Title:

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

Author list:

Jia-Ying Su[1-5], Yun-Lin Wang[1,5], Yu-Tung Hsieh[1], Yu-Chi Chang[1], Cheng-Han Yang[1], YoonSoon Kang[1], Yen-Tsung Huang[2], Chien-Ling Lin[1†]

Affiliations:

[1]Institute of Molecular Biology, Academia Sinica, Taipei, 115 Taiwan

[2]Institute of Statistical Science, Academia Sinica, Taipei, 115 Taiwan

[3]Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, 115 Taiwan

[4]Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, 112 Taiwan

[5]These authors contributed equally

[†]To whom correspondence should be addressed.

Email: mbcllin@gate.sinica.edu.tw

Fax: 886-2-2782-6085

Present Address: Institute of Molecular Biology, Academia Sinica, No. 128, Sec. 2, Academia Road, Nangang District, Taipei City 115014 TAIWAN

Running Title: High TA-diNT Reduces RNA Stability in human cells

Keywords: untranslated region, RNA stability, UTR variants, massively parallel reporter assay, statistical learning

ABSTRACT

UTRs contain crucial regulatory elements for RNA stability, translation and localization, so their integrity is indispensable for gene expression. It has been estimated that ~3.7% of disease-associated genetic variants are located in UTRs. However, functional interpretation of UTR variants is largely incomplete because efficient means of experimental or computational assessment are lacking. 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 multiple cell lines, and then applied LASSO regression to model the influential regulators of RNA stability. We found that TA dinucleotides are the most prominent destabilizing element. Gain of TA dinucleotide outlined mutant UTRs with reduced stability. Studies on endogenous transcripts indicate that high TA-dinucleotide ratios in UTRs promote RNA degradation. Conversely, elevated GC content and protein binding on TA dinucleotides protect high-TA RNA from degradation. Further analysis reveals polarized roles of TA-dinucleotide-binding proteins in RNA protection and degradation. Furthermore, the TA-dinucleotide ratio of both UTRs is a common characteristic of genes in innate immune response pathways, implying that the global transcriptomic regulon involves stability coordination via UTRs. We also demonstrate that stability-altering UTRs are associated with changes in biobank-based health indices, providing evidence that UTR-mediated RNA stability contributes to establishing robust gene networks and potentially enabling disease-associated UTR variants to be classified for 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 (Pesole et al., 2001), 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).

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.

Multiplexed reporter assays to elucidate UTR stability control

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). Since post-transcriptional regulation is an extremely sophisticated network mechanism, regulatory features deduced from endogenous steady-state RNA levels can be camouflaged by other dominant factors. For example, studies of cellular RNA degradation followed by transcriptional inhibition have revealed that the length of coding regions and ribosome occupancy are major RNA stability determinants (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, 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).

Although the link between UTR variants and disease is evident, the effects of 5' and 3' UTR sequence variation on RNA stability regulation remains unclear. Efforts have been made to examine how 3' UTR variants impact steady-state RNA levels, but there have been no systematic assessments of the explicit effects of variants of both UTRs on stability regulation. To examine potentially pathogenic UTR mutations and their links to RNA stability, we developed a massively parallel reporter assay in which coding regions and human 5'/3' UTRs with disease-relevant mutations were generated *in vitro* 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 TA 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 GFP 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 wild-type (WT) 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 ($t_{1/2}$) 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 WT 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. 1D,E; 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 WT 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 according to specific sequence content. Based on the definition of AREsite2 (http://nibiru.tbi.univie.ac.at/AREsite2), we categorized AREs as either WTTTW or its longer derivatives, TTTGTTT or AWTAAA (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, ATTTA/ATTA-containing AREs destabilized RNA when present in either UTR type, whereas in SH-SY5Y cells, extremely T-dominant AREs ($T_{8\text{-}10}A_{1\text{-}2}$) stabilized it **(Fig. 2B)**, similar to the stabilizing effect of T(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. To do so, we examined the presence of all 7-mers in our MPRA library against 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 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, GT-rich sequences in 5' UTRs stabilized RNAs **(Fig. 3B,D)**. In contrast, CA- and TG-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, TA dinucleotides and other TA-rich sequences—such as WWWWWW (W=A/T; a potential binding motif of Peptidylprolyl Isomerase E (PPIE)), ATTTA (a potential binding motif of ELAV Like RNA Binding Protein 1 (ELAVL1)), and TTTATA (a potential binding motif of hnRNPA1)—are strongly destabilizing **(Figs. 3B-E & 4A,C)**. TA 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, TA dinucleotides in both UTRs overwhelmed other factors in robustly destabilizing RNAs in SH-SY5Y cells **(Fig. 3D,E)**. Therefore, TA dinucleotides seem to be a universal destabilizing motif, so we investigated how TA dinucleotides regulate RNA stability in further detail.

TA dinucleotides are the most common and effective RNA destabilizing factor

TA dinucleotides proved to be the strongest stability determinant for both UTR types in SH-SY5Y cells. 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)**. TA-rich motifs (in the same cluster as TA dinucleotides) behave similarly to TA dinucleotides in regulating RNA stability, whereas GC-rich motifs have the opposite effect **(Fig. 4B,D; Supplemental Fig. S5A,B)**. Within the same cluster, TA 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 TA-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 TA dinucleotide/WWWWWW cluster **(Supplemental Fig. S5C)**, indicating a universal destabilizing effect of TA-rich sequences. Next,

to examine if there is a region-specific effect of TA and closely-related AT dinucleotides, we used a sliding window to establish the localization-associated relationship between the TA/AT dinucleotide ratio and RNA half-life. Correlation coefficients between TA/AT dinucleotide ratios and UTR stability were calculated for each window, and we assumed that regions displaying a strong correlation between TA/AT dinucleotide ratios and stability rank hosted TA/AT dinucleotides that control RNA stability **(Supplemental Fig. S5D)**. We found that TA/AT 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 TA dinucleotides are a common and prominent RNA destabilizing motif.

Next, we examined the effect of mutating the most effective destabilizing TA 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 TA dinucleotide mutations, compared to stabilizing or non-significant mutations **(Fig. 4E)**. To further validate the impact of TA-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 TA-dinucleotide, resulting in a discrepancy of one TA-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 TA-dinucleotide resulted in the destabilization of RNAs **(Fig. 4F**, Left: 0 to 1 TA-dinucleotide; Right: 1 to 2 TA-dinucleotides**)**. Moreover, to explore the influence of the TA-destabilizing effect on endogenous mRNA stability, we assessed the TA-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 TA-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 TA 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 wild-type counterparts are with significantly longer or shorter RNA half-life, respectively **(Supplemental Fig. S6A)**. This intrinsic difference in effects on RNA stability for WT UTRs was observed for both cell lines and it was amplified in their mutant counterparts **(Supplemental Fig. S6B,C)**. A motif analysis revealed that WT 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.

GC content RBP- and ribosome-binding hinders the destabilizing effect of TA 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 TA dinucleotides are strongly associated with RNA half-life, with the 3' UTR library from SH-SY5Y cells displaying the strongest association. However, overall, TA dinucleotides are more strongly correlated with RNA half-life than GC content **(Fig. 4B,D; Supplemental Fig. S7A)**. We reasoned that GC content and TA 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 ($t_{1/2}$ ~ TA-diNT% + GC% + diNT% × GC%). For both 5' UTRs and 3' UTRs, we observed that TA 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 TA dinucleotides **(Supplemental Fig. S7A)**. Moreover, the interaction term (TA-dinucleotide*GC-content) 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 TA dinucleotides. To further explore this interaction, we stratified degrees of GC content and examined their effects on TA dinucleotides. Notably, despite a somewhat anti-correlation between TA dinucleotides and GC content, they do not simply oppose each other. Substantial counts of TA dinucleotides are observed in regions characterized by high GC content. We found that TA 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 TA 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 TA dinucleotide ratios but was barely detectable for low TA-dinucleotide ratios **(Supplemental Fig. S7B)**. Thus, the RNA destabilizing effect is most pronounced under conditions of a low TA dinucleotide ratio and low GC content **(Supplemental Fig. S7C)**.

Moreover, the stabilizing or destabilizing effects of TA deletion or addition, respectively, were only observed for low GC content, further supporting that high GC content hinders the destabilizing effect of TA dinucleotides **(Fig. 5B, Supplemental Fig. S7D)**.

To corroborate the interplay between the TA 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 TA-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 TA dinucleotides and RNA P-body localization, especially for those occurring in 3' UTRs **(Fig. 5D)**. For both UTR types, we observed a greater TA 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 TA dinucleotides from being recognized by nucleases, though other physical hindrances may also diminish the destabilizing effect of TA dinucleotides. Therefore, we examined the influence of RBP binding on the destabilizing effect of TA dinucleotides. In Figures 5F and G, we calculated the number of RBP species binding to TA 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 TA dinucleotide. In doing so, we revealed that the RNA group with multiple binding partners indeed 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 TA-destabilization, we selected four 3' UTRs (APC, WDR35, SH3TC2 and MTR) with a TA 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 (ρ = 0.8) in accordance with numbers of RBP binding sites per TA dinucleotide. Together, the results argue a protective role of RBP binding on TA 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 TA-dinucleotide-binding proteins with RNA half-life, we identified TA-binding RBPs that play roles in either safeguarding or promoting RNA degradation **(Fig. 5J; Supplemental Fig. S7E; Supplemental Table S5)**. Notably, the protective TA-containing motifs were found to be T-rich, reinforcing the observation that T-stretch sequences may enhance RNA stability. Thus, we have identified interplay between GC content, RBP-binding and the destabilizing effect on RNAs of TA 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)**.

TA dinucleotide ratio of UTRs reflects functional enrichment

Since we identified the TA dinucleotides of UTRs as a major stability determinant, we argued that if TA 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 TA 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 TA 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 TA dinucleotide ratios **(Fig. 6B)**. To investigate how TA dinucleotide localization contributes to biological functions, we utilized a sliding window approach to identify UTR regions with TA dinucleotides above or below genomic background levels **(Fig. 6C,D)**. TA 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 TA dinucleotide-enriched windows, whereas genes responsible for regulating cell proliferation had most TA dinucleotides-depleted windows **(Fig. 6C)**. For 3' UTRs, RNA translation and immune-related functions proved to be the GO terms with most TA dinucleotide-enriched windows **(Fig. 6D)**. Since 5' and 3' UTRs may synergistically regulate gene expression, we examined genes for which the TA dinucleotide ratio in both UTRs significantly differed from the genomic background, which revealed that the TA 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 TA dinucleotide ratios in both UTRs being above background levels, are all related to the immune response. It is also noteworthy that the high TA-dinucleotide ratio of genes regulating appetite, synaptic signaling and the immune response reflects the transient expression nature of these gene groups, further supporting that TA dinucleotides exert a destabilizing effect on RNA. Together, this genome-wide analysis indicates that the TA-dinucleotide ratio of UTRs reflects global regulation of gene expression dynamics.

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 induce 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 TA-rich sequences are the most influential stability determinants **(Fig. 3)**.

Finally, to establish a causal relationship 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. 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 TA dinucleotides, as well as WWWWWW (W:A/T) and ATTTA (ELAV1 binding site) motifs, all highlighting the importance of TA rich sequences in UTRs for RNA stability **(Fig. 3)**. Among these destabilizing factors, TA dinucleotides were best correlated with RNA half-life **(Fig. 4A-D)**. The TA 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 TA 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 TA dinucleotides for RNA stability

Although it was intuitive to infer a negative correlation between TA dinucleotides and GC content, the best-known stability regulator, we found that TA dinucleotides cannot simply be viewed as an inverse proxy for GC content. Instead, the TA 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 TA dinucleotide ratio **(Supplemental Fig. S7)**. Consequently, TA 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 TA dinucleotides is generally hampered by RBP binding **(Fig. 5F,G)**. We reason that, similar to structural hindrance, recognition of TA dinucleotides can be blocked by the physical occupancy of RBPs, resulting in an overall protective effect of RBPs against TA 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 unexpectedly 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 that those in 3' UTRs caused stability changes **(Fig. 1D)**. 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 TA 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 TA 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 are 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, wild type and mutant, which only differs only by the variant position. The variant will be 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 ith 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 represented wild-type (WT) 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 signal value < 1 or *p* value > 0.05 will be 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 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 first clustered by hierarchical clustering tree. 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:

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

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

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

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.

The height for cutting the tree was gradually increased until the VIF values of all representative features were less than 5. 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 x_{ij} 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.

***In vivo* validation**

To validate TA-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

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

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

where $N_{i(t)}$ and $N_{i(t=0)}$ are T to C conversion rate at time points t and 0. R lm() function was used to estimate λ and calculate transcript half-life $t_{1/2}$.

TA-diNT count were calculated by R Biostrings::countPattern and sequence length by nchar(). TA-diNT ratio in each UTR sequences were TA-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.

TA-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.

TA-dinucleotide ratio in each UTR sequence was calculated by (TA count ÷ (sequence length - 1)). For every GO annotation containing > 20 qualified transcripts, TA-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 TA-dinucleotide ratio, TA-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 TA-ratio to 100 fragments and the mean sliding TA-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 TA-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&hgtgroup_compGeno_close=0#compGenoGroup

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/

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

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.

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.

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/ARTN e05005
10.7554/eLife.05005

Agarwal, V., & Kelley, D. R. (2022). The genetic and biochemical determinants of mRNA degradation rates in mammals. *Genome Biology*, *23*(1). https://doi.org/ARTN 245
10.1186/s13059-022-02811-x

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

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*(21), 3613-3634. https://doi.org/10.1007/s00018-012-0990-9

Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, *136*(2), 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*(1). https://doi.org/ARTN e3001956
10.1371/journal.pbio.3001956

Beffagna, G., Occhi, G., Nava, A., Vitiello, L., Ditadi, A., Basso, C., Bauce, B., Carraro, G., Thiene, G., Towbin, J. A., Danieli, G. A., & Rampazzo, A. (2005). Regulatory mutations in transforming growth factor-beta 3 gene cause arrhythmogenic right ventricular cardiomyopathy type 1. *Cardiovascular Research*, *65*(2), 366-373. https://doi.org/10.1016/j.cardiores.2004.10.005

Benjamin Neymotin, V. E., David Gresham. (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*(22), 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*(1). https://doi.org/ARTN 30
10.1186/s12915-021-00949-x

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 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/ARTN 474
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*. https://doi.org/10.1038/s41594-022-00844-1
Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T. S., Malta, T. M., Pagnotta, S. M., Castiglioni, I., Ceccarelli, M., Bontempi, G., & Noushmehr, H. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, *44*(8). https://doi.org/ARTN e71
10.1093/nar/gkv1507
Courel, M., Clement, Y., Bossevain, C., Foretek, D., Cruchez, O. V., Yi, Z., Benard, M., Benassy, M. N., Kress, M., Vindry, C., Ernoult-Lange, M., Antoniewski, C., Morillon, A., Brest, P., Hubstenberger, A., Crollius, H. R., Standart, N., & Weil, D. (2019). GC content shapes mRNA storage and decay in human cells. *Elife*, *8*. https://doi.org/ARTN e49708
10.7554/eLife.49708
Dai, H., & Guan, Y. T. (2020). The Nubeam reference-free approach to analyze metagenomic sequencing reads. *Genome Research*, *30*(9), 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*(4), 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*(12), 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*(D1), 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), 1-22. https://doi.org/DOI 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*(2), 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/ARTN baw035
10.1093/database/baw035
Goyal, S., Tanigawa, Y., Zhang, W. H., Chai, J. F., Almeida, M., Sim, X. L., Lerner, M., Chainakul, J., Ramiu, J. G., Seraphin, C., Apple, B., Vaughan, A., Muniu,

J., Peralta, J., Lehman, D. M., Ralhan, S., Wander, G. S., Singh, J. R., Mehra, N. K., . . . Sanghera, D. K. (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*(1). https://doi.org/ARTN 113
10.1186/s12944-021-01531-8

Griesemer, D., Xue, J. R., Reilly, S. K., Ulirsch, J. C., Kukreja, K., Davis, J. R., Kanai, M., Yang, D. K., Butts, J. C., Guney, M. H., Luban, J., Montgomery, S. B., Finucane, H. K., Novina, C. D., Tewhey, R., & Sabeti, P. C. (2021). Genome-wide functional screen of 3 ' UTR variants uncovers causal variants for human disease and evolution. *Cell*, *184*(20), 5247-+. 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/ARTN 840369
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*(12), 1009-1021. https://doi.org/10.1016/j.tibs.2019.07.001

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T. S., Harte, R. A., Hsu, F., Hillman-Jackson, J., Kuhn, R. M., Pedersen, J. S., Pohl, A., Raney, B. J., Rosenbloom, K. R., Siepel, A., Smith, K. E., Sugnet, C. W., . . . Kent, W. J. (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*(8), 1-23. <Go to ISI>://WOS:000261526300001

Huntzinger, E., & Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature Reviews Genetics*, *12*(2), 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*(9), 814-+. 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*(4), 357-U121. https://doi.org/10.1038/Nmeth.3317

Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., & Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database-the Journal of Biological Databases and Curation*. https://doi.org/ARTN bar030
10.1093/database/bar030

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., & Maglott, D. R. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, *44*(D1), D862-D868.

https://doi.org/10.1093/nar/gkv1222
Lee, B. T., Barber, G. P., Benet-Pages, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A. S., Lee, C. M., Muthuraman, P., Nassar, L. R., Nguy, B., Pereira, T., Perez, G., Raney, B. J., Rosenbloom, K. R., Schmelter, D., Speir, M. L., . . . Kent, W. J. (2022). The UCSC Genome Browser database: 2022 update. *Nucleic Acids Research*, *50*(D1), 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*(6), 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/Artn 26
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*(12). https://doi.org/ARTN 550
10.1186/s13059-014-0550-8
Luo, E. C., Nathanson, J. L., Tan, F. E., Schwartz, J. L., Schmok, J. C., Shankar, A., Markmiller, S., Yee, B. A., Sathe, S., Pratt, G. A., Scaletta, D. B., Ha, Y. C., Hill, D. E., Aigner, S., & Yeo, G. W. (2020). Large-scale tethered function assays identify factors that regulate mRNA stability and translation. *Nature Structural & Molecular Biology*, *27*(10), 989-+. https://doi.org/10.1038/s41594-020-0477-6
Luo, Y. H., Hitz, B. C., Gabdank, I., Hilton, J. A., Kagda, M. S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., Baymuradov, U. K., Graham, K., Litton, C., Miyasato, S. R., Strattan, J. S., Jolanki, O., Lee, J. W., Tanaka, F. Y., Adenekan, P., . . . Cherry, J. M. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, *48*(D1), D882-D889. https://doi.org/10.1093/nar/gkz1062
MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., & Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, *45*(D1), 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*(6472), 1470-+. https://doi.org/ARTN eaav1741
10.1126/science.aav1741
Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biology*, *3*(3). https://doi.org/ARTN 0004
10.1186/gb-2002-3-3-reviews0004
Mitchell, P., & Tollervey, D. (2001). MRNA turnover. *Current Opinion in Cell Biology*, *13*(3), 320-325. https://doi.org/Doi 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*(1), 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*(10), 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*(22), 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*(1). https://doi.org/ARTN 851
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*(1-2), 73-81. https://doi.org/Doi 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/ARTN 1170
10.3389/fphar.2019.01170

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 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*(6), 1083-+. 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*(7), 803-+. 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*(6), 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*(8), 677-679. <Go to ISI>://WOS:000082063000002

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*(1). https://doi.org/ARTN jkab404
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/ARTN 529
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*(11), 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*(6), 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*(4). https://doi.org/ARTN e1474
10.1002/wrna.1474
van der Loo, M. P. J. (2014). The stringdist Package for Approximate String Matching. *R Journal*, *6*(1), 111-122. <Go to ISI>://WOS:000343788100012
Vejnar, C. E., Messih, M. A., Takacs, C. M., Yartseva, V., Oikonomou, P., Christiano, R., Stoeckius, M., Lau, S., Lee, M. T., Beaudoin, J. D., Musaev, D., Darwich-Codore, H., Walther, T. C., Tavazoie, S., Cifuentes, D., & Giraldez, A. J. (2019). Genome wide analysis of 3 ' UTR sequence elements and proteins regulating mRNA stability during maternal-to-zygotic transition in zebrafish. *Genome Research*, *29*(7), 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*(D1), 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/ARTN 177
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*(7). 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*(4), 387-U259. https://doi.org/10.1038/nbt.2851

FIGURES AND FIGURE LEGENDS

Figure 1

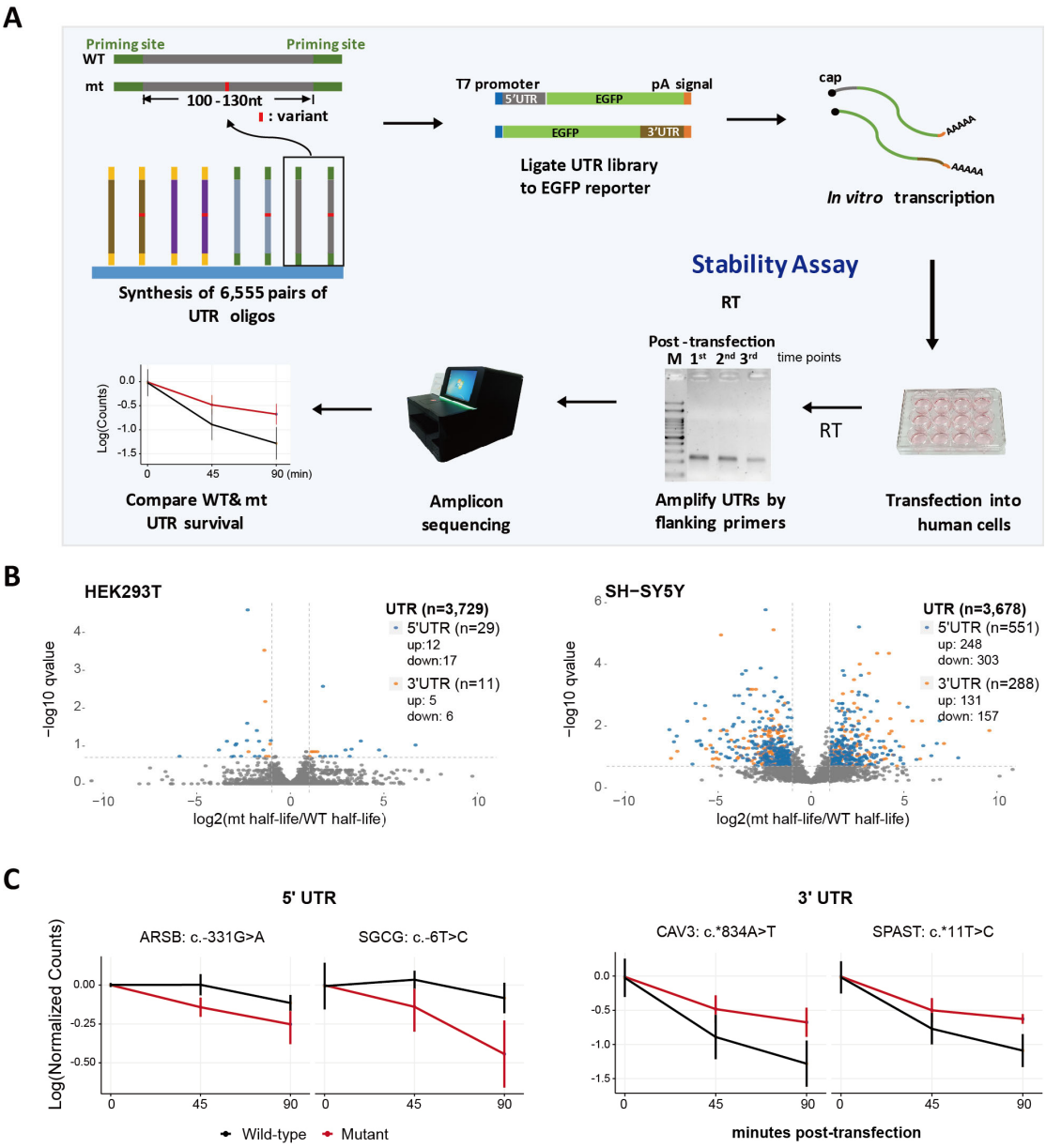

Figure 1. Massively parallel reporter assay (MPRA) to determine the effects of UTR variants on RNA stability. (A) MPRA workflow. In brief, 6,555 WT and mutant UTR pairs were synthesized in bulk, ligated with promoters and reporter sequences, *in vitro*-transcribed into capped and tailed RNAs, transfected 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.

Figure 2

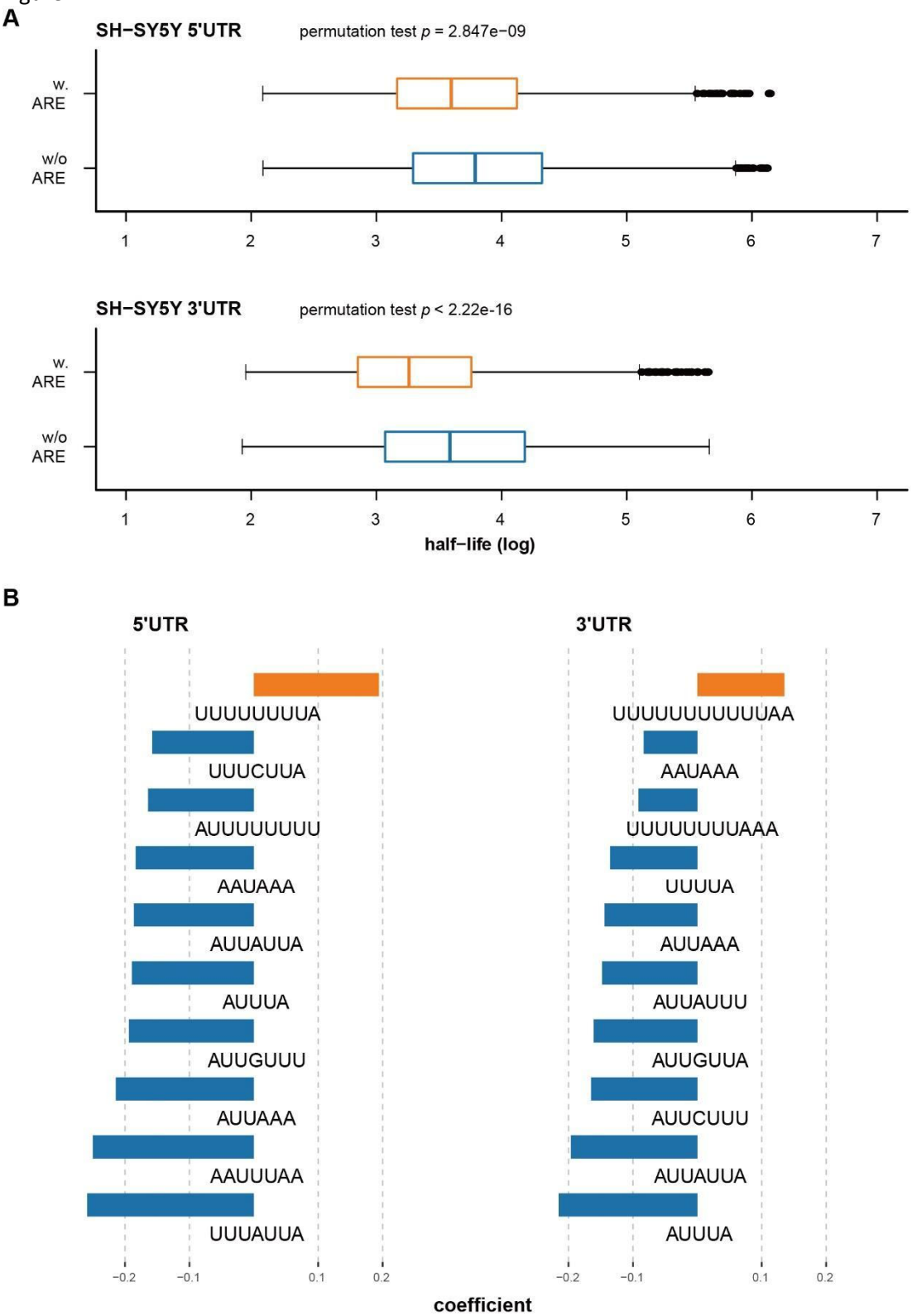

Figure 2. AREs generally destabilize RNAs (except extremely U-rich AREs). (A) AREs of both UTR types destabilize RNA. **(B)** The ten most influential AREs in terms of RNA stability. Coefficients are determined by regression analysis, representing the effect size of each motif.

Figure 3

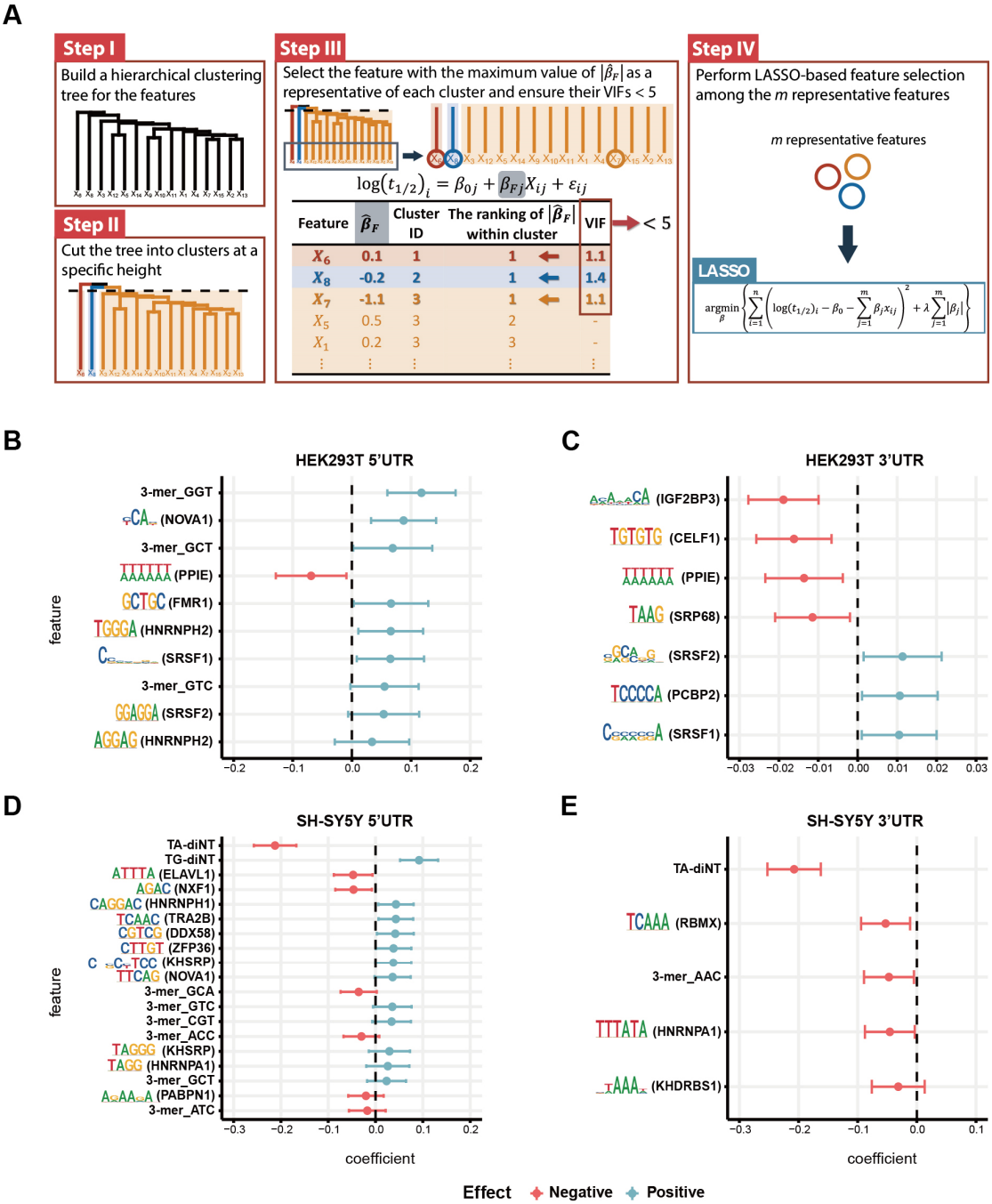

Figure 3. Inferential statistical analysis of RNA stability determinants. (A) Workflow of variable selection to build models of influencers of RNA stability. **(B-E)** Influential regulators for the 5' UTR library from HEK293T cells (B), the 3' UTR library from HEK293T cells (C), the 5' UTR library from 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).

Figure 4

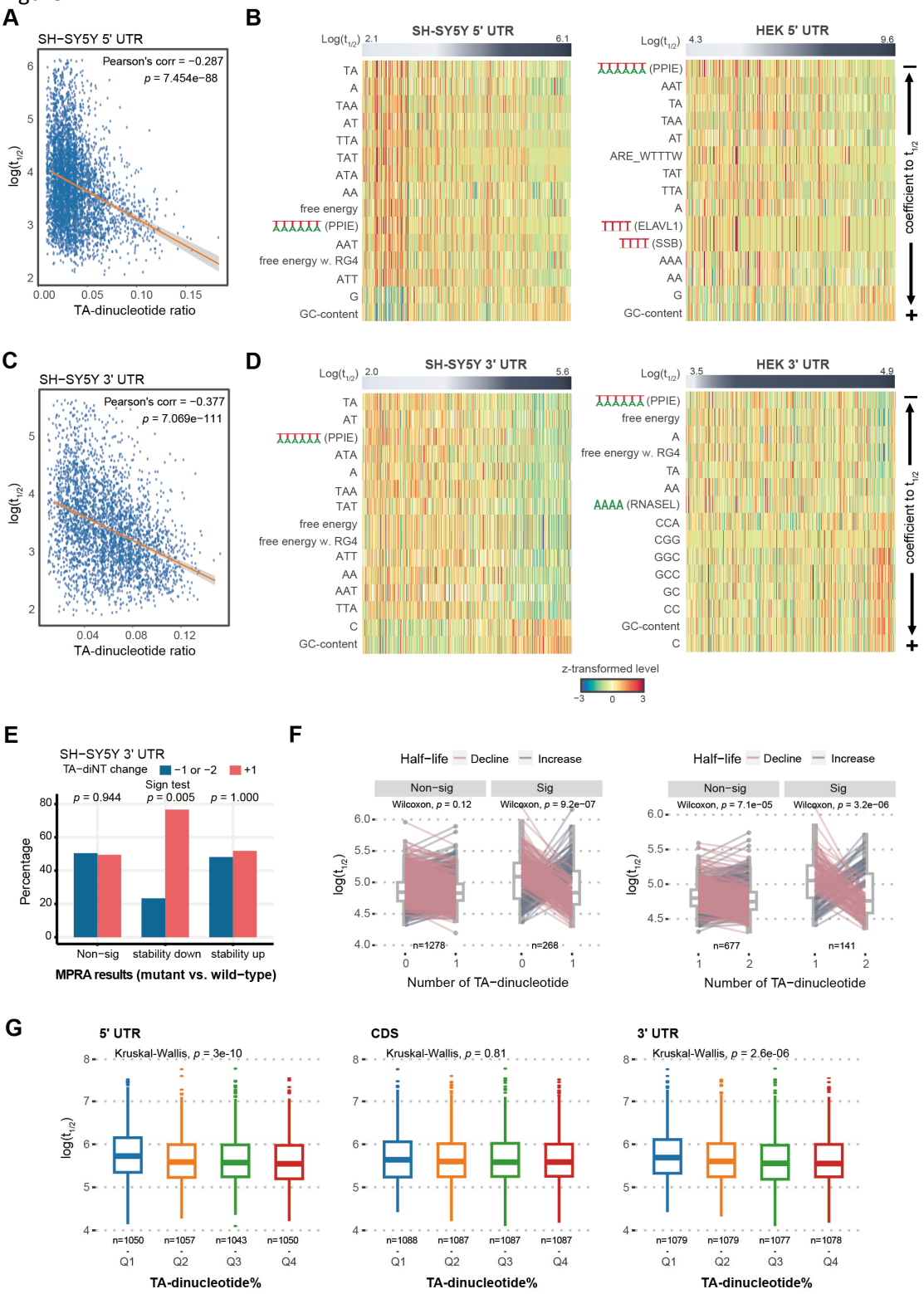

Figure 4. The UTR TA dinucleotide ratio is the most common and influential RNA destabilizing factor. (A) Correlation of the 5' UTR TA dinucleotide ratio and half-life. **(B)** Top 15 influential factors in the TA cluster of 5' UTR. UTRs are arranged by half-life, and factors by their coefficient to half-life. Note that there are destabilizing factors (such as TA and AT dinucleotides) as well as stabilizing factors (such as GC content and G monomers) in this cluster. TA dinucleotide and WWWWWW (PPIE) (where W represents A/T) are representative

of the cluster for modeling UTR stability in SH-SY5Y and HEK293T cells, respectively. **(C)** Correlation of the 3’ UTR TA dinucleotide ratio and half-life. **(D)** Top 15 influential factors in the TA cluster of 3’ UTR. **(E)** Mutational gain of a TA dinucleotide by 3’ UTRs significantly reduces RNA stability (lower panel). **(F)** Gain of TA 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 TA-nucleotide ratios of both UTRs reduce endogenous RNA stability in HEK293 cells. Q1-Q4 denote quantile groups categorized based on the TA-dinucleotide ratio.

Figure 5

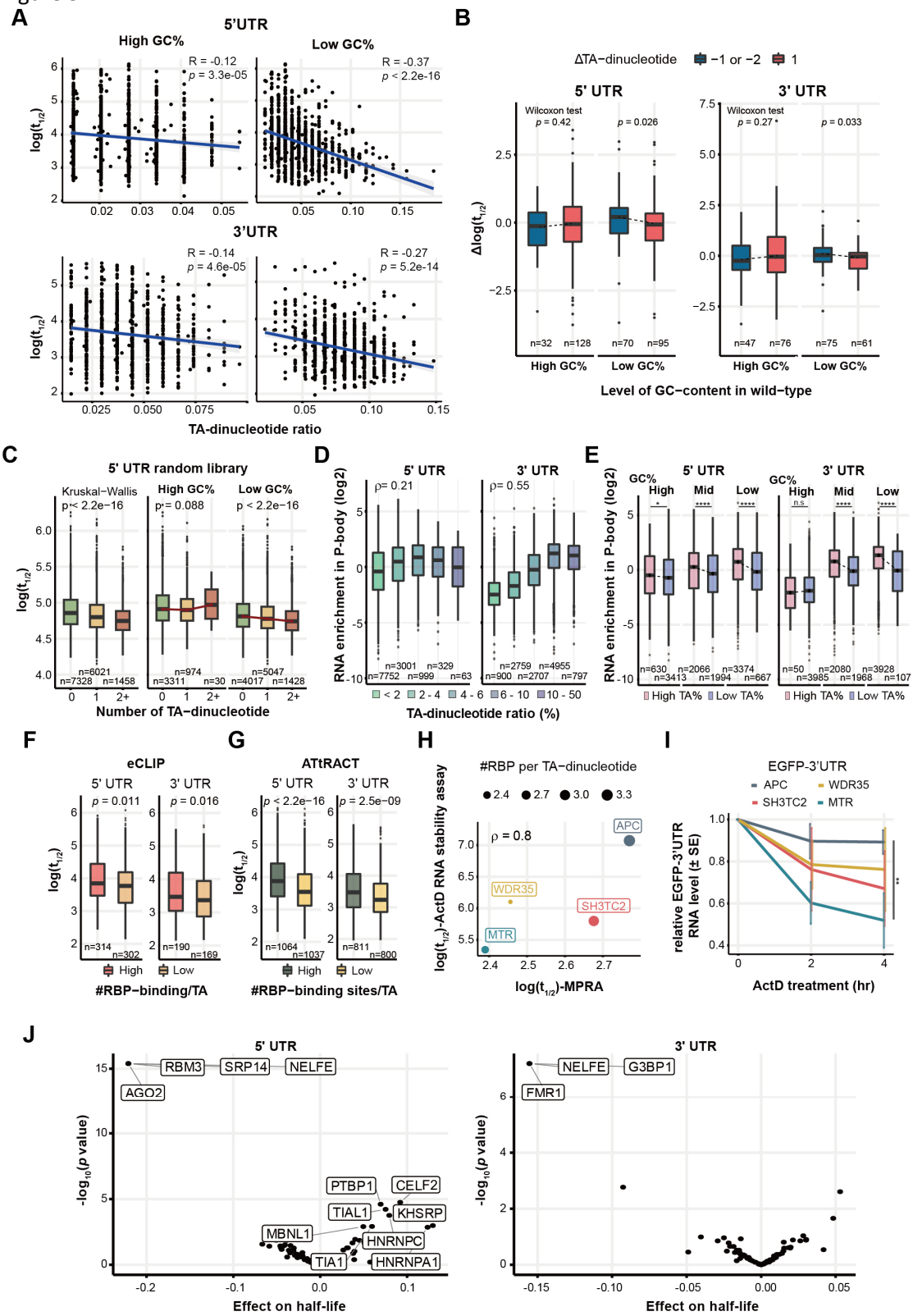

Figure 5. GC content, RBP and ribosome binding shields RNA from the destabilizing effect of TA dinucleotides. (A) MPRA data stratified according to the GC content (GC%) of UTRs. The data was divided into high and low groups according to the median of GC%. In both UTRs, the destabilizing effect of TA 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 TA dinucleotides in mutant UTRs. The destabilizing effect of TA-addition (blue) and the stabilizing effect of TA-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 TA dinucleotide is observed with 5' UTR random library. High GC content hinders the TA-destabilizing effect. **(D)** TA dinucleotides are enriched in P-body-resident mRNAs. ρ represents Spearman's correlation coefficient. **(E)** High GC content inhibits enrichment of TA dinucleotide-hosting mRNAs in P-bodies. For medium or low GC%, a high TA 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 TA dinucleotide are more stable. The high and low groups was stratified based on the median of number of RBPs per TA. **(G)** UTRs harboring more predicted RBP-binding sites per TA 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-TA 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 TA-dinucleotide-binding protein motifs with RNA half-life in SH-SY5Y cells. Note that TA-binding RBPs can have both positive and negative effects on RNA stability.

Figure 6

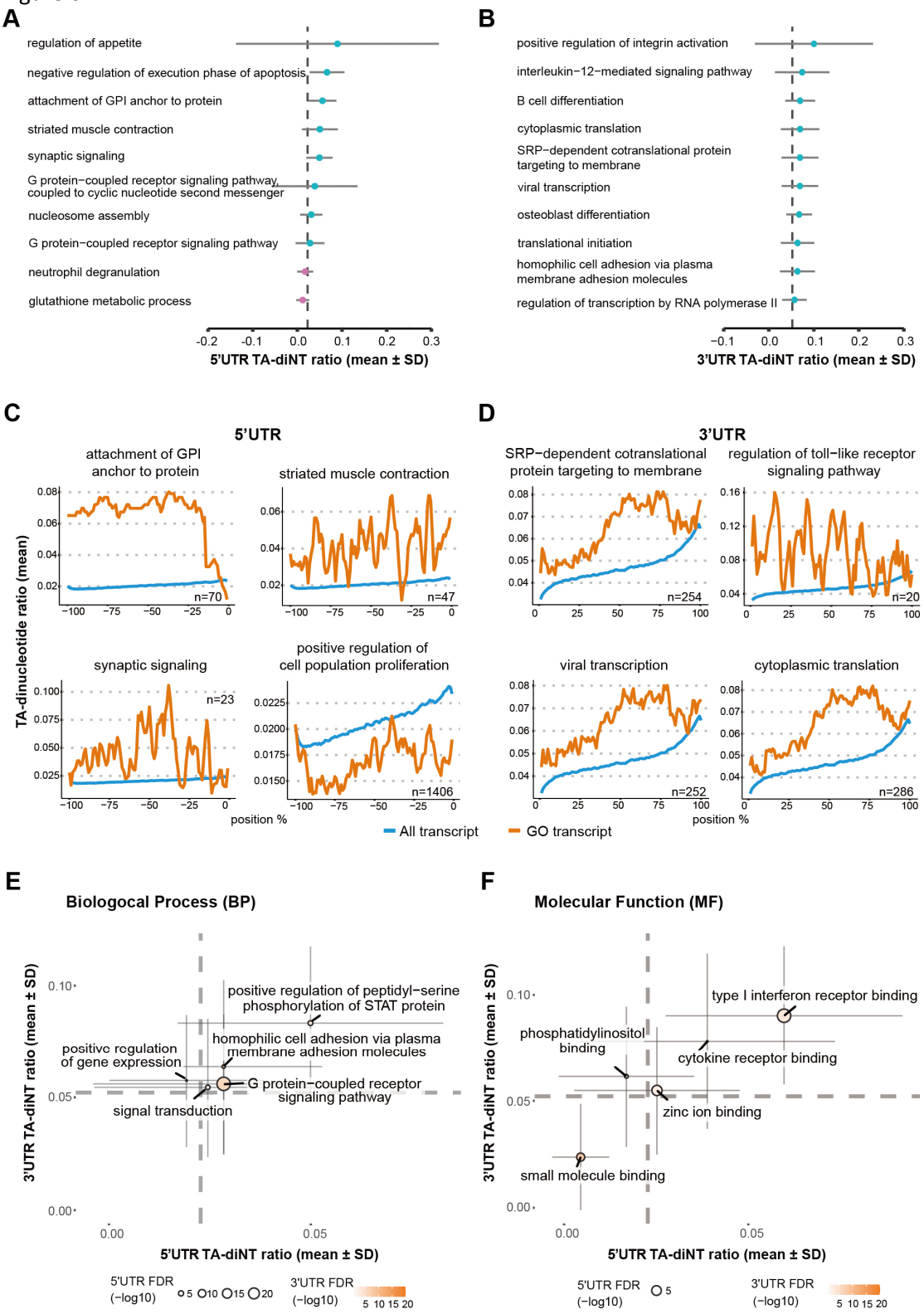

Figure 6. TA dinucleotides delineate functional gene groups. (A) The top ten biological processes for which the 5' UTR TA dinucleotide ratio most significantly deviated from the genomic background. **(B)** The top ten biological processes for which the 3' UTR TA dinucleotide ratio most significantly deviated from the genomic background. **(C)** Functional gene groups for which the 5' UTR TA 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 TA dinucleotide ratio was significantly above or below the genomic background in more than ten sliding windows. **(E)** Biological processes for RNAs in which the TA dinucleotide ratios of both 5’ and 3’ UTRs are significantly different from the genomic background. **(F)** Molecular functions for RNAs in which the TA dinucleotide ratios of both 5’ and 3’ UTRs are significantly different from the genomic background.

Figure 7

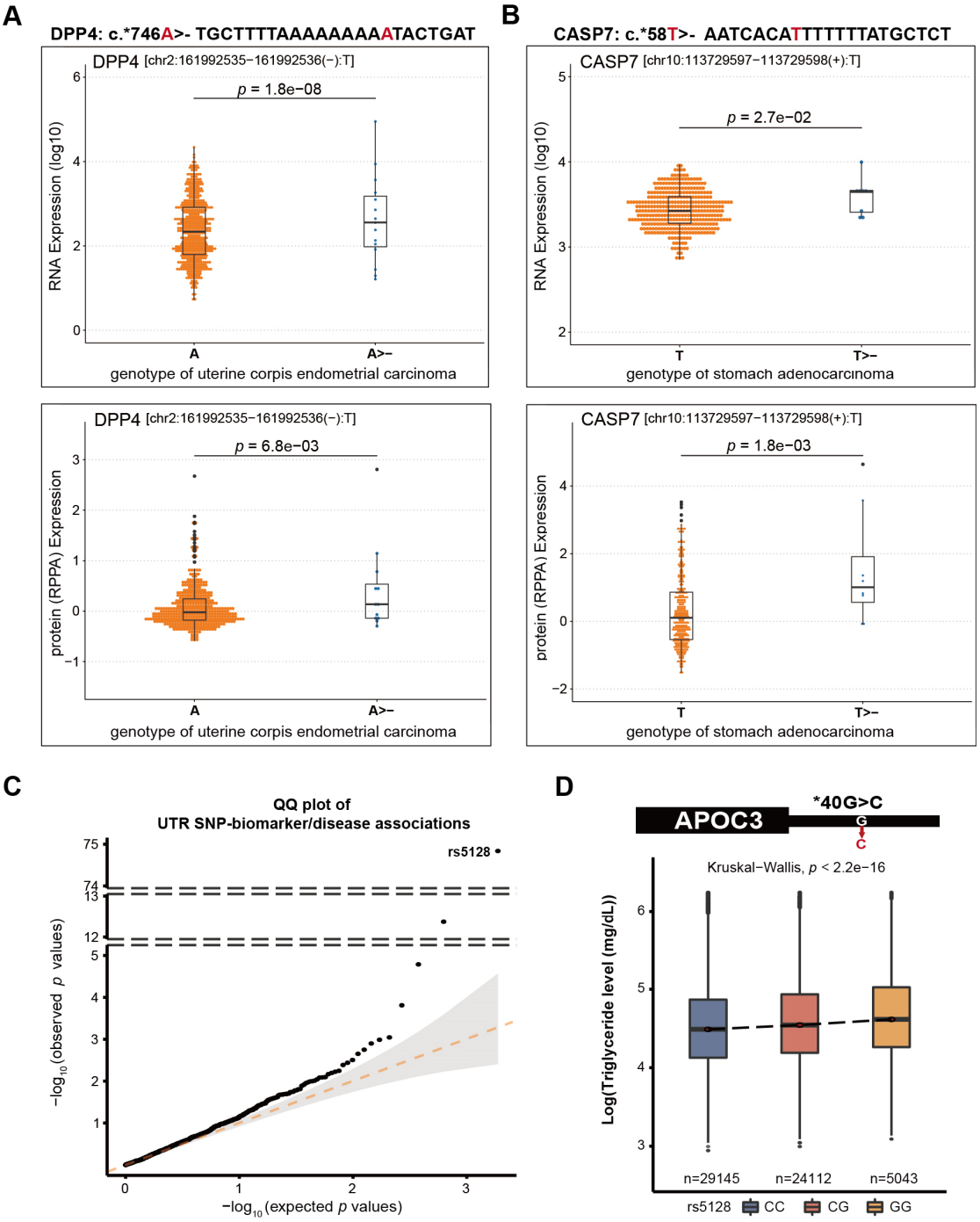

Figure 7. UTR variants associated with disease. (A-B) 3' UTR mutations that increase RNA and protein expression in carcinoma samples. Protein level was determined by reverse phase protein arrays (RPPA). **(C)** QQ plot of the *p* value distribution of stability-altering UTR variants in association with 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).