Genetics and Genomics
Computational and Systems Biology

Reviewed Preprint
v2 • January 16, 2025
Revised by authors

Reviewed Preprint
v1 • June 18, 2024

Multiplexed Assays of Human Disease-relevant Mutations Reveal UTR Dinucleotide Composition as a Major Determinant of RNA Stability

Jia-Ying Su, Yun-Lin Wang, Yu-Tung Hsieh, Yu-Chi Chang, Cheng-Han Yang, YoonSoon Kang, Yen-Tsung Huang, Chien-Ling Lin

Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan • Institute of Statistical Science, Academia Sinica, Taipei, Taiwan • Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan • Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, Taiwan

eLife Assessment

This **valuable** study combines massively parallel reporter assays and regression analysis to identify sequence features in untranslated regions contributing to the stability of in vitro transcribed mRNA delivered to cells. The strength of evidence presented is **solid**, although some points about half-life measurements and the relevance of identified sequence features to native transcript stability will inform future discussion surrounding the present study. Taken together, the work will be of interest to a broad swath of colleagues studying post-transcriptional gene regulation and especially to those using massively parallel reporter assays.

https://doi.org/10.7554/eLife.97682.2.sa4

Abstract

UTRs contain crucial regulatory elements for RNA stability, translation and localization, so their integrity is indispensable for gene expression. Approximately 3.7% of genetic variants associated with diseases occur in UTRs, yet a comprehensive understanding of UTR variant functions remains limited due to inefficient experimental and computational assessment methods. To systematically evaluate the effects of UTR variants on RNA stability, we established a massively parallel reporter assay on 6,555 UTR variants reported in human disease databases. We examined the RNA degradation patterns mediated by the UTR library in two cell lines, and then applied LASSO regression to model the influential regulators of RNA stability. We found that UA dinucleotides and UA-rich motifs are the most prominent destabilizing element. Gain of UA dinucleotide outlined mutant UTRs with reduced stability. Studies on endogenous transcripts indicate that high UA-dinucleotide ratios in UTRs promote RNA degradation. Conversely, elevated GC content and protein binding on UA dinucleotides protect high-UA RNA from degradation. Further analysis reveals polarized roles of UA-dinucleotide-binding proteins in RNA protection and degradation. Furthermore, the UA-

dinucleotide ratio of both UTRs is a common characteristic of genes in innate immune response pathways, implying a coordinated stability regulation through UTRs at the transcriptomic level. We also demonstrate that stability-altering UTRs are associated with changes in biobank-based health indices, underscoring the importance of precise UTR regulation for wellness. Our study highlights the importance of RNA stability regulation through UTR primary sequences, paving the way for further exploration of their implications in gene networks and precision medicine.

Introduction

Untranslated regions of RNAs are indispensable for post-transcriptional regulation of gene expression

A mature RNA consists of three regions - a 5' untranslated region (5' UTR), the protein coding region, and a 3' untranslated region (3' UTR) (Mignone et al., 2002). UTRs are indispensable for gene expression. For most mRNAs of higher eukaryotes, the 5' UTR is essential for ribosome entry and the 3' UTR is responsible for polyadenylation to stabilize the RNA and enhance its translation efficiency. The average length of the 5' and 3' UTRs in human is ~210 nucleotides (nt) and ~1030 nt, respectively, with mean 3' UTR length being more diverse among species and higher eukaryotes hosting longer 3' UTRs (Pesole et al., 2001). These structural differences among species are consistent with genomic complexity and more complex post-transcriptional regulatory mechanisms. The UTRs contain *cis*-regulatory elements that contribute to post-transcriptionally regulating gene expression, such as via protein translation control, subcellular mRNA localization, and mRNA stability upon interactions with *trans*-acting factors such as RNA-binding proteins (RBPs) and microRNAs (miRNAs) (Barrett et al., 2012).

UTRs control RNA stability

RNA stability regulation serves as an mRNA quality control mechanism (e.g., nonsense- mediated mRNA decay) to gate protein production (Schoenberg & Maquat, 2012). Decades of research have shown that the *cis*-regulatory elements in UTRs affect mRNA stability. These sequence elements control mRNA integrity and can trigger mRNA degradation pathways by interacting with RBPs or small regulatory RNAs such as miRNAs and small interfering RNAs (siRNAs) (Garneau et al., 2007; Mitchell & Tollervey, 2001). For instance, the miRNA- Argonaute complex recruits the CCR4-NOT (Carbon Catabolite Repression—Negative On TATA-less) complex to initiate deadenylation and decay (Huntzinger & Izaurralde, 2011). Another example is AU-rich elements (AREs, sequence elements rich in adenosine and uridine) present in the 3' UTRs of many mRNAs that provide binding sites for ARE-binding proteins that trigger the RNA degradation pathway (Garneau et al., 2007). Many ARE-binding proteins have been characterized to date that are involved in stability regulation of ARE-hosting mRNAs (Barreau et al., 2005), including Tristetrapolin (TTP), Butyrate Response Factor 1 (BRF1), Heterogeneous Nuclear Ribonucleoprotein (hnRNP) D (also known as AU-Rich Element- Binding Protein 1, AUF1), and KH-type splicing regulatory protein (KSRP), all of which destabilize mRNA, unlike ELAV Like RNA Binding Protein 1 (ELAVL1, also known as HuR) that stabilizes it (Schoenberg & Maquat, 2012). This regulatory mechanism necessitates physical access to the sequence elements, so structural contexts are critical (Paschoud et al., 2006). Complex secondary structures such as RNA G-quadruplexes (RG4s) and pseudoknots may also play a role in stability regulation. RG4s are enriched in UTRs where they regulate many post- transcriptional regulatory processes, including RNA stability (Dumas et al., 2021), though the detailed mechanisms remain to be elucidated. Although 3' UTR-mediated regulation gains more attention, 5' UTRs may also contribute to mRNA stability regulation. For instance, upstream open reading frames (uORF) in 5' UTRs can facilitate RNA decay in a translation- dependent manner, and RG4s in 5' UTRs reduce RNA stability in a ribosome-independent manner (Jia et al., 2020).

Multiplexed reporter assays to elucidate UTR stability control

Efforts have been made to elucidate RNA stability regulation on a genome-wide scale to understand its general regulatory mechanisms. However, due to the complexity of post-transcriptional regulation, features inferred from endogenous steady-state RNA levels can be obscured by other dominant factors. For instance, studies examining cellular RNA degradation alongside transcriptional inhibition have shown that coding region length and ribosome occupancy are key determinants of RNA stability (Benjamin Neymotin, 2015). Additionally, RNA stability inferred from steady-state RNA concentrations normalized against transcription rates has indicated that splice junction density is a major factor promoting RNA stability (Agarwal & Kelley, 2022; Blumberg et al., 2021). Nonetheless, these studies have not systematically decoded the influence of primary sequences on RNA stability regulation. Therefore, efforts have been made to establish massively parallel reporter assays for bulk- synthesized UTRs to unveil their governance of RNA regulation in various species. Bidirectional promoters driving a control transcript and a green fluorescence protein (GFP) hosting various test UTRs were first used to evaluate the effect of the UTRs on fluorescence signals (Oikonomou et al., 2014; Sample et al., 2019; Slutskin et al., 2018; Wissink et al., 2016; Zhao et al., 2014). Alternatively, various UTRs have been inserted into plasmids prior to cellular expression, and then the DNA and RNA levels of each construct have been compared to infer the effect of the UTRs on RNA expression (Griesemer et al., 2021; Litterman et al., 2019; Siegel et al., 2022). Nevertheless, because the steady-state level is a result of production and decay, these approaches cannot differentiate the effect of the UTRs on transcription, stability and, in some cases, even protein production, greatly limiting scientific interpretation. Injections or transfections of *in vitro*-transcribed RNAs have been used to study RNA stability. These studies have identified AREs and miRNA-binding sites as destabilizing elements and U-rich sequences as stabilizing elements in zebrafish embryos (Rabani et al., 2017; Vejnar et al., 2019). Similar studies in human cell lines have shown that RG4 structures and A-rich sequences in 5' UTRs promote RNA decay (Jia et al., 2020).

UTR mutations and disease

UTR sequence variation affects mRNA stability, translation and localization. RNA dysregulation arising from mutations in UTRs significantly and negatively affects gene regulation, which can promote phenotypical and even pathological change. According to the NHGRI-EBI GWAS Catalog, genome-wide association studies up to 2018 had uncovered that ~3.7% of disease risk/quantitative trait-associated genetic variants are located in UTRs (MacArthur et al., 2017; Steri et al., 2018). Indeed, certain studies have provided evidence that alterations to even a single nucleotide in a UTR can impact mRNA translation or transcript half-life in disease contexts. For example, a single nucleotide substitution of the 36th position in the 5' UTR of *transforming growth factor-β3* (*TGFβ3*; OMIM# 190230) or its 1723th 3' UTR position is associated with arrhythmogenic right ventricular cardiomyopathy (Beffagna et al., 2005). Similarly, point mutation of the *GFPT1* 3' UTR results in congenital myasthenic syndrome. GFPT1 (Glutamine-Fructose-6-Phosphate Transaminase 1) is the rate-limiting enzyme for hexosamine biosynthesis, and mutation of its 3' UTR results in a 90% reduction in protein production, potentially due to gain of a miRNA binding site (Dusl et al., 2015). Collectively, these studies support that the precise RNA regulation exerted by UTRs plays a critical role in controlling gene expression, with UTR mutations potentially eliciting divergent phenotypes and even severe disease.

Despite their disease relevance, a comprehensive overview of the pathogenic effects of UTR mutations is still lacking. The medical genetics community has earnestly advocated for consideration of "UTR variants in genetic diagnostic procedures" (Dusl et al., 2015). Specifically, the impact of 5' and 3' UTR sequence variations on RNA stability regulation remains unclear. While there have been efforts to examine how 3' UTR variants affect steady- state RNA levels, systematic assessments of the explicit effects of disease-relevant variants in both UTRs on stability regulation are still absent. To examine potentially pathogenic UTR mutations and their links to

RNA stability, we developed a massively parallel reporter assay in which human 5'/3' UTRs with disease-relevant mutations were generated *in vitro*, ligated with the enhanced green fluorescence protein (EGFP) coding region, and then directly transfected into human cell lines to assess their decay patterns by next-generation sequencing. Taking redundancy and interdependency of regulatory features into consideration, our approach identified that UA dinucleotides are the most influential destabilizing sequence element for RNA stability. Moreover, we found that joint regulation by 5' and 3' UTRs shapes the expression kinetics of functional gene groups. Our study unveils RNA stability determinants and delineates the importance of precise UTR control in maintaining harmonious genetic networks for human health.

Results

Massively parallel reporter assay (MPRA) for RNA stability

Our above-described genome-wide analysis prompted the hypothesis that UTR variants that disrupt critical RNA regulatory elements may be linked to pathogenicity. Since one of the major roles of UTRs is to control RNA stability, we hypothesized that disease-relevant UTR variants may alter RNA stability. Therefore, we designed 6,555 pairs of 155-nt UTR fragments centering on the variant collected from the HGMD and ClinVar disease databases, and performed time-course assays to examine the relative stability. First, we fused UTRs to the EGFP coding regions, transcribed them *in vitro*, and then transfected them into human embryonic kidney cells (HEK293T) or neuroblastoma cells (SH-SY5Y), considering pervasive neurological diseases in the mutation collection. Then, we monitored the relative abundance of the reference (ref) and mutant (mt) alleles by amplicon sequencing over a time course (30, 75, 120 min for HEK293T; 20, 40, 60 min for SH-SY5Y) (**Fig. 1A**). Primers targeting common reporter regions were utilized for the retrieval of UTR sequences at each time point. We estimated the decay constant and half-life (t1/2) of each UTR according to its relative abundance over time (see Methods). We defined stability-altering variants as those for which the decay constants significantly changed relative to their ref counterparts, as determined by weighted linear regression (see Methods). We observed that variants in both 5' and 3' UTRs significantly altered RNA half-life, with slightly more variants having a negative impact on RNA stability (**Fig. 1B&C**; **Supplemental Table S1**).

Our results from three independent experiments are highly consistent, with a Spearman's correlation coefficient > 0.93 (Supplemental Fig. S1A,B). From among 3,700 pairs of valid comparisons, 40 (1.1%) and 839 (22.8%) variants displayed significantly altered stability compared to their ref counterparts in HEK293T and SH-SY5Y cells, respectively (**Fig. 1B&C**). Thus, we observed a significant effect of UTR variation on RNA stability, but their regulatory impact was strikingly divergent between the two tested cell lines. We attribute this divergence to potential differences in translation capacity, as well as variations in the composition and concentration of RNA-binding proteins and RNases between the two cell lines.

Impact of bi-functional AREs on RNA stability

AREs are well-recognized regulatory motifs controlling RNA stability. Accordingly, we examined if the regulatory effect of AREs could be captured by our MPRA approach. Multiple approaches have revealed AREs as exerting a destabilizing effect on RNA stability (Barreau et al., 2005). However, ARE motifs and ARE-binding proteins are diverse, so the impact of binding may vary considerably. Therefore, we examined the effect of AREs on RNA stability of the ref alleles according to specific sequence content. Based on the definition of AREsite2 (*http://nibiru.tbi.univie.ac.at/AREsite2*), we categorized AREs as either WUUUW or its longer derivatives, UUUGUUU or AWUAAA (W:A/G) (Fallmann et al., 2016). We observed that AREs in either the 5' or 3' UTRs generally destabilized RNA (**Fig. 2A**; Supplemental Fig. S2; Table S2). More specifically, AUUUA/AUUA-containing AREs are associated with RNA destabilization when present in either UTR type, whereas in SH-SY5Y cells, extremely U-dominant AREs (U8- 10A1-2) stabilized it (**Fig. 2B**), similar to the stabilizing

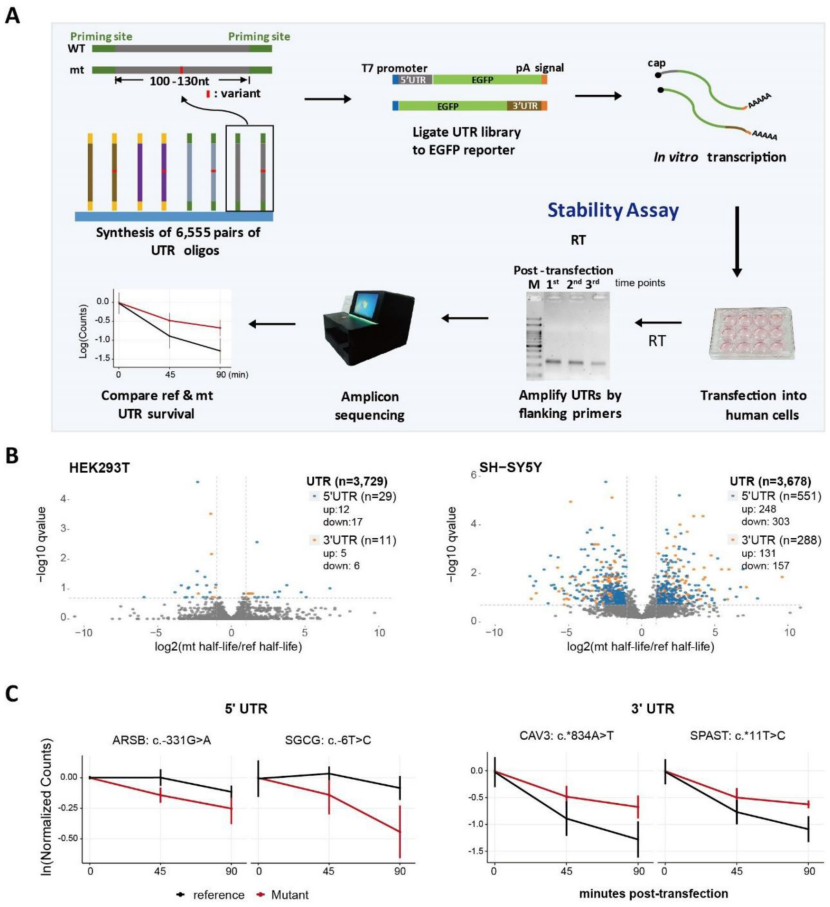


Figure 1.

Massively parallel reporter assay (MPRA) to determine the effects of UTR variants on RNA stability.

(A) MPRA workflow. In brief, 6,555 reference (ref) and mutant (mt) UTR pairs were synthesized in bulk, ligated with promoters and reporter sequences, *in vitro*-transcribed into capped and tailed RNAs, transfected into human cell lines, and then the remaining RNAs were collected over a time-course. The collected RNAs were reverse-transcribed, amplified and sequenced to resolve the genotype of each UTR. The unique sequences were used to calculate RNA half-life. Mutational effects were inferred from those pairs significantly differing in half-life (see Methods). **(B)** Volcano plot of MPRA data from three repeated experiments. The colored dots indicate significant stability-altering variants. **(C)** Examples of significant stability-altering UTR mutations in both UTR types.

effect of U-stretches described for zebrafish (Rabani et al., 2017; Vejnar et al., 2019). These results suggest that although mostly destabilizing, AREs can play dual roles in regulating RNA stability by recruiting binding proteins of diverse functions.

Modeling the impact of UTR-mediated regulation on RNA stability

Given our discovery that the effect of AREs is heavily dependent on sequence content, we decided to further explore the effects of other sequence elements, i.e., beyond known regulatory motifs, in more detail. Since most reported RBP motifs are 6-mers, we initiated a search for novel motifs by analyzing the presence of all 7-mers in our massively parallel reporter assay (MPRA) library, correlating their occurrence with mRNA half-life. For those with significant stabilizing or destabilizing effects, we clustered similar ones into motifs (Supplemental Fig. S3). The motifs suggest a G-rich stabilizing profile and an A-rich destabilizing profile, with the latter being more pronounced for the 3' UTR. Next, to gain a comprehensive understanding of the contextual effect of each sequence element, we took advantage of LASSO regression, which minimizes coefficients of explanatory factors to select the most influential factors. We considered as many factors as possible to explain the half-life of our ref UTR libraries, including primary sequences, RBP binding sites (ATtRACT database (Giudice et al., 2016)), miRNA seed sites, secondary structures, and folding energy. Furthermore, to avoid collinearity confounding our model, e.g., the effects of very similar factors (such as 'AA' and 'AAA' sequences), we clustered the factors according to their properties, and then only one representative factor from within a cluster (i.e., the one with the highest correlation to half-life within a cluster) was subjected to LASSO regression (**Fig. 3A**, Supplemental Fig. S4, Table S3 and Methods). LASSO regression renders as zero the coefficients of factors with minimal explanatory power (see Methods for details). Overall, we started with 1,231 (5' UTR) or 1,475 (3' UTR) factors, but only 5-19 factors were selected ultimately for each trained model (**Fig. 3B-E**; Supplemental Table S4). The selected explanatory factors all represent small kmer motifs (k=2 to 3) and RBP binding motifs. Unexpectedly, we identified some unique regulatory factors in each cell line, indicating that RNA decay pathways are typically shared but can be strongly influenced by the cellular environment. Overall, motifs that are at least two nucleotides long proved critical for RNA stability, supporting the sequence specificity of the decay process.

In both of the cell lines we tested, GU-rich sequences in 5' UTRs stabilized RNAs (**Fig. 3B,D**). In contrast, CA- and UG-repeat sequences—potential binding sites for Insulin Like Growth Factor 2 mRNA Binding Protein 3 (IGF2BP3) and CUGBP Elav-Like Family Member 1 (CELF1)—in 3' UTRs proved the most destabilizing factors in HEK293T cells (**Fig. 3C**). Moreover, we noticed that for both UTRs, UA dinucleotides and other UA-rich sequences—such as WWWWWW (W=A/U; a potential binding motif of Peptidylprolyl Isomerase E (PPIE)), AUUUA (a potential binding motif of ELAV Like RNA Binding Protein 1 (ELAVL1)), and UUUAUA (a potential binding motif of hnRNPA1)—are strongly destabilizing (**Figs. 3B-E** & **4A,C**). UA dinucleotides and WWWWWW belong to the same cluster, but they were respectively selected for LASSO regression in the two cell lines because they displayed the highest explanatory power (largest coefficient by univariate regression) for RNA half-life in each cell line (**Figs. 3A** & **4B,D**). Most prominently, UA dinucleotides in both UTRs overwhelmed other factors in robustly destabilizing RNAs in SH-SY5Y cells (**Fig. 3D,E**). Therefore, UA dinucleotides seem to be a universal destabilizing motif, so we investigated how UA dinucleotides regulate RNA stability in further detail.

UA dinucleotides and UA-rich motifs are the most common and effective RNA destabilizing factor

UA dinucleotides proved to be the strongest stability determinant for both UTR types in SH- SY5Y cells. UA dinucleotides alone present a negative correlation with RNA stability, with a Pearson's correlation coefficient of -0.287 for 5' UTRs and -0.377 for 3' UTRs (**Fig. 4A,C**). UA- rich motifs (in the same cluster as UA dinucleotides) behave similarly to UA dinucleotides in regulating RNA

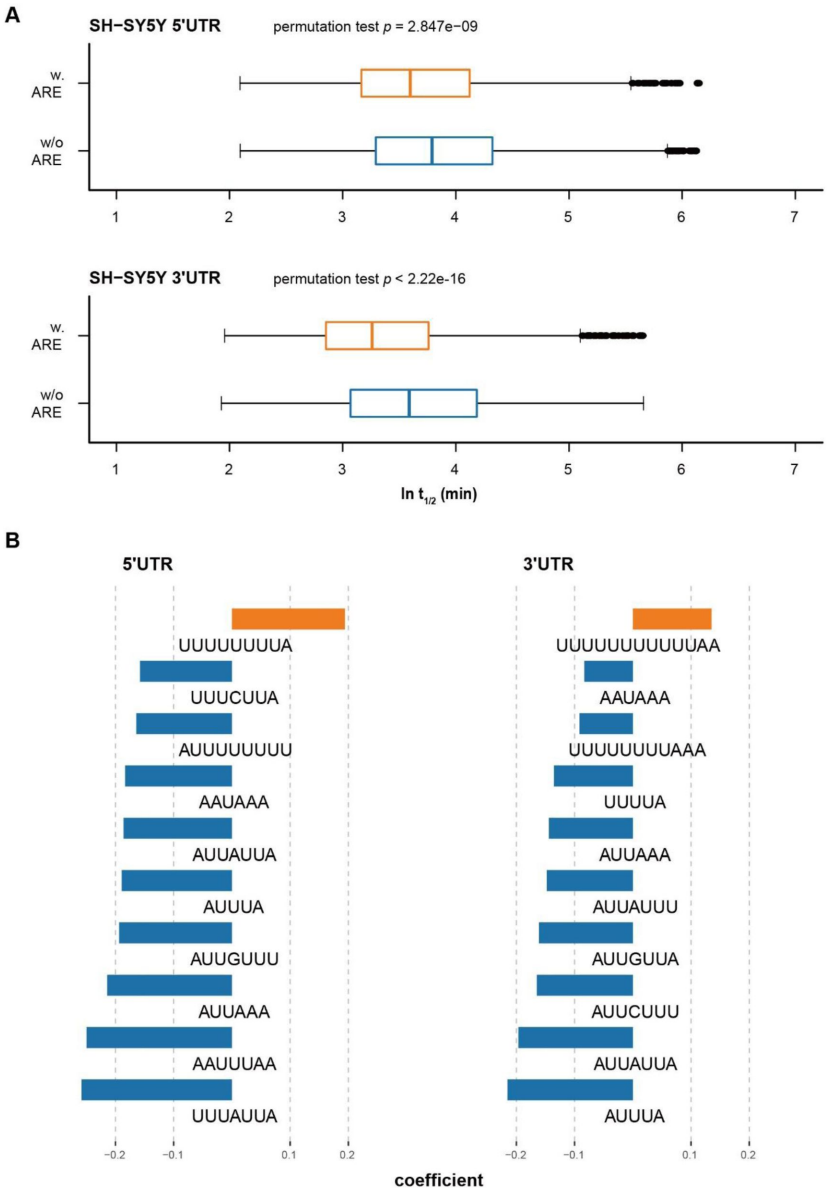


Figure 2.

AREs generally destabilize RNAs (except extremely U-rich AREs).

(A) AREs of both UTR types destabilize RNA. **(B)** The ten most influential AREs in terms of RNA stability in SH- SY5Y cells. Coefficients are determined by regression analysis, representing the effect size of each motif.

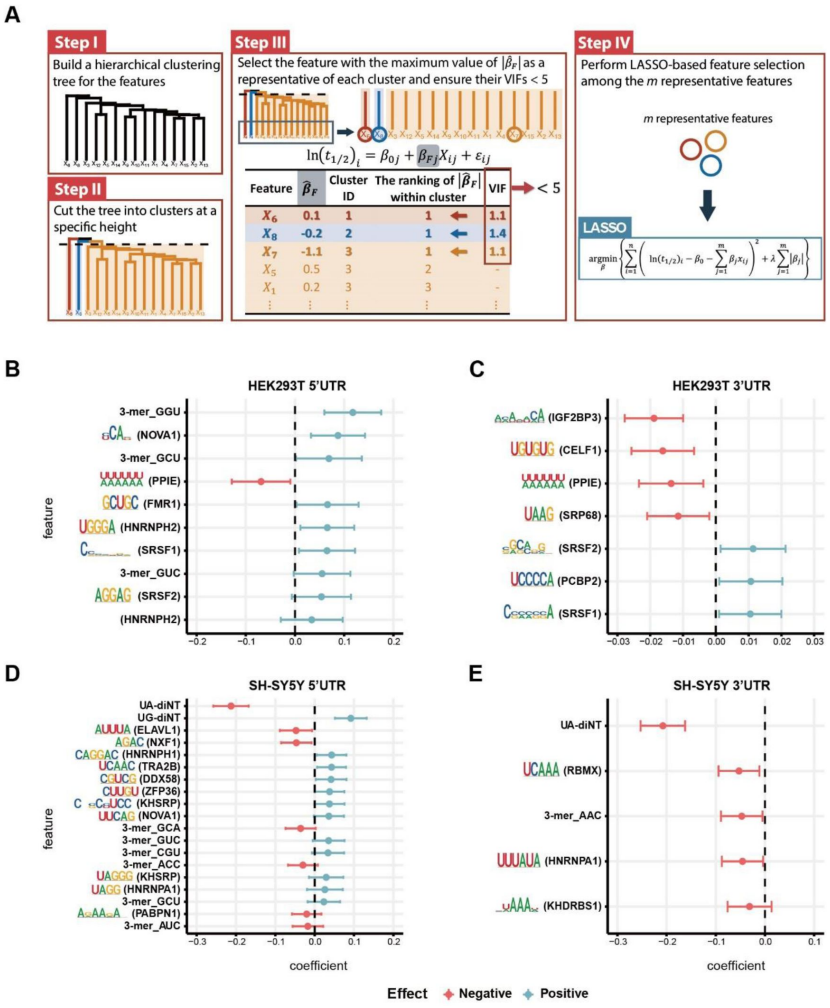


Figure 3.

Inferential statistical analysis of RNA stability determinants.

(A) Workflow of variable selection to build models of influencers of RNA stability. **(B-E)** Influential regulators for the 5′ UTR library from HEK293T cells (B), the 3′ UTR library from HEK293T cells (C), the 5′ UTR library from SH-SY5Y cells (D), and the 3′ UTR library from SH-SY5Y cells (E). The error bars represent 95% confidence intervals of the coefficients. Note that the factors presented on the figure are representative of their respective clusters (see Methods and Supplemental Fig. S3).

stability, whereas GC-rich motifs have the opposite effect (**Fig. 4B,D**; Supplemental Fig. S5A,B). Within the same cluster, UA dinucleotides and the WWWWWW motif were the strongest RNA stability regulators in each cell line (Supplemental Table S3). Given the strong destabilizing effect of factors in the UA-associated cluster for both UTR types and both cell lines, we further analyzed their commonalities. An UpSet analysis revealed that all features contributing to RNA stability across four experimental groups (HEK293T 5' UTRs, HEK293T 3' UTRs, SH-SY5Y 5' UTRs, SH-SY5Y 3' UTRs) occur in the UA dinucleotide/WWWWWW cluster (Supplemental Fig. S5C), indicating a universal destabilizing effect of UA-rich sequences. Next, to examine if there is a region-specific effect of UA and closely-related AU dinucleotides, we used a sliding window to establish the localization- associated relationship between the UA/AU dinucleotide ratio and RNA half-life. Correlation coefficients between UA/AU dinucleotide ratios and UTR stability were calculated for each window, and we assumed that regions displaying a strong correlation between UA/AU dinucleotide ratios and stability rank hosted UA/AU dinucleotides that control RNA stability (Supplemental Fig. S5D). We found that UA/AU dinucleotides in the UTRs of SH-SY5Y cells were generally strongly correlated with RNA stability, but only weakly associated with RNA stability in HEK293T cells (apart from a relatively strong correlation at the ends of 3' UTRs, implying a protective role against exonuclease digestion) (Supplemental Fig. S5E). Together, these results support that UA dinucleotides are a common and prominent RNA destabilizing motif.

Next, we examined the effect of mutating the most effective destabilizing UA dinucleotide (resulting in dinucleotide gain or loss) in terms of altering RNA stability. We found a clear propensity for 3' UTRs with stability loss to accumulate gain of UA dinucleotide mutations, compared to stabilizing or non-significant mutations (**Fig. 4E**). To further validate the impact of UA dinucleotides, we curated a subset of oligo pairs from a 5' UTR random library (Jia et al., 2020) with the sole difference being the presence of one additional UA dinucleotide, resulting in a discrepancy of one UA dinucleotide between them. This selection allowed us to investigate whether these variations influenced RNA half-lives. We defined significant pairs as those with half-life differences greater than or equal to a 1.5-fold change. Notably, the acquisition of an additional UA dinucleotide resulted in the destabilization of RNAs (**Fig. 4F**, Left: 0 to 1 UA dinucleotide; Right: 1 to 2 UA dinucleotides). Moreover, to explore the influence of the UA-destabilizing effect on endogenous mRNA stability, we assessed the UA- dinucleotide ratio in relation to RNA half-life of HEK293 and human erythroleukemia K562 cells. While exon junction density and transcript length have been suggested as major determinants of RNA half-life *in vivo* (Agarwal & Kelley, 2022; Blumberg et al., 2021), our findings indicate that elevated UA-dinucleotide ratios in both UTRs significantly promote RNA decay (**Fig. 4G** and Supplemental Fig. S5F). This effect is specific, as such ratios in the coding region are inconsequential. Thus, we have identified UA dinucleotides as being the strongest RNA destabilizing feature in UTRs, with their mutational addition proving the most prevalent cause of reduced RNA stability.

Intrinsic features of UTRs determine RNA stability

Apart from their shared regulatory mechanisms, our MPRA revealed distinct UTR behaviors. Despite undergoing exactly the same procedure, library RNAs in SH-SY5Y cells degraded much faster than those in HEK293T cells (Supplemental Fig. S1D). Distinct RNA stability control in neuronal cells by GC content of transcripts has been reported (Guvenek et al., 2022). Additionally, cell-specific contexts, such as expression of RBPs and miRNAs, may have intensified the discrepancy in stability control between the two cell lines. Moreover, for mutant UTRs that significantly destabilized or stabilized RNAs, we found that their ref counterparts are with significantly longer or shorter RNA half-life, respectively (Supplemental Fig. S6A). This intrinsic difference in effects on RNA stability for ref UTRs was observed for both cell lines and it was amplified in their mutant counterparts (Supplemental Fig. S6B,C). A motif analysis revealed that ref UTRs whose mutant counterparts significantly altered RNA stability tended to harbor more NOVA1 (NOVA Alternative Splicing Regulator 1) and PPIE binding sites, but fewer FMR1 (fragile X mental retardation 1) binding sites (Supplemental Fig. S6D). These observations indicate that intrinsic properties of the UTRs where mutations occur greatly influence the mutational effect.

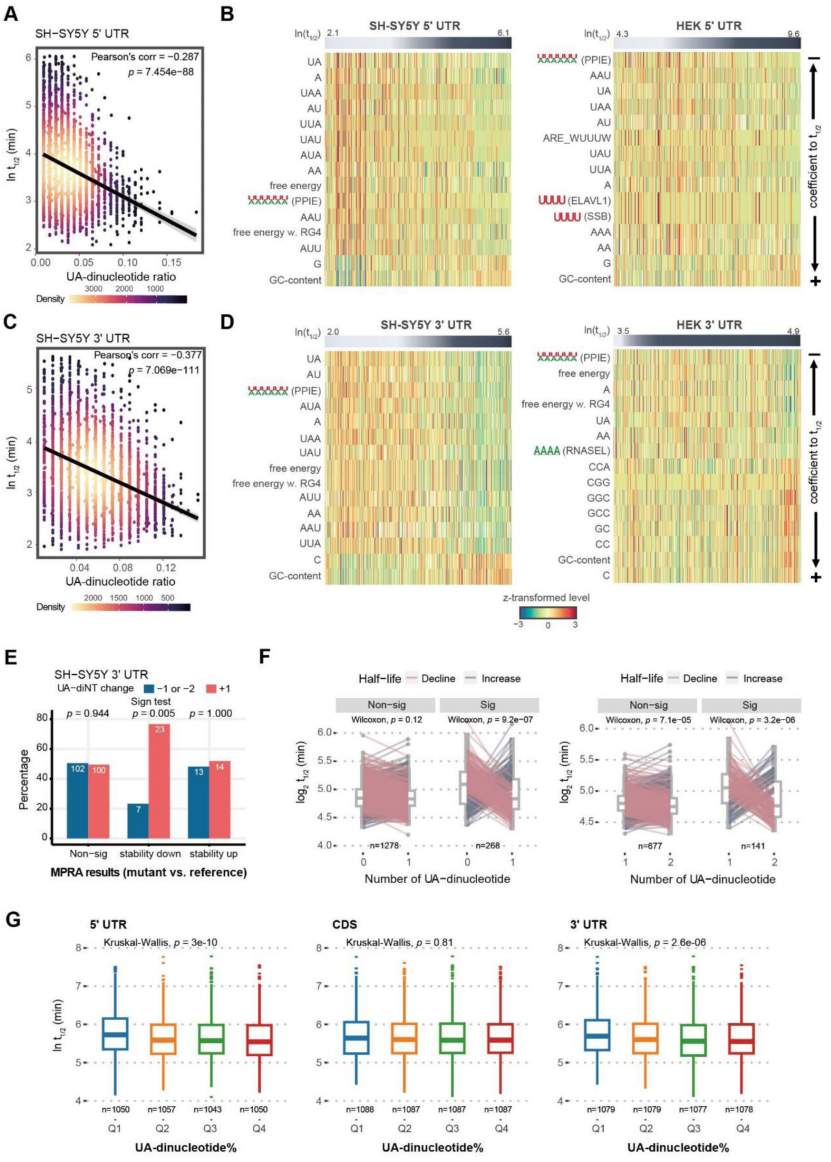


Figure 4.

The UTR UA dinucleotides and UA-rich motifs are the most common and influential RNA destabilizing factor.

(A) Correlation of the 5′ UTR UA dinucleotide ratio and half-life. **(B)** Top 15 influential factors in the UA cluster of 5′ UTR. UTRs are arranged by half-life, and factors by their coefficient to half-life. Note that there are destabilizing factors (such as UA and AU dinucleotides) as well as stabilizing factors (such as GC content and G monomers) in this cluster. UA dinucleotide and WWWWWW (PPIE) (where W represents A/U) are representative of the cluster for modeling UTR stability in SH-SY5Y and HEK293T cells, respectively. **(C)** Correlation of the 3′ UTR UA dinucleotide ratio and half-life. **(D)** Top 15 influential factors in the UA cluster of 3′ UTR. **(E)** Mutational gain of a UA dinucleotide by 3′ UTRs significantly reduces RNA stability (lower panel). **(F)** Gain of UA dinucleotides in a random 5′ UTR library led to RNA destabilization. We categorized pairs with a ≥ 1.5-fold change as 'significant' (Sig) and those with less than this threshold as 'non-significant' (Non-sig). **(G)** High UA-nucleotide ratios of both UTRs reduce endogenous RNA stability in HEK293 cells. Q1-Q4 denote quantile groups categorized based on the UA-dinucleotide ratio.

GC content RBP- and ribosome-binding hinders the destabilizing effect of UA dinucleotides

Many previous studies have reported GC content to be a major RNA stability determinant (Courel et al., 2019; Litterman et al., 2019; Zhao et al., 2014). A univariate regression analysis on our UTR libraries revealed that both GC content and UA dinucleotides are strongly associated with RNA half-life, with the 3' UTR library from SH-SY5Y cells displaying the strongest association. However, overall, UA dinucleotides are more strongly correlated with RNA half-life than GC content (**Fig. 4B,D**; Supplemental Fig. S7A). We reasoned that GC content and UA dinucleotides may represent confounding factors in our models. Therefore, to further dissect their respective contributions to RNA stability, we examined their relationship by regressing both factors as well as their interaction term against RNA half-life (t1/2 ~ UA-diNT% + GC% + UA-diNT% × GC%). For both 5' UTRs and 3' UTRs, we observed that UA dinucleotides exhibited a stronger association (i.e., smaller p values) with RNA half-life than GC content. Indeed, the link between GC content and RNA half-life became non- significant when we also accounted for UA dinucleotides (Supplemental Fig. S7A). Moreover, the interaction term (UA-diNT% × GC%) proved significant for three out of four experimental groups (SH-SY5Y 5' UTR: p = 0.18, 3' UTR: p = 0.00047, HEK293T 5' UTR: p = 0.0062, 3' UTR: p = 1.2e-7), supporting that GC content influences the effect of UA dinucleotides. To further explore this interaction, we stratified degrees of GC content and examined their effects on UA dinucleotides. Notably, despite a somewhat anti-correlation between UA dinucleotides and GC content, they do not simply oppose each other. Substantial counts of UA dinucleotides are observed in regions characterized by high GC content. We found that UA dinucleotides strongly destabilized RNA in the context of low GC content (bottom 50%), but their effects were neutralized somewhat under scenarios of high GC content (top 50%) (**Fig. 5A**). The GC protective effect also supports the observation that change of UA dinucleotides in high GC-content 5' UTR did not always translate into a change of stability (**Fig. 4F**). Conversely, the protective effect of GC content was remarkably strong for high UA-dinucleotide ratios but was barely detectable for low UA-dinucleotide ratios (Supplemental Fig. S7B). Thus, the RNA destabilizing effect is most pronounced under conditions of a low UA-dinucleotide ratio and low GC content (Supplemental Fig. S7C). Moreover, the stabilizing or destabilizing effects of UA deletion or addition, respectively, were only observed for low GC content, further supporting that high GC content hinders the destabilizing effect of UA dinucleotides (**Fig. 5B**, Supplemental Fig. S7D).

To corroborate the interplay between the UA-dinucleotide ratio and GC content, we investigated their combined impact on RNA stability using a fixed-length 5' UTR random library (Jia et al., 2020). Our findings reveal a negative association between the UA-dinucleotide count within the library and RNA half-life, with this effect being attenuated under conditions of high GC content (top 50%) (**Fig. 5C**). As another layer of validation, we further explored their relative contributions to mRNA enrichment in P-bodies (Courel et al., 2019). P-bodies are membraneless granules for RNA storage and turnover (Beadle et al., 2023). We uncovered a positive correlation between UA dinucleotides and RNA P-body localization, especially for those occurring in 3' UTRs (**Fig. 5D**). For both UTR types, we observed a greater UA dinucleotide-stabilizing effect in P-bodies when GC content is low (**Fig. 5D**), consistent with our MPRA dataset.

We hypothesized that the protective effect of GC content arises from extensive intramolecular interactions that shield UA dinucleotides from being recognized by nucleases, though other physical hindrances may also diminish the destabilizing effect of UA dinucleotides. Therefore, we examined the influence of RBP binding on the destabilizing effect of UA dinucleotides. In **Figures 5F and G**, we calculated the number of RBP species binding to UA dinucleotides using both experimental data (enhanced crosslinking immunoprecipitation or eCLIP) and predicted RBP-binding motifs (ATtRACT). We then organized our MPRA-derived results based on whether they exhibited a high (top 50%) or low (bottom 50%) number of RBP binding partners per UA dinucleotide. In doing so, we revealed that the RNA group with multiple binding partners indeed

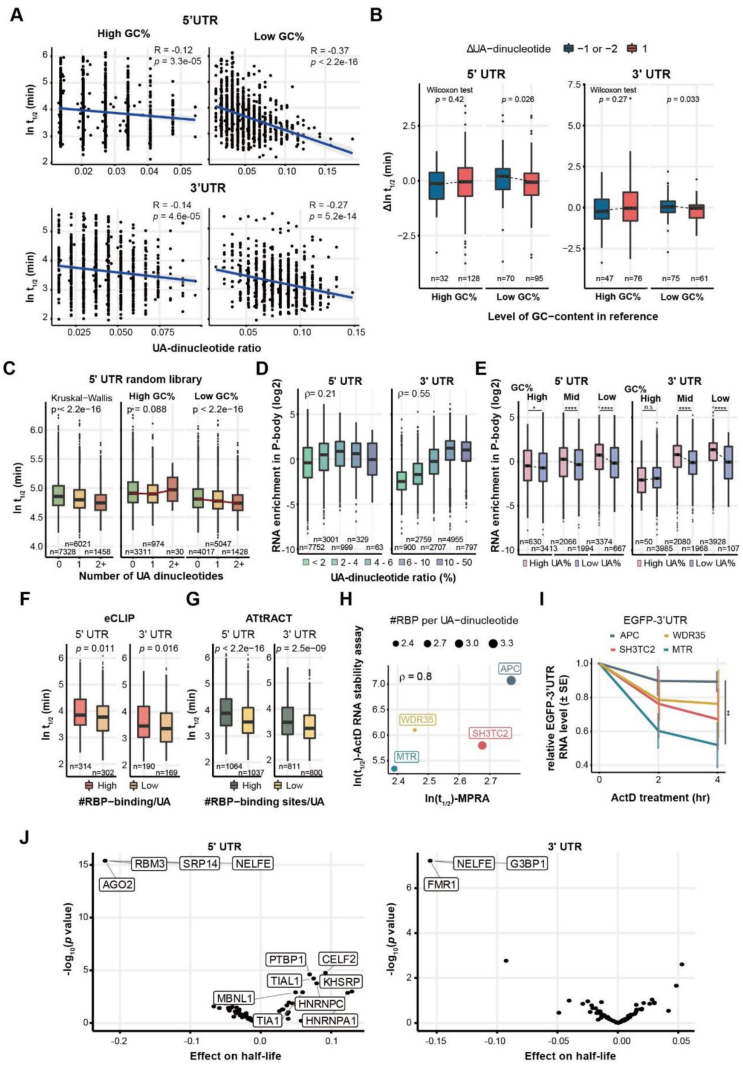


Figure 5.

GC content, RBP and ribosome binding shields RNA from the destabilizing effect of UA dinucleotides.

(A) MPRA data of SH-SY5Y cells stratified according to the GC content (GC%) of UTRs. The data was divided into high and low groups according to the median of GC%. In both UTRs, the destabilizing effect of UA dinucleotides is more evident in the context of low GC content (right panels). *P* values were determined by linear regression. **(B)** High GC content hinders the effect of altered UA dinucleotides in mutant UTRs. The destabilizing effect of UA- addition (blue) and the stabilizing effect of UA-deletion (crimson) are only observed under the condition of low GC content. *P* values were determined by a two-sided Wilcoxon rank sum test. **(C)** Destabilizing effect of UA dinucleotide is observed with 5′ UTR random library. High GC content hinders the UA-destabilizing effect. **(D)** UA dinucleotides are enriched in P-body- resident mRNAs. *ρ* represents Spearman's correlation coefficient. **(E)** High GC content inhibits enrichment of UA dinucleotide-hosting mRNAs in P-bodies. For medium or low GC%, a high UA-dinucleotide ratio promoted the P-body localization of mRNAs, but this was not the case for high GC%. This effect was more prominent for 3′ UTRs. *P* values were determined by a two-sided Wilcoxon rank sum test. **(F)** UTRs with more eCLIP RBP binding signals per UA dinucleotide are more stable. The high and low groups was stratified based on the median of number of RBPs per UA. **(G)** UTRs harboring more predicted RBP-binding sites per UA dinucleotide are more stable, as determined by MPRA. *P* values were determined by two- sided Wilcoxon rank sum test (F and G). **(H)** Comparison of RNA half-life of high-UA UTRs determined by MPRA and transcription inhibition with actinomycin D (ActD) (H). *ρ* represents Spearman's correlation coefficient. **(I)** RNA stability assay with Actinomycin D treatment. Error bars are standard errors computed from three experimental replicates. **: FDR-adjusted *p* value = 0.002. **(J)** Association of UA dinucleotide-binding protein motifs with RNA half-life in SH-SY5Y cells. Note that UA-binding RBPs can have both positive and negative effects on RNA stability.

displayed longer half-life compared to the group with few binding partners. This result proved consistent for both UTR types and based on experimental (**Fig. 5F**) or predicted (**Fig. 5G**) RBP binding data. To further validate the RBP protective effect against UA- destabilization, we selected four 3' UTRs (APC, WDR35, SH3TC2 and MTR) with a UA- dinucleotide ratio greater than 90% quantile and assessed their RNA stability by transcription inhibition with actinomycin D (**Fig. 5H,I**). Half-life obtained by the actinomycin D treatment were highly correlated with the MPRA result ($\rho = 0.8$) in accordance with numbers of RBP binding sites per UA dinucleotide. Together, the results argue a protective role of RBP binding on UA dinucleotides.

In acknowledgment of the dual potential of RBPs to either promote or inhibit RNA degradation, we conducted a detailed analysis of the impact of each RBP's binding on RNA stability. By correlating the binding of UA dinucleotide-binding proteins with RNA half-life, we identified UA-binding RBPs that play roles in either safeguarding or promoting RNA degradation (**Fig. 5J**; Supplemental Fig. S7E; Supplemental Table S5). Notably, the protective UA-containing motifs were found to be U-rich, reinforcing the observation that U-stretch sequences may enhance RNA stability. Thus, we have identified interplay between GC content, RBP-binding and the destabilizing effect on RNAs of UA dinucleotides. The fact that these regulatory factors may control RNA stability via synergistic or antagonistic mechanisms emphasizes the need to consider all contributory factors simultaneously, as achieved by our modeling approach, to gain a complete overview of the regulatory network (**Fig. 3**).

UA-dinucleotide ratio of UTRs reflects functional enrichment

Since we identified the UA dinucleotides of UTRs as a major stability determinant, we argued that if UA dinucleotides do indeed represent a functional motif regulating RNA stability, then this property could be a proxy of expression dynamics for classifying genes into functional groups. Therefore, we calculated the average UA-dinucleotide ratio of each gene and compared this distribution within a given GO term against the entire genome background. We found that the 5' UTRs of genes responsible for regulating appetite, apoptosis and synaptic signaling display the highest UA-dinucleotide ratios and those involved in glutathione metabolism exhibit the lowest (**Fig. 6A**; Supplemental Table S6). In terms of 3' UTRs, genes linked to integrin activation, the interleukin-mediated pathway and B-cell differentiation have the highest UA-dinucleotide ratios (**Fig. 6B**). To investigate how UA dinucleotide localization contributes to biological functions, we utilized a sliding window approach to identify UTR regions with UA dinucleotides above or below genomic background levels (**Fig. 6C,D**). UA dinucleotide in each 10-nt window with 1-nt step was calculated and normalized to the UTR length (Methods). For 5' UTRs, we found that synaptic signaling and striated muscle contraction pathways represent the functional groups with genes having most UA dinucleotide-enriched windows, whereas genes responsible for regulating cell proliferation had most UA dinucleotides-depleted windows (**Fig. 6C**). For 3' UTRs, RNA translation and immune-related functions proved to be the GO terms with most UA dinucleotide-enriched windows (**Fig. 6D**). Since 5' and 3' UTRs may synergistically regulate gene expression, we examined genes for which the UA-dinucleotide ratio in both UTRs significantly differed from the genomic background, which revealed that the UA-dinucleotide ratio in both UTRs was consistently either above or below background values (**Fig. 6E,F**). Plotting these positive or negative effects two-dimensionally, we observed that the gene groups mostly lie in the first and third quadrants, which we interpret as indicative of a synergistic stabilizing or destabilizing effect of both UTRs. Notably, the gene groups in the first quadrant, reflecting UA-dinucleotide ratios in both UTRs being above background levels, are all related to the immune response. It is also noteworthy that the high UA-dinucleotide ratio of genes regulating appetite, synaptic signaling and the immune response reflects the transient expression nature of these gene groups, further supporting that UA dinucleotides exert a destabilizing effect on RNA. Together, this genome-wide analysis indicates that the UA-dinucleotide ratio of UTRs reflects global regulation of gene expression dynamics.

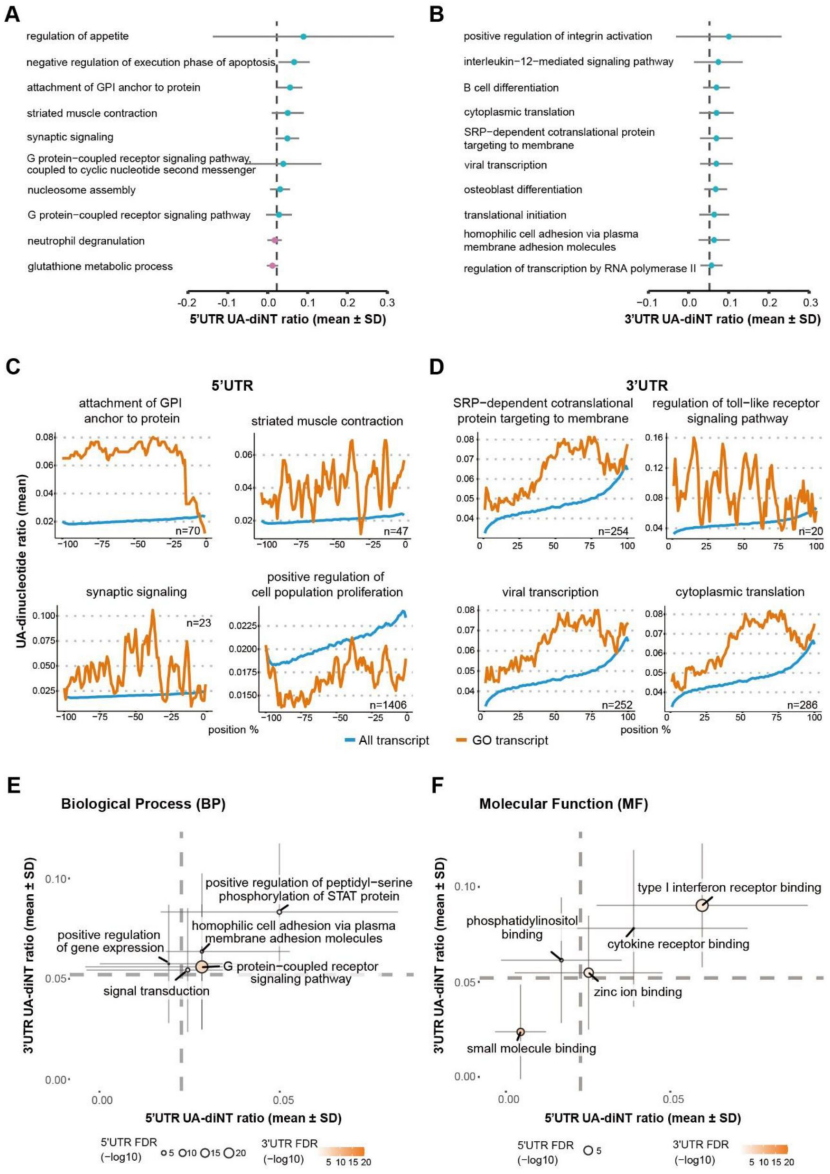


Figure 6.

UA dinucleotides delineate functional gene groups.

(A) The top ten biological processes for which the 5′ UTR UA-dinucleotide ratio most significantly deviated from the genomic background (dashed line). **(B)** The top ten biological processes for which the 3′ UTR UA-dinucleotide ratio most significantly deviated from the genomic background. **(C)** Functional gene groups for which the 5′ UTR UA-dinucleotide ratio was significantly above or below the genomic background in more than ten sliding windows. **(D)** Functional gene groups for which the 3′ UTR UA-dinucleotide ratio was significantly above or below the genomic background in more than ten sliding windows. **(E)** Biological processes for RNAs in which the UA-dinucleotide ratios of both 5′ and 3′ UTRs are significantly different from the genomic background (dashed lines). **(F)** Molecular functions for RNAs in which the UA-dinucleotide ratios of both 5′ and 3′ UTRs are significantly different from the genomic background (dashed lines). The thin solid lines represent the standard deviation of the UA-dinucleotide ratio within the gene group.

UTR variants associated with disease

The results from our MPRA stability assay and the genome-wide functional classification support the hypothesis that UTR-mediated RNA stability and gene expression may be disrupted by SNPs within the UTRs. To expand our findings from controlled MPRA experiments to human physiological conditions, we explored the effect of genetic variations in UTRs by surveying human disease databases and biobanks.

Since dysregulated RNA stability is known to contribute to cancer progression (Perron et al., 2022), we first investigated UTR mutations in samples taken from cancer patients (Harmonized Cancer Datasets: *https://portal.gdc.cancer.gov/*). We identified several SNPs in UTRs that correlated with aberrant RNA expression and/or protein expression (Supplemental Table S7). Interestingly, two 3' UTR mutations that resulted in aberrant expression of both RNA and proteins (*DDP4* [Dipeptidyl Peptidase 4] and CASP7 [Caspase 7], respectively) were A/T deletions in TA-rich sequences with increased RNA and protein expression levels (**Fig. 7A,B**), in agreement with our findings that UA-rich sequences are the most influential stability determinants (**Fig. 3**).

Finally, to establish a correlation between UTR variants and health outcomes, we examined if stability-altering UTR variations identified in our MPRA experiments are associated with abnormal physiological/pathological presentations among the Taiwanese community (Taiwan Biobank (TWB) data). We employed a quantile-quantile (Q-Q) plot to compare the *p*-value distribution of significant stability-altering UTR SNPs associated with skewed biochemical indices or self-reported diseases against theoretical *p* values (**Fig. 7C**). We observed that *p* values were skewed towards the stability-altering variants, indicating that TWB subjects harboring stability-altering UTR variants are more likely to display abnormal biochemical phenotypes. The most significant association was detected for the 3' UTR of APOC3 (apolipoprotein C3 c.*40G>C) and blood triglyceride levels ($p = 3.5 \times 10^{-74}$ as determined by linear regression) (**Fig. 7C,D**), total cholesterol ($p = 7.6 \times 10^{-12}$) and self-reported hyperlipidemia ($p = 0.00038$) (Supplemental Table S8). APOC3 is involved in metabolizing triglyceride-rich lipoproteins, and its mutation has been associated with low plasma triglyceride levels (Boren et al., 2020; Goyal et al., 2021). Our findings provide compelling evidence that regulation of RNA turnover by UTRs controls metabolic equilibrium, which can be perturbed by SNPs within the UTRs. Overall, we have demonstrated that pathogenic UTR variants are enriched in critical regulatory regions and may elicit disease.

Discussion

Small kmers in UTRs determine RNA stability

The function of UTRs in regulating RNA stability has long been recognized. However, very few reports have systematically addressed the functional impact of genetic variations in UTRs (Griesemer et al., 2021; Sample et al., 2019). Therefore, UTR variants are typically classified as being benign or of unknown functions, without experimental support. To methodically dissect the impact of such UTR variants, we have established an MPRA to test the effect of disease- related UTR variants on RNA stability. Unlike previous assays that measure the RNA:DNA ratio or protein output to infer RNA stability, our assay directly measures RNA survival and thereby circumvents confounding effects on transcription and/or protein production. However, while our approach effectively assesses the stability of synthesized RNA in human cells, it may not fully capture the decay dynamics of nuclear-synthesized RNA, which can be influenced by endogenous modifications and *trans*-acting RNA binding factors.

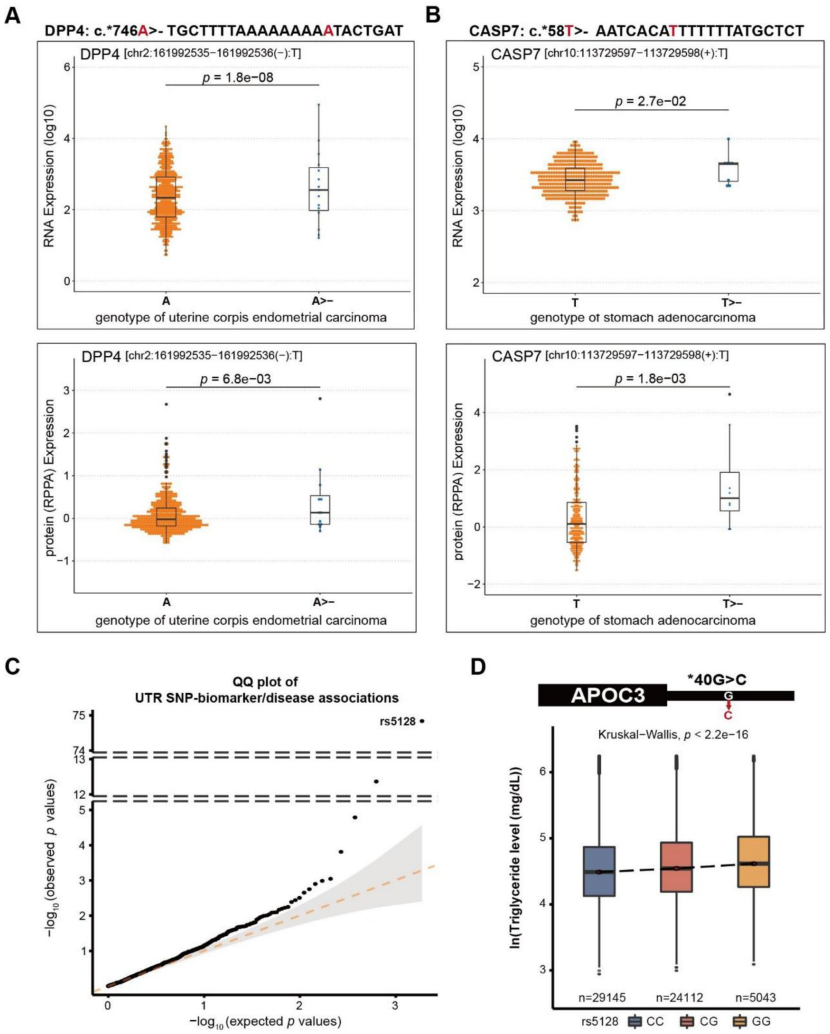


Figure 7.

UTR variants associated with disease.

(A-B) 3′ UTR mutations that increase RNA and protein expression in carcinoma samples. Protein level was determined by reverse phase protein arrays (RPPA). **(C)** QQ plot of the *p* value distribution of stability-altering UTR variants in association with health biomarkers or self-reported diseases against a theoretical distribution. **(D)** The G allele of the most significant UTR variant (rs5128) identified in (C) is associated with plasma triglyceride levels in the Taiwanese population (TWB dataset).

From among almost 1,500 potential stability regulators, we applied LASSO regression to select 5-19 independent factors that best explained RNA half-life. The major destabilizing factors proved to be UA dinucleotides, as well as WWWWWW (W:A/U) and AUUUA (ELAV1 binding site) motifs, all highlighting the importance of UA-rich sequences in UTRs for RNA stability (**Fig. 3**). Among these destabilizing factors, UA dinucleotides were best correlated with RNA half- life (**Fig. 4A-D**). The UA-dinucleotide ratio rather than GC content or folding energy explained our RNA stability data better, implying that specific sequence recognition by *trans* factors and not simply regulation according to structuredness is the underlying control mechanism. Similarly, MPRA on 3' UTR variants identified mono- or di-nucleotide composition as primary factors controlling RNA expression (Griesemer et al., 2021). This di-nucleotide specificity may partially be attributed to the sequence preference of a human ribonuclease superfamily in which eight catalytically-active RNases (numbered 1-8) all share homology with bovine pancreatic ribonuclease A. RNase A cleaves 3'-5' phosphodiester bonds with a specificity for pyrimidines (U/C) at the main anchoring site and purines (A/G) at the secondary site (Sorrentino, 2010). Kinetic analysis has revealed a >100-fold preference for UpA than UpG substrates for the human RNase A family (RNase 1-7, RNase 8 was not studied in this report) (Prats-Ejarque et al., 2019), consistent with our finding that UA dinucleotides represent the most destabilizing sequence motifs. While most of the studies on RNA degradation mechanism concerned deadenylation and decapping followed by exonuclease digestion, the contribution of endonucleases on overall RNA degradation might be underestimated. It was reported that transcript and UTR length is negatively correlated with RNA stability (Benjamin Neymotin, 2015; Blumberg et al., 2021), implying that increased RNA mass could be susceptible to endonuclease attack. Thus, our findings and those of others demonstrate that specific sequence recognition, especially di-nucleotide composition, determines RNA stability.

Interplay of GC content, RBP binding and UA dinucleotides for RNA stability

Although it was intuitive to infer a negative correlation between UA dinucleotides and GC content, the best-known stability regulator, we found that UA dinucleotides cannot simply be viewed as an inverse proxy for GC content. Instead, the UA-dinucleotide ratio more adequately explained our RNA stability data, and the protective effect of GC content on stability was almost abolished in the context of a low UA-dinucleotide ratio (Supplemental Fig. S7). Consequently, UA dinucleotides destabilize RNA more robustly under conditions of low GC content (**Fig. 5A**; Supplemental Fig. S7). This reciprocal interaction was also apparent in independent RNA stability assays and P-body transcriptomic analysis that inferred RNA degradation *in vivo* (**Fig. 5C-E**). Thus, the effect of GC content was also impacted by the sequence context, potentially explaining discrepancies among previous reports on the effects of GC content on RNA stability (Courel et al., 2019; Griesemer et al., 2021; Litterman et al., 2019; Zhao et al., 2014). In contrast, although RBPs do not always protect RNA from degradation, with their effect on RNA stability depending on the factors recruited (**Fig. 5J**), the destabilizing effect of UA dinucleotides is generally hampered by RBP binding (**Fig. 5F,G**). We reason that, similar to structural hindrance, recognition of UA dinucleotides can be blocked by the physical occupancy of RBPs, resulting in an overall protective effect of RBPs against UA dinucleotide-mediated degradation that may overwhelm the destabilizing effect of a subset of ARE-binding proteins.

UTR variants linked to RNA stability and population health

We identified many crucial regulatory features in 5' UTRs. Thus, our results evidence that 5' UTRs are an indispensable region for controlling RNA stability. Our MPRA data revealed that more 5' UTR variants than those in 3' UTRs caused stability changes (**Fig. 1B**). Moreover, we found that stability-altering variants of both UTR types are associated with skewed biomarker or disease accessions in the TWB. Although we did not intend to dissect translation-dependent or -independent decay pathways in our system, it demonstrated the critical contribution of both UTR types to regulating RNA stability. Significantly, our presentation of UA dinucleotide enrichment or depletion in the UTRs of functional gene groups indicates that both UTR types may jointly regulate

RNA stability to achieve kinetic control of a functional pathway (**Fig. 6E,F**). At the sequence level, our results indicate that mutational gain of UA dinucleotides is linked to diminished RNA half-life (**Fig. 4E**). This simple rule can be adopted as a primary screening for UTR mutation-mediated pathologies and a principle for sequence design of synthetic RNA to be expressed in human cells. Further validation of the sequence effects can be undertaken in pertinent cell or tissue types. Collectively, our findings have uncovered the RNA stability regulation exerted by the sequence composition of UTRs, revealed health-related UTR variants, and provided a foundation for precise diagnosis of non-coding genetic variants.

Methods

Metagene analysis

Both variants on all UTR and uORF use dbSNP version 151 (Sherry et al., 1999) as baseline, disease variants, UTR sources are as described above. uORF definitions were downloaded from TIS-db (Wan & Qian, 2014), and their genome coordinates were lifted from hg19 to hg38 before analysis.

Disease variant collection, UTR library construction

Variant collection

Disease-related variants were collected from ClinVar (Landrum et al., 2016) and HGMD database (Stenson et al., 2017). Variants located in the coding region or labeled as benign variants were removed. The UTRs were defined by NCBI RefSeq and ENCODE V27. We then intersected selected variant positions with the UTR regions by using BEDTools (Quinlan & Hall, 2010) to collect UTR variants.

Library construction

For each variant, we extracted 115 bp of sequence around variant position, and built one oligo pair, reference and mutant, which only differs only by the variant position. The variant was placed in the middle of the sequence, unless it is located near the boundary of UTR. UTR specific primers were then added on each sequence. (5' UTR-F: 5'-CGCTAGGGATCCTCTAGTCA- 3', 5' UTR-R: 5'-ACCGGTCGCCACCATGGTGA-3'; 3' UTR-F: 5'-GGACGAGCTGTACAAGTAAA-3', 3' UTR-R: 5'-GCGGCCGCGCAATAACTAGC-3'). Overall, we built an oligo library containing 12,472 sequences (6,555pairs) of both 5' and 3' UTR, and synthesized them by CustomArray Inc. (U.S.).

UTR-library DNA templates were assembled by overlap extension PCR with Herculase II Fusion Enzyme (Agilent Technologies). Initially, the oligonucleotide library sequences were double-strandized and amplified by PCR. After being subjected to PCR clean-up (Qiagen), the UTR- library amplicons were then appended with EGFP CDS (derived from EGFP-N1 vector) and T7 promoter sequence through overlapping PCR. Eventually, the assembled full-length DNA templates (T7-5' UTR-EGFP or T7-EGFP-3' UTR) were subjected to PCR clean-up again for the following *in vitro* transcription.

In vitro transcription and polyadenylation

200 ng of the PCR product was subject to *in vitro* transcription using MEGAscript™ T7 Transcription Kit (Thermo Fisher Scientific). To cap the RNA product, m7G(5')ppp(5')G RNA Cap Structure Analog (New England Biolabs, 4:1 to GTP) was supplemented to the *in vitro* transcription reaction. After 3 hours' incubation in 37°C, DNA template was removed by 1 µl Turbo DNase at 37°C for 15 min. The RNA product was purified by illustra microspin G-50 column (GE Healthcare

Life Sciences). 10 μg of the purified RNA was then polyadenylated by 4U poly(A) polymerase (New England Biolabs) at 37°C for 1 hr and purified again by illustra microspin G-50 column or Direct-zol RNA Miniprep kits (Zymo Research) before transfection (Supplemental Fig. S1C).

Transfection

Seed HEK293T or SH-SY5Y cells 20 hrs before transfection. Transfect 500 ng capped and polyadenylated RNA per well into a 12-well plate by Lipofectamine 3000 reagent (Thermo Fisher Scientific). Wash the cells twice before collecting the first time point (30 min for HEK293T, 20 min for SH-SY5Y), and then harvest RNA along the time course.

Amplicon preparation for sequencing

RNA was extracted using Trizol reagent (Invitrogen) and total RNA was extracted with Direct- zol RNA Miniprep kits (Zymo Research). Afterwards, cDNA was reverse transcribed from 1 μg of RNA using library-specific primers (5' UTR: EGFP3-f_UMI_5Lib-RT; 3' UTR: CMV3- f_UMI_M13-rev, see Supplemental Table S9) with SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific). The cDNA was amplified first by primers EGFP3-f and CMV3-f 15 cycles by Phusion™ High-Fidelity DNA Polymerase (Thermo Fisher Scientific) with GC buffer. The resultant product was cleaned up using QIAGEN PCR Purification kit and subject to second round of PCR with primers Forward/P5_fractionsID_CMV3f and Reverse/P7_#N_EGFP3f for 5' UTR library, and Forward/P5_fractionsID_EGFP3f and Reverse/P7_#N_CMV3f (Supplemental Table S9) for 3' UTR library by 10 cycles. The PCR product was purified again by QIAGEN PCR Purification kit for Illumina® NextSeq paired-end 150 sequencing.

RNA-seq data processing

QC

PCR duplication of single-end raw fastq files (150bp) was first removed by nubeam-dedup (Dai & Guan, 2020), then adapter and low-quality base (sliding mean quality < 20) was trimmed by trimmomatic (Bolger et al., 2014). Trimmed reads of length less than 80bp were discarded.

Alignment

FASTQ data alignment was done by HISAT2 (Kim et al., 2015), genome information was built from the UTR library.

Detail parameters are as below, HISAT2: hisat2 --no-spliced-alignment --score-min L,0,-0.7 -x index -U trimmed_fastq.gz

Count

After alignment, BEDTools multicov (Quinlan & Hall, 2010) was used to build the count matrix. Since accuracy is critical in this project, alignment results were filtered by mapping quality (HISAT2: MAPQ = 60), and count only "exactly one alignment" result. Reads were normalized by DESeq2 (Love et al., 2014).

Half-life estimation

For each oligonucleotide, we estimated decay constant λ and half-life ($t_{1/2}$) by the following equations:

$$ln\left(\frac{C_{i(t)}}{C_{i(t=0)}}\right) = -\lambda t_i$$
$$t_{1/2} = \frac{ln\ (2)}{\lambda}$$

where $C_{i(t)}$ and $C_{i(t=0)}$ are read count values of the i^{th} replicate at time points t and 0. For a more accurate estimate of λ, we used the first time point (HEK293T: 30min, SH-SY5Y: 20min) as the time point of t = 0, and the models were weighted by their inverse standard deviations. All the read count values were normalized by the average of $C_{i(t=0)}$ before constructing the linear models. After that, half-life λ was estimated by linear regression function "lm" of R. We only selected oligonucleotide that R^2 > 0.5 and *MSE(mean squared error)* < 1 for further analysis.

Statistic method to detect mutation effect on stability

To determine variants that changed oligonucleotide stability, we used the normalized counts to build a linear model weighted with the inverse of their standard deviations as follows:

$$ln\left(\frac{C_{i(t)}}{C_{i(t=0)}}\right) = \beta t_i + \theta t_i G_i$$

where G_i representedreference (ref) or mutant (mt).

We tested the significance of θ to determine the effect of the variant. We then corrected the p value by FDR.

All statistical analyses were performed by R language (version 4.0.5/4.2.0), linear models were constructed by "lm" function and p value adjustment was conducted by "p.adjust" function and "qvalue" function from the R package "qvalue".

In-silico feature prediction

Free Energy

Free Energy was estimated by the ViennaRNA package RNAfold command (Lorenz et al., 2011), which predicted minimum free energy (MFE) by default.

GC content and kmer ratio

GC content, the ratio of G/C nucleotide of a sequence, and short kmer (k=1~3) were calculated by letterFrequency and oligonucleotideFrequency function of the R package Biostrings (*https://bioconductor.org/packages/Biostrings*).

Motifs

For motif scanning such as PUM motif, RBP motif, ARE and novel motif, were done by Biostrings functions, countPWM or countPattern (*https://bioconductor.org/packages/Biostrings*). PUM motif was retrieved from (Rabani et al., 2017). RBP motif PWMs were download from ATtRACT database (Giudice et al., 2016) and ARE definition followed by AREsite database (Fallmann et al., 2016), including ATTTA, WTTTW, WWTTTWW, WWWTTTWWW, WWWWTTTWWWW, WWWWWTTTWWWWW, TTTGTTT, GTTTG, AWTAAA.

KOZAK/uAUG

KOZAK sequence was defined as in the previous study (Hernandez et al., 2019), optimal: GCCRCCAUGG; strong: NNNRNNAUGG; moderate_H: NNNRNNAUG(A/C/U); moderate_Y: NNN(C/U)NNAUGG and weak: NNN(C/U)NNAUG(A/C/U). We scanned these motifs on the library sequence by R package Biostrings countPattern function.

miRNA

miRNA binding site was predicted by TargetScan 7.0 (Agarwal et al., 2015; McGeary et al., 2019). miRNA expression information was downloaded from miRmine database (Panwar et al., 2017), only those miRNAs that have at least 1 RPM in HEK293T or SH-SY5Y cells were selected as features. Since the 6-mer seed sequence is known to be less effective (Bartel, 2009), only predicted 7mer (7mer-m8, 7mer-1a) and 8mer seed sequences were kept as features.

RNA binding protein binding

eCLIP (enhanced crosslinking immunoprecipitation) data were downloaded from ENCODE (Y. H. Luo et al., 2020) and GSE117290 datasets (E. C. Luo et al., 2020), and selected only eCLIP regions identified in two biological replicates. eCLIP signals with value < 1 or p value > 0.05 were removed. All genome coordinates were lifted to hg38 by liftover command (Hinrichs et al., 2006). RBP bindings were determined by intersecting eCLIP signals and variant coordinates by BEDTools intersect command (Quinlan & Hall, 2010).

G-quadruplex (RG4)

G-quadruplex structures were predicted by the RNAfold command from the ViennaRNA package (Lorenz et al., 2011), with -g parameter.

Conservation level

phastCons scores (pC4way-pC100way) were downloaded from UCSC genome browser (Lee et al., 2022) and intersected with UTR genome coordinates by the intersect function of BEDTools (Quinlan & Hall, 2010). An average phastCons score was calculated for each UTR.

Novel 7-mer motifs

Primer sequences were first removed from the UTR library sequence. Then a 7-mer table was built by manual R script (R-Core-Team, 2021). UTR library half-life was transformed by log2() and the extreme values were filtered out. 7-mers occurring less than 20 times in the UTR library were filtered out. Regress 7-mers to log2(half-life) by glmnet::glmnet() LASSO selection (Friedman et al., 2010) to obtain the coefficients. Repeat the regression 2000 times with bootstrap resampling and select 7-mers that were selected > 1,600 times for further analysis. The non-zero distribution of coefficients of a 7-mer was then tested by permutation coin::oneway_test (Hothorn et al., 2008), and the resulting p values were adjusted by p.adjust() Bonferroni adjustment. 7-mers with non-zero coefficients (adjusted p values < 0.05) were further divided into positive and negative effects according to their mean coefficients. To generate motifs, the text distance in each group was calculated using stringdist::stringdistmatrix() function with “lv” distance (van der Loo, 2014) and clustered by hclust(). Finally, use cutree(), h=2 to subgroup the 7-mers. 7-mers of a subgroup were combined to build PWM according to the weight of their mean coefficients.

Feature selection

A LASSO model was utilized to perform feature selection. Features with a high proportion of 0 (≥ 90%) were excluded to ensure the robustness of the selection model. In addition, to avoid multicollinearity caused by similar features that perturb feature selection, all features were clustered using single-linkage hierarchical clustering with the distance metric defined as one minus the absolute value of the Spearman correlation coefficient. We cut the tree at a specific height, and the feature that had the greatest influence on RNA stability, which was examined using a simple linear regression model, was selected to be the representative of each cluster. Then we calculated the variance inflation factor (VIF) value of the representative features. The VIFs were obtained by the following linear model and equations:

$$\hat{X}_{ij} = \hat{\alpha}_{0(j)} + \sum_{k \neq j} \hat{\alpha}_{k(j)} X_{ik}$$

$$R_j^2 = \frac{\sum_i \left(\hat{X}_{ij} - \bar{X}_{\cdot j}\right)^2}{\sum_i \left(X_{ij} - \bar{X}_{\cdot j}\right)^2}$$

$$VIF_j = \frac{1}{1 - R_j^2}$$

where $\hat{X}_{ij}$ and X_{ik} are the estimated value of the jth feature and the value of the kth feature of the ith UTR (note that the kth feature is a feature other than the jth feature), $\hat{\alpha_{0(j)}}$ and $\hat{\alpha_{k(j)}}$ are the intercept and the regression coefficients of the linear model that regressed the jth feature on the other remaining features, and $X_{\cdot j}$ is the mean level of the jth feature of all UTRs.

The height for cutting the tree was gradually increased until the VIF values of all representative features were less than 5. By identifying and excluding highly collinear variables, we aimed to minimize multicollinearity and improve the accuracy of our regression models (Akinwande et al., 2015). Finally, the LASSO-based feature selection was conducted on the representative features (Supplemental Fig. S3). The LASSO model was carried out using R package glmnet (Friedman et al., 2010) and can be written as follows:

$$\underset{\beta}{\operatorname{argmin}} \left\{ \sum_i \left(ln(t_{1/2})_i - \beta_0 - \sum_j \beta_j X_{ij} \right)^2 + \lambda \sum_j |\beta_j| \right\}$$

where β_0 and β_j are the intercept and regression coefficients of the jth feature, and ith is the level of the jth feature of the ith UTR. λ was determined through a ten-fold cross-validation process, which yielded the minimum mean cross-validated error. We performed a 2:1 split of the data into training and testing sets to validate the robustness of the model. The performance metrics, including the correlation coefficient between observed and predicted values, mean average error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and R-squared, are provided in Supplemental Table S4.

In vivo validation

To validate UA-diNT effect in *in vivo* data, we collected public SLAM-seq data set GSE126523 (K562) and GSE214396 (HEK293) to estimate the half-life of endogenous transcripts. The first step was to remove adapters and low-quality sequences (sliding window average sequencing quality < 20 and read length < 50) by trimmomatic. SLAM-seq fastq was then aligned to transcript sequences by HISAT-3N (Zhang et al., 2021) with annotation file gencode.v43.primary_assembly. After alignment, 'exactly one alignment' result was further used to calculate T to C conversion by hisat-3n-table. Nucleotide positions with less than 10 coverage or potential SNPs (T to C conversion rate larger than 0.8) were removed. Each transcript T to C conversion rate were calculated by Total

TtoC/(Total TtoC + Total T). Transcripts which has at least 2 batches in each timepoint were kept for further analysis. After all time points T to C conversion rate normalized to time 0 average T to C conversion rate, normalized data then fit

$$ln\left(\frac{N_{i(t)}}{N_{i(t=0)}}\right) = -\lambda t_i$$
$$t_{1/2} = \frac{ln\ (2)}{\lambda}$$

where $N_{i(t)}$ and $N_{i(t=0)}$ are T to C conversion rate of the i^{th} transcript at time points t and 0. R lm() function was used to estimate λ and calculate transcript half-life t1/2.

UA-diNT counts were calculated by R Biostrings::countPattern and sequence length by nchar(). UA-diNT ratios in each UTR sequences were UA-diNT counts normalized to the sequence length.

RNA stability assay with actinomycin D treatment

SH-SY5Y cells were seeded approximately $4x10^5$ cells per well in 12-well plates before the day of the transfection. The cells were transfected with 800 ng pEGFP-3' UTR plasmids by using Viromer® ONE RED (Lipocalyx GmbH). 24 hr post-transfection, the cells per 12-well were equally divided into 3 wells in 24-well plate. After 16-hour incubation, cells were treated with actinomycin D (5 μg/ml) and were harvested by TRIzol at the respective time points (0, 2, 4 hours after stopping the transcription). The RNA was purified by the Direct-zol™ RNA Microprep kits (Zymo Research). Equal amount of the total RNA for each sample were reverse transcribed using SuperScript™ IV (ThermoFisher) with random hexamer. Real-time PCR was performed with FastSYBR Green Plus Master Mix (Applied Biosystems) and QuantStudio™ 12K Flex Real-Time PCR System (Applied Biosystems). Relative mRNA abundance in different time points were normalized to time points 0 hr ($2^{-(\triangle Ct - \triangle Ct0)}$) for each pEGFP-3' UTR.

Construction of UTR-library DNA Templates

UTR-library DNA templates were assembled by overlap extension PCR with Herculase II Fusion Enzyme (Agilent Technologies). Initially, the oligonucleotide library sequences (CustomArray, Inc. USA) were double-strandized and amplified by PCR. After subjected to PCR clean-up (Qiagen), the UTR-library amplicons were then appended with EGFP CDS and CMV promoter sequence (derived from EGFP-N1 vector) sequentially through overlapping PCR. Eventually, the assembled full-length DNA templates (CMVP-5' UTR-EGFP or CMVP-EGFP-3' UTR) were subjected to PCR clean-up again for the following transfection.

UA-ratio analysis for GO annotation

GO annotations were downloaded from QuickGO (Binns et al., 2009) and Ensembl BioMart (Kinsella et al., 2011). Full Human genome sequence and transcript annotation were from R Bioconductor package BSgenome.Hsapiens.UCSC.hg38 v.1.4.4 (UCSC version hg38, based on GRCh38.p13) and TxDb.Hsapiens.UCSC.hg38.knownGene v.3.15.0 . UTR length < 10 nucleotides were removed from analysis.

UA-dinucleotide ratio in each UTR sequence was calculated by (UA count ÷ (sequence length - 1)). For every GO annotation containing > 20 qualified transcripts, UA-ratio of those transcripts were compared with all transcripts by Fisher-Pitman permutation test (R coin package (Hothorn et al., 2008)). Bonferroni corrected p value < 0.01 was defined as significant enrichment/depletion.

For the sliding UA-dinucleotide ratio, UA-ratio of the most 5' 10-nt window was first calculated and the window was slid by 1 nt in each movement to the 3'end for each transcript. Then each UTR was normalized to length by splitting sliding UA-ratio to 100 fragments and the mean sliding UA-

dinucleotide ratio in each fragment was calculated. Finally calculate the mean ratio in each fragment across all transcripts within a GO term was calculated to represent the sliding UA-dinucleotide ratio.

TCGA Data

TCGA data were downloaded by R/Bioconductor package TCGABiolinks (Colaprico et al., 2016) with the following query code:

RNA-seq: GDCquery(TCGA-id, data.category = “Transcriptome Profiling”, experimental.strategy = “RNA-Seq”, data.type = “Gene Expression Quantification”,workflow.type = “STAR - Counts”)
Protein: GDCquery(project = TCGA-id, data.type = “Protein expression quantification”, legacy = TRUE, data.category = “Protein expression”, platform = “MDA_RPPA_Core”)

Only donor IDs and variants valid in both RNA/Protein datasets underwent further analyses.

The Taiwan Biobank Data, TWB

In order to identify the association between RNA stability altering variants and common chronic diseases and biochemical indices, the genotyping and phenotypic data of 68,978 Taiwanese people were obtained from the TWB (*https://www.biobank.org.tw/*). The information on the diseases was self-reported and collected through questionnaires. Each participant was genotyped on the Affymetrix Axiom genome-wide TWB 2.0 array containing 752,921 SNP (single nucleotide polymorphism) probes. The study was approved by the Institutional Review Board of Academia Sinica.

We conducted statistical analyses to assess the association of 21 variants significantly affecting RNA stability with 23 reported traits and the levels of 24 biochemical indices. Before carrying out association tests, the quality control for genotyping data was performed as described previously to ensure its reliability (Chiang et al., 2022). Linear regression and logistic regression were performed to examine the association between each SNP and continuous biochemical index and dichotomous trait, respectively. Multinomial logistic regression followed by a likelihood ratio test was used to determine p-values for biochemical indices containing more than two levels and was fitted by R package nnet (v 7.3-17). All SNPs were tested as co-dominant genetic models. The square root of age, gender, dwelling place and the batch of array were included in the regression models to adjust for potential confounding effects. The first ten principal components were also included as covariates in all regression models to control the population stratification.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; *https://www.ncbi.nlm.nih.gov/geo/*) under accession number GSE217518 (reviewer token snspaakujtsdpcv).

Codes used for the analysis in this study have been deposited at *https://github.com/chienlinglin/modeling-UTR-variants-stability/* and as Supplemental Code.

Other databases used in the study:

UCSC PhyloP: https://genome.ucsc.edu/cgi-bin/hgTracks?hgsid=1351580935_14MOQtNDW7V78RaXEDp3Yy4m4PTb&c=chr2&hgTracksConfigPage=configure&hgtgrou

ARESite2: *http://nibiru.tbi.univie.ac.at/AREsite2* ATtRACT: *https://attract.cnic.es/download*

Ensembl: *https://asia.ensembl.org/info/data/ftp/index.html*

Harmonized Cancer Datasets: *https://portal.gdc.cancer.gov/*

Taiwan Biobanks: *https://www.biobank.org.tw/*

Acknowledgements

We thank the Genomics Core and the Bioinformatics Core of the Institute of Molecular Biology (IMB), Academia Sinica, for performing the amplicon sequencing and for providing computing resources. We thank all members of IMB, particularly Drs. Jun-Yi Leu and Hung-Lun Chiang, for tremendous help and support.

Additional information

Funding

This work was supported by Career Development Award and Multidisciplinary Health Cloud Research Program of Academia Sinica (AS-CDA-108-M03 and AS-PH-109-01-3), Career Development Award of National Health Research Institute, Taiwan (NHRI-EX112-10908BC) and Excellent Young Scholar Research Grants and Ta-You Wu Memorial Award of National Science and Technology Council, Taiwan (MOST 111-2628-B-001-003 and 108-2118-M-001- 013-MY5).

Author contributions statement

Y-L W and J-Y S carried out modeling and analyses. Y-C C, C-H Y, YS K, and C-L L established the MPRA. Y-T Hsieh performed the validation experiment. Y-T Huang supervised the statistical analysis. Y-L W, J-Y S, Y-C C, C-H Y and C-L L wrote the manuscript.

Additional files

supplemental information

Supplemental Table S1

Supplemental Table S2

Supplemental Table S3

Supplemental Table S4

Supplemental Table S5

Supplemental Table S6

Supplemental Table S7

Supplemental Table S8

Supplemental Table S9

References

Agarwal V., Bell G. W., Nam J. W., Bartel D. P (2015) **Predicting effective microRNA target sites in mammalian mRNAs** *Elife* **4** https://doi.org/10.7554/eLife.05005

Agarwal V., Kelley D. R (2022) **The genetic and biochemical determinants of mRNA degradation rates in mammals** *Genome Biology* **23** https://doi.org/10.1186/s13059-022-02811-x

Akinwande M. O., Dikko H. G., Samson A (2015) **Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis** *Open Journal of Statistics* **5**:754–767 https://doi.org/10.4236/ojs.2015.57075

Barreau C., Paillard L., Osborne H. B (2005) **AU-rich elements and associated factors: are there unifying principles?** *Nucleic Acids Research* **33**:7138–7150 https://doi.org/10.1093/nar/gki1012

Barrett L. W., Fletcher S., Wilton S. D (2012) **Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements** *Cellular and Molecular Life Sciences* **69**:3613–3634 https://doi.org/10.1007/s00018-012-0990-9

Bartel D. P (2009) **MicroRNAs: Target Recognition and Regulatory Functions** *Cell* **136**:215–233 https://doi.org/10.1016/j.cell.2009.01.002

Beadle L. F., Love J. C., Shapovalova Y., Artemev A., Rattray M., Ashe H. L. (2023) **Combined modelling of mRNA decay dynamics and single-molecule imaging in the embryo uncovers a role for P-bodies in 5' to 3' degradation** *Plos Biology* **21** https://doi.org/10.1371/journal.pbio.3001956

Beffagna G. *et al.* (2005) **Regulatory mutations in transforming growth factor-beta 3 gene cause arrhythmogenic right ventricular cardiomyopathy type 1** *Cardiovascular Research* **65**:366–373 https://doi.org/10.1016/j.cardiores.2004.10.005

Benjamin Neymotin V. E., Gresham David (2015) **Global determinants of mRNA degradation rates in Saccharomyces cerevisiae [New Results]** *bioRxiv* https://doi.org/10.1101/014845

Binns D., Dimmer E., Huntley R., Barrell D., O'Donovan C., Apweiler R (2009) **QuickGO: a web-based tool for Gene Ontology searching** *Bioinformatics* **25**:3045–3046 https://doi.org/10.1093/bioinformatics/btp536

Blumberg A., Zhao Y. X., Huang Y. F., Dukler N., Rice E. J., Chivu A. G., Krumholz K., Danko C. G., Siepel A (2021) **Characterizing RNA stability genome-wide through combined analysis of PRO-seq and RNA-seq data** *Bmc Biology* **19** https://doi.org/10.1186/s12915-021-00949-x

Bolger A. M., Lohse M., Usadel B (2014) **Trimmomatic: a flexible trimmer for Illumina sequence data** *Bioinformatics* **30**:2114–2120 https://doi.org/10.1093/bioinformatics/btu170

Boren J., Packard C. J., Taskinen M. R (2020) **The Roles of ApoC-III on the Metabolism of Triglyceride-Rich Lipoproteins in Humans** *Frontiers in Endocrinology* **11** https://doi.org/10.3389/fendo.2020.00474

Chiang H. L., Chen Y. T., Su J. Y., Lin H. N., Yu C. H. A., Hung Y. J., Wang Y. L., Huang Y. T., Lin C. L (2022) **Mechanism and modeling of human disease-associated near-exon intronic variants that perturb RNA splicing** *Nature Structural & Molecular Biology* **29** https://doi.org/10.1038/s41594-022-00844-1

Colaprico A. *et al.* (2016) **TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data** *Nucleic Acids Research* **44** https://doi.org/10.1093/nar/gkv1507

Courel M. *et al.* (2019) **GC content shapes mRNA storage and decay in human cells** *Elife* **8** https://doi.org/10.7554/eLife.49708

Dai H., Guan Y. T (2020) **The Nubeam reference-free approach to analyze metagenomic sequencing reads** *Genome Research* **30**:1364–1375 https://doi.org/10.1101/gr.261750.120

Dumas L., Herviou P., Dassi E., Cammas A., Millevoi S (2021) **G-Quadruplexes in RNA Biology: Recent Advances and Future Directions** *Trends in Biochemical Sciences* **46**:270–283 https://doi.org/10.1016/j.tibs.2020.11.001

Dusl M., Senderek J., Muller J. S., Vogel J. G., Pertl A., Stucka R., Lochmuller H., David R., Abicht A (2015) **A 3'-UTR mutation creates a microRNA target site in the GFPT1 gene of patients with congenital myasthenic syndrome** *Human Molecular Genetics* **24**:3418–3426 https://doi.org/10.1093/hmg/ddv090

Fallmann J., Sedlyarov V., Tanzer A., Kovarik P., Hofacker I. L (2016) **AREsite2: an enhanced database for the comprehensive investigation of AU/GU/U-rich elements** *Nucleic Acids Research* **44**:D90–D95 https://doi.org/10.1093/nar/gkv1238

Friedman J., Hastie T., Tibshirani R (2010) **Regularization Paths for Generalized Linear Models via Coordinate Descent** *Journal of Statistical Software* **33**:1–22 https://doi.org/10.18637/jss.v033.i01

Garneau N. L., Wilusz J., Wilusz C. J (2007) **The highways and byways of mRNA decay** *Nature Reviews Molecular Cell Biology* **8**:113–126 https://doi.org/10.1038/nrm2104

Giudice G., Sanchez-Cabo F., Torroja C., Lara-Pezzi E (2016) **ATtRACT-a database of RNA-binding proteins and associated motifs** *Database-the Journal of Biological Databases and Curation* https://doi.org/10.1093/database/baw035

Goyal S. *et al.* (2021) **APOC3 genetic variation, serum triglycerides, and risk of coronary artery disease in Asian Indians, Europeans, and other ethnic groups** *Lipids in Health and Disease* **20** https://doi.org/10.1186/s12944-021-01531-8

Griesemer D. *et al.* (2021) **Genome- wide functional screen of 3 ' UTR variants uncovers causal variants for human disease and evolution** *Cell* **184** https://doi.org/10.1016/j.cell.2021.08.025

Guvenek A., Shin J., De Filippis L., Zheng D. H., Wang W., Pang Z. P. P., Tian B. (2022) **Neuronal Cells Display Distinct Stability Controls of Alternative Polyadenylation mRNA Isoforms, Long Non-Coding RNAs, and Mitochondrial RNAs** *Frontiers in Genetics* **13** https://doi.org/10.3389/fgene.2022.840369

Hernandez G., Osnaya V. G., Perez-Martinez X (2019) **Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes** *Trends in Biochemical Sciences* **44**:1009–1021 https://doi.org/10.1016/j.tibs.2019.07.001

Hinrichs A. S. *et al.* (2006) **The UCSC Genome Browser Database: update 2006** *Nucleic Acids Research* **34**:D590–D598 https://doi.org/10.1093/nar/gkj144

Hothorn T., Hornik K., van de Wiel M. A. V., Zeileis A. (2008) **Implementing a Class of Permutation Tests: The coin Package** *Journal of Statistical Software* **28**:1–23

Huntzinger E., Izaurralde E (2011) **Gene silencing by microRNAs: contributions of translational repression and mRNA decay** *Nature Reviews Genetics* **12**:99–110 https://doi.org/10.1038/nrg2936

Jia L. F., Mao Y. H., Ji Q. Q., Dersh D., Yewdell J. W., Qian S. B (2020) **Decoding mRNA translatability and stability from the 5 ' UTR** *Nature Structural & Molecular Biology* **27** https://doi.org/10.1038/s41594-020-0465-x

Kim D., Landmead B., Salzberg S. L (2015) **HISAT: a fast spliced aligner with low memory requirements** *Nature Methods* **12**:357–U121 https://doi.org/10.1038/Nmeth.3317

Kinsella R. J. *et al.* (2011) **Ensembl BioMarts: a hub for data retrieval across taxonomic space** *Database-the Journal of Biological Databases and Curation* https://doi.org/10.1093/database/bar030

Landrum M. J. *et al.* (2016) **ClinVar: public archive of interpretations of clinically relevant variants** *Nucleic Acids Research* **44**:D862–D868 https://doi.org/10.1093/nar/gkv1222

Lee B. T. *et al.* (2022) **The UCSC Genome Browser database: 2022 update** *Nucleic Acids Research* **50**:D1115–D1122 https://doi.org/10.1093/nar/gkab959

Litterman A. J., Kageyama R., Le Tonqueze O., Zhao W. X., Gagnon J. D., Goodarzi H., Erle D. J., Ansel K. M. (2019) **A massively parallel 3 ' UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization** *Genome Research* **29**:896–906 https://doi.org/10.1101/gr.242552.118

Lorenz R., Bernhart S. H., Siederdissen C. H. Z., Tafer H., Flamm C., Stadler P. F., Hofacker I. L (2011) **ViennaRNA Package 2.0** *Algorithms for Molecular Biology* **6** https://doi.org/10.1186/1748-7188-6-26

Love M. I., Huber W., Anders S (2014) **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2** *Genome Biology* **15** https://doi.org/10.1186/s13059-014-0550-8

Luo E. C. *et al.* (2020) **Large-scale tethered function assays identify factors that regulate mRNA stability and translation** *Nature Structural & Molecular Biology* **27** https://doi.org/10.1038/s41594-020-0477-6

Luo Y. H. *et al.* (2020) **New developments on the Encyclopedia of DNA Elements (ENCODE) data portal** *Nucleic Acids Research* **48**:D882–D889 https://doi.org/10.1093/nar/gkz1062

MacArthur J. *et al.* (2017) **The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)** *Nucleic Acids Research* **45**:D896–D901 https://doi.org/10.1093/nar/gkw1133

McGeary S. E., Lin K. S., Shi C. Y., Pham T. M., Bisaria N., Kelley G. M., Bartel D. P (2019) **The biochemical basis of microRNA targeting efficacy** *Science* **366** https://doi.org/10.1126/science.aav1741

Mignone F., Gissi C., Liuni S., Pesole G (2002) **Untranslated regions of mRNAs** *Genome Biology* **3** https://doi.org/10.1186/gb-2002-3-3-reviews0004

Mitchell P., Tollervey D (2001) **MRNA turnover** *Current Opinion in Cell Biology* **13**:320–325 https://doi.org/10.1016/S0955-0674(00)00214-3

Oikonomou P., Goodarzi H., Tavazoie S (2014) **Systematic Identification of Regulatory Elements in Conserved 3 ' UTRs of Human Transcripts** *Cell Reports* **7**:281–292 https://doi.org/10.1016/j.celrep.2014.03.001

Panwar B., Omenn G. S., Guan Y. F (2017) **miRmine: a database of human miRNA expression profiles** *Bioinformatics* **33**:1554–1560 https://doi.org/10.1093/bioinformatics/btx019

Paschoud S., Dogar A. M., Kuntz C., Grisoni-Neupert B., Richman L., Kuhn L. C (2006) **Destabilization of interleukin-6 mRNA requires a putative RNA stem-loop structure, an AU-rich element, and the RNA-binding protein AUF1** *Molecular and Cellular Biology* **26**:8228–8241 https://doi.org/10.1128/Mcb.01155-06

Perron G., Jandaghi P., Moslemi E., Nishimura T., Rajaee M., Alkallas R., Lu T. Y., Riazalhosseini Y., Najafabadi H. S (2022) **Pan-cancer analysis of mRNA stability for decoding tumour post-transcriptional programs** *Communications Biology* **5** https://doi.org/10.1038/s42003-022-03796-w

Pesole G., Mignone F., Gissi C., Grillo G., Licciulli F., Liuni S (2001) **Structural and functional features of eukaryotic mRNA untranslated regions** *Gene* **276**:73–81 https://doi.org/10.1016/S0378-1119(01)00674-6

Prats-Ejarque G., Lu L., Salazar V. A., Moussaoui M., Boix E (2019) **Evolutionary Trends in RNA Base Selectivity Within the RNase A Superfamily** *Frontiers in Pharmacology* **10** https://doi.org/10.3389/fphar.2019.01170

Quinlan A. R., Hall I. M (2010) **BEDTools: a flexible suite of utilities for comparing genomic features** *Bioinformatics* **26**:841–842 https://doi.org/10.1093/bioinformatics/btq033

R-Core-Team. (2021) **R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna** https://www.R-project.org

Rabani M., Pieper L., Chew G. L., Schier A. F (2017) **A Massively Parallel Reporter Assay of 3 ' UTR Sequences Identifies In Vivo Rules for mRNA Degradation** *Molecular Cell* **68** https://doi.org/10.1016/j.molcel.2017.11.014

Sample P. J., Wang B., Reid D. W., Presnyak V., McFadyen I. J., Morris D. R., Seelig G (2019) **Human 5 ' UTR design and variant effect prediction from a massively parallel translation assay** *Nature Biotechnology* **37** https://doi.org/10.1038/s41587-019-0164-5

Schoenberg D. R., Maquat L. E (2012) **Regulation of cytoplasmic mRNA decay** *Nature Reviews Genetics* **13**:448–448 https://doi.org/10.1038/nrg3254

Sherry S. T., Ward M. H., Sirotkin K (1999) **dbSNP - Database for single nucleotide polymorphisms and other classes of minor genetic variation** *Genome Research* **9**:677–679

Siegel D. A., Le Tonqueze O., Biton A., Zaitlen N., Erle D. J. (2022) **Massively parallel analysis of human 3 ' UTRs reveals that AU-rich element length and registration predict mRNA destabilization** *G3-Genes Genomes Genetics* **12** https://doi.org/10.1093/g3journal/jkab404

Slutskin I. V., Weingarten-Gabbay S., Nir R., Weinberger A., Segal E (2018) **Unraveling the determinants of microRNA mediated regulation using a massively parallel reporter assay** *Nature Communications* **9** https://doi.org/10.1038/s41467-018-02980-z

Sorrentino S (2010) **The eight human "canonical" ribonucleases: Molecular diversity, catalytic properties, and special biological actions of the enzyme proteins** *Febs Letters* **584**:2194–2200 https://doi.org/10.1016/j.febslet.2010.04.018

Stenson P. D., Mort M., Ball E. V., Evans K., Hayden M., Heywood S., Hussain M., Phillips A. D., Cooper D. N (2017) **The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies** *Human Genetics* **136**:665–677 https://doi.org/10.1007/s00439-017-1779-6

Steri M., Idda M. L., Whalen M. B., Orru V (2018) **Genetic variants in mRNA untranslated regions** *Wiley Interdisciplinary Reviews-Rna* **9** https://doi.org/10.1002/wrna.1474

van der Loo M. P. J. (2014) **The stringdist Package for Approximate String Matching** *R Journal* **6**:111–122

Vejnar C. E. *et al.* (2019) **Genome wide analysis of 3 ' UTR sequence elements and proteins regulating mRNA stability during maternal-to-zygotic transition in zebrafish** *Genome Research* **29**:1100–1114 https://doi.org/10.1101/gr.245159.118

Wan J., Qian S. B (2014) **TISdb: a database for alternative translation initiation in mammalian cells** *Nucleic Acids Research* **42**:D845–D850 https://doi.org/10.1093/nar/gkt1085

Wissink E. M., Fogarty E. A., Grimson A (2016) **High-throughput discovery of post-transcriptional cis-regulatory elements** *Bmc Genomics* **17** https://doi.org/10.1186/s12864-016-2479-7

Zhang Y., Park C., Bennett C., Thornton M., Kim D (2021) **Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N** *Genome Research* **31** https://doi.org/10.1101/gr.275193.120

Zhao W. X., Pollack J. L., Blagev D. P., Zaitlen N., McManus M. T., Erle D. J (2014) **Massively parallel functional annotation of 3 ' untranslated regions** *Nature Biotechnology* **32**:387–U259 https://doi.org/10.1038/nbt.2851

Author information

Jia-Ying Su*

Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, Taiwan, Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, Taiwan

ORCID iD: 0000-0001-5934-5458

*These authors contributed equally

Yun-Lin Wang*
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

*These authors contributed equally

Yu-Tung Hsieh
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

Yu-Chi Chang
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

Cheng-Han Yang
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

YoonSoon Kang
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan

Yen-Tsung Huang
Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

Chien-Ling Lin†
Institute of Molecular Biology, Academia Sinica, Taipei, Taiwan
ORCID iD: 0000-0002-5730-799X

For correspondence: mbcllin@gate.sinica.edu.tw
†Present Address: Institute of Molecular Biology, Academia Sinica, Taipei City, Taiwan

Editors

Reviewing Editor
John Calarco
University of Toronto, Toronto, Canada

Senior Editor
Alan Moses
University of Toronto, Toronto, Canada

Reviewer #1 (Public review):

In the manuscript by Su et al., the authors present a massively parallel reporter assay (MPRA) measuring the stability of in vitro transcribed mRNAs carrying wild-type or mutant 5' or 3' UTRs transfected into two different human cell lines. The goal presented at the beginning of the manuscript was to screen for effects of disease-associated point mutations on the stability of the reporter RNAs carrying partial human 5' or 3' UTRs. However, the majority of the manuscript is dedicated to identifying sequence components underlying the differential stability of reporter constructs. This analysis showed that UA dinucleotides are the most predictive feature of RNA stability in both cell lines and both UTRs.

The effect of AU rich elements (AREs) on RNA stability is well established in multiple systems, and the present study confirms this general trend, but points out variability in the consequence of seemingly similar motifs on RNA stability. For example, the authors report

that a long stretch of Us has extreme opposite effects on RNA stability depending on whether it is preceded by an A (strongly destabilizing) or followed by an A (strongly stabilizing). While the authors interpretation of a context-dependence of the effect is certainly well-founded, it seems counterintuitive that the preceding or following A would be the (only) determining factor. This points to a generally reductionist approach taken by the authors in the analysis of the data and in their attempt to dissect the contribution of "AU rich sequences" to RNA stability, with a general tendency to reduce the size and complexity of the features (e.g. to dinucleotides). While this certainly increases the statistical power of the analysis due to the number of occurrences of these motifs, it limits the interpretability of the results. How do UA dinucleotides per se contribute to destabilizing the RNA, both in 5' and 3' UTRs, but (according to limited data presented) not in coding sequences? What is the mechanism? RBPs binding to UA dinucleotide containing sequences are suggested to "mask" the destabilizing effect, thereby leading to a more stable RNA. Gain of UA dinucleotides is reported to have a destabilizing effect, but again no hypothesis is provided as to the underlying molecular mechanism. In addition to reducing the motif length to dinucleotides, the notion of "context dependence" is used in a very narrow sense.

The present MPRA measures the effect of UTR sequences in one specific reporter context and using one experimental approach (following the decay of in vitro transcribed and transfected RNAs). While this method certainly has its merits compared to other approaches, it also comes with some caveats: RNA is delivered naked, without bound RBPs and no nuclear history, e.g. of splicing (no EJCs), editing and modifications. Therefore, it remains to be seen whether UA dinucleotide frequency is a substantial factor in determining the half-lives of endogenous mRNAs.

The authors conclude their study with a meta-analysis of genes with increased UA dinucleotides in 5' and 3'UTRs, showing that specific functional groups are overrepresented among these genes. In addition, they provide evidence for an effect of disease-associated UTR mutations on endogenous RNA stability. While these elements link back to the original motivation of the study (screening for effects of point mutations in 5' and 3' UTRs), they provide only a limited amount of additional insights.

In summary, this manuscript presents an interesting addition to the long-standing attempts at dissecting the sequence basis of RNA stability in human cells. The analysis is in general comprehensive and sound; however, it remains unclear to what extent the findings can be generalized beyond the method and the experimental system used here.

Comments on revisions:

Parts of my original comments have been adequately addressed by the reviewers.
After reading the revised manuscript and the rebuttal, my main concern is related to the figure comparing the half-lives as measured in the two different cell lines that was included in the response to reviewer 2, but not in the revised manuscript. The complete lack of correlation between the half-lives of the 3'UTR library measured in the two cell lines is concerning. While variability and cell type-specific effects can be expected, some principles should be the same (such as the effect of UA dinucleotides that the authors report), leading to at least some correlation.
In addition, it is unclear to me why the half-lives measured for the two libraries in HEK cells are shifted (median ln(t 1/2)=6-7 for the 5'UTR library and ln(t 1/2)=4-4.5 for the 3'UTR library), but not in SH.

I feel that this figure contains important information that should be included in the final manuscript.

https://doi.org/10.7554/eLife.97682.2.sa3

Reviewer #2 (Public review):

Summary of goals:

Untranslated regions are key cis-regulatory elements that control mRNA stability, translation, and translocation. Through interactions with small RNAs and RNA binding proteins, UTRs form complex transcriptional circuitry that allows cells to fine-tune gene expression. Functional annotation of UTR variants has been very limited, and improvements could offer insights into disease relevant regulatory mechanisms. The goals were to advance our understanding of the determinants of UTR regulatory elements and characterize the effects of a set of "disease-relevant" UTR variants.

Strengths:

The use of a massively parallel reporter assay allowed for analysis of a substantial set (6,555 pairs) of 5' and 3' UTR fragments compiled from known disease associated variants. Two cell types were used.

The findings confirm previous work about the importance of AREs, which helps show validity and adds some detailed comparisons of specific AU-rich motif effects in these two cell types.

Using a Lasso regression, TA-dinucleotide content is identified as a strong regulator of RNA stability in a context dependent manner based on GC content and presence of RNA binding protein binding motifs. The findings have potential importance, drawing attention to a UTR feature that is not well characterized.

The use of complementary datasets, including from half-life analyses of RNAs and from random sequence library MRPA's, is a useful addition and supports several important findings. The finding the TA dinucleotides have explanatory power separate from (and in some cases interacting with) GC content is valuable.

The functional enrichment analysis suggests some new ideas about how UTRs may contribute to regulation of certain classes of genes.

Weaknesses:

In this section, original reviewer comments about the initial submission and the responses of the authors are listed together with new reviewer responses to the authors:

Reviewer original comment 1: It is difficult to understand how the calculations for half-life were performed. The sequencing approach measures the relative frequency of each sequence at each time point (less stable sequences become relatively less frequent after time 0, whereas more stable sequences become relatively more frequent after time 0). Since there is no discussion of whether the abundance of the transfected RNA population is referenced to some external standard (e.g., housekeeping RNAs), it is not clear how absolute (rather than relative) half-lives were determined.

Author response: [The authors showed the equations used to calculate half lives based on read counts.] They stated that "The absolute abundance was not required for the half-life calculation."

Reviewer response to authors: The methods section states that DESeq2 was used to normalize read counts. DESeq2 normalization assumes that levels of most RNAs are not different between samples. That assumption is not valid here, since RNAs in the library are introduced into cells at time 0 and all RNAs decrease over time. If DESeq2 is applied without modification to normalize across timepoints, normalized reads from less stable RNAs will decrease over time (as expected) but normalized reads from more stable RNAs will increase. Can the

authors please clarify in the methods how the read counts were normalized to account for this issue?

Reviewer original comment 2: Fig. S1A and B are used to assess reproducibility. They show that read counts at a given time point correlate well across replicate experiments. However, this is not a good way to assess reproducibility or accuracy of the measurements of t1/2 are. (The major source of variability in read counts in these plots - especially at early time points - is likely starting abundance of each RNA sequence, not stability.) This creates concerns about how well the method is measuring t1/2. Also creating concern is the observation that many RNAs are associated with half-lives that are much longer than the time points analyzed in the study. For example, based upon Figure S1 and Table S1 correctly, the median t1/2 for the 5' UTR library in HEK cells appears to be >700 minutes. Given that RNA was collected at 30, 75, and 120 minutes, accurate measurements of RNAs with such long half lives would seem to be very difficult.

Author response: ... The calculation of the half-life involves first determining the decay constant λ, which represents a constant rate of decay. Since λ is a constant, it is possible to accurately calculate it without needing data over the entire decay range. Our experimental design considers this by selecting appropriate time points to ensure a reliable estimation of λ, and thus, the half-life. To determine the most suitable time points, we conducted preliminary experiments using RT-PCR. These experiments indicated that 30, 75, and 120 minutes provided an effective range for capturing the decay dynamics of the transcripts.

Reviewer response to author comments: Based on Fig. S1D, for 3' UTRs in both cell types and for 5' UTRs in SH-SY5Y cells, median t1/2 is in the range of ~30 to 90 minutes (corresponding to ln t1/2 = 3.5 to 4.5). Measuring RNAs at 30, 75, and 120 minutes would therefore be a good choice for these cases, However, median t1/2 in HEK cells appears to be ~600 minutes (corresponding to ln t1/2 ~6.4) for HEK cells. For t1/2 of 600 minutes, RNA levels at the final time point (120 minutes) would be 90% of the those at the first time point (30 minutes), which illustrates why the method would need to be able to reliably capture very small changes in RNA abundance to accurately measure t1/2 for transcripts with half-lives much longer than 120 minutes. As suggested in our original review, this concern could be addressed by showing the correlation of half-lives across replicates for the 5' and 3' UTR libraries in both cell types. Alternatively, the authors could show other measures of reproducibility for the half-life measurements across replicates. This requires no additional experimentation and can be done using the data from replicate runs shown in Fig. S1A and B. We remain concerned that for sequences with very long half-lives, extrapolating the half-life from small changes between 30 and 120 minutes will lead to imprecise measurements.

Reviewer original comment 3: There is no direct comparison of t1/2 between the two cell types studied for the full set of sequences studied. This would be helpful in understanding whether the regulatory effects of UTRs are generally similar across cell lines (as has been shown in some previous studies) or whether there are fundamental differences. The distribution of t1/2's is clearly quite different in the two cell lines, but it is important to know if this reflects generally slow RNA turnover in HEK cells or whether there are a large number of sequence-specific effects on stability between cell lines. A related issue is that it is not clear whether the relatively small number of significant variant effects detected in HEK cells versus SH-SY5Y cells is attributable to real biological differences between cell types or to technical issues (many fewer read counts and much longer half lives in HEK cells).

Author response: For both cell lines, we selected oligonucleotides with R2 > 0.5 and mean squared error (MSE) < 1 for analysis when estimating half-life (λ) by linear regression. This selection criterion was implemented to minimize the effect of experimental noise. After quality control, we selected common UTRs and compared the RNA half-lives of the two cell lines using a scatter plot. The figure below shows that RNA half-lives are quite different between the cell lines, with a moderate similarity observed in the 5' UTRs (R = 0.21), while the

correlation in the 3' UTRs is non-significant. Despite the low correlation of mRNA half-life between the two cell lines, UA-dinucleotide and UA-rich sequences consistently emerge as the most significant destabilizing features, suggesting a shared regulatory mechanism across diverse cellular environments.

Reviewer response to author comments: We appreciate that the authors shared this additional analysis of the data. We believe that this is an important finding and that the additional figure showing correlations of half-lives across cell types should be included in the manuscript or supplement. Discussion of this result in the manuscript would also be useful for readers. This result is surprising to us since we would have expected that widely expressed RNA-binding proteins would have led to more similar effects between the two cell types, as previously found using other approaches (e.g., studies of 3' UTR effects in MPRAs). It would also be appropriate to discuss that differences seen between the two cell types indicate that caution is warranted when trying to generalize the results of this study to other cell types.

Reviewer original comment 4 has been addressed adequately in the revised manuscript.

Appraisal and impact:

Reviewer original comment 1: The work adds to existing studies that previously identified sequence features, including AREs and other RNA binding protein motifs, that regulate stability and puts a new emphasis on the role of "TA" (better "UA") dinucleotides. It is not clear how potential problems with the RNA stability measurements discussed above might influence the overall conclusions, which may limit the impact unless these can be addressed.

It is difficult to understand whether the importance of TA dinucleotides is best explained by their occurrence in a related set of longer RBP binding motifs (see Fig 5J, these motifs may be encompassed by the "WWWWWW cluster") or whether some other explanation applies. Further discussion of this would be helpful. Does the LASSO method tend to collapse a more diverse set of longer motifs that are each relatively rare compared to the dinucleotide? It remains unclear whether TA dinucleotides are associated with less stability independent of the presence of the known larger WWWWWW motif. As noted above, the importance of TA dinucleotides in the HEK experiments appears to be less than is implied in the text.

Author response: To ensure the representativeness of the features entered into the LASSO model, we pre-selected those with an occurrence greater than 10% among all UTRs. There is no evidence to support a preference for dinucleotides by LASSO. To address whether the destabilizing effect of UA dinucleotides is part of the broader WWWWWW motif, we divided UA dinucleotides into two groups: those within the WWWWWW motif and those outside of it. Specifically, we divided UTRs into two categories: 'at least one UA within a WWWWWW motif' and 'no UA within a WWWWWW motif,' and visualized the results using a boxplot. As shown in [figures provided to the reviewers], the destabilizing trend still remains for UA dinucleotides outside of the WWWWWW motif, although the effect appears to be more pronounced when UA is within the WWWWWW motif. This suggests that while UA dinucleotides have a destabilizing effect independently, their impact is amplified when they are part of the broader WWWWWW motif.

Reviewer response to authors: These are useful additional analyses, and we suggest that the additional figure and discussion should be included in the manuscript/supplement so that readers can benefit from them.

Reviewer original comment 2: The inclusion of more than a single cell type is an acknowledgement of the importance of evaluating cell type-specific effects. The work suggests a number of cell type-specific differences, but due to technical issues (especially with the HEK data, as outlined above) and the use of only two cell lines, it is difficult to understand cell type effects from the work.

The inclusion of both 3' and 5' UTR sequences distinguishes this work from most prior studies in the field. Contrasting the effects of these regions on stability is of interest, although the role of these UTRs (especially the 5' UTR) in translational regulation is not assessed here.

Author response: We examined the role of UTR and UTR variants in translation regulation using polysome profiling. By both univariate analysis and an elastic regression model, we identified motifs of short repeated sequences, including SRSF2 binding sites, as mutation hotspots that lead to aberrant translation. Furthermore, these polysome-shifting mutations had a considerable impact on RNA secondary structures, particularly in upstream AUG-containing 5' UTRs. Integrating these features, our model achieved high accuracy (AUROC > 0.8) in predicting polysome-shifting mutations in the test dataset. Additionally, metagene analysis indicated that pathogenic variants were enriched at the upstream open reading frame (uORF) translation start site, suggesting changes in uORF usage underlie the translation deficiencies caused by these mutations. Illustrating this, we demonstrated that a pathogenic mutation in the IRF6 5' UTR suppresses translation of the primary open reading frame by creating a uORF. Remarkably, site-directed ADAR editing of the mutant mRNA rescued this translation deficiency. Because the regulation of translation and stability does not converge, we illustrate these two mechanisms in two separate manuscripts (this one and doi.org/10.1101/2024.04.11.589132).

Reviewer response to authors: This is useful context. No further comment.

https://doi.org/10.7554/eLife.97682.2.sa2

Reviewer #3 (Public review):

Summary:

In their manuscript titled "Multiplexed Assays of Human Disease-relevant Mutations Reveal UTR Dinucleotide Composition as a Major Determinant of RNA Stability" the authors aim to investigate the effect of sequence variations in 3'UTR and 5'UTRs on the stability of mRNAs in two different human cell lines.

To do so, the authors use a massively parallel reporter assay (MPRA). They transfect cells with a set of mRNA reporters that contain sequence variants in their 3' or 5' UTRs, which were previously reported in human diseases. They follow their clearance from cells over time relative to the matching non-variant sequence. To analyze their results, they define a set of factors (RBP and miRNA binding sites, sequence features, secondary structure etc.) and test their association with differences in mRNA stability. For features with a significant association, they use clustering to select a subset of factors for LASSO regression and identify factors that affect mRNA stability.
They conclude that the TA dinucleotide content of UTRs is the strongest destabilizing sequence feature. Within that context, elevated GC content and protein binding can protect susceptible mRNAs from degradation. They also show that TA dinucleotide content of UTRs affects native mRNA stability and that it is associated with specific functional groups. Finally, they link disease associated sequence variants with differences in mRNA stability of reporters.

Strengths:

(1) This work introduces a different MPRA approach to analyze the effect of genetic variants. While previous works in tissue culture use DNA transfections that require normalization for transcription efficiency, here the mRNA is directly introduced into cells at fixed amounts, allowing a more direct view of the mRNA regulation.

(2) The authors also introduce a unique analysis approach, which takes into account multiple factors that might affect mRNA stability. This approach allows them to identify general sequence features that affect mRNA stability beyond specific genetic variants, and reach important insights on mRNA stability regulation. Indeed, while the conclusions to genetic variants identified in this work are interesting, the main strength of the work involves general effect of sequence features rather than specific variants.

(3) The authors provide adequate support for their claims and validate their analysis using both their reporter data and native genes. For the main feature identified, TA di-nucleotides, they perform follow-up experiments with modified reporters that further strengthen their claims, and also validate the effect on native cellular transcripts (beyond reporters), demonstrating its validity also within native scenarios.

(4) The work provides a broad analysis of mRNA stability, across two mRNA regulatory segments (3'UTR and 5'UTR) and is performed in two separate cell-types. Comparison between two different cell-types is adequate, and the results demonstrate, as expected, the dependence of mRNA stability on the cellular context. Analysis of 3'UTR and 5'UTR regulatory effects also shows interesting differences and similarities between these two regulatory regions.

Weaknesses:

In their revised manuscripts, the authors successfully address many of the weaknesses raised in the original review, including the effect of possible confounding effects, and additional methodology details. Notably, two of the issues raised in the original report, have only been partially addressed in the revision.

(1) The analysis and regression models built in this work are not thoroughly investigated relative to native genes within cells.
While using MPRAs indeed allows to isolate regulatory effects that are less influential in-vivo, the resulting effects still provide some regulatory function in-vivo. The goal of such an analysis would not be to demonstrate the predictive power of the models, or to make any claims regarding using these models to fully explain or predict the stability of native transcripts. Clearly, additional more prominent factors could function in controlling endogenous RNA stability.
Instead, the goal of such an investigation is to simply assess the fraction of in-vivo regulation that the factors identified in this work contribute in native contexts, and what is the relative contribution of the phenomena captured by the well-controlled MPRA study.
This reviewer believes that even if the effects identified by the current MPRA study only contribute a small fraction of in-vivo variation, an analysis that aim to estimate what this fraction is, will be very relevant to this study for several reasons. First, in order to appreciate the results of this study within their in-vivo context. Second, in light of the questions raised as motivation for this study, and particularly the need to identify the effect of disease-associated 3'UTR variants, which clearly have an in-vivo effect.

(2) Methodology validation can be performed with simulated data (generated in-silico by the authors) to provide an independent support for the ability of the current methodology to correctly extract regulatory effects from the data.

https://doi.org/10.7554/eLife.97682.2.sa1

Author response:

The following is the authors' response to the original reviews.

Public Reviews:

Reviewer #1 (Public Review):

In the manuscript by Su et al., the authors present a massively parallel reporter assay (MPRA) measuring the stability of in vitro transcribed mRNAs carrying wild-type or mutant 5' or 3' UTRs transfected into two different human cell lines. The goal presented at the beginning of the manuscript was to screen for effects of disease-associated point mutations on the stability of the reporter RNAs carrying partial human 5' or 3' UTRs. However, the majority of the manuscript is dedicated to identifying sequence components underlying the differential stability of reporter constructs. This shows that TA dinucleotides are the most predictive feature of RNA stability in both cell lines and both UTRs.

The effect of AU rich elements (AREs) on RNA stability is well established in multiple systems, and the present study confirms this general trend but points out variability in the consequence of seemingly similar motifs on RNA stability. For example, the authors report that a long stretch of Us has extreme opposite effects on RNA stability depending on whether it is preceded by an A (strongly destabilizing) or followed by an A (strongly stabilizing). While the authors interpretation of a context- dependence of the effect is certainly well-founded, it seems counterintuitive that the preceding or following A would be the (only) determining factor. This points to a generally reductionist approach taken by the authors in the analysis of the data and in their attempt to dissect the contribution of "AU rich sequences" to RNA stability, with a general tendency to reduce the size and complexity of the features (e.g. to dinucleotides). While this certainly increases the statistical power of the analysis due to the number of occurrences of these motifs, it limits the interpretability of the results. How do TA dinucleotides per se contribute to destabilizing the RNA, both in 5' and 3' UTRs, but (according to limited data presented) not in coding sequences? What is the mechanism? RBPs binding to TA dinucleotide containing sequences are suggested to "mask" the destabilizing effect, thereby leading to a more stable RNA. Gain of TA dinucleotides is reported to have a destabilizing effect, but again no hypothesis is provided as to the underlying molecular mechanism. In addition to reducing the motif length to dinucleotides, the notion of "context dependence" is used in a very narrow sense; especially when focusing on simple and short motifs, a more extensive analysis of the interdependence of these features (beyond the existing analysis of the relationship between TA- diNTs and GC content) could potentially reveal more of the context dependence underlying the seemingly opposite behavior of very similar motifs.

(We have used UA instead of TA, as per the reviewer's suggestion)

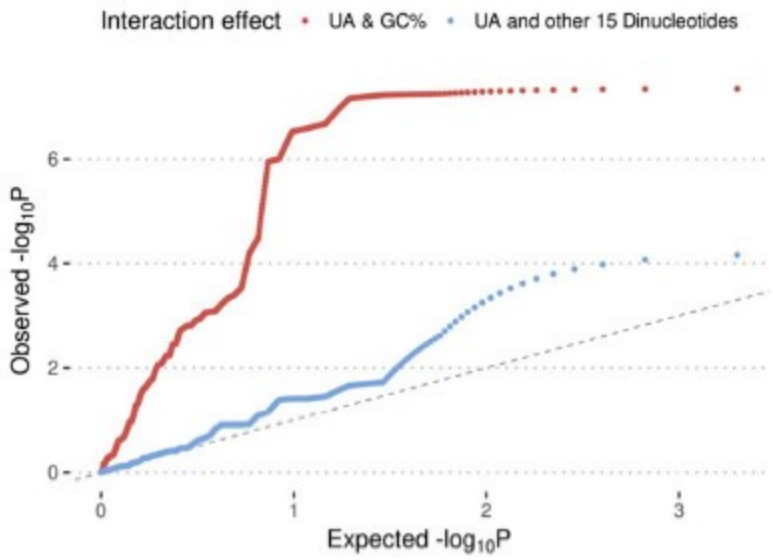
The contribution of coding region sequence to RNA stability has been extensively discussed (For example: doi.org/10.1016/j.molcel.2022.03.032; doi.org/10.1186/s13059-020-02251-5; doi.org/10.15252/embr.201948220; doi.org/10.1371/journal.pone.0228730; doi.org/10.7554/eLife.45396). While UA content at the third codon position (wobble position) has been implicated as a pro-degradation signal, codon optimality has emerged as the most prominent determinant for RNA stability. This indicates that the role of coding regions in RNA stability differs from that of UTRs due to the involvement of translation elongation. We did not intend to suggest that UA-dinucleotides in UTRs and coding regions have the same effect.

To ensure the representativeness of the features entered into the LASSO model, we pre-selected those with an occurrence greater than 10% among all UTRs. As a result, while motifs with very low occurrences were excluded from the analysis, there is no evidence to indicate a preference for dinucleotides by the LASSO model.

We hypothesize that UA-dinucleotide may recruit endonucleases RNase A family, whose catalytic pockets exhibit a strong bias for UA dinucleotide (doi.org/10.1016/j.febslet.2010.04.018). Structures or protein bindings that block this recognition might stabilize RNAs. To gain further insight into the motif interactions, we investigated the interactions between UA and other 15 dinucleotides through more detailed analyses. We conducted a linear regression analysis investigating interactions between UA and the other 15 dinucleotides. The formula used below includes UA:

$\ln(t_{1/2})_i = \beta_0 + \beta_{UA}X_{UA_i} + \beta_{DiNT}X_{DiNT_i} + \beta_{GC\%}X_{GC\%_i} + \beta_{UAxGC\%}X_{UA_i}X_{GC\%_i} + \beta_{UAxDiNT}X_{UA_i}X_{DiNT_i} + \epsilon_i$, where all β terms represent the regression coefficients, and X_{UA_i}, X_{DiNT_i}, and $X_{GC\%_i}$ represent the number of UA dinucleotides, the number of other dinucleotides (other than UA), and the GC content of the i^{th} UTR, respectively, and ϵ_i denotes the error term. For each dinucleotide, we tested the significance of $\beta_{UAxGC\%}$ and $\beta_{UAxDiNT}$, and compared their p-values using a quantile-quantile (QQ) plot. Author response image 1 shows that the interaction effect of UA dinucleotides with GC% is much more significant than interactions with the other 15 dinucleotides, as indicated by the inflated QQ plot of p-values. This suggests that GC content is a more critical contextual factor influencing UA dinucleotides' impact on RNA stability.

Author response image 1.



> *The present MPRAs measures the effect of UTR sequences in one specific reporter context and using one experimental approach (following the decay of in vitro transcribed and transfected RNAs). While this approach certainly has its merits compared to other approaches, it also comes with some caveats: RNA is delivered naked, without bound RBPs and no nuclear history, e.g. of splicing (no EJCs), editing and modifications. One way to assess the generalizability of the results as well as the context dependence of the effects is to perform the same analysis on existing datasets of RNA stability measurements obtained through other methods (e.g. transcription inhibition). Are TA dinucleotides universally the most predictive feature of RNA half-lives?*

Our system studies the stability control of RNA synthesized in vitro and delivered into human cells. While we did not intend to generalize our conclusions to endogenous RNAs, our

approach contributes to the understanding of in vitro synthesized RNA used for cellular expression, such as in vaccines. It is known that endogenous RNAs undergo very different regulation. The most prominent factors controlling endogenous RNA stability are the density of splice junctions and the length of UTRs (doi.org/10.1186/s13059-022-02811-x; doi.org/10.1186/s12915-021-00949-x). To decipher the sequence regulation, these factors are controlled in our experiments. Therefore, we do not expect the dinucleotide features found by our approach to be generalized as the most predictive feature of RNA half-life in vivo.

> *The authors conclude their study with a meta-analysis of genes with increased TA dinucleotides in 5' and 3'UTRs, showing that specific functional groups are overrepresented among these genes. In addition, they provide evidence for an effect of disease-associated UTR mutations on endogenous RNA stability. While these elements link back to the original motivation of the study (screening for effects of point mutations in 5' and 3' UTRs), they provide only a limited amount of additional insights.*

We utilized the Taiwan Biobank to investigate whether mutations significantly affecting RNA stability also impact human biochemical measurements. Our findings indicate that these mutations indeed have a significant effect on various biochemical indices. This highlights the importance of our study, as it bridges basic science with potential applications in precision medicine. By linking specific UTR mutations with measurable changes in biochemical indices, our research underscores the potential for these findings to inform targeted medical interventions in the future.

> *In summary, this manuscript presents an interesting addition to the long-standing attempts at dissecting the sequence basis of RNA stability in human cells. The analysis is in general very comprehensive and sound; however, at times the goal of the authors to find novelty and specificity in the data overshadows some analyses. One example is the case where the authors try to show that TA-dinucleotides and GC content are decoupled and not merely two sides of the same coin.*
>
> *They claim that the effect of TA dinucleotides is different between high- and low-GC content contexts but do not control for the fact that low GC-content regions naturally will contain more TA dinucleotides and therefore the effect sizes and the resulting correlation between TA-diNT rate and stability will be stronger (Fig. 5A). A more thorough analysis and greater caution in some of the claims could further improve the credibility of the conclusions.*

Low GC content implies a higher UA content but does not directly equate to a high UA-dinucleotide ratio. For instance, the sequence AUUGAACCUU has a lower GC content (0.3) compared to UAUAGGCCGC (0.6), yet it also has a lower UA-dinucleotide ratio (0 vs. 0.22). To address this concern more rigorously, we performed a stratified analysis based on UA-diNT rate. As shown in our Fig. S7C, even after stratifying by UA- dinucleotide ratio (upper panel high UA- dinucleotide ratio / lower panel low UA- dinucleotide ratio), we still observe that the destabilizing effect of UA is stronger in the low GC content group.

> ***Reviewer #2 (Public Review):***
>
> *Summary of goals:*
>
> *Untranslated regions are key cis-regulatory elements that control mRNA stability, translation, and translocation. Through interactions with small RNAs and RNA binding proteins, UTRs form complex transcriptional circuitry that allows cells to fine-tune gene expression. Functional annotation of UTR variants has been very limited, and improvements could offer insights into disease relevant regulatory mechanisms. The*

> *goals were to advance our understanding of the determinants of UTR regulatory elements and characterize the effects of a set of "disease-relevant" UTR variants.*
>
> *Strengths:*
>
> *The use of a massively parallel reporter assay allowed for analysis of a substantial set (6,555 pairs) of 5' and 3' UTR fragments compiled from known disease associated variants. Two cell types were used.*
>
> *The findings confirm previous work about the importance of AREs, which helps show validity and adds some detailed comparisons of specific AU-rich motif effects in these two cell types.*
>
> *Using a Lasso regression, TA-dinucleotide content is identified as a strong regulator of RNA stability in a context dependent manner based on GC content and presence of RNA binding protein binding motifs. The findings have potential importance, drawing attention to a UTR feature that is not well characterized.*
>
> *The use of complementary datasets, including from half-life analyses of RNAs and from random sequence library MRPA's, is a useful addition and supports several important findings. The finding the TA dinucleotides have explanatory power separate from (and in some cases interacting with) GC content is valuable.*
>
> *The functional enrichment analysis suggests some new ideas about how UTRs may contribute to regulation of certain classes of genes.*
>
> *Weaknesses:*
>
> *It is difficult to understand how the calculations for half-life were performed. The sequencing approach measures the relative frequency of each sequence at each time point (less stable sequences become relatively less frequent after time 0, whereas more stable sequences become relatively more frequent after time 0). Since there is no discussion of whether the abundance of the transfected RNA population is referenced to some external standard (e.g., housekeeping RNAs), it is not clear how absolute (rather than relative) half-lives were determined.*

We estimated decay constant λ and half-life ($t_{1/2}$) by the following equations:

$$ln\left(\frac{C_{i(t)}}{C_{i(t=0)}}\right) = -\lambda t_i$$

$$t_{1/2} = \frac{ln\,(2)}{\lambda}$$

where $C_{i(t)}$ and $C_{i(t=0)}$ are read count values of the ith replicate at time points t and 0 (see also Methods). The absolute abundance was not required for the half-life calculation.

> *Fig. S1A and B are used to assess reproducibility. They show that read counts at a given time point correlate well across replicate experiments. However, this is not a good way to assess reproducibility or accuracy of the measurements of t1/2 are. (The major source of variability in read counts in these plots - especially at early time points - is likely the starting abundance of each RNA sequence, not stability.) This creates concerns about how well the method is measuring t1/2. Also creating concern is the observation that many RNAs are associated with half-lives that are much longer than the time points analyzed in the study. For example, based upon Figure S1 and Table S1 correctly, the median t1/2 for the 5' UTR library in HEK cells appears to be >700 minutes. Given that*

> *RNA was collected at 30, 75, and 120 minutes, accurate measurements of RNAs with such long half lives would seem to be very difficult.*

We estimated the half-life based on the following equations:

$$ln\left(\frac{C_{i(t)}}{C_{i(t=0)}}\right) = -\lambda t_i$$
$$t_{1/2} = \frac{ln\,(2)}{\lambda}$$

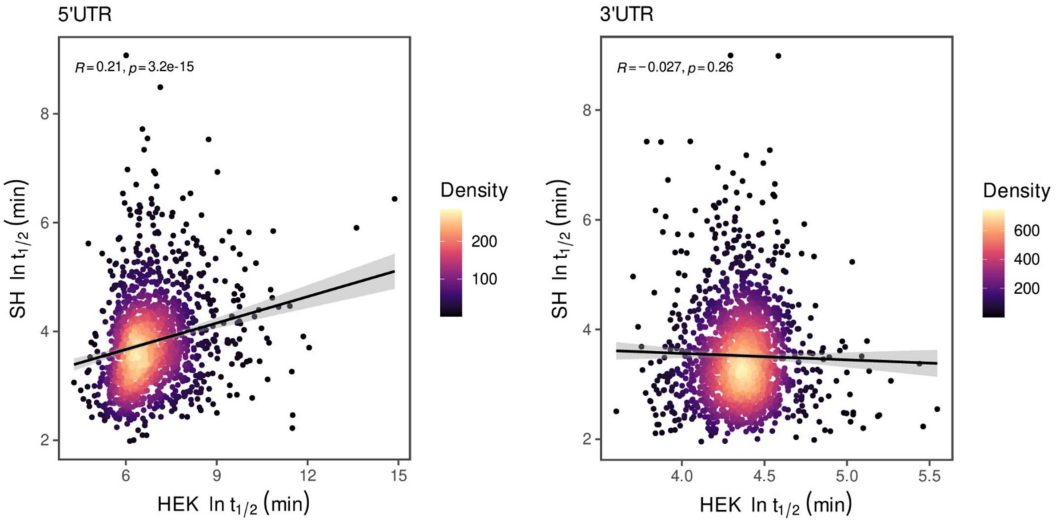
where $C_{i(t)}$ and $C_{i(t=0)}$ are read count values of the ith replicate at time points t and 0 (see also Methods). The calculation of the half-life involves first determining the decay constant λ, which represents a constant rate of decay. Since λ is a constant, it is possible to accurately calculate it without needing data over the entire decay range. Our experimental design considers this by selecting appropriate time points to ensure a reliable estimation of λ, and thus, the half-life. To determine the most suitable time points, we conducted preliminary experiments using RT-PCR.

These experiments indicated that 30, 75, and 120 minutes provided an effective range for capturing the decay dynamics of the transcripts.

> *There is no direct comparison of t1/2 between the two cell types studied for the full set of sequences studied. This would be helpful in understanding whether the regulatory effects of UTRs are generally similar across cell lines (as has been shown in some previous studies) or whether there are fundamental differences. The distribution of t1/2's is clearly quite different in the two cell lines, but it is important to know if this reflects generally slow RNA turnover in HEK cells or whether there are a large number of sequence-specific effects on stability between cell lines. A related issue is that it is not clear whether the relatively small number of significant variant effects detected in HEK cells versus SH-SY5Y cells is attributable to real biological differences between cell types or to technical issues (many fewer read counts and much longer half lives in HEK cells).*

For both cell lines, we selected oligonucleotides with $R^2 > 0.5$ and mean squared error (MSE) < 1 for analysis when estimating half-life (λ) by linear regression. This selection criterion was implemented to minimize the effect of experimental noise. After quality control, we selected common UTRs and compared the RNA half-lives of the two cell lines using a scatter plot. Author response image 2 shows that RNA half-lives are quite different between the cell lines, with a moderate similarity observed in the 5' UTRs (R = 0.21), while the correlation in the 3' UTRs is non-significant.

Author response image 2.



Despite the low correlation of mRNA half-life between the two cell lines, UA-dinucleotide and UA-rich sequences consistently emerge as the most significant destabilizing features, suggesting a shared regulatory mechanism across diverse cellular environments.

> *The general assertion is made in many places that TA dinucleotides are the most prominent destabilizing element in UTRs (e.g., in the title, the abstract, Fig. 4 legend, and on p. 12). This appears to be true for only one of the two cell lines tested based on Fig. 3.*

UA-dinucleotides and other UA-rich sequences exhibit similar effects on RNA stability, as illustrated in Fig. S5A-C. In two cell lines, UA-dinucleotide and WWWWWW sequences were representatives of the same stability-affecting cluster. While the impact of UA-dinucleotides can be generalized, we have rephrased some statements for clarification to avoid any potential misunderstanding. For examples:

Abstract: "...We found that UA dinucleotides and UA-rich motifs are the most prominent destabilizing element."

p.10: "UA dinucleotides and UA-rich motifs are the most common and effective RNA destabilizing factor"

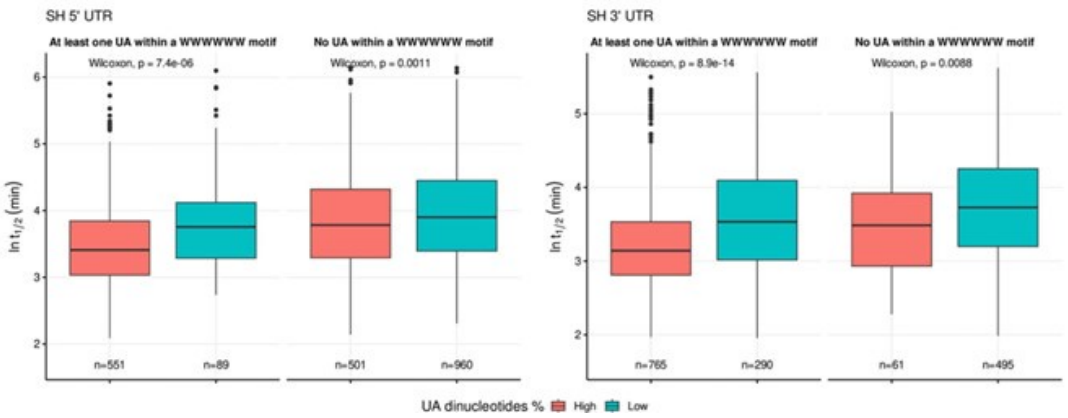
Figure 4: "The UTR UA dinucleotides and UA-rich motifs are the most common and influential RNA destabilizing factor."

> *Appraisal and impact:*
>
> *The work adds to existing studies that previously identified sequence features, including AREs and other RNA binding protein motifs, that regulate stability and puts a new emphasis on the role of "TA" (better "UA") dinucleotides. It is not clear how potential problems with the RNA stability measurements discussed above might influence the overall conclusions, which may limit the impact unless these can be addressed.*
>
> *It is difficult to understand whether the importance of TA dinucleotides is best explained by their occurrence in a related set of longer RBP binding motifs (see Fig 5J, these motifs may be encompassed by the "WWWWWW cluster") or whether some other explanation applies. Further discussion of this would be helpful. Does the LASSO method tend to*

collapse a more diverse set of longer motifs that are each relatively rare compared to the dinucleotide? It remains unclear whether TA dinucleotides are associated with less stability independent of the presence of the known larger WWWWWWW motif. As noted above, the importance of TA dinucleotides in the HEK experiments appears to be less than is implied in the text.

To ensure the representativeness of the features entered into the LASSO model, we pre-selected those with an occurrence greater than 10% among all UTRs. There is no evidence to support a preference for dinucleotides by LASSO. To address whether the destabilizing effect of UA dinucleotides is part of the broader WWWWWW motif, we divided UA dinucleotides into two groups: those within the WWWWWW motif and those outside of it. Specifically, we divided UTRs into two categories: 'at least one UA within a WWWWWW motif' and 'no UA within a WWWWWW motif,' and visualized the results using a boxplot. As shown in Author response image 3, the destabilizing trend still remains for UA dinucleotides outside of the WWWWWW motif, although the effect appears to be more pronounced when UA is within the WWWWWW motif. This suggests that while UA dinucleotides have a destabilizing effect independently, their impact is amplified when they are part of the broader WWWWWW motif.

Author response image 3.



The inclusion of more than a single cell type is an acknowledgement of the importance of evaluating cell type-specific effects. The work suggests a number of cell type-specific differences, but due to technical issues (especially with the HEK data, as outlined above) and the use of only two cell lines, it is difficult to understand cell type effects from the work.

The inclusion of both 3' and 5' UTR sequences distinguishes this work from most prior studies in the field. Contrasting the effects of these regions on stability is of interest, although the role of these UTRs (especially the 5' UTR) in translational regulation is not assessed here.

We examined the role of UTR and UTR variants in translation regulation using polysome profiling. By both univariate analysis and an elastic regression model, we identified motifs of short repeated sequences, including SRSF2 binding sites, as mutation hotspots that lead to aberrant translation. Furthermore, these polysome-shifting mutations had a considerable impact on RNA secondary structures, particularly in upstream AUG-containing 5’ UTRs. Integrating these features, our model achieved high accuracy (AUROC > 0.8) in predicting polysome-shifting mutations in the test dataset. Additionally, metagene analysis indicated that pathogenic variants were enriched at the upstream open reading frame (uORF)

translation start site, suggesting changes in uORF usage underlie the translation deficiencies caused by these mutations. Illustrating this, we demonstrated that a pathogenic mutation in the IRF6 5' UTR suppresses translation of the primary open reading frame by creating a uORF. Remarkably, site-directed ADAR editing of the mutant mRNA rescued this translation deficiency. Because the regulation of translation and stability does not converge, we illustrate these two mechanisms in two separate manuscripts (this one and doi.org/10.1101/2024.04.11.589132).

> ***Reviewer #3 (Public Review):***
>
> *Summary:*
>
> *In their manuscript titled "Multiplexed Assays of Human Disease-relevant Mutations Reveal UTR*
>
> *Dinucleotide Composition as a Major Determinant of RNA Stability" the authors aim to investigate the effect of sequence variations in 3'UTR and 5'UTRs on the stability of mRNAs in two different human cell lines.*
>
> *To do so, the authors use a massively parallel reporter assay (MPRA). They transfect cells with a set of mRNA reporters that contain sequence variants in their 3' or 5' UTRs, which were previously reported in human diseases. They follow their clearance from cells over time relative to the matching non-variant sequence. To analyze their results, they define a set of factors (RBP and miRNA binding sites, sequence features, secondary structure etc.) and test their association with differences in mRNA stability. For features with a significant association, they use clustering to select a subset of factors for LASSO regression and identify factors that affect mRNA stability.*
>
> *They conclude that the TA dinucleotide content of UTRs is the strongest destabilizing sequence feature. Within that context, elevated GC content and protein binding can protect susceptible mRNAs from degradation. They also show that TA dinucleotide content of UTRs affects native mRNA stability, and that it is associated with specific functional groups. Finally, they link disease associated sequence variants with differences in mRNA stability of reporters.*
>
> *Strengths:*
>
> *This work introduces a different MPRA approach to analyze the effect of genetic variants. While previous works in tissue culture use DNA transfections that require normalization for transcription efficiency, here the mRNA is directly introduced into cells at fixed amounts, allowing a more direct view of the mRNA regulation.*
>
> *The authors also introduce a unique analysis approach, which takes into account multiple factors that might affect mRNA stability. This approach allows them to identify general sequence features that affect mRNA stability beyond specific genetic variants, and reach important insights on mRNA stability regulation. Indeed, while the conclusions to genetic variants identified in this work are interesting, the main strength of the work involve general effect of sequence features rather than specific variants.*
>
> *The authors provide adequate supports for their claims, and validate their analysis using both their reporter data and native genes. For the main feature identified, TA di-nucleotides, they perform follow-up experiments with modified reporters that further strengthen their claims, and also validate the effect on native cellular transcripts (beyond reporters), demonstrating its validity also within native scenarios.*
>
> *The work provides a broad analysis of mRNA stability, across two mRNA regulatory segments (3'UTR and 5'UTR) and is performed in two separate cell-types. Comparison*

> *between two different cell-types is adequate, and the results demonstrate, as expected, the dependence of mRNA stability on the cellular context. Analysis of 3'UTR and 5'UTR regulatory effects also shows interesting differences and similarities between these two regulatory regions.*
>
> *Weaknesses:*
>
> *(1) The authors fail to acknowledge several possible confounding factors of their MPRA approach in the discussion.*
>
> *First, while transfection of mRNA directly into cells allows to avoid the need to normalize for differences in transcription, the introduction of naked mRNA molecules is different than native cellular mRNAs and could introduce biases due to differences in mRNA modifications, protein associations etc. that may occur co-transcriptionally.*
>
> *Second, along those lines, the authors also use in-vitro polyadenylation. The length of the polyA tail of the transfected transcripts could potentially be very different than that of native mRNAs and also affect stability.*

The transcripts used in our study were polyadenylated in vitro with approximately 100 nucleotides

(Fig. S1C), similar to the polyA tail lengths typically observed in vivo (dx.doi.org/10.1016/j.molcel.2014.02.007). Additionally, these transcripts were capped to emulate essential mRNA characteristics and to minimize immune responses in recipient cells. This design allows us to study RNA decay for in vitro-synthesized RNA delivered into human cells, akin to RNA vaccines, but it does not necessarily extend to endogenous RNAs. As mentioned, endogenous RNAs undergo nuclear processing and are decorated by numerous trans factors, resulting in distinct regulatory mechanisms. We therefore provided a more discussion on these differences and their implications in the revised manuscript: “However, while our approach effectively assesses the stability of synthesized RNA in human cells, it may not fully capture the decay dynamics of nuclear-synthesized RNA, which can be influenced by endogenous modifications and trans-acting RNA binding factors. (p. 18)”

> *(2) The analysis approach used in this work for identifying regulatory features in UTRs was not previously used. As such, lack of in-depth details of the methodology, and possibly also more general validation of the approach, is a drawback in convincing the reader in the validity of this approach and its results.*
>
> *In particular, a main point that is not addressed is how the authors decide on the set of "factors" used in their analysis? As choosing different sets of factors might affect the results of the analysis.*

In our study, we employed the calculation of the Variance Inflation Factor (VIF) as a basis for selecting variables. This well-established method is widely used to detect variables with high collinearity, thus ensuring the robustness and reliability of our analysis. By identifying and excluding highly collinear variables, we aimed to minimize multicollinearity and improve the accuracy of our regression models. For more detailed information on the use of VIF in regression analysis, please refer to Akinwande, M., Dikko, H., and Samson, A. (2015). Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. Open Journal of Statistics, 5, 754-767. doi: 10.4236/ojs.2015.57075. We have included the method details in the revised manuscript (p. 28) :”... to avoid multicollinearity caused by similar features that perturb feature selection, all features were clustered using single-linkage hierarchical clustering with the distance metric defined as one minus the absolute value of the Spearman correlation coefficient. We cut the tree at a specific height, and the feature that had the greatest influence on RNA stability, which was examined using a simple

linear regression model, was selected to be the representative of each cluster. Then we calculated the variance inflation factor (VIF) value of the representative features. The VIFs were obtained by the following linear model and equations:

$$\hat{X}_{ij} = \hat{\alpha}_{0(j)} + \sum_{k \neq j} \hat{\alpha}_{k(j)} X_{ik}$$

$$R_j^2 = \frac{\sum_i \left(\hat{X}_{ij} - \bar{X}_{\cdot j}\right)^2}{\sum_i \left(X_{ij} - \bar{X}_{\cdot j}\right)^2}$$

$$VIF_j = \frac{1}{1 - R_j^2}$$

where $\hat{X}_{ij}$ and X_{ik} are the estimated value of the jth feature and the value of the kth feature of the ith UTR (note that the kth feature is a feature other than the jth feature), $\hat{\alpha}_{0(j)}$ and $\hat{\alpha}_{k(j)}$ are the intercept and the regression coefficients of the linear model that regressed the jth feature on the other remaining features, and $\bar{X}_{\cdot j}$ is the mean level of the jth feature of all UTRs."

> *For example, the choice to use 7-mer sequences within the factors set is not explained, particularly when almost all motifs that are eventually identified (Figure 3B-E) are shorter.*

The known RBP motifs are primarily 6-mer. To explore the possibility of discovering novel motifs that could significantly impact our model, we started with 7-mer sequences. However, our analysis revealed that including these additional variables did not improve the explanatory power of the model; instead, it reduced it. Consequently, our final model focuses on motifs shorter than 7-mer. We explained the motif selections in the revised manuscript (p. 9): "Given our discovery that the effect of AREs is heavily dependent on sequence content, we decided to further explore the effects of other sequence elements, i.e., beyond known regulatory motifs, in more detail. Since most reported RBP motifs are 6-mers, we initiated a search for novel motifs by analyzing the presence of all 7-mers in our massively parallel reporter assay (MPRA) library, correlating their occurrence with mRNA half-life."

> *In addition, the authors do not perform validations to demonstrate the validity of their approach on simulated data or well-established control datasets. Such analysis would be helpful to further convince the reader in the usefulness and robustness of the analysis.*

We acknowledge the importance of validating our approach on simulated data or well-established control datasets to demonstrate its robustness and reliability. However, to the best of our knowledge, there are currently no well-established control datasets available that perfectly correspond to our specific study context. Despite this, we will continue to search for any relevant datasets that could be utilized for this purpose in future work. This effort will help to further reinforce the confidence in our methodology and its findings.

> *(3) The analysis and regression models built in this work are not thoroughly investigated relative to native genes within cells. The effect of sequence "factors" on native cellular transcripts' stability is not investigated beyond TA di-nucleotides, and it is unclear to what degree do other predicted factors also affect native transcripts.*

Our system studies the stability control of RNA synthesized in vitro and delivered into human cells. While we validated the UTR UA-dinucleotide effect in vivo, we did not intend to conclude that this is the most influential regulation for endogenous RNAs. It is known that endogenous RNAs undergo very different regulation. The most prominent factors controlling endogenous RNA stability are the density of splice junctions and the length of UTRs (doi.org/10.1186/s13059-022-02811-x; doi.org/10.1186/s12915-021-00949-x). To decipher the sequence regulation, we controlled for these factors in our experiments. Therefore, we acknowledge that several endogenous features, which were excluded by our approach, may serve as predictive features of RNA half-life in vivo.

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations For The Authors):***
>
> *Specific comments:*
>
> *Some references are missing, e.g for the sentence:*

Please see the response below.

"Similarly, point mutation of the GFPT1 3' UTR results in congenital myasthenic syndrome." (p5)

The reference has been added to the text:

Dusl, M., Senderek, J., Muller, J. S., Vogel, J. G., Pertl, A., Stucka, R., Lochmuller, H., David, R., & Abicht, A. (2015). A 3'-UTR mutation creates a microRNA target site in the GFPT1 gene of patients with congenital myasthenic syndrome. Human Molecular Genetics, 24(12), 34183426. https://doi.org/10.1093/hmg/ddv090

> *"...but there have been no systematic assessments of the explicit effects of variants of both UTRs on stability regulation." (not true in the current phrasing; e.g. PMIDs 32719458, 36156153, 34849835)*

These references have been added to the text. However, we have to point out that these studies do not focus on the effects of the disease-relevant variants. To clarify, we modified the sentence to "... systematic assessments of the explicit effects of disease-relevant variants in both UTRs on stability regulation are still absent."

> *"Multiple approaches have revealed AREs as exerting a destabilizing effect on RNA stability (Barreau et al., 2005). (p8)*

The reference has been added to the text:

Barreau, C., Paillard, L., & Osborne, H. B. (2005). AU-rich elements and associated factors: are there unifying principles? Nucleic Acids Research, 33(22), 7138-7150. https://doi.org/10.1093/nar/gki1012

> *"This effect is specific, as such ratios in the coding region are inconsequential." (p12)*

This refers to our findings of Fig. 4G and Supplemental Fig. S5F.

> *What are the sequences at the 5' and 3'UTR without insertion of a library? 5'UTR library (especially in SH) has much longer half-life compared to 3'utr library (Fig S1D).*

There is no designed 5'UTR of the 3'UTR library, only the Kozak sequence derived from the pEGFPC1 vector. This may partially underlie the shorter half-life of the 3' UTR library.

> *Fig2A: What are the units? "half-life (log)" Do the numbers correspond to log10(min)?*

It represents ln (min). To clarify, we now use 'ln $t_{1/2}$ (min)' in all figures.

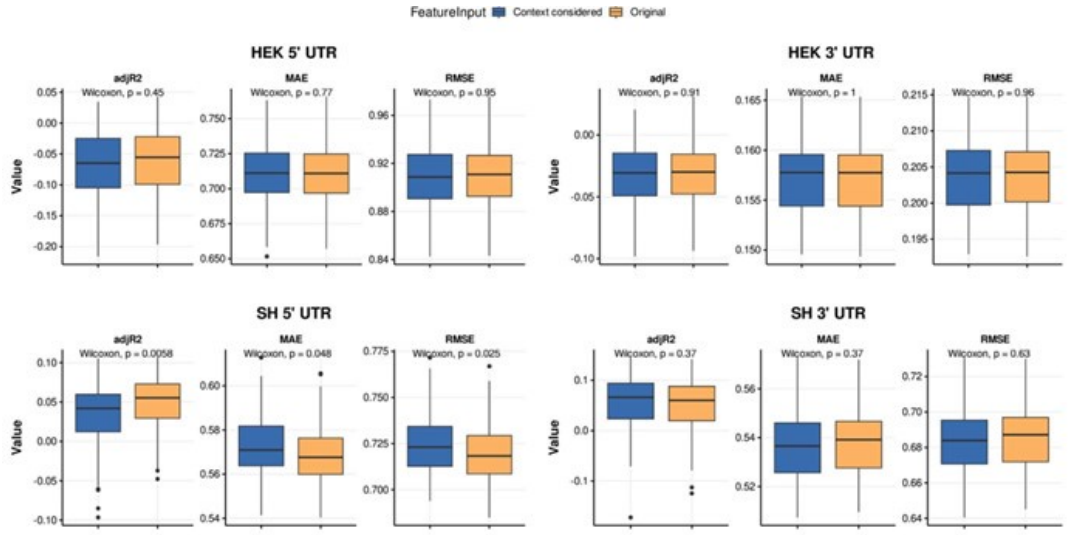
> *Fig 2 and 3: This was done only on the wild-type sequences? Or all tested sequences together, wt and mut?*

It was done only on the wild-type sequences. To clarify, we modified the text to "we examined the effect of AREs on RNA stability of the ref alleles according to specific sequence content.... (p.8)" and "We considered as many factors as possible to explain the half-life of our ref UTR libraries,.... (p.9)". 'ref' stands for reference.

> *"Furthermore, to avoid collinearity confounding our model, e.g., the effects of very similar factors (such as 'AA' and 'AAA' sequences), we clustered the factors according to their properties, and then only one representative factor from within a cluster (i.e., the one with the highest correlation to halflife within a cluster) was subjected to LASSO regression": Given the observed context dependence, e.g. in the case of poly-U stretches: Isn't this clustering leading to similar/identical motifs with different context being grouped together (such as polyU preceded by an A (strongly destabilizing, according to Fig 2B) or followed by one (strongly stabilizing, according to Fig 2B), resulting in ignoring the context or using one potential outcome while a motif from the same cluster can have the opposite effect?*

Thank you very much for pointing this out. To determine if considering different contextual effects within each feature cluster would enhance model performance, we modified our feature selection by choosing both the feature with the largest positive and the largest negative effect on RNA half-life in Step III of Figure 3A. We then split the data into a 2:1 training and testing set and repeated this process 100 times. Model performance was evaluated using mean average error (MAE), root mean squared error (RMSE), and adjusted R-squared. From Author response image 4, we observed no significant improvement in model performance using this new approach. Notably, in the SH-SY5Y 5' UTR model, our original method even outperformed the modified one, with statistically lower MAE and RMSE and a higher adjusted R-squared. Therefore, we believe our current approach remains appropriate.

Author response image 4.



> *"Overall, motifs that are at least two nucleotides long proved critical for RNA stability, supporting the sequence specificity of the decay process." Unclear why this supports the "sequence specificity"*

No monomers were selected as an explanatory factor. On the contrary, specific sequence combinations and order are important for the regulation. These findings suggest sequence-specific recognition for the decay process.

> *Fig3: The same features were used in both cell lines? If yes: Since they were selected for their highest correlation with half-life, how was a common set chosen? If no: problematic to compare.*

Thank you for your question regarding feature selection across cell lines. Initially, the features were collected uniformly for both cell lines. However, subsequent feature selection steps were cell-type specific, focusing on identifying features with the greatest impact on RNA half-life in each context. This approach allows us to still compare model performance and discuss the similarities and differences in selected features across cell types. By maintaining a consistent starting point, we ensure that any observed differences reflect cell-specific regulatory dynamics.

> *uORFs were not used as features?*

Thank you for pointing this out. At the beginning of our study, we investigated the impact of Kozak sequence strength (categorized as weak, moderate, strong, or optimal) on RNA half-life. However, we found that this feature performed poorly in predicting RNA stability, and as a result, we decided not to include upstream open reading frames (uORFs) or Kozak sequences in our subsequent analyses.

> *Experimental reproducibility: Only correlations between replicates for the same time point is shown, but no comparison between time points or between decay rates. How reproducible were the paired differences between mut/wt?*

The decay rate was calculated by modeling the slope of a linear regression of all time points. Therefore, there is only one decay rate associated with a genotype. To rule out inconsistent data, we excluded any regression with a mean square error greater than 1, as this indicates a poor fit of the data points.

> *Fig 7C/p17: This does not establish a "causal relationship" as the authors claim.*

We agree with the reviewer's suggestion. We have modified the text on p.17 to "to establish a correlation between UTR variants and health outcomes,….."

> *In the discussion, the authors claim that TA-diNTs are not only an opposite of the GC percentage and base this on Fig 5A.*

Fig 5A: The range of TA-diNTs is naturally much higher in the low GC group. To make the high and low GC content comparable (as the authors aim to do), the correlation should be assessed for the same range of TA dint in both cases.

To address this concern more rigorously, we performed a stratified analysis based on UA-diNT rate. As shown in our Fig. S7C, even after stratifying by UA- dinucleotide ratio (upper panel high UA- dinucleotide ratio / lower panel low UA- dinucleotide ratio), we still observe that the destabilizing effect of UA is stronger in the low GC content group.

Supplemental Figure S7. Interplay of GC content and TA dinucleotide on stability regulation, related to Figure 5. (C) Stratifications of both TA dinucleotide ratio and GC content showed that the destabilizing effect of TA dinucleotide is the most prominent under conditions of low TA dinucleotide ratio and low GC content. The same trend was observed for 5' UTR (left) and 3' UTR (right).

> *The injection of in vitro transcribed and polyA/capped RNA certainly has advantages over other methods, but delivering naked mRNA without nuclear history might also lead to artifacts. The caveats of the approach should be discussed more extensively.*

We appreciate the suggestion and have hence added the following in the Discussion (p.18): "However, while our approach effectively assesses the stability of synthesized RNA in human cells, it may not fully capture the decay dynamics of nuclear-synthesized RNA, which can be influenced by endogenous modifications and trans-acting RNA binding factors."

> *"We unexpectedly identified many crucial regulatory features in 5' UTRs." Why was this unexpected?*

We initially thought the 3' UTR would play a major role in stability regulation. To avoid confusion, we have removed the word 'unexpected' from the text (p. 20): "We identified many crucial regulatory features in 5' UTRs."

> *"...a massively parallel reporter assay in which coding regions and human 5'/3' UTRs with diseaserelevant mutations were generated in vitro and then directly transfected into human cell lines to assess their decay patterns by next-generation sequencing": also coding regions?*

Thanks for the question. Indeed, the coding region was not synthesized together with the UTR library. Therefore, we modified the text of p. 6 to "…we developed a massively parallel reporter assay in which human 5'/3' UTRs with disease-relevant mutations were generated in vitro, ligated with the enhanced green fluorescence protein (EGFP) coding region, and then

directly transfected into human cell lines to assess their decay patterns by next-generation sequencing."

> ***Reviewer #2 (Recommendations For The Authors):***
>
> *Nomenclature: When discussing RNA sequences, "U" should be used in place of "T" (e.g., "UA dinucleotide").*

We have replaced the RNA sequence "T" with "U" of the text and figures.

> *Abstract: "We examined the RNA degradation patterns mediated by the UTR library in multiple cell lines" - It would be clearer to state that two cell lines (rather than multiple) were used.*

We appreciate the suggestion. We have modified the abstract as suggested: "We examined the RNA degradation patterns mediated by the UTR library in two cell lines…"

> *The manuscript refers to "wild-type (WT) and mutant (mt) alleles." (p. 7 and elsewhere). It would be better to use "reference" instead of "wild type" given that these are human populations.*

We appreciate the suggestion. All instances of 'wild-type' or 'WT' in the text and figures have been replaced with 'reference' or 'ref'.

> *In the introduction, it is stated that traditional MPRAs "cannot differentiate the effect of the UTRs on transcription, stability and, in some cases, even protein production, greatly limiting scientific interpretation." This is confusing, since these assays can and have been used in association with both RNA decay measurements and measurements of reporter protein levels that allow assessment of effects on stability and protein production (including in the cited references).*

We reason that the RNA steady-state level (e.g., sequencing the overall RNA normalized to DNA) or protein steady-state level (e.g., detecting the fluorescence signal) does not precisely reveal the decay kinetics of the RNA. Steady-state level is a result of production and decay, both of which UTRs contribute to. Similarly, the protein level is not a perfect estimate of the RNA decay.

To clarify, we have modified the introduction (p. 5) to "Nevertheless, because the steady-state level is a result of production and decay, these approaches cannot differentiate the effect of the UTRs on transcription, stability and, in some cases, even protein production, greatly limiting scientific interpretation."

Adding raw and normalized read count data from individual experiments (e.g., to Table S1) would make it more likely for others to use this dataset to address additional questions.

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE217518 (reviewer token snspaakujtsdpcv).

> *The manuscript would benefit from further clarification about model selection. Additional details regarding how the features were clustered, and the actual clusters themselves should be included.*
>
> *It should be discussed why Lasso was chosen vs Ridge or Elastic Net, in the context of handling multicollinearity. Often, data is subsetted for training and validation, and model performance metrics are presented.*

Thank you for pointing out the need for further clarification on model selection. The features were clustered using single-linkage hierarchical clustering with the distance metric defined as one minus the absolute value of the Spearman correlation coefficient (this information has been added to the manuscript on p. 28: "…to avoid multicollinearity caused by similar features that perturb feature selection, all features were clustered using single-linkage hierarchical clustering with the distance metric defined as one minus the absolute value of the Spearman correlation coefficient."). The resulting feature clusters are available in Supplemental Table S3.

Regarding model selection, we chose LASSO over ridge and elastic net primarily for feature selection, as ridge does not perform feature selection. Elastic net is essentially a hybrid of ridge regression and LASSO regularization, but we opted for LASSO for its simplicity and effectiveness in selecting a sparse set of important features.

We also performed a 2:1 training and testing set analysis and have included these details in the manuscript. Model performance metrics, including correlation coefficient between observed and predicted values in the testing set, mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and R-squared, are provided in new Supplemental Table S4.

> *Recommend reviewing and correcting verb tenses in the methods section.*

We appreciate the reviewer's suggestion. We have corrected verb tenses in the methods section, which includes "The UTRs were defined by NCBI RefSeq and ENCODE V27. (p.21)", "The variant was placed in the middle of the sequence….(p.22)", and "eCLIP signals with value < 1 or p value > 0.05 were removed. (p.26)"

> *Please add information about which cell type(s) are being used in each of the figure legends (e.g., in Figs. 2B and 5).*

We appreciate the reviewer's suggestion. We have added the cell type information in the figure legends: "Figure 2…. (B) The ten most influential AREs in terms of RNA stability in SH-SY5Y cells." And "Figure 5…..(A) MPRA data of SH-SY5Y cells stratified according to the GC content (GC%) of UTRs."

> *Recommend review of axis labels and consistency in formatting the log(half-lives) and including the base of the log and the time unit (minutes). Even better, converting axis labels from log minutes to minutes would make this easier to understand.*

Thank you for the suggestion regarding axis labels and consistency. We have unified the half-life label to 'ln $t_{1/2}$ (min)' in all figures. We chose not to convert the axis from logarithmic minutes to minutes because the original scale is highly skewed, which would hinder clear data visualization.

> *The discussion refers to Figure 1D but Figure 1 only has A-C*

Thank you for pointing out this mistake. 'Fig. 1D' has been changed to 'Fig. 1B' in the text (p. 7 and p. 20).

> *The analyses in Fig. 2 are interpreted as demonstrating that AREs destabilize RNAs. These analyses are examining associations, so it would be more appropriate to say that AREs are associated with destabilization (since it is formally possible that other sequences that are present in these UTR fragment cause destabilization). A similar issue arises on p. 10: "TA dinucleotides alone can negatively regulate RNA stability, with a Pearson's*

> *correlation coefficient of -0.287 for 5' UTRs and -0.377 for 3' UTRs (Fig. 4A,C)." This is an association and does not establish causation. Again on p. 17: "We identified several SNPs in UTRs that induce aberrant RNA expression and/or protein expression (Supplemental Table S7)." These may be causal but may simply be in LD with other variants that are causal.*

We agree that the association observed is not proven to be causal. Therefore, we modified the text as suggested:

“AUUUA/AUUA-containing AREs are associated with RNA destabilization.” (p. 8)

“UA dinucleotides alone present a negative correlation with RNA stability, with a Pearson’s correlation coefficient of -0.287 for 5’ UTRs and -0.377 for 3’ UTRs.” (p.10)

“We identified several SNPs in UTRs that correlated with aberrant RNA expression and/or protein expression.” (p. 17)

> *Figure 4C is important in that it examines whether variant sequences that differ in a manner that changes the number of dinucleotide repeats affect stability. Please show the number (not just the percentage) of sequences in each category.*

Thank you for your insightful comment. We believe the figure you referred to is Figure 4E. We have updated the figure to include the number of sequences in each category.

> *Figure 6A and B: The horizontal axes appear to be misaligned since the dotted vertical lines do not cross at 0. ?*

The dotted vertical lines represent the genomic background of the UA-diNT ratio. To clarify it, we have modified the legend to: “Figure 6……(A) The top ten biological processes for which the 5’ UTR UA-dinucleotide ratio most significantly deviated from the genomic background (dashed line).”

> *It may be helpful to state what the dashed and solid lines represent on Figure 6 E/F. Please correct spelling of "Biological" in 6E.*

As per the reviewer’s suggestions, we have modified the legend of Figure 6 to: “……..(E) Biological processes for RNAs in which the UA-dinucleotide ratios of both 5’ and 3’ UTRs are significantly different from the genomic background (dashed lines). (F) Molecular functions for RNAs in which the UA-dinucleotide ratios of both 5’ and 3’ UTRs are significantly different from the genomic background (dashed lines). The thin solid lines represent the standard deviation of the UAdinucleotide ratio within the gene group.”

In addition, the spelling of “Biological” in Fig. 6E has been corrected.

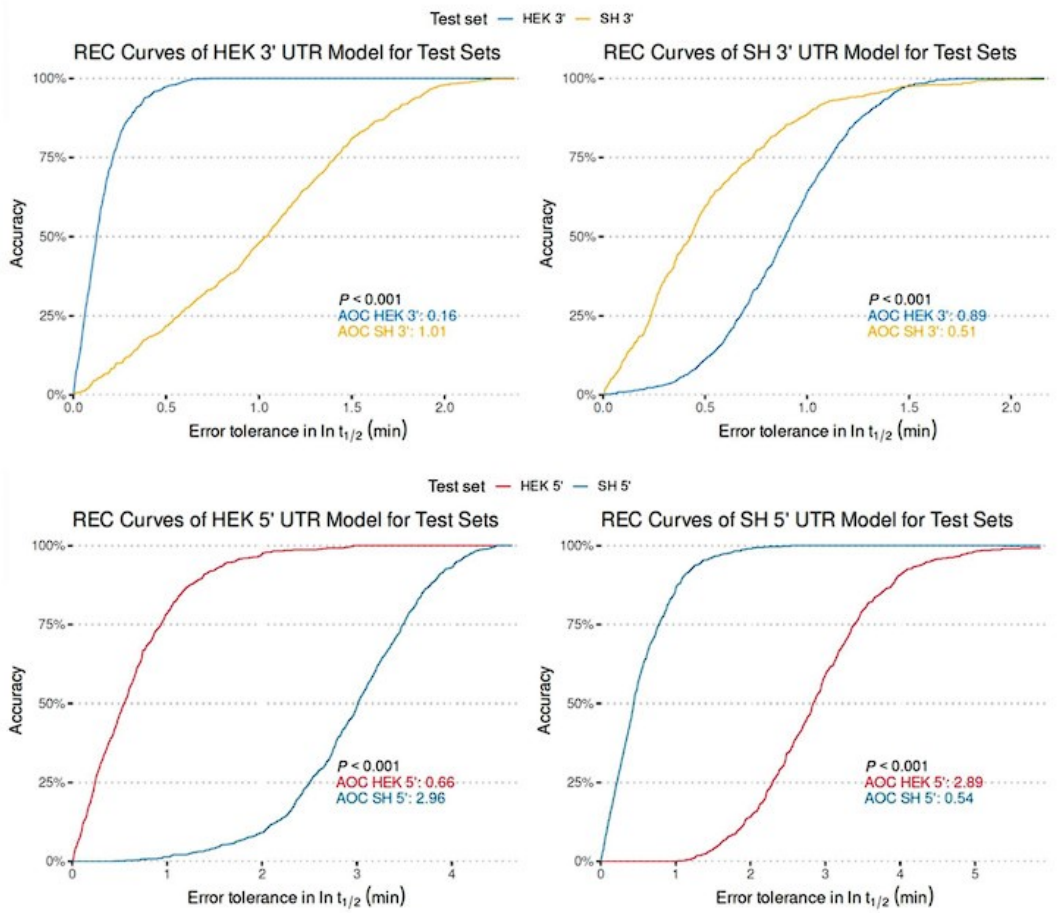
> ***Reviewer #3 (Recommendations For The Authors):***
>
> *I have 3 points that I think could improve science and its presentation within the manuscript.*
>
> *(1) Most importantly, how well do LASSO regression models predict the stability of native transcripts? Such analysis can also be useful for comparison between two different cell-types. How well does the regression model learned (on reporters) within one cell-type predict mRNA stability (of reporters and native genes) in this cell-type and in the other cell-type? Similarly, models can also help to analyze the effects of 5'UTR and 3'UTR sequences on mRNA stability. In particular, how well does the regression model of each separate regulatory sequence (3'UTR or 5'UTR) is able to predict the stability of native*

> *genes in the cell? Can the predictions be improved by combining both 3'UTR and 5'UTR sequence features within the regression models?*

The decay model for native transcripts has been established in prior research (doi.org/10.1186/s13059-022-02811-x; doi.org/10.1186/s12915-021-00949-x), which indicates that exon junction density and transcript length are the primary determinants of RNA stability. Based on these findings, we designed the MPRA with fixed length and without splicing to focus on the contribution of primary sequences. We validated the destabilizing effect of UA dinucleotide on endogenous RNAs (Fig. 4G and Supplemental Fig. S5F) but do not recommend using our model to fully explain or predict the stability of native transcripts.

To assess the model's cross-cell type predictive performance for RNA half-life, we employed the Regression Error Characteristic (REC) curve (Bi & Bennett, 2003). Similar to the receiver operating characteristic (ROC) curve, the REC curve illustrates the trade-off between error tolerance and accuracy, with better performance indicated by curves trending toward the upper left. We also computed the Area Over the Curve (AOC) as a performance metric, where lower values indicate better predictive ability. From Author response image 5, the REC curves reveal that cross-cell type prediction performance is suboptimal. The y-axis represents prediction accuracy, while the x-axis denotes error tolerance for the natural logarithm of RNA half-life ($\ln(t_{1/2})$, in minutes).

Author response image 5.



In response to the suggestion of combining 5' and 3' UTR sequence features in the regression model, we believe this approach may not be ideal. As shown in Figure S1D, the distribution of

RNA half-lives between 5' and 3' UTRs is significantly different, reflecting their distinct regulatory roles. Additionally, the base composition differs, with 5' UTRs having a higher GC content compared to 3' UTRs. Combining these datasets would likely make the origin of the sequence (5' or 3' UTR) the most predictive feature, thereby reducing the model's interpretability. Furthermore, our MPRA results, derived from separate 5' or 3' UTR library, do not support a combined model, further suggesting this approach may not be suitable with our data.

> *The conclusions regarding genetic variants are interesting, yet the main strength of the work involves identifying general sequence features that affect mRNA stability rather than specific variants. I wonder if the authors have considered to shift the focus of the motivation part to reflect that?*

We appreciated the reviewer's suggestion. We have revised the abstract and introductions to emphasize the general UTR regulation. Here is the revised abstract:

UTRs contain crucial regulatory elements for RNA stability, translation and localization, so their integrity is indispensable for gene expression. Approximately 3.7% of genetic variants associated with diseases occur in UTRs, yet a comprehensive understanding of UTR variant functions remains limited due to inefficient experimental and computational assessment methods. To systematically evaluate the effects of UTR variants on RNA stability, we established a massively parallel reporter assay on 6,555 UTR variants reported in human disease databases. We examined the RNA degradation patterns mediated by the UTR library in two cell lines, and then applied LASSO regression to model the influential regulators of RNA stability. We found that UA dinucleotides and UA-rich motifs are the most prominent destabilizing element. Gain of UA dinucleotide outlined mutant UTRs with reduced stability. Studies on endogenous transcripts indicate that high UA-dinucleotide ratios in UTRs promote RNA degradation. Conversely, elevated GC content and protein binding on UA dinucleotides protect high-UA RNA from degradation. Further analysis reveals polarized roles of UA-dinucleotide-binding proteins in RNA protection and degradation. Furthermore, the UA-dinucleotide ratio of both UTRs is a common characteristic of genes in innate immune response pathways, implying a coordinated stability regulation through UTRs at the transcriptomic level. We also demonstrate that stability-altering UTRs are associated with changes in biobank-based health indices, underscoring the importance of precise UTR regulation for wellness. Our study highlights the importance of RNA stability regulation through UTR primary sequences, paving the way for further exploration of their implications in gene networks and precision medicine.

> *Plots presenting correlations (e.g., Figure 4A, 4C) are more informative when plotted as density plots (i.e., using colorscale to show density of the dots at each part of the plot).*

We greatly appreciate the reviewer's insightful suggestion regarding the use of density plots for presenting correlations. We have modified Figures 4A and 4C in the revised manuscript to implement density plotting. The updated figures now utilize a colorscale that highlights areas of high and low data density.

https://doi.org/10.7554/eLife.97682.2.sa0