Title
Systematic Identification of Regulatory Elements in Conserved 3′ UTRs of Human Transcripts

Journal
Cell Reports, 7(1)

ISSN
2639-1856

Authors
Oikonomou, Panos
Goodarzi, Hani
Tavazoie, Saeed

Publication Date
2014-04-01

DOI
10.1016/j.celrep.2014.03.001

Peer reviewed

HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2015 May 14.

Published in final edited form as:

Cell Rep. 2014 April 10; 7(1): 281–292. doi:10.1016/j.celrep.2014.03.001.

Systematic identification of regulatory elements in conserved 3'-untranslated regions of human transcripts

Panos Oikonomou[1,*], **Hani Goodarzi**[1,*], and **Saeed Tavazoie**[1,2,3,†]

[1]Joint Centers for Systems Biology, Columbia University, New York, NY 10032 USA

[2]Department of Biochemistry and Molecular Biology, Columbia University, New York, NY 10032 USA

[3]Department of Systems Biology, Columbia University, New York, NY 10032 USA

Summary

Post-transcriptional regulatory programs governing diverse aspects of RNA biology remain largely uncharacterized. Understanding the functional roles of RNA *cis*-regulatory elements is essential for decoding complex programs that underlie the dynamic regulation of transcript stability, splicing, localization, and translation. Here, we describe a combined experimental/computational technology to reveal a catalogue of functional regulatory elements embedded in 3'-untranslated regions (3'UTRs) of human transcripts. We used a bidirectional reporter system coupled with flow cytometry and high-throughput sequencing to measure the effect of short, non-coding vertebrate-conserved RNA sequences on transcript stability and translation. Information-theoretic motif analysis of the resulting sequence-to-gene-expression mapping revealed linear and structural RNA *cis*-regulatory elements that positively and negatively modulate the post-transcriptional fates of human transcripts. This combined experimental/computational strategy can be used to systematically characterize the vast landscape of post-transcriptional regulatory elements controlling physiological and pathological cellular state transitions.

Introduction

Gene expression is highly regulated in order to achieve the requisite repertoire of RNA and protein products across the vast space of possible cellular phenotypes encompassing physiological and developmental contingencies. In the past decade, there has been a concerted effort to better understand how DNA binding proteins regulate transcription at a genome-wide level (Consortium et al., 2012). However, gene expression can also be heavily influenced by the fate of mRNAs post-transcriptionally. Messenger RNAs pass through

†To whom correspondence should be addressed. st2744@columbia.edu.
*These authors contributed equally to this work

Accession Numbers

The sequencing data for the sorted subpopulations have been deposited in the Gene Expression Omnibus (GEO ID GSE55396).

several steps of regulation in the nucleus and cytoplasm that control their processing, surveillance, localization, translation and stability (Martin and Ephrussi, 2009; Moore, 2005). Such post-transcriptional processing allows the cell to fine-tune gene expression in a fast, precise and cost effective manner (Keene, 2007). The mechanisms involved, such as alternative polyadenylation or alternative splicing have been widely studied (Licatalosi and Darnell, 2010; Maniatis and Tasic, 2002). Trans-factors including microRNAs (Bartel, 2004; He and Hannon, 2004) and RNA-binding proteins (Glisovic et al., 2008) can exert significant influence on the stability and the translation of their target mRNAs. Also, many species rely on post-transcriptional regulation to maintain mRNAs in a translationally silent state in different cell-types and developmental stages (Carrington and Ambros, 2003; Farh et al., 2005). As such, perturbations to these regulatory programs can lead to disease, including neurodegenerative disorders and cancer (Cooper et al., 2009; Lukong et al., 2008).

However, while there have been extensive studies of microRNA regulation (Bartel, 2009; Carthew and Sontheimer, 2009), the vast majority of mammalian RNA-binding proteins still remain uncharacterized both in terms of molecular function and the biological processes they affect (Cook et al., 2011; Ray et al., 2013). Many known post-transcriptional mechanisms act through non-coding *cis*-elements in the 3'UTRs of genes (Chan et al., 2005; Varani, 2001). The *cis*-regulatory elements recognized by RNA binding proteins are usually short (4–8 nucleotides; Lunde et al., 2007), highly degenerate (Zhang et al., 2013), may associate with competing factors (Mukherjee et al., 2011) and require specific secondary structures for recognition and binding (Hatzis et al., 2011). These characteristics make them difficult to predict and experimentally validate. The characterization of functional elements has benefited from computational strategies that predict a large number of putative elements (Foat and Stormo, 2009; Kazan et al., 2010; Li et al., 2010; Rabani et al., 2011) along with *in vitro* methods that systematically identify binding site specificities of RBPs for short RNAs (Martin et al., 2012; Tenenbaum et al., 2000) and engineer new RNA-protein interactions (Chen et al., 2008). However, current *in vitro* and *in vivo* genome-wide experimental methods characterizing RNA regulation, including SELEX (Tuerk and Gold, 1990), RNACompete (Ray et al., 2013), PARCLIP (Hafner et al., 2010), and HITS-CLIP (Darnell, 2010) are labor intensive and require knowledge of specific RNA-binding proteins.

Recently, high-throughput reporter assays have been developed to determine the functionality of regulatory elements in yeast promoters (Sharon et al., 2012) and human enhancers (Kheradpour et al., 2013; Melnikov et al., 2012; Patwardhan et al., 2012). These studies allowed the experimental dissection of transcriptional regulatory roles for thousands of sequences in parallel. Our goal was to extend these functional surveys to the domain of posttranscriptional regulation. To enable a systematic and unbiased analysis of post-transcriptional regulation in human cells, we have developed a new experimental approach that quantifies the effect of short, conserved RNA regulatory sequences on gene expression with respect to transcript stability and translation. We started by generating a large library of conserved 3'UTR sequences which was cloned into a single chromosomal site immediately downstream (3'UTR) of a fluorescent reporter. Our study was designed to reveal the functional role of these isolated short RNA sequences that can potentially work as building blocks in their full-length 3'UTR context. Fluorescence-activated cell sorting (FACS) of the

resulting population allowed us to isolate members of the library at distinct intervals along the distribution of fluorescent reporter expression. Utilizing next-generation sequencing, we characterized the sequence composition of the library across different ranges of reporter expression. In order to catalogue the *cis*regulatory elements involved, we took advantage of a suite of computational algorithms, previously described by our group, that explore the immense space of linear and small structural RNA elements and reveal motifs that are significantly informative of genome-wide measurements of RNA behavior (Elemento et al., 2007; Goodarzi et al., 2012). Employing these computational analyses in conjunction with measurements of the post-transcriptional regulatory effect for library sequences enabled the identification of representative linear and structural RNA motifs whose predicted effects were subsequently experimentally validated. Together, this interdisciplinary framework provides a principled and unbiased approach for decoding mammalian post-transcriptional regulatory programs.

Results

Regulatory elements tend to be evolutionarily conserved between related organisms (Elemento and Tavazoie, 2005; Pritsker et al., 2004; Xie et al., 2005). Therefore, to guide our search in finding functional non-coding elements, we performed pair-wise local alignments of the annotated 3'UTR regions of human genes to those of their orthologs in three different vertebrates, namely *P. troglodytes*, *M. musculus* and *G. gallus*. The analysis produced 3500 significant alignments (expectation value <0.1) of various lengths that were conserved across all four species. MicroRNAs and RNA-binding proteins recognize short sequences, either linear or structural. Thus, in order to enable chemical synthesis of these conserved regions while maximizing the likelihood of capturing individual regulatory elements, we split the conserved regions into 34 nucleotide segments with 50% overlaps.

This approach resulted in a diverse library of 16,332 unique 34-nt subsequences (Table S1) which were synthesized using a custom Agilent 44K microarray as previously described (Ray et al., 2009). To construct the post-transcriptional reporter library, the generated pool of DNA was cloned downstream of a fluorescent reporter system stably integrated into a single chromosomal locus in the Flp-In 293 cell line, a human embryonic kidney cell line with a single Flp Recombination Target (FRT) site (Life Technologies). This was a critical requirement in order to minimize gene expression variance caused by epigenetic context of the integration site. Figure 1 presents a schematic overview of this approach and details are provided in Experimental Procedures. The resulting library, named Conserved 3'UTR (C3U) Flp-In293 Library (Fig.1), expressed both GFP and mCherry coding sequences driven by a bidirectional promoter. The mCherry sequence is additionally controlled by the 3'UTR library inserts. We obtained over 50,000 independent stably expressing clones which were pooled together, representing over 3x coverage of the original library. This human cell line library allowed us to compare, in parallel, the effects of different 3'UTR sequences on mCherry expression in an unbiased manner using the GFP signal as the endogenous control.

The expression of the reporter system was quantified using flow cytometry. Briefly, for each cell we measured the ratio of mCherry fluorescence intensity relative to GFP, termed Dual-reporter Intensity Ratio (DIR, see Experimental Procedures for details). For a clonal cell line

containing the empty bidirectional construct (i.e. without any additional 3'UTR sequences downstream of the mCherry) the correlation coefficient between mCherry and GFP signal was 0.75±0.015. As expected, the stably transfected C3U-library exhibited a broader distribution of DIR measurements compared to the clonal control cell line (5% increase in CV, $p<10^{-12}$), highlighting the presence of functional modulators of expression within the cloned sequence population (Fig. S1A). We used FACS to separate cells based on their individual DIR measurements. The cells were sorted into a total of eight bins, each consisting of ~ 10% of the total population (four bins with lower and four with higher DIRs; Fig. 2A). To avoid admitting cells in which the reporter system is not being properly expressed, only GFP positive cells were sorted. A large sample of GFP-positive cells was also collected to serve as the background population for normalization. We generated two biological replicates of the C3U-library and for each bin and each replicate we sorted twice for four total replicates of each subpopulation. Upon sorting the subpopulations, we grew the cells and extracted genomic DNA for parallel quantification by high-throughput sequencing.

Identification of functional 3'UTR sequences

We observed that the sorted subpopulations exhibited a stable trend towards their selected DIR for multiple generations (at least 20 doublings; Fig. 2B). Thus, DIR measurement is a heritable trait that can be effectively selected and enriched for (Fig. S1B). To ensure that the cloned 3'UTR sequences—rather than genetic or epigenetic modulations in the background — were causing the observed shift in the DIRs, we extracted genomic DNA from the subpopulations, amplified the 3'UTR inserts downstream of mCherry and re-cloned them into fresh Flp-In 293 cells using the same transfection procedure as before. The resulting cell lines preserved the DIR trends observed for the original subpopulations indicating that the 3'UTR inserts indeed were modulating gene expression (Fig. 2C).

We utilized high-throughput sequencing to estimate the frequency of each 3' UTR tag in the background and sorted subpopulations. In short, we used the universal adapters to minimally amplify the cloned 3'UTRs while adding the appropriate sequencing adapters. We then subjected the samples to next-generation sequencing and achieved on average 2 million reads per sample (technical replicates were consolidated prior to analysis). From the resulting sequences we retained those containing both terminal adapters, upon removal of which the sequences were aligned to the original synthetic library. Over- and under-representation patterns were estimated by normalizing the frequencies of each library sequence in the sorted populations to those of the background population for each replicate. We defined a measure of over- or under- representation in each consecutive subpopulation, which reflects whether a specific 3'UTR element acts as an activator or repressor. For example, if a 3'UTR element were over-represented in populations with low DIR and under-represented in populations with high DIR, its effect on gene expression was predicted to be repressive. Alternatively, 3'UTR elements that were under-represented in populations with low DIR and over-represented in populations with high DIR were predicted to be activators. For each sequence we required the difference in average frequency in the high and low DIR bins to be more than 2-fold and retained the most significant over- and under-representation patterns (q-value<0.05; see Experimental Procedures). Applying these analytical criteria, we have identified 1038 putative repressors and 1012 putative activators (see Table S2 for the

complete list). Therefore, we have characterized the functionality of a large number of RNA elements (>2000) out of those present in our library utilizing a single human embryonic kidney cell line. Some 3'UTR sequences that appear not to have a significant functional role may participate in the regulatory program in different cell-types or external conditions. However, by focusing on conserved sequences, we ensured a higher probability that these sequences are broadly functional and not limited to the cell type they were tested in.

Experimental validation of the identified functional sequences

To independently validate our approach, we individually examined a subset of 3'UTR sequences that were predicted to be either suppressors or activators of gene expression based on our measurements. We selected three inserts that were predicted to be suppressors and two that were predicted to be activators representing a wide range of q-values (Fig. 3). We cloned each of the five sequences back into our reporter construct along with shuffled sequences as controls for each case. For clonal cell lines stably expressing each of these constructs, we measured their DIR relative to the shuffled control (Fig. 3A). In addition, in order to establish whether the observed differences in expression were due to modulation at the transcript level, as opposed to translation, we performed quantitative PCR for mCherry transcript relative to GFP in each cell line. Of the five constructs tested, two suppressors and two activators showed significant expression modulations (15% to 47% shifts in DIRs with *p*-values ranging from 0.002 to $<10^{-16}$, see Fig.3A and Fig. S2A–C). The remaining sequence also showed a trend towards the predicted DIR (6% shift, p=0.08, Fig. S2A–C). In addition, a subset of these elements was validated in an independent reporter assay using a dual luciferase construct (Fig. S2D). Hence, our approach correctly captures the regulatory consequences of individual 3'UTR sequences. Moreover, our findings highlight the magnitude at which these short sequences can modulate gene expression in either direction.

For example, C3U-R120, a 34-nt sequence in our library, was over-represented up to 5-fold in the low DIR bins and under-represented over 2-fold in all high DIR bins (q-value=4×10^{-6}, Fig.3). This sequence was derived from the 3'UTR of *RAB12*, a GTPase, and is complementary to the 8-nt seed region of miR-20/miR-17. *RAB12* has a predicted target site for miR-20/miR-17 at this specific locus (John et al., 2004). DIR measurement for a reporter construct of C3U-R120 exhibited a 25% decrease compared to its shuffled control (*p*-value=3×10^{-5}). Qualitative PCR showed a similar decrease of transcript abundance (20% decrease, *p*-value=0.04).

As another example, C3U-R840, a 34-nt sequence in our library, was over-represented 2- to 16-fold in the low DIR bins and under-represented up to 5-fold in the high DIR ones (q-value= 0.02, Fig.3). This sequence was derived from the 3'UTR of *LPPR5*, a lipid phosphate phosphatase gene, and contains UAUUUAU[AU]—a known AU-rich element—which reportedly destabilizes transcripts in many cell lines including 293 cells (Pham et al., 2008). Consistently, quantitative RT-PCR results showed a significant decrease of transcript abundance in the presence of this sequence (50% decrease, *p*-value=0.003; Fig.3A). The UAUUUAU[AU] element is known to be bound by KH-type regulatory protein (KHSRP; Gherzi et al., 2004) which is expressed in 293 cells (GSM630953). Consistently, KHSRP knock-down in mouse C2C12 cells (GSE38907) was shown to result in a significant up-

regulation of the transcripts that carry UAUUUAU[AU] instances in their 3' UTR (*p*-value=2×10^{-4}; Fig. S3A).

On the other hand, sequence C3U-A626 was over-represented 2- to 5-fold in high DIR populations and under-represented up 2- to 5-fold in low DIR ones (q-value= 7×10^{-3}; Fig. 3A). This sequence, derived from the 3'UTR of transcription factor TCF21, increased DIR levels of the reporter construct by 27% (*p*-value=10^{-10}) and showed significant increase of transcript abundance (87%; *p*-value=4×10^{-4}; Fig.3A) as measured by quantitative RT-PCR. This sequence contains several potential binding sites for known RBPs (DAZPA1, G3BP2, RMB46; Ray et al., 2013), however the function of these RNA binding proteins is not well characterized. These validation surveys demonstrate that our framework successfully captures known and novel RNA *cis*-regulatory elements with correctly predicted effects on gene expression.

Post-transcriptional Regulatory Element Discovery

The identification of individual *cis*-regulatory elements and their effect on mRNA stability and translation provides a powerful resource for annotating RNA regulatory regions. However, we posited that the size and diversity of our post-transcriptional reporter library may enable the *de novo* discovery of compact motif representations that are targeted by regulatory *trans*-factors. In order to accomplish this, we first assigned a score to every member of the library by considering the normalized frequency for that sequence and for each sorted population. The sum of all the contributions was then used to group the sequences into 8 co-represented clusters (see Experimental Procedures and Fig. S3B), ranging from sequences that are predominantly over-represented in populations with low DIRs (cluster 0) to those predominantly over-represented in populations with high DIRs (cluster 7) and under-represented elsewhere (Fig. S3B).

We observed that many known targets of RBPs are informative of our DIR measurements (Fig. 4A). For example, both the motif UAUUUAU[AU] which is bound by KHSRP and the element UGUA[ACU]AUA which is targeted by PUM1/PUM2 are known to promote mRNA degradation. Consistently, transcripts containing copies of either motif are over-represented in the populations with low, and under-represented in populations with high DIRs (Fig. 4A). In view of the evidence that the seed regions of miRNAs are primarily responsible for targeting transcripts, we tested whether these 8-nt seed regions of known miRNAs were informative of DIR measurements in the eight co-represented clusters. Indeed, we observed that many known miRNA recognition sites showed significant expected under- and over-representations across the DIR profiles (Fig. 4B). Moreover, sequences that match at least one miRNA seed were enriched in the top third ranked of repressors (55%), while they were depleted in the top third ranked activators (44%, *p*-value=0.005). The enrichment and depletion is even more significant if we consider the sequences that match more than one miRNA seed (26% for repressors and 13% for activators; *p*-value=10^{-4}).

The set of eight clusters with coherent DIR profiles was then used for *de novo* motif discovery using computational methods previously developed by our group aimed at finding linear (FIRE; Elemento et al., 2007) and structural (TEISER; Goodarzi et al., 2012) RNA

elements. The application of these two algorithms revealed a total of 14 linear and 8 structural RNA motifs that were significantly informative of the observed post-transcriptional effects across the library (FDR<0.1). In Fig. 5 we have shown the most significant of these elements (for the complete list see Fig. S3C and Fig. S3D). Not surprisingly, some elements were highly enriched among the suppressor sequences and others enriched among activators. The most significant elements, thus, showed a gradient of enrichment and depletion across the DIR bins, which further supports their biological relevance. Two of the linear motifs we discovered matched known miRNAs. For example, C3U-LM1, which is enriched in the repressive clusters, matches let-7, which is known to be active in HEK-293 cells (Schmitter et al., 2006). Most of these elements were also significantly informative of mRNA stability measurements in human and mouse highlighting their functionality in various hosts (Fig. S4A–D) and their functional conservation in distantly related mammals (Fig. S4E–H). Like many established RNA regulatory sequences, these discovered motifs are short and degenerate making them likely to appear with high frequencies in the human transcriptome. The additional structural constraint for the motifs shown in Fig. 5A makes them likely to appear in 1–6% of the human 3'UTRs, similarly to some of the less degenerate RBP target sites (Ray et al., 2013).

From the list of discovered informative elements we chose the most significant structural motif (C3U-SM1; *z*-score=14.7) to further elucidate its regulatory role (Fig. 6). C3U-SM1 instances in the human 3'UTRs have lower GC content (0.27) compared to that of the human transcriptome (0.43, Fig. S5A). Even though C3U-SM1 has low GC content and may resemble an ARE element, it does not function as a linear element and relies on the information provided by its secondary structure for its enrichment/depletion pattern across the DIR measurements in the C3U Library. If we consider only matches to C3U-SM1's sequence pattern that do not satisfy the secondary structure requirement, the enrichment patterns disappear and the motif ceases to be informative in our dataset (*z*-score=-0.14; Fig. S5B). In addition, we observed that natural instances of C3U-SM1 in human 3'UTRs showed lower *in silico* folding energies compared to their shuffled counterparts highlighting their propensity to form more stable local secondary structures (*p*-value=0.02; Fig. S5C).

In order to probe the functionality of C3U-SM1, we cloned two individual instances of this motif into our reporter construct. Shuffled control sequences were also synthesized and cloned in parallel (Table S3). For stably expressed cell lines containing these constructs, DIR was significantly lower for the C3U-SM1 instances relative to the shuffled controls (Fig. 6B). For a first set of instances (C3U-SM1v1, Fig. 6B upper panel and Fig. 6D) the construct showed a ~ 35% reduction in DIR compared to the controls (p-value<10^{-16}), while for the second set (C3U-SM1v2, Fig. 6B lower panel) we observed a 30% reduction in expression (p-value=2.7×10^{-12}). To ensure that the regulatory effect stems from the motif instance, we also synthesized a control construct for C3U-SM1v1 where only the motif instance was shuffled (36% reduction in expression, see Fig. S5D). Additionally, we transiently transfected FlpIn-293 cells with the C3U-SM1v1 construct and used quantitative PCR to measure the relative quantity of mCherry transcript relative to GFP (Fig. 6C). Transcript levels appeared significantly lower than the control (p-value<3×10^{-5}). C3U-SM1v1 and C3U-SM1v2 were further validated in a context independent of the bidirectional

reporter construct using a dual luciferase vector (Fig. S5E). Overall, our findings demonstrate that the presence of the C3U-SM1 element results in a significant post-transcriptional down-regulation of the host transcript.

Discussion

The strategy presented here provides a novel framework for high-throughput characterization of RNA-regulatory elements *in vivo*. We utilized this technology to uncover the functionality of short, conserved 3'UTR elements included in our library. We validated a small set of these sequences by measuring their individual effects on expression and transcript levels. As expected, the amplitude and direction of the observed regulatory effects varied across the library. While some of the sequences functioned as suppressors, others functioned as activators. Surprisingly, we found an equal number of repressing and activating elements given the repressive role of miRNAs and the association of longer 3'UTRs with low expression levels. However, we tested the functionality of isolated short RNA elements, which allowed us to reveal unexpected regulatory contributions that might have been masked otherwise. Indeed, stabilizing *trans*-factors might be common among the vast majority of RBPs that have yet to be characterized (Goodarzi et al., 2012; Yugami et al., 2007). In addition to the functional characterization of novel 3'UTR regulatory sequences, our approach was successful in recapturing the regulatory consequences of known RNA binding proteins and miRNA recognitions sites. Furthermore, we generated a catalog of post-transcriptional regulatory motifs with significant positive or negative contributions to gene expression. As expected from the conserved nature of the 3'UTR library sequences, many of the identified linear and structural motifs are consistently informative of mRNA stability measurements in mouse cells in addition to other human cell lines further highlighting their broad functionality. Finally, we investigated the role of one of the novel motifs, a structural element termed C3U-SM1, which was predicted to function as a repressor. Using C3U-SM1-carrying reporter constructs we showed that instances of this motif result in a significant and substantial down-regulation of the host transcript.

Apart from being informative of the reporter's expression levels, C3U-SM1 potentially plays a role in a variety of cellular processes and pathways (Fig. S6A–B). For example, the levels of transcripts containing the C3U-SM1 motif were positively correlated with the doubling times of breast cancer cell lines they were measured in (Fig. 7A). We should also note that cell-cycle genes showed a significant enrichment among the C3U-SM1-carrying transcripts (*p*value= 0.01; Fig. S6A). While, the enrichment and depletion patterns shown in Fig. 7 are merely correlative in nature, they suggest a potential role for this element and its underlying regulon in cell proliferation. It remains to be determined whether a putative binding factor actively participates in cancer regulation. Higher cell proliferation rate is one of the hallmarks of cancer; consistently, we observed the aggregate expression level of C3U-SM1-carrying transcripts to be lower for more advanced stage tumors (Fig. 7B). Using the aggregate expression of these transcripts (with the median as the threshold) to stratify a large cohort of breast cancer patients (Korpal et al., 2011), we observed that C3U-SM1 was also significantly informative of clinical outcome (Fig. 7C). More importantly, C3U-SM1 was also informative of gene expression patterns in human tissue-specific and cancer whole-genome datasets (Fig. S6C–D). The observation that this newly discovered motif is

informative of expression patterns or disease outcomes in independent datasets hints at a potentially broader significance and a critical role in the regulatory programs that govern gene expression.

Post-transcriptional *cis*-regulatory elements often participate in combinatorial functions to modulate gene expression (Han et al., 2005; Hogan et al., 2008). In order to reveal such regulatory interactions, we used mutual information to capture significant co-localization of these elements in the 3’ UTRs of human transcripts (Fig. S7). Remarkably, many of the linear motifs we discovered appear to function in two large groups, one consists mostly of putative repressors and the other of putative activators. The clustering of these elements suggests a coordinated and concerted underlying post-transcriptional regulatory network.

The combined experimental/computational framework presented here allows parallel, *in vivo* identification of functional post-transcriptional regulatory elements in mammalian transcripts with an effect on mRNA stability or translation. As demonstrated, high-throughput parallel approaches can annotate large sets of previously uncharacterized functional non-coding sequences and, therefore, enhance our systems-level understanding of gene regulation. It should be noted however that our method was applied in the context of a single cell-line under one growth condition. Additional cell lines and diversified internal and external conditions may yield a more complete picture of the regulatory role for such elements and capture a different subset of the elements as functional. Our study was designed to characterize the functional effect of individual short 3’UTR sequences isolated from their full-length 3’UTR context. However, it is possible to use the bidirectional assay presented here to test the effect of a library of full-length 3’UTRs and measure the functionality of individual elements in their native combinatorial context. It will be of interest to design and test a variety of 3’UTR libraries, whether it is comprehensive sets of k-mers or specific collections of RNAs with given structural or sequence specifications. It is also possible to systematically study the regulatory consequences of different parameters, such as 3’UTR length (Sandberg et al., 2008) and positional bias of a given element. Furthermore, our method can be used to validate, in parallel, the functionality of large sets of *in-vitro* predicted recognition specificities for RBPs and micro-RNAs. Beyond decoding post-transcriptional regulation, these global surveys produce catalogues of functional elements for precise expression of transgenic elements in a given cell type. As such they will be a critical resource for synthetic biology applications.

Experimental Procedures

Reporter construct for testing the functionality of 3’UTR sequences

The vector pBI-CMV2 (Clontech) was used to create a bidirectional construct with GFP and mCherry driven by the same CMV promoter along with a Gateway cloning site downstream of mCherry (in its 3′UTR). Both the GFP and the mCherry sequences were preceded by the Kozak consensus sequence for optimal expression. A Gateway cloning site was inserted downstream of the mCherry in its 3’UTR prior to the polyadenylation signal. The bidirectional construct was then cloned into the FRT-based pcDNA5/FRT/TOPO expression vector (Life Technologies). Universal adapters were used to amplify each 3’UTR sequence under investigation: 5’-CTAATACGACTCACTATTAG and 5’-

CTGTATCCGCTCGCTCTTCA. Appropriate Gateway sequences (Life Technologies) were added during PCR amplification. A two-step recombination reaction (first a BP recombination into pDONR221, followed by an LR reaction) was used to transfer the sequence into the Gateway cloning site of the reporter construct. The pcDNA5/FRT bidirectional reporter construct was cloned into Flp-In 293 (Life Technologies), a human cell line related to the Human Embryonic Kidney 293 cells which contains a single integrated Flp Recombination Target (FRT) site at a transcriptionally active genomic locus. Co-transfection of the bidirectional construct with pOG44 plasmid, which expresses the Flp-In recombinase, results in cleavage of the FRT site in the genome and integration of the reporter gene in the same locus in every cell (Life Technologies). Transfections were performed using Lipofectamine 2000 according to the manufacturer's recommendations (Life Technologies). Stable cell lines expressing the reporter system were selected for 10–12 days under 50μg/ml of Hygromycin.

Library construction

The library of 16,332 3'UTR elements was synthesized on a custom Agilent 44K microarray designed for this purpose (Array Design ID 030938) as described previously (Ray et al., 2009). Each array contained duplicate sets of all library sequences. A universal adapter sequence was included upstream of each sequence on the array which allowed for primer extension. As shown in Figure 1, the adapter oligonucleotide was annealed to the microarray probes, and primer extension was performed to make the probes double stranded. A common DNA linker was ligated to all the probes on the microarray, after which single-stranded DNAs containing the universal adapters were stripped from the array. The entire library was PCR amplified using adapter sequences which led to the addition of flanking Gateway sequences for subsequent cloning. The Gateway site was then used to clone the library into the plasmid via recombination. For each Gateway recombination step we performed bacterial transformations in large enough volumes and efficiencies to ensure at least 30× coverage of the library of elements present on the array. After transfecting the construct in Flp-In293 cells we obtained 50,000 independent clones, approximately 3× coverage of the original library sequences. All the clones were pooled together into a single population. This process was repeated twice to produce two independently generated libraries (C3U Library A and B).

Fluorescence activated cell sorting and analysis

Cells were harvested at 90–95% confluency for sorting and analysis on a BD FACSaria sorter. The sorted cells were gated so that they were always GFP positive. The distribution of mCherry to GFP ratios, DIR, was calculated. For sorting the C3U library into subpopulations, we gated the population into bins each containing 10% of the total number of cells. We collected cells for the top four high DIR bins (H10, H20, H30, H40) and the bottom four low DIR bins (L10, L20, L30, L40). 500,000 cells were collected for each bin to ensure sufficient representation of every 3'UTR sequence in the population in two replicates each. Background populations for the GFP positive cells were also collected. The process was repeated for both C3U libraries A and B. Subpopulations were grown in media with 50 μg/ml Hygromycin for at least 2 passes. For each subpopulation we extracted at least 5μg of genomic DNA, which contains approximately 1.5 million copies of our reporter system from

different genomes, such that each member of the original 3'UTR library was represented at least 100 times and there is sufficient dynamic range for quantification.

Sequencing data analysis and clustering

We PCR amplified the cloned 3'UTR sequences from each subpopulation's total genomic DNA and multiplexed using standard Illumina adapters (TruSeq adapters 2, 4, 5, 6, 7, 9, 12). Additionally we included 6 custom designed adapters (ATCACG, TTAGGCG, ACTTGACG, GATCAGTAG, TAGCTTACAG, GGCTACGAGTG) at the 5' end of the amplified sequences from each subpopulation, which were used to further multiplex but also to stagger the library sequences to achieve the necessary variability in the initial bases (since all our library sequences included an identical universal adapter at the 5' end). Samples were run on two lanes and sequencing data were filtered for sequences containing both terminal universal adapters. The sequence was then matched to the library probes allowing for sequence shifts. Out of 16,332 sequences in the designed library, we found 9302 sequences in C3U Library A and 8496 sequences in C3U Library B, with a combined representation of 14141 sequences present. Prior to further analysis, we merged data from technical replicates and from adjacent bins to increase read-counts in each group (L10 & L20; L30 & L40; H10 & H20; H30 & H40). This resulted in 4 merged populations for each C3U library. A frequency was calculated for every probe in every merged population and it was normalized to the frequency in the corresponding background population. We considered only sequences with over 20 counts in a given sample and filtered out sequences that appeared only in the background samples. For further analysis, we calculated the *z*-score for the log-normalized frequency of each sequence within each subpopulation. The corresponding *p*-values for each *z*-score were combined using Fisher's method and considering whether it is over-represented in all low DIR bins and under-represented in all high DIR bins or the opposite. Benjamini-Hochberg-corrected q-values were then calculated for each sequence. A sequence was predicted to be repressive if q-value<0.05 and the average frequency in the low DIR bins was at least two times higher than the average frequency for high DIR bins. The opposite was required for sequences predicted to be activators. In order to cluster the sequences into groups with coherent representation in DIR subpopulations, we calculated a score for each library member. A positive score for a given sequence signifies enrichment in high DIR populations or depletion in low DIR populations. A negative score indicates enrichment in low DIR or depletion in high DIR populations. We constructed the score by considering the over- or under- representation for each sequence in every merged population compared to the background. For every sequence, high DIR populations contributed to the score positively if the sequence was over-represented and negatively if the sequence was under-represented (three points for more than 4-fold and one point for more than 1.4-fold over- or under-representation). Conversely, a population with low DIR contributed to the score negatively if the sequence was over-represented and positively if the sequence was under-represented. For every sequence we calculated the sum of all contributions as the final score. This score was then used to sort the sequences into co-expressed clusters (Fig. S3B).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank members of the Tavazoie lab for stimulating discussions and we are also thankful to Sohail Tavazoie and Sasan Amini for useful comments on the project. We gratefully acknowledge the help of Kristie Gordon, Sandra Tetteh (Columbia University HICCC) and Tina DeCoste (Princeton University) for their kind help with FACS experiments. The work was supported by a NHGRI Award to Saeed Tavazoie (2R01HG003219).

References

Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004; 116:281–297. [PubMed: 14744438]

Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009; 136:215–233. [PubMed: 19167326]

Carrington JC, Ambros V. Role of MicroRNAs in Plant and Animal Development. Science. 2003; 301:336–338. [PubMed: 12869753]

Carthew RW, Sontheimer EJ. Origins and Mechanisms of miRNAs and siRNAs. Cell. 2009; 136:642–655. [PubMed: 19239886]

Chan CS, Elemento O, Tavazoie S. Revealing posttranscriptional regulatory elements through network-level conservation. PLoS computational biology. 2005; 1:e69. [PubMed: 16355253]

Chen Y, Mandic J, Varani G. Cell-free selection of RNA-binding proteins using in vitro compartmentalization. Nucleic Acids Res. 2008; 36:e128. [PubMed: 18790803]

Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. RBPDB: a database of RNA-binding specificities. Nucleic Acids Res. 2011; 39:D301–D308. [PubMed: 21036867]

Cooper TA, Wan L, Dreyfuss G. RNA and disease. Cell. 2009; 136:777–793. [PubMed: 19239895]

Darnell RB. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley interdisciplinary reviews. RNA. 2010; 1:266–286. [PubMed: 21935890]

Elemento O, Slonim N, Tavazoie S. A universal framework for regulatory element discovery across all genomes and data types. Molecular cell. 2007; 28:337–350. [PubMed: 17964271]

Elemento O, Tavazoie S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome biology. 2005; 6:R18. [PubMed: 15693947]

Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science. 2005; 310:1817–1821. [PubMed: 16308420]

Foat BC, Stormo GD. Discovering structural cis-regulatory elements by modeling the behaviors of mRNAs. Molecular systems biology. 2009; 5:268. [PubMed: 19401680]

Gherzi R, Lee KY, Briata P, Wegmuller D, Moroni C, Karin M, Chen CY. A KH domain RNA binding protein, KSRP, promotes ARE-directed mRNA turnover by recruiting the degradation machinery. Mol Cell. 2004; 14:571–583. [PubMed: 15175153]

Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and posttranscriptional gene regulation. FEBS letters. 2008; 582:1977–1986. [PubMed: 18342629]

Goodarzi H, Najafabadi HS, Oikonomou P, Greco TM, Fish L, Salavati R, Cristea IM, Tavazoie S. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. Nature. 2012; 485:264–268. [PubMed: 22495308]

Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jungkamp A-C, Munschauer M. PAR-CliP-a method to identify transcriptome-wide the binding sites of RNA binding proteins. Journal of visualized experiments: JoVE. 2010

Han K, Yeo G, An P, Burge CB, Grabowski PJ. A combinatorial code for splicing silencing: UAGG and GGGG motifs. PLoS biology. 2005; 3:e158. [PubMed: 15828859]

Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, Vidaurre T, Holmes F, Souchon E, Wang HK, et al. A Genomic Predictor of Response and Survival Following Taxane-Anthracycline Chemotherapy for Invasive Breast Cancer. Jama-J Am Med Assoc. 2011; 305:1873–1881.

He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. Nature reviews. Genetics. 2004; 5:522–531. [PubMed: 15211354]

Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS biology. 2008; 6:e255. [PubMed: 18959479]

John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. PLoS biology. 2004; 2:e363. [PubMed: 15502875]

Kazan H, Ray D, Chan ET, Hughes TR, Morris Q. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLoS computational biology. 2010; 6:e1000832. [PubMed: 20617199]

Keene JD. RNA regulons: coordination of post-transcriptional events. Nature Reviews Genetics. 2007; 8:533–543.

Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. 2013; 23:800–811. [PubMed: 23512712]

Korpal M, Ell BJ, Buffa FM, Ibrahim T, Blanco MA, Celià-Terrassa T, Mercatali L, Khan Z, Goodarzi H, Hua Y. Direct targeting of Sec23a by miR-200s influences cancer cell secretome and promotes metastatic colonization. Nature medicine. 2011; 17:1101–1108.

Li X, Quon G, Lipshitz HD, Morris Q. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA. 2010; 16:1096–1107. [PubMed: 20418358]

Licatalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. Nature Reviews Genetics. 2010; 11:75–87.

Lukong KE, Chang KW, Khandjian EW, Richard S. RNA-binding proteins in human genetic disease. Trends in genetics : TIG. 2008; 24:416–425. [PubMed: 18597886]

Lunde BM, Moore C, Varani G. RNA-binding proteins: modular design for efficient function. Nature reviews. Molecular cell biology. 2007; 8:479–490. [PubMed: 17473849]

Maniatis T, Tasic B. Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature. 2002; 418:236–243. [PubMed: 12110900]

Martin KC, Ephrussi A. mRNA localization: gene expression in the spatial dimension. Cell. 2009; 136:719–730. [PubMed: 19239891]

Martin L, Meier M, Lyons SM, Sit RV, Marzluff WF, Quake SR, Chang HY. Systematic reconstruction of RNA functional motifs with high-throughput microfluidics. Nature methods. 2012; 9:1192–1194. [PubMed: 23142872]

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, 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]

Moore MJ. From birth to death: the complex lives of eukaryotic mRNAs. Science. 2005; 309:1514–1518. [PubMed: 16141059]

Mukherjee N, Corcoran DL, Nusbaum JD, Reid DW, Georgiev S, Hafner M, Ascano M Jr, Tuschl T, Ohler U, Keene JD. Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. Mol Cell. 2011; 43:327–339. [PubMed: 21723170]

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee SI, Cooper GM, et al. Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol. 2012; 30:265–270. [PubMed: 22371081]

Pham DH, Moretti PA, Goodall GJ, Pitson SM. Attenuation of leakiness in doxycycline-inducible expression via incorporation of 3' AU-rich mRNA destabilizing elements. BioTechniques. 2008; 45:155–156. 158–160. passim. [PubMed: 18687064]

Pritsker M, Liu YC, Beer MA, Tavazoie S. Whole-genome discovery of transcription factor binding sites by network-level conservation. Genome Res. 2004; 14:99–108. [PubMed: 14672978]

Rabani M, Kertesz M, Segal E. Computational prediction of RNA structural motifs involved in post-transcriptional regulatory processes. Methods in molecular biology. 2011; 714:467–479. [PubMed: 21431758]

Ray D, Kazan H, Chan ET, Castillo LP, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nature biotechnology. 2009; 27:667–670.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013; 499:172–177. [PubMed: 23846655]

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]

Schmitter D, Filkowski J, Sewer A, Pillai RS, Oakeley EJ, Zavolan M, Svoboda P, Filipowicz W. Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells. Nucleic Acids Res. 2006; 34:4801–4815. [PubMed: 16971455]

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat Biotechnol. 2012; 30:521–530. [PubMed: 22609971]

Tenenbaum SA, Carson CC, Lager PJ, Keene JD. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. Proceedings of the National Academy of Sciences. 2000; 97:14085–14090.

Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science. 1990; 249:505–510. [PubMed: 2200121]

Varani G. Delivering messages from the 3' end. Proc Natl Acad Sci U S A. 2001; 98:4288–4289. [PubMed: 11296278]

Xie X, Lu J, Kulbokas E, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature. 2005; 434:338–345. [PubMed: 15735639]

Yugami M, Kabe Y, Yamaguchi Y, Wada T, Handa H. hnRNP-U enhances the expression of specific genes by stabilizing mRNA. FEBS Lett. 2007; 581:1–7. [PubMed: 17174306]

Zhang C, Lee K-Y, Swanson MS, Darnell RB. Prediction of clustered RNA-binding protein motif sites in the mammalian genome. Nucleic acids research. 2013

Highlights

- Functional characterization of short, conserved 3'UTR sequences
- Discovery of linear and structural RNA motifs controlling gene expression
- Validation of the repressive regulatory effect of C3U-SM1, a structural RNA motif

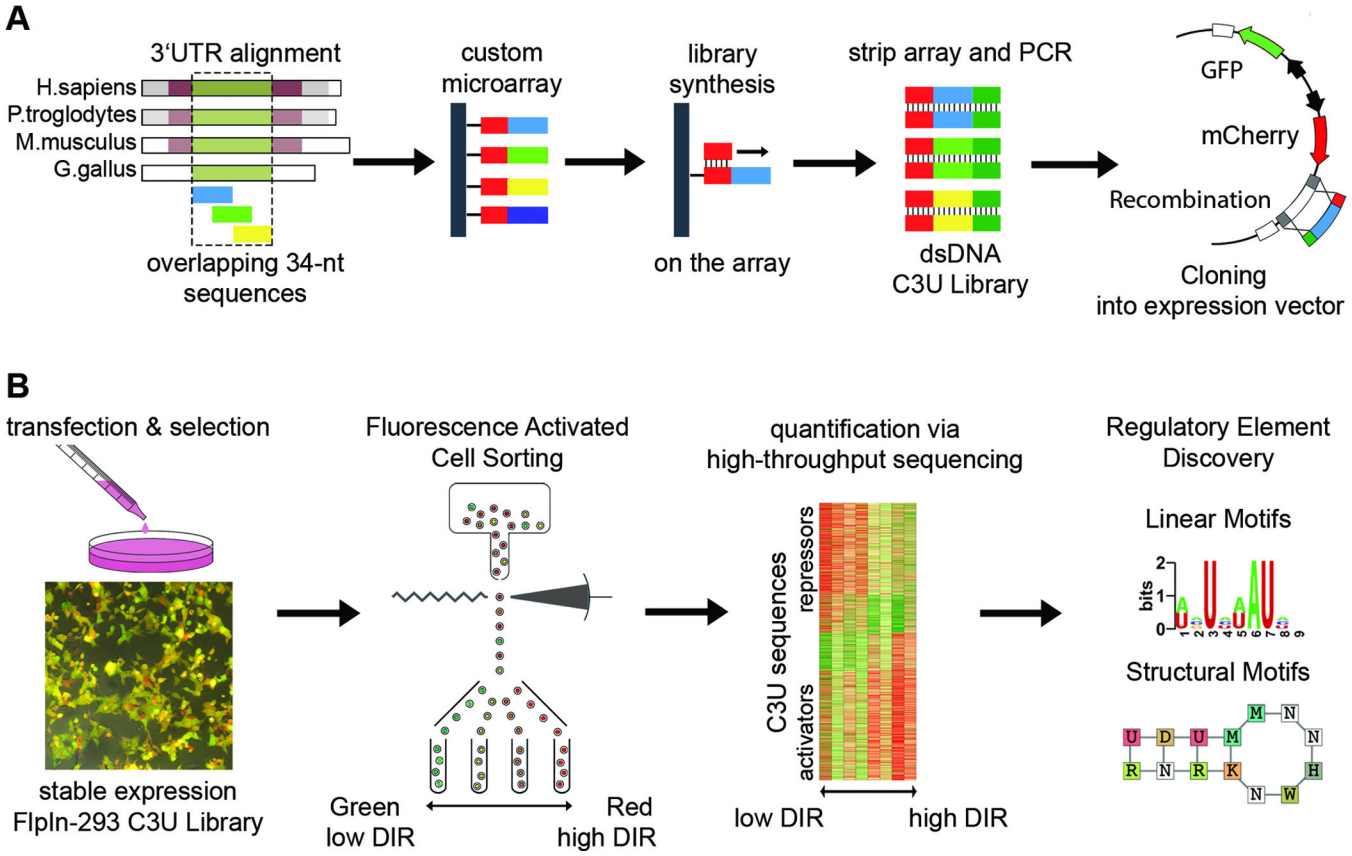

Figure 1. Identification of 3'UTR regulatory elements in human transcripts

(A) 34-nt conserved 3'UTR sequences were identified and synthesized on a custom microarray. Universal adapter primers were annealed on each probe, followed by primer extension. Single stranded DNA sequences were stripped and PCR amplified using universal adapters. Library sequences were cloned downstream of a fluorescent mCherry reporter in a bidirectional construct via recombination. (B) Transfection of the vector into human FlpIn-293 cells produced a library of FlpIn293 cells which stably expressed the bidirectional construct controlled by the 34-nt conserved 3'UTR sequences. Cells from this library were FACS-sorted into expression bins and analyzed via high-throughput sequencing. Based on over- and under- representation patterns for each sequence in each expression bin, sequences were predicted to be either gene expression repressors or activators. Results were further analyzed to identify new linear and structural RNA regulatory motifs. See also Table S1.

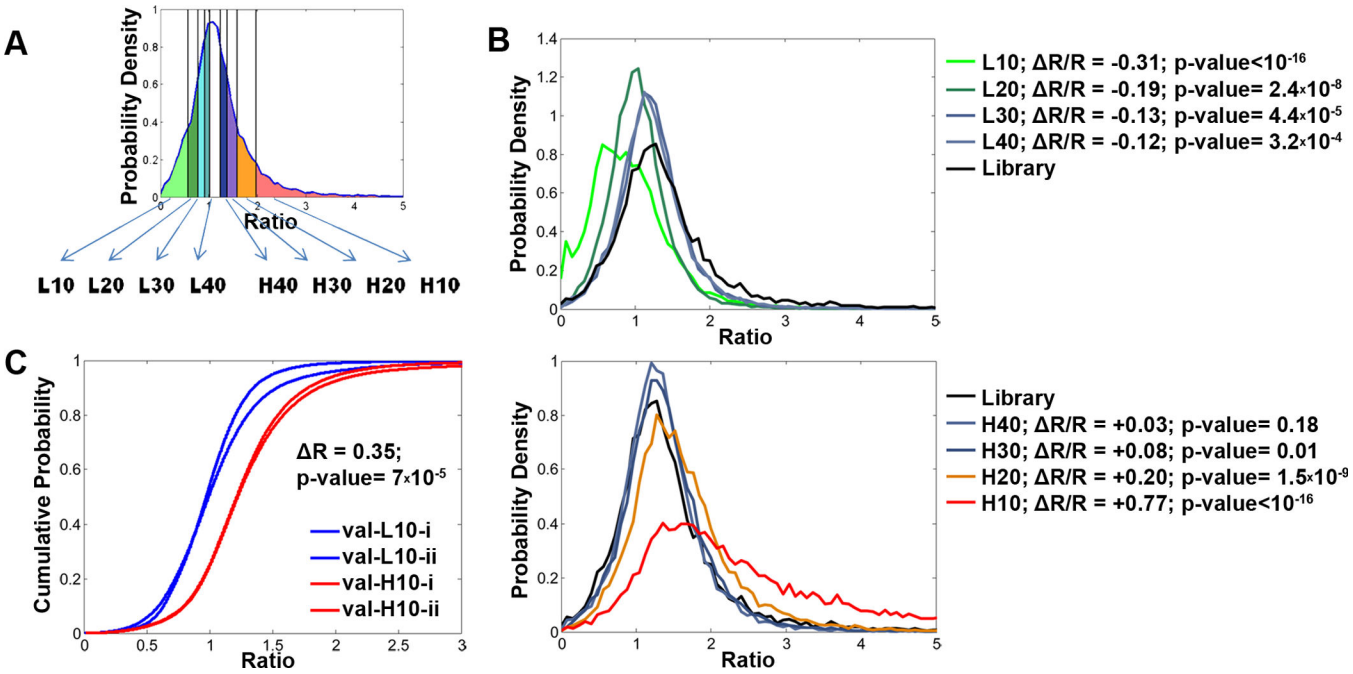

Figure 2. Sorting of Flp-In293 C3U Library into expression bins

(A) Cells were sorted into expression bins based on the Dual-reporter Intensity Ratio (DIR) of mCherry to GFP fluorescence. Each bin contains ~ 10% of the initial library. (B) Distributions of DIR for sorted populations exhibited a stable trend towards the sorted bin for the four low DIR bins (L10, L20, L30, L40) and the four high DIR bins (H10, H20, H30, H40). The differences between the median intensity ratios between each population and the library are shown. (C) Cumulative DIR distributions for validation populations with bidirectional reporters with re-cloned inserts from the initial sorted bins. Fifty independent clones were pooled together for each validation population. Shown are two replicates for each of two populations, cloned with inserts from low DIR sub-populations (val-L10) and H10 high DIR sub-populations (val-H10). Consistent with the origin of the inserts, the average DIR of val-L10 is significantly lower than the DIR of val-H10 populations ($\Delta R = 0.35$; p-value$=7\times10^{-5}$). See also Figure S1.

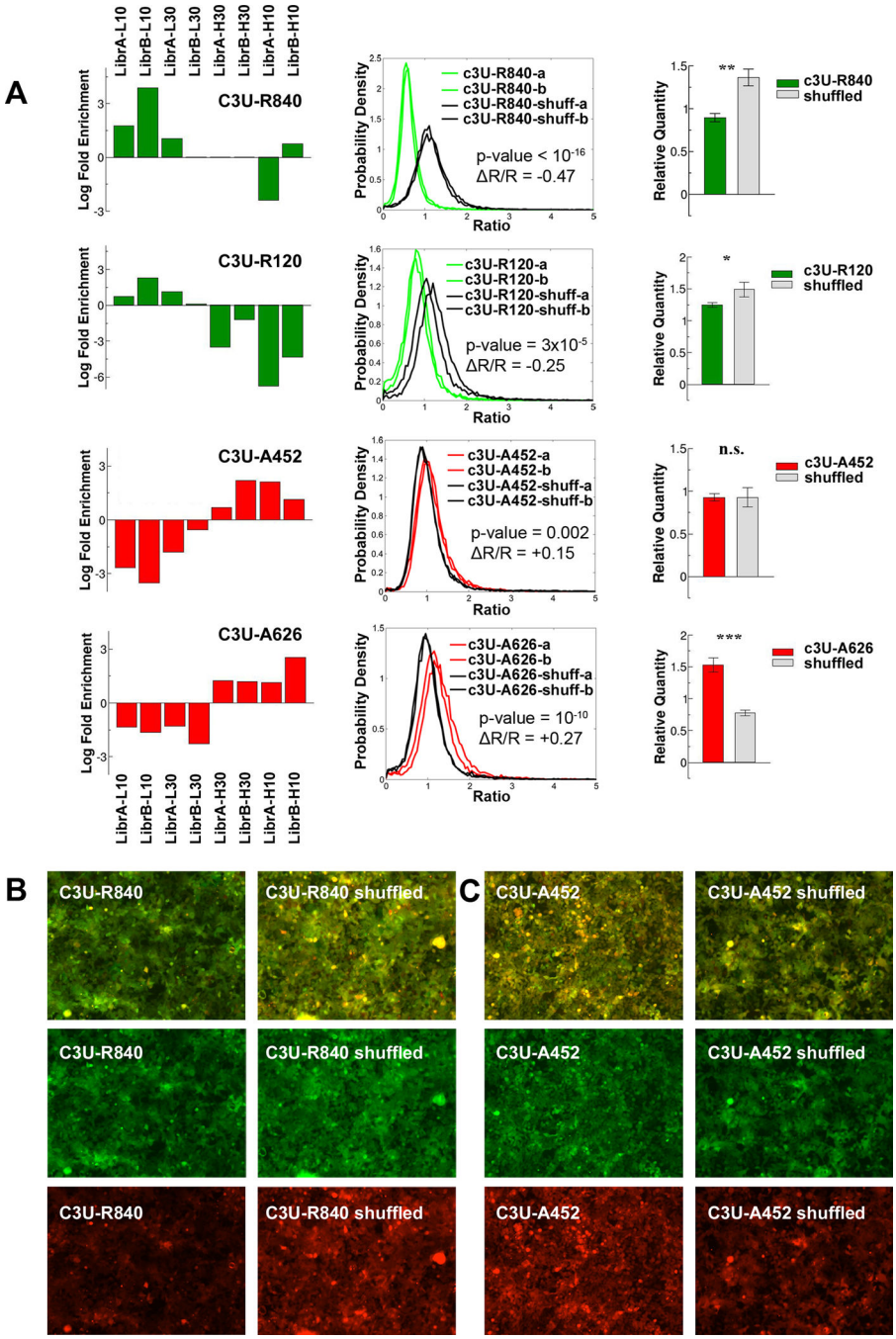

Figure 3. Validation of functional regulatory sequences as suppressors or activators of gene expression

(A) Each row represents a different 3'UTR sequence from the original library. *1st column:* Log-fold over and under representation patterns in sorted expression bins for each sequence. The frequency of each sequence in a given sub-population was calculated from the high-throughput sequencing results and normalized to its frequency in the background population. Green distributions represent putative repressors (C3U-R840, q-value= 0.02; and C3U-R120, q-value=4×10^{-6}) while red distributions represent putative activators (C3U-A452 q-

value= 1.4×10^{-3}; C3U-A626 q-value= 7×10^{-3}). *2nd column:* DIR distribution plots for clonal cell lines expressing the bidirectional reporter system with a single 3'UTR sequence (C3U-R840, C3U-R120, C3U-A452, C3U-A626). Distributions for each sequence (in color) compared to control cell lines containing shuffled versions of each sequence (black, C3U-R840-shuff, C3U-R120-shuff, C3U-A452-shuff, C3U-A626-shuff). Two replicate cell lines were produced for each sequence and its control. The difference, ΔR, in median DIR between the 3'UTR sequence under investigation and its shuffled control is reported in each panel. *3rd column:* Quantitative PCR showing up- or down-regulation in mRNA levels relative to the shuffled controls for each sequence (data are represented as mean ±SEM, significant differences are marked by stars, *, $p<0.05$; **, $p<0.01$; ***, $p<0.001$.) (B,C) Representative microscopy images for C3U-R840 (B) and C3U-A452 (C) and their shuffled controls. GFP (green) and mCherry (red) images are shown individually and overlapped. See also Figure S2 and Table S3.

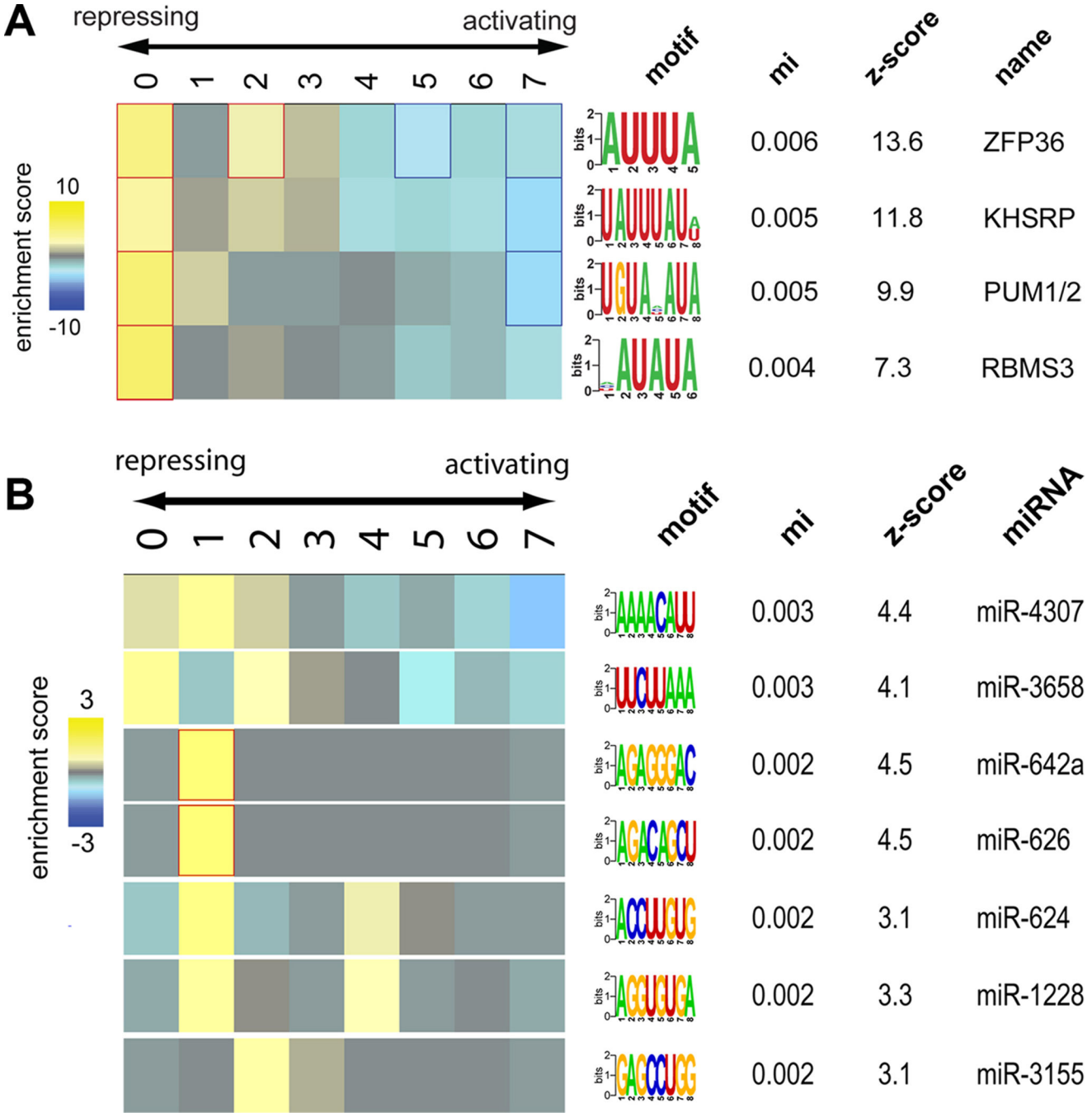

Figure 4. Known post-trascriptional regulatory target sites are informative of gene expression in the C3U library

Shown are the over and under representation patterns for binding motifs of known RBPs (A) and sequences complementary to the first 8 nucleotides of known microRNAs (B) that were informative of gene expression in the C3U library. Sequences were clustered into eight sets, from those enriched in low-expression populations (left) to those enriched in high-expression populations (right). Reported are each motif's primary sequence, the mutual information values and *z*-scores associated with a randomization-based statistical test.

Yellow entries denote enrichment while blue entries denote significant depletion of a given motif in the corresponding cluster (for details see Elemento et al., 2007; also see Figure S3).

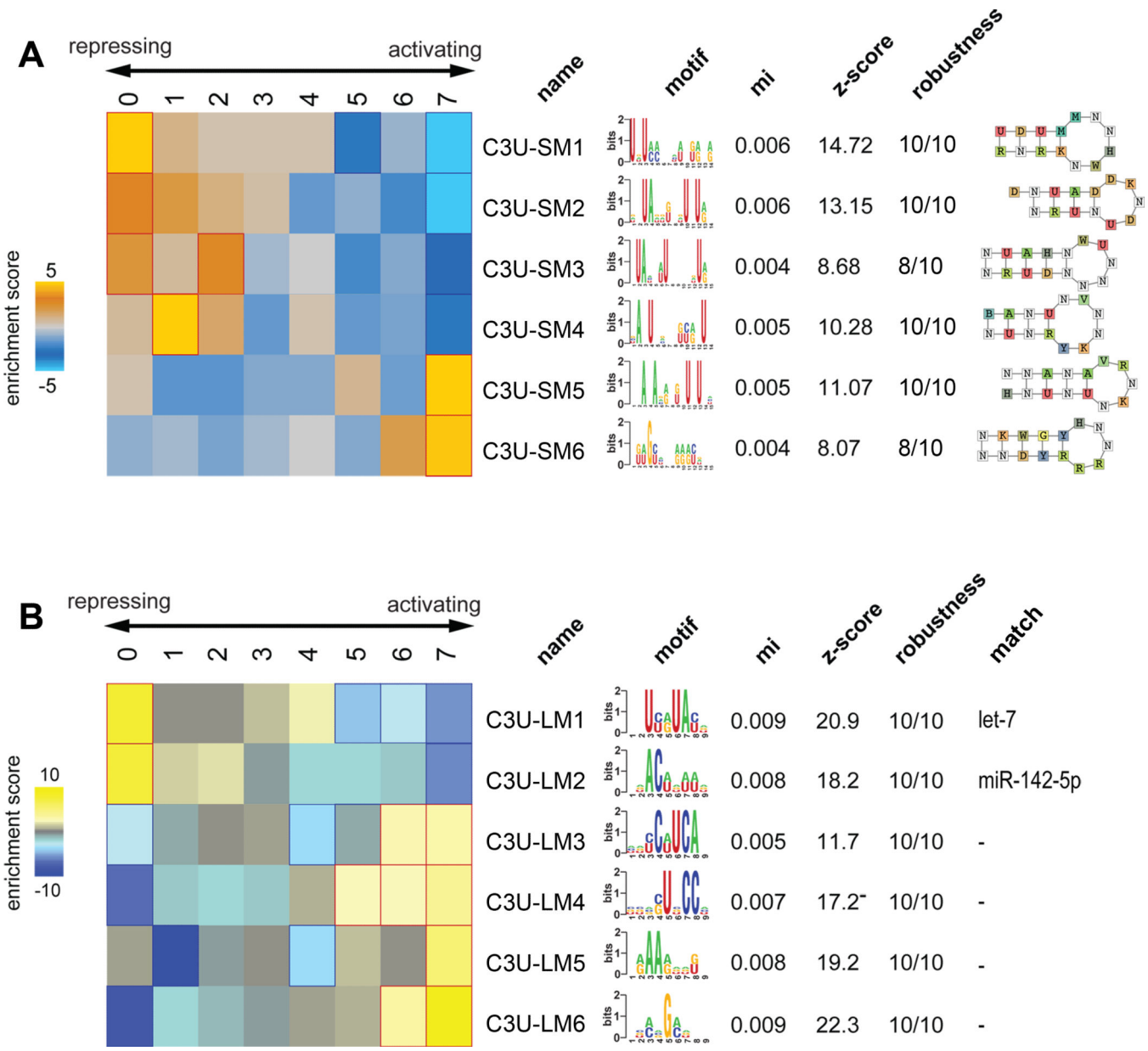

Figure 5. Discovery of Informative Post-transcriptional Regulatory Motifs

A partial list of structural RNA motifs discovered by TEISER (A) and linear RNA motifs discovered by FIRE (B) within the 3'UTR sequence library. For the complete sets see Fig. S3C and Fig. S3D. Sequences were clustered into eight sets as in Fig.4. Over representation (orange/yellow) and under representation (blue) patterns are shown for each discovered motif in the corresponding cluster. Reported are each motif's assigned name, its primary sequence and for structural elements an illustration of its secondary structure using the following single letter nucleotide code: Y = [UC], R = [AG], K = [UG], M = [AC], S = [GC], W = [AU], B = [GUC], D = [GAU], H = [ACU], V = [GCA] and N = any nucleotide. Also shown are the mutual information values, *z*-scores associated with a randomization-based statistical test, robustness scores from a three-fold jackknifing test, and matches to

target sequences of known regulators (for details see Elemento et al., 2007; Goodarzi et al., 2012). Also see Figures S3 and S4.

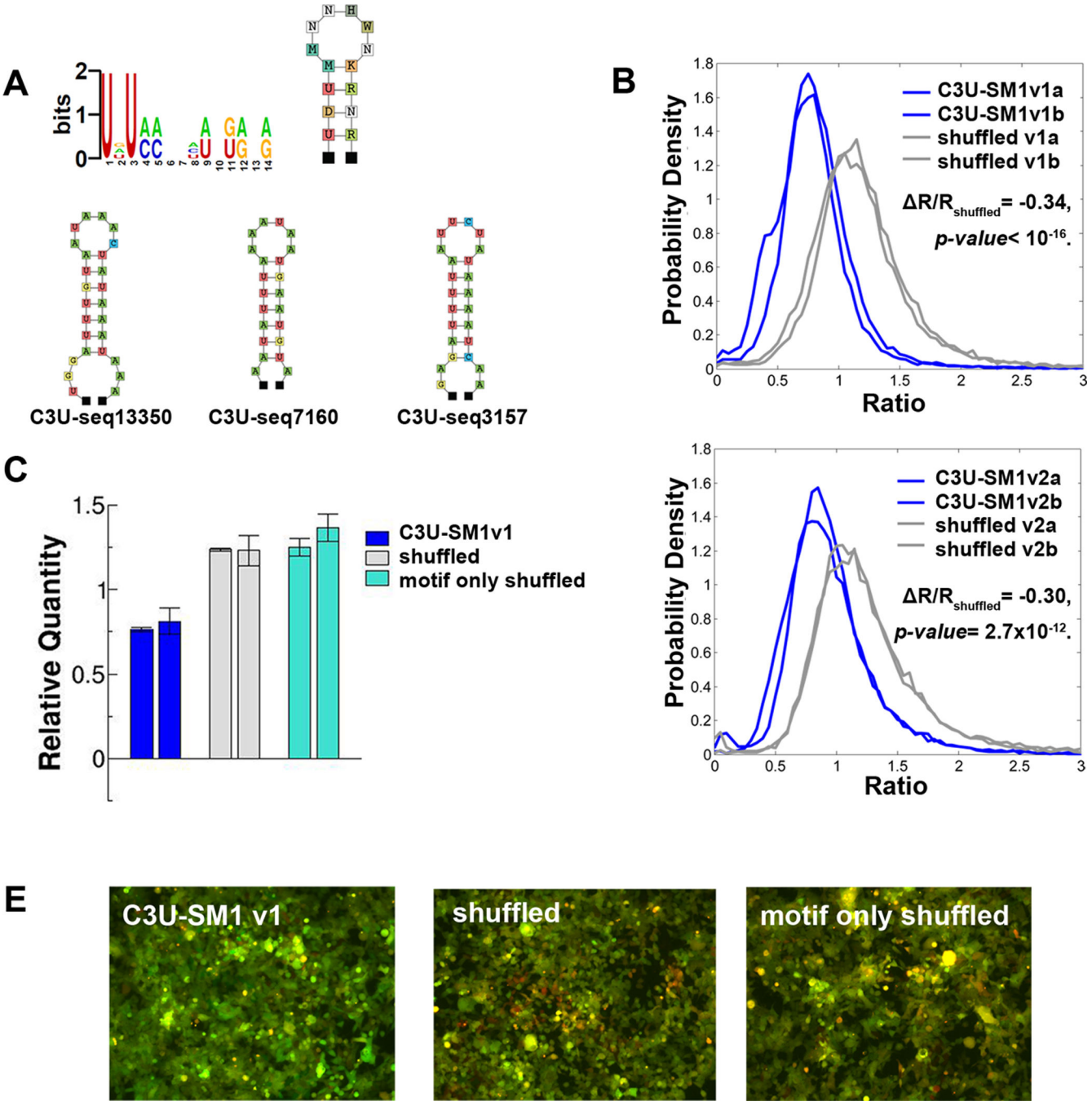

Figure 6. The regulatory consequences of the structural RNA motif C3U-SM1 on transcript abundance

(A) C3U-SM1's sequence logo, secondary structure and three endogenous instances. (B) The effect of C3U-SM1 on gene expression. Two sets (C3U-SM1v1 and C3U-SM1v2) each comprising two different instances of C3U-SM1 from the C3U Library were cloned into the bidirectional reporter construct along with two different shuffled versions as controls (see Experimental Procedures). Two clonal populations were generated for each construct and the distribution of mCherry to GFP ratio (DIR) was measured using flow cytometry. (C) The mCherry transcript levels, as measured by quantitative PCR, were significantly lower in

cells transiently transfected with C3U-SM1 constructs compared to the shuffled controls (data are represented as mean ±SEM, *p*-values = 4×10^{-5} and 1.5×10^{-5} respectively). (D) Representative microscopy images for bidirectional reporter cell lines containing C3U-SM1 instances vs. shuffled controls. GFP (green) and mCherry (red) images are shown overlapped. See also Figure S5 and Table S3.

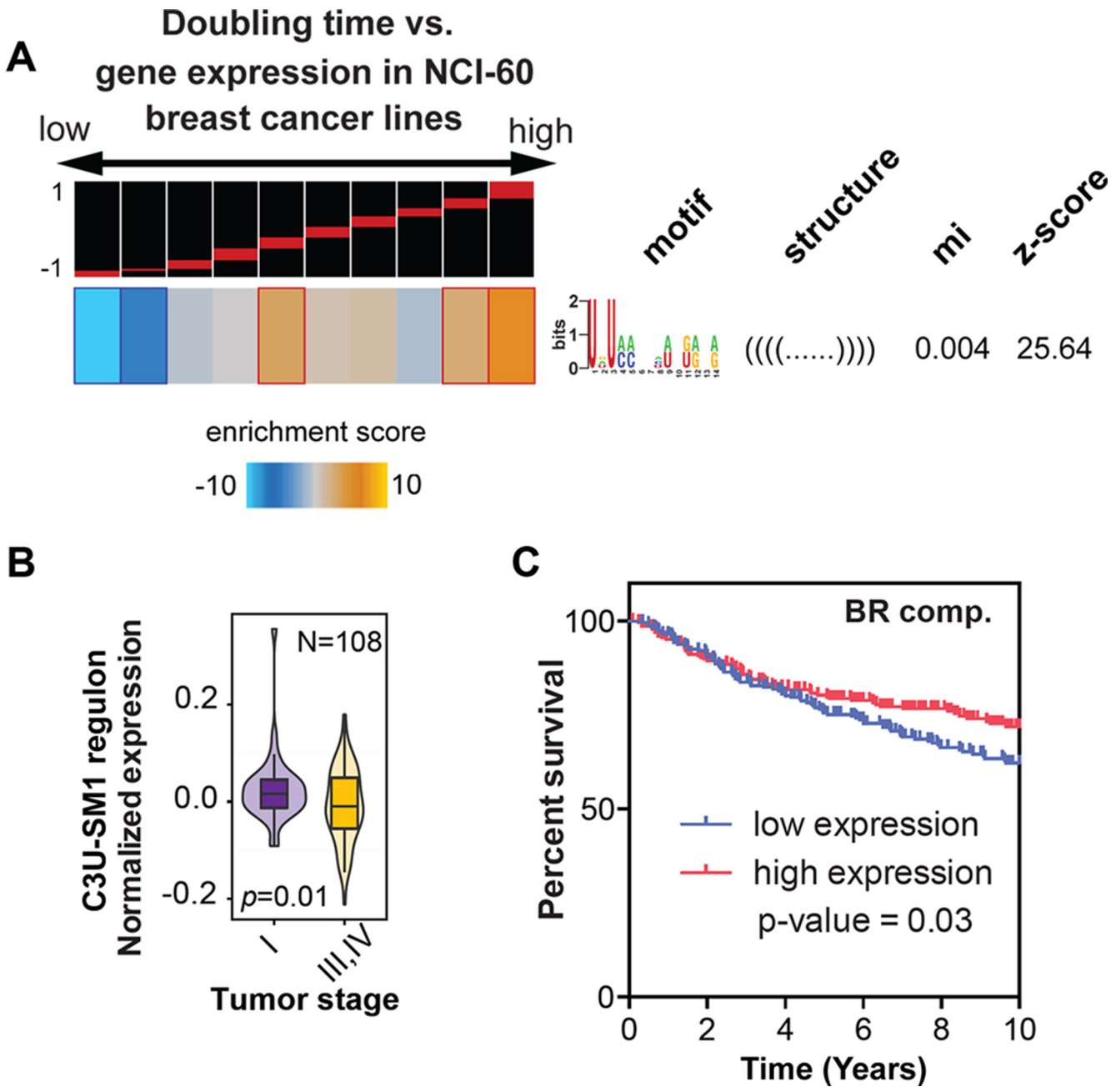

Figure 7. C3U-SM1 is informative of cancer cell proliferation rates and patient outcome

(A) Gene expression profiles across five breast cancer cell lines were correlated with cell line doubling times. The resulting values were analyzed using TEISER and over- and under-representation patterns of transcripts are shown for C3U-SM1. Instances of C3U-SM1 are over represented in bins with positive correlation values for breast cancer cell lines. For genes with positive correlation values, high expression indicates low proliferation rates (*p*-value=10^{-7}). (B) The aggregate expression level of transcripts with C3U-SM1 instances is lower for more advanced stage tumors (C) Patients with high aggregate expression levels in transcripts carrying C3U-SM1 in their 3'UTR showed better survival outcomes (N=459).

The combined *p*-value for these three independent observations (Fisher's method) is 10^{-8}. See also Figure S6 and S7.