Winter 12-15-2017

Mechanism of Gene Regulation by Coding PolyA Tracks

Laura Lea Arthur
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Molecular Genetics and Genomics

Dissertation Examination Committee:
Sergej Djuranovic, Chair
Nicholas Davidson
Tim Schedl
Jim Skeath
Matthew Walter

Mechanism of Gene Regulation by Coding PolyA Tracks

by
Laura Arthur

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2017
St. Louis, Missouri

Table of Contents

List of Figures iv
List of Tables vi
List of Abbreviations vii
Acknowledgments ix
Abstract xi
Chapter 1: Introduction 1
1.1 Poly(A) tail mediated regulation 2
1.2 Evidence for ribosomal stalling by polybasic sequences 6
1.3 Coding polyA sequences repress translation more than polybasic sequences 9
1.4 Perspectives 15
1.5 Figures 17
1.6 References 23
Chapter 2: Translational control by lysine-encoding A-rich sequences 27
Preface 27
2.1 Abstract 28
2.2 Introduction 28
2.3 Results 30
2.4 Conclusion 38
2.4 Materials and Methods 39
2.4.1 Experimental Protocols 39
2.4.2 Bioinformatic Analysis 42
2.5 Figures 45
2.6 Tables 86
2.7 References 133
Chapter 3: Rapid generation of hypomorphic mutations 135
Preface 135
3.1 Abstract 136
3.2 Introduction 136
3.3 Results 138

3.4 Discussion 152
3.5 Methods 156
3.6 Figures 168
3.7 References 181
Chapter 4: Coding polyA tracks induce mRNA surveillance pathways 186
4.1 Abstract 186
4.2 Introduction 186
4.2 Results 188
4.3 Discussion 193
4.4 Methods 194
4.5 Figures 197
4.6 References 206
Chapter 5: Conclusions and Future Directions 208
5.1 Summary 208
5.2 Future Directions 208

List of Figures

Figure 1.1 Normal and aberrant translation. 18
Figure 1.2 Potential outcomes of programmed ribosome frameshifting on coding polyA tracks. 20
..... 21
Figure 1.3 Model of effects of synonymous mutations within coding polyA tracks 22
Figure 2.1. Effects of different lysine codons on mCherry reporter expression and mRNA stability. 46
Figure 2.2. The effect of codon usage in polylysine tracks on translation and protein levels. 49
Figure 2.3. Native poly(A) tracks control reporter mRNA and protein levels. 52
Figure 2.4. The effect of synonymous mutations in poly(A) tracks of human genes. 55
Figure 2.5. Putative mechanisms through which poly(A) tracks exert their function. 58
Supplemental Figure 2.1. 59
Supplemental Figure 2.2 60
Supplemental Figure 2.3 61
Supplemental Figure 2.4 63
Supplemental Figure 2.6. 67
Supplemental Figure 2.7. 69
Supplemental Figure 2.8. 70
Supplemental Figure 2.9. 71
Supplemental Figure 2.10. 72
Supplemental Figure 2.11. 74
Supplemental Figure 2.12. 75
Supplemental Figure 2.13. 76
Supplemental Figure 2.14. 77
Supplemental Figure 2.15. 82
Supplemental Figure 2.16. 83
Supplemental Figure 2.17. 84
Supplemental Figure 2.18. 85
Figure 3.1. Design and mechanism of polyA track tag regulated gene expression. 169
Figure 3.2. Regulation of reporter gene by polyA tracks in the single cell prokaryotic and eukaryotic organisms. 171
Figure 3.3. Regulation of reporter gene by polyA tracks in the eukaryotic tissue cultures. 174
Figure 3.4. PolyA tracks regulate mCherry reporter gene expression in different organs of *D. melanogaster*. 176
Figure 3.5. PolyA tracks regulate mCherry reporter expression independently of the promoter strength. 178
Figure 3.6. Regulation of drug resistance and metabolic survival by insertion of polyA track tags in genes from *E. coli* and *S. cerviseae*. 180
Figure 4.1. Effect of mutation within polyA tracks of endogenous genes on expression. 198

Figure 4.2. Effect of lysine incorporation in the nascent peptide on ribosomal frameshifting... 200
Figure 4.4. Rescue of polyA track reporter by knockdown of NGD and NMD factors. 203
Figure 4.5. Production of alternative frameshifted protein. 205

List of Tables

Supplemental Table 2.1. Statistics of occurrences of transcripts containing polyA tracks in different organisms. 86
Supplemental Table 2.2. Gene Ontology terms for polyA track genes. 88
Supplemental Table 2.3. PolyA track genes that are potential NMD targets. 96
Supplemental Table 2.4. Peptides arising from possible frame-shifting on polyA tracks..... 116
Supplemental Table 2.5. Table of genes with mutations within polyA region reported in COSMIC database. 132

List of Abbreviations

CAM, chloramphenicol

CNS, central nervous system

CRISPR, clustered regularly interspersed short palindromic repeats

DNA, deoxyribonucleic acid

FS, frameshift

GFP, green fluorescence protein

HDF, human dermal fibroblast

mChr, mCherry fluorescence protein

mRNA, messenger RNA

NGD, no-go decay

NSD, nonstop decay

ORF, open reading frame

PCR, polymerase chain reaction

PTC, premature termination codon

PV, proventriculus

RNA, ribonucleic acid

RQC, ribosome-associated quality control

SG, salivary gland

STR, short tandem repeats

TALEN, transcription activator-like effector nuclease

tRNA, transfer RNA

UTR, untranslated region

WT, wild type

YFP, yellow fluorescence protein

Acknowledgments

Advances in scientific knowledge are not accomplished in isolation. I am grateful for the professional and personal guidance I have received from the Washington University community as I endeavored to make my first contributions to Biology.

The Djuranovic lab was a perfect fit. My thesis advisor, Sergej Djuranovic, was and continues to be an exceptional source scientific knowledge. More importantly, his enthusiasm for reseach and unwavering positive outlook insprired and calmed me in times of personal doubt. My labmate, Kyle, was more like a second PI for me. I could ask him anything and get a well-informed, thorough answer.

Laura Arthur

Washington University in St. Louis

December 2017

Dedicated to my nieces and nephews

ABSTRACT OF THE DISSERTATION

Mechanism of Gene Regulation by Coding PolyA Tracks

Laura Arthur

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics and Genomics

Washington University in St. Louis, 2017

Dr. Sergej Djuranovic, Chair

Regulation of gene expression is essential for cellular development and survival. The great variety and complexity of regulatory mechanisms underscores this fact. Messenger RNA stability and translational efficiency are often key determinants of gene expression. mRNA surveillance pathways, discovered for their role in degradation of aberrant mRNA, are now known to be instrumental in the regulation of physiologically correct mRNA stability. Thus, the study of cis elements in a transcript that can induce mRNA surveillance pathways has become an area of particular interest.

Here I report on the mechanism of gene regulation by coding polyA tracks, defined as a sequence of at least 12 consecutive nucleotides in which all but one are adenosines. When a polyA track is present in the open reading frame of an mRNA, the translating ribosome stalls and frameshifts due to interactions with the polyA sequence. These events lead to degradation of the transcripts by the mRNA surveillance pathways no-go decay and nonsense mediated decay. As a consequence of the polyA sequence, less protein is expressed from these transcripts. Approximately 2% of genes in most genomes, including humans, contain a polyA track and are potentially regulated by this mechanism.

Chapter 1: Introduction

All organisms must be able to regulate gene expression and abundance to develop, reproduce, and respond to their environments. In the flow of information from DNA to RNA to protein, regulatory mechanisms have evolved to increase or decrease the rate of essentially every biochemical reaction. As the intermediate molecule in this process, the abundance and translational efficiency of mRNA plays a central role in the final abundance of protein. Changes to translation result in more rapid changes in protein expression than changes to transcription. Consequently, the life cycle of mRNA and its translational efficiency is under strict regulation.

Much of the research into these post-transcriptional regulation mechanisms focuses on modulation of translation initiation and mRNA stability by *cis* elements in the 5' and 3' untranslated regions (UTRs) and the *trans* factors that interact with them. For example, the 5' terminal oligopyrimidine (TOP) motif consists of 4-15 pyrimidine nucleotides near the transcription start site that is bound by La-related protein 1 (LARP1) to stimulate translation in response to mammalian target of rapamycin (mTOR) to promote cell growth[1]. TOP motifs are found in most ribosomal proteins and may play a role in translational upregulation of these genes during the M phase of mitotic cell division[2,3]. Similarly, the PUF family of proteins bind specific sequences in the 3'UTRs of many transcripts to decrease expression of those genes[4]. See references [5] and [6] for a more in depth review of 5' and 3' UTR mediated regulation, respectively. While control of translation initiation and modulation of mRNA stability through binding of *trans* factors play an important role in specifying the amount of translated protein products, the contribution of translation elongation in overall gene regulation has been largely overlooked. Over the last several

years, however, regulation of the rate and fidelity of elongation has emerged as one of the key determinants of mRNA stability and translation efficiency and thus protein abundance[7].

Here, we discuss a novel translation elongation regulatory mechanism mediated by short stretches of consecutive adenosine residues in the open reading frame (ORF) of an mRNA, termed polyA tracks. Coding polyA tracks are *cis*-acting elements that cause ribosomal stalling and programmed ribosomal frameshifting, decreasing mRNA stability and reducing protein synthesis. PolyA tracks are found endogenously in 1-2% of genes in most organisms, including humans, meaning a significant portion of the genome is potentially regulated by this mechanism[8]. The similarities of coding polyA tracks to both the poly(A)-tail and poly-basic amino acid tracks warrant a brief discussion and comparison of these mechanisms as well.

1.1 Poly(A) tail mediated regulation

1.1.1 The poly(A) tail is a key regulator of gene expression

The most well-known polyadenylate sequence, the polyA tail, plays an essential role in the regulation of mRNA stability and translation. Nearly all eukaryotic mRNAs undergo 3' end cleavage and polyadenylation to generate the polyA tail. The poly(A) tail globally stimulates translation by binding cystolic poly(A)-binding protein 1 (PABP1c) which interacts with eukaryotic initiation factor 4G (eIF4G) to form a pseudo-ring structure and promote efficient translation initiation[9,10].

Poly(A) tail length can be dynamically regulated by cytoplasmic deadenylases and polyadenylases in a developmental, tissue specific, or cell cycle dependent manner to influence translation efficiency. Many groups have shown that during early embryogenesis in *Xenopus* and *Drosophila* oocytes, translationally silent maternal mRNA with short poly(A) tails are activated

by extension of their polyA tails[11]. More recently, work in HeLa cells showed that the poly(A) tail on a subset of transcripts is shortened or extended during the M phase of the mitotic cell cycle. Transcripts with poly(A) tails shortened to fewer than 20 nt were less efficiently translated, possibly because of the reduced ability to bind PABPC1[3]. Analysis of transcripts with poly(A) tail lengths shorter than 30 nt consistently find that they are translationally silenced and destabilized[3,12,13], however, studies on the translation of transcripts with poly(A) tails longer than 30 nt have come to conflicting conclusions. Initial studies have found a positive correlation of tail length to translational efficiency generally[14] only in embryonic cells[3,15,16]. More recently, however, a negative correlation of tail length to translation efficiency was found in *C. elegans*[13]. These discrepancies may be due to different experimental design, or reflect a complexity of regulation that is cell type and developmentally dependent.

In yeast, transcripts that lack a poly(A) tail, such as those cleaved by an endonuclease, are rapidly degraded by the cytoplasmic exosome[17]. Additionally, transcripts lacking a polyA tail, with or without a stop codon, are translated with lower efficiency than those that are polyadenylated[17]. These observations are consistent with the role of the poly(A) tail in stimulating translation and indicates that the lack of poly(A) tail signals to the cell that the transcript is aberrant.

Alternative polyadenylation generates transcripts with alternative 3' coding sequences or 3' untranslated regions (UTRs)[18]. This regulated process can influence the expression of the transcript by including or excluding protein binding sites in the alternative 3' UTR. Yet another way that alternative polyadenylation can influence expression is by placing the poly(A) tail before the canonical stop codon, effectively creating a transcript with no in-frame stop codon, also known as a nonstop transcript.

1.1.2 Translation of the poly(A) tail induces mRNA and nascent protein decay

Typically, the ribosome does not encounter the polyA tail since it is found downstream of the termination codon. Aberrant or premature polyadenylation (APA) or mutations that eliminate the stop codon, however, allow the ribosome to continue translation into the polyA tail. In fact, premature polyadenylation appears to occur in approximately 1% of yeast and human transcripts[19,20]. These nonstop transcripts are less stable and produce less protein than properly polyadenylated transcripts[19,21], but the molecular mechanisms underlying this instability and the role, if any, the polyA tail plays in it was unclear until recently. More recent studies have identified a role for ribosome associated quality control in eliminating those sequences. Furthermore, these studies revealed new mechanisms of ribosome mediated gene regulation, including regulation by short, coding polyA tracks.

Eukaryotic mRNAs lacking a stop codon are rapidly decayed by the translation dependent mRNA surveillance pathway nonstop decay (NSD). This pathway was initially characterized in yeast and requires the exome complex of 3'-to5' exoribonucleases involved in general cytoplasmic mRNA decay[19,21]. However it is distinct from general mRNA decay because it does not require deadenylation prior to degradation and involves endonuclease activity of the exosome, rather than just the exonuclease activity[19,21]. A cofactor of the exosome, Ski7p, recognizes ribosomes stalled on nonstop transcripts. This function requires the GTPase C-terminal domain of Ski7 in a manner homologous to the GTPase domain of eRF3 recognizing the stop codon of a normal transcript[21]. Since Ski7 tightly binds to the exosome, it is recruited to transcripts recognized as nonstop[21]. The C-terminal domain of Ski7 is the only known factor that is specific to the NSD pathway. Ski7, however, is found only in some species of *Saccharomyces*. A homologous protein, HBS1, was found to function in the NSD pathway in other eukaryotes[22,23]. HBS1 protein, together with Dom34

(human Pelota), was also found to be one of the mediators of another mRNA surveillance pathway, termed No-go decay (NGD)[24].

NGD is an mRNA surveillance pathway that detects ribosomes stalled by mRNA or nascent protein features other than the poly(A) tail. Stable secondary structure in the mRNA, rare codons, and chemical damage to the mRNA have all been shown to induce the NGD pathway[24,25]. Stalled ribosomes are recognized by the HBS1:Dom34 complex which promotes ribosome subunit dissociation from the mRNA to resolve the stall[24,26]. The complete ribosome recycling is enabled by action of ABCE1 (yeast Rli) a highly conserved ABC-type ATPase[27,28]. Endonucleolytic cleavage of the stall-inducing mRNA by an unknown endonuclease and subsequent ribosome recycling are followed by further degradation of cleaved transcript using the canonical mRNA decay pathways with the exonuclease activity of exosome in the 3' to 5' direction and Xrn1 in the 5' to 3' direction[24].

Aberrant mRNA transcripts induce both NSD and NGD during translation while nascent peptides are synthesized. Like the destabilization of aberrant mRNA transcripts by NSD and NGD, the nascent protein must also be degraded to protect the cell from the potential harmful effects of aberrant or truncated proteins. The ribosome-associated quality control (RQC) complex, originally discovered and characterized in yeast, fulfills this role by targeting the nascent peptide to the proteasome for degradation[29–31]. The RQC consists of four known components - Ltn1p, Rqc1p, Rqc2p, and Cdc48p that act together to identify and target the stalled nascent peptide for degradation. Briefly, the Dom34:Hbs1 complex dissociates the large and small subunits of the ribosome without hydrolyzing the peptidyl-tRNA. The nascent peptide remains associated with the 60S subunit of the ribosome[26], creating a distinct substrate compared to normal translation termination in which eRF1 hydrolyses the peptidyl-tRNA and triggers its dissociation from the

60S subunit. Rqc1 recognizes the 60S associated with peptidyl-tRNA, and this complex promotes the association of the E3 ligase, Ltn1p (yeast homologue of mammalian Listerin), with the ribosome near the exit channel and stalled peptide. Ltn1p then ubiquitylates the stalled nascent peptide, targeting it to the CDC48 and subsequently to proteasome[32]. Activation of the RQC pathway removes the peptidyl-tRNA from the 60S subunit to complete ribosome recycling and insures the degradation of nascent peptides translated from aberrant, nonstop transcripts.

1.2 Evidence for ribosomal stalling by polybasic sequences

Translation of the poly(A) tail results in the synthesis of a poly-lysine sequence at the C-terminus of the nascent peptide. This poly-lysine tail was predicted to be the cause of ribosome stalling events observed on poly(A) tails. In fact, research into the poly(A) tail mediated activation of NSD and NGD, as well as studies on the RQC pathway, led to the discovery that clusters of basic amino acids, either lysine or arginine, can act as a stalling sequence even when encoded before the stop codon. It was experiments in yeast, prompted by the above question of non-stop transcripts, that revealed that polybasic sequences of greater than 10 residues in length in the coding sequence are sufficient to cause reduction in protein abundance[33]. In vitro studies, using rabbit reticulocyte lysates indicated that runs of positively charged lysine or arginine residues impede translation rates through electrostatic interactions of positively charged peptide with the negatively charged ribosome exit channel[34]. As such translation of the polybasic sequence are cause for ribosomal stalling even in the context outside of polyA tail translation targeting sequences found in endogenous genes in yeast[35] Thus, stretches of poly-basic residues within protein coding genes are yet another mechanism that controls protein abundance.

Research in several labs verified the regulatory importance of polybasic stretches. Brandman *et al.* identified putative RQC substrates with six or more basic amino acids in a stretch of 10. Notably, one of the putative RQC substrates identified in that study was Rqc1, which contains a well conserved polybasic region. Mutation analysis revealed that this polybasic region serves as an autoregulatory element. When Rqc1 expression is high, the RQC pathway is highly active and constrains expression of Rqc1 transcripts through its polybasic track[29]. These polybasic stretches appear to trigger both transcript and nascent peptide degradation consistent with stalling on the polyA tail: via activation of mRNA surveillance and the RQC.

Several groups recently identified and further characterized mammalian homologs of RQC factors. ZNF598 is a human RING domain protein that shares homology with Hel2, the yeast RING domain that promotes read-through of polybasic sequences[36]. Knockdown of ZNF598 in a terminal stalling assay increased read-through of a sequence of at least 20 AAA codons to produce full length reporter proteins[36,37]. The increase in full length protein rather than an increase in truncated protein fragments suggests that the ribosomes did not terminally stall in the absence of ZNF598. A similar experiment in ZNF598 knockout HEK cells finds consistent ZNF598-dependent repression of polyA reporters[38]. The N-terminal RING domain of ZNF598 acts as an E3 ligase that was shown to ubiquitylate RPS20, RPS10, and RPS3[37,38]. These ubiquitylation events are necessary to promote terminal stalling on consecutive AAA codons.

A genetic screen in *S. cerevisiae* revealed that mutations in ASC1, the yeast homolog of RACK1, increased expression of transcripts containing 12 consecutive arginine or lysine codons in a translation dependent manner.[39] RACK1 is a highly conserved component of the eukaryotic 40S subunit that acts as a scaffold protein in a wide variety of signal transduction pathways.[40] Knockdown of RACK1 in human cells also increased expression of transcripts with consecutive

AAA lysine codons.[37] RACK1 facilitates ubiquitylation of certain ribosomal proteins in response to translation of consecutive AAA codons. In a similar role, ZNF598 targets RPS3 and RPS20. Since double knockdown of RACK1 and ZNF598 did not have an additive effect, it seems that these proteins act in the same pathway to promote stalling on consecutive AAA codons. The discovery that knockdown of ZNF598 and RACK1 can promote read-through of consecutive lysine codons indicates that the ribosome does not inherently stall, but must be induced to stall by ZNF598-mediated and RACK1-assisted ubiquitylation of ribosomal proteins[36–38]. After the ribosome is ubiquitylated by ZNF598 and RACK1, it is terminally stalled and recognized by the Hbs1:Dom34 complex which splits the ribosomal subunits[41]. Nuclear export mediator factor (NEMF, mammalian homologue of Rqc2p), stabilizes Listerin, the mammalian homolog of yeast Ltn1p, which ubiquitylates the nascent peptide to mark it for degradation by the proteasome, completing the RQC pathway[42].

Note that characterization of RQC factors in mammalians cells relied on poly-lysine reporters encoded exclusively with AAA codons, since AAG lysine and poly arginine sequences did not cause efficient stalling in human cell culture. This distinction raises the possibility that stalling due to polybasic sequences in the nascent peptide does not affect mammalian ribosomes as strongly as yeast ribosomes and that consecutive AAA codons may be stalling the ribosome in a different manner. It is also possible that the nature of the reporters used in these studies accounts for the difference in stalling potentials. Both Hegde and Bennett's group used a dual fluorescence stalling system in which GFP and RFP are separated by two viral 2A peptide sequences flanking the stalling sequence[36,37].

1.3 Coding polyA sequences repress translation more than polybasic sequences

1.3.1 Codon usage in poly-lysine tracks show an underrepresentation of the AAA codon

Sequences of six or more consecutive polybasic amino acids, lysine and arginine, are underrepresented in the transcriptome of multiple organisms compared to runs of other amino acids, suggesting a selective pressure against polybasic amino acid sequences [43]. Further analysis of codon usage within poly-lysine peptides, however, reveals that there is a lower than expected frequency of AAA codons compared to AAG codons in runs of four consecutive lysine residues in the coding regions of genes[8]. Similar discrepancies in lysine codon usage patterns are present in over 150 prokaryotic and eukaryotic genomes that have been analyzed[44]. Interestingly, when a poly-lysine track does contain several consecutive AAA codons, orthologous proteins across species have comparable lysine codon usage for these tracks[8,45]. This apparent conservation and evolutionary selection of codon usage in such sequences suggests a functional significance for coding polyA tracks.

1.3.2 PolyA in the most efficient stalling sequence

Careful biochemical analysis of poly-lysine stalling sequences in both prokaryotic and eukaryotic systems revealed that consecutive AAA lysine codons exhibit a greater delay in ribosome movement than an equivalent number of AAG lysine codons[8,45,46]. Reporters with consecutive AAA codons produce significantly less protein than those with AAG, despite coding for the same polybasic amino acid sequence. This discordance in translation likely does not result from differences in tRNA abundance or binding affinity, as there are multiple cognate and

isodecoder tRNAs for lysine codons in higher organisms. *E. coli* uses a single tRNA for both lysine codons, a UUU* modified anti-codon for both AAA and AAG, but still show reduced translational efficiency for iterated AAA codons[47]. The presence of this phenomenon in *E. coli* similarly discounts the possibility that interactions with polyA bind protein (PABP) decrease translation efficiency of consecutive AAA codons, as PABP is not found in prokaryotes[48]. Instead, it argues for the possibility that the interactions between the ribosome and the mRNA sequence itself cause the difference in translational efficiency. The role of the lysine residues was examined by kinetic analysis of lysine incorporation in an *in vitro* translation system which showed that the rate of second lysine addition in an iterated sequence of either AAA or AAG codons is slower than typical elongation rates of heteropolymeric sequences. Consecutive AAA codons, however, slowed incorporation to a greater extent than AAG codons[45]. Together these studies suggest that the polyA sequence itself contributes to decreased protein expression, but whether the lysine residues translated from the mRNA sequence facilitates or enables this decrease is unknown.

There are conflicting reports about the effect of polybasic stalling in human tissue culture, but consistent evidence suggests that polyA tracks are a potent negative regulator of expression. In two recent reports, HEK293 cell line ribosomes are resistant to stalling on many of the stalling sequences characterized in yeast, including poly arginine and select stem loop structures[36,37]. Even poly-lysine reporters encoded by as many as 20 AAG codons were reported to have little to no decrease in protein expression compared to no-insert controls[36,37]. Early investigations of the difference between AAG vs AAA induced stalling in human dermal fibroblasts found that 12 iterated AAG codons were sufficient to induce some mRNA instability and decreased protein expression in tissue culture[8], consistent with several studies of polybasic sequences in yeast[32,33,35,49] . In all studies, however, the coding polyA sequence mediates more potent

translational regulation across many species with as few as four consecutive AAA codons reducing protein expression by 50% of control constructs[8,36,37,46]. These studies indicate that the stalling effect is specific for lysine residues encoded by polyA sequences rather than for poly-lysine stretches in general. Four AAG codons had no effect on protein expression in any reports. Aggregate ribosome profiling data show that in a sequence of four consecutive lysine codons, with three or more AAA codons, one can see increased ribosomal occupancy, indicative of increased ribosomal stalling. The same effect is not seen with four lysine tracks with fewer than three AAA codons[8]. The polyA induced stalling likely triggers activation of mRNA surveillance and the RQC pathway to decrease mRNA stability and protein expression in a manner similar to polybasic induced stalling[8].

1.3.3 PolyA sequences cause frameshifting in addition to stallin

Investigations of the stalling potential of consecutive AAA codons revealed that the ribosome can also frameshift while elongating over the polyA track[8,45]. In an *in vitro* translation system, as few as six consecutive adenosine residues were sufficient to cause a change in translation frame and the addition of an extra lysine residue in the nascent peptide when programed reactions were missing downstream charged-tRNAs or termination factors associated with stop codons [45]. When all factors are present, truncated products that result from frameshifting were detected *in vitro* with a minimum of three consecutive AAA codons[45]. Ribosomal frameshifting was also seen with reporters expressing human genes with endogenous polyA tracks[8]. This behavior is surprising because frameshifting typically requires the presence of a downstream stimulatory element such as a pseudoknot or stem loop structures that slow ribosome movement on mRNAs[50]. It is possible that slowed incorporation of lysine residues in consecutive tracks may

stall the ribosome sufficiently to allow frameshifting to occur, substituting for physical stall introduced by pseudoknot structure typically seen in programed frameshifting.

When the ribosome frameshifts on a polyA track, it will continue the elongation cycle out of frame and likely encounter a premature termination codon downstream. Like other programmed ribosomal frameshifting events, this can lead to the activation of nonsense mediated decay (NMD) which will result in degradation of the transcript and nascent peptide. Knockout of the major NMD factor, Upf1, in *S. cerevisiae* partially restores expression of reporters containing 12 AAA codons, demonstrating that polyA track transcripts are degraded by the NMD pathway in yeast[45]. It's worth mentioning that mRNA levels of the reporter with polyA track in the second exon of beta-globin gene are notably similar to insertion of a stop codon at the same position[8]. The frameshifted or stalled product in this case would be efficiently degraded and was not observed. No frameshifting has been reported for other polybasic stall sequences. In fact, it was shown that ribosomes do not normally frameshift on rare CGA codon repeats in wild type yeast even though they cause significant ribosomal stalling[51]. This characteristic is perhaps the reason that polyA tracks have a much stronger effect on protein expression and mRNA stability. PolyA tracks probably stall and elicit NGD to the same extent as other polybasic tracks, but then undergo additional destabilization by NMD because of frameshifting events.

RACK1, the ribosome associated protein discussed above for its role in stalling on AAA encoded poly-lysine tracks, is also known to promote frame maintenance during translation elongation. Knockdown of the yeast homolog of RACK1, Asc1, resulted in frameshifting on CGA codon repeats which stall in wild type yeast strains but do not frameshift[52]. If human RACK1 also functions to maintain the coding frame, then reduced levels would lead to more frameshifting, and thus, more degradation by NMD. However, as discussed above, it also allows more read through

of the stall sequence and less degradation by NGD or NSD. Given these contradictory predications, RACK1 likely has complex effects on the expression of transcripts with polyA tracks.

1.3.4 Endogenous coding polyA sequences regulate gene expression

Bioinformatic analysis of the human genome revealed that more than 450 genes contain a polyA track – defined as a sequence of 12 nucleotides in which at least 11 are adenosine[8]. A similar proportion of polyA track genes were found in other vertebrate genomes[8]. Since the definition of polyA tracks depends on the nucleotide sequence and not the amino acid sequence, it does not necessarily encode four consecutive lysine residues. Additionally, the interrupting nucleotide in the sequence can fall anywhere among the adenosines. It is not known how the exact composition of the polyA track in endogenous genes affects stalling and frameshifting efficiencies given the possibilities for mRNA modifications and alternative splicing of these genes. Insertion of gene-derived polyA tracks into a reporter significantly decrease protein expression and mRNA stability compared to both no insertion and consecutive AAG controls[8]. These experiments demonstrate that polyA tracks play a role in controlling gene expression by attenuating translation efficiency and likely promoting degradation of the transcripts by NGD and NMD.

In addition to degradation by one of the mRNA surveillance pathways, there is another possible outcome for transcripts harboring a polyA track. If a polyA track induces frameshifting in the last exon of a transcript, it is likely that the NMD pathway will be inefficient, since it is known that PTCs less than 50 nt upstream of the last exon-exon junction do not initiate strong NMD response[53]. Instead, the translation cycle could terminate normally on the out of frame stop codon. The protein synthesized in this scenario would have an altered, frameshifted amino acid sequence after the polyA track. Depending on position of the out of frame stop codon, the novel

protein would have either a C-terminal extension or truncation compared to translation in annotated 0 frames. No such proteins have been identified *in vivo* yet, however, bioinformatics analyses suggests that some polyA track genes may escape NMD and result in a novel C-termini of considerable length[8]. Altered C-termini could add regulatory or functional domains to known proteins, adding to the complexity of the proteome. It is expected that these hypothetical proteins would be produced in low amounts, however, and hard to distinguish from their annotated 0 frame counterparts due to the minimal sequence specificity that comes only from the unique C-termini.

1.1.2 Synonymous mutations in polyA tracks significantly impact gene expression

The discrepancy between translational efficiency of consecutive AAA and AAG codons raises the possibility of changes in gene expression due to synonymous mutations. Synonymous mutations alter a gene sequence without changing the sequence of the encoded protein, therefore, typically leaving its function unchanged. For this reason, synonymous mutations are often considered inconsequential when analyzing sequencing data for cancer or other genetic diseases to identify causal variants. *Cis* regulatory gene sequences, however, are sensitive to synonymous mutations. Mutations in exonic motifs that define splice sites are a well-recognized example of synonymous variants causing changes to protein expression and function[54]. Similarly, synonymous mutations that interrupt a polyA track can significantly affect gene expression without affecting the protein sequence. This effect has been demonstrated with both reporter genes with artificial polyA tracks inserted and with genes that endogenously contain a polyA track. When the lysine codons are mutated from AAA to AAG codons to decrease the length of the polyA track, protein expression increases significantly. The converse is true for AAG to AAA mutations which increase the length of the polyA track[8]. In addition to changing gene dosage, a synonymous mutation that increased the length of the polyA track in the *ZCRB1* gene resulted in the production of a potential

frameshifted protein[8], a result which supports the prediction of proteins with novel c-terminals produced from polyA track genes.

1.4 Perspectives

The discovery of coding polyA track mediated gene regulation adds to the growing list of post-transcriptional gene regulatory mechanisms. It has become clear that ribosome mediated translation control is a pathway commonly exploited by the cell to selectively modulate expression of genes. This is done through engagement of mRNAs, tRNAs, ribosomes, nascent polypeptide chain and pathways that control both protein and mRNA quality. PolyA tracks are another mechanism that regulates translation of physiologically correct mRNA at the step of elongation by eliciting responses from both mRNA surveillance pathways and RQC. Nonetheless, open questions remain about the molecular mechanism and biological significance of polyA induced stalling and frameshifting.

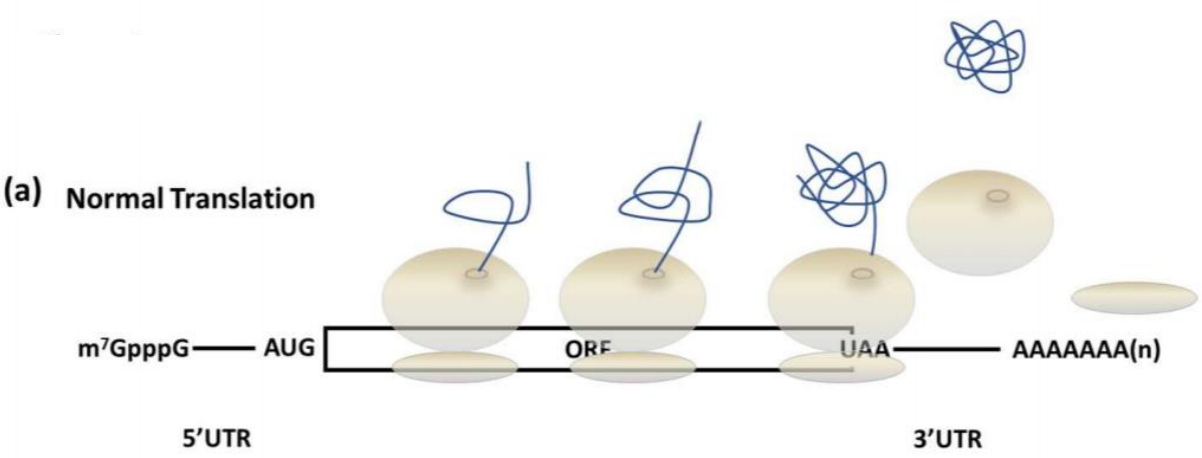
ZNF598 and RACK1 are the most upstream factors shown to act in the ribosome stalling and rescue pathway, meaning their recruitment and ubiquitylation activity may promote terminal stalling on sequences that would otherwise transiently pause the ribosome. If so, regulation of the abundance of these factors would change the stalling efficiency of polyA tracks, allowing for dynamic regulation of these genes. RACK1 is already known to function in cell signaling pathways and has increased translational efficiency during the M phase of the mitotic cell cycle when translation is globally suppressed[3].

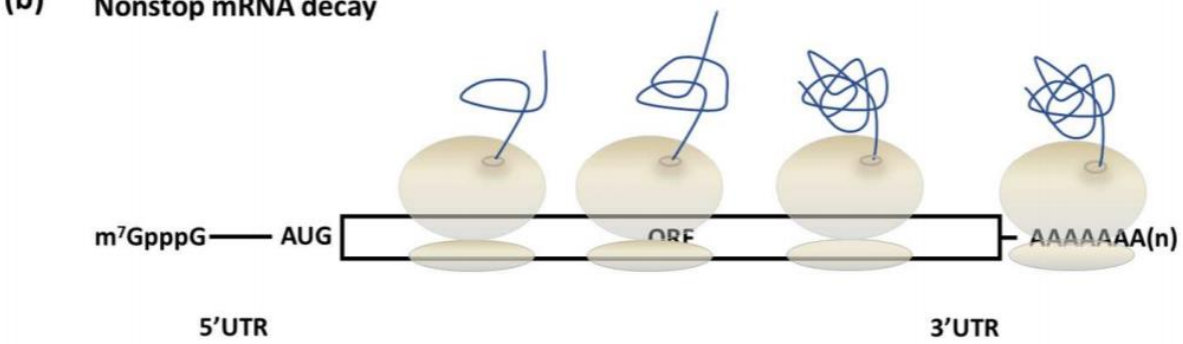
The possibility of alternative proteins rising from frameshifts on polyA tracks,as well as contribution of synonymous mutations in cellular health, are open questions with implication for human health. The frameshifted products as well as synonymous mutations in polyA tracks could

be cell type specific as many of the endogenous polyA track genes are subject to alternative splicing, often excluding polyA track containing exons. This would further add to complexity of gene regulation by polyA tracks resulting in deferential gene expression across cell types.

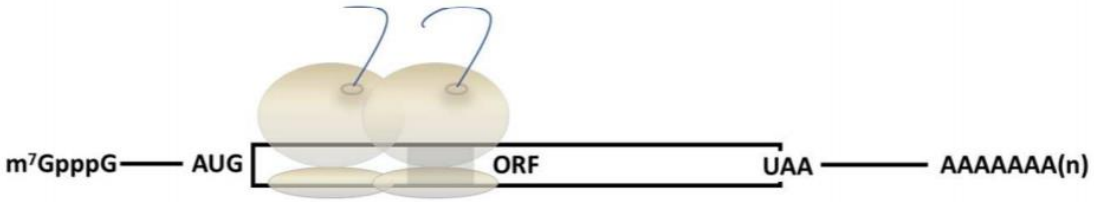
Finally, a majority of the sequenced genomes contain 1-2% of polyA containing transcripts[44]. This distribution represents a small, but significant, number of genes potentially regulated by this mechanism. However, organisms with very AT rich genomes, such as *Plasmodium* species with more than 60% of total transcripts with polyA tracks, may have evolved different mechanisms that enable productive and correct synthesis of lysine-rich peptides from long polyA tracks. If so, elucidating the differences between these ribosomes may provide insight into ribosome evolution and potential targets for therapeutics.

1.5 Figures





(c)



(d)

Stall inducing	RACK1 (Asc1), ZNF598 (Hel2)
Ribosome release and recycling	Hbs1, Pelota (Dom34), ABCE1 (Rli1)
Nascent protein ubiquitination	Listerin (Ltn1), NEMF (Rqc2)

Figure 1.1 Normal and aberrant translation.

(a) Scheme of translation of normal mRNA with 5' cap (m^7GpppG), 3' polyA tail, 3' and 5' untranslated regions (UTR), and open reading frame (ORF) indicated. Translation proceeds in three phases; initiation, elongation, and termination. (b) In the absence of an in-frame stop codon, the ribosome continues elongation into the poly(A) tail. This activates nonstop decay and the recruitment of the GTPase Ski7 in *Saccharomyces cerevisiae.* In other eukaryotes, Dom34 and Hbs1 are recruited. (c) Peptide or mRNA induced stalls (indicated by gray box on mRNA scheme) during the elongation cycle activate the no-go mRNA decay pathway. The rescue factors Dom34 and Hbs1 are recruited to recycle the ribosome in this pathway as well. Both nonstop decay and no-go decay result in endonucleolytic cleavage and degradation of the mRNA and ubquitylation and degradation of the nascent peptide. (d) Table of factors involved in ribosome stall induction, ribosome release, and nascent protein degradation discussed in text.

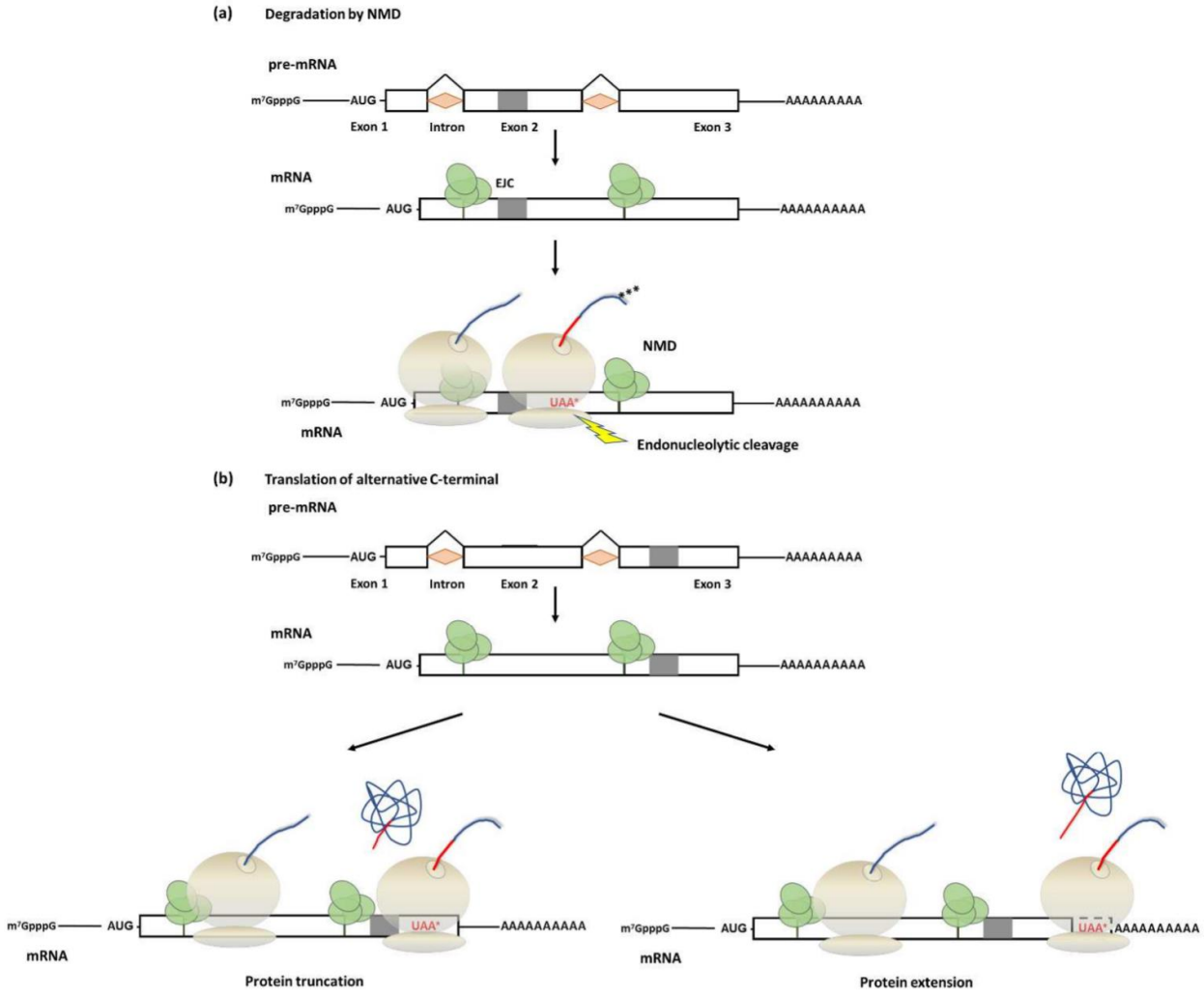
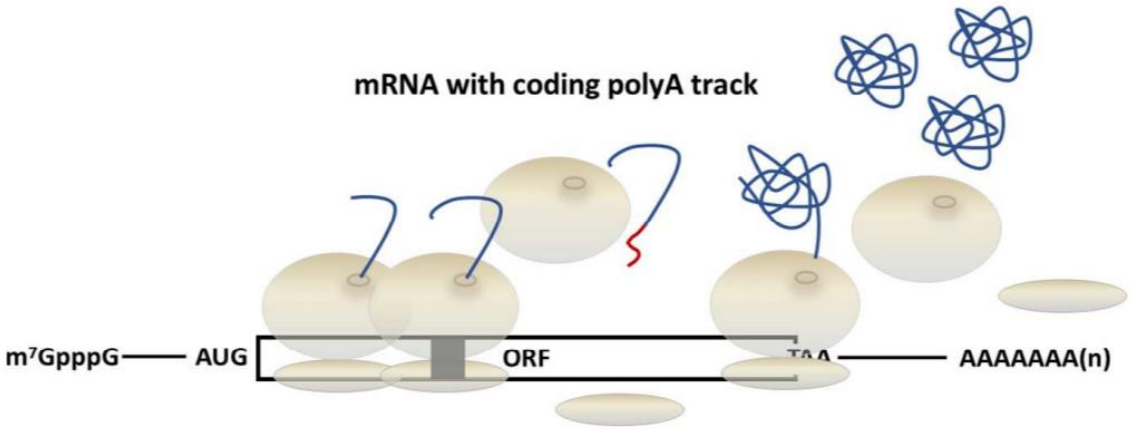
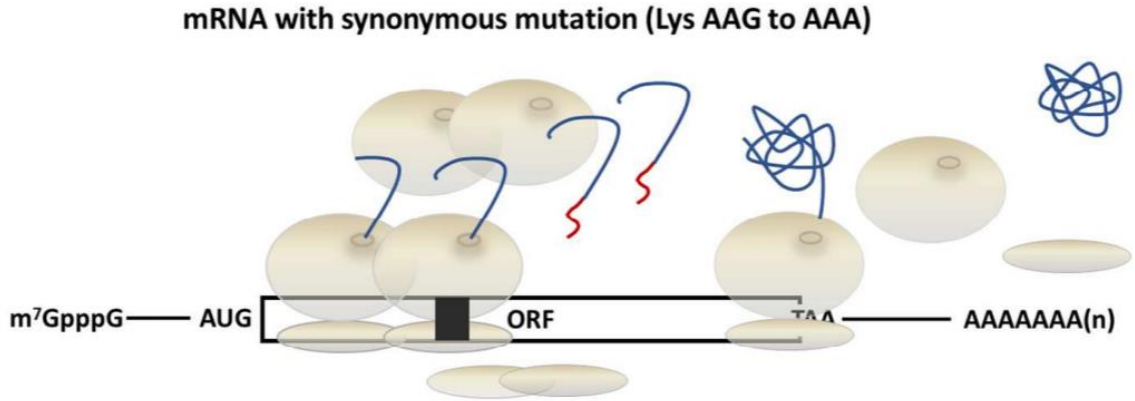

(a) Degradation by NMD
pre-mRNA
m7GpppG
AUG
AAAAAAAAA
Exon 1
Intron
Exon 2
Exon 3
mRNA
EJC
NMD
UAA*
Endonucleolytic cleavage
(b) Translation of alternative C-terminal
Protein truncation
Protein extension

Figure 1.2 Potential outcomes of programmed ribosome frameshifting on coding polyA tracks.

(a) Ribosomes that frameshift on a polyA track or other recoding element will continue elongation until reaching a stop codon in the new frame (protein sequence from non-zero frame indicated by red line). Encountering a premature termination codon will initiate the mRNA surveillance pathway, nonsense mediated decay to degrade both the mRNA and nascent peptide. If the premature termination codon occurs before the last exon junction complex (EJC), degradation by NMD is stimulated. (b) If the frameshift happens in the last exon, the transcript may evade NMD and instead synthesize a nascent peptide with an alternative C-terminus coded by the non-zero frame (indicated by UAA*) after the polyA track and either truncated (top) or extended (bottom) depending on the position of the out of frame termination codon.


mRNA with coding polyA track
m7GpppG
AUG
ORF
AAAAAAA(n)


mRNA with synonymous mutation (Lys AAG to AAA)
m7GpppG
AUG
ORF
AAAAAAA(n)

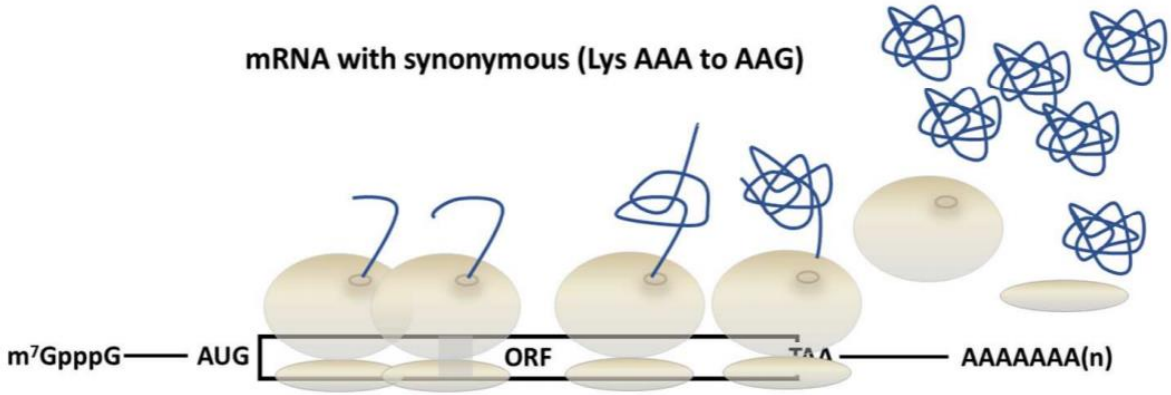

mRNA with synonymous (Lys AAA to AAG)
m7GpppG
AUG
ORF
AAAAAAA(n)

Figure 1.3 Model of effects of synonymous mutations within coding polyA tracks

(a) Scheme of translation of mRNA with polyA track indicated by the grey segment in the open reading frame. The translation efficiency of mRNAs with a polyA track are sensitive to synonymous mutations. (b) Mutations that increase the number of adenosine residues, Lys AAG to AAA, will increase stalling and frameshifting on the polyA track, leading to reduced WT protein expression from the mRNA and greater frequency of production of alternative C-terminus proteins. (c) The opposite mutation, Lys AAA to AAG, will increase WT protein expression and decrease frequency of alternative C-terminus protein production by decreasing the frequency of stalling and frameshifting on the polyA tracks.

1.6 References

1. Amaldi, F. & Pierandrei-Amaldi, P. TO P Genes : A Translationally Controlled Class of Genes Including Those Coding for Ribosomal Proteins. *Prog. Mol. Subcell. Biol.* **18,** (1997).
2. Yamashita, R. *et al.* Comprehensive detection of human terminal oligo-pyrimidine (TOP) genes and analysis of their characteristics. *Nucleic Acids Res.* **36,** 3707–3715 (2008).
3. Park, J. *et al.* Regulation of Poly(A) Tail and Translation during the Somatic Cell Cycle. *Mol. Cell* **62,** 462–471 (2016).
4. Wickens, M., Bernstein, D. S., Kimble, J. & Parker, R. A PUF family portrait : 3′UTR regulation as a way of life. *Trends Genet.* **18,** 150–157 (2002).
5. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science (80-.).* **352,** 1413–1416 (2016).
6. Szostak, E. & Gebauer, F. Translational control by 3'-UTR-binding proteins. *Brief. Funct. Genomics* **12,** 58–65 (2012).
7. Presnyak, V. *et al.* Codon optimality is a major determinant of mRNA stability. *Cell* **160,** 1111–1124 (2015).
8. Arthur, L. L. *et al.* Translational control by lysine-encoding A-rich sequences. *Sci. Adv.* **1,** e1500154 (2015).
9. Tarun, S. Z. & Sachs, A. B. Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G. *EMBO J.* **15,** 7168–77 (1996).
10. Tarun, S. Z., Wells, S. E., Deardorff, J. A. & Sachs, A. B. Translation initiation factor eIF4G mediates in vitro poly(A) tail-dependent translation. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 9046–51 (1997).
11. Wickens, M. In the beginning is the end: regulation of poly(A) addition and removal during early development. *Trends Biochem. Sci.* **15,** 320–324 (1990).
12. Chorghade, S. *et al.* Poly(A) tail length regulates PABPC1 expression to tune translation in the heart. *Elife* **e24139,** 1–19 (2017).
13. Lima, S. A. *et al.* Short poly(A) tails are a conserved feature of highly expressed genes. *Nat. Struct. Mol. Biol.* (2017). doi:10.1038/nsmb.3499
14. Weill, L., Belloc, E., Bava, F. A. & Méndez, R. Translational control by changes in poly(A) tail length: Recycling mRNAs. *Nat. Struct. Mol. Biol.* **19,** 577–585 (2012).
15. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508,** 66–71 (2014).
16. Chang, H., Lim, J., Ha, M. & Kim, V. N. TAIL-seq: Genome-wide determination of poly(A) tail length and 3' end modifications. *Mol. Cell* **53,** 1044–1052 (2014).

17. Meaux, S. & Hoof, A. Van. Yeast transcripts cleaved by an internal ribozyme provide new insight into the role of the cap and poly (A) tail in translation and mRNA decay. *RNA* **12,** 1323–1337 (2006).

18. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat Rev Mol Cell Biol* **18,** 165–187 (2017).

19. Frischmeyer, P. a *et al.* An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. *Science (80-.).* **295,** 2258–61 (2002).

20. Ozsolak, F. *et al.* Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143,** 1018–1029 (2010).

21. van Hoof, A., Frischmeyer, P. a, Dietz, H. C. & Parker, R. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science (80-.).* **295,** 2262–4 (2002).

22. Saito, S., Hosoda, N. & Hoshino, S. I. The Hbs1-Dom34 protein complex functions in non-stop mRNA decay in mammalian cells. *J. Biol. Chem.* **288,** 17832–17843 (2013).

23. Hashimoto, Y., Takahashi, M., Sakota, E. & Nakamura, Y. Nonstop-mRNA decay machinery is involved in the clearance of mRNA 5′-fragments produced by RNAi and NMD in Drosophila melanogaster cells. *Biochem. Biophys. Res. Commun.* **484,** 1–7 (2017).

24. Doma, M. K. & Parker, R. Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. *Nature* **440,** 561–4 (2006).

25. Gandhi, R., Manzoor, M. & Hudak, K. a. Depurination of Brome mosaic virus RNA3 in vivo results in translation-dependent accelerated degradation of the viral RNA. *J. Biol. Chem.* **283,** 32218–28 (2008).

26. Shoemaker, C. J., Eyler, D. E. & Green, R. Dom34:Hbs1 promotes subinut dissociation and peptidyl-tRNA drop off to initiate no-go decay. *Science (80-.).* **330,** 369–372 (2010).

27. Pisareva, V. P., Skabkin, M. A., Hellen, C. U. T., Pestova, T. V. & Pisarev, A. V. Dissociation by Pelota, Hbs1 and ABCE1 of mammalian vacant 80S ribosomes and stalled elongation complexes. *EMBO J.* **30,** 1804–1817 (2011).

28. Shoemaker, C. J. & Green, R. Kinetic analysis reveals the ordered coupling of translation termination and ribosome recycling in yeast. *Proc. Natl. Acad. Sci.* **108,** E1392–E1398 (2011).

29. Brandman, O. *et al.* A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell* **151,** 1042–1054 (2012).

30. Defenouillère, Q. *et al.* Cdc48-associated complex bound to 60S particles is required for the clearance of aberrant translation products. (2013). doi:10.1073/pnas.1221724110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1221724110

31. Yonashiro, R. *et al.* The Rqc2/Tae2 subunit of the Ribosome-Associated Quality Control (RQC) complex marks ribosome-stalled nascent polypeptide chains for aggregation. *Elife*

5, (2016).

32. Bengtson, M. H. & Joazeiro, C. A. P. Role of a ribosome-associated E3 ubiquitin ligase in protein quality control. *Nature* **467,** 470–3 (2010).

33. Ito-Harashima, S., Kuroha, K., Tatematsu, T. & Inada, T. Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast. *Genes Dev.* **21,** 519–24 (2007).

34. Lu, J. & Deutsch, C. Electrostatics in the ribosomal tunnel modulate chain elongation rates. *J. Mol. Biol.* **384,** 73–86 (2008).

35. Dimitrova, L. N., Kuroha, K., Tatematsu, T. & Inada, T. Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. *J. Biol. Chem.* **284,** 10343–52 (2009).

36. Juszkiewicz, S. & Hegde, R. S. Initiation of Quality Control during Poly(A) Translation Requires Site-Specific Ribosome Ubiquitination. *Mol. Cell* **65,** 1–8 (2017).

37. Sundaramoorthy, E. *et al.* ZNF598 and RACK1 Regulate Mammalian Ribosome-Associated Quality Control Function by Mediating Regulatory 40S Ribosomal Ubiquitylation. *Mol. Cell* **65,** 1–10 (2017).

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

39. Kuroha, K. *et al.* Receptor for activated C kinase 1 stimulates nascent polypeptide-dependent translation arrest. *EMBO Rep.* **11,** 956–61 (2010).

40. Gerbasi, V. R., Weaver, C. M., Hill, S., Friedman, D. B. & Link, A. J. Yeast Asc1p and mammalian RACK1 are functionally orthologous core 40S ribosomal proteins that repress gene expression. *Mol. Cell. Biol.* **24,** 8276–8287 (2004).

41. Tsuboi, T. *et al.* Dom34: Hbs1 Plays a General Role in Quality-Control Systems by Dissociation of a Stalled Ribosome at the 3' End of Aberrant mRNA. *Mol. Cell* **46,** 518–529 (2012).

42. Shao, S., Brown, A., Santhanam, B. & Hegde, R. S. Structure and assembly pathway of the ribosome quality control complex. *Mol. Cell* **57,** 433–445 (2015).

43. Karlin, S., Brocchieri, L., Bergman, A., Mrazek, J. & Gentles, A. J. Amino acid runs in eukaryotic proteomes and disease associations. *PNAS* **99,** 333–338 (2002).

44. Habich, M., Djuranovic, S. & Szczesny, P. PATACSDB—The database of polyA translational attenuators in coding sequences. *PeerJ Comput. Sci.* **2,** e45 (2016).

45. Koutmou, K. S. *et al.* Ribosomes slide on lysine-encoding homopolymeric A stretches. *Elife* **4,** 1–18 (2015).

46. Arthur, L. L. *et al.* Rapid generation of hypomorphic mutations. *Nat. Commun.* **8,** 1–15 (2017).

47. Chan, P. P. & Lowe, T. M. GtRNAdb : A database of transfer RNA genes detected in

genomic sequence. *Nucleic Acids Res.* **37,** 93–97 (2009).

48. Mangus, D. A., Evans, M. C. & Jacobson, A. Poly(A)-binding proteins : multifunctional scaffolds for the post- transcriptional control of gene expression. *Genome Biol.* **4,** (2003).

49. Brandman, O. *et al.* A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell* **151,** 1042–54 (2012).

50. Ketteler, R. On programmed ribosomal frameshifting: The alternative proteomes. *Front. Genet.* **3,** 1–10 (2012).

51. Letzring, D. P., Wolf, A. S., Brule, C. E. & Grayhack, E. J. Translation of CGA codon repeats in yeast involves quality control components and ribosomal protein L1. *RNA* **19,** 1208–1217 (2013).

52. Wolf, A. S. & Grayhack, E. J. Asc1, homolog of human RACK1, prevents frameshifting in yeast by ribosomes stalled at CGA codon repeats. *RNA* **21,** 935–945 (2015).

53. Nagy, E. & Maquat, L. E. A rule for termination-codon position within intron-containing genes: When nonsense affects RNA abundance. *Trends Biochem. Sci.* **23,** 198–199 (1998).

54. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156,** 1324–35 (2014).

Chapter 2: Translational control by lysine-encoding A-rich sequences

Preface

The following work was performed by me, Sergej Djuranovic, Slavia Pavlovic-Djuranovic, Kristin Smith-Koutmou, Rachel Green, and Pawel Szczesny. S. D. and R. G. conceived and supervised the project. P.S. analyzed and interpreted bioinformatic data. L.L.A, S.P-D, and K.S-K designed and conducted experiments and analyzed and interpreted data. All authors contributed to drafting and revising the manuscript.

This chapter is published in its entirety [Arthur, L.L., Pavlovic-Djuranovic, S., Koutmou, K., Green, R., Szczesny P., & Djuranovic, S. Translational control by lysine-encoding A-rich sequences. Sci. Adv. 1, e1500154 (2015).] and is available at http://advances.sciencemag.org/content/1/6/e1500154. This article is published under the Creative Commones Attribution-NonCommercial license which permits reproduction in any medium.

We thank D. Owyoung, J. T. Mendell, J. Coller, and T. Schedle for helpful comments during preparation of the manuscript. This work was supported by the NIH (grant T32 GM: 007067 to L.L.A and grant F32 GM100608 to K.S-K) and the American Cancer Society (grant IRG-58-010-58-2 to S.D.).

2.1 Abstract

Regulation of gene expression involves a wide array of cellular mechanisms that control the abundance of the RNA or protein products of that gene. Here we describe a gene-regulatory mechanism that is based on poly(A) tracks that stall the translation apparatus. We show that creating longer or shorter runs of adenosine nucleotides, without changes in the amino acid sequence, alters the amount of protein output and the stability of mRNA. Sometimes these changes result in the production of an alternative "frame-shifted" protein product. These observations are corroborated using reporter constructs and in the context of recombinant gene sequences. Approximately two percent of genes in the human genome may be subject to this uncharacterized, yet fundamental form of gene regulation. The potential pool of regulated genes encodes many proteins involved in nucleic acid binding. We hypothesize that the genes we identify are part of a large network whose expression is fine-tuned by poly(A)-tracks, and we provide a mechanism through which synonymous mutations may influence gene expression in pathological states.

2.2 Introduction

Gene expression in cells is a multistep process that involves transcription of genetic material from DNA to RNA and ultimately translation of mRNA into protein. These processes are subject to stringent control at all levels. Translational regulation generally controls the amount of protein generated from a given mRNA. While a majority of translational regulation mechanisms target the recruitment of ribosomes to the initiation codon, the protein synthesis machinery can also modulate translation elongation and termination (1, 2).

Pausing during the translational cycle ─ so-called ribosome stalling ─ is one mechanism by which the level of translation elongation can be regulated. Ribosome stalling is recognized by components of mRNA surveillance pathways, no-go decay (NGD) and non-stop decay (NSD), resulting in endonucleolytic cleavage of the stalled mRNA, ribosome rescue and proteolytic degradation of incomplete protein products (3). NGD and NSD act on aberrant mRNAs that result in translational arrest, as observed with damaged bases, stable stem-loop structures (4), rare codons (5) or mRNAs lacking stop codons (6). However, these mechanisms also act on more specific types of translational pauses, such as runs of codons that encode consecutive basic amino acids (7, 8). It is thought that polybasic runs, as well as translation of the poly(A) tail in the case of non-stop mRNAs, cause ribosome stalling through interaction of the positively charged peptide with the negatively charged ribosome exit channel (9). Presumably, the strength of the stall is dependent on the length and composition of the polybasic stretch, and thus the impact on overall protein expression might vary (3). Given this logic, it seems plausible that such an amino acid motif may act as a gene regulatory element that would define the amount of protein translated and the stability of the mRNA. For example, structural and biophysical differences between lysine and arginine residues as well as potential mRNA sequence involvement could act to further modulate this process.

Most studies investigating the effects of polybasic sequences during translation have used reporter sequences in *E. coli* (10), yeast (8, 11) or in vitro rabbit reticulocyte lysate (9). However, detailed mechanistic information about the nature of the stall in endogenous targets through genome-wide analyses have not yet been conducted. Here we report on translational regulation induced by polyA coding sequences in human cells, demonstrating that, unexpectedly, these sequences induce ribosome pausing directly, without a role for the encoded basic peptide.

2.3 Results

Bioinformatic analysis can be used as an initial approach to ask whether there are evolutionary constraints that limit the abundance of polybasic amino acid residues. Runs of polybasic residues in coding sequences of genes from many eukaryotic organisms are underrepresented when compared to runs of other amino acids (12). Interestingly, polyarginine runs have a similar abundance to polylysine runs at each segment length across multiple organisms (Supplementary Fig. S1). We developed a series of mCherry reporters to evaluate the effects of polybasic sequences on translation efficiency (output). The reporter construct consists of a double HA-tag, a run of control or polybasic sequences, followed by the mCherry reporter sequence (HA-mCherry, Fig. 1A.). As a control for DNA transfection and in vivo fluorescence measurements, we also created a construct with green fluorescent protein (GFP). We used our reporters to ask whether the polybasic sequences influence translation of reporter sequences in neonatal human fibroblasts (HDFs) as well as in Drosophila S2 cells and Chinese hamster ovary cells (CHO) (Fig. 1B, C and Supplementary Fig S2 and S3). We followed expression of the mCherry reporter using fluorescence at 610nm in vivo or western blot analyses of samples collected 48 hours after transfection (Fig. 1B, C). The stability of reporter mRNAs was determined using previously described quantitative reverse transcription polymerase chain reaction assay (qRT-PCR, (13)) (Fig. 1D). By careful primer design, this method allows us to estimate the level of endonucleolytic cleavage on mRNAs with stalled ribosome complexes.

The results of DNA transfections indicate that polylysine codons specifically inhibit translation and decrease the stability of the mCherry reporter mRNA while up to 12 arginine codons (AGG and CGA) have much less, if any effect, on either translation or mRNA stability

(Fig. 1B-D and Supplementary Fig S2 and S3). The potency of translational repression by lysine codons is clearly seen with as few as six AAA coded lysines (AAA6) and increases with the length of the homo-polymeric amino acid run. We also note that the levels of expressed mCherry reporters (Fig 1B and C) correlate with the stability of their mRNAs (Fig. 1D), consistent with earlier published observations (4, 6, 11). To control for possible transcriptional artifacts due to the effects of homo-polymeric sequence on transcription by RNA polymerase, we also electroporated mRNAs synthesized in vitro by T7 RNA polymerase directly into HDF cells. Previous studies established that T7 RNA polymerase is able to transcribe such homopolymeric sequences with high fidelity (10,13). Results of our mRNA electroporation work reproduced DNA transfection experiments, consistent with models of translational repression triggered by lysine codons (Supplementary Fig. S4). To assess whether the stability of polylysine reporter mRNAs is dependent on translation, we introduced the translation initiation inhibitor harringtonine (15) into HDF cells prior to mRNA electroporation. In this case, we did not observe any significant change in mRNA stability between wild type and polylysine mCherry constructs (Supplementary Fig. S5) indicating that accelerated decay of polylysine mCherry mRNAs is dependant on translation. As a further support for this observation insertion of the 36 nucleotide sequence equivalent to twelve lysine codons (36As) after the stop codon, in the 3'UTRregion, did not have any effect on expression level or mRNA stability of assayed construct (Suplementary Fig S6). Insertion of polylysine codons at different positions along the coding sequence drastically reduced reporter expression and mRNA levels independently of the position in construct. As such, it follows that the observed changes in mRNA stability (Fig. 1D) are result from translation-dependent processes.

The most striking observation from these data is that the production of polylysine constructs is codon dependent; runs of polylysine residues coded by AAA codons have a much larger effect on the protein output from reporter constructs than an equivalent run of lysine AAG codons (Fig. 1B-D and Supplementary Fig S2-7)). This effect is unlikely to be driven by the intron-less nature of our reporter since the constructs containing human beta globin gene with two introns showed the same effect on protein output and RNA stability (Suplementray Fig S7). We also note that this effect is unlikely to be driven by tRNALys abundance, since relative protein expression and mRNA stability are comparable in cells from various species that do not share similar tRNA abundance profiles (http://gtrnadb.ucsc.edu/; Fig. 1 and Supplementary Fig. S2-7). Furthermore, the human genome encodes a comparable number of tRNA genes for AAA and AAG codons (http://gtrnadb.ucsc.edu/Hsapi19/) and general codon usage is similar (0.44 vs 0.56, AAA vs AAG). The generality of codon-dependent polylysine protein production was recently documented in *E. coli* cells where a single tRNALys(UUU) decodes both AAA and AAG codons (10).

In light of these experimental observations, we systematically explored codon usage and the distribution of lysine codons in polylysine tracks in various species (Supplementary Fig. S8). Remarkably, we find a strong underrepresentation of poly(A) nucleotide runs in regions coding for iterated lysine (even with as few as three lysines) in human genes (Supplementary Fig. S8). When there are four iterated lysine residues, the difference between expected (from data for all lysine residues) and observed codon usage for four AAA codons in a row is over one order of magnitude (Supplementary Fig. S9). Notably, similar patterns of codon usage in lysine poly(A) tracks is observed in other vertebrates (Supplementary Fig. S10). A cumulative analysis of three ribosome profiling datasets from human cells for regions encoding 4 lysines in a row revealed

that the occupancy pattern for 4 lysines encoded by three AAA and one AAG codon is different from the pattern for two, three and four AAG codons in 4 lysine tracks (Fig 2A). The latter three resemble the occupancy pattern for tracks of arginines (Supplementary Fig. S11), which is similar to the ribosome stalling on runs of basic aminoacids observed by other researchers (16). This indicates that the observed effect has to be dependent on nucleotide, not aminoacid sequence. Additionally, the first example (with three AAA and one AAG codon) has a region of increased ribosome occupancy found additionally after the analyzed region (Fig. 2A). All of these suggests that attenuation of translation for poly(A) nucleotide tracks occurs via a different mechanism than interaction of positively charged residues with negatively charged ribosomal exit tunnel.

In order to probe the potential impact of the observed disparities in codon distribution for runs of three and four consecutive lysine codons, we inserted runs of three lysine resides with various numbers of consecutive As (A9-A13) into our mCherry reporter construct (Fig 2B). As in the previous experiments (Figs. 1B and 1C), we followed the expression of the mCherry reporter as well as the stability of the mRNA (Fig. 2C-E). We find that the insertion of sequences with 12 or more consecutive As attenuates mCherry reporter expression by more than 50% with comparable effects on mRNA stability. Importantly, in each construct, no more than three lysines are encoded so the increasing effect on protein output must result from consecutive As, not Ks.

Next, we asked whether polylysine sequences from naturally occurring genes have the same general effect on expression of reporter protein. To take an unbiased approach, we selected different lengths of homopolymeric lysine runs and various distributions of AAA and AAG codons (Fig 3 A). Reporter constructs with lysine runs were electroporated into HDF cells and relative amounts of reporter expression and mRNA stability were evaluated (Fig 3B-C. As with

the designed sequences in Fig. 2B, the observed decreases in reporter protein expression and mRNA stability correlated with the number of consecutive A nucleotides and not with total number of lysine codons in the chosen sequences. Our reporter experiments together (Fig 1B-D, Fig 2B-E, Fig 3A-C and Supplementary Fig. S2-7) argue that the repressive effects of polylysine sequence are caused by iterated poly(A) tracks rather than by runs of encoded lysine residues. Similar effects were recently documented in in vivo and in vitro experiments with *E. coli* cells or purified translational system, respectively (10). The differences that we observe in expression of reporter sequences with poly(A) nucleotide tracks from human genes favor the possibility that such regions in natural genes play a "translational attenuator" role that can modulate overall protein expression.

Based on our results with insertion of 12 consecutive A nucleotides (Fig. 2C) and endogenous A-rich sequences (Fig. 3B), we suggest that runs of 11As in a stretch of 12 nucleotides (12A-1 pattern) will typically yield a measurable effect on protein expression. Since we did not require any particular codon phase, these do not necessarily encode four consecutive lysines. As such, we have used the 12A-1 pattern to search the cDNA sequence database for multiple organisms (NCBI RefSeq resource (17)). This query revealed over 1800 mRNA sequences from over 450 human genes; the proportion was similar in other vertebrates (Supplementary Table S1). Gene ontology analyses revealed an overrepresentation of nucleic acid binding proteins, especially RNA binding and poly(A) RNA binding proteins (Supplementary Table S2). The positions of poly(A) tracks are distributed uniformly along these identified sequences with no significant enrichment towards either end of the coding region (Fig 4A). The proteins encoded by these mRNAs are often conserved among various eukaryotes; of the 7636 protein isoforms coded by mRNA with poly(A) tracks from human, mouse, rat, cow,

frog, zebrafish and fruit fly, 3877 are classified as orthologous between at least two organisms. These orthologous proteins share very similar codon usage in the poly-lysine track, as seen in the example of the RASAL2 tumor suppressor protein (18) (Supplementary Fig. S12). These observations are consistent with the idea that poly(A) tracks may regulate specific sets of genes in these different organisms. Additional analyses of the ribosome profiling data for mRNAs from selected pools of genes (12A-1 pattern genes) showed an increased number of ribosome footprints (RPFs) in sequences following the polyA tracks (Supplementary Fig. S11). The observed pattern was similar, albeit more pronounced, to the pattern observed for 4 lysine tracks encoded by 3 AAA codons and one AAG(Fig. 2A), despite the fact that in many cases the selected pattern did not encode four lysines.

Given the strong sequence conservation and possible role in modulation of protein expression, we further explored the effects of mutations in poly(A) tracks. We used our reporter constructs containing poly(A) nucleotide tracks from endogenous genes (ZCRB1, MTDH and RASAL2) to evaluate effects of synonymous lysine mutations in these poly(A) tracks on protein expression (Figure 4B-D, and Supplementary Figs. S13-14). In each construct, we made mutations that changed selected AAG codons to AAA, increasing the length of consecutive As. Alternatively, we introduced AAA to AAG changes to create interruptions in poly(A) tracks. Reporter constructs with single AAG-to-AAA changes demonstrate consistent decreases in protein expression and mRNA stability. Conversely, AAA-to-AAG changes result in increases in protein expression and mRNA stability (Fig. 4C-D and Supplementary Figs. S13, S14).

We next asked whether the same synonymous mutations have similar effects when cloned in the full-length coding sequence of the ZCRB1 gene (Figure 4E-G, Supplementary Fig. S15). Indeed, the effects on protein and mRNA levels that we observed with the mCherry

reporter sequences are reproduced within the context of the complete coding sequence of the ZCRB1 gene (and mutated variant). Mutation of single AAG-to-AAA codons in the polyA track of the ZCRB1 gene (K137K; 411G>A) resulted in a significant decrease in both protein expression and mRNA stability (Fig. 4F and G, Supplementary Fig. S15); substitution of two AAA codons with synonymous AAG codons (K136K:408 A>G; K139K:417A>G) resulted in increases in both recombinant ZCRB1 protein output and mRNA stability. Generally, mutations resulting in longer poly(A) tracks reduced protein expression and mRNA stability, while synonymous substitutions that result in shorter poly(A) nucleotide tracks increased both protein expression and mRNA stability. From these observations we suggest that synonymous mutations in poly(A) tracks could have modulatory effects on protein production from these genes.

Poly(A) tracks resemble ribosome "slippery" sequences that have been associated with translational frame-shifts (19, 20). Recent studies suggest that polyA tracks can induce "sliding" of *E. coli* ribosomes resulting in frameshifting (10). Therefore, we looked for potential frame-shifted products of overexpressed ZCRB1 variants by immuno-precipitation using an engineered N-terminally located HA-tag. We observed the presence of a protein product of the expected size that results from possible frame-shifting in our construct with increased length A tracts (ZCRB K137K (411G>A) mutant) (Fig. 5A). The presence of potential frame-shifted protein products was not observed in WT or control double synonymous mutations K136K(408 A>G): K139K(417A>G). Interestingly, we note that the K137K-synonymous change represents a recurrent cancer mutation found in the COSMIC database (COSMIC stands for Catalogue Of Somatic Mutations In Cancer, http://cancer.sanger.ac.uk, (21)) for ZCRB1 gene (http://cancer.sanger.ac.uk/cosmic/mutation/overview?id=109189). Similar results were obtained when we compared immuno-precipitations of overexpressed and HA-tagged wild type MTDH

gene and a K451K (1353 G>A) variant, yet another cancer-associated mutation (http://cancer.sanger.ac.uk/cosmic/mutation/overview?id=150510; Supplementary Fig. S16).

To further document the extent and direction of frame-shifting in the ZCRB1 transcript, we introduced polyA tracks from WT ZCRB1 and a K137K ZCRB1 mutant into a Renilla luciferase reporter gene. We introduced single or double nucleotide(s) downstream in the reporter sequence following the A track, thus creating +1 and -1 frame-shift (FS) constructs, respectively (Fig. 5B). When compared to wild type ZCRB1 polyA track, the G>A mutant shows decreases in full length luciferase protein expression (approximately 40% reduction in "zero" frame); additionally, the G>A mutant exhibits an increase in expression of -1FS frame construct (which is not observed in the wild type ZCRB1 poly(A) track -1FS construct). (Fig. 5C). The total amount of luciferase protein activity from the -1FS ZCRB1 G>A mutant construct is approximately 10% of that expressed from the "zero" frame mutant construct (Fig. 5C and Supplementary Fig S17). No significant change in luciferase expression was detected in samples electroporated with +1FS constructs where expression from these constructs resulted in background levels of luciferase activity (Supplementary Fig S17).

Frame-shifting and recognition of out-of-frame premature stop codons can lead to nonsense-mediated mRNA decay (NMD) that results in targeted mRNA decay (22, 23). Our recent data suggests that NMD may play a role in polyA track-containing mRNA stability. Deletion of NMD factor Upf1p in yeast cells partially rescues mRNA levels from constructs with poly(A) tracks (10). We have analyzed the complete set of human poly(A) track-containing genes to see whether they would be likely targets for NMD as a result of frameshifting on the poly(A) track (based on the usual rules for NMD, (24–27). Based on the position of the poly(A) track, and its relation to possible PTCs in the -1 and +1 frame, and the location of downstream

exon-intron boundaries, we find that only 20% of our genes of interest would likely be targeted by NMD as a result of frame-shifting during poly(A)-mediated stalling (these transcripts and position of PTCs are listed in Supplementary Table S3). The majority of human poly(A) track genes may not elicit NMD response since PTCs in both -1 and +1 frame following poly(A) tracks are more than 55 nucleotides away from established exon-intron boundaries. While the majority of frame-shift events seem to lead to proteins that would be truncated immediately after poly(A) tracks, in a few cases a novel peptide chain of substantial length may be produced (Suplementary Table S4). As such, the outcome of poly(A) track stalling and slipping may include a scenario in which a frame-shifted protein product is synthesized in addition to the full-length gene product (scheme shown in Figure 5D). The possible role and presence of such fragments from poly(A) track genes and their variants is still to be elucidated.

2.4 Conclusion

In conclusion, we present evidence that lysine coding poly(A) nucleotide tracks in human genes may act as translational attenuators. We show that the effect is dependent on nucleotide not aminoacid sequence and the attenuation occurs in a different manner from previously described polybasic amino acid runs. These “poly(A) translational attenuators” are highly conserved across vertebrates, implying that they might play an important role in balancing gene dosage. Presence of such a regulatory function is further supported by negative selection against single nucleotide variants in human poly(A) segments both in dbSNP and COSMIC databases (Supplementary Data D1, Supplementary Table S5 and Fig. S18). However, it is not yet clear what the effects stemming from synonymous mutation in poly(A) tracks are. Our results point to either alterations in protein-levels (altered gene dosage) or to the production of frame-shifted

products in the cell. As such, these translational attenuation mechanisms may supplement the already large number of mechanisms through which synonymous mutations can exert biological effects (reviewed in (28)).

2.4 Materials and Methods

2.4.1 Experimental Protocols

Cell Culture

HDF cells were cultured in Dulbecco's modified Eagle's medium (DMEM) (Gibco) and supplemented with 10% fetal bovine serum, 5% minimum essential medium nonessential amino acids (100×, Gibco), 5% penicillin and streptomycin (Gibco), and L-glutamine (Gibco). T-Rex-CHO cells were grown in Ham's F12K medium (American Type Culture Collection) with the same supplements. Drosophila S2 cells were cultured in Express Five SFM Medium (Invitrogen) supplemented with penicillin (100 U/ml), streptomycin (100 U/ml) (Gibco), and 45 ml of 200 mM L-glutamine (Gibco) per 500 ml of medium.

Plasmids and mRNA were introduced to the cells by the Neon Transfection System (Invitrogen) with 100-µl tips according to cell-specific protocols (www.lifetechnologies.com/us/en/home/life-science/cell-culture/transfection/transfection---selection-misc/neon-transfection-system/neon-protocols-cell-line-data.html). Cells electroporated with DNA plasmids were harvested after 48 hours if not indicated differently. Cells electroporated with mRNA were harvested after 4 hours, if not indicated differently. All transfections in S2 cells were performed using Effectene reagent (Qiagen).

DNA constructs

mCherry reporter constructs were generated by PCR amplification of an mCherry template with forward primers containing the test sequence at the 5′ end and homology to mCherry at the 3′ end. The test sequence for each construct is listed in the following table. The PCR product was purified by NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) and integrated into the pcDNA-DEST40, pcDNA-DEST53, or pMT-DEST49 expression vector by the Gateway cloning system (Invitrogen). Luciferase constructs were generated by the same method.

Whole gene constructs were generated by PCR amplification from gene library database constructs from Thermo (MTDH clone ID: 5298467) or Life Technologies GeneArt Strings DNA Fragments (ZCRB1) and cloned in pcDNA-DEST40 vector for expression. Synonymous mutations in the natural gene homopolymeric lysine runs were made by site-directed mutagenesis. Human β-globin gene (delta chain; HBD) was amplified from genomic DNA isolated from HDF cells. Insertions of poly(A) track, AAG codons, or premature stop codon in HBD constructs were made by site-directed mutagenesis. The sequences of inserts are given in table S6.

In vitro mRNA synthesis

Capped and polyadenylated mRNA was synthesized in vitro using a mMESSAGE mMACHINE T7 Transcription Kit (Life Technologies) following the manufacturer's procedures. The quality of mRNA was checked by electrophoresis and sequencing of RT-PCR products.

RNA extraction and qRT-PCR

Total RNA was extracted from cells using the RiboZol RNA extraction reagent (Amresco) according to the manufacturer's instructions. RiboZol reagent (400 μl) was used in

each well of 6- or 12-well plates for RNA extraction. Precipitated nucleic acids were treated with Turbo deoxyribonuclease (Ambion), and total RNA was dissolved in ribonuclease-free water and stored at −20°C. RNA concentration was measured by NanoDrop (OD260/280). iScript Reverse Transcription Supermix (Bio-Rad) was used with 1 μg of total RNA following the manufacturer's protocol. iQ SYBR Green Supermix (Bio-Rad) protocol was used for qRT-PCR on the CFX96 Real-Time system with Bio-Rad CFX Manager 3.0 software. Cycle threshold (C_t) values were normalized to the neomycin resistance gene expressed from the same plasmid.

Western blot analysis

Total cell lysates were prepared with passive lysis buffer (Promega). Blots were blocked with 5% milk in 1× tris-buffered saline–0.1% Tween 20 (TBST) for 1 hour. Horseradish peroxidase–conjugated or primary antibodies were diluted according to the manufacturer's recommendations and incubated overnight with membranes. The membranes were washed four times for 5 min in TBST and prepared for imaging, or secondary antibody was added for additional 1 hour of incubation. Images were generated by Bio-Rad Molecular Imager ChemiDoc XRS System with Image Lab software by chemiluminescence detection or by the LI-COR Odyssey Infrared Imaging System. Blots imaged by the LI-COR system were first incubated for 1 hour with Pierce DyLight secondary antibodies.

Immunoprecipitaion

Total cell lysates were prepared with passive lysis buffer (Promega) and incubated with Pierce anti-HA magnetic beads overnight at 4°C. Proteins were eluted by boiling the beads with 1× SDS sample buffer for 7 min. Loading of protein samples was normalized to total protein amounts.

Cell imaging

HDF cells were electroporated with the same amount of DNA plasmids and plated in six-well plates with optically clear bottom. Before imaging, cells were washed with fresh DMEM without phenol red and incubated for 20 min with DMEM containing 0.025% Hoechst 33342 dye for DNA staining. Cells were washed with DMEM and imaged in phenol red–free medium with an EVOS FL microscope using a 40× objective. Images were analyzed using EVOS FL software.

2.4.2 Bioinformatic Analysis

Sequence data and variation databases

Sequence data were derived from a NCBI RefSeq resource (18) on February 2014. Two variations of databases were used: dbSNP (31), build 139 and COSMIC, build v70 (23).

mRNA mapping

Because we observed some inconsistencies between transcripts and proteins in some of the sequence databases, before starting the analyses, we mapped protein sequences to mRNA sequences using the exonerate tool (32), using protein2genome model and requiring a single best match. In case of multiple best matches (when several transcripts had given identical results), the first one was chosen because the choice of corresponding isoform (this was the most common reason for multiple matches) did not influence downstream analyses.

Ribosome profiling data

Three independent studies of ribosome profiling data from human cells were analyzed: (i) GSE51424 prepared by Gonzalez and co-workers (33), from which samples SRR1562539, SRR1562540, and SRR1562541 were used; (ii) GSE48933 prepared by Rooijers and co-workers

(34), from which samples SRR935448, SRR935449, SRR935452, SRR935453, SRR935454, and SRR935455 were used; and (iii) GSE42509 prepared by Loayza-Puch and co-workers (35), from which samples SRR627620 to SRR627627 were used. The data were analyzed similarly to the original protocol created by Ingolia and co-workers (36), with modifications reflecting the fact that reads were mapped to RNA data instead of genome.

Raw data were downloaded and adapters specific for each experiments were trimmed. Then, the reads were mapped to human noncoding RNAs with bowtie 1.0.1 (37) (bowtie -p 12 -t --un), and unaligned reads were mapped to human RNAs (bowtie -p 12 -v 0 -a -m 25 --best --strata --suppress 1,6,7,8). The analysis of occupancy was originally done in a similar way to Charneski and Hurst (17); however, given that genes with poly(A) were not highly expressed and the data were sparse (several positions with no occupancy), instead of mean of 30 codons before poly(A) position, we decided to normalize only against occupancy of codon at the position 0 multiplied by the average occupancy along the gene. Occupancy data were visualized with R and ggplot2 library using geom_boxplot aesthetics. On all occupancy graphs, the upper and lower hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The upper and lower whiskers extend from hinges at 1.5*IQR of the respective hinge.

Variation analysis

To assess the differences in single-nucleotide polymorphisms (SNPs) in poly(A) regions versus random regions of the same length in other genes, we needed to use the same distribution of lengths in both cases. The distribution of lengths for poly(A) regions identified as mentioned above (12 A's allowing for one mismatch) up to length 19 (longer are rare) is presented in fig. S19. Using the same distribution of lengths, we selected one random region of length drawn from the distribution randomly placed along each gene from all human protein coding RNAs.

The distributions of the number of SNPs per segment for all poly(A) segments and for one random segment for each mRNA were compared using Welch's two-sample t test, Wilcoxon rank sum test with continuity correction, and two-sample permutation test with 100,000 permutations.

Abundance of polytracks in protein sequences

Abundance was expressed by the following equation:

$$\text{Abundance} = \frac{1}{-\log_{10} \frac{N_P}{N_R}}$$

where N_P is the number of proteins with K+ polytrack (at least 2, at least 3, etc.) and N_R is the total number of occurrences of a particular amino acid. This is to normalize against variable amino acid presence in different organisms. All isoforms of proteins were taken into account.

Other analyses

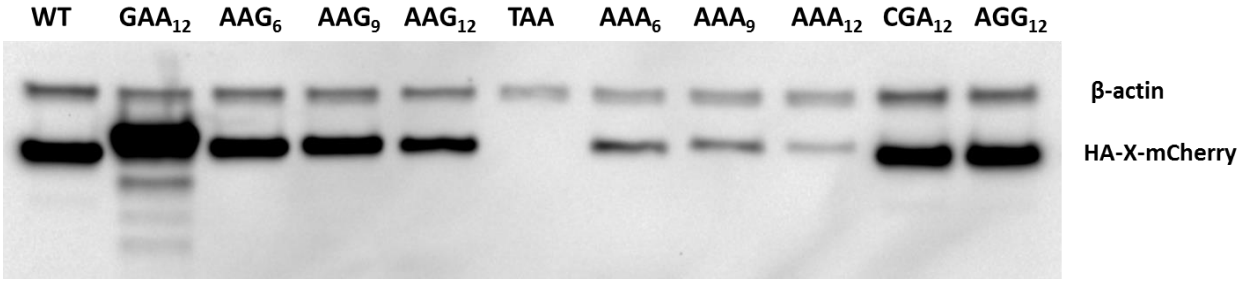
The list of human essential genes was obtained from the work of Georgi and co-workers (38). Gene Ontology analyses were done using Term Enrichment Service at http://amigo.geneontology.org/rte. Most of the graphs were prepared using R and ggplot2 library. For Fig. 3A, the values of the y axis were computed by one-dimensional Gaussian kernel density estimates implemented in the R software. Custom Perl scripts were used to analyze and merge the data

2.5 Figures

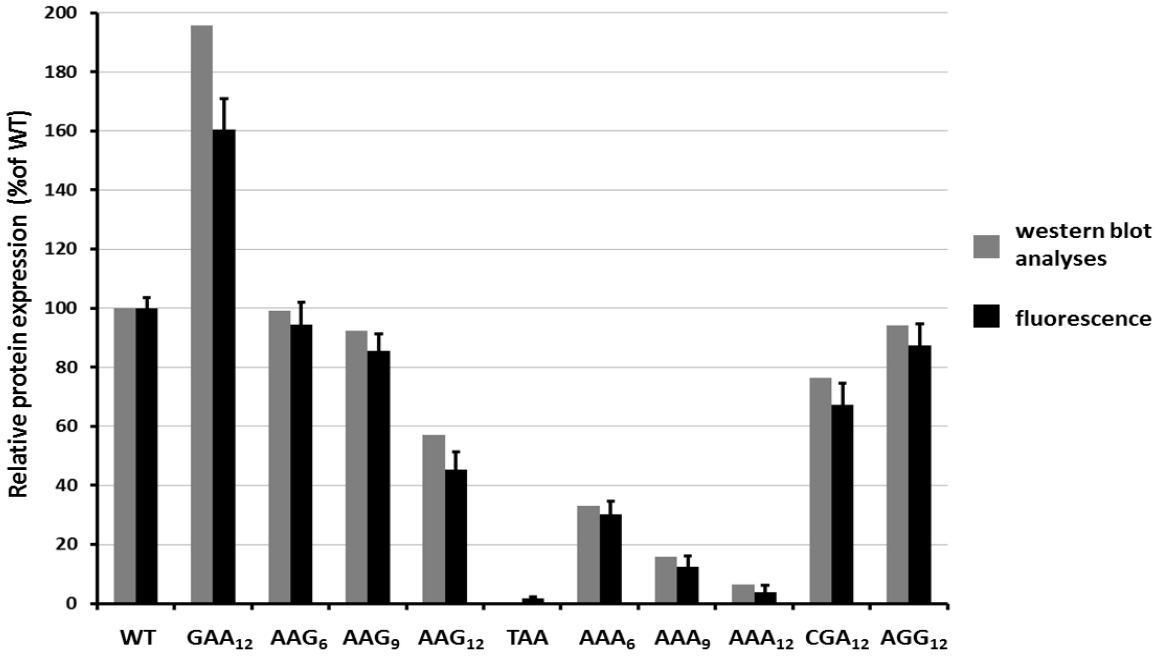
A.

$X = (AAA)_{6\text{-}12}, (AAG)_{6\text{-}12}, (GAA)_{12}, (CGA)_{12}, (AGG)_{12}$, TAA

B.



C.



D.

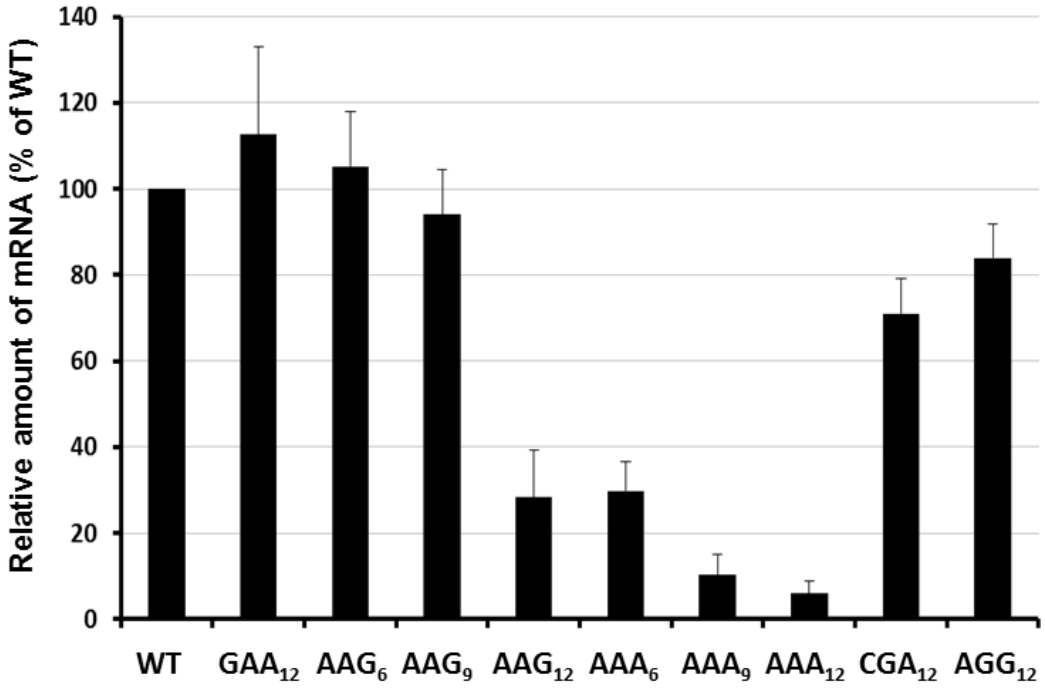
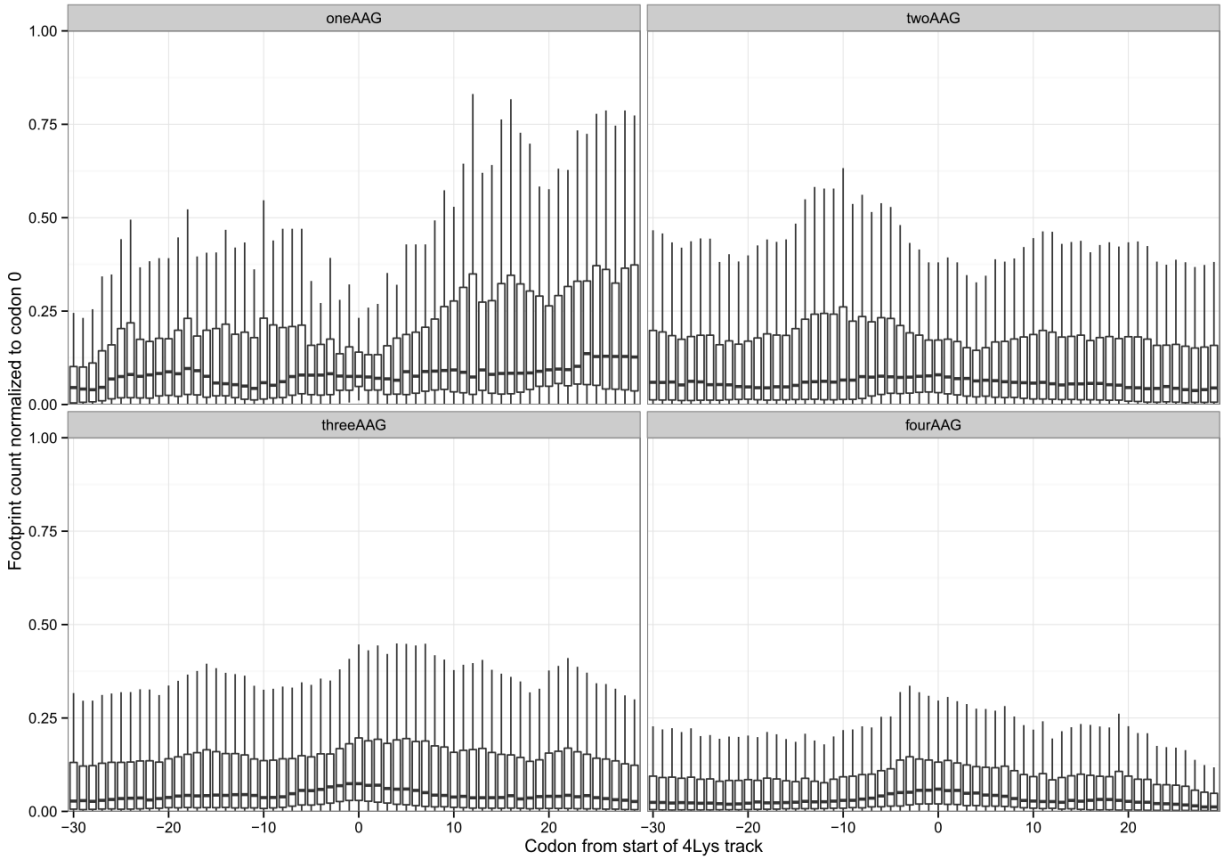


Figure 2.1. Effects of different lysine codons on mCherry reporter expression and mRNA stability.

(A) Cartoon of reporter constructs used in electroporation experiments. (B) Western blot analyses of HA-X-mCherry constructs 48 hours after electroporation (HA and β-actin antibodies). (C) Normalized protein expression using LI-COR Western blot analyses or in vivo mCherry fluorescence measurement. β-Actin or fluorescence of coexpressed GFP construct was used for normalization of the data. Each bar represents the percentage of wild-type mCherry (WT) expression/fluorescence. (D) Normalized RNA levels of HA-X-mCherry constructs. Neomycin resistance gene was used for normalization of qRT-PCR data. Each bar represents the percentage of wild-type mCherry (WT) mRNA levels.

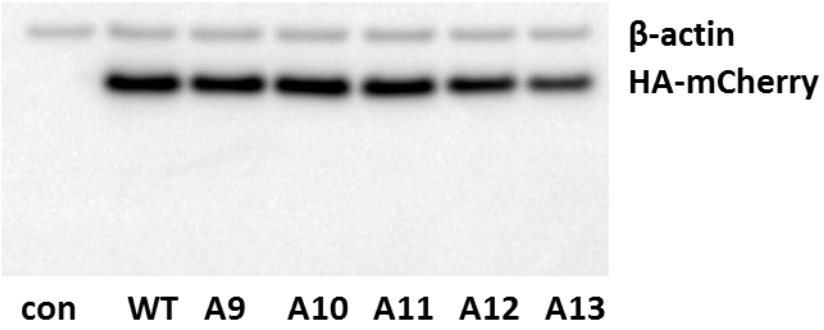
A.



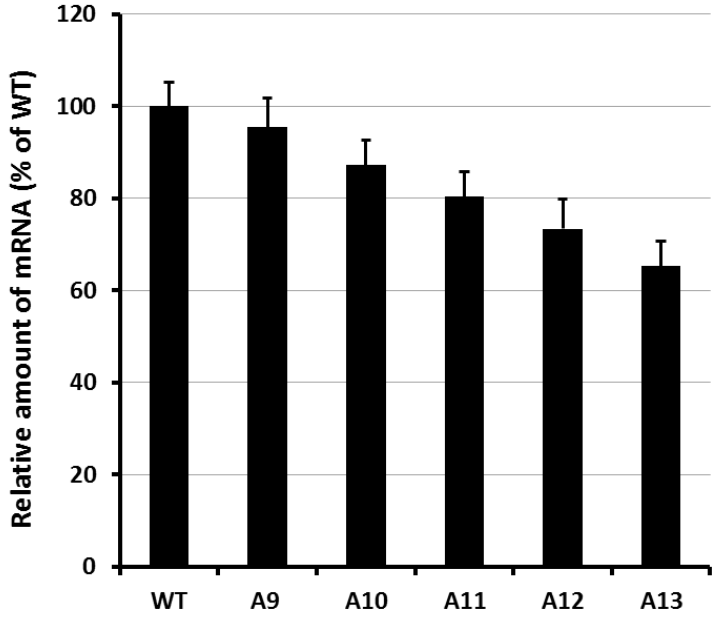
B.

WT	GCAGCGGTGAGC	A A V S
A9	GCAGCGAAAAAAAAAGTGAGC	A A K K K V S
A10	GCAGCAAAAAAAAAAGTGAGC	A A K K K V S
A11	GCAGCAAAAAAAAAACCGTG	A A K K K T V
A12	GCAGCAAAAAAAAAAACGTG	A A K K K N V
A13	GCAGAAAAAAAAAAAACGTG	A E K K K N V

C.


β-actin
HA-mCherry
con
WT
A9
A10
A11
A12
A13

D.


Relative amount of mRNA (% of WT)
120
100
80
60
40
20
0
WT
A9
A10
A11
A12
A13

E.

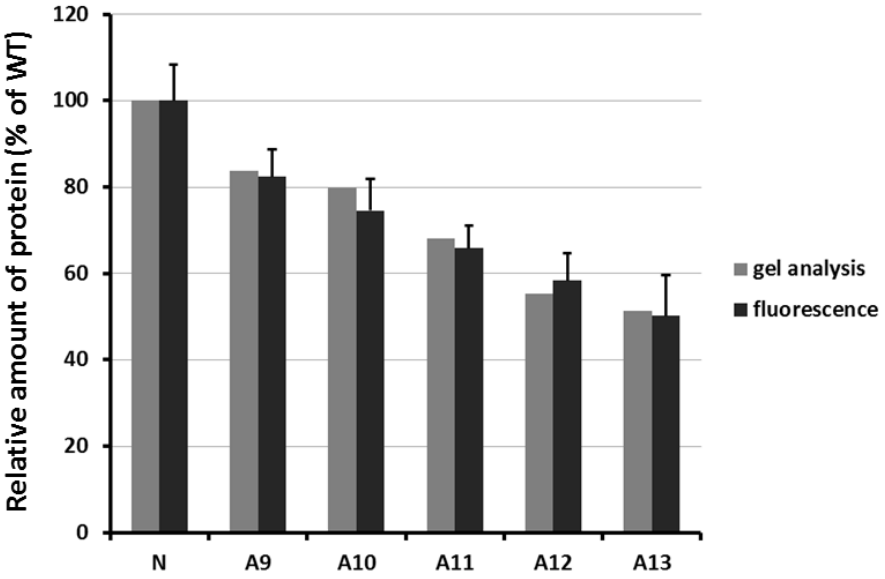
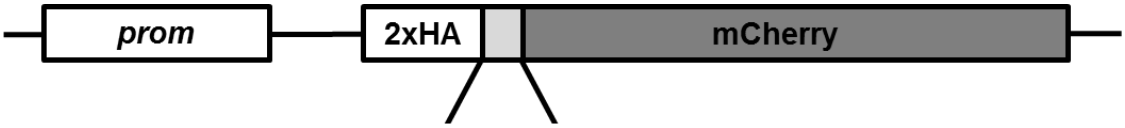

Relative amount of protein (% of WT)
120
100
80
60
40
20
0
N
A9
A10
A11
A12
A13
gel analysis
fluorescence

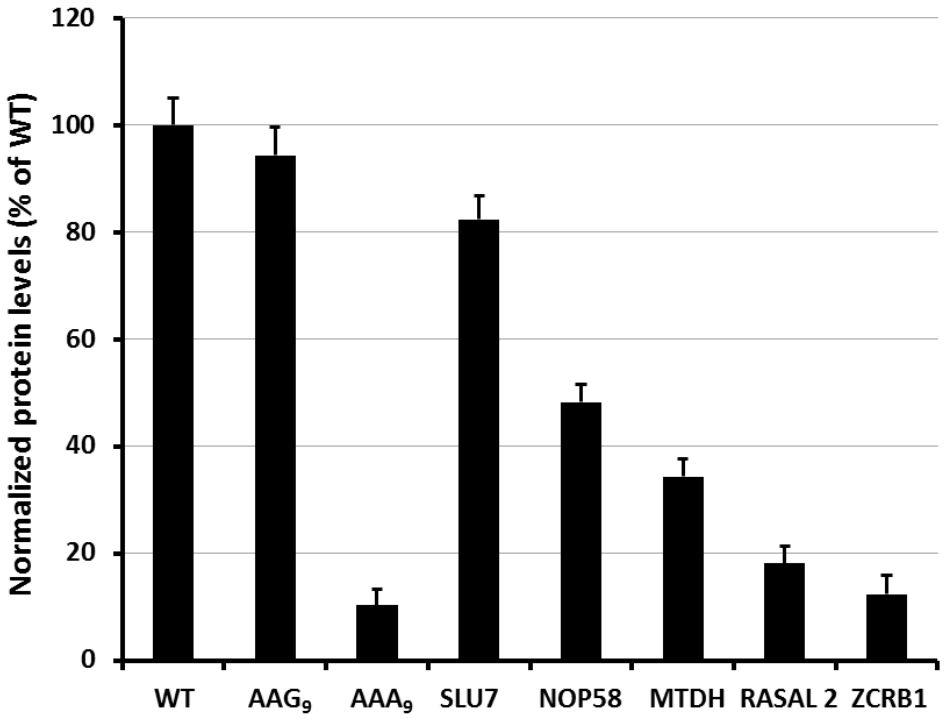
Figure 2.2. The effect of codon usage in polylysine tracks on translation and protein levels. (**A**) Occupancy of ribosomal footprints for regions around different codon combinations for four lysine tracks. All combinations of one, two, three, and four AAG codons per group are shown. Data for four AAA codons are not shown because only a single gene has such a sequence. The upper and lower "hinges" correspond to the first and third quartiles (the 25th and 75th percentiles). The upper and lower whiskers extend from hinges up or down at a maximum of 1.5*IQR (interquartile range) of the respective hinge. (**B**) Sequences of HA-(A9–A13)-mCherry constructs used in electroporation experiments. (**C**) Western blot analyses of HA-(A9–A13)-mCherry constructs 48 hours after electroporation (HA and β-actin antibodies). (**D**) Normalized protein expression using LI-COR Western blot analyses or in vivo mCherry fluorescence measurement. β-Actin or fluorescence of coexpressed GFP construct was used for normalization of the data. Each bar represents the percentage of wild-type mCherry (WT) expression/fluorescence. (**E**) Normalized RNA levels of HA-X-mCherry constructs. Neomycin resistance gene was used for normalization of qRT-PCR data. Each bar represents the percentage of wild-type mCherry (WT) mRNA levels.

A.

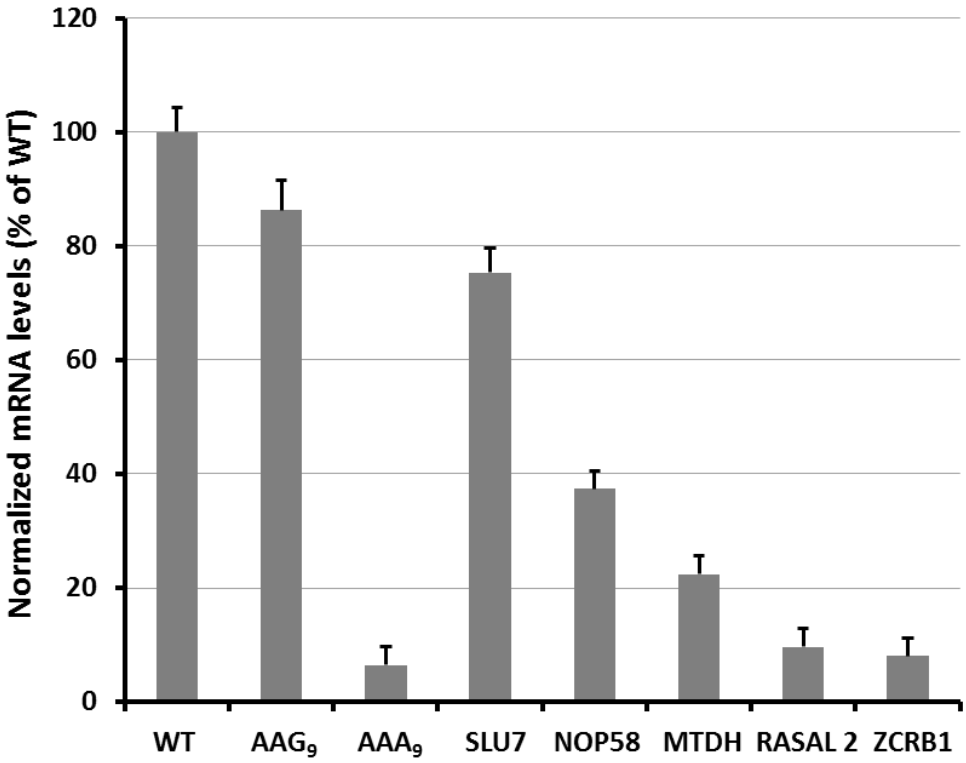


SLU7	GAGAAGAAGAAGAAGAAAAAGAAGAAGAAGAAGCAT	10 Lys (9AAG/1AAA)
MTDH	TCCAAAAAGAAAAAAAAGAAAAAGAAGAAGCAAGGT	9 Lys (5AAG/4AAA)
NOP58	GAGAAAAAGAAGAAAAAGAAAAAAAAGAGAGAGAGA	8 Lys (4AAG/4AAA)
ZCRB1	CCAAAGAAGAAAGAAAAAAAGAAAAAAAAAGAAAGCT	6 Lys (2AAG/4AAA)
RASAL2	GTGGAAAAAAAGAAAAAAAAGGACAAGAATAATTAT	5 Lys (2AAG/3AAA)

B.



C.


Normalized mRNA levels (% of WT)
120
100
80
60
40
20
0
WT
AAG9
AAA9
SLU7
NOP58
MTDH
RASAL 2
ZCRB1

D.

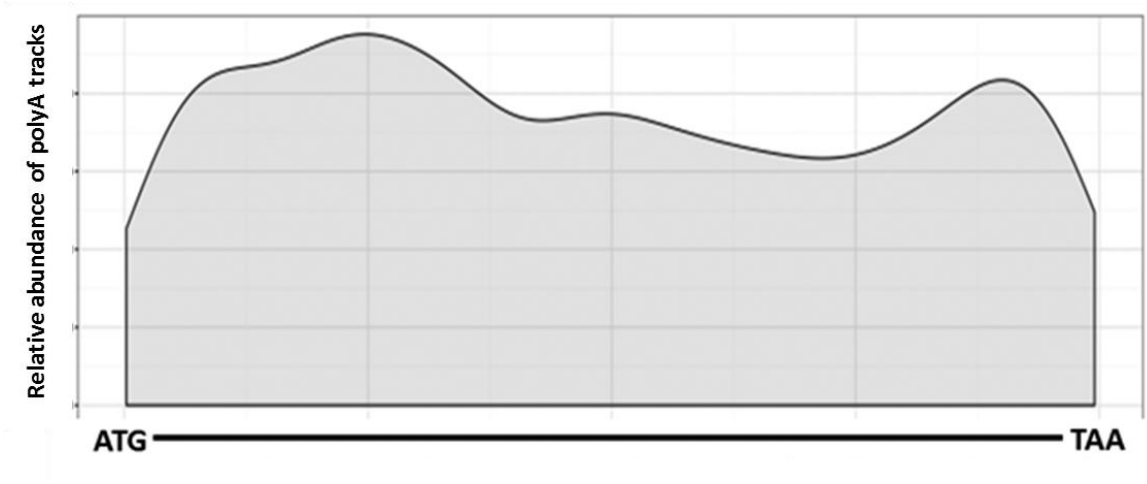
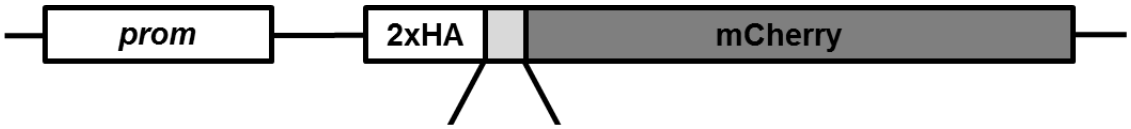

Relative abundance of polyA tracks
ATG
TAA

Figure 2.3. Native poly(A) tracks control reporter mRNA and protein levels.

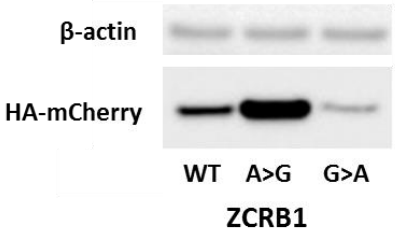
(**A**) Sequences of polylysine runs from human genes incorporated into HA-X-mCherry constructs. Continuous runs of lysine residues are labeled. The number of lysine residues and the ratio of AAG and AAA codons for each construct are indicated. (**B**) Normalized protein expression using in vivo mCherry reporter fluorescence. Fluorescence of cotransfected GFP was used to normalize the data. Each bar represents the percentage of wild-type mCherry (WT) expression/fluorescence. (**C**) Normalized RNA levels of HA-X-mCherry constructs. Neomycin resistance gene was used for normalization of qRT-PCR data. Each bar represents the percentage of wild-type mCherry (WT) mRNA levels. (**D**) Smoothed Gaussian kernel density estimate of positions of poly(A) tracks along the gene. Position of poly(A) segment is expressed as a ratio between the number of the first residue of the poly(A) track and the length of the gene.

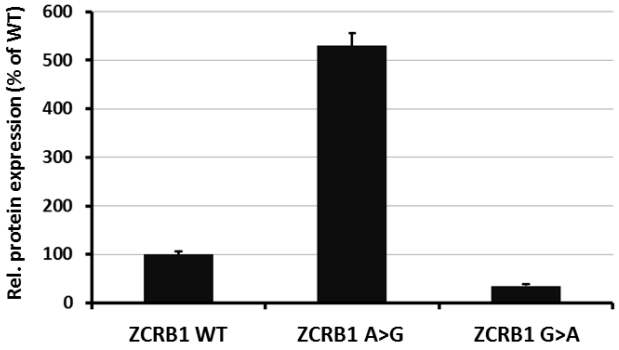
A.



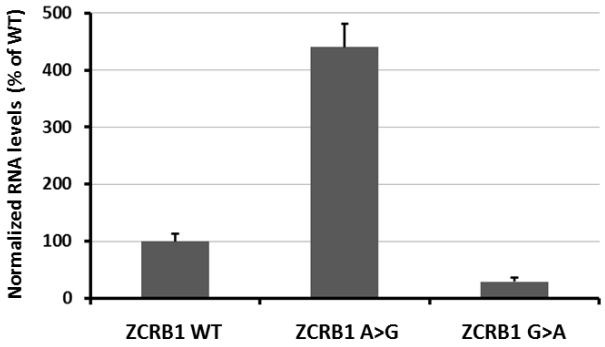
	Lys	Lys	Lys	Glu	Lys	Lys	Lys	Lys	Lys	Lys
ZCRB1 WT	AAG	AAG	AAA	GAA	AAA	AAG	AAA	AAA	AAG	AAA
ZCRB1 A>G	AAG	AAG	AAA	GAA	AAG	AAG	AAA	AAG	AAG	AAA
ZCRB1 G>A	AAG	AAG	AAA	GAA	AAA	AAA	AAA	AAA	AAG	AAA

B.





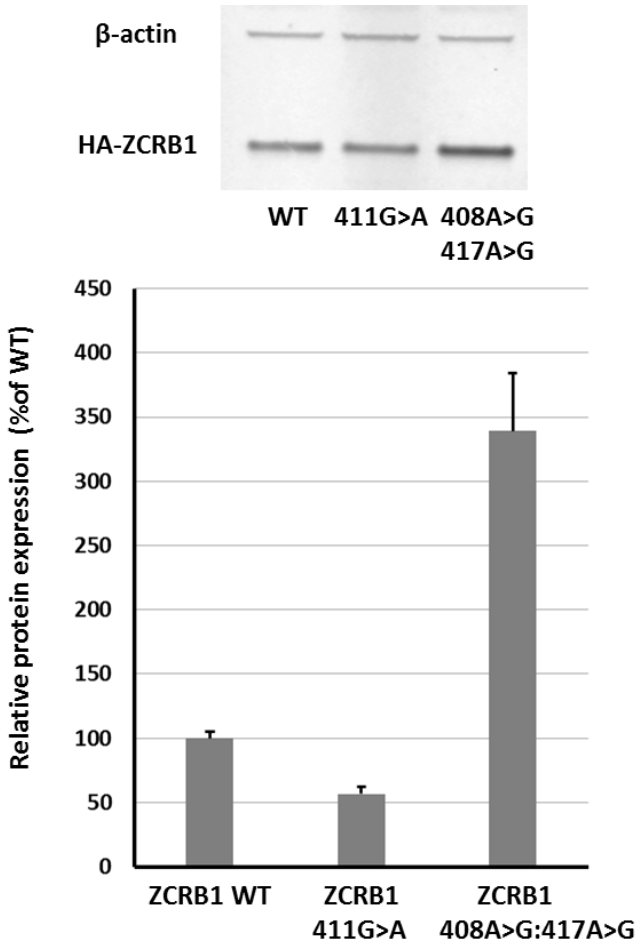
C.



D.

	Pro	Lys	Lys	Lys	Glu	Lys	Lys	Lys	Lys	Lys	Lys	Ala
ZCRB1 WT	CCA	AAG	AAG	AAA	GAA	AAA	AAG	AAA	AAA	AAG	AAA	GCT
ZCRB1 411G>A	CCA	AAG	AAG	AAA	GAA	AAA	AAA	AAA	AAA	AAG	AAA	GCT
ZCRB1 408A>G;417A>G	CCA	AAG	AAG	AAA	GAA	AAG	AAG	AAA	AAG	AAG	AAA	GCT

E.


β-actin
HA-ZCRB1
WT
411G>A
408A>G
417A>G
Relative protein expression (%of WT)
450
400
350
300
250
200
150
100
50
0
ZCRB1 WT
ZCRB1 411G>A
ZCRB1 408A>G:417A>G

F.

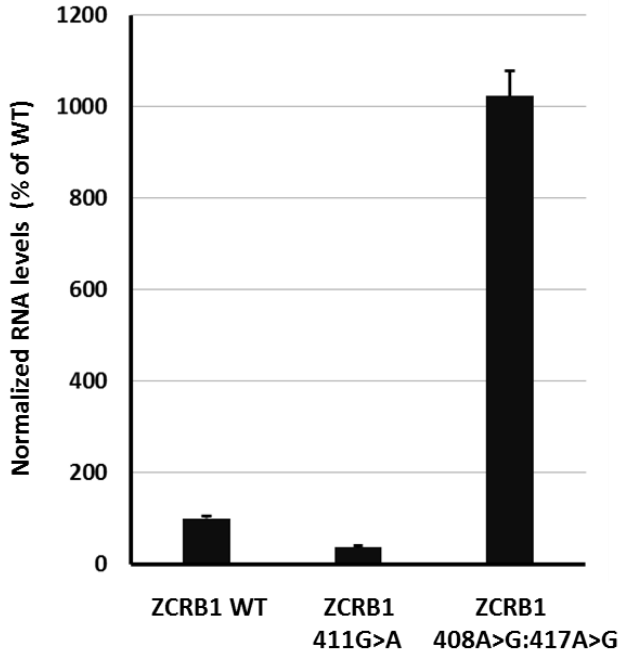
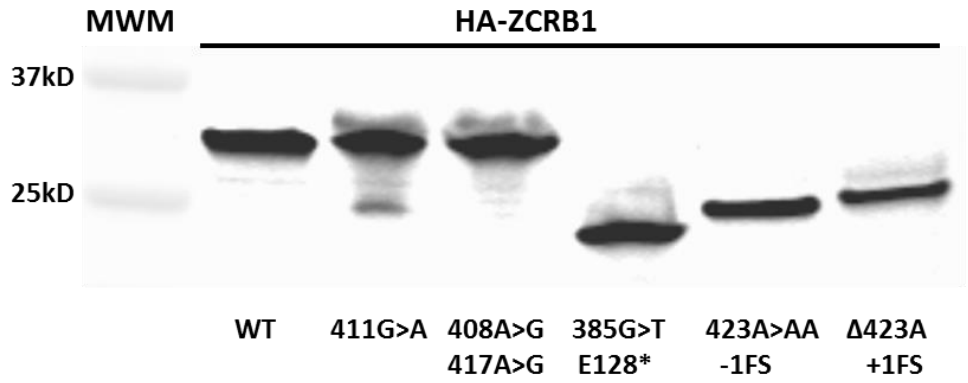

Normalized RNA levels (% of WT)
1200
1000
800
600
400
200
0
ZCRB1 WT
ZCRB1 411G>A
ZCRB1 408A>G:417A>G

Figure 2.4. The effect of synonymous mutations in poly(A) tracks of human genes.

(**A**) Scheme of constructs with ZCRB1 gene poly(A) tracks used for analyses of synonymous mutations. (**B**) Western blot analyses and normalized protein expression of ZCRB1 reporter constructs with synonymous mutations (HA and β-actin antibodies). Each bar represents the percentage of wild-type ZCRB1-mCherry (WT) expression. (**C**) Normalized RNA levels of ZCRB1 reporter constructs with synonymous mutations. Neomycin resistance gene was used for normalization of qRT-PCR data. Each bar represents the percentage of wild-type ZCRB1-mCherry construct (WT) mRNA levels. (**D**) Scheme of full-length HA-tagged ZCRB gene constructs. Position and mutations in poly(A) tracks are indicated. (**E**) Western blot analysis and normalized protein expression of ZCRB1 gene constructs with synonymous mutations. Each bar represents the percentage of wild-type HA-ZCRB1 (WT) expression. (**F**) Normalized RNA levels of ZCRB1 gene constructs. Neomycin resistance gene was used for normalization of qRT-PCR data.

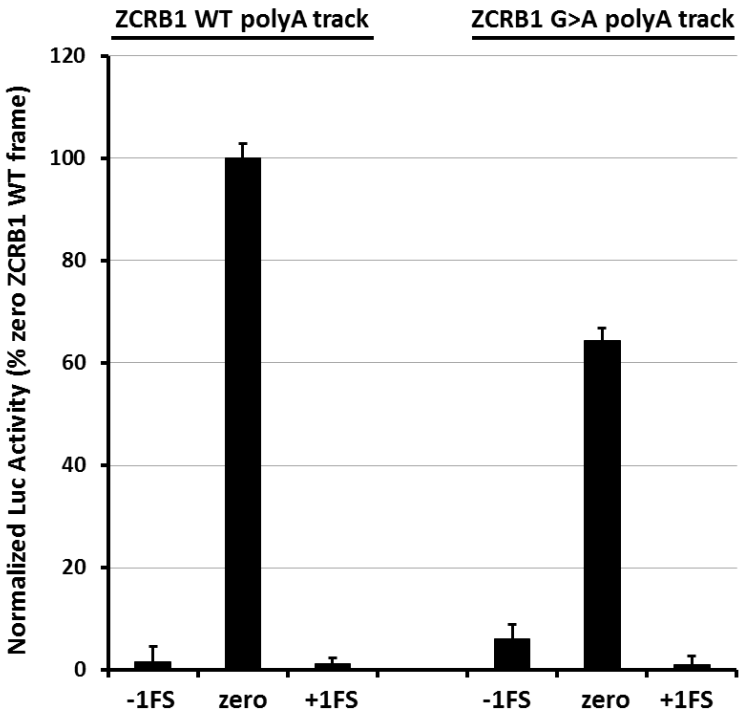
A.



B.

LUC con	ATG-2XHA-LUC
ZCRB WT - LUC	ATG-2xHA-AAGAAGAAAGAAAAAAAGAAAAAAAAGAAAGCT-LUC
ZCRB -1 - LUC	ATG-2xHA-AAGAAGAAAGAAAAAAAGAAAAAAAAGAAAGCTCTAA-LUC
ZCRB +1 - LUC	ATG-2xHA-AAGAAGAAAGAAAAAAAGAAAAAAAAGAAAGCTCCTAA-LUC
ZCRB 417G>A - LUC	ATG-2xHA-AAGAAGAAAGAAAAAAAAAAAAAAAAGAAAGCT-LUC
ZCRB 417G>A -1 - LUC	ATG-2xHA-AAGAAGAAAGAAAAAAAAGAAAAAAAAGAAAGCTCTAA-LUC
ZCRB 417G>A +1 - LUC	ATG-2xHA-AAGAAGAAAGAAAAAAAAGAAAAAAAAGAAAGCTCCTAA-LUC

C.



D.

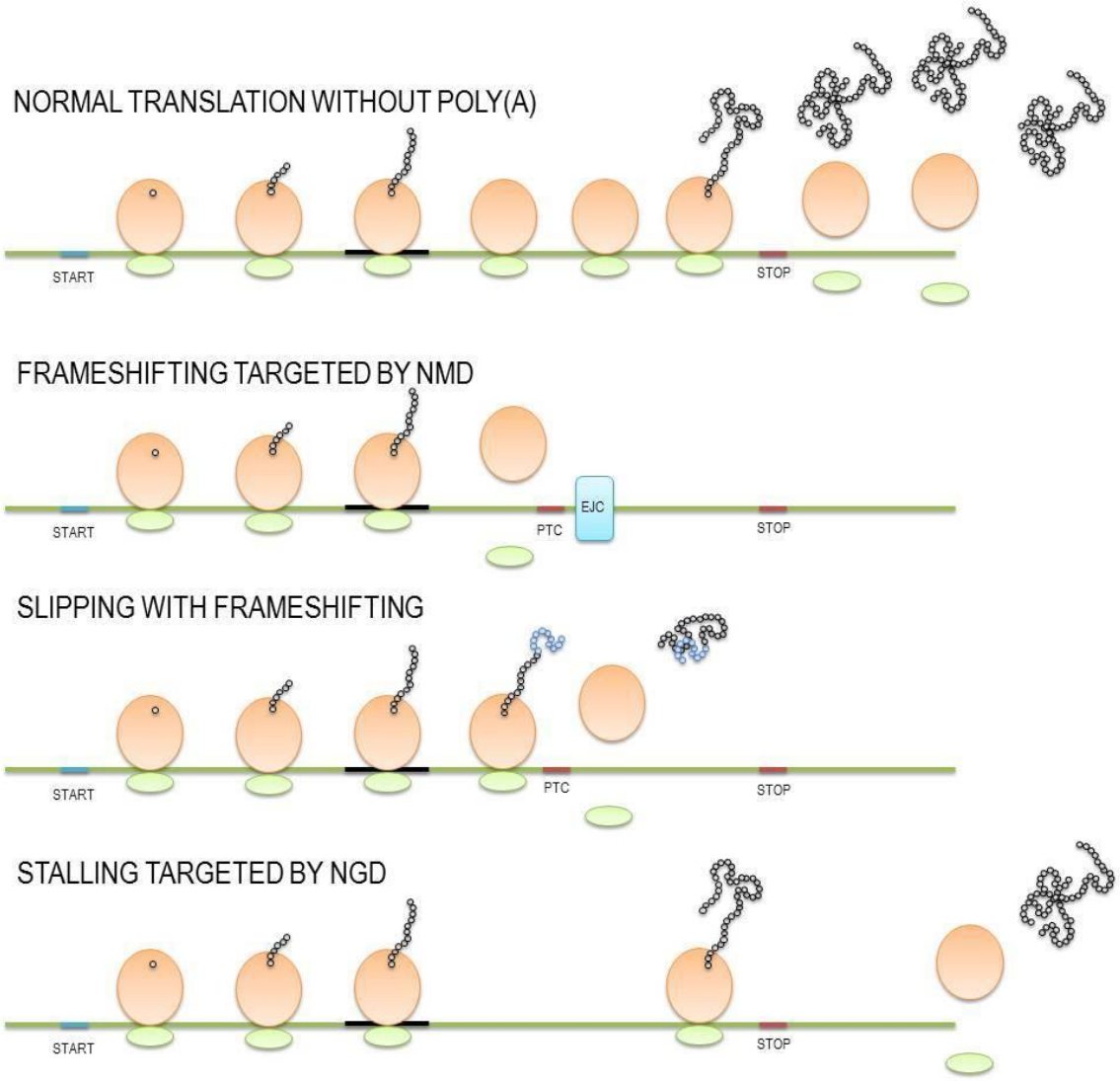
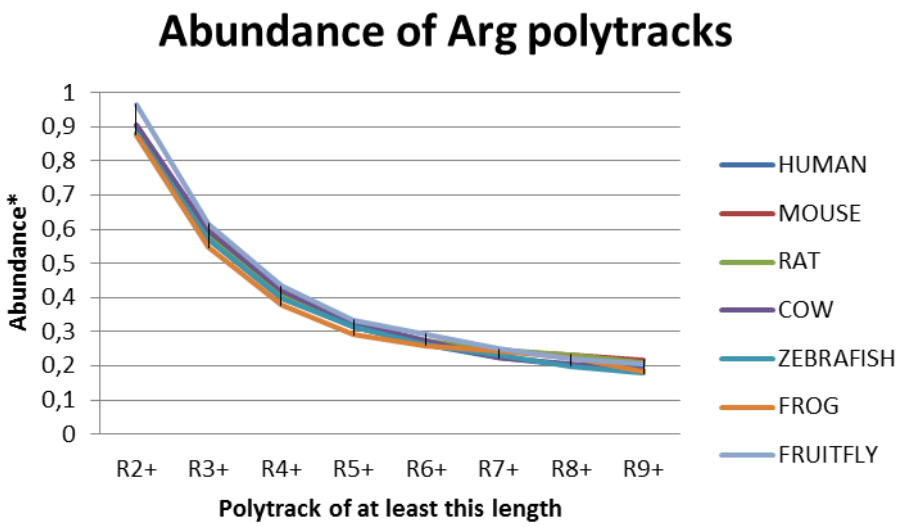

NORMAL TRANSLATION WITHOUT POLY(A)
START
STOP
FRAMESHIFTING TARGETED BY NMD
START
PTC
EJC
STOP
SLIPPING WITH FRAMESHIFTING
START
PTC
STOP
STALLING TARGETED BY NGD
START
STOP

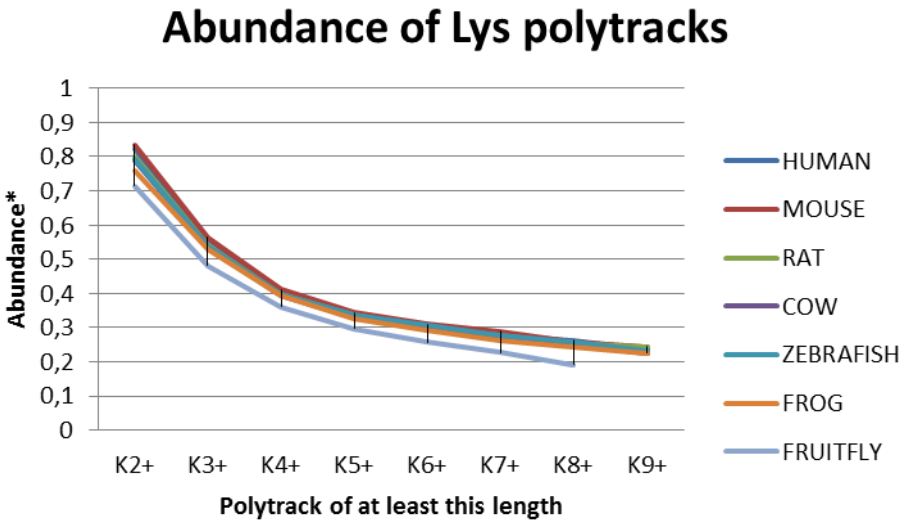
Figure 2.5. Putative mechanisms through which poly(A) tracks exert their function.

(**A**) Immunoprecipitation of HA-ZCRB gene constructs using anti-HA magnetic beads. ZCRB1 WT, synonymous (single 411G>A or double 408A>G; 417A>G), nonsense [385G>T, insertion of stop codon before poly(A) track], deletion (423ΔA, equivalent to +1 frameshift), and insertion (423A>AA, equivalent to −1 frameshift) mutant constructs are labeled. (**B**) Scheme of luciferase constructs used to estimate frameshifting potential for ZCRB1 WT and 411G>A mutant poly(A) tracks. (**C**) Luciferase levels (activity) from −1, "zero," and +1 frame constructs of wild-type and G>A mutant ZCRB1 poly(A) tracks are compared. Bars represent the normalized ratio of ZCRB1 G>A and ZCRB1 WT poly(A) tracks, elucidating changes in the levels of luciferase expression in all three frames. (**D**) Model for function of poly(A) tracks in human genes. Poly(A) tracks lead to three possible scenarios: frameshifting consolidated with NMD, which results in reduced output of wild-type protein; frameshifting with synthesis of both out-of-frame and wild-type protein; and nonresolved stalling consolidated by endonucleolytic cleavage of mRNA and reduction in wild-type protein levels, as in the NGD pathway. Scheme for translation of mRNAs without poly(A) tracks is shown for comparison.

(a)

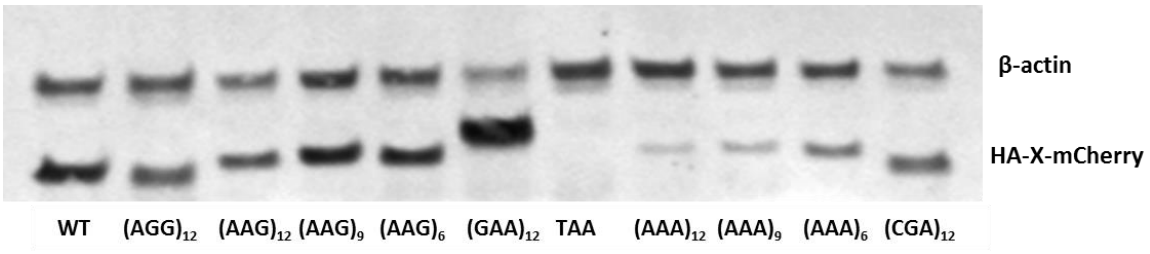


(b)

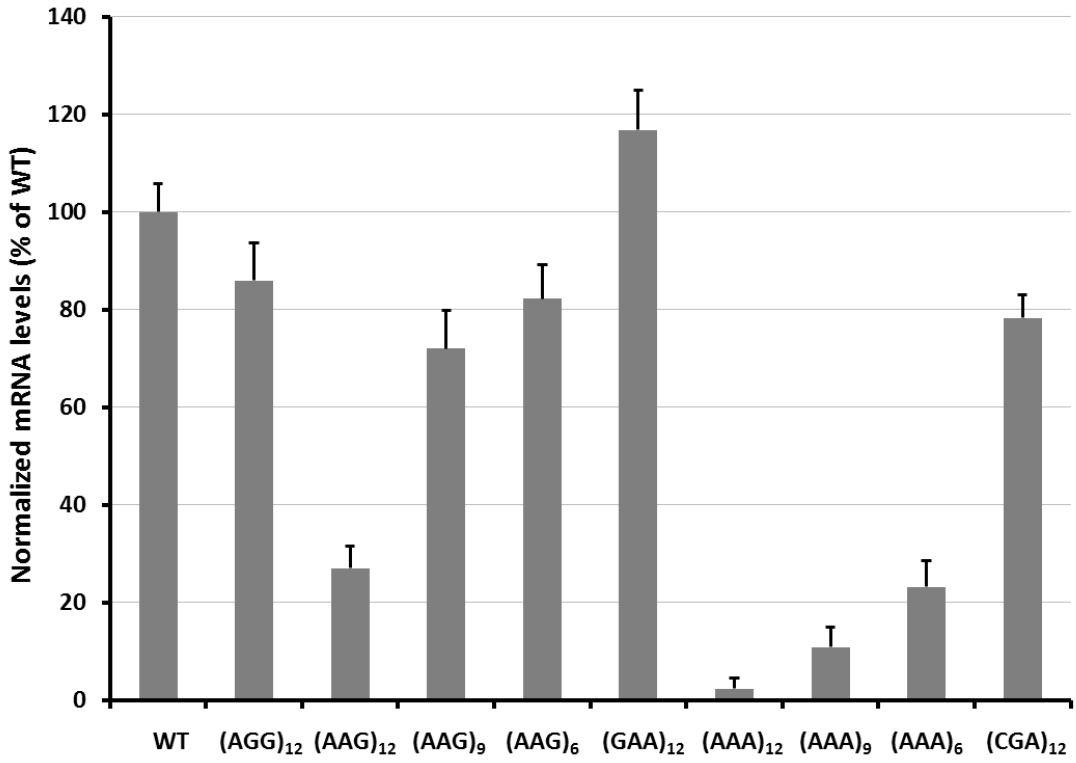


Supplemental Figure 2.1. Distribution of polyarginine (a) and polylysine (b) runs of different length in several organisms. Abundance is normalized to the number of residues of certain kind in all protein isoforms (see Methods).

(a)

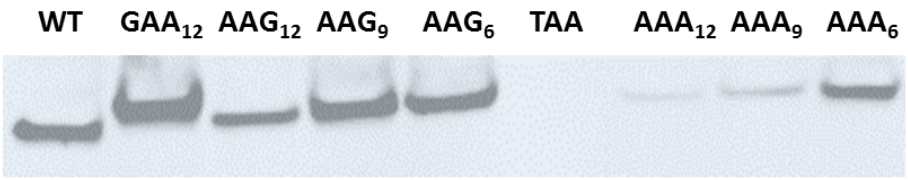


(b)

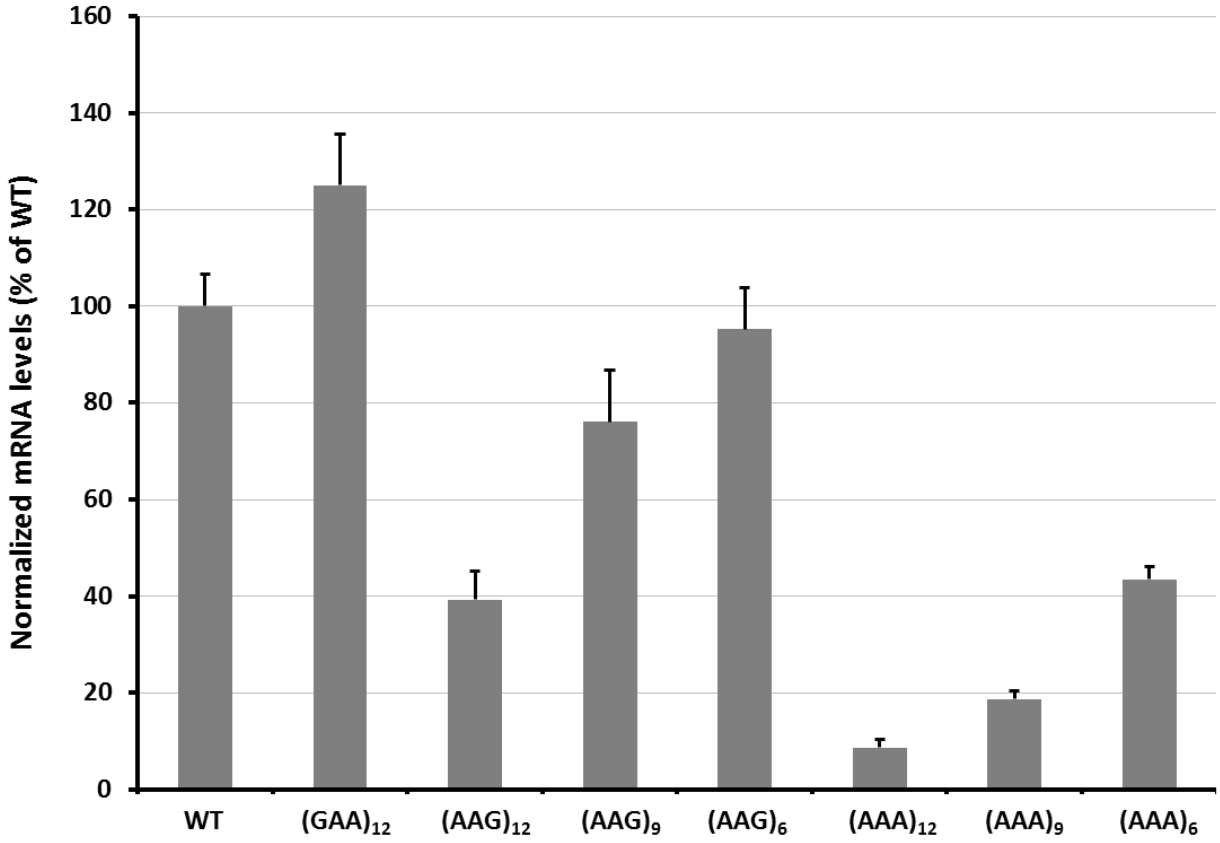


Supplemental Figure 2.2. Expression of HA-X-mCherry reporters in Chinese Hamster Ovary cells. (a) Western blot analysis of reporter expression was normalized to β-actin levels. (b) qRT-PCR analyses of mRNA abundance was normalized to neomycin resistance gene and presented as fraction of mRNA levels for mCherry construct without insert.

(a)

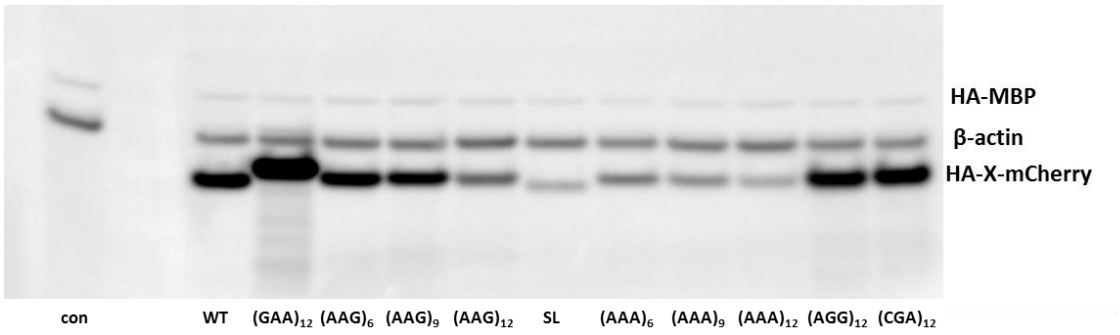


(b)

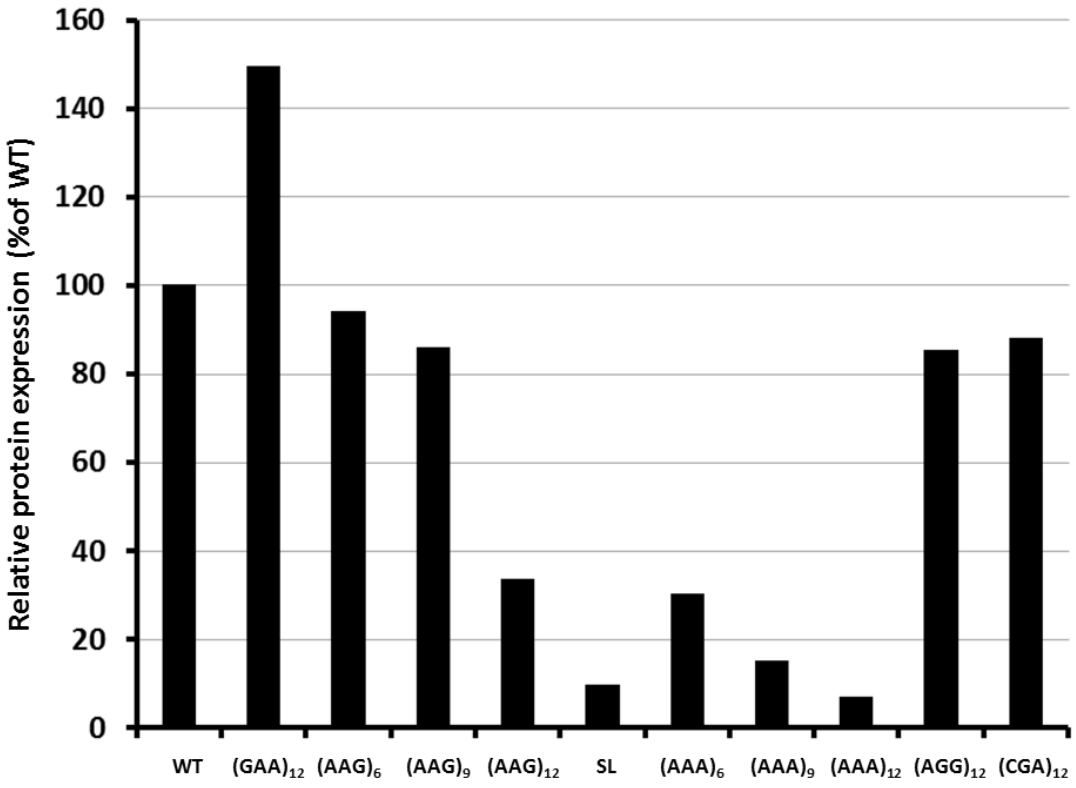


Supplemental Figure 2.3. Expression of HA-X-mCherry reporters in Drosophila S2 cells. (a) Western blot analysis of reporter expression was normalized to total protein amount. (b) qRT-PCR analyses of mRNA abundance was normalized to levels of endogenous GAPDH mRNA and presented as fraction of mRNA levels for mCherry construct without insert.

(a)

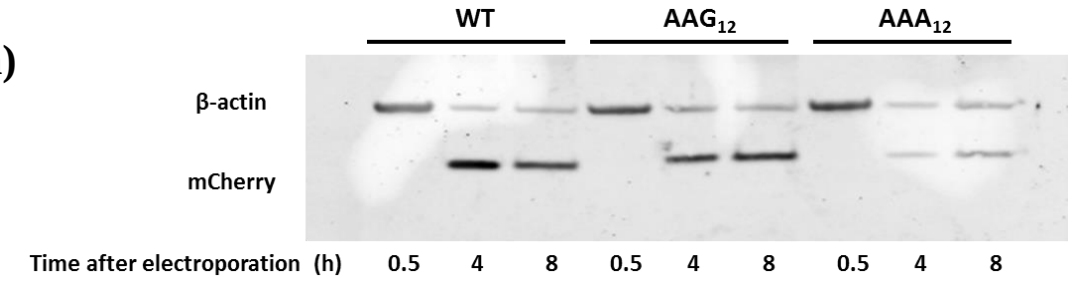

HA-MBP
β-actin
HA-X-mCherry
con
WT
(GAA)12
(AAG)6
(AAG)9
(AAG)12
SL
(AAA)6
(AAA)9
(AAA)12
(AGG)12
(CGA)12

(b)

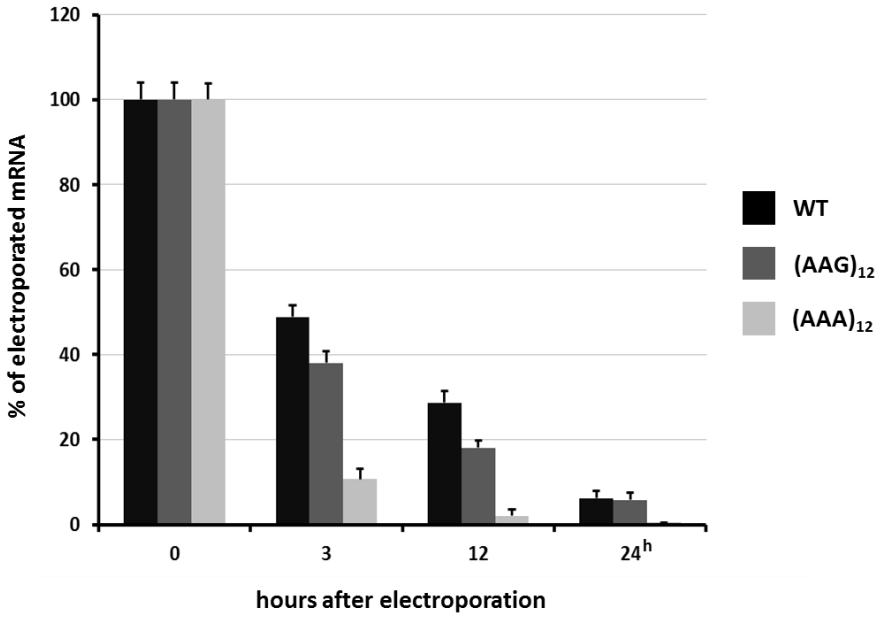

Relative protein expression (%of WT)
160
140
120
100
80
60
40
20
0
WT
(GAA)12
(AAG)6
(AAG)9
(AAG)12
SL
(AAA)6
(AAA)9
(AAA)12
(AGG)12
(CGA)12

Supplemental Figure 2.4. Expression of HA-X-mCherry reporters from T7-RNA polymerase in vitro transcribed mCherry mRNAs in HDFs. (a) HA-MBP mRNAs were in vitro transcribed and co-electroporated into HDFs as a control for electroporation efficiency and western blot normalization. β-actin was used as a control for the total protein amounts. (b)Each lane was subjected to Bio-Rad quantification analyses to determine the levels of expression shown on the graph below.

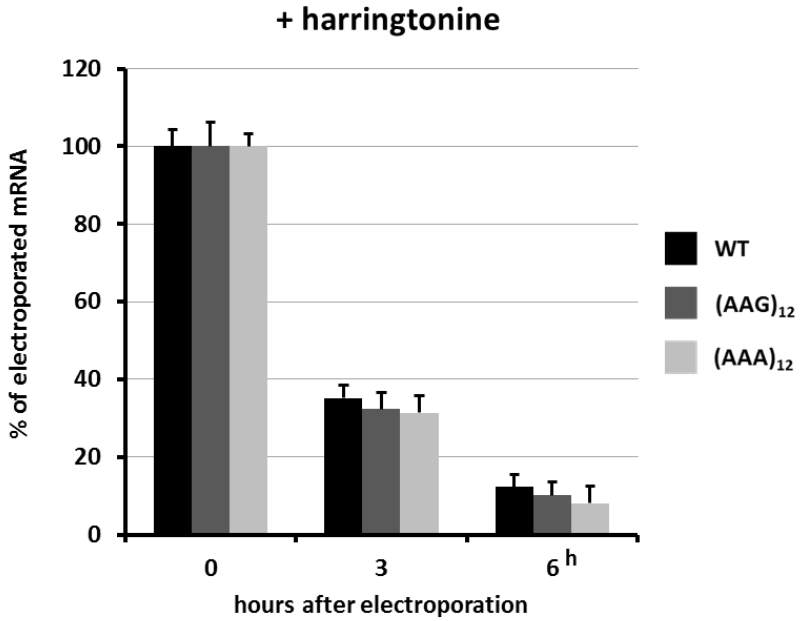
(a)


WT
AAG12
AAA12
β-actin
mCherry
Time after electroporation (h) 0.5 4 8 0.5 4 8 0.5 4 8

(b)

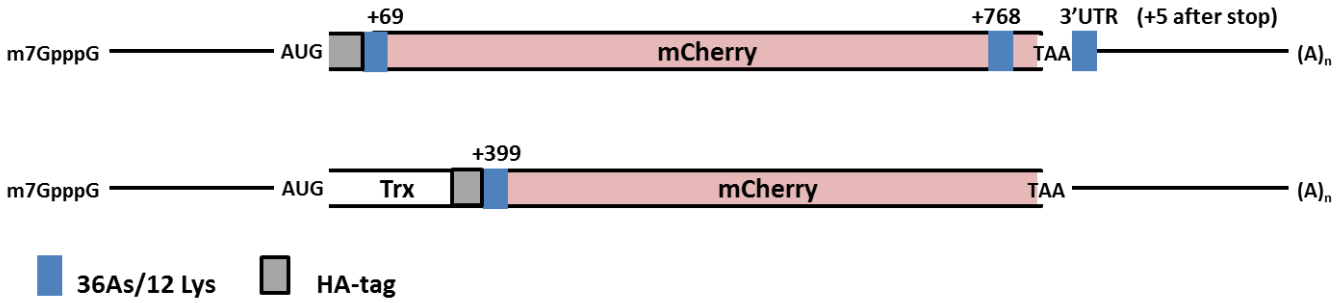

% of electroporated mRNA
120
100
80
60
40
20
0
0
3
12
24h
hours after electroporation
WT
(AAG)12
(AAA)12

(c)

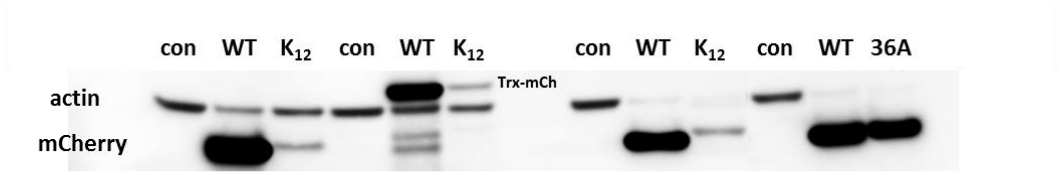

+ harringtonine
% of electroporated mRNA
120
100
80
60
40
20
0
0
3
6 h
hours after electroporation
WT
(AAG)12
(AAA)12

Supplemental Figure 2.5. Differential stability of electroporated mRNAs from HA-X-mCherry reporters is translation dependant. In vitro transcribed mCherry mRNAs were electroporated in HDFs. (a)Protein and (b)mRNA levels were assessed by western blot analyses or qRT-PCR. HA-(AAA)12-mCherry construct shows significant reduction in protein levels as well as in mRNA stability. (c) Addition of translation initiation inhibitor, herringtonine, completely abolishes effect on mRNA stability.

(a)

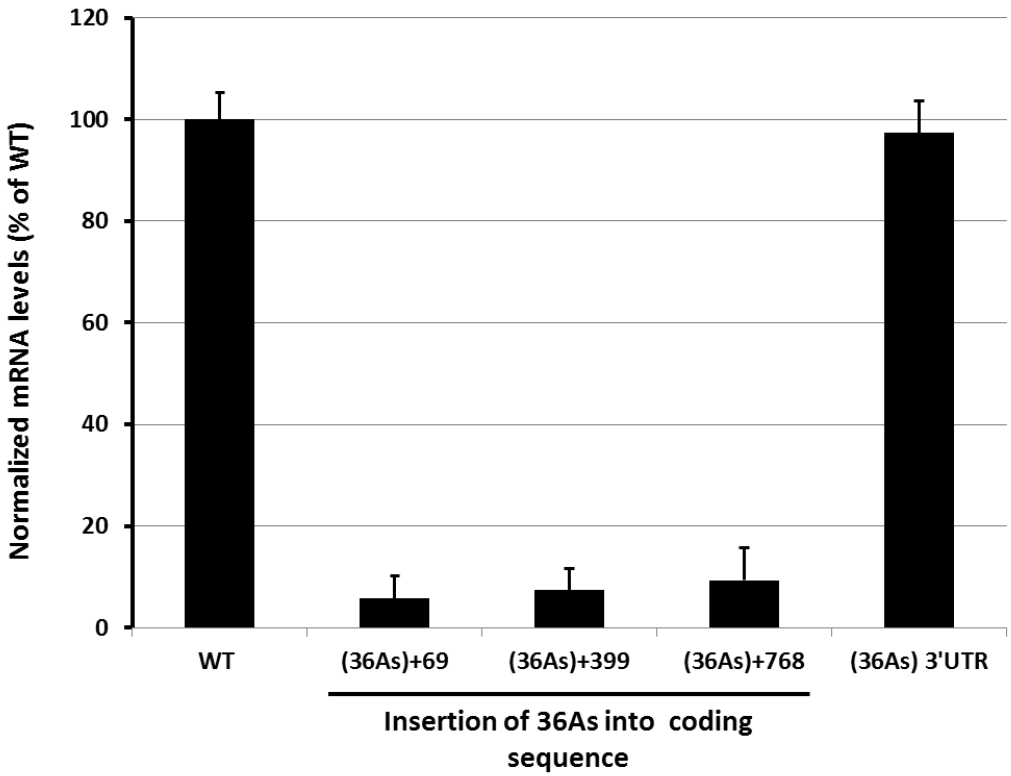


(b)



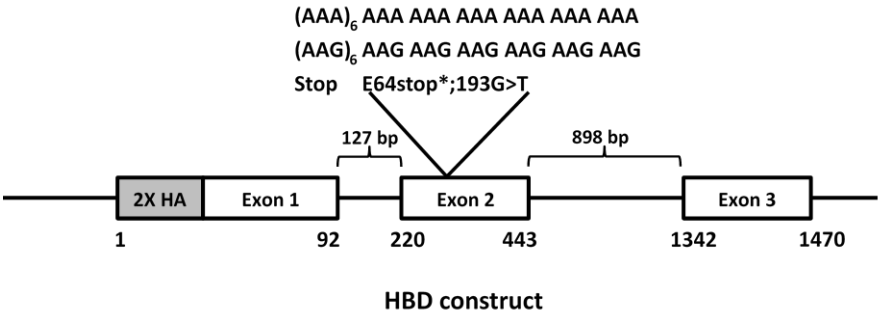
Position of inserted A_{36} tract (bp) +69 +399 * +768 3'UTR (+5 after stop)

(c)



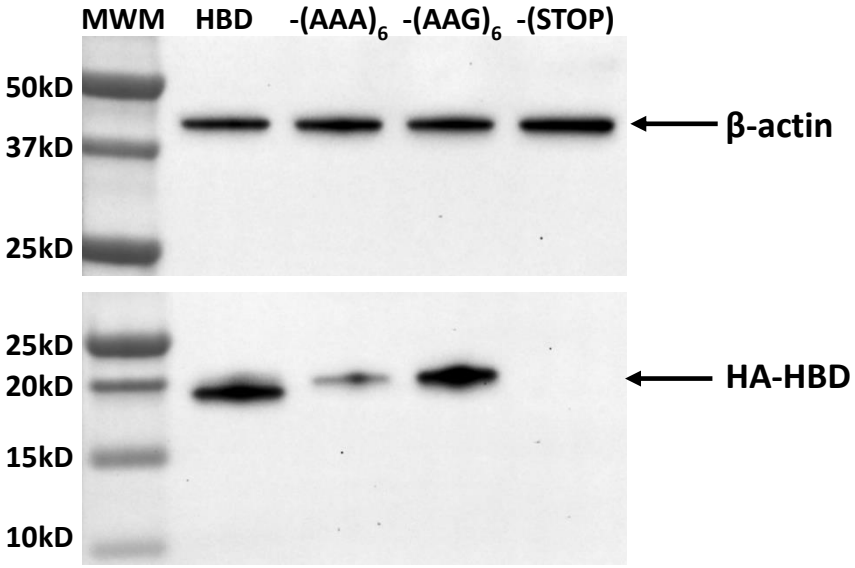
Supplemental Figure 2.6. Insertion of polylysine mCherry constructs in the coding sequence results in the same protein reduction and decreased mRNA stability. (a) Scheme of assayed mCherry constructs. Thioredoxin (Trx) fusion protein was used instead of insertion of polylysine run in the middle of mCherry gene. Positions of 12 lysine (36As) insertions and HA-tag in the constructs are labeled with blue and gray boxes, respectively. Numbers above reporter indicate distance of 36As insertion from the first nucleotide in the coding sequence. (b)Protein expression was monitored by western blot analyses using HA and beta-actin antibody. (c) qRT-PCR analyses of mRNA abundance was normalized to neomycin resistance gene and presented as fraction of mRNA levels for WT mCherry construct without insert.

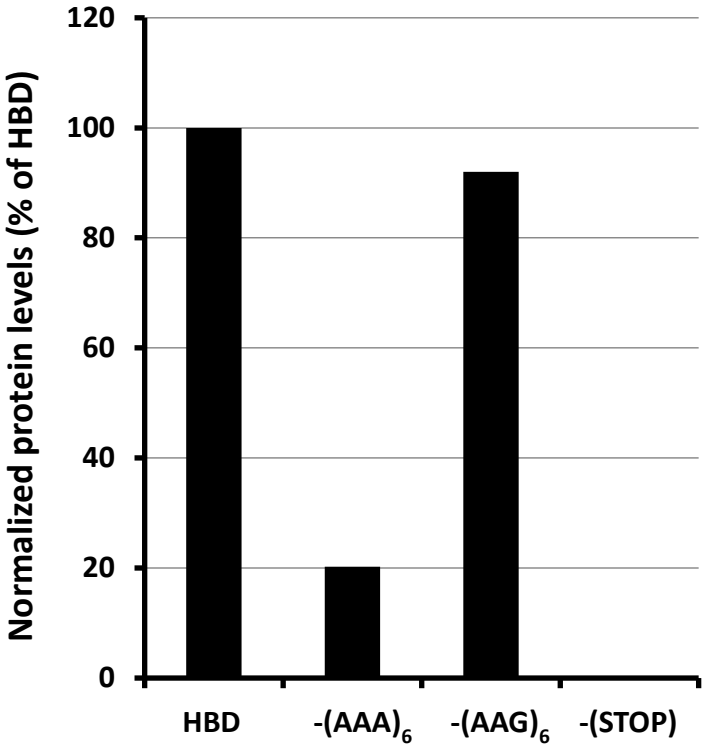
(a)



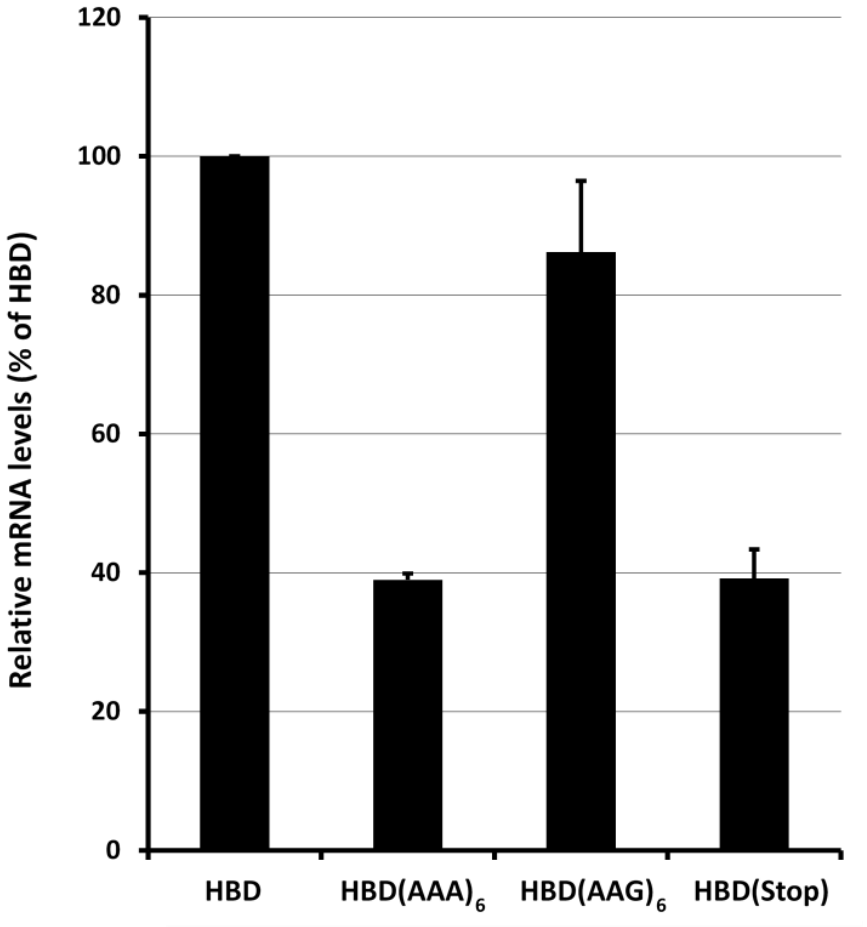
HBD construct

(b)

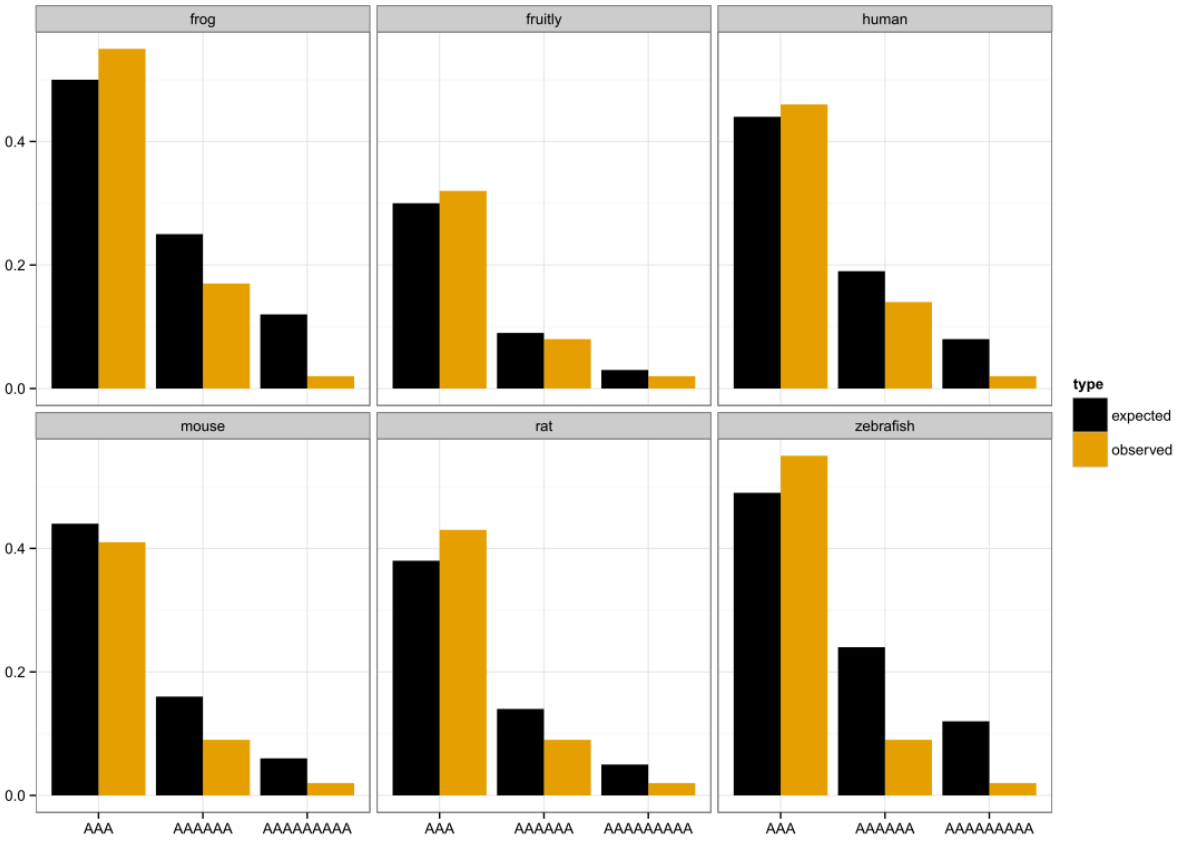




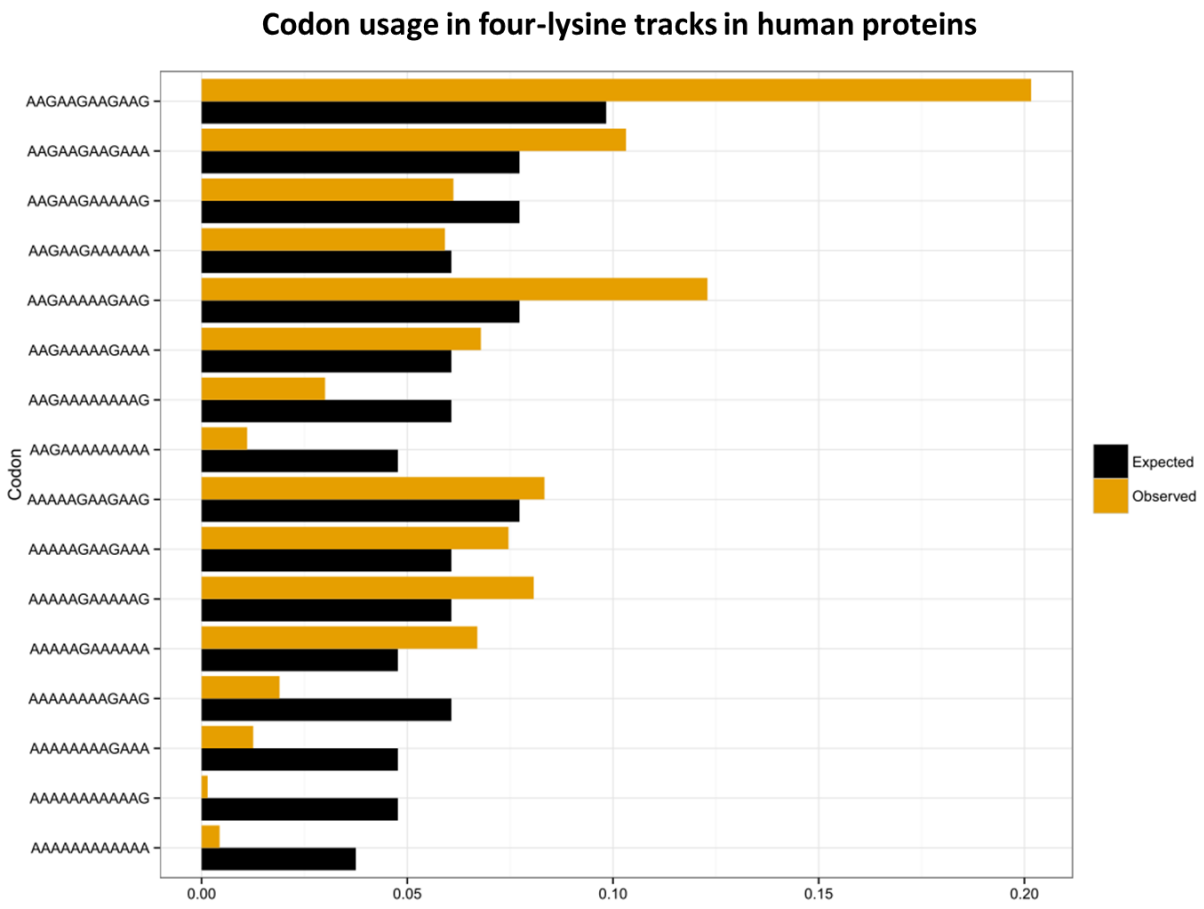
(c)



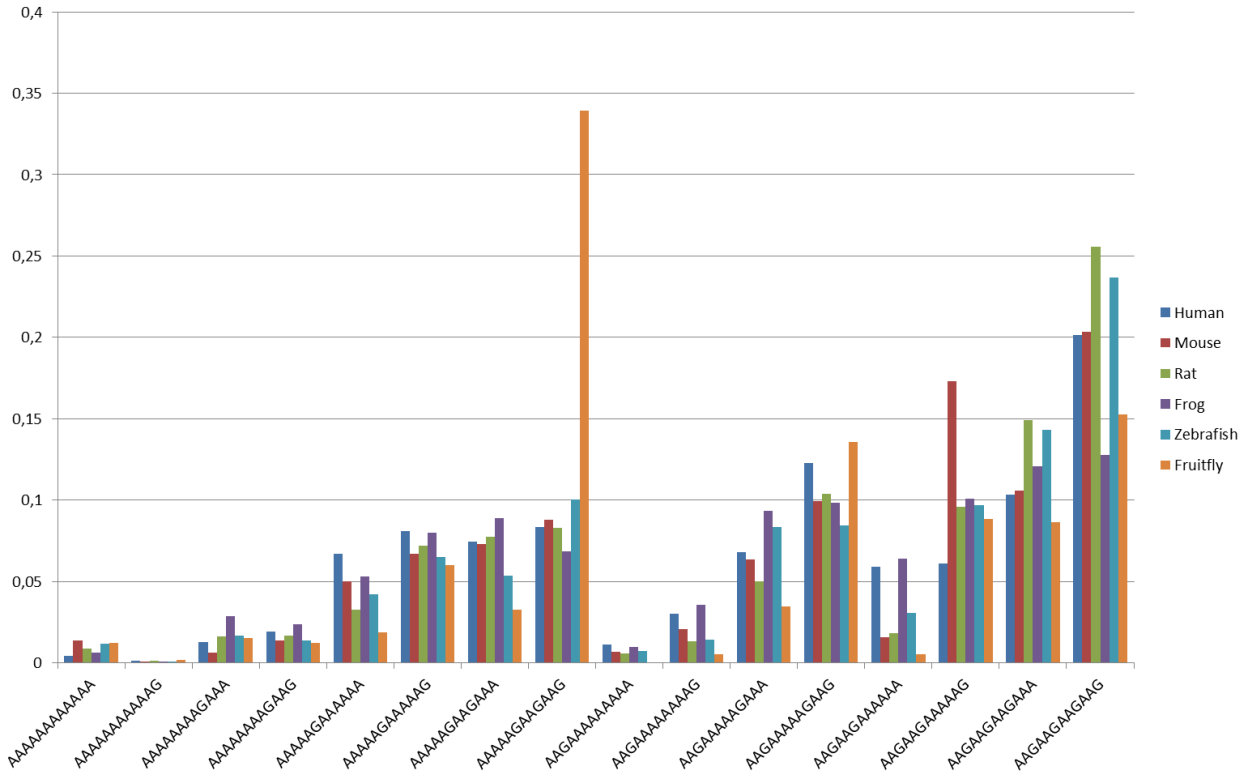
Supplemental Figure 2.7. Expression of HA-tagged beta globin (delta)(HBD) constructs with natural introns in HDF cells. (a) scheme of beta-globin gene with position of poly lysine and stop codon insertions. Position and length of introns as well as exons in HBD constructs are indicated. (b) Western blot analysis of HA-HBD construct expression normalized to β-actin levels. (c) qRT-PCR analyses of mRNA abundance was normalized to neomycin resistance gene and presented as fraction of mRNA levels for WT HBD construct without insert.



Supplemental Figure 2.8. Comparison of usage of AAA in single, double and triple lysine runs across several organisms. Expected values (black bars) are based on Kazusa database, while observed (yellow bars) are calculated from all isoforms of proteins available in NCBI RefSeq database.

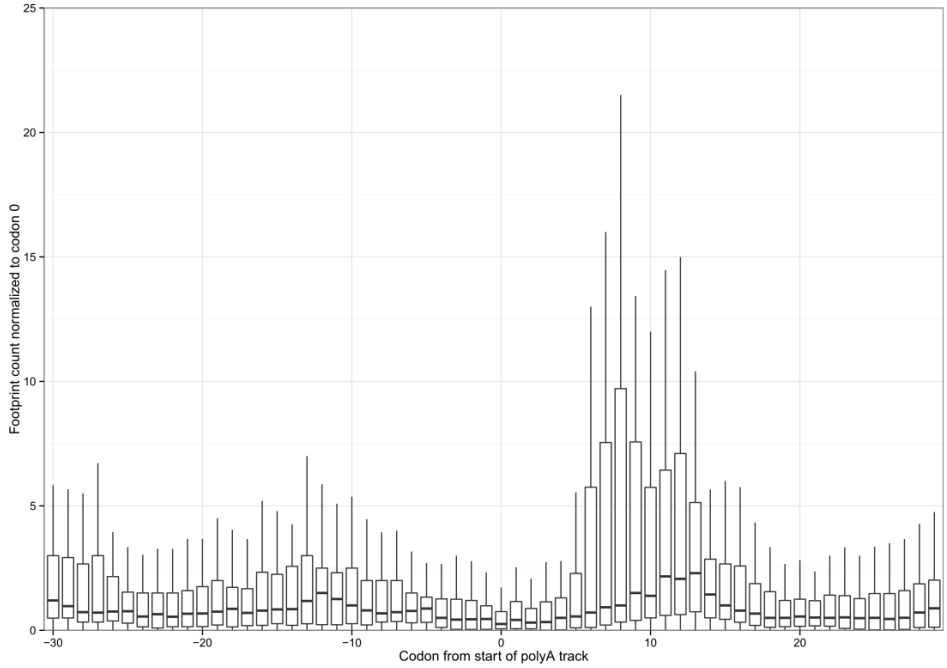


Supplemental Figure 2.9. Observed codon usage in all isoforms of human proteins vs expected (based on the proportions 0.44 to 0.56, AAA to AAG for all lysines) in the tracks of four consecutive lysines.

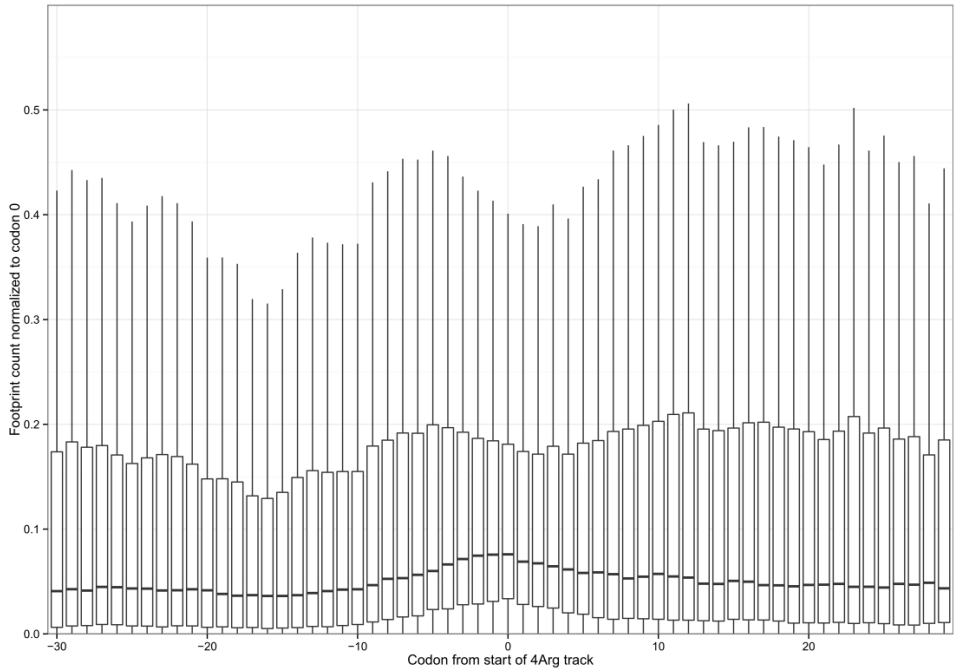


Supplemental Figure 2.10. Codon distribution in four-lysine tracks in different organisms. All protein isoforms sequences and sequences of corresponding mRNAs were taken into account. The script checks all tracks of four consecutive lysines, even when overlapping (if there is a track of five lysines, it will report two nucleotide strings of length 12).

(a)

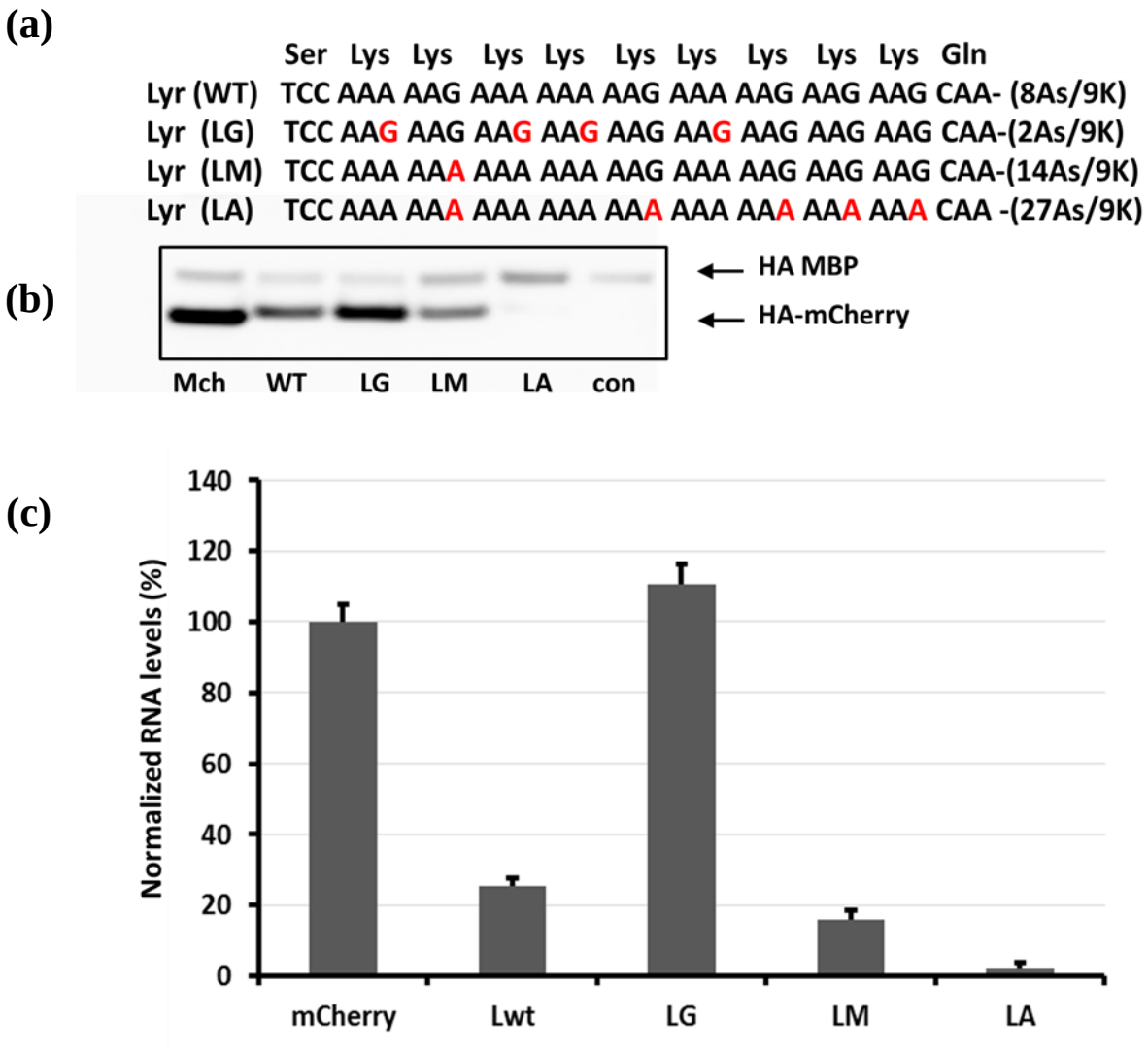

Footprint count normalized to codon 0
Codon from start of polyA track

(b)

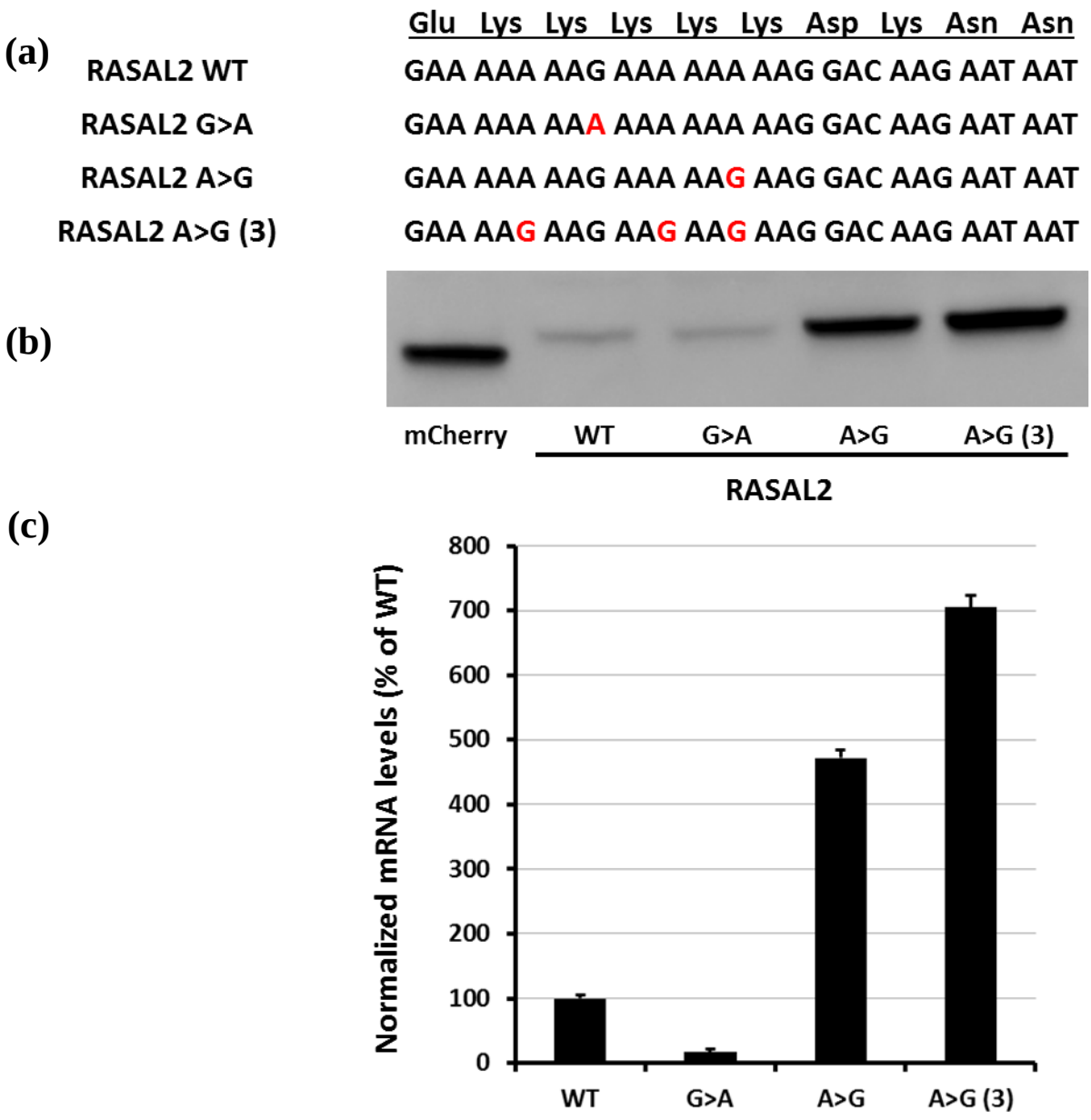

Footprint count normalized to codon 0
Codon from start of 4Arg track

Supplemental Figure 2.11. Occupancy of ribosomal footprints from three different datasets : (a) – region around polyA tracks; (b) – region around 4 arginine tracks, all codons combinations together. . The upper and lower "hinges" correspond to the first and third quartiles (the 25th and 75th percentiles). The upper and lower whiskers extend from hinges at 1.5*IQR of the respective hinge.

Supplemental Figure 2.12. Sequence conservation of RAS Activating-Like protein 2 gene (RASAL2) at DNA and protein sequence. Polylysine sequence and nucleotides forming polyA track are in indicated in red and bold letters, respectively.



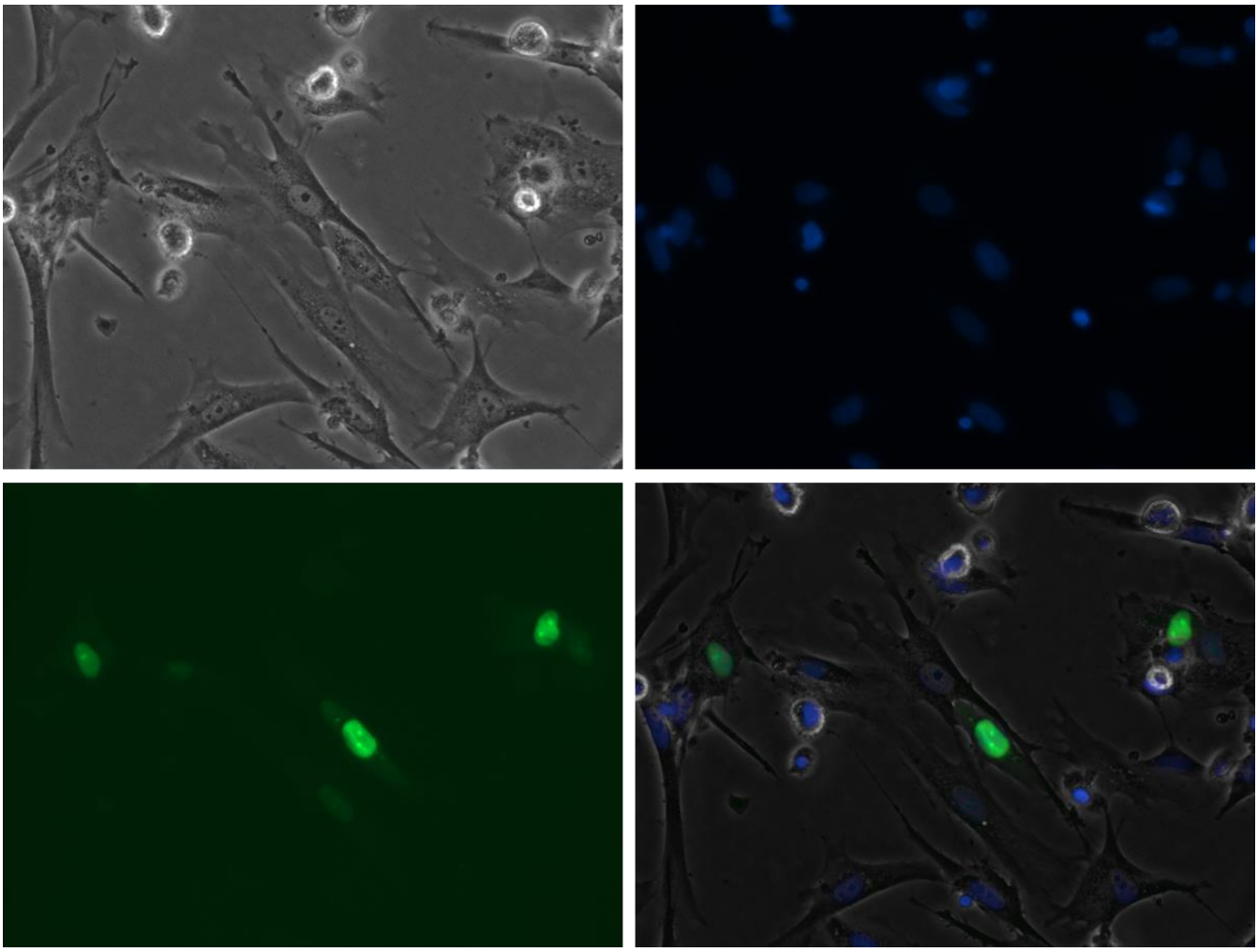
Supplemental Figure 2.13. Synonymous mutations in mCherry reporter with metadherin (MTDH, Lyric(Lyr)) polyA track. (a) scheme of reporter sequences with G>A and A>G synonymous mutations. (b) western blot analyses of reporter constructs with synonymous mutations. (c) normalized mRNA levels for reporter sequences with wil type MTDH polyA-track (Lwt) and corresponding mutants. mRNA levels are represented as fractions of wild type mCherry levels.



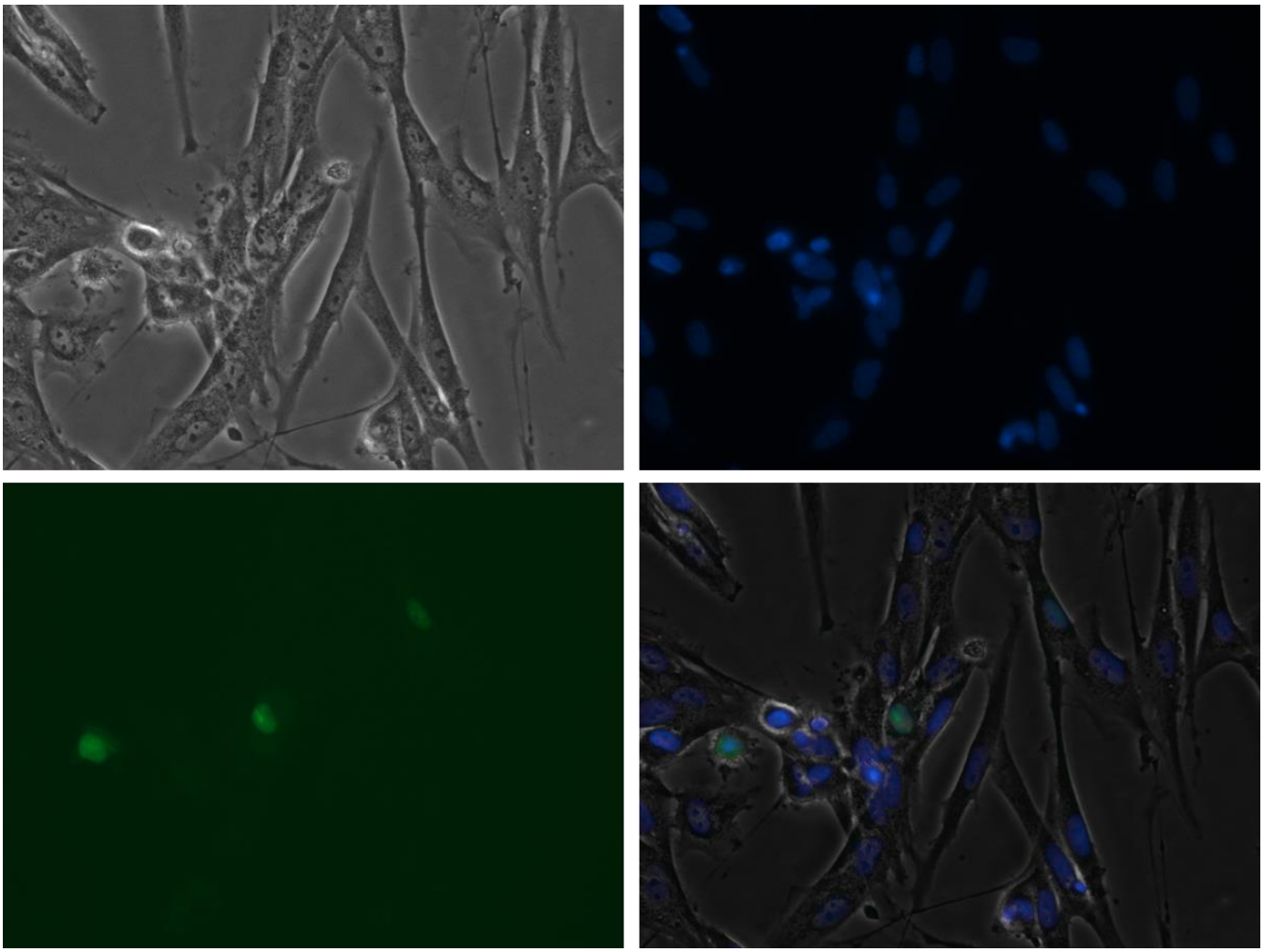
Supplemental Figure 2.14. Synonymous mutations in mCherry reporter with RASAL2 polyA track. (a) scheme of reporter sequences with G>A and A>G synonymous mutations. (b) western blot analyses of reporter constructs with synonymous mutations. (c) normalized mRNA levels

for reporter sequences with wild type RASAL2 polyA-track (Lwt) and corresponding mutants. mRNA levels are represented as fractions of wild type mCherry levels.

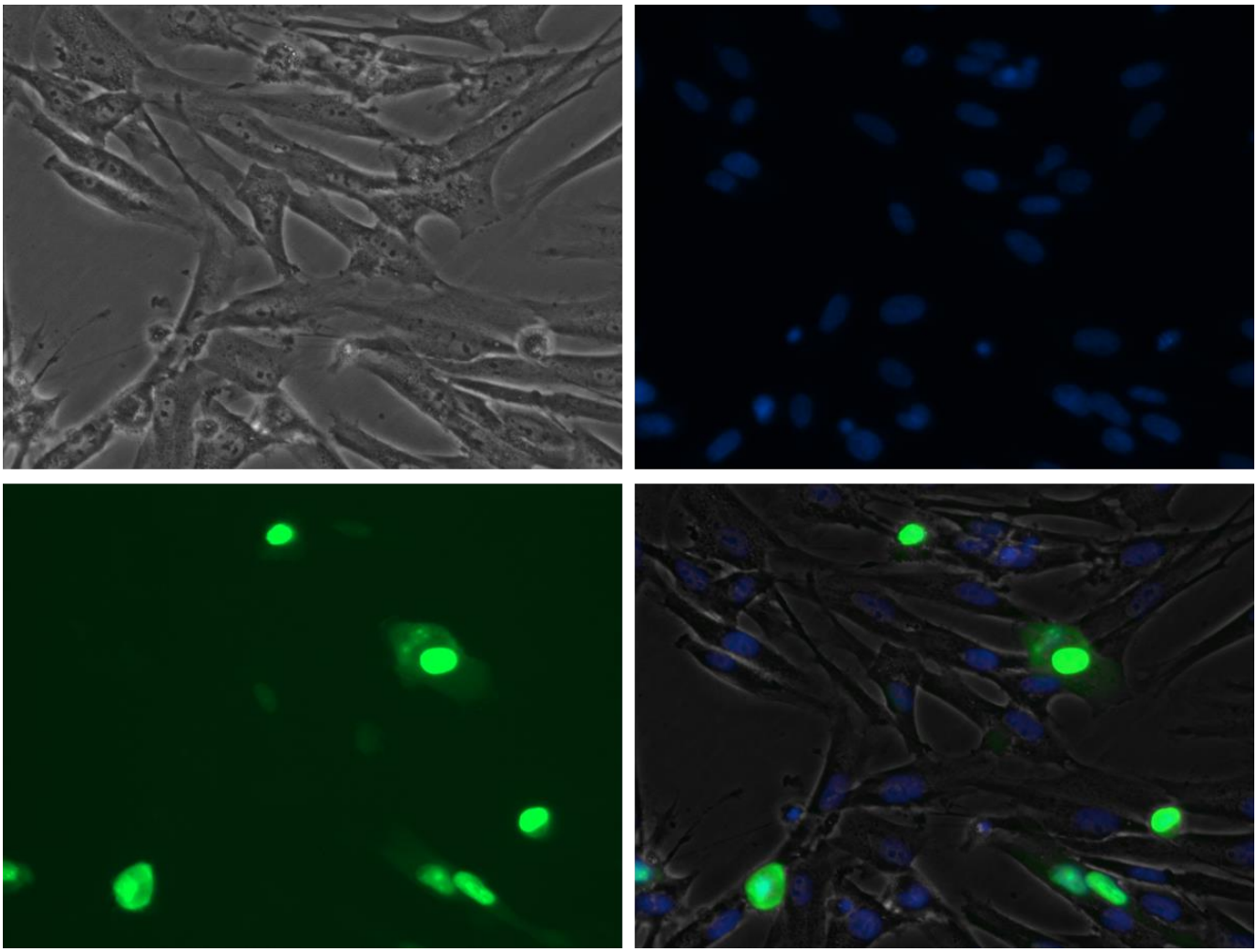
(a)



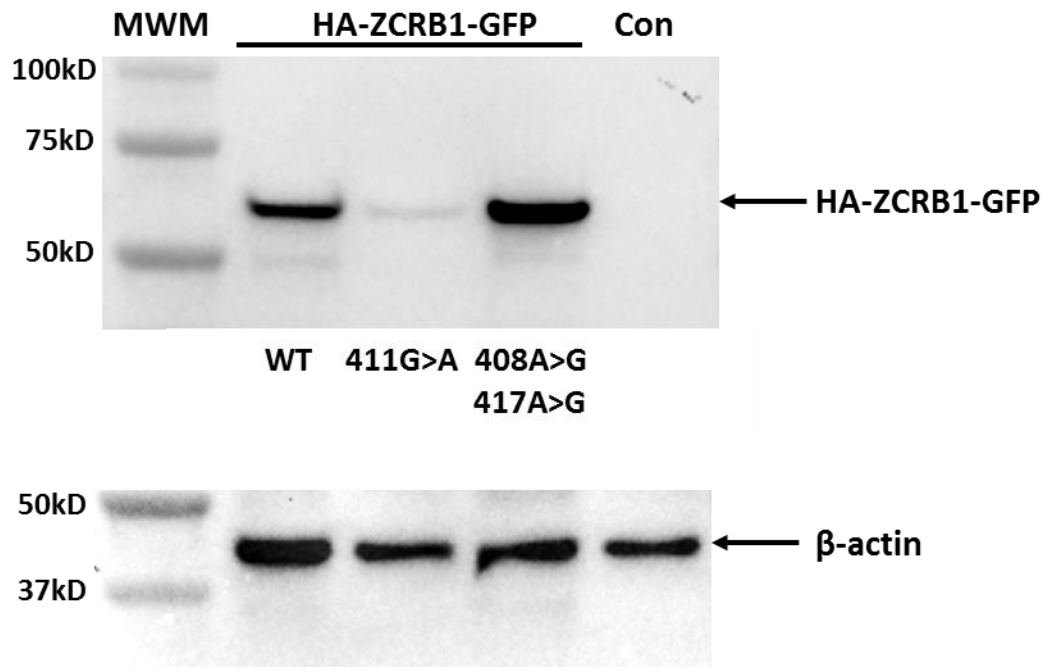
(b)



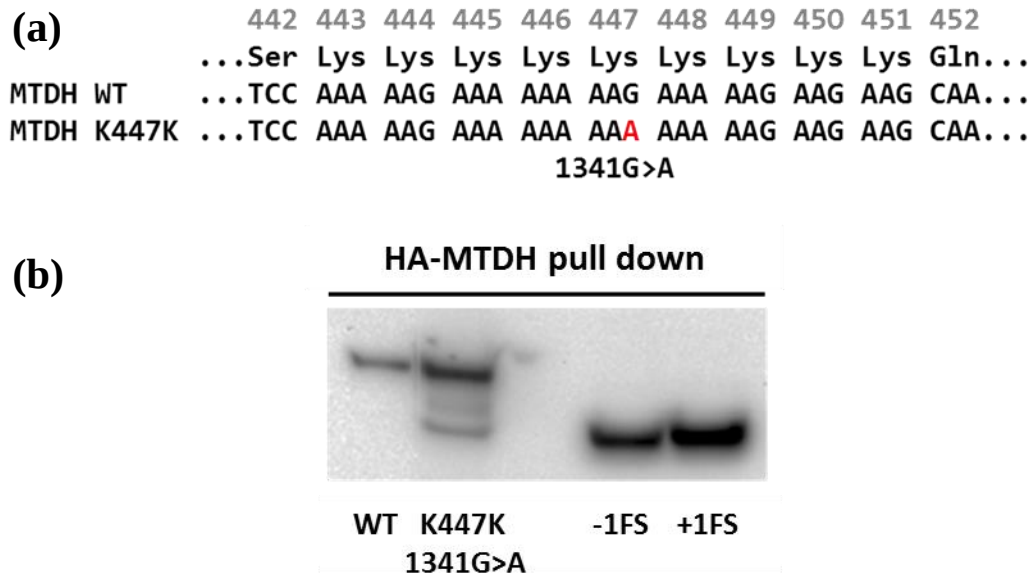
(c)



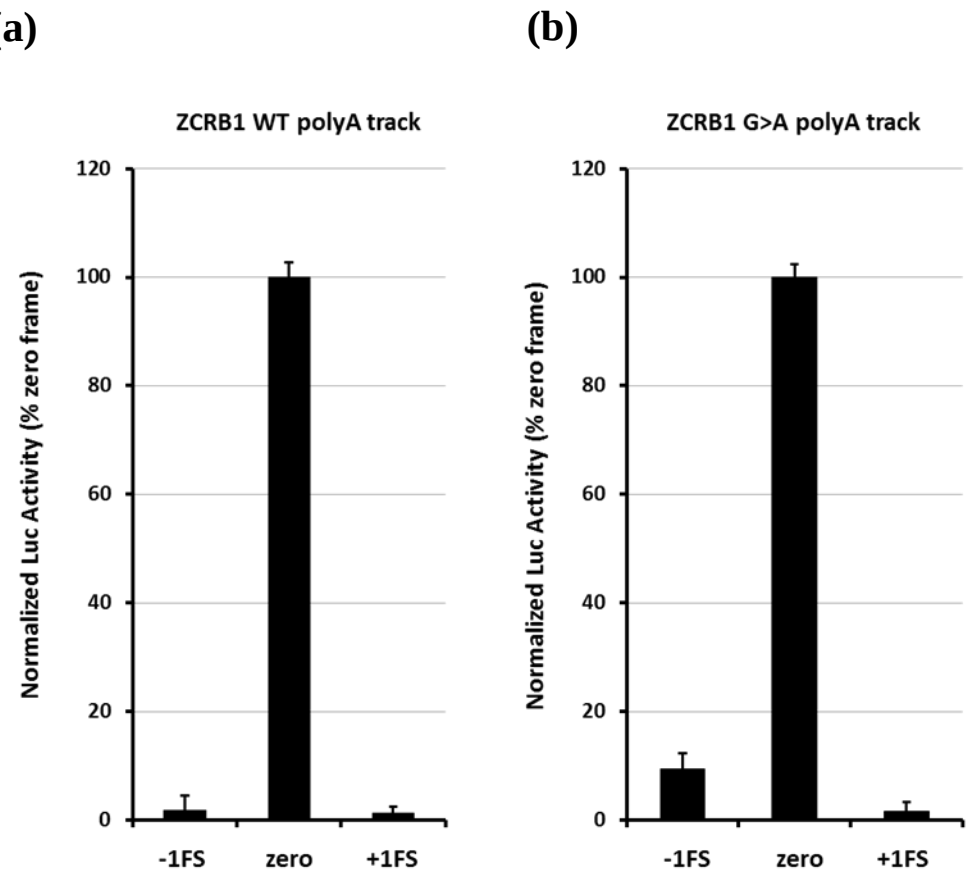
(d)



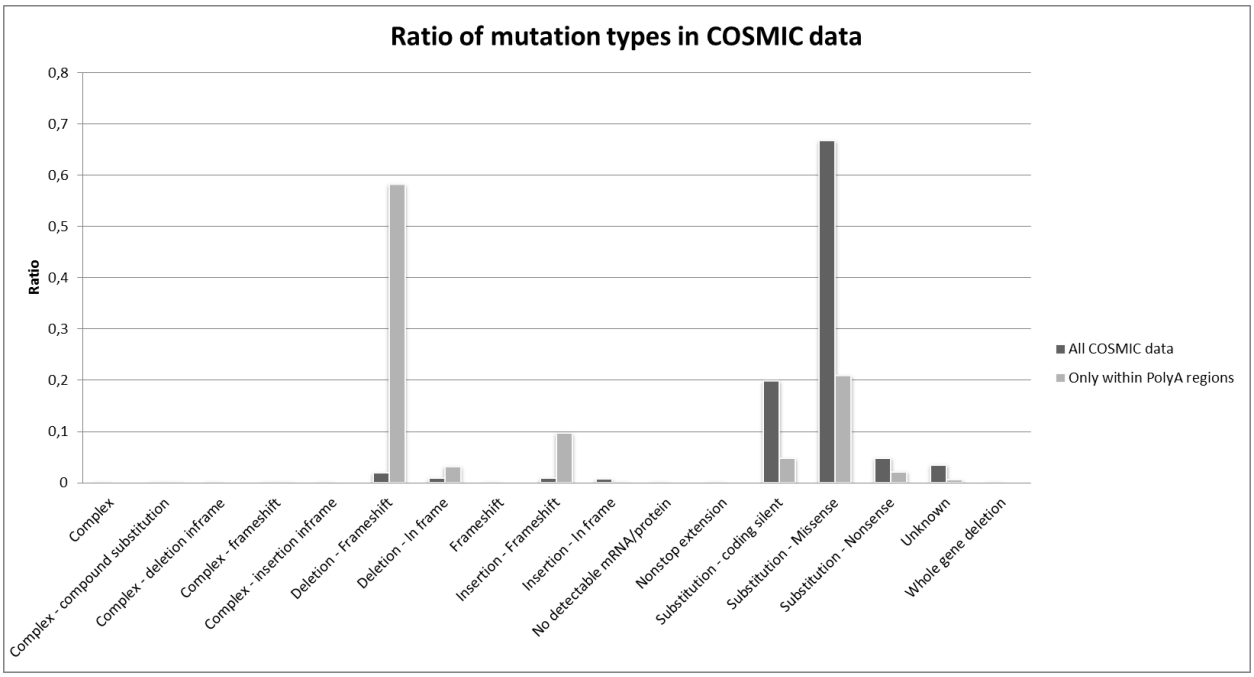
Supplemental Figure 2.15. Expression analysis of N-terminally HA- and C-terminaly GFP-tagged ZCRB1 gene and its synonymous mutants in HDF cells using Evos-FL microscop. Cell images were taken 24 hours post electroporation using same optical settings. Cell nuclei were made visible using Hoechst 33342 dye. Images of HDF cells expressing double -tagged ZCRB1 wild type (WT) protein (a) ZCRB1 K137K;411G>A (b) and ZCRB1 K136K:408 A>G; K139K:417A>G (c) mutants. Images for each channel (trans, DAPI and GFP) were taken separately and overlay image was composed using EVOS FL digital software. (d) Western blot analyses of HA-ZCRB1-GFP proteins from HDF cells using HA-antibody. Western blot analyses were normalized using beta-actin levels as loading controls.



Supplemental Figure 2.16. Introduction of COSMIC database reported synonymous mutation K447K (1341G>A) in full length recombinant MTDH gene. (a) Sequence of the wild type and K447K (G>A) mutant of MTDH gene. (b) Western blot analyses of HA-tagged WT and K474K mutant MTDH proteins. Major additional protein product corresponds to frame-shifted MTDH protein products (-1 and +1 FS) created by insertion or deletion of one nucleotide following the last nucleotide in polyA-track.



Supplemental Figure 2.17. Frame-shifting efficiency of polyA tracks from ZCRB1 WT (a) and ZCRB G>A mutant (b) measured by luciferase activity. Values for -1 and +1 frame-shifts (FS) for WT and mutant polyA track are presented as fractions of luciferase activity coming from expression from zero frame construct of WT(a) or mutant G>A sequence(b).



Supplemental Figure 2.18. Proportion of mutation types in polyA segments vs all mutation types. Data has been generated from COSMIC database. There is a dramatic shift in the distribution of mutations in polyA segments from substitutions (all COSMIC data) to frameshifts (polyA segments).

2.6 Tables

Organism	Total number of protein isoforms	Number of proteins encoded by mRNAs with polyA region	Percentage of polyA-affected protein isoforms
Saccharomyces cerevisiae	6717	37	0,5
Bos Taurus	63075	1262	2,0
Xenopus laevis	28827	923	3,2
Drosophila melanogaster	27777	317	1,1
Danio rerio	43136	784	1,8
Rattus norvegicus	60631	1036	1,7
Mus musculus	77936	1458	1,9
Homo sapiens	71479	1867	2,6

Supplemental Table 2.1. Statistics of occurrences of transcripts containing polyA tracks in different organisms.

Term	Background frequency	Sample frequency	Expected	+/-	P-value
nucleic acid binding (GO:0003676)	3799	149	7.614e+01	+	6.651e-15
heterocyclic compound binding (GO:1901363)	5648	189	1.132e+02	+	5.339e-13
RNA binding (GO:0003723)	1500	79	3.006e+01	+	7.430e-13
organic cyclic compound binding (GO:0097159)	5717	190	1.146e+02	+	8.553e-13
poly(A) RNA binding (GO:0044822)	1122	62	2.249e+01	+	1.388e-10
binding (GO:0005488)	12444	316	2.494e+02	+	5.325e-09
DNA binding (GO:0003677)	2322	87	4.654e+01	+	1.491e-06
protein binding (GO:0005515)	8307	219	1.665e+02	+	3.725e-05
ion binding (GO:0043167)	5844	166	1.171e+02	+	3.815e-05
chromatin binding (GO:0003682)	409	26	8.197e+00	+	6.569e-05
zinc ion binding (GO:0008270)	1181	49	2.367e+01	+	2.713e-04
molecular_function (GO:0003674)	15480	353	3.103e+02	+	3.165e-04
transition metal ion binding (GO:0046914)	1417	52	2.840e+01	+	3.844e-03
ATP binding (GO:0005524)	1430	52	2.866e+01	+	4.880e-03
nucleotide binding (GO:0000166)	2264	73	4.538e+01	+	6.077e-03
nucleoside phosphate binding (GO:1901265)	2265	73	4.540e+01	+	6.164e-03
adenyl ribonucleotide binding (GO:0032559)	1465	52	2.936e+01	+	9.081e-03
adenyl nucleotide binding (GO:0030554)	1483	52	2.972e+01	+	1.235e-02
protein serine/threonine kinase activity (GO:0004674)	434	22	8.698e+00	+	1.523e-02
nucleoside-triphosphatase activity (GO:0017111)	707	30	1.417e+01	+	2.103e-02

purine ribonucleoside triphosphate binding (GO:0035639)	1760	58	3.527e+01	+	2.439e-02
purine ribonucleotide binding (GO:0032555)	1801	59	3.610e+01	+	2.486e-02
purine ribonucleoside binding (GO:0032550)	1769	58	3.545e+01	+	2.783e-02
purine nucleoside binding (GO:0001883)	1772	58	3.551e+01	+	2.908e-02
ribonucleoside binding (GO:0032549)	1773	58	3.553e+01	+	2.950e-02
ribonucleotide binding (GO:0032553)	1816	59	3.640e+01	+	3.087e-02
purine nucleotide binding (GO:0017076)	1821	59	3.650e+01	+	3.315e-02
nucleoside binding (GO:0001882)	1783	58	3.574e+01	+	3.409e-02
protein kinase activity (GO:0004672)	591	26	1.184e+01	+	3.421e-02
helicase activity (GO:0004386)	145	11	2.906e+00	+	3.556e-02
small molecule binding (GO:0036094)	2539	76	5.089e+01	+	4.353e-02

Supplemental Table 2.2. Gene Ontology terms for polyA track genes. Overrepresentation of Gene Ontology terms for 456 genes containing polyA tracks in their coding regions up to P-value of 0.05.

mRNA GI number	end of polyA	nearest stop codon	location of stop codon	location of downstream intron-exon boundary
315360663	465	TAG	472	491
315360663	465	TGA	467	491
54792083	710	TAG	839	849
61743937	294	TAA	310	327
61743937	294	TAG	324	327
61743937	294	TGA	306	327
41350197	346	TGA	385	398
42542385	1274	TAA	1312	1342
115430109	97	TAA	139	171
115430109	97	TAG	321	340
115430109	97	TGA	162	171
38327633	431	TGA	452	470
148596997	396	TGA	411	416
146133847	192	TAA	473	499
146133847	192	TAG	200	246
189163500	2545	TAG	2788	2813
214010231	475	TAA	748	780
284795238	986	TAG	997	1014
284795238	986	TGA	1001	1014
332205944	3430	TAA	3688	3732
283549152	927	TGA	996	1015
325197190	472	TAA	484	515
325197190	472	TGA	502	515
325197194	1713	TAA	1725	1756
325197194	1713	TGA	1743	1756
325197181	768	TAA	780	811
325197181	768	TGA	798	811
332205942	3451	TAA	3709	3753
378786661	207	TAA	304	352
378786661	207	TAG	351	352
378786661	207	TGA	364	408
367460089	3008	TAA	3256	3288
378786660	315	TAA	412	460
378786660	315	TAG	459	460
378786660	315	TGA	472	516
387528005	1266	TAG	1416	1437
57863258	885	TAG	938	947
380420368	574	TAG	821	842
527317397	1917	TGA	1941	1958
543173130	2544	TAG	2847	2872
347954818	953	TGA	1010	1038
564473350	2511	TAA	2513	2521
585420399	1248	TAG	1381	1407
585420411	916	TAG	1049	1075
256542287	192	TAA	473	499
256542287	192	TAG	200	246
223634474	186	TAA	467	493
223634474	186	TAG	194	240
585866348	1248	TAG	1381	1407
110349755	889	TGA	926	951
110349753	889	TGA	926	951
544583528	866	TAA	980	1012
93277100	828	TAA	855	903
145699132	2771	TAA	2772	2777
93141224	645	TAG	649	667
527317380	612	TAA	693	696
32967604	1594	TAG	1683	1697
72377390	216	TAG	281	318
153251835	867	TAA	901	905
153251835	867	TGA	872	905
215272324	361	TGA	421	464
46255020	219	TGA	226	247
116063563	3480	TAA	3534	3538
116063563	3480	TGA	3509	3538
194473999	931	TAA	998	1004
56550119	511	TAA	525	559
56550119	511	TAG	531	559
564473377	1659	TAA	1661	1669
151301227	2638	TAA	2698	2723
451172080	719	TAG	775	808
451172080	719	TGA	770	808
94538358	192	TAA	473	499
94538358	192	TAG	200	246
119637838	1257	TAG	1275	1304
119637838	1257	TGA	1288	1304
49640008	6092	TAA	6105	6154
49640008	6092	TGA	6114	6154
49640010	4801	TAA	4814	4863
49640010	4801	TGA	4823	4863
239787903	1353	TAA	1428	1472
239787903	1353	TAG	1469	1472

239787903	1353	TGA	1442	1472
55749880	458	TGA	519	549
225579091	2672	TAA	2683	2705
225579091	2672	TGA	2695	2705
125987600	747	TAA	858	889
187828563	1100	TAA	1112	1153
451172082	569	TGA	733	764
451172084	533	TAG	589	622
451172084	533	TGA	584	622
451172086	533	TGA	697	728
270483794	1195	TGA	1196	1203
209870068	878	TAA	940	985
72377376	305	TAG	370	407
63054847	785	TGA	789	820
573459699	766	TAA	781	801
573459699	766	TGA	893	915
116642876	1399	TAG	1450	1493
116642876	1399	TGA	1457	1493
188219548	1077	TAA	1084	1113
188219548	1077	TAG	1103	1113
544583488	906	TAA	1020	1052
543173126	2619	TAG	2922	2947
544346128	1063	TAG	1145	1189
544346128	1063	TGA	1232	1257
170650704	667	TAA	807	830
170650704	667	TAG	798	830
544583450	918	TAA	1032	1064
289547540	894	TAA	992	1033
367460086	2981	TAA	3229	3261
51243064	631	TGA	638	656
215599550	501	TAG	662	689
475505353	3017	TAA	3256	3290
122937226	703	TAA	731	738
122937226	703	TAG	855	869
122937395	987	TAA	1025	1028
122937395	987	TAG	1032	1072
122937395	987	TGA	1020	1028
57863256	971	TAG	1024	1033
451172081	569	TAG	625	658
451172081	569	TGA	620	658
87298936	1370	TAA	1453	1487
138175816	1579	TAA	1583	1610
138175816	1579	TGA	1580	1610
150036261	475	TAA	575	602
150036261	475	TGA	476	493
315360661	566	TAG	573	592
315360661	566	TGA	568	592
38373672	3210	TAG	3474	3489
38373672	3210	TGA	3244	3290
544346132	1174	TAG	1256	1300
544346132	1174	TGA	1343	1368
116805347	219	TGA	261	278
157502183	1552	TAA	1882	1906
23111022	714	TAG	721	740
23111022	714	TGA	716	740
32967602	1594	TAG	1683	1697
573459727	352	TAA	367	387
573459727	352	TGA	479	501
573459714	617	TAA	632	652
573459714	617	TGA	744	766
574275032	622	TGA	629	633
574275429	622	TGA	629	633
574275427	622	TGA	629	633
574275776	622	TGA	629	633
574275778	557	TGA	564	568
169234948	317	TAG	360	370
169234948	317	TGA	336	370
557636701	3791	TAA	3834	3870
557636701	3791	TGA	3826	3870
344925844	1128	TGA	1241	1282
115298681	736	TGA	746	787
61743939	295	TAA	310	327
61743939	295	TAG	324	327
61743939	295	TGA	306	327
289547543	894	TAA	992	1033
7705934	473	TAG	718	750
209870072	451	TAA	514	559
209870078	917	TAA	979	1024
209870070	704	TAA	766	811
215599561	320	TAG	481	508
544583540	918	TAA	1032	1064
284172494	1574	TAA	1651	1685
239582713	995	TAG	1343	1371
225579094	2815	TAA	2826	2848

225579094	2815	TGA	2838	2848
110347419	1175	TAA	1181	1221
110347419	1175	TAG	1192	1221
110347419	1175	TGA	1187	1221
284795233	1121	TAG	1132	1149
284795233	1121	TGA	1136	1149
544583452	918	TAA	1032	1064
399498564	848	TAG	902	943
399498564	848	TGA	906	943
399498565	851	TAA	863	904
223005861	4124	TGA	4130	4134
155029543	670	TAG	810	848
155029541	801	TAG	941	979
157739935	1035	TAA	1036	1063
38202203	1157	TAA	1292	1305
388240807	1129	TAA	1135	1175
388240807	1129	TAG	1146	1175
388240807	1129	TGA	1141	1175
67782329	488	TAG	1321	1324
542133168	2773	TAA	2877	2887
542133168	2773	TAG	2838	2887
542133173	2334	TAA	2438	2448
542133173	2334	TAG	2399	2448
542133166	2690	TAA	2794	2804
542133166	2690	TAG	2755	2804
542133167	2827	TAA	2931	2941
542133167	2827	TAG	2892	2941
542133171	2646	TAA	2750	2760
542133171	2646	TAG	2711	2760
542133176	1321	TAA	1425	1435
542133176	1321	TAG	1386	1435
542133175	1388	TAA	1492	1502
542133175	1388	TAG	1453	1502
542133172	2571	TAA	2675	2685
542133172	2571	TAG	2636	2685
365192531	3074	TAA	3322	3354
239787905	1468	TAA	1543	1587
239787905	1468	TAG	1584	1587
239787905	1468	TGA	1557	1587
393185909	180	TAG	191	237
393185909	180	TGA	216	237
542133170	2639	TAA	2743	2753

542133170	2639	TAG	2704	2753
95147341	3226	TAA	3230	3264
95147341	3226	TAG	3238	3264
95147341	3226	TGA	3247	3264
34577113	1789	TAG	1857	1898
542133169	2639	TAA	2743	2753
542133169	2639	TAG	2704	2753
422398891	320	TAA	417	438
422398891	320	TAG	526	529
284795244	1103	TAG	1114	1131
284795244	1103	TGA	1118	1131
301500638	762	TAA	887	905
151301203	1982	TAA	2027	2050
151301203	1982	TGA	2030	2050
573459695	1552	TAA	1882	1906
422398886	320	TAA	417	438
296080784	530	TGA	663	703
300244517	499	TAA	535	547
300244517	499	TAG	517	547
300244517	499	TGA	502	547
296278216	1064	TAA	1146	1187
527317401	2486	TGA	2588	2594
313661424	3736	TGA	3751	3795
52145311	2119	TAA	2152	2183
52145311	2119	TGA	2139	2183
300797236	693	TAG	800	808
422398893	337	TAA	434	455
422398893	337	TAG	543	546
574274904	622	TGA	629	633
296278222	341	TAA	423	464
422398882	337	TAA	434	455
544346208	1454	TAG	1536	1580
544346208	1454	TGA	1623	1648
291084569	1795	TAG	2235	2249
291084569	1795	TGA	1825	1860
284795242	1103	TAG	1114	1131
284795242	1103	TGA	1118	1131
313661429	3451	TGA	3466	3510
324711032	1276	TAG	1294	1323
324711032	1276	TGA	1307	1323
313661427	3610	TGA	3625	3669
555943904	605	TAG	608	654

555943904	605	TGA	718	744
301500639	738	TAA	863	881
302191685	381	TAA	404	441
422398895	337	TAA	434	455
422398895	337	TAG	543	546
422398884	320	TAA	417	438
422398884	320	TAG	544	547
300244513	239	TAA	275	287
300244513	239	TAG	257	287
300244513	239	TGA	242	287
313661433	3604	TGA	3619	3663
300244516	430	TAA	466	478
300244516	430	TAG	448	478
300244516	430	TGA	433	478
527317400	2615	TGA	2717	2723
422398870	337	TAA	434	455
422398870	337	TAG	561	564
555943722	605	TAG	608	654
555943722	605	TGA	718	744
284004933	616	TAA	618	625
296278218	560	TAA	642	683
574275184	577	TGA	584	588
574275536	666	TGA	673	677
386781586	575	TAG	592	638
300797255	693	TAG	800	808
386781549	946	TAG	963	1009
557947981	412	TAA	718	752
557947981	412	TGA	430	435
110347424	1113	TAA	1119	1159
110347424	1113	TAG	1130	1159
110347424	1113	TGA	1125	1159
110347417	1104	TAA	1110	1150
110347417	1104	TAG	1121	1150
110347417	1104	TGA	1116	1150
53729338	2711	TAA	2714	2749
53729336	2290	TAA	2293	2328
94538369	841	TAG	883	905
94538369	841	TGA	857	905
160707983	2206	TAA	2238	2272
160707983	2206	TAG	2264	2272
154146192	1034	TAA	1086	1093
154146192	1034	TGA	1050	1093

145275203	1841	TAA	1907	1911
154146222	954	TAA	1006	1013
154146222	954	TGA	970	1013
120953299	4732	TAG	4829	4832
120953299	4732	TGA	4784	4832
189409139	1021	TAA	1074	1109
189409139	1021	TAG	1096	1109
239787837	233	TAG	377	396
239787843	177	TAG	321	340
217035143	1437	TAG	1467	1509
188035876	1539	TAA	1553	1558
188035876	1539	TGA	1545	1558
312283645	1473	TAA	1503	1531
291084574	1802	TAG	2242	2256
291084574	1802	TGA	1832	1867
312434029	1760	TAA	1790	1818
21536424	476	TAA	508	553
21536424	476	TGA	515	553
312283632	1760	TAA	1790	1818
393185915	487	TAG	498	544
393185915	487	TGA	523	544
187761329	1522	TGA	1553	1582
557786140	263	TAA	265	269
296278211	1102	TAA	1184	1225
296278220	493	TAA	575	616
257467646	2736	TGA	2785	2797
385648269	3155	TAA	3253	3263
385648269	3155	TAG	3260	3263
27894332	928	TAG	1071	1117
27894332	928	TGA	942	981
350606366	375	TAA	464	481
350606366	375	TAG	434	481
317008579	4447	TGA	4458	4488
209915559	770	TAG	825	852
296278213	618	TAA	700	741
350606370	621	TAA	710	727
350606370	621	TAG	680	727
350606369	438	TAA	527	544
350606369	438	TAG	497	544
385648267	3393	TAA	3491	3501
385648267	3393	TAG	3498	3501
61744441	448	TGA	466	488

312283650	1601	TAA	1631	1659
589811545	2995	TGA	3044	3056
387527983	928	TGA	942	981
197245388	940	TAA	1013	1042
197245388	940	TGA	968	982
262399360	1971	TAG	2015	2045
262399359	1962	TAG	2006	2036
283806678	8679	TAG	8790	8839
189242611	609	TAG	944	984
365812503	1512	TAG	1515	1536
365812503	1512	TGA	1522	1536
153792693	6916	TAG	6952	6987
70980548	254	TAG	312	322
38683845	918	TAA	924	937
38683845	918	TAG	959	999
38683845	918	TGA	981	999
148612837	2289	TAG	2586	2599
47419908	2793	TAA	2807	2832
47419908	2793	TAG	2821	2832
47419908	2793	TGA	2795	2832
21361597	1222	TGA	1225	1252
291190786	1506	TAG	1514	1528
67944632	259	TAG	391	414
67944631	259	TAG	391	414
67944633	415	TAG	547	570
122114650	3206	TAA	3430	3450
109715821	865	TAA	991	1017
109715821	865	TAG	1082	1108
109715821	865	TGA	917	923
154354989	1512	TAG	1515	1536
154354989	1512	TGA	1522	1536
47419910	2895	TAA	2909	2934
47419910	2895	TAG	2923	2934
47419910	2895	TGA	2897	2934
216548486	1594	TGA	1632	1658
189491764	1859	TGA	1943	1944
189242610	609	TAG	944	984
315360659	761	TAG	768	787
315360659	761	TGA	763	787
114155141	2355	TAA	2463	2469
114155141	2355	TAG	2450	2469
75677575	956	TAA	959	1006
75677575	956	TAG	971	1006
325197180	604	TAA	616	647
325197180	604	TGA	634	647
189409141	1081	TAA	1134	1169
189409141	1081	TAG	1156	1169
91208427	6031	TAA	6289	6333
115527086	2080	TGA	2087	2130
125656164	2810	TAG	2926	2932
61635914	967	TGA	974	981
183227692	581	TAA	618	652
183227692	581	TGA	625	652
51873042	1122	TAG	1187	1197
189163502	2545	TAG	2788	2813
148746212	2045	TAG	2114	2133
148746212	2045	TGA	2099	2133
122937397	8679	TAG	8790	8839
284795235	1103	TAG	1114	1131
284795235	1103	TGA	1118	1131
392307008	2145	TAA	2393	2428
124107605	853	TAA	861	905
124107605	853	TGA	864	905
385648266	3527	TAA	3625	3635
385648266	3527	TAG	3632	3635
325053711	956	TAA	959	1006
325053711	956	TAG	971	1006
325053712	956	TAA	959	1006
325053712	956	TAG	971	1006
392307006	2628	TAA	2876	2911
459215048	2072	TAG	2147	2151
388490157	3286	TGA	3323	3332
32483358	2382	TAA	2485	2498
32483358	2382	TGA	2449	2498
291290967	5965	TAG	6092	6118
291084578	1907	TAG	2347	2361
291084578	1907	TGA	1937	1972
307746920	769	TAG	802	814
32483365	2349	TAA	2452	2465
32483365	2349	TGA	2416	2465
531990837	1144	TAA	1153	1195
531990837	1144	TGA	1184	1195
32483360	2382	TAA	2485	2498
32483360	2382	TGA	2449	2498

22027649	707	TAG	787	830
22027649	707	TGA	782	830
350606365	491	TAA	580	597
350606365	491	TAG	550	597
459215047	2072	TAG	2147	2151
291575138	897	TAG	1023	1025
531990838	858	TAA	868	910
531990838	858	TGA	899	910
353411933	747	TAG	751	769
313661425	3742	TGA	3757	3801
257467647	3543	TGA	3592	3604
388490159	3786	TGA	3823	3832
531990840	822	TAA	832	874
531990840	822	TGA	863	874
354682004	903	TAG	1487	1508
354682006	1473	TAA	1583	1592
354682006	1473	TAG	1484	1505
354682006	1473	TGA	1572	1592
32483364	2349	TAA	2452	2465
32483364	2349	TGA	2416	2465
305410828	910	TAG	942	954
467091961	1974	TAA	1981	1992
467091961	1974	TAG	2146	2157
467091961	1974	TGA	1990	1992
401664559	8145	TAG	8151	8189
163792200	1425	TAG	1542	1560
119829186	1344	TAA	1374	1392
119829186	1344	TGA	1363	1392
305410830	1079	TAG	1111	1123
284004924	591	TAA	593	600
224994180	967	TGA	974	981
270483791	1363	TGA	1364	1371
305410832	747	TAG	780	792
307746901	1016	TAG	1048	1060
270132520	4540	TAA	4667	4716
153791627	1121	TAG	1132	1149
153791627	1121	TGA	1136	1149
166235162	541	TAA	591	637
193788631	841	TAG	883	905
193788631	841	TGA	857	905
194353965	6584	TAG	6589	6592
217330569	1221	TAG	1261	1294
217330573	1408	TAG	1448	1481
51873044	1026	TAG	1091	1101
119395733	9174	TAG	9307	9345
217330567	1513	TAG	1553	1586
51873046	1122	TAG	1187	1197
299782586	1712	TGA	1735	1757
289547522	1454	TAG	1494	1527
574269958	1379	TAA	1458	1466
574269958	1379	TAG	1385	1396
574272532	1583	TAA	1662	1670
574272532	1583	TAG	1589	1600
574272304	1192	TAA	1271	1279
574272304	1192	TAG	1198	1209
574271714	1478	TAA	1557	1565
574271714	1478	TAG	1484	1495
574271316	1379	TAA	1458	1466
574271316	1379	TAG	1385	1396
574273241	1478	TAA	1557	1565
574273241	1478	TAG	1484	1495
574269522	1427	TAA	1506	1514
574269522	1427	TAG	1433	1444
558472849	1026	TAG	1091	1101
156523967	855	TAA	867	889
156523967	855	TAG	864	889
156523967	855	TGA	873	889
112382209	1093	TAA	1108	1117
112382209	1093	TAG	1174	1218
112382209	1093	TGA	1097	1117
111120330	2052	TAA	2059	2070
111120330	2052	TAG	2224	2235
111120330	2052	TGA	2068	2070
88703042	692	TAA	713	762
109255233	1177	TAG	1255	1287
88703040	587	TAA	608	657
112789561	683	TGA	836	840
126032349	10872	TAG	10941	10960
62244047	946	TAG	963	1009
226437633	186	TAA	452	455
226437633	186	TAG	496	502
226437633	186	TGA	209	231
199559531	1216	TAA	1223	1231
189163519	4030	TAA	4062	4074

189163519	4030	TAG	4116	4164
189163519	4030	TGA	4051	4074
199559437	1154	TAA	1161	1169
199559489	1301	TAA	1308	1316
199558961	1216	TAA	1223	1231
205360986	300	TAA	372	404
205360986	300	TAG	387	404
205360986	300	TGA	304	306
305410834	1016	TAG	1048	1060
32479524	1140	TAG	1659	1670
215982795	767	TAG	823	843
349501056	1573	TAG	1613	1616
305410827	909	TAG	942	954
349501055	1573	TAG	1613	1616
349501059	916	TAG	956	959
349501060	1573	TAG	1613	1616
32479526	1373	TAG	1892	1903
167860144	1425	TAA	1448	1496
167860144	1425	TGA	1473	1496
305410836	693	TAG	726	738
386781570	946	TAG	963	1009
167830435	2605	TGA	2714	2728
167830432	658	TAG	857	895
167830432	658	TGA	660	685
594191052	1137	TAA	1416	1457
594191052	1137	TAG	1236	1279
55956799	426	TAA	429	440
113204614	8646	TAG	8716	8747
113204616	8646	TAG	8716	8747
296939603	2391	TAA	2823	2867
296939601	2391	TAA	2823	2867
27765081	2101	TAA	2199	2221
27765081	2101	TAG	2288	2299
27765081	2101	TGA	2196	2221
157266316	2444	TAA	2455	2464
157266316	2444	TAG	2447	2464
27765075	428	TAA	526	548
27765075	428	TAG	615	626
27765075	428	TGA	523	548
207113159	4478	TGA	4515	4549
304376300	4364	TGA	4401	4435
207113158	4247	TGA	4284	4318
207113161	4367	TGA	4404	4438
207113163	4250	TGA	4287	4321
236463299	559	TAA	595	608
236463299	559	TAG	574	608
236463299	559	TGA	587	608
236463163	559	TAA	595	608
236463163	559	TAG	574	608
236463163	559	TGA	587	608
349501083	1110	TAG	1150	1153
315360665	835	TAG	842	861
315360665	835	TGA	837	861
387527982	915	TAG	1058	1104
387527982	915	TGA	929	968
402513678	237	TAA	246	266
402513678	237	TAG	257	266
305410841	927	TAG	959	971
305410838	905	TAG	937	949
305410844	1074	TAG	1106	1118
574957022	893	TAA	901	935
574957022	893	TAG	1028	1077
574957022	893	TGA	920	935
574957031	1965	TAA	1973	2007
574957031	1965	TAG	2100	2149
574957031	1965	TGA	1992	2007
386268034	971	TAG	1012	1032
386268035	971	TAG	1012	1032
386268037	971	TAG	1012	1032
223555916	1674	TAA	1693	1709
223555916	1674	TGA	1687	1709
396578113	1329	TAA	1399	1403
396578113	1329	TGA	1452	1479
396578116	1117	TAA	1187	1191
396578116	1117	TGA	1240	1267
95147334	3551	TAA	3660	3706

Supplemental Table 2.3. PolyA track genes that are potential NMD targets. Table of mRNAs that have intron-exon boundary closer than 50 nucleotides downstream from a stop codon arising from frameshifting over polyA tracks. These genes would fall in category of "classical" NMD targets if frameshifting occurs on polyA track.

Frameshift	Gene name	Sequence	ELMs
minus one	ALS2CR12	TKKFEMESGEE	ELME000117 ELME000085 ELME000064
minus one	APTX	LEFFQYRIL	ELME000355 ELME000370 ELME000120
minus one	ASTE1	EETEYQLF	ELME000370 ELME000236 ELME000120 ELME000352
minus one	BRCA1	QPNASQAQQKPTTHGR	ELME000202 ELME000353 ELME000239 ELME000070
minus one	C10orf90	RTKEKGDLTK	ELME000351
minus one	CASP5	DVLLYDTIFQIFNNRNCLS	ELME000020 ELME000336 ELME000370 ELME000335 ELME000120 ELME000352
minus one	CCDC146	SWNKKSKR	ELME000011 ELME000285 ELME000278
minus one	CCDC148	LGQEKTEVARNG	ELME000355
minus one	CCDC168	KKVSKLGSQGWRNQ	ELME000202 ELME000351 ELME000008 ELME000053
minus one	CDYL	RQSIWFGGKA	ELME000351 ELME000197
minus one	CENPQ	HLKDLSSEGQTKH	ELME000334 ELME000053
minus one	CEP290	HYQLQVQELTDL	ELME000086 ELME000163
minus one	CHEK1	YLNPWKKIDSAPLALL	ELME000182 ELME000085 ELME000084 ELME000355 ELME000081
minus one	CHRM5	EEKLYWQGNSKLP	ELME000352 ELME000137
minus one	CNTLN	MLKMTRNGCCTFRNFLK DSFLLPHIY	ELME000146 ELME000336 ELME000079

			ELME000337 ELME000355 ELME000370 ELME000369
minus one	CNTRL	AAQTRLSEL	ELME000086 ELME000285
minus one	CPNE3	TRIQVLSV	ELME000333 ELME000091 ELME000365
minus one	DHX36	TFGILKCSISEKSCLRMECKRNW	ELME000336 ELME000162 ELME000368 ELME000063 ELME000103 ELME000360 ELME000370 ELME000108 ELME000064 ELME000100
minus one	DIAPH3	TDFFVLWKASGTIQFNCK	ELME000368 ELME000085 ELME000370 ELME000328
minus one	DYNC1I2	TRRRKLLLLCKKNQILKKKGEKLKHCFKAWG	ELME000146 ELME000149 ELME000106 ELME000108 ELME000231 ELME000012 ELME000233 ELME000102
minus one	EIF2AK2	LLQKLLSKK	ELME000045 ELME000355
minus one	FAM133A	SKDETEKEKD	ELME000285 ELME000220 ELME000064
minus one	FBXO38	DVYPSCSSTTASTVGNSSSHNTASQSPDF	ELME000136 ELME000202 ELME000063 ELME000159 ELME000197 ELME000239 ELME000070 ELME000352 ELME000053

minus one	FILIP1	EILLLAQNEPCPQSQLLHF PERRLQKVEEAHLQTGPH PLFR	ELME000202 ELME000173 ELME000335 ELME000108 ELME000012 ELME000352 ELME000102
minus one	GON4L	TKRKRDGRGQEGTLAYD LKLDDMLDRTLEDGAKQ HN	ELME000147 ELME000108 ELME000220 ELME000365 ELME000100 ELME000102
minus one	GOPC	KEAQLEAEVKLLRKENEA L	ELME000232 ELME000351 ELME000335 ELME000365 ELME000089 ELME000102
minus one	GPATCH4	KEAERGGRSYSI	ELME000091 ELME000351
minus one	HMGXB4	RRTKREREEKSQKRRTCR PTRCS	ELME000271 ELME000202 ELME000101 ELME000351 ELME000062 ELME000108 ELME000220 ELME000276 ELME000100 ELME000008 ELME000102 ELME000053
minus one	IFI16	PKKRLDPKG	ELME000106 ELME000108 ELME000100
minus one	IQCA1	EEEKGKTTQESQKTKERN KGEK	ELME000117 ELME000202 ELME000063 ELME000220 ELME000064 ELME000352 ELME000053
minus one	JARID2	SHSQYHLSPPG	ELME000367 ELME000249 ELME000136

			ELME000159 ELME000285
minus one	KNOP1	HQEGDALPGHSKPSRSMESS	ELME000063 ELME000287 ELME000239 ELME000053
minus one	KRCC1	NKAKKGQRRKCFGTSLFLD	ELME000336 ELME000353 ELME000108 ELME000102
minus one	LPIN2	GERNTNRT	ELME000062
minus one	MAST4	SGKVTKSLSASALSLMIPG DMFAVSPLGSPMSPHSLS SDPSSSRDSSPSRDSSAAS ASPHQPIVIHSSGKNYGFT IRAIRVYVGDSDIYTVHHI VWNVEEGSPACQAGLKA GDLITHINGEPVHGLVHT EVIELLLKSG	ELME000155 ELME000182 ELME000367 ELME000336 ELME000136 ELME000149 ELME000147 ELME000337 ELME000085 ELME000063 ELME000159 ELME000173 ELME000335 ELME000062 ELME000285 ELME000313 ELME000153 ELME000365 ELME000148 ELME000239 ELME000052 ELME000321 ELME000053
minus one	MED19	KRILNGKGRRKRRRKRRI DIVQTTQVWAAPRPAAA AAY	ELME000271 ELME000146 ELME000202 ELME000101 ELME000103 ELME000093 ELME000351 ELME000108 ELME000278 ELME000012 ELME000233 ELME000369

			ELME000052 ELME000276 ELME000100 ELME000270 ELME000102
minus one	MGA	MGSDEFDISPRISKQQEGS SASSVDLGQMF	ELME000136 ELME000085 ELME000063 ELME000159 ELME000062 ELME000365 ELME000197 ELME000053
minus one	MIS18BP1	SIPTYVKKRKTTNHSSQM TVH	ELME000182 ELME000271 ELME000202 ELME000063 ELME000062 ELME000285 ELME000108 ELME000070 ELME000100 ELME000008 ELME000270 ELME000102 ELME000053
minus one		MKNK1	HQQPRACVY
minus one	MORC4	IITEDSLPSLEAILNYSIFNR ENDLLAQFDAIPGKKGTR VLIWNIRR	ELME000336 ELME000149 ELME000147 ELME000063 ELME000355 ELME000106 ELME000093 ELME000120 ELME000341 ELME000012 ELME000287 ELME000069 ELME000365 ELME000233 ELME000070 ELME000052 ELME000008 ELME000137

minus one	MYH10	QAHIQDLEEQLDEEEGAR QKLQLEKVTAEAKIKKM EEEILLLEDQNSKFIKEKK LMEDRIAECSSQLAEEEE KAKNLAKIR	ELME000342 ELME000146 ELME000117 ELME000149 ELME000202 ELME000106 ELME000333 ELME000335 ELME000353 ELME000365 ELME000233 ELME000239
minus one	NEK3	QQNQDSFGK	ELME000353
minus one	NGFRAP1	LIMANIHQENEEMEQPMQ NGEEDRPLGG	ELME000355
minus one	NIPBL	ILTHRRLGVLQE	ELME000355 ELME000106 ELME000108 ELME000012 ELME000102
minus one	NPIPB15	KTKQNPRSKNK	ELME000351
minus one	NR3C1	FSRPLQESHKKP	ELME000117 ELME000336 ELME000355
minus one	NUP85	RWCQVAPLSISS	ELME000351
minus one	OSBPL1A	LSEALETLA	ELME000086 ELME000355
minus one	PA2G4	GLQDCRECHQWGNIRRK	ELME000108 ELME000012 ELME000060 ELME000102
minus one	PHLPP1	RRIHGQYIHCHAKETWNC WAEAWWCRCPLSYQA	ELME000182 ELME000160 ELME000091 ELME000351 ELME000108 ELME000012 ELME000365 ELME000102
minus one	PLXNC1	QILTSYIFGKQTAFLFASG	ELME000182 ELME000198 ELME000353 ELME000197 ELME000052
minus one	PPP1R10	CHLRLPSQAP	ELME000354 ELME000367

			ELME000202 ELME000334
minus one	PRPF40A	QAKQLRKRNWEA	ELME000271 ELME000353 ELME000108 ELME000278 ELME000012 ELME000100 ELME000102
minus one	PXK	FSSKEVKTICS	ELME000146 ELME000355
minus one	RNF145	AAKEKLEAV	ELME000285 ELME000365 ELME000089
minus one	SENP7	YPRVSCYFQVITRKDTQSY	ELME000182 ELME000368 ELME000337 ELME000202 ELME000370 ELME000062 ELME000120 ELME000220 ELME000365 ELME000197 ELME000239 ELME000008 ELME000102 ELME000053
minus one	SGOL1	VPQKKMHKSVSS	ELME000336
minus one	SH3RF1	FVEVAFWRLH	ELME000368 ELME000355 ELME000370
minus one	SLC46A3	SPIFAFQEEVQKKVSR	ELME000117 ELME000011 ELME000285
minus one	SMAD5	CHGGTGESLEQSRTAE	ELME000354 ELME000336 ELME000053
minus one	SPATA16	ATSNCSAK	ELME000085 ELME000285 ELME000070
minus one	TAF1D	RYQPTGRPRGRPEGRRNPIYS	ELME000103 ELME000093 ELME000351 ELME000122 ELME000108

			ELME000097 ELME000012 ELME000095 ELME000102
minus one	TAOK1	SGFQSRRRIYSISKQKKK	ELME000011 ELME000285 ELME000108 ELME000012 ELME000065 ELME000051 ELME000061 ELME000102
minus one	TDRD5	HNRRFARS	ELME000108 ELME000012 ELME000102
minus one	TFDP2	SGLACLPILLRNVRIWR	ELME000285
minus one	TMEM254	KKIEAKNGDPNDCSEFLR SVWVVFWPQSIPYQNLGP LGPFTQYLVDHHHTLLCN GYWLAWLIHVGESLYAIV LCK	ELME000020 ELME000155 ELME000182 ELME000317 ELME000336 ELME000079 ELME000368 ELME000160 ELME000202 ELME000084 ELME000351 ELME000370 ELME000335 ELME000120 ELME000081 ELME000052
minus one	U2SURP	IWNSSKKN	ELME000355 ELME000070
minus one	ULK4	RECWAVPLAAYTV	ELME000091 ELME000351 ELME000369
minus one	VEZF1	KLHLCALTA	ELME000091 ELME000351
minus one	ZC3H13	WKSQERENLGLIS	ELME000149 ELME000355 ELME000333 ELME000231
minus one	ZMYM5	IDAAEHRLYENEKNDGVL LLYT	ELME000149 ELME000084

			ELME000355 ELME000321
minus one	ZRANB2	QRESSWSCIY	ELME000080 ELME000199 ELME000368 ELME000147 ELME000063 ELME000370 ELME000062 ELME000353
Frameshift	**Gene name**	**Sequence**	**ELMs**
plus one	ABCC2	SLGPKKMFQNPG	ELME000146 ELME000285
plus one	ALS2CR12	MTKKFEMESGEE	ELME000117 ELME000085 ELME000064
plus one	ANKHD1-EIF4EBP3	GTEKETGRR	ELME000093
plus one	ANKHD1	GTEKETGRR	ELME000093
plus one	ANKRD49	GKRPKQIASLGC	ELME000091 ELME000108 ELME000100 ELME000270
plus one	APAF1	ITNLSRLVVRPHTDAVYHACF	ELME000155 ELME000355 ELME000371 ELME000070 ELME000052
plus one	APTX	IGILSIQNTS	ELME000355 ELME000333 ELME000053
plus one	APTX	WNSFNTEYF	ELME000063 ELME000355 ELME000089
plus one	BBX	KKSKMDRHG	ELME000351
plus one	BEND2	YQPCCIGICR	ELME000355
plus one	BLM	VSSKSVSEGRDG	ELME000063 ELME000064 ELME000321
plus one	BRCA1	YNQMPVRHSRNLQLME	ELME000355 ELME000106 ELME000062 ELME000120
plus one	C10orf90	RTKEKGDLTK	ELME000351
plus one	C16orf45	KLQKQREDE	ELME000351

plus one	CAPN3	SPSSSFRTEQTATRSWVW TRSQRRAKAKQA	ELME000146 ELME000011 ELME000147 ELME000337 ELME000202 ELME000063 ELME000285 ELME000108 ELME000365 ELME000197 ELME000239 ELME000102 ELME000053
plus one	CASP5	DVLLYDTIFQIFNNRNCLS	ELME000020 ELME000336 ELME000370 ELME000335 ELME000120 ELME000352
plus one	CASP5	RMCCFMTPSSRYSTTATA SV	ELME000358 ELME000136 ELME000063 ELME000159 ELME000351 ELME000062 ELME000365 ELME000239 ELME000053
plus one	CCDC122	VLFNLKNELHELEKEIAAI SAE	ELME000085 ELME000002 ELME000365
plus one	CCDC146	CLGTRSQN	ELME000354
plus one	CCDC148	WAKKKQKWQEME	ELME000271 ELME000201 ELME000355 ELME000278
plus one	CEP290	IIINFKCRSLQIF	ELME000355 ELME000106
plus one	CHD9	RRRYRREAI	ELME000101 ELME000103 ELME000351 ELME000108 ELME000012 ELME000089 ELME000102

plus one	CHRM5	WKRSCTGRGTASY	ELME000355 ELME000173 ELME000062 ELME000108 ELME000100 ELME000053
plus one	CNTRL	QHKLDYQNC	ELME000084 ELME000353
plus one	EIF2AK2	FYRNYSQRN	ELME000202 ELME000084 ELME000355 ELME000070
plus one	EIF5B	TEGQKTEF	ELME000086
plus one	ERC2	ATGPHRREGDTGR	ELME000285 ELME000108 ELME000102
plus one	ERO1LB	ARERLFSLLQG	ELME000045 ELME000368 ELME000149 ELME000370 ELME000285 ELME000231 ELME000061
plus one	EXOC1	NCFLCATVTTERPVQ	ELME000336 ELME000353
plus one	FAM133A	SQRMKQRKKRM	ELME000271 ELME000011 ELME000101 ELME000285 ELME000108 ELME000100 ELME000270 ELME000102
plus one	FAM227B	DSSFVSIYTHLWENVPRIF EALLIMESK	ELME000182 ELME000368 ELME000063 ELME000370 ELME000120 ELME000047 ELME000365 ELME000352
plus one	FAM81B	TEIVFQKYQIYKK	ELME000370
plus one	FILIP1	WRYYSWPRTSHVPSHNY YIFQREDSRKWKRRICRQ AHIPYS	ELME000182 ELME000271 ELME000355 ELME000103

			ELME000370 ELME000062 ELME000120 ELME000108 ELME000278 ELME000012 ELME000048 ELME000239 ELME000052 ELME000100 ELME000163 ELME000102 ELME000053
plus one	FOXP3	MRTPPHPVIISAHTHRKKF GLLEERGLRLPHRTAWFF FSV	ELME000358 ELME000155 ELME000367 ELME000085 ELME000063 ELME000106 ELME000333 ELME000091 ELME000351 ELME000370 ELME000365 ELME000102
plus one	GGNBP2	KKKKSKILKCDEHIQKLG SCITDP	ELME000271 ELME000146 ELME000007 ELME000351 ELME000173 ELME000278 ELME000008
plus one	GYPC	GVETPPAKSAEKK	ELME000358 ELME000136 ELME000085 ELME000159 ELME000239
plus one	HMGXB4	REGQRERERRKAKKEEH VGLPGVL	ELME000155 ELME000271 ELME000146 ELME000101 ELME000103 ELME000351 ELME000108 ELME000002 ELME000278

			ELME000233 ELME000102
plus one	HYDIN	EIESDFLATTNTTKAQEEQTSS	ELME000117 ELME000336 ELME000070 ELME000352 ELME000053
plus one	IGHMBP2	QRTSGHRSAHGGGL	ELME000085 ELME000062 ELME000353 ELME000053
plus one	IL1R2	DHSCDHFPPQDHISFSGVKTDNPV	ELME000368 ELME000085 ELME000370 ELME000352 ELME000053
plus one	IQCA1	KKKKKEKQPKKAKKQKKGTKEK	ELME000271 ELME000146 ELME000351 ELME000278 ELME000276 ELME000008
plus one	JARID2	GPTLSTISALL	ELME000085 ELME000063 ELME000365
plus one	KCNC1	KHIPRPPQLGSPNY	ELME000155 ELME000199 ELME000136 ELME000159 ELME000351 ELME000005
plus one	KDM4D	HDCGGVSPFGKQ	ELME000136 ELME000159
plus one	KNOP1	TRREMPSQATPSPPGPWRAA	ELME000358 ELME000155 ELME000006 ELME000136 ELME000202 ELME000063 ELME000159 ELME000108 ELME000102
plus one	LARP7	DRVEASSLPEVRTGKRKRSSSEDAESLAPRSK	ELME000271 ELME000093 ELME000173 ELME000062

			ELME000108 ELME000365 ELME000064 ELME000100 ELME000061 ELME000008 ELME000352 ELME000270 ELME000102
plus one	LOC101929870	QKSFPSEGQRQRRSLFLDSNSRENLGWLAKLTREQNILPEAEKPHALSGGG	ELME000146 ELME000336 ELME000085 ELME000063 ELME000101 ELME000106 ELME000062 ELME000353 ELME000108 ELME000231 ELME000012 ELME000365 ELME000233 ELME000051 ELME000102
plus one	LPIN2	KEKEIQTGQ	ELME000351
plus one	MPP3	PPMSPACEDTAAPFDEQQQEMAASAAFIDRHYGHLVDAVLVKEDLQGAYSQLKVVLEKLSKDTHWVPVSWVR	ELME000146 ELME000336 ELME000136 ELME000368 ELME000147 ELME000202 ELME000085 ELME000063 ELME000159 ELME000333 ELME000370 ELME000120 ELME000326 ELME000313 ELME000365 ELME000197 ELME000233 ELME000052
plus one	NAA35	SPIEPRDHNEPSISEHVCWNV	ELME000147 ELME000285

			ELME000365 ELME000064
plus one	NCOA7	SQTKCRDSLSCSYKDSYWEGR	ELME000336 ELME000162 ELME000062 ELME000285 ELME000064 ELME000321 ELME000053
plus one	NEK3	PSRIRIALG	ELME000146 ELME000012 ELME000365
plus one	NHLRC2	CTTLAGTGDT	ELME000354
plus one	NIPBL	KKRKAYEPK	ELME000271 ELME000351 ELME000108 ELME000278 ELME000100 ELME000102
plus one	NIPBL	RFLPTGGWGCYRR	ELME000106 ELME000351 ELME000370 ELME000012
plus one	NPIPB15	KKQNKTHAPKTN	ELME000351 ELME000070
plus one	NR3C1	NSAGHYRSLTRNL	ELME000146 ELME000085 ELME000353 ELME000120 ELME000334
plus one	OSBPL1A	CQKHWRRWP	ELME000354 ELME000160 ELME000108 ELME000012 ELME000102
plus one	PDZD9	ERGVSNKVKTSVHNLSKTQQTKLTV	ELME000336 ELME000202 ELME000091 ELME000365 ELME000070 ELME000352
plus one	PEG10	GVEEGARIQASIPTE	ELME000365 ELME000051 ELME000060
plus one	PPFIA2	RLGQLRGFMETEAAAQESLG	ELME000117 ELME000351

			ELME000140 ELME000239
plus one	PPP1R10	TVTYGCQAKPL	ELME000163
plus one	PRR14L	CEENVCRS	ELME000354 ELME000079
plus one	QTRTD1	LGKTGDHTMDIPGCLLYT KTGSAPHLTHHTL	ELME000182 ELME000336 ELME000147 ELME000085 ELME000355 ELME000173 ELME000052 ELME000053
plus one	RALGPS2	SSAPNAVAFTRRFNH	ELME000085 ELME000285 ELME000328 ELME000108 ELME000012 ELME000102
plus one	RBPJ	MERDGCSEQESQPCAFIG	ELME000117 ELME000202 ELME000064 ELME000352 ELME000053
plus one	RNF10	RNRSSCSAPQSSTP	ELME000085 ELME000063 ELME000351 ELME000062 ELME000239 ELME000070 ELME000053
plus one	RNF145	WLQRRNWRQC	ELME000355 ELME000108 ELME000102
plus one	RYR1	RKISQSAQT	ELME000202 ELME000085 ELME000351 ELME000008
plus one	SENP7	HIRGCPVTSKSSPERQLKV MLTNVLWTDLGRKFRKT LPRND	ELME000146 ELME000336 ELME000136 ELME000063 ELME000106 ELME000159 ELME000093 ELME000062

			ELME000153 ELME000278 ELME000365 ELME000064 ELME000239 ELME000052 ELME000102 ELME000053
plus one	SENP7	YPRVSCYFQVITRKGLTTK	ELME000182 ELME000146 ELME000368 ELME000337 ELME000370 ELME000062 ELME000120 ELME000365 ELME000197 ELME000239 ELME000102
plus one	SGOL1	FPKKKCTNLSVP	ELME000271 ELME000367 ELME000365 ELME000233 ELME000070 ELME000008
plus one	SLC26A8	KKGPERAPFLVSFVQ	ELME000336 ELME000149 ELME000351 ELME000328
plus one	SLC46A3	AQFLHSRRKFRKKCHV	ELME000271 ELME000336 ELME000011 ELME000285 ELME000108 ELME000278 ELME000102
plus one	SLC4A7	KEEAERMLQDDDDTVHLPFEGGSLLQIPVKA	ELME000155 ELME000367 ELME000149 ELME000147 ELME000091 ELME000351 ELME000335 ELME000240 ELME000052

plus one	SLCO5A1	SVDAVSDDDVLKEKSNN SEQADKKVSSMGFGKDV RDLPRAAVRI	ELME000063 ELME000106 ELME000091 ELME000198 ELME000285 ELME000365 ELME000197 ELME000233 ELME000070 ELME000008
plus one	SLCO5A1	SVDAVSDDDVLKEKSNN SEQADKKVSSMGFGKDV RGVIIVPSAGVGIVLGGYI	ELME000085 ELME000063 ELME000198 ELME000285 ELME000365 ELME000197 ELME000233 ELME000070 ELME000008
plus one	SPG11	IFLKKRKEL	ELME000011 ELME000355 ELME000108 ELME000100 ELME000102
plus one	SYCP1	QENTNIFIGNT	ELME000353
plus one	TCF25	EKQEKQHGRSIGKRTRRY RSHPRED	ELME000101 ELME000103 ELME000093 ELME000108 ELME000278 ELME000012 ELME000048 ELME000100 ELME000352 ELME000102
plus one	TDRD5	ATTEDLQEA	ELME000285 ELME000052
plus one	TERF1	SRRATESRIPVSKSQP	ELME000146 ELME000062 ELME000285 ELME000108 ELME000239 ELME000008 ELME000102
plus one	TFDP2	QVDWPAYQFCSGMSESG	ELME000085 ELME000063

			ELME000370 ELME000353
plus one	THOC2	KERCTALQDKLLEEEKKQMEHVQRVLQRLKLEKDNWL	ELME000351 ELME000062 ELME000002
plus one	TMEM254	KENRSQEWRPK	ELME000202 ELME000351 ELME000070
plus one	TNRC6B	ATQKVTEQKTKVPE	ELME000011 ELME000285 ELME000053
plus one	TRAPPC10	QHPSPNLYCGQ	ELME000182 ELME000136 ELME000159 ELME000353 ELME000163
plus one	TRDN	CRSRTTQGKKTGKERKTCGTSKVTKERTLRN	ELME000354 ELME000147 ELME000202 ELME000063 ELME000093 ELME000173 ELME000062 ELME000220 ELME000064 ELME000052 ELME000102 ELME000053
plus one	TRDN	RNKDTGERTEES	ELME000351 ELME000053
plus one	TRPC1	IASGIPGFVLIYIDVWPVQL	ELME000020 ELME000182 ELME000085 ELME000355 ELME000333 ELME000091 ELME000120 ELME000083 ELME000047 ELME000365 ELME000081
plus one	ULK4	LESAGLFPWLHTQC	ELME000086 ELME000085 ELME000355
plus one	VEZF1	QNFICVHLLQ	ELME000353

plus one	WNK1	QEESSLKQQVEQSSASQT GIKQLPSASTGIPTASTTS ASVSTQVE	ELME000147 ELME000202 ELME000085 ELME000063 ELME000106 ELME000353 ELME000365 ELME000064 ELME000239 ELME000053
plus one	ZCRB1	LLNQKKKLRK	ELME000271 ELME000011 ELME000355 ELME000278 ELME000102
plus one	ZDHHC3	DEHESRFWPPLLSRLGQP LCHARPREGRPVPVCGLK DPDRHGHSDTSPHHSTTV PSVLMNV	ELME000136 ELME000368 ELME000063 ELME000159 ELME000370 ELME000173 ELME000365 ELME000233 ELME000239 ELME000369 ELME000352 ELME000053
plus one	ZFHX3	STPFSFHS	ELME000285
plus one	ZMYM5	LLMLQNTDYMKMRKMM VCCCCT	ELME000355 ELME000122 ELME000120 ELME000095 ELME000102

Supplemental Table 2.4. Peptides arising from possible frame-shifting on polyA tracks.

Direction of frame-shift, gene name and peptide sequences following polyA track are shown in table. Additional analyses of possible eukaryotic linear motifs(ELMs) found in these peptides is included.

Gene name	mutation (nucleotide)	mutation (protein)	Type of mutation
AASDH	c.342delA	p.K114fs*14	Deletion - Frameshift
ABCA5	c.733G>A	p.E245K	Substitution - Missense
ABCA5	c.742_743insA	p.I248fs*12	Insertion - Frameshift
ABCA5	c.742A>T	p.I248L	Substitution - Missense
ABCA5	c.742delA	p.I248fs*1	Deletion - Frameshift
ABCC2	c.882G>C	p.K294N	Substitution - Missense
ACBD3	c.568delA	p.R190fs*43	Deletion - Frameshift
ADAL	c.644delA	p.E218fs*33	Deletion - Frameshift
ADAL_ENST00000428046	c.563delA	p.E191fs*33	Deletion - Frameshift
AHI1	c.910_911insA	p.T304fs*6	Insertion - Frameshift
AHI1	c.910delA	p.T304fs*23	Deletion - Frameshift
AHI1	c.911C>A	p.T304K	Substitution - Missense
AIM2	c.1027A>C	p.T343P	Substitution - Missense
AIM2	c.1027delA	p.T343fs?	Deletion - Frameshift
AKD1_ENST00000424296	c.2091A>G	p.K697K	Substitution - coding silent
AKD1_ENST00000424296	c.2098_2099delGA	p.E700fs*33	Deletion - Frameshift
AKD1_ENST00000424296	c.2098G>A	p.E700K	Substitution - Missense
AL118506.1	c.48G>A	p.K16K	Substitution - coding silent
ALS2CR12	c.777A>C	p.K259N	Substitution - Missense
ALS2CR12	c.778A>G	p.K260E	Substitution - Missense
ANKHD1	c.4385A>G	p.K1462R	Substitution - Missense
ANKHD1	c.4386G>A	p.K1462K	Substitution - coding silent
ANKHD1-EIF4EBP3	c.4385A>G	p.K1462R	Substitution - Missense
ANKHD1-EIF4EBP3	c.4386G>A	p.K1462K	Substitution - coding silent
ANKRD12	c.2806_2809delAAAC	p.K936fs*22	Deletion - Frameshift
ANKRD1	c.215_216insA	p.K73fs*10	Insertion - Frameshift
ANKRD1	c.216G>T	p.K72N	Substitution - Missense
ANKRD1	c.223C>T	p.L75L	Substitution - coding silent
ANKRD26	c.1340_1341insA	p.N447fs*5	Insertion - Frameshift
ANKRD32_ENST00000265140	c.987A>C	p.E329D	Substitution - Missense
ANKRD36C	c.2517_2518insA	p.Q840fs*11	Insertion - Frameshift
ANKRD36C	c.2713A>G	p.K905E	Substitution - Missense
ANKRD36C	c.2716T>C	p.C906R	Substitution - Missense
ANKRD36C_ENST00000420871	c.2517_2518insA	p.Q840fs*11	Insertion - Frameshift
ANKRD36C_ENST00000420871	c.2713A>G	p.K905E	Substitution - Missense
ANKRD36C_ENST00000420871	c.2716T>C	p.C906R	Substitution - Missense
ANKRD49	c.200delA	p.M70fs*32	Deletion - Frameshift
APAF1	c.1798_1799delAA	p.N602fs*23	Deletion - Frameshift

APAF1	c.1798delA	p.N602fs*8	Deletion - Frameshift
APAF1	c.1799A>G	p.K600R	Substitution - Missense
ARHGAP18	c.1418T>G	p.M473R	Substitution - Missense
ARHGAP18	c.492delA	p.K164fs*54	Deletion - Frameshift
ASH1L	c.2134delA	p.R712fs*23	Deletion - Frameshift
ASTE1	c.1884_1885delAA	p.R632fs*10	Deletion - Frameshift
ASTE1	c.1892A>G	p.K631R	Substitution - Missense
ASTE1	c.1894_1895insA	p.R632fs*11	Insertion - Frameshift
ASTE1	c.1894delA	p.R632fs*33	Deletion - Frameshift
ATAD2	c.354_355insA	p.E119fs*18	Insertion - Frameshift
ATAD2	c.354delA	p.E119fs*8	Deletion - Frameshift
ATL1	c.1665G>T	p.K555N	Substitution - Missense
ATR	c.2320_2321insA	p.I774fs*3	Insertion - Frameshift
ATR	c.2320delA	p.I774fs*5	Deletion - Frameshift
BARD1	c.623_624insA	p.K209fs*5	Insertion - Frameshift
BARD1	c.623delA	p.K208fs*4	Deletion - Frameshift
BARD1_ENST00000260947	c.623_624insA	p.K209fs*5	Insertion - Frameshift
BARD1_ENST00000260947	c.623delA	p.K208fs*4	Deletion - Frameshift
BAT2D1	c.464delA	p.E158fs*66	Deletion - Frameshift
BAT2D1_ENST00000392078	c.464delA	p.E158fs*66	Deletion - Frameshift
BEND5	c.545A>G	p.K182R	Substitution - Missense
BEND5	c.545delA	p.K182fs*15	Deletion - Frameshift
BEND5	c.546G>A	p.K182K	Substitution - coding silent
BEND5_ENST00000371833	c.1052A>G	p.K351R	Substitution - Missense
BEND5_ENST00000371833	c.1052delA	p.K351fs*15	Deletion - Frameshift
BEND5_ENST00000371833	c.1053G>A	p.K351K	Substitution - coding silent
BPTF	c.2874delA	p.I961fs*1	Deletion - Frameshift
BPTF_ENST00000335221	c.3252delA	p.I1087fs*1	Deletion - Frameshift
BRCA1	c.1960_1961insA	p.Y655fs*18	Insertion - Frameshift
BRCA1	c.1961_1961delA	p.K654fs*47	Deletion - Frameshift
BRCA1	c.1961_1962insA	p.Y655fs*18	Insertion - Frameshift
BRCA1	c.1961delA	p.K654fs*47	Deletion - Frameshift
BRCA1_ENST00000471181	c.1961delA	p.K654fs*47	Deletion - Frameshift
BRCA2	c.8941G>A	p.E2981K	Substitution - Missense
C10orf68	c.555G>A	p.K185K	Substitution - coding silent
C10orf6	c.1002A>G	p.E334E	Substitution - coding silent
C10orf90	c.1991A>T	p.K664M	Substitution - Missense
C10orf90	c.1992_1993GA>TT	p.K664_K665>N*	Complex - compound substitution
C10orf90	c.1998delA	p.E667fs*7	Deletion - Frameshift
C10orf96	c.605A>C	p.E202A	Substitution - Missense
C12orf45	c.544delA	p.K184fs*>2	Deletion - Frameshift
C13orf40	c.2236_2237insA	p.I746fs*40	Insertion - Frameshift

C13orf40	c.2236delA	p.I746fs*1	Deletion - Frameshift
C14orf102	c.261G>T	p.K87N	Substitution - Missense
C14orf102	c.268_269insA	p.R90fs*7	Insertion - Frameshift
C14orf102	c.268delA	p.R90fs*69	Deletion - Frameshift
C14orf23	c.342_343insA	p.T117fs*20	Insertion - Frameshift
C14orf23	c.342delA	p.T117fs*8	Deletion - Frameshift
C14orf23	c.345A>C	p.K115N	Substitution - Missense
C14orf23	c.346_347insAAC	p.K116_T117insQ	Insertion - In frame
C14orf38	c.1890A>G	p.K630K	Substitution - coding silent
C14orf38	c.1894_1895insA	p.N632fs*6	Insertion - Frameshift
C14orf38	c.1895_1896insA	p.N632fs*6	Insertion - Frameshift
C16orf45	c.310G>T	p.E104*	Substitution - Nonsense
C16orf45	c.317C>A	p.T106N	Substitution - Missense
C16orf88	c.647A>C	p.K216T	Substitution - Missense
C16orf88	c.652delA	p.I218fs*41	Deletion - Frameshift
C18orf34	c.874delA	p.M292fs*3	Deletion - Frameshift
C18orf34_ENST00000383096	c.874delA	p.M292fs*3	Deletion - Frameshift
C1orf131	c.416G>A	p.R139K	Substitution - Missense
C1orf9	c.850G>T	p.E284*	Substitution - Nonsense
C1orf9_ENST00000367723	c.1327G>T	p.E443*	Substitution - Nonsense
C2orf77	c.407A>G	p.K136R	Substitution - Missense
C2orf77_ENST00000447353	c.407A>G	p.K136R	Substitution - Missense
C3orf77	c.1885delA	p.A632fs*5	Deletion - Frameshift
C3orf77_ENST00000309765	c.1885delA	p.A632fs*5	Deletion - Frameshift
C6orf103_ENST00000367493	c.1813delA	p.K607fs*>6	Deletion - Frameshift
C6orf10	c.1543G>A	p.E515K	Substitution - Missense
CAMKK2_ENST00000392474	c.1601C>A	p.T534K	Substitution - Missense
CAMSAP1L1	c.3748_3749insA	p.Q1253fs*12	Insertion - Frameshift
CAMSAP1L1	c.3749delA	p.K1252fs*19	Deletion - Frameshift
CAPN3_ENST00000397163	c.1788_1789insA	p.T599fs*33	Insertion - Frameshift
CASP5	c.153_154delAA	p.K51fs*3	Deletion - Frameshift
CASP5	c.154delA	p.T52fs*26	Deletion - Frameshift
CASP5_ENST00000393141	c.240_241delAA	p.K80fs*3	Deletion - Frameshift
CASP5_ENST00000393141	c.241delA	p.T81fs*26	Deletion - Frameshift
CCBL1_ENST00000427720	c.375_376delAA	p.K125fs*>34	Deletion - Frameshift
CCDC108_ENST00000295729	c.438delA	p.K146fs*3	Deletion - Frameshift
CCDC148	c.1260delA	p.K420fs*15	Deletion - Frameshift
CCDC150	c.838_839insA	p.E284fs*13	Insertion - Frameshift
CCDC150	c.839delA	p.E284fs*14	Deletion - Frameshift
CCDC150	c.847A>G	p.K283E	Substitution - Missense
CCDC150	c.850G>A	p.E284K	Substitution - Missense

CCDC175	c.1890A>G	p.K630K	Substitution - coding silent
CCDC34_ENST00000328697	c.720A>G	p.L240L	Substitution - coding silent
CCDC34_ENST00000328697	c.731_732insA	p.N244fs*3	Insertion - Frameshift
CCDC34_ENST00000328697	c.731delA	p.N244fs*28	Deletion - Frameshift
CCT8L1	c.1642_1643insA	p.I552fs*6	Insertion - Frameshift
CCT8L2	c.1654_1655insA	p.I552fs*6	Insertion - Frameshift
CCT8L2	c.1654delA	p.I552fs*4	Deletion - Frameshift
CD46	c.509A>G	p.N170S	Substitution - Missense
CDHR3_ENST00000542731	c.2259-2dela	p.?	Unknown
CDKL2	c.222_223insA	p.R75fs*7	Insertion - Frameshift
CDKL2	c.222delA	p.K74fs*5	Deletion - Frameshift
CDYL	c.216_217insA	p.G75fs*13	Insertion - Frameshift
CDYL	c.217delA	p.K76fs*25	Deletion - Frameshift
CDYL	c.219A>G	p.K73K	Substitution - coding silent
CEP164	c.336_337insA	p.E117fs*88	Insertion - Frameshift
CEP164	c.337delA	p.K116fs*22	Deletion - Frameshift
CEP164	c.347A>G	p.K116R	Substitution - Missense
CEP290	c.828delA	p.E277fs*16	Deletion - Frameshift
CHD2	c.3724_3725insA	p.Y1246fs*13	Insertion - Frameshift
CHD2	c.3725delA	p.K1245fs*4	Deletion - Frameshift
CHD2_ENST00000394196	c.3725delA	p.K1245fs*4	Deletion - Frameshift
CHD7	c.1922A>T	p.K641I	Substitution - Missense
CHD7_ENST00000423902	c.1922A>T	p.K641I	Substitution - Missense
CHEK1	c.668A>C	p.E223A	Substitution - Missense
CHEK1	c.668delA	p.T226fs*14	Deletion - Frameshift
CHEK1	c.675A>G	p.K225K	Substitution - coding silent
CIR1	c.865delA	p.I289fs*51	Deletion - Frameshift
CNTRL_ENST00000373855	c.1328delA	p.I446fs*1	Deletion - Frameshift
COL17A1	c.1170delA	p.E391fs*12	Deletion - Frameshift
COL17A1	c.1171G>C	p.E391Q	Substitution - Missense
CPNE3	c.771G>A	p.K257K	Substitution - coding silent
CPNE3	c.771G>T	p.K257N	Substitution - Missense
CWC27	c.995delA	p.V335fs*1	Deletion - Frameshift
CWF19L2_ENST00000282251	c.279G>T	p.K93N	Substitution - Missense
CWF19L2_ENST00000282251	c.287delA	p.K96fs*41	Deletion - Frameshift
DCLRE1C	c.1706A>C	p.K569T	Substitution - Missense
DCLRE1C	c.1708delA	p.R570fs*6	Deletion - Frameshift
DCLRE1C_ENST00000378278	c.2051A>C	p.K684T	Substitution - Missense
DCLRE1C_ENST00000378278	c.2053delA	p.R685fs*6	Deletion - Frameshift
DDX18	c.327G>T	p.K109N	Substitution - Missense

DDX18	c.333G>T	p.K111N	Substitution - Missense
DDX59	c.1294A>G	p.K432E	Substitution - Missense
DDX59_ENST00000331314	c.1294A>G	p.K432E	Substitution - Missense
DENR	c.317delA	p.K108fs*10	Deletion - Frameshift
DHX36	c.2560delA	p.R854fs*4	Deletion - Frameshift
DHX36	c.2564A>T	p.K855I	Substitution - Missense
DHX36	c.460_461insA	p.M154fs*3	Insertion - Frameshift
DHX36	c.460delA	p.M154fs*27	Deletion - Frameshift
DHX36	c.461_462insA	p.M154fs*3	Insertion - Frameshift
DHX36	c.578delA	p.N193fs*25	Deletion - Frameshift
DIAPH2	c.208G>T	p.E70*	Substitution - Nonsense
DIAPH3	c.952A>C	p.K318Q	Substitution - Missense
DIAPH3	c.958delA	p.I320fs*20	Deletion - Frameshift
DNAH6	c.6440G>T	p.R2147I	Substitution - Missense
DNAJC1	c.578_579insA	p.T194fs*18	Insertion - Frameshift
DNAJC2	c.590_591insA	p.N197fs*4	Insertion - Frameshift
DNAJC2	c.590delA	p.N197fs*8	Deletion - Frameshift
DSEL	c.2905_2906insA	p.R969fs*4	Insertion - Frameshift
DSEL	c.2905delA	p.R969fs*16	Deletion - Frameshift
DSEL	c.2910A>C	p.K970N	Substitution - Missense
DSEL	c.2910A>T	p.K970N	Substitution - Missense
DYNC1I2	c.97delA	p.E36fs*34	Deletion - Frameshift
DYNC2H1_ENST00000398093	c.832A>C	p.N278H	Substitution - Missense
EEA1	c.2428A>C	p.K810Q	Substitution - Missense
EFCAB7	c.1138_1139insA	p.I380fs*7	Insertion - Frameshift
EHBP1	c.1026delA	p.N344fs*2	Deletion - Frameshift
EIF2AK2	c.1531G>T	p.E511*	Substitution - Nonsense
EIF3J	c.222G>C	p.K74N	Substitution - Missense
EIF3J	c.223delA	p.I77fs*1	Deletion - Frameshift
EIF3J	c.229delA	p.I77fs*1	Deletion - Frameshift
EML6	c.4063delA	p.K1357fs*9	Deletion - Frameshift
EML6	c.4071G>T	p.K1357N	Substitution - Missense
ENSG00000121031	c.10811delA	p.N3604fs*48	Deletion - Frameshift
ENSG00000121031	c.496_497insA	p.I166fs*11	Insertion - Frameshift
ENSG00000121031	c.496delA	p.I166fs*6	Deletion - Frameshift
ENSG00000174501	c.4764_4765insA	p.Q1589fs*11	Insertion - Frameshift
ENSG00000174501	c.4960A>G	p.K1654E	Substitution - Missense
ENSG00000174501	c.4963T>C	p.C1655R	Substitution - Missense
ENSG00000188423	c.651A>G	p.K217K	Substitution - coding silent
ENSG00000188423	c.658_659delGA	p.E220fs*33	Deletion - Frameshift
ENSG00000188423	c.658G>A	p.E220K	Substitution - Missense
ENSG00000225516	c.313A>C	p.K105Q	Substitution - Missense

ENSG00000268852	c.52A>T	p.K18*	Substitution - Nonsense
ERC2	c.1528delA	p.T510fs*21	Deletion - Frameshift
ERC2_ENST00000288221	c.1528delA	p.T510fs*21	Deletion - Frameshift
ERCC4	c.1461G>C	p.K487N	Substitution - Missense
ERICH1	c.487delA	p.R163fs*3	Deletion - Frameshift
ERO1LB	c.188A>C	p.K63T	Substitution - Missense
ESCO2	c.1117G>C	p.D373H	Substitution - Missense
F5	c.3096G>C	p.K1032N	Substitution - Missense
F5	c.3102A>C	p.K1034N	Substitution - Missense
F8	c.3632A>C	p.K1211T	Substitution - Missense
F8	c.3637delA	p.I1213fs*5	Deletion - Frameshift
F8	c.3638T>G	p.I1213S	Substitution - Missense
F8_ENST00000360256	c.3632A>C	p.K1211T	Substitution - Missense
F8_ENST00000360256	c.3637delA	p.I1213fs*5	Deletion - Frameshift
F8_ENST00000360256	c.3638T>G	p.I1213S	Substitution - Missense
FAM133A	c.150T>G	p.N50K	Substitution - Missense
FAM178A	c.1002A>G	p.E334E	Substitution - coding silent
FAM186A_ENST00000327337	c.2261delA	p.K754fs*2	Deletion - Frameshift
FAM200B	c.170_173delAAGT	p.S59fs*9	Deletion - Frameshift
FAM200B_ENST00000422728	c.170_173delAAGT	p.S59fs*9	Deletion - Frameshift
FAM83A_ENST00000536633	c.1065delA	p.K357fs*8	Deletion - Frameshift
FAM9A	c.477_478insA	p.Q160fs*8	Insertion - Frameshift
FAM9A	c.477delA	p.K159fs*3	Deletion - Frameshift
FAM9A	c.480A>G	p.Q160Q	Substitution - coding silent
FASTKD1	c.2213G>T	p.R738I	Substitution - Missense
FASTKD1	c.2216A>G	p.K739R	Substitution - Missense
FASTKD1	c.2220delA	p.K740fs*10	Deletion - Frameshift
FAT4	c.12401delA	p.K4136fs*17	Deletion - Frameshift
FBXO38	c.2083delA	p.N697fs*32	Deletion - Frameshift
FBXO38	c.2085A>G	p.K695K	Substitution - coding silent
FERMT2	c.452A>C	p.K151T	Substitution - Missense
FERMT2	c.455delA	p.K152fs*4	Deletion - Frameshift
FERMT2	c.456G>A	p.K152K	Substitution - coding silent
FERMT2	c.456G>AG	p.K153fs*5	Complex - frameshift
FEZ2	c.837G>T	p.K279N	Substitution - Missense
FEZ2	c.838A>G	p.K280E	Substitution - Missense
FLG	c.476_477insA	p.E160fs*10	Insertion - Frameshift
FLG	c.477_478insA	p.E160fs*10	Insertion - Frameshift
FLJ45831	c.312G>T	p.K104N	Substitution - Missense
FRA10AC1	c.694_695insA	p.R232fs*4	Insertion - Frameshift
FRA10AC1	c.694delA	p.R232fs*>84	Deletion - Frameshift

FRA10AC1_ENST00000371426	c.694_695insA	p.R232fs*4	Insertion - Frameshift
FRA10AC1_ENST00000371426	c.694delA	p.R232fs*67	Deletion - Frameshift
GIMAP7	c.776delA	p.I261fs*1	Deletion - Frameshift
GOLGA4	c.4091delA	p.V1367fs*10	Deletion - Frameshift
GPR110	c.95_96insA	p.E33fs*6	Insertion - Frameshift
GPR110	c.95_96insT	p.K32fs*7	Insertion - Frameshift
GPR110_ENST00000371243	c.95_96insA	p.E33fs*6	Insertion - Frameshift
GRK4_ENST00000398052	c.656delA	p.R222fs*2	Deletion - Frameshift
GRK4_ENST00000398052	c.666_669delAATA	p.I223fs*10	Deletion - Frameshift
GRK4_ENST00000398052	c.668T>A	p.I223K	Substitution - Missense
GRLF1	c.570A>T	p.K190N	Substitution - Missense
GRLF1_ENST00000317082	c.570A>T	p.K190N	Substitution - Missense
GRLF1_ENST00000317082	c.578A>C	p.K193T	Substitution - Missense
HELLS	c.454delA	p.N154fs*29	Deletion - Frameshift
HERC5	c.409_410insA	p.I140fs*19	Insertion - Frameshift
HERC5	c.410delA	p.I140fs*1	Deletion - Frameshift
HERC5	c.416A>T	p.K139I	Substitution - Missense
HMGXB4	c.1163delA	p.K391fs*33	Deletion - Frameshift
HMGXB4	c.1173G>A	p.K391K	Substitution - coding silent
HMMR	c.1990_1991delAA	p.K666fs*3	Deletion - Frameshift
HMMR	c.1990delA	p.K666fs*11	Deletion - Frameshift
IQGAP2	c.4364G>T	p.R1455I	Substitution - Missense
ITIH5_ENST00000397146	c.2020C>A	p.Q674K	Substitution - Missense
ITPR2	c.3123-1G>A	p.?	Unknown
ITPR2	c.3127G>T	p.E1043*	Substitution - Nonsense
JMJD1C	c.5357A>G	p.E1786G	Substitution - Missense
JMJD1C	c.5364delA	p.E1789fs*45	Deletion - Frameshift
JMJD1C_ENST00000399262	c.6068A>G	p.E2023G	Substitution - Missense
JMJD1C_ENST00000399262	c.6075delA	p.E2026fs*45	Deletion - Frameshift
KCNC1	c.1362_1363insA	p.K458fs*16	Insertion - Frameshift
KCNC1	c.1363delA	p.K457fs*20	Deletion - Frameshift
KCNC1_ENST00000265969	c.1362_1363insA	p.K458fs*16	Insertion - Frameshift
KCNC1_ENST00000265969	c.1363delA	p.K457fs*20	Deletion - Frameshift
KCNQ1	c.1257G>T	p.K419N	Substitution - Missense
KCNQ1	c.1258delA	p.K422fs*10	Deletion - Frameshift
KDM2B_ENST00000377071	c.75A>C	p.K25N	Substitution - Missense
KDM2B_ENST00000377071	c.77delA	p.K26fs*81	Deletion - Frameshift
KDM2B_ENST00000377071	c.82_83delAC	p.T28fs*8	Deletion - Frameshift
KDM4D	c.271delA	p.K93fs*4	Deletion - Frameshift
KIAA1279	c.1509delA	p.I506fs*1	Deletion - Frameshift
KIAA1731	c.1649G>A	p.R550K	Substitution - Missense

KIAA1731_ENST00000325212	c.1649G>A	p.R550K	Substitution - Missense
KIAA2018	c.3045A>C	p.K1015N	Substitution - Missense
KIAA2018	c.3046_3047delAA	p.N1016fs*8	Deletion - Frameshift
KIAA2018	c.3046_3047insA	p.N1016fs*9	Insertion - Frameshift
KIAA2018	c.3046A>C	p.N1016H	Substitution - Missense
KIAA2018	c.3047delA	p.N1016fs*23	Deletion - Frameshift
KIAA2026_ENST00000399933	c.2069delA	p.K690fs*3	Deletion - Frameshift
KIAA2026_ENST00000399933	c.2074T>C	p.L692L	Substitution - coding silent
KIF5B	c.1045G>T	p.E349*	Substitution - Nonsense
KIF6	c.1458G>C	p.K486N	Substitution - Missense
KIF6_ENST00000287152	c.1458G>C	p.K486N	Substitution - Missense
KRCC1	c.715A>G	p.K239E	Substitution - Missense
LAMB4	c.4803A>T	p.K1601N	Substitution - Missense
LMOD2_ENST00000458573	c.1432A>C	p.K478Q	Substitution - Missense
LMOD2_ENST00000458573	c.1437G>T	p.K479N	Substitution - Missense
LRRC17	c.597A>C	p.K199N	Substitution - Missense
LRRIQ1	c.818A>G	p.E273G	Substitution - Missense
LRRIQ1_ENST00000393217	c.4615_4616insA	p.I1542fs*8	Insertion - Frameshift
LRRIQ1_ENST00000393217	c.4616delA	p.I1542fs*17	Deletion - Frameshift
LRRIQ1_ENST00000393217	c.4623A>G	p.K1541K	Substitution - coding silent
LRRIQ1_ENST00000393217	c.4624A>G	p.I1542V	Substitution - Missense
LRRIQ1_ENST00000393217	c.4625T>A	p.I1542N	Substitution - Missense
LRRIQ1_ENST00000393217	c.5095_5096insA	p.N1702fs*13	Insertion - Frameshift
LRRIQ1_ENST00000393217	c.5096delA	p.N1702fs*20	Deletion - Frameshift
LTN1	c.1597_1598delAA	p.N536fs*2	Deletion - Frameshift
LTN1	c.1607delA	p.N536fs*33	Deletion - Frameshift
LTN1	c.1608T>G	p.N536K	Substitution - Missense
MAP7D3	c.2567A>C	p.K856T	Substitution - Missense
MAP9	c.1733A>C	p.K578T	Substitution - Missense
MARCKS	c.454delA	p.K155fs*12	Deletion - Frameshift
MCF2	c.773T>A	p.I258K	Substitution - Missense
MCF2	c.774A>T	p.I258I	Substitution - coding silent
MCF2	c.780_781insA	p.L261fs*6	Insertion - Frameshift
MCF2_ENST00000370573	c.773T>A	p.I258K	Substitution - Missense
MCF2_ENST00000370573	c.774A>T	p.I258I	Substitution - coding silent
MCF2_ENST00000370573	c.780_781insA	p.L261fs*6	Insertion - Frameshift
MCF2_ENST00000370578	c.1208T>A	p.I403K	Substitution - Missense
MCF2_ENST00000519895	c.953T>A	p.I318K	Substitution - Missense
MCF2_ENST00000519895	c.954A>T	p.I318I	Substitution - coding silent
MCF2_ENST00000519895	c.960_961insA	p.L321fs*6	Insertion - Frameshift
MIS18BP1	c.471delA	p.K157fs*24	Deletion - Frameshift

MLH3	c.1755_1756insA	p.E586fs*3	Insertion - Frameshift
MLH3	c.1755delA	p.E586fs*24	Deletion - Frameshift
MLH3	c.1756G>T	p.E586*	Substitution - Nonsense
MLH3_ENST00000355774	c.1755_1756insA	p.E586fs*3	Insertion - Frameshift
MLH3_ENST00000355774	c.1755delA	p.E586fs*24	Deletion - Frameshift
MLH3_ENST00000355774	c.1756G>T	p.E586*	Substitution - Nonsense
MORC1	c.2634A>C	p.E878D	Substitution - Missense
MORC1	c.2641_2642insA	p.I881fs*11	Insertion - Frameshift
MORC1	c.2641delA	p.I881fs*1	Deletion - Frameshift
MPP3	c.1538C>T	p.T513M	Substitution - Missense
MPP6	c.910delA	p.K306fs*4	Deletion - Frameshift
MTDH	c.1341G>A	p.K447K	Substitution - coding silent
MTDH	c.1342A>T	p.K448*	Substitution - Nonsense
MTIF2	c.1975T>A	p.F659I	Substitution - Missense
MYCBP2	c.1124delA	p.K375fs*4	Deletion - Frameshift
MYCBP2_ENST00000357337	c.1124delA	p.K375fs*4	Deletion - Frameshift
MYCBP2_ENST00000407578	c.1238delA	p.K413fs*4	Deletion - Frameshift
MYT1L	c.182G>T	p.R61I	Substitution - Missense
MYT1L	c.187A>G	p.T63A	Substitution - Missense
NAA16	c.1883G>C	p.R628T	Substitution - Missense
NAA35	c.1692G>A	p.K564K	Substitution - coding silent
NAA35	c.1693delA	p.K567fs*6	Deletion - Frameshift
NEK1	c.3156A>G	p.K1052K	Substitution - coding silent
NEK1_ENST00000507142	c.3156A>G	p.K1052K	Substitution - coding silent
NHLRC2	c.1517_1518insA	p.N508fs*13	Insertion - Frameshift
NHLRC2	c.1522A>C	p.N508H	Substitution - Missense
NIPBL	c.1507delA	p.R505fs*35	Deletion - Frameshift
NIPBL_ENST00000448238	c.1507delA	p.R505fs*35	Deletion - Frameshift
NKRF	c.700A>C	p.K234Q	Substitution - Missense
NKRF_ENST00000542113	c.745A>C	p.K249Q	Substitution - Missense
NKTR	c.1297_1298insA	p.V436fs*2	Insertion - Frameshift
NOL7	c.707delA	p.K238fs*16	Deletion - Frameshift
NOL7	c.716delA	p.N240fs*14	Deletion - Frameshift
NOP58	c.1564delA	p.K524fs*>6	Deletion - Frameshift
NUFIP1	c.494delA	p.K165fs*38	Deletion - Frameshift
NUP85	c.127+2T>C	p.?	Unknown
OR6C76	c.921_922insA	p.H312fs*>2	Insertion - Frameshift
OR6C76	c.921C>A	p.H307Q	Substitution - Missense
OR6C76	c.922_923delAA	p.K311fs*>2	Deletion - Frameshift
OR6C76	c.922delA	p.K311fs*>2	Deletion - Frameshift
OR6C76	c.932A>C	p.K311T	Substitution - Missense

PA2G4	c.1108delA	p.K372fs*16	Deletion - Frameshift
PA2G4	c.1116G>T	p.K372N	Substitution - Missense
PARP14	c.4200G>T	p.K1400N	Substitution - Missense
PARP14_ENST00000474629	c.4689G>T	p.K1563N	Substitution - Missense
PCDH7	c.2685A>G	p.K895K	Substitution - coding silent
PCDH7_ENST00000361762	c.2826A>G	p.K942K	Substitution - coding silent
PCDHA12	c.552delA	p.D187fs*8	Deletion - Frameshift
PCDHA12	c.559G>T	p.D187Y	Substitution - Missense
PDCL2	c.253A>G	p.K85E	Substitution - Missense
PDCL2	c.262A>C	p.K88Q	Substitution - Missense
PKD2L2	c.904delA	p.I304fs*1	Deletion - Frameshift
PKD2L2	c.916G>T	p.E306*	Substitution - Nonsense
PKD2L2_ENST00000508883	c.916G>T	p.E306*	Substitution - Nonsense
PLXNC1	c.4192delA	p.I1400fs*21	Deletion - Frameshift
PLXNC1	c.4196A>C	p.K1399T	Substitution - Missense
PNISR	c.806A>G	p.K269R	Substitution - Missense
PNISR	c.807_808insA	p.A270fs*6	Insertion - Frameshift
PPFIA2	c.2527A>G	p.K843E	Substitution - Missense
PPFIA2	c.2529A>G	p.K843K	Substitution - coding silent
PPP1R10	c.924A>G	p.K308K	Substitution - coding silent
PPP1R10	c.930delA	p.V311fs*79	Deletion - Frameshift
PPP2R3C	c.67_68insA	p.S23fs*2	Insertion - Frameshift
PPP2R3C	c.67delA	p.S23fs*5	Deletion - Frameshift
PRKDC	c.10814delA	p.N3605fs*48	Deletion - Frameshift
PRKDC	c.496_497insA	p.I166fs*11	Insertion - Frameshift
PRKDC	c.496delA	p.I166fs*6	Deletion - Frameshift
PRPF40A	c.1167+3A>T	p.?	Unknown
PRPF40A_ENST00000359961	c.1560+3A>T	p.?	Unknown
PRPF40A_ENST00000410080	c.1479+3A>T	p.?	Unknown
PRR11	c.58delA	p.E23fs*9	Deletion - Frameshift
PTHLH_ENST00000354417	c.557delA	p.K186fs*12	Deletion - Frameshift
PTPLAD1	c.1074G>T	p.K358N	Substitution - Missense
PTPLAD1	c.1077A>C	p.K359N	Substitution - Missense
PTPRC	c.2404G>T	p.E802*	Substitution - Nonsense
PTPRZ1	c.5636delA	p.K1879fs*34	Deletion - Frameshift
PTPRZ1_ENST00000393386	c.5636delA	p.K1879fs*34	Deletion - Frameshift
PXK	c.1367_1368insA	p.R459fs*15	Insertion - Frameshift
PXK_ENST00000356151	c.1367_1368insA	p.R459fs*15	Insertion - Frameshift
PYHIN1	c.416_417insA	p.P142fs*3	Insertion - Frameshift
PYHIN1	c.423_424insA	p.P142fs*3	Insertion - Frameshift
PYHIN1	c.424C>A	p.P142T	Substitution - Missense

PYHIN1_ENST00000368135	c.424C>A	p.P142T	Substitution - Missense
PYHIN1_ENST00000392254	c.424C>A	p.P142T	Substitution - Missense
Q99543-2	c.590_591insA	p.N197fs*4	Insertion - Frameshift
Q99543-2	c.590delA	p.N197fs*8	Deletion - Frameshift
RAD23B_ENST00000457811	c.312_313delAA	p.K108fs*38	Deletion - Frameshift
RALGAPA1	c.5582delA	p.N1861fs*6	Deletion - Frameshift
RALGAPA1_ENST00000307138	c.5582delA	p.N1861fs*6	Deletion - Frameshift
RASAL2	c.1097A>G	p.E366G	Substitution - Missense
RASAL2	c.1104_1105insA	p.D372fs*4	Insertion - Frameshift
RASAL2	c.1104G>C	p.K368N	Substitution - Missense
RASAL2	c.1105_1106insA	p.D372fs*4	Insertion - Frameshift
RASAL2	c.1105delA	p.K371fs*7	Deletion - Frameshift
RASAL2	c.1111_1112insA	p.D372fs*4	Insertion - Frameshift
RASAL2	c.1112_1113insA	p.D372fs*4	Insertion - Frameshift
RASAL2_ENST00000367649	c.1151A>G	p.E384G	Substitution - Missense
RASAL2_ENST00000367649	c.1158_1159insA	p.D390fs*4	Insertion - Frameshift
RASAL2_ENST00000367649	c.1158G>C	p.K386N	Substitution - Missense
RASAL2_ENST00000367649	c.1159_1160insA	p.D390fs*4	Insertion - Frameshift
RASAL2_ENST00000367649	c.1159delA	p.K389fs*7	Deletion - Frameshift
RASAL2_ENST00000367649	c.1165_1166insA	p.D390fs*4	Insertion - Frameshift
RASAL2_ENST00000367649	c.1166_1167insA	p.D390fs*4	Insertion - Frameshift
RASAL2_ENST00000462775	c.707A>G	p.E236G	Substitution - Missense
RASAL2_ENST00000462775	c.714_715insA	p.D242fs*4	Insertion - Frameshift
RASAL2_ENST00000462775	c.714G>C	p.K238N	Substitution - Missense
RASAL2_ENST00000462775	c.715_716insA	p.D242fs*4	Insertion - Frameshift
RASAL2_ENST00000462775	c.715delA	p.K241fs*7	Deletion - Frameshift
RBM43	c.205G>T	p.E69*	Substitution - Nonsense
RBM43	c.213_214insA	p.V72fs*18	Insertion - Frameshift
RBM43	c.213delA	p.V72fs*13	Deletion - Frameshift
RBMX2	c.491delA	p.K166fs*29	Deletion - Frameshift
RBMX2	c.498_499insA	p.K170fs*30	Insertion - Frameshift
RBMX2	c.503A>C	p.K168T	Substitution - Missense
RBMX2	c.505A>G	p.K169E	Substitution - Missense
RBMX2	c.506A>G	p.K169R	Substitution - Missense
RBMX2	c.511_514delGAAA	p.E171fs*23	Deletion - Frameshift
RBPJ	c.202delA	p.E71fs*21	Deletion - Frameshift
RBPJ	c.204A>G	p.K68K	Substitution - coding silent
RBPJ_ENST00000348160	c.205delA	p.E72fs*21	Deletion - Frameshift
RBPJ_ENST00000348160	c.207A>G	p.K69K	Substitution - coding silent
RDX	c.628A>C	p.N210H	Substitution - Missense
REV3L	c.4543G>A	p.E1515K	Substitution - Missense
REV3L	c.4543G>T	p.E1515*	Substitution - Nonsense

REV3L	c.4550T>C	p.I1517T	Substitution - Missense
REV3L_ENST00000358835	c.4777G>T	p.E1593*	Substitution - Nonsense
REV3L_ENST00000358835	c.4784T>C	p.I1595T	Substitution - Missense
RG9MTD1	c.384delA	p.K131fs*3	Deletion - Frameshift
RIF1	c.4507delA	p.K1505fs*18	Deletion - Frameshift
RNASEH2B	c.917A>C	p.K306T	Substitution - Missense
RNASEH2B	c.926T>A	p.I309N	Substitution - Missense
RNASEH2B	c.926T>C	p.I309T	Substitution - Missense
RNF145	c.68A>G	p.K23R	Substitution - Missense
RNF145	c.68delA	p.K23fs*17	Deletion - Frameshift
RNF145	c.69G>A	p.K23K	Substitution - coding silent
RNF145	c.70A>G	p.K24E	Substitution - Missense
RNF145	c.71A>G	p.K24R	Substitution - Missense
RNF145	c.79_80delAA	p.N27fs*43	Deletion - Frameshift
RNF145	c.80delA	p.N27fs*13	Deletion - Frameshift
RNF145	c.81C>A	p.N27K	Substitution - Missense
RNPC3	c.346_347insA	p.R120fs*3	Insertion - Frameshift
RNPC3	c.347delA	p.R120fs*18	Deletion - Frameshift
ROCK1_ENST00000399799	c.149A>C	p.K50T	Substitution - Missense
RPL9	c.150G>A	p.K50K	Substitution - coding silent
RPL9	c.158A>G	p.K53R	Substitution - Missense
RPL9	c.159G>T	p.K53N	Substitution - Missense
RSPO3	c.659_660insA	p.P223fs*2	Insertion - Frameshift
RYR1	c.8508G>T	p.K2836N	Substitution - Missense
RYR1	c.8509delA	p.T2839fs*89	Deletion - Frameshift
SAT1_ENST00000379251	c.439delA	p.N150fs*13	Deletion - Frameshift
SCAF11	c.2995_2999delGAAAA	p.E999fs*2	Deletion - Frameshift
SCAF11	c.3002_3003insA	p.N1001fs*2	Insertion - Frameshift
SCAF11	c.3002delA	p.N1001fs*5	Deletion - Frameshift
SCAPER	c.2603delA	p.N868fs*8	Deletion - Frameshift
SCAPER	c.2605A>T	p.K869*	Substitution - Nonsense
SCAPER	c.2613delA	p.A872fs*4	Deletion - Frameshift
SCAPER_ENST00000538941	c.1867A>T	p.K623*	Substitution - Nonsense
SCAPER_ENST00000538941	c.1875delA	p.A626fs*4	Deletion - Frameshift
SEC63	c.1586delA	p.K529fs*4	Deletion - Frameshift
SEC63	c.1587G>T	p.K529N	Substitution - Missense
SENP7	c.230A>G	p.K77R	Substitution - Missense
SENP7_ENST00000394095	c.230A>G	p.K77R	Substitution - Missense
SEPT7_ENST00000469679	c.683C>A	p.A228E	Substitution - Missense
SH3RF1	c.2151G>A	p.K717K	Substitution - coding silent
SHPRH	c.495_496insA	p.E166fs*7	Insertion - Frameshift
SHPRH	c.495delA	p.E166fs*3	Deletion - Frameshift

SLC16A12_ENST00000371790	c.83G>C	p.R28T	Substitution - Missense
SLC22A9	c.1005A>C	p.K335N	Substitution - Missense
SLC22A9	c.995delA	p.K335fs*67	Deletion - Frameshift
SLC45A2	c.865A>C	p.K289Q	Substitution - Missense
SLC46A3	c.154A>C	p.K52Q	Substitution - Missense
SLC46A3_ENST00000380814	c.154A>C	p.K52Q	Substitution - Missense
SLC4A7	c.3450G>C	p.K1150N	Substitution - Missense
SLCO5A1	c.1148A>G	p.K383R	Substitution - Missense
SLCO5A1	c.1150T>G	p.F384V	Substitution - Missense
SLTM	c.1539delA	p.E514fs*8	Deletion - Frameshift
SLTM	c.1540G>T	p.E514*	Substitution - Nonsense
SMAD5	c.137A>G	p.K46R	Substitution - Missense
SMC1B_ENST00000357450	c.2444C>G	p.T815S	Substitution - Missense
SMC2	c.814delA	p.I274fs*1	Deletion - Frameshift
SMC2	c.822A>G	p.I274M	Substitution - Missense
SMC2_ENST00000303219	c.822A>G	p.I274M	Substitution - Missense
SMC2L1	c.814delA	p.I274fs*1	Deletion - Frameshift
SMC2L1	c.822A>G	p.I274M	Substitution - Missense
SMNDC1	c.567C>A	p.N189K	Substitution - Missense
SNX6	c.538A>T	p.K180*	Substitution - Nonsense
SP100_ENST00000341950	c.1349_1350insA	p.K455fs*>20	Insertion - Frameshift
SPATA1	c.1345delA	p.I452fs*1	Deletion - Frameshift
SPEF2	c.711G>A	p.K237K	Substitution - coding silent
SPEF2	c.712A>G	p.K238E	Substitution - Missense
SPEF2_ENST00000356031	c.2649_2650delAA	p.K886fs*2	Deletion - Frameshift
SPEF2_ENST00000356031	c.2649delA	p.K886fs*8	Deletion - Frameshift
SPEF2_ENST00000356031	c.711G>A	p.K237K	Substitution - coding silent
SPEF2_ENST00000356031	c.712A>G	p.K238E	Substitution - Missense
SPINK5	c.2459delA	p.K823fs*101	Deletion - Frameshift
SREK1IP1	c.363A>C	p.K121N	Substitution - Missense
SYCP1	c.2891_2892insA	p.L968fs*5	Insertion - Frameshift
SYCP1	c.2892delA	p.K967fs*2	Deletion - Frameshift
SYCP2	c.3062_3063insT	p.N1024fs*3	Insertion - Frameshift
SYCP2	c.3071delA	p.N1024fs*26	Deletion - Frameshift
TAF1B	c.187_188delAA	p.N66fs*3	Deletion - Frameshift
TAF1B	c.187_189delAAA	p.K65delK	Deletion - In frame
TAF1B	c.187delA	p.N66fs*26	Deletion - Frameshift
TAF1B	c.198C>A	p.N66K	Substitution - Missense
TAF1D	c.281_282insA	p.K95fs*28	Insertion - Frameshift
TAF1D	c.281delA	p.K94fs*26	Deletion - Frameshift
TAF7L	c.1331A>C	p.Q444P	Substitution - Missense

TAF7L	c.1333A>C	p.K445Q	Substitution - Missense
TAOK1_ENST00000261716	c.1738G>A	p.E580K	Substitution - Missense
TCF25	c.384_385insA	p.Q132fs*27	Insertion - Frameshift
TCF25	c.385delA	p.K131fs*17	Deletion - Frameshift
TCF25	c.393_394insA	p.Q132fs*27	Insertion - Frameshift
TCOF1	c.4127delA	p.E1379fs*>33	Deletion - Frameshift
TCOF1_ENST00000504761	c.4358delA	p.E1456fs*>33	Deletion - Frameshift
TCP1	c.730A>C	p.T244P	Substitution - Missense
TDRD5	c.1243G>A	p.E415K	Substitution - Missense
TET1_ENST00000373644	c.57C>A	p.N19K	Substitution - Missense
TET1_ENST00000373644	c.58delA	p.K22fs*23	Deletion - Frameshift
TEX10	c.11_13delAAA	p.K4delK	Deletion - In frame
TEX10	c.13delA	p.R5fs*10	Deletion - Frameshift
TEX15	c.5122delA	p.R1708fs*3	Deletion - Frameshift
TFAM	c.432delA	p.E148fs*2	Deletion - Frameshift
TFAM	c.441+2T>G	p.?	Unknown
TFAM_ENST00000395377	c.376delA	p.T126fs*6	Deletion - Frameshift
THOC2	c.2531A>G	p.K844R	Substitution - Missense
THOC2_ENST00000245838	c.2768A>G	p.K923R	Substitution - Missense
TIF1	c.2574G>A	p.K858K	Substitution - coding silent
TMEM97	c.518_519insA	p.*177fs?	Insertion - Frameshift
TMEM97	c.519delA	p.K176fs?	Deletion - Frameshift
TNRC6A	c.104A>C	p.K35T	Substitution - Missense
TNRC6B_ENST00000301923	c.179A>G	p.K60R	Substitution - Missense
TNRC6B_ENST00000301923	c.180G>A	p.K60K	Substitution - coding silent
TNRC6B_ENST00000301923	c.182A>C	p.K61T	Substitution - Missense
TNRC6B_ENST00000454349	c.113A>G	p.K38R	Substitution - Missense
TNRC6B_ENST00000454349	c.114G>A	p.K38K	Substitution - coding silent
TNRC6B_ENST00000454349	c.116A>C	p.K39T	Substitution - Missense
TPR	c.2055A>C	p.E685D	Substitution - Missense
TPR	c.2056A>T	p.K686*	Substitution - Nonsense
TRDN	c.1145A>G	p.K382R	Substitution - Missense
TRIM24	c.2676G>A	p.K892K	Substitution - coding silent
TRIM59	c.594G>T	p.Q198H	Substitution - Missense
TRIM59	c.604delA	p.S202fs*3	Deletion - Frameshift
TRMT6	c.458delA	p.K153fs*9	Deletion - Frameshift
TRPC1	c.509A>G	p.K170R	Substitution - Missense
TTF1	c.1007_1008insA	p.K337fs*9	Insertion - Frameshift
TTF1	c.1007delA	p.K336fs*87	Deletion - Frameshift
TWISTNB	c.932_933insA	p.K312fs*6	Insertion - Frameshift
TWISTNB	c.932delA	p.K311fs*15	Deletion - Frameshift
TWISTNB	c.933G>T	p.K311N	Substitution - Missense

TWISTNB	c.935_936insG	p.R313fs*5	Insertion - Frameshift
TWISTNB	c.938G>A	p.R313K	Substitution - Missense
ULK4_ENST00000301831	c.1778delA	p.K593fs*17	Deletion - Frameshift
USP36	c.2874_2879delGAAAAA	p.K959_K960delKK	Deletion - In frame
USP36_ENST00000312010	c.2874_2879delGAAAAA	p.K959_K960delKK	Deletion - In frame
USP36_ENST00000312010	c.2874G>A	p.K958K	Substitution - coding silent
USP40	c.3468G>T	p.K1156N	Substitution - Missense
USP40	c.3477_3478insA	p.Q1160fs*20	Insertion - Frameshift
USP40	c.3477delA	p.K1159fs*12	Deletion - Frameshift
USP40_ENST00000450966	c.3468G>T	p.K1156N	Substitution - Missense
USP40_ENST00000450966	c.3477_3478insA	p.Q1160fs*20	Insertion - Frameshift
USP40_ENST00000450966	c.3477delA	p.K1159fs*12	Deletion - Frameshift
VAX1	c.473A>C	p.K158T	Substitution - Missense
VAX1	c.477A>G	p.K159K	Substitution - coding silent
VAX1	c.477delA	p.K159fs*>28	Deletion - Frameshift
VAX1	c.478C>G	p.Q160E	Substitution - Missense
VAX1	c.480A>C	p.Q160H	Substitution - Missense
VAX1	c.483G>T	p.K161N	Substitution - Missense
WDHD1	c.1827G>A	p.K609K	Substitution - coding silent
WNK1	c.1738_1749ins?	p.?	Unknown
WNK1	c.1739delA	p.K583fs*11	Deletion - Frameshift
WNK1	c.1740A>G	p.E580E	Substitution - coding silent
WNK1	c.1749G>T	p.K583N	Substitution - Missense
WNK1_ENST00000537687	c.1739delA	p.K583fs*11	Deletion - Frameshift
WNK1_ENST00000537687	c.1740A>G	p.E580E	Substitution - coding silent
WNK1_ENST00000537687	c.1749G>T	p.K583N	Substitution - Missense
ZC3H13	c.3719G>C	p.S1240T	Substitution - Missense
ZC3H13	c.3719G>T	p.S1240I	Substitution - Missense
ZCCHC9	c.189G>T	p.K63N	Substitution - Missense
ZCCHC9	c.196G>T	p.E66*	Substitution - Nonsense
ZCRB1	c.411G>A	p.K137K	Substitution - coding silent
ZCRB1	c.419_420insA	p.K141fs*4	Insertion - Frameshift
ZCRB1	c.419delA	p.K140fs*13	Deletion - Frameshift
ZFHX3	c.1031A>T	p.N344I	Substitution - Missense
ZFR	c.1074_1075insA	p.E359fs*27	Insertion - Frameshift
ZFR	c.1074delA	p.E359fs*4	Deletion - Frameshift
ZMAT1	c.1276A>T	p.K426*	Substitution - Nonsense
ZMAT1	c.1277A>T	p.K426I	Substitution - Missense
ZMAT1	c.1278A>G	p.K426K	Substitution - coding silent
ZMAT1_ENST00000372782	c.1789A>T	p.K597*	Substitution - Nonsense
ZMAT1_ENST00000372782	c.1790A>T	p.K597I	Substitution - Missense

ZMAT1_ENST00000372782	c.1791A>G	p.K597K	Substitution - coding silent
ZMYM5_ENST00000337963	c.1934delA	p.N645fs*>25	Deletion - Frameshift
ZNF236	c.1373delA	p.M461fs*1	Deletion - Frameshift
ZNF236_ENST00000543926	c.1373delA	p.M461fs*1	Deletion - Frameshift
ZNF34	c.664delA	p.T222fs*15	Deletion - Frameshift
ZNF518A	c.2777delA	p.T929fs*2	Deletion - Frameshift
ZNF518A_ENST00000371192	c.2777delA	p.T929fs*2	Deletion - Frameshift
ZNF600	c.194_195insA	p.L66fs*4	Insertion - Frameshift
ZNF644	c.871delA	p.R291fs*7	Deletion - Frameshift
ZNF644	c.872G>T	p.R291I	Substitution - Missense

Supplemental Table 2.5. Table of genes with mutations within polyA region reported in COSMIC database.

2.7 References

1. J. D. Dinman, M. J. Berry, 22 Regulation of Termination and Recoding. *Cold Spring Harb. Monogr. Arch.* 48 (2007) (available at https://cshmonographs.org/index.php/monographs/article/view/3291).

2. J. W. B. Hershey, N. Sonenberg, M. B. Mathews, Principles of translational control: an overview. *Cold Spring Harb. Perspect. Biol.* **4** (2012).

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

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

5. D. P. Letzring, K. M. Dean, E. J. Grayhack, Control of translation efficiency in yeast by codon-anticodon interactions. *RNA N. Y. N.* **16**, 2516–2528 (2010).

6. L. N. Dimitrova, K. Kuroha, T. Tatematsu, T. Inada, Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. *J. Biol. Chem.* **284**, 10343–10352 (2009).

7. K. Kuroha *et al.*, Receptor for activated C kinase 1 stimulates nascent polypeptide-dependent translation arrest. *EMBO Rep.* **11**, 956–961 (2010).

8. O. Brandman *et al.*, A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell*. **151**, 1042–1054 (2012).

9. J. Lu, C. Deutsch, Electrostatics in the ribosomal tunnel modulate chain elongation rates. *J. Mol. Biol.* **384**, 73–86 (2008).

10. K. S. Koutmou, A. P. Schuller, J. L. Brunelle, A. Radhakrishnan, S. Djuranovic, R. Green. eLIFE 10.7554/eLife.05534 (2015).

11. T. Tsuboi *et al.*, Dom34:hbs1 plays a general role in quality-control systems by dissociation of a stalled ribosome at the 3’ end of aberrant mRNA. *Mol. Cell*. **46**, 518–529 (2012).

12. S. Karlin, L. Brocchieri, A. Bergman, J. Mrazek, A. J. Gentles, Amino acid runs in eukaryotic proteomes and disease associations. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 333–338 (2002).

13. S. Djuranovic, A. Nahvi, R. Green, miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science*. **336**, 237–240 (2012).

14. J. N. Barr, G. W. Wertz, Polymerase slippage at vesicular stomatitis virus gene junctions to generate poly(A) is regulated by the upstream 3’-AUAC-5’ tetranucleotide: implications for the mechanism of transcription termination. *J. Virol.* **75**, 6901–6913 (2001).

15. M. Fresno, A. Jiménez, D. Vázquez, Inhibition of translation in eukaryotic systems by harringtonine. *Eur. J. Biochem. FEBS*. **72**, 323–330 (1977).

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

17. K. D. Pruitt *et al.*, RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756–763 (2014).

18. S. K. McLaughlin *et al.*, The RasGAP gene, RASAL2, is a tumor and metastasis suppressor. *Cancer Cell*. **24**, 365–378 (2013).

19. E. J. Belfield, R. K. Hughes, N. Tsesmetzis, M. J. Naldrett, R. Casey, The gateway pDEST17 expression vector encodes a -1 ribosomal frameshifting sequence. *Nucleic Acids Res.* **35**, 1322–1332 (2007).

20. J. Chen *et al.*, Dynamic pathways of -1 translational frameshifting. *Nature*. **512**, 328–332 (2014).

21. S. A. Forbes *et al.*, COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* (2014).

22. A. T. Belew, V. M. Advani, J. D. Dinman, Endogenous ribosomal frameshift signals operate as mRNA destabilizing elements through at least two molecular pathways in yeast. *Nucleic Acids Res.* **39**, 2799–2808 (2011).

23. A. T. Belew *et al.*, Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway. *Nature*. **512**, 265–269 (2014).

24. J. Lykke-Andersen, M. D. Shu, J. A. Steitz, Human Upf proteins target an mRNA for nonsense-mediated decay when bound downstream of a termination codon. *Cell*. **103**, 1121–1131 (2000).

25. H. Le Hir, D. Gatfield, E. Izaurralde, M. J. Moore, The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. *EMBO J.* **20**, 4987–4997 (2001).

26. Y.-F. Chang, J. S. Imam, M. F. Wilkinson, The nonsense-mediated decay RNA surveillance pathway. *Annu. Rev. Biochem.* **76**, 51–74 (2007).

27. M. W.-L. Popp, L. E. Maquat, Organizing principles of mammalian nonsense-mediated mRNA decay. *Annu. Rev. Genet.* **47**, 139–165 (2013).

28. R. C. Hunt, V. L. Simhadri, M. Iandoli, Z. E. Sauna, C. Kimchi-Sarfaty, Exposing synonymous mutations. *Trends Genet. TIG*. **30**, 308–321 (2014).

Chapter 3: Rapid generation of hypomorphic mutations

Preface

The following work was performed by me, Joyce J. Chung, Preetam Janakirama, Kathryn M. Keefer, Igor Kolotilin, Slavica Pavlovic-Djuranovic, Douglas Chalker, Vojislava Grbic, Rachel Green, Rima Menassa, Heather L. True, James B. Skeath and Sergej Djuranovoc. R.G. and S.D. are responsible for the initial idea and design of the study. S.D. and S.P.D. generated and validated initial expression constructs. S.D. and L.L.A. performed experiments in *E. coli* and human cell cultures. L.L.A. performed experiments in D. melanogaster with the guidance of J.B.S. J.J.C. and D.L.C. conducted experiments in T. thermophila. P.J., V.G., I.K. and R.M. designed experiments in N. benthamiana; P.J. and I.K. completed the experiments. K.M.K. and H.L.T. performed experiments in S. cerevisiae. All authors contributed to writing the manuscript.

This chapter is published in its entirety [Arthur, L. L., Chung, J. J., Jankirama, P., Keefer, K. M., Kolotilin, I., Pavlovic-djuranovic, S., … Djuranovic, S. (2017). Rapid generation of hypomorphic mutations. Nature Communications, 8, 1–15. http://doi.org/10.1038/ncomms14112]

and is available at https://www.nature.com/articles/ncomms14112.

We thank D. Piston, M. Jovanovic, T. Schedl and P. Szczęsny and members of Djuranovic lab for helpful comments. This work was supported by the NIH grant T32

GM007067 to LLA and KMK, National Science Foundation grant # 1412336 (MCB, PI: Chalker) to JCC and DLC, the NSERC-DG to VG, NIH Grant R01 NS036570 to JBS, NIH Grant R01 GM072778 to HT, NIH Grant R01 GM112824 and the American Cancer Society (grant IRG-58-010-58-2) to SD. Subject to the U.S. Provisional Patent Application, Serial No. 62/361,307.

3.1 Abstract

Hypomorphic mutations are a valuable tool for both genetic analysis of gene function and for synthetic biology applications. However, current methods to generate hypomorphic mutations are limited to a specific organism, change gene expression unpredictably, or depend on changes in spatial-temporal expression of the targeted gene. Here we present a simple and predictable method to generate hypomorphic mutations in model organisms by targeting translation elongation. Adding consecutive adenosine nucleotides, so-called polyA tracks, to the gene coding sequence of interest will decrease translation elongation efficiency, and in all tested cell cultures and model organisms, this decreases mRNA stability and protein expression. We show that protein expression is adjustable independent of promoter strength and can be further modulated by changing sequence features of the polyA tracks. These characteristics make this method highly predictable and tractable for generation of programmable allelic series with a range of expression levels.

3.2 Introduction

Manipulation of gene activity is a standard genetic approach to gain insight into gene function in single and multicellular organisms[1]. In many cases, complete loss of gene function (null allele or knockout of the locus) will provide the most valuable information about gene

function. However, for essential genes, complete loss of function leads to lethality, which usually precludes obtaining functional information for later cellular or developmental stages. Similarly, for genes that function in multiple cellular/developmental processes and have pleiotropic null mutant phenotypes, it can be difficult to distinguish primary from secondary effects. In many of these cases, however, partial loss of function or hypomorphic mutations can overcome lethality and pleiotropy, allowing later stage cells and organisms to be examined for phenotypic consequences. Furthermore, hypomorphic mutations, because they retain residual gene activity and partial phenotypes, are used in suppressor or enhancer genetic screens to identify other genes that act in the same biological process. One example of a hypomorphic condition is a 50% reduction in gene activity from heterozygosity for a null allele, which for some genes can display mutant phenotypes (called a haploinsufficency). However, for other genes a 50% reduction in gene activity is sufficient for normal function, and thus hypomorphic mutations with a further reduction are required to observe a phenotype. For these reasons, it is important to have the ability to generate hypomorphic mutations with a range of loss of gene activity.

Hypomorphic mutations are traditionally obtained in forward genetic chemical mutagenesis screens. The hypomorphic allele needs to be isolated, identified, and then genetically and biochemically characterized in order to be further used in analysis to deduce gene function. As a consequence, this process is time consuming. In addition, because of evolution, discovered alleles can be species specific. These difficulties and the importance of hypomorphic alleles have prompted the development of several methods to generate hypomorphic mutations directly in various model organisms[2,3,4,5,6,7,8,9,10,11]. However, these methods again are usually specific to one organism, can have unpredictable alterations in gene activity (separation of function, gain of function), or can change other aspects of regulation that affect interpretation of

phenotype, such as spatial-temporal gene expression. With the rise of gene editing systems, such as CRISPR/Cas9 (ref.12) or TALEN technology[13], and more extensive use of synthetic and systems biology approaches[14,15], there is an increasing interest in generating hypomorphic mutations of a target gene through simpler, more systematic and rapid approaches.

Here we present a method for the generation of hypomorphic mutations that produces a range of reduced levels of nearly wild type protein through the use of disrupted translational elongation. This method is based on polyA tracks, a novel cis regulatory element that decreases gene expression by disrupting messenger RNA (mRNA) translation[16,17]. Insertion of consecutive adenosine nucleotides into the open reading frame of an mRNA will decrease protein expression by decreasing the efficiency of the translation elongation phase leading to diminished production of protein and mRNA destabilization, and thus to diminished mRNA levels.

3.3 Results

3.3.1 Use of polyA tracks to generate hypomorphic mutants

We have recently identified polyA tracks as a regulator of gene expression[16,17]. This mechanism is used endogenously in most eukaryotic genomes and regulates ~2% of human genes[16,18]. The polyA track causes ribosomal stalling and frameshifting during translation elongation, leading to mRNA instability and degradation of nascent protein products[16,17]. The translation elongation cycle is an ideal target for a universal method of gene regulation because it is the most highly conserved step in protein biosynthesis between bacteria and eukaryotes[19]. Therefore, we reasoned that polyA tracks, because of their versatility in lengths and sequence composition, can be used as a system to create programmable hypomorphic mutants and regulate gene expression in a wide variety of model organisms (Fig. 1).

We have generated a fluorescent reporter gene that has an insertion of defined polyA tracks in order to control the amount of expression (Fig. 1a). The reporter consists of either a constitutive or inducible promoter driving expression of the mCherry fluorescence and other reporter proteins. A double HA-tag was added at the beginning of the coding sequence for detection through western blot analysis. The polyA track is inserted directly after the HA-tag. The length of the polyA track varies from 9 to 36 consecutive adenosine nucleotides, adding 3–12 lysine residues to the protein sequence (Fig. 1a). To assess the effects of polybasic peptide arising from sequential lysine residues[20,21], we also generated control reporters with consecutive lysine AAG codons. We hypothesized that as the length of the polyA track is increased, expression of the reporter gene products will decrease (Fig. 1b,c). These reporters can be transiently transfected, recombined or inserted into the genome of cell cultures or whole organisms. Likewise, endogenous genes can be edited to include a polyA track in their open reading frames (ORF's) using genome editing methodology.

3.3.2 PolyA tracks can regulate gene expression in *Escherichia coli*

We first tested whether polyA tracks can be used in single-cell model organisms to attenuate gene expression from a defined reporter gene. Previous attempts to control gene expression in Escherichia coli cells have used degeneration of ribosome binding sequence[2] (rbs) or spacer sequence between rbs sequence and starting AUG codon[22], but did not focus on the coding sequence. To show that polyA tracks can be used to control gene expression in *E. coli* cells, we created a set of reporters with increasing length of polyA tracks under the arabinose-inducible promoter pBAD (Supplementary Fig. 1). We transformed *E. coli* cells with plasmids expressing HA-mCherry, HA-$(AAG)_n$-mCherry or HA-$(AAA)_n$-mCherry. All *E. coli* cell cultures were induced at the same optical density and monitored for both cell growth and

fluorescence of the mCherry constructs during induction. While *E. coli* containing wild type and LysAAG controls (6×, 9× and 12× lysine AAG codons) show no significant differences in the amount of mCherry fluorescence, cell cultures containing constructs with polyA tracks show progressively less fluorescence with increasing length of the polyA track (Fig. 2a). Addition of 9 and 12As in a row (3 or 4 Lys_{AAA} codons) consistently reduced fluorescence of mCherry reporter by 15–35%. Further additions in the length of the polyA track resulted in continuing decreases in mCherry reporter fluorescence to the point where 36 consecutive adenosine nucleotides resulted in barely visible expression of the reporter (<5% of wild type). Western blot analyses of equal amounts of *E. coli* cell lysates expressing different polyA track and control reporters were consistent with the mCherry fluorescence data and indicated again that protein abundance of reporter proteins strongly depends on the length of the polyA track (Fig. 2b). Reporters with 9 and 12As in a row (3 and 4Lys_{AAA} codons, respectively) show reduction in protein abundance in the range of 20–40% of the wild type mCherry, and constructs with >27As in a row (9 and more Lys_{AAA} codons) were nearly undetectable by western blot analyses.

To further fine tune expression of the reporter gene with polyA tracks, we tested whether insertion of non-lysine codons in the middle of long polyA tracks (33As, 11 AAA codons) would result in change of the reporter accumulation in *E. coli* cells. Indeed, addition of such codons did not change levels of the fluorescent reporter drastically but fine-tuned them to the fluorescent values between constructs with 30As or 33As in a row (Supplementary Fig. 2). Additional analysis of XAA and AAY codons, where X and Y denote C/G or T/C/G nucleotides respectively, for programming the length of polyA tracks revealed that gene attenuation by polyA tracks does not depend necessarily on lysine codons but other codons can contribute to the overall length and effects of polyA tracks (Supplementary Fig. 2). In a similar manner,

endogenous genes with different arrangement of A–rich codons (mainly lysine encoding AAA and AAG codons) have shown differential gene regulation in our previous study[17].

3.3.4 PolyA tracks can be used in protozoan *T. thermophila*

We previously showed that polyA tracks can influence expression of the reporter genes in Saccharomyces cerevisiae cells[17]. To test whether polyA tracks can regulate gene expression in another model, single-cell eukaryotic organism, we monitored the effect of various length tracks on YFP expression in the protozoan Tetrahymena thermophila. The genome of *T. thermophila* has extremely high AT content (>75%) and has been extensively used as a microbial animal model[23]. Our *T. thermophila* reporter contained the coding sequence of a Macronucleus-Localized Protein of unknown function (MLP1, TTHERM_00384860) fused to eYFP protein (Supplementary Fig. 3). The fusion with MLP1 directed YFP to *Tetrahymena* macronuclei to allow easier quantification of YFP levels (Fig. 2c). These two proteins were fused, separated by linkers containing an HA-tag (MLP1-HA-YFP (WT)) and polyA tracks of 18, 27 or 36As, $(AAA)_6$, $(AAA)_9$ or $(AAA)_{12}$, respectively, or 12 Lys_{AAG} $(AAG)_{12}$ codons inserted as a control. All constructs were expressed upon cadmium-induction of the upstream MTT1 promoter. Just as in our *E. coli* experiments with mCherry reporter, the YFP gene containing increasing lengths of polyA tracks exhibited a progressive decrease in total protein accumulation, measured by fluorescence, relative to the HA-linker fusion or Lys_{AAG} insertion controls (Fig. 2d). The construct with 18As in a row ($6Lys_{AAA}$) showed ~50% reduction in protein fluorescence, while constructs with 27 and 36As exhibited nearly undetectable levels of fluorescence and required 20 times longer exposure for detection of YFP by microscopy. The construct with $12Lys_{AAG}$ codons showed fluorescence that was 4–5 fold lower than the WT construct. This effect was comparable to the polybasic peptide stalling that was observed earlier in *S. cerevisiae* cells[17,20,21]. We further

supported our fluorescence results by western blot analyses (Fig. 2e). The MLP1-HA-YFP-fusion (WT) was readily visible, whereas the polyA track-YFP fusions exhibited attenuated expression. Insertion of 18A's ($6Lys_{AAA}$) showed expression levels that were ~35% of the HA-YFP control, while 27 and 36A's (9 and $12Lys_{AAA}$) constructs were at the limit of detection (Fig. 2e). Insertion of $12Lys_{AAG}$ codons in the fusion protein resulted in a 6- to 8-fold reduction in expression of fusion protein. Since the different fusion proteins were all transcribed from the MTT1 promoter under identical induction conditions we reasoned that the amount of mRNA produced for each construct should be equivalent. We used qRT-PCR to quantify the steady-state level of YFP mRNA for each fusion (Fig. 2f). The steady-state mRNA levels of polyA track-YFP constructs reflected the decreasing YFP protein accumulation relative to the WT consistent with earlier reports in *S. cerevisiae*[17,20]. Insertion of 18As ($6Lys_{AAA}$ codons) reduced mRNA levels to approximately 30–35% of WT levels, while insertion of 27 and 36As (9 and $12Lys_{AAA}$ codons, respectively) reduced mRNA levels to <5% of the HA-YFP construct (Fig. 2d). Interestingly, while attenuation at the protein level was stronger for the insertion of $12Lys_{AAG}$ codons than for $6Lys_{AAA}$ codons, the trend was not consistent at the mRNA level. The observed discrepancies between the protein levels and mRNA degradation in the case of polybasic peptide (12 AAG codon) are likely attributed to the two distinct pathways employed in regulation of polybasic peptide stalling and polyA track induced stalling and frameshifting in eukaryotic cells[16,17,24]. Nevertheless, the effects of polyA tracks on regulation of the reporter gene showed a consistent relationship between mRNA levels and protein accumulation levels. Most importantly, we were able to control expression of reporter genes using polyA tracks in *T. thermophila*, a single-cell AT-rich protozoan, as we previously reported in *S. cerevisiae* and the above-mentioned *E. coli* cells[17].

3.3.5 PolyA tracks can regulate gene expression in plant tissues

To test whether polyA tracks can attenuate gene expression in plants, we transiently co-expressed HA-$(AAA)_n$-mCherry with an YFP construct as an internal control in the model plant Nicotiana benthamiana (Supplementary Fig. 4). The expression of mCherry and YFP was assessed by fluorescence imaging (Fig. 3a). Like cell cultures, *N. benthamiana* epidermal cells showed attenuated mCherry fluorescence proportional to the length of the polyA tracks (6,9 and 12 Lys_{AAA} codons) compared with the HA-mCherry and 12 Lys_{AAG} control constructs (Fig. 3a). The fluorescence data for each construct revealed the same trend of gene expression regulation as in *T. thermophila* cells (Fig. 2d). As fluorescence in this assay was not quantifiable, protein abundance was determined by semi-quantitative western blot analysis of *N. benthamiana* leaves infiltrated with the HA-$(AAA)_n$-mCherry. The levels of HA-mCherry proteins were normalized to levels of the cis-linked selectable marker phosphinotricin acetyl transferase (BAR) in the same sample (Fig. 3b,c). The addition of a polyA track with 18As (6 Lys_{AAA}) decreased protein accumulation to ~70% of HA-mCherry levels. Further reduction of mCherry protein accumulation, to 30% and below the detection limit was observed in 9 Lys_{AAA} and 12 Lys_{AAA} constructs, respectively (Fig. 3c). Insertion of 12 Lys_{AAG} codons displayed ~50% reduction in the reporter expression compared with WT construct. Parallel analyses of steady-state mRNA levels of transcripts with increasing lengths of polyA tracks showed progressively reduced levels of polyA track mRNAs when compared with transcript levels of the HA-mCherry and AAG-containing control constructs (Fig. 3d). mRNA levels were reduced to ~50–55% of WT expression for 6Lys_{AAA} transcripts, while 9 and 12Lys_{AAA} constructs had reduced mRNA levels to ~30 and 20% of control, respectively (Fig. 3d). Insertion of 12 Lys_{AAG} codons exhibited marginal effects on mRNA levels and again showed the already observed discrepancy between

mRNA levels and protein accumulation as a result of polybasic peptide stalling. However, these results indicate once again that polyA tracks affect both mRNA and protein levels more consistently and can be used to regulate gene expression in plants.

3.3.6 PolyA tracks can regulate genes in human tissue cultures

To further assess the universality of polyA tracks on protein expression, we tested our reporter series in human tissue cultures using HeLa cells. Plasmids with HA-mCherry, HA-$(AAG)_{12}$-mCherry and HA-$(AAA)_n$-mCherry reporters, driven by the constitutively active CMV promoter, were electroporated into HeLa cells for transient expression. Protein abundances were assessed by western blot analyses 24 h after electroporation (Fig. 3e). As in our previous study on expression of endogenous and synthetic polyA tracks in various human tissue cultures16, constructs with increasing length of polyA tracks (6, 9 and 12 Lys_{AAA}) were expressed at lower levels than control constructs and the reductions in protein expression were proportional to the length of polyA track. The construct with 18As ($6Lys_{AAA}$) displayed an ~3-fold reduction in expression compared with the WT construct. Insertion of 27 and 36As (9 and $12Lys_{AAA}$, respectively), exhibited a 6- and 25-fold reduction of HA-mCherry expression compared with WT (Fig. 3f). The control construct with $12Lys_{AAG}$ codons did not show any reduction in protein levels compared with the WT construct (Fig. 3e,f). These results again argue for differences between translational stalling induced by polybasic peptides[16,17,20,21], which seems to be cell- or organism-specific and unpredictable, and polyA track-induced ribosomal stalling and frameshifting[16,17], which is clearly dependent on the length of polyA tracks. Importantly, this latter pathway appears to be conserved between multiple organisms. Altogether with our previous study[16], our results indicate that polyA tracks can readily be used to regulate expression of reporters or genes transiently transfected in diverse eukaryotic tissues and cultured cell

systems, such as *N. benthamiana* and human cell cultures, as well as other mammalian or insect tissue culture systems[16].

3.3.7 PolyA tracks can regulate gene expression in model organisms

We next sought to test whether polyA tracks can be used to regulate reporter gene expression in complex, multicellular organisms. We chose the fruit fly, *Drosophila melanogaster*, because of the well-developed tools in the manipulation of endogenous genetic loci, as well as for the ready assessment of the mCherry reporter activity. Using the PhiC31-integrase approach[25], we generated single transgene insertions of the HA-mCherry and HA-$(AAG)_{12}$-mCherry controls, and HA-$(AAA)_n$-mCherry (6, 9 and 12Lys_{AAA}) constructs in the identical genomic location in the third chromosome (see methods; Supplementary Fig. 5). All constructs contained an Upstream Activation Sequence (UAS) followed by the HSP70 promoter which actively transcribes mCherry reporter mRNAs in response to expression of GAL4 protein[26]. To drive expression of mCherry in all tissues, each transgenic line was crossed to a driver line with Tub-GAL4 to express GAL4 protein in all tissues. In addition, the driver line carried a UAS-linked GFP transgene, which allowed us to use GFP expression for normalization of the mCherry reporter genes (Supplementary Fig. 5).

Expression of mCherry was assessed by fluorescence imaging of formaldehyde fixed salivary glands (SG), central nervous system (CNS) and proventriculus (PV) dissected from otherwise wild type third instar larvae (Fig. 4a). Wild type HA-mCherry expressed well in all imaged tissues. Addition of a polyA track with 18A's (6Lys_{AAA}) reduced mCherry expression to ~30% of the wild type construct in all three tissues (Fig. 4b–d; Supplementary Fig. 6). Constructs with 27As and 36As (9 and 12Lys_{AAA} codons, respectively) reduced expression of

mCherry in all assayed tissues to ~20 and 10% of wild type levels, respectively (Fig. 4b–d; Supplementary Fig. 6). Western blot analyses on cell lysates produced from five fruit fly larvae for each independent construct were consistent with our quantification of fluorescence imaging data (Supplementary Fig. 7). As in the previous experiments with *T. thermophila* and tissue culture systems (Figs 2 and 3), mRNA stability of polyA track constructs in fruit fly larvae exhibited an inverse correlation with the length of polyA track (Supplementary Fig. 8) and concordance with protein abundances measured by western blot analyses. Insertion of $12Lys_{AAG}$ codons had a moderate effect on levels of mCherry mRNA and protein and was similar to the expression levels for the 18As ($6Lys_{AAA}$) insertion construct (Fig. 4b–d; Supplementary Figs 7 and 8). Our data indicate that individual tissues of a complex multicellular organism, such as fruit fly, are equally sensitive to gene expression attenuation mediated by polyA tracks. Therefore, one can use polyA track constructs to create hypomorphic alleles and allelic series, in complex multicellular organisms with similar relative gene expression attenuation efficiencies as observed in the isolated tissues.

One of the potential limitations of polyA tracks could be their homopolymeric nature and observed hypermutability of short tandem repeats (STRs) and homopolymeric sequences that was previously observed in different systems[22,27,28]. To control for the cell population heterogeneity that may arise from hypermutability of polyA tracks, we have analysed the mutation rate of a single locus insertion of our longest polyA track insertion (36As) in the D. melanogaster genome. The sequence of the mCherry reporter gene from genomic DNA isolated from a whole adult fly after more than a year of homozygote crosses, approximately 30 generations of fruit flies, was compared with the original DNA construct that was used for generation of the transgenic insect. Both Sanger and Illumina sequencing of the amplified polyA

track (36As) regions from the transgenic fruit fly genome and original DNA vector revealed differences between two data sets (Supplementary Fig. 9). In ~8% of the cases by Illumina sequencing we have observed loss of the polyA track because of the possible recombination event indicated by the similarity of the amplified sequence to the genomic sequence located on the chromosome X and 3L of the fruit fly genome. Illumina sequencing was not reliable enough to show differences in the polyA tracks length because of the difficulty associated with sequencing of the homopolymers. The Sanger sequencing of the amplified polyA track regions from plasmid and genomic DNA used for the Illumina sequencing revealed a low frequency of mutations in the polyA tracks. We observed insertions as well as polyA track shortening, which is in good concordance with previously published data from different cellular systems[22,27,28,29]. This is in agreement with our previous analysis of endogenous polyA track genes, which showed rather strong conservation of polyA track sequences in multiple analysed genomes. Taken together, these data indicate that polyA tracks can be used for stable control of gene expression over a multiple generations of model organisms and for an extended period of time without a strong mutational drift in cell mutability.

3.3.8 PolyA track control is independent of promotor strength

Our data from the fruit fly experiment indicated that the ratio between reporters with polyA track insertion and control is maintained in all tissues (Fig. 4b–d; Supplementary Figs 6–8). This suggests that inserted polyA tracks maintain their capacity for gene regulation independent of the strength of DNA transcription, which is known to have a large dynamic range across genes and cell types[30].

To systematically evaluate how differences in the strength of transcription would affect gene regulation and hypomorphic expression of reporters with polyA track insertion, we used human Flp-In T-REx 293 cell lines (Thermo Fisher Scientific). Using a protocol for generation of stable and inducible expression cell lines, we generated cells with a single insertion of our HA-mCherry control and HA-$(AAA)_{12}$-mCherry polyA construct in a defined chromosomal locus (Supplementary Fig. 10). The strength of transcription in these cell lines was varied by use of increasing concentrations of doxycycline (0.001–0.1 mg μl^{-1} of Dox) added to the growth media, and levels of transcription were assayed in relation to constitutively expressed hygromycin B phosphotransferase (Supplementary Fig. 10). The dose-dependent response of the doxycycline-inducible CMV promoter for both the polyA track and control mCherry transcript ranged over two orders of magnitude. At the same time, relative expression of the polyA track construct was constant; 12–17% expression relative to the control construct based on the western blot analysis (Fig. 5a,b). Moreover, relative mRNA levels of control and polyA track constructs did not change under different transcriptional regimes (Fig. 5c). The steady-state amount of the polyA track construct mRNA was consistently in a range between 1 and 3% of the normalized control construct. The same results are obtained using stable cell lines that express HA-tagged human haemoglobin (delta chain, WT-HBD) and an 18As HBD construct (HBD-6Lys_{AAA}) with polyA track inserted in the second exon of the HBD coding sequence (Supplementary Fig. 11). Expression of the HBD-6Lys_{AAA} protein was 2- to 3-fold reduced compared with the WT-HBD construct, based on western blot analysis (Supplementary Fig. 12), and mRNA levels were ~5-fold lower than HBD-WT mRNA levels, measured by qRT-PCR (Supplementary Fig. 13). The relative ratios of WT-HBD and HBD-6Lys_{AAA} protein and mRNA levels were constant for all doxycycline induction levels. Altogether with previous data, showing regulated expression of

mCherry reporter in different tissues of the transgenic fruit fly, these data demonstrate that polyA tracks can control gene expression independently of the promoter strength associated with the assayed gene.

3.3.10 PolyA tracks create hypomorphic mutants in functional genes

The polyA tracks are mainly composed of lysine residues, AAA or AAG codons, which can be problematic when expressed as tags because of their charge and propensity for acquiring specific modifications (ubiquitination, acetylation, SUMOylition and hydroxylation). These features of poly-lys chains can potentially influence protein function as well as cell homoeostasis[24,31,32,33].

We tested our ability to regulate gene expression of functional proteins in both the bacterial and eukaryotic cell systems. In *E. coli*, the chloramphenicol acetyltransferase (CAT) gene confers resistance to the broad spectrum antibiotic chloramphenicol (CAM) in a dose-dependent manner[34]. Moreover, the CAT protein is functional only as a homotrimer which has a rather complex protein structure (PDB: 3CLA)[35]. As such, regulation of this gene and subsequent protein folding could be challenging because of the additional lysine residues introduced either by polyA track or control AAG constructs. To show that we can regulate expression of CAT protein by insertion of polyA tracks, we assessed *E. coli* survival under increasing concentrations of CAM in comparison with the wild type CAT gene. To control for the influence of additional lysine residues in the N-terminus of the CAT protein, we also inserted 10LysAAG codons in the N-terminus of CAT. Expression of WT-CAT, $(AAG)_{10}$-CAT and $(AAA)_n$-CAT constructs was driven by the inducible arabinose promoter (pBAD, Supplementary Fig. 14). All *E. coli* cultures were pulse induced, with addition of 0.1% arabinose, and growth was monitored on LB plates

using different concentrations of CAM in the media. WT-CAT and AAG_{10}-CAT control constructs were able to survive CAM selection to the same extent (75 mg ml^{-1} CAM, Fig. 6a). Therefore, the function of CAT protein is not affected by the addition of 10 consecutive Lys residues. By contrast, polyA track constructs led to increased CAM sensitivity of *E. coli* cells which correlated nicely with the length of the polyA tracks inserted in the CAT gene (Fig. 6a). While the majority of constructs could grow on minimal addition of CAM in the media (15 mg ml^{-1}), constructs with 24, 27 or 30As (8, 9 and $10Lys_{AAA}$) were unable to grow on LB-plates with a CAM concentration of 30 mg ml^{-1}. Furthermore, survivability of *E. coli* cells with CAT constructs having 15, 18 and 21As (5, 6 and $7Lys_{AAA}$) on one hand, and 9 and 12As (3 and $4Lys_{AAA}$) on the other hand, was impaired when cells were grown on LB-plates with final CAM concentrations of 50 mg ml^{-1} or 75 mg ml^{-1}, respectively (Fig. 6a). The survivability of *E. coli* cultures with different CAT constructs was in concordance with expression levels of CAT protein assayed by western blot analyses (Supplementary Fig. 15). The insertion of $10Lys_{AAG}$ codons in the CAT gene did not affect *E. coli* cell growth on CAM selective media or levels of CAT protein expression, arguing that insertion of multiple lysine residues in the N-terminus is not detrimental for the function, structure and stability of CAT protein. These data demonstrate that polyA tracks can regulate levels of certain enzyme expression (CAT) in *E. coli* cells proportionally with their length.

To test the ability of polyA tracks to regulate expression and function of protein in a eukaryotic cell we monitored how polyA tracts affect expression of N-succinyl-5-aminoimidazole-4-carboxamide ribotide synthetase (Ade1, SAICAR) in *S. cerevisiae* (Fig. 6b). The Ade1 protein is a monomeric protein with multiple domains in the structure[36] (PDB: 1A48) and requires strict folding of an active site for catalysis of the seventh step in the purine

biosynthesis pathway. Disruption of the ADE1 gene results in the storage of a red pigment because of the buildup of a metabolic byproduct of the adenine biosynthesis pathway. Colonies of yeast that are *ade1Δ* are a dark red colour; reintroduction of functional Ade1 protein restores the wild type white coloration in a dose-dependent manner. Such differences in colony colour and ability to grown on adenine dropout media (SD-Ade) have been utilized to differentiate strains of yeast prions[37], assess mitotic stability[38] and monitor gene expression[39].

To survey how polyA tracts affect expression of Ade1, we transformed *ade1Δ* strains of *S. cerevisiae* with single copy plasmids (p416) encoding polyA-ADE1-FLAG, with the polyA tracks containing 18, 27 or 36As (6, 9 or 12 Lys_{AAA}, Supplementary Fig. 16). Control plasmids contained no insertions (WT) or 12 Lys_{AAG} codons, which showed moderate attenuation of reporter expression in our previous study[17]. Transformants were spotted onto plates to monitor colour phenotypes and growth on media lacking adenine (SD-Ade, Fig. 6b). The empty vector control exhibited a dark red coloration and inability to grow on SD-Ade, consistent with disruption of the ADE1 locus, while the wild type Ade1-FLAG restored both the white phenotype on media containing adenine (SD-Ura) and growth on SD-Ade. Yeast containing constructs with polyA track lengths of 18, 27 and 36As showed progressively pinker coloration and poorer growth on SD-Ade; however, the control 12 Lys_{AAG} construct conferred a nearly-WT white colour and strong growth on SD-Ade (Fig. 6b). Dot blot analysis of Ade1 protein expression, normalized to total protein, was in accordance with our phenotypic results and revealed visibly reduced amounts of expression for constructs with insertion of 9 and 12 Lys_{AAA} codons (Supplementary Fig. 16). Expression of Ade1 protein with 12 lysine residues at the N-terminus, as in the case of 12 Lys_{AAG}, did not impair folding or function of the assayed protein (Ade1) and exhibited similar phenotypes relative to the insertion of 6 Lys_{AAA} codons. Therefore,

addition of polyA tracks to functional genes in both *E. coli* and *S. cerevisiae* preserved protein structure and function but regulated protein abundance. While behaviour of some proteins will of course be sensitive to the inclusion of multiple lysines within their structure, polyA tracks can potentially be quite broadly used in the creation of hypomorphic gene mutants with fixed levels of protein expression.

3.4 Discussion

We have presented a rapid method of generating hypomorphic mutations in a reporter or gene of interest. Insertion of a polyA track into a coding sequence exhibits predictable and robust attenuation of gene expression in all tested cell culture and model organism systems. The length and the sequence of the polyA track can be manipulated to achieve full range of expression levels, allowing for the generation of an allelic series from the level of a complete knockout to wild type expression for the study of gene function. This method can also be used in synthetic biology applications that require precise gene control and modelling of metabolic and signalling networks[14,15].

The use of polyA tracks overcomes many of the challenges present in current methods of generating hypomorphic mutations and controllable gene expression. For instance, an approach tested for attenuating gene expression in *E. coli* involves mutagenesis of the Shine-Dalgarno sequence (ribosome binding sequence, rbs) in the gene of interest[2,22]. The expression levels from all possible six-mer Shine-Dalgarno sequences were experimentally determined and the information is available in the EMOPEC database[2]. However, this valuable resource would have to be generated anew to use this approach in other bacteria and it cannot be applied to eukaryotic systems. In addition, many orthogonal translation systems that are used to perturb gene networks

rely on modified rbs (refs 40, 41, 42, 43). Use of an orthogonal translation system would prohibit use of the rbs for expression regulation. A similar method has overcome this problem by using tandem repeats to change the distance between the rbs and starting ATG codon and, thereby, controlling gene expression levels[22]. However, the restricted use in one system, *E. coli* cells, still limits application of this method. The polyA track system of gene regulation that we describe here for the creation of hypomorphic mutations overcomes these issues because of its dependency on regulation of translation elongation cycle which is well conserved between bacteria and eukaryotes[19].

Hypomorphic mutations have been generated in eukaryotic cell systems by insertion of an antibiotic resistance gene into an intron[10] or the 3′-untranslated region of a gene[3]. Insertion of the neomycin resistance gene (neo) into an intron introduces a cryptic splice site that results in aberrant splicing of transcripts, effectively reducing gene expression[10]. The reliance on stochastic cryptic splicing events leads to unpredictable changes in transcript expression and is rather gene dependent. Insertion of neo in various genes have resulted in expression of a functionally null allele[44], hypomorphic expression[10,45] or no change in expression[46]. Our system of polyA tracks gives predictable gene expression attenuation in a variety of different eukaryotic systems, and furthermore shows similar relative gene expression attenuation efficiencies in different tissues of the same organism.

We have primarily introduced polyA tracks at the N-terminal regions of reporter genes because of the uniformity of the construct design and to reduce potential frameshifting effects[16,17]. We do not anticipate this to be a major limitation of this method. Our *Tetrahymena* reporters place the polyA tracts at the N-terminus of YFP, but at the C-terminus of the linked *Tetrahymena* gene (Supplementary Fig. 3). Furthermore, insertion of polyA track in the second

exon of the human beta globin gene (HBD) gene, an unstructured loop of the protein, argue that polyA tracks can be effectively introduced at various positions in the gene (Supplementary Figs 11–13). In addition, we have shown previously that naturally occurring polyA track sequences exist in the human genome and that potential frameshifted products are efficiently degraded by non-sense mediated decay mechanisms[16].

The addition of a polyA track to the target gene will result in additional lysine residues in the protein product. Like any protein tag, it is important to consider the effects of the additional residues when studying the functionality of the protein. We have shown that the function and stability of two structurally diverse proteins, CAT and Ade1, are not affected by up to 12 additional lysine residues. To control for possible effects of the poly-lysine tracks, investigators can create an allele with the same number of lysine residues encoded by AAG codons. The AAG codons will have minimal effect on expression levels while encoding a synonymous protein. Furthermore, the flexibility in polyA track placement within the coding sequence allows investigators to choose the most suitable insertion site for the protein of interest.

The conservation of the polyA track sequences in the multiple genes across vertebrates as well as our analysis of mutation rates of polyA tracks (36As) inserted in the defined locus of *D. melanogaster* genome argue that polyA tracks can be used to create stable hypomorphic gene alleles. Our results are in the range of already described hypermutability (~8%) of the short tandem repeats (STRs) and BAT-40 microsatellite (40As) located in the second intron of the 3-beta-hydroxysteroid dehydrogenase gene[28]. The distinction is that our data show general mutation rates for the whole fruit fly after >30 generations, while in the case of the mentioned study[28] the mutation rate is dependent on the cell type. An additional study found that the mutation rate in polyA region, 10As in this case, is in the range of 10−4 per cell per generation[29].

As such, the authors argue that ∼1% of cells will be affected by a polyA region mutation in 100 generations. Similar rates were observed in the other studies with ∼10−6–10−2 mutation rate for the different lengths of homopolymeric regions or STRs (refs 22, 47, 48). PolyA tracks used in our study tend to operate on the shorter side of the length distribution of STRs and as such should have similar if not even lower rates of mutations.

PolyA tracks that are used endogenously in eukaryotic genomes are typically interrupted by other nucleotides at various positions within the A-rich sequence, which further reduces potential hypermutability effects. We have observed that the position of the interrupting nucleotide or codon, in combination with the length of the A-rich sequence, modulates gene expression (Supplementary Fig. 2)16. These observations suggest that polyA-mediated regulation can be further developed for even more precise control of gene expression. Last, ∼2% of human genes are endogenously regulated by polyA tracks, including many well-studied, disease-associated genes, such as BRCA1, TCOF1 and MTDH among others[16,18]. As we showed in our previous study, synonymous mutations of the internal polyA track of such genes can allow investigators to dramatically change expression levels of these genes without manipulation of protein sequence or gene regulatory elements such as promoters and enhancers[16]. The use of our method is not restricted only to these genes, and we feel that the synthetic biology field will benefit from this application. Control of biosynthetic pathways for production of useful secondary metabolites[49], antibiotics[50] or recombinant antibodies[51], as well as introduction of controllable retrosynthetic and fully engineered pathways[52] or ultimate control of metabolic pathways in the modelling of diseases are just a few among the multiple possible applications of this method in the future.

3.5 Methods

3.5.1 *Escherichia coli* experiments

mCherry reporter constructs used for expression in *E.coli* cells were subcloned using LR clonase recombination (Thermo Fisher Scientific) from pENTR/D-Topo constructs used in this study or in previous studies[16,17]. The resulting pBAD-DEST49 vector constructs express Thioredoxin (Thrdx) fusion protein as Thrdx-HA tag-polyA-mCherry. For assaying expression of mCherry reporter all constructs were expressed in 2 ml *E.coli* Top10 strain grown in LB-Carbencilin (LB-Carb; final concentration 100 μgmL^{-1}). The cells were grown to optical density at 600nm (OD_{600}) of 0.4 at 37°C and induced with addition of arabinose (0.5% w v^{-1}). Fluorescence of mCherry reporter for each construct was measured in triplicates 2 to 4 hours after induction using Biotek Synergy H4 plate reader (Excitation 475±9, Emission 620±9). The amount of fluorescence was normalized to number of cells measured by OD_{600}. To additionally check for expression of fusion proteins, 200µl of the cells was harvested 2 hr post-induction, resuspended in 100µl of 2xSDS sample buffer and analyzed by SDS-PAGE followed by western blot analysis using HA-tag specific probe (Santa Cruz Biotechnology Inc.). Images of western blot analyses were generated by Bio-Rad Molecular Imager ChemiDoc XRS System with Image Lab software for chemi-luminescence detection.

The chloramphenicol acetyltransferase (CAT) constructs used for functional protein studies were created by amplification of the CAT gene from pENTR/D-Topo vector (Thermo Fisher Scientific) using primers listed in Table 1. Constructs were subcloned into pBAD-DEST49 vector for use in functional assays. *E. coli* Top10 cells freshly transformed with pBAD-DEST49 plasmids expressing CAT reporters with different polyA tracks as well as CAT control

reporters were grown in liquid LB-Carb media (100 μg mL^{-1}). For the chloramphenicol (CAM) survivability assay *E. coli* cells were grown to OD_{600} = 0.4 and non-induced fractions were spotted on LB-Carb plates (Carb 100 μg mL^{-1}) without chloramphenicol. The residual amount of the cells was induced for 1 hour with arabinose (final concentration 0.1% (w v^{-1})). Fraction of cells, 100μl, was harvested after 30 minutes of induction, resuspended in 50μl of 2xSDS sample buffer and analyzed by SDS-PAGE followed by western blot analysis using HA-tag specific probe (Santa Cruz Biotechnology Inc.). Reminder of cells were washed twice in M9 minimal media, resuspended in the starting volume of LB-Carb media and 5ul of cells was spotted as induced fraction on LB-Carb or to LB-Carb/CAM plates with raising amount of chloramphenicol in the media (CAM 15-100 μg mL^{-1}). Plates were incubated overnight at 37°C and imaged 24 hours post-induction using Bio-Rad Molecular Imager ChemiDoc XRS System.

3.5.2 *Saccharomyces cerevisiae* experiments

In order to conduct functional studies with polyA track-induced hypomorphic attenuation in *S. cerevisiae* cells the *ADE1* locus was deleted from 74-D964 yeast strain via homologous recombination. Resultant *ade1Δ* strains (Table 2) were transformed with an empty vector or a plasmid-based *ADE1* containing variable length of polyA tracks as well as WT and $12Lys_{AAG}$ insertion constructs. Constructs were generated by performing PCR on *ADE1* isolated from the yeast genomic tiling library (Open Biosystems) with primers listed in Table 3. PCR products were digested and ligated into a p416 vector backbone containing the *ADE1* endogenous promoter. Clones were verified via sequencing and correct constructs were transformed into *ade1Δ* deletions strains via the PEG-LiOAc method (Table 2). To generate dilution spottings, three colonies were picked from each transformation plate and grown overnight in selective

media. In the morning, cultures were normalized to OD_{600} = 1.0 and 10µl of cells were spotted onto rich media, SD-Ura, and SD-Ade for phenotypic analysis.

Relative protein abundance was determined via dot blot. Briefly, yeast transformants were picked from selection plates to inoculate 10mL of SD-Ura and grown overnight to ~OD=0.6. In the morning, cells were harvested and lysed in buffer (25 mM Tris-HCl pH 7.5, 50 mM KCl, 1 mM EDTA, Roche protease inhibitor cocktail) via mechanical disruption with acid-washed glass beads (Sigma). Total protein was normalized to 1 mg mL^{-1} via Bradford assay, and 20 µg of total protein was spotted onto a nitrocellulose membrane. Western blotting was performed by overnight incubation with anti-Flag (Sigma M2, 1:1,000 in 5% milk) and goat anti-rabbit (Sigma, 1:10,000 in 5% milk) antibodies, followed by detection with chemiluminescence (Amersham ECL).

3.5.3 *Tetrahymena thermophila* experiments

T. thermophila strain B2086 (II) was used for all experiments reported. Similar results were obtained with strain CU428 [(VII) mpr1-1/mpr1-1]. To assess the effect of polyA-tracks on protein accumulation, we modified a fluorescent protein tagging vector, pBSICY-gtw[53] so as to fuse YFP to the carboxyl-terminus of a Macronucleus-Localized Protein 1 of unknown function (MLP1, TTHERM_00384860), separated by a Gateway recombination cassette (Invitrogen/Life Technologies, Inc), and expressed from the cadmium inducible *MTT1* promoter[54]. The *MLP1* gene coding region was amplified with oligonucleotides 5’ ALM Bsi’ 5’ - CAC CCG TAC GAA TAA AAT GAG CAT TAA TAA AGA AGA AGT-3’ and 3’ ALM RV 5’- GAT ATC TTC AAT TTT AAT TTT TCT TCG AAG TTG C 3’ and cloned into pENTR-D in a topoisomerase mediated reaction prior to digesting with BsiWI and EcoRV and inserting into *BsiWI/PmeI*

digested pBSICY-gtw. Subsequently, LR Clonase II was used to insert a linker containing the sequence coding for an HA epitope tag alone (WT) or the tag plus different length of polyA tracks or AAG insertions in place of the Gateway cassette.

The expression cassette is located within the 5' flanking region of a cycloheximide resistant allele of the rpL29 gene to direct its integration into this genomic locus. These constructs were linearized with *PvuI* and *SacI* in the region flanking the *Tetrahymena* rpl29 sequences and introduced into starved *Tetrahymena* cells by biolistic transformation[55,56]. Transformants were selected in 1x SPP medium containing 12.5 μg mL^{-1} cycloheximide. To control for copy number, PCR assays with primers MTT2386 5'-TCTTAGCTACGTGATTCACG-3'and Chx-117, 5'- ATGTGTTATTAATCGATTGAT-3' and Chx85r, 5'- TCTCTTTCATGCATGCTAGC - 3' verified that all rpL29 loci contained the integrated expression construct.

Transgene expression was induced by addition of 0.6 μg mL^{-1} $CdCl_2$ and cells were grown 12-16 hours before monitoring protein accumulation. YFP accumulation was visualized by epifluorescence microscopy as previously described[57]. Whole cells extracts were generated by boiling concentrated cell pellets (2 x 10^6) in 120μl of 1x Laemmli sample buffer, followed by fractionation on 10% SDS polyacrylamide gels and transfer to nitrocellulose. YFP accumulation was a monitored with mouse anti-GFP antisera (G28R anti-GFP (OAEA00007) antibody, Aviva Systems Biology) and normalized to acetylated Rabbit anti-Histone H3 trimethyl-lysine (Upstate Biotechnologies/Millipore, NY). qPCR analysis was done using 5' – AGGCCTACAAGACCAAGGGT – 3' and 5'- AGAGCGGTTTTGACGTTGGA – 3' primers for *T. thermophila* ribosomal protein L21 (rpl21) which was used for normalization. Primers 5' – CCCGTATGACGTACCGGATTATG – 3' and 5' – ACTTCAGGGTCAGCTTGCC – 3' were

used for detection and estimation of fusion protein transcript levels using SybrGreen master mix and CFX96 Touch™ Real time PCR Detection System (BioRad). Normalized ΔCt values were used to calculate fold ratio between WT, $12Lys_{AAG}$ and polyA track constructs.

3.5.4 *Nicotiana benthamiana* experiments

Constructs for expression of HA-tagged mCherry reporters that were already cloned in pEntryD-TOPO vector were sub-cloned to pEarleyGate 100 (ABRC stock number CD3-724) through LR reaction using LR clonase (Invitrogen™). The mCherry reporter constructs, pEARLY100 and pBIN61 plasmids were individually electroporated into *Agrobacterium tumefaciens* strains GV3101[58]. The strain carrying pBIN61 construct expressing p19 protein from tomato bushy stunt virus was co-infiltrated with the reporter constructs to suppress post-transcriptional gene silencing[59]. The *Agrobacterium* suspensions carrying the reporter constructs were infiltrated into the leaves of 5- to 6-week-old *N. benthamiana* plants as described in Joensuu *et al*., 2010[60]. Briefly, saturated over-night cultures were spun-down and resuspended in the infiltration solution (3.2 g L^{-1} Gamborg's B5 plus vitamins, 20 g L^{-1} sucrose, 10 mM MES pH 5.6, 200 μM 4′-Hydroxy-3′,5′-dimethoxyacetophenone) to a final OD_{600} = 1.0; *Agrobacterium* suspensions carrying the reporter constructs were individually mixed with suspensions carrying the pBIN61 construct in 1:1 ratio prior to infiltrations. These suspensions were infiltrated into separate segments of two young leaves on each of eight different *N. benthamiana* plants, which served as biological replicates. For control, 1:1 suspension of *A. tumefaciens* carrying pEARLY 100 with no insert along with pBIN61 was used. The infiltrated plants were maintained in a controlled growth chamber conditions at 22°C, with a 16h photoperiod**.**

Samples of the abaxial epidermis of *N. benthamiana* leaves infiltrated with different mCherry reporter constructs were collected 6 days post-infiltration. Infiltration was performed as described in the previous section, with the addition of an YFP-expressing construct pEARLY104 (ABRC stock number CD3-686), which served as infiltration control. The samples were visualized for fluorescence by confocal laser-scanning microscopy using a Leica TCS SP2 confocal microscope. Samples for RNA and total soluble protein (TSP) extraction were separately collected from the infiltrated plants 6 days post-infiltration using a cork borer (7.1 mm in diameter); each sample contained equal amounts of leaf tissue (2 leaf discs) collected from each of the segments on the two leaves infiltrated with the same construct.

Analysis of mCherry protein accumulation was carried out by western blot as described in Gutiérrez *et al.*, 2013 and Conley *et al.*, 2009[61,62]. Briefly, phosphate-buffered saline (PBS: 8 g L^{-1} NaCl, 1.16 g L^{-1} Na_2HPO_4, 0.2 g L^{-1} KH_2PO_4, 0.2 g L^{-1} KCl, pH 7.4), supplemented with 1 mM EDTA, 1 mM phenylmethanesulfonylfluoride (PMSF), 1 µg mL^{-1} leupeptin 0.1% Tween-20 and 100 mM sodium L-ascorbate was used for total soluble protein (TSP) extractions. Bradford assay (Biorad) was used to quantify TSP in the extracts using a standard curve (r^2=0.99) of known concentrations of Bovine Serum Albumin (BSA). Sample extracts (25 µg TSP for mCherry and 5 µg TSP for phosphinotricin acetyl transferase [BAR] protein detection) were separated by SDS-PAGE, blotted onto nitrocellulose membrane and probed with a primary anti-HA tag antibody (Genscript) for mCherry, or anti-Phosphinotricin acetyl transferase antibody (Abcam) for BAR, both at 1:2,000 dilution, followed by HRP-conjugated secondary antibody (Biorad) at 1:5,000 dilution. The blots were washed (3 times X 10 min) in 1XTris-buffered Saline (TBS**,** 50 mM Tris, 150 mM NaCl, pH 7.5) containing 0.1% Tween (Sigma) and images were obtained after 1 min incubation with the enhanced chemiluminescence (ECL) detection

system (GE Healthcare)**.** Numerical values for protein accumulation were derived from the detected band intensities on the analyzed images using TotalLab TL 100 software (Nonlinear Dynamics, Durham, USA). The mCherry accumulation values were normalized for BAR accumulation detected in the same sample. Normalized values of the mCherry protein accumulation for each reporter construct were presented as the mean of eight biological replicates ± SE; Tukey's honest significance test (JMP software, SAS Institute Inc.) was used to identify significantly different means (α=0.05).

For quantitative RT–PCR (qPCR), total RNA was extracted using an RNeasy plant mini kit coupled with DNase treatment (Qiagen). The purified RNA (500 ng) was reverse-transcribed using the Maxima first-strand cDNA synthesis kit (Thermo Fisher Scientific). The resulting cDNA (2 ng μl^{-1}) was quantified by qPCR using the Maxima SYBR Green/ROX qPCR master mix (Thermo Fisher Scientific) and CFX384 Touch™ Real-Time PCR Detection System (Biorad**).** Cycle threshold (Ct) values were normalized to phosphinothricin N-acetyltransferase (*BAR*) gene expressed from the same plasmid used for transient expression. Primer sequences used: For mCherry - mCherryFWD: 5'-GGCTACCCATACGATGTTCC-3'; mCherryREV: 5'-CCTCCATGTGCACCTTGAAG-3'; for BASTA - BAR-F3: 5'-TCAAGAGCGTGGTCGCTG-3' and BAR-R3: 5'-CAAATCTCGGTGACGGGCAG-3'.

3.5.5 *Drosophila melanogaster* experiments

Reporter gene expression was achieved with the GAL4/UAS system. The UAS-mCherry transgene plasmids were constructed from the phiC31 integrase plasmid, pJFRC28-10XUAS-IVS-GFP-p10 (Addgene plasmid #36431)[63]. GFP was removed by digestion with KpnI and XbaI and replaced with HA-mCherry and HA-polyA-mCherry. Transgenic fly lines were obtained by

injecting P{CaryP}attP2 embryos with each pJFRC28 mCherry construct to achieve site-specific, single insertion on the third chromosome at the attP2 landing site (Rainbow Transgenic Flies, Inc.). Injected G_0 adult flies were backcrossed to w^{1118} flies. Red-eyed progeny indicated successful germline integration of the UAS-mCherry expression cassette. Male red-eye progeny were crossed to female w;TM3 Sb/TM6 Tb flies followed by sib-crosses of the F_1 progeny to generate homozygous UAS-mCherry transgenic lines. Insertion was confirmed by Sanger sequencing of PCR amplified mCherry from genomic DNA of individual flies. Each mCherry transgenic fly line was crossed to a TubGal4 UAS-GFP driver line (derived from BSC42734) to achieve mCherry expression in all tissues. GFP expression was used for normalization. All flies were maintained at 25°C.

Third instar larvae from each cross were fixed in formaldehyde and dissected to recover the salivary glands (SG), intact central nervous system (CNS), and proventriculus (PV). The tissues were mounted on glass cover slips and confocal images were taken on a Zeiss Imager 2 upright microscope using identical parameters for all images of each tissue type. Fluorescence intensity of the mCherry and GFP were quantified with Zen 9 software.

Total RNA was extracted from each cross by pooling 5 third instar larvae in 1.5 ml RNase-free Eppendorf tubes which were then frozen in dry ice. Frozen samples were homogenized using 1.5 ml pestles (Fisherbrand, RNase- and DNase-free). After homogenization, 1 ml RiboZol reagent (Amresco) was added and extraction was completed according to manufacturer's instructions. Total RNA samples were treated with TURBO DNA-free kit (Ambion) to remove potential genomic DNA. cDNA synthesis was performed with iScript Reverse Transcription Supermix (Bio-Rad) with 1 µg of total RNA in a 20 µl reaction. RT-qPCR

was performed in the Bio-Rad CFX96 Real-Time System with iQ SYBR Green Supermix (Bio-Rad). The mCherry transcript was detected with the following primers: 5'-TGACGTACCGGATTATGCAA-3' and 5'-ATATGAACTGAGGGGACAGG-3'. Cycle threshold (Ct) values were normalized to EF1 with the following primers: 5'-GCGTGGGTTTGTGATCAGTT-3' and 5'-GATCTTCTCCTTGCCCATCC-3')[64].

For western blot analysis, five third instar larvae from each cross were collected and frozen in dry ice. Frozen samples were homogenized using 1.5 ml pestles (Fisherbrand, RNase- and DNase-free). After homogenization, SDS sample buffer was added and the samples were boiled for 10 minutes. Anti-HA (Santa Cruz Biotechnology Inc.) was used to detect mCherry expression. Samples were normalized with anti-GFP (Clontech).

For Illumina sequencing analysis, genomic DNA was isolated from a single adult HA-(AAA)12-mCherry fly approximately 30 generations after transgene insertion. The QIAGEN DNeasy Blood and Tissue Kit were used, following the supplementary protocol for total DNA extraction from insect cells. The polyA track of the mCherry transgene was amplified from both the genomic DNA and the plasmid DNA that was used to generate the transgenic fly line. The primers (5'- GGCTGCTCGAGCCGTATGACGTACCGGATTATGC-3' and 5'-CCTTGTGATCAGGCCATGTTATCCTCCTCGC-3') added XhoI and SpeI restriction enzyme sites flanking the polyA track region. PCR products were purified with NucleoSpin Gel and PCR Clean-up (Macherey-Nagel) and digested with XhoI and SpeI to generate overhangs for ligation of Illuminia adapter sequences. To enrich for amplicons with both adapter sequences, the product was amplified with the following primers: 5'- -3' and 5'- -3'. Samples were run as a spike-in on an Illumina HISeq machine. Reads that matched the first 32 or 33 expected nucleotides,

depending on the indexing barcode used, were counted. These resulted in approximately 670,000 genomic reads and 1,000,000 plasmid reads after removing sequences that had fewer than 50 reads.

For Sanger sequencing analysis, the polyA track from the genomic DNA extracted for Illumina sequencing and the original plasmid were amplified with the following primers: 5'-GGCTGCTCGAGCCGTATGACGTACCGG-3' and 5'-ACATGAACTGAGGGGACAGG-3'. The PCR products were purified with NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) and ligated into the pCR-Blunt vector using the Zero Blunt PCR Cloning Kit (Thermofisher). The vector was transformed into Max Efficiency DH5α competent cells (Thermofisher) and 60 colonies from both the genomic DNA and plasmid DNA were sent for sequencing (Genewiz). Sequences with good quality scores were used for downstream analysis.

3.5.6 *Homo sapiens* cell culture experiments

mCherry reporter constructs used for transient expression in human cells were subcloned using LR clonase recombination (Thermo Fisher Scientific) from pEntryD-Topo constructs used in other experiments or in previous studies[16]. DNA fragments for constructs used for creation of inducible and stable cell lines were PCR amplified, purified and ligated into pcDNA 5/FRT/TO vector (Thermo Fisher Scientific).

HeLa cells were cultured in Dulbecco's modified Eagle's medium (DMEM) (Gibco) and supplemented with 10% fetal bovine serum, 5% minimum essential medium nonessential amino acids (100×, Gibco), 5% penicillin and streptomycin (Gibco), and L-glutamine (Gibco). Flp-In™ T-REx™ 293 cells were grown in the same media with addition of 5 $\mu g mL^{-1}$ of blastocidin and

100 μg mL^{-1} of Zeocin for non-recombined cells, or 5 μg mL^{-1} of blastocidin and 100 μg mL^{-1} of hygromycin for growth of stable cell lines expressing mCherry or HBD constructs.

Plasmids were introduced to the tissue culture cells by the Neon Transfection System (Thermo Fisher Scientific) using 100-μl tips according to cell-specific protocols (www.lifetechnologies.com/us/en/home/life-science/cell-culture/transfection/transfection---selection-misc/neon-transfection-system/neon-protocols-cell-line-data.html). Hela cells, used for transient expression, were electroporated with 1.5 μg of DNA plasmids and were harvested 24 hours after the electroporation. Flp-In™ T-REx™ 293 cells were electroporated with plasmids, selected for positive clones as described by protocol (https://tools.thermofisher.com/content/sfs/manuals/flpinsystem_man.pdf). Expression of polyA track and control constructs was induced by addition of various amounts of doxycycline from a common stock (1mg mL^{-1}) and harvested 24 or 48 hours after induction, if not indicated differently.

Total RNA was extracted from cells using the RiboZol RNA extraction reagent (Amresco) according to the manufacturer's instructions or using GenEluteTM Direct. RiboZol reagent (500 μl) was used in each well of 6- or 12-well plates for RNA extraction. Precipitated nucleic acids were treated with Turbo deoxyribonuclease (Ambion), and total RNA was dissolved in ribonuclease-free water and stored at −20°C. RNA concentration was measured by NanoDrop (OD260/280). iScript Reverse Transcription Supermix (Bio-Rad) was used with 1 μg of total RNA following the manufacturer's protocol. RT-qPCR was performed in the Bio-Rad CFX96 Real-Time System with iQ SYBR Green Supermix (Bio-Rad). For both transient expression samples and stable cell line samples, the mCherry transcript was detected with the following primers: 5'-TGACGTACCGGATTATGCAA-3' and 5'-

ATATGAACTGAGGGGACAGG-3'. Cycle threshold (Ct) values were normalized to the neomycin resistance gene expressed from the same plasmid for transient expression (5'-CTGAATGAACTGCAGGACGA-3' and 5'-ATACTTTCTCGGCAGGAGCA-3') or hygromycin (5'-GATGTAGGAGGGCGTGGATA-3' and 5'-ATAGGTCAGGCTCTCGCTGA-3' or actin gene for stable cell lines (5'-AGAAAATCTGGCACCACACC-3' and 5'-AGAGGCGTACAGGGATAGCA-3'.

Total cell lysates were prepared with passive lysis buffer (Promega). Blots were blocked with 5% milk in 1× tris-buffered saline with 0.1% Tween 20 (TBST) for 1 hour. Horseradish peroxidase–conjugated or primary antibodies (anti-β-actin (Santa Cruz Biotechnology Inc.), anti δ-tubulin (Sigma-Aldrich)) were diluted according to the manufacturer's recommendations and incubated overnight with the membranes. The membranes were washed four times for 5 min in TBST and prepared for imaging, or secondary antibody was added for an additional 1 hour incubation. Images were generated by Bio-Rad Molecular Imager ChemiDoc XRS System with Image Lab software by chemiluminescence detection or by the LI-COR Odyssey Infrared Imaging System. Blots imaged by the LI-COR system were first incubated for 1 hour with Pierce DyLight secondary antibodies.

3.5.7 Data availability

The data that support the findings of this study are available from the authors on reasonable request; see author contributions for specific data sets.

3.6 Figures

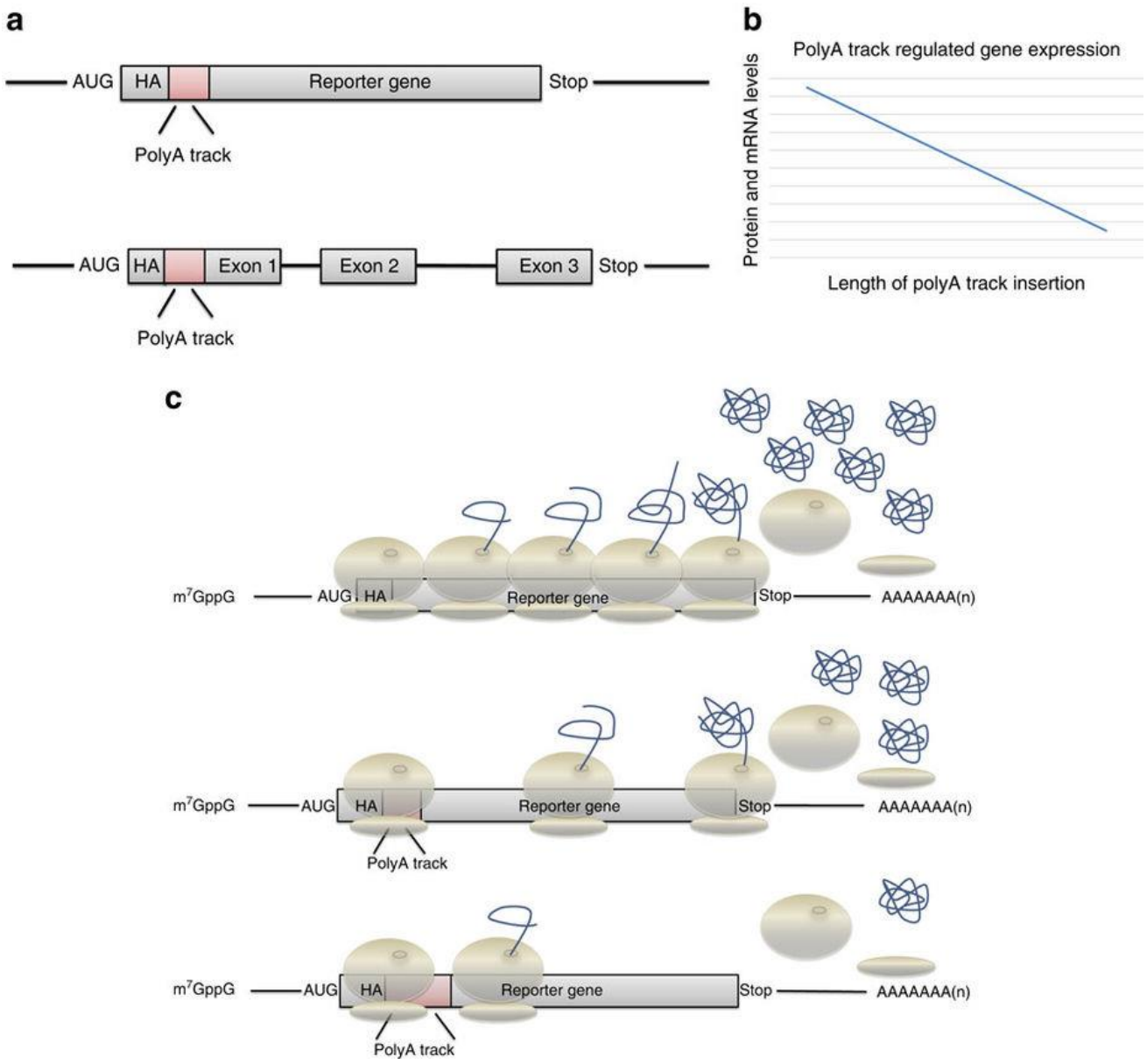

a
AUG
HA
Reporter gene
Stop
PolyA track
AUG
HA
Exon 1
Exon 2
Exon 3
Stop
PolyA track
b
PolyA track regulated gene expression
Protein and mRNA levels
Length of polyA track insertion
c
m7GppG
AUG
HA
Reporter gene
Stop
AAAAAAA(n)
m7GppG
AUG
HA
Reporter gene
Stop
AAAAAAA(n)
PolyA track
m7GppG
AUG
HA
Reporter gene
Stop
AAAAAAA(n)
PolyA track

Figure 3.1. Design and mechanism of polyA track tag regulated gene expression.

(a) Scheme of inserted polyA tracks in the reporter genes used in this study. Hemagglutinin (HA) tag (gray) and polyA tracks (red) were introduced in the coding region of the reporter genes next to the start AUG codon. Exon boundaries as well as termination codon (Stop) are indicated. (b) Proposed correlation between gene products levels, mRNA and protein, and the length of inserted polyA track tags. The reduction in levels of both reporter protein and mRNA is dependent on increasing length of consecutive adenosine nucleotides in the coding sequence. (c) Scheme of translation of eukaryotic reporter mRNA with or without inserted polyA tracks. The length of inserted polyA track tag determines the protein output of the regulated reporter gene as indicated by the number of globular protein structures. Features of the eukaryotic mRNAs (m7GpppG - cap, AUG - start codon, Stop - termination codon and poly(A) tail), as well as HA-tag, position of the polyA track tag, ribosome and nascent polypeptide chain are illustrated in the scheme.

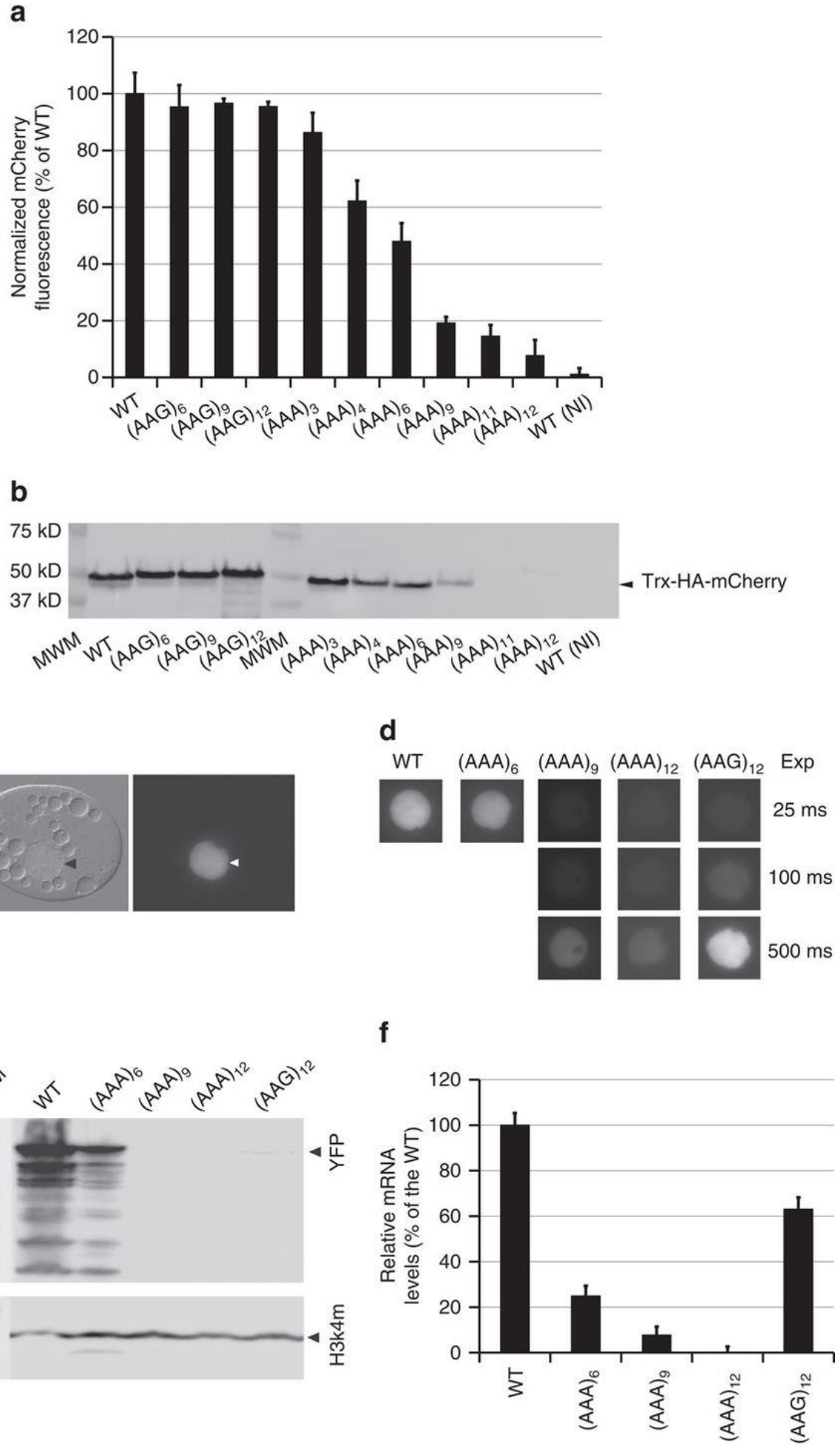

a
Normalized mCherry fluorescence (% of WT)
120
100
80
60
40
20
0
WT
(AAG)6
(AAG)9
(AAG)12
(AAA)3
(AAA)4
(AAA)6
(AAA)9
(AAA)11
(AAA)12
WT (NI)
b
75 kD
50 kD
37 kD
Trx-HA-mCherry
MWM
WT
(AAG)6
(AAG)9
(AAG)12
MWM
(AAA)3
(AAA)4
(AAA)6
(AAA)9
(AAA)11
(AAA)12
WT (NI)
c
d
WT
(AAA)6
(AAA)9
(AAA)12
(AAG)12
Exp
25 ms
100 ms
500 ms
e
kDa
MWM
WT
(AAA)6
(AAA)9
(AAA)12
(AAG)12
130
100
70
55
40
25
15
YFP
H3k4m
f
Relative mRNA levels (% of the WT)
120
100
80
60
40
20
0
WT
(AAA)6
(AAA)9
(AAA)12
(AAG)12

Figure 3.2. Regulation of reporter gene by polyA tracks in the single cell prokaryotic and eukaryotic organisms.

(a) Percentage of mCherry fluorescence of tested Lys_{AAG} $((AAG)_{6-12})$ and $Lys_{AAA}((AAA)_{3-12})$ insertion constructs compared to wild type fluorescence (WT, no insertion construct). mCherry fluorescence was assayed at excitation wavelength of 475±9nm and emission was detected at 620±9nm. Error bars indicate mean mCherry fluorescence values ± standard deviation for three individual *E. coli* colonies for each construct. Background levels of mCherry expression can be estimated from the fluorescence of the non-induced wild type construct (WT(NI)). (b) Western blot analysis of mCherry constructs expressed in *E. coli* cells. Equal amounts of *E. coli* cell lysates with Thioredoxin(Trx) fusion proteins were used for analysis. Fusion proteins were detected using HA-tag specific antibody. Positions of the fusion protein (Trx-HA-mCherry) and sizes of molecular weight markers (MWM) are indicated. (c) Representative differential interference contrast microscopy (left panel) and the corresponding fluorescence image (right panel -25 msec exposure) of a *T. thermophila* cell expressing the wild type (WT) MLP1-HA-YFP fusion. Arrowheads denote the position of the macronucleus. (d) *MPL1*-HA-YFP accumulation within macronuclei of live *T. thermophila* cells expressing an allelic series of fusion proteins -- WT, $(AAA)_{6-12}$, and $(AAG)_{12}$ -- was visualized by epifluorescence microscopy. Different exposures times are indicated on the right to demonstrate the relative accumulation of each variant. (e) Western blot analysis was performed with whole cell lysates made from *T. thermophila* cells expressing the MLP-HA-YFP fusion proteins. Protein from equivalent cell numbers was loaded in each lane and detected using YFP specific antibody (top panel) and normalized to the nuclear histone species, histone H3 trimethyl-lysine 4 (H3K4m) (bottom panel). Positions of the full-length fusion protein (YFP), normalization control (H3k4m), and sizes of molecular weight markers (MWM) are indicated. Degradation of excess fusion protein is readily apparent as faster

migrating species below the full-length MLP1-HA-YFP. (f) Steady state levels of fusion gene constructs measured by qRT-PCR. Relative levels of the mRNA for $(AAG)_{12}$ and $(AAA)_{6\text{-}12}$ are presented as percentage of the wild type (WT) construct mRNA levels. Error bars represent mean± standard deviation values (n=3).

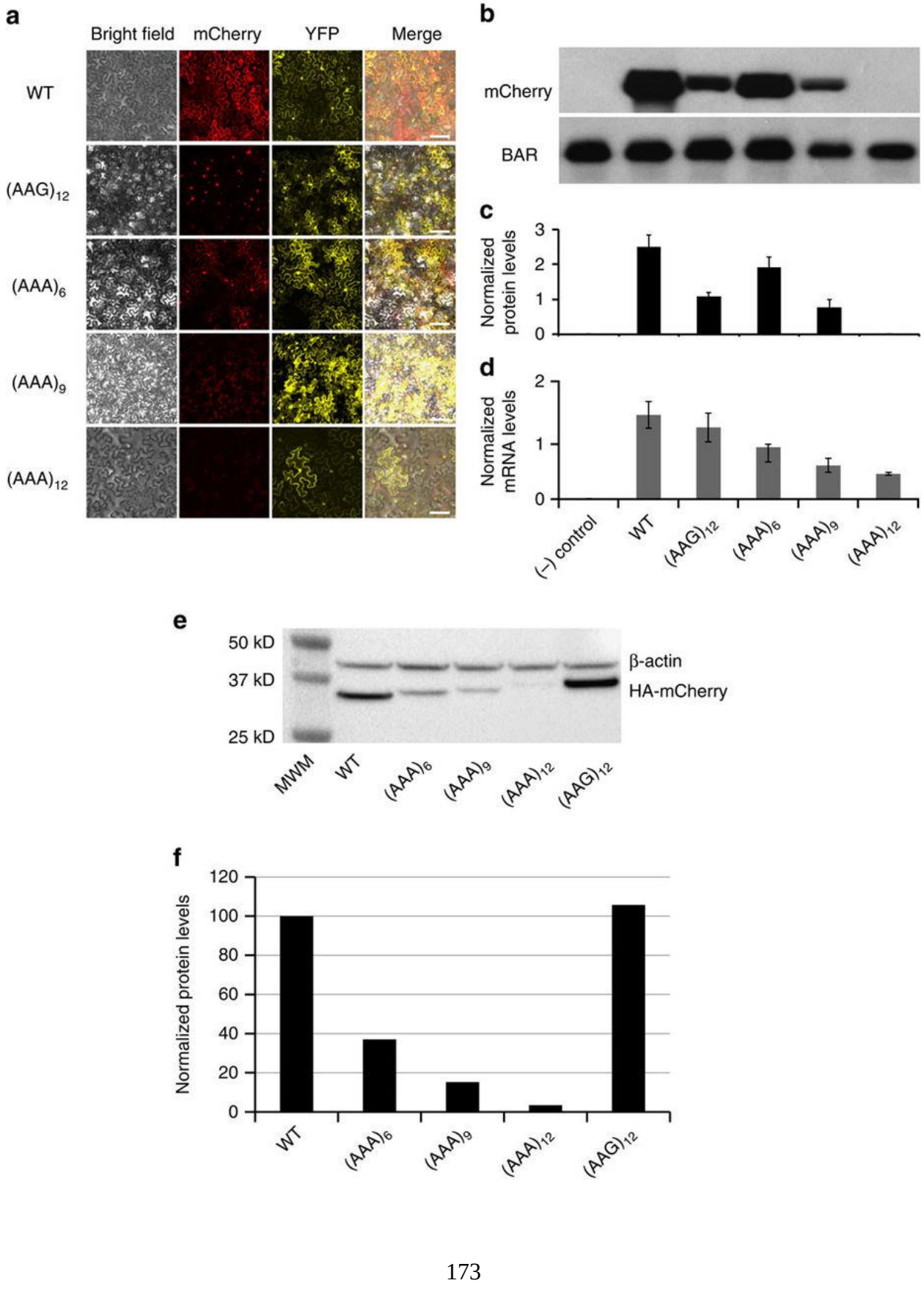

a
Bright field
mCherry
YFP
Merge
WT
(AAG)12
(AAA)6
(AAA)9
(AAA)12
b
mCherry
BAR
c
Normalized protein levels
3
2
1
0
d
Normalized mRNA levels
2
1
0
(–) control
WT
(AAG)12
(AAA)6
(AAA)9
(AAA)12
e
50 kD
37 kD
25 kD
β-actin
HA-mCherry
MWM
WT
(AAA)6
(AAA)9
(AAA)12
(AAG)12
f
Normalized protein levels
120
100
80
60
40
20
0
WT
(AAA)6
(AAA)9
(AAA)12
(AAG)12

Figure 3.3. Regulation of reporter gene by polyA tracks in the eukaryotic tissue cultures.

(a) Fluorescence images of *N. benthamiana* epidermal cells transiently expressing wild type (WT), $(AAG)_{12}$ and $(AAA)_{6\text{-}12}$ mCherry constructs. YFP expression was used as a transfection control. (b) - Western blot analysis, (c) – protein level estimate and (d) – mRNA levels for transfected (-) insert control and WT, $(AAG)_{12}$ and $(AAA)_{6\text{-}12}$ mCherry constructs expressed transiently in *N. benthamiana* epidermal cells. (b) Primary HA-tag antibody was used for detection of HA-mCherry constructs. Phosphinotricin acetyl transferase (BAR) specific antibody was used as a loading and normalization control (c). Levels of mCherry protein from different constructs were derived from detected band intensities normalized for BAR accumulation detected in the same sample. Error bars represent mean values ± standard error from biological replicates (n=8). (d) mRNA levels for different mCherry constructs were calculated as cycle threshold (Ct) values and normalized to *BAR* gene mRNA values. Error bars represent mean values ± standard error from biological replicates (n=3). (e) Western blot analysis of transient mCherry constructs expression in HeLa cells. WT, 12 Lys_{AAG} ($(AAG)_{12}$) and 6-12 Lys_{AAA} ($(AAA)_{6\text{-}12}$) mCherry proteins were detected using HA-tag specific primary antibody. β-actin was used as a loading control and was detected using specific antibody. Positions of the fusion protein (HA-mCherry), normalization control (β-actin) and sizes of molecular weight markers (MWM) are indicated. (f) Quantification of the mCherry protein levels from detected western blot intensities. Levels of mCherry were normalized to β-actin band intensities and represented as a percentage of the wild type construct values.

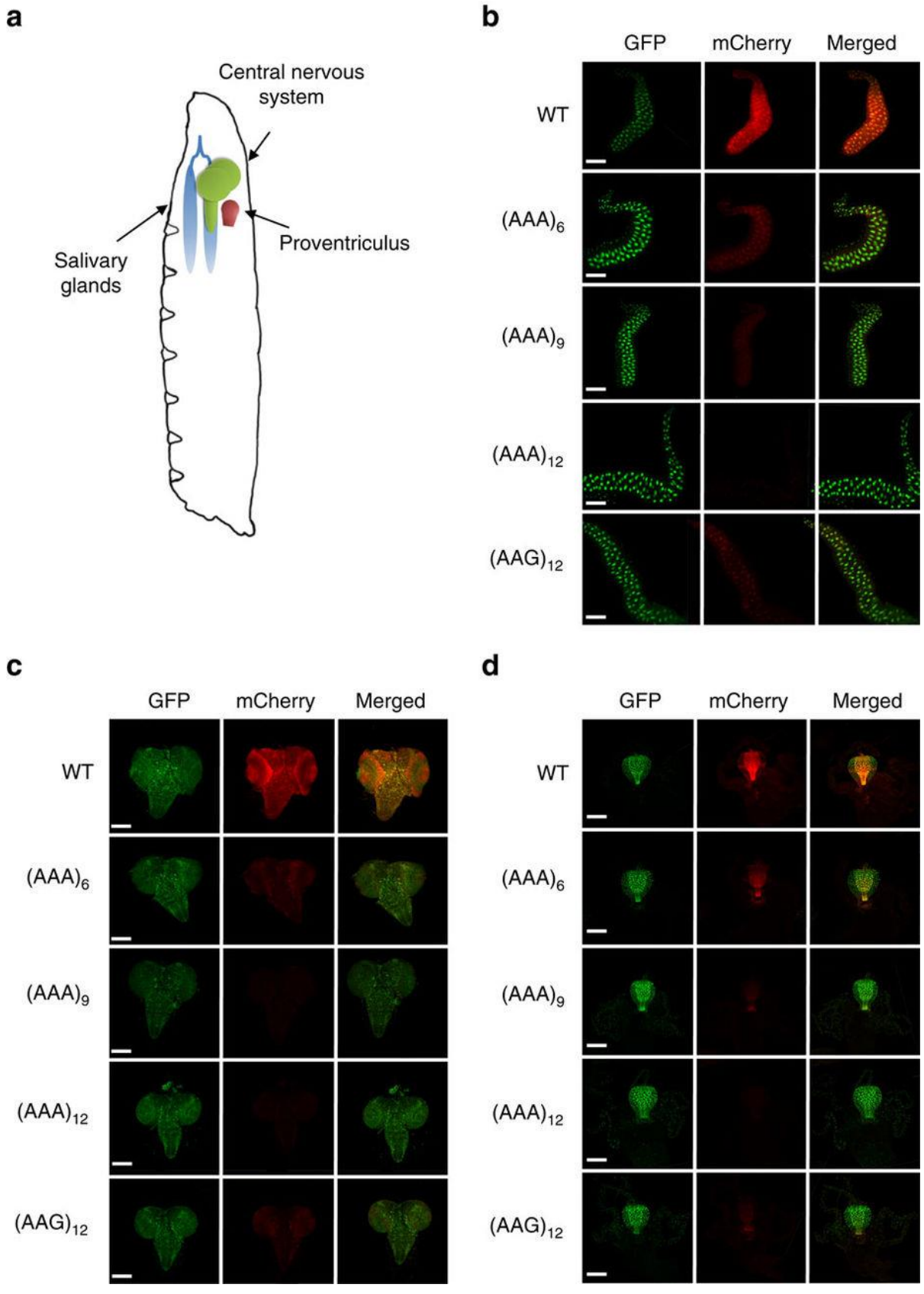
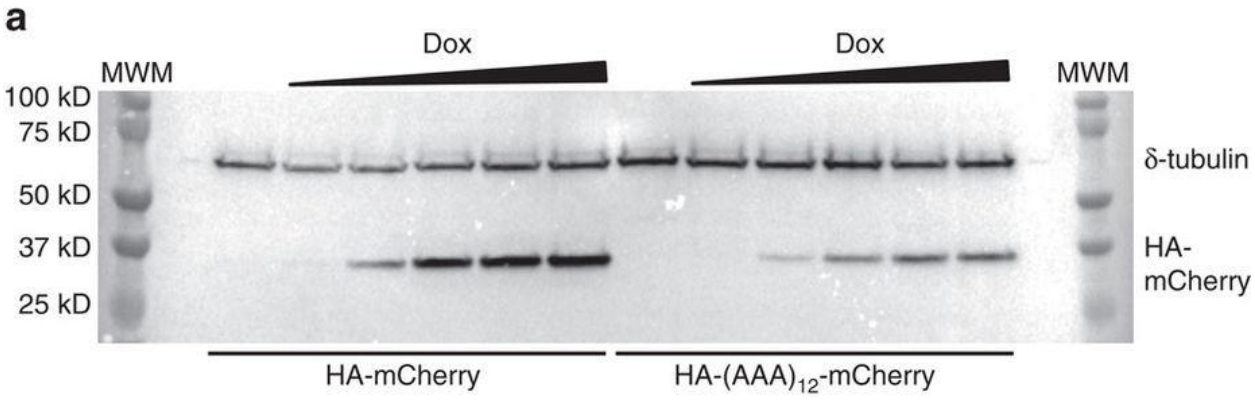
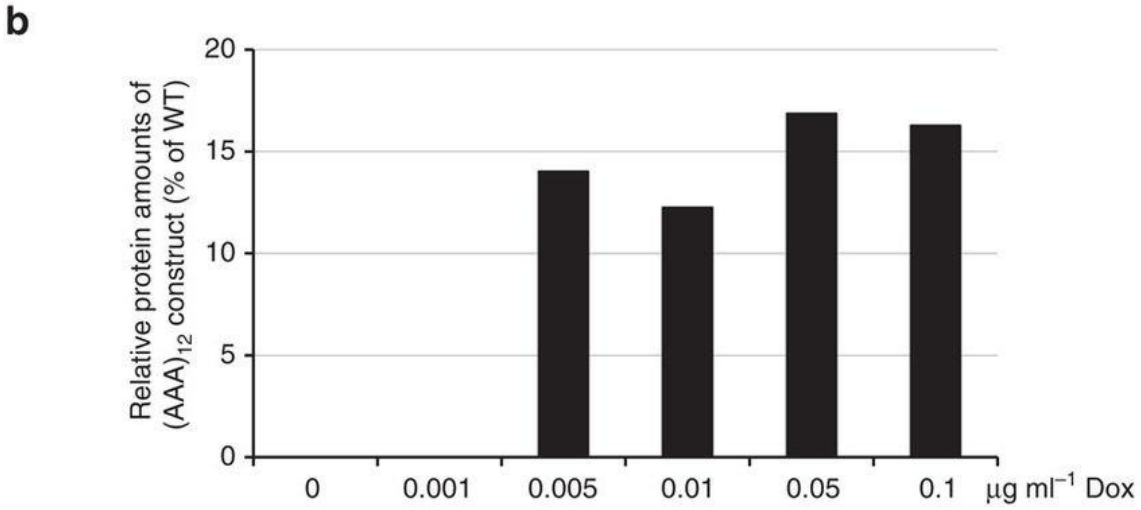

a
Central nervous system
Proventriculus
Salivary glands
b
GFP
mCherry
Merged
WT
(AAA)6
(AAA)9
(AAA)12
(AAG)12
c
GFP
mCherry
Merged
WT
(AAA)6
(AAA)9
(AAA)12
(AAG)12
d
GFP
mCherry
Merged
WT
(AAA)6
(AAA)9
(AAA)12
(AAG)12

Figure 3.4. PolyA tracks regulate mCherry reporter gene expression in different organs of *D. melanogaster*.

(a) Diagram of third instar fruit fly larva showing approximate location of salivary glands (SG, blue), central nervous system (CNS, green) and proventriculus (PV, red). Fluorescence imaging of formaldehyde fixed SG (b), CNS (c) and PV (d), dissected from larvae expressing wild type (WT), $(AAG)_{12}$ and $(AAA)_{6\text{-}12}$ mCherry constructs. mCherry and GFP indicate images acquired by selective fluorescence filter setting. Overlay of mCherry and GFP fluorescence is shown in the merged panel.


a
MWM
Dox
Dox
MWM
100 kD
75 kD
50 kD
37 kD
25 kD
δ-tubulin
HA-
mCherry
HA-mCherry
HA-(AAA)12-mCherry


b
Relative protein amounts of (AAA)12 construct (% of WT)
20
15
10
5
0
0
0.001
0.005
0.01
0.05
0.1
μg ml−1 Dox

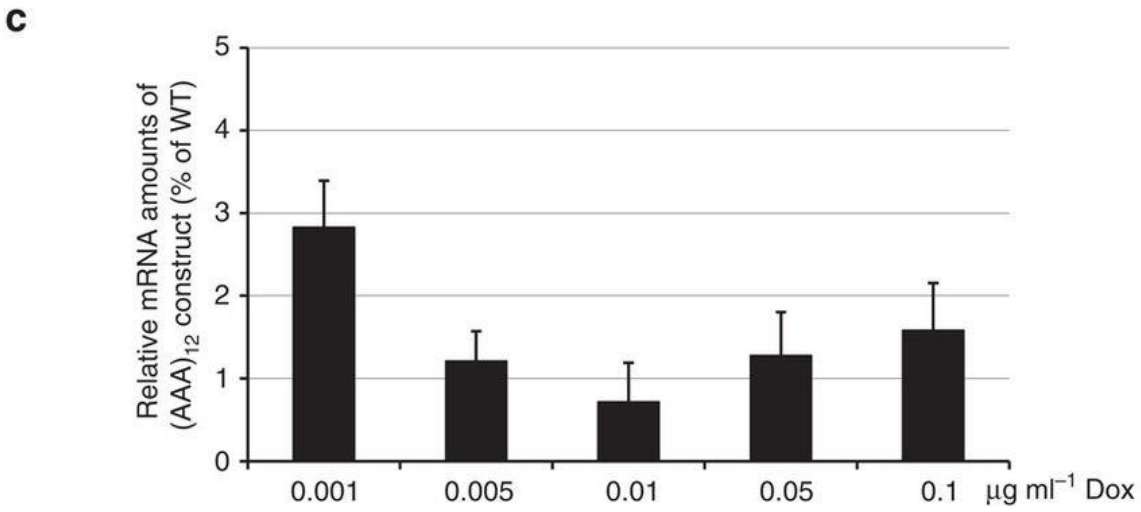

c
Relative mRNA amounts of (AAA)12 construct (% of WT)
5
4
3
2
1
0
0.001
0.005
0.01
0.05
0.1
μg ml−1 Dox

Figure 3.5. PolyA tracks regulate mCherry reporter expression independently of the promoter strength.

(a) Western blot analysis of the cell lysates from Flp-In^{TM} T-REx^{TM} 293 stable cell lines expressing doxycycline (Dox) inducible wild type (HA-mCherry) and 12 Lys_{AAA} insertion construct (HA-$(AAA)_{12}$-mCherry) from a single locus. Dox concentration in the media was varied from 0 to 0.1 μg mL^{-1}. Constitutively expressed δ-tubulin was used as a loading control and was detected using specific antibody. Positions of the fusion protein (HA-mCherry), normalization control (δ-tubulin) and sizes of molecular weight markers (MWM) are indicated. (b) Quantification of the mCherry protein levels from detected western blot intensities. Levels of mCherry were normalized to δ-tubulin band intensities and represented as a percentage of the wild type construct values at each Dox concentration. Numbers indicate concentration of Dox in the media. (c) Steady state mRNA levels of the 12 Lys_{AAA} insertion construct ($(AAA)_{12}$) measured by qRT-PCR. Relative levels of the mRNA for $(AAA)_{12}$ are presented as percentage of the wild type (WT) construct mRNA level at each Dox concentration. Error bars represent mean± standard deviation values (n=3). Numbers indicate final concentration of Dox in the media.

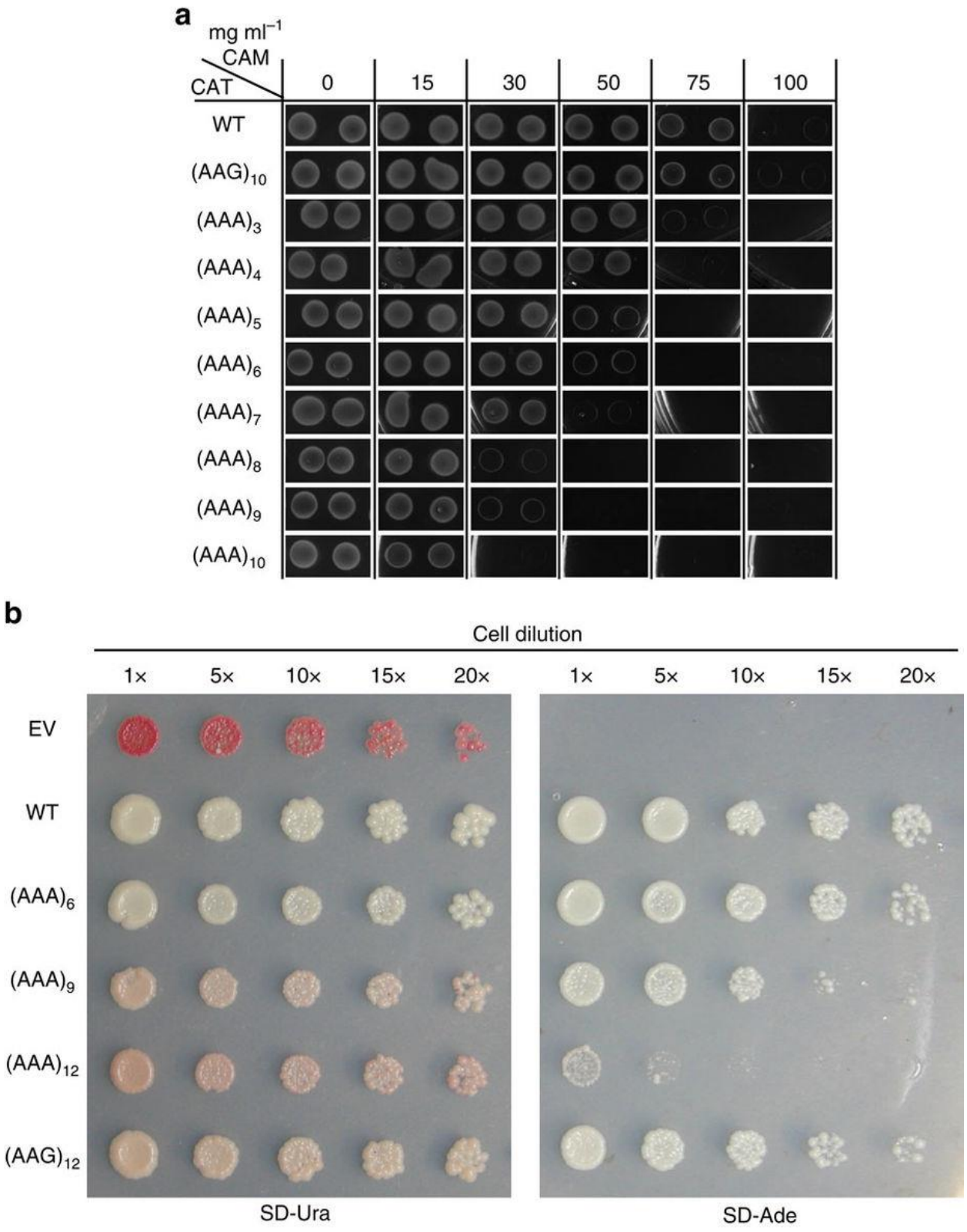

a
mg ml−1
CAM
CAT
0
15
30
50
75
100
WT
(AAG)10
(AAA)3
(AAA)4
(AAA)5
(AAA)6
(AAA)7
(AAA)8
(AAA)9
(AAA)10
b
Cell dilution
1×
5×
10×
15×
20×
1×
5×
10×
15×
20×
EV
WT
(AAA)6
(AAA)9
(AAA)12
(AAG)12
SD-Ura
SD-Ade

Figure 3.6. Regulation of drug resistance and metabolic survival by insertion of polyA track tags in genes from *E. coli* and *S. cerviseae*.

(a) Survival of *E. coli* cells expressing wildtype (WT), 10Lys_{AAG} ($(AAG)_{10}$) and 3-10 Lys_{AAA} $(AAA)_{3-10}$ chloramphenicol acetyltransferase (CAT) constructs on chloramphenicol (CAM) selective media. Pulse induced *E. coli* cells, expressing different CAT constructs, were plated on selective antibiotic plates with varying amounts of CAM in the media (0-100 mg mL^{-1}). Two independent clones were assessed for each construct. *E. coli* colonies were imaged 16 hours after plating. (b) Assays for *ADE1* gene regulation by polyA tracks ($(AAA)_{6-12}$). Ability of *S. cerevisiae ade1*Δ cells to produce sufficient levels of functional Ade1 protein were assayed by reintroduction of single copy vector with wild type (WT), 12 Lys_{AAG} ($(AAG)_{12}$) and 6-12 Lys_{AAA} ($(AAA)_{6-12}$) Ade1 construct. Empty vector (EV) served as a negative control. Yeast colonies show differential red coloration, on the selective SD-Ura media, which is proportional to the activity of Ade1 protein. Adenine dropout media (SD-Ade) selects for yeast cells expressing sufficient amounts of functional Ade1 protein. Dilutions of the yeast cultures showing relative survival and growth are indicated.

3.7 References

1. Muller, H. J. Further Studies on the Nature and Causes of Gene Mutations. *Proc. 6th Int. Congr. Genet.* **1,** 213–255 (1932).
2. Bonde, M. T. *et al.* Predictable tuning of protein expression in bacteria. *Nat. Methods* **13,** (2016).
3. Breslow, D. K. *et al.* A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat. Methods* **5,** 711–718 (2008).
4. Chappell, J., Watters, K. E., Takahashi, M. K. & Lucks, J. B. A renaissance in RNA synthetic biology: New mechanisms, applications and tools for the future. *Curr. Opin. Chem. Biol.* **28,** 47–56 (2015).
5. Dawlaty, M. M. & van Deursen, J. M. Gene targeting methods for studying nuclear transport factors in mice. *Methods* **39,** 370–378 (2006).
6. Ferri, A. L. *et al.* Sox2 deficiency causes neurodegeneration and impaired neurogenesis in the adult mouse brain. *Development* **131,** 3805–3819 (2004).
7. Garí, E., Piedrafita, L., Aldea, M. & Herrero, E. A set of vectors with a tetracycline-regulatable promoter system for modulated gene expression in Saccharomyces cervisiae. *Yeast* **13,** 837–848 (1997).
8. Goto, T., Hara, H., Nakauchi, H., Hochi, S. & Hirabayashi, M. Hypomorphic phenotype of Foxn1 gene-modified rats by CRISPR/Cas9 system. *Transgenic Res.* **25,** 533–544 (2016).
9. LaFave, M. C. & Sekelsky, J. Transcription initiation from within P elements generates hypomorphic mutations in Drosophila melanogaster. *Genetics* **188,** 749–752 (2011).
10. Meyers, E. N., Lewandoski, M. & Martin, G. R. An Fgf8 mutant allelic series generated by Cre- and Flp-mediated recombination. *Nat. Genet.* **18,** 136–41 (1998).
11. Redden, H., Morse, N. & Alper, H. S. The synthetic biology toolbox for tuning gene expression in yeast. *FEMS Yeast Res.* **15,** 1–10 (2015).
12. Doudna, J. A. & Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. *Science (80-.).* **346,** 1258096–1258096 (2014).
13. Joung, J. K. & Sander, J. D. TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol* **14,** 49–55 (2013).
14. Chuang, H.-Y., Hofree, M. & Ideker, T. A decade of systems biology. *Annu. Rev. Cell Dev. Biol.* **26,** 721–44 (2010).
15. Xie, M., Haellman, V. & Fussenegger, M. Synthetic biology—application-oriented cell engineering. *Curr. Opin. Biotechnol.* **40,** 139–148 (2016).

16. Arthur, L. L., Pavlovic-Djuranovic, Slavica Koutmou, K. S., Green, R., Szczesny, P. & Djuranovic, S. Translational control by lysine-encoding A-rich sequences. *Sci. Adv.* **1,** e1500154 (2015).

17. Koutmou, K. S. *et al.* Ribosomes slide on lysine-encoding homopolymeric A stretches. *Elife* **4,** 1–18 (2015).

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

19. Melnikov, S. *et al.* One core, two shells: bacterial and eukaryotic ribosomes. *Nat. Struct. Mol. Biol.* **19,** 560–567 (2012).

20. Kuroha, K. *et al.* Receptor for activated C kinase 1 stimulates nascent polypeptide-dependent translation arrest. *EMBO Rep.* **11,** 956–61 (2010).

21. Brandman, O. *et al.* A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell* **151,** 1042–1054 (2012).

22. Egbert, R. G. & Klavins, E. Fine-tuning gene networks using simple sequence repeats. *Proc. Natl. Acad. Sci.* **109,** 16817–16822 (2012).

23. Collins, K. & Gorovsky, M. a. Tetrahymena thermophila. *Curr. Biol.* **15,** R317–R318 (2005).

24. Brandman, O. *et al.* A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell* **151,** 1042–54 (2012).

25. Groth, A. C., Fish, M., Nusse, R. & Calos, M. P. Construction of Transgenic Drosophila by Using the Site-Specific Integrase from Phage phiC31. *Genetics* **166,** 1775–1782 (2004).

26. Duffy, J. B. GAL4 system in Drosophila: a fly geneticist's Swiss army knife. *Genesis* **34,** 1–15 (2002).

27. Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J. & Rolf, B. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet.* **62,** 1408–1415 (1998).

28. Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat. Rev.* **5,** 435–445 (2004).

29. Chung, H. *et al.* Mutation rates of TGFBR2 and ACVR2 coding microsatellites in human cells with defective DNA mismatch repair. *PLoS One* **3,** 2–10 (2008).

30. Li, J. & Zhang, Y. Relationship between promoter sequence and its strength in gene expression. *Eur. Phys. J. E* **37,** 1–6 (2014).

31. Dimitrova, L. N., Kuroha, K., Tatematsu, T. & Inada, T. Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. *J. Biol. Chem.* **284,** 10343–52 (2009).

32. Choe, Y.-J. *et al.* Failure of RQC machinery causes protein aggregation and proteotoxic

stress. *Nature* **531,** 191–195 (2016).

33. Yonashiro, R. *et al.* The Rqc2/Tae2 subunit of the Ribosome-Associated Quality Control (RQC) complex marks ribosome-stalled nascent polypeptide chains for aggregation. *Elife* **5,** (2016).

34. Shaw, W. V. & Leslie, A. G. W. Chloramphenicol acetyl transferase w. *Annu. Rev. Chem. Biomol. Eng. Vol 3* **20,** 363–386 (1991).

35. Leslie, A. G. W. Refined crystal structure of type III chloramphenicol acetyltransferase at 1??75 ?? resolution. *J. Mol. Biol.* **213,** 167–186 (1990).

36. Levdikov, V. M. *et al.* The structure of SAICAR synthase: an enzyme in the de novo pathway of purine nucleotide biosynthesis. *Structure* **6,** 363–76 (1998).

37. Liebman, S. W. & Chernoff, Y. O. Prions in yeast. *Genetics* **191,** 1041–1072 (2012).

38. Hieter, P., Mann, C., Snyder, M. & Davis, R. W. Mitotic stability of yeast chromosomes: A colony color assay that measures nondisjunction and chromosome loss. *Cell* **40,** 381–392 (1985).

39. Mano, Y., Kobayashi, T. J., Nakayama, J. ichi, Uchida, H. & Oki, M. Single Cell Visualization of Yeast Gene Expression Shows Correlation of Epigenetic Switching between Multiple Heterochromatic Regions through Multiple Generations. *PLoS Biol.* **11,** (2013).

40. Hui, A. & de Boer, H. a. Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **84,** 4762–6 (1987).

41. Hui, A. *et al.* Directing Ribosomes to a Single mRNA Species: A Method to Study Ribosomal RNA Mutations and Their Effects on Translation of a Single MEssenger in Escherichia coli. *Methods Enzymol.* **153,** 432–452 (1987).

42. Lee, K., Holland-Staley, C. A. & Cunningham, P. R. Genetic analysis of the Shine-Dalgarno interaction: Selection of alternative functional mRNA-rRNA combination. *RNA* **2,** 1270–1285 (1996).

43. Rackham, O. & Chin, J. W. A network of orthogonal ribosome·mRNA pairs. *Nat. Chem. Biol.* **1,** 159–166 (2005).

44. Nagy, A. *et al.* Dissecting the role of N-myc in development using a single targeting vector to generate a series of alleles. *Curr. Biol.* **8,** 661–664 (1998).

45. Hirotsune, S. *et al.* Graded reduction of Pafah1b1 (Lis1) activity results in neuronal migration defects and early embryonic lethality. *Nat. Genet.* **19,** 333–339 (1998).

46. Wolpowitz, D. *et al.* Cysteine-rich domain isoforms of the neuregulin-1 gene are required for maintenance of peripheral synapses. *Neuron* **25,** 79–91 (2000).

47. Blake, C., Tsao, J. L., Wu, a & Shibata, D. Stepwise deletions of polyA sequences in mismatch repair-deficient colorectal cancers. *Am. J. Pathol.* **158,** 1867–70 (2001).

48. Zhivotovsky, L. a *et al.* The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am. J. Hum. Genet.* **74,** 50–61 (2004).

49. O'Connor, S. E. Engineering of Secondary Metabolism. *Annu. Rev. Genet.* **49,** 71–94 (2015).

50. Braff, D., Shis, D. & Collins, J. J. Synthetic biology platform technologies for antimicrobial applications. *Adv. Drug Deliv. Rev.* **105,** 35–43 (2015).

51. Wu, C.-Y., Rupp, L. J., Roybal, K. T. & Lim, W. A. Synthetic Biology Approaches to Engineer T Cells. *Curr. Opin. Immunol.* **33,** 123–130 (2015).

52. Carbonell, P., Planson, A.-G., Fichera, D. & Faulon, J.-L. A retrosynthetic biology approach to metabolic pathway design for therapeutic production. *BMC Syst. Biol.* **5,** 122 (2011).

53. Motl, J. A. & Chalker, D. L. Zygotic expression of the double-stranded RNA binding motif protein Drb2p is required for DNA elimination in the ciliate Tetrahymena thermophila. *Eukaryot. Cell* **10,** 1648–1659 (2011).

54. Shang, Y. *et al.* A robust inducible-repressible promoter greatly facilitates gene knockouts, conditional expression, and overexpression of homologous and heterologous genes in Tetrahymena thermophila. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 3734–9 (2002).

55. Cassidy-Hanley, D. *et al.* Germline and somatic transformation of mating *Tetrahymena thermophila* by particle bombardment. *Genetics* **146,** 135–47 (1997).

56. Bruns, P. J. & Cassidy-Hanley, D. Biolistic transformation of macro- and micrnuclei. *Methods cell biolgy* **62,** 303–305 (2000).

57. Matsuda, A., Shieh, A. W. Y., Chalker, D. L. & Forney, J. D. The conjugation-specific Die5 protein is required for development of the somatic nucleus in both Paramecium and Tetrahymena. *Eukaryot. Cell* **9,** 1087–1099 (2010).

58. Koncz, C. & Schell, J. The promoter of TL-DNA gene 5 controls the tissue-specific expression of chimaeric genes carried by a novel type of Agrobacterium binary vector. *MGG Mol. Gen. Genet.* **204,** 383–396 (1986).

59. Voinnet, O., Rivas, S., Mestre, P. & Baulcombe, D. An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus. *Plant J.* 949–956 (2003).

60. Joensuu, J. J. *et al.* Hydrophobin fusions for high-level transient protein expression and purification in Nicotiana benthamiana. *Plant Physiol.* **152,** 622–33 (2010).

61. Gutiérrez, S. P., Saberianfar, R., Kohalmi, S. E. & Menassa, R. Protein body formation in stable transgenic tobacco expressing elastin-like polypeptide and hydrophobin fusion proteins. *BMC Biotechnol.* **13,** 40 (2013).

62. Conley, A. J., Joensuu, J. J., Jevnikar, A. M., Menassa, R. & Brandle, J. E. Optimization of elastin-like polypeptide fusions for expression and purification of recombinant proteins

in plants. *Biotechnol. Bioeng.* **103,** 562–573 (2009).

63. Pfeiffer, B. D., Truman, J. W. & Rubin, G. M. Using translational enhancers to increase transgene expression in Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 6626–31 (2012).

64. Ponton, F., Chapuis, M. P., Pernice, M., Sword, G. A. & Simpson, S. J. Evaluation of potential reference genes for reverse transcription-qPCR studies of physiological responses in Drosophila melanogaster. *J. Insect Physiol.* **57,** 840–850 (2011).

Chapter 4: Coding polyA tracks induce mRNA surveillance pathways

4.1 Abstract

The efficiency and fidelity of mRNA translation are key determinants in genes expression. Recently, coding sequences that initiate the mRNA surveillance pathways have been identified as regulatory elements. We have previously reported on coding polyA tracks as a potential regulator of translation elongation for approximately 2% of human genes that contain them. Here, we provide further insight into the mechanism of gene regulation by polyA tracks. We find that iterated adenosine residues induce mRNA degradation by no-go decay and to a greater extent by nonsense mediated decay following ribosomal frameshifting. Additionally, when frameshifting occurs but does not induce nonsense mediated decay, a novel protein is synthesized with a frameshifted amino acid sequence in the c-terminal.

4.2 Introduction

Ribosome associated gene regulation has emerged as an additional role of the mRNA surveillance pathways in recent years. The three surveillance pathways, nonsense mediated decay (NMD), nonstop decay (NSD), and no-go decay (NGD) were all discovered for their role in degrading aberrant mRNA as a quality control mechanism. It is now well known, however, that NMD is also employed by the cell to regulate gene expression. It can be induced by several different molecular mechanisms to degrade physiologically correct mRNAs (reviewed in reference 1). For instance, programmed ribosomal frameshifting in the human *CCR5* mRNA, directs the ribosome to a premature termination codon (PTC) and, thereby, inducing degradation

by NMD. The rate of frameshifting is regulated by miRNAs meaning the rate at which the mRNA is targeted by NMD can be controlled by the cell[2]. Another mechanism to induce NMD is regulated unproductive splicing, a mechanism in which an alternative splicing isoform contains an in-frame premature termination codon (PTC) and is, therefore, targeted for degradation by NMD. The SR family of splicing regulators are autoregulated by this mechanism[3]. mRNA surveillance pathways are an effective method of gene regulation.

Regulatory elements that induce the NSD and NGD pathways are less well understood. Cryptic poly(A) sites in some genes result in transcripts with no stop codon due to premature cleavage and polyadenylation[4,5]. Translation into the premature poly(A) tail results in degradation by the nonstop decay pathway, triggered by ribosome stalling events when the poly(A) tail is translated into a poly lysine peptide sequence[6,7]. In a related mechanism, coding poly lysine or poly arginine sequences that are at least 12 residues long have also been shown to cause ribosomal stalling and degradation by NGD[7,8]. Several labs verified the regulatory potential of poly basic sequences by identifying a small number of genes with stall inducing poly basic sequences in the open reading frame[9,10].

One class of poly basic sequences has proven to be more efficient at decreasing protein expression - poly lysine encoded by the AAA codon. Work by our lab and others found that when a transcript contains at least 4 consecutive AAA codons, mRNA stability and protein expression is significantly decreased compared to a transcript with an equivalent number of AAG codons[11–15]. This observation is true in all model organisms that have been tested, including both prokaryotes and eukaryotes[13]. The discrepancy between expression of transcripts with poly lysine sequences encoded by AAA vs AAG suggest that the mRNA sequence plays a role in destabilization rather than just the nascent peptide sequence. Indeed, experiments in an *in*

vitro translation system showed that the ribosome can frameshift while elongating over consecutive AAA codons but not AAG[11], a result confirmed by several independent reporter assays in *E. coli* and human tissue culture[12,14,15]. Like other programmed ribosomal frameshifting events, polyA induced frameshifting is likely to lead to mRNA destabilization by NMD and degradation of the nascent peptide. The mechanics of the frameshifting event are unclear, however, because efficient ribosomal frameshifting typically requires a secondary stimulatory element such as a stem loop[16], which is not found in genes that endogenously contain coding polyA tracks. One possibility is that the stalling induced by incorporation of iterated lysine residues into the nascent peptide is sufficient to allow frameshifting to occur efficiently.

Here we examine the molecular mechanism of gene regulation by coding polyA tracks. Surprisingly, we find that frameshifting on consecutive AAA codons occurs independently of the charged lysine residues incorporated in the nascent peptide. This very clearly distinguishes polyA mediated genes regulation from regulation by poly basic tracks. Additionally, we show that transcripts harboring a polyA track are targeted for degradation by both NGD and NMD. Knockdown of these pathways individually can deconvolute degradation prompted by peptide mediated stalling and degradation prompted by the ribosome encountering a PTC following frameshifting. Finally, we provide evidence that regulation by polyA tracks can lead to production of alternative, frameshifted proteins when NMD is not induced.

4.2 Results

4.2.1 Mutations in polyA tracks impact gene expression

Our previous work defined coding polyA tracks as a sequence of at least 12 consecutive nucleotides in which all but one are adenosines (12A-1 pattern). The rationale for this definition

was based on expression of reporters with various artificial and endogenous polyA tracks which revealed that the 12A-1 pattern decreased protein expression to approximately 50% of no insertion controls[17].

Analysis of mutations within polyA track in SPIDEX, the largest database that evaluates the impact mutation on gene expression, now provides independent confirmation of the accuracy of the 12A-1 pattern as the most distinguishing between A-rich sequences with measurable effect on gene expression and A-rich sequences with lesser effects. SPIDEX is focusing on alternative splicing events and in theory its data should be reflecting only that kind of regulatory mechanism. However, it was trained using over 500 parameters for each known mutation, we therefore expected that some mutations were assigned as changing splicing event by mistake.

Expression of transcripts that have a 12A-1 pattern was compared to expression of transcripts that have mutations within a 12A-1 polyA track (Fig 1a). Expression change is measured by PSI z-score. Positive z-score indicates increase in expression, measured as the abundance of all transcripts carrying the site, negative z-score is a decrease in expression. Zero change means that mutation was neutral in terms of number of A's, for instance, G to C. Mutations decreasing the length of polyA has a positive impact on expression. The converse is also true. This is consistent with mutation analysis of reporter constructs performed previously[17]. The distributions on the left and right side of zero are significantly different (p-value < 10e-16).

We next asked at which length the difference between left and right side is the largest. In other words, we sought after statistical validation of the choice of 12A-1. The difference in distributions of PSI expression values between mutations that decrease and mutations that increase polyA tracks length was compared for polyA tracks of different lengths (Fig 1b). The -

log(P-value) calculated for each length indicates the difference in expression distributions for mutations that decrease vs increase polyA track length. The largest difference in distributions is found when you divide the tracks into groups: 11A-1 or shorter and 12A-1 or longer. This independently supports the choice of the 12-1A pattern as the regulatory coding polyA motif.

4.2.2 Frameshifting is independent of lysine charge

To determine if frameshifting on polyA tracks requires the incorporation of lysine residues into the nascent peptide, we used a reconstituted *E. coli in vitro* translation system on a series of short, lysine-encoding mRNAs (Fig. 2a). *E. coli* is an ideal system because it uses a single tRNA to decode both the AAA and AAG lysine codons. Electrophoretic TLC (eTLC) resolved the reactions products which were radio labeled with either initiator S^{35}-methionine to visualize all intermediates and full-length products or elongator S^{35}-methionine to visualize only products that incorporate a second methionine (Fig 2b). The frameshift (FS) reporter transcripts are designed so there is not a second methionine encoded in the 0 frame, but there is a methionine codon if there is -1 or +1 frameshifting in the -1 and +1 FS reporters, respectively. The reactions were completed in the presence of Lys-$tRNA^{Lys}$ or Val-$tRNA^{Lys}$ generated with the flexizyme system to mischarge $tRNA^{Lys}$ with a valine residue. Mischarging the tRNA allowed us to decouple the effects of the polyA sequence of the mRNA from the effects of the lysine residues in the nascent peptide.

We find that synthesis of full length and intermediate products proceeds and TLC can resolve MK, MK_2, MK_3, and MK_{4+} as distinct spots on the blot (first blot). In the presence of Lys-$tRNA^{Lys}$, reactions labeled with elongator S^{35}-methionine show methionine incorporation in both $4K_{AAG}$ and 4V controls. Translation of the -1FS-$4K_{AAA}$ also produces peptides that incorporate methionine, indicating that a frameshift occurred in the -1 direction during

translation of the polyA sequence. However, there are no products that incorporate methionine for the +1FS-4K_{AAA} or either 4K_{AAG} transcripts. These results indicate that consecutive lysine residues in the nascent peptide are not alone sufficient to induce ribosomal frameshifting, they must be encoded by the AAA codon. In the presence of Val-tRNA^{Lys}, we observe products from both 4K_{AAG} and 4V controls. Importantly, the peptide translated from the 4K_{AAG} reporter migrates to the same position as the peptide from the 4V reporter. The mischarged tRNA efficiently incorporated valine residues from lysine codons. Despite the absence of lysine residues in the peptide, products are still observed from the -1FS-4K_{AAA} reporter. Ribosomal frameshifting occurred on consecutive AAA codons independent of peptide charge. No frameshifting was observed for +1FS-4K_{AAA} or either 4K_{AAG} reporters.

4.2.3 PolyA track induced frameshifting is directional

Our previous results suggest that frameshifting on polyA tracks occurs only in the -1 direction. We used two additional *in vitro* translation systems to test if the directionality of frameshifting was an artifact of the reconstituted system. A series of GFP reporters were created with four N-terminal lysine codons with either the AAA or AAG codon (Fig. 3a). The FS reporters were programmed to include a stop codon immediately following the lysine codons in two of the three coding frames, so GFP is produced only if frameshifting occurs in the -1 or +1 direction for 4K_{AAA}-(-1FS)-GFP and 4K_{AAG}-(+1FS)-GFP, respectively.

The constructs were transcribed using T7 polymerase and equal amount of mRNA were translated in the rabbit reticulocyte translation system and products analyzed by electrophoresis on SDS page gel (Fig 3b). The presence of four consecutive AAA codons reduced GFP expression in the 0-frame 2-fold compared to WT GFP. We observed production of GFP from both the 4K_{AAA}-(-1FS)-GFP and 4K_{AAG}-(+1FS)-GFP constructs at a lower efficiency than the 0-

frame. Interestingly, there is a 5-fold reduction in GFP in the -1-frame compared to the 0-frame, and another 5-fold reduction in the +1-frame compared to the -1-frame. This result suggests the rather than a true +1 frameshift, the products from $4K_{AAG}$-(+1FS)-GFP may be the result of two -1 frameshifts on the polyA track. We can resolve these two possibilities by determining which amino acids are incorporated in the peptide by mass spectrometry. The mass spec results reveal that the $4K_{AAG}$-(+1FS)-GFP products are the result of two -1 frameshifting events (not shown). The same constructs were translated in *E. coli* NEB system as well as in *E. coli* cells, giving approximately the same results.

4.2.3 Knockdown of NGD and NMD pathways partially rescues polyA track expression

To support our model that coding polyA tracks induced stalling leads to degradation by NGD and NMD, we expressed the human hemoglobin gene (delta chain: HBD) which has six consecutive AAA codons in the second exon, referred to as $HBD_{(AAA)6}$ (Figure 4a) in HDF cells while targeting key factors of both pathways with siRNA. We previously reported that this insertion shows a significant decrease in protein expression and mRNA stability compared to wildtype HBD and an $(AAG)_6$ control insertion[12]. For NGD factors, western blot analysis of the *in vitro* translated prodicts shows an increase in expression of $HBD_{(AAA)6}$ when Pelota, the human homolog of Dom34, is targeted (Fig 4b). This result indicates that transcripts with polyA tracks are at least partially degraded by the Pelota dependent NGD pathway. A similar analysis which targets UPf1, the key factor in NMDl, with siRNA shows a similar increase in $HBD_{(AAA)6}$ expression (Fig 4c). Together these results show that polyA tracks induce mRNA degradation by two mRNA surveillance pathways, NGD and NMD.

4.2.4 PolyA track in last exon produces an alternative protein

NMD is most efficiently activated if the PTC is at least 50 nt 5' of an exon junction. Consequently, if a PTC is encountered in the last exon, NMD is unlikely to be activated and the translation cycle will terminate normally. In the event of frameshifting in the last exon, this will result in a c-terminally extended if the PTC falls after the in-frame stop codon. Therefore, we hypothesize that polyA tracks in the last exon of a gene can cause production of an alternative frameshifting product. We use mCherry reporter constructs with 11 lysine codons, either AAA or AAG, to examine the production of alternative proteins (Fig 5a). We see that protein production of both mCh-11K_{AAA} and mCh-11K_{AAG} is significantly decreased compared to WT mCh, with slightly less expression for mCh-11K_{AAA} (Fig 5b). This result is expected because of the large number of lysine codons in this construct. Eleven lysines are sufficient to cause poly basic peptide stalling of the ribosome so both the AAA and AAG contructs will be degraded by NGD. The mRNA abundance is also approximately equal between the two constructs. The western blot with a long exposure, however, shows that there is an additional product at the expected molecular weight for a frameshift after the polyA sequence.

4.3 Discussion

In this study, we examined the molecular mechanism behind polyA mediated gene expression. Consistent with earlier models, we found that coding polyA tracks induce mRNA degradation by both NGD and NMD, but can also result in the production of a novel protein with a frameshifted amino acid sequence after the polyA. This dual mechanism explains why consecutive AAA codons result in greater down regulation of protein expression compared to other poly basic sequences which induce NGD to the same extent, but do not frameshift to induce NMD or alternative protein production.

Importantly, frameshifting on the consecutive AAA codons is independent of lysine charge, clearly distinguishing this polyA mediated regulation from poly basic mediated gene regulation. Furthermore, this results places polyA tracks among the few sequences which can induce programmed ribosomal frameshifting without a secondary stimulatory element. Identification of the mechanical forces that allow for frameshifting on iterated adenosine residues will inform our understanding of ribosome frame maintenance. We also provide evidence that polyA tracks in endogenous genes potentially produce novel frameshifted proteins, adding complexity to the proteome.

4.4 Methods

In vitro translation from *Escherichia coli* purified components

*In vitro t*ranslation of polyA and control transcripts was performed as previously described[11]. Briefly, mRNAs were transcribed from DNA plasmid templates with T7 polymerase. *E. coli* ribosomes were programed with mRNAs and f-[^{35}S]-Met-tRNA^{Met} or unlabeled f-tRNA^{Met} in the P site to form 70S initiation complexes (ICs). Translation was initiated by adding equal volumes of ternary complex to ICs. [^{35}S]-Met-tRNA^{Met} was included for when the IC was unlabeled to label final products. Reactant, intermediates, and products were run on electrophoretic TLC and visualized by phosophorimaging.

NEB PURExpress *in vitro* translation system

GFP frameshift reporters were transcribed by T7 polymerase to generate mRNA from DNA plasmids. *In vitro* translation was performed according to the manufacturer's protocol.

Cell Culture

HDF cells were cultured in Dulbecco's modified Eagle's medium (DMEM) (Gibco) and supplemented with 10% fetal bovine serum, 5% minimum essential medium nonessential amino acids (100×, Gibco), 5% penicillin and streptomycin (Gibco), and L-glutamine (Gibco). Plasmids (1 ug) and siRNA (25 pg) were introduced to the cells by the Neon Transfection System (Invitrogen) with 100-µl tips according to cell-specific protocols (www.lifetechnologies.com/us/en/home/life-science/cell-culture/transfection/transfection---selection-misc/neon-transfection-system/neon-protocols-cell-line-data.html). Cells electroporated with DNA plasmids and siRNA were harvested after 48 hours.

DNA constructs

mCherry reporter constructs were generated by PCR amplification of an mCherry template with forward primers containing the test sequence at the 5′ end and homology to mCherry at the 3′ end. The test sequence for each construct is listed in the following table. The PCR product was purified by NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel) and integrated into the pcDNA-DEST40.

Whole gene TAF1D constructs were generated by PCR amplification from Life Technologies GeneArt Strings DNA Fragments and cloned in pcDNA-DEST40 vector for expression. Synonymous mutations in the natural gene homopolymeric lysine runs were made by site-directed mutagenesis. Human β-globin gene (delta chain; HBD) was amplified from genomic DNA isolated from HDF cells. Insertions of poly(A) track was made by site-directed mutagenesis.

RNA extraction and qRT-PCR

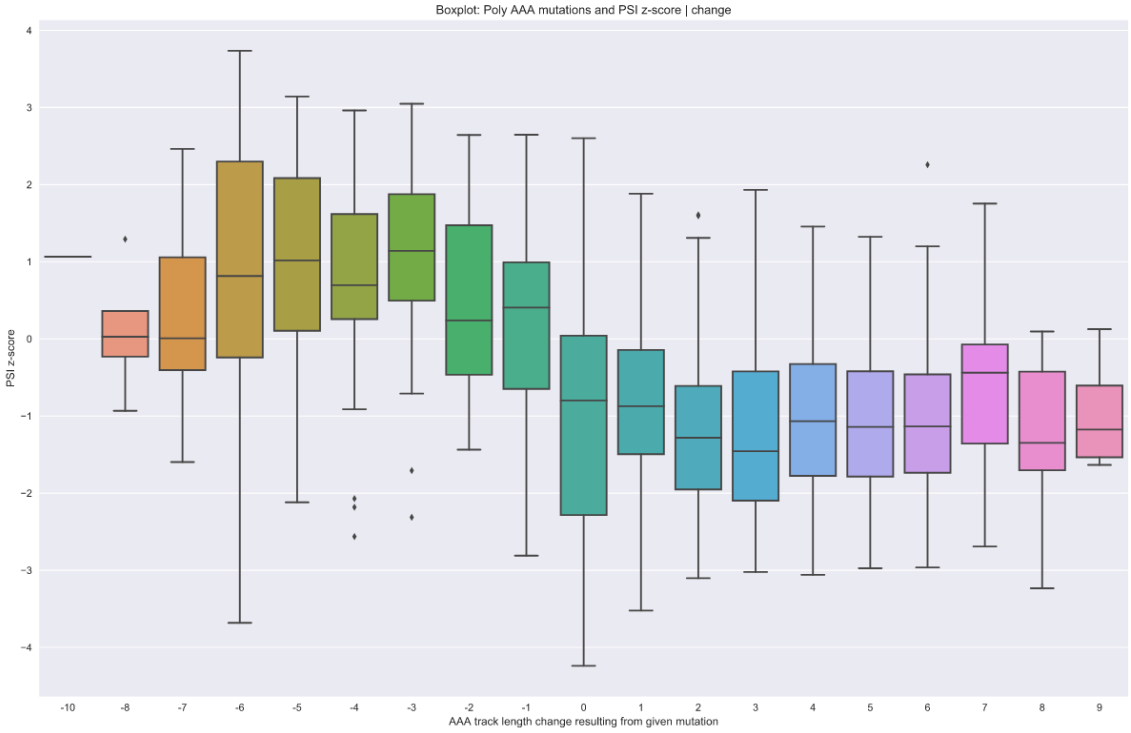
Total RNA was extracted from cells using the RiboZol RNA extraction reagent (Amresco) according to the manufacturer's instructions. RiboZol reagent (400 μl) was used in each well of 6- or 12-well plates for RNA extraction. Precipitated nucleic acids were treated with Turbo deoxyribonuclease (Ambion), and total RNA was dissolved in ribonuclease-free water and stored at −20°C. RNA concentration was measured by NanoDrop (OD260/280). iScript Reverse Transcription Supermix (Bio-Rad) was used with 1 μg of total RNA following the manufacturer's protocol. iQ SYBR Green Supermix (Bio-Rad) protocol was used for qRT-PCR on the CFX96 Real-Time system with Bio-Rad CFX Manager 3.0 software. Cycle threshold (C_t) values were normalized to the neomycin resistance gene expressed from the same plasmid.

Western blot analysis

Total cell lysates were prepared with passive lysis buffer (Promega). Blots were blocked with 5% milk in 1× tris-buffered saline–0.1% Tween 20 (TBST) for 1 hour. Horseradish peroxidase–conjugated or primary antibodies were diluted according to the manufacturer's recommendations and incubated overnight with membranes. The membranes were washed four times for 5 min in TBST and prepared for imaging, or secondary antibody was added for additional 1 hour of incubation. Images were generated by Bio-Rad Molecular Imager ChemiDoc XRS System with Image Lab software by chemiluminescence detection.

4.5 Figures

A.



B. \

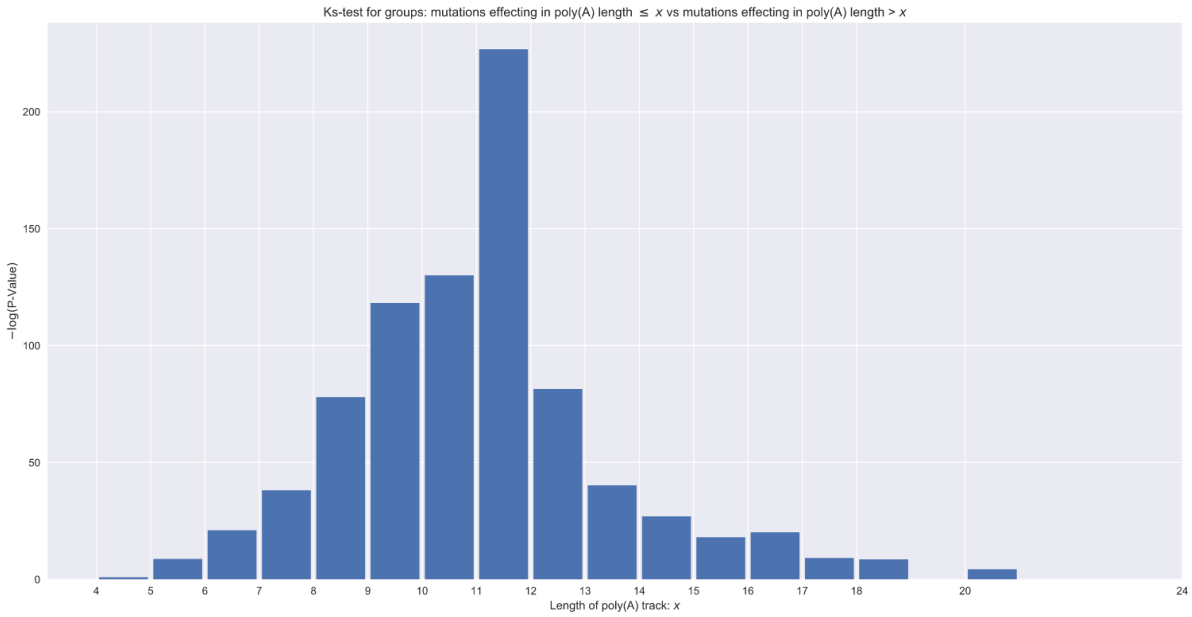


Figure 4.1. Effect of mutation within polyA tracks of endogenous genes on expression. (A) Expression of polyA track transcripts in the SPIDEX database measured as the abundance of all transcripts carrying the mutation plotted by change in polyA track length. Zero change means the mutation does not impact polyA track length. A positive z-score indicated expression increases while a negative score means expression decreases. (B) Measure of difference in expression distributions of mutated transcripts that increase polyA track length compared to those that decrease it. The difference is measured as -log(P-Value) and is plotted by starting polyA track length.

A.

Lane	Transcript and peptide sequence	Translation Frame
1	AUG AAG AAG AAG AAG AUG UAA	
	Met Lys Lys Lys Lys Met stop	0
2	AUG GUG GUG GUG GUG AUG UAA	
	Met Val Val Val Val Met stop	0
3	AUG AAA AAA AAA AAA ugg uaa	
	Met Lys Lys Lys Lys Trp stop	0
	Met Lys Lys Lys Lys Met Val	-1
4	AUG AAA AAA AAA AAA CAU GUA UAA	
	Met Lys Lys Lys Lys His Trp stop	0
	Met Lys Lys Lys Lys Met Try	+1
5	AUG AAG AAG AAG AAG ugg uaa	
	Met Lys Lys Lys Lys Trp stop	0
	Met Lys Lys Lys Lys Met Val	-1
	tgfrrr	
6	AUG AAG AAG AAG AAG CAU GUA UAA	
	Met Lys Lys Lys Lys His Trp stop	0
	Met Lys Lys Lys Lys Met Try	+1

B.

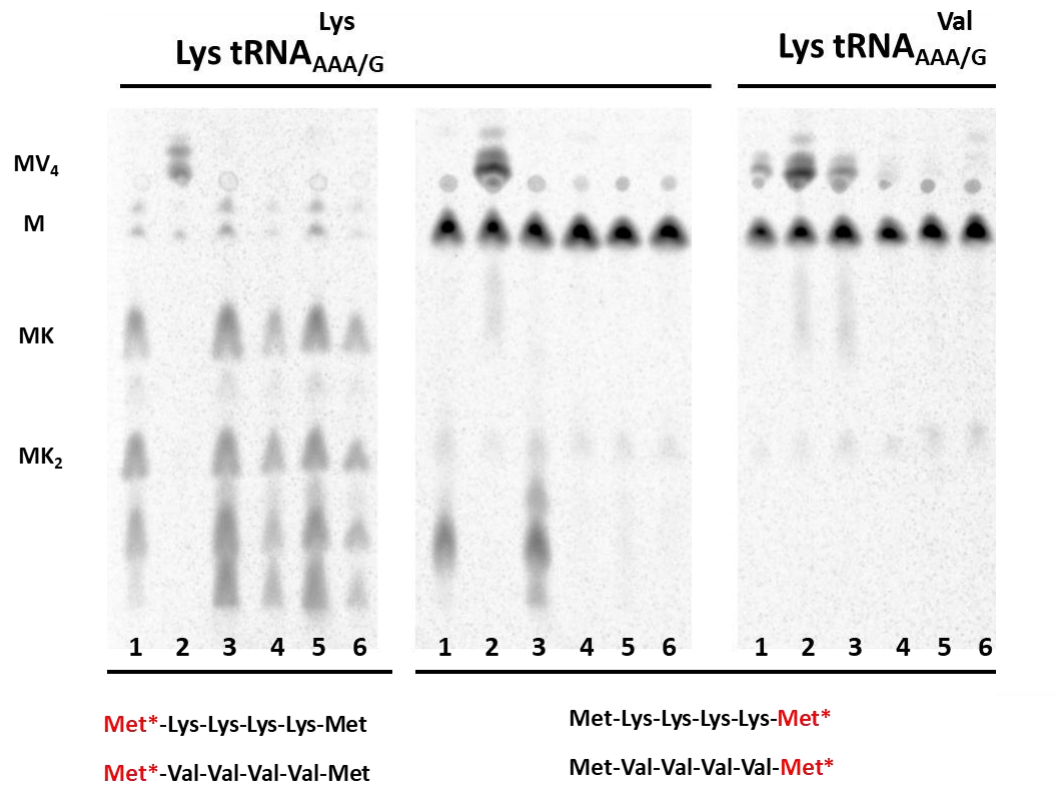


Figure 4.2. Effect of lysine incorporation in the nascent peptide on ribosomal frameshifting. (A) Sequences of MK_4 and control mRNA used for *in vitro* translation. The lane, mRNA sequence and expected amino acid sequence are indicated. The 0 frame amino acid sequence relative to the start codon and either -1 or +1 frame after the polyA sequence are included depending on which frame is designed to have the second Met. (B) eTLC displaying the *in vitro* translated products of MK4 and control transcripts.

A.

B.

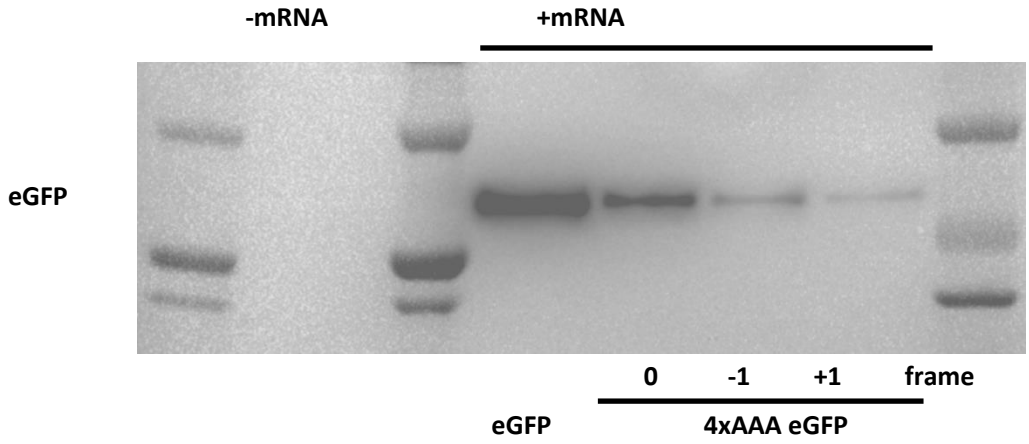
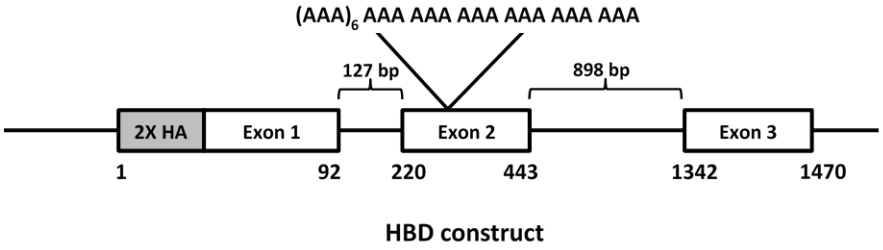


Figure 3. Directional frameshifting on polyA tracks.

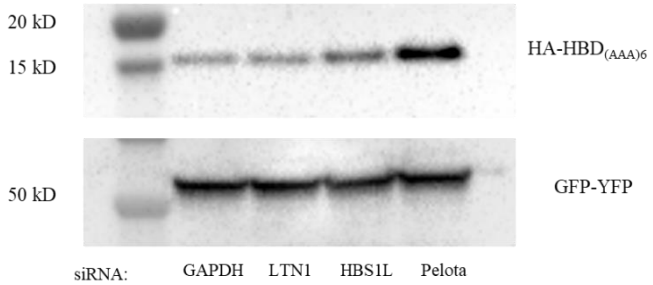
(A) Diagram of frameshifted GFP constructs. Frame of stop codons inserted before GFP sequence are labeled directly before the HA tag. Frame expected to express GFP is indicated to the right. (B) Western blot analysis of *in vitro* translated products. Probed with anti-GFP.

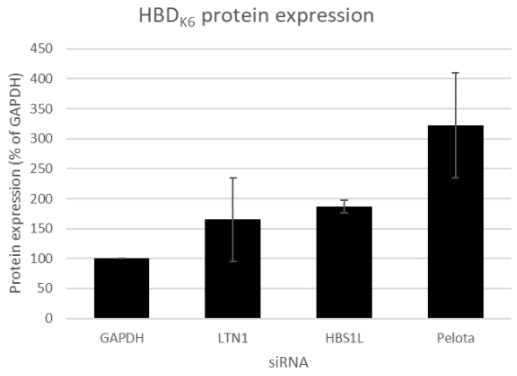
A.



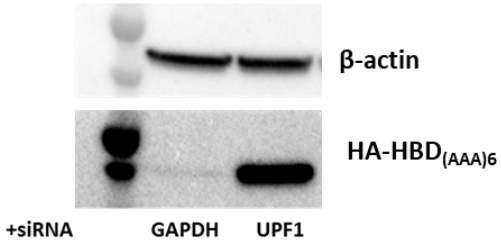
HBD construct

B.





C.



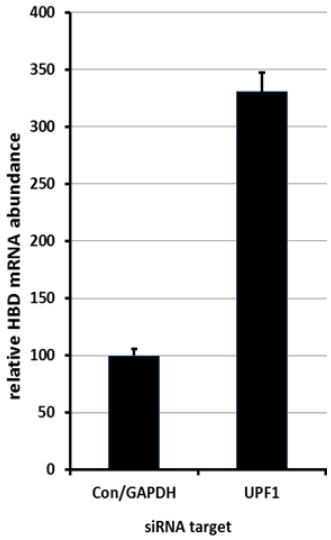
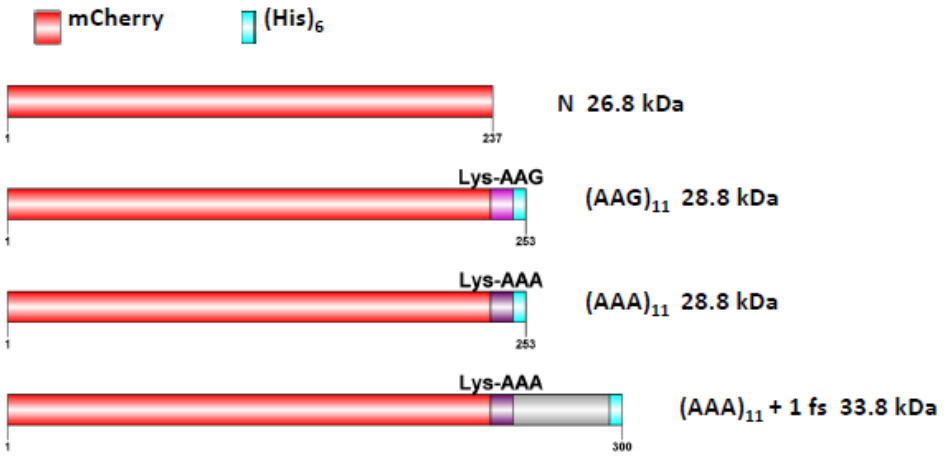


Figure 4.4. Rescue of polyA track reporter by knockdown of NGD and NMD factors.

(A) Diagram of hemoglobin delta chain with position of polyA insertion indicated. (B) Western blot analysis and quantification of $HBD_{(AAA)6}$ expressed in HDF cells with various siRNA to knock down factors of the NGD pathway. (C) Western blot analysis of $HBD_{(AAA)6}$ expression in HDF cells with siRNA targeting the NMD factor, UPF1 and mRNA stability measured by qPCR.

A.


mCherry
(His)6
N 26.8 kDa
Lys-AAG
(AAG)11 28.8 kDa
Lys-AAA
(AAA)11 28.8 kDa
Lys-AAA
(AAA)11 + 1 fs 33.8 kDa
1
237
253
300

B.

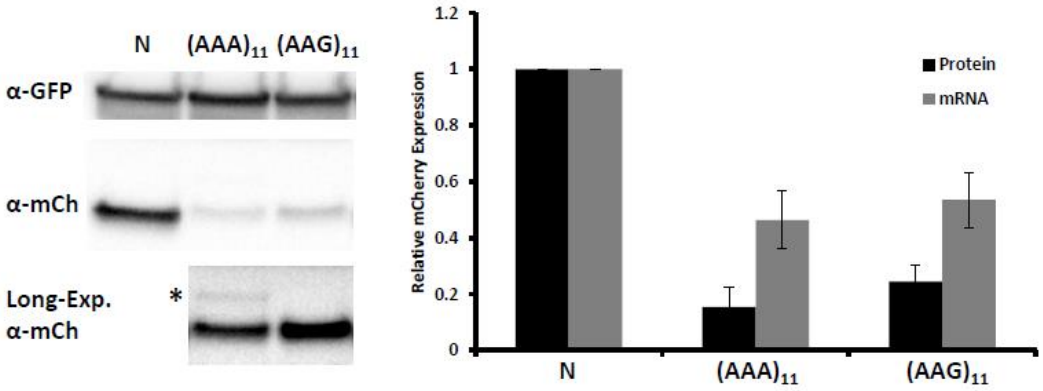

N
(AAA)11
(AAG)11
α-GFP
α-mCh
Long-Exp.
α-mCh
*
Relative mCherry Expression
Protein
mRNA
1.2
1
0.8
0.6
0.4
0.2
0

Figure 4.5. Production of alternative frameshifted protein.

(A) Diagram of mCherry constructs with poly lysine insertion in the c-terminus. The length of the poly lysine sequence and codon is indicated along with the expected molecular weight of full length protein. The third construct removes the in-frame stop codon to demonstrate the size of expected frameshifted protein. (B) Western blot analysis of c-terminal polyA constructs expressed in HDF cells. A long-exposure is included to show the faint band present at the expected molecular weight of a protein frameshifted after the polyA sequence.

4.6 References

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

2. Belew, A. T. *et al.* Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway. *Nature* **512,** 265–269 (2014).

3. Lareau, L. F., Inada, M., Green, R. E., Wengrod, J. C. & Brenner, S. E. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature* **446,** 926–929 (2007).

4. Frischmeyer, P. a *et al.* An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. *Science (80-.).* **295,** 2258–61 (2002).

5. Ozsolak, F. *et al.* Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell* **143,** 1018–1029 (2010).

6. Ito-Harashima, S., Kuroha, K., Tatematsu, T. & Inada, T. Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast. *Genes Dev.* **21,** 519–24 (2007).

7. Lu, J. & Deutsch, C. Electrostatics in the ribosomal tunnel modulate chain elongation rates. *J. Mol. Biol.* **384,** 73–86 (2008).

8. Bengtson, M. H. & Joazeiro, C. a P. Role of a ribosome-associated E3 ubiquitin ligase in protein quality control. *Nature* **467,** 470–3 (2010).

9. Dimitrova, L. N., Kuroha, K., Tatematsu, T. & Inada, T. Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. *J. Biol. Chem.* **284,** 10343–52 (2009).

10. Brandman, O. *et al.* A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell* **151,** 1042–1054 (2012).

11. Koutmou, K. S. *et al.* Ribosomes slide on lysine-encoding homopolymeric A stretches. *Elife* **4,** 1–18 (2015).

12. Arthur, L. L., Pavlovic-Djuranovic, Slavica Koutmou, K. S., Green, R., Szczesny, P. & Djuranovic, S. Translational control by lysine-encoding A-rich sequences. *Sci. Adv.* **1,** e1500154 (2015).

13. Arthur, L. L. *et al.* Rapid generation of hypomorphic mutations. *Nat. Commun.* **8,** 1–15 (2017).

14. Juszkiewicz, S. & Hegde, R. S. Initiation of Quality Control during Poly(A) Translation Requires Site-Specific Ribosome Ubiquitination. *Mol. Cell* 1–8 (2017).

doi:10.1016/j.molcel.2016.11.039

15. Sundaramoorthy, E. *et al.* ZNF598 and RACK1 Regulate Mammalian Ribosome-Associated Quality Control Function by Mediating Regulatory 40S Ribosomal Ubiquitylation. *Mol. Cell* **65,** 1–10 (2017).

16. Ketteler, R. On programmed ribosomal frameshifting: The alternative proteomes. *Front. Genet.* **3,** 1–10 (2012).

17. Arthur, L. L., Pavlovic-Djuranovic, Slavica Koutmou, K. S., Green, R., Szczesny, P. & Djuranovic, S. Translational control by lysine-encoding A-rich sequences. *Sci. Adv.* **1,** e1500154 (2015).

Chapter 5: Conclusions and Future Directions

5.1 Summary

Gene expression must be strictly regulated to ensure correct cell development and survival. In fact, improper gene expression is a hallmark of cellular pathology and disease, making it a criticle area of study. In this work, I have shown that coding polyA tracks are a potent form of gene regulation that rely on the dual mechanism of ribosomal stalling and frameshifting. By distrupting the translation elongation cycle, the presence of polyA tracks in a gene sequences leads to decreased mRNA stability and protein expression. The process of translation elongation is highly conserved across evolutionarily distant organisms, meaning any regulation mechanism that targets it will be nearly universal. Indeed, polyA tracks negatively impact expression of genes in all model organisms that have been investigated. This predictability makes insertiong of polyA tracks into a gene of interest a convienient tool for creation of hypomorphic mutations by investigators. While much of the molecular mechanism has been determined, the exact physical interactions between the ribosomes and consecutive adenosines that recruit stall promoting factors and allow frameshifting is still uknown. Additional studies are necessary to continue to define the mechanism and biological significance of coding polyA mediated regulation.

5.2 Future Directions

5.2.1 Dynamic regulation of polyA mediated stalling and frameshifting

Disregarding potential mutations to the DNA sequence, the length and composition of a polyA track in a gene is static. Even across species, the polyA track is typically conserved. How, then, does the cell increase expression of polyA containing genes? It is possible that increased transcription is sufficient to compensate for degradation by the mRNA surveillance pathways, but this would be an inefficient use of resources for the cell. The discovery of ZNF598 and RACK1 ubiquitylation of the ribosome to promote stalling raises a more likely possibility – regulation of these factors can increase or decrease expression of polyA track genes. It is already known that RACK1 is regulated by many cell signaling pathways and during the mitotic cell cycle. Investigations of the expression pattern of ZNF598 are necessary to reveal if it is similarly regulated.

5.2.2 Role of synonymous mutation of polyA tracks in cancer

Coding polyA mediated regulation depends on the mRNA sequence rather than the translated peptide sequence. Beucase of this characteristic, synonymous mutations that do not change the peptide sequence can have dramatic effect on protein expression or change the rate at which an alternative, frameshifting product is synthesized. Moving forward, it will be important to examine the role of mutations in coding polyA genes for their role in human disease, such as cancer.