HHS Public Access

Author manuscript

Sci Adv. Author manuscript; available in PMC 2015 August 28.

Published in final edited form as:

Sci Adv. 2015 July ; 1(6): . doi:10.1126/sciadv.1500154.

Translational control by lysine-encoding A-rich sequences

Laura Arthur[1,*], **Slavica Pavlovic-Djuranovic**[1,*], **Kristin Smith-Koutmou**[2], **Rachel Green**[2,3], **Pawel Szczesny**[4], and **Sergej Djuranovic**[1]

[1]Washington University School of Medicine, Department of Cell Biology and Physiology. 600 South Euclid Avenue, Campus Box 8228, St. Louis, MO 63110 [2]Johns Hopkins School of Medicine, Department of Molecular Biology and Genetics. 725 N. Wolfe Street, Baltimore, MD 21205 [3]Howard Hughes Medical Institute [4]Institute of Biochemistry and Biophysics Polish Academy of Sciences, Department of Bioinformatics, Pawinskiego 5a, 02-106 Warsaw, Poland

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 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.

Keywords

Lysine; ribosome stalling; gene regulation; mRNA stability; synonymous mutations

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).

Correspondence to: Pawel Szczesny; Sergej Djuranovic.

*These authors contributed equally to this work.

Competing Interests: The authors declare that they have no competing interests.

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 trigger translational arrest, as observed with damaged bases, stable stem-loop structures (4), rare codons (5) or mRNAs lacking stop codons (non-stop mRNAs) (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 poly(A) coding sequences in human cells, demonstrating that, these sequences unexpectedly induce ribosome pausing directly, without a role for the encoded basic peptide.

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 under-represented 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 standard 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 strings of lysine 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 (AAA_6) 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 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-encoding mCherry constructs (Supplementary Fig. S5); these data indicate that accelerated decay of polylysine mCherry mRNAs is dependent on translation. Consistent with this observation, the insertion of 36As (sequence equivalent to twelve lysine AAA codons) after the stop codon, in the 3'UTR region, did not affect the protein expression level or mRNA stability of the assayed construct (Supplementary Fig S6). Insertion of polylysine codons at different positions along the coding sequence drastically reduced reporter expression and mRNA levels independent of the relative position in the construct. As such, it follows that the observed changes in mRNA stability (Fig. 1D) result from a 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 constructs containing human hemoglobin gene (delta chain, HBD) 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 simly due to $tRNA^{Lys}$ abundance, since the 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 $tRNA^{Lys(UUU)}$ 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 under-representation of poly(A) nucleotide runs in regions coding for iterated lysines (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 are observed in other vertebrates (Supplementary Fig. S10).

Ribosome profiling data have the potential to reveal features of pausing on polybasic stretches throughout the genome (16). A cumulative analysis of three ribosome profiling datasets from human cells for regions encoding 4 lysines in a row revealed that the occupancy pattern on 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 amino acids observed by other researchers (17). This suggests that the observed effect on protein output and mRNA stability is dependent on nucleotide, not simply the amino acid 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). Together, these data suggest that attenuation of translation on poly(A) nucleotide tracks occurs via a different mechanism than just the interaction of positively charged residues with the 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 reduces 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 3A). 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 a 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 propose that a run 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 the A string to begin in any particular codon frame, the sequence may 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 (18)). 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 over-representation 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 3D). The proteins encoded by these mRNAs are often conserved among 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 (19) (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 poly(A) 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 4A–C, 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. 4B–C 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 4D–F, 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 poly(A) track of the *ZCRB1* gene (K137K; 411G>A) resulted in a significant decrease in both protein expression and mRNA stability (Fig. 4E and F, 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 modulate protein production from these genes.

Poly(A) tracks resemble ribosome “slippery” sequences that have been associated with translational frame-shifts (20,21). Recent studies suggest that polyA tracks can induce ”sliding” of *E. coli* ribosomes resulting in frameshifting (10,22). 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 **C**atalogue **O**f **S**omatic **M**utations **I**n **C**ancer, http://cancer.sanger.ac.uk, (23)) 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 (24, 25). Our recent data suggests that NMD may play a role in determining the stability of poly(A) track-

containing mRNAs. Deletion of NMD factor Upf1p in yeast cells partially rescues mRNA levels from constructs with simple 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, (26–29)). Based on the position of the poly(A) tracks, and their position relative to possible PTCs in the −1 and +1 frame, and the location of downstream exon-intron boundaries, we find that a part 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 considerable number of human poly(A) track genes may not elicit NMD response since PTCs in both −1 and +1 frame following poly(A) tracks are less than 50 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 (Supplementary 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.

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 amino acid sequence and the attenuation occurs in a distinct 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 (30)).

Materials and methods

Experimental protocols

Cell culture—HDF cells were cultured in Dulbecco’s modified Eagle’s medium (Gibco) and supplemented with 10% fetal bovine serum, 5% MEM non-essential amino acids (100×, Gibco), 5% penicillin and streptomycin (Gibco), and L-glutamine (Gibco). T-Rex-CHO cells were grown in Ham’s F12K medium (ATCC) with the same supplements. Drosophila S2 cells were cultured in Express Five SFM Medium (Invitrogen) supplemented with 100 units per milliliter penicillin, 100 units per milliliter streptomycin (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 (http://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 CloneId:5298467) or Life Technologies GeneArt Strings DNA Framents (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 beta-globin gene (HBD, delta chain) was amplified from genomic DNA isolated from HDF cells. Insertions of poly(A)-track, AAG-codons or pre-mature stop codon in HBD constructs were made by site directed mutagenesis. Sequences of inserts are in the Table S6.

***In vitro* mRNA synthesis—**Capped and polyadenylated mRNA was synthesized *in vitro* using mMessage mMachine T7 Transcription Kit (LifeTechnologies) following manufacturers 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. 400 µl of Ribozol reagent was used per well of 6 or 12 well plates for RNA extraction. Precipitated nucleic acids were treated by Turbo DNAse (Ambion) and total RNA was dissolved in RNAse-free water and stored at −20°C. RNA concentration was measured by Nanodrop (OD260/280). iScript Reverse Transcription Supermix (Biorad) was used with 1 µg of total RNA following the manufacturer's protocol. iQ SYBR Green Supermix (Biorad) protocol was used for qRT-PCR on the CFX96 Real-Time system with Bio-Rad CFX Manager 3.0 software. Cycle threshold (Ct) 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 × TBS 0.1% Tween-20 (TBST) for 1 hour. HRP-conjugated or primary antibodies were diluted by manufacturer recommendations and incubated overnight with membranes. Membranes were washed 4 times for 5 minutes in TBST and prepared for imaging or secondary antibody was added for

additional one 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 hr with Pierce DyLight secondary antibodies.

Immunoprecipitation—Total cell lysates were prepared with passive lysis buffer (Promega) and incubated with Pierce anit-HA magnetic beads overnight at 4°C. Proteins were eluted by boiling the beads with 1× SDS sample buffer for 7 minutes. 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 6 well plates with the optically clear bottom. Prioir to imaging cells were washed with a fresh DMEM media without Phenol-Red and incubated 20 minutes with DMEM media containing 0.025% Hoechst 33342 dye for DNA staining. Cells were washed with DMEM media and imaged in Phenol-Red free media using EVOS-FL microscop using 40× objective. Images were analyzed using EVOS-FL software.

Bioinformatics analysis

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

mRNA mapping—As we observed some inconsistences between transcripts and proteins in some of the sequence databases, before starting the analyses we mapped protein sequences to mRNA sequences using 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), a first one was chosen, as the choice of corresponding isoform (as 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. These were:

- GSE51424 prepared by Gonzales and coworkers (33) from which samples: SRR1562539, SRR1562540 and SRR1562541 were used;
- GSE48933 prepared by Rooijers and coworkers (34) from which samples: SRR935448, SRR935449, SRR935452, SRR935453, SRR935454 and SRR935455 were used;
- GSE42509 prepared by Loayza-Puch and coworkers (35) from which samples SRR627620–SRR627627 were used.

The data were analyzed similarly to the original protocol created by Ingolia and coworkers (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 polyA were not highly expressed and the data were sparse (several positions with no occupancy), instead of mean of 30 codons prior to polyA 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 SNPs in polyA regions vs random region of the same length in other genes, we needed to use the same distribution of lengths in both cases. The distribution of lengths for polyA regions identified as mentioned above (12 As allowing for one mismatch) up to length 19 (longer are rare) is presented in the Fig. S19. Using the same distribution of lengths, we selected one random region of length drawed from the distribution randomly placed along each gene from all human protein coding RNAs. Distributions of number of SNPs per segment for all polyA segments and for one random segment for each mRNA were compared using Welch Two Sample t-test, Wilcoxon rank sum test with continuity correction and two sample permutation test with 100000 permutations.

Abundance of polytracks in protein sequences—Abundance was expressed by a following equation:

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

where N_P is 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 aminoacid. It is to normalize against variable aminoacid presence in different organisms. All isoforms of proteins were taken into account.

Other analyses—List of human essential genes was obtained from the work of Georgi and coworkers (38). Gene Ontology analyses were done using Term Enrichment Service at http://amigo.geneontology.org/rte. Most of graphs were prepared using R and GGPLOT2 library. For the Fig. 3A, the values of the Y-axis are computed by 1D gaussian kernel density estimates implemented in R software. Custom Perl scripts were used to analyze and merge the data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank D. Owyoung, J. T. Mendell, J. Coller and T. Schedl for helpful comments. We would also like to thank the National Institutes of Health for funding (T32 GM: 007067 to LA and F32 GM100608 to KSK). This work was supported by a grant from the American Cancer Society IRG-58-010-58-2 to SD.

References

1. Dinman JD, Berry MJ. 22 Regulation of Termination and Recoding. Cold Spring Harb. Monogr. Arch. 2007; 48 (available at https://cshmonographs.org/index.php/monographs/article/view/3291).
2. Hershey JWB, Sonenberg N, Mathews MB. Principles of translational control: an overview. Cold Spring Harb. Perspect. Biol. 2012; 4
3. Shoemaker CJ, Green R. Translation drives mRNA quality control. Nat. Struct. Mol. Biol. 2012; 19:594–601. [PubMed: 22664987]
4. Doma MK, Parker R. Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. Nature. 2006; 440:561–564. [PubMed: 16554824]
5. Letzring DP, Dean KM, Grayhack EJ. Control of translation efficiency in yeast by codon-anticodon interactions. RNA N. Y. N. 2010; 16:2516–2528.
6. Dimitrova LN, Kuroha K, Tatematsu T, Inada T. Nascent peptide-dependent translation arrest leads to Not4p-mediated protein degradation by the proteasome. J. Biol. Chem. 2009; 284:10343–10352. [PubMed: 19204001]
7. Kuroha K, et al. Receptor for activated C kinase 1 stimulates nascent polypeptide-dependent translation arrest. EMBO Rep. 2010; 11:956–961. [PubMed: 21072063]
8. Brandman O, et al. A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. Cell. 2012; 151:1042–1054. [PubMed: 23178123]
9. Lu J, Deutsch C. Electrostatics in the ribosomal tunnel modulate chain elongation rates. J. Mol. Biol. 2008; 384:73–86. [PubMed: 18822297]
10. Koutmou KS, Schuller AP, Brunelle JL, Radhakrishnan A, Djuranovic S, Green R. eLIFE. 2015
11. 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. 2012; 46:518–529. [PubMed: 22503425]
12. Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. Amino acid runs in eukaryotic proteomes and disease associations. Proc. Natl. Acad. Sci. U. S. A. 2002; 99:333–338. [PubMed: 11782551]
13. Djuranovic S, Nahvi A, Green R. miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. Science. 2012; 336:237–240. [PubMed: 22499947]
14. Barr JN, Wertz GW. 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. 2001; 75:6901–6913. [PubMed: 11435570]
15. Fresno M, Jiménez A, Vázquez D. Inhibition of translation in eukaryotic systems by harringtonine. Eur. J. Biochem. FEBS. 1977; 72:323–330.
16. Ingolia NT. Ribosome profiling: new views of translation, from single codons to genome scale. Nat. Rev. Gen. 2014; 15:205–213.
17. Charneski CA, Hurst LD. Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. PLoS Biol. 2013; 11:e1001508. [PubMed: 23554576]
18. Pruitt KD, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014; 42:D756–D763. [PubMed: 24259432]
19. McLaughlin SK, et al. The RasGAP gene, RASAL2, is a tumor and metastasis suppressor. Cancer Cell. 2013; 24:365–378. [PubMed: 24029233]
20. Belfield EJ, Hughes RK, Tsesmetzis N, Naldrett MJ, Casey R. The gateway pDEST17 expression vector encodes a −1 ribosomal frameshifting sequence. Nucleic Acids Res. 2007; 35:1322–1332. [PubMed: 17272299]

21. Chen J, et al. Dynamic pathways of −1 translational frameshifting. Nature. 2014; 512:328–332. [PubMed: 24919156]
22. Yan S, Wen JD, Bustamante C, Tinoco I Jr. Ribosome excursions during mRNA translocation mediate broad branching of frameshift pathways. Cell. 2015; 160(5):870–881. [PubMed: 25703095]
23. Forbes SA, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2014
24. Belew AT, Advani VM, Dinman JD. Endogenous ribosomal frameshift signals operate as mRNA destabilizing elements through at least two molecular pathways in yeast. Nucleic Acids Res. 2011; 39:2799–2808. [PubMed: 21109528]
25. Belew AT, et al. Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway. Nature. 2014; 512:265–269. [PubMed: 25043019]
26. Lykke-Andersen J, Shu MD, Steitz JA. Human Upf proteins target an mRNA for nonsense-mediated decay when bound downstream of a termination codon. Cell. 2000; 103:1121–1131. [PubMed: 11163187]
27. Le Hir H, Gatfield D, Izaurralde E, Moore MJ. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay. EMBO J. 2001; 20:4987–4997. [PubMed: 11532962]
28. Chang Y-F, Imam JS, Wilkinson MF. The nonsense-mediated decay RNA surveillance pathway. Annu. Rev. Biochem. 2007; 76:51–74. [PubMed: 17352659]
29. Popp MW-L, Maquat LE. Organizing principles of mammalian nonsense-mediated mRNA decay. Annu. Rev. Genet. 2013; 47:139–165. [PubMed: 24274751]
30. Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. Exposing synonymous mutations. Trends Genet. TIG. 2014; 30:308–321. [PubMed: 24954581]
31. Sherry ST, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29:308–311. [PubMed: 11125122]
32. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005; 6:31. [PubMed: 15713233]
33. Gonzalez C, et al. Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. J. Neurosci. Off. J. Soc. Neurosci. 2014; 34:10924–10936.
34. Rooijers K, Loayza-Puch F, Nijtmans LG, Agami R. Ribosome profiling reveals features of normal and disease-associated mitochondrial translation. Nat. Commun. 2013; 4
35. Loayza-Puch F, et al. p53 induces transcriptional and translational programs to suppress cell proliferation and growth. Genome Biol. 2013; 14:R32. [PubMed: 23594524]
36. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat. Protoc. 2012; 7:1534–1550. [PubMed: 22836135]
37. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10:R25. [PubMed: 19261174]
38. Georgi B, Voight BF, Bućan M. From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. PLoS Genet. 2013; 9:e1003484. [PubMed: 23675308]

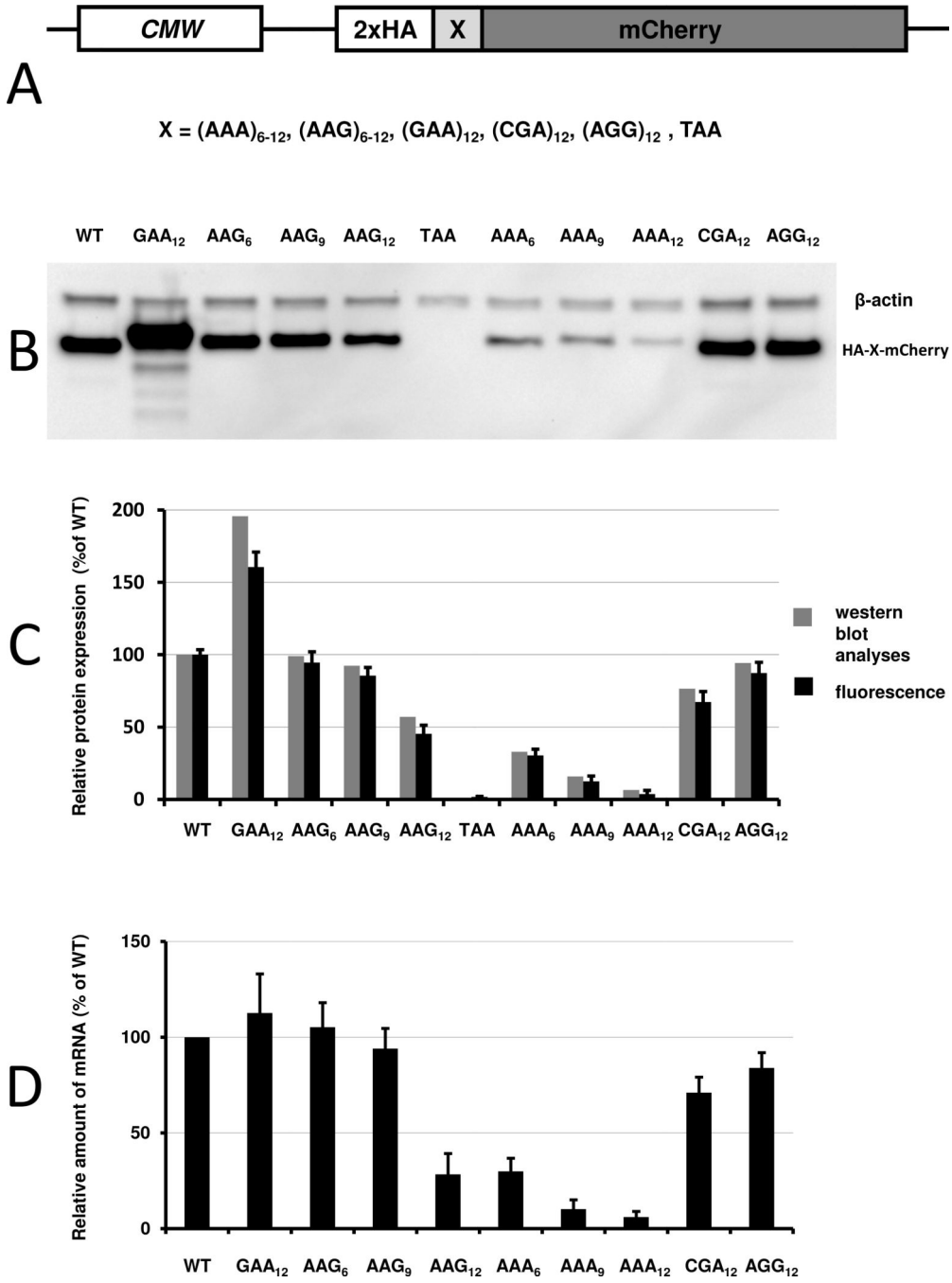

Fig. 1.

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 Licor western blot analyses or *in vivo* mCherry fluorescence measurement; β-actin or fluorescence of co-expressed GFP construct were used for normalization of the data, respectively. Each bar represents 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 percentage of wild type mCherry (WT) mRNA levels.

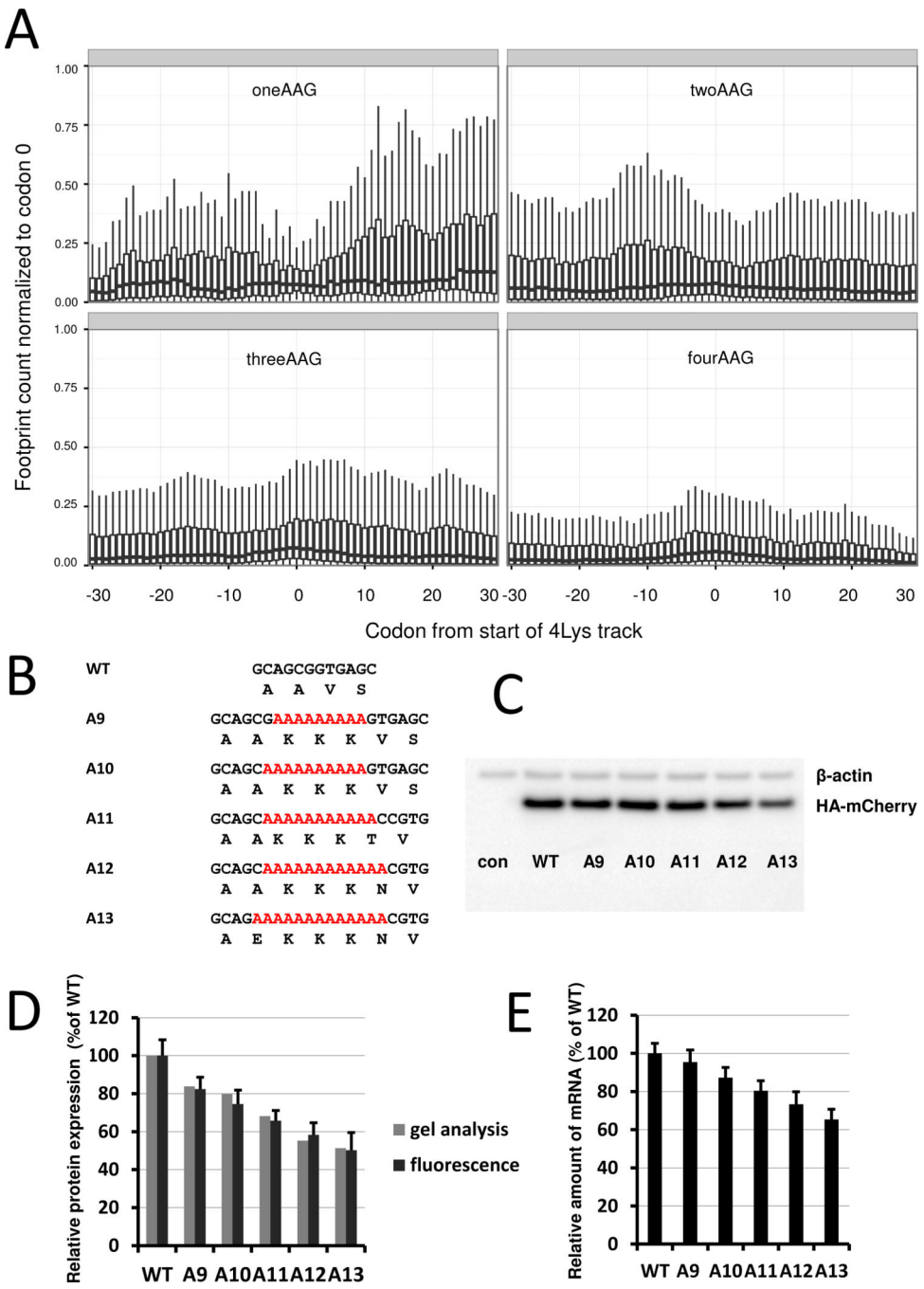

Fig. 2.

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 is 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 maximum of 1.5*IQR 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 Licor western blot analyses or in vivo mCherry fluorescence measurement. β-actin or fluorescence of co-expressed GFP construct were used for normalization of the data, respectively. Each bar represents 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 percentage of wild type mCherry (WT) mRNA levels..

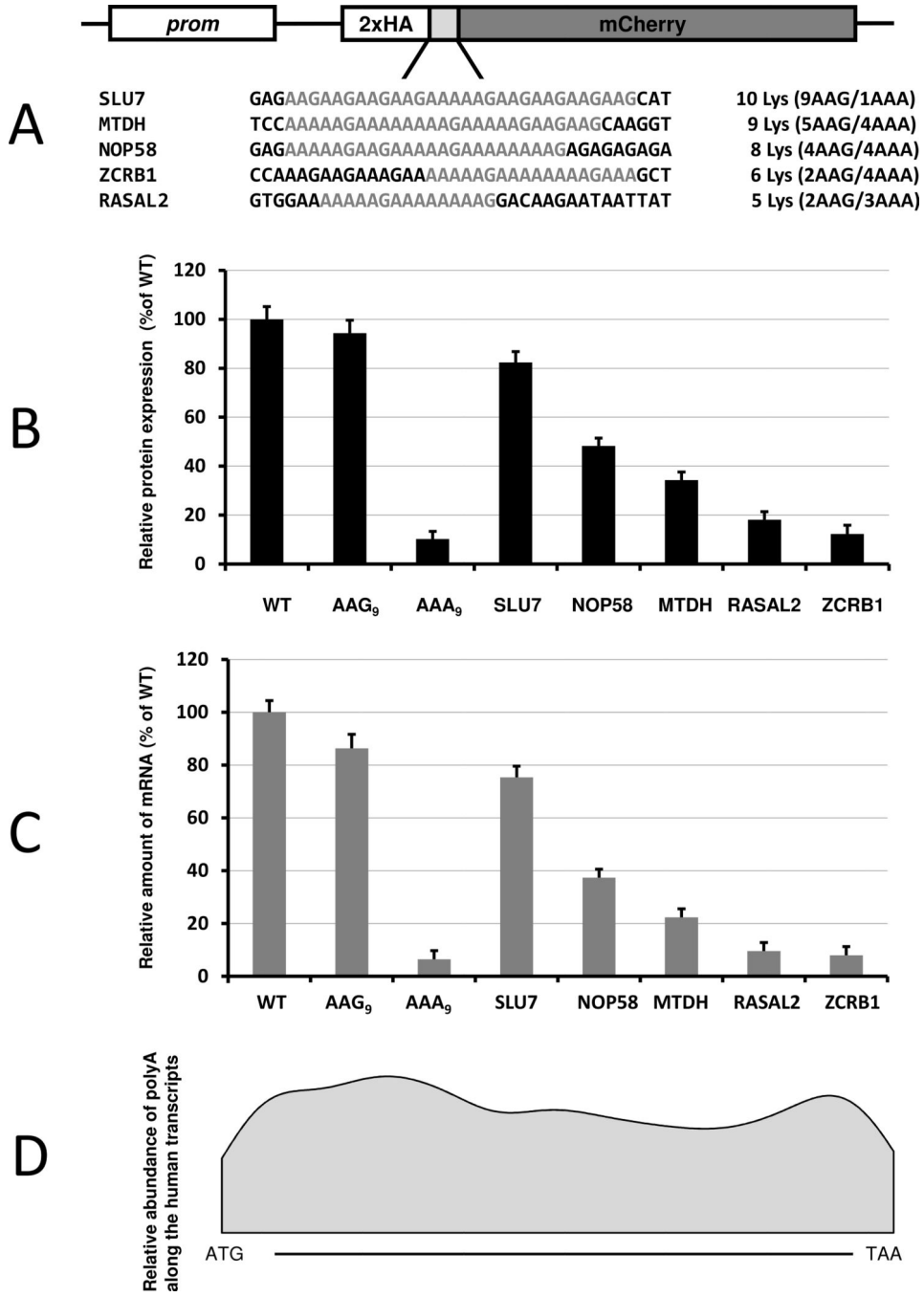

Fig. 3.

A – Sequences of polylysine runs from human genes incorporated into HA-X-mCherry constructs. Continuous runs of lysine residues are labeled. Number of lysine residues and ratio of AAG and AAA codons for each constructs are indicated. B – Normalized protein expression using *in vivo* mCherry reporter fluorescence. Fluorescence of co-transfected GFP was used to normalize the data. Each bar represents 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

percentage of wild type mCherry (WT) mRNA levels. D - Smoothed Gaussian kernel density estimate of positions of polyA tracks along the gene. Position of polyA segment is expressed as a ratio between number of first residue of polyA track and length of a gene.

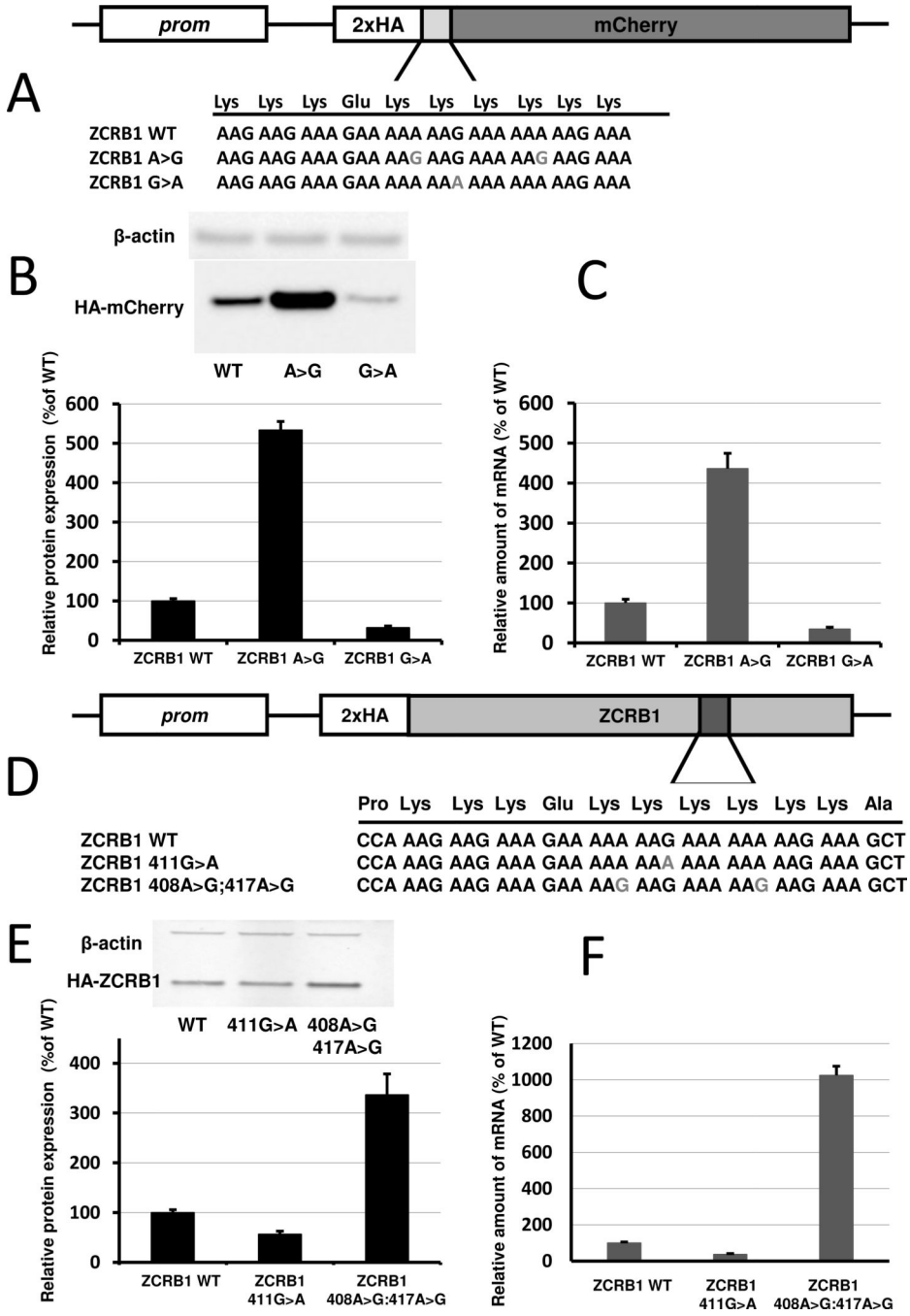

Fig. 4.

A– Scheme of constructs with ZCRB1 gene polyA 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 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 percentage of wild type ZCRB1-mCherry construct (WT) mRNA levels. D – Scheme of full-length HA-tagged

ZCRB gene constructs. Position and mutations in polyA tracks are indicated. E - Western blot analysis and normalized protein expression of ZCRB1 gene constructs with synonymous mutations. Each bar represents 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.

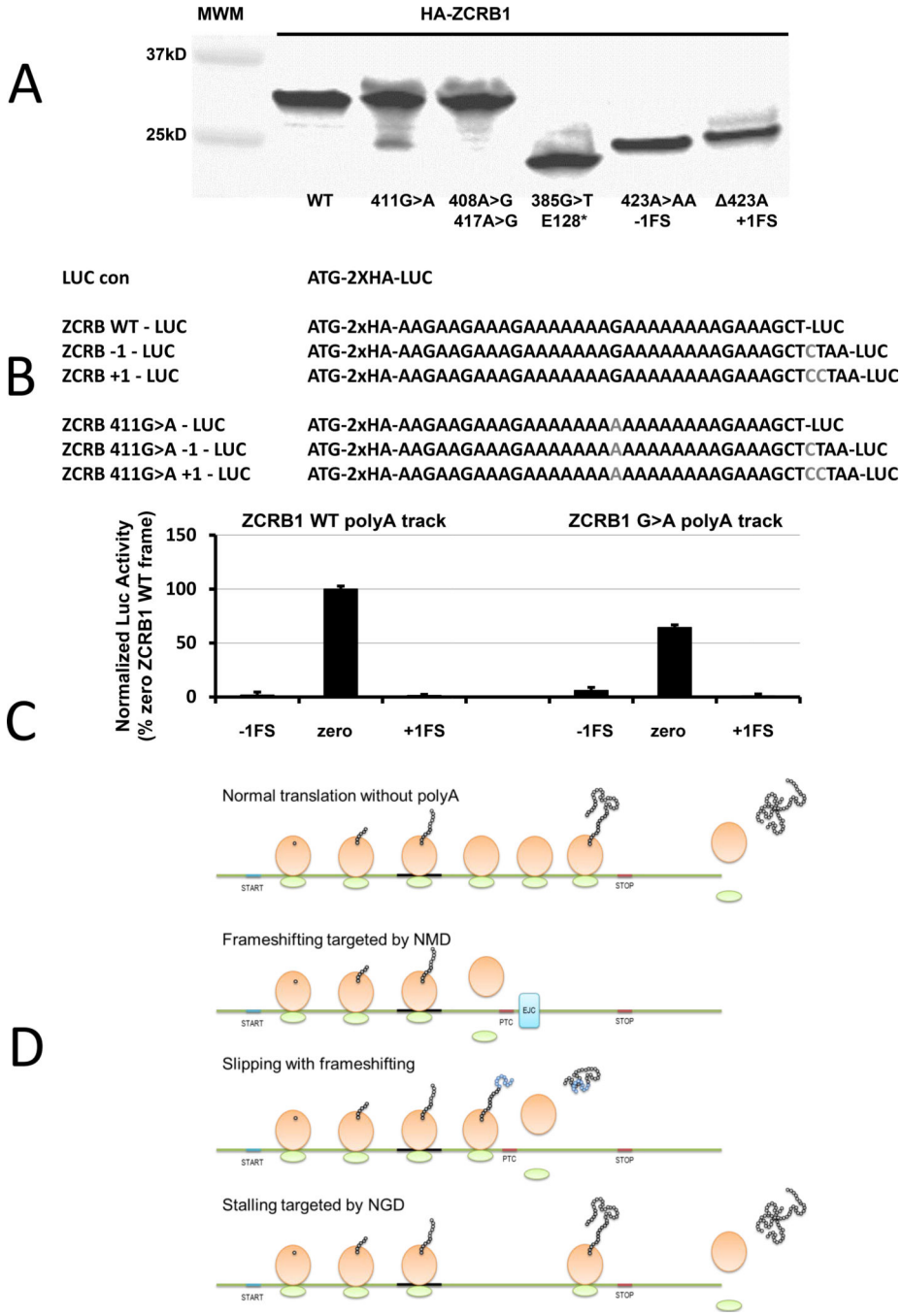

Fig. 5.

A - Immunoprecipitation of HA-ZCRB gene constructs using anti-HA magnetic beads. ZCRB1 WT, synonymous (single 411G>A or double 408A>G; 417A>G), non-sense (385G>T, insertion of stop codon prior poly(A) track), deletion (423ΔA, equivalent to +1 frame-shift) or insertion (423A>AA, equivalent to −1 frame-shift) mutant constructs are labelled respectively. B – Scheme of luciferase constructs used to estimate frame-shifting potential for ZCRB1 WT and 411G>A mutant polyA tracks. C - Luciferase levels (activity) from −1, “zero” and +1 frame constructs of wild type and G>A mutant ZCRB1 polyA track

are compared. Bars represent normalized ratio of ZCRB1 G>A and ZCRB1 WT poly(A) tracks elucidates 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 non-resolved stalling consolidated by endonucleolytic cleavage of mRNA and reduction in wild type protein levels, as in NGD pathway. Scheme for translation of mRNAs without poly(A) tracks is shown for comparison.