OXFORD G3 Genes | Genomes | Genetics

G3, 2022, 12(1), jkab404
https://doi.org/10.1093/g3journal/jkab404
Advance Access Publication Date: 29 November 2021
Investigation

Massively parallel analysis of human 3′ UTRs reveals that AU-rich element length and registration predict mRNA destabilization

David A. Siegel ,[1,*,†] Olivier Le Tonqueze,[1,†] Anne Biton,[1,2,†] Noah Zaitlen,[1] and David J. Erle[1]

[1]Department of Medicine, Lung Biology Center, University of California San Francisco, San Francisco, CA 94158, USA
[2]Hub de Bioinformatique et Biostatistique—Département Biologie Computationnelle, Institut Pasteur, Université de Paris, Bioinformatics and Biostatistics Hub, F-75015 Paris, France

*Corresponding author: Email: David.Siegel@UCSF.edu
†These authors contributed equally to this work.

Abstract

AU-rich elements (AREs) are 3′ UTR cis-regulatory elements that regulate the stability of mRNAs. Consensus ARE motifs have been determined, but little is known about how differences in 3′ UTR sequences that conform to these motifs affect their function. Here, we use functional annotation of sequences from 3′ UTRs (fast-UTR), a massively parallel reporter assay (MPRA), to investigate the effects of 41,288 3′ UTR sequence fragments from 4653 transcripts on gene expression and mRNA stability in Jurkat and Beas2B cells. Our analyses demonstrate that the length of an ARE and its registration (the first and last nucleotides of the repeating ARE motif) have significant effects on gene expression and stability. Based on this finding, we propose improved ARE classification and concomitant methods to categorize and predict the effect of AREs on gene expression and stability. Finally, to investigate the advantages of our general experimental design we examine other motifs including constitutive decay elements (CDEs), where we show that the length of the CDE stem-loop has a significant impact on steady-state expression and mRNA stability. We conclude that fast-UTR, in conjunction with our analytical approach, can produce improved yet simple sequence-based rules for predicting the activity of human 3′ UTRs.

Keywords: AU-rich element; mRNA decay; mRNA stability; GC content; constitutive decay element; MPRA; 3′ UTR

Introduction

3′ untranslated regions play an important role in regulating mRNA fate by complexing with RNA binding proteins that help control mRNA localization, translation, and stability (Glisovic *et al.* 2008; Castello *et al.* 2012; Mayr 2017). Identification of a consensus UUAUUUAU sequence in the 3′ UTRs of human and mouse mRNAs encoding tumor necrosis factor (TNF-α) and a variety of other inflammatory mediators led to the suggestion that these AU-rich elements AREs) could be important for regulating gene expression (Caput *et al.* 1986). Subsequent studies confirmed that these and other AREs interact with ARE-binding proteins such as AUF1 (also known as hnRNPD), HuR and other Hu family proteins, and the CCCH zinc finger-containing RBPs ZFP36 (tristetraprolin), ZFP36L1, and ZFP36L2 (Hodson *et al.* 2010), to alter mRNA degradation and protein expression (Barreau 2005). In most cases, AREs have been reported to destabilize mRNAs, although in some cellular contexts certain AREs and ARE-binding proteins have been shown to stabilize mRNAs (Peng 1998; Barreau 2005). Subsequent analyses of the human genome concluded that as many as 58% of human genes code for mRNAs that contain AREs (Bakheet 2001, 2003; Bakheet *et al.* 2018), suggesting that these elements play a major role in regulating expression of a large group of genes.

An initial classification of AREs was proposed based upon studies of human 3′ UTR sequences and analyses of mutation effects and the activity of synthetic AREs (Chen and Shyu 1995). AUUUA motifs were recognized as critical for destabilizing effects of many 3′ UTRs, and in many cases destabilizing AREs contained two or more overlapping repeats of this motif. However, some AUUUA motif-containing sequences were not destabilizing and some AU-rich sequences that lacked the AUUUA motif had potent mRNA destabilizing activity. Based upon these observations, Chen and Shyu (Chen and Shyu 1995) divided AREs into two classes of AUUUA-containing AREs and a third class of non-AUUUA AREs. Class I AUUUA-containing AREs had 1-3 copies of scattered AUUUA motifs coupled with a nearby U-rich region or U stretch, whereas class II AUUUA-containing AREs had at least two overlapping copies of the nonamer UUAUUUA(U/A)(U/A) in a U-rich region. Non-AUUUA AREs had a U-rich region and other unknown features, and the relationship of these sequences to AUUUA-containing AREs

Received: May 31, 2021. **Accepted:** October 13, 2021
© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

remains poorly understood. Subsequent studies based on analyses of a set of 4884 AUUUA-containing AREs led to a new classification based primarily on the number of overlapping AUUUA-repeats (Bakheet 2001, 2003; Bakheet *et al.* 2018). This classification system, with five clusters distinguished by the number of repeats, was used to identify AUUUA-containing AREs in the human genome. AREs identified using this classification were found to be abundant in 3′ UTRs of human genes. However, the functional activities of this large set of ARE-containing 3′ UTR sequences remains mostly unknown.

To address this shortcoming we relied on novel experimental and analytical strategies. First, we leveraged massively parallel reporter assays (MPRAs), which are capable of simultaneously characterizing the regulatory impact of thousands of mRNA sequence fragments for functional analysis (Zhao *et al.* 2014; Kreimer *et al.* 2017; Rabani *et al.* 2017; Avsec *et al.* 2021; Sample *et al.* 2019). We previously developed a massively parallel method for functional annotation of sequences from 3′ UTRs (fast-UTR) (Zhao *et al.* 2014). In that study, we used fast-UTR to analyze a set of 3000 160 nt sequences from human 3′ UTRs that were highly conserved across mammalian species and confirmed that AREs and constitutive decay elements (CDEs) in these sequences are important contributors to 3′ UTR-mediated mRNA destabilization. However, due to the limited number of sequences included in that study, we were unable to identify rules that governed the activity of sequences that conformed to these motifs.

To address this issue, in this study, we designed, produced, and analyzed a fast-UTR library containing a comprehensive set of ARE sequences from human mRNA 3′ UTRs in human cell lines. Although the activity of specific AREs can differ between cell types, our goal was to discover a general rule that could be applied in different cells. We began our analysis by studying Jurkat T cells, given the importance of AREs in T cell development (Hodson *et al.* 2010) and autoimmunity. We then studied the same set of 3′ UTR sequences in Beas2B airway epithelial cells to investigate whether rules discovered in Jurkat T cells were replicated in a different cell type. In total, the MPRA in our current study included 41,288 sequences from 4653 transcripts from each of two different cell lines. Applying novel analytic approaches to this set of sequences (?) allowed us to identify rules that predict functional effects of AREs on mRNA stability and stead-state expression.

To verify that our combined fast-UTR approach and analytic strategy was not uniquely successfully in the context of AREs, we also considered a second important class of 3′ UTR regulatory elements known as CDEs (Caput *et al.* 1986; Leppek *et al.* 2013). CDEs are conserved stem loop motifs that bind to the proteins Roquin and Roquin2, resulting in increased mRNA decay (Leppek *et al.* 2013). CDEs include an upper stem-loop sequence of the form UUCYRYGAA flanked by lower stem sequences. Lower stem sequences are formed by 2–5 nt pairs of reverse-complementary sequences (*e.g.*, CCUUCYRYGAAGG has a lower stem length of 2). We similarly generated a fast-UTR library containing a comprehensive set of CDEs. We again obtained novel sequence based rules predicting CDE function. Finally, to show the potential of our approach to aid functional interpretation of arbitrary 3′ UTR sequence fragments, we conclude by developing statistical models to predict the activity of entire 160 nt 3′ UTR sequence fragments in our full MPRA data set.

Materials and methods

Sequence design

Definition of 3′ UTR regions and regulatory regions

We segmented 3′ UTRs from human RefSeq transcripts (v68) into 160 nt sliding windows with a shift of 80 nt. Only regions of at least 20 nucleotides were included; segments shorter than 160 nt were padded with a sequence from the CXCL7 3′ UTR (NM_002704, 475-602) that had minimal regulatory effects in prior experiments (Zhao *et al.* 2014). We then identified those 160 nt 3′ sequence segments that contain suspected ARE motifs and CDE motifs, as well as several other features that are described in the Supplementary Methods. The ARE motifs targeted for inclusion were those defined according to the ARED website (Bakheet 2001) as of Fall 2014, generally containing one or more repeat of AUUUA, following rules that are precisely defined in the Supplementary Methods (Supplementary Table S1). Since the design of our experiment, the ARED group has revised their ARE class definitions into "clusters" in a more recent publication (ARE Plus) (Bakheet *et al.* 2018), and our analysis follows those updated conventions. The CDE motifs contain a central sequence of UUCYRYGAA, surrounded by a "lower stem" of 2–5 bp and then an unpaired nt.

In addition to the reference regions containing ARE and CDE motifs, we designed segments that contain mutated versions of these motifs. For AREs, we mutated the central U of the AUUUA pentamer to a C or G, alternating down the sequence (for example AUUUAUUUA was mutated to AUcUAUgUa). Since several features were targeted for inclusion, we noticed that there are several regions in the dataset that contain core AUUUA pentamers that were shorter than the ARED definitions (for instance, a region containing a CDE might also contain an "AUUUA" within the 160 nt window), so we have designed segments where 429 of these smaller AUUUAs were mutated into AUcUAs (Bakheet 2001). For CDEs, two types of mutations were introduced: (1) UUCYRYGAA was mutated to UagYRYGAA, and (2) the CDE sequence was shuffled. In total, the assay contained 41,288 segments from 13,334 3′′ UTR regions from 4653 RefSeq transcripts.

Fast-UTR assay

An oligonucleotide pool containing the full set of library segments was produced by massively parallel synthesis (Agilent), amplified by PCR, and cloned into the BTV lentiviral plasmid as previously described (Zhao *et al.* 2014). Barcodes were introduced using random 8-mer sequences incorporated into the PCR primer in order to obtain systematic estimates of technical variability. A lentiviral library was used to transduce Jurkat T cells or Beas2B airway epithelial cells expressing tetracycline transactivator tTA, allowing for doxycycline (Dox)-regulated reporter transcription. Cells were maintained in culture for 2 weeks after transduction and were not stimulated. Cells were left untreated (t_0) or treated with Dox (1 μg/ml) for 4 h to inhibit reporter transgene expression (t_4) prior to isolation of DNA and RNA using the AllPrep DNA/RNA/Protein Mini Kit (Qiagen# 80004) according to the manufacturers protocol. RNA was reverse transcribed to cDNA, and both genomic DNA and cDNA were amplified to produce sequencing libraries. Sequencing was performed using an Illumina HiSeq 4000. After processing, we were left with an average of 22 clones per segment for the Jurkat data at t_0, 18 clones per segment for the Jurkat data at $t_4/(t_4 + t_0)$, 9 clones per segment for the Beas2B

data at t_0, and seven clones per segment for the Beas2B data at $t_4/(t_4 + t_0)$.

Data analysis

Steady state expression and stability

We quantify mRNA activity in two ways, by measuring "steady-state expression" and "mRNA stability." To quantify steady state expression we use the count-normalized ratio before the addition of Dox:

$$\text{steady state expression} \equiv \frac{RNA}{(RNA + DNA)}\Big|_{t_0} \tag{1}$$

To quantify mRNA stability, we use a ratio of ratios, comparing the expression 4 h after the addition of Dox (t_4) to the steady-state expression (t_0):

$$\text{mRNA stability} \equiv \frac{\frac{RNA}{(RNA+DNA)}\Big|_{t_4}}{\frac{RNA}{(RNA+DNA)}\Big|_{t_4} + \frac{RNA}{(RNA+DNA)}\Big|_{t_0}} \tag{2}$$

This is not intended to be a perfect measurement of the decay time, but should give some measure of the stability with consistent use. For example, if a segment contains a binding site for an mRNA degrading protein, the steady state expression and stability will be small, and a mutation to that binding site will lead to greater stability; whereas a segment containing a binding site for a stabilizing protein will result in a larger estimate of steady state expression and stability. These ratios do not diverge when RNA or DNA counts are small. In a previous study (Zhao *et al.* 2014), we showed that the concentration decays exponentially, and we used the 4 h timepoint because it contained the most information and the widest dynamic range.

To investigate the effect of a mutation on the activity of a UTR segment we examine the change (Δ) in steady state expression or stability by subtracting the reference value from the mutant value: $\Delta \equiv \text{Mutant} - \text{Reference}$. If a functional element (ARE or CDE) is destabilizing, then mutating it will likely lead to a positive change in stability and expression by disrupting the corresponding mRNA binding proteins.

To set a minimum threshold for the data quality of a 3′ UTR sequence segment, we required segments to be represented by at least 5 clones with more than 5 counts of DNA each, and for each segment to have at least 1 count of RNA in at least 1 clone. While low RNA counts are generally indicative of low gene expression, we believe the measurement of zero RNA counts to most likely be a technical artifact, since the correlations across sequencing replicates are greatly increased when segments with zero RNA counts are removed.

GC content

GC content has been suggested to impact measurements of gene expression and stability (Benjamini and Speed 2012; Litterman *et al.* 2019). Because the primary purpose of this study is to examine the relative contributions of known sequence motifs to the expression and stability of mRNAs, confounding effects of GC content are a critical consideration.

Therefore, to account for GC-content in the fast-UTR data, we fit 5th-order polynomials to the steady state expression and stability as a function of GC-content, and subtract them from the steady state expression and stability. We assessed the performance of Nth order polynomial fits ($N \in \{1, \ldots, 7\}$) through leave-one-out cross validation and found that the prediction accuracy (Pearson correlation between predicted and actual steady state expression and stability) did not improve significantly after $N = 5$. See Supplementary materials for further details.

Evaluation of ARE classification methods

To examine the quality of previously proposed as well as novel classification systems for AREs, we propose assessment via prediction quality. We assess the quality of predictions by determining correlations between measured and predicted RNA expression and stability values using a leave-one-chromosome-out approach. Broadly, we argue that better classification approaches will more accurately predict the affect of individuals AREs.

Leaving out one chromosome eliminates some sources of bias, such as generating predictions from overlapping genomic regions. Therefore we split the data into a training set of 3′ UTR segments from 23 chromosomes with which we generated our model parameters, and a test set of segments from the 24th chromosome (treating Y as a separate chromosome from X). We repeated this process for each chromosome to generate a prediction for every ARE-containing segment in the dataset. When predicting the effect of a designed mutation on a mutant and reference sequence pair, we only consider mutations that disrupt pentamers of the ARE. For categorical predictions, we calculated the categorical means from the training set, then used those means as predictors for the test set. For example, if the mean for category 3 AREs was 0.4 in the training set, we assigned a "predicted value" of 0.4 to any category 3 ARE in the test set. For regression-based methods, we performed the regressions on the training set to generate model parameters, then applied those models to the test set to produce predictions. The correlation between predicted and measured values is then reported.

The rules for five ARE prediction methods are detailed below, and additional methods are given in the Supplementary materials. Unless stated otherwise, a segment is classified by its longest ARE if more than one is present:

i) ARE Plus: We used the five cluster motifs described by Bakheet *et al.* (2018) to create a categorical prediction variable. The mean value of each category (cluster) in the training set was used as the predicted value in the test set. Clusters 1 and 2 motifs total 13 nucleotides, with AU-rich segments flanking one or two AUUUA core motifs, respectively. Clusters 3, 4, and 5 include 3, 4, or 5 exact AUUUA repeats respectively. This system differs somewhat from the earlier ARED definitions described by the same group (Bakheet 2001) which we used for the initial design.
ii) AREScore: We used the AREScore website created by Spasic *et al.* (2012) to give each sequence fragment in the training and test sets an ARE "score." We then treated AREScore as a continuous variable and performed linear regression. Regression of steady state and stability measurements on the AREScores created predictions for a given AREScore in the test set.
iii) Naive Effective Length Pentamers: Pentamers were classified by the "effective length" according to the formula floor((length(nt) + registration − 2)/4). "Registration" refers to the starting nucleotides of the ARE within the initial AUUUA pentamer: an ARE that starts AUUU*=0, UUUA*=1, UUAU*=2, and UAUU*=3. No mismatches allowed. The number of pentamers was treated as a categorical prediction variable. The mean value of each category in the training set was used as the prediction value in the test set.

iv) Effective Length (nt): AREs classified by the "effective length" (length(nt) + starting registration), rather than class of pentamer. No mismatch allowed. Effective length is treated as a categorical variable, where the mean of each category in the training set is used to predict the value in the test set.

v) Lasso K-Mer Regression: Each 160 nt segment was broken down into a list of 156 5-mers, which were used in a Lasso (Tibshirani 1996) regression scheme following (Rabani *et al.* 2017), solving the linear model y = Xb+c for effect sizes b (y is the outcome vector containing steady state expression or stability, and X is the frequency matrix). Further details are given in the Supplementary materials. Here the training set is not limited to segments with AREs, and includes every segment in the full MPRA; but the test set is limited to the same set of segments with AREs as Figure 3 (any AUUUA pentamer with length 6 or higher, plus the length-5 pentamer "AUUUA"). To predict the effect of mutations, we simply subtract the predicted expression or stability of the wild-type from the predicted expression or stability of the mutant; we do not train and test on the difference data directly.

Results

An MPRA to investigate the effect of AREs on gene expression and stability

We used our previously developed fast-UTR MPRA to test effects of a large set of human 3′ UTR segments containing elements conforming to previously defined ARE motifs on mRNA steady-state levels and mRNA stability (Figure 1). In our initial analysis, we observed a significant effect of GC content of the 3′ UTR segments on steady state expression, with increased GC content associated with reduced expression (Supplementary Figure S1A). We recently reported a similar finding in a fast-UTR-based analysis of ≈27,000 70-nt RNA binding protein binding 3′ UTR sequences tested in mouse primary T cells (Litterman *et al.* 2019). GC content also had an apparent effect on mRNA stability (Supplementary Figure S1B). The GC content effects are not due to AU-rich elements, since their effect is to do the opposite: AU-rich elements have low GC-content but generally lower steady-state expression and stability. In subsequent analyses, we adjusted for GC content as described in the Materials and Methods (Supplementary Table S2 and Figure S1).

Measurements were performed at several time points with respect to the addition of Dox: before Dox was added (t_0), and 4 h after Dox. Four hours was chosen because evaluating RNA/(RNA+DNA) at $t_4/(t_4 + t_0)$ showed greater sequence-based variation than using t_2 or t_6 in preliminary data, and because it had a greater dynamic range in our past study of fast-UTR (Zhao *et al.* 2014). We did perform similar analyses of some of the data using t_2 and t_6 for comparison; see Supplementary Figures S13, S14, and Table S10. Data analyzed at t_2 generally replicated the results analyzed at t_4. The t_6 data have limited value for assessing less stable transcripts, likely because the amount of residual transcript is low and difficult to measure accurately, and low levels of transcription that persist in the presence of Dox are sufficient to interfere with the measurements of stability.

Established categories of ARE predict gene expression and stability

We found that ARED-Plus classification was associated with functional activity of 3′ UTR segments. When 3′ UTR segments were classified according to the presence of sequences conforming to the ARED clusters, ARE cluster membership had a significant impact on both steady state RNA level (Figure 2A) and RNA stability (Figure 2B). Both steady state level and stability decreased as the number of ARE repeats increased. Segments that contained the minimal AUUUA sequence only (without any ARED-Plus motifs) had a slightly lower steady state level and stability than segments with no AUUUA or ARED-Plus motif ($p = 7 \times 10^{-6}$, 4×10^{-30}, 7×10^{-3}, and 3×10^{-23} for Figure 2, A–D, respectively, by two-sided Welch's *t*-test). There was a consistent decrease in steady state RNA and stability with cluster number from clusters 1 through 3. Cluster 5 motif-containing segments were more active than cluster 3; cluster 4 segments were also

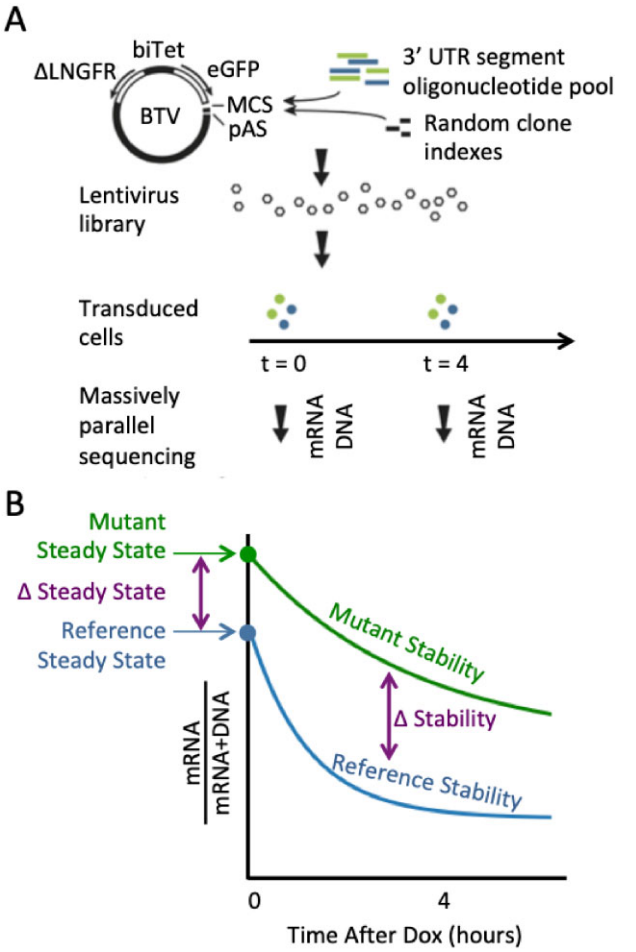

Figure 1 The fast-UTR MPRA. (A) The BTV plasmid includes a bidirectional tetracycline regulated promoter (biTet) that drives expression of enhanced green fluorescent protein (EGFP) and a reference protein (truncated low-affinity nerve growth factor receptor, ΔLNGFR). The EGFP reporter transgene includes a multiple cloning site (MCS) for insertion of 3′ UTR test sequences and a polyadenylation signal (pAS). Pools of 160-mer oligonucleotides containing 3′ UTR segments were inserted into BTV together with random octamer indexes used to identify each clone. Cells were transduced with BTV lentiviral libraries and massively parallel sequencing was used to measure 3′ UTR segment sequences in genomic DNA and mRNA isolated from cells. (B) Steady state mRNA levels were determined from clone read counts for mRNA samples before the addition of Dox. mRNA stability was estimated from mRNA read counts obtained before and 4 h after the addition of Dox to inhibit transcription. The blue line represents a 3′ UTR segment with an element that promotes rapid mRNA decay and the green line represents a sequence with an inactivating mutation of the destabilizing element that increases steady-state mRNA levels and reduces the decay rate.

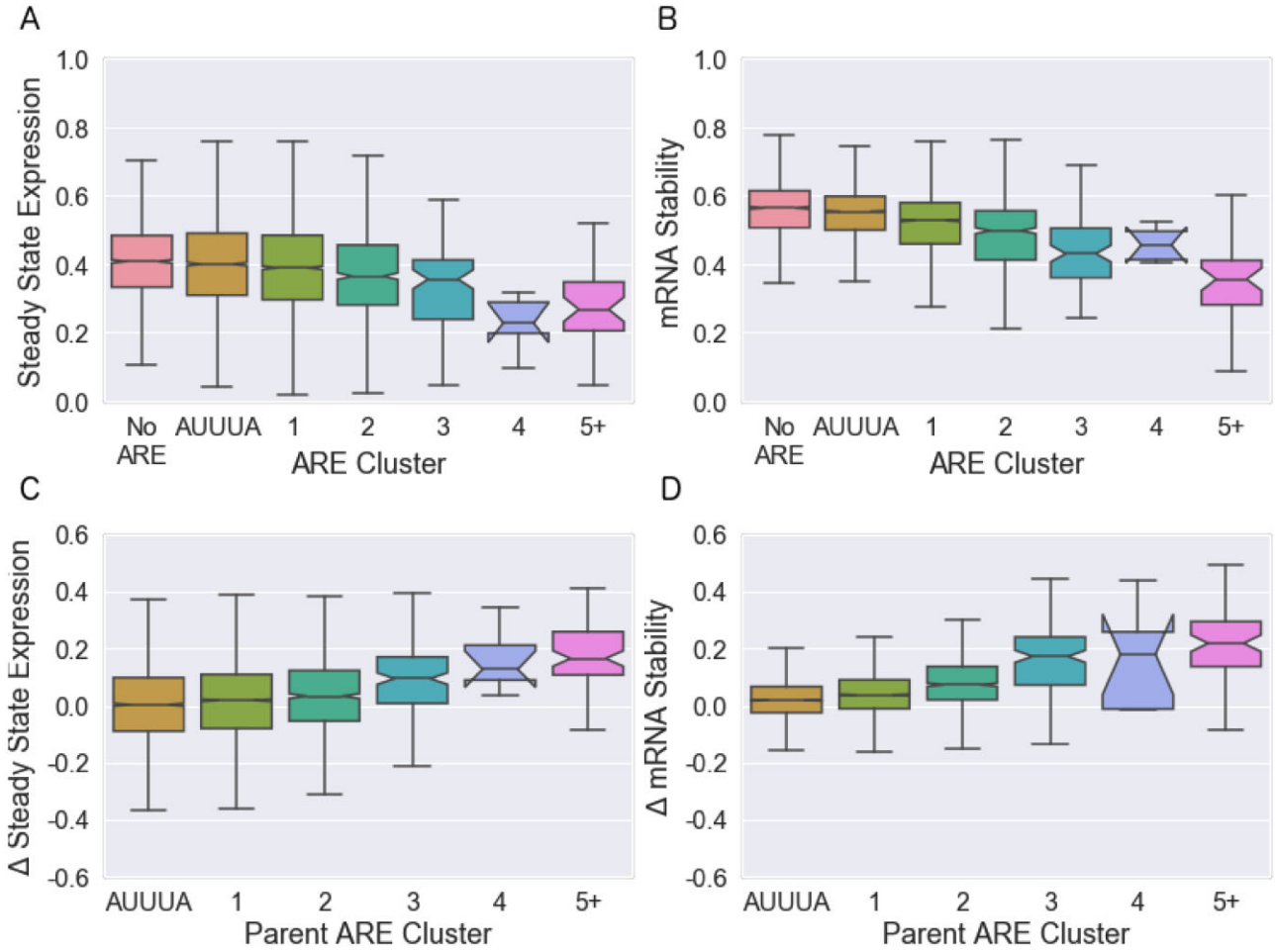

Figure 2 Effect of established ARE clusters on mRNA expression and stability in Jurkat T cells. Boxplots show medians and quantiles, and notches indicate 95% confidence intervals. (A, B) Effects of 3′ UTR segments containing ARED motifs on (A) steady-state RNA expression and (B) RNA stability. (C, D) Effects of mutations in ARED motifs on (C) steady-state expression and (D) RNA stability. **Δ** represents values for segments containing mutations in the ARE motif minus values for the corresponding reference segment with intact ARE motifs. For the 3% of segments with more than one ARE, we categorize them according to the largest ARE present. The "AUUUA" cluster consists of segments that contain the minimal "AUUUA" sequence but lack the flanking sequences required for ARED-Plus. ARE cluster membership was significantly associated with changes in all four measures ($p = 3 \times 10^{-22}$ for A, 6×10^{-149} for B, 1×10^{-31} for C, and 8×10^{-177} for D by linear regression).

associated with low expression and stability although the number of segments in this cluster was small and the confidence interval was larger than for other clusters.

The 160-nt 3′ UTR regions that we tested contained both AREs and surrounding sequences that could also affect reporter expression. To more directly examine the effects of the AREs on expression levels and RNA stability, we leveraged our novel technology to systematically examine the effect of mutations that disrupted the AREs (Figure 2, C and D). As expected, disrupting AREs increased steady state RNA and stability. The effects of mutations were clearly related to the number of AUUUA repeats, as represented by ARED-Plus cluster.

Although Figure 2 shows a clear relationship between ARED-Plus cluster and the effects of mutating AREs, Figure 2 also shows a significant amount of variation within each cluster. While the 95% confidence intervals in Figure 2 are often very small, the interquartile ranges typically vary from 0.1 and 0.2. As we continued to explore the behavior of AREs, we first wished to know how much of the variation within each cluster was caused by technical sources of variability such as sampling error, and how much of this variation could theoretically be improved by some different set of rules for the activity of AREs.

To address this issue, we developed MPRAudit, a novel method to determine the fraction of variance explained by sequence variation in MPRAs and other barcoded assays (?). The premise of MPRAudit is that the technical variation from sequence to sequence can be estimated from the variation from clone to clone for a given sequence segment. If this technical variability can be determined, then we attribute the remaining variability between sequences to be due to sequence variation. When MPRAudit is applied to pairs of segments where the ARE motif has been deliberately mutated, it can determine the fraction of variance caused by the type of mutation and the flanking sequence. We call this quantity the "explainability," denoted b^2. In terms of the technical and total variances, $b^2 \equiv \frac{total - technical}{total}$. When b^2 is close to 0, the technical variance is the source of all variation. When b^2 is close to 1, the variation from sequence to sequence is the primary source of variation and technical sources are small. Table 1 shows that there is a significant amount of sequence variation remaining ($b^2 > 0$) within most of these established groups. This

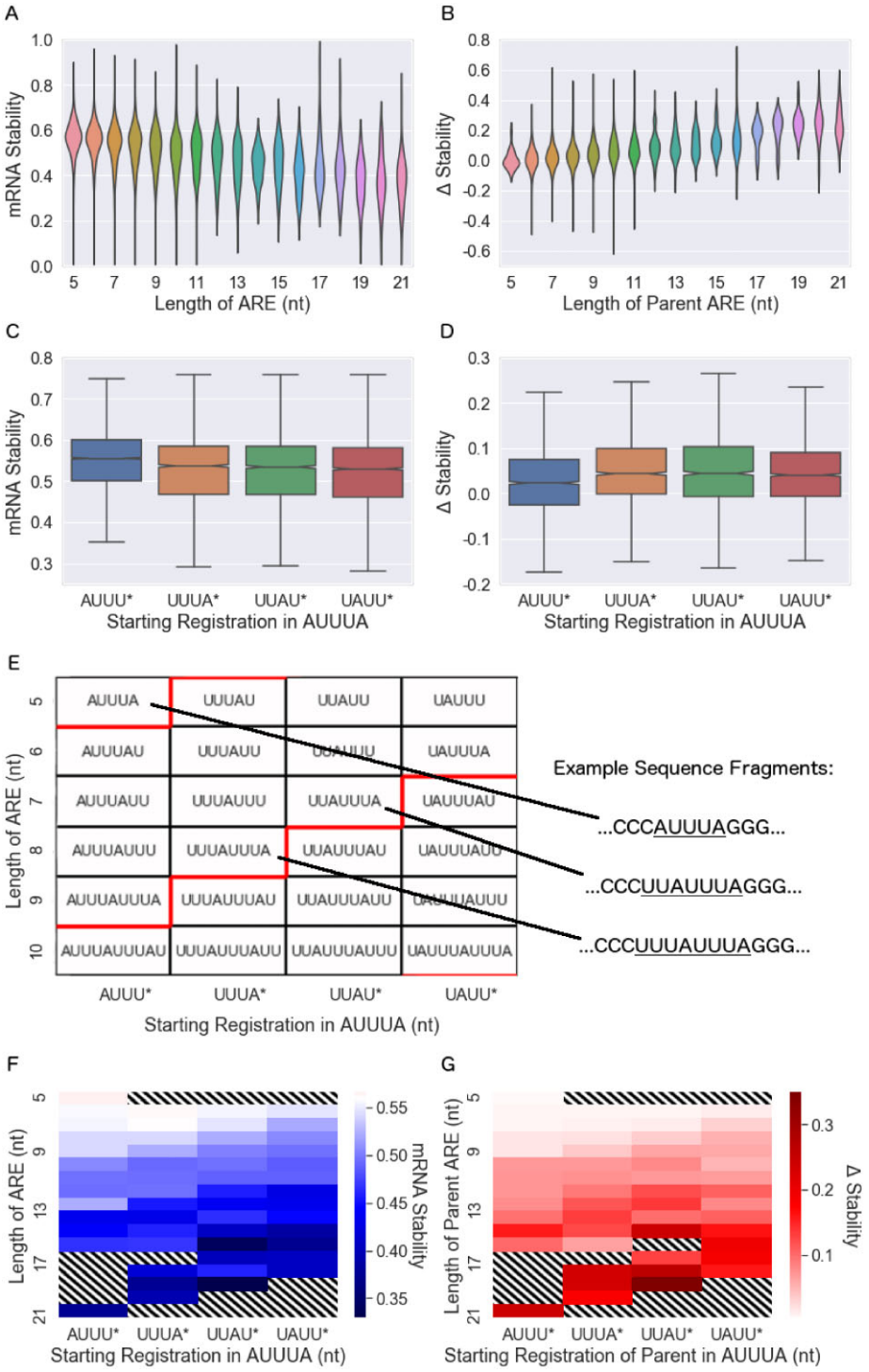

Figure 3 mRNA stability is associated with ARE length and registration. (A, B) Length of the ARE (in nt) is associated with mRNA stability for the segment containing the ARE (A) and with the change in stability that resulted from mutation of the ARE (B). (C, D) Starting registration of the ARE (starting nucleotide of the "AUUUA" pattern) is associated with mRNA stability (C) and mutation-induced change in stability (D). Notches in the boxplots that show confidence intervals are barely visible at this scale. Linear regression shows that the slopes of these datasets (treating the x-axis as a continuous variable) are nonzero, with two-sided *P*-values of $< 1 \times 10^{-300}$, 9×10^{-268}, 2×10^{-58}, and 7×10^{-3}, respectively. (E) Classification of AREs by length and registration, a schematic for panels (F) and (G). In (E–G), the x-axis gives the starting registration of the ARE (starting nucleotide of the repeating "AUUUA" pattern), while the y-axis gives the length of the ARE. In (E), red lines indicate the step between ending registrations (ending nucleotide of the repeating" AUUUA" pattern) of *UUUA (boxes above the red line) and *UUAU (boxes below the red line). (F) mRNA stability is given by the shade of blue according to the length and registration of AREs, following the schematic shown in (E). The mean mRNA stability for segments lacking an ARE is 0.56, represented as white. The diagonal pattern of increasing shades in (F) and (G) show the importance of ARE registration to an ARE of a given length. (G) Mutation-induced change in mRNA stability is given by the shade of red according to length and registration (white corresponds to no change). Standard errors for each entry in each panel are given in the Supplementary Figures S6 and S7. In panels F and G, boxes that weren't studied, boxes that had no corresponding segments in the dataset, and boxes with large standard errors (≥0.07) are excluded and filled with diagonal stripes.

Table 1 Fraction of variance explained by sequence variation within the clusters of AREs defined using ARED-Plus motifs (Bakheet *et al.* 2018)

Cluster	b^2 Δ steady state	b^2 Δ stability	N_{Seqs} steady state	N_{Seqs} stability
AUUUA	0.2269 ± 0.0006	0.057 ± 0.002	4098	3714
ARED-Plus Cluster 1	0.2271 ± 0.0004	0.116 ± 0.001	5278	4725
ARED-Plus Cluster 2	0.214 ± 0.002	0.307 ± 0.005	1122	1025
ARED-Plus Cluster 3	0.32 ± 0.01	0.41 ± 0.02	158	155
ARED-Plus Cluster 4	0.09 ± 0.15	0.16 ± 0.21	9	9
ARED-Plus Cluster 5	0.35 ± 0.02	0.50 ± 0.01	77	73

If b^2 is zero then the variation is due to technical factors; if b^2 is greater than zero then the data support more refined subgroupings of AREs with different functional properties within each class. N_{Seqs} gives the number of distinct sequences that pass our quality control filters.

suggests that there are different categorical subgroupings that can explain more of the effects of ARE-containing sequences than established clusters.

ARE registration and length affect mRNA stability

To create a new ARE classification system that could explain more of the observed effects of ARE-containing sequence segments, we examined three parameters that we hypothesized would correlate with ARE activity: the length of the ARE, the starting registration of the ARE (the nucleotide on which the AUUUA pattern starts), and the conservation status of the ARE. For this analysis, we looked only at perfect matches to a repeating "AUUUA" pattern of length at least 6, plus exact matches to the "AUUUA" pentamer itself.(Lagnado *et al.* 1994; Zubiaga *et al.* 1995; Wiklund *et al.* 2002) Hence, we examined "UUAUUU," despite the fact that it does not contain an "AUUUA," but did not include "UAUUU" because it is too short.

As expected, the length of the ARE was associated with activity. We found a gradual increase in ARE activity as a function of the ARE length (in nucleotides) without obvious abrupt increases at 5-nt (AUUUA pentamer) intervals (Figure 3, A and B shows mRNA stability; Supplementary Figure S2, A and B shows steady state). This suggests that the length of the ARE may provide additional information beyond that provided by the ARED-Plus clusters, which depend on the number of pentamers.

We next considered the starting point of the ARE within the AUUUA pentamer, which we term the "starting registration." We define an ARE starting with "A" to have starting registration 0, and each time a nucleotide is removed from the front of the first pentamer the registration increases by 1. For instance, the sequence "AUUUAUU" has length 7 and registration 0, while "UUUAUUU" has length 7 and registration 1 (and we consider this an ARE despite the absence of an "AUUUA"). Since overlapping pentamers have a periodicity of four nucleotides, an ARE may have one of four possible starting registrations (and one of four possible ending registrations, see Supplementary Figures S6 and S7). We found that starting registration did have an effect on ARE activity, with sequences beginning with "AUUU" being less active than sequences with other registrations (Figure 3, C and D shows mRNA stability; Supplementary Figure S2, C and D shows steady state). We conclude that both ARE length and registration have significant associations with ARE activity in univariate analyses.

In contrast to ARE length and registration, we found that conservation status had little effect on ARE activity. We used phastCons, which identifies conserved elements from 100 vertebrate species (Siepel 2005), to classify AREs as conserved (overlap with phastCons conserved regions) or non-conserved. Surprisingly, there was little difference between the activity of conserved and non-conserved AREs (Supplementary Figure S5). We also considered whether more highly conserved sequences might be more active than other conserved sequences. However, within the set of conserved sequences, we found no significant association of the phastCons lod score with ARE activity in our fast-UTR assay.

Since ARE length and registration were each independently associated with ARE activity, we explored the relationship between these two parameters further. Any sequence of AUUUA pentamer repeats can be classified by its length in nucleotides and its starting (or ending) registration. Figure 3E shows the sequence motifs of several example AREs, organized according to ARE length along the y-axis and ARE starting registration along the x-axis. Motifs of constant ending registration appear along the diagonals; for instance, the boxes directly above (or below) the red lines have the same ending registration. Figure 3, F and G are similar to 3E, except the color of each box represents the mean stability across sequence segments or change in stability with mutation, respectively. As expected, activity increases with increasing ARE length (moving from top to bottom of heat maps). The diagonal contours arise due to effects of registration. Similar effects are apparent when analyzing steady state RNA levels rather than stability, and when using ending registration rather than the starting registration as a parameter (Supplementary Results and Supplementary Figures S6 and S7). These results suggest that registration, in additional to length, should be an integral part of ARE classification and provide insights into underlying mechanisms.

ARE classification methods development and comparison with prior methods

We anticipated that our enhanced understanding of how the length and registration (starting or ending nucleotide) of AREs affect gene expression and stability would allow us to make improvements to existing categories of AREs and to create better prediction methods. To quantify the performance of these categories, we used leave-one-chromosome-out cross validation to train and test predictions for each of the segments with AREs in our dataset. Leaving out one chromosome avoids overfitting from overlapping 3′ UTR segments, since overlapping segments will be within the same chromosome.

We examined correlations between out-of-chromosome predictions and measured data for several classifications of AREs and linear models as described in the Methods section. Using the 5 clusters defined by ARED-Plus (Bakheet *et al.* 2018) (as in Figure 2, A and B) led to a modest correlation with steady state mRNA level and a somewhat higher correlation with mRNA stability (columns 1 and 2 of Table 2). Correlations were somewhat higher for mutation-induced changes in mRNA level and stability (columns 3 and 4), since these measures depend more critically on the targeted ARE itself than on the other sequences within the 3′ UTR segments. We also examined the "AREScore" scoring

Table 2 Out-of-chromosome correlations between predictions and measured data for existing and novel ARE categories (i–iii) and a prediction method (iv)

#	Category or method	1. Steady-state expression	2. Stability	3. Δ Steady-state	4. Δ Stability
i	ARED-Plus (Bakheet *et al.* 2018) Clusters	0.09	0.21	0.14	0.29
ii	AREScore (Spasic *et al.* 2012) Algorithm	0.09	0.22	−0.04	−0.01
iii	Naive Effective Length Pentamers	0.11	0.30	0.15	0.35
iv	Effective Length (nt) (Length + Registration)	0.14	0.31	0.16	0.37
v	Lasso K-Mer Regression ($K = 5$)	0.26	0.49	0.13	0.34

See Supplementary Table S5 for further categories and Supplementary Table S6 for Beas2B data.

Table 3 Fraction of sequence-driven variance explained by given classification systems, as calculated by MPRAudit

Categorization	b^2 Expression	b^2 Stability	b^2 Δ Expression	b^2 Δ Stability
ARED-plus	0.00 ± 0.02	0.05 ± 0.02	0.05 ± 0.09	0.39 ± 0.12
Effective length pentamers	0.02 ± 0.02	0.06 ± 0.02	0.06 ± 0.09	0.54 ± 0.13
Effective length	0.03 ± 0.02	0.07 ± 0.02	0.08 ± 0.09	0.64 ± 0.13

$b^2 = 1$ would imply a perfect model of ARE behavior. The b^2 statistic is first calculated for the entire dataset, ignoring groupings, then calculated within groups, and compared. The fraction explained by categories is calculated as $\frac{b^2_{Total} - b^2_{Groups}}{b^2_{Total}}$.

algorithm (Spasic *et al.* 2012), which performs as well as ARED-Plus on predictions of steady state expression and stability (columns 1 and 2 of Table 2), but poorly predicts the change in steady state expression and stability due to the deliberate mutations in our study (columns 3 and 4). Including both ARE length (in nt) and registration (the starting nucleotide) to form an "effective length" (as suggested by Figure 3, E–G) resulted in significant increases in correlations, making 50% improvements to predictions of both steady-state expression and mRNA stability. To verify statistical significance, we compared the squared out-of-chromosome residuals generated by the ARED-Plus categories to the squared out-of-chromosome residuals generated by the "effective length" method. A Mann-Whitney *U*-test finds that the out-of-chromosome residuals were smaller for the "effective length" categories with *P*-values of $P = 0.010$, 5.1×10^{-30}, and 0.43, 3.1×10^{-8} for steady state mRNA, stability, change in steady state, and change in stability due to mutation, respectively. We also considered a set of related methods based on ARE length, starting or ending registration, and allowance of a mismatch to the strict "AUUUA" motif, but found none that outperformed the "effective length" (length and starting registration) method (Supplementary Table S5).

The wealth of data generated by this protocol also opens the possibility of more complicated sequence-based prediction approaches to 3′ UTR function. As a first step in this direction, we report the results of a method that makes use of k-mer decomposition and regression. This use of k-mers is different from the other methods we have attempted, in that it is agnostic to the presence of AREs in the training data and might be sensitive to the presence of other active elements in the 3′ UTR aside from AREs. Perhaps for this reason we found that the k-mer method had the best performance on predictions of steady state expression and stability. On the other hand, the finite length of the k-mer limits the scope of the model and may prevent it from learning rules for longer and more active AREs. Therefore its performance on predicting the change in steady state and stability with ARE mutation was slightly worse than the performance of the effective length method. The distribution of k-mer effect sizes, the identity of top k-mers, and the effect sizes of ARE-related k-mers are given in Supplementary Figure S9 and Table S3 and S4, respectively. AUUUA is one of the most negative 5-mers, and the other ARE component sequences (UUUAU, UUAUU, and UAUUU) have negative lasso amplitudes as well. On the other hand, substituting a C or G for one of the nucleotides of these four ARE component sequences results in fewer negative amplitudes: only 2 out of 40 of the mutated 5-mers are negative ($p = 1 \times 10^{-4}$ using a Fisher exact test). The mutated 5-mers are also more positive than the 980 remaining 5-mers in the dataset ($P = 0.003$ using a chi-square test with Yates' correction for continuity).

Overall, these results show that the new knowledge introduced by our analysis is helpful for categorizing and classifying AREs and does a better job of predicting gene expression and the effects of mutations than previously established categories.

Table 2 shows that our methods improve predictions, but Table 3 shows that there is still room for further improvement. MPRAudit allows us to calculate an upper limit to the performance of prediction methods, and also to calculate the fraction of variance explained by ARE categories, out of the total possible variance in the dataset caused by sequence variation (removing the fraction caused by known technical factors). It does this by calculating the fraction of sequence variation that remains within each group (restricting the analysis to sequences with AREs) before and after the groupings of Table 2 are applied. We do not apply this technique to the method of k-mers because the method of k-mers does not create groupings and it has a very large number of variables. We find that our predictions of steady state expression and stability explain a small fraction of the total variation caused by differences in sequence, which might be expected by the relatively small correlations in the first three columns of Table 2 (the squared correlation is related to the fraction of variance explained). On the other hand, our novel groupings explain up to two-thirds of the sequence-based variation for the effects of ARE mutations on sequence stability, as the technical sources of variation make up a sizeable fraction of the total variance. All told, we conclude that human 3′ UTRs have many additional unknown regulatory mechanisms, and ARE-mediated decay is just one contributor to 3′ UTR effects on mRNA stability.

As an additional test of our ability to predict mRNA stability from the length and registration of AREs, we analyzed full-length 3'UTRs from publicly available measurements of mRNA half-

lives.(Dölken *et al.* 2010; Tani *et al.* 2012) Since full-length 3'UTRs are usually significantly longer than the 160 nt sequences in our assay, in addition to the "effective length" of the largest ARE, we also considered the "total effective length," the sum of the effective lengths of all AREs in the UTR. We compare these "effective length" methods to AREScore in Supplementary Table S7. In this case, the methods perform with similar accuracy, perhaps because of the added emphasis on very long UTRs with many AREs.

CDE steady-state expression and stability vary with stem length

To determine if the fast-UTR approach could also be applied to other known 3′ UTR elements we next considered CDEs. The fast-UTR library we constructed contained all 345 3′ UTR segments with sequences conforming to a previously-defined degenerate CDE stem-loop motif (Leppek *et al.* 2013). We found that the destabilizing effects of the CDE motif increased with increasing stem length (Figure 4).

As with AREs, we deliberately introduced mutations into the CDE sequences to isolate the effects of these motifs from the overall sequence. For CDEs, two types of mutations were introduced: (1) the central motif UUCYRYGAA was mutated to UagYRYGAA, or (2) the entire motif and surrounding stem were shuffled. For both cases, the differences between mutant and reference sequences were recorded. Since these two types of mutations had similar effects on gene expression and stability (Supplementary Figure S9), we combined them for further analysis. We found that the effect of CDE mutation increased with the length of the outer stem (Figure 4, C and D), consistent with the effects on steady state and stability measurements in Figure 4, A and B: increasing the length of the CDE stem leads to a decrease in steady-state expression and the stability of the mRNA, and the effects of mutations increase with increasing stem length.

Other regulatory elements

Although our analysis focuses on the behavior of AREs and CDEs, Table 3 suggests that there may be many other regulatory elements in the human 3' UTR. To investigate additional possibilities, we analyzed the motifs that were highlighted by Rabani et al. (2017) as being stabilizing or destabilizing in zebrafish and that also happened to be present in our dataset. Supplementary Figure S10, A and B shows that only a few of these motifs are active in human cell lines, most notably the Pumilio and miR430 motifs (in addition to AREs). miR430 does not exist in humans, but the human miRNA miR-302a shares the same seed sequence (Rosa *et al.* 2009) and therefore might account for this finding. Comparison with Supplementary Figure S10, C and D shows that GC-residualization plays an important role in determining the activity of these sequences.

Following up on the potential importance of miRNA targets, we examined the effects of 3' UTR sequences containing predicted targets for miRNAs that were relatively abundant in these cells. We previously found that miRNAs represented by 1% of miRNA reads in small RNA-sequencing were associated with decreased expression of their predicted targets, whereas less abundant miRNAs were not (Zhao *et al.* 2014). Using the Targetscan database (Agarwal *et al.* 2015) to identify associated target sequences, we compared the steady state expression and stability of segments with these target sequences to segments without them. miRNA read counts for the two cell types are given in Supplementary Table S8, and the results of this analysis are given in Supplementary Table S9. Sequence segments had lower steady state expression and mRNA stability for several of the miRNAs we investigated, which confirms the importance of miRNA targets as mRNA regulatory elements. Visualizations of miRNA activity are given as volcano plots in Supplementary Figure S22. A comparison of the activity of miRNAs to ARE cluster and CDE stem length is given in Supplementary Table S13.

The position or relative position of an ARE in a sequence segment has been shown to play a role in its activity in the past (Cottrell *et al.* 2018), but we do not find strong supporting evidence for this in our data (see Supplementary Figure S15 and Table S11).

Analysis of MPRA in Beas2B Cells Confirms Major Findings in Jurkat Cells

To determine whether insights obtained from our studies of Jurkat T cells would apply to another cell type, we used fast-UTR to study the same 3′ UTR segment library in Beas2B human bronchial epithelial cells. The steady-state expression and mRNA stability of individual sequences correlate modestly between the datasets (see Supplementary Figure S21), but analysis of the Beas2B data confirms that the effects of ARE length and registration and CDE stem length are also seen in these cells. In both cell lines, the steady state expression and stability decrease as the cluster of ARE increases in length (Figure 2 and Supplementary Figure S17), and the change in steady state expression and mRNA stability increase with increasing ARE cluster as well. In both cell lines diagonal patterns due to the starting and ending nucleotides of an ARE are observed in Figure 3 and Supplementary Figure S7, and increasing effects for longer AREs are observed in Figure 3 and Supplementary Figure S3. In both cell lines, the steady state expression and mRNA stability decrease as a function of an increasing CDE stem length (Figure 4 and Supplementary Figure S18). And in both cell lines, the "Effective Length" method and "Lasso K-Mer Regression" outperform existing methods for predicting the activity of AREs in out-of-chromosome predictions (compare Table 2, Supplementary Tables S5 and S6).

One difference between these sets of results is that the ARE predictions are better for the Beas2B data than the Jurkat data, while the effects of mutations in Supplementary Figure S17 are also larger than those in Figure 2. In general, the ARE activity appears to be higher in our Beas2B data (see also Supplementary Figure S19), but whether these differences are due to batch effects or cell types is not clear since our experimental design did not include technical replicates for these two cell types.

Discussion

In this study, we have developed and analyzed a massive experimental system for examining 3′ UTR biology. Our approach uncovered novel features that affect the stability and steady-state expression of AREs and CDEs in 3′ UTRs. We show that the length of an ARE, as well as the starting or ending nucleotide, has an effect on gene expression and stability, and verify these findings through designed mutations of the active motif. In CDEs, we similarly show that longer stem loops have an effect on the activity of the motif. Using our recently developed method MPRAudit, we show that a model consisting of ARE length and registration explains up to 64% of the effects of mutations on the stability of AREs.

One way to assess the strength of different active elements is to estimate the effect of a single nucleotide mutation in these elements. As an estimate of the per-nucleotide effect of GC content, we note that in Supplementary Figure S1C stability decreases from roughly 0.5–0 as GC content increases from roughly 0.3 to

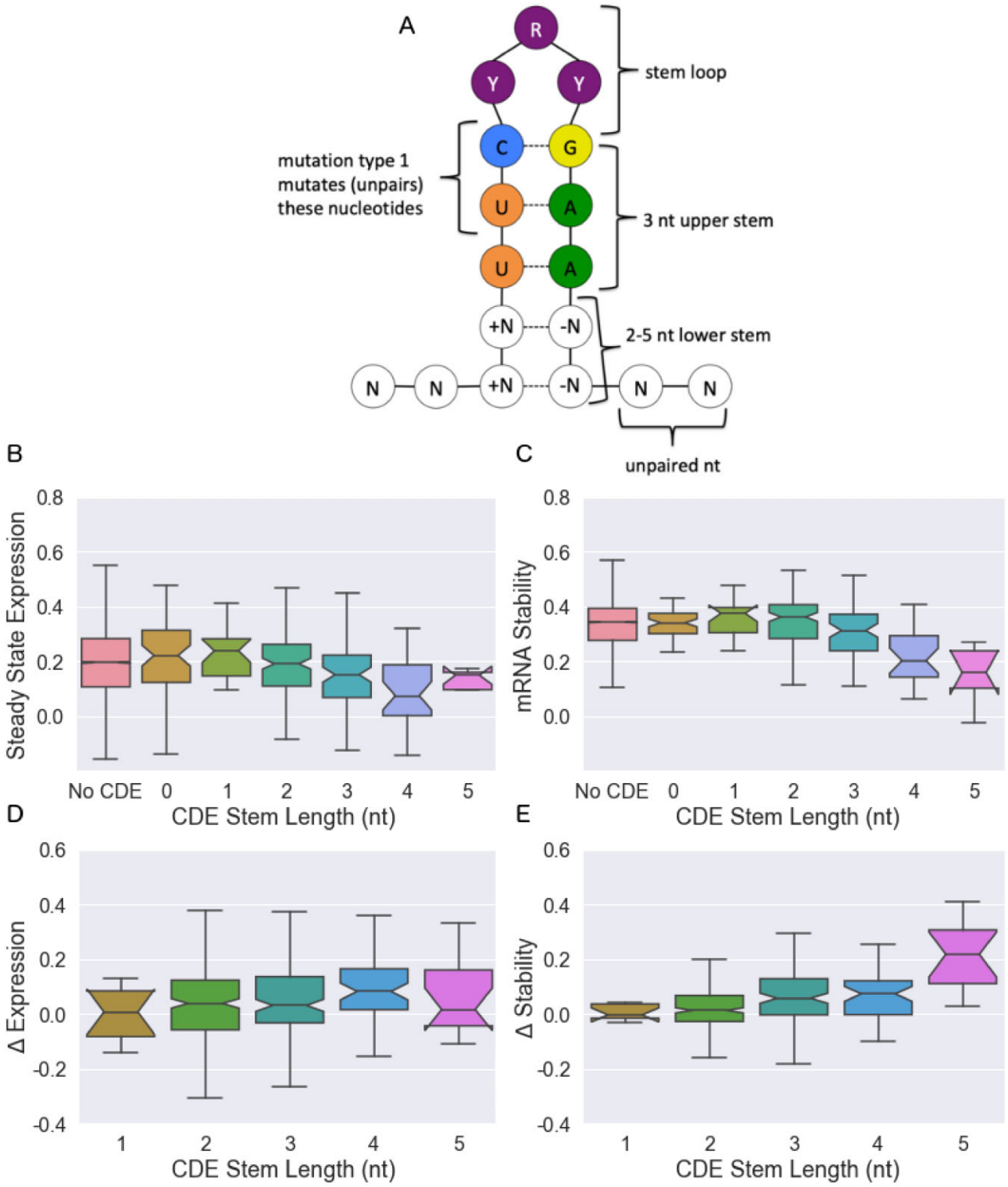

Figure 4 Effect of CDE stem length on mRNA expression and stability. (A) Schematic of the CDE stem loop in our investigation. (B) Steady-state expression, (C) stability, (D) change of steady-state expression with CDE mutation, and (E) change of stability with CDE mutation. Boxplots give medians and quantiles, notches denote 95% confidence intervals. While lower stem lengths of 2–5 nt are considered active, CDE motifs with stem lengths of 0–1 are also contained within the dataset and shown here for reference. Linear regression gives slopes that differ from zero with $p = 1.2 \times 10^{-6}$, 2.8×10^{-7}, 3.2×10^{-3}, and 2.9×10^{-13}, for panels (B–E), respectively.

0.7. This corresponds to a change of 64 nucleotides, or a change of roughly 0.0078 per nucleotide. To estimate the per-nucleotide effect of a mutated ARE, we note that the change in stability for a category 3 cluster in Figure 2D is 0.158, and it is caused by a mutation of just 3 nucleotides (the middle U in each AUUUA pentamer), for a change in stability of 0.053 per nucleotide. Therefore

the ARE mutations have a larger effect than GC content by a factor of approximately 7. To estimate the per-nucleotide effect of a mutated CDE, we note that the average change in stability for 2 nt mutations for length 5 CDEs was 0.2 in Supplementary Figure S9, which gives a change in stability of 0.1 per nucleotide, which is nearly twice as large as the ARE mutation, the largest overall. We can also note for comparison that the most significant 5-mer in our K-mer analysis of mRNA stability (Supplementary Table S3) had a Lasso amplitude of −0.0194, or approximately −0.0038 per nucleotide. Overall, while the effect of GC content is important, we find that mutations to the active ARE and CDE elements have much larger effects on a per-nucleotide basis.

One of our major findings is that the start and end positions of the ARE within the AUUUA pentamer have a significant effect on mRNA stability and steady-state expression. Table 2 and Supplementary Table S5 show that adding the length and registration of the ARE to form an "effective length" improves the prediction correlation by almost 50% for steady state expression and stability, which implies that our new model explains more than twice as much of the variation in the dataset; but the effect of mutations on steady state expression remains difficult to predict. We also found that allowing one mismatch generally gives worse performance than requiring a perfect ARE match. Lasso regression on k-mers performs best on predictions of expression and stability, since it incorporates features from outside the ARE; but it makes slightly worse predictions for the effects of targeted ARE mutations than our methods that rely on ARE length and registration alone. AREScore is unable to predict the effects of our targeted mutations.

The steady state expression and stability of CDEs are consistent with our understanding of their secondary structure. If the central motif of the CDE (UUCYRYGAA) is a binding site for destabilizing proteins, increasing the prominence of its stem loop could have an impact on the binding affinity of the CDE.

This research has many other limitations that suggest directions for future work. We analyze a subset of 3′ UTRs and not the whole genome; and we focus on the behavior of regulatory sequences, but ignore the important contribution of secondary structure. Although we study the behavior of two cell lines, a wider study of more cell types including primary cells or *in vivo* studies with technical replicates for each cell type would enable a wider investigation of general and cell type-specific regulatory effects. Further studies will be required to understand how stimuli that can affect the activity of AREs, for example by affecting the expression or function of ARE binding proteins, contribute to ARE-mediated regulation. Within this dataset, a significant amount of variance is explained by the categories of sequences that we have uncovered, but a significant amount of variance remains unexplained by our work. While our work has produced some general rules for predicting the steady state expression and stability of mRNAs, we note that there are many ARE binding proteins and we are unable to disentangle their effects from one another. It should also be noted that while we have analyzed the activity of several sequence motifs in our dataset, our assay was primarily designed to focus on AREs and CDEs. By making the raw data in this work publicly available, we hope machine learning researchers, statisticians, and geneticists will make further improvements to the models we have devised here.

Data availability

Data has been made available at https://datadryad.org/stash/share/PHVzsyOxRrGcteykzcBvHKO1c_K-FrWD3jAmZe6uTIo.

Data and python notebooks that demonstrate the generation of the main figures in the manuscript are available on our github: https://github.com/david-a-siegel/AU-Rich-Elements.

Supplementary methods, tables, and figures are available online.

Supplementary material is available at *G3* online.

Acknowledgments

The authors would like to thank Wenxue Zhao, Mark Ansel, Steven Brenner, and Ilias Soares for useful discussions.

Funding

D.A.S., A.B., and N.Z. were funded by NIH grants K25 HL121295, U01 HG009080, R01 HG006399, R01 CA227237, R03 DE025665, R01 CA227466, R01 ES029929, R01 GM110251, R01 HL124285, and DoD grant W81XWH-16-2-0018. D.J.E. and O.L.T. were funded by NIH grants R35 HL145235, R01 GM110251, and R01 HL124285.

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

Agarwal V, Bell GW, Nam JW, Bartel DP. 2015. Predicting effective microRNA target sites in mammalian mRNAs. eLife. 4:e05005.

Avsec Ž, Weilert M, Shrikumar A, Alexandari A, Krueger S, *et al.* 2021. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nature Genetics 53:354–366.

Bakheet T, Frevel M, Williams BR, Greer W, Khabar KS. 2001. ARED: human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins. Nucleic Acids Res. 29:246–254.

Bakheet T, Williams BRG, Khabar KSA. 2003. ARED 2.0: an update of AU-rich element mRNA database. Nucleic Acids Res. 31:421–423.

Bakheet T, Hitti E, Khabar KSA. 2018. ARED-plus: an updated and expanded database of AU-rich element-containing mRNAs and pre-mRNAs. Nucleic Acids Res. 46:D218–D220.

Barreau C, Paillard L, Osborne HB. 2005. AU-rich elements and associated factors: are there unifying principles? Nucleic Acids Res. 33:7138–7150.

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 40:e72.

Caput D, Beutler B, Hartog K, Thayer R, Brown-Shimer S, *et al.* 1986. Identification of a common nucleotide sequence in the 3'-untranslated region of mRNA molecules specifying inflammatory mediators. Proc Natl Acad Sci USA. 83:1670–1674.

Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, *et al.* 2012. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell. 149:1393–1406.

Chen CYA, Shyu AB. 1995. AU-rich elements: characterization and importance in mRNA degradation. Trends Biochem Sci. 20: 465–470.

Cottrell KA, Chaudhari HG, Cohen BA, Djuranovic S. 2018. PTRE-seq reveals mechanism and interactions of RNA binding proteins and miRNAs. Nat Commun. 9:301.

Dölken L, Malterer G, Erhard F, Kothe S, Friedel CC, *et al.* 2010. Systematic analysis of viral and cellular MicroRNA targets in cells latently infected with human γ-herpesviruses by RISC immunoprecipitation assay. Cell Host Microbe. 7:324–334.

Glisovic T, Bachorik JL, Yong J, Dreyfuss G. 2008. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett. 582: 1977–1986.

Hodson DJ, Janas ML, Galloway A, Bell SE, Andrews S, *et al.* 2010. Deletion of the RNA-binding proteins ZFP36l1 and ZFP36l2 leads to perturbed thymic development and t lymphoblastic leukemia. Nat Immunol. 11:717–724.

Kreimer A, Zeng H, Edwards MD, Guo Y, Tian K, *et al.* 2017. Predicting gene expression in massively parallel reporter assays: a comparative study. Hum Mutat. 38:1240–1250.

Lagnado CA, Brown CY, Goodall GJ. 1994. AUUUA is not sufficient to promote poly(a) shortening and degradation of an mRNA: the functional sequence within AU-rich elements may be UUAUUUA(u/a)(u/a). Mol Cell Biol. 14:7984–7995.

Leppek K, Schott J, Reitter S, Poetz F, Hammond MC, *et al.* 2013. Roquin promotes constitutive mRNA decay via a conserved class of stem-loop recognition motifs. Cell. 153:869–881.

Litterman AJ, Kageyama R, Tonqueze OL, Zhao W, Gagnon JD, *et al.* 2019. A massively parallel 3′ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. Genome Res. 29:896–906.

Mayr C. 2017. Regulation by 3′-untranslated regions. Annu Rev Genet. 51:171–194.

Peng SS, Chen CY, Xu N, Shyu AB. 1998. RNA stabilization by the AU-rich element binding protein, HuR, an ELAV protein. Embo J. 17:3461–3470.

Rabani M, Pieper L, Chew GL, Schier AF. 2017. A massively parallel reporter assay of 3′ UTR sequences identifies in vivo rules for mRNA degradation. Mol Cell. 68:1083–1094.e5.

Rosa A, Spagnoli FM, Brivanlou AH. 2009. The miR-430/427/302 family controls mesendodermal fate specification via species-specific target selection. Dev Cell. 16:517–527.

Sample PJ, Wang B, Reid DW, Presnyak V, McFadyen IJ, *et al.* 2019. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. Nat Biotechnol. 37:803–809.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, *et al.* 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.

Spasic M, Friedel CC, Schott J, Kreth J, Leppek K, *et al.* 2012. Genome-wide assessment of AU-rich elements by the AREScore algorithm. PLoS Genet. 8:e1002433.

Tani H, Mizutani R, Salam KA, Tano K, Ijiri K, *et al.* 2012. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. Genome Res. 22:947–956.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. J R Statist Soc B. 58:267–288.

Wiklund L, Sokolowski M, Carlsson A, Rush M, Schwartz S. 2002. Inhibition of translation by UAUUUAU and UAUUUUUAU motifs of the AU-rich RNA instability element in the HPV-1 late 3′ untranslated region. J Biol Chem. 277:40462–40471.

Zhao W, Pollack JL, Blagev DP, Zaitlen N, McManus MT, *et al.* 2014. Massively parallel functional annotation of 3′ untranslated regions. Nat Biotechnol. 32:387–391.

Zubiaga AM, Belasco JG, Greenberg ME. 1995. The nonamer UUAUUUAUU is the key AU-rich sequence motif that mediates mRNA degradation. Mol Cell Biol. 15:2219–2230.

Communicating editor: B. Andrews