The role of the coding sequence in regulating human mRNA stability

by

Ashrut Mohan Narula

A thesis submitted in conformity with the requirements
for the degree of Master of Science

Department of Molecular Genetics
University of Toronto

The role of the coding sequence in regulating human mRNA stability

Ashrut Mohan Narula

Master of Science

Department of Molecular Genetics
University of Toronto

2018

Abstract

Proper control of mRNA stability is key in the regulation of gene expression, with half-lives varying considerably transcriptome-wide. While sequence elements regulating mRNA decay are classically attributed to the untranslated regions (UTRs) of the transcript, recent work across a range of non-human model systems has uncovered a conserved decay mechanism triggered by codon usage within the coding sequence (CDS). Here, I use the human ORFeome (hORFeome) collection, a library of human CDSs flanked by universal synthetic UTRs, to demonstrate that hORF constructs with different CDSs possess a wide range of stability comparable to that of endogenous transcripts. This CDS-mediated diversity in mRNA decay requires translation. I further identify codon and amino acid usage to be key elements driving this variation in stability and outline their relative impacts in yeast and humans. Overall, the hORFeome collection represents a powerful framework for deconvoluting the effects of CDSs and UTRs on stability transcriptome-wide.

Acknowledgements

First and foremost, thanks to my supervisors Dr. Olivia Rissland and Dr. James Ellis. To Olivia, you have been an exemplary mentor, going out of your way to foster the growth of your trainees not just by providing valuable scientific direction through the ups and downs of this project, but also by inculcating in your students a historical perspective of science, creating leadership opportunities, and making sure stomachs never went empty (mini-fridge!). I'm sincerely grateful that you have always made yourself available to provide guidance and ensured the momentum of my project never wavered despite external factors. To James, I am grateful that you went out of your way to make sure I was welcome in the lab – it is rare to find a set of trainees that exclusively have positive things to say about their supervisor, so the decision to finish this project in your lab was a no-brainer. Your strong sense of ethics and overall trust in your trainee's intrinsic drive to advance their projects are key traits that I will keep in mind if I ever serve in a managerial role.

I would also like to extend my appreciation to my committee members Dr. Craig Smibert and Dr. Mikko Taipale, who I chose for this role explicitly for their openness in sharing scientific advice and their commitment to mentoring graduate students. To Craig, I admire your breadth of knowledge and your ability to ask an insightful question for every situation, and will always be indebted for your routine supply of donuts during marathon undergrad lab sessions. To Mikko, I have been inspired by your generosity of ideas and the openness with which you let me complete key aspects of this project in your lab. I would also like to thank Dr. Julie Claycomb, Dr. Andrew Spence, and members of the Claycomb lab for insightful discussion at lab meetings.

Thanks to all those that have played a collaborative role and contributed toward this project. Dr. Jeff Coller and his lab, particularly Megan Forrest and Gavin Hanson, for their openness in sharing and discussing data, methods, and conclusions that have been critical for the direction of this project. Dr. Jason Moffat and his lab, particularly Olga Sizova and Amy Tong for facilitating the transfer of the hORFeome collection, as well as Patricia Mero for guidance in lentiviral production. Dr. Mikko Taipale and his lab, specifically Nader Alerasool for generating pilot stable cell lines, guidance with lentivirus production, and overall great discussion. I also owe a debt of gratitude to all funding sources that have contributed to this project, including grants from the University of Toronto, CIHR, CF Canada, and the Hospital for Sick Children.

Thanks to all members of the Rissland and Ellis labs. First, to Dr. Beth Nicholson and Peter Pasceri, your hard work and dedication to keeping everything running smoothly may seem unnoticed at times but is very evident and much appreciated. To my computational counselors Andrew Lugowski and Marat Mufteev, thank you for so patiently guiding a wet lab biologist into the world of bioinformatics and data science. To Dr. Deivid Rodrigues for your guidance in the lab and post-transcriptional discussion outside of it. And finally, to all lab members for invaluable helping hands, constant support and friendship, and overall generosity with baked goods.

Finally, I would like to thank all the numerous mentors that have played a part in my academic journey at the University of Toronto. Specifically, Maria, Dave, Serge, and others for all the help through past projects and for inspiring me to stay the course and take the plunge into graduate school. To all the fellow students that have so greatly shaped my graduate experience and allowed me to vent about this project over numerous beers. And finally, to my close friends and family for their invaluable role in founding who I am today.

Table of Contents

List of Figures

Chapter 1 Introduction

1.1 Post transcriptional regulation

1.1.1 An overview of post transcriptional regulation

In eukaryotic cells, genetic information is stored in the form of DNA molecules in the nucleus. RNA Polymerase II transcribes DNA into messenger RNA (mRNA) transcripts, which are exported to the cell's cytoplasmic compartment and translated into proteins by ribosomal machinery (Crick, 1970). Proteins then act as effector molecules and carry out a wide range of functional activities. Cells have evolved complex mechanisms to regulate the spatial and temporal abundance of proteins, allowing for finer control of functional activity. One way by which cells regulate how much protein is expressed is by controlling the cytoplasmic fate of mRNAs: mRNAs can be differentially localized to the various regions of the cell (Holt and Bullock, 2009; Meignin and Davis, 2010), translated into proteins at variable rates (Kong and Lasko, 2012), sequestered in specialized compartments when not needed (Decker and Parker, 2012), and degraded in a controlled manner (Pérez-Ortín et al., 2013; Schoenberg and Maquat, 2012; Wu and Brewer, 2012).

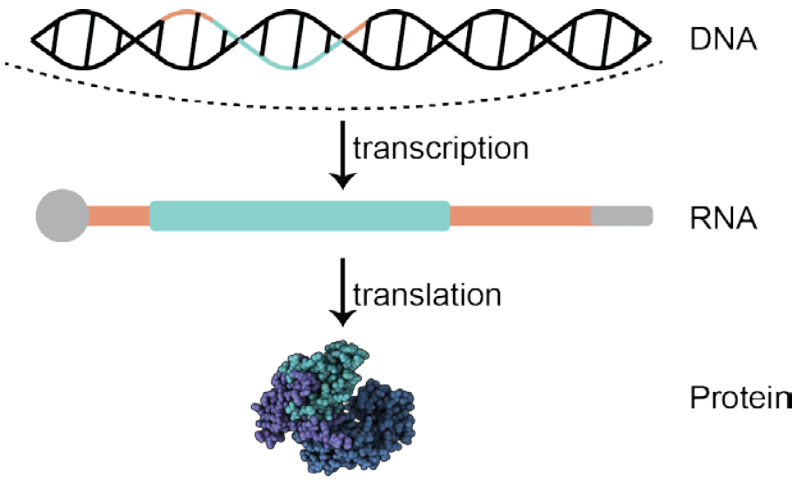

Figure 1. Central dogma of molecular biology. Genetic information is stored in the nuclear compartment in the form of DNA, transcribed into RNA, transported to the cytoplasm, and translated into proteins by ribosomal machinery.

1.1.2 Structure of the mature mRNA

Following transcription, the vast majority of pre-mRNA molecules undergo 5′ capping, intron splicing, and 3′ cleavage and polyadenylation to produce a mature transcript. The 5′ 7-methylguanosine cap and 3′ poly-adenosine (poly(A)) tail are near-universal mRNA features. They play key roles in translation and decay, serving as binding sites for proteins involved in initiating translation, while also protecting mRNAs from exonucleolytic decay factors (Rissland, 2016). At the sequence level, mRNAs can be divided into a 5′ untranslated region (5′ UTR), coding sequence (CDS) (also known as open reading frame (ORF)), and a 3′ untranslated region (3′ UTR). UTRs harbor a variety of cis regulatory sequence elements that can be bound by trans acting factors such as RNA binding proteins (RBPs) and microRNAs (miRNAs) to regulate mRNA localization, translation, and stability (Halbeisen et al., 2008). In contrast, the CDS dictates the order in which amino acids (AAs) are added to synthesize a protein product and are not traditionally thought to play a regulatory role. This thesis explores the under-characterised regulatory potential of the mRNA CDS.

Figure 2. Structure of a canonical mRNA. The majority of mature mRNAs have a 5′ 7-methylguanosine cap, followed by a 5′ untranslated region, coding sequence, 3′ untranslated region, and poly(A) tail.

1.1.3 mRNA translation – the initiation, elongation, and termination phases

Translation refers to the process by which the information carried in the CDS is converted into a protein molecule, and it can be subcategorised into initiation, elongation, and termination phases. The vast majority of eukaryotic translation is initiated by the interaction of translation initiation factors (eIFs) with the 5′ cap. First, an initiator Met-tRNAi is loaded into the 40S small ribosomal subunit along with additional eIFs. This pre-initiation complex assembles at the 5′ cap via eIF4F, a complex comprising the cap-binding protein eIF4E, a DEAD-Box RNA helicase eIF4A, and a large scaffold protein called eIF4G. The small subunit then scans the mRNA for the first AUG start codon, a process facilitated by additional helicases. At the AUG, eIFs dissociate and the 60S large ribosomal subunit assembles to form the complete 80S ribosome, primed to begin protein synthesis (Jackson et al., 2010).

The information in the CDS consists of adjacent, nonoverlapping triplet nucleotides called codons, which are recognised by transfer RNA (tRNA) adaptor molecules with complementary

anticodon sequences. An amino acid (AA) corresponding to the anticodon is covalently attached to each tRNA's 3′ end by aminoacyl tRNA synthetases through a process called charging. Ribosomes have three major binding sites for tRNAs: the A (aminoacyl), P (peptidyl), and E (exit) sites. With each elongation cycle, the ribosomal machinery recruits a charged tRNA complementary to the codon in the A-site by a process called decoding. In the second translocation step, a peptide bond forms between the AAs carried by tRNAs at the A- and P-sites, and the growing polypeptide chain is transferred to the P-site. Finally, after a third translocation, this tRNA transfers the nascent protein to the next tRNA, and leaves the E-site of the ribosome to be charged again in the cytosol. Elongation factors (EFs) enter and leave the ribosome during each cycle of elongation to facilitate these processes. (Dever et al., 2018).

Translation termination occurs when the ribosome reaches one of three stop codons (UAA, UAG, UGA). These codons are not recognised by a tRNA but instead are bound by release factors. Together, these factors free the carboxyl end of the growing peptide chain to produce a mature protein and split the ribosome so that a new translation cycle can begin again. (Dever and Green, 2012).

1.1.4 Regulation of translation initiation modulates translation efficiency

Translation efficiency (TE) is defined as the number of protein molecules produced per mRNA in a unit time (Humphreys, 1969). One common way of measuring TE is through deep sequencing of ribosome-protecting mRNA fragments (ribosome profiling) to measure mRNA translation *in vivo*. Together with older microarray and northern-blotting based studies, this technique has conclusively demonstrated that mRNAs from different genes are translated at different rates with considerable variation genome-wide (Fields et al., 2015; Ingolia et al., 2009). In eukaryotic cells, translation initiation is considered the rate-limiting factor determining TE (Shah et al., 2013; Weinberg et al., 2016) and is thus tightly controlled to regulate protein output (Sonenberg and Hinnebusch, 2009).

Considerable regulation of initiation occurs in the 5′ UTR. One mechanism of controlling translation comes from secondary structure, where increased structure reduces translational efficiency (Kertesz et al., 2010). The sequence context of the AUG start codon can also alter AUG recognition efficiency by the scanning ribosome, with Kozak consensus sequences increasing TE in eukaryotes (Kozak, 1987). In addition, upstream open reading frames (uORFs)

are short CDSs within the 5′ UTR that exist in ~50% of mammalian mRNAs and compete for scanning ribosomes, thus altering the translation of downstream CDSs (Hinnebusch et al., 2016).

The 5′ cap and its binding partners provide additional regulatory opportunities. The most well studied translational repressors are eIF4E binding proteins (4EBPs), which block eIF4E-eIF4G binding in the absence of mTOR activation, preferentially inhibiting a sub-class of mRNAs containing a 5′ Terminal Oligopyrimidine Tract (5′ TOP) motif. When phosphorylated by mTOR, 4EBPs release eIF4E to allow translation to initiate on these transcripts (Fonseca et al., 2014). The 3′UTR also serves as a binding site for an array of RBPs and miRNAs that function in diverse modes of translational regulation (Szostak and Gebauer, 2013). Overall, translation initiation is a stringently controlled process with various mechanisms modulating protein output.

1.1.5 Codon usage regulates translation elongation and translation efficiency

With 61 non-stop codons encoding the entire pool of 20 AAs (Nirenberg et al., 1965), the genetic code holds considerable redundancy. Most AAs are encoded by multiple (i.e., synonymous) codons; however corresponding tRNAs are often expressed at different levels, which can lead to different rates of decoding. Moreover, most tRNA anticodon triplets recognize more than one synonymous codon through non-Watson-Crick “wobble” base pairing at the third position of the codon (Agris et al., 2007; Crick, 1966). These different tRNAs encoding the same amino acid are termed isoacceptors. In fact, some isoacceptors are entirely absent from the genome so that decoding of certain codons relies entirely on wobble interactions at the third position.

Given this large variation in the abundance and decoding rates of tRNAs for each set of synonymous codons, it was hypothesised that certain “optimal” codons are more efficiently and accurately decoded than their counterparts. Evidence supporting this codon-dependent regulation of translation elongation rates has come from recent translation inhibitor-free ribosome profiling experiments, which suggest that optimal codons have reduced ribosome occupancy *in vivo* (Hussmann et al., 2015; Weinberg et al., 2016). Furthermore, synthetic codon-optimised reporters demonstrate faster translation elongation rates in cell-free systems (Yu et al., 2015), *in cellulo* (Yan et al., 2016), and in ribosome run-off experiments (Presnyak et al., 2015). Consistent with the importance of decoding for overall elongation speed, recent work indicates that the ribosomal A-site decoding rate is more influential than AA-specific peptide bond formation rates in regulating translation elongation in *S. cerevisiae* (Hanson et al., 2018).

While the effects of codon usage on translation elongation are evident, it has also been found that codon-optimised mRNAs have higher TEs (Hanson et al., 2018; Tuller et al., 2010). One hypothesis is that translation elongation rates propagate back to modulate translation initiation and affect TE. In fact, it has been shown that codon-dependent slowing of ribosome translocation around the start codon results in crowding, and restricts translation initiation rates (Chu et al., 2014). Further, given that free ribosomes are predicted to be under limited supply (Chu and Haar, 2012; Shah et al., 2013), it has also been proposed that faster translation elongation may allow for increased ribosome turnover and more efficient loading and initiation by the resultant free ribosome onto the mRNA (Rodnina, 2016). It is important to note however, that CDS codon optimality and elongation rates have likely co-evolved with initiation rates in the 5′ UTR to regulate total protein synthesis. It is thus important to deconvolute the effects of these two processes on protein output in a systematic manner. Nevertheless, there is considerable evidence that protein output (TE) is higher for mRNAs enriched in optimal codons, and lower for non-optimal mRNAs.

Given that synonymous codons can have differential effects on elongation and TE without affecting the protein product produced, a clear opportunity arises for cells to co-opt codons for additional regulatory purposes. Different subsets of genes use codons at different frequencies, presumably to control translation rates (Akashi, 2003; Hanson and Coller, 2017). In fact, while genomes and transcriptomes exhibit an overall bias toward optimal codons, highly expressed “housekeeping” genes incorporate them at a particularly high frequency (Sharp and Li, 1985). Interestingly, the distribution of synonymous codon usage across the CDS is also not entirely random. Non-optimal codons have been postulated as being used as a mechanism by which cells can control co-translational protein folding (Buhr et al., 2016; Zhou et al., 2015), while highly conserved regions of mRNAs associated with key structural domain residues tend to show increased codon optimality (Saunders and Deane, 2010; Zhou et al., 2007). Further, non-optimal codon-dependent slowing of translation elongation also exposes hidden AA residues, allowing for them to be post-translationally modified and subject to faster degradation (Zhang et al., 2010). Taken together, codon usage is a key tool to control gene expression.

1.1.6 tRNA pools shape gene expression, and their misregulation drives disease

As mentioned, tRNA levels appear to vary between different tissues (Dittmar et al., 2006) and species (Plotkin and Kudla, 2011), and codon usage in these different cells appears to have co-

evolved with its corresponding tRNA milieu (Plotkin et al., 2004). For instance, brain tissues appear to have a particularly unique profile of codon usage and tRNA levels (Dittmar et al., 2006; Plotkin et al., 2004). As such, a hypothesis has arisen suggesting that tRNA levels are an important regulator of gene expression.

In support of this hypothesis, studies of ribosome occupancy in *E. coli* suggest that AA starvation alters tRNA concentrations, resulting in increased ribosomal pausing and reduced global protein synthesis, as well as an upregulation of a small subset of genes (Dittmar et al., 2005; Subramaniam et al., 2014). In *S. cerevisiae*, oxidative stress reprograms the tRNA modification status, which enhances the translation of certain codons, and increases TE of genes required for effective resistance to oxidative conditions (Chan et al., 2012). In human cell lines as well, optimal transcripts are more sensitive to starvation-dependent alterations in charged tRNA pools, while the translation of non-optimal mRNAs is favoured to activate stress-response genes (Saikia et al., 2016). Cells thus appear to modulate tRNA pools in response to stresses to enable expression of stress-response genes, while repressing the genes adapted to the typical pool of charged tRNAs.

Interestingly, deviations in the amounts of charged tRNAs are associated with disease. Alteration of neuron-specific tRNA levels due to defects in tRNA synthetases (Jordanova et al., 2006; Lee et al., 2006; Vester et al., 2013), pre-tRNA splicing (Karaca et al., 2014), or loss-of-function mutations in tRNA genes (Ishimura et al., 2014), drive the production and accumulation of misfolded proteins. Due to the neuron-specific expression of several of these tRNA biogenesis factors, this mistranslation causes a range of neurological disorders including intellectual disabilities and neurodegeneration. In fact, it has been proposed that codon usage is shaped by selection against this mistranslation-induced protein misfolding (Drummond and Wilke, 2008).

In addition, certain tRNAs also appear to be upregulated in human cancer cells (Gingold et al., 2014). Two rare codon tRNAs were identified to be upregulated during human breast cancer cell proliferation and were found to enhance the stability and translation of transcripts enriched in their cognate codons, driving cancer progression (Goodarzi et al., 2016). Interestingly, cancerous cells also adapt their codon usage to these misregulated tRNA pools over time (Son et al., 2017). Overall, understanding the effects of this mirsregulated tRNA-codon relationship on various cellular processes is important in expanding our understanding of the progression of multiple diseases.

1.1.7 Eukaryotic mRNA decay – proteins and pathway

Another key way cells regulate protein abundance is by tightly controlling the levels of translationally available mRNA. mRNA abundance is a function of both synthesis and decay, and while transcription has historically been well studied as a modulator of differential gene expression, our understanding of mRNA decay pathways in this context has only emerged more recently. A host of decay pathways function to degrade both normal and aberrant transcripts. Nuclear surveillance pathways retain and degrade erroneous mRNA, however this thesis will focus on cytoplasmic decay processes.

For mRNA decay to occur, an exonuclease must first gain access to the body of mature transcripts that are normally protected by their 5′ cap and 3′ poly(A) tail. Deadenylation is carried out by deadenylase complexes (Pan2/Pan3 complex, the CCR4-NOT complex, and PARN), and occurs at variable rates specific to each mRNA molecule (Cao and Parker, 2001). Removal of the 5′ cap follows, wherein a host of decapping activators promote the activity of decapping complexes such as Dcp1-Dcp2. Once the cap is removed, the body of the transcript is degraded 5′→3′ by the exonuclease Xrn1. 3′→5′ cytoplasmic exonucleases such as the cytoplasmic exosome and Dis3L2 exist as well, but their role in general mRNA decay is less clear. Because exonucleases merely require a 5′-monophosphate (in the case of Xrn1) or a 3′-OH (in the case of the exosome and Dis3L2), endonucleolytic cleavage also provides access to the transcript body and initiates rapid clearance (Coller and Parker, 2004; Ghosh and Jacobson, 2010).

Several cytoplasmic surveillance pathways have evolved to identify and degrade defective mRNA molecules. These pathways function to prevent the cell from expending energy translating aberrant mRNAs and minimize the expression of non-functional proteins. Each of these quality control mechanisms use translation to identify defective transcripts. For instance, mRNAs with premature translation termination codons are identified during the first round of translation and degraded by non-sense mediated mRNA decay (NMD) machinery (Schoenberg and Maquat, 2012), mRNAs lacking a stop codon entirely are targeted when ribosomes read through the poly(A) tail, and mRNAs with stalled ribosomes are directed for endonucleolytic cleavage and decay by the no-go decay pathway (Doma and Parker, 2007; Isken and Maquat, 2007).

1.1.8 Regulatory factors determining mRNA stability

mRNAs from different genes display stabilities that range across orders of magnitude (Keene, 2007; Keene and Tenenbaum, 2002; Morris et al., 2010). Regulatory machineries must exist to identify transcript-specific features and tightly control their decay. As with other post-transcriptional regulatory mechanisms, *cis* sequence elements are recognized by *trans* factors to regulate transcript stability. Regulation of stability is important for the cell to respond to dynamic biological processes (Elkon et al., 2010) and is particularly prevalent in developmental contexts (Chen and Coller, 2016). *Trans* factors are often differentially expressed in tissues, developmental time-points, and sub-cellular compartments resulting in context-specific regulation of transcripts. Interestingly, *cis* target sequences can also be context-specific, with different UTR splicing or polyadenylation isoforms expressed in different cell types. These isoforms can demonstrate variable stability, thus allowing for finer control in the abundance of the resultant protein without altering CDS sequence (Berkovits and Mayr, 2015; Mayr and Bartel, 2009; Nam, Rissland et al., 2014).

In general, the majority of *cis* regulatory sequences reside in 3′ UTRs (Geisberg et al., 2014). These *cis* elements recruit two main classes of *trans* factors: miRNAs and RBPs. miRNAs are short non-coding RNAs (ncRNAs) that bind complementary sites in the UTR and recruit the miRNA-induced silencing complex (RISC) to facilitate translational repression and destabilization (Fabian and Sonenberg, 2012). The human genome encodes over 2500 miRNAs that are predicted to target most human mRNAs, with ~700 expressed at levels sufficient to regulate gene expression (Friedman et al., 2009). RBPs are the other diverse class of regulators, representing approximately 10% of the human proteome, and generally bind mRNAs at specific RNA binding motifs or structures (Gerber et al., 2008). Well characterized RBPs include Pumilio (Etten et al., 2012), Smaug (Smibert et al., 1996; Tadros et al., 2007), and HuR (Barreau et al., 2018), each of which repress expression of specific transcripts. Also of note, mRNA secondary structure plays a key role in determining stability (Wu and Bartel, 2017). Despite a good understanding of transcript-specific regulation of stability through UTR regulatory elements, the role of the CDS in modulating global stability is under-characterized and is the focus of this thesis.

1.1.9 Techniques to measure mRNA stability

In order to dissect the molecular mechanisms underlying the regulation of mRNA stability, reliable methodologies (Pérez-Ortín et al., 2013) are required to estimate decay rates of a given transcript. Historically, these methods employ one of two main strategies. The first strategy involves the inhibition of RNA polymerase-dependent transcription with the addition of drugs (Bensaude, 2011) or the generation of conditional mutants (Herrick et al., 1990). Following transcription inhibition, the rate of reduction in mRNA levels over time is measured by Northern blot, RT-qPCR, or RNA-sequencing. While global transcription shut-off experiments have been invaluable in measuring decay rates genome-wide, inhibition for an extended period alters overall cellular physiology and can result in major side-effects (Nonet et al., 1987). This strategy can also be applied to single genes using drug-responsive regulatable promoters such as the Tet-off system, wherein the addition of doxycycline shuts off transcription from this promoter (Bellí et al., 1998). While this system doesn't impact overall cellular physiology, it is challenging to expand to genome-wide studies and requires the generation of non-endogenous chimeric genes.

The second strategy to measure stability relies on *in vivo* labelling of mRNAs with modified nucleotide precursors, followed by pulse or pulse-chase assays to determine incorporation kinetics over time (Dolken et al., 2008; Lugowski et al., 2017; Munchel et al., 2011; Neymotin et al., 2014; Rabani et al., 2011; Tani et al., 2012). Historically, radioactive 32P nucleotide precursors have been used in conjunction with Northern blots to study the regulation of individual genes (Ross, 1995). More recently, this strategy has been adapted for use with modified uridine triphosphate (UTP) precursors, such as the nucleobase 4-thiouracil (4tU) or the nucleoside 4-thiouridine (4SU), with multiple methods of detecting incorporation discussed below. While pulse-chase labelling strategies have been used to estimate genome-wide kinetic parameters (Munchel et al., 2011), labeled nucleotides are subject to internal recycling (Nikolov and Dabeva, 1985) and can result in inaccurate estimates of decay rates. In contrast, approach to equilibrium strategies estimate decay kinetics by fitting the rate at which transcripts approach steady state over time (Greenberg, 1972; Lugowski et al., 2017; Neymotin et al., 2014), however require the constant supply of nucleotide precursors.

It is important to acknowledge that several strategies exist to isolate and detect 4SU-labeled transcripts from the unlabeled pool. The first strategy is a biochemical separation method involving reversible biotinylation of 4SU, and subsequent fractionation with streptavidin coated

magnetic beads (Dolken et al., 2008; Duffy et al., 2015). This method requires considerable starting material, can present issues with low signal-to-noise performance due to limited biotinylation efficiency (Duffy et al., 2015), and requires complex spike-in strategies or normalization procedures (Lugowski et al., 2018; Neymotin et al., 2014; Sun et al., 2012). Recently, fractionation-free approaches have emerged that rely on reverse-transcription dependent thymine to cytosine (T>C) conversions in a high-throughput sequencing compatible manner (Herzog et al., 2017; Schofield et al., 2018). While these alternate approaches require less starting material and are spike-in independent, measured half-lives are not yet highly comparable with previously existing methods as they detect 4SU labeling at the nucleotide as opposed to transcript-level resolution. Overall, selecting the method of choice is a key decision during the course of any study examining mRNA stability, as it often affects the absolute half-life values obtained (Lugowski et al., 2017; Rabani et al., 2011).

1.2 A role for the Coding Sequence in modulating mRNA stability

1.2.1 Initial links between mRNA translation and stability

The processes of translation and mRNA decay have long been known to be intimately linked (Bicknell and Ricci, 2017; Radhakrishnan and Green, 2016; Roy and Jacobson, 2013). Initial observations suggested that inhibition of translation with cycloheximide treatment or introduction of tRNA nucleotidyltransferase mutants resulted in an overall stabilization of mRNAs (Herrick et al., 1990; Peltz et al., 1992). Further studies in *S. cerevisiae* demonstrated that dissociation of translation initiation factors produced an increased rate of deadenylation, decapping, and destabilization of transcripts (Lagrandeur and Parker, 1999; Schwartz and Parker, 1999, 2000). *In vitro* work recapitulated this finding, with the translation initiation complex component eIF4E found to inhibit decapping rates by binding the cap and blocking access (Ramirez et al., 2002). Further, all three major mRNA surveillance pathways also depend on active translation of the defective transcripts for efficient quality control and mRNA decay (Shoemaker and Green, 2012).

Transcriptome-wide correlations further support this relationship – mRNAs with increased ribosomal density, and hence translation, tend to be more stable (Edri and Tuller, 2014; Hanson et al., 2018; Neymotin et al., 2016). Importantly however, these correlations may have arisen due

to co-evolution of translation and stability to maximize gene expression, and do not imply causality. Nevertheless, the process of translating an mRNA appears to have a considerable impact on its stability.

1.2.2 Coding sequence and mRNA stability

Specific coding sequences have been previously identified in a small number of mRNAs as regulating the stability of the transcript they are on (Lee and Gorospe, 2011; Schnall-Levin, Rissland et al., 2011). These coding region determinants (CRDs) of instability generally function through deadenylation and decapping in a translation-dependent manner. A possible function for this translation-dependent decay might be to ensure that protein production remains transient for these transcripts, and that gene expression is tightly controlled with a transcriptional response. In fact, inhibition of translation appears to selectively stabilize and upregulate predominantly early inducible inflammatory genes including cytokines and transcription factors (Coleclough et al., 1990; Yamazaki and Takeshige, 2008). Specific regions within certain gene CDSs such as that of Fos (Schiavi et al., 1994), c-Myc (Lemm and Ross, 2002), or LARP4 (Mattijssen et al., 2017) have been well characterised, but remain anecdotal examples.

Interestingly, CRDs were identified as inducing rapid decay through rare codon-mediated ribosomal pausing in both yeast (Caponigro et al., 1993; Hoekema et al., 1987) and mammalian genes (Lemm and Ross, 2002; Mattijssen et al., 2017), with codon replacement experiments demonstrating re-stabilization of transcripts. As with many advances in molecular biology, a global link between codon usage and transcript stability was first established in the *S. cerevisiae* model system (Presnyak et al., 2015). Certain codons were preferentially enriched in either stable transcripts or unstable transcripts. Enrichment was quantified by the Codon Stability Coefficient (CSC) – the correlation between the frequency of a particular codon in a given transcript, and the stability of that transcript. These CSCs are highly correlated with codon optimality. In fact, reporter genes encoded by synonymous codons were able to recapitulate the entire range of endogenous mRNA decay rates. Since this landmark study, attempts to model stability across *S. cereivisiae* genes have identified codon usage, ribosome density, and other translation related factors as major correlators with mRNA decay rates (Cheng et al., 2017; Neymotin et al., 2016).

This global link between codon usage and mRNA stability has also been extended to other organisms. Similar investigations in *E. coli* indicate that inefficiently translated transcripts are

rapidly degraded in a translation-dependent manner (Boël et al., 2016). In addition, analysis of previously published datasets demonstrate that this link between codon usage and stability is conserved in *S. pombe* as well (Harigaya and Parker, 2016). Further, two recent studies in the trypanosomatid model system, where regulation of mRNA levels is dominated by post-transcriptional mechanisms, identify codon usage bias as a key predictor of mRNA levels (Jeacock et al., 2018a; Nascimento et al., 2018). A role for this translation-dependent pathway has also been identified during metazoan development using zebrafish as a model, where codon identity has been demonstrated to be a major force regulating the clearance of maternal transcripts during the maternal to zygotic transition (Bazzini et al., 2016; Mishima and Tomari, 2016). Transcripts enriched in optimal codons were also identified as having increased steady state levels across human somatic tissues (Bazzini et al., 2016), and increased stability in mouse fibroblast cell lines (Radhakrishnan and Green, 2016), suggesting that codon-mediated regulation of transcript stability may be a highly conserved process.

Investigation into the factors and pathways underlying this mechanism have implicated the yeast ortholog of human DEAD box helicase DDX6, Dhh1, as a coupler of codon composition and mRNA stability (Radhakrishnan et al., 2016). Dhh1 was previously found to promote decapping by directly slowing ribosome movement when tethered to a transcript, and accelerating degradation when rare codons were engineered into a reporter (Sweet et al., 2012). We now know that Dhh1 preferentially binds and destabilizes non-optimal mRNAs transcriptome-wide, and that overexpression of Dhh1 results in ribosomal accumulation on non-optimal codons. While Dhh1-mediated regulation depends on translation, it importantly functions independent of quality control pathway components. Dhh1 is thus thought to sense slowed ribosomal elongation on transcripts enriched in non-optimal codons, and facilitate their deadenylation in a translation-dependent manner (Radhakrishnan et al., 2016).

Nevertheless, a global relationship between codon composition and mRNA stability has not systematically been demonstrated in the human context, and my graduate work aims to address this gap in the literature.

1.3 Developing a massively parallel reporter assay to investigate the CDS-stability relationship

1.3.1 Massively parallel reporter assays as a tool to explore novel facets of molecular biology

Traditionally, *cis* regulatory elements have been identified one at a time by studying individual reporter genes, limiting the discovery of novel elements genome wide. With the advent of high-throughput oligonucleotide synthesis and sequencing, massively parallel reporter assays (MPRAs) have been utilized to identify regulatory elements on a genome-wide scale. Generally, these assays consist of a vector containing a reporter gene such as luciferase or GFP, a promoter, and potential regulatory sequences inserted into the plasmid at some site. These plasmids are either transiently transfected or integrated into the cell's genome, and expression of reporters is measured using assays such as fluorescence-activated cell sorting (FACS). Through the use of unique barcodes or by directly sequencing the inserts, MPRAs can then identify which oligonucleotide sequences modulate reporter gene expression. Historically, MPRAs have been invaluable for identifying *cis* regulatory DNA elements such as enhancers and promoters (Inoue and Ahituv, 2016; Melnikov et al., 2014; White, 2016), and have even been applied *in vivo* (Mogno et al., 2013; Shen et al., 2016). With advances in transcriptomics, MPRAs have been extended from studies of transcriptional regulation to the domain of post-transcriptional regulation.

Given that 3′ UTRs contain *cis* regulatory elements that control mRNA translation and stability, MPRAs have predominantly focused on how this region affects the expression of reporter genes. Studies thus far have relied on the insertion of sequences into the 3′ UTRs of fluorescence reporter genes, and identified their effects on mRNA abundance, protein synthesis, and mRNA stability. These libraries range from randomly synthesized 8mers (Wissink et al., 2016), to targeted libraries based on conservation (Cottrell et al., 2018; Oikonomou et al., 2014; Zhao et al., 2014), to the random fragmentation of UTRs with *in vivo* activity (Yartseva et al., 2017). Similar insertion strategies have been applied to 5′ UTR-based MPRAs as well, and have been able to predict the relative effects of Kozak sequences, uORFs, and secondary structure on overall expression of a HIS3 reporter gene (Cuperus et al., 2017). While these strategies are useful for identifying discrete elements, some MPRAs also use 3′ UTRs in full (Wolter et al., 2015) to better understand how the UTR as a whole affects gene expression. Note however that

plasmid delivery of long sequences can be a limiting factor and can introduce considerable biases to such analysis. Overall, MPRA analyses have been useful in identifying regulatory sequences and RBP or miRNA interactions with UTRs (Cottrell et al., 2018; Vainberg Slutskin et al., 2018).

Certain MPRAs have also been applied to investigate the post-transcriptional effects of sequences within the CDS. For instance, 154 variants of GFP differing in synonymous codon usage displayed a 250-fold variance in protein level and considerable variation in mRNA degradation patterns when expressed in *E. coli* (Kudla et al., 2009). Expansion to >14,000 synthetic reporters of GFP with variable promoters, ribosome binding sites, and N-terminal codons confirmed codon-dependent changes in expression levels (Goodman et al., 2013). In *S. cerevisiae* as well, analysis of >35,000 GFP variants with the insertion of 3 adjacent randomized codons identified 17 codon pairs that were associated with low protein expression due to substantially reduced elongation rates (Gamble et al., 2016).

Interestingly, these inhibitory codon pairs were subsequently identified to correlate with faster mRNA decay (Harigaya and Parker, 2017), adding to the growing evidence in support of the link between translation and stability. This link was further solidified by Bazzini et al. (Bazzini et al., 2016), who constructed a CDS library with identical UTR sequences, and demonstrated that inhibition of translation stabilized mRNAs enriched in optimal codons in zebrafish and *Xenopus* embryos. Importantly, while the synthetic CDS sequences in this study were generated by randomly fragmenting the transcriptome into ~300-500nt lengths, and were to some degree representative of endogenous genes, they do not look at individual CDSs as a whole.

1.3.2 The Human ORFeome collection

ORFeome collections refer to libraries of vectors housing all annotated open reading frames (ORFs) of a given organism. These collections were first developed in *C. elegans* to simplify the large-scale heterologous expression of proteins, with the aim of facilitating high-throughput reverse proteomic approaches such as yeast two-hybrid screens, immunoprecipitation mass spectrometry screens, and gain of function overexpression screens (Reboul et al., 2003). Collections were subsequently expanded to the ORFs of other species including *S. cerevisiae* (Gelperin et al., 2005), *Drosophila* (Bischof et al., 2014), *Xenopus* (Grant et al., 2015), and humans (Rual et al., 2004).

The human ORFeome (hORFeome) collection initially comprised ~8000 ORFs (Rual et al., 2004), and with iterated expansions (Lamesch et al., 2007), now contains over 16,000 ORFs mapping to ~14,000 genes (Yang et al., 2011). ORFs were cloned into flexible Gateway entry vectors, allowing for the subsequent cloning into a host of expression/destination vectors (Walhout et al., 2000). Further, these clones are encoded in a lentiviral expression vector, and can hence be delivered to most cell types. The hORFeome has been applied to a wide array of proteomic screens (Jain et al., 2018; Lievens et al., 2016; Taipale et al., 2012; Zhong et al., 2016), however we propose here to adapt it to an MPRA addressing post-transcriptional regulation.

1.4 Rationale and hypothesis

The overarching goal of this thesis is to determine the extent to which coding sequence regulates mRNA stability in human cells. Given the convincing data from a host of organisms discussed above, I hypothesize that human coding sequence plays a role in regulating mRNA stability in a translation-dependent manner. A major challenge with testing this hypothesis with data from endogenous mRNAs is that the UTRs flanking CDSs convolute the identification of relationships between coding sequence and stability. To address this, I have developed an experimental strategy to directly determine the impact of human ORFs on mRNA stability genome-wide. My framework utilizes the human ORFeome (hORFeome) collection, wherein each ORF possess universally common 5′ and 3′ UTRs. hORFeome-wide stability measurements allow for both a comparison across hORF constructs to interrogate the net impact of individual CDSs on stability, as well as a comparison of hORFs to their endogenous counterparts to determine the impact of endogenous gene UTR sets on decay rates. Overall, mRNA stability data obtained from these constructs would more definitively highlight the contribution of each ORF sequence to stability genome-wide, helping to ascertain whether any features in the ORF sequence are correlated with human mRNA stability.

Chapter 2: Material and Methods

2.1 Materials

2.1.1 Cell lines and cell culture

Human HEK293T cells were obtained from Dr. Mikko Taipale's lab and cultured in Dulbecco's Modified Eagle Media (DMEM) (Lonza) supplemented with 10% fetal bovine serum (FBS) (VWR Seradigm) and 1% penicillin-streptomycin solution (BioShop). Cell lines were cultured at 37°C in a humidified incubator with 5% CO_2. Cell lines were maintained under ~80% confluence by regular trypsinization with HyClone Trypsin Protease (Fisher Scientific) and re-plating. All cells were cultured in Greiner Cellstar® cell culture dishes.

Drosophila melanogaster Schneider 2 (S2) cells (Thermo Fisher Scientific R69007) were cultured in ExpressFive SFM media (Thermo Fisher Scientific) supplemented with 10% heat-inactivated FBS (Wisent) and 20mM L-Glutamine (Life Technologies) at 28°C.

S. cerevisiae USY006 was grown in YPD liquid or plates at 30°C. Cultures were obtained from Dr. John Rubinstein's lab.

2.1.2 human ORFeome collection

The human ORFeome collection version 8.1 (ccsbBroad304) cloned into lentiviral vector pLX304 was obtained from Dr. Jason Moffat's lab. Clones were transferred as a series of 96-well overnight bacterial cultures, with ~12 96-well plates transferred each day. Equal volumes of bacterial cultures were pooled into 36 pools comprising ~576 clones each. Plasmid DNA was isolated using the GeneJET Plasmid Midiprep Kit (Thermo Scientific) as per manufacturer's instructions, yielding an average of ~70ug plasmid DNA per pool. The 36 isolated pools were further combined into 6 unique pools for downstream cell line generation.

2.2 Biological Methods

2.2.1 Generation of hORFeome lentiviral libraries and cell lines

Each of the 6 unique virus pools were packaged by lipofectamine 2000 (Life Technologies) transfection using 15µg of each Lentivirus pLX304 pool, 10.8µg psPAX2 packaging vector (obtained from Dr. Mikko Taipale), and 1.8µg pVSV-G envelope vector (obtained from Dr. Mikko Taipale), according to manufacturer's instructions. Cells were cultured with transfection media for ~8 hours. Media was then removed and switched to 15mL harvest media (DMEM + 10% FBS + 1.1g/100mL BSA (7.5% solution, Life Technologies)). Cells were left for 2 days to complete virus production. Media was then collected from the plate and filtered through a 0.45µm filter (Acrodisc) by syringe. Harvested viruses were aliquoted.

Freshly thawed HEK293T cells were grown in 10cm dishes to reach ~30-50% confluence for the day of infection. Media was removed and 9mL pre-warmed infection media (DMEM + 10% FBS + 8µg/mL Polybrene) was added to cells. 2mL of each of the 6 pools of freshly harvested virus were added to 1 plate of HEK293T cells each and incubated overnight. Cells were then trypsinized and expanded into 15cm dishes. Successfully infected cells were selected using selection media (DMEM + 10% FBS + 6µg/mL Blasticidin (BioShop)) for 6 days. Selection media was changed every day. Cells were then frozen in cell freezing medium (Sigma-Aldrich) and stored in liquid nitrogen.

2.2.2 PCR validation of hORFeome cell lines

Samples were harvested from hORFeome infected cell lines by trypsinising cells and spinning at 1,000G for 2 minutes at 4°C. Cell pellets were washed with cold PBS and resuspended in 300uL QuickExtract^{TM} solution (EpicentreBio). DNA was extracted according to the manufacturer's instructions. Extracted DNA was used as the template for PCR amplification. For 3′ UTR amplification, AN009 (GGCGCGTTAAGTCGACAATC) and AN010 (CCACATAGCGTAAAAGGAGCAAC) were used; for CDS amplification, AN213 (CGCAAATGGGCGGTAGGCGTG) and AN010 were used. PCR products were run on a ~1.5% agarose gel (FroggaBio) and visualized with SafeView dye (Applied Biological Materials).

2.2.3 Immunoblotting

~1 million cells were harvested by trypsinizaiton and pelleted by centrifugation at 1,000G for 2 minutes at 4°C. Cell pellets were resuspended in 500µL Lysis Buffer A (100 mM KCl, 0.1 mM EDTA, 20 mM HEPES-KOH pH 7.6, 0.4% NP-40, 10% glycerol, 1 mM DTT, complete mini EDTA-free protease inhibitors (Roche)) and clarified at 21,000G for 5 minutes at 4°C. 250µL supernatant was mixed with 20µL 4x Bolt LDS sample buffer (Invitrogen), 8µL 10x Bolt sample reducing agent (Invitrogen) and proteins were denatured at 75°C for 10 minutes. Protein samples were loaded into Bolt 4-12% Bis-TRIS Plus gels (Invitrogen) and run at 160V for ~1 hour. The gel was transferred onto an Amersham Hybond PVDF membrane (GE Healthcare) at 20V for ~1 hour and blocked in PBST (1XPBS with 1% Tween-20 (Sigma-Aldrich)) with 5% milk (BioBasic) for 30 minutes. Primary antibodies were added at 1: 10,000 concentration for α-V5 antibodies (Sigma-Aldrich V8012), 1: 10,000 for α-Puromycin antibodies (Kerafast 3RH11), and 1: 5,000 for α-tubulin antibodies (Sigma-Aldrich T5168). Blots were incubated shaking in primary antibody overnight at 4°C.

Blots were then washed 3x with PBST for 5 minutes each and incubated with 1: 10,000 concentration α-mouse secondary antibody (NEB 7076) for 1 hour at room temperature. Blots were washed with PBST 3x for 5 minutes each. Blots were imaged using ECL Prime Western Blotting Detection Reagent (GE Healthcare) and exposed on Amersham Hyperfilm (GE Healthcare). Quantification of western blots was performed using the ImageJ Fiji distribution (Schindelin et al., 2012).

2.2.4 Polysome fractionation

hORF cell line 1 was grown for 24 hours in the presence of either DMSO or 100µM 4EGI-1 (Cedarlane). Cells were treated with 100µg/mL cycloheximide (CHX) (BioShop) to arrest translating ribosomes and incubated at 37°C for 10 minutes. Cells were harvested on ice by washing 2x with ice-cold PBS containing 100 µg/mL CHX, and lysing with 500µL ice-cold filter-sterilized lysis buffer (10 mM Tris-HCl (pH 7.4), 5 mM MgCl2, 100 mM KCl, 1% Triton X-100, 2 mM DTT, 500 U/ml RNasin (Promega), 100 µg/ml CHX, Protease inhibitor (1X complete, EDTA-free, Roche)). Cells were scraped off the dish into tubes and sheared gently 4x with a 26-guage needle. Lysed cells were centrifuged at 1,300G for 10 minutes at 4°C and clarified supernatant was isolated.

A 10/50% sucrose gradient was created by combining heavy and light solutions on a BioComp Gradient Master™. Heavy and light solutions consisted of 20 mM HEPES-KOH (pH 7.4), 5 mM MgCl2, 100 mM KCl, 2 mM DTT, 100 µg/ml CHX, and 20 U/ml SUPERaseIn, and either 10% or 50% sucrose (w/v) respectively. 300µL of samples were layered on sucrose gradients and centrifuged in a pre-cooled Beckman Ultracentrifuge L-90K using SW41 rotor at 36,000 RPM (221632.5G) for 2 hours at 4°C.

The gradient was fractionated using the BioComp Piston Gradient Fractionator™ and absorbance measurements were made using an Econo EM-1 UV Monitor (BioRad). Polysome to monosome ratios were determined by calculating areas under the corresponding peaks using custom R scripts in RStudio version 3.3.1.

2.2.5 Puromycin incorporation assay

hORF cell line 1 was grown for 1, 8, or 24 hours in the presence of either DMSO, 100µM 4EGI-1, or 5µg/mL cycloheximide. Cells were pulsed with 1.5µg/mL Puromycin Dihydrochloride (Gibco) for 10 minutes at 37ºC. Cells were then harvested and lysed as per 2.2.3 above and probed by western blot with α-Puromycin antibody (Kerafast 3RH11) to detect overall incorporation.

2.2.6 Generation of spike-in RNA

Two sets of spike-in RNA were generated. An unlabeled *S. cereivisae* spike-in is used to determine the enrichment of 4SU-labeled RNA over unlabeled RNA. *S. cereivisae* strain USY006 was grown in YPD liquid culture at 30°C, and RNA was isolated using hot acidic phenol method (Rissland and Norbury, 2009). Yeast RNA was obtained from Andrew Lugowski. A 4SU labeled *D. melanogaster* spike-in was also generated by supplementing S2 culture media with 100µM 4SU for 24 hours prior to harvesting (although note that for this thesis, reads mapping to these spike-ins were disregarded for analysis). RNA was extracted using TRI-reagent (Molecular Research Center) as per manufacturer's instructions.

2.2.7 Metabolic labeling of hORFeome cell lines

Freshly thawed HEK293T hORFeome cell lines were cultured for 3-4 passages and seeded into DMEM + FBS culture media in 15cm dishes such that they attained ~50% confluence on the day of the time course. Media was replaced with DMEM + 10% FBS + 100µM 4SU (Sigma-Aldrich)

reconstituted in DMSO. Cells were harvested at 1, 2, 4, 8, 12, and 24 hours after addition of 4SU. Harvesting was performed by dislodging cells off the plate during two washes with cold PBS followed by spinning at 1,000G for 5 minutes at 4°C. Cell pellets were resuspended in 1mL TRI-Reagent (Molecular Research Center) and extracted according to manufacturer instructions.

For translation inhibition experiments, hORF cell line 1 or cell line 4 cells growing in 10mL DMEM + FBS in 10cm dishes were pre-treated with either 0.1% DMSO or 100µM 4EGI-1 (Cedarlane) dissolved in DMSO for 1 hour. Following this, 100µM 4SU was added to media for all plates and the time course was performed as described above.

2.2.8 Reversible biotinylation and fractionation of 4SU-labeled mRNAs

Following the extraction of 4SU labeled RNA from hORF cell lines, 100µg of total hORF RNA was mixed with 10µg unlabeled *S. cerevisiae* RNA (i.e., 10% w/w) and 10µg 4SU labeled S2 *D. melanogaster* RNA (i.e., 10% w/w). Water was added to bring the volume up to 120µL. 1mg/mL HPDP-biotin (Thermo Fisher Scientific) was reconstituted in dimethylformamide by shaking at 37°C for 30 minutes at 300RPM. 120µL of 2.5x Citrate buffer (25 mM citrate, pH 4.5, 2.5 mL EDTA) and 60µL of 1mg/mL HPDP-biotin were added to the pre-mixed RNA sample for each time point. This solution was incubated at 37°C for 2 hours at 300RPM on an Eppendorf ThermoMixer F1.5 in the dark to conjugate biotin to 4SU. Samples were then extracted twice with two times acid phenol, pH 4.5 (Invitrogen), and once with chloroform. RNA was precipitated with 18µL 5M NaCl, 750µL 100% ethanol, and 2µL 15mg/mL GlycoBlue co-precipitant (Invitrogen) overnight at -20°C. Precipitated RNA was pelleted for 30 minutes at 21,000G at 4°C. The RNA pellet was resuspended in 200µL of 1x wash buffer (10 mM Tris-Cl, pH 7.4, 50 mM NaCl, 1 mM EDTA).

Biotinylated RNA was then purified using the µMACS Streptavidin microbeads system (Miltenyi Biotec). 50µL Miltenyi beads per sample were pre-blocked with 48µL 1x wash buffer and 2µL yeast tRNA (Invitrogen), rotating for 20 minutes at room temperature. µMACS microcolumns were washed 1x with 100µL nucleic acid equilibration buffer (Miltenyi Biotec), followed by 5x washes with 100µL 1x wash buffer. Beads were applied to microcolumns in 100µL aliquots, and again washed 5x with 100µL 1x wash buffer. Beads were demagnetized and eluted off the column with 2x 100µL 1x wash buffer, and columns were placed back on the

magnetic stand. 200µL beads were mixed with each sample of biotinylated RNA and rotated at room temperature for 20 minutes.

Samples were then applied to their respective microcolumns in 100µL aliquots, washed 3x with 400µL wash A buffer (10mM Tris-Cl, pH 7.4, 6M urea, 10mM EDTA) pre-warmed to 65°C, and then washed 3x with 400µL wash B buffer (10mM Tris-Cl, pH7.4, 1M NaCl, 10mM EDTA). RNA was eluted with 5x 100µL of 1x wash buffer supplemented with 0.1M DTT, and flow through was collected in a tube. Purified RNA was precipitated with 30µL 5M NaCl, 2µL Glycoblue, and 1mL 100% ethanol, incubated at –20°C overnight. Samples were then spun at 21,000G at 4°C for 30 minutes and resuspended in 20µL water. RNA quality was assessed by running 3µL of samples on a ~1.5% agarose gel. 10µL of fractionated RNA samples were then used to generate sequencing libraries.

2.2.9 Generation of next generation sequencing libraries and RNA-sequencing

10µL of purified 4SU labeled RNA or unpurified total RNA from the 24-hour time point was used to prepare sequencing libraries using the TruSeq Stranded mRNA Sample Preparation Kit (Illumina), according to manufacturer's instructions. Adapter-ligated fragments were enriched with 14x PCR cycles. ~16-22 samples were multiplexed on a single lane in an Illumina HiSeq 2500 at The Centre for Applied Genomics (TCAG, SickKids) to obtain ~10 million 50bp single-end reads per sample.

2.3 Bioinformatic processing and statistical analysis

2.3.1 Obtaining and combining reference genomes

Human (hg38, http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.2bit), *D. melanogaster* (dm6, http://hgdownload.soe.ucsc.edu/goldenPath/dm6/bigZips/dm6.2bit), and *S. cerevisiae* (sacCer3, http://hgdownload.soe.ucsc.edu/goldenPath/sacCer3/bigZips/sacCer3.2bit) genomes were obtained in 2bit format using the UCSC Table Browser (Kent, 2004). 2bit files were converted to FASTA using the kentUtils command twoBitToFa, and GTF annotations were downloaded using the kentUtils command genePredToGtf. The three genomes were combined using custom bash scripts to make a hg38+dm6+sacCer3 genome.

In addition, an intron-specific GTF file was generated from the human GTF annotations using the published process_GTF.R script (https://github.com/risslandlab/DRUID) (Lugowski et al., 2017). Briefly, introns are taken from the longest transcript of each gene and removed if they overlap with any annotated exons or an intron from another gene on the same strand.

2.3.2 Initial processing of sequencing reads

Library quality was assessed using FastQC v0.11.5 (Andrews S. (2010), http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Reads were trimmed and clipped for Illumina adapters using Trimmomatic v0.36 (Bolger et al., 2014) using the following settings: -phred33 ILLUMINACLIP: TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36.

2.3.3 Genome mapping and counting

Trimmed reads were aligned to the indexed hg38+dm6+sacCer3 genome using STAR version 2.5.2 (Dobin et al., 2013) with the following non-default settings: --outFilterMultimapNmax 10 --outFilterMismatchNoverLmax 0.05 --outFilterScoreMinOverLread 0.75 --outFilterMatchNminOverLread 0.85 --alignIntronMax 1 --outFilterIntronMotifs RemoveNoncanonical --outSAMtype BAM SortedByCoordinate --quantMode GeneCounts.

HTSeq version 0.6.1 (Anders et al., 2015) was then used to quantify gene counts from aligned BAM files using the following settings: --order=pos --stranded=reverse --minaqual=10 --mode=intersection-strict. Note that counting of features from human CDS and intronic GTF files were performed separately.

Computations for initial processing, genome mapping, and counting were performed on the gpc supercomputer at the SciNet HPC Consortium. SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto.

2.3.4 Defining hORF genes

Gene counts were loaded into RStudio version 3.3.1. Dplyr package (Wickham et al., 2017) was used for all data manipulation and filtering. Steady state RNA-sequencing counts mapping to human CDS features were obtained from each cell line and normalized to library size to allow

comparisons across samples. For a given cell line X, genes were described as “detectable hORF genes” if they met each of the following conditions:

1. They were in the list of hORFs infected into cell line X;
2. Normalized steady state RNA-sequencing for cell line X was greater than 3-fold that in its paired cell line;
3. Normalized steady state RNA-sequencing for cell line X was greater than 4 reads

Genes that were not infected into cell line X were described as endogenous genes.

2.3.5 Calculation of mRNA half-lives

All half-life calculations were performed in RStudio version 3.3.1. Introns were first filtered for those that exhibit unstable dynamics as described in DRUID (Lugowski et al., 2017). Briefly, k-means clustering was used with 4 defined clusters, and the cluster exhibiting expected decay behavior was manually selected. Read counts for mature human mRNAs were then filtered such that each gene had at least 1 read mapped to its CDS at each time point, and at least 5 reads mapped at any (at least 1 of 6) timepoint. CDS-mapping reads for each gene at each timepoint were then normalized to the sum of all corresponding filtered intron-mapping reads.

Half-lives were calculated by fitting these normalized read counts at each timepoint to a bounded growth equation using weighted nonlinear least squares. The bounded growth equation has been previously described (Lugowski et al., 2017). Briefly, the equation states:

$$y(t) = y_{eq} \times (1 - e^{kt})$$

where y(t) is the amount of a given transcript remaining at time t, y_{eq} is the amount of that transcript at steady state, and k is the transcript-specific decay constant. Note that because we are normalizing to internally produced spike-ins, we do not control for the dilution due to growth.

The nls() function in the stats package was used to fit the timepoints to the equation above, with settings equivalent to the following:

- start = c(y_{eq} = max(y), k = -0.5)
- algorithm = “port”
- weights = 1/y(t)
- lower = c(y_{eq} = 0, k = -Inf), upper = c(y_{eq} = Inf, k = 0)

If the data did not converge, a value of NA was returned.

Once the equation is fit, the half-life of each transcript is then obtained using the following equation:

$$HL = \frac{\ln(2)}{k}$$

For half-lives derived from other studies, published half-lives were directly used. Human HEK293 half-life data was obtained from GEO accession number GSE99517 (Lugowski et al., 2018). Yeast half-life data was obtained from Presnyak et al., 2015.

2.3.6 Calculation of CSCs, AASCs, and AA stretches

Codon usage frequency was calculated from each gene's coding sequence using the seqinr package's uco function (Charif, D. and Lobry, J.R., 2007). Amino acid usage was calculated using Biostrings package's (Pagès H et al., 2018) translate function to convert coding sequence into AA sequence, and then using Biostrings alphabetFrequency function to count AAs per CDS. AA usage frequency was determined by dividing AA count by CDS length. Note that for frame shift controls, codon usage and AA usage were calculated after shifting the frame by +1 (removing positions 1, n-2, and n-1 from CDS of length n) and +2 (removing positions 1, 2, and n-1 from CDS of length n). AA stretches were calculated using stringr package's str_count function with various patterns listed in standard regular expression notations.

CSCs from a given half-life dataset were calculated by determining the Spearman correlation between the codon frequency for each codon in a transcript with the measured half-lives of that transcript. Stop codons were excluded from CSC calculations. AASCs were similarly calculated by determining the Spearman correlation between the AA frequency for each AA in a transcript with the measured half-lives of that transcript.

2.3.7 Calculation of local secondary structure

To measure local secondary structure, each gene's coding sequence was assayed by sliding 100bp windows, each starting 3bp apart. Each window was folded using ViennaRNA version 2.2.8 RNAfold function (Lorenz et al., 2011) using default parameters. Computations were performed on the supercomputer at the SickKids Centre for Computational Medicine High Performance Computing Facility cluster. Minimum folding energy (MFE) for each 100bp sequence was extracted from output files using custom bash scripts. Median and minimum MFEs

across each CDS were determined using group_by and summarise functions in the dplyr package in RStudio.

2.3.8 Miscellaneous statistical analysis

Calculations comparing statistics between endogenous genes or hORF constructs and a given variable were performed by randomly sampling 500 genes from each half-life dataset 1000 times, determining values for each statistic within each subsample, and taking the mean of these statistics across each subsample.

For variance comparisons between endogenous and hORF genes, 200 half-lives were randomly sampled from each dataset 1000 times, and the variance across each subsample was calculated. The median value of this variance was calculated, and ratios of variance between the two sets of genes are presented.

Several additional R packages were used for data manipulation and figure generation (Charif and Lobry, 2007; Harrell and Dupont, 2016; Pagès et al., 2017; S, 2010; Team, 2013; Wickham; Wickham et al., 2017).

Chapter 3: Results Part I: Coding sequence regulates mRNA stability

3.1 Generating hORFeome expressing lines and detecting hORF construct abundance

The overarching goal of my thesis is to determine the extent to which coding sequence (CDS) regulates mRNA stability in human cells by systematically measuring the mRNA decay rates of hORF constructs. I first generated cell lines expressing the hORFeome collection. To do this, I obtained the V5-tagged hORFeome collection version 8.1 cloned into lentiviral expression vectors, representing ~16,000 human ORFs (Yang et al., 2011). These ORFs are under the transcriptional control of a constitutively active CMV promoter, harbour a stabilising Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE) in the 3′ UTR, and include a Blasticidin marker to select for successfully infected cells.

To ensure that I maintained library complexity, I divided the collection into 6 unique pools comprising ~3,000 clones each. I then packaged these hORFs into lentiviral particles, and infected human embryonic kidney (HEK293T) cell lines at a titre of ~1 unique viral particle infecting each individual cell (Fig3).

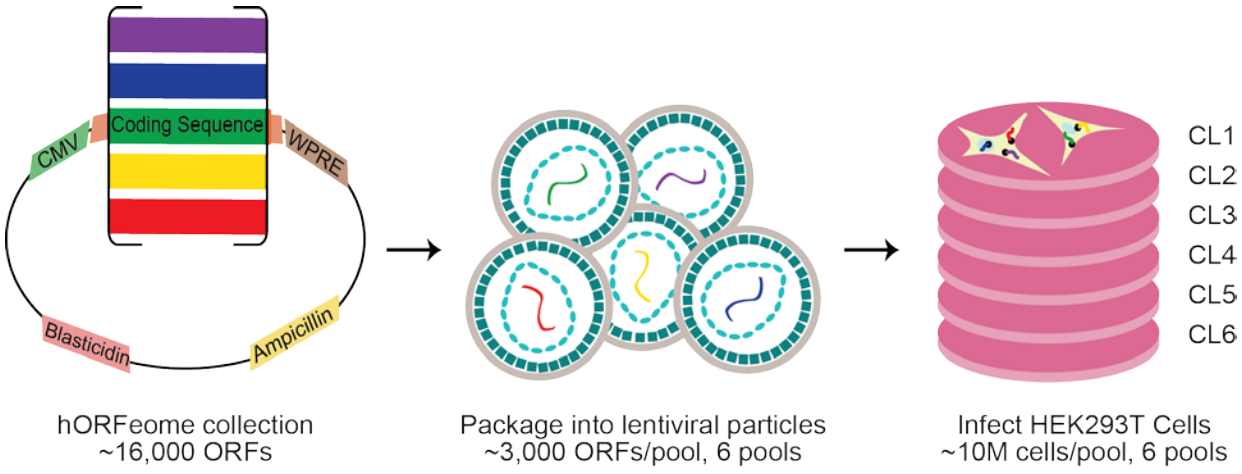

Figure 3. Experimental workflow for generating human ORFeome expressing stable cell lines. Initially, the hORFeome lentiviral expression library was isolated from bacterial stocks and pooled into 6 sets comprising ~3,000 clones each. Lentiviral particles were then packaged in HEK293T cells and harvested. Viruses were used to infect and integrate hORF genes into HEK293T cells, generating 6 unique hORFeome expressing cell lines. hORF genes were driven by a constitutively active CMV promoter and harboured a WPRE element in the 3′ UTR.

To characterize these lines, I first extracted genomic DNA from infected cells, and PCR amplified hORFs using primers flanking the constructs. I observed a smear representing a large range of ORF sizes in each pooled cell line, while detecting no amplification in uninfected negative controls (Fig4A). To confirm that these ORFs were transcribed and translated into detectable proteins, I isolated protein lysates and performed western blotting against the V5 epitope tag at the C-terminus of all hORF constructs. Once again, I observed a large range of sizes in each of the 6 stable cell lines, with no detectable protein in the uninfected negative control (Fig4B), indicating that hORF proteins are being expressed. Note that some of the larger hORF constructs in the library were not detectable by western blotting and may have dropped out of the collection due to an upper size limit during viral packaging and delivery.

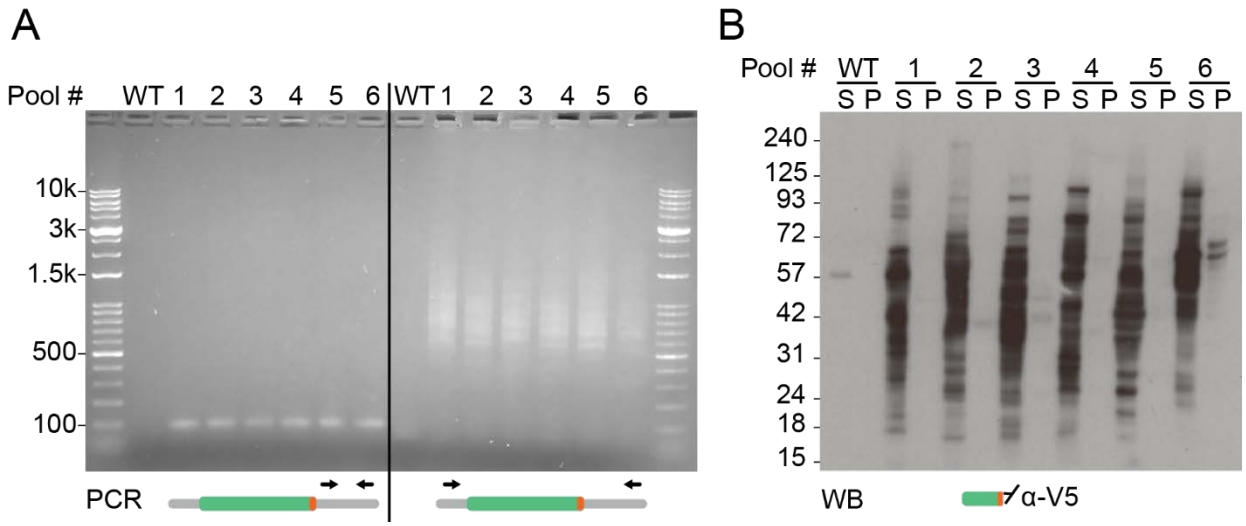

Figure 4. Confirmation of hORFeome integration and expression. (A). DNA was extracted from wild-type uninfected 293T and hORFeome cell lines, and PCR amplified using the indicated primers. Gel electrophoresis was conducted to separate amplified fragments and confirm successful integration of hORF constructs at the DNA level. Ladder DNA fragment sizes are indicated in kilo base pair units. (B). Protein lysates were prepared from wild-type uninfected 293T and hORFeome cell lines, separated into soluble (S) and precipitate (P) fractions, and electrophoretically separated on a denaturing polyacrylamide gel. Proteins were transferred to a PVDF membrane and probed with α-V5 antibodies by western blot to confirm successful expression of hORF proteins. Ladder fragment sizes are indicated in kilo Daltons.

To test the expression levels of hORF mRNAs, I performed high-throughput RNA sequencing (RNA-seq) on each hORFeome cell line. As expected, endogenous transcripts were found to have similar steady-state expression profiles between cell lines ($r_s = 0.98$, $p < 10^{-16}$) (Fig5A). In contrast, hORF constructs showed differential expression signatures, expressed at a higher abundance in the cell line into which they were infected, while remaining unexpressed in the opposing cell line (Fig5A).

The major challenge with using short-read RNA-seq methods to measure hORF transcript abundance is that sequencing reads mapping to the CDS of a particular gene cannot distinguish between a hORF construct and its endogenous counterpart. To get around this issue, I took advantage of the fact that wild-type HEK293T cells express up to one third of their genome at a very low or undetectable level, and hypothesized that reads mapping to these lowly-expressed genes were instead derived from hORF constructs. By restricting my analysis to these hORF genes with lowly expressed endogenous counterparts, I would thus be able quantify hORF transcript abundance.

To define which hORF constructs are accurately detectable in the corresponding cell line (but not the others), I devised cutoffs wherein reads were considered hORF-derived if their abundance was three-fold higher in the infected cell line compared to the corresponding cell line pair (solid black lines, Fig5A). I selected these cutoffs to maintain a false discovery rate (defined by the fraction of endogenous genes occurring above the cutoff) of ~6-17%. These hORF-derived transcripts have significantly higher expression in the expected cell lines (Fig5B). This approach thus allowed me to quantify mRNA expression of a subset of hORF constructs.

3.2 Coding sequence differentially regulates mRNA turnover

Having developed a pipeline for quantifying the abundance of a subset of hORF mRNAs, I next measured the decay kinetics for hORFeome- and endogenously-derived transcripts. To do so, I used an approach-to-equilibrium *in vivo* 4SU-labelling strategy as opposed to transcriptional shutoff assays to minimise perturbations to cellular physiology. Briefly, I added 4SU to culture media and then harvested cells at a range of timepoints post-labelling (Lugowski et al., 2017; Neymotin et al., 2014). I then enriched 4SU labeled transcripts by reversible biotinylation, streptavidin pull-down, and elution. Resultant RNA was quantified using high-throughput RNA

sequencing, and the rate at which each gene approaches equilibrium was calculated to determine decay rates and half-lives (Fig6A).

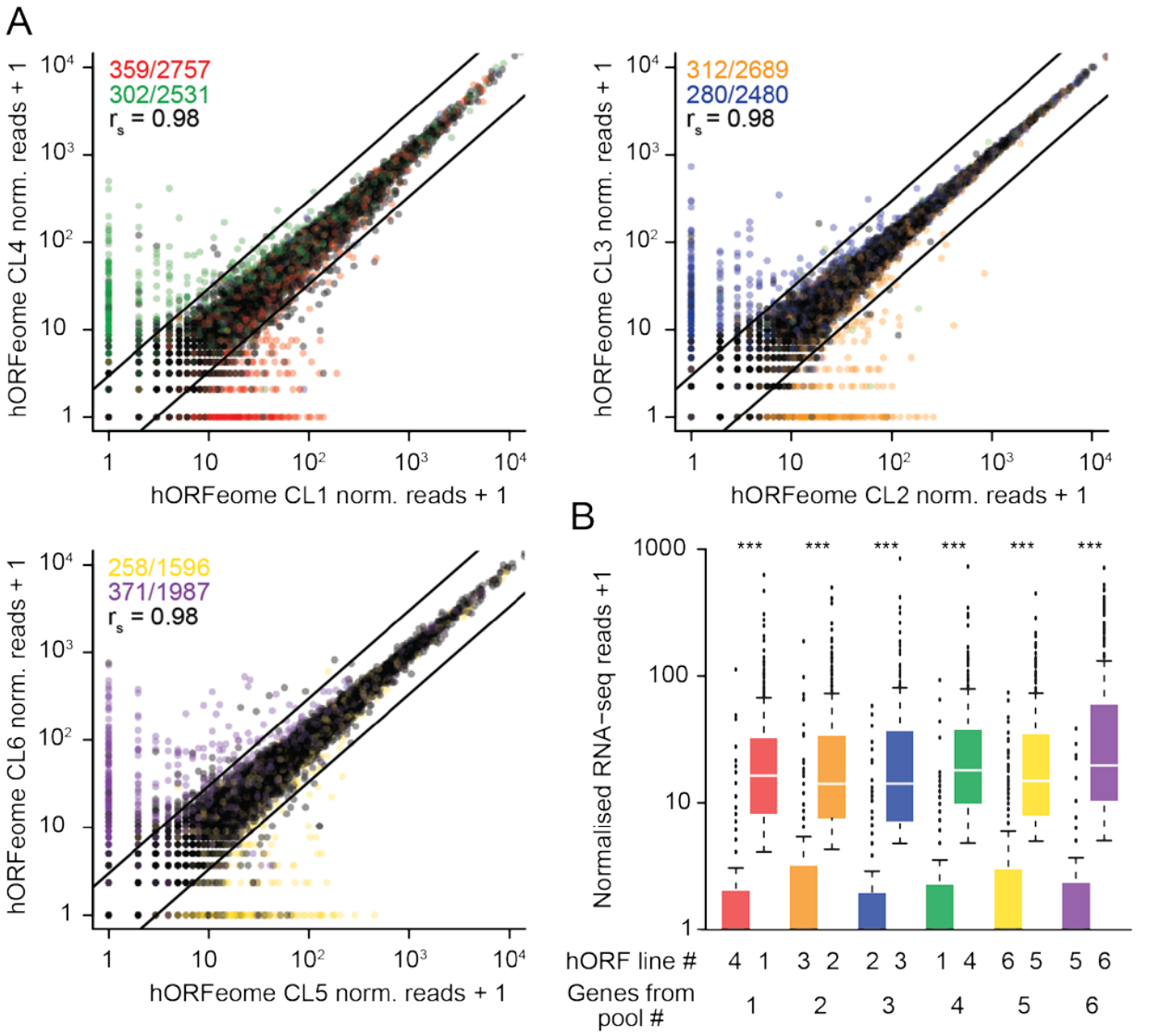

Figure 5. hORF genes are enriched in infected cell lines. (A). RNA sequencing was performed on hORFeome expressing cell lines 1-6. Steady state read counts normalised to library size are plotted for each experimental pair of cell lines, with genes in hORF pools coloured as indicated. Detectable hORF genes are defined as those with a 3-fold higher transcript abundance relative to its corresponding cell line pair, as indicated by the points outside of the solid black cutoff lines. The number of hORF genes passing the cutoff are indicated, along with total hORF constructs infected into the cell line. The Spearman correlation (r_s) between endogenous genes is listed for each cell line pair. (B). RNA sequencing read counts in the cell line of interest and its corresponding cell line pair are plotted for all "detectable" hORF genes infected into the cell line of interest. (*** $p < 2.2x10^{-16}$, Wilcoxon rank-sum test).

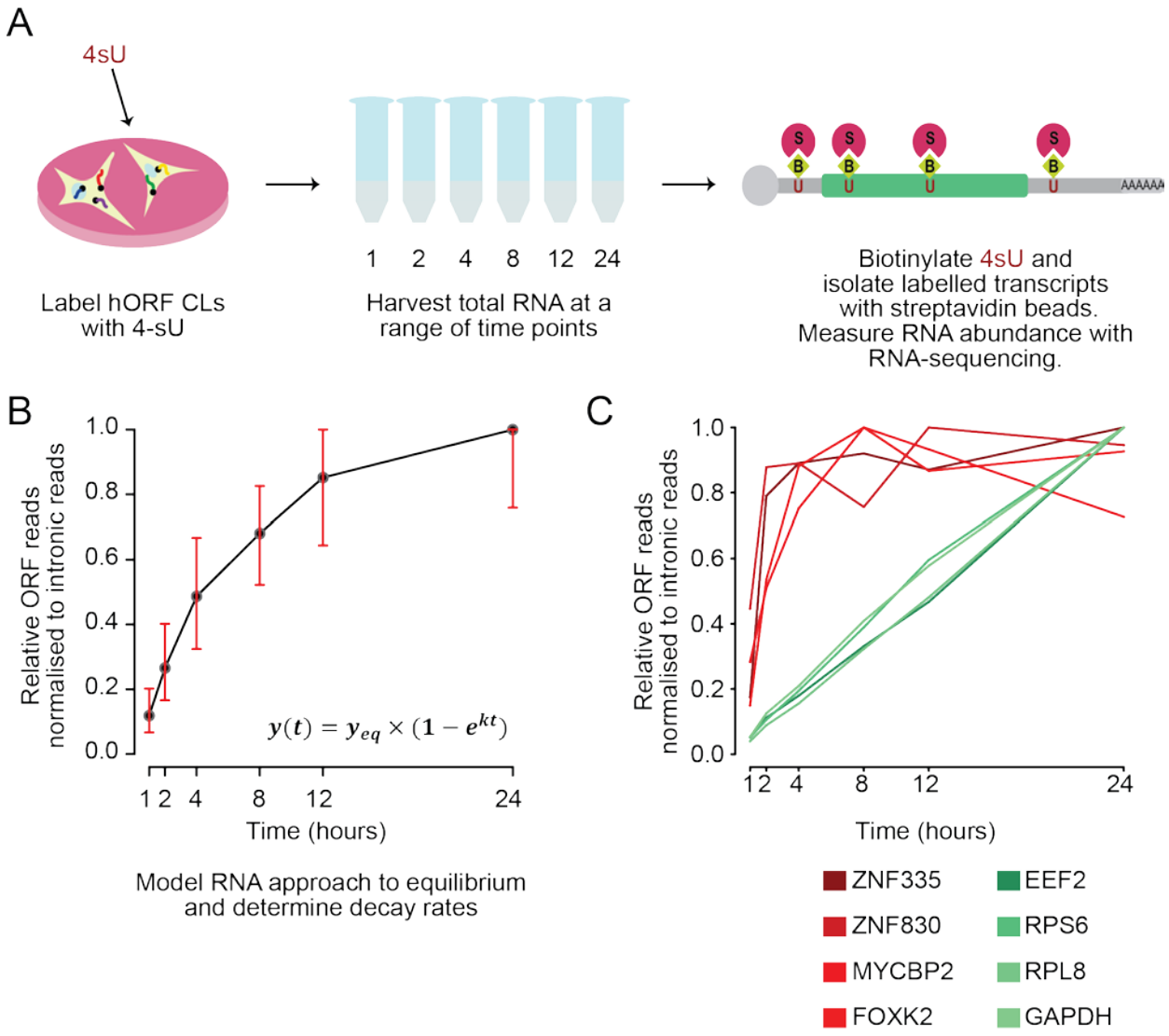

Figure 6. Approach to equilibrium-based methods for measuring RNA stability in human cells. (A). hORFeome expressing cell lines were pulsed with 4SU and RNA was extracted across a range of timepoints. 4SU labeled transcripts were conjugated to biotin, isolated with streptavidin magnetic beads, and used as templates for poly(A) enriched stranded mRNA sequencing. Resultant reads were mapped to human CDSs and normalised to intron mapping reads at each timepoint. Decay rates were determined by modelling the approach to equilibrium kinetics for each transcript. (B). ORF-mapping read counts per gene were normalised to a subset of intron-mapping reads at each timepoint. These gene counts were then normalised to the timepoint with the highest value. Genes with the lowest 20th percentile expression were filtered out. Median count values at each timepoint are plotted, along with the 25th and 75th percentile count values shown with red bars. Decay rates are modelled using the indicated equation for each individual transcript. (C). Intron-normalised ORF read counts, normalised to the timepoint with the highest value, are plotted for the indicated short-lived and long-lived transcripts at each timepoint.

To determine half-lives, I normalised CDS-mapping read counts to total intron-mapping reads using a method previously developed in the lab (called DRUID; Lugowski et al., 2017, 2018). These introns act as an internal normalization control that allow for comparison between time points so that half-lives can be determined. To focus on the quantification of hORF constructs over endogenous genes, I restricted read mapping to human CDS sequences only, as opposed to the entire human genome. Finally, decay rates and half-lives are then calculated by fitting a weighted nonlinear regression to estimate the kinetic parameters in the indicated half-life equation (Fig6B).

Overall, I observed labeled transcripts saturating over time with canonical approach-to-equilibrium kinetics (Fig6B). As expected, short-lived transcripts such as transcription factors become completely labeled in a short period of time, while longer-lived transcripts such as ribosomal protein genes reach equilibrium at a slower rate (Fig6C).

As expected, endogenous transcript half-lives were highly correlated between biological replicate cell line pairs measured in parallel (r_s = 0.842, 0.879, 0.871 respectively, $p < 10^{-16}$ for each) (Fig7A). In addition, endogenous mRNA half-lives from hORFeome HEK293T cell lines shared a strong correlation (r_s = 0.698, $p < 10^{-16}$) with published HEK293 half-lives measured using the same method (Lugowski et al., 2018) (Fig7B). Furthermore, genes involved in similar functions tend to possess similar mRNA decay rates (Keene, 2007). For instance, in my half-life datasets, mRNAs encoding transcription factors and tRNA modification factors are more unstable, while genes involved in translation or glycolysis have longer half-lives as expected (Fig7C). Taken together, these results validate the half-lives measured for endogenously expressed transcripts in hORF cell lines.

I next calculated half-lives for hORF-derived mRNAs. After restricting my analysis to the previously defined “detectable” hORFs, I was able to determine half-lives for ~10% of these constructs (Fig8A). hORF mRNAs were more stable relative to endogenous transcripts overall (two-sided Kolmogorov-Smirnov test (KS test) $p < 10^{-16}$), likely because of the stabilizing WPRE element in their 3′ UTRs (Fig8B).

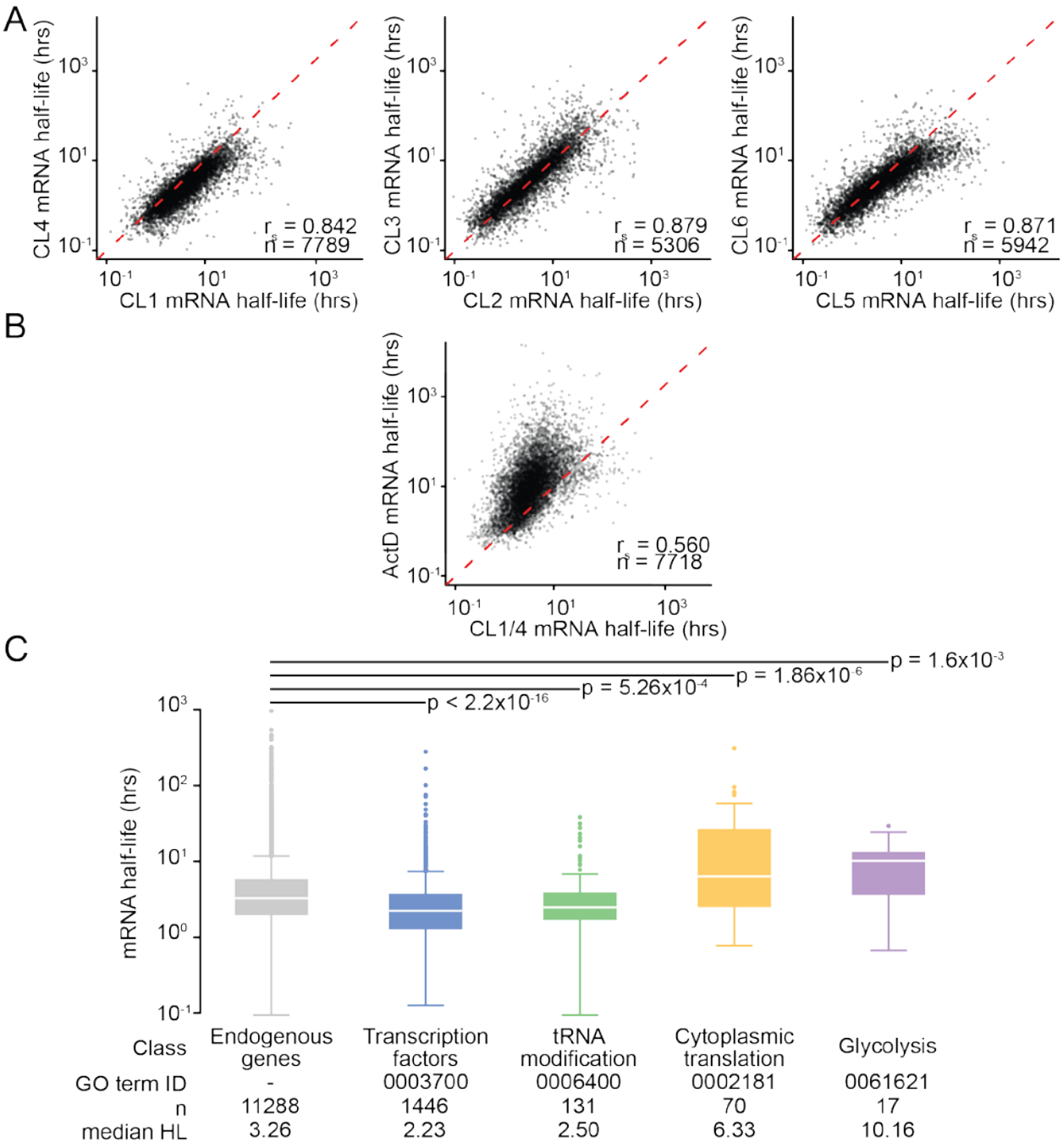

Figure 7. Endogenous transcript half-lives calculated from hORF-expressing cell lines align with published data. (A). Endogenous gene half-lives calculated from hORF cell lines using 4SU RATE-seq methods are highly correlated between replicates. Spearman correlation (r_s) and number of genes (n) with half-lives are indicated. (B). Endogenous gene half-lives calculated from 293T hORF cell lines correlate well with previously published 4SU half-life datasets. Spearman correlations (r_s) and number of genes with half-lives (n) are indicated. (C). Endogenous gene half-lives calculated from hORF cell lines were binned into functional categories known to be enriched in stable or unstable transcripts. Binning was performed using the indicated GO terms, and boxplot representations of endogenous half-lives are plotted. P-values represent the difference between the indicated group of transcripts from all endogenous transcripts (Wilcoxon rank-sum test).

Strikingly, hORF mRNAs also demonstrated considerably different rates of turnover (Fig8C). Endogenous mRNAs vary in their CDS and UTRs; however, hORFeome-derived transcripts, despite only differing in their CDS, sampled a similar range in stabilities. I next quantified the difference in variation between endogenous and hORF mRNA stability. Because of the large differences in sample size between endogenous and hORF transcripts, I calculated variance by repeatedly sampling 200 genes from each dataset 1,000 times (Fig8D) and found that hORF transcripts showed a marginal but significant increase in half-life variance (Endogenous σ^2 = 0.94, hORF σ^2= 1.36, hORF:endogenous σ^2 = 1.45, $p < 10^{-16}$). Thus, changing the coding sequence of an mRNA is able to recapitulate the entire variance in stability seen across the endogenous transcriptome, suggesting that the human CDS has a major impact on mRNA stability.

3.3 The importance of translation in coding sequence-mediated stability regulation

Having identified a link between coding sequence and mRNA turnover in human cell lines, I set out to determine the extent to which this relationship is dependent on translation, as has been investigated in other model systems (Bazzini et al., 2016; Mishima and Tomari, 2016; Radhakrishnan et al., 2016). To do so, I used 4EGI-1, a small molecule that disrupts the interaction between initiation factors eIF4E and eIF4G to block cap-dependent translation initiation (Fig9A) (Moerke et al., 2007; Sekiyama et al., 2015). To quantify the reduction in translational engagement, I first demonstrated that polysomes are reduced following a 24hr treatment with 4EGI-1 as compared to a DMSO control (Fig9B). Furthermore, I performed a puromycin labelling assay, wherein brief incubation with puromycin leads to incorporation of the antibiotic into nascent polypeptide chains. After lysis, puromycin incorporation is determined by western blotting, and puromycin signal thus directly reflects the global rate of mRNA translation (Schmidt et al., 2009). This assay, in conjunction with the polysome profile, also demonstrates a reduction in translation (although not a complete shut off) (Fig9C).

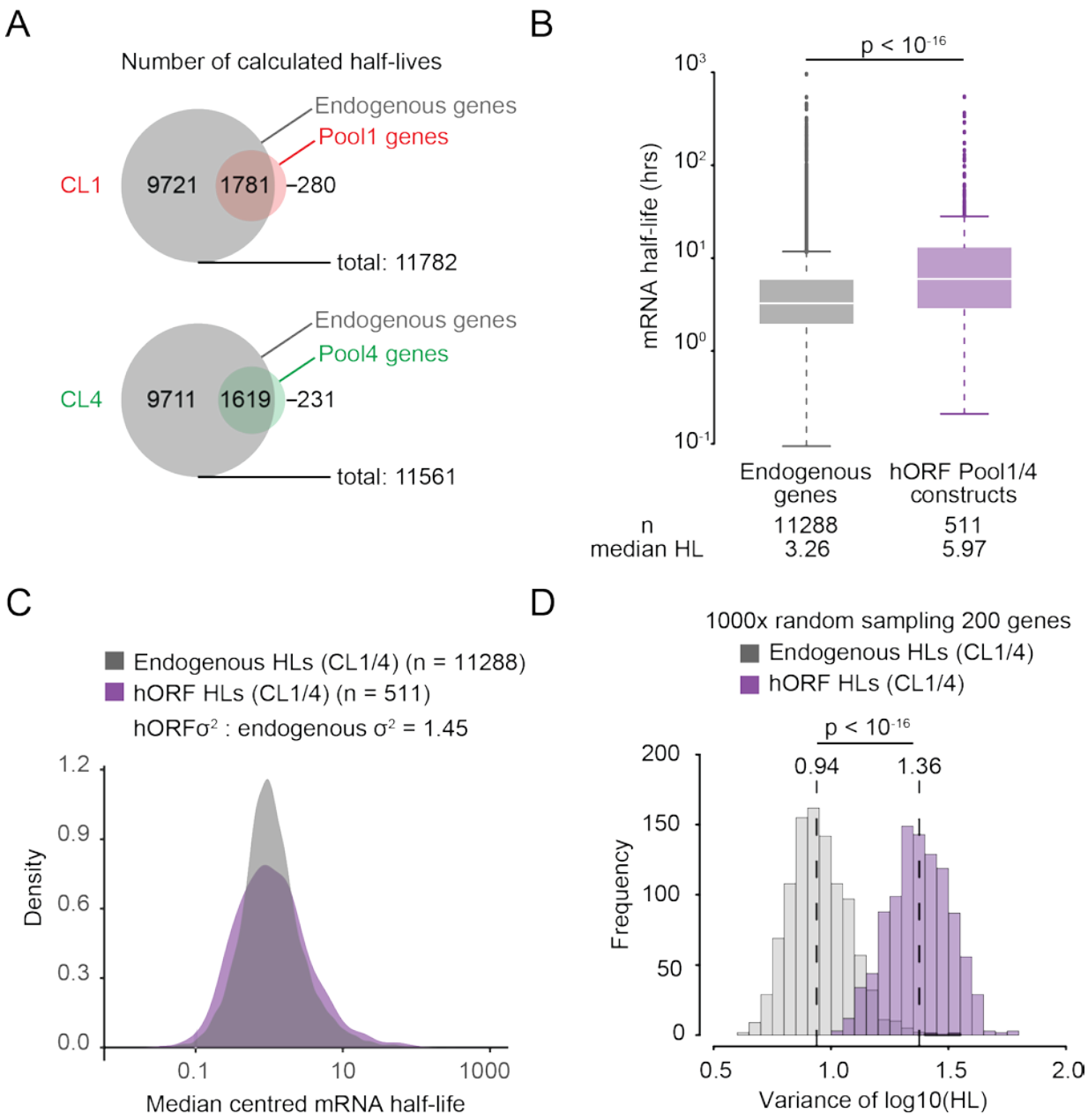

Figure 8. Coding sequence regulates mRNA stability. (A). Venn diagram representation of number of calculated half-lives. For each cell line, the number of half-lives measured for endogenous-only or hORF-only genes are indicated. (B). Endogenous genes and hORF constructs demonstrate differential stability. Half-lives for endogenous genes were individually measured from cell line 1 and cell line 4 and averaged. Half-lives for hORF constructs are only calculated for genes not endogenously expressed above the previously described cutofffs, and combined from CL1 and CL4. Boxplot representations of half-lives are plotted, with number of half-lives and median half-life indicated. P-value was determined using KS test. (C). Median-centred density plots comparing the variation in endogenous and hORF half-lives for the CL1-CL4 paired experiment. Ratios of calculated variance are indicated. (D). Variance calculations were performed by randomly sampling 200 genes from endogenous or hORF datasets 1000 times and determining the median variance, indicated by dashed lines. Difference between variance distributions was calculated using the KS test.

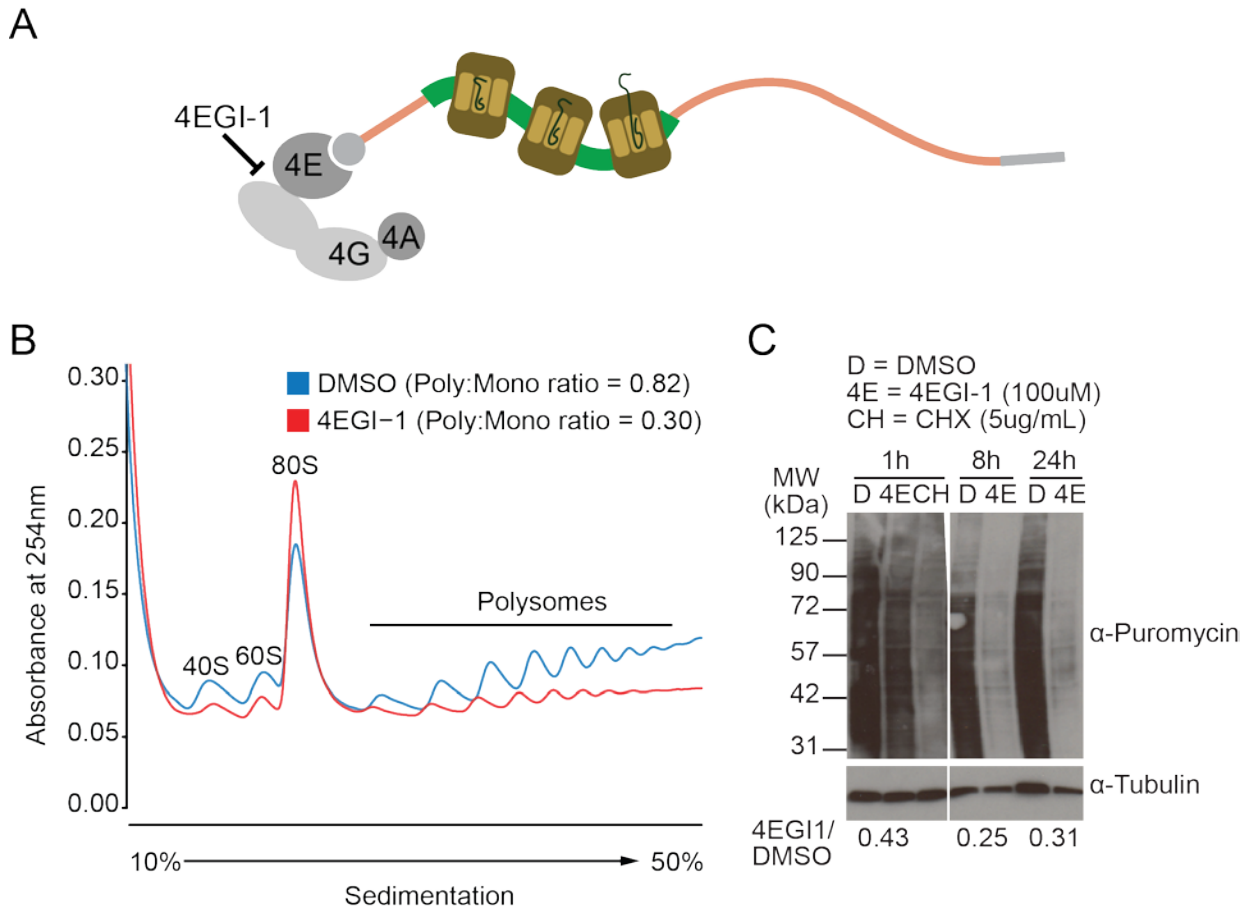

Figure 9. 4EGI-1 treatment inhibits translation. (A). Schematic representation of 4EGI-1 inhibition of cap-dependent translation. 4EGI-1 inhibits the interaction between eIF4E and eIF4G, thus preventing cap-dependent translation initiation. (B). Polysome profiles following 4EGI-1 treatment. hORF cell line 1 was gown for 24hrs in the presence of either DMSO or 100µM 4EGI-1. Cells were treated with 100µg/mL cycloheximide to arrest translating ribosomes, lysed, and subjected to 10/50% sucrose gradient centrifugation. The gradient was fractionated and absorbance detected. Polysome to monosome ratios were determined by calculating areas under the corresponding peaks. (C). Puromycin incorporation assay following 4EGI-1 treatment. hORF cell line 1 was treated with either DMSO, 100µM 4EGI-1, or 5µg/mL cycloheximide for the indicated times. Cells were then pulsed with 1.5µg/mL puromycin for 10 minutes at 37°C and lysed. Lysates were probed with antibodies targeting puromycin to detect overall incorporation. Translation in 4EGI-1 was determined relative to DMSO treatment by quantifying overall intensity.

Cell lines treated with the drug solvent DMSO displayed a relatively similar expression profile (r_s = 0.954, 0.964, $p < 10^{-16}$) when compared with previous RNA-seq measurements on untreated cells (Fig10A), highlighting that DMSO treatment alone at the concentrations used does not alter gene expression. Endogenous transcript expression was also strongly correlated between cell lines treated with either DMSO (r_s = 0.969, $p < 10^{-16}$) or 4EGI-1 (r_s = 0.971, $p < 10^{-16}$) when measurements were conducted in parallel (Fig10B). In contrast, I observed larger changes in mRNA expression profiles when comparing DMSO-treated and 4EGI-1-treated cell lines (r_s = 0.935, 0.936, $p < 10^{-16}$ each), particularly for a subset of transcripts (Fig10C). This result indicates that prolonged translation inhibition with 4EGI-1 leads to the alteration of steady state mRNA abundances.

To determine whether the variation in stability for hORFeome mRNAs depends on translation, I measured half-lives in cell lines 1 and 4 following treatment with either DMSO or 4EGI-1, as before. Like steady-state abundance, mRNA half-lives are also highly altered upon treatment with 4EGI-1 (r_s = 0.404, $p < 10^{-16}$ for endogenous genes, r_s = 0.377, $p < 10^{-11}$ for hORF constructs) (Fig11A). Further, while variance in endogenous gene half-lives is reduced ~1.4-fold (Endogenous DMSO σ^2 = 0.83, Endogenous 4EGI-1 σ^2= 0.58, 4EGI-1: DMSO σ^2 = 0.70, $p < 10^{-16}$) upon 4EGI-1 treatment, this reduction is on the order of ~4-fold (hORF DMSO σ^2 = 1.06, hORF 4EGI-1 σ^2= 0.26, 4EGI-1: DMSO σ^2 = 0.24, $p < 10^{-16}$) for hORF constructs (Fig11B). Thus, I conclude that the large range of hORF stability requires translation. Translation is also a major mediator of stability for endogenous genes, but 3′ UTR-mediated regulation is also likely a key driver of differences in stability between transcripts.

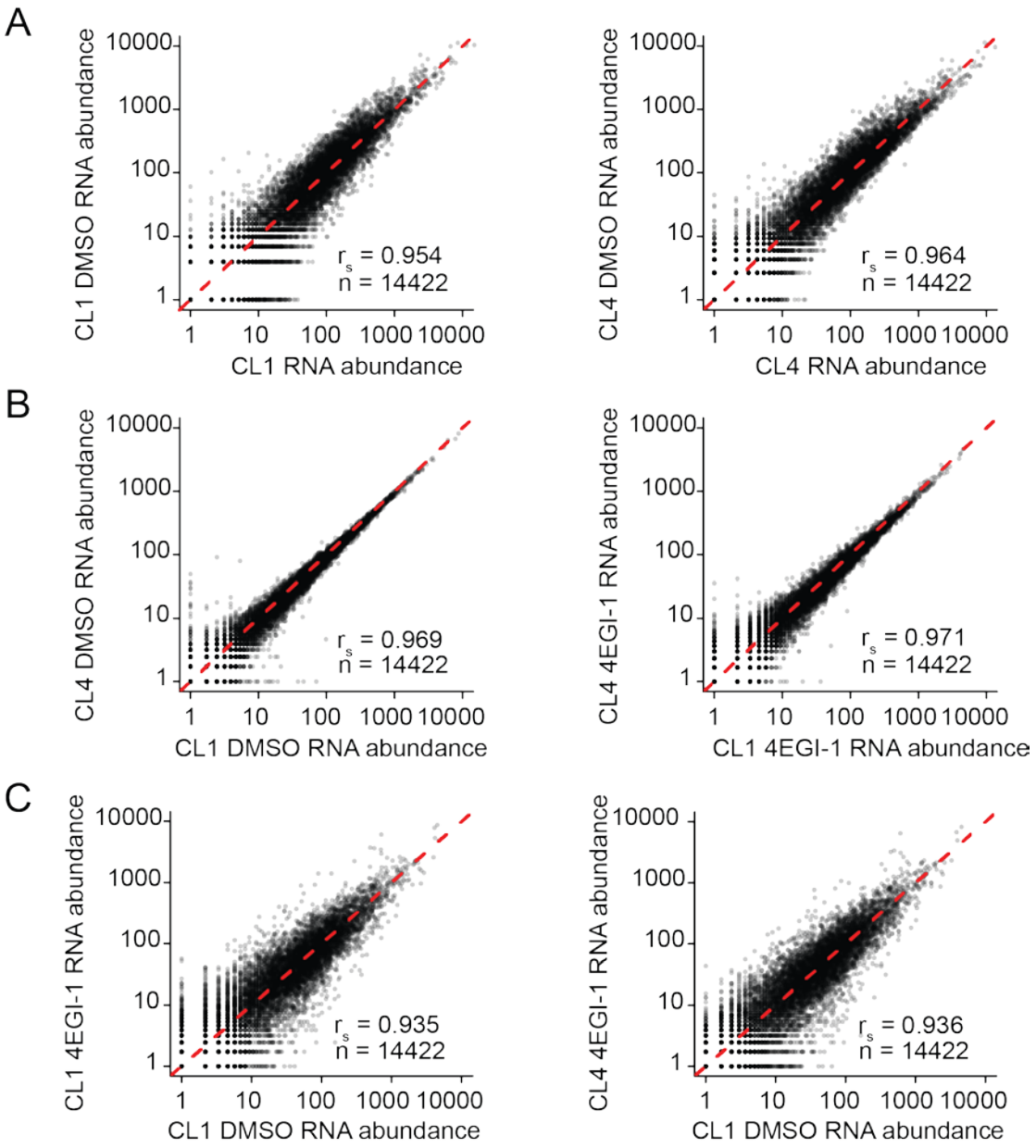

Figure 10. Translation inhibition alters steady state mRNA expression profiles. (A). Correlations of steady state RNA-sequencing expression profiles of untreated or DMSO treated cell line 1 and cell line 4. Spearman correlation (r_S) and number of endogenous genes measured (n) are indicated. (B). Correlation of steady state RNA-sequencing expression profiles of DMSO-treated or 4EGI-1-treated cells for transcripts in cell line 1 with cell line 4. (C) Correlations of DMSO-treated mRNA expression profiles with 4EGI-1-treated expression profiles in cell line 1 or cell line 4.

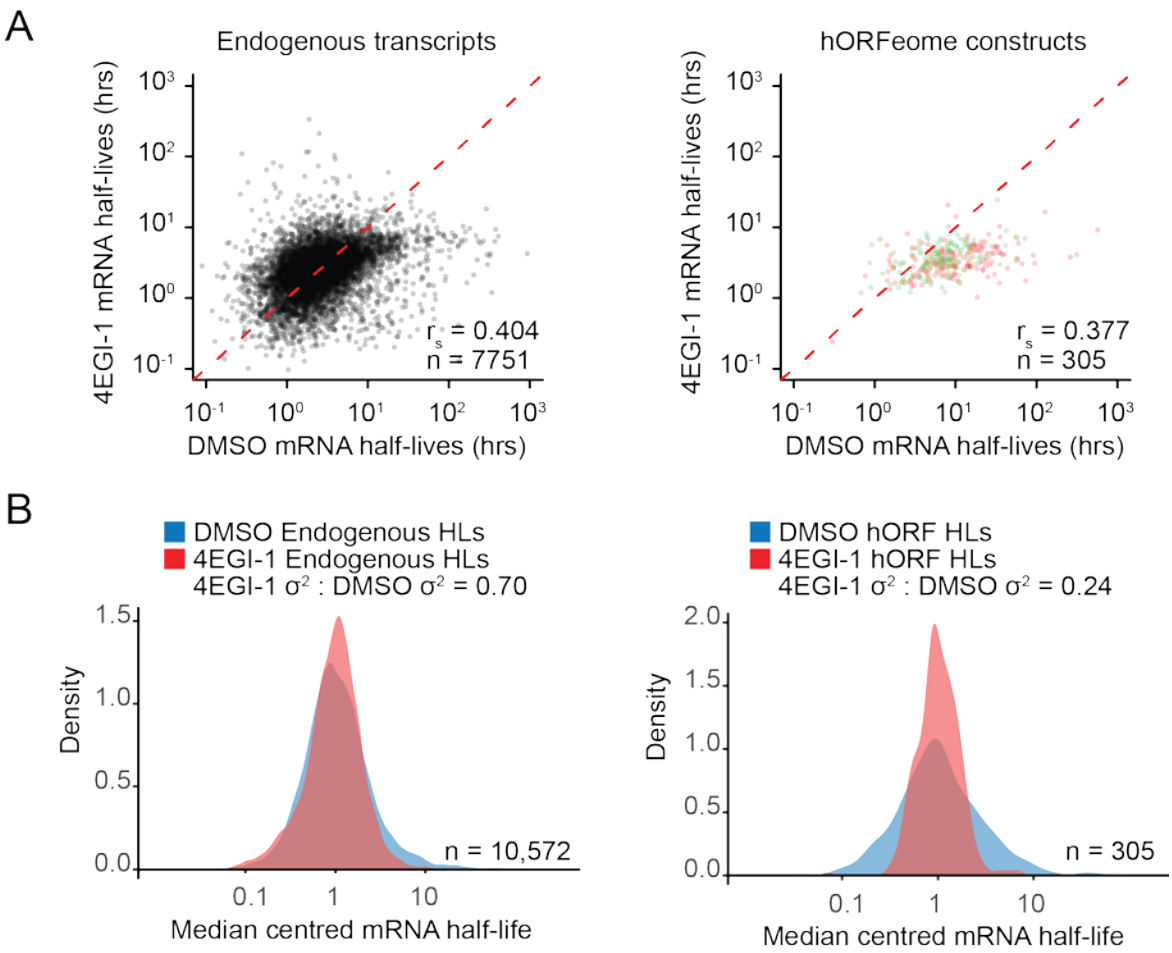

Figure 11. Coding sequence mediated regulation of mRNA stability is translation dependent. (A). mRNA half-lives are altered upon treatment with 4EGI-1. Endogenous transcript half-lives are calculated by averaging half-lives measured in cell line 1 or cell line 4. hORF construct half-lives are displayed for the subset of detectable hORF transcripts, with pool 1 and pool 4 genes indicated in red and green respectively. (B). Median-centred density plots comparing the variation in endogenous and hORF half-lives for the CL1-CL4 paired experiment following treatment with either DMSO or 4EGI-1. Ratios of calculated variance are indicated. Variance calculations were performed by randomly sampling 200 genes from endogenous or hORF datasets 1000 times and determining the median variance. The ratio of variances in 4EGI-1 to DMSO treatments is listed.

Chapter 4: Results Part II: Coding sequence determinants of mRNA stability

4.1 Longer coding sequences do not destabilize mRNAs

Having identified the mRNA coding sequence as a mediator of transcript stability, I next investigated which elements might drive these effects. Previous work in organisms ranging from bacteria (Dressaire et al., 2013) to humans (Duan et al., 2013) has identified transcript length as an intrinsic property of mRNAs that correlates strongly with transcript decay – shorter transcripts are considerably more stable. In fact, in budding yeast, the length of the CDS specifically has been found to be the single strongest predictive feature of degradation rates, explaining ~30% of the variance in endogenous half-lives (Geisberg et al., 2014; Neymotin et al., 2016). It has been posited that the mechanism behind this relationship might result from additional instability elements in the longer CDSs or from increased susceptibility to endonucleolytic attack (Feng and Niu, 2007).

I observed a similar relationship for endogenously derived mRNAs. There, a significant negative correlation is observed between stability and overall length (r_s = -0.265, $p < 10^{-16}$) (Fig12A), as well as CDS length (r_s = -0.200, $p = 9.56x10^{-6}$) (Fig12B). In contrast, no such relationship with CDS length is seen in the hORFeome (r_s = -0.060, p = 0.183) (Fig12B). Thus, I conclude that CDS length does not directly mediate increased instability and likely co-evolved with other features that directly stimulate mRNA decay. In support of this hypothesis, I observed a negative correlation between hORFeome stability and endogenous 3′ UTR length (r_s = -0.125, $p = 5.35x10^{-3}$) (Fig12B), possibly suggesting that 3′ UTR length may have co-evolved with other features within the CDS that can negatively control half-life. Overall, this finding suggests that CDS length does not directly change mRNA stability and highlights the power of the hORFeome approach for deconvoluting direct and indirect effects on gene regulation.

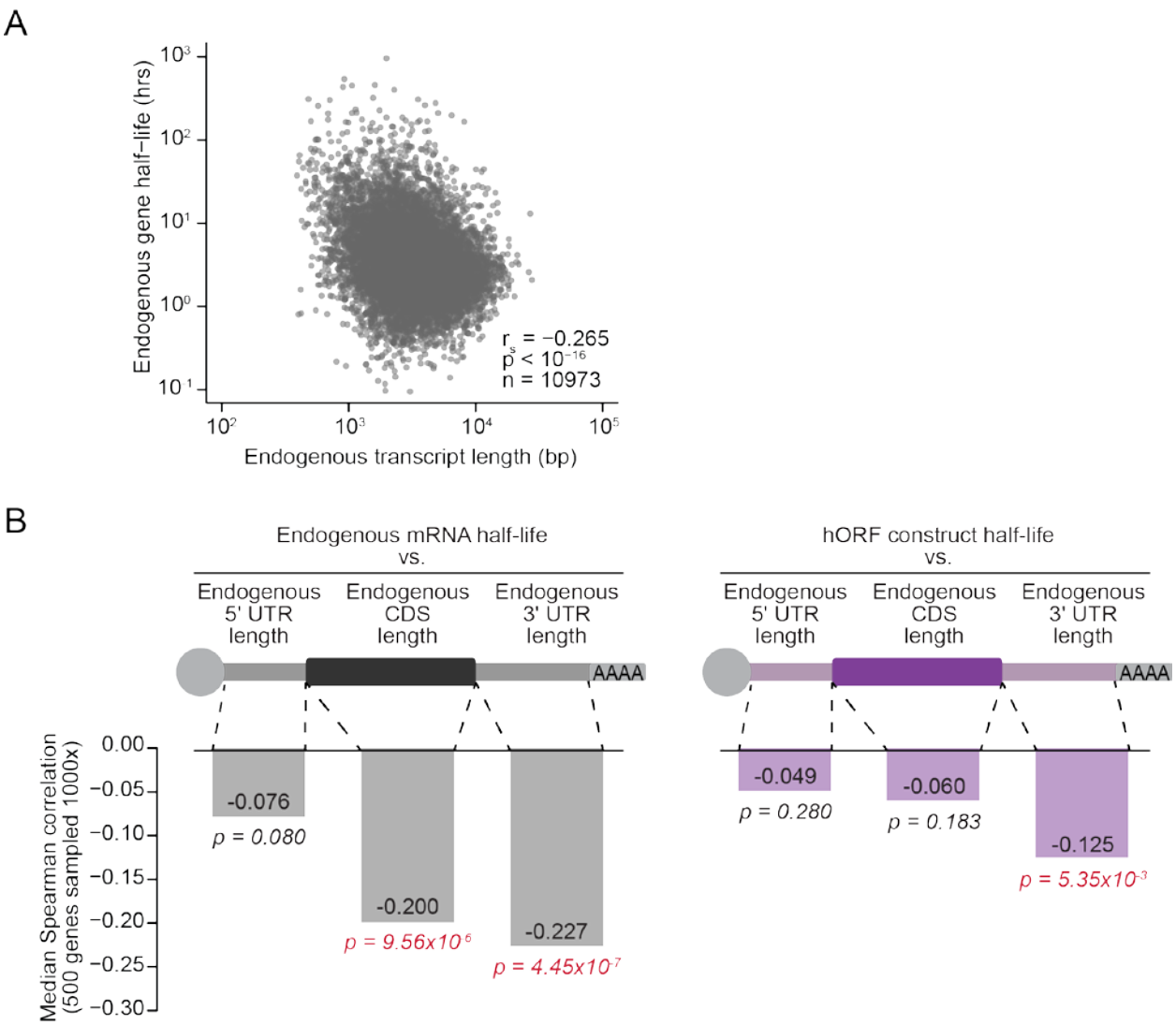

Figure 12. Coding sequence length drives the correlation between transcript length and decay for endogenous genes, but not for hORF constructs. (A). Transcript length for each endogenous gene was plotted against the mRNA half-life of that endogenous gene. Spearman correlation (r_s), p-value, and sample size (n) for this relationship is indicated. (B). Spearman correlations (r_s) were determined for the relationship between endogenous 5′UTR, CDS, or 3′UTR lengths, and endogenous or hORF gene half-lives. P-values of these Spearman correlations are indicated below, with $p < 0.05$ highlighted in red. Note that to compare across different sample sizes, Spearman correlations and corresponding p-values were calculated by subsampling 500 data points 1000 times and measuring the median Spearman correlation and p-value for each subset.

4.2 Coding sequence secondary structures are associated with mRNA stability

Complex secondary structures along the transcript body are also known to influence mRNA stability. In general, UTR folding energy (ΔG) is positively correlated with stability, i.e. transcripts with more folded UTRs (more negative ΔG) are less stable (Duan et al., 2013; Geisberg et al., 2014). In particular, stable structures in the 5′ UTR of budding or fission yeast transcripts strongly decrease both translation and stability of that transcript (Cheng et al., 2017), with some evidence of reduced ribosomal elongation caused by structures in the CDS as well (Chen et al., 2013). Given this, I set out to examine whether secondary structures within the CDS led to variation in hORF stability.

I focused on local secondary structures within the CDS, reasoning that these local structures would be most relevant for the elongating ribosome. I determined the folding energy for a sliding window of 100bp (sliding by 3 bp) across each CDS using ViennaRNA RNAfold (Lorenz et al., 2011) (Fig13A). Note that both median and minimum ΔG values across the CDS correlated with each other ($r_s = 0.740$, $p < 10^{-16}$), although they differed in absolute magnitude.

While secondary structure in the CDS only has a minor influence on endogenous mRNA stability ($r_s = 0.041$, p-value = $2x10^{-5}$), hORF constructs with more stable secondary structures within the CDS tend to be less stable ($r_s = 0.157$, p-value = $4x10^{-4}$) (Fig13B). In fact, when I restricted this analysis to endogenous genes with short 3′ UTRs (<60bp), CDS secondary structures now recapitulated this relationship first observed in the hORFeome ($r_s = 0.164$, p-value = 0.04) (Fig13C). Overall, this suggests that secondary structure elements in the CDS appear to play a role in regulating the stability of mRNAs, however their contribution to stability is likely masked by additional UTR-mediated regulation.

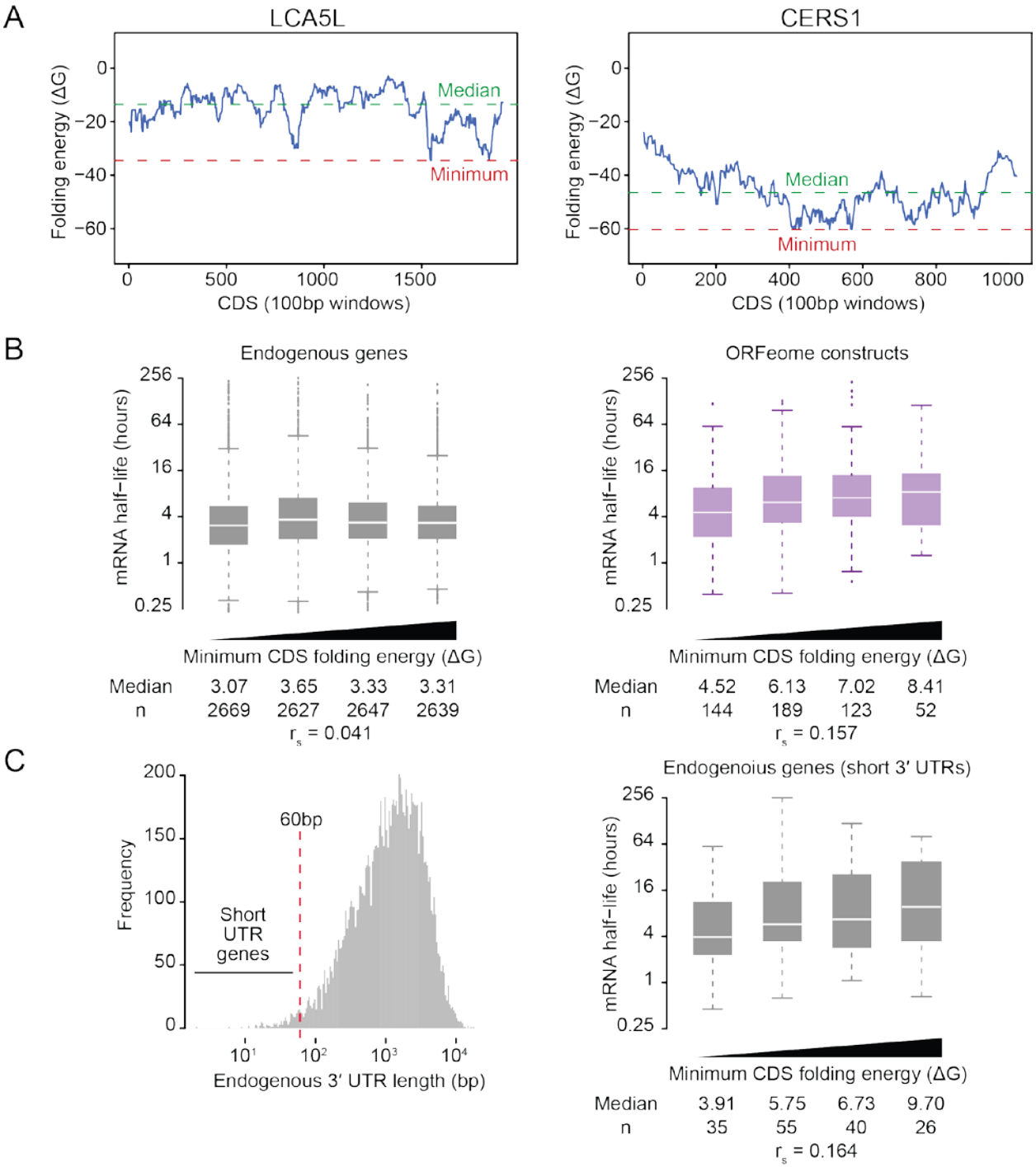

Figure 13. Secondary structures within the CDS are correlated with transcript stability. (A). Each gene's CDS was split into 100bp bins starting 3bp apart. 100bp sequences were folded using ViennaRNA RNAfold, and folding energy is plotted for two example genes. Windows with median and minimum folding energy are indicated with green and red lines respectively. (B). Endogenous mRNAs and hORF constructs were binned into quartiles based on the folding energy of the most structured (minimum ΔG) 100bp window within their CDS, and mRNA half-lives within each bin were plotted. Median half-life and sample size of each bin are indicated, along with the Spearman correlation (r_s) between the two variables. (C). Histogram representing the distribution of 3′UTR lengths across all transcripts, with the 60bp cutoff indicated. Endogenous transcripts with short 3′UTRs (<60bp) were then plotted as boxplots after binning as in (B).

4.3 CDS codon usage bias is a determinant of human mRNA stability

Codon usage has been consistently identified as a regulator of stability across a host of non-human model organisms, wherein certain codons are preferentially enriched in stable or unstable transcripts (Bazzini et al., 2016; Jeacock et al., 2018b; Mishima and Tomari, 2016; Nascimento et al., 2018; Neymotin et al., 2016; Presnyak et al., 2015). To determine which codons were associated with stable transcripts in human cells, I calculated the Codon Stabilization Coefficient (CSC), defined as the Spearman correlation between codon frequency and transcript stability (Fig14A, B). These calculations were reproducible across replicates (Fig14C). hORF constructs with a high frequency of these stabilizing codons had a higher half-life ($r_s = 0.301$, $p = 4x10^{-12}$), while a computational frameshift during calculation of codon frequency eliminates this relationship ($r_s = 0.003$, $p = 0.9$), indicating that these codon usage effects are independent of the general effects of sequence or structure composition.

Optimal codons are thought to be determined in large part by tRNA abundance, where more abundant cognate tRNAs lead to faster decoding than less abundant tRNAs. I asked whether differences in tRNA expression level could explain the variation in CSC values I observed. However, although tRNA gene copy number (GCN) is a good proxy for charged tRNA abundance in yeast, determining human tRNA GCN is non-trivial due to the overall large number of tRNA-like unexpressed pseudogenes. As such, a key step in defining optimality in humans is to accurately measure tRNA levels. I thus made use of several recently published datasets that employ various sequencing methods to detect tRNA abundance in humans (Cozen et al., 2015; Gogakos et al., 2017; Mattijssen et al., 2017; Qin et al., 2015; Shigematsu et al., 2017; Zheng et al., 2015). I observed considerable variation in tRNA read counts, and each metric weakly correlated with the frequency of codon usage in the HEK293T transcriptome (Fig15A). For the purposes of my analyses, I proceeded with the tRNA level dataset derived from HEK293T cells (Zheng et al., 2015), the same background as the hORFeome expressing cell lines.

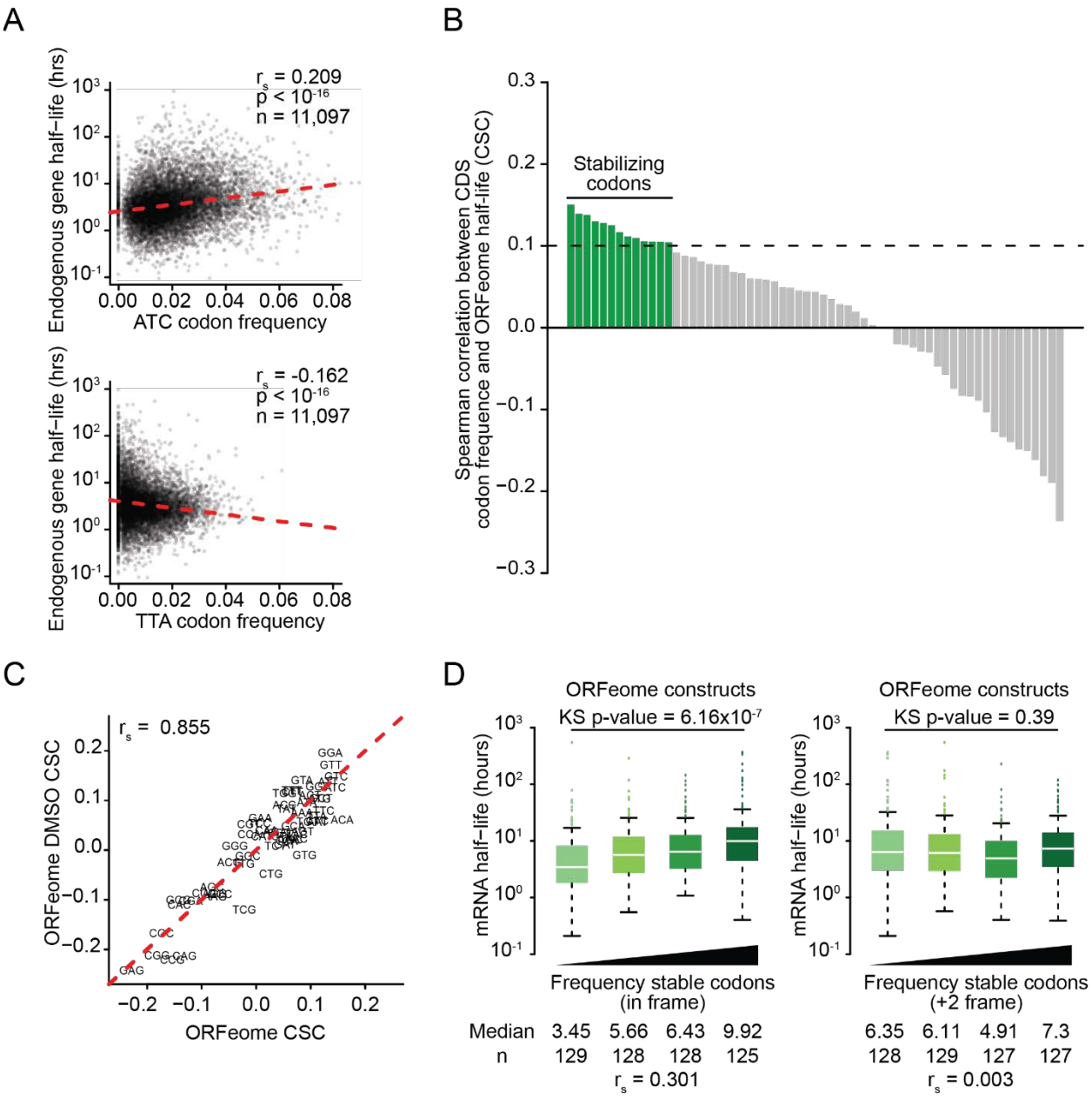

Figure 14. Certain codons are enriched on stable mRNAs. (A). The codon stabilization coefficient (CSC) of a codon represents the Spearman correlation of the frequency of a particular codon on a given mRNA with the half-life of that mRNA. CSC calculations for codons ATC and TTA derived from endogenous half-lives are indicated, along with corresponding Spearman correlation (r_s) and p-value. (B). Distribution of hORFeome-derived CSCs, with all codons above 0.10 CSC defined as stabilizing. (C). Scatterplot representation of CSCs calculated from two independent replicates of ORFeome half-lives, with Spearman correlation (r_s) indicated. (D). hORF mRNAs were separated into quartiles based on frequency of stabilizing codons in the CDS, determined either in or out (+2) of frame. Median half-lives and sample size (n) are indicated for each bin, and p-value is calculated by performing the KS test across bin 1 and bin 4. Spearman correlation (r_s) is listed.

Strikingly, hORF mRNAs with a high frequency of optimal codons (the 13 codons with highest cognate tRNA read counts) were significantly more stable than mRNAs depleted of these codons (KS test p = 0.03) (Fig15B). This codon-mediated stabilisation of hORF mRNAs was lost following treatment with 4EGI-1 (KS test p = 0.80) (Fig15B), highlighting its dependence on translation. Further, this relationship was lost upon a computational frameshift of the CDS, suggesting that the effects of codon usage are translation dependent rather than general effects of sequence or structure (KS test p = 0.20, 0.44) (Fig15B). Overall, codons with higher tRNA decoding rates tend to be associated with mRNAs that are stable in the presence of translation.

4.4 Amino acid usage is a determinant of human mRNA stability

Nonetheless, codon usage did not explain a large amount of the variation in hORF mRNA stabilities. Amino acid identity can also affect elongation rates due to different peptide bond formation rates or interactions with the ribosomal exit tunnel (Johansson et al., 2011; Lareau et al., 2014; Tanner et al., 2009). For instance, stretches of positively charged amino acids (lysine, arginine, histidine) in the nascent polypeptide show a length-dependent additive trend in slowing ribosome translocation, possibly due to their interaction with the negatively charged ribosome exit tunnel (Charneski and Hurst, 2013; Lu and Deutsch, 2008; Sabi and Tuller, 2015). In addition, polyproline motifs also slow peptide-bond formation, and proline is generally associated with slower translation elongation speeds and ribosome stalling (Artieri and Fraser, 2014; Gardin et al., 2014; Pavlov et al., 2009; Pelechano and Alepuz, 2017). I thus hypothesized that amino acid use might mediate differences in hORF mRNA stability.

I calculated the amino acid stabilization coefficient (AASC), defined as the Spearman correlation between the frequency of an amino acid on a given transcript and the stability of that transcript (Fig16A), and I observed a wide range of stabilization effects (Fig16B). While AASCs calculated from independently derived half-life datasets were highly correlated (r_s = 0.913, p = $3x10^{-6}$) (Fig16C), AASCs derived from endogenous mRNAs shared some differences with those from hORF transcripts (r_s = 0.570, p = 0.01) (Fig16D), presumably due to masking effects of endogenous regulatory elements.

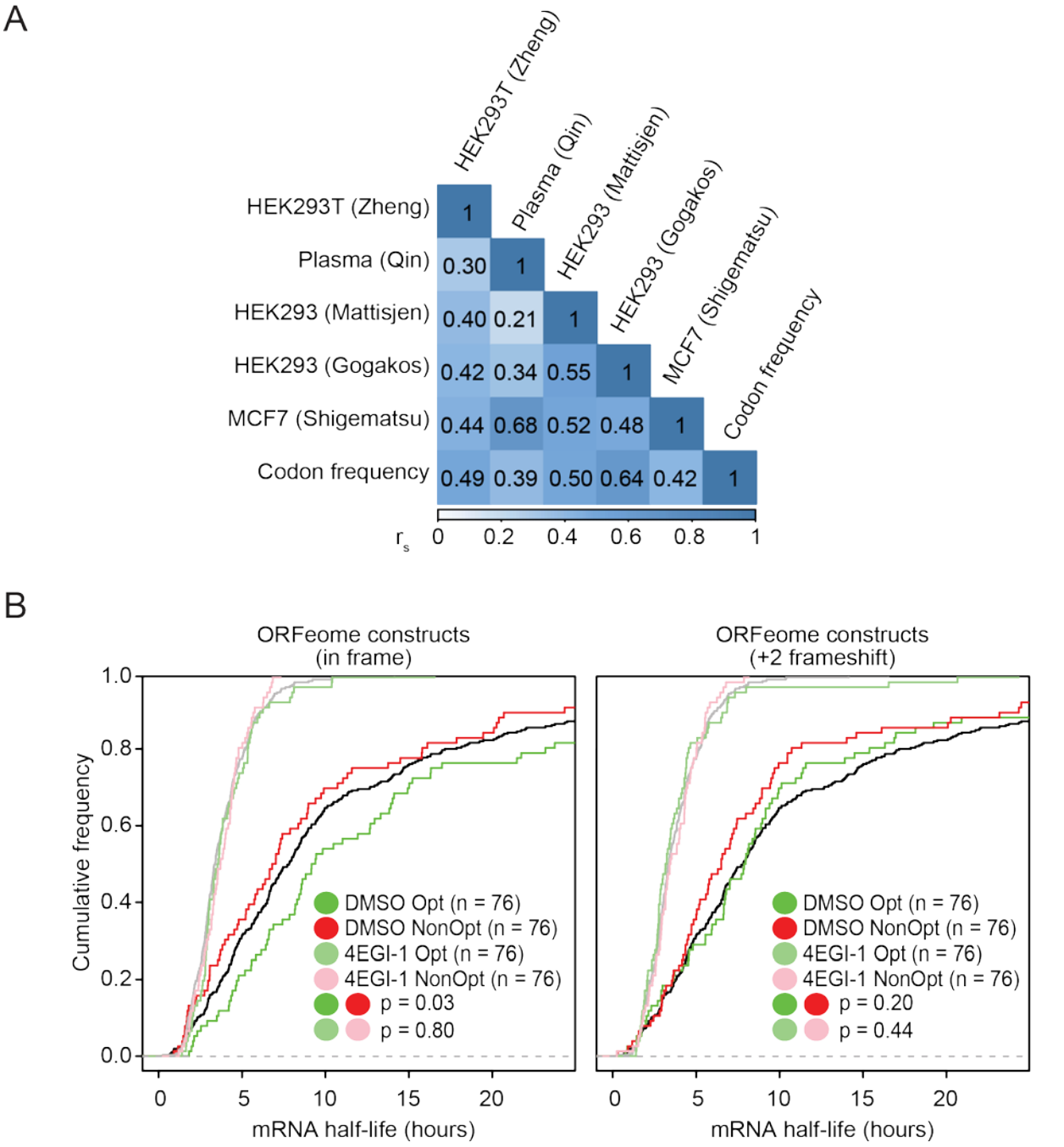

Figure 15. Optimal codons stabilize mRNAs in a translation dependent manner. (A). Spearman correlations of tRNA sequencing read counts obtained from the indicated cell type and publication, as well as human coding sequence codon frequency. (B). Cumulative frequency distribution plots of hORF half-lives calculated following DMSO or 4EGI-1 treatment. Optimal codons were defined as those with the 13 highest read counts in the Zheng *et al*. dataset. Genes were split into optimal, non-optimal, or other based on the frequency of optimal codons in their CDS. P-values are calculated using the KS test and are indicated for the comparisons shown by the coloured dots. Computational frameshifts are shown with similar analyses performed.

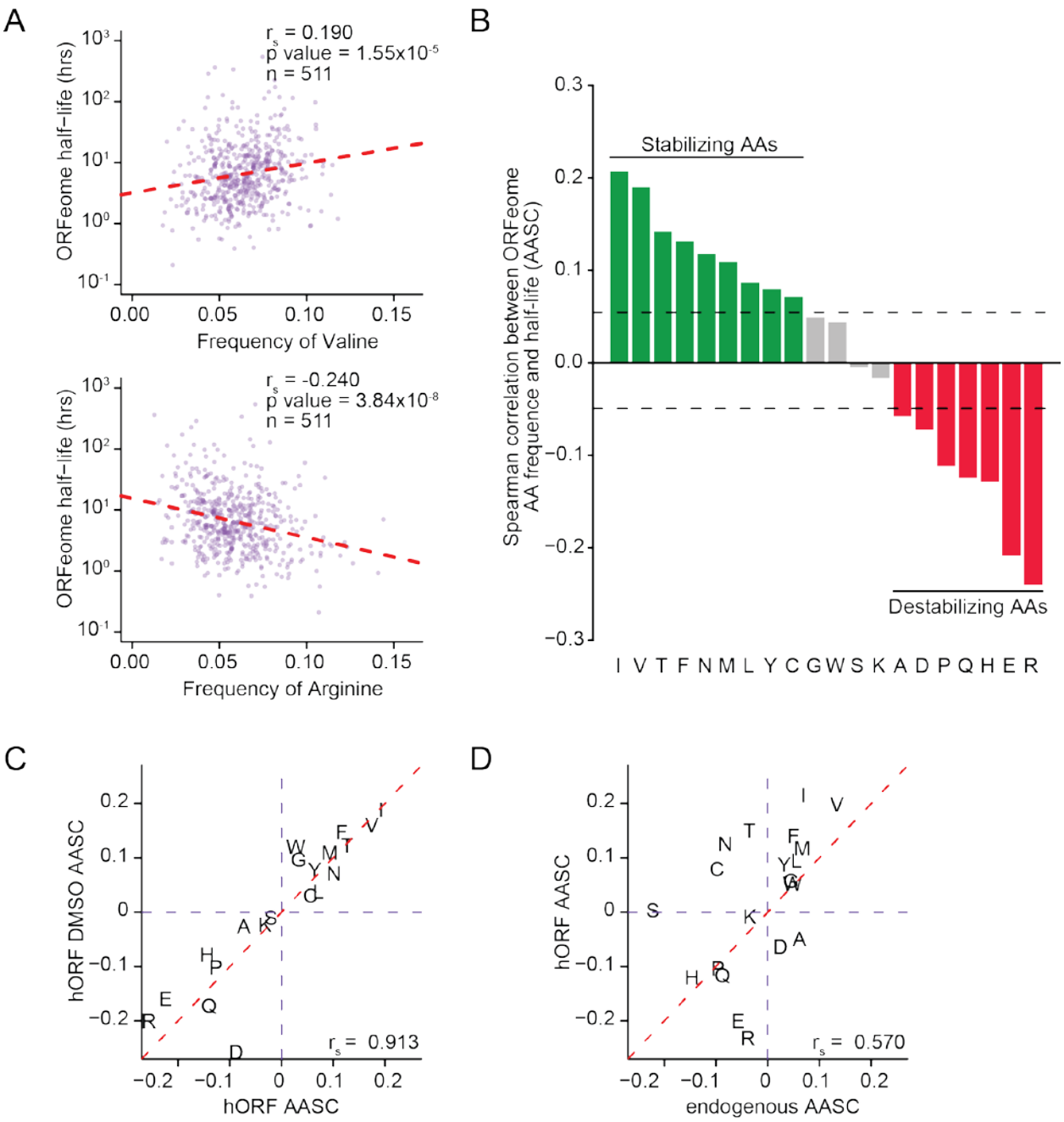

Figure 16. Certain amino acids are enriched on stable hORF constructs. (A). Amino Acid Stabilization Coefficients (AASCs) were calculated by determining Spearman correlations between frequency of an amino acid on a given transcript and the stability of that transcript. AASC calculations for Valine and Arginine derived from hORFeome half-lives are indicated, along with corresponding Spearman correlation (r_s), sample size (n), and p-value. (B). Distribution of hORFeome-derived AASCs, with all amino acids above 0.05 defined as stabilizing. (C). Comparison of AASCs derived from independent hORFeome half-life measurements, with Spearman correlation (r_s) listed. Amino acids are represented by their one-letter IUPAC code. (D). Comparison of AASCs derived from endogenous gene half-lives or hORF construct half-lives, with Spearman correlation (r_s) listed.

I observed that positively and negatively charged amino acids, as well as proline, had destabilizing tendencies, consistent with a model where slow translation elongation leads to mRNA destabilization (Fig17A). Strikingly, hORF-derived AASCs shared a very strong correlation with amino acid hydrophobicity (r_s = 0.741, p = $2x10^{-4}$) (Kyte and Doolittle, 1982) (Fig17A/B). Translation frame-shift controls (r_s = 0.341, p = 0.1, and r_s = 0.165, p = 0.5) demonstrated that this relationship was dependent on reading frame and is thus likely independent of confounding factors such as GC content (Fig17B).

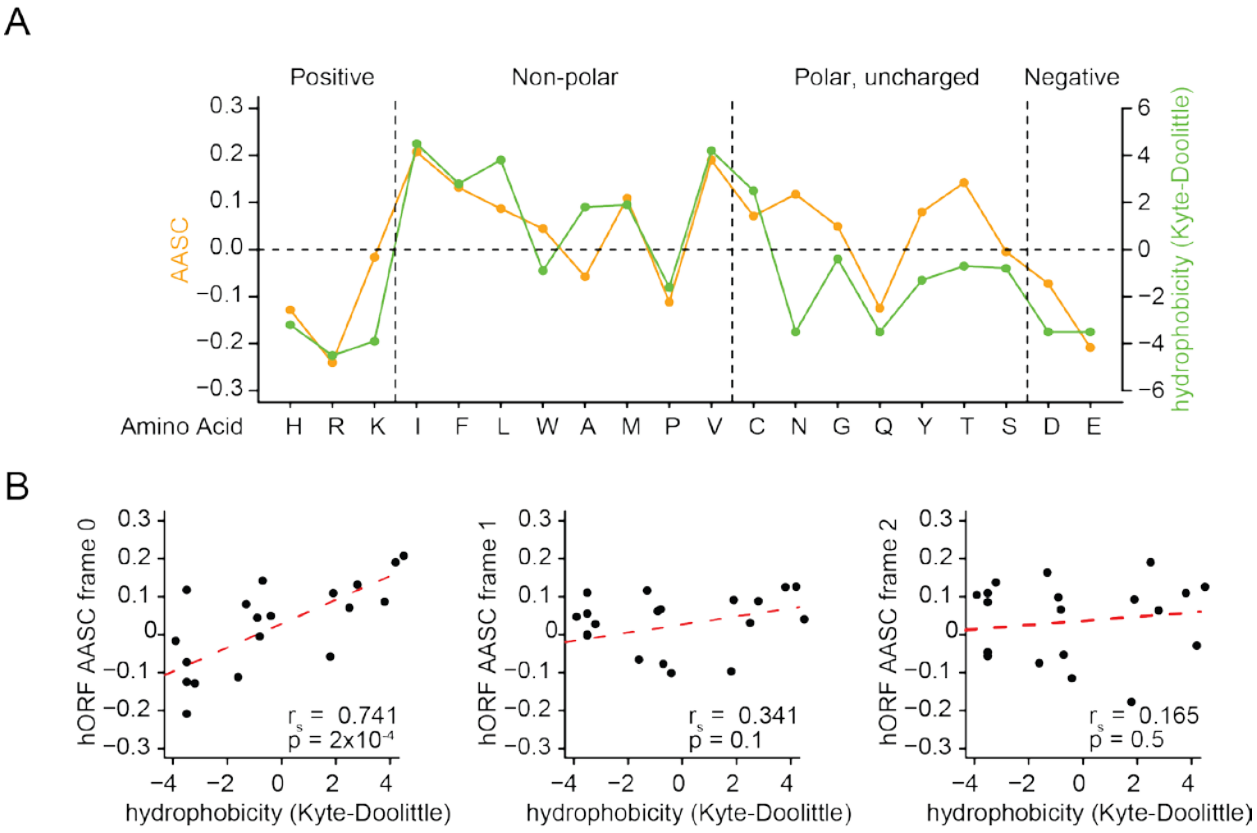

Figure 17. Hydrophobic amino acids are enriched in stable hORFs. (A). Line plot depicting the correlation between hORF-derived AASCs and hydrophobicity as measured by the Kyte-Doolittle scale. Amino acid classes are indicated. (B). Scatterplots highlighting correlation between hydrophobicity and AASCs, with amino acid frequencies calculated using different translation frames. Spearman correlation (r_s) and corresponding p-value are listed.

In addition, amino acid stretches magnified this relationship. Transcripts with a stretch of 5 hydrophobic residues tended to be more stable (r_s = 0.194) (Fig18), while transcripts with stretches of 3 charged residues or 2 proline residues were less stable (r_s = –0.189, -0.106 respectively) (Fig18). Overall, these results suggest that amino acid usage also plays a role in global CDS-mediated stability regulation, possibly by impacting ribosomal translocation rates.

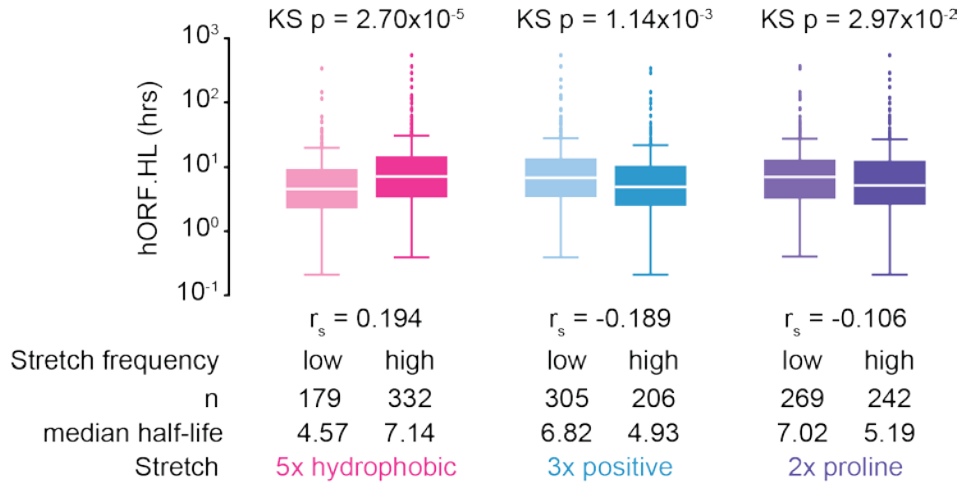

Figure 18. Amino acid stretches magnify stability effects. Counts of the indicated stretches were calculated across all human CDSs and normalized to the CDS length to obtain stretch frequency. All human genes were divided equally into two groups with low or high frequency of the indicated stretch pattern. Half-lives for hORF constructs falling within these two groups are represented as boxplots, with median half-life values and sample size (n) indicated. Spearman correlation (r_s) between stretch frequency and hORF half-life is listed. P-values represent the significance of the differences between the two groups of transcripts, as calculated by the KS test.

4.5 Relative contributions of codon and amino acid usage to human mRNA stability

Previous work in budding yeast had identified codon usage as a major predictor of mRNA stability (Presnyak et al., 2015), while my work in the hORFeome collection has identified both codon and amino acid usage as major contributors. I next set about to determine their relative contributions in humans and yeast.

I first measured the variance in CSCs across each of the 2-6 codons encoding an individual amino acid. Interestingly, in yeast, each amino acid had a wide range of CSCs, with most residues having the choice between stable or unstable codons (Fig19A top, Fig19B left). In other words, in yeast, AASC is not a good predictor of CSC (r_s = 0.301, p = 0.02). On the other hand, CSCs from the human ORFeome collection are almost entirely a function of AASC (r_s = 0.714, p

$= 10^{-10}$), suggesting that the choice of amino acid usage in itself constrains that stability contributions of the codons encoding it (Fig19A bottom, Fig19B right).

Thus, while codon usage impacts translation elongation during tRNA decoding steps, amino acid choice impacts peptide bond formation rates and passage through the ribosomal exit tunnel to modulate translation. Different organisms appear to be more constrained at different steps – tRNA decoding plays a broader role in budding yeast, while amino acid effects are more predominant in humans.

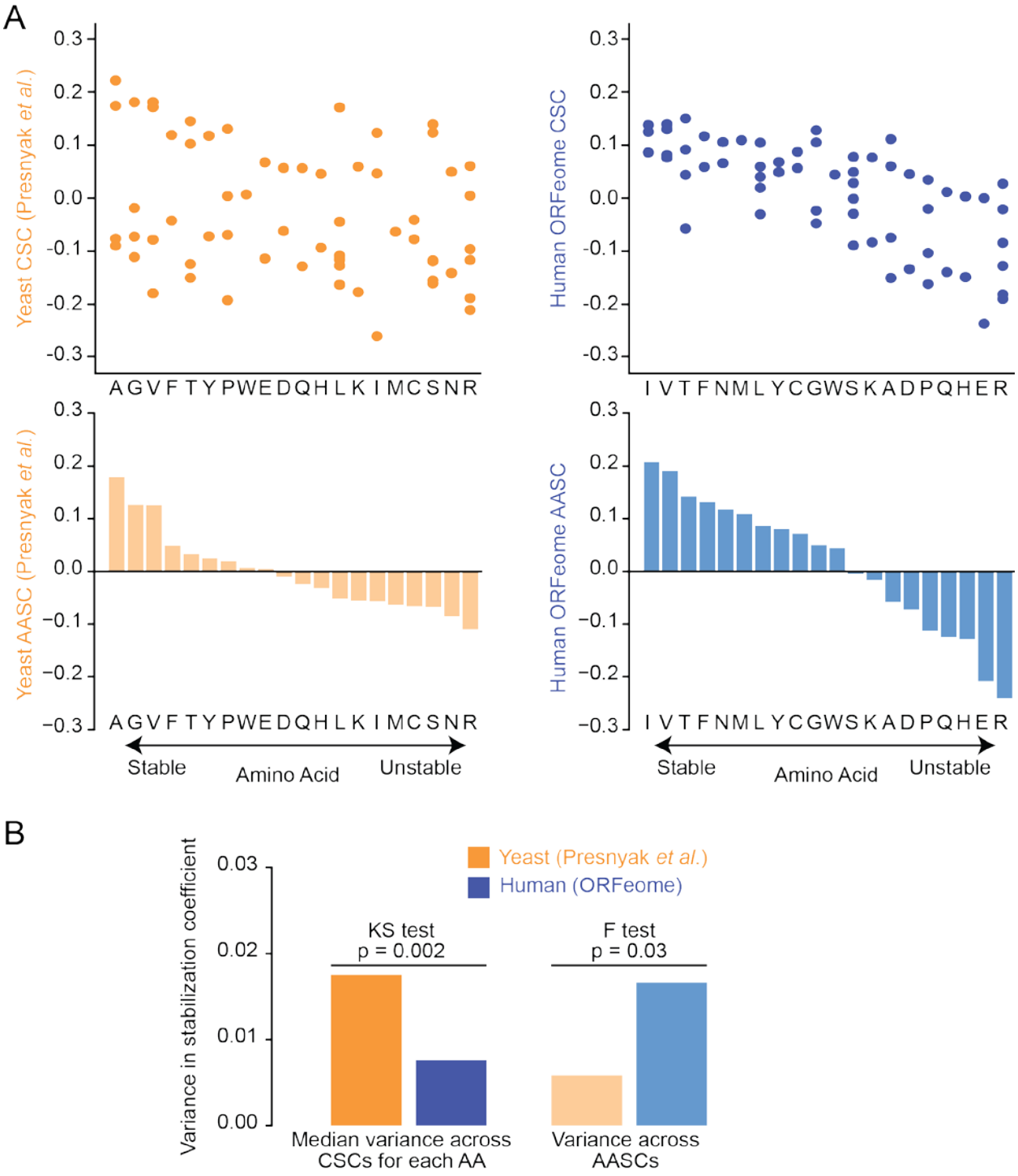

Figure 19. Relative importance of codon and amino acid usage in yeast and humans. (A). CSCs and AASCs were calculated from yeast half-life data (Presnyak *et al*., 2005) and human ORFeome half-lives. Amino acids were ordered by descending AASC for each organism (bottom), and CSCs for all codons encoding the indicated amino acid were plotted as dots (top). (B). Variance was calculated across all CSCs for each amino acid with more than 1 codon, and the median of these values is plotted as a bar (left). P-value comparing distribution of CSC per AA variance was calculated using the KS test. Variance was also calculated across all AASCs (right). P-value comparing variances was calculated using F-test.

Chapter 5: Discussion

5.1 Key findings from hORFeome stability measurements

The CDS and UTRs of each transcript have co-evolved to cooperatively regulate gene expression across many different contexts. Given these complex relationships involving multiple variables, it becomes challenging to determine the extent to which each individual CDS and UTR element directly causes differences in transcript stability. Probing a collection of human ORFeome mRNAs has the unique advantage of deconvoluting the effects of CDS elements from those of its corresponding UTRs. Importantly, in contrast to smaller scale reporter analyses, an hORFeome-wide examination of stability has the advantage of a larger sample size to identify relationships at a global level. In addition, reporter analyses often alter only one aspect of a CDS at a time which, while useful for isolating the impact of sequences at a finer resolution, miss the context of the remainder of the CDS. In contrast, hORF constructs probe entire CDSs, hence any long-range effects on stability (for instance prolonged slowing of ribosome elongation, or long-range RNA-RNA interactions) are still detected.

The first major finding from hORFeome stability measurements is that changes in coding sequence alone can produce a wide range in half-lives, as diverse as those measured across endogenous genes (Fig8C). Furthermore, this large range in stability collapses in the absence of translation (Fig11B), suggesting that translation recognizes CDS features to either stabilize or degrade transcripts. These results are in line with reporter assays in a range of different organisms (Bazzini et al., 2016; Boël et al., 2016; Mishima and Tomari, 2016; Presnyak et al., 2015; Radhakrishnan et al., 2016).

Second, this work begins to identify features within the CDS that can explain this variation. I focused on several different possible explanations: length, secondary structure, codon use, and amino acid use.

Although CDS length is generally thought to be strongly correlated with transcript stability (Feng and Niu, 2007; Geisberg et al., 2014; Neymotin et al., 2016), this relationship is lost in hORF constructs (Fig12B). A possible explanation for this observation is that CDS length does not directly cause instability, but that it is instead correlated with other elements in the UTR that drive this relationship. This is a good example of the power of using the hORFeome to

deconvolute correlated CDS and UTR features to identify which is more directly responsible for stability effects.

Secondary structure in the UTR has previously been linked with decay (Duan et al., 2013; Geisberg et al., 2014), but a systematic analysis of the effects of CDS structure has not been performed. Strong secondary structures within the CDS might impede ribosome translocation and lead to decay (Tunney et al., 2018), however only a very weak relationship was observed with endogenous mRNA stability (Fig13B). This relationship was considerably stronger across hORF constructs – less structured CDSs were found to be more stable (Fig13B). Interestingly, these conclusions from hORF constructs are particularly relevant to endogenous genes with short 3′ UTRs (Fig13C).

Pioneering work in budding yeast identified codon usage as a key factor for mRNA stability (Neymotin et al., 2016; Presnyak et al., 2015; Radhakrishnan et al., 2016). Specifically, transcripts with more optimal codons (i.e., those with a high tRNA supply to demand ratio) tend to be more stable. Extending this analysis to humans has been particularly challenging due to the difficulty in identifying tRNA gene copy number from human genome sequencing data. Further, tRNA-sequencing methods to directly measure abundance are not very advanced, with high variation observed between datasets from different groups and tissues (Fig15A). Nevertheless, I do observe that certain codons are correlated with stable hORFs in a translation dependent manner (Fig14, Fig15), however there is a much smaller range in stabilization effects as compared with yeast (Fig19).

This finding drove me to identify additional features that can explain the large variance in hORFeome stability. AA usage within the CDS has been widely linked to translation elongation rate (Johansson et al., 2011; Lareau et al., 2014; Tanner et al., 2009), but has never been linked with mRNA stability in humans. Recent work in yeast (Hanson et al., 2018) and zebrafish (Bazzini et al., 2016) has examined AA usage in this context, however both studies found that codon usage was the main driver of stability. Our data are among the first to indicate that certain AAs stabilize CDSs in human cells (Fig16, Fig17, Fig18), with hydrophobic residues in particular being correlated with stable transcripts. In contrast to budding yeast, AA usage plays a broader role in modulating stability in humans (Fig19). These findings suggest a model wherein the slowing of translation elongation by codon and AA usage results in the destabilization of transcripts in human cells (Fig20).

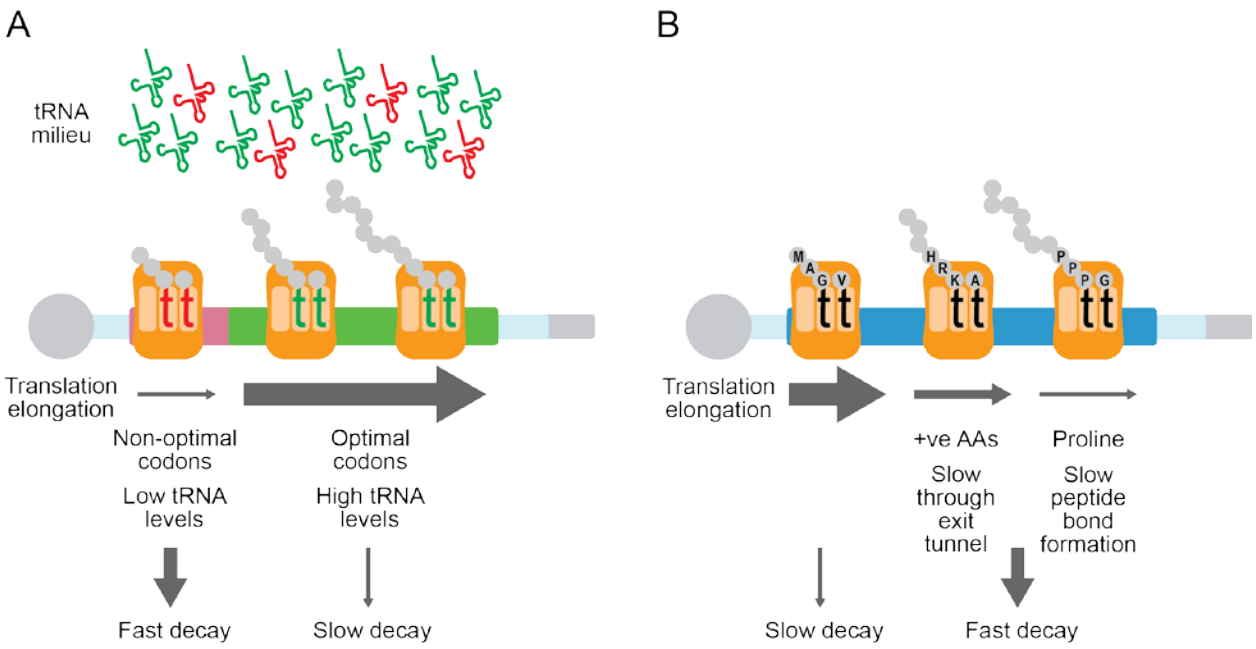

Figure 20. Codon and amino acid usage impact mRNA stability in humans. Translation elongation is a function of both codon and AA usage. (A). Codons are decoded at variable rates by tRNAs, dependant on the supply to demand ratio of tRNA molecules. (B). Amino acids modulate elongation as well, with each AA having variable peptide bond formation rates and movement through the ribosomal exit tunnel. Slowing of translation elongation is sensed by as yet unidentified factors that trigger the destabilization of the transcript.

5.2 Limitations of the current hORFeome stability datasets

When drawing conclusions from this hORF stability dataset, it is important to keep in mind certain caveats and considerations. First, there are biases in the hORFeome library that arise during cell line generation. Due to upper size limits in lentiviral packaging, a considerable number of hORF constructs with long CDS lengths are lost from the pooled virus preparation (Fig4), resulting in a library biased toward short CDSs.

Second, due to the sequencing method used, I restricted analysis to hORF constructs that are not endogenously expressed. I chose to maintain a relatively low false discovery rate of ~10%, at the cost of reduced sample size. For instance, of the over ~3,000 hORFs infected into cell line 1, I was limited to a small sample size of 359 detectable hORF constructs, of which I was only able to measure reliable half-lives for 280 constructs (Fig8A). To mitigate differences in sample size, I sub-sampled data points from each of the endogenous and hORF stability datasets numerous times and calculated the average statistics across each sub-sample (Fig8D). In addition, analysing only non-endogenously expressed hORFs biased the dataset away from conserved "housekeeping" genes which tend to be highly expressed, stable, and enriched in optimal codons,

and toward non-essential, low-optimality CDSs. Most importantly, this restriction also prevented us from analysing biologically interesting hORF constructs such as those corresponding to structural or ribosomal protein genes. We were unable to measure stability of both endogenous and hORF transcripts for the same gene, which would have also been very valuable in determining the net effect of endogenous UTRs.

Third, the Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element (WPRE) in the 3′ UTR of hORFeome transcripts is generally included in lentiviral vectors to enhance gene expression by reducing read-through transcription, increasing nuclear export, and stabilising mRNAs (Zufferey et al., 1999). WPREs are common in viral constructs lacking any introns, since these transcripts do not undergo splicing-dependent deposition of proteins that function to increase expression. In future iterations of hORFeome-wide stability screens, a mini intron could be inserted into the 5′ UTR to facilitate this splicing-dependent deposition of expression-enhancing protein complexes. Following this, a range of different 3′ UTRs can be examined in the context of the hORFeome to confirm that our conclusions are in fact independent of the WPRE.

Finally, translation inhibition by 4EGI-1 likely results in indirect effects of the drug treatment. While it would be expected from previously published data (Bazzini et al., 2016; Mishima and Tomari, 2016; Radhakrishnan et al., 2016) that a reduction in translation rates during 4EGI-1 treatment would result in reduced variance across hORF stability, an alternate explanation is that blocking translation could be reducing the abundance of key proteins that may be involved in mediating CDS-dependent regulation of decay rates. These concerns are somewhat mitigated by including computational-frameshift controls (Fig15B). However, an additional strategy would be to generate an hORFeome library with a translation inhibitory sequence in the 5′ UTR, such as the drug-responsive RocA-eIF4A target sequence (Iwasaki et al., 2016), to allow for the selective repression of translation initiation. Stability measurements using this modified hORFeome collection could then provide additional evidence that translation is required for CDS-mediated stability regulation.

In spite of these limitations with this first iteration of hORFeome-wide stability measurements, this system has provided a powerful framework to demonstrate that coding sequence alone can result in a wide range of stability and has allowed for the identification of novel CDS elements that may be involved in this regulation. Knowing the limitations of this dataset provides insight

into the development of future iterations of hORFeome cell lines and stability data to continue to increase the impact of this system.

5.3 Future directions

The first step in expanding the power of this framework is to measure hORFeome-wide stability from currently existing hORFeome cell lines. This would expand sample size from only non-endogenously expressed genes to all ~16,000 hORF constructs, as well as allow for a comparison of hORF genes with their endogenous counterparts. To this end, I set out to develop a novel library preparation strategy to selectively detect hORF constructs while excluding their endogenous counterparts. Lexogen's QuantSeq^{TM} 3′ mRNA-Seq Kit is a commercially available library preparation protocol designed to generate Illumina-compatible sequencing libraries for the 3′ end of polyadenylated RNAs by using Oligo-dT primers during first strand synthesis. To selectively detect hORF constructs, I have replaced this Oligo-dT primer with a primer specific to the hORF construct 3′ UTR, hence first strand synthesis is only carried out on hORF mRNAs while endogenous genes are excluded (Fig21A). Thus far, I have shown that hORF-specific Quant-seq is an excellent method for enriching hORF constructs, however this technique requires further optimization to increase both measurement accuracy and precision.

After optimizing detection of hORF constructs using modified Quant-seq, I am also interested in implementing alternate methods for measuring half-lives. SLAM-seq is a recently published method for measuring RNA kinetics (Herzog et al., 2017), which relies on reverse-transcription dependent T-to-C conversion to detect 4SU incorporation over time in a high throughput sequencing compatible manner (Fig21B). The advantage of this method is that it requires considerably lower input RNA compared with the fractionation methods utilised in this thesis. Thus far, I have tested SLAM-seq and observed a large percentage of T-to-C conversions (Fig21B), with increasing conversion rates over the 4SU labelling time course. Unfortunately, SLAM-seq measurements yield half-lives that are weakly correlated with previous measurements. Additional experiments are required to determine whether this discrepancy is due to varying sensitivity to extracellular 4SU concentration changes over time. Overall, Quant-seq coupled with SLAM-seq while promising at this stage, requires additional optimization to obtain hORFeome-wide stability measurements. Once a larger dataset of hORF half-lives is available, a more quantitative model of all CDS elements predicting mRNA stability can be built.

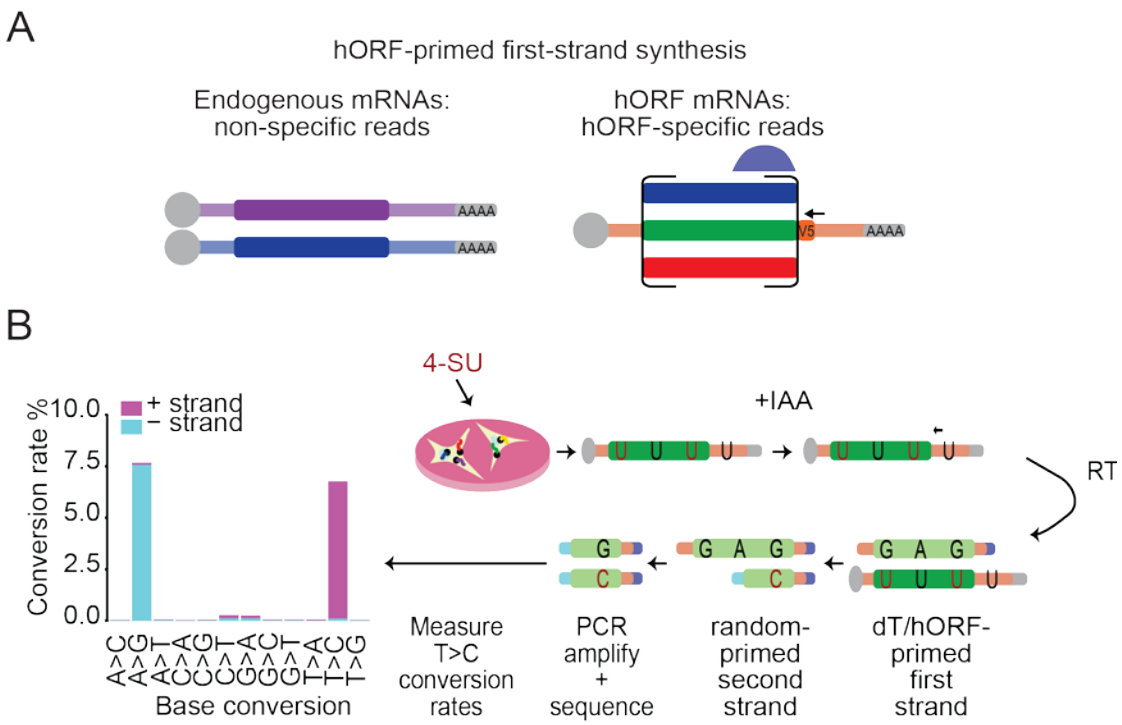

Figure 21. Measuring mRNA stability hORFeome-wide. (A). Schematic representation of hORF-primed Quant-seq library preparation. During first strand synthesis, hORF UTR-specific primers are used to generate cDNA exclusively from hORF transcripts. Random priming is performed for second strand synthesis, and resultant libraries are PCR-amplified and sequenced. Reads are expected to pile up at the 3′ region of hORF CDSs. Endogenous mRNAs will not be primed. (B). Schematic representation of SLAM-seq, an alternate method to measure 4-SU incorporation over time. Incorporated 4SU is modified using Iodoacetamide and is misread as a C by reverse transcriptase during first strand synthesis. Percentage conversion rates for each base transition in a SLAM-seq library were measured and plotted, for reads mapping to the - and + DNA strands.

In addition to increasing sample size, it is critical to confirm that the relationship with codons and amino acids are causal, as opposed to correlative. In collaboration with Dr. Jeff Coller, we have shown that firefly luciferase variants with different synonymous codon usage have different stabilities. Consistent with my analysis, optimal variants tend to be more stable (unpublished data, M. Forrest and J. Coller, Case Western Reserve University). In the future, additional reporter analyses where CDS length, secondary structure, and amino acid usage are changed will be necessary.

With increased sample size, an analysis of positional effects of codons within the CDS would also be interesting to examine. In general, codon usage near the 5′ end of the transcript is thought to impact translation initiation (Chu et al., 2014), while codon usage near the 3′ end of the CDS has the largest effect on stability (Mishima and Tomari, 2016; Radhakrishnan et al., 2016).

Examining hORFeome-wide measurements would help ascertain the extent to which this division exists in humans.

A biologically relevant test of the importance of codon and AA usage in the hORFeome would be to modify the intracellular milieu by perturbing the levels of charged tRNAs and AAs. One simple way to induce these changes is by depleting AAs in human cell culture media, which has been shown to result in codon usage-mediated differential translation of mRNAs (Saikia et al., 2016). We would predict that CDS stability effects would also be altered during starvation. Another interesting comparison would be to examine CDS effects on stability in a more diverse set of cell lines. In collaboration with Dr. James Ellis' lab, we are measuring endogenous mRNA stability on cells at varying stages of neuronal development. Upon obtaining this data, determining the effects of coding sequence on different cell types will be of interest, particularly if tRNA pools are altered over neurodevelopment.

After gaining an understanding of CDS-mediated regulation of stability, it becomes key to explore the mechanisms governing this relationship. With Dhh1 identified as a key regulator of this pathway in yeast (Radhakrishnan et al., 2016), the next step would be to investigate its human ortholog DDX6 as a candidate regulatory factor. Previously published DDX6 eCLIP datasets from the ENCODE consortium (Encode Consortium, 2012) in HepG2 and K562 cell lines can be analysed to determine whether DDX6 shows preferential binding to optimal or non-optimal transcripts. Reporters with variable codon and AA usage can also be used for these mechanistic studies, including DDX6 occupancy analysis and tethering experiments. Through these and other tests, I hope to begin to decipher the mechanisms governing CDS mediated mRNA decay regulation in humans.

References

Agris, P.F., Vendeix, F.A.P., and Graham, W.D. (2007). tRNA ’ s Wobble Decoding of the Genome : 40 Years of Modification. 1–13.

Akashi, H. (2003). Translational selection and yeast proteome evolution. Genetics *164*, 1291–1303.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

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

Barreau, C., Paillard, L., and Osborne, H.B. (2018). AU-rich elements and associated factors : are there unifying principles ? Nucleic Acids Res. *33*, 7138–7150.

Bazzini, A.A., Viso, F., Moreno-mateos, M.A., Johnstone, T.G., and Charles, E. (2016). Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. EMBO J. 1–17.

Bellí, G., Garí, E., Piedrafita, L., Aldea, M., and Herrero, E. (1998). An activator / repressor dual system allows tight tetracycline-regulated gene expression in budding yeast. Nucleic Acids Res. *26*, 942–947.

Bensaude, O. (2011). Inhibiting eukaryotic transcription. Which compound to choose? How to evaluate its activity? Transcription *1264*.

Berkovits, B.D., and Mayr, C. (2015). Alternative 3’ UTRs act as scaffolds to regulate membrane protein localization. Nature *522*, 363–367.

Bicknell, A.A., and Ricci, E.P. (2017). When mRNA translation meets decay. Biochem. Soc. Trans. *45*, 339–351.

Bischof, J., Sheils, E.M., Björklund, M., and Basler, K. (2014). Generation of a transgenic ORFeome library in Drosophila. Nat. Protoc. *9*, 1607–1620.

Boël, G., Letso, R., Neely, H., Price, N., Wong, K., Su, M., Luff, J.D., Valecha, M., Hunt, J.F., Everett, J.K., et al. (2016). Codon influence on protein expression in E . coli correlates with mRNA levels. Nature.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M. V., and Komar, A.A. (2016). Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. Mol. Cell *61*, 341–351.

Cao, D., and Parker, R. (2001). Computational modeling of eukaryotic mRNA turnover. RNA 1192–1212.

Caponigro, G., Muhlrad, D., and Parker, R.O.Y. (1993). A Small Segment of the M ATl Transcript Promotes mRNA Decay in Saccharomyces cerevisiae : a Stimulatory Role for Rare Codons. Mol. Cell. Biol. *13*, 5141–5148.

Chan, C.T.Y., Pang, Y.L.J., Deng, W., Babu, I.R., Dyavaiah, M., Begley, T.J., and Dedon, P.C. (2012). Reprogramming of tRNA modifications controls the oxidative stress response by codon-biased translation of proteins. Nat. Commun. *3*.

Charif, D., and Lobry, J.R. (2007). Seqin{R} 1.0-2: a contributed package to the {R} project for statistical computing devoted to biological sequences retrieval and analysis.

Charneski, C.A., and Hurst, L.D. (2013). Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. PLoS Biol. *11*.

Chen, Y.-H., and Coller, J. (2016). A Universal Code for mRNA Stability? Trends Genet. *xx*, 1–2.

Chen, C., Zhang, H., Broitman, S.L., Reiche, M., Farrell, I., Cooperman, B.S., and Goldman, Y.E. (2013). Dynamics of translation by single ribosomes through mRNA secondary structures. Nat. Struct. Mol. Biol. *20*, 582–588.

Cheng, J., Maier, K.C., Avsec, Z., Rus, P., and Gagneur, J. (2017). Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast. RNA.

Chu, D., and Haar, T. Von Der (2012). The architecture of eukaryotic translation. Nucleic Acids Res. *40*, 10098–10106.

Chu, D., Kazana, E., Singh, T., and Tuite, M.F. (2014). Translation elongation can control translation initiation on eukaryotic mRNAs. *33*, 21–34.

Coleclough, C., Kuhnt, L., and Lefkovitst, I. (1990). Regulation of mRNA abundance in activated T lymphocytes : Identification of mRNA species affected by the inhibition of protein synthesis. Proc Natl Acad Sci U S A *87*, 1753–1757.

Coller, J., and Parker, R. (2004). EUKARYOTIC mRNA DECAPPING. Annu. Rev. Biochem.

Cottrell, K.A., Chaudhari, H.G., Cohen, B.A., and Djuranovic, S. (2018). PTRE-seq reveals mechanism and interactions of RNA binding proteins and miRNAs. Nat. Commun. *9*, 1–13.

Cozen, A.E., Quartley, E., Holmes, A.D., Hrabeta-Robinson, E., Phizicky, E.M., and Lowe, T.M. (2015). ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments. Nat. Methods *12*, 879–884.

Crick, F. (1966). Codon-Anticodon Pairing: The Wobble Hypothesis. J. Mol. Biol. 548–555.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561–563.

Cuperus, J.T., Groves, B., Kuchina, A., Rosenberg, A.B., Jojic, N., Fields, S., Seelig, G., and Science, C. (2017). Deep learning of the regulatory grammar of yeast 5 ′ untranslated regions from 500 , 000 random sequences. Genome Res. 1–10.

Decker, C.J., and Parker, R. (2012). P-Bodies and Stress Granules : Possible Roles in the Control of Translation and mRNA Degradation. 1–16.

Dever, T.E., and Green, R. (2012). The Elongation, Termination, and Recycling Phases of Translation in Eukaryotes. Cold Spring Harb Perspect Biol.

Dever, T.E., Dinman, J.D., and Green, R. (2018). Translation Elongation and Recoding in Eukaryotes. Cold Spring Harb. Perspect. Biol. a032649.

Dittmar, K.A., Sørensen, M.A., Elf, J., Ehrenberg, M., and Pan, T. (2005). Selective charging of tRNA isoacceptors induced by amino-acid starvation. EMBO Rep. *6*, 151–157.

Dittmar, K.A., Goodenbour, J.M., and Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. PLoS Genet. *2*, 2107–2115.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dolken, L., Ruzsics, Z., Ra, B., Mages, R.G., Hoffmann, R., Dickinson, P., and Forster, T. (2008). High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. RNA 1959–1972.

Doma, M.K., and Parker, R. (2007). RNA Quality Control in Eukaryotes. Cell.

Dressaire, C., Picard, F., Redon, E., Loubière, P., Queinnec, I., Girbal, L., and Cocaign-Bousquet, M. (2013). Role of mRNA Stability during Bacterial Adaptation. PLoS One *8*.

Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell *134*, 341–352.

Duan, J., Shi, J., Ge, X., Dölken, L., Moy, W., He, D., Shi, S., Sanders, A.R., Ross, J., and Gejman, P. V. (2013). Genome-wide survey of interindividual differences of RNA stability in human lymphoblastoid cell lines. Sci. Rep. *3*, 3–7.

Duffy, E.E., Catherine, D., Kitchen, R.R., Mark, B., Simon, M.D., Duffy, E.E., Rutenberg-schoenberg, M., Stark, C.D., Kitchen, R.R., Gerstein, M.B., et al. (2015). Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry Technology Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry. Mol. Cell *59*, 858–866.

Edri, S., and Tuller, T. (2014). Quantifying the Effect of Ribosomal Density on mRNA Stability. PLoS One *9*.

Elkon, R., Zlotorynski, E., Zeller, K.I., and Agami, R. (2010). Major role for mRNA stability in shaping the kinetics of gene induction. BMC Genomics.

Encode Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Etten, J. Van, Schagat, T.L., Hrit, J., Weidmann, C.A., Brumbaugh, J., Coon, J.J., and

Goldstrohm, A.C. (2012). Human Pumilio Proteins Recruit Multiple Deadenylases to Efficiently Repress Messenger RNAs *. J. Biol. Chem. *287*, 36370–36383.

Fabian, M.R., and Sonenberg, N. (2012). The mechanics of miRNA-mediated gene silencing : a look under the hood of miRISC. Nat. Struct. Mol. Biol. *19*, 586–593.

Feng, L., and Niu, D.K. (2007). Relationship between mRNA stability and length: An old question with a new twist. Biochem. Genet. *45*, 131–137.

Fields, A.P., Rodriguez, E.H., Jovanovic, M., Stern-Ginossar, N., Haas, B.J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S.A., Ingolia, N.T., et al. (2015). A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. Mol. Cell *60*, 816–827.

Fonseca, B.D., Smith, E.M., Yelle, N., Alain, T., Bushell, M., and Pause, A. (2014). The ever-evolving role of mTOR in translation. Semin. Cell Dev. Biol. *36*, 102–112.

Friedman, R.C., Farh, K.K., Burge, C.B., and Bartel, D.P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 92–105.

Gamble, C.E., Brule, C.E., Dean, K.M., Fields, S., Grayhack, E.J., Gamble, C.E., Brule, C.E., Dean, K.M., Fields, S., and Grayhack, E.J. (2016). Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast Article Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast. Cell *166*, 1–12.

Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., and Futcher, B. (2014). Measurement of average decoding rates of the 61 sense codons in vivo. Elife *3*, 1–20.

Geisberg, J. V., Moqtaderi, Z., Fan, X., Ozsolak, F., and Struhl, K. (2014). Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. Cell *156*, 812–824.

Gelperin, D.M., White, M. a, Wilkinson, M.L., Kon, Y., Kung, L. a, Wise, K.J., Lopez-Hoyo, N., Jiang, L., Piccirillo, S., Yu, H., et al. (2005). Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. Genes Dev. *19*, 2816–2826.

Gerber, P., Herschlag, D., Brown, P.O., Hogan, D.J., and Riordan, D.P. (2008). Diverse RNA-

Binding Proteins Interact with Functionally Related Sets of RNAs , Suggesting an Extensive Regulatory System. PLoS Biol. *6*.

Ghosh, S., and Jacobson, A. (2010). RNA decay modulates gene expression and controls its fidelity. Wiley Interdiscip Rev RNA.

Gingold, H., Tehler, D., Christoffersen, N.R., Nielsen, M.M., Asmar, F., Kooistra, S.M., Christophersen, N.S., Christensen, L.L., Borre, M., Sørensen, K.D., et al. (2014). A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation. Cell *158*, 1281–1292.

Gogakos, T., Brown, M., Garzia, A., Meyer, C., Hafner, M., and Tuschl, T. (2017). Characterizing Expression and Processing of Precursor and Mature Human tRNAs by Hydro-tRNAseq and PAR-CLIP. Cell Rep. *20*, 1463–1475.

Goodarzi, H., Nguyen, H.C.B., Zhang, S., Dill, B.D., Molina, H., and Tavazoie, S.F. (2016). Modulated expression of specific tRNAs drives gene expression and cancer progression. Cell *165*, 1416–1427.

Goodman, D.B., Church, G.M., and Kosuri, S. (2013). Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. Science (80-.). 475–480.

Grant, I.M., Balcha, D., Hao, T., Shen, Y., Trivedi, P., Patrushev, I., Fortriede, J.D., Karpinka, J.B., Liu, L., Zorn, A.M., et al. (2015). The Xenopus ORFeome: A resource that enables functional genomics. Dev. Biol. *408*, 345–357.

Greenberg, J.R. (1972). High Stability of Messenger RNA in Growing Cultured Cells. Nature.

Halbeisen, R.E., Galgano, A., Scherrer, T., and Gerber, A.P. (2008). Review Post-transcriptional gene regulation : From genome-wide studies to principles. *65*, 798–813.

Hanson, G., and Coller, J. (2017). Codon optimality, bias and usage in translation and mRNA decay. Nat. Rev. Mol. Cell Biol.

Hanson, G., Alhusaini, N., Morris, N., Sweet, T., and Coller, J. (2018). Translation elongation and mRNA stability are coupled through the ribosomal A- site. BioRXiv Prepr.

Harigaya, Y., and Parker, R. (2016). Analysis of the association between codon optimality and

mRNA stability in Schizosaccharomyces pombe. BMC Genomics 1–16.

Harigaya, Y., and Parker, R. (2017). The link between adjacent codon pairs and mRNA stability. BMC Genomics *18*, 364.

Harrell, F., and Dupont, C. (2016). Hmisc: Harrell Miscellaneous. R package.

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.

Herzog, V.A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T.R., Wlotzka, W., von Haeseler, A., Zuber, J., and Ameres, S.L. (2017). Thiol-linked alkylation of RNA to assess expression dynamics. Nat. Methods.

Hinnebusch, A.G., Ivanov, I.P., and Sonenberg, N. (2016). Translational control by 5 ′ -untranslated regions of eukaryotic mRNAs. Science (80-.). *352*, 1413–1416.

Hoekema, A., Kastelein, R.O.B.A., Vasser, M., and Boer, H.A.D.E. (1987). Codon Replacement in the PGKI Gene of Saccharomyces cerevisiae : Experimental Approach To Study the Role of Biased Codon Usage in Gene Expression. Mol. Cell. Biol. *7*, 2914–2924.

Holt, C.E., and Bullock, S.L. (2009). Subcellular mRNA localization in animal cells and why it matters. Science *326*, 1212–1216.

Humphreys, T. (1969). Efficiency of translation of messenger-RNA before and after fertilization in sea urchins. Dev. Biol. *20*, 435–458.

Hussmann, J.A., Patchett, S., Johnson, A., Sawyer, S., and Press, W.H. (2015). Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. PLoS Genet. *11*, 1–25.

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

Inoue, F., and Ahituv, N. (2016). Decoding enhancers using massively parallel reporter assays Fumitaka. Genomics *106*, 87–92.

Ishimura, R., Nagy, G., Dotu, I., Zhou, H., Yang, X.L., Schimmel, P., Senju, S., Nishimura, Y.,

Chuang, J.H., and Ackerman, S.L. (2014). Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. Science (80-.). *345*, 455–459.

Isken, O., and Maquat, L.E. (2007). Quality control of eukaryotic mRNA : safeguarding cells from abnormal mRNA function. Genes Dev.

Iwasaki, S., Floor, S.N., and Ingolia, N.T. (2016). Rocaglates convert DEAD-box protein eIF4A into a sequence-selective translational repressor. Nature *534*, 558–561.

Jackson, R.J., Hellen, C.U.T., and Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. Nat. Rev. Mol. Cell Biol. *11*, 113–127.

Jain, P., Boso, G., Langer, S., Soonthornvacharin, S., De Jesus, P.D., Nguyen, Q., Olivieri, K.C., Portillo, A.J., Yoh, S.M., Pache, L., et al. (2018). Large-Scale Arrayed Analysis of Protein Degradation Reveals Cellular Targets for HIV-1 Vpu. Cell Rep. *22*, 2455–2468.

Jeacock, L., Faria, J., and Horn, D. (2018a). Codon usage bias controls mRNA and protein abundance in trypanosomatids. Elife *7*, 1–20.

Jeacock, L., Faria, J., and Horn, D. (2018b). Codon usage bias controls mRNA and protein abundance in 1 trypanosomatids 2 3. 1–20.

Johansson, M., Ieong, K.-W., Trobro, S., Strazewski, P., Aqvist, J., Pavlov, M.Y., and Ehrenberg, M. (2011). pH-sensitivity of the ribosomal peptidyl transfer reaction dependent on the identity of the A-site aminoacyl-tRNA. Proc. Natl. Acad. Sci. *108*, 79–84.

Jordanova, A., Irobi, J., Thomas, F.P., Van Dijck, P., Meerschaert, K., Dewil, M., Dierick, I., Jacobs, A., De Vriendt, E., Guergueltcheva, V., et al. (2006). Disrupted function and axonal distribution of mutant tyrosyl-tRNA synthetase in dominant intermediate Charcot-Marie-Tooth neuropathy. Nat. Genet. *38*, 197–202.

Karaca, E., Weitzer, S., Pehlivan, D., Shiraishi, H., Gogakos, T., Hanada, T., Jhangiani, S.N., Wiszniewski, W., Withers, M., Campbell, I.M., et al. (2014). Human CLP1 mutations alter tRNA biogenesis, Affecting both peripheral and central nervous system function. Cell *157*, 636–650.

Keene, J.D. (2007). RNA regulons : coordination of post-transcriptional events. Nat. Rev. *8*,

533–543.

Keene, J.D., and Tenenbaum, S.A. (2002). Eukaryotic mRNPs May Represent Posttranscriptional Operons. Mol. Cell *9*, 1161–1167.

Kent, D.K.A.S.H.T.S.F.K.M.R.C.W.S.D.H.W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res. *32*, 493D–496.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. Nature *467*, 103–107.

Kong, J., and Lasko, P. (2012). Translational control in cellular and developmental processes. Nat. Rev. Genet. *13*, 383–394.

Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger rNAS. Nucleic Acids Res. *15*, 8125–8148.

Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009). Coding-Sequence Determinants of Gene Expression in Escherichia coli. Science (80-.). 255–259.

Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

Lagrandeur, T., and Parker, R.O.Y. (1999). The cis acting sequences responsible for the differential decay of the unstable MFA2 and stable PGK1 transcripts in yeast include the context of the translational start codon. RNA 420–433.

Lamesch, P., Li, N., Milstein, S., Fan, C., Hao, T., Szabo, G., Hu, Z., Venkatesan, K., Bethel, G., Martin, P., et al. (2007). hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. Genomics *89*, 307–315.

Lareau, L.F., Hite, D.H., Hogan, G.J., and Brown, P.O. (2014). Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. 1–16.

Lee, E.K., and Gorospe, M. (2011). Coding region , the neglected post-transcriptional code. RNA Biol. *8*, 44–48.

Lee, J.W., Beebe, K., Nangle, L.A., Jang, J., Longo-Guess, C.M., Cook, S.A., Davisson, M.T., Sundberg, J.P., Schimmel, P., and Ackerman, S.L. (2006). Editing-defective tRNA synthetase

causes protein misfolding and neurodegeneration. Nature *443*, 50–55.

Lemm, I., and Ross, J. (2002). Regulation of c-myc mRNA Decay by Translational Pausing in a Coding Region Instability Determinant. Mol. Cell. Biol. *22*, 3959–3969.

Lievens, S., Van der Heyden, J., Masschaele, D., De Ceuninck, L., Petta, I., Gupta, S., De Puysseleyr, V., Vauthier, V., Lemmens, I., De Clercq, D.J.H., et al. (2016). Proteome-scale Binary Interactomics in Human Cells. Mol. Cell. Proteomics *15*, 3624–3639.

Lorenz, R., Bernhart, S.H., zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). {ViennaRNA} Package 2.0. Algorithms Mol. Biol. *6*, 26.

Lu, J., and Deutsch, C. (2008). Electrostatics in the Ribosomal Tunnel Modulate Chain Elongation Rates. J. Mol. Biol. *384*, 73–86.

Lugowski, A., Nicholson, B., and Rissland, O.S. (2017). Determining mRNA half-lives on a transcriptome-wide scale. Methods 4–12.

Lugowski, A., Nicholson, B., and Rissland, O.S. (2018). DRUID: A pipeline for transcriptome-wide measurements of mRNA stability. Rna rna.062877.117.

Mattijssen, S., Arimbasseri, A.G., Iben, J.R., Gaidamakov, S., Lee, J., Hafner, M., and Maraia, R.J. (2017). LARP4 mRNA codon-tRNA match contributes to LARP4 activity for ribosomal protein mRNA poly(A) tail length protection. Elife *6*, 1–33.

Mayr, C., and Bartel, D.P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell *138*, 673–684.

Meignin, C., and Davis, I. (2010). Transmitting the message: intracellular mRNA localization. Curr. Opin. Cell Biol. *22*, 112–119.

Melnikov, A., Zhang, X., Rogov, P., Wang, L., and Mikkelsen, T.S. (2014). Massively Parallel Reporter Assays in Cultured Mammalian Cells. J. Vis. Exp. 1–8.

Mishima, Y., and Tomari, Y. (2016). Codon Usage and 3′ UTR Length Determine Maternal mRNA Stability in Zebrafish. Mol. Cell *61*, 874–885.

Moerke, N.J., Aktas, H., Chen, H., Cantel, S., Reibarkh, M.Y., Fahmy, A., Gross, J.D.D., Degterev, A., Yuan, J., Chorev, M., et al. (2007). Small-Molecule Inhibition of the Interaction

between the Translation Initiation Factors eIF4E and eIF4G. Cell *128*, 257–267.

Mogno, I., Kwasnieski, J.C., and Cohen, B. a (2013). Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. Genome Res. 1908–1915.

Morris, A.R., Mukherjee, N., and Keene, J.D. (2010). Systematic analysis of posttranscriptional gene expression. WIREs Syst. Biol. Med. 9–12.

Munchel, S.E., Shultzaberger, R.K., Takizawa, N., Weis, K., and Matera, A.G. (2011). Dynamic profiling of mRNA turnover reveals gene-specific and system-wide regulation of mRNA decay. Mol. Biol. Cell *22*, 2787–2795.

Nam, J.W., Rissland, O.S., Koppstein, D., Abreu-Goodger, C., Jan, C., Agarwal, V., Yildirim, M.A., Rodriguez, A., and Bartel, D.P. (2014). Global analyses of the effect of different cellular contexts on microRNA targeting. Mol. Cell *53*, 1031–1043.

Nascimento, J. de F., Kelly, S., Sunter, J., and Carrington, M. (2018). Codon choice directs constitutive mRNA levels in trypanosomes. Elife *7*, 1–26.

Neymotin, B., Athanasiadou, R., and Gresham, D. (2014). Determination of in vivo RNA kinetics using RATE-seq. RNA *20*, 1645–1652.

Neymotin, B., Ettore, V., and Gresham, D. (2016). Multiple Transcript Properties Related to Translation Affect mRNA Degradation Rates in Saccharomyces cerevisiae. G3: Genes|Genomes|Genetics *6*, 3475–3483.

Nikolov, E.N., and Dabeva, M.D. (1985). Re-utilization of pyrimidine nucleotides during rat liver regeneration. Biochem. J. *228*, 27–33.

Nirenberg, B.Y.M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., Rottmant, F., and Neal, C.O. (1965). RNA codewords and protein synthesis, vii. On the general nature of the RNA code. Proc Natl Acad Sci U S A *53*, 1161–1168.

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.

Oikonomou, P., Goodarzi, H., and Tavazoie, S. (2014). Systematic identification of regulatory

elements in conserved 3’ UTRs of human transcripts. Cell Rep. *7*, 281–292.

Pagès, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2017). Biostrings: Efficient manipulation of biological strings. R Packag. Version 2.46.0.

Pavlov, M.Y., Watts, R.E., Tan, Z., Cornish, V.W., Ehrenberg, M., and Forster, A.C. (2009). Slow peptide bond formation by proline and other N-alkylamino acids in translation. Proc. Natl. Acad. Sci. *106*, 50–54.

Pelechano, V., and Alepuz, P. (2017). EIF5A facilitates translation termination globally and promotes the elongation of many non polyproline-specific tripeptide sequences. Nucleic Acids Res. *45*, 7326–7338.

Peltz, S.W., Donahue, J.L., and Jacobson, A. (1992). A Mutation in the tRNA Nucleotidyltransferase Gene Promotes Stabilization of mRNAs in Saccharomyces cerevisiae. Mol. Cell. Biol. *12*, 5778–5784.

Pérez-Ortín, J.E., Alepuz, P., Chávez, S., and Choder, M. (2013). Eukaryotic mRNA decay: Methodologies, pathways, and links to other stages of gene expression. J. Mol. Biol. *425*, 3750–3775.

Plotkin, J.B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. *12*, 32–42.

Plotkin, J.B., Robins, H., and Levine, A.J. (2004). Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci USA *101*, 12588–12591.

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R., et al. (2015). Codon Optimality Is a Major Determinant of mRNA Stability. Cell *160*, 1111–1124.

Qin, Y., Yao, J.U.N., Wu, D.C., Nottingham, R.M., Mohr, S., Hunicke-smith, S., and Lambowitz, A.M. (2015). High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. Rna 1–18.

Rabani, M., Levin, J.Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., et al. (2011). Metabolic labeling of RNA uncovers

principles of RNA production and degradation dynamics in mammalian cells. Nat. Biotechnol. *29*, 436–442.

Radhakrishnan, A., and Green, R. (2016). Connections underlying translation and mRNA stability. J. Mol. Biol. *428*, 3558–3564.

Radhakrishnan, A., Chen, Y.-H., Martin, S., Alhusaini, N., Green, R., Coller, J., Anderson, J.S., Parker, R.P., Barbee, S.A., Estes, P.S., et al. (2016). The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. Cell *0*, 1497–1506.

Ramirez, C.V., Vilela, C., Berthelot, K., and Mccarthy, J.E.G. (2002). Modulation of Eukaryotic mRNA Stability via the Cap-binding Translation Complex eIF4F. J. Mol. Biol. *2836*, 951–962.

Reboul, J., Vaglio, P., Rual, J.F., Lamesch, P., Martinez, M., Armstrong, C.M., Li, S., Jacotot, L., Bertin, N., Janky, R., et al. (2003). C. elegans ORFeome version 1.1: Experimental verification of the genome annotation and resource for proteomescale protein expression. Nat. Genet. *34*, 35–41.

Rissland, O.S. (2016). The organization and regulation of mRNA-protein complexes. Wiley Interdiscip. Rev. RNA.

Rissland, O.S., and Norbury, C.J. (2009). Decapping is preceded by 3′ uridylation in a novel pathway of bulk mRNA turnover. Nat. Struct. Mol. Biol. *16*, 616–623.

Rodnina, M. V. (2016). The ribosome in action: Tuning of translational efficiency and protein folding. Protein Sci. *25*, 1390–1406.

Ross, J. (1995). mRNA Stability in Mammalian Cells. Microbiol. Rev. *59*, 423–450.

Roy, B., and Jacobson, A. (2013). The intimate relationships of mRNA decay and translation. Trends Genet. *29*, 691–699.

Rual, J.-F., Hirozane-kishikawa, T., Hao, T., Bertin, N., Li, S., Dricot, A., Li, N., Rosenberg, J., Lamesch, P., Vidalain, P., et al. (2004). Human ORFeome Version 1 . 1 : A Platform for Reverse Proteomics. Genome Res. *1*, 2128–2135.

S, A. (2010). FastQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Sabi, R., and Tuller, T. (2015). A comparative genomics study on the effect of individual amino

acids on ribosome stalling. BMC Genomics *16*, 1–12.

Saikia, M., Wang, X., Mao, Y., Wan, J.I., Pan, T.A.O., and Qian, S. (2016). Codon optimality controls differential mRNA translation during amino acid starvation. Rna 1–9.

Saunders, R., and Deane, C.M. (2010). Synonymous codon usage influences the local protein structure observed. *38*, 6719–6728.

Schiavi, S.C., Wellington, C.L., Shyus, A., Chens, A., Greenberg, M.E., and Belascos, J.G. (1994). Multiple Elements in the c-fos Protein-coding Region Facilitate mRNA Deadenylation and Decay by a Mechanism Coupled to Translation. J. Biol. Chem. *269*, 3441–3448.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: An open-source platform for biological-image analysis. Nat. Methods *9*, 676–682.

Schmidt, E.K., Clavarino, G., Ceppi, M., and Pierre, P. (2009). SUnSET, a nonradioactive method to monitor protein synthesis. Nat. Methods *6*, 275–277.

Schnall-Levin, M., Rissland, O.S., Johnston, W.K., Perrimon, N., Bartel, D.P., and Berger, B. (2011). Unusually effective microRNA targeting within repeat-rich coding regions of mammalian mRNAs. Genome Res. *21*, 1395–1403.

Schoenberg, D.R., and Maquat, L.E. (2012). Regulation of cytoplasmic mRNA decay. Nat. Rev. Genet. *13*, 448–448.

Schofield, J.A., Duffy, E.E., Kiefer, L., Sullivan, M.C., and Simon, M.D. (2018). TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. Nat. Methods *15*, 221–225.

Schwartz, D.C., and Parker, R. (1999). Mutations in Translation Initiation Factors Lead to Increased Rates of Deadenylation and Decapping of mRNAs in Saccharomyces cerevisiae. Mol. Cell. Biol. *19*, 5247–5256.

Schwartz, D.C., and Parker, R. (2000). mRNA Decapping in Yeast Requires Dissociation of the Cap Binding Protein , Eukaryotic Translation Initiation Factor 4E. Mol. Cell. Biol. *20*, 7933–7942.

Sekiyama, N., Arthanari, H., Papadopoulos, E., Rodriguez-Mias, R.A., Wagner, G., and Léger-Abraham, M. (2015). Molecular mechanism of the dual activity of 4EGI-1: Dissociating eIF4G from eIF4E but stabilizing the binding of unphosphorylated 4E-BP1. Proc. Natl. Acad. Sci. U. S. A. *112*, E4036-45.

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

Sharp, P.M., and Li, W.-H. (1985). The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential implicaitons. *13*, 3021–3030.

Shen, S.Q., Myers, C.A., Hughes, A.E.O., Byrne, L.C., Flannery, J.G., and Corbo, J.C. (2016). Massively parallel cis -regulatory analysis in the mammalian central nervous system. Genome Res. 238–255.

Shigematsu, M., Honda, S., Loher, P., Telonis, A.G., Rigoutsos, I., and Kirino, Y. (2017). YAMAT-seq: An efficient method for high-throughput sequencing of mature transfer RNAs. Nucleic Acids Res. *45*, e70.

Shoemaker, C.J., and Green, R. (2012). Translation drives mRNA quality control. Nat. Struct. Mol. Biol. *19*, 594–601.

Smibert, C.A., Wilson, J.E., Kerr, K., and Macdonald, P.M. (1996). smaug protein represses translation of unlocalized nanos mRNA in the Drosophila embryo. Genes Dev. 2600–2609.

Son, H., Kang, H., Kim, H.S., and Kim, S. (2017). Somatic mutation driven codon transition bias in human cancer. Sci. Rep. *7*, 1–11.

Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. Cell *136*, 731–745.

Subramaniam, A.R., Zid, B.M., and O'Shea, E.K. (2014). An integrated approach reveals regulatory controls on bacterial translation elongation. Cell *159*, 1200–1211.

Sun, M., Schulz, D., Pirkl, N., Etzold, S., Larivie, L., Maier, K.C., Seizl, M., Tresch, A., and Cramer, P. (2012). Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. Genome Res. 1350–1359.

Sweet, T., Kovalak, C., and Coller, J. (2012). The DEAD-box protein Dhh1 promotes decapping by slowing ribosome movement. PLoS Biol. *10*, e1001342.

Szostak, E., and Gebauer, F. (2013). Translational control by 3'-UTR-binding proteins. Brief. Funct. Genomics *12*, 58–65.

Tadros, W., Goldman, A.L., Babak, T., Menzies, F., Vardy, L., Orr-weaver, T., Hughes, T.R., Westwood, J.T., Smibert, C.A., and Lipshitz, H.D. (2007). SMAUG Is a Major Regulator of Maternal mRNA Destabilization in Drosophila and Its Translation Is Activated by the PAN GU Kinase. Dev. Cell 143–155.

Taipale, M., Krykbaeva, I., Koeva, M., Kayatekin, C., Westover, K.D., Karras, G.I., and Lindquist, S. (2012). Quantitative analysis of Hsp90-client interactions reveals principles of substrate recognition. Cell *150*, 987–1001.

Tani, H., Mizutani, R., Salam, K.A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y., and Akimitsu, N. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res. 947–956.

Tanner, D.R., Cariello, D.A., Woolstenhulme, C.J., Broadbent, M.A., and Buskirk, A.R. (2009). Genetic identification of nascent peptides that induce ribosome stalling. J. Biol. Chem. *284*, 34809–34818.

Team, R.C. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

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

Tunney, R.J., McGlincy, N.J., Graham, M.E., Naddaf, N., Pachter, L., and Lareau, L. (2018). Accurate design of translational output by a neural network model of ribosome distribution. Nat. Struct. Mol. Biol. 201517.

Vainberg Slutskin, I., Weingarten-Gabbay, S., Nir, R., Weinberger, A., and Segal, E. (2018). Unraveling the determinants of microRNA mediated regulation using a massively parallel reporter assay. Nat. Commun. *9*.

Vester, A., Velez-Ruiz, G., McLaughlin, H.M., Nisc Comparative Sequencing Program, Lupski, J.R., Talbot, K., Vance, J.M., Züchner, S., Roda, R.H., Fischbeck, K.H., et al. (2013). A Loss-of-Function Variant in the Human Histidyl-tRNA Synthetase (HARS) Gene is Neurotoxic In Vivo. Hum. Mutat. *34*, 191–199.

Walhout, A.J.M., Temple, G.F., Brasch, M.A., Hartley, J.L., Lorson, M.A., van den Heuvel, S., and Vidal, M. (2000). GATEWAY recombinational cloning: Application to the cloning of large numbers of open reading frames or ORFeomes. Methods Enzymol. *328*, 575-IN7.

Weinberg, D.E., Shah, P., Eichhorn, S.W., Hussmann, J.A., Plotkin, J.B., and Bartel, D.P. (2016). Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. Cell Rep. *14*, 1787–1799.

White, M.A. (2016). Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. Genomics *106*, 165–170.

Wickham, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Wickham, H., Francois, R., Henry, L., and Müller, K. (2017). A Grammar of Data Manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr.

Wissink, E.M., Fogarty, E.A., and Grimson, A. (2016). High-throughput discovery of post-transcriptional cis-regulatory elements. BMC Genomics *17*, 177.

Wolter, J.M., Kotagama, K., Babb, C.S., and Mangone, M. (2015). Detection of miRNA Targets in High-throughput Using the 3'LIFE Assay. J. Vis. Exp. 1–9.

Wu, X., and Bartel, D.P. (2017). Widespread Influence of 3′-End Structures on Mammalian mRNA Processing and Stability. Cell *169*, 905–917.e11.

Wu, X., and Brewer, G. (2012). The regulation of mRNA stability in mammalian cells: 2.0. Gene *500*, 10–21.

Yamazaki, S., and Takeshige, K. (2008). Protein synthesis inhibitors enhance the expression of mRNAs for early inducible inflammatory genes via mRNA stabilization. Biochim. Biophys. Acta *1779*, 108–114.

Yan, X., Hoek, T.A., Vale, R.D., and Tanenbaum, M.E. (2016). Dynamics of Translation of

Single mRNA Molecules in Vivo. Cell *165*, 976–989.

Yang, X., Boehm, J.S., Yang, X., Salehi-Ashtiani, K., Hao, T., Shen, Y., Lubonja, R., Thomas, S.R., Alkan, O., Bhimdi, T., et al. (2011). A public genome-scale lentiviral expression library of human ORFs. Nat. Methods *8*, 659–661.

Yartseva, V., Takacs, C.M., Vejnar, C.E., Lee, M.T., and Giraldez, A.J. (2017). RESA identifies mRNA-regulatory sequences at high resolution. Nat. Methods *14*, 201–207.

Yu, C., Dang, Y., Zhao, F., Matthew, S., Yu, C., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M.S., et al. (2015). Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Article Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. Mol. Cell *59*, 744–754.

Zhang, F., Saha, S., Shabalina, S.A., and Kashina, A. (2010). Differential Arginylation of Actin Sequence – Dependent Degradation. Science (80-.). *1065*, 1534–1537.

Zhao, W., Pollack, J.L., Blagev, D.P., Zaitlen, N., McManus, M.T., and Erle, D.J. (2014). Massively parallel functional annotation of 3’ untranslated regions. Nat. Biotechnol. *32*, 387–391.

Zheng, G., Qin, Y., Clark, W.C., Dai, Q., Yi, C., He, C., Lambowitz, A.M., and Pan, T. (2015). Efficient and quantitative high-throughput tRNA sequencing. Nat. Methods *12*, 835–837.

Zhong, Q., Pevzner, S.J., Hao, T., Wang, Y., Mosca, R., Menche, J., Taipale, M., Tasan, M., Fan, C., Yang, X., et al. (2016). An inter-species protein-protein interaction network across vast evolutionary distance. Mol. Syst. Biol. *12*, 865–865.

Zhou, M., Wang, T., Fu, J., Xiao, G., and Liu, Y. (2015). Nonoptimal codon usage influences protein structure in intrinsically disordered regions. Mol. Microbiol. *97*, 974–987.

Zhou, T., Weems, M., and Wilke, C.O. (2007). Translationally Optimal Codons Associate with Structurally Sensitive Sites in Proteins.

Zufferey, R., Donello, J.E., Trono, D., and Hope, T.J. (1999). Woodchuck hepatitis virus posttranscriptional regulatory element enhances expression of transgenes delivered by retroviral vectors. J. Virol. *73*, 2886–2892.