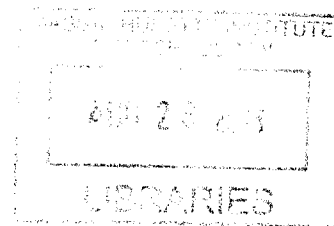


Transcript Leaders: Annotation and Insight into Functions in Translation

ARCHIVES

By
Joshua A. Arribere

B.A. Molecular and Cellular Biology
B.A. Applied Mathematics
The University of California at Berkeley, 2008



SUBMITTED TO THE DEPARTMENT OF BIOLOGY IN PARTIAL FUFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

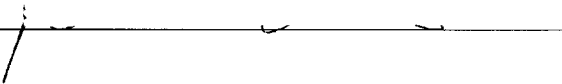
DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

July 2013

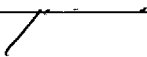
© 2013 Massachusetts Institute of Technology. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper
and electronic copies of this thesis document in whole or in part in any medium now
known or hereafter created

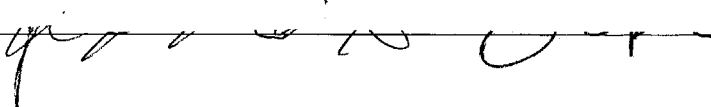
Signature of Author:


Department of Biology
August, 2013

Certified by:


Wendy V. Gilbert
Associate Professor of Biology
Thesis Supervisor

Accepted by:


Stephen Bell
Professor of Biology
Co-Chair, Graduate Committee

Transcript Leaders: Annotation and Insight into Functions in Translation

By

Joshua A. Arribere

Submitted to the Department of Biology on August 30, 2013 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Biology

Abstract

For a eukaryotic mRNA to be properly expressed, it undergoes a series of several steps, including transcription, modification, splicing, packaging, export, localization, translation, and decay. Of these steps transcription is the most extensively studied, though the remaining steps are also indispensable for proper protein production. While we understand many of these steps in biochemical detail *in vitro*, we have a much poorer knowledge of how they occur and are regulated for a given gene *in vivo*. Post-transcriptional regulation is carried out primarily through the noncoding portions of the mRNA: the Transcript Leader (TL or 5'UTR) upstream of the Open Reading Frame (ORF), and the 3'Untranslated Region (3'UTR) downstream. To understand how these regions affect post-transcriptional gene expression, it is critical to have precise annotations of the mRNA(s) produced from a gene.

In Chapter 2 I describe the development of Transcript Leader Sequencing (TL-seq), a technique to annotate TLs, and demonstrate its utility in yeast. TL-seq annotations reveal interesting TL-dependent regulation, including transcription within ORFs and short TLs that are associated with translation initiation at the second AUG of the ORF. To further study the roles of TLs in translation, I develop Translation-Associated Transcript Leader Sequencing (TATL-seq). TATL-seq works by applying TL-seq across fractions of a polysome gradient, generating TL-specific translational measurements. This approach demonstrates a widespread inhibitory function for upstream AUGs (uAUGs), and that ~6% of yeast genes express multiple TL species with distinct translational activities. This demonstrates that alternative TLs are prevalent and functional even in a relatively simple eukaryote like yeast.

My interest in alternative TLs prompted me to explore TL variation in mammals, where many thousands of genes are known to have alternative TLs. In Chapter 3 I enumerate the contributions of alternative mRNA processing events to alternative TLs in mice. I observe alternative TLs produced by alternative Transcription Start Sites (TSSs), and also demonstrate that alternative splicing events, such as skipped exons and alternative splice sites, contribute substantially to functional TL diversity. To facilitate the future study of alternative TLs in mammals, in Appendix I I modify TL-seq to sequence longer TL fragments and optimize TL-seq's enzymatic steps to reduce input RNA requirements.

This thesis is concerned with understanding post-transcriptional mRNA expression both globally and gene-specifically. In particular, I seek to understand the role the Transcript Leader has in affecting translation and degradation of its transcript. The findings detailed here define and analyze discernable features of TLs that relate to translational properties of the downstream message. Furthermore, the techniques developed enable analyses of TLs and translation that could not be carried out with previous technologies.

Thesis Supervisor: Wendy V. Gilbert
Title: Associate Professor of Biology

Acknowledgments

This thesis would not have been possible without several people. First of all, I would like to thank my advisor Wendy Gilbert. Her creativity, guidance, and endless energy have made me a better scientist, and without her I would not be where I am today.

Thanks are due to my thesis committee members Chris Burge and Angelika Amon, and members of their lab. I owe a special thanks to Jeremy Rock in the Amon Lab for introducing me to yeast, and Charles Lin and Jason Merkin in the Burge Lab for many helpful discussions. Also, I am grateful to Phil Sharp and Fred Winston for serving on my Thesis Defense Committee.

I would also like to thank all the members of the Gilbert Lab who have made the lab what it is today. Thanks to Boris Zinshteyn for many helpful conversations regarding computational analyses and programming. A very warm thank you to my baymate, Mary Kay Thompson, who has challenged me to be a better scientist, inspired me with her persistence, and encouraged me to think critically. Thank you to Thomas Carlile, Pavan Vaidyanathan, Maria Rojas-Duran, Audra Amasino, and Kristen Bartoli with whom I've had the pleasure of interacting with scientifically and socially.

Thank you to my friends and classmates whom I've gotten to know over the years. Nathaniel Schafheimer, Andrew Nager, and Mary Kay Thompson were gracious enough to drag me out of lab on Sunday nights to refine my tastes in cinema and wine, and stimulate my brain with great conversation. Thank you to Lorraine Ling and Robert Dorkin for being great roommates and kind friends.

I owe a large debt of gratitude to the BioREFS program, its members (past and present), and those people who I have interacted with by virtue of my affiliation with this organization. This organization has opened my eyes to the complexity of human interaction and the importance of humanity in all that we do, especially science.

The MIT Biology program as a whole deserves a great deal of credit. I would like to thank Betsey Walsh and Janice Chang for keeping the department running and helping me avoid fees. Also, Steve Bell and Tania Baker for running the Graduate Committee, and Alan Grossman for encouraging me to come to MIT all those years ago.

Lastly, I would like to thank my family. Thank you Mom and Dad for fostering my curiosity at such a young age and always encouraging me to do my best. My parents were always unrelenting advocates for my education, and without them I would not be here. Thank you to my sisters and the rest of my family. Whether they've realized it or not, they've been a great source of support and will continue to be in the future.

Biographical Note

Education

- 2008-2013 Ph.D. (Biology), Massachusetts Institute of Technology, Cambridge, MA
- 2004-2008 B.A. (MCB, Applied Mathematics) University of California, Berkeley, Berkeley, CA

Research Experience

- 2009-2013 Graduate Studies
Laboratory of Dr. Wendy Gilbert, MIT
Transcript Leaders and Translation, Translation Regulation Under Stress
- 2006-2008 Undergraduate Research
Laboratory of Dr. John Conboy, Lawrence Berkeley National Lab
Laboratory of Dr. Sharon Amacher, University of California, Berkeley
Fox RNA-Binding Proteins in Alternative Splicing in Zebrafish Develop.

Teaching Experience

- Spring, 2012 Head Teaching Assistant, Biochemistry (7.05), MIT
- Fall, 2009 Teaching Assistant, Introductory Biology (7.01), MIT
- Spring, 2008 Teaching Assistant, Biophysical Chemistry, UC Berkeley

Awards

- 2009-2013 Graduate Research Fellow, National Science Foundation
- 2008 MCB High Honors, UC Berkeley
- 2008 Spencer W. Brown award in Genetics, Genomics, and Development, UC Berkeley
- 2007-2008 Robert & Colleen Haas Scholar, UC Berkeley

Activities

- 2008-2013 BioREFS (Resource for Easing Friction and Stress)

Independent peer resource for support and mediation during stress

2008-2010 Chipperfield Committee
Graduate committee to invite distinguished speakers for a named
lectureship

Publications

Arribere JA, Gilbert WV. Roles for Transcript Leaders in Translation and mRNA Decay Revealed by Transcript Leader Sequencing. *Genome Res.* 2013 Jun;23(6):977-87.

Arribere JA, Doudna JA, Gilbert WV. Reconsidering Movement of Eukaryotic mRNAs Between Polysomes and P bodies. *Mol Cell.* 2011 Dec 9;44(5):745-58.

Gallagher TL, Arribere JA, Geurts PA, Exner CR, McDonald KL, Dill KK, Marr HL, Adkar SS, Garnett AT, Amacher SL, Conboy JG. Rbfox-Regulated Alternative Splicing is Critical for Zebrafish Cardiac and Skeletal Muscle Functions. *Dev Biol.* 2011 Nov 15;359(2):251-61.

Das D, Clark TA, Schweitzer A, Yamamoto M, Marr H, Arribere J, Minovitsky S, Poliakov A, Duchak I, Blume JE, Conboy JG. A correlation with exon expression approach to identify cis-regulatory elements for tissue-specific alternative splicing. *Nucleic Acids Res.* 2007;35(14):4845-57.

Table of Contents

Title Page	1
Abstract	2
Acknowledgements	4
Biographical Note	5
Table of Contents	7
List of Figures and Tables	12
Chapter One: Introduction	15
Transcript Leaders	16
Mechanism of Eukaryotic Cap-Dependent Translation Initiation	16
Examples of TL-Dependent Translation Regulation	19
<i>IRE/IRP-1</i>	19
<i>TL Length-Dependent Effects</i>	20
<i>IRESeS</i>	22
<i>uORFs</i>	23
Regulation via Intragenic TL Heterogeneity	27
TL Annotation Techniques	29
Evidence for Widespread Differences in Translational Activity and Possible Causal Role of TLs	33
Thesis Overview	34
References	35
Chapter Two: Roles for Transcript Leaders in Translation and mRNA Decay Revealed by Transcript Leader Sequencing	46
Abstract	47
Introduction	48
Materials and Methods	50
<i>Yeast strains and Growth Conditions</i>	50
<i>RNA Isolation and Polysome Gradient Fractionation</i>	50
<i>Transcript Leader Sequencing</i>	51
<i>Peak-Calling Algorithm</i>	52
<i>uAUG analysis</i>	52
<i>Shape Index</i>	53
<i>TATL-seq</i>	54
<i>Read Assignment</i>	54

<i>Peak RPKM</i>	55
<i>Ribosome Footprint Profiling</i>	55
<i>Nucleosome Analysis</i>	56
<i>Comparison with other TSS annotations</i>	56
<i>Peak Filters</i>	56
<i>Luciferase Reporter Assays</i>	57
<i>Normalized Read Density</i>	58
Results	59
<i>Defining Transcript Leaders</i>	59
<i>Transcription Initiation Within ORFs</i>	65
<i>Genes with Short TLs Exhibit Inefficient Start Codon Recognition</i>	71
<i>uAUGs Are Conserved Inhibitory Elements for Translation</i>	75
<i>Intragenic TL heterogeneity</i>	80
<i>Intragenic TL Heterogeneity Can Have Consequences for Translation</i>	82
Discussion	86
Acknowledgements	91
References	91
Chapter Three: Alternative Splicing Contributes to Functional TL Diversity	101
Abstract	102
Introduction	102
Materials and Methods	106
<i>Annotations</i>	106
<i>Percent Spliced In Analysis</i>	106
Results	106
<i>Mouse Alternative TLs are Produced by Alternative Promoters and Alternative Splicing</i>	107
<i>Alternative Splicing in TLs Regulates uAUG Abundance</i>	109
Discussion	110
Acknowledgements	111
References	111
Chapter Four: Conclusions and Future Directions	116

Introduction	117
Annotating TLs	117
Computational Tools for TL Analysis	119
Unexpected TSSs Identified by TL-seq	120
TATL-seq: Directly Analyzing TLs in Translation	122
Alternative TLs	124
Final Remarks	127
References	127
Appendix I: TL-seq Optimization for TL Identification in Mammalian Genomes	134
Abstract	135
Introduction	135
Materials and Methods	137
<i>Transcript Leader Sequencing</i>	137
<i>Biotinylated Oligo Capture</i>	138
<i>Annotations</i>	138
Results	138
<i>TL Fragments Hundreds of Nucleotides Long are Necessary to Define Mouse TLs</i>	138
<i>TL-seq Modification and Optimization for Mammalian Systems</i>	141
Discussion	142
References	145
Appendix II: Investigating the Translational Consequences of Acute eIF4E Depletion	148
Abstract	149
Introduction	149
Materials and Methods	152
<i>Yeast strains and Growth Conditions</i>	152
<i>S35 Incorporation</i>	152
<i>Polysome gradients and Ribo-seq</i>	153
<i>Western Blots</i>	153
<i>uORF analysis</i>	153
Results	154
<i>Techniques for Acute Reduction of eIF4E Activity</i>	154

<i>Assaying the Effect of eIF4E Depletion on Translation and Growth</i>	155
<i>Assaying the Effect of eIF4E Depletion on Gene-Specific Translation</i>	157
<i>GCN4 is Translationally Upregulated in cdc33-42</i>	158
<i>eIF4E Depletion Leads to an Increase in TL-mapping Ribo-seq Reads</i>	161
Discussion	164
Future Directions	166
Acknowledgments	168
References	168
Appendix III: Reconsidering Movement of Eukaryotic mRNAs Between Polysomes and P-bodies	173
Abstract	174
Introduction	174
Materials and Methods	177
<i>Yeast Strains and Culture</i>	177
<i>Extract Preparation, Polysome Gradient Fractionation and RNA isolation</i>	177
<i>cDNA Synthesis and Labeling, Microarray fabrication and Hybridization</i>	178
<i>Image Analysis and Microarray Data Processing</i>	178
<i>Clustering Analysis</i>	179
<i>Plasmids</i>	180
<i>Quantitative RNA Analysis</i>	180
<i>Microscopy</i>	181
Results	181
<i>Changes in Polysomal mRNA Levels Closely Parallel Changes in Transcript Abundance</i>	184
<i>Relationships Between Changes in Transcript Levels and Ribosome Occupancy</i>	189
<i>Functionally Distinct Groups of Genes Are Co-Regulated at the Post-Transcriptional Level</i>	193
<i>Only a Subset of mRNAs Can Return to Polysomes Following Starvation and Re-Feeding</i>	199
<i>Molecular Insights Into Selective Preservation of Non-Translating RPG Transcripts</i>	206

Discussion	212
<i>Translation is Reduced, Not 'Inhibited', Following Glucose Withdrawal</i>	212
<i>Most mRNAs are Depleted Coincident with Translational Repression</i>	213
<i>Mechanisms of Post-transcriptional Regulation in Glucose-starved Cells</i>	213
<i>Implications for the Interpretation of P-bodies</i>	216
<i>Bet-Hedging</i>	217
Acknowledgements	219
References	219

List of Figures and Tables

Figure 1.1	Mechanism of Eukaryotic Cap-Dependent Translation Initiation	18
Figure 1.2	Examples of TL-dependent Translation Regulation	26
Figure 1.3	TL Identification Techniques	32
Table 2.1	Strains Used For In Vivo Translation	58
Table 2.2	Oligonucleotides Used In This Study	59
Figure 2.1	TL-seq Preferentially Recovers Capped 5' Ends	61
Figure 2.2	TL-seq Produces TSS Annotations	64
Figure 2.3	TL-seq Detects Internal Peaks with a TSS-like Nucleosome Distribution	66
Table 2.3	Genes With Similar Internal TSSs Between TL-seq and (Miura et al. 2006)	67
Figure 2.4	Three Types of Internal TSSs Identified by TL-seq	69
Figure 2.5	Northern Validation of Internal Peak-Containing Genes	70
Figure 2.6	TL-seq Internal Peaks Are Also Peak-like in RNA-seq Libraries Made Using Similar 5'-Capture Techniques	71
Figure 2.7	Short TL Genes Are Enriched for NMD Targets	74
Figure 2.8	Validation of Steady State mRNA Fold Changes in NMD-Deficient Yeast	75
Figure 2.9	TATL-seq Quantifies Translation Activity of TLs In Vivo	76
Figure 2.10	uAUGs Are an Underrepresented and Conserved Sequence Element Associated with Decreased Translation	78
Figure 2.11	There Are No Detectable Effects of Near-uAUG Codons on Translation or NMD	79
Figure 2.12	There is Intragenic TL Heterogeneity	81
Figure 2.13	Distribution of RNA-seq Reads About Short/Long TL Pairs	83

Figure 2.14	Intragenic TL Heterogeneity Leads to Different Translation Behavior In Vivo	84
Figure 2.15	Distribution of TL Peaks Across a Polysome Gradient for Genes Shown in Figure 2.14D	85
Table 2.5	Summary Information From Peaks for Pooled TATL-seq Libraries	86
Figure 3.1	Alternative Events that Contribute to TL Diversity	108
Figure 3.2	Alternative Splicing Affects the Abundance of uAUGs	110
Figure I.1	TL Length and Exonic Structure For Mouse Transcripts	140
Figure I.2	TL-seq Modification to Sequence Longer TL Fragments With Less Input	142
Figure II.1	4EGI-1 Does Not Inhibit Cell Growth in <i>S. cer</i>	155
Figure II.2	<i>cdc33-42</i> Exhibits Decreased Growth and Protein Synthesis at 32C	156
Figure II.3	Gene Expression is Correlated between <i>cdc33-42</i> and <i>CDC33</i>	158
Figure II.4	<i>GCN4</i> Exhibits eIF2 α -Independent Translational Upregulation in <i>cdc33-42</i>	160
Figure II.5	Global Distribution of Read Locations in <i>cdc33-42</i> and <i>CDC33</i>	162
Figure II.6	Global Increase in uORF-Mapping Reads in Multiple Ribo-seq Datasets, Regardless of Initiation Codon	163
Figure II.7	Cycloheximide-Dependent Increase in non-AUG uORF-Mapping Reads	164
Figure III.1	Regulation of Transcription and Translation in Glucose-Starved Cells	183
Figure III.2	Induction Kinetics of <i>Adr1</i> , <i>Cat8</i> , <i>Rgt1</i> and <i>Mig1</i> Target Genes	186
Figure III.3	Changes in Total mRNA Levels Exceed Changes in Polysomal mRNA Levels	188
Figure III.4	Gene Groupings Produced by k-means Clustering at Different Values of <i>k</i>	190
Figure III.5	Ribosome Occupancy and mRNA Abundance are Divergently	

	Regulated	191
Figure III.6	RPGs and RBGs Differ in their Post-Transcriptional Responses to Glucose Withdrawal	196
Figure III.7	Highly Expressed mRNAs are Preferentially Retained in the Non-Translating Pool	197
Figure III.8	Quantitative RT-PCR Validation of Select Genes' P/T ratios -Glucose	198
Figure III.9	Translational Resurrection is Restricted to a Subset of Genes for a Limited Time	200
Figure III.10	P-bodies are Present Under the Conditions Examined by Microarray	202
Figure III.11	Quantitative RT-PCR Validation of Select Genes' mRNA Abundance in Polysome Fractions Following Glucose Starvation and Re-Feeding	204
Figure III.12	RPGs Have Relatively Short Half-Lives in Glucose Replete Conditions	208
Figure III.13	Resurrection-Competent mRNAs Associate with Pab1 and Have Longer Poly(A) Tails	209
Figure III.14	Ribosome Occupancy and Ribosome Density are Essentially Unrelated	210

Chapter 1

Introduction

Transcript Leaders

Transcript Leaders (TLs, or 5'Untranslated Regions) are a feature of mRNAs that are heterogeneous and important for gene expression. In eukaryotes, the TL is defined as the region of an mRNA from the methyl-7-guanosine (m7G) 5'-5' triphosphate cap up to the start codon (AUG) of the ORF. The 5' end of the TL is made shortly after transcription initiation as the nascent eukaryotic pre-mRNA 5' end undergoes conversion from a triphosphate to a methyl-7-guanosine (m7G) 5'-5' triphosphate cap, a process known as capping (Rasmussen and Lis 1993). Most genes have at least one TL, and many have intragenic TL heterogeneity (multiple TLs). TLs vary in length and sequence composition within and between genes and across species. They range from an average of 20-30 nts in yeast to 100-200 nts in mammalian genomes, though even within an organism there is a high variance to this distribution (Calvo et al. 2009; Nagalakshmi et al. 2008). TLs' heterogeneous sequence composition leads to differences with respect to RNA secondary structure and RNA binding protein (RBP) sites. First I will describe how the TL plays a central role in translation initiation, and then how TL heterogeneity is exploited for translational regulation.

Mechanism of Eukaryotic Cap-Dependent Translation Initiation

To initiate the process of translation, a TL must engage the numerous components of the eukaryotic translation initiation machinery. During translation initiation, the m7G cap of the mRNA is bound by the cytoplasmic cap-binding complex, eIF4F, composed of eIF4G, eIF4A, eIF4B, and the cap-binding protein eIF4E (Figure 1.1, reviewed in (Jackson et al. 2010)). Interaction of eIF4G with the TL is stabilized by interactions with

the Poly(A)-Binding Protein (Pab1 in yeast) that is bound to the poly(A) tail. Meanwhile a small ribosomal subunit (40S) is bound by initiator tRNA and eukaryotic initiation factors (eIFs) 1, 1A, 2, 3, and 5, each of which is composed of 1 to 13 subunits, forming a 43S complex. Through a poorly understood mechanism eIF4F is thought to facilitate recruitment of a 43S complex downstream of the cap yet near the 5' end of the mRNA. Subsequent to loading, the 43S occupies a large footprint of RNA, covering >12 nt upstream of the peptidyl transfer site (P-site) and ~16nt downstream (Legon 1976).

Once bound to the mRNA, the translation machinery must then locate the ORF start codon and commence translation elongation. To locate the start codon (AUG) of the ORF, the 43S complex scans the TL linearly in a net 5' to 3' direction, inspecting sequential nucleotide triplets using the anticodon of the associated initiator tRNA. Scanning is a processive process facilitated by RNA helicases such as eIF4A and can cover hundreds or thousands of nucleotides until a start codon is recognized by the 43S. Once an AUG is recognized, the 43S undergoes irreversible conformational changes triggering eIF rearrangement and halting the scanning process. Finally, eIF5B facilitates association of a large subunit (60S) with the small subunit to form an 80S ribosome, thus completing translation initiation and beginning translation elongation.

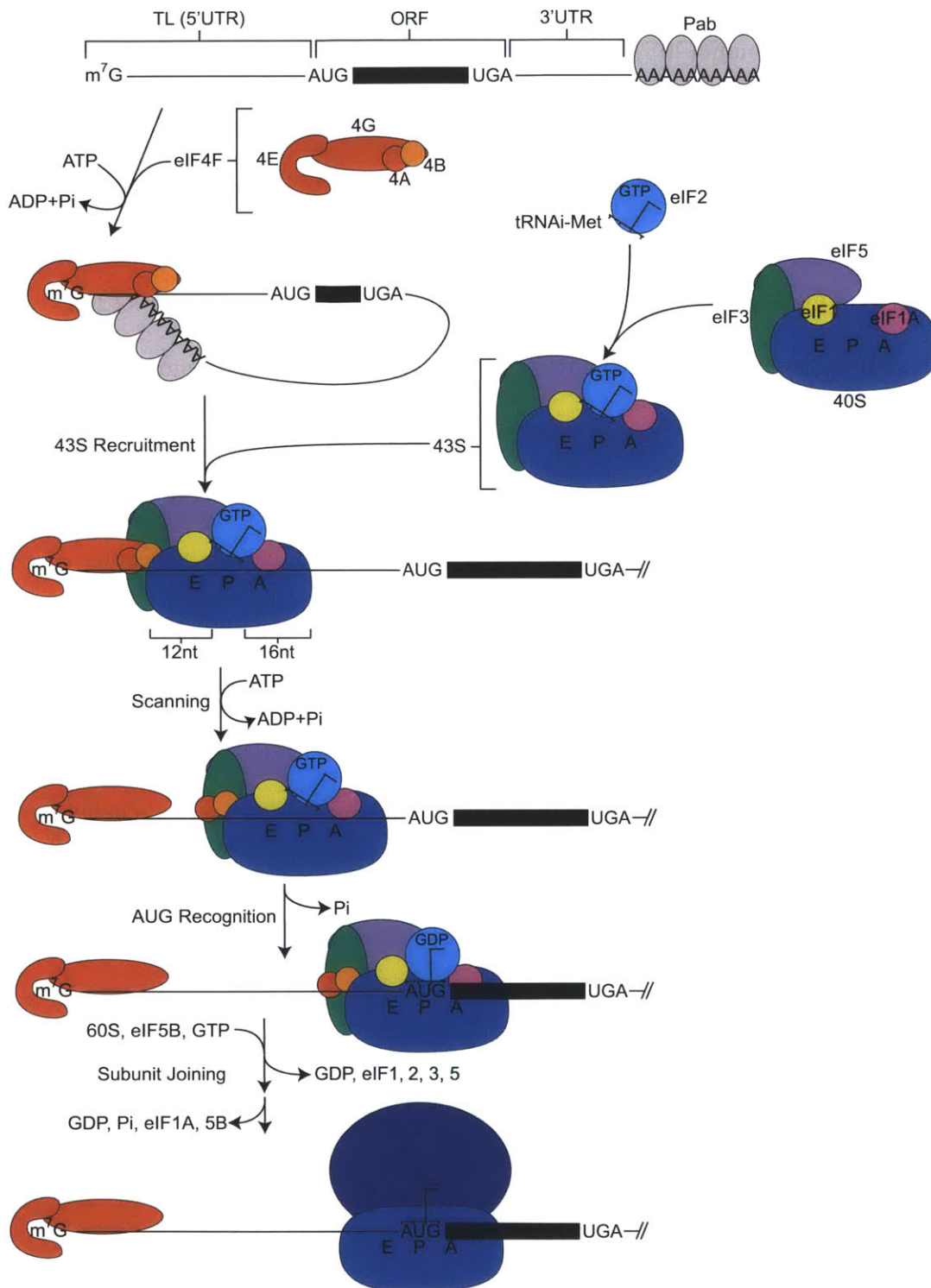


Figure 1.1: Mechanism of Eukaryotic Cap-Dependent Translation Initiation

This figure was adapted with modification from (Jackson et al. 2010). While we show eIF4A/B traveling with the small subunit, it is unknown whether eIF4F or its subunits maintain interactions with the 43S complex during scanning (for a discussion of this, see

(Jackson et al. 2010)). For simplicity the mRNA downstream of the ORF is omitted after 43S binding and drawings are not to scale.

Translation initiation is the rate-limiting step in translation and is subject to extensive regulation both globally and gene-specifically. Global changes in translation can be effected by altering the activities or levels of the eIFs. Many of the aforementioned eIFs are dynamically modified phosphoproteins: eIF4G is modified in response to serum starvation (Raught et al. 2000); eIF2 is phosphorylated in response to amino acid starvation (Dever et al. 1992); and eIF4E is a target of MNK1/2 kinases (Pyronnet et al. 1999). eIF regulation can also lead to gene-specific translational changes as well (see *GCN4* below), and both global and gene-specific effects are often mediated by the TL. Gene-specific effects can also occur via RNA secondary structure and RBP sites contained in the TL. In the next section I will detail several examples illustrating how the TL is known to impact translation of the downstream ORF.

Examples of TL-Dependent Translation Regulation

IRE/IRP-1

TLs can exert physiologically relevant translation regulation by obstructing the initiation machinery. For example, in the mammalian *ferritin* mRNA, the TL region adjacent to the cap contains an RNA motif known as the Iron Response Element (IRE), which mediates iron-regulated translational repression (Aziz and Munro 1987). The IRE forms a hairpin secondary structure that binds the Iron Response Protein (IRP-1) and serves as a barrier to 43S recruitment, thus preventing translation of the downstream ORF (Figure 1.2A) (Muckenthaler et al. 1998). If the IRE is artificially placed further downstream from the

cap, the 43S can then load onto the mRNA, yet the IRE/IRP-1 blocks the 43S from scanning (Paraskeva et al. 1999). Interestingly, IRP-1 binding to the IRE is iron-dependent: when intracellular iron is high, IRP-1 dissociates from the IRE, enabling translation of ferritin protein, an iron storage protein that sequesters iron and thus protects the cell from the deleterious effects of high intracellular iron. Other proteins also demonstrate the impediment mode of regulation typified by IRP-1/IRE. For example, in neurons, activity-dependent phosphorylation of GRB7 decreases its binding to the 5' end of the *kor* mRNA and allows for eIF4E binding (Tsai et al. 2007). Additionally, RNA structures alone can serve to block initiation: artificial mRNA hairpins obstruct recruitment and/or scanning of a 43S complex *in vitro* and *in vivo* in a stability-dependent manner (Babendure et al. 2006). Thus TLs can affect translation of the downstream ORF by regulating initiation factor binding or procession.

TL Length-Dependent Effects

Numerous studies argue that long TLs can contribute to translational regulation. TLs hundreds of nucleotides long are known to cause poor translation for genes that encode regulatory functions such as signal transduction and cell cycle control, arguing for an important role in regulation of gene expression (Willis 1999; Morrish and Rumsby 2001). The poor translation of these long TLs is thought to be due to inefficient scanning over long distances, possibly because of an increased propensity for long TLs to form secondary structures that can obstruct 43S scanning. In support of this, knockdown of RNA helicases such as DDX3 preferentially decrease translation of mRNAs with long TLs (Lai et al. 2008). It is ambiguous whether this is due to helicase activity directly

driving 43S scanning over long distances, indirectly enhancing scanning by unwinding TL secondary structure, or both. Using different TL sequences, other studies have demonstrated no loss in translation activity for TLs ranging from 43 nt to over 1,000 nt in length (Kozak 1991b; Berthelot et al. 2003). These seemingly contradictory results can be reconciled if one hypothesizes that TL sequence and secondary structure, and not length alone, has an important role in determining translation activity.

At the opposite end of the spectrum, short TLs can decrease translation efficiency. While many mRNAs have been found that have naturally short TLs (e.g. 2% of TLs are <12nt in (Xu et al. 2009)), most of what we know about translation of genes with short TLs comes from artificial short TL reporter constructs and naturally short viral RNAs. Because the 43S complex protects at least 12 nts upstream of its P-site (see above), it is unclear how eIF4F might recruit a 43S to an AUG <12 nt from the cap. Consistent with this view of the spatial constraints of ribosome recruitment, artificial shortening of TLs to only a few nucleotides in length decreases the efficiency of initiation at the cap-proximal AUG in mammals and yeast (Figure 1.2B) (van den Heuvel et al. 1989; Pestova and Kolupaeva 2002; Sedman et al. 1990; Kozak 1991a). Concomitant with a decrease in cap-proximal initiation, an increase in initiation at the next downstream AUG codon is observed, demonstrating that 43S complexes still load and are competent for initiation, though fail to recognize the cap-proximal AUG. While the physical constraints of the translation machinery might anticipate full functionality for TL lengths greater than 12 nt, in practice more space appears to be required: TL lengths up to 20nt appear to have defects in initiation at the cap-proximal AUG (van den Heuvel et al. 1989). In Chapter 2 I

will describe hundreds of examples of short TLs found in yeast, and discuss how their length influences the translation and decay of the mRNA as a whole *in vivo*.

IRESeS

Internal Ribosomal Entry Sites (IRESeS) are sequences that substitute one or more of the eIFs' activities in initiation and recruit ribosomes internal to the 5' end of the mRNA. The impetus for their discovery was the revelation that picornavirus and poliovirus RNAs are efficiently translated, and yet their TLs are highly structured RNAs that lack an m7G cap at their 5' end (Jang et al. 1988; Pelletier and Sonenberg 1988). Such viral IRESeS initiate translation by substituting eIF binding and/or function with structural RNA elements. In the case of the poliovirus and EMCV IRESeS, an RNA structure in the TL binds and recruits eIF4G, which then carries out its initiation functions in the absence of eIF1, 1A, and 4E (Pestova et al. 1996; Lomakin et al. 2000). Additional IRESeS exhibit distinct factor requirements: the HCV IRES binds directly to eIF3 and uses none of eIF4G/A/B/E, and at the extreme end, the CrPV IRES requires no eIFs (Kieft et al. 2001; Wilson et al. 2000).

IRESeS also occur in some cellular TLs of stress-induced genes. One notable example of this is *YMR181C*, encoding a protein required for yeast starvation response (Gilbert et al. 2007). *YMR181C* is transcribed as a bicistronic transcript continuous with its upstream gene, and thus, akin to the viral situation, lacks an m7G cap near its ORF (one is present, but is >1.1kb upstream). A 12nt adenosine-rich tract upstream of the *YMR181C* start codon is necessary and sufficient for its cap-independent translational activity, and this region is thought to recruit Pab1p, then eIF4G, to the mRNA. IRESeS

have been reported in numerous cellular mRNAs important for cellular stress responses, indicating a potentially important role in cellular adaptation to stress (reviewed in (Komar 2005)).

uORFs

TL-contained AUG codons are a notable feature of mRNAs that directly compete with the downstream ORF for ribosomes. Because scanning is linear and directional from the cap, AUG codons in the TL (upstream AUGs, uAUGs) that are distinct from the AUG of the ORF can also serve as initiation sites for scanning 43S complexes (Kozak 1989). Additionally, uAUGs are followed by a short peptide and termination codon, making an upstream ORF (uORF). Most commonly, ribosomes terminate after translating a uORF and dissociate from the mRNA, thus decreasing translation of the downstream ORF.

Due to their simple sequence characteristics, uAUGs and uORFs are one of the most prevalent and readily studied feature in TLs. Between 40 and 50% of all annotated human and mouse mRNAs have at least one uORF (Calvo et al. 2009). Consistent with their known inhibitory function, uORFs are associated globally with a decrease in protein expression from the downstream ORF (Calvo et al. 2009). Due to a lack of TL annotations, the prevalence of uORFs in yeast has remained unknown—genome-wide estimates of uORF prevalence range from 15% (Lawless et al. 2009) to 30% (Marija Cvijović 2007). Also, due to a lack of translational datasets, the global effect of uORFs on steady-state protein levels has not been analyzed in yeast.

uORFs provide unique and fascinating regulatory opportunities for an mRNA. In addition to regulating ORF start codon recognition, some uORFs encode regulatory

peptides that stall ribosomes in *cis*. During translation of the yeast *CPA1* gene, the Arginine Attenuator Peptide encoded by a uORF in the TL stalls translating ribosomes in an arginine concentration-dependent manner (Wang et al. 1999). The *CLN3* uORF is subject to leaky scanning and is only recognized by a fraction of the ribosomes that scan over it (Polymenis and Schmidt 1997). A minority of uORFs also cause reinitiation after their translation. Once ribosomes translate the first uORF in the TL of yeast *GCN4*, they reinitiate scanning and initiate translation at one of three other uORFs or the *GCN4* ORF (Figure 1.2C, reviewed in (Hinnebusch 2005)). However, in order to recognize an AUG codon, the reinitiating ribosome must acquire eIF2-tRNA-Met (Ternary Complex, TC). When free TC levels are high, reinitiating ribosomes rapidly bind TC, translate uORF 2, 3, or 4, and dissociate from the transcript. When TC levels are low, reinitiating ribosomes scan past the uORFs before being bound by TC, leading to initiation at the *GCN4* ORF. Importantly, the levels of free TC are subject to regulation by eIF2 α phosphorylation, which increases during stresses including amino acid starvation. Thus while uORFs decrease translation generally, they also allow for selective translational regulation of the downstream protein.

Because uORFs lead to translation termination near the 5' end of a transcript, some uORF-containing mRNAs are regulated by nonsense-mediated mRNA decay (NMD). NMD is a process by which ribosomes terminating at premature termination codons (PTCs) are recognized as aberrant and the PTC-containing mRNA is degraded (reviewed in (Kervestin and Jacobson 2012)). In yeast PTC recognition is thought to occur by a distance-sensing mechanism—ribosomes terminating far from the 3' end of the mRNA are flagged as premature (Kebaara and Atkin 2009). Such prematurely

terminating ribosomes can result from errors in gene expression, though increasingly NMD has come to be recognized as an opportunity for post-transcriptional regulation. Ribosomes terminating at a uORF can elicit NMD, making uORFs important regulators of both translation and decay of their mRNA (Gaba et al. 2005). However, NMD does not affect all uORF-containing mRNAs: while uORFs in *CPA1*, *EST1*, and *PET130* are thought to trigger NMD, those in *GCN4* and *YAPI* do not (He et al. 2003; Vilela et al. 1998; Ruiz-Echevarría and Peltz 2000). With knowledge of only a few anecdotal examples, it has remained unknown the extent to which uORFs elicit NMD genome-wide in yeast.

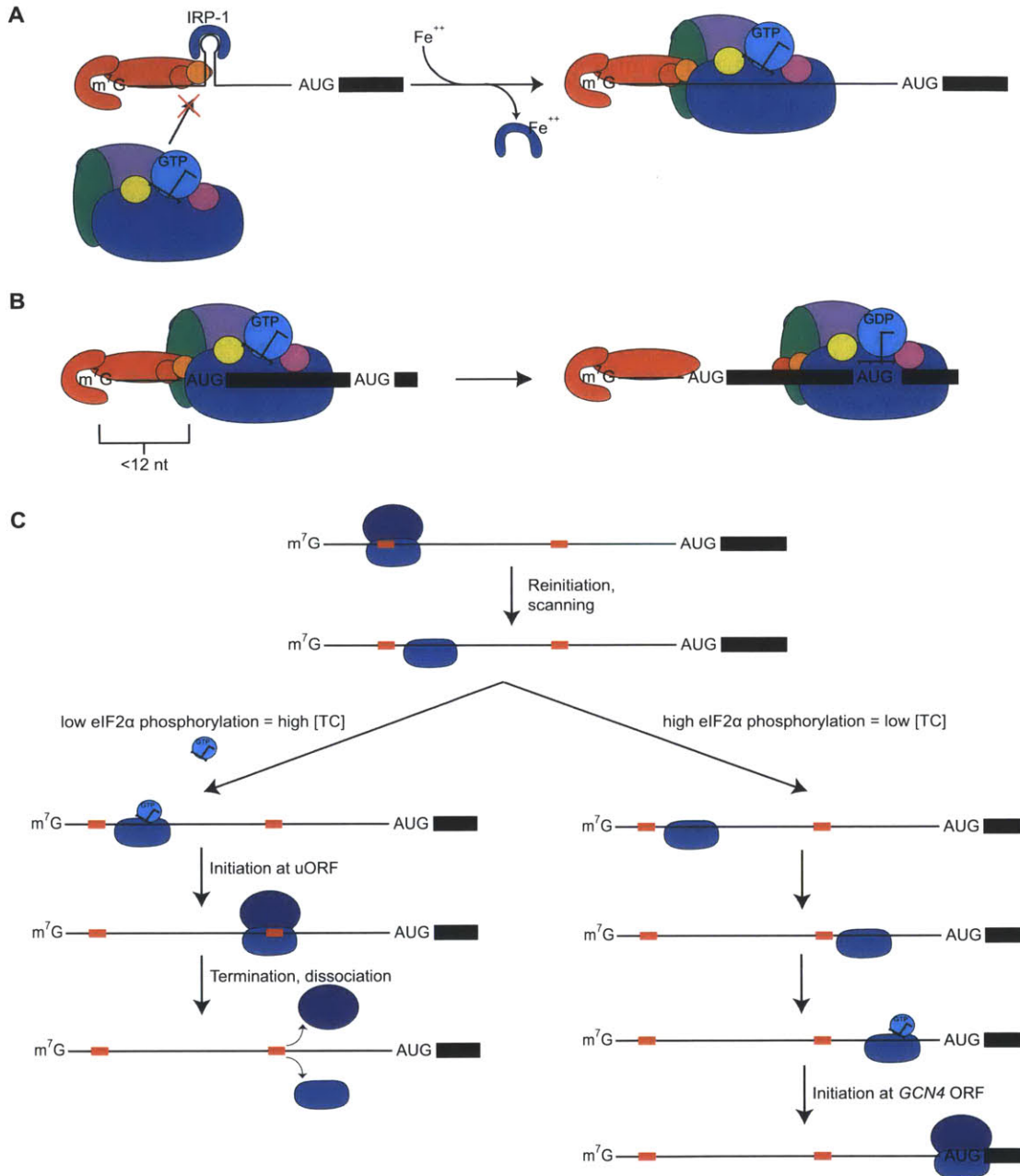


Figure 1.2: Examples of TL-dependent Translation Regulation

(A) The IRE blocks recruitment of 43S to the mRNA via iron-dependent binding of IRP-1. eIFs are colored as in Figure 1.1.

(B) Short TLs (<12nt) are known to exhibit reduced initiation from cap-proximal AUGs due to the physical constraints of the initiation machinery. 43S complexes load, however, and initiate at downstream AUGs (right).

(C) Reinitiation after translation of uORF1 in the *GCN4* TL is regulated by eIF2 α phosphorylation. Low eIF2 α phosphorylation causes high levels of free TC (eIF2-tRNA-Met), leading to translation at uORFs 2-4 (left arrow). High eIF2 α phosphorylation reduces levels of free eIF2-tRNA-Met, causing scanning past the uORFs and initiation at

the *GCN4* start codon (right arrow). uORFs are indicated by red boxes, *GCN4* ORF by black box. For simplicity, all eIFs except eIF2 are omitted, and only uORFs 1 and 4 are shown.

The above examples all demonstrate how a single TL can engage the translational machinery to regulate its downstream ORF. However, oftentimes an ORF will not have a single TL, but several TL species that differ by a few or hundreds of nucleotides. These distinct TL species often exhibit different translational behaviors through mechanisms outlined above. Thus, as will be discussed in the next section, to understand the composite post-transcriptional behavior of an ORF, one must appreciate the contributions of its individual TL species.

Regulation via Intragenic TL Heterogeneity

TL heterogeneity is a pervasive feature of eukaryotic genes. Evidence from small-scale studies of Transcription Start Site (TSS) locations in yeast shows a majority of genes observed have TSS heterogeneity generating TLs that vary in length from a few bases to several dozen nucleotides (Hahn et al. 1985). Genome-wide studies argue that >95% of yeast genes have more than one TL (Zhang and Dietrich 2005; Miura et al. 2006).

Heterogeneity is also prevalent in mammalian systems, and genes with more TL heterogeneity tend to encode regulatory proteins such as transcription factors, suggesting that TL variation is important for regulation of gene expression (Resch et al. 2009). In humans alone it is estimated there are at least 10,000 alternative first exons, making alternative TLs a prominent contributor to mRNA isoform diversity (Wang et al. 2008). Despite our ever-expanding appreciation of TSS heterogeneity, only a handful of the alternative TLs they generate have been assessed for functionality.

Alternative TLs can function by changing the N-terminus of the encoded protein. Such functional TL heterogeneity was first shown via studies of the yeast gene for invertase, *SUC2*, which encodes both a secreted and intracellular form of the sucrose hydrolyzing enzyme (Carlson and Botstein 1982). Production of both proteins occurs via alternative TLs: a distal TL isoform encodes an N-terminal secretion signal absent from the proximal TL. Additional examples of TL-mediated N-terminal alternative coding events include yeast *HTS1*, which encodes both cytoplasmic and mitochondrial histidine tRNA synthetase, yeast *FUM1*, which encodes both cytoplasmic and mitochondrial fumarase, and the mammalian cannabinoid receptor, which encodes two distinct extracellular termini (Chiu et al. 1992; Wu and Tzagoloff 1987; Shire et al. 1995).

Alternative TLs commonly confer distinct translational behaviors on their downstream ORF. A recent study examining translation of TLs from nine yeast genes with alternative TLs observed that most conferred significantly different translation activity in a reporter *in vitro* and *in vivo* (Rojas-Duran and Gilbert 2012). In this study, some TL variants differed by as little as nine nucleotides, demonstrating that even small differences in TSS choice and TL length can have consequences for translation.

Alternative TLs are also functional in higher eukaryotes: the human *NOD2* gene, which is genetically implicated in Crohn's disease, encodes two alternative TL variants whose translation responds differently to the drug rapamycin (Rosenstiel et al. 2007). Similarly, the human estrogen receptor beta (*ERβ*) gene encodes at least four alternative TLs whose translation differs and expression changes in tissues and cancer (Smith et al. 2009; 2010). The mammalian neurotrophin *BDNF* gene contains at least nine alternative TLs with distinct expression patterns and functions in humans (Baj and Tongiorgi 2008; Pruunsild

et al. 2007). In each of the above cases in which it has been assayed, alternative TLs confer distinct translational activities.

Although TL heterogeneity has clear effects on gene expression, it remains understudied, partially for technical reasons. Most techniques, whether genome-wide or gene-by-gene, utilize sequences in the ORF to quantitate or follow the activity of a gene, and these sequences are largely constant between alternative TL variants, rendering the above regulation invisible. Thus to study alternative TLs, and TLs in general, techniques must exist which can reliably distinguish and readily define them.

TL Annotation Techniques

Whole-length cDNA sequencing is a low-throughput technique for mapping TSSs, though it can provide false 5' end information. In this approach, reverse transcriptase (RT) primes off an oligonucleotide poly-dT bound to the poly(A) tail of an mRNA and elongates through to the 5' end of the message. Large-scale cDNA sequencing efforts were instrumental in annotation of metazoan genomes, in particular human and mouse, though such efforts were not undertaken in many organisms including yeast.

Unfortunately because RT has poor processivity (superscript III RT, for example, polymerizes only ~40nt per template association) (Huber et al. 1989; Katz and Skalka 1994), prematurely truncated cDNAs are frequently generated. If the RT does reach the 5' end, it can add untemplated bases, frustrating efforts for single-nucleotide TSS identification. Also, since cDNA sequencing requires reading out long stretches of DNA, it lacks the throughput conferred by current next generation short-read sequencers.

Because TLs often change in response to extracellular stimuli (Law et al. 2005), throughput is important when seeking to examine TLs in multiple cellular conditions.

To take advantage of the recent rise in genome-wide technologies, computational methods have been developed to extract mRNA boundaries from microarray tiling and RNA sequencing data. In microarray tiling approaches, oligonucleotides that tile the genome are hybridized to fluorescently labeled cDNA (Figure 1.3A) (David et al. 2006; Xu et al. 2009). The 5' end of a gene is determined computationally by a segmentation algorithm that divides the genome into contiguous regions of consistent probe intensity. Similarly, to estimate TSSs from RNA-seq, read density upstream of a gene can be fed into a machine learning algorithm trained on 5' ends empirically determined via gene-specific 5'RACE (Figure 1.3B, see description of 5'RACE below) (Nagalakshmi et al. 2008). The net effect of both of these approaches is to identify the genomic position where signal (probe intensity or read density) drops off, interpreted as a switch from a transcribed region to a non-transcribed one. While both of these approaches yield genome-wide TSS annotations, the resultant TSS estimates are imprecise and fail to appreciate TL heterogeneity by imposing a one gene, one TL rule.

Targeted techniques exist that directly sequence only the 5' ends of mRNAs and can thus observe TSS heterogeneity. Most notable among these techniques are 5'Rapid Amplification of cDNA Ends (5'RACE) and Capped Analysis of Gene Expression (CAGE) (Figure 1.3C,D) (Maruyama and Sugano 1994; Shiraki et al. 2003). In both these techniques the unique chemical properties of the m⁷G-triphosphate linkage are used to isolate and sequence the exact 5' ends of mRNAs. For 5'RACE, the cap is removed, an adaptor is attached via ligation, and prior knowledge of the mRNA sequence is used to

design primers that specifically amplify the TSS of a gene of interest. In CAGE, capped RNAs are biotinylated, selected from a pool of other RNAs, and a linker with an MmeI restriction site is added to the 5' end. MmeI cleaves 20nt downstream, generating a pool of DNA molecules each containing the first ~20nts of a transcript. In the first iterations of CAGE, these molecules were concatamerized and sequenced as a plasmid; in modern iterations deep sequencing is employed (Plessy et al. 2010). In principle these techniques should comprehensively identify TSSs, but in practice this is not always the case. Biases in 5' end capture caused by the enzymes involved can lead to failure to detect all 5' ends of a gene (false negatives) and/or capture of internal, non-TSS regions of the mRNA (false positives). Despite these imperfections, 5'RACE and CAGE are widely used techniques that are useful for both gene-specific (5'RACE) and global (CAGE) studies of TLs. The chemistry behind these techniques has been employed to create many related techniques, including PEAT, CapSeq, nanoCAGE, and RAMPAGE (Ni et al. 2010; Pelechano et al. 2013; Plessy et al. 2010; Batut et al. 2013; Gu et al. 2012).

In Chapter 2 I will describe a technique I developed to annotate TLs genome-wide, Transcript Leader Sequencing (TL-seq), and its application to yeast. This technique takes advantage of next generation sequencing, but with a simpler library procedure and longer read length than conventional CAGE. I also describe computational tools I developed to reduce non-TSS artifacts, enabling genome-wide analysis of TLs. Such accurate genome-wide annotations will be instrumental in globally understanding the post-transcriptional behavior of mRNAs.

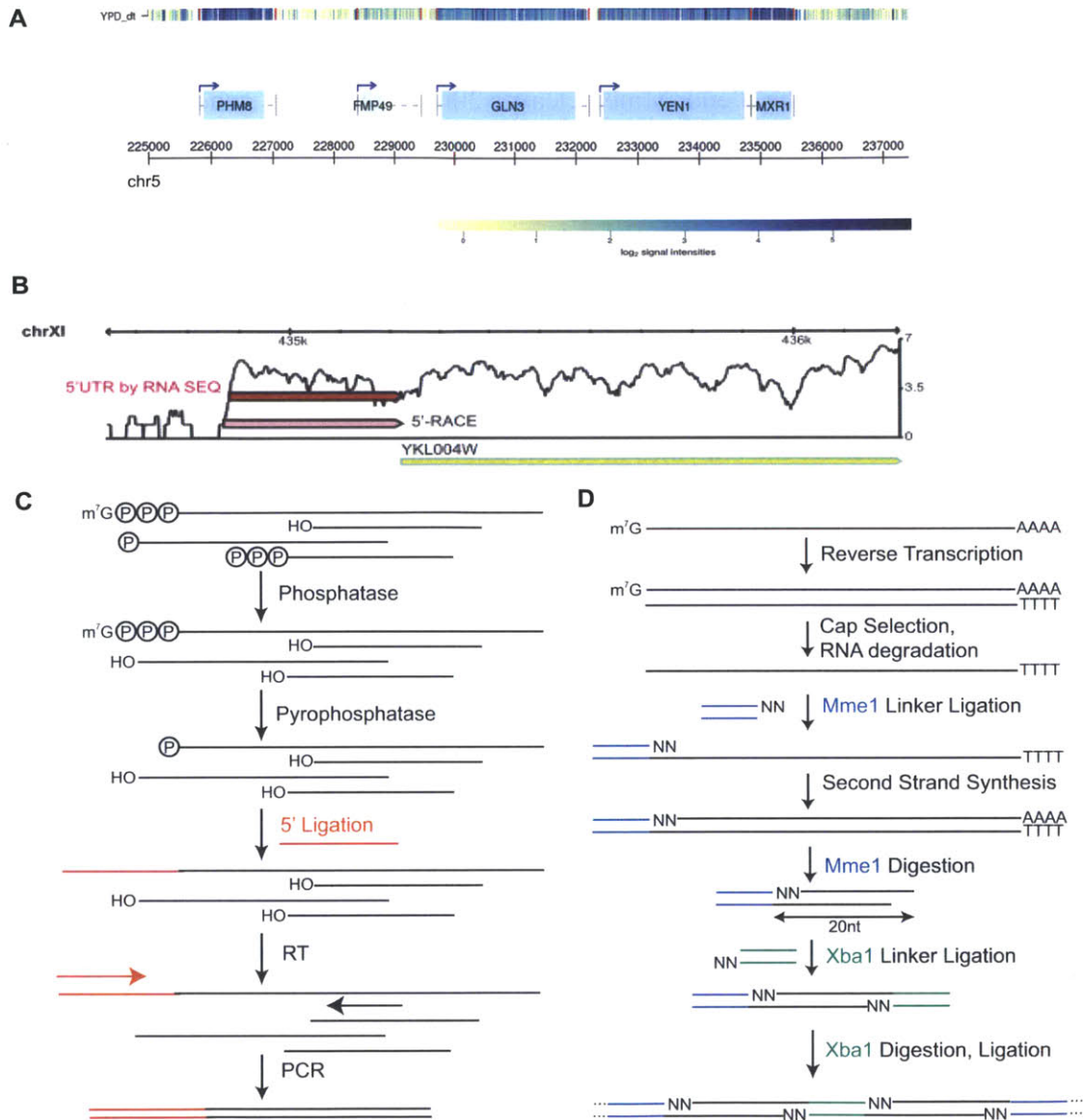


Figure 1.3: TL Identification Techniques

(A) Microarray tiling approach to determine transcript boundaries for a region of chromosome 5. YPD_dt shows the signal intensity for microarray hybridization, and red lines indicate transcript boundaries identified via segmentation. Blue arrows indicate TSSs. Figure made from the online genome viewer (<http://steinmetzlab.embl.de/NFRsharing/>) from (Xu et al. 2009).

(B) RNA-seq approach to determine TL boundaries for *YKL004W*. Shown in pink is the TSS identified by 5'-RACE, and in red is the boundary identified by computational analysis of RNA-seq data. Figure was taken directly from (Nagalakshmi et al. 2008), without modification.

(C) 5'RACE allows for selective TSS identification for a gene of interest. A 5'adaptor (red) of known sequence is ligated to formerly capped RNAs. PCR using a forward

primer in the adaptor (red arrow) and a reverse gene-specific primer (black arrow) allows for selective amplification of a gene of interest.

(D) CAGE captures the first 20nt of a transcript. Cap selection includes oxidation, biotinylation, and selection of capped species with streptavidin as described in (Carninci and Hayashizaki 1999). Here, 20nt tags are shown concatamerized, and are then subject to sanger sequencing.

Evidence for Widespread Differences in Translational Activity and Possible Causal

Role of TLs

Recent technological developments have enabled a more quantitative and global view of translation. Ribosome footprint profiling (Ribo-seq) provides a snapshot of ribosome abundance and distribution for all translated mRNAs (Ingolia et al. 2009). During Ribo-seq, ribosomes in the act of translating an mRNA are harvested and subject to mild ribonuclease treatment. The nuclease cleaves the unbound portions of the mRNA, while the ribosome protects a ~28nt stretch of the RNA (the so-called “footprint”). Footprints are then harvested and captured for high-throughput sequencing, yielding 28nt reads. The translation efficiency for any given gene can then be defined as the density of Ribo-seq reads to RNA-seq (total transcript) reads.

Numerous Ribo-seq studies have demonstrated that translational differences are both pervasive and regulated. Translation efficiencies are gene-specific and vary genome-wide by >100 fold, and this wide breadth of translation activity has been observed in every system in which it has been examined, from *E. coli* to human cells (Guo et al. 2010; Ingolia et al. 2009; 2011; Stadler and Fire 2011; Oh et al. 2011). Furthermore, gene-specific translation efficiencies change in different cellular states, such as starvation in yeast and stem cell differentiation in mice, demonstrating that translation has a large dynamic range and is subject to extensive regulation (Ingolia et al. 2009; 2011).

We are now in a position of trying to understand and study how such differences in translation arise. Extrapolating from studies of individual genes (above), it stands to reason that translation efficiency is determined in large part by the non-coding sequences of mRNAs, in particular TLs. However, in order to relate TL properties to translational measures, one must first know the precise sequence of the TL for any given gene. Also, since many techniques (Ribo-seq included) provide ORF-based measures of translation, they cannot quantify the relative contributions of alternative TLs to the translational behavior of a given gene. To study such TLs, additional techniques are required that directly monitor alternative TLs.

Thesis Overview

This thesis studies the role of TLs in translation. In Chapter 2 I develop TL-seq, a technique that can annotate TLs *de novo*, and demonstrate its application to yeast. Using these annotations, I identify hundreds of mRNAs with TLs <12nt long, and demonstrate out-of-frame initiation leading to NMD on these messages. In conjunction with polyribosome fractionation, I directly monitor the activity of individual TLs in translation. This approach demonstrates that uAUGs have a negative impact on translation and that alternative TLs have distinct translational activities. In Appendix I, I revise TL-seq to produce paired-end read information and optimize it to work with low input requirements, making it more practical for mammalian samples. Using extant annotations, I show in Chapter 3 that alternative splicing in TLs increases the prevalence of uAUGs, thus creating functionally distinct TL variants. Finally, I describe my efforts to understand the global and gene-specific translational consequences of acute depletion

of eIF4E in yeast. All together, this thesis presents techniques to study TL-dependent translational regulation and establishes their utility by analyzing TLs genome-wide.

References

Aziz N, Munro HN. 1987. Iron regulates ferritin mRNA translation through a segment of its 5' untranslated region. *Proc Natl Acad Sci USA* **84**: 8478–8482.

Babendure JR, Babendure JL, Ding J-H, Tsien RY. 2006. Control of mammalian translation by mRNA structure near caps. *RNA* **12**: 851–861.

Baj G, Tongiorgi E. 2008. BDNF splice variants from the second promoter cluster support cell survival of differentiated neuroblastoma upon cytotoxic stress. *Journal of Cell Science* **122**: 156–156.

Batut P, Dobin A, Plessy C, Carninci P, Gingeras TR. 2013. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Research* **23**: 169–180.

Berthelot K, Muldoon M, Rajkowitsch L, Hughes J, McCarthy JEG. 2003. Dynamics and processivity of 40S ribosome scanning on mRNA in yeast. *Molecular Microbiology* **51**: 987–1001.

Calvo SE, Pagliarini DJ, Mootha VK. 2009. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci USA* **106**: 7507–7512.

Carlson M, Botstein D. 1982. Two differentially regulated mRNAs with different 5' ends

- encode secreted with intracellular forms of yeast invertase. *Cell* **28**: 145–154.
- Carninci P, Hayashizaki Y. 1999. High-efficiency full-length cDNA cloning. *Meth Enzymol* **303**: 19–44.
- Chiu MI, Mason TL, Fink GR. 1992. HTS1 encodes both the cytoplasmic and mitochondrial histidyl-tRNA synthetase of *Saccharomyces cerevisiae*: mutations alter the specificity of compartmentation. *Genetics* **132**: 987–1001.
- David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. 2006. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci USA* **103**: 5320–5325.
- Dever TE, Feng L, Wek RC, Cigan AM, Donahue TF, Hinnebusch AG. 1992. Phosphorylation of initiation factor 2 alpha by protein kinase GCN2 mediates gene-specific translational control of GCN4 in yeast. *Cell* **68**: 585–596.
- Gaba A, Jacobson A, Sachs MS. 2005. Ribosome occupancy of the yeast CPA1 upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. *Molecular Cell* **20**: 449–460.
- Gilbert WV, Zhou K, Butler TK, Doudna JA. 2007. Cap-Independent Translation Is Required for Starvation-Induced Differentiation in Yeast. *Science* **317**: 1224–1227.
- Gu W, Lee H-C, Chaves D, Youngman EM, Pazour GJ, Conte D, Mello CC. 2012. CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as *C. elegans* piRNA precursors. *Cell* **151**: 1488–1500.

- Guo H, Ingolia NT, Weissman JS, Bartel DP. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**: 835–840.
- Hahn S, Hoar ET, Guarente L. 1985. Each of three “TATA elements” specifies a subset of the transcription initiation sites at the *CYC-1* promoter of *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **82**: 8562–8566.
- He F, Li X, Spatrick P, Casillo R, Dong S, 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**: 1439–1452.
- Hinnebusch AG. 2005. Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol* **59**: 407–450.
- Huber HE, McCoy JM, Seehra JS, Richardson CC. 1989. Human immunodeficiency virus 1 reverse transcriptase. Template binding, processivity, strand displacement synthesis, and template switching. *J Biol Chem* **264**: 4669–4678.
- Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**: 218–223.
- Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* **147**: 789–802.
- Jackson RJ, Hellen CUT, Pestova TV. 2010. The mechanism of eukaryotic translation

initiation and principles of its regulation. 1–15.

Jang SK, Kräusslich HG, Nicklin MJ, Duke GM, Palmenberg AC, Wimmer E. 1988. A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *J Virol* **62**: 2636–2643.

Katz RA, Skalka AM. 1994. The retroviral enzymes. *Annu Rev Biochem* **63**: 133–173.

Kebaara BW, Atkin AL. 2009. Long 3'-UTRs target wild-type mRNAs for nonsense-mediated mRNA decay in *Saccharomyces cerevisiae*. *Nucleic Acids Res* **37**: 2771–2778.

Kervestin S, Jacobson A. 2012. NMD: a multifaceted response to premature translational termination. *Nat Rev Mol Cell Biol* **13**: 700–712.

Kieft JS, Zhou K, Jubin R, Doudna JA. 2001. Mechanism of ribosome recruitment by hepatitis C IRES RNA. *RNA* **7**: 194–206.

Komar AA. 2005. Internal Ribosome Entry Sites in Cellular mRNAs: Mystery of Their Existence. *Journal of Biological Chemistry* **280**: 23425–23428.

Kozak M. 1991a. A short leader sequence impairs the fidelity of initiation by eukaryotic ribosomes. *Gene Expr* **1**: 111–115.

Kozak M. 1991b. Effects of long 5' leader sequences on initiation by eukaryotic ribosomes in vitro. *Gene Expr* **1**: 117–125.

Kozak M. 1989. The scanning model for translation: an update. *The Journal of Cell Biology* **108**: 229–241.

- Lai M-C, Lee Y-HW, Tam W-Y. 2008. The DEAD-box RNA helicase DDX3 associates with export messenger ribonucleoproteins as well as tip-associated protein and participates in translational control. *Mol Biol Cell* **19**: 3847–3858.
- Law GL, Bickel KS, Mackay VL, Morris DR. 2005. The undertranslated transcriptome reveals widespread translational silencing by alternative 5' transcript leaders. *Genome Biol* **6**: R111.
- Lawless C, Pearson RD, Selley JN, Smirnova JB, Grant CM, Ashe MP, Pavitt GD, Hubbard SJ. 2009. Upstream sequence elements direct post-transcriptional regulation of gene expression under stress conditions in yeast. *BMC Genomics* **10**: 7.
- Legon S. 1976. Characterization of the ribosome-protected regions of 125I-labelled rabbit globin messenger RNA. *J Mol Biol* **106**: 37–53.
- Lomakin IB, Hellen CU, Pestova TV. 2000. Physical association of eukaryotic initiation factor 4G (eIF4G) with eIF4A strongly enhances binding of eIF4G to the internal ribosomal entry site of encephalomyocarditis virus and is required for internal initiation of translation. *Mol Cell Biol* **20**: 6019–6029.
- Marija Cvijović DDEBGJKPS. 2007. Identification of putative regulatory upstream ORFs in the yeast genome using heuristics and evolutionary conservation. *BMC Bioinformatics* **8**: 295.
- Maruyama K, Sugano S. 1994. Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. *Gene* **138**: 171–174.

- Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, Ito T. 2006. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci USA* **103**: 17846–17851.
- Morrish BC, Rumsby MG. 2001. The 5' UTR of Protein Kinase C α Confers Translational Regulation in Vitro and in Vivo. *Biochemical and Biophysical Research Communications* **283**: 1091–1098.
- Muckenthaler M, Gray NK, Hentze MW. 1998. IRP-1 binding to ferritin mRNA prevents the recruitment of the small ribosomal subunit by the cap-binding complex eIF4F. *Molecular Cell* **2**: 383–388.
- Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**: 1344–1349.
- Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J. 2010. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature Publishing Group* **7**: 521–527.
- Oh E, Becker AH, Sandikci A, Huber D, Chaba R, Gloge F, Nichols RJ, Typas A, Gross CA, Kramer G, et al. 2011. Selective Ribosome Profiling Reveals the Cotranslational Chaperone Action of Trigger Factor In Vivo. *Cell* **147**: 1295–1308.
- Paraskeva E, Gray NK, Schläger B, Wehr K, Hentze MW. 1999. Ribosomal pausing and scanning arrest as mechanisms of translational regulation from cap-distal iron-responsive elements. *Mol Cell Biol* **19**: 807–816.

- Pelechano V, Wei W, Steinmetz LM. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**: 127–131.
- Pelletier J, Sonenberg N. 1988. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature* **334**: 320–325.
- Pestova TV, Hellen CU, Shatsky IN. 1996. Canonical eukaryotic initiation factors determine initiation of translation by internal ribosomal entry. *Mol Cell Biol* **16**: 6859–6869.
- Pestova TV, Kolupaeva VG. 2002. The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection. *Genes Dev* **16**: 2906–2922.
- Plessy C, Bertin N, Takahashi H, Simone R, Salimullah M, Lassmann T, Vitezic M, Severin J, Olivarius S, Lazarevic D, et al. 2010. Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nature Publishing Group* **7**: 528–534.
- Polymenis M, Schmidt EV. 1997. Coupling of cell division to cell growth by translational control of the G1 cyclin CLN3 in yeast. *Genes Dev* **11**: 2522–2531.
- Pruunsild P, Kazantseva A, Aid T, Palm K, Timmusk T. 2007. Dissecting the human BDNF locus: Bidirectional transcription, complex splicing, and multiple promoters. *Genomics* **90**: 397–406.
- Pyronnet S, Imataka H, Gingras AC, Fukunaga R, Hunter T, Sonenberg N. 1999. Human

- eukaryotic translation initiation factor 4G (eIF4G) recruits mnk1 to phosphorylate eIF4E. *EMBO J* **18**: 270–279.
- Rasmussen EB, Lis JT. 1993. In vivo transcriptional pausing and cap formation on three *Drosophila* heat shock genes. *Proc Natl Acad Sci USA* **90**: 7923–7927.
- Raught B, Gingras AC, Gygi SP, Imataka H, Morino S, Gradi A, Aebersold R, Sonenberg N. 2000. Serum-stimulated, rapamycin-sensitive phosphorylation sites in the eukaryotic translation initiation factor 4G1. *EMBO J* **19**: 434–444.
- Resch AM, Ogurtsov AY, Rogozin IB, Shabalina SA, Koonin EV. 2009. Evolution of alternative and constitutive regions of mammalian 5'UTRs. *BMC Genomics* **10**: 162.
- Rojas-Duran MF, Gilbert WV. 2012. Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA* **18**: 2299–2305.
- Rosenstiel P, Huse K, Franke A, Hampe J, Reichwald K, Platzer C, Roberts RG, Mathew CG, Platzer M, Schreiber S. 2007. Functional characterization of two novel 5' untranslated exons reveals a complex regulation of NOD2 protein expression. *BMC Genomics* **8**: 472.
- Ruiz-Echevarría MJ, Peltz SW. 2000. The RNA binding protein Pub1 modulates the stability of transcripts containing upstream open reading frames. *Cell* **101**: 741–751.
- Sedman SA, Gelembiuk GW, Mertz JE. 1990. Translation initiation at a downstream AUG occurs with increased efficiency when the upstream AUG is located very close to the 5' cap. *J Virol* **64**: 453–457.

- Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, et al. 2003. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci USA* **100**: 15776–15781.
- Shire D, Carillon C, Kaghad M, Calandra B, Rinaldi-Carmona M, Le Fur G, Caput D, Ferrara P. 1995. An amino-terminal variant of the central cannabinoid receptor resulting from alternative splicing. *J Biol Chem* **270**: 3726–3731.
- Smith L, Brannan RA, Hanby AM, Shaaban AM, Verghese ET, Peter MB, Pollock S, Satheesha S, Szykiewicz M, Speirs V, et al. 2009. Differential regulation of oestrogen receptor β isoforms by 5' untranslated regions in cancer. *Journal of Cellular and Molecular Medicine* **14**: 2172–2184.
- Smith L, Coleman LJ, Cummings M, Satheesha S, Shaw SO, Speirs V, Hughes TA. 2010. Expression of oestrogen receptor β isoforms is regulated by transcriptional and post-transcriptional mechanisms. *Biochem J* **429**: 283–290.
- Stadler M, Fire A. 2011. Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA* **17**: 2063–2073.
- Tsai N-P, Bi J, Wei L-N. 2007. The adaptor Grb7 links netrin-1 signaling to regulation of mRNA translation. *EMBO J* **26**: 1522–1531.
- van den Heuvel JJ, Bergkamp RJ, Planta RJ, Raué HA. 1989. Effect of deletions in the 5'-noncoding region on the translational efficiency of phosphoglycerate kinase mRNA in yeast. *Gene* **79**: 83–95.

- Vilela C, Linz B, Rodrigues-Pousada C, McCarthy JE. 1998. The yeast transcription factor genes YAP1 and YAP2 are subject to differential control at the levels of both translation and mRNA stability. *Nucleic Acids Res* **26**: 1150–1159.
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476.
- Wang Z, Gaba A, Sachs MS. 1999. A highly conserved mechanism of regulated ribosome stalling mediated by fungal arginine attenuator peptides that appears independent of the charging status of arginyl-tRNAs. *J Biol Chem* **274**: 37565–37574.
- Willis AE. 1999. Translational control of growth factor and proto-oncogene expression. *Int J Biochem Cell Biol* **31**: 73–86.
- Wilson JE, Pestova TV, Hellen CU, Sarnow P. 2000. Initiation of protein synthesis from the A site of the ribosome. *Cell* **102**: 511–520.
- Wu M, Tzagoloff A. 1987. Mitochondrial and cytoplasmic fumarases in *Saccharomyces cerevisiae* are encoded by a single nuclear gene FUM1. *J Biol Chem* **262**: 12275–12282.
- Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**: 1033–1037.
- Zhang Z, Dietrich FS. 2005. Mapping of transcription start sites in *Saccharomyces*

cerevisiae using 5' SAGE. *Nucleic Acids Res* **33**: 2838–2851.

Chapter 2

Roles for Transcript Leaders in Translation and mRNA Decay Revealed by Transcript

Leader Sequencing*

*This research was originally published in Genome Research and has been edited for presentation here. Arribere JA, Gilbert WV. Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. Genome Res. 2013 Jun;23(6):977-87. doi: 10.1101/gr.150342.112

Abstract

Transcript leaders (TLs) can have profound effects on mRNA translation and stability. To map TL boundaries genome-wide, we developed TL-sequencing (TL-seq), a technique combining enzymatic capture of m⁷G-capped mRNA 5'-ends with high-throughput sequencing. TL-seq identified mRNA start sites for the majority of yeast genes and revealed many examples of intragenic TL heterogeneity. Surprisingly, TL-seq identified transcription initiation sites within 6% of protein-coding regions, and these sites were concentrated near the 5' ends of ORFs. Furthermore, ribosome density analysis showed these truncated mRNAs are translated. Translation Associated TL-seq (TATL-seq), which combines TL-seq with polysome fractionation, enabled annotation of TLs and simultaneously assayed their function in translation. Using TATL-seq to address relationships between TL features and translation of the downstream ORF, we observed that upstream AUGs (uAUGs), and no other upstream codons, were associated with poor translation and nonsense-mediated mRNA decay (NMD). We also identified hundreds of genes with very short TLs, and demonstrated that short TLs were associated with poor translation initiation at the annotated start codon, and increased initiation at downstream AUGs. This frequently resulted in out-of-frame translation and subsequent termination at premature termination codons, culminating in NMD of the transcript. Unlike previous approaches, our technique enabled observation of alternative TL variants for hundreds of genes, and revealed significant differences in translation in genes with distinct TL isoforms. TL-seq and TATL-seq are useful tools for annotation and functional characterization of TLs, and can be applied to any eukaryotic system to investigate TL-mediated regulation of gene expression.

Introduction

Regulation of gene expression controls cellular fate and fitness. Post-transcriptional regulation of messenger RNAs (mRNAs) can have large effects on gene expression (via protein output) by modulating mRNA translation, stability and localization. For example, gene-specific translational efficiencies vary over 100-fold genome-wide (Ingolia et al. 2009), and mRNA half-lives range from a few minutes to many hours. Despite the increasing evidence for pervasive post-transcriptional regulation of gene expression in eukaryotes, most genome-wide studies to date have focused on transcriptional aspects of gene expression. Quantitative genome-scale assays for post-transcriptional control mechanisms are beginning to transform our understanding of the regulation of gene expression, but are not yet able to capture several important aspects.

Post-transcriptional regulation of gene expression is largely governed by features of the noncoding portions of mRNA, both downstream (3'UTR and poly(A) tail) and upstream (TL or 5'UTR) of the open reading frame (ORF). Because some TLs contain upstream ORFs (uORFs) that are translated, it is more accurate to refer to "5'UTRs" as TLs. TLs are particularly important for translation initiation. During translation initiation in eukaryotes, a cap-binding complex (eIF4F) binds to the TL via the 5' methyl-7-guanosine (m^7G) cap and facilitates recruitment of a small ribosomal subunit and its associated eukaryotic initiation factors (eIFs) to form a pre-initiation complex (PIC). The PIC scans in a net 5' to 3' direction until it locates a start codon (AUG), triggering complex rearrangements that eventually result in formation of an elongating 80S ribosome (reviewed in (Jackson et al. 2010)). Scanning PICs can be captured by upstream

AUGs (uAUGs), leading to decreased initiation from the main protein-coding ORF. A few uAUGs have well-characterized translational regulatory functions, including those found in the TLs of the stress-responsive transcription factors *GCN4* and *ATF4* (reviewed in (Hinnebusch 2005)). Other TLs allow specific genes to be efficiently translated under conditions of widespread translational inhibition (Gilbert et al. 2007). Although most examples of TL-mediated translational control come from small-scale studies, recent developments in genome-wide technologies have revealed widespread post-transcriptional regulation by TLs (Calvo et al. 2009; Thoreen et al. 2012).

Understanding the full range of TLs' impact on post-transcriptional regulation of gene expression will require accurate, genome-wide annotations. Previous efforts to define TLs in yeast include full-length cDNA sequencing (Miura et al. 2006), 5' Serial Analysis of Gene Expression (5'SAGE (Zhang and Dietrich 2005)), and computational identification of transcript boundaries from measurements using tiling microarrays (Xu et al. 2009) or RNA-seq (Nagalakshmi et al. 2008). Importantly, the latter two approaches limited each gene to only one TL, whereas the first two approaches observed widespread TL heterogeneity. More than 99% of genes analyzed by (Miura et al. 2006) and 95% of genes in (Zhang and Dietrich 2005) had more than one TL. Such heterogeneity is consistent with studies of individual genes (Hahn et al. 1985), indicating that one TL per gene is an oversimplification in many cases.

Here we introduce a method to study TLs on a genomic scale, TL-seq, and demonstrate its utility in *S. cerevisiae*, identifying one or more TLs for the majority of genes. Surprisingly, we observed hundreds of genes with very short TLs, and showed that this feature leads to initiation at downstream AUGs, often culminating in nonsense-

mediated mRNA decay (NMD). Of TLs identified by TL-seq, <15% contained at least one upstream AUG (uAUG), significantly fewer than expected by chance. When TLs contain uAUGs, they tend to be conserved, reduce translation, and target the transcript for NMD. In addition, we determined the extent of intragenic TL heterogeneity and identified many new examples in yeast, including ORF-internal transcription start sites (TSSs) that may produce alternative protein variants. Finally, using TATL-seq, we identified hundreds of cases where one gene encodes multiple TL isoforms, and showed that the majority of these variants are associated with distinct translational activities *in vivo*.

Materials and Methods

Yeast strains and Growth Conditions

Yeast cultures (Sigma 1278b *MATa ura3 leu2 trp1 his3* and BY4742 *Mata his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*) were grown to mid-log (OD₆₀₀~0.5-1.0) phase in YPAD (1% yeast extract, 2% peptone, 0.01% adenine hemisulfate, 2% glucose) at 30°C in flasks with vigorous shaking.

RNA Isolation and Polysome Gradient Fractionation

Total RNA was isolated from yeast cell pellets by hot phenol extraction as described (Clarkson et al. 2010). Polysome gradients were prepared and RNA was extracted from gradient fractions as described (Arribere et al. 2011) (also see Appendix III). For Northern blots, RNA was separated on 1.1% agarose, 6% formaldehyde gels and blotted as previously described (Carlile and Amon 2008). Primer sequences for probes are given in Table 2.2. qPCR was performed as previously described (Arribere et al. 2011) (also

see Appendix III). Fold change over RNA levels in mutant strains was determined by first normalizing RNA levels to 18S rRNA, then dividing by this same quantity from wild type yeast.

Transcript Leader Sequencing

Polyadenylated mRNA (oligo dT cellulose purified as described (Sambrook et al. 2001) was fragmented by alkaline hydrolysis. RNA fragments of ~50-80 nts were gel purified and dephosphorylated with 30 U Calf-Intestinal Phosphatase (CIP, NEB) in 50 μ l reactions at 37°C for 60 min, followed by phenol:chloroform extraction and isopropanol precipitation. Purified CIP-treated fragments were treated with 25 U Tobacco Acid Pyrophosphatase (TAP, Epicentre) in 50 μ l at 37°C for two hours, then precipitated. Next a 5'RNA adaptor was added via ligation in a 20 μ l reaction with 20 U of T4 RNA Ligase (NEB) for one hour at 37°C. Gel purification of a higher molecular weight species yielded the ligated RNA, which was then 3'end captured via poly(A) tailing (TL-seq, as previously described in (Ingolia et al. 2009)) or ligation with preadenylated adaptor (TL-seq biological replicates and TATL-seq, as previously described in (Mayr and Bartel 2009)). cDNA was prepared from ligated RNA (Superscript III, Invitrogen), amplified by 10-12 cycles of PCR (Phusion, Finnzymes), and sequenced on an Illumina Genome Analyzer II (TL-seq and TATL-seq) or HiSeq (TL-seq replicates).

The computationally pooled TATL-seq libraries were used for Figures 2.1, 2.3-2.15, as these libraries gave more data for more genes. Analyses gave similar results using TL-seq or pooled TATL-seq data.

Peak-Calling Algorithm

The peak-calling algorithm was developed with modifications from (Johnson et al. 2007). For each gene an expected background density of reads at a given nucleotide assuming a uniform distribution throughout the feature is

$$\lambda = \frac{N}{L}$$

where N is the total number of reads mapping to that feature (including up- and downstream boundaries), and L is the total length of the feature (including up- and downstream boundaries). The observed read density, x , at each position was calculated by scanning along the feature using an n -nucleotide window. For analyses here, $n=50$ nt; using smaller n yielded a higher fraction of artifactual peaks while larger n yielded fewer peaks overall. If a window contained more than 5 times the expected number of reads (that is $\geq 5n\lambda$), a p-value for the enrichment was calculated based on the Poisson distribution:

$$F(x) = \frac{(n\lambda)^x e^{-n\lambda}}{x!}$$

Consecutive windows of enrichment ($p < 0.01$) $\geq n$ nucleotides in length were defined as a peak. The exact nucleotide position of the peak was defined as the mode read in that region.

Monoppeak TL-seq genes (used in Figures 2.7, 2.10B-D, 2.2C, 2.11B,C) were those genes with exactly one TL-seq peak upstream of the annotated ORF start codon (2,619 out of 3,434 total ORFs with TL-seq peaks).

uAUG analysis

The expected frequency of uAUGs in TLs was determined by randomizing TL lengths between genes and calculating the fraction of uAUG-containing TLs for 10,000 randomizations. Alternatively, individual TL sequences were shuffled preserving di- or mononucleotide frequency within a given TL sequence using the algorithm described in (Altschul and Erickson 1985). P-values were calculated using a z-statistic approximating a normal distribution from the randomizations.

For conservation analysis, TL positions were included if an ungapped genomic alignment existed amongst four yeast species: *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus*. A trinucleotide within a *S. cerevisiae* TL was deemed “conserved” if the same trinucleotide was present at the same position in the three other yeast species. A trinucleotide was deemed “non-conserved” if one or more yeast species contained a mutation anywhere in that trinucleotide at the aligned position. For each trinucleotide, the frequency of conservation was defined as the number of conserved instances divided by the total number of eligible instances of that trinucleotide (conserved plus non-conserved).

Shape Index

Shape Index (as defined by (Hoskins et al. 2011)) was used to quantify TL heterogeneity within genes and is defined as

$$SI = \sum_{i|f_i \neq 0} f_i \log f_i$$

where f_i is the frequency of reads from a given nucleotide i . For Gene Ontology analysis, SI of all genes within a given GO category (SGD annotations) was compared to

the overall distribution of SIs, and GO categories with a $p < 0.001$ (Mann Whitney U Bonferroni-corrected p-value) were deemed significantly different.

TATL-seq

Polysome gradients were fractionated and RNA was collected from each of seven fractions. Purified RNA was poly(A) selected and fragmented, as per the TL-seq protocol. To quantify individual TLs across a polysome gradient, peak-calling was performed on the computationally pooled TATL-seq libraries. The abundance of each peak in each fraction was then quantified by calculating the density of reads (RPKM). The relative abundance of a TL in a fraction is equal to its read density in that fraction divided by its read density over all fractions. Specifically, the relative abundance of a TL x in fraction i was defined as

$$\frac{\frac{x_i}{F_i}}{\frac{\sum_{i=1}^7 x_i}{\sum_{i=1}^7 F_i}}$$

where x_i is the number of reads mapping to the TL and F_i is the library size of fraction i .

Read Assignment

The yeast genome and annotations were downloaded from the Saccharomyces Genome Database (yeastgenome.org) on May 26, 2010. 35 nt reads were mapped to the genome using Bowtie (Langmead et al. 2009) allowing for three mismatches. For multiply mapping reads, read counts were divided by the total number of places where they mapped. The majority of reads mapped intergenically; for annotation purposes reads were assigned to the nearest protein-coding gene. The intergenic distance between ORF

boundaries irrespective of strand was bisected, and reads mapping up- or downstream of this boundary were assigned to the up- or downstream gene, respectively. Reads in the sense orientation mapping within a feature's boundaries were assigned to that gene, and antisense-oriented reads (~9% of TL-Seq library) were ignored for the analyses shown here.

Peak RPKM

To determine peak abundances (Fig 2.2 and 2.9), RPKM was calculated. The numerator ("RPK") was the number of reads per kilobase of peak (50 nt or 0.05 for most peaks) and the denominator ("M") was the total number of mRNA-mapping reads. Analyses were also performed with "M" as peak-mapping reads and gave similar results.

Ribosome Footprint Profiling

To identify AUG initiation codon peaks, Ribo-seq was performed on Sigma 1278b yeast starved for glucose for 3 hours and incubated with cycloheximide *in vivo*. Ribo-seq and analysis was performed as previously described (Ingolia et al. 2009). For metagene plots, 5' ends of reads were offset by 12 nts to indicate the position corresponding to the ribosomal P-site. Gene-specific translation efficiency was defined as \log_2 of the ratio of footprint read density (footprint reads per kilobase per million (RPKM)) to RNA-seq read density (total RPKM) for all genes with at least 64 reads in each library, excluding the first 8 codons of the ORF. Excluding the first 8 codons yields the most reproducible Ribo-seq density; the results shown here are the same if the first 8 codons are included in

the calculation. For all analyses of published Ribo-seq data, translation efficiency was recalculated according to this formula.

Nucleosome Analysis

Each nucleosome was modeled with a Gaussian probability density function using the positions and parameters (nucleosome intensity, standard deviation) given from nucleosome ChIP-seq (Mavrich et al. 2008). An average nucleosome distribution was calculated from all TL-seq peaks upstream of ORFs. For each TL-seq peak, a nucleosome signature score was defined by comparing that peak's nucleosome distribution to the average distribution for all genes and quantitated with Spearman's ρ .

Comparison with other TSS annotations

Transcript boundaries from computational analysis of tiling microarray arrays were compared to their respective genome annotations to determine TL lengths (David et al. 2006; Xu et al. 2009). TL lengths were compared among these datasets and correlation was determined.

Peak Filters

To avoid spurious conclusions from artifactual peaks, peaks were subject to an *in silico* subtractive peak approach or filtered by nucleosome signature. For the *in silico* subtractive peak approach, the peak-calling algorithm was applied to an RNA-seq library prepared with T4 RNA ligase (Geisler et al. 2012). Any TL-seq peak that overlapped with an RNA-seq peak was removed from further analyses. For the nucleosome signature

approach, each TL-seq peak's surrounding nucleosome density (from (Mavrich et al. 2008)) was compared to the average nucleosome distribution of all TL-seq peaks upstream of an ORF (shown in Figure 2.2A). Spearman's ρ was used to assess the degree to which a peak's nucleosome density agreed with the genomic average. All of these approaches preferentially removed ORF-internal TL-seq peaks. Analyses shown were repeated with *in silico* subtracted TL-seq peaks or all peaks passing a nucleosome signature score $\rho \geq 0.1$, 0.3, or 0.5; all results were qualitatively the same. All analyses shown were of unfiltered TL-seq peaks.

Luciferase Reporter Assays

To assess the translation activity of an individual TL, we created an inducible system for assaying translation *in vivo*. pRS415 with the *GALI* promoter and *CYCI* terminator was obtained from (Mumberg et al. 1994). A NotI site was introduced upstream of the *CYCI* terminator and a BglIII site was inserted at the TSS of the *GALI* promoter by site-directed mutagenesis. Firefly luciferase was inserted as a HindIII-NotI fragment, thus generating the vector pWG554. Each TL was PCR amplified from genomic DNA and cloned as a BamHI-NcoI fragment into pWG554. All constructs were confirmed by restriction analysis and sequencing. Plasmids were introduced into BY4742 by LiAc transformation to make the strains listed in Table 2.1.

Strains were grown overnight in SC -Leu, 2% Raffinose, 2% Galactose. At OD₆₀₀=0.5-1.0, cells were harvested by centrifugation and flash frozen. Lysates were prepared by vortexing with glass beads in cold lysis buffer (1x PBS, 1mM PMSF, +protease inhibitors), and clarified by centrifugation for 30" at 16,000 RCF, followed by

5' at 16,000 RCF at 4°C. The supernatant was removed to a new tube and luciferase activity was assayed with Bright-Glo (Promega) in a Centro XS Luminescence Microplate Reader (Berthold Technologies). Luciferase was normalized to lysate concentration based on absorbance at 260nm. RNA was extracted from lysate by phenol-chloroform and *Fluc* mRNA levels normalized to 18S rRNA levels were determined by RT-qPCR .

Normalized Read Density

At a given position in a gene, the normalized read density was defined as the read count at that position divided by the read density across the whole gene. Read density was defined as the total number of reads mapping to that gene divided by the length of that gene. For a group of genes, the normalized read density is simply the average of all individual genes in that group.

Table 2.1: Strains Used For In Vivo Translation

<i>E. coli</i> strain	Plasmid
	pRS415GAL1 pGAL1-CYC1 <i>LEU</i>
BWG554	pRS415GAL1 pGAL1-FLUC-CYC1 <i>LEU</i>
BWG557	pWG554 pGAL-tl <i>CMP2</i> [-63,-3]-FLUC-CYC1
BWG558	pWG554 pGAL-tl <i>CMP2</i> [-125,-3]-FLUC-CYC1
BWG559	pWG554 pGAL-tl <i>KAP95</i> [-138,-3]-FLUC-CYC1
BWG560	pWG554 pGAL-tl <i>KAP95</i> [-202,-3]-FLUC-CYC1
BWG561	pWG554 pGAL-tl <i>PIRI</i> [-104,-3]-FLUC-CYC1
BWG562	pWG554 pGAL-tl <i>PIRI</i> [-383,-3]-FLUC-CYC1
BWG563	pWG554 pGAL-tl <i>CRZI</i> [-120,-3]-FLUC-CYC1
BWG564	pWG554 pGAL-tl <i>CRZI</i> [-330,-3]-FLUC-CYC1
BWG565	pWG554 pGAL-tl <i>YDR089W</i> [-46,-3]-FLUC-CYC1
BWG566	pWG554 pGAL-tl <i>YDR089W</i> [-191,-3]-FLUC-CYC1
BWG567	pWG554 pGAL-tl <i>YDL129W</i> [-37,-3]-FLUC-CYC1

BWG568	pWG554 pGAL-tl <i>YDL129W</i> [-163,-3]-FLUC-CYC1
BWG569	pWG554 pGAL-tl <i>AGPI</i> [-46,-3]-FLUC-CYC1
BWG570	pWG554 pGAL-tl <i>AGPI</i> [-98,-3]-FLUC-CYC1
BWG571	pWG554 pGAL-tl <i>AGPI</i> [-281,-3]-FLUC-CYC1

Yeast Strain	Relevant Genotype
YWG11	BY4742 Mat α <i>his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0</i>
YWG733	YWG11 transformed with pWG557
YWG734	YWG11 transformed with pWG558
YWG735	YWG11 transformed with pWG559
YWG736	YWG11 transformed with pWG560
YWG737	YWG11 transformed with pWG561
YWG738	YWG11 transformed with pWG562
YWG739	YWG11 transformed with pWG563
YWG740	YWG11 transformed with pWG564
YWG741	YWG11 transformed with pWG565
YWG742	YWG11 transformed with pWG566
YWG743	YWG11 transformed with pWG567
YWG744	YWG11 transformed with pWG568
YWG745	YWG11 transformed with pWG569
YWG746	YWG11 transformed with pWG570
YWG747	YWG11 transformed with pWG571

Table 2.2: Oligonucleotides Used In This Study

Name	Purpose	Oligonucleotide Sequence
<i>YKR078W</i>	Northern	AAGTAGGTGCACTTGAACCAAAGG
<i>EPL1</i>	Northern	CAGTCTGGTGATATAGTCCCTACG
<i>PDR10</i>	Northern	AGTTCCTGGTGCAGCACCTACTAT
TLSeqAdaptor1	TL-seq, TATL-seq	AAUGAUACGGCGACCACCGACAGGUUCAGA GUUCUACAGUCCGACG
TLSeqAdaptor2	TL-seq Replicates	AAUGAUACGGCGACCACCGACAGGUUCAGA GUUCUACAGUCCGACGAUC

Results

Defining Transcript Leaders

To facilitate identification of TLs on a genomic scale, we developed TL-seq, an adaptation of 5' RACE for deep-sequencing platforms. TL-seq takes advantage of the unique m⁷G-protected 5'-5' triphosphate linkage in the mRNA cap to biochemically

distinguish and physically separate mRNA 5' ends from other RNA species (Figure 2.1A). Fragmented and size-selected RNA is first treated with a phosphatase, reducing the majority of 5'RNA ends to hydroxyl groups, leaving m⁷G-capped 5' ends intact. Subsequent treatment with a pyrophosphatase cleaves two of the three phosphates from the cap, yielding a 5' monophosphorylated RNA. This RNA species is a substrate for RNA ligase, enabling selective ligation to the 5' monophosphate (formerly capped) RNA fragments and not the 5' hydroxyl fragments. 5' adaptor ligation causes an increase in RNA fragment length that can be resolved by PAGE, enabling purification (Figure 2.1B). Finally, these ligated RNA fragments are converted to DNA and deep-sequenced using standard techniques.

As anticipated, TL-seq functioned in a strongly pyrophosphatase-dependent manner. Omission of pyrophosphatase from the enzymatic steps substantially decreased the yield of ligated material (Figure 2.1B). Pyrophosphatase treatment resulted in an enrichment of reads upstream of annotated yeast ORFs compared to the untreated sample (Figure 2.1C). Metagene analysis revealed a pyrophosphatase-dependent density of reads directly upstream of the start codon of ORFs. The maximum of this read density is 25-30 nt upstream of annotated ORF start codons (Figure 2.1D), on par with previous estimates of TL lengths in yeast (Miura et al. 2006; Nagalakshmi et al. 2008).

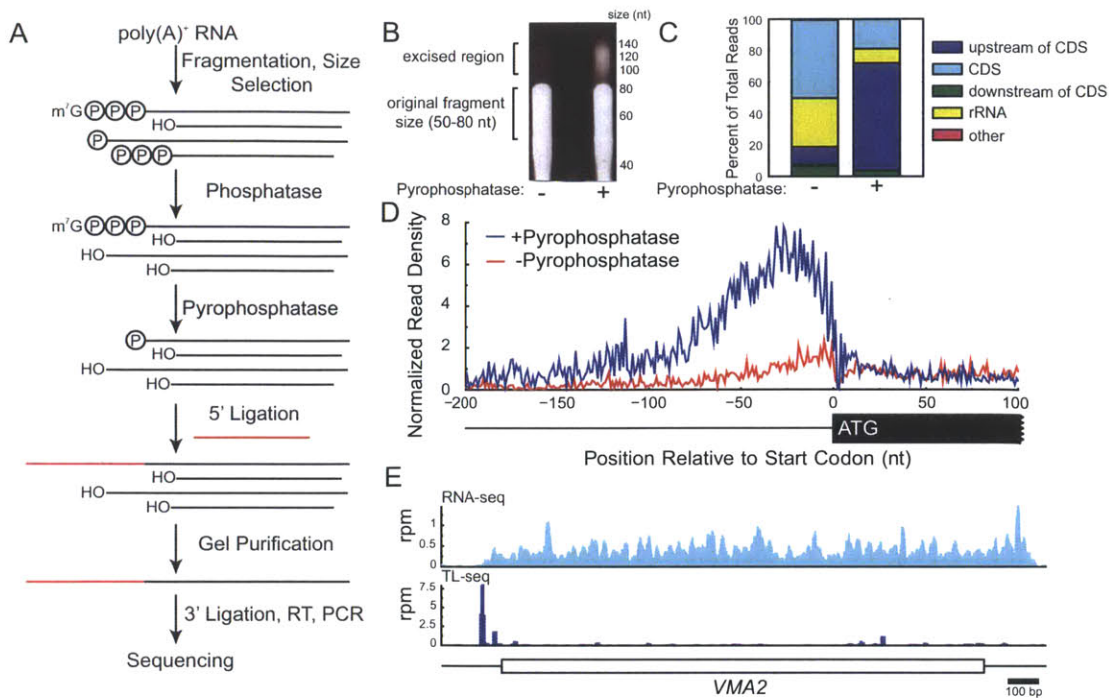


Figure 2.1: TL-seq Preferentially Recovers Capped 5' Ends

(A) Schematic of TL-seq.

(B) Fragmented RNA (50-80nt) was treated or mock-treated with pyrophosphatase. Subsequently both reactions were treated with RNA ligase and a 45nt adaptor. The gel is overexposed to visualize the shift (bracket). Size markers for a DNA ladder are indicated to the right of the gel.

(C) Distribution of where 5' ends of reads map with and without inclusion of pyrophosphatase.

(D) Genes were aligned by their annotated translation start codon and the distribution of reads calculated with or without pyrophosphatase.

(E) Comparison of RNA-seq and TL-seq profiles for *VMA2*. Ordinate is in reads per million (rpm), scale bar at lower right.

While 68% of reads mapped upstream of ORFs, others mapped to presumably uncapped transcripts, (e.g. rRNA, 9%), indicating the presence of non-TSS-generated background reads in the library. To increase the signal-to-noise ratio in our TL data, we implemented a peak-calling algorithm that, for each gene, defines an expected background distribution and then identifies regions with significantly higher read density

than expected. This method of computationally filtering background reads is analogous to methods commonly used to identify sites of significant enrichment in ChIP-seq data (Johnson et al. 2007). The identified regions of high read density (peaks) were then analyzed.

Peaks upstream of ORFs exhibited known TSS characteristics. There is a stereotypical nucleosome distribution about some TSSs in yeast (Yuan et al. 2005). Comparing TSSs predicted by TL-seq with genome-scale nucleosome density maps showed the expected features of this characteristic distribution, including periodic placement of nucleosomes both up- and downstream of the TSS as well as a 5'-Nucleosome Free Region (5'NFR) (Figure 2.2A). In addition, a peak of Reb1 binding sites ~100nt upstream of TL-seq peaks was apparent and consistent with reports of a fixed-distance relationship between binding of this transcription factor and TSSs in yeast (Figure 2.2B, (Koerber et al. 2009; Rhee and Pugh 2011)). For 2,619 ORFs TL-seq identified a single peak upstream of the annotated AUG. The TL annotations for these monopeak genes were highly reproducible in both length and abundance between biological replicates (Spearman's $\rho \sim 0.94$, $\rho \sim 0.98$, respectively, Figure 2.2C) and in good agreement with TL lengths determined by high-density tiling array analysis (Spearman's $\rho \sim 0.6$, Figure 2.2D) (Xu et al. 2009). Unlike tiling array methods, which cannot reliably distinguish alternative TSSs, TL-seq readily identified genes with at least two distinct TL peaks upstream of their ORF, which included 6.3% of genes. This number should be treated as a lower bound as it requires that intragenic TL variants be spaced >50 nt apart, a limitation imposed by the peak-calling algorithm. By simply examining read abundance, >99% of genes had reads from more than one position

upstream of the CDS, most separated by only a few nucleotides. Collectively, these data show that TL-seq and an associated peak-calling algorithm are useful tools for genome-scale identification of TSSs in eukaryotic cells.

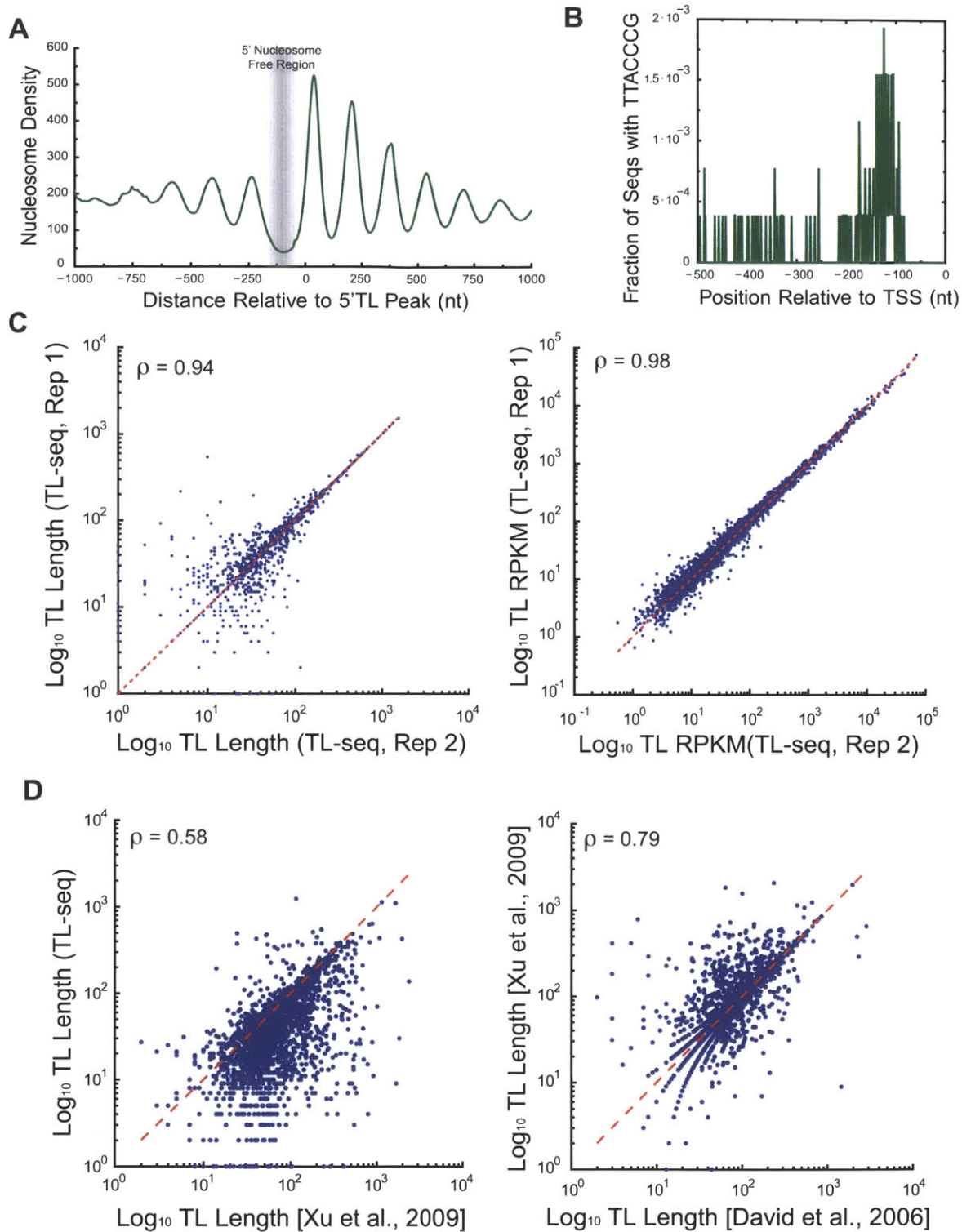


Figure 2.2: TL-seq Produces TSS Annotations

(A) TL-seq peaks show nucleosome distributions expected for TSSs. Average nucleosome distribution about TL-seq peaks upstream of ORFs, where each nucleosome was modeled as a Gaussian according to parameters and positions given in (Mavrich et al.

2008).

(B) Reb1 binding sites relative to TL-seq peaks upstream of ORFs. Ordinate shows the fraction of all TL-seq peaks containing a match to the consensus Reb1p motif TTACCCG.

(C) TL-seq TL lengths and abundances are correlated between biological replicates. TL-seq was performed on poly(A)-selected RNA from two independent liquid cultures grown in parallel. Results show TL lengths annotated by TL-seq (left), and those TL peaks' abundances (right).

(D) TL-seq TL lengths are correlated with other annotations. Spearman's ρ is indicated on each graph for TL lengths annotated by TL-seq or tiling microarray (Xu et al. 2009; David et al. 2006). The left graph shows the line $y=x$, demonstrating that TL-seq tends to call shorter TLs whereas (Xu et al. 2009) tend to call longer TLs. Notably, the correlation between TL lengths determined by different methods was only slightly less than the reproducibility of TL lengths from two studies using the same tiling microarray approach.

Transcription Initiation Within ORFs

Unexpectedly, 41% of TL-seq peaks mapped within ORF boundaries. While such internal peaks were generally of lower abundance and significance (p-value from peak-calling), they persisted as a substantial fraction of peaks even at more stringent significance cutoffs (Figure 2.3A). In bulk, these internal peaks showed a nucleosome signature with the same characteristics as canonical TSSs, including a periodicity both up- and downstream and a 5' NFR although this signal was decreased compared to 5' TL peaks (Figure 2.3B). Using peak-specific nucleosome signature scores as an orthogonal metric of TSSs, we estimate that 6% of TSSs are internal to ORF boundaries (Figure 2.3C, Methods), and this number is in good agreement with previous observations of the fraction of internal TSSs (6%) estimated from a large-scale cDNA sequencing approach (Miura et al. 2006). Furthermore, there was significant overlap between internal-TSS-containing genes ($p=2.9 \times 10^{-22}$, Fisher's exact test) as well as internal TSS positions (Table 2.3) identified by the two studies. This overlap is notably high, given that the previous study (Miura et al. 2006) pooled cDNAs sequenced from yeast in rich media

and undergoing meiosis. Thus, our results support the existence of a substantial number of internal peaks within ORFs.

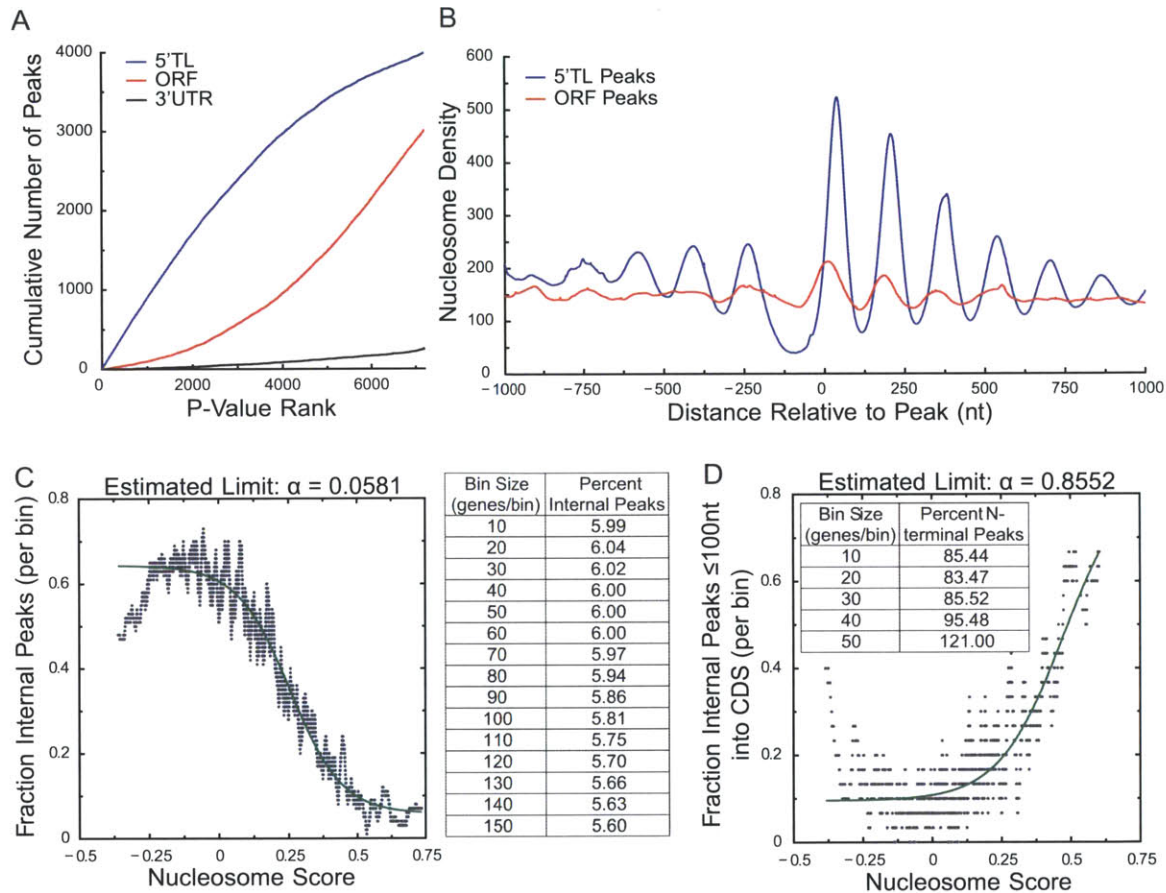


Figure 2.3: TL-seq Detects Internal Peaks with a TSS-like Nucleosome Distribution

(A) The cumulative number of peaks in each class (5'TL/ORF/3'UTR) versus p-value rank from peak-calling algorithm is shown. Even at stringent p-value cutoffs (left side of graph), a substantial number of internal peaks persist.

(B) Internal peaks exhibit a nucleosome signature characteristic of TSSs. The average nucleosome signature of groups of peaks (upstream of genes or ORF-internal) is shown.

(C) Nucleosome score plotted against fraction of internal peaks, binned by nucleosome score using a bin size of 100 peaks. For each TL-seq peak, a nucleosome score was defined according to that peak's nucleosome distribution's concordance with the genome-wide average (Spearman's ρ). Under the assumption that TL-seq peaks with a higher nucleosome signature score are more likely to represent true TSSs, we examined the distribution of peaks at different scores. The "true" fraction of internal peaks is the limit as the data approach a nucleosome signature score of 1, where the fraction of true positives approaches one. At high nucleosome scores (e.g. ≥ 0.5), we observed that the fraction of internal peaks levels off at ~6%. A sigmoid function was fit to the data and the limit inferred. Table shows the same calculation performed at different bin sizes,

demonstrating that the limit is not a function of the arbitrary bin size.

(D) Nucleosome score plotted against fraction of all internal peaks that are N-terminal (within 100 nts of the 5' boundary of an ORF). Analysis similar to part (C), but with the ordinate as fraction of all internal peaks being N-terminal. Table shows results for different bin sizes. Due to the paucity of data at high nucleosome signature scores, the function yielded a poor fit and anomalously high percentages (bin sizes 50 and above). Thus the fraction of true internal peaks that are ≤ 100 nt into the ORF is 85-100%.

Gene	Empirical P-value	Miura TSS	TL-seq TSS	Distance (nt)
<i>YNR008W</i>	0.0057	1672	1678	6
<i>YML021C</i>	0.0087	21	26	5
<i>YNL284C</i>	0.0183	42	51	9
<i>YPL108W</i>	0.0490	27	14	13
<i>YGL212W</i>	0.0386	27	8	19
<i>YGR079W</i>	0.0209	105	117	12
<i>YML126C</i>	0.0022	65	67	2
<i>YML050W</i>	0.0473	35	12	23
<i>YLR092W</i>	0.0235	1549	1581	32
<i>YKL219W</i>	0.0199	59	72	13
<i>YIL120W</i>	0.0270	859	836	23
<i>YIL064W</i>	0.0015	52	51	1
<i>YAL059W</i>	0.0015	59	58	1
<i>YGR233C</i>	0.0192	3448	3414	34
<i>YNL216W</i>	0.0182	12	35	23
<i>YHL006C</i>	0.0416	361	371	10
<i>YNL240C</i>	0.0007	1344	1343	1
<i>YAR031W</i>	0.0224	151	162	11
<i>YJL060W</i>	0.0023	8	10	2
<i>YKL022C</i>	0.0126	102	86	16
<i>YJR074W</i>	0.0259	91	82	9

Table 2.3: Genes With Similar Internal TSSs Between TL-seq and (Miura et al. 2006)

To assess whether a gene contained closely spaced TSSs between the two studies, all genes with a single internal TSS in both studies were identified. The distance between those two TSSs was compared to 100,000 randomly chosen pairs of points in the ORF. P-value indicates the fraction of randomly chosen pairs whose distance was less than the observed distance from the two studies.

The internal peaks were recovered from a protocol that enriches for RNA species containing both poly(A) tails and m⁷G caps, the hallmarks of translatable eukaryotic mRNAs. To assay for translation initiation on these 5' truncated transcripts, we examined ribosome distributions on internal TL peak-containing ORFs by ribosome footprint

profiling (Ribo-seq). Ribo-seq uses deep sequencing of ribosome-protected mRNA fragments to reveal the precise positions of mRNA-associated ribosomes (Ingolia et al. 2009). Ribo-seq of glucose-starved yeast shows a high density of ribosomes at the initiation codon of all genes (Figure 2.4A), but not at internal AUGs (Figure 2.4A, inset), thus providing a tool for the identification of the sites of initiating ribosomes (Vaidyanathan et. al., manuscript in preparation). Consistent with the internal TL-seq peaks existing as the 5' ends of translated mRNAs, ribosome footprint density was elevated at the first downstream AUG, regardless of whether it is in-frame or out-of-frame to the annotated full-length ORF (Figure 2.4B, C). For those internal transcripts where the predicted first encountered AUG is out-of-frame, the density of ribosomes at this AUG was much higher than the first in-frame AUG (Figure 2.4C, inset), as expected from the 5' to 3' scanning model of translation initiation.

Closer examination of the putative internal transcripts on a gene-by-gene basis identified four broad categories of internal TL peaks. The first category was that of 5'-misannotation (Figure 2.4D). In these cases, the annotated translation initiation codon is likely incorrect, as RNA-seq and Ribo-seq data, as well as decreased amino acid conservation corroborate the TL-seq annotation of a predominant TSS 3' to the annotated translation start site. A second category was that of extreme N-terminal peaks in which the major peak identified by TL-seq mapped within the first 100 bases of the ORF. This class includes genes that are known to produce internal transcripts generating N-terminal protein variants, such as *KAR4* (Figure 2.4E, (Gammie et al. 1999), *FUM1* (Wu and Tzagoloff 1987), and *HTSI* (Chiu et al. 1992). The majority ($\geq 85\%$) of internal TSSs were N-terminal (≤ 100 nt into an ORF) (Figure 2.3D). A third potentially interesting

class of internal peaks mapped to loci that appear to generate much shorter distinct second transcripts (Figure 2.4F). Northern analysis for three genes in this category revealed multiple transcripts of distinct sizes (Figure 2.5). A fourth class of ORF-internal TL-seq peaks could be attributed to RNA-ligase dependent capture bias (Figure 2.6) and were subsequently filtered (see Methods).

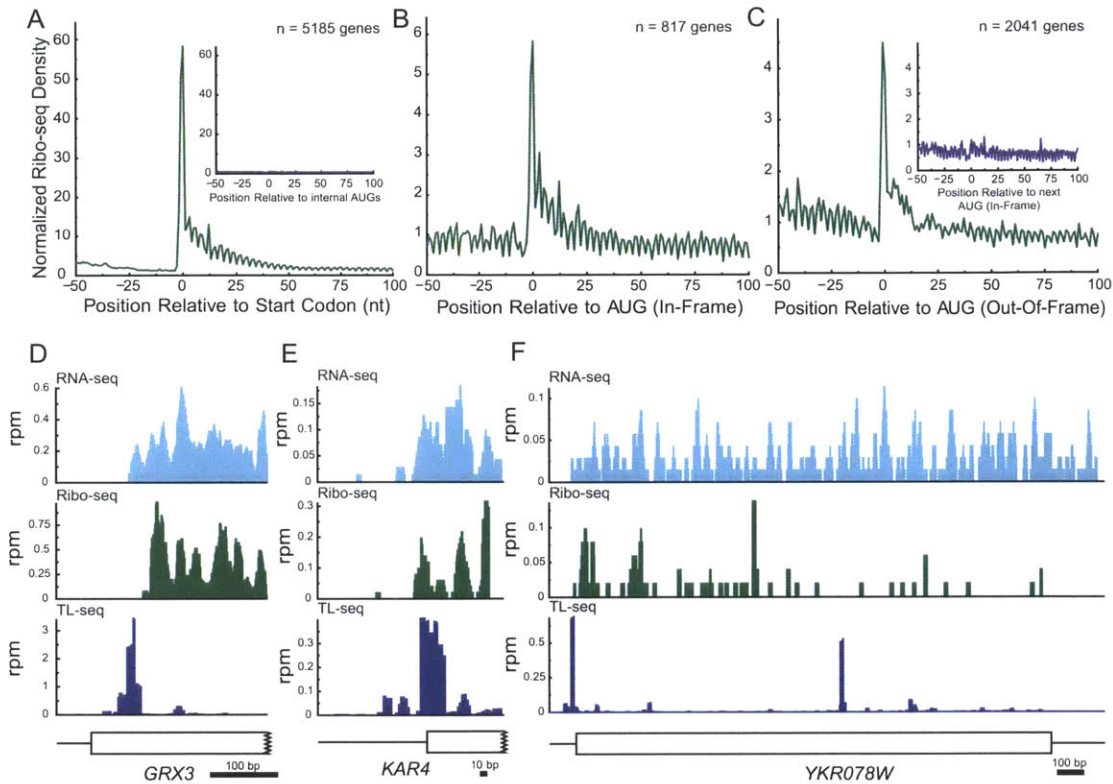


Figure 2.4: Three Types of Internal TSSs Identified by TL-seq

- (A) Ribosome footprint density aligned relative to annotated start codons for Ribo-seq from glucose-starved yeast. Ribosomes accumulate at initiation AUGs but not internal AUGs (inset).
- (B) Ribosome footprint density for internal TL genes whose first AUG is in-frame with the annotated start codon.
- (C) Ribosome footprint density for internal TL genes whose first AUG is out-of-frame with the annotated start codon. Inset is the first in-frame AUG for these same peaks.
- (D) Misannotated N-termini. RNA-seq, Ribo-seq, and TL-seq support a TSS starting internal to the annotated AUG.
- (E) N-terminal peak. TL-seq called a TSS just inside of the annotated ORF. RNA-seq and Ribo-seq support such an internal TSS.
- (F) TL-seq identified a second internal TSS.

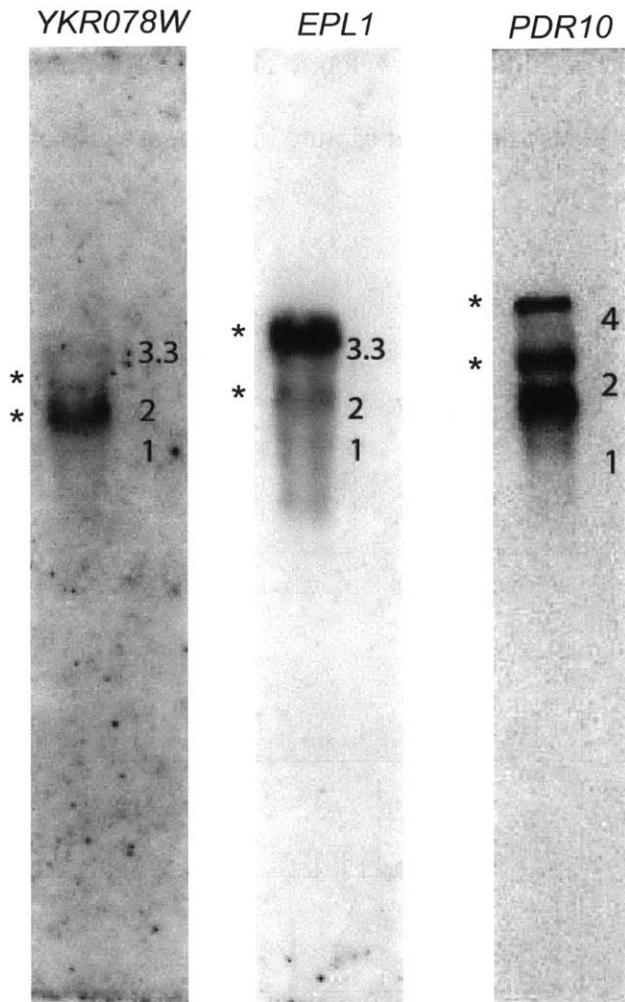


Figure 2.5: Northern Validation of Internal Peak-Containing Genes

The expected differences in transcript sizes (in kb) from TL-seq are: 1.0 (*YKR078W*); 1.4 (*EPL1*); and 1.9 (*PDR10*). The lower *PDR10* species was not predicted by TL-seq. Sizes of markers (in kb) are indicated to the right of each gel. Predicted bands indicated with asterisks.

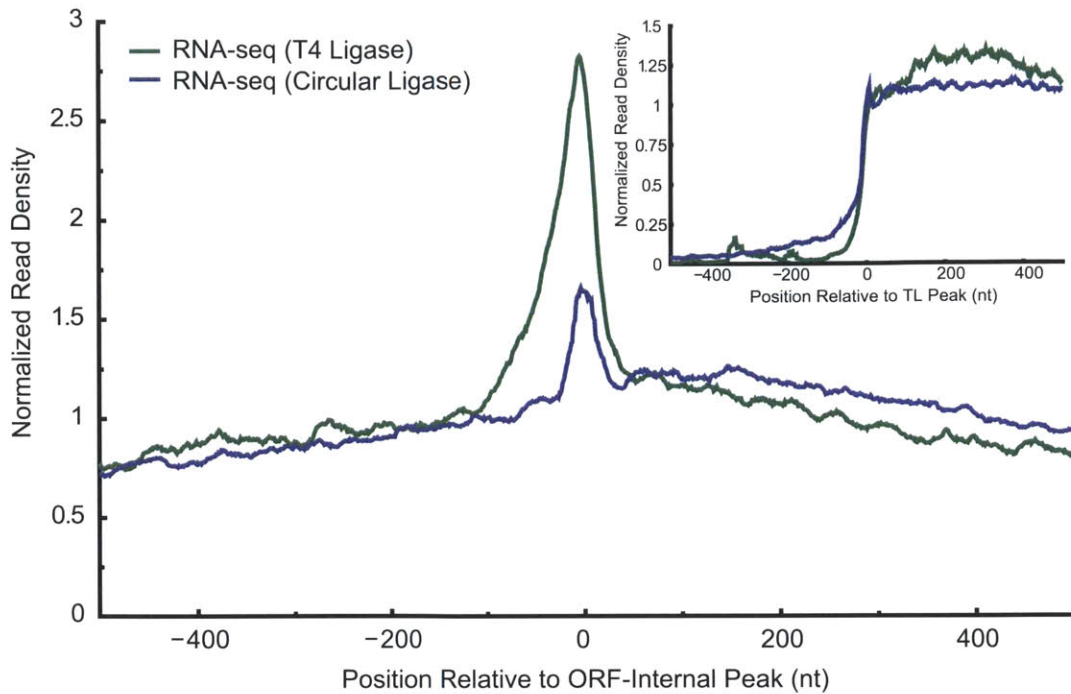


Figure 2.6: TL-seq Internal Peaks Are Also Peak-like in RNA-seq Libraries Made Using Similar 5'-Capture Techniques

The distribution of RNA-seq reads about internal TL-seq peak positions was analyzed for RNA-seq libraries prepared by different protocols. While TL peaks upstream of an ORF exhibit the distribution expected of TSSs (inset), the ORF-internal peaks do not. Furthermore, the peak-like distribution in the RNA-seq read distribution about ORF-internal TL-seq peaks is much greater when T4 RNA Ligase was used for library preparation. Reads were averaged in a 20 nt sliding window.

Genes with Short TLs Exhibit Inefficient Start Codon Recognition

TL-seq revealed an unexpected class of genes with very short TLs (≤ 12 nts), which are interesting from a translation initiation perspective. A ribosomal small subunit correctly positioned at an initiation codon protects at least 12 nts 5' to the AUG-occupied ribosomal P-site (Legon 1976). In addition, the capped 5' ends of most translating mRNAs are thought to be bound by eIFs that facilitate ribosome recruitment. Given the physical constraints of the factors involved, it seemed likely that short TL genes would initiate translation inefficiently at the cap-proximal AUG. Consistent with this prediction,

artificial 5' truncation of the *PGK1* TL to 21 nts or fewer has been shown to reduce the efficiency of translation from the first start codon in yeast (van den Heuvel et al. 1989). We reasoned that the frame of the second downstream AUG would determine the susceptibility of a short TL mRNA to nonsense-mediated decay (NMD) (Figure 2.7A). If a ribosome missed the cap-proximal AUG and initiation occurred at a downstream AUG in-frame with the annotated ORF, termination would occur at the annotated stop codon. However, if the first recognized AUG was out-of-frame, 98% of the time ribosomes would terminate at a premature termination codon (PTC), which leads to degradation by the NMD pathway in yeast (Leeds et al. 1991).

According to the above logic, short TL genes should be targets of the NMD pathway only when the second AUG within the ORF is out-of-frame. To test this prediction, we analyzed published microarray data for changes in transcript levels in yeast deleted for each of three factors required for NMD (Upf1, Upf2, or Upf3) (He et al. 2003). Genes with short TLs exhibited a significant shift towards increased steady-state mRNA levels in NMD-deficient *upf1Δ* yeast strains, consistent with our model ($p=0.0009$, Figure 2.7B). As predicted, a significant shift was observed only when the second AUG was out-of-frame (Figure 2.7B). We confirmed the microarray results by qPCR: 8/9 genes behaved as expected (Figure 2.8). Increased mRNA levels for short TL genes were observed in all examined NMD-deficient strains (Figure 2.8), and also for short TL genes identified by other TL annotations ((Xu et al. 2009), $p=10^{-7}$).

Some short TL genes' mRNA levels were more affected by inactivation of NMD than others. According to our model, NMD sensitivity should relate to the extent of in-frame vs. out-of-frame initiation. As predicted, genes whose mRNA levels increased in

*upf1*Δ exhibited decreased ribosome density at the first, annotated AUG concomitant with increased ribosomal density at the downstream out-of-frame AUG (Figure 2.7C). This result is consistent with a direct role for out-of-frame translation leading to NMD of short TL genes. We note that while TL-seq identified a TL ≤12 nt for short TL genes, the Ribo-seq density, as well as RNA-seq density indicate the existence of longer TL isoforms for these genes as well, likely a result of TL heterogeneity. Although we focused our analysis on TLs ≤12 nt, which is the minimum length between the ribosomal P site and mRNA exit site, it is likely that more than 12 nts are required for simultaneous interaction of eIF4F with the cap and 40S subunits with the mRNA (Kozak and Shatkin 1977; Lazarowitz and Robertson 1977; Legon 1976). Consistent with this view, analysis of TL lengths up to 20 nts yielded similar results. Thus, mRNAs with short TLs represent a new class of NMD substrate, revealing a new functional role for NMD.

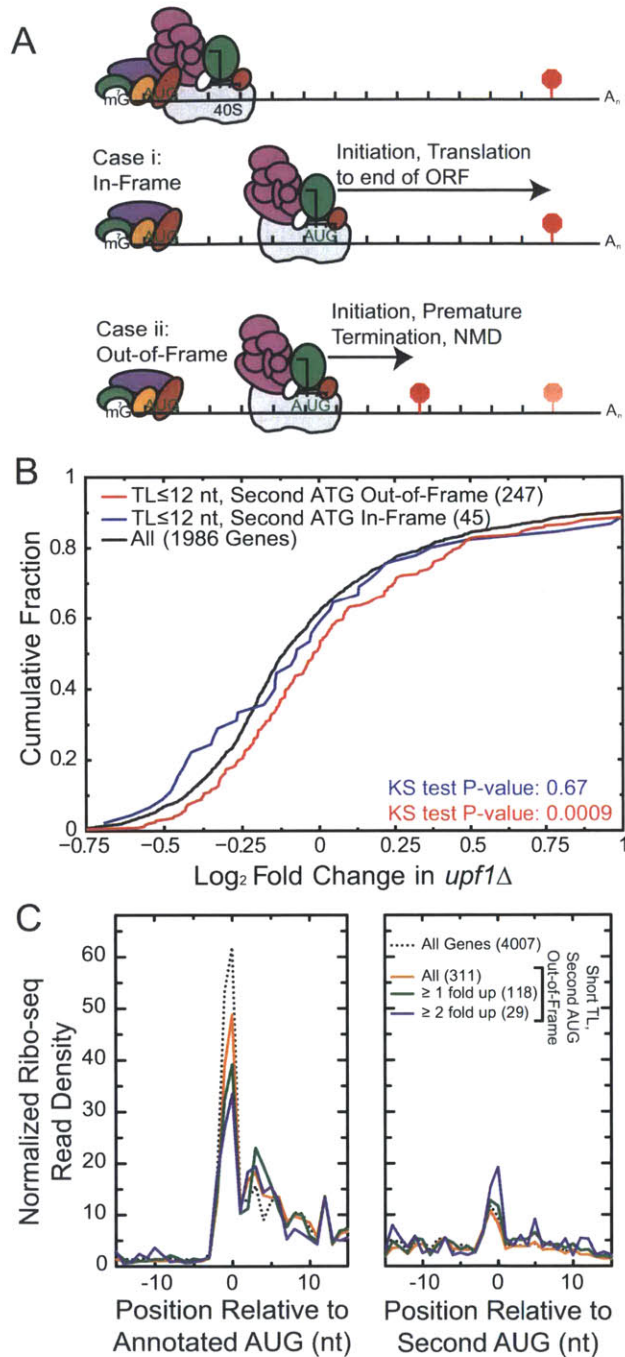


Figure 2.7: Short TL Genes Are Enriched for NMD Targets

(A) Model predicting why short TL genes with a second out-of-frame AUG are NMD targets (not to scale). Failure to identify the cap-proximal AUG in short TLs results in scanning and recognition of a second, downstream AUG. If the second AUG is out-of-frame, it results in premature termination and NMD. For simplicity the eIF4F complex is drawn on the cap during scanning, though some or all of its subunits may remain associated with the small ribosomal subunit (Aitken and Lorsch 2012; Jackson et al. 2010).

(B) Fold change in steady-state mRNA levels for short TL genes in *upf1*Δ cells. Genes

with short TLs exhibit a significant shift towards increased RNA levels only when the next AUG encountered is out-of-frame. Number of genes in each group is indicated in parentheses.

(C) Ribosome density analysis of genes with a second AUG out-of-frame, using Ribo-seq from glucose-starved yeast. The dotted line shows all genes; solid lines show short TL genes with second AUG out-of-frame. Fold up indicates genes whose mRNAs are increased in the *upf1Δ* microarray data.

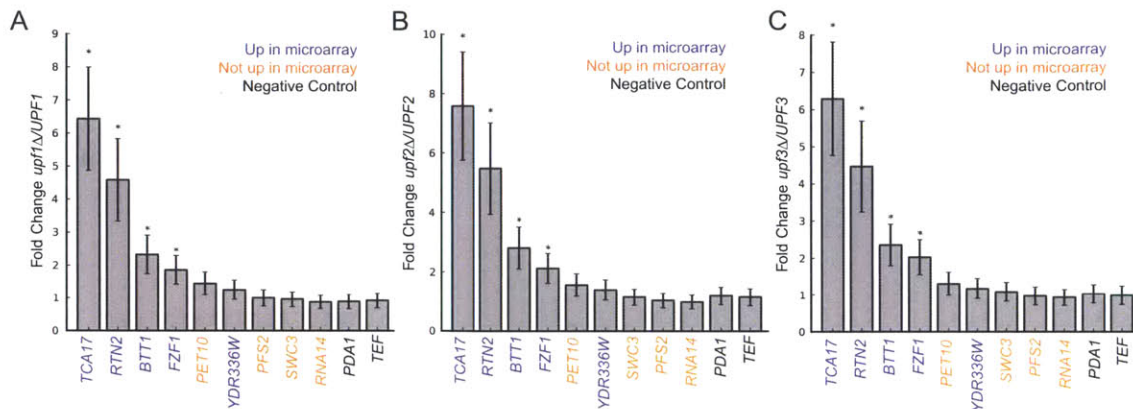


Figure 2.8: Validation of Steady State mRNA Fold Changes in NMD-Deficient Yeast

(A) qPCR validation of microarray results. 4/5 genes predicted to increase based on microarray results in *upf1Δ* do so (purple), whereas 4/4 genes predicted not to increase do not (orange). *PDA1* and *TEF1* are negative controls. Bars are mean and standard deviation. * $p < 0.05$.

(B) Same for *upf2Δ*.

(C) Same for *upf3Δ*.

uAUGs Are Conserved Inhibitory Elements for Translation

Having established TL-seq as a method for defining TSSs, we examined the relationships between TL features and translation activity genome-wide. To systematically investigate the translational activity of TLs, we developed Translation-Associated Transcript Leader Sequencing (TATL-seq) (Figure 2.9A). TL-seq was performed on each of seven fractions across a polysome gradient, which differentially sediments mRNAs according to the number of ribosomes bound. Because translation initiation is thought to be rate-limiting for translation of most genes, more efficiently translated mRNA isoforms associate with

heavier polysomes. The distribution of 3,916 TL peaks for 3,651 genes was quantified across a polysome gradient, thus determining TL-isoform specific sedimentation, and presumptively, information about the translational activity of individual TL isoforms. As expected, TL abundance was highly correlated between adjacent gradient fractions and less correlated between fractions with large differences in translation activity (e.g. non-translating mRNAs in fraction 1 and efficiently translated mRNAs in fraction 7) (Figure 2.9B, C). TATL-seq thus enables de novo TL annotation while simultaneously testing those TLs' translational activity in a single experiment.

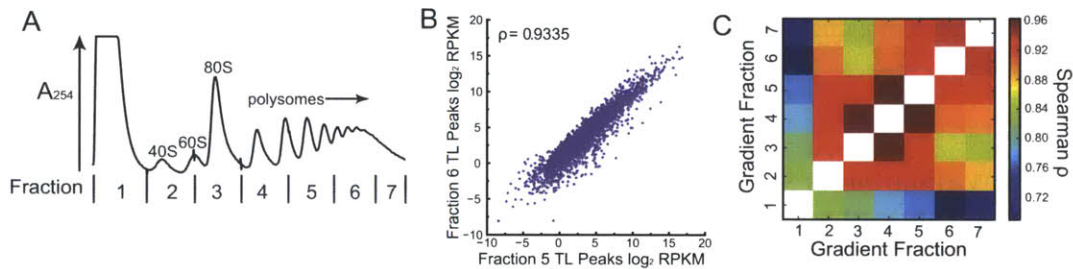


Figure 2.9: TATL-seq Quantifies Translation Activity of TLs In Vivo

- (A) TATL-seq was performed on each of 7 fractions across a polysome gradient.
- (B) Peaks were called on the computationally pooled TATL-seq fractions, then RPKMs were computed for each fraction individually. The Spearman correlation between two gradient fractions is shown.
- (C) Heatmap of Spearman's ρ for peak abundance in different TATL-seq fractions.

uAUGs are thought to negatively affect the efficiency of translation initiation at the downstream ORF, but the generality of this effect in yeast has been a subject of debate, partially because previous analyses were restricted to a handful of uAUGs (Lawless et al. 2009; Marija Cvijović 2007). To assess the generality of uAUG-mediated translational repression in rapidly dividing yeast, we examined the polysomal distribution of 773 TL species that contained one or more uAUGs compared to 3,143 TL species that lacked uAUGs. uAUG-containing mRNAs showed a substantial and significant increase

in sedimentation outside of polysomal fractions (Mann Whitney $p=8.7 \times 10^{-34}$ for fraction 1, $p=1.1 \times 10^{-14}$ for fraction 2, Figure 2.10A). In addition to providing a sink for scanning ribosomes, uAUGs that are followed by short coding regions (uORFs) lead to translation termination near the 5' end of the transcript, which has been shown to elicit mRNA degradation via the NMD pathway (Oliveira and McCarthy 1995). Consistent with uORF-stimulated NMD being a general mechanism, uAUG-containing mRNAs' steady-state levels were significantly increased in NMD-deficient yeast (mean fold change 1.72, $p < 10^{-16}$, Figure 2.10B).

An estimated ~15% of genes with a single TL identified by TL-seq contained ≥ 1 uAUG, which is consistent with previous estimates of uAUG prevalence in yeast (Lawless et al. 2009). This frequency was much lower than the expected value predicted from randomizing gene-specific TL lengths (~23%, $p=10^{-69}$ Figure 2.10C). The observed frequency of uAUGs was also significantly lower than expected by chance given the dinucleotide composition of TLs ($p < 10^{-310}$). Thus uAUGs are underrepresented in yeast TLs. Nevertheless, some yeast uAUGs are known to have biological functions as translational regulators of gene expression (e.g. *GCN4* (Mueller and Hinnebusch 1986) and *CPA1* (Wang et al. 1999)). Consistent with potential regulatory importance, uAUG is the most conserved of all possible uNNN codons in the TL region (Figure 2.10D). Taken together, these data show that uAUGs are uncommon in yeast TLs, but when uAUGs are present, they tend to be both conserved and functional.

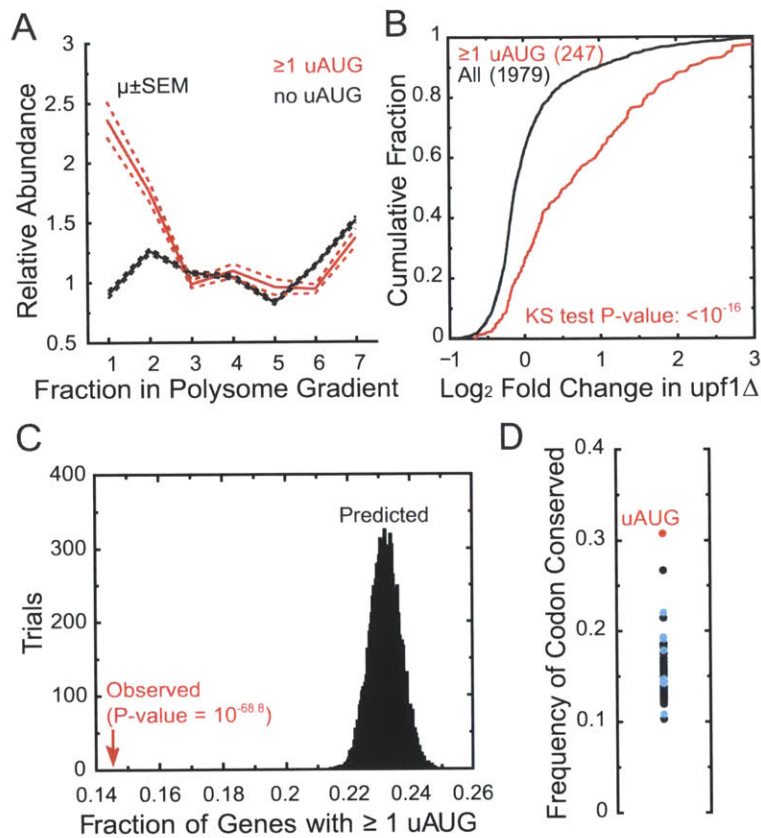


Figure 2.10: uAUGs Are an Underrepresented and Conserved Sequence Element Associated with Decreased Translation

(A) TATL-seq sedimentation pattern for uAUG-containing and all TLs. Relative abundance (ordinate) is the abundance of a given TL in a fraction divided by its abundance across the entire gradient.

(B) Fold change in mRNA steady state levels for single TL genes, either with uAUG-containing TLs or all TLs. Numbers in parentheses indicate number of genes.

(C) The fraction of uAUG-containing TLs was calculated based on observed TL lengths for single TL genes (red arrow) and 10,000 randomizations of gene-specific TL length (histogram). P-value based on z-score of the observed value compared to histogram.

(D) Conservation of each of 64 possible uNNN trinucleotides in the TL region was calculated using a genome-wide alignment of *S. paradoxus*, *S. mikatae*, *S. bayanus*, and *S. cerevisiae* and single TL genes. The ordinate is the number of conserved instances over the total occurrences of that trinucleotide. Near-uAUG trinucleotides are highlighted in blue.

Although translation initiation can occur at near-AUG codons under some circumstances (Chang 2004; Tang 2004; Ingolia et al. 2009), our data do not support a widespread functional role for non-AUG initiation. Upstream-near-AUGs (near-uAUGs)

are not highly conserved, unlike uAUGs (Figure 2.10D). Furthermore, near-uAUGs were not associated with higher sedimentation in a polysome gradient, nor was there a significant shift in translation efficiency (TE) by Ribo-seq (Figure 2.11A,B). Similar results were obtained using gene-specific TE values from cells subjected to amino acid starvation, a condition in which near-uAUG initiation was proposed to be a frequent event (Ingolia et al. 2009). Finally, genes with near-uAUG codons in their TLs were not shifted towards increased steady state mRNA levels in NMD-deficient yeast (Figure 2.11C). Restricting these analyses to near-uAUG codons in favorable initiation contexts did not affect the results Thus near-uAUG codons do not have an identifiable genome-scale functional role akin to that observed for uAUGs.

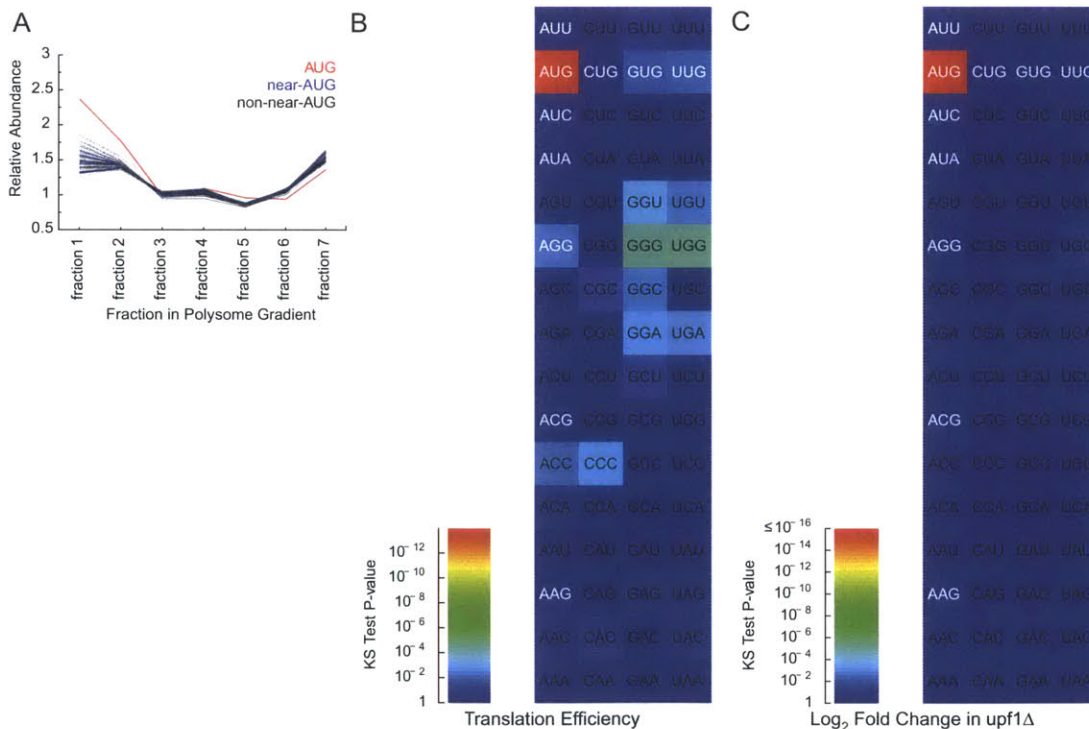


Figure 2.11: There Are No Detectable Effects of Near-uAUG Codons on Translation or NMD

(A) TATL-seq sedimentation pattern for all uNNN-containing TLs on the same plot. (B) TE from Ribo-seq was compared for uNNN-containing and all TLs. Significant

assessed with KS test. While three near-uAUG codons (GUG, AGG, and UUG) were associated with a slight decrease in TE in rich media, these changes were not significant after correction for multiple hypothesis testing ($p < 0.0007$). Near-uAUG codons highlighted in white.

(C) KS test p-value for differences in fold change of mRNA steady state levels in *upf1Δ*. Only uAUG is significant. Similar results were obtained for *upf2Δ* and *upf3Δ*.

Intragenic TL heterogeneity

The TL peak-calling algorithm reduces a distribution of reads to a single peak, thus simplifying the data for downstream analyses. However, this simplification ignored potentially interesting differences in the distribution of reads within TL peaks. In fact, the distribution of reads within TL peaks was variable, with some peaks being more heterogeneous than others. Unlike microarray-based techniques for identifying TSSs, TL-seq allowed us to directly observe and study such variability. To quantify gene-specific TL heterogeneity, we utilized Shape Index (SI), a metric that incorporates fractional read abundance at every position in a defined region to provide a measure of the heterogeneity therein (Hoskins et al. 2011). The maximum SI score is 0, which indicates that all reads in the region were generated from a single position. Deviations from this distribution result in a decrease in SI score. Using this metric to quantify intragenic TL heterogeneity, the genome-wide distribution of SI scores (Figure 2.12A, top) was centered about a modestly heterogeneous average (e.g. *SUI3* Figure 2.12B), with a substantial fraction of genes being much more or less heterogeneous (e.g. *BSC5* and *MSS116*, respectively, Figure 2.12B).

To investigate whether differences in TL peak heterogeneity might be functionally significant, we examined the distribution of SI scores among Gene Ontology (GO) categories. Notably, TLs from genes encoding proteins predicted to have regulatory

functions (GO categories “specific transcriptional repressor activity”, “sequence-specific DNA binding”, “response to stress”) were significantly shifted towards greater heterogeneity (Figure 2.12B). For these genes, TL heterogeneity might represent an underappreciated mechanism for regulation. In mammals, genes with alternative TL variants are also enriched for regulatory functions (Resch et al. 2009), consistent with what was observed here in yeast. Conversely, genes encoding ribosomal proteins and factors required for ribosome biogenesis and translation (GO categories “ribosome”, “translation”, “pseudouridine synthase activity”) were significantly shifted towards more homogeneous TLs. For such mRNAs, a homogeneous TL population might ensure high levels of expression and/or concerted post-transcriptional regulation of the downstream ORFs.

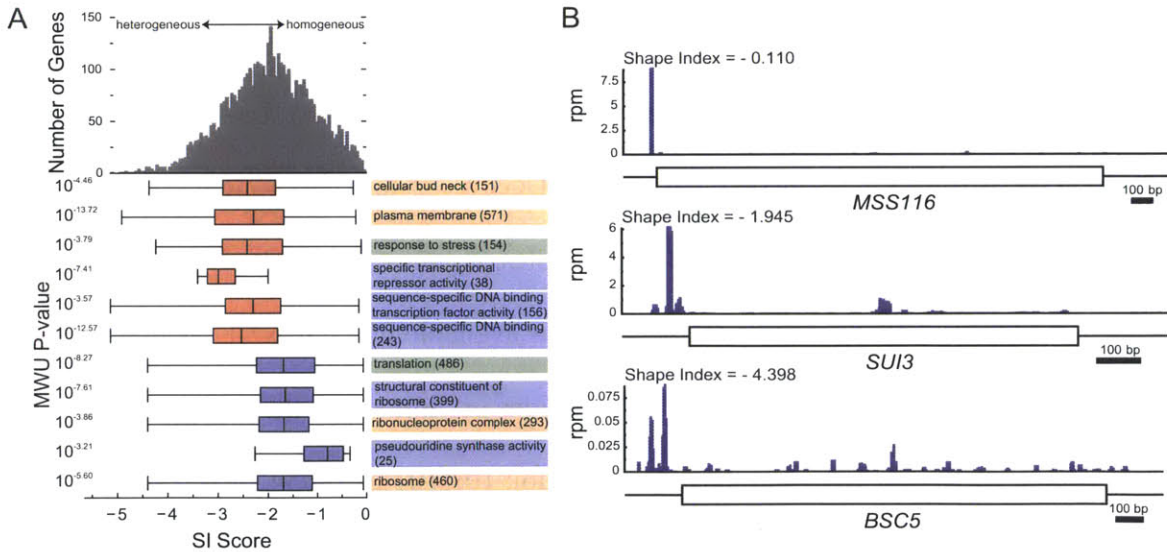


Figure 2.12: There is Intragenic TL Heterogeneity

(A) SI scores of GO categories with ≥ 10 genes were compared to all genes (top histogram). GO categories with a significantly different SI score distribution (Bonferroni-corrected Mann Whitney U $p < 0.001$) are shown, with number of genes in parentheses. Box and whisker plots indicate quartiles and range, red/blue shading indicates decreased/increased SI. Shading of GO categories indicates Process (green), Component (orange), or Function (blue).

(B) Three examples of genes with different Shape Index (SI) scores.

Intragenic TL Heterogeneity Can Have Consequences for Translation

230 yeast genes exhibited more than one well-separated peak of TL reads upstream of their annotated ORF. Such alternative TLs have been shown to confer differences in translation efficiency and regulation in a handful of cases though the generality of this phenomenon is unknown (Rosenstiel et al. 2007; Smith et al. 2009). Many of these 230 genes showed distinct polysome sedimentation patterns. To confirm that these TLs exist as full-length alternative TL mRNAs with the same ORF, we examined RNA-seq density and observed an increase in read density at each position that did not decrease at further downstream positions (Figure 2.13). Of genes with two TL species, the longer TL appeared more poorly translated on average, being 1.6-fold more abundant than the shorter isoform in the non-translating pool ($p=1.1 \times 10^{-4}$ Mann Whitney, Figure 2.14A). Many of these apparent differences in translation are likely due to uAUGs—nearly 35% of the long TL isoforms contained at least one uAUGs, a >2-fold enrichment over the genome-wide frequency of uAUGs in TLs. Examples of genes with multiple TL species is shown in Figures 2.14B,C, and 2.15. A full list is available in Table S2.1 (available on CD at MIT's Institute Archives & Special Collections or online with this article as Table S4 at Genome Research. 2013 Jun;23(6):977-87. doi: 10.1101/gr.150342.112.), and a summary of these and other findings are presented in Table 2.5.

To determine whether the alternative TL sequences were sufficient to alter translation activity, translation efficiencies (protein produced per mRNA) were determined for alternative TL variants for six genes using Firefly luciferase (*Fluc*) reporters. Each TL was inserted upstream of *Fluc* and downstream of the *GALI*

promoter, yielding constructs for inducible expression of mRNAs that differed solely in the TL region. *Fluc* activity per mRNA varied by almost 25-fold (Figure 2.14D), demonstrating that TLs were sufficient to confer large differences in translation efficiency *in vivo*. These results are consistent with *in vitro* data showing large TL-dependent differences in cap-dependent translation efficiency (Rojas-Duran and Gilbert 2012). The largest fold difference between variants of a single gene was seen for *CRZI*; the shorter isoform was translated more than 19-fold more efficiently than the longer isoform. In five of six cases including *CRZI*, intragenic TL variants showed significantly different translation efficiencies (Figure 2.14D). In four such cases, the TL isoform that was predicted to have higher translation activity by TATL-seq was in fact better translated in the TL construct. For these genes, we conclude that differences in the TL region alone are sufficient to confer the observed differences in translation efficiency. The remaining cases may represent instances where other elements (e.g. 3'UTRs, promoter-dependent mRNP assembly differences) or combinations of elements contribute to the differential translation activities observed by TATL-seq.

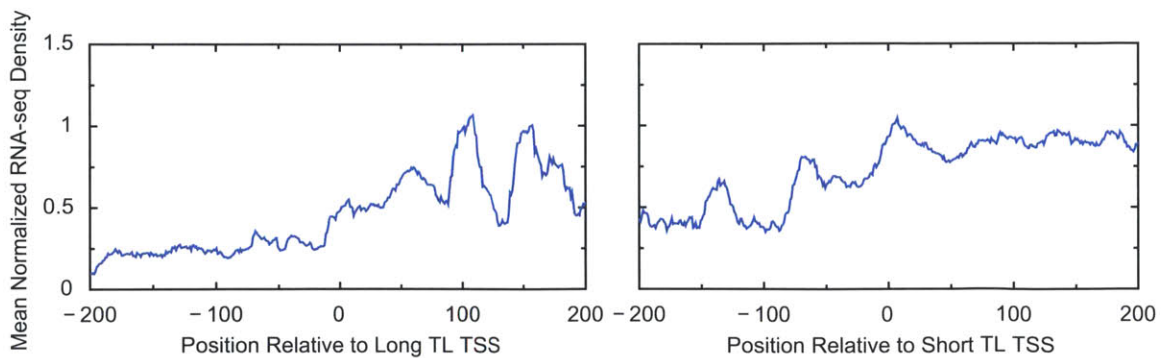


Figure 2.13: Distribution of RNA-seq Reads About Short/Long TL Pairs

RNA-seq read counts were normalized per gene, and averaged over a 21nt window about each position. If long TLs are part of short, truncated transcripts, there should be a drop off in RNA-seq read density downstream of the long TL TSS and upstream of the short TL TSS.

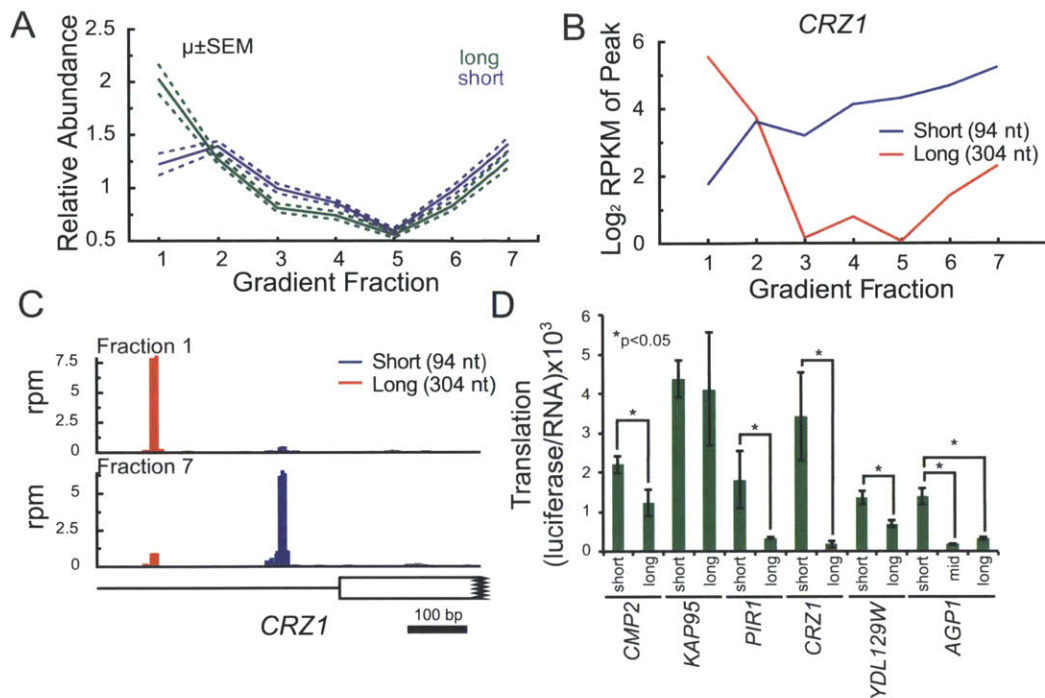


Figure 2.14: Intrinsic TL Heterogeneity Leads to Different Translation Behavior In Vivo

(A) Average sedimentation pattern for 204 short/long TL pairs.

(B) *CRZ1* is an example of a gene with multiple TLs showing distinct sedimentation patterns in a polysome gradient.

(C) TATL-seq profile of *CRZ1* from fractions 1 and 7.

(D) TLs are sufficient to confer the translational behavior predicted from TATL-seq (4/6 genes). *In vivo* translation (ordinate) was determined as luciferase activity per unit RNA. Mean and standard deviation of biological triplicates is shown, $*p < 0.05$, Student's t-test.

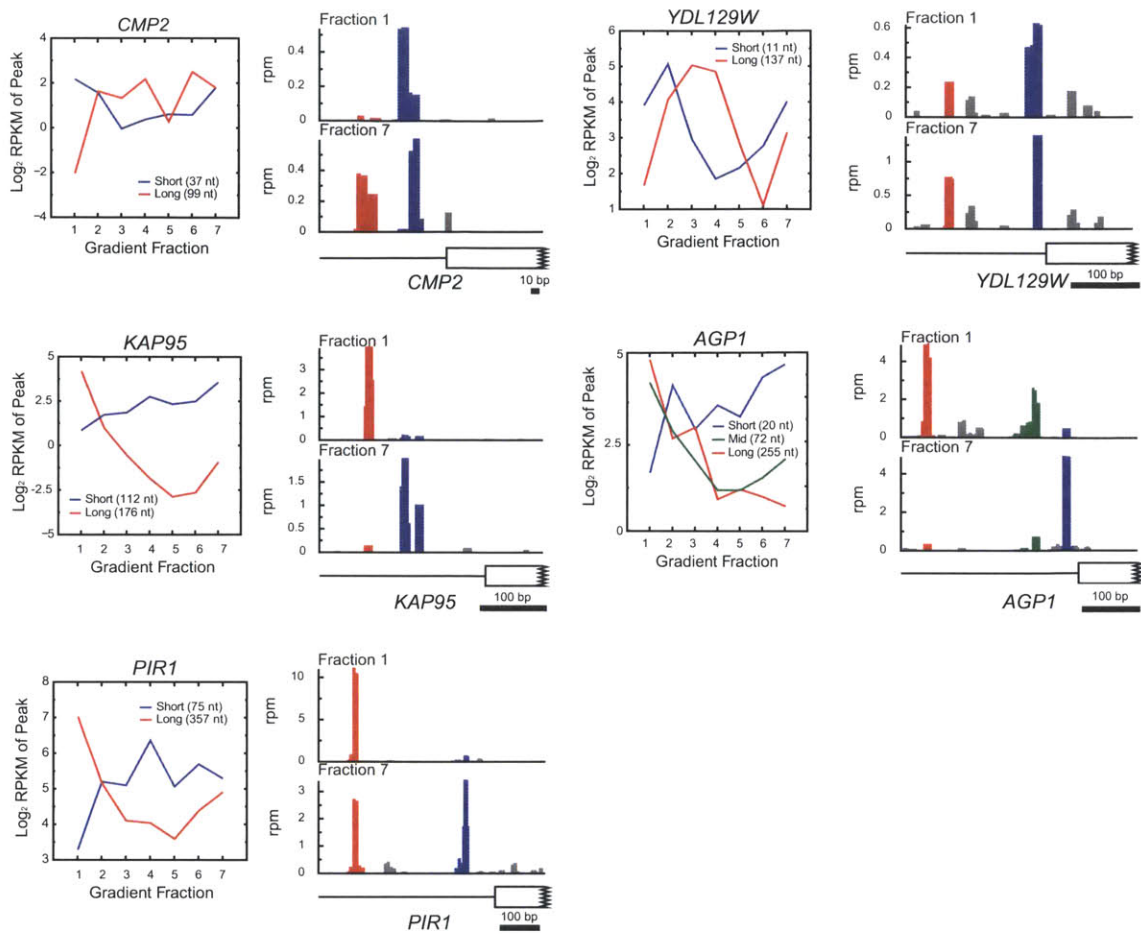


Figure 2.15: Distribution of TL Peaks Across a Polysome Gradient for Genes Shown in Figure 2.14D

To determine the predicted relative translation between constructs, we compared their relative sedimentation with polysomes, that is, the ratio of RPKMs in the translating pool (fractions 3-7) to the non-translating pool (fraction 1).

Category	Number of Events	Notes
Peaks called by TL-seq	7,254	
Peaks upstream of a CDS	4,002	Nucleosome distribution shown in Fig 2.2A
Genes with at least one internal TL-seq peak	2,171	True fraction likely 6% of genes, >85% of which are <100 nt into CDS (Fig 2.3C,D)
Genes with one TL peak called, none elsewhere	2,668	Used for analyses in Fig 2.7, 2.10B-D, S2.2B, 2.11
Genes with one TL peak called, one or more in ORF/3'UTR	3,729	
Genes with TL-seq peak ≤ 12 nt upstream of CDS	412	Enriched for NMD targets when second AUG is out-of-frame (Fig 2.7)
Genes with at least two TL-seq peaks >50 nt apart upstream of their CDS	230	These show distinct sedimentation on average by TATL-seq (Fig 2.14A)

Table 2.5: Summary Information From Peaks for Pooled TATL-seq Libraries

Discussion

Mapping of TSSs and TLs is a critical component of functional genome annotation. Here we present techniques for genome-wide, high resolution TSS identification and functional characterization of TLs. Applying the TL-seq and TATL-seq approaches to yeast revealed TLs that alter protein-coding potential, translation efficiency, and susceptibility to NMD, thus demonstrating the methods' potential to illuminate TL-mediated post-transcriptional regulation of gene expression. TL-seq has a fast and straightforward library preparation procedure that can be applied to any capped transcriptome, i.e. any RNA pool for which 5' RACE is useful. Because TL-seq and TATL-seq are sequencing-based approaches, they are applicable in any organism with a sequenced genome and do not require prior knowledge of transcribed regions.

The TL-seq approach enabled visualization of TSSs within other annotated transcripts. Such internal TSSs exhibited characteristic nucleosome signatures similar to canonical TSSs, suggesting they are true TSSs and not merely artifacts of the technique.

Furthermore, protein products from internal initiation have recently been identified by mass spectrometry, thus corroborating our findings here that such internal transcripts are translated (Fournier et al. 2012). Analysis of transcript 5' ends also yielded a significant number of ORF-internal reads in *Drosophila* (16% (Ni et al. 2010)) and human cells (2-3% (Kanamori-Katayama et al. 2011)). These findings raise the possibility that diverse eukaryotes express internal transcripts, the functions of which are largely uncharacterized.

The majority of internal TSSs fell into the category of N-terminal peaks. The phenomenon of internal TSSs encoding functionally distinct N-terminal protein isoforms was first described for the *SUC2* invertase gene (Carlson and Botstein 1982). The N-terminal sequence that is missing from the shorter mRNA isoform encodes a secretory peptide, and thus the upstream TL variant encodes a secreted Suc2p while the downstream TL encodes cytoplasmic Suc2p. Similar cases have been described for *FUM1* (Wu and Tzagoloff 1987), *LEU4* (Beltzer et al. 1986), *HTS1* (Chiu et al. 1992), *TRM1* (Ellis et al. 1987), *VASI* (Chatton et al. 1988), and *KAR4* (Gammie et al. 1999). Our data suggest that N-terminal TSSs might be a more common way of creating and regulating protein diversity than previously appreciated.

Approximately one percent of TSSs were ≥ 100 nt internal to the ORF. Similar internal initiation events, termed “cryptic initiation”, have previously been observed in yeast mutants with defects in chromatin structure and/or transcription elongation as well as in wild type cells subject to starvation (Kaplan 2003; Cheung et al. 2008). For certain genes including *FLO8* and *STE11*, increased H3K4 trimethylation, increased H4 acetylation, and decreased H3K36 methylation correlate with an increase in cryptic

initiation events (Carrozza et al. 2005). Comparing these chromatin modifications averaged across internal peak-containing genes versus all genes, we did not observe a significant difference in H4 acetylation nor H3K36 methylation, and in fact observed a slight decrease in H3K4 trimethylation ($p=0.011$, Mann Whitney U test) (using data from (Pokholok et al. 2005)). It is thus unclear whether the internal TSS events observed here are mechanistically related to previously described instances of cryptic initiation.

Genome-wide microarray studies of cryptic transcription in yeast proposed that at least 17% of genes have the potential for cryptic initiation under some circumstances (Cheung et al. 2008).

TL-seq also identified an unexpected class of genes with extremely short TLs. The physical constraints of translation initiation on short TLs suggest these features pose problems for the initiation machinery, and our results verify this to be the case. Previously it was noted that short TLs of viral and/or artificial mRNAs were associated with decreased initiation at the cap-proximal AUG and increased initiation at downstream AUGs both *in vivo* and *in vitro* (Sedman et al. 1990; Pestova and Kolupaeva 2002; Kozak 1991). We demonstrate for the first time that short TLs on natural cellular mRNAs lead to inefficient initiation at the cap-proximal AUG, increased initiation at downstream AUGs *in vivo*, and targeting of the transcript for decay by NMD. Together, these observations indicate a novel form of TL-mediated post-transcriptional regulation and reveal a new functional role for NMD in yeast.

Why might short TL mRNAs exist, if only to be degraded? Short TL genes present unique post-transcriptional regulatory opportunities for a cell. Their degradation via NMD could be regulated, as conditions such as hypoxia and amino acid starvation

have been reported to stabilize NMD substrates in humans (Gardner 2008; Mendell et al. 2004). For reporter mRNAs with short TLs, the efficiency of cap-proximal AUG recognition *in vitro* can be controlled by the levels of eIF1 (Pestova and Kolupaeva 2002), raising the possibility that the levels or activity of eIF1 *in vivo* may control the production of full-length protein (and the extent of out-of-frame initiation) for short TL genes. Alternatively these genes may produce alternate, longer TLs in altered cellular conditions, as has been observed for some yeast genes following nitrogen starvation, pheromone response, and osmotic stress (Law et al. 2005), thus leading to less out-of-frame initiation and less NMD of those mRNAs. The balance between initiation at cap-proximal and downstream AUGs, and the recognition of the latter by NMD represents an opportunity for concerted regulation of this class of transcripts.

In all systems in which it has been examined, gene-specific translation efficiency has been shown to vary substantially (Ingolia et al. 2009; Guo et al. 2010; Stadler and Fire 2011; Arava et al. 2003; Hendrickson et al. 2009; Thoreen et al. 2012). While we do not yet understand all the factors contributing to this wide (>100-fold) variation, here we demonstrate that TLs can have a direct role in explaining some of the observed variation in translation genome-wide.

TATL-seq has great potential to illuminate the mechanisms responsible for translation activity differences and regulation. Since it is a direct measurement of TL isoform-specific translation, it is particularly useful for discerning the relative contributions of multiple isoforms to the overall translation of a gene. Isoform differences are invisible to most conventional approaches (e.g. RNA-seq, polysome microarray,

ribosome footprint profiling) and may confound efforts to relate TL properties to ORF-based measurements of translation activity.

Our systematic investigation of intragenic TL variation showed that such variation has significant consequences for translation. These findings support and extend the conclusions from low-throughput studies of yeast and human genes with alternative TLs. We detected TL variants with differing translation even under standard growth conditions in *S. cerevisiae*, which has relatively low levels of mRNA isoform diversity. In contrast, TL diversity is quite common in mammals; in fact, it was suggested that TSS selection contributes more to mRNA isoform diversity than alternative splicing in some tissues (Pal et al. 2011). Furthermore, in a diverse panel of human tissues, the total number of alternative TLs observed was similar to the numbers of alternative 3'UTRs and alternatively spliced internal exons (Wang et al. 2008). Intriguingly, the majority of TL variants showed tissue-specific expression patterns. Importantly, because most intragenic TL variants do not change the coding potential of the mRNA, their influences must be felt during the post-transcriptional life of the mRNA, namely, during translation, localization, and/or decay. We anticipate that the TL-seq and TATL-seq methods described here will enable systematic studies of TL regulation and function in eukaryotes.

Data Access

The TL-seq and TATL-seq data, as well as processed TL-lengths, and links to UCSC and SGD genome browser-formatted data from these experiments have been submitted to the NCBI Gene Expression Omnibus (GEO) ([http:// www.ncbi.nlm.nih.gov/geo/](http://www.ncbi.nlm.nih.gov/geo/)) under

series accession no. GSE39074. The data is also viewable under “Select Tracks” in GBrowse at SGD (<http://yeastgenome.org>).

Acknowledgements

We thank Stuart Levine and the BioMicro Center for performing the sequencing and for discussion of ChIP-seq peak-calling algorithms; Uttam RajBhandary, Jason Merkin, and Charles Lin for insightful discussions; Chris Burge, Stirling Churchman, and members of the Gilbert Lab for critical reading of the manuscript. This work was supported by a National Science Foundation Graduate Research Fellowship award to J.A.A., and National Institute of General Medical Sciences Grant GM081399 to W.V.G.

References

- Aitken CE, Lorsch JR. 2012. A mechanistic overview of translation initiation in eukaryotes. *Nat Struct Mol Biol* **19**: 568–576.
- Altschul SF, Erickson BW. 1985. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol Biol Evol* **2**: 526–538.
- Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D. 2003. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **100**: 3889–3894.
- Arribere JA, Doudna JA, Gilbert WV. 2011. Reconsidering Movement of Eukaryotic mRNAs between Polysomes and P Bodies. *Molecular Cell* **44**: 745–758.

- Beltzer JP, Chang LF, Hinkkanen AE, Kohlhaw GB. 1986. Structure of yeast LEU4. The 5' flanking region contains features that predict two modes of control and two productive translation starts. *J Biol Chem* **261**: 5160–5167.
- Calvo SE, Pagliarini DJ, Mootha VK. 2009. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci USA* **106**: 7507–7512.
- Carlile TM, Amon A. 2008. Meiosis I Is Established through Division-Specific Translational Control of a Cyclin. *Cell* **133**: 280–291.
- Carlson M, Botstein D. 1982. Two differentially regulated mRNAs with different 5' ends encode secreted with intracellular forms of yeast invertase. *Cell* **28**: 145–154.
- Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, Lee KK, Shia W-J, Anderson S, Yates J, Washburn MP, et al. 2005. Histone H3 Methylation by Set2 Directs Deacetylation of Coding Regions by Rpd3S to Suppress Spurious Intragenic Transcription. *Cell* **123**: 581–592.
- Chang KJ. 2004. Translation Initiation from a Naturally Occurring Non-AUG Codon in *Saccharomyces cerevisiae*. *Journal of Biological Chemistry* **279**: 13778–13785.
- Chatton B, Walter P, Ebel JP, Lacroute F, Fasiolo F. 1988. The yeast VAS1 gene encodes both mitochondrial and cytoplasmic valyl-tRNA synthetases. *J Biol Chem* **263**: 52–57.
- Cheung V, Chua G, Batada NN, Landry CR, Michnick SW, Hughes TR, Winston F. 2008.

- Chromatin- and transcription-related factors repress transcription from within coding regions throughout the *Saccharomyces cerevisiae* genome. *Plos Biol* **6**: e277.
- Chiu MI, Mason TL, Fink GR. 1992. HTS1 encodes both the cytoplasmic and mitochondrial histidyl-tRNA synthetase of *Saccharomyces cerevisiae*: mutations alter the specificity of compartmentation. *Genetics* **132**: 987–1001.
- Clarkson BK, Gilbert WV, Doudna JA. 2010. Functional overlap between eIF4G isoforms in *Saccharomyces cerevisiae*. *PLoS ONE* **5**: e9114.
- David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. 2006. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci USA* **103**: 5320–5325.
- Ellis SR, Hopper AK, Martin NC. 1987. Amino-terminal extension generated from an upstream AUG codon is not required for mitochondrial import of yeast N2,N2-dimethylguanosine-specific tRNA methyltransferase. *Proc Natl Acad Sci USA* **84**: 5172–5176.
- Fournier CT, Cherny JJ, Truncali K, Robbins-Pianka A, Lin MS, Krizanc D, Weir MP. 2012. Amino Termini of Many Yeast Proteins Map to Downstream Start Codons. *J Proteome Res* 121121094456004.
- Gammie AE, Stewart BG, Scott CF, Rose MD. 1999. The two forms of karyogamy transcription factor Kar4p are regulated by differential initiation of transcription, translation, and protein turnover. *Mol Cell Biol* **19**: 817–825.

- Gardner LB. 2008. Hypoxic inhibition of nonsense-mediated RNA decay regulates gene expression and the integrated stress response. *Mol Cell Biol* **28**: 3729–3741.
- Geisler S, Lojek L, Khalil AM, Baker KE, Collier J. 2012. Decapping of Long Noncoding RNAs Regulates Inducible Genes. *Molecular Cell* **45**: 279–291.
- Gilbert WV, Zhou K, Butler TK, Doudna JA. 2007. Cap-Independent Translation Is Required for Starvation-Induced Differentiation in Yeast. *Science* **317**: 1224–1227.
- Guo H, Ingolia NT, Weissman JS, Bartel DP. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**: 835–840.
- Hahn S, Hoar ET, Guarente L. 1985. Each of three “TATA elements” specifies a subset of the transcription initiation sites at the *CYC-1* promoter of *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **82**: 8562–8566.
- He F, Li X, Spatrick P, Casillo R, Dong S, 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**: 1439–1452.
- Hendrickson DG, Hogan DJ, McCullough HL, Myers JW, Herschlag D, Ferrell JE, Brown PO. 2009. Concordant Regulation of Translation and mRNA Abundance for Hundreds of Targets of a Human microRNA ed. P.D. Zamore. *Plos Biol* **7**: e1000238.
- Hinnebusch AG. 2005. Translational regulation of *GCN4* and the general amino acid control of yeast. *Annu Rev Microbiol* **59**: 407–450.
- Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C,

- Booth BW, Zhang D, Wan KH, et al. 2011. Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Research* **21**: 182–192.
- Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**: 218–223.
- Jackson RJ, Hellen CUT, Pestova TV. 2010. The mechanism of eukaryotic translation initiation and principles of its regulation. 1–15.
- Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**: 1497–1502.
- Kanamori-Katayama M, Itoh M, Kawaji H, Lassmann T, Katayama S, Kojima M, Bertin N, Kaiho A, Ninomiya N, Daub CO, et al. 2011. Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Research* **21**: 1150–1159.
- Kaplan CD. 2003. Transcription Elongation Factors Repress Transcription Initiation from Cryptic Sites. *Science* **301**: 1096–1099.
- Koerber RT, Rhee HS, Jiang C, Pugh BF. 2009. Interaction of Transcriptional Regulators with Specific Nucleosomes across the *Saccharomyces* Genome. *Molecular Cell* **35**: 889–902.
- Kozak M. 1991. A short leader sequence impairs the fidelity of initiation by eukaryotic ribosomes. *Gene Expr* **1**: 111–115.
- Kozak M, Shatkin AJ. 1977. Sequences of two 5'-terminal ribosome-protected fragments

- from reovirus messenger RNAs. *J Mol Biol* **112**: 75–96.
- Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
- Law GL, Bickel KS, Mackay VL, Morris DR. 2005. The undertranslated transcriptome reveals widespread translational silencing by alternative 5' transcript leaders. *Genome Biol* **6**: R111.
- Lawless C, Pearson RD, Selley JN, Smirnova JB, Grant CM, Ashe MP, Pavitt GD, Hubbard SJ. 2009. Upstream sequence elements direct post-transcriptional regulation of gene expression under stress conditions in yeast. *BMC Genomics* **10**: 7.
- Lazarowitz SG, Robertson HD. 1977. Initiator regions from the small size class of reovirus messenger RNA protected by rabbit reticulocyte ribosomes. *J Biol Chem* **252**: 7842–7849.
- Leeds P, Peltz SW, Jacobson A, Culbertson MR. 1991. The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing a premature translational termination codon. *Genes Dev* **5**: 2303–2314.
- Legon S. 1976. Characterization of the ribosome-protected regions of 125I-labelled rabbit globin messenger RNA. *J Mol Biol* **106**: 37–53.
- Marija Cvijović DDEBGJKPS. 2007. Identification of putative regulatory upstream ORFs in the yeast genome using heuristics and evolutionary conservation. *BMC Bioinformatics* **8**: 295.

- Mavrich TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, Qi J, Schuster SC, Albert I, Pugh BF. 2008. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Research* **18**: 1073–1083.
- Mayr C, Bartel DP. 2009. Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* **138**: 673–684.
- Mendell JT, Sharifi NA, Meyers JL, Martinez-Murillo F, Dietz HC. 2004. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nat Genet* **36**: 1073–1078.
- Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, Ito T. 2006. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci USA* **103**: 17846–17851.
- Mueller PP, Hinnebusch AG. 1986. Multiple upstream AUG codons mediate translational control of GCN4. *Cell* **45**: 201–207.
- Mumberg D, Müller R, Funk M. 1994. Regulatable promoters of *Saccharomyces cerevisiae*: comparison of transcriptional activity and their use for heterologous expression. *Nucleic Acids Res* **22**: 5767–5768.
- Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**: 1344–1349.

- Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J. 2010. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature Publishing Group* **7**: 521–527.
- Oliveira CC, McCarthy JE. 1995. The relationship between eukaryotic translation and mRNA stability. A short upstream open reading frame strongly inhibits translational initiation and greatly accelerates mRNA degradation in the yeast *Saccharomyces cerevisiae*. *J Biol Chem* **270**: 8936–8943.
- Pal S, Gupta R, Kim H, Wickramasinghe P, Baubet V, Showe LC, Dahmane N, Davuluri RV. 2011. Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Research* **21**: 1260–1272.
- Pestova TV, Kolupaeva VG. 2002. The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection. *Genes Dev* **16**: 2906–2922.
- Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, Lee TI, Bell GW, Walker K, Rolfe PA, Herbolsheimer E, et al. 2005. Genome-wide Map of Nucleosome Acetylation and Methylation in Yeast. *Cell* **122**: 517–527.
- Resch AM, Ogurtsov AY, Rogozin IB, Shabalina SA, Koonin EV. 2009. Evolution of alternative and constitutive regions of mammalian 5'UTRs. *BMC Genomics* **10**: 162.
- Rhee HS, Pugh BF. 2011. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* **147**: 1408–1419.

- Rojas-Duran MF, Gilbert WV. 2012. Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA* **18**: 2299–2305.
- Rosenstiel P, Huse K, Franke A, Hampe J, Reichwald K, Platzer C, Roberts RG, Mathew CG, Platzer M, Schreiber S. 2007. Functional characterization of two novel 5' untranslated exons reveals a complex regulation of NOD2 protein expression. *BMC Genomics* **8**: 472.
- Sambrook J, Russell DW, Irwin N, Janssen K. 2001. *Molecular Cloning*. 3rd ed. eds. J. Argentine, N. Irwin, K. Janssen, S. Curtis, M. Zierler, M. Dickerson, I. Sialiano, N. McNerny, D. Brown, S. Schaefer, et al. Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
- Sedman SA, Gelembiuk GW, Mertz JE. 1990. Translation initiation at a downstream AUG occurs with increased efficiency when the upstream AUG is located very close to the 5' cap. *J Virol* **64**: 453–457.
- Smith L, Brannan RA, Hanby AM, Shaaban AM, Verghese ET, Peter MB, Pollock S, Satheesha S, Szykiewicz M, Speirs V, et al. 2009. Differential regulation of oestrogen receptor β isoforms by 5' untranslated regions in cancer. *Journal of Cellular and Molecular Medicine* **14**: 2172–2184.
- Stadler M, Fire A. 2011. Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA* **17**: 2063–2073.
- Tang HL. 2004. Translation of a Yeast Mitochondrial tRNA Synthetase Initiated at Redundant non-AUG Codons. *Journal of Biological Chemistry* **279**: 49656–49663.

- Thoreen CC, Chantranupong L, Keys HR, Wang T, Gray NS, Sabatini DM. 2012. A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature* **485**: 109–113.
- van den Heuvel JJ, Bergkamp RJ, Planta RJ, Raué HA. 1989. Effect of deletions in the 5'-noncoding region on the translational efficiency of phosphoglycerate kinase mRNA in yeast. *Gene* **79**: 83–95.
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476.
- Wang Z, Gaba A, Sachs MS. 1999. A highly conserved mechanism of regulated ribosome stalling mediated by fungal arginine attenuator peptides that appears independent of the charging status of arginyl-tRNAs. *J Biol Chem* **274**: 37565–37574.
- Wu M, Tzagoloff A. 1987. Mitochondrial and cytoplasmic fumarases in *Saccharomyces cerevisiae* are encoded by a single nuclear gene FUM1. *J Biol Chem* **262**: 12275–12282.
- Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**: 1033–1037.
- Yuan G-C, Liu Y-J, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ. 2005. Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* **309**: 626–630.

[http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?dbfrom=pubmed&id=15961632
&retmode=ref&cmd=prlinks.](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?dbfrom=pubmed&id=15961632&retmode=ref&cmd=prlinks)

Zhang Z, Dietrich FS. 2005. Identification and characterization of upstream open reading frames (uORF) in the 5' untranslated regions (UTR) of genes in *Saccharomyces cerevisiae*. *Curr Genet* **48**: 77–87.

Chapter 3

Alternative Splicing Contributes to Functional TL Diversity in Mouse

Abstract

Transcript Leaders (TLs, or 5'UTRs) are important determinants of post-transcriptional regulation of gene expression. Through the action of alternative promoters and alternative splicing, cells produce a wide array of alternative TL sequences. Through computational analyses of published annotations, we show that there are at least 14,580 alternative TL pairs in the mouse genome. Furthermore, we demonstrate that alternative splicing contributes substantially to TL diversity, generating nearly 6,000 alternative TLs. Notably, TL-contained alternative exons are enriched for upstream AUGs (uAUGs), suggesting that a main function of alternative splicing in TLs is to regulate translation. These results lay the groundwork for further genome-wide studies of alternative mammalian TLs and their functions in post-transcriptional regulation.

Introduction

Post-transcriptional regulation of gene expression is carried out primarily through the noncoding portions of an mRNA: the Transcript Leader (TL, or 5'UTR) upstream of the open reading frame (ORF), and the 3'UTR downstream. The TL is the portion of the mRNA from the 5' methyl-7-guanosine (m7G) cap to the start codon of the ORF. The TL has a critical role during translation initiation—through the action of the cytoplasmic cap-binding complex, a small ribosomal subunit is recruited to the TL, then linearly scans until it finds the AUG of the ORF. As discussed in the introduction (Chapter 1), this process is subject to extensive regulation by the TL.

TLs in mammals are known to control the translation of the downstream ORF, sometimes with phenotypic consequences. Upstream AUGs (uAUGs) are a notable

example of a TL-contained sequence that is present in ~40% of mammalian transcripts and is associated with a decrease in production of the downstream protein (Calvo et al. 2009). TL sequences are important for a normal cellular phenotype: a mutation that introduces a uORF in the TL of *Cdkn2A* causes a predisposition to melanoma, and a mutation that ablates a uORF in thrombopoietin leads to an increase in thrombopoietin protein production causing hereditary thrombocythemia (Liu et al. 1999; Wiestner et al. 1998). uAUGs in the TL of the transcription factor *Atf4* are important for its translational upregulation in response to dsRNA, ER stress, and nutritional stresses (Ron and Harding 2007). TL-dependent translational regulation can also occur via TL-contained RNA secondary structures (Babendure et al. 2006), RNA-binding proteins (Tsai et al. 2007), or a combination of both (Muckenthaler et al. 1998). These examples highlight the capacity of TLs to regulate translation, and the importance of this regulation for normal cellular physiology.

Many mammalian genes produce transcripts with alternative TLs with distinct activities and functions. Human *NOD2*, *ERβ*, and *DICER1* all have alternative TLs whose expression differs across a panel of tissues, and furthermore each of these TL variants confers a different translational efficiency on the downstream ORF (Smith et al. 2009; Rosenstiel et al. 2007; Singh et al. 2005). The neurotrophin *BDNF* has at least nine distinct TLs with non-redundant functions (Baj and Tongiorgi 2008; Pruunsild et al. 2007). Alternative TLs can also code for distinct N-termini, as has been shown in the cannabinoid receptor *CBI* (Shire et al. 1995), as well as numerous examples in yeast (*SUC2* (Carlson and Botstein 1982), *HTS1* (Chiu et al. 1992), and *KAR4* (Gammie et al.

1999)). These examples highlight the importance of alternative TLs and their distinct translational activities.

Known contributors to alternative TL diversity include alternative promoters and alternative splicing. Throughout this chapter we distinguish between alternative promoters, which create TL diversity by alternative Transcription Start Sites (TSSs), and alternative splicing, which creates alternative isoforms by inclusion/exclusion of sequences downstream of the TSS. In a panel of human tissues, >10,000 alternative promoter pairs were detected, making alternative promoters a substantial contributor to TL isoform diversity (Wang et al. 2008). Furthermore, genes with alternative promoters are enriched for regulatory functions, suggesting that alternative promoters and TLs play a crucial role in cellular adaptation (Resch et al. 2009). There are also many anecdotal examples where alternative splicing contributes to TL isoform diversity and translation regulation. A retained intron in the *CBI* gene encodes an N-terminal variant (Shire et al. 1995). Skipped exons in the TLs of *NOD2*, *ERβ*, *BDNF*, and *DICER1* generate functional TL variants (Smith et al. 2009; Rosenstiel et al. 2007; Singh et al. 2005; Pruunsild et al. 2007). However, the role of alternative splicing in generating TL diversity genome-wide has not been examined.

In this chapter we undertake a computational analysis of alternative TLs in mouse. Using existing annotations, we identify over 14,580 alternative TL pairs, the majority of which are produced by alternative promoters. We identify 5,801 alternative TL pairs created through alternative splicing, demonstrating that alternative splicing is a significant contributor to TL diversity genome-wide. We show that skipped exons in TLs are enriched for uAUGs, demonstrating that alternative splicing events affecting TLs are

likely functional. These results enumerate alternative mRNA processing events as they pertain to TLs, and suggest that alternative splicing may be important for TL-dependent translational regulation.

Materials and Methods

Annotations

ENSEMBL Transcript annotations for mouse (*Mus musculus*, mm10) were downloaded from the UCSC genome browser. For analysis of alternative events, coding transcripts (gene_biotype = protein_coding) that shared at least one splice site were pooled and the resulting group was analyzed for each of the events shown. For transcripts that lacked open reading frame (ORF) annotations, ORFs were annotated by identifying the longest ORF in a transcript longer than 100 amino acids.

Percent Spliced In Analysis

Percent Spliced In (PSI, Ψ) was calculated as fragments per kilobase per million (fpkm) of transcripts that included the exon divided by the fpkm of all transcripts that spanned the exon, as previously described in (Merkin et al. 2012). Average Ψ was the mean of all Ψ values across a panel of nine mouse tissues: heart, brain, spleen, liver, testes, colon, kidney, lung, and muscle (Merkin et al. 2012). All analyses were performed with custom python scripts and plotted using PyX.

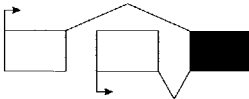
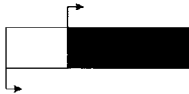
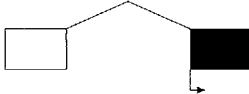
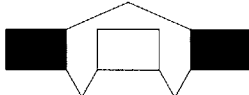
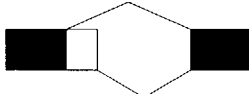
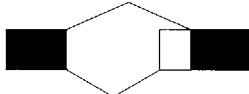

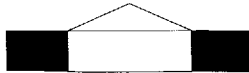
Results

Mouse Alternative TLs are Produced by Alternative Promoters and Alternative Splicing

To understand the contributions of alternative mRNA processing events to alternative TLs, we examined the number of alternative promoters affecting TLs in mouse. On par with what was observed in humans (Wang et al. 2008), we observed 9,049 alternative TLs generated by Alternative First Exons (AFEs) (Figure 3.1). Tandem TSSs and skipped first exons together were the largest contributors to TL diversity, generating 34,447 alternative TLs. However, a substantial number of these events are possibly due to annotation errors as the 5' end of annotations are notoriously inaccurate (Dike 2004), and events in both classes are difficult to differentiate from internal capture artifacts. Ignoring tandem TSSs and skipped first exons, AFEs alone generate over 9,000 alternative TL pairs.

Because ~40% of mouse TLs span more than one exon (Appendix I, Figure I.1), it is feasible that alternative splicing events also contribute to TL isoform diversity. Indeed, we observed 5,801 alternative TL pairs generated by alternative splicing events (Figure 3.1 sum of Skipped Exon, Alternative 5' Splice Site, Alternative 3' Splice Site, Retained Intron, Mutually Exclusive Exons). Furthermore, 20% of all alternative splicing events anywhere in a transcript occurred in TLs, indicating that a substantial amount of alternative splicing directly affects TL diversity. However, genome-wide studies of alternative splicing have primarily focused on the regulation of ORF-internal exons (Ellis et al. 2012; Merkin et al. 2012; Lewis et al. 2003). We note that the number of alternative TLs generated by alternative splicing (5,801) is over half the number of alternative TLs generated by AFEs (9,049), demonstrating that alternative splicing contributes

significantly to the production of alternative TLs. Thus alternative splicing in TLs is a pervasive and under-studied aspect of mRNA processing that may have widespread impacts on the regulation of gene expression in mammals.

Event Type (Number)	Schematic	ORF Start Site Location Breakdown
Alternative First Exon (AFE, 9049)		Start Site Downstream: 4800 Start Site in One Alt Exon: 2676 Start Site in Both Alt Exons: 1573
Tandem Transcription Start Site (33705)		Start Site Downstream: 30785 Start Site in Alt Exon: 2920
Skipped First Exon (742)		Start Site Downstream: 287 Start Site in Alt Exon: 455
Skipped Exon (SE, 1512) 8881 events anywhere in a transcript		Start Site Downstream: 1255 Start Site in Alt Exon: 257
Alternative 5'SS (2336) 5047 events anywhere in a transcript		Start Site Downstream: 1996 Start Site in Alt Exon: 340
Alternative 3'SS (819) 4667 events anywhere in a transcript		Start Site Downstream: 705 Start Site in Alt Exon: 114
Mutually Exclusive Exons (114) 758 events anywhere in a transcript		Start Site Downstream: 79 Start Site in Alt Exon: 23 Both Start Sites in Alt Exons: 12
Retained Intron (1020) 6354 events anywhere in a transcript		Start Site Downstream: 954 Start Site in Alt Exon: 66

Total Events: 49297

Figure 3.1: Alternative Events that Contribute to TL Diversity

For each event, the constitutive regions are shaded and alternative regions outlined. All events that change at least one base in the TL are included. Arrows indicate start of

transcription; for alternative splicing events, transcription began upstream and is not indicated in the diagram. For all depicted events, the ORF start codon lies in the alternative region or downstream of it. The last column indicates the number of event types further broken down by location of the annotated ORF start codon for both isoforms.

Alternative Splicing in TLs Regulates uAUG Abundance

To determine if TL-contained alternative splicing might be functional, we analyzed the prevalence of known TL functional elements in alternative exons. Alternative splicing in ORFs is known to regulate exons affecting nonsense-mediated mRNA decay (NMD), protein-protein interactions, and protein phosphorylation (Lewis et al. 2003; Ellis et al. 2012; Merkin et al. 2012). Upstream AUGs are a readily identifiable functional element that regulate translation of downstream ORFs by direct competition for ribosomes. In mammals and yeast, uAUGs also regulate transcripts by inducing NMD (Hurt et al. 2013) (Chapter 2). We examined the prevalence of uAUGs in TL-contained skipped exons (SEs) and observed an enrichment relative to upstream constitutive exons (Figure 3.2A, $p\text{-value}=2.85e-45$, Fisher's Exact Test). Furthermore, we also observed that SEs that are efficiently included (high average percent spliced in, PSI, Ψ) tended to a lower frequency of uAUGs than SEs that were more commonly spliced out (Figure 3.2B). Together this shows that constitutive exons tend to have a lower uAUG frequency than alternative exons in TLs. Given the known widespread functional role of uAUGs, the skipped exons harboring them likely confer functional differences. Thus alternative splicing generates TL variants with strongly predicted distinct functions that influence translation and mRNA decay of their transcripts.

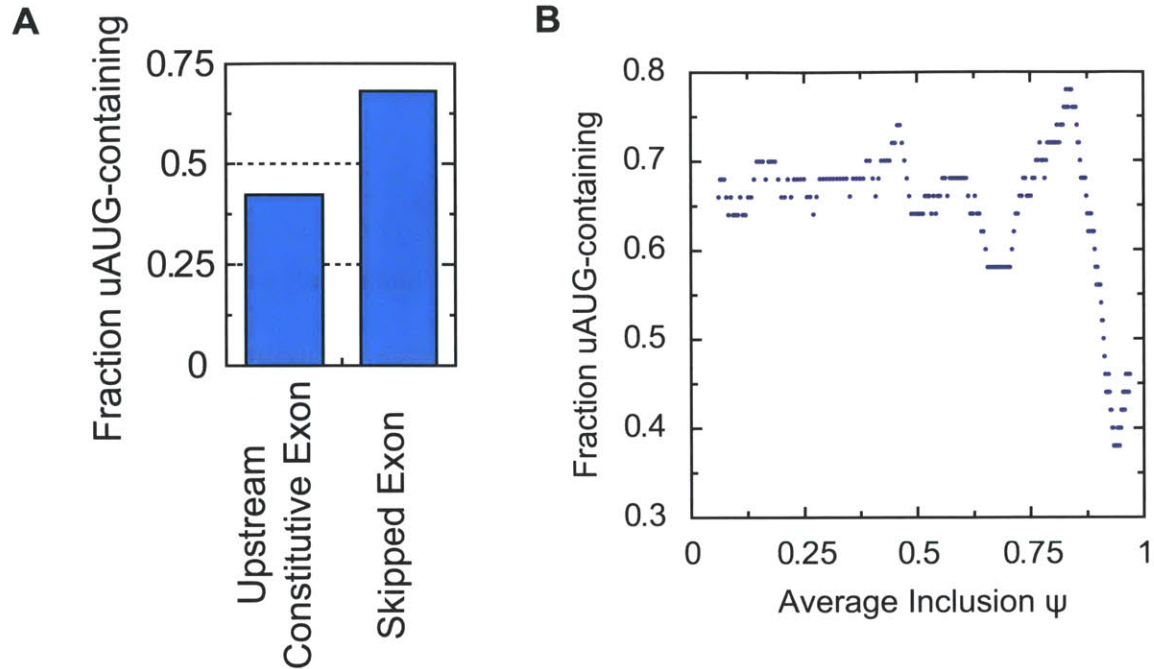


Figure 3.2: Alternative Splicing Affects the Abundance of uAUGs

(A) Fraction of exons containing at least one upstream AUG (uAUG). The fraction uAUG-containing was calculated for skipped exons or their upstream constitutive exons. (B) Fraction of skipped exons containing at least one uAUG versus percent spliced in (Ψ). For each skipped exon, its average Ψ was calculated across a panel of nine mouse tissues (Merkin et al. 2012). Exons were sorted by average Ψ , binned into 50 exon groups, and fraction uAUG-containing calculated.

Discussion

Alternative TLs are a prominent feature of higher eukaryotes and contribute substantially to mRNA isoform diversity. Through analysis of extant annotations, we enumerate different mRNA processing events that create TL diversity. Similar to what was observed in humans, AFEs generate 9,049 alternative TLs in mouse. In previous reports, we show >10,000 mouse TLs generated by Alternative First Exons. We also observed almost 6,000 alternative TL pairs generated by alternative splicing, thus showing that alternative splicing contributes substantially to TL diversity. TL-contained skipped exons are enriched for uAUGs, suggesting an important functional role for alternative splicing in

regulating translation. All together, these results suggest that alternative splicing is an important contributor to post-transcriptional regulation through production of alternative TLs.

Our results here demonstrate that there are at least 14,850 alternative TLs in mouse, providing a rich regulatory landscape for translation. Nine nucleotide differences in TL lengths are sufficient to confer distinct translation efficiencies in yeast (Rojas-Duran and Gilbert 2012), and thus in mammals, where alternative TLs often differ by a hundred or more nucleotides, translation behaviors driven by alternative TLs are likely to be extremely diverse. Consistent with this, in every example in which it's been assessed in mammals, alternative TLs confer distinct translational efficiencies (Smith et al. 2009; 2010) as well as different translational responses to cellular signaling pathways such as mTOR (Rosenstiel et al. 2007). Here we show alternative splicing generates many thousand alternative TL pairs differing by one or more uAUGs, thus presenting evidence for widespread regulation by alternative TLs.

Acknowledgements

We thank Jason Merkin for helpful discussion regarding computational analysis and splicing.

References

Babendure JR, Babendure JL, Ding J-H, Tsien RY. 2006. Control of mammalian translation by mRNA structure near caps. *RNA* **12**: 851–861.

Baj G, Tongiorgi E. 2008. BDNF splice variants from the second promoter cluster

- support cell survival of differentiated neuroblastoma upon cytotoxic stress. *Journal of Cell Science* **122**: 156–156.
- Calvo SE, Pagliarini DJ, Mootha VK. 2009. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci USA* **106**: 7507–7512.
- Carlson M, Botstein D. 1982. Two differentially regulated mRNAs with different 5' ends encode secreted with intracellular forms of yeast invertase. *Cell* **28**: 145–154.
- Chiu MI, Mason TL, Fink GR. 1992. HTS1 encodes both the cytoplasmic and mitochondrial histidyl-tRNA synthetase of *Saccharomyces cerevisiae*: mutations alter the specificity of compartmentation. *Genetics* **132**: 987–1001.
- Dike S. 2004. The mouse genome: Experimental examination of gene predictions and transcriptional start sites. *Genome Research* **14**: 2424–2429.
- Ellis JD, Barrios-Rodiles M, Çolak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O'Hanlon D, Kim PM, et al. 2012. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular Cell* **46**: 884–892.
- Gammie AE, Stewart BG, Scott CF, Rose MD. 1999. The two forms of karyogamy transcription factor Kar4p are regulated by differential initiation of transcription, translation, and protein turnover. *Mol Cell Biol* **19**: 817–825.
- Hurt JA, Robertson AD, Burge CB. 2013. Global analyses of UPF1 binding and function reveals expanded scope of nonsense-mediated mRNA decay. *Genome Research*.

- Lewis BP, Green RE, Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci USA* **100**: 189–192.
- Liu L, Dilworth D, Gao L, Monzon J, Summers A, Lassam N, Hogg D. 1999. Mutation of the CDKN2A 5' UTR creates an aberrant initiation codon and predisposes to melanoma. *Nat Genet* **21**: 128–132.
- Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science* **338**: 1593–1599.
- Muckenthaler M, Gray NK, Hentze MW. 1998. IRP-1 binding to ferritin mRNA prevents the recruitment of the small ribosomal subunit by the cap-binding complex eIF4F. *Molecular Cell* **2**: 383–388.
- Pruunsild P, Kazantseva A, Aid T, Palm K, Timmusk T. 2007. Dissecting the human BDNF locus: Bidirectional transcription, complex splicing, and multiple promoters. *Genomics* **90**: 397–406.
- Resch AM, Ogurtsov AY, Rogozin IB, Shabalina SA, Koonin EV. 2009. Evolution of alternative and constitutive regions of mammalian 5'UTRs. *BMC Genomics* **10**: 162.
- Rojas-Duran MF, Gilbert WV. 2012. Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA* **18**: 2299–2305.
- Ron D, Harding HP. 2007. 13 eIF2 α Phosphorylation in Cellular Stress Responses and Disease. *Cold Spring Harbor Monograph Archive* **48**: 345–368.

- Rosenstiel P, Huse K, Franke A, Hampe J, Reichwald K, Platzer C, Roberts RG, Mathew CG, Platzer M, Schreiber S. 2007. Functional characterization of two novel 5' untranslated exons reveals a complex regulation of NOD2 protein expression. *BMC Genomics* **8**: 472.
- Shire D, Carillon C, Kaghad M, Calandra B, Rinaldi-Carmona M, Le Fur G, Caput D, Ferrara P. 1995. An amino-terminal variant of the central cannabinoid receptor resulting from alternative splicing. *J Biol Chem* **270**: 3726–3731.
- Singh S, Bevan SC, Patil K, Newton DC, Marsden PA. 2005. Extensive variation in the 5'-UTR of Dicer mRNAs influences translational efficiency. *Biochemical and Biophysical Research Communications* **335**: 643–650.
- Smith L, Brannan RA, Hanby AM, Shaaban AM, Verghese ET, Peter MB, Pollock S, Satheesha S, Szykiewicz M, Speirs V, et al. 2009. Differential regulation of oestrogen receptor β isoforms by 5' untranslated regions in cancer. *Journal of Cellular and Molecular Medicine* **14**: 2172–2184.
- Smith L, Coleman LJ, Cummings M, Satheesha S, Shaw SO, Speirs V, Hughes TA. 2010. Expression of oestrogen receptor β isoforms is regulated by transcriptional and post-transcriptional mechanisms. *Biochem J* **429**: 283–290.
- Tsai N-P, Bi J, Wei L-N. 2007. The adaptor Grb7 links netrin-1 signaling to regulation of mRNA translation. *EMBO J* **26**: 1522–1531.
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes.

Nature **456**: 470–476.

Wiestner A, Schlemper RJ, van der Maas AP, Skoda RC. 1998. An activating splice donor mutation in the thrombopoietin gene causes hereditary thrombocythaemia. *Nat Genet* **18**: 49–52.

Chapter 4

Conclusions and Future Directions

Introduction

This research focuses on understanding the roles of TLs in translation and developing tools to explore TL-mediated translational regulation in eukaryotic systems. We develop Transcript Leader Sequencing (TL-seq), a technique to annotate TLs in any eukaryotic organism. Using TL-seq in yeast, we find ORF-internal TSS events that increase the number of N-terminally truncated proteins that are generated. We also identify short TLs on hundreds of endogenous genes and show that this feature is associated with inefficient initiation at cap-proximal AUGs concomitant with initiation at further downstream AUGs, leading to NMD of the transcript. Using Translation-Associated TL-seq (TATL-seq), we demonstrate a genome-wide function for uAUGs in yeast in reducing protein expression from the downstream ORF and triggering NMD of the transcript. We also identify hundreds of genes with alternative TLs that exhibit distinct translational activities in yeast. By computational analysis of existing annotations, we note that TL isoform variants are a prominent feature of mammalian genomes, with contributions from both alternative promoters and alternative splicing. We demonstrate a role for alternative splicing in regulating the abundance of uAUG-containing TL isoforms, and thus translational control by TLs. In the following chapter we review these findings and highlight potential avenues of future research.

Annotating TLs

Dissatisfied with the available techniques to annotate TLs, we developed TL-seq, a straightforward and versatile protocol to identify TLs. TL-seq works on the same principle as 5'RACE: replacement of the cap with an adaptor of known sequence, then

recovery of adaptor-containing RNAs. In Chapter 2 we demonstrated that TL-seq enriched for the 5' ends of transcripts and used this protocol to study TLs in yeast. The potential for TL-seq to study TLs in other systems prompted us to apply TL-seq to mammalian RNAs as well. In Appendix I, we modified and optimized TL-seq, making it compatible with paired-end sequencing and reducing input requirements. All together, these efforts have led to the creation of a novel technique capable of isolating and identifying TLs from a pool of eukaryotic mRNAs.

There are currently many TL identification techniques, including TL-seq (similar to CIP-TAP (Gu et al. 2012)), nanoCAGE, PEAT, RAMPAGE, and TIF-seq, and these techniques differ in input amount requirements, TL enrichment steps, and transcript information produced (Ni et al. 2010; Plessy et al. 2010; Batut et al. 2013; Pelechano et al. 2013; Arribere and Gilbert 2013). For example, TL-seq requires one microgram of mRNA and uses adaptor ligation, RAMPAGE requires a few micrograms of total RNA and uses template switching and cap biotinylation, and TIF-seq uses adaptor ligation with cDNA circularization to determine both the 5' and 3' ends of transcripts. There are obvious instances where one protocol is preferred: because TIF-seq (and the related RNA-PET (Ruan and Ruan 2012)) is the only protocol of these that can simultaneously annotate the 5' and 3' ends of transcripts it is the only choice for annotation of both transcript ends. There are also distinct biases in each protocol: T4 RNA ligase is known to exhibit sequence preferences (Sorefan et al. 2012) and template switching produces chimeric DNAs (Tang et al. 2013). Direct comparisons of these techniques will help illuminate the biases inherent in each method, how these biases impact the nature of the data produced, and how biases can be reduced for future applications. A direct

comparison would also help researchers determine which of these techniques is best suited for TL annotation in their sample.

Computational Tools for TL Analysis

While TL-seq enriched for 5'ends, non-TSS species were also recovered, motivating our development of a peak-calling algorithm to identify TSSs. In Chapter 2 we employed a peak-calling algorithm that identified regions of high read density normalized to mRNA expression level and demonstrated the resulting peaks exhibit TSS characteristics. We presented just one example of a peak-calling algorithm, and like ChIP-seq peak-callers, algorithms based on the shape of read distributions (e.g. (Zhang et al. 2008)) or enrichment relative to control libraries (e.g. (Rozowsky et al. 2009)) are also possible. Indeed, we observed that TL-seq peaks upstream of ORFs were broader than those inside of ORFs, suggesting that shape-based peak-calling algorithms may be useful. In addition to peak-callers, in Chapter 2 we utilized ChIP-seq data and TL-seq minus pyrophosphatase to reduce background and increase confidence in the TSSs identified. Since all TSS identification techniques have a non-TSS background, peak-calling algorithms and filters are broadly applicable and indispensable *in silico* approaches to discern TSSs from non-TSS artifacts.

In Chapter 2 we used Shape Index (described in (Hoskins et al. 2011)) to quantify and put a single numeric value on the heterogeneity of the TSS (TL) distribution for a gene. Consistent with what was seen in mammals (Resch et al. 2009), we observed that yeast genes encoding regulatory functions tended to have heterogeneous TLs. While we were unable to correlate TSS heterogeneity with chromatin features, this may have been a

technical failure as yeast ChIP-seq and ChIP-chip datasets are often low resolution. Future endeavors with higher resolution chromatin immunoprecipitation datasets, or TL-seq in other systems, may prove informative in identifying the transcriptional cause of TL heterogeneity. Regardless of the cause of TL heterogeneity, metrics such as Shape Index are useful in quantifying and analyzing TL heterogeneity. Shape Index can detect single-nucleotide TL heterogeneity, whereas our peak-calling algorithm filtered out lower abundance TSSs and could only detect TLs separated by at least 50 nts. Thus Shape Index is a less biased metric for studying TL heterogeneity.

Unexpected TSSs Identified by TL-seq

Surprisingly, we found that ~6% of TSSs occurred within annotated ORF boundaries, over 85% of which were within the first 100nts of an ORF. Internal TSSs included many previously known examples that produce N-terminal truncations (Gammie et al. 1999; Wu and Tzagoloff 1987; Chiu et al. 1992). While a few internal TSSs were misannotation events, in a majority of internal TSS cases, we noted amino acid conservation upstream of the major internal TSS. While TL-seq, RNA-seq, and Ribo-seq supported a major TSS internal to the annotated ORF start codon, for a majority of cases we also observed a minority of reads from upstream positions, indicating the existence of lower abundance, longer TL isoforms. Taken together, this data argues that the extended N-terminus may be functional, but not highly expressed under the conditions examined. The ratio of long to internal TSS isoforms and N-terminal protein variants may change in a condition-specific manner, as has been observed previously (Law et al. 2005; Gammie et al. 1999;

Carlson and Botstein 1982). Our data suggest N-terminal TSSs might be a more common way of creating and regulating protein diversity than previously appreciated.

Protocols such as TL-seq will likely prove useful in future endeavors to look at cryptic initiation events genome-wide. Here, <1% of TSSs were >100nts into the ORF, and while it is unclear whether these events are mechanistically related to cryptic initiation (see Chapter 2 Discussion), TL-seq and similar TSS identification techniques may prove useful in studying cryptic initiation in the future. An obvious issue is the high number of internal, probably artifactual peaks (Chapter 2, Figure 2.6), though this may be ameliorated with further technical improvements to TL-seq such as additional cap-selection steps (e.g. cap-trapper, (Carninci and Hayashizaki 1999)). Even with its current internal artifacts, TL-seq may prove informative for defining TSSs in conjunction with other high throughput datasets, such as RNA-seq or ChIP-seq. For example, in Chapter 2 we used nucleosome positions to estimate the number of internal TSSs detected by TL-seq. Currently only low-throughput methods exist to look at cryptic transcription events in a targeted way, but it is likely only a matter of time until a global view of cryptic transcription is possible.

We also identified several hundred yeast mRNAs with TLs less than 12 nt long, and showed that these mRNAs often induce NMD. We demonstrated in Chapter 2 that there is a substantial population of normal cellular mRNAs in yeast with short TLs that cause downstream initiation and, when the downstream AUG is out-of-frame, NMD. As expected, the amount of out-of-frame translation initiation was correlated with the extent of NMD of the mRNA. Why might a cell produce short TLs if the mRNAs are only to be degraded? Short TLs present the potential for concerted post-transcriptional regulation of

a class of mRNAs. Conditions such as hypoxia regulate efficiency of NMD targeting in yeast (Gardner 2008), and in mammals the efficiency of cap-proximal recognition on short TLs is controlled by eIF1 levels (Pestova and Kolupaeva 2002). It's also possible that longer TLs are generated from short TL-containing genes during stress, as TSS locations change in different cell states in yeast (Law et al. 2005). Thus short TLs present an opportunity for TL-mediated post-transcriptional regulation.

It is unclear to what extent short TLs affect gene expression in organisms other than yeast. Poor initiation efficiency at cap-proximal AUGs has been observed in mammalian systems, along with a tendency to initiate at downstream AUGs (Sedman et al. 1990; Kozak 1991; Pestova and Kolupaeva 2002), raising the possibility that short TL genes may be NMD targets there as well. However, a subset of mammalian short TL genes initiate translation efficiently and do so through a specialized scanning-independent pathway that is characterized by a distinct non-Kozak AUG context, the so-called "TISU element", present in 4.5% of protein-coding transcripts (Elfakess and Dikstein 2008; Elfakess et al. 2011). Due to their efficient cap-proximal initiation, TISU-containing mRNAs would not be expected to be NMD targets. It is possible that such transcripts would be subject to NMD if they were translated via canonical initiation, perhaps under conditions in which the alternative TISU initiation pathway is less active. Thus the potential for regulation of short TL genes by NMD may be more complicated in mammalian systems.

TATL-seq: Directly Analyzing TLs in Translation

To enable the direct study of TLs in translation, in Chapter 2 we developed Translation-Associated TL-seq (TATL-seq). We fractionated mRNAs according to translation activity on a polysome gradient and identified the pool of TLs in each fraction, providing information on TL features that differ between well and poorly translated mRNAs. This approach provides two distinct and complementary measures of translation: the fraction of a gene's mRNA associated with ribosomes (ribosome occupancy) and the average density of ribosomes on an mRNA (ribosome density) (as defined by (Arava et al. 2003)). We can use these metrics to study TL-mediated translation effects directly, rather than trying to relate ORF-based measurements to TL properties *in silico*. The latter approach is especially problematic when the TL sequence is unknown or there are multiple TLs adjoined to the same ORF sequence. TATL-seq is unique in that it provides both *de novo* annotation of TLs and an assay for TL functionality in translation.

Upstream AUGs (uAUGs) are a notable TL feature we observed to be an important TL-contained functional element. We show that uAUGs have a negative impact on translation and stimulate mRNA decay by NMD genome-wide in yeast. Importantly, while we observe that uAUGs are selected against in TLs, we also observe that they are conserved, arguing they serve an important function. Downregulation of protein production is an important function of uAUGs (Liu et al. 1999), though it is also possible that uAUGs are important for some as yet unidentified regulatory phenomenon. Possibilities include regulation via uORF peptide-mediated ribosomal stalling, reinitiation, or uORF recognition (Polymenis and Schmidt 1997; Wang et al. 1999; Mueller and Hinnebusch 1986). Since there are few known characteristic sequence signatures of these events, it is difficult to look for them genome-wide. Future studies

will hopefully elucidate the gene-specific and genome-wide functions of uORFs in post-transcriptional regulation.

Our results argue against a widespread physiological role in translation regulation or NMD for translation initiation at non-AUG codons in TLs. Initiation at non-AUG codons was proposed to be a frequent event in TLs, esp. under conditions including amino acid starvation (Ingolia et al. 2009). However, unlike uAUGs, other upstream codons are not associated with NMD and decreased translation (Chapter 2, Figure 2.11). Similar to amino acid starvation, an increase in Ribo-seq reads at non-AUG codons was also observed in glucose starvation and a hypomorphic eIF4E mutant (Appendix II). At least for glucose starvation, the increase in TL reads at non-AUG codons was cycloheximide-dependent. We conclude that Ribo-seq TL-mapping reads are more likely an effect of incubating cells with a translation elongation inhibitor than a global relaxation on initiation specificity. Future endeavors should focus on determining a method to more accurately study translation in TLs, thus enabling direct observation of uORF translation.

Alternative TLs

Using TATL-seq in yeast, we identified over 200 yeast genes with at least two TLs, and showed that these TLs are associated with distinct translational behaviors. This is surprising for a eukaryote that is neither known nor prized for its mRNA isoform diversity. Why might the cell produce so many long TLs when they are, on average, less well translated than the shorter isoforms? Having more TLs affords a greater regulatory potential, possibly preparing the cell for environmental stressors. Consistent with this

view, the longer isoforms exhibited a two-fold higher uAUG frequency, suggesting that uORF-dependent regulation may play a role in their expression. A uORF-containing TL may enable the continued expression of the downstream ORF under conditions of global inhibition of translation, as is known to happen for *GCN4* during amino acid starvation.

While alternative TLs occur for a few hundred genes in yeast, they are a pervasive feature of mammalian genomes, and we anticipate that TATL-seq will be useful for their study. There are over 10,000 alternative TLs in human (Wang et al. 2008) and at least as many in mouse (Chapter 3). Genes with alternative promoters and alternative TLs are enriched for regulatory functions, suggesting an important role in cellular adaptation (Resch et al. 2009). Because many TL variants do not change the coding potential of the ORF, they are invisible to ORF-based measurements such as Ribo-seq. Furthermore, noncoding TL variants must exert their effects during the post-transcriptional life of an mRNA. Techniques such as TATL-seq will likely prove useful in elucidation of the molecular function of alternative TLs. The potential for TATL-seq to visualize and analyze alternative TLs prompted us to adapt this technique to mammalian systems (Appendix I). TATL-seq now provides a basis for studying alternative TLs as they affect the translational activity of their respective mRNA.

To understand the diversity of alternative TLs in mammals, we enumerated TL-affecting alternative mRNA processing events and found a large number of functionally distinct TLs produced via alternative splicing. While alternative TLs are often generated by alternative promoters, we observed that nearly 6,000 alternative TL pairs are generated through alternative splicing (Chapter 3). Alternative splicing is known to be functionally important genome-wide for the regulation of protein-protein interactions,

protein phosphorylation sites, and NMD isoforms (Ellis et al. 2012; Merkin et al. 2012; Lewis et al. 2003). In Chapter 3 we add translation regulation to this list of processes affected by alternative splicing: skipped exons are enriched for uAUGs, suggesting that TL-contained alternative splicing regulates the capacity of TLs to regulate translation. Just as accurate identification of the TSS is necessary to define the TL, so too are splicing events. These observations motivate the use of TSS sequencing protocols that provide sequence information hundreds of nucleotides beyond the TSS (e.g. TL-seq, RAMPAGE), thus defining the entire sequence of a TL.

While we identified 14,850 alternative TLs in mouse, this number is likely an underestimate. There are many promoters and exons that remain to be identified: recent large-scale sequencing across nine mouse tissues discovered tens of thousands of novel exons (Merkin et al. 2012). Also, the splicing and promoter events we listed in Chapter 3 (Figure 3.1) are not exhaustive. Alternative splicing involving multiple alternative exons, as well as alternative splicing combined with alternative promoters, may combinatorially increase the diversity of TLs. Here we ignore alternative TLs generated via tandem TSSs in mammals for technical concerns regarding false positives in TSS annotations.

However, hundreds of yeast genes with tandem TSSs produce alternative TLs with different translational behaviors (Chapter 2, Figure 2.14), demonstrating that at least in yeast, tandem TSSs occur and are often functional. Additionally, functionally relevant alternative TL isoforms in mammals are also generated via tandem TSSs (Smith et al. 2009; Singh et al. 2005). Future analyses should focus on reliably differentiating true TSSs from artifactual ones so as to detect and study tandem TSS events. With all of this

in mind, we anticipate that the true number of alternative TLs is likely tens of thousands, making alternative TLs a significant contributor to mammalian mRNA isoform diversity.

Final Remarks

This work deepens our understanding of post-transcriptional regulation by TLs. We demonstrate the functionality of both novel and known TL characteristics that influence the translation and overall stability of their mRNAs genome-wide in yeast. We show that TL isoform variation is produced by both transcription and alternative splicing, and alternative TLs are a pervasive and functional feature of eukaryotic gene expression. Many more TL characteristics that affect post-transcriptional regulation remain to be discovered, and we anticipate the techniques and work put forth here will be helpful in future TL and post-transcriptional analyses.

References

- Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D. 2003. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **100**: 3889–3894.
- Arribere JA, Gilbert WV. 2013. Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Research* **23**: 977–987.
- Batut P, Dobin A, Plessy C, Carninci P, Gingeras TR. 2013. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Research* **23**: 169–180.

- Carlson M, Botstein D. 1982. Two differentially regulated mRNAs with different 5' ends encode secreted with intracellular forms of yeast invertase. *Cell* **28**: 145–154.
- Carninci P, Hayashizaki Y. 1999. High-efficiency full-length cDNA cloning. *Meth Enzymol* **303**: 19–44.
- Chiu MI, Mason TL, Fink GR. 1992. HTS1 encodes both the cytoplasmic and mitochondrial histidyl-tRNA synthetase of *Saccharomyces cerevisiae*: mutations alter the specificity of compartmentation. *Genetics* **132**: 987–1001.
- Elfakess R, Dikstein R. 2008. A Translation Initiation Element Specific to mRNAs with Very Short 5'UTR that Also Regulates Transcription. *PLoS ONE* **3**: e3094.
- Elfakess R, Sinvani H, Haimov O, Svitkin Y, Sonenberg N, Dikstein R. 2011. Unique translation initiation of mRNAs-containing TISU element. *Nucleic Acids Res* **39**: 7598–7609.
- Ellis JD, Barrios-Rodiles M, Çolak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O'Hanlon D, Kim PM, et al. 2012. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular Cell* **46**: 884–892.
- Gammie AE, Stewart BG, Scott CF, Rose MD. 1999. The two forms of karyogamy transcription factor Kar4p are regulated by differential initiation of transcription, translation, and protein turnover. *Mol Cell Biol* **19**: 817–825.
- Gardner LB. 2008. Hypoxic inhibition of nonsense-mediated RNA decay regulates gene expression and the integrated stress response. *Mol Cell Biol* **28**: 3729–3741.

- Gu W, Lee H-C, Chaves D, Youngman EM, Pazour GJ, Conte D, Mello CC. 2012. CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as *C. elegans* piRNA precursors. *Cell* **151**: 1488–1500.
- Hoskins RA, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, et al. 2011. Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Research* **21**: 182–192.
- Ingolia NT, Ghacmmaghami S, Newman JRS, Weissman JS. 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**: 218–223.
- Kozak M. 1991. A short leader sequence impairs the fidelity of initiation by eukaryotic ribosomes. *Gene Expr* **1**: 111–115.
- Law GL, Bickel KS, Mackay VL, Morris DR. 2005. The undertranslated transcriptome reveals widespread translational silencing by alternative 5' transcript leaders. *Genome Biol* **6**: R111.
- Lewis BP, Green RE, Brenner SE. 2003. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci USA* **100**: 189–192.
- Liu L, Dilworth D, Gao L, Monzon J, Summers A, Lassam N, Hogg D. 1999. Mutation of the CDKN2A 5' UTR creates an aberrant initiation codon and predisposes to melanoma. *Nat Genet* **21**: 128–132.

- Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science* **338**: 1593–1599.
- Mueller PP, Hinnebusch AG. 1986. Multiple upstream AUG codons mediate translational control of GCN4. *Cell* **45**: 201–207.
- Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J. 2010. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature Publishing Group* **7**: 521–527.
- Pelechano V, Wei W, Steinmetz LM. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**: 127–131.
- Pestova TV, Kolupaeva VG. 2002. The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection. *Genes Dev* **16**: 2906–2922.
- Plessy C, Bertin N, Takahashi H, Simone R, Salimullah M, Lassmann T, Vitezic M, Severin J, Olivarius S, Lazarevic D, et al. 2010. Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nature Publishing Group* **7**: 528–534.
- Polymenis M, Schmidt EV. 1997. Coupling of cell division to cell growth by translational control of the G1 cyclin CLN3 in yeast. *Genes Dev* **11**: 2522–2531.
- Resch AM, Ogurtsov AY, Rogozin IB, Shabalina SA, Koonin EV. 2009. Evolution of alternative and constitutive regions of mammalian 5'UTRs. *BMC Genomics* **10**: 162.

- Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat Biotechnol* **27**: 66–75.
- Ruan X, Ruan Y. 2012. Genome wide full-length transcript analysis using 5' and 3' paired-end-tag next generation sequencing (RNA-PET). *Methods Mol Biol* **809**: 535–562.
- Sedman SA, Gelembiuk GW, Mertz JE. 1990. Translation initiation at a downstream AUG occurs with increased efficiency when the upstream AUG is located very close to the 5' cap. *J Virol* **64**: 453–457.
- Singh S, Bevan SC, Patil K, Newton DC, Marsden PA. 2005. Extensive variation in the 5'-UTR of Dicer mRNAs influences translational efficiency. *Biochemical and Biophysical Research Communications* **335**: 643–650.
- Smith L, Brannan RA, Hanby AM, Shaaban AM, Verghese ET, Peter MB, Pollock S, Satheesha S, Szykiewicz M, Speirs V, et al. 2009. Differential regulation of oestrogen receptor β isoforms by 5' untranslated regions in cancer. *Journal of Cellular and Molecular Medicine* **14**: 2172–2184.
- Sorefan K, Pais H, Hall AE, Kozomara A, Griffiths-Jones S, Moulton V, Dalmay T. 2012. Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence* **3**: 1–1.
- Tang DTP, Plessy C, Salimullah M, Suzuki AM, Calligaris R, Gustincich S, Carninci P. 2013. Suppression of artifacts and barcode bias in high-throughput transcriptome

analyses utilizing template switching. *Nucleic Acids Res* **41**: e44–e44.

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

Wang Z, Gaba A, Sachs MS. 1999. A highly conserved mechanism of regulated ribosome stalling mediated by fungal arginine attenuator peptides that appears independent of the charging status of arginyl-tRNAs. *J Biol Chem* **274**: 37565–37574.

Wu M, Tzagoloff A. 1987. Mitochondrial and cytoplasmic fumarases in *Saccharomyces cerevisiae* are encoded by a single nuclear gene FUM1. *J Biol Chem* **262**: 12275–12282.

Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Appendices

Appendix I

TL-seq Optimization for TL Identification in Mammalian Genomes

Abstract

Transcript Leaders (TLs or 5'UTRs) are an important feature of mammalian mRNAs, yet techniques to accurately define them have lagged. While many protocols exist to identify Transcription Start Sites (TSSs), these protocols often do not unambiguously define TLs, which can be hundreds of nucleotides long, span multiple exons, and encompass alternative splicing events. To enable unambiguous TL annotation in mouse, we revise the TL-seq method to sequence TL fragments of arbitrary size. The improved TL-seq protocol should increase the number of TLs detected and better discern between TL splice variants. We also optimize enzymatic steps to make TL-seq compatible with the lower input requirements of mammalian samples. The revised TL-seq protocol we present is a useful tool ready to define TLs in any eukaryotic RNA sample.

Introduction

Post-transcriptional regulation of gene expression is carried out primarily through the noncoding portions of an mRNA: the Transcript Leader (TL, or 5'UTR) upstream of the open reading frame (ORF), and the 3'UTR downstream. The TL is the portion of the mRNA from the 5' methyl-7-guanosine (m7G) cap to the start codon of the ORF and has critical roles in many steps of gene expression, including translation (Chapter 1). In order to understand the contribution of TLs to post-transcriptional regulation *in vivo*, accurate and genome-wide annotations are necessary. Importantly, since TLs are known to change in different cellular conditions (Law et al. 2005), techniques that can easily identify TLs in a sample *de novo* are desirable. To identify TLs, most protocols to-date have focused on identifying the 5' end of the TL, the Transcription Start Site (TSS).

While identifying TSSs is often sufficient to define TLs in organisms such as yeast, in mammals TSS identification alone does not define the TL sequence. Due to its compact genome and essentially complete coding annotations, yeast Transcription Start Sites (TSSs) can readily be assigned to the closest ORF on the same strand. Additionally, because splicing events are constitutive and known, one can simply look up a TL sequence in the genome given a TSS and its ORF. However, in mammalian genomes genes are spread out, with TSSs occurring tens or hundreds of kilobases away from the coding exons of their transcript. Alternative splicing is rampant and annotations are incomplete: of the mouse exons discovered by a recent large-scale study, ~13% were novel (Merkin et al. 2012). Thus, while in yeast knowledge about the location of a TSS is often sufficient to define a TL, in mammals the situation is more complex.

Given the prevalence and importance of alternative TLs, protocols that can readily identify and study them are desirable. Using TL-seq in yeast, we previously observed hundreds of TL variants with distinct translational behaviors (see Chapter 2). From the handful of known examples in mammals, it is anticipated that many alternative mammalian TLs will be functional as well (Rosenstiel et al. 2007; Smith et al. 2009; Baj and Tongiorgi 2008). However, techniques such as Capped Analysis of Gene Expression (CAGE) only sequence the first 20 bases of a TL and will be unable to discern alternative TLs that differ at sequences downstream of the cap. Thus techniques to differentiate between TL variants generated by alternative splicing in mouse must provide more information than just the TSS. Techniques compatible with paired-end sequencing such as RAMPAGE have demonstrated the utility of sequencing longer portions of the TL by

unambiguously linking promoter sequences with their downstream gene and increasing the number of annotated TLs (Batut et al. 2013).

In this appendix, we revisit the design of TL-seq and optimize it to annotate TLs in mammalian genomes. Because mammalian TLs are long and potentially span multiple exons, TL-seq was modified to sequence longer RNA fragments. We anticipate the modified TL-seq protocol can define ~90% of mouse TLs, and will better capture alternative TLs generated by alternative splicing. By optimizing the enzymatic steps we also decrease input requirements to ~1ug of mRNA, making TL-seq practical for application to a wide array of eukaryotic samples.

Materials and Methods

Transcript Leader Sequencing

TL-seq was performed as previously described with the following modifications. To obtain longer capped fragments, fragmentation time was shortened from 60' to 2'. Previously, after 5' adaptor ligation, ligated material was observed and purified via gel shift. To obviate the need for this step and the fragment size restriction it imposes, we switched to 5' ligation with a biotinylated adaptor. Ligated fragments were then purified by selection on streptavidin beads. For 5' ligation, phosphatase and pyrophosphatase-treated fragments were incubated with T4 RNA Ligase and 10ug biotinylated adaptor in 1X ligase buffer (50mM Tris pH 7.5, 10mM DTT, 10mM MgCl₂) supplemented with 1mM ATP. For 5' ligation there needs to be vast molar excess (here ~30-100X) of 5' adaptor to inhibit insert-insert ligation. For 3' ligation, RNA was incubated with

50pmole preadenylated adaptor and T4 RNA Ligase in 1X ligase buffer supplemented with PEG8000 to 15% without ATP.

Biotinylated Oligo Capture

After 5' ligation, biotinylated oligonucleotides were captured on MyOne Streptavidin C1 magnetic beads (Dyna) per the manufacturer's protocol with the following modifications: after sample binding and three washes with 1X B&W Buffer, the beads were twice washed with Stringent Wash Buffer (10mM Tris pH7.5, 0.5mM EDTA, 4M Urea, 0.01% Tween). In our hands the stringent wash was necessary to remove all detectable non-specific binding of RNAs to the beads. After stringent washes, biotinylated species were eluted from the beads with water for 2' at 95C.

Annotations

ENSEMBL Transcript annotations for mouse (*Mus musculus*, mm10) were downloaded from the UCSC genome browser. For analysis of alternative events, coding transcripts (gene_biotype = protein_coding) that shared at least one splice site were pooled and the resulting group was analyzed for each of the events shown. For transcripts that lacked open reading frame (ORF) annotations, ORFs were annotated by identifying the longest ORF in a transcript longer than 100 amino acids. All analyses were performed with custom python scripts and plotted using PyX.

Results

TL Fragments Hundreds of Nucleotides Long are Necessary to Define Mouse TLs

The goal of TL-seq is to define TL sequences genome-wide and enable their study. This requires identification of the TSS as well as the mRNA sequence between the TSS and start codon of the ORF. Since the m7G triphosphate cap structure is conserved among all eukaryotes, the enzymatic steps of TL-seq should work similarly and define TSSs in any eukaryotic organism, including mouse. However, the median TL length in mouse is ~140nt (Calvo et al. 2009) and can span multiple exons, much different than the average gene in yeast, where a 20-30nt TL is continuous in genomic space with its ORF. Thus in mammals, it is unclear whether TL-seq as previously implemented with a 50-80nt insert would be useful in defining the entire TL sequence.

We sought to determine the number of TLs that would be unambiguously defined by TL-seq using a single 50-80nt read by analyzing TL length and exonic structure of annotated mouse transcripts. For ~60% of transcripts, the TSS is in the same exon as the ORF start codon, and thus identifying the TL is similar to the yeast case: once the TSS has been identified, the TL can be determined by looking up the intervening sequence between the TSS and start codon (Figure I.1). For the remaining ~40% of transcripts where the TL spans at least two exons, we analyzed the total length of all exons upstream of the exon containing the ORF start codon. Here, a majority of TLs will be incompletely covered if only the first 50-80nts of the transcript is sequenced (Figure I.1, orange line). This becomes especially problematic if the TL spans alternative splice sites, annotated or not. We anticipate that such transcripts will prove problematic for TL annotation using TL-seq and many other TSS-identification techniques using single-end short reads. However, with paired-end sequencing using an RNA fragment size 300nts or longer (Figure I.1, blue line), TL-seq could unambiguously define ~90% of mouse TLs.

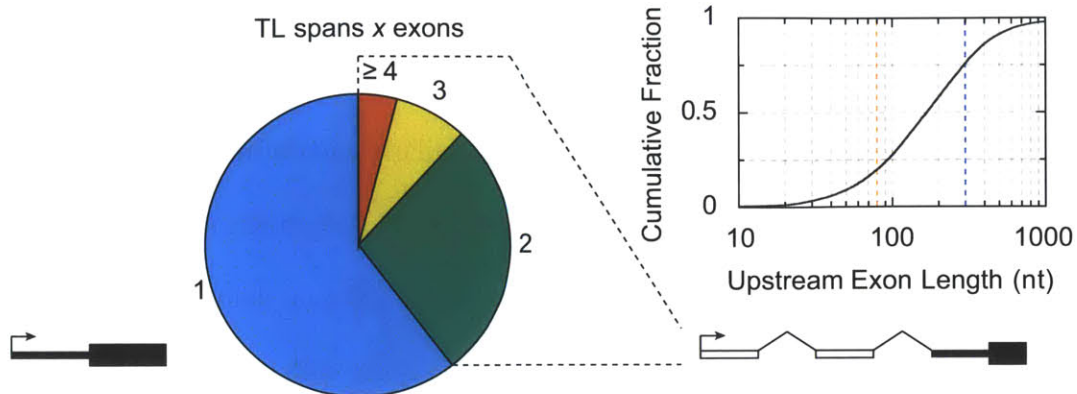


Figure I.1: TL Length and Exonic Structure For Mouse Transcripts

Fraction of transcripts with the TL spanning 1 to ≥ 4 exons (left). For those transcripts where the TL spans more than one exon, the total length of all upstream TL exons (open boxes) is shown in the cumulative distribution function (right). This does not include the TL sequence in the exon containing the ORF start (shaded box). On the graph at right, 80nt is highlighted in orange, 300nt highlighted in blue (see text).

There are at least 14,580 alternative TL pairs in mouse (Chapter 3), and it is important that TL-seq be able to identify these variants. With a single read of 50-80nt, TL-seq would in theory be able to distinguish TL variants generated by alternative promoters, but most of the TL-contained alternative splicing events would occur beyond the 3' end of such a short read and thus be undetectable. (We distinguish between alternative promoters, which create TL diversity by alternative Transcription Start Sites (TSSs), and alternative splicing, which creates alternative isoforms by inclusion/exclusion of sequences downstream of the TSS.) TL-contained alternative splicing produces nearly 6,000 alternative TLs and alternative exons are enriched for upstream AUGs (uAUGs), a known functional element (Chapter 3). Thus capturing TL-contained alternative splicing will likely be important for functional characterization of alternative TLs. If paired-end sequencing were undertaken with a library containing RNA fragments 300nt or longer, a majority of TL-contained alternative splicing events would

be defined. Thus, to comprehensively annotate alternative mammalian TLs, TL-seq should be adapted to sequence longer RNA fragments.

TL-seq Modification and Optimization for Mammalian Systems

We switched from a gel-based to a bead-based purification scheme to enable capture of longer TL fragments (Figure I.2). The strategy of TL-seq relies on selective adaptor ligation of capped RNA fragments and purification of those fragments by gel shift (Chapter 2, Figure 2.1B). A longer fragment size can be selected, but the width of the fragment distribution must still be shorter than the length of the 5' adaptor (45nts), otherwise the gel shift will overlap with the unligated fragment distribution. In the modified TL-seq scheme, a biotinylated 5' adaptor is used and the post-ligation gel purification step replaced with streptavidin selection. The biotin-streptavidin reaction is highly specific, with femtomolar affinity and low non-specific binding (Green 1975). The size distribution of captured TLs is determined only by the initial fragmentation reaction; thus this protocol can deal with capped fragments of much larger size than were previously isolated from yeast.

The input requirements for the previous implementation of TL-seq were impractical for mammalian samples, so we optimized product yield for each step in TL-seq. Using quantitative radioactive assays, we observed phosphatase and pyrophosphatase treatment to be close to 100% efficient. While ligation was extremely inefficient (5-10%), we were able to increase efficiency to ~30-60% by changing buffer conditions (addition of PEG8000 to 15% for 3' ligation) or adding excess biotinylated adaptor (5' ligation). The net improvement of these reactions should allow for a ~40-fold

reduction in the amount of input material, making TL-seq possible on ~1ug poly(A)+ RNA. Consistent with these estimates, we have successfully performed TL-seq on 1ug of input yeast RNA yielding inserts 140-200nt in length, and TOPO clones indicate the library has enriched for TLs.

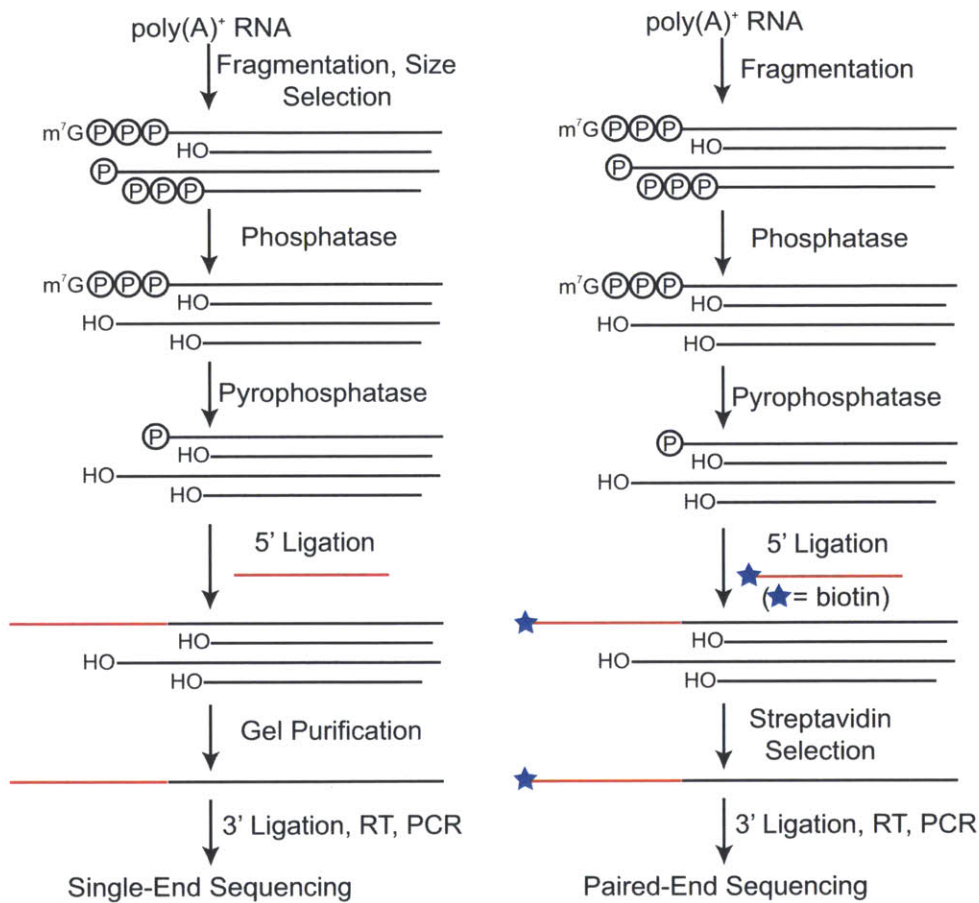


Figure I.2: TL-seq Modification to Sequence Longer TL Fragments With Less Input Shown at left is the original TL-seq protocol and on right the modified TL-seq protocol.

Discussion

In order to study TLs genome-wide, techniques need to exist which can easily annotate them. Here we show that ~40% of mouse TLs span multiple exons, and in these cases

simply identifying the TSS will incompletely define the TL sequence. Alternative TLs generated by alternative splicing in particular are not defined by TSS identification alone. To address these issues, we modify the TL-seq protocol to make it capable of sequencing long stretches of TL. We also undertake a careful quantification of enzymatic efficiencies in TL-seq, and reduce input requirements by almost two orders of magnitude. The final TL-seq technique provides a directed approach for TL identification in mammals and other eukaryotes.

TL-seq enables studies to look selectively at transcript architecture at 5' ends with greater sequencing depth and accuracy than RNA-seq. RNA-seq is 3' biased, producing fewer reads and less information at the 5' end of transcripts. Due to issues such as low processivity of reverse transcriptase, mRNA end information is imprecise in RNA-seq libraries. This imprecision may confound TL studies since differences of as few as nine nucleotides in TL length confer distinct translational efficiencies in yeast (Rojas-Duran and Gilbert 2012). Thus even with infinite sequencing power, RNA-seq will not yield the TSS information of TL-seq. RNA-seq still detect tens of thousands of novel exons in well-studied organisms such as mouse (Merkin et al. 2012), demonstrating that our list of exons is incomplete. Additional TL-contained exons that produce novel TL isoforms likely remain to be discovered. Detection of low abundance TLs are important; for example, yeast alternative TLs that are only ~10% of total transcript levels can exhibit robust translation and contribute to protein production for an ORF (Rojas-Duran and Gilbert 2012). Because it can identify and quantify novel and low abundance TL isoforms, TL-seq provides a targeted approach to study TLs more comprehensively and accurately than RNA-seq.

There are now multiple TSS identification techniques compatible with paired-end sequencers including TL-seq and RAMPAGE, and their technical details vary. TL-seq works by oligo-capping with a microgram of input mRNA. RAMPAGE captures RNA fragments by template switching and has two selection steps to enrich for capped fragments: RNase digestion of uncapped fragments, followed by cap biotinylation and selection with streptavidin (Batut et al. 2013). Both protocols have known biases: RNA ligation (TL-seq) is known to have sequence preferences (Sorefan et al. 2012), and template switching (RAMPAGE) can produce chimeric DNAs (Tang et al. 2013). Each protocol has its advantages as well. RAMPAGE can take a few micrograms of total RNA as input, making it the only option when the amount of sample is extremely low. During template switching, RAMPAGE adds untemplated G/C residues past the 5' end of the mRNA, potentially frustrating single nucleotide TSS identification. TL-seq, however, identifies the first nucleotide downstream of the cap. A direct side-by-side comparison of TL-seq and RAMPAGE will be useful for assessing the performance and pitfalls of the two techniques.

TL-seq is a useful tool that can be applied to any biological sample, potentially in conjunction with function-dependent fractionation, to study TLs and TSSs in any eukaryotic system. TL sequences bound by a RNA-binding protein of interest can be determined by performing RNA immunoprecipitation followed by TL-seq. This may prove useful in assessing the role of the RNA helicases eIF4A, Ded1, and Dbp1 in yeast, which are thought to function non-redundantly on different subsets of TLs (Chuang 1997). TL-seq could provide a more in-depth look at TSSs, alternative promoters, and their association with nearby splicing events. In conjunction with RNA isolation

following chromatin immunoprecipitation (Bittencourt and Auboeuf 2011), TL-seq can elucidate what TSSs are associated with a given chromatin factor. We have already demonstrated the utility of TL-seq to elucidate the translational activity of alternative TLs *in vivo* (TATL-seq, Chapter 2), and we anticipate that future studies employing TL-seq will provide important insight as to the roles of TLs in many aspects of eukaryotic gene expression.

References

- Baj G, Tongiorgi E. 2008. BDNF splice variants from the second promoter cluster support cell survival of differentiated neuroblastoma upon cytotoxic stress. *Journal of Cell Science* **122**: 156–156.
- Batut P, Dobin A, Plessy C, Carninci P, Gingeras TR. 2013. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Research* **23**: 169–180.
- Bittencourt D, Auboeuf D. 2011. Analysis of Co-transcriptional RNA Processing by RNA-ChIP Assay. In *Methods in Molecular Biology*, Vol. 809 of *Methods in Molecular Biology*, pp. 563–577, Springer New York, New York, NY.
- Calvo SE, Pagliarini DJ, Mootha VK. 2009. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci USA* **106**: 7507–7512.
- Chuang RY. 1997. Requirement of the DEAD-Box Protein Ded1p for Messenger RNA Translation. *Science* **275**: 1468–1471.

- Green NM. 1975. Avidin. *Adv Protein Chem* **29**: 85–133.
- Law GL, Bickel KS, Mackay VL, Morris DR. 2005. The undertranslated transcriptome reveals widespread translational silencing by alternative 5' transcript leaders. *Genome Biol* **6**: R111.
- Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science* **338**: 1593–1599.
- Rojas-Duran MF, Gilbert WV. 2012. Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA* **18**: 2299–2305.
- Rosenstiel P, Huse K, Franke A, Hampe J, Reichwald K, Platzer C, Roberts RG, Mathew CG, Platzer M, Schreiber S. 2007. Functional characterization of two novel 5' untranslated exons reveals a complex regulation of NOD2 protein expression. *BMC Genomics* **8**: 472.
- Smith L, Brannan RA, Hanby AM, Shaaban AM, Verghese ET, Peter MB, Pollock S, Satheesha S, Szykiewicz M, Speirs V, et al. 2009. Differential regulation of oestrogen receptor β isoforms by 5' untranslated regions in cancer. *Journal of Cellular and Molecular Medicine* **14**: 2172–2184.
- Sorefan K, Pais H, Hall AE, Kozomara A, Griffiths-Jones S, Moulton V, Dalmay T. 2012. Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence* **3**: 1–1.
- Tang DTP, Plessy C, Salimullah M, Suzuki AM, Calligaris R, Gustincich S, Carninci P.

2013. Suppression of artifacts and barcode bias in high-throughput transcriptome analyses utilizing template switching. *Nucleic Acids Res* **41**: e44–e44.

Appendix II

Investigating the Translational Consequences of Acute eIF4E Depletion

Abstract

While eukaryotic initiation factors (eIFs) have important roles generally in translation, they also have gene-specific functions. eIF2 α phosphorylation reduces protein synthesis globally and increases *GCN4* translation specifically. Overexpression of eIF4E is associated with cellular transformation in mammals, and a reduction of eIF4E activity is accompanied by cell cycle arrest in yeast. However, the molecular basis for differential translational sensitivity of mRNAs to eIF4E remains poorly known. To address this question, we monitor translation in *Saccharomyces cerevisiae* following a reduction in eIF4E activity. We observe a reduction in protein synthesis globally, and assay gene-specific changes by ribosome footprint profiling (Ribo-seq). Intriguingly, we observe eIF2 α -independent upregulation of *GCN4*. We observe a large increase in reads mapping to TLs, though this is likely cycloheximide-dependent. This work provides a foray into genome-wide translational studies of eIF4E and discusses important technical considerations for future endeavours.

Introduction

Translation initiation is a regulated process by which mRNAs recruit eukaryotic initiation factors (eIFs) and ribosomes to translate their Open Reading Frame (ORF). During translation initiation, the methyl-7-guanosine (m7G) cap of the mRNA is bound by the cytoplasmic cap-binding complex eIF4F, composed of eIF4G, 4A, 4B and cap-binding 4E. eIF4F then recruits a small subunit with associated initiation factors (eIF1, 1A, 2, 3, 5, and initiator tRNA-Met) to the mRNA (reviewed in (Jackson et al. 2010)). This initiation complex then scans the transcript leader (TL or 5'UTR) linearly in a net 5' to 3'

direction until an initiation codon (AUG) is found. Once an AUG is found, a large subunit is recruited and an elongating 80S is formed. This process is subject to extensive regulation—several of the aforementioned eIFs are dynamically regulated phosphoproteins (Raught et al. 2000; Dever et al. 1992; Pyronnet et al. 1999). While the model of cap-dependent translation initiation might anticipate global changes in translation in response to regulation of eIF activities, there are also gene-specific changes with important physiological consequences.

A well-studied example of how eIFs can confer both global and gene-specific changes in translation is eIF2 α phosphorylation via the *GCN2* kinase, which reduces cellular translation and also enables selective translational upregulation of *GCN4* (reviewed in (Hinnebusch 2005), also see Chapter 1, Figure 1.2). The 5' UTR of *GCN4* encodes 4 upstream ORFs (uORFs 1-4) and during initiation, ribosomes translate uORF1, terminate, and then resume scanning. However, to initiate at a downstream AUG, the scanning ribosome must acquire Ternary Complex (TC), composed of eIF2 and initiator tRNA-Met. When eIF2 α phosphorylation is low, TC levels are high, and reinitiation frequently occurs at the nearby downstream uORF2, 3, or 4. However, when Gcn2p is activated (e.g. during amino acid starvation), eIF2 α phosphorylation increases, TC levels decrease, ribosomes scan past the uORFs, then bind TC and ultimately initiate at the AUG of *GCN4*. Thus when translation is globally downregulated via eIF2 α phosphorylation, *GCN4* is selectively upregulated.

Regulation of the cap-binding protein eIF4E is also known to have global and gene-specific effects on translation. In mammalian cells, eIF4E is a proto-oncogene whose overexpression is associated with tumorigenesis, and phosphorylation of eIF4E at

Ser 209 by MNK1/2 has been shown to be important for this phenotype (Pyronnet et al. 1999; Wendel et al. 2007; Lazaris-Karatzas et al. 1990). Genes important for tumorigenesis, including c-Myc and VEGF, are known to be translationally sensitive to eIF4E levels and exhibit increased translation in eIF4E-overexpressing cells (Clemens and Bommer 1999). In yeast, the gene encoding eIF4E (*CDC33*) was originally identified in a screen for cell division cycle (*cdc*) mutants (Reid and Hartwell 1977). Subsequently it has been shown that decreased translation of the cyclin *CLN3* is responsible for the *cdc* phenotype, and mutation of the uORF in the *CLN3* TL is sufficient to rescue the cell cycle defect of a *cdc33* strain (Polymenis and Schmidt 1997). All of these findings highlight an important role for eIF4E in influencing the expression of a physiologically relevant subset of mRNAs across eukaryotes, though the molecular mechanism by which differential eIF4E sensitivity occurs is largely unknown.

To understand gene-specific regulation by eIF4E *in vivo*, we monitored translation in a hypomorphic, heat-sensitive *cdc33-42* strain compared to *CDC33* under the same conditions. We observe a ~50% reduction in global translation at the nonpermissive temperature, and profile gene-specific changes using ribosome footprint profiling (Ribo-seq). We observe a 5-fold upregulation of *GCN4*, though we did not detect an increase in eIF2 α phosphorylation, indicating an eIF2 α -independent translational upregulation mechanism. We observe a stark increase in the number of Ribo-seq reads mapping to TLs globally. Analysis of Ribo-seq TL reads revealed an increase in abundance at uORFs beginning with AUG and non-AUG codons. Importantly, increased read abundance at non-AUG uORFs is cycloheximide-dependent

and not physiologically relevant under the conditions assayed. These experiments lay the groundwork for further translational studies of eIF4E in yeast.

Materials and Methods

Yeast strains and Growth Conditions

Yeast cultures (YWG9 (*CDC33::LEU2, his3, ade2, ura3, trp1, pCDC33:TRP1*), YWG10 (*CDC33::LEU2, his3, ade2, ura3, trp1, pcdc33-42:TRP1*), and *pdr5Δ* (W303, Mata, *ade2-1, leu2-3, ura3, trp1-1, his3-11, can1-100, GAL, psi+, ade1::KAN, pdr5::TRP1*)) were grown to mid-log (OD₆₀₀~0.7) phase in YPAD (1% yeast extract, 2% peptone, 0.01% adenine hemisulfate, 2% glucose) in a 25°C water bath in flasks with vigorous shaking. Temperature shift experiments were conducted by media swap via centrifugation with pre-warmed YPAD at the desired temperature. Growth curves for *cdc33-42* and *CDC33* were performed in 96 well format with a Bio-TEK Synergy HT plate reader.

S35 Incorporation

Cells were grown in synthetic complete media with glucose and no methionine (SC-Met) overnight to OD~0.5-0.7. 0.1ul EasyTag L-[³⁵S]-Methionine was added per ml of cells, and samples were removed every 15' for 90'. For each time point, cells were assayed for OD₆₀₀ and 100ul of cells were spotted onto filter paper and incubated at 65C for at least 15' or until all time points were done. Filter papers were pooled and washed for 10' with 5% cold TCA, then 10' with cold 95% ethanol. Filters were dried and radioactivity determined by scintillation counting. Protein synthesis was determined as counts per million (CPM) per OD₆₀₀ per unit time.

Polysome gradients and Ribo-seq

Lysis and polysome gradients were performed as previously described (Arribere et al. 2011) (also see Appendix III) except that triton was omitted from all buffers. Ribosome footprint profiling (Ribo-seq) was performed as previously described (Ingolia et al. 2009). Reads per kilobase per million (RPKM) was calculated excluding the first eight codons of the ORF.

Western Blots

TCA extracts with the same number of OD600 units loaded per lane were run using SDS-PAGE (12%). Western blotting was performed using primary antibody (polyclonal FL-315, Santa Cruz Biotechnology) at 1:1000 and secondary anti Rabbit at 1:200 in TBST with 1% NFM.

uORF analysis

uORFs were identified using the PSSM-based scoring approach described previously (Ingolia et al. 2009). Briefly, a position-specific scoring matrix (PSSM) was derived from the top 100 most highly translated yeast ORFs by Ribo-seq. Using the TL boundaries from (Nagalakshmi et al. 2008), the initiation context of each trinucleotide in the TL was scored. Trinucleotides with an AUG_CAI score >0.6162 were considered for further analyses, as this corresponded to the same score cutoff as (Ingolia et al. 2009). An in-frame stop codon was identified for each potential initiation codon, thus defining an

upstream ORF (uORF). Reads were then iteratively assigned to uORFs, starting with AUG uORFs, then near-AUG uORFs, then non-near-AUG uORFs.

Results

Techniques for Acute Reduction of eIF4E Activity

In order to assay the differential requirement of mRNAs for eIF4E *in vivo*, we sought a way to acutely inhibit eIF4E function. The inhibitor 4EGI-1 is reported to inhibit growth and the eIF4E-4G interaction in mammalian cells at micromolar concentrations (Moerke et al. 2007), so we assayed its effect on yeast growth. Some drugs are ineffective in yeast due to export via the multidrug transporter Pdr5p, and deletion of this gene confers drug hypersensitivity (Leppert et al. 1990). Growth of the *pdr5Δ* strain was insensitive to 4EGI-1 however (Figure II.1), indicating the drug is not efficiently taken up by yeast or that it inefficiently targets the yeast eIF4E-4G interaction. To inhibit eIF4E, we also considered knockdown of eIF4E with a conditional degron or promoter, though depletion of other eIFs by these methods takes several hours, likely due to the long half-life of these proteins in cells (Jivotovskaya et al. 2006). Several hypomorphic alleles of eIF4E exist that exhibit normal growth at the permissive temperature and slow growth and decreased protein synthesis at the restrictive temperature (Altmann and Trachsel 1989). Of the known 4E alleles, *cdc33-42* was chosen as it has normal growth at the permissive temperature of 25C and the greatest defect in protein synthesis and growth at restrictive temperatures (32C or 37C) compared to all other *cdc33* alleles.

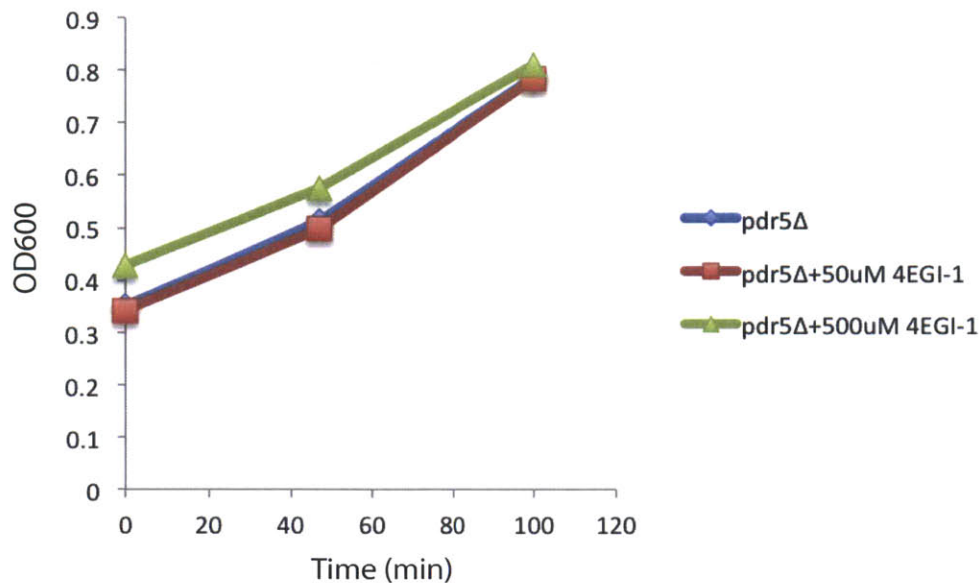


Figure II.1: 4EGI-1 Does Not Inhibit Cell Growth in *S. cer*

Cells were grown to OD600~0.3 and drug added directly to liquid media. Samples were taken at the indicated times and OD600 assayed.

Assaying the Effect of eIF4E Depletion on Translation and Growth

To assess the global effect of acute eIF4E depletion, we monitored growth and protein synthesis before and after temperature shift from 25C to 32C. As reported previously, *cdc33-42* and *CDC33* had similar doubling time at the permissive temperature, 25C (Figure II.2A). However, *cdc33-42* exhibited a significantly longer lag phase. This effect may be a result of underinoculation of *cdc33-42* relative to *CDC33* owing to the larger cell size of *cdc33-42* compared to *CDC33* (see discussion). At higher temperatures (30C, 32C), we observed a decrease in doubling time for *CDC33*, and an increase in doubling time for *cdc33-42* (Figure II.2A). At 32C, the mutant doubling time is over twice that of *CDC33*. We chose to conduct future experiments at 32C instead of 37C because there is a diminished heat shock response (Gasch et al. 2000) and cell cycle defect at 32C, both of

which could lead to secondary effects on translation that confound interpretation of direct eIF4E-dependent events. To assay the effect of *cdc33-42* on protein synthesis, we monitored steady-state S35 incorporation at 25C and 32C (Figure II.2B). At 25C, *cdc33-42* has 25% less S35 incorporation, and at 32C the mutant strain had 48% of the protein synthesis of the control. We also followed protein synthesis by analytical polysome profiling at 25C and after shift to 32C. In *cdc33-42* there was a gradual loss of polysomes after shift to 32C, consistent with a reduction in protein synthesis, while in *CDC33* there was an opposite effect (Figure II.2C). These results confirm the temperature sensitivity of *cdc33-42* and demonstrate its impact on growth and global translation.

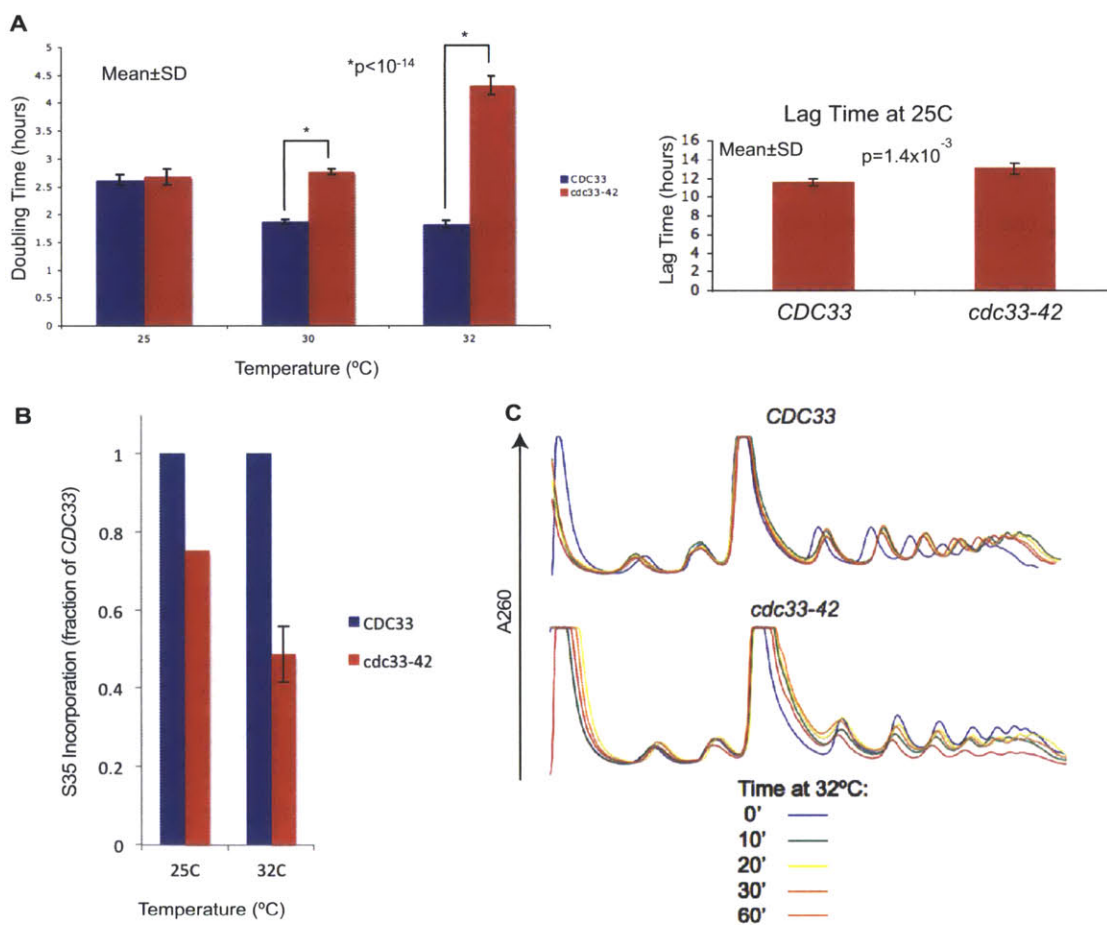


Figure II.2: *cdc33-42* Exhibits Decreased Growth and Protein Synthesis at 32C

(A) Growth characteristics of *CDC33* and *cdc33-42* cells. Saturated cultures were diluted back to OD₆₀₀~0.01 and OD monitored until the culture again reached saturation. Doubling time (left) is during log-phase growth, and lag time (right) is the delay it takes for cells to achieve this growth rate. Experiments performed in quintuplicate (*CDC33*) and sextuplicate (*cdc33-42*). Error bars indicate standard deviation. P-values from Student's T-test.

(B) Steady state protein synthesis for cells grown at 25C or 32C. S35 incorporation was monitored over time and protein synthesis determined as S35 incorporated per OD₆₀₀ per unit time. Values are presented as fraction of *CDC33* S35 incorporation.

Incorporation at 32C was done in biological duplicate and standard deviation shown.

(C) Analytical polysome profiles for cells grown at 25C then shifted to 32C for the indicated time. Traces were shifted along the abscissa so as to facilitate comparison.

Assaying the Effect of eIF4E Depletion on Gene-Specific Translation

To assess the gene-specific translational consequences of eIF4E depletion, we performed ribosome footprint profiling (Ribo-seq (Ingolia et al. 2009)) 30' after a temperature shift from 25C to 32C. Consistent with the known role of eIF4E as a general translation factor for all mRNAs, gene-specific Ribo-seq read density (reads per kilobase per million, RPKM) values were well correlated between the mutant and control strains (Spearman $R > 0.95$) (Figure II.3). While some genes lay above and below the diagonal, in the absence of replicate information the statistical significance of such changes remains unknown. No large off-diagonal groups of genes were apparent, arguing against a significant population of genes with differential translation from the norm in *cdc33-42* under the conditions assayed.

Because eIF4E has a known role in regulating translation of *CLN3*, we looked at the change in translation of this cyclin. Contrary to expectations, we observed an apparent slight increase (1.1 fold) in translation of the *CLN3* ORF in *cdc33-42*, though the significance of this result is unclear. First, in the absence of replicate data, it is difficult to ascertain the statistical significance of this fold change. Secondly, because

Ribo-seq and RNA-seq do not provide absolute measures of gene expression, it is likely that absolute *CLN3* translation is decreased in *cdc33-42*, even though it is slightly increased relative to the average gene. Considering that bulk translation in *cdc33-42* is half that of *CDC33* (Figure II.2B), RPKM values of genes in *cdc33-42* should be corrected by a factor of two. Thus we estimate the amount of *CLN3* translation is reduced approximately 45% in *cdc33-42* relative to *CDC33*.

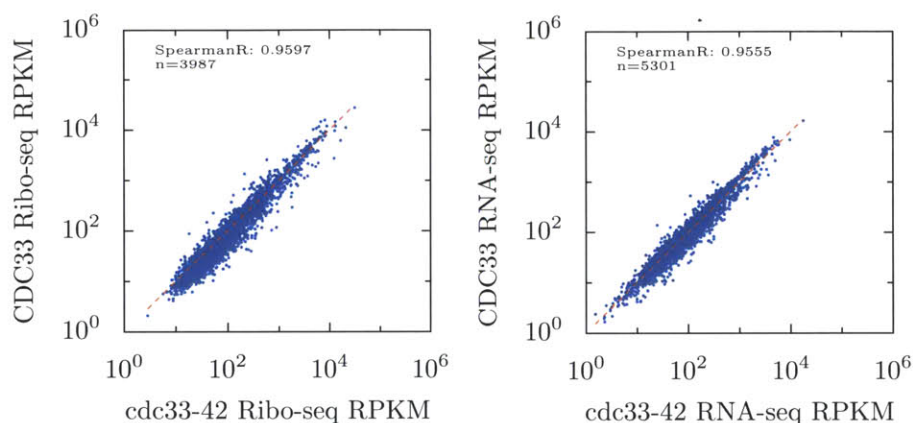


Figure II.3: Gene Expression is Correlated between *cdc33-42* and *CDC33*

Ribo-seq (left) and RNA-seq (right) libraries from *cdc33-42* (abscissa) and *CDC33* (ordinate). Spearman’s correlation and the line $y=x$ are shown.

GCN4* is Translationally Upregulated in *cdc33-42

Given the known role of eIF4E in regulating *CLN3* translation via its uORF, we looked for other interesting uORF-dependent regulation. Surprisingly, we found that *GCN4* translation was increased ~5-fold in *cdc33-42*. We examined the Ribo-seq read distribution in the *GCN4* TL and observed a pattern consistent with the known eIF2 α - and uORF-dependent upregulation mechanism of *GCN4* (Figure II.4A). In *CDC33*, there is a high density of Ribo-seq reads at all uORFs, and few at the *GCN4* ORF. However, in

cdc33-42, Ribo-seq reads are less abundant at uORFs 2-4, yet significantly increased at the *GCN4* ORF. Thus in *cdc33-42*, there is a pattern of uORF translation characteristic of uORF-dependent *GCN4* translational upregulation.

One possibility to explain the observed uORF-dependent upregulation of *GCN4* in *cdc33-42* is through an indirect effect of eIF4E on eIF2 α phosphorylation. To test this, we performed western blots against phosphorylated eIF2 α before and after temperature shift. We consistently detected a slight decrease in the phosphorylation of eIF2 α in *cdc33-42* relative to *CDC33* at both permissive and restrictive temperatures (Figure II.4B), the opposite direction expected for eIF2 α -dependent regulation in *cdc33-42*. Such a result does not support a role for eIF2 α phosphorylation in causing the translational upregulation of *GCN4* in *cdc33-42*. Thus eIF4E is involved in the translational regulation of *GCN4* through an as yet unidentified mechanism.

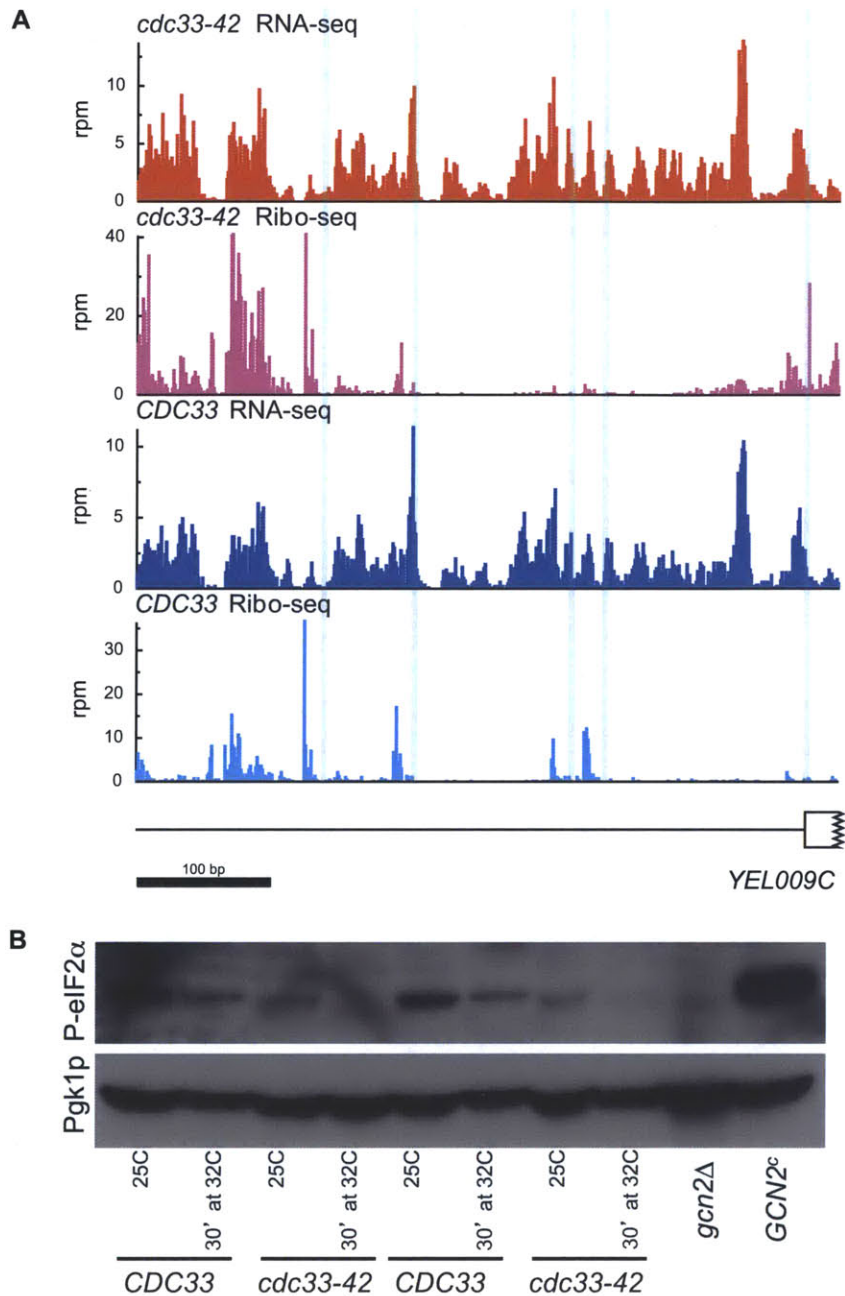


Figure II.4: *GCN4* Exhibits eIF2 α -Independent Translational Upregulation in *cdc33-42*

(A) AUG codons are highlighted (light blue), indicating position of uORFs and *GCN4* start codon. Ordinate shows reads per million (rpm) for each library. The positions of the 5' ends of reads are shown.

(B) eIF2 α phosphorylation is not increased in *cdc33-42*. Cells were grown at 25C and shifted to 32C for 30' and harvested. Control strains (*gcn2* Δ and *GCN2*^c) were grown at 30C. *GCN2*^c is a constitutively active allele of the eIF2 α kinase *GCN2* (Ramirez et al. 1992). Equivalent amounts of cells were loaded in each lane, loading control shown. Technical duplicates are of TCA preparation from the same culture.

eIF4E Depletion Leads to an Increase in TL-mapping Ribo-seq Reads

Our results suggested a role for uORFs in regulating translation in *cdc33-42*, so we analyzed TL-mapping reads globally. Strikingly, we observed a ~75% increase in Ribo-seq reads mapping to TLs in *cdc33-42* (Figure II.5). A large increase in TL-mapping reads was also shown during amino acid starvation in yeast, where it was proposed that ribosomes initiated at both AUG and near-AUG uORFs (codons that have one mismatch to AUG) (Ingolia et al. 2009). To determine whether an increase in near-AUG uORF translation also occurred in *cdc33-42*, we analyzed the positions of Ribo-seq reads with respect to uORFs. Indeed, we observed an increase in reads mapping to AUG-uORFs and near-AUG uORFs, indicating the possibility of translation initiation in TLs at non-cognate codons (Figure II.6A). As a negative control for these analyses, we also looked at non-near-AUG-uORFs (the remaining 43 codons with more than one base difference from AUG) and, surprisingly, observed a similar fold increase. Initiation at near-AUG codons is known to occur for two yeast genes, *grs1* and *ala1* (Chang 2004; Tang 2004), though initiation at non-near-AUG codons is unprecedented on yeast mRNAs.

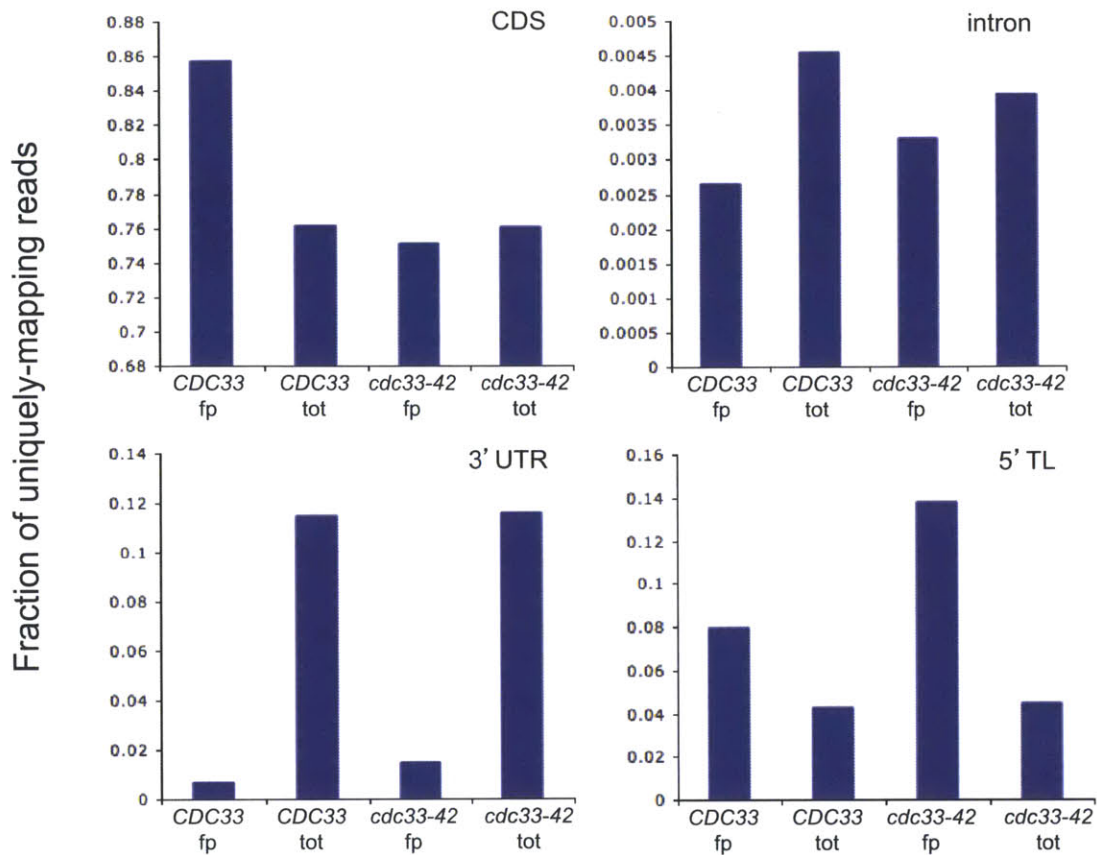


Figure II.5: Global Distribution of Read Locations in *cdc33-42* and *CDC33*

Reads were mapped to *S. cer* genome and the fraction of uniquely mapping reads with 5' ends falling in a given feature determined. TL/3'UTR boundaries taken from (Nagalakshmi et al. 2008), fp=Ribo-seq, tot=RNA-seq. Note ordinate scale changes between graphs.

We next determined whether the increase in Ribo-seq reads at non-near-AUG uORFs was peculiar to the *cdc33-42* mutant. Near-AUG initiation was proposed to be a frequent event during amino acid starvation in yeast (Ingolia et al. 2009), though an increase at non-near-AUG codons was not reported. Analyzing existing amino acid starvation data, we observe a ~6-fold increase in reads mapping to near-AUG-uORFs (Figure II.6B), consistent with what was reported previously. Surprisingly, we observed the same fold increase in reads mapping to non-near-AUG uORFs. Thus Ribo-seq reads

at non-near-AUG-uORFs are not specific to *cdc33-42*, but also increase during amino acid starvation.

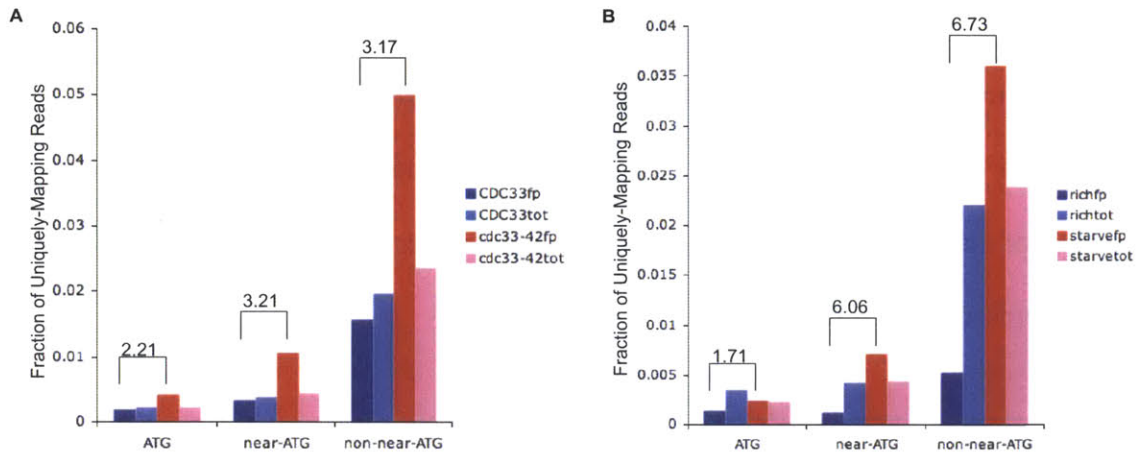


Figure II.6: Global Increase in uORF-Mapping Reads in Multiple Ribo-seq Datasets, Regardless of Initiation Codon

Fraction of uniquely mapping reads mapping to uORFs starting with AUG, near-AUG, and non-near-AUG codons for *cdc33-42* (A) and amino acid starvation (B). In (B), “rich” is rich media and “starve” is amino acid starvation; data for amino acid starvation was taken from (Ingolia et al. 2009). Reads were assigned to uORFs in the order of AUG, near-AUG, and non-near-AUG. The fold increase between Ribo-seq libraries is indicated above brackets.

Faced with the biological implications of such a large relaxation on initiation specificity, we sought alternative explanations for the increase in TL-mapping non-AUG reads. Ribo-seq of glucose-starved yeast demonstrated a ~4-fold increase in reads mapping to TLs and non-AUG uORFs (Vaidyanathan et. al., unpublished). Amino acid starvation, glucose starvation, and *cdc33-42* all have reduced translation initiation efficiency and a large pool of free ribosomes. We considered the possibility that the increase in TL reads was due to continued loading of ribosomes onto the TL after elongating ribosomes were blocked with cycloheximide. Under this hypothesis, we

expect a cycloheximide-dependent increase in TL reads that further increases in cells compromised for translation initiation. To test this hypothesis, we performed Ribo-seq during glucose starvation without cycloheximide and, surprisingly, observed the increase in TL reads at near-AUG uORFs was cycloheximide-dependent (Figure II.7).

Importantly, omission of cycloheximide led to a greater loss of TL-mapping Ribo-seq reads from near-AUG uORFs than from AUG-uORFs. Due to the strong cycloheximide-dependence of many TL-mapping reads, it is unclear whether reads generated from such positions represent *in vivo* translation.

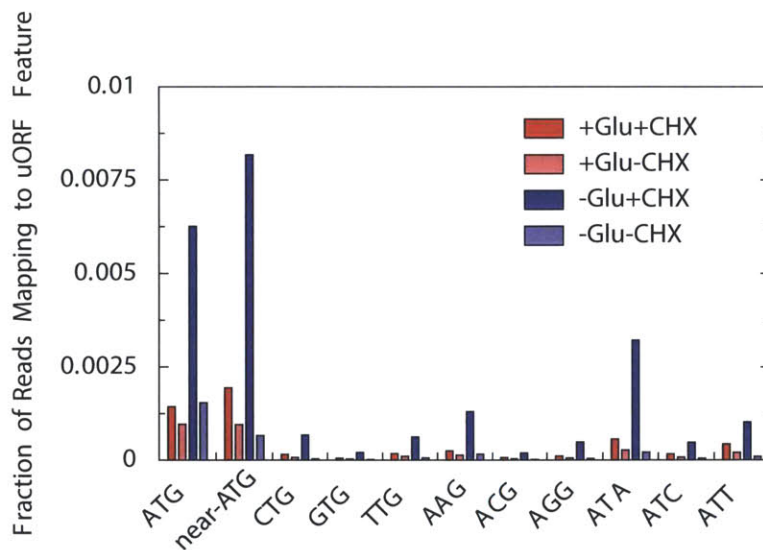


Figure II.7: Cycloheximide-Dependent Increase in non-AUG uORF-Mapping Reads

The fraction of uniquely mapping reads mapping to uORFs starting with uAUG or near-uAUG, +/- glucose and +/- cycloheximide. The cycloheximide-dependent increase in Ribo-seq reads mapping to near-AUG uORFs is also shown broken down by near-AUG codon.

Discussion

eIF4E is a critical component of the translational machinery and is important for normal cell division in yeast. Here we show that eIF4E inhibition has both global and gene-

specific effects, including a previously unreported eIF2 α -independent enhancement of *GCN4* translation. We observe a global reduction in translation across a majority of cellular messages, consistent with the known role of eIF4E as a general translation factor. A two-fold increase in TL-mapping reads is also apparent, though Ribo-seq experiments conducted with and without cycloheximide under glucose starvation argue that this effect is cycloheximide-dependent.

Translational upregulation of *GCN4* in *cdc33-42* was apparent though the mechanism remains unknown. We show here that while the increase in *GCN4* translation appears uORF-dependent, it does not occur through eIF2 α phosphorylation, indicating an atypical mechanism of upregulation. Another known mechanism of *GCN4* upregulation is through decreased abundance of free initiator tRNA-Met (Dever et al. 1995); preliminary experiments were inconclusive whether this might be the case in *cdc33-42*. Another possibility is that eIF4E has a role in reinitiation, a process that is poorly understood at a molecular level. Translational regulation of *GCN4* is a well-studied system with numerous genetic tools useful for dissecting the contributions of its TL and uORFs to regulation, and those tools may prove useful for understanding eIF4E-dependent upregulation as well. Additionally, translation in *cdc33-42* extracts is strongly eIF4E-dependent (Altmann and Trachsel 1989), providing a system to analyze a direct role for eIF4E in translational regulation of *GCN4* in the absence of cellular feedback.

The most notable global feature in *cdc33-42* Ribo-seq libraries was an almost two-fold increase in TL reads, though we do not think translation occurs to this extent in TLs *in vivo*. These reads occur without a discernable sequence preference and in numerous other libraries, all of which have a known decrease in initiation. Furthermore,

in glucose starvation the TL-mapping reads are predominantly cycloheximide-dependent. Upstream ORFs are known to decrease translation of the downstream ORF and trigger nonsense-mediated mRNA decay (NMD) of their transcript (Gaba et al. 2005). Consistent with translation at uAUGs, there is both a global decrease in translation and an increase in steady-state mRNA levels for uAUG-containing mRNAs in yeast (Chapter 2, Figure 2.11). However, uORFs initiated with any other trinucleotide failed to exhibit these same changes. The same is true of mouse mRNAs as well (data not shown). To unify all of these observations with a single explanation, we consider it likely that the increase at non-AUG codons in Ribo-seq does not occur to the same extent under physiological conditions in cells, but is an effect of inhibiting translation elongation with cycloheximide. If correct, we would expect the same increase in TL reads in any cells compromised in initiation and incubated with a translation elongation inhibitor, cycloheximide or otherwise. Thus far every dataset, published and unpublished, has behaved as expected under this hypothesis.

Future Directions

The *cdc33-42/CDC33* strains are in an unknown background, possibly confounding comparisons with other yeast datasets, which are primarily in S288C. The stress-responsive transcription factor *GCN4* is among the lowest expressed genes in S288C, however by RNA-seq in *CDC33* it is among the highest two hundred expressed mRNAs in the cell and even higher in *cdc33-42*. Given the induction of *GCN4* in response to numerous cellular stress pathways, it is likely that the *CDC33* background is in a very different cellular state than S288C. Furthermore, the plasmid-based expression of eIF4E

used here leads to a population of cells with heterogeneous eIF4E expression. To contextualize future eIF4E studies with other *S. cer* datasets, eIF4E alleles should be integrated into the S288C background.

Because eIF4E depletion causes arrest in G1, comparisons between eIF4E-depleted and control samples will undoubtedly be confounded by secondary effects due to different translational activities during the cell cycle. Translational activities of individual messages are known to change during both mitosis (Polymenis and Schmidt 1997) and meiosis (Brar et al. 2012), and hundreds of genes' mRNA levels fluctuate over mitosis (Spellman et al. 1998). When considering other temperatures and times to inhibit eIF4E activity in the *cdc33-42* strain in the future, experimenters would be wise to assay the quantitative effect on cell cycle by flow cytometry and protein synthesis by S35 incorporation. If there is an observable defect in cell division, elutriation may be useful for correcting for such differences and thus simplifying downstream analyses.

OD600 is commonly used as a proxy for cell number for yeast cultures, though it may be inaccurate for *cdc33-42*. OD600 is a measure of absorbance and is affected by cell size and flocculation, and such properties are not always constant between experimental and control samples. For instance, *cdc33-42* has a known cell-cycle defect leading to arrest of cells in G1. G1 cells are larger and thus absorb more light per cell, causing *cdc33-42* cultures with the same number of cells to absorb more light and appear to be at a higher OD than *CDC33* cultures. Thus at the same OD there are more *CDC33* cells than *cdc33-42*, and this likely causes underinoculation leading to an apparent delay in entering log-phase growth (Figure II.2A). Because cell density affects the availability of numerous extracellular molecules such as nutrients and secreted proteins, future

experiments should determine an appropriate way of normalizing for an effect of cell size so as to ensure a similar number of cells in a similar nutritional state are analyzed. Again, this will minimize the study of indirect phenotypic consequences of eIF4E inhibition.

Comparisons between Ribo-seq libraries, like all sequencing-based assays, are relative, and future eIF4E studies would benefit from knowledge of absolute changes. Global changes in the abundance or translation of all mRNAs will be invisible and this is important for the interpretation of datasets such as *cdc33-42*. Here we show that under the conditions assayed, *cdc33-42* has half the protein synthetic capacity of *CDC33* (Figure II.2B). Naively we might expect half as many reads in all ORFs, though absolute information is lost and we see a correlation of RPKM values centered around $y=x$ (Figure II.3). However, if a small amount of non-yeast ribosomes (e.g. mammalian) were added to the same amount of lysate from experiment and control samples, it may be possible to know absolute changes as well. Absolute translational measurements are essential for accurate interpretation of gene-specific and global changes in mRNA translation.

Acknowledgements

Libraries for glucose starvation experiments were prepared by Boris Zinshteyn. Thanks to Dave Bartel for helpful discussions regarding the effects of cycloheximide on Ribo-seq.

References

- Altmann M, Trachsel H. 1989. Altered mRNA cap recognition activity of initiation factor 4E in the yeast cell cycle division mutant *cdc33*. *Nucleic Acids Res* **17**: 5923–5931.
- Arribere JA, Doudna JA, Gilbert WV. 2011. Reconsidering Movement of Eukaryotic

- mRNAs between Polysomes and P Bodies. *Molecular Cell* **44**: 745–758.
- Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS. 2012. High-Resolution View of the Yeast Meiotic Program Revealed by Ribosome Profiling. *Science* **335**: 552–557.
- Chang KJ. 2004. Translation Initiation from a Naturally Occurring Non-AUG Codon in *Saccharomyces cerevisiae*. *Journal of Biological Chemistry* **279**: 13778–13785.
- Clemens MJ, Bommer UA. 1999. Translational control: the cancer connection. *Int J Biochem Cell Biol* **31**: 1–23.
- Dever TE, Feng L, Wek RC, Cigan AM, Donahue TF, Hinnebusch AG. 1992. Phosphorylation of initiation factor 2 alpha by protein kinase GCN2 mediates gene-specific translational control of GCN4 in yeast. *Cell* **68**: 585–596.
- Dever TE, Yang W, Aström S, Byström AS, Hinnebusch AG. 1995. Modulation of tRNA(iMet), eIF-2, and eIF-2B expression shows that GCN4 translation is inversely coupled to the level of eIF-2.GTP.Met-tRNA(iMet) ternary complexes. *Mol Cell Biol* **15**: 6351–6363.
- Gaba A, Jacobson A, Sachs MS. 2005. Ribosome occupancy of the yeast CPA1 upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. *Molecular Cell* **20**: 449–460.
- Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. 2000. Genomic expression programs in the response of yeast cells to

- environmental changes. *Mol Biol Cell* **11**: 4241–4257.
- Hinnebusch AG. 2005. Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol* **59**: 407–450.
- Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**: 218–223.
- Jackson RJ, Hellen CUT, Pestova TV. 2010. The mechanism of eukaryotic translation initiation and principles of its regulation. 1–15.
- Jivotovskaya AV, Valásek L, Hinnebusch AG, Nielsen KH. 2006. Eukaryotic translation initiation factor 3 (eIF3) and eIF2 can promote mRNA binding to 40S subunits independently of eIF4G in yeast. *Mol Cell Biol* **26**: 1355–1372.
- Lazaris-Karatzas A, Montine KS, Sonenberg N. 1990. Malignant transformation by a eukaryotic initiation factor subunit that binds to mRNA 5' cap. *Nature* **345**: 544–547.
- Leppert G, McDevitt R, Falco SC, Van Dyk TK, Ficke MB, Golin J. 1990. Cloning by gene amplification of two loci conferring multiple drug resistance in *Saccharomyces*. *Genetics* **125**: 13–20.
- Moerke NJ, Aktas H, Chen H, Cantel S, Reibarkh MY, Fahmy A, Gross JD, Degterev A, Yuan J, Chorev M, et al. 2007. Small-Molecule Inhibition of the Interaction between the Translation Initiation Factors eIF4E and eIF4G. *Cell* **128**: 257–267.
- Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The

Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing.

Science **320**: 1344–1349.

Polymenis M, Schmidt EV. 1997. Coupling of cell division to cell growth by translational control of the G1 cyclin CLN3 in yeast. *Genes Dev* **11**: 2522–2531.

Pyronnet S, Imataka H, Gingras AC, Fukunaga R, Hunter T, Sonenberg N. 1999. Human eukaryotic translation initiation factor 4G (eIF4G) recruits mnk1 to phosphorylate eIF4E. *EMBO J* **18**: 270–279.

Ramirez M, Wek RC, Vazquez de Aldana CR, Jackson BM, Freeman B, Hinnebusch AG. 1992. Mutations activating the yeast eIF-2 alpha kinase GCN2: isolation of alleles altering the domain related to histidyl-tRNA synthetases. *Mol Cell Biol* **12**: 5801–5815.

Raught B, Gingras AC, Gygi SP, Imataka H, Morino S, Gradi A, Aebersold R, Sonenberg N. 2000. Serum-stimulated, rapamycin-sensitive phosphorylation sites in the eukaryotic translation initiation factor 4GI. *EMBO J* **19**: 434–444.

Reid BJ, Hartwell LH. 1977. Regulation of mating in the cell cycle of *Saccharomyces cerevisiae*. *The Journal of Cell Biology* **75**: 355–365.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* **9**: 3273–3297.

- Tang HL. 2004. Translation of a Yeast Mitochondrial tRNA Synthetase Initiated at Redundant non-AUG Codons. *Journal of Biological Chemistry* **279**: 49656–49663.
- Wendel HG, Silva RLA, Malina A, Mills JR, Zhu H, Ueda T, Watanabe-Fukunaga R, Fukunaga R, Teruya-Feldstein J, Pelletier J, et al. 2007. Dissecting eIF4E action in tumorigenesis. *Genes Dev* **21**: 3232–3237.

Appendix III

Reconsidering Movement of Eukaryotic mRNAs Between Polysomes and P-bodies*

*This research was originally published in Molecular Cell and has been edited for presentation here. Arribere JA, Doudna JA, Gilbert WV. Reconsidering movement of eukaryotic mRNAs between polysomes and P bodies. Mol Cell. 2011 Dec 9;44(5):745-58. doi: 10.1016/j.molcel.2011.09.019.

Abstract

Cell survival in changing environments requires appropriate regulation of gene expression, including post-transcriptional regulatory mechanisms. From reporter gene studies in glucose-starved yeast, it was proposed that translationally silenced eukaryotic mRNAs accumulate in P-bodies and can return to active translation. We present evidence contradicting the notion that reversible storage of non-translating mRNAs is a widespread and general phenomenon. First, genome-wide measurements of mRNA abundance, translation, and ribosome occupancy following glucose withdrawal show that most mRNAs are depleted from the cell coincident with their depletion from polysomes. Second, only a limited sub-population of translationally repressed transcripts, comprising fewer than 400 genes, can be reactivated for translation upon glucose re-addition in the absence of new transcription. This highly selective post-transcriptional regulation could be a mechanism for cells to minimize the energetic costs of reversing gene-regulatory decisions in rapidly changing environments by transiently preserving a pool of transcripts whose translation is rate-limiting for growth.

Introduction

Cells respond to changing environments by regulating gene expression. Regulation can occur at the level of transcription and/or post-transcriptionally during processes including pre-mRNA splicing, mRNA export, translation and mRNA decay. In some embryonic cells, gene regulation during early development is entirely post-transcriptional and involves temporally and spatially controlled translation of maternally deposited mRNAs (Johnstone and Lasko 2001; Richter 1991). More typically, cells employ a combination

of transcriptional and post-transcriptional regulatory strategies. The logical and mechanistic relationships between transcriptional and post-transcriptional regulation are poorly understood, if indeed such relationships exist.

Various hypotheses have been proposed for the role of translational regulation in contexts where transcriptional regulatory mechanisms are also active. For example, translational activation of pre-existing mRNAs can produce new protein faster than transcriptional activation of the same genes, and may therefore be important in situations that demand rapid responses. In addition, translational mechanisms can control where proteins are produced within the cell. Furthermore, translational regulation has been suggested to act globally as an amplifier of the effects of transcriptional gene control, increasing the protein output from transcriptionally induced genes and further decreasing the protein output from transcriptionally repressed genes (Melamed et al. 2008; Preiss et al. 2003). On the other hand, translational attenuation has also been proposed to act as a global dampener of transcriptional noise in gene expression (Blake et al. 2003; Ozbudak et al. 2002; Raser 2005).

We set out to determine the relationship between the programs of transcriptional and translational response to stress. We further sought to determine the biological logic behind selection of specific mRNAs for translational regulation, and the molecular differences between genes controlled at transcriptional versus translational levels. The glucose starvation response in yeast is an appropriate model system because glucose withdrawal induces widespread changes in both transcription and translation. Transcriptional changes are mediated by well-characterized signaling pathways and transcription factors (Zaman et al. 2008). Translation activity changes by an incompletely

understood mechanism requiring genes that have been variously implicated in deadenylation-dependent mRNA decapping and decay, mRNA sub-cellular localization, and the formation of translationally repressed mRNPs (Ashe et al. 2000; Brengues 2005; Coller and Parker 2004; 2005; Holmes et al. 2004; Teixeira et al. 2005). In response to glucose starvation, yeast initiate a cellular differentiation program known as haploid invasive growth, which is thought to function as a cellular foraging response (Cullen and Sprague 2000). Because this cellular adaptive response to glucose starvation requires new protein synthesis, the ‘global’ repression of translation must either be short-lived or in fact affect only a subset of genes.

Here we used DNA microarrays to investigate changes in mRNA abundance, translation activity and ribosome occupancy during a two-hour time course following glucose withdrawal. We found that the view of ‘global’ translational repression is over-generalized. While ‘bulk’ translation was greatly reduced, hundreds of newly transcribed mRNAs associated with polysomes within ten minutes of glucose withdrawal. Functionally coherent groups of genes were co-regulated at the post-transcriptional as well as transcriptional level. Using computational approaches, we related gene-specific post-transcriptional changes to underlying mRNA properties by exploiting recent genome-wide studies of yeast mRNA characteristics including abundance, half-life, translational efficiency, poly(A) tail length and association with various RNA-binding proteins. Following a lead generated by this analysis, we examined whether all genes or only a specific sub-population of genes are capable of returning to active translation in the absence of new transcription. In contradiction of the prevailing model in the field, we found that the capacity for translational reactivation is narrowly restricted to a limited

subset of mRNAs. Transient preservation of these select mRNAs, whose translation is rate-limiting for growth in rich media, could act as a buffer to minimize the fitness costs associated with "false alarms" caused by transient depletion of glucose or by noise in glucose-sensitive signaling pathways.

Materials and Methods

Yeast Strains and Culture

All experiments used Sigma 1278b background yeast (*MATa ura3 leu2 trp1 his3*), strain F1950 (a gift from Hiten Madhani). Yeast cultures were grown in YPAD (1% Yeast extract, 2% Peptone, 0.01% Adenine hemisulfate, 2% Dextrose) at 30°C in baffled flasks with vigorous shaking. Glucose-starved cultures were prepared from YPAD cultures grown to $OD_{600} = 1.0 - 1.1$, harvested by centrifugation (5 min at 12,000 x g), resuspended in pre-warmed YPA medium lacking glucose and returned to shaking at 30°C for various times. Transcription by RNA polymerase II was inhibited where indicated by the addition of thiolutin (Sigma) to the culture medium to a final concentration of 3 $\mu\text{g}/\text{mL}$.

Extract Preparation, Polysome Gradient Fractionation and RNA isolation

Yeast polysomes were prepared as described (Clarkson et al. 2010). Briefly, cycloheximide was added to cell cultures to a final concentration of 0.1 mg/mL for 2 minutes before harvesting cells by centrifugation (5 min at 12,000 x g). Cells were lysed by vortexing with glass beads in cold 1X PLB (20 mM HEPES-KOH, pH 7.4, 2 mM Magnesium Acetate, 100 mM Potassium Acetate, 0.1 mg/mL cycloheximide, 3 mM

DTT). The crude extracts were clarified (20 minutes at 15,500 x g) and the resulting supernatants were applied to linear 10-50% sucrose gradients in 1X PLB. Lysates were typically 100-200 OD₂₆₀ U/mL. For large-scale gradients, 50 OD₂₆₀ U were applied to 40 mL gradients and spun for 4 hours at 27,000 rpm in a Beckman SW28 rotor. For small-scale gradients, 20 OD₂₆₀ U were applied to 11 mL gradients and spun for 3 hours at 35,000 rpm in a Beckman SW41 rotor.

RNA was purified from polysomal gradient samples by denaturation with guanidine HCl followed by successive isopropanol and ethanol precipitations. Total RNA was isolated from clarified lysates by hot phenol extraction followed by isopropanol precipitation (Clarkson et al. 2010).

cDNA Synthesis and Labeling, Microarray fabrication and Hybridization

Production of Cy3 and Cy5 labeled cDNA from total and polysomal RNA samples, microarray fabrication, hybridization and washing were performed as described (Clarkson et al. 2010). cDNA synthesis was performed with a 1:1 mixture of oligo dT and random nonamer primers to permit detection of mRNAs with short poly(A) tails. Labeled cDNA was hybridized to custom microarrays containing 70mer oligonucleotide probes to all SGD-annotated ORFs. Details about the Operon YBOX and AROS probe sets are available at the website: <http://omad.operon.com/download/index.php>.

Image Analysis and Microarray Data Processing

A single ratio value (median pixel intensities for the 635 nm and 532 nm images of each spot) was determined for all microarray features by averaging within-array spot replicates

from all biological and technical replicate array experiments. Spot ratio data were normalized within each array using global loess regression as described (Pleiss et al. 2007). Arrays with high background or spatial bias were discarded and the experiment repeated. A maximum of 8 ratio measurements were obtained for each feature: 2 spots x 2 biological replicate experiments x 2 [dye-flipped] technical replicate microarrays [duplicate cDNA synthesis, labeling and microarray hybridization]. Prior to feature identification, poor quality spots were excluded from further analysis if they showed visible defects due to dust or printing irregularities.

Clustering Analysis

Hierarchical clustering was performed on timecourse data of averaged feature ratio values using centroid linkage and Euclidean distance as the distance measure. A standard *k*-means clustering algorithm was used to identify groups of genes displaying similar changes in total mRNA relative abundance (T/T) or ribosome occupancy (P/T). Briefly, *k*-means clustering proceeds by first randomly assigning genes to one of an arbitrarily chosen number of groups. The mean vector for all genes in each group is computed, and the genes are then reassigned to the group whose center is closest to the gene. In our analysis, each vector was comprised of five T/T or six P/T values. We used Euclidian distance as the distance metric for evaluating closeness. Clustering proceeds by repeating these two steps until the optimal solution is found. The optimal solution is the one with the lowest possible within-group sum of distances. All clustering analysis was performed using Cluster 3.0.

Plasmids

pRP1186 (Dcp2-RFP, URA3) and pPS2037 (PGK1-U1A, URA3) were previously described (Teixeira et al. 2005; Brodsky and Silver 2000). pWG510 (U1A-GFP, HIS3) was made by subcloning the TEF1 promoter in place of the Gal promoter in pPS2045 (Brodsky and Silver 2000). Additional mRNA reporters pWG511 (RRP4-U1A), pWG512 (LSG1-U1A), pWG513 (RPS26A-U1A) and pWG514 (MFA2-U1A) consist of each gene's promoter and coding region followed by the 3' region from pPS2037 (U1A, PGK1 3'UTR, ADH2 terminator). U1A reporters were constructed by PCR from genomic DNA with *XhoI* and *BamHI* linkers inserted into pPS2037. pWG515 (MFA2-pG-U1A) was constructed via insertion of a *BglII* site after the stop codon followed by cloning of 18 consecutive guanosine residues flanked by *BglII* sites. pWG515 contains two copies of the U1A hairpin while all others contain 16; this difference is incidental and arises during introduction of the *BglII* site into pWG514. All plasmids were confirmed by sequencing.

Quantitative RNA Analysis

Yeast polysome gradients and RNA isolation were performed as described in Experimental Procedures. Prior to RNA extraction, in vitro transcribed Firefly luciferase mRNA was added to each 1 mL gradient fraction to permit normalization of qRT-PCR values between fractions. Reverse transcription and quantitative PCR was performed using SuperScript III and real-time PCR reagents (Invitrogen) according to manufacturer's instructions using a Roche LightCycler 480. Gene-specific primer sequences are available upon request.

Microscopy

Cells were grown in SC media supplemented with amino acids to OD 0.6-0.8. Aliquots of cells were spun down and resuspended in SC plus glucose or SC minus glucose for 10 minutes. Cells were visualized with a Nikon Eclipse Ti microscope and subsequent to capture, brightness and contrast of whole images were adjusted in Adobe Photoshop.

Accession Numbers

The microarray data have been deposited at the Gene Expression Omnibus (GEO) repository under the accession number GSE31393.

Results

We studied wild type invasive-growth competent yeast subjected to acute glucose starvation. Cells were starved for 0, 10, 20, 30, 60, or 120 minutes before processing. For each time point, polysome profiles were generated to monitor global translation. In agreement with previous reports (Ashe et al. 2000; Kuhn et al. 2001), a collapse of the polysome region and concomitant increase in the 80S monosome peak occurred after 10 minutes of glucose withdrawal, indicating a bulk reduction in translation initiation.

Previous investigations of the translational response to acute glucose withdrawal using either bulk measurements (Ashe et al. 2000; Holmes et al. 2004) or examination of a few genes (Bregues 2005) suggested that little translation initiation occurs 10-20 minutes following glucose removal. In light of prior studies of glucose-stimulated transcriptional regulation, as well as genetic evidence that many of the genes induced upon glucose

depletion are required for growth in the absence of glucose, we reasoned that such transcripts must at some time associate with polysomes in glucose-starved cells (Zaman et al. 2008). We extended the time course of the experiment to learn when bulk translation recovers as cells adapt to growth in low glucose concentrations. Polysomes remained low between 10 and 30 minutes post glucose withdrawal, and showed signs of partial recovery after 60 minutes, with further increases by 120 minutes (Figure III.1A).

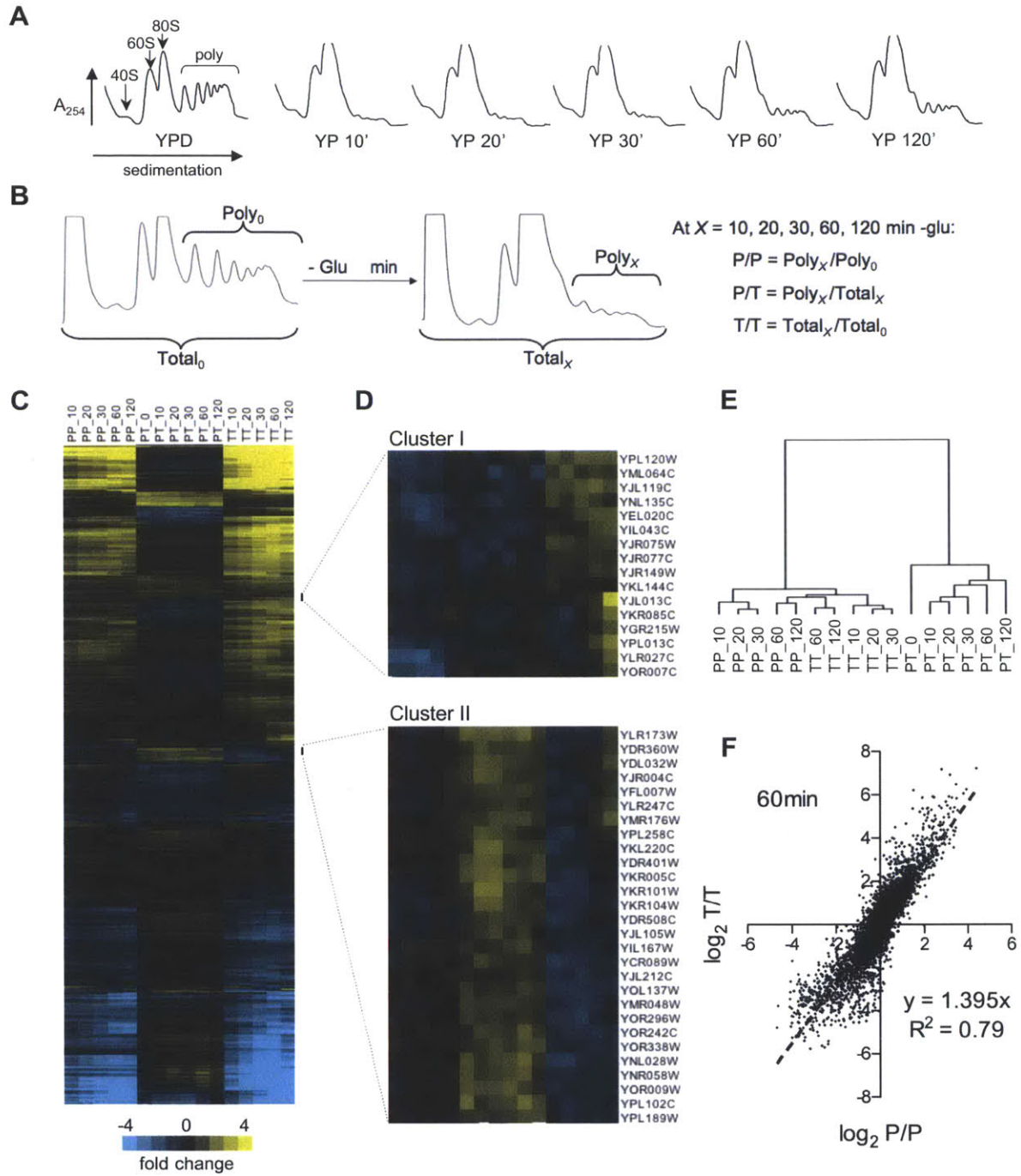


Figure III.1: Regulation of Transcription and Translation in Glucose-Starved Cells

(A) Polysome profiles of yeast starved for glucose by transfer from YPD to YP media. (B) Schematic of the microarray comparisons performed. (C) Time-resolved gene expression profiles resulting from the comparisons shown in B. Ratio values are indicated by color scale and represent the average of eight measurements (2 biological replicates X 2 technical replicates X 2 probes per gene). Rows (genes) are ordered according to the results of hierarchical clustering based on Euclidean distance with P/T values assigned

twice the weight of P/P or T/T. (D) Unusual gene clusters for which polysomal mRNA abundance diverged from total mRNA abundance. (E) P/T ratios are more similar to each other than to T/T or P/T for any time point. Branch lengths represent Spearman rank correlation coefficients between data columns from C. (F) Total mRNA abundance changes parallel and exceed polysomal mRNA abundance changes after 60 minutes minus glucose. Dotted line indicates best fit by linear regression. See also Figures III.2 and III.3.

Changes in Polysomal mRNA Levels Closely Parallel Changes in Transcript

Abundance

To examine the kinetics of gene-specific translation activity following glucose withdrawal, we isolated RNA from polysome fractions and total cell lysates, and prepared fluorescently labeled cDNA for competitive hybridization on custom microarrays. In order to determine relative changes in translation activity, mRNA abundance, and the fraction of total mRNA associated with polysomes (ribosome occupancy), we performed the following comparisons for each time point after glucose withdrawal: polysomal RNA starved with polysomal RNA mock-starved (P_x/P_0); total RNA starved with total RNA mock-starved (T_x/T_0); and polysome starved x -min with total starved x -min (P_xT_x), respectively (Figure III.2B). All experiments were performed twice (biological replicates), with each RNA sample processed in duplicate (technical replicates). ‘Noisy’ genes for which mRNA abundance or polysome association varied by > 1.5-fold between biological replicate experiments in unstarved cells were omitted from further analysis (495 genes). Reproducible results were obtained for 5,590 genes.

Glucose withdrawal led to changes in the relative mRNA abundance of hundreds of genes within 10 minutes, consistent with prior studies of carbon source-mediated regulation of yeast transcription (Zaman et al. 2008). These changes persisted and were amplified over the course of 120 minutes of starvation. Our data do not distinguish

between transcriptional induction/repression and mRNA stabilization/destabilization as the mechanism responsible for changing total mRNA abundance. Both mechanisms likely contribute. Many of the genes that showed relatively increased mRNA levels following glucose withdrawal are known targets of glucose regulated transcription factors (Figure III.2). For most genes, changes in their relative mRNA abundance in polysomes (P/P) mirrored changes in overall mRNA levels (T/T) (Figure III.1C). Notably, for the more than 1,000 genes whose relative mRNA abundance increased by 2-fold or more after glucose withdrawal, relative polysomal abundance similarly increased. Furthermore, most induced genes did not show a noticeable lag between the increase of mRNA in the total RNA pool and appearance in the polysomal fraction. Thus, despite the reduction in the rate of 'bulk' protein synthesis, translation initiation occurred on transcriptionally up-regulated mRNAs as early as 10 minutes following glucose withdrawal. Reduction in polysomal mRNA closely paralleled reduction in total mRNA levels for almost all down-regulated genes. These data contradict the model that glucose withdrawal leads to widespread sequestration of mRNAs in stable translationally repressed mRNPs (Bregues and Parker 2007; Bregues 2005; Hoyle et al. 2007; Teixeira et al. 2005).

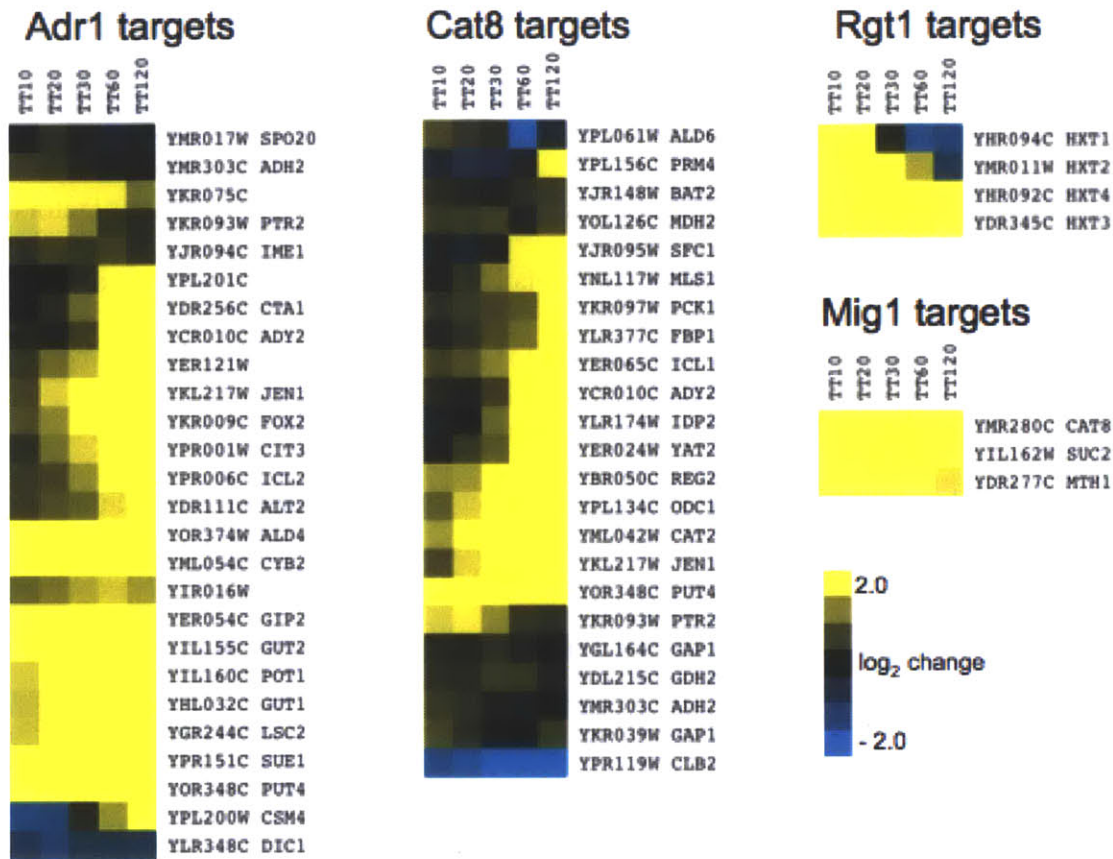


Figure III.2: Induction Kinetics of Adr1, Cat8, Rgt1 and Mig1 Target Genes

Time-resolved gene expression profiles of known targets of glucose-regulated transcription factors. Transcriptional targets were identified by Young *et al.* 2003. Our data are consistent with the rapid derepression of Mig1 targets, as well as a later induction of Cat8 targets due to the delayed accumulation of Cat8 protein.

Unsupervised hierarchical clustering revealed small groups of genes that appear to be regulated primarily at the post-transcriptional level: cluster I genes showed relatively reduced polysomal mRNA levels despite increased total mRNA abundance due to low ribosome occupancy; conversely, cluster II genes had relatively unaffected polysomal mRNA levels despite reduced total mRNA abundance due to high ribosome occupancy (Figure III.1D). Nevertheless, there was strong overall agreement between the relative increase or decrease of an mRNA in the cell at any time point after glucose withdrawal

(T/T) and the change in polysome association at that time point (P/P). In contrast, ribosome occupancy (P/T) at any time point was more highly correlated with ribosome occupancy at other times than with either P/P or T/T at the same time point after glucose withdrawal (Figure III.1E), suggesting that it is largely an intrinsic property of each mRNA, although some starvation-induced changes in P/T were observed. At the global scale, fold-changes in polysomal mRNA levels were somewhat compressed compared to changes in total mRNA levels (Figure III.1F, III.3). Thus, our data do not indicate widespread 'potentiation', whereby changes in total mRNA levels are amplified by homodirectional changes in translation efficiency, as was suggested in studies of yeast subject to rapamycin treatment or heat shock (Preiss et al. 2003). Stress specificity of 'potentiation' was previously noticed in comparison of the translational responses to amino acid starvation and butanol stress (Smirnova et al. 2005).

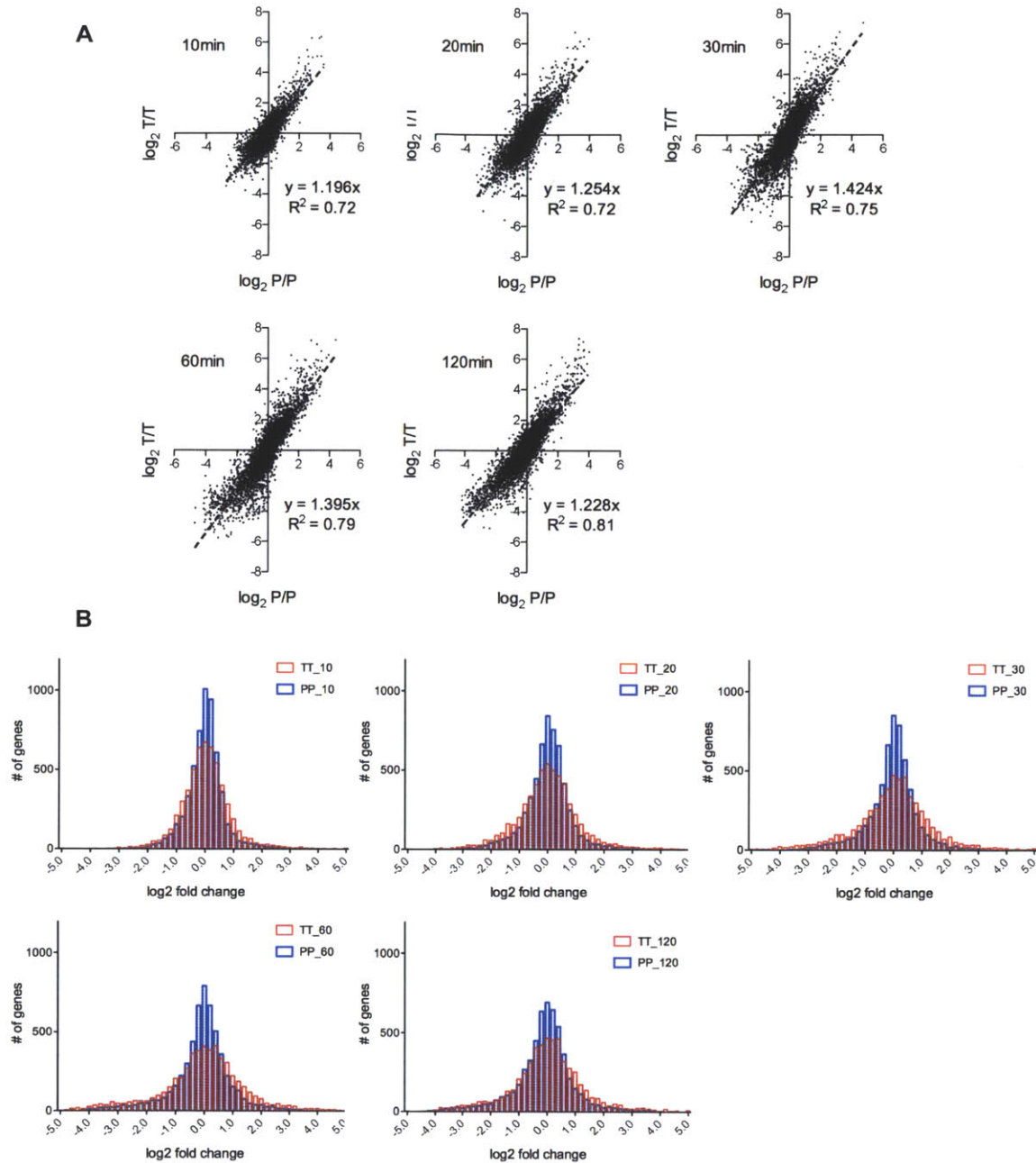


Figure III.3: Changes in Total mRNA Levels Exceed Changes in Polysomal mRNA Levels

(A) Total mRNA abundance changes parallel and exceed polysomal mRNA abundance changes after glucose withdrawal. Each data point represents a single gene. Dotted line indicates best fit by linear regression. Slope and goodness of fit (R^2) are displayed.

(B) Histogram of all genes showing a given \log_2 fold change in Total (TT) or polysomal (PP) mRNA levels at each timepoint following glucose withdrawal.

Relationships Between Changes in Transcript Levels and Ribosome Occupancy

The complex groupings of genes produced by combining the analysis of transcriptional and post-transcriptional regulatory behavior using unsupervised hierarchical clustering did not readily reveal the logic underlying the relationship between the two modes of regulation. To investigate this relationship more directly, changes in total mRNA levels (T/T) and changes in ribosome occupancy (P/T) were analyzed separately using k -means clustering to identify groups of genes displaying similar behavior for each mode of regulation. Experimenting with various group numbers ($k = 2-20$) revealed that $k = 7$ for the T/T comparisons and $k = 3$ for the P/T comparisons gave robust solutions reflecting a reasonable compromise between preserving the complexity of the data and simplifying the subsequent analysis (Figure III.4). Clustering genes by their ribosome occupancy (P/T) produced simple divisions into groups with high, low, and 'neutral' P/T ratios (Figure III.5A).

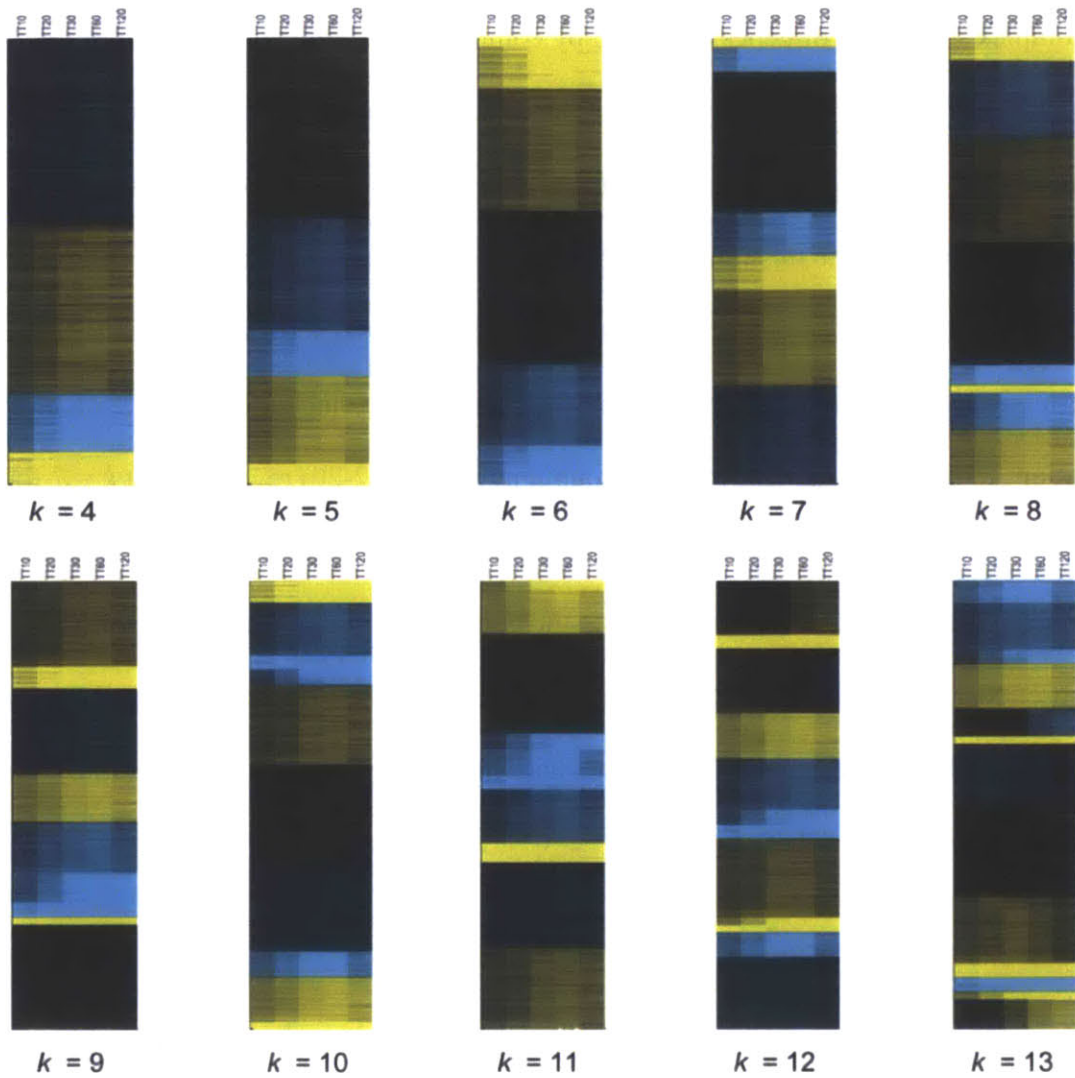


Figure III.4: Gene Groupings Produced by k-means Clustering at Different Values of k

Genes were clustered by T/T ratios. Data and color scale are the same as Figure III.5. Results shown are the best solution after 1,000 trials at a given value of k . Values of $k > 10$ revealed some segregation of genes according to the kinetics of gene expression changes, but these solutions were not robust (found only 2-4 times out of 1,000 trials).

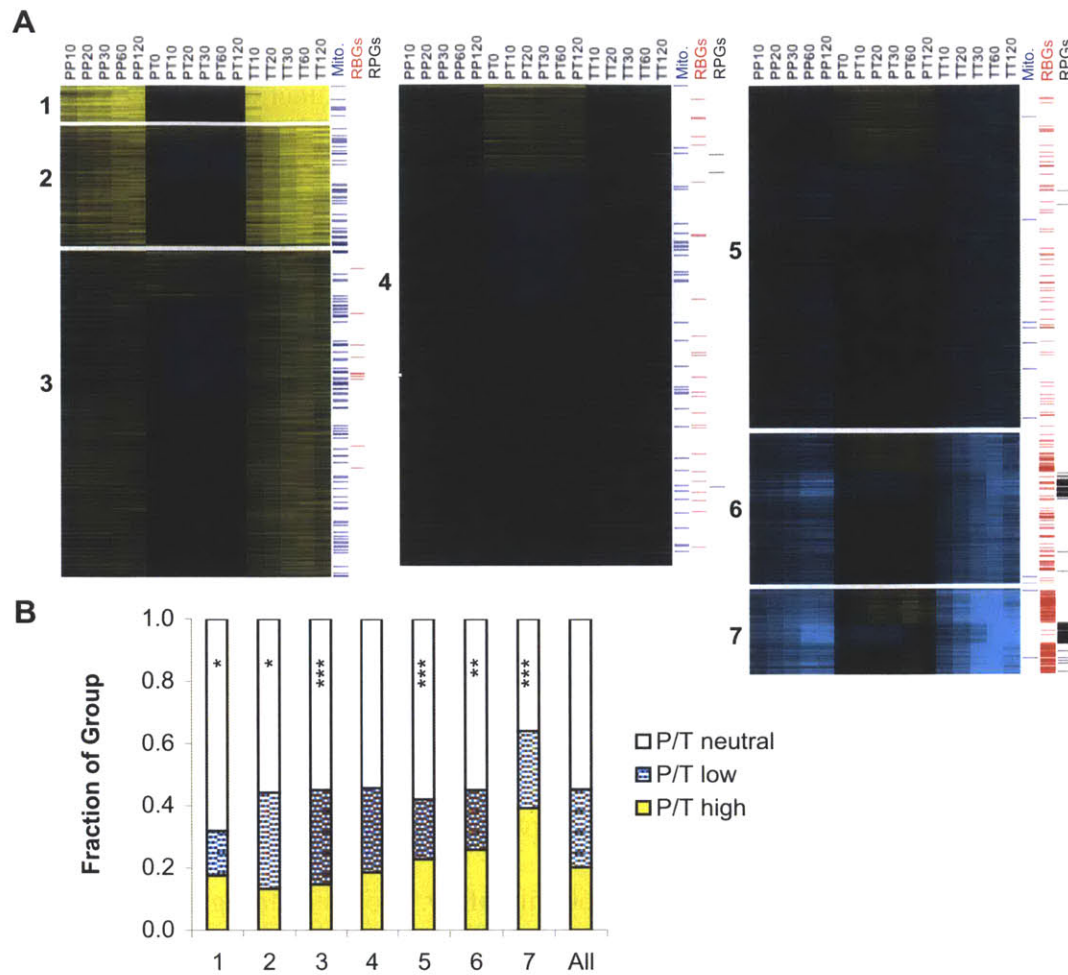


Figure III.5: Ribosome Occupancy and mRNA Abundance are Divergently Regulated

(A) The seven groups of genes identified by k-means clustering of T/T ratios include different proportions of genes with high, low, or neutral P/T ratios. Data and color scale are as in Figure III.1C, with rows (genes) re-ordered to highlight the differences in ribosome occupancy (P/T) among genes with similar glucose-withdrawal induced changes in total mRNA abundance (T/T). P/P ratios are displayed for comparison but were not considered during clustering. Selected GO categories enriched in specific P/T groups are indicated. (“mito.” – mitochondrial; “RBG” – ribosome biogenesis; “RPG” – ribosomal protein gene) (B) P/T ratios are not equally distributed among T/T groups. Asterisks (*) indicate significant deviations from the distributions predicted by chance (Fisher’s exact test 2-tailed p-value, * < 0.05; ** < 0.005; *** < 0.0005). See also Figure III.4.

Although we anticipated that kinetic analysis of the total mRNA changes in response to glucose starvation would reveal temporal distinctions between genes, the groups of co-regulated genes identified by *k*-means clustering at *k* = 7 differed primarily by the magnitude of relative increase/decrease rather than by the timing of the maximal changes in gene expression (Figure III.5A). Clustering with values of *k* > 12 did reveal kinetically distinct patterns of mRNA accumulation that are consistent with current understanding of transcriptional responses to glucose withdrawal (Figure III.4). For example, genes subject to repression by Mig1 in the presence of glucose (e.g. *SUC2* and *CAT8*) were de-repressed within 10 minutes following glucose withdrawal, whereas known targets of the Cat8 transcription factor (e.g. *PCK1*, *ICL1*, and *FBP1*) accumulated mRNA strongly only after 120 minutes (Figure III.2). Groups of genes clustered based on starvation-induced changes in total relative mRNA levels ranged from strongly induced (119 genes, median induction at t = 60 min of 19.7-fold) to strongly repressed (311 genes, median repression at t = 60 min of 11.9-fold). The use of multiple time points as well as both biological and technical replicates allowed the confident identification of genes displaying modest yet consistent changes in mRNA levels. The most weakly induced category includes more than 100 genes encoding mitochondrial proteins known to be important for growth in the absence of glucose, highlighting the potential biological significance of coordinated small changes in gene expression.

Changes in mRNA relative abundance and ribosome occupancy were not independent of one another ($\chi^2 = 179$, $p < 0.0001$ for T/T vs. P/T). Starvation-induced genes were more likely to show low ribosome occupancy after glucose withdrawal, and repressed genes were more likely to show high ribosome occupancy (Figure III.5B).

Despite the statistical interdependence of changes in total mRNA levels and ribosome occupancy, for each group of genes having similarly induced/repressed total mRNA levels there were many genes displaying each of the possible post-transcriptional regulatory behaviors.

Functionally Distinct Groups of Genes Are Co-Regulated at the Post-Transcriptional Level

To investigate the possibility that differences in post-transcriptional regulatory behavior are biologically significant, the function of genes in each category was examined by gene ontology (GO) analysis. Notably, the GO terms that were significantly enriched ($p < 0.01$ with Bonferroni correction for multiple hypothesis testing) for the seven mRNA abundance-based groups (highly induced, moderately induced, weakly induced, strongly repressed, moderately repressed, weakly repressed, and unchanged) segregated within these groups along post-transcriptional regulatory divisions. A complete list of significant GO terms for each of the 21 regulatory groups (7 T/T x 3 P/T) is provided in Supplemental Table III.1 available on CD at MIT's Institute Archives & Special Collections or online with this article as Table S1 at Molecular Cell. 2011 Dec 9;44(5):745-58. doi: 10.1016/j.molcel.2011.09.019. Rarely were GO categories split between multiple post-transcriptional (P/T) regulatory groups, and where such a split occurred, the GO category spanned two out of three most similar groups – 'high' and 'neutral' or 'low' and 'neutral', not 'high' and 'low'. This suggests that the post-transcriptional behavior of a gene is related to its biological function in the starved cell.

The post-transcriptional partitioning of functionally related genes may derive from mechanistic similarities in their gene expression pathways. For example, the nuclear-encoded mitochondrial protein genes showed consistently low ribosome occupancy despite the fact that these genes are transcriptionally up-regulated in response to glucose withdrawal and encode proteins required for the cellular adaptation to low glucose conditions. mRNAs of some nuclear-encoded mitochondrial proteins are translated on cytosolic ribosomes that associate with mitochondria (Marc et al. 2002; Sylvestre et al. 2003). Subcellular localization of these mRNAs is driven by cis-acting elements in their 3'UTRs, and requires both the trans-acting RNA-binding protein Puf3 and the translocase of the mitochondrial outer membrane complex for full localization (Eliyahu et al. 2009; Saint-Georges et al. 2008). The apparent lag between the appearance of these transcriptionally induced mRNAs in the cell and their association with polysomes in our experiments could be explained by translational silencing of messages in transit from the nucleus to the periphery of mitochondria. In light of this explanation for the low P/T ratios of mRNAs encoding mitochondrial proteins, it is interesting to consider whether other transcriptionally induced yet poorly translated mRNAs identified in our analysis might also be subject to localization-dependent translational control.

The most striking example of gene function partitioning according to post-transcriptional regulatory behavior was the separation of cytoplasmic ribosomal protein genes (RPGs) from ribosome biogenesis factors (RBGs) (Figure III.5A, III.6, III.7). Genes from both functional groups were moderately to strongly reduced in both the total and polysomal mRNA pools after glucose withdrawal. This down-regulation is likely due to the greatly reduced demands for new ribosome synthesis as the cells transition from

rapid growth and division, requiring the assembly of ~200,000 new ribosomes every 90 minutes (Warner 1999), to cellular differentiation and slower growth in the invasive filamentous form (Cullen and Sprague 2000). The two groups diverged in their post-transcriptional responses to glucose starvation. The P/T ratios of RBG mRNAs increased after glucose withdrawal and remained relatively high throughout the two-hour experiment. In contrast, the mRNAs encoding RPGs showed very low P/T ratios between 10 and 30 minutes after glucose withdrawal, and these ratios increased between 30 and 120 minutes (Figure III.5A). Low P/T ratios indicate that a population of non-translating mRNA exists in the cell. To test this interpretation directly, we performed qRT-PCR on polysome gradient fractions after 10 minutes of glucose starvation. RPG mRNAs (low P/T genes) accumulated in ribosome-free fractions at the top of the gradient, whereas mRNAs from genes with high P/T ratios did not (Figure III.8). In principle, the apparent increase in ribosome occupancy for RPGs after 60-120 minutes could result from improved translation (increased P) or from degradation of the non-polysomal pool of mRNA. Given that P/P and T/T ratios for RPGs were divergent at early times minus glucose and converged after 60 minutes of starvation (Figure III.6), the second interpretation is more plausible, suggesting that this sub-population of mRNA is only transiently stable as a non-translating pool. Notably, the messages that were most dramatically down-regulated upon glucose withdrawal and had low P/T values in acutely starved cells are the most abundant (Figure III.7). Transient preservation of these mRNAs in a non-translating pool (Figure III.8) may reflect a bet-hedging strategy (see Discussion).

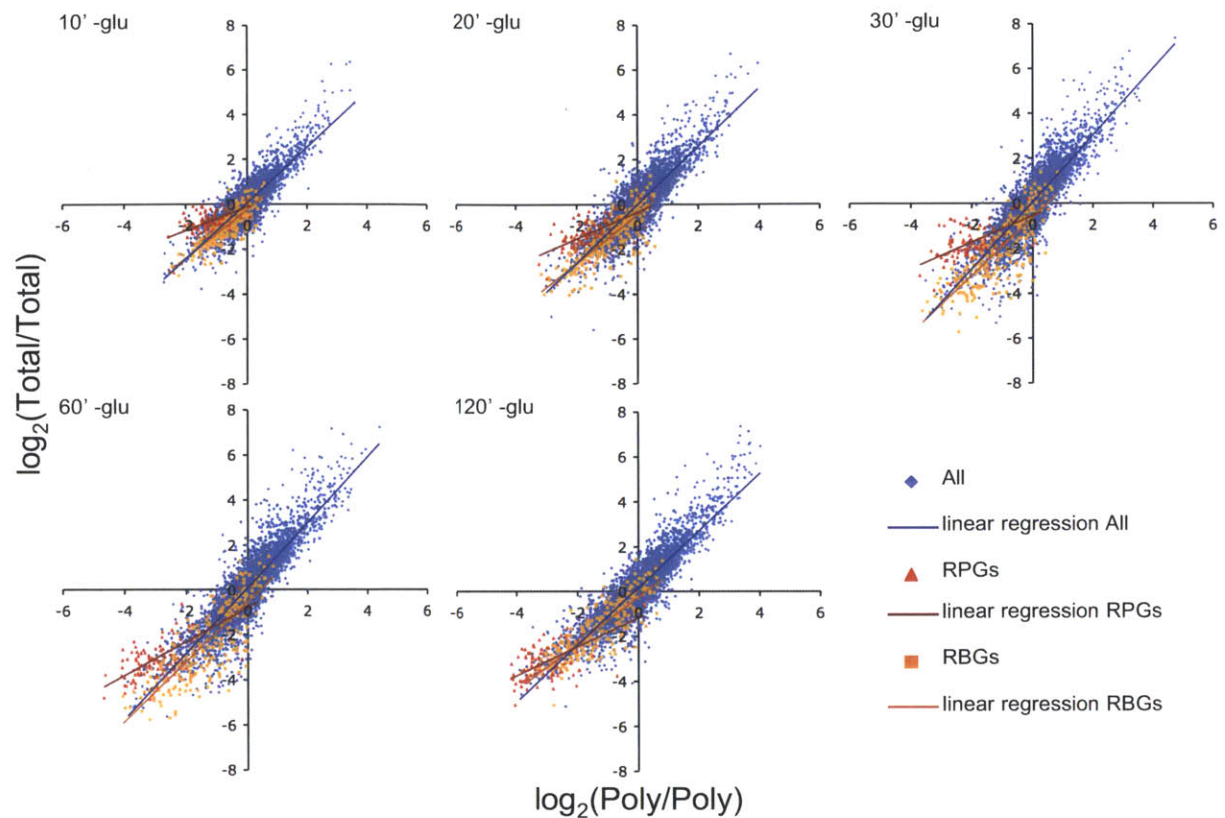


Figure III.6: RPGs and RBGs Differ in their Post-Transcriptional Responses to Glucose Withdrawal

Kinetics of total mRNA abundance changes compared to polysomal mRNA abundance changes following glucose withdrawal. RPG mRNAs (red triangles) were preferentially depleted from the polysomal RNA pool compared to the total mRNA pool at early times. RBGs (gold squares), like most genes (blue diamonds), disappeared from the totals in concert with their loss from polysomes. After 120 minutes –glucose, the RPGs T/T versus P/P ratios more closely resembled the population as a whole, indicating a loss of a non-polysomal pool of mRNA. The results of linear regression analysis for each group of genes are shown by color-coded lines. See also Figure III.8.

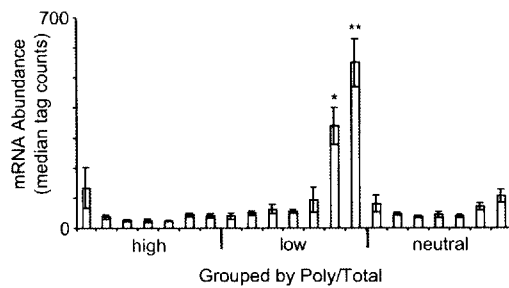


Figure III.7: Highly Expressed mRNAs are Preferentially Retained in the Non-Translating Pool

Mean mRNA abundance by group, as determined for each ORF by tag counts from next-generation sequencing of mRNA from cells grown in rich media (Nagalakshmi et al. 2008). Groups are as in Figure III.5A. The two T/T groups that were strongly decreased following glucose withdrawal and also showed low P/T ratios are comprised of significantly more abundant mRNAs ($p < 0.05$) than all other groups. Significance was assessed by Student's t-test with Bonferroni correction. Within each P/T classification, groups are arranged from left to right from highest T/T ratio to lowest. Error bars indicate standard error of the mean (SEM).

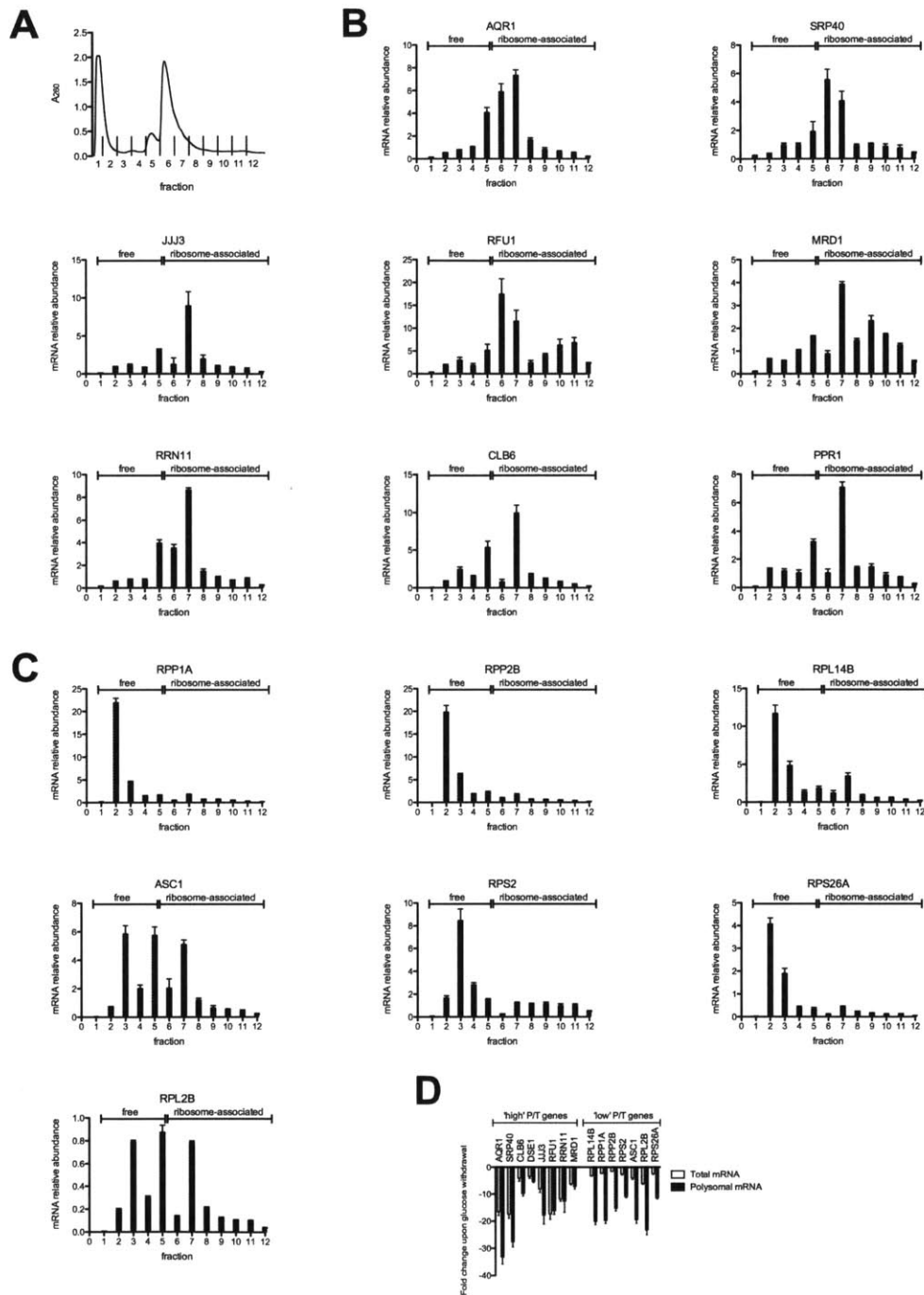


Figure III.8: Quantitative RT-PCR Validation of Select Genes' P/T ratios -Glucose
(A) Polysome gradient fractions from glucose-starved cells (10 minutes). **(B, C)** mRNA abundance per fraction relative to 1/12th input (set equal to 1). **(B)** high P/T genes by microarray. **(C)** low P/T genes. **(D)** qRT-PCR comparison of mRNA abundance before and after 10 minutes of glucose starvation in total or pooled polysomal mRNA samples.

Only a Subset of mRNAs Can Return to Polysomes Following Starvation and Re-Feeding

Previous reports showed that the global reduction in protein synthesis caused by ten minutes of glucose withdrawal can be reversed within five minutes of glucose re-addition (Ashe et al. 2000; Brengues 2005). It was proposed that this rapid recovery of translation is due to mobilization of stored, translationally silenced mRNPs from P-bodies (cytoplasmic RNP granules) to polysomes. Only two reporter genes were tested for the capacity to return to polysomes in the absence of new transcription (Brengues 2005). Our data suggest that most mRNAs are depleted from the total mRNA pool coincident with their loss from polysomes (Figure III.1C, III.8D), and that the capacity for translational resurrection of pre-existing messages is thus narrowly restricted to a select sub-population that includes the RPG mRNAs. To test this hypothesis directly, we examined global as well as gene-specific recovery of translation following glucose re-addition to starved cells. New transcription was inhibited with the drug thiolutin (Grigull et al. 2004), to assess the capacity of pre-existing mRNAs to return to polysomes. Treatment with thiolutin slightly blunted translational recovery upon glucose re-addition, resulting in fewer heavy polysomes, more disomes, and persistence of a larger 80S monosome peak (Figure III.9A). These results indicate that initiation of translation on newly synthesized mRNAs accounts for some of the ‘recovery’ of translation after glucose re-addition, even after only 10 minutes of glucose starvation.

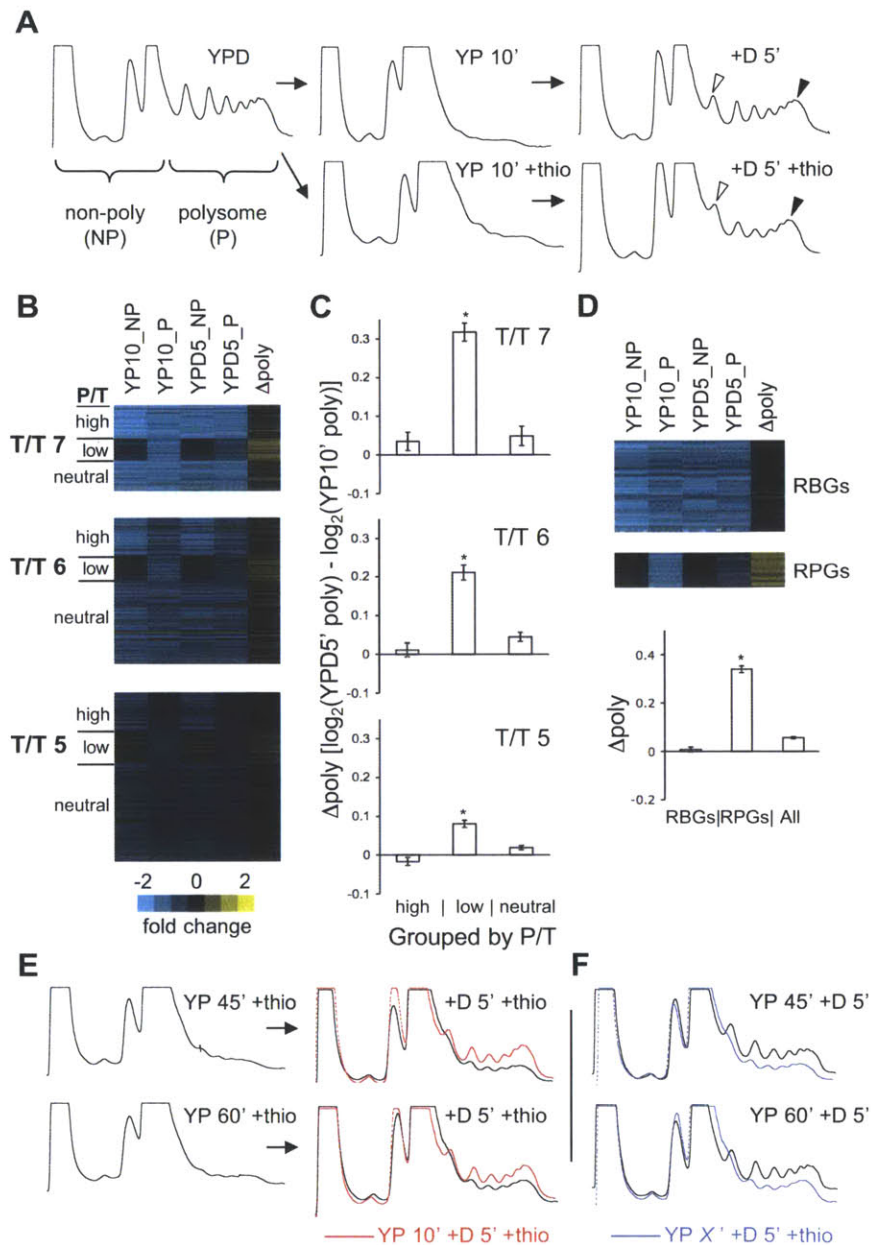


Figure III.9: Translational Resurrection is Restricted to a Subset of Genes for a Limited Time

(A) Polysome profiles of cells subjected to 10 minutes of glucose starvation followed by 5 minutes of glucose repletion in the presence (+T) or absence of thiolutin to inhibit new transcription. Thiolutin treatment slightly reduced polysome recovery. Note the relative heights of the disome (open arrowhead) and polysome (filled arrowhead) peaks and the widths of the monosome peaks \pm thio. Analytical polysome assays (A, E, F) were repeated 2-4 times. Representative traces are shown for each. (B) Ratio values shown by color scale from microarray analysis of non-polysomal (NP) and polysomal (P) mRNA

from cells starved for 10 minutes (YP 10' +thio) or starved and re-fed for 5 minutes (+D 5' +thio). Polysomal mRNA from YPD cultures served as the reference sample for each array. Genes are organized according to T/T and P/T groups from Figure III.5. Group 5 is shown at 50% vertical scale. Genes that showed low P/T ratios in starved cells were less depleted from the non-polysomal fraction than genes with high or neutral P/T ratios and showed greater mobilization into polysomes upon re-feeding. "Δpoly" values were derived from the ratio of ratios (YPD5_P versus YP10_P). (C) Graphs show quantification of the Δpoly values from B. Error bars in C and D indicate SEM. Asterisks (*) indicate Student's t-test p-value < 0.0001. (D) RPGs as a class preferentially recovered in polysomes upon glucose re-addition. "All" = all genes from T/T 5-7. (E) Prolonged glucose starvation in the absence of new transcription leads to reduced polysome recovery upon re-feeding. Cells were starved for 45 (top row) or 60 (bottom row) minutes in the presence of thiolutin before re-feeding for 5 minutes. Polysome profiles of recovery (+D 5') after only 10 minutes -glucose +thio are overlaid for comparison (red lines). (F) New transcription contributes substantially to polysome recovery after prolonged starvation. Polysome profiles of recovery (+D 5') after 45' or 60' of starvation are shown in black. Polysome recovery after starvation for 45 or 60 minutes +thiolutin is shown (blue lines) for comparison. See also Figure III.10.

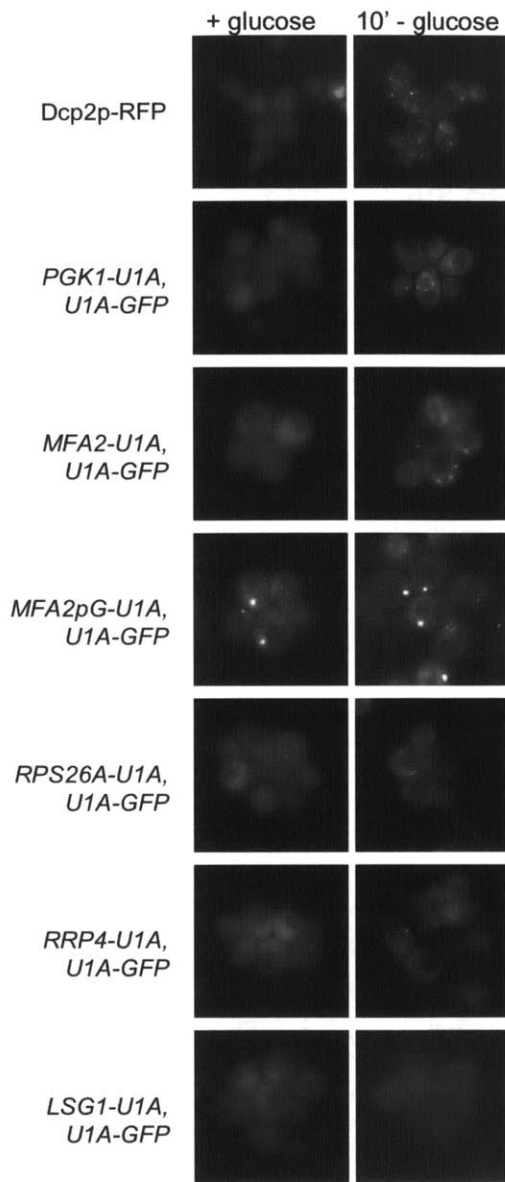


Figure III.10: P-bodies are Present Under the Conditions Examined by Microarray

Cells carrying previously characterized P-body protein (Dcp2) or RNA (PGK1-U1A, MFA2-U1A, and MFA2pG-U1A co-expressed with U1A-GFP) markers on plasmids were imaged in glucose-containing or glucose-depleted media as indicated. Localization of mRNAs from low P/T (RPS26A) and high P/T (RRP4, LSG1) genes was similarly assessed. Representative images are shown.

Nevertheless, substantial polysome recovery occurred even when new transcription was inhibited (Figure III.9A). To determine which genes participate in this

recovery, the polysomal and non-polysomal mRNA pools were examined using microarrays following 10 minutes of glucose starvation and again after 5 minutes of glucose repletion, with thiolutin present throughout. A gene's P/T ratio during glucose starvation predicted its ability to return to polysomes upon glucose re-addition. Genes that showed high or neutral P/T ratios following glucose removal (Figure III.5A) were depleted from the non-polysomal as well as the polysomal mRNA pools after 10 minutes of glucose starvation, and showed very little mobilization into polysomes 5 minutes after glucose re-addition (Figure III.9B, C). This argues that genes like the RBGs, which have high P/T ratios under conditions of decreasing P, do so because of rapid depletion of the non-polysomal mRNA. Consistent with this interpretation, we found by qRT-PCR that 80-95% of the total mRNA from these high P/T genes is gone from the total mRNA pool by 10 minutes after glucose starvation. In contrast, the genes that showed low P/T after glucose withdrawal were preferentially depleted from the polysomal compared to the non-polysomal pool (Fig III.8D). This group of genes also showed significantly greater mobilization into polysomes upon glucose re-addition by microarray ($p < 0.0001$), and by qRT-PCR analysis of select genes' mRNA abundance in polysome gradient fractions (Figure III.11). Similar results were observed for more moderately down-regulated genes, with the low P/T sub-population showing significantly greater recovery than either the high or neutral P/T genes, although the extent of recovery was somewhat less. The RPGs as a class showed ~6-fold more recovery than all other starvation-repressed genes (T/T clusters 5, 6 and 7 from Figure III.5A) combined ($p < 6.0 \times 10^{-54}$), whereas RBGs did not recover (Figure III.9D).

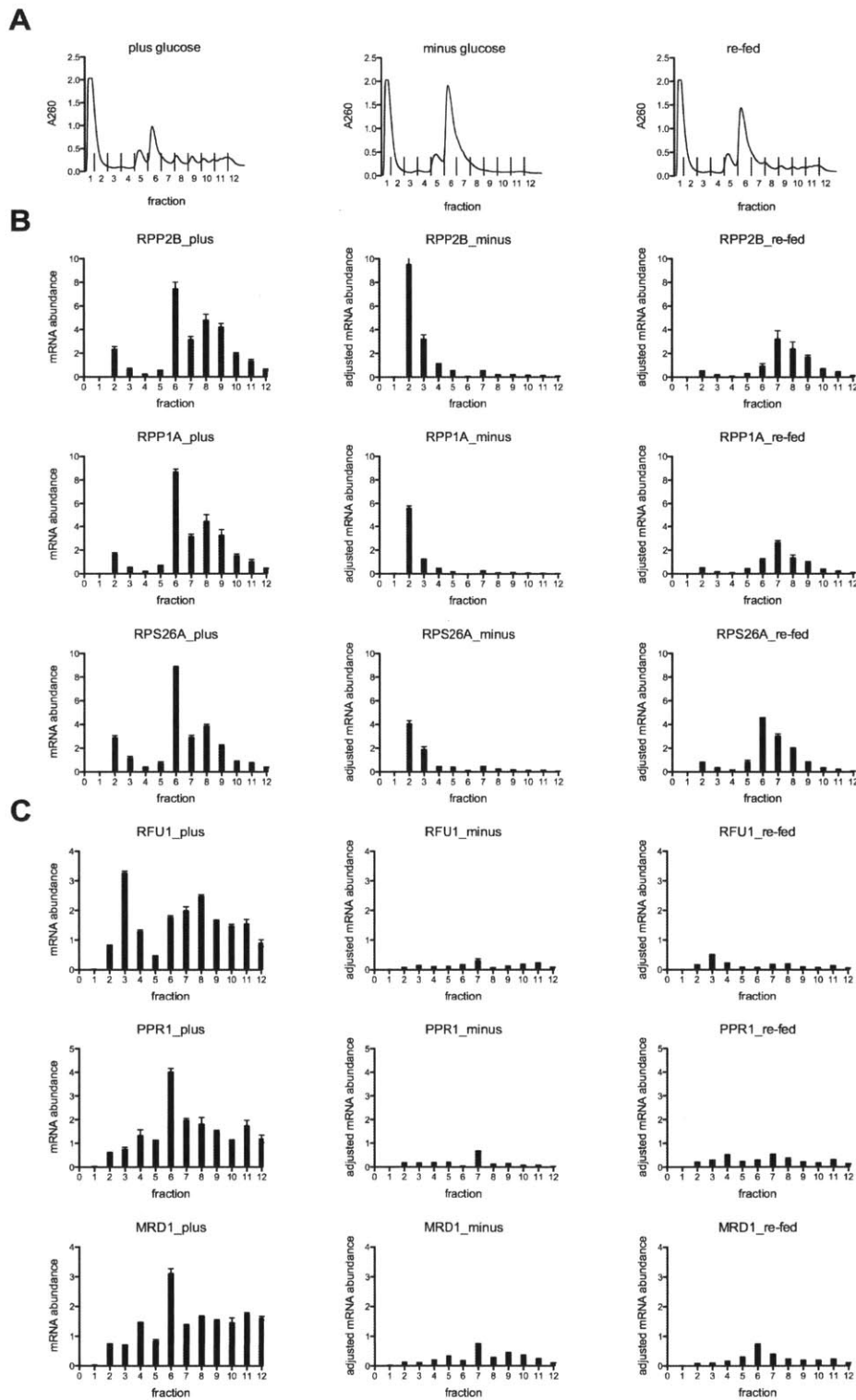


Figure III.11: Quantitative RT-PCR Validation of Select Genes' mRNA Abundance in Polysome Fractions Following Glucose Starvation and Re-Feeding

(A) Polysome gradient fractions from plus glucose (left), starved (10 minutes minus glucose, with thiolutin, center), and re-fed (10 minutes minus glucose and 5 minutes plus glucose, with thiolutin, right). (B, C) (left) mRNA abundance per fraction, relative to 1/12th input and normalized to Fluc dope-in control RNA. (center, right) Adjusted mRNA abundance per fraction determined by qRT-PCR comparison with plus glucose fractions, shown on the same scale as plus glucose samples. Error bars indicate standard deviation of the mean of three replicates.

We verified that P-bodies formed under these conditions based on the localization of previously characterized protein and RNA reporters (Figure III.10). In addition, we examined the localization of RPS26A-U1A, a low P/T gene capable of returning to polysomes upon re-feeding, and of RRP4-U1A and LSG1-U1A, high P/T genes that are largely depleted within 10 minutes of glucose withdrawal and are not capable of returning to polysomes in the absence of new transcription. We detected P-body localization for RPS26A-U1A and RRP4-U1A. We did not detect LSG1-U1A, which was very lowly expressed by Northern blot (data not shown). These reporter experiments don't distinguish whether the P-body localized portion of RPS26A mRNA is the fraction of intact mRNA that returns to polysomes upon re-feeding (~50%, Figure III.11B), or alternatively, if the P-body localized portion is comprised of decay intermediates of the fraction (~50%, Figure III.8D) that disappeared from the total mRNA pool, based on qRT-PCR using primers within the ORF. We found that stabilizing 3'UTR decay intermediates led to prominent P-body localization (Figure III.10), which is consistent with previous reports (Sheth and Parker 2003). Thus, P-bodies likely contain endogenous mRNAs undergoing decay. Whether or not they also contain stored translationally repressed RPG mRNAs remains an open question. Wherever translationally repressed mRNAs reside in the cell, only select mRNAs are able to persist in a non-translating pool and resume active translation following a short period of glucose starvation.

The glucose starvation microarray time course data suggest that the non-translating sub-population of mRNAs that accumulates at early times is depleted after more prolonged glucose starvation (Figure III.5A, III.6). This interpretation of the data predicts a turning point after which any recovery of polysomes upon glucose re-addition would require new transcription. To test this prediction, we subjected cells to varying periods of glucose starvation, with and without inhibition of new transcription by the drug thiolutin, and examined global translation activity after five minutes of glucose re-addition. After 45 or 60 minutes of starvation in the absence of new transcription, polysome recovery was greatly reduced compared to cells starved for only 10 minutes (Figure III.9E). If new transcription was allowed to occur during extended starvation, translational recovery upon glucose re-addition increased (Figure III.9F). These observations, together with the indication that specific mRNAs (RPGs) are lost between the 10- and 60-minute time points (Figure III.6), suggests that translational activation of these silenced mRNAs contributes substantially to polysome recovery upon re-feeding of briefly starved cells. These data argue that cells' inability to rapidly restore translation without new transcription after longer periods of starvation is due to the disappearance of a population of non-translating mRNAs. Between 30 and 60 minutes appears to be a switch point for P/T ratios in our data, regardless of whether the ratios are increasing or decreasing. This suggests that the timing of the 'turning point' for cells to recover translation of stored RPGs mRNAs may relate to widespread changes in the activity of factors that influence P/T through effects on mRNA stability.

Molecular Insights Into Selective Preservation of Non-Translating RPG Transcripts

How are certain mRNAs able to persist in the cell, even transiently, following glucose starvation, when the majority of mRNAs are depleted from the total RNA pool coincident, within the time resolution of our experiments, with their loss from polysomes? Comparison with genome-wide mRNA half-life measurements indicates that stability in glucose-replete conditions is probably not the determining factor for an mRNA's capacity to persist in a stable non-translating pool after glucose withdrawal (Figure III.12). In particular, the RPGs as a class have short half-lives in glucose-replete conditions compared to most genes (Grigull et al. 2004; Wang et al. 2002).

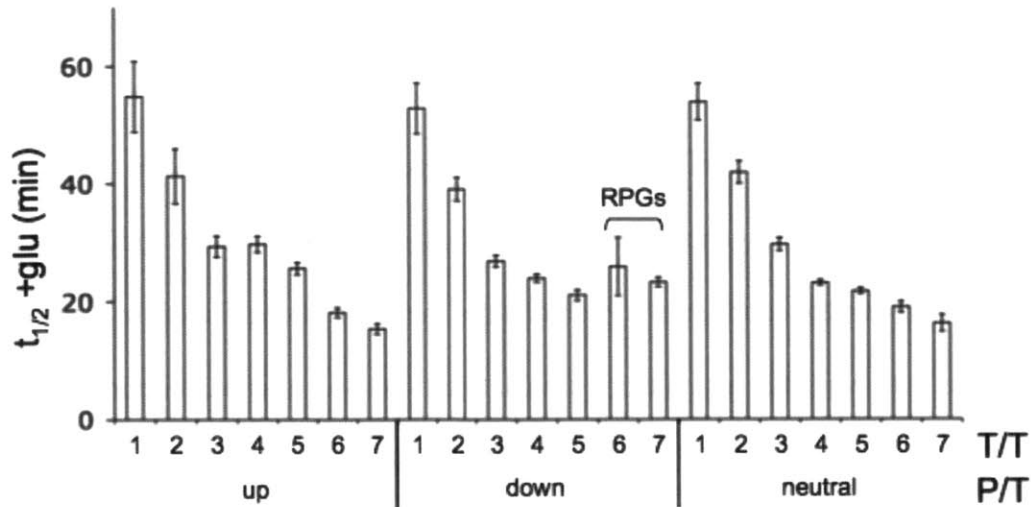


Figure III.12: RPGs Have Relatively Short Half-Lives in Glucose Replete Conditions

Mean mRNA half-lives by group, as determined for cells grown in rich media (Wang et al. 2002). Within each P/T classification, groups are arranged from left to right from highest T/T ratio to lowest. T/T group numbers correspond to Figure III.5. The two groups that include the RPGs are indicated. Error bars indicate SEM.

Alternatively, the ‘post-transcriptional operon’ hypothesis posits a role for RNA-binding proteins (RBPs) in coordinating the post-transcriptional fate of functionally related genes (Keene 2007; Keene and Tenenbaum 2002). To investigate the possibility that specific RBPs affect the post-transcriptional behavior of mRNAs following glucose withdrawal, we examined data from a recent genome-wide association study that identified mRNA targets of 40 yeast RBPs (Hogan et al. 2008). Comparing the ‘transcriptional’ (changes in total mRNA abundance) and post-transcriptional (P/T) regulatory groups identified in our study with RBP-mRNA target groups revealed concordance between RBP association patterns and P/T but not T/T (Figure III.13A). For simplicity, only the seven RBPs that showed significant enrichment or de-enrichment ($p < 0.01$, Bonferroni corrected) of target mRNAs within at least one group are displayed. Similar to GO terms, which rarely spanned opposing P/T groups, RBP-association

profiles were similar for “high” and “neutral” P/T groups for many RBPs, and for “low” and “neutral” for Puf3, whereas the “high” and “low” groups had almost no enrichments in common and frequently appeared to be mirror images of each other. In contrast, organizing RBP-association patterns by T/T group resulted in a ‘checkerboard’ appearance, despite the fact that T/T groups contain functionally and cytologically related genes. This suggests that the coherence of the association between P/T behavior and particular RBPs might be the result of direct effects of these RBPs on mRNA translatability and/or mRNA stability outside of polysomes.

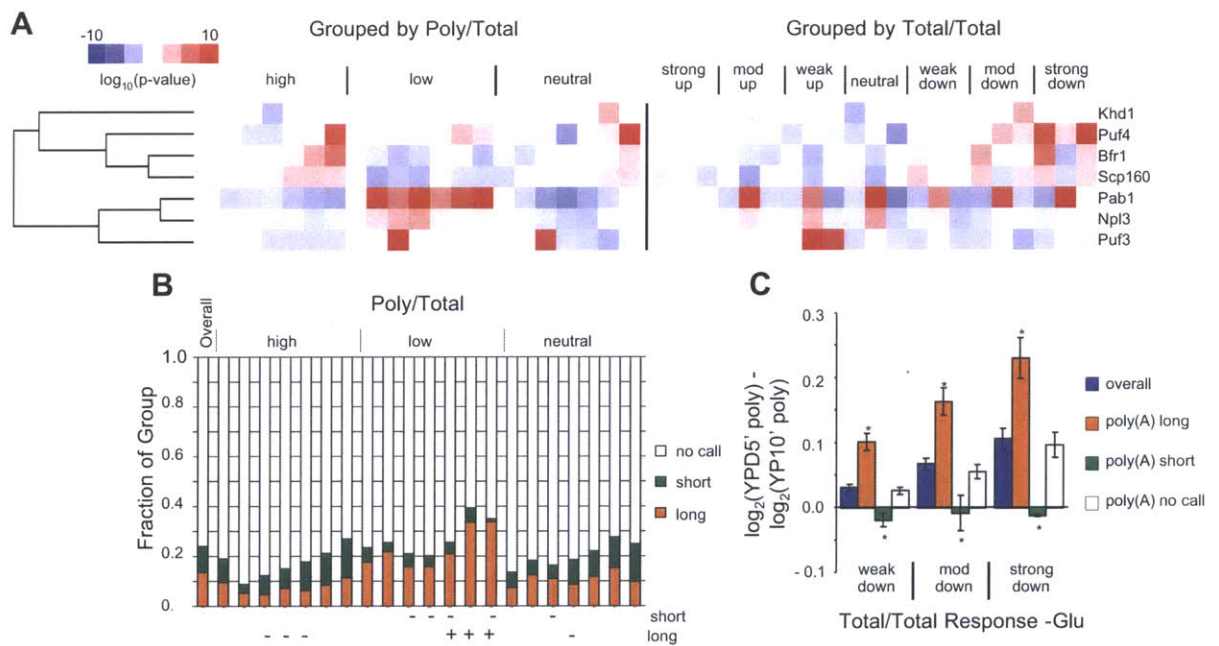


Figure III.13: Resurrection-Competent mRNAs Associate with Pab1 and Have Longer Poly(A) Tails

(A) Comparison of mRNA behavior following glucose withdrawal (7 T/T groups x 3 P/T groups) with RBP association profiles [Hogan et al.]. RBP target lists (rows) were clustered according to p-values for enrichment or de-enrichment within a given mRNA regulatory group. P-values were obtained using Fisher’s exact test with Bonferroni correction for multiple hypothesis testing. Columns are grouped by P/T (left) or T/T (right) similarity, ordered from left to right: high, low, neutral, and from most increased to most decreased T/T. Note that the 'strong down' T/T category includes genes that are

not strongly reduced until after 60 minutes, although they are strongly reduced in the P/P at earlier times. (B) mRNAs with low P/T ratios have longer poly(A) tails as a group than mRNAs with high P/T. This effect is most pronounced for the group of genes that was most strongly down regulated (T/T) following glucose withdrawal. Poly(A) tail lengths were determined genome-wide and classified as ‘long’, ‘short’, or ‘no call’ (neither long nor short compared to most genes) [Beilharz and Preiss]. + and – indicate the presence of significantly too many or too few long- or short-tailed mRNAs within each of the 21 mRNA groups. Significance was assessed by Fisher’s exact test (Bonferroni corrected p-value < 0.01). (C) mRNAs with long poly(A) tails show greater capacity for translational recovery upon glucose re-addition than mRNAs with short or average length tails. Polysome recovery was assessed as described in Figure III.9. Error bars indicate SEM. Asterisks (*) indicate p < 0.05 (Bonferroni corrected, Student’s t-test for difference compared to all other transcripts in the same T/T group). See also Figures III.12 and III.14.

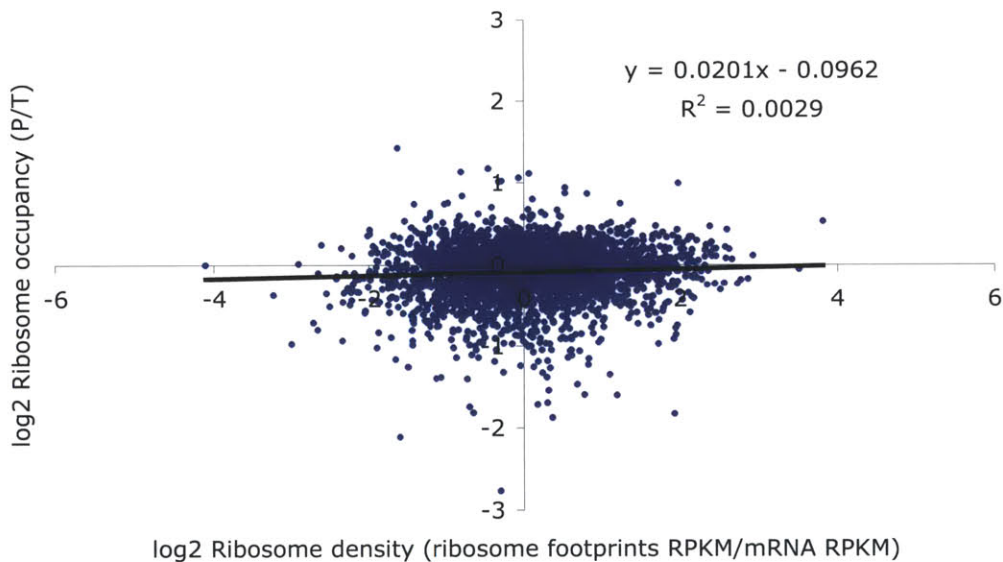


Figure III.14: Ribosome Occupancy and Ribosome Density are Essentially Unrelated

Ribosome occupancy (P/T) values from plus glucose samples (Figure III.5) compared with ribosome density as determined by ribosome footprint profiling (Ingolia et al. 2009). Each spot represents a single gene.

Pab1-associated mRNAs were notably enriched with groups having low P/T ratios, and correspondingly depleted in groups with high and neutral P/T ratios (Figure III.13A). Comparison of P/T groups with genome-wide measurements of poly(A) tail

lengths (Beilharz and Preiss 2007) revealed a lack of short-tailed and an enrichment of long-tailed genes among mRNAs with low P/T values. Conversely, genes with long poly(A) tails were significantly under-represented among groups with high P/T ratios (Figure III.13B). The association of Pab1 and long poly(A) tails with groups of mRNAs with low P/T ratios was counterintuitive given Pab1's role in enhancing translation initiation and the positive correlation between poly(A) tail length and ribosome density (Beilharz and Preiss 2007). However, genome-wide ribosome occupancy (P/T), as measured in this study or by others (Arava et al. 2003), does not correlate with measures of translational efficiency (number of ribosomes/mRNA) determined by polysome fractionation (Arava et al. 2003) or by ribosome footprint profiling (Ingolia et al. 2009) (Figure III.13). Thus, we interpret the low P/T values of RPGs and high P/T values of RBGs after 10-30 minutes without glucose as consequences of differences in the stabilities of non-translating mRNAs, rather than as differences in translational activity following glucose withdrawal.

Consistent with a role for Pab1/poly(A) in stabilizing non-translating mRNAs, poly(A) tail length was positively correlated with the capacity for translational resurrection following glucose re-addition (Figure III.13C). The intersection between the two groups that show increased total mRNA levels, low P/T ratios, and enrichment with Pab1 (as well as Puf3) is dominated by nuclear-encoded mitochondrial protein genes, previously described as having long poly(A) tails (Beilharz and Preiss 2007) and thought to transit the cytoplasm in a translationally repressed state before being translated on peri-mitochondrial ribosomes (Eliyahu et al. 2009; Marc et al. 2002; Sylvestre et al. 2003). A

potentially unifying explanation for the enrichment of Pab1 with various low P/T groups is that Pab1 association may stabilize non-translating mRNAs.

Discussion

Translation is Reduced, Not ‘Inhibited’, Following Glucose Withdrawal

Since the initial discovery of a rapid and dramatic signal-mediated reduction in polysomes following glucose withdrawal (Ashe et al. 2000), subsequent studies have investigated the mechanisms responsible for ‘global inhibition’ of translation in glucose-starved cells (for example, (Bregues and Parker 2007; Bregues 2005; Hilgers 2006; Holmes et al. 2004; Hoyle et al. 2007)). Thus, we were initially surprised to observe widespread agreement between the relative abundance of mRNAs in the polysomal and total mRNA pools following glucose withdrawal. The data reported here show that mRNAs from more than 1,000 genes rapidly increased their relative association with polysomes in glucose-starved cells. Many of these genes are known to be transcriptional targets of well-characterized glucose-regulated transcription factors, and encode proteins required for survival and growth in the absence of glucose. Furthermore, the kinetics of accumulation of these mRNAs in polysomes was nearly indistinguishable from their appearance/accumulation in the total mRNA pool. The simplest interpretation of these data is that cells respond to glucose starvation by up-regulating expression of genes whose products promote adaptation to low glucose environments. This view agrees with extensive genetic and genome-wide gene expression studies (reviewed in (Zaman et al. 2008)), and is consistent with an earlier study, which reported (as data not shown) that protein synthesis, measured by [³⁵S]methionine incorporation rates, decreased to 20%

after 5 minutes without glucose, and returned to 40% by 15 minutes without glucose (Kuhn et al. 2001). We conclude that there is widespread translation at a reduced level in glucose-starved cells.

Most mRNAs are Depleted Coincident with Translational Repression

Previous studies suggested that widespread degradation of mRNAs is not part of the response to glucose withdrawal based on Northern blots of specific genes (Ashe et al. 2000). Examining the behavior of those genes in our genome-wide experiments reveals that their behavior is not typical. Three of the genes examined, *PAB1*, *ACT1* and *PGK1*, were barely reduced in polysomes after 10 minutes without glucose (12%, 13% and 5% reductions, respectively), and the fourth gene, *RPL28/CYH2*, is unusual in its capacity to survive as a non-translating mRNA. Among more than 2,000 genes showing consistently reduced mRNA levels in polysomes following glucose withdrawal, only a small minority (< 20%), dominated by the RPGs, persisted in a stable non-translating pool that could return to polysomes if glucose was restored. Most mRNAs were depleted from the total mRNA population as quickly as they were depleted from polysomes. This observation is consistent with reports showing increased decapping and decay of specific mRNAs in response to decreased translation initiation (LaGrandeur and Parker 1999; Schwartz and Parker 1999).

Mechanisms of Post-transcriptional Regulation in Glucose-starved Cells

As our data are consistent with widespread decay of translationally repressed mRNAs, we reconsidered the evidence for and against a model of widespread mRNA decapping as

the cause of translational repression following glucose starvation. Deletion of either component of the decapping enzyme complex (*dcp1Δ* or *dcp2Δ*) prevents polysome reduction following glucose withdrawal (Holmes et al. 2004). Moreover, the ability to inhibit protein synthesis following glucose removal correlated with the extent of residual decapping activity retained by various *dcp1* mutants. Our genome-wide results are equally consistent with the model that activation of decapping causes widespread translational repression and mRNA decay, or alternatively, that translational repression occurs first and mRNA decay follows soon after. We favor the first model because inhibition of decapping prevents translational repression, whereas no known translational repressive mechanism is required for polysome collapse, despite extensive efforts to identify such a mechanism (Holmes et al. 2004). Nevertheless, decapping and 5' to 3' mRNA decay can occur on polysome-associated mRNAs (Hu et al. 2009). Thus, for the majority of genes whose mRNAs are depleted from the total pool coincident, within the time resolution of our experiments, with their disappearance from polysomes, we cannot conclude whether they are first evicted from polysomes and then degraded, or whether the initiation of decay prevents subsequent loading of ribosomes (and permits run-off of previously loaded ribosomes).

If widespread decapping is in fact responsible for most of the translational repression in glucose-starved cells, three major questions remain. First, what is the trigger that activates decapping within one minute following glucose withdrawal? The rapidity of the response suggests that it involves post-translational modification of some factor. Mutations that disrupt either the glucose derepression signaling pathway or the Ras/PKA pathway interfere with translational repression (Ashe et al. 2000), but no direct molecular

connections to regulation of the decapping machinery have yet been characterized. Second, how do some messages escape decapping? There are two types of escapees: the ‘privileged’ mRNAs, including the RPG messages, that are present at the moment of glucose withdrawal and shift out of polysomes, and all the mRNAs that are produced afterwards. Third, what mechanism accounts for the translational repression of the RPGs? Our results implicate the unusually long poly(A) tails of some messages and their association with the poly(A) binding protein as likely determinants for the preservation of non-translating mRNA. This is an attractive model given the evidence that deadenylation is the rate-limiting step in most mRNA decay (Parker and Song 2004). For the second class of escapees – mRNAs whose abundance increases in glucose-starved cells – it is conceivable that they are endowed with longer poly(A) tails in the nucleus and/or that they are assembled into RNPs with features that render them inherently resistant to decapping. We do not favor such a model because these mRNAs are depleted from polysomes within 5 minutes following glucose re-addition (see complete array data in Supplemental Table SIII.2 available on CD at MIT’s Institute Archives & Special Collections or online with this article as Table S2 at Molecular Cell. 2011 Dec 9;44(5):745-58. doi: 10.1016/j.molcell.2011.09.019.). An alternative explanation is that the ‘special’ feature of newly transcribed mRNAs is simply that they enter the cytoplasm after the decapping machinery is sequestered in P-bodies.

How are the RPG mRNAs translationally repressed? The observation that this group of mRNAs readily returns to active translation upon glucose re-addition strongly suggests that they are not decapped. Furthermore, our model that their long poly(A) tails and stable association with Pab1 are important for escaping decapping argues that they

retain all of the features thought to be required for ribosome recruitment under normal conditions. The mTOR pathway is implicated in nutrient regulation of RPG translation in mammalian cells (Hamilton et al. 2006), but no such mechanism was thought to exist in yeast (Warner 1999). Although the molecular details of the yeast mechanism almost certainly differ, our data suggest that a translational control mechanism targeting RPGs does exist. Genome-wide investigations of yeast translational responses to other stresses including high salinity (Melamed et al. 2008), transfer to a non-fermentable carbon source (Kuhn et al. 2001), and rapamycin treatment (Preiss et al. 2003) have also reported translational repression of RPGs, suggesting that this mechanism acts downstream of a common stress response signaling pathway.

Implications for the Interpretation of P-bodies

Glucose starvation in yeast has been extensively studied as a model stress to investigate the relationship between translational repression and mRNA localization to cytoplasmic granules or P-bodies. A central tenet of the model is that P-bodies are not just sites of mRNA decay, but can also be sites for storage of translationally silent intact mRNAs capable of return to the actively translating pool in response to altered cellular conditions (Bregues 2005; Teixeira et al. 2005). Because core P-body components and their roles in mRNA metabolism are well conserved, this model has far-reaching implications for the regulation of gene expression in eukaryotic cells (Anderson and Kedersha 2006; Parker and Sheth 2007).

Our findings refute the notion that reciprocal movement between a non-translating mRNA pool and polysomes is a general property of eukaryotic mRNAs (Bregues 2005).

This model was based on previous observations of two reporter mRNAs following glucose starvation and repletion. We found that less than 20% of genes have the capacity to persist in a translationally silent state in glucose-starved cells. This raises the possibility that the majority of the proteins accumulated in P-bodies under these conditions are engaged with messages undergoing irreversible decay, although we cannot rule out the possibility that a subset of P-body associated endogenous mRNAs can escape and resume translation. If so, the RPG messages are good candidates.

Bet-Hedging

What advantage might transient preservation of RPG messages confer, and why should their behavior differ from the RBGs? Although the ribosome biogenesis factors and ribosomal proteins (RPs) function together in the assembly of new ribosomes, the factors act catalytically whereas the RPs are required stoichiometrically. Accordingly, the mRNAs encoding the RPs are among the most abundant messages in the cell, comprising up to 30% of cellular mRNA during rapid growth in rich media and requiring a significant investment of energy for their synthesis (Holstege et al. 1998; Nagalakshmi et al. 2008; Warner 1999). If the cell were to degrade these messages immediately following release from polysomes, reversing this decision would be energetically costly. By retaining these mRNAs in a non-translating state for 10-30 minutes after the perception of glucose withdrawal, the cell might preserve the capacity for rapid resumption of growth should glucose return or the starvation signal turn out to have been a false alarm. Given that the production of new ribosomes largely determines the growth rate and generation time of yeast in nutrient replete conditions (Warner 1999; Zaman et al.

2008), it is tempting to speculate that the selection of RPG mRNAs for transient preservation is a bet-hedging strategy by which the cell retains the capacity for a return to rapid growth and division while simultaneously initiating preparations for prolonged starvation. Previous investigations of global relationships between transcriptional and translational changes following other environmental stresses (high salt shock, transfer to a non-fermentable carbon source, or treatment with rapamycin to mimic nutrient deprivation) also described translational repression of RPGs through measurements that required the existence of a stable non-translating population of mRNA (Kuhn et al. 2001; Melamed et al. 2008; Preiss et al. 2003). The transient preservation of translationally repressed RPGs may be a general feature of yeast cells' responses to growth-inhibiting stresses. Although translational activity was not determined, exploration of the common response to diverse environmental stresses revealed reduction in both RBG and RPG mRNAs, with the RPGs showing a delay (Gasch et al. 2000).

An interesting and unexplained phenomenon revealed by our data is the existence of a turning point, beyond which the cell appears to give up on hedging its bets and commits fully to the starvation gene expression program. How is the timing of this turning point determined molecularly? Does the timer run at a constant speed, or can it be adjusted according to features of the environment that might predict how likely it is that the cell will experience a reversal of fortunes in the near future? Do all cells hedge a little, transiently storing 30-50% of their RPG mRNAs, or do some cells store 100% and others none? These questions remain for future studies.

In answer to our original question – why would the cell choose to regulate gene expression at the level of translation – we conclude in part: because translational

repression is reversible at relatively low cost. Direct testing of this adaptive fitness model awaits the identification of the molecular mechanism(s) responsible for transient translational repression of the RPGs, in order to assess its impact on competitive fitness under fluctuating conditions.

Acknowledgements

We thank Adam Carrol, Paige Nittler, and Manlin Luo for technical assistance with the microarrays; Gregg Whitworth for advice and custom software for data processing; Joel Greenwood for computer support; Hiten Madhani for yeast strains; Roy Parker and Pam Silver for plasmids; and Megan Bergkessel, Chris Burge, Bryan Clarkson and members of the Gilbert lab for helpful discussions and comments on the manuscript. This work was supported by grant R00GM081399 (NIGMS) to W.G.

References

- Anderson P, Kedersha N. 2006. RNA granules. *The Journal of Cell Biology* **172**: 803–808.
- Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D. 2003. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **100**: 3889–3894.
- Ashe MP, De Long SK, Sachs AB. 2000. Glucose depletion rapidly inhibits translation initiation in yeast. *Mol Biol Cell* **11**: 833–848.
- Beilharz TH, Preiss T. 2007. Widespread use of poly(A) tail length control to accentuate

- expression of the yeast transcriptome. *RNA* **13**: 982–997.
- Blake WJ, KAEm M, Cantor CR, Collins JJ. 2003. Noise in eukaryotic gene expression. *Nature* **422**: 633–637.
- Bregues M. 2005. Movement of Eukaryotic mRNAs Between Polysomes and Cytoplasmic Processing Bodies. *Science* **310**: 486–489.
- Bregues M, Parker R. 2007. Accumulation of polyadenylated mRNA, Pab1p, eIF4E, and eIF4G with P-bodies in *Saccharomyces cerevisiae*. *Mol Biol Cell* **18**: 2592–2602.
- Brodsky AS, Silver PA. 2000. Pre-mRNA processing factors are required for nuclear export. *RNA* **6**: 1737–1749.
- Clarkson BK, Gilbert WV, Doudna JA. 2010. Functional overlap between eIF4G isoforms in *Saccharomyces cerevisiae*. *PLoS ONE* **5**: e9114.
- Coller J, Parker R. 2004. Eukaryotic mRNA decapping. *Annu Rev Biochem* **73**: 861–890.
- Coller J, Parker R. 2005. General translational repression by activators of mRNA decapping. *Cell* **122**: 875–886.
- Cullen PJ, Sprague GF. 2000. Glucose depletion causes haploid invasive growth in yeast. *Proc Natl Acad Sci USA* **97**: 13619–13624.
- Eliyahu E, Pnueli L, Melamed D, Scherrer T, Gerber AP, Pines O, Rapaport D, Arava Y. 2009. Tom20 Mediates Localization of mRNAs to Mitochondria in a Translation-Dependent Manner. *Mol Cell Biol* **30**: 284–294.

- Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11**: 4241–4257.
- Grigull J, Mnaimneh S, Pootoolal J, Robinson MD, Hughes TR. 2004. Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Mol Cell Biol* **24**: 5534–5547.
- Hamilton TL, Stoneley M, Spriggs KA, Bushell M. 2006. TOPs and their regulation. *Biochem Soc Trans* **34**: 12–16.
- Hilgers V. 2006. Translation-independent inhibition of mRNA deadenylation during stress in *Saccharomyces cerevisiae*. *RNA* **12**: 1835–1845.
- Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. 2008. Diverse RNA-Binding Proteins Interact with Functionally Related Sets of RNAs, Suggesting an Extensive Regulatory System. *Plos Biol* **6**: e255.
- Holmes LEA, Campbell SG, De Long SK, Sachs AB, Ashe MP. 2004. Loss of translational control in yeast compromised for the major mRNA decay pathway. *Mol Cell Biol* **24**: 2998–3010.
- Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**: 717–728.

of the translational start codon. *RNA* **5**: 420–433.

Marc P, Margeot A, Devaux F, Blugeon C, Corral-Debrinski M, Jacq C. 2002. Genome-wide analysis of mRNAs targeted to yeast mitochondria. *EMBO Rep* **3**: 159–164.

Melamed D, Pnueli L, Arava Y. 2008. Yeast translational response to high salinity: Global analysis reveals regulation at multiple levels. *RNA* **14**: 1337–1351.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**: 1344–1349.

Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A. 2002. Regulation of noise in the expression of a single gene. *Nat Genet* **31**: 69–73.

Parker R, Sheth U. 2007. P Bodies and the Control of mRNA Translation and Degradation. *Molecular Cell* **25**: 635–646.

Parker R, Song H. 2004. The enzymes and control of eukaryotic mRNA turnover. *Nat Struct Mol Biol* **11**: 121–127.

Pleiss JA, Whitworth GB, Bergkessel M, Guthrie C. 2007. Transcript Specificity in Yeast Pre-mRNA Splicing Revealed by Mutations in Core Splicosomal Components. *Plos Biol* **5**: e90.

Preiss T, Baron-Benhamou J, Ansorge W, Hentze MW. 2003. Homodirectional changes in transcriptome composition and mRNA translation induced by rapamycin and heat shock. *Nat Struct Mol Biol* **10**: 1039–1047.

- Raser JM. 2005. Noise in Gene Expression: Origins, Consequences, and Control. *Science* **309**: 2010–2013.
- Richter JD. 1991. Translational control during early development. *Bioessays* **13**: 179–183.
- Saint-Georges Y, Garcia M, Delaveau T, Jourden L, Le Crom S, Lemoine S, Tanty V, Devaux F, Jacq C. 2008. Yeast Mitochondrial Biogenesis: A Role for the PUF RNA-Binding Protein Puf3p in mRNA Localization ed. J. Bähler. *PLoS ONE* **3**: e2293.
- Schwartz DC, Parker R. 1999. Mutations in translation initiation factors lead to increased rates of deadenylation and decapping of mRNAs in *Saccharomyces cerevisiae*. *Mol Cell Biol* **19**: 5247–5256.
- Sheth U, Parker R. 2003. Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. *Science* **300**: 805–808.
- Smirnova JB, Selley JN, Sanchez-Cabo F, Carroll K, Eddy AA, McCarthy JEG, Hubbard SJ, Pavitt GD, Grant CM, Ashe MP. 2005. Global gene expression profiling reveals widespread yet distinctive translational responses to different eukaryotic translation initiation factor 2B-targeting stress pathways. *Mol Cell Biol* **25**: 9340–9349.
- Sylvestre J, Vialette S, Corral-Debrinski M, Jacq C. 2003. Long mRNAs coding for yeast mitochondrial proteins of prokaryotic origin preferentially localize to the vicinity of mitochondria. *Genome Biol* **4**: R44.
- Teixeira D, Sheth U, Valencia-Sanchez MA, Brengues M, Parker R. 2005. Processing bodies require RNA for assembly and contain nontranslating mRNAs. *RNA* **11**: 371–

- Raser JM. 2005. Noise in Gene Expression: Origins, Consequences, and Control. *Science* **309**: 2010–2013.
- Richter JD. 1991. Translational control during early development. *Bioessays* **13**: 179–183.
- Saint-Georges Y, Garcia M, Delaveau T, Jourden L, Le Crom S, Lemoine S, Tanty V, Devaux F, Jacq C. 2008. Yeast Mitochondrial Biogenesis: A Role for the PUF RNA-Binding Protein Puf3p in mRNA Localization ed. J. Bähler. *PLoS ONE* **3**: e2293.
- Schwartz DC, Parker R. 1999. Mutations in translation initiation factors lead to increased rates of deadenylation and decapping of mRNAs in *Saccharomyces cerevisiae*. *Mol Cell Biol* **19**: 5247–5256.
- Sheth U, Parker R. 2003. Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. *Science* **300**: 805–808.
- Smirnova JB, Selley JN, Sanchez-Cabo F, Carroll K, Eddy AA, McCarthy JEG, Hubbard SJ, Pavitt GD, Grant CM, Ashe MP. 2005. Global gene expression profiling reveals widespread yet distinctive translational responses to different eukaryotic translation initiation factor 2B-targeting stress pathways. *Mol Cell Biol* **25**: 9340–9349.
- Sylvestre J, Vialette S, Corral-Debrinski M, Jacq C. 2003. Long mRNAs coding for yeast mitochondrial proteins of prokaryotic origin preferentially localize to the vicinity of mitochondria. *Genome Biol* **4**: R44.
- Teixeira D, Sheth U, Valencia-Sanchez MA, Brengues M, Parker R. 2005. Processing bodies require RNA for assembly and contain nontranslating mRNAs. *RNA* **11**: 371–

382.

Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO. 2002. Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci USA* **99**: 5860–5865.

Warner JR. 1999. The economics of ribosome biosynthesis in yeast. *Trends in Biochemical Sciences* **24**: 437–440.

Zaman S, Lippman SI, Zhao X, Broach JR. 2008. How *Saccharomyces* Responds to Nutrients. *Annu Rev Genet* **42**: 27–81.