Noncoding translation mitigation

Jordan S. Kesner

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

© 2022

Jordan S. Kesner

All Rights Reserved

Abstract

Noncoding translation mitigation

Jordan S. Kesner

In eukaryotes, sequences that code for the amino acid structure of proteins represent a small fraction of the total sequence space in the genome. These are referred to as coding sequences, whereas the remaining majority of the genome is designated as noncoding. Studies of translation, the process in which a ribosome decodes a coding sequence to synthesize proteins, have primarily focused on coding sequences, mainly due to the belief that translation outside of canonical coding sequences occurs rarely and with little impact on a cell. However, recently developed techniques such as ribosome profiling have revealed pervasive translation in a diverse set of noncoding sequences, including long noncoding RNAs (lncRNAs), introns, and both the 5' and 3' UTRs of mRNAs. Although proteins with amino acid sequences derived partially or entirely from noncoding regions may be functional, they will often be nonfunctional or toxic to the cell and therefore need to be removed. Translation outside of canonical coding regions may further expose the noncoding genome to selective pressure at the protein level, leading to the generation of novel functional proteins over evolutionary timescales. Despite the potentially significant impact of these processes on the cell, the cellular mechanisms that function to detect and triage translation in diverse noncoding regions, as well as how peptides that escape triage may evolve into novel functional proteins, remain poorly understood.

This thesis will describe novel findings that offer new insight into the process of noncoding translation mitigation revealed by a combination of high-throughput systems-based approaches and validated by biochemical and genetic approaches. Chapter 1 will discuss general

concepts in the translation of noncoding sequences and the relevant cellular systems and impacts on human health. Chapter 2 will discuss the results of a high-throughput reporter assay investigating translation in thousands of noncoding sequences from diverse sources. The results discussed in this chapter revealed two factors involved in the mitigation of proteins derived from noncoding sequences: C-terminal hydrophobicity and proteasomal degradation. Chapter 3 will build on Chapter 2 and discuss the results of a genome-wide CRISPR/Cas9 knockout screen that identified the BAG6/TRC35/RNF126 membrane protein chaperone complex as a key cellular pathway in the detection and degradation of proteins with translated noncoding sequences. Having identified the BAG6 complex as targeting a specific reporter of translation of the 3’ UTR in the AMD1 gene, a series of knockout cell lines validated these results and demonstrated the participation of two additional genes, SGTA and UBL4A. Through coimmunoprecipitation western blots and rescue assays with flow cytometry as a readout, we confirmed physical interaction between BAG6 and the 3’ UTR of AMD1, and a similar experiment confirmed interaction between BAG6 and a readthrough mutant of the SMAD4 tumor suppressor gene. Finally, by combining our high-throughput reporter library with our BAG6 knockout cell line, we demonstrated that BAG6 targets hydrophobic C-terminal tails in many noncoding sequences of diverse origin. Finally, Chapter 4 will discuss the evolutionary perspective of noncoding translation through analyses of the sequence content of human and mouse genomes. The findings of this chapter demonstrate a significant trend for increased uracil content in noncoding regions of the genome, which frequently results in the translation of hydrophobic amino acids. We also find that many functional translated noncoding peptides localize to membranes, providing a theoretical link between the shuttling of translated noncoding sequences to a protein complex involved in membrane protein quality control and the emergence of newly evolving proteins from the noncoding genome.

Table of Contents

List of Figures

Acknowledgments

I would first like to acknowledge my advisor, Dr. Xuebing Wu, who I have the great fortune of having as my mentor. Throughout my time in his lab, Dr. Wu has shown a great personal interest in my development as a scientist and has always been willing to share his considerable knowledge and expertise. I am immensely grateful to have had the opportunity to train with Dr. Wu, and only through his guidance and leadership would this work have been possible. I have learned much from Dr. Wu, I thank him for his outstanding mentorship, and I thoroughly look forward to continuing to learn from him in the future.

I thank the members of my committee, Dr. Peter Sims, Dr. Chaolin Zhang, Dr. Natura Myeku, and Dr. Mohammed AlQuraishi, all of whom I have had the pleasure of working with and who have been invaluable sources of support and feedback.

I am grateful to have worked with many talented members of the Columbia community, including all of the members of the Wu lab and the Cardiometabolic Genomics Program. I thank Dr. Jeremy Worley, Dr. Prem Subramaniam, Dr. Mikko Turunen, and Adina Grunn for the many skills they taught me.

I thank Dr. Peter Sims, Dr. Ronald Liem, Dr. Donna Farber, and Zaia Sivo Nouara for their dedication to guiding me through the sometimes complex process of PhD training.

To my friends and family, I thank you for your support and advice in all things science and beyond. I thank all of my friends from the Columbia Integrated program and the Master's Program in Biotechnology.

To my parents, Harvey and Renee, my siblings Devin and Josh, my partner Yocelyn, and my confidant Topias, for their unwavering love, support, and friendship throughout the years.

Dedication

Dedicated to Leo and Helen Kesner, for inspiring and nurturing a love of science from the beginning.

Chapter 1: Introduction

1.1 Translation in coding and noncoding sequences

1.1.1 Noncoding sequences in transcription and translation

For eukaryotic cells, the path along the central dogma of DNA to RNA to protein presents many complexities. These complexities are reflected in the structure and nature of the genome, where only about 1% of the total genomic sequence is predicted to encode the amino acid sequence of functional proteins. The transcriptional machinery is known to have relatively low specificity, resulting in widespread production of transcripts from noncoding regions of the genome[1–3]. However, accumulation of these transcripts is prevented in cells through a fail-safe mechanism that recognizes abundant poly(A) signals in the noncoding genome and functions to suppress pervasive transcription in mammalian cells by degrading transcripts[4,5]. Nonetheless, noncoding transcripts such as long noncoding RNAs (lncRNAs) are prevalent in the transcriptome and are often translated[6]. Further, aberrant RNA splicing and polyadenylation can generate mRNAs containing noncoding sequences derived from introns and UTRs within their open reading frames (ORFs)[7–9]. This presents a potential problem to the cell, as noncoding sequences contained in lncRNAs and aberrant mRNAs that escape quality control pathways are now subject to translation by the ribosome.

Several studies have found through analysis of ribosome profiling data that translation outside of canonical coding sequences (CDS) is widespread[10,11]. Most cytoplasmic lncRNAs in mouse embryonic stem cells were found to have ribosome footprints indistinguishable from footprints in mRNA CDS, indicating active translation of lncRNAs[6]. In human cells, it was estimated that up to 40% of lncRNAs, 35% of mRNA 5' UTRs, and 4% of mRNA 3' UTRs are actively translated[12]. Other studies have found evidence for translation in mRNAs with retained introns and mRNAs resulting from intronic polyadenylation[13,14].

Defects in mRNA quality control and processing pathways can cause an accumulation of aberrant mRNAs within the cell, potentially increasing the rate of noncoding translation. Such defects commonly manifest in various pathological conditions, such as cancer, aging, and neurodegenerative disorders[14–26]. Similarly, nonstop mutations in some genetic disorders and cancers can lead to translation of mRNA 3' UTRs due to readthrough of the mutated stop codon[27,28]. 3' UTR translation can also occur due to malfunction of the translational machinery and as a side effect of aminoglycoside drugs[29,30].

Despite the relevance of noncoding translation to human health, little is known about the cellular mechanisms of surveillance and triage of proteins produced as a result of translation in noncoding sequences. Noncoding sequences are unlikely to be exposed to the same selective constraints as canonical coding sequences, increasing the chances that proteins containing translated noncoding sequences are nonfunctional, toxic, or prone to aggregation. Removing these potentially harmful proteins is therefore critical to maintaining proteostasis within the cell.

Studies to date investigating mechanisms of noncoding protein surveillance have utilized limited sets of reporters of 3' UTR translation and have come to very different conclusions as to how cells handle translated noncoding sequences[28,31–34]. Models proposed from these various studies regarding the fate of proteins with translated noncoding sequences include proteasomal degradation, aggregation in lysosomes, and translational arrest induced by ribosome stalling. The lack of consensus from these studies highlights the need for more systematic studies that could potentially identify common principles and mechanisms involved in the cellular surveillance and triage of proteins with translated noncoding sequences.

1.1.2 Translation in canonical coding sequences

In eukaryotes, capped, polyadenylated, and spliced mRNA molecules are targeted for translation by the ribosome in a highly regulated and complex process[35]. mRNAs generally consist of 5' and 3' untranslated regions (UTRs) as well as a coding sequence (CDS) that is delimited by a start (AUG) and stop codon[35]. The coding sequence defines a set of codons that encode a chain of amino acids to be translated by the ribosome[35]. In contrast, the 5' and 3' UTRs generally serve regulatory functions related to transcription and translation[36].

Translation is composed of three main stages: initiation, elongation, and termination. In the initiation phase, a complex of eukaryotic initiation factors (eIFs) recognizes the 7-methylguanosine (m^7G) cap at the 5' end of the mRNA molecule[35]. Another group of eIFs, along with a 40S ribosomal subunit and a Met-tRNAi, form a complex known as the 43S pre-initiation complex (43S PIC), which then attaches to the activated mRNA and begins a process of scanning along the 5' UTR for the start codon of the CDS[35]. The PIC completes scanning upon recognition of the AUG codon in the P-site, at which point a GTP is hydrolyzed to GDP, allowing the large 60S ribosomal subunit to join the complex[35]. The full 80S initiation complex is now ready to begin the process of elongation within the coding sequence[35].

During the elongation phase of translation, the 80S ribosome synthesizes the amino acid chain, starting with the methionine residue attached to the initiator tRNA[37]. The 80S ribosome first decodes the codon by matching it to its complementary tRNA carrying an amino acid[37]. With the correct tRNA positioned in the A site of the ribosome, a peptidyl bond is formed between the amino acid carried by the tRNA and the previous amino acid in the chain[37]. The growing chain is then translocated back to the P site, at which point the process repeats, starting with the next codon in the coding sequence[37].

Upon recognizing a stop codon in the coding sequence, the ribosome initiates the process of translation termination so that the full-length amino acid chain can be released[37]. When a termination codon is detected in the A site, it is recognized by the termination factor eRF1, which coordinates with the GTPase eRF3 to catalyze the termination of the elongation phase and allow the release of the fully synthesized peptide from the ribosome[37]. Following release of the peptide, the 80S complex is disassembled from the mRNA and its subunits are released to be recycled for translation of additional mRNA molecules[37].

1.1.3 Quality control pathways in translation

Many things can go wrong during translation, for which cells have evolved specific quality control mechanisms that function to fix or terminate faulty instances of translation. Many factors can induce ribosome stalling, resulting in the temporary or complete stoppage of a translating ribosome[38,39]. Factors that may cause stalling in ribosomes include the presence of two or more adjacent proline codons in a coding sequence, insufficient access to specific tRNAs, or stable secondary structure in mRNA[39,40]. Aberrant mRNAs that lack stop codons due to truncation or improper polyadenylation can similarly result in ribosome stalling[39,40].

Mechanisms that function to detect and triage ribosome stalling can differ depending on the cause of ribosome stalling. Stalling induced due to the difficulty ribosomes have in forming proline-proline peptidyl bonds, for example, can be rescued by a protein known as eIF5A, which helps to catalyze the formation of the difficult peptide bond[38]. Ribosomal stalling that is caused by the lack of a stop codon in an aberrant mRNA (Nonstop), or due to structural or codon features of the ORF (No-go), activates the nonstop mRNA decay pathway (NSD) and the no-go decay pathway (NGD) respectively, both of which result in degradation of the offending mRNA[40]. Both NSD and NGD require translation of the faulty mRNA to be detected by the cell, and thus they will necessarily produce an aberrant peptide prior to being degraded[40]. These peptides are detected and targeted by the ribosome quality control complex (RQC), whose activation can lead to ribosome recycling or degradation, degradation of the peptide, and activation of cellular stress responses[39].

1.1.4 Translation outside of canonical coding sequences

In addition to translation of canonical coding sequences, ribosomes can also translate noncoding sequences. The mechanism by which this noncoding translation occurs differs depending on the type of noncoding sequence being translated.

1.1.5 Non-canonical translation: long non-coding RNA (lncRNA)

Long non-coding RNAs (lncRNAs) are a class of RNA that share many similarities with mRNA, except that they typically do not contain open reading frames or may contain short open reading frames[41]. Like mRNA, lncRNAs undergo splicing, capping at their 5’ ends, and polyadenylation at their 3’ ends[41]. It has been estimated that up to 40% of identified lncRNAs in humans are translated, although it is likely that many of these translation events do not result in functional peptides due to their difficulty of detection[12]. Despite this, several examples in the literature have identified peptides derived from lncRNA translation with functional roles in development[42,43], physiology[44], and cell migration[45]. It is likely that other functional peptides derived from the translation of lncRNAs have yet to be discovered.

1.1.6 Non-canonical translation: 5’ UTR upstream ORF (uORF)

Upstream ORFs (uORF) are defined as small open reading frames located within the 5’ UTR of an mRNA, originating from an upstream start codon (uAUG) that is 5’ to the start codon of the canonical coding sequence[46]. These uORFs are widespread in eukaryotic genomes and function as a mechanism of translational regulation of protein expression[46]. uORFs are believed to reduce the expression of the corresponding CDS by ‘sequestering’ the translational activity of ribosomes[46]. A scanning PIC in the 5’ UTR of an mRNA that contains an uORF may recognize the uAUG and initiate translation of the full uORF, only after which translation of the canonical CDS can begin[46]. In this scenario, translation of the uORF will also interfere with the ability of additional scanning PICs to initiate translation at the canonical CDS[46]. Alternatively, a ribosome translating a uORF may fail to properly terminate at the uORF stop codon, in which case the stalled ribosome will trigger nonsense-mediated decay (NMD) of the mRNA[46].

From the perspective of the uAUG, a uORF can be placed into one of three classes depending on the positioning of the next in-frame stop codon. A uORF with an in-frame stop codon located upstream of the start codon of the canonical coding sequence is classified as a nonoverlapping uORF[46]. If the stop codon is downstream of the canonical start codon but not in the same frame, it is an out-of-frame overlapping uORF[46]. Finally, if the next in-frame stop codon is the stop codon of the CDS, the uORF is classified as an N-terminal extension[46].

Studies leveraging ribosome profiling have suggested widespread translation of uORFs occurs in humans, with estimates as high as 35% of all uORFs undergoing translation[12]. By analyzing the codon optimality and evolutionary conservation of uORFs, it has been suggested that most translational events in uORFs do not lead to functional peptides and instead serve only as a mechanism to suppress translation of the downstream CDS[47,48]. One previous study examined the RNA expression level of approximately 8000 transcripts containing zero, one, two, or three predicted uORFs, and found that a large majority of the uORF-containing transcripts had lower levels of expression than those that did not contain any uORFs, and that expression tended to decrease further as the number of uORFs in a transcript increased[48]. However, some examples of functional peptides derived from the translation of 5’ UTRs have been found[11].

1.1.7 Non-canonical translation: 3’ UTR readthrough

The 3’ UTR of an mRNA molecule begins immediately after the stop codon of the CDS. During translation of a given CDS, recognition of a stop codon by the ribosome and proper execution of the termination phase of translation will ensure that the nucleotide sequence of the 3’ UTR is not decoded by the ribosome and added to the growing peptide chain. Mechanistically, translation of the 3’ UTR sequence by a ribosome originating from the CDS start codon can occur due to malfunction of the translational machinery or due to a missense mutation converting the native stop codon of a transcript into a sense codon[28,29]. Ribosomal readthrough of the native stop codon can occur stochastically, is influenced by the sequence content surrounding the stop codon[29], can be controlled by specific cellular pathways[49], and can be induced by treatment with drugs such as aminoglycosides[29]. Several studies have also used engineered tRNA variants to bypass nonstop mutations[50,51].

Translation in 3’ UTRs can result in a number of outcomes. Readthrough of the native stop codon of a coding sequence may yield a protein variant extended at its C-terminal end with its length determined by the positioning of the next in-frame stop codon in the 3’ UTR[28]. In the absence of a downstream in-frame stop codon, the ribosome will eventually translate the poly(A) tail at the 3’ end of the UTR, which will trigger targeting of the mRNA for degradation by the non-stop decay mediated mRNA surveillance pathway[38]. For C-terminal extended readthrough proteins that are released, there may be various effects on the protein and cell. Depending on the specific gene, length of the tail extension, and amino acid content of the extension, the resulting protein may be degraded[27], may acquire cellular toxicity[52], or may become prone to aggregation[53]. One study proposed a model in which stop codon readthrough in the AMD1 gene causes ribosome stalling in the downstream 3’ UTR, eventually leading to the formation of a ribosomal queue that extends into the coding sequence and inhibits translation of the affected mRNA[34].

1.1.8 Non-canonical translation: Introns

Translation of intronic sequences can originate from aberrant mRNAs produced through intronic retention or intronic polyadenylation[7–9]. While intron retention can occur due to malfunction of splicing, it has also been shown to play a role in regulating gene expression and can lead to the generation of novel protein isoforms[54–56]. Although mRNA with retained introns are subject to degradation via quality control mechanisms such as NMD, they are detectable in the transcriptome using modern sequencing technologies[57]. Interestingly, increased retention of introns is a common feature of many types of cancers, neurodegenerative disorders, and aging, likely due to increasing dysfunction of the spliceosome[18–20]. Activation of the hypoxia stress response in tumor microenvironments inhibits NMD and can further increase the accumulation of intron-containing mRNAs[16,58,59]. Translation of aberrant mRNAs containing retained introns has previously been detected via ribosome profiling[13], and widespread translation of mRNAs resulting from intronic polyadenylation has been demonstrated by detecting truncated proteins in leukemia[14].

1.1.9 Non-canonical translation: Downstream ORFs (dORF)

Similar to upstream ORFs in mRNA 5’ UTRs, downstream ORFs (dORF) residing in the 3’ UTRs of mRNAs have been characterized. Translation of dORFs has been shown by recent studies involving ribosome profiling[11,60]. Another study identified an IRES in the 3’ UTR of *GTP cyclohydrolase1* (*GCH1*), which was able to initiate translation in a manner that is resistant to cap-dependent translation initiation inhibitors (Torin 1)[61]. Another study analyzed the transcriptomes of both humans and zebrafish and determined that in both cases, greater than 80%

of transcripts contained potential small ORFs (10-100 aa) in their 3' UTRs[62]. By further analyzing ribosome profiling data, this same study identified evidence of translation in 1,406 human and 1,153 zebrafish putative dORFs[62]. Interestingly, this study suggested that translation of dORFs acts as a mechanism to enhance translation of the primary ORF in a transcript, in direct contrast to the function of most uORFs[63]. Amino acid sequence conservation is only seen in a small number of these dORFs (6 conserved, 141 weakly conserved), suggesting that in a majority of cases, dORFs function as a mechanism of translational regulation that depends on translation in the dORF itself rather than through the production of a functional peptide[62]. However, this does not exclude the possibility that some dORFs result in functional micropeptides, and two such peptides with functions related to cell proliferation have recently been characterized[11]. Importantly, as this appears to be a relatively widespread mechanism of translational regulation, the cellular fate of micropeptides derived from translation of dORFs, regardless of whether they are functional or not, has yet to be characterized.

1.2 Proteostasis and pathways that maintain the proteome

1.2.1 Protein folding and molecular chaperones

Proteins are significant contributors to the processes that regulate and maintain cellular homeostasis. As cells expend a considerable fraction of their available energy in the process of protein synthesis, it is critical that proteins are correctly produced, folded, trafficked, and maintained so that they are able to function properly[64–66]. Proteins are highly complex molecules with a vast array of functions and may require post-translation modification, chaperoned folding or trafficking, targeted transport to subcellular locations, multimerization, and cleavage to function correctly. Accordingly, many cellular processes and mechanisms are dedicated to ensuring the fidelity of the processes involved in proper proteostasis. In general, cells utilize three primary methods of protein quality control for maintaining proteostasis: re-folding of proteins, degradation of proteins, and sequestration of proteins[64]. The proteostatic processes involved are primarily regulated by a class of proteins known as chaperones, and there is significant crosstalk between the three approaches[64]. These protein chaperones help newly synthesized proteins fold correctly, guide transport of proteins across cellular membranes, make attempts to refold misfolded protein substrates, and, if impossible, target misfolded proteins for degradation or sequestration[64].

One core component of maintaining proteostasis is ensuring the proper folding of newly synthesized proteins[66]. In addition to ensuring correct protein folding, cells must constantly contend with the propensity of newly synthesized proteins to form insoluble aggregates[67]. Unchecked protein aggregation within the cell can result in several adverse effects, including loss of the function of the aggregated protein and the formation of potentially toxic collections of insoluble protein aggregates such as amyloid fibrils[66]. The propensity for protein aggregate formation prior to folding into the correct tertiary structure arises from exposed hydrophobic residues and unstructured domains that are subject to interaction and association with other proteins in solution. Once fully folded, these regions are protected from aggregation-prone interactions by being buried within the structure of the protein[66]. In cells, the tendency of unfolded proteins to aggregate through these types of interactions is further exacerbated by factors including the relatively high level of molecular crowding, as well as the presence of polyribosome complexes[66].

To counteract the potential for nascent peptides to form aggregates, cells expend considerable energy to ensure the proper folding of newly synthesized peptides through the protein chaperone network. Proteins within this network can be broadly classified into two groups based on the mechanism through which they aid in the folding of newly synthesized proteins. Chaperones such as the members of the nascent polypeptide-associated complex (NAC) associate with ribosomes and stabilize nascent polypeptide chains during translation[66]. On the other hand, chaperones such as the members of chaperonin complexes form large structures that capture and insulate unfolded proteins from the cellular cytosol after translation is completed, providing a protected chamber within which the captured protein can fold into its native structure[66]. Other types of chaperones protect proteins from misfolding due to environmental factors, such as increases in temperature[68]. Yet other chaperones possess some ability to disassociate small groups of aggregated proteins[64].

1.2.2 Cellular systems for removing aberrant proteins

Proteins that fail to fold correctly despite the activity of the chaperone network will acquire misfolded structures and require removal from the cell to prevent any potential adverse effects. Removal of these misfolded proteins is primarily achieved through their degradation in the ubiquitin-proteasome system (UPS)[64]. Misfolded proteins destined for degradation by the UPS are tagged with a chain of a small protein called ubiquitin by proteins known as E3 ubiquitin ligases[69]. These poly-ubiquitinated proteins are subsequently transported to a protein complex known as the 26S proteasome[69]. The 26S proteasome cleaves the polyubiquitinated protein into short peptide chains, thus removing the misfolded protein from the cell[69].

A separate pathway for removing unwanted proteins from the cell is known as the autophagy-lysosomal pathway (ALP), which can target protein aggregates, misfolded proteins, and other cellular molecules[70]. Within this pathway, proteins destined for degradation are first sequestered from the cell by the formation of a membrane-bound organelle known as an autophagosome[70]. These autophagosomes subsequently fuse with larger membrane-bound organelles known as lysosomes[70]. The lumen of lysosomes is highly acidic and contains a multitude of enzymes that are active at low pH and degrade the contents of the lysosome[70]. The products of lysosomal degradation can then be released from the lysosome and recycled by the cell[70].

Misfolded or aggregated proteins unable to be cleared by the UPS or the ALP may be spatially sequestered by cells into specialized compartments known as inclusions[64]. In addition to preventing potentially harmful interactions between healthy components of the cell and misfolded or aggregated proteins, inclusions may maintain proteostasis by sequestering proteins that require refolding or degradation so as not to overwhelm the other branches of the proteostatic machinery[64].

1.3 The relevance of translation in noncoding sequences to human health

1.3.1 Aging

In humans, aging is associated with a progressive and far-reaching decline in the litany of systems that maintain cellular health. Although the impact on an individual's overall health from the increasing dysfunction of any specific system varies widely, human health generally tends to deteriorate along with the deterioration of the molecular machinery as aging progresses. In the context of noncoding translation, there are several salient ways in which cellular aging influences human health due to the deterioration of specific molecular systems.

A study investigating detectable changes in the transcriptome due to aging among a cohort of 698 human subjects found that the most significant age-related changes were in genes involved in RNA processing pathways, including splicing and polyadenylation[71]. Other studies have similarly reported increasing malfunction of RNA processing pathways in aging cells[19,21]. In humans, one study identified are-related increases in intron retention in 43 tissues across a cohort of 948 subjects[21]. The impairment of RNA processing due to aging-related processes may lead to the accumulation of aberrant mRNAs with retained introns or polyadenylated introns, which ribosomes can translate.

1.3.2 Cancer

Cancer is one of the most widespread and important diseases affecting human health, and the relationship between noncoding translation and cancers is multifaceted. One recent publication analyzed data from the COSMIC database spanning 62 tumor types in search of nonstop extension mutations, which are genetic mutations that bypass the native stop codon of a CDS and result in a C-terminally extended protein with a peptide tail encoded by the 3’ UTR[27]. Among the results of this study, the most frequently mutated protein was found to be the tumor suppressor protein SMAD4, the loss of which plays a role in various cancers[27]. Readthrough of the *SMAD4* stop codon induced by this set of mutations leads to a 40 amino-acid extension at the C-terminal end of the protein, which in turn causes it to be targeted for proteasomal degradation resulting in almost undetectable levels of the protein[27].

Several factors associated with cancers affect the splicing process and can subsequently increase translation of noncoding sequences. SF3B1, a protein that is a critical component of the spliceosome, was found to be mutated in 15% of patients diagnosed with Chronic Lymphocytic Leukemia (CLL)[72]. Among patients with mutations in SF3B1, there are significant changes in pre-mRNA splicing patterns[72]. A study focusing on cancers driven by the *MYC* oncogene identified synthetic lethality between mutated *MYC* and *BUD31*, another core component of the spliceosome[15]. Mutations in the oncogene *KRAS* in human lung epithelial cells caused significant differences in mRNA isoform expression compared to cells expressing wild-type *KRAS*[73]. Another study focusing on myelodysplasia found frequent mutations in several genes involved in 3’ splice site recognition, leading to abnormalities in RNA splicing[17]. Furthermore, induction of the hypoxic stress response in tumor microenvironments inhibits the NMD RNA quality control pathway and leads to further accumulation of aberrant mRNAs[58].

Interestingly, a number of studies have found that antigens specific to tumor cells are primarily derived from translation occurring in noncoding sequences and that the detection of these peptides is correlated with poor patient prognosis[74–77]. Further, antigens derived from noncoding sequences represent a potential source of cancer-specific targets for the development of novel therapeutics[74].

1.3.3 Genetic diseases

Translation of noncoding regions due to the inheritance of genetic mutations has been linked to many genetic disorders. A readthrough mutation in the *REEP1* gene that causes an aggregation-promoting C-terminal extension of the protein has been identified in Charcot-Marie-Tooth disease, a condition affecting neurons in the peripheral nervous system[53]. Malfunctioning of the *RAB39B* gene has been shown to be involved in Autism Spectrum Disorder (ASD) and Intellectual Disability (ID), and a nonstop mutation causing a 21 amino acid C-terminal extension to *RAB39B* was recently identified in a family with a history of both ASD and severe ID[78]. A homozygous nonstop mutation in the *PDE6C* gene found in a consanguineous family affected by Cone-rod dystrophy (CORD) has been proposed to cause the disease phenotype[79]. A nonstop mutation in the *MITF* gene implicated in causing type 2 Waardenburg syndrome (WS2) extends the protein by 33 amino acids and results in both nuclear and cytoplasmic localization, whereas the wild-type *MITF* is located exclusively in the nucleus[80]. Alterations to the localization of *MITF* were shown to induce haploinsufficiency and result in the WS2 phenotype[80]. In addition to the examples listed above, disease-causing genetic mutations resulting in C-terminal extensions have been identified in the genes CRYM, DBT, ITM2B, SH2D1A, PAX6, MOCS2, CTSK, FKRP, RUNX2, IKBKG, PNPO, HBA2, SHOX, NHP2, and FHL1[28,81–86].

Nonsense mutations convert sense codons to stop codons, resulting in the introduction of a premature termination codon (PTC) in protein-coding sequences[29]. Interestingly, such PTCs tend to be read through by ribosomes more efficiently than native stop codons, with the likelihood of readthrough influenced by factors such as the specific stop codon identity as well as the sequence content surrounding the stop codon[29]. Such mutations have been found to account for approximately 11% of all inherited human genetic diseases[29]. Introducing a PTC in a gene coding sequence often results in a decrease or complete loss of protein function, either through destabilization of the mutated mRNA or expression of a truncated protein[29]. The large number of genetic diseases caused by nonsense mutations makes them an attractive therapeutic target, and some recent studies have investigated methods for rescuing proteins containing PTC[29]. One recent study investigated the use of aminoglycoside antibiotics to stimulate stop codon readthrough, a potential mechanism for restoring full-length protein production from disease-causing genes with PTC[29]. Although the aminoglycoside drugs were effective at inducing readthrough of PTC, they also caused a global increase in readthrough of native stop codons, resulting in widespread translation of gene 3’ UTRs[29]. Potential drugs developed to induce readthrough of PTC for genetic diseases will have to consider the effects of translation into the 3’ UTRs of off-target genes[29].

1.3.4 Neurodegenerative diseases

The accumulation of aberrant mRNAs in the transcriptome is connected to various neurodegenerative diseases primarily through disruptions in splicing[19,20,24,87–94]. Alzheimer’s disease (AD) is a progressive neurodegenerative disorder in which the disease-causing mechanism is believed to be the aggregation of a protein known as Tau, causing neuronal death[20]. Interestingly, several components of the U1 snRNP, which plays a critical role in splicing by binding to the 5’ exon-intron junction in pre-mRNA, have been found to aggregate with Tau in AD[20]. The aggregation of U1 snRNP components with Tau disrupts splicing, resulting in the accumulation of aberrant mRNAs with retained introns and cryptic splice junctions[20].

Amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) are related neurodegenerative diseases caused by a dipeptide repeat in *C9orf72*[25]. The mutant C9orf72 protein has been shown to inhibit the NMD RNA quality control pathway by disrupting the RNA helicase UPF1[25]. Disrupting this pathway allows aberrant mRNA to accumulate in ALS/FTD, a known hallmark of the disease[25]. Interestingly, restoring the function of UPF1 has been found to protect against the neurotoxic effects of mutant C9orf72, suggesting that the malfunction of the NMD pathway is a critical component of the disease pathology[26].

Many neurodegenerative diseases tend to begin or worsen with age, likely due to the increasing levels of malfunction in various cellular processes associated with aging. The progressively increasing dysfunction in cellular functions is likely to make cells more susceptible to the disease-causing processes that underly neurodegenerative diseases. In one study that utilized ribosome profiling, aging neurons were found to be actively translating the 3’ UTRs of hundreds of different genes[30]. Translation of these 3’ UTRs was further found to be correlated with oxidative stress and resulted in the production of many peptides derived from 3’ UTR sequences of unknown function[30].

1.4 A lack of consensus on the cellular mechanisms of triage of proteins with translated noncoding sequences

Studies investigating the cellular surveillance mechanisms involved in mitigating the buildup of nonfunctional or toxic proteins resulting from the translation of noncoding sequences have primarily focused on a small number of reporters of 3' UTR translation[25–30]. This section will summarize the findings and methods of these studies relating to noncoding translation mitigation.

1.4.1 Arribere et al. 2016

Arribere et al. utilized a paired reporter system and investigated reporter loss due to translation of 3' UTRs. They selected nine genes in C. elegans designed to represent a variety of cellular functions, where each gene was required to have an in-frame stop codon at least 30 amino acids from the native stop codon and also upstream of any known polyadenylation site. In most of these nine cases, they found that there was at least a ten-fold reduction in the reporter with the translated UTR as compared to the control reporter. They further found that synonymous mutations in the UTR sequences did not rescue the reporter, as well as finding that loss of the reporter required physical linkage between the reporter and the UTR amino acid sequences. These results strongly suggested that loss of the reporter occurs at the protein level, after or co-translationally, rather than due to some mechanism that functions prior to translation.

To exclude the possibility of these findings being the result of artifacts derived from the use of reporter systems, the authors next identified a set of endogenous genes in C. elegans for which readthrough translation could be studied. A group of five genes was identified with prerequisites similar to the first nine genes tested but were also required to have an observable phenotypic readout. Comparable losses of translated protein were observed in these endogenous cases in C. elegans. By comparing protein levels to RNA abundance and ribosome loading in one of the genes tested (unc-54), it was concluded that loss of a readthrough protein does not require a reduction in mRNA abundance or ribosomal load. Summarizing these observations, the authors concluded that 3' UTRs encode a signal detected at the amino acid level by cells that marks the translated protein for degradation either during translation or post-translationally.

This study investigated whether readthrough into 3' UTRs also results in protein loss in human cells. Using K562 cells and a dual-color reporter with UTR sequences from 13 human genes, they found that in 9 cases, readthrough into the 3' UTR resulted in between a 3 to 30-fold reduction in the level of the readthrough protein.

The authors of this study hypothesized that targeted degradation of 3' UTR readthrough products might function as a buffering system in cells to mitigate toxicity induced by suppressor tRNA or ribosomal frameshifting. While they did not identify a specific mechanism of protein loss in readthrough products, they did importantly note a correlation between peptide hydrophobicity and degradation in both C. elegans and human cells.

1.4.2 Kramarski et al. 2020

Kramarski et al. first used tRNA suppression utilizing N-tert-butyloxycarbonyl lysine to suppress translation termination at amber stop codons (UAG) in HEK293T cells, finding significant differences in levels of protein products detected via mass spectrometry. Following this, they cloned the 3' UTRs of 13 human genes selected at random into a dual-color mCherry-T2A-eGFP reporter with the UTR sequence inserted at the 3' end of eGFP and upstream of the stop codon. By comparing the ratio of mCherry to GFP by either fluorescence or immunoblot intensity, they observed a range of effects on the level of reporter eGFP, with the weakest effect causing about a 10% decrease in the mCherry/eGFP ratio, while more than half of 3' UTRs caused more than a 10-fold decrease. By shuffling the amino acids in some of these UTRs and also adding fragments of coding sequence upstream to the UTR, and observing the effects on the mCherry/eGFP ratio, this study concluded that depletion of terminally extended proteins does not depend on the amino acid composition of the extension or any specific consensus sequence(s).

The authors of this study next aimed to identify the cellular mechanism leading to the loss of the readthrough proteins. Using MG-132 to suppress the proteasome in HEK293T cells transfected with readthrough reporters, they noted a decrease in the level of eGFP rather than an increase as would be expected if readthrough products were sent to the proteasome for degradation. Interestingly, it has been noted in the literature that inhibition of the proteasome may result in upregulation of the lysosomal machinery as a compensatory mechanism for maintaining proteostasis[31]. However, lysosomal inhibition with chloroquine did not result in an increase in the cellular levels of eGFP in the soluble cell fraction. In light of neither proteasomal nor lysosomal inhibition rescuing eGFP, the authors hypothesized that reporters with readthrough extensions were not degraded but rather lost from the soluble fraction of the cell by aggregation and accumulation in lysosomes. This hypothesis was supported by the results of immunoblotting on soluble and insoluble cell fractions, in addition to confocal microscopy showing the formation of punctae. They also found an association between the intrinsic disorder of a terminal extension (lack of hydrophobicity) and its depletion in soluble cell fractions due to an increased tendency to form aggregates.

1.4.3 Shibata et al. 2015

Shibata et al. investigated a readthrough mutation in the mouse cellular FLICE-like apoptosis inhibitory protein (cFLIP-L), resulting in a terminally extended protein containing an additional 46 amino acids. It was observed that the terminally extended cFLIP-L protein contains a degron in the terminal extension, which is targeted for degradation via the UPS and mediated by the E3 ligase Trim21. It was observed that proteasomal inhibition with MG-132 caused significant accumulation of the extended cFLIP-L protein when expressed in HeLa cells, and the extended mutant was highly ubiquitinated compared to the wild-type protein. Importantly, they found that mRNA abundance of the readthrough mutant was not significantly reduced in mice compared to the wild-type mRNA, suggesting loss of the protein post-translationally. Similar mechanisms of proteasomal degradation were found to be responsible for terminally extended protein variants of the human pyridoxamine 5-phosphate oxidase (PNPO) and 3β-hydroxysteroid dehydrogenase type II (HSD3B2) proteins.

1.4.4 Yordanova et al. 2018

Yordanova et al. first identified a significant peak in the 3’ UTR of the human adenosylmethionine decarboxylase 1 (AMD1) gene in ribosomal footprinting data 384 nucleotides downstream of the CDS stop codon. This peak suggested an unusual accumulation of ribosomes in the 3’ UTR of AMD1. The authors identified a high degree of evolutionary conservation of the 3’ UTR sequence up to the proposed ribosomal peak and a similar peak in ribosome profiling data from mouse, rat, frog, and fish samples. It was hypothesized that the ribosomal peak in the 3’ UTR of AMD1 resulted from ribosomes that stochastically read through the CDS stop codon and stall at a downstream ribosomal stalling sequence. While no extended readthrough product was detected with immunoblotting of an HA-tagged AMD1, mutating the native stop codon of AMD1 for 100% ribosomal readthrough led to a complete loss of protein. Further, a 65-fold decrease in reporter intensity was noted when attaching the AMD1 3’ UTR to a dual-luciferase system expressing both firefly and renilla luciferase from a single mRNA.

Neither proteasomal inhibition with MG132 nor lysosomal inhibition with concanamycin A resulted in the rescue of a GFP reporter with a terminal extension derived from the AMD1 3' UTR. A model was proposed in which ribosomes that translate through the AMD1 stop codon stall at a stalling sequence downstream within the 3' UTR. Successive instances of ribosomal readthrough result in the formation of a ribosome queue that will eventually reach into the coding sequence, inhibiting the translation of AMD1 from a given mRNA when a threshold of readthrough events has occurred.

1.5 Thesis Objectives

The primary objectives of this dissertation are to uncover potentially unifying principles that govern cellular mechanisms of surveillance and mitigation of noncoding translation and that dictate the fate of the protein products of noncoding translation within the cell. Chapter 2 will focus on a high-throughput screen that identified proteasomal degradation as the mechanism of loss of a wide range of translated noncoding sequences from various contexts. This screen further identifies C-terminal hydrophobicity as a key determinant of proteasomal degradation among these sequences. Chapter 3 will focus on a genome-wide knockout screen with a reporter of 3’ UTR translation that identified the BAG6 complex as the molecular detector of translated noncoding sequences that targets them for proteasomal degradation. In Chapter 4, I explore the potential evolutionary impacts of translation in noncoding sequences and propose a mechanism through which novel membrane-bound proteins may evolve. Finally, considering the significant and wide-ranging relationship between noncoding translation and human health, I aim to encourage future investigation into noncoding translation mitigation that may lead to therapeutic advances.

Chapter 2: A high-throughput reporter assay identifies proteasomal degradation and C-terminal hydrophobicity in noncoding translation mitigation

The work in this chapter is adapted in part from the following preprint:

Kesner, Jordan S. Chen, Ziheng. Aparicio, Alexis A. Wu, Xuebing. A unified model for the surveillance of translation in diverse noncoding sequences. Available at bioRxiv: https://www.biorxiv.org/content/10.1101/2022.07.20.500724v1

2.1 Introduction

There is currently no consensus on the fate of proteins resulting from translation in diverse classes of noncoding sequences encompassing 5' UTRs, 3' UTRs, introns, and lncRNAs. Although several studies have investigated these questions, conclusions drawn from the results of these studies are limited in that they are based only on translation of 3' UTR sequences from a small number of genes[28,31,33,34]. The alternative models proposed by these studies suggested that translation in 3' UTR sequences can result in targeting the translated peptide to the proteasome, aggregation and accumulation of the protein in lysosomes, or translational arrest induced by ribosome stalling and subsequent formation of a ribosome queue. In this chapter, I will discuss the results of a series of experiments that take advantage of a high-throughput reporter assay allowing us to study translation in thousands of unique noncoding sequences representing 5' UTRs, 3' UTRs, introns, and lncRNAs. The basis of the ability of this reporter assay to simulate translation in diverse noncoding sequences relies on a feature that is shared among all the noted types of noncoding sequences: translation results in a protein with a C-terminal region encoded by a noncoding sequence.

2.2 Results

Although translation in different types of noncoding sequences occurs via unique mechanisms, one shared feature is that translation of noncoding sequences will result in the C-terminal tail of the resulting protein being derived from a noncoding sequence (**Figure 2.1A**). We therefore reasoned that translation of a diverse set of noncoding sequences could be investigated through a reporter system that appended a given noncoding sequence to the C-terminal end of a reporter protein. For the purposes of our study, we chose to utilize a dual color reporter system to investigate noncoding translation (**Figure 2.1B**). This reporter contains mCherry and EGFP on the same transcript, separated by a self-cleaving T2A sequence, and to which a given sequence of interest is appended to the end of EGFP. A stop codon is present either at the C-terminus of EGFP (control reporter) or the end of the appended sequence (experimental reporter). Both mCherry and EGFP will be translated from the same mRNA in this reporter system, while the T2A sequence separates the proteins during synthesis allowing mCherry and EGFP to separate physically. This system has several advantages, including allowing for the investigation of EGFP (reporter) abundance in single cells while also allowing for the normalization of transfection efficiency and total expression levels in individual cells (mCherry). By calculating a ratio of EGFP/mCherry in individual cells in a population and then comparing the ratios between the control reporter and the experimental, one can accurately estimate the post-translational loss of a given reporter due to the addition of a C-terminal tail derived from a noncoding sequence.

We first set out to validate the proper functioning of this reporter system before adapting it for a high-throughput screen. We choose three reporters derived from three human genes, *HSP90B1*, *ACTB*, and *GAPDH*, which have previously been used to model 3’ UTR readthrough translation, translation of an intronic polyadenylation product, and an intronic retention product, respectively[31,95]. The noncoding sequence from these three genes was cloned into the dual-color reporter at the C-terminus of EGFP, and the ratio of EGFP/mCherry was calculated for each (**Figure 2.2A**). By comparing to control for each reporter, we found a 9.5-fold decrease in EGFP/mCherry for the *HSP90B1* reporter, an 18.1-fold decrease for the *ACTB* reporter, and a 4.2-fold decrease for the *GAPDH* reporter. These results were consistent with previously published findings and confirmed the proper functioning of the dual-color reporter system for downstream experiments (**Figure 2.2B**).

Given that previous studies have suggested that degradation of readthrough polypeptides occurs by either the proteasome[27,28] or the lysosome[33], we treated cells expressing the *ACTB* intron reporter with either the proteasome inhibitor lactacystin or the lysosome inhibitor chloroquine. While lysosome inhibition had a minimal effect, proteasome inhibition almost completely rescued the loss of EGFP caused by translation of the *ACTB* intron (1.4-fold loss of EGFP/mCherry ratio relative to control), suggesting the *ACTB* intron-coded peptide is primarily degraded by the proteasome (**Figure 2.2C**).

We next asked what could be learned about the peculiarities of noncoding translation by employing a high-throughput, systems-based assay to test thousands of translated noncoding sequences at once. To this end, we synthesized a library on the dual-color backbone containing 12,000 unique 90-nucleotide sequences appended to the EGFP tail (**Figure 2.3A**). Three thousand of these sequences are derived from coding sequences (2000 internal, 1000 terminal), which serve as a type of control in the library. The remaining 9000 sequences are derived from a variety of noncoding regions, including 5000 intronic sequences, 2000 3’ UTR sequences, 1000 5’ UTR sequences, and 1000 lncRNA sequences. We similarly synthesized a separate library containing approximately 500,000 random 39 nucleotide sequences appended to the tail of EGFP in the dual-color reporter (**Figure 2.3A**).

To determine if many of these sequences would result in loss of EGFP upon translation, we set up a high-throughput screen with flow cytometry as a readout. Both libraries were transduced into HEK293T cells at a low MOI such that any individual cell had only one reporter. In each case, a significant decrease in EGFP signal is seen while mCherry remains relatively stable in a high percentage of the cellular population when assayed via flow cytometry (**Figure 2.3B and 2.3C**). These results suggest that EGFP loss due to the addition of a tail of up to 30 amino acids derived from either random or noncoding sequences is likely to result in protein loss for many sequences without a significant change in mRNA abundance.

As described above, there is a lack of consensus on what endpoint mechanism is responsible for the eventual degradation or loss of proteins containing translated noncoding sequences. We therefore tried to determine if proteasomal or lysosomal degradation was responsible for the significant loss of EGFP observed in our high-throughput screen. To this end, we performed proteasomal and lysosomal inhibition with several different drugs in a population of cells expressing our Pep30 noncoding library. Inhibition of the lysosome had little to no effect on rescuing the EGFP/mCherry ratio in this population. In contrast, proteasomal inhibition resulted in a significant rescue of the lost EGFP (**Figure 2.3D**). These results suggest that for a large percentage of the noncoding translation reporters, the endpoint of protein loss is at the proteasome, and very few are degraded in the lysosome.

Having identified that many of these reporters undergo protein loss in the proteasome, we next set out to determine if we could identify characteristics of the amino acid sequence in the individual reporters that could be correlated to their loss. We again transduced the Pep30 library into HEK293T cells at a low MOI such that any given cell expressed only a single reporter and used FACS to separate the cells into an EGFP high and EGFP low population (**Figure 2.4A**). The 90-nucleotide tails were isolated and sequenced from these two populations. We then compared the abundance of each tail across the EGFP high and EGFP low populations to assign a degradation score to each reporter based on how likely it was to end up in the EGFP high or EGFP low bin. Tails that are highly degraded by the cell will show a higher abundance in the EGFP low bin population and vice versa. We assign a degradation score for each reporter, which is calculated as the log2 ratio of the abundance of the tail in the EGFP low bin over the EGFP high bin (**Figure 2.4A**). A higher degradation score thus indicates that a given reporter is more likely to be degraded in the cells, while a lower degradation score indicates that the reporter is more stable.

Several interesting results were found from this analysis. We first compared the distribution of degradation scores based on the length of the tails in the reporter library (**Figure 2.4B**). Within the library, approximately 4000 of the unique reporters will result in translation of the full 30 amino acids encoded by the 90-nucleotide sequence. The remaining sequences all contain in-frame stop codons at some position within the sequence. When looking at the relationship between the length of a given reporter tail and its level of degradation, it appears that there is a minimum length before which the appended amino acid chain can cause degradation of the reporter. Reporters that terminate before ~13 amino acids when added to the EGFP reporter generally have very low levels of degradation, indicating that peptides of this short length do not significantly change the behavior of the protein as a whole, at least in terms of targeted degradation by the cell. Additionally, beyond this minimum tail length of about 13 amino acids, there is no strong relationship between tail length and degradation. In other words, a tail of more than 13 amino acids appears just as likely to result in protein degradation, whether it is 13 amino acids long or the full 30 amino acids.

An interesting pattern also emerged in relation to the class of sequence attached to the reporter (i.e., intron, 3’ UTR, 5’ UTR, CDS, etc.). Looking at each group of reporters individually, we found that terminal coding sequences had by far the lowest degradation scores, indicating that these sequences are unusually stable (**Figure 2.4C**). This is in line with the idea that the C-terminus of functional proteins is under selection to avoid sequences that will be targeted for degradation[96]. Proteins not under this type of selection are likely to be targeted for degradation by their C-terminal ends, and many C-terminal degrons that mark proteins for degradation are well documented[96]. Alternatively, proteins which are required to be short lived may select for such degrons in their C-terminal tails[96]. Among sequences with positive degradation scores, sequences derived from introns had the highest average degradation, while those derived from 5’ UTRs had the lowest average degradation. Sequences derived from internal CDS, 3’ UTRs, and lncRNAs had similar levels of degradation. Interestingly, frameshifted CDS-derived sequences had the second highest levels of degradation among the different categories.

To further understand the characteristics of noncoding sequences that trigger degradation, we next looked at the full-length reporters (4726 reporters with a tail length of 30 amino acids, no in-frame stop codon) and calculated a correlation between degradation and a spectrum of predicted properties of the amino acid tail (**Figure 2.5A**). The properties examined include average hydrophobicity, hydrophobic moment, transmembrane potential, molecular weight, net charge, mass to charge ratio, c degron prediction, instability, intrinsic disorder, coil, alpha helix, beta sheet, interaction potential, and counts of aliphatic, aromatic, basic, charged, small, acidic, and polar amino acids. Among physicochemical properties, the average hydrophobicity of the tail had the strongest correlation with degradation, showing a spearman correlation of approximately 0.67. Transmembrane potential also showed a moderate correlation of approximately 0.3. Prediction of C-end degrons only showed a weak correlation of about 0.2, indicating that degrons are not primarily responsible for the loss of the reporters. Interestingly, intrinsic disorder showed a strong negative correlation of about -0.65 with degradation, likely explained by its anti-correlation to hydrophobicity. A plot of all the full-length reporters showing degradation on the y-axis and average hydrophobicity of the tail on the x-axis illustrates the strong correlation between hydrophobicity and degradation on a reporter-by-reporter basis (**Figure 2.5B**).

To uncover the exact nature of the degradation signal, we examined the amino acid composition and various physicochemical and structural properties of the tail peptides. Using the kpLogo tool for position-specific sequence analysis[97], we performed a Student's t-test for every amino acid at every position in the 30-aa tail to test if the presence of a given amino acid at a particular position is associated with stronger degradation (**Figure 2.5C**). Strikingly, we found that almost all hydrophobic residues are associated with increased degradation at most positions in the 30-aa tail. The only exception is alanine (A), the least hydrophobic of the nine hydrophobic residues, and is only associated with degradation at the last two positions, consistent with its function as a C-terminal end degron that is recognized by Cullin-RING E3 ubiquitin ligases[96,98]. We also confirmed two other C-degrons, arginine (R) at the third to last position and glycine (G) at the last position. However, a 30-variable regression model using A/G/R residues in the last ten positions is only weakly predictive of degradation (Spearman correlation coefficient, Rs = 0.22).

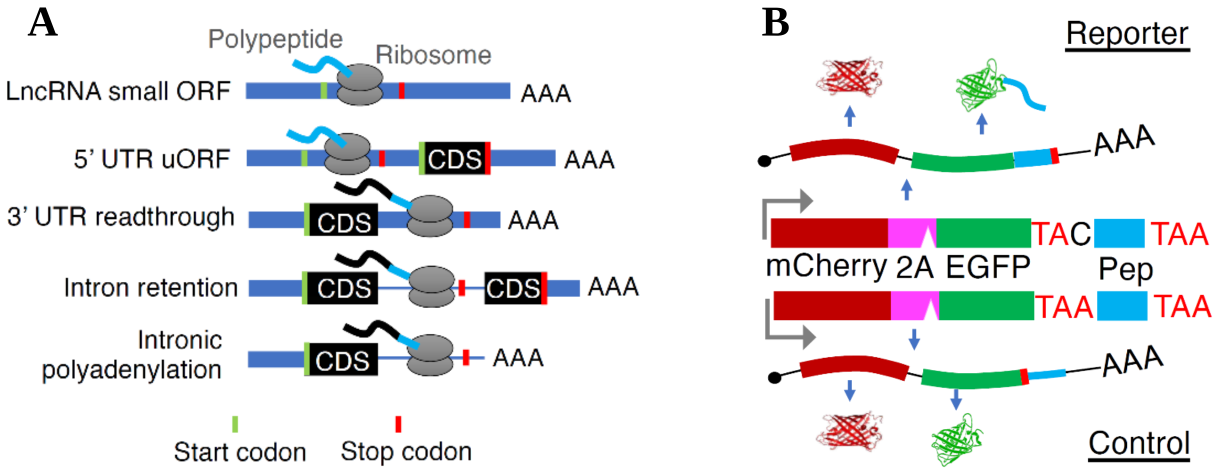
2.3 Discussion

The noncoding sequences that can be translated are heterogeneous at three levels: they are located differently relative to annotated coding regions (i.e., lncRNAs, 5’ UTRs, 3’ UTRs, and introns of mRNAs); they are translated when different quality control mechanisms fail (e.g., mis-splicing, mis-polyadenylation, and stop codon readthrough), and they are very diverse in terms of their primary nucleotide sequence and therefore codon usage and RNA structures. Although several studies have investigated noncoding translation in 3’ UTRs, it remains unclear whether a common mechanism is used for the surveillance of unintended translation in such heterogeneous sequences.

The results presented in this chapter highlight and take advantage of a positional feature shared in noncoding translation regardless of the origin of the underlying sequence: the C-terminal region of peptides resulting from noncoding translation is encoded by noncoding sequences. This unifying feature was utilized in our high-throughput assay to probe features of noncoding translation surveillance that are common to translation in all the noted classes of noncoding sequence. In doing so, we identified several key shared aspects of translation in diverse noncoding sequences; translation of the noncoding sequence often leads to protein loss, noncoding protein loss is achieved through proteasomal degradation, and loss of the noncoding protein is significantly correlated to C-terminal hydrophobicity.

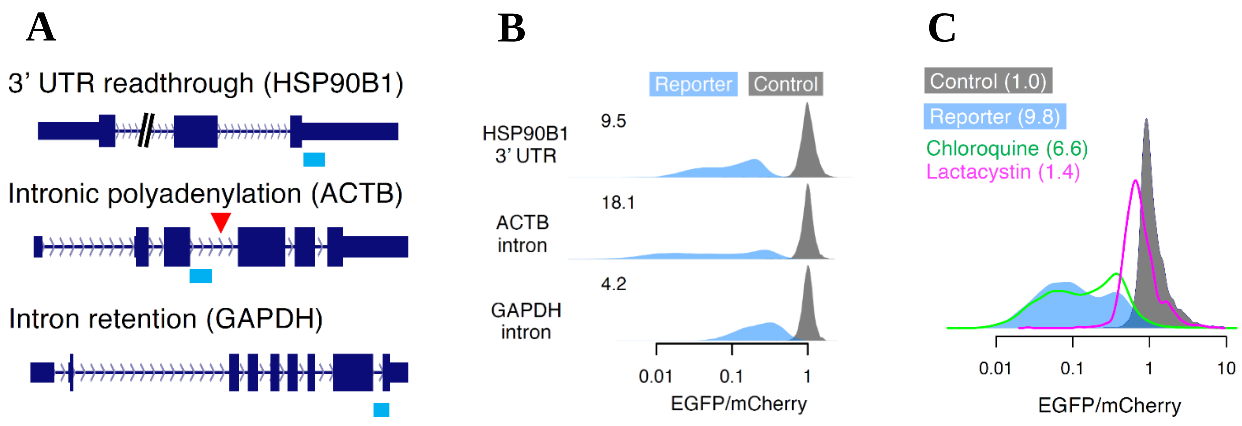
2.4 Figures

Figure 2.1: A reporter system for investigating diverse classes of noncoding translation



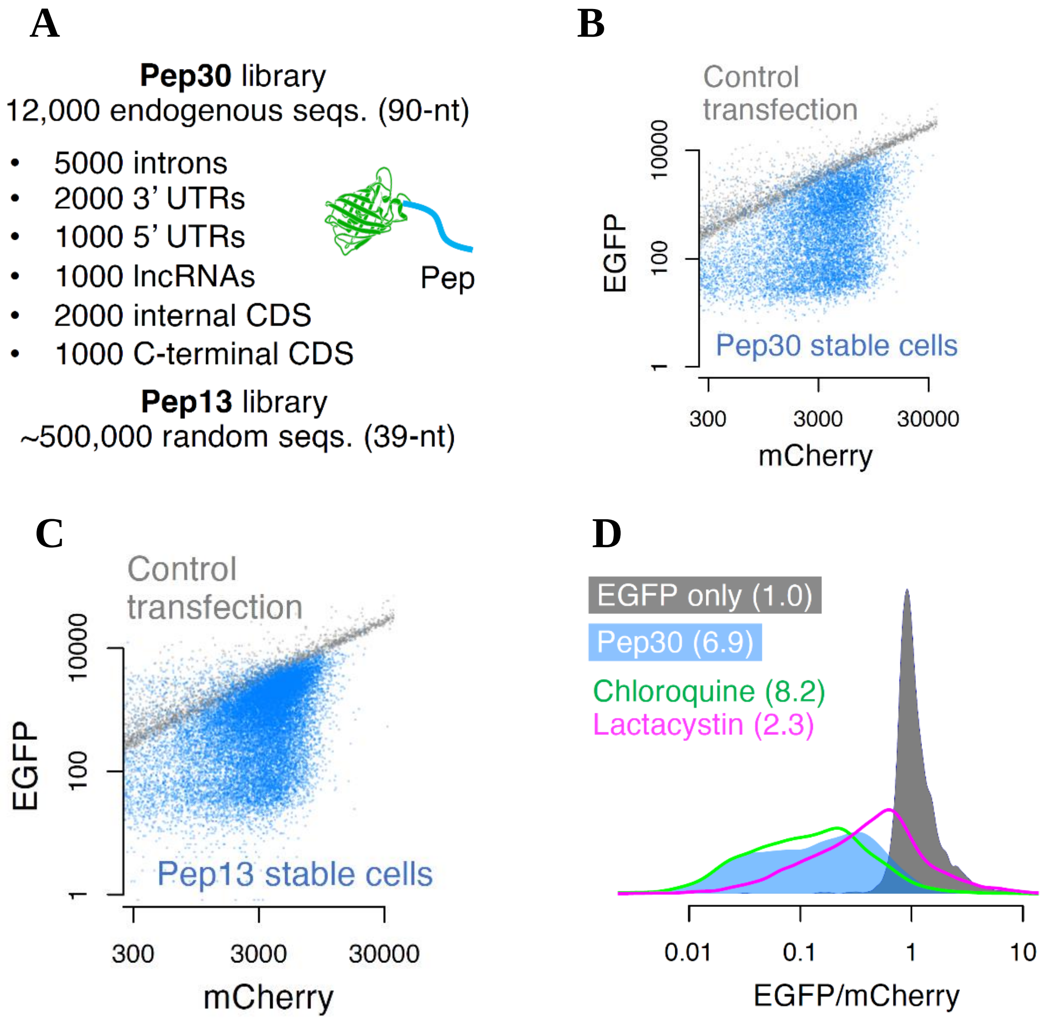
(A) Schematic showing the mechanism of translation in various types of noncoding regions. All forms of noncoding translation shown here result in the C-terminal region of the synthesized peptide being derived from noncoding sequences. (B) Diagram of the dual-color reporter used to study noncoding translation. 2A is a self-cleaving peptide sequence that uncouples the mCherry and EGFP fluorescent reporters during translation. Pep represents the C-terminal addition to the EGFP reporter comprised of noncoding sequence. The control and reporter versions differ only in the placement of the stop codon, which is at the C-terminal end of EGFP in the control and at the C-terminal end of the noncoding sequence in the reporter.

Figure 2.2: Validating the function of the dual-color reporter using previously studied translated noncoding sequences



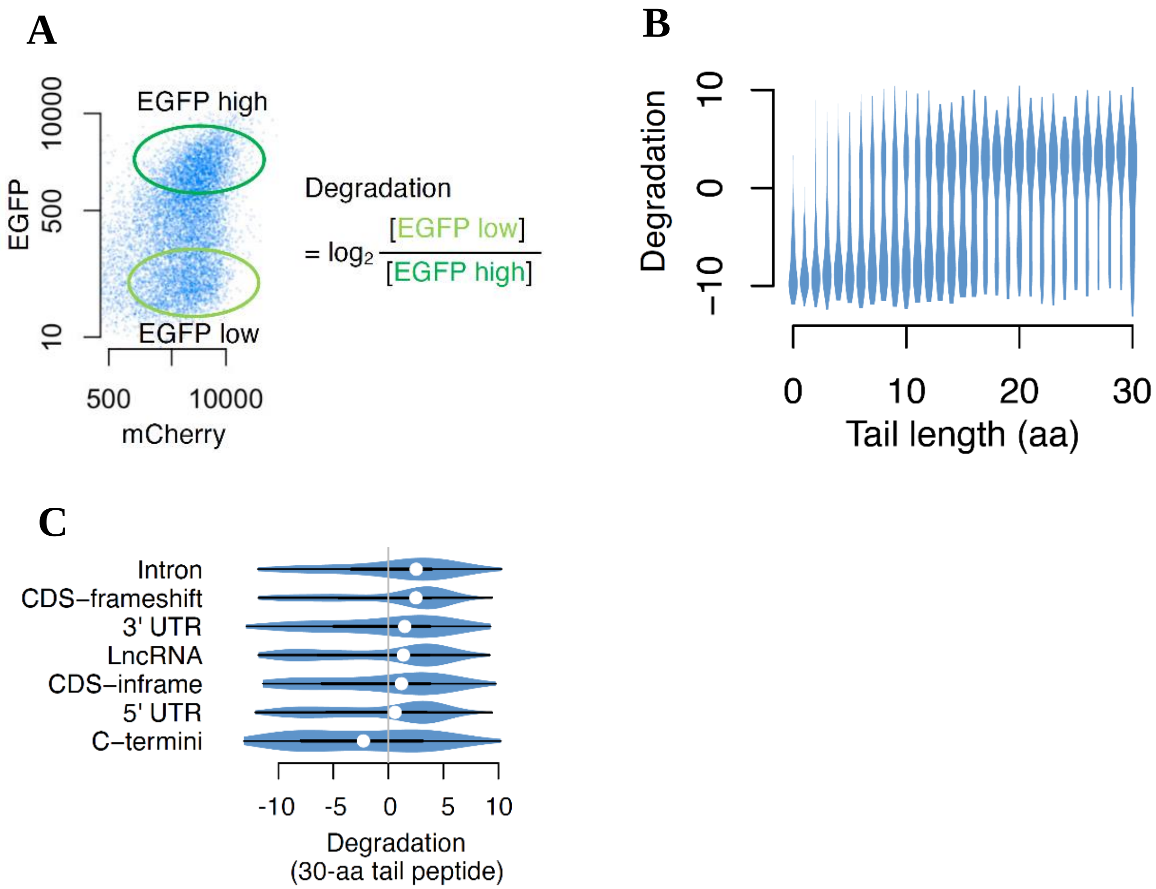
(A) Diagram of three genes known to undergo different types of noncoding translation. Translated noncoding sequences originating from 3' UTR readthrough in HSP90B1, intronic polyadenylation in ACTB, and intron retention in GAPDG were cloned into the dual color reporter. (B) Reporter and control variants of the dual-color reporter containing noncoding sequences from the three genes in (A) were transfected into HEK293T cells and analyzed with flow cytometry. By calculating the EGFP/mCherry for each reporter, we can observe a significant loss in EGFP levels when each noncoding sequence is translated, ranging from a 4.2 to 18.1-fold difference compared to control. The numbers indicate the median fold loss of EGFP/mCherry relative to control. (C) EGFP/mCherry ratio for cells transfected with either the control or the ACTB intron reporter, alone or with simultaneous treatment of either proteasome inhibitor (lactacystin) or lysosome inhibitor (chloroquine). The numbers indicate the median fold loss of EGFP/mCherry relative to control.

Figure 2.3: A high-throughput screen reveals drastic loss of protein due to translation of diverse noncoding sequences



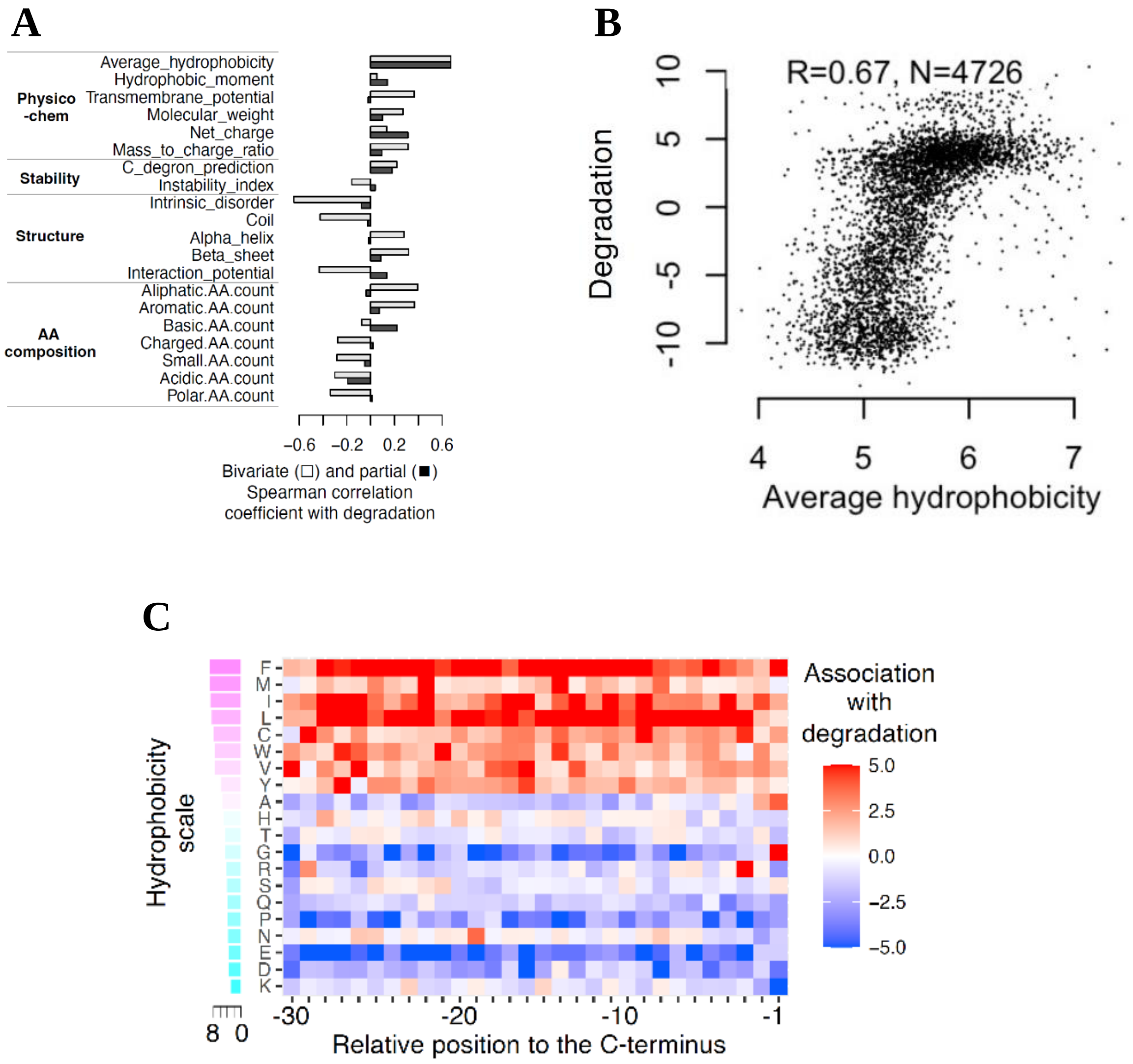
(A) The dual-color reporter was used to generate two reporter libraries of noncoding translation. The Pep30 library comprises 12,000 unique sequences, including randomly selected sequences from 9000 noncoding and 3000 coding regions in the human transcriptome (30 aa). The Pep13 library comprises approximately 500,000 randomly generated sequences (13 aa). (B) The Pep30 library or a control reporter was transfected into HEK293T cells and analyzed with flow cytometry. (C) The Pep13 library or a control reporter was transfected into HEK293T cells and analyzed with flow cytometry. (D) The Pep30 library or a control reporter was transfected into HEK293T cells and treated with either a proteasome inhibitor (Lactacystin) or a lysosome inhibitor (Chloroquine).

Figure 2.4: Analysis of sequence enrichment in high and low EGFP cell populations carrying the Pep30 library



(A) Pep30 stable cells were sorted into high and low EGFP bins and the tail sequences (DNA) were cloned and sequenced. The degradation score for each sequence is calculated as the log2 ratio of read counts in EGFP-low vs. EGFP-high bin. (B) Violin plots of degradation score for tails of varying lengths. (C) Violin plots comparing degradation of 30-aa tails encoded by various types of sequences.

Figure 2.5: Analysis of sequence properties associated with degradation from the Pep30 library



(A) Spearman correlation coefficient (light bar) between various properties of the tail peptides and degradation. Dark bar: partial correlation conditioned on average hydrophobicity. (B) A hydrophobicity-vs-degradation scatter plot for tails of 30-aa length. (C) A heatmap visualizing the association (Student's t-test statistics capped at 5.0) between degradation and the presence of each amino acid at every position in the Pep30 library. Amino acids (rows) are sorted by hydrophobicity (Miyazawa scale)

2.5 Methods

Plasmids

HSP90B1, *ACTB*, *GAPDH*, and *SMAD4* reporters: the 3' UTR of *HSP90B1*, intron 3 of *ACTB*, the last intron of *GAPDH*, and the 3' UTR of *SMAD4* were PCR-amplified from the genomic DNA of HEK293T cells with primers listed in Table S3. The PCR products were then either digested with NotI and SbfI (*GAPDH* and *SMAD4*) or NsiI-HF/PspOMI (*ACTB* and *HSP90B1*), which generate the same overhangs. The inserts were then ligated with NotI/SbfI-digested pJA291 (Addgene #74487) (Arribere et al., 2016)

AMD1 reporters

The AMD1 readthrough reporter (Fig. 4A) was generated by inserting genomic DNA-amplified fragment into pJA291 using NotI/SbfI sites. Overlap extension PCR (OEP) cloning was used to insert a P2A sequence between EGFP and the translated AMD1 3' UTR in the readthrough reporter (Fig. 4B). Systematic deletion of individual or combinations of hydrophobic regions from the readthrough reporter were done using NEB Q5 Site-Directed Mutagenesis (SDM) Kit (#E0554) (Fig. 4C and Fig. S4). The AMD1 roadblock reporter (Fig. 4F) was generated using OEP cloning. OEP cloning was again used to delete the putative ribosome pausing signal from the roadblock reporter (Fig. 4G), or replace the AMD1 sequence with a poly(A) sequence (Fig. 4E). Deletion of the ribosome stalling signal from the readthrough reporter was also generated by OEP cloning (Fig. 4D). All primers used were listed in Table S3. All plasmids were transformed into NEB Stable Competent E. coli (C3040) according to the manufacturer's protocol. Positive clones were confirmed via sanger sequencing.

Cell culture

HEK293T cells used in this study were purchased from ATCC. Cells were cultured in DMEM with 4.5 g/L D-Glucose supplemented with 10% fetal bovine serum, 1% penicillin/streptomycin was added except when producing lentivirus. Low passage number cells were used and maintained under 90% visual confluency. Cells were maintained at 5% CO2 and 37 °C. HEK293T cells used in this study were confirmed to be negative for Mycoplasma Contamination and routinely tested using the MycoAlertTM Mycoplasma Detection Kit (Lonza, LT07-418). For experiments involving the *SMAD4* gene, clonal cell lines harboring *SMAD4* readthrough mutations as well as the parental HEK293T cells were obtained as a generous gift from Dr. Sven Diederichs. Transfection of plasmids was done using Lipofectamine 2000 or Lipofectamine 3000 according to the manufacturer's instructions. Flow cytometry analyses of transfected cells were typically performed 24 or 48 hours after.

Lentivirus and stable cell line generation

For generating lentivirus, 750,000 HEK293T cells were seeded in 6-well plates with DMEM supplemented with 10% FBS. After 24 hours, the cells were transfected with the second-generation lentiviral packaging plasmids and the lentiviral plasmid of interest using Lipofectamine 3000. The virus-containing media was collected 48 and 72 hours after transfection, combined, clarified by centrifugation at 500 RCF for 5 minutes, and then passed through a 45 µM PVDF filter. The purified virus was stored at 4°C for short-term use or aliquoted and frozen at -80°C. For the generation of stable cell lines, HEK293T cells were reverse transduced in 6-well plates in media with 10 µg/mL polybrene using purified virus such that <30% of the cells are transduced. Twenty-four hours after transduction, the virus-containing media is removed, and fresh media is added. After another 24 hours, the cells are collected, and transduction efficiency is confirmed via flow cytometry. Transduced cells are then selected with puromycin at 2 μg/mL for 48 hours or via flow cytometry to generate a stable cell line for downstream analysis.

Flow cytometry analysis

Cells were collected and resuspended in 1-4 mL of fresh media and passed through a 35 μM mesh cell strainer immediately prior to flow cytometry. Flow cytometry was performed on either a Bio-Rad ZE5 or NovoCyte Quanteon analyzer. Gating of samples and export of data for downstream analysis was done using the FCS Express software.

Massively parallel reporter assays in HEK293T cells

For the Pep30 library, a pool of 12,000 oligos were synthesized by Twist Bioscience, each containing a 90-nt variable sequence flanked by a 15-nt constant sequence on each side. The left constant sequence TACTGCGGCCGCTAC carries a NotI site, whereas the right constant sequence TGACTAGCTGACCTG contains stop codons in all three reading frames, followed by a SbfI site (extended into the vector backbone) for cloning. The variable sequences were picked from a set of randomly selected lncRNAs (Hezroni et al., 2015), as well as the following regions in coding mRNAs (RefSeq): the 5' end of coding exons, introns, 3' UTRs, 5' UTR ORFs, and the 3' end of the last coding exon. Regions annotated to multiple classes or overlapping with each other on either strand were discarded. For introns and 3’ UTRs, the first 90 nt was used. For lncRNAs and 5’ UTRs, the first AUG was identified, and the next 90 nt were used. For the C-termini of CDS, the last 90nt of the ORF (excluding the stop codon) were used. For internal CDS, the first 90 nt were used, with about one-third being in-frame with the EGFP ORF. The oligo pool was PCR-amplified and then cloned into pJA291 using the NotI/SbfI sites and primers listed in Table S3. The Pep13 library was cloned into pJA291 using NEB Q5 Site-Directed Mutagenesis Kit (#E0554). The Pep30 and Pep13 libraries were then used to generate stable cell libraries using lentiviral transduction such that each cell was integrated with at most one virus. Cells were then sorted into EGFP-high (top 20%) or EGFP-low (bottom 20%) bins, and the variable regions of the reporter were then cloned and sequenced.

Correlation between mitigation and physiochemical and structural properties of tail peptides

Secondary structures of each peptide were predicted using S4PRED (Moffat and Jones, 2021), which outputs a vector indicating whether each residue is in an α-helix, β-sheet, or coil. The number of residues in each of the secondary structure motifs in a peptide is used to calculate the correlation with mitigation. Protein intrinsic disorder was calculated using the program IUPred3, specially for short disorder analysis without smoothing. The disorder score for each residue in a peptide is added together, and the total disorder score is used to calculate the correlation with mitigation. All other properties were calculated using the following functions in the R package Peptides (Osorio et al., 2015): Average_hydrophobicity: hydrophobicity using the Miyazawa scale (Miyazawa and Jernigan, 1985) unless otherwise noted. Hydrophobic_moment: hmoment , Amino acid composition(*.AA.count): aacomp, Mass-to-charge ratio: mz, Molecular_weight: mw, Net charge: charge, Interaction_potential: boman, Instability_index: instaIndex, and Transmembrane_potential: membpos.

Chapter 3: Systems-based approaches identify the BAG6 membrane protein triage complex in widespread surveillance of noncoding peptides

The work in this chapter is adapted in part from the following preprint:

Kesner, Jordan S. Chen, Ziheng. Aparicio, Alexis A. Wu, Xuebing. A unified model for the surveillance of translation in diverse noncoding sequences. Available at bioRxiv: https://www.biorxiv.org/content/10.1101/2022.07.20.500724v1

3.1 Introduction

Following the findings outlined in Chapter 2, we next sought to apply further systematic approaches to identify the molecular pathways that actuate the targeting of noncoding proteins for proteasomal degradation. The gene AMD1 experiences a high level of stochastic readthrough compared to most genes in the human genome, occurring at a rate of approximately 1.6%[34]. A previous study investigating the outcome of this readthrough had found that a ribosome queue will form due to a stalling sequence in the downstream 3' UTR, which will eventually result in translational arrest due to a physical blockage of the transcript from the queued ribosomes[34]. This study concluded that this is a conserved mechanism that limits the amount of AMD1 proteins that can be translated from a given AMD1 transcript and that loss of AMD1 production is due to this translational arrest. The basis of this study was the observation of a signal peak in the purported ribosome stalling sequence in the AMD1 3' UTR in ribosome profiling data combined with conservation of the 3' UTR sequence. However, no ribosome profiling signal is seen in the

3’ UTR tail itself, as would be expected of regions that are actively translated comparably to a coding sequence.

For these purposes, we chose to focus on AMD1 3’ UTR translation as a reporter of noncoding translation. Several factors make AMD1 an attractive choice as a model of noncoding translation, including its high rate of stochastic endogenous readthrough, as well as uncertainty in the literature as to the consequences of translating the AMD1 3’ UTR[29,34].

3.2 Results

To test the idea that a ribosome queue forms in the tail of AMD1, we cloned the full-length AMD1 3’ UTR into our dual color reporter (**Figure 3.2A**). As expected, when transfected into HEK293T cells, we found that the control reporter had a 19.4-fold higher ratio of EGFP/mCherry than the readthrough reporter, confirming that translation of the AMD1 3’ UTR was indeed leading to loss of EGFP (**Figure 3.1A**). Interestingly, however, treatment of this reporter with the proteasome inhibitor MG-132 resulted in a strong rescue, lowering the fold difference in the ratio to 1.9 (**Figure 3.1A**). This result strongly suggests that the translation of the AMD1 tail results in protein loss through a proteasome-dependent mechanism and not a stalling mechanism, as suggested previously.

We performed a series of follow-up experiments to confirm that ribosome stalling was not the likely mechanism of protein loss in the case of AMD1 3’ UTR translation. We first inserted a 2A sequence in the AMD1 reporter between EGFP and the AMD1 tail, uncoupling these two sequences at the protein level (**Figure 3.1B**). In the case that a ribosome queue was forming, translation of this reporter should still result in a significant loss in EGFP similar to the parental reporter, as the ribosome queue will not be lost due to the presence of the 2A sequence on the mRNA. Alternatively, if it is the translated peptide sequence itself that causes protein loss, this reporter should show a large rescue as the AMD1 tail peptide will be physically separated from the EGFP protein during translation. Indeed, we find that a large rescue is achieved, even greater than in the case of proteasome inhibition, when the AMD1 tail and EGFP are separated by the 2A self-cleaving sequence, with the ratio now at only 1.2-fold higher EGFP/mCherry in the control compared to experimental reporter (**Figure 3.1B**).

We also noticed five regions of high localized hydrophobicity in the translated AMD1 tail (**Figure 3.2A**). In accordance with our previous findings in Chapter 2 that C-terminal hydrophobicity plays a role in the loss of translated noncoding sequences, we next generated a series of deletion mutants in these hydrophobic regions. While we found that deletion of any individual hydrophobic region did not result in a large rescue (**Figure 3.2B**), deletion of the last three hydrophobic regions did result in a rescue, reducing the ratio difference to 2.1 (**Figure 3.1C**). These results further suggest that similarly to the reporters in our original high-throughput screen, C-terminal hydrophobicity in the translated AMD1 tail plays a role in its loss.

Several other experiments with modified AMD1 reporters strongly argue against the model of loss due to the formation of a ribosome queue. Deletion of the purported stalling signal downstream in the AMD1 3’ UTR did not result in any rescue of the EGFP/mCherry ratio and, in fact, resulted in a larger difference of 21.3 (**Figure 3.1D**). Insertion of a poly(A) roadblock sequence between two 2A sequences in between mCherry and EGFP resulted in a massive ratio difference of 136.3, rescued by MG-132 treatment to 50.3 (**Figure 3.1E**). However, when the poly(A) sequence is replaced with the full-length AMD1 tail sequence, this ratio is found to be 2.3, suggesting that the AMD1 sequence does not induce ribosome stalling in a similar mechanism to a poly(A) roadblock sequence (**Figure 3.1F**). Deletion of the purported ribosome pausing sequence in the AMD1 tail in the previous reporter results only in a minimal change to the ratio, from 2.3 to 2 (**Figure 3.1G**), suggesting there is little to no effect on translation due to this signal, which was proposed to be critical for the formation of a ribosome queue in the previously proposed model.

To investigate the potential cellular mechanisms by which the translated *AMD1* tail is recognized and targeted by the cell for degradation, we performed a genome-wide CRISPR-Cas9 based knockout screen[99]. We first generated a stable cell line from HEK293T cells which carried the dual-color reporter with the full-length *AMD1* 3' UTR at a single copy per cell by lentiviral transduction at a low MOI (**Figure 3.3A)**. This population of cells was then transduced with a second lentiviral library expressing Cas9 and a library of guides with ten unique guides targeting each gene in the human genome, also at a low MOI such that each cell received only one unique guide[99]. The population of cells carrying the reporter and Cas9 plus guides was then sorted with flow cytometry and separated into a low EGFP and high EGFP population while normalizing to the levels of mCherry. Cells in the high EGFP population have a ratio of mCherry to EGFP closer to 1, indicating rescue of the EGFP-AMD1 tail protein loss, and cells in the low EGFP population have a low mCherry/EGFP ratio, indicating high levels of loss of the EGFP-AMD1 reporter. Genomic DNA was then isolated from these two populations, and the guide sequences were determined by sequencing. Enrichment of the guides across the two populations indicative of rescue or further loss of the reporter was then determined using the MAGeCK package[100]. Several interesting results were found from the guide enrichments (**Figure 3.3B**).

Of the top 20 guide hits (FDR <0.01, enriched in high EGFP), 17 are distinct proteasomal components. These results further support the model of proteasomal degradation for the EGFP-AMD1 reporter. Interestingly, the remaining 3 top hits with FDR < 0.01, *BAG6*(*BAT3*), *TRC35*(*GET4*), and *RNF126*, are all key components of the highly conserved BAG6 pathway for membrane protein triage in the cytosol (**Figure 3.3C**). The BAG6 pathway is embedded as a quality control module in the Transmembrane domain Recognition Complex (TRC) pathway, also called Guided Entry of Tail-anchored proteins (GET) pathway, for the triage of tail-anchored (TA) proteins. Similar to noncoding translation products, TA proteins have a hydrophobic C-terminal tail that functions as a transmembrane domain (TMD) while also serving as the membrane targeting signal. Unlike most membrane proteins with an N-terminal signal peptide mediating co-translational targeting to membranes, TA proteins can only be targeted post-translationally, after the C-terminal targeting signal has emerged from the ribosome exit tunnel. Immediately after being released from the ribosome, TA proteins are captured by the ribosome-associated co-chaperone SGTA, which binds and shields the hydrophobic TMD in nascent TA proteins[101–106]. SGTA then delivers the substrate to the BAG6-UGL4A-TRC35 heterotrimeric complex by binding to UBL4A[107,108]. Authentic TA proteins will be transferred directly from SGTA to TRC40, which is associated with the trimeric complex via TRC35, and are committed to membrane targeting. Defective TA proteins, however, will be released from SGTA and re-captured by BAG6, which recruits the E3 ubiquitin ligase RNF126 that catalyzes the ubiquitination of the substrate, committing it to proteasomal degradation[109,110]. The BAG6 pathway also mediates the degradation of misfolded ER proteins extracted to the cytosol by p97/VCP in the ER-associated degradation (ERAD) pathway[108].

To validate the results of the screen, we firsts set out to generate a series of knockout cell lines in the top genes of interest enriched in the high EGFP population in HEK293T cells. We decided to include in our validation knockout clones of *SGTA* and *UBL4A*, as we suspected they may play a role in this process and could potentially be false negatives in the screen due to poor guide performance. We chose the top 2 performing guide sequences for each of these genes from the CRISPR screen and cloned the guide sequences into a Cas9 and guide expressing lentiviral vector, which was then transfected into HEK293T cells. This population of cells was selected with puromycin, individual clones were sorted into 96-well plates, and those clones were then allowed to grow out to obtain a suitable population number for genomic DNA extraction. Individual clones were initially screened by Sanger sequencing and the ICE CRISPR analysis tool was used to detect clones with compound heterozygous frameshift mutations, which should yield a complete knockout of the target protein. After identifying several clones that appeared to be complete knockouts by sanger sequencing, western blots were performed with wild-type HEK293T cells as a control against each of the five proteins of interest to validate the full loss of the target protein in each of our selected clones (**Figure 3.4A**).

With our knockout clones validated, we next set out to validate the results of the screen both in terms of our positive hits (*BAG6, TRC35, RNF126*) and our suspected false-negative hits (*SGTA* and *UBL4A*). To do this, we transfected the control *AMD1* and readthrough *AMD1* reporter into each of our five knockout clones and the parental wild-type HEK293T control. We then calculated the fold difference in the EGFP/mCherry ratio via flow cytometry (**Figure 3.4B**). As expected, in the wild-type control, there was a 16.2-fold difference in the ratio of EGFP/mCherry, indicating a significant loss of the EGFP-AMD1 protein due to translation of the *AMD1* tail. Among our top three positive screen hits, all resulted in a strong rescue of EGFP/mCherry, with TRC35 at 4.3, BAG6 at 3.5, and RNF126 at 2.9. It is interesting that knockout of these proteins does not result in a complete rescue, which is likely indicative that there is a compensatory pathway activated in the absence of BAG6 functionality or that there is cooperation in targeting the *AMD1* tail from another quality control pathway even under normal conditions. However, as proteasome inhibition does result in a nearly complete rescue (19.4 to 1.9), we do not believe that lack of complete rescue in the knockout of the BAG6 pathway indicates there is a non-proteasomal based mechanism making a significant contribution to the loss of the EGFP-*AMD1* reporter. Interestingly, we found a moderate rescue in both the *SGTA* and *UBL4A* knockout clones, at ratios of 5.1 and 7,8, respectively. In light of these results, these genes were likely false negatives from the CRISPR screen due either to their relatively weaker effect than knockout of the other three BAG6 complex genes or due simply to the peculiarities of the individual guides used in the screen to target these two genes. It is also plausible that protein half-life may have played a role, as the screen was performed over a shorter time span than the outgrowth and validation of the clonal knockout lines.

In the TRC/GET pathway, BAG6 captures substrates by directly binding to their C-terminal hydrophobic transmembrane domains. We therefore performed a co-immunoprecipitation involving a pulldown using an antibody specific to EGFP and subsequently performed a western blot with an antibody specific to BAG6 (**Figure 3.5A and 3.5B**). In this experiment, we included both the readthrough reporter with the full-length AMD1 tail appended to EGFP, as well as the deletion mutant, which removes the final three identified hydrophobic regions within the tail that was previously shown to result in a strong rescue of the reporter. Both reporters were transfected into wild-type HEK293T cells prior to the pulldown. As expected, BAG6 is strongly detected in the population expressing the full-length AMD1 tail and is barely detectable with the hydrophobic region deletion mutant (**Figure 3.5B**). These results demonstrate physical interaction between BAG6 and the *AMD1* tail and further demonstrate that this interaction largely depends on the presence of the highly hydrophobic regions within the translated AMD1 tail.

Although demonstrating that BAG6 and its associated proteins play a key role in detecting and targeting the translated AMD1 tail for proteasomal degradation, the results of our experiments discussed thus far do not reveal the scope of this surveillance activity by the BAG6 complex. We therefore next set out to determine if the BAG6 complex was responsible for the targeting of translated noncoding sequences on a larger scale. We first focused our attention on the well-known tumor suppressor gene *SMAD4*. A recent study identified a series of recurring mutations in the native *SMAD4* stop codon by analyzing patient data derived from the COSMIC database[27]. These nonstop mutations result in translational readthrough of the SMAD4 stop codon into the 3' UTR, extending the SMAD4 protein by 40 amino acids[27]. The study demonstrated that this extended form of SMAD4 almost completely degraded in the cell, that this degradation is proteasome-dependent, and that a ten amino acid long hydrophobic degron is necessary for the degradation of the SMAD4 mutant protein[27]. The *SMAD4* readthrough mutant thus shares the three critical characteristics identified in our studies of the AMD1 readthrough mutant for targeting by BAG6 for degradation. As such, we first generated a reporter construct using the same dual-color parental vector that expressed the *SMAD4* 3' UTR tail appended to EGFP (**Figure 3.6B**). When transfected into wild-type HEK293T cells, there is a 20.5-fold difference in the EGFP/mCherry ratio between the stop control and readthrough reporter, indicating a significant loss of the reporter due to translation of the SMAD4 3' UTR, as expected. When transfected into the BAG6 knockout HEK293T clonal cell line, this ratio drops to 7.9, indicating significant rescue of the *SMAD4* readthrough reporter, though somewhat weaker than the rescue seen with the AMD1 readthrough reporter. To determine if we could detect an effect on SMAD4 readthrough protein endogenously, we also generated a clonal *BAG6* knockout cell line from a parental strain with a homozygous mutation in the *SMAD4* stop codon, causing it to express only the *SMAD4* extended readthrough mutant (**Figure 3.6A**). As seen by western blot, knockout of BAG6 in this endogenous readthrough SMAD4 mutant does result in a detectable stabilization of the endogenous SMAD4 readthrough mutant, further suggesting BAG6 is at least partially responsible for the loss of the endogenous SMAD4 readthrough mutant (**Figure 3.6A**).

To determine if BAG6 also binds directly to the translated *SMAD4* 3’ UTR, we performed a co-immunoprecipitation experiment similar to the one performed for *AMD1* (**Figure 3.6C and 3.6D**). Using the endogenous *SMAD4* readthrough HEK293T mutant, we performed a pulldown with an antibody specific to SMAD4 and subsequently performed a western blot with an antibody specific for BAG6. In this immunoprecipitation, we included a set of samples treated with the proteasome inhibitor Bortezomib, as the SMAD4 readthrough product is nearly undetectable under normal conditions. As expected, we could detect BAG6 in association with the SMAD4 readthrough product and detected no BAG6 associating with the wild-type SMAD4 (no translation of *SMAD4* 3’ UTR) (**Figure 3.6D**). These results show that BAG6 physically interacts with the readthrough product of SMAD4. Aside from demonstrating that the BAG6 complex targets a second protein with translated noncoding sequences and in an endogenous context, these results potentially have implications for several cancers in which SMAD4 inactivation is a critical aspect of tumorigenesis.

In further attempts to determine the full scope of BAG6 targeting of translated noncoding sequences beyond our test cases of AMD1 and SMAD4, we turned again to our original Pep30 noncoding library discussed in Chapter 2. We aimed to repeat our initial screen to determine the scope and scale of rescue that could be achieved from the knockout of BAG6 among our 12,000 unique reporters, 9000 of which are derived from noncoding sequences. To do so, we transduced the Pep30 library into wild-type HEK293T and BAG6 knockout cells and sorted each population into four bins based on EGFP/mCherry ratio using FACS (**Figure 3.7A**). The ratio gating was generated from the wild-type population such that the lowest EGFP/mCherry bin contained approximately 40% of the cell population, and the three bins with increasingly higher ratios of EGFP/mCherry contained approximately 20% of the cell population each. These same gates were then applied to the BAG6 knockout cells, which immediately demonstrated significant differences in the cell population and an overall increase in levels of EGFP expression. After sorting, genomic DNA was extracted from each bin for each sample and sequenced to determine the abundance of each reporter in every sample. We then assigned a degradation score to each reporter, defined as the abundance in the lowest EGFP/mCherry ratio bin (most highly degraded) over the abundance in all bins (**Figure 3.7B**). When the degradation scores of each reporter are plotted against each other in the wild-type and BAG6 knockout cell population, it is immediately apparent that a large number of reporters are stabilized to differing levels in the BAG6 knockout (lower degradation score when comparing the same reporter to wild-type) (**Figure 3.7B**). Notably, the degree to which a given reporter is stabilized appears to be highly dependent on the average hydrophobicity of the translated tail (**Figure 3.7C**). These results suggest that the activity of the BAG6 complex in targeting products of noncoding translation is widespread and depends on the presence of C-terminal hydrophobicity.

3.3 Discussion

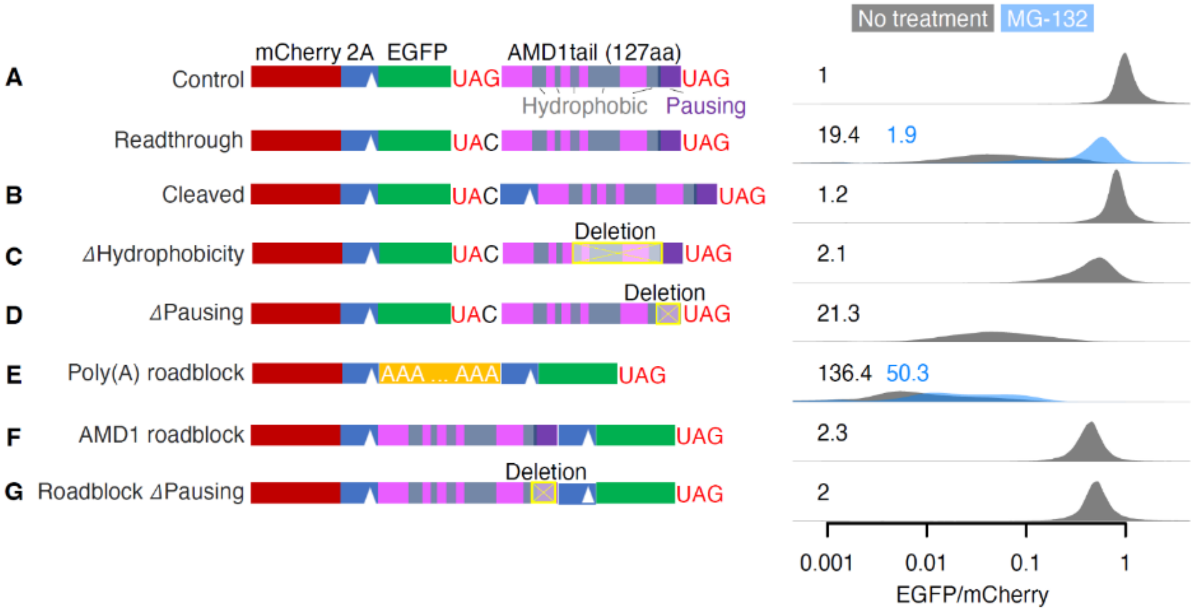
The results presented in this chapter strongly argue against the previously proposed model of ribosome queuing in the *AMD1* 3’ UTR and instead suggest that proteasomal degradation is responsible for protein loss due to translation of the *AMD1* 3’ UTR. This finding is also consistent with the conclusions of Chapter 2, as the *AMD1* 3’ UTR has several regions of high hydrophobicity. The results of our genome-wide knockout screen both supported the model of proteasomal degradation and uncovered the participation of the BAG6 membrane protein quality control complex in detecting and targeting the AMD1 reporter for proteasomal degradation.

Three features of the BAG6 pathway make it especially appealing for the surveillance of noncoding translation. First, the pathway recognizes C-terminal hydrophobic tails, a defining feature of noncoding translation products that is also associated with their degradation. Second, multiple components of this pathway, including BAG6, TRC35, and SGTA, have all been shown to be physically associated with the ribosome[101,103,104,111], positioning the complex for rapid surveillance of noncoding translation products before they are released to the cytoplasm. Consistent with this, it has also been reported that BAG6 is associated with polyubiquitinated nascent polypeptides and targets them for proteasomal degradation[112], although the identity of these nascent polypeptides remains unknown.

We further demonstrated that the role of the BAG6 complex in targeting translated noncoding sequences extends beyond AMD1 and may comprise a general mechanism by which most noncoding proteins with hydrophobic C-terminal tails are detected and targeted by the cell for degradation in the proteasome. This idea is supported by the physical interaction and rescue of the SMAD4 readthrough mutant protein and the results of our high-throughput screen combined with BAG6 knockout. Widespread targeting of translated noncoding sequences by BAG6, a protein with a primary function of chaperoning nascent tail-anchored proteins for insertion into cellular membranes, also suggests a possible mechanism by which noncoding sequences in the genome can be exposed to selection at the protein level, resulting in the emergence of novel functional proteins over evolutionary timescales. This proposed model will be discussed in more detail in Chapter 4.

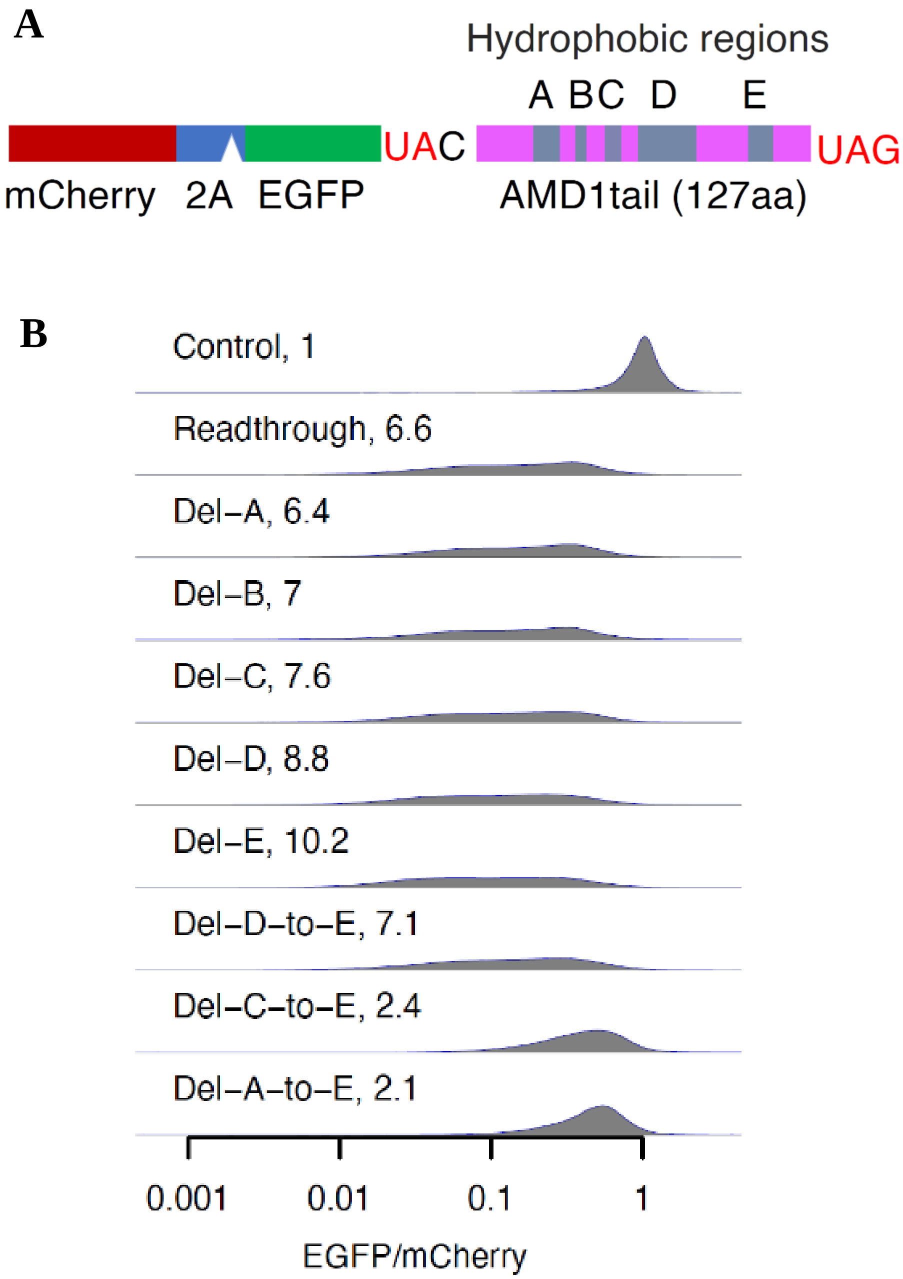
3.4 Figures

Figure 3.1: Protein loss due to translation of the AMD1 3′ UTR is rescued by proteasome inhibition and deletion of hydrophobic regions



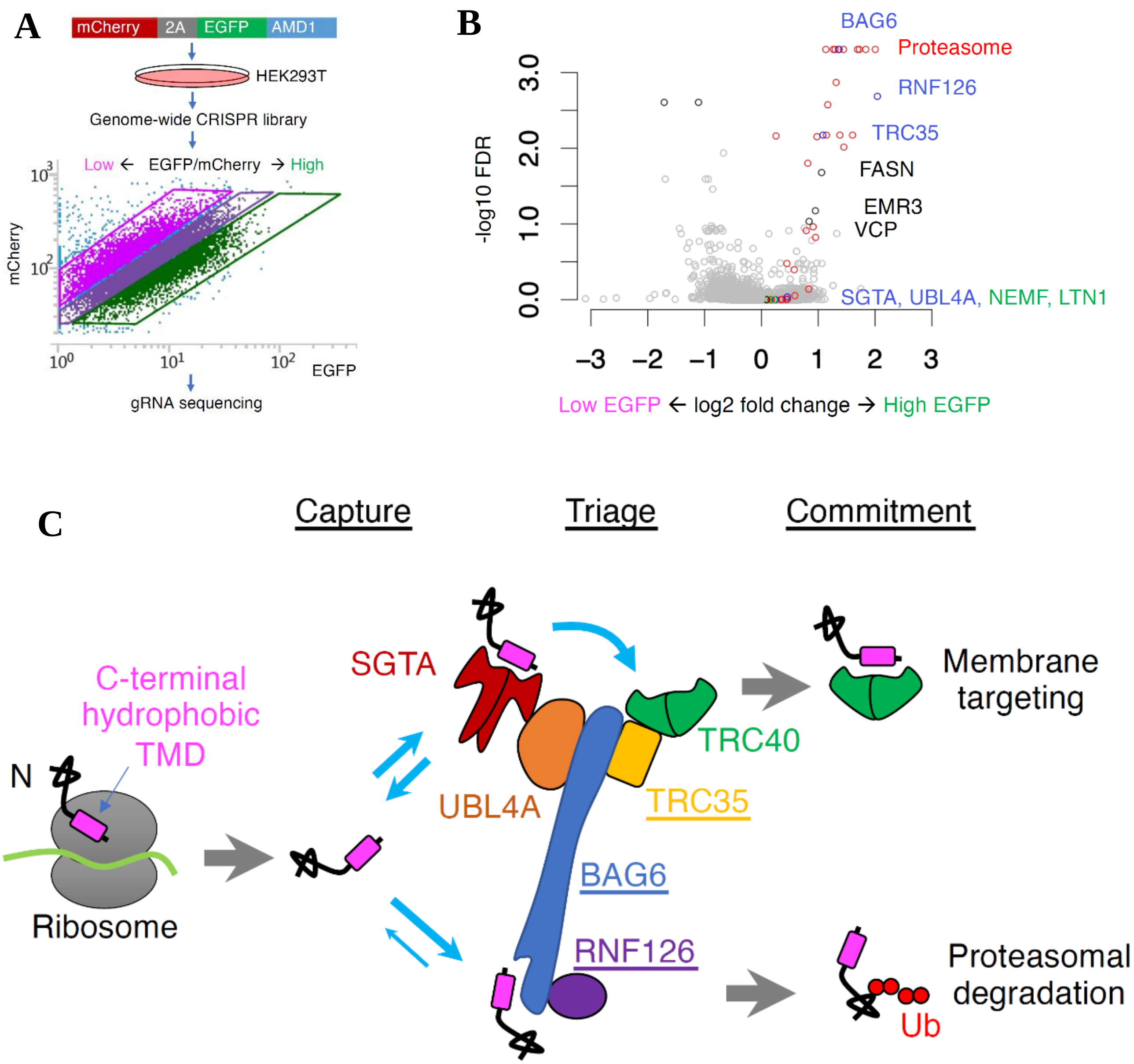
(A-G) Reporter constructs shown on the left were transfected into HEK293T cells. The EGFP/mCherry ratio was quantified in individual cells using flow cytometry with distributions shown on the right on a log-10 scale. The number in each plot is the median fold-decrease of the EGFP/mCherry ratio. Data from cells treated with the proteasome inhibitor MG-132 are shown in blue.

Figure 3.2: Protein loss due to translation of the AMD1 3′ UTR is rescued by deletion of multiple hydrophobic regions



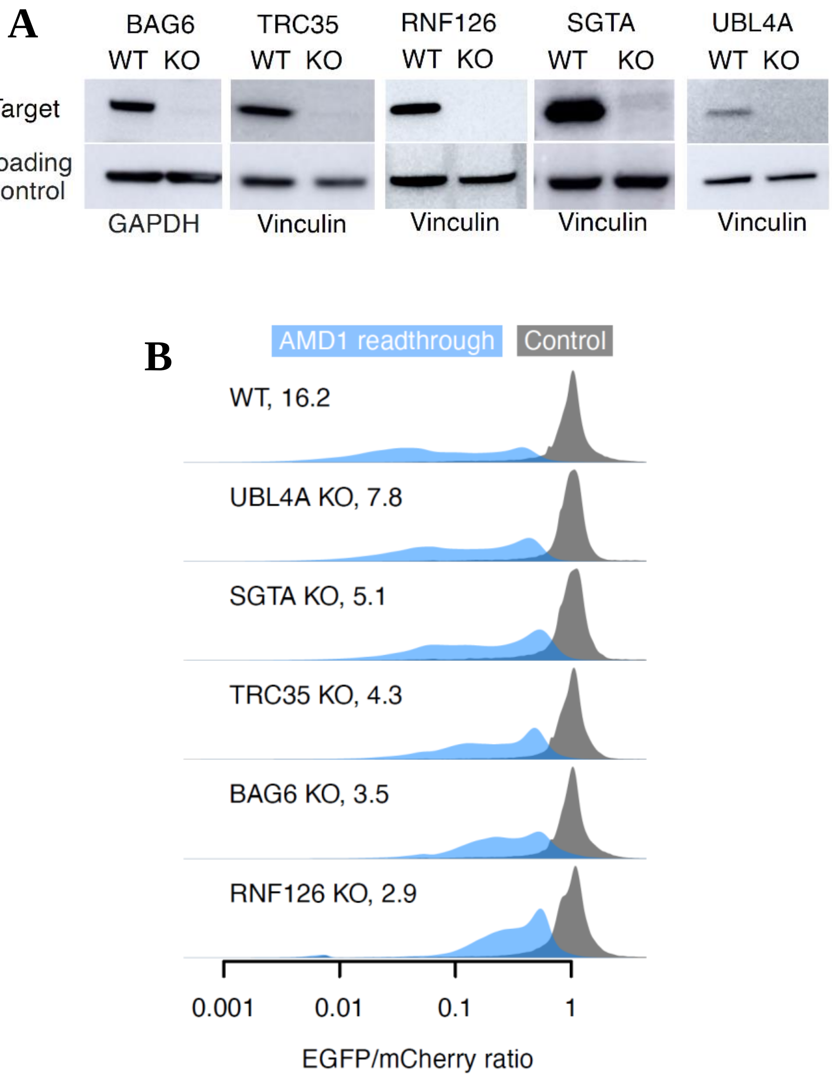
(A) Schematic showing five identified regions of high hydrophobicity in the amino acid sequence of the AMD1 3’ UTR. Lengths of individual hydrophobic regions are: A = 12 aa, B = 4 aa, C = 7 aa, D = 26 aa, E = 12 aa. (B) Deletion mutants of the AMD1 reporter show little rescue when individual hydrophobic regions are deleted but strong rescue when the final three regions are deleted simultaneously.

Figure 3.3: A genome-wide CRISPR/Cas9 knockout screen identifies the BAG6 complex in targeting of the translated AMD1 3′ UTR



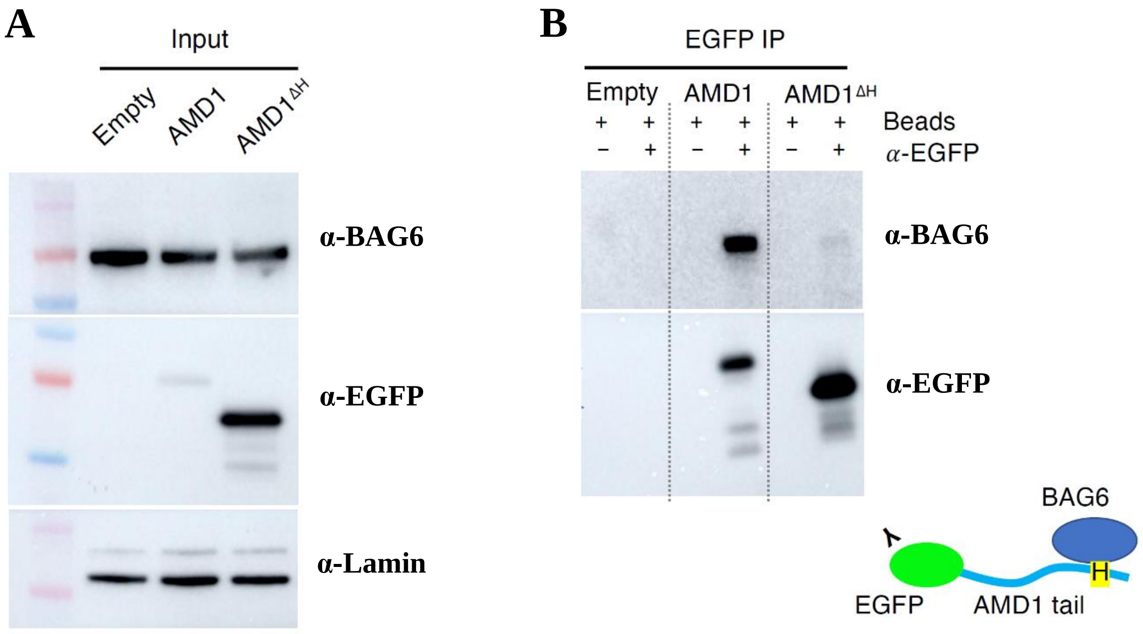
(A) A schematic of the genome-wide CRISPR screen using the AMD1 3’ UTR reporter construct. (B) Volcano plot showing the top hits identified in rescue of the AMD1 reporter from analysis of guide enrichment performed by the MAGeCK package. (C) Schematic showing the known functions of the BAG6 complex and associated proteins in membrane protein biogenesis and quality control

Figure 3.4: Knockout of genes in the BAG6 pathway validates the results of the CRISPR screen and identifies participation of SGTA and UBL4A in targeting the AMD1 tail



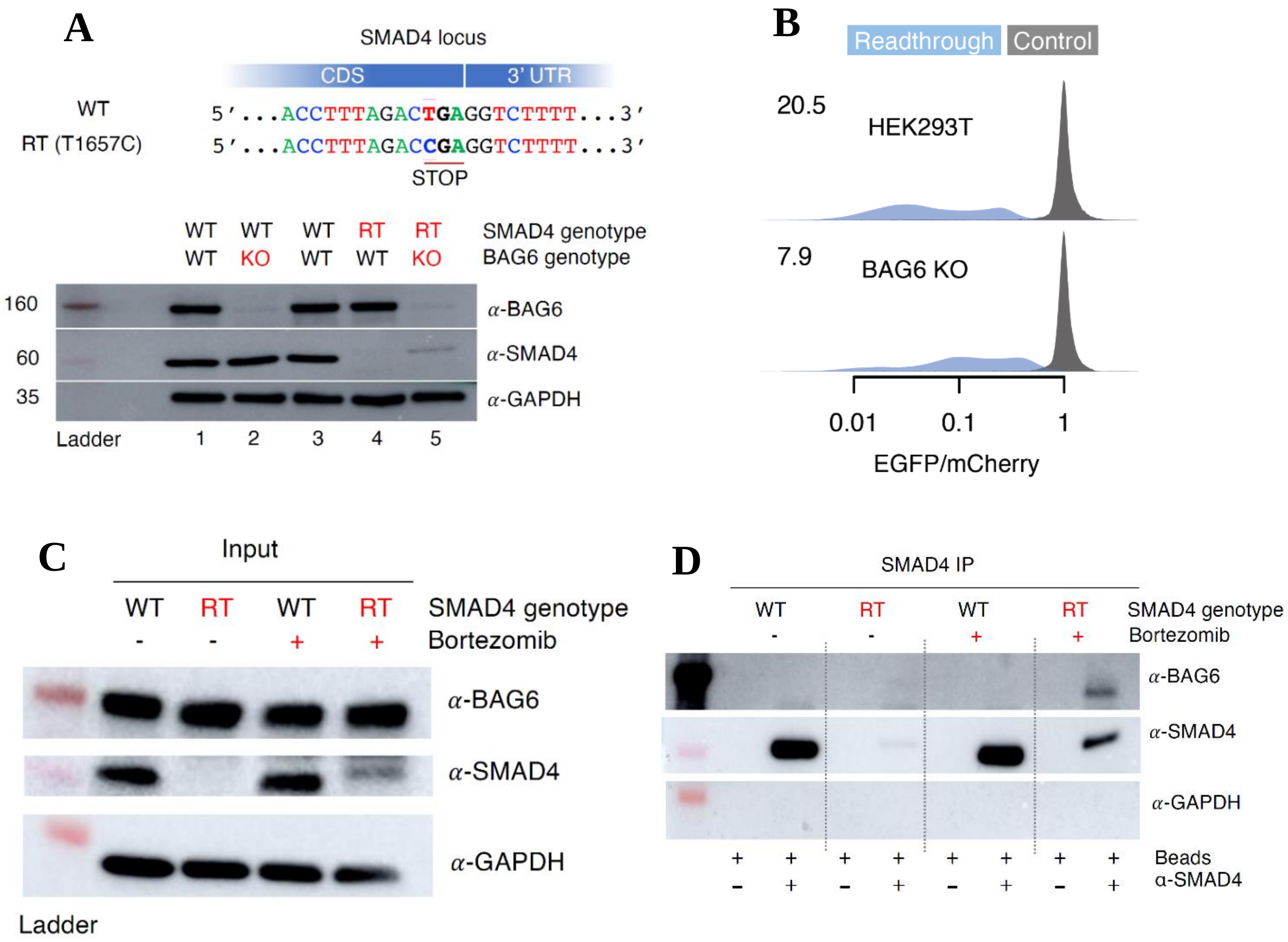
(A) Western blots showing loss of the target protein in five knockout clones isolated from HEK293T cells. (B) EGFP/mCherry ratio of the AMD1 reporter in WT and KO cells.

Figure 3.5: Co-immunoprecipitation demonstrates protein-protein interaction between BAG6 and the translated AMD1 3′ UTR



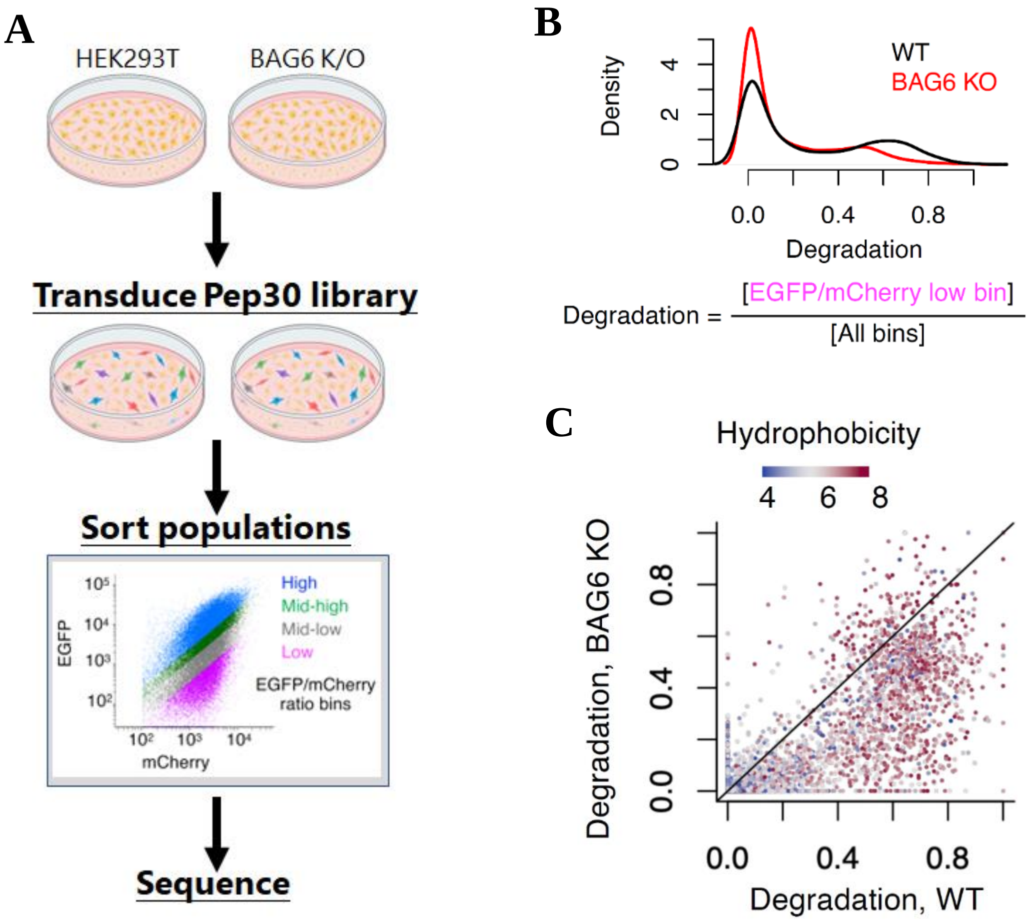
(A) Input of the BAG6 co-IP with EGFP-AMD1tail or the mutant without the C-terminal hydrophobic region (AMD1ΔH). (B) BAG6 co-immunoprecipitates with EGFP-AMD1tail but not AMD1ΔH.

Figure 3.6: Readthrough mutants of the SMAD4 tumor suppressor protein are targeted by BAG6



(A) A homozygous nonstop T1657C mutation in HEK293T cells causes readthrough (RT) translation of SMAD4, which is barely detectable in BAG6 wild type (WT) cells (lane 4) but is stabilized in BAG6 KO cells (lane 5). RT: readthrough. (B) BAG6 knockout results in partial rescue of the EGFP/mCherry ratio in a reporter expressing the translated SMAD4 3' UTR appended to EGFP. Analysis done by flow cytometry (C) Input of the BAG6 co-IP with SMAD4 readthrough product. Bortezomib: proteasome inhibitor. (D) Co-IP of BAG6 with SMAD4 readthrough products.

Figure 3.7: A high-throughput screen using the noncoding Pep30 reporter library in BAG6 knockout cells



(A) Schematic of the screen setup investigating rescue of noncoding sequences in the Pep30 library with BAG6 knockout. (B) A degradation score was assigned to each reporter detected from the sorted bins defined as the fraction of total reporters in the lowest EGFP/mCherry ratio bin over all bins. The distribution of reporter degradation scores in both wild-type HEK293T and the BAG6 knockout cell line is shown here. (C) Scatter plot of degradation score of individual reporters in the wild-type sample compared to the BAG6 knockout sample. The average hydrophobicity of each reporter is shown by its color.

3.5 Methods

Plasmids

CRISPR guide RNA plasmids: The parental lentiCRISPR v2 plasmid (Addgene # 52961) was digested with BsmBI and purified using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel). Forward and reverse oligos containing the guide sequence of interest were phosphorylated, annealed, and ligated into the parental plasmid with T4 PNK and T4 DNA ligase. Targeting and non-targeting guide sequences are derived from the CRISPR KO library described previously (Wang et al., 2014).

All plasmids were transformed into NEB Stable Competent *E. coli* (C3040) according to the manufacturer's protocol. Positive clones were confirmed via sanger sequencing.

Cell culture

HEK293T cells used in this study were purchased from ATCC. Cells were cultured in DMEM with 4.5 g/L D-Glucose supplemented with 10% fetal bovine serum, 1% penicillin/streptomycin was added except when producing lentivirus. Low passage number cells were used and maintained under 90% visual confluency. Cells were maintained at 5% CO2 and 37 °C. HEK293T cells used in this study were confirmed to be negative for Mycoplasma Contamination and routinely tested using the MycoAlertTM Mycoplasma Detection Kit (Lonza, LT07-418). For experiments involving the SMAD4 gene, clonal cell lines harboring SMAD4 readthrough mutations as well as the parental HEK293T cells were obtained as a generous gift from Dr. Sven Diederichs. Transfection of plasmids was done using Lipofectamine 2000 or Lipofectamine 3000 according to the manufacturer's instructions. Flow cytometry analyses of transfected cells was typically performed 24 or 48 hours after.

Lentivirus and stable cell line generation

For generating lentivirus, 750,000 HEK293T cells were seeded in 6-well plates with DMEM supplemented with 10% FBS. After 24 hours, the cells were transfected with the second-generation lentiviral packaging plasmids as well as the lentiviral plasmid of interest using Lipofectamine 3000. The virus-containing media was collected 48 and 72 hours after transfection, combined, clarified by centrifugation at 500 RCF for 5 minutes, and then passed through a 45 μM PVDF filter. The purified virus was stored at 4°C for short-term use or aliquoted and frozen at -80°C. For the generation of stable cell lines, HEK293T cells were reverse transduced in 6-well plates in media with 10 μg/mL polybrene using purified virus such that <30% of the cells are transduced. Twenty-four hours after transduction, the virus-containing media is removed, and fresh media is added. After another 24 hours, the cells are collected, and transduction efficiency is confirmed via flow cytometry. Transduced cells are then selected with puromycin at 2 μg/mL for 48 hours or via flow cytometry to generate a stable cell line for downstream analysis.

Flow cytometry analysis

Cells were collected and resuspended in 1-4 mL of fresh media and passed through a 35 μM mesh cell strainer immediately prior to flow cytometry. Flow cytometry was performed on either a Bio-Rad ZE5 or NovoCyte Quanteon analyzer. Gating of samples and export of data for downstream analysis was done using the FCS Express software.

Massively parallel reporter assays comparing WT and BAG6 KO HEK293T cells

HEK293T and a clonal BAG6 knockout cell line were reverse transduced with the Pep30 library such that less than 30% of cells were transduced (thus are most likely a single integration per cell). The virus-containing media was removed after 24 hours, and fresh media with 10% FBS and 1% PenStrep was added to the plates. After another 24 hours, transduced cells were purified based on their expression of mCherry. The transduced populations were returned to culture and allowed to grow out for an additional six days, with passaging as necessary to maintain confluence below 80%. After six days, both populations were sorted into four bins based on the ratio of EGFP/mCherry expression (High, mid-high, mid-low, and low) using a FACSAria cell sorter. The same mCherry/EGFP ratio gates were used for both WT and BAG6 KO cells. Sorted cells were spun down at 500 RCF for 5 minutes, washed once with 1000 uL PBS, spun down again, then frozen at -20 as a cell pellet. Genomic DNA was subsequently isolated from the cell populations using a Macherey-Nagel NucleoSpin Tissue kit, and genomic DNA was eluted in 50 uL of elution buffer. Libraries were then amplified using PCR with custom Illumina adapters, using Q5 high-fidelity PCR mix with 1000 ng input gDNA per sample. Libraries were amplified for a total of 24-27 cycles. After amplification, libraries were cleaned up using SPRISelect beads at a ratio of 0.7x. Purified library size was confirmed via gel, and libraries were quantified using the KAPA qPCR Illumina library quantification kit. Libraries were subsequently pooled in a ratio based on the total cells collected from each sample. The pooled library was sequenced on a NextSeq 550 with 2.5% PhiX spike-in, using the 75-cycle high-output kit with 80 cycles in read 1 and 8 cycles in index read 1. Reads were aligned to a custom index for the Pep30 library generated with the command bowtie-build in bowtie version 1.2.3 and the option -v 3 --best (best alignment with up to 3 mismatches). The counts of each Pep30 sequence were extracted from the alignment with the bash command cut -f 3 | sort | uniq -c. The mitigation index of each sequence in a sample is calculated by dividing the number of reads in the low EGFP/mCherry bin by the sum of read counts in all bins of the same sample.

Genome-wide CRISPR screen

The Human Activity-Optimized CRISPR Knockout Library (3 sub-libraries in lentiCRISPRv1) was obtained from addgene (https://www.addgene.org/pooled-library/sabatini-crispr-human-high-activity-3-sublibraries/) and prepared according to the standard protocol. Library lentivirus was produced using Mirus LT1 transfection reagent and second-generation packaging plasmids. 9.2x107 HEK293T cells carrying the stable AMD1-EGFP reporter were reverse transduced with the CRISPR library with 8 μg/mL polybrene. Media was changed 24 hours after transduction. Selection with 2 μg/mL puromycin was initiated 48 hours after transduction. After 48 hours of puromycin selection, cells were collected and sorted, and sorted cell populations were frozen at -80 °C. Libraries were prepared for Illumina sequencing from the sorted cell populations as described in Joung et. al., 2017. Libraries were amplified for a total of 28 PCR cycles, purified using the Zymo DNA Clean & Concentrator-5 kit, and the correct-sized band was subsequently purified by gel extraction. Fragment sizes of the libraries were confirmed by bioanalyzer and concentrations were determined using the KAPA qPCR library quantification kit. The pooled library was then sequenced on a NextSeq 550 with 86 cycles in Read 1 and 6 cycles in Index Read 1.

Co-immunoprecipitation

HEK293T cells were seeded in 10-cm plates with 3x106 cells per plate. Reporters were transfected into the cells 24 hours after seeding using Lipofectamine 3000. Forty-eight hours after transfection, cells were treated with DMSO (vehicle) or 0.1 μM Bortezomib. After 24 hours of drug treatment, cells were collected, washed twice in cold PBS, and resuspended in lysis buffer (0.025 M Tris pH 7.4, 0.15 M NaCl, 0.001 M EDTA, 1% NP-40 alternative, 5% Glycerol). Lysates were incubated at 4°C with rotation for 30 minutes, centrifuged at 12,000

RCF at 4°C for 20 minutes, and the supernatant was collected. The pulldowns were performed using Novex DYNAL Dynabeads Protein G conjugated with a primary antibody according to the manufacturer's protocol. Following coimmunoprecipitation, western blots were performed as described below.

Generation of knockout cell lines

HEK293T cells (7.5 x 10^5) were seeded in 6-well plates and transfected the next day with 4 μg of the lentiCRISPR v2 plasmid (https://www.addgene.org/52961/) containing a sgRNA sequence specific to the targeted gene. After 24 hours, cells were passaged into media containing 2 μg/mL puromycin. After two days of puromycin selection, cells were collected, and single cells were sorted into 96-well plates. Individual clones were allowed to grow for 1-4 weeks and then passaged into 6-well plates. Clones were then screened for frameshift mutations in both alleles of the target gene using sanger sequencing and the ICE CRISPR analysis tool (https://www.synthego.com/products/bioinformatics/crispr-analysis). Complete knockout of the target genes was then verified using western blotting. Additionally, for BAG6 KO cells, the target locus was PCR-amplified and cloned into plasmids. Sanger sequencing of ten clones confirmed two frameshifting alleles, one with a 5-nt deletion and the other with a 11-nt deletion (Fig. 5SA).

Western blotting

Cells were cultured and transfected where applicable as described above. Cells were collected on ice and washed with cold PBS and subsequently lysed in RIPA buffer supplemented with a 1X protease inhibitor cocktail for 30 minutes at 4 °C on a rotator. Lysates were then cleared by centrifugation at 16,000 RCF and 4 °C for 20 minutes. Protein concentrations were determined using a BCA assay, and samples were then prepared using LDS sample buffer supplemented with sample reducing agent and heated to 70 C for 10 minutes. Samples were then run on an SDS-PAGE gel and transferred to an activated PVDF membrane for 90 minutes at 30 volts or overnight at 10 volts. Membranes were blocked with 5% BSA in PBS-T for 1 hour at room temperature or overnight at 4 °C. Membranes were then cut and incubated with the appropriate primary antibody in blocking buffer supplemented with 0.02% sodium azide for 1 hour at room temperature or overnight at 4 °C. Secondary antibodies were added at a 1:10,000 dilution and incubated for 1 hour at room temperature. Immobilon ECL Ultra Western HRP Substrate was then added to the membranes and blots were visualized using an Amersham Imager 600.

Massively parallel reporter assays comparing WT and BAG6 KO HEK293T cells

HEK293T and a clonal BAG6 knockout cell line were reverse transduced with the Pep30 library such that less than 30% of cells were transduced (thus are most likely a single integration per cell). The virus-containing media was removed after 24 hours and fresh media with 10% FBS and 1% PenStrep was added to the plates. After another 24 hours, transduced cells were purified based on their expression of mCherry. The transduced populations were returned to culture and allowed to grow out for an additional six days, with passaging as necessary to maintain confluence below 80%. After six days, both populations were sorted into four bins based on the ratio of EGFP/mCherry expression (High, mid-high, mid-low, and low) using a FACSAria cell sorter. The same mCherry/EGFP ratio gates were used for both WT and BAG6 KO cells. Sorted cells were spun down at 500 RCF for 5 minutes, washed once with 1000 uL PBS, spun down again, then frozen at -20 as a cell pellet.

Genomic DNA was subsequently isolated from the cell populations using a Machery Nagel Nucleospin Tissue kit, and genomic DNA was eluted in 50 uL of elution buffer. Libraries were then amplified using PCR with custom Illumina adapters, using Q5 high-fidelity PCR mix with 1000 ng input gDNA per sample. Libraries were amplified for a total of 24-27 cycles. After amplification, libraries were cleaned up using SPRISelect beads at a ratio of 0.7x. Purified library size was confirmed via gel and libraries were quantified using the KAPA qPCR Illumina library quantification kit. Libraries were subsequently pooled in a ratio based on the number of total cells collected from each sample. The pooled library was sequenced on a NextSeq 550 with 2.5% PhiX spike in, using the 75-cycle high-output kit with 80 cycles in read 1 and 8 cycles in index read 1.

Reads were aligned to a custom index for the Pep30 library generated with the command bowtie-build in bowtie version 1.2.3 and the option -v 3 --best (best alignment with up to 3 mismatches). The counts of each Pep30 sequence were extracted from the alignment with the bash command cut -f 3 | sort | uniq -c. The mitigation index of each sequence in a sample is calculated by dividing the number of reads in the low EGFP/mCherry bin by the sum of read counts in all bins of the same sample.

Chapter 4: Insights into the evolution of novel proteins from the noncoding genome

The work in this chapter is adapted in part from the following preprint:

Kesner, Jordan S. Chen, Ziheng. Aparicio, Alexis A. Wu, Xuebing. A unified model for the surveillance of translation in diverse noncoding sequences. Available at bioRxiv: https://www.biorxiv.org/content/10.1101/2022.07.20.500724v1

4.1 Introduction

The results discussed thus far highlight a fascinating question: for what purpose are translated noncoding sequences channeled into a pathway that is well established to function in the biogenesis and quality control of membrane proteins? While it could be argued that the cell has simply chosen to use the quality control aspect of this pathway as an already established system for triaging rare noncoding translation events, there is some fascinating evidence that this process as a whole may play a role in the evolution of novel proteins, which would likely first be directed to cellular membranes but could evolve further from there. While the data presented here suggest that on a physiological timescale, translated noncoding sequences will be targeted and directed to the proteasome for degradation through the BAG6 pathway, on an evolutionary timescale, this same process may allow for the exposure of the noncoding genome to natural selection at the protein level, a process that would potentiate the evolution of novel proteins from the noncoding genome. In this context, it makes perfect sense that widespread translation of noncoding sequences at a low level by the mechanisms discussed in Chapter 1 would occur. Regardless of whether this process occurs due to imperfect mechanisms of translation or is somehow programmed into the cell genetically, low-level but widespread translation of the noncoding genome allows genomes to expose novel sequences to natural selection while simultaneously preventing quality control mechanisms from being overwhelmed by a large number of potentially toxic or aggregation-prone polypeptides.

Precedent for such a mechanism exists from prior studies investigating proto-genes in yeast cells[113]. For example, previous studies in yeast have shown hundreds of potential proto-genes derived from translated noncoding sequences are potentially functional and subject to selection through various assays and analyses[113]. These proto-genes also tend to contain putative transmembrane domains[113,114]. We manually curated 64 translated functional noncoding peptides in humans from the available literature to determine if there was a preference for localization to cellular membranes[11,115–133]. Indeed, 47 of the 64 translated functional noncoding peptides we curated appear to localize to cellular membranes. In light of this evidence, we set out in this chapter to determine if there is evidence for the evolution of novel proteins from noncoding sequences in the human genome.

4.2 Results

To determine if C-terminal hydrophobicity underlies the aforementioned differential stability between canonical protein C-termini and all other sequences, including internal protein sequences and peptides derived from noncoding sequences, we performed genome-wide *in silico* analysis of C-terminal hydrophobicity in both the canonical proteome and the predicted noncoding proteome. Specifically, we calculated the average hydrophobicity for each of the last 100 residues coded by both the annotated coding sequences (CDS, n = 40,324 unique amino acid sequences, >= 200-aa) and predicted peptides (>= 30aa) from various noncoding sequences, including in-frame ORFs extended into introns (n=200,284) and 3' UTRs (n = 14,057) as well as the longest ORFs in 5' UTRs (n = 11,790) and lncRNAs (n = 29,788). Indeed, we found that hydrophobic residues are progressively depleted towards the C-terminal end of canonical proteins (CDS), especially the last 30 aa, whereas the opposite trend is present for all noncanonical peptides (**Figure 4.1A**). Notably, the very C-termini of peptides from introns, 3' UTRs, and lncRNAs have a hydrophobicity approaching that of entirely random amino acid sequences, suggesting that by default, unevolved nonfunctional proteins will have a relatively high average hydrophobicity, and are subjected to proteasomal degradation. The difference in hydrophobicity disappears further away from the very C-termini (50-100aa upstream) of proteins. Given that only longer ORFs (> 50-aa) were used in calculating the average hydrophobicity in the upstream region, these results suggest that longer noncanonical ORF peptides are either also under selection to deplete hydrophobicity and thus may be functional, or they are in fact alternative or mis-annotated isoforms of functional proteins.

Further supporting the evolutionary selection against protein C-tail hydrophobicity, we found that in humans and mice, evolutionarily young protein-coding genes tend to have higher hydrophobicity at the C-terminal tail (last 30aa) than evolutionarily older genes (**Figure 4.1B**). For example, human-specific genes - the youngest human genes that originated after the human-chimpanzee divergence 4 to 6 million years ago[134] - have the highest C-terminal hydrophobicity as a group than older genes in the human genome. A strong negative correlation (R_s = -0.97, $p < 10^{-15}$) is observed between estimated gene age and average protein C-tail hydrophobicity in the mouse genome, supporting the idea that as genes evolve, they progressively lose hydrophobic residues in the C-terminal tail, potentially resulting in longer protein half-lives. A similar albeit weaker trend is observed in the human genome, especially for genes originating within the last 100 million years.

To further understand the propensity of noncoding sequences to code for hydrophobic amino acids, we first used kpLogo to test if hydrophobic residues are associated with nucleotide bias in the genetic code, as has been suggested previously[135,136]. We confirmed that codons coding for hydrophobic residues are more likely to have Uracil (U) at all three positions, and especially at the center position of the codon (**Figure 4.2A**). Indeed, all 16 codons with U at the center code for highly hydrophobic amino acids (**Figure 4.2C**).

Because canonical coding sequences have evolved to be GC-rich / AT-poor (47.0% AT) relative to the AT-rich genome background (54.6% AT), sequences outside of functional coding regions are thus T/U-rich and will tend to code for more hydrophobic residues. Indeed, we found a strong agreement between U-content and C-tail hydrophobicity across different regions (**Figure 4.3A and 4.3E**). For example, introns have the highest U-content (31.0%) and also have the highest C-tail hydrophobicity, whereas 5’ UTRs have a U-content comparable to coding regions and are also associated with moderate hydrophobicity. The high GC-content in 5’ UTRs is largely due to the presence of CpG islands in most human gene promoters[137].

Our combined results suggest a model in which peptides derived from the translation of noncoding sequences are likely to contain hydrophobic C-terminal domains through which they will be recognized and captured by the BAG6 complex. While many of these peptides will likely be targeted for proteasomal degradation, some may escape this path and instead be targeted by the BAG6 complex for membrane insertion. This creates a mechanism through which noncoding regions of the genome can be exposed to selection at the protein level, allowing for the emergence of novel membrane-bound proteins over evolutionary time scales. In light of this proposed model, we asked if we could find evidence that functional peptides originating from noncoding sequences are preferentially localized to cellular membranes. We searched the available literature for examples of functional peptides derived from the translation of 5' UTRs or previously annotated lncRNAs in mammalian cells for which localization of the peptide had been experimentally determined. Indeed, a large majority (47 of 64 total peptides) of the identified peptides demonstrated localization to various cellular membranes[11,115–130,132,133,138] (**Figure 4.3A**).

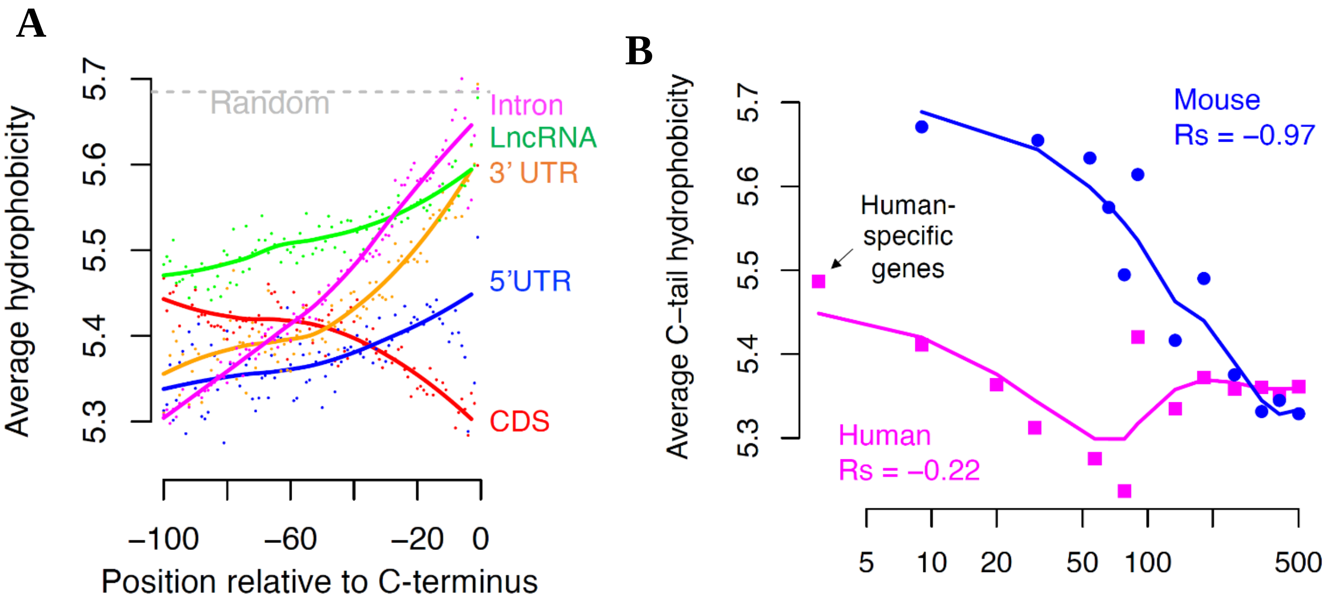
4.3 Discussion

Combined with our previous findings that the protein products of translation in noncoding regions of the genome can be fed into the BAG6 membrane protein quality control pathway and are targeted via hydrophobic C-terminal domains, the results discussed in this chapter further suggest a plausible mechanism by which novel membrane proteins may evolve over time. As this process requires selective pressure at the protein level, BAG6 provides a necessary link between the noncoding nucleotide sequences in the genome and selective pressure of potential novel functional proteins. Previous studies in yeast have suggested similar mechanisms by which novel proteins may evolve from proto genes and show a tendency to form transmembrane domains[113,114]. In this chapter, we show that the intrinsic nucleotide bias in the noncoding genome and in the genetic code will frequently result in protein translated from noncoding sequences to harbor hydrophobic C-terminal tails. These tails are then subject to targeting by the BAG6 complex, which may occasionally mistake the protein for a genuine tail-anchored protein and shuttle them for insertion into cellular membranes. To what degree noncoding proteins are targeted for membrane insertion as opposed to proteasomal degradation, and what features may influence this decision remain unclear.

Our curation of the literature further showed that the majority of functional proteins translated from noncoding sequences localize to cellular membranes[11,115–130,132,133,138]. One question that remains to be answered is what underlies this preferential membrane targeting of noncoding proteins. For example, it is plausible that targeting unknown and potentially aggregation-prone proteins to membranes functions to sequester the hydrophobic domain away from the cellular cytosol while still exposing the soluble protein domain to the cytosol to allow for interaction with existing proteins or other cellular components. However, further investigation is required to determine if this is the case or if another mechanism underlies this tendency aside from the nucleotide bias in the noncoding genome.

4.4 Figures

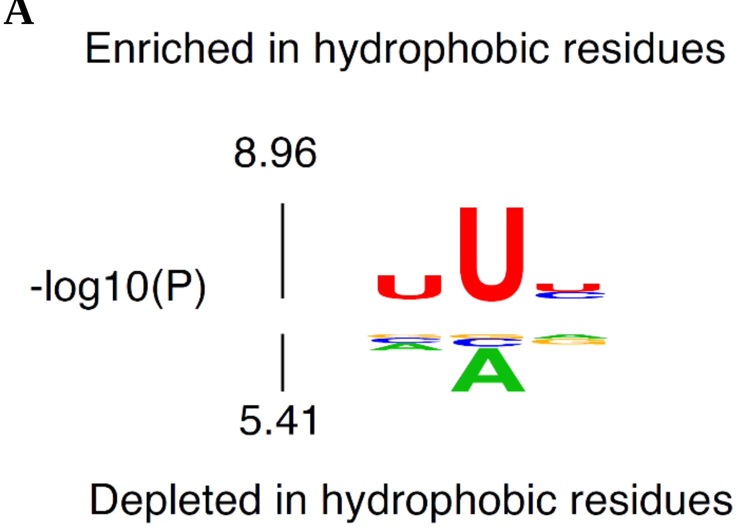
Figure 4.1: Depletion of C-terminal hydrophobicity in annotated proteins



(A) Genome-scale average hydrophobicity at each residue within the last 100 aa of peptides encoded by coding (>= 200 aa) and various noncoding sequences (>= 30 aa). (B) Average C-tail (last 30 aa) hydrophobicity of human (magenta) and mouse (blue) genes grouped by age based on time of origination estimated from vertebrate phylogeny. The lines are a loess fit of the dots.

Figure 4.2: Nucleotide bias in both the genetic code and the genome drives hydrophobicity in noncanonical peptides

A



B

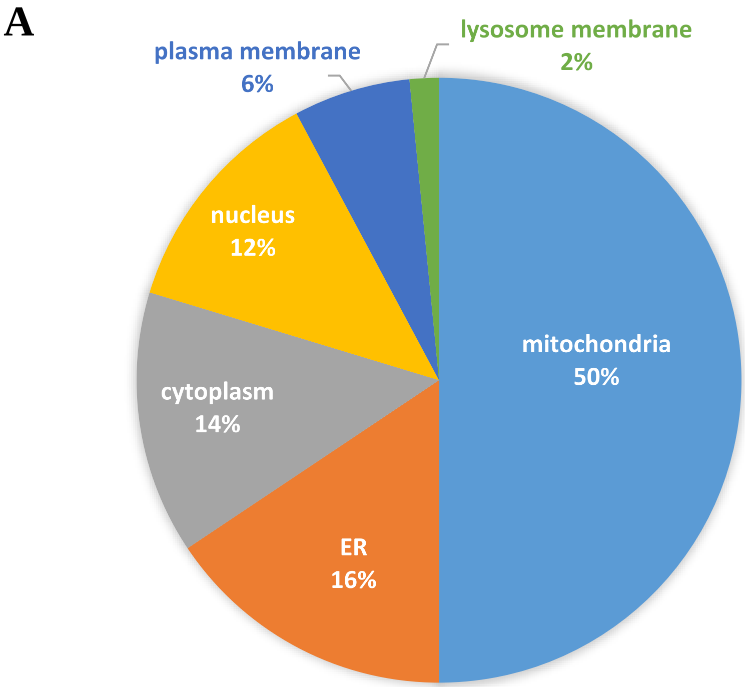
	U/T	A	C	G
Intron	31.0%	28.4%	19.9%	20.7%
3' UTR	28.3%	26.8%	22.5%	22.4%
Genome	26.6%	28.0%	22.7%	22.7%
LncRNA	26.6%	28.0%	22.7%	22.7%
Coding	21.7%	25.2%	26.5%	26.5%
5' UTR	21.7%	22.6%	27.7%	28.0%

C

F	UUU	9.03
F	UUC	9.03
M	AUG	8.95
I	AUU	8.83
I	AUC	8.83
I	AUA	8.83
L	UUA	8.47
L	UUG	8.47
L	CUU	8.47
L	CUC	8.47
L	CUA	8.47
L	CUG	8.47
C	UGU	7.93
C	UGC	7.93
W	UGG	7.66
V	GUU	7.63
V	GUC	7.63
V	GUA	7.63
V	GUG	7.63
Y	UAU	5.89
Y	UAC	5.89
A	GCU	5.33
A	GCC	5.33
A	GCA	5.33
A	GCG	5.33
H	CAU	5.1
H	CAC	5.1
U	ACU	4.49
U	ACC	4.49
U	ACA	4.49
U	ACG	4.49
G	GGU	4.48
G	GGC	4.48
G	GGA	4.48
G	GGG	4.48
R	CGU	4.18
R	CGC	4.18
R	CGA	4.18
R	CGG	4.18
R	AGA	4.18
R	AGG	4.18
S	UCU	4.09
S	UCC	4.09
S	UCA	4.09
S	UCG	4.09
S	AGU	4.09
S	AGC	4.09
P	CCU	3.87
P	CCC	3.87
P	CCA	3.87
P	CCG	3.87
Q	CAA	3.87
Q	CAG	3.87
N	AAU	3.71
N	AAC	3.71
E	GAA	3.65
E	GAG	3.65
D	GAU	3.59
D	GAC	3.59
K	AAA	2.95
K	AAG	2.95

(A) kpLogo plot visualizing the association between nucleotides at each position and amino acid hydrophobicity. (B) Nucleotide composition in different types of regions in the human genome. (C) Codons ranked by the hydrophobicity of the corresponding amino acids.

Figure 4.3: Localization of functional peptides derived from noncoding sequences to cellular membranes



(A) Curation of the available literature for functional peptides derived from translation of 5' UTRs or previously annotated lncRNAs in mammalian cells for which localization of the peptide had been experimentally determined. A majority of the identified peptides (47/64) are localized to various cellular membranes

4.5 Methods

Genome-scale hydrophobicity analysis

We systematically compared C-terminal hydrophobicity of proteins encoded by coding and noncoding sequences. The coding sequences (CDS) of annotated proteins were downloaded from Ensembl (Homo_sapiens.GRCh38.cds.all.fa) and translated into proteins using BioPython. Only proteins with more than 200 aa were used for downstream analysis. The cDNA sequences for protein-coding and long noncoding RNA transcripts(lncRNA) were obtained from GENCODE v37. From the coding transcripts the 5’ UTR and 3’ UTR sequences were extracted. For both 5’ UTR and lncRNA, the longest ORF was translated into peptides. For 3’ UTR and introns, the first in-frame stop codon marks the end of the tail ORF and only those with at least 30 codons were used. Noncoding sequence encoded peptides were removed if found in the canonical proteome. For each group, the average hydrophobicity at each position relative to the last amino acid (the most C-terminal) was calculated using the *hydrophobicity* function in the R package *Peptides.*

Correlation between C-tail hydrophobicity and gene age

Gene age was inferred by a previous study (Zhang et al., 2010). Briefly, human and mouse genes were assigned to branches of the vertebrate phylogenetic tree based on the presence and absence of orthologs in various species. The age of the genes in a branch is calculated as the middle point of each branch. The average hydrophobicity of the last 30 aa of all genes in a branch was calculated using the R package described above.

Chapter 5: Conclusions

5.1 Conclusions

Recent studies have shown that translation outside of canonical coding sequences is pervasive in cells. Proteins resulting from the translation of noncoding sequences can potentially be toxic or prone to aggregation; consequently, cellular mechanisms must be in place to detect and remove these proteins. Understanding these mechanisms not only has implications for basic science but also has potential therapeutic value. The relationship between noncoding translation and human health is multifaceted, in that noncoding translation influences various genetic diseases and cancers while, at the same time, many pathological conditions can result in a cellular environment in which the translation of noncoding sequences is likely to occur. Despite this, investigations into the surveillance mechanisms of noncoding proteins have been limited in scope and lack a consensus among their findings. In this work, we set out to use systems-based approaches to identify a general mechanism for the surveillance of translation in noncoding sequences.

In chapter 2, we identified a feature common to translation in diverse types of noncoding sequences, including 5’ UTRs, 3’ UTRs, introns, and lncRNAs. That is, translation in any of these regions will result in a protein with a C-terminal region derived from noncoding sequence. We leveraged this shared feature to design a high-throughput assay that could be used to investigate translation in thousands of sequences derived from 5’ UTRs, 3’ UTRs, introns, and lncRNAs simultaneously. The results of this experiment revealed that translation in diverse types of noncoding sequences commonly resulted in protein loss, which strongly correlated with the translated sequence's hydrophobicity. Follow-up experiments identified the proteasome as the likely endpoint of protein loss in many of these sequences and suggested that lysosomal degradation did not play a significant role.

Chapter 3 focused on identifying the mechanisms that function to link the translation of noncoding proteins to their eventual degradation in the proteasome. A genome-wide CRISPR/Cas9 knockout screen in a cell line expressing a reporter of 3’ UTR translation that causes protein loss identified the BAG6/TRC35/RNF126 membrane protein quality control complex in the rescue of the reporter. The BAG6 complex is known to recognize tail-anchored membrane proteins through their hydrophobic C-terminal tails and target them either for membrane insertion or proteasomal degradation, making it a good candidate for the surveillance of noncoding proteins. Through further experiments using a cell line knockout model of BAG6, we found that BAG6 targeting of noncoding proteins is widespread and includes a readthrough mutant of the SMAD4 tumor suppressor protein and a large fraction of reporters with high hydrophobicity from our high-throughput noncoding translation library.

In chapter 4, we investigated a potential link between the evolution of novel proteins from the noncoding genome and the features of noncoding translation identified in chapters 2 and 3. Genome-wide analysis of the average hydrophobicity within 100 residues of the C-terminus in coding and noncoding peptides showed a significant trend toward higher hydrophobicity in noncoding peptides, and lower hydrophobicity in coding peptides as the C-terminus is approached.

In conclusion, this work presents several key findings related to mechanisms used by cells to mitigate translation in noncoding regions of the genome. These findings include the signal cells use to detect translated noncoding sequences, the proteins responsible for detecting that signal, and the endpoint of degradation for noncoding proteins. Through experimental methods and computational analysis of genomic sequences, we show that the surveillance mechanism targeting noncoding proteins for degradation is widespread and likely applies to many instances of translation in diverse types of noncoding sequences. We further propose an adjacent mechanism through which this system may expose noncoding sequences in the genome to natural selection at the protein level, potentially resulting in the generation of novel functional membrane proteins over evolutionary timescales. Summarizing these findings, we propose a unified model for the surveillance of translation in diverse noncoding sequences that can be stated as follows: due to the nucleotide composition of the noncoding genome, proteins resulting from the translation of noncoding sequences are likely to have hydrophobic C-terminal tails, which are captured by the BAG6 membrane protein quality control complex and targeted for proteasomal degradation or membrane insertion. This represents a fail-safe mechanism through which cells can both prevent the accumulation of potentially toxic proteins while also providing a pathway for exposing noncoding sequences in the genome to selection at the protein level.

References

1. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
2. Jensen, T. H., Jacquier, A. & Libri, D. Dealing with Pervasive Transcription. *Molecular Cell* **52**, 473–484 (2013).
3. Selinger, D. W. *et al.* RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. *Nature Biotechnology* **18**, 1262–1268 (2000).
4. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).
5. Ntini, E. *et al.* Polyadenylation site–induced decay of upstream transcripts enforces promoter directionality. *Nature Structural & Molecular Biology* **20**, 923–928 (2013).
6. Ingolia, N. T. *et al.* Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Reports* **8**, 1365–1379 (2014).
7. Derti, A. *et al.* A quantitative atlas of polyadenylation in five mammals. *Genome Research* **22**, 1173–1183 (2012).
8. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* **40**, 1413–1415 (2008).
9. Wang, E. T. *et al.* Alternative Isoform Regulation in Human Tissue Transcriptomes. *Nature* **456**, 470–476 (2008).
10. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nature Reviews Genetics* **15**, 205–213 (2014).
11. Chen, J. *et al.* Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146 (2020).

12. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08890 (2015).

13. Weatheritt, R. J., Sterne-Weiler, T. & Blencowe, B. J. The ribosome-engaged landscape of alternative splicing. *Nature Structural & Molecular Biology* **23**, 1117–1123 (2016).

14. Lee, S.-H. *et al.* Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature* **561**, 127–131 (2018).

15. Hsu, T. Y.-T. *et al.* The spliceosome is a therapeutic vulnerability in MYC-driven cancer. *Nature* **525**, 384–388 (2015).

16. Wang, D. *et al.* Inhibition of nonsense-mediated RNA decay by the tumor microenvironment promotes tumorigenesis. *Molecular and cellular biology* **31**, 3670–80 (2011).

17. Yoshida, K. *et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64–69 (2011).

18. Dvinge, H. & Bradley, R. K. Widespread intron retention diversifies most cancer transcriptomes. *Genome Medicine* **7**, 45 (2015).

19. Adusumalli, S., Ngian, Z., Lin, W., Benoukraf, T. & Ong, C. Increased intron retention is a post-transcriptional signature associated with progressive aging and Alzheimer's disease. *Aging Cell* **18**, e12928 (2019).

20. Hsieh, Y.-C. *et al.* Tau-Mediated Disruption of the Spliceosome Triggers Cryptic RNA Splicing and Neurodegeneration in Alzheimer's Disease. *Cell reports* **29**, 301-316.e10 (2019).

21. Mariotti, M., Kerepesi, C., Oliveros, W., Mele, M. & Gladyshev, V. N. Deterioration of the human transcriptome with age due to increasing intron retention and spurious splicing. *bioRxiv* 2022.03.14.484341 (2022) doi:10.1101/2022.03.14.484341.

22. Mazin, P. *et al.* Widespread splicing changes in human brain development and aging. *Molecular Systems Biology* **9**, 633–633 (2013).

23. Son, H. G. *et al.* RNA surveillance via nonsense-mediated mRNA decay is crucial for longevity in daf-2/insulin/IGF-1 mutant C. elegans. *Nature Communications* **8**, 14749 (2017).

24. Bai, B. *et al.* U1 small nuclear ribonucleoprotein complex and RNA splicing alterations in Alzheimer's disease. *Proceedings of the National Academy of Sciences* **110**, 16562–16567 (2013).

25. Sun, Y., Eshov, A., Zhou, J., Isiktas, A. U. & Guo, J. U. C9orf72 arginine-rich dipeptide repeats inhibit UPF1-mediated RNA decay via translational repression. *Nature Communications* **11**, 3354 (2020).

26. Xu, W. *et al.* Reactivation of nonsense-mediated mRNA decay protects against C9orf72 dipeptide-repeat neurotoxicity. *Brain* **142**, 1349–1364 (2019).

27. Dhamija, S. *et al.* A pan-cancer analysis reveals nonstop extension mutations causing SMAD4 tumour suppressor degradation. *Nature Cell Biology* **22**, 999–1010 (2020).

28. Shibata, N. *et al.* Degradation of Stop Codon Read-through Mutant Proteins via the Ubiquitin-Proteasome System Causes Hereditary Disorders*. *Journal of Biological Chemistry* **290**, 28428–28437 (2015).

29. Wangen, J. R. & Green, R. Stop codon context influences genome-wide stimulation of termination codon readthrough by aminoglycosides. *eLife* **9**, e52611 (2020).

30. Sudmant, P. H., Lee, H., Dominguez, D., Heiman, M. & Burge, C. B. Widespread Accumulation of Ribosome-Associated Isolated 3′ UTRs in Neuronal Cell Populations of the Aging Brain. *Cell Reports* **25**, 2447-2456.e4 (2018).

31. Arribere, J. A. *et al.* Translation readthrough mitigation. *Nature* **534**, 719–723 (2016).

32. Hashimoto, S., Nobuta, R., Izawa, T. & Inada, T. Translation arrest as a protein quality control system for aberrant translation of the 3′-UTR in mammalian cells. *FEBS Letters* **593**, 777–787 (2019).

33. Kramarski, L. & Arbely, E. Translational read-through promotes aggregation and shapes stop codon identity. *Nucleic Acids Research* **48**, 3747–3760 (2020).

34. Yordanova, M. M. *et al.* AMD1 mRNA employs ribosome stalling as a mechanism for molecular memory formation. *Nature* **553**, 356–360 (2018).

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

36. Barrett, L. W., Fletcher, S. & Wilton, S. D. Untranslated Gene Regions and Other Non-coding Elements. *SpringerBriefs in Biochemistry and Molecular Biology* 1–56 (2013) doi:10.1007/978-3-0348-0679-4_1.

37. Dever, T. E. & Green, R. The Elongation, Termination, and Recycling Phases of Translation in Eukaryotes. *Cold Spring Harbor Perspectives in Biology* **4**, a013706 (2012).

38. Schuller, A. P. & Green, R. Roadblocks and resolutions in eukaryotic translation. *Nature Reviews Molecular Cell Biology* **19**, 526–541 (2018).

39. Brandman, O. & Hegde, R. S. Ribosome-associated protein quality control. *Nature Structural & Molecular Biology* **23**, 7–15 (2016).

40. Joazeiro, C. A. P. Ribosomal Stalling During Translation: Providing Substrates for Ribosome-Associated Protein Quality Control. *Annual review of cell and developmental biology* **33**, 343–368 (2017).

41. Fernandes, J. C. R., Acuña, S. M., Aoki, J. I., Floeter-Winter, L. M. & Muxel, S. M. Long Non-Coding RNAs in the Regulation of Gene Expression: Physiology and Disease. *Non-Coding RNA* **5**, 17 (2019).

42. Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A. & Couso, J. P. Peptides Encoded by Short ORFs Control Development and Define a New Eukaryotic Gene Family. *PLoS Biology* **5**, e106 (2007).

43. Kondo, T. *et al.* Small Peptides Switch the Transcriptional Activity of Shavenbaby During Drosophila Embryogenesis. *Science* **329**, 336–339 (2010).

44. Magny, E. G. *et al.* Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames. *Science* **341**, 1116–1120 (2013).

45. Pauli, A. *et al.* Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors. *Science* **343**, 1248636–1248636 (2014).

46. Zhang, H., Wang, Y. & Lu, J. Function and Evolution of Upstream ORFs in Eukaryotes. *Trends in Biochemical Sciences* **44**, 782–794 (2019).

47. Zhang, H. *et al.* Genome-wide maps of ribosomal occupancy provide insights into adaptive evolution and regulatory roles of uORFs during Drosophila development. *PLoS Biology* **16**, e2003903 (2018).

48. Johnstone, T. G., Bazzini, A. A. & Giraldez, A. J. Upstream ORF s are prevalent translational repressors in vertebrates. *The EMBO Journal* **35**, 706–723 (2016).

49. Beznosková, P., Wagner, S., Jansen, M. E., Haar, T. von der & Valášek, L. S. Translation initiation factor eIF3 promotes programmed stop codon readthrough. *Nucleic acids research* **43**, 5099–111 (2015).

50. Wang, J. *et al.* AAV-delivered suppressor tRNA overcomes a nonsense mutation in mice. *Nature* **604**, 343–348 (2022).

51. Ko, W., Porter, J. J., Sipple, M. T., Edwards, K. M. & Lueck, J. D. Efficient suppression of endogenous CFTR nonsense mutations using anticodon-engineered transfer RNAs. *Molecular therapy. Nucleic acids* **28**, 685–701 (2022).

52. Tawde, M., Bior, A., Feiss, M., Teng, F. & Freimuth, P. A polypeptide model for toxic aberrant proteins induced by aminoglycoside antibiotics. *PLOS ONE* **17**, e0258794 (2022).

53. Bock, A. S. *et al.* A nonstop variant in REEP1 causes peripheral neuropathy by unmasking a 3′UTR-encoded, aggregation-inducing motif. *Human Mutation* **39**, 193–196 (2018).

54. Schmitz, U. *et al.* Intron retention enhances gene regulatory complexity in vertebrates. *Genome Biology* **18**, 216 (2017).

55. Braunschweig, U. *et al.* Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research* **24**, 1774–1786 (2014).

56. Gontijo, A. M., Miguela, V., Whiting, M. F., Woodruff, R. C. & Dominguez, M. Intron retention in the Drosophila melanogaster Rieske iron sulphur protein gene generated a new protein. *Nature Communications* **2**, 323 (2011).

57. Wong, J. J. -L., Au, A. Y. M., Ritchie, W. & Rasko, J. E. J. Intron retention in mRNA: No longer nonsense. *BioEssays* **38**, 41–49 (2016).

58. Gardner, L. B. Hypoxic inhibition of nonsense-mediated RNA decay regulates gene expression and the integrated stress response. *Molecular and cellular biology* **28**, 3729–41 (2008).

59. Popp, M. W. & Maquat, L. E. Nonsense-mediated mRNA Decay and Cancer. *Current Opinion in Genetics & Development* **48**, 44–50 (2018).

60. Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal* **33**, 981–993 (2014).

61. Nobuta, R. *et al.* eIF4G-driven translation initiation of downstream ORFs in mammalian cells. *Nucleic Acids Research* **48**, 10441–10455 (2020).

62. Dodbele, S. & Wilusz, J. E. Ending on a high note: Downstream ORFs enhance mRNA translational output. *The EMBO Journal* **39**, e105959 (2020).

63. Wu, Q. *et al.* Translation of small downstream ORFs enhances translation of canonical main open reading frames. *The EMBO Journal* **39**, e104763 (2020).

64. Chen, B., Retzlaff, M., Roos, T. & Frydman, J. Cellular Strategies of Protein Quality Control. *Cold Spring Harbor Perspectives in Biology* **3**, a004374 (2011).

65. Liberek, K., Lewandowska, A. & Ziętkiewicz, S. Chaperones in control of protein disaggregation. *The EMBO Journal* **27**, 328–335 (2008).

66. Hartl, F. U. & Hayer-Hartl, M. Molecular Chaperones in the Cytosol: from Nascent Chain to Folded Protein. *Science* **295**, 1852–1858 (2002).

67. Tartaglia, G. G., Pechmann, S., Dobson, C. M. & Vendruscolo, M. Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends in Biochemical Sciences* **32**, 204–206 (2007).

68. Haslbeck, M., Franzmann, T., Weinfurtner, D. & Buchner, J. Some like it hot: the structure and function of small heat-shock proteins. *Nature Structural & Molecular Biology* **12**, 842–846 (2005).

69. Ciechanover, A. The ubiquitin–proteasome pathway: on protein death and cell life. *The EMBO Journal* **17**, 7151–7160 (1998).

70. Finkbeiner, S. The Autophagy Lysosomal Pathway and Neurodegeneration. *Cold Spring Harbor Perspectives in Biology* **12**, a033993 (2019).

71. Harries, L. W. *et al.* Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing. *Aging Cell* **10**, 868–878 (2011).

72. Wang, L. *et al.* SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia. *The New England Journal of Medicine* **365**, 2497–2506 (2011).

73. Lo, A., McSharry, M. & Berger, A. Oncogenic KRAS alters splicing factor phosphorylation and alternative splicing in lung cancer. *bioRxiv* 2022.05.20.492866 (2022) doi:10.1101/2022.05.20.492866.

74. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Science Translational Medicine* **10**, (2018).

75. Smart, A. C. *et al.* Intron retention is a source of neoepitopes in cancer. *Nature biotechnology* **36**, 1056–1058 (2018).

76. Xiang, R. *et al.* Increased expression of peptides from non-coding genes in cancer proteomics datasets suggests potential tumor neoantigens. *Communications Biology* **4**, 496 (2021).

77. Dong, C. *et al.* Intron retention-induced neoantigen load correlates with unfavorable prognosis in multiple myeloma. *Oncogene* **40**, 6130–6138 (2021).

78. Mignogna, M. L. *et al.* Clinical characterization of a novel RAB39B nonstop mutation in a family with ASD and severe ID causing RAB39B downregulation and study of a Rab39b knock down mouse model. *Human molecular genetics* **31**, 1389–1406 (2022).

79. Nasiri, S., Talebi, F., Asl, J. M. & Mardasi, F. G. Identification of a Novel Non-Stop Mutation in PDE6C Gene in an Iranian Family With Con-Rod Dystrophy. *ACTA MEDICA IRANICA* (2020) doi:10.18502/acta.v58i6.4059.

80. Sun, J. *et al.* Functional analysis of a nonstop mutation in MITF gene identified in a patient with Waardenburg syndrome type 2. *Journal of Human Genetics* **62**, 703–709 (2017).

81. Abe, S. *et al.* Identification of CRYM as a Candidate Responsible for Nonsyndromic Deafness, through cDNA Microarray Analysis of Human Cochlear and Vestibular Tissues * * Nucleotide sequence data reported herein are available in the DDBJ/EMBL/GenBank databases; for details, see the Electronic-Database Information section of this article. *The American Journal of Human Genetics* **72**, 73–82 (2003).

82. Yao, W., Yin, T., Tambini, M. D. & D'Adamio, L. The Familial dementia gene ITM2b/BRI2 facilitates glutamate transmission via both presynaptic and postsynaptic mechanisms. *Scientific Reports* **9**, 4862 (2019).

83. Sylla, B. S. *et al.* The X-linked lymphoproliferative syndrome gene product SH2D1A associates with p62dok (Dok1) and activates NF-κB. *Proceedings of the National Academy of Sciences* **97**, 7470–7475 (2000).

84. Leimkühler, S. *et al.* Ten novel mutations in the molybdenum cofactor genes MOCS1 and MOCS2 and in vitro characterization of a MOCS2 mutation that abolishes the binding ability of molybdopterin synthase. *Human Genetics* **117**, 565–570 (2005).

85. Mills, P. B. *et al.* Neonatal epileptic encephalopathy caused by mutations in the PNPO gene encoding pyridox(am)ine 5′-phosphate oxidase. *Human Molecular Genetics* **14**, 1077–1086 (2005).

86. Binder, G. *et al.* SHOX Haploinsufficiency and Leri-Weill Dyschondrosteosis: Prevalence and Growth Failure in Relation to Mutation, Sex, and Degree of Wrist Deformity. *The Journal of Clinical Endocrinology & Metabolism* **89**, 4403–4408 (2004).

87. Lee, Y. J. & Rio, D. C. Analysis of altered pre-mRNA splicing patterns caused by a mutation in the RNA binding protein hnRNPA1 linked to amyotrophic lateral sclerosis. *bioRxiv* 2022.02.03.479052 (2022) doi:10.1101/2022.02.03.479052.

88. Tollervey, J. R. *et al.* Analysis of alternative splicing associated with aging and neurodegeneration in the human brain. *Genome Research* **21**, 1572–1582 (2011).

89. Tan, Q. *et al.* Extensive cryptic splicing upon loss of RBM17 and TDP43 in neurodegeneration models. *Human Molecular Genetics* **25**, ddw337 (2016).

90. Ong, C.-T. & Adusumalli, S. Increased intron retention is linked to Alzheimer's disease. *Neural Regeneration Research* **15**, 259–260 (2019).

91. Raj, T. *et al.* Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nature Genetics* **50**, 1584–1592 (2018).

92. Vance, C. *et al.* Mutations in FUS, an RNA Processing Protein, Cause Familial Amyotrophic Lateral Sclerosis Type 6. *Science* **323**, 1208–1211 (2009).

93. Zhang, Z. *et al.* SMN Deficiency Causes Tissue-Specific Perturbations in the Repertoire of snRNAs and Widespread Defects in Splicing. *Cell* **133**, 585–600 (2008).

94. Ling, J. P., Pletnikova, O., Troncoso, J. C. & Wong, P. C. TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. *Science* **349**, 650–655 (2015).

95. Lindeboom, R. G. H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nature genetics* **48**, 1112–1118 (2016).

96. Koren, I. *et al.* The Eukaryotic Proteome Is Shaped by E3 Ubiquitin Ligases Targeting C-Terminal Degrons. *Cell* **173**, 1622-1635.e14 (2018).

97. Wu, X. & Bartel, D. P. kpLogo: positional k-mer analysis reveals hidden specificity in biological sequences. *Nucleic Acids Research* **45**, W534–W538 (2017).

98. Lin, H.-C. *et al.* C-Terminal End-Directed Protein Elimination by CRL2 Ubiquitin Ligases. *Molecular Cell* **70**, 602-613.e3 (2018).

99. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science* **343**, 80–84 (2014).

100. Li, W. & Liu, S. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biology* (2014).

101. Hessa, T. *et al.* Protein targeting and degradation are coupled for elimination of mislocalized proteins. *Nature* **475**, 394–397 (2011).

102. Leznicki, P. & High, S. SGTA antagonizes BAG6-mediated protein triage. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 19214–9 (2012).

103. Leznicki, P. & High, S. SGTA associates with nascent membrane protein precursors. *EMBO reports* **21**, e48835 (2020).

104. Mariappan, M. *et al.* A Ribosome-Associating Factor Chaperones Tail-Anchored Membrane Proteins. *Nature* **466**, 1120–1124 (2010).

105. Shao, S., Rodrigo-Brenni, M. C., Kivlen, M. H. & Hegde, R. S. Mechanistic basis for a molecular triage reaction. *Science* **355**, 298–302 (2017).

106. Wunderley, L., Leznicki, P., Payapilly, A. & High, S. SGTA regulates the cytosolic quality control of hydrophobic substrates. *Journal of Cell Science* **127**, 4728–4739 (2014).

107. Mock, J.-Y. *et al.* Bag6 complex contains a minimal tail-anchor–targeting module and a mock BAG domain. *Proceedings of the National Academy of Sciences* **112**, 106–111 (2015).

108. Xu, Y., Cai, M., Yang, Y., Huang, L. & Ye, Y. SGTA Recognizes a Noncanonical Ubiquitin-like Domain in the Bag6-Ubl4A-Trc35 Complex to Promote Endoplasmic Reticulum-Associated Degradation. *Cell Reports* **2**, 1633–1644 (2012).

109. Hu, X. *et al.* RNF126-Mediated Reubiquitination Is Required for Proteasomal Degradation of p97-Extracted Membrane Proteins. *Molecular Cell* **79**, 320-331.e9 (2020).

110. Rodrigo-Brenni, M. C., Gutierrez, E. & Hegde, R. S. Cytosolic quality control of mislocalized proteins requires RNF126 recruitment to Bag6. *Molecular cell* **55**, 227–37 (2014).

111. Zhang, Y. *et al.* Cotranslational Intersection between the SRP and GET Targeting Pathways to the Endoplasmic Reticulum of Saccharomyces cerevisiae. *Molecular and cellular biology* **36**, 2374–83 (2016).

112. Minami, R. *et al.* BAG-6 is essential for selective elimination of defective proteasomal substrates. *Journal of Cell Biology* **190**, 637–650 (2010).

113. Carvunis, A.-R. *et al.* Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).

114. Vakirlis, N. *et al.* De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nature Communications* **11**, 781 (2020).

115. Senís, E. *et al.* TUNAR lncRNA Encodes a Microprotein that Regulates Neural Differentiation and Neurite Formation by Modulating Calcium Dynamics. *Frontiers in Cell and Developmental Biology* **9**, 747667 (2021).

116. Li, M. *et al.* A putative long noncoding RNA-encoded micropeptide maintains cellular homeostasis in pancreatic β-cells. *bioRxiv* 2020.05.12.091728 (2020) doi:10.1101/2020.05.12.091728.

117. Wang, L. *et al.* The micropeptide LEMP plays an evolutionarily conserved role in myogenesis. *Cell Death & Disease* **11**, 357 (2020).

118. Bhatta, A. *et al.* A Mitochondrial Micropeptide Is Required for Activation of the Nlrp3 Inflammasome. *The Journal of Immunology* **204**, 428–437 (2019).

119. Zhang, C. *et al.* Micropeptide PACMP inhibition elicits synthetic lethal effects by decreasing CtIP and poly(ADP-ribosyl)ation. *Molecular Cell* **82**, 1297-1312.e8 (2022).

120. Dangelmaier, E. A. *et al.* An Evolutionarily Conserved AU-Rich Element in the 3' Untranslated Region of a Transcript Misannotated as a Long Noncoding RNA Regulates RNA Stability. *Molecular and Cellular Biology* **42**, e00505-21 (2022).

121. Anderson, D. M. *et al.* A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* **160**, 595–606 (2015).

122. Nelson, B. R. *et al.* A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* **351**, 271–275 (2016).

123. Bi, P. *et al.* Control of muscle formation by the fusogenic micropeptide myomixer. *Science* **356**, 323–327 (2017).

124. Makarewich, C. A. *et al.* MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid β-Oxidation. *Cell reports* **23**, 3701–3709 (2018).

125. Anderson, D. M. *et al.* Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. *Science Signaling* **9**, ra119 (2016).

126. Matsumoto, A. *et al.* mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* **541**, 228–232 (2017).

127. Song, H. *et al.* HNF4A-AS1-encoded small peptide promotes self-renewal and aggressiveness of neuroblastoma stem cells via eEF1A1-repressed SMAD4 transactivation. *Oncogene* **41**, 2505–2519 (2022).

128. Polenkowski, M. *et al.* Identification of Novel Micropeptides Derived from Hepatocellular Carcinoma-Specific Long Noncoding RNA. *International Journal of Molecular Sciences* **23**, 58 (2021).

129. Li, M. *et al.* Micropeptide MIAC Inhibits HNSCC Progression by Interacting with Aquaporin 2. *Journal of the American Chemical Society* **142**, 6708–6716 (2020).

130. Huang, N. *et al.* An Upstream Open Reading Frame in Phosphatase and Tensin Homolog Encodes a Circuit Breaker of Lactate Metabolism. *Cell Metabolism* **33**, 128-144.e9 (2021).

131. Oishi, N. *et al.* XBP1 mitigates aminoglycoside-induced endoplasmic reticulum stress and neuronal cell death. *Cell Death & Disease* **6**, e1763–e1763 (2015).

132. Sang, Y. *et al.* Mitochondrial micropeptide STMP1 promotes G1/S transition by enhancing mitochondrial complex IV activity. *Molecular Therapy* (2022) doi:10.1016/j.ymthe.2022.04.012.

133. Lin, Y.-F. *et al.* A novel mitochondrial micropeptide MPM enhances mitochondrial respiratory activity and promotes myogenic differentiation. *Cell Death & Disease* **10**, 528 (2019).

134. Zhang, Y. E., Vibranovski, M. D., Landback, P., Marais, G. A. B. & Long, M. Chromosomal Redistribution of Male-Biased Genes in Mammalian Evolution with Two Bursts of Gene Gain on the X Chromosome. *PLoS Biology* **8**, e1000494 (2010).

135. Prilusky, J. & Bibi, E. Studying membrane proteins through the eyes of the genetic code revealed a strong uracil bias in their coding mRNAs. *Proceedings of the National Academy of Sciences* **106**, 6662–6666 (2009).

136. Wolfenden, R. V., Cullis, P. M. & Southgate, C. C. F. Water, Protein Folding, and the Genetic Code. *Science* **206**, 575–577 (1979).

137. Vavouri, T. & Lehner, B. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. *Genome Biology* **13**, R110–R110 (2012).

138. Wu, S. *et al.* A micropeptide XBP1SBM encoded by lncRNA promotes angiogenesis and metastasis of TNBC via XBP1s pathway. *Oncogene* **41**, 2163–2172 (2022).