Published in final edited form as:

Nat Biotechnol. 2014 April ; 32(4): 387–391. doi:10.1038/nbt.2851.

Massively parallel functional annotation of 3' untranslated regions

Wenxue Zhao[1], **Joshua L. Pollack**[1], **Denitza P. Blagev**[1], **Noah Zaitlen**[1], **Michael T. McManus**[2,3], and **David J. Erle**[1]

[1]Lung Biology Center, University of California San Francisco, San Francisco, California, USA

[2]Diabetes Center, University of California San Francisco, San Francisco, California, USA

[3]Department of Microbiology and Immunology, University of California San Francisco, San Francisco, California, USA

Functional characterization of noncoding sequences is crucial for understanding the human genome and learning how genetic variation contributes to disease. 3' untranslated regions (UTRs) are an important class of noncoding sequences but their functions remain largely uncharacterized[1]. We developed a method for massively parallel functional annotation of sequences from 3' UTRs (fast-UTR) and used this approach to measure effects of a total of >450 kb of 3' UTR sequences from >2000 human genes on steady-state mRNA abundance, mRNA stability and protein production. We found widespread regulatory effects on mRNA levels that were coupled to effects on mRNA stability and protein production. Furthermore, we discovered 87 novel cis-regulatory elements and measured the effects of genetic variation within known and novel 3' UTR motifs. This work shows how massively parallel approaches can improve functional annotation of noncoding sequences, advance our understanding of cis-regulatory mechanisms and quantify effects of human genetic variation.

3' UTRs contain cis-regulatory elements that control mRNA stability and translation by interacting with RNA binding proteins, such as AU-rich element (ARE) binding proteins[2] and Pumilio[3], and with multiprotein RNA-induced silencing complexes containing miRNAs[4]. The presence of specific regulatory elements, rather than the length of the 3' UTR, seems to be the major determinant of 3' UTR regulatory activity[5]. 3' UTR elements have been discovered by fine-mapping individual 3' UTRs[6], by identifying sequence[7] or structural[8] motifs enriched in 3' UTRs and by sequencing RNA fragments associated with RNA binding proteins[9]. In addition, miRNA target sites can be predicted using miRNA and 3' UTR sequences[10]. Approaches based on 3' UTR sequence analysis are well suited for

Correspondence should be addressed to: D.J.E. (david.erle@ucsf.edu).

AUTHOR CONTRIBUTIONS
D.J.E and W.Z. conceived of key aspects of the project and designed the experiments. W.Z. and D.B carried out the experimental work. M.T.M. contributed to the design and interpretation of the experiments. D.J.E., J.P., N.Z., and W.Z. analyzed the data. D.J.E and W.Z. wrote the manuscript. All authors reviewed and commented upon the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

detecting elements with common and highly conserved motifs but may fail to identify biologically relevant sequences that lack such motifs. Experimental approaches based on protein binding have been valuable but require prior knowledge of relevant RNA binding proteins. Furthermore, none of these approaches directly quantify the effects of 3' UTR cis-regulatory elements on gene expression or accurately predict how sequence variation affects gene expression. Recent progress in massively parallel oligonucleotide synthesis and massively parallel sequencing provides opportunities for direct functional analysis of non-coding sequences. For example, these technologies have allowed systematic functional dissection of several core promoter[11] and enhancer[12, 13] sequences. Here we develop a massively parallel approach to investigate how 3' UTR sequences regulate gene expression and identify 3' UTR cis-regulatory elements that are sensitive to the effects of sequence variation.

Fast-UTR is based on a bidirectional tetracycline-regulated viral reporter (BTV) that measures the effects of 3' UTR sequences on mRNA and protein production (Fig. 1a **and** Supplementary Figs. 1 and 2). We used massively parallel synthesis to produce pools of 200-mer oligonucleotides containing sequences from 3' UTRs and used massively parallel sequencing to quantify these sequences in RNA and DNA samples isolated from transduced cells. To test fast-UTR we fine-mapped the *CXCL2* 3' UTR, which destabilizes mRNA and reduces protein production (Supplementary Figs. 1 and 2). We first used an oligonucleotide pool containing all possible single nucleotide polymorphisms (SNPs) in a *CXCL2* 3' UTR segment that contains a highly active ARE (ARE1)[14]. We used the resulting fast-UTR library to measure cis-regulatory activity in BEAS-2B immortalized human bronchial epithelial cells. Most SNPs within ARE1 increased steady state mRNA, whereas SNPs outside ARE1 rarely did (Fig. 1b). Ten of the 201 possible SNPs that we tested have been detected in human populations; two of these SNPs had large effects (>50% increase in mRNA, $p < 10^{-9}$), two had smaller effects and six had no detectable effects (Supplementary Table 1 and Supplementary Data 1). These results illustrate how fast-UTR can be used to measure functional consequences of human genetic variation.

To identify other *CXCL2* 3' UTR elements that influence mRNA levels, we designed overlapping oligonucleotides to cover the entire 3' UTR and transduced BEAS-2B cells with a fast-UTR library produced from these oligonucleotides. Massively parallel sequencing revealed that the 42,230 clones contained spontaneous mutations (total of 10,968 deletions and 2,620 point mutations). We compared clones containing spontaneous mutations within each 8 nucleotide sequence interval with clones that had no mutations in that interval (Fig. 1c). In addition to ARE1, mutations in another predicted ARE (ARE4) and two novel elements (N1 and N2) also caused large increases in mRNA levels (>50%). Three other predicted AREs had smaller (ARE2, ARE5) or undetectable (ARE3) effects. Each of the elements with strong activity is highly conserved in vertebrates (Fig. 1c) and in the *CXCL2* paralog *CXCL3* (Supplementary Fig. 3a). Deletion of any of these elements from the full-length *CXCL3* 3' UTR increased reporter protein production, indicating that each element makes a non-redundant contribution to the reduction in protein expression (Supplementary Fig. 3b). We conclude that fast-UTR is effective for identifying functional elements, including novel elements not predicted by available computational approaches.

A systems-level understanding of 3' UTR function requires measurements of the effects of a large set of 3' UTR sequences on steady state mRNA level, mRNA stability and protein production. We designed a set of 160 nucleotide oligonucleotides containing 3' UTR segments covering 3000 highly conserved sequences from 2089 genes (Supplementary Data 2). In total, these segments cover >450 kb (1.5% of all annotated human 3' UTR sequence) and we used fast-UTR libraries to analyze their effects (Supplementary Note 1 and Supplementary Data 3). The 3' UTR segments had a 60-fold range of effects on steady state mRNA levels (Fig. 2a). Most 3' UTR segments reduced reporter mRNA levels below the level seen with the empty BTV vector (no 3' UTR test sequence), but some 3' UTR segments produced modest increases in mRNA levels. We also found substantial effects of 3' UTRs on mRNA stability (Fig. 2b). Although many 3' UTR segments were destabilizing (minimum half-life 1 hour), no individual segment was as destabilizing as the full-length *CXCL2* 3' UTR (half-life 0.5 hour, Supplementary Figure 1), probably because this full-length 3' UTR has multiple destabilizing elements (Supplementary Figure 3). Steady state mRNA levels had a highly significant ($p < 10^{-197}$) correlation with mRNA stability (Fig. 2c). However, nine 3' UTR segments gave low mRNA steady state levels (ranging from 4–20% of the median value for all segments) but did not reduce stability (mRNA half-lives ≥ 3 h), suggesting that these 3' UTR sequences affected mRNA production. Decreased mRNA production could result from altered mRNA processing (for example, alternative polyadenylation, which would produce a transcript that would not be amplified using the fast-UTR PCR primers) or by reduced transcription.

We next used fast-UTR with flow cytometric sorting to assess the effects of 3' UTR sequences on protein production (Fig. 2d). Mean normalized reporter protein level in cells transduced with the conserved 3' UTR library was 80% of the level seen in cells transduced with empty BTV (Supplementary Fig. 4). We used flow cytometric sorting to enrich for 3' UTR sequences associated with relatively high or low reporter protein levels compared with other sequences in the fast-UTR library. Massively parallel sequencing showed that 305 segments (11%) were enriched by ≥10-fold in cells with lower reporter protein levels (Supplementary Data 4). Individual flow cytometric analysis of 21 randomly selected segments from this group showed that these segments had substantial effects (median 5.3-fold lower protein levels than empty BTV) (Fig. 2e). 568 segments (20%) were enriched by ≥10-fold in cells with relatively high reporter protein levels (Supplementary Data 4); the median increase in protein levels for 17 randomly selected segments from this group was 1.2-fold (Fig. 2e). 3' UTR sequences enriched in cells with relatively high reporter protein production tended to give higher levels of reporter mRNA and longer mRNA half-lives, whereas 3' UTR sequences enriched in cells with relatively low reporter protein production tended to give lower levels of reporter mRNA and shorter half-lives (Fig. 2f,g, $p < 10^{-30}$ for all comparisons). Our findings with this large and diverse set of 3' UTR sequences suggest coupling between effects on mRNA stability, steady state mRNA levels and protein production. Similar coupling has been described for specific miRNA targets[15–17]; our work suggests that this coupling is a general characteristic of many 3' UTR regulatory elements. Although we did not directly measure translation, our results strongly suggest that 3' UTRs rarely had isolated effects on translation that were not coupled to mRNA destabilization and a reduction in steady state mRNA levels.

We used fast-UTR to measure effects in two other cell types and found highly significant ($p \leq 10^{-243}$) correlations between 3' UTR sequence activities across cell types, although global differences between cell types were also apparent (Fig. 2h–j). For example, 12% of 3' UTR segments increased mRNA levels by ≥2.5-fold over median levels in WiDr colorectal adenocarcinoma cells but no 3' UTR sequences had this effect in BEAS-2B cells. Unexpectedly, 3' UTR sequences had similar effects in WiDr cells and Jurkat T cell leukemia cells despite the distinct lineages of these cells (Fig. 2j). Using an unbiased approach for *de novo* motif discovery, we identified a set of related AU-rich motifs that were associated with differential 3' UTR segment activity in the three cell types (Online methods; Supplementary Fig. 5). Different ARE binding proteins have distinct effects on mRNA stability[2] and it is possible that differences in RNA binding protein expression or activity contributed to the observed cell type-dependent effects. As fast-UTR can be used in diverse cell types, this approach will be useful for studying how different cellular contexts affect 3' UTR regulatory activity.

To discover functional cis-elements in the conserved 3' UTR segments, we used error-prone PCR to generate large numbers of mutations and measured their effects on mRNA stability (Supplementary Note 1). Many active, mutation-sensitive elements identified by fast-UTR precisely co-localized with known RNA binding protein motifs or predicted miRNA targets (Fig. 3a). As expected, mutations in motifs recognized by destabilizing RNA binding proteins (AREs, constitutive decay element stem-loop motifs and the canonical human Pumilio motif) generally increased mRNA stability whereas mutations in the destabilizing CU-rich element motif[18] decreased stability (Fig. 3b and Supplementary Fig. 6). We also discovered that mutations in the UGUACAG motif increased mRNA stability (Fig. 3b **and** Supplementary Fig. 7a,b). This motif is similar to the canonical human Pumilio motif (UGUAAAUA)[9, 19, 20] and corresponds precisely to a *Drosophila* Pumilio motif identified by RNACompete[21], therefore it seems likely that this motif serves as an alternative binding site for Pumilio proteins or related RNA binding proteins in human cells. Mutations in short conserved hairpin sequences predicted by EvoFold[22] also tended to increase stability (Fig. 3b). Analysis of 1503 miRNA targets predicted by TargetScan[23] showed that effects of mutations in these sites depended on levels of the corresponding miRNAs (as estimated by RNA sequencing, Supplementary Data 5) and the miRNA target sequence context and conservation (assessed by the TargetScan context+ score) (Fig. 3c). In some cases, fast-UTR showed marked differences in activity among predicted miRNA targets with the same seed sequence match (e.g., miR-17 family targets in Fig. 3a) or sequences containing identical RNA binding protein motifs (e.g., Pumilio motifs in Supplementary Fig. 7b, c). These differences suggest that fast-UTR is sensitive to effects of sequences flanking miRNA seed sequence binding sites and core protein-binding motifs. Such flanking sequences might modulate element activity by participating in interactions with trans-regulatory factors or affecting mRNA secondary structure and regulatory element accessibility.

Since fast-UTR does not depend upon prior knowledge of sequence motifs, it is well suited for *de novo* identification of 3' UTR cis-regulatory elements. We identified 106 destabilizing elements and 44 stabilizing elements within the conserved 3' UTR segments (5% false discovery rate, Supplementary Data 6 and 7). 55% (83/150) of these elements did not contain predicted miRNA targets or motifs known to be recognized by RNA binding proteins (Supplementary Table 2). We tested individually six novel elements and two elements with known RNA binding motifs and found that seven of the eight reduced reporter protein production in all three cell types studied (Fig. 3d). Thus, fast-UTR can identify many previously unannotated regulatory elements, which might be non-canonical binding sites for known miRNAs or RNA-binding proteins or sites recognized by unknown trans-regulatory factors.

The role of 3' UTRs in gene regulation is an area of ongoing research. A strength of fast-UTR is that it directly measures the effects of 3' UTR sequences and sequence variation on gene expression at high throughput. The identification of many known miRNA targets and RNA binding protein motifs using fast-UTR supports the relevance of fast-UTR results to endogenous mRNAs. An analysis of correlations between endogenous mRNA stability and fast-UTR stability measurements for 3' UTR segments from the same mRNAs provides further evidence in support of the relevance of fast-UTR (Supplementary Note 2 and Supplementary Table 3). However, complementary approaches are required to understand how elements discovered by fast-UTR are affected by endogenous mRNA secondary structure, interactions with neighboring 3' UTR elements, and alternative polyadenylation. Several studies have examined cases in which alternative polyadenylation increases stability through loss of destabilizing elements[24, 25], although a recent report identified many cases in which shorter isoforms produced by use of proximal polyadenylation sites were less stable, suggesting a loss of stabilizing elements[26]. This is of interest given the substantial number of stabilizing elements we identified with fast-UTR. Cell type-specific alternative polyadenylation[27] could cause cell type-specific effects on 3' UTR regulatory activity that would not be measured by fast-UTR.

Moving from genetic variant associations to causation is a major challenge facing human genetics, and fast-UTR could help to identify variants with a causal role in disease as 3' UTRs contain many variants that are associated with human diseases[28]. Fast-UTR data also promises to be useful for refining computational methods for predicting functional effects of sequence variation and for designing and testing 3' UTR regulatory sequences that could be useful for synthetic biology. Furthermore, fast-UTR can be used in different cell types to investigate tissue-specific gene regulation and could be modified to analyze 5' UTR sequences. In conclusion, our work provides an example of how massively parallel functional assays provide powerful tools for genome annotation, for investigation of cis-regulatory mechanisms and for direct measurement of functional effects of human genetic variation.

METHODS

Construction of the BTV 3' UTR reporter

We constructed the BTV reporter plasmid by replacing the constitutive bidirectional promoter in BdLV[29] with a bidirectional tetracycline responsive promoter[30]. We inserted a linker with *Mlu*I, *Sbf*I and *Pac*I sites between the EGFP open reading frame and the polyA signal for subcloning of 3' UTR sequences. To test full-length *CXCL2* and *CXCL3* 3' UTRs (beginning after the stop codon and ending before the poly A signal sequence), we amplified these sequences from human genomic DNA (G304A, Promega) and cloned the products into BTV. To analyze mutant *CXCL3* 3' UTRs, we deleted selected elements by PCR mutagenesis. For individual analysis of short 3' UTR segments containing active elements, we designed appropriate oligonucleotide pairs, annealed them, and ligated the annealed oligonucleotides into BTV. We used short segments with mutations within the active elements as controls. All constructs were validated by DNA sequencing. Sequences of oligonucleotides used for generating these constructs are included in Supplementary Table 4.

Lentivirus production and cell transduction

We produced lentiviruses by co-transfecting 293T cells with 3 μg of BTV reporter together with 1 μg each of pMDL, p-RSV, and p-VSV-G using Fugene HD (Roche). We harvested conditioned medium 48 h later and used it immediately or froze it at −80°C for later use. We used lentiviral preparations to transduce BEAS-2B, Jurkat or WiDr cells carrying a tetracycline transactivator (tTA) transgene. BEAS-2B-tTA cells[31] were a generous gift from A. Shyu. We produced tTA-expressing Jurkat T cells and WiDr colorectal adenocarcinoma cells by transducing parental lines (obtained from the UCSF Cell Culture Facility) with a tTA lentivirus and then screening single clones for tTA activity. For transductions, we added lentivirus-containing conditioned medium diluted in working medium (1:1) with polybrene (8 μg/ml final) to tTA-expressing cells. Lentivirus-containing medium was replaced with fresh medium after 24 h. For analysis of mRNA stability, cells were cultured in medium alone or medium supplemented with doxycycline (1 μg/ml) for 0, 2, 4, or 8 h. Cell lines were not authenticated or tested for mycoplasma contamination.

Analysis of cis-regulatory effects of individual 3' UTR sequences

To analyze effects of individual full-length 3' UTRs or 3' UTR segments, we harvested transduced cells 72 h after infection, stained cells with Alexa647-conjugated ME20-4 anti-LNGFR antibody, fixed cells with 1% paraformaldehyde, and analyzed cells using a FACSCanto flow cytometer (Becton Dickinson). Flow cytometric analyses were performed in triplicate. We normalized the median ratio of GFP/LNGFR fluorescence for transduced (LNGFR-positive) cells relative to ratios obtained using the empty BTV reporter (no 3' UTR test sequence inserted, defined as 100%) and a version of the reporter lacking the GFP transgene (defined as 0%). To analyze effects on reporter mRNA stability, we extracted RNA (Qiagen RNeasy Mini/Midi Kit), reverse transcribed RNA to cDNA (Invitrogen SuperScript III First-Strand Synthesis System), and analyzed cDNA by SYBR green quantitative real-time PCR using primers for GFP and LNGFR (used as a reference transcript). We calculated reporter mRNA levels using the ΔΔCt method[32].

Design of the *CXCL2* and *CXCL3* fast-UTR libraries

We analyzed a highly active segment of *CXCL2* (nucleotides 589–716 from NM_002089) using a fast-UTR systematic mutagenesis library that included all 201 possible single base substitutions within a 67-nt region containing a 21 nt ARE (ARE1) and 46 nt of flanking sequence. To analyze the complete 696 nt *CXCL2* 3' UTR, we produced another library with 205 overlapping 128 nt segments spaced at intervals of 4 nucleotides. To improve coverage of sequences near the 5' and 3' ends of the 3' UTR, we included oligonucleotides containing <128 nucleotides (4–124 nucleotides) of sequence from the ends of the 3' UTR. To maintain a constant 128 nt test sequence size, we padded these oligonucleotides by adding a sequence from the *CXCL7* 3' UTR (NM_002704, 475–602) that had minimal regulatory effects in preliminary experiments. We used a similar approach to produce the *CXCL3* 3' UTR library. For each chemokine library, each oligonucleotide included 128 nucleotides of 3' UTR test sequence, 5’ and 3' primer recognition sequences used for PCR amplification, and a 12-nt unique index that could be used to identify each test sequence.

Design of the conserved 3' UTR segment library

We used PhastCons (downloaded from the UCSC Genome Browser, phastConsElements17way track based on the NCBI36/hg18 human genome assembly) to identify conserved elements within 3' UTRs. We selected the 1000 most conserved 3' UTR sequence elements from each of three size classes (21–40, 41–80, and 81–160 nt) for analysis. The resulting 3000 elements were drawn from a total of 2089 human genes. We designed oligonucleotides with 160-nt 3' UTR segments containing the conserved elements. In some cases, two or more elements from the same 3' UTR could be included in a single oligonucleotide. The complete set of 2828 segment sequences is shown in Supplementary Data 1. 20-nt primer recognition sequences were added to the 5' and 3' ends of each segment for PCR amplification.

Production of fast-UTR libraries

We used oligonucleotide pools that were produced by massively parallel synthesis using Agilent microarrays. Oligonucleotides (0.05 pmol) were amplified by PCR (Kapa Biosystems HiFi Library Amplification Kit; 98 °C for 45 s followed by 15 cycles of 98 °C for 15 s, 68 °C for 30 s, and 72 °C for 30 s followed by 72 °C for 60 s). For one experiment, the conserved 3' UTR segment oligonucleotide pool was amplified using error-prone PCR (Agilent GeneMorph II Random Mutagenesis Kit, 20 cycles) to introduce mutations at a higher frequency. PCR primer sequences are shown in Supplementary Table 4.

PCR products resulting from amplification of oligonucleotide pools were digested with *Mlu* I and *Sbf* I (which recognized sites in the PCR primers), purified from a 2% TAE gel, and ligated into BTV plasmid with T4 DNA ligase. Ligation mixtures were introduced into XL-1 blue cells by electroporation. After overnight culture at 37 °C, all colonies were harvested together by rinsing plates with liquid LB for preparation of plasmid libraries. For the chemokine and high fidelity conserved libraries, a second ligation step was performed to introduce random octamer indexes for identification of individual clones. For the error-prone PCR conserved segment library, random octamer sequences were incorporated into one of the PCR primers, eliminating the requirement for a second ligation step. We used plasmid libraries for lentivirus library production.

Fast-UTR assays

To analyze effects of 3' UTR segments on steady state mRNA levels, we transduced 10^6 tTA-expressing BEAS-2B, Jurkat, WiDr cells with fast-UTR lentiviral libraries. We analyzed transduced cells by flow cytometry to ensure adequate transduction efficiency

(>50% LN-GFR$^+$ cells). We passaged cells several times with extensive washing to remove residual lentiviral RNA before harvesting cells for isolation of cellular RNA and genomic DNA. To analyze effects of 3' UTR segments on mRNA stability, we harvested replicate plates of cells after 0, 2, 4, and 8 h treatment with doxycycline (1 μg/ml). To enrich for 3' UTR segments associated with higher or lower reporter protein production, we transduced 10^8 BEAS-2B cells with the conserved 3' UTR segment lentiviral library at a concentration that resulted in ~5% transduction efficiency as measured by LNGFR staining to minimize the number of cells with >1 lentiviral reporter. We enriched for reporter-containing cells using anti-LNGFR antibody and anti-mouse IgG1 MACS MicroBeads (Miltenyi). We then used a Becton Dickinson FACSAria III flow cytometer to sort 5×10^6 cells based on expression of GFP and LNGFR. DNA was isolated from cells with high reporter protein levels (top 15% of GFP/LNGFR ratios) and low reporter protein levels (bottom 15% of GFP/LNGFR ratios). After amplification using PCR primers recognizing sequences flanking the 3' UTR test segments, we re-cloned segments from the high and low sort gates into BTV for a second round of transduction and sorting.

Massively parallel sequencing

We used massively parallel sequencing to analyze RNA and DNA from cells transduced with fast-UTR lentiviral libraries. RNA was transcribed to cDNA. cDNA and genomic DNA were amplified using PCR primers that included multiplexing indexes, sequencing primer recognition sites, and attachment sequences (Supplementary Table 4). Amplified material was gel purified and sequenced using an Illumina HiSeq 2000 sequencer. We used a custom read 1 primer (GGTGTTCAGTGTACCAGTTCGCGTAGGTTCAGA) together with standard read 2 and multiplex index read primers.

Fast-UTR data analysis

We used paired end reads (105 nt per read) to analyze the test 3' UTR sequence and assign each read pair to a specific clone (using the random octamer clone index). We aligned reads and determined consensus sequences with Bowtie2[33] and samtools[34] and then aligned each consensus sequence to the designed sequence using needle[35] to identify mutations. We then used read counts to estimate 3' UTR effects on steady state mRNA, mRNA stability, and protein production. For all clones in each RNA or DNA sample, we normalized read numbers by dividing by the total number of reads for that sample. Clones with few reads (mean <10 reads/sample) were excluded from further analysis.

To estimate steady state mRNA effects, we determined the number of reads from cDNA (made from RNA prepared from cells not treated with doxycycline). To account for differences in lentiviral titer and transduction efficiency, we normalized this value by dividing it by the number of reads from genomic DNA obtained from a replicate culture. The effects of each segment were determined from the median mRNA/DNA ratios of all clones containing that segment. Segments represented by fewer than 10 clones were excluded from analysis.

To estimate mRNA stability, we calculated the median ratios of RNA reads from samples collected after 2, 4, or 8 h of doxycycline treatment to RNA reads from a sample not treated with doxycycline. We used qRT-PCR (with GFP PCR primers) to measure the decrease in overall amount of reporter mRNA following doxycycline treatment, and then multiplied RNA read ratios by the amount of reporter mRNA in each sample to determine the fraction of each 3' UTR segment remaining at each time point. We estimated reporter mRNA half-lives γ for each segment by fitting to an exponential decay model $mRNA(t) = mRNA(0) \bullet 2^{-(t/\gamma)}$ (model 1). Since doxycycline did not completely prevent reporter transcription (>90% reduction), we did not use low mRNA ratios $(mRNA(t)/mRNA(0)) < 1/8)$ in half-life calculations for model 1. Alternatively, we calculated half-lives by using a model $mRNA(t) = (mRNA(0) - l) \bullet 2^{-(t/\gamma)} + l$ that incorporated a constant transcriptional leak l and retained all mRNA ratios (model 2). Models 1 and 2 gave similar results ($R = 0.87$ with $l = 0.03$), but model 2 estimates for mRNAs with short half-lives were quite sensitive to the value used for the leak parameter. Since the amount of leak is difficult to determine precisely and may vary between experiments or between different cells from the same experiment, we used model 1 to generate the estimates presented here.

To identify 3' UTR segments that were enriched in sorted cell populations, we determined the ratio of normalized read counts ([reads from high reporter protein population]/[reads from low reporter protein population]) for each 3' UTR segment. We classified segments with ratios ≥ 10:1 as enriched in the high reporter protein population and those with ratios ≤ 1:10 as enriched in the low reporter protein population. 3' UTR segments with <500 total normalized read counts in these two sorted populations were excluded.

Discovery of motifs with different activity in different cell types

We used the Discriminative Regular Expression Motif Elicitation (DREME)[36] to search for motifs that were present in segments with different activity in the three cell types used for fast-UTR measurements of steady-state mRNA. We rank-transformed steady-state mRNA levels (i.e., for each cell type, we ranked segments from lowest to highest steady-state mRNA level). For each segment, the difference in activity between two cells *A* and *B* was defined as ($\text{rank}_A - \text{rank}_B$). DREME was used to identify octamer motifs that were differentially represented in segments with relatively low mRNA levels in cell *A* (bottom 15% of $\text{rank}_A - \text{rank}_B$ values) versus segments with relatively low mRNA levels in cell *B* (top 15% of $\text{rank}_A - \text{rank}_B$ values). Similar motifs were identified using larger segment sets (e.g., 33% versus 15%, not shown). The –k 8 and –norc options were used to direct DREME to search for octamer motifs on the appropriate strand only; other options were left at default values. For each pairwise comparison, we present the motif associated with the lowest *p* value. Each comparison also revealed other motifs with significant but substantially higher *p* values; all of these motifs were AU-rich (not shown).

Analysis of mutation effects

We identified functional elements by comparing mutant clones to clones lacking mutations. For analysis of designed single nucleotide mutations in the highly active region of *CXCL2*, we compared all clones with a given mutation to all wild type clones (no designed mutations). Mutations introduced by synthesis of by standard or error-prone PCR generally did not produce sufficient numbers of clones with specific mutations to determine the effects of individual mutations. Instead, we used an 8 nt sliding window and classified all clones as either mutant or wild type depending upon the presence of mutations in that interval. We then measured the median differences between mutant and wild type clones. We measured effects on mRNA levels (represented by the ratio of steady state mRNA to DNA) and/or mRNA stability (represented by the ratio of mRNA after 4 h of doxycycline treatment to steady state mRNA). Since values were not always normally distributed, we applied a two-sided Wilcoxon rank sum test to mutant and wild type ratios to estimate mutation effects (determined from the location shift) and generate *p* values. For analysis of effects of mutations on elements matching known motifs, we excluded elements represented by fewer than 5 mutant and 5 wild type clones.

For de novo element discovery, we considered all sequence intervals of length 6, 8, 10, 12, 14, 16, 18 and 20 nt, excluding those represented by fewer than 20 mutant and 20 wild type clones. We calculated mRNA stability (defined as [RNA reads from cells treated with doxycycline for 4 h]/[RNA reads from untreated sample]) for each clone. For each interval, we determined the median stability difference between mutant and wild type clones and computed the *p* value using the two-sided Wilcoxon rank sum test. We calculated *q* values using the false discovery rate method. When multiple sequences with significant *q* values overlapped, we selected the element with the best false discovery rate (lowest *q* value).

Identification of elements conforming to previously determined motifs

We used TargetScan[23] version 6 to predict miRNA targets (predictions for each chromosome downloaded from http://www.targetscan.org/vert_61/ucsc/hg19/hg19ConsChr1.bed through http://www.targetscan.org/vert_61/ucsc/hg19/hg19ConsChrY.bed on July 28, 2013). We obtained ARE motifs from the AREsite database (http://rna.tbi.univie.ac.at/cgi-bin/AREsite.cgi). We used a recently identified 13 nt stem-loop motif[37] to identify CDEs. We used the UGUANAUA motif identified in human Pumilio immunoprecipitation experiments[9, 19, 20] to predict elements recognized by Pumilio. We used the $(C/U)CCAN_xCCC(U/A)(C/U)_yUC(C/U)CC$ consensus sequence that has been shown to increase mRNA stability[38] to identify CU-rich elements. We obtained a set of 193 RNA binding domain motifs identified by RNAcompete, SELEX, and immunoprecipitation from the Catalog of Inferred Sequence Binding Preferences of RNA binding proteins (http://cisbp-rna.ccbr.utoronto.ca/, accessed August 17, 2013) and used FIMO[39] to identify matching elements in the 3' UTR segments included in the fast-UTR libraries. We also searched for the sRSM1 stabilizing stem-loop motif (recognized by HNRPA2B1) and 5 other 3' UTR stem-loop motifs recently found to be enriched in stable (sRSM2-4) or unstable (sRSM7-8) mRNAs[8]. We identified all cases in which elements identified by fast-UTR had a ≥5 nt overlap with miRNA targets predicted by TargetScan or known RNA binding motifs. We used the TruSeq Small RNA Sample Preparation Kit (Illumina) to profile miRNAs in BEAS-2B cells. Sequencing reads were deposited in the NCBI Short Read Archive (accession number SRX463338; http://www.ncbi.nlm.nih.gov/sra/?term=SRX463338).

Comparison of fast-UTR results with endogenous mRNA stability

Microarray-based measurements of endogenous mRNA stability were obtained from two previous reports: Yang *et al.*[40] (mRNA levels after actinomycin D treatment of HepG2 hepatocellular carcinoma cells) and Goodarzi *et al.*[8] (4-thiouridine labeling of mRNA in MDA-MB-231 breast cancer cells). For the Yang *et al.* data set, we used mRNA half-lives reported in Supplementary Table 9 of that publication. For the Goodarzi *et al.* data set, we downloaded microarray data from NCBI Gene Expression Omnibus (GEO accession number GSE35800) and calculated relative stability values as the slope of log signal intensity versus time (0, 1, 2 and 4 h). When there were multiple microarray probes for a single gene symbol, we used the mean value for all probes. We matched mRNA probes with fast-UTR segments by gene symbol and calculated Spearman correlations for pairwise comparisons of fast-UTR stability values with each of the two endogenous mRNA datasets. We also determined the correlation between the two endogenous mRNA datasets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank K. M. Ansel, R. Barbeau, A. Barczak, S. E. Brenner, K. Chin, C. Eisley, E. Ostrin, S.-W. Park, A. Sayce, T. Wang, and G. Zhang for advice and technical assistance. This work was supported by research grants (D.J.E.) and a training grant (D.B.) from the NIH.

References

1. Zhao W, Blagev D, Pollack JL, Erle DJ. Toward a systematic understanding of mRNA 3' untranslated regions. Proc Am Thorac Soc. 2011; 8:163–166. [PubMed: 21543795]
2. Barreau C, Paillard L, Osborne HB. AU-rich elements and associated factors: are there unifying principles? Nucleic Acids Res. 2005; 33:7138–7150. [PubMed: 16391004]
3. Wickens M, Bernstein DS, Kimble J, Parker R. A PUF family portrait: 3'UTR regulation as a way of life. Trends Genet. 2002; 18:150–157. [PubMed: 11858839]
4. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009; 136:215–233. [PubMed: 19167326]
5. Sharova LV, et al. Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. DNA Res. 2009; 16:45–58. [PubMed: 19001483]
6. Chen CY, Shyu AB. AU-rich elements: characterization and importance in mRNA degradation. Trends Biochem Sci. 1995; 20:465–470. [PubMed: 8578590]
7. Xie X, et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature. 2005; 434:338–345. [PubMed: 15735639]
8. Goodarzi H, et al. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. Nature. 2012; 485:264–268. [PubMed: 22495308]
9. Hafner M, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell. 2010; 141:129–141. [PubMed: 20371350]
10. Thomas M, Lieberman J, Lal A. Desperately seeking microRNA targets. Nat Struct Mol Biol. 2010; 17:1169–1174. [PubMed: 20924405]
11. Patwardhan RP, et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat Biotechnol. 2009; 27:1173–1175. [PubMed: 19915551]
12. Patwardhan RP, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012; 30:265–270. [PubMed: 22371081]
13. Melnikov A, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012; 30:271–277. [PubMed: 22371084]

14. Numahata K, et al. Analysis of the mechanism regulating the stability of rat macrophage inflammatory protein-2 mRNA in RBL-2H3 cells. J Cell Biochem. 2003; 90:976–986. [PubMed: 14624457]
15. Bazzini AA, Lee MT, Giraldez AJ. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science. 2012; 336:233–237. [PubMed: 22422859]
16. Djuranovic S, Nahvi A, Green R. miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. Science. 2012; 336:237–240. [PubMed: 22499947]
17. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature. 2010; 466:835–840. [PubMed: 20703300]
18. Wang S, et al. Nitric oxide activation of Erk1/2 regulates the stability and translation of mRNA transcripts containing CU-rich elements. Nucleic Acids Res. 2006; 34:3044–3056. [PubMed: 16757573]
19. Galgano A, et al. Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the miRNA regulatory system. PLoS One. 2008; 3:e3164. [PubMed: 18776931]
20. Morris AR, Mukherjee N, Keene JD. Ribonomic analysis of human Pum1 reveals cis-trans conservation across species despite evolution of diverse mRNA target sets. Mol Cell Biol. 2008; 28:4093–4103. [PubMed: 18411299]
21. Ray D, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013; 499:172–177. [PubMed: 23846655]
22. Pedersen JS, et al. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol. 2006; 2:e33. [PubMed: 16628248]
23. Garcia DM, et al. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol. 2011; 18:1139–1146. [PubMed: 21909094]
24. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science. 2008; 320:1643–1647. [PubMed: 18566288]
25. Mayr C, Bartel DP. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell. 2009; 138:673–684. [PubMed: 19703394]
26. Spies N, Burge CB, Bartel DP. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. Genome Res. 2013; 23:2078–2090. [PubMed: 24072873]
27. Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. Genes Dev. 2013; 27:2380–2396. [PubMed: 24145798]
28. Schork AJ, et al. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet. 2013; 9:e1003449. [PubMed: 23637621]

Online methods references

29. Amendola M, Venneri MA, Biffi A, Vigna E, Naldini L. Coordinate dual-gene transgenesis by lentiviral vectors carrying synthetic bidirectional promoters. Nat Biotechnol. 2005; 23:108–116. [PubMed: 15619618]
30. Gossen M, Bujard H. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. Proc Natl Acad Sci U S A. 1992; 89:5547–5551. [PubMed: 1319065]
31. Chen CY, et al. Versatile applications of transcriptional pulsing to study mRNA turnover in mammalian cells. RNA. 2007; 13:1775–1786. [PubMed: 17728382]
32. Yuan JS, Reed A, Chen F, Stewart CN Jr. Statistical analysis of real-time PCR data. BMC Bioinformatics. 2006; 7:85. [PubMed: 16504059]
33. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9:357–359. [PubMed: 22388286]

34. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]

35. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000; 16:276–277. [PubMed: 10827456]

36. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics. 2011; 27:1653–1659. [PubMed: 21543442]

37. Leppek K, et al. Roquin promotes constitutive mRNA decay via a conserved class of stem-loop recognition motifs. Cell. 2013; 153:869–881. [PubMed: 23663784]

38. Holcik M, Liebhaber SA. Four highly stable eukaryotic mRNAs assemble 3' untranslated region RNA-protein complexes sharing cis and trans components. Proc Natl Acad Sci U S A. 1997; 94:2410–2414. [PubMed: 9122208]

39. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011; 27:1017–1018. [PubMed: 21330290]

40. Yang E, et al. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. Genome Res. 2003; 13:1863–1872. [PubMed: 12902380]

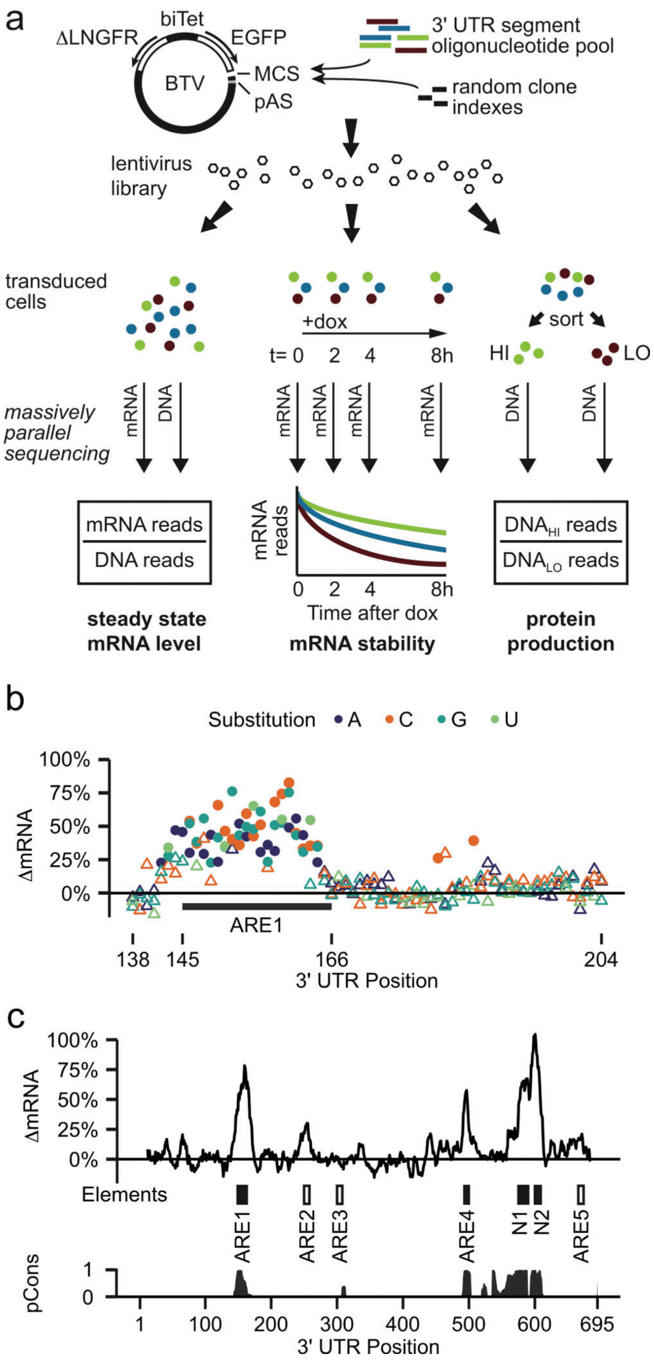

Figure 1. The fast-UTR system

(a) The fast-UTR method uses a bidirectional tetracycline-regulated viral (BTV) reporter plasmid that includes an enhanced green fluorescent protein (EGFP) reporter transgene with a multiple cloning site (MCS) for insertion of 3' UTR test sequences and a polyadenylation signal (pAS). A bidirectional tetracycline regulated promoter (biTet) drives expression of EGFP and a reference protein (truncated low-affinity nerve growth factor receptor, ΔLNGFR). Pools of 200-mer oligonucleotides containing 3' UTR segments were synthesized, amplified by PCR and inserted into BTV along with random octamer indexes used to identify each clone. BTV lentiviral libraries were used to transduce cells and massively parallel sequencing was used to quantify 3' UTR segment sequences in genomic DNA and mRNA isolated from cells and to identify mutations. Steady state mRNA levels were determined from clone read counts for mRNA samples, normalized according to DNA read counts. mRNA stability was estimated from mRNA read counts determined before and after addition of doxycycline (dox) to inhibit transcription. Blue represents segments that have minimal effects on mRNA and protein, green represents segments that increase steady state mRNA and protein levels by stabilizing mRNA and dark red represents segments that reduce steady state mRNA and protein levels by destabilizing mRNA. Flow cytometric sorting was used to enrich for 3' UTR segments that increased (HI) or decreased (LO) reporter protein production. **(b)** Fast-UTR was used to study effects of all 201 possible SNPs within a 67-nt region of the *CXCL2* 3' UTR containing an active ARE (ARE1: UAUUUAUUUAUUUAUUUAUUUAU). Values are median differences in steady state mRNA levels (ΔmRNA) between mutant clones (minimum 72 mutant clones for each SNP) and wild type clones; positive values indicate mutations that increase mRNA level. Filled circles represent SNPs with significant effects on mRNA (nominal $p < 5 \times 10^{-5}$ and Bonferroni adjusted $p < 0.01$ by Wilcoxon rank sum test); other SNPs are represented by open triangles. **(c)** Functional effects of spontaneous mutations across the complete *CXCL2* 3' UTR. Values represent median effects of mutations within 8 nucleotide intervals (minimum of 216 mutant clones in each interval). Two AREs (ARE1 and ARE4) and 2 novel elements (N1 and N2) were highly active (ΔmRNA > 50%, each with nominal $p < 10^{-71}$ by Wilcoxon rank sum test, represented by solid boxes). Three less active predicted AREs are shown as open boxes. pCons represents phastCons conservation-based estimates of the probability that each nucleotide belongs to a conserved element based on an alignment of 46 vertebrate species.

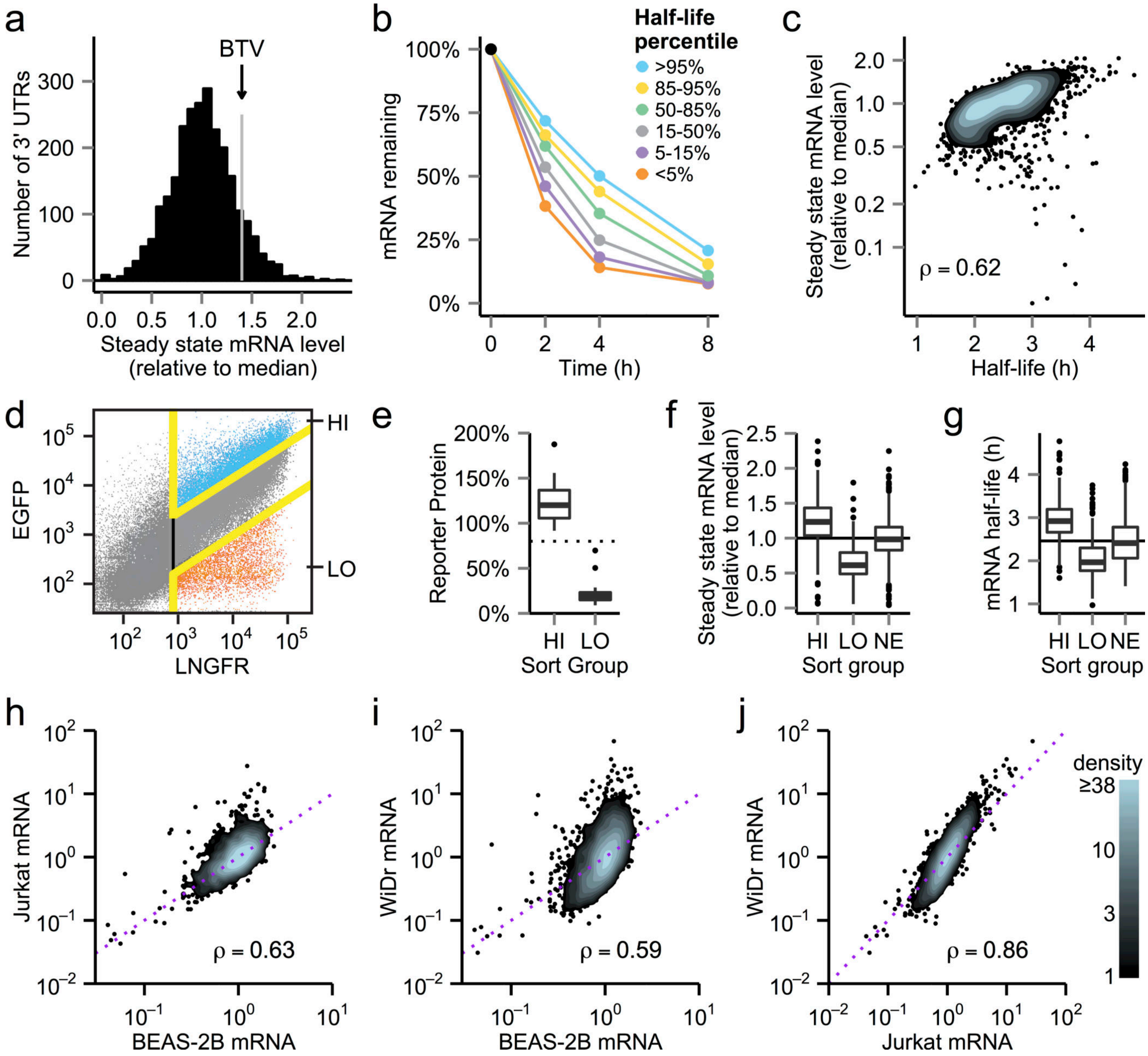

Figure 2. Effects of conserved 3' UTR sequences on steady state mRNA levels, mRNA stability and protein production

(a) 3' UTR effects on steady state mRNA levels in BEAS-2B cells. mRNA abundance for each 3' UTR segment was normalized to genomic DNA and is plotted relative to the median level of all the segments. **(b)** 3' UTR effects on mRNA decay. Half-lives were calculated for each 3' UTR segment and each segment was assigned to one of six clusters based on half-life. For example, >95% indicates the 5% of segments with the longest half-lives and <5% indicates the 5% of segments with the shortest half-lives. Points represent median percentage of mRNA remaining after doxycycline addition for all segments in a given cluster. **(c)** 3' UTR effects on mRNA stability versus steady state mRNA levels. ρ, Spearman correlation coefficient. **(d)** Flow cytometric sorting of BEAS-2B cells transduced with the conserved 3' UTR segment library. Cells were sorted based on levels of the EGFP reporter relative to the LNGFR reference. Cells with the highest EGFP/ΔLNGFR ratios (top 15% of transduced cells) were sorted into the “HI” gate and cells with the lowest ratios (bottom 15%) were sorted into the “LO” gate. **(e)** Boxplot of functional activity of 3' UTR test segments enriched in sorted cells (17 segments from HI and 21 segments from LO). The whiskers extend to the highest and lowest value within 1.5× the interquartile range. Points outside the whiskers represent outliers. The dotted line indicates mean reporter protein level (calculated from EGFP/ΔLNGFR ratios) in unsorted cells. **(f,g)** Steady state mRNA levels (f) and half-lives (g) for 3' UTR segments enriched in cells with high or low reporter protein levels and for segments that were not enriched (NE) by sorting. Horizontal lines indicate median values for all segments. **(h-j)** Conserved 3' UTR segment effects on steady state mRNA levels in BEAS-2B, Jurkat and WiDr cells. Diagonal dotted lines are lines of identity (i.e. identical levels of mRNA in the two cell types).

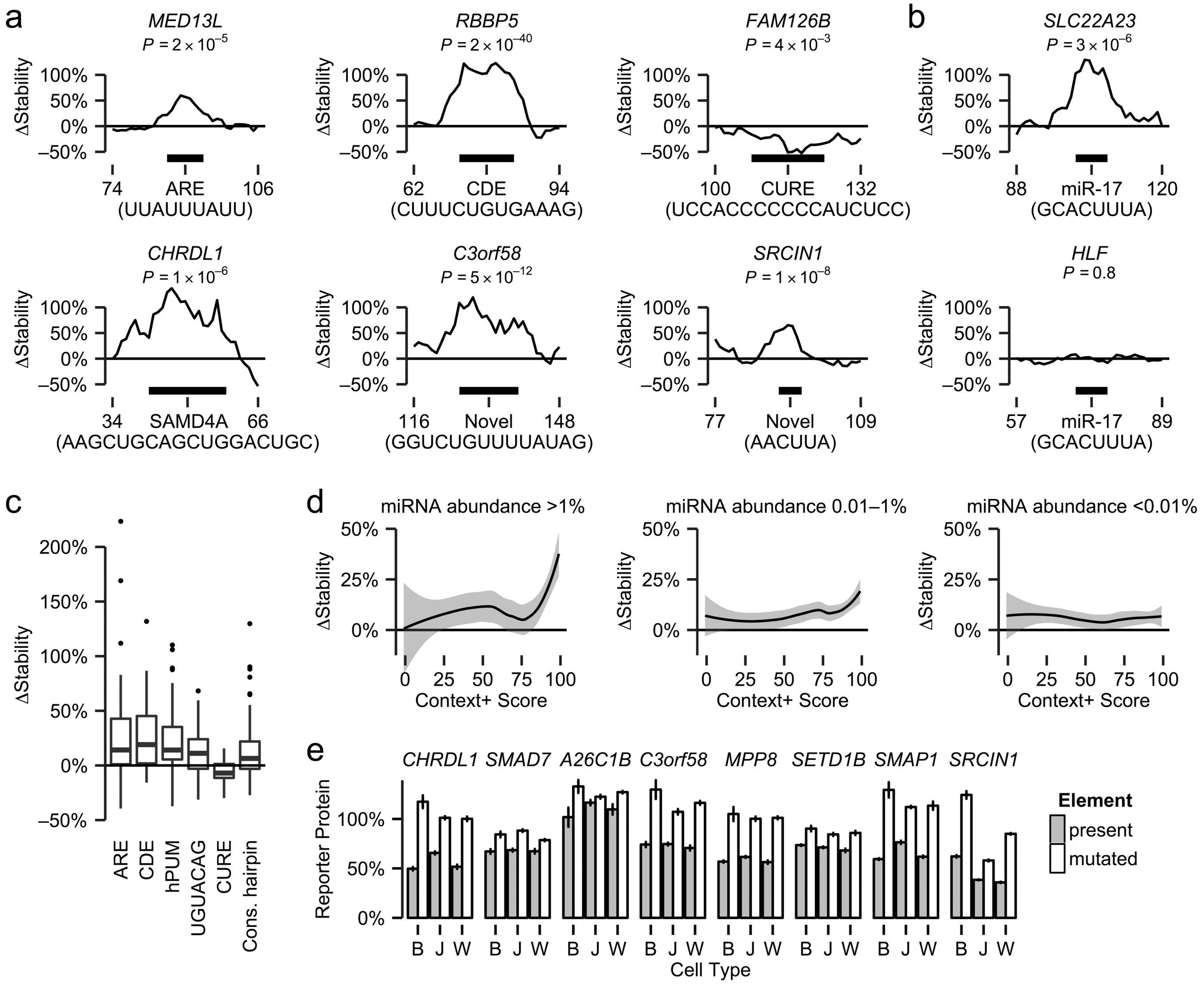

Figure 3. Functional elements identified using fast-UTR

(a) Examples of known and novel elements. ΔStability values represent median differences in the fraction of mRNA remaining after 4 h of doxycycline treatment between mutant and wild type clones for 8-mer sliding windows. Black bars below the graphs show positions of known motifs or novel elements identified by fast-UTR. AREs, constitutive decay element (CDEs), and the sterile alpha motif domain containing 4A (SAMD4A) motif are associated with destabilizing RNA binding proteins and CU-rich elements (CUREs) are associated with stabilizing RNA binding proteins. Novel elements did not contain known RNA binding protein motifs or predicted miRNA targets. *P* values were calculated by comparing clones with mutations in motifs to clones without motif mutations using the Wilcoxon rank sum test. Horizontal axis values represent nucleotide positions within each 160 nt segment. Each active element shown here and all other elements identified by fast-UTR are listed in Supplementary Data 6, which includes the numbers of mutant and wild type clones used to produce statistics for each active element. **(b)** Two predicted targets for the miR-17 family have identical seed sequences and similar TargetScan context+ scores (97 for *SLC22A23*, 95 for *HLF*) but only one has detectable activity by fast-UTR. Statistics for the inactive predicted miR-17 family target site in *HLF* were computed using data from 48 clones with mutations and 432 wild type clones. **(c)** Systematic analysis of mutations within known RNA binding protein recognition motifs. Boxplots represent 77 AREs, 13 CDEs, 72 canonical human Pumilio motifs (hPUM), 120 UGUACAG motifs (identical to a previously identified *Drosophila* Pumilio motif), 20 CUREs and 67 conserved short hairpin structures identified by EvoFold (Cons. hairpin). Mutations in each type of motif had significant effects on mRNA stability ($p < 10^{-7}$ for AREs, hPUM, and UGUACAG motifs; $p < 0.002$ for conserved short hairpins; and $p < 0.05$ for CDEs and CUREs by Wilcoxon signed rank test). **(d)** Predicted miRNA target activity measured by fast-UTR depends on miRNA abundance and TargetScan context+ score. Lines represent loess estimates and shaded areas represent 95% confidence intervals. miRNA abundance was determined by RNA sequencing. **(e)** Effects of 50-mer segments containing 8 destabilizing elements were compared to control segments containing mutant elements by flow cytometry. The element in *CHRDL1* has a predicted SAMD4A motif, the element in *SMAD7* has a predicted SRSF1 motif and the other 6 elements are novel. Values represent means ± standard error of the mean, determined from triplicate flow cytometric analyses. B, BEAS-2B; J, Jurkat; W, WiDr cells. All elements significantly reduced protein production ($p < 0.05$ for control versus mutant by two-sided Student’s t-test) except for the *A26C1B* element in BEAS-2B and Jurkat cells.