mRNA sequence features determine the efficiency of translation termination and association of the nonsense-mediated mRNA decay machinery with elongating ribosomes

A Dissertation Presented

By

Kotchaphorn Mangkalaphiban

Submitted to the Faculty of the Morningside Graduate School of Biomedical Sciences at UMass Chan Medical School in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY

October 20, 2023

GENETICS
mRNA sequence features determine the efficiency of translation termination and association of the nonsense-mediated mRNA decay machinery with elongating ribosomes

A Dissertation Presented
By
Kotchaphorn Mangkalaphiban

This work was undertaken in the Morningside Graduate School of Biomedical Sciences Interdisciplinary Program

Under the mentorship of

___________________________________________________________
Allan Jacobson, Ph.D., Thesis Advisor

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

___________________________________________________________
Jill Moore, Ph.D., Member of Committee

___________________________________________________________
Joel Richter, Ph.D., Member of Committee

___________________________________________________________
Wendy Gilbert, Ph.D., External Member of Committee

___________________________________________________________
Andrei Korostelev, Ph.D., Chair of Committee

___________________________________________________________
Mary Ellen Lane, Ph.D.,
Dean of the Morningside Graduate School of Biomedical Sciences

October 20, 2023
Dedication

To, ป้าแม่ และ น้อง ๆ ทั้งสอง

ป้า, I wish you could have seen this in person. I hope you see this in spirit.
Acknowledgments

I would like to thank the following people, whom this dissertation would not have been possible without. First, I would like to thank Allan for the opportunity to work under his guidance. I joined the lab with no experience in bioinformatics and did not know what I was doing. Allan, I will be forever grateful for your patience while I was fumbling around, for believing that I can do it, for your care and support in every step of this journey, even things as trivial as drafting an email, for providing a wonderful learning environment not just for the sciences but also my professional development, for pushing me beyond the limit and beyond what I thought I was capable of, for all the fun discussions over every graph I made (ugly or pretty), and for diligently adding “the” and deleting “the” from my writing. (This is the only thing my language instinct just cannot grasp for some reason, but I promise I have improved on everything else!) Thank you for being an example and an inspiration as a scientist and as a leader. Thank you so much.

I would like to thank all the lovely members of the Jacobson Lab, Robin Ganesan, Feng He, and Chan Wu. Robin, thank you for mentoring me since I entered as a rotation student and continuing to mentor me throughout the years, for always being there to answer my incessant questions about ribosome profiling, for maintaining the lab environment in its best condition for everyone to work in, and for teaching me about work-life balance. Feng, thank you for always clarifying things for me whenever I felt like my head would explode from too much information, for showing me how to tackle complex results, and for your advice
both in sciences and in life. Chan, I probably would not have survived graduate school without you. Thank you for taking me places all the time, for guiding me through scientific and personal hiccups, showing me what life-long learning is like, and for being such a great friend as well as colleague. You all have made the lab such a wonderful place to be in.

I would like to thank my Thesis Research Advisory Committee (TRAC), Dr. Andrei Korostelev, Dr. Zhiping Weng, and Dr. Sean Ryder for their invaluable advice to my thesis project and for making sure I was doing well. Dr. Richard Baker, thank you for helping me learn bioinformatics and for answering even the most basic questions about coding or statistics. Dr. Jill Moore, thank you for your insights and advice about machine learning and statistics, which are central to the development of my thesis project. Dean Mary Ellen Lane, thank you for having our (graduate students’) best interests in mind and ensuring we all have a smooth graduate school career.

I am deeply grateful to Dr. Ashley Hart from the Student Counseling Service for teaching me the skills I need to get myself through hard times, school-related or not, in a healthy (and scientific!) way, an invaluable lesson that will stay with me throughout my life.

I would like to thank all the new friends I have made at UMass, for the camaraderie and support as fellow students in the cohort ever since our first year until now. I would like to thank all my long-time friends from home and college for watching me grow and believing in me. Special thanks to Amrisha for always
making sure I eat well, have at least one funny video/joke in my chat weekly, endless talks about random things from nothing to everything, and for refusing a gift from me until I graduate.

Lastly, this Ph.D. journey would not have even started, let alone finished, without the support from my family. Thank you ป้า for putting your children’s education as your first priority always and for providing all the support we need to succeed. I hope I have made you proud. Thank you แม่ for always listening to me even though you didn’t understand anything I was saying, providing strong mental support, and letting me be a kid again when I want to. Thank you น้องหวาน and น้องพี for the sacrifices you have unknowingly made in order for me to pursue my dreams. I would like to thank my American parents, Michael and Elynne Gaffney, for being another foundation of my educational journey and for always welcoming me with open arms.
Abstract

Translation of mRNA into protein is terminated when the ribosome encounters one of the three stop codons (UAA, UAG, and UGA) at the end of an open reading frame (ORF). Infrequently, stop codons are decoded by a near-cognate tRNA, allowing “readthrough” of the stop codon and synthesis of an extended polypeptide. When termination occurs prematurely, the mRNA is degraded by the nonsense-mediated mRNA decay (NMD) pathway. Premature and normal termination appear to differ in their efficiency, but the exact “rules” of how NMD distinguishes them mechanistically remain to be elucidated. Using ribosome profiling and bioinformatics analyses, this study aims to understand, at a transcriptome-wide level, the cis-acting elements that influence termination efficiency and how premature termination is recognized by Upf1, a key NMD factor. Analyses of yeast and human mRNA sequences in both normal and readthrough-inducing conditions revealed largely conserved roles of identities of the stop codon, the following nucleotide, P-site codon, and 3’-UTR length in readthrough efficiency regulation. The analyses of yeast mRNAs associated with Upf1-bound ribosomes demonstrated that Upf1 binds ribosomes in two distinct complexes across all mRNA ORFs, suggesting that Upf1 associates with the ribosome during translation elongation before premature termination takes place. Together, these results provide insights into the regulation of termination and the early steps of NMD at the transcriptome-wide level.
Table of Contents

Dedication .......................................................................................................................... iii
Acknowledgments .............................................................................................................. iv
Abstract ............................................................................................................................... vii
Table of Contents ................................................................................................................ viii
List of Tables ........................................................................................................................ xiv
List of Figures ........................................................................................................................ xvi
List of Copyrighted Materials Produced by Author ......................................................... xxii
List of Multimedia Files ..................................................................................................... xxii
CHAPTER I: Introduction ...................................................................................................... 1
Overview ............................................................................................................................... 1
Translation termination ......................................................................................................... 2
The release factors: eRF1 and eRF3 ....................................................................................... 2
Mechanisms of stop codon recognition and peptide release .............................................. 3
Stop codon readthrough ...................................................................................................... 7
Definition and mechanism of stop codon readthrough ....................................................... 7
Evolution of stop codon readthrough .................................................................................. 9
Cis-regulatory elements of readthrough efficiency ................................................................. 13
Stop codons are not created equal ....................................................................................... 14
Nucleotides immediately upstream and downstream of the stop codon ....................... 15
mRNA secondary structure in the 3'-UTR ........................................................................ 18
Trans-regulatory elements of readthrough efficiency ......................................................... 20
mRNA-specific trans-regulatory factors ........................................................................... 20
Global trans-regulatory factors ......................................................................................... 21
Nonsense-mediated mRNA decay (NMD) .......................................................................... 28
The Upf factors: Upf1, Upf2, and Upf3 ............................................................................. 29
Targets of NMD .................................................................................................................. 32
Mechanisms of NMD substrate targeting by Upf1 ............................................................ 39
Mechanisms of NMD activation and substrate degradation ................................................ 44
Nonsense mutations and human diseases ......................................................................... 50
Therapeutic strategies targeting termination: promoting stop codon readthrough .......... 51
Therapeutic strategies targeting NMD: increasing the amount of mRNAs available for translation ........................................................................................................... 62
Challenges for current therapeutic strategies ................................................................... 65
Scope of thesis ......................................................................................................................... 67

CHAPTER II: Transcriptome-wide cis-regulatory elements of normal termination codon readthrough in yeast ................................................................. 69

Abstract .................................................................................................................................. 70

Introduction .............................................................................................................................. 71

Results ...................................................................................................................................... 75

Inactivation of eRF1 promotes readthrough of both normal and premature stop codons and accumulation of ribosomes at those codons ............................................................................................................. 75

Analysis pipeline development ............................................................................................... 85

Random forest analyses reveal stop codon identity, P-site amino acid identity, and 3'-UTR length as important factors in predicting readthrough efficiency .............................................................................. 91

Readthrough-promoting stop codons and nucleotides at positions +4 to +9 occur at a genome-wide level ................................................................................................. 95

Specific combinations of nucleotides and tRNA in the ribosomal P site may influence readthrough efficiency ......................................................................................... 103

Readthrough efficiency increased with 3'-UTR lengths in the eRF1 mutants, but not in wild-type or recycling factor mutant cells ....................................................................... 107

Discussion .............................................................................................................................. 109

Materials and Methods ......................................................................................................... 113

Yeast strains and culture growth conditions ........................................................................... 113

Library preparation and sequencing ....................................................................................... 114

Reads processing and alignment ............................................................................................ 115

Public ribosome profiling data sets ........................................................................................ 118

Data analyses .......................................................................................................................... 119

Random forest models .......................................................................................................... 120

Data Availability ..................................................................................................................... 122

Acknowledgments .................................................................................................................. 122

Author Contributions ............................................................................................................ 122

Declaration of Interests ........................................................................................................... 122

CHAPTER III: Transcriptome-wide cis-regulatory elements of normal termination codon readthrough in human cells ......................................................... 123

Abstract .................................................................................................................................. 124

Introduction .............................................................................................................................. 125

Results ...................................................................................................................................... 127

Random forest models identify mRNA features predictive of readthrough efficiency ................. 127
Known stop codon context influences readthrough efficiency .................. 130
Readthrough-promoting stop codon and nt +4 occur together less often than expected ................................................................. 133
P-site codons have mild effects on readthrough efficiency ....................... 137
Readthrough efficiency increases with 3'-UTR length ............................... 139
Random forest model can accurately predict outcomes of PTC readthrough measured by Dual-Luciferase assay in G418-treated cells .................................................. 141
Three codons flanking each side of a PTC comprise a sufficient context to experimentally measure readthrough efficiency .................. 145
Stop codon context in CFTR PTC alleles ................................................. 149
Mutations that result in UAA PTCs do not respond well to G418 treatment ................................................................. 152
Discussion ............................................................................................ 154
Near-cognate tRNA decoding in the A-site likely influences the efficiency of translation termination ................................................................. 154
Potential applications of machine learning models in readthrough prediction ........................................................................ 155
Materials and Methods ........................................................................ 158
Data acquisition ..................................................................................... 158
Random forest models and statistical analyses ........................................... 158
Generation of dual-luciferase hCFTR PTC reporters ............................... 160
Dual-luciferase reporter assay to measure PTC readthrough in HEK293 cells ........................................................................ 160
Data Availability .................................................................................... 161
Acknowledgments .................................................................................. 161
Author Contributions ............................................................................ 162
Declaration of Interests .......................................................................... 162
CHAPTER IV: Recognition of inefficient translation termination and NMD: Upf1’s action on ribosomes ....................................... 163
Abstract ............................................................................................... 164
Introduction ............................................................................................ 164
Results ..................................................................................................... 167
  Rationale for the methodological approaches ......................................... 167
  Purification of FLAG-Upf1- and Upf1-FLAG-associated ribosomes yields stoichiometric recovery of ribosomal proteins ......................... 175
  Upf1 association with 80S ribosomes in CHX-treated cells promotes the formation of atypical ribosome-protected mRNA fragments .... 181
  Upf1 function but not full activity of the NMD pathway is required for formation of atypical footprints by Upf1-associated ribosomes .......... 185
L footprints are generated by additional protection of mRNA 5’ to the normal ribosome-protected segment ................................................................. 186
L footprints reflect an early phase of Upf1 association with polyribosomes ........................................................................................................ 188
Analysis of mRNAs enriched in Upf1-associated ribosomes supports Upf1’s role in mRNA surveillance ............................................................. 193
Ribosomes associated with upf1DE572AA accumulate downstream of normal termination codons ............................................................... 198

Discussion ................................................................................................................................. 207
Upf1 functions while associated with ribosomes ................................................................. 207
The Upf1:ribosome complex detected in CHX-treated cells comprises a step that precedes commitment to NMD ........................................ 208
ATPase-deficient Upf1 interferes with normal translation termination ......................... 210
Upf1 functions in multiple complexes .................................................................................. 211

Materials and Methods ......................................................................................................... 213
Yeast Strains .......................................................................................................................... 213
Oligonucleotides ................................................................................................................... 213
Plasmids ............................................................................................................................... 214
Polyribosome analysis, protein detection, and quantitation ............................................. 218
Ribosome purification ......................................................................................................... 219
Negative-stain transmission electron microscopy ............................................................ 219
Immunopurification .............................................................................................................. 219
Mass spectrometry ............................................................................................................. 220
mRNA Decay Analysis ....................................................................................................... 224
RNA-Seq and Ribosome Profiling Library Preparation from Cells Expressing N-terminal FLAG-tagged Upf1 Protein ........................................ 224
RNA-Seq and Ribosome Profiling Library Preparation from Cells Expressing C-terminal FLAG-tagged Upf1 Protein ...................................... 225
High-throughput Sequencing of RNA-Seq and Ribosome Profiling Libraries ....................... 227
Sequence Alignment, Transcript Quantification, and Data Analyses ................................. 228
Analyses of Changes in Transcript Abundance ................................................................. 233
Analysis of A-site codon occupancy .................................................................................. 233
Codon optimality .................................................................................................................. 234
Statistical analyses and data visualization ......................................................................... 235

Data Availability ....................................................................................................................... 235

Acknowledgments .................................................................................................................. 236
Author Contributions ............................................................................................................. 236
Declaration of Interests ............................................................................................................ 236

Chapter V: Discussion .............................................................................................................. 237
## List of Tables

### Chapter I

**Table 1.1.** List of trans-acting factors that have been implicated in the global regulation of termination and/or readthrough ........................................... 22

### Chapter II

**Table 2.1.** Ribosome profiling data and yeast strains ............................................ 77  
**Table 2.2.** Assessed X variables and their modes of measurement ...................... 88  
**Table 2.3.** Number of genes involved in statistical analyses ................................. 92  
**Table 2.4.** Comparison of nt -3 to +9 effects (based on Figures 2.8 and 2.9) to previous data ................................................................................... 100  
**Table 2.5.** Ribosome profiling reads processing and alignment statistics ............. 118  
**Table 2.6.** Offset from 5’ and 3’ ends of footprint to the first nucleotide in the P-site for each footprint length in each sample .......................................... 119

### Chapter IV

**Table 4.1.** Comparison of M vs L footprints abundance 5’ and 3’ of the mRNA coding region (CDS) by Fisher’s exact test .............................................. 193  
**Table 4.2.** Yeast strains used in this study ............................................................. 213  
**Table 4.3.** Oligonucleotides used in this study .................................................... 213  
**Table 4.4.** Plasmids used in this study ................................................................. 214

### Appendix I

**Table A1.1.** Statistical test results of pairwise comparison between Reference, Up, and Down TE groups ................................................................. 289  
**Table A1.2.** Yeast strains used in this study ......................................................... 306  
**Table A1.3.** Oligonucleotides used in this study .................................................. 306
Table A1.4. Ribosome profiling reads processing and alignment statistics........312
**List of Figures**

**Chapter II**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page number</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Reproducibility between replicates of sup45 temperature-shift ribosome profiling and RNA-Seq datasets</td>
<td>76</td>
</tr>
<tr>
<td>2.2</td>
<td>Ribosome occupancy at and beyond the stop codon increases in yeast cells defective in translation termination or ribosome recycling</td>
<td>79</td>
</tr>
<tr>
<td>2.3</td>
<td>Evidence for readthrough of premature termination codons (PTCs) upon eRF1 inactivation</td>
<td>81</td>
</tr>
<tr>
<td>2.4</td>
<td>Readthrough events were characterized by maintenance of reading frame when ribosomes bypass the stop codon</td>
<td>84</td>
</tr>
<tr>
<td>2.5</td>
<td>Definition of X and Y Variables</td>
<td>87</td>
</tr>
<tr>
<td>2.6</td>
<td>Correlation matrix of relationships between readthrough efficiency, gene expression, translation efficiency, codon optimality, and transcript length</td>
<td>90</td>
</tr>
<tr>
<td>2.7</td>
<td>Random forest identified stop codon identity, P-site amino acid identity, and 3'-UTR length as factors critical in readthrough efficiency prediction</td>
<td>94</td>
</tr>
<tr>
<td>2.8</td>
<td>Known individual effects of stop codon and surrounding nucleotide identities on readthrough efficiency occur at a genome-wide level</td>
<td>97</td>
</tr>
<tr>
<td>2.9</td>
<td>Readthrough-permissive combinations of stop codon and nt +4 identities have additive effects on readthrough efficiency and are avoided in the genome</td>
<td>102</td>
</tr>
<tr>
<td>2.10</td>
<td>Specific codons in the P site were associated with higher or lower readthrough efficiencies</td>
<td>106</td>
</tr>
<tr>
<td>2.11</td>
<td>Readthrough efficiency increased with 3'-UTR lengths in eRF1 mutants, but not in wild-type or ribosome recycling factor mutant cells</td>
<td>108</td>
</tr>
</tbody>
</table>
Chapter III

Figure 3.1. mRNA features predictive of readthrough efficiency..........................129

Figure 3.2. Effects of the stop codon and flanking nucleotides identities on readthrough efficiency.................................................................132

Figure 3.3. Frequency of stop codon and nt +4 in each sample.........................136

Figure 3.4. Effects of triplet codons in the ribosomal P-site on readthrough efficiency............................................................................................138

Figure 3.5. Readthrough efficiency increases with 3’-UTR length.......................140

Figure 3.6. Schematic of dual-luciferase (Dual-Luc) reporter used to measure readthrough of CFTR PTC allele.................................................................142

Figure 3.7. Random forest model can accurately predict readthrough of CFTR PTC alleles in G418-treated cells.................................................................144

Figure 3.8. Extended stop codon context is sufficient to accurately predict readthrough efficiency.............................................................................................147

Figure 3.9. Stop codon context of CFTR PTC alleles.........................................151

Figure 3.10. UAA PTC alleles do not respond well to G418-treatment..............153

Chapter IV

Figure 4.1. Electron micrographs of RNase I-digested 80S ribosomes..............170

Figure 4.2. Polysomal distributions of wild-type and mutant forms of Upf1.......171

Figure 4.3. NMD phenotypes of UPF1 mutants..............................................173

Figure 4.4. NMD substrates were stabilized in episomally expressed FLAG-tagged UPF1 strains.......................................................................................174

Figure 4.5. Stoichiometric recovery of FLAG-tagged Upf1 with ribosomal proteins.................................................................................................178

Figure 4.6. Stoichiometric recovery of FLAG-tagged Upf1 with ribosomal proteins.................................................................................................180
Figure 4.7. Immunopurification of ribosomes associated with FLAG-tagged Upf1.................................................................................................................. 180

Figure 4.8. Distribution of footprint length for each strain and cycloheximide treatment condition............................................................. 183

Figure 4.9. Abundance of ribosome footprints relative to canonical start or stop codons for immunopurified (IP) or total ribosomes (total)............. 184

Figure 4.10. Atypically large footprints are attributable to a 5' extension of the mRNA region protected by the ribosome.............................. 187

Figure 4.11. Upf1 progressively associates with ribosomes across mRNA coding regions except when Upf1 association forms L footprints......... 190

Figure 4.12. Upf1 progressively associates with ribosomes across mRNA coding regions except when Upf1 association forms L footprints........... 192

Figure 4.13. Characteristics of mRNAs enriched or depleted for Upf1-associated ribosomes.......................................................................... 197

Figure 4.14. Codon occupancy for individual codon identity in the A-site and its surrounding codon positions in C-terminally FLAG-tagged UPF1 libraries..... 202

Figure 4.15. The relationship between codon optimality and ribosome occupancy at a codon is disrupted in the presence of cycloheximide........... 203

Figure 4.16. Percentage of footprints in the 3'-UTR region............................................... 204

Figure 4.17. Fractions of footprints in ribosome profiling libraries in particular mRNA regions or reading frames.................................................. 205

Figure 4.18. Reproducibility between replicates of ribosome profiling libraries.. 230

Figure 4.19. Reproducibility between replicates of RNA-Seq libraries................. 232

Chapter V

Figure 5.1. Cis-acting elements modulating readthrough efficiency are mostly conserved between yeast and human........................................ 239

Figure 5.2. Yeast vs. human 3'-UTR lengths...................................................................... 240

Figure 5.3. Examples of machine learning models.............................................................. 246
Appendix I

Figure A1.1. Replicate reproducibilities ................................................................. 259

Figure A1.2. PAB1 gene deletion leads to ribosome accumulation at mRNA start codons and in UTR regions ................................................................. 263

Figure A1.3. PAB1 gene deletion has significant effects on the transcriptome and the proteome ................................................................................................. 266

Figure A1.4. PAB1 gene deletion changes mRNA and protein levels of many components of the translation machinery ................................................................. 270

Figure A1.5. Pab1 regulates readthrough efficiency in a distance-dependent manner ................................................................................................................. 273

Figure A1.6. Translation initiation defects in response to PAB1 deletion may be related to reduction in eIF4G level ........................................................................ 276

Figure A1.7. 5'-UTR features of mRNAs with differential changes in translation efficiency ........................................................................................................ 281

Figure A1.8. Influences of start codon context in changes of translation efficiency (TE) .............................................................................................................. 283

Figure A1.9. Evidence for communication of mRNA 5'-3' ends in efficient initiation .................................................................................................................. 288

Figure A1.10. Changes in the abundance of decapping activator substrates by different consequences of PAB1 deletion ............................................................... 294

Figure A1.11. mRNA features determining AUG occupancy in different yeast strains .................................................................................................................. 302

Figure A1.12. Nucleotides determinant of AUG occupancy reside in the ribosome’s mRNA channel ......................................................................................... 303

Figure A1.13. The inverse relationship between AUG occupancy and mRNA length is abolished in the absence of Pab1 ...................................................................... 305

Appendix II

Figure A2.1. Ribosome profiling library preparation workflow .................................. 324
Figure A2.2. Ribosome profiling library alignment statistics..........................327

Figure A2.3. Relationships between number of PCR cycles and percentage of reads at each sequence alignment processing step.................330
List of Copyrighted Materials Produced by Author


List of Multimedia Files

*Interactive Figure A1.1.* is available for download at [https://github.com/Jacobson-Lab/Pab1_deletion/blob/main/pab1_mass_spec_gene_ontology_pab1d-vs-pbp1d.html](https://github.com/Jacobson-Lab/Pab1_deletion/blob/main/pab1_mass_spec_gene_ontology_pab1d-vs-pbp1d.html). Downloaded file can be opened in any internet browser except for Safari.
CHAPTER I: Introduction

Overview

Translation is the process by which genetic information encoded in mRNA is converted to protein by the ribosome in four stages: initiation, elongation, termination, and ribosome recycling (Schuller and Green 2018; Hellen 2018). Translation termination occurs when the ribosome encounters a stop codon (UAA, UAG, or UGA) and the growing nascent polypeptide chain is released from the translation complex, ensuring the correct synthesis of protein. However, prematurely-occurring termination is undesirable. Approximately 10-11% of all human genetic disease cases arise from nonsense mutations, which introduce a premature termination codon (PTC) within a gene’s coding region (Mort et al. 2008; Benhabiles et al. 2016). As a result, the protein product is truncated and often destroyed, and the associated mRNA is degraded by a quality control process called the nonsense-mediated mRNA decay (NMD) pathway. To treat diseases caused by nonsense mutations, attempts have been made to induce bypass of the PTC during translation and antagonize NMD. However, a deeper understanding of both processes is needed to enhance therapeutic benefits and safety. In this introduction, I summarize the current understanding of the mechanisms and regulation of translation termination and NMD, point out areas where much more information is required, and discuss challenges for current therapeutic strategies.
Translation termination

Termination occurs when the ribosome encounters a stop codon in its aminoacyl transfer RNA (tRNA) binding site (A site) and, in eukaryotes, is carried out by two release factors, eRF1 and eRF3 (Frolova et al. 1994; Stansfield et al. 1995; Zhouravleva et al. 1995; Frolova et al. 1996; Salas-Marco and Bedwell 2004; Alkalaeva et al. 2006). In the budding yeast Saccharomyces cerevisiae, eRF1 and eRF3 are encoded by the SUP45 and SUP35 genes, respectively (Kushnirov et al. 1988; Stansfield et al. 1995). In mammals, there are two isoforms of eRF3, eRF3a and eRF3b, where the latter is ubiquitously expressed (Hoshino et al. 1998; Chauvin et al. 2005; Hellen 2018).

The release factors: eRF1 and eRF3

The domains and functions of the release factors were largely discovered and described by genetics studies and in vitro reconstitution and translation systems. Precise mechanistic details of these factors have been elucidated by structural biology approaches, including X-ray crystallography and cryo-electron microscopy (cryo-EM). Structural resemblance between eRF1 and tRNA molecules (Song et al. 2000; Cheng et al. 2009; Brown et al. 2015), as well as homology between major parts of eRF3 and elongation factors eEF1A/EF-Tu (Kong et al. 2004; Cheng et al. 2009), indicated that their respective functions likely mimic the molecular mechanisms of translation elongation (Schuller and Green 2018). eRF1 consists of three distinct domains. The amino-terminal (N) domain, corresponding to a tRNA’s anticodon loop, contains the highly conserved NIKS
motif as well as YxCxxxF and GTS motifs involved in stop codon recognition (Bertram et al. 2000; Song et al. 2000; Chavatte et al. 2002; Frolova et al. 2002; Bulygin et al. 2010; Wong et al. 2012; des Georges et al. 2014; Blanchet et al. 2015; Brown et al. 2015). The middle (M) domain contains the conserved GGQ motif responsible for nascent peptide release from the peptidyl-tRNA (Frolova et al. 1999; Song et al. 2000; Shao et al. 2016). And the carboxy-terminal (C) domain contains binding sites for eRF3, the ribosome recycling factor ABCE1 (yeast Rli1), and an NMD factor, Upf1 (Ito et al. 1998; Merkulova et al. 1999; Frolova et al. 2000; Song et al. 2000; Akhmaloka et al. 2008; Ivanov et al. 2008; Mantsyzov et al. 2010; Preis et al. 2014; Shao et al. 2016). eRF3 consists of N and M domains that are neither well conserved nor functionally essential for termination, but contain a binding site for poly(A)-binding protein (PABP) (Hoshino et al. 1999a; Uchida et al. 2002; Kononenko et al. 2010; Kozlov and Gehring 2010; Roque et al. 2015). In yeast, the N-M domains are involved in the self-aggregation and prion properties of the yeast [PSI+] state (Ter-Avanesyan et al. 1993; Paushkin et al. 1996). The conserved C-terminal region consists of the GTP-binding (G) domain and two β-barrel domains (domain 2 and 3) containing an eRF1 binding site (Kong et al. 2004; Cheng et al. 2009; des Georges et al. 2014) as well as a Upf1 binding site (Ivanov et al. 2008).

**Mechanisms of stop codon recognition and peptide release**

eRF1 and eRF3 have been shown to interact and form a complex both with and without the ribosome in structural studies (Cheng et al. 2009; Taylor et al.
The latter results, along with previous biochemical studies showing that stable eRF3-GTP binding requires eRF1-eRF3 interaction (Hauryliuk et al. 2006; Mitkevich et al. 2006; Pisareva et al. 2006; Kononenko et al. 2007), support a model for a pre-termination step where release factors are thought to have formed an eRF1-eRF3-GTP ternary complex prior to arriving at the ribosomal A-site. A recent study employing single-molecule fluorescence spectroscopy also demonstrated that eRF1-eRF3-GTP preferentially binds the ribosome as a ternary complex, with a rate similar to that of aminoacyl-tRNA binding to its cognate A-site codon during elongation, rather than as individual components (Lawson et al. 2021). This observation corroborates a previous study suggesting that eRF3 serves as an eRF1 chaperone, delivering eRF1 to the ribosome (Eyler et al. 2013). However, an alternative model of sequential assembly has been suggested for specific conditions such as stress response, where eRF3 and ABCE1 wait at the ribosome for Dbp5-mediated delivery of eRF1 (Beißel et al. 2019).

Stop codon recognition by eRF1 follows ribosome-binding by the eRF1-eRF3-GTP complex. Structural studies have shed light on how conserved motifs of eRF1’s N domain distinguish stop codons from other sense codons. The K$_{63}$ residue of the NIKS motif forms a hydrogen bond with uridine (U), which is not possible with other nucleotides due to either steric hindrance or inability to form such bonds, imposing the requirement for U as the first nucleotide of a codon (position +1) to be recognized by eRF1 (Chavatte et al. 2002; Brown et al. 2015;
Matheisl et al. 2015). The C\textsubscript{127} residue of YxCxxxxF forms two hydrogen bonds with the flipped-out A\textsubscript{1825} of 18S ribosomal RNA (rRNA), resulting in its stacking with the +2 and +3 bases – this configuration is only compatible if the +2 and +3 bases are purines, thus providing another requirement for stop codon recognition (Brown et al. 2015; Matheisl et al. 2015; Shao et al. 2016). Additionally, E\textsubscript{55} of eRF1 forms a hydrogen bond with purines at +2 and/or +3 positions, but not when there are two consecutive G’s (e.g., in the UGG codon) because their repulsion to each other and to E\textsubscript{55} inhibits the desired interactions. The T\textsubscript{32} residue of the GTS motif works cooperatively with the YxCxxxxF motif, adopting two different conformations for UAG and UGA. For UAG, T\textsubscript{32} forms a hydrogen bond with the +3 G. For UGA, a slight movement of the YxCxxxxF motif accommodates the +2 G, pushing away T\textsubscript{32} from the stop codon (Brown et al. 2015). Remarkably, the structures also revealed that the stacking of flipped-out A\textsubscript{1825} of 18S rRNA with the +2 and +3 bases results in mRNA compaction that allows the +4 base to enter the ribosomal A-site, making a +4 base stack with G\textsubscript{626} of 18S rRNA (Brown et al. 2015; Matheisl et al. 2015; Shao et al. 2016) and +5 base stack with C\textsubscript{1698} of 18S rRNA that protrudes into the mRNA channel (Shao et al. 2016). These stacking interactions are more stable with purines, which explains why stop codons followed by a purine at +4 and/or +5 positions have higher termination efficiency (see “Cis-regulatory elements of readthrough efficiency” below).

Upon stop codon recognition, eRF3 hydrolyzes GTP, inducing a conformational change in eRF1 and peptide hydrolysis activity (Salas-Marco and
Similar to the requirement for GTP binding, the GTPase activity of eRF3 is dependent on interactions with eRF1’s M-C domains as well as the ribosome (Frolova et al. 1996; Kononenko et al. 2007), most likely through the sarclin-ricin loop of the large ribosomal subunit (60S) like other translational GTPases (des Georges et al. 2014; Preis et al. 2014; Shao et al. 2016). Prior to GTP hydrolysis, eRF1 is in a pre-accommodated state where its GGQ motif is further away (>80 Å) from the P-site tRNA ester bond in the peptidyl transferase center (Shao et al. 2016). After GTP hydrolysis, the M-C domains of eRF1 are rotated, putting eRF1 in an extended, accommodated state and placing the GGQ motif at the peptidyl transferase center, with the Q_{185} residue close to the ester bond that links peptidyl tRNA to the nascent peptide (Preis et al. 2014; Brown et al. 2015; Matheisl et al. 2015; Shao et al. 2016). This configuration allows accommodation of a water molecule that hydrolyzes the ester bond and releases the nascent peptide, completing the translation termination process (Frolova et al. 1999; Schuller and Green 2018; Hellen 2018).

After GTP hydrolysis, eRF3 dissociates from the termination complex, and ABCE1 binds to eRF1 at the same site as eRF3 (Pisarev et al. 2010; Preis et al. 2014; Young et al. 2015; Shao et al. 2016; Schuller and Green 2018; Hellen 2018). Although ABCE1 is not required for termination, it has been shown to enhance the rate of peptide release (Khoshnevis et al. 2010; Shoemaker and Green 2011; Young and Guydosh 2022). The major role of ABCE1 is facilitation of post-termination ribosome recycling. Dissociation of the large ribosomal subunit (60S)
is mediated by ABCE1 (Rli1) and an accessory factor eIF3j (Hcr1) (Pisarev et al. 2010; Young et al. 2015; Young and Guydosh 2019), followed by dissociation of the small ribosomal subunit (40S) by eIF2D (Tma64) and the MCT-1 (Tma20) / DENR (Tma22) heterodimer (Young et al. 2018, 2021). Split ribosomal subunits enter the pool of free ribosomes available for new rounds of translation.

Stop codon readthrough

Definition and mechanism of stop codon readthrough

When a stop codon appears in the ribosomal A-site, eRF1 is in competition with tRNAs for A-site binding and stop codon recognition. Although eRF1 is highly efficient, a near-cognate tRNA whose anticodon matches with 2 out of 3 bases of the stop codon can sometimes outcompete eRF1 in stop codon decoding, resulting in an erroneous insertion of an amino acid and continuation of translation elongation until the next in-frame stop codon is encountered. This stop codon miscoding phenomenon is termed stop codon readthrough or nonsense suppression (Brenner et al. 1965; Dabrowski et al. 2015). At normal termination codons (NTCs) at the end of mRNA open reading frames (ORFs), the rate of stop codon misreading by a near-cognate tRNA is very low, estimated at 1 in 100,000 codons (Rodnina 2023).

The near-cognate tRNAs that decode stop codons have been systematically determined in yeast and human cells to be those with a codon-anticodon mismatch usually at either the 1st or 3rd base (Blanchet et al. 2014; Roy et al. 2015, 2016; Xue et al. 2017). The same amino acids are inserted at UAA and
UAG, although at different frequencies, while different amino acids are inserted at UGA. In yeast, UAA and UAG are decoded by tRNA\textsuperscript{Tyr} (codon UAU/UAC), tRNA\textsuperscript{Gln} (codon CAA/CAG), and tRNA\textsuperscript{Lys} (codon AAA/AAG), while UGA is decoded by tRNA\textsuperscript{Trp} (codon UGG), tRNA\textsuperscript{Arg} (codon CGA/AGA), and tRNA\textsuperscript{Cys} (codon UGU/UGC) (Blanchet et al. 2014; Roy et al. 2015, 2016). In humans, the amino acids identified are similar, with additional 2\textsuperscript{nd} base mismatches occurring occasionally, including tRNA\textsuperscript{Trp} (codon UGG) decoding UAG and tRNA\textsuperscript{Leu} (codon UUA) decoding UGA (Roy et al. 2016; Xue et al. 2017). The choice of the tRNA decoding a stop codon can be influenced by: i) readthrough-promoting aminoglycoside treatment as evidenced by differences in amino acid or tRNA identities/frequencies observed in the same construct in untreated vs. treated cells (Roy et al. 2015, 2016; Beznosková et al. 2019), ii) identities of nucleotides surrounding the stop codon, as evidenced by differences in amino acid identities/frequencies observed in different constructs that had the same stop codon (Blanchet et al. 2014; Xue et al. 2017), and iii) tRNA modifications, as evidenced by differences in amino acid frequencies upon deletion of a gene encoding a tRNA modification enzyme, making another tRNA more competitive (Blanchet et al. 2018; Beznosková et al. 2019). The latter study in yeast also elucidated the mechanism by which unconventional base pairing (mismatch) between the tRNA anticodon and the stop codon can occur due to specific modifications of the tRNA anticodon loop and adjacent nucleotides, such as
pseudouridylation of the tRNA$^{\text{Tyr}}$ anticodon, which does not affect decoding of tyrosine sense codons (Blanchet et al. 2018).

Stop codons can also be decoded by cognate tRNAs. Nonsense suppressor mutations have been identified in yeast and E. coli, e.g., a mutation that alters the anticodon of a normal tRNA$^{\text{Tyr}}$ to render it cognate to the UAA stop codon, resulting in high readthrough at a transcriptome-wide level (Goodman et al. 1968, 1977; Pierce et al. 1987; Guydosh and Green 2014; Young et al. 2015). There are also specialized cognate tRNAs that carry non-canonical albeit natural amino acids (Rodnina et al. 2020; Rodnina 2023). UGA recoding by a cognate tRNA$^{\text{Sec}}$ carrying selenocysteine can occur in all domains of life and is regulated by the existence of a selenocysteine insertion sequence (SECIS) element in the mRNA 3'-UTR that forms a stem-loop structure (Labunskyy et al. 2014). UAG recoding by a cognate tRNA$^{\text{Pyl}}$ carrying pyrrolysine only occurs in methanogenic archaea (Rodnina 2023).

Since the latter special cases of readthrough have their own dedicated biosynthetic pathways and regulation, which have been described extensively (Fekner and Chan 2011; Labunskyy et al. 2014; Wan et al. 2014), I only focus here on the current understanding of stop codon readthrough by near-cognate tRNA miscoding.

**Evolution of stop codon readthrough**

Two competing hypotheses consider the evolution of stop codon readthrough: the adaptive hypothesis posits that readthrough is utilized in a regulated way to expand proteome diversity, while the error hypothesis proposes
that most readthrough events are nonadaptive molecular errors and cells may evolve to avoid them (Li and Zhang 2019; Zhang and Xu 2022). In support of the adaptive hypothesis, viruses have long been known to exploit readthrough to increase the protein-coding potential of their compact genomes, occasionally producing polyproteins (Pelham 1978; Strauss et al. 1983; Skuzeski et al. 1991; Wills et al. 1991; Li and Rice 1993; Schueren and Thoms 2016). Readthrough in eukaryotic cells is more recently described. Natural readthrough products, namely C-terminally extended isoforms of proteins, are more prevalent than previously anticipated (Rodnina 2023), as their discovery has been facilitated by an increasing availability of genomics data and genome-wide experimental approaches, such as ribosome profiling, mass spectrometry, and phylogenetic analyses (Schueren and Thoms 2016). Experimental evidence for readthrough or phylogenetic evidence for potential readthrough of normal protein-coding mRNAs exist for yeast (Saccharomyces cerevisiae) (Namy et al. 2002; Namy 2003; Williams 2004; Kleppe and Bornberg-Bauer 2018), ciliates (Heaphy et al. 2016; Swart et al. 2016), plants (Nyikó et al. 2017; Sahoo et al. 2022), fruit flies (Drosophila melanogaster) (Steneberg et al. 1998; Steneberg and Samakovlis 2001; Jungreis et al. 2011; Dunn et al. 2013; Karki et al. 2022), frogs, and mammals, including humans (Chittum et al. 1998; Ingolia et al. 2011; Jungreis et al. 2011; Eswarappa et al. 2014; Loughran et al. 2014; Schueren et al. 2014; Stiebler et al. 2014; Hofhuis et al. 2016; De Bellis et al. 2017; Loughran et al. 2018; Yamaguchi and Baba 2018; Yordanova et al. 2018; Rajput et al. 2019; Sapkota et
C-terminally extended proteins can change protein function, localization, or expression. For example: i) the readthrough isoform of \( VEGFA \) diminishes its angiogenesis function compared to the canonical non-readthrough isoform (Eswarappa et al. 2014), ii) readthrough isoforms of \( LDHB \) and \( MDH1 \) gain a peroxisomal localization signal in their C-terminal extensions, changing their cellular distributions (Schueren et al. 2014; Stiebler et al. 2014; Hofhuis et al. 2016), and iii) the double-readthrough isoform of \( MTCH2 \) changed both the localization (from mitochondria to cytoplasm) and the stability of the respective proteins (Manjunath et al. 2020). Such functional and regulatory consequences of readthrough of mRNAs encoded by mammalian genes have been reviewed recently (Manjunath et al. 2022).

In addition to established functional readthrough, computational analyses of C-terminally extended regions of readthrough candidate genes revealed that these regions are more conserved than their downstream counterparts (Jungreis et al. 2011; Dunn et al. 2013; Jungreis et al. 2016), possess higher intrinsic structural order (Kosinski and Masel 2020), contain more binding motifs (in the case of structural disorder) (Pancsa et al. 2016), and have stronger bias in synonymous single nucleotide variants compared to the same-sized non-coding controls (Jungreis et al. 2016). These results suggest that readthrough is allowed to happen to facilitate protein evolution, and the resulting product is under either positive selection to gain a new functional/regulatory role or purifying selection to limit deleterious consequences.
Deleterious, nonadaptive consequences of readthrough that is expected of most stochastic stop codon miscoding is the basis for the error hypothesis, which asserts that the readthrough phenomenon itself would be selected against in order to limit cytotoxicity of readthrough products (Li and Zhang 2019). Several lines of evidence support this hypothesis. First, C-terminally extended proteins have been shown to have reduced expression and be degraded by the ubiquitin-proteasome system, as well as targeted to the lysosome, the cellular pathways that prevent accumulation of toxic, faulty proteins and control protein expression (Shibata et al. 2015; Arribere et al. 2016; Kramarski and Arbely 2020). Second, although C-terminal extension regions are more conserved than their downstream counterparts in readthrough candidate genes, the extent of this conservation is no different from the same regions examined in non-readthrough candidate genes of similar expression level, suggesting that the conservation is likely unrelated to readthrough (Li and Zhang 2019). Third, RNA-Seq and ribosome profiling experiments revealed that gene expression is inversely correlated with readthrough efficiency (Kleppe and Bornberg-Bauer 2018; Li and Zhang 2019), meaning that highly expressed genes have evolved to minimize readthrough for several reasons: i) since genes important to cellular fitness are usually highly expressed, the effects of their readthrough products are potentially more harmful than readthrough of genes expressed at low levels and ii) readthrough of highly expressed genes would result in a higher number of readthrough products compared to poorly expressed genes of the same readthrough rate, potentially
overwhelming the protein degradation pathways and wasting cellular energy (Belinky et al. 2018; Li and Zhang 2019; Zhang and Xu 2022). If readthrough is largely adaptive, these scenarios would not apply, and the inverse relationship between gene expression and readthrough efficiency would not be expected (Li and Zhang 2019). Lastly, readthrough-inhibiting sequence motifs are over-represented while readthrough-promoting motifs under-represented in highly expressed genes (Bonetti et al. 1995; McCaughan et al. 1995; Trotta 2016; Belinky et al. 2018; Li and Zhang 2019). The identities of these motifs are discussed in detail below (see “*Cis*-regulatory elements of readthrough efficiency”).

Ultimately, the need for both adaptiveness and error minimization may be at play in the evolution of stop codon readthrough. Although UAA, the stop codon with the highest fidelity, is concentrated among mRNAs of highly expressed and essential genes, UGA, the most leaky stop codon especially when followed by a pyrimidine, is still the most common stop codon in the human genome (Trotta 2016). Collectively, these studies indicate that cells have evolved to limit deleterious consequences of readthrough while also allowing readthrough to happen at a sufficient level to facilitate development of programmed readthrough for beneficial C-terminal extensions.

*Cis*-regulatory elements of readthrough efficiency

Studies of termination generally involve reporter assays that rely on the inverse relationship between termination and readthrough efficiencies, where readthrough products are measured to infer termination efficiency. Different
mRNAs result in varying levels of readthrough products, with readthrough ranging from very low/almost undetectable to programmed readthrough rates of 0.1-0.3 relative to expression of the upstream coding region (Rodnina et al. 2020). This observation, along with the >100-fold differences in readthrough among cellular mRNAs in ribosome profiling data of various organisms (Dunn et al. 2013; Baudin-Baillieu et al. 2014; Wangen and Green 2020), indicate that the efficiency of readthrough, and by extension the efficiency of termination, is subjected to mRNA sequence-specific regulation.

Stop codons are not created equal

Experiments comparing readthrough of different stop codons in identical reporters established that readthrough efficiency differs among the three stop codons, with UGA allowing the highest readthrough, followed by UAG and UAA (Bonetti et al. 1995; Howard et al. 2000; Manuvakhova et al. 2000; Floquet et al. 2012; Loughran et al. 2014; Cridge et al. 2018). Consistent with these results, the majority of the identified mammalian readthrough genes harbor UGA stop codons in their mRNAs, only one has UAG (Yamaguchi and Baba 2018), and none has UAA. Importantly, the trend for UGA > UAG > UAA regarding their readthrough propensity has been confirmed for endogenous mRNAs in analyses of ribosome profiling of a human cell line (Wangen and Green 2020). The basis for UAA’s high termination fidelity may reflect two possible hydrogen bonding interactions between E55 of eRF1 and both A’s at the +2 or +3 nucleotide positions, as opposed to one A at either position for UAG and UGA (Brown et al. 2015).
Nucleotides immediately upstream and downstream of the stop codon

Stop codon identities cannot sufficiently explain variations in readthrough efficiency. In fact, the idea that the stop “signal” is a tetranucleotide instead of a triplet codon was proposed very early as a consequence of experiments demonstrating that the nucleotide immediately after the stop codon, designated nt +4, can significantly change readthrough efficiency, as well as computational analyses showing that genomes are biased for certain tetranucleotides that are strong terminators (Brown et al. 1990; McCaughan et al. 1995). Although there are slight variations between studies, the general trend and consensus for the most to least readthrough permissive nt +4 is C > U > G > A (Rodnina et al. 2020). These results are consistent with structural studies revealing that the stacking of the +4 base with G_{626} of 18S rRNA is more stable when the base is a purine (A and G) (Brown et al. 2015; Matheisl et al. 2015; Shao et al. 2016), making termination more efficient.

Additional nucleotides downstream of the stop codon, particularly nt +5 to +9, have also been examined for their influence on readthrough efficiency in various organisms and experimental systems. Such studies have examined the effects of reporter point mutations in vivo, in vitro translation screens, and in silico analyses of genomes (Skuzeski et al. 1991; Cassan and Rousset 2001; Namy et al. 2001; Williams 2004; Pacho et al. 2011; Floquet et al. 2012; Loughran et al. 2014; Schueren et al. 2014; Cridge et al. 2018; Tate et al. 2018; Anzalone et al. 2019; Sokolova et al. 2020). Purines at position +5 are more stable stacking with
C\textsubscript{1698} of 18S rRNA protruding into the mRNA channel (Shao et al. 2016), in line with experiments demonstrating that A or G at +5 decreased readthrough and increased termination efficiency (McCaughan et al. 1995; Schueren et al. 2014). However, numerous experiments also showed that A (or U) at +5 is readthrough-promoting compared to other nucleotides, and this has become the current consensus (Namy et al. 2001; Loughran et al. 2014; Anzalone et al. 2019; Rodnina et al. 2020). Additionally, a CUAG context for nt +4 to +7 exists in half (7 out of 14) of the documented mRNAs of mammalian readthrough genes (Manjunath et al. 2022) and is reflected in regression models of multiple reporter contexts (Schueren et al. 2014). Differences in results are likely due to different reporter sequences used, which indicate how the effect of individual nucleotides is not simply additive but indicate complex interactions among nucleotide context positions. Although the exact mechanism of downstream nucleotides +5 to +9 in the regulation of termination is still unclear, it is possible that these nucleotides interact with the rRNAs or ribosomal proteins because they appear to reside in the ribosome's mRNA channel. The latter conclusion follows from their protection from RNASel digestion in ribosome profiling experiments as well as inferences of structural studies (Ingolia et al. 2009; Jenner et al. 2010; Ben-Shem et al. 2011; Cridge et al. 2018; Tate et al. 2018). A recent study using a reconstituted translation system further demonstrated that readthrough efficiency varied in a context-dependent manner in the absence of eRF1 while termination efficiency did not differ between standard and “weak” context, leading to a proposition that downstream nucleotides
affect the rate of near-cognate tRNA incorporation (rather than eRF1’s stop codon recognition or peptide release activity) through the involvement of ribosomes and other factors usually engaged in translation elongation (Biziaev et al. 2022).

In contrast to nucleotides downstream of the stop codon, nucleotides upstream of the stop codon have been explored less, results have been more convoluted, and mechanisms more difficult to determine because mutating a nucleotide could also change the tRNA decoding a codon and/or the amino acid residing in the ribosome’s exit tunnel. Up to 18 upstream nucleotides have been examined experimentally, with the closest six corresponding to nucleotides in the ribosomal P- and E-sites having the most data (Mottagui-Tabar et al. 1998; Cassan and Rousset 2001; Tork et al. 2004; Pacho et al. 2011; Floquet et al. 2012; Schueren et al. 2014; Loughran et al. 2023). An early study testing these six nucleotides as two triplets suggested that the amino acid identity corresponding to codon -2 (E-site) affected readthrough, with acidic residues favoring termination, while the tRNA was responsible for the effect of codon -1 (P-site) on readthrough, as different codons encoding the same amino acid resulted in varied readthrough efficiency (Mottagui-Tabar et al. 1998). However, other studies found no relation between readthrough efficiency and amino acid or tRNA, but A at nt -1 and/or -2 immediately upstream of the stop codon were found to be readthrough-promoting (Cassan and Rousset 2001; Tork et al. 2004). Another study did not find nt -1 and -2 to be important for basal readthrough, but instead nt -6 (Floquet et al. 2012). On the other hand, a recent study found both A at nt -1 and the P-site tRNA to be
important in readthrough efficiency differences between two genes with the same proximal downstream context (Loughran et al. 2023). The disparity in results again is likely due to differences in surrounding sequences used in different studies and reasserts that multiple nucleotide positions together may affect readthrough in a way that is not simply due to additive effects of individual positions.

Although systematic mutagenesis of upstream nucleotides that are inside the ribosome’s mRNA channel have been fully covered across multiple studies, upstream nucleotides representative of the nascent peptide in the ribosome’s exit tunnel have not. At least 30 amino acids, equivalent to 90 upstream nucleotides, can fit into the exit tunnel (Wilson et al. 2016). In human cytomegalovirus, the expression of gp48/UL4 mRNA is inhibited by translation of a 22-codon upstream ORF called uORF2. Interaction of the nascent peptide product of uORF2 with the exit tunnel perturbs the peptidyl transferase center, inhibits eRF1’s peptide hydrolysis activity, induces ribosome stalling, and prevents translation of the downstream gp48/UL4 (Janzen et al. 2002; Matheisl et al. 2015). This type of termination regulation has not been reported in mammalian cells. However, given that ribosomes can stall during elongation, e.g., as a consequence of inhibitory codon pairs and consecutive prolines (Gamble et al. 2016; Schuller and Green 2018), stalling peptide sequences for termination may yet be described.

**mRNA secondary structure in the 3’-UTR**

For specific mRNA readthrough, mRNA secondary structures formed in the mRNA’s 3’-UTR region downstream of the stop codon have been shown to
increase readthrough efficiency in various species. These include a pseudoknot downstream of gag mRNA in Moloney murine leukemia virus (Wills et al. 1991), a hairpin loop 3 nt downstream of hdc mRNA in Drosophila (Steneberg and Samakovlis 2001), and stem loops in the VEGFA mRNA in human cells (Eswarappa et al. 2014; Wagner 2019; Manjunath et al. 2022). Experimental evidence for the involvement of mRNA structure was generally obtained by mutating nucleotides that were predicted as key contributors to the structure, such as at the stem, and then monitoring changes in measured readthrough efficiency. Additionally, computational analyses predicted stable RNA structures downstream of stop codons of readthrough candidate genes, some of which are conserved across closely-related species (Firth et al. 2011; Jungreis et al. 2011), as well as in top-ranked readthrough constructs from an in vitro translation screen (Anzalone et al. 2019). Proving that these structures actually form in vivo is difficult, as the formation of mRNA secondary structure is dependent on various environmental factors, e.g., salt concentration (Firth et al. 2011); however, this dependency on the right environment means that the structure can serve a regulatory role controlling readthrough efficiency in different environmental conditions, providing functional readthrough product when needed. Although direct evidence for these structures’ precise mechanisms is limited, current mechanistic explanations for their readthrough enhancement property include induction of ribosome stalling or conformational rearrangement that disfavors termination, physical hindrance of release factor access to the ribosomal A-site, and recruitment of a trans-acting
factor that promotes readthrough (Firth et al. 2011; Eswarappa et al. 2014; Rodnina et al. 2020).

**Trans-regulatory elements of readthrough efficiency**

Termination and readthrough efficiencies change when the levels of any of the key components in the termination process change. Readthrough efficiency has been shown to increase in cells with overexpression of near-cognate tRNAs (Beznosková et al. 2021) or depletion of functional eRF1 or eRF3 through genetic mutation (Stansfield et al. 1997; Roy et al. 2015), mRNA knockdown (Carnes 2003), protein degradation (Sharma et al. 2021a), or existence of prion-forming eRF3 that leads to aggregation of the release factors and inhibition of their functions (Baudin-Baillieu et al. 2014). In addition to these essential factors, many studies have identified other *trans*-acting factors that can modulate the efficiency of termination and readthrough globally or in an mRNA-specific manner.

**mRNA-specific trans-regulatory factors**

For a *trans*-regulatory element to be mRNA-specific, it must be linked to a specific *cis*-regulatory element of an mRNA. So far, there are only two documented cases of *trans*-acting factors in mRNA-specific programmed readthrough. The requirement for readthrough of VEGFA includes mRNA stem loop structures, which in turn interact with the RNA-binding protein hnRNPA2/B1 (Eswarappa et al. 2014; Wagner 2019). The involvement of hnRNPA2/B1 is supported by decreased readthrough efficiency observed when its binding site was mutated or when its expression was reduced via siRNA-mediated knock-down (Eswarappa et al.
Readthrough of AGO1 is controlled by the microRNA (miRNA) let-7a, which induced readthrough upon binding to a complementary sequence in the AGO1 3’-UTR (Singh et al. 2019). The involvement of let-7a is also supported by the decrease in readthrough efficiency observed when its binding site was mutated or moved or when a let-7a inhibitor impeded its function. Additional support for its role follows from the increased readthrough efficiency observed when let-7a was overexpressed (Singh et al. 2019). As many known readthrough mRNAs still lack known mechanisms (Manjunath et al. 2022), not to mention many more readthrough mRNAs that remain undiscovered, it is likely that there exist more mRNA-specific trans-acting regulators of readthrough that have yet to be identified.

**Global trans-regulatory factors**

For a trans-regulatory factor to regulate readthrough efficiency non-specifically, it must either interact or modify termination machineries (e.g., the release factors or the ribosome), associate with mRNAs in a non-sequence-specific manner, or bind specific sequences that exist ubiquitously in most mRNAs, such as the mRNA poly(A) tail. Currently implicated global (or apparently global) trans-acting factors and their effects on termination and/or readthrough are listed in Table 1.1.
<table>
<thead>
<tr>
<th>Protein (organism)</th>
<th>Main/Known function(s)</th>
<th>Effect(s) on termination / readthrough</th>
<th>Possible mechanism(s) of action</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCE1 (human) / Rli1 (yeast)</td>
<td>Ribosome recycling / subunits splitting</td>
<td>Promotes termination; increases rate of peptide release in vitro</td>
<td>Binds eRF1 after eRF3 dissociates, locking eRF1 in an accommodated state and promoting peptide release</td>
<td>(Khoshnevis et al. 2010; Shoemaker and Green 2011; Young and Guydosh 2022)</td>
</tr>
<tr>
<td>PABPC (human) / Pab1 (yeast)</td>
<td>Regulation of mRNA stability, translation initiation</td>
<td>Promotes termination; readthrough increases in <em>pab1Δ</em> cells (stop codon mid-ORF)</td>
<td>Likely release factor recruitment through interaction with eRF3</td>
<td>(Roque et al. 2015; Ivanov et al. 2016; Wu et al. 2020)</td>
</tr>
<tr>
<td>PAIP1, PAIP2 (human)</td>
<td>Interactors of PABPC, modulate PABPC's role in translation initiation</td>
<td>Promote readthrough; readthrough increases in the presence of PAIP1/PAIP2 in cell-free translation (stop mid-ORF)</td>
<td>Compete with eRF3 for PABPC binding</td>
<td>(Ivanov et al. 2019a)</td>
</tr>
<tr>
<td>DDX19 (human) / Dbp5 (yeast)</td>
<td>mRNA export through nuclear pore complex</td>
<td>Promotes termination; readthrough increases in Dbp5 mutant cells (dual reporter assay)</td>
<td>Delivers eRF1 to eRF3 and the ribosome, promoting efficient pre-termination complex formation</td>
<td>(Gross et al. 2007; Beißel et al. 2019; Mikhailova et al. 2017; Beißel et al. 2020)</td>
</tr>
<tr>
<td>Gle1</td>
<td>mRNA export through nuclear pore complex</td>
<td>Promotes termination; readthrough increases in Gle1 mutant cells (dual reporter assay, stop codon mid-ORF)</td>
<td>Activates Dbp5, which is involved in pre-termination complex formation</td>
<td>(Bolger et al. 2008; Alcázar-Román et al. 2010)</td>
</tr>
<tr>
<td>Pub1 (yeast)</td>
<td>Post-transcriptional regulation, stress granule formation, maintenance of tubulin cytoskeleton</td>
<td>Promotes readthrough; readthrough decreases for certain tetrancleotides in <em>pub1Δ [PSI+]</em> cells (Dual-Luc assay, stop codon mid-ORF)</td>
<td>Unclear, but likely through interaction with eRF3's prion domain; also prevents excessive/toxic eRF3 prion formation</td>
<td>(Urakov et al. 2017, 2018)</td>
</tr>
<tr>
<td>OGFOD1 (human) / Tpa1 (yeast)</td>
<td>Prolyl hydroxylase, translation</td>
<td>Promotes termination; readthrough</td>
<td>Hydroxylates Rps23, which is part of the</td>
<td>(Keeling et al. 2006; Henri et al. 2010)</td>
</tr>
<tr>
<td><strong>Rps23</strong></td>
<td>40S ribosomal subunit decoding center</td>
<td>Controls fidelity of termination; readthrough differs in cells with modified vs. unmodified Rps23 in a context-dependent manner (Dual-Luc assay, stop codon mid-ORF)</td>
<td>Likely eRF1 or near-cognate tRNA accommodation in the decoding center</td>
<td>(Loenarz et al. 2014)</td>
</tr>
<tr>
<td><strong>Ett1 (yeast)</strong></td>
<td>Unknown</td>
<td>Promotes termination; readthrough increases in ett1Δ cells (Dual-Luc assay, stop codon mid-ORF)</td>
<td>Unknown</td>
<td>(Henri et al. 2010)</td>
</tr>
<tr>
<td><strong>Caf1 (human) / Pop2 (yeast)</strong></td>
<td>Cytoplasmic deadenylation</td>
<td>Promotes termination; readthrough increases in pop2Δ cells (Dual-Luc assay)</td>
<td>Unknown</td>
<td>(Keeling et al. 2006)</td>
</tr>
<tr>
<td><strong>eIF3</strong></td>
<td>Translation initiation, pre-initiation complex, release of mRNA and tRNA from post-termination complex</td>
<td>Promotes readthrough; readthrough decreases in eIF3 subunit mutant cells (Dual-Luc assay)</td>
<td>Unclear, but likely through interaction with ribosomal protein Rps3, eRF1, and pre-termination complex, modulating precision of eRF1's stop codon recognition at the third wobble position</td>
<td>(Beznosková et al. 2013, 2015; Poncová et al. 2019)</td>
</tr>
<tr>
<td><strong>uS3 (human) / Rps3 (yeast)</strong></td>
<td>Translation, part of a ribosome</td>
<td>Controls fidelity of termination; different mutations have opposite readthrough phenotypes (Dual-Luc assay)</td>
<td>(See eIF3)</td>
<td>(Poncová et al. 2019)</td>
</tr>
</tbody>
</table>
### eIF5A
- Translation initiation, translation elongation of stalling sequences
- Promotes translation termination; increased accumulation of ribosomes (with nascent peptide) stalled at stop codons in eIF5A depleted cells (ribosome profiling, 5PSeq)
- Stimulates peptide hydrolysis activity of eRF1 *in vitro*  
  (Pelechano and Alepuz 2017; Schuller et al. 2017; Lawson et al. 2021)

### eEF3
- Translation elongation, catalyzes tRNA release from E-site
- Promotes termination; readthrough increases when eEF3 is removed from cell-free system (size-based assay, stop codon mid-ORF), but this is not seen *in vivo* (ribosome profiling)
- Unclear, but can stimulate peptide release and post-termination complex disassembly  
  (Kurata et al. 2013; Kasari et al. 2019; Kobayashi et al. 2023)

### PDCD4 (human)
- Tumor suppressor, cell cycle progression, transcription, translation initiation
- Promotes termination
- Stimulates termination complex formation and peptide release *in vitro*  
  (Shuvalova et al. 2021)

It is important to note that, depending on which step of termination it acts upon, the stop codon recognition step (eRF1/near-cognate tRNA incorporation) or the peptide release step, a *trans*-acting factor may be a regulator of both termination and readthrough, or only termination. For instance, ABCE1/Rli1 can promote termination by enhancing the rate of peptide release (Khoshnevis et al. 2010; Shoemaker and Green 2011), but its depletion does not lead to stop codon readthrough because stop codon recognition has already been completed, peptide release can still happen, but ribosome recycling fails, resulting in random re-initiation downstream of the stop codon (Young et al. 2015). Thus, ABCE1/Rli1 can be considered a *trans*-acting regulator of termination but not readthrough. In fact,
ABCE1/Rli1 and eIF3j/Hcr1 were initially thought to also regulate readthrough because their respective depletion induced signal from the downstream ORF in dual-ORF reporter assays, e.g., dual-luciferase (Dual-Luc) or lacZ-Luc (Khoshnevis et al. 2010; Beznosková et al. 2013), but later ribosome profiling experiments showed that they function in ribosome recycling because their depletion produced ribosome footprints in all reading frames as opposed to one dominant reading frame in-frame with the coding region (Young et al. 2015; Young and Guydosh 2019). Dual-ORF or bicistronic reporters, where a sequence containing the stop codon is inserted between two ORFs, have been the mainstay of readthrough research for a long time due to highly quantitative properties (high sensitivity and dynamic range), internal control for mRNA level and transcription level, and ease of use. Readthrough efficiency is usually defined as the ratio of downstream to upstream signals from the two ORFs. However, the read-outs for readthrough and re-initiation that happens to be in-frame with the upstream ORF are indistinguishable. Thus, the apparent roles of many trans-acting factors (as well as cis-acting factors) in readthrough regulation that has been derived from only dual-ORF reporter results must be interpreted with caution.

One trans-acting factor is of particular interest with regard to premature termination occurring in mRNAs derived from genes harboring nonsense mutations: cytoplasmic poly(A)-binding protein (PABP or PABPC, Pab1 in yeast). PABPC is an RNA binding protein that binds the mRNA poly(A) tail and has multiple roles in mRNA expression regulation, which have been extensively reviewed
Mammals have multiple isoforms of PABPCs, of which the most ubiquitous isoform is PABPC1, whereas yeast has only one PABPC, called Pab1 (Brambilla et al. 2019; Passmore and Coller 2022). In both yeast and higher eukaryotes, PABPC is known to regulate poly(A) tail length, which in turn controls mRNA stability, through its interaction with deadenylase complexes, and translation initiation through its interaction with initiation factor eIF4G. PABPC’s role in translation termination is thought to be through its interaction with eRF3.

PABPC’s conserved structure consists of four RNA recognition motif (RRM) domains, a proline-rich (P) linker domain, and a C-terminal mademoiselle (MLLE) domain that interacts with other proteins via their PABP-interacting motif 2 (PAM2) (Brambilla et al. 2019; Passmore and Coller 2022; Qi et al. 2022). In metazoans, PABPC’s C domain interacts with eRF3’s PAM2 motifs in the N domain (Hoshino et al. 1999a, 1999b; Cosson et al. 2002a; Kozlov et al. 2001; Uchida et al. 2002; Kononenko et al. 2010; Kozlov and Gehring 2010; Jerbi et al. 2016). In yeast, PABPC-eRF3 interaction involves broader regions of both proteins, as yeast eRF3 lacks the PAM2 motif; PABPC’s P-C domains interact with eRF3 N-M domains (Cosson et al. 2002b; Kozlov et al. 2002; Roque et al. 2015). The absence of PABPC or its eRF3-interacting domain has been shown to impact termination efficiency in vitro and in vivo. In a reconstituted mammalian in vitro translation system, PABPC promoted eRF3 binding to a pre-termination complex, stop codon recognition, and peptidyl-tRNA hydrolysis only with full-length eRF3, but not with
the C-terminal region of eRF3, which lacks a PABPC binding site (Ivanov et al. 2016). Termination efficiency was also reduced in vitro in the presence of PABP-interacting proteins PAIP1 and PAIP2 that compete with eRF3 for free PABPC binding (Ivanov et al. 2019a). Interestingly, deletion of yeast eRF3’s N-M domains or Pab1’s P-C domains decreased readthrough efficiency in vivo as measured by Dual-Luc reporter and ade2-1 selective growth assays (Roque et al. 2015). However, readthrough in this study was measured in the presence of suppressor tRNA, in which the mechanism differs from the competition between near-cognate tRNAs and eRF1. Another yeast in vivo study showed that deletion of Pab1 (in a pbp1Δupf1Δ background to suppress lethality and control for mRNA level), or Pab1’s P-C domains, increased readthrough efficiency of premature termination codons (PTCs) across the LUC (luciferase) ORF (Wu et al. 2020), supporting the notion that Pab1 promotes termination by recruiting the release factors to the stop codon. Direct in vivo evidence for PABPC’s role in termination in mammalian cells has yet to be demonstrated.

The study from Wu et al. also provided important evidence regarding Pab1’s role in regulating termination, namely that readthrough efficiency decreases and termination efficiency increases as PTCs get closer to the mRNA 3’ end (where Pab1 resides) and this trend is disrupted in the absence of Pab1 (Wu et al. 2020). In the same vein, readthrough was shown to increase with 3’-UTR lengthening (Wu et al. 2020). These results are consistent with the extent of stop codon re-assignment as sense codons in a position-dependent manner in ciliates, where
closer proximity to 3’ end-localized PABPC has been shown to enhance termination (Heaphy et al. 2016; Swart et al. 2016; Záhonová et al. 2016). Moreover, these results hint at how the nonsense-mediated mRNA decay (NMD) pathway may distinguish between premature and normal termination in the faux 3’-UTR model (Amrani et al. 2004; Passmore and Coller 2022). Tethering PABPC or Pab1 downstream of PTCs has been shown to antagonize NMD activity, as with tethering eRF3 (Amrani et al. 2004; Ivanov et al. 2008). Since premature termination appears to be less efficient or slower than normal termination (Amrani et al. 2004, 2006), these results are consistent with PABPC’s role in termination enhancement.

**Nonsense-mediated mRNA decay (NMD)**

Eukaryotic cells possess multiple quality control pathways that monitor the fidelity of translation to minimize the accumulation of aberrant mRNAs, prevent the production of faulty proteins from such mRNAs, and rescue stalled ribosomes (Schuller and Green 2018). The three well characterized forms of these specialized “mRNA surveillance” mechanisms are no-go decay (NGD), non-stop decay (NSD), and nonsense-mediated decay (NMD) (Shoemaker and Green 2012; Monaghan et al. 2023). All three pathways share the core canonical mRNA decay machineries, namely deadenylation by Pan2-Pan3 and Ccr4-Not complexes, decapping by Dcp1-Dcp2 holoenzyme, endonucleolytic cleavage, 5′-3′ exonucleolytic decay by Xrn1, and 3′-5′ exonucleolytic decay by the exosome (Shoemaker and Green 2012). They differ in dictating their mRNA targets and the
combinations of core machineries used to degrade the mRNAs. NGD targets mRNAs containing stem loop structures, stretches of non-optimal codons, or other elongation-hindering sequences that cause ribosome stalling (Doma and Parker 2006). NSD targets mRNAs with no stop codons that cause ribosomes to translate into mRNA poly(A) tails (Frischmeyer et al. 2002; Van Hoof et al. 2002). Both NGD and NSD involve a complex of ribosome-binding protein Pelota (Dom34 in yeast) and the GTPase Hbs1 (Pisareva et al. 2011; Guydosh and Green 2014, 2017) in the ribosome decoding center, structurally homologous to eRF1-eRF3 complex (Shao et al. 2016). NMD targets mRNAs undergoing premature termination and requires the conserved Upf factors, Upf1, Upf2, and Upf3 (Leeds et al. 1991; He et al. 1997; He and Jacobson 2015b; Kurosaki et al. 2019). In metazoans, additional Smg complex and exon junction complex (EJC) proteins are involved in NMD (Kurosaki et al. 2019).

**The Upf factors: Upf1, Upf2, and Upf3**

The core Upf proteins, Upf1, Upf2, and Upf3, are evolutionarily conserved in eukaryotes, from yeast to human. Upf1 is the principal NMD factor, responsible for both substrate targeting and promoting mRNA degradation, and its activation is stimulated by Upf2 and Upf3.

Upf1 is a superfamily I RNA helicase consisting of two major domains: a cysteine-/histidine-rich (CH) N-terminal domain and a helicase/ATPase C-terminal domain (Leeds et al. 1991; Altamura et al. 1992; Koonin 1992; Czaplinski et al. 1995; Clerici et al. 2009; Chakrabarti et al. 2011). The CH domain binds the C-
terminal region of Upf2 (He et al. 1997), the ribosomal protein Rps26 (Min et al. 2013), the decapping enzyme catalytic subunit Dcp2 (He and Jacobson 1995, 2015a, 2022), and the C-terminal domain of eRF3 (Ivanov et al. 2008). The Upf1 CH domain also interacts with its own helicase domain and self-associates (Fiorini et al. 2013; He et al. 2013). These various overlapping interactions suggest that Upf1 is involved in multiple, sequential steps of NMD and may dimerize during some step(s). Two distinct complexes have already been identified as Upf1-2-3 and Upf1-decapping complexes (Dehecq et al. 2018). In metazoans, the Upf1 N-terminal domain is extended upstream of the CH domain by an unstructured region and the C-terminal domain is extended downstream of the helicase domain by a serine-/glutamine-rich (SQ) region, both of which contain multiple phosphorylation sites and interaction sites for Smg proteins and Dcp2 (Perlick et al. 1996; Applequist et al. 1997; Lykke-Andersen 2002; Okada-Katsuhata et al. 2012; Loh et al. 2013; Nicholson et al. 2014; He and Jacobson 2015b; Kim and Maquat 2019).

In both yeast and metazoans, the C-terminal domain has the RNA-dependent ATPase and helicase activities required for NMD activation (Weng et al. 1996a, 1996b; Franks et al. 2010) as well as a binding site for eRF1 (Ivanov et al. 2008; Zhouravleva and Gryzina 2012). In yeast, the helicase domain additionally binds the L1 stalk of the 25S rRNA, potentially placing Upf1 at the ribosomal E-site (Schuller et al. 2018).

Upf2 is an acidic protein containing three tandem eIF4G-like (Middle portion of eIF4G, MIF4G) domains and a C-terminal domain with Upf1 and eRF3 binding
sites (He and Jacobson 1995; Serin et al. 2001; Wang et al. 2001; Clerici et al. 2014; Fourati et al. 2014; López-Perrote et al. 2016; Zhouravleva et al. 2022). MIF4G-1 and MIF4G-2 act as a scaffold for MIF4G-3, which binds Upf3 and, in metazoans, Smg1, and is required for triggering NMD (Ponting 2000; Clerici et al. 2014; Fourati et al. 2014; Bufton et al. 2022). Upf2 is critical for Upf1’s conformational change and helicase activity. When unbound to Upf2, the intramolecular interaction between Upf1’s CH and helicase domains puts Upf1 in a closed conformation, inhibiting its helicase and ATPase activities (Chakrabarti et al. 2011; Fiorini et al. 2013). Upon Upf2 binding to the CH domain, Upf1 adopts an open conformation, which results in decreased RNA binding activity (Chamieh et al. 2008; Chakrabarti et al. 2011; Xue et al. 2023).

Upf3 is a basic protein with one isoform in yeast and Caenorhabditis elegans and two paralogs in vertebrates, Upf3a and Upf3b (also known as Upf3X), the latter of which is more prevalent (Lee and Culbertson 1995; Serin et al. 2001; Kunz et al. 2006). Upf3 contains an RNA recognition motif (RRM) domain that binds Upf2 and eRF3, and a C-terminal domain that binds the EJC (Lee and Culbertson 1995; Serin et al. 2001; Wang et al. 2001; Kim et al. 2001; Gehring et al. 2003; Kadlec et al. 2004; Buchwald et al. 2010; Neu-Yilik et al. 2017; Bufton et al. 2022; Zhouravleva et al. 2022). Upf3 is required for NMD in yeast, but its exact role in the pathway is still unclear. Unlike in yeast, metazoan Upf3 appears to be less important and perhaps dispensable to NMD because NMD remained partially
active upon null mutation or deletion of both paralogs (Avery et al. 2011; Yi et al. 2022).

**Targets of NMD**

*Nature of NMD substrates*

Early studies in yeast and mammalian cells indicated that PTC-containing mRNAs were unstable and subject to accelerated decay (Benz and Forget 1971; Losson and Lacroute 1979; Maquat et al. 1981), but the basis of that accelerated decay was thought to be a consequence of ribosome drop-off that rendered mRNAs susceptible to cytoplasmic endonucleases. The experiments of Leeds et al. identified the yeast *UPF1* gene as a specific regulator of the stability of PTC-containing mRNAs but not WT mRNAs, thus demonstrating that a specific decay pathway selectively targeted the former mRNAs (Leeds et al. 1991). This conclusion was reinforced by experiments showing that two additional yeast proteins, Upf2 and Upf3, joined Upf1 in a complex that targeted PTC-containing mRNAs for decapping and 5' to 3' degradation, as well as by experiments identifying similar regulatory components in *C. elegans* and eventually in mammals and plants (Hodgkin et al. 1989; Leeds et al. 1991; Altamura et al. 1992; Pulak and Anderson 1993; He and Jacobson 1995; Lee and Culbertson 1995; Peltz et al. 1994; Perlick et al. 1996; He et al. 1997; Serin et al. 2001; Gatfield 2003; Arciga-Reyes et al. 2006; Kertész et al. 2006). The pathway was dubbed NMD when it was shown that its substrates included intron-containing mRNAs that entered the cytoplasm, i.e., its activity was not limited to the transcripts of genes harboring
nonsense mutations (He et al. 1993). The discovery of more NMD targets followed from the detection of mRNAs that were upregulated upon Upf factor depletion (He et al. 2003; Mendell et al. 2004; Rehwinkel et al. 2005; Yepiskoposyan et al. 2011; Celik et al. 2017a; Colombo et al. 2017; Muir et al. 2018). NMD substrates can be categorized into two major types: traditional and probabilistic NMD substrates (Celik et al. 2017b), i.e., transcripts with “hard-wired” PTCs and transcripts in which ribosomes encountered PTCs as a consequence of aberrant transcription, pre-mRNA processing, or translation.

Traditional NMD substrates are mRNAs that have encoded early in-frame, stop codons, which include the transcripts of genes with nonsense or frameshift mutations (Benz and Forget 1971; Losson and Lacroute 1979; Maquat et al. 1981; Leeds et al. 1991; He et al. 2003; Celik et al. 2017a), unspliced pre-mRNAs incorrectly exported into the cytoplasm (He et al. 1993; Sayani et al. 2008; Celik et al. 2017a), mRNA products of alternative splicing (e.g., exon skipping, PTC-containing exon inclusion, intron retention) (Mendell et al. 2004; Lareau et al. 2007; Ni et al. 2007; Jaillon et al. 2008; McGlincy and Smith 2008; Weischenfeldt et al. 2012; Ge and Porse 2014; Lykke-Andersen et al. 2014; Muir et al. 2018; Karousis et al. 2021; Embree et al. 2022), mRNAs with programmed frameshifting (He et al. 2003; Celik et al. 2017b), and transcribed pseudogenes and nonproductively rearranged genes (Li and Wilkinson 1998; He et al. 2003; McGlincy and Smith 2008; Celik et al. 2017a; Muir et al. 2018). An additional class of traditional substrates includes some but not all of the cellular mRNAs that have uORFs (He
et al. 2003; Mendell et al. 2004; Gaba et al. 2005; Nyikó et al. 2009; Yepiskoposyan et al. 2011; Arribere and Gilbert 2013; Celik et al. 2017a; Colombo et al. 2017; Cai et al. 2021; Russell et al. 2023), atypically long 3′-UTRs or 3′-UTR introns that make their NTCs resemble PTCs (Mühlrad and Parker 1999; Bühler et al. 2006; Singh et al. 2008; Kebara and Atkin 2009; Kertesz et al. 2010; Yepiskoposyan et al. 2011; Zaborske et al. 2013; Colombo et al. 2017; Murtha et al. 2019), and normal noncoding RNAs (e.g., long noncoding RNAs or the transcripts of transposable elements) that happen to get translated (He et al. 2003; Thompson and Parker 2007; Kurihara et al. 2009; Tani et al. 2013; Smith et al. 2014; Celik et al. 2017a; Colombo et al. 2017; Andjus et al. 2021).

Probabilistic NMD substrates are RNAs that lack in-frame PTCs, but may contain features that promote out-of-frame translation that ultimately lead to encounters with PTCs (Celik et al. 2017b). These include mRNAs with a propensity for leaky scanning where translation initiates downstream and out-of-frame of the canonical site (Welch 1999; He et al. 2003), mRNAs generated from errors in transcription start site, and mRNAs with stretches of non-optimal codons that can lead to unexpected frameshifting during translation (Celik et al. 2017a). Up to 10-15% of normal-looking, polysome-associated mRNAs in the transcriptome (some of which lack these known features) have also been found as NMD substrates (He et al. 2003; Rehwinkel et al. 2005; Celik et al. 2017a). These probabilistic substrates establish that NMD is a quality control pathway that continues to monitor translation throughout an mRNA’s life cycle, not just the first round of
translation (Celik et al. 2017b). Consistent with this notion, a single-molecule imaging study showed that the decay kinetics of a PTC-containing reporter mRNA were described poorly by the mathematical simulation for NMD occurring only during the first round translation but fit the simulation for multiple rounds of translation with equal probability of NMD induction for each ribosome (Hoek et al. 2019). Moreover, NMD efficiency was similar on cap-binding complex-bound mRNAs (first round of translation) and eIF4E-bound mRNAs (subsequent rounds of translation) (Hoek et al. 2019).

**Differences between premature and normal termination**

An important long-standing question about NMD is how the pathway discriminates between premature and normal termination and selectively targets mRNAs containing PTCs. Premature termination must differ in some way from normal termination. The occurrence of toeprinting signals in cell-free extracts at PTCs but not at NTCs and the occurrence at NTCs only when eRF1 was inactivated suggest that premature termination or recycling is slower and perhaps less efficient than normal termination (Amrani et al. 2004, 2006; Peixeiro et al. 2012). This conclusion is supported by observations that PTCs exhibit higher readthrough than NTCs, as demonstrated by higher detectability of PTC readthrough product compared to NTC readthrough product from reporter constructs (Wu et al. 2020) and stronger responses to readthrough-inducing drugs in various studies of nonsense suppression therapies (Peltz et al. 2013; Keeling et al. 2014; Roy et al. 2015) (also see “Therapeutic strategies targeting termination:
promoting stop codon readthrough”). In contrast with these observations, a recent study found that the extent of ribosome stalling as measured by ribosome profiling on termination codons of NMD-sensitive vs. NMD-insensitive mRNAs did not differ; however, that study incorrectly focused on NTCs instead of PTCs of NMD substrates (Karousis et al. 2020).

A mechanistic explanation for inherent differences in termination efficiency between PTCs and NTCs may be related to nucleotide context near the stop codon (discussed in “Cis-regulatory elements of readthrough efficiency”). Since PTCs occur randomly and are not subject to the same evolutionary pressure as NTCs to be efficient, PTCs may be in a context that is more prone to basal readthrough and respond more readily to readthrough-inducing drugs. However, the relationship between stop codon context-mediated termination efficiency and NMD susceptibility has not been directly investigated.

Another, perhaps more obvious, explanation for inherent differences in termination efficiency between PTCs and NTCs is the faux-UTR model, first proposed in yeast, which posits that premature termination lacks or has diminished interactions with termination-enhancing factors localized in the mRNA 3’ end, namely PABPC (Amrani et al. 2004). PABPC’s interaction with eRF3 is thought to stimulate the release factor complex recruitment to the stop codon, increasing termination efficiency (Hoshino et al. 1999b; Cosson et al. 2002b; Ivanov et al. 2016) (see “Trans-regulatory elements of readthrough efficiency”). Thus, PTCs which are further away from PABPC are thought to benefit less from PABPC’s
termination-enhancing activity. In support of this model, mRNAs with abnormally long 3'-UTRs are substrates of NMD (see above) and 3'-UTR lengthening by a mutation of a polyadenylation site made the mRNA sensitive to NMD (Muhlrad and Parker 1999). A recent study demonstrated that termination efficiency increased while readthrough efficiency decreased as PTCs near the mRNA 3' end, and this trend was dependent on the presence of PABPC (Wu et al. 2020). Moreover, PABPC has been directly shown to antagonize NMD also in a proximity-dependent manner, evidenced by mRNA stabilization upon tethering PABPC or eRF3 close to PTCs (Amrani et al. 2004; Behm-Ansmant et al. 2007; Eberle et al. 2008; Ivanov et al. 2008; Silva et al. 2008; Singh et al. 2008; Kervestin et al. 2012). It is also postulated that when PABPC is near the stop codon, eRF3-PABPC interaction possibly outcompetes eRF3-Upf1 and/or eRF3-Upf3 interactions, hence inhibiting NMD (Ivanov et al. 2008; Singh et al. 2008; Kervestin et al. 2012). While this model is reasonable for yeast, the much longer 3'-UTRs that exist in mammals (average 1278 nt in mammals vs. 121 nt in yeast) makes the distance-dependent discrimination of PTC vs. NTC more challenging (Muñoz et al. 2023). The existence of mRNAs with long 3'-UTRs that are NMD-resistant raises a question of exactly how long would be considered too long. Additionally, the fact that longer 3'-UTRs should have a higher propensity to be spatially non-linear and form secondary structures make the actual distance difficult to determine. However, this non-linear structure can result in PABPC being spatially close to the stop codon despite the long 3'-UTR, which is actually in line with the proximity concept of this
model. This idea is speculated to be the reason some mRNAs with long 3’-UTRs can escape NMD (Singh et al. 2008).

Despite Upf proteins interaction with the release factors (Czaplinski et al. 1995; Kashima et al. 2006; Singh et al. 2008), it is unclear whether they influence termination efficiency at PTCs or not, but Upf1 seems to have a role in ribosome recycling (Franks et al. 2010; Ghosh et al. 2010; Durand et al. 2016; Serdar et al. 2020). Initially, Upf proteins were thought to help recruit release factors because deletion of UPF genes increased readthrough of PTCs in vivo (Weng et al. 1996b; Maderazo et al. 2000; Wang et al. 2001). However, this observation appears to be an indirect consequence of stabilization of two NMD substrates, the uORF-containing ALR1 and ALR2 mRNAs that encode Mg^{2+} transporters (Johansson and Jacobson 2010). Deletion of ALR1 counteracted the readthrough phenotype of UPF1 deletion (Johansson and Jacobson 2010). Additionally, changing of cellular Mg^{2+} concentration interfered with the function of the ribosome recycling factor ABCE1 (Barthelme et al. 2011; Sims and Igarashi 2012), which would lead to an apparent increase readthrough phenotype. Deletion or mutation of UPF1, however, led to defects in post-termination ribosome release at PTCs (Amrani et al. 2004; Franks et al. 2010; Ghosh et al. 2010; Serdar et al. 2016, 2020), suggesting that termination complex disassembly is usually slow or inefficient at PTCs. Upf1 is thought to mediate 60S subunit dissociation, at least in yeast, because low-level AUG-dependent reinitiation consistent with the canonical 40S scanning mechanism of translation initiation was observed near PTCs in a Upf1-
dependent manner (Ghosh et al. 2010). Consistent with this notion, non-translating ribosomes have been shown to accumulate on 3’ mRNA decay intermediates (downstream of PTCs), blocking exonucleolytic decay, when Upf1’s ATPase activity was deficient (Franks et al. 2010; Serdar et al. 2016, 2020). Thus, PTCs appear to have less efficient termination and recycling than NTCs, and Upf1 seems to recognize as well as aid at least one of these processes.

Mechanisms of NMD substrate targeting by Upf1

Another long-standing question about NMD is when and how Upf1 is recruited to the prematurely terminating mRNP. Conflicting results from different organisms and experimental setups lead to multiple hypotheses of Upf1 recruitment.

One hypothesis is that Upf1 stochastically pre-associates with either the mRNA or the ribosome, poised to be readily activated upon a premature termination event. Pre-association with the mRNA is proposed in the 3’-UTR sensing model for metazoans, which was conceived based on the observations that Upf1 was enriched in mRNA 3’-UTRs in a sequence-independent manner in RNA-based affinity purification and CLIP-Seq experiments (Hogg and Goff 2010; Hurt et al. 2013; Zünd and Mühlemann 2013; Zünd et al. 2013; Kurosaki and Maquat 2013; Gregersen et al. 2014; Kurosaki et al. 2014; He and Jacobson 2015b; Embree et al. 2022). The model postulates that Upf1 binds mRNA throughout its entire length prior to translation but is displaced by translating ribosomes, hence Upf1 is only found in the 3’-UTR upon normal termination. When
termination occurs prematurely, the not-yet-displaced Upf1 results in the higher local Upf1 concentration and, in combination with the reduced eRF3-PABPC interactions, enhance Upf1’s interaction with the release factors at the PTC, triggering NMD. Consistent with this model, PTC readthrough and reinitiation, where translation would have continued to displace mRNA-bound Upf1, antagonize NMD (Keeling et al. 2004; Hogg and Goff 2010; Neu-Yilik et al. 2011; Cohen et al. 2019; Embree et al. 2022). However, the observed 3’-UTR enrichment could be related to Upf1’s function in other decay pathways (Kim and Maquat 2019), since this model seems contradictory to the translation-dependent nature of NMD (Atkin et al. 1995, 199, 1997; Zhang et al. 1997; Hu et al. 2010). Upf1 binding throughout the mRNA and its displacement by translocating ribosomes would actually make Upf1 most concentrated on untranslated mRNAs. Additionally, the depletion of mRNA coding sequences in the experiments implies that once Upf1 is displaced during the first round of translation, it is difficult for Upf1 to bind the mRNA again, which means this model cannot sufficiently explain how NMD targets probabilistic substrates that undergo premature termination during subsequent rounds of translation.

An alternative explanation to the depletion of mRNA coding sequences identified in the affinity purification and CLIP-Seq experiments is that Upf1 pre-associates with ribosomes. Ribosomal proteins as well as translation factors were in fact present in the affinity-purified fraction (Hogg and Goff 2010; He and Jacobson 2015b) and ribosome complexes were likely too large to be efficiently
resolved through SDS-PAGE in the CLIP-Seq protocol, not to mention the fact that reads mapped to rRNAs are routinely discarded bioinformatically from sequencing data analysis. In yeast, evidence for direct Upf1-translating ribosome interaction includes Upf1’s co-sedimentation with polyribosomes in a nuclease-resistant and high salt-resistant manner (Atkin et al. 1995, 1997; Mangus and Jacobson 1999; Min et al. 2013; Celik 2017) as well as the identification of Upf1’s CH domain interactions with Rps26 on the 40S ribosomal subunit (Min et al. 2013). Mammalian Upf1 is also polyribosome-associated (Nott et al. 2004; López-Perrote et al. 2016; Kurosaki et al. 2018; Yoshikawa et al. 2018). In addition, the level of cellular Upf1 molecules is at least ten times higher than that of Upf2 and Upf3 (Maderazo et al. 2000; Kurosaki et al. 2019), arguing against the idea that all three factors are only recruited upon premature termination and supporting the idea that Upf1 likely has an additional role not associated with Upf2 and Upf3. While it is tempting to conclude that these results mean Upf1 binds elongating ribosomes, it remains to be determined whether Upf1 actually binds only prematurely terminating ribosomes, and its co-migration with polysomes is detected because translation is still active on that transcript. Thus, direct evidence for this hypothesis has been limited, but recent studies suggest that it occurs and involves the ATPase activity of Upf1. Lee et al. proposed that Upf1 undergoes rounds of associating and dissociating with all mRNAs, and efficient dissociation from non-targets is ATPase-dependent, as ATPase-deficient Upf1 accumulated on non-targets as much as NMD targets, losing its substrate specificity (Lee et al. 2015). On NMD targets, the
ATPase activity may be delayed until after recruitment of other NMD factors, and together with the helicase activity, promotes mRNP remodeling and dissociation. In support of this model, ATPase and helicase activities of Upf1 can be decoupled and the ATPase activity alone was sufficient for dissociation and for functional NMD (Chapman et al. 2022), suggesting that ATPase-mediated dissociation during its surveillance function is possible. However, this model does not explain why there would be Upf1 detected in mRNA 3'-UTRs.

The problem with either pre-association hypothesis (Upf1-mRNA or Upf1-ribosome) is the disproportionate Upf1-ribosome stoichiometry, making it impossible for Upf1 to be on every ribosome or mRNA. The cellular Upf1 level is only 1-3% of the level of ribosomes in yeast (Maderazo et al. 2000; Ghaemmaghami et al. 2003; Celik et al. 2015), and while the levels are more even in mammalian cells (Pal et al. 2001), Upf1 has multiple roles in other decay pathways (Kim and Maquat 2019). It is then likely that Upf1 only associates with ribosomes undergoing premature termination. This hypothesis is applicable to the faux-UTR model as well as exon junction complex (EJC)-dependent NMD in metazoans. EJCs are mRNA splicing complexes left over 24 nt upstream of exon-exon junctions on spliced mRNAs and are thought to be removed by ribosomes and/or Pym1 protein during the first round of translation (Gehring et al. 2009; Woodward et al. 2017; Schlautmann and Gehring 2020). Thus, unlike normal termination, where all EJCs would have already been removed, premature termination that occurs before the last exon would still be in the presence of at
least one downstream EJC. The downstream EJC is thought to link to NMD through a Upf2-Upf3 complex (Le Hir et al. 2001; Kim et al. 2001; Gehring et al. 2003, 2005; Ivanov et al. 2008; Buchwald et al. 2010; López-Perrote et al. 2016), which then recruits Upf1 (Chamieh et al. 2008; Kurosaki et al. 2019). In support of this model, tethering EJC components near a stop codon activated NMD (Lykke-Andersen et al. 2001; Gehring et al. 2003, 2005). Additionally, the presence of an EJC and the distance of a PTC to an EJC are two of the strongest determinants of NMD efficiency in the human transcriptome (Lindeboom et al. 2016). However, this model limits NMD targeting to only the subset of NMD substrates that have more than one exon and to only the first round of translation, thus excluding probabilistic NMD substrates. EJCs may also enhance NMD indirectly by enhancing translation, as NMD is translation-dependent. The higher expression of spliced mRNAs compared to unspliced mRNAs have been attributed to EJCs through interactions between core EJC/ EJC-bound proteins and initiation factor eIF3/48S pre-initiation complex, although the exact mechanism of translation enhancement is still unknown (Nott et al. 2003, 2004; Ma et al. 2008; Chazal et al. 2013; Schlautmann and Gehring 2020). Thorough studies to isolate the two effects of EJC on NMD are needed to determine the extent of EJC-mediated NMD independent of its translation-inducing function.

Despite the different hypotheses in Upf1 recruitment to targeted mRNAs, targeting is still the most elusive step of NMD. These models are also not mutually exclusive, e.g., the faux-UTR model and EJC model work with both the pre-
association hypothesis as well as the recruitment hypothesis. However, yeast and metazoans have many different features that determine which models are more likely (see below).

**Mechanisms of NMD activation and substrate degradation**

Yeast and metazoan NMD pathways are quite different in the possible targeting mechanisms, activation, and choices of degradation pathways. I first discuss the current model of the simpler yeast NMD pathway and then the more complicated metazoan NMD pathway.

**Current model of yeast NMD**

Yeast Upf1 has been implicated to play a role in multiple steps of NMD, not just substrate targeting. This idea is supported by the observations that: i) Upf1 is polyribosome-associated via direct interaction of its CH domain with Rps26 (Atkin et al. 1995, 1997; Mangus and Jacobson 1999; Min et al. 2013), ii) Upf proteins interact with release factors (He and Jacobson 1995; Wang et al. 2001; Ivanov et al. 2008), iii) Upf1 can dimerize (He et al. 2013), iv) Upf1 exists in at least two distinct complexes, the “detector” complex containing Upf1-Upf2-Upf3 and the “effector” complex containing Upf1-decay components (Dehecq et al. 2018), v) Upf1 binds two non-overlapping sites in the C-terminal region of decapping enzyme subunit Dcp2 (He and Jacobson 1995, 2015a; He et al. 2018, 2022), but deletion of the two Upf1 binding sites on Dcp2 did not result in the same extent of substrate stabilization as deletion of *UPF1* itself (2-3 fold vs. > 10 fold), indicating that decapping is not rate-limiting and that Upf1 has a major upstream function (He
et al. 2022), vi) decapping of NMD substrates occurs while the mRNAs are polyribosome-associated (Hu et al. 2010), and vii) deletion of Upf1 or ATPase-deficient Upf1 result in defects in ribosome recycling (Franks et al. 2010; Ghosh et al. 2010; Serdar et al. 2016, 2020). Based on these results, the yeast NMD pathway likely involves the following steps.

Upf1’s consistent, direct association with polyribosomes suggests that Upf1 may bind ribosomes stochastically during the course of translation, “surveying” for premature termination. Because yeast does not have many intron-containing mRNAs and lacks EJCs, Upf1 most likely senses premature termination by the inherent inefficiency of premature termination (i.e., the faux-UTR model). Upon a ribosome’s encounter with a PTC in its A-site, the slow recruitment of the release factors due to non-optimal termination context leads to Upf2-Upf3 binding (“detector” complex), stabilizing the Upf1-ribosome complex and activating Upf1’s ATPase activity. ATP hydrolysis promotes the dissociation of the 60S ribosomal subunit, the release factors, Upf2, and Upf3. This step is likely the rate-limiting step of NMD that commits the mRNA to decay. Next, the “effector” complex is created when Upf1 recruits the decapping enzyme to the remaining 40S mRNP via its interaction with Dcp2, likely involving two points of contacts: 40S-bound Upf1 dimerizes with the free Dcp2-bound Upf1, at the same time allowing 40S-bound Upf1 to bind to the second Dcp2 binding site. Upon decapping complex recruitment, the mRNA 5’ cap is removed and the mRNA is digested 5’ to 3’ by the Xrn1 exonuclease (He and Jacobson 2022). This deadenylation-independent
decay is the predominant decay pathway for NMD substrates in yeast. However, in the absence of decapping, NMD can also utilize the alternative deadenylolation followed by exosome-mediated 3′-5′ decay via binding of Upf1 with Ski7, a part of the Ski complex that interacts with exosome (Mitchell and Tollervey 2003; Takahashi et al. 2003).

In addition to directly demonstrating Upf1 pre-association with ribosomes, this model is far from complete owing to a few contradictory observations. Although the “detector” complex should theoretically contain the release factors and ribosomal proteins, they were not enriched in the Upf-TAP purification (Dehecq et al. 2018). It is proposed that this result is due to either the short duration of transient interactions between the Upf factors and the terminating mRNP or the interactions did not survive through the purification procedures (Dehecq et al. 2018). The former explanation seems consistent with a role of Upf1 facilitating ribosome release. However, Ghosh et al. showed that the defect in ribosome release was consistent with a 40S scanning mechanism, suggesting that only the 60S subunit was released by Upf1 (Ghosh et al. 2010). If that is the case, ribosomal proteins of the 40S subunit should have been identified as well. Thus, the latter explanation of procedural limitation might apply. Further experiments are needed to consolidate these observations and determine the precise role of Upf1 in ribosome recycling.

Another aspect of the data that has not been incorporated into the model is the presence of Nmd4 and Ebs1 in the “effector” complex (Dehecq et al. 2018). Nmd4, the yeast homolog of metazoan Smg6, was identified in a screen for NMD
components (He and Jacobson 1995), but has not been further characterized. Ebs1, identified as the yeast homolog of Smg5/7, was shown to play a Smg-like role in NMD and interact with Upf1 (Luke et al. 2007). These results suggest that an endonucleolytic pathway and additional route to deadenylation-dependent pathway may also exist in yeast. (These pathways are discussed for metazoans below).

**Current models of metazoan NMD**

Unlike yeast, there are multiple branches of NMD activation and multiple branches of mRNA degradation in metazoans, with much more complex data (Lejeune et al. 2003; Kurosaki et al. 2019; Yi et al. 2021). Currently, modes of NMD activation can be divided into two major branches, EJC-dependent and EJC-independent, considering most genes have multiple exons in metazoans. EJC-dependent NMD is the better understood and the more efficient branch. EJC-s located downstream of the ribosome undergoing premature termination are thought to serve as a platform for Upf2-Upf3b binding via direct interaction with Upf3b (Buchwald et al. 2010). Since Upf3 is also found in the nucleus and can shuttle between the nucleus and the cytoplasm (Serin et al. 2001), it is even thought that the Upf2-Upf3b-EJC complex, termed 3′-UTR EJC, might form in the nucleus during pre-mRNA splicing and be exported along with the mRNA into the cytoplasm. Upon premature termination, the 3′-UTR EJC complex is thought to recruit Upf1 and Smg1 to the terminating ribosome, making the Smg1-Upf1-eRFs (SURF) complex (Kashima et al. 2006). Smg1 is a phosphatidylinositol 3-kinase-
related protein kinase that phosphorylates serine/threonine-glutamine (S/TQ) motifs on Upf1 (Yamashita et al. 2001; Durand et al. 2016), and only does so when recruited to NMD-targeted mRNPs. Prior to recruitment, Smg1 is in a complex with Smg8 and Smg9, co-factors that inhibit its kinase activity (Yamashita et al. 2009; Arias-Palomo et al. 2011). Smg1-mediated phosphorylation is enhanced by the RNA helicase DHX34, which acts as a scaffold protein and promotes mRNP remodeling from surveillance mode into decay-inducing (DECID) mode (Hug and Cáceres 2014; Melero et al. 2016).

SURF complex formation, Upf1 phosphorylation by Smg1, and mRNP remodeling to form the DECID complex are also part of EJC-independent NMD, only that Upf1 recruitment mechanism is through other means, likely sensing the inherent mechanistic differences between premature and normal termination. This branch would take care of premature termination that occurs during subsequent rounds of translation as well as mRNAs transcribed from genes lacking exons.

Upf1 phosphorylation creates non-overlapping binding sites for Smg6 and Smg5-Smg7 proteins that initiate the decay mechanisms, although phosphorylation-independent association has also been observed (Okada-Katsuhata et al. 2012; Chakrabarti et al. 2014; Nicholson et al. 2014). Immediate recruitment of the decapping complex to the mRNPs, as occurs in yeast, has been demonstrated through direct Upf1-Dcp2 interaction and Upf1-Smg5-Pnrc2 interaction, but this pathway seems to be minor in metazoans (Lykke-Andersen 2002; Cho et al. 2009; Lai et al. 2012; Cho et al. 2013; Nicholson et al. 2018; Yi et
The major pathways are the endonucleolytic pathway involving Smg6 and the deadenylation/decapping-dependent pathway involving Smg5-Smg7 heterodimer.

Smg6 is an endonuclease that cleaves mRNA at or nearby the PTC, creating two mRNA fragments (Huntzinger et al. 2008; Eberle et al. 2009; Boehm et al. 2014; Lykke-Andersen et al. 2014; Schmidt et al. 2015; Kim et al. 2022). The 5' fragment is degraded 3' to 5' by the exosome while the 3' fragment that contains Upf1 and (sometimes) EJC(s) is degraded 5' to 3' by Xrn1 (Eberle et al. 2009; Kurosaki et al. 2019). It is still unclear when ATP hydrolysis by Upf1 and ribosome release occurs, but it appears that Smg6 cleavage activity is not dependent on disassembly of mRNPs. When Upf1 lacked ATPase activity, the 3' fragment but not the 5' fragment accumulated, suggesting that Smg6 could still cleave the mRNA but Xrn1's access to the 3' fragment was blocked, likely by the presence of unreleased ribosomes (Franks et al. 2010). Recently, ribosome stalling at Smg6 cleavage sites was detected through ribosome profiling, raising a question whether ribosome release occurs before or after Smg6 cleavage (Kim et al. 2022). In another pathway, a Smg5-Smg7 heterodimer recruits the Ccr4-Not deadenylase complex via direct interaction between Smg7 and Pop2 (Unterholzner and Izaurralde 2004; Loh et al. 2013; Nicholson et al. 2018). The Ccr4-Not complex removes poly(A) tails and recruits the decapping complex, ensuring both exosome-mediated 3'-5' decay and Xrn1-mediated 5'-3' decay (Unterholzner and Izaurralde 2004; Yamashita et al. 2005; Kurosaki et al. 2019). The two pathways were first
thought to be distinct and partially redundant because they target the same transcripts and individual depletion of each Smg only partially stabilized NMD substrates (Metze et al. 2013; Boehm et al. 2016; Colombo et al. 2017), but recent evidence for the dependence of Smg6’s cleavage activity on Smg5 (Nelson et al. 2018) and/or Smg7 (Boehm et al. 2021), as well as possible roles of Smg5/Smg7 outside of the heterodimer (Huth et al. 2022), indicate that they may be more interdependent than anticipated, with slight variation in target specificity (Ottens et al. 2017).

Despite the identification of the key players of the NMD pathway, the outstanding fundamental questions for both yeast and metazoan NMD are how exactly Upf1 distinguishes its substrates from non-substrates and is recruited to premature termination. Answers to these questions are important for further development of target-specific therapeutics for nonsense mutation diseases.

**Nonsense mutations and human diseases**

Nonsense mutations change amino acid-coding sense codons into stop codons in the middle of mRNA coding regions. Patients with nonsense mutations often have severe disease phenotypes due to essentially complete loss of functional protein expression. Such deficient expression has two fundamental causes: i) the truncated protein is usually nonfunctional or toxic, and ii) the associated nonsense-containing mRNAs are degraded by NMD. Since nonsense mutations account for approximately 10-11% of human genetic disease cases (Mort et al. 2008; Benhabiles et al. 2016), therapeutics targeting termination and/or
NMD have the potential to be beneficial to a large number of patients across various diseases, including but not limited to cystic fibrosis (CF), Duchenne muscular dystrophy (DMD), neurofibromatosis, and cancers (Temaj et al. 2023). So far, only one readthrough-promoting small molecule drug, ataluren (PTC124), has been approved for clinical use in Europe to treat DMD (Welch et al. 2007; Peltz et al. 2013; Li et al. 2023), but it is not effective for CF (Aslam et al. 2017, 2023). Many drugs have reached clinical trial stages but have not been approved for clinical use due to either insufficient therapeutic activity or notable toxicity. Although there are alternative therapeutic strategies in development for treating nonsense mutation diseases, such as gene editing or exon skipping, I focus on strategies that directly and mechanistically address premature termination, which are those that target either the termination process or NMD.

**Therapeutic strategies targeting termination: promoting stop codon readthrough**

One way to treat diseases caused by nonsense mutations is to induce stop codon readthrough, an approach known as therapeutic nonsense suppression (Peltz et al. 2013; Keeling et al. 2014; Nagel-Wolfrum et al. 2016; Bidou and Allamand 2018; Dabrowski et al. 2018; Martins-Dias and Romão 2021; Palma and Lejeune 2021; Porter et al. 2021; Manjunath et al. 2022; Li et al. 2023; Temaj et al. 2023; Wagner et al. 2023). The number of drug candidates is vast, with both known and unknown mechanisms. Here, I discuss performances and challenges of major approaches with known mechanisms of therapeutic nonsense
suppression, including: i) depletion of release factor levels or supplementation of certain tRNAs to increase the chance of near-cognate tRNA incorporation, ii) enhancement of near-cognate tRNA binding to ribosomal A sites harboring a stop codon, iii) expression of suppressor tRNAs that are cognate to stop codons, and iv) interference with the translation machinery undergoing (premature) termination.

Changes in expression levels of endogenous factors

Early studies of termination showed that decreased amounts of functional release factors, e.g., by mutation of the gene encoding eRF1 or prionization of eRF3, can increase readthrough efficiency (Stansfield et al. 1997; Akhmaloka et al. 2008; Baudin-Baillieu et al. 2014). Thus, knocking down expression of either factor appears to be a potentially attractive therapeutic strategy. Targeted degradation of eRF1 or eRF3 mRNA by small interfering RNA (siRNA), short hairpin RNA (shRNA), or antisense oligonucleotides (ASOs) have also been shown to increase readthrough efficiency of reporter constructs in human cells (Carnes 2003; Janzen and Geballe 2004). The use of small RNAs in patients is, however, challenging due to their instability, inefficiency in crossing the cell membrane, and immunogenicity, but delivery methods are being explored and such RNA products have proven to be clinically useful for reducing the levels of “normal” mRNAs that direct the synthesis of disease-causing proteins (Tatiparti et al. 2017; Adams et al. 2018; Raal et al. 2020; Hu et al. 2020; Friedrich and Aigner 2022). The use of ASOs to knock down eRF1 and eRF3 expression in mice has been tried, resulting in a moderate increase in the readthrough of a disease-causing PTC (Huang et al.
2019). The mice were able to tolerate lowering release factor levels to 30-40% but not lower, as determined by body weight, organ weight, and levels of liver enzymes (Huang et al. 2019). In another approach, small molecules that promote proteasome-mediated degradation of eRF1 and eRF3a/b have been shown to increase readthrough of PTCs in five different genetic diseases (Baradaran-Heravi et al. 2021; Sharma et al. 2021a; Lee et al. 2022). Treatment of patient-derived primary bronchial epithelial cells with an eRF1 degrader in combination with the aminoglycoside G418 has been shown to restore function to the CFTR protein, the protein whose defects are responsible for cystic fibrosis (Sharma et al. 2021a). As with depletion of eRF1 or eRF3, overexpression of near-cognate tRNAs would be expected to enhance the extent of readthrough. Indeed, increased expression of tyrosine and tryptophan tRNAs increased readthrough of reporters in human cells (Beznosková et al. 2021). Overexpressing tRNA^{Trp} also increased readthrough levels sufficient to restore p53 protein function (Beznosková et al. 2021). While changing expression of the release factors and/or near-cognate tRNAs has proved successful, off-target effects (that may not be immediately evident) are a serious concern since the effects are global and not likely to discriminate between PTCs and NTCs.

**Enhancement of near-cognate tRNA binding to the ribosomal A site**

Improvement of natural near-cognate tRNA incorporation is another approach being investigated as a means to promote sufficient readthrough to rescue protein function. Post-transcriptional modifications of tRNA anticodons and
adjacent nucleotides have been shown to change readthrough efficiency in yeast (Blanchet et al. 2014). These modifications, such as pseudouridylation, are purportedly why mismatched base pairing can occur (Blanchet et al. 2014). In yeast, pseudouridylation of uridine residues of stop codons has been shown to increase PTC readthrough in vitro and in vivo in a manner that is independent of stop codon context (Karijolich and Yu 2011; Adachi and Yu 2020). This independence of stop codon context and the high specificity of the pseudouridylation mechanism (i.e., limited off-target effects) make this approach a promising therapeutic strategy. The mechanism of endogenous pseudouridylation of tRNAs or rRNAs involves the box H/ACA snoRNPs system, which consists of core modification enzymes, a group of non-coding, hairpin-forming snoRNAs, and a complementary small guide RNA sequence. Targeted pseudouridylation of specific stop codons has been achieved in yeast by supplying customized small guide RNA sequences complementary to the sites of interest (Huang et al. 2012), and has been tested in human cells with varying degrees of success (Nir et al. 2022; Adachi et al. 2023; Song et al. 2023). One study included a variation where a minor isoform of one of the core proteins, pseudouridine synthase DKC1 that lacks a nuclear localization signal enhances the efficiency of the H/ACA snoRNP system (RESTART) (Song et al. 2023).

Another approach related to stop codon modification is the use of a nucleoside analog. Clitocine is a natural adenosine nucleoside analog originally isolated from mushrooms for its insecticidal property (Kubo et al. 1986). Clitocine
appeared in a high-throughput screen for small molecule readthrough drugs that also identified ataluren/PTC124 (Welch et al. 2000). Further characterization revealed that clitocine is incorporated into mRNAs, substituting for adenosine during transcription (Friesen et al. 2017). Clitocine substitution induced readthrough of PTCs in luciferase reporters with high potency against all three stop codons. It is most effective against UAA and UAG, then UGA, rescued full length production from a nonsense codon-containing p53 mRNA in cell lines, and inhibited growth of xenografted tumors in mice (Friesen et al. 2017). Clitocine substitution also affected the identity and frequency of near-cognate tRNAs decoding the stop codon, with an increase in first position mispairing and a decrease in third position mispairing (Friesen et al. 2017). These observations led to a hypothesis that clitocine also likely reduces eRF1’s ability to recognize and bind to the stop codon, reducing termination and increasing readthrough efficiencies (Friesen et al. 2017; Li et al. 2023), but the exact molecular mechanism has not been studied. Since clitocine incorporation is transcriptome-wide and not site-specific, the effect of clitocine on sense codon decoding was determined and found to be negligible (Friesen et al. 2017). However, clitocine is highly toxic, so it has not been pursued further as a readthrough-promoting drug (Li et al. 2023).

**Nonsense suppressor tRNAs**

A drawback of near-cognate tRNA overexpression or stop codon modifications is that the inserted amino acid might not be correct, i.e., it may not be the same amino acid that existed in the wild-type protein. Thus, a more useful
approach may be to use suppressor tRNAs, in which the anticodon of a tRNA carrying the amino acid of interest is modified to be cognate to the stop codon of interest. Suppressor tRNAs were first characterized in certain *E. coli* or yeast strains (Goodman et al. 1968, 1977). Subsequently, import of *E. coli*-derived suppressor tRNAs, site-directed mutagenesis of mammalian tRNAs, and synthetically constructed tRNAs were used to suppress nonsense mutations in mammalian cell culture, demonstrating suppression of all three stop codons (UAA suppressors tRNA\textsuperscript{Tyr} and tRNA\textsuperscript{Ser}, UAG suppressor tRNA\textsuperscript{Lys}, UGA suppressor tRNA\textsuperscript{Arg}) as well as full-length protein synthesis and rescue of several disease-causing mutations across multiple diseases (including synthesis of β-globin/β-thalassemia, XPA/Xeroderma pigmentosum, dystrophin/DMD, CDH1/E-cadherin/hereditary diffuse gastric cancer, CFTR/CF) (Temple et al. 1982; Panchal et al. 1999; Buvoli et al. 2000; Köhrer et al. 2001; Kiselev et al. 2002; Bordeira-Carriço et al. 2014; Lueck et al. 2019; Ko et al. 2022). Further modifications beyond the tRNA anticodon can also be beneficial: additional alteration in the tRNA T\textsubscript{Ψ}C stem increases its binding affinity to translation elongation factor eEF1A, enhancing the efficacy of suppressor tRNA in nonsense suppression (Albers et al. 2023). This finding is consistent with a previous study in yeast where loss of critical modifications of UAA suppressor tRNA \textit{SUP4} can be rescued by elevating eEF1A level (Klassen and Schaffrath 2018). Consistent with this notion, the structure of suppressor tRNA in the ribosomal A-site resolved in *E. coli* showed that stop codon
decoding by suppressor tRNA structurally resemble elongating tRNA rather than release factors during termination (Albers et al. 2021).

Challenges for suppressor tRNA therapeutics include its delivery and efficacy in vivo in rescuing protein function. Earlier studies simply used direct injection and/or electroporation of suppressor tRNAs into specific tissues of mice, including muscles and hearts (Buvoli et al. 2000; Kiselev et al. 2002; Lueck et al. 2019). Their in vivo efficacies were then evaluated by monitoring levels and activities of readthrough products of the co-injected nonsense-containing reporters (Buvoli et al. 2000; Lueck et al. 2019) or the actual disease-causing allele, such as dystrophin levels in mdx mice (a DMD model) (Kiselev et al. 2002). Co-injection of suppressor tRNAs and human nonsense-containing dystrophin cDNA into mouse muscles resulted in full-length dystrophin production in 2.5% of fibers (Kiselev et al. 2002). However, direct cDNA injection is not ideal in actual therapies because the tissue or organ of interest may not be accessible. Other approaches such as viral delivery have been discussed but not demonstrated until recently (Porter et al. 2021). One study delivered suppressor tRNAs packaged in recombinant adeno-associated virus (rAAV) vector through mouse tail vein injection (Wang et al. 2022b) while another delivered suppressor tRNAs encapsulated by lipid nanoparticles (LNPs) through intravenous injection (Albers et al. 2023).

Another huge concern regarding suppressor tRNA as a therapeutic strategy has been its off-target effects on NTCs. Multiple studies have investigated these effects at a transcriptome-wide level using ribosome profiling with mixed results. In
a study by Lueck et al., most of their efficacious anticodon edited tRNAs (ACE-tRNAs) identified in a high-throughput cell-based screen appeared to have only small effects on NTCs in HEK293 cells, with the highest effect being a two-fold increase in NTC readthrough compared to more than 100-fold for PTC readthrough (Lueck et al. 2019). Wang et al. found that their UAG-suppressor tRNAs have negligible effects on NTC readthrough in vitro (patient-derived cells) but induced NTC readthrough in vivo (liver from treated mice) in a UAG-specific manner (Wang et al. 2022b). Albers et al. found that the number of transcripts undergoing NTC readthrough upon suppressor tRNA introduction is similar to that at basal level from ribosome profiling experiments derived from whole lungs and livers of mice (Albers et al. 2023). Despite how minimal the effects on NTC readthrough are compared to PTC readthrough, the sustained minimal NTC readthrough of certain endogenous mRNAs may still be toxic to cells and careful evaluations of long-term effects of suppressor tRNA therapeutics are still needed.

*Interference with the translation machinery during termination*

Some readthrough-promoting compounds can induce readthrough by interfering with the translation machinery during translation termination. The most well studied class of compounds with this mechanism are the aminoglycosides, a class of antibiotics that strongly bind the prokaryotic ribosome’s decoding center, leading to misincorporation of near-cognate tRNAs at both sense and nonsense codons and defects in protein synthesis (Keeling et al. 2014). Although aminoglycosides preferentially bind prokaryotic ribosomes, some of them can bind
eukaryotic ribosomes and thus have been explored extensively as readthrough-inducing agents (Keeling et al. 2014). These include but are not limited to G418 (geneticin), gentamycin, paromomycin, amikacin, kanamycin, and neomycin, and their activity has been established in numerous in vitro and in vivo experiments (Martins-Dias and Romão 2021; Li et al. 2023).

Among these drugs, G418 is the most potent and has been commonly used as a benchmark for other readthrough-inducing drug candidates. However, G418 itself has not reached clinical application due to its cytotoxicity and its induction of transcriptome-wide NTC readthrough (Wangen and Green 2020; Schilff et al. 2021; Li et al. 2023). Gentamicin is another commonly studied aminoglycoside that is less potent than G418 and also less toxic. Gentamicin has been tested in pre-clinical disease models and reached clinical trial stages for various diseases (Wilschanski et al. 2000; Du et al. 2002; Wilschanski et al. 2003; Martins-Dias and Romão 2021; Mosallaei et al. 2022; Martínez-Santamaría et al. 2022; Li et al. 2023), but the results were mixed and overall improvement of disease phenotypes only occurred in a fraction of patients. Since aminoglycosides have been shown to induce readthrough in a stop codon context-dependent manner (Howard et al. 2000; Manuvakhova et al. 2000; Floquet et al. 2012), it is possible that patients with permissive contexts may respond more positively than those with non-permissive contexts. This aspect of readthrough therapeutics needs to be explored further.
Several strategies to improve the outcome of aminoglycoside treatment have been considered. Synthetic aminoglycoside derivatives have been developed that maintain or increase potency while reducing toxicity (Nudelman et al. 2009, 2010; Vecsler et al. 2011; Shulman et al. 2014; Bidou et al. 2017; Kerem 2020; Popadyynec et al. 2021). One aminoglycoside derivative, ELX-02, has reached phase I and II clinical trial stages for treatment of CF (NCT04126473 and NCT04135495) and Alport syndrome (NCT05448755). Additionally, novel compounds than can enhance the efficacy of existing aminoglycosides have been identified and tested against disease-causing alleles in cells with promising results (Baradaran-Heravi et al. 2016; Wong et al. 2023).

Non-aminoglycoside compounds have also been examined for their readthrough-inducing potential, one of which, ataluren (PTC124), has been shown to be quite effective across various disease models (Du et al. 2008; Finkel 2010; Li et al. 2014; Bedwell et al. 2015; Long et al. 2020; Wang et al. 2022a; Wu et al. 2022). Ataluren has been tested in multiple clinical trials (Hirawat et al. 2007; Kerem et al. 2008; Finkel et al. 2013; Kerem et al. 2014) (comprehensive lists and reviews are in Martins-Dias and Romão 2021; Li et al. 2023), many of which are on-going or recently started (e.g., NCT04336826), and the drug (marketed as Translarna) was approved for treatment of DMD in Europe. Ataluren was identified in a high-throughput screen of ~800,000 small molecules for readthrough induction of nonsense-containing luciferase reporters (Welch et al. 2007). Importantly, ataluren is bioavailable and induces readthrough of PTCs without apparent effects.
on NTCs (Welch et al. 2007; Peltz et al. 2013), making it a safer drug than aminoglycosides in that respect. The exact molecular mechanism by which ataluren promotes readthrough is still unclear, but all emerging lines of evidence are related to the translation machinery. Unlike single, strong binding of aminoglycosides to the ribosome, ataluren (and ataluren-like compounds) binds multiple weaker sites, including two sites on rRNA close to the ribosome’s decoding center and peptidyl transferase center, and a third site on the release factor complex (Ng et al. 2018; Huang et al. 2022). Ataluren appears to inhibit release factor complex activity before or at the peptide hydrolysis step without affecting near-cognate tRNA complex binding (Huang et al. 2022). This result is consistent with a previous study showing that while ataluren induces PTC readthrough, the identity and proportion of amino acids inserted at PTCs are the same between endogenous basal readthrough in untreated cells and ataluren-mediated readthrough (Roy et al. 2016).

Despite its success in DMD patients, ataluren has not resulted in satisfying improvement of other diseases, likely due to an insufficient amount of readthrough product being made to reach the threshold for being beneficial (Konstan et al. 2020). The efficacy of ataluren-mediated readthrough appears to be sequence context-dependent, so a deeper understanding of how ataluren treatment is influenced by stop codon context may be needed. To increase potency, several derivatives of ataluren have been developed with promising preliminary results (Pibiri et al. 2015, 2016, 2018; Moosajee et al. 2016; Bezzerrri et al. 2022).
Therapeutic strategies targeting NMD: increasing the amount of mRNAs available for translation

A major challenge of nonsense suppression therapy is to increase the level of full-length readthrough protein sufficiently to be beneficial to cells. Thus, a supplemental approach to increasing readthrough efficiency is to increase the level of mRNAs available for readthrough induction by antagonizing NMD. I discuss NMD inhibitors in two broad strategies: i) depletion of NMD factors and ii) interference with the functions of NMD factors or translation machineries.

Depletion of NMD factors

As with depletion of release factors to improve PTC readthrough, depletion of NMD factors has been attempted to increase the level of PTC-containing mRNAs. As one of the core factors, Upf1 has been the most frequent target for early in vitro studies using siRNA- and ASO-mediated depletion in models of multiple diseases, including Ullrich congenital muscular dystrophy, cystic fibrosis, DMD, LQT2 syndrome, and hemophilia B (Usuki et al. 2006; Linde et al. 2007; Gong et al. 2011; Huang et al. 2018). In all cases, the PTC-containing mRNAs were upregulated and/or functional expression of full-length proteins were restored (Usuki et al. 2006; Linde et al. 2007; Gong et al. 2011; Huang et al. 2018). In addition to Upf1, knocking down of Upf2, Upf3, components of the Smg complex, and release factors have been systematically tested (Usuki et al. 2013; Huang et al. 2018). While Upf1 and Smg1 depletion produced the best results, the effects were shown to be cytotoxic and cause embryonic lethality (Usuki et al. 2013;
Martins-Dias and Romão 2021). In the Ullrich disease study, the factor that produced the best results while having no cytotoxicity among the 15 NMD factors tested in human cells was a subunit of the Smg1 complex, Smg8 (Usuki et al. 2013). In another study that screened 10 mouse NMD factors, Upf3b knockdown was found to be well tolerated in normal mice while sufficiently stabilizing PTC-containing mRNA levels in respective DMD and hemophilia B mouse models (Huang et al. 2018).

In addition to RNA interference, depletion of all Upf proteins by a compound has also been demonstrated. Curcumin, a natural compound found in turmeric with therapeutic potential for a variety of diseases, was found to reduce the occupancy of histone H3 and RNA polymerase II at the promoter region of human Upf1, subsequently reducing its transcription (Feng et al. 2015). Consequently, PTC-containing mRNAs were stabilized.

**Interference with the functions of NMD factors or translation machineries**

Because NMD is dependent on translation and the occurrence of premature termination, inhibitors of any of the steps of protein synthesis can also be NMD inhibitors. Many nonsense suppression approaches, such as eRF3b degraders (Baradaran-Heravi et al. 2021), pseudouridylation (Adachi et al. 2023), and suppressor tRNAs (Albers et al. 2023), already have been shown to also inhibit NMD, possibly indirectly through decreasing the number of premature termination events that would trigger NMD. Translation inhibitors that bind ribosomes such as cycloheximide, puromycin, and emetine can inhibit NMD, but they are toxic
(Martins-Dias and Romão 2021). Pateamine A, a small molecule compound that interacts with initiation factor eIF4A, inhibits NMD by inhibiting translation initiation as well as interfering with eIF4A3-containing EJC complex (Dang et al. 2009; Naineni et al. 2023), but it is also toxic.

There are efforts to identify direct inhibitors of NMD. Inhibition of Smg1 has been demonstrated for caffeine, wortmannin, and pyrimidine derivatives (Usuki et al. 2004; Gopalsamy et al. 2012; Keeling et al. 2013). Unfortunately, caffeine and wortmannin are non-specific, as they can also affect other kinases (Usuki et al. 2004; Martins-Dias and Romão 2021), while modifications to pyrimidine derivatives to be Smg1-selective have been described (Gopalsamy et al. 2012). Regardless, caffeine has been tested in combination with a readthrough-promoting drug. NMD attenuation by caffeine increased PTC-containing CFTR mRNA level and, in combination with ataluren, improved recovery of CFTR full-length protein in vitro (Lentini et al. 2019). Several other compounds were identified through small molecule screens. NMDI-1 changes cellular distribution of Upf1, stabilizes its hyperphosphorylated form, and blocks its interaction with Smg5 (Durand et al. 2007). NMDI-1 was shown to be effective in a mouse model of the lysosomal storage disease mucopolysaccharidosis I-Hurler (MPS I-H), increasing PTC-containing Idua mRNA level and enhancing activity of alpha-L-iduronidase (product of Idua) when combined with gentamicin (Keeling et al. 2013). In another study, NMDI14 was identified as an inhibitor of Upf1-Smg7 interaction, rescuing
full-length p53 protein expression when used in combination with G418 in vitro (Martin et al. 2014).

**Challenges for current therapeutic strategies**

Since both translation termination and NMD are important global cellular processes, the biggest challenges to overcome in developing nonsense suppression and NMD inhibition as therapeutic strategies are off-target effects. While NMD inhibition stabilizes the mRNA of a gene harboring a nonsense mutation, hundreds of other endogenous NMD substrates, including potentially faulty mRNAs arising from upstream processing errors, would also accumulate and would have the potential to be harmful to cells. Similarly, while nonsense suppression induces readthrough of PTCs, readthrough of perhaps thousands of NTCs might also occur albeit at lower levels in most cases. Although the much smaller increases in NTC readthrough compared to the increase in PTC readthrough promoted by current treatments is encouraging, long-term effects need to be carefully studied to ensure safety.

The success of ataluren in treatment of nonsense mutations in DMD patients suggests that there may be an optimal readthrough level along the scale of potency vs. safety, and this effect varies for different diseases and mutation contexts. Ataluren may not be as potent as aminoglycosides, but it was sufficient to rescue the level of full-length DMD proteins required to improve disease phenotypes while not potent enough to induce detectable NTC readthrough and cause toxicity. Unfortunately, the levels of full-length proteins synthesized upon
ataluren treatment for other diseases are not sufficient to rescue the disease phenotypes. Optimal effects may exist for other readthrough-inducing compounds matched with specific diseases.

Among all the major approaches discussed for nonsense suppression, only pseudouridylation has a site-specific modification system which can ensure readthrough of a specific PTC on a specific mRNA, reducing off-target toxicity (though the delivery method may be a challenge). However, another challenge for pseudouridylation and all other nonsense suppression therapies except for engineered tRNAs is that the amino acid incorporated at the PTC may not be the same one that exists in wild-type protein. Depending on protein chemistry, this may be tolerated or deleterious. Fortunately, it may be possible to determine beforehand how often the desired amino acid will be inserted and whether the treatment will be appropriate for specific patients based on their PTC contexts because these contexts have been shown to influence which tRNA decodes the stop codon and thus the amino acid incorporated into the protein. The challenge of inserting the correct amino acid is overcome by engineering suppressor tRNAs; however, this approach again is not PTC-specific and can produce off-target effects.

An ideal therapy would be one that is safe, potent, PTC site-specific, and produces the wild-type protein. Thus, a deeper understanding of translation termination and NMD is needed to further fine-tune current therapeutic strategies as well as facilitate development of novel strategies for treatment of diseases.
caused by nonsense mutations. Specifically, the mechanistic differences between PTCs and NTCs and how NMD distinguishes between them will be beneficial.

**Scope of thesis**

This dissertation describes my work in identifying specific properties of mRNA sequences that determine termination efficiency and recognition of premature termination by NMD at a transcriptome-wide level. The efficiency of termination at PTCs appears to be less efficient than at NTCs, but what contributes to this efficiency difference has not been demonstrated fully at a global level. In Chapters II and III, I investigate the effects of cis-acting elements on termination efficiency through comprehensive bioinformatics analyses of extended stop codon contexts in relation to endogenous NTC readthrough efficiencies measured by ribosome profiling of yeast (Chapter II) and human cells (Chapter III) in normal and readthrough-inducing conditions. I expand the roles of many known elements to a global scale and their conservation across species as well as identify novel regulators of readthrough. In Chapter III, my collaborators and I also demonstrate the value of using machine learning models trained on readthrough efficiency from ribosome profiling data to predict readthrough efficiency of new unseen sequence context.

In Chapter IV, I present a collaborative work describing early ribosome-associated steps of NMD and properties of mRNAs associated with Upf1-bound ribosomes. Using selective ribosome profiling, we discover two distinct Upf1-ribosome complexes, one of which is dependent on Upf1’s ATPase function, that
accumulate differently across mRNA ORFs. We find that Upf1 likely already binds ribosomes prior to the occurrence of premature termination, acting as a surveillance factor, rather than being recruited upon premature termination.
CHAPTER II:

Transcriptome-wide \textit{cis}-regulatory elements of normal
termination codon readthrough in yeast

The work in this chapter has been published as:

Mangkalaphiban, K. \textit{et al}. Transcriptome-wide investigation of stop codon
Abstract

Translation of mRNA into a polypeptide is terminated when the release factor eRF1 recognizes a UAA, UAG, or UGA stop codon in the ribosomal A site and stimulates nascent peptide release. However, stop codon readthrough can occur when a near-cognate tRNA outcompetes eRF1 in decoding the stop codon, resulting in the continuation of the elongation phase of protein synthesis. At the end of a conventional mRNA coding region, readthrough allows translation into the mRNA 3’-UTR. Previous studies with reporter systems have shown that the efficiency of termination or readthrough is modulated by cis-acting elements other than stop codon identity, including two nucleotides 5’ of the stop codon, six nucleotides 3’ of the stop codon in the ribosomal mRNA channel, and stem-loop structures in the mRNA 3’-UTR. It is unknown whether these elements are important at a genome-wide level and whether other mRNA features proximal to the stop codon significantly affect termination and readthrough efficiencies in vivo. Accordingly, we carried out ribosome profiling analyses of yeast cells expressing wild-type or temperature-sensitive eRF1 and developed bioinformatics strategies to calculate readthrough efficiency, and to identify mRNA and peptide features which influence that efficiency. We found that the stop codon (nt +1 to +3), the nucleotide after it (nt +4), the codon in the P site (nt -3 to -1), and 3’-UTR length are the most influential features in the control of readthrough efficiency, while nts +5 to +9 had milder effects. Additionally, we found low readthrough genes to have shorter 3’-UTRs compared to high readthrough genes in cells with thermally
inactivated eRF1, while this trend was reversed in wild-type cells. Together, our results demonstrated the general roles of known regulatory elements in genome-wide regulation and identified several new mRNA or peptide features affecting the efficiency of translation termination and readthrough.

**Introduction**

Translation of an mRNA into a polypeptide is terminated when the release factor eRF1 interacts with a UAA, UAG, or UGA stop codon in the ribosomal A site and another release factor, eRF3, hydrolyzes GTP and stimulates the polypeptide release activity of eRF1 (Hellen 2018; Salas-Marco and Bedwell 2004; Schuller and Green 2018). However, at low frequency, a near-cognate tRNA (nc-tRNA; a tRNA with one base pair mismatch in its anticodon) outcompetes eRF1 in decoding the stop codon, resulting in the continuation of translation elongation. When the stop codon is located at its normal site at the end of an open reading frame (ORF) elongation will continue into the mRNA 3’-untranslated region (3’-UTR), producing a C-terminally extended polypeptide product. Such events are termed stop codon readthrough or nonsense suppression (Hellen 2018; Dabrowski et al. 2015; Rodnina et al. 2020; Brenner et al. 1965).

The process of readthrough was first characterized primarily in viruses, which utilize this mechanism to increase the protein-coding capacity of their compact genomes (Schueren and Thoms 2016; Skuzeski et al. 1991). It is only quite recently that C-terminally extended yet functional protein isoforms have been discovered in the proteomes of higher eukaryotes, and their expression is often
attributable to readthrough levels that are 10 to 100-fold higher than basal readthrough frequencies (De Bellis et al. 2017; Eswarappa et al. 2014; Hofhuis et al. 2016; Loughran et al. 2014, 2018; Namy et al. 2002; Rajput et al. 2019; Schueren et al. 2014; Steneberg and Samakovlis 2001; Yamaguchi and Baba 2018). In addition, ribosome profiling and phylogenetic investigation of flies, humans, and yeast cells revealed that readthrough is more prevalent than previously anticipated, with efficiencies varying by more than 100-fold for distinct mRNAs (Loughran et al. 2014; Dunn et al. 2013; Jungreis et al. 2011; Kleppe and Bornberg-Bauer 2018). These observations indicate that the efficiency of translation termination can be subjected to transcript-specific regulation, and that a key to understanding this regulation may lie in specific cis-acting mRNA sequences.

Previous studies on termination using reporter systems mainly relied on the inverse relationship between termination and readthrough efficiencies, where the level of readthrough protein product was quantified to infer the efficiency of termination. These studies have shown that the extent of readthrough, and thus the efficiency of termination, is modulated by multiple cis-acting elements (Dabrowski et al. 2015; Rodnina et al. 2020; Tate et al. 2018). The three stop codons themselves differ in termination efficiency, with UGA resulting in the most readthrough, and UAA the least (Bonetti et al. 1995; Loughran et al. 2014). Nucleotides in the immediate vicinity of the stop codon, including two nucleotides 5’ of the stop codon (Mottagui-Tabar et al. 1998; Tork et al. 2004) and up to six
nucleotides 3’ of the stop codon in the ribosomal mRNA channel (Skuzeski et al. 1991; Schueren et al. 2014; Anzalone et al. 2019; Cridge et al. 2018; Namy et al. 2001), have been shown to modulate readthrough efficiency. Additionally, stem-loop structures in the mRNA 3’-UTR are enriched in readthrough-prone transcripts (Steneberg and Samakovlis 2001; Jungreis et al. 2011; Anzalone et al. 2019; Firth et al. 2011). It remains to be determined whether the effects of these elements are applicable at a genome-wide level for endogenous mRNAs in vivo. Transcriptome-wide ribosome profiling studies of flies, human, and yeast cells have detected readthrough of individual mRNAs (Dunn et al. 2013; Kleppe and Bornberg-Bauer 2018; Ingolia et al. 2009; Li and Zhang 2019; Wangen and Green 2020), but detailed dissection of the relationships between readthrough efficiency and cis-acting mRNA sequences have only been minimally investigated.

While the stop codon and its immediate 3’ flanking sequences have been studied quite extensively, with structural evidence showing certain nucleotide preferences for optimal interactions with eRF1 motifs and rRNAs (Brown et al. 2015; Shao et al. 2016; Tate et al. 2018), other proximal regions have not. Novel readthrough regulatory elements are likely to exist because readthrough has been observed in genes that do not have any of the known readthrough-promoting signals (although there are genes that have readthrough-promoting elements, but do not show detectable readthrough experimentally) (Eswarappa et al. 2014; Jungreis et al. 2011; Rajput et al. 2019). Based on studies of other translational control events, such as ribosome stalling (Hellen 2018; Wilson et al. 2016), several
other mRNA features could affect termination and readthrough. The nascent peptide sequences in the ribosomal exit tunnel may affect termination or readthrough efficiency via their interference with the peptidyl transferase center. The proximity of the stop codon to the poly(A)-binding protein (PABP) is also thought to influence readthrough, as PABP is known to interact with eRF3 and enhance termination both in vitro and in vivo (Ivanov et al. 2016; Swart et al. 2016; Wu et al. 2020). These possibilities have yet to be explored with regard to normal termination at a global scale.

Accordingly, we have carried out ribosome profiling of yeast cells expressing a wild-type or temperature-sensitive mutant allele of eRF1 (Sup45 in yeast) and developed bioinformatics strategies to measure readthrough efficiency of individual mRNAs on a genome-wide scale, and to identify mRNA and peptide features which influence that efficiency. Our results demonstrated the general roles of known regulatory elements, such as the identities of the stop codon and the surrounding nucleotides, in genome-wide regulation and identified several new mRNA features that appear to play a role in translation termination and readthrough, including the penultimate codon in the P site (when a stop codon is in the A site) and the length of the 3’-UTR.
Results

**Inactivation of eRF1 promotes readthrough of both normal and premature stop codons and accumulation of ribosomes at those codons**

We performed ribosome profiling and total RNA sequencing of isogenic wild-type (WT) and eRF1 mutant yeast cells. The latter harbor the *sup45*-2 (subsequently referred to as *sup45*-ts) mutation that minimally affects eRF1 function when cells are grown at 25°C, but renders eRF1 inactive within 30 minutes of growth at 37°C (Stansfield et al. 1997). Reads between replicates were reproducible, with Pearson’s correlation coefficients (r) of 0.96 and 0.99 on average for ribosome profiling and RNA-Seq, respectively (Figure 2.1). In addition to the profiling data obtained from these experiments, we also analyzed published ribosome profiling data obtained from yeast cells depleted of eRF1 via specific transcription shutoff (Wu et al. 2019) (Table 2.1). For comparison to ribosome occupancies in the 3’-UTR that are not due to termination defects, we included analyses of ribosome profiling data obtained from cells that are depleted of the ribosome recycling factor, Rli1 (Young et al. 2015) (Table 2.1). In these cells, full-length newly synthesized protein is properly released, but the ribosomes fail to recycle, then reinitiate randomly downstream of the stop codon, and generally produce a peptide independent of the original ORF (Young et al. 2015). To evaluate 3’-UTR translation, ribosome footprints mapped to genes that have no experimental UTR annotations and genes that overlap by more than 18 base pairs on the same mRNA strand (where read assignment is ambiguous) were discarded.
from further analyses. With these filters, we were left with 2,693 genes, or ~41% of the yeast genome, for analyses. This set of genes is referred as the reference gene set.

**Figure 2.1.** Reproducibility between replicates of *sup45* temperature-shift ribosome profiling and RNA-Seq datasets.
Correlation between ribosome density (top) or mRNA abundance (bottom) in reads per kilobase per million (RPKM) of a pair of replicates for each yeast strain and growth temperature. Pearson’s correlation coefficient ($r$) was calculated and reported for each pair of replicates.
Table 2.1. Ribosome profiling data and yeast strains.

<table>
<thead>
<tr>
<th>Sample</th>
<th>SRR Run</th>
<th>Yeast strain Notation in this paper</th>
<th>Genotype</th>
<th>Culture media</th>
<th>CHX in lysis buffer</th>
</tr>
</thead>
<tbody>
<tr>
<td>eRF1 temperature-sensitive mutant (This study): GSE162780</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM4959785, GSM4959786, GSM4959787, GSM4959788</td>
<td>SRR13208091, SRR13208092, SRR13208093, SRR13208094</td>
<td>WT (HFY114)</td>
<td>SUP45 MATa ade2-1 his3-11,15 leu2-3,112 trp1-1 ura3-1 can1-100</td>
<td>YPD</td>
<td>No</td>
</tr>
<tr>
<td>eRF1 ts mutant (HFY1218)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM4959789, GSM4959790, GSM4959791, GSM4959792</td>
<td>SRR13208095, SRR13208096, SRR13208097, SRR13208098</td>
<td>sup45-ts</td>
<td>MATa ade2-1 his3-11,15 leu2-3,112 trp1-1 ura3-1 can1-100 sup45-2</td>
<td>YPD</td>
<td>No</td>
</tr>
<tr>
<td>eRF1 depletion (Wu et al., 2019): GSE115162</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM3168380, GSM3168381</td>
<td>SRR7241903, SRR7241904</td>
<td>WT (yCW30)</td>
<td>SUP45-D MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 HO::ADH1p-OsTIR1-URA3</td>
<td>YPGR → YPD + 0.5mM auxin</td>
<td>Yes</td>
</tr>
<tr>
<td>eRF1 depleted (yKW13)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM3168385, GSM3168386</td>
<td>SRR7241908, SRR7241909</td>
<td>eRF1 depleted (yKW13)</td>
<td>sup45-d MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 KanMX4::P_GAL1-SUP45</td>
<td>YPGR → YPD</td>
<td>Yes</td>
</tr>
<tr>
<td>Rli1 depletion (Young et al., 2015): GSE69414</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM1700885</td>
<td>SRR2046309, SRR2046310</td>
<td>WT (BY4741)</td>
<td>RL1Δ-D MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0</td>
<td>YPG → YPD</td>
<td>Yes</td>
</tr>
<tr>
<td>Rli1 depleted (YDH369)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSM1700886, GSM1700891</td>
<td>SRR2046311, SRR2046312, SRR2046319</td>
<td>Rli1 depleted (YDH369)</td>
<td>rl1-d MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 lys2Δ0 rli1Δ::kanMX4 pDH181 (P_GAL1-UBI-R-FH-RL11, LEU2)</td>
<td>YPG → YPD</td>
<td>Yes</td>
</tr>
</tbody>
</table>

To explore the consequence of eRF1 inactivation on translation, ribosome footprints were mapped relative to their respective start or stop codons based on their P-site locations and quantified (Figure 2.2A-D). In all strains, footprints in the coding (CDS) region showed 3-nt periodicity, a profile expected of translating ribosomes. The absence of footprints at the penultimate codon of the open reading frames (ORFs) in SUP45 cells at 25°C indicated that, with the stop codon in the A site, termination was completed, and ribosomes were dissociated from the mRNAs before they could be captured. The presence of footprints at this position in sup45-ts cells at 25°C suggested that the mutation slowed down termination enough for the ribosomes to be captured. When eRF1 was inactivated in sup45-ts cells grown at 37°C, or depleted in sup45-d cells, a high number of footprints was observed at
the penultimate codon, indicating that ribosomes were stalling as they awaited successful termination. Similarly, the high number of footprints in *rli1-d* cells was also consistent with ribosome stalling, as the stop codon had entered the ribosomal A site, but a factor needed for successful recycling was lacking. Additionally, the peaks at approximately 30 nt upstream of the stop codon in *sup45-ts* cells at 37°C and *sup45-d* samples identified ribosomes stacking against those stalled at the stop codon. As a consequence of failed termination or recycling, the mutant libraries, especially from the *sup45-ts* cells at 37°C, had increased footprint reads in the 3′-UTR region compared to those of WT samples (Figure 2.2A-D, insets), implying stop codon readthrough, frameshifting, or reinitiation events. Although inactivation (*sup45-ts* cells at 37°C) and depletion (*sup45-d*) of eRF1 both caused loss of eRF1 function, these two different strategies likely resulted in loss of eRF1 function to different extent and yielded slightly different metagene profiles. While the *sup45-ts* strain lost normal eRF1 function within 30 minutes of the temperature shift, the *sup45-d* strain needed 9 hours of growth after transcription shutoff to gradually dilute the functional eRF1 protein level. In eRF1 depleted cells, although the level of functional eRF1 was significantly decreased, small amount of eRF1 still existed and was fully capable of promoting termination. The residual functional eRF1 in *sup45-d* cells likely resulted in a diminished quantity of 3′-UTR reads compared to those obtained with the *sup45-ts* strain. In addition to increased reads near the stop codons, all the mutant samples also showed increased footprint reads at the start codon and decreased footprint reads in the CDS compared to
their WT counterparts. These observations are in agreement with the notion that termination and recycling steps are linked to translation initiation (Schuller and Green 2018).

**Figure 2.2.** Ribosome occupancy at and beyond the stop codon increases in yeast cells defective in translation termination or ribosome recycling.  
**A-D.** Ribosome footprints (2,693 genes with 3’-UTR annotations and without sequence overlapping) normalized to the total number of footprints in the indicated nt window (-20 to 40 nt around the start codon and -40 to 80 nt around the stop codon) aligned at their start or canonical stop codons. Footprints were plotted by the position of their P-sites. Each panel contains the data from WT and its respective mutant. The elevated number of footprints at the penultimate codon of the ORFs in mutant samples demonstrates ribosome stalling when the stop codon was in the A site. (The footprints at this codon and their shift to reading frame +1 in other yeast strains obtained from published data may be attributable to the differences in strain background or sequencing library preparation procedures.) Inset: Magnified view of the region in the box, showing increased 3’-UTR ribosome occupancy in sup45-ts cells at 37°C, sup45-d, and rli1-d cells relative to their respective WT.
We also examined the effects of eRF1 inactivation on premature translation termination. Two alleles in our strain background, ade2-1 and can1-100, contain premature termination codons (PTCs). Their gene specific profiles revealed an accumulation of ribosomes 5’ to the respective PTCs and increased ribosome density in the part of the CDS following the PTCs in sup45-ts cells grown at 37°C (compared to sup45-ts at 25°C or to WT cells at either temperature) (Figure 2.3A-B). This observation is further supported by computing the ratio of in-frame footprint counts downstream vs. upstream of PTCs, a measurement of readthrough efficiency that automatically takes into account the differences in mRNA levels between samples (Figure 2.3C). For the PTC in ade2-1, sup45-ts cells grown at 37°C showed ~32-fold and 17-fold higher readthrough efficiencies relative to WT cells grown at 37°C and sup45-ts cells grown at 25°C, respectively. For can1-100, sup45-ts cells grown at 37°C showed >100-fold and 26-fold higher readthrough efficiencies relative to WT cells grown at 37°C and sup45-ts cells grown at 25°C, respectively. Together, these data show that readthrough of PTCs also occurs in the absence of functional eRF1. Further evidence for PTC readthrough includes the gene-specific profiles for RPL28 and RPS0A (Figure 2.3 D-E). Intron-containing transcripts from both of these genes enter the cytoplasm and their translation stalls and triggers NMD (nonsense-mediated mRNA decay) when ribosomes encounter PTCs within the respective introns (He and Jacobson 2015b; Celik et al. 2017a). Figure 2.3D-E shows that, for both transcripts, ribosome profiling reads that were generally absent from the respective intron regions in
sup45-ts cells at 25°C or in WT cells at either temperature were abundant in sup45-ts cells grown at 37°C and accumulated mostly at in-frame stop codons (yellow rectangles).
Figure 2.3. Evidence for readthrough of premature termination codons (PTCs) upon eRF1 inactivation.

A-B. Ribosome footprints normalized to the respective mRNA levels mapped across the ORF of two PTC-containing alleles, ade2-1 (A) and can1-100 (B), in SUP45 and sup45-ts strains at 25°C and 37°C.

C. Average readthrough efficiency of PTC readthrough in the ade2-1 and can1-100 alleles. PTC readthrough efficiency was calculated by dividing frame 0 footprint count downstream of the PTC to that upstream of the PTC. Average ± standard deviation of two replicates were plotted for each sample, with black dots indicating the individual data points.

D-E. Read coverage tracks from the Integrative Genomics Viewer (IGV) (Robinson et al. 2011) showing coverage of ribosome profiling reads for the intron-containing genes, RPL28 (D) and RPS0A (E) in SUP45 and sup45-ts strains at 25°C and 37°C. Yellow rectangles indicate the position of termination codons in frame with the respective initiation codons under conditions where the introns are translated. Full scale for D and E equals 50 reads.

To determine whether the increased ribosome density in regions following the stop codons in eRF1 mutants was due to stop codon readthrough rather than frameshifting or reinitiation events, we calculated the fraction of reads in each of the three reading frames in different mRNA regions: 5′-UTR, CDS, the 3′-UTR region between the canonical stop codon and the first downstream in-frame stop codon (“extension”), and the rest of the 3′-UTR (“distal 3′-UTR”) (Figure 2.4A). If readthrough occurred, we expected i) the amount of reading frame 0 footprints, which represent in-frame translation in the CDS, to also be the dominant reading frame in the extension and ii) the amount of frame 0 footprints in the extension from sup45-ts cells grown at 37°C to be higher than that of their WT counterparts. Indeed, we saw that frame 0, which was dominant in the CDS, was also dominant in the extension, but not in the distal 3′-UTR of most strains (Figure 2.4B). It is notable that the amount of reading frame 0 footprints in the extension region was less pronounced than that in the CDS, possibly because other recoding or
recycling events, such as frameshifting or reinitiation, could also happen at the stop codon that contributed to the fraction of out-of-frame footprints. Nevertheless, compared to the reading frame fraction in the extension region of their WT counterparts, sup45-ts at 37°C displayed notable enrichment of frame 0 reads while sup45-d and sup45-ts at 25°C showed only slight increases of frame 0 reads relative to their WT counterparts (Figure 2.4B). On the other hand, rli1-d had a reading frame preference comparable to its WT counterpart (Figure 2.4B). These results indicate that, in cells with inactivated eRF1, a considerable proportion of footprints in the extension region of the mRNAs are caused by stop codon readthrough.

In summary, the increased in-frame footprints in the extension region of the 3'-UTR in eRF1 mutants indicated that a lack of functional eRF1 led to stop codon readthrough globally. Hence, we developed formal criteria for the quantitation of readthrough of individual mRNAs and a bioinformatic scheme to assess the genome-wide positive and negative cis-acting effects of neighboring sequences to the readthrough of stop codons.

**Figure 2.4.** Readthrough events were characterized by maintenance of reading frame when ribosomes bypass the stop codon. 
**A.** Example of ribosome footprint count plot with labeled mRNA regions of interest. **B.** Fractions of reads in each of the three reading frames for different mRNA regions. The results of each sample were the average of 2 or 3 replicates, with black dots indicating individual data points. The dashed line drawn at 0.33 is the theoretical fraction at which all 3 reading frames are represented in equal amounts. 
(Next page)
Analysis pipeline development

To dissect the relationships between readthrough efficiency (Y variable) and mRNA or nascent peptide features (X variables), we used the random forest machine learning approach (Breiman 2001; Liaw and Wiener 2002) to first narrow down the candidate features. We defined each of the variables as follows:

**Defining readthrough efficiency (Y variable)**

Since termination efficiency is not directly measurable by ribosome profiling we calculated readthrough efficiency and utilized the generally inverse relationship of termination and readthrough (shown experimentally by Wu et al. (2020) (Wu et al. 2020)) to infer termination efficiency. Our subsequent interpretation of the results was based on this relationship, where high readthrough efficiency would point to low termination efficiency and vice versa in this experimental context. Readthrough efficiency in the context of ribosome profiling can be loosely defined as the ratio of ribosome density (footprint count normalized to the length of the region) in the 3’-UTR to ribosome density in the CDS; however, the nature of the mutants should be taken into account to ensure accuracy and minimize noise. First, due to ribosome stalling and queuing at the start and stop codons in mutant strains (Figure 2.2A), ribosome density in the CDS may be overestimated; therefore, we excluded footprints whose P-sites fall into the first 15 and the last 33 nucleotides of the CDS of each gene in all samples. Although the stalling of the ribosomes also occurred at the first downstream stop codon in the 3’-UTR, we could not use the same rules of footprint exclusion here because the lengths of
many extension regions are smaller than 33 nucleotides. Thus, the apparent readthrough level could be somewhat overestimated. Second, as readthrough is by definition in-frame with the CDS (Figure 2.4B), only footprints that were in frame 0 were considered. The adjusted readthrough efficiency calculation, log-transformed, is shown in Figure 2.5A.

Additional sources of noise in the data include: 1) poorly-expressed genes that result in an exaggeration of the readthrough efficiency ratios and 2) genes lacking footprints in the 3'-UTR region, for which it was challenging to distinguish lack of readthrough from insufficient sequencing depth. Inclusion of these genes confounded previous analyses and led to conflicting results. For instance, genes belonging in the “leaky” set (genes that had 3'-UTR reads, i.e. readthrough) were highly expressed compared to genes in the “non-leaky set” (genes lacking 3'-UTR reads), but within the “leaky” set, readthrough rate was negatively correlated with gene expression (Kleppe and Bornberg-Bauer 2018). Therefore, we discarded genes with both RPKM of CDS < 5 and RPKM of 3'-UTR < 0.5 (Young et al. 2015) from further analyses.

**Defining mRNA and nascent peptide features (X variables)**

To date, mRNA features affecting stop codon readthrough have frequently been identified 3’ of the stop codon, but we also considered the possibility that mRNA sequence on the 5’ side of the stop codon and the nascent peptide in the exit tunnel may contribute to variations in readthrough efficiencies (Figure 2.5B). These features can either be directly documented or inferred from the mRNA
sequences (Table 2.2). Therefore, we used established algorithms to predict RNA secondary structure (Lorenz et al. 2011) and nascent peptide $\alpha$-helix formation (Drozdetskiy et al. 2015).

\[
\text{Readthrough efficiency} = \log_2 \left( \frac{\text{# of frame 0 footprints}_{\text{extension}}}{\text{length (kb)}_{\text{extension}}} \div \frac{\text{# of frame 0 footprints}_{\text{Adjusted CDS}}}{\text{length (kb)}_{\text{Adjusted CDS}}} \right)
\]
Figure 2.5. Definition of X and Y Variables.
A. Readthrough efficiency calculation formula, which takes into account the ribosome pile-up at NTCs and reading frame preference. B. mRNA and nascent peptide features of interest in the context of the ribosome and related translation termination components. Previously identified variables are colored and labeled. Hypothesized variables are in purple and marked with “?”. Details of how each variable is measured are in Table 2.2. Information regarding the effects of particular nucleotides on readthrough efficacy is taken from a recent review, as is the general format of the schematic (Rodnina et al. 2020).

Table 2.2. Assessed X variables and their modes of measurement.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotides in the mRNA channel</td>
<td>Stop codon (nt +1 to +3) in the A site UAA, UAG, UGA</td>
</tr>
<tr>
<td></td>
<td>nt -1 to -15 and nt +4 to +9 A, C, G, U</td>
</tr>
<tr>
<td></td>
<td>Amino acids in the upper tunnel (aa 3-9 from PTC)</td>
</tr>
<tr>
<td></td>
<td>Amino acids at the constriction site (aa 10-12 from PTC)</td>
</tr>
<tr>
<td></td>
<td>Amino acids in the central tunnel (aa 13-19 from PTC)</td>
</tr>
<tr>
<td></td>
<td>Amino acids in the lower tunnel (aa 20-30 from PTC)</td>
</tr>
<tr>
<td>3’-UTR secondary structure</td>
<td>Tendency for 3’-UTR sequence to form a secondary structure The lowest minimum free energy (MFE) for a structure in 150 nt scanning window of entire 3’-UTR sequence, predicted by RNALFold function in ViennaRNA package (Lorenz et al. 2011)</td>
</tr>
<tr>
<td>Proximity of PABP to stop codon</td>
<td>Distance of structure from stop codon Number of nt from stop codon to the base of the structure</td>
</tr>
<tr>
<td>Negative controls</td>
<td>Distance of PABP (i.e., poly(A)-tail) to stop codon 3’-UTR length (nt)</td>
</tr>
<tr>
<td></td>
<td>Randomly assigned factor (categorical) A, C, G, U</td>
</tr>
<tr>
<td></td>
<td>Randomly assigned number (numeric) 1-100</td>
</tr>
</tbody>
</table>
Other than the mRNA features listed in Table 2.2, many characteristics of the mRNAs (such as expression levels, codon optimality, translation efficiency, and gene length) were previously correlated with readthrough efficiency measured by ribosome profiling (Kleppe and Bornberg-Bauer 2018; Li and Zhang 2019). We performed similar analyses involving pairwise correlation between readthrough efficiency, gene expression level (determined by RSEM of RNA-Seq data), translation efficiency (ribosome density divided by gene expression level), length of transcript and mRNA regions, and codon optimality (tRNA adaptation index (Pechmann and Frydman 2013; Reis et al. 2004)) (Figure 2.6). We found that, except for sup45-ts at 37°C, readthrough efficiency was negatively correlated with gene expression and codon optimality, consistent with previous results (Kleppe and Bornberg-Bauer 2018; Li and Zhang 2019). This correlation may have an evolutionary, rather than a mechanistic, explanation; for example, the inverse relationship between gene expression and readthrough efficiency may result from more deleterious effects of readthrough from highly expressed genes than from poorly expressed genes, and thus highly expressed genes evolved to have lower readthrough efficiency (Li and Zhang 2019). Since gene expression level was positively correlated with codon optimality across the ORF, where mRNAs with optimal codons are more stable and hence more abundant than those with non-optimal codons (Presnyak et al. 2015), we could not make a distinction whether the observed negative correlation between readthrough efficiency and codon optimality was simply due to increased mRNA abundance or was instead
mechanistically relevant to the readthrough process. We therefore did not include gene expression and codon optimality in further analyses. Moreover, readthrough measurements in general should be interpreted with caution and due consideration given the caveats described above.

**Figure 2.6.** Correlation matrix of relationships between readthrough efficiency, gene expression, translation efficiency, codon optimality, and transcript length. Spearman’s correlation coefficient (ρ) was calculated and reported for each pair of variables.
Random forest analyses reveal stop codon identity, P-site amino acid identity, and 3'-UTR length as important factors in predicting readthrough efficiency

In order to identify without bias the mRNA and nascent peptide features affecting readthrough efficiency, we used two independent approaches involving the random forest algorithm (Breiman 2001; Liaw and Wiener 2002): 1) random forest regression to predict readthrough efficiency based on the X variables, and 2) random forest classification to differentiate between “High” (top 15%) and “Low” (bottom 15%) readthrough genes based on the X variables. The number of genes in each type of analysis can be found in Table 2.3. A regression model and a classification model with 5-fold cross-validation were created for each sample using the same tuning parameters. For the regression model, we used Normalized Root Mean Squared Error (NRMSE) as a measurement of predictive ability of X variables. We observed NRMSE of approximately 0.14 across all samples, indicating that the average difference between predicted and actual readthrough efficiency is 14% (Figure 2.7A). For the classification model, Area Under the Receiver Operating Characteristics (AUROC) was used as a measurement of model capability at distinguishing between the “High” and “Low” groups. AUROC values ranging from 0.72-0.82, 0.76 on average, were observed across all samples; therefore, the models had ~76% chance of classifying genes into the correct group (Figure 2.7B).
Table 2.3. Number of genes involved in statistical analyses.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Regression</th>
<th></th>
<th>Classification</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High</td>
<td>Low</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>SUP45, 25°C</td>
<td>829</td>
<td>125</td>
<td>250</td>
<td></td>
</tr>
<tr>
<td>SUP45, 37°C</td>
<td>565</td>
<td>85</td>
<td>170</td>
<td></td>
</tr>
<tr>
<td>sup45-ts, 25°C</td>
<td>1347</td>
<td>202</td>
<td>404</td>
<td></td>
</tr>
<tr>
<td>sup45-ts, 37°C</td>
<td>2484</td>
<td>373</td>
<td>746</td>
<td></td>
</tr>
<tr>
<td>SUP45-D</td>
<td>975</td>
<td>147</td>
<td>294</td>
<td></td>
</tr>
<tr>
<td>sup45-d</td>
<td>2146</td>
<td>322</td>
<td>644</td>
<td></td>
</tr>
<tr>
<td>RLI1-D</td>
<td>937</td>
<td>141</td>
<td>282</td>
<td></td>
</tr>
<tr>
<td>rli1-d</td>
<td>2007</td>
<td>301</td>
<td>602</td>
<td></td>
</tr>
</tbody>
</table>

To determine which of the X variables were most responsible for the predictions, we permuted each X variable for the regression and classification models and assessed its importance by calculating either percent increase in mean squared error (%IncMSE) or Mean Decrease Accuracy (MDA). %IncMSE is the calculation of percent increase in prediction error, while MDA is the percent increase in misclassification, when the X variable was permuted. Thus, the higher the %IncMSE or MDA value is for an X variable, the more critical that variable is in predicting readthrough efficiency. We performed hierarchical clustering of these values across samples and found that in general, the eRF1 mutants were more similar to each other than to other strains in both regression and classification approaches (Figure 2.7C-D).

In order to determine which X variables exhibited meaningful %IncMSE or MDA values, we established their baselines (negative controls) by randomly assigning one numeric value between 1-100 (random number) and one of four categorical values (random factor) to each gene before model training. Because the assignment was random, these two variables should have no influence on the
prediction. Indeed, random number and random factor were among the X variables with %IncMSE and MDA values close to zero (Figure 2.7C-D). Reassuringly, in both the regression and classification approaches, the variables previously known to affect readthrough efficiency, such as the identity of a stop codon, nucleotide (nt) +4 and +5 were among the variables with %IncMSE and MDA values above the baseline in most samples, arguing that our pipeline was capable of distinguishing relevant mRNA features. In addition to the stop codon identities, 3’-UTR length and the P-site amino acid (or tRNA or codon) also had one of the highest %IncMSE and MDA values in most samples. The nt -3, and -2 in the P-site only had high MDA values in the ribosome recycling mutant, rli1-d, suggesting that they may play a role in recycling; however, this is only true in the classification model.

Interestingly, the first downstream stop codon in the 3’-UTR (“1st 3’-UTR stop” in Figure 2.7C-D) had high %IncMSE and MDA values only in sup45-ts at 37°C and sup45-d samples. This observation is due to ribosome pile-up at those stop codons, as with canonical stop codons (Figure 2.2B-C). The identity of this stop codon affects the read count in the extension region – its high %IncMSE and MDA value reinforces the role of stop codon identity in readthrough but not mechanistic relevance to readthrough occurring specifically at the canonical stop codon.
In summary, random forest analyses reveal identities of the stop codon, P-site amino acid (or tRNA or nucleotides), and 3’-UTR length as mRNA features that control genome-wide readthrough efficiency.
Figure 2.7. Random forest identified stop codon identity, P-site amino acid identity, and 3’-UTR length as factors critical in readthrough efficiency prediction. 

A. Performance metric for random forest regression displaying average Normalized Root Mean Squared Error (NRMSE) ± standard deviation across 5-fold cross-validation. B. Performance metric for random forest classification displaying average Area Under the Receiving Operator Curve (AUROC) ± standard deviation across 5-fold cross-validation. The value of 0.5 (dashed line) means the classification by the input X variables is no better than random chance. C-D. Feature importance extracted from the random forest models. The relative importance of a feature is represented by percent increase in mean squared error (%IncMSE) for regression (C) and Mean Decrease Accuracy (MDA) for classification (D), which is percent increase in error that results from permuting the feature. The higher the %IncMSE or MDA is, the more crucial the feature is in predicting readthrough efficiency or correctly classifying genes into groups. Arbitrary continuous and discrete values randomly assigned to the genes (random number and random factor) are used as negative controls.

Readthrough-promoting stop codons and nucleotides at positions +4 to +9 occur at a genome-wide level

Previous studies using synthetic dual reporters have shown that the UGA stop codon and certain nucleotides at positions -2, -1, +4 to +9 are the most permissive features for stop codon readthrough (Figure 2.5B) (Rodnina et al. 2020; Skuzeski et al. 1991; Schueren et al. 2014; Tate et al. 2018; Bonetti et al. 1995; Mottagui-Tabar et al. 1998; Tork et al. 2004; Anzalone et al. 2019; Cridge et al. 2018; Namy et al. 2001). Since the random forest analyses also identified the stop codon identity and some of these nucleotides as features that, when permuted, led to higher prediction error than baseline, we wondered whether these same stop codon and nucleotide identities were also readthrough-permissive or readthrough-inhibiting at a genome-wide level. To answer this question, the genes were divided into 3 groups for analysis of stop codon identity or 4 groups for analyses of
individual nucleotides at each position and Wilcoxon’s rank sum test was used to compare the median of genes in each group to the overall sample median. The differences between group median and sample median were shown as a heatmap (Figure 2.8A). The significance values from Wilcoxon’s rank sum test were represented by tile size, where a larger tile depicted p-value < 0.05. As expected, the distribution of random factor variables was not significantly different between the groups and sample median and showed no particular pattern of higher or lower readthrough efficiency across all samples.

The variables that showed the strongest pattern and significance were the stop codon and the nucleotide immediately downstream (nt +4). Consistent with the notion that UGA is the most readthrough-permissive stop codon and UAA is the strongest terminator (Bonetti et al. 1995), genes with UAA as stop codons had significantly lower readthrough efficiencies while genes with UGA had significantly higher readthrough efficiencies compared to the sample median in all samples except for \textit{rli1-d} cells. For nt +4, genes with C at this position had significantly higher readthrough efficiencies, and those with G had slightly lower readthrough efficiencies. These observations are in line with previous data where C was reported to allow the most readthrough and G the least (Rodnina et al. 2020; Bonetti et al. 1995; Anzalone et al. 2019; Cridge et al. 2018) and are supported by structural evidence showing preferential base stacking of purines at nt +4 with 18S rRNA, enhancing the stop codon recognition by eRF1 (Brown et al. 2015).
Figure 2.8. Known individual effects of stop codon and surrounding nucleotide identities on readthrough efficiency occur at a genome-wide level.

A. Heatmap of median readthrough efficiency of genes containing particular stop codon or nucleotide relative to median readthrough efficiency of all genes in the sample. Positive value (red) indicates that the group of genes had higher readthrough efficiencies compared to the sample median, while negative value (blue) indicates lower readthrough efficiencies. Wilcoxon’s rank sum test was used to determine whether the difference in group and sample media was significant. Significant difference was represented as a larger tile.

B. Heatmap demonstrating over-representation (red) or under-representation (blue) of a stop codon or nucleotide in “High” or “Low” readthrough groups relative to the frequency in the reference gene set. \( \chi^2 \) goodness of fit test with Bonferroni correction was performed to compare usage distribution between the “High” readthrough group,
“Low” readthrough group, and the reference. Significant difference was represented as a larger tile. Grey tiles signify that there was zero observation of that nucleotide (essentially under-represented), and log₂ calculation and statistical analysis could not be performed.

For nucleotides further into the mRNA channel 3’ of the stop codon (nt +5 to +9), only nt +9 showed clear trends: genes with A had higher readthrough efficiencies (i.e., readthrough-permissive), consistent with previous reports, and genes with U had lower readthrough efficiencies (i.e., readthrough-inhibiting). At position +5, although the differences between groups' readthrough efficiencies and overall readthrough efficiencies were not significant in every sample, genes containing adenine had a uniform pattern of higher readthrough efficiency compared to sample median (unlike random factor), indicating that A at nt +5 was relatively readthrough-permissive.

Adenine at nt -1 and -2 has been shown to be readthrough-permissive (Mottagui-Tabar et al. 1998; Tork et al. 2004), but this was not reflected in our current analysis (Figure 2.8A). Instead, we observed that U at nt -1 was readthrough-inhibiting, and A and C but not G and U at nt -3 were readthrough-permissive. It is noteworthy that nt -3 was identified as having a significant role along with nt -2; this may mean that the amino acid they encode or its tRNA, rather than the nucleotides themselves, mechanistically affect the termination process.

Thus, we also explored these three nucleotides as a codon (see below).

It is notable that most of the patterns for nt -3 and -2 in rli1-d sample did not match with other samples. In fact, some of them appeared to be the opposite. Since other samples, even the sup45 mutants, had wild-type Rli1, the opposite
patterns in rli1-d sample may suggest that nt -3 and -2 (or the amino acid they encode) played a role in a ribosome recycling step rather than the termination step. This also influenced the number of ribosomes translated into the 3’-UTR in rli1-d, as shown in the feature importance plot (Figure 2.7D).

To see whether we could extract more comprehensive information on these nucleotides, we used a different approach involving only the “High” and “Low” readthrough genes used for the classification models. \(\chi^2\) goodness of fit tests with Bonferroni correction were used to compare the frequencies of UGA, UAG, UGA, A, C, G, and U at each nucleotide position between the “High” readthrough group, “Low” readthrough group, and the reference gene set (2,693 genes). For data presentation purposes, the frequencies were converted into fractions and \(\log_2\) ratios of “High” or “Low” groups to the reference were computed, as shown in the heatmap (Figure 2.8B). A positive \(\log_2\) ratio (red) indicates an over-representation while a negative \(\log_2\) ratio (blue) indicates an under-representation of X in the group compared to reference. The patterns, especially those of the “Low” readthrough group, corroborated our previous approach (Figure 2.8A). For instance, genes with C at the nt +4 position had higher readthrough efficiencies than the sample median (Figure 2.8A), and within the “Low” readthrough group (Figure 2.8B), C was severely under-represented (blue tile) or even absent (grey tile). Moreover, we were able to rank the nucleotides from readthrough-permissive to readthrough-inhibiting using this approach. Although the patterns of either groups were not significantly different from reference based on the \(\chi^2\) analysis,
uniform patterns could be observed across all samples for most nucleotide positions (compared to random factor). For nt +5, the pattern allowed us to extrapolate that following readthrough-permissive G, were increasingly readthrough-inhibiting A, U, and C (which was most associated with termination). This information was difficult to obtain in the dual reporter assays because readthrough events were so low that they were indistinguishable from each other.

Information compiled for other nucleotide positions is shown in Table 2.4.

**Table 2.4. Comparison of nt -3 to +9 effects (based on Figures 2.8 and 2.9) to previous data (Rodnina et al. 2020).**

The order of nucleotide/stop codon is from most to (“>”) least permissive to readthrough.

<table>
<thead>
<tr>
<th>Position</th>
<th>Literature (Rodnina et al. 2020)</th>
<th>This study</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3</td>
<td>-</td>
<td>A/C &gt; G &gt; U</td>
</tr>
<tr>
<td>-2</td>
<td>A</td>
<td>A &gt; G &gt; U &gt; C</td>
</tr>
<tr>
<td>-1</td>
<td>A</td>
<td>A &gt; G &gt; C &gt; U</td>
</tr>
<tr>
<td><strong>Stop codon (+1, +2, +3)</strong></td>
<td>UGA &gt; UAG &gt; UAA</td>
<td>UGA &gt; UAG &gt; UAA</td>
</tr>
<tr>
<td>+4</td>
<td>C &gt; U &gt; G &gt; A</td>
<td>C &gt; U &gt; A &gt; G</td>
</tr>
<tr>
<td>+5</td>
<td>A</td>
<td>A &gt; G &gt; U &gt; C</td>
</tr>
<tr>
<td>+6</td>
<td>A/G</td>
<td>A/C &gt; G/U</td>
</tr>
<tr>
<td>+7</td>
<td>N</td>
<td>U</td>
</tr>
<tr>
<td>+8</td>
<td>U/C/G</td>
<td>G &gt; A/U &gt; C</td>
</tr>
<tr>
<td>+9</td>
<td>A</td>
<td>A &gt; C/G &gt; U</td>
</tr>
</tbody>
</table>

Next, we asked how particular combinations of stop codon and nt +4 identities, the two strongest variables among all the nucleotides, affect readthrough efficiency together. Genes were divided into groups based on their stop codon and nt +4 identities and the same analysis as described in Figure 2.8A was performed. We found that genes containing UGA, the most readthrough-permissive stop codon, followed by C, the most readthrough-permissive nt +4, had the highest readthrough efficiency relative to the overall median. UGA followed by G, the least
readthrough-permissive nt +4, reduced readthrough efficiency but not to the level of overall median (Figure 2.9A). On the other hand, genes containing the most readthrough-inhibiting combination, UAAG, had the lowest readthrough efficiencies while UAAC brought readthrough up to a level that was equal to that of the overall median. With regard to UGAC association with the highest readthrough efficiency, we noticed that this trend was not true for the sup45-ts strain at 37°C where, unlike in other strains, genes containing UGAC did not have higher readthrough efficiency than the median (Figure 2.9A). This result suggests that in the absence of functional eRF1, some combinations of stop codon and nt +4 identities had limited effects on readthrough efficiency variation.

To test whether particular stop codon and nt +4 identities tend to occur together at different readthrough rates, a $\chi^2$ test of independence was performed using all genes (“All”), genes in the “High” readthrough group, or genes in the “Low” readthrough group (Figure 2.9B). We found that in general (“All”), the association between stop codon and nt +4 identities was significant in all samples. Readthrough-inhibiting features, UAA and G, appeared together more frequently than the expected frequency (positive residuals) while readthrough-promoting features, UGA and C, tended to repel each other (negative residuals) (Figure 2.9B, left panel). Similar but weaker associations were seen among genes with “Low” readthrough, but no association was observed among those with “High” readthrough. From an evolutionary standpoint, this analysis supported a previous proposal stating that genes have evolved to minimize stop codon readthrough (Li
and Zhang 2019), as low readthrough features tended to coexist but high readthrough features tended to exclude each other.

**Figure 2.9.** Readthrough-permissive combinations of stop codon and nt +4 identities have additive effects on readthrough efficiency and are avoided in the genome.

**A.** Heatmap of median readthrough efficiency of genes containing particular combination of stop codon and nt +4 relative to median readthrough efficiency of all genes in the sample. Statistical analysis as in A. **B.** Heatmap showing residuals of $\chi^2$ test of independence between stop codon and nt +4 identities. Significant values ($p < 0.05$), represented by larger tile size, indicate the significant association between stop codon and nt +4 identities. Positive (red) residuals specify positive association (attraction) and negative (blue) residuals negative association (repulsion) between variables.
To summarize, we showed that readthrough-promoting stop codon and nucleotides at positions +4 to +9 previously determined in reporter assays also occur at a genome-wide level, with some combinations of stop codon and nt +4 identities selected for or against each other. Additionally, we identified readthrough-inhibiting nucleotides refractory to previous studies due to the detection limits of reporter assays.

**Specific combinations of nucleotides and tRNA in the ribosomal P site may influence readthrough efficiency**

Random forest identified the P-site amino acid to be a key factor in readthrough efficiency prediction, and we observed patterns of nucleotide usages at position -3, -2, and -1; however, it was unclear whether readthrough was influenced by the nucleotide themselves, the amino acid they encode, the tRNA that decoded them, or a combination of these three possibilities. Therefore, we performed analysis as described in Figure 2.8A, but as codons instead of individual nucleotide positions (Figure 2.10A) and were able to identify codons associated with genes having significant increases or decreases in readthrough efficiencies compared to the sample median.

From this analysis, two particular nucleotide combinations were likely responsible for low readthrough efficiencies, namely the C/G and U combinations and the UU and A/C combinations. In the first group, CGU and GCU were codons that appeared in genes with significantly lower readthrough efficiencies in most samples even though they code for different amino acids and were clearly distinct
from their respective synonymous codons. Remarkably, genes containing CUG (leucine), which had exactly the same nucleotide composition, showed higher readthrough efficiencies compared to sample median. These opposite trends suggested that not only specific combinations of nucleotides, but also specific positional arrangements have to be optimal for efficient termination. In the second group genes containing UUA and UUC had lower readthrough efficiencies but only in WT conditions (WT strains and sup45-ts at 25°C). Similar to the C/GU case, UUA and UUC had the same pattern despite coding for different amino acids and were distinct from their respective synonymous codons.

In both cases, the amino acid identity was ruled out as a possible mechanism. We wondered whether these two groups of codons share certain properties that may explain their occurrence in low readthrough genes and explored P-site codon usage frequency in the reference set, tRNA adaptation index (tAI) (Pechmann and Frydman 2013; Reis et al. 2004), tRNA anti-codon modifications (Johansson et al. 2008), and wobble pairs (Figure 2.10B). None of these properties were shared exclusively among the four codons, suggesting the mechanisms by which UUA/UUC and CGU/GCU influence readthrough efficiency were different. Nevertheless, the role of tRNA properties could not be ruled out. It is likely that the nucleotides and tRNAs together interact with the translation termination components in such a way that is optimal for a termination event to occur.
At the other end of the spectrum, codons that appeared in genes with significantly higher readthrough efficiencies include AUA, ACA, ACC, CUG, and GAC. Arguments similar to those made for the low readthrough set of codons applied here. First, the amino acid identity could be ruled out. Although ACA and ACC encode the same amino acid (threonine), it could be ruled out as a possible mechanism because other threonine codons showed different patterns. All five codons did not share any tRNA properties in common; however, individualized nucleotide-tRNA property combinations could not be ruled out as possible mechanisms.

Next, we performed a $\chi^2$ test of independence to determine whether the P-site codons with prominent effects on readthrough efficiencies tended to appear in conjunction with particular stop codons. Although statistical analysis here suffered from small sample size and the results should be interpreted with caution, we could still observe a similar trend as seen with stop codon and nt +4 in that readthrough-inhibiting features attracted one another. We found the most readthrough-inhibiting P-site codons, CGU and GCU, coexisted more frequently with UAA than UAG or UGA stop codons (Figure 2.10C). On the other hand, ACA, which was found in genes having high readthrough efficiencies, coexisted more frequently with UAG than UAA.

As with nt -3 and -2, many P-site codons (e.g., GAC, CUA, CUC, UUA) in the rli1-d mutant had either the opposite effect on readthrough efficiencies or opposite level of significance compared to the eRF1 mutants and WT cells. Thus,
it is possible that these codons affect the recycling step rather than termination step.
**Figure 2.10.** Specific codons in the P site were associated with higher or lower readthrough efficiencies.

**A.** Heatmap of median readthrough efficiency of genes containing a particular triplet codon in the ribosomal P site. Positive value (red) indicates that the group of genes had higher readthrough efficiencies compared to sample median, while negative value (blue) indicates lower readthrough efficiencies. Wilcoxon’s rank sum test was used to determine whether the difference in group and sample media was significant. Significant difference was represented as a larger tile. **B.** General information regarding the P-site codons. The Y-axis shows frequency of codon usage in the P site among the reference gene set. Color of the bar represents tRNA adaptation index (tAI), a measurement for codon optimality where 0 is non-optimal and 1 is optimal (Pechmann and Frydman 2013; Reis et al. 2004). Below the bar plots are wobble pair information and anti-codon modifications. **C.** Heatmap showing residuals of $\chi^2$ test of independence between stop codon and P-site codon identities. Significant values (p < 0.05), represented by larger tile size, indicate the significant association between stop codon and P-site codon identities. Positive (red) residuals specify positive association (attraction) and negative (blue) residuals negative association (repulsion) between variables. Interpretation of the results should be done with caution as the analysis suffered from small sample size.

In summary, we identified four codons (CGU, GCU, UUA, and UUC) associated with lower readthrough efficiencies and five codons (AUA, ACA, ACC, CUG, and GAC) associated with higher readthrough efficiencies. It is unlikely that amino acid identity by itself is the reason for this observation; rather, specific combinations of nucleotide and tRNA properties likely influence termination and readthrough.

**Readthrough efficiency increased with 3’-UTR lengths in the eRF1 mutants, but not in wild-type or recycling factor mutant cells**

Poly(A)-binding protein (PABP, Pab1 in yeast) has been shown previously to interact with eRF3, enhancing termination efficiency *in vitro* (Ivanov et al. 2016). Consistent with these observations, deletion of the *PAB1* gene in yeast, or extension of the mRNA 3’-UTR length downstream of a PTC, leads to inefficient
termination and readthrough *in vivo* (Wu et al. 2020). Accordingly, we asked whether readthrough efficiency and 3’-UTR length are positively correlated. A weak but significant positive correlation between readthrough efficiency and 3’-UTR length was observed in \textit{sup45-ts} cells at 37°C and \textit{sup45-d} cells while a negative correlation was observed in the respective WT strains (Figure 2.11). These results hinted that Pab1’s expected role in aiding termination became prominent only when eRF1 activity was inefficient. Although this result may be prone to artifacts, as 3’-UTR length also correlates with other features of the mRNAs (Figure 2.6), our finding was consistent with multiple studies linking the distance from the stop codon to Pab1 and termination efficiency, such as recent studies of context dependent translation termination in Ciliates (Swart et al. 2016).

\begin{figure}
\centering
\includegraphics[width=\textwidth]{scatter_plots}
\caption{Readthrough efficiency increased with 3’-UTR lengths in eRF1 mutants, but not in wild-type or ribosome recycling factor mutant cells. Scatter plots of readthrough efficiency vs. 3’-UTR length. Spearman’s correlation coefficient ($\rho$) and p-value were calculated for each sample.}
\end{figure}
Discussion

In this study, we analyzed ribosome profiling data of WT and mutant yeast strains defective in translation termination or ribosome recycling to identify mRNA or peptide features that influence readthrough efficiency at a genome-wide level. First, we characterized the phenotypes of eRF1 mutant cells, which included ribosomes stalling at start and stop codons and increased in-frame footprints in 3’-UTRs compared to wild-type cells (Figures 2.2 and 2.4B). We took these phenotypes into account when calculating readthrough efficiency of each transcript (Figure 2.5A) and excluded any transcripts that lack 3’-UTR annotation or sufficient footprints in the 3’-UTR from further analyses. Although these considerations left us with a smaller number of analyzable data points, we were able to reduce noise in the data that confounded a previous analysis attempt (Kleppe and Bornberg-Bauer 2018). With our analysis strategies, we demonstrated that the readthrough-promoting stop codon and proximal nucleotides previously determined in reporter assays (Anzalone et al. 2019; Cridge et al. 2018; Namy et al. 2001; Schueren et al. 2014) are not reporter-specific, but also occur at a genome-wide level in yeast for endogenous mRNAs (Figures 2.8 and 2.9). Moreover, our analyses identify readthrough-inhibiting nucleotides that were refractory to previous studies using reporter gene assays because the low level readthrough was below the detection threshold.

Two novel mRNA features that we found to be major determinants of readthrough efficiency were the codon in the ribosomal P site (when a stop codon
is in the A site) and 3’-UTR length. For the P-site codon (nt -3, -2, and -1), our results did not support previous conclusions of AA being the most readthrough-permissive nucleotides at position -2 and -1 (Rodnina et al. 2020; Mottagui-Tabar et al. 1998; Tork et al. 2004) when we analyzed the data by nucleotide position (Figures 2.8 and 2.9 and Table 2.4). Moreover, when we analyzed nt -3, -2, and -1 together as a codon in the P site, we did not find codons NAA to be significantly associated with higher readthrough. Instead, we found that genes with AUA, ACA, ACC, CUG, and GAC as their penultimate codon had higher readthrough efficiencies compared to the sample median (Figure 2.10A), suggesting that there may be properties in the P site other than adenines that could influence readthrough. Previous studies ruled out properties of the last amino acid residue (encoded by the P-site codon) as features affecting readthrough, but reported conflicting results for tRNA in the P site (Mottagui-Tabar et al. 1998; Tork et al. 2004). We were also able to rule out amino acid properties, and although we could not pinpoint the importance of tRNA properties, we could conclude that no single tRNA property (such as specific wobble pair or modification) was responsible for high readthrough. Nevertheless, we could not rule out synergistic effects of both nucleotide and tRNA on readthrough efficiency. The same arguments applied to our discovery of readthrough-inhibiting candidates: CGU, GCU, UUA, and UUC. Further experiments are needed to deconstruct the roles of nucleotide and tRNA in mediating readthrough efficiency.
Although ribosome recycling was not the focus of our study, our use of profiling data from recycling factor-depleted cells (Young et al. 2015) allowed us to observe that some nucleotides and codons in the P site have specific effects in rli1-d cells but not in wild-type or eRF1 mutant cells (Figures 2.8-2.10). Since wild-type and eRF1 mutant cells were both wild-type for recycling factors, these differences may be attributed to the recycling step. As failure to recycle leads to ribosomes reinitiating randomly within <10 nt downstream of the stop codon (Young et al. 2015), there is ~33% chance that reinitiation stays in-frame with the ORF, and there was no way for us to distinguish these footprints from actual readthrough footprints. Thus, we speculated that the differences in “readthrough” efficiency between recycling factor-depleted cells and other samples could mean these nucleotides or codons affect the recycling step. If this is true, cautious use and interpretation of dual/bi-cistronic reporter assays in measuring readthrough efficiency may need to be exercised, as reinitiation that is in-frame with the original reading frame could still produce a functional readthrough reporter. This is evidenced by previous claims borne out of dual reporter experiments stating that functions of Rli1 and Hcr1 were to control termination and readthrough (Beznosková et al. 2013; Khoshnevis et al. 2010), while ribosome profiling data and subsequent experiments demonstrated their ribosome recycling properties (Young and Guydosh 2019; Young et al. 2015). Unless the readthrough product was defined by the combined size of both internal control and readthrough reporters, enzymatic or colorimetric assays that separately detect the internal
control reporter and readthrough reporter may not be ideal methods for studying readthrough contexts.

Another mRNA feature that we found to be a significant determinant of readthrough efficiency was 3′-UTR length. Intriguingly, readthrough efficiency decreased with 3′-UTR length in WT cells but increased with 3′-UTR length in eRF1 mutant cells. The trend in eRF1 mutant cells is consistent with the hypothesis that proximity of a stop codon to PABP enhances termination efficiency (Wu et al. 2020). The fact that only eRF1 mutant cells displayed such a trend implied that PABP’s expected role in enhancing termination is principally observable when eRF1 function is not efficient. This result is consistent with recent work from our laboratory demonstrating that deletion of the yeast *PAB1* gene led to a reduced accumulation of premature termination products, increased readthrough at multiple PTCs, and increased readthrough of an otherwise “weak” PTC in parallel with 3′-UTR lengthening of the respective transcripts (Wu et al. 2020). We surmise that the proximity of the stop codon to PABP, which is thought to enhance termination efficiency through interaction of PABP with eRF3 (Ivanov et al. 2016; Roque et al. 2015), is masked in the ribosome profiling data of WT cells where eRF1 is fully functional, but when this protein became less efficient or absent, PABP’s role in recruiting eRF3 or other factors could be distinguished. The relationship between PABP’s proximity to a stop codon and the efficiency of termination parallels that of PABP and mRNA stability in the “faux 3′-UTR” model for NMD, where increasing PABP’s distance from a stop codon led to aberrant
termination and destabilized the mRNAs (He and Jacobson 2015b; Amrani et al. 2004).

Although normal and premature termination processes have been shown to differ in efficiency, they share the fundamental requirement for release factors and a stop codon (Amrani et al. 2004; He and Jacobson 2015b; Wu et al. 2020). While our study largely involved normal termination codons, we provided additional features of the mRNA sequences that influence readthrough, and hence termination efficiencies, which could be useful in designing or understanding therapeutic approaches for the large number of diseases caused by nonsense mutations (Dabrowski et al. 2015).

**Materials and Methods**

**Yeast strains and culture growth conditions**

Yeast strains used in this study are in the W303 background. The temperature-sensitive sup45-2 strain (HFY1218) was derived from the wild-type strain (HFY114) (He et al. 2003) by the pop-in/pop-out gene replacement technique. Cells were grown in 1 L of YPD at 25°C with shaking. When the OD$_{600}$ of the culture reached 0.6-0.8, its temperature was shifted as follows: Cells were collected by centrifugation at 5,000 rpm for 5 min at room temperature (RT) using a JA10 rotor in a Beckman Coulter Avanti J-E centrifuge, resuspended in 400 ml of fresh YPD media, transferred to a new flask, and incubated in a 25°C water bath with shaking. After 20 min of incubation, 400 ml of pre-warmed (57°C) YPD were added to the flask, and cells were incubated in a 37°C water bath for 30 minutes.
with shaking. As a control for the temperature shift procedure, the same protocol was carried out except that YPD was pre-warmed to 25°C and the subsequent 30-minute incubation was at 25°C.

**Library preparation and sequencing**

Ribosome profiling and RNA-seq libraries were prepared as described previously in our library preparation protocol for immunoprecipitated ribosomes (Ganesan et al. 2019). Briefly, 1 L of cells of OD$_{600}$ between 0.6 and 0.8 were harvested by rapid vacuum filtration and flash-frozen in liquid nitrogen in the presence of Footprinting Buffer (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 5 mM MgCl$_2$) plus 1% TritonX-100, 0.5 mM DTT, 1 mM phenylmethylsulfonyl fluoride (PMSF) and 1X protease inhibitors. Cells were lysed in a Cryomill (Retsch) at 5 Hz for 2 min, then 10 Hz for 15 min. Lysates were clarified by ultracentrifugation in a Beckman Coulter Optima L-90K Ultracentrifuge at 18,000 rpm for 10 min at 4°C, using a 50Ti rotor. Centrifugation was repeated for the supernatant at 18,000 rpm for 15 min at 4°C.

For ribosome profiling (Ingolia et al. 2012), lysates were digested with RNaseI (Invitrogen, #AM2294), then layered onto a 1 M sucrose cushion in Footprinting Buffer plus 0.5 mM DTT and centrifuged in a Beckman Optima TLX Ultracentrifuge at 60,000 rpm for 1 hour at 4°C using a TLA100.3 rotor to isolate 80S ribosomes. Ribosome-protected fragments were extracted using a miRNeasy kit (QIAGen, #217004) and depleted of rRNA using a Yeast Ribo-Zero Gold rRNA removal kit (Illumina, #MRZY1324) according to the manufacturer’s protocol.
Multiplexed cDNA libraries were prepared with the NEXTFlex Small RNA-Seq Kit v3 (BIOO Scientific, #NOVA-5132-06) according to the manufacturer’s protocol and sequenced (single-end, 75 cycles) on a NextSeq500.

For RNA-Seq, total RNAs were extracted from lysates using a miRNeasy kit (QIAgen, #217004), depleted of genomic DNA contamination using Baseline-Zero DNase (Epicentre, #DB0715K), and depleted of rRNA using a Yeast Ribozero Gold rRNA removal kit (Illumina, #MRZY1324) according to the manufacturer’s protocol. Multiplexed cDNA libraries were prepared with a TruSeq Stranded mRNA sample prep kit (Illumina, #RS-122-2101) according to the manufacturer’s protocol and sequenced (single-end, 75 cycles) on a NextSeq500.

**Reads processing and alignment**

The following software packages were used to process RNA- and ribo-seq library sequences: fastqc v0.10.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/); cutadapt v1.9 (Martin 2011); samtools v0.1.19 (Li et al. 2009); bowtie v1.0.0 (Langmead et al. 2009), cufflinks v2.2.1 (Trapnell et al. 2010); bam2fastq v1.1.0 (https://gsl.hudsonalpha.org/information/software/bam2fastq); RSEM v1.3.0 (Li and Dewey 2011).

A customized yeast transcriptome was constructed from the S288C reference genome sequence and annotations downloaded from the Saccharomyces Genome Database (https://www.yeastgenome.org) on September 10, 2015 (S288C_reference_genome_R64-2-1_20150113).
Annotations for the following classes of transcripts were parsed from the genomic gff3-formatted annotation file and processed separately: protein coding genes (6551); intron-containing genes (272); genes with 5'-UTR introns (24); frameshifted genes (47); blocked and pseudogenes (18), non-coding RNAs (ncRNAs) (16). For each gene in all but the last class, 5’- and 3’-UTR were added to the individual gff files employing the longest UTR entry among annotations from multiple studies (Zhang and Dietrich 2005; Miura et al. 2006; Nagalakshmi et al. 2008; Xu et al. 2009; Yassour et al. 2009) downloaded from the YeastMine database on July 3, 2017. 5’- and 3’-UTR information was available for 2840 and 2849 genes, respectively. In cases where no UTR was annotated, we used the 75th percentile lengths of 97 and 173 nt, respectively. Unspliced versions (pre-mRNAs) of all intron-containing genes were defined as a continuous exon extending from the start of the 5’-UTR to end of the 3’-UTR. After editing the individual parsed gff files, they were concatenated, and a fasta file with sequences of all 6952 transcripts was generated using the cufflinks gffread program. Information included in the fasta header lines generated by gffread was used to construct a transcript-to-gene-map file for use with RSEM, which was used for gene- and isoform-level quantification of transcripts. The fasta file is available at https://github.com/Jacobson-Lab/yeast_transcriptome_v5. A modified transcriptome was used for riboWaltz (see Data Analyses section). The riboWaltz transcriptome contained only the spliced transcripts of protein coding genes (6551) and genes with 5’-UTR introns (24). In addition, the riboWaltz transcriptome
considered the stop codon as a part of the 3'-UTR region rather than the CDS region, thus the length of CDS and 3'-UTR regions of each gene were adjusted accordingly.

For ribosome profiling libraries, de-multiplexed sequences were first processed by cutadapt to filter out low quality reads and read lengths shorter than 23 nt while trimming adapter sequences (cutadapt -a TGGAATTCTCGGGTGCCAAGG -q 10 --trim-n -m 23). Reads aligning to non-protein coding RNAs (rRNA, tRNA, snoRNA, and other non-protein coding RNA sequences) were removed by aligning with bowtie (-m 100 -3 4 -5 4 -n 2 -l 15 --un) and retaining only unaligned reads. The random 4-mers introduced on both sides of each RPF during library preparation were trimmed, concatenated, and stored as an identifier “barcode” used to remove PCR duplicates by aligning reads to the transcriptome (bowtie -m 10 -n 2 -l 15 -S --best --strata) and retaining only a single barcode (and associated read) at each position where multiple reads aligned. After duplicates removal, the SAM-formatted alignment file was converted either to BAM format using samtools or to fastq format using bam2fastq for use in downstream analyses. Transcript abundance measurements were determined by RSEM with settings --seed-length 15 --bowtie-m 10. Reads processing and alignment statistics can be found in Table 2.5.

The TruSeq stranded RNA-Seq libraries were processed using RSEM with no additional pre-processing.
Public ribosome profiling data sets

Public ribosome profiling data of *sup45*-d (Wu et al. 2019), *rli1*-d (Young et al. 2015), and their WT counterparts were downloaded from NCBI’s Gene Expression Omnibus (GEO) database. Yeast strain genotypes, cell growth conditions, accession numbers of the data series and specific sequencing runs are outlined in S1 Table. After download, reads were adapter-trimmed and length-filtered by cutadapt (cutadapt -a CTGTAGGCACCATCAAT -q 10 --trim-n -m 15), after which reads aligning to non-protein coding RNAs were removed (bowtie -m 100 -n 2 -l 15 --un) and remaining reads aligned to the transcriptome exactly as described above, with the omission of the PCR duplicate removal step. Reads processing and alignment statistics can be found in Table 2.5.

Table 2.5. Ribosome profiling reads processing and alignment statistics. *ncRNAs refer to non protein-coding RNAs.

<table>
<thead>
<tr>
<th>Sample</th>
<th>SRR Run</th>
<th>Total reads processed</th>
<th>Reads ≥ 15 bp &amp; contained adapters</th>
<th>Reads not aligned to ncRNAs*</th>
<th>Duplicate reads</th>
<th>Reads aligned to transcriptome</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUP45, 25°C (replicate 1)</td>
<td>SRR13208091</td>
<td>67,532,129</td>
<td>57,853,137</td>
<td>25,787,398</td>
<td>3,441,725</td>
<td>20,407,888</td>
</tr>
<tr>
<td>SUP45, 25°C (replicate 2)</td>
<td>SRR13208092</td>
<td>61,556,833</td>
<td>51,451,140</td>
<td>18,867,321</td>
<td>7,016,422</td>
<td>10,101,506</td>
</tr>
<tr>
<td>SUP45, 37°C (replicate 1)</td>
<td>SRR13208093</td>
<td>49,805,337</td>
<td>44,462,117</td>
<td>14,274,514</td>
<td>1,866,021</td>
<td>11,195,529</td>
</tr>
<tr>
<td>SUP45, 37°C (replicate 2)</td>
<td>SRR13208094</td>
<td>64,613,280</td>
<td>55,971,740</td>
<td>24,366,262</td>
<td>5,660,607</td>
<td>16,685,974</td>
</tr>
<tr>
<td>sup45-ts, 25°C (replicate 1)</td>
<td>SRR13208095</td>
<td>91,931,471</td>
<td>70,972,671</td>
<td>34,574,953</td>
<td>5,747,485</td>
<td>24,437,810</td>
</tr>
<tr>
<td>sup45-ts, 25°C (replicate 2)</td>
<td>SRR13208096</td>
<td>60,891,476</td>
<td>49,432,024</td>
<td>25,181,918</td>
<td>4,148,599</td>
<td>18,939,808</td>
</tr>
<tr>
<td>sup45-ts, 37°C (replicate 1)</td>
<td>SRR13208097</td>
<td>77,154,540</td>
<td>55,550,470</td>
<td>21,139,968</td>
<td>3,701,109</td>
<td>14,378,321</td>
</tr>
<tr>
<td>sup45-ts, 37°C (replicate 2)</td>
<td>SRR13208098</td>
<td>66,416,665</td>
<td>48,944,777</td>
<td>17,330,578</td>
<td>3,701,109</td>
<td>10,930,634</td>
</tr>
<tr>
<td>SUP45-D (replicate 1)</td>
<td>SRR7241903</td>
<td>54,157,010</td>
<td>53,979,859</td>
<td>39,261,094</td>
<td>N/A</td>
<td>39,261,094</td>
</tr>
<tr>
<td>SUP45-D (replicate 2)</td>
<td>SRR7241904</td>
<td>32,548,315</td>
<td>32,188,657</td>
<td>21,139,968</td>
<td>N/A</td>
<td>21,139,968</td>
</tr>
<tr>
<td>sup45-d (replicate 1)</td>
<td>SRR7241908</td>
<td>46,687,648</td>
<td>45,226,887</td>
<td>34,527,078</td>
<td>N/A</td>
<td>34,527,078</td>
</tr>
<tr>
<td>sup45-d (replicate 2)</td>
<td>SRR7241909</td>
<td>32,916,353</td>
<td>32,750,862</td>
<td>20,325,474</td>
<td>N/A</td>
<td>20,325,474</td>
</tr>
<tr>
<td>RLI1-D (replicate 1)</td>
<td>SRR2046309</td>
<td>52,021,032</td>
<td>51,768,936</td>
<td>17,545,223</td>
<td>N/A</td>
<td>17,545,223</td>
</tr>
<tr>
<td>RLI1-D (replicate 2)</td>
<td>SRR2046310</td>
<td>52,026,256</td>
<td>49,283,307</td>
<td>41,880,568</td>
<td>N/A</td>
<td>41,880,568</td>
</tr>
<tr>
<td>rli1-d (replicate 1)</td>
<td>SRR2046311</td>
<td>46,386,847</td>
<td>45,694,917</td>
<td>15,064,692</td>
<td>N/A</td>
<td>15,064,692</td>
</tr>
<tr>
<td>rli1-d (replicate 2)</td>
<td>SRR2046312</td>
<td>158,583,030</td>
<td>155,589,606</td>
<td>53,259,434</td>
<td>N/A</td>
<td>53,259,434</td>
</tr>
<tr>
<td>rli1-d (replicate 3)</td>
<td>SRR2046319</td>
<td>35,129,965</td>
<td>33,633,653</td>
<td>28,658,467</td>
<td>N/A</td>
<td>28,658,467</td>
</tr>
</tbody>
</table>
**Data analyses**

Analyses of ribosome profiling data were performed in the R software environment (versions 3.5 and 3.6). RPFs of 20-23 nt and 27-32 nt in length were used for the analyses. P-site offsets of reads in ribosome profiling libraries were assigned using the R package riboWaltz (Lauria et al. 2018), the results of which are provided in Table 2.6. After assigning P-site location, reads from replicate libraries were combined. In addition to riboWaltz, the following packages were used for data analysis and visualizations: data.table, dplyr, reshape2, seqinr, ggplot2, and scales. Built-in base R functions, rcompanion, and ggpubr were used for statistical analyses. randomForest and caret were used to generate and analyze random forest models.

**Table 2.6. Offset from 5’ and 3' ends of footprint to the first nucleotide in the P-site for each footprint length in each sample.**

<table>
<thead>
<tr>
<th>Sample</th>
<th>P-site Offset from 5'/3' ends for each footprint length (nt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUP45, 25°C (replicate 1)</td>
<td>12/7 13/7 13/8 14/8 11/15 12/15 13/15 13/16 13/17 13/18</td>
</tr>
<tr>
<td>SUP45, 25°C (replicate 2)</td>
<td>12/7 13/7 13/8 14/8 11/15 12/15 13/15 13/16 13/17 13/18</td>
</tr>
<tr>
<td>SUP45, 37°C (replicate 1)</td>
<td>12/7 13/7 13/8 13/9 11/15 12/15 13/15 13/16 13/17 13/18</td>
</tr>
<tr>
<td>SUP45, 37°C (replicate 2)</td>
<td>12/7 13/7 13/8 14/8 11/15 12/15 13/15 13/16 13/17 13/18</td>
</tr>
<tr>
<td>sup45-ts, 25°C (replicate 1)</td>
<td>12/7 12/8 13/8 14/8 11/15 12/15 13/15 13/16 13/17 13/18</td>
</tr>
<tr>
<td>sup45-ts, 25°C (replicate 2)</td>
<td>12/7 13/7 13/8 14/8 11/15 12/15 13/15 13/16 13/17 13/18</td>
</tr>
<tr>
<td>sup45-ts, 37°C (replicate 1)</td>
<td>12/7 12/8 13/8 14/8 11/15 12/15 13/15 13/16 13/17 13/18</td>
</tr>
<tr>
<td>sup45-ts, 37°C (replicate 2)</td>
<td>12/7 13/7 13/8 14/8 11/15 12/15 13/15 13/16 13/17 13/18</td>
</tr>
<tr>
<td>SUP45-D (replicate 1)</td>
<td>13/6 14/6 15/6 16/6 11/15 12/15 13/15 14/15 14/16 14/17</td>
</tr>
<tr>
<td>SUP45-D (replicate 2)</td>
<td>13/6 14/6 15/6 16/6 11/15 12/15 13/15 14/15 15/15 16/15</td>
</tr>
<tr>
<td>sup45-d (replicate 1)</td>
<td>13/6 14/6 15/6 16/6 11/15 12/15 13/15 14/15 14/16 14/17</td>
</tr>
<tr>
<td>sup45-d (replicate 2)</td>
<td>13/6 14/6 15/6 16/6 11/15 12/15 13/15 14/15 15/15 16/15</td>
</tr>
<tr>
<td>rl1-d (replicate 1)</td>
<td>4/15 5/15 6/15 7/15 11/15 12/15 13/15 14/15 14/16 14/17</td>
</tr>
<tr>
<td>rl1-d (replicate 2)</td>
<td>4/15 5/15 6/15 7/15 11/15 12/15 13/15 14/15 14/16 14/17</td>
</tr>
</tbody>
</table>
For analyses results that were visualized in a heatmap form, extreme results were made to fit the indicated scale in order to not jeopardize the visualization of most of the results. The actual results for each figure and associated R codes used to generate the figure are provided for reference at https://github.com/Jacobson-Lab/sup45-ts_readthrough.

**Random forest models**

Random forest analyses (Breiman 2001) were implemented in the R environment using R packages randomForest (Liaw and Wiener 2002) and caret (Kuhn 2008). A classification (predicts readthrough groups) and a regression (predicts readthrough efficiency) model were created for each sample. Each model was trained with 100 trees (parameter ntree = 100) and 81 mRNA features. The number of features chosen to split at each tree node varied (parameter mtry ranging from 1 to 81, increasing by 10), with the number resulting in the highest accuracy automatically selected by the program. Model accuracy was evaluated by 5-fold cross-validation using receiver operating characteristic (ROC) as the performance metric for classification model and root mean squared error (RMSE) for regression model. To report performance metric for classification, area under the ROC curve (AUROC) for each fold was extracted, and the average ± standard deviation across 5 folds was calculated. For regression, RMSE was normalized by the range of the Y variable (readthrough efficiency) for each fold, and the average ± standard deviation across 5 folds was calculated.
Interpretation of each performance metric is as follows. AUROC is a common metric used to assess classification model. An ROC curve is generated by plotting true positive rates (TPR) against false positive rates (FPR) of multiple classification thresholds. A straight diagonal line indicates that TPR equals to FPR, resulting in area under the curve (AUROC) of 0.5, and means the chance that the model can classify genes into the correct groups is 50% – this is no better than random chance. Thus, the higher AUROC is than 0.5, the better the model performs, and AUROC of 1 indicates a perfect classification model (James et al. 2013). Acceptable value for AUROC depends on the context of the study. For our study, we think the AUROC of ~0.7-0.8 is acceptable.

Root mean squared error (RMSE), one of the performance metrics for regression, is the average of the distance between the predicted values and the actual values. The unit of RMSE is the same as the Y variable. Whether the error is considered large or small depends on the range of the data. For instance, an average error of ± 0.5 may be considered small for the data ranging from 0-100, but considered large for the data ranging from 0-1. Therefore, in order to easily assess the errors across samples, we normalized the RMSE by the range of the data, where the resulting normalized RMSE (NRMSE) displays the average fraction of error relative to data range.

After the random forest models were trained, feature importance for each X variable was extracted with the importance() function. We selected Mean Decrease Accuracy (MDA) and % Increase in Mean Squared Error (%IncMSE) as
measures of feature importance for classification and regression models, respectively.

**Data Availability**

The raw sequencing data are deposited and available at Gene Expression Omnibus (GEO) under accession number GSE162780. All data files underlying figures and R codes to plot them are here: https://github.com/Jacobson-Lab/sup45-ts_readthrough.

**Acknowledgments**

This work was supported by grants to A.J. (5R01 GM27757-37 and 1R35GM122468-04) from the U.S. National Institutes of Health and from the Cystic Fibrosis Foundation (CFFKOROST20GO). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank Jill Moore, Zhiping Weng, Andrei Korostelev, Barry Cooperman, and Sean Ryder for helpful discussions.

**Author Contributions**

K.M., F.H., R.G., and A.J. conceived and designed the experiments, F.H. constructed the sup45-2 yeast strain, K.M. carried out the experiments and data analyses, R.B. wrote data processing scripts, K.M. and A.J. wrote the paper with input from all authors, and A.J. obtained funding for the study.

**Declaration of Interests**

A.J. is co-founder, director, and research consultant for PTC Therapeutics Inc.
CHAPTER III:

Transcriptome-wide cis-regulatory elements of normal termination codon readthrough in human cells

The work in this chapter is currently under review for publication.

This is a collaborative work with Dr. David Bedwell's lab in the Department of Biochemistry and Molecular Genetics, Heersink School of Medicine, at the University of Alabama at Birmingham. I performed bioinformatics analyses. Lianwu Fu, Ming Du, Kari Thrasher, Kim M. Keeling, and David Bedwell conceived, designed, and carried out the dual-luciferase hCFTR readthrough experiments, the results of which were critical proof of the predictive ability of machine learning model to predict readthrough efficiency.
Abstract

Protein synthesis terminates when the ribosome encounters a stop codon in its A-site. Although translation termination efficiency is high, stop codon readthrough can occur when a near-cognate tRNA outcompetes release factors during decoding, allowing continued translational elongation. Understanding readthrough efficiency regulation is important for development of therapeutics targeted to diseases caused by nonsense mutations. Cis-acting regulators of readthrough efficiency have been identified in several organisms and comparisons of parallel findings can inform the mechanism of readthrough regulation. Using a machine learning approach, we analyzed readthrough efficiency data from published HEK293T ribosome profiling experiments and compared it to that obtained from similar yeast studies. We obtained evidence for the conservation of identities of stop codon, its context, and 3'-UTR length (when termination is compromised), but not P-site codon, suggesting a tRNA role in readthrough regulation. Further, we demonstrated that models trained on data from cells treated with the readthrough-promoting drug, G418, accurately predict readthrough of premature termination codons arising from CFTR nonsense alleles that cause cystic fibrosis. This predictive ability has the potential to aid development of nonsense suppression therapies by predicting readthrough efficiency or a patient’s likelihood of improvement in response to drugs given the sequence context of their nonsense mutations.
Introduction

Termination of protein synthesis occurs when the ribosome encounters one of the three stop codons (UAA, UAG, and UGA) at the end of an mRNA open reading frame (ORF). The release factor complex, comprised of eRF1, eRF3, and GTP, recognizes a stop codon in the ribosomal A-site and facilitates nascent peptide release, then the ribosome is recycled by Rli1/ABCE1 for another round of translation (Hellen 2018; Schuller and Green 2018). The termination process is highly efficient and has a low error rate (Dabrowski et al. 2015), but an error can still occur when the stop codon is decoded by a near-cognate tRNA instead of eRF1, resulting in continued translation elongation into the mRNA 3’-UTR, a process termed “stop codon readthrough” (Dabrowski et al. 2015; Rodnina et al. 2020). An understanding of the details of this process is likely to be beneficial for the development of therapeutics that target diseases caused by nonsense mutations, where readthrough is desirable at premature termination codons (PTCs) but not normal termination codons (NTCs) (Welch et al. 2007).

Cis-acting elements influencing the efficiency of termination and readthrough have been identified in several organisms. These modulators of readthrough efficiency include the identity of the stop codon and flanking nucleotides (Anzalone et al. 2019; Bonetti et al. 1995; Cridge et al. 2018; Namy et al. 2001; Schueren et al. 2014), stem loop structures in the mRNA 3’-UTR (Anzalone et al. 2019; Firth et al. 2011), and specific RNA binding protein motifs (Eswarappa et al. 2014; Manjunath et al. 2022). Among these, the stop codon and
the nucleotide following it (nt +4) are the most studied across species and yield the most consistent results. Structural insights into their mechanism have been demonstrated with cryo-electron microscopy (Brown et al. 2015; Shao et al. 2016) and their significance has been extended to a transcriptome-wide level in human and yeast cells by ribosome profiling experiments (Mangkalaphiban et al. 2021; Wangen and Green 2020). The conservation of other elements, however, is more difficult to determine due to variation in experimental conditions, analysis strategies, and (for endogenous mRNAs) existing nucleotide usage biases in different species. The diversity in mRNA sequences in transcriptomics data provides opportunities to determine whether cis-acting elements are conserved for endogenous mRNAs between yeast and human cells as well as to explore how they can be utilized to predict readthrough efficiency given new sequences.

Therefore, to gain parallel insights into cis-acting elements affecting readthrough efficiency at a human transcriptome-wide level, we re-analyzed published readthrough efficiency data for HEK293T cells (Wangen and Green 2020) in the same manner as we did for yeast cells (Mangkalaphiban et al. 2021). Our analysis revealed that, in addition to the previously established importance of stop codon context, readthrough efficiency increased with 3′-UTR length in HEK293T cells under readthrough-promoting conditions, consistent with our yeast results. We also found that the patterns of P-site codon triplets associated with high or low readthrough efficiency in HEK293T cells are largely different from those observed in yeast. Further, we demonstrated that a machine learning model
trained on these cis-acting elements can predict with high accuracy the readthrough efficiency of PTCs arising from nonsense mutations found in cystic fibrosis patients. Collectively, we obtained evidence for conservation of cis-acting elements modulating readthrough efficiency among human and yeast cells at a transcriptome-wide level, derived insights into the mechanism of translation termination that may involve tRNA properties, and presented potential applications of ribosome profiling data coupled with machine learning approaches in readthrough prediction and nonsense suppression therapies.

Results

Random forest models identify mRNA features predictive of readthrough efficiency

To study cis-acting elements affecting transcriptome-wide readthrough efficiency in human cells, we applied the analysis approaches we previously developed with yeast ribosome profiling data (Mangkalaphiban et al. 2021) to readthrough efficiency data generated from ribosome profiling experiments of HEK293T cells treated with various readthrough-promoting aminoglycosides (Wangen and Green 2020). For each sample, mRNAs with detectable readthrough were identified and random forest models (Breiman 2001; Liaw and Wiener 2002) were trained to predict readthrough efficiency of these mRNAs using mRNA or nascent peptide features (Figure 3.1). The feature importance score of each feature extracted from the model indicates the predictive ability of that feature (Figure 3.1A-B). As expected, the negative control features (randomly assigned
numbers or letters) have low feature importance scores for both random forest regression (predicting readthrough efficiency values) (Figure 3.1A, negative control = “NC” columns) and classification (predicting extremely “high” and “low” – top and bottom 15% – readthrough) (Figure 3.1B, negative control = “NC” columns) approaches. Among the features with high scores that are predictive of readthrough efficiency are the identities of the stop codon and the nucleotide immediately after it (nt +4) (Figure 3.1A-B), each of which has previously been shown to affect termination and readthrough efficiency (Dabrowski et al. 2015; Rodnina et al. 2020; Tate et al. 2018). Other features with high scores include 3’-UTR length and the identity of P-site amino acid (Figure 3.1A-B), both of which were previously recognized in yeast (Wu et al. 2020; Mangkalaphiban et al. 2021). For the stop codon and nt +4, the feature importance scores for both are especially prominent in G418-treated samples, where readthrough levels were much higher than those seen in HEK293T cells treated with other aminoglycosides and 3’-UTR ribosome footprints showed 3-nt periodicity indicative of continued translation into the 3’-UTR from the CDS region (Wangen and Green 2020).
A

Nascent peptide in the exit tunnel

<table>
<thead>
<tr>
<th>aa 20-30 from PTC</th>
<th>aa 13-19 from PTC</th>
<th>aa 10-12 from PTC</th>
<th>aa 3-9 from PTC</th>
<th>nt from stop</th>
<th>nt from stop</th>
<th>NC</th>
</tr>
</thead>
<tbody>
<tr>
<td>untreated</td>
<td>Amikacin</td>
<td>G418 (2)</td>
<td>G418 (0.5)</td>
<td>G418 (0.5, 10 min)</td>
<td>Gentamicin</td>
<td>Neomycin</td>
</tr>
<tr>
<td>- charged</td>
<td>- charged</td>
<td>- charged</td>
<td>- charged</td>
<td>- charged</td>
<td>- charged</td>
<td>- charged</td>
</tr>
<tr>
<td>No charge</td>
<td>No charge</td>
<td>No charge</td>
<td>No charge</td>
<td>No charge</td>
<td>No charge</td>
<td>No charge</td>
</tr>
<tr>
<td>+ charged</td>
<td>+ charged</td>
<td>+ charged</td>
<td>+ charged</td>
<td>+ charged</td>
<td>+ charged</td>
<td>+ charged</td>
</tr>
<tr>
<td>Aromatic</td>
<td>Aromatic</td>
<td>Aromatic</td>
<td>Aromatic</td>
<td>Aromatic</td>
<td>Aromatic</td>
<td>Aromatic</td>
</tr>
<tr>
<td>Nonpolar</td>
<td>Nonpolar</td>
<td>Nonpolar</td>
<td>Nonpolar</td>
<td>Nonpolar</td>
<td>Nonpolar</td>
<td>Nonpolar</td>
</tr>
<tr>
<td>Hydrophlic</td>
<td>Hydrophlic</td>
<td>Hydrophlic</td>
<td>Hydrophlic</td>
<td>Hydrophlic</td>
<td>Hydrophlic</td>
<td>Hydrophlic</td>
</tr>
<tr>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
</tr>
<tr>
<td>Hydrophobic</td>
<td>Hydrophobic</td>
<td>Hydrophobic</td>
<td>Hydrophobic</td>
<td>Hydrophobic</td>
<td>Hydrophobic</td>
<td>Hydrophobic</td>
</tr>
</tbody>
</table>

B

Nascent peptide in the exit tunnel

<table>
<thead>
<tr>
<th>aa 20-30 from PTC</th>
<th>aa 13-19 from PTC</th>
<th>aa 10-12 from PTC</th>
<th>aa 3-9 from PTC</th>
<th>nt from stop</th>
<th>nt from stop</th>
<th>NC</th>
</tr>
</thead>
<tbody>
<tr>
<td>untreated</td>
<td>Amikacin</td>
<td>G418 (2)</td>
<td>G418 (0.5)</td>
<td>G418 (0.5, 10 min)</td>
<td>Gentamicin</td>
<td>Neomycin</td>
</tr>
<tr>
<td>- charged</td>
<td>- charged</td>
<td>- charged</td>
<td>- charged</td>
<td>- charged</td>
<td>- charged</td>
<td>- charged</td>
</tr>
<tr>
<td>No charge</td>
<td>No charge</td>
<td>No charge</td>
<td>No charge</td>
<td>No charge</td>
<td>No charge</td>
<td>No charge</td>
</tr>
<tr>
<td>+ charged</td>
<td>+ charged</td>
<td>+ charged</td>
<td>+ charged</td>
<td>+ charged</td>
<td>+ charged</td>
<td>+ charged</td>
</tr>
<tr>
<td>Aromatic</td>
<td>Aromatic</td>
<td>Aromatic</td>
<td>Aromatic</td>
<td>Aromatic</td>
<td>Aromatic</td>
<td>Aromatic</td>
</tr>
<tr>
<td>Nonpolar</td>
<td>Nonpolar</td>
<td>Nonpolar</td>
<td>Nonpolar</td>
<td>Nonpolar</td>
<td>Nonpolar</td>
<td>Nonpolar</td>
</tr>
<tr>
<td>Hydrophlic</td>
<td>Hydrophlic</td>
<td>Hydrophlic</td>
<td>Hydrophlic</td>
<td>Hydrophlic</td>
<td>Hydrophlic</td>
<td>Hydrophlic</td>
</tr>
<tr>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
<td>Neutral</td>
</tr>
<tr>
<td>Hydrophobic</td>
<td>Hydrophobic</td>
<td>Hydrophobic</td>
<td>Hydrophobic</td>
<td>Hydrophobic</td>
<td>Hydrophobic</td>
<td>Hydrophobic</td>
</tr>
</tbody>
</table>

C

NRMSE

D

AUROC
**Figure 3.1.** mRNA features predictive of readthrough efficiency.

**A-B.** Feature importance scores, % increase in mean square error (%IncMSE) for regression (A) and Mean Decrease Accuracy (MDA) for classification (B), were extracted from random forest models. The higher the feature importance score, the more important a feature is in predicting readthrough efficiency or distinguishing between “high” and “low” readthrough mRNAs. Numbers or categorical values randomly assigned to the mRNAs are used as negative controls (“NC”) and to set the baseline feature importance scores as “unimportant” features.

**C-D.** Random forest model performance metrics. Normalized Root Mean Square Error (NRMSE) shows prediction errors of random forest regression models (C). The area under the receiver operating characteristic (AUROC) reports the capability of the model to distinguish between “high” and “low” readthrough mRNAs, which is higher than random chance (0.5, red horizontal line) in all samples (D). For both metrics, the values were average ± standard deviation of 5-fold cross-validation.

**Known stop codon context influences readthrough efficiency**

To understand how mRNA features affect readthrough efficiency prediction, we grouped mRNAs based on the usage of stop codons, nt +4, or other nucleotides flanking the stop codons, and compared median readthrough efficiency of that group to the sample median (Figure 3.2A). When mRNAs were grouped by the identity of the negative control random letters, no significant differences between sample median and each group’s median were observed in any samples (Figure 3.2A, “Random”). However, when mRNAs were grouped by the stop codon identity, we found that mRNAs that use UGA as the stop codon had higher readthrough efficiency, while those using UAG and UAA had lower readthrough efficiency than the overall sample median in most samples (Figure 3.2A, “Stop”). This observation is consistent with results from Wangen and Green despite the difference in statistical analysis approaches, as well as with results from previous studies showing that UGA is the most readthrough-permissive stop codon, while UAA is
the least (Bonetti et al. 1995; Dabrowski et al. 2015; Loughran et al. 2014; Rodnina et al. 2020). Using our approach, we can extrapolate from the general trends across samples that the readthrough-promoting motif is UGACANNNNA while the readthrough-inhibitory motif is UAAG(G/C)NNNC in these HEK293T cells.

We further asked whether the strongest features, the stop codon and nt +4, together have additive effects on readthrough efficiency as shown in previous studies. As expected, when mRNAs were grouped by the combination of these two features, the most readthrough-permissive combination, UGAC, had the highest increase in readthrough efficiency relative to sample median, while the most readthrough-inhibiting combination, UAAG, has the highest decrease in readthrough efficiency across most samples (Figure 3.2B). The trends for each feature when another feature is held constant (Figure 3.2B) are also mostly in line with analyses performed for each feature independently (Figure 3.2A). When nt +4 is held constant, the order of most to least readthrough-promoting stop codon is UGA > UAG > UAA (Figure 3.2B, compare “A” columns from different stop codon panels to each other, and so on). When the stop codon is held constant, the order most to least readthrough-promoting nt +4 is C > U > A/G (Figure 3.2B, compare different nt +4 columns to each other within a specific stop codon panel).
Figure 3.2. Effects of the stop codon and flanking nucleotides identities on readthrough efficiency.

A-B. Differences between median readthrough efficiency of all mRNAs in the sample and median readthrough efficiency of a group of mRNAs containing particular stop codon or nucleotide (A), or stop codon with nt +4 as quadruplet (B). Positive (red) and negative (blue) values indicate that the group of mRNAs had higher and lower median readthrough efficiency compared to the sample median, respectively. Two-tailed Wilcoxon’s rank sum test with Benjamini-Hochberg method for multiple testing correction was used to determine whether the difference was significant. C. Standardized residuals of $\chi^2$ test of independence determining association between stop codon and nt +4 identities. Positive residuals (pink) indicate that the pair occurs together more often than expected (attraction) while negative residuals (green) less often than expected. For all panels, significant result ($p < 0.05$) is represented as a larger tile.
Readthrough-promoting stop codon and nt +4 occur together less often than expected

Unlike studies of readthrough using reporters, endogenous mRNAs have been subjected to evolutionary pressures to limit deleterious consequences of readthrough but also allow chances for protein adaptation (Jungreis et al. 2016; Li and Zhang 2019; Manjunath et al. 2022). Studying readthrough of endogenous mRNAs using ribosome profiling further biases the pool of data due to a detection threshold for 3'-UTR reads, which depends on sufficiently high expression and translation of the mRNA CDS region, the existence of readthrough-promoting features that allow readthrough of certain mRNAs to be “seen” more often, and aminoglycoside treatment (Figure 3.3A). Indeed, we observed that, on average, ~54% of mRNAs in each sample had UGA stop codons; these frequencies are only slightly higher than the 49% determined from the Reference group, which encompasses mRNAs from all samples combined regardless of whether readthrough is detected or not (Figure 3.3B). For nt +4, A and G are the most common with comparable frequencies, followed closely by C, while U is depleted (Figure 3.3B). The high abundance of readthrough-promoting UGA but depletion in readthrough-promoting nt +4 suggests that a combination of readthrough-promoting features may have been under a negative selection.

To statistically determine if certain pairs of stop codon and nt +4 occur more or less often than the expected frequencies given existing biases in the data, we performed a $\chi^2$ test of independence (Figure 3.2C), testing whether the observed
distribution of nt +4 among stop codons (Figure 3.3C-D) is different from the expected frequencies where nt +4 would be distributed evenly among stop codons (Figure 3.3E-F). We found that UGAC is over-represented in Untreated, Amikacin-treated, and Tobramycin-treated samples while UGAG is over-represented in the rest of the samples as well as in the Reference (Figure 3.2C, pink in UGA panel). Coincidentally, the three samples have the least number of mRNAs with detectable readthrough (Figure 3.3A). These results indicate that when overall readthrough levels are very low, detectable readthrough in Untreated, Amikacin-treated, and Tobramycin-treated samples are in part biased towards mRNAs with readthrough-promoting features. Among the aminoglycoside-treated samples, G418-treated samples show the closest results to the Reference (Figure 3.2C); readthrough induced by G418 led to a high-enough number of mRNAs with detectable readthrough (Figure 3.3A) such that the diversity of stop codon and nt +4 combinations is sufficiently representative of the pool of mRNAs expressed in HEK293T cells.

For the Reference and G418-treated samples, we determined that UGAG being noticeably the most common quadruplet in the data (Figure 3.3C) is not solely due to the naturally high abundance of UGA and G in the data, but their association is truly over-represented (Figure 3.2C). The same is true for UAAA being the most common among the UAA and nt +4 combinations (Figure 3.3C-D and Figure 3.2C). These observations suggest that a combination of readthrough-promoting features tend to be selected against.
Although UAAU also occurs more often than expected (Figure 3.2C), the low prevalence of U in general results in UAAU being relatively infrequent overall, not that much more frequent than UAAC or UAAG (Figure 3.3C). Similarly, although UAGC occurs more often than expected (Figure 3.2C), its frequency is no more than UAGA or UAGG (Figure 3.3C). Because $\chi^2$ analysis compares proportions, it is expected that when G is over-represented in one group, other nucleotides like U or C would be under-represented in that group and would appear over-represented in another group. This is likely the reason for the positive associations seen for UAGC and UAAU rather than them being evolutionarily selected for, because their actual frequencies in the data are quite low (Figure 3.3C).

Taken together, we found that UGAG and UAAA occur more often than expected. This diminished association between readthrough-promoting features is likely the consequence of evolutionary pressure to limit high readthrough among mRNAs expressed in HEK293T cells.

**Figure 3.3. Frequency of stop codon and nt +4 in each sample.**

A. Number of mRNAs with detectable readthrough in ribosome profiling data of HEK293T cells treated with different aminoglycosides. Reference sample consists of all mRNAs detected by ribosome profiling in all samples combined regardless of whether readthrough was detectable or not. B. Percentage of stop codon (left) and nt +4 (right) observed in each sample. C. Percentage of stop codon and nt +4 as quadruplets observed in each sample. D. Percentage of nt +4 for each stop codon observed in each sample. E. Expected percentage of stop codon and nt +4 as quadruplets in each sample if nts +4 were distributed evenly among the stop codons. F. Expected percentage of nt +4 for each stop codon in each sample if nts +4 were distributed evenly among the stop codons.

(Next page)
P-site codons have mild effects on readthrough efficiency

Some nucleotide positions upstream of the stop codon seem to be associated with different readthrough efficiency levels. For example, mRNAs with guanine at position -1 and those with uracil at position -3 had lower and higher readthrough efficiencies than the sample median, respectively (Figure 3.2A). These nucleotide positions encode the amino acid immediately prior the stop codon, so when the stop codon enters the ribosomal A-site, these nucleotides are in the P-site. Concurrently, the P-site amino acid has feature importance scores higher than baseline in the random forest models (Figure 3.1A-B).

To determine whether readthrough efficiency is modulated by the P-site codon nucleotides, encoded amino acids, or decoding tRNAs, we grouped mRNAs by the identity of their P-site codon triplets and compared each group’s median readthrough efficiency to the sample median (Figure 3.4). If readthrough efficiency is strongly influenced by the nucleotides, we expect the codons with the same nucleotide composition at certain positions to have the same results regardless of the amino acid they encode. On the other hand, if readthrough efficiency is influenced by the amino acid, we expect the codons that encode the same amino acid to have the same results despite their differences in nucleotide composition. Results that do not follow these expectations suggest that tRNA properties may be involved or indicate a more complex interplay of at least two of the three P-site features.
Figure 3.4. Effects of triplet codons in the ribosomal P-site on readthrough efficiency.

Differences between median readthrough efficiency of a group of mRNAs containing particular P-site codon and median readthrough efficiency of all mRNAs in the sample. Positive (red) and negative (blue) values indicate that the group of mRNAs had higher and lower median readthrough efficiency compared to the sample median, respectively. Two-tailed Wilcoxon’s rank sum test with Benjamini-Hochberg multiple testing correction was used to determine whether the difference was significant. Significant difference (p < 0.05) is represented as a larger tile.

Although no codons show significant results consistently across samples, the direction of association remains uniform across samples for many codons, as opposed to non-uniform patterns of negative control random letters (Figure 3.2A, “Random”). For example, mRNAs that have AAG in the P-site generally have lower readthrough efficiency than the sample median, while those that have UGG have higher readthrough efficiency than the sample median (Figure 3.4). Because UGG is the only codon encoding tryptophan (W), it is not discernable whether this specific codon nucleotide combination, the decoding tryptophan tRNA, or the encoded tryptophan is responsible for higher readthrough efficiency. It is notable, however, that the first nucleotide of UGG corresponds to U at position -3, which is associated with high readthrough efficiency (Figure 3.2A) and most other codons that begin with U also show the general trend of higher readthrough efficiency.
(regardless of significance level) (Figure 3.4). Therefore, U at position -3 may be mechanistically responsible for high readthrough efficiency in these mRNAs.

**Readthrough efficiency increases with 3′-UTR length**

Random forest models identified 3′-UTR length as one of the important predictors of readthrough efficiency. To understand the relationship between 3′-UTR length and readthrough efficiency, we calculated Spearman’s correlation coefficient between these two values. Readthrough efficiency is positively correlated with 3′-UTR length in all samples (Figure 3.5A). To rule out the possibility that the observed correlation is skewed by extremely short or long 3′-UTRs, we repeated the analysis with only those mRNAs whose 3′-UTRs are longer than 100 nt and shorter than 5,000 nt. The positive correlations are maintained in all samples (Figure 3.5B). The positive correlation between readthrough efficiency and 3′-UTR length is probably linked to the role of poly(A)-binding protein (PABP) in enhancing translation termination, which has been demonstrated both *in vivo* and *in vitro* in other eukaryotes (Ivanov et al. 2016; Wu et al. 2020). Further, in a reporter gene assay, the distance of PABP to the stop codon also correlated with readthrough efficiency in yeast (Wu et al. 2020).
**Figure 3.5.** Readthrough efficiency increases with 3’-UTR length.  

A. Readthrough efficiency vs. 3’-UTR length for all mRNAs (that have UTR annotations) in the sample. B. Readthrough efficiency vs. 3’-UTR length for mRNAs with 3’-UTR lengths longer than 100 nt but shorter than 5,000 nt. For all panels, Spearman’s correlation coefficient ($\rho$) and the associated $p$-value is reported for each sample.
Random forest model can accurately predict outcomes of PTC readthrough measured by Dual-Luciferase assay in G418-treated cells

Nonsense mutations introduce PTCs in the middle of mRNA coding regions, resulting in the production of nonfunctional, truncated proteins and severe diseases (Mort et al. 2008). Determining readthrough efficiency of PTCs is crucial in understanding disease phenotypes and designing therapeutic approaches to these diseases. We tested whether the random forest models trained on endogenous mRNA features and transcriptome-wide readthrough efficiency measured by ribosome profiling can accurately predict readthrough efficiency of 15 PTCs derived from CFTR nonsense alleles that are found in cystic fibrosis patients.

To experimentally measure PTC readthrough, we created a dual-luciferase (Dual-Luc) reporter construct for each PTC allele by fusing a DNA fragment containing the PTC and 9 nucleotides (3 codons) flanking the 5’ and 3’ sides of the PTC in-between the Renilla and Firefly luciferase genes, where the upstream gene (Renilla) lacks a stop codon and the downstream gene (Firefly) lacks a start codon (Figure 3.6). The construct was then transfected into HEK293 cells, treated or not treated with 0.1 mg/mL G418, and expression of the two reporter genes was quantified after 24 hours. Readthrough efficiency was determined as the percentage of Firefly luciferase normalized to the signal for Renilla luciferase.
To predict readthrough, we recorded mRNA features exactly as appeared in the construct for each PTC allele and used the trained random forest models (one trained on data from untreated cells and another on data from cells treated with 0.5 mg/mL G418) to determine readthrough efficiency values based on the features. The predicted readthrough efficiency is positively correlated with average readthrough efficiency measured in both treatment conditions, with Spearman’s rank correlation coefficient of 0.4 in untreated and 0.74 in the G418-treated condition (Figure 3.7A). Excluding the apparent outliers, W882X and W216X, improves the correlation to 0.87 in the G418-treated condition. Predictions from the untreated model performed worse because the model trained on untreated data had poorer performance, or higher NMRSE, than the model trained on G418-treated data (Figure 3.1C). Since the same mRNA features were used to train both models, the poorer performance is likely due to the lower number of mRNAs with detectable readthrough for model training (Figure 3.3A) as well as poorer accuracy of readthrough efficiency values derived from ribosome profiling data of untreated

**Figure 3.6.** Schematic of dual-luciferase (Dual-Luc) reporter used to measure readthrough of CFTR PTC allele. PTC context (stop codon and flanking 3 codons) was inserted between Renilla and Firefly genes. Readthrough efficiency is Firefly signal normalized to Renilla signal in percentage.

| CFTR-G542X: | AUG GUU CUU UGA GAA GGU GGA |
| CFTR-R553X: | GGA GGU CAA UGA GCA AGA AUU |
| CFTR-R1162X: | UCU GUG AGG UGA GUC UUU AAG |
| CFTR-W1282X: | UUG CAA CAG UGA AGG AAA GCC |

Readthrough efficiency = \( \frac{\text{Firefly}}{\text{Renilla}} \times 100\% \)
cells. Unlike G418 treatment where readthrough was induced, evidenced by ribosome footprints in the 3’-UTR predominantly in the same reading frame as in the CDS region, ribosome footprints in the 3’-UTR from untreated cells showed no preference in reading frame (Wangen and Green 2020). Thus, the “noise” from non-readthrough ribosome footprints was proportionally higher in untreated cells than in G418-treated cells, resulting in poorer predictive ability. Thus, it is unsurprising that when we compared changes in readthrough efficiency upon G418 treatment (G418-treated/Untreated) between predicted and measured values, the correlation is weak (Figure 3.7B).
Figure 3.7. Random forest model can accurately predict readthrough of CFTR PTC alleles in G418-treated cells.

A. Readthrough measured by Dual-Luc assay (average of 4-7 replicates) vs. readthrough predicted by random forest model using CFTR PTC allele in Dual-Luc reporter’s sequence.

B. Response to G418 treatment is defined as fold-change of readthrough in G418-treated to untreated condition from A., measured vs. predicted.

C. Comparison of two readthrough prediction schemes: predicted using CFTR PTC allele in reporter’s sequence or predicted using CFTR PTC allele’s native sequence.

D. As in A, but readthrough was predicted using CFTR PTC allele’s native sequence.

E. As in B, but readthrough was predicted using CFTR PTC allele’s native sequence. For all panels, Spearman's correlation coefficient ($\rho$) and the associated p-value is reported.
Three codons flanking each side of a PTC comprise a sufficient context to experimentally measure readthrough efficiency

The random forest models were trained on 75 mRNA features that encompass identities of nucleotides inside and outside the terminating ribosome, properties of amino acids in the ribosome’s exit tunnel, and 3’-UTR length (Figure 3.1A-B and Figure 3.8A, “Full”). On the other hand, the PTC context used in the assay is more limited, only covering 21 mRNA features from the original CFTR mRNA sequence (Figure 3.8A, dark green boxes) and partially affecting 10 nascent peptide tunnel features (Figure 3.8A, light green boxes), meaning 44 of the 75 mRNA features used in the prediction were those of the reporter mRNAs that are constant across different alleles. For broader application of the model, predicting readthrough with native sequence features is desirable. Comparing predictions derived from native CFTR sequences to alleles with reporter sequences revealed very strong correlations, 0.95 for the G418-treated model (Figure 3.7C), suggesting that the 44 mRNA features of the original CFTR mRNA that were not copied to the assay had minimal contributions in the prediction. This is expected since the context used in the assays encompasses all the important features except one, the 3’-UTR length (Figure 3.8A). Additionally, predicted readthrough derived from native CFTR sequences still reflected the Dual-Luc assay measurements quite well in the G418-treated condition (Figure 3.7D), but not in the untreated condition or in response to G418 treatment (Figure 3.7E). The
reduction in correlation compared to the analysis derived from reporter sequences was minimal in G418-treated condition (Figure 3.7D vs. 3.7A), possibly due to the 3'-UTR length feature that varied between PTC alleles in the native sequence but was kept constant in the assay. These results indicate that PTC context used in the assay may be sufficient to measure readthrough efficiency.

To further ensure that mRNA features outside of the context used in Dual-Luc PTC alleles indeed did not contribute to readthrough efficiency determination, we created additional random forest models lacking those mRNA features and compared the prediction errors to those from the full models for G418-treated condition (Figure 3.8A-C). The “Reduced” model contains the combination of features used in the assay and the important features, minus the exit tunnel features (Figure 3.8A, Reduced). Encouragingly, the Reduced model performed as well as the Full model in both regression and classification approaches, as evidenced by unchanged NRMSE and AUROC values (Figure 3.8B-C, Full vs. Reduced). As controls, we created three “Mock” models that had the same number of features as the Reduced model but only include features with low importance scores (Figure 3.8A, Mock1-3). In contrast to the Reduced model, Mock models performed significantly worse than the Full model in both approaches, as evidenced by increases in NRMSE and decreases in AUROC values (Figure 3.8B-V, Full vs. Mock). Because the Dual-Luc assay did not take into account the 3'-UTR length, which is identified as an important predictor of readthrough efficiency in random forest models (Figure 3.1A-B), we next asked whether excluding 3'-UTR
length from the Reduced model significantly changed model performance (Figure 3.8A, Assay). This model performed slightly worse than the Full model, but not significantly (Figure 3.8B-C, Full vs. Assay). These results show that stop codon context alone is sufficient to predict readthrough efficiency.

Together, our findings not only demonstrated broader application of the model in prediction readthrough efficiency but also indicated that three codons flanking each side of a PTC is a sufficient context to experimentally measure readthrough efficiency.
Figure 3.8. Extended stop codon context is sufficient to accurately predict readthrough efficiency.

A. Comparison of mRNA features important in readthrough efficiency prediction and those captured in PTC context in Dual-Luc assay. Feature importance scores are from random forest regression model (%IncMSE) trained on G418 (0.5) data (“Full”), as in Figure 1A. mRNA features varied between different PTC alleles in Dual-Luc assay are in dark green boxes, while those that were combinations of PTC context and reporter sequence are in light green boxes. mRNA features used in different random forest models (y-axis) are indicated by black “X”. Number of mRNA features used in each model excluding the two negative controls, NC) is indicated in parentheses. Compared to the “Full” model, the “Reduced” model omits features unimportant in readthrough efficiency prediction but also maintains minimal features used in Dual-Luc assay. The three “Mock” models serve as controls for the “Reduced” model, having the same number of features as the “Reduced” model. Red "X" for “Mock2” model are features not in the original “Full” model at all but have data type and range similar to the indicated feature to control for model behavior: identities of P- and E-site amino acids were replaced by amino acids 5 and 10 codons upstream of stop codon; 3’-UTR length was replaced by 5’-UTR length. The “Assay” model represents minimal features captured in Dual-Luc assay. B-C. Performance metrics for each random forest model from the regression approach, NRMSE (C), and classification approach, AUROC (D). Purple horizontal line is average of 5-fold cross-validated metric values. Two-tailed Student’s t-test was used to compare each model to the “Full” model, with Benjamini-Hochberg method for multiple-testing correction. Significance levels are indicated by the following: (ns) not significant, (*) p < 0.05, (**) p < 0.01, (***) p < 0.001, (****) p < 0.0001.
Stop codon context in CFTR PTC alleles

Next, we asked what contexts existed in PTC alleles from highest to lowest readthrough measured in the G418-treated condition and whether they are consistent with the results derived from transcriptome-wide ribosome profiling data of G418-treated cells (Figure 3.9). Notably, the allele with highest readthrough, S434X, has a UGAC context (Figure 3.9A). The highest readthrough among UAG and UAA alleles also has C at the nt +4 position (Y1092X and S434X) (Figure 3.9B). Comparing to their UAA counterparts with identical context, S434X-UGA and Y1092X-UAG both have higher readthrough than S434X-UAA and Y1092X-UAA, respectively, as expected (Figure 3.9). These relative comparisons are also reflected in predicted readthrough of these alleles by the Dual-Luc assay (Figure 3.7A).

Among the four UAA alleles, E585X unsurprisingly has the lowest readthrough because it has the most readthrough-inhibiting context UAAA (Figure 3.9B). However, three of the UAA alleles actually have very similar immediate context, UAACU, but their measured readthrough varied, suggesting the roles of other nucleotide positions such as -7, -1, and +9 in mediating readthrough efficiency (Figure 3.9B). In a similar situation, the context of the lowest UGA readthrough allele (W1282X) may not seem that different from the second highest UGA readthrough allele (E60X), but their difference at position -1 may have contributed to their differing readthrough levels (Figure 3.9B). For UAG alleles, the trend is more difficult to discern, but if the two outliers (W882X and W216X) where
the predicted and measured readthrough do not quite agree (Figure 3.7A) are excluded from consideration, the rank order of Y1092X, E60X, and Q493X suggests the role of nt +4 and +9 (Figure 3.9B).

Although W1282X being the lowest UGA readthrough allele is not surprising considering its readthrough-inhibiting context, it is surprising that it actually has even lower readthrough than the lowest UAA readthrough allele, E585X, as evidenced by the discrepancy between the measured and predicted readthrough ranking of the two alleles (Figure 3.7A). This discrepancy, among other outliers, suggested that there exist other features that influence readthrough efficiency that were not included in the model (see Discussion). Nevertheless, the overall high correlation of predictions using the current model and assay measurements (Figure 3.7), as well as the generally expected preference for readthrough-promoting or -inhibiting contexts in highest and lowest readthrough alleles (Figure 3.9), demonstrate value in application of machine learning in readthrough efficiency prediction and validate the use of limited context in measuring readthrough experimentally.

*Figure 3.9. Stop codon context of CFTR PTC alleles.*

**A.** Nucleotide sequences of PTC alleles colored at individual nucleotide position (x-axis) by the effects of nucleotide identity on readthrough efficiency in G418 (0.5) sample as in Figure 2A. PTC alleles are ordered by readthrough efficiency measured by Dual-Luc assay in G418-treated condition (y-axis). **B.** As in A, but grouped by stop codon identity.

(Next page)
Mutations that result in UAA PTCs do not respond well to G418 treatment

One of the major questions in nonsense suppression therapeutics is how responsive a PTC is to drug treatment. Because the model did not give accurate predictions in the untreated condition (Figures 3.7B and 3.7E) to use as baseline readthrough efficiency, we investigated response to G418 treatment using Dual-Luc results. We first examined the relationship between response to G418 treatment and basal readthrough level (Figure 3.10A). It appears that S434X (both UGA and UAA alleles), which already exhibited high basal readthrough in untreated cells, did not respond well to G418 treatment (Figure 3.10A). There was no discernable trend related to basal readthrough for the rest of the alleles (Figure 3.10A). G550X showed the best response to G418 treatment, showing the highest increase in readthrough, although it was impossible to determine the reason without further experiments or more data. However, we observed a trend that UGA alleles had the highest and UAA alleles the lowest increase in readthrough upon G418 treatment (Figure 3.10A-B). Thus, in addition to being readthrough-inhibiting in general, UAA also seems to inhibit G418-mediated induction of readthrough.
Figure 3.10. UAA PTC alleles do not respond well to G418-treatment.
Response to G418 treatment is defined as log₂ fold-change of readthrough in G418-treated to untreated condition. **A.** Response to G418 treatment vs. basal readthrough level in untreated condition. **B.** Response to G418 treatment vs. stop codon identity. Box-plot center line, median; lower and upper hinges, first and third quartiles (the 25th and 75th percentile); whiskers, 1.5x interquartile range; points, actual data. Two-tailed Wilcoxon’s rank sum test with Benjamini-Hochberg method for multiple testing correction was used to perform pairwise comparison. Significance levels are indicated by the following: (ns) not significant, (*) p < 0.05, (**) p < 0.01, (***) p < 0.001, (****) p < 0.0001. For both panels, grey dashed lines at 0 indicate no changes.
Discussion

In this study, we investigated the roles of extended stop codon context, 3′-UTR length, and nascent peptide features in the regulation of readthrough efficiency using random forest machine learning algorithm and published ribosome profiling data of HEK293T cells treated with different aminoglycosides. We confirmed the importance of stop codon and nt +4 identity and found that 3′-UTR length and P-site amino acids also have contributions in readthrough regulation. Moreover, comparing the model’s predictions of new data to actual measurements by Dual-Luc assay revealed potential applications of readthrough prediction in future basic research as well as therapeutics development of diseases caused by nonsense mutations.

Near-cognate tRNA decoding in the A-site likely influences the efficiency of translation termination

Termination vs. readthrough is a competition between eRF1 and a near-cognate tRNA in A-site binding. Other than stop codon contexts, other factors that influence the outcome of this competition include tRNA concentration (Roy et al. 2015) which is not included in the random forest models. This missing information may be one of the reasons for the existence of outliers in the comparison of predicted and measured readthrough efficiency (Figure 3.7). In particular, we noticed that the predictions are the least accurate for UAG PTC alleles; Spearman’s correlation coefficients for the data in Figure 3.7A, G418-treated panel, for each stop codon are 0.94 for UGA, 0.2 for UAG, and 0.8 for UAA.
Although the same amino acids are usually inserted at UAA and UAG stop codons, the near-cognate tRNAs that decode them are not always the same (Roy et al. 2015, 2016; Blanchet et al. 2014; Xue et al. 2017). Perhaps the concentration of some tRNAs that decode UAG are much higher or lower than average, making it a rate-determining or rate-limiting step in the eRF1 vs. tRNA competition and overpowering noticeable influences of other mRNA features. To add to the complexity of this process, stop codon context not only influences readthrough efficiency, but also dictates which tRNA and thus amino acid to be inserted at a PTC (Xue et al. 2017). An additional consideration is the stop codon and near-cognate tRNA base-pairing kinetics. Previous studies in both yeast and human cells demonstrated that the 2\textsuperscript{nd} base (middle) is almost always cognate while mismatches occur at either the 1\textsuperscript{st} or 3\textsuperscript{rd} base (Roy et al. 2015, 2016; Blanchet et al. 2014; Xue et al. 2017). This means that UAA and UGA have the same nucleotides for the 3\textsuperscript{rd} base mismatch, A, while UAG has G. Again, the base-pairing kinetics with G may be rate-determining or rate-limiting for readthrough, resulting in the predictions based on other mRNA features alone being less accurate.

**Potential applications of machine learning models in readthrough prediction**

Apart from a few outliers, we showed that random forest models trained on readthrough data of endogenous mRNAs derived from ribosome profiling experiments can accurately predict readthrough of CFTR PTCs measured by a Dual-luciferase assay, particularly in G418-treated cells (Figure 3.7). The significance of this result is three-fold.
First, this result indicates that readthrough efficiency at NTCs and PTCs share the same “rules” for cis-regulatory elements. Because nonsense mutations happen randomly and are not subject to the same evolutionary pressure as that of NTCs, there has been a concern that information obtained from studying NTC readthrough may not be relevant to studying disease-related nonsense mutations. However, we showed that despite biases in certain context to appear more often or occur less frequently in the transcriptome (Figure 3.2C), the endogenous mRNA NTC contexts are diverse enough to allow accurate prediction of readthrough of PTCs (Figure 3.7).

Second, this result validates the use of a small fragment of native PTC context in studying readthrough. Studies of readthrough usually involve only a fragment of native sequence near the stop codon, which varies between studies. It was unclear whether this small fragment would be sufficient to mimic readthrough of PTCs with the whole-gene native sequence that exist in patients. Our trained model includes larger native sequence contexts than are used in the Dual-Luciferase assay, but the prediction and measured readthrough efficiency are still consistent with each other. Hence, we provided evidence that minimal context that contains the most influential mRNA features is sufficient to experimentally measure readthrough.

Lastly, our results demonstrate that a machine learning model trained on “big data” can predict an assay outcome. Studying readthrough of nonsense mutations in disease patients has mostly required difficult-to-estimate, indirect
measurements of readthrough (McDonald et al. 2013, 2022; Aslam et al. 2017, 2023). The Dual-Luciferase assay, although commercialized and streamlined for ease of use, can still be tedious if there are many mutations and contexts to test. Thus, the high predictive ability of a machine learning model may help with future readthrough experimental design, saving time, labor, and resources. Although basal readthrough prediction is not as accurate, making studying fold-change in drug treatment more difficult to predict, it might not matter because the most important information for therapeutic purposes is readthrough level after drug treatment, which we showed to be accurate. Thus, the application of readthrough prediction may also extend to aiding clinical trial design for nonsense suppression therapies, predicting whether a patient will respond well to a readthrough drug given the nonsense mutation’s sequence context.
Materials and Methods

Data acquisition

Readthrough efficiency data for HEK293T cells treated with different aminoglycosides were downloaded from Wangen and Green, eLife (2020), Figure 2 – source data 1 (https://cdn.elifesciences.org/articles/52611/elife-52611-fig2-data1-v2.xlsx) (Wangen and Green 2020) and log2-transformed for all analyses.

Sequences of spliced mRNAs were downloaded from Ensembl using R package biomaRt (Durinck et al. 2009) according to Ensembl Transcript ID provided in Figure 2 – source data 1. The final HEK293T mRNA isoforms selection was described in details in the original publication (Wangen and Green 2020) and mRNA region lengths provided in Figure 2 – source data 1 were adjusted accordingly to reflect actual lengths as well as to consider the stop codon as part of the 3’-UTR. For yeast, the longest UTR entry was chosen for mRNAs with multiple annotations across studies (Zhang and Dietrich 2005; Miura et al. 2006; Nagalakshmi et al. 2008; Xu et al. 2009; Yassour et al. 2009) deposited in the YeastMine database (as of July 3, 2017) and mRNA region lengths were adjusted accordingly to consider the stop codon as part of the 3’-UTR. Other HEK293T mRNA sequence features were defined as described previously for yeast in Table 2.2 of Chapter II (Mangkalaphiban et al. 2021).

Random forest models and statistical analyses

Analyses were performed in the R programming environment with the following R packages: readxl, data.table, dplyr, reshape2, biomaRt, randomForest,
caret, Biostrings, seqinr, rstatix, ggplot2, ggpubr, ggh4x, ggrepel, patchwork, and Cairo.

For each sample, mRNAs with too little read coverage and undetectable readthrough, where RPKM of CDS < 5 and RPKM of the extension region (3’-UTR region between the canonical stop codon and next in-frame downstream stop codon) < 0.5, were discarded from further analyses. The remaining data was used to create a random forest regression model and for comparative analyses. The top 15% and bottom 15% of the remaining data ranked by readthrough efficiency were assigned as “high” and “low” readthrough mRNAs, respectively, and used to create a random forest classification model.

Random forest models were trained to predict readthrough efficiency (regression) or readthrough groups (classification) with 100 trees and 5-fold cross-validation to optimize the number of features allowed for splitting at each node (mtry hyperparameter). For additional models in Figure 3.8, the same parameters were applied except mtry was kept as default. Performance metrics extracted from each model are root mean squared error (RMSE) normalized to the range of Y variable (readthrough efficiency) for the regression model and area under the receiver operating characteristic (AUROC) for the classification model. Feature importance metrics, which indicate the predictive ability of each mRNA feature, extracted from the final model are % increase in mean squared error (%IncMSE) for the regression model and Mean Decrease Accuracy (MDA) for the classification model.
Comparative analyses between median readthrough efficiency of a group of mRNAs (defined by identity of stop codon, nucleotide, or codon triplet) and median readthrough efficiency of all mRNAs in a sample (“sample median”) were performed by two-tailed Wilcoxon’s rank sum test with Benjamini-Hochberg multiple testing correction.

**Generation of dual-luciferase hCFTR PTC reporters**

Sense/antisense oligonucleotides containing hCFTR PTC contexts (PTC plus 3 codons of upstream and downstream hCFTR context) were annealed and cloned into the Ascl/Sbf1 sites of dual-luc Ascl G542X 11 codon SbfI/pcDNA3.1 Zeo+ (pDB1497). The annealed oligonucleotides were inserted between upstream Renilla and downstream firefly luciferase genes. The firefly activity can only be detected when readthrough of the PTC occurs. A total of 15 dual-luciferase reporters containing different hCFTR PTC contexts (6 UGAs, 5 UAGs, and 4 UAAs) were generated.

For readthrough prediction of hCFTR PTC reporters, the sequences of luciferase genes were used to determine mRNA features outside of the PTC contexts. The 3’-UTR length is defined as the number of nucleotides from PTC to the bGH poly(A) signal in the reporter.

**Dual-luciferase reporter assay to measure PTC readthrough in HEK293 cells**

The dual-luciferase reporter constructs were transiently transfected into HEK293 cells (CLS Cat# 300192/p777_HEK293, RRID:CVCL_0045) to test readthrough efficiency. HEK293 cells were seeded into 96-well plates at 2x10^4
cells/well. Twenty-four hours after seeding, the cells were transfected with 0.1 µg DNA/well using Lipofectamine LTX regents (ThermoFisher Cat# 15338500). Three hours after transfection, the cells were treated with 100 µg/ml G418. After 24 hours of treatment, dual luciferase assays were performed using a Dual-Luciferase Reporter Assay System (Promega Cat# E1960) with a Glomax Discover Microplate Reader (Promega) to measure both Renilla and firefly activities. The firefly activity normalized to the Renilla activity was calculated as the readthrough level. Three parameters were reported from this assay: 1) the basal readthrough in untreated cells; 2) the readthrough induced by 100 µg/ml G418; and 3) the fold-increase in readthrough when comparing G418-treated and untreated samples. Each experiment included treated and untreated samples assayed in quadruplicate wells in a 96-well plate. Four to seven independent experiments were performed for each hCFTR PTC dual-luc reporter.

Data Availability

Processed data tables and scripts used to acquire, analyze, and visualize data are available at https://github.com/Jacobson-Lab/AG_readthrough.

Acknowledgments

This work was supported by grants to: A.J. from the U.S. National Institutes of Health (1R35GM122468), Andrei Korostelev and A.J. from the Cystic Fibrosis Foundation (CFFKOROST20GO), and D.B. from the Cystic Fibrosis Foundation (BEDWEL19G0 and Southern Research subaward 858345). We thank members of the Jacobson lab for comments on the manuscript.
Author Contributions

K.M. and A.J. conceived and designed the analysis; K.M. carried out the bioinformatics analysis; L.F., M.D., K.T., and K.K. conceived and designed the dual-luciferase hCFTR readthrough experiments; L.F., M.D., and K.T. carried out the dual-luciferase hCFTR experiments; K.M., L.F., M.D., K.T., K.K., D.B., and A.J. analyzed the data; K.M. and A.J. wrote the paper; A.J. and D.B. obtained funding for the study.

Declaration of Interests

A.J. is co-founder, director, and consultant for PTC Therapeutics Inc. D.B. is a consultant for PTC Therapeutics Inc. K.M., L.F., M.D., K.T., and K.K declare no competing interests.
CHAPTER IV:

Recognition of inefficient translation termination and NMD: Upf1’s action on ribosomes

The work in this chapter has been published as:


§ Robin Ganesan and Kotchaphorn Mangkalaphiban are co-first authors with equivalent contributions. R. G. created yeast strains, generated high-throughput sequencing data, performed all wet lab experiments, and wrote the paper. I optimized rRNA depletion strategy for ribosome profiling library preparation, performed extensive bioinformatics and statistical analyses, and wrote the paper.
Abstract

Upf1, Upf2, and Upf3, the central regulators of nonsense-mediated mRNA decay (NMD), appear to exercise their NMD functions while bound to elongating ribosomes, and evidence for this conclusion is particularly compelling for Upf1. Hence, we employed selective profiling of yeast Upf1:ribosome association to define that step in greater detail, understand whether the nature of the mRNA being translated influences Upf1:80S interaction, and elucidate the functions of ribosome-associated Upf1. Our approach has allowed us to clarify the timing and specificity of Upf1 association with translating ribosomes, obtain evidence for a Upf1 mRNA surveillance function that precedes the activation of NMD, identify a unique ribosome state that generates 37-43 nt ribosome footprints whose accumulation is dependent on Upf1’s ATPase activity, and demonstrate that a mutated form of Upf1 can interfere with normal translation termination and ribosome release. In addition, our results strongly support the existence of at least two distinct functional Upf1 complexes in the NMD pathway.

Introduction

Nonsense-mediated mRNA decay (NMD) is a eukaryotic translation-dependent mRNA quality control pathway whose central regulators are the three Upf proteins, Upf1, Upf2, and Upf3 (He and Jacobson 2015b; Kurosaki et al. 2019). NMD is initiated in response to atypical translation termination events, e.g., when an elongating ribosome encounters a termination codon that occurs prematurely within an open reading frame or in a context that otherwise renders termination
inefficient (Amrani et al. 2004; Wu et al. 2020). Although NMD’s existence has been known for decades (He et al. 1993; Leeds et al. 1991; Losson and Lacroute 1979; Maquat 2000), and the pathway has been studied extensively (He and Jacobson 2015b; Kurosaki et al. 2019), the precise mechanism by which the Upf proteins recognize a ribosome undergoing an atypical termination event and respond to it by triggering accelerated degradation of the associated mRNA remains unknown. Several observations have indicated that the NMD functions of the yeast Upf proteins are exercised while these factors are bound to ribosomes, and the evidence for this conclusion is particularly compelling for Upf1. This includes experiments demonstrating: i) co-sedimentation of Upf1 with polyribosomes (Atkin et al. 1997; Kashima et al. 2006; Mangus and Jacobson 1999), ii) retained association of Upf1 with ribosomes when cytoplasmic extracts are digested with RNase A (Atkin et al. 1995) or micrococcal nuclease (Mangus and Jacobson 1999), iii) substantial resistance of Upf1:ribosome association to prior treatment with high salt (Celik 2017; Ghosh et al. 2010; Mangus and Jacobson 1999; Min et al. 2013), iv) two-hybrid interaction of Upf1 with the 40S ribosomal subunit protein Rps26 and elimination of this interaction by specific Upf1 CH domain C62Y and C84S mutations (Min et al. 2013), v) interaction of Upf1 with the release factors (Czapliński et al. 1998; Ivanov et al. 2008; Kashima et al. 2006; Wang et al. 2001), and vi) a requirement for Upf1 in post-termination ribosome reutilization (Ghosh et al. 2010). As in yeast, Upf1 migrates with polyribosomal sucrose gradient fractions in mammalian cell extracts (Kurosaki et al. 2018; López-Perrote et al. 2016; Nott
et al. 2004; Yoshikawa et al. 2018). However, human Upf1 appears to bind mRNAs and be displaced from their coding regions to their 3'-UTRs by elongating ribosomes (Hogg and Goff 2010; Hurt et al. 2013; Kurosaki and Maquat 2013; Zünd et al. 2013). The latter observations may reflect the involvement of Upf1 in numerous mammalian mRNA decay pathways other than NMD (Kim and Maquat 2019) or imply that Upf1 may respond \textit{in trans} to an NMD-activating event during translation termination, possibly as part of a “two-factor authentication” process (Boehm et al. 2021).

In yeast, NMD substrates are generally thought to be decapped without prior deadenylation and to be subsequently degraded by the 5' to 3' exonuclease, Xrn1 (Muhlrad and Parker 1994). Recently, we demonstrated that Upf1 acts as a decapping activator for yeast NMD, controlling both substrate specificity and activation of the decapping enzyme by binding to two specific regulatory motifs in the Dcp2 C-terminal domain (He et al. 2022). Significantly, the same study demonstrated that decapping is not rate-limiting for NMD substrates and that the major function for Upf1 in NMD is imparted at a step upstream of mRNA decapping. These observations indicate that there are substantial unknown aspects of Upf1 association with mRNAs or prematurely terminating ribosomes and the consequences of such association for determining a transcript’s eligibility for NMD and/or for committing it to the NMD pathway.

Regardless of where Upf1 and the other Upf factors reside while awaiting their activation and involvement in mRNA decay, it appears that the steps in which
these proteins are committed to NMD functions are likely to be localized to ribosomes (He and Jacobson 2015b; Kurosaki et al. 2019). Hence, we employed selective ribosome profiling (Becker et al. 2013) of yeast Upf1 to define the step and the nature of the mRNA being translated on Upf1:80S interaction, and to elucidate the functions of ribosome-associated Upf1. Our approach has allowed us to clarify the timing and specificity of Upf1 association with translating ribosomes, obtain evidence for a Upf1 mRNA surveillance function that precedes the activation of NMD, identify a unique ribosome state that generates 37-43 nt footprints whose accumulation is dependent on Upf1’s ATPase activity, and demonstrate that a mutated form of Upf1 can interfere with normal translation termination.

**Results**

**Rationale for the methodological approaches**

To characterize Upf1 association with polyribosomes, we carried out selective ribosome profiling analyses (Becker et al. 2013) of yeast cells expressing FLAG-tagged alleles of *UPF1*. To avoid recovery of Upf1 associated with extra-ribosomal complexes (Atkin et al. 1995; Mangus and Jacobson 1999), FLAG-based immunopurification was applied to RNase I-digested ribosomes that had been pelleted through a 1M sucrose cushion, a procedure yielding relatively pure 80S ribosome preparations (Figure 4.1). To address possible methodological or biological variables arising from tag location, *UPF1* constructs with N- or C-terminal FLAG epitopes were examined. Although several protocols (Becker et al. 2013;
Döring et al. 2017; Wagner et al. 2020) recommend that selective ribosome profiling be carried out with C-terminally tagged proteins to preclude sequencing reads generated by immunopurification of ribosomes with a nascent N-terminally tagged protein, we simply eliminated all Upf1-specific reads bioinformatically. To increase the possibility of characterizing steps in NMD upstream of the actual activation of mRNA decay, we examined cells with or without full NMD activity, including cells harboring upf mutations and cells that were briefly treated or not treated with cycloheximide (CHX) during harvesting and lysis. CHX inhibits translation elongation and CHX treatment has been shown to inhibit NMD in vivo (Zhang et al. 1997) and to block the formation of toeprints specific to premature termination codons in vitro (Amrani et al. 2004).

Pilot experiments in which FLAG-tagged Upf1 was expressed from centromeric plasmids did not yield sufficient immunopurified ribosomes for construction of substantive ribosome profiling libraries. Hence, episomally expressed UPF1 genes were used for all experiments. The increased expression inherent to such constructs raises the possibility that results obtained with their use might not reflect bona fide Upf1 functions or distribution across an mRNA coding region. However, several results mitigate this concern: i) in vitro incubation of yeast Upf1 and 80S ribosomes at Upf1:80S ratios considerably higher than those achieved by in vivo episomal expression nevertheless yielded highly specific Upf1:80S interaction (Schuller et al. 2018), ii) the distribution of WT Upf1-FLAG across different polysome fractions was comparable when the protein was
expressed from centromeric or episomal vectors in WT *UPF2* strains, even when free Upf1 was increased (Figure 4.2, middle, fractions 11 and 12, episomal Upf1-FLAG), whereas increased migration of Upf1-FLAG into heavier polysomes was observed when *UPF2* was deleted, and iii) episomal expression of *UPF1* displayed a similar NMD phenotype compared to centromeric *UPF1*, causing only slight inhibition of NMD relative to endogenous *UPF1* (Figure 4.3). Consistent with the latter northern blotting results, our RNA-Seq analyses showed that episomally expressed *UPF1* manifested negligible dominant-negative activity, i.e., yeast NMD substrates (Celik et al. 2017a) showed only small increases in abundance as compared to the substantial increases in abundance seen in *upf1DE572AA* and *upf2Δ* strains in which NMD is inactivated (Figure 4.4). Cells harboring episomally expressed *UPF1* thus appear to maintain near normal NMD function, a conclusion further supported by additional experiments described below, including those in which we: i) episomally expressed all three Upf proteins simultaneously, ii) employed *upf1* or *upf2* mutations to determine if a particular result depended on Upf1 or NMD function, and iii) tested whether Upf1:ribosome complexes manifested stoichiometric relationships between Upf1 and ribosomal proteins.
Figure 4.1. Electron micrographs of RNase I-digested 80S ribosomes.
Representative micrographs of negatively stained 80S ribosomes from FLAG-UPF1 (prepared with or without cycloheximide), and UPF1-FLAG, upf1DE572AA-FLAG, and UPF1-FLAG/upf2Δ (prepared with cycloheximide) were captured at 87,000X magnification. Scale bar of 200nm is shown.
Figure 4.2. Polysomal distributions of wild-type and mutant forms of Upf1.
Top, Representative polsome trace from a upf1Δ strain containing centromerically expressed UPF1-FLAG across a 7-47% (right to left) sucrose gradient. Middle, comparison of Upf1 polysomal distribution between centromeric (light blue) and episomal (dark blue) UPF1-FLAG. Western blots were probed with α-FLAG antibody and the mean percent total Upf1-FLAG signal per fraction across sucrose gradient fractions 1-10 from three replicate experiments was calculated and plotted with standard error of the mean. Bottom, representative western blots of centromeric UPF1-FLAG and episomal UPF1-FLAG, upf1DE572AA-FLAG and UPF1-FLAG in a upf2Δ strain background. Blots were probed with α-FLAG antibody. Arrow denotes migration of Upf1 protein.
Figure 4.3. NMD phenotypes of UPF1 mutants.
Top, Northern blots of mRNA isolated from two replicates of WT or upf1Δ cells containing an empty vector (EV) or from upf1Δ cells expressing UPF1-FLAG or its mutant alleles from centromeric (top panel) or episomal vectors (bottom panel). Blots probed for CYH2 mRNA and pre-mRNA and for SCR1 as a loading control. Bottom, Ratio of CYH2 pre-mRNA:mRNA for the strains depicted in the top panels, mean and range. WT EV, wild-type UPF1 strain containing empty vector; upf1Δ EV, upf1Δ strain containing empty vector.
Figure 4.4. NMD substrates were stabilized in episomally expressed FLAG-tagged UPF1 strains.

Differential expression analysis was performed by DESeq2 to determine changes in total mRNA abundance between each FLAG-tagged UPF1 strain and its isogenic WT + EV (empty vector) for each CHX condition. Analyzed differential expression data from three replicates each of upf1Δ vs WT, upf2Δ vs WT, and upf3Δ vs WT strains were obtained from Celik et al., 2017. Mean of normalized counts were plotted against log₂ fold change between the strain of interest and its associated WT, with grey dashed line at log₂ fold change of zero to indicate no change. The plot’s y-axis is limited to -10 and 10 for better visualization. Data points with log₂ fold change higher than 10 were plotted at 10 as triangles, and data points with log₂ fold change lower than -10 were plotted at -10 as inverted triangles. Only NMD substrates that showed significant changes (FDR-adjusted p-value < 0.01) between the two strains were shown. The number of significantly differentially expressed NMD substrates on each plot is provided.
Purification of FLAG-Upf1- and Upf1-FLAG-associated ribosomes yields stoichiometric recovery of ribosomal proteins

Confirmation of the specificity of recovery of Upf1-associated ribosomes followed from mass spectrometry and intensity-based absolute quantification of protein abundance (iBAQ) (Schwanhäusser et al. 2011, 2013) analyses of ribosomes from two biological replicates of FLAG-UPF1 strains or three biological replicates of UPF1-FLAG strains harvested with or without addition of CHX. The normalized iBAQ values for the total and immunopurified samples were compared to each other in scatterplots and volcano plots (Figure 4.5). Using ribosomes from cells untreated or briefly treated with CHX during cell harvesting and lysis, immunopurification resulted in substantial enrichment of FLAG-tagged Upf1 (Figure 4.5B, red dots) such that, in all cases, its increased abundance after immunopurification approximated stoichiometry with the ribosomal proteins recovered in the samples (Figure 4.5A, blue dots, and Figure 4.6), a result consistent with copurification of Upf1 with ribosomes. Specificity of the FLAG immunopurification procedure was demonstrated by greatly reduced recovery of ribosomes after immunopurification (Figure 4.7, α-Rps6), and the lack of increased Upf1:ribosomal protein stoichiometry, when the same procedure was applied to ribosomes from control cells expressing 6XHis-tagged UPF1 (Figure 4.5A and Figure 4.6). The stoichiometry of Upf1:ribosomal proteins after immunopurification of ribosomes from cells expressing FLAG-UPF1 treated or not treated with CHX at the time of harvesting was similar (Figure 4.5A, red dots, Y axis values). Likewise,
Upf1 derived from the C-terminally FLAG-tagged UPF1 allele expressed in different genetic backgrounds was also recovered stoichiometrically with ribosomal proteins, indicating that tag location, deletion of UPF2, and the ATPase-inactivating upf1DE572AA mutation did not affect Upf1:ribosome association or its recovery. As expected for ribosomes subjected to RNase I digestion, immunopurification of Upf1-associated ribosomes did not lead to uniform recovery of all ribosomal proteins. Among those that are depleted in most IP samples are Rpp2A and Rpp2B (Figures 4.5A and B), components of the ribosomal stalk (Hanson et al. 2004).

With the exception of the FLAG-tagged UPF1 gene, the plasmid selective markers, and the gene disruption cassette selective markers in upf1Δ and upf2Δ strains, all genes in these strains were present at their normal copy numbers. Thus, it was not surprising that immunopurification of FLAG-Upf1 or Upf1-FLAG did not lead to co-recovery of the interacting Upf factors, Upf2 and Upf3. However, detectable levels of Upf2 and Upf3 were recovered after immunopurification of ribosomes from FLAG-UPF1 cells in which both untagged UPF2 and UPF3 were also expressed from episomal vectors (Figure 4.5A, FLAG-UPF1 UPF2/3 EE +CHX), demonstrating that their lack of detection in other samples was due to lower relative abundance of endogenous Upf2 and Upf3 and not the consequence of loss during the immunopurification procedure. Immunopurification of ribosomes from the FLAG-UPF1 UPF2/3 EE +CHX cells also led to co-recovery of the release factor eRF1 and enrichment for the mRNA decapping factors Edc3 and Dcp2 (Figure 4.5A), factors that were not recovered in any of the other IP samples.
However, many other proteins including translation initiation and elongation factors, and proteins unrelated to translation, mRNA decay, or protein folding were also enriched only in samples from this strain (Figure 4.5B), so it is uncertain whether co-recovery of release and decapping factors was specific to the presence of Upf2 and Upf3 on the ribosomes or conversely, whether simultaneous increased expression of all three Upf factors resulted in non-specific recovery of these and other proteins. Having demonstrated the specificity of our immunopurification procedure, subsequent experiments employed ribosome profiling and RNA-Seq analyses of cells expressing FLAG-tagged Upf1 episomally.
Figure 4.5. Stoichiometric recovery of FLAG-tagged Upf1 with ribosomal proteins.

Input (Total) and immunopurified ribosomes (IP) from different lysates were analyzed by mass spectrometry. Intensity based absolute quantification (iBAQ) was performed and normalized across biosamples. A. Average iBAQ of biological replicates, log_{10}-transformed, of proteins identified from IP were plotted against those from the Total ribosomes. Proteins exclusively identified in either Total or IP are plotted on the x- or y-axis, respectively. B. Differential abundance analysis was performed for proteins detected in both IP and Total samples using R package limma (Kammers et al. 2015; Smyth 2004). Negative log_{10} p-values adjusted by Benjamini-Hochberg method were plotted against log_2 fold change in protein abundance in IP over Total. Gray vertical dashed line indicates log_2 fold change of zero (no change). Positive log_2 fold change (right of vertical line) are proteins enriched in IP. Negative log_2 fold change (left of vertical line) are proteins depleted in IP. Gray horizontal dashed line indicates the cutoff of adjusted p-value at 0.05; proteins above this cutoff have significant changes. Two biological replicates of FLAG-UPF1 strains, three biological replicates of UPF1-FLAG strains, and one sample of the negative control experiment (6XHis-UPF1) were analyzed.
Figure 4.6. Stoichiometric recovery of FLAG-tagged Upf1 with ribosomal proteins.
Average normalized iBAQ values, quantified from mass spectrometry data, of 100 most abundant proteins (in descending order) in each sample. Red arrow denotes Upf1 protein.

Figure 4.7. Immunopurification of ribosomes associated with FLAG-tagged Upf1.
Representative western blot of aliquots of total and immunopurified ribosomes (IP) derived from strains expressing episomal UPF1-FLAG or 6XHis-Upf1. Ribosomes were immunopurified with α-FLAG resin. Equal amounts of input ribosomes (by A_{260}) and equal % recovered material (as a fraction of final volume) were loaded into the respective lanes. Blot was probed with α-FLAG and α-Rps6 antibody. All panels are from the same blot.
Upf1 association with 80S ribosomes in CHX-treated cells promotes the formation of atypical ribosome-protected mRNA fragments

Ribo-Seq libraries were prepared from both immunopurified ribosomes and the respective pre-purification total ribosomes derived from the complete set of strains expressing FLAG-tagged Upf1, as well as from wild-type cells harboring an empty vector (EV) control. We first examined the nature of the ribosome protected fragments recovered in all libraries. Analyses of ribosome protected footprint length distribution in FLAG-UPF1 libraries from cells without CHX treatment showed that footprints of ~20-23 nt in length (“small” size, hereafter denoted as “S”) were predominant, followed by ~27-32 nt footprints (“medium” size, hereafter denoted as “M”; Figure 2A, green lines). These previously detected (Wu et al. 2019) footprint sizes were also present in control cells lacking FLAG-tagged UPF1 and in samples with or without prior immunopurification of Upf1-associated ribosomes. In the presence of CHX, the predominant footprint sizes from total or immunopurified ribosomes were the M size class; this is expected since CHX is known to freeze ribosomes with occupied A sites, yielding footprints ~28 nt in length (Lareau et al. 2014; Wu et al. 2019) (Figure 4.8A, red lines). Notably, ribosomes from CHX-treated strains expressing FLAG-UPF1 yielded an atypical footprint size class of ~37-43 nt (“large” size, hereafter denoted as “L”) (Figure 4.8A, red solid lines). The L footprints represent translating ribosomes (i.e., they are neither RNA binding complexes nor scanning ribosomes), as do the typical S and M footprints, because they exhibited the 3-nt periodicity expected of translating
ribosomes (Figure 4.9). The ratio of L:M footprints was increased in the FLAG-UPF1 IP libraries compared to total libraries (Figure 4.8A, red solid lines), and the L footprints were not detectable in libraries prepared from cells without CHX treatment (Figure 4.8A, green lines).

We considered the possibility that the atypical L footprints might be a consequence of non-stoichiometric expression of UPF2 or UPF3 relative to UPF1 in cells expressing episomal FLAG-UPF1, but concurrent episomal expression of UPF2, UPF3, and FLAG-UPF1 still yielded L footprints from immunopurified ribosomes (Figure 4.8A, bottom panel). Likewise, we considered the possibility that the L footprints were caused by the 5'-FLAG epitope on Upf1, but analysis of libraries generated from cells expressing UPF1-FLAG again showed recovery of the L footprints in immunopurified ribosomes (Figure 4.8B). Therefore, this unique class of footprints is most likely specific to Upf1 association with the 80S ribosome.
Figure 4.8. Distribution of footprint length for each strain and cycloheximide treatment condition.
Fractions of each footprint length (nt) were calculated and averaged among replicates. Green and red lines indicate libraries prepared in the absence and presence of cycloheximide, respectively. Dashed and solid lines represent total and immunopurified (IP) ribosomes, respectively. Gray shaded areas highlight different classes of footprint size: from left to right, 20-23 nt = small (S), 27-32 nt = medium (M), and 37-43 nt = large (L). A. Data from libraries prepared from two biological replicates of N-terminally FLAG-tagged strains and their WT+EV control. B. Data from libraries prepared from three biological replicates of C-terminally FLAG-tagged strains and their WT+EV control.
A

Footprint size

<table>
<thead>
<tr>
<th>S</th>
<th>M</th>
<th>L</th>
</tr>
</thead>
</table>

Distance from start (nt) | Distance from stop (nt)

| Distance from start (nt) | Distance from stop (nt) |

IP | Total

B

Footprint size

<table>
<thead>
<tr>
<th>S</th>
<th>M</th>
<th>L</th>
</tr>
</thead>
</table>

Distance from start (nt) | Distance from stop (nt)

| Distance from start (nt) | Distance from stop (nt) |

IP | Total
Figure 4.9. Abundance of ribosome footprints relative to canonical start or stop codons for immunopurified (IP) or total ribosomes (total).

A. WT+EV and FLAG-UPF1 strains treated or untreated with CHX. B. WT+EV, UPF1-FLAG, UPF1-FLAG/upf2Δ, and upf1DE572AA-FLAG strains treated with CHX. For each footprint size, the number of footprint P-sites at each nucleotide position relative to start or stop codon of an mRNA is normalized to the total number of footprints of that size in the library and converted to percentage. The percentage of footprint counts in the first and last 100 nt of mRNAs were plotted. mRNAs whose coding region is less than 200 nt in length were disregarded from the plot to avoid repeated counting. Vertical grey solid lines signify the first nucleotide of the start and stop codons. Vertical grey dashed lines occur every 3 nucleotides to indicate reads in-frame with the start codon.

Upf1 function but not full activity of the NMD pathway is required for formation of atypical footprints by Upf1-associated ribosomes

To understand the origin of the L footprints, we determined whether their formation required Upf1 function, as well as function of the entire NMD pathway. Total and immunopurified ribosomes from CHX-treated upf1Δ strains harboring the upf1DE572AA-FLAG allele, or UPF2 or upf2Δ strains harboring the UPF1-FLAG allele, were subjected to ribosome profiling and analysis of footprint length distribution (Figure 4.8B). The L footprints were undetectable in any libraries from total ribosomes but were recovered in libraries prepared from immunopurified ribosomes isolated from UPF2 or upf2Δ cells expressing UPF1-FLAG. The results indicate that full functionality of the NMD pathway is not required to form L footprints and that the position of the FLAG epitope does not influence L footprint formation. However, the L footprints were undetectable in libraries prepared from the upf1DE572AA-FLAG strain (Figure 4.8B), indicating that full Upf1 function is required for their formation. Upf1-DE572AA is present on polysomes (Figure 4.2)
and is able to interact with Rps26 in a two-hybrid assay (Min et al. 2013); therefore, the loss of the L footprints in cells expressing upf1DE572AA-FLAG is likely due to the inability of this mutant Upf1 to promote a ribosome-associated function that requires Upf1’s ATP hydrolysis activity.

**L footprints are generated by additional protection of mRNA 5’ to the normal ribosome-protected segment**

The recovery of L footprints from CHX-treated cells expressing FLAG-UPF1 or UPF1-FLAG (Figures 4.8A and B) suggests that a fraction of Upf1 is bound to ribosomes in a configuration that inhibits RNase I digestion of mRNA that usually occurs at the edge of the ribosome during library preparation. Mapping of the 5’ and 3’ ends of the M and L footprints recovered from Upf1-associated ribosomes using the start and stop codons as reference points showed that the 3’ ends but not the 5’ ends of M and L footprints are aligned at the same nucleotide location (Figure 4.10). Thus, the size difference between M and L footprints is entirely attributable to an extension on the 5’ side of the normal ribosome-protected fragment. The 5’ ends of the typical M footprints are ~12-13 nt upstream from the reference codon (start or stop codon), while the 5’ ends of the atypical L footprints are ~23-25 nt upstream from the reference codon, regardless of the N- or C-terminal placement of the FLAG epitope, confirming that the extra 10-13 nt that extend nuclease protection do so on the 5’ side of the fragment. This corresponds to the region of the mRNA near the exit channel of the ribosome, the same region in which Upf1:80S interaction was observed *in vitro* (Schuller et al. 2018).
Figure 4.10. Atypically large footprints are attributable to a 5’ extension of the mRNA region protected by the ribosome. Mapping of the 5’ and 3’ ends of footprint lengths 25-45 nt relative to the start and stop codons (red dashed lines) from IP libraries of FLAG-UPF1, UPF1-FLAG, and UPF1-FLAG/upf2Δ (+CHX). For each footprint length, footprint count at each nucleotide position relative to the start or stop reference codons were normalized to the total footprint count in the library, and these normalized counts from biological replicates were averaged. Different footprint lengths align at the 3’ ends but diverge at the 5’ ends. L and M fragment size classes are as indicated for each sample. Lines at the bottom of the figure indicate the approximate 5’ and 3’ limits of the L and M fragments over the start and stop reference codons.
L footprints reflect an early phase of Upf1 association with polyribosomes

Metagene analyses comparing the distributions of S, M, and L ribosome-protected fragments across normalized coding regions from total and immunopurified ribosomes from FLAG-UPF1 cells show that, in the absence of CHX, the S and M footprints from immunopurified ribosomes are markedly underrepresented in approximately the first half of the coding region and become overrepresented in the second half of the coding region compared to total ribosomes (Figure 4.11A; Figures 4.12A and 4.12D; Figure 4.9A). The same pattern holds for the M footprints in libraries from CHX-treated cells (Figures 4.11B and 4.11C, top panels; Figures 4.9A, 4.12B and 4.12C, top panels, 4.12E and 4.12F). These observations suggest that Upf1 is not stripped off the coding region by translating ribosomes (Hogg and Goff 2010; Kurosaki and Maquat 2013; Zünd et al. 2013) and that Upf1 association with ribosomes occurs routinely during the course of translation elongation. The metagene distribution of S and M footprints recovered from immunopurified ribosomes was nearly identical regardless of the presence or absence of CHX (Figure 4.11D, bottom panel), except for a small peak over the start codon in M footprints in either condition (Figure 4.9A), indicating that the observed progressive increase in Upf1-associated ribosomes that form S and M footprints across the coding region is not CHX-dependent.

The L footprints, which are only detectable with Upf1-associated ribosomes in CHX-treated cells, accumulate rapidly at the beginning of mRNA coding regions and slowly decrease their relative accumulation across the entire coding region in
libraries prepared from immunopurified ribosomes from FLAG-UPF1, UPF1-FLAG, or UPF1-FLAG/upf2Δ strains (Figures 4.11B and 4.11C, bottom panels; Figures 4.12B and 4.12C, bottom panels; Table 4.1). The relative accumulation of L footprints toward the 5' end of mRNAs is significantly higher than the relative accumulation of M footprints (Table 4.1). These patterns, and their expected 3-nt periodicity, are also evident in higher resolution metagene plots of the first and last 100 nt of the coding regions (Figure 4.9). Notably, L footprints do not accumulate over the start codon as do M footprints (Figure 4.9, green lines). The rapid appearance of L footprints relative to ribosome progression across mRNA coding regions (Figures 4.11, Figure 4.9) indicates that their formation commences early during translation and is either stabilized by the CHX-mediated elongation block or that Upf1 may prefer to bind to ribosomes which are in the pre-translocation state trapped by CHX (Wu et al. 2019; Zhao et al. 2021).
**Figure 4.11.** Upf1 progressively associates with ribosomes across mRNA coding regions except when Upf1 association forms L footprints.
Distribution of footprint abundance across the coding region (CDS) for each footprint size class from total (purple) or IP (orange) ribosomes profiling libraries. 

**A.** Small (S) and medium (M) footprint distribution from strains expressing WT+EV and FLAG-UPF1 in the absence of CHX treatment. 

**B.** Medium (M) and large (L) footprint distribution from strains expressing WT+EV and FLAG-UPF1 in the presence of CHX. 

**C.** Medium (M) and large (L) footprint distribution from CHX-treated cells expressing WT+EV, UPF1-FLAG, UPF1-FLAG/upf2Δ, or upf1DE572AA-FLAG. Each gene’s CDS was divided into 100 bins and the percentage of footprints’ P-sites belonging to each bin in all genes was calculated. Grey dashed line signifies a theoretical number where each percentage of CDS contains an equivalent number of footprints for a sum of 100% across all 100 bins. 

**D.** Distribution of dominant footprint (red = Medium, M for +CHX; green = Small, S for -CHX) abundance across the coding region. Data is the same as used in **A** and **B** top panels but re-plotted in such a way to directly observe the differences or similarity in footprint distribution between the absence and presence of CHX during the library preparation of each strain.
Figure 4.12. Upf1 progressively associates with ribosomes across mRNA coding regions except when Upf1 association forms L footprints.

A-C. Distribution of footprint abundance across the coding region (CDS) for each footprint size class from total (purple) or IP (orange) ribosomes profiling libraries. Each transcript's CDS was divided into 4 bins and the percentage of footprints' P-sites belonging to each bin in each transcript was calculated and plotted. Grey dashed line signifies a theoretical number where each bin of CDS contains an equivalent number of footprints for a sum of 100% across all 4 bins. D-F. Difference between footprint abundance in IP and that in Total ribosomes for each footprint size and bin (calculated in part A-C).

Table 4.1. Comparison of M vs L footprints abundance 5' and 3' of the mRNA coding region (CDS) by Fisher’s exact test.

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>M</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FLAG-UPF1 (+CHX): p-value &lt; 2.2e-16, Odds ratio = 5.341332</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of mRNAs with higher abundance in the 5' half of CDS</td>
<td>2,644</td>
<td>987</td>
<td>3,631</td>
</tr>
<tr>
<td>Number of mRNAs with higher abundance in the 3' half of CDS</td>
<td>1,593</td>
<td>3,177</td>
<td>4,770</td>
</tr>
<tr>
<td>Total</td>
<td>4,237</td>
<td>4,164</td>
<td>8,401</td>
</tr>
<tr>
<td><strong>UPF1-FLAG (+CHX): p-value &lt; 2.2e-16, Odds ratio = 7.180561</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of mRNAs with higher abundance in the 5' half of CDS</td>
<td>3,048</td>
<td>948</td>
<td>3,996</td>
</tr>
<tr>
<td>Number of mRNAs with higher abundance in the 3' half of CDS</td>
<td>1,710</td>
<td>3,820</td>
<td>5,530</td>
</tr>
<tr>
<td>Total</td>
<td>4,758</td>
<td>4,768</td>
<td>9,526</td>
</tr>
<tr>
<td><strong>UPF1-FLAG/upf2Δ (+CHX): p-value &lt; 2.2e-16, Odds ratio = 7.6964</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of mRNAs with higher abundance in the 5' half of CDS</td>
<td>3,119</td>
<td>902</td>
<td>4,021</td>
</tr>
<tr>
<td>Number of mRNAs with higher abundance in the 3' half of CDS</td>
<td>1,754</td>
<td>3,905</td>
<td>5,659</td>
</tr>
<tr>
<td>Total</td>
<td>4,873</td>
<td>4,807</td>
<td>9,680</td>
</tr>
</tbody>
</table>

Analysis of mRNAs enriched in Upf1-associated ribosomes supports Upf1’s role in mRNA surveillance

To gain insight into the targets of Upf1 binding, we used DESeq2 to compare the distribution of reads in libraries prepared from IP’d and total ribosomes and identified mRNAs enriched, depleted, or unchanged in IP libraries. We found a larger number of mRNAs identified as having significant differences between IP’d
and total ribosomes from the C-terminally FLAG-tagged strains than those from N-terminally FLAG-tagged strains (Figure 4.13A). Because false discovery rate (FDR) in differential expression analyses takes into account replicate variability, the N-terminally FLAG-tagged data set, which had two instead of three biological replicates and smaller sequencing library size than the C-terminally FLAG-tagged data set, had higher variability between replicates, resulting in a smaller number of mRNAs with significant adjusted p-values compared to the C-terminally FLAG-tagged data set. Thus, we focused on the C-terminally FLAG-tagged data set for subsequent analyses.

In the C-terminally FLAG-tagged data set, we found that both NMD substrates and non-NMD substrates can be enriched, depleted, or unchanged in all IP’d samples (Figure 4.13A). These observations suggest that Upf1 can bind ribosomes translating any mRNA and that Upf1 binding to a ribosome by itself does not trigger NMD of the associated mRNA. However, the enrichment and depletion of some transcripts after IP suggests that specific mRNA features may enhance or reduce recruitment of Upf1 to ribosomes.

To characterize mRNAs enriched or depleted in the libraries from immunopurified ribosomes, we performed comparative analyses of ribosome occupancy, coding sequence length, and codon optimality between mRNAs enriched, depleted, or unchanged in IP vs total ribosomes from *UPF1-FLAG*, *UPF1-FLAG/upf2Δ*, and *upf1DE572AA-FLAG* strains (Figures 4.13B-D). Consistent with our earlier observations that NMD substrates have significantly
lower ribosome occupancy than non-NMD substrates (Celik et al. 2017a), we found that mRNAs enriched in Upf1-associated ribosomes for both NMD substrate and non-NMD substrate categories in all strains have significantly lower ribosome occupancy than those depleted in Upf1-associated ribosomes (Figure 4.13B) (Celik et al. 2017a; He et al. 2018). This observation suggests that Upf1:ribosome association occurs more frequently when an mRNA's translation dynamics are similar to those of NMD substrates, and less frequently when an mRNA's translation dynamics are opposite those of NMD substrates. However, the fact that this pattern is observed in both NMD and non-NMD substrates and in strains in which the NMD pathway is inactivated (UPF1-FLAG/upf2Δ and upf1DE572AA-FLAG) further suggests that Upf1 binding to ribosomes is likely to be a surveillance step prior to NMD commitment.

NMD activation is triggered by premature termination events, including out of frame translation ending at a premature stop codon, an event that may occur by chance more frequently in mRNAs with relatively long open reading frames. Therefore, we analyzed the coding sequence length for all mRNAs recovered in our ribosome profiling libraries. We found that mRNAs enriched in IP have significantly longer coding sequence lengths than mRNAs in both the unchanged and depleted groups and that mRNAs depleted in IP have significantly shorter coding sequence lengths than mRNAs in the other two groups (Figure 4.13C). However, there was no difference in these patterns between NMD and non-NMD substrates nor between strains with an active or inactive NMD pathway. As such,
it is unlikely that this observation is related to Upf1’s NMD function but rather reflects the steady ribosomal accumulation of Upf1 across the coding region observed in Figure 4.9, Figure 4.11, and Figure 4.12, and the increased opportunities for Upf1 association on longer coding sequences, a result that again supports Upf1:ribosome association prior to NMD commitment.

Previous studies (Presnyak et al. 2015; Radhakrishnan et al. 2016) showed that mRNAs targeted for decay tend to have a high number of non-optimal codons, with corresponding decoding tRNAs that are low in abundance leading to slow elongation. We also showed previously that NMD substrates have slightly lower codon optimality scores than non-NMD substrates (Celik et al. 2017a). Thus, we wondered whether this trend is also true for transcripts enriched in IP libraries. We calculated a codon optimality score for each transcript by determining the geometric mean of codon optimality scores of individual codons in the coding sequence (Reis et al. 2004; Tuller et al. 2010b). We found no difference in mean codon optimality scores among the three mRNA groups in any strains (Figure 4.13D), again supporting our hypothesis that Upf1 associates generally with ribosomes while playing a surveillance function prior to NMD commitment. It is possible, however, that Upf1 could be recruited to localized regions of non-optimal codons in an mRNA, and thus a single codon optimality score of the entire coding region cannot capture this phenomenon.
**Figure 4.13.** Characteristics of mRNAs enriched or depleted for Upf1-associated ribosomes.

**A.** Results of differential expression analyses by DEseq2 between mRNA abundance in IP vs Total ribosome profiling libraries of each strain and CHX treatment. False discovery rate (FDR) with threshold of 0.05 was used to determine significant differential expression. mRNAs with adjusted p-value < 0.05 and positive log₂(IP / Total) were considered enriched in IP libraries (orange) while those with adjusted p-value < 0.05 and negative log₂(IP / Total) were considered depleted in IP libraries (purple); otherwise, their abundance did not differ (“unchanged”) between IP and Total (grey).

**B-D.** Comparative analyses of ribosome occupancy (**B**), coding sequence length (**C**), and codon optimality score (**D**) between mRNA groups identified in **A**. Two-tailed Wilcoxon’s rank sum test with FDR method for multiple testing correction was used to compare each pair of mRNA groups. Symbols for levels of significance are ns: p > 0.05, *: p < 0.05, **: p < 0.01, ***: p < 0.001, ****: p < 0.0001. Number of mRNAs in each group is provided. Outlier mRNAs (those beyond the whiskers of box plots) were included in the analyses but omitted from plotting.

### Ribosomes associated with upf1DE572AA accumulate downstream of normal termination codons

To determine whether Upf1-associated ribosomes are found at specific codons, we calculated the mean relative A-site codon occupancy of immunopurified and total ribosomes for any given codon and its surrounding region (Figure 4.14A). Because ribosomal pauses are related to translation elongation dynamics and the abundance of tRNAs in the cell (Dana and Tuller 2014; Tuller et al. 2010a), and their relationships have been shown to be disrupted by CHX treatment (Hussmann et al. 2015; Sharma et al. 2021b), we tested whether CHX affects codon occupancy analyses in our libraries by computing Spearman’s rank correlation of A-site mean relative occupancy with the inverse of tRNA adaptation index (tAI) (Reis et al. 2004; Tuller et al. 2010b) for a given sense codon identity in each library, as described by Hussmann et al. (Hussmann et al. 2015) (Figure...
4.15). Consistent with results from the Hussmann et al. (Hussmann et al. 2015) analysis, we found positive correlations for both S and M footprints for libraries untreated with CHX (Figure 4.15, open circles), demonstrating that the expected relationship between translation dynamics and codon optimality is maintained. Moreover, we found negative correlations for both M and L footprints for libraries derived from cells treated with CHX (Figure 4.15, circles with plus sign), indicating that CHX disrupted the ability to accurately measure the impact of codon optimality on ribosomal enrichment or depletion at sense codons.

When we compared log$_2$ foldchange in mean relative occupancy of immunopurified vs. total ribosomes across libraries prepared from the three C-terminally FLAG-tagged strains, we found that their patterns of enrichment or depletion around stop codons vary between strains (Figure 4.14B). In strains expressing WT _UPF1-FLAG_, whether in _UPF2_ or _upf2Δ_ background, we found that mean relative occupancy of immunopurified ribosomes were depleted compared to total ribosomes in the 3'-UTR region downstream of all three in-frame stop codons (Figure 4.14B, purple highlighting). In contrast, we found that in libraries from strains expressing _upf1DE572AA-FLAG_, the mean relative occupancy of immunopurified ribosomes in the 3'-UTR region was increased compared to total ribosomes (Figure 4.14B, orange highlighting). These differences in the amount of 3'-UTR footprints between IP and total libraries were also evident when we calculated the overall percentage of footprints in the libraries that mapped to the 3'-UTR region (Figure 4.16A and Figure 4.17B). Unlike other strains containing WT
UPF1, where the 3'-UTR footprints were slightly reduced in IP compared to total libraries, the upf1DE527AA-FLAG strain IP libraries show an increase to almost double the amount of 3'-UTR footprints compared the total libraries, and this increase is statistically significant (Figure 4.16A and Figure 4.17B). The upf1DE572AA mutation has been observed to produce mRNP-protected decay intermediates downstream of premature stop codons in transcripts from reporter constructs (Franks et al. 2010; Serdar et al. 2016, 2020), and these decay intermediates were later found to be bound by a ribosome at their 5' terminus (Serdar et al. 2020). Our observation that footprints from Upf1-bound ribosomes are selectively enriched downstream of normal termination codons in the upf1DE572AA-FLAG strain is surprising in light of the generally accepted notion that Upf1 functions during premature termination (He and Jacobson 2015b; Kurosaki et al. 2019).

Footprints in the 3'-UTR of an mRNA can be caused by several events, such as nonsense codon readthrough, reinitiation, and frameshifting. Previously, ribosomes associated with 3' decay intermediates in the upf1DE572AA mutant strain were determined not to be engaged in canonical translation, neither readthrough nor reinitiation (Serdar et al. 2020). Therefore, their footprints would be expected to have a random reading frame and the percentage of in-frame 3'-UTR footprints in IP libraries prepared from the upf1DE572AA-FLAG strain should be reduced compared to total ribosome libraries or libraries prepared from strains harboring WT UPF1. Thus, we calculated percentages of the three reading frames
in the 3’-UTR region of all libraries (Figure 4.16B and Supplemental Figure S4.17D) and used values from the total ribosome libraries to establish baseline level of these events. We found that from total libraries of all C-terminally FLAG-tagged strains, reading frame 0 accounts for approximately 50-55% of the footprints in the 3’-UTR. In IP libraries of strains containing UPF1-FLAG, the percentage of frame 0 footprints is reduced slightly to around 45-49%. However, in IP libraries of upf1DE572AA-FLAG strain, the percentage of frame 0 drops to 37%, approximately equivalent to frames 1 and 2. This evidence suggests that these footprints, like the 3’ decay intermediates observed previously (Serdar et al. 2020), are not generated by ribosomes undergoing canonical translation but rather are a consequence of downstream ribosome migration and defective ribosome recycling attributable to ribosome-bound ATPase-deficient Upf1. Further, while we recognize that this unexpected detection of Upf1’s action at normal termination codons could be due to episomal expression of UPF1, the detection of these footprints only in IP libraries of upf1DE572AA-FLAG cells supports the biological significance of our findings.
**Figure 4.14.** Codon occupancy for individual codon identity in the A-site and its surrounding codon positions in C-terminally FLAG-tagged *UPF1* libraries. A. Mean relative occupancy values were calculated as described in Materials and Methods. Briefly, read count of major footprint size (M in this case) at each codon position relative to the codon of interest in the A-site was normalized by the average count in the 60-codon window. Thus, mean relative occupancy of 1 (white) indicates that A-site occupancy at that position is no different from the window average. Mean relative occupancy > 1 (red) and < 1 (blue) respectively indicate enrichment and depletion of ribosomal A-site at that position compared to window average. Only codons -10 to 10 were plotted. B. Differences in mean relative occupancy between IP and total libraries. Ratios of mean relative occupancy (from A) of IP to Total, log₂-transformed. Thus, log₂(IP/Total) of 0 (white) indicates no difference in mean relative occupancy between IP and Total at that codon position, while positive (orange) and negative (purple) log₂(IP/Total) indicate respectively higher and lower occupancy in IP compared to total at that codon position.

**Figure 4.15.** The relationship between codon optimality and ribosome occupancy at a codon is disrupted in the presence of cycloheximide. Spearman’s rank correlation was computed of mean relative occupancy at 61 sense codons with their associated inverse of tRNA adaptation index (1/tAI). Positive correlation indicates the expected translation dynamic where the ribosomes spend longer time translating non-optimal codons and less time translating optimal codons.
**Figure 4.16. Percentage of footprints in the 3'-UTR region.**

**A.** Percentage of footprints whose P-sites fall into the 3'-UTR region for each library. **B.** Percentage of frame 0 footprints within the 3'-UTR region. Horizontal grey dashed line indicates a theoretical percentage of 33% where all three reading frames would be equally represented. For both A and B, percentage calculated from each replicate was plotted as a grey dot. Bar plot represents the average among the three replicates. Two-tailed paired t-test was used to determine significant differences between percentages from IP and Total libraries.
Figure 4.17. Fractions of footprints in ribosome profiling libraries in particular mRNA regions or reading frames.  
A-B. Fractions of footprints in mRNA regions (5’-UTR, coding, 3’-UTR) in which ribosome footprint P-sites are located for each library. P-site location of each read was determined by testing whether its coordinate was located before the coordinate of the first nucleotide of the start codon (5’-UTR region), after the coordinate of the last nucleotide of the penultimate codon (3’-UTR region), or at or in between the two mentioned coordinates (coding region, CDS). The number of P-sites in each of the 3 regions was tabulated and normalized to the total number of footprints in the library.  
C-D. Fractions of ribosome footprints in each of the three reading frames within different mRNA regions (5’-UTR, CDS, 3’-UTR). The reading frame of each footprint was calculated by dividing its coordinate relative to the first nucleotide of the start codon then recording the remainder. Thus, reading frame “0” refers to translation that is in-frame with the start codon. The number of footprints in each reading frame in a particular mRNA region was tabulated and normalized to the total number of footprints in that mRNA region.
Discussion

**Upf1 functions while associated with ribosomes**

Given the uncertainties underlying the association of Upf1 with elongating ribosomes (see Introduction), we sought to elucidate the issue by combining *UPF1* genetics with selective ribosome profiling. While recognizing the caveat that our observations may be influenced by episomal expression of Upf1, the results of these experiments strongly support the notion that Upf1 functions while associated with ribosomes engaged in elongation. In support of this conclusion we have shown that: i) immunopurification of FLAG-tagged Upf1 from a sample of total cellular ribosomes yields stoichiometric recovery of Upf1 and ribosomal proteins that is specific for the FLAG tag regardless of its 5' or 3' location in the respective *UPF1* ORF or the status of NMD activity in the cells from which ribosomes were isolated (Figure 4.5); ii) Upf1-associated ribosomes purified from cells treated or not treated with CHX respectively yield the expected predominant 27-32 nt (“M”) or 20-23 nt (“S”) footprint size classes and these footprints manifest appropriate 3 nt periodicity characteristic of translating ribosomes (Figures 4.8 and 4.9); iii) Upf1 association with ribosomes in CHX-treated cells leads to formation of an additional 5’-extended 37-43 nt (“L”) footprint and it, too, manifests 3 nt periodicity (Figures 4.8, 4.9, and 4.10; see also next section); iv) S and M footprints from immunopurified ribosomes are relatively underrepresented in approximately the first half of normalized coding regions and overrepresented in the second half (Figures 4.11 and 4.12) whereas L footprints rapidly accumulate in normalized
ORFs and maintain their presence throughout mRNA coding regions, diminishing only slightly from the start of translation until its termination (Figures 4.11 and 4.12); and v) ribosomes associated with Upf1 that harbors the DE572AA amino acid substitutions terminate or recycle improperly at normal termination codons and are retained in mRNA 3'-UTR regions (Figures 4.14, 4.16 and 4.17). These results lead us to conclude that, at least in yeast, Upf1 is not displaced by elongating ribosomes, but is carried by them routinely during the course of translation elongation. Since NMD substrates and non-NMD substrates were found to be enriched, depleted, or unchanged in all immunopurified ribosome samples (Figure 4.13) it appears that stochastic binding of Upf1 to translating ribosomes is insufficient to trigger NMD of the associated mRNA and that the observed association may well be part of a translation surveillance mechanism in which Upf1 is sensing termination efficiency.

The Upf1:ribosome complex detected in CHX-treated cells comprises a step that precedes commitment to NMD

In CHX-treated cells, the footprints recovered from immunopurified Upf1-associated ribosomes include the expected (Wu et al. 2019) M footprints of ~27-32 nt as well as novel 37-43 nt ribosome-protected fragments that we designate as L footprints (Figure 4.8A). The latter footprints may be formed by a CHX-induced ribosome collision and subsequent endonucleolytic cleavage (Guydosh and Green 2017) or by the combined mRNA protection effects of Upf1 plus a ribosome to which it is bound (Figure 4.10). The latter additive effect would be consistent with
the size of ribosome footprints and the observation that human Upf1 protects 8-11 nt of RNA in RNA binding and unwinding assays (Chakrabarti et al. 2011). L footprints accumulate early during translation of mRNA ORFs, well before distal regions of mRNA are recovered in M or S footprints from Upf1-associated ribosomes (Figures 4.11 and 4.12). L footprints are not recovered from Upf1-associated ribosomes purified from cells without CHX treatment (Figure 4.8A), and their formation does not depend on specific placement of the FLAG epitope tag to one end of Upf1 (Figure 4.8). Formation of L footprints does depend on Upf1 function because they are absent when Upf1 harbors DE572AA substitutions, but they are not dependent on a functional NMD pathway or Upf2 activity because they are still recovered from Upf1-associated ribosomes in CHX-treated upf2Δ cells (Figure 4.8B). This combination of properties suggests that CHX treatment has trapped ribosomes and Upf1 in an otherwise transient state preceding a commitment to NMD, i.e., L footprints are Upf1-specific, not NMD-specific. The existence of such a transient state implies that Upf1 can interact with ribosomes in modes that either do or do not lead to NMD. The early appearance of L footprints (i.e., their recovery from 5’ regions of mRNA ORFs), and their persistence throughout most of the coding region (Figures 4.9 and 4.11), implies that this potential surveillance for functional targets (e.g., a ribosome undergoing premature termination) is likely to be active during most translation elongation events. In addition, the uniformly delayed recovery of S and M footprints (the majority of the footprints recovered) from Upf1-associated ribosomes (Figures 4.9 and 4.11), is
synchronous with the onset of decline in L footprint formation, suggesting distinct states of Upf1 interaction with ribosomes.

**ATPase-deficient Upf1 interferes with normal translation termination**

Our analyses of the mean relative codon occupancy of immunopurified vs. total ribosomes in libraries from the C-terminally FLAG-tagged strains (treated with CHX) indicated that the relative ribosomal enrichment or depletion around all three stop codons was not comparable for the three strains (Figure 4.14B). Libraries prepared from immunopurified ribosomes derived from cells expressing WT UPF1-FLAG, with or without functional Upf2, were quite similar and manifested a relative depletion of ribosomes in the 3’-UTR region downstream of all three in-frame stop codons (Figure 4.14B, purple highlighting). This result was consistent with ongoing translation termination and ribosome release at the normal ends of essentially all mRNA ORFs. In contrast, cells expressing upf1DE572AA-FLAG manifested a deficiency in normal termination and ribosome release, i.e., the accumulation of ribosomes downstream of normal termination codons without regard to reading frame (Figures 4.14B, 4.16A, 4.17B).

Upf1 was originally identified as a regulator of the stability of mRNAs undergoing premature translation termination (He et al. 1993; Leeds et al. 1991) and this role for Upf1 has been substantiated by three decades of additional studies (He and Jacobson 2015b; Kurosaki et al. 2019) that describe a cascade of Upf1-driven events commencing with premature termination and culminating with mRNA decapping or endonucleolytic cleavage and subsequent exonuclease
digestion (Colombo et al. 2017; Eberle et al. 2009; He and Jacobson 2001, 2015b; He et al. 2022; Huntzinger et al. 2008; Kurosaki et al. 2019; Loh et al. 2013; Muhlrad and Parker 1994; Nelson et al. 2018; Ottens et al. 2017). It was thus surprising that cells expressing upf1DE572AA exhibited an apparent defect in normal termination, particularly since a recent study of reporter mRNAs containing premature termination codons (PTCs) in upf1DE572AA yeast cells also accumulated unreleased ribosomes downstream of stop codons, but these stop codons were all PTCs (Serdar et al. 2020). Our results are consistent with the notion that ribosome-bound Upf1 normally monitors the status of translation termination events and exerts its activity only when those events are aberrant, i.e., when they occur prematurely or in premature context, but suggest that the DE572AA mutation may in some instances interfere with normal termination as well as premature termination. For example, the Upf1 DE572AA mutant protein may be able to target elongating ribosomes but cannot dissociate from them efficiently because of its disrupted ATPase cycle. This mutant Upf1 protein could thus persist on ribosomes, even at normal termination codons, and interfere with normal ribosome recycling, causing ribosome accumulation in mRNA 3'-UTRs.

**Upf1 functions in multiple complexes**

Finally, there remains the question of the specific function of ribosome-bound Upf1 in the NMD pathway. Our earlier results suggested that Upf1 may promote the release of an otherwise poorly dissociable premature termination complex (Ghosh et al. 2010). While the results of Figures 4.14, 4.16, and 4.17
support this possibility, our recent study on the role of decapping activators in the targeting and activation of the decapping enzyme suggests an additional function (He et al. 2022). In that study, we showed that specific cis-binding elements in the Dcp2 C-terminal domain control the substrate specificity of the decapping enzyme by orchestrating the formation of target-specific decapping complexes. The latter include a Upf1-containing decapping complex that targets NMD substrates (He et al. 2022). Since deletion of UPF1 promotes extensive stabilization of NMD substrates but Upf1-mediated recruitment of the decapping enzyme only makes a minor contribution to the overall decay of NMD substrates (He et al. 2022), it is likely that Upf1 carries out a major function upstream of decapping enzyme recruitment. In light of the fact that Upf1 has two nearly identical binding sites in the unstructured Dcp2 C-terminal domain (He et al. 2022), and that Upf1 is known to dimerize (He et al. 2013), it is possible that ribosome-associated Upf1 remodels an mRNP to render it susceptible to decapping and then facilitates decapping by dimerizing with a monomer of Upf1 present in an NMD specific decapping complex. Consistent with this proposition, Upf1 has been found in two distinct complexes, a “detector” complex containing Upf1, Upf2, and Upf3, and an “effector” complex containing Dcp1, Dcp2, and Edc3, as well as Nmd4 and Ebs1 (Dehecq et al. 2018).
Materials and Methods

Yeast Strains

Yeast strains used in this study are listed in Table 4.2. Strains containing complete gene deletions of UPF1 and UPF2 (HFY871, HFY861, HFY467) were described previously (He et al. 1997; He and Jacobson 1995).

Table 4.2. Yeast strains used in this study.

<table>
<thead>
<tr>
<th>Name</th>
<th>Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>HFY114</td>
<td>\textit{MAT} a ade2-1 his3-11,15 leu2-3,112 trp1-1 ura3-1 can1-100 UPF1 NMD2 UPF3</td>
</tr>
<tr>
<td>HFY871</td>
<td>\textit{MAT} a ade2-1 his3-11,15 leu2-3,112 trp1-1 ura3-1 can1-100 upf1::HIS3 NMD2 UPF3</td>
</tr>
<tr>
<td>HFY467</td>
<td>\textit{MAT} a ade2-1, his3-11,15, leu2-3,112, trp1-1, ura3-1, can1-100, nmd2::HIS3, upf1::URA3</td>
</tr>
</tbody>
</table>

Oligonucleotides

Oligonucleotides used in this study were obtained from Eurofins Operon or Integrated DNA Technologies (IDT) and are listed in Table 4.3. Gene fragments were synthesized by Quintara Biosciences and are described below in Plasmids.

Table 4.3. Oligonucleotides used in this study.

<table>
<thead>
<tr>
<th>Oligo</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-N-His-UPF1</td>
<td>TTGTCATTATCAAATGGCCATTACATCATCATCATGTGCTTCCGTTCTCACACTC</td>
</tr>
<tr>
<td>5-N-HIS-UPF1-r</td>
<td>GAGTGTGAAGAAACGGAACCATGATGATGATGATGGTGTACCATTGATAATGACAA</td>
</tr>
<tr>
<td>5Upf1C62Ymut</td>
<td>GCCTTCAGCTTCCCAGACAAATTCATATGCCTATGTTGGG</td>
</tr>
<tr>
<td>3Upf1C62Ymut</td>
<td>CCACAATACGCAATGAAATTTGTCGAACTGAGGAGG</td>
</tr>
<tr>
<td>K436F</td>
<td>CCC ACCAGGGCACTTGTAGAAGCATGGTTTACTTACGAG A</td>
</tr>
<tr>
<td>K436R</td>
<td>TGCTGAAGTACTTCTACCTACATGCGTTGGG</td>
</tr>
<tr>
<td>Upf1AKS-HPAmut</td>
<td>GCCCTCAGCTTCAACATTTCATATGCCTATGTTGGG</td>
</tr>
<tr>
<td>UPF1AKS-HPAmutrev</td>
<td>CGTATTATGAAGAGAGAGGTAAGTGGGAGGGTGCTCTTCCGCTCTGGTGGTCT</td>
</tr>
<tr>
<td>5Upf1RRAAmut</td>
<td>AGACCGTAGTCTAGACAGGAGGCCAGAGTGGTGTGGTGGCTC</td>
</tr>
<tr>
<td>3Upf1RRAAmut</td>
<td>GGCCATTGTGTGTTTCACTGAGTGGCTCTTCTAAGCAGGAGTTGGT</td>
</tr>
<tr>
<td>Upf1R779C</td>
<td>TGGTTGCGTGGGCACAAAACACAGATAAGATTATGTAATCCTTCTT</td>
</tr>
<tr>
<td>Upf1R779Crev</td>
<td>GAAAAGGATTACATATCTTATCGTGTTTGGTCCATGAAACAAA</td>
</tr>
<tr>
<td>Plasmid</td>
<td>Description</td>
</tr>
<tr>
<td>------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>YEplac112</td>
<td>Yeast episomal plasmid</td>
</tr>
<tr>
<td>pG1-FLAG-UPF1</td>
<td>Yeast episomal plasmid containing entire UPF1 coding region as a 3.6kb BamHI fragment; FLAG-tag fused at the N-terminus</td>
</tr>
<tr>
<td>pRS425</td>
<td>Yeast episomal plasmid</td>
</tr>
<tr>
<td>pRS316</td>
<td>Yeast centromeric plasmid</td>
</tr>
<tr>
<td>YEplac195-TPI-UPF3</td>
<td>Yeast episomal plasmid containing entire UPF3 coding region; expressed under the TPI promoter</td>
</tr>
<tr>
<td>YEplac112-6XHis-UPF1</td>
<td>Yeast episomal plasmid containing entire UPF1 coding region; 6XHis-tag fused at the N-terminus</td>
</tr>
<tr>
<td>pRS425-UPF1-FLAG</td>
<td>Yeast episomal plasmid containing entire UPF1 coding region; FLAG-tag fused at the C-terminus; expressed under the TDH3 promoter</td>
</tr>
<tr>
<td>pRS425-C62Y-FLAG</td>
<td>Same as pRS425-UPF1-FLAG, containing C62Y mutation</td>
</tr>
<tr>
<td>pRS425-K436E-FLAG</td>
<td>Same as pRS425-UPF1-FLAG, containing K436E mutation</td>
</tr>
<tr>
<td>pRS425-DE572AA-FLAG</td>
<td>Same as pRS425-UPF1-FLAG, containing DE572AA mutation</td>
</tr>
<tr>
<td>pRS425-AKS484HPA-FLAG</td>
<td>Same as pRS425-UPF1-FLAG, containing AKS484HPA mutation</td>
</tr>
<tr>
<td>pRS425-R779C-FLAG</td>
<td>Same as pRS425-UPF1-FLAG, containing R779C mutation</td>
</tr>
<tr>
<td>pRS425-RR793AA-FLAG</td>
<td>Same as pRS425-UPF1-FLAG, containing RR793AA mutation</td>
</tr>
<tr>
<td>pRS316-UPF1-FLAG</td>
<td>Yeast centromeric plasmid containing entire UPF1 coding region; FLAG-tag fused at the C-terminus; expressed under the TDH3 promoter</td>
</tr>
<tr>
<td>pRS316-C62Y-FLAG</td>
<td>Same as pRS316-UPF1-FLAG, containing C62Y mutation</td>
</tr>
<tr>
<td>pRS316-K436E-FLAG</td>
<td>Same as pRS316-UPF1-FLAG, containing K436E mutation</td>
</tr>
<tr>
<td>pRS316-DE572AA-FLAG</td>
<td>Same as pRS316-UPF1-FLAG, containing DE572AA mutation</td>
</tr>
</tbody>
</table>
The following plasmids were published previously: YEplac112 (Gietz and Akio 1988); pG1-FLAG-UPF1 (Czaplinski et al. 1995); pRS425 and pRS316 (Sikorski and Hieter 1989).

YEplac112-6Xhis-UPF1: pRS314-UPF1 (He and Jacobson 1995) was digested with BamHI, Sall, and Xbal to release a 4.2kb BamHI-Sall fragment of UPF1. This fragment was ligated into pGEM-3Zf (+) (Promega) that had been digested with BamHI and Sall and dephosphorylated with calf intestinal alkaline phosphatase (NEB). 6Xhis tag was added to the N-terminus of Upf1 in pGEM-3Zf (+) using QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent) with oligonucleotides 5-N-His-UPF1 and 5-N-HIS-UPF1-r to yield pGEM3Zf(+)-6Xhis-UPF1. A ~4.2kb Sacl/PstI fragment was isolated from pGEM3Zf(+)-6Xhis-UPF1, ligated into Yeplac112 which had been digested with Sacl and PstI and dephosphorylated with calf intestinal alkaline phosphatase to yield Yeplac112-6Xhis-UPF1.

pRS425-UPF1-FLAG: gene fragments were synthesized by Quintarabio with flanking restriction sites 1) UPF1-Xho-Sph, a 1593 bp fragment containing the TDH3 (glyceraldehyde-3-phosphate dehydrogenase, GAPDH) promoter inserted
into a BamHI site plus ggc (ggatccggc) immediately upstream of the start codon of the UPF1 coding region, then from the start codon to the SphI site at position 914 of the UPF1 coding sequence; 2) UPF1-Sph-Nco, a 1102 bp fragment from the SphI site in the UPF1 coding sequence to the Ncol site in the UPF1 coding sequence; 3) UPF1-Nco-Sac, a 1167 bp fragment from the Ncol site in the UPF1 coding sequence to a SacI site in the UPF1 3’UTR followed by a spacer with a SacI restriction site (caccgcggtggagctc), and FLAG epitope sequence (gattacaaggatgacgacgataag) inserted immediately upstream of the stop codon. These three gene fragments were ligated into pRS425 at the XhoI-SacI sites of the MCS.

Mutations of UPF1 were placed into UPF1-FLAG as follows:

C62Y: The UPF1-Sph-Nco and UPF1-Nco-Sac gene fragments described above were cloned in to the SphI-SacI sites in the pGEMT-EZ (Promega) MCS to generate pGEMT-EZ-Sph-UPF1-FLAG-Sac. The UPF1-Xho-Sph gene fragment in plasmid vector pQ (Quintarabio) was mutagenized with oligonucleotides 5Upf1C62Ymut and 3Upf1C62Ymut using the QuikChange II XL site directed mutagenesis kit (Agilent). The UPF1Xho-C62Y-SphI fragment was isolated and ligated with the 2.28 kb Sph-UPF1-FLAG-Sac fragment from pGEMT-EZ-Sph-UPF1-FLAG-Sac into the XhoI-SacI site of the MCS of pRS425.

K436E: pGEM3Z-6Xhis-UPF1 was mutagenized with oligonucleotides K436F and K436R using the QuikChange II XL site directed mutagenesis kit, the 1116 bp SphI-Ncol fragment containing the K436E mutation was isolated and substituted for the
UPF1-Sph-Nco fragment as described above for the construction of pRS425-UPF1-FLAG.

**AKS484HPA, RR793AA and R779C:** pGEMT-EZ-Sph-UPF1-FLAG-Sac was mutagenized with oligonucleotides Upf1AKS-HPAmut and UPF1AKS-HPAmutrev, 5Upf1RRAAmut and 3Upf1RRAAmut, or Upf1R779C and Upf1R779Crev, respectively, using the QuikChange II XL site directed mutagenesis kit. The 2.28 kb Sphl-Sacl fragments were isolated and were ligated with the UPF1-Xho-Sph gene fragment into the Xhol-Sacl site of the pRS425 MCS.

**DE572AA:** the 1116bp Sphl-Ncol fragment from pACTII-UPF1-TH4-3-2-DE572AA (He et al., 2013) was isolated and substituted for the UPF1-Sph-Nco fragment described above.

**UPF1-FLAG** and mutants were placed under the **UPF1** promoter into pRS316 as follows: a 340 bp fragment containing the **UPF1** promoter sequence flanked by Xhol and BamHl restriction sites was generated by PCR from pRS313-UPF1 (E-B) with oligonucleotides 5Upf1XhoProm and 3Upf1BamProm and digested with Xhol and BamHl. A 3.195 kb BamHl-Sacl fragment from pRS425-UPF1-FLAG or mutants was isolated and ligated with the digested PCR product above into the Xhol-Sacl site of the pRS316 MCS.

pRS425-UPF2: a 4946 bp fragment generated by PCR with primers Upf2-425-fwd and Upf2-425-rev containing a the entire UPF2 gene was inserted in to pRS425 cut with NotI and Sall using the NEBuilder High-Fidelity DNA Assembly Cloning Kit (New England Biolabs).
yEplac195-TPI-UPF3: A 1164 bp SalI-XbaI fragment was generated by PCR from genomic DNA with primers 5’Sal-UPF3 and 3’Xba-UPF3 and restriction digested with Sall and Xbal. yEplac195-TPI-FLAG-UPF3 was digested with Sall and XbaI and the vector backbone was ligated to the digested PCR product to generate yEplac195-TPI-UPF3.

**Polyribosome analysis, protein detection, and quantitation**

For polyribosome analysis, cells were grown in selective media and harvested as described previously (Mangus and Jacobson 1999). Lysates were loaded onto 7-47% sucrose gradients, subjected to ultracentrifugation, fractions collected, and individual fractions were precipitated with trichloroacetic acid (TCA). Samples from total or immunopurified ribosomes or aliquots from TCA precipitated sucrose gradient fractions (fractions 1-10, undiluted; fractions 11 and 12, diluted 1:10) were run in 1X Laemmlili buffer on 4-20% Mini-PROTEAN TGX Precast Protein Gels (BioRad) in 1X Tris/Glycine/SDS buffer (BioRad), subjected to western blotting onto Immobilon™-P PVDF Transfer Membranes (EMD Millipore) by electrotransfer using a Trans-Blot® SD Semi-Dry Transfer Cell (Biorad). Protein was visualized with rabbit α-FLAG polyclonal antibody (Sigma) or rabbit monoclonal α-Rps6 antibody (Cell Signaling) followed by HRP-conjugated donkey α-rabbit (GE) antibody. Polysomal distribution of tagged WT and mutant Upf1 protein and tagged protein signal from total or immunopurified ribosomes was visualized by western blotting on Amersham Hyperfilm ECL film (GE). Band quantitation from scanned images was performed using Multigauge V3.0 (Fuji).
For quantitative analysis of the polysome distribution of tagged proteins, the antibody signal for a given fraction was calculated as the percent of its total signal across fractions 1-10 of an individual sucrose gradient western blot. Values from western blots of replicate gradient fractions were averaged and a bar graph generated to display the mean percent signal across sucrose gradient fractions 1-10 from all replicates of that sample with standard error of the mean displayed.

**Ribosome purification**

Yeast cultures were grown, whole cell yeast lysates were prepared and digested with rNaseI, and ribosomes were recovered as described previously (Ganesan et al. 2019).

**Negative-stain transmission electron microscopy**

Ribosomes were prepared for electron microscopy using the conventional negative staining procedure. Formvar-carbon-coated copper grids were treated by glow discharge using a PELCO easiGlow glow discharge cleaning system for 30 seconds at 20 mA. The ribosomes were diluted in footprinting buffer (20mM Tris, pH7.4, 150mM NaCl, 5mM MgCl$_2$) and negatively stained with 1% (w/v) aqueous uranyl acetate. All images were taken with an FEI Tecnai Spirit 12 (FEI Company, Hillsboro, OR), operated at 120 kV. Micrographs were recorded with a Gatan Rio 9 CMOS camera.

**Immunopurification**

FLAG-tagged protein-associated ribosomes were isolated by immunopurification using anti-FLAG M2 Affinity gel as described previously.
(Ganesan et al. 2019). Eluate was collected by centrifugation and spun through Amicon Ultra 2ml 100K centrifugal filter units (Millipore) until the volume was <200μl. The concentrate from one reaction set was pooled into one Amicon Ultra 2ml 100K centrifugal filter unit and spun until the total volume was approximately 160μl. For the \textit{UPF1-FLAG}, \textit{upf1DE572AA-FLAG} and \textit{UPF1-FLAG/upf2Δ} ribosome profiling libraries, immunopurified ribosomes were flash frozen and stored at -80°C until RNA purification. For \textit{FLAG-UPF1} libraries and for all mass spectrometry analyses, the immunopurified concentrate was layered on top of a 1M sucrose cushion in IP buffer plus 1X protease inhibitors, 10U/ml Superase-In, and spun at 166,180g for 100 minutes at 4°C in a TLA100 rotor. Supernatant was removed and the pellet was resuspended in 60μl IP buffer plus 0.5mM DTT, 1X protease inhibitors, 20U/ml Superase-In; and \textit{A}_{260} was determined on a spectrophotometer.

**Mass spectrometry**

Mass spectrometry analyses were performed by the Mass Spectrometry Facility of the UMass Chan Medical School.

**Sample preparation**

Samples were prepared for mass spectrometry by disrupting total or immunopurified ribosomes in 1X Laemmli buffer and running on a 4-20% Mini-PROTEAN TGX Precast Protein Gel in 1X Tris/Glycine/SDS buffer (BioRad) for 5 minutes. Gels were stained using the NOVEX Colloidal Blue Staining Kit (ThermoFisher Scientific) and destained in water. Fragments were excised and cut
into 1x1 mm pieces, transferred to microfuge tubes, and each added 1 mL of water, followed by a solution of 20 µL of 45 mM of 1,4 dithiothreitol (DTT) in 200 µL of 250 mM ammonium bicarbonate. Samples were incubated at 50°C for 30 minutes, cooled to room temperature, added 20 µL of 100 mM iodoacetamide (IAA) and incubated for 30 min. Excessive DTT and IAA was removed, and the gel pieces washed with water (3x, 1 mL each), followed by 1mL of a 1:1 solution of 50 mM ammonium bicarbonate:acetonitrile, quenched with 200 µL of acetonitrile and dried in a Speed Vac. Gel pieces were rehydrated in a mixture of 4 ng/µL trypsin (Promega, Madison, WI) and 0.01% ProteaseMAX (Promega) in 50 µL of 50 mM ammonium bicarbonate and incubated for 18 hours at 37°C. Supernatants were collected and further extraction was performed by adding 200 µL of 80:20 solution of acetonitrile:1% (v/v) formic acid in water. Supernatants were combined, peptides were lyophilized in a Speed Vac and resuspended in 25 µL of 5% acetonitrile with 0.1% (v/v) formic acid for mass spectrometry analysis.

Data acquisition

Data was acquired using a NanoAcquity UPLC (Waters Corporation, Milford, MA) coupled to a Q Exactive hybrid mass spectrometer (ThermoFisher Scientific) (N-terminal FLAG-tagged datasets) or an Orbitrap Fusion Lumos Tribrid (ThermoFisher Scientific) mass spectrometer (C-terminal FLAG-tagged datasets). Peptides were trapped and separated using an in-house 100 µm I.D. fused-silica precolumn (Kasil frit) packed with 2 cm ProntoSil (Bischoff Chromatography, DE) C18AQ (200Å, 5µm) media and configured to an in-house packed 75 µm I.D.
fused-silica analytical column (gravity-pulled tip) packed with 25 cm Magic (Bruker, Billerica, MA) C18AQ (100Å, 3µm) media, respectively. Mobile phase A was 0.1 % (v/v) formic acid in water and mobile phase B was 0.1 % (v/v) formic acid in acetonitrile. Following a 3.8 µL sample injection, peptides were trapped at flow rate of 4 µL/min with 5% B for 4 min, followed by gradient elution at a flow rate of 300 nL/min from 5-35% B in 60 min, wash and reconditioning (total run time ~90 min). Electrospray voltage was delivered by liquid junction electrode (1.4 kV) located between the columns and the transfer capillary to the mass spectrometer was maintained at 275°C. Mass spectra were acquired over m/z 300-1750 Da with a resolution of 60,000 (m/z 200), maximum injection time of 30 ms, and an AGC target of 700,000. Tandem mass spectra were acquired using data-dependent acquisition (2 sec cycle) with an isolation width of 1.2 Da, HCD collision energy of 30%, resolution of 15,000, maximum injection time of 100 ms, and an AGC target of 10,000. Biological triplicates of each strain (input and IP) were analyzed in technical triplicate.

**Database searches**

Raw data were processed using Proteome Discoverer (ThermoFisher Scientific, version 2.1.1.21) and searched against Uniprot Yeast (downloaded 07/2021) database using Mascot (Matrix Science, version 2.6.2). Search parameters were as follows: full tryptic specificity with up to 2 missed cleavages; precursor mass tolerance 10 ppm; fragment mass tolerance 0.05 Da; cysteine carbamidomethylation considered as a fixed modification, while protein N-terminal
acetylation, methionine oxidation, peptide N-terminal (E,Q) pyroglutamate conversion, serine/threonine phosphorylation and lysine ubiquitination (GG) were specified as variable modifications. Peptide and protein validation and annotation was done in Scaffold 4.8.9 (Proteome Software, Portland, OR) employing Peptide Prophet (Keller et al. 2002) and Protein Prophet (Nesvizhskii et al. 2003) algorithms. Peptides were filtered at a 1% FDR, while protein identification threshold was set to greater than 99% probability and with a minimum of 2 identified peptides per protein. Intensity-based absolute quantification (iBAQ) (Schwanhäusser et al. 2011, 2013), calculated and normalized across biosamples in Scaffold, is used as a measure of protein abundance.

**Protein Abundance Analysis**

Differential abundance analysis of protein levels between IP and Total ribosomes were performed in R environment using data exported from Scaffold. For proteins with identification probability < 99% and Total Spectrum Count < 2, their iBAQs were set to zero. For each protein in a sample, at least two replicates were required to have positive iBAQ values for further differential abundance analysis; otherwise, the identification of the protein was deemed inconsistent. iBAQs of biological replicates were averaged and log\(_{10}\)-transformed for plotting Figure 4.5A. Proteins that were exclusively identified in Total or IP samples (i.e., those with missing values in one of them) were plotted on the axes. For proteins that were identified in both Total and IP samples, their difference in abundance were further analyzed using an R package limma (Kammers et al. 2015; Smyth
First, iBAQs in each replicate samples were log_{10}-transformed. Missing iBAQ values within a sample were empirically imputed from the left-shifted Gaussian distribution of iBAQs of that replicate sample. The model was fit using limma, and the resulting log_2 fold change between IP and Total protein abundance with the associated p-value adjusted by the Benjamini-Hochberg method were used to plot Figure 4.5B.

**mRNA Decay Analysis**

mRNA decay phenotypes were assessed by northern blotting of total RNA prepared from cell pellets by the hot phenol method (Herrick et al. 1990), probing for CYH2 mRNA and pre-mRNA as described previously (He and Jacobson 1995; He et al. 1993; Herrick et al. 1990). Blots were visualized on a Fuji phosphorimager and quantitated with Multigauge V3.0 (Fuji). Northern blot images in Figure 4.3 were subjected to a Multigauge noise reduction filter, kernel size 3X3, applied equally across the entire image for display purposes only.

**RNA-Seq and Ribosome Profiling Library Preparation from Cells Expressing N-terminal FLAG-tagged Upf1 Protein**

RNA was isolated from clarified lysates (60 μl), and total (15 μl) or immunopurified ribosomes (entire recovered volume) using the miRNeasy mini kit (Qiagen) according to manufacturer’s standard protocol. RNA was eluted in 30 μl (lysate) or 14 μl (ribosomes), then treated with RiboZero Magnetic Gold (Yeast) Kit (Illumina) as described previously (Ganesan et al. 2019). Ribosome protected RNA fragments were 3’ dephosphorylated with T4 polynucleotide kinase (NEB) in 150
mM MES-NaOH pH 5.5, 450 mM NaCl, 15 mM MgCl₂, 15 mM β-mercaptoethanol, 0.3U/μl Superase-In at 37°C, 2 hours; 65°C, 20min; and RNA was selectively recovered with RNA clean and concentrator-5 (Zymo Research). RNA fragments were then 5' phosphorylated with T4 polynucleotide kinase in T4 PNK buffer, 1mM ATP, 1U/μl Superase-In at 37°C, 1 hour; 60°C, 10 min; and RNA was selectively recovered with RNA Clean and Concentrator-5 (Zymo Research) according to the manufacturer's protocol. Ribosome profiling libraries were prepared using the NEXTflex Small RNA-Seq Kit v3 (Perkin Elmer).

Total mRNA libraries were prepared using the TruSeq Stranded mRNA LT Sample Prep Kit (Illumina) according to manufacturer’s protocol.

Over the course of library preparation, the amounts of RNA and final libraries were quantified by Qubit 3.0 Fluorometer with the Qubit RNA HS Assay Kit (ThermoFisher Scientific) and the Qubit dsDNA HS Assay Kit (ThermoFisher Scientific), respectively. Assessments of rRNA depletion, RNA quality, and final libraries were done by Fragment Analyzer capillary electrophoresis system (Advanced Analytical) at the UMass Chan Medical School Molecular Biology Core Labs (RRID: SCR_018263).

RNA-Seq and Ribosome Profiling Library Preparation from Cells Expressing C-terminal FLAG-tagged Upf1 Protein

Because the RiboZero Magnetic Gold (Yeast) Kit (Illumina) was discontinued midway through this study, we developed an alternative rRNA depletion method based on oligonucleotide blocking of adapter ligation and cDNA
synthesis reactions. To maximize the amount of ribosome protected mRNA fragments input into the libraries, we altered immunopurification and RNA isolation protocols for the C-terminally FLAG tagged ribosome profiling libraries as follows by: 1) eliminating the final pelleting through a sucrose cushion following immunopurification, to avoid any loss of immunopurified ribosomes, and 2) using the small (<200nt) RNA preparation workflow in the miRNeasy kit according to manufacturer’s instructions, to remove as much extraneous RNA (either rRNA or large mRNA fragments) as possible prior to ribosome profiling library preparation. Following RNA isolation, 10 μl RNA from total or immunopurified ribosomes were 3’ dephosphorylated and 5’ phosphorylated as described above (Ganesan et al. 2019) and was selectively recovered with RNA Clean and Concentrator-5 (Zymo Research) according to the manufacturer’s instructions, using adjusted RNA Binding Buffer diluted 1:2 with ethanol, into 13 μl water. 8.5 μl RNA was incubated in a thermocycler with a preheated lid with 2 μl oligonucleotides (for RNA from total ribosomes) or 1 μl 1:10 diluted oligonucleotides (for RNA from immunopurified ribosomes) from the QIAseq FastSelect –rRNA Yeast Kit (Qiagen) in 10.5 μl total at 75°C, 2 min; 70°C, 2 min; 65°C, 2 min; 60°C, 2 min; 55°C, 2 min; 37°C, 2 min; 25°C, 2 min; 4°C, hold. Ribosome profiling libraries were prepared using the NEXTflex Small RNA-Seq Kit v3 (Perkin Elmer) but eliminating the 70 °C denaturation step prior to 3’ 4N Adenylated Adapter ligation. Adapters were undiluted for RNA from total ribosomes and 1:4 diluted for RNA from immunopurified ribosomes. PCR cycles were performed until a library appeared
by Fragment Analyzer analysis, 15 cycles for libraries from total ribosomes or 15-18 cycles for libraries from immunopurified ribosomes, and libraries were selectively recovered using the gel-free size selection and cleanup protocol according to manufacturer’s protocol.

For RNA-Seq library preparation, RNAs from clarified lysates (60 μl) were isolated using the miRNeasy mini kit (Qiagen) according to manufacturer’s standard protocol. The use of TruSeq Stranded mRNA LT Sample Prep Kit (Illumina) was altered as follows: 1μg of RNA, 14.5 μl of Fragment, Prime, Finish mix (FPF), 1μl of oligonucleotides from the QIAseq FastSelect –rRNA Yeast Kit (Qiagen) in a total of 20.5 μl were incubated in a thermocycler with a preheated lid at 94°C, 8 min; 75°C, 2 min; 70°C, 2 min; 65°C, 2 min; 60°C, 2 min; 55°C, 2 min; 37°C, 2 min; 25°C, 2 min; 4°C, hold; followed by first strand cDNA synthesis and the remainder of the protocol according to manufacturer’s instructions, except RNA Adapters were diluted 1:2 at the ligation step. If excess adapter dimers were present, it was necessary to reconstruct the library and dilute the RNA Adapter 1:4 or 1:8 and increase the PCR cycle number to 17, as was done for libraries T5019, T5033, T5035 and T5338.

**High-throughput Sequencing of RNA-Seq and Ribosome Profiling Libraries**

RNA-Seq libraries were sequenced on either a HiSeq4000 (Illumina) at Beijing Genomics Institute (BGI) or a NextSeq500 (Illumina) in-house. Ribosome profiling libraries were sequenced on a NextSeq500 in-house.
Sequence Alignment, Transcript Quantification, and Data Analyses.

RNA-Seq and ribosome profiling reads were aligned to a yeast transcriptome (available at https://github.com/Jacobson-Lab/yeast_transcriptome_v5) using bowtie (Langmead et al. 2009) and transcript abundance were determined using RSEM (Li and Dewey 2011), as described previously (Mangkalaphiban et al. 2021). For intron-containing genes, spliced (“mRNA”) and unspliced (“pre-mRNA”) isoforms were indexed as separate entries. Ribosome profiling libraries were pre-processed with adapter trimming and removal of reads aligned to non-protein coding RNAs. Because some of the constructs were tagged at their N-termini, a fraction of reads recovered were from ribosomes associated with the Upf factor nascent peptide. Therefore, reads arising from UPF factor transcripts were discarded from these and all subsequent ribosome profiling libraries analyzed. PCR duplicates identified via the 4N barcodes on either side of the read were removed (Mangkalaphiban et al. 2021). The remaining unique reads were further processed in two ways:

1) Transcript abundance was determined by RSEM either with all reads or a subset of reads of desired read lengths using a transcriptome containing pre-mRNA entries (rsem-calculate-expression --strandedness forward --fragment-length-mean [vary] --fragment-length-sd [vary] --seed-length 15 --bowtie-m 10). Reproducibility of biological replicate libraries is shown as correlation matrix (Figure 4.18A) and PCA plot (Figure 4.18B).
2) Reads were aligned to a separate transcriptome where only spliced mRNAs were considered (bowtie -m 10 -n 2 -l 15). The resulting aligned reads were processed by R package riboWaltz (Lauria et al. 2018) for initial visual inspection and calculation of P-site offsets, which were manually checked and modified for accuracy and used as the basis for subsequent reading frame calculations, periodicity, and other metagene plots. Analysis of ribosome footprints from both CHX treated and untreated cells and from total and Upf factor associated ribosomes showed the footprints mapped primarily to coding regions, most P-sites mapped to the “0” reading frame in coding regions, and displayed 3-nucleotide periodicity, indicative of translating ribosomes (Figures 4.9 and 4.17). As expected, CHX addition resulted in accumulation of ~27-32 nt footprints over the start codon, as described previously (Gerashchenko and Gladyshev 2014; Ingolia 2010; Lareau et al. 2014). Replicates were averaged unless otherwise specified.

RNA-Seq libraries were aligned to the transcriptome containing pre-mRNA entries and transcript abundance determined by RSEM (rsem-calculate-expression --strandedness reverse --fragment-length-mean 200 --fragment-length-sd 50 --bowtie-m 10) without any pre-processing steps. Reproducibility of biological replicate RNA-Seq libraries is shown as a correlation matrix (Figure 4.19A) and PCA plot (Figure 4.19B).
### Table

<table>
<thead>
<tr>
<th>Strain</th>
<th>Library</th>
<th>PCA: 11.9% variance</th>
<th>PCA: 43.9% variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT + EV (N)</td>
<td>Total -CHX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAG-UPF1</td>
<td>Total +CHX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAG-UPF1 -UPF2/3 EE</td>
<td>Total IP -CHX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WT + EV (C)</td>
<td>Total IP +CHX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UPF1-FLAG</td>
<td>Total IP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UPF1-FLAG/Upf2</td>
<td>IP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ΔDE572AA-FLAG</td>
<td>IP</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Diagram

#### A

The diagram shows a heatmap with Pearson's r values ranging from 0.4 to 1.0. The x-axis represents different libraries, and the y-axis represents different strains. The legend indicates the presence of -CHX and +CHX conditions.

#### B

The scatter plot displays PCA analysis with 11.9% and 43.9% variance explained. The markers represent different libraries and strains with specific conditions indicated in the legend.
Figure 4.18. Reproducibility between replicates of ribosome profiling libraries.
A. Correlation matrices showing pairwise Pearson’s correlation (r) of transcript abundance (log_{10}-transformed, non-zero RPKM values) between total and IP ribosome profiling libraries. B. Principal Component Analyses (PCA) on the main footprint abundance (S and M for -CHX libraries; M for +CHX libraries) show abundance variation across multiple libraries. Libraries with smaller distance to one another are more similar than ones that are further apart. Biological replicates tend to cluster together. Variation in transcript abundance between samples are mostly explained by whether Upf1-associated ribosomes were selectively recovered, as IP and Total libraries are segregated along the x-axis.
### Table

<table>
<thead>
<tr>
<th>Strain</th>
<th>Library</th>
<th>PCA (77% variance)</th>
<th>PCA (8.1% variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT + EV (N)</td>
<td>-CHX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAG-UPF1</td>
<td>+CHX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAG-UPF1 UPF2/3 EE</td>
<td>-CHX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WT + EV (C)</td>
<td>+CHX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UPF1-FLAG</td>
<td>-CHX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UPF1-FLAG/UPF2 Δ</td>
<td>+CHX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE572AA-FLAG</td>
<td>+CHX</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Diagram

- **A**: Heat map showing Pearson's r values.
- **B**: PCA plot showing variance.
**Figure 4.19. Reproducibility between replicates of RNA-Seq libraries.**

A. Correlation matrix showing pairwise Pearson’s correlation ($r$) of transcript abundance (log$_{10}$-transformed, non-zero RPKM values) between RNA-Seq libraries. B. Principal Component Analysis (PCA) on reads abundance shows variation across multiple libraries. Libraries with smaller distance to one another are more similar than ones that are further apart. Biological replicates tend to cluster together. Variations in gene expression between strains are mostly explained by the difference in library preparation methods between N- and C-terminal FLAG-tagged strains (sample segregation along the x-axis).

**Analyses of Changes in Transcript Abundance**

All analyses involving transcript abundance changes (changes in mRNA abundance: log$_2$(FLAG-tagged / WT+EV); ribosome occupancy: log$_2$(Total Ribo-Seq / RNA-S-seq); log$_2$(IP / Total)) were performed in the R programming environment using an R package DESeq2 (Love et al. 2014). The “expected_count” columns in the RSEM file output “isoforms.results” were used as input. Default parameters were used in all analyses with false discovery rate (FDR) to adjust p-value for multiple testing correction. For differential expression analysis between RNA-Seq libraries, adjusted p-value cutoff of 0.01 was used to determine significant changes. For all other analyses, adjusted p-value cutoff of 0.05 was used to determine significant changes.

**Analysis of A-site codon occupancy**

Calculations of relative A-site codon occupancy values were adapted from Hussmann et al. (Hussmann et al. 2015). P-site coordinates were added by 3 to obtain A-site coordinates of the reads. For each mRNA sequence in the transcriptome that had any reads mapped to them, the coordinates of a codon of interest were determined. Reads that had their A-sites mapped within 60-codon
window (-30 to +30 codon positions; -90 to +90 nt positions) of the codon were tabulated. After tabulating reads for each nucleotide position around the codon, reads were summed and divided by number of codons positions to obtain average read density. An entry with average density < 0.1 (i.e., less than 1 read per 10 codons) were discarded from further analyses.

For codon-resolution analysis, reads at nucleotide positions that belong to the same codon position were summed and nucleotide positions were converted to codon positions. Relative occupancy at each position was calculated by dividing the number of reads at that position by the average read density. Relative occupancy of 1 would mean the A-site occupancy at that position is not different from the average across the 30-codon window, where relative occupancy higher and lower than 1 would mean enrichment and depletion, respectively, relative to the average. Then, mean relative occupancy for each codon position was calculated by averaging relative occupancy values of all occurrences of a codon in the transcriptome.

For comparison of mean relative occupancy between IP and Total ribosome samples, mean relative occupancy at each position from IP was divided by that of Total to get a ratio, and log₂ transformed.

**Codon optimality**

We used tRNA adaptation index (tAI), derived from tRNA gene copy numbers and wobble base-pairing penalty (Reis et al. 2004; Tuller et al. 2010b), as a measurement of codon optimality for a given codon identity. Codon optimality
score for a coding sequence was calculated by taking geometric mean of tAIs associated with all codons in the coding sequence (Reis et al. 2004).

**Statistical analyses and data visualization**

Due to the non-normal nature of the data two-tailed Wilcoxon’s rank sum test with false discovery rate (FDR) for multiple testing correction was used to compare ribosome occupancy, coding sequence length, and codon optimality between IP-enriched, IP-depleted, and unchanged groups of mRNAs.

Fisher’s exact test was performed by fisher.test() function and logistic regression by glm() function with “family” parameter = binomial(link = “logit”) in R stats package.

R packages used for data preparation, statistical analyses, and visualization were dEseq2, riboWaltz, limma, stats, rstatix, MASS, ggplot2, ggpubr, ggrepel, ggh4x, scales, patchwork, Cairo, readxl, data.table, dplyr, and reshape2.

**Data Availability**

All custom scripts have been made available at https://github.com/Jacobson-Lab/Upf1-ribosomes. Sequencing data that support this study have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) with the accession code GSE186795 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE186795). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (Perez-Riverol et al., 2019) with the dataset identifier PXD029577 and 10.6019/PXD029577.
Acknowledgments

This work was supported by a grant to A.J. (1R35GM122468) from the U.S. National Institutes of Health. We thank EiEi Min for providing ribosomes used early in the development of the selective ribosome profiling protocol; Alper Celik for his generous gifts of yeast strains and plasmids; Chan Wu, Alper Celik, EiEi Min, Scott Shaffer, Andrei Korostelev, Zhiping Weng, and Sean Ryder for helpful discussions; members of the UMass Chan Medical School Electron Microscopy Core Facility for electron micrographs; members of the UMass Chan Medical School Mass Spectrometry Core Facility for proteomics analyses; and the UMass Chan Medical School Molecular Biology Core Labs (RRID: SCR_018263) for RNA and sequencing library quality control analyses.

Author Contributions

R.G. and A.J. conceived and designed the experiments, R.G. and K.M. carried out the experiments, R.B. and K.M. wrote data processing scripts, K.M., R.B., R.G., F.H., and A.J. analyzed the data, R.G., K.M., and A.J. wrote the paper with input from all authors, and A.J. obtained funding for the study.

Declaration of Interests

A.J. is co-founder, director, and Scientific Advisory Board chair of PTC Therapeutics Inc. All other authors declare no competing interests.
Chapter V: Discussion

Yeast vs. human cis-regulatory elements of readthrough efficiency

For decades, it has been known that the efficiency of termination and stop codon readthrough is regulated by more than just the stop codon identity. Cis-acting elements influencing these efficiencies have been identified in several organisms, and include the identity of the stop codon in combination with flanking nucleotides (Anzalone et al. 2019; Bonetti et al. 1995; Cridge et al. 2018; Namy et al. 2001; Schueren et al. 2014), stem-loop structures in the mRNA 3’-UTR (Anzalone et al. 2019; Firth et al. 2011), and specific RNA-binding protein motifs (Eswarappa et al. 2014; Manjunath et al. 2022). Among these, the stop codon and the nucleotide following it (nt +4) are the most studied across species and yield the most consistent results. Structural insights into their mechanism have been demonstrated with cryo-electron microscopy (Brown et al. 2015; Shao et al. 2016) and their significance has been extended to a transcriptome-wide level in human cells by ribosome profiling experiments (Wangen and Green 2020). However, two aspects of cis-regulatory elements of readthrough that are further away from the stop codon had not been investigated: i) the importance of these elements at a transcriptome-wide level and ii) their conservation across species. The latter had been difficult to determine due to variations in experimental conditions, analysis strategies, and (for endogenous mRNAs) existing nucleotide usage biases in different species.
Major *cis*-acting elements modulating transcriptome-wide translation termination and readthrough efficiency are conserved

In Chapter II and III of this thesis I developed a bioinformatics strategy involving the unbiased random forest machine learning algorithm to investigate complex interplay of all known and hypothesized *cis*-regulatory elements of readthrough at a transcriptome-wide level using ribosome profiling to measure readthrough efficiency of endogenous mRNAs. I applied the same strategy to the yeast data that I generated (Mangkalaphiban et al. 2021) as well as to published readthrough efficiency data derived from ribosome profiling of HEK293T cells (Wangen and Green 2020). The unbiased random forest approaches identified the same elements modulating readthrough efficiency in both human and yeast cells, which include the stop codon, nt +4, 3’-UTR length, and P-site amino acids (Figures 2.7, 3.1, and 5.1). Among these elements, the stop codon and nt +4 (as well as readthrough promoting nucleotides at +5 and +9), which all have high (or relatively high) feature importance scores, influence readthrough efficiency in the same manner in both organisms (Figures 2.7, 3.1, and 5.1). These results are in line with previous knowledge of *cis*-acting elements that modulate termination, extending that knowledge to a transcriptome-wide level and reenforcing the notion that the mechanism of termination is conserved in eukaryotes.
**Figure 5.1.** *Cis*-acting elements modulating readthrough efficiency are mostly conserved between yeast and humans.

Comparison of results analyzed in a similar manner with data from yeast W303 (Mangkalaphiban et al. 2021) (Chapter II) and human HEK293T cells (Wangen and Green 2020) (Chapter III). mRNA features are shaded based on relative importance in readthrough efficiency prediction (red shading). The details of how each feature inhibits or promotes readthrough are ordered from left to right (green shading).
One difference between the data from HEK293T cells and yeast is the relationship between readthrough efficiency and 3'-UTR length in wild-type conditions. Under circumstances where termination is rendered less efficient, either by release factor inactivation in yeast or by aminoglycoside treatment in HEK293T cells, readthrough efficiency increases with 3'-UTR length, implying that the closer proximity of PABP to the stop codon enhances termination efficiency when the normal termination process is compromised (Figures 2.11, 3.5, and 5.1). Under normal circumstances where termination is not hindered, readthrough still increases with 3'-UTR length in untreated HEK293T cells (Figures 3.5 and 5.1), but decreases with 3'-UTR length in wild-type yeast cells (Figures 2.11 and 5.1) (Mangkalaphiban et al. 2021). This discrepancy may be related to the fact that yeast mRNA 3'-UTRs are naturally much shorter than those of human cells (Figure 5.2), limiting possible additional 3'-UTR interactions and regulation. This reduced binding opportunity for additional regulatory proteins to the 3'-UTR and increased proximity of poly(A) tail (and poly(A) tail-associated proteins) to the stop codon may suggest that these regulatory factors have either stronger or weaker influence on 3'-UTR function under normal circumstances in yeast compared to human cells.

Figure 5.2. Yeast vs. human 3'-UTR lengths. Distribution of 3'-UTR lengths for mRNAs that have UTR annotations in the untreated HEK293T data set (pink) (Wangen and Green 2020) (Chapter III) and the WT yeast (SUP45, 25°C) data set (yellow) (Mangkalaphiban et al. 2021) (Chapter II).
P-site tRNA abundance and/or properties may influence the efficiency of translation termination

Contrary to other mRNA contexts where the trends are similar between human and yeast data, the identity of P-site codons that are associated with high or low readthrough efficiency mostly differ between human and yeast cells (Figures 2.10, 3.4, and 5.1). Except for the GCU codon that is associated with low readthrough almost uniformly across samples in both species, no other codons in human cells showed patterns consistent with those found to be significant in yeast cells. Because the nucleotide and amino acid properties do not vary between species, the differences in results could be attributable to structural variation between yeast and human ribosomes or the difference in tRNA pools and properties. For the latter scenario, even though the mechanism underlying the influence of P-site position in readthrough efficiency may be conserved, the variation in tRNA abundance and modifications between species can result in differences in the identity of the codon identified in our P-site codon analyses. Although the P-site tRNA has been claimed to have no effect on readthrough levels, this conclusion was derived from observed changes in readthrough levels when the third wobble base of the codon was mutated while keeping the same decoding tRNA for both the original and the mutated codons unchanged (Floquet et al. 2012). Additional experiments that can control for nucleotide and amino acid identities while varying the tRNA properties will be needed to pinpoint the underlying mechanism behind the P-site codon’s effect on readthrough efficiency.
Future directions in readthrough efficiency prediction

In Chapter III, my collaborators from the Bedwell lab and I demonstrated that random forest models trained on NTC readthrough efficiency values derived from ribosome profiling data of endogenous mRNAs can predict PTC readthrough efficiency with high accuracy (Figure 3.7). This result opens up possibilities of how mining high-throughput sequencing data can help facilitate both basic research and therapeutics development involving readthrough, from optimization of experimental design to prediction of disease improvement outcome in patients upon drug treatment. However, additional tests for the most appropriate algorithm and additional data might need to be considered to improve prediction accuracy.

Advantages and disadvantages of the random forest algorithm

Random forest (Breiman 2001) is an ensemble of decision trees (Figure 5.3A-B). A decision tree is a supervised learning algorithm that sequentially splits data into two groups based on a decision rule at each split. For its application in readthrough that I used in Chapters II and III, a decision tree splits readthrough efficiency values (Y or response variable) into groups based on mRNA features (X or predictor variables), as exemplified in Figure 5.3A. Because each mRNA feature is not limited to one decision rule, which would be the case in some other algorithms such as linear regression where a variable is given one coefficient, decision trees can handle non-linear relationships and complex interactions well. An example for complex interaction is when a role of one variable depends on a
value of another variable (e.g., variable Z only affects readthrough significantly if the stop codon is UGA), which can be captured by the decision tree (e.g., Z is never chosen to be a decision rule until UGA is already its own branch). However, one decision tree is prone to overfitting the data and lacks predictive accuracy (James et al. 2013), which is improved by random forest. Random forest aggregates hundreds of decision trees (i.e., making a forest!) with added randomness: i) for each tree, only about 2/3 of the data is randomly selected for training / building the tree while about 1/3 of the data (out-of-bag or OOB data) is used for testing, and ii) for each split, only a subset of predictors is kept for decision rule consideration (Figure 5.3B). Prediction of a new data point is an average of predictions from all trees for the regression model or a majority vote across predictions from all trees for the classification model. Thus, the random forest approach is robust to noise variables and/or outliers, reduces overfitting, and increases prediction accuracy (Hastie et al. 2009; James et al. 2013). Another advantage of random forest is the variable importance measure. OOB data used to test the model gives a model performance metric as a prediction error, called OOB error. By permuting or shuffling the values of one predictor at a time, random forest records a new OOB error in response to variable permutation and compares the new error to the original error. The changes in OOB errors are quantitative. The predictor whose permutation increases OOB error and decreases model accuracy the most is considered the most important predictor in the model. Hence, I used
random forest to identify important mRNA features in predicting readthrough efficiency.

Despite its various advantages, random forest has its disadvantages, mostly in the realm of new data prediction. As seen in Figure 5.3A, if the new data point is missing a value for a predictor, it cannot go through the decision tree. Giving an arbitrary value for that predictor might not be possible in all cases. For example, if the missing value predictor is gene expression, the fact that the value is missing might mean expression is not detected at all and the value of zero can be assigned. However, if the missing value is the identity of nt -7 because the stop codon (+1 to +3) appears so early, there is no baseline value that can be given at that position (or positions further upstream) because the four choices of values (i.e., A, C, G, or U) have equal weights; choosing one over another has no biological meaning and could skew the result. Since the models in Chapters II and III include up to 90 nts upstream of the stop codon for nascent peptide features, this means readthrough of PTC or NTC that appears less than 90 nts away from the AUG codon cannot be predicted, at all or accurately. In comparison, in linear regression where contributions of predictors are independent and additive (Figure 5.3C, left), the missing predictor is simply dropped from the prediction formula and readthrough efficiency can still be predicted, but this is not possible with random forest. In the same vein, predictions from random forest will only be the average of values that already exist in the training data. In other words, random forest cannot extrapolate outside of the training set. For instance, a decision rule may state that
mRNAs with 3'-UTR lengths > 256 nt will be given a readthrough efficiency value of 3.4; otherwise it will be 1.2. In this case, two mRNAs with differing 3'-UTR lengths of 300 nt and 3000 nt will be predicted to have the same readthrough efficiency of 3.4. In linear regression where the importance of a predictor (i.e., coefficient) is multiplicative of the predictor value itself, the readthrough efficiency prediction for 3000 nt may be higher than 3.4. However, this disadvantage of random forest may not be too problematic if the training data set is vast and encompasses the distribution of “real world” data.

Nevertheless, a comparison of different machine learning algorithms is still needed to find an algorithm with the best performance in readthrough prediction. If all the important predictors have approximated linear relationships with readthrough efficiency, perhaps multiple linear regression or its regularized variations will outperform decision tree-based models (James et al. 2013). If not, then other non-linear models may be considered, including but not limited to i) generalized additive model (GAM), an extension of linear models to include non-parametric, non-linear functions but maintain additive relationships between predictors (James et al. 2013) (Figure 5.3, right), and ii) neural network, an algorithm that mimics the neurons and synapses in human brain, which are functions and connections, respectively (Hastie et al. 2009).
**Figure 5.3. Examples of machine learning models.**

**A.** A simple example of a decision tree and how decision rules might look when trained to predict readthrough efficiency (response variable, Y) based on mRNA features (predictor variables, X).

**B.** Visual presentation of random forest, an ensemble of decision trees.

**C.** Examples of alternative models that can be used to predict readthrough efficiency. Linear regression and generalized additive model (GAM) are both additive models where individual contributions of predictors are added together, but the relationship between the predictor and outcome is linear for linear regression (represented by coefficients) while it can be non-linear and/or non-monotonous for GAM (represented by functions) (James et al. 2013).
Additional considerations to improve readthrough prediction and extend its application

Other than finding the best performing algorithm, another factor that contributes to prediction accuracy is the quality and quantity of the input data. As discussed earlier in this chapter as well as in Chapter III, P-site and A-site tRNAs likely contribute to the efficiency of termination and readthrough processes. Including tRNA properties, such as their abundance, type of modifications, and site of modifications as predictors likely will increase the accuracy of readthrough prediction, ideally when they are determined in parallel with ribosome profiling experiments (from the same set of cells). A recent multi-omics study of bacterium *Mycoplasma pneumoniae* demonstrated values of time-course RNA-Seq, ribosome profiling, tRNA-Seq, and quantitative mass spectrometry in providing comprehensive, high-quality quantitative data for mathematical modeling of translation initiation and elongation rates (Weber et al. 2023). However, this study also concluded that there exist unidentified mechanisms of translation regulation other than measured here. Similarly for translation termination, it is highly likely that there exist regulatory elements of readthrough that have yet to be discovered. Computational approaches utilized in pattern discovery or pattern recognition may be used to find new motifs from the mRNA sequences.

The current models showed that sequence context proximal to the stop codon helps to mediate readthrough efficiency. It is possible then that for readthrough drug candidates that failed clinical trials because they failed to
improve disease phenotypes in a sufficient number of patients, the drug may have been effective for patients with readthrough-permissive context but not patients with readthrough-inhibiting context. Thus, the models are potentially useful in predicting disease outcomes upon drug treatment and in aiding clinical trial design. However, additional factors will need to be considered for that purpose. Although readthrough efficiency is deterministic of the level of full-length protein produced, the protein product is not necessarily fully functional. In many cases, the amino acid inserted at a PTC is different from the original, non-mutated one. Since stop codon context also affects the identity and frequency of amino acid insertion (Blanchet et al. 2014; Xue et al. 2017), including such information as predictors may improve a model's predictive ability. However, for this information to be most useful, it is critical to understand how the inserted amino acid affects protein function. Therefore, protein chemistry is another layer of information to consider in order to extend the model application to prediction of disease phenotypes.

One caveat of the models trained on data derived from ribosome profiling of G418-treated cells is that G418 is a double-edged sword. While its potency allows an elevated number of readthrough events to be detected with high accuracy, G418 is also toxic, possibly because of its effects on NTCs (Wangen and Green 2020; Schilff et al. 2021; Li et al. 2023). A safe drug ideally would have negligible effects on NTCs, which makes developing a model and learning from it based on NTC readthrough data from endogenous mRNAs not possible. However, other high-throughput approaches to generate readthrough data from diverse
mRNA sequences can be utilized. An approach involving an mRNA display library coupled with \textit{in vitro} translation selection has been developed, where only readthrough ribosomes are selectively isolated followed by high-throughput sequencing of the associated RNAs, and the abundance of sequences post-selection vs. pre-selection is analyzed (Anzalone et al. 2019). A similar readthrough-selective approach coupling an mRNA library with live cells \textit{in vivo} may be possible. For example, if cells express a fluorescent reporter that can only be expressed upon readthrough are sorted by fluorescence activated cell sorting (FACS) followed by targeted high-throughput sequencing of the reporters, then the abundance of sequences post-FACS vs. pre-FACS can be analyzed. An mRNA library with live cells followed by targeted ribosome profiling (e.g., using primers to amplify only reporter library sequences during sequencing library preparation) is also an option. These systems can be applied to study sequence contexts that regulate readthrough in untreated vs. drug-treated cells.

**Limitations of current methods and future directions in studying premature termination recognition by NMD**

In an attempt to understand how NMD, or Upf1 particularly, recognizes premature termination and distinguishes it from normal termination, we carried out selective ribosome profiling of Upf1-associated ribosomes to capture Upf1’s recognition of premature termination as it is happening during translation. These experiments provide insights into Upf1’s actions on the ribosome, showing that there are at least two distinct Upf1-80S complexes, that Upf1 is not displaced by
ribosomes during translation but is carried by them, and that Upf1 can act on normal termination presumably when it is inefficient (Chapter IV).

However, Upf1 needed to be overexpressed in order to yield sufficient material for sequencing library preparation. Overexpression of Upf1 may have led to non-specific ribosome binding, although data from Upf1 ATPase mutant cells helped determine Upf1-specific results. Nonetheless, data from endogenously expressed Upf1 is ideal and could give more accurate insights into its function. Strategies to improve recovery of endogenously expressed Upf1-bound ribosomes include using better epitope tags for immunopurification and utilizing different Upf1 mutants to increase the yield of Upf1-80S complexes trapped at different stages of premature termination recognition.

Another drawback of our experiments is that ribosome footprints for premature termination events, either in endogenous PTC alleles or those that happen randomly during translation, were all too infrequent for meaningful analyses, even with overexpressing Upf1 and deleting Upf2 to increase yield. We had expected the diversity of premature termination events occurring naturally in cells would further our understanding of premature termination recognition; however, their occurrence was masked by the much more abundant normal translation captured in the library. So, it may be more beneficial for now to sacrifice the diversity aspect by using a few highly overexpressed PTC-containing reporters, so that the yield of ribosome footprints is high enough for meaningful analyses. In addition to ribosome profiling, cryoEM is a powerful approach that can help
visualize the molecular mechanism during premature termination recognition by the Upf factors.
APPENDIX I:

Direct and indirect consequences of PAB1 deletion in the regulation of translation initiation, translation termination, and mRNA decay

Abstract

Many human diseases are caused by having too much or too little of certain cellular proteins. The amount of an individual protein is influenced by the level of its messenger mRNA (mRNA) and the efficiency of translation of the mRNA into a polypeptide chain by the ribosomes. Cytoplasmic poly(A)-binding protein (PABPC) plays numerous roles in the regulation of this multi-staged process, but understanding its specific role has been challenging because it is sometimes unclear whether experimental results are related to PABPC’s direct role in a specific biochemical process or to indirect effects of its other roles, leading to conflicting models of PABPC’s functions between studies. In this study, we characterized defects of each stage of protein synthesis in response to loss of PABPC in yeast cells by measuring whole-cell levels of mRNAs, ribosome-associated mRNAs, and proteins. We demonstrated that defects in most steps of protein synthesis other than the last can be explained by reduced levels of mRNAs that code for proteins important for that step in addition to loss of PABPC’s direct role on that step. Our data and analyses serve as resources for the design of future studies of PABPC’s functions.
Introduction

Eukaryotic mRNAs are subject to complex post-transcriptional regulation by RNA binding proteins (RBPs) that control protein output. Cytoplasmic poly(A)-binding proteins (PABPCs) are RBPs that bind polyadenylated tails at mRNA 3’-ends and subsequently play roles in multiple stages of cytoplasmic mRNA regulation, from translation initiation to termination and mRNA decay (Mangus et al. 2004; Brambilla et al. 2019; Passmore and Coller 2022). The numerous functions of PABPC are attributed to its ability to interact with not only mRNAs but also various other proteins to form messenger ribonucleoprotein (mRNP) complexes (Brambilla et al. 2019). PABPC’s conserved structure consists of three regions: i) four RNA recognition motif (RRM) domains, two of which (RRM1 and RRM2) bind 12 adenosines with high affinity (Sachs et al. 1987; Passmore and Coller 2022), while the protein overall covers 27 adenosines (Baer and Kornberg 1983), ii) a proline-rich (P) linker domain, and iii) a C-terminal mademoiselle (MLLE) domain that interacts with other proteins via their PABP-interacting motif 2 (PAM2) (Brambilla et al. 2019; Passmore and Coller 2022; Qi et al. 2022). Mammals have multiple isoforms of PABPCs, of which the most ubiquitous isoform as well as the most studied is PABPC1, whereas the yeast Saccharomyces cerevisiae has only one PABPC, Pab1 (Sachs et al. 1986; Brambilla et al. 2019; Passmore and Coller 2022; Qi et al. 2022).

The current model for general mRNA decay in yeast involves biphasic poly(A) tailing shortening by Pan2-Pan3 and Ccr4-Not deadenylase complexes,
decapping by the Dcp1/Dcp2 holoenzyme, and exonucleolytic Xrn1-mediated 5’-3’ degradation or exosome-mediated 3’-5’ degradation (Passmore and Coller 2022; Liu et al. 2022; He and Jacobson 2022). PABPC appears to have a paradoxical role in these processes. On the one hand, PABPC1/Pab1 stimulates deadenylation by recruiting Pan2-Pan3 to the poly(A) tail through its interaction with the PAM2 motif on Pan3 (Mangus et al. 2004; Siddiqui et al. 2007; Schäfer et al. 2019; Passmore and Coller 2022; Liu et al. 2022). Consistent with this model, mRNAs in yeast cells harboring a deletion of PAB1 or mutations of Pab1’s C-terminal Pan3-interacting domain had longer poly(A) tails than their counterparts in wild-type cells (Sachs and Davis 1989; Caponigro and Parker 1995; Brown and Sachs 1998; Mangus et al. 2004; Brambilla et al. 2019; Passmore and Coller 2022). On the other hand, PABPC has been shown to protect mRNAs from exonucleases (Passmore and Coller 2022; Liu et al. 2022).

Importantly, mRNAs with longer poly(A) tails are generally more stable and better translated than their short-tailed or un-tailed counterparts (Doel and Carey 1976; Deshpande et al. 1979; Rosenthal et al. 1983; Jacobson and Favreau 1983; Palatnik et al. 1984; Drummond et al. 1985; Galili et al. 1988; Gallie et al. 1989; Munroe and Jacobson 1990a; Passmore and Coller 2022), observations that led us to propose a role for PABPC in the enhancement of translation initiation by the possible formation of an mRNA “closed-loop” (Munroe and Jacobson 1990a, 1990b; Jacobson 1996). Current elaborations of this model postulate that mRNAs could be circularized by a chain of interactions between poly(A)-associated PABPC
and 5’ end-localized initiation factors, giving rise to a poly(A) tail-PABPC-eIF4G-eIF4E-5’cap network (Tarun and Sachs 1996; Wells et al. 1998; Amrani et al. 2008; Borman 2000; Dever et al. 2016; Brambilla et al. 2019). eIF4G has been shown to interact with PABPC’s RRM2 domain (Tarun and Sachs 1996; Imataka 1998; Kessler and Sachs 1998; Melamed et al. 2015), and disrupting this interaction reduced translation (Borman 2000; Michel et al. 2000; Qi et al. 2022). Thus, PABPC can stabilize the cap-binding complex and aid the recruitment of the 43S pre-initiation complex to the mRNA (Merrick and Pavitt 2018; Brambilla et al. 2019; Passmore and Coller 2022). However, this closed-loop arrangement is not required for all mRNAs or all conditions, raising the question of how else 5’-3’ communication is facilitated or whether it is a universal step (Costello et al. 2015; Dever et al. 2016; Thompson and Gilbert 2017; Vicens et al. 2018; Kershaw et al. 2023). Depletion of PABPC1 in mammalian cells had minimal effects on transcriptome-wide translation efficiency (Xiang and Bartel 2021; Kajjo et al. 2022), suggesting that the stimulatory effect of PABPC on translation initiation may be restricted to circumstances where translation initiation efficiency is rate-limited.

In addition to its interactions with initiation factors, PABPC also interacts with the release factor eRF3 via its PAM2 motif in metazoans and its P-C domains in yeast (Hoshino et al. 1999a; Cosson et al. 2002a, 2002b; Roque et al. 2015; Brambilla et al. 2019; Qi et al. 2022). Translation termination involves stop codon recognition in the ribosomal A-site and nascent peptide release by eRF1, whose hydrolysis function and conformational change are stimulated by eRF3’s GTPase
activity (Hellen 2018; Salas-Marco and Bedwell 2004). PABPC is thought to enhance termination efficiency by promoting the recruitment of the eRF1-eRF3 release factor complex to the stop codon. The role of PABPC in termination is also inferred from: i) the observation that tethering PABPC1 or Pab1 downstream of premature termination codons (PTCs) antagonized nonsense-mediated mRNA decay (NMD), an mRNA decay pathway thought to be activated by the reduced termination efficiency of premature translation termination (Amrani et al. 2004; Ivanov et al. 2008) and ii) the increased termination efficiency observed with proximity of a stop codon to the mRNA 3’ end (Swart et al. 2016; Wu et al. 2020). Direct evidence for PABPC’s ability to enhance termination includes: i) addition of PABPC1 to an in vitro termination assay improved termination efficiency (Ivanov et al. 2016), ii) PABP-interacting protein PAIP1 and PAIP2 competed with eRF3 for free PABPC binding, reducing termination efficiency of PTCs in vitro (Ivanov et al. 2019b), and iii) deletion of PAB1 or Pab1’s P-C domains in vivo increased stop codon readthrough efficiency of reporter PTCs in a proximity-dependent manner (Wu et al. 2020). However, as with initiation, a full understanding of PABPC’s role during termination of endogenous mRNAs in vivo is still lacking.

PABPC’s involvement in many major stages of mRNA regulation and translation complicate attempts to define a specific role for PABPC in vivo by deleting, depleting, overexpressing, or mutating the protein and have led to conflicting models of PABPC’s function. Therefore, to specifically assess direct and indirect consequences of deleting PABPC, we generated and analyzed RNA-Seq,
ribosome profiling (Ingolia et al. 2009), and mass spectrometry data from yeast cells lacking Pab1. As expected, mRNA and protein abundance changed substantially in \( pab1\Delta \) cells. We found that deleting \( PAB1 \) resulted in a translation termination defect that appears to be at least partially due to reduced eRF3 protein level as well as to loss of Pab1’s stimulatory function on termination. In addition, translation initiation defects and changes in relative translation efficiency in \( pab1\Delta \) cells may be confounded by reduced initiation factor levels, especially eIF4G and eIF1. Further, an analysis of decapping activator substrates revealed that increased levels of certain mRNA subgroups may be partially caused by reduced levels of a specific decapping activator or components of Ccr4-Not deadenylation complex. Together, our results demonstrate direct and indirect consequences of \( PAB1 \) deletion and illustrate the complexity of Pab1’s role in transcriptome-wide regulation of translation and mRNA decay.

**Results**

**Deletion of \( PAB1 \) promotes transcriptome-wide accumulation of ribosomes downstream of normal stop codons**

Deletion of \( PAB1 \) has been shown to decrease translation termination efficiency and increase stop codon readthrough efficiency of reporter PTCs *in vivo* (Wu et al. 2020). PTC readthrough occurs when a near-cognate tRNA outcompetes eRF1 in stop codon decoding, resulting in continued in-frame translation elongation and production of a C-terminally extended polypeptide (Dabrowski et al. 2015; Rodnina et al. 2020). To determine whether decreased
termination efficiency and increased readthrough efficiency could also be observed at normal termination codons (NTCs) of endogenous mRNAs when Pab1 is absent, we performed ribosome profiling and RNA-Seq analyses of yeast cells harboring a PAB1 deletion. Because PAB1 is an essential gene, the deletion was done in a pbp1Δ background, which suppresses pab1Δ lethality (Mangus et al. 1998), and we used the PAB1/pbp1Δ strain as our wild-type PAB1 control. Detailed descriptions of pbp1Δ cells have been described elsewhere (Mangus et al. 1998; Tuong Vi et al. 2021; Van De Poll et al. 2023). Sequencing data obtained from three biological replicates of each strain were reproducible, as evidenced by high Pearson’s correlation coefficients between replicates (Figure A1.1A).

**Figure A1.1. Replicate reproducibilities.**

A. Correlation matrix showing Pearson correlation coefficients (r) of transcript abundance (non-zero RPKM values, log₁₀-transformed) between pairs of sequencing libraries, RNA-seq and ribosome profiling libraries. B. Correlation matrix showing Pearson correlation coefficients (r) of log₂ intensity values from mass spectrometry data between pairs of samples.
### A

**Pearson's r**

<table>
<thead>
<tr>
<th></th>
<th>RNA WT</th>
<th>pbp1Δ</th>
<th>pbp1Δpab1Δ</th>
<th>RPF WT</th>
<th>pbp1Δ</th>
<th>pbp1Δpab1Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
<td>0.96</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td>2</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
<td>0.96</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td>3</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>0.95</td>
<td>0.93</td>
<td>0.94</td>
</tr>
</tbody>
</table>

**Protein**

<table>
<thead>
<tr>
<th></th>
<th>RNA WT</th>
<th>pbp1Δ</th>
<th>pbp1Δpab1Δ</th>
<th>RPF WT</th>
<th>pbp1Δ</th>
<th>pbp1Δpab1Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
<td>0.98</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>2</td>
<td>0.95</td>
<td>0.93</td>
<td>0.93</td>
<td>0.96</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>3</td>
<td>0.95</td>
<td>0.95</td>
<td>0.94</td>
<td>0.95</td>
<td>0.95</td>
<td>0.94</td>
</tr>
</tbody>
</table>

### B

**Pearson's r**

<table>
<thead>
<tr>
<th></th>
<th>Protein WT</th>
<th>pbp1Δ</th>
<th>pbp1Δpab1Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.99</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td>0.95</td>
<td>1.00</td>
<td>0.97</td>
</tr>
<tr>
<td>3</td>
<td>0.95</td>
<td>0.99</td>
<td>0.97</td>
</tr>
</tbody>
</table>
In all strains, P-sites of ribosome profiling reads mostly map to the coding region of mRNAs, correspond to one major reading frame (frame 0), and show the expected 3-nucleotide periodicity indicative of translating ribosomes (Figure A1.2A-C, “CDS” region). Notably, the relative amount of ribosome footprints found in mRNA 3’-UTR regions is increased in \( pab1\Delta pbp1\Delta \) cells compared to \( pbp1\Delta \) or WT cells (Figure A1.2A inset and B), demonstrating that cells lacking Pab1 manifest an apparent termination defect. This defect occurred not only at canonical stop codons, as evidenced by an increase of footprints in the “extension” region (the 3’-UTR region from the canonical stop codon to the next in-frame stop codon), but also at the first in-frame stop codons downstream of the canonical stop codon, as manifested by an increase of footprints in the distal 3’-UTR region (Figure A1.2B).

The presence of ribosomes in the 3’-UTR can arise from stop codon readthrough, ribosome frameshifting, or reinitiation. To determine the primary driver of increased ribosome footprints in the 3’-UTR of mRNAs in \( pab1\Delta pbp1\Delta \) cells, analyses of reading frame proportions in different mRNA regions were carried out. Stop codon readthrough would yield footprints predominantly in reading frame 0 in the extension region, while frameshifting or reinitiation events would not show this bias. Indeed, the proportion of ribosome footprints in reading frame 0 increases in the extension region in \( pab1\Delta pbp1\Delta \) cells compared to \( pbp1\Delta \) or WT cells (Figure A1.2C) and this increase is comparable to that observed in ribosome profiling data of cells depleted of functional eRF1 (Mangkalaphiban et al.)
2021; Wu et al. 2019). These results demonstrate that a notable portion of 3'-UTR footprints in pab1Δpbp1Δ cells arises from stop codon readthrough, thus suggesting that translation termination is less efficient in the absence of Pab1.

To verify that the increase in stop codon readthrough is transcriptome-wide, i.e., that the results of Figures 1A-C were not derived from a limited number of mRNAs, we calculated readthrough efficiency for each mRNA by dividing the density of frame 0 footprints in the extension by that in the CDS region. The number of mRNAs with detectable readthrough is increased in pab1Δpbp1Δ cells, almost double that observed in WT, and overall readthrough efficiency in pab1Δpbp1Δ cells is significantly higher than in the other two strains (Figure A1.2D).

Curiously, the pbp1Δ strain also shows higher overall readthrough efficiency than the WT strain (Figure A1.2D), raising the question of whether Pbp1 plays a role in termination and contributes to increased readthrough in the pab1Δpbp1Δ strain. However, this observation is most likely not due to stop codon readthrough. Footprints in the 3'-UTR region in the pbp1Δ strain are only higher than in WT in the few codons immediately following the canonical stop codon, while in the pab1Δpbp1Δ strain, the increase in 3'-UTR footprints extend substantially beyond that (Figure A1.2A, inset). More importantly, the proportion of frame 0 footprints in the extension region in pbp1Δ cells does not differ from that in WT cells (Figure A1.2C). Although we limited our readthrough efficiency calculation to only frame 0 footprints to ensure as accurate a calculation as possible, we still cannot completely exclude other events, such as reinitiation, that happened to also
produce ribosome footprints in frame 0. Thus, despite the fact that deletion of \textit{PBP1} did not result in proportionally higher stop codon readthrough events compared to WT, a slight increase in footprints immediately following the stop codon leads to an apparent increase in the calculated readthrough efficiency values in the \textit{pbp1Δ} strain.

In short, \textit{PAB1} deletion resulted in an apparent translation termination defect, manifested as increased footprints in mRNA 3'-UTR regions. The preference for frame 0 footprints in the extension region and the higher median readthrough efficiency values calculated for individual mRNAs observed in \textit{pab1Δpbp1Δ} cells compared to the two controls indicate that stop codon readthrough increases transcriptome-wide in the absence of Pab1.
**Figure A1.2. PAB1 gene deletion leads to ribosome accumulation at mRNA start codons and in UTR regions.**

A. Ribosome footprints (replicate libraries were pooled) were counted by their P-site positions in the indicated nucleotide window around the canonical start and stop codons of annotated ORFs. Raw footprint counts were normalized by the total footprint count in the windows. Inset: Magnified view of the 3'-UTR region (Distance from stop > 0 nt). B-C. Percentage of footprints in sequencing library belonging in different mRNA regions (B) and percentage of frame 0 footprints in each mRNA region, where grey dashed line indicates a theoretical 33% at which all 3 reading frames are equally represented (C). “Start” region includes the canonical AUG and 3 flanking nucleotides on each side. “Stop” region includes the canonical stop codon and 3 flanking nucleotides on each side. “Extension” is the region following the “Stop” until (but not including) the first in-frame stop codon in the 3'-UTR. “Distal 3'-UTR” is the 3'-UTR region following “Extension.” Percentages from individual replicate libraries (grey points) were averaged (bar plot and reported value above it). Unpaired Student’s t-test with Benjamini-Hochberg method for multiple-testing correction was used to compare values between pairwise strains. D. Readthrough efficiency distribution in each strain (see Methods for calculation). Footprints from replicate libraries were pooled. Two-sided Wilcoxon’s rank sum test with Benjamini-Hochberg method for multiple-testing correction was used to compare values between pairwise strains. For B-D, only significant comparisons were reported as the following: (*) p < 0.05, (***) p < 0.01, (****) p < 0.001, (***** p < 0.0001. For all panels, only reads belonging to genes with UTR annotations and minimally overlapping sequences (less than 18 bp overlap with another gene on the same strand) were included in the analyses (2,693 genes).
Depletion of release factors may account for the increased ribosome footprints in the 3’-UTR upon PAB1 deletion

Because PABPC plays major roles in the regulation of mRNA decay and translation, the termination defect and increased readthrough observed in \( \textit{pab1}\Delta \textit{pbp1}\Delta \) cells can be due to: i) loss of Pab1’s direct function in termination via its interaction with eRF3 (Roque et al. 2015; Ivanov et al. 2016) or ii) changes in the stability of release factor mRNAs or changes in their translation that result in depletion of the respective proteins, which in turn affect global termination efficiency. To test for the latter scenario, we analyzed mRNA and protein abundance changes between strains using RNA-Seq and mass spectrometry data. As expected of Pab1’s essential and extensive role in mRNA stability regulation, steady-state mRNA and protein levels changed drastically when Pab1 was absent (Figure A1.3A-B) while the absence of Pbp1 had minimal effects (Figure A1.3A-B). For proteins that were detectable by mass spectrometry, changes in their mRNA and protein abundance are quite consistent with each other, with Spearman’s correlation of 0.62-0.74 (Figure A1.3C). To study the effects of \( \textit{PAB1} \) deletion, we focused further analyses on abundance changes in \( \textit{pab1}\Delta \textit{pbp1}\Delta \) cells relative to \( \textit{pbp1}\Delta \) cells.
Figure A1.3. PAB1 gene deletion has significant effects on the transcriptome and the proteome.

A. Volcano plots of changes in transcriptome (RNA-Seq data) between strains. Orange, purple, and grey dots represent mRNAs with higher abundance (positive log₂ fold change, adjusted p-value < 0.01), lower abundance (negative log₂ fold change, adjusted p-value < 0.01), and no change (adjusted p-value ≥ 0.01), respectively. B. Volcano plots of changes in proteome (mass spectrometry data) between strains. Orange, purple, and grey dots represent proteins with higher abundance (positive log₂ fold change, adjusted p-value < 0.015), lower abundance (negative log₂ fold change, adjusted p-value < 0.015), and no change (adjusted p-value ≥ 0.015), respectively. C. Comparison of log₂ fold change in transcriptome and proteome, with Spearman’s correlation coefficient. Grey, genes whose mRNA and protein abundance remained unchanged. Blue, genes whose protein but not mRNA abundance changed significantly. Red, genes whose mRNA but not protein abundance changed significantly. Green, genes whose mRNA and protein abundance both changed significantly.

Among the detectable proteins related to mRNA translation, the pab1Δ mutation had the strongest effect on the levels of ribosomal proteins (Figure A1.4, bottom left) and multiple initiation factors (Figure A1.4, top left), but only appeared to have small effects on elongation factors (Figure A1.4, top right). This result is corroborated by gene ontology analysis, which identified translation and ribosome biogenesis as two of the many pathways enriched in the down-regulated group of proteins (Interactive Figure A1.1).

For the two release factors, eRF1 (Sup45) showed a slight reduction in its mRNA level but not its protein level, while eRF3 (Sup35) showed significant reduction in both mRNA and protein levels in the absence of Pab1 (Figure A1.4, bottom right), where the eRF3 protein level was 80% of that in pbp1Δ cells. However, the reduction in release factor levels may not necessarily result in reduced termination efficiency. If overall translation is also reduced, the normal
stoichiometry of supply and demand for release factors may still be maintained or supply may even exceed demand. The latter conclusion follows from observations that PABPC depletion substantially reduces overall protein synthesis such that almost all heavy polysomes are lost (Mangus et al. 1998; Kajjo et al. 2022) and protein synthesis is limited by the amount of free ribosomes (Shah et al. 2013; Dever et al. 2016; Rogers et al. 2017; Fernandes et al. 2017). Hence, since termination can only occur after initiation and elongation, the larger reductions in ribosomal proteins and initiation factors (the most reduced ribosomal protein and initiation factor are respectively reduced to 45% and 64% of their normal levels) would most likely be limiting and release factors would thus be expected to still be in excess.

We also considered whether increased ribosome footprints in the 3'-UTR arise from a reduced level of Rli1, a ribosome recycling factor. However, since recycling can only occur after termination, Rli1 may still be in excess stoichiometrically in the context of overall reduced translation in \( \textit{pab1} \Delta \textit{pbp1} \Delta \) cells. More importantly, \( \textit{pab1} \Delta \textit{pbp1} \Delta \) cells show a preference for frame 0 in the extension region (Figure A1.2C), consistent with stop codon readthrough, while Rli1 depletion cells resulted in reinitiation in the 3'-UTR in all three reading frames (Young et al. 2015).

Termination occurs when the eRF1/eRF3 complex outcompetes the near-cognate aminoacyl-tRNA (aa-tRNA)/eEF1A complex in binding to the ribosomal A-site. Thus, in addition to the shift in equilibrium between the demand and supply of
release factors, the shift in equilibrium between the levels of the eRF1/eRF3 complex and the aa-tRNA/eEF1A complex also affects termination efficiency. In \textit{pab1}\textsubscript{Δ}\textit{pbp1}\textsubscript{Δ} cells, the relative level of eEF1A mRNA is unchanged but eEF1A protein is at 129\% relative to the level in \textit{pbp1}\textsubscript{Δ} cells (Figure A1.4, top right). This increase in eEF1A to eRF3 ratio may increase the chance of stop codon decoding by aa-tRNA/eEF1A, reducing termination efficiency and increasing ribosome footprints in the 3’-UTR regions.

Together, previous studies and our results here suggest that a significant reduction in translation is reducing the demand for release factors and recycling factors, making them likely to still be in excess. Therefore, the observed decreased termination efficiency or increased stop codon readthrough in \textit{pab1}\textsubscript{Δ}\textit{pbp1}\textsubscript{Δ} cells is not because the release factor levels are limiting but is likely due to an increase in competitive advantage of aa-tRNA/eEF1A in stop codon decoding in addition to a loss of Pab1’s stimulatory function on termination.

\textbf{Figure A1.4.} \textit{PAB1} gene deletion changes mRNA and protein levels of many components of the translation machinery. As in Figure A1.3C for \textit{pab1}\textsubscript{Δ}\textit{pbp1}\textsubscript{Δ}/\textit{pbp1}\textsubscript{Δ} comparison, with the focus on translation-related genes. The ribosomal proteins depicted here account for 94\% of all ribosomal proteins that make up the 40S and 60S subunits. (Next page)
3'-UTR length is no longer predictive of readthrough efficiency when *PAB1* is deleted

Pab1 has been shown to affect NMD-sensitivity and readthrough efficiency of PTC-containing mRNAs in a manner dependent on PTC proximity to mRNA-associated Pab1 (Amrani et al. 2004, 2006; Wu et al. 2020). As would be expected from this relationship, readthrough of PTCs in reporter mRNAs increased in response to 3'-UTR lengthening, but this trend was lost when *PAB1* was deleted (Wu et al. 2020). Recently, we investigated the *cis*-regulatory elements of transcriptome-wide stop codon readthrough using a random forest machine learning approach and found that 3'-UTR length was an important predictor of readthrough, where mRNAs with short 3'-UTRs had lower readthrough than those
with long 3′-UTRs when eRF1’s functionality was compromised, but the trend was the opposite in WT cells (Mangkalaphiban et al. 2021). These data led us to further assess the involvement of Pab1 and its proximity to the stop codon as a predictor of readthrough efficiency. If Pab1 is involved, we expected that the relationship between readthrough efficiency and 3′-UTR length would disappear or weaken when PAB1 is deleted. Thus, we applied the same random forest approaches to identify mRNA features that influence the prediction of readthrough efficiencies in WT, pbp1Δ, pab1Δpbp1Δ strains (Figure A1.5A). As expected, the negative control features (NC) have no influence on readthrough efficiency prediction in any strain, while the identity of the stop codon is an important predictor of readthrough in all strains (Figure A1.5A). The length of the 3′-UTR is an important predictor of readthrough efficiency in WT and pbp1Δ cells but is no longer important in pab1Δpbp1Δ cells (Figure A1.5A). The relationship between 3′-UTR length and readthrough efficiency is slightly negatively correlated in WT and pbp1Δ strains, but it is weakened in the pab1Δpbp1Δ strain (Figure A1.5B, “all”). The fact that the correlation is not completely eliminated even in the absence of Pab1 could be due to technical limitations of readthrough efficiency calculation, where frame 0 non-readthrough 3′-UTR footprints were inevitably included. Nevertheless, when other mRNA features were included in the analysis and controlled for (i.e., random forest regression model), this weak correlation became insignificant in predicting readthrough efficiency in pab1Δpbp1Δ cells.
To further see how 3′-UTR length synergistically regulates readthrough with the strongest feature, stop codon identity, we grouped mRNAs by their stop codon identities and then performed the correlation analysis (Figure A1.5B). We found that the correlation between readthrough efficiency and 3′-UTR length is closer to zero in the absence of Pab1 than those observed in the other two strains for mRNAs with UAA as the stop codon, which happens to be the most common stop codon in the yeast transcriptome, but this trend isn’t observed for UAG and UGA (Figure A1.5B). This result suggests that mRNAs with UAG and UGA, although allowing higher readthrough (Figure A1.5C), are less sensitive to the proximity of Pab1 to the stop codon in this readthrough measurement, possibly because termination is slower and rate-limiting, unlike UAA where termination is faster. Since stop codon identity is a more important feature affecting readthrough efficiency, the effect caused by loss of Pab1 is somewhat masked.

Overall, we find that the proximity of Pab1 to the stop codon, as measured by 3′-UTR length, plays a role in readthrough efficiency prediction in combination with other known mRNA features, and the predictive ability of 3′-UTR length is lost in \( pab1\Delta pbp1\Delta \) cells. Because 3′-UTR length was still predictive of readthrough efficiency in cells depleted of functional eRF1 (Mangkalaphiban et al. 2021), the fact that it no longer predicts readthrough of mRNAs in \( pab1\Delta pbp1\Delta \) cells further supports the notion that readthrough occurring in \( pab1\Delta pbp1\Delta \) cells is likely due to loss of Pab1’s direct function in termination rather than reduction in functional release factor levels.
**Figure A1.5.** Pab1 regulates readthrough efficiency in a distance-dependent manner.

**A.** Average feature importance scores (percent increase in mean squared error (%IncMSE)) extracted from 25 random forest models (5-fold cross-validation, repeated 5 times) trained for each strain to predict readthrough efficiency. Higher feature importance score (red) means the prediction error is high when that feature is permuted. As negative controls (NC), each mRNA was assigned arbitrary continuous and discrete values (Random number and Random factor). Features with significant importance (empirical p-value < 0.05 in at least 15 out of 25 models) are represented as bigger tiles. **B.** Spearman’s correlation coefficient of the relationship between readthrough efficiency and 3’-UTR length using all available data (“all”) or split data by stop codon usage. Number and dot size reflect the number of mRNAs in each correlation coefficient. **C-D.** Heatmaps of median readthrough efficiency of mRNA groups, grouped by the identity of the stop codon or the identity of nucleotide at positions near the stop codon (C) or the identity of P-site codon (D), relative to median readthrough efficiency of all mRNAs in the sample. Positive (red) and negative (blue) values indicate that the group has higher and lower readthrough efficiency than the sample median, respectively. Two-sided Wilcoxon’s rank sum test with Benjamini-Hochberg method for multiple-testing correction was used to compare a group’s median readthrough efficiency to the sample median. Significant results (p < 0.05) are represented as bigger tiles.

**Deletion of PAB1 leads to translation initiation defects but has minimal effects on translation efficiency**

Pab1 is thought to promote translation initiation by aiding the association of the 40S ribosomal subunit and the eIF4F complex with the mRNA 5’ cap through a direct interaction with eIF4G (Tarun and Sachs 1996; Borman 2000; Dever et al. 2016; Brambilla et al. 2019). Translation initiation defects may thus be expected when PAB1 is deleted and, consistent with this hypothesis, we observed increased accumulation of ribosomes at the canonical AUG start codon and in the 5’-UTR region in the pab1Δpbp1Δ strain (Figure A1.2A-B).

To determine whether deletion of PAB1 affected initiation rates of all mRNAs equally, we assessed the change in relative translation efficiency (TE) of each
mRNA between *pab1Δpbp1Δ* and *pbp1Δ* strains (Figure A1.6A). Although ~500 mRNAs show substantive increases or decreases in TE, most mRNAs (>90% of the transcriptome) do not show significant changes in relative TE (Figure A1.6A, grey). These results indicate that the absence of Pab1 affects the initiation process of most mRNAs to the same extent, such that the number of ribosomes recovered for a particular mRNA ORF remains proportional to the mRNA level. This observation is consistent with previous reports of human cells depleted of PABPC (Xiang and Bartel 2021; Kajjo et al. 2022). However, we cannot rule out an indirect effect of *PAB1* deletion on translation initiation factor levels. As shown in the analysis of transcriptome and proteome changes in response to *PAB1* deletion, some initiation factors were reduced at both mRNA and protein levels (Figure A1.4). These disproportionate initiation factor levels could potentially cause a global reduction in translation and the appearance of unchanged relative mRNA TE. Notably, among the most reduced initiation factors are the two paralogs of eIF4G (eIF4G1/Tif4631 and eIF4G2/Tif4632), a binding partner of Pab1.
A. Volcano plot of changes in relative translation efficiency (TE) between $pab1\Delta pbp1\Delta$ and $pbp1\Delta$ strains. Orange, purple, and grey dots represent mRNAs with increased (positive log$_2$ fold change, adjusted p-value < 0.05), decreased (negative log$_2$ fold change, adjusted p-value < 0.05), and unchanged TE (adjusted p-value $\geq$ 0.05), respectively, in the $pab1\Delta pbp1\Delta$ strain.

B. Distribution of TE changes between eIF4G depleted (eIF4Gd) cells vs. isogenic WT cells (Zinshteyn et al. 2017) in each mRNA group from A. Two-sided Wilcoxon’s rank sum test with Benjamini-Hochberg method for multiple-testing correction was used to compare values between pairwise groups. Only significant comparisons were reported as the following: (*) p < 0.05, (**) p < 0.01, (***) p < 0.001, (****) p < 0.0001. C. Proportion and number of mRNAs from groups in A that were enriched (positive log$_2$ fold change, FDR < 0.05), depleted (negative log$_2$ fold change, FDR < 0.05), or unchanged (FDR $\geq$ 0.05) in RIP-seq data (Costello et al. 2015). Pairwise $\chi^2$ test with Benjamini-Hochberg method for multiple-testing correction was used to compare between Reference, Up, and Down groups. p < 0.05 in all pairwise comparisons (exact values provided in Table A1.1). For B and C., Reference (Ref.) group includes all mRNAs regardless of TE changes (Up + Down + Unchanged) to recapitulate the general distribution of measured values in the transcriptome.

**Figure A1.6.** Translation initiation defects in response to $PAB1$ deletion may be related to reduction in eIF4G level.
Reduction of initiation factor levels confound the effects of PAB1 deletion on translation efficiency

To further investigate how deletion of PAB1 impacts translation initiation, we focused on 470 completely spliced mRNAs that showed significant increases ("Up") or decreases ("Down") in TE in response to PAB1 deletion and compared them with two published data sets. First, we investigated TE changes obtained from ribosome profiling and RNA-Seq data of cells depleted for eIF4G through a degron system inducible by growth media and temperature shifts (Zinshteyn et al. 2017). We found that mRNAs in the Down group upon PAB1 deletion also have overall lowered TE upon 2 hours of eIF4G depletion compared to the Up group and the general distribution in the transcriptome (Reference "Ref." group) (Figure A1.6B, 2hr). This trend is even observed before depletion and immediately after depletion was initiated (Figure A1.6B, preshift and 0hr), which is unsurprising because the degron strain showed decreased eIF4G levels even in uninduced conditions (Zinshteyn et al. 2017). These results suggested that TE changes observed upon PAB1 deletion may be partially mediated by reduction in eIF4G level in pab1Δpbp1Δ cells. For the second analysis, we utilized RIP-seq data of mRNAs associated with immunoprecipitated (IP’d) TAP-tagged eIF4G and Pab1 (Costello et al. 2015). If our data was indirectly influenced by the reduction of eIF4G level, we would see that mRNAs enriched in IP of eIF4G, which was hypothesized to be more dependent on eIF4G for translation initiation, would be most sensitive to eIF4G reduction – they would have decreased TE and therefore be found in our
Down group. Indeed, we found that mRNAs that were enriched in IPs of eIF4G1 or eIF4G2 are over-represented in the Down group compared to the Reference group (Figure A1.6C, red, and Table A1.1). Moreover, mRNAs that were depleted in IPs of eIF4G1 or eIF4G2, which were hypothesized to rely less on cap-dependent initiation, are over-represented in the Up group and under-represented in the Down group (Figure A1.6C, blue, and Table A1.1). Notably, this trend is stronger than what is observed in Pab1 IP (Figure A1.6C, and Table A1.1). Overall, these results indicate that reduction in eIF4G levels cannot be ruled out as a possible explanation for the observed changes in TE when PAB1 is deleted.

Despite the apparent effect of eIF4G on different mRNA subgroups, eIF4G still does not explain the entirety of the data. For example, although the mRNAs enriched in eIF4G2 IP are proportionally over-represented in the Down group compared to Reference (35% vs. 18%), 65% of the mRNAs in the Down group are those not highly dependent on eIF4G (Figure A1.6C). Thus, we further characterized these mRNAs by exploring their 5’-UTR features.

First, we observed that mRNAs in the Down group tend to have longer 5’-UTRs than those in the Up or Reference groups (Figure A1.7A). A longer 5’-UTR increases the chance for motifs or upstream ORFs (uORFs) that could interfere with initiation complex assembly or the scanning mechanism. The presence of uORFs is generally thought to suppress translation initiation of the main ORF (Hinnebusch et al. 2016). We found that the proportion of mRNAs with at least one uORF is higher in the Down group than in other groups (Figure A1.7B, left panel,
and Table A1.1). We further asked if the results were due to uORFs that are completely upstream or uORFs that are overlapping with the main ORF by conducting the same analysis for the two types of uORFs separately. The presence of overlapping uORFs is quite rare in our data set and the proportions are not different between groups, but the results for upstream uORFs mimicked those of all uORFs analyzed together (Figure A1.7B and Table A1.1). The sensitivity of uORF-containing mRNAs to PAB1 deletion could be related to the relative reduction in eIF1 (Sui1) level in the \textit{pab1Δpbp1Δ} strain (Figure A1.4). eIF1 has a role in start codon recognition, discriminating against suboptimal start sites in favor of the optimal one (Jackson et al. 2010; Hinnebusch 2014, 2017; Zhou et al. 2020). Therefore, it is not surprising that mRNAs with decreased TE tend to have uORFs (Figure A1.7B). Consistent with this notion, ribosome footprints in the 5’-UTR region are increased in the \textit{pab1Δpbp1Δ} strain relative to the other two strains (Figure A1.2B).

Next, we considered motif-dependent regulation of translation initiation. First, recruitment of Pab1 to poly(A) tracts in the 5’-UTR was reported to induce internal cap-independent initiation in yeast (Gilbert et al. 2007). Therefore, mRNAs with this motif would be expected to have decreased TE in the absence of Pab1 and thus be found more often in our Down group. However, we did not find differences in the proportions of mRNAs containing poly(A) tracts in the 5’-UTR between groups (Figure A1.7C and Table A1.1). Second, oligo(U) longer than 7 nt in the 5’-UTR was identified as eIF4G1’s preferential binding motif and can
promote initiation (Zinshteyn et al. 2017; Niederer et al. 2022). However, we did not find differences in the proportions of mRNAs containing oligo(U) in the 5'-UTR between groups (Figure A1.7D and Table A1.1). In a parallel approach, we used the motif discovery tool STREME (Bailey 2021) to identify motif(s) enriched in 5'-UTR sequences of either Up and Down group relative to the Reference. Neither poly(A) tract nor oligo(U) is enriched in either group compared to sequences from the Reference group, consistent with our direct analyses (Figure A1.7C-D), but the AUG motif is enriched in the Down group (Figure A1.7E), consistent with our uORF analysis (Figure A1.7B). No other novel motif was identified.
Figure A1.7. 5’-UTR features of mRNAs with differential changes in translation efficiency.

A. Distribution of 5’-UTR length in each mRNA group from Fig. 4A. Two-sided Wilcoxon’s rank sum test with Benjamini-Hochberg method for multiple-testing correction was used to compare values between pairwise groups. Only significant comparisons were reported as the following: (*) p < 0.05, (**) p < 0.01, (***) p < 0.001, (****) p < 0.0001.

B-D. Proportion and number of mRNAs from groups in A with (“Yes”) or without (“No”) uORF (B), poly(A) tract (C), or oligo(U) (D) in the 5’-UTR. Pairwise Fisher’s exact test with Benjamini-Hochberg method for multiple-testing correction was used to compare between Reference, Up, and Down groups. p-values are provided in S2 Table.

E. Motif enriched in 5’-UTR sequences of mRNAs in the Down group relative to Reference, identified by STREME. For all panels, analyses were limited to mRNAs with existing UTR annotations. Reference (Ref.) group includes all mRNAs regardless of TE changes (Up + Down + Unchanged) to recapitulate the general distribution of measured values in the transcriptome.
We also asked whether nucleotide context around the start codon and near the 5’ cap influence changes in TE in response to PAB1 deletion by comparing proportions of nucleotides at each position from each group relative to Reference (Figure A1.8). The optimal context for translation in yeast has been determined as AA(A/G)AAUGUCU, with position -3 (3rd nucleotide upstream of AUG, which is considered as positions +1 +2 +3) being the most important and conserved (Hamilton et al. 1987; Hinnebusch et al. 2016; Li et al. 2022). We did not find mRNAs in the Up or Down group to have biases in nucleotide usage at position -3 compared to Reference or each other (Figure A1.8A). However, other positions in this window that show significant differences relative to the Reference are consistent with the consensus, namely, A/C are enriched in the Up group at position -4 (Figure A1.8A) and U is depleted in the Down group at position +4 (Figure A1.8B). Beyond the immediate AUG context, G is enriched and U is depleted in the Down group at position +18 (Figure A1.8B). From the 5’ cap, U is enriched and A is depleted in the Down group at position +6 (Figure A1.8C).

Overall, our results show that studying Pab1’s role on global translation initiation through deleting or depleting Pab1 can be confounded by the reduction in initiation factor levels, especially eIF4G and eIF1.
Figure A1.8. Influences of start codon context in changes of translation efficiency (TE).

A-B. Relative proportions of nucleotide usage upstream (A) and downstream (B) of main ORF’s AUG (positions +1 +2 +3) in Up and Down groups relative to Reference. C. Relative proportions of nucleotide usage from the mRNA 5’ cap (first 18 nucleotides of the 5’-UTR sequences) in Up and Down groups relative to Reference. In all panels, Reference (Ref.) group includes all mRNAs regardless of TE changes (Up + Down + Unchanged) to recapitulate the general proportions in the transcriptome. Positive (red) and negative (blue) log_2 relative proportion indicates that the nucleotide is over-represented and under-represented, respectively, in the group compared to the Reference. Pairwise χ^2 test with Benjamini-Hochberg method for multiple-testing correction was used to compare the nucleotide frequencies between Reference, Up, and Down groups. p < 0.05 for Up or Down vs. Reference is represented by a big tile, while non-significant results are represented by a small Tile. p < 0.05 for Up vs. Down is represented by an asterisk (“*”).
Properties of the mRNAs with differential TE in response to \textit{PAB1} deletion support the notion that mRNA 5’ and 3’ ends communicate

Changes in an mRNA’s poly(A) tail length can change the extent of its commitment to translation initiation, observations which led to the model of 5’-3’ communication of mRNA ends in translation (Jacobson and Favreau 1983; Palatnik et al. 1984; Munroe and Jacobson 1990a, 1990b). Pab1 is thought to play an important role in facilitating this closed-loop mRNA structure, bridging the interaction with both the poly(A) tail and eIF4G (Wells et al. 1998; Borman 2000; Amrani et al. 2008). Shorter mRNAs are thought to form more stable structures than longer mRNAs (Amrani et al. 2006, 2008), but the closed-loop structure may not apply to every mRNA as not all mRNPs contain the closed-loop components (Vicens et al. 2018; Costello et al. 2015; Kershaw et al. 2023) and for those enriched for closed-loop components, they have variable translation efficiency, not just high efficiency (Kershaw et al. 2023).

Although it is uncertain on which mRNAs the closed-loop occurs, the possibility and efficiency of 5’-3’ communication may still be explained through simple proximity of 5’ and 3’ ends (Vicens et al. 2018). Efficient 5’-3’ communication allows efficient feedback of ribosomes recycled from termination to a new round of initiation, and this efficiency should be gene length-dependent, as diffusion of ribosomes between the ends would be expected to be more efficient for shorter mRNAs than longer mRNAs (Rogers et al. 2017; Fernandes et al. 2017), even without Pab1 facilitating the closed-loop. This is especially relevant when the
availability of ribosomes is limiting and ribosome recruitment becomes more dependent on recycled ribosomes than the limited free ribosomes (Rogers et al. 2017; Fernandes et al. 2017), which may be the case in our data with the reduction in ribosomal protein levels in \( \text{pab1}\Delta \text{pbp1}\Delta \) cells (Figure A1.4). When comparing CDS and entire mRNA lengths between groups, we found that mRNAs in the Down group tend to be longer than both the Up group and the Reference while those in the Up group tend be shorter (Figure A1.9A-B). These results are consistent with the proximity-based 5'-3' communication model.

Consistent with the notion that efficient recycling of ribosomes at the 3' end promotes efficient translation initiation at the 5' end, we found that mRNAs with increased TE have lower readthrough efficiency (i.e., more efficient termination) while those with decreased TE have higher readthrough efficiency (Figure A1.9C). We limited our analysis to mRNAs with detectable readthrough due to the cyclic nature of translation and the detection limit of readthrough ribosomes. Since the amount of readthrough ribosomes depends on the amount of translation of the CDS, for mRNAs with the same readthrough efficiency, readthrough may not be detectable for mRNAs with lower TE (so readthrough efficiency appears to be zero for them) but remain or become detectable for mRNAs with higher TE. Thus, the fact that the proportion of mRNAs with detectable readthrough in the Down group is lower than the Up group (Figure A1.9D and Table A1.1) does not necessarily mean that readthrough efficiency is lower in the Down group. In sum, we found
that mRNAs with increased TE had lower readthrough efficiency and those with decreased TE had higher readthrough efficiency (Figure A1.9C).

To see if Pab1’s function at termination influences the distinction between Up and Down TE groups in response to \textit{PAB1} deletion, we compared 3’-UTR lengths among groups, but found no significant differences (Figure A1.9E). This result is perhaps not surprising, as 3’-UTR length is not the strongest feature in WT conditions and no longer matters in the \textit{pab1Δpbp1Δ} strain (Figure A1.5A). We next asked whether a specific sequence motif is enriched in either Up or Down group. No differences between groups were detected in terms of the presence of poly(A) tracts in the 3’-UTR region (Figure A1.9F and Table A1.1), which can serve as a binding site of Pab1 in addition to the poly(A) tail, and this result was confirmed by the lack of poly(A) motif in the motif discovery approach, STREME. However, STREME identified the UAKGUA motif enriched in the Up group relative to Reference (Figure A1.9G). The UA…UA sequence may indicate two consecutive strong stop codons, either UAA or UAG, where the second stop codon (although out-of-frame with the first) acts as a fail-safe stop codon in case of failed termination or recycling at the first stop codon. Moreover, the enrichment of G following the first stop codon is consistent with the observation that mRNAs with G at this position had the lowest readthrough efficiency (Figure A1.5C, “+4”). This motif being enriched in mRNAs in the Up group, along with the shorter mRNA length and lower readthrough efficiency of this group, is consistent with the
hypothesis that more efficient ribosome recycling promotes efficient new rounds of translation initiation.

**Figure A1.9. Evidence for communication of mRNA 5’-3’ ends in efficient initiation.**

A-B. Distribution of CDS length (A) or mRNA length (B) in each mRNA group from Fig. 4A. C. Distribution of readthrough efficiency for mRNAs with detectable readthrough in the pab1Δpbp1Δ strain in each group. D. Proportion and number of mRNAs with detectable (“Yes”) or not detectable (“No”) readthrough in the pab1Δpbp1Δ strain. E. Distribution of 3’-UTR length in each mRNA group. F. Proportion and number of mRNAs with (“Yes”) or without (“No”) poly(A) tracts in the 3’-UTR. For A, B, C, and E, two-sided Wilcoxon’s rank sum test with Benjamini-Hochberg method for multiple-testing correction was used to compare values between pairwise groups. Only significant comparisons were reported as the following: (*) p < 0.05, (**) p < 0.01, (***) p < 0.001, (****) p < 0.0001. For D and F, Pairwise Fisher’s exact test with Benjamini-Hochberg method for multiple-testing correction was used to compare between groups. p-values are provided in Supplemental Table S2. G. Motif enriched in the 3’-UTR sequences of mRNAs in the Up group relative to Reference, identified by STREME. For all panels, analyses were limited to mRNAs with existing UTR annotations. Reference (Ref.) group includes all mRNAs regardless of TE changes (Up + Down + Unchanged) to recapitulate the general distribution of measured values in the transcriptome. (Next page)
A. CDS length (nt) 

B. mRNA length (nt) 

C. Readthrough efficiency 

D. Readthrough detected: 

F. poly(A) tract present: 

G. Enriched in 3'-UTR sequences

Enrichment plot shows enriched sequences at positions 1-6 with an enrichment score of 2.0, indicating a significant difference in 3'-UTR sequences between the Up and Ref. conditions, with a p-value of 0.0325.
Table A1.1. Statistical test results of pairwise comparison between Reference, Up, and Down TE groups

<table>
<thead>
<tr>
<th>Number of mRNAs enriched, unchanged, or depleted in RIP-seq experiment from Costello et al. (2015)</th>
<th>Pairwise Chi-square test with Benjamini-Hochberg method for multiple testing correction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>eIF4G1 RIP-seq</strong></td>
<td></td>
</tr>
<tr>
<td>group1</td>
<td>group2</td>
</tr>
<tr>
<td>Reference</td>
<td>Up</td>
</tr>
<tr>
<td>Reference</td>
<td>Down</td>
</tr>
<tr>
<td>Up</td>
<td>Down</td>
</tr>
<tr>
<td><strong>eIF4G2 RIP-seq</strong></td>
<td></td>
</tr>
<tr>
<td>group1</td>
<td>group2</td>
</tr>
<tr>
<td>Reference</td>
<td>Up</td>
</tr>
<tr>
<td>Reference</td>
<td>Down</td>
</tr>
<tr>
<td>Up</td>
<td>Down</td>
</tr>
<tr>
<td><strong>Pab1 RIP-seq</strong></td>
<td></td>
</tr>
<tr>
<td>group1</td>
<td>group2</td>
</tr>
<tr>
<td>Reference</td>
<td>Up</td>
</tr>
<tr>
<td>Reference</td>
<td>Down</td>
</tr>
<tr>
<td>Up</td>
<td>Down</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of mRNAs with poly(A) tracts (i.e. at least 10 consecutive A's, allowing at most 2 mismatches within 10 A's)</th>
<th>Pairwise Fisher's exact test with Benjamini-Hochberg method for multiple testing correction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In the 5'-UTR</strong></td>
<td></td>
</tr>
<tr>
<td>group1</td>
<td>group2</td>
</tr>
<tr>
<td>Reference</td>
<td>Up</td>
</tr>
<tr>
<td>Reference</td>
<td>Down</td>
</tr>
<tr>
<td>Up</td>
<td>Down</td>
</tr>
</tbody>
</table>

| **In the 3'-UTR**                                                                                                  |                                                                                     |
| group1 | group2 | n | estimate | p       | conf.low | con.high | p.adj |
| Reference | Up      | 2688 | 0.994    | 1       | 0.682    | 1.44     | 1     |
| Reference | Down    | 2633 | 0.654    | 0.0978  | 0.383    | 1.09     | 0.274 |
| Up       | Down    | 207  | 0.659    | 0.183   | 0.345    | 1.24     | 0.274 |

<table>
<thead>
<tr>
<th>Number of mRNAs with oligo(U) (i.e. at least 7 consecutive U's, no mismatch allowed)</th>
<th>Pairwise Fisher's exact test with Benjamini-Hochberg method for multiple testing correction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>In the 5'-UTR</strong></td>
<td></td>
</tr>
<tr>
<td>group1</td>
<td>group2</td>
</tr>
<tr>
<td>Reference</td>
<td>Up</td>
</tr>
<tr>
<td>Reference</td>
<td>Down</td>
</tr>
<tr>
<td>Up</td>
<td>Down</td>
</tr>
</tbody>
</table>

| **In the 3'-UTR**                                                                                                  |                                                                                     |
| group1 | group2 | n | estimate | p       | conf.low | con.high | p.adj |
| Reference | Up      | 2688 | 1.3      | 0.214   | 0.83     | 1.99     | 0.642 |
| Reference | Down    | 2633 | 0.948    | 1       | 0.486    | 1.73     | 1     |
| Up       | Down    | 207  | 0.729    | 0.485   | 0.331    | 1.54     | 0.728 |

<table>
<thead>
<tr>
<th>Number of mRNAs with uORF in the 5'-UTR</th>
<th>Pairwise Fisher's exact test with Benjamini-Hochberg method for multiple testing correction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>uORF: Total</strong></td>
<td></td>
</tr>
<tr>
<td>group1</td>
<td>group2</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>Reference</td>
<td>Up</td>
</tr>
<tr>
<td>Reference</td>
<td>Down</td>
</tr>
<tr>
<td>Up</td>
<td>Down</td>
</tr>
</tbody>
</table>

**uORF: completely upstream**

<table>
<thead>
<tr>
<th>group1</th>
<th>group2</th>
<th>n</th>
<th>estimate</th>
<th>p</th>
<th>conf.low</th>
<th>con.high</th>
<th>p.adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>Up</td>
<td>2688</td>
<td>0.721</td>
<td>0.262</td>
<td>0.387</td>
<td>1.26</td>
<td>0.262</td>
</tr>
<tr>
<td>Reference</td>
<td>Down</td>
<td>2633</td>
<td>4.05</td>
<td>3.19E-08</td>
<td>2.45</td>
<td>6.63</td>
<td>9.57E-08</td>
</tr>
<tr>
<td>Up</td>
<td>Down</td>
<td>207</td>
<td>5.57</td>
<td>7.71E-07</td>
<td>2.64</td>
<td>12.2</td>
<td>1.16E-06</td>
</tr>
</tbody>
</table>

**uORF: overlapping with main ORF**

<table>
<thead>
<tr>
<th>group1</th>
<th>group2</th>
<th>n</th>
<th>estimate</th>
<th>p</th>
<th>conf.low</th>
<th>con.high</th>
<th>p.adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>Up</td>
<td>2688</td>
<td>0.696</td>
<td>0.551</td>
<td>0.218</td>
<td>1.71</td>
<td>0.605</td>
</tr>
<tr>
<td>Reference</td>
<td>Down</td>
<td>2633</td>
<td>1.23</td>
<td>0.605</td>
<td>0.383</td>
<td>3.09</td>
<td>0.605</td>
</tr>
<tr>
<td>Up</td>
<td>Down</td>
<td>207</td>
<td>1.77</td>
<td>0.503</td>
<td>0.393</td>
<td>7.97</td>
<td>0.605</td>
</tr>
</tbody>
</table>

**Number of mRNAs with detectable readthrough in pab1Δpbp1Δ strain**

Pairwise Fisher's exact test with Benjamini-Hochberg method for multiple testing correction

<table>
<thead>
<tr>
<th>group1</th>
<th>group2</th>
<th>n</th>
<th>estimate</th>
<th>p</th>
<th>conf.low</th>
<th>con.high</th>
<th>p.adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>Up</td>
<td>2688</td>
<td>1.96</td>
<td>4.22E-04</td>
<td>1.32</td>
<td>2.94</td>
<td>4.22E-04</td>
</tr>
<tr>
<td>Reference</td>
<td>Down</td>
<td>2633</td>
<td>0.397</td>
<td>1.57E-04</td>
<td>0.232</td>
<td>0.66</td>
<td>2.35E-04</td>
</tr>
<tr>
<td>Up</td>
<td>Down</td>
<td>207</td>
<td>0.205</td>
<td>1.39E-07</td>
<td>0.105</td>
<td>0.389</td>
<td>4.17E-07</td>
</tr>
</tbody>
</table>
Substrates of decapping activators Pat1/Lsm1 and Upf1/Upf2/Upf3 tend to be more increased than decreased in response to PAB1 deletion

Among its many functions, PABPC also has important regulatory roles in mRNA decay (Mangus et al. 2004; Brambilla et al. 2019; Passmore and Coller 2022). Hence, we asked how the absence of Pab1 impacts the levels of mRNAs that are substrates of different decapping activators, namely Dhh1, Pat1/Lsm1, and the Upf factors of the NMD pathway. Dhh1, Pat1/Lsm1, and NMD substrates are defined respectively as mRNAs whose levels were increased in dhh1Δ cells, commonly increased in pat1Δ and lsm1Δ cells, and commonly increased in upf1Δ, upf2Δ, and upf3Δ cells, relative to WT (Celik et al. 2017a; He et al. 2018). Approximately half of the mRNAs in each group showed significant changes in abundance in pab1Δpbp1Δ relative to pbp1Δ cells. Of mRNAs that showed significant changes, those that are targets of either Pat1/Lsm1 or NMD or both tend to be increased rather than decreased (Figure A1.10A, top and middle panels, columns 3-8, compare Up (orange) to Down (purple)). On the other hand, mRNAs that are only substrates of Dhh1 are somewhat comparably increased and decreased (Figure A1.10A, top and middle panels, column 2, compare Up (orange) to Down (purple)).

To determine whether increases in the abundance of Pat1/Lsm1 and Dhh1 substrates, which follow the canonical deadenylation-dependent pathway, was due to loss of Pab1’s role in deadenylase recruitment or dysregulation of decay pathway components, we investigated changes in mRNA and protein levels of
genes involved in decapping, deadenylation, and decay (Figure A1.10B). We found that while the decapping enzyme (Dcp1 and Dcp2), Pat1, Lsm1, Dhh1, exonuclease Xrn1, and exosome components are not significantly enriched or depleted at the protein level, two proteins that are components of the Ccr4-Not deadenylase complex, Mot2 (also commonly known as Not4) and Not5, are significantly depleted to ~77-80% of the amount in pbp1Δ cells (Figure A1.10B). Not4 and Not5 are thought to link slow translation elongation of non-optimal codons to deadenylation as well as deadenylation to decapping (Passmore and Coller 2022). Specifically, Dhh1’s association with the ribosome requires Not5’s ribosome binding and Not4’s E3 ligase activity to ubiquitinate 40S ribosomal subunit protein eS7 (Buschauer et al. 2020). Not5 has also been shown to bind Pat1 to promote decapping (Alhusaini and Coller 2016). Thus, stabilization of Pat1/Lsm1 and some Dhh1 substrates may be partially attributable to dysregulated Ccr4-Not components, in addition to loss of Pab1’s role in deadenylase recruitment.

Because decapping of NMD substrates is usually a deadenylation-independent mechanism that is triggered by premature translation termination (Muhrad and Parker 1994; He and Jacobson 2001, 2022), stabilization of NMD substrates in pbp1Δpbp1Δ cells is likely unrelated to Pab1’s role in protecting the poly(A) tail or deadenylase recruitment. Rather, it is likely due to decreases in translation, which leads to decreased frequency of premature termination. To test whether NMD substrates have a decreased initiation rate in the absence of Pab1,
we compared log₂ fold change in TE between \( pab1\Delta pbp1\Delta \) and \( pbp1\Delta \) strains for each group of mRNAs to the value of unchanged TE (log₂ fold change = 0) (Figure A1.10A, bottom panel). Consistent with our hypothesis, we found that NMD substrates that are significantly stabilized or have unchanged mRNA levels in \( pab1\Delta pbp1\Delta \) have relatively lower TE while the minority that are depleted have relatively unchanged TE (Figure A1.10A, bottom panel, column 4). However, reduced TE is probably not the only contribution to NMD substrate enrichment, as the protein level of the key NMD protein Upf1 is significantly reduced in \( pab1\Delta pbp1\Delta \) cells to ~85% of the level in \( pbp1\Delta \) cells (Figure A1.10B).

**Figure A1.10.** Changes in the abundance of decapping activator substrates by different consequences of \( PAB1 \) deletion.

**A.** mRNA abundance and translation efficiency changes between \( pab1\Delta pbp1\Delta \) and \( pbp1\Delta \) strains for Dhh1, Pat/Lsm1, and NMD substrates. **(Top)** Panel indicating substrate status (green) of panels below. A substrate is defined as an mRNA that is upregulated upon decapping deactivator gene deletion (He et al. 2018; He and Jacobson 2022). Dhh1: mRNAs upregulated in a \( dhh1\Delta \) strain relative to WT (He et al. 2018); Pat1/Lsm1: mRNAs commonly upregulated in \( pat1\Delta \) and \( lsm1\Delta \) strains relative to WT (He et al. 2018); NMD: mRNAs commonly upregulated in \( upf1\Delta \), \( upf2\Delta \), and \( upf3\Delta \) strains relative to WT (Celik et al. 2017a). **(Middle)** Proportions of mRNA abundance changes from Figure A1.3A: Up (orange), Down (purple), and Unchanged (grey), separated into columns by substrate status. **(Bottom)** Distribution of log₂ fold change in TE between \( pab1\Delta pbp1\Delta \) and \( pbp1\Delta \) strains for mRNA groups from the middle panel. Two-sided Wilcoxon’s rank sum test with Benjamini-Hochberg method for multiple-testing correction was used to compare median log₂ fold change in TE of zero (no change). Significant levels were reported as the following: (ns) not significant, (*) \( p < 0.05 \), (**) \( p < 0.01 \), (***) \( p < 0.001 \), (****) \( p < 0.0001 \). **B.** Comparison of log₂ fold change in transcriptome and proteome as in Fig 2C, with the focus on factors related to mRNA decay. 

(Next page)
A

<table>
<thead>
<tr>
<th>Substrate</th>
<th>Dhh1</th>
<th>Pat1/Lsm1</th>
<th>NMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 4702</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 521</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 325</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 752</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 428</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 53</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Percentage

<table>
<thead>
<tr>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
</tr>
<tr>
<td>75</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

mRNA abundance change in \( \text{pab1} \Delta \text{pbp1} / \text{pbp1} \Delta \)

### Decapping/5'-3' Decay

- **DCP1**
- **EAP1**
- **LSM1**
- **PAT1**
- **RRP3**

### Deadenylation/3'-5' Decay

- **CCR4**
- **CSA**
- **MTC2**
- **NOT3**

B

mRNA vs. protein changes:

- Neither
- Protein only
- mRNA only
- Both
Discussion

Numerous biochemical, structural, *in vitro*, and *in vivo* studies of specific mRNAs have identified pleiotropic roles for PABPC in cytoplasmic mRNA deadenylation, translation initiation, and translation termination (Passmore and Coller 2022). However, because of PABPC’s multiple apparent functions, defining its transcriptome-wide roles has been difficult. High-throughput approaches exploring transcriptome-wide effects of PABPC depletion have been carried out in mammalian cells (Xiang and Bartel 2021; Kajjo et al. 2022), but none have been done in yeast. Consistent with observations in mammalian cells, we found that deletion of yeast PAB1 resulted in major changes in the transcriptome (Figure 1.3A), but only minimal changes in relative translation efficiency (Figure A1.6A). Further, we showed that pab1Δ cells also drastically changed their proteome (Figure 1.3B) and provided the first evidence for a transcriptome-wide translation termination defect (Figure A1.2).

Our proteomics data have provided insights to the direct vs. indirect consequences of PAB1 deletion and suggested a new layer of complexity to interpreting genome-wide gene expression alterations in PABPC-depleted cells. PABPC’s role in translation initiation has been elusive, partly because the extent of its activity is dependent on the stoichiometry between PABPC, poly(A) tracts, and basal translation levels (Xiang and Bartel 2021), and these experimental conditions frequently vary between studies. Moreover, Pab1 appears to have preferential association with certain mRNAs (Costello et al. 2015). Hence, it seems
counterintuitive that depleting PABPC/deleting \textit{PAB1} reduces translation overall, yet relative TE is unchanged for most mRNAs (Xiang and Bartel 2021; Kajjo et al. 2022). These observations are akin to the effects of depleting eIF4G in yeast (Park et al. 2011). Since \textit{PAB1} deletion also reduced eIF4G mRNA and protein levels (Figure A1.4), it is possible that effects attributed to the absence of Pab1 are at least partially due to dysregulation of eIF4G mRNA stability and the consequent reduction of eIF4G, which in turn reduced global translation initiation. This sequence of events is also likely for ribosomal proteins, as PABPC depletion has been shown to cause accelerated decay of mRNAs with short poly(A) tails (Xiang and Bartel 2021), which usually are characteristics of highly expressed, highly translated mRNAs, including those encoding ribosomal proteins (Lima et al. 2017; Passmore and Coller 2022). Even when we focused our analyses on mRNAs that did have significant changes in TE, where their drastic changes may be due to specific factors outside of the global regulators of translation initiation, their properties are still related to eIF4G-dependent pre-initiation complex recruitment, eIF1-mediated start codon recognition, and efficiency of ribosome recycling (Figures A1.6, A1.7, A1.9). Due to their intrinsic properties, these mRNAs with significant increases or decreases in TE are respectively more or less dependent on these processes than most mRNAs and are thus more sensitive to reduction in these factors in \textit{pab1Δ} cells (Figure A1.4). As a result, the direct role of Pab1 in initiation is masked by the changes in these core components of initiation, i.e.,
reduction of initiation factor levels cannot be completely ruled out as an explanation for the translation initiation defects observed in \( pab1\Delta \) cells.

Similar reasoning might be applied to translation termination since there is a slight reduction in release factor levels in \( pab1\Delta \) cells (Figure A1.4). The increased stop codon readthrough observed in \( pab1\Delta \) cells is unlikely to be due to the release factor level becoming limited, since reduced initiation factors and ribosomal proteins are more limiting than release factors, skewing the usual stoichiometry of demand vs. supply for release factors towards the supply, but this observation is likely due to a shift in stoichiometry of eRF1/eRF3 vs. aa-tRNA/eEF1A towards aa-tRNA/eEF1A, promoting readthrough. However, Pab1’s direct role in termination is still implied because i) 3’-UTR length, which approximates the distance of Pab1 to the stop codon, lost its ability to predict readthrough efficiency in \( pab1\Delta \) cells as opposed to WT (Figure A1.5A) or eRF1 mutant cells (Mangkalaphiban et al. 2021) (i.e., 3’-UTR length should have still been predictive of readthrough efficiency if increased readthrough was a consequence of release factor depletion alone), and ii) deletion of Pab1’s eRF3 interacting domain only (\( pab1\Delta C \)), which affects termination but not mRNA decay (Roque et al. 2015), yields significant stop codon readthrough of reporter mRNAs (Wu et al. 2020). Nevertheless, it remains to be determined whether a 20% reduction in termination factors in a \( PAB1 \) strain would yield the same termination defect as a \( pab1\Delta \) mutation and whether \( pab1\Delta C \) cells significantly change their release factor levels. Because near cognate tRNAs compete with release factors
in stop codon decoding, changes in aminoacyl-tRNA levels, synthesis, and modifications should also be considered.

Since deletion of *PAB1* in yeast is lethal, we also deleted *PBP1*, which suppresses this lethality. Surprisingly, single deletion *pbp1Δ* cells showed an apparent increase in readthrough efficiency compared to WT cells (Figure 1.2D), raising the question of whether Pbp1 plays a role in translation termination. However, this is unlikely for several reasons. First, we noticed that the amount of footprints in the 3'-UTR from *pbp1Δ* cells is only relatively higher than WT in the first few codons after the stop codon, and from only one of the three biological replicates, compared to *pab1Δpbp1Δ* cells in which the higher amount of footprints sustains throughout the 80-nt window (Figure A1.2A). This observation is clearer through quantifications (Figure A1.2B). Second, footprints in the extension region from *pbp1Δ* cells do not show frame 0 preference as do those from *pab1Δpbp1Δ* cells (Figure A1.2C). Despite the lack of frame 0 preference, these footprints unfortunately cannot be selectively discarded for readthrough efficiency calculation, defined as frame 0 footprint density in the extension relative to that in the CDS. Thus, by this definition, the calculation results in an apparent increase in readthrough efficiency in *pbp1Δ* cells (Figure A1.2D). It is important to note that the same apparent “increase” in “readthrough efficiency” was also observed in cells lacking recycling factor Rli1 (Mangkalaphiban et al. 2021), which caused random reinitiation downstream of stop codons and not readthrough (Young et al. 2015).
We showed that the sets of mRNA substrates of Pat1/Lsm1-mediated decay and NMD both manifest relative increases in abundance and are thus likely to be partially stabilized in \( pab1\Delta \) cells despite their different decay mechanisms, while substrates of Dhh1-mediated decay did not have a tendency to increase or decrease (Figure A1.10A). Stabilization of Pat1/Lsm1 substrates may be attributed to the loss of Pab1’s role in stimulating deadenylation of mRNAs usually subject to deadenylation-dependent decay. However, there is also a slight depletion of Ccr4-Not components, Not4 and Not5 (Figure A1.10B), and the extent of how much this depletion contributes to substrate stabilization is unknown. The consequences of Not4 and Not5 depletion may also apply to some Dhh1 substrates (i.e., those that are increased) but not all, possibly due to Dhh1’s involvement in multiple decapping complexes which can lead to alternative degradation pathways (He and Jacobson 2022). Some complexes may require Pab1’s role in deadenylase recruitment while for others, Dhh1’s communication with the Ccr4-Not complex may be enhanced by lower Pab1 level (Webster et al. 2018). In contrast to Pat1/Lsm1 substrates, partial increases in the levels of NMD substrates most likely arises from the combination of decreased translation and decreased Upf1. The reduction in overall translation would result in fewer instances of termination and, together with the slight reduction in Upf1, would lower the likelihood that a nonsense-containing mRNA would be targeted by NMD.

Collectively, our observations of the indirect effects of \( PAB1 \) deletion may help explain the discrepancies in previous studies of PABPC’s functions and our
multi-omics data can be helpful resources for the design of future experiments involving genetic manipulation, depletion, or overexpression of PABPC.

**Supplemental Discussion**

Due to the relative increase in accumulation of ribosomes over start codons upon $PAB1$ deletion (Figure A1.2A-B), we wondered if any *cis*-regulatory elements of transcriptome-wide AUG occupancy responded to $PAB1$ deletion. We first defined AUG occupancy as the number of ribosome profiling footprints with the canonical AUG occupying their P-sites in frame 0 normalized to the mRNA level from Ribo-Seq experiments ($\log_2$-transformed). We then trained a random forest regression model to predict AUG occupancy using identities of nucleotides near the start codon, features of uORFs, features of the 5'-UTRs, and mRNA region lengths. Strikingly, random forest revealed that the identities of the nt -12 (nt 12 positions upstream of AUG) and a few nearby nucleotides as well as CDS length are important predictors of AUG occupancy (Figure A1.11A).

Analyses of nucleotide identities (Figure A1.11B-C) showed that mRNAs with G at positions -13 to -8 tend to have lower AUG occupancy than average while those with U have higher AUG occupancy than average (Figure A1.11B). Additionally, mRNAs with C at position -12 tend to have higher AUG occupancy (Figure A1.11B). The trends are the same in all yeast strains, hence these nucleotides are likely unrelated to Pab1 but general regulators of AUG occupancy. We further asked where these nucleotides reside in the ribosomes occupying AUG. Since the sudden spike in footprint count near the start of the mRNA ORF usually
indicates initiating ribosomes with AUG in their P-sites, the locations of footprint 5' and 3' ends will determine how many nucleotides were protected by the ribosomes (i.e., reside in the mRNA channel) during the RNaseI digestion step of sequencing library preparation. The quantities of mapped footprints' 5' and 3' ends at each position relative to the AUG codon showed that high amount of footprints' 5' ends starts at -13, indicating that nt -13 to -8 reside in the ribosome's mRNA channel (Figure A1.12).

However, a concern was raised regarding whether this observation is an artifact of library preparation. Enzymes used during library preparation to ligate footprints to adapter sequences or to digest unprotected RNAs can have differential activities among the four nucleotides, lead to over- or under-representation of footprints, and can affect analyses of codon occupancy (Mok et al. 2023). Although our protocol involved randomized 4N at the adapter ligation site to reduce ligation biases and used RNaseI, which displays no nucleotide preference and yields the highest quality footprints for yeast (Delcardayre and Raines 1995; Gerashchenko and Gladyshev 2017), we cannot completely rule out these biases specifically. Since AUG occupancy calculation only involved ribosome occupancy at one position, the bias may be stronger than a collective quantification across the ORF. Thus, we used changes in relative TE to investigate translation initiation (Figures A1.6-A1.9). Comparison of TE of the same mRNA, i.e., identical sequence, in different yeast strains control for sequence-based biases that we might have in our libraries.
**Figure A1.11.** mRNA features determining AUG occupancy in different yeast strains.

A. Average feature importance scores (percent increase in mean squared error (%IncMSE)) extracted from 25 random forest models (5-fold cross-validation, repeated 5 times) trained for each strain to predict AUG occupancy. Higher feature importance score (red) means the prediction error is high when that feature is permuted. As negative controls (NC), each mRNA was assigned arbitrary continuous and discrete values (Random number and Random factor). Features with significant importance (empirical p-value < 0.05 in at least 15 out of 25 models) are represented as bigger tiles. B-C. Heatmaps of median AUG occupancy of mRNA groups, grouped by the identity of the nucleotides upstream (B) or downstream (C) of the start codon, relative to median AUG occupancy of all mRNAs in the sample. Positive (red) and negative (blue) values indicate that the group has higher and lower AUG occupancy than the sample median, respectively. Two-sided Wilcoxon’s rank sum test with Benjamini-Hochberg method for multiple-testing correction was used to compare a group’s median readthrough efficiency to the sample median. Significant results (p < 0.05) are represented as bigger tiles.
Figure A1.12. Nucleotides determinant of AUG occupancy reside in the ribosome’s mRNA channel.
Mapping of the 5’ and 3’ ends of footprint lengths 18–34 nt relative to the start codon (red dashed line). For each footprint length, footprint count at each nucleotide position relative to the start codon were normalized to the total footprint count in the library.
We also explored the relationship between CDS length and AUG occupancy. We found a slightly negative but significant correlation between CDS length and AUG occupancy in WT and \( pbp1\Delta \) but not \( pab1\Delta pbp1\Delta \) cells (Figure A1.13A). Since the CDS region usually makes up almost the entire mRNA, we repeated the same analysis with mRNA length and found the same results (Figure A1.13B). These results indicate that shorter mRNAs tend to have higher AUG occupancy than longer mRNAs in the presence of Pab1, consistent with the notion that short mRNAs may be able to form more stable Pab1-dependent closed-loop and have more efficient ribosome recycled to new rounds of initiation (Amrani et al. 2008). However, because the correlation is quite weak, the closed-loop likely does not apply to all mRNAs, consistent with previous studies showing that some highly translated mRNAs were depleted of closed-loop components (Costello et al. 2015; Vicens et al. 2018; Kershaw et al. 2023).
Figure A1.13. The inverse relationship between AUG occupancy and mRNA length is abolished in the absence of Pab1.

A-B. Scatter plot of AUG occupancy vs. CDS length (A) or mRNA length (B). Spearman’s correlation coefficient was calculated and reported for each sample.
Materials and Methods

Yeast strain construction

Yeast strains used in this study, listed in Table A1.2, are in the W303 background. Gene deletions were achieved by the PCR-based method (Longtine et al. 1998) and high efficiency transformation (Schiestl and Gietz 1989) of fragments amplified by oligonucleotides listed in Table A1.3, synthesized by Integrated DNA Technologies (IDT). Three biological replicates (isolates) of each mutant strain and three technical replicates of isogenic wild-type (WT) strain were used for all experiments.

Table A1.2. Yeast strains used in this study.

<table>
<thead>
<tr>
<th>Strain</th>
<th>Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>HFY114</td>
<td>MATa ura3-1 leu2-3,112 his3-11,15 trp1-1 ade2-1 can1-100</td>
</tr>
<tr>
<td>KMY01, KMY02, KMY06</td>
<td>MATa ura3-1 leu2-3,112 his3-11,15 trp1-1 ade2-1 can1-100 pbp1::URA3</td>
</tr>
<tr>
<td>KMY07, KMY10, KMY11</td>
<td>MATa ura3-1 leu2-3,112 his3-11,15 trp1-1 ade2-1 can1-100 pbp1::URA3 pab1::kanMX</td>
</tr>
</tbody>
</table>

Table A1.3. Oligonucleotides used in this study.

<table>
<thead>
<tr>
<th>Oligonucleotide</th>
<th>Sequence (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBP1_5F (NotI)</td>
<td>TAAGCAGCGGCGCCATTTGTTCAGAGTTACGATTAGA</td>
</tr>
<tr>
<td>PBP1_5R_v2 (BamHI)</td>
<td>TCGGATCGGATCCATTGTTTTCAATGTAATGAACCTCTT</td>
</tr>
<tr>
<td>PBP1_3F_v2 (EcoRI)</td>
<td>ATCGATGAATTCTGAAGGAAGCACGTCAGAAAAAGC</td>
</tr>
<tr>
<td>PBP1_3R (Sall)</td>
<td>TGCTTAGTCGACATTTGCTCTATTTCACAGACCTCGA</td>
</tr>
<tr>
<td>URA3_5F_v2 (BamHI)</td>
<td>GAAACAATGGATCCGATCCGAGCTTTCAATTCATC</td>
</tr>
<tr>
<td>URA3_3R_v2 (EcoRI)</td>
<td>CCTTCGAAATTTCACGATCCAGTTTCAATTTCATC</td>
</tr>
<tr>
<td>PBP1_start_F</td>
<td>TCCGAAATTTCAGAAGGGGAAGC</td>
</tr>
</tbody>
</table>
The *pbp1Δ* strain was made by replacing the coding sequence of the *PBP1* gene in a WT strain (HFY114) with the *URA3* gene. The *URA3* cassette was obtained from plasmid HFSE1380 (He et al. 2018) by PCR using *URA3_5F_v2* and *URA3_3R_v2* primers. Homology arms flanking the *PBP1* coding sequence were amplified from genomic DNA by PCR using *PBP1_5F*, *PBP1_5R_v2*, *PBP1_3F_v2*, and *PBP1_3R* primers. DNA fragments consisting of the homology arms flanking the *URA3* cassette were constructed by PCR and transformed into competent WT yeast cells. Verification of successful gene replacement was
confirmed by PCR followed by Sanger sequencing using primers listed in Table A1.3.

Subsequent deletion of the PAB1 gene from the pbp1Δ strain was carried out by replacing the coding sequence of PAB1 with kanMX, resulting in a pab1Δpbp1Δ strain. The kanMX cassette was obtained from the pCAS plasmid (Addgene, #60847) by PCR using PAB1_KanMX_F and PAB1_KanMX_R primers. Homology arms flanking the PAB1 coding sequence were amplified from genomic DNA by PCR using PAB1_5H_ext_F2, PAB1_5H_ext_R, PAB1_3H_ext_F, and PAB1_3H_ext_R primers. DNA fragments consisting of the homology arms flanking the kanMX cassette were constructed by PCR and transformed into competent pbp1Δ (KMY01) yeast cells. Verification of successful gene replacement was confirmed by PCR followed by DNA sequencing using primers listed in Supplemental Table S4 as well as by examining growth phenotype. Doubling time for these strains grown in YEPD at 30°C were 1.5 hours for WT and pbp1Δ strains and 3.5-4.5 hours for the pab1Δpbp1Δ strain.

**Cell growth and harvest**

Cells were grown in 1 L of YEPD at 30°C with shaking. When the OD$_{600}$ of the culture reached 0.6–0.8, cells were collected by rapid vacuum filtration, flash-frozen in liquid nitrogen in the presence of Footprinting Buffer (20 mM Tris-HCl pH 7.4, 150 mM NaCl, 5 mM MgCl$_2$) plus 1% TritonX-100, 0.5 mM DTT, 1 mM phenylmethylsulfonyl fluoride (PMSF), and 1X protease inhibitors, and lysed in a Cryomill (Retsch) (5 Hz, 2 min; 10 Hz, 15 min). Cell lysates were clarified by
ultracentrifugation in a Beckman Coulter Optima L-90K Ultracentrifuge at 18,000 rpm for 10 min at 4°C, using a 50Ti rotor. Centrifugation was repeated for the supernatant at 18,000 rpm for 15 min at 4°C. Lysates were stored at -80°C in aliquots.

**Ribosome profiling library preparation and sequencing**

Ribosome profiling libraries were prepared as described previously (Ganesan et al. 2022). Lysates were digested with RNase I (Invitrogen, #AM2294) for 1 hour at 25°C with shaking at 700 rpm, and the reaction was stopped using SUPERase-In RNase Inhibitor (Invitrogen, #AM2694). RNase I-treated lysates were then layered onto a 1 M sucrose cushion in Footprinting Buffer plus 0.5 mM DTT and centrifuged in a Beckman Optima TLX Ultracentrifuge at 60,000 rpm for 1 hour at 4°C using a TLA100.3 rotor to isolate 80S ribosomes. Ribosome-protected fragments (RPFs) were extracted from pelleted 80S ribosomes using a miRNeasy kit (QIAGen, #217004) following the manufacturer’s protocol for enriched recovery of small RNAs (<200 nt). RNAs larger than 200 nt, which include ribosomal RNAs (rRNAs) and other large RNAs, were discarded. Small RNAs were 3’ dephosphorylated and 5’ phosphorylated with T4 polynucleotide kinase (T4PNK, NEB, #M0201S), and purified with RNA Clean and Concentrator-5 (Zymo Research, #R1013) according to the manufacturer’s instructions that separately recover small and large RNA fractions. Large RNA fractions were discarded. Approximately 1 μg of small RNAs in 8.5 μl were incubated with 2 μl of QIAseq FastSelect –rRNA Yeast Kit (Qiagen, #334215) at 75°C, 2 min; 70°C, 2 min; 65°C,
2 min; 60°C, 2 min; 55°C, 2 min; 37°C, 2 min; 25°C, 2 min; 4°C, hold. This step hybridized any remaining rRNAs in the sample to the rRNA oligonucleotides, creating duplexes that would fail to ligate to sequencing adapters or fail to be reversed-transcribed into cDNA for sequencing library preparation. Sequencing libraries were prepared from the 10.5 μl reactions using the NEXTflex Small RNA-Seq Kit v3 (Perkin Elmer/Bioo Scientific, #NOVA-512-05) according to the manufacturer’s protocol, except for the RNA denaturation step (70°C, 2 min incubation) before 3’ 4N Adenylated Adapter ligation, which was skipped. Based on the manufacturer’s instructions for optimization, adapters were undiluted, PCR was performed for 15 cycles, and the library was purified using the manufacturer’s gel-free size-selection and cleanup protocol. Extra rounds of cleanup were performed if the amount of PCR primers was still high compared to the amount of library, as analyzed on a Fragment Analyzer. Three libraries were multiplexed according to NEXTflex’s recommended combinations of barcodes (index sequences) and sequenced (single-end, 75 cycles) in-house on Illumina NextSeq 500 or NextSeq 550 sequencers.

**RNA-Seq library preparation and sequencing**

RNA-Seq libraries were prepared and sequenced as described previously (Ganesan et al. 2022). Briefly, total RNAs were extracted from lysates using a miRNeasy kit (QIAgen, #217004) following the manufacturer’s protocol for recovery of total RNAs (standard protocol). Genomic DNA contamination was depleted using Baseline-Zero DNase (Lucigen/Epicentre, #DB0715K) according
to manufacturer’s instructions. Approximately 1 μg of DNase-treated RNAs were used to prepare a sequencing library. The rRNA depletion strategy using QIAseq FastSelect –rRNA Yeast Kit (Qiagen, #334215) was integrated into the RNA fragmentation step of the TruSeq Stranded mRNA Library Prep kit (Illumina, #20020594) according to the QIAseq FastSelect’s manual. Three libraries were multiplexed using recommended combinations of TruSeq RNA Single Indexes Set A (Illumina, #20020492) and sequenced (single-end, 75 cycles) on an Illumina NextSeq 500 sequencer.

Sequence alignment

The yeast transcriptome used for sequence alignment was from https://github.com/Jacobson-Lab/yeast_transcriptome_v5, the generation of which was described previously (Mangkalaphiban et al. 2021). Reads pre-processing, alignment, and quantification were performed on the University of Massachusetts Green High Performance Computing Cluster using the following provided software packages: cutadapt v1.9, bowtie v1.0.0, fastqc v0.10.1, samtools v0.1.19, bedtools v2.26.0, UMI-tools v1.1.1, and RSEM v1.3.0.

Ribosome profiling reads were pre-processed, aligned to the transcriptome, and transcript abundance quantified as described previously, except for the PCR duplicate removal step, where the UMI-tools software package was employed (Smith et al. 2017). The UMI-tools’ “extract” function was used to record four nucleotides at each end of a read, which were introduced during library preparation by the NEXTflex Small RNA-Seq Kit v3. UMI-tools’ “dedup” function with the default
(“directional”) method was used to identify and remove PCR duplicates based on the extracted UMIs. Number of reads processed, number of PCR duplicates removed, number of remaining unique reads, and other relevant sequencing statistics for each library are provided in Table A1.4.

**Table A1.4. Ribosome profiling reads processing and alignment statistics.**

<table>
<thead>
<tr>
<th>Strain</th>
<th>Genotype</th>
<th>Total reads processed</th>
<th>Reads ≥ 15 bp &amp; contained adapters</th>
<th>Reads not aligned to ncRNAs*</th>
<th>PCR duplicates</th>
<th>Reads failed to align to transcriptome</th>
<th>Unique reads aligned to transcriptome</th>
</tr>
</thead>
<tbody>
<tr>
<td>HFY114_1</td>
<td>WT</td>
<td>80,205,117</td>
<td>76,029,629</td>
<td>49,238,479</td>
<td>34,771,961</td>
<td>1,567,309</td>
<td>12,899,209</td>
</tr>
<tr>
<td>HFY114_2</td>
<td>WT</td>
<td>92,678,801</td>
<td>88,267,878</td>
<td>53,454,200</td>
<td>37,419,529</td>
<td>1,964,428</td>
<td>14,070,243</td>
</tr>
<tr>
<td>HFY114_3</td>
<td>WT</td>
<td>134,083,694</td>
<td>127,571,813</td>
<td>75,311,898</td>
<td>57,900,962</td>
<td>2,822,287</td>
<td>14,588,649</td>
</tr>
<tr>
<td>KMY01</td>
<td><em>pbp1Δ</em></td>
<td>128,966,425</td>
<td>120,181,076</td>
<td>71,016,854</td>
<td>51,831,713</td>
<td>2,358,236</td>
<td>16,826,905</td>
</tr>
<tr>
<td>KMY02</td>
<td><em>pbp1Δ</em></td>
<td>135,671,151</td>
<td>128,475,480</td>
<td>79,067,635</td>
<td>57,741,459</td>
<td>2,983,491</td>
<td>18,342,685</td>
</tr>
<tr>
<td>KMY06</td>
<td><em>pbp1Δ</em></td>
<td>130,383,470</td>
<td>121,385,659</td>
<td>70,796,308</td>
<td>52,119,824</td>
<td>2,756,169</td>
<td>15,920,315</td>
</tr>
<tr>
<td>KMY07</td>
<td><em>pab1Δpbp1Δ</em></td>
<td>83,358,755</td>
<td>78,651,457</td>
<td>43,957,590</td>
<td>27,001,278</td>
<td>2,566,212</td>
<td>14,390,100</td>
</tr>
<tr>
<td>KMY10</td>
<td><em>pab1Δpbp1Δ</em></td>
<td>133,625,820</td>
<td>125,135,403</td>
<td>76,328,094</td>
<td>47,733,967</td>
<td>9,988,438</td>
<td>18,605,689</td>
</tr>
<tr>
<td>KMY11</td>
<td><em>pab1Δpbp1Δ</em></td>
<td>97,744,225</td>
<td>90,739,165</td>
<td>38,799,378</td>
<td>23,235,271</td>
<td>7,113,924</td>
<td>8,450,183</td>
</tr>
</tbody>
</table>

RNA-Seq reads were aligned to the transcriptome and transcript abundance quantified using RSEM without any pre-processing.

**Mass spectrometry (LC-MS/MS)**

**Sample preparation**

Protein concentrations in cell lysates were determined by Pierce BCA Protein Assay according to the manufacturer’s protocol (Thermo Scientific). Aliquots of cell lysates containing 50 µg total protein were snap frozen, lyophilized in a SpeedVac, then reduced, alkylated, and digested following the S-Trap digestion protocol (ProtiFi). In brief, lyophilized lysates were first resuspended in 23 µl Lysis buffer (5% SDS in 50mM Triethyl ammonium bicarbonate (TEAB)). Resuspended protein extracts were reduced by adding 1 µl 200mM TCEP and
incubating at 55 °C for 1 hour, then alkylated by adding 1 µl 375mM iodoacetamide (IAA) and incubating at room temperature for 30 minutes, protected from light. To further denature and trap proteins, samples were mixed with 2.5 µl of 27.5% phosphoric acid (H₃PO₄ in water) and 165 µl of binding/wash buffer (100mM TEAB in 90% methanol). The mixtures were applied to S-Trap columns and centrifuged at 4,000 g for 30 seconds. Columns were washed 5 times, each by 150 µl of binding/wash buffer and centrifugation at 4,000 g for 30 seconds. To digest proteins, 25 µl of digestion buffer containing 1 µg trypsin (5 µl of 0.2 µg/µl trypsin in 50mM TEAB + 20 µl 50mM TEAB) was added to each column and samples were incubated at 37 °C overnight. Digested peptides were collected by 3 subsequent centrifugations at 4,000 rpm for 1 min following the addition of these elution buffers for each collection: 1) 40 µl 50mM TEAB in water, 2) 40 µl 0.2% formic acid in water, and 3) 40 µl 50% acetonitrile in water. All flowthroughs from the same sample were pooled, lyophilized in a SpeedVac, and stored at -80 °C.

Samples were labeled using a TMT10plex labeling kit (Thermo Scientific). Lyophilized, digested peptides were resuspended in 50 µl 100mM TEAB and incubated with 10 µl of TMT10plex reagent (equilibrated to room temperature and resuspended in acetonitrile) at room temperature for 1 hour. Reactions were quenched with 5 µl of 5% hydroxylamine at room temperature for 15 minutes. Equal amounts (55 µl) of each sample were pooled together; 50 µl of the pooled reaction was saved for direct shotgun analysis and the rest for high-pH fractionation. For the latter, pooled samples were dried in a SpeedVac,
resuspended in 300 µl of 0.1% trifluoroacetic acid (TFA) in water, and fractionated using a Pierce High pH Reversed-Phase Peptide Fractionation Kit (Thermo Scientific), collecting 1 flowthrough fraction, 1 wash fraction, and 6 step-gradient fractions.

**Data acquisition**

Mass spectrometry data was acquired using an Orbitrap Fusion Lumos Tribrid Mass Spectrometer (Thermo Scientific). Dried peptides were resuspended in 18 µl of 5% acetonitrile with 0.1% formic acid in water, vortexed for 2 minutes, and centrifuged at 16,000 rpm for 16 minutes. For mass spectrometry, 3.8 µl of the resuspended peptides were injected into the Mass Spectrometer. Peptides were trapped for 4 minutes at a flow rate of 4.0 µl/min onto a 100 µm I.D. fused-silica precolumn (Kasil frit) packed with 2 cm of 5 µm ReproSil-Pur 120 C18-AQ (dr-maisch.com), and eluted and separated in 120 minutes at a flow rate of 300 nl/min by an in-house made 75 µm I.D fused silica analytical column (gravity-pulled tip) packed with 25 cm of 3 µm ReproSil-Pur 120 C18-AQ (dr-maisch.com). Mobile phases were A (water (0.1% (v/v) formic acid) and B (acetonitrile (0.1% (v/v) formic acid). The biphasic elution program was as follows: 0-100 min (10-35% B); 100-120 min (35-65% B); 120-121 min (65-95% B); 121-126 min (95% B); 126-127 min (95-5% B); 145 min (STOP).

The MS data acquisition was performed in positive electrospray ionization mode (ESI+), within the mass range of 375-1500 Da with the Orbitrap resolution of 120,000 (m/z 200) and a maximum injection time of 50 milliseconds. Data
dependent acquisition (ddMS2) was carried out with a 1.2 Da isolation window, a resolution of 30,000 (m/z 200), maximum injection time of 110 milliseconds, and the customed AGC target with a 38% of HCD collision energy.

**Data analysis**

Raw data files were processed with Proteome Discoverer (version 2.1.1.21, Thermo Scientific) and searched against the Uniprot Saccharomyces cerevisiae database (downloaded 06/28/2021) using Mascot Server (version 2.8, Matrix Science). Search parameters included full trypsin, with variable modifications of oxidized methionine, pyroglutamic acid (from Q), and N terminal acetylation. Fixed modifications were carbamidomethylation on cysteine and TMT10plex on peptide N-terminus and lysine side chain. Assignments were made using a 10ppm mass tolerance for the precursor and 0.05 Da mass tolerance for the fragments. Peptide and protein validation and annotation was done in Scaffold (version 5, Proteome Software, Inc.) using Peptide Prophet (Keller et al. 2002) and Protein Prophet (Nesvizhskii et al. 2003) algorithms. Peptides were filtered at a 1% FDR, while protein identification threshold was set to greater than 99% probability and with a minimum of two identified peptides per protein. Protein clustering analysis was applied to increase the probability of protein identification for proteins that share peptides (e.g. paralogs). Quantitative analyses, including TMT label-based quantification, median normalization of log₂ intensity values, and log₂ fold change calculation were carried out in Scaffold Q+S.
Data acquired from flowthrough and wash fractions were used to determine the success of high-pH fractionation. Further analysis was based on data acquired from 6 step-gradient fractions.

**Bioinformatics and statistical analyses**

Data analyses and visualization were performed in the R software environment versions 3.5 and 4.2 using the following R packages: data.table, dplyr, reshape2, readxl, openxlsx, caret, randomForest, rPermute, rstatix, rcompanion, DESeq2, limma, Biostrings, seqinr, riboWaltz, ORFik, gprofiler2, scales, ggplot2, ggpubr, ggh4x, ggrepel, ggVennDiagram, ggseqlogo, patchwork, and Cairo.

*Ribosome profiling analysis and readthrough efficiency calculation*

Aligned reads of ribosome profiling libraries were processed by R package riboWaltz (Lauria et al. 2018) for initial diagnostic, read length filter (retain reads 20-23 nt and 27-32 nt in length), and determination of read’s P-site offsets, which were manually checked and modified for accuracy (for all samples, the following offsets from 5’ and 3’ end [nt | nt] were used for each read length: 20 [12 | 7], 21 [13 | 7], 22 [13 | 8], 23 [14 | 8], 27 [11 | 15], 28 [12 | 15], 29 [13 | 15], 30 [13 | 16], 31 [13 | 17], 32 [13 | 18]). Read counts belonging to different mRNA regions (5’-UTR, CDS, extension, and distal 3’-UTR), read’s reading frame, and metagene analysis assessing periodicity were based on the read’s P-site location of mRNAs with annotated UTRs.

Readthrough efficiency was calculated for each mRNA as follows:
Readthrough efficiency = \frac{\text{frame 0 read count in extension}}{\text{frame 0 read count in CDS}} \times \frac{\text{length of extension (nt)}}{\text{length of CDS (nt)}}

where the first 15 bp of the CDS region were excluded to avoid bias in ribosome accumulation over or near the start codon, and the extension region was defined as the 3'-UTR region from the canonical stop codon (inclusive) to the next in-frame stop codon (exclusive).

**Random forest models**

Random forest analyses were carried out with R packages caret (Kuhn 2008), randomForest (Breiman 2001; Liaw and Wiener 2002), and rfPermute (Archer 2021). For each sample, a random forest regression approach was trained to use mRNA features (previously defined in Mangkalaphiban et al. 2021) to predict readthrough efficiency values with 5-fold cross-validation, repeated 5 times, resulting in a total of 25 models. Each model was trained with 100 trees, the default number of features to split at each tree node (square root of number of features), and 1,000 permutation replicates to empirically determine p-value for feature importance. Feature importance score, percent increase in mean squared error (\%IncMSE), was an average of scores extracted from 25 models. A feature was considered significantly important predictor of readthrough efficiency if its empirical p-value was less than 0.05 in at least 15 out of 25 models. Model performance metric was reported as an average of root mean squared error normalized to the range of readthrough efficiency values (NRMSE) across 25 regression models.
Analysis of transcript abundance changes

All analyses involving transcript abundance changes were performed with the R package DESeq2 (Love et al. 2014). The "expected_count" columns in the RSEM file output "isoforms.results" were used as input raw read count. Results were extracted with automatic independent filtering applied at significant cutoff (alpha) of 0.01. The false discovery rate (FDR) method was used to adjust the P-value.

For differential expression analysis of RNA-Seq libraries, mRNAs with adjusted P-value < 0.01 were considered significantly differentially expressed between samples, regardless of magnitude of log₂ fold change. For Figures. 2C-D, S2C, and 6B, where RNA-Seq log₂ fold change were plotted against mass spectrometry log₂ fold change, expected_count of mRNAs in the same protein cluster were added together and RNA-seq analysis was carried out as described.

For relative changes in translation efficiency (TE), ribosome profiling reads whose P-site locations were in the 5'-UTR, 3'-UTR, or the first 15 bp and the last 3 bp of the CDS (ribosomes paused over canonical start codon, translational ramp, and canonical stop codon) were discarded and transcript abundance for the rest of reads was re-quantified by RSEM. TE was defined as Ribo-Seq reads in CDS normalized to RNA-Seq reads. mRNAs with adjusted P-value < 0.05 were considered to have significant changes in TE between samples, regardless of the magnitude of log₂ fold change.
**Analysis of protein abundance changes**

Differential abundance analyses of protein levels between samples were performed with the R package limma (Smyth 2004; Kammers et al. 2015) on log2 normalized intensity data exported from Scaffold. The Benjamini-Hochberg method was used to adjust the P-value. Proteins with adjusted P-value < 0.015 were considered to have significant changes in abundance between samples, regardless of magnitude of log2 fold change.

Gene ontology analysis of proteins enriched (“Up”) or depleted (“Down”) was carried out by the R package gprofiler2’s gost function with default parameters (Raudvere et al. 2019; Kolberg et al. 2020).

**mRNA features**

mRNA features related to analysis of stop codon readthrough efficiency were defined as previously described (Mangkalaphiban et al. 2021).

Identification of uORFs in the 5’-UTR was done with the findORFs function in the R package ORFik (Tjeldnes et al. 2021), limiting uORF’s start codon to be AUG only and no minimum uORF length.

Poly(A) tracts in the 5’-UTR and 3’-UTR were defined as stretches of at least 10 consecutive adenines (the A in AUG of main ORF included), allowing at most 2 other nucleotides in the 10 A’s window.

Oligo(U) in the 5’-UTR and 3’-UTR was defined as a stretch of at least 7 consecutive uracils, allowing no other nucleotides in the window.
Motif discovery

Identification of sequence motifs enriched in 5'-UTR (excluding the start codon) or 3'-UTR (including the stop codon) sequences of mRNA TE groups compared to the Reference was carried out by the STREME algorithm from the MEME Suite 5.5.0 (https://meme-suite.org/meme/tools/streme) (Bailey et al. 2015; Bailey 2021). A general search for 3-15 nt-long motifs was performed as well as a focused search for shorter 3-6 nt-long motifs. Short motifs had to be identified in both searches to be regarded as significant enrichments.

Statistical analyses

Statistical analyses were performed using the R packages rstatix, rcompanion, and ggpubr. Specific statistical parameters, multiple-testing correction method, statistical significance (p-value or symbols representing ranges of p-values), and sample size (n) are reported accordingly on the figures, in the figure legends, or in Table A1.1.

Data Availability

Raw sequencing data have been deposited at the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) under accession numbers GSE229691 and GSE229692. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Perez-Riverol et al. 2022) partner repository with the dataset identifier PXD041495 and 10.6019/PXD041495. Analysis scripts and data used to generate figures are available at https://github.com/Jacobson-Lab/Pab1_deletion.
Acknowledgments

This work was supported by a grant to A.J. (1R35GM122468) from the U.S. National Institutes of Health. We thank Feng He, Chan Wu, Jill Moore, Zhiping Weng, Andrei Korostelev, Sean Ryder, and Alan Hinnebusch for helpful discussions. We thank members of the UMass Chan Medical School Mass Spectrometry Core Facility for proteomics data acquisition, the UMass Chan Medical School Molecular Biology Core Labs for Fragment Analyzer service, and the UMass Chan Medical School Bioinformatics Core and Information Technology for the high-performance computing platform.

Author Contributions

K.M. and A.J. conceived and designed the experiments, K.M. and R.G. carried out the experiments, K.M. wrote data processing scripts, K.M. and A.J. analyzed the data, K.M., and A.J. wrote the paper, and A.J. obtained funding for the study.

Declaration of Interests

A.J. is co-founder, director, and consultant for PTC Therapeutics Inc. K.M. and R.G. declare no competing interests.
APPENDIX II:

Optimization of rRNA depletion method for yeast

ribosome profiling library preparation

Introduction

Ribosome profiling, first developed by Ingolia and colleagues in 2009, is a powerful approach to study transcriptome-wide translation at nucleotide resolution in cell populations (Ingolia et al. 2009). The principle of ribosome profiling is to isolate ribosomes from a cell lysate, digest naked RNAs with an RNase, extract ribosome-protected mRNA fragments (RPFs), subject the fragments to deep sequencing, and map the resulting sequences back to the genome to investigate where and how much translation occurs. One problem with the original ribosome profiling protocol was the large amount of rRNA in the sequencing library, which was up to 84% of all reads in the original paper (Ingolia et al. 2009). Depleting these reads before deep sequencing would increase the yield of actual mRNA translation footprints and reduce sequencing cost. Since then, various depletion methods that had been developed and commercialized mostly for total RNA-Seq have been adapted to ribosome profiling. Our lab had been using Illumina’s Yeast Ribo-Zero Gold rRNA removal kit with success. Libraries prepared with Ribo-Zero yielded five-fold more reads mapped to the mRNA sequences compared to libraries prepared without any rRNA depletion method. However, Ribo-Zero for yeast was discontinued by Illumina in 2020 due to lower demand for yeast products.
compared to other model organisms. Thus, in order to continue our experiments, we set out to optimize a rRNA depletion method to integrate with our current ribosome profiling library preparation protocol.

Results and Discussion

Ribosome profiling library preparation workflow

The workflow of our ribosome profiling library preparation is summarized in Figure A2.1. After RNA extraction from ribosomes, a mixture of RPFs and rRNAs exists in the sample. First, RNAs are dephosphorylated at the 3' end and phosphorylated at the 5' end (Step 1). Then, we used Perkin Elmer’s NEXTFLEX Small RNA-Seq Kit v3 for library preparation, which supplied Illumina-compatible 5' and 3' adapters each linked with 4 randomized nucleotides (4N). The unique combinations of 8N allow for downstream bioinformatics analysis to distinguish PCR duplicates from real unique footprints that map to the same location. Major library preparation steps include ligation of adapters to either ends of the RNAs (Step 2-3), reverse transcription to produce the first cDNA strand (Step 4), and polymerase chain reaction (PCR) to amplify the cDNA followed by a final clean-up step to remove enzymes and excess adapters/primers (Step 5). Without prior removal, rRNAs will be in the final sequencing library along with RPFs due to their similar chemistry. To remove rRNAs, Ribo-Zero was applied before Step 1. The kit contains biotinylated RNA oligonucleotides with sequence complementary to rRNA sequences, so that after rRNA binding to these oligonucleotides, unbound mRNAs can be washed out and collected for library preparation.
Figure A2.1. Ribosome profiling library preparation workflow.
Schematic of major ribosome profiling library preparation steps (Left) and strategies attempted to remove rRNA contamination in the library (Right, triangle bullets) before or during the indicated step.
QIAseq FastSelect (QIAGEN) was the most promising among the tested commercial kits

To find a new rRNA depletion method that would work well with our current library preparation procedure, I collected ribosomes and extracted total RNAs from WT yeast cells, then started testing other commercial yeast kits according to the manufacturers’ protocols, which at the time were largely intended for total RNA-Seq, with minimal modifications. riboPOOL from siTOOLs implemented the same strategy as Ribo-Zero, but with biotinylated DNA oligonucleotides. NEBNext from New England Biolabs (NEB) used single-stranded DNA (ssDNA) oligonucleotides and utilized the unique property of endonuclease RNaseH that specifically hydrolyzes any RNA hybridized to DNA, followed by DNasel digestion of ssDNA oligonucleotides. NEBNext had a subsequent bead-based RNA purification step that involved discarding fragments smaller than 100 nt, which had to be replaced by the ethanol precipitation method because otherwise all ribosome footprints (~20-32 nt) would have been discarded. Both riboPOOL and NEBNext were applied before Step 1 (Figure A2.1). On the other hand, QIAseq FastSelect from QIAGEN used locked nucleic acid (LNA) oligonucleotides, which upon binding to complementary rRNAs would block the reverse transcription reaction for rRNAs (Step 4), subsequently “depleting” them from library preparation steps (Figure A2.1).

After library preparation and sequencing, raw sequencing reads were processed in the following steps: i) adapter trimming and read length filtering, ii)
alignment of reads to non-protein coding RNAs (ncRNAs: rRNAs, tRNAs, snoRNAs, etc.) and their removal, iii) alignment of the remaining reads to the protein-coding transcriptome, and iv) PCR duplicates removal (Mangkalaphiban et al. 2021). I recorded the number of reads at each step for analysis of library quality (Figure A2.2). As a baseline, the library prepared without any rRNA depletion method yielded 6% of all reads as usable reads, i.e., unique reads mapped to the transcriptome (Figure A2.2, *None_0.5ug). The library prepared with Ribo-Zero yielded 30% of all reads as usable (Figure A2.2, *Ribo-Zero_0.5ug). Among the three new kits, the library prepared with NEBNext performed worse than no rRNA depletion (Figure A2.2, *NN_0.5ug_1ul_r), yielding only 0.5% of usable reads. This library had a strikingly high percentage of reads that were shorter than 15 nt, which were unmappable due to too many matches in the transcriptome. These were likely products of incomplete digestion from RNaseH/DNaseI. These products would have been discarded at a size-selection clean-up step (manufacturer’s recommendation for RNA-Seq), but this step had to be skipped to avoid losing ribosome footprints. For the other two kits, the yield from the library prepared with riboPOOL was no better than when no depletion method was used (Figure A2.2, *riboPOOL_0.5ug) while that from QIAseq FastSelect was the best, but only about half as effective as Ribo-Zero (Figure A2.2, *QF_0.5ug_1ul_r). From this set of pilot experiments, QIAseq FastSelect’s strategy of blocking the reverse transcription step seemed the most promising.
Figure A2.2. Ribosome profiling library alignment statistics.

Raw sequencing reads from ribosome profiling libraries (y-axis, grouped by the step at which rRNA depletion was applied as shown in Figure A2.1) were processed and aligned to the yeast transcriptome. Percentages of reads at each processing step were calculated (x-axis) and specifically labeled for those aligned to non-protein coding RNAs and final unique reads that mapped to the transcriptome (“Usable reads”). Library names were generally derived from the kit used (Ribo-Zero, riboPOOL, NN = NEBNext, QF = QIAseq FastSelect), starting RNA amount (in µg), oligonucleotide amount (in µl if used straight from the provided concentration or in pmol if diluted), and hybridization temperature (r = ramp-down, c = constant). Additional labels are i = isopropanol precipitation instead of column-based purification for the clean-up step and S = small RNA selection during RNA extraction. Asterisk (*) indicates a library prepared following the manufacturer's recommendation of rRNA depletion protocol. Statistics for Ribo-Zero is from a replicate 1 library made from WT cells grown at 25°C from Chapter II (Mangkalaphiban et al. 2021).
Optimization of the “oligo block” strategy

I further optimized QIAseq FastSelect’s “oligo block” strategy by varying the type of oligonucleotides (QIAseq FastSelect’s LNA or NEBNext’s ssDNA), the step at which the oligonucleotides would be blocking (before Step 1, 2, 3, or 4), the hybridization temperature (constant or sequential ramp-down), the amount of oligonucleotides used (recommended amount, higher, or lower), and the amount of starting RNAs. I reasoned that the earlier the intervention of blocking oligonucleotides, the more downstream steps would also be blocked, hence the likelihood of rRNA depletion would be higher. After all, the cDNA synthesis step at which QIAseq FastSelect was recommended to be incorporated was an early step in most RNA-Seq protocols. I also hypothesized that the amount of oligonucleotides needed would likely be higher than the recommended amount because RPFs are expected to have much higher amounts of rRNA contamination than total RNAs. For the same reason, the amount of starting RNAs might also have to be higher than recommended, but the ratio of oligonucleotides to RNA would also have to be proportionally adjusted. I made more libraries with different combinations of these parameters, but not all of them resulted in libraries with sufficient materials or quality for sequencing. Additional adjustments to the protocol were made to accommodate the changed parameters. For instance, the 5’-phosphorylation step (Step 1B) was followed by reaction clean-up using a Zymo spin column, which has a binding capacity of 10 µg. Oligonucleotides introduced before this step could clog the column and decrease the yield of actual RNAs, thus
either multiple columns were used and their eluates were pooled or RNAs were instead recovered by isopropanol precipitation. If oligonucleotides were introduced prior to 3’ adapter ligation (Step 2), the denaturation step in the NEXTFLEX protocol (heating samples at 80°C) was skipped to prevent uncoupling of the already hybridized rRNA-oligonucleotide duplex. Another adjustment was size selection for RNAs smaller than 200 bp during RNA extraction (QIAGEN’s miRNeasy kit with Appendix A protocol) from pelleted ribosomes and during RNA purification after 3’-dephosphorylation/5’-phosphorylation steps (Zymo’s RNA and Concentrator-5 kit Appendix protocol) in an attempt to minimize rRNA level prior to the use of oligonucleotides.

In the end, 22 libraries yielded sufficient materials of satisfactory quality for sequencing. Among them, the library prepared with QIAseq FastSelect’s oligonucleotides at twice the manufacturer’s recommended amount, hybridized to size-selected small RNAs before the 3’ adapter ligation step yielded the highest percentage of usable reads at 36%, which was comparable or even better than Ribo-Zero (Figure A2.1, QF_1ugS_2ul_r). We used this optimized protocol for ribosome profiling experiments in parts of Chapter IV and Appendix I.

During library preparation, I tried to minimize the number of PCR duplicates by minimizing the required number of PCR cycles. I started out at 10 cycles and measured the amount of materials (i.e., PCR products) using a Fragment Analyzer. If the 1:4 dilution of PCR products was undetectable, 2-3 additional cycles were performed at a time until they became detectable. As expected, libraries that
required a higher number of PCR cycles tended to have higher amounts of PCR duplicates (Figure A2.2 and A2.3). However, libraries that required a lower number of PCR cycles also tended to have higher amounts of rRNA contamination (Figure A2.2 and A2.3). These observations mean that when rRNA depletion was not effective, there were more RNAs in the sample and thus a smaller number of PCR cycles was required. Therefore, when the amount of starting small RNAs is not 1 µg as tested here, the amount of oligonucleotides may have to be adjusted to maintain the optimal RNA to oligonucleotide ratio in order to achieve the optimal “spot” where rRNA depletion was effective (i.e., not underused) but also not overused to the point of requiring a higher number of PCR cycles to achieve sufficient yield for sequencing.
Figure A2.3. Relationships between number of PCR cycles and percentage of reads at each sequence alignment processing step.
Scatter plots of PCR cycle number vs. reads percentage at each alignment step. Each dot represents a sequencing library. Spearman’s correlation coefficient ($r$) and the associated p-value is reported.

Materials and Methods

Yeast strain and growth condition

The wild-type yeast strain used in this study, HFY114, has the W303 background. Cells were grown at 30°C in 1 L of YEPD with shaking and harvested during exponential growth phase, when the $\text{OD}_{600}$ was 0.6-0.8.

Ribosome profiling

Procedures and materials (products with catalog numbers) used for ribosome profiling library preparation, including isolation of 80S ribosomes, were previously described (Ganesan et al. 2019). The final, optimized ribosome profiling protocol was described in full detail in Chapter IV’s Materials and Methods section and the associated publication (Ganesan et al. 2022).

Modifications of protocols for each strategy tested

Specific steps that were modified during optimization experiments are described below for each library (using library names in Figure A2.2). In all experiments, section 2.7.3 in Ganesan et al. (2019) were omitted since it described the use of the Ribo-Zero protocol.

Pilot experiments using manufacturers’ protocols

*riboPOOL_0.5ug: rRNA depletion by riboPOOLs (siTOOLs, Trial kit) was carried out after RNA extraction according to manufacturer’s recommendation,
followed by column-based RNA purification using RNA Clean and Concentrator-5 (Zymo Research, Catalog # R1016). Purified RNAs were then subjected to 3’-dephosphorylation, 5’-phosphorylation, and NEXTFLEX library preparation workflow.

*NN_0.5ug_1ul_r: rRNA depletion by NEBNext (New England Biolabs) was carried out after RNA extraction according to manufacturer’s recommendation with a minor modification. After DNaseI digestion, the bead-based RNA purification step was replaced by ethanol precipitation. Purified RNAs were then subjected to 3’-dephosphorylation, 5’-phosphorylation, and NEXTFLEX library preparation workflow.

*QF_0.5ug_1ul_r: rRNA depletion by QIAseq FastSelect (QIAGEN, Catalog #334215) was incorporated into the NEXTFLEX Small RNA-Seq Kit v3 workflow (Perkin Elmer) at the reverse transcription step (Step E). Adapter-ligated RNAs (25 µl) were mixed with 13 µl NEXTflex RT Buffer and 1 µl QIAseq FastSelect oligonucleotides, then incubated in a thermocycler at 37°C for 2 min and 25°C for 2 min. Then, 2 µl M-MuLV Reverse Transcriptase was added to the reaction and the remaining steps were carried out according to the manufacturer’s protocol.

**Oligonucleotide block at Step 1A**

NN_0.5ug_1ul_r, NN_1ug_1ul_50pmol_r, NN_1ug_1ul_100pmol_r, NN_1ug_1ul_200pmol_r: The indicated amount of RNA and NEBNext ssDNA oligonucleotides were mixed with 2 µl NEBNext Probe Hybridization Buffer and the reaction was brought to 15 µl with nuclease-free water. The reaction was incubated
in a thermocycler at 95°C for 2 min, then 95°C to 22°C at ramping speed 0.1 C/sec (2% ramping), and, finally, 22°C for 5 min. The reaction was cleaned up using RNA Clean and Concentrator-5.

NN_0.5ug_2ul_c: RNA (0.5µg) was mixed with 2 µl NEBNext ssDNA oligonucleotides and brought to 15 µl with nuclease-free water. The reaction was incubated in a thermocycler at 25°C for 20 min.

QF_1ug_1ul_r: RNA (1 µg) was mixed with 1 µl QIAseq FastSelect oligonucleotides and brought to 15 µl with nuclease-free water. The reaction was incubated in a thermocycler at 75°C for 2 min, 70°C for 2 min, 65°C for 2 min, 60°C for 2 min, 55°C for 2 min, 37°C for 2 min, 25°C for 2 min, and then held at 4°C. The reaction was cleaned up using RNA Clean and Concentrator-5.

For all procedural variations, purified RNAs were subjected to 3'-dephosphorylation, 5'-phosphorylation, and NEXTFLEX library preparation workflow with the omission of steps involving high temperature to avoid denaturation of the rRNA-oligonucleotide duplexes. The omitted steps included denaturation steps at the beginning of 3'-dephosphorylation (80°C for 2 min) and the NEXTFLEX workflow (70°C for 2 min), and T4PNK enzyme inactivation during 3'-dephosphorylation (65°C for 20 min) and 5'-phosphorylation (60°C for 10 min).

Oligonucleotide block at Step 1B

NN_1ug_1ul_50pmol_r, NN_1ug_1ul_100pmol_r, NN_1ug_1ul_200pmol_r, NN_2ug_2ul_r, NN_4ug_2ul_r: The indicated amount of RNA was 3'-
dephosphorylated and purified with RNA Clean and Concentrator-5. Purified 3’-dephosphorylated RNAs were then mixed with the indicated amount of NEBNext ssDNA oligonucleotides and 2 µl NEBNext Probe Hybridization Buffer, and the reaction was brought to 15 µl with nuclease-free water. The reaction was incubated in a thermocycler at 95°C for 2 min, 95°C to 22°C at ramping speed 0.1 C/sec (2% ramping), and then 22°C for 5 min. The reaction was cleaned up using RNA Clean and Concentrator-5.

NN_4ug_2ul_ri: The same procedure used for NN_4ug_2ul_r was performed except for the clean-up step post-hybridization, which was carried out by isopropanol precipitation.

QF_1ug_1ul_r, QF_2ug_2ul_r, QF_4ug_2ul_r: The indicated amount of RNA was 3’-dephosphorylated and purified with RNA Clean and Concentrator-5. Purified 3’-dephosphorylated RNA was then mixed with the indicated amount of QIAseq FastSelect oligonucleotides. The reaction was incubated in a thermocycler at 75°C for 2 min, 70°C for 2 min, 65°C for 2 min, 60°C for 2 min, 55°C for 2 min, 37°C for 2 min, 25°C for 2 min, and then held at 4°C. The reaction was cleaned up with RNA Clean and Concentrator-5.

QF_4ug_2ul_ri: The same procedure used with QF_4ug_2ul_r was performed except for the clean-up step post-hybridization, which was carried out by isopropanol precipitation.

For all procedural variations, purified RNAs were then subjected to 5’-phosphorylation and NEXTFLEX library preparation workflow with the omission of
steps involving high temperature to avoid denaturation of rRNA-oligonucleotide duplexes, which included T4PNK enzyme inactivation during the 5'-phosphorylation step (60°C for 10 min) and the denaturation step (80°C for 2 min) at the beginning of the NEXTFLEX workflow.

Oligonucleotide block at Step 2

QF_1ugS_1ul_r, QF_1ugS_2ul_r: Small RNAs (1 µg) were subjected to 3'-dephosphorylation and 5'-phosphorylation with a size-selection for small RNAs during the reaction clean-up (RNA Clean and Concentrator-5, Appendix protocol). Purified RNAs were then mixed with the indicated amount of QIAseq FastSelect oligonucleotides and brought to 10.5 µl with nuclease-free water. The reaction was incubated in a thermocycler at 75°C for 2 min, 70°C for 2 min, 65°C for 2 min, 60°C for 2 min, 55°C for 2 min, 37°C for 2 min, 25°C for 2 min, and then held at 4°C. Then, the reaction was subjected to the NEXTFLEX library preparation workflow, skipping the initial denaturation step (80°C for 2 min).

Oligonucleotide block at Step 3

NN_0.5ug_1ul_c: RNA (0.5 µg) was subjected to 3'-dephosphorylation, 5'-phosphorylation, and NEXTFLEX library preparation workflow up until the Excess (3') Adapter Inactivation step (Step C). The reaction (14 µl) was then mixed with 1 µl NEBNext ssDNA oligonucleotides and incubated in a thermocycler at 25°C for 20 min. The rest of the workflow was proceeded from 5' 4N Adapter Ligation (Step D).
NN_0.5ug_2ul_r: RNA (0.5 µg) was subjected to 3'-dephosphorylation, 5'-phosphorylation, and NEXTFLEX library preparation workflow up until the Excess (3') Adapter Inactivation step (Step C). The reaction was cleaned up using RNA Clean and Concentrator-5. Purified RNA was mixed with 2 µl NEBNext ssDNA oligonucleotides and 2 µl NEBNext Probe Hybridization Buffer, and brought to 15 µl with nuclease-free water, and incubated in a thermocycler at 95°C for 2 min and 22°C for 5 min. The reaction was cleaned up using RNA Clean and Concentrator-5, then the rest of the workflow was proceeded from 5' 4N Adapter Ligation (Step D).

Oligonucleotide block at Step 4

QF_0.5ug_1ul_c: The library was prepared as described for *QF_0.5ug_1ul_r, except that the hybridization was carried out at 25°C for 20 min.

Acknowledgments

I thank New England Biolabs for providing the NEBNext rRNA depletion kit for yeast, which at the time was under development and not commercially available; Perkin Elmer's R&D team, especially former Senior Application Scientist Kevin Allen, for their suggestions and advice regarding the incorporation of the rRNA depletion method into their NEXTFLEX Small RNA-Seq Kit v3 workflow; the Molecular Biology Core Labs (MBCL) at UMass Chan Medical School for their Fragment Analyzer service; and Robin Ganesan for her unending support, advice, and helpful discussions throughout the optimization process.
Bibliography


Andjus S, Morillon A, Wery M. 2021. From Yeast to Mammals, the Nonsense-Mediated mRNA Decay as a Master Regulator of Long Non-Coding RNAs Functional Trajectory. ncRNA 7: 44.


Cridge AG, Crowe-McAuliffe C, Mathew SF, Tate WP. 2018. Eukaryotic translational termination efficiency is influenced by the 3′ nucleotides within the ribosomal mRNA channel. *Nucleic Acids Res* **46**: 1927–1944.


Kononenko AV, Mitkevich VA, Atkinson GC, Tenson T, Dubovaya VI, Frolova LY, Makarov AA, Hauryliuk V. 2010. GTP-dependent structural rearrangement
of the eRF1:eRF3 complex and eRF3 sequence motifs essential for PABP binding. *Nucleic Acids Research* **38**: 548–558.


Martínez-Santamaría L, Maseda R, De Arriba MDC, Membrilla JA, Sigüenza AI, Mascías J, García M, Quintana L, Esteban-Rodríguez I, Hernández-


Mitkevich VA, Kononenko AV, Petrushanko IYu, Yanvarev DV, Makarov AA, Kisselev LL. 2006. Termination of translation in eukaryotes is mediated by the quaternary eRF1•eRF3•GTP•Mg2+ complex. The biological roles of eRF3 and prokaryotic RF3 are profoundly distinct. *Nucleic Acids Research* **34**: 3947–3954.


Paushkin SV, Kushnirov VV, Smirnov VN, Ter-Avanesyan MD. 1996. Propagation of the yeast prion-like [psi+] determinant is mediated by oligomerization of


Pibiri I, Lenti L, Melfi R, Tutone M, Baldassano S, Ricco Galluzzo P, Di Leonardo A, Pace A. 2018. Rescuing the CFTR protein function: Introducing 1,3,4-


Eukaryotic Translation Termination Complexes Containing eRF1-eRF3 or eRF1-ABCE1. *Cell Reports* **8**: 59–65.


Tate WP, Cridge AG, Brown CM. 2018. 'Stop' in protein synthesis is modulated with exquisite subtlety by an extended RNA translation signal. *Biochemical Society Transactions* BST20180190.


efficiency in a genome-reduced bacterium. Molecular Systems Biology e11301.


Yamaguchi Y, Baba H. 2018. Phylogenetically Conserved Sequences Around Myelin P0 Stop Codon are Essential for Translational Readthrough to Produce L-MPZ. *Neurochemical Research* **43**: 227–237.


Young DJ, Meydan S, Guydosh NR. 2021. 40S ribosome profiling reveals distinct roles for Tma20/Tma22 (MCT-1/DENR) and Tma64 (elF2D) in 40S subunit recycling. *Nat Commun* 12: 2976.


