the name suggests, enable an OmniPrep-constructed library to undergo paired-end sequencing. This type of sequencing is not ideal, however, for two reasons. First, the 5' 21 nucleotides of the reverse read will sequence the ligation adaptor, wasting sequencing space. Second, the quality of the reverse read will be low, because the Illumina machines interpret incorporation of the same nucleotide across the entire slide as an indication of error. In order for OmniPrep to be fully compatible with paired-end sequencing, the adaptor sequences need to change. The ideal adaptor sequences are Illumina's multiplexing adaptors, as they work on all Illumina platforms. Additionally, they have built-in barcodes which are read in a separate sequencing reaction.

Initial experiments utilizing multiplex adaptor sequences resulted in multiple RT product bands. Due to the degeneracy of these adaptor sequences, the RT primer is able to self-prime during reverse transcription to produce spurious products. In an attempt to minimize this effect, the length of the 3' ligation adaptor was extended to 29 nts and the RT primer shortened so it no longer anneals adjacent to the RNA-DNA ligation site, but 7 nucleotides further back into the 3' adaptor. These changes should minimize RT primer self-hybridization, but extensive experimentation will be performed to ensure that these changes do not alter the efficiency of OmniPrep.

Measuring the Lower Limits of OmniPrep

One main goal of OmniPrep was to create a method which could accept a wide range of RNA input amounts, as the Moore lab prepares RNA sequencing libraries from a wide variety of samples. We tested the lower limits of our method and detected a linear relationship between RNA input and RT product signal across a wide range of input concentrations(3.3 nM to 133 nM input; see Figure 3.4F). However, this only reports on RNA capture and RT product synthesis for a fixed RNA input amount. To measure capture efficiency across a wide range of RNA expression levels within a single sample, however, would be a more sophisticated experiment to truly measure of the robustness of OmniPrep. With this type of RNA input, we could compare the capture efficiencies of abundant RNAs to those of lowly expressed RNAs and test the reproducibility of capture at the lower end of the expression range by comparing signal across experimental replicates.

Perspective

Undoubtedly, further enzymatic and technological developments will reduce the already small bias that results from OmniPrep. Given that most biases originate from RNA library construction, the development of a direct RNA sequencing platform requiring no amplification would enable complete and bias-free profiling of the entire transcriptome. Though significant technological developments are necessary, it is possible that nanopore sequencing will help to realize this dream (Ozsolak and Milos, 2011).

OmniPrep represents a significant improvement over other library generation methods because it introduces minimal bias and works for all types of RNA. This bias can have large affects on the sequence populations of small RNAs and RNA footprint libraries, and the direct abundance comparison of 2 species within the same library should be avoided. However, as these biases are systematic, comparing abundance counts between similarly-prepared libraries should accurately reflect the original sample differences.

The converse of this statement is also true: libraries prepared with different construction methods will have different biases, and comparing abundance counts between differently-prepared libraries may be inaccurate. Though important, this point is often overlooked, especially when quantifying the abundance of a unique RBP or RNP footprint. To determine if a specific footprint is enriched in a sample, the data should be normalized to RNA-Seq data made using (a) RNA fragments roughly the same length as the footprints (Jackson and Standart, 2015) and (b) the same library preparation method. Then, only by comparing the footprinting libraries to the RNA-Seq libraries is it possible to determine if a specific RNA sequence is enriched in the footprinting libraries or simply preferentially captured and amplified during library construction. If significantly longer RNA fragments are used to generate the control RNA-Seq libraries, the effects of local nucleotide content on library bias will be minimized, and the libraries will not be comparable. Unfortunately, producing footprint-length fragments for RNA-Seq libraries often results in tRNA contamination. As most footprints are shorter than full-length tRNA, fragmenting RNA to these shorter lengths will also generate tRNA fragments. In my case, preparing RNA-Seq libraries to complement ribosome profiling data resulted in a significant amount of tRNA contamination (most of the contaminating ncRNA sequences in the footprinting libraries were rRNA, while in the RNA-Seq libraries more were tRNA; see Table 2.2).

Despite the abundance of publications demonstrating bias in RNA deep sequencing library preparation, these issues are largely ignored by the field. Many companies are producing kits for library construction, which can increase throughput and reduce the challenges of library preparation facing a novice user. However, development of these kits has focused on decreasing the total time of library preparation. Therefore, each enzymatic reaction is relatively short and likely incomplete. It is up to the conscientious scientist to be aware of the limitations of library preparation methods and limit data analyses to those justified by the experimental system.

Bibliography

Abaza, I. and Gebauer, F. (2008). Trading translation with RNA-binding proteins. *RNA*, 14(3):404–409.

Aeschimann, F., Xiong, J., Arnold, A., Dieterich, C., and Großhans, H. (2015). Transcriptome-wide measurement of ribosomal occupancy by ribosome profiling. *Methods (San Diego, Calif.)*.

Agirrezabala, X., Lei, J., Brunelle, J. L., Ortiz-Meoz, R. F., Green, R., and Frank, J. (2008). Visualization of the hybrid state of tRNA binding promoted by spontaneous ratcheting of the ribosome. *Molecular Cell*, 32(2):190–197.

Akopian, D., Shen, K., Zhang, X., and Shan, S.-o. (2013). Signal recognition particle: an essential protein-targeting machine. *Annual Review of Biochemistry*, 82:693–721.

Alon, S., Vigneault, F., Eminaga, S., Christodoulou, D. C., Seidman, J. G., Church, G. M., and Eisenberg, E. (2011). Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Research*, 21(9):1506–1511.

Amrani, N., Ganesan, R., Kervestin, S., Mangus, D. A., Ghosh, S., and Jacobson, A. (2004). A faux 3'-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. *Nature*, 432(7013):112–118.

Anders, S., Pyl, P. T., and Huber, W. (2014). HTSeq–A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169.

Anderson, D. M., Anderson, K. M., Chang, C.-L., Makarewich, C. A., Nelson, B. R., et al. (2015). A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell*, 160(4):595–606.

Arava, Y., Boas, F. E., Brown, P. O., and Herschlag, D. (2005). Dissecting eukaryotic translation and its control by ribosome density mapping. *Nucleic Acids Research*, 33(8):2421–2432.

Arava, Y., Wang, Y., Storey, J. D., Liu, C. L., Brown, P. O., and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7):3889–3894.

Archer, S. K., Shirokikh, N. E., and Preiss, T. (2014). Selective and flexible depletion of problematic sequences from RNA-seq libraries at the cDNA stage. *BMC Genomics*, 15:401.

Armour, C. D., Castle, J. C., Chen, R., Babak, T., Loerch, P., et al. (2009). Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nature Methods*, 6(9):647–649.

Arribere, J. A. and Gilbert, W. V. (2013). Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Research*, 23(6):977–987.

Artieri, C. G. and Fraser, H. B. (2014). Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Research*, 24(12):2011–2021.

Aspden, J. L., Eyre-Walker, Y. C., Phillips, R. J., Amin, U., Mumtaz, M. A. S., Brocard, M., and Couso, J.-P. (2014). Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife*, 3:e03528.

Baer, C., Claus, R., and Plass, C. (2013). Genome-wide epigenetic regulation of miRNAs in cancer. *Cancer Research*, 73(2):473–477.

Baker, K. E. and Parker, R. (2004). Nonsense-mediated mRNA decay: terminating erroneous gene expression. *Current Opinion in Cell Biology*, 16(3):293–299.

Barbosa, C., Peixeiro, I., and Romão, L. (2013). Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. *PLoS Genetics*, 9(8):e1003529.

Bartel, D. P. and Szostak, J. W. (1993). Isolation of new ribozymes from a large pool of random sequences. *Science*, 261(5127):1411–1418.

Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., et al. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal*, 33(9):981–993.

Benes, V., Blake, J., and Doyle, K. (2011). Ribo-Zero Gold Kit: improved RNA-seq results after removal of cytoplasmic and mitochondrial ribosomal RNA. *Nature Methods*, pages iii–iv.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59.

Bhattacharya, A., McIntosh, K. B., Willis, I. M., and Warner, J. R. (2010). Why Dom34 stimulates growth of cells with defects of 40S ribosomal subunit biosynthesis. *Molecular and Cellular Biology*, 30(23):5562–5571.

Bicknell, A. A., Cenik, C., Chua, H. N., Roth, F. P., and Moore, M. J. (2012). Introns in UTRs: why we should stop ignoring them. *BioEssays*, 34(12):1025–1034.

Bissels, U., Wild, S., Tomiuk, S., Holste, A., Hafner, M., Tuschl, T., and Bosio, A. (2009). Absolute quantification of microRNAs by using a universal reference. *RNA (New York, N.Y.)*, 15(12):2375–2384.

Bonafoux, B., Lejeune, M., Piquemal, D., Quere, R., Baudet, A., Assaf, L., Marti, J., Aguilar-Martinez, P., and Commes, T. (2004). Analysis of remnant reticulocyte mRNA reveals new genes and antisense transcripts expressed in the human erythroid lineage. *Haematologica*, 89(12):1434–1438.

Braat, A. K., Yan, N., Arn, E., Harrison, D., and Macdonald, P. M. (2004). Localization-dependent oskar protein accumulation; control after the initiation of translation. *Developmental Cell*, 7(1):125–131.

Brar, G. a., Yassour, M., Friedman, N., Regev, A., Ingolia, N. T., and Weissman, J. S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science*, 335(6068):552–557.

Breese, M. R. and Liu, Y. (2013). NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics (Oxford, England)*, 29(4):494–496.

Brenner, S., Jacob, F., and Meselson, M. (1961). An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. *Nature*, 190(4776):576–581.

Britten, R. J. and Roberts, R. B. (1960). High-Resolution Density Gradient Sedimentation Analysis. *Science*, 131(3392):32–33.

Brown, A., Shao, S., Murray, J., Hegde, R. S., and Ramakrishnan, V. (2015). Structural basis for stop codon recognition in eukaryotes. *Nature*, 524(7566):493–496.

Budkevich, T. V., Giesebrecht, J., Behrmann, E., Loerke, J., Ramrath, D. J. F., et al. (2014). Regulation of the mammalian elongation cycle by subunit rolling: a eukaryotic-specific ribosome rearrangement. *Cell*, 158(1):121–131.

Buschmann, T. and Bystrykh, L. V. (2013). Levenshtein error-correcting barcodes for multiplexed DNA sequencing. *BMC Bioinformatics*, 14:272.

Bystrykh, L. V. (2012). Generalized DNA barcode design based on Hamming codes. *PloS One*, 7(5):e36852.

Cajigas, I. J., Tushev, G., Will, T. J., tom Dieck, S., Fuerst, N., and Schuman, E. M. (2012). The local transcriptome in the synaptic neuropil revealed by deep sequencing and high-resolution imaging. *Neuron*, 74(3):453–466.

Calvo, S. E., Pagliarini, D. J., and Mootha, V. K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proceedings of the National Academy of Sciences*, 106(18):7507–7512.

Cannon, M., Krug, R., and Gilbert, W. (1963). The Binding of S-RNA by *Escherichia coli* Ribosomes. *Journal of Molecular Biology*, 7:360–378.

Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J. M., and Pascual-Montano, A. (2007). GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biology*, 8(1):R3.

Chang, J. C. and Kan, Y. W. (1979). beta 0 thalassemia, a nonsense mutation in man. *Proceedings of the National Academy of Sciences of the United States of America*, 76(6):2886–2889.

Chang, Y.-F., Imam, J. S., and Wilkinson, M. F. (2007). The nonsense-mediated decay RNA surveillance pathway. *Annual Review of Biochemistry*, 76:51–74.

Chen, D. and Patton, J. T. (2001). Reverse transcriptase adds nontemplated nucleotides to cDNAs during 5'-RACE and primer extension. *BioTechniques*, 30(3):574–582.

Chen, W., Shulha, H. P., Ashar-Patel, A., Yan, J., Green, K. M., Query, C. C., Rhind, N., Weng, Z., and Moore, M. J. (2014). Endogenous U2· U5· U6 snRNA complexes in S. pombe are intron lariat spliceosomes. *RNA*, 20:1–13.

Chu, Q., Ma, J., and Saghatelian, A. (2015). Identification and characterization of sORF-encoded polypeptides. *Critical Reviews in Biochemistry and Molecular Biology*, pages 1–8.

Ciandrini, L., Stansfield, I., and Romano, M. C. (2013). Ribosome traffic on mRNAs maps to gene ontology: genome-wide quantification of translation initiation rates and polysome size regulation. *PLoS Computational Biology*, 9(1):e1002866.

Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, 5(7):613–619.

Cosgrove, J. W., Heikkila, J. J., and Brown, I. R. (1982). Translation of mRNA associated with monosomes and residual polysomes following disaggregation of brain polysomes by LSD and hyperthermia. *Neurochemical Research*, 7(4):505–518.

Crappé, J., Van Criekinge, W., and Menschaert, G. (2014). Little things make big things happen: A summary of micropeptide encoding genes. *EuPA Open Proteomics*, 3:128–137.

Crappé, J., Van Criekinge, W., Trooskens, G., Hayakawa, E., Luyten, W., Baggerman, G., and Menschaert, G. (2013). Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics*, 14(1):648.

Crick, F. H. (1958). On protein synthesis. *Symp. Soc. Exp. Biol.*, pages 138–163.

Culbertson, M. R. and Neeno-Eckwall, E. (2005). Transcript selection and the recruitment of mRNA decay factors for NMD in Saccharomyces cerevisiae. *RNA*, 11(9):1333–1339.

Czaplinski, K. and Singer, R. H. (2006). Pathways for mRNA localization in the cytoplasm. *Trends in Biochemical Sciences*, 31(12):687–693.

Dever, T. E. and Green, R. (2012). The elongation, termination, and recycling phases of translation in eukaryotes. *Cold Spring Harbor Perspectives in Biology*, 4(7):a013706.

Eguen, T., Straub, D., Graeff, M., and Wenkel, S. (2015). MicroProteins: small size–big impact. *Trends in Plant Science*.

Elbashir, S. M., Lendeckel, W., and Tuschl, T. (2001). RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes & Development*, 15(2):188–200.

Ellington, A. and Pollard, J. D. (2001). Synthesis and purification of oligonucleotides. *Current Protocols in Molecular Biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 2:Unit2.11.

Eminaga, S., Christodoulou, D. C., Vigneault, F., Church, G. M., and Seidman, J. G. (2013). Quantification of microRNA Expression with Next-Generation Sequencing. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 4:Unit4.17.

Epicentre Technologies Corporation (2012). ARTseq Ribosome Profiling Kit Manual.

Favaudon, V. and Pochon, F. (1976). Magnesium dependence of the association kinetics of Escherichia coli ribosomal subunits. *Biochemistry*, 15(18):3903–3912.

Fersht, A. (1985). *Enzyme Structure and Mechanism*. WH Freeman & Co., New York, 2nd edition edition.

Fiechter, V., Cameroni, E., Cerutti, L., De Virgilio, C., Barral, Y., and Fankhauser, C. (2008). The evolutionary conserved BER1 gene is involved in microtubule stability in yeast. *Current Genetics*, 53(2):107–115.

Frank, J. and Agrawal, R. K. (2000). A ratchet-like inter-subunit reorganization of the ribosome during translocation. *Nature*, 406(6793):318–322.

Fraser, C. S. (2015). Quantitative studies of mRNA recruitment to the eukaryotic ribosome. *Biochimie*, 114:58–71.

Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K., et al. (2012). Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Research*, 22(11):2208–2218.

Fuchs, R. T., Sun, Z., Zhuang, F., and Robb, G. B. (2015). Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. *PloS One*, 10(5):e0126049.

Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A., and Couso, J.-P. (2007). Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLOS Biology*, 5(5):e106.

Gao, Q., Das, B., Sherman, F., and Maquat, L. E. (2005). Cap-binding protein 1-mediated and eukaryotic translation initiation factor 4E-mediated pioneer rounds of translation in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12):4258–4263.

Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., and Futcher, B. (2014). Measurement of average decoding rates of the 61 sense codons in vivo. *eLife*, 3:e03735.

Gebauer, F. and Hentze, M. W. (2004). Molecular mechanisms of translational control. *Nature Reviews Molecular Cell Biology*, 5(10):827–835.

Gerard, G. F., Fox, D. K., Nathan, M., and D'Alessio, J. M. (1997). Reverse transcriptase. The use of cloned Moloney murine leukemia virus reverse transcriptase to synthesize DNA from RNA. *Molecular Biotechnology*, 8(1):61–77.

Gerashchenko, M. V. and Gladyshev, V. N. (2014). Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Research*, 42(17):e134.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K., and Weissman, J. S. (2003). Global analysis of protein expression in yeast. *Nature*, 425:737–741.

Gierer, A. (1963). Function of aggregated reticulocyte ribosomes in protein synthesis. *Journal of Molecular Biology*, 6:148–157.

Giorgi, C., Yeo, G. W., Stone, M. E., Katz, D. B., Burge, C., Turrigiano, G., and Moore, M. J. (2007). The EJC factor eIF4AIII modulates synaptic strength and neuronal protein expression. *Cell*, 130(1):179–191.

Gould, P. S., Bird, H., and Easton, A. J. (2005). Translation toeprinting assays using fluorescently labeled primers and capillary electrophoresis. *BioTechniques*, 38(3):397–400.

Green, R. E., Malaspinas, A.-S., Krause, J., Briggs, A. W., Johnson, P. L. F., et al. (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, 134(3):416–426.

Gros, F., Hiatt, H., Gilbert, W., Kurland, C. G., Risebrough, R. W., and Watson, J. D. (1961). Unstable ribonucleic acid revealed by pulse labelling of Escherichia coli. *Nature*, 190:581–585.

Guan, Q., Zheng, W., Tang, S., Liu, X., Zinkel, R. A., Tsui, K.-W., Yandell, B. S., and Culbertson, M. R. (2006). Impact of nonsense-mediated mRNA decay on the global expression profile of budding yeast. *PLoS Genetics*, 2(11):e203.

Guydosh, N. R. and Green, R. (2014). Dom34 rescues ribosomes in 3' untranslated regions. *Cell*, 156(5):950–962.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., et al. (2010). Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, 141(1):129–141.

Hafner, M., Renwick, N., Brown, M., Mihailovic, A., Holoch, D., et al. (2011). RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA (New York, N.Y.)*, 17(9):1697–1712.

Hafner, M., Renwick, N., Farazi, T. A., Mihailovic, A., Pena, J. T. G., and Tuschl, T. (2012). Barcoded cDNA library preparation for small RNA profiling by next-generation sequencing. *Methods (San Diego, Calif.)*, 58(2):164–170.

Hamady, M., Walker, J. J., Harris, J. K., Gold, N. J., and Knight, R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nature Methods*, 5(3):235–237.

Han, Y., Gao, X., Liu, B., Wan, J., Zhang, X., and Qian, S.-B. (2014). Ribosome profiling reveals sequence-independent post-initiation pausing as a signature of translation. *Cell Research*, 24(7):842–851.

Hansen, K. D., Brenner, S. E., and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research*, 38(12):e131.

Harrison, P. M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. (2002). A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Research*, 30(5):1083–1090.

Hartz, D., McPheeters, D. S., Traut, R., and Gold, L. (1988). Extension inhibition analysis of translation initiation complexes. *Methods in Enzymology*, 164:419–425.

He, F., Li, X., Spatrick, P., Casillo, R., Dong, S., and Jacobson, A. (2003). Genome-wide analysis of mRNAs regulated by the nonsense-mediated and 5' to 3' mRNA decay pathways in yeast. *Molecular Cell*, 12(6):1439–1452.

He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N., and Kinzler, K. W. (2008). The antisense transcriptomes of human cells. *Science*, 322(5909):1855–1857.

Heyer, E. E., Özadam, H., Ricci, E. P., Cenik, C., and Moore, M. J. (2015). An optimized kit-free method for making strand-specific deep sequencing libraries from RNA fragments. *Nucleic Acids Research*, 43(1):e2.

Hinnebusch, A. G. (2005). Translational regulation of GCN4 and the general amino acid control of yeast. *Annual Review of Microbiology*, 59:407–450.

Hinnebusch, A. G. (2014). The scanning mechanism of eukaryotic translation initiation. *Annual Review of Biochemistry*, 83:779–812.

Hinnebusch, A. G. and Lorsch, J. R. (2012). The mechanism of eukaryotic translation initiation: new insights and challenges. *Cold Spring Harbor Perspectives in Biology*, 4(10).

Hoagland, M. B., Stephenson, M. L., Scott, J. F., Hecht, L. I., and Zamecnik, P. C. (1958). A soluble ribonucleic acid intermediate in protein synthesis. *Journal of Biological Chemistry*, 231(1):241–257.

Holt, C. E. and Schuman, E. M. (2013). The central dogma decentralized: new perspectives on RNA function and local translation in neurons. *Neuron*, 80(3):648–657.

Hu, W., Petzold, C., Coller, J., and Baker, K. E. (2010). Nonsense-mediated mRNA decapping occurs on polyribosomes in Saccharomyces cerevisiae. *Nature Structural & Molecular Biology*, 17(2):244–247.

Illumina (2013). Technical Note: Using a PhiX Control for HiSeq Sequencing Runs.

Ingolia, N. T. (2010). Genome-wide translational profiling by ribosome footprinting. *Methods in Enzymology*, 470:119–142.

Ingolia, N. T., Brar, G. a., Rouskin, S., McGeachy, A. M., and Weissman, J. S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature Protocols*, 7(8):1534–1550.

Ingolia, N. T., Brar, G. a., Rouskin, S., McGeachy, A. M., and Weissman, J. S. (2013). Genome-wide annotation and quantitation of translation by ribosome profiling. *Current Protocols in Molecular Biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 4:Unit4.18.

Ingolia, N. T., Brar, G. a., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E., Wills, M. R., and Weissman, J. S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell reports*, 8(5):1365–1379.

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

Ingolia, N. T., Lareau, L. F., and Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802.

Irier, H. A., Shaw, R., Lau, A., Feng, Y., and Dingledine, R. (2009). Translational regulation of GluR2 mRNAs in rat hippocampus by alternative 3' untranslated regions. *Journal of Neurochemistry*, 109(2):584–594.

Ishigaki, Y., Li, X., Serin, G., and Maquat, L. E. (2001). Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20. *Cell*, 106(5):607–617.

Jackson, R. and Standart, N. (2015). The awesome power of ribosome profiling. *RNA (New York, N.Y.)*, 21(4):652–654.

Jackson, R. J., Hellen, C. U. T., and Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular Cell Biology*, 11(2):113–127.

Jagannathan, S., Reid, D. W., Cox, A. H., and Nicchitta, C. V. (2014). De novo translation initiation on membrane-bound ribosomes as a mechanism for localization of cytosolic protein mRNAs to the endoplasmic reticulum. *RNA*, 20(10):1489–1498.

Jan, C. H., Williams, C. C., and Weissman, J. S. (2014). Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science*, 346(6210):1257521.

Jayaprakash, A. D., Jabado, O., Brown, B. D., and Sachidanandam, R. (2011). Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic Acids Research*, 39(21):e141.

Johansson, M. J. O., He, F., Spatrick, P., Li, C., and Jacobson, A. (2007). Association of yeast Upf1p with direct substrates of the NMD pathway. *Proceedings of the National Academy of Sciences*, 104(52):20872–20877.

Jung, H., O'Hare, C. M., and Holt, C. E. (2011). Translational regulation in growth cones. *Current Opinion in Genetics & Development*, 21(4):458–464.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(Database issue):D493–6.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., et al. (2005). Antisense transcription in the mammalian transcriptome. *Science*, 309(5740):1564–1566.

Kemp, G. and Cymer, F. (2014). Small membrane proteins - elucidating the function of the needle in the haystack. *Biological Chemistry*, 395(12):1365–1377.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, 12(6):996–1006.

Kervestin, S. and Jacobson, A. (2012). NMD: a multifaceted response to premature translational termination. *Nature Reviews Molecular Cell Biology*, 13(11):700–712.

Kirsch, J. F., Siekevitz, P., and Palade, G. E. (1960). Amino acid incorporation in vitro by ribonucleoprotein particles detached from guinea pig liver microsomes. *Journal of Biological Chemistry*, 235:1419–1424.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72–74.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology*, 17(7):909–915.

Kozak, M. (1998). Primer extension analysis of eukaryotic ribosome-mRNA complexes. *Nucleic Acids Research*, 26(21):4853–4859.

Krishnan, K., Ren, Z., Losada, L., Nierman, W. C., Lu, L. J., and Askew, D. S. (2014). Polysome profiling reveals broad translatome remodeling during endoplasmic reticulum (ER) stress in the pathogenic fungus Aspergillus fumigatus. *BMC Genomics*, 15:159.

Kurien, B. T. and Scofield, R. H. (2002). Extraction of nucleic acid fragments from gels. *Analytical Biochemistry*, 302(1):1–9.

Kwok, C. K., Ding, Y., Sherlock, M. E., Assmann, S. M., and Bevilacqua, P. C. (2013). A hybridization-based approach for quantitative and low-bias single-stranded DNA ligation. *Analytical Biochemistry*, 435(2):181–186.

Landry, C. R., Zhong, X., Nielly-Thibault, L., and Roucou, X. (2015). Found in translation: functions and evolution of a recently discovered alternative proteome. *Current Opinion in Structural Biology*, 32:74–80.

Langevin, S. A., Bent, Z. W., Solberg, O. D., Curtis, D. J., Lane, P. D., Williams, K. P., Schoeniger, J. S., Sinha, A., Lane, T. W., and Branda, S. S. (2013). Peregrine: A rapid and unbiased method to produce strand-specific RNA-Seq libraries from small quantities of starting material. *RNA Biology*, 10(4):502–515.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25.

Lareau, L. F., Hite, D. H., Hogan, G. J., and Brown, P. (2014). Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *eLife*.

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294(5543):858–862.

Lauressergues, D., Couzigou, J.-M., Clemente, H. S., Martinez, Y., Dunand, C., Bécard, G., and Combier, J.-P. (2015). Primary transcripts of microRNAs encode regulatory peptides. *Nature*.

Lécuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T. R., Tomancak, P., and Krause, H. M. (2007). Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, 131(1):174–187.

Lee, R. C. and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294(5543):862–864.

Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B., and Qian, S.-B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*, 109(37):E2424–32.

Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., Gnirke, A., and Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods*, 7(9):709–715.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools.

Bioinformatics, 25(16):2078–2079.

Li, S., Dong, X., and Su, Z. (2013). Directional RNA-seq reveals highly complex condition-dependent transcriptomes in E. coli K12 through accurate full-length transcripts assembling. *BMC Genomics*, 14:520.

Linsen, S. E. V., de Wit, E., Janssens, G., Heater, S., Chapman, L., et al. (2009). Limitations and possibilities of small RNA digital gene expression profiling. *Nature Methods*, 6(7):474–476.

Llácer, J. L., Hussain, T., Marler, L., Aitken, C. E., Thakur, A., Lorsch, J. R., Hinnebusch, A. G., and Ramakrishnan, V. (2015). Conformational Differences between Open and Closed States of the Eukaryotic Translation Initiation Complex. *Molecular Cell*, 59(3):399–412.

Lorenz, R. and Bernhart, S. (2011). ViennaRNA Package 2.0. *. . . for Molecular Biology*.

Losson, R. and Lacroute, F. (1979). Interference of nonsense mutations with eukaryotic messenger RNA stability. *Proceedings of the National Academy of Sciences of the United States of America*, 76(10):5134–5137.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., et al. (2005). MicroRNA expression profiles classify human cancers. *Nature*, 435(7043):834–838.

Lui, W.-O., Pourmand, N., Patterson, B. K., and Fire, A. (2007). Patterns of known and novel small RNAs in human cervical cancer. *Cancer Research*, 67(13):6031–6043.

Madabhushi, R., Gao, F., Pfenning, A. R., Pan, L., Yamakawa, S., et al. (2015). Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes. *Cell*, 161(7):1592–1605.

Maderazo, A. B., Belk, J. P., He, F., and Jacobson, A. (2003). Nonsense-containing mRNAs that accumulate in the absence of a functional nonsense-mediated mRNA decay pathway are destabilized rapidly upon its restitution. *Molecular and Cellular Biology*, 23:842–851.

Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS Letters*, 583(24):3966–3973.

Mamanova, L. and Turner, D. J. (2011). Low-bias, strand-specific transcriptome Illumina sequencing by on-flowcell reverse transcription (FRT-seq). *Nature Protocols*, 6(11):1736–1747.

Maquat, L. E., Kinniburgh, A. J., Rachmilewitz, E. A., and Ross, J. (1981). Unstable β-globin mRNA in mRNA-deficient β 0 thalassemia. *Cell*, 27:543–553.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., et al. (2005).

Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, 17(1):10–12.

Maxam, A. M. and GILBERT, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–564.

Miller, W., Drautz, D. I., Ratan, A., Pusey, B., Qi, J., et al. (2008). Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, 456(7220):387–390.

Moazed, D. and Noller, H. F. (1989). Intermediate states in the movement of transfer RNA in the ribosome. *Nature*, 342(6246):142–148.

Morin, R. D., O'Connor, M. D., Griffith, M., Kuchenbauer, F., Delaney, A., et al. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research*, 18(4):610–621.

Morris, D. R. and Geballe, A. P. (2000). Upstream Open Reading Frames as Regulators of mRNA Translation. *Molecular and Cellular Biology*, 20(23):8635–8642.

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

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

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., et al. (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 39(13):e90.

Nandakumar, J., Shuman, S., and Lima, C. D. (2006). RNA ligase structures reveal the basis for RNA specificity and conformational changes that drive ligation forward. *Cell*, 127(1):71–84.

Noeske, J. and Cate, J. H. D. (2012). Structural basis for protein synthesis: snapshots of the ribosome in motion. *Current Opinion in Structural Biology*, 22(6):743–749.

Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C., Yang, X., Tirado, F., Carazo, J. M., and Pascual-Montano, A. (2009). GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Research*, 37(Web Server issue):W317–22.

Noll, H. (2008). The discovery of polyribosomes. *BioEssays*, 30(11-12):1220–1234.

Noll, M., Hapke, B., Schreier, M. H., and Noll, H. (1973). Structural dynamics of bacterial ribosomes. I. Characterization of vacant couples and their relation to complexed ribosomes. *Journal of Molecular Biology*, 75(2):281–294.

Nottrott, S., Simard, M. J., and Richter, J. D. (2006). Human let-7a miRNA blocks protein production on actively translating polyribosomes. *Nature Structural & Molecular Biology*, 13(12):1108–1114.

Nürenberg, E. and Tampé, R. (2013). Tying up loose ends: ribosome recycling in eukaryotes and archaea. *Trends in Biochemical Sciences*, 38(2):64–74.

Oh, E., Becker, A. H., Sandikci, A., Huber, D., Chaba, R., et al. (2011). Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell*, 147(6):1295–1308.

Olsen, P. H. and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Developmental Biology*, 216(2):671–680.

Ostroff, L. E., Fiala, J. C., Allwardt, B., and Harris, K. M. (2002). Polyribosomes Redistribute from Dendritic Shafts into Spines with Enlarged Synapses during LTP in Developing Rat Hippocampal Slices. *Neuron*, 35(3):535–545.

Ozsolak, F. and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Publishing Group*, 12(2):87–98.

Palade, G. E. (1955). A small particulate component of the cytoplasm. *The Journal of Biophysical and Biochemical Cytology*, 1(1):59–68.

Pan, T. and Uhlenbeck, O. C. (1992). In vitro selection of RNAs that undergo autolytic cleavage with lead (2+). *Biochemistry*.

Paquin, N. and Chartrand, P. (2008). Local regulation of mRNA translation: new insights from the bud. *Trends in Cell Biology*, 18(3):105–111.

Parameswaran, P., Jalili, R., Tao, L., Shokralla, S., Gharizadeh, B., Ronaghi, M., and Fire, A. Z. (2007). A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Research*, 35(19):e130–e130.

Pauling, L. and Corey, R. B. (1953). A Proposed Structure For The Nucleic Acids. *Proceedings of the National Academy of Sciences of the United States of America*, 39(2):84–97.

Peccarelli, M. and Kebaara, B. W. (2014). Regulation of natural mRNAs by the nonsense-mediated mRNA decay pathway. *Eukaryotic Cell*, 13(9):1126–1135.

Pelechano, V., Chávez, S., and Pérez-Ortín, J. E. (2010). A complete set of nascent transcription rates for yeast genes. *PloS One*, 5(11):e15442.

Pelechano, V. and Steinmetz, L. M. (2013). Gene regulation by antisense transcription. *Nature Publishing Group*, 14(12):880–893.

Pelechano, V., Wei, W., and Steinmetz, L. M. (2013). Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, 497(7447):127–131.

Pelechano, V., Wei, W., and Steinmetz, L. M. (2015). Widespread Co-translational RNA Decay Reveals Ribosome Dynamics. *Cell*, 161(6):1400–1412.

Perry, R. B. and Fainzilber, M. (2014). Local translation in neuronal processes–in vivo tests of a "heretical hypothesis". *Developmental Neurobiology*, 74(3):210–217.

Petersen, C. P., Bordeleau, M.-E., Pelletier, J., and Sharp, P. A. (2006). Short RNAs repress translation after initiation in mammalian cells. *Molecular Cell*, 21(4):533–542.

Pfeffer, S., Lagos-Quintana, M., and Tuschl, T. (2005). Cloning of small RNA molecules. *Current Protocols in Molecular Biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 26:Unit 26.4.

Pisarev, A. V., Skabkin, M. A., Pisareva, V. P., Skabkina, O. V., Rakotondrafara, A. M., Hentze, M. W., Hellen, C. U. T., and Pestova, T. V. (2010). The role of ABCE1 in eukaryotic posttermination ribosomal recycling. *Molecular Cell*, 37(2):196–210.

Potter, M. D. and Nicchitta, C. V. (2002). Endoplasmic reticulum-bound ribosomes reside in stable association with the translocon following termination of protein synthesis. *Journal of Biological Chemistry*, 277(26):23314–23320.

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., et al. (2015). Codon Optimality Is a Major Determinant of mRNA Stability. *Cell*, 160(6):1111–1124.

Quail, M. A., Otto, T. D., Gu, Y., Harris, S. R., Skelly, T. F., McQuillan, J. A., Swerdlow, H. P., and Oyola, S. O. (2012). Optimal enzymes for amplifying sequencing libraries. *Nature Methods*, 9(1):10–11.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramamurthi, K. S. and Storz, G. (2014). The small protein floodgates are opening; now the functional analysis begins. *BMC Biology*, 12:96.

Ravasi, T., Suzuki, H., Pang, K. C., Katayama, S., Furuno, M., et al. (2006). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Research*, 16(1):11–19.

Reboll, M. R. and Nourbakhsh, M. (2014). Identification of Actively Translated mRNAs. In *link.springer.com*, pages 173–178. Methods in Molecular Biology, New York, NY.

Regalado, A. (2014). EmTech: Illumina Says 228,000 Human Genomes Will Be Sequenced in 2014. *MIT Technology Review*.

Reid, D. W. and Nicchitta, C. V. (2015). Diversity and selectivity in mRNA translation on the endoplasmic reticulum. *Nature Reviews Molecular Cell Biology*, 14(4):221–231.

Rheinberger, H.-J., Sternbach, H., and Nierhaus, K. H. (1981). Three tRNA binding sites on Escherichia coli ribosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 78(9):5310–5314.

Ricci, E. P., Kucukural, A., Cenik, C., Mercier, B. C., Singh, G., Heyer, E. E., Ashar-Patel, A., Peng, L., and Moore, M. J. (2014). Staufen1 senses overall transcript secondary structure to regulate translation. *Nature Structural & Molecular Biology*, 21(1):26–35.

Risebrough, R. W., Tissieres, A., and Watson, J. D. (1962). Messenger-RNA attachment to active ribosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 48:430–436.

Rosenfeld, N., Aharonov, R., Meiri, E., Rosenwald, S., Spector, Y., et al. (2008). MicroRNAs accurately identify cancer tissue origin. *Nature Biotechnology*, 26(4):462–469.

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., and Albà, M. M. (2014). Long non-coding RNAs as a source of new peptides. *eLife*, 3:e03523.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467.

Sayani, S., Janis, M., Lee, C. Y., Toesca, I., and Chanfreau, G. F. (2008). Widespread impact of nonsense-mediated mRNA decay on the yeast intronome. *Molecular Cell*, 31(3):360–370.

Seiser, R. M. and Nicchitta, C. V. (2000). The fate of membrane-bound ribosomes following the termination of protein synthesis. *Journal of Biological Chemistry*, 275(43):33820–33827.

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., and Plotkin, J. B. (2013). Rate-limiting steps in yeast protein translation. *Cell*, 153(7):1589–1601.

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*, 15:284.

Sheng, M. and Hoogenraad, C. C. (2007). The postsynaptic architecture of excitatory synapses: a more quantitative view. *Annual Review of Biochemistry*, 76:823–847.

Sheng, M. and Kim, E. (2011). The postsynaptic organization of synapses. *Cold Spring Harbor Perspectives in Biology*, 3(12).

Shirokikh, N. E., Alkalaeva, E. Z., Vassilenko, K. S., Afonina, Z. A., Alekhina, O. M., Kisselev, L. L., and Spirin, A. S. (2010). Quantitative analysis of ribosome-mRNA complexes at different translation stages. *Nucleic Acids Research*, 38(3):e15.

Singh, G., Kucukural, A., Cenik, C., Leszyk, J. D., Shaffer, S. A., Weng, Z., and Moore, M. J. (2012). The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. *Cell*, 151(4):750–764.

Singh, G., Ricci, E. P., and Moore, M. J. (2014). RIPiT-Seq: A high-throughput approach for footprinting RNA:protein complexes. *Methods (San Diego, Calif.)*, 65(3):320–332.

Siwiak, M. and Zielenkiewicz, P. (2010). A comprehensive, quantitative, and genome-wide model of translation. *PLoS Computational Biology*, 6(7):e1000865.

Skabkin, M. A., Skabkina, O. V., Hellen, C. U. T., and Pestova, T. V. (2013). Reinitiation and Other Unconventional Posttermination Events during Eukaryotic Translation. *Molecular Cell*, 51(2):249–264.

Slayter, H. S., Warner, J. R., Rich, A., and Hall, C. E. (1963). The Visualization of Polyribosomal Structure. *Journal of Molecular Biology*, 7:652–657.

Smith, J. E., Alvarez-Dominguez, J. R., Kline, N., Huynh, N. J., Geisler, S., Hu, W., Coller, J., and Baker, K. E. (2014). Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. *Cell Reports*, 7(6):1858–1866.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B., and Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679.

Sonenberg, N. and Hinnebusch, A. G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, 136(4):731–745.

Sorefan, K., Pais, H., Hall, A. E., Kozomara, A., Griffiths-Jones, S., Moulton, V., and Dalmay, T. (2012). Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence*, 3(1):4.

Staehelin, T., Wettstein, F. O., and Noll, H. (1963). Breakdown of rat-liver ergosomes in vivo after actinomycin inhibition of messenger RNA synthesis. *Science*, 140(3563):180–183.

Steitz, J. A. (1969). Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. *Nature*, 224(5223):957–964.

Sterling, C. H., Veksler-Lublinsky, I., and Ambros, V. (2015). An efficient and sensitive method for preparing cDNA libraries from scarce biological samples. *Nucleic Acids Research*, 43(1):e1.

Steward, O. and Schuman, E. M. (2003). Compartmentalized synthesis and degradation of proteins in neurons. *Neuron*, 40(2):347–359.

Storz, G., Wolf, Y. I., and Ramamurthi, K. S. (2014). Small proteins can no longer be ignored. *Annual Review of Biochemistry*, 83:753–777.

Sutton, M. A. and Schuman, E. M. (2006). Dendritic protein synthesis, synaptic plasticity, and memory. *Cell*, 127(1):49–58.

Tabas-Madrid, D., Nogales-Cadenas, R., and Pascual-Montano, A. (2012). GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acids Research*, 40(Web Server issue):W478–83.

Thomas, J. D. and Johannes, G. J. (2007). Identification of mRNAs that continue to associate with polysomes during hypoxia. *RNA*, 13(7):1116–1131.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2):344–354.

Twiss, J. L. and Fainzilber, M. (2009). Ribosomes in axons–scrounging from the neighbors? *Trends in Cell Biology*, 19(5):236–243.

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R. B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302(5648):1212–1215.

Van Der Kelen, K., Beyaert, R., Inzé, D., and De Veylder, L. (2009). Translational control of eukaryotic gene expression. *Critical Reviews in Biochemistry and Molecular Biology*, 44(4):143–168.

Van Nieuwerburgh, F., Soetaert, S., Podshivalova, K., Ay-Lin Wang, E., Schaffer, L., Deforce, D., Salomon, D. R., Head, S. R., and Ordoukhanian, P. (2011). Quantitative Bias in Illumina TruSeq and a Novel Post Amplification Barcoding Strategy for Multiplexed DNA and Small RNA Deep Sequencing. *PloS One*, 6(10):e26969.

Viollet, S., Fuchs, R. T., Munafo, D. B., Zhuang, F., and Robb, G. B. (2011). T4 RNA ligase 2 truncated active site mutants: improved tools for RNA analysis. *BMC Biotechnology*, 11:72.

Vivancos, A. P., Güell, M., Dohm, J. C., Serrano, L., and Himmelbauer, H. (2010). Strand-specific deep sequencing of the transcriptome. *Genome Research*, 20(7):989–999.

Vogel, C. and Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Publishing Group*, 13(4):227–232.

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

Voorhees, R. M. and Ramakrishnan, V. (2013). Structural basis of the translational elongation cycle. *Annual Review of Biochemistry*, 82:203–236.

Wang, D. O., Kim, S. M., Zhao, Y., Hwang, H., Miura, S. K., Sossin, W. S., and Martin, K. C. (2009). Synapse- and stimulus-specific local translation during long-term neuronal plasticity. *Science*, 324(5934):1536–1540.

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476.

Wang, X., Zhao, B. S., Roundtree, I. A., Lu, Z., Han, D., Ma, H., Weng, X., Chen, K., Shi, H., and He, C. (2015). N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell*, 161(6):1388–1399.

Warner, J., Knopf, P., and Rich, A. (1963). A multiple ribosomal structure in protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 49:122–129.

Warner, J. R. (1999). The economics of ribosome biosynthesis in yeast. *Trends in Biochemical Sciences*, 24(11):437–440.

Warner, J. R. and Knopf, P. M. (2002). The discovery of polyribosomes. *Trends in Biochemical Sciences*, 27(7):376–380.

Warner, J. R. and Rich, A. (1964). The number of soluble RNA molecules on reticulocyte polyribosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 51:1134–1141.

Watson, J. D. and Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738.

Wettstein, F. O. and Noll, H. (1965). Binding of Transfer Ribonucleic Acid to Ribosomes engaged in Protein Synthesis: Number and Properties of Ribosomal Binding Sites. *Journal of Molecular Biology*, 11:35–53.

Wettstein, F. O., Staehelin, T., and Noll, H. (1963). Ribosomal aggregate engaged in protein synthesis: characterization of the ergosome. *Nature*, 197:430–435.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

Wilke, C. O. (2015). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. R package version 0.4.0.

Williams, C. C., Jan, C. H., and Weissman, J. S. (2014). Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling. *Science*, 346(6210):748–751.

Xing, Y. and Lee, C. (2006). Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. *Nature Reviews Genetics*, 7(7):499–509.

Yu, X. and Warner, J. (2001). Expression of a micro-protein. *Journal of Biological Chemistry*, 276(36):33821–33825.

Zhang, P., Kang, J.-Y., Gou, L.-T., Wang, J., Xue, Y., et al. (2015). MIWI and piRNA-mediated cleavage of messenger RNAs in mouse testes. *Cell Research*, 25(2):193–207.

Zhang, Z., Lee, J. E., Riemondy, K., Anderson, E. M., and Yi, R. (2013). High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. *Genome Biology*, 14(10):R109.

Zhang, Z., Theurkauf, W. E., Weng, Z., and Zamore, P. D. (2012). Strand-specific libraries for high throughput RNA sequencing (RNA-Seq) prepared without poly(A) selection. *Silence*, 3(1):9.

Zhao, S., Zhang, Y., Gordon, W., Quan, J., Xi, H., Du, S., von Schack, D., and Zhang, B. (2015). Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genomics*, 16:675.

Zhuang, F., Fuchs, R. T., Sun, Z., Zheng, Y., and Robb, G. B. (2012). Structural bias in T4 RNA ligase-mediated 3'-adapter ligation. *Nucleic Acids Research*, 40(7):e54.

Zhulidov, P. A., Bogdanova, E. A., Shcheglov, A. S., Vagner, L. L., Khaspekov, G. L., et al. (2004). Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Research*, 32(3):e37.

Zivraj, K. H., Tung, Y. C. L., Piper, M., Gumy, L., Fawcett, J. W., Yeo, G. S. H., and Holt, C. E. (2010). Subcellular profiling reveals distinct and developmentally regulated repertoire of growth cone mRNAs. *The Journal of Neuroscience*, 30(46):15464–15478.