ABSTRACT

Title of Document: mRNA SUICIDE: DESTABILIZATION BY PROGRAMMED RIBOSOMAL FRAMESHIFTING

Jonathan L. Jacobs, Ph.D., 2006

Directed By: Associate Professor Jonathan D. Dinman, Department of Cell Biology & Molecular Biology

Cis-acting mRNA elements that promote programmed -1 ribosomal frameshifting (-1 PRF) redirect a fraction of translating ribosomes into a new translational reading frame. In viruses, these signals typically direct the translation of alternative protein products from a single mRNA. However, programmed frameshifts could also direct elongating ribosomes to premature termination codons, in which case the mRNAs could become targets for degradation by the nonsense mediated mRNA decay pathway (NMD). Computational analyses revealed the presence of 10,340 consensus -1 PRF signals in the *Saccharomyces cerevisiae* genome. Of the 6,353 yeast open reading frames (ORFs) included in this study, 1,275 are predicted to have at least one strong and statistically significant -1 PRF signal. In contrast to viral frameshifting, the predicted outcomes of nearly all of these genomic frameshift signals would direct ribosomes to premature termination codons, in theory making these mRNAs substrates for NMD. Nine of these predicted -1 PRF signals were tested empirically,

eight of which promoted efficient levels of PRF *in vivo*. This study also demonstrates that viral -1 PRF signals are sufficient to target a reporter mRNA for degradation via NMD. Furthermore, several of -1 PRF signals from the yeast genome were also shown to act as NMD-dependent mRNA destabilizing element. Importantly, these signals are found in genes whose mRNAs are known to be natural targets for NMD. These findings support the hypothesis that PRF may be used by cellular mRNAs to initiate "mRNA suicide". A model is presented in which programmed frameshifting acts as a general post-transcriptional regulatory mechanism to control gene expression by regulating mRNA abundance.

mRNA SUICIDE: DESTABILIZATION BY PROGRAMMED RIBOSOMAL FRAMESHIFTING

By

Jonathan L. Jacobs

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2006

Advisory Committee:
Associate Professor Jonathan Dinman, Chair
Associate Professor Eric Baehrecke
Associate Professor Lian-Yong Gao
Associate Professor Steve Mount
Professor Dorothy Beckett, Dean's Representative

Dedication

This work is dedicated to my son, Aidan. He was a gift to me long before I finished this work, but will be with me long after I have forgotten it.

Acknowledgements

This work was supported by a generous fellowship to me from the National Libraries of Medicine, and various grants to my advisor from the National Institutes of Health and the National Science Foundation. I also want to acknowledge the helpful technical support and computing resources donated from the National Cancer Institute's Advanced Biomedical Computing Center in Frederick, MD.

I would also like to thank my advisory committee. Eric Baehrecke for career advice and general motivational support. Steve Mount for more career advice and reminding me that I should not "forget my thesis". Lian-Yong Gao for stepping in and serving as a committee advisor at the last minute. Dorothy Beckett for her excellent editorial comments and questions during the defense.

I also acknowledge my advisor, Jonathan Dinman, who has provided years of support. His insight, depth of knowledge, and guidance provided me with a constant source of motivation without ever limiting the scope of my work.

And finally, I thank my wife Kristi. My closest friend and life-long partner in science, without her support none of this work would have been possible.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Name	Description
ABCC	The Advanced Biomedical Computer Center, a section of the National Cancer Institute located in Frederick, MD.
CDS	The protein coding sequence of a gene.
cPRF	Candidate programmed ribosomal frameshift signal.
DLR	Dual luciferase assay system (Harger and Dinman, 2003).
DSE	Downstream element.
HDOA	High-density oligonucleotide array.
kD	Kilodaltons.
LULA	Last universally conserved ancestor.
MCS	Multiple cloning site.
MFE	Minimum free energy, usually measured in kcal/mol, for a given RNA secondary structure.
NMD	Nonsense mediated mRNA decay.
PERL	Practical Extraction and Report Language. A common "scripting" programming language used by bioinformatics and Internet developers.
PPCC	Probability plot correlation coefficient.
PRF	Programmed ribosomal frameshift(ing).
PRFdb	The website that connect to the Programmed -1 Ribosomal Frameshifting Database, available via the Internet at http://dinmanlab.umd.edu/prfdb.
PTC	Premature termination codon.

rORF	random open reading frame.
RSV	*Rous sarcoma* virus.
RTC	Read through control.
UDE	Upf-dependent element.
uORF	Upstream open reading frame.
UTR	Untranslated region.
ZFC	Zero frame control.
z_R	z-score of cPRF signals when compared to a distribution of randomly shuffled sequences.

Chapter 1: Literature Review

Introduction

The broad based integration of previously divergent disciplines has positioned the life sciences as a hub of multidisciplinary scientific research. As a result, the mechanisms governing cellular processes, their regulation, and the flow of information at the molecular level within the cell are being revealed in increasingly greater levels of detail and complexity. Interdisciplinary discoveries in the last decade have revealed a new view of the cell; one that represents an entangled system where most processes, genes, and pathways seem to have multifunctional roles. An emerging view of the cell is one where the cellular processes themselves are pleiotropic in addition to the genes which serve as their principle components. As entwined as these networks are, however, technological advances are helping to make the cell into a glass box, and these new connections are opening avenues that allow an even deeper understanding of the regulation gene expression. This dissertation seeks to describe one of these new connections, joining together two previously unrelated processes in the eukaryotic cell: programmed -1 ribosomal frameshifting, and the nonsense mediated mRNA decay pathway.

Programmed -1 Ribosomal Frameshifting

Background

Programmed ribosomal frameshifting (PRF) is a translational recoding phenomenon historically associated with viruses and retrotransposons. A PRF signal stochastically redirects translating ribosomes into a new reading frame (i.e. by +1 or -1 nucleotide) and, in the typical viral context, these signals allow ribosomes to bypass the usual in-frame stop codon and continue synthesis of a C-terminally extended fusion protein, as shown in Figure 1 below. Although many well-characterized methods of translational recoding have been identified, this dissertation focuses solely on programmed -1 ribosomal frameshifting as the target. Other methods of recoding exist and are reviewed elsewhere, such as programmed +1 ribosomal frameshifting, translational hoping, and ribosome shunting (Farabaugh, 1996; Harger et al., 2002; Namy et al., 2004).

Stop codon bypass by actively translating ribosomes via frameshifting was first suggested by Jamjoom et al. (1977) as a possible mechanism capable of producing two proteins from a single mRNA in eukaryotes (Jamjoom et al., 1977). It was not for several years, however, until Jacks & Varmus (1986) demonstrated that ribosomal frameshifting was indeed the mechanism used by *Rous sarcoma* virus (RSV) to direct the translation of both the *gag* structural protein and the *gag-pol* polyprotein from a single mRNA. They were able to generate a [^{35}S]-labeled 76-kilodalton (kD) *gag* protein and a 108-kD *gag-pol* polyprotein using a rabbit reticulocyte lysate system for *in vitro* transcription and translation that had been charged with unspliced RSV mRNA. Furthermore, deletion

analysis of the RSV transcriptional cassette identified a 34-nucleotide window that was responsible for ribosomal slippage. Later studies also demonstrated that other retroviruses used programmed -1 ribosomal frameshifting, including the human immunodeficiency virus 1, HIV-1 (Jacks et al., 1988b), and the mouse mammary tumor virus, MMTV (Jacks et al., 1987). It was not until three years after the initial study of RSV frameshifting, that the "simultaneous slippage model" (SSM) was proposed as the mechanism for -1 PRF (Jacks et al., 1988a). In this later study, Jacks et. al. (1988) used radiolabeled amino acid sequencing, coupled with site-directed mutagenesis, to identify a "slippery" heptamer sequence in the region of overlap between the gag and pol ORFs of RSV. Interpretation of the data led them to propose that the mechanism of -1 PRF involved the simultaneous slippage of both the A- and P-site tRNAs occupying the ribosome during a pause in translation elongation. Deletion analysis confirmed that a second element required for -1 PRF was a strong secondary structure present in the viral mRNA immediately downstream (3') from the slippery site. In this same publication, Jacks et al. (1988) stated that -1 PRF is not expected to be limited to viral genes and that the heptameric slippery sites required for ribosomal slippage are found in many cellular viral and genes. In line with their predictions, it has become increasingly apparent that -1 PRF is much more widespread and is likely employed by organisms representing every branch in the tree of life (Baranov et al., 2002; Cobucci-Ponzano et al., 2005; Harger et al., 2002; Namy et al., 2004). Thus, the seminal work in the 1980's paved the way for future studies into -1 PRF and helped opened the door further into research directed towards

understanding ribosome structure function, translational fidelity, and post-transcriptional of gene expression.

An important benefit reaped from investigations into the various mechanisms of PRF in general has been a vast improvement in our understanding of the kinetics of translation elongation and of key structures in the ribosome important for maintaining fidelity. The various mechanisms that promote programmed frameshifting (e.g. -1 or +1) shed light on the kinetics of translation in unique ways. They have provided conceptual platforms from which the translational community has found a unifying model of translation elongation (Harger et al., 2002). Towards this end, the SSM originally proposed by Jacks et al. (1988) has been a subject of intensive study. Interestingly, after almost two decades of research, the basic tenants of the SSM are still essentially correct (Namy et al., 2004).

-1 PRF Motifs and the Simultaneous Slippage Model

The most well defined -1 PRF phenomena are directed by an mRNA sequence motif composed of three important elements:

1. a "slippery site" composed of seven nucleotides where the translational shift in reading frame actually takes place;
2. a short spacer sequence of usually less than 12 nucleotides; and
3. a downstream stimulatory structure, usually a pseudoknot.

A typical -1 PRF signal is shown in Figure 2 below. In eukaryotic viruses, the slippery site has the heptameric motif N NNW WWH (Harger et al., 2002). Current models posit that aminoacyl- and peptidyl-tRNAs are positioned on this sequence while the ribosome

pauses at the downstream secondary structure (Kontos et al., 2001; Lopinski et al., 2000; Plant and Dinman, 2005; Plant et al., 2003; Somogyi et al., 1993). The nature of the slippery sequence enables re-pairing of the non-wobble bases of both the aminoacyl- and peptidyl-tRNAs with the -1 frame codons, as shown in Figure 3 (Plant et al., 2003). While it is generally accepted that mRNA pseudoknots are the most common type of downstream stimulatory structures, other mRNA structures are capable of filling this role as well (Baril et al., 2003; Kollmus et al., 1996b). Nonetheless, it is thought that the essential function of the stimulatory structure is to provide a kinetic barrier to a translating ribosome to promote a momentary pause in translation (Lopinski et al., 2000).

Nonsense Mediated mRNA Decay

Background

The rapid and specific destruction of a nonsense containing mRNA was first discovered by Losson & Lacroute (1979) (Losson and Lacroute, 1979) in a mutant allele of the orotidine 5-phosphate decarboxylase gene (*URA3)* in *Saccharomyces cerevisiae*. A similar phenomenon was also revealed a few years later in human cell-lines (Kinniburgh et al., 1982) where a nonsense mutation in ß-globin resulted in substantially lower mRNA levels and a specific disease phenotype, ßº-thalassemia (Maquat et al., 1981). Additional examples of "nonsense mediated decay" (NMD) soon followed in additional yeast (Pelsy and Lacroute, 1984) and mammalian genes (Daar and Maquat, 1988), implying that a general mRNA decay pathway was responsible for the surveillance of multiple gene transcripts. NMD-like processes were soon discovered in a

number of additional model organisms including *Caenorhabditis elegans* (Pulak and Anderson, 1993), *Drosophila melanogaster* (Chia et al., 1985; Kreitman, 1983), *Gylcin max* (Jofuku et al., 1989), *Oryza sativa* (Isshiki et al., 2001), and *Phaseolus vulgaris* (Voelker et al., 1990). It is now widely accepted that NMD is an ancient, broadly conserved pathway likely to be central to the normal processing and surveillance of mRNAs in most, if not all, eukaryotes. What is most interesting, however, are the details and variations of the how NMD actually functions in each of these organisms.

The NMD Model in Yeast: Core Surveillance Conserved

The mechanism of how nonsense-mediated mRNA decay operates has been an intensely studied topic since the early 1990's. Several models have been proposed, and subsequently debunked. Even today there are two fundamentally different models for NMD with evidence for and against both; referred to here as the EJC model (Maquat, 2004b), and the *faux* UTR model (Amrani et al., 2004). Fortunately, the major division in the two models lies in the fact that the first model is generally applied to "higher eukaryotes" and the second applies more generally to *S. cerevisiae*[1]. It remains to be demonstrated clearly in plants how NMD functions and how this agrees with what is known about NMD in other eukaryotes. The yeast model of NMD is shown in Figure 4 (Gonzalez et al., 2001).

[1] As a result, readers are cautioned that the literature is replete with references to of NMD "in yeast", or "in mammals".

Regardless of which model is "correct", an important feature of NMD in general is the requirement of three proteins, encoded by the *UPF1*, *UPF2*, and *UPF3* genes that serve as the core of the pathway. This core is functionally conserved in all systems in which NMD has been identified. The model of how these proteins interact, and the role they play in targeting mRNAs for decay, is also generally conserved. In yeast, the system under study in this dissertation, premature termination is immediately followed by the disassembly of poly-ribosome mRNPs, the association of the premature termination codon containing (PTC+) mRNA with the surveillance complex[2], and the rapid degradation of the transcript. The association of the surveillance complex with the PTC+ mRNA is thought to be mediated by an interaction of Upf1p with Hrp1p (Gonzalez et al., 2000), which binds a weakly defined downstream *cis*-element (DSE) or by Upf1p-mRNA interaction. Once an mRNA has been marked as aberrant by NMD there is a complete reorganization of the mRNP which in turn results in the mRNA exiting the pool of actively translated transcripts (Muhlrad and Parker, 1999) and the subsequent sequestering of the transcript in specific processing bodies found elsewhere in the cytoplasm (Brengues et al., 2005; Sheth and Parker, 2003). In yeast, NMD facilitates mRNA decay primarily by accelerating the deadenylation and subsequent decapping of mRNAs, followed by their rapid 5' and 3' degradation by Xrn1p and the exosome complex respectively (Hagan et al., 1995; Mitchell and Tollervey, 2003; Muhlrad and

[2] The surveillance complex refers to the interacting complex of *UPF1*, *UPF2*, and *UPF3* proteins that serves as the core molecular machinery of NMD.

Parker, 1994). It should be noted, however, that the mechanisms and transcript selection methods of NMD vary across species (Maquat, 2004a).

In yeast, the *faux* UTR model best describes our current understanding of how mRNAs are targeted to NMD (Amrani et al., 2006). Briefly, the *faux* UTR model posits that the kinetics of translation termination at a premature termination codon, PTC, is extremely inefficient relative to the rate of termination at the proper stop codon. Furthermore, the model suggests that this inefficiency is related to the proximity of the stop codon to the downstream untranslated region, 3'-UTR (Cao and Parker, 2003). Inefficient premature termination is due to the lack of the proper termination "context" and the absence of factors bound to the 3'-UTR that are known to facilitate efficient termination events. Using mRNA toe-printing assays (Sachs et al., 2002), Amrani et al. (2004) demonstrated that ribosomes terminating translation at premature termination codons were stalled much longer than those that terminated at the normal downstream stop codon, and they postulated that the difference in termination kinetics acts as the driving force for recruitment of the mRNA by NMD. Furthermore, by introducing a *faux* 3'-UTR with a tethered poly(A)-binding protein immediately downstream from a PTC, it was possible to have nonsense containing mRNAs evade the NMD machinery (Amrani et al., 2004). Finally, Amrani et al. (2004) suggest that exon-junction complexes in metazoans or the yeast protein Hrp1p, may serve simply as functional analogues to factors bound to the 3'UTR of mRNAs in yeast.

Natural mRNA Targets of NMD

It has been argued for nearly two decades that the primary role of NMD is to serve as a control point for unwanted expression of truncated or otherwise aberrant proteins (Maquat, 2004b). This view has been supported by the "pioneer-round" model of translation (Chiu et al., 2004; Ishigaki et al., 2001; Lejeune et al., 2003) in which PTC+ mRNAs are subjected to NMD only during the first round of translation. Hence, the general view of NMD has been one of quality control and mRNA transcript surveillance. However, this view has been challenged recently by several groups that have shown that NMD is responsible for the normal expression of genes in *C. elegans*, *S. cerevisiae* and in mammalian cell systems (Dahlseid et al., 2003; Green et al., 2003; He et al., 2003; Hillman et al., 2004; Kebaara et al., 2003; Lelivelt and Culbertson, 1999; Mendell et al., 2004; Rehwinkel et al., 2005; Wittmann et al., 2006). These studies have demonstrated that there is a broad class of mRNAs, estimated between 10 - 20% of the eukaryotic transcriptome, that are primary and secondary targets for NMD-dependent expression. Furthermore, in an evolutionary context, NMD significantly increases the tolerance and retention of PTC+ containing genes in diploid organisms, thereby reducing negative selection against them, primarily because a second copy of the gene is still present and functional (Lynch and Kewalramani, 2003; Xing and Lee, 2004). The components of NMD have also been traced back to the "last universally common ancestor" (LUCA) and the pathway as a whole is believed to have been present in the earliest of eukaryotes (Anantharaman et al., 2002; Mendell et al., 2004; Wilkinson,

2005). Together, these findings strongly suggest that the role of NMD in the cell is much more than just a transcript surveillance mechanism.

In general terms, five classes of mRNA substrates have been identified as so-called natural NMD targets in *S. cerevisiae*. First, transframe or missense exonic mutations that result in nonsense containing mature mRNAs are the most widely studied target for NMD (Leeds et al., 1991). Second, inefficiently or erroneously spliced pre-mRNAs, such as those encoded by *CYH2* (He et al., 1993), are targeted to NMD upon translation in the cytoplasm. Third, mRNAs with small upstream open reading frames, (uORFs) have been shown to be stability regulated NMD substrates (Ruiz-Echevarria and Peltz, 2000). Fourth, primary NMD targets are mRNAs whose stability is directly influenced by the activity of the core NMD proteins despite any identifiable sequence aberrations, such as the mRNA encoded *EST2* (Dahlseid et al., 2003). Fifth, secondary targets of NMD are mRNAs whose downstream transcript levels are affected by the expression of a primary NMD target, such *URA1*, *URA2*, and *URA4* in yeast whose transcription is under the regulation of the transactivator Ppr1p (Losson et al., 1983). Interestingly, the mRNA encoded by *PPR1* has been shown to be a direct target of the NMD pathway, although the mechanism of this regulation has yet to be unraveled (Kebaara et al., 2003).

The first genomic study that identified natural targets for NMD applied high-density oligonucleotide microarrays (HDOA) against RNA isolated from wild-type, *upf1*Δ, *upf2*Δ and *upf3*Δ yeast strains (Lelivelt and Culbertson, 1999). Lelivelt and Culbertson found that 529 mRNAs, approximately 9% of the transcriptome, were

between 1.2- and 11-fold overexpressed in a *UPF*-dependent manner. Interestingly, there was no correlated effect with the presence or absence of introns in each of these genes which suggested that alternative splicing is not the primary source of NMD substrates in yeast[3]. After directly testing several candidate NMD targets by Northern analysis, Lelivelt and Culbertson concluded, however, that the majority of upregulated mRNAs were the result of secondary effects from the misregulation of transcription transactivators. This earlier study nonetheless fueled the search for primary NMD targets and additional genomic studies aimed at identifying them. A second genomics study in yeast identified 545 ORFs being identified as NMD-dependent (He et al., 2003)[4]. Using additional genetic knockouts of *DCP1* and *XRN1*[5], it was surprising that He et al. (2003) concluded that the majority of mRNAs targeted for NMD were primary targets. Recently, Atkin and colleages (Taylor et al., 2005) have sought to settle these conflicting views of natural NMD targets by analyzing and integrating the results from the two previous microarray studies and identifying coregulated sets of genes based on known transcription factors and their putative binding sites. They concluded that there are

[3] It should be noted that *S. cerevisiae* is known to have relatively few intron containing genes compared to other eukaryotes and that, as previously mentioned, NMD is thought have played a significant role in the proliferation of intron containing genes in metazoans.

[4] Interestingly, the intersection of the two studies by the Culbertson and Jacobson groups did not produce an identical list of genes whose expression was NMD-dependent.

[5] *DCP1* and *XRN1* are genes encoding the decapping enzyme and the primary 5'-3' exonuclease in yeast, both of which are downstream from NMD transcript selection. Using these knockouts in conjunction with *upf1*Δ, *upf2*Δ and *upf3*Δ, He et al. (2003) were able to identify targets through epistasis.

indeed primary and secondary targets of NMD, but the lack of sequence motifs for transcription binding sites limited them to the analysis of genes with known functions. Since approximately 46% of the transcripts upregulated in the He et al. (2003) data set have unknown function, the results from Atkin and colleages remains largely inconclusive. Finally, similar reports spurred by the relative success in *S. cerevisiae* have since been published detailing the NMD-dependent repression of large classes of mRNAs in human cell lines (Mendell et al., 2004), *C. elegans* (Mitrovich and Anderson, 2000), and *D. melanogaster* (Rehwinkel et al., 2005).

Summary

Described above are two examples of biological processes thought to be involved in the expression of genes that are "exceptions to the rule". Programmed -1 ribosomal frameshifting was thought to be primarily a viral phenomenon associated specifically with bicistronic mRNAs encoding transframe ORFs. Nonsense mediated mRNA decay is a pathway that specifically targets aberrantly spliced or nonsense containing mRNAs for rapid degradation. Both of these processes, however, have been demonstrated to be involved in the normal expression of many genes, albeit in unusual and as yet uncharacterized ways. This dissertation seeks to connect these two fields together. Chapter 2 presents evidence that programmed -1 frameshift signals are widespread in the yeast genome and that the expected result of these frameshift signals is the premature termination of translation. The statistical properties of an assay system to measure frameshifting *in vivo* are explored in Chapter 3. Following in Chapter 4 is empirical evidence supporting the notion that many of these predicted frameshift signals are functional *in vivo* and act as mRNA destabilizing elements dependent on the functioning of the nonsense mediated decay pathway. Finally, in Chapter 5, we conclude with a model of "mRNA suicide" which describes how functional frameshift signals can direct mRNAs to NMD. The coupling of these two processes poses a potentially novel mode of post-transcriptional control of gene expression.

Chapter 1 Figures

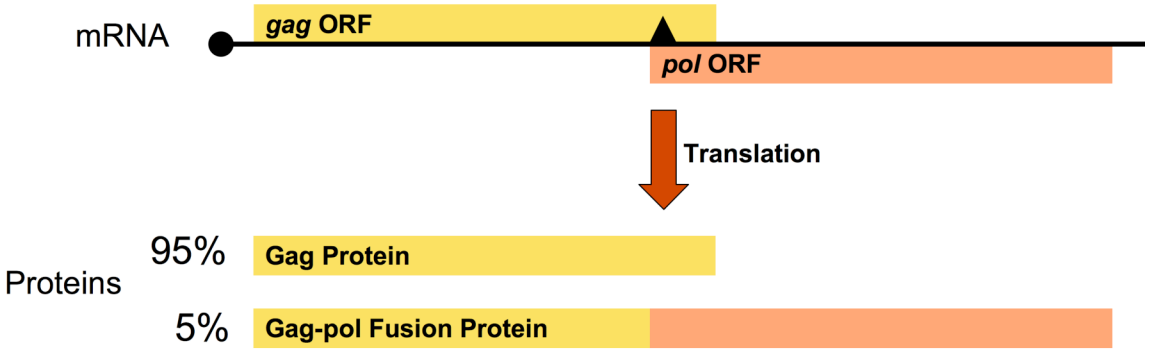

Figure 1: The Viral Context for Programmed Ribosomal Frameshifting

In the above example, the *gag* and *pol* ORFs occupy the same mRNA and are out of frame with respect to one another. A programmed frameshift signal, indicated by the black triangle, directs 5% of translating ribosomes to bypass the zero-frame stop codon at the end of the *gag* ORF and continue translation into the *pol* ORF. The translation products from this mRNA are two proteins whose stoichiometric ratio are determined by the efficiency of the frameshift signal itself.

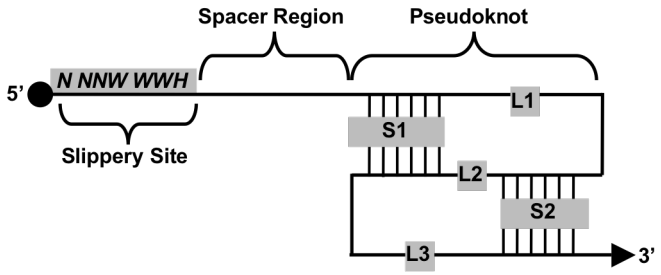

Figure 2: A Typical -1 PRF Signal

Typical -1 PRF signals consist of a heptameric slippery site that fits the motif N NNW WWH (spaces indicate zero frame codons), a short spacer region and an mRNA pseudoknot with two stem and three loop regions (S1, S2, and L1 – L3 respectively).

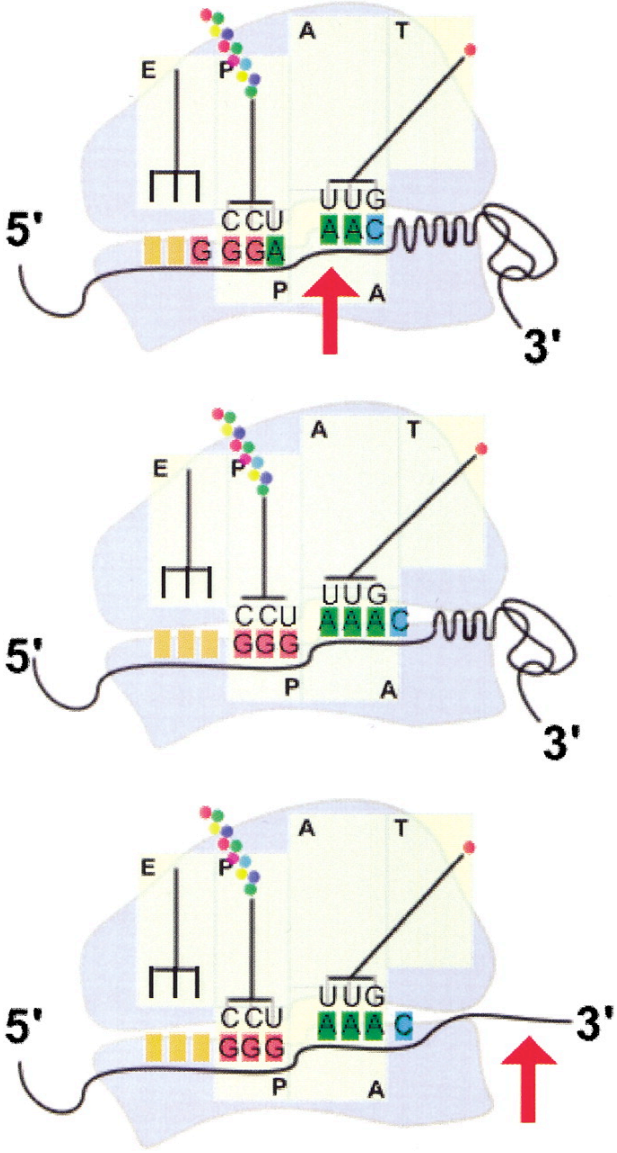

Figure 3: Overview of Programmed -1 Ribosomal Frameshifting

This figure and legend adapted from Plant *et. al.* (2003). (Top) An mRNA pseudoknot induces elongating ribosomes to pause with their A- and P-site tRNAs positioned over the heptameric X XXY YYZ ‘‘slippery site’’ (red arrow). The incoming frame is indicated by spaces. (Middle) While paused at the slippery site, if the ribosome shifts by 1 base in the 5' direction, the non-wobble bases of both the A- and P-site tRNAs can re-pair with the new –1 frame codons. (Bottom) The mRNA pseudoknot is denatured (arrow), and elongation continues in new reading frame.

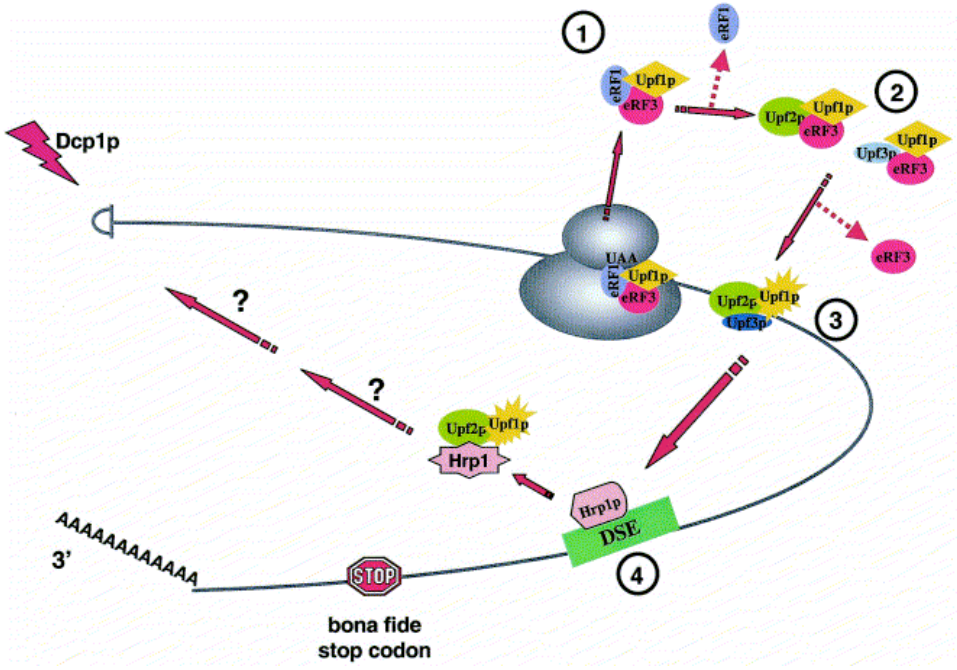

Figure 4: A Model of Nonsense Mediated mRNA Decay in Yeast

This figure and the following legend have been adapted from Gonzalez et al. (2001) (1) Upf1p associates with translation release factors eRF1 and eRF3 during the termination. (2) Dissociation of eRF1 from the ribosome allows Upf2p to bind the eRF3-Upf1p complex. (3) Upf3p joins the complex and displaces eRF3. Failure to displace Hrp1 from the DSE during translation allows the trimeric Upf complex to recognize the DSE marker as a signal that promotes the rapid decaping and subsequent exonucleolytic degradation.

Chapter 2: Computational Identification of Programmed -1 Ribosomal Frameshift Signals in *Saccharomyces cerevisiae*

Introduction

A growing number of examples now exist of PRF signals in expressed eukaryotic genes (Bekaert et al., 2005; Manktelow et al., 2005; Matsufuji et al., 1996; Morris and Lundblad, 1997; Shigemoto et al., 2001; Wills et al., 2006). The existence of these PRF signals in a wide variety of viral and prokaryotic genomes suggests an ancient and possibly universal mechanism for controlling the expression of actively translated mRNAs. There have been several published reports aimed at the *in silico* identification of "recoding signals" using a wide variety of computational approaches (Bekaert et al., 2003; Bekaert et al., 2005; Gao et al., 2003; Gurvich et al., 2003; Hammell et al., 1999; Harrison et al., 2002; Moon et al., 2004; Namy et al., 2003; Shah et al., 2002). While the methodologies of each study covered a broad range of bioinformatics techniques, with the exception of Hammell *et al.* (1999), the general goal of each of these was fundamentally the same. Searches were directed to first find out-of-frame ORFs followed by the identification of PRF signals in the overlapping region between them that could act to potentially redirect ribosomes from the upstream ORF into the downstream ORF, thereby resulting in the translation of a fusion protein. The results of these investigations suggest that PRF signals are more widespread than previously anticipated and that their distribution is not limited to viral or prokaryotic genomes. This computational strategy is based on the assumption that PRF outcomes should mimic

those observed in viral genomes. The strength of this approach is that it can identify new classes of *cis*-acting signals capable of directing efficient PRF. However, its weakness is that it cannot identify new functional outcomes of frameshifting.

In contrast, while an "outcome-neutral" approach using mRNA motifs known to promote efficient PRF cannot identify new frameshift signals, it can enable an expansion of our understanding of functional uses for PRF. In this vein, the first computational search for eukaryotic -1 PRF signals (Hammell et al., 1999) did not focus solely on identifying two overlapping out-of-frame ORFs, but instead aimed to find these motifs throughout the entire CDS of the yeast genome. This early study identified some 260 putative -1 PRF signals in the annotated portion of the *Saccharomyces cerevisiae* genome. An unexpected finding was that the vast majority of "genomic" -1 PRFs were dramatically different from viral frameshifts: greater than 99% of such events were predicted to direct elongating ribosomes to premature termination codons. This observation engendered the hypothesis that PRF could be used to post-transcriptionally regulate gene expression through the nonsense-mediated mRNA decay pathway. Proof of this principle was later demonstrated in yeast by monitoring the effects of a well-defined viral -1 PRF signal on the stability of the resulting reporter genes' mRNA, the details of which are in Chapter 4 below and Plant *et al.* (2004). The shortcomings of the Hammell et al. (1999) study were its limitation by incomplete annotation of the yeast genome and relatively insufficient computational resources available at the time (ca. 1995-98). Thus, in order to achieve a more comprehensive approach, a new set of

informatics tools were developed and applied using faster and more robust computational platforms. The results of the bioinformatics presented here in Chapter 2 show that:

1. pattern matching approaches coupled with a predictive method for folding RNA sequences provide a dramatic improvement in the results;
2. -1 PRF motifs are widespread in the genome of *S. cerevisiae*; and
3. many of the putative signals identified have predicted secondary structures with statistically significant measures of free energy.

This method of identifying putative signals was also tested on the genome of a recently sequenced virus, the SARS coronavirus, and successfully identified a novel -1 PRF signal that was later confirmed experimentally. Finally, the results from the yeast genome are stored in the PRFdb, a publicly accessible Internet resource, created to house the computational results of this search for "context-neutral" -1 PRF signals in yeast

Materials & Methods

Hardware & Software Used

All software was compiled and run on one or more of the following systems: Dell Precision 620, 2 x PIII XEON 866 MHz running Mandrake Linux 10.x; Apple Power Macintosh, 2 x G4 1.4 GHz PowerPC running OS X Tiger; SGI Cluster 64x MIPS R14K 600MHz running Irix 6.5; SGI Altix 3000, 64x 1.5 GHz Itanium II running Linux-64. Supercomputing resources were made available courtesy of The National Cancer

Institute's Advanced Biomedical Computing Center (ABCC)[6]. Unless otherwise noted, data mining and analysis was carried out using scripts written in PERL, each of which are available on request. In all cases, data is stored in a mySQL relational database referred to as the PRFdb. The permanent Internet address of the PRFdb is http://dinmanlab.umd.edu/prfdb.

Pattern Matching

RNAMotif (Macke et al., 2001) was utilized for finding subsequences in the coding regions of *Saccharomyces cerevisiae* that serve as potential translational frameshift signals. The descriptor of the putative programmed frameshift signal motif was created *ad hoc* from analysis of 56 known viral -1 PRF signals from the RECODE (Baranov et al., 2003) database. The *RNAMotif* descriptor had the following requirements:

1. define slippery sites as "N NNW WWH", where spaces indicate zero-frame codon boundaries. N, W and H are standard IUPAC codes and represent triplet repeats;
2. allow any sequence between 0 – 12 nucleotides in length to serve as the spacer between the slippery site and the pseudoknot;
3. allow G:U base pairing in pseudoknot stems;
4. each stem in the pseudoknot must be between 4 – 20 nucleotides in length;
5. stem 1 must have at least 50% GC content;

[6] The computational resources of the ABCC are available to any NIH funded research program. Their Internet address is http://www.abcc.ncifcrf.gov.

6. the first loop must be between 1 – 3 nucleotides in length;
7. the second loop is optional and can be no longer than 3 nucleotides
8. the third loop must be at least as long as one-half the length of the first stem and no longer than 100 nucleotides.

The relationship of these features with regards to a typical -1 PRF signal is illustrated in Figure 2 above on page 15.

Whole Genome Randomization Methods

The complete coding sequence (CDS) of *S. cerevisiae* was randomized 100 times using seven different methods for sequence randomization[7]. Each method of randomization was conducted such that each genome had the same number of ORFs of identical lengths of the natural *S. cerevisiae* genome. In addition, each random genome was generated such that stop codons were only present in the terminal 3' position[8]. Beyond these similarities, the seven methods for randomization included:

1. noBias - randomized ORFs with unbiased nucleotide bias;
2. nShuffle - nucleotides from each natural ORF are shuffled by triplicate mononucleotide permutations;
3. nBias - randomized ORFs using the natural CDS single-nucleotide frequency;
4. cShuffle - codons from each natural ORF are shuffled by triplicate monocodon permutation;

[7] 700 random genomes total.

[8] In other words, no in-frame termination codons were allowed for any of the randomly generated ORFs.

5. sBias - a silent bias where the codons are randomized in place so as to maintain protein coding sequence;
6. cBias - randomized ORFs using the observed CDS codon usage bias; and
7. dnNuc - randomized ORFs generated using the observed CDS dinucleotide frequency.

As was done for the natural *S. cerevisiae* genome, *RNAMotif* was used to search each of these randomized genomes.

RNA Secondary Structure Prediction

Pknots (Rivas and Eddy, 1999) was used to predict the minimum free energy "fold" of each motif hit identified by *RNAMotif*. Each motif hit identified by *RNAMotif* was folded by *pknots* and assigned a predicted minimum free energy value (MFE in kcal/mol) and a predicted secondary structure. The structural calculations and predictions were carried out using the supercomputing hardware at the ABCC.

Data Redundancy Elimination

PERL scripts were created and parsed the entire PRFdb for structurally redundant -1 PRF signals; i.e. any record that was structurally identical with any other record associated with the same slippery site in the same gene was removed. Specifically, the criteria for record elimination was simply that to find any sequence that was a complete subset of any other sequence for the same PRF signal was removed, leaving the larger of the two in the PRFdb. This reduced the overall size of the PRFdb from 173,452 sequence windows, motifs first identified by *RNAMotif*, to a smaller dataset of 66,842 sequence

windows that were non-redundant in terms of their predicted secondary structures and MFE values.

PRF Signal Randomization & Statistical Analysis

Each folded motif hit was randomly shuffled and refolded 100 times using *pknots*, producing a distribution of random MFEs specific for each of the motif hits. Distributions of random MFE values using *pknots* with pseudoknots folding disabled were in general not statistically different from those generated using *pknots* with this option enabled (data not shown), but had considerably shorter generation time, identical energy parameters, and could be run on the same computing platform. Motif hits were then compared to the resulting distribution and assigned a *z*-score:

$$z_R = \frac{X - \bar{x}}{\sigma} \qquad [\ 1\ ]$$

where X is the predicted MFE value for each sequence, $\bar{x}$ is the estimate of the mean for the distribution of MFE values obtained from 100 randomizations, and σ is the standard deviation of random structure MFE values. The normalized value of z_R (*z*-random) obtained provides an estimate of the statistical significance and uniqueness of the predicted structure for the natural sequence: *i.e.* is the sequence more or less stable than we might expect by chance (Chamary and Hurst, 2005; Freyhult et al., 2005; Le et al., 1989; Le et al., 2001; Le et al., 2002; Schultes et al., 1999; Seffens and Digby, 1999; Tuplin et al., 2002).

Computational Analyses of the SARS-CoV -1 PRF Signal

The SARS-associated coronavirus -1 PRF signal was identified from the complete genome sequence, using a combined approach. First, the pattern matching descriptor of known -1 PRF signals was used in conjunction with *RNAMotif* to identify the nucleotide sequence corresponding to the frameshift signal's slippery site. Second, *pknots* was employed to fold the motif hits immediately 3′ to the slippery site and to produce a predicted MFE value in kilocalories per mole for the sequence. The statistical significance of the lowest energy MFE value of the sequence window was tested by generating 500 randomly shuffled sequences derived from the native sequence, refolding each of these, and calculating their MFE values using *pknots*. This resulted in a normal distribution of MFE values, against which the native sequence could be compared and a z_R score calculated. *FASTA3* (Pearson, 2000) was used to initially identify sequences homologous to the SARS -1 PRF signal based on primary sequence similarity. The search space included 1,724 viral genome sequences downloaded using the National Center for Biotechnology Information's Entrez Taxonomy Browser (Wheeler et al., 2000). The resulting pairwise alignments produced by *FASTA3* were collated to produce a multiple-sequence alignment using *Clustal W* (Thompson et al., 1994). An unrooted phylogenetic tree was created from this alignment and visualized using *Tree View* (Page, 1996).

Results

RNAMotif and the canonical -1 PRF signal

The main differences between this study and the previous work by Hammell *et al.* (1999) are:

1. the availability of a completely annotated yeast genome;
2. significantly more powerful computational resources;
3. application of more sophisticated statistical analyses; and
4. a different parameter was employed for the -1 PRF motif.

RNAMotif (Macke et al., 2001) was exploited, and an appropriate albeit somewhat relaxed, "descriptor" of known viral -1 PRF signals was developed by analysis of a database of experimentally confirmed recoding signals (Baranov et al., 2003). The results of this pattern matching approach identified 10,340 slippery sites in the 6,353 annotated coding sequences (CDS) of the yeast genome, 6,016 of which are followed by at least one pseudoknot motif. In total, *RNAMotif* identified 173,452 sequence windows that matched the specified parameters[9].

Whole Genome Randomization, prevalence in yeast

To determine the statistical significance of the above results, they were compared to what would be expected by chance. One method of identifying statistically significant

[9] The large number of motif hits made by *RNAMotif* is the result of many overlapping sequence windows for each match. Each sequence fulfills the criteria for a -1 PRF signal in multiple ways.

motifs in nucleic acid sequences is to repeat the initial motif search using a large set of randomized sequences. The frequency of finding the motif in randomized sequences can provide some insight into the likelihood that a match in a natural sequence occurs by chance. In this chapter, a conservative approach was applied by randomizing the entire yeast CDS genome using seven different strategies so as to not introduce bias due to the choice of any one randomization method. All of the randomized genomes contained the same number of ORFs (rORF) as the natural yeast genome and the same number of total nucleotides in the CDS sequence space. Furthermore, rORFs with in-frame premature termination codons were discarded and randomly re-generated until full length read-through sequences were obtained. A total of 100 randomized replicate genomes were generated for each of the seven methods. Each random genome was then searched for the occurrence of potential -1 PRF signals with *RNAMotif* using the same descriptor described in the Materials & Methods. The results in Table 1 below show that the actual number of motif hits found is statistically different when compared to any of the seven randomized datasets[10]; suggesting that the prevalence of -1 PRF signals may be under multiple selective pressures.

Each of the randomization types that seek to mimic the natural CDS of yeast (cShuffle, sBias, cBias, and dnBias) generated genomes that harbored more -1 PRF signals than were actually found in the natural genome. This suggests selective pressure against the acquisition of spurious -1 PRF signals in yeast; *i.e.* the yeast genome would be expected to have more -1 PRF signals than were actually observed. This is consistent

[10] Each of the tests had a p-value ≤ 0.02 by a two-sample Student's t-test (Devore J. L. ,2000).

with the notion that -1 PRF signals can lead to aberrant translation and, most likely, dysfunctional proteins. In contrast, randomization strategies that mimicked the overall genome-wide or individual CDS nucleotide bias (nBias and nShuffle respectively) produced random genomes with significantly fewer -1 PRF signals than were actually observed. If there were strong and genome-wide evolutionary pressures against the presence of any -1 PRF signals, then they would be expected to be relatively non-existent in the yeast genome and statistically indistinguishable from the nShuffle and nBias randomization datasets. This is, however, not the case. This set of comparisons suggest that there may be evolutionary pressure for the maintenance of certain classes of existing frameshift signals. In addition, randomized genomes using an unbiased nucleotide frequency (noBias) were generated as a negative control. These random genomes contained far fewer -1 PRF signals than observed for the actual yeast genome and far less than any of the other randomization strategies, supporting the hypothesis that the function of -1 PRF has been positively selected for. In sum, the number of slippery sites followed by at least one pseudoknot motif (6,016) present in the actual yeast genome is statistically significant when compared to the number of expected -1 PRF signals for all of the randomization strategies employed. Therefore, although unexpectedly large, the number of putative signals identified can not be ruled out as entirely artifactual.

Secondary Structure Prediction

The next step was to assign additional layers of predictive metrics to the dataset so as to enhance the ability to identify functional -1 PRF signals for empirical testing. The first task was to assign a minimum free energy (MFE) value to each motif hit

identified by *RNAMotif* (Macke et al., 2001). This was not a trivial matter since nearly all known -1 PRF signals require an mRNA pseudoknot (Plant et al., 2003) and RNA pseudoknot prediction represents a well known and computationally difficult problem (Lyngso and Pedersen, 2000). However, *pknots* (Rivas and Eddy, 1999), an algorithmic extension of the popular *mfold* (Mathews et al., 1999), is capable of predicting pseudoknots of the type that are generally found associated with functional -1 PRF signals. Coupled with a set of scripts written in PERL (Wall et al., 2000), *pknots* was able to fold every potential *RNAMotif* hit[11] in approximately 5000 CPU hrs. using the computational resources available at the ABCC. Once the initial folding was completed, the dataset was then reduced to a structurally non-redundant dataset of 66,842 structures through the use of several additional scripts written in PERL. The nearly 3-fold reduction in the data was possible due to the huge number of overlapping motif hits initially made by *RNAMotif*. These analyses provide each non-redundant *RNAMotif* match with a predicted RNA secondary structure and MFE value.

The overall distribution of all MFE values determined by *pknots* for the most stable predicted secondary structures[12] for each structure 3' from the slippery motif is

[11] All 173,452 sequence windows that were initially identified by *RNAMotif*.

[12] The most stable subsequence immediately downstream of a given slippery site is the sequence window of RNA that, when folded by *pknots*, results in the lowest MFE value in kcal/mol as compared to every other sequence window associate with the same slippery site.

shown in Figure 5 below and fits a normal distribution[13]. The distribution of base-pair counts for each structure fits an extreme-value distribution and is shown in Figure 6 below. The statistical correlation of MFE, length, base pair counts and z_R scores for the entire set of candidate PRF signals is shown in Table 2. The summary statistics of the same structures are shown in Table 3.

Randomization of RNA structures

To identify statistically significant motif hits, *z*-scores (z_R) were calculated for each predicted RNA secondary structure folded by *pknots*. For each candidate signal, the MFE value of the predicted structure was compared to the distribution of MFE values obtained from 100 permutations of the same sequence using an implementation in PERL of a similar algorithm previously described (Seffens and Digby, 1999). The randomization approach disrupts the nucleotide base order and any potential secondary structure for each input sequence but preserves the exact mononucleotide count of each base within the shuffling window. Significance scores derived from permutation shuffling approaches such as this have previously been successful in finding biologically meaningful RNA structures from primary sequence data both by this author (Plant et al., 2005) and several other research groups (Barrette et al., 2001; Le et al., 1989; Le et al., 2001; Seffens and Digby, 1999). Furthermore, it is expected that this measure of

[13] Probability plot-correlation coefficient (PPCC) ≥ 0.98. See Chapter 3 and Filliben (1975) for PPCC values and their use in determining how well a distribution of observed data actually fits a normal distribution.

significance is sufficient since functional secondary structures in mRNA sequences are considered more stable than random sequence and are under selective pressure (Chamary and Hurst, 2005; Katz and Burge, 2003; Ringner and Krogh, 2005; Schultes et al., 1999). It should be noted, however, that several reports have indicated that this randomization strategy is not accurate for estimating the significance of RNA secondary structures in-general and that a superior method of randomization lies in preserving both mono- and dinucleotide ratios (Clote et al., 2005; Freyhult et al., 2005; Rivas and Eddy, 2000; Workman and Krogh, 1999). Nonetheless, for the purposes of this study the randomization strategy employed for the calculation of z_R was adequate. For this dataset, the randomization step was limited to 100 permutations per sequence due to the sheer number of input sequences that required z_R scores. This reasonably estimated a normal distribution of MFE values for each input sequence and a probability plot correlation coefficient goodness-of-fit test (Filliben, 1975) was carried out for each distribution to statistically verify each estimation of a normal distribution. A PPCC $\geq$ 0.98 was found for greater than 99% of all the candidate signals in the database indicating that 100 random shuffles was sufficient for good estimates of z_R.

Any $z_R \leq$ -1.65 indicates a structure with a p-value $\leq$ 0.05 and is therefore more stable than expected by chance. The distribution of z_R scores for all candidate PRF signals fits a normal distribution with a PPCC $\geq$ 0.98 and is shown below in Figure 7. A total of 3,228 candidate signals out of 66,842 non-redundant structures include putative structures that meet or exceed the criteria for significance, having z_R scores in the range of $z_R = [-7.10, -1.65]$. These significant structures are distributed among 2,025 ORFs.

A total of 1,203 individual slippery sites in 751 ORFs are found to have more than one significant structure immediately downstream. Each of these statistically significant structures and the associated 5' slippery sites are considered candidate -1 PRF signals (cPRF) open for further investigation.

An interesting finding from this analysis is that statistically significant motif hits do not necessarily have low MFE values; a result that was previously shown to be true for structural RNAs in general (Le et al., 1989; Schultes et al., 1999; Seffens and Digby, 1999). We therefore sought to filter the list of putative -1 PRF signals further by comparing z_R scores and MFE values, which are shown to be only weakly correlated in Table 2 below with a correlation coefficient of 0.53. Comparing these two features is similar to an approach previously employed (Le et al., 2001). In Figure 8 below, energetically strong candidates with statistically significant predicted secondary structures are in the lower-left quadrant. From this analysis, 1,706 strong candidate signals were identified with significant $z_R \leq$ -1.65 and whose MFE values are in the lowest 25% (MFE $\leq$ 17.3 kcal/mol). These strong candidate signals are distributed among 1,275 individual ORFs, where 320 ORFs have two or more strong signals.

Nearly all PRFs result in termination

Analysis of the cPRF signals from the perspective of alternative recoding, shown in Figure 9 below, reveals that greater than 99% of the expected outcomes of -1 PRF would result in premature termination. The prevalence of out-of-frame termination signals is not unexpected. The average distance a ribosome can continue elongating in an alternative reading frame is ~6 codons in either the +1 or -1 frame for all CDS in yeast, as

shown in Figure 10. Only 10 -1 PRF signals out of 10,340 potential slippery sites shown in Figure 9 are predicted to bypass the normal zero-frame termination codon, i.e. -1 PRF in the viral context, and encode an alternative C-terminal extension. However, BLAST analyses (Altschul et al., 1990) revealed that none of these extensions are predicted to encode functional alternative protein domains[14]. This suggests that although potential -1 PRF signals are widespread in the yeast genome, they are almost uniformly predicted to direct ribosomes to a premature termination signals. Thus, these signals would be expected to target their native transcripts to the nonsense-mediated mRNA decay pathway (Plant et al., 2004).

Computational Identification of a Novel Viral -1 PRF Signal

As detailed in Chapter 1, programmed -1 ribosomal frameshift signals typically have a tripartite organization. From 5′ to 3′, these are composed of a heptameric slippery site, a spacer region, and a stable mRNA secondary structure, typically an mRNA pseudoknot. The structural arrangement of these features is shown in Figure 2 on page 15. A previous analysis of the SARS-CoV -1 PRF signal demonstrated that a sequence spanning nucleotide positions 13392 - 13472 satisfied these three requirements and was able to promote efficient -1 PRF in rabbit reticulocyte lysates (Thiel et al., 2003). The -1 PRF signal presented in that study contained a typical mRNA pseudoknot composed of two double-helical, Watson–Crick base paired stems connected by two single-stranded loops, shown in Figure 11A below.

[14] Data not shown.

The presence of a long, 29-nt loop 2 seemed to be unusual, prompting additional computational analyses in an effort to further define the structure of this mRNA pseudoknot. The nucleotide sequence suspected of featuring a -1 PRF signal between *ORF1a* and *ORF1b* was scanned by *RNAMotif* (Macke et al., 2001), using the same pattern-based description of -1 PRF signals detailed above in the Materials & Methods. As expected, a so-called slippery site "U UUA AAC" and a large H-type pseudoknot were identified; the two primary stimulating elements required for efficient ribosomal slippage. This analysis was subsequently coupled with *pknots* (Rivas and Eddy, 1999) as described above and the most thermodynamically stable structure was predicted for each *RNAMotif* hit found 3' of the identified slippery site. The predicted structure for the SARS-CoV frameshift signal was extremely stable, with a calculated MFE of −26.68 kcal/mol. The surprising result was that the 29-nt sequence designated loop 2 by Thiel *et al.* (2003) was predicted to form a third helix, nested within the sequences defined by stems 1 and 2, as shown in Figure 11B. Though a small, internally nested third helix (helix-3) has been shown to be present in the HIV-1 group O frameshift signal(Baril et al., 2003), such an extensive base pairing pattern had not been demonstrated for any other viral frameshift signal. To determine the statistical significance of this finding, a distribution of MFE values taken from 500 randomly shuffled SARS-CoV frameshift signals was created. Each of the randomly shuffled sequences was folded using *pknots*, as described above. The resulting normal distribution had a mean MFE of -21.12 ± 2.67

kcal/mol, revealing that the predicted three-stemmed pseudoknot structure of the native sequence is highly significant[15] with a z_R score of -2.05 and $p = 0.02$.

Conservation of mRNA Pseudoknots in Coronaviruses

To address the question of whether the potential to form a three-stemmed mRNA pseudoknot is unique to the SARS-CoV, structures were searched in all of the known viral -1 PRF signals listed in the RECODE 2003 database (Baranov et al., 2003), as well as the putative frameshift signals in all of the sequenced members of the Order *Nidovirales*[16]. The SARS-CoV frameshift signal itself is homologous to all of the nine other frameshift signals for coronaviruses whose genomes have been fully sequenced. A multiple sequence alignment of the ten coronavirus frameshift signals is presented in Figure 12. This shows that both stems 1 and 2 are highly conserved, with a strong conservation of base complementation in the cores of both stems 1 and 2; blue and red sequences respectively. This analysis also shows all of the coronavirus frameshift signals have the potential to form a third helix, although the structures and sequences are less well conserved, as shown in Figure 12 in green. In addition, the potential of sequences located approximately 200 nt downstream of the slippery site to form long-range "kissing loop" interactions with the 5′ half of stem 2 was previously noted for HCoV-229E (Herold and Siddell, 1993) and TEGV (Eleouet et al., 1995) viruses. The alignment in Figure 12 suggests this property was only conserved among the group 2 coronaviruses. A

[15] p value calculated using a one-tailed Student's t-test Devore J.L. (2000).

[16] This Order includes both coronaviruses and arteriviruses.

phylogenetic tree of the -1 PRF signals constructed from the multiple sequence alignment is presented in Figure 13. As expected, the group 1 and group 2 coronaviruses cluster together, and neither the SARS-CoV nor the avian infectious bronchitis virus (AIBV) frameshift signals cluster with either group. Of particular interest, however, is that very similar mRNA pseudoknot structures are predicted to occur within groups, but not between them.

The PRFdb

The PRFdb (http://dinmanlab.umd.edu/prfdb) is a publicly available database that stores the results of the bioinformatics data presented. This online resource allows interested researchers to search for and analyze candidate -1 PRF signals in the genome of *S. cerevisiae*. The PRFdb also contains a list of strong candidate -1 PRF signals that may warrant further empirical investigations.

Discussion

Programmed ribosomal frameshifting was first identified as a translational phenomenon in the *Rous sarcoma* virus over two decades ago (Jacks and Varmus, 1985). Since then, it has been shown to be a general mechanism of gene regulation utilized by a wide variety of RNA viruses (Baranov et al., 2002; Cobucci-Ponzano et al., 2005; Harger et al., 2002; Namy et al., 2004). Frameshifting has also been demonstrated to be functionally important for the expression of a growing list of prokaryotic (Blinkowa and Walker, 1990; Sekine and Ohtsubo, 1989; Tsuchihashi and Kornberg, 1990), archaeal(Cobucci-Ponzano et al., 2003), and eukaryotic genes (Shigemoto et al., 2001;

Wills et al., 2006). Thus, it is becoming increasingly apparent that PRF is a fundamental mechanism of post-transcriptional gene regulation and is present in every branch of the tree of life. The need to identify PRF signals in higher organisms has grown in importance as we have become more aware of their prevalence. In response, there have been numerous computational studies aimed at identifying PRF signals (Bekaert et al., 2003; Gao et al., 2003; Gurvich et al., 2003; Hammell et al., 1999; Harrison et al., 2002; Moon et al., 2004; Namy et al., 2003; Shah et al., 2002). Furthermore, at least one study has aimed to find PRF signals present in chromosomal intergenic regions (Bekaert et al., 2005). Each study has met with varying degrees of success, but empirical testing of predicted PRF signals suggest that there are indeed functional, and previously unannotated, PRF signals in a variety of contexts within the coding regions of genes derived from higher organisms. With the exception of the earliest study by Hammell *et al.*, all of these studies have focused on recoding in the "viral-context": *i.e.* they were aimed towards finding PRF signals predicted to direct ribosomes into a new reading frame so as to produce functional alternative C-terminal extensions of the native proteins. The study by Hammell *et al.*, was context neutral, focusing instead on searching for mRNA motifs that resembled known viral -1 PRF signals. While the current study revisits the original question posed by Hammell *et al.* (*i.e.* how often are functional -1 PRF signals present in the yeast genome?), it also asks an important second question: are genome encoded -1 PRF signals capable of promoting -1 PRF, and if so, how does this affect the expression of the mRNAs encoding them?

The outcome independent approach taken in this study searched for -1 PRF signals irrespective of the expected result of translation after the frameshift. A model for functional -1 PRF signals was developed from analysis of the RECODE database, and *RNAMotif* was employed to find all relevant motif hits in the yeast genome. A virtual avalanche of results was returned by the initial *RNAMotif* search, unveiling over 170,000 successful matches in the yeast genome. The resulting pattern matches were then each folded using *pknots*, a secondary structure prediction algorithm that "folds" RNA sequences so as to minimize the overall free energy value of each sequence. This software has the distinct advantage of being able to fold RNA sequences into pseudoknot conformations, a feature that is missing from the more popular *mfold* algorithm. This added benefit, which was essential for this study, comes with a significant penalty in terms of time and computational resources. Fortunately, *pknots* was run on a large super-computing cluster that allowed it to complete its calculations for every motif hit identified by *RNAMotif* in less than 6 months. While more time efficient heuristic algorithms have become more recently developed (Dirks and Pierce, 2004), *pknots* was the only algorithm capable of such computations available at the start of this project.

Once *pknots* had completed folding every motif hit, the dataset was reduced to approximately one-third of its original size by the automated removal and reduction of redundant structures occupying the same sequence space. This "boiling down" of the data resulted in a non-redundant dataset of some 66,842 structures located downstream (3') from 10,340 slippery sites. The "strongest" structure with the lowest MFE value immediately downstream of each of the slippery sites were dubbed "candidate -1 PRF

signals", cPRF, and were marked for further study. With over ten thousand candidate signals to consider for empirical testing, a second layer of predictive metrics was applied to aid in further filtering the dataset. Each of the candidate -1 PRF signals were randomly permutated and refolded 100 times to produce a distribution of randomized MFE values specific for each native sequence. The MFE value from each folded native sequence was compared each distribution to assess the significance, or uniqueness, of each fold and assigned a z_R score. The combination of MFE value and z_R score, coupled with the additional feature statistics from folding each structure permitted further reduction of the dataset to a smaller list of 1,706 strong candidate -1 PRF signals distributed among 1,275 ORFs.

As a proof of principle, these techniques were applied to identify the specific location and nature of the -1 PRF signal present between *ORF1a* and *ORF1b* in the recently sequenced SARS Coronavirus genome (Marra et al., 2003). Surprisingly, not only was the approach presented in this Chapter able to correctly identify the SARS-CoV -1 PRF signal, but it predicted a completely new structure unlike any previously characterized. Furthermore, these computational predictions were later validated and the novel three-stemmed pseudoknot structure was confirmed by nuclease mapping and 2D nuclear magnetic resonance studies (Plant et al., 2005). In short, this venture into studying the SARS-CoV frameshift signal demonstrated that the techniques presented in this Chapter were robust enough to legitimize empirical investigation of computationally identified -1 PRF signals in the yeast genome.

As a final note, the PRFdb was constructed to serve as a repository for all the predicted structures, slippery sites and statistical data gathered from analysis of the yeast genome. A website and interactive database designed to supplement this Chapter is accessible via the Internet, http://dinmanlab.umd.edu/prfdb. Currently, visitors are limited to searching for putative -1 PRF signals in the yeast genome only, or to download a batch text-file of candidate PRF signals. It is expected, however, that under the supervision of Dr. Jonathan Dinman, the methodology presented in this chapter will continue to evolve over time and that the scope will be expanded to include additional genomes including: seven additional budding yeast species, the human genome, and several other "model system" genomes.

Chapter 2 Tables

	Match Count	***stdev***	***p*-value**
S. cerevisiae	6,016	-	-
noBias	3,044	64.07	< 0.01
nShuffle	4,567	70.84	< 0.01
nBias	4,660	65.89	< 0.01
cShuffle	6,551	85.13	0.02
sBias	6,580	82.13	0.02
cBias	6,639	86.52	0.02
dnBias	6,774	88.16	0.01

Table 1: The Number of -1 PRF Motifs in Yeast is Not Random

The yeast genome has a significant number of putative programmed -1 ribosomal frameshift signals compared to randomized genomes created using any one of seven different randomization strategies when searched using *RNAMotif* with a defined descriptor of functional -1 PRF signals (Macke et al., 2001). The 6,016 matches for the *S. cerevisiae* genome represent the number of slippery sites that are followed by at least one pseudoknot motif. The *p*-value for each method is the result of a two-sample Student's t-test (Devore, 2000). The description of each randomization strategy can be found in the Materials & Methods section of Chapter 2 on page 22.

	Correlation		
	z_R	MFE	Length
Pairs	-0.32	-0.81	0.86
Length	-0.21	-0.73	
MFE	0.53		

Table 2: Correlation Coefficients of Secondary Structure Feature Statistics

Correlation of statistical features for the most stable (lowest MFE) predicted structures immediately 3' of the 10,340 individual slippery sites identified in the yeast CDS. z_R, predicted secondary structure significance of each native RNA sequence as compared to 100 shuffled permutations; MFE, minimum free energy value; Length, the total number of nucleotides folded by *pknots* (Rivas and Eddy, 1999) corresponding to the sequence window initially identified by *RNAMotif* (Macke et al., 2001); Pairs, the total number of AT, GC, or GU base pairs present in each predicted secondary structure.

	mean	stdev	[min, max]
Pairs	19.0	6.0	[1, 35]
Length	73.1	16.3	[14, 92]
MFE	-13.6	5.4	[-34.1, 1.5]
z_R	-1.2	1.3	[-7.1, 75.5]

Table 3: Descriptive Statistics for Predicted Structure Features

The mean, standard deviation, minimum and maximum values for features gathered from analysis of 10,340 of the most stable secondary structures found immediately 3' of each slippery site.

Chapter 2 Figures

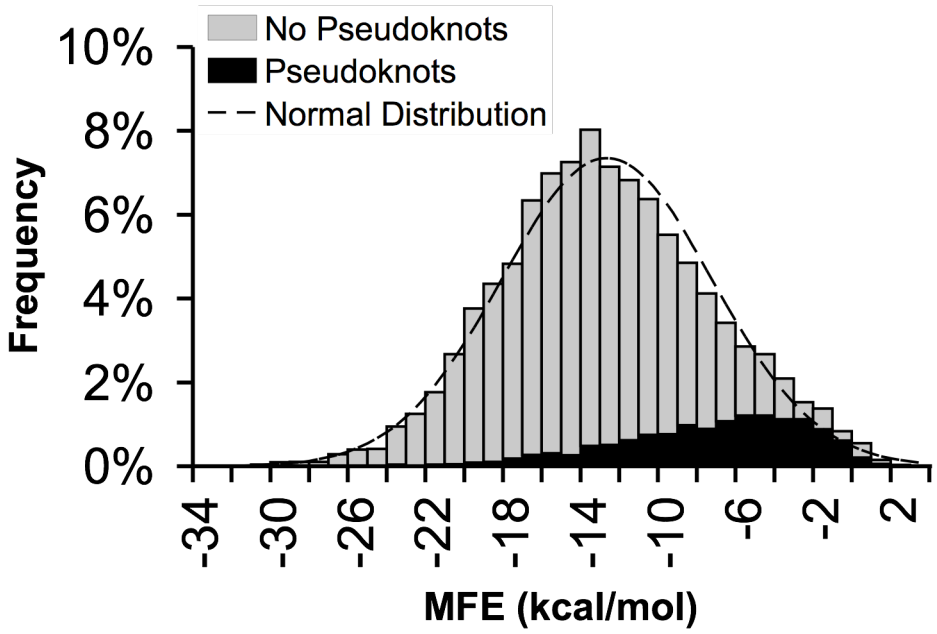

Figure 5: The Distribution of MFE Values

The distribution of minimum free energy values (MFE, in kcal/mol) for the lowest energy secondary structure predictions found immediately 3' of each slippery site in the PRFdb closely fits a normal distribution (dotted line) with a PPCC $\geq$ 0.98. Structures with pseudoknotted base pairing positions are shown in black.

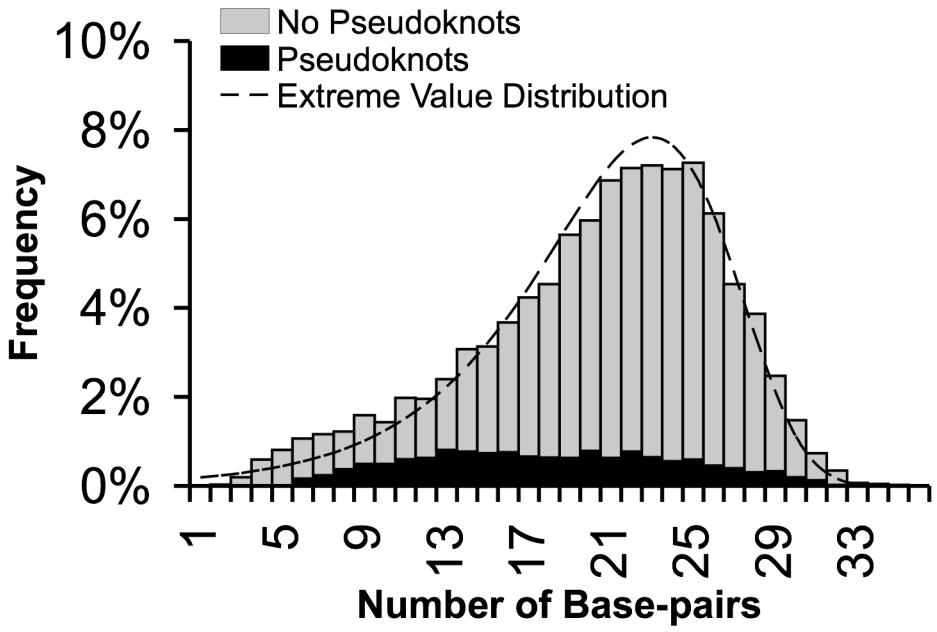

Figure 6: The Distribution of Base-Pair Counts

The number of base pairs for the most stable 3' predicted structures fits an extreme value distribution (dotted line) with a PPCC $\geq$ 0.97. Structures with pseudoknotted base pairing positions are shown in black.

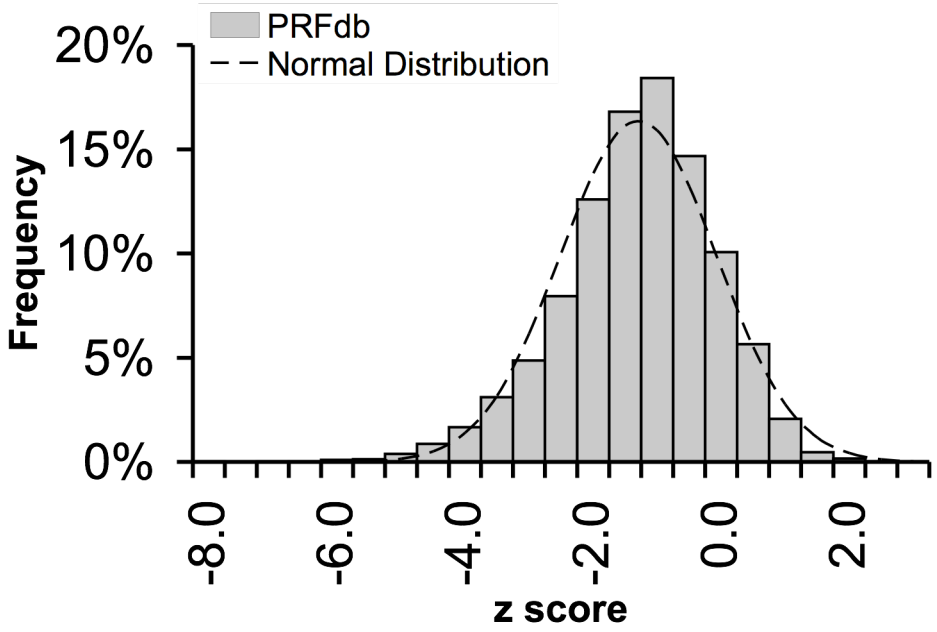

Figure 7: The Distribution of z_R Scores

z_R for all the strongest structures found immediately 3' of each slippery site in the PRFdb fits a normal distribution (dotted line).

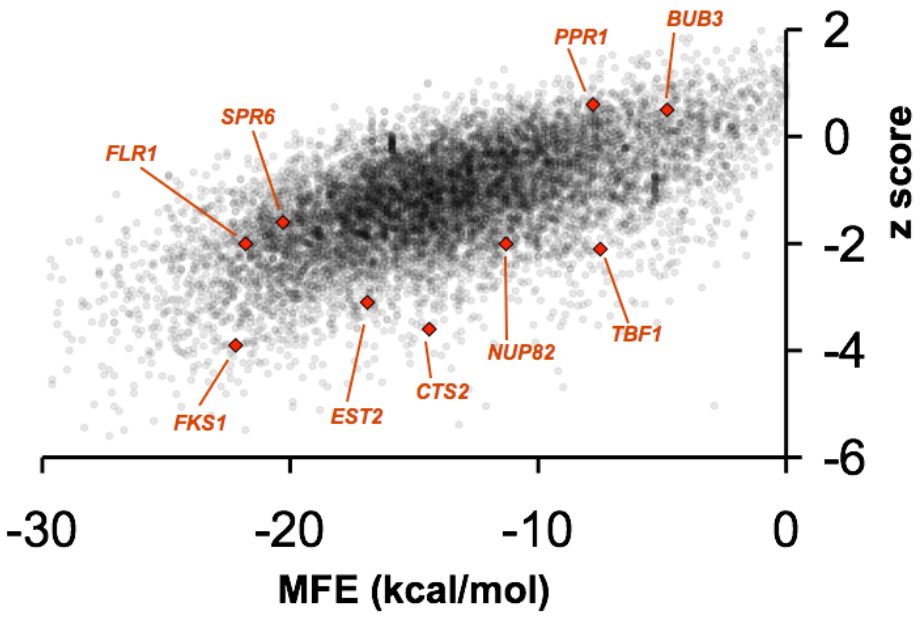

Figure 8: Scatterplot of MFE vs. z_R Score

Scatter plot of MFE values vs. z_R scores for 10,340 candidate -1 PRF signals demonstrates the weak correlation between these two feature statistics. The red diamonds and associated labels indicate the location and parental gene names of nine sequences empirically tested for frameshifting.

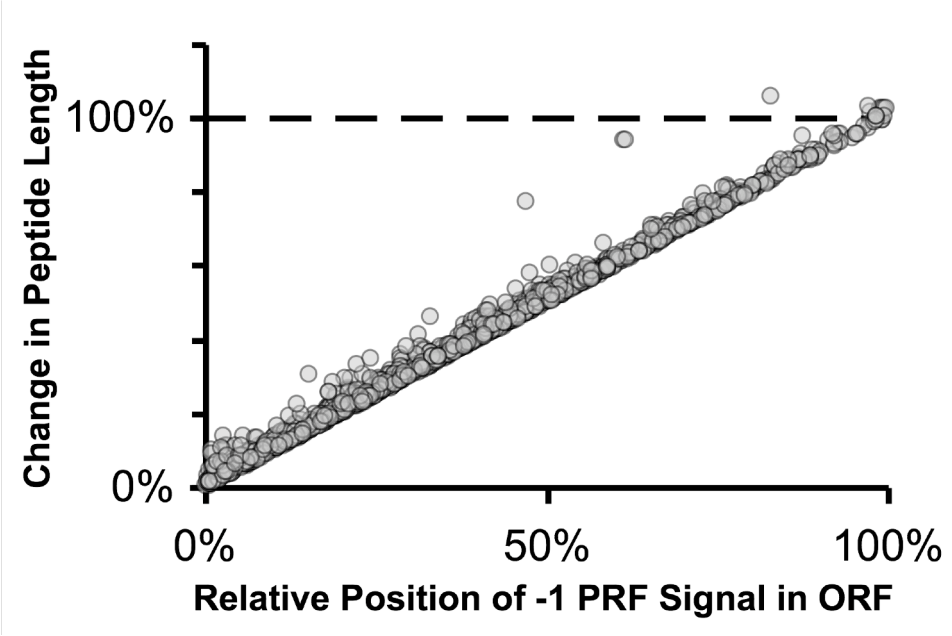

Figure 9: Frameshifting Outcomes Result in Premature Termination

The CDS of *S. cerevisiae* is not prone to lengthy out-of-frame translation. The relative positions of candidate -1 PRF signals from the start codon of each ORF compared to the expected overall change in peptide length if a frameshifting event were to occur. These data indicate that there are no examples in the PRFdb of frameshifting into a functional alternative protein coding sequence.

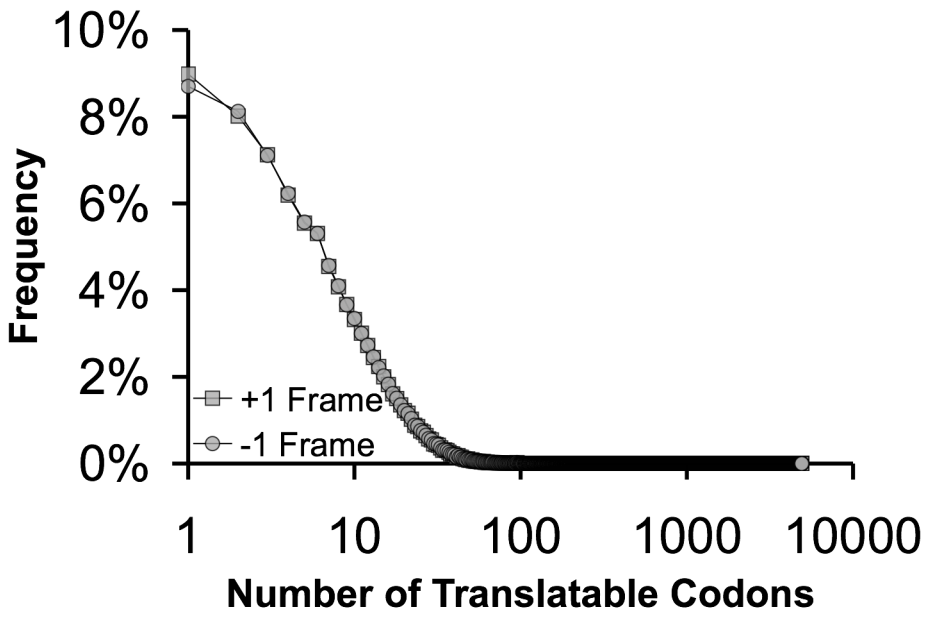

Figure 10: Frequency of Lengths for Translatable Out-of-Frame Codons

Frequency of lengths for out-of-frame translation for the CDS of *S. cerevisiae*. Independent of specific sites of translational frameshifting, the number of out of frame codons a ribosome can translate rarely exceeds ten.

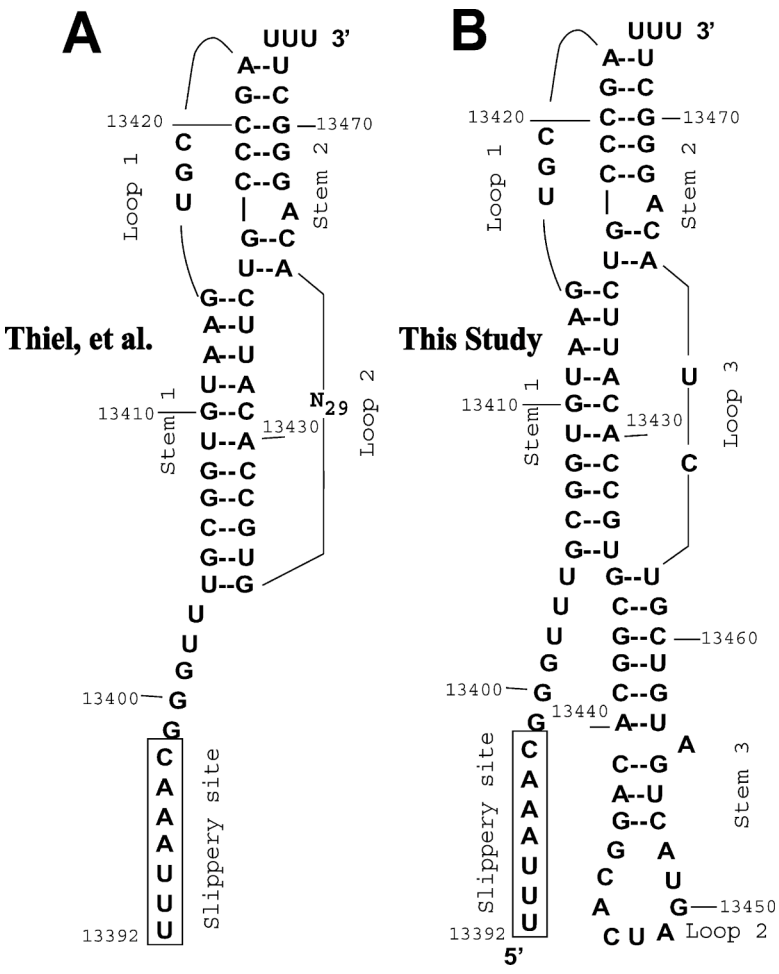

Figure 11: Different Representations of the SARS-CoV Frameshift Signal

Alternative views of the SARS-CoV -1 PRF secondary structure. This caption and the figure above have been adapted from Plant *et al.* (2005). (A) Two-stemmed H-type mRNA pseudoknot proposed by Thiel *et al.* (2003). (B) Three-stemmed mRNA pseudoknot structure investigated in this study.

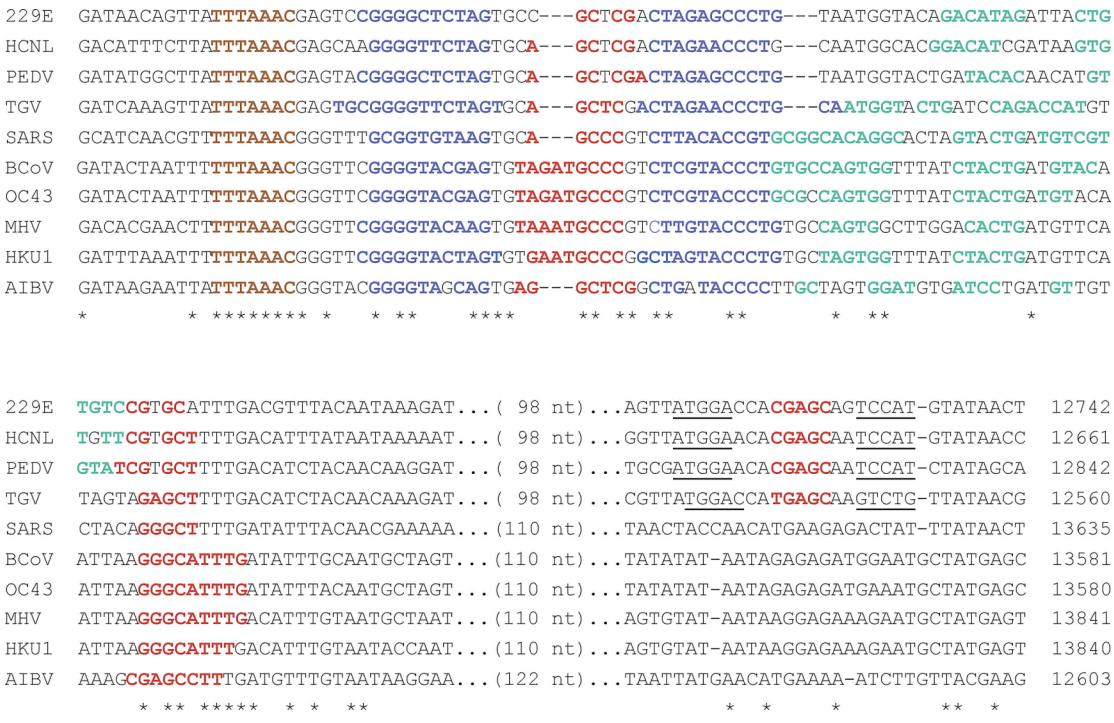

Figure 12: Multiple Sequence Alignment of Coronavirus −1 PRF Signals

Adapted from Plant *et al.* (2005). AIBV, avian infectious bronchitis virus; BCoV, bovine coronavirus; HCoV-229E, human coronavirus 229E; HCoV-HKU1; HCoV-NL63, human coronavirus NL63; HCoV-OC43, human coronavirus OC43; MHV, murine hepatitis virus; PEDV, porcine epidemic diarrhea virus; SARS, SARS coronavirus; TGV, transmissible gastroenteritis virus. Slippery sites are indicated in brown; dashes indicate gaps in the alignments; base pairing positions involved in the consensus first, second, and third helices are denoted by blue, red, and green nucleotides, respectively

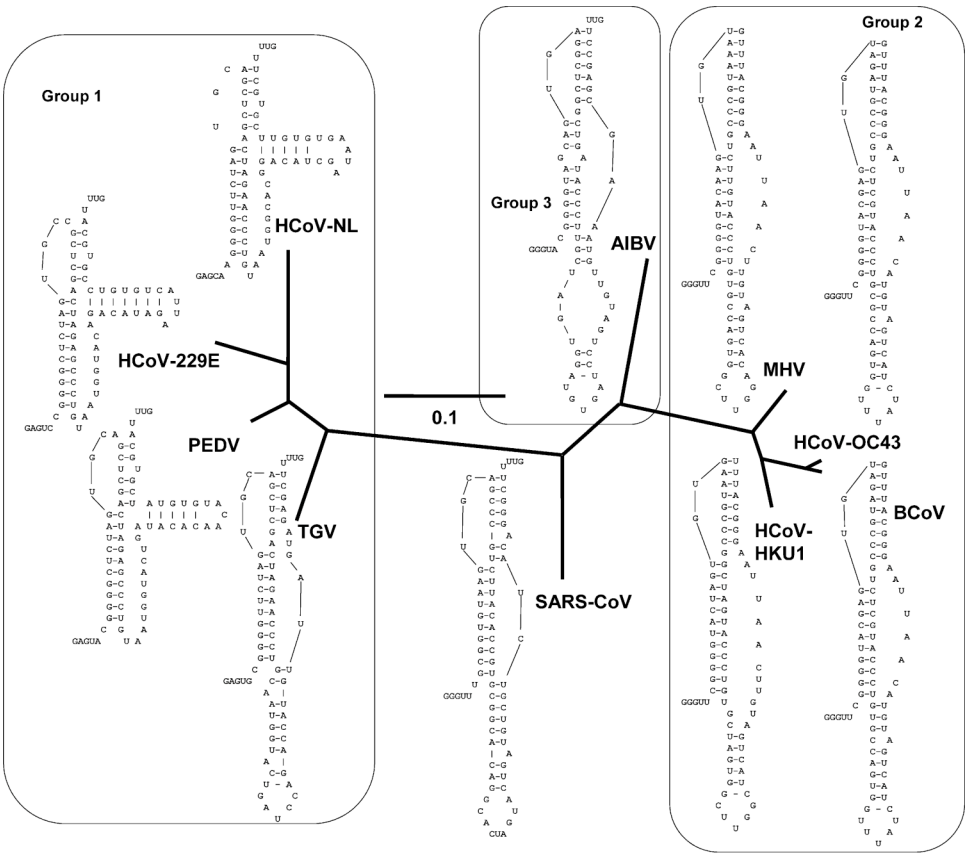

Figure 13: Phylogenetic Tree of Coronavirus −1 PRF Signals

Unrooted tree constructed based on the multiple sequence alignment from Figure 12. This caption and the figure above have been adapted from Plant *et al.* (2005).

Chapter 3: The Statistics of Bicistronic Assay Systems

Introduction

In the last decade, polycistronic reporter assays have generally become a mainstay in molecular biology. In particular, various bicistronic systems have been widely adopted as standard experimental techniques in the fields of translation initiation (Coleman et al., 2003; Imbert et al., 2003; Venkatesan et al., 2003), elongation (Meskauskas et al., 2003; Novac et al., 2004), recoding (Grentzmann et al., 1998; Harger and Dinman, 2003; Horsburgh et al., 1996; Kollmus et al., 1996a), and termination (Keeling et al., 2004). The ratiometric nature of the data produced from these experiments requires careful statistical treatment that is often lacking in the literature. The goal of the work presented in this Chapter was to propose a standardized statistical analysis pipeline for polycistronic reporter data and to provide researchers with a solid foundation on which to build their analyses.

Towards this end, we applied rigorous statistical methods to datasets originating from several sets of dual luciferase assays designed to measure the efficiency of -1 PRF signals in *Saccharomyces cerevisiae*. A -1 PRF signal is a *cis*-acting mRNA element that redirects translating ribosomes into a new reading frame after encountering a so-called

"slippery site"[17]. The efficiency of frameshifting depends on the PRF signal in question, typically between 1 - 10%, and can be measured *in vivo* using a dual luciferase reporter (DLR) assay system (Harger and Dinman, 2003).

Briefly, the dual luciferase assay (DLA) simultaneously measures the luminescence (e.g. expression) of both the *Renilla* and firefly luciferase enzymes synthesized from a single bicistronic mRNA. In an experimental frameshift reporter construct, the two genes are separated by a functional -1 PRF signal and the downstream firefly gene is placed into the -1 frame relative to the upstream *Renilla* gene. The relative expression of firefly to *Renilla* is normalized by a zero-frame control plasmid that lacks frameshift signal and has firefly in the zero frame. The resulting ratiometric data from our DLA is inherently sensitive to the propagation of error and therefore requires a careful statistical workup. The data are similar to the ratiometric data produced by other bicistronic assay systems despite the dissimilarity between the actual protocols producing it. This allows the methods presented in this report to be applied and extended to any polycistronic system that produces ratiometric data. Our analysis pipeline is designed to:

1. systematically identify and eliminate erroneous outliers;
2. confirm that the data is normally distributed;
3. establish the minimum number of replicates for each data set;
4. minimize the propagation of error when calculating ratiometric statistics; and
5. provide a solid statistical foundation for comparing datasets from different experiments.

[17] For reviews, see Harger et al. (2002) and Plant et al. (2003).

We have supplemented this study with a set of Microsoft Excel spreadsheets that automate the analysis and an online tutorial to help guide the reader through the analysis pipeline (http://dinmanlab.umd.edu/statistics). It is our hope that the methods presented in here will be adopted by researchers who utilize bicistronic reporters.

Materials & Methods

Genetic methods and plasmid construction

Escherichia coli strain *DH5α* was used to amplify plasmids, and *E. coli* transformations were performed using the high efficiency method(Inoue et al., 1990). YPAD and synthetic complete medium (H-) were used as described previously (Dinman and Wickner, 1994). Yeast strain JD932 *(MATa ade2-1 trp1-1 ura3-1 leu2-3,112 his3-11,15 can1-100*) (Peltz et al., 1999) was used for *in vivo* measurement of programmed -1 ribosomal frameshifting. Yeast cells were transformed using the alkali cation method (Ito et al., 1983). The dual luciferase reporter plasmid pJD375 was used as a zero-frame control as it does not contain a functional frameshift signal. The plasmid pJD376 was used as a positive control frameshifting and contains the -1 PRF signal from the endogenous yeast L-A virus (Harger and Dinman, 2003). Putative frameshift signals from *Saccharomyces cerevisiae* genes *BUB3* and *TBF1* were cloned into the multiple-cloning site (MCS) of pJD375. The construction of these two new plasmids, pJD519 and pJD478, was done using the following strategy. Each pair of forward and reverse oligonucleotides shown in Table 9 on page 132 were combined in 1:1 molar ratios, heated to 95ºC, and allowed cool to room temperature. The annealed doubled stranded DNA duplexes were

subsequently purified on 2% agarose by gel extraction[18]. Annealing the forward and reverse oligonucleotides left overhanging single stranded DNA complementary to *SalI* and *BamHI* restriction sites[19]. The resulting double stranded DNA was ligated into p2mci (Grentzmann et al., 1998), thus creating pJD519 and pJD478. The frameshift signal was sub-cloned as a *SalI-EcoRI* fragment into similarly digested pJD375. The ORF *1a-1b* frameshift signal from the SARS-associated Coronavirus was also cloned; oligonucleotides were annealed, gel purified and cloned into *BamHI* and *SacI* restricted p2mc (Grentzmann et al., 1998). This was further subcloned into a pJD375-based plasmid where the reading frame was corrected using site directed mutagenesis[20] to add a cytosine downstream of the *BamHI* restriction site to produce plasmid pSARS. *In vivo* dual luciferase assays for programmed -1 ribosomal frameshifting were performed as previously described in yeast strain JD1158 (Harger and Dinman, 2003), detailed in Table 7 on page 129. Luminescence readings were obtained using a TD20/20 luminometer[21]. Reactions were carried out using the Dual-Luciferase® Reporter Assay

[18] Gel extraction and purification of duplex DNA was done using the QIAEX II Gel Extraction Kit from Qiagen Inc., Valencia, CA.

[19] All DNA restriction digests were carried out using enzymes and standard protocols from Fermentas Inc., Hanover, MD. unless otherwise noted.

[20] The QuikChange® II XL Site-Directed Mutagenesis Kit from Stratagene Inc. (La Jolla, CA.) was used with a standard protocol.

[21] Turner Designs Inc. Sunnyvale, CA.

System[22]. Yeast cells were grown in the absence or presence of 20 μg/ml of anisomycin[23] in the appropriate media.

Calculation of Luminescence Ratios

The relative expression ratio of firefly luminescence (F_{RLU}) to *Renilla* luminescence (R_{RLU}) for the dual-luciferase reporter assay system is given by:

$$x_i = \frac{F_{RLU}}{R_{RLU}} \qquad [2]$$

where each x_i is the ratio obtained from an individual luminometer reading. For each of the frameshift reporters studied in this chapter, the values of x_1 - x_n comprise pooled datasets of size n. The statistics of this report are based on sets of ratiometric luminescence values (x_1-x_n) taken from multiple experiments and cell lysates.

Identification & Exclusion of Outliers

For outlier exclusion we first determine the boundaries of each of the four quartiles within each DLR dataset: the maximum (Q_{max}), the 75th percentile (Q_{75}), the median ($\tilde{x}$), the 25th percentile (Q_{25}), and the minimum values (Q_{min}) for each dataset of x_1- x_n is shown in Table 4. The fourth spread (fs) is calculated by

$$fs = Q_{75} - Q_{25} \qquad [3]$$

The standard upper and lower outlier boundaries are then calculated by

[22] Promega Corporation, Madison, WI.

[23] Anisomycin was obtained from Sigma-Aldrich, St. Louis, MO.

$$O_U = \tilde{x} + (1.5 \times fs) \quad [4]$$

$$O_L = \tilde{x} - (1.5 \times fs) \quad [5]$$

Data points that lie above or below these boundaries are considered outliers. For example, the solid data points in Figures 8 – 12 below are considered outliers that are excluded from further analysis (Devore, 2000).

Descriptive Statistics

We use standardized statistical expressions for the calculation of sample mean ($\bar{x}$), sample median ($\tilde{x}$), sample variance (s_{N-1}^2), sample standard deviation (s_{N-1}), and the standard error of the sample mean (s_e) from each single variable dataset of size n (Devore, 2000):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad [6]$$

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, \text{ if } n \text{ is odd} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, \text{ if } n \text{ is even} \end{cases} \quad [7]$$

$$s_{N-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \quad [8]$$

$$s_{N-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \quad [9]$$

$$s_e = \frac{s_{N-1}}{\sqrt{n}} \quad [10]$$

Probability Plot Correlation Coefficients

Probability plots were constructed and the corresponding normal probability plot correlation coefficients (*PPCC*) were determined for each set of DLR data (Chambers, 1983; Filliben, 1975). Briefly, the ratiometric values of firefly to *Renilla* luminescence (x_i - x_n, Eq. [2] above) are rank-ordered within each dataset and each ratio is assigned a standard normal observed, z_{Obs}, according to the following expression:

$$z_{Obs} = \frac{x_i - \bar{x}}{s_{N-1}} \qquad [\ 11\ ]$$

In addition, the expected *z*-score, z_{Exp}, for each value of x_i is calculated from the inverse standard normal distribution function for a given percentile rank[24] of x_i. The paired data, x_i and z_{Exp}, is then plotted on a graph. Linear least squares regression is used to plot a linear trend line fitted onto the data (Devore, 2000). The trend line's derived formula provides an expected ratio value (y_i) for each observed value (x_i) for a given value of z_{Exp}.

[24] The expected z-score for each data point is simply a measure of the standard deviation from the mean the i^{th} value has for any distribution of values that are rank-ordered, smallest to largest, with respect to one another. For example, if there were 100 values in a dataset that was perfectly normally distributed (PPCC = 1.0), then, after rank-ordering the dataset, the 50^{th} value would have a z-score of -0.01, the 51^{st} value a z-score of +0.01. Furthermore, the smallest and largest values in this dataset would have a z-scores of -2.33 or +2.33 respectively. Thus, by comparing the observed data to a hypothetical dataset of expected values it then becomes possible to observe the degree of correlation between the two. The correlation coefficient, PPCC, represents how well the observed data fits a idealized normal distribution with the same mean and standard deviation.

The correlation between the observed and expected values is given by the probability plot correlation coefficient, or *PPCC*:

$$PPCC(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad [\ 12\ ]$$

Where X and Y are the paired sets of expected and observed luciferase ratios, $\bar{x}$ and $\bar{y}$ are sample means, and a $PPCC = 1.0$ would indicate a perfect correlation between X and Y; i.e. a perfect normally distributed dataset. Another method for calculating the same *PPCC* value uses the correlation of paired values of z_{Obs} and z_{Exp} directly without the need for constructing a probability plot[25]. The *PPCC* is compared to a lookup table of critical values for a specified significance level[26] and sample size (n) in order to accept or reject the hypothesis that the data is normally distributed (Filliben, 1975). The *PPCC* critical values for rejection are shown in Table 13, Appendix E: PPCC Critical Values.

Minimum Sample Size

For a given confidence level (e.g. $\alpha = 0.05$) and predetermined limit on the numerical error, the minimum uncorrected sample size for a given dataset is calculated as follows (Devore, 2000):

[25] A tutorial of how to use an alternative, non-graphical method for PPCC calculation is available on the Internet at http://dinmanlab.umd.edu/statistics.

[26] For this study, a 5% significance level was set *a priori*.

$$\tilde{N} = \left\lceil \left(2z_{\alpha/2} \times \frac{s_{N-1}}{E} \right)^2 \right\rceil \qquad [13]$$

where $\tilde{N}$ is the minimum uncorrected sample size, $z_{\alpha/2}$ is the standard normal coefficient for a given value of $\alpha/2$, s_{N-1} is the sample standard deviation, and E is the amount of acceptable error in estimating the mean, usually 10% of $\bar{x}$). It was previously shown that the use of expression [13] for minimum sample size estimation substantially underestimates the number of trials needed for a given confidence interval (Kupper and Hafner, 1989). However, once $\tilde{N}$ is calculated, the minimum corrected sample size (N^*) can be found by consulting Table 14 in Appendix F: Minimum Corrected Sample Size. Each dataset must have no fewer than N^* replicates in order for further analysis to be well substantiated.

Ratiometric Statistics

The relative expression ($\bar{x}_R$) of each experimental reporter and its corresponding control is:

$$\bar{x}_R = \frac{\bar{x}_E}{\bar{x}_C} \qquad [14]$$

where in the case of dual luciferase assays $\bar{x}_E$ and $\bar{x}_C$ are the sample mean firefly to *Renilla* luminescence ratios for experimental and control reporters respectively. The estimated sample variance (s_R^2) for $\bar{x}_R$ is given by (Kendall et al., 1994):

$$s_R^2 = \frac{s_E^2}{\left(\bar{x}_C\right)^2} + \frac{\left(\bar{x}_E\right)^2 s_C^2}{\left(\bar{x}_C\right)^4} \qquad [15]$$

where $\bar{x}_E$ and $\bar{x}_C$ are the sample means from expression [6] above and the sample variances s_E^2 and s_C^2 are from expression [8]. Expression [15] makes the assumption that $\bar{x}_C \geq 0$ and the sample variances (s_E^2 and s_C^2) do not overlap zero (Kendall et al., 1994). Researchers should take care to make sure these are valid for each dataset. From expression [15], it follows that the sample standard deviation (s_R) of $\bar{x}_R$ is (Fersht, 1999):

$$s_R = \bar{x}_R \times \sqrt{\left(\frac{s_E^2}{\bar{x}_E}\right)^2 + \left(\frac{s_C^2}{\bar{x}_C}\right)^2} \qquad [16]$$

Finally, the standard error $s_e(\bar{x}_R)$ of $\bar{x}_R$ is calculated using the following expression, which correctly accounts for the propagation of error for independent samples of different sizes (Koopman, 2004):

$$s_e(x_R) = \bar{x}_R \times \sqrt{\frac{s_E^2 / N_E}{(\bar{x}_E)^2} + \frac{s_C^2 / N_C}{(\bar{x}_C)^2}} \qquad [17]$$

The number of replicates from each DLA dataset is specified by N_E and N_C for the experimental and control datasets respectively.

Comparing Datasets

We are often interested in finding the statistical significance of two differing experimental conditions, *a* and *b*. For bicistronic reporter assay data, it is appropriate to use an unpaired two-sample *t*-test since it is designed for small, continuous datasets (Devore, 2000). For example, when comparing experiments *a* and *b*, the degrees of freedom for the *t*-test, $v_{a,b}$, can be estimated, even if the number of replicates differs between the two datasets, using Expression 18:

$$v_{a,b} = \left\lfloor \frac{\left(\frac{s_{R_a}^2}{n_a} + \frac{s_{R_b}^2}{n_b} \right)^2}{\frac{\left(s_{R_a}^2 / n_a \right)^2}{n_a - 1} + \frac{\left(s_{R_b}^2 / n_b \right)^2}{n_b - 1}} \right\rfloor \quad [18]$$

The *t* statistic is then calculated by

$$t_{a,b} = \left| \frac{\bar{x}_{R_a} - \bar{x}_{R_b}}{\sqrt{\frac{s_{R_a}^2}{n_a} + \frac{s_{R_b}^2}{n_b}}} \right| \quad [19]$$

The values of $s_{R_a}^2$ and $s_{R_b}^2$ are the estimated sample variances from Expression [15] for each ratio $\bar{x}_{R_a}$ and $\bar{x}_{R_b}$. The sample sizes n_a and n_b correspond to the sample sizes of each dataset for the experimental frameshift reporters under each experimental condition *a* and *b*. Once the *t* statistic is computed, it can then be compared to a table of critical values, such as those found in (Devore, 2000), to either accept or reject each hypothesis.

Results

Data Visualization

The first step for data post-processing is to visualize the raw data. At the very minimum, good visualization techniques provide a qualitative understanding of the data's robustness before any descriptive or inferential statistics have been calculated. Here, the quality and linearity of the data can be ascertained immediately by plotting the relative luminescence units, *Renilla* RLU vs. firefly RLU, from each trial for a set of identical experiments such as those found in Figure 14 - Figure 17 below. The linear relationship

between *Renilla* and firefly expression in the context of the dual luciferase assay system has been well characterized and can be used as a first-hand measure of data quality (Harger and Dinman, 2003; Sherf et al., 1996). For example, the linearity of the assay can be clearly seen in the datasets from the "zero-frame" control (ZFC) pJD375 and in the L-A frameshift pJD376 reporters, despite the large differences in scale. Three outliers are also immediately and visually identifiable in the pJD519 frameshift reporter data shown in Figure 16. Furthermore, the pJD478 frameshift reporter data in Figure 17 demonstrates an unexpected scattering of the data.

Identification & Exclusion of Outliers

While some outliers occasionally can be identified at the visualization step, it is usually preferred to use a statistical based method to quantitatively rule them out. This is useful because standardized methods eliminate human bias across data sets, and they make no assumptions about the underlying distribution of the data. In our analysis, we use expressions [3], [4], and [5] above to exclude data beyond the bounds of the standard outlier boundaries O_U and O_L. Outliers identified using this method can be seen as solid data points in Figure 14 – 12. The resulting data is hereafter considered "trimmed" from a statistical point of view. This provides a simple and consistent method to identify outliers and, when applied uniformly, some data points can be identified as outliers that may have not been previously obvious from simple visual inspection, e.g. so called "hidden outliers", as is the case in Figure 14 below.

Descriptive Statistics

We employ standardized expressions for the usual descriptive statistics on each data set. This includes determining the sample mean ($\bar{x}$), sample median ($\tilde{x}$), sample variance (s^2_{N-1}), sample standard deviation (s_{N-1}), and the standard error of the sample mean (s_e) for samples of size n (Devore, 2000). Each of these statistics are presented in Table 4 on page 77 for the data relative to this chapter.

Probability Plots

After outliers have been excluded, the next step is to determine if the data is normally distributed. This is an essential step because all of the subsequent statistical measures depend on the assumption that the data comes from a normal distribution. A χ^2 goodness-of-fit test for normality to either reject or accept this hypothesis is often used for this calculation (Croarkin and Tobias, 2004). However, this is not an appropriate test for bicistronic data because:

1. there are typically too few data points for the χ^2 to be valid; and
2. whereas a χ^2 test is generally only appropriate for discrete data, bicistronic data is continuous.

A simple solution is to construct a histogram of the ratiometric data and visually inspect each set's distribution. While histograms provide a qualitative view of the data and a visual estimate for the goodness-of-fit of the data to a normal distribution, they do not provide a quantitative means for excluding, i.e. rejecting, any particular dataset.

For a more rigorous approach, a normal probability plot is created for each dataset and a normal probability plot correlation coefficient (*PPCC*) is determined (Chambers, 1983; Filliben, 1975). This coefficient allows for the formal rejection or acceptance of the hypothesis that a potentially small, continuous dataset is normally distributed by comparing the value of the *PPCC* to a table of critical values. A sufficiently high coefficient indicates that the data is normally distributed. Using this approach, the data collected from experiments done with plasmids pJD375, pJD376, and pJD519 is acceptable because each dataset has a PPCC that passes the critical value in Table 4. In contrast, the data from pJD478 is rejected because its PPCC value does not meet or exceed the critical value. Rejection can occur for many reasons, including poor-data collection, corrupted experimental conditions, or insufficient sample size. The probability plots for each of these datasets is shown in Figure 18 – 21. The strong "heavy tail" of Figure 21 is a clear indicator that the data are not normally distributed as it does not align with its expected values.

Minimum Sample Size

Experiments in molecular biology are often limited to three replicate trials due to limitations in time, financial resources or experimental complexity. Nonetheless, triplicate experiments do not typically satisfy the requirements of proper statistical analysis. Thus, the question remains as to how many replicate experiments should be done. Expression [12] is commonly used to answer this question (Devore, 2000), but Kupper and Hafner (20) previously showed that the use of this expression for sample size estimation greatly underestimates the number of trials needed for a desired confidence

interval, further exacerbating the problem. The *corrected* minimum sample size (N^*) can be found by consulting Table 14 in Appendix F: Minimum Corrected Sample Size, which assumes

1) the data are normally distributed;
2) a desired confidence level has been determined *a priori*; and
3) the amount of experimental error was decided *a priori*.

Generally, the acceptable amount of error for the estimate of the mean is 5% – 10% of its true value. For example, the pJD376 dataset has a sample mean ($\bar{x}$) and sample standard deviation (s_{N-1}) of 0.0263 and 0.0017 respectively. Our goal is to perform enough trials to be at least 95% confident that the sample mean is at least within 10% of the true value of the mean. Using expression [13], we find that the minimum uncorrected number of trials is $\tilde{N} = 7$. However, using Kupper & Hafner's method for sample size correction, the minimum corrected sample size is $N^* = 13$. With 40 samples, the pJD376 dataset is of sufficient size. The values of $\tilde{N}$, N^*, and the actual sample size N for each dataset are reported in Table 4.

Ratiometric Statistics

Once each ratiometric dataset has been trimmed of outliers, passed a test for normalcy, and found to be of sufficient size, it is then possible to begin calculating the ratiometric efficiency ($\bar{x}_R$) of an experimental reporter relative to that of its corresponding control reporter (see expression [14]). The reporters we use in our laboratory typically measure translational frameshifting: thus, in this case, $\bar{x}_R$ is a measure of the frameshift efficiency of the -1 PRF signal present in the experimental

DLR reporter constructs pJD376, pJD519, pJD478, pSARS. However, in other translational contexts $\bar{x}_R$ could be e.g., the frequency of IRES-promoted initiation, or read-through suppression. A serious pitfall associated with $\bar{x}_R$ is the potential for the propagation of error in its estimation since it is derived from a ratio of two estimates, $\bar{x}_E$ and $\bar{x}_C$ which are each ratios themselves. The correct reporting of the error on this measurement and its estimated variance should therefore be treated with care. Expressions [15]-[17] take the propagation of error into account and determine best-estimates for the sample variance s_R^2, sample standard deviation s_R, and the standard error $s_e(\bar{x}_R)$ of the sample mean $\bar{x}_R$. Each Exp. [15]-[17] assumes two, independent and normally distributed data sets that are related by the ratio $\bar{x}_R$ and each component dataset has potentially unequal sample sizes *N*. The importance of the estimation of s_R^2 in Exp. [15] cannot be overstated. This value is of particular importance when determining the statistical difference between two experiments; *e.g.* it is used in the *t*-test below.

Comparing Datasets

The final stage is to determine if two experiments, each with their own respective values of $\bar{x}_R$ and s_R^2, are statistically different. The published record of studies utilizing various bicistronic reporters shows a wide variety of methods including fold-change, *z*-tests, or χ^2-tests. For comparisons between datasets, a *z*-test is appropriate only for larger datasets with at least 40 samples each. Datasets for bicistronic reporter systems are usually not this large. Furthermore, a χ^2-test is inappropriate as it requires both large sample sizes and that the data be separated of into discrete categorical values. We instead

use the unpaired two-sample *t*-test, Exp. [18] and Exp. [19], since it is more appropriate for relatively small continuous datasets (Devore, 2000). The requirements of this test are that the data must be normally distributed and independent, which are satisfied by the bicistronic assay datasets presented here. The hypothesis tested against states that two datasets (*X* & *Y*) come from the same population. A rejected hypothesis therefore affirms that the two datasets are indeed statistically different at some predefined confidence level. In the context of the experiments presented here, a 95% confidence level was used. The *p*-value obtained from this test is an estimation of the probability of an incorrect conclusion (Devore, 2000).

Two Working Examples

Our first set of frameshift reporters comprises pJD376, pJD519, and pJF478, each of which was compared to the zero-frame control reporter pJD375 to measure the efficiency of programmed -1 frameshifting attributable to each frameshift signal. Each of these experimental reporters contains a -1 PRF signal that was either previously characterized (Dinman et al., 1991) or that was computationally identified. Our results show that the PRF signals in reporters pJD376 and pJD519 are "well-behaved" in that they pass several tests for statistical reliability. The frameshift efficiency of pJD376, the L-A virus *gag-pol* PRF signal, is shown in Table 4 to promote relatively high levels of frameshifting (8.0% ± 0.2%). By contrast, the PRF signal in from the *BUB3* gene, (pJD519) is shown to be less efficient with only 0.70% ± 0.02% frameshifting. Furthermore, the efficiency of frameshifting for signal present in pJD478 from the *TBF1* gene was not calculated in Table 4 because the data itself failed two important statistical

reliability tests[27]. Without the techniques presented in this report, the recoding efficiency of pJD478 may have be erroneously calculated and subsequently reported incorrectly as being ~1.86%[28]. This important, and often overlooked, aspect of the reliability testing experimental measurements demonstrates the importance of quantitatively determining the linearity, minimum sample size, and normalcy of each dataset studied.

In our second example, we begin with two dual-luciferase reporters: a zero-frame control, pJD375, and a frameshift reporter, pSARS, representing the functional SARS-associated Coronavirus ORF *1a-1b* frameshift signal identified in Chapter 2. The experiment is designed to study the efficiency of ribosomal frameshifting in the presence or absence of the drug anisomycin. This well characterized translational inhibitor is known to suppress programmed -1 ribosomal frameshifting *in vivo* (Dinman et al., 1997). The initial dataset of raw luminescence values for each construct (with and without drug) was plotted as described in the Materials & Methods and the raw data was found to be linear. Outliers were then identified and excluded. Furthermore, each data set passed the *PPCC* test for being normally distributed. The values of $\bar{x}_R \pm s_e(\bar{x}_R)$, i.e. -1 PRF efficiency, are 2.6% ± 0.2% in the presence of 20 µg/mL anisomycin and 3.3% ± 0.5% in

[27] Subsequently, pJD478 was reampliflied from freshly transformed E. coli and 10 individual clones were resequenced by Macrogen Inc.. Furthermore, resequencing of the pJD478 plasmid sample used for the yeast transformants in this chapter showed degenerate sequence present in the -1 PRF sequence region, most likely as a result of contamination. Data from all subsequent experiments with pJD478 presented in Chapter 4 was found to be "well-behaved" in that it was normally distributed and passed the PPCC test.

[28] The -1 PRF efficiency of pJD478 was later determined to be ~5.2%; see Chapter 4.

its absence. If we had simply relied on "fold-change" statistics, we would have only reported an approximate 21% reduction in recoding efficiency. Furthermore, if we had calculated the frameshift efficiencies for pSARS with and without anisomycin using previously published techniques (Grentzmann et al., 1998; Harger and Dinman, 2003), the observed ~21% reduction in PRF efficiency would have been determined to be statistically insignificant (p = 0.804, data not shown). However, using the unpaired two-sample t-test in Exp. [18] and [19], we find t = 8.92 with 18 degrees of freedom (v = 18) for the effects of anisomycin on -1 PRF. A significance level of α = 0.001 indicates a critical value of t = 3.92 (Devore, 2000). Thus our results soundly reject the null hypothesis in favor of the alternative hypothesis that anisomycin affects programmed -1 ribosomal frameshifting. Numerical computation of the p-value of this finding yields $p = 5.04 \times 10^{-8}$; an highly significant result.

Online Tutorial

A tutorial detailing each of the statistical methods presented in this report has been made available on the Internet at http://dinmanlab.umd.edu/statistics. The tutorial provides step-by-step instructions and screen-shots on how to use these methods using Microsoft Excel.

Discussion

In this chapter, a statistical analysis pipeline has been outlined for ratiometric data potentially derived from a wide variety of polycistronic reporter assay systems. As an example, the methods outlined above were successfully applied to eight datasets

originating from a series of dual luciferase assays designed to measure programmed -1 ribosomal frameshifting. The reporter plasmids vary only in the nature of recoding element positioned between the *Renilla* and firefly open reading frames. This statistical analysis pipeline can be applied to other dual reporter systems and easily extended to any polycistronic assay system that relies on ratiometric data. The importance of the proper statistical analysis of any dataset cannot be overstated. At a minimum, this chapter brings to light the statistical issues surrounding bi- or polycistronic reporter data and opens the door to more rigorous treatment of this particular data type. It is hoped that the synthesis of methodologies presented here will serve as a white paper for researchers who utilize polycistronic reporter systems in general. Addressed below are several key features for analysis of bicistronic data in particular and a summary of the findings.

First, the nature of most bicistronic reporter assays present researchers with two components of information for each experiment that are further combined into a ratio. The data are most often reported as a ratio of gene X to gene Y expression. The goal is usually to measure the expression ratio of genes X/Y in an experimental construct and observe any differences in ratio of genes X/Y expression compared with a known control. Since the data are both ratiometric and continuous in nature, propagation of error in the datasets is a primary issue that must be addressed carefully. We address this issue with expressions [15] – [17] for estimates of the sample variance, sample standard deviations, and the standard error of the sample mean for a ratio of two normally distributed sample means. Only once an appropriate measure of the combined variance and corresponding

error is found, is it then possible to determine if two independent datasets are statistically different.

Second, methods designed to systematically rule out certain data points as outliers have largely gone unreported in the life sciences literature; suggesting that outlier data is all too often dealt with on an *ad hoc* basis. Outliers can severely impact the quality and subsequent analysis of any dataset. Thus, their systematic exclusion should be an important first step in any analysis pipeline. Presented in this chapter is a simple, standardized method for outlier exclusion that makes no assumptions about the underlying distribution of the data using Exp. [3] – [5]. By exploiting the property of fourth spreads (Exp. [3]), we are able to systematically exclude data points that are significantly above or below the median. This method does not necessarily always result in the exclusion of data; frequently the maximum or minimum values for any dataset are well within the outlier boundaries. The net result is a robust, trimmed dataset that is less affected by the presence of a few outliers; a vulnerability in means calculated from untrimmed datasets.

Third, a common assumption is that data are normally distributed. This is necessary because common statistical analyses rely on this assumption in order to remain valid. However, biological data is often *not* normally distributed due to the tendency of living cells to either maximize or minimize the efficiency of any given process. Surprisingly, there has not been a single publication utilizing a bicistronic reporter assay system that has reported attempts to check the validity of this assumption. This chapter presents a procedure for constructing probability plots of each dataset, and a statistically

sound method for determining the normalcy of the data using probability plot correlation coefficients. No subsequent statistical analysis that is fundamentally based on the properties of a normal distribution would be valid without first confirming that the data actually fits a normal distribution. Failure to do this quantitatively could lead researchers to reach false conclusions.

Fourth, as a rule-of-thumb molecular biology experiments are typically carried out in triplicate. This is often a reality that is expected and unavoidable because many experiments are time consuming, expensive, or both (e.g. blots, gel shift assays, etc). We suggest that the "Three's a Charm" rule-of-thumb should be reconsidered when experiments are relatively simple and rapid. Most bicistronic reporter assays fit these criteria because they usually take advantage of the specific activity of a pair of easily assayable enzymes. In expression [13], we present a straightforward method to calculate the minimum corrected sample size (N^*) needed to achieve a specific level of confidence in the results. The researcher needs only to decide *a priori* what the acceptable level of error is for their data.

Using a metric to determine minimum sample size, however statistically sound, may seem unreasonable or simply cost prohibitive to some, particularly for smaller labs with limited resources. However, consider the following example. Typically, with respect to the dual luciferase assay system in *E. coli* (Grentzmann et al., 1998) or *S. cerevisiae* (Harger and Dinman, 2003), it is not unusual for cell lysates to be collected over a course of three days and for three luminescence readings (firefly and *Renilla*) to be averaged together on each day. This produces only a single luminescence ratio each day for each

reporter. Not only does this approach inadvertently create another layer of error propagation, i.e. an average of averages at the experiments end, but it is both cost and time prohibitive if the goal is to gather enough data points to satisfy Kupper & Hafner's test for minimum corrected sample size. A suitable compromise is to pool individual reads from each lysate into a larger data set before excluding outliers and calculating any statistics. In this case, the scenario outlined above would produce 9 data points each for *Renilla* and firefly luciferase; 3 for each cell lysate for each of 3 days. If the cell types, reporters used, and experimental conditions are identical, pooling the data in this way builds a rigorous data set that is more resistant to experimental bias. Furthermore, if three separate cell cultures were grown in parallel on each day, then 27 data points would then be collected for each experimental condition in same amount of days. By pooling the raw data together, it becomes possible to build a larger data set in less time.

The rigorous statistical analysis presented here also has significance for the field of frameshifting because the confirmation that anisomycin inhibits this process is important in helping to define the mechanism of PRF. We previously proposed a mechanistic model based on structural and biochemical data in which the -1 frameshift occurs after accommodation of the aminoacyl-tRNA (aa-tRNA) into the ribosomal A-site (the A/A hybrid state), and prior to peptidyltransfer (Plant et al., 2003). Recently, another group suggested that the shift occurs prior to accommodation when the aa-tRNA occupies the A/T hybrid state, i.e. while the anticodon of the aa-tRNA is in the decoding center A-site, but the 3' acceptor end has not yet been positioned into the peptidyltransferase center (Leger et al., 2004). Anisomycin binds in the A-site of the peptidyltransferase

center (Hansen et al., 2003) inhibiting binding of the acceptor end of the aa-tRNA into the peptidyltransferase center (Carrasco et al., 1973; Grollman, 1967). The observation that -1 PRF is inhibited by anisomycin is consistent with our model in that inhibiting the formation of the proposed substrate for the shift (i.e. inhibiting formation of the aa-tRNA in the A/A hybrid state) decreased the frequency of the reaction. In contrast, anisomycin does not affect the interaction of the aa-tRNA anticodon with the decoding center, i.e. does not impact on the formation of the aa-tRNA in the A/T hybrid state, and would not be predicted to affect -1 PRF if this were the substrate for the shift. In sum, the application of the rigorous statistical analyses to genetic data reinforces prior structural and biochemical analyses, strengthening the argument that programmed -1 ribosomal frameshifting occurs after accommodation of the aa-tRNA into the A/A hybrid state.

Chapter 3 Tables

Table 4: DLR Data for Development of the Statistical Analysis Protocol

Summary of the dual luciferase reporter assay datasets. C_1 and C_2, zero-frame control reporters; F_1 - F_4, frameshift reporters; Q_{max}, maximum ratio; Q_{75}, 75th percentile; Q_{25}, 25th percentile; Q_{min}, minimum ratio; $\tilde{x}$, median; *fs*, fourth spread; O_U, standard upper outlier boundary; O_L, standard lower outlier boundary; *PPCC*, normal probability plot correlation coefficient; $\bar{x}$, sample mean; s^2_{N-1}, sample variance; s_{N-1}, sample standard deviation; s_e, standard error of the sample mean; $\tilde{N}$, minimum uncorrected sample size; $N*$, minimum corrected sample size; N, actual sample size; $\bar{x}_R$, estimate of sample mean for the ratio of the $\bar{x}$ of experimental frameshift reporter to $\bar{x}$ of control reporter (i.e. frameshift efficiency); s^2_R, sample variance for $\bar{x}_R$; s_R, sample standard deviation of $\bar{x}_R$; $s_e(\bar{x}_R)$, standard error of the sample mean $\bar{x}_R$; n/c, no calculated.

Table 4: DLR Data for Development of the Statistical Analysis Protocol

(continued from previous page)

	Sample Data				No Drug		20 µg/mL Anisomycin	
	pJD375	**pJD376**	**pJD519**	***pJD478***	***C_2***	***F_4***	***C_2***	***F_4***
Q_{MAX}	0.485	0.033	0.0028	0.078	0.347	0.0119	0.346	0.0091
Q_{75}	0.379	0.028	0.0025	0.025	0.331	0.0112	0.333	0.0087
$\tilde{x}$	**0.320**	**0.027**	**0.0023**	**0.011**	**0.310**	**0.0104**	**0.329**	**0.0086**
Q_{25}	0.285	0.025	0.0021	0.005	0.307	0.0099	0.324	0.0083
Q_{MIN}	0.252	0.023	0.0005	0.004	0.260	0.0076	0.310	0.0078
fs	0.092	0.003	0.0004	0.020	0.024	0.0013	0.009	0.0004
O_U	0.459	0.031	0.0029	0.041	0.346	0.0123	0.343	0.0091
O_L	0.182	0.022	0.0017	0.000	0.274	0.0085	0.316	0.0080
Outliers?	3	3	3	6	3	1	3	4
$\bar{x}$	**0.330**	**0.026**	**0.002**	**0.013**	**0.316**	**0.011**	**0.327**	**0.009**
s^2_{N-1}	$2.9\text{x}10^{-3}$	$3.0\text{x}10^{-6}$	$5.9\text{x}10^{-8}$	$1.1\text{x}10^{-4}$	$2.9\text{x}10^{-4}$	$6.6\text{x}10^{-7}$	$3.7\text{x}10^{-5}$	$5.4\text{x}10^{-8}$
s_{N-1}	$5.4\text{x}10^{-2}$	$1.7\text{x}10^{-3}$	$2.4\text{x}10^{-4}$	$1.0\text{x}10^{-2}$	$1.7\text{x}10^{-2}$	$8.1\text{x}10^{-4}$	$6.1\text{x}10^{-3}$	$2.3\text{x}10^{-4}$
s_e	$5.9\text{x}10^{-3}$	$2.8\text{x}10^{-4}$	$4.7\text{x}10^{-5}$	$1.4\text{x}10^{-3}$	$4.4\text{x}10^{-3}$	$2.0\text{x}10^{-4}$	$1.6\text{x}10^{-3}$	$6.2\text{x}10^{-5}$
$\tilde{N}$	42	7	18	939	5	10	1	2
$N*$	54	13	26	433	11	17	6	7
N	84	40	27	51	15	17	15	14
Sufficient Sampling?	YES	YES	YES	NO	YES	YES	YES	YES
PPCC	0.98	0.99	0.99	0.92	0.93	0.99	0.98	0.97
Cut-off	0.98	0.96	0.94	0.97	0.91	0.92	0.91	0.90
Normal?	YES	YES	YES	NO	YES	YES	YES	YES
$\bar{x}_R$	**-**	**0.080**	**0.007**	**n/c**	**-**	**0.034**	**-**	**0.026**
s^2_R	-	$2.0\text{x}10^{-4}$	$1.8\text{x}10^{-6}$	n/c	-	$9.8\text{x}10^{-6}$	-	$7.4\text{x}10^{-7}$
s_R	-	$1.4\text{x}10^{-2}$	$1.4\text{x}10^{-3}$	n/c	-	$3.1\text{x}10^{-3}$	-	$8.6\text{x}10^{-4}$
$s_e(\bar{x}_R)$	-	$1.7\text{x}10^{-3}$	$1.9\text{x}10^{-4}$	n/c	-	$7.8\text{x}10^{-4}$	-	$2.3\text{x}10^{-4}$

Chapter 3 Figures

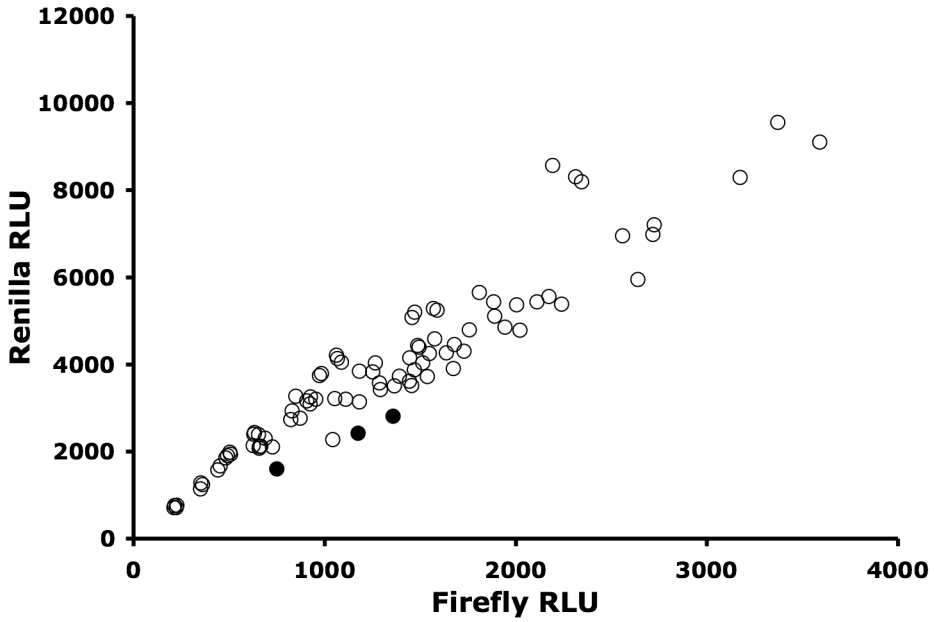

Figure 14: Comparing Luminescence Values from pJD375

Visualization of *Renilla* and firefly luminescence data from control reporter pJD375 in a wild-type yeast strain JD1158. Outliers are shown by solid data points.

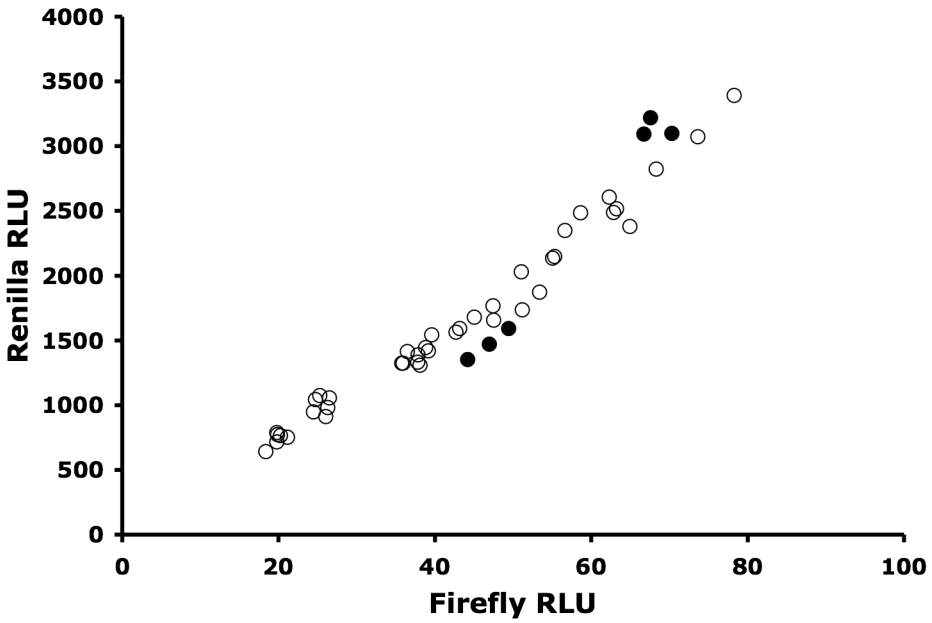

Figure 15: Comparing Luminescence Values from pJD376

Visualization of *Renilla* and firefly luminescence data from L-A viral frameshift reporter, pJD376, in a yeast strain JD1158. Outliers are shown by solid data points.

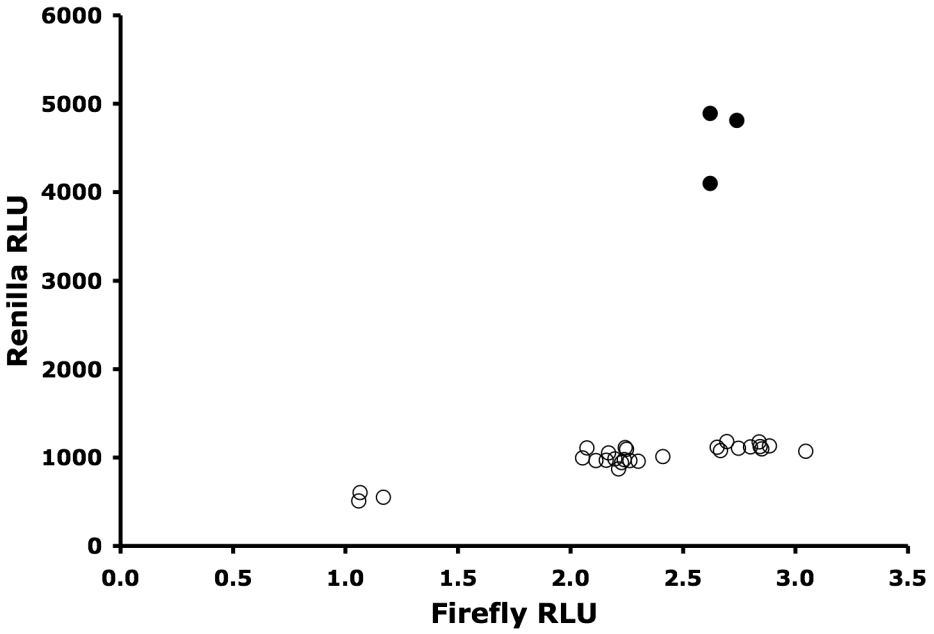

Figure 16: Comparing Luminescence Values from pJD519

Visualization of *Renilla* and firefly luminescence data using a putative -1 PRF signal from *BUB3* (pJD519) in a strain JD1158. Outliers are shown by solid data points.

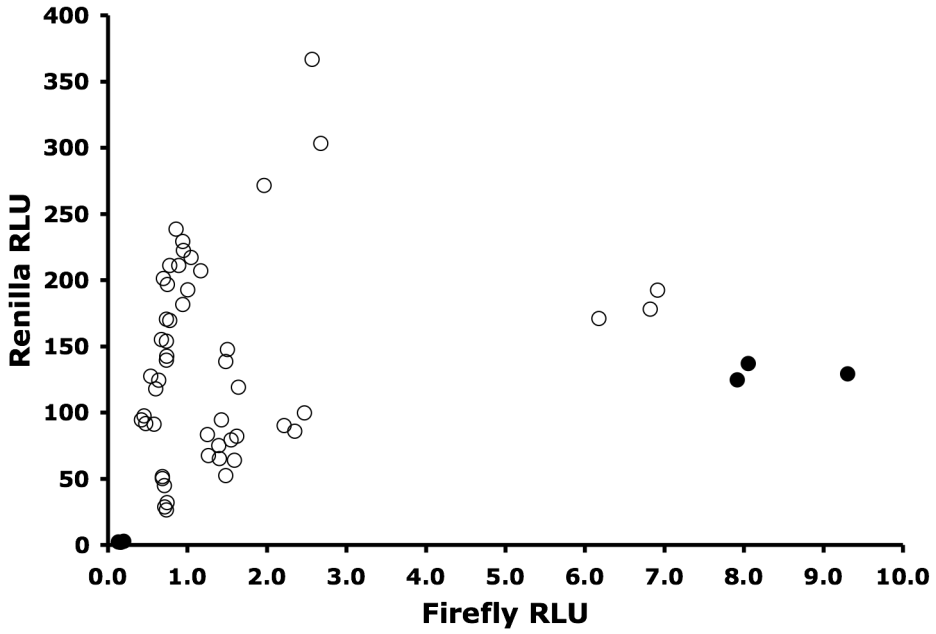

Figure 17: Comparing Luminescence Values from pJD478

Visualization of *Renilla* and firefly luminescence data using a putative -1 PRF signal from *TBF1* (pJD478) in a strain JD1158. Non-parametrically determined outliers are shown by solid data points; although it is apparent that these data display a high degree of nonsystematic error.

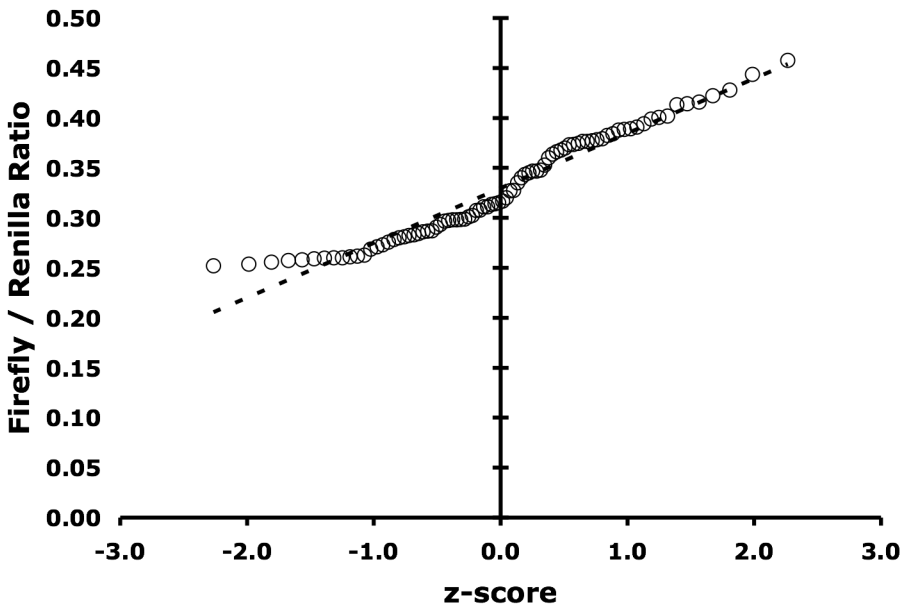

Figure 18: Probability Plot of Luminescence Ratios for pJD375

Visualization of *Renilla* and firefly luminescence ratios from control reporter pJD375 in wild-type yeast strain JD1158. After outlier exclusion, the linearity of the above probability plot demonstrates that the ratiometric values for pJD375 are normally distributed since the PPCC for this data is 0.98, which is at the cut-off level for normalcy.

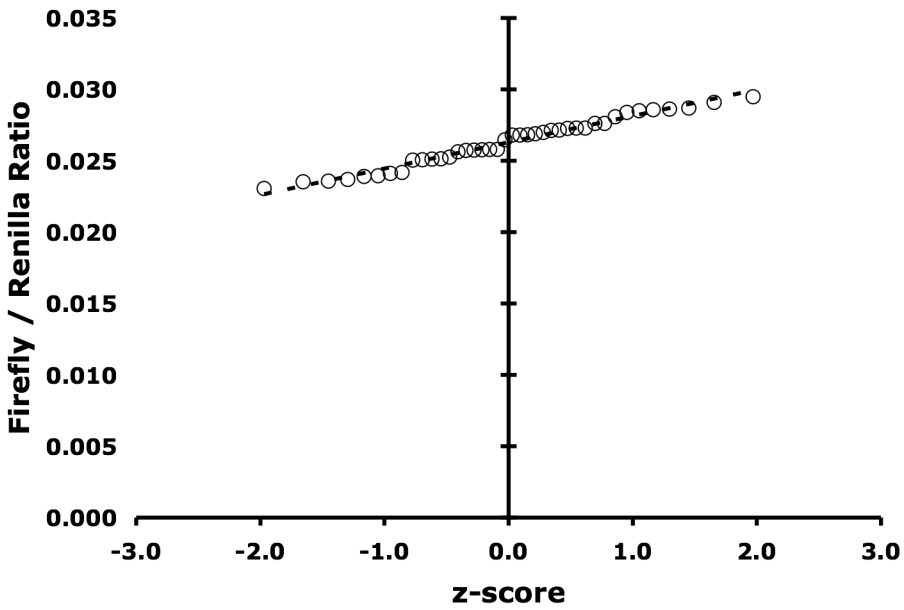

Figure 19: Probability Plot of Luminescence Ratios for pJD376

Visualization of *Renilla* and firefly luminescence ratios from the L-A frameshift reporter pJD376 in strain JD1158. After outlier exclusion, the linearity of the above probability plot demonstrates that the ratios for pJD376 are normally distributed.

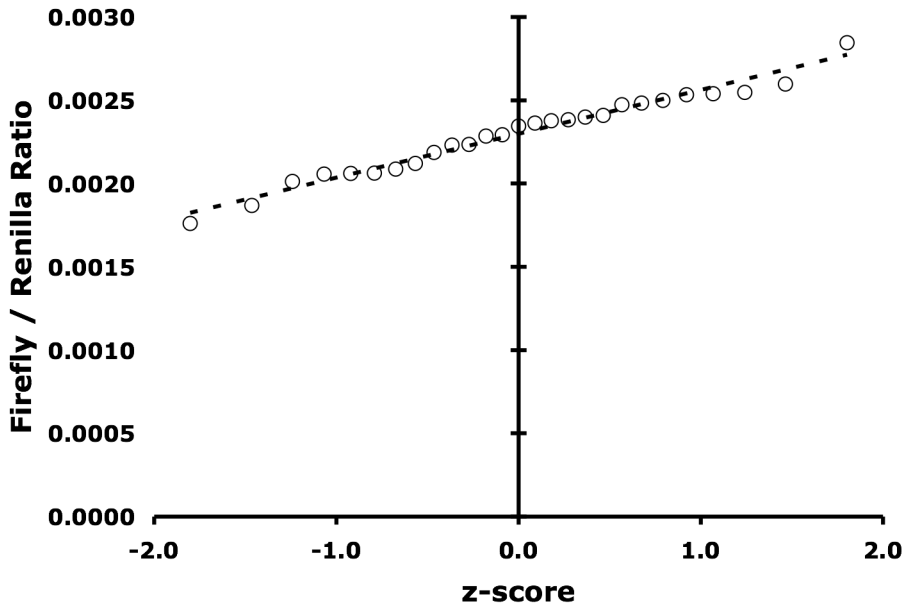

Figure 20: Probability Plot of Luminescence Ratios for pJD519

Visualization of *Renilla* and firefly luminescence ratios from a putative -1 PRF signal cloned from *BUB3* (pJD519) and expressed in strain JD1158. After outlier exclusion, the linearity of the above probability plot demonstrates that the ratios for pJD519 are normally distributed.

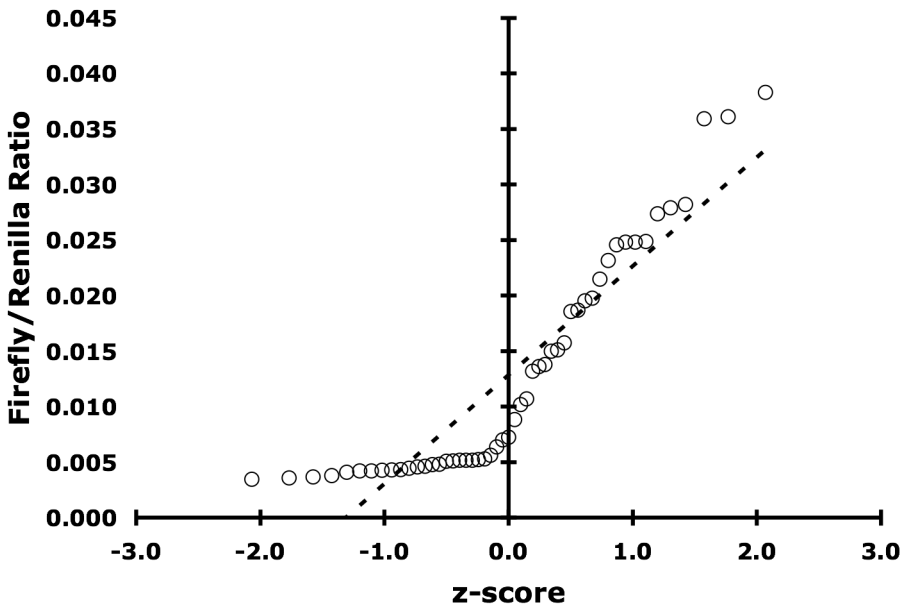

Figure 21: Probability Plot of Luminescence Ratios for pJD478

Visualization of *Renilla* and firefly luminescence ratios from a putative -1 PRF signal cloned from *TBF1* (pJD478) in strain JD1158. After outlier exclusion, the above probability plot systematically demonstrates that the data for pJD478 is *not* normally distributed, and therefore should be used for further consideration. On Table 4 above is a PPCC = 0.92 for this data, but with a cut-off of 0.97 it fails to meet the minimum criteria for being normally distributed.

Chapter 4: Computationally Identified -1 PRF Signals Can Promote Efficient Frameshifting & Function as mRNA Destabilizing Elements

Introduction

The results of the bioinformatics search presented in Chapter 2 provided strong evidence that functional -1 PRF signals are widespread throughout the yeast genome. Coupled with the results of several additional independent bioinformatics studies (Bekaert et al., 2003; Bekaert et al., 2005; Cobucci-Ponzano et al., 2005; Gurvich et al., 2003; Hammell et al., 1999; Manktelow et al., 2005; Moon et al., 2004; Shah et al., 2002), the data suggest that not only are -1 PRF signals present yeast, but also are prevalent in many distantly related organisms. This suggests a uniform means-to-an-end, whereby -1 PRF signals are functioning to control post-transcriptional gene expression. In Chapter 1, we reviewed the basic molecular process of -1 PRF in eukaryotes as well as the mechanism of nonsense mediated decay in yeast. This chapter will bring these two processes together and provide evidence that -1 PRF signals are not limited to recoding protein translation, but can also act to destabilize mRNAs in an NMD-dependent manner. The data below demonstrates that:

1. a functional viral -1 PRF signal acts as an NMD-dependent mRNA destabilizing element;
2. computationally identified -1 PRF signals endogenous to a variety of *S. cerevisiae* genes promote efficient recoding when tested *in vivo*; and
3. several yeast -1 PRF signals are also capable of acting as mRNA destabilizing elements in a similar manner as the viral -1 PRF signal.

The significance of these findings suggest a new model for post-transcriptional control of gene expression, "mRNA suicide", that couples genome encoded PRF signals with the rapid degradation of native mRNAs.

Materials & Methods

Genetic methods and plasmid construction

Plasmid amplification and transformations were performed as previously described in Chapter 3. YPAD and synthetic complete medium (H^-) were used as described previously (Dinman and Wickner, 1994). Isogenic ResGen yeast strains, detailed in Table 7, derived from BY4742 were used for *in vivo* measurement of programmed -1 ribosomal frameshifting and for quantitation of steady state *PGK1* reporter mRNA. All yeast cells were transformed using the alkali cation method (Ito et al., 1983). Dual luciferase plasmids pJD375 and pJD376 have been described previously (Harger and Dinman, 2003).

Computationally identified putative -1 PRF signals derived from *BUB3*, *CTS2*, *EST2*, *FKS1*, *FLR1*, *NUP82*, *PPR1*, *SPR6*, and *TBF1* were designed with the appropriate

restriction sites on the 5' and 3' ends. Naturally occurring termination codons were eliminated from the -1 reading frame by shortening the spacer region between slippery site and the putative downstream stimulatory structure by a single nucleotide when appropriate. PAGE purified oligonucleotides corresponding to each -1 PRF signal were annealed and gel purified. The oligonucleotides used for this construction are shown in Table 10 on page 133. pJD375, the "zero-frame" control dual-luciferase frameshift reporter plasmid, was used as a vector backbone and each putative PRF signal was cloned into unique *SalI* and *BamHI* restriction sites located in the MCS between the *Renilla* and firefly luciferase open reading frames. The resulting new PRF-reporter vectors were verified by DNA sequencing[29].

Plasmid pJD741[30] was used as the vector backbone for construction of *PGK1* reporter mRNA plasmids containing putative -1 PRF signals. Oligonucleotide primers for the 5' and 3' flanking positions of each PRF signal[31] were used to PCR amplify fragments from each parental gene directly from JD1158 genomic DNA. The primers were engineered with restriction sites (*BamHI* and *SalI*) that allowed amplified fragments to be sub-cloned into pJD375 as described above. A second set of oligonucleotide primers were designed with overhanging *KpnI* restriction sites such that PCR reactions with the newly constructed pJD375-derived clone used as template would result in amplification of a fragment for cloning into pJD741. These amplified fragments that

[29] Macrogen Inc. Seoul, Korea.

[30] pJD741 was previously reported in the literature as pW9 (Plant et al., 2004).

[31] The sequence of these primers can be found in Appendix C: Oligonucleotides.

contained a *KpnI* restriction sites on the 5' end, 36 nucleotides from 3' portion of the *Renilla* ORF, 29 nucleotides from the 5' end of the firefly ORF, and a second *KpnI* restriction site. This final fragment was then cloned into a unique *KpnI* restriction site found in *PGK1* in plasmid pJD741. Restriction analysis, directional PCR, and DNA sequencing verified the correct orientation of each cloned insert. The "read-through control" (RTC) was constructed in a similar manner directly from the dual-luciferase zero-frame control (ZFC) vector pJD375. The premature termination codon (PTC) containing construct was engineered directly from the RTC by cutting at a unique *AvrII* restriction site in the MCS, backfilling the overhanging ends using Klenow fragment, followed by blunt-end ligation. This resulted in a PTC clone identical to the RTC clone only with the addition of an in-frame stop codon. In all of these -1 PRF signal containing clones, frameshifting results in premature termination at a -1 frame termination codon present in the amplified genomic fragment immediately 3' of the signal.

Accession Numbers

The primary SGD accession numbers of genes from which the entire CDS or a subsequence of the CDS was used in this study are: *BUB3* (#S000005552), *CTS2* (#S000002779), *EST2* (#S000004310), *FLR1* (#S000000212), *FKS1* (#S000004334), *NUP82* (#S000003597), *PGK1* (#S000000605), *PPR1* (#S000004004), *SPR6* (#S000000917), and *TBF1* (#S000006049).

NMD modeling

In silico modeling of NMD was carried out in a manner similar to that of Cao & Parker (2001) with the additional goal of separating out the pioneer round of translation for each message from subsequent rounds. Thus, decay of cellular mRNAs was modeled according to each 'round of translation' as opposed to overall time of decay (Cao and Parker, 2001; Cao and Parker, 2003). One successful round of translation represents, in terms of the *in silico* model presented here, the complete cycle of initiation, elongation and termination events. The following assumptions were devised for simplification and reduction of parameters:

1. rates of these individual translational events were considered uniform for each mRNA in the pool;
2. premature termination of a ribosome directs an mRNA to the NMD pathway with an efficiency of 98% (Cao and Parker, 2003);
3. mRNAs in the pool are exposed to a 1% constitutive rate of decay (pCRD) independent of translational accuracy or fidelity; and
4. the probability of frameshifting (pPRF) per translational round ranged from 1 to 4% for each message.

Modeling of the pioneer round of translation was also done using the same parameters, except that pPRF was set to 0% after the first round of translation. The degradation of wild-type mRNAs assumed pPRF = 0% for those messages. A background rate of non-programmed frameshifting was introduced at 0.01% (pNPRF = 0.0001) (Dinman et al.,

1991). For each simulation, a starting pool of 10,000 mRNAs was used. The computational model was developed in PERL and the source code is available on request.

Dual Luciferase Assay System

Renilla to firefly luciferase ratios, i.e. frameshift efficiency, was determined for nine putative -1 PRF signal containing dual luciferase reporters, as well as the L-A viral frameshift control reporter pJD376. All calculated -1 PRF efficiencies were normalized to the pJD375 zero-frame control reporter as previously describe in Chapter 3 and in (Harger and Dinman, 2003). A minimum of 12 replicate assays were carried out for each candidate -1 PRF signal. Statistical analyses of each luciferase dataset followed the protocol established in Chapter 3 aimed at identifying outliers and at validating the statistical assumptions implicit in the DLR system (Jacobs and Dinman, 2004). All of the datasets passed the required statistical tests with a varying number of replicate experiments for each.

Preparation of RNA and cDNA Samples

Strains JD1158 and JD1181 were transformed with each of the *PGK1-PRF* vectors, the RTC and PTC vector. Cell cultures were grown at 30ºC in synthetic dropout media (ura-) for 16 - 24 hours until reached exponential growth and an OD_{595} between 1.0 and 2.0. A 1 mL aliquot of cells was collected from each culture, immediately centrifuged, decanted and frozen in liquid N_2. All collected aliquots of cells were then

stored at -80ºC. Total RNA was later isolated from frozen aliquots[32] and diluted to final total RNA concentration of 100 ng/μL. RNA samples were checked for quality by gel electrophoresis followed by ethidium bromide staining. Synthesis of cDNA was carried out using the isolated RNA as template with a single round reverse-transcriptase reaction that utilizing random hexamer primers[33].

Quantitative Real Time PCR

The relative quantitation of *PGK1* reporter mRNA from each sample was determined by using quantitative real-time PCR (qPCR) with prepared cDNA as starting template[34]. Specific primers for qPCR, shown in Table 12 in Appendix C: Oligonucleotides, were designed to take advantage of a 37 nucleotide fragment of the *Renilla* ORF present in the *PGK1* reporter mRNA. Control experiments using pJD741 as an empty vector control or cDNA amplified from untransformed yeast cells never resulted in amplification of the endogenous copy of *PGK1* on in any other reproducible secondary amplicon. A second set of qPCR primers were designed for 18S rRNA, which was used for normalization. Input cDNA concentrations and primer concentration were first optimized for linearity using the RTC, pJD753. The system was found to be linear

[32] RNA was purified on silica bead filters using the RNeasy® Mini Kit from Qiagen Inc., Valencia, CA.

[33] The iScript cDNA Synthesis Kit from Bio-Rad Laboratories (Hercules, CA) is specifically designed for the amplification of cDNAs for qPCR experiments.

[34] qPCR experiments were carried out using an ABI7700 Prism Sequence Detector from Applied Biosystems Inc. (Foster City, CA). The iTaq SYBR Green Supermix with ROX system for qPCR were supplied by Bio-Rad Laboratories (Hercules, CA).

across 6 logs of starting RNA concentrations, as shown in Figure 22, with 50 nM of primers per reaction. Similar results were found for the detection of 18S rRNA, although the C_T values were substantially lower[35].

Melting curves of qPCR end-products were carried out in order to ensure that amplification was specific and only a single amplicon was produced from each reaction. The melting curves were determined as follows. Completed qPCR reactions were cooled to room temperature and then gradually heated to 95ºC over a period of 20 minutes. SYBR Green fluorescence was recorded over time between 60ºC and 95ºC. Rapid reductions in the total observable SYBR Green fluorescence indicated the melting of double stranded DNA present in the reaction mixture. By plotting the rate of fluorescence change across a range of temperatures, a spike represents rapid duplex melting. In Figure 23, are representative melting curves of RTC and 18S rRNA amplicons generated from ten qPCR samples. The results indicate that non-specific amplification does not occur with either primer set for any of the qPCR results presented here. The data collected from each qPCR experiment was analyzed using the previously described comparative C_T method (Livak and Schmittgen, 2001).

[35] Data not shown.

Results

Computational modeling of PRF-dependent NMD

Although the *in vivo* efficiency of the L-A frameshift is only 2–9% depending on the assay system used (Dinman et al., 1991; Harger and Dinman, 2003; Jacobs and Dinman, 2004), a *PGK1* reporter mRNA containing this signal was only 2-fold more stable than a zero frame nonsense-containing mRNA (Plant et al., 2004). An *in silico* approach was devised to address how such low levels of frameshifting could have such strong effects on mRNA stability. Theoretical mRNA decay rates based on several different models were computationally generated to address this question. Beginning with several pools of 10,000 identical mRNAs of equal length, each mRNA in each pool was assumed to be subject to a constitutive rate of decay, independent of translation, that was arbitrarily set at 1% of the messages capable of entering a decay pathway after each successful round of translation. This model also assumes all messages in a given pool have a number of identical features, including maximal ribosome load, one -1 PRF signal (except wild-type, which has none), and that if any of the loaded ribosomes shifted frame there would be a 98% chance that the mRNA will be recognized as aberrant and degraded by NMD (Cao and Parker, 2003). Finally, the process of translational elongation was considered error-free for the purposes of this simulation.

Several computationally generated solutions are depicted in Figure 24, each depending on whether or not NMD remains active after the "pioneer round" (Ishigaki et al., 2001) of translation. The decay profile predicted for a wild-type message, i.e. no PRF, is shown in Figure 24 as solid black lines and follows a shallow, approximately linear

negative slope. A previous *in vitro* study suggested that approximately half of ribosomes that pause at the wild-type L-A -1 PRF signal actually shifted (Lopinski et al., 2000). Assuming that 50% of ribosomes shift during the pioneer round of translation, and that NMD may be inactivated after this first round (Ishigaki et al., 2001), the calculated decay profile shows that, although roughly half of the mRNAs are eliminated from the pool at the first round of translation, the trajectories and decay rates of the remaining mRNAs parallel that of the wild-type mRNA (yellow lines). In contrast, if -1 PRF efficiency occurs at an efficiency of 2%, and if NMD remains active beyond the pioneer round (Maderazo et al., 2003; Keeling et al., 2004), then calculated decay profile is observed to fit an exponential trajectory and the theoretical decay rate of such an mRNA is significantly greater than that of the wild-type mRNA (green lines). Further, the data in Figure 24 predicts that rates of mRNA decay would follow an inverse proportionality relationship with -1 PRF efficiency, as shown by the red and blue lines. This is based on the notion that efficient PRF signals would direct messages to the NMD pathway with greater effectiveness.

Figure 25 plots the empirical decay profiles of the *PGK1* reporter mRNAs shown in Plant et al. (2004), Figures 1 and 3. The data for the in-frame control p3131 follows the predicted shallow linear negative slope. The decay profiles for the frameshift reporters[36] follow the typical 'biphasic' decay profiles of nonsense-containing mRNAs observed with the nonsense-containing controls p3082 and pJD255 (Leeds et al., 1991). The first phase of these decay profiles follow the logarithmic trajectories predicted by the model of

[36] pJD269, pJD274 and pJD273 are all described previously (Plant et. al. 2004).

continuous NMD, providing independent support for the model that the NMD apparatus remains active beyond the pioneer round of translation in yeast. It is notable, however, that whereas logarithmic mRNA decay proceeds to zero in the computational model, this decay function abates at ~30% of time zero, after which the shallow negative linear function is observed. Given that NMD is only active on actively translating ribosomes, we suggest that only 70% of the *PGK1* test mRNAs were initially present in the pool of actively translated mRNAs. The remaining 30% would not be actively translated, and thus only subject to degradation by non-NMD processes.

These results establish that a functional -1 PRF signal can act as an mRNA destabilizing element, and that this effect is largely dependent on NMD (Plant et al., 2004). Therefore, it was important to test the hypothesis that an endogenous -1 PRF signal, such as any of the ones identified computationally in Chapter 2, could also have the same effect on it's encoding mRNA.

The Selection of Candidates for Empirical Testing

Bioinformatics studies generally benefit from the infusion of experimental bench data. To this end, nine candidate signals possessing a wide range of feature statistics were selected from the PRFdb for empirical testing. First and foremost, -1 PRF signals were selected from genes having scorable phenotypes when under- or over-expressed. Second, eight of the nine candidate signals chosen are predicted to fold into a pseudoknot, the exception being the signal chosen from *FKS1*. Nearly all known functional -1 PRF signals described in the literature have this requirement for a pseudoknot structure. Third, not all the selected signals should fully meet the criteria

outlined in Chapter 2 for strong candidate signals. For example, the two signals from *FLR1* and *SPR6* met all of the criteria for strong-candidate -1 PRF signals having $z_R \leq -1.65$ and predicted MFE values in the lowest 25% of all structures in the PRFdb[37]. Signals identified in the genes *CTS2*, *EST2*, *NUP82*, and *TBF1* meet less stringent criteria in that, although they are not in the first quartile of the most stable structures, they nonetheless are considered significant with $z_R \leq -1.65$. Candidate signals from *BUB3* and *PPR1* were chosen because they specifically do not meet any of the criteria above except that they have a predicted slippery site and pseudoknot structure. The predicted slippery sites and associated secondary structures for each of these candidate signals are shown in Figure 26. The feature statistics of each candidate signal are summarized in Table 5 on page 107.

Testing for Frameshifting

Each of the nine candidate -1 PRF signals were cloned into pJD375, the zero-frame dual-luciferase frameshift reporter, and the ability of each signal to promote -1 PRF was measured in a wild-type yeast strain, JD1158, as previously described (Harger and Dinman, 2003). Briefly, the ratio of firefly to *Renilla* luciferase expression promoted by -1 PRF signal containing reporters is normalized to a "zero-frame" control reporter pJD375, and these ratios are statistically tested for reliability as previously described (Jacobs and Dinman, 2004). At least ten replicate experiments were carried out for each reporter. The results, shown in Figure 27 and Table 6 below, indicate that every signal

[37] The reader is referred to pages 28 and 30 in Chapter 2.

containing a predicted mRNA pseudoknot promoted -1 PRF at levels that significantly exceeded non-programmed (or background) frameshifting. In contrast, the sequence derived from *FKS1*, which is not predicted to contain a pseudoknot, did not promote frameshifting to any measurable degree. In a broad sense, the experimental data divides the signals into high- and low-efficiency -1 PRF signals. The signals cloned from *CTS2*, *EST2*, and *PPR1* promoted -1 PRF at approximately 64%, 56%, and 43% respectively. The remaining functional signals promoted -1 PRF at levels between 0.4% - 5.2%. For purposes of comparison, the well-characterized -1 PRF signal from the yeast L-A virus promoted 9.1% frameshifting.

Some, but not all, PRF signals can destabilize mRNA

It was previously demonstrated that the -1 PRF signal derived from the yeast L-A virus could function as an mRNA destabilizing element when cloned into a *PGK1* reporter mRNA (Plant et al., 2004). To examine whether functional -1 PRF signals derived from chromosomally encoded genes could have similar *in vivo* activities, six of the nine such signals were cloned into an episomal *PGK1* reporter plasmid. Specific detection of the *PGK1* reporter mRNA was carried out using quantitative real-time PCR for each clone in isogenic wild-type and *upf3*Δ yeast strains[38]. In Figure 28, the three signals from *EST2*, *PPR1*, and *SPR6* effectively reduced the steady state abundance of the reporter *PGK1* mRNA to as little as ~30% of the read-through control. In addition, as shown in Figure 29, these three signals had no distinguishable effects on reporter *PGK1*

[38] The reader is referred to page 93 in the Materials & Methods, Quantitative Real Time PCR, above.

mRNA levels as compared to the read-through control (RTC) in the NMD deficient *upf3*Δ strain. Interestingly, while the strong -1 PRF signal derived from *CTS2* did not act to reduce steady mRNA levels in the wild-type strain (94% of control), there was a strong and reproducible derepression of *PGK1-CTS2* mRNA levels in the *upf3*Δ strain ($p \leq 0.05$; Student's T-test). The nearly 2-fold increase in PGK1-CTS2 levels indicates that this message is a substrate for NMD in general, despite the strong constitutive expression of the *PGK1* promoter present in reporter plasmid. Unexpectedly, the -1 PRF signals from *BUB3* and *TBF1* not only seemed to increase the steady-state levels of *PGK1* reporters into which they were cloned, but the resulting mRNAs were also insensitive to NMD. Overall, four of the six *PGK1* reporter mRNAs tested were shown to be derepressed between 1.2 – 2.2 fold in the *upf3*Δ strain when compared to the degree of derepression of the control mRNA, as shown in Figure 30. While modest, these changes nonetheless indicate that the reporter mRNAs are targeted for degradation by the NMD pathway at steady state. Interestingly, as shown in Table 5, the native mRNAs from the genes containing the above -1 PRF signals have all been previously shown to be natural substrates for NMD (He et al., 2003; Lelivelt and Culbertson, 1999) and are among the least stable in yeast transcriptome (Wang et al., 2002). Together, these findings suggest that four out of six genomic -1 PRF signals tested may have the capacity to act as regulatory elements by directing mRNAs to the NMD pathway for degradation.

Discussion

This Chapter demonstrates that computationally identified endogenous -1 PRF signals found in the yeast genome can promote efficient frameshifting *in vivo*, and that a

subset of these can act to destabilize mRNAs in a manner that is dependent on the functionality of the NMD pathway. Support of this finding originates from data that answers four fundamental questions:

1. Does the NMD pathway act on messages beyond the "pioneer" of translation?
2. Can a -1 PRF signal act as an NMD-dependent mRNA destabilizing element?
3. Do putative, computationally identified -1 PRF signals promote efficient frameshifting *in vivo*?
4. Can an endogenous -1 PRF signal from yeast act as an NMD-dependent mRNA destabilizing element?

The affirmative answers to each of these questions, and the supporting data presented previously in this Chapter, is discussed further below.

The extremely efficient nature of the NMD apparatus on nonsense-containing mRNAs has hampered our ability to determine whether the NMD can happen on mature mRNAs after the first round of translation. Here, we have addressed the issue by inserting -1 PRF signals into a *PGK1* reporter mRNA so that ribosomes encounter nonsense codons at low frequencies; effectively creating a conditional PTC+ mRNA. Comparison of the resulting reporter mRNA decay profiles with computationally modeled ones supports the findings of Maderazo et al. (2003) and Keeling et al. (2004) by providing an independent, less invasive way to address the question of whether or not NMD can remain active after the pioneer round of translation.

The observed mRNA decay profiles shown in Figure 25 are also important because they address the question of whether an mRNA pseudoknot can re-form on an

mRNA after it has been denatured by an elongating ribosome. For example, a pioneer ribosome that does not shift and continues to translate the message in the 0-frame would denature the mRNA pseudoknot. If the pseudoknot were not able to re-form, then the –1 PRF signal would be rendered non-functional and the mRNAs would be stable. Similar to the scenario described above for the case of NMD confined to the pioneer round of translation, if this were the case then the observed decay plots would follow linear rather than exponential trajectories. The observed mRNA decay profiles clearly show continuous frameshifting on –1 PRF-competent mRNAs. Thus, even if the first ribosome fails to shift and denatures the mRNA pseudoknot, the ability of subsequent ribosomes to shift demonstrates that the mRNA pseudoknot is able to re-form. This hypothesis is also in accordance with recent findings that actively translated mRNAs are not maximally loaded with ribosomes (Arava et al., 2005; Arava et al., 2003) and that considerable secondary structure is present in the coding regions of actively translated mRNAs (Chamary and Hurst, 2005; Katz and Burge, 2003; Meyer and Miklos, 2005).

Nine candidate signals were chosen for empirical testing of frameshift efficiency. Each was chosen for a variety of reasons that reflected:

1. the diversity of feature statistics from the PRFdb;
2. genes whose native transcripts were relatively unstable; and
3. genes who native transcripts were found to be upregulated in the absence of a functioning NMD pathway.

All but one candidate -1 PRF signal promoted efficient frameshifting *in vivo*[39]. In addition, six of the nine signals were also examined for their effects on mRNA stability. Four of these were shown to have from weak to strong NMD-dependent mRNA destabilizing activities. We believe that the powerful *PGK1* promoter that was used to drive transcription in the reporter plasmids may have significantly masked the destabilizing effects of the -1 PRF signals. A brief consideration of these four signals follows.

The signal from *CTS2* was considered a strong candidate -1 PRF signal since it is predicted to feature a spacer of an appropriate size, a better than expected number of base pairs, and a statistically significant MFE value. The full-length mRNA of *CTS2* is naturally unstable and was found to be derepressed in the absence of a functioning NMD pathway, indicating that its mRNA is likely a natural target for NMD (He et al., 2003; Lelivelt and Culbertson, 1999; Wang et al., 2002). The candidate -1 PRF signal from *CTS2* was found to promote frameshifting at very high levels, 63.7%, and targeted the *PGK1* reporter mRNA to the NMD pathway as evidenced by the apparent strong degree of depression in a *upf3*Δ strain.

The -1 PRF signal cloned from *EST2* exhibits many of the same features listed above for *CTS2*, and was also found to promote frameshifting at surprisingly high levels, 56.4%. The signal from *EST2* had only a weak effect on PGK1 reporter mRNA abundance in a wild-type strain, reducing the mRNA to 77% of RTC levels[40]. However,

[39] The exception being the signal from *FKS1*, which is the only one not predicted to fold into a pseudoknot.

[40] two-sample Student's t-test indicates $p = 0.06$, which is marginally significant at best.

in a *upf3*Δ strain, the levels of the PGK1-EST2 mRNA was upregulated to approximately 93% of the RTC levels, making it statistically indistinguishable from the RTC[41]. Overall, this represented a reproducible, albeit weak, 22% repression by the NMD apparatus .

The sequence from *PPR1* promoted 43.2% frameshifting despite containing a relatively weak and statistically insignificant pseudoknot structure[42]. The *PPR1* candidate signal's ability to direct highly efficient -1 PRF is likely due to the presence of three tandem, overlapping, slippery sites which give translating ribosomes three opportunities to shift reading frame. Thus, when the effects of mRNA destabilization were assayed using qPCR, this signal was shown to reduce the levels of PGK1-PPR1 mRNA to 55% of the RTC in a wild-type strain[43]. These mRNA levels rose to 86% of the RTC in a *upf3*Δ which translates to a 1.6-fold increase relative to wild-type. The *PPR1* mRNA has previously been shown to be upregulated in an NMD deficient strain background and is known to be an extremely difficult to detect transcript in wild-type strains (He et al., 2003; Kebaara et al., 2003; Lelivelt and Culbertson, 1999; Wang et al., 2002). Also interesting is that this -1 PRF signal lies in a sequence region of *PPR1* previously indicated to be responsible for a 3.6 fold derepression of full-length mRNA transcript levels in a *upf1*Δ strain (Kebaara et al., 2003).

Examination of the signal derived from the *SPR6* gene led it to be considered a strong candidate -1 PRF signal for many of the same reasons cited for the signals derived

[41] $p = 0.45$ by two-sample Student's t-test.

[42] The pseudoknot from the *PPR1* signal is predicted to have an MFE of -7.8 kcal/mol, and $z_R = 0.6$.

[43] $p \leq 0.01$ by two-sample Student's t-test.

from *CTS2* and *EST2*. The *SPR6* signal was not, however, a "high efficiency" frameshift signal when tested *in vivo* with the dual-luciferase reporter assay system as it only promoted 0.4% frameshifting. However, it is possible that the necessary base deletion in the predicted spacer region of this PRF signal negatively affected recoding efficiency as measured by the DLR system. Nonetheless, this signal reduced the levels of steady state PGK1-SPR6 mRNA to 30% of the RTC mRNA in a wild-type strain background[44], and in a *upf3*Δ strain these levels rose to 66% of the RTC mRNA. This represents a 2.2-fold derepression of mRNA levels, indicating that this signal specifically targets the reporter transcript for degradation by NMD.

The findings presented by this study suggest that PRF signals can function efficiently in a number of different ways. For example, while sequences that are predicted to fold into strong, statistically significant, pseudoknotted mRNA structures serve as efficient stimulators of -1 PRF (*e.g.* signals from *CTS2*, *EST2*), the presence of multiple overlapping slippery sites can also have an equally strong effect, even if the stimulatory structure is not ideal (*e.g.* signals from *BUB3*, *PPR1*, *NUP82*). Most importantly, it appears that the presence of a pseudoknot as the "most stable" structure following a slippery site is critical, providing further support for the "torsional restraint" model of -1 PRF (Plant and Dinman, 2005). This is further evidenced by the fact that the very energetically favorable and highly significant structure derived from *FKS1* failed to promote detectable -1 PRF *in vivo*. This last point may be because there is no predicted

[44] $p = 0.02$ by two-sample Student's t-test.

pseudoknot structure with an MFE value lower than that of the predicted non-pseudoknotted structure immediately following the slippery site of interest in *FKS1*.

Chapter 4 Tables

Table 5: Features of Nine Candidate -1 PRF Signals

ORF, the parental ORF for the putative -1 PRF signal; *pos*, the starting nucleotide position immediately after the slippery site relative to the ORF's start codon; *Slippery Site,* the heptameric slippery sequence expected to be the site of translational recoding, spaces between nucleotides indicate zero-frame codons (*BUB3* and *NUP82* contain two overlapping slippery sites, and *PPR1* contains three); *Spacer Length,* the distance between the last nucleotide of each slippery site and the first nucleotide involved in base pairing of the downstream predicted secondary structure, multiple entries indicate spacers for each possible slippery site; *Pairs,* the predicted number of base pairs for the downstream structure; *MFE,* the predicted minimum free energy in kcal/mol; z_R, the statistical significance of the MFE value compared to 100 randomized sequences; *PK,* indicates whether the predicted structure a pseudoknot; *NMD*Δ_1, independently determined fold change of mRNA levels for the corresponding gene averaged across three strains defective in nonsense mediated mRNA decay (He et al., 2003); *NMD*Δ_2, an earlier measurement of fold change for mRNA levels in an nmdΔ strain (Lelivelt and Culbertson, 1999); $T_{1/2}$, independently determined half-life for the full-length mRNA for each ORF after transcriptional arrest (Wang et al., 2002). The average half-life for all mRNAs was 26 minutes; n.d., not detected; n.a., data was not available.

ORF	*pos*	*Slippery Site*	*Spacer Length*	*Pairs*	*MFE*	z_R	*PK?*	NMD_1	NMD_2	$T_{1/2}$
BUB3	858	**A AAA AAU UUC**	6 , 9	14	-4.8	0.5	Yes	0.8	2.9	10 m
CTS2	1245	**A AAA AAU**	7	20	-14.4	-3.6	Yes	2.2	1.3	15 m
EST2	1653	**A AAA AAU**	6	27	-16.9	-3.1	Yes	1.5	2.4	n.d.
FKS1	3768	**A AAA AAC**	3	23	-22.2	-3.9	No	n.a.	1.4	n.d.
FLR1	228	**A AAA AAU**	5	30	-21.8	-2.0	Yes	1.2	3.9	17 m
NUP82	1545	**U UUA AAA AAC**	7, 10	16	-11.3	-2.0	Yes	1.1	1.0	n.d.
PPR1	1182	**U UUU UUU UUA AAC**	3, 6, 9	18	-7.8	0.6	Yes	1.6	2.0	n.d.
SPR6	279	**A AAA AAA**	8	27	-20.3	-1.6	Yes	1.1	2.7	16 m
TBF1	1521	**A AAU UUA**	5	11	-8.3	-2.1	Yes	0.8	2.9	5 m

Table 5: Features of Nine Candidate -1 PRF Signals (cont.)

	ZFC	**L-A**	***BUB3***	***CTS2***	***EST2***	***FKS1***
pJD#	pJD375	pJD376	pJD519	pJD485	pJD522	pJD523
$\bar{x}$	0.2896	0.0263	0.0023	0.1845	0.1632	2.6×10^{-5}
s^2_{N-1}	6.6×10^{-4}	3.0×10^{-6}	4.9×10^{-8}	6.5×10^{-5}	0.0040	1.1×10^{-10}
N	61	40	26	11	31	27
$\bar{x}_R$	-	9.08%	0.80%	63.71%	56.4%	<0.01%
$s_e(\bar{x}_R)$	-	0.14%	0.02%	1.11%	3.96%	<0.01%

	FLR1	***NUP82***	***PPR1***	***SPR6***	***TBF1***
pJD#	pJD522	pJD477	pJD476	pJD520	pJD478
$\bar{x}$	0.0015	0.0025	0.1252	0.0012	0.0152
s^2_{N-1}	3.4×10^{-8}	5.5×10^{-8}	5.5×10^{-5}	5.3×10^{-8}	1.3×10^{-5}
N	28	24	18	21	12
$\bar{x}_R$	0.53%	0.85%	43.24%	0.42%	5.23%
$s_e(\bar{x}_R)$	0.01%	0.02%	0.78%	0.02%	0.36%

Table 6: Frameshifting Statistics of Yeast -1 PRF Signals

Summary of the dual luciferase reporter assay data from nine endogenous -1 PRF signals identified in *S. cerevisiae*. ZFC, zero-frame control pJD375 to which the other Renilla/firefly ratios were normalized; L-A, data for the -1 PRF signal from the yeast L-A virus. *BUB3*, *CTS2*, *EST2*, *FKS1*, *FLR1*, *NUP82*, *PPR1*, *SPR6*, and *TBF1* are the encoding genes from which each -1 PRF signal originates; $\bar{x}$, sample mean of *Renilla* to firefly luciferase activity ratios; s^2_{N-1}, sample variance; N, sample size; $\bar{x}_R$, normalized frameshift efficiency of each; $s_e(\bar{x}_R)$, standard error of frameshifting.

Chapter 4 Figures

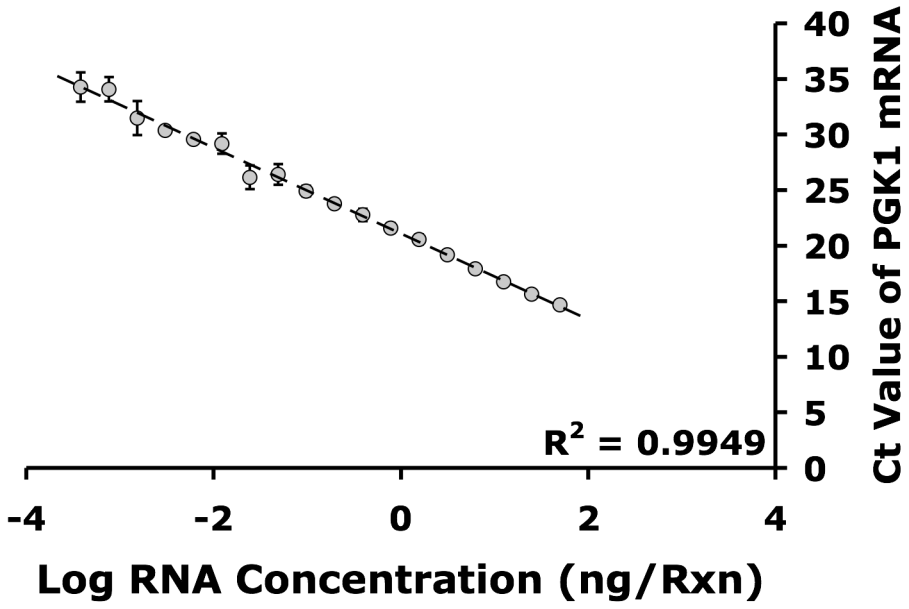

Figure 22: Total RNA Template Dilutions for qPCR of PGK1

Total RNA was isolated from strain JD1158 as described in the Materials & Methods of Chapter 4. The qPCR system for detection of the reporter copy of PGK1 was determined be linear between 0.4 pg of 50 ng.

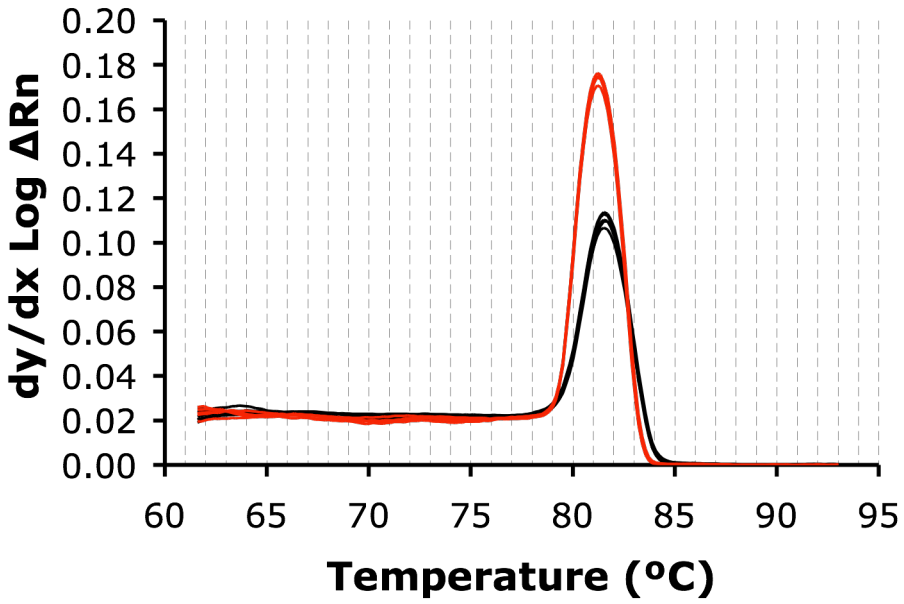

Figure 23: Melting Curves for PGK1 and 18S qPCR Amplicons

Data from five representative experiments are presented for both PGK1 (in red) and 18 rRNA (in black) amplicon. The results above indicate that only a specific PCR amplification product was produced by each primer set. X axis – temperature; Y axis – Rate of change of baseline subtracted florescence intensity.

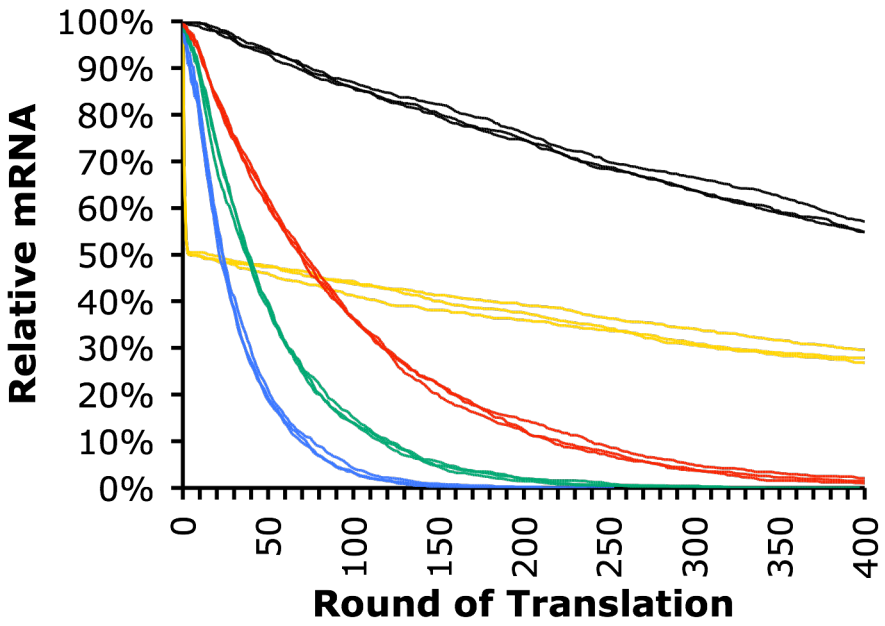

Figure 24: *In Silico* Modeling of NMD

The predicted decay profile of an mRNA lacking a PRF signal (0% PRF) is shown as black lines. The decay profile of an mRNA in which 50% of ribosomes shift at the pioneer round, and where NMD was inactivated after this first round, is shown as yellow lines. The calculated decay profile of an mRNA where –1 PRF efficiency was set at 2% efficiency and NMD remained active after the pioneer round is depicted by green lines. When NMD remained active after the pioneer round of translation and –1 PRF efficiencies were set at 4% or 1%, the calculated decay profiles followed the trajectories shown as blue and red lines respectively.

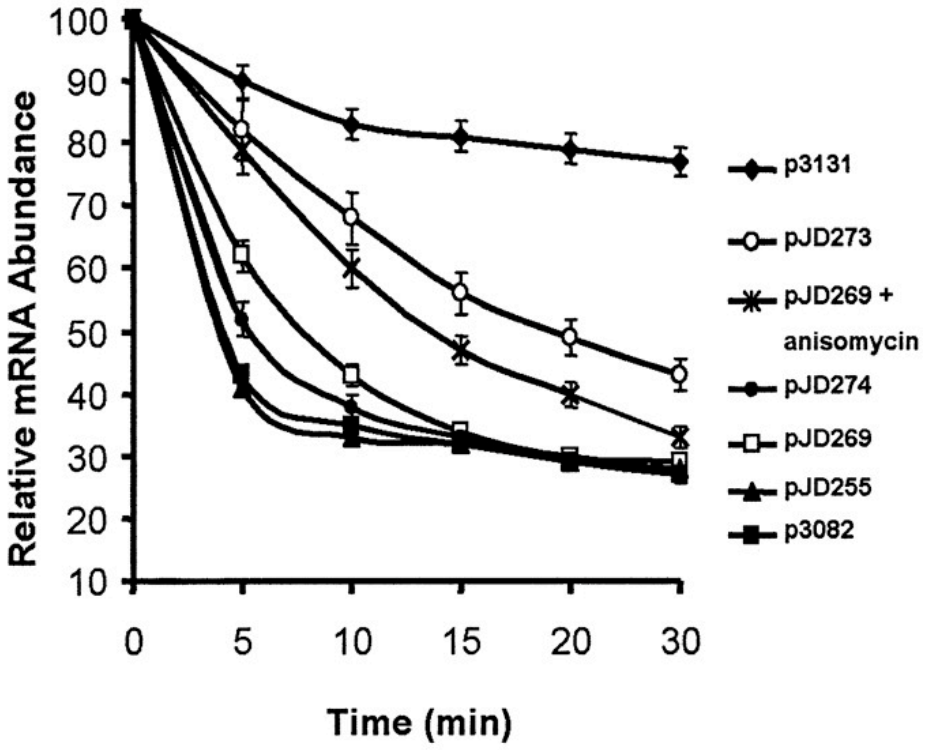

Figure 25: NMD Remains Active After the Pioneer Round of Translation

PGK1 reporter mRNA decay profiles corresponding PTC, -1 PRF signal, and control plasmids. p3131, read-through control message. p3082, nonsense containing message. pJD255, nonsense containing mRNA with functional -1 PRF signal. pJD269, read through mRNA with a -1 PRF signal. pJD274, read through mRNA with a high-efficiency -1 PRF signal. pJD273, read through mRNA with a low efficiency -1 PRF signal. pJD269 + anisomycin, read through mRNA with a -1 PRF signal and cells grown in the presence of 4 µg/mL of anisomycin. The figure above and this legend are adapted from Plant et. al. (2004) Figure 5B.

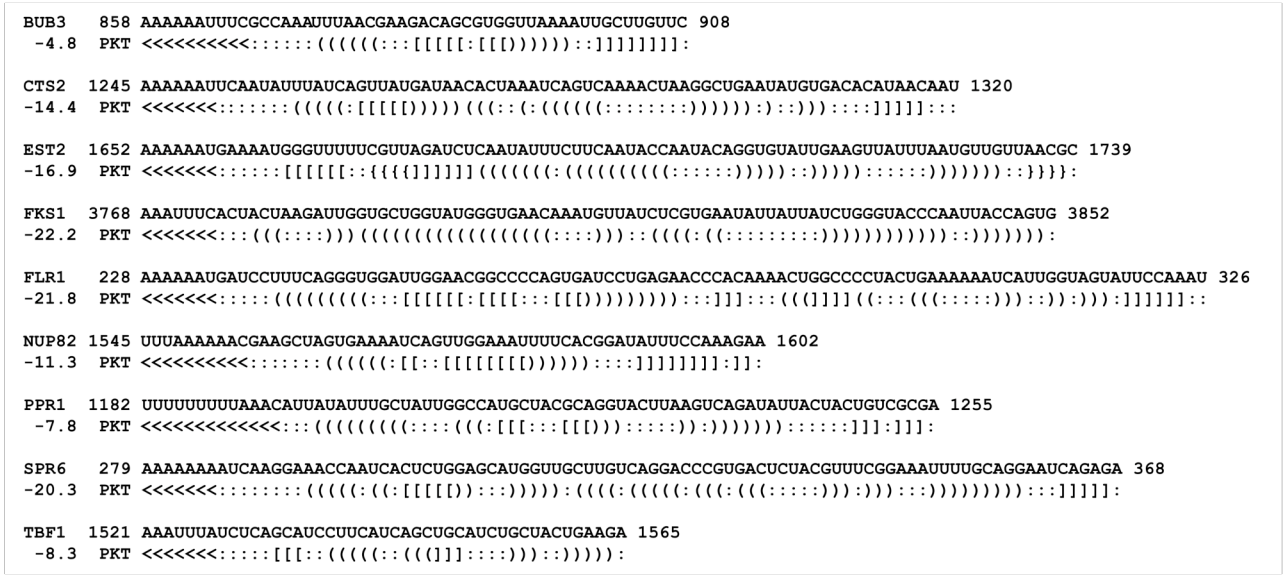

Figure 26: Nine Putative -1 PRF Signals from Yeast

Nine examples of candidate -1 PRF signals chosen to generally represent the diversity of features present in the PRFdb. Gene names are shown with RNA sequence and corresponding CDS nucleotide start and stop locations. The predicted structure with corresponding MFE value (in kcal/mol) is shown for each empirically tested candidate signal in bracket notion; *pknots* predicted structure (PKT); **<<<<<<<** Slippery site; brackets **( ) [ ] { }** represent base pairing partners; **:** represent unpaired base positions. See Table 5: Features of Nine Candidate -1 PRF Signals on page 107 for statistical data.

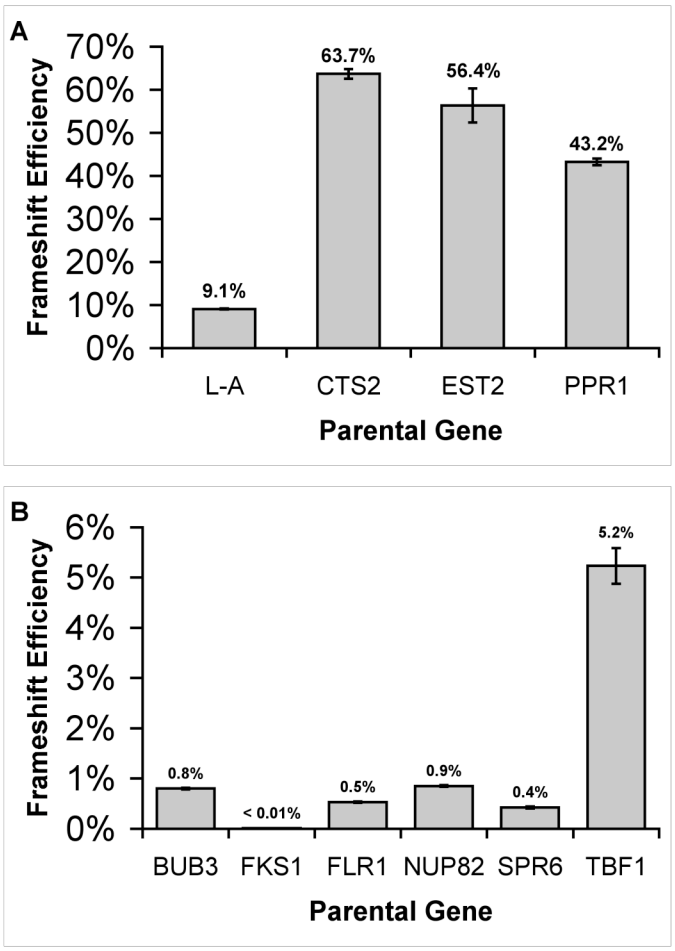

Figure 27: Frameshift Efficiencies of Nine Candidate -1 PRF Signals

A) High-efficiency and B) Low-efficiency frameshifting. The parental genes of each signal are indicated with the -1 PRF efficiency as was measured using a dual-luciferase reporter assay system (Harger and Dinman, 2003; Jacobs and Dinman, 2004).

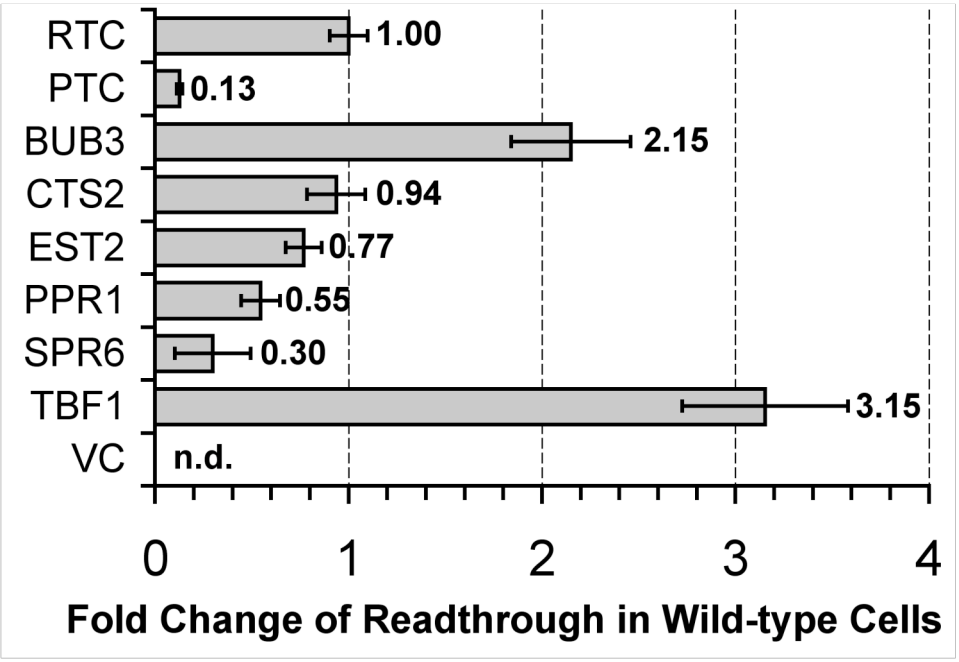

Figure 28: Wild-type *UPF1* Repression of Reporter *PGK1* mRNA

Reporter mRNAs harboring a premature termination codon (PTC) or one of the candidate -1 PRF signals from *EST2*, *PPR1*, and *SPR6* are measurably down regulated in a wild-type strain background JD1158 as measured by quantitative real-time PCR. No change was detected in the read through control reporter (RTC). The empty-vector control *PGK1* reporter (VC) was not detectable. All data was normalized to the levels of 18S rRNA present in each sample.

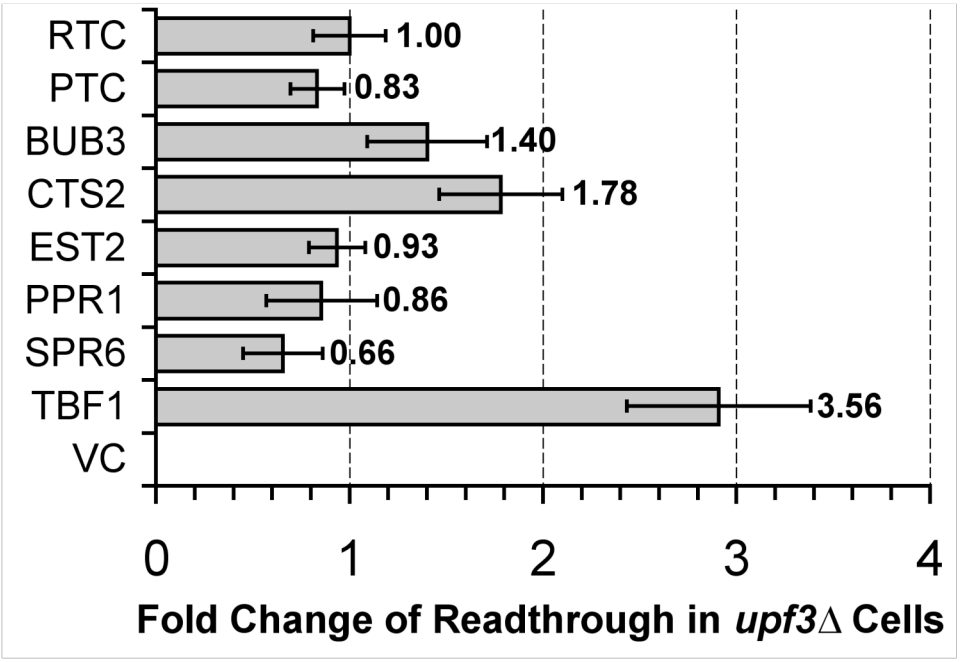

Figure 29: *PGK1* Reporter mRNA in a *upf3*Δ Strain

Reporter mRNAs encoding a premature termination codon (PTC) or one of the candidate -1 PRF signals from *CTS2*, *EST2*, *PPR1*, or *SPR6* increase in a *upf3*Δ strain. No change was detected in the read through control reporter (RTC). The empty-vector control *PGK1* reporter (VC) was not detectable. All data was normalized to the levels of 18S rRNA present in each sample.

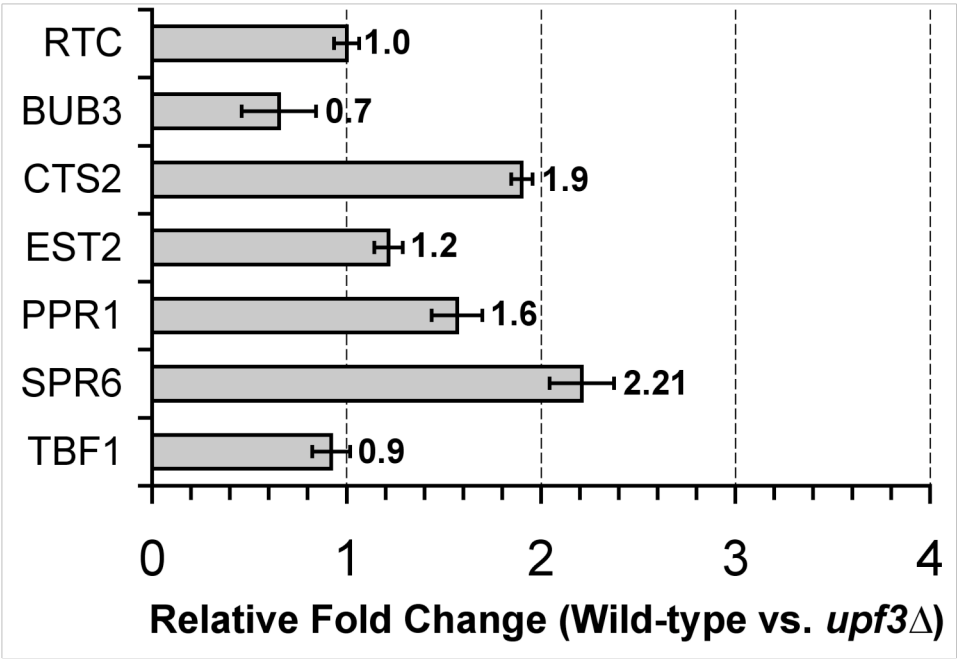

Figure 30: Relative Derepression of Reporter *PGK1* mRNA in *upf3*Δ

The relative fold-change of expression between wild-type and *upf3*Δ strains for reporter mRNAs comparing the data between Figure 28 and Figure 29. The levels of read-through control mRNA (RTC) between yeast strains was unchanged. The steady state levels of PTC containing reporter mRNA was derepressed approximately 6.5 fold in a *upf3*Δ strain.

Chapter 5: mRNA Suicide

Future Directions

Empirical studies are an important requirement for testing broad-based bioinformatics predictions. With regard to the research presented in this dissertation, there are many possible directions along which this work could be extended that serve this purpose. Most importantly, however, this author expects the greatest short-term benefit to this project will come from two specific experiments.

Functional Genomics of mRNA Suicide

Just as microarray experiments have been used to identify natural mRNAs targeted by NMD, it is not difficult to envision an experimental set up that utilizes similar approaches to identify transcripts that are subject to mRNA suicide. For example, a microarray time course assay can be designed to measure mRNA half-lives in transcriptionally arrested cells using both a wild-type and NMD deficient strains. Despite the excellent work of Culbertson, Jacobson and their respective colleages, they failed to measure mRNA half-lives for the microarray experiments they conducted (He et al., 2003; Lelivelt and Culbertson, 1999). Despite attempts to sort out primary and secondary targets from the existing data using bioinformatics approaches, measuring mRNA half-lives is a critical experiment that must be carried out in order to clearly define transcripts

that are primary targets of NMD (Taylor et al., 2005). Primary NMD targets should not only be upregulated in a *upf1*Δ, *upf2*Δ or *upf3*Δ strain but also should have vastly different *rates* of decay.

In addition, numerous mutant alleles of various components of the large ribosomal subunit have been identified in the Dinman lab. Many of these alleles have specific *trans*-dominant effects on programmed frameshifting in a wild type strain, which makes the experimental setup straight forward. Although, the transcript levels of PRF signal containing mRNAs are not currently known in strains harboring these mutations, we expect them to be inversely affected by changes in PRF efficiency. Plasmids encoding *trans*-dominant alleles of 5S rRNA (Kiparisov et al., 2005) can simply be transformed in both wild-type and NMDΔ strains. Transcripts whose steady state levels are epistatically affected by the functionality of the NMD pathway could be identified in this way. Thus, data of this sort would provide valuable insight into which cellular transcripts are influenced by mRNA suicide.

Finally, the two antibiotics anisomycin and sparsomycin are known to affect the ribosome's ability to maintain reading frame (Dinman et al., 1997). The introduction of these drugs into cell culture media is expected to differentially affect the decay rates of mRNAs subject to mRNA suicide because their respective PRF signals would have a altered efficiencies. Microarray time-course experiments aimed at measuring the differential rates of decay in the presence and absence of both +1 and -1 frameshift specific antibiotics could identify transcripts whose stability is dependent on the ribosomes ability to maintain reading frame.

Data garnered from the above three microarray experiments could then be integrated and compared across experiments. The resulting data set would aid in the identification of transcripts whose decay rates are the result of mRNA suicide. These candidates would be expected to have the following characteristics:

1. steady state derepression in a *upf1*Δ, *upf2*Δ, or *upf3*Δ strain background;
2. mRNA half-lives would be markedly reduced in an NMD deficient strain;
3. ribosome mutants exhibiting frameshifting defects would demonstrate altered steady state levels of PRF signal containing transcripts which would be suppressed for mRNA suicide transcripts in strains lacking NMD; and
4. mRNA suicide transcripts would have altered half-lives in the presence of antibiotics that interfere with translation elongation which would be largely abrogated in the absence of NMD.

Cessation of transcription would be essential for the experiments outlined above. Fortunately, in yeast, there are several genetic mutants that allow for transcriptional arrest after heat-shock, *RPO21* (Nonet et al., 1987) and *RPB4* (Miyao et al., 2001)[45]. In addition, to avoid possible complications from initiation-mediated mRNA decay as a result from heat-shock induction (Heikkinen et al., 2003), antibiotics such as actinomycin-D (Schindler and Davies, 1975) or thiolutin (Herrick et al., 1990) could be used to force transcriptional arrest as a secondary method. These experiments would no doubt yield a wealth of information, not only about the probable targets of mRNA

[45] *RPO21* has been recently renamed. It was previously known as *RPB1*.

suicide, but also in terms of how translation in general affects the stability of mRNAs themselves.

Analysis of Model Targets of mRNA Suicide

Functional genomic studies will aid in the identification of transcripts that are subject to mRNA suicide. Analysis of individual mRNAs will also be critical for understanding the interplay of PRF and NMD in terms of the control of their expression. Finding good candidate genes to serve as model targets will likely be easier once the genomics experiments described above are carried out. Fortunately, however, there is independent evidence for two candidate genes whose mRNAs may turn out to be targets for mRNA suicide, *PPR1* and *EST2*.

The *PPR1* gene, a zinc-finger transcriptional activator of the uracil biogenesis pathway, encodes an approximately 2.6-kb mRNA (Losson et al., 1983). This large transcript is highly unstable and only one or two copies are thought to be present in wild-type yeast cells at any given time. The extremely low abundance of the *PPR1* mRNA usually results in its absence from microarray datasets because it is thought to be below the signal to noise ratio. Nonetheless, using other methods of detection, the mRNA encoded by *PPR1* has been found to be a direct target for nonsense mediated decay (Kebaara et al., 2003; Lelivelt and Culbertson, 1999). The exact mechanism is thought to be the result of a short six codon uORF that overlaps with the natural AUG start codon (Pierrat et al., 1993). Using heterologous reporters of *PPR1*, Atkin and colleages have identified that not only is the 5' UTR of the mRNA required for NMD-dependent decay, but the first 92 nucleotides of the CDS is required as well (Kebaara et al., 2003).

Furthermore, they also reported that inclusion of the first 1250 nucleotides exacerbated this effect and they observed a 3.6-fold increase in *upf1*Δ strain. They conclude that the predominate effect of NMD-dependent decay is the result of a Upf-dependent element, UDE, in the 5' end of the gene and could not explain the stronger depression in the reporter possessing the large 1250 nucleotide fragment.

Interestingly, the CDS of *PPR1* is predicted to have eight slippery sequences, three of which are tandemly overlapping at positions 1182, 1185, and 1188. The details of all of *PPR1* slippery sequences are available in the PRFdb. The predicted secondary structure positioned downstream from the slippery sites at positions 1182-1188, i.e. the tandem slippery site, is shown in Figure 26 on page 114. The PRF efficiency of the tandem slippery site was shown to be highly efficient in this study and this signal alone is capable of targeting transcripts to NMD. It may be that the stability of the *PPR1* mRNA is controlled by two elements, uORFs and PRF signals, that are both targeting it for destruction. This would also explain it's extremely low transcript copy number.

A second potential model transcript may lie in the mRNA produced by *EST2*, the principle telomerase subunit in yeast. The *EST2* mRNA levels, and telomere length in general (Lew et al., 1998), has been previously shown to be under the influence of NMD (Lelivelt and Culbertson, 1999). Berman and colleages have also demonstrated that *EST2* mRNA is a direct target for NMD (Dahlseid et al., 2003), but do not provide an explanation for NMD-dependent regulation.

Similar to what was found in *PPR1*, the *EST2* CDS has multiple slippery sites and several highly promising putative PRF signals. As described in Chapter 4, the PRF signal

in *EST2* at position 1653 promotes highly efficient PRF and is capable of affecting the stability of mRNAs in an NMD-dependent way. Mutagenesis of individual slippery sites or of putative structural elements in the full length gene could be useful in determining if PRF in the natural transcript is indeed a substrate for NMD.

Conclusions

The results obtained by this study provide evidence that functional -1 PRF signals are present in the yeast genome and that they do not function to solely direct ribosomes into alternative reading frames. Moreover, functional -1 PRF signals present in the coding regions of the yeast genome can function to destabilize their encoding mRNAs in a manner that is dependent on the functioning of the nonsense-mediated mRNA decay pathway. This process is illustrated by a novel model, "mRNA suicide", which couples the ability of the mRNA to redirect ribosomes into alternative reading frames with the nonsense-mediated mRNA decay pathway as shown in Figure 31 below. Implicit in this model is that modulation of PRF efficiency could be used as the controlling effecter of transcript degradation rates. The efficiency of frameshifting could, in theory, be regulated by several means including:

1. alterations in the interactions between *trans*-acting factors and the ribosome (Cui et al., 1998; Dinman and Kinzy, 1997; Muldoon-Jacobs and Dinman, 2006);
2. changes in post-transcriptional modifications of rRNAs (Baxter-Roshek J.L. and Dinman J.D., unpublished data), or in the post-translational modification status of specific ribosomal proteins (Mazumder et al., 2003; Williamson et al., 1997);
3. expression of alternative forms of 5S rRNA (Kiparisov et al., 2005);

4. *trans*-acting factors that interact directly with the mRNA to either stabilize or destabilize stimulatory structures capable of affecting translation (Kollmus et al., 1996b);
5. alterations to translational accuracy or ribosome fidelity in response to changes in environmental stimuli (Stahl et al., 2004; Barak et al., 1996); or by
6. modulation of the sensitivity of the NMD apparatus to premature termination events at various locations throughout an mRNA (Weil and Beemon, 2006).

Regulation models such as these can be applied to a variety of biological examples where the stability of individual mRNAs or whole classes of mRNAs require a flexible stability threshold responsive to environmental cues. Furthermore, natural mRNA substrates for NMD that do not contain PTCs have been discovered in both the yeast and human transcriptomes (He et al., 2003; Kim et al., 2005; Lelivelt and Culbertson, 1999; Mendell et al., 2004; Moriarty et al., 1998; Pan et al., 2006; Taylor et al., 2005; Wittmann et al., 2006), suggesting that regulation of mRNA expression by NMD is broadly conserved and is used to regulate a variety of physiological processes.

It has also recently been suggested that specific sequences present in coding regions of mRNAs are capable of translationally stalling ribosomes long enough to direct them to be endonucleolytically cleaved and specifically degraded in both prokaryotic and eukaryotic organisms (Doma and Parker, 2006; Sunohara et al., 2004), a process termed "No-go decay". Many of the predicted candidate -1 PRF signals identified in the current work are predicted to be more stable than the mRNA structure used by Doma & Parker (data not shown) and thus would be expected to be similarly capable to stall translating

ribosomes. Thus, it is also possible that mRNAs containing these extremely stable mRNA structures identified in the current study may also be targeted for No-go decay independent of PRF. Conversely, it is possible that the presence of a slippery site just upstream from strong secondary structures may be the specific feature that allows for such mRNAs to evade being subject to this pathway (Plant et al., 2003). Nevertheless, it is reasonable to envision that a general function of PRF signals in the coding regions of eukaryotic mRNAs is to act as post-transcriptional capacitors of gene expression.

Chapter 5 Figures

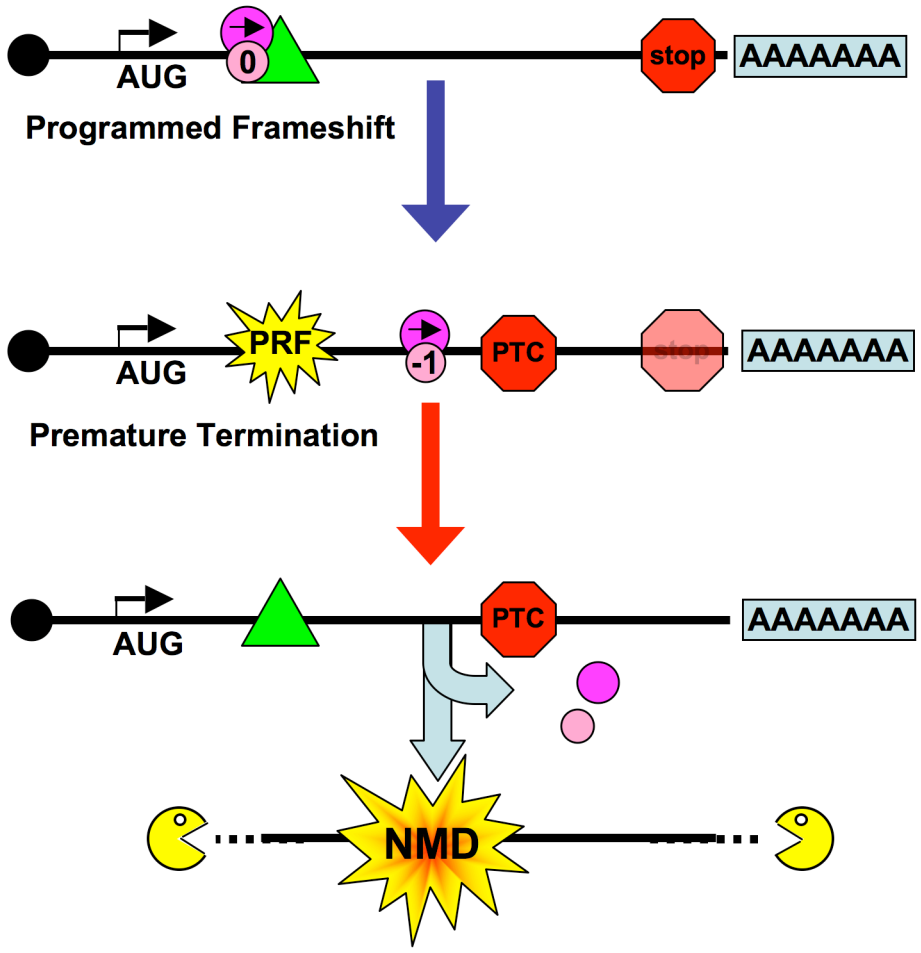

Figure 31: A Model of mRNA Suicide

The simplistic model above proposes that actively translating ribosomes may shift reading frame at functional recoding signals present in the mRNA, which in turn forces the premature termination of translation and the targeting of the transcript for degradation by NMD.

Appendix A: Yeast Strains

The table below summarizes the strains of *S. cerevisiae* referenced throughout this thesis. Please refer to the relevant pages for their specific use.

Table 7: Yeast Strains

Strain	Description
BY4742	*MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*
JD932	*MATa ade2-1 trp1-1 ura3-1 leu2-3,112 his3-11,15 can1-100*
JD1158	*MATα his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*
JD1181	*MATα upf3::KanR his3Δ1 leu2Δ0 lys2Δ0 ura3Δ0*

Appendix B: Plasmids

The table below summarizes the cloning plasmids referenced throughout this thesis. Please refer to the relevant pages for their specific use.

Table 8: Plasmids

Plasmid Name	Parental Vector	Description
pJD375	p2mc[46]	A zero-frame control (ZFC) dual luciferase reporter (DLR) expressing a bicistronic mRNA encoding a fusion of *Renilla* and firefly luciferase. This vector was previously described in (Plant et al., 2004).
pJD376	p2mci[46]	A viral frameshift signal containing dual luciferase reporter (DLR) expressing a bicistronic mRNA. This vector encodes a fusion of *Renilla* and firefly luciferase. This vector was previously described in (Plant et al., 2004).
pJD476	pJD375	DLR vector with a synthetically derived *PPR1* -1 PRF signal cloned into the multiple cloning site (MCS) between *Renilla* and firefly luciferase.
pJD477	pJD375	DLR vector with a synthetically derived *NUP82* -1 PRF signal cloned into the MCS between *Renilla* and firefly luciferase.
pJD478	pJD375	DLR vector with a synthetically derived *TBF1* -1 PRF signal cloned into the MCS between *Renilla* and firefly luciferase.
pJD485	pJD375	DLR vector with a synthetically derived *CTS2* -1 PRF signal cloned into the MCS between *Renilla* and firefly luciferase.
pJD519	pJD375	DLR vector with a synthetically derived *BUB3* -1 PRF signal cloned into the MCS between *Renilla* and firefly luciferase.
pJD520	pJD375	DLR vector with a synthetically derived *SPR6* -1 PRF signal cloned into the MCS between *Renilla* and firefly luciferase.

[46] Grentzmann, G., Ingram, J. A., Kelly, P. J., Gesteland, R. F., and Atkins, J. F. (1998). A dual-luciferase reporter system for studying recoding signals. Rna *4*, 479-486.

Plasmid Name	Parental Vector	Description
pJD521	pJD375	DLR vector with a synthetically derived *EST2* -1 PRF signal cloned into the MCS between *Renilla* and firefly luciferase.
pJD522	pJD375	DLR vector with a synthetically derived *FLR1* -1 PRF signal cloned into the MCS between *Renilla* and firefly luciferase.
pJD523	pJD375	DLR vector with a synthetically derived *FKS1* -1 PRF signal cloned into the MCS between *Renilla* and firefly luciferase.
pJD741	pW9[47]	Full-length *PGK1* gene on a low-copy *URA3*-based vector. Identical to the previously described pW9 vector (Plant et al., 2004). Used as starting material for vectors aimed studying the NMD-dependent effects of -1 PRF signals using qPCR. Also used directly as a negative empty-vector control for qPCR.
pJD753	pJD741	The MCS of pJD375 was cloned into the *KpnI* site of *PGK1* as described in Chapter 4: Materials and Methods on page 93. The insertion of the MCS of pJD375 included the 3' end of *Renilla* and the 5' end of firefly luciferase in order to specifically detect the *PGK1* mRNA generated from these vectors in the presence of the background expression of *PGK1* from the endogenous gene. This vector served as a read-through control (RTC) for *PGK1* expression for all qPCR experiments.
pJD748	pJD753	An mRNA stability vector derived from pJD753. The PRF signal region of *SPR6* was amplified from genomic DNA isolated from yeast strain JD1158 and cloned into the MCS pJD753 *PGK1*.
pJD754	pJD753	An mRNA stability vector derived from pJD753. The PRF signal region of *EST2* was amplified from genomic DNA isolated from yeast strain JD1158 and cloned into the MCS pJD753 *PGK1*.
pJD755	pJD753	An mRNA stability vector derived from pJD753. The PRF signal region of *BUB3* was amplified from genomic DNA isolated from yeast strain JD1158 and cloned into the MCS pJD753 *PGK1*.
pJD756	pJD753	An mRNA stability vector derived from pJD753. The PRF signal region of *TBF1* was amplified from genomic DNA isolated from yeast strain JD1158 and cloned into the MCS pJD753 *PGK1*.

[47] Plant, E. P., Wang, P., Jacobs, J. L., and Dinman, J. D. (2004). A programmed -1 ribosomal frameshift signal can function as a cis-acting mRNA destabilizing element. Nucleic Acids Res *32*, 784-790.

Plasmid Name	Parental Vector	Description
pJD757	pJD753	An mRNA stability vector derived from pJD753. The PRF signal region of *CTS2* was amplified from genomic DNA isolated from yeast strain JD1158 and cloned into the MCS pJD753 *PGK1*.
pJD759	pJD753	An mRNA stability vector derived from pJD753. The PRF signal region of *PPR1* was amplified from genomic DNA isolated from yeast strain JD1158 and cloned into the MCS pJD753 *PGK1*.
pJD765	pJD753	An mRNA stability vector derived from pJD753. A restriction site in the MCS was digested with *AvrII*, back-filled with Klenow fragment, and re-circularized with T4 Ligase. This created an in-frame premature termination codon in *PGK1* mRNA which serves as a positive control for NMD.
pJD766	pJD753	An mRNA stability vector derived from pJD753. The PRF signal region of *EST2* was amplified from genomic DNA isolated from yeast strain JD1158 and cloned into the MCS pJD753 *PGK1*.

Appendix C: Oligonucleotides

The table below summarizes the various oligonucleotides and primers referenced throughout this thesis. All oligonucleotides shown were ordered from Integrated DNA Technologies (Skokie, IL). Please refer to the relevant pages for their specific use.

Table 9: Oligonucleotides Used in Chapter 3

Name	Sequence
pJD519 Forward	TCGACAAAAAATCATCTTTCAGGGTGGATTGGAACGGCCCCAGTGATCCTGAGAACCCACAA AACTGGCCCG
pJD519 Reverse	GATCCGGGCCAGTTTTGTGGGTTCTCAGGATCACTGGGGCCGTTCCAATCCACCCTGAAAGA TGATTTTTTG
pJD478 Forward	CGACAAATTTATCTCAAGCATCCTTCATCAGCTGCATCTGCTACTGAAGAG
pJD478 Reverse	GATCCTCTTCTGTAGCAGATGCAGCTGAAGAAGGATGCTGAGATAAATTTG
pSARS Forward	GATCCTTTTTAAACGGGTTTGCGGTGTAAGTGCAGCCCGTCTTACACCGTGCGGCACAGGCA CTAGTACTGATGTCGTCTACAGGGCTTTTGAGCT
pSARS Reverse	CAAAAGCCCTGTAGACGACATCAGTACTAGTGCCTGTGCCGCACGGTGTAAGACGGGCTGCA CTTACACCGCAAACCCGTTTAAAAAG

Table 10: Oligonucleotides Used in Chapter 4

The oligonucleotides below were used in the construction of PRF signal containing dual-luciferase plasmids. Upper case, genomic sequence; Underline, restrictions sites; *, positions deleted to shift -1 frame stop codons into an alternative reading frame such that frameshifting does not result in premature termination; Bold, substitutions made to eliminate -1 frame PTCs.

Parent Gene	Sequence
BUB3 Forward	ccccgtcgacAAAAAATTTCGCC*AATTTAACGAAGACAGCGTGGTTAAAATTGCTTGTTCGGACggatccccc
EST2 Forward	ccccgtcgacAAAAAAT**C***AAATGGGTTTTTCGTTAGATCTCAATATTTCTTCAATACCAATACAGGTGTATTGAAGTTATTTAATGTTGTTAACGCTggatccccc
CTS2 Forward	ccccgtcgacAAAAAAT*CAATATTTATCAGTTATGATAACACTAAATCAGTCAAAACTAAGGCTGAATATGTGACACATAACAATggatccccc
FKS1 Forward	ccccgtagacAAATTTC**C**ACTACTAAGATTGGTGCTGGTATGGGTGAACAAATGTTATCTCGTGAATATTATTATCTGGGTACCCAATTACCAGTACggatccccc
FLR1 Forward	ccccgtcgacAAAAAAT**C**AT*CTTTCAGGGTGGATTGGAACGGCCCCAGTGATCCTGAGAACCCACAAAACTGGCCCggatccccc
NUP82 Forward	ccccgtcgacTTTAAAAAACGAAG*TAGTGAAAATCAGTTGGAAATTTTCACGGATATTTCCAAAGAAggatccccc
PPR1 Forward	ccccgtcgacTTTTTTTTTAAACAT*ATATTTGCTATTGGCCATGCTACGCAGGTACTTAAGTCAGATATTACTACTGTCGCGACAggatccccc
SPR6 Forward	ccccgtcgacAAAAAAAAT*AAGGAAACCAATCACTCTGGAGCATGGTTGCTTGTCAGGACCCGTGACTCTACGTTTCGGAAATTTTGCAGGAATCAGAGAggatccccc
TBF1 Forward	ccccgtcgacAAATTTATCTCAAGCATCCTTCATCAGCTGCATCTGCTACTGAAGAGGggatccccc

Table 11: Primers for Cloning

PCR primers used in this study for the amplification of genome encoded -1 PRF signals from a variety of yeast genes. Each amplicon was cloned into pJD375. DLR primers were then used to PCR amplify fragments from the dual luciferase reporter which where then subcloned into the *KpnI* site of *PGK1* (pJD741). Shown is the final amplicon size inserted into *PGK1* at position 497 (relative to the AUG start codon).

Parent Gene	Forward Primer (5' → 3')	Reverse Primer (3' ← 5')	Amplicon Size
BUB3	CCCCGTCGACCCTATACACGGCTG GCTCTG	CCCCGGATCCATCAGAAG TTGCCAGACATAG	222
EST2	CCCCGTCGACCATACCAAGGATGG AATGTATG	CCCCGGATCCTGATAAAT GAACCGTCCTCAC	267
CTS2	CCCCGTCGACGGAGGGCATCTGGC CTTAC	CCCCGGATCCAGCTTCAC CACATGACTCC	282
PPR1	CCCCGTCGACGTCAACATGCGTCC GATAGTTG	CCCCGGATCCCAATGCCT CCAACCTGTCTG	298
SPR6	CCCCCGTCGACGATGAGTCCAAGT AGGAAG	CCCCCGGATCCGGTAGCT TGCTGACATGCAC	198
TBF1	CCCCGTCGACTCAATCTCCAAATT CGTCAAC	CCCCGGATCCCATCCCAT CTTCTAAATGAGG	276
DLR	CCCCGGTACCTCGTTCGTTGAGCG AGTTC	CCCCGGTACCGGCGTCTT CCATGAGCTC	Varied

Table 12: Primers Used for Quantitative Real-Time PCR

Parent Gene	Forward Primer (5' → 3')	Reverse Primer (3' ← 5')	Amplicon Size
PGK1- *Renilla*	GTCGGTCCAGAAGTTGAAGC	TGAGAACTCGCTCAACGAAC	226
18S rRNA	GGAATTCCTAGTAAGCGCAAG	GCCTCACTAAGCCATTCAATC	99

Appendix E: PPCC Critical Values

Critical values for probability plot correlation coefficients used to determine if data is normally distributed. Shown are critical values for both the 1% and 5% significance levels for datasets of up to 100 samples (Filliben, 1975).

Table 13: PPCC Critical Values

N	0.01	0.05	N	0.01	0.05
4	0.8234	0.8666	**33**	0.9505	0.9661
5	0.8240	0.8786	**34**	0.9521	0.9671
6	0.8351	0.8880	**35**	0.9530	0.9678
7	0.8474	0.8970	**36**	0.9540	0.9686
8	0.8590	0.9043	**37**	0.9551	0.9693
9	0.8689	0.9115	**38**	0.9555	0.9700
10	0.8765	0.9173	**39**	0.9568	0.9704
11	0.8838	0.9223	**40**	0.9576	0.9712
12	0.8918	0.9267	**41**	0.9589	0.9719
13	0.8974	0.9310	**42**	0.9593	0.9723
14	0.9029	0.9343	**43**	0.9609	0.9730
15	0.9080	0.9376	**44**	0.9611	0.9734
16	0.9121	0.9405	**45**	0.9620	0.9739
17	0.9160	0.9433	**46**	0.9629	0.9744
18	0.9196	0.9452	**47**	0.9637	0.9748
19	0.9230	0.9479	**48**	0.9640	0.9753
20	0.9256	0.9498	**49**	0.9643	0.9758
21	0.9285	0.9515	**50**	0.9654	0.9761
22	0.9308	0.9535	**55**	0.9683	0.9781
23	0.9334	0.9548	**60**	0.9706	0.9797
24	0.9356	0.9564	**65**	0.9723	0.9809
25	0.9370	0.9575	**70**	0.9742	0.9822
26	0.9393	0.9590	**75**	0.9758	0.9831
27	0.9413	0.9600	**80**	0.9771	0.9841
28	0.9428	0.9615	**85**	0.9784	0.9850
29	0.9441	0.9622	**90**	0.9797	0.9857
30	0.9462	0.9634	**95**	0.9804	0.9864
31	0.9476	0.9644	**100**	0.9814	0.9869
32	0.9490	0.9652			

Appendix F: Minimum Corrected Sample Size

The corrected minimum sample size for normally distributed data can be estimated from the table below, assuming a 95% confidence interval and less than 100 samples in total. For uncorrected sample sizes larger than 100 or for a different confidence interval, please consult the previously published table in Kupper & Hafner (1989).

Table 14: Corrected Minimum Sample Size

Uncorrected, $\tilde{N}$	Corrected, N^*
5	10
10	16
15	22
20	28
25	34
30	40
35	46
40	51
45	57
50	62
55	68
60	74
65	79
70	85
75	90
80	95
85	101
90	106
95	112
100	117

Bibliography

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Amrani, N., Dong, S., He, F., Ganesan, R., Ghosh, S., Kervestin, S., Li, C., Mangus, D. A., Spatrick, P., and Jacobson, A. (2006). Aberrant termination triggers nonsense-mediated mRNA decay. Biochem Soc Trans *34*, 39-42.

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

Anantharaman, V., Koonin, E. V., and Aravind, L. (2002). Comparative genomics and evolution of proteins involved in RNA metabolism. Nucleic Acids Res *30*, 1427-1464.

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

Arava, Y., Wang, Y., Storey, J. D., Liu, C. L., Brown, P. O., and Herschlag, D. (2003). Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *100*, 3889-3894.

Barak, Z., Gallant, J., Lindsley, D., Kwieciszewki, B., and Heidel, D. (1996). Enhanced ribosome frameshifting in stationary phase cells. J Mol Biol *263*, 140-148.

Baranov, P. V., Gesteland, R. F., and Atkins, J. F. (2002). Recoding: translational bifurcations in gene expression. Gene *286*, 187-201.

Baranov, P. V., Gurvich, O. L., Hammer, A. W., Gesteland, R. F., and Atkins, J. F. (2003). Recode 2003. Nucleic Acids Res *31*, 87-89.

Baril, M., Dulude, D., Gendron, K., Lemay, G., and Brakier-Gingras, L. (2003). Efficiency of a programmed -1 ribosomal frameshift in the different subtypes of the human immunodeficiency virus type 1 group M. Rna *9*, 1246-1253.

Barrette, I., Poisson, G., Gendron, P., and Major, F. (2001). Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis and pattern searching. Nucleic Acids Res *29*, 753-758.

Bekaert, M., Bidou, L., Denise, A., Duchateau-Nguyen, G., Forest, J. P., Froidevaux, C., Hatin, I., Rousset, J. P., and Termier, M. (2003). Towards a computational model for -1 eukaryotic frameshifting sites. Bioinformatics *19*, 327-335.

Bekaert, M., Richard, H., Prum, B., and Rousset, J. P. (2005). Identification of programmed translational -1 frameshifting sites in the genome of Saccharomyces cerevisiae. Genome Res *15*, 1411-1420.

Blinkowa, A. L., and Walker, J. R. (1990). Programmed ribosomal frameshifting generates the Escherichia coli DNA polymerase III gamma subunit from within the tau subunit reading frame. Nucleic Acids Res *18*, 1725-1729.

Brengues, M., Teixeira, D., and Parker, R. (2005). Movement of eukaryotic mRNAs between polysomes and cytoplasmic processing bodies. Science *310*, 486-489.

Cao, D., and Parker, R. (2001). Computational modeling of eukaryotic mRNA turnover. Rna *7*, 1192-1212.

Cao, D., and Parker, R. (2003). Computational modeling and experimental analysis of nonsense-mediated decay in yeast. Cell *113*, 533-545.

Carrasco, L., Barbacid, M., and Vazquez, D. (1973). The trichodermin group of antibiotics, inhibitors of peptide bond formation by eukaryotic ribosomes. Biochim Biophys Acta *312*, 368-376.

Chamary, J. V., and Hurst, L. D. (2005). Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol *6*, R75.

Chambers, J. M. (1983). Graphical methods for data analysis (Boston: Duxbury Press).

Chia, W., Savakis, C., Karp, R., Pelham, H., and Ashburner, M. (1985). Mutation of the Adh gene of Drosophila melanogaster containing an internal tandem duplication. J Mol Biol *186*, 679-688.

Chiu, S. Y., Lejeune, F., Ranganathan, A. C., and Maquat, L. E. (2004). The pioneer translation initiation complex is functionally distinct from but structurally overlaps with the steady-state translation initiation complex. Genes Dev *18*, 745-754.

Clote, P., Ferre, F., Kranakis, E., and Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. Rna *11*, 578-591.

Cobucci-Ponzano, B., Rossi, M., and Moracci, M. (2005). Recoding in archaea. Mol Microbiol *55*, 339-348.

Cobucci-Ponzano, B., Trincone, A., Giordano, A., Rossi, M., and Moracci, M. (2003). Identification of an archaeal alpha-L-fucosidase encoded by an interrupted gene. Production of a functional enzyme by mutations mimicking programmed -1 frameshifting. J Biol Chem *278*, 14622-14631.

Coleman, H. M., Brierley, I., and Stevenson, P. G. (2003). An internal ribosome entry site directs translation of the murine gammaherpesvirus 68 MK3 open reading frame. J Virol *77*, 13093-13105.

Croarkin, C., and Tobias, P. (2004). NIST/SEMATECH e-Handbook of Statistical Methods.

Cui, Y., Dinman, J. D., Kinzy, T. G., and Peltz, S. W. (1998). The Mof2/Sui1 protein is a general monitor of translational accuracy. Mol Cell Biol *18*, 1506-1516.

Daar, I. O., and Maquat, L. E. (1988). Premature translation termination mediates triosephosphate isomerase mRNA degradation. Mol Cell Biol *8*, 802-813.

Dahlseid, J. N., Lew-Smith, J., Lelivelt, M. J., Enomoto, S., Ford, A., Desruisseaux, M., McClellan, M., Lue, N., Culbertson, M. R., and Berman, J. (2003). mRNAs encoding

telomerase components and regulators are controlled by UPF genes in Saccharomyces cerevisiae. Eukaryot Cell *2*, 134-142.

Devore, J. L. (2000). Probability and statistics for engineering and the sciences, 5th edn (Pacific Grove, CA: Duxbury).

Dinman, J. D., Icho, T., and Wickner, R. B. (1991). A -1 ribosomal frameshift in a double-stranded RNA virus of yeast forms a gag-pol fusion protein. Proc Natl Acad Sci U S A *88*, 174-178.

Dinman, J. D., and Kinzy, T. G. (1997). Translational misreading: mutations in translation elongation factor 1alpha differentially affect programmed ribosomal frameshifting and drug sensitivity. Rna *3*, 870-881.

Dinman, J. D., Ruiz-Echevarria, M. J., Czaplinski, K., and Peltz, S. W. (1997). Peptidyl-transferase inhibitors have antiviral properties by altering programmed -1 ribosomal frameshifting efficiencies: development of model systems. Proc Natl Acad Sci U S A *94*, 6606-6611.

Dinman, J. D., and Wickner, R. B. (1994). Translational maintenance of frame: mutants of Saccharomyces cerevisiae with altered -1 ribosomal frameshifting efficiencies. Genetics *136*, 75-86.

Dirks, R. M., and Pierce, N. A. (2004). An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. J Comput Chem *25*, 1295-1304.

Doma, M. K., and Parker, R. (2006). Endonucleolytic cleavage of eukaryotic mRNAs with translation elongation stalls. Nature *in press*.

Eleouet, J. F., Rasschaert, D., Lambert, P., Levy, L., Vende, P., and Laude, H. (1995). Complete sequence (20 kilobases) of the polyprotein-encoding gene 1 of transmissible gastroenteritis virus. Virology *206*, 817-822.

Farabaugh, P. J. (1996). Programmed translational frameshifting. Annu Rev Genet *30*, 507-528.

Fersht, A. (1999). Structure and mechanism in protein science : a guide to enzyme catalysis and protein folding (New York: W.H. Freeman).

Filliben, J. J. (1975). Probability Plot Correlation Coefficient Test for Normality. Technometrics *17*, 111-117.

Freyhult, E., Gardner, P. P., and Moulton, V. (2005). A comparison of RNA folding measures. BMC Bioinformatics *6*, 241.

Gao, X., Havecker, E. R., Baranov, P. V., Atkins, J. F., and Voytas, D. F. (2003). Translational recoding signals between gag and pol in diverse LTR retrotransposons. Rna *9*, 1422-1430.

Gonzalez, C. I., Bhattacharya, A., Wang, W., and Peltz, S. W. (2001). Nonsense-mediated mRNA decay in Saccharomyces cerevisiae. Gene *274*, 15-25.

Gonzalez, C. I., Ruiz-Echevarria, M. J., Vasudevan, S., Henry, M. F., and Peltz, S. W. (2000). The yeast hnRNP-like protein Hrp1/Nab4 marks a transcript for nonsense-mediated mRNA decay. Mol Cell *5*, 489-499.

Green, R. E., Lewis, B. P., Hillman, R. T., Blanchette, M., Lareau, L. F., Garnett, A. T., Rio, D. C., and Brenner, S. E. (2003). Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. Bioinformatics *19 Suppl 1*, i118-121.

Grentzmann, G., Ingram, J. A., Kelly, P. J., Gesteland, R. F., and Atkins, J. F. (1998). A dual-luciferase reporter system for studying recoding signals. Rna *4*, 479-486.

Grollman, A. P. (1967). Inhibitors of protein biosynthesis. II. Mode of action of anisomycin. J Biol Chem *242*, 3226-3233.

Gurvich, O. L., Baranov, P. V., Zhou, J., Hammer, A. W., Gesteland, R. F., and Atkins, J. F. (2003). Sequences that direct significant levels of frameshifting are frequent in coding regions of Escherichia coli. Embo J *22*, 5941-5950.

Hagan, K. W., Ruiz-Echevarria, M. J., Quan, Y., and Peltz, S. W. (1995). Characterization of cis-acting sequences and decay intermediates involved in nonsense-mediated mRNA turnover. Mol Cell Biol *15*, 809-823.

Hammell, A. B., Taylor, R. C., Peltz, S. W., and Dinman, J. D. (1999). Identification of putative programmed -1 ribosomal frameshift signals in large DNA databases. Genome Res *9*, 417-427.

Hansen, J. L., Moore, P. B., and Steitz, T. A. (2003). Structures of five antibiotics bound at the peptidyl transferase center of the large ribosomal subunit. J Mol Biol *330*, 1061-1075.

Harger, J. W., and Dinman, J. D. (2003). An in vivo dual-luciferase assay system for studying translational recoding in the yeast Saccharomyces cerevisiae. Rna *9*, 1019-1024.

Harger, J. W., Meskauskas, A., and Dinman, J. D. (2002). An "integrated model" of programmed ribosomal frameshifting. Trends Biochem Sci *27*, 448-454.

Harrison, P., Kumar, A., Lan, N., Echols, N., Snyder, M., and Gerstein, M. (2002). A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. J Mol Biol *316*, 409-419.

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

He, F., Peltz, S. W., Donahue, J. L., Rosbash, M., and Jacobson, A. (1993). Stabilization and ribosome association of unspliced pre-mRNAs in a yeast upf1- mutant. Proc Natl Acad Sci U S A *90*, 7034-7038.

Heikkinen, H. L., Llewellyn, S. A., and Barnes, C. A. (2003). Initiation-mediated mRNA decay in yeast affects heat-shock mRNAs, and works through decapping and 5'-to-3' hydrolysis. Nucleic Acids Res *31*, 4006-4016.

Herold, J., and Siddell, S. G. (1993). An 'elaborated' pseudoknot is required for high frequency frameshifting during translation of HCV 229E polymerase mRNA. Nucleic Acids Res *21*, 5838-5842.

Herrick, D., Parker, R., and Jacobson, A. (1990). Identification and comparison of stable and unstable mRNAs in Saccharomyces cerevisiae. Mol Cell Biol *10*, 2269-2284.

Hillman, R. T., Green, R. E., and Brenner, S. E. (2004). An unappreciated role for RNA surveillance. Genome Biol *5*, R8.

Horsburgh, B. C., Kollmus, H., Hauser, H., and Coen, D. M. (1996). Translational recoding induced by G-rich mRNA sequences that form unusual structures. Cell *86*, 949-959.

Imbert, I., Dimitrova, M., Kien, F., Kieny, M. P., and Schuster, C. (2003). Hepatitis C virus IRES efficiency is unaffected by the genomic RNA 3'NTR even in the presence of viral structural or non-structural proteins. J Gen Virol *84*, 1549-1557.

Inoue, H., Nojima, H., and Okayama, H. (1990). High efficiency transformation of Escherichia coli with plasmids. Gene *96*, 23-28.

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

Isshiki, M., Yamamoto, Y., Satoh, H., and Shimamoto, K. (2001). Nonsense-mediated decay of mutant waxy mRNA in rice. Plant Physiol *125*, 1388-1395.

Ito, H., Fukuda, Y., Murata, K., and Kimura, A. (1983). Transformation of intact yeast cells treated with alkali cations. J Bacteriol *153*, 163-168.

Jacks, T., Madhani, H. D., Masiarz, F. R., and Varmus, H. E. (1988a). Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region. Cell *55*, 447-458.

Jacks, T., Power, M. D., Masiarz, F. R., Luciw, P. A., Barr, P. J., and Varmus, H. E. (1988b). Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. Nature *331*, 280-283.

Jacks, T., Townsley, K., Varmus, H. E., and Majors, J. (1987). Two efficient ribosomal frameshifting events are required for synthesis of mouse mammary tumor virus gag-related polyproteins. Proc Natl Acad Sci U S A *84*, 4298-4302.

Jacks, T., and Varmus, H. E. (1985). Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting. Science *230*, 1237-1242.

Jacobs, J. L., and Dinman, J. D. (2004). Systematic analysis of bicistronic reporter assay data. Nucleic Acids Res *32*, e160.

Jamjoom, G. A., Naso, R. B., and Arlinghaus, R. B. (1977). Further characterization of intracellular precursor polyproteins of Rauscher leukemia virus. Virology *78*, 11-34.

Jofuku, K. D., Schipper, R. D., and Goldberg, R. B. (1989). A frameshift mutation prevents Kunitz trypsin inhibitor mRNA accumulation in soybean embryos. Plant Cell *1*, 427-435.

Katz, L., and Burge, C. B. (2003). Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res *13*, 2042-2051.

Kebaara, B., Nazarenus, T., Taylor, R., Forch, A., and Atkin, A. L. (2003). The Upf-dependent decay of wild-type PPR1 mRNA depends on its 5'-UTR and first 92 ORF nucleotides. Nucleic Acids Res *31*, 3157-3165.

Keeling, K. M., Lanier, J., Du, M., Salas-Marco, J., Gao, L., Kaenjak-Angeletti, A., and Bedwell, D. M. (2004). Leaky termination at premature stop codons antagonizes nonsense-mediated mRNA decay in S. cerevisiae. Rna *10*, 691-703.

Kendall, M. G., Stuart, A., Ord, J. K., and O'Hagan, A. (1994). Kendall's advanced theory of statistics, Vol 1, 6th edn (New York: Halsted Press).

Kim, Y. K., Furic, L., Desgroseillers, L., and Maquat, L. E. (2005). Mammalian Staufen1 recruits Upf1 to specific mRNA 3'UTRs so as to elicit mRNA decay. Cell *120*, 195-208.

Kinniburgh, A. J., Maquat, L. E., Schedl, T., Rachmilewitz, E., and Ross, J. (1982). mRNA-deficient beta o-thalassemia results from a single nucleotide deletion. Nucleic Acids Res *10*, 5421-5427.

Kiparisov, S., Petrov, A., Meskauskas, A., Sergiev, P. V., Dontsova, O. A., and Dinman, J. D. (2005). Structural and functional analysis of 5S rRNA in Saccharomyces cerevisiae. Mol Genet Genomics *274*, 235-247.

Kollmus, H., Flohe, L., and McCarthy, J. E. (1996a). Analysis of eukaryotic mRNA structures directing cotranslational incorporation of selenocysteine. Nucleic Acids Res *24*, 1195-1201.

Kollmus, H., Hentze, M. W., and Hauser, H. (1996b). Regulated ribosomal frameshifting by an RNA-protein interaction. Rna *2*, 316-323.

Kontos, H., Napthine, S., and Brierley, I. (2001). Ribosomal pausing at a frameshifter RNA pseudoknot is sensitive to reading phase but shows little correlation with frameshift efficiency. Mol Cell Biol *21*, 8657-8670.

Koopman, R. (2004). Personal Communication.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature *304*, 412-417.

Kupper, L. L., and Hafner, K. B. (1989). How Appropriate Are Popular Sample-Size Formulas. American Statistician *43*, 101-105.

Le, S. Y., Chen, J. H., and Maizel, J. V. (1989). Thermodynamic stability and statistical significance of potential stem-loop structures situated at the frameshift sites of retroviruses. Nucleic Acids Res *17*, 6143-6152.

Le, S. Y., Liu, W. M., Chen, J. H., and Maizel, J. V., Jr. (2001). Local thermodynamic stability scores are well represented by a non-central student's t distribution. J Theor Biol *210*, 411-423.

Le, S. Y., Zhang, K., and Maizel, J. V., Jr. (2002). RNA molecules with structure dependent functions are uniquely folded. Nucleic Acids Res *30*, 3574-3582.

Leeds, P., Peltz, S. W., Jacobson, A., and Culbertson, M. R. (1991). The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing a premature translational termination codon. Genes Dev *5*, 2303-2314.

Leger, M., Sidani, S., and Brakier-Gingras, L. (2004). A reassessment of the response of the bacterial ribosome to the frameshift stimulatory signal of the human immunodeficiency virus type 1. Rna *10*, 1225-1235.

Lejeune, F., Li, X., and Maquat, L. E. (2003). Nonsense-mediated mRNA decay in mammalian cells involves decapping, deadenylating, and exonucleolytic activities. Mol Cell *12*, 675-687.

Lelivelt, M. J., and Culbertson, M. R. (1999). Yeast Upf proteins required for RNA surveillance affect global expression of the yeast transcriptome. Mol Cell Biol *19*, 6710-6719.

Lew, J. E., Enomoto, S., and Berman, J. (1998). Telomere length regulation and telomeric chromatin require the nonsense-mediated mRNA decay pathway. Mol Cell Biol *18*, 6121-6130.

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods *25*, 402-408.

Lopinski, J. D., Dinman, J. D., and Bruenn, J. A. (2000). Kinetics of ribosomal pausing during programmed -1 translational frameshifting. Mol Cell Biol *20*, 1095-1103.

Losson, R., Fuchs, R. P., and Lacroute, F. (1983). In vivo transcription of a eukaryotic regulatory gene. Embo J *2*, 2179-2184.

Losson, R., and Lacroute, F. (1979). Interference of nonsense mutations with eukaryotic messenger RNA stability. Proc Natl Acad Sci U S A *76*, 5134-5137.

Lynch, M., and Kewalramani, A. (2003). Messenger RNA surveillance and the evolutionary proliferation of introns. Mol Biol Evol *20*, 563-571.

Lyngso, R. B., and Pedersen, C. N. (2000). RNA pseudoknot prediction in energy-based models. J Comput Biol *7*, 409-427.

Macke, T. J., Ecker, D. J., Gutell, R. R., Gautheret, D., Case, D. A., and Sampath, R. (2001). RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res *29*, 4724-4735.

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

Manktelow, E., Shigemoto, K., and Brierley, I. (2005). Characterization of the frameshift signal of Edr, a mammalian example of programmed -1 ribosomal frameshifting. Nucleic Acids Res *33*, 1553-1563.

Maquat, L. E. (2004a). Nonsense-mediated mRNA decay: A comparative analysis of different species. Current Genomics *5*, 175-190.

Maquat, L. E. (2004b). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol *5*, 89-99.

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

Marra, M. A., Jones, S. J., Astell, C. R., Holt, R. A., Brooks-Wilson, A., Butterfield, Y. S., Khattra, J., Asano, J. K., Barber, S. A., Chan, S. Y.*, et al.* (2003). The Genome sequence of the SARS-associated coronavirus. Science *300*, 1399-1404.

Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol *288*, 911-940.

Matsufuji, S., Matsufuji, T., Wills, N. M., Gesteland, R. F., and Atkins, J. F. (1996). Reading two bases twice: mammalian antizyme frameshifting in yeast. Embo J *15*, 1360-1370.

Mazumder, B., Sampath, P., Seshadri, V., Maitra, R. K., DiCorleto, P. E., and Fox, P. L. (2003). Regulated release of L13a from the 60S ribosomal subunit as a mechanism of transcript-specific translational control. Cell *115*, 187-198.

Mendell, J. T., Sharifi, N. A., Meyers, J. L., Martinez-Murillo, F., and Dietz, H. C. (2004). Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. Nat Genet *36*, 1073-1078.

Meskauskas, A., Harger, J. W., Jacobs, K. L., and Dinman, J. D. (2003). Decreased peptidyltransferase activity correlates with increased programmed -1 ribosomal frameshifting and viral maintenance defects in the yeast Saccharomyces cerevisiae. Rna *9*, 982-992.

Meyer, I. M., and Miklos, I. (2005). Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. Nucleic Acids Res *33*, 6338-6348.

Mitchell, P., and Tollervey, D. (2003). An NMD pathway in yeast involving accelerated deadenylation and exosome-mediated 3'-->5' degradation. Mol Cell *11*, 1405-1413.

Mitrovich, Q. M., and Anderson, P. (2000). Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans. Genes Dev *14*, 2173-2184.

Miyao, T., Barnett, J. D., and Woychik, N. A. (2001). Deletion of the RNA polymerase subunit RPB4 acts as a global, not stress-specific, shut-off switch for RNA polymerase II transcription at high temperatures. J Biol Chem *276*, 46408-46413.

Moon, S., Byun, Y., Kim, H.-J., Jeong, S., and Han, K. (2004). Predicting genes expressed via -1 and +1 frameshifts. Nucl Acids Res *32*, 4884-4892.

Moriarty, P. M., Reddy, C. C., and Maquat, L. E. (1998). Selenium deficiency reduces the abundance of mRNA for Se-dependent glutathione peroxidase 1 by a UGA-dependent mechanism likely to be nonsense codon-mediated decay of cytoplasmic mRNA. Mol Cell Biol *18*, 2932-2939.

Morris, D. K., and Lundblad, V. (1997). Programmed translational frameshifting in a gene required for yeast telomere replication. Curr Biol *7*, 969-976.

Muhlrad, D., and Parker, R. (1994). Premature translational termination triggers mRNA decapping. Nature *370*, 578-581.

Muhlrad, D., and Parker, R. (1999). Recognition of yeast mRNAs as "nonsense containing" leads to both inhibition of mRNA translation and mRNA degradation: implications for the control of mRNA decapping. Mol Biol Cell *10*, 3971-3978.

Muldoon-Jacobs, K. L., and Dinman, J. D. (2006). Specific Effects of Ribosome-Tethered Molecular Chaperones on Programmed -1 Ribosomal Frameshifting. . Eukaryotic Cell *in press*.

Namy, O., Duchateau-Nguyen, G., Hatin, I., Hermann-Le Denmat, S., Termier, M., and Rousset, J. P. (2003). Identification of stop codon readthrough genes in Saccharomyces cerevisiae. Nucleic Acids Res *31*, 2289-2296.

Namy, O., Rousset, J. P., Napthine, S., and Brierley, I. (2004). Reprogrammed genetic decoding in cellular gene expression. Mol Cell *13*, 157-168.

Nonet, M., Scafe, C., Sexton, J., and Young, R. (1987). Eucaryotic RNA polymerase conditional mutant that rapidly ceases mRNA synthesis. Mol Cell Biol *7*, 1602-1611.

Novac, O., Guenier, A. S., and Pelletier, J. (2004). Inhibitors of protein synthesis identified by a high throughput multiplexed translation screen. Nucleic Acids Res *32*, 902-915.

Page, R. D. (1996). TreeView: an application to display phylogenetic trees on personal computers. Comput Appl Biosci *12*, 357-358.

Pan, Q., Saltzman, A. L., Kim, Y. K., Misquitta, C., Shai, O., Maquat, L. E., Frey, B. J., and Blencowe, B. J. (2006). Quantitative microarray profiling provides evidence against widespread coupling of alternative splicing with nonsense-mediated mRNA decay to control gene expression. Genes Dev *20*, 153-158.

Pearson, W. R. (2000). Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol *132*, 185-219.

Pelsy, F., and Lacroute, F. (1984). Effect of Ochre Nonsense Mutations on Yeast Ura1 Messenger-Rna Stability. Current Genetics *8*, 277-282.

Peltz, S. W., Hammell, A. B., Cui, Y., Yasenchak, J., Puljanowski, L., and Dinman, J. D. (1999). Ribosomal protein L3 mutants alter translational fidelity and promote rapid loss of the yeast killer virus. Mol Cell Biol *19*, 384-391.

Pierrat, B., Lacroute, F., and Losson, R. (1993). The 5' untranslated region of the PPR1 regulatory gene dictates rapid mRNA decay in yeast. Gene *131*, 43-51.

Plant, E. P., and Dinman, J. D. (2005). Torsional restraint: a new twist on frameshifting pseudoknots. Nucleic Acids Res *33*, 1825-1833.

Plant, E. P., Jacobs, K. L., Harger, J. W., Meskauskas, A., Jacobs, J. L., Baxter, J. L., Petrov, A. N., and Dinman, J. D. (2003). The 9-A solution: how mRNA pseudoknots promote efficient programmed -1 ribosomal frameshifting. Rna *9*, 168-174.

Plant, E. P., Perez-Alvarado, G. C., Jacobs, J. L., Mukhopadhyay, B., Hennig, M., and Dinman, J. D. (2005). A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. PLoS Biol *3*, e172.

Plant, E. P., Wang, P., Jacobs, J. L., and Dinman, J. D. (2004). A programmed -1 ribosomal frameshift signal can function as a cis-acting mRNA destabilizing element. Nucleic Acids Res *32*, 784-790.

Pulak, R., and Anderson, P. (1993). mRNA surveillance by the Caenorhabditis elegans smg genes. Genes Dev *7*, 1885-1897.

Rehwinkel, J., Letunic, I., Raes, J., Bork, P., and Izaurralde, E. (2005). Nonsense-mediated mRNA decay factors act in concert to regulate common mRNA targets. Rna *11*, 1530-1544.

Ringner, M., and Krogh, M. (2005). Folding Free Energies of 5'-UTRs Impact Post-Transcriptional Regulation on a Genomic Scale in Yeast. PLoS Comput Biol *1*, e72.

Rivas, E., and Eddy, S. R. (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol *285*, 2053-2068.

Rivas, E., and Eddy, S. R. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics *16*, 583-605.

Ruiz-Echevarria, M. J., and Peltz, S. W. (2000). The RNA binding protein Pub1 modulates the stability of transcripts containing upstream open reading frames. Cell *101*, 741-751.

Sachs, M. S., Wang, Z., Gaba, A., Fang, P., Belk, J., Ganesan, R., Amrani, N., and Jacobson, A. (2002). Toeprint analysis of the positioning of translation apparatus components at initiation and termination codons of fungal mRNAs. Methods *26*, 105-114.

Schindler, D., and Davies, J. (1975). Inhibitors of macromolecular synthesis in yeast. Methods Cell Biol *12*, 17-38.

Schultes, E. A., Hraber, P. T., and LaBean, T. H. (1999). Estimating the contributions of selection and self-organization in RNA secondary structure. J Mol Evol *49*, 76-83.

Seffens, W., and Digby, D. (1999). mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res *27*, 1578-1584.

Sekine, Y., and Ohtsubo, E. (1989). Frameshifting is required for production of the transposase encoded by insertion sequence 1. Proc Natl Acad Sci U S A *86*, 4609-4613.

Shah, A. A., Giddings, M. C., Parvaz, J. B., Gesteland, R. F., Atkins, J. F., and Ivanov, I. P. (2002). Computational identification of putative programmed translational frameshift sites. Bioinformatics *18*, 1046-1053.

Sherf, B. A., Navarro, S. L., Hannah, R. R., and Wood, K. V. (1996). Dual-Luciferase Reporter Assay: An Advanced Co-Reporter Technology Integrating Firefly and Renilla Luciferase Assays, In Promega Notes Magazine, pp. 2.

Sheth, U., and Parker, R. (2003). Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. Science *300*, 805-808.

Shigemoto, K., Brennan, J., Walls, E., Watson, C. J., Stott, D., Rigby, P. W., and Reith, A. D. (2001). Identification and characterisation of a developmentally regulated mammalian gene that utilises -1 programmed ribosomal frameshifting. Nucleic Acids Res *29*, 4079-4088.

Somogyi, P., Jenner, A. J., Brierley, I., and Inglis, S. C. (1993). Ribosomal pausing during translation of an RNA pseudoknot. Mol Cell Biol *13*, 6931-6940.

Stahl, G., Salem, S. N., Chen, L., Zhao, B., and Farabaugh, P. J. (2004). Translational accuracy during exponential, postdiauxic, and stationary growth phases in Saccharomyces cerevisiae. Eukaryot Cell *3*, 331-338.

Sunohara, T., Jojima, K., Tagami, H., Inada, T., and Aiba, H. (2004). Ribosome stalling during translation elongation induces cleavage of mRNA being translated in Escherichia coli. J Biol Chem *279*, 15368-15375.

Taylor, R., Kebaara, B. W., Nazarenus, T., Jones, A., Yamanaka, R., Uhrenholdt, R., Wendler, J. P., and Atkin, A. L. (2005). Gene set coregulated by the Saccharomyces cerevisiae nonsense-mediated mRNA decay pathway. Eukaryot Cell *4*, 2066-2077.

Thiel, V., Ivanov, K. A., Putics, A., Hertzig, T., Schelle, B., Bayer, S., Weissbrich, B., Snijder, E. J., Rabenau, H., Doerr, H. W.*, et al.* (2003). Mechanisms and enzymes involved in SARS coronavirus genome expression. J Gen Virol *84*, 2305-2315.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-4680.

Tsuchihashi, Z., and Kornberg, A. (1990). Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. Proc Natl Acad Sci U S A *87*, 2516-2520.

Tuplin, A., Wood, J., Evans, D. J., Patel, A. H., and Simmonds, P. (2002). Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus. Rna *8*, 824-841.

Venkatesan, A., Sharma, R., and Dasgupta, A. (2003). Cell cycle regulation of hepatitis C and encephalomyocarditis virus internal ribosome entry site-mediated translation in human embryonic kidney 293 cells. Virus Res *94*, 85-95.

Voelker, T. A., Moreno, J., and Chrispeels, M. J. (1990). Expression analysis of a pseudogene in transgenic tobacco: a frameshift mutation prevents mRNA accumulation. Plant Cell *2*, 255-261.

Wall, L., Christiansen, T., and Orwant, J. (2000). Programming Perl, 3rd edn (Beijing ; Cambridge, Mass.: O'Reilly).

Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D., and Brown, P. O. (2002). Precision and functional specificity in mRNA decay. Proc Natl Acad Sci U S A *99*, 5860-5865.

Weil, J. E., and Beemon, K. L. (2006). A 3' UTR sequence stabilizes termination codons in the unspliced RNA of Rous sarcoma virus. Rna *12*, 102-110.

Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A., and Rapp, B. A. (2000). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res *28*, 10-14.

Wilkinson, M. F. (2005). A new function for nonsense-mediated mRNA-decay factors. Trends Genet *21*, 143-148.

Williamson, N. A., Raliegh, J., Morrice, N. A., and Wettenhall, R. E. (1997). Post-translational processing of rat ribosomal proteins. Ubiquitous methylation of Lys22 within the zinc-finger motif of RL40 (carboxy-terminal extension protein 52) and tissue-specific methylation of Lys4 in RL29. Eur J Biochem *246*, 786-793.

Wills, N. M., Moore, B., Hammer, A., Gesteland, R. F., and Atkins, J. F. (2006). A functional -1 ribosomal frameshift signal in the human paraneoplastic Ma3 gene. J Biol Chem.

Wittmann, J., Hol, E. M., and Jack, H. M. (2006). hUPF2 silencing identifies physiologic substrates of mammalian nonsense-mediated mRNA decay. Mol Cell Biol *26*, 1272-1287.

Workman, C., and Krogh, A. (1999). No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. Nucleic Acids Res *27*, 4816-4822.

Xing, Y., and Lee, C. J. (2004). Negative selection pressure against premature protein truncation is reduced by alternative splicing and diploidy. Trends Genet *20*, 472-475.