A Nascent Peptide Code for Translational Control of mRNA Stability in Human Cells

Phillip Cannon Burke

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctorate of Philosophy, University of Washington

2022

Reading Committee:

Jesse Bloom, Chair

Rasi Subramaniam

Adam Geballe

Program Authorized to Offer Degree:

Microbiology

University of Washington

Abstract

A Nascent Peptide Code for Translational Control of mRNA Stability in Human Cells

Phillip Cannon Burke

Chair of the Supervisory Committee:

Professor Jesse Bloom

Department of Microbiology

Stability of eukaryotic mRNAs is associated with their codon, amino acid, and GC content. Yet, coding sequence motifs that predictably alter mRNA stability in human cells remain poorly defined. Here, we develop a massively parallel assay to measure mRNA effects of thousands of synthetic and endogenous coding sequence motifs in human cells. We identify several families of simple dipeptide repeats whose translation triggers mRNA destabilization. Rather than individual amino acids, specific combinations of bulky and positively charged amino acids are critical for the destabilizing effects of dipeptide repeats. Remarkably, dipeptide sequences that form extended β strands in silico and in vitro slowdown ribosomes and reduce mRNA levels in vivo. The resulting nascent peptide code underlies the mRNA effects of hundreds of endogenous peptide sequences in the human proteome. Our work suggests an intrinsic role for the ribosome as a selectivity filter against the synthesis of bulky and aggregation-prone peptides.

Contents

List of Figures and Tables

Chapter 2

Chapter 3

Acknowledgements

Thank you to my brilliant wife, Kaitlyn LaCourse - you are my inspiration, and I never would have come this far without you. Thanks to my parents, Mary Ellen Cannon and Tom Burke, for all your support in my education (and in life!)

Thank you to both the Bloom and Subramaniam labs for providing such a wonderful and collaborative scientific environment to do research in. And finally, thank you Rasi for your mentorship these last six years.

Chapter 1. Introduction

I. Overview of eukaryotic gene expression

"Life is the mode of action of proteins." — Friedrich Engels[1]

Proteins are macromolecules made up of chains of amino acids that carry out diverse functions within our cells. It is difficult to overstate how important proteins are to living beings - they maintain our cellular structure, perform the biochemical reactions required for metabolism, and facilitate both intercellular signaling pathways that determine cell behavior and extracellular communication required for multicellular life. Our cells function by producing the correct proteins, in the appropriate amounts, at exactly the right time. It stands to reason that organisms across all kingdoms of life have evolved sophisticated mechanisms for controlling protein production. In this dissertation I will focus on the regulation of eukaryotic protein synthesis, though many of these regulatory mechanisms also exist or have functional parallels in bacteria and archaea[2–4].

Genetic information flows from our DNA, which is transcribed to messenger RNA (mRNA), which is then translated into proteins. This overarching control system for protein production is considered the central dogma of molecular biology[5], and each stage contributes to the protein landscape within our cells. The human genome provides coding potential for ~20,000 annotated proteins, and undergoes evolutionary selection and genetic drift in both coding and non-coding regions[6–9]. While having the appropriate gene is a prerequisite for producing a given protein, making the choice to express that gene is of equal importance for protein synthesis. Accordingly, transcription of our genes into mRNAs is highly regulated. In eukaryotes, DNA is located in the nucleus of the cell, mostly wrapped around histone proteins in a protein-DNA complex known as chromatin. To begin transcription, RNA polymerase II (Pol II) binds to a promoter region upstream of a given gene and synthesizes an RNA message complementary to the DNA sequence. RNA Pol II binding is mediated by various transcription factors, and can be modified by protein cofactor binding and

epigenetic modifications of the nearby chromatin[10,11].

During transcription, the eukaryotic pre-mRNA transcript undergoes several processing steps to facilitate its subsequent translation. First, the 5' end of the pre-RNA is capped with an N7-methylated guanosine linked to the first nucleotide of the RNA via a reverse 5′ to 5′ triphosphate linkage, followed by methylation of the 2'O ribose of the first nucleotide[12,13]. This is often followed by additional cap-adjacent modifications such as 2'O methylation of the second nucleotide, N6 methylation of the first transcribed adenosine nucleotide in the mRNA, and likely other modifications as well[14]. Most eukaryotic genes are interspersed with non-coding intron sequences, which are co-transcriptionally spliced out of the transcript at intron-exon junctions (defined by specific DNA sequence motifs), leaving behind a protein complex at the spliced exon-exon site called the exon junction complex (EJC)[15]. The pre-mRNA is then cleaved near the 3' end and a tail of adenosine residues, usually around 200 nucleotides in length, is added to the 3' end by polyadenylation machinery proteins. The processed mRNA is then exported from the nuclease and, in the process, the nuclear cap binding proteins on the 5' cap are replaced by eukaryotic initiation factor (eIF) proteins which form the eIF4F cap-binding complex[16–18]. Any errors in these processing steps can interfere with mRNA export, resulting in the mRNA signal being untranslated[12,19]. In addition, the 5' cap, correctly spliced coding sequence (CDS), and poly-A tail are all elements of the mRNA that are critical for eukaryotic translation.

Once a processed mRNA has been exported from the nucleus to the cytosol it can be translated into protein by the ribosome. The ribosome is a large macromolecular complex, formed by a small and large subunit, each composed of dozens of proteins held together by scaffolds of ribosomal RNA (rRNA)[20]. In eukaryotes, these subunits are referred to as the 40S and 60S subunits respectively, based on their sedimentation coefficients — combined, these subunits are referred to as the 80S monosome (one ribosome). During translation, mRNA is threaded through a channel between the two ribosome subunits; the 40S subunit reads out the mRNA sequence as triplet nucleotide

codons which correspond to transfer RNAs (tRNAs) charged with specific amino acids, and the 60S subunit accepts these tRNAs and catalyzes peptide bond formation between their amino acids to extend the growing peptide chain. The mammalian cell has over 10 million ribosomes and translation as a whole is the most energy consumptive process in cells[21–23]. Accordingly, translation is highly regulated to allow fine-tuned control of protein production, allowing cells to maintain cellular homeostasis and rapidly respond to environmental stimuli[24–27]. Translation also serves as a quality control checkpoint, with the ribosome acting as both a scaffold for quality control factors, and as sensor of faulty mRNAs itself[28].

Finally, both mRNA and protein stability are determinants of protein levels, and are influenced by factors such as subcellular localization, post-transcriptional or post-translational modifications, and cell state[29]. My research focuses on the interplay between translation, quality control, and mRNA stability, and I discuss these concepts in greater depth below.

II. mRNA translation

Initiation and the 5' UTR

mRNA translation is a cyclical process that can be divided into three general stages: ribosome scanning and initiation, mRNA decoding and nascent peptide elongation, and termination and ribosome recycling. Translation initiation occurs in a series of steps where the ribosome identifies an AUG start codon and decodes this using a methionyl tRNA (transfer RNA) specialized for initiation, termed Met-tRNAi. The start codon is identified by a scanning mechanism wherein the 43S pre-initiation complex (PIC), a 40S small ribosomal subunit preloaded with the GTP-eIF2-Met-tRNAi ternary complex and several other eIF proteins, is recruited to the capped 5' end of mRNA by the eIF4F cap binding complex. The PIC then scans down the mRNA, inspecting successive nucleotide triplets as for complementarity to the anticodon of Met-tRNAi. Once the PIC encounters an AUG start codon, it halts and irreversibly hydrolyses the GTP of the GTP-eIF2-Met-tRNAi ternary

complex, resulting in release of GDP-eIF2 and other eIFs[24,30]. The large 60S subunit then joins to form an 80S initiation complex, ready to accept the next aminoacyl-tRNA and begin peptide bond synthesis[24,30].

Initiation is normally the rate limiting step for transition, thus this step is highly regulated and is a major determinant of protein output from a given mRNA.[cite: 24,31] In addition to the AUG start codon, the nucleotides surrounding the start codon are critical determinants of initiation efficiency[32,33]. mRNA secondary structures in the 5' untranslated region (5' UTR) influence ribosome scanning and initiation[34,35]. Cell state also affects initiation differently for different mRNAs. For example, some mRNAs have pyrimidine-rich 5' leader sequences known as 5'TOPs, which cause these mRNAs to undergo differential initiation in response to multiple trans-acting pathways that sense nutrient starvation[36,37]. Another canonical example of stress signaling differentially impacting initiation is for the ATF4 transcript. During multiple types of cellular stress, the alpha subunit of eIF2 becomes phosphorylated, preventing efficient formation of the GTP-eIF2-Met-tRNAi ternary complex and resulting in globally reduced translation initiation rate. While this reduces expression for most proteins, the ATF4 mRNA and a subset of similar mRNAs have complex 5' UTRs that contain multiple open reading frames; reduced initiation efficiency causes the ribosome to skip past one or more upstream open reading frames, increasing the chance that it initiates on the correct open reading frame for protein expression[38,39].

Elongation and the protein coding sequence

Ribosomes contain three tRNA-binding sites called the aminoacyl (A), peptidyl (P) and exit (E) sites. Translation initiation results in an elongation-competent 80S ribosome, with the initiator Met-tRNA base-paired with the AUG start codon in the ribosomal P-site[40]. At this stage, the second codon of the mRNA, positioned in the A-site, can be paired with the cognate anticodon of an aminoacyl-tRNA (a tRNA charged with an amino acid)[42]. Codon-anticodon matching at the A-site in the 40S subunit stimulates the peptidyl transferase reaction performed in the 60S subunit, resulting

in peptide bond synthesis between the Met-tRNAi and the incoming aminoacyl-tRNA, and leaving a deacylated Met-tRNAi (an uncharged tRNA with no amino acid) in the P-site. The ribosome then ratchets the mRNA by one codon, shifting the peptidyl-tRNA (a peptide-bound tRNA) to the P-site, and the deacylated tRNA to the E-site, which frees up the A-site for the next incoming cognate aminoacyl-tRNA[43]. The next cognate aminoacyl-tRNA is then accepted at the A-site, the deacylated tRNA is cleared from the E-site, and the cycle of peptide bond synthesis and ribosome translocation repeats[44]. As elongation progresses the growing nascent poly-peptide chain extends through a channel in the ribosome known as the exit tunnel and ultimately out into the cytosol, where the nascent peptide can then undergo cotranslational modifications and complex folding, prior to being released as a fully synthesized protein.

Both peptide bond synthesis and ribosome translation reactions are facilitated by a multitude of eukaryotic elongation factors (eEFs), and require large amounts of GTP, making elongation by far the most energy consuming step of translation[43]. Thus, it stands to reason that this step is also highly regulated[40,45,46]. While initiation is often thought to be the main regulatory step of translation, recent studies have shown that changes in elongation rate can cause dramatic changes in protein expression in all domains of life[47]. Elongation rate is not uniform; the decoding speed of any given codon is influenced by a variety of different factors. Charged tRNA availability for a given codon influences how quickly the codon is decoded. The cell must have a supply of free amino acids for tRNA charging, and nutrient starvation for specific amino acids can have a strong negative effect on protein expression[48,49].

tRNA availability also leads to the concept of codon optimality, wherein codons that are frequently used in an organism's genome often have higher copy number of cognate tRNAs, and are thus decoded more rapidly[50]. While the optimality of any given codon differs by organism, codon optimality is a major determinant of protein levels across diverse species[55]. Poorly translated codons have also been implicated in regulation of mRNA stability in several other organisms [cite] Peptide

bond synthesis rate also differs between amino acids pairs, and the presence of specific dipeptide combinations in the CDS has been demonstrated to reduce protein expression in *Saccharomyces cerevisiae* (brewer's yeast)[56]. In specific cases, the nascent peptide itself can cause ribosomes to slow down via electrostatic or steric interactions with the proteins and rRNA that form the ribosome exit tunnel[57,58]. In addition, strong mRNA secondary structures such as G4 quadruplexes can also slow elongation rate, and multiple accessory proteins exist to help the ribosome clear such secondary structures[59]. Changes in translation elongation rate can modulate protein output[45,60,61], facilitate cotranslational protein folding[62], and sometimes result in the activation of quality control-off pathways that limit protein expression[46,47]. I will discuss these pathways and their intersection with mRNA stability below.

Termination and the 3' UTR

Open reading frames in mRNA begin with the AUG start codon and are punctuated by stop codons, encoded by TGA, TAG, and TAA nucleotide triplets. When the elongating ribosome encounters a stop codon in the open reading frame, rather than being decoded by a tRNA, a complex of the eukaryotic release factors 1 (eRF1) and 3 (eRF3) enter the A-site. eRF1 decodes the stop codon, which causes GTP hydrolysis and dissociation of eRF3 and repositioning of eRF1 in the A-site, where it then coordinates nucleophilic attack of the nascent peptide by a water molecule, resulting release of the nascent peptide[63]. The ribosome is then removed from the transcript by the ribosome recycling factor ABCE1, and the fully synthesized protein is released to perform its function within the cell[63,64]. While termination isn't generally considered a regulatory step for protein synthesis, given that the mRNA is already fully translated at this stage, termination does affect mRNA stability in a number of ways. Inefficient termination can cause ribosomes to read-through the stop codon, resulting in translation of the nucleotides in the 3' untranslated region (3' UTR). 3'UTR sequences often have non-optimal codon usage or difficult to translate sequences, and their translation can trigger several distinct quality control pathways which result in degradation of the mRNA.

The 3'UTR also contains regulatory sequence motifs which can be bound by RNA binding proteins (RBPs) that directly regulate mRNA translation and mRNA stability[65]. For example, the poly-A tail is bound by poly-A binding protein (PABP), which interacts with eIF proteins in the mRNA cap to facilitate mRNA circularization and promote translation initiation[66]. Another example is the many immune signaling pathway mRNAs that have ARE motifs (A/T Rich Elements) in their 3' UTRs that actively destabilize the transcript. Having a short halflife is beneficial for these immune signaling mRNAs, as this allows quick quenching of immunostimulatory signals that might cause extensive tissue damage if maintained long-term[67]. As a general principle, mRNA halflife is another variable that can be tuned to maintain appropriate protein levels, and mRNA degradation is intimately coupled to translation.

III. mRNA stability and quality control

Canonical mRNA turnover

Differential regulation of mRNA degradation allows cells to further tune their levels of mRNA expression[29,68]. Canonical turnover of cytoplasmic mRNAs is initiated by the shortening of the poly-A tail (deadenylation), followed by removal of the 5' cap (decapping), and finally degradation of the mRNA starting from the now unprotected 5' and 3' termini. Degradation rates of individual mRNAs can vary by over an order of magnitude due to differences in the rates of these deadenylation, decapping, and degradation steps[69]. Deadenylation is carried out by the Ccr4/Pop2/Not and the Pan2/Pan3 protein complexes. The Ccr4 and Pan2 deadenylases are influenced by the binding of PABP to the poly (A) tail; because PABP-mRNA interactions are influenced by translation, this leads to deadenylation being coupled to aspects of translation[69,70]. Removal of the poly-A tail results in loss of PABP on the mRNA, which interferes with closed-loop translation and allows the decapping enzyme Dcp2 to hydrolyze the 5' cap[71]. Uncapped mRNAs are then degraded by cytoplasmic 5' to 3' exonuclease XRN1, which preferentially degrades single stranded RNAs with a 5' monophosphate[71,72]. Following deadenylation, mRNAs can also

be subjected to 3' to 5' degradation by the cytoplasmic RNA exosome, however this pathway is less active than 5' to 3' XRN1 mediated degradation in normal mRNA turnover[71,73,74]. Instead the exosome is critical for the decay of cleaved mRNA products resulting from an array of quality control pathways[75].

Nonsense-mediated mRNA decay

Not all messages should be translated. Mutations in the CDS, incorrect splicing, and UV or chemical damage can all result in faulty mRNAs that encode for incorrect and potentially deleterious protein products[2,76,77] Cells have evolved intricate quality control mechanisms for recognizing and eliminating faulty mRNAs and their protein products[2,65]. Many mRNA quality control pathways depend on ongoing translation; in this regard, the ribosome acts as a quality control hub, sensing mRNA defects and recruiting the appropriate quality control factors[28].

One of the most well studied examples of mRNA quality control is the nonsense-mediated mRNA decay (NMD) pathway. This pathway is activated when the ribosome encounters a premature termination codon (PTC) occurring before the end of an mRNA's canonical CDS, usually due to an error in splicing or a nonsense or frameshift mutation in the CDS[78]. Ribosomes are able to recognize mispositioned premature stop codons because they occur far upstream from the poly-A tail, such that the termination reaction doesn't occur in proximity to PABP, which normally would facilitate efficient termination. Inefficient termination results in the ribosome pausing for an extended period of time on the stop codon — as we'll also see in later sections, ribosome pausing is a general trigger for quality control, though the exact nature of pauses can differ, resulting in different outcomes. In the case of NMD this extended pause results in recruitment of the main NMD factor, UPF1, and a host of additional factors including eRF1, which form into a complex on the mRNA that triggers mRNA degradation. After assembly on the mRNA, phosphorylation of UPF1 results in recruitment of additional protein factors that promote RNA decapping and deadenylation and cause endonucleolytic cleavage, leading to the exposure of the transcript ends to

cellular exonucleases[79]. While not required for NMD, recruitment of UPF1 to PTCs is greatly enhanced by bridging protein factors that interact with the exon-junction complex (EJC) deposited during mRNA splicing[78]. EJCs are removed by the first ribosome to translate a given mRNA (in what is known as the "pioneering round" of translation), however if the ribosome encounters a stop codon upstream of an exon junction, it terminates before knocking off the downstream EJC[80]. Stop codons generally only occur in the last exon, so a downstream exon indicates that the transcript has been miss-spliced and should be degraded; accordingly the presence of a downstream EJC greatly enhances NMD activity[80].

Non-Stop and No-Go decay

In the opposite scenario, transcripts can also be prematurely-polyadenylated, misspliced, mutated, or damaged in such a way that they no longer contain an in-frame stop codon at all. Without a termination signal, ribosomes would remain stuck on the 3' end of the transcript forever. Accordingly, there is a quality control pathway termed non-stop decay (NSD) to prevent this eventuality. In humans, NSD occurs when release factor-like protein HSB1 enters the empty A-site of ribosomes that are stalled at the 3' end of nonstop mRNAs[64,65,81]. HSB1 recruits the protein PELOTA (Dom34 in yeast), a specialized ribosome rescue factor which functions similarly to ABCE1, which dissociates the ribosomal subunits from the mRNA[64]. Activation of this pathway also recruits several SKI complex proteins which facilitate degradation of the mRNA fragment through interactions with the exosome[59,64,82].

Under normal circumstances, ribosomes are unlikely to even reach the end of the mammalian transcript in eukaryotic mRNAs. In the event that a transcript lacks a stop codon, or the ribosome reads through the stop codon, it must first translate the 3' UTR and poly-A tail before reaching the end of the transcript. 3'UTRs are not optimized for protein coding, and thus may contain suboptimal codon usage or amino acid combinations that are difficult to translate, however the real challenge for the ribosome is the poly-A tail. The nucleotide triplet AAA encodes lysine, so the

poly-A tail is translated as consecutive lysine residues, the positive electrostatic charge of which is thought to slow ribosome by interactions with the negatively charged rRNA of the ribosome exit tunnel[57,83]. Poly-A sequences of approximately 12 nucleotides or longer can also cause ribosome slippage on the mRNA, wherein the ribosome tracks backwards by one or more nucleotides, causing frameshifting and slowing ribosome elongation further[83,84]. Ribosomes stalling while translating poly-A sequences results in the activation of an alternate quality control pathway known as no-go decay (NGD)[51,81,85,86]. The NSD factors HSB1 and PELOTA also play a similar role in ribosome rescue for NGD[51,81,85,86], but aside from this overlap, the yeast NSD and NGD pathways are mediated by different sets of well-characterized accessory protein factors, and triggered by distinct coding sequence motifs[45]. While non-stop decay occurs when the ribosome reaches the 3' end of a message, no-go decay is triggered by ribosome stalls, and acts downstream of the ribosome associated-quality control (RQC) pathway[86–88]. Ultimately the yeast NGD pathway results in mRNA cleavage by the recently characterized endonuclease Cue2[89]. While humans have a Cue2 homolog, NEMF, the activity of this protein remains uncharacterized, and generally NGD is less well understood in mammalian cells. The predominantly characterized human NGD factors are HSB1 and PELOTA, and the only known activator is poly-A stretches[81]. Due to this overlap in factors and targets (prematurely polyadenylated mRNAs), human NSD and NGD are sometimes considered the same pathway[65,86,90].

Ribosome-associated quality control

In yeast, no-go decay is a downstream outcome of the ribosome-associated quality control pathway, which senses collided ribosomes. When a trailing ribosome collides with a leading ribosome that has stalled, the two ribosomes' 40S subunits create a unique interface that is recognized by the E3 ligase ZNF598 (Hel2 in yeast), which ubiquitinates eS10 (uS3 in yeast) and uS10 residues of the 40S subunits[87,91,92]. This ubiquitination triggers a cascade of additional ubiquitination events that ultimately result in ribosome rescue by HSB1 and PELOTA, and activation of the NGD pathway

in yeast[93–95]. This process of premature translation termination also results in ubiquitination of the nascent peptide by the factor LTN1, tagging the peptide (which is presumed faulty) for degradation by the proteasome[96,97].

RQC machinery is well conserved from yeast to human cells and much of the initial characterization of this pathway has been performed in yeast[98]. One key difference between the yeast and mammalian RQC systems is that yeast ribosomes engage RQC at a variety of mRNA motifs, including rare codon repeats, poly-basic residues, and strong stem loops, whereas, until recently, mammalian ribosomes had only been demonstrated to induce RQC when they translate four or more consecutive AAA lysine codons[88,92]. Poly-A sequences are strongly depleted in human coding regions, and the human genome encodes only 27 transcripts with 12-consecutive adenines[99]. This strong negative selection is likely because stretches of adenines longer than 12 nucleotides not only trigger RQC, but also cause ribosome frameshifting, which can result in aberrant protein expression or dysregulated transcript silencing due to nonsense-mediated mRNA decay[83,84]. Additionally, while collisions at poly-A sequences reduce mammalian protein expression through RQC, ribosome collisions have not been demonstrated to reduce mammalian mRNA stability[91,100]. Recent work has also demonstrated ribosome collisions that trigger RQC on the endogenous mammalian XBP1u mRNA sequence, which encodes a difficult to translate nascent peptide sequence that causes ribosome arrest, however the effects of RQC on XBP1u mRNA stability remain uncharacterized[101,102]. Interestingly, while the relationship between mammalian mRNA stability and RQC remains unknown, the human RQC pathway is coupled with inhibition of translation initiation through interactions between the proteins GIGYF2 and EDF1, and the 4EBP1 cap binding protein[100,103–105] Reduced initiation and elongation rates are also associated with reduced mRNA halflife, so the feedback loop from ribosome collisions to translation inhibition may also indirectly impact mRNA stability by predisposing RQC substrate mRNAs towards canonical mRNA degradation pathways[100,103,104].

Codon mediated mRNA decay

The lack of known human coding sequences that trigger RQC has been a central question in the translation quality control field since the pathway was mechanistically characterized in 2017[106–108]. However, an opposite but equally perplexing question has vexed the field for even longer; we have known since 2015 that codon usage is a critical determinant of mRNA stability in eukaryotes, but mechanisms of codon-mediated mRNA instability still remain incompletely understood[54,109]. The observation that codon usage correlated with mRNA stability was first made in yeast[109], and soon after in zebrafish[52], but it wasn't until late 2019 that this observation was extended to mammalian systems[110–112]. The factors that recognize codon optimality and destabilize mRNA remained elusive until 2020 when Buschauer and colleagues discovered that the Ccr4-Not complex, which functions in the polyadenylation step of canonical mRNA decay, also monitors the translating ribosome for codon optimality[113,114]. The Ccr4-Not complex accessory protein Not5 (CNOT3 in humans) probes the ribosome A-site during translation elongation and slow decoding kinetics due to suboptimal codon usage result in enhanced mRNA degradation in a CNOT4-dependent manner[113,114]. In addition to codon usage, amino acid usage, GC content, and GC3 content (the presence of a G or C in the third position of a codon) are correlated endogenous human mRNA stability[110–112,115].

V. Thesis Objectives

Rational

The goal of my work is to address a central question in the translation field: What are the coding sequence motifs that regulate mRNA stability in human cells?

In the past few years, several groups have implicated coding sequence metrics such as codon, amino acid, and GC content as regulators of mRNA stability in human cells[110–112,115]. These studies relied on correlation with measured stability of endogenous human mRNAs. However, since the sequences of endogenous mRNAs are the result of complex evolutionary pressures, we have little

mechanistic insight into how the implicated metrics influence mRNA stability. Parallel studies of the few known coding sequence motifs that stall ribosomes in human cells using reporters failed to find any effects on mRNA stability. These prior approaches are limited by a biological constraint: The space of coding sequence motifs that can impact mRNA stability is simply too large ($>4^{30}$) to be characterized by using a few reporters or even by measuring the stability of the entire human transcriptome. Thus, there is a critical need in the translation field for a bottom-up high throughput approach for systematically mapping the effects of increasingly complex coding sequence motifs on mRNA stability.

Approach

To address this problem, I developed a massively parallel assay to measure the mRNA levels of thousands of coding sequence motifs in human cells. By pairing this assay with deep learning models and in vitro experiments, I was able to decipher general structural principles for nascent peptide-mediated gene regulation that have been inaccessible to direct approaches such as cryo-EM. Using this approach, I found that nascent peptides with a combination of β-strand structures and bulky and positively charged sequences trigger acute mRNA instability by stalling ribosomes. The human proteome contains hundreds of short peptide motifs that decrease mRNA levels through the above described nascent peptide code. I then extended my high-throughput sequencing assay to profile the mRNA stability effects of the entire coding sequence of RNA viruses. As a proof of concept, I profiled three RNA viruses that are of great importance to global human health; SARS-CoV-2, influenza A virus, and HIV. Using this approach, I uncovered both known and previously uncharacterized viral sequence motifs that dramatically influence mRNA stability when translated.

Chapter 2: A Nascent Peptide Code for Translational Control of mRNA Stability in Human Cells

Portions published in: Burke P., Park H., Subramaniam AR. A Nascent Peptide Code for Translational Control of mRNA Stability in Human Cells. *Nature Communications* 2022, https://doi.org/10.1038/s41467-022-34664-0.

I. Introduction

Protein expression is determined by a balance between the translation rate and stability of mRNAs. In human cells, mRNA stability is often regulated by sequence motifs in the 3′ untranslated region such as microRNA-binding sites and AU-rich elements[65]. Additionally, the protein coding region has been recently recognized as a critical determinant of eukaryotic mRNA stability[50,98]. The role of the coding sequence in mRNA stability is best understood in the budding yeast *S. cerevisiae* where poorly translated codons and nascent peptide motifs with positively charged residues can destabilize mRNAs[51,88,109,116]. Poorly translated codons have also been implicated in regulation of mRNA stability in several other organisms[52–55].

Coding sequence features regulating mRNA stability in human cells are less clear. Several recent studies examined the coding sequence determinants of endogenous mRNA stability in human cells and arrived at differing conclusions. Two studies implicated synonymous codon choice as the primary determinant of mRNA stability in human cells[110,111]. Another found GC and GC3 (wobble base GC) content as major factors regulating mRNA stability[115]. A fourth study identified amino acid content to be an important contributor[112]. Extended amino acid motifs and G-quadruplexes in coding regions have also been implicated as triggers of specific mammalian mRNA decay pathways[59,100]. The associations reported in these studies relied on endogenous human coding sequences. Since human mRNAs differ from each other in codon, amino acid, and GC content as well as in their length and the presence of specific sequence motifs, it is challenging to identify the contribution

of each factor to mRNA stability. Further, reporters used in the above studies for validation differ extensively in their nucleotide or amino acid content, which complicates their interpretation.

Here, we developed a massively parallel assay to measure the mRNA effects of thousands of coding sequence motifs in human cells. We designed our assay with the initial goal of systematically delineating the individual contribution of mRNA features implicated in previous studies. Instead, we unexpectedly uncovered a potent role for the sequence and structure of the nascent peptide in regulating mRNA stability and ribosome elongation rate. The resulting nascent peptide code regulates the ribosome stalling and mRNA destabilizing effects of hundreds of endogenous peptide sequences from the human proteome. Our results point to an unappreciated role for the ribosome as a selectivity filter against the synthesis of bulky and aggregation-prone peptide sequences.

II. Results

A massively parallel assay for mRNA levels in human cells

We reasoned that coding sequence motifs that alter mRNA stability should be identifiable through their effects on steady state mRNA levels. To study the effect of coding sequence motifs on mRNA levels in an unbiased manner, we designed a library of 4,096 oligonucleotides made of all possible codon pairs (Fig. 2.1). We repeated each codon pair as a tandem 8× repeat with the rationale that their effects will be amplified and readily measurable. We cloned the oligonucleotide library as a pool into a dual fluorescence reporter vector separated by 2A linkers – a design widely used for studying ribosome stalling motifs in human cells[91,94,101,105–107]. We added multiple random 24nt barcodes without stop codons 3′ of each oligonucleotide insert and linked the barcode sequences to the corresponding insert by high-throughput sequencing. Most studies of coding sequence motifs use transient transfection or lentiviral integration of reporters, which makes measurement of steady state effects on mRNA levels across a large pool difficult. To avoid this, we stably integrated the reporter pool at the *AAVS1* locus of HEK293T cells using CRISPR Cas9-mediated homologous

recombination. We extracted mRNA and genomic DNA from the pooled cells and counted each barcode by high-throughput sequencing. Normalization of the total barcode count in the mRNA by the corresponding count in the genomic DNA for each of the 4,096 inserts provides a relative measure of the steady-state mRNA level of that insert. We examined whether our assay captured the effects of known mRNA-destabilizing motifs. We first calculated the effect of individual codons on mRNA level, by averaging across all possible neighboring codons as well as across the first and second positions of each codon within the repeat (Fig. 2.2a). While our library reporter does not encode any splice sites, the variable insert region is upstream of a ~700 nt eYFP cassette, thus stop codons in the variable region are perceived as being upstream of a long 3'UTR, which can trigger nonsense-mediated mRNA decay[79]. Accordingly, stop codons in either the first or second position of the codon pair repeat decrease mRNA levels (Fig. 2.2a, Fig. 2.3a), consistent with a mRNA destabilizing effect due to NMD[117–119]. We also observe a mild correlation between our measured effects of codons on mRNA level and published codon stabilization coefficients calculated from endogenous mRNA stability (Fig. 2.2c)[112]. However, mRNA levels in our assay show little correlation with GC and GC3 content (Fig. 2.2b) or with binary measures of codon optimality (Fig. 2.2a)[110–112,115]. Instead, the strongest differences in mRNA abundance in our assay are seen at the amino acid level, with effects spanning a 2-fold range in relative abundance (Fig. 2.2d, Fig. 2.3a). Among the twenty amino acids, the positively charged amino acids lysine and arginine cause the largest average decreases in mRNA levels (Fig. 2.3a). The known association between positively charged residues in the nascent peptide and slow elongation[57,120–123] suggests that the decrease in steady-state mRNA levels observed in our assay is caused by ribosome slowdown at these residues.

Specific dipeptide repeats trigger decrease in mRNA levels

We wondered whether the average effects of amino acids on mRNA levels (Fig. 2.3a) belie larger effects driven by specific amino acid combinations. We assessed the effect of each pairwise amino

acid combination on mRNA abundance and found that these combinations span over a 16-fold range in relative abundance in our assay (Fig. 2.3b). While lysine and arginine reduce mRNA levels on average, unexpectedly, these amino acids have mild or no negative effect on mRNA levels on their own (Fig. 2.3b: Lys-Lys, Arg-Arg, Arg-Lys). Rather, the effects of lysine and arginine are primarily driven by co-occurrence with bulky amino acids[124] (ratio of side chain volume to length > 18Å^2) such as valine, isoleucine, leucine, phenylalanine, and tyrosine (Fig. 2.3b). Likewise, most bulky amino acids decrease mRNA levels in combination with lysine and arginine, but not on their own (Fig. 2.3b). Oddly, we don't observe a reduction in mRNA level when arginine and lysine are combined with the bulky amino acid tryptophan. However we got relatively low sequencing depth for the tryptophan-encoding inserts in our library, resulting in higher than average noise for these inserts (evidenced by the variability of Trp effects in Fig. 2.3a), which may obscure mRNA level effects caused by co-occurrence with arginine or lysine. In addition, a few dipeptides that contain certain positively charged amino acids (Arg-His) or bulky amino acids (Phe-Ser) also have a strong negative effect on steady-state mRNA levels (Fig. 2.3b). The combinatorial effect of positively charged and bulky amino acids on mRNA level is captured by a linear statistical model (Fig. 2.3c): Isoelectric point[124] (pI, a measure of positive charge) and bulkiness[124] of amino acids are positive correlates of mRNA level, while an interaction term between these two physical properties is a negative correlate of mRNA level [mRNA = (0.31 × pI) + (0.20 × bulkiness) – (0.03 × pI × bulkiness), Adjusted R^2 = 0.25]. By contrast, ignoring the interaction between pI and bulkiness results in negative or no correlation of these properties with mRNA level (mRNA = – 0.18 × pI, Adjusted R^2 = 0.21), which is in line with Fig. 2.3a. The effects of dipeptide repeats in the translated +0 frame strongly correlates with the codon-matched +3 frame, but only weakly with the codon-mismatched +1 and +2 frames (Fig. 2.3d). The high correlation between the +0 and +3 frames is also seen from the diagonal symmetry of Fig. 2.3b and arises from similarity of the encoded peptides (for example $(XY)_8$ and $(YX)_8$ are identical except at their termini). These frame correlations are consistent with

the mRNA effects arising at the translational level as opposed to transcriptional or RNA processing differences. Together, our results show that translation of bulky and positively charged amino acids is critical for their negative effect on mRNA level.

Primary sequence of dipeptide repeats regulates mRNA stability

Several observations suggest that translation of specific dipeptide repeats is a general trigger of mRNA instability in human cells. To characterize various dipeptide repeats influence mRNA levels, we picked three distinct dipeptides to test further; the charged and bulky valine-lysine, highly charged arginine-histidine, and bulky but uncharged serine-phenylalanine combinations.

Multiple human cell lines show lower mRNA levels of these dipeptide repeats relative to their frameshifted controls (HEK293T, HeLa, HCT116, and K562; Fig. 2.4a), pointing to the generality of the observed effects. Upon actinomycin D treatment to inhibit transcription, transcripts from reporters with mRNA level-reducing dipeptides decay faster than their frameshifted controls (Fig. 2.4b). This confirms that the decrease in steady-state mRNA levels caused by dipeptide repeats arises from reduction in mRNA stability.

We wondered if translation of dipeptide inserts that reduce mRNA levels and mRNA stability also cause premature translation termination[106]. To test this, we used fluorescence-activated cell sorting followed by genomic DNA barcode sequencing (FACS-seq) on the 8× codon pair library (Fig. 2.1). This reporter library encodes 2A-linked upstream RFP and downstream YFP cassettes surrounding the variable dipeptide sequence, such that inserts that cause premature translation termination will produce RFP but not YFP fluorescence signal (Fig. 2.1). We sorted cells that had low YFP signal relative to RFP (`low-YFP gate` in Fig. 2.4c and Supplementary Fig. 4d), and then measured the enrichment of each dipeptide insert in this low-YFP population relative to the unsorted population (Fig. 2.4d). Inserts encoding stop codons between RFP and YFP are enriched in the low-YFP population, indicating that our assay robustly identifies inserts that cause premature termination. Similarly, inserts with lower mRNA levels (< 2-fold below median in Fig. 2.3b) are also

significantly enriched in the low-YFP gate relative to all other dipeptide inserts, indicating that such inserts also cause premature termination in addition to reducing mRNA levels.

Finally, to decipher the effect of dipeptide repetition on mRNA levels, we systematically varied the number of several destabilizing dipeptides identified in our initial assay (Fig. 2.5a). As the number of dipeptide repeats increases from 1 to 8, each dipeptide starts decreasing reporter mRNA levels at a distinct repeat number between 4 and 7 (Fig. 2.4a). We then altered the periodicity of dipeptide repeats by intermixing dipeptides with their reversed counterparts such that the overall amino acid composition remains unchanged (Fig. 2.5b). Even minor perturbations of RH repeats abrogate their negative effect on mRNA levels (Fig. 2.5b). By comparison, VK repeats had a gradual negative effect on mRNA levels as their periodicity is increased, while SF repeats show an intermediate trend (Fig. 2.5b). These experiments reveal that the primary sequences of destabilizing dipeptide repeats encode critical regulatory information beyond the identity of the amino acid pairs forming the repeats.

Secondary structure of dipeptide repeats mediates mRNA effects

Since dipeptide sequences are known to form distinct secondary structures based on their periodicity[125,126], we asked whether mRNA-destabilizing dipeptide repeats adopt specific secondary structures. Using a deep neural network model for secondary structure prediction[127], we find that many dipeptide repeats that strongly reduce mRNA levels *in vivo* are computationally predicted to form β strands with a high probability (Fig. 2.6a). We next assigned all dipeptide repeats in the library to either α helices or β strands if their respective prediction probabilities are greater than 0.5. We find that dipeptide repeats predicted to form β strands have a significantly lower mRNA level on average than those predicted to form α helices (Fig. 2.6b, $P < 0.001$, two-sided Mann-Whitney test). This observation is consistent with the destabilizing amino acids lysine and arginine predominantly occurring in β strands or unstructured peptides in our library (See Burke et al. 2022 for extended data)[128]. Among dipeptides containing the positively charged

amino acids lysine or arginine, the measured propensity of the second amino acid to occur in a β strand[129] ('Chou-Fasman propensity') is highly correlated with mRNA instability (Fig. 2.6c). This correlation is not observed with α helix propensities of the same amino acids (Fig. 2.6c) suggesting that β strand formation promotes mRNA instability, as opposed to α helix formation stabilizing mRNAs in our assay. mRNA levels of dipeptide repeats containing the negatively charged amino acid glutamate, which are also predicted to form β strands with high probability when combined with bulky amino acids do not show significant correlation with β strand or α helix propensities (data available in Burke et al. 2022)[128]. Thus, a combination of bulky and positively charged amino acids in the primary sequence and β strand in the secondary structure are strong and significant predictors of the mRNA-destabilizing effects of dipeptide repeats [mRNA = (0.30 × pI) + (0.23 × bulkiness) – (0.03 × pI × bulkiness) – (0.52 × β-strand-propensity), Adjusted R^2 = 0.27].

Extended β strands slow ribosome elongation and reduce mRNA levels

To test the causal role of β strands in nascent peptide-mediated translational control, we combined the mRNA-destabilizing dipeptides VK, KV, SF, and FS pairwise into 16 amino acid-long peptides. Even though the four constituent dipeptides are strongly predicted to form β strands on their own (Fig. 2.6a), their combinations can form either β strands or α helices with high probability (Fig. 2.7a). Importantly, all combinations are encoded by the same set of four amino acids to control for amino acid composition. We commercially synthesized two 16 amino acid peptides and used circular dichroism to confirm their secondary structure *in vitro* (Fig. 2.7b, left panel). As predicted (Fig. 2.7a), 4×SVKF primarily forms β strands in aqueous solution, while 4×SKVF forms α helices in the presence of trifluoroethanol (TFE) as a co-solvent[130–132] (Fig. 2.7b, right panel). We then measured the transit time of ribosomes on mRNAs encoding 16 amino acid inserts preceding a nanoluciferase reporter in a rabbit reticulocyte lysate (RRL) *in vitro* translation system (Fig. 2.7c). The β strand-forming 4×SVKF and 4×VKFS inserts slow ribosome elongation relative to the α

helix-forming 4xSKVF and 4xKVFS inserts, with a 200 s difference in *in vitro* transit time (Fig. 2.7c). Strikingly, all β strand peptides decrease mRNA levels over 8-fold relative to α helix controls when tested *in vivo* using our cDNA/gDNA barcode-sequencing reporter assay (Fig. 2.7d). We observe similar effects on mRNA level due to β strand formation in HeLa, HCT116, and K562 cells (data available in Burke et al. 2022)[128]. We also tested the translation kinetics of the β stranded VK8 insert by RRL nanoluciferase assay and found that this insert slows ribosome transit time by 100 s relative to its frameshifted control (data available in Burke et al. 2022)[128]. Thus, nascent peptides that contain positively charged and bulky amino acids and that are predicted to form β strands trigger ribosome slowdown in human cells. This observation agrees with disome profiling results on endogenous mRNAs, where R-X-K motifs (R – Arg, X – any amino acid, K – lysine) are highly enriched at E, P, and A sites respectively of the lead ribosome[101]. Notably, several R-X-K motifs with the highest disome density have interspersed bulky residues such as phenylalanine, isoleucine, and leucine[101].

Dipeptide motifs in the human genome reduce mRNA levels

We sought to identify endogenous sequences in the human genome that regulate mRNA levels based on the dipeptide code identified above. To do this, we scanned all annotated human protein coding sequences for destabilizing dipeptide combinations of bulky and positively charged amino acids (Fig. 2.8a). Using a heuristic peptide score (Fig. 2.8a, top), we identified the 16 amino acid long peptide within each coding sequence that has the maximum density of destabilizing dipeptides. To test whether these endogenous motifs above can reduce mRNA levels, we cloned 1,201 such motifs into our reporter and measured their mRNA levels by high throughput sequencing (Fig. 2.8b). Motifs with high destabilizing dipeptide content result in lower mRNA levels than control motifs ($P < 0.01$, Fig. 2.8b, left panel). Among destabilizing motifs, those predicted to form β strands result in lower mRNA levels than the remaining motifs ($P < 0.05$, Fig. 2.8b, right panel). To confirm the destabilizing role of the specific dipeptides identified in our study, we disrupted them by

moving the bulky and positively charged amino acids to opposite ends without changing the amino acid composition in 1,079 endogenous motifs (Fig. 2.8c, top). As predicted, the resulting mutations increase mRNA levels (median $\log_2$ ΔmRNA = 0.38) with 783 mutated motifs having significantly higher mRNA levels ($P < 0.05$) than their wild-type counterparts (Fig. 2.8c, bottom). Examination of destabilizing motifs with annotated β strand structures in the Protein Data Bank (PDB) shows that these β strands are part of antiparallel β sheets and are significantly longer than the 5-6 residue length of typical β strands[133] (Fig. 2.8d). Together, these results show that β-stranded endogenous motifs containing bulky and positively charged dipeptides can reduce mRNA levels.

III. Discussion

Here, we identify a combinatorial code composed of bulky, positively charged, and extended β strand nascent peptides that regulates translation and mRNA stability in human cells. We demonstrate that a minimal combination of these sequence and structural elements is sufficient to induce ribosome slowdown and cause changes in gene expression, and is widespread in the human proteome. As discussed below, elements of the code uncovered here allow us to synthesize a large body of observations on nascent peptide-mediated slowdown of ribosomes and regulation of mRNA stability in human cells. Our results also point to a role for the ribosome as a post-synthesis filter against nascent peptide sequences that are bulky and aggregation prone.

The nascent peptide code for mRNA stability described here is significantly more complex and localized along the mRNA than previously associated sequence features such as codons, amino acids, and GC content[110–112,115]. We don't observe large effects on mRNA levels due to codon optimality or GC content in our assay (Supplementary Fig. 1). This is likely because the 48 nucleotide inserts constitute only ~3% of the 1725 nucleotide coding sequence of our library reporters (Fig. 2.1), which limits the impact changing these motifs can have on overall reporter composition. Nevertheless, some individual codon and amino acid signatures in our data agree with the findings of previous studies (Fig. 2.2a, Fig. 2.3a). For example, bulky amino acids such as Leu, Ile,

Val, and Phe are stabilizing on average, though their codon-specific effects vary across previous studies[110–112]. The amino acid serine shows prominent codon-specific effects, with AGU and AGC codons reducing mRNA level more than the remaining codons[110–112,115]. The methionine AUG start codon and the near-cognate start codons (CUG, GUG, UUG) all promote mRNA stability[110–112,115], possibly through effects on increased downstream translation[134]. With the exception of arginine, lysine, and glycine, our amino acid level effects correlate with the amino acid stability coefficient calculated from endogenous mRNA stability (Fig. 2.2d)[112]. While glycine codons generally stabilize endogenous mRNAs in prior studies, all four glycine codons decrease mRNA levels in our assay, suggesting that glycine dipeptides also cause nascent peptide-mediated ribosome slowdown and mRNA instability. Indeed, we find that Gly-Gly dipeptides reduce mRNA levels (Fig. 2.3b) consistent with previous observations that poly-glycine motifs slowdown ribosomes[135]. In our data, glycine has the largest effects on mRNA levels when in combination with Leu and Phe, suggesting a nascent peptide-mediated destabilization mechanism akin to that of the biochemically similar Ser-Phe dipeptides. While positive charge in the nascent peptide can slow ribosomes[57,120], our results show that positive charge by itself is insufficient to induce changes in gene expression in human cells. The importance of bulky amino acids for mRNA effects observed here is in line with the role of side chain bulk in ribosome-associated quality control in *S. cerevisiae*[136]. Further, bulky synthetic amino acid analogs in the nascent peptide and small molecules that add bulk to the exit tunnel can both reduce ribosome elongation rate[137–140]. Ribosome profiling in *S. cerevisiae* and human cells shows that tripeptide combinations of bulky and positively charged amino acids are enriched at sites of increased ribosome density[101,141]. Bulky and positively charged amino acids also play critical roles in many known ribosome-arresting peptides[102,116,142–144], and several human arrest peptide sequences stall ribosomes specifically in the presence of small molecule metabolites or drugs in the ribosome exit tunnel[145,146]. Structural studies of arrest peptides suggest that bulky and positively charged amino acids might slow down ribosomes by altering the geometry of

the peptidyl-transferase center (PTC) and/or by steric interactions with the constriction point in the exit tunnel formed by the uL4 and uL22 proteins[102,116,147,148].

Our work shows that extended β strand motifs in nascent peptides contribute to ribosome slowdown and mRNA instability in human cells. This role of a simple secondary structural motif like β strand is surprising given that cryo-EM studies of stalled ribosome nascent chain complexes reveal a diverse range of extended conformations, turns, and helices that are specific to each arrest peptide[116,140,144,149,150]. This comparison is complicated by the fact that cryo-EM studies are performed on post-arrest complexes where the nascent chain might have already undergone extensive conformational rearrangements. Further, while several motifs uncovered here form β strands *in silico* and *in vitro* in isolation, they might have a significantly different structure within the confined geometry of the ribosome exit tunnel[132,151–153]. At the molecular level, β strands in nascent chains could contribute to ribosome slowdown as an allosteric relay that communicates steric interactions between the nascent chain and the distal portions of the ribosome exit tunnel such as the uL4/uL22 constriction to the PTC[116,138,148,149,154]. This possibility is supported by our observation that destabilizing dipeptide repeats are at least 10-12 amino acids long (Fig. 2.5a), which is consistent with the distance between the uL4/uL22 constriction and the PTC.

In addition to the sequence and structural determinants of nascent peptide-mediated ribosome slowdown studied here, several classes of nascent peptide sequences that slowdown ribosomes might not be revealed by our assay. For example, poly-prolines do not emerge as destabilizing motifs in our assay even though they are known to slowdown ribosomes[101]. This is likely because poly-proline stalls are resolved without triggering quality control or mRNA instability[59]. While extended β strands are the primary structural motif associated with ribosome slowdown here, we also find motifs with unstructured regions that nevertheless reduce mRNA levels (Figs. 2.6b, 2.8b). This might be in part due to limitations of existing computational methods[127] to predict secondary structures or their limited relevance to secondary structures forming inside the ribosome. It is also likely

that the combinatorial code of positively charged, bulky, and β strand sequences uncovered here underlies some, but not all, classes of nascent peptides that have the potential to slowdown ribosomes and effect changes in gene expression. For example, the arginine-histidine dipeptide repeat destabilizes mRNA and causes premature termination similar to the β stranded Val-Lys and Ser-Phe inserts (Fig. 2.4a-d). Unlike the latter inserts, Arg-His effects require a longer insert length and strict dipeptide periodicity (Fig. 2.5a-b), and occur with no predicted β strand formation (Fig. 2.6a). The 8×Arg-His repeats are reminiscent of dipeptide repeat expansions in the human C9ORF72 gene, which cause neurological disease in humans[155–157]. Alternate initiation in the C9ORF72 ORF results in translation of extended Arg-Gly and Arg-Pro repeats, which stall ribosomes and cause premature termination in a length dependent manner, with 20× dipeptide repeats being the minimal length required to to stall ribosomes[158,159]. Unsurprisingly, we do not observe marked effects from 8× Arg-Gly or Arg-Pro in our assay, as 10× repeats of these dipeptides do not cause premature termination[158]. However, to our knowledge, Arg-His dipeptide repeats have never been tested in this manner prior to our work. It may be that Arg-His repeats impact translation through a similar mechanism as Arg-Gly and Arg-Pro repeats, but with a more acute effect on ribosome elongation that requires fewer repeats to trigger. Notably, Arg-rich peptides without Gly or Pro dipeptide periodicity (for example 12×Arg) do not stall ribosomes[158,160]. This agrees with our finding that positively charged dipeptide repeats composed of 8×RR and 8×RK have little effect on mRNA levels (Fig. 2.3b).

Nascent peptides that slowdown ribosomes might exert their effects on mRNA stability through distinct cellular pathways compared to the ones sensing codon, amino acid, and GC content of mRNAs[89,105,113,161]. In this vein, poly-lysine sequences encoded by poly-A and the *Xbp1* arrest sequence are among the few known nascent peptide motifs with intrinsic ability to stall ribosomes in human cells[86,162,163]. Both poly-A runs and the *Xbp1* arrest sequence are substrates of the ribosome-associated quality control (RQC) pathway, which causes premature translation

termination in response to ribosome collisions, limiting production of the proteins encoding these motifs[91,101,102,106,107]. The RQC pathway is most well studied in yeast, where it also destabilizes the mRNA encoding the stalling motif through activity of the endonuclease Cue2, in a process termed No-Go decay[51,85,89]. While the effects of the human RQC pathway on mRNA stability are not fully characterized, humans have a Cue2 homolog, N4BP2, which suggests that this pathway could reduce mRNA level in addition to limiting protein production[100]. There also are examples of pathological peptide repeat sequences that cause ribosome slowdown and premature termination which are not subject to the RQC pathway[158]. This includes Arg-Gly and Arg-Pro dipeptides from the C9ORF72 ORF, which cause amyotrophic lateral sclerosis and frontotemporal dementia[155,156], and poly-glutamine repeats translated from CAG nucleotide repeat expansions in the *mHtt* gene, which cause Huntington's disease[164–166].

Interestingly, although the RQC pathway isn't demonstrated to act directly on these toxic repeats, expression of RQC pathway components is associated with lower disease severity in both instances[166,167]. As the destabilizing peptide sequences we identify in this study cause ribosome slowdown (Fig. 2.7c) and premature termination (Fig. 2.4d), we suspect that some inserts may be directly repressed by RQC, in a manner similar to the stalling XBP1u nascent protein[102], whereas others may be resistant to RQC repression, as is the case with Arg-rich dipeptides[158]. In addition, it is likely that the effects of endogenous nascent peptide motifs on ribosome slowdown and mRNA stability are modulated by other cotranslational events such as nascent protein folding outside the ribosome[168,169], membrane insertion[170,171], and multiprotein assembly[172,173].

The nature of the nascent peptide code uncovered here has important implications for cellular homeostasis and disease. Ribosome slowdown and mRNA destabilization induced by bulky and extended β strands, which are highly aggregation prone[174,175], implies that the ribosome has an intrinsic ability to throttle the synthesis of such proteins. Ribosome slowdown at extended β strands could serve as a quality control mechanism, testing the ability of long β strands (10 amino acids

or greater in length) to eventually fold into antiparallel β sheets outside the ribosome, and thus avoid aggregation. This ribosomal selectivity filter would act before other cotranslational mechanisms such as codon optimality that help avoid aggregation after β strands emerge from the ribosome[62,176]. Slow translation elongation without mRNA decay can also help recruit protein chaperones, which may be important to properly fold β strands[177,178]. Finally, the gene regulatory potential of the dipeptide motifs uncovered here suggests that disease-causing missense mutations occuring at these motifs might exert their phenotype by altering protein expression *in cis* rather than protein activity.

IV. Materials and Methods

Plasmid construction

Plasmids, oligonucleotides, and cell lines used in this study are listed in Supplementary Tables S1-S3.

Parent vector construction

The AAVS1-targeting parent vector pPBHS285 used for this study was constructed using Addgene plasmid #68375[179] as a backbone. The PGK1 promoter was replaced with the CMV promoter and the native pCMV 5′ UTR region. The coding sequence was replaced by a codon-optimized mKate2 and eYFP fusion cassette, linked with two 2A linker sequences. These 2A sequences surround a cassette encoding an EcoRV restriction site, Illumina R1 sequencing primer binding site, and a T7 promoter. The R1 primer binding and T7 sequences are in reverse orientation (3′ - 5′) for *in vitro* transcription and sequencing of inserts and barcode sequences at the EcoRV site.

Variable oligo pool design

Four oligo pools were designed for this study.

Pool 1 (Fig. 2.2, Fig. 2.3, Fig. 2.6b-c) encodes all possible dicodon (6 nt) combinations, for a total of 4096 codon pairs. These 6nt dicodon inserts were repeated eight times to create 8× dicodon

repeat inserts, each 48nt in length.

Pool 2 (Fig. 2.5a,b, Fig. 2.7d) encodes several dipeptide combinations identified in Library 1 as causing mRNA instability. For Fig. 2.5b, the number of dipeptide repeats was systematically reduced from 8 to 1. Repeats were replaced with a Ser-Gly linker, shown to be not destabilizing in Library 1, to maintain a constant 48nt insert length. For Fig. 2.5b, periodicity of dipeptides was by interspersing 1, 2, or 4 tandem repeats of each dipeptide with an equal number of its sequence-reversed counterpart. For Fig. 2.7d, destabilizing dipeptides KV and SF were combined and rearranged to form either α helices or β strands, as predicted by S4PRED.

Pool 3 (Fig. 2.8) encodes the 16 amino acid nascent peptide motifs from the human proteome identified as potentially destabilizing by the scoring method described in Fig. 2.8a along with 4 flanking codons on either side. The library encodes the top 1079 predicted stalling motifs with a peptide score > 9, and 122 control motifs with a peptide score < 3. The library also includes the mutants with reordered amino acids from the 1,079 endogenous destabilizing dipeptide motifs, which were designed as described in Fig. 2.8c.

Pool 4 (Fig 2.4a,b encodes 8 inserts: 3 destabilizing dipeptide repeats $(RH)_8$, $(VK)_8$, $(SF)_8$, their respective frameshift controls $(PS)_8$,$(QS)_8$,$(FQ)_8$, the β strand peptide $(SVKF)_4$, and the α helix peptide $(SKVF)_4$.

Oligo pools 1-3 were synthesized by Twist Biosciences with flanking sequences for PCR and cloning into the EcoRV site of the parent pPBHS285 vector. Oligo pool 4 was cloned by PCRing individual inserts and pooling them before cloning.

Plasmid library construction

Parent vector pPBHS285 was digested with EcoRV. The oligo pools described above were PCR amplified using primers oHJ01 and either oPB348 (Library 1) or oPB409 (Libraries 2–4). oPB348 and oPB409 both encode a 24 nt random barcode region, comprised of 8×VNN repeats to exclude

in-frame stop codons (where V is any nucleotide except T). Barcoded oligo pools were cloned into pPBHS285 by Gibson assembly. Assembled plasmid pools were transformed with high efficiency into NEB10Beta *E.coli*. For pools 1-3, the transformed plasmid pools were extracted from 15-50 *E.coli* colonies per insert in the library, thus bottlenecking the number of unique barcodes present in each plasmid pool. Resulting plasmid pools contained between 60,000–400,000 unique barcode sequences for pools 1-3. For pool 4, the transformed library was bottlenecked to around 150 barcodes per insert, and 6 such pools with distinct barcodes were extracted for multiplexed library preparation of different cell lines.

The plasmid libraries corresponding to pools 1-4 are pPBHS286, pPBHS309, pHPHS296, and pHPHS406, respectively. Variable insert and barcode sequences for each plasmid library are provided as part of the data analysis code.

CRISPR vectors

The CLYBL-targeted Cas9-BFP expression vector pHPHS15 was constructed by Golden Gate assembly of either entry plasmids or PCR products with pHPHS11 (MTK0_047[180] Addgene #123977) as backbone, pHPHS3 (MTK2_007[180] Addgene #123702) for the pEF1a promoter, pADHS5[49] (pU6-(BbsI)_CBh-Cas9-T2A-BFP[181] Addgene #64323) for the Cas9-2A-BFP insert cassette, and pHPHS6 (MTK4b_003[180] Addgene #123842) for the rabbit β-globin terminator. sgRNA vectors pPBHS320 (gRNA_AAVS1-T1 Addgene #41817) and pADHS4[49] (eSpCas9(1.1)_No_FLAG_AAVS1_T2 Addgene #79888) were used for insertion at the AAVS1 locus. pASHS16 (MTK234_030 spCas9-sgRNA1-hCLYBL[180] Addgene #123910) was used for insertion at the CLYBL locus.

Cell line maintenance and generation

HEK293T cells (RRID:CVCL_0063, ATCC CRL-3216), HCT116 cells (RRID:CVCL_0291, NCI60 cancer line panel), and HeLa cells (RRID:CVCL_0030, ATCC CCL-2) were grown in DMEM

(Thermo 11965084). K562 cells (RRID:CVCL_0004, ATCC CCL-243) were grown in IMDM (Thermo 12440053). Media for all cells was supplemented with 10% FBS (Thermo 26140079). Cells were grown at 37C in 5% CO2. All transfections into HEK293T, HCT116, and HeLa cells were performed using Lipofectamine 3000 (Thermo L3000015). Transfections into K562 cells were performed using an Amaxa Nucleofector V kit (Lonza VCA-1003). HEK293T cells that stably express Cas9 (hsPB80) were generated by transfecting the CLYBL::Cas9-BFP vector pHPHS15 and spCas9 sgRNA1 hCLYBL vector, and selecting with 200μg/mL hygromycin.

CRISPR integration of plasmid libraries

hsPB80 CLYBL::Cas9-BFP HEK293T cells were seeded to 50% confluency on 15 cm dishes for all library transfections. 10 μg of library plasmid (pPBHS286, pPHBS309, or pHPHS296) and 1.5 μg of each AAVS1 targeting CRISPR vector were transfected per 15 cm dish. pPBHS286, and pPBHS309 were each transfected into a single 15 cm dish. pHPHS296 was transfected into three 15 cm dishes. pHPHS406 pools with different barcodes were transfected into single 10 cm dishes of hsPB80, HeLa, HCT116 and 2 million cells of K562. Cells were selected with 2 μg/mL puromycin, added 48 hours post-transfection. Cells from the three pHPHS296 transfections were combined at the start of selection. Puromycin selection was removed after 6-10 days, once cells were growing robustly in selection. 24 hours after removing puromycin selection, stable library cells were plated into two separate 15cm dishes, to reach 75% confluency the next day, for matched mRNA and gDNA harvests. For pHPHS406, libraries were maintained in two 10 cm dishes or T75 flasks (for K562).

mRNA stability measurement

hsPB80 cells containing the stably integrated pHPHS406 library were seeded to 50% confluence in a 6-well plate. Actinomycin D (ActD) power was dissolved in DMSO at 1 mM (1.25 mg/mL) and added to each well of the 6-well plate to a final concentration of 5 μg/mL. Before harvesting, 1

million HeLa cells containing the pHPHS406 library were lysed in 6 mL of Trizol reagent, to create a Trizol lysis solution containing a set number of mRNAs with different barcodes than those in the hsPB80 pHPHS406 pool, for barcode count normalization across samples. ActD treated hsPB80 wells were harvested at 0, 0.5, 1, 2, 4, and 6 hours after the addition of ActD by adding 0.75 mL of the Trizol lysis solution above to wells at each timepoint, then following the manufacturer's Trizol mRNA extraction protocol.

Library Genomic DNA extraction

Reporter library genomic DNA was harvested from one 75% confluent 15 cm or 10 cm dish of stably expressing library cells. Genomic DNA was harvested using Quick-DNA kit (Zymo D3024), following the manufacturer's instructions, with 3 mL of genomic DNA lysis buffer per 15 cm plate, and 1 ml of the same buffer per 10 cm plate. Between 0.5-10 μg of purified genomic DNA from each library sample was sheared into ~350 nucleotide length fragments by sonication for 10 min on ice using a Diagenode Bioruptor. Sheared gDNA was then *in vitro* transcribed into RNA (denoted gRNA below and in analysis code) starting from the T7 promoter region in the insert cassette, similar to previous approaches[182,183], using a HiScribe T7 High Yield RNA Synthesis Kit (NEB E2040S). Transcribed gRNA was treated with DNase I (NEB M0303S) and cleaned using an RNA Clean and Concentrator kit (Zymo R1013).

Library mRNA extraction

Reporter library mRNA was harvested from one 75% confluent 15 cm or 10 cm dish of stably expressing library cells. mRNA was harvested by using 3 mL of Trizol reagent (Thermo) to lyse cells directly on the plate, and then following the manufacturer's mRNA extraction protocol. Purified mRNA was then DNaseI (NEB M0303S) treated and cleaned using an RNA Clean and Concentrator kit (Zymo R1013).

mRNA and genomic DNA barcode sequencing

Between 0.5-10 μg of DNaseI-treated mRNA and gRNA for each library was reverse transcribed into cDNA using Maxima H Minus Reverse Transcriptase (Thermo EP0752) and a primer annealing to the Illumina R1 primer binding site (oPB354). A 170-nucleotide region surrounding the 24-nucleotide barcode was PCR amplified from the resulting cDNA in two rounds, using Phusion Flash High-Fidelity PCR Master Mix mastermix (Thermo F548L). Round 1 PCR was carried out for 10 cycles, with cDNA template comprising 1/10th of the PCR reaction volume, using primers oPB361 and oPB354. Round 1 PCRs were cleaned using a 2× volume of Agencourt Ampure XP beads (Beckman Coulter A63880) to remove primers. Cleaned samples were then used as template for Round 2 PCR, carried out for 5-15 cycles, using a common reverse primer (oAS111) and indexed forward primers for pooled high-throughput sequencing of different samples (oAS112-135 and oHP281-290). Amplified samples were run on a 1.5% agarose gel and fragments of the correct size were purified using ADB Agarose Dissolving Buffer (Zymo D4001-1-100) and UPrep Micro Spin Columns (Genesee Scientific 88-343). Concentrations of gel-purified samples were measured using a Qubit dsDNA HS Assay Kit (Q32851) with a Qubit 4 Fluorometer. Samples were sequenced using an Illumina HiSeq 2500 or Illumina NextSeq 2000 in 1×50, 2×50, or 1×100 mode (depending on other samples pooled with the sequencing library).

Insert-barcode linkage sequencing

Plasmid library pools 1-4 (pPBHS286, pPBHS309, pHPHS296, and pHPHS406) were diluted to 10 ng/μL. A 240-nucleotide region surrounding the 48-nucleotide variable insert sequence and the 24-nucleotide barcode was PCR amplified from these pools in two rounds, using Phusion Flash High-Fidelity PCR Master Mix mastermix (Thermo F548L). Round 1 PCR was carried out for 10 cycles, with 10 ng/μL plasmid pool template comprising 1/10th of the PCR reaction volume, using primers oPB29 and oPB354. Round 1 PCRs were digested with DpnI (Thermo FD1704) at 37°C for 30 minutes to remove template plasmid and cleaned using a 2× volume of Agencourt Ampure XP beads (Beckman Coulter A63880) to remove primers and enzyme. Cleaned samples

were used as template for Round 2 PCR, for 5 cycles, using oAS111 and indexed forward primers (oAS112-135 and oHP281-290). Amplified Round 2 PCR products were purified after size selection and quantified as described above for barcode sequencing. Samples were sequenced using an Illumina MiSeq or Illumina NextSeq 2000 in 2×50 or 1×100 mode.

Fluorescence-activated cell sorting and genomic DNA sequencing assay

Two 15 cm dishes of 75% confluent hsPB80 cells stably expressing the pHPHS286 library were used as input for fluorescence-activated cell sorting, using a BD FACSAria II flow cytometer. Fluorescence values of the first 50,000 sorted cells are plotted for reference in Fig. 2.4c. Fluorescence gates were determined using hsPB80 cells containing the pHPHS285 no-insert parent vector and untransfected hsPB80 cells as positive and negative controls for RFP and YFP fluorescence. Full gating strategy for the pHPHS286 library cells and pHPHS285 no-insert cells is available in Burke et al. 2022. 2.5M cells with ~10-fold or greater RFP expression relative to YFP were sorted into the low-YFP gate and gDNA was extracted from these cells, as well as from 2.5M unsorted cells from the same suspension, using 3 mL of gDNA lysis buffer. 4 μg of gDNA from each sample was used as input for gDNA barcode sequencing, following the procedures detailed above. Barcodes in each sample were quantified as described in the computational methods below. Low-YFP gate enrichment for each dipeptide insert was calculated as the log2 ratio of the summed low-YFP barcode counts to the summed unsorted barcode counts.

Rabbit reticulocyte nanoluciferase transit time assay

DNA fragments encoding 4×KVFS and 4×SKVF (α helix) and 4×VKFS and 4×SVKF (β strand) peptides were generated by PCR-amplifying overlapping oligos that encode each sequence in the forward and reverse direction (oPB470-473 and oPB488-491). Nanoluciferase cassette was amplified from an IDT gBlock (oPN204) using oAS1287 and oPB465. Insert sequences and the Nanoluciferase cassette were combined by overlap PCR using oPB464 and oPB462, which add

a 5′ T7 promoter site and a 3′ polyA tail to the amplified reporter template, with oAS1545 used to bridge oPB462 annealing. Resulting insert-Nanoluciferase cassette sequences were confirmed by Sanger sequencing. The PCR products were transcribed into mRNA using a HiScribe T7 High Yield RNA Synthesis Kit (NEB E2040S). mRNA was cleaned using an RNA Clean and Concentrator kit (Zymo R1013). *In vitro* Nanoluciferase reporter translation reactions were performed as described in Susorov et al. 2020[184]. Reaction mixture containing 50% of nuclease-treated rabbit reticulocyte lysate (RRL) (PRL4960, Promega) was supplemented with 30 mM Hepes-KOH (pH = 7.5), 50 mM KOAc, 1.0 mM $Mg(OAc)_2$, 0.2 mM ATP and GTP, 0.04 mM of 20 amino acids (PRL4960, Promega), and 2 mM DTT. Nanoluciferase substrate furimazine (PRN1620, Promega) was added to the mixture at 1%. 15 μL aliquots of the mixture were placed in a 384-well plate and incubated at 30°C for 5 min in a microplate reader (Tecan INFINITE M1000 PRO). Translation reactions were started by simultaneous addition of 3 μL mRNA, to a final concentration of 10 ng/μL, and luminescence signal was recorded every 10 seconds over a period of 25 minutes.

Circular dichroism

4×SKVF (α helix) and 4×SVKF (β strand) peptides were commercially synthesized (Genscript) at >90% purity level. Peptides were dissolved in water to 400 μM concentration, then diluted into 10 mM sodium-phosphate buffer (pH = 7.5) and 0, 20, or 40 volumetric percent of 2,2,2-trifluoroethanol (TFE) to final concentrations ranging between 15-30 μM. CD spectra were measured at 25C using a Jasco J-815 Circular Dichroism Spectropolarimeter. The CD spectra were recorded between 180-260 nm with a resolution of 0.5 nm for both peptides and blank buffer solutions in 1 mm cuvettes.

Computational analyses

Pre-processing steps for high-throughput sequencing were implemented as `Snakemake` workflows[185]. `Python` (v3.7.4) and `R` (v3.6.2) programming languages were used for all analyses unless mentioned otherwise. In the description below, files ending in `.py` refer to `Python` scripts

and files ending in .Rmd or .R refer to R Markdown or R scripts. All scripts are provided as a Supplementary file (code.tar.gz).

Barcode to insert assignment

The raw data from insert-barcode linkage sequencing are in .fastq format. If the inserts and barcodes were on paired-end reads instead of single-end reads, the reads were renamed in increasing numerical order starting at 0 to enable easy matching of insert and barcode reads. This was done in rename_fastq_paired_reads.py. The oligo pools were used to create a reference fasta file in create_reference_for_aligning_library.R. A bowtie2[186] (v2.4.2) reference was created from the fasta file using the bowtie2-build command with default options. The insert read was aligned to the bowtie2 reference using bowtie2 command with options -N 1 -L 22 --end-to-end with the --trim5 and --trim3 options set to include only the region corresponding to the insert. The alignments were sorted and indexed using samtools[187] (v1.11) commands sort and index with default options. The alignments were filtered to include only reads with simple cigar strings and a MAPQ score greater than 20 in filter_alignments.R. The barcodes corresponding to each filtered alignment were parsed and tallied in count_barcode_insert_pairs.py. Depending on the sequencing depth, only barcodes that were observed at least 4-10 times were included in the tally. The tallied barcodes were aligned against themselves using bowtie2-build with default options and bowtie2 with options -L 24 -N 1 --all --norc. The self-alignment was used to exclude barcodes that are linked to distinct inserts or ones that are linked to the same barcode but are aligned against each other by bowtie2. In the latter case, the barcode with the lower count is discarded. The final list of insert barcode pairs is written as a tab-delimited .tsv.gz file for aligning barcodes from genomic DNA and mRNA sequencing below.

Barcode counting in genomic DNA and mRNA

The raw data from sequencing barcodes in genomic DNA and mRNA is in .fastq format. The

filtered barcodes `.tsv.gz` file from the insert-barcode linkage sequencing is used to create a reference `fasta` file in `create_bowtie_reference.R`. A `bowtie2` (v2.4.2) reference was created from the `fasta` file using the `bowtie2-build` command with default options. The barcodes were aligned to the `bowtie2` reference using `bowtie2` command with options `-N 1 -L 20 --norc` with the `--trim5` and `--trim3` options set to include only the region corresponding to the barcode. The alignments were sorted, indexed, and tallied using the `samtools` commands `sort`, `index`, `idxstats` with default options. GNU `awk` (v4.1.4) was used for miscellaneous processing of tab-delimited data between pre-processing steps. The final list of counts per barcode in each sample of genomic DNA or mRNA is written as a tab-delimited `.tsv.gz` file for calculating mRNA levels below.

mRNA quantification

All barcode counts corresponding to each insert in each sample were summed. Only inserts with a minimum of 200 reads and 6 barcodes summed across the mRNA and gRNA samples were included. Otherwise the data were designated as missing. mRNA levels were calculated as the log2 ratio of the summed mRNA barcode counts to the summed gRNA barcode counts. mRNA levels were median-normalized within each library. For mRNA stability measurements, the summed mRNA counts for each insert at each time point were normalized by the total barcode counts for the spiked-in HeLa cells at the same time point. Then, the spike-in normalized mRNA levels for each insert were further normalized to the time 0 value.

Linear statistical modeling of mRNA levels

Amino acid scales for isoelectric point pI, bulkiness, and secondary structure propensity were taken from prior studies[124,129,188]. The median-normalized mRNA levels for lysine, arginine, or glutamate dipeptides were modeled as a function of amino acid scales (as indicated in the figures) using the `R` function `lm` with default parameters. Only fit coefficients significantly different from zero ($P < 0.05$)

are reported for each linear model.

Secondary structure prediction

Secondary structure was predicted solely from the amino acid sequence using the default single sequence model in S4PRED[127] (downloaded from https://github.com/psipred/s4pred on Apr 17, 2021) and the neural network was used without any modification in `predict_secondary_structure.py`.

Cartoons of 4×SVKF and 4×SKVF in Fig. 2.7a were predicted using the PEP-FOLD3 server[189] with default parameters and the resulting PDB files were visualized using PyMOL software (Schrodinger).

Calculation of secondary structure content from circular dichroism

The raw circular dichroism data (Fig. 2.7b, left panel) were converted to the two-column spectrum file format as required for `SESCA`[190] (v095, downloaded from `https://www.mpibpc.mpg.de/sesca` on Jul 28, 2021). Secondary structure was estimated using the `SESCA` script `SESCA_deconv.py` using the pre-computed basis set `Map_BB_DS-dTSC3.dat` and options `@err 2 @rep 100`. The output `.txt` file was parsed to extract the α helix, β strand, and random coil content shown in Fig. 2.7b, right panel.

Calculation of ribosome transit time

The raw luminescence vs. time data (Fig. 2.7c, middle panel) were fit to a straight line in the linear regimes (600s < t < 900s for 4×SKVF and 4×KVFS, 900s < t < 1200s for 4×SVKF and 4×VKFS) using the `R` function `lm`. The `intercept` term from the fit was used as the transit time of ribosomes across the full transcript and its mean and standard error across technical replicates is shown in the Fig. 2.7c right panel.

Statistical analyses

For barcode sequencing, error bars were calculated as the standard deviation of 100 bootstrap samples of barcodes across the gRNA and mRNA samples. The standard deviation was measured

for the log2 mRNA levels calculated as described in the *mRNA quantification* section. For all other experiments, the standard error of the mean was calculated using the `std.error` function from the `plotrix` `R` package. P-values for statistical significant differences were calculated using the `t.test` or `wilcox.test` `R` functions as appropriate for each figure (see figure captions).

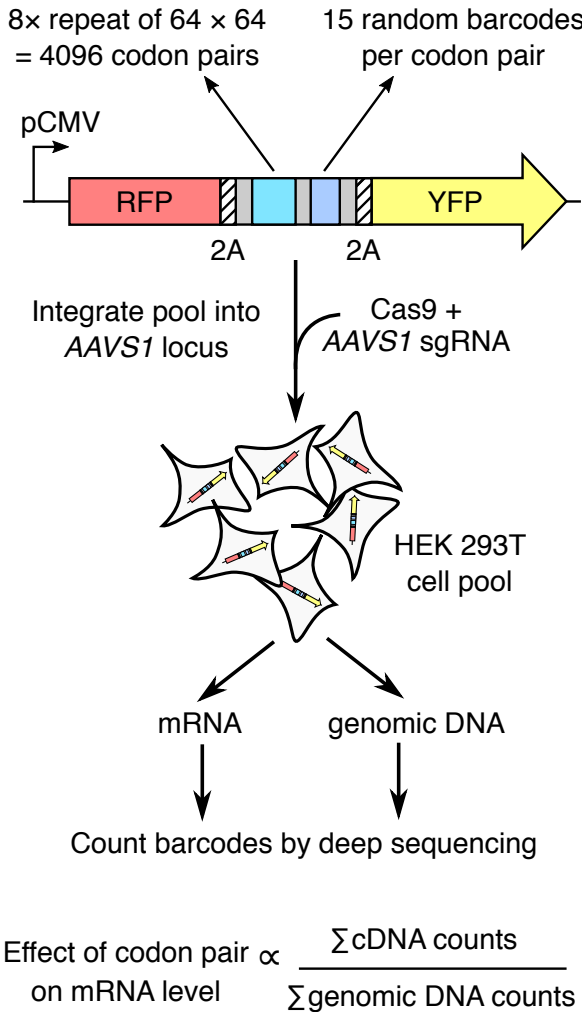
Data Availability

The raw sequencing data generated in this study have been deposited in the Sequence Read Archive under BioProject "PRJNA785998". Raw data from circular dichroism and luciferase assays are available at https://github.com/rasilab/burke_2022.

All code to reproduce figures in the manuscript starting from raw data is publicly available at https://github.com/rasilab/burke_2022.

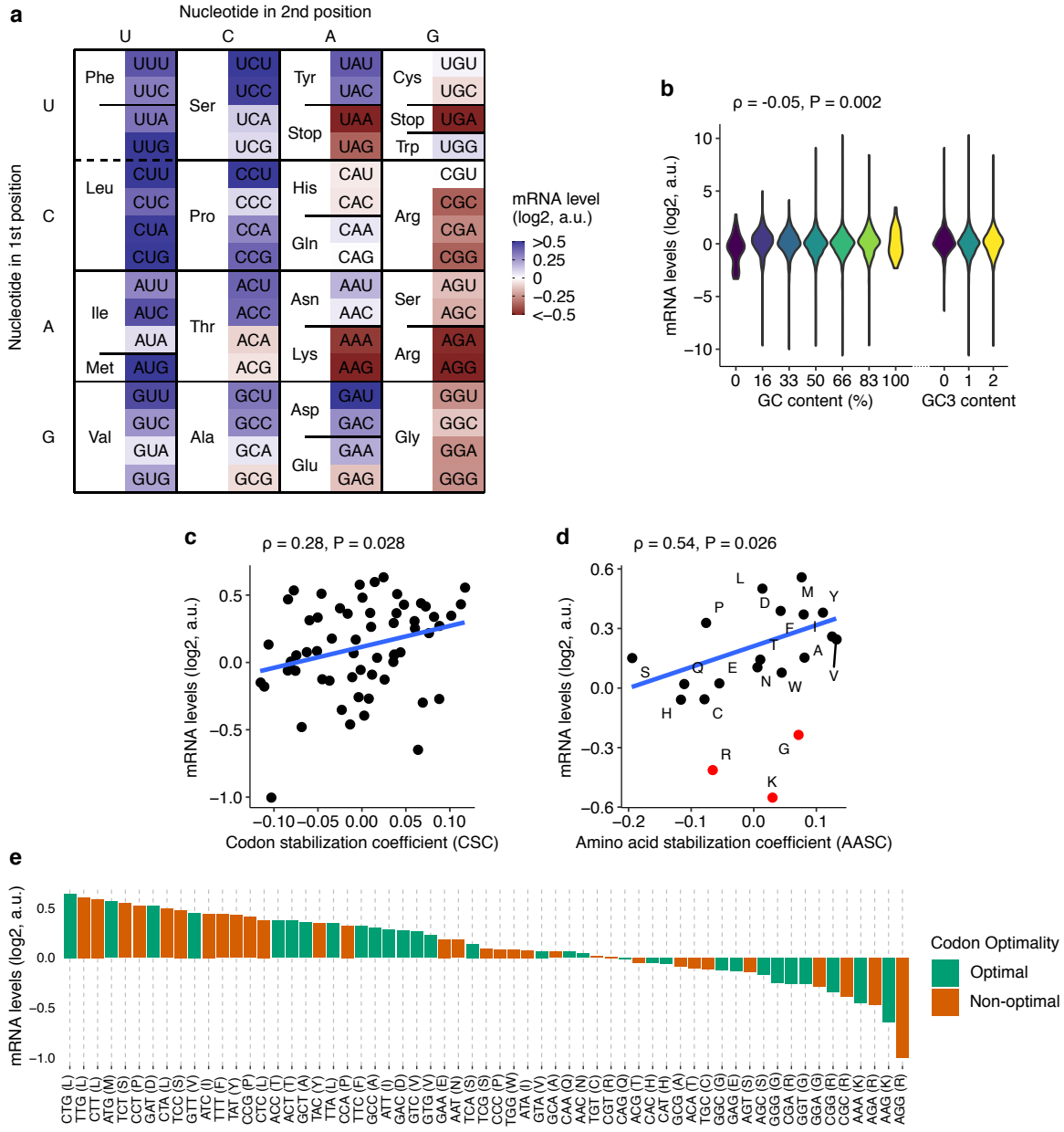
V. Figures

Figure 2.1



A massively parallel assay for mRNA levels in human cells
8× repeats of all 4096 codon pairs are synthesized as pooled oligonucleotides, linked in-frame to 24nt random barcodes, and cloned between RFP and YFP reporters with intervening 2A sequences. Each insert has a median of 15 random barcodes without in-frame stop codons. Reporter cassettes are integrated as a pool at the *AAVS1* locus in HEK293T cells by Cas9-mediated homologous recombination and constitutively expressed off the CMV promoter. Steady state mRNA level of each insert is determined by sequencing corresponding barcodes in the cDNA and the genomic DNA and normalizing the summed cDNA read counts by the genomic DNA read counts.

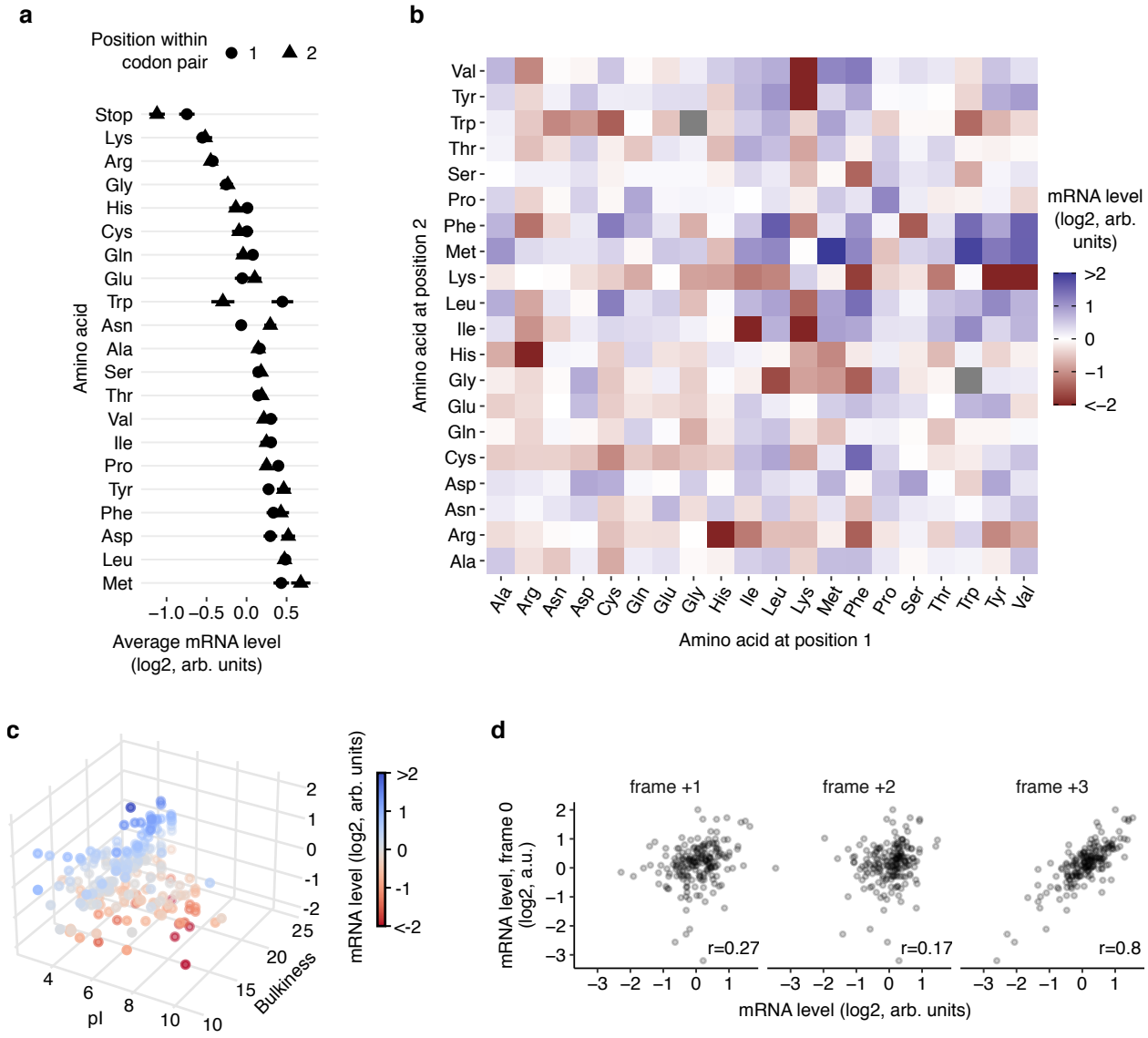
Figure 2.2



Effects of primary mRNA sequence on mRNA levels

(a) Heatmap of mean effect of each codon on mRNA levels averaged across both positions of 8x dicodon repeat. Values >0.50 were set to 0.51, and values <-0.5 were set to -0.51 to highlight differences in intermediate values. **(b)** mRNA levels of dicodon repeats as a function of their GC content (left) or the number of GC3s in the dicodon (right). Spearman rank correlation coefficient ρ and its P-value are shown for GC content (two-sided t-test); GC3 content had no significant correlation with mRNA levels. **(c)** Average codon effects on mRNA levels as a function of their mean codon stabilization coefficient[111]. Pearson correlation coefficient ρ and its P-value are shown (two-sided t-test). **(d)** Average amino acid effects on mRNA levels as a function of their mean amino acid stabilization coefficient[112]. Pearson correlation coefficient ρ and it's P-value are shown, calculated from the points in black. Points in red (arginine, lysine, and glycine) are excluded from this Pearson correlation calculation, as these amino acids' average effects are likely driven by specific dipeptide combinations (see Discussion section). **(e)** Average codon effects on mRNA levels plotted from high to low, with codons colored by HEK293T codon optimality[111].

Figure 2.3



Dipeptide repeats reduce mRNA levels

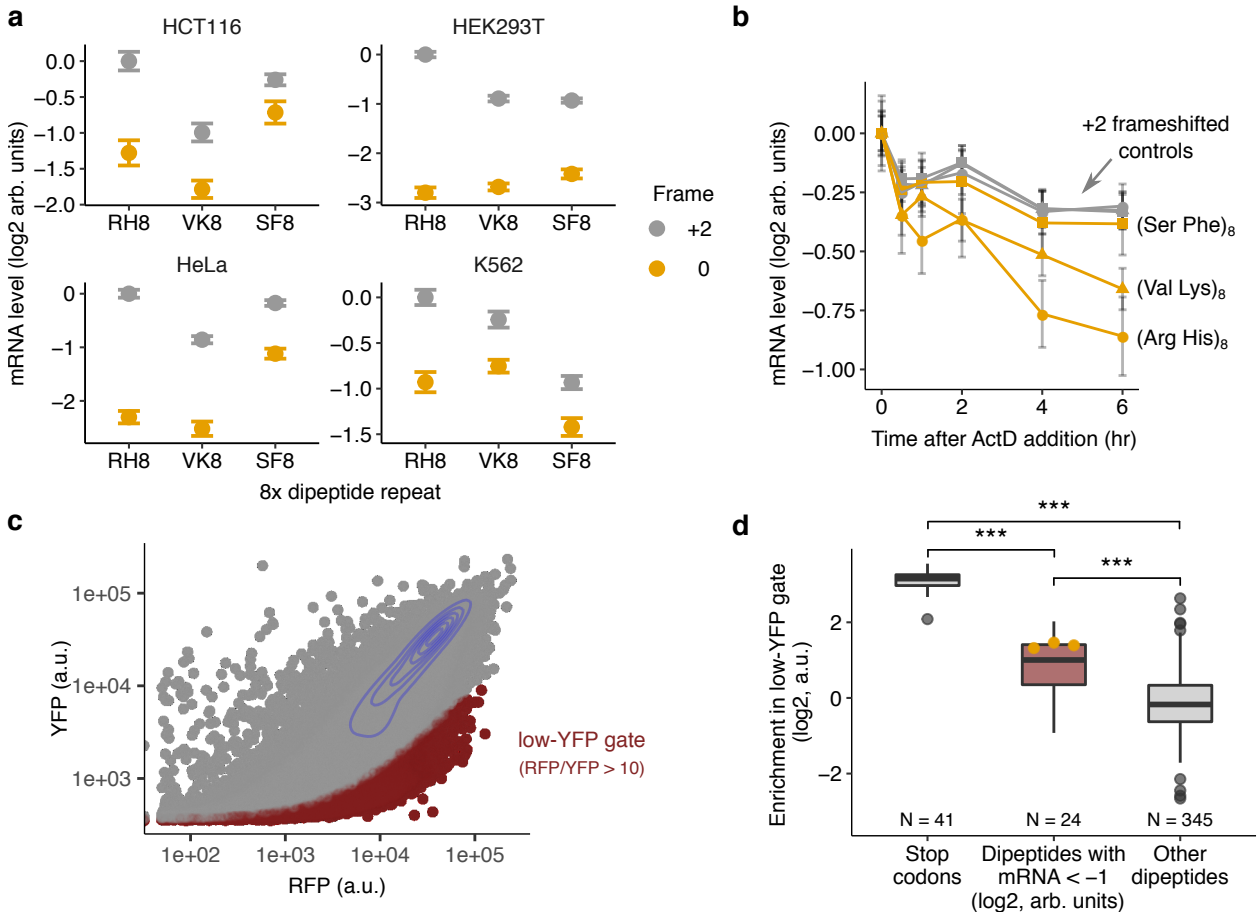
(a) mRNA level of reporters with codons encoding one of the twenty amino acids or a stop codon in position 1 (circles) or position 2 (triangles) of the 8× codon pair insert shown in Fig. 2.1.

(b) mRNA level of reporters encoding 400 different dipeptide repeats. Amino acids encoded by the first or second position in the dipeptide are shown along the horizontal or vertical axis respectively. Two dipeptide repeats with missing values are shown in grey.

(c) mRNA level of dipeptide repeat-encoding reporters plotted as a function of the average isoelectric point (pI) and the bulkiness[124] of the two amino acids in the dipeptide.

(d) mRNA level of reporters encoding 190 different dipeptide repeats (excluding reversed repeats) in the correct reading frame (frame 0, vertical axis) or in reading frames shifted by +1, +2, or +3 nucleotides (horizontal axes). *r* is the Pearson correlation coefficient between frame 0 and the frameshifted mRNA levels. mRNA levels in **a–d** are in arbitrary units (arb. units) and are normalized to the median value across all dipeptide repeats. Data in **a** are presented as mean values and error bars represent +/- standard error of measurement (SEM) over a median of 15 barcodes per insert calculated using 100 bootstrap samples. Most error bars in **a** are smaller than data markers.

Figure 2.4



Dipeptide repeats reduce mRNA stability and cause premature termination in human cells

(a) mRNA levels of reporters with dipeptide repeats (orange): $(Arg\ His)_8$, $(Val\ Lys)_8$, $(Ser\ Phe)_8$, or the three +2 frameshift controls (grey): $(Pro\ Ser)_8$, $(Gln\ Ser)_8$, $(Phe\ Gln)_8$ in 4 different cell lines: HCT116, HEK293T, HeLa, and K562.

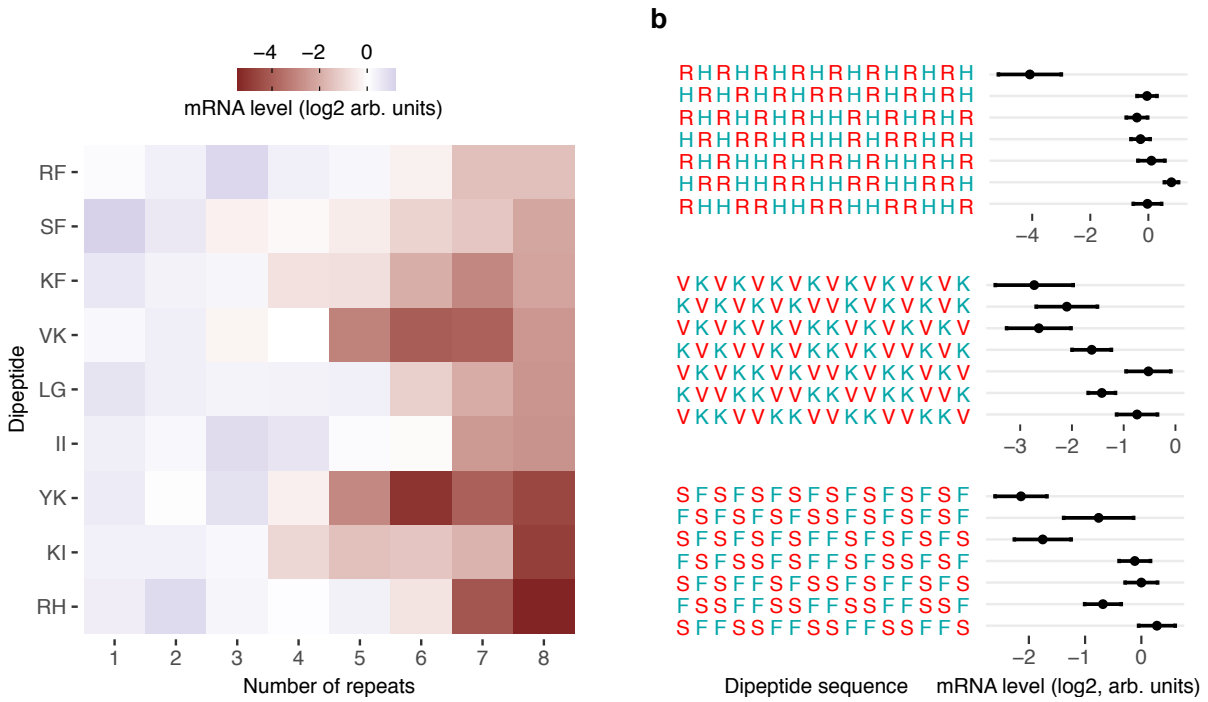
(b) mRNA stability of reporters from **a** in HEK293T cells. Reporter mRNA levels are measured at indicated time points after Actinomycin D-induced transcriptional shut off. Most points for the three frameshift controls overlap with each other.

(c) Gating strategy for fluorescence-activated cell sorting of HEK293T cells expressing the stably integrated 8× dicodon library from Fig. 2.1. Cells with low ratio of YFP/RFP were sorted into the "low-YFP" bin (dark red points).

(d) Enrichment of dipeptide inserts in the low-YFP gate is shown for three subgroups; inserts encoding in-frame stop codons, dipeptide repeats with mRNA level < -1 (log2, a.u.) in Fig. 2.3b, and all other inserts. Enrichment values for the RH8, VK8, and SF8 dipeptides are highlighted in orange (RH8, VK8, SF8; left to right). The bounds of the box plots are the upper and lower quartile with the median value in the center. Whiskers extend to the most extreme data point no more than [1.5] times the length of the box away from the box. Outliers extending further than the whiskers are shown as individual data points. ***: $P < 0.001$ (two-sided Mann-Whitney U test) for differences between subgroups. Stop vs mRNA < -1 log2 arbitrary units; P value = 2e-16. Stop vs Other; P value = 2e-16. mRNA < -1 log2 arbitrary units vs Other; P value = 2e-7.

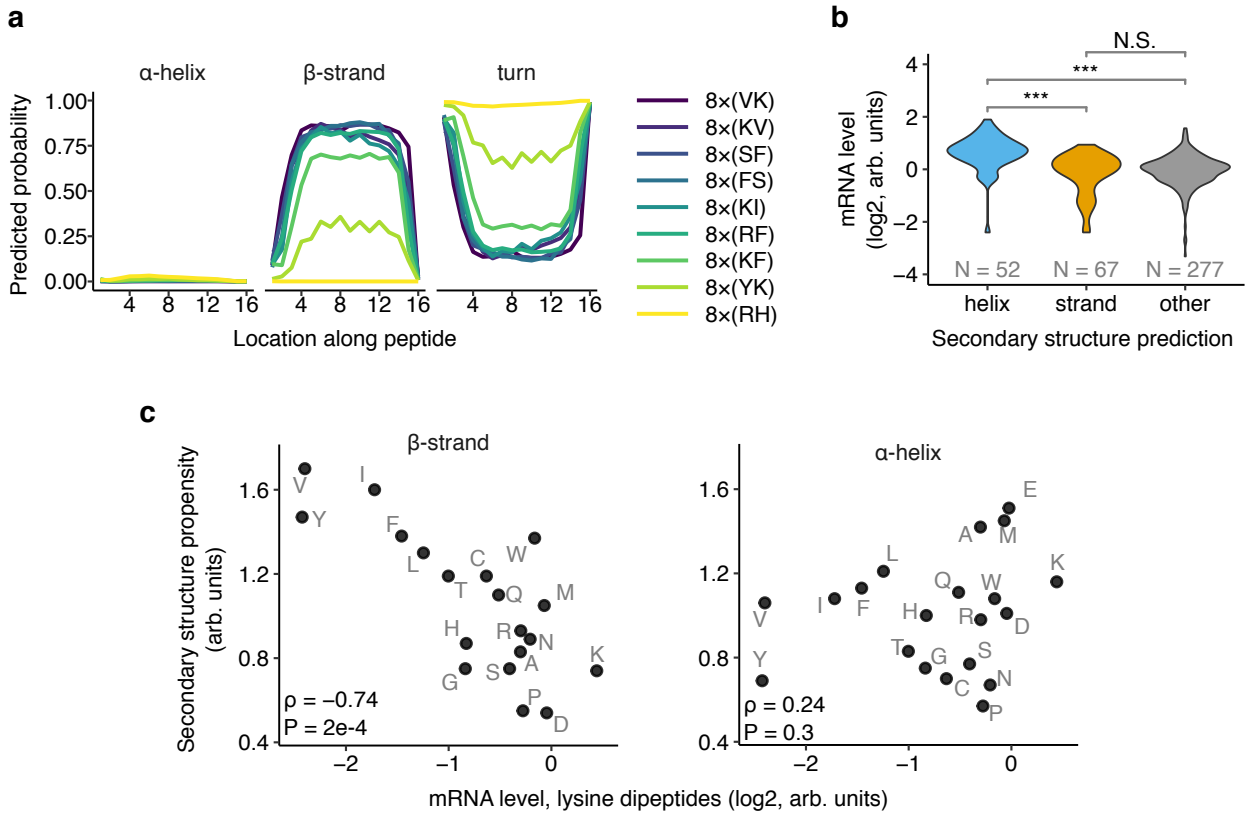
mRNA levels are measured using the pooled sequencing assay in Fig. 2.1 and normalized by the median value across all inserts in the pool. Amino acids in **a** and **b** are labeled by their one-letter codes. Data in **a** and **b** are mean values +/- SEM, calculated over a median of 550 and 370 barcodes per insert respectively, using 1000 bootstrap samples each.

Figure 2.5



Nascent peptide primary sequence modulates mRNA level effects
(a) mRNA levels of dipeptide-encoding reporters with different dipeptide repeat length. Missing values shown in grey. **(b)** mRNA levels of dipeptide-encoding reporters with different dipeptide repeat periodicity. mRNA levels are measured using the pooled sequencing assay in Fig. 2.1 and normalized by the median value across all inserts in the pool. Amino acids are labeled by their one-letter codes. Data in **b** are mean values +/- SEM, calculated over a median of 15 barcodes per insert using 100 bootstrap samples.

Figure 2.6

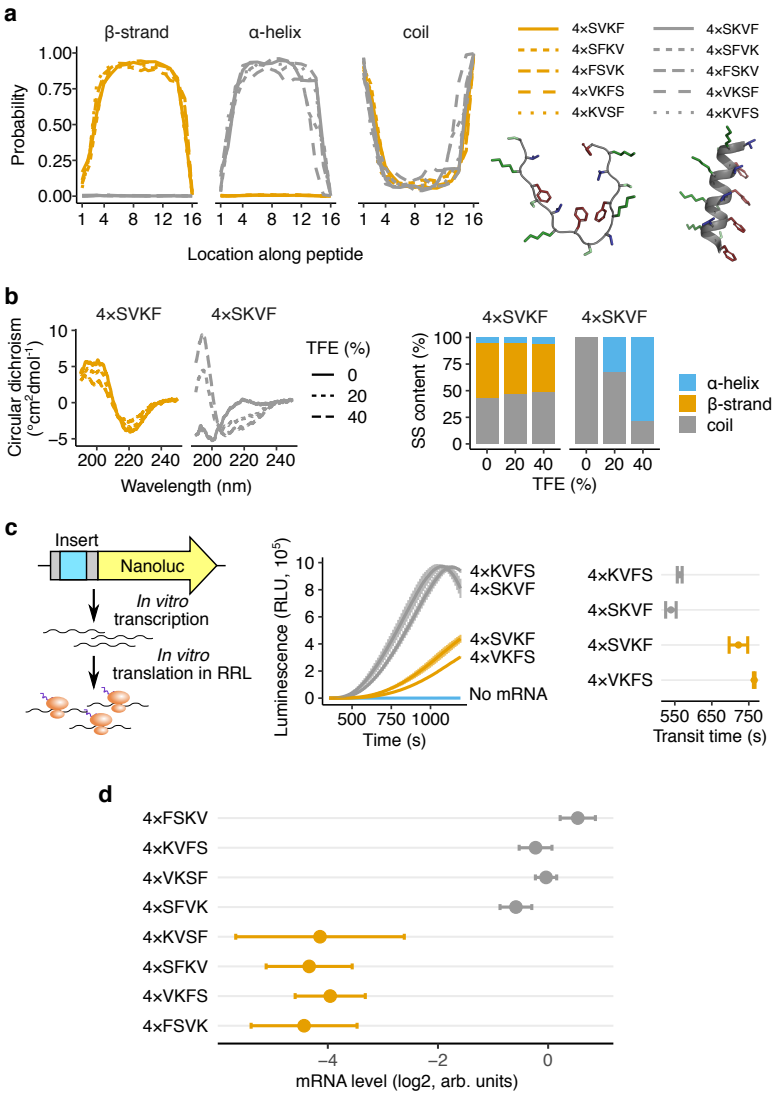


Secondary structure of dipeptide repeats mediates effects on mRNA levels

(a) Computationally predicted secondary structure probability along 16 amino acid-long peptide sequences encoded by destabilizing dipeptides. Secondary structure probabilities are predicted using S4PRED[127].

(b) Distribution of mRNA levels of dipeptide repeat-encoding reporters from Fig. 2.3b partitioned by predicted protein secondary structure. Per residue probability of secondary structure formation are predicted using S4PRED. Inserts with >50% average prediction probability of forming α helix or β strand are classified as such, or else grouped as 'other'. *N* is the number of dipeptide repeats predicted to be in each category. ***: $P < 0.001$, N.S.: not significant (two-sided Mann-Whitney U test). Helix vs Strand; P value = 2.2e-9. Helix vs Other; P value = 7.3e-14. Strand vs Other; P value = 0.92. **(c)** mRNA levels of dipeptide repeat-encoding reporters with lysines in one position and one of twenty amino acids in the other position of the repeat (labeled in grey) shown on horizontal axes. Propensity[129] of the second amino acid to occur in a β strand or an α helix is shown on vertical axes. ρ is the Spearman correlation coefficient between the two axes with the indicated *P* value (two-sided Spearman rank correlation test).

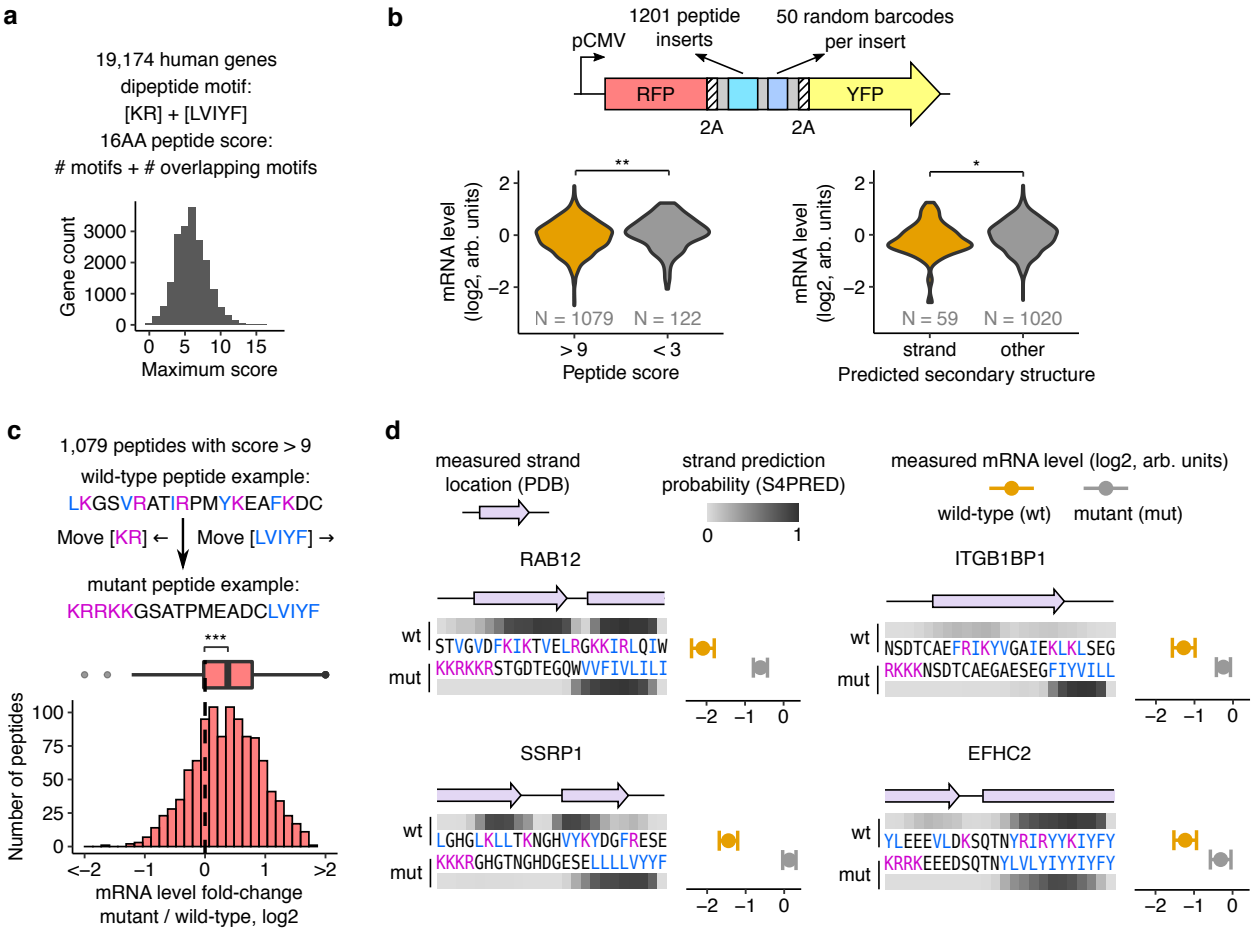
Figure 2.7



Extended β strands slowdown ribosomes and reduce mRNA levels

(a) Computationally predicted secondary structure probability along 16 amino acid-long peptide sequences encoded by alternating VK or KV dipeptides with SF or FS dipeptides identified as destabilizing in Fig. 2.3. Secondary structure probabilities are predicted using S4PRED. The 10 different peptide sequences are 4× repeats of the dipeptide combinations shown in the legend (for example, 4×SVKF: SVKF SVKF SVKF SVKF). Predicted β strand and α helix structures of 4×SVKF and 4×SKVF respectively using PEP-FOLD3[189] are shown below the legends. **(b)** Measured circular dichroism spectra of *in vitro* synthesized 4×SVKF or 4×SKVF peptides (left). Measurements are performed with 0, 20, 40% of Trifluoroethanol (TFE) as co-solvent in 10 mM sodium phosphate buffer (pH = 7.5). Relative content of different secondary structures is estimated by linear deconvolution of the measured spectra in B from a pre-computed basis set using SESCA[190]. **(c)** *In vitro* measurements of ribosome transit time on mRNAs encoding β strand- or α helix-forming peptides followed by Nanoluciferase. Luminescence is measured as a function of time after addition of *in vitro* transcribed mRNAs to rabbit reticulocyte lysate (RRL) at t=0*s* (E). Standard error of measurement across three technical replicates is shown as a shaded area on either side of the mean. Ribosome transit times are estimated by measuring the X-intercept of the linear portion of the raw luminescence signal. **(d)** *In vivo* mRNA levels of reporters encoding one of eight different dipeptide combinations. mRNA levels are measured using the reporter constructs and pooled sequencing assay in Fig. 2.1. Data are presented as mean values +/- SEM over a median of 550 barcodes per insert calculated using 1000 bootstrap samples.

Figure 2.8



Dipeptide motifs in the human genome reduce mRNA levels

(a) Scoring methodology for dipeptide motifs in human CDS. Destabilizing dipeptides formed by lysine (K) or arginine (R) with an adjacent leucine (L), valine (V), isoleucine (I), tyrosine (Y), or phenylalanine (F) are given a score of 1. If two such dipeptides overlap, an additional score of 1 is given. The 16 amino acid peptide window with the maximum score is identified in each CDS, and the distribution of these peptide scores across all genes is shown in the lower panel.

(b) mRNA levels of destabilizing (peptide score > 9) and control motifs (peptide score < 3) are measured by pooled cloning (1,201 total inserts) into a reporter construct followed by deep sequencing as in Fig. 2.1. Left panel: Distribution of measured mRNA levels of destabilizing dipeptide motifs compared to control motifs (P = 0.004). Right panel: Distribution of measured mRNA levels of destabilizing dipeptide motifs in the left panel partitioned by predicted secondary structure (P = 0.022). 59 motifs with an average β strand prediction probability > 0.5 using S4PRED are classified as β strands.

(c) Increase in measured mRNA levels upon reordering amino acids in 1,079 endogenous destabilizing dipeptide motifs from C (median Δlog2 mRNA = 0.38, P = 2.2e-16). All codons encoding K or R are moved to the 5′ end of the mutated motif and codons encoding L, V, I, Y, or F are moved to the 3′ end. mRNA levels of motifs are measured using the pooled reporter assay in B. ***: P < 0.001, **: P < 0.01, *: P < 0.05 using two-sided Mann-Whitney U test in B and C.

(d) Examples of endogenous destabilizing motifs with known β-stranded secondary structure. The measured secondary structure of each wild-type motif from PDB is shown as a purple ribbon diagram. Prediction probability for β strands using S4PRED is shown as a grayscale heatmap for wild-type and mutant motifs. Measured mRNA levels of wild-type (orange) and mutant (grey) motifs are shown on the right within each panel. mRNA levels of wild-type and mutant motifs are measured using the pooled reporter assay in C, and presented as mean values +/- SEM over a median of 50 barcodes per insert calculated using 1000 bootstrap samples.

Chapter 3: RNA Viruses Encode mRNA Destabilizing Sequence Motifs

I. Introduction

Viruses pose a large global health burden, recently and acutely emphasized by the global pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[191]. Viral mRNAs are translated into protein by their host-cell translation machinery, which means human viruses are subject to the same pathways and rules that regulate translation of our own mRNAs[192]. RNA viruses have high mutation rates, allowing them to quickly adopt beneficial mutations, such as synonymous mutations that might improve their mRNA translation[193,194]. Despite this, many RNA viruses maintain suboptimal codon usage or GC-content for human translation, either in specific open reading frames[195,196], or genome-wide[197,198]. This raises the question of whether viral mRNAs are sensitive to mRNA degradation pathways in human cells in sequence dependent manner. Specific viral RNA sequences and proteins are required for viral replication, thus some viral sequence motifs are evolutionarily restricted, particularly at the peptide level[199–201]. Highly destabilizing sequences that are nevertheless present in these viral genomes may be important motifs for viral fitness, thereby presenting potential therapeutic targets or sites of particular evolutionary interest. In this regard, the high-throughput sequencing assay we developed in Burke et al. 2022 presents a unique method to profile viral genomes, allowing unbiased identification of mRNA destabilizing RNA or peptide motifs which may be of biological importance[128].

II. Results and Discussion

A massively parallel assay for mRNA level effects of viral sequence motifs

We hypothesized that identifying viral coding sequence regions that reduce the mRNA stability when translated might reveal sequence motifs that are important for viral fitness. To test this, we extended our recently published high-throughput sequencing assay from Burke et al. 2022 (Fig. 2.1) to study the effects of viral coding sequence motifs on mRNA levels in an unbiased manner[128].

We designed a library of 4500 oligonucleotides encompassing the full coding sequences of three RNA viruses important to human health; SARS-CoV2, human immunodeficiency virus (HIV), and influenza A virus (IAV). Oligos contained 75nt (25 amino acid) coding sequence fragments from these viruses, with each fragment having 66nt (22 amino acids) of tiled overlapping region with its neighboring upstream motif, such that our library encoded the entire CDS of each virus in stepwise increments of 9nt (3 amino acids). Viruses often encode multiple translated open reading frames within the same transcript, and sometimes these ORFs can overlap one another in separate reading frames. To account for this we treated each canonically annotated open reading frame as a separate instance in our library design. For example, in the case of overlapping ORFs in separate reading frames, we designed oligos starting in-frame at the start codon for both ORFs. This approach enabled us to profile the entire coding sequence of each viral ORF, known as the viral "ORFeome".

We followed the same library approach as in Burke et al. 2022, using a dual fluorescence reporter vector separated by 2A linkers and a median of 50 random 24nt barcodes without stop codons 3′ of each oligonucleotide insert and linked to the corresponding insert by high-throughput sequencing[128]. We stably integrated the reporter pool at the *AAVS1* locus of HEK293T cells using CRISPR Cas9-mediated homologous recombination. We extracted mRNA and genomic DNA from the pooled cells and counted each barcode by high-throughput sequencing. Normalization of the total barcode count in the mRNA by the corresponding count in the genomic DNA for each of the inserts provides a relative measure of the steady-state mRNA level of that insert. We then examined the measured mRNA level effects across each viral ORFeome to identify coding sequence motifs with the strongest deleterious effects on mRNA level, median normalized across all motifs in the library.

While the majority (92%) of all characterized motifs had effects within 2-fold (+/- 1 log2 a.u.) of the library median, many individual motifs displayed marked effects on mRNA levels. Overall the three

viral ORFeomes had motifs spanning over a 30-fold range of effects on mRNA level, with some motifs increasing relative mRNA levels up to 4-fold, and other motifs reducing mRNA levels by more than 8-fold. SARS-CoV2 effects are shown in Fig. 3.2, and motifs with mRNA levels < -1.5 log2 a.u (~3-fold reduced) are listed in Table 3.1. HIV effects are shown in Fig. 3.3, and motifs with mRNA levels < -1.75 log2 a.u are listed in Table 3.2. IAV effects are shown in Fig. 3.4, and motifs with mRNA levels < -1.5 log2 a.u are listed in Table 3.3. From these data we picked three highly mRNA destabilizing motifs from SARS-CoV-2 to explore further, based on known biology and high confidence from this ORFeome-wide assay. While we chose to focus on a few destabilizing motifs from SARS-CoV-2 for initial characterization, these unbiased measurements of ORFeome mRNA level effects may prove to be a useful resource for other researchers studying these viruses. Full mRNA level data are available upon request.

The role of Nsp1 in mRNA level effects

The first SARS-CoV2 motif we looked at was the ORF1ab_460 motif, which reduced mRNA levels by 6.5-fold relative to the median (Table 3.1). This motif is located at the C-terminus of the Nsp1 protein, which is one of the first proteins expressed during SARS-COV2 infection and plays a critical role in host-cell immune suppression[202,203]. Nsp1 inhibits mRNA translation through interactions between its C-terminal polypeptide region and the ribosome exit tunnel[202,204]. In addition to blocking translation, Nsp1 also globally affects RNA export and reduces host-cell mRNA stability[205–207]. ORF1ab_460 encodes the peptide sequence **Y**ED**F**QENWNT**KH**SSGVTRELM**R**ELN; mutational studies of the Nsp1 C-terminus demonstrated that the five residues highlighted in bold in the peptide sequence above are required for Nsp1 ribosome binding[204]. In addition, expression of this Nsp1 C-terminal region alone is sufficient for the ribosome binding and translation inhibition[202,204]. Based on this, we hypothesize that this ORF1ab_460 peptide motif represents the minimal functional unit for Nsp1 ribosome binding and host translation shutoff activity. Expression of this short inhibitory peptide in our assay could interfere with translation and globally reduce cellular mRNA

levels in *trans*, by increasing mRNA turnover through canonical decay or quality control pathways. In our assay, global reduction in mRNA levels would be read out as reduced cDNA counts relative to gDNA for the cells expressing this peptide. In addition, this peptide may interact with the exit tunnel and interfere with its own translation in *cis*, in a manner similar to known ribosome arrest peptides[149]. If translation of the Nsp1 C-terminus does reduce its encoding mRNA levels in *cis*, the resulting reduction of Nsp1 C-terminus protein expression could potentially abrogate any global effects on mRNA stability in cells, but in either scenario we would expect to see reduced mRNA levels from ORF1ab_460. It is also worth noting that SARS-CoV-2 is capable of expressing Nsp1 during infection regardless of any *cis* or *trans* acting effects this peptide might have on its own translation. Overall, the clean signal derived from this Nsp1 C-terminal peptide region demonstrates the ability of our high-throughput assay to identify viral sequence motifs that impact mRNA stability and are critical for viral function in an unbiased manner.

The role of Nsp10 in mRNA level effects

We next looked at the 3' end of ORF1ab, which contains an extended 129nt region where all the tiled motifs were more than 2-fold destabilizing. This destabilizing region starts from ORF1ab_12979 and ends at ORF1ab_13033 and encodes the 43 amino acid sequence CLYCRCHIDHPNPKGFCDLKGKYVQIPTTCANDPVGFTLKNTV, which corresponds to the C-terminal region of the Nsp10 peptide. Nsp10 is a scaffolding protein required for the activity of the Nsp14 and Nsp16 methyltransferases, which are essential for SARS-CoV-2 replication. Based on tiled motif effects, the amino acids LKGKYVQ in the center of this sequence are the only residues common to all destabilizing peptides in this region. The structure of Nsp1 protein has been solved in complex with Nsp14 and Nsp16[208]. The YVQ residues of LKGKYVQ mark the start of a 5 amino acid long beta strand, and the upstream LKGKY residues display charged and bulky dipeptide periodicity we previously found to be associated with ribosome slowdown and mRNA instability[128,208]. We hypothesize that this C-terminal Nsp10 region may cause nascent-peptide

mediated stalling, resulting in mRNA instability. During infection, this effect may be abrogated by upstream protein folding that occurs during translation of the full length Nsp10 protein. However, it is also possible that this motif cotranslationally limits SARS-CoV-2 mRNA levels, either by design or necessity for viral fitness.

The role of ORF3a in mRNA level effects

We next looked at the N-terminal region of ORF3a, encoding the destabilizing 28 amino acid motif MDLFMRIFTIGTVTLKQGEIK**D**ATPSDF. Interestingly, this N-terminal region overlaps with the recently discovered ORF3c, which begins in the +1 reading frame at the aspartic acid residue (D) highlighted in bold in the peptide motif above. Of note, ORF3c was not annotated as a SARS-CoV-2 ORF at the time of our library design, so it was not included in our profiling approach. ORF3c is predicted to be translated through a leaky scanning mechanism where the ribosome fails to initiate at the start of ORF3a[209,210]. However, a mechanistic basis for stochastic failure of initiation at ORF3a has not been characterized. Based on the tiled ORF3a_1 and ORF3a_10 motifs both having effects, the 22 amino acid FMRIFTIGTVTLKQGEIKDATP peptide appears to be the minimal unit required for mRNA effects. We hypothesize that ribosomes pause while synthesizing the C-terminal threonine-proline peptide bond, facilitated by upstream nascent peptide interactions with the ribosome exit tunnel. While ribosome slowdown leads to mRNA instability in our assay, ribosome pausing is also a known mechanism for enabling the translation of alternate open reading frames[34]. This potential ORF3a_1 stalling mechanism is reminiscent of the human cytomegalovirus (hCMV) uORF2, wherein a combination of nascent peptide interactions with the exit tunnel cause ribosomes to pause at a Pro-Pro-Stop motif while translating this uORF, and leaky scanning past the uORF2 AUG initiation codon results in increased expression of a downstream cistron[39,149,162,211]. In the case of SARS-CoV-2 ORF3a, ribosome pausing at this destabilizing peptide motif may facilitate translation of ORF3c, potentially by a ribosome queueing mechanism resulting in leaky scanning[212], or by ribosome collision induced frameshifting[213,214].

Testing effects of SARS-CoV2 peptide sequences on mRNA level

The abundance of positively charged and bulky amino acids in these destabilizing viral motifs (Tables 3.1, 3.2, 3.3), implies that these viral sequences may be subject to the same nascent peptide-mediated mechanisms that we find limit mRNA levels for synthetic and endogenous human sequence motifs[128]. To further characterize the effects of these SARS CoV2 motifs on mRNA levels, we designed mutant versions of these peptides where we rearranged the bulky and positively charged residues to the peptide termini (Fig 3.5a). This allowed us to maintain the codon and amino acid usage of these motifs but disrupt the mRNA and peptide sequence secondary structure. We tested both the wildtype and mutant inserts cloned as a pool both into our established *AAVS1* integration reporter, and into an inducible expression reporter designed for BxbI serine recombinase-mediated integration into engineered landing pad cells (Fig 3.5a). mRNA level effects were max normalized within the oligo pool, which also contained other wildtype and mutant viral motifs, as well as several control sequences with destabilizing or neutral effects on mRNA level, previously tested in the pooled AAVS1 assay[128].

As expected, the wildtype viral motif sequences reduced mRNA levels in both the AAVS1 and BxbI landing pad reporter systems (Fig 3.5a). Interestingly, while mutations of each peptide rescued mRNA levels relative to their wildtype counterparts, these mutant sequences were also somewhat destabilizing. This implies a hybrid mechanism where these motifs' effects are partially driven by codon and amino acid usage, while the full effect of each requires the correct primary mRNA or peptide sequence. This is somewhat surprising for the ORF1ab_460 motif, which we expected to globally reduce mRNA levels in a peptide-dependent manner by binding the ribosome exit tunnel and causing promiscuous stalling throughout the translatome. However, this motif does encode a high abundance of both charged and bulky residues, which could result in translation inhibition in *cis*, regardless of the order of the residues. Overall, these results suggest that codon and amino acid usage, as well as primary mRNA and peptide sequence, all contribute to the mRNA stability

effects of these SARS-CoV-2 motifs.

Future Directions

Identification of destabilizing sequence motifs in the SARS-CoV-2, HIV, and IAV genomes is just the starting line for this research. Our ultimate goal for this project is to test the requirement of these motifs for viral fitness, by generating viruses with mutations at these motifs. The large-scale peptide rearrangements we made to test these viral motifs were useful for initial validation of peptide effects, however these mutations are not ideal for testing viral fitness, as adjusting a peptide sequence so severely would likely have a deleterious effect on encoded protein function (regardless of the effects of the mutation on mRNA stability in *cis*). Instead, the ideal scenario would be to test single amino acid substitutions for each motif that rescue mRNA levels in our high-throughput assay. To find single amino acid mutations that can abrogate the mRNA destabilizing effects of these motifs, we will perform a deep mutational scan of the three SARS-CoV-2 motifs highlighted in Chapter 3, as well as the most destabilizing HIV and IAV sequence motifs. We will then test the effects of these mutations on viral fitness in collaboration with the Bloom, Emmerman, and Hyde labs in the University of Washington Department of Microbiology.

III. Materials and Methods

Plasmid construction

Information on plasmids, oligonucleotides, and cell lines used in this study are available on github: https://github.com/rasilab/viral_mrna_stability.

Parent vector construction

The AAVS1-targeting parent vector pPBHS285 used for this study was constructed using Addgene plasmid #68375[179] as a backbone. The PGK1 promoter was replaced with the CMV promoter and the native pCMV 5′ UTR region. The coding sequence was replaced by a codon-optimized mKate2 and eYFP fusion cassette, linked with two 2A linker sequences. These 2A sequences surround a

cassette encoding an EcoRV restriction site, Illumina R1 sequencing primer binding site, and a T7 promoter. The R1 primer binding and T7 sequences are in reverse orientation (3′ - 5′) for *in vitro* transcription and sequencing of inserts and barcode sequences at the EcoRV site.

The Landing Pad integration parent vector pHPHS482 was generated by amplifying the mKate2-2A-EcoRV-2A insert from pPBHS285 (including the R1 and T7 sequence region) using oPB497 + oPB498, and cloning this upstream of eYFP in a pTET-EcoRI-eYFP-Bxb1attB-PuroR vector (pH-PHS476), originally constructed using the Bxb1 integration plasmid pKAM33 which was a kind gift from the Fowler lab[215].

Variable oligo pool design

Two oligo pools were designed for this study.

Pool 1 (Fig. 3.1) encodes 25 amino acid nascent peptide motifs from the SARS-CoV-2, IAV, and HIV ORFeomes (all annotated open reading frames). In frame coding sequences from all viral ORFs were tiled by 3 amino acids, starting at the start coding on each ORF, such that the full coding sequence of each virus was profiled in increments of 3 codons.

Pool 2 (Fig. 3.5) contains 20 top destabilizing wildtype viral coding sequence motifs picked from Pool1 profiling (Figs. 3.2, 3.3, 3.4), as well as mutants with reordered amino acids for each of these 20 sequences (designed as described in Fig. 3.5a). This pool included the three SARS-CoV-2 motifs highlighted in Fig. 3.5b). In addition, this pool contained 8 control inserts characterized in Burke et al. 2022: 3 destabilizing dipeptide repeats $(RH)_8$, $(VK)_8$, $(SF)_8$, their respective frameshift controls $(PS)_8$,$(QS)_8$,$(FQ)_8$, the β strand peptide $(SVKF)_4$, and the α helix peptide $(SKVF)_4$.

Oligo pools were synthesized by Twist Biosciences with flanking sequences for PCR and cloning into the EcoRV site of the parent pPBHS285 or pHPHS482 vector. Plasmids encoding the control sequences in Pool 2 were cloned and pooled previously[128], and were spiked into the Pool 2 viral motif oligo pool at a 1:5 concentration ratio.

Final plasmid pool identities: pHPHS296 is the viral ORFeome *AAVS1* integration plasmid pool; pHPHS491 is the Pool 2 landing pad integration plasmid pool; pHPHS492 is the Pool 2 *AAVS1* integration plasmid pool.

Integration vectors

The CLYBL-targeted Cas9-BFP expression vector pHPHS15 was constructed by Golden Gate assembly of either entry plasmids or PCR products with pHPHS11 (MTK0_047[180] Addgene #123977) as backbone, pHPHS3 (MTK2_007[180] Addgene #123702) for the pEF1a promoter, pADHS5[49] (pU6-(BbsI)_CBh-Cas9-T2A-BFP[181] Addgene #64323) for the Cas9-2A-BFP insert cassette, and pHPHS6 (MTK4b_003[180] Addgene #123842) for the rabbit β-globin terminator. sgRNA vectors pPBHS320 (gRNA_AAVS1-T1 Addgene #41817) and pADHS4[49] (eSpCas9(1.1)_No_FLAG_AAVS1_T2 Addgene #79888) were used for insertion at the AAVS1 locus. pASHS16 (MTK234_030 spCas9-sgRNA1-hCLYBL[180] Addgene #123910) was used for insertion at the CLYBL locus. pHPHS115 was used for expressing BxbI recombinase for landing pad integration (pCAG-NLS-HA-Bxb1[216], Addgene #51271).

Cell line maintenance and generation

HEK293T cells (RRID:CVCL_0063, ATCC CRL-3216) were grown in DMEM (Thermo 11965084) supplemented with 10% FBS (Thermo 26140079). Cells were grown at 37C in 5% CO2. All transfections were performed using Lipofectamine 3000 (Thermo L3000015). HEK293T cells that stably express Cas9 (hsPB80) were generated by transfecting the CLYBL::Cas9-BFP vector pHPHS15 and spCas9 sgRNA1 hCLYBL vector, and selecting with 200μg/mL hygromycin. HEK293T cells containing an AttP landing pad site in the AAVS1 locus (hsPB126) were generated by co-transfecting pHPHS232 (the landing pad donor construct with AAVS1 homology arms) and the AAVS1 targeting CRISPR vectors, and selecting cells with blasticidin for 2 weeks. Plasmid sequences and cell lines available upon request.

Integration of plasmid libraries into cell lines

hsPB80 CLYBL::Cas9-BFP HEK293T cells were seeded to 50% confluency on 10 cm or 15 cm dishes for all library transfections. 10 μg of Pool 1 library plasmid pHPHS296 and 1.5 μg of each AAVS1 targeting CRISPR vector were transfected into three 15 cm dishes. 4 μg of Pool 2 library plasmid pHPHS492 and 0.5 μg of each AAVS1 targeting CRISPR vector were transfected into a single 10 cm dish. hsPB126 AAVS1::AttP_landing_pad HEK293T cells were seeded to 50% confluence on a single 10 cm dish, and transfected with 4 μg of library plasmid pHPHS491 and 1 μg of pHPHS115 BxbI expression vector.

Cells were selected with 2 μg/mL puromycin, added 72 hours post-transfection. Puromycin selection was removed after 6-10 days, once cells were growing robustly in selection. 24 hours after removing puromycin selection, stable library cells were plated into two separate 15 cm dishes (Pool 1 cells) or 10 cm dishes (Pool 2 cells), to reach 75% confluency the next day, for matched mRNA and gDNA harvests.

Library Genomic DNA extraction

Reporter library genomic DNA was harvested from one 75% confluent 15 cm or 10 cm dish of stably expressing library cells. Genomic DNA was harvested using Quick-DNA kit (Zymo D3024), following the manufacturer's instructions, with 3 mL of genomic DNA lysis buffer per 15 cm plate, and 1 ml of the same buffer per 10 cm plate. Between 0.5-10 μg of purified genomic DNA from each library sample was sheared into ~350 nucleotide length fragments by sonication for 10 min on ice using a Diagenode Bioruptor. Sheared gDNA was then *in vitro* transcribed into RNA (denoted gRNA below and in analysis code) starting from the T7 promoter region in the insert cassette, similar to previous approaches[182,183], using a HiScribe T7 High Yield RNA Synthesis Kit (NEB E2040S). Transcribed gRNA was treated with DNase I (NEB M0303S) and cleaned using an RNA Clean and Concentrator kit (Zymo R1013).

Library mRNA extraction

Reporter library mRNA was harvested from one 75% confluent 15 cm or 10 cm dish of stably expressing library cells. mRNA was harvested by using 3 mL of Trizol reagent (Thermo) to lyse cells directly on the plate, and then following the manufacturer's mRNA extraction protocol. Purified mRNA was then DNaseI (NEB M0303S) treated and cleaned using an RNA Clean and Concentrator kit (Zymo R1013).

mRNA and genomic DNA barcode sequencing

Between 0.5-10 μg of DNaseI-treated mRNA and gRNA for each library was reverse transcribed into cDNA using Maxima H Minus Reverse Transcriptase (Thermo EP0752) and a primer annealing to the Illumina R1 primer binding site (oPB354). A 170-nucleotide region surrounding the 24-nucleotide barcode was PCR amplified from the resulting cDNA in two rounds, using Phusion Flash High-Fidelity PCR Master Mix mastermix (Thermo F548L). Round 1 PCR was carried out for 10 cycles, with cDNA template comprising 1/10th of the PCR reaction volume, using primers oPB361 and oPB354. Round 1 PCRs were cleaned using a 2× volume of Agencourt Ampure XP beads (Beckman Coulter A63880) to remove primers. Cleaned samples were then used as template for Round 2 PCR, carried out for 5-15 cycles, using a common reverse primer (oAS111) and indexed forward primers for pooled high-throughput sequencing of different samples (oAS112-135 and oHP281-290). Amplified samples were run on a 1.5% agarose gel and fragments of the correct size were purified using ADB Agarose Dissolving Buffer (Zymo D4001-1-100) and UPrep Micro Spin Columns (Genesee Scientific 88-343). Concentrations of gel-purified samples were measured using a Qubit dsDNA HS Assay Kit (Q32851) with a Qubit 4 Fluorometer. Samples were sequenced using an Illumina HiSeq 2500 or Illumina NextSeq 2000 in 1×50, 2×50, or 1×100 mode (depending on other samples pooled with the sequencing library).

Insert-barcode linkage sequencing

Plasmid library pools pHPHS296, pHPHS491, and pHPHS492 were diluted to 10 ng/μL. A 240-nucleotide region surrounding the 48-nucleotide variable insert sequence and the 24-nucleotide barcode was PCR amplified from these pools in two rounds, using Phusion Flash High-Fidelity PCR Master Mix mastermix (Thermo F548L). Round 1 PCR was carried out for 10 cycles, with 10 ng/μL plasmid pool template comprising 1/10th of the PCR reaction volume, using primers oPB29 and oPB354. Round 1 PCRs were digested with DpnI (Thermo FD1704) at 37°C for 30 minutes to remove template plasmid and cleaned using a 2× volume of Agencourt Ampure XP beads (Beckman Coulter A63880) to remove primers and enzyme. Cleaned samples were used as template for Round 2 PCR, for 5 cycles, using oAS111 and indexed forward primers (oAS112-135 and oHP281-290). Amplified Round 2 PCR products were purified after size selection and quantified as described above for barcode sequencing. Samples were sequenced using an Illumina MiSeq or Illumina NextSeq 2000 in 2×50 or 1×100 mode.

Computational analyses

Pre-processing steps for high-throughput sequencing were implemented as `Snakemake` workflows[185]. `Python` (v3.7.4) and `R` (v3.6.2) programming languages were used for all analyses unless mentioned otherwise. In the description below, files ending in `.py` refer to `Python` scripts and files ending in `.Rmd` or `.R` refer to `R Markdown` or `R` scripts. All scripts are provided as a Supplementary file (`code.tar.gz`).

Barcode to insert assignment

The raw data from insert-barcode linkage sequencing are in `.fastq` format. If the inserts and barcodes were on paired-end reads instead of single-end reads, the reads were renamed in increasing numerical order starting at 0 to enable easy matching of insert and barcode reads. This was done in `rename_fastq_paired_reads.py`. The oligo pools were used to create a reference `fasta` file in `create_reference_for_aligning_library.R`. A `bowtie2`[186] (v2.4.2) reference was created from

the fasta file using the bowtie2-build command with default options. The insert read was aligned to the bowtie2 reference using bowtie2 command with options -N 1 -L 22 --end-to-end with the --trim5 and --trim3 options set to include only the region corresponding to the insert. The alignments were sorted and indexed using samtools[187] (v1.11) commands sort and index with default options. The alignments were filtered to include only reads with simple cigar strings and a MAPQ score greater than 20 in filter_alignments.R. The barcodes corresponding to each filtered alignment were parsed and tallied in count_barcode_insert_pairs.py. Depending on the sequencing depth, only barcodes that were observed at least 4-10 times were included in the tally. The tallied barcodes were aligned against themselves using bowtie2-build with default options and bowtie2 with options -L 24 -N 1 --all --norc. The self-alignment was used to exclude barcodes that are linked to distinct inserts or ones that are linked to the same barcode but are aligned against each other by bowtie2. In the latter case, the barcode with the lower count is discarded. The final list of insert barcode pairs is written as a tab-delimited .tsv.gz file for aligning barcodes from genomic DNA and mRNA sequencing below.

Barcode counting in genomic DNA and mRNA

The raw data from sequencing barcodes in genomic DNA and mRNA is in .fastq format. The filtered barcodes .tsv.gz file from the insert-barcode linkage sequencing is used to create a reference fasta file in create_bowtie_reference.R. A bowtie2 (v2.4.2) reference was created from the fasta file using the bowtie2-build command with default options. The barcodes were aligned to the bowtie2 reference using bowtie2 command with options -N 1 -L 20 --norc with the --trim5 and --trim3 options set to include only the region corresponding to the barcode. The alignments were sorted, indexed, and tallied using the samtools commands sort, index, idxstats with default options. GNU awk (v4.1.4) was used for miscellaneous processing of tab-delimited data between pre-processing steps. The final list of counts per barcode in each sample of genomic DNA or mRNA is written as a tab-delimited .tsv.gz file for calculating mRNA levels

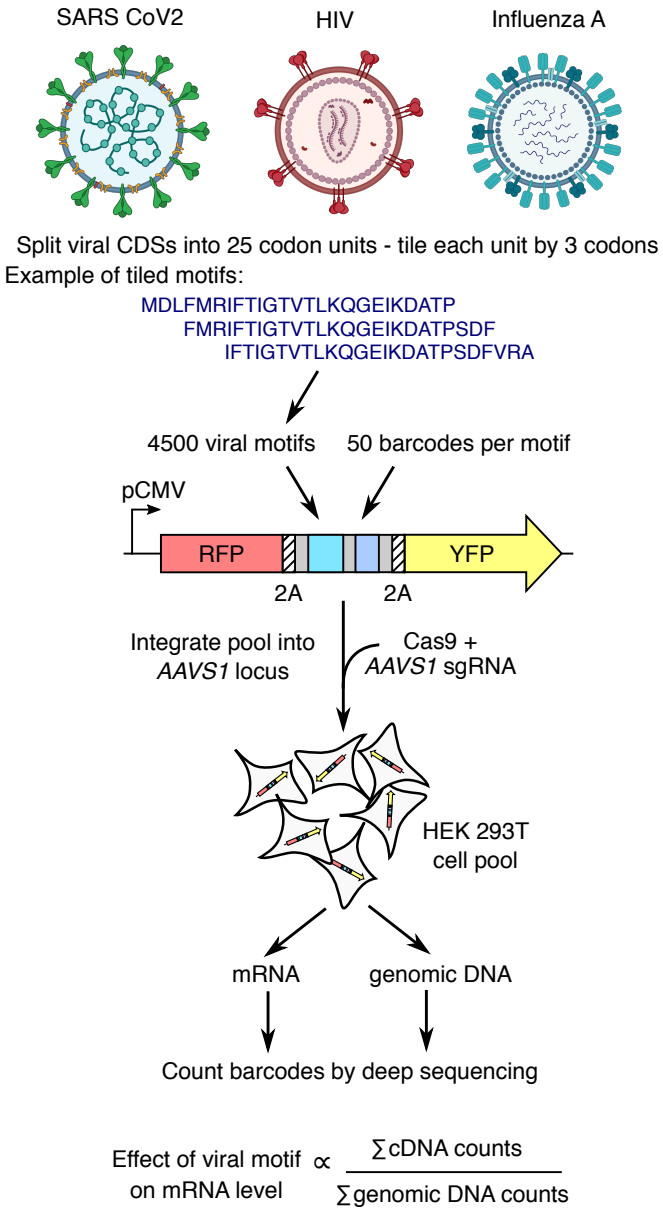
below.

mRNA quantification

All barcode counts corresponding to each insert in each sample were summed. Only inserts with a minimum of 200 reads and 6 barcodes summed across the mRNA and gRNA samples were included. mRNA levels were calculated as the log2 ratio of the summed mRNA barcode counts to the summed gRNA barcode counts. mRNA levels were median-normalized within the Pool 1 library, and max normalized to the highest value within the Pool 2 library.

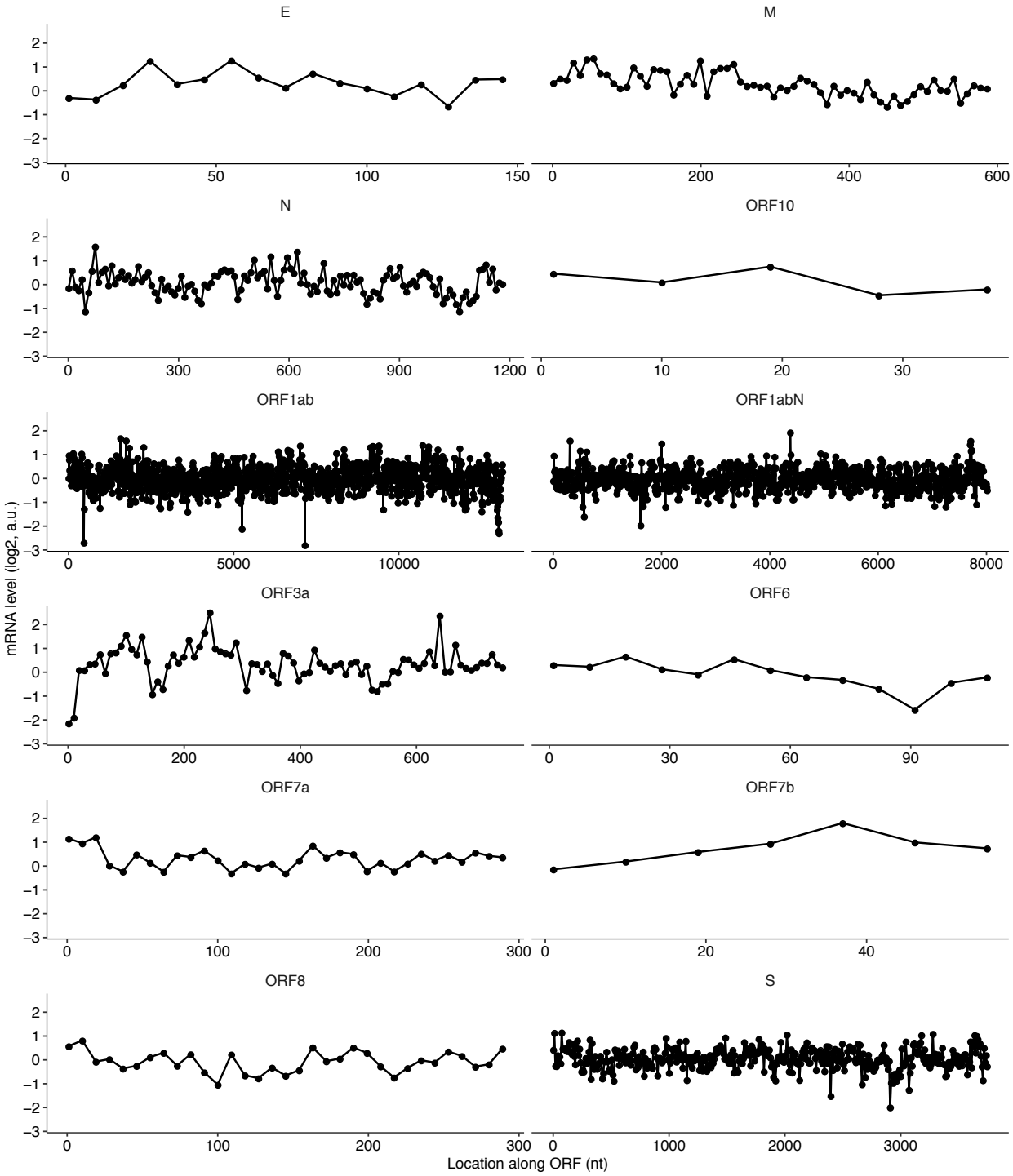
IV. Figures

Figure 3.1



A massively parallel assay for viral motif effects on mRNA levels
The full coding sequences of SARS-CoV2, HIV, and IAV were synthesized as a pool of 4500 oligonucleotide fragments. Each oligo contains a 75nt variable insert region encoding 25 amino acids tiled from the start codon of each annotated viral open reading frame in 3 amino acid (9nt) increments. Oligos were linked in-frame to 24nt random barcodes and cloned between RFP and YFP reporters with intervening 2A sequences. Each viral motif has a median of 50 random barcodes without in-frame stop codons. Reporter cassettes are integrated as a pool at the *AAVS1* locus in HEK293T cells by Cas9-mediated homologous recombination and constitutively expressed off the CMV promoter. Steady state mRNA level of each insert is determined by sequencing corresponding barcodes in the cDNA and the genomic DNA and normalizing the summed cDNA read counts by the genomic DNA read counts.

Figure 3.2



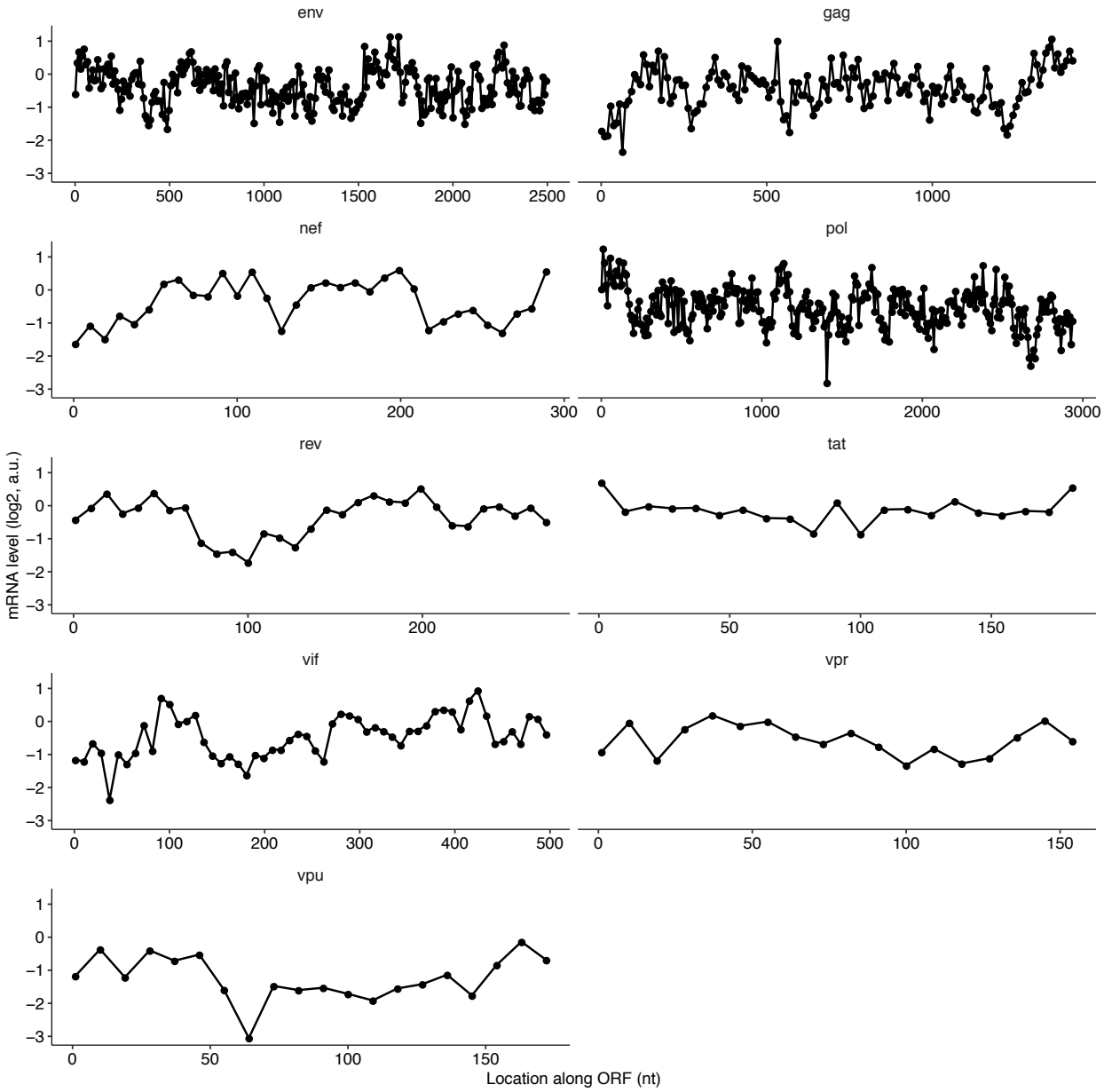
mRNA level measurements of the SARS-CoV-2 ORFeome
mRNA level of reporters encoding tiled insert sequences spanning the SARS-CoV-2 ORFeome. Points represent individual insert sequences. Position of each insert sequence in the ORF is shown in nucleotides on the x-axis.

Table 3.1

SARS-CoV-2 sequence motifs with mRNA levels < -1.5 log2 a.u below median

ORF	Location	mRNA level (log2 a.u.)	Motif Amino Acid Sequence
ORF1ab	7156	-2.79	FFASFYYVWKSYVHVVDGCNSSTCM
ORF1ab	460	-2.69	YEDFQENWNTKHSSGVTRELMRELN
ORF1ab	13033	-2.29	LKGKYVQIPTTCANDPVGFTLKNTV
ORF1ab	13024	-2.19	FCDLKGKYVQIPTTCANDPVGFTLK
ORF3a	1	-2.14	MDLFMRIFTIGTVTLKQGEIKDATP
ORF1ab	5248	-2.11	VVCKTCGQQQTTLKGVEAVMYMGTL
S	2908	-1.99	FGAISSVLNDILSRLDKVEAEVQID
ORF1abN	1612	-1.96	AISAKNRARTVAGVSICSTMTNRQF
ORF3a	10	-1.90	FMRIFTIGTVTLKQGEIKDATPSDF
ORF1ab	13015	-1.82	PKGFCDLKGKYVQIPTTCANDPVGF
ORF1ab	12997	-1.63	HIDHPNPKGFCDLKGKYVQIPTTCA
ORF1abN	568	-1.59	AGIVGVLTLDNQDLNGNWYDFGDFI
ORF6	91	-1.57	YIINLIIKNLSKSLTENKYSQLDEE
S	2395	-1.51	GFNFSQILPDPSKPSKRSFIEDLLF

Figure 3.3



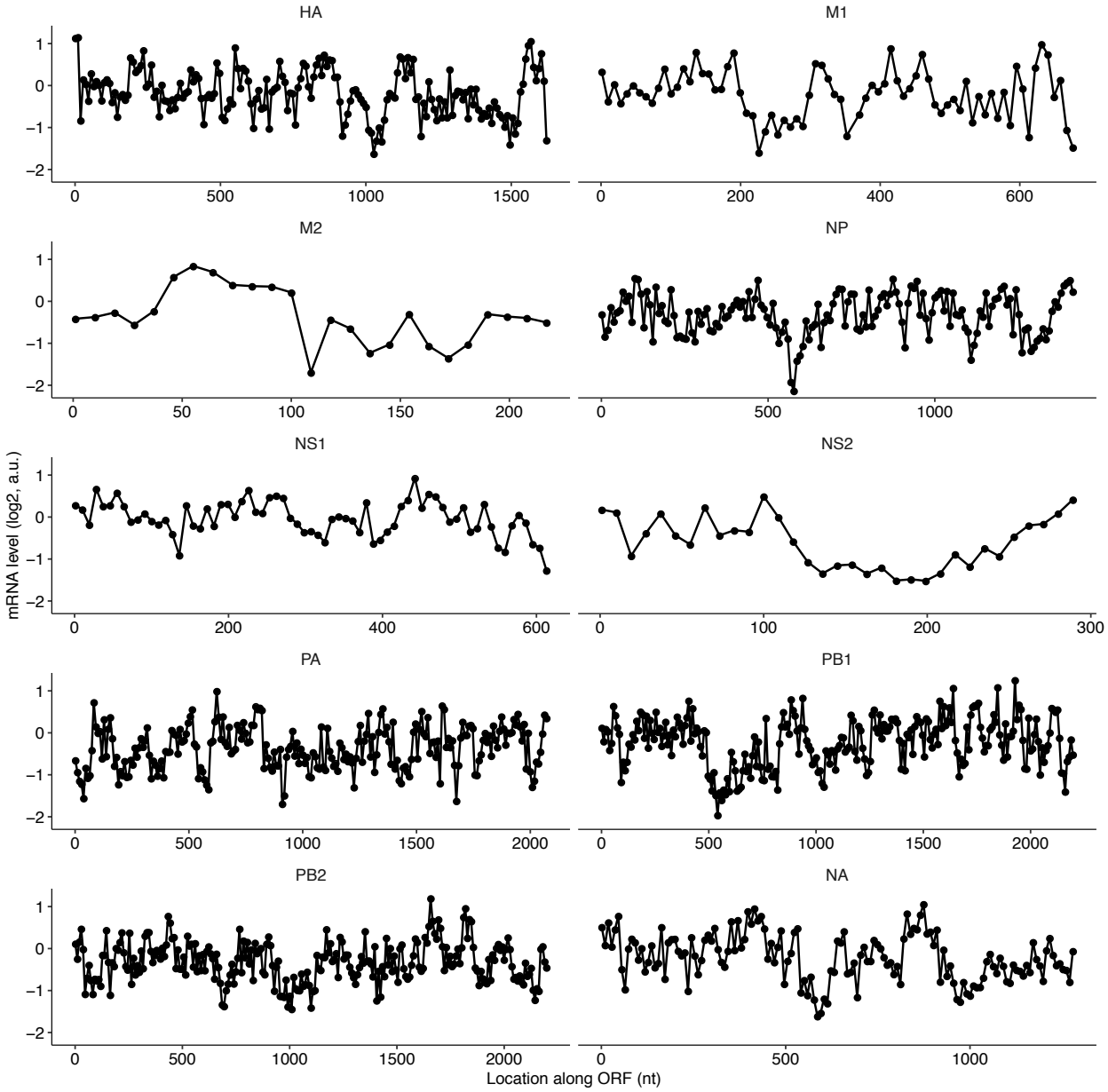
mRNA level measurements of the HIV ORFeome
mRNA level of reporters encoding tiled insert sequences spanning the HIV ORFeome. Points represent individual insert sequences. Position of each insert sequence in the ORF is shown in nucleotides on the x-axis.

Table 3.2

HIV sequence motifs with mRNA levels < -1.75 log2 a.u below median

ORF	Location	mRNA level (log2 a.u.)	Motif Amino Acid Sequence
vpu	64	-3.05	VWSIVIIEYRKILRQRKIDRLIDRL
pol	1405	-2.81	VHGVYYDPSKDLIAEIQKQGQGQWT
vif	37	-2.37	VDRMRIRTWKSLVKHHMYVSGKARG
gag	64	-2.34	RPGGKKKYKLKHIVWASRELERFAV
pol	2674	-2.29	QMAVFIHNFKRKGGIGGYSAGERIV
pol	2701	-2.06	KRKGGIGGYSAGERIVDIIATDIQT
pol	2665	-2.05	TAVQMAVFIHNFKRKGGIGGYSAGE
pol	2683	-2.00	VFIHNFKRKGGIGGYSAGERIVDII
vpu	109	-1.91	RKIDRLIDRLIERAEDSGNESEGEI
gag	10	-1.87	RASVLSGGELDRWEKIRLRPGGKKK
gag	19	-1.85	VLSGGELDRWEKIRLRPGGKKKYKL
gag	1225	-1.82	RKKGCWKCGKEGHQMKDCTERQANF
pol	2863	-1.81	KLLWKGEGAVVIQDNSDIKVVPRRK
pol	2692	-1.80	HNFKRKGGIGGYSAGERIVDIIATD
pol	2071	-1.78	VPAHKGIGGNEQVDKLVSAGIRKVL
vpu	145	-1.76	RAEDSGNESEGEISALVEMGVEMGH

Figure 3.4



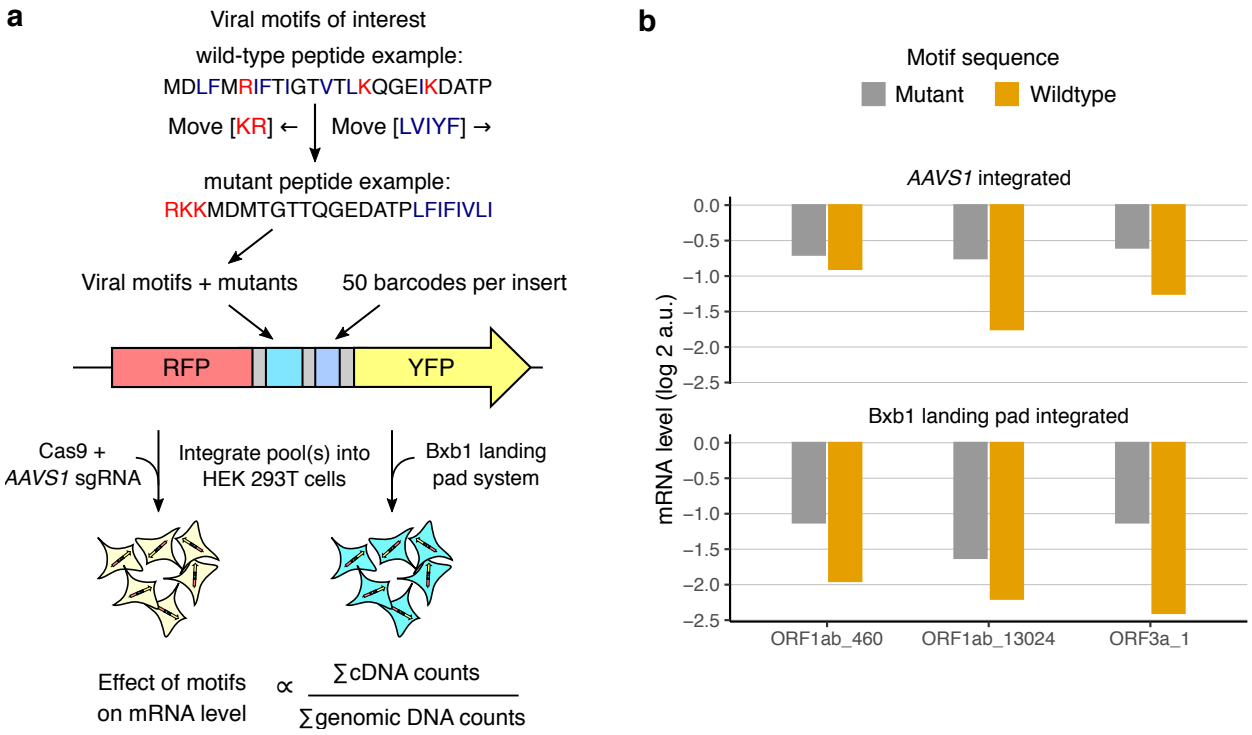
mRNA level measurements of the IAV ORFeome
mRNA level of reporters encoding tiled insert sequences spanning the HIV ORFeome. Points represent individual insert sequences. Position of each insert sequence in the ORF is shown in nucleotides on the x-axis.

Table 3.3

IAV sequence motifs with mRNA levels < -1.5 log2 a.u below median

ORF	Location	mRNA level (log2 a.u.)	Motif Amino Acid Sequence
NP	577	-2.13	MVRMIKRGINDRNFWRGENGRKTRI
PB1	541	-1.96	ITTHFQRKRRVRDNMTKKMITQRTM
NP	568	-1.92	VMEMVRMIKRGINDRNFWRGENGRK
M2	109	-1.69	HLTLWILDRLFFKCIYRRFKYGLKG
PA	910	-1.69	LYDAIKCMRTFFGWKEPNVVKPHEK
HA	1027	-1.62	RGLFGAIAGFIEGGWTGMIDGWYGY
PA	1675	-1.62	RPMFLYVRTNGTSKIKMKWGMEMRR
NA	586	-1.61	IITETIKSWRKKILRTQESECACVN
PB1	559	-1.60	RKRRVRDNMTKKMITQRTMGKKKQR
M1	226	-1.59	RRRFVQNALNGNGDPNNMDKAVKLY
PA	37	-1.55	IVELAEKTMKEYGEDLKIETNKFAA
NA	595	-1.53	ETIKSWRKKILRTQESECACVNGSC
NS2	199	-1.52	EQLGQKFEEIRWLIEEVRHKLKITE
NS2	181	-1.51	RNEKWREQLGQKFEEIRWLIEEVRH

Figure 3.5



Testing the effects of coding sequence rearrangements on destabilizing viral motifs
(a) Destabilizing SARS-CoV2 motifs were mutated to reorder their amino acids. All codons encoding K or R are moved to the 5′ end of the mutated motif and codons encoding L, V, I, Y, or F are moved to the 3′ end. ORF3a wildtype and mutant peptides are shown as an example. Wildtype and mutant viral motifs were cloned between RFP and YFP reporters with a median of 50 random downstream barcodes without in-frame stop codons. Reporter cassettes were integrated as a pool at the *AAVS1* locus in HEK293T cells by Cas9-mediated homologous recombination and constitutively expressed from a CMV promoter, or integrated by serine recombinase at a BxbI site in HEK293T cells using and expressed from a Tet-promoter induced by 2 ug/mL doxycycline for 24 hours prior to harvest.
(b) Steady state mRNA level of each insert is determined by sequencing corresponding barcodes in the cDNA and the genomic DNA and normalizing the summed cDNA read counts by the genomic DNA read counts. mRNA levels were max normalized to non-destabilizing insert sequences included in the oligo pool, which were set to lfc = 0.

Chapter 4: Conclusions

In this dissertation, I report the development of a massively parallel assay to profile the mRNA effects of over 10,000 endogenous and synthetic coding sequence motifs in human cells. The unprecedented scale of this in vivo assay, which takes advantage of pooled oligo synthesis and Cas9-mediated genome editing, allowed me to discover a complex nascent peptide code for mRNA stability. This discovery transforms our understanding of the ribosome from being a passive translator of the genetic code to an active filter against aggregation-prone peptides such as bulky β strands that are implicated in neurodegeneration. The large peptide-driven effects on mRNA stability found in my work will re-orient the translation field from its current focus on simple RNA motifs such as codon usage and GC content, and spur efforts to identify the genetic regulators that couple ribosome motion to mRNA stability in human cells. The finding that certain amino acid substitutions can impact mRNA stability in *cis* will lead to a major rethinking in the area of disease genetics where missense mutations are normally assumed to alter protein activity and not protein expression. My comprehensive profiling of viral coding sequences adds more depth to our established peptide code for cotranslational mRNA stability effects, and highlights regulatory mRNA sequence motifs that may be critical for viral fitness. Finally, the nascent peptide effects on mRNA stability and protein expression uncovered here will be of practical relevance to the design of mRNA-based therapeutics, which have taken center stage in the post-COVID era.

References

1. Luisi, P. L. About Various Definitions of Life. 10 (1997).

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

3. Filbeck, S., Cerullo, F., Pfeffer, S. & Joazeiro, C. A. P. Ribosome-associated quality-control mechanisms from bacteria to humans. *Molecular Cell* **82**, 1451–1466 (2022).

4. Leedom, S. L. & Keiler, K. C. Ribosome collisions: New ways to initiate ribosome rescue. *Current Biology* **32**, R469–R472 (2022).

5. Crick, F. Central Dogma of Molecular Biology. 3 (1970).

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

7. Ponomarenko, E. A. *et al.* The Size of the Human Proteome: The Width and Depth. *International Journal of Analytical Chemistry* **2016**, 1–6 (2016).

8. King, M.-C. & Wilson, A. C. Evolution at Two Levels in Humans and Chimpanzees: Their macromolecules are so alike that regulatory mutations may account for their biological differences. *Science* **188**, 107–116 (1975).

9. Joshi, M., Kapopoulou, A. & Laurent, S. Impact of Genetic Variation in Gene Regulatory Sequences: A Population Genomics Perspective. *Front. Genet.* **12**, 660899 (2021).

10. Woychik, N. A. & Hampsey, M. The RNA Polymerase II Machinery. *Cell* **108**, 453–463 (2002).

11. Roeder, R. G. 50+ years of eukaryotic transcription: an expanding universe of factors and mechanisms. *Nat Struct Mol Biol* **26**, 783–791 (2019).

12. Proudfoot, N. J., Furger, A. & Dye, M. J. Integrating mRNA Processing with Transcription. *Cell* **108**, 501–512 (2002).

13. Ramanathan, A., Robb, G. B. & Chan, S.-H. mRNA capping: biological functions and applications. *Nucleic Acids Res* **44**, 7511–7526 (2016).

14. Galloway, A. & Cowling, V. H. mRNA cap regulation in mammalian cell function and fate. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1862**, 270–279 (2019).

15. Black, D. L. Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annu. Rev. Biochem.* **72**, 291–336 (2003).

16. Shatkin, A. J. & Manley, J. L. The ends of the affair: Capping and polyadenylation. *nature structural biology* **7**, 5 (2000).

17. Tian, B. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research* **33**, 201–212 (2005).

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

19. Maquat, L. E. & Carmichael, G. G. Quality Control of mRNA Function. *Cell* **104**, 173–176 (2001).

20. Verschoor, A., Warner, J. R., Srivastava, S., Grassucci, R. A. & Frank, J. Three-dimensional structure of the yeast ribosome. *Nucleic Acids Research* **26**, 655–661 (1998).

21. Buttgereit, F. & Brand, M. D. A hierarchy of ATP-consuming processes in mammalian cells. *Biochemical Journal* **312**, 163–167 (1995).

22. Rolfe, D. F. & Brown, G. C. Cellular energy utilization and molecular origin of standard metabolic rate in mammals. *Physiological Reviews* **77**, 731–758 (1997).

23. Leslie, M. There are millions of protein factories in every cell. Surprise, they're not all the same. *Science* (2017) doi:10.1126/science.aan6994.

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

25. Lindqvist, L. M., Tandoc, K., Topisirovic, I. & Furic, L. Cross-talk between protein synthesis, energy metabolism and autophagy in cancer. *Current Opinion in Genetics & Development* **48**, 104–111 (2018).

26. Wek, R. C. Role of eIF2α Kinases in Translational Control and Adaptation to Cellular Stress. *Cold Spring Harb Perspect Biol* **10**, a032870 (2018).

27. Hershey, J. W. B., Sonenberg, N. & Mathews, M. B. Principles of Translational Control. *Cold Spring Harb Perspect Biol* **11**, a032607 (2019).

28. Pechmann, S., Willmund, F. & Frydman, J. The Ribosome as a Hub for Protein Quality Control. *Molecular Cell* **49**, 411–421 (2013).

29. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**, 227–232 (2012).

30. Pisarev, A. V., Unbehaun, A., Hellen, C. U. T. & Pestova, T. V. Assembly and Analysis of Eukaryotic Translation Initiation Complexes. in *Methods in Enzymology* vol. 430 147–177 (Elsevier, 2007).

31. Smith, R. C. L. *et al.* Translation initiation in cancer at a glance. *Journal of Cell Science* **134**, jcs248476 (2021).

32. Kozak, M. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *Journal of Molecular Biology* **196**, 947–950 (1987).

33. Noderer, W. L. *et al.* Quantitative analysis of mammalian translation initiation sites by FACS -seq. *Mol Syst Biol* **10**, 748 (2014).

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

35. Li, J. J., Chew, G.-L. & Biggin, M. D. Quantitative principles of cis-translational control by general mRNA sequence features in eukaryotes. *Genome Biol* **20**, 162 (2019).

36. Hsieh, A. C. *et al.* The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature* **485**, 55–61 (2012).

37. Farooq, Z. *et al.* The amino acid sensor GCN2 suppresses terminal oligopyrimidine (TOP) mRNA translation via La-related protein 1 (LARP1). *Journal of Biological Chemistry* **298**, 102277 (2022).

38. Baird, T. D. & Wek, R. C. Eukaryotic Initiation Factor 2 Phosphorylation and Translational Control in Metabolism. *Advances in Nutrition* **3**, 307–321 (2012).

39. Bottorff, T. A., Park, H., Geballe, A. P. & Subramaniam, A. R. Translational buffering by ribosome stalling in upstream open reading frames. *PLoS Genet* **18**, e1010460 (2022).

40. Knight, J. R. P. *et al.* Control of translation elongation in health and disease. *Disease Models & Mechanisms* **13**, dmm043208 (2020).

41. Browning, K. S. & Bailey-Serres, J. Mechanism of Cytoplasmic mRNA Translation. *The Arabidopsis Book* **13**, e0176 (2015).

42. Dever, T. E., Dinman, J. D. & Green, R. Translation Elongation and Recoding in Eukaryotes. *Cold Spring Harb Perspect Biol* **10**, a032649 (2018).

43. Behrmann, E. *et al.* Structural Snapshots of Actively Translating Human Ribosomes. *Cell* **161**, 845–857 (2015).

44. Ferguson, A. *et al.* Functional Dynamics within the Human Ribosome Regulate the Rate of Active Protein Synthesis. *Molecular Cell* **60**, 475–486 (2015).

45. Schuller, A. P. & Green, R. Roadblocks and resolutions in eukaryotic translation. *Nat Rev Mol Cell Biol* **19**, 526–541 (2018).

46. Collart, M. A. & Weiss, B. Ribosome pausing, a dangerous necessity for co-translational events. *Nucleic Acids Research* **48**, 1043–1055 (2020).

47. Richter, Joel D. & Coller, J. Pausing on Polyribosomes: Make Way for Elongation in Translational Control. *Cell* **163**, 292–300 (2015).

48. Subramaniam, A. R., Pan, T. & Cluzel, P. Environmental perturbations lift the degeneracy of the genetic code to regulate protein levels in bacteria. *Proceedings of the National Academy of Sciences* **110**, 2419–2424 (2013).

49. Darnell, A. M., Subramaniam, A. R. & O'Shea, E. K. Translational Control through Differential Ribosome Pausing during Amino Acid Limitation in Mammalian Cells. *Molecular Cell* **71**, 229–243.e11 (2018).

50. Hanson, G. & Coller, J. Codon optimality, bias and usage in translation and mRNA decay. *Nat Rev Mol Cell Biol* **19**, 20–30 (2018).

51. Doma, M. K. & Parker, R. Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. *Nature* **440**, 561–564 (2006).

52. Bazzini, A. A. *et al.* Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *EMBO J* **35**, 2087–2103 (2016).

53. de Freitas Nascimento, J., Kelly, S., Sunter, J. & Carrington, M. Codon choice directs constitutive mRNA levels in trypanosomes. *eLife* **7**, e32467 (2018).

54. Harigaya, Y. & Parker, R. Codon optimality and mRNA decay. *Cell Res* **26**, 1269–1270 (2016).

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

56. Gamble, C. E., Brule, C. E., Dean, K. M., Fields, S. & Grayhack, E. J. Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast. *Cell* **166**, 679–690 (2016).

57. Lu, J. & Deutsch, C. Electrostatics in the Ribosomal Tunnel Modulate Chain Elongation Rates. *Journal of Molecular Biology* **384**, 73–86 (2008).

58. Ito, K. & Chiba, S. Arrest Peptides: *Cis* -Acting Modulators of Translation. *Annu. Rev. Biochem.* **82**, 171–202 (2013).

59. Tuck, A. C. *et al.* Mammalian RNA Decay Pathways Are Highly Specialized and Widely Linked to Translation. *Molecular Cell* **77**, 1222–1236.e13 (2020).

60. Frenkel-Morgenstern, M. *et al.* Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. *Mol Syst Biol* **8**, 572 (2012).

61. Saikia, M. *et al.* Codon optimality controls differential mRNA translation during amino acid starvation. *RNA* **22**, 1719–1727 (2016).

62. Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat Struct Mol Biol* **20**, 237–243 (2013).

63. Nürenberg-Goloub, E. & Tampé, R. Ribosome recycling in mRNA translation, quality control, and homeostasis. *Biological Chemistry* **401**, 47–61 (2019).

64. Saito, S., Hosoda, N. & Hoshino, S. The Hbs1-Dom34 Protein Complex Functions in Non-stop mRNA Decay in Mammalian Cells. *Journal of Biological Chemistry* **288**, 17832–17843 (2013).

65. Heck, A. M. & Wilusz, J. The Interplay between the RNA Decay and Translation Machinery in Eukaryotes. *Cold Spring Harb Perspect Biol* **10**, a032839 (2018).

66. Park, E.-H. *et al.* Multiple elements in the eIF4G1 N-terminus promote assembly of eIF4G1•PABP mRNPs *in vivo*: Functionally redundant elements in the eIF4G1 N-terminus. *The EMBO Journal* **30**, 302–316 (2011).

67. Espel, E. The role of the AU-rich elements of mRNAs in controlling translation. *Seminars in Cell & Developmental Biology* **16**, 59–67 (2005).

68. Yang, E. *et al.* Decay Rates of Human mRNAs: Correlation With Functional Characteristics and Sequence Attributes. *Genome Res.* **13**, 1863–1872 (2003).

69. Parker, R. RNA Degradation in *Saccharomyces cerevisae*. *Genetics* **191**, 671–702 (2012).

70. Yi, H. *et al.* PABP Cooperates with the CCR4-NOT Complex to Promote mRNA Deadenylation and Block Precocious Decay. *Molecular Cell* **70**, 1081–1088.e5 (2018).

71. Łabno, A., Tomecki, R. & Dziembowski, A. Cytoplasmic RNA decay pathways - Enzymes and mechanisms. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1863**, 3125–3147 (2016).

72. Hsu, C. L. & Stevens, A. Yeast cells lacking 5'-->3' exoribonuclease 1 contain mRNA species that are poly(A) deficient and partially lack the 5' cap structure. *Mol Cell Biol* **13**, 4826–4835 (1993).

73. Mitchell, P., Petfalski, E., Shevchenko, A., Mann, M. & Tollervey, D. The Exosome: A Conserved Eukaryotic RNA Processing Complex Containing Multiple 3′□5′ Exoribonucleases. *Cell* **91**, 457–466 (1997).

74. Anderson, J. S. J. The 3' to 5' degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the SKI2 DEVH box protein and 3' to 5' exonucleases of the exosome complex. *The EMBO Journal* **17**, 1497–1506 (1998).

75. Zinder, J. C. & Lima, C. D. Targeting RNA for processing or destruction by the eukaryotic RNA exosome and its cofactors. *Genes Dev.* **31**, 88–100 (2017).

76. Lykke-Andersen, J. & Bennett, E. J. Protecting the proteome: Eukaryotic cotranslational quality control pathways. *Journal of Cell Biology* **204**, 467–476 (2014).

77. Rivas, M. A. *et al.* Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**, 666–669 (2015).

78. He, F. & Jacobson, A. Nonsense-Mediated mRNA Decay: Degradation of Defective Transcripts Is Only Part of the Story. *Annu. Rev. Genet.* **49**, 339–366 (2015).

79. Nickless, A., Bailis, J. M. & You, Z. Control of gene expression through the nonsense-mediated RNA decay pathway. *Cell Biosci* **7**, 26 (2017).

80. Maquat, L. E., Tarn, W.-Y. & Isken, O. The Pioneer Round of Translation: Features and Functions. *Cell* **142**, 368–374 (2010).

81. Ikeuchi, K., Yazaki, E., Kudo, K. & Inada, T. Conserved functions of human Pelota in mRNA quality control of nonstop mRNA. *FEBS Lett* **590**, 3254–3263 (2016).

82. Arribere, J. A. & Fire, A. Z. Nonsense mRNA suppression via nonstop decay. *eLife* **7**, e33292 (2018).

83. Arthur, L. L. *et al.* Translational control by lysine-encoding A-rich sequences. *Sci. Adv.* **1**, e1500154 (2015).

84. Koutmou, K. S. *et al.* Ribosomes slide on lysine-encoding homopolymeric A stretches. *eLife* **4**, e05534 (2015).

85. Guydosh, N. R. & Green, R. Translation of poly(A) tails leads to precise mRNA cleavage. *RNA* **23**, 749–761 (2017).

86. Chandrasekaran, V. *et al.* Mechanism of ribosome stalling during translation of a poly(A) tail. *Nat Struct Mol Biol* **26**, 1132–1140 (2019).

87. Simms, C. L., Yan, L. L. & Zaher, H. S. Ribosome Collision Is Critical for Quality Control during No-Go Decay. *Molecular Cell* **68**, 361–373.e5 (2017).

88. Park, H. & Subramaniam, A. R. Inverted translational control of eukaryotic gene expression by ribosome collisions. *PLoS Biol* **17**, e3000396 (2019).

89. D'Orazio, K. N. *et al. The endonuclease Cue2 cleaves mRNAs at stalled ribosomes during No Go Decay*. http://biorxiv.org/lookup/doi/10.1101/671099 (2019) doi:10.1101/671099.

90. Karamyshev, A. L. & Karamysheva, Z. N. Lost in Translation: Ribosome-Associated mRNA and Protein Quality Controls. *Front. Genet.* **9**, 431 (2018).

91. Juszkiewicz, S. *et al.* ZNF598 Is a Quality Control Sensor of Collided Ribosomes. *Molecular Cell* **72**, 469–481.e7 (2018).

92. Ikeuchi, K. *et al.* Collided ribosomes form a unique structural interface to induce Hel2-driven quality control pathways. *EMBO J* **38**, (2019).

93. Matsuo, Y. *et al.* Ubiquitination of stalled ribosome triggers ribosome-associated quality control. *Nat Commun* **8**, 159 (2017).

94. Juszkiewicz, S., Speldewinde, S. H., Wan, L., Svejstrup, J. Q. & Hegde, R. S. The ASC-1 Complex Disassembles Collided Ribosomes. *Molecular Cell* **79**, 603–614.e8 (2020).

95. Narita, M. *et al.* A distinct mammalian disome collision interface harbors K63-linked polyubiquitination of uS10 to trigger hRQT-mediated subunit dissociation. *Nat Commun* **13**, 6411 (2022).

96. Brandman, O. *et al.* A Ribosome-Bound Quality Control Complex Triggers Degradation of Nascent Peptides and Signals Translation Stress. *Cell* **151**, 1042–1054 (2012).

97. Shao, S., Brown, A., Santhanam, B. & Hegde, Ramanujan S. Structure and Assembly Pathway of the Ribosome Quality Control Complex. *Molecular Cell* **57**, 433–444 (2015).

98. D'Orazio, K. N. & Green, R. Ribosome states signal RNA quality control. *Molecular Cell* **81**, 1372–1383 (2021).

99. Habich, M., Djuranovic, S. & Szczesny, P. PATACSDB—the database of polyA translational attenuators in coding sequences. *PeerJ Computer Science* **2**, e45 (2016).

100. Weber, R. *et al.* 4EHP and GIGYF1/2 Mediate Translation-Coupled Messenger RNA Decay. *Cell Reports* **33**, 108262 (2020).

101. Han, P. *et al.* Genome-wide Survey of Ribosome Collision. *Cell Reports* **31**, 107610 (2020).

102. Shanmuganathan, V. *et al.* Structural and mutational analysis of the ribosome-arresting human XBP1u. *eLife* **8**, e46267 (2019).

103. Hickey, K. L. *et al.* GIGYF2 and 4EHP Inhibit Translation Initiation of Defective Messenger RNAs to Assist Ribosome-Associated Quality Control. *Molecular Cell* **79**, 950–962.e6 (2020).

104. Juszkiewicz, S. *et al.* Ribosome collisions trigger cis-acting feedback inhibition of translation initiation. *eLife* **9**, e60038 (2020).

105. Sinha, N. K. *et al.* EDF1 coordinates cellular responses to ribosome collisions. *eLife* **9**, e58828 (2020).

106. Sundaramoorthy, E. *et al.* ZNF598 and RACK1 Regulate Mammalian Ribosome-Associated Quality Control Function by Mediating Regulatory 40S Ribosomal Ubiquitylation. *Molecular Cell* **65**, 751–760.e4 (2017).

107. Juszkiewicz, S. & Hegde, R. S. Initiation of Quality Control during Poly(A) Translation Requires Site-Specific Ribosome Ubiquitination. *Molecular Cell* **65**, 743–750.e4 (2017).

108. Garzia, A. *et al.* The E3 ubiquitin ligase and RNA-binding protein ZNF598 orchestrates ribosome quality control of premature polyadenylated mRNAs. *Nat Commun* **8**, 16056 (2017).

109. Presnyak, V. *et al.* Codon Optimality Is a Major Determinant of mRNA Stability. *Cell* **160**, 1111–1124 (2015).

110. Narula, A., Ellis, J., Taliaferro, J. M. & Rissland, O. S. Coding regions affect mRNA stability in human cells. *RNA* **25**, 1751–1764 (2019).

111. Wu, Q. *et al.* Translation affects mRNA stability in a codon-dependent manner in human cells. *eLife* **8**, e45396 (2019).

112. Forrest, M. E. *et al.* Codon and amino acid content are associated with mRNA stability in mammalian cells. *PLoS ONE* **15**, e0228730 (2020).

113. Buschauer, R. *et al.* The Ccr4-Not complex monitors the translating ribosome for codon optimality. *Science* **368**, eaay6912 (2020).

114. Gillen, S. L. *et al.* Differential regulation of mRNA fate by the human Ccr4-Not complex is driven by coding sequence composition and mRNA localization. *Genome Biol* **22**, 284 (2021).

115. Hia, F. *et al.* Codon bias confers stability to human mRNA s. *EMBO Rep* **20**, (2019).

116. Matsuo, Y. *et al.* RQT complex dissociates ribosomes collided on endogenous RQC substrate SDD1. *Nat Struct Mol Biol* **27**, 323–332 (2020).

117. Amrani, N. *et al.* A faux 3′-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. *Nature* **432**, 112–118 (2004).

118. Eberle, A. B., Lykke-Andersen, S., Mühlemann, O. & Jensen, T. H. SMG6 promotes endonucleolytic cleavage of nonsense mRNA in human cells. *Nat Struct Mol Biol* **16**, 49–55 (2009).

119. Singh, G., Rebbapragada, I. & Lykke-Andersen, J. A Competition between Stimulators and Antagonists of Upf Complex Recruitment Governs Human Nonsense-Mediated mRNA Decay. *PLoS Biol* **6**, e111 (2008).

120. Charneski, C. A. & Hurst, L. D. Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. *PLoS Biol* **11**, e1001508 (2013).

121. Requião, R. D., de Souza, H. J. A., Rossetto, S., Domitrovic, T. & Palhano, F. L. Increased ribosome density associated to positively charged residues is evident in ribosome profiling experiments performed in the absence of translation inhibitors. *RNA Biology* **13**, 561–568 (2016).

122. Lu, J., Kobertz, W. R. & Deutsch, C. Mapping the Electrostatic Potential within the Ribosomal Exit Tunnel. *Journal of Molecular Biology* **371**, 1378–1391 (2007).

123. Nissley, D. A. *et al.* Electrostatic Interactions Govern Extreme Nascent Protein Ejection Times from Ribosomes and Can Delay Ribosome Recycling. *J. Am. Chem. Soc.* **142**, 6103–6110 (2020).

124. Zimmerman, J. M., Eliezer, N. & Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *Journal of Theoretical Biology* **21**, 170–201 (1968).

125. Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hechtt, M. H. Protein Design by Binary Patterning of Polar and Nonpolar Amino Acids. **262**, 7 (1993).

126. Xiong, H., Buckwalter, B. L., Shieh, H. M. & Hecht, M. H. Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides. *Proceedings of the National Academy of Sciences* **92**, 6349–6353 (1995).

127. Moffat, L. & Jones, D. T. Increasing the accuracy of single sequence prediction methods using a deep semi-supervised learning framework. *Bioinformatics* **37**, 3744–3751 (2021).

128. Burke, P. C., Park, H. & Subramaniam, A. R. A nascent peptide code for translational control of mRNA stability in human cells. *Nat Commun* **13**, 6829 (2022).

129. Chou, P. Y. & Fasman, G. D. Empirical Predictions of Protein Conformation. *Annu. Rev. Biochem.* **47**, 251–276 (1978).

130. Luo, P. & Baldwin, R. L. Mechanism of Helix Induction by Trifluoroethanol: A Framework for Extrapolating the Helix-Forming Properties of Peptides from Trifluoroethanol/Water Mixtures Back to Water. *Biochemistry* **36**, 8413–8421 (1997).

131. Jasanoff, A. & Fersht, A. R. Quantitative Determination of Helical Propensities from Trifluoroethanol Titration Curves. *Biochemistry* **33**, 2129–2135 (1994).

132. Kolář, M. H. *et al. Folding of VemP into translation-arresting secondary structure is driven by the ribosome exit tunnel*. http://biorxiv.org/lookup/doi/10.1101/2021.04.15.440051 (2021) doi:10.1101/2021.04.15.440051.

133. Watkins, A. M. & Arora, P. S. Anatomy of β-Strands at Protein–Protein Interfaces. *ACS Chem. Biol.* **9**, 1747–1754 (2014).

134. Wu, C. C.-C., Peterson, A., Zinshteyn, B., Regot, S. & Green, R. Ribosome Collisions Trigger General Stress Responses to Regulate Cell Fate. *Cell* **182**, 404–416.e14 (2020).

135. Chyżyńska, K., Labun, K., Jones, C., Grellscheid, S. N. & Valen, E. Deep conservation of ribosome stall sites across RNA processing genes. *NAR Genomics and Bioinformatics* **3**, lqab038 (2021).

136. Mizuno, M. *et al.* The nascent polypeptide in the 60S subunit determines the Rqc2-dependency of ribosomal quality control. *Nucleic Acids Research* (2021) doi:10.1093/nar/gkab005.

137. Ramu, H. *et al.* Nascent peptide in the ribosome exit tunnel affects functional properties of the A-site of the peptidyl transferase center. *Mol Cell* **41**, 321–330 (2011).

138. Lu, J., Hua, Z., Kobertz, W. R. & Deutsch, C. Nascent Peptide Side Chains Induce Rearrangements in Distinct Locations of the Ribosomal Tunnel. *Journal of Molecular Biology* **411**, 499–510 (2011).

139. Po, P. *et al.* Effect of Nascent Peptide Steric Bulk on Elongation Kinetics in the Ribosome Exit Tunnel. *Journal of Molecular Biology* **429**, 1873–1888 (2017).

140. Li, W. *et al.* Structural basis for selective stalling of human ribosome nascent chain complexes by a drug-like molecule. *Nature Structural & Molecular Biology* **26**, 501–509 (2019).

141. Sabi, R. & Tuller, T. Computational analysis of nascent peptides that induce ribosome stalling and their proteomic distribution in *Saccharomyces cerevisiae*. *RNA* **23**, 983–994 (2017).

142. Parola, A. L. & Kobilka, B. K. The peptide product of a 5' leader cistron in the beta 2 adrenergic receptor mRNA inhibits receptor synthesis. *Journal of Biological Chemistry* **269**, 4497–4505 (1994).

143. Reynolds, K., Zimmer, A. M. & Zimmer, A. Regulation of RAR beta 2 mRNA expression: evidence for an inhibitory peptide encoded in the 5'-untranslated region. *Journal of Cell Biology* **134**, 827–835 (1996).

144. Matheisl, S., Berninghausen, O., Becker, T. & Beckmann, R. Structure of a human translation termination complex. *Nucleic Acids Res* **43**, 8615–8626 (2015).

145. Lintner, N. G. *et al.* Selective stalling of human translation through small-molecule engagement of the ribosome nascent chain. *PLOS Biology* **15**, e2001882 (2017).

146. Ivanov, I. P. *et al.* Polyamine Control of Translation Elongation Regulates Start Site Selection on Antizyme Inhibitor mRNA via Ribosome Queuing. *Molecular Cell* **70**, 254–264 (2018).

147. Bhushan, S. *et al.* SecM-Stalled Ribosomes Adopt an Altered Geometry at the Peptidyl Transferase Center. *PLOS Biology* **9**, e1000581 (2011).

148. Seidelt, B. *et al.* Structural Insight into Nascent Polypeptide Chain–Mediated Translational Stalling. *Science* **326**, 1412–1415 (2009).

149. Wilson, D. N., Arenz, S. & Beckmann, R. Translation regulation via nascent polypeptide-mediated ribosome stalling. *Current Opinion in Structural Biology* **37**, 123–133 (2016).

150. Su, T. *et al.* The force-sensing peptide VemP employs extreme compaction and secondary structure formation to induce ribosomal stalling. *eLife* **6**, e25642 (2017).

151. Hardesty, B. & Kramer, G. Folding of a nascent peptide on the ribosome. in *Progress in Nucleic Acid Research and Molecular Biology* vol. 66 41–66 (Academic Press, 2000).

152. Woolhead, C. A., Johnson, A. E. & Bernstein, H. D. Translation Arrest Requires Two-Way Communication between a Nascent Polypeptide and the Ribosome. *Molecular Cell* **22**, 587–598 (2006).

153. Lu, J. & Deutsch, C. Secondary Structure Formation of a Transmembrane Segment in Kv Channels. *Biochemistry* **44**, 8230–8243 (2005).

154. Yap, M.-N. & Bernstein, H. D. The Plasticity of a Translation Arrest Motif Yields Insights into Nascent Polypeptide Recognition inside the Ribosome Tunnel. *Molecular Cell* **34**, 201–211 (2009).

155. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron* **72**, 245–256 (2011).

156. Renton, Alan E. *et al.* A Hexanucleotide Repeat Expansion in C9ORF72 Is the Cause of Chromosome 9p21-Linked ALS-FTD. *Neuron* **72**, 257–268 (2011).

157. Mizielinska, S. *et al.* *C9orf72* repeat expansions cause neurodegeneration in *Drosophila* through arginine-rich proteins. *Science* **345**, 1192–1194 (2014).

158. Kriachkov, V. *et al.* *Arginine-rich C9ORF72 ALS Proteins Stall Ribosomes in a Manner Distinct From a Canonical Ribosome-Associated Quality Control Substrate*. http://biorxiv.org/lookup/doi/10.1101/2022.02.09.479805 (2022) doi:10.1101/2022.02.09.479805.

159. Loveland, A. B. *et al.* Ribosome inhibition by C9ORF72-ALS/FTD-associated poly-PR and poly-GR proteins revealed by cryo-EM. *Nat Commun* **13**, 2776 (2022).

160. Kanekura, K. *et al.* Characterization of membrane penetration and cytotoxicity of C9orf72-encoding arginine-rich dipeptides. *Sci Rep* **8**, 12740 (2018).

161. Glover, M. L. *et al.* NONU-1 Encodes a Conserved Endonuclease Required for mRNA Translation Surveillance. *Cell Reports* **30**, 4321–4331.e4 (2020).

162. Cao, J. & Geballe, A. P. Mutational Analysis of the Translational Signal in the Human Cytomegalovirus gpUL4 (gp48) Transcript Leader by Retroviral Infection. *Virology* **205**, 151–160 (1994).

163. Yanagitani, K., Kimata, Y., Kadokura, H. & Kohno, K. Translational Pausing Ensures Membrane Targeting and Cytoplasmic Splicing of XBP1u mRNA. *Science* **331**, 586–589 (2011).

164. Yang, J., Hao, X., Cao, X., Liu, B. & Nyström, T. Spatial sequestration and detoxification of Huntingtin by the ribosome quality control complex. *eLife* **5**, e11792 (2016).

165. Zheng, J. *et al.* Role of the ribosomal quality control machinery in nucleocytoplasmic translocation of polyQ-expanded huntingtin exon-1. *Biochemical and Biophysical Research Communications* **493**, 708–717 (2017).

166. Aviner, R. *et al.* *Ribotoxic collisions on CAG expansions disrupt proteostasis and stress responses in Huntington's Disease*. http://biorxiv.org/lookup/doi/10.1101/2022.05.04.490528 (2022) doi:10.1101/2022.05.04.490528.

167. Park, J. *et al.* ZNF598 co-translationally titrates poly(GR) protein implicated in the pathogenesis of *C9ORF72* -associated ALS/FTD. *Nucleic Acids Research* **49**, 11294–11311 (2021).

168. Cymer, F. & von Heijne, G. Cotranslational folding of membrane proteins probed by arrest-peptide–mediated force measurements. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 14640–14645 (2013).

169. Nilsson, Ola B. *et al.* Cotranslational Protein Folding inside the Ribosome Exit Tunnel. *Cell Reports* **12**, 1533–1540 (2015).

170. Ismail, N., Hedman, R., Schiller, N. & von Heijne, G. A biphasic pulling force acts on transmembrane helices during translocon-mediated membrane integration. *Nat Struct Mol Biol* **19**, 1018–1022 (2012).

171. Karamyshev, Andrey L. *et al.* Inefficient SRP Interaction with a Nascent Chain Triggers a mRNA Quality Control Pathway. *Cell* **156**, 146–157 (2014).

172. Shiber, A. *et al.* Cotranslational assembly of protein complexes in eukaryotes revealed by ribosome profiling. *Nature* **561**, 268–272 (2018).

173. Bertolini, M. *et al.* Interactions between nascent proteins translated by adjacent ribosomes drive homomer assembly. *Science* **371**, 57–64 (2021).

174. Stanger, H. E. *et al.* Length-dependent stability and strand length limits in antiparallel β-sheet secondary structure. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 12015–12020 (2001).

175. Richardson, J. S. & Richardson, D. C. Natural -sheet proteins use negative design to avoid edge-to-edge aggregation. *Proceedings of the National Academy of Sciences* **99**, 2754–2759 (2002).

176. Chaney, J. L. *et al.* Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLoS Comput Biol* **13**, e1005531 (2017).

177. Stein, K. C. & Frydman, J. The stop-and-go traffic regulating protein biogenesis: How translation kinetics controls proteostasis. *Journal of Biological Chemistry* **294**, 2076–2084 (2019).

178. Zhao, T. *et al.* Disome-seq reveals widespread ribosome collisions that promote cotranslational protein folding. *Genome Biology* **22**, 16 (2021).

179. Dalvai, M. *et al.* A Scalable Genome-Editing-Based Approach for Mapping Multiprotein Complexes in Human Cells. *Cell Reports* **13**, 621–633 (2015).

180. Fonseca, J. P. *et al.* A Toolkit for Rapid Modular Construction of Biological Circuits in Mammalian Cells. *ACS Synth. Biol.* **8**, 2593–2606 (2019).

181. Chu, V. T. *et al.* Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat Biotechnol* **33**, 543–548 (2015).

182. Muller, R., Meacham, Z. A., Ferguson, L. & Ingolia, N. T. CiBER-seq dissects genetic networks by quantitative CRISPRi profiling of expression phenotypes. *Science* **370**, eabb9662 (2020).

183. McGlincy, N. J. *et al.* A genome-scale CRISPR interference guide library enables comprehensive phenotypic profiling in yeast. *BMC Genomics* **22**, 205 (2021).

184. Susorov, D., Egri, S. & Korostelev, A. A. Termi-Luc: a versatile assay to monitor full-protein release from ribosomes. *RNA* **26**, 2044–2050 (2020).

185. Koster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

186. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. 4 (2012).

187. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

188. Gasteiger, E. *et al.* Protein Identification and Analysis Tools on the ExPASy Server. in *The Proteomics Protocols Handbook* (ed. Walker, J. M.) 571–607 (Humana Press, 2005). doi:10.1385/1-59259-890-0:571.

189. Lamiable, A. *et al.* PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic Acids Research* **44**, W449–W454 (2016).

190. Nagy, G., Igaev, M., Jones, N. C., Hoffmann, S. V. & Grubmüller, H. SESCA: Predicting Circular Dichroism Spectra from Protein Molecular Structures. *J. Chem. Theory Comput.* **15**, 5087–5102 (2019).

191. Gebru, A. A. *et al.* Global burden of COVID-19: Situational analyis and review. *HAB* **29**, 139–148 (2021).

192. Hogg, J. R. Viral Evasion and Manipulation of Host RNA Quality Control Pathways. *J Virol* **90**, 7010–7018 (2016).

193. Lauring, Adam S., Acevedo, A., Cooper, Samantha B. & Andino, R. Codon Usage Determines the Mutational Robustness, Evolutionary Capacity, and Virulence of an RNA Virus. *Cell Host & Microbe* **12**, 623–632 (2012).

194. Duffy, S. Why are RNA virus mutation rates so damn high? *PLoS Biol* **16**, e3000003 (2018).

195. Jenkins, G. M. & Holmes, E. C. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Research* **92**, 1–7 (2003).

196. Pandit, A. & Sinha, S. Differential Trends in the Codon Usage Patterns in HIV-1 Genes. *PLoS ONE* **6**, e28889 (2011).

197. Pavon-Eternod, M. *et al.* Vaccinia and influenza A viruses select rather than adjust tRNAs to optimize translation. *Nucleic Acids Research* **41**, 1914–1921 (2013).

198. Mogro, E. G., Bottero, D. & Lozano, M. J. Analysis of SARS-CoV-2 synonymous codon usage evolution throughout the COVID-19 pandemic. *Virology* **568**, 56–71 (2022).

199. Babar, M. M. & Zaidi, N.-S. S. Protein sequence conservation and stable molecular evolution reveals influenza virus nucleoprotein as a universal druggable target. *Infection, Genetics and Evolution* **34**, 200–210 (2015).

200. Voitenko, O. S., Dhroso, A., Feldmann, A., Korkin, D. & Kalinina, O. V. Patterns of amino acid conservation in human and animal immunodeficiency viruses. *Bioinformatics* **32**, i685–i692 (2016).

201. Cagliani, R., Forni, D., Clerici, M. & Sironi, M. Coding potential and sequence conservation of SARS-CoV-2 and related animal viruses. *Infection, Genetics and Evolution* **83**, 104353 (2020).

202. Simeoni, M., Cavinato, T., Rodriguez, D. & Gatfield, D. I(nsp1)ecting SARS-CoV-2–ribosome interactions. *Commun Biol* **4**, 715 (2021).

203. Fisher, T. *et al.* Parsing the role of NSP1 in SARS-CoV-2 infection. *Cell Reports* **39**, 110954 (2022).

204. Schubert, K. *et al.* SARS-CoV-2 Nsp1 binds the ribosomal mRNA channel to inhibit translation. *Nat Struct Mol Biol* **27**, 959–966 (2020).

205. Kamitani, W. *et al.* Severe acute respiratory syndrome coronavirus nsp1 protein suppresses host gene expression by promoting host mRNA degradation. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 12885–12890 (2006).

206. Finkel, Y. *et al.* SARS-CoV-2 uses a multipronged strategy to impede host protein synthesis. *Nature* **594**, 240–245 (2021).

207. Zhang, K. *et al.* Nsp1 protein of SARS-CoV-2 disrupts the mRNA export machinery to inhibit host gene expression. *Sci. Adv.* **7**, eabe7386 (2021).

208. Krafcikova, P., Silhan, J., Nencka, R. & Boura, E. Structural analysis of the SARS-CoV-2 methyltransferase complex involved in RNA cap creation bound to sinefungin. *Nat Commun* **11**, 3717 (2020).

209. Firth, A. E. A putative new SARS-CoV protein, 3c, encoded in an ORF overlapping ORF3a. *Journal of General Virology* **101**, 1085–1089 (2020).

210. Jungreis, I., Sealfon, R. & Kellis, M. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. *Nat Commun* **12**, 2642 (2021).

211. Cao, J. & Geballe, A. P. Translational inhibition by a human cytomegalovirus upstream open reading frame despite inefficient utilization of its AUG codon. *J Virol* **69**, 1030–1036 (1995).

212. Sachs, M. S. & Geballe, A. P. Downstream control of upstream open reading frames: Figure 1. *Genes Dev.* **20**, 915–921 (2006).

213. Simms, C. L., Yan, L. L., Qiu, J. K. & Zaher, H. S. Ribosome Collisions Result in +1 Frameshifting in the Absence of No-Go Decay. *Cell Reports* **28**, 1679–1689.e4 (2019).

214. Houston, L., Platten, E. M., Connelly, S. M., Wang, J. & Grayhack, E. J. Frameshifting at collided ribosomes is modulated by elongation factor eEF3 and by integrated stress response regulators Gcn1 and Gcn20. 21.

215. Matreyek, K. A., Stephany, J. J., Chiasson, M. A., Hasle, N. & Fowler, D. M. An improved platform for functional assessment of large protein libraries in mammalian cells. *Nucleic Acids Research* gkz910 (2019) doi:10.1093/nar/gkz910.

216. Hermann, M. *et al.* Binary recombinase systems for high-resolution conditional mutagenesis. *Nucleic Acids Research* **42**, 3894–3907 (2014).