Savinov *et al. Genome Biology* (2021) 22:293
https://doi.org/10.1186/s13059-021-02509-6

Genome Biology

RESEARCH Open Access

Effects of sequence motifs in the yeast 3′ untranslated region determined from massively parallel assays of random sequences

Andrew Savinov[1,2], Benjamin M. Brandsen[1,3], Brooke E. Angell[1,4], Josh T. Cuperus[1*] and Stanley Fields[1,5*]

* Correspondence: cuperusj@uw.edu; fields@uw.edu
[1]Department of Genome Sciences, University of Washington, Box 355065, Seattle, WA 98195, USA
Full list of author information is available at the end of the article

Abstract

Background: The 3′ untranslated region (UTR) plays critical roles in determining the level of gene expression through effects on activities such as mRNA stability and translation. Functional elements within this region have largely been identified through analyses of native genes, which contain multiple co-evolved sequence features.

Results: To explore the effects of 3′ UTR sequence elements outside of native sequence contexts, we analyze hundreds of thousands of random 50-mers inserted into the 3′ UTR of a reporter gene in the yeast *Saccharomyces cerevisiae*. We determine relative protein expression levels from the fitness of transformants in a growth selection. We find that the consensus 3′ UTR efficiency element significantly boosts expression, independent of sequence context; on the other hand, the consensus positioning element has only a small effect on expression. Some sequence motifs that are binding sites for Puf proteins substantially increase expression in the library, despite these proteins generally being associated with post-transcriptional downregulation of native mRNAs. Our measurements also allow a systematic examination of the effects of point mutations within efficiency element motifs across diverse sequence backgrounds. These mutational scans reveal the relative in vivo importance of individual bases in the efficiency element, which likely reflects their roles in binding the Hrp1 protein involved in cleavage and polyadenylation.

Conclusions: The regulatory effects of some 3′ UTR sequence features, like the efficiency element, are consistent regardless of sequence context. In contrast, the consequences of other 3′ UTR features appear to be strongly dependent on their evolved context within native genes.

Keywords: 3′ untranslated region, mRNA processing, Efficiency element, Puf protein, Massively parallel reporter assay

Background

The regulation of gene expression is central to biology, enabling functions ranging from environmental adaptation to animal development. However, deciphering the underlying logic of this regulation is difficult using only natural genetic elements because the relevant sequences in any organism vastly under-sample sequence space. For example, the roughly 6000 genes of the yeast *Saccharomyces cerevisiae* or 20,000 human protein-coding genes are dwarfed by even the set of possible 20-mer DNA sequences ($\sim 1.1 \times 10^{12}$), let alone the set of possible sequences approaching the lengths of regulatory sequences, which can span hundreds or thousands of base pairs. In addition, the regulatory sequences sampled by evolution are only a small number of the possible outcomes. Thus, additional facets of gene regulation might be learned by systematically interrogating the functional consequences of libraries of random synthetic sequences whose size vastly exceeds the number of an organism's genes. Enabled by advances in high-throughput sequencing and oligonucleotide synthesis, this approach has been taken to develop a deeper understanding of 5′ untranslated regions (UTRs) of mRNAs [1, 2], promoters [3, 4], and splicing [5].

Here, we extend this massively parallel approach to the regulatory grammar of 3′ UTR sequences in the model eukaryote *S. cerevisiae*. The 3′ UTR plays important roles in mRNA metabolism, affecting mRNA stability, translation, and localization [6]. These activities are mediated by proteins that bind to sequence and structural features of 3′ UTRs. High-throughput studies of naturally occurring 3′ UTRs from yeast [7, 8], humans [9–13], and zebrafish [14] have identified sequence motifs that significantly affect mRNA abundance, mRNA stability, and protein production. Additional work has revealed sequence motifs that determine sites of polyadenylation [15, 16]. In yeast, work based largely on a few well-studied genes [17–19], especially *CYC1* [20–22], has identified three sequence elements in the 3′ UTR that play large roles in determining gene expression levels as well as 3′ end cleavage and polyadenylation. These sequence features are termed the efficiency element (consensus UAUAUA), positioning element (consensus AAWAAA, with W an A or U), and cleavage and polyadenylation site (YA_N, with Y a C or U) [22]. Biochemical and structural investigations have shown that the efficiency element binds Hrp1 [23, 24], which in turn recruits the rest of the cleavage factor I (CF I) complex. This complex is required for efficient cleavage and polyadenylation and includes the Rna15 protein, which associates with the positioning element in the context of this complex [25, 26].

Measurements of the protein levels associated with ~ 13,000 3′ UTR sequences, largely from the yeast transcriptome as well as mutant versions of 217 native sequences, demonstrated a major role for the efficiency element [8]. Studies have also interrogated yeast mRNA stabilities transcriptome-wide [27–30]. One such study [21] suggested that poly(U) elements near the 3′ end of 3′ UTRs are important determinants of stability, and hence gene expression levels, an effect thought to be mediated by formation of RNA hairpins with the poly(A) tail. Investigations of native yeast genes have also suggested stabilizing and destabilizing roles for sequence motifs associated with binding by various RNA-binding proteins, most notably the Puf family of proteins [31–34], often via experiments that deleted or over-expressed Puf protein genes. In yeast, Puf proteins primarily function as repressors of gene expression via mRNA destabilization [35, 36].

However, as Puf proteins act via recruitment of additional factors, some Puf protein binding sites lead to mRNA localization or increased translation [6, 37, 38].

As naturally occurring 3′ UTR sequences have evolved to function in specific biological contexts, measuring the effect of a sequence element in a native 3′ UTR sequence context is complicated by the possible effects of co-evolved sequence features. Thus, we sought to build on the foundational studies of native yeast 3′ UTRs by performing a high-throughput assay of the expression of a single reporter gene under the regulatory control of hundreds of thousands of random 3′ UTR sequences. We determined that the efficiency element is the major regulator of gene expression, independent of sequence context. On the other hand, the positioning element and poly(U) motifs had only modest effects on expression. Three Puf protein binding site sequences were associated with substantially enhanced expression in these random sequence backgrounds, opposite to their effect in native mRNAs, pointing to a predominant role for sequence context in Puf protein-based regulation. The large number of 3′ UTR sequence variants analyzed in these experiments also allowed us to determine the effects of single base changes in 3′ UTR elements across diverse random sequence backgrounds, indicating the relative importance of each base in efficiency element sequences.

Results

Library and assay design

To assay the effects of random 50-base elements (N50) within a 3′ UTR, we generated two libraries in the context of the *HIS3* gene coding sequence and the *CYC1* gene promoter and 3′ UTR sequences, using a low copy number centromeric vector that carries a *LEU2* selection marker (see "Methods"). The random sequence was synthesized from equal ratios of the four nucleotides at each position. In one library (termed N50-EPC), we replaced the first 102 bases of the *CYC1* 3′ UTR with the N50 element. This N50 element was positioned between the *HIS3* termination codon and a region of 50 bases of *CYC1* that includes the efficiency and positioning elements, the cleavage site where polyadenylation occurs, and 101 bases of constant sequence that constitute the remaining region of the *CYC1* terminator (Fig. 1a). In the other library (termed N50-C), the sequence 3′ of the N50 element included only the cleavage site and the same downstream constant sequence derived from the 3′ end of *CYC1* as in N50-EPC (Fig. 1a). Based on estimates of the number of unique transformants, the N50-EPC library consisted of 2.1 million variants and the N50-C library consisted of 2.5 million variants. Our rationale for generating these two N50 libraries was that the N50-EPC library should provide a reasonably high baseline of faithful 3′ processing through the use of the canonical *CYC1* elements, allowing the identification of random elements that would modulate gene expression around this baseline; the N50-C library, lacking invariant efficiency and positioning elements, was intended to have low baseline expression and thereby reveal sequence features that increase expression levels.

The plasmid-borne *HIS3* reporter gene was transformed into a yeast *his3* and *leu2* deletion mutant by leucine selection in order to ensure that transformation and plasmid maintenance did not confound the histidine-based growth selection readout. Following transformation, we selected for growth in media lacking both leucine and

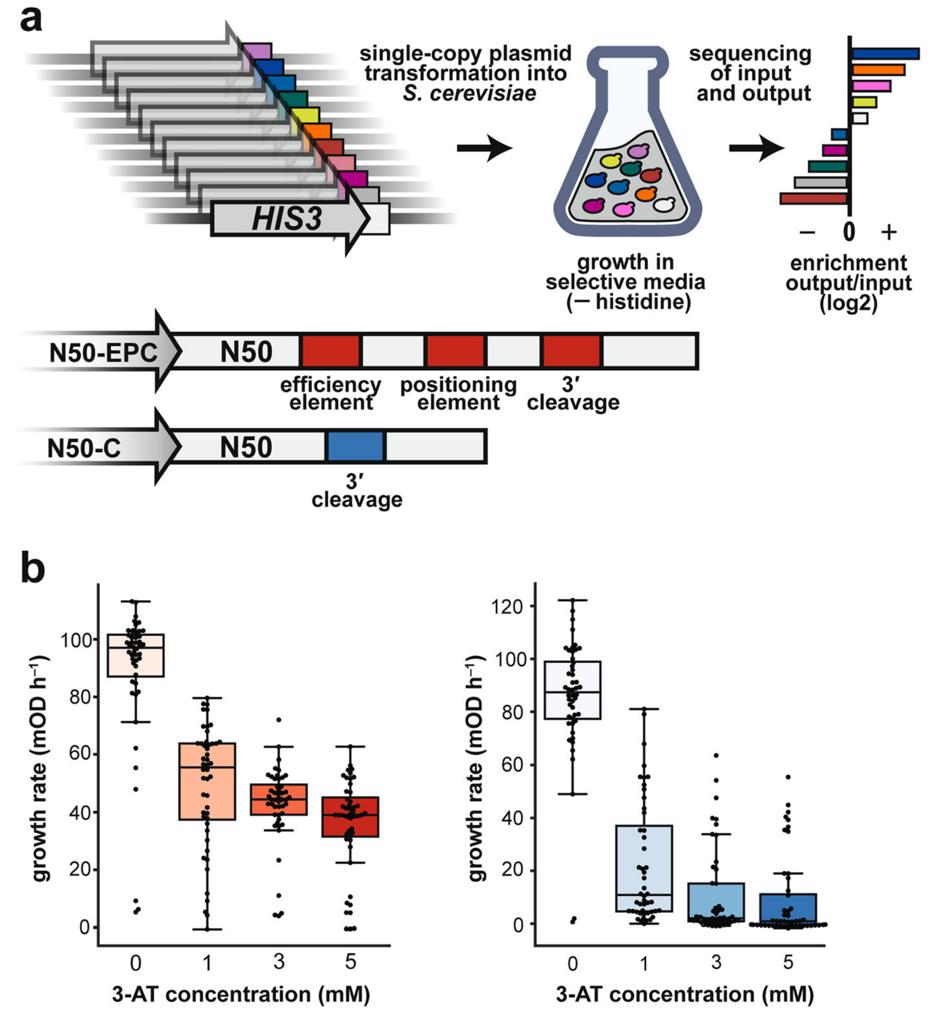

Fig. 1 A massively parallel assay of the effects of random 50 bp sequences on expression mediated by 3′ UTRs. **a** Library design and selection assay (upper) and library layout schematic, including known 3′ UTR motifs present in the two libraries by design (lower). **b** Determination of optimal 3-AT concentrations through growth rate measurements of 45 random variants each from the N50-EPC (left) and N50-C (right) libraries

histidine, and supplemented with 3-amino-1,2,4-triazole (3-AT), a competitive inhibitor of His3, which allowed us to read out the relative expression of library variants at the His3 protein level (see "Methods"). By testing 45 variants from each of the two libraries over a range of 3-AT concentrations, we established that 1 mM yielded the greatest dynamic range of growth rates (Fig. 1b). The use of the His3 selection provided a continuous readout of protein expression that did not depend on a FACS binning strategy, as would be necessitated by fluorescence-based readouts [3, 8, 39, 40]. The library design and selection strategy was derived from previous work to investigate 5′ UTR sequence variant effects [1] and validated in that work to report faithfully on relative His3 protein levels and growth rates of individual variants. We sequenced the N50 elements of each library prior to selection and after ~ 24 h (N50-EPC) or ~ 30 h (N50-C) of growth to OD_{600} = 1.0 in the absence of histidine and in the presence of 1 mM 3-AT; a single massively parallel growth selection was performed for each library (see "Methods").

The relative change in abundance of each variant is presented throughout the text as a $\log_2$ enrichment, *Enr*, equal to $\log_2(f_{\text{post-selection}}/f_{\text{pre-selection}})$, where $f_{\text{pre-selection}}$ and $f_{\text{post-selection}}$ denote population frequencies of the variant before and after selection. We filtered these data for minimum read counts to improve our confidence in the input and output variant frequencies, leaving ~ 590,000 N50-C sequences and ~ 280,000 N50-EPC sequence for which enrichment in the growth selection was quantified (see "Methods").

Overall properties of the N50-C and N50-EPC libraries

An initial analysis of sequences in the N50-C library revealed a correlation between overall AU content of the N50 element and His3 protein expression (Pearson's r = 0.27; Fig. 2a). In contrast, the same analysis performed for the N50-EPC library showed a striking lack of correlation (Pearson's r = − 0.021; Fig. 2b). These results hinted at a greater sequence dependence of expression in the N50-C context compared to the N50-EPC context. We thus sought to identify other 3′ UTR sequence features besides AU content that act as determinants of expression in the randomized N50 sequence

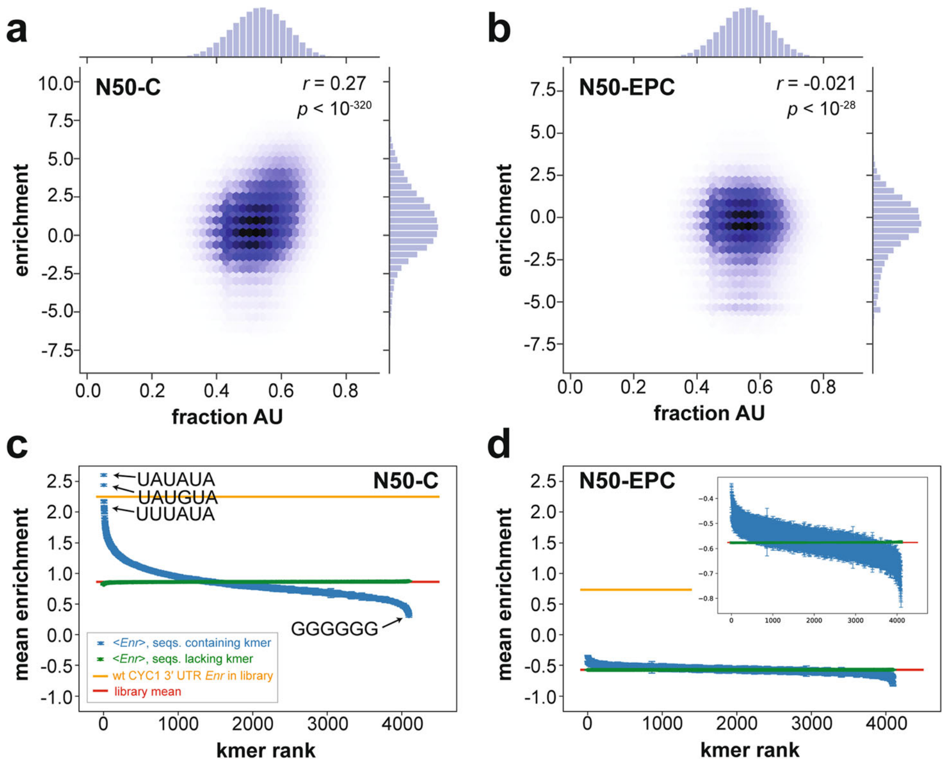

Fig. 2 Comparison of AU content and k-mer effects for the N50-C and N50-EPC libraries. **a, b** Enrichment scores of the N50-C library (**a**) and the N50-EPC library (**b**) as a function of 50-mer sequence AU content, with values of Pearson's *r* indicated. **c**, **d** Plots of average expression effects of all possible 6-mer sequences across the N50-C (**c**) and N50-EPC (**d**) libraries. The horizontal axis displays 6-mer sequence "rank" based on level of expression of N50 sequences containing each 6-mer (i.e., the 6-mer associated with the highest expression is assigned rank 1). Blue data, average enrichment across all library sequences containing the 6-mer; green data, average enrichment across all library sequences lacking the 6-mer; error bars, s.e.m.; red line, average enrichment across all library sequences; orange line, enrichment of plasmid constructs bearing the wild-type (wt) *CYC1* 3′ UTR sequence, with no random 50-mer. The identities of several individual example 6-mers are indicated. Inset in **d** shows 6-mer effects in the N50-EPC library on an expanded scale

backgrounds, initially by carrying out a systematic analysis of the effects of all possible 6-mer RNA sequences in the libraries (Fig. 2c, d). We found that the average $\log_2$ enrichment (*Enr*) of library sequences carrying a given 6-mer ranged from 0.30 to 2.60 in the N50-C library (library mean *Enr* of 0.86), but only − 0.76 to − 0.39 (library mean *Enr* of − 0.57) in the N50-EPC library. The range of 6-mer effects in the N50-EPC library was comparable to the uncertainties in the mean effects of each 6-mer (Fig. 2d, inset). Although the distributions of variant growth rates (Fig. 1b) led to *Enr* ranges that differed in the N50-C and N50-EPC libraries, the effects can be compared between the two libraries by considering the enrichment relative to the library mean (Fig. 2c,d red). Another point of comparison is provided by the enrichment of plasmids carrying the high-expression wild-type (no N50 inserted) *CYC1* 3′ UTR sequence, which were present in both libraries; these wild-type *CYC1* 3′ UTR plasmids yielded *Enr* = 2.25 in the N50-C library and *Enr* = 0.73 in the N50-EPC library (Fig. 2c, d, orange). Thus, the wild-type *CYC1*-normalized mean enrichment (*CYC1* enrichment subtracted from mean enrichment) is similar between the two libraries (N50-C, − 1.39; N50-EPC, − 1.31), as expected. Given the minimal expression consequences of N50 sequence content in the N50-EPC library, we focused our subsequent analyses on the N50-C library data.

The number of sequences containing any given 6-mer sequence in the N50-C library was substantial, leading to good estimates of the average effects of 6-mers on protein expression in a random sequence context; each hexamer was carried by at least 846 and typically thousands (mean ± s.d., 6438 ± 2956) of N50-C sequences. In the N50-C library, the 6-mer producing the highest average expression was UAUAUA (average *Enr* of 2.60, corresponding to an average ~ 6-fold enrichment in the selection across ~ 14,000 random sequences containing this hexamer; Fig. 2c). UAUAUA is the consensus efficiency element, and the next five highest-ranked 6-mer features (down to an average *Enr* of 2.09) were all point mutants of this motif. These six sequences were followed in rank by AUAUAU (*Enr* of 2.06) and the related sequence AUAUAA (*Enr* of 2.05). Lower-ranked 6-mers generally contained an increasing proportion of G and C bases. The most detrimental 6-mer was GGGGGG (average *Enr* of 0.30), with sequences such as GGAGGG, GGGAGG, and GGGGGA having similar effects (average *Enr* ~ 0.33). These results demonstrate that the growth selection assay was capable of detecting sequence features associated with reduced, as well as enhanced, protein expression.

We performed equivalent analyses of the average enrichment of N50-C sequences containing all possible 4-, 5-, 7-, and 8-mers, and obtained similar results (Additional File 1: Fig. S1; see also the Source Data [41]). The most highly enriched sequences in all cases were $(UA)_N$ repeats and their point mutants; the lowest-ranking sequences were consistently poly(G) stretches and variations on this theme with single non-G bases included. Across all *k*-mers considered in this analysis, a preference for AU-rich sequences was maintained (Additional File 1: Fig. S2a). We also performed analyses of 4-mers, 5-mers, 7-mers, and 8-mers found in N50-EPC library sequences. For this library, *k*-mers present in the most enriched sequences did not have higher levels of A and U bases (Additional File 1: Fig. S2b), consistent with our observation that enriched sequences in the N50-EPC library were not AU-rich. The same minimal consequences of expression seen for 6-mers were observed for 4-, 5-, 7-, and 8-mers, although the 8-mer data were noisy such that they are difficult to interpret (Additional File 1: Fig. S3, Source Data [41]).

To further explore potential motifs in these random libraries, we used kpLogo [42] to find enriched and depleted bases within the N50 sequence. We subsampled 50,000 random sequences from the N50-C and N50-EPC libraries and, weighting each sequence by its enrichment score, determined the relative enrichment and depletion associated with specific bases at each position in the N50 region (Additional File 1: Fig. S4a). In sequences from the N50-C library, U and A bases found near the beginning of the N50 region were associated with high enrichment, and G was associated with depletion at most positions in the sequence. In sequences from the N50-EPC library, only modest enrichment and depletion were associated with bases at any position in the sequence (Additional File 1: Fig. S4b), consistent with our observation that the N50 sequence only minimally affected a variant's enrichment score in the N50-EPC context.

We also sought to identify sequence elements or motifs in these data de novo using the MEME Suite program STREME [43]. Searching for motifs associated with either high or low enrichment (> 2.5σ away the mean), and using a randomly sampled set of library sequences of the same size as the background, we found that by far the most significant and frequently occurring motif was an N50-C motif associated with enrichment and containing a UAUAUA 6-mer (p = 1.12 × 10^{-43}); the few other motifs reported by STREME had barely significant p values (especially given the large library size) and did not occur often (Additional File 1: Fig. S5; see also Source Data [41]). Examining these motifs using the MEME Suite motif search tool Tomtom [44] using the RNA binding sites database, which contains only four *S. cerevisiae* proteins [45], resulted in a highly significant alignment to Hrp1 for the enriched N50-C motif containing UAUAUA, as well as two other enrichment-associated motifs; the other motifs did not yield significant matches to these four RNA-binding proteins (Additional File 1: Fig. S5). The occurrence of a depletion-associated motif rich in G bases in the N50-C library could indicate a structural phenomenon, similar to that observed in yeast 5′ UTRs [1].

Overall, our findings that the consensus efficiency element UAUAUA and closely related sequences predominated among high-expression variants in the N50-C library, and that no *k*-mer tested had any significant effect on expression in the N50-EPC library, demonstrate that the efficiency element is the dominant 3′ UTR feature for setting the protein expression level. Perhaps due to the strong effects of efficiency element motifs, other sequence motifs were difficult to discover de novo in this library, similar to the results of Shalem et al. [8]. However, other important 3′ UTR sequence elements and their functions in a biological context have been established based on extensive studies of native yeast genes. Thus, the N50-C library presented a novel opportunity to develop an understanding of the functions of these key motifs outside of a biologically evolved sequence context.

Sequence determinants of efficiency element function

We sought to analyze the effects of specific motifs on relative protein expression levels, beginning with the core sequence elements previously found to be involved in cleavage and polyadenylation [22]. We first analyzed the average expression of N50 sequences carrying the canonical consensus efficiency element UAUAUA (Fig. 3a, middle), which as noted was associated with the largest expression boost of any 6-mer sequence (Fig. 2c). In contrast, shuffled sequences derived from this motif (Hamming distance of ≥ 3; see “Methods”), which have the same AU content, had a smaller effect on expression (average ~ 2.8-fold enrichment; Fig. 3a, right), demonstrating that the specific sequence of UAUAUA is necessary for maximal effect. This conclusion is also evident from histograms revealing the distributions of motif effects in the N50-C library (Additional File 1: Fig. S6a). These results confirm the generality of this motif’s importance, which had been inferred largely from native sequences and select synthetic contexts [8, 21, 22]. In particular, our findings demonstrate that a UAUAUA efficiency element increases protein expression regardless of sequence background, without reliance on nearby co-evolved sequence features.

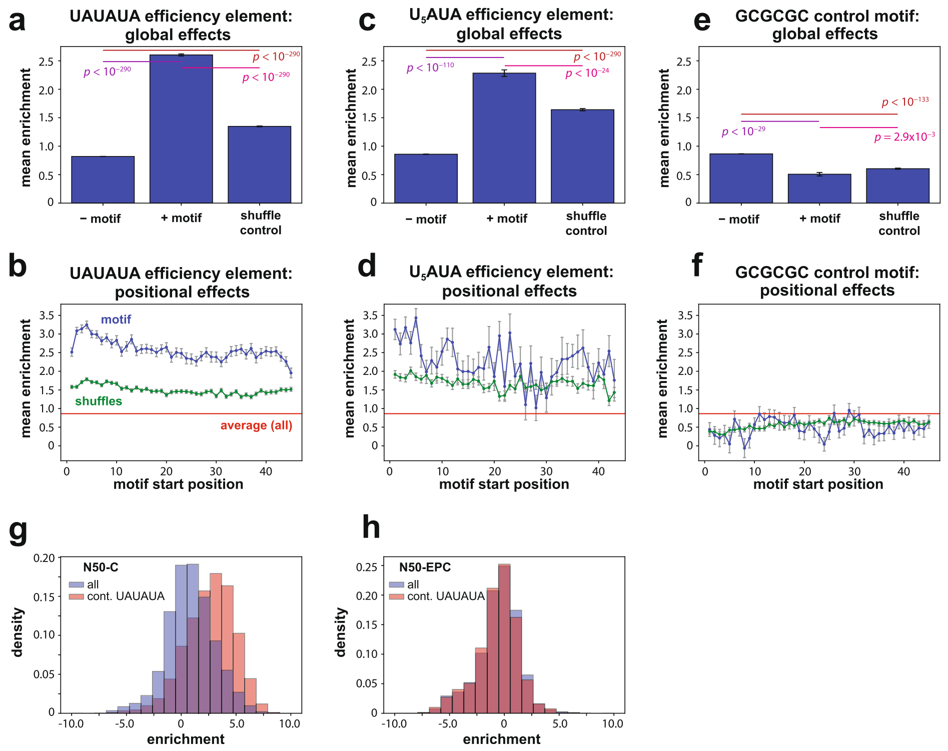

Fig. 3 Effects of efficiency element sequences in a random context. **a** Average effects of UAUAUA sequence motifs on growth selection enrichment across the N50-C library. Bars correspond to (left to right): mean across all sequences lacking the indicated motif, mean across all sequences containing the indicated motif, and mean across sequences containing shuffles of the motif but not the motif itself (Hamming distance minimum = half of motif length, see “Methods”). Error bars indicate standard error of the mean (s.e.m.); *p* values shown for each pairwise comparison are from Welch’s two-sided *t*-tests. **b** Average effects in the growth selection of UAUAUA sequence elements with 5′ end of the motif located at each position in the N50; blue, sequences containing the motif; green, sequences containing shuffles of the motif but not the motif itself (see “Methods”); red, average enrichment across all N50-C library sequences. Error bars indicate s.e.m. **c, d** As for **a** and **b**, respectively, but for the alternative efficiency element U_5AUA. **e**, **f** As for **a** and **b**, respectively, but for the control hexamer sequence GCGCGC. Bars in **a**, **c**, and **e** are derived from the following numbers of 3′ UTR variants: from left to right, in **a**, $N = 575757$, $N = 14267$, and $N = 5593$; in **c**, $N = 588534$, $N = 1490$, and $N = 1037$; and in **e**, $N = 585773$, $N = 4251$, and $N = 4608$. **g, h** Comparison of the effects of the consensus efficiency element motif UAUAUA in the N50-C library (**g**) and the N50-EPC library (**h**); enrichment score histograms of all sequences shown in blue, and of all sequences containing UAUAUA shown in tan. See also Supplementary Figure 6a–c for distributions of motif effects corresponding to panels **a**, **c**, and **e**

The large size of the library also allowed us to determine the average effects of the consensus efficiency element when its 5′ end is located at each position in the random 50-mer. Expression mediated by the efficiency element depended on its sequence location, with N50-C variants carrying this motif generally displaying higher expression levels when the element was localized further upstream within the 3′ UTR (Fig. 3b, blue). However, this consensus element remained beneficial for expression at all sequence locations. The effects of shuffled hexamers derived from UAUAUA showed far less position dependence (Fig. 3b, green), suggesting that the shuffled sequences may largely reflect generic benefits of higher AU content (Fig. 2a).

To determine which effects observed with the consensus element UAUAUA generalized to other efficiency element variants, we next considered the alternative efficiency element U_5AUA [20]. Compared to UAUAUA, U_5AUA had similar, but weaker, expression effects, both on the average enrichment across all sequences containing this motif (*Enr* = 2.28, or ~ 4.9-fold, across ~ 1500 sequences; Fig. 3c, middle; Additional File 1: Fig. S6b, red) and as a function of location within the random 50-mer (Fig. 3d, blue). Shuffled-sequence controls showed that the increase in expression associated with U_5AUA rose above AU content effects (Fig. 3c, right; Fig. 3d, green; Additional File 1: Fig. S6b, white), but to a lesser extent than for UAUAUA. As an additional control, we examined the effects of the sequence GCGCGC, the alternative pyrimidine and purine analog of the UAUAUA element. As expected, this GC-rich sequence was associated with lower-than-average enrichment (average *Enr* of 0.51, falling in the bottom 3.2% of 6-mer sequences), in a sequence context-independent and position-independent manner (Fig. 3e,f; Additional File 1: Fig. S6c).

We also examined the influence of the consensus efficiency element UAUAUA on the distribution of growth selection enrichments in both the N50-C and N50-EPC libraries. Compared to the distributions across all sequences, library variants containing UAUAUA yielded a shift towards higher protein expression across the N50-C library (Fig. 3g). In contrast, no such increase in expression was observed in the N50-EPC context, which contains an efficiency element in its constant sequence (Fig. 3h). In fact, there was a small reduction in expression when an additional efficiency element was present in the N50 sequence across the N50-EPC library (mean ± standard error of the mean (s.e.m.), *Enr* = − 0.575 ± 0.004 across all N50-EPC sequences, vs. *Enr* = − 0.648 ± 0.027 across N50-EPC sequences containing UAUAUA), suggesting that an extra efficiency element might be slightly detrimental in the context of UTRs containing efficiency and positioning elements and a cleavage site by reducing the efficiency of cleavage and polyadenylation. These findings appear to be specific to the optimized N50-EPC context; among the 81 sequences in the N50-C library containing two (non-overlapping) UAUAUA motifs, an additional efficiency element was associated with further increased protein expression (average *Enr* = 3.38 ± 0.23) compared to sequences carrying a single efficiency element (average *Enr* = 2.60 ± 0.03). Overall, these results suggest a “threshold model” for 3′ UTR gene regulation, in which an optimized efficiency element–positioning element–cleavage and polyA site architecture largely sets the expression level, to the exclusion of other regulatory sequences.

The selection assay results from the large 3′ UTR library effectively contained mutational scans of sequence motifs across diverse random sequence backgrounds. We sought to leverage these measurements to systematically investigate the functional role of each base in a number of motifs (Fig. 4; Additional File 1: Fig. S7), beginning with the consensus efficiency element UAUAUA. Considering first point mutations of UAUAUA that maintained AU content, we found that no such mutation yielded a larger boost in expression than the consensus sequence, as shown previously by the 6-mer analysis. Point mutations present at the 5′ and 3′ ends of the efficiency element were most detrimental to expression level compared to N50 sequences containing the unmutated consensus element, whereas the central bases were the least sensitive to mutation (Fig. 4a). These results suggest that the most important sequence-specific binding interactions of this element with the Hrp1 protein occur at the termini. Structural work suggests that Hrp1 makes binding contacts with all six bases of the efficiency element [24], and this mutational scan informs on the relative importance and specificity of these interactions in vivo. Results were similar with single mutations of UAUAUA that conserve pyrimidine or purine identity instead of AU content, although a G was superior to a U at position 4 (Fig. 4b), with this variant being the second highest-ranked hexamer sequence (Fig. 2c).

We performed a similar analysis for the alternative efficiency element U_5AUA. In this case, any AU content-maintaining point mutation at bases 3–8 of the motif reduced His3 expression substantially, apart from the U_4A mutation that yields a consensus UAUAUA efficiency element; mutations to the first two bases had no effect (Fig. 4c). Furthermore, changes to bases towards the 3′ end of the motif tended to reduce expression slightly more. These findings suggest that in the case of the U_5AUA element, the sequence U_3AUA is in fact responsible for Hrp1 binding, consistent with the same 6-mer binding mode as observed for the consensus efficiency element. Our analysis of *k*-mer effects showed that U_3AUA was the hexamer associated with the sixth highest expression level across the library (average *Enr* = 2.09; Fig. 2c), likely accounting for much of the activity in the U_5AUA context (average *Enr* = 2.28).

A mutational scan of an AU element initiating with an A rather than a U, AUAUAU (excluding all sequences that also contain UAUAUA due to a U preceding this motif), showed that single mutations had no effect at any site (Fig. 4d). This result suggests that this permuted efficiency element motif may be a poor site for Hrp1 recruitment in vivo, despite containing a nearly complete consensus site UAUAU, further highlighting the essential role of the bases at the 5′ and 3′ termini. The insensitivity of AUAUAU to point mutations, indicating a likely lack of specificity, is striking in the context of the relatively high expression conferred by this motif—nominally the seventh highest-ranked hexamer, average *Enr* = 2.06, and falling around rank 61 (in the top 1.5% of 6-mers), average *Enr* = 1.55, when UAUAUA-containing sequences are excluded.

We next considered variants of the consensus efficiency element containing additional (UA) repeats. Increasing the number of dinucleotide repeats to four or five did not further increase expression (Fig. 4e, f), suggesting that there is no increase in Hrp1 binding by elongated versions of the UAUAUA motif, with three UA repeats providing maximal affinity. In contrast, certain $(UA)_5$ mutants in which the dinucleotide repeat is broken up by point mutations did increase expression level beyond that associated with UAUAUA, with $(UA)_3U_3A$ providing the highest enrichment among variants investigated (~ 9-fold enrichment; Fig. 4f). Based on these findings, $(UA)_3U_3A$ may prove to be a useful generic efficiency element for achieving increased protein expression in

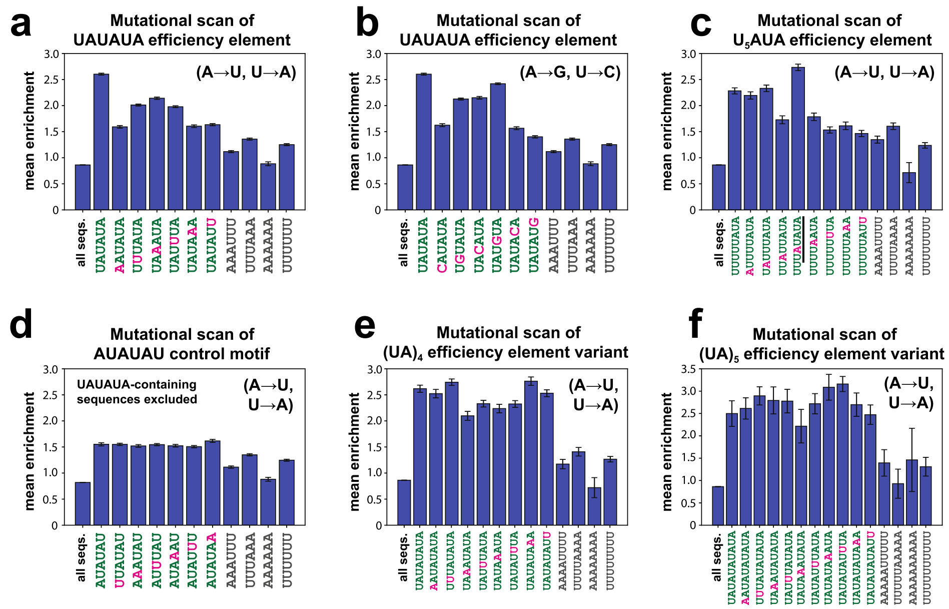

Fig. 4 Mutational scans reveal properties of efficiency element motifs. **a** to **f** Average effects of sequence motifs (green text) and point mutants of those motifs (magenta text) on enrichment scores, compared to all sequences in the dataset ("all seqs.") and controls with equivalent AU content (gray text). Searches for sequences containing point mutants of each motif also exclude sequences containing the consensus (unmutated) motif. **a** Single A⟶U and U⟶A substitutions and controls for the consensus efficiency element UAUAUA. **b** Single U⟶C or A⟶G substitutions and controls for the consensus efficiency element UAUAUA. **c** Single A⟶U and U⟶A substitutions and controls for the alternative efficiency element U_5AUA. Underline highlights the generation of a UAUAUA efficiency element through one of the point mutants of U_5AUA. **d** Single A⟶U and U⟶A substitutions and controls for the "inverted consensus" sequence AUAUAU. **e** Single A⟶U and U⟶A substitutions and controls for the extended efficiency element $(UA)_4$. **f** Single A⟶U and U⟶A substitutions and controls for the efficiency element $(UA)_5$. Throughout, error bars indicate s.e.m. See also Supplementary Figure 7 for distributions of motif effects corresponding to the mutational scans plotted in **a–f**

yeast. The limited increase in enrichment conferred by this motif over UAUAUA may partially be a consequence of the growth of the yeast becoming saturating under our selection conditions, as $(UA)_3U_3A$ may increase expression further than the measured enrichments reflect.

Effects of positioning element motifs and the optimal arrangement of efficiency and positioning elements

We next analyzed the positioning element, which plays a role in determining the site of cleavage and polyadenylation; mutations in this element lead to imprecise cleavage [21]. This element in yeast is A-rich, and its consensus sequence of AAWAAA bears striking similarity to the AAUAAA element found in 3′ UTRs of metazoans. Although changes in the precision by which cleavage and polyadenylation occur might be expected to affect expression by altering mRNA stability, we found that the presence of an AAWAAA element in the N50 had only modest effects on expression in the N50-C library (average *Enr* = 1.20, ~ 2.3-fold enrichment; Fig. 5a, middle), not substantially higher than a hexamer of equivalent AU content, AAAUUU (Fig. 5a, right; see also Additional File 1: Fig. S6d). AAAAAA and AAUAAA, which match the AAWAAA

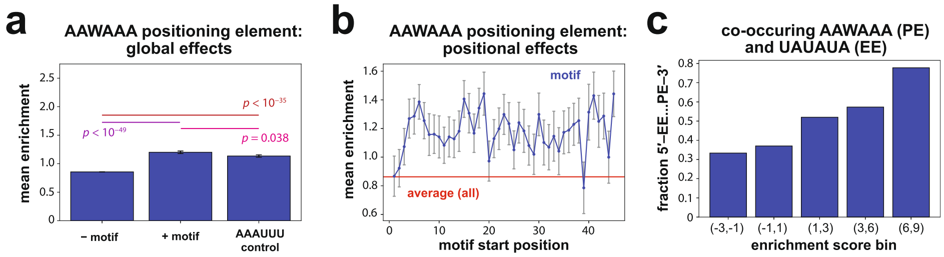

Fig. 5 Effects of the consensus positioning element in a random sequence context. **a** Effects of the consensus positioning element AAWAAA on growth selection enrichment across all sequences containing the motif in the N50-C library. Bars correspond to (left to right): mean across sequences in the N50-C library lacking the indicated motif; mean across sequences containing indicated motif; and mean across sequences containing AAAUUU, a control sequence of equivalent AU content. Error bars indicate s.e.m.; *p* values shown for each pairwise comparison are from Welch's two-sided *t*-tests. See also Supplementary Figure 6d for corresponding distributions of motif effects. **b** Average effects in the growth selection of sequences containing the AAWAAA element with 5′ end of the motif located at each position in the N50. Error bars indicate s.e.m. **c** Among N50-C library sequences containing both a consensus efficiency element (EE) (UAUAUA) and a consensus positioning element (PE) (AAWAAA), plot of the fraction with the EE located 5′ of the PE for sequences in each enrichment score bin. Bars in **a** and **c** are derived from the following numbers of 3′ UTR variants: from left to right, in **a**, *N* = 580254, *N* = 9770, and *N* = 10218; and in **c**, *N* = 9, *N* = 27, *N* = 50, *N* = 89, and *N* = 9

motif, had average enrichments of *Enr* = 0.91 and 1.36, respectively. However, AAUAAA falls within the top 4% of hexamer sequences despite its modest effect size, reflecting the rapid drop in associated enrichment with hexamer rank (Fig. 2c). A positional analysis of the effects of AAWAAA in random sequence backgrounds showed that the positioning element generally had similar effects when found at sites throughout the N50 sequence (Fig. 5b), consistent with observations that the location of a positioning element, while altering mRNA isoform distributions, minimally affected total mRNA abundance [22].

The positioning element might be expected to have strongly context-dependent effects on expression, given that binding of Rna15 protein to this element requires Hrp1 binding to a nearby efficiency element, allowing formation of the CF I complex which incorporates Rna15 [26]. To investigate the generalizability and properties of the efficiency element–positioning element interaction, we considered the fraction of N50 sequences containing the canonical consensus forms of both elements (UAUAUA and AAWAAA) in which the efficiency element is 5′ of the positioning element, across 3′ UTRs falling into different enrichment score bins. The fraction of sequences with this arrangement was larger in bins of increasingly higher enrichment scores (Fig. 5c), suggesting that the stereotyped arrangement of these elements derived from biologically occurring sequences is generally optimal for expression in any sequence context.

Some Puf protein binding sites increase protein expression in a random sequence context

We examined the results of the N50-C library selection on another class of 3′ UTR sequence elements—Puf protein binding sites—including binding site motifs for Puf1 and Puf2, Puf3, Puf4, Puf5, and Puf6. Although we did not discover Puf-associated motifs de novo in enriched or depleted sequences (Additional File 1: Fig. S5), we hypothesized that the large effects of the efficiency element motif may mask signals from other motifs, consistent with previous observations [8]. However, directly interrogating the effects of Puf motifs individually should nonetheless reveal their regulatory consequences in a random sequence background. By investigating shuffled versions of these motifs, we found that the Puf1 and Puf2 motif (UAAUNNNUAAU [46]) did not significantly impact His3 expression (beyond the effects of its concomitant AU content) (Fig. 6a,b; Additional File 1: Fig. S6e). In contrast, Puf3 (UGUANAUA [31, 38, 47]) (Fig. 6c,d; Additional File 1: Fig. S6f), Puf4 (UGUANANUA [31, 48, 49]) (Fig. 6e,f; Additional File 1: Fig. S6g), and Puf5 (UGUANNNNUA [31, 48]) (Fig. 6g,h; Additional File 1: Fig. S6h) motifs were associated with significantly enhanced protein expression, and the Puf6 site (UUGU [50, 51]) was associated with a weak but statistically significant increase in expression (beyond the shuffled sequence control) (Fig. 6i,j; Additional File 1: Fig. S6i). The strongest increases in expression were associated with Puf motifs located closer to the 5′ end of the 3′ UTR (Fig. 6d,f,h,j). These findings stand in contrast to the traditional view of yeast Puf proteins as repressive elements acting mainly through mRNA destabilization (reviewed in refs. 26, 27), and suggest that in the absence of additional co-evolved sequence features some Puf binding sites increase expression.

In the case of the Puf3 binding site, the associated enhancement of expression may be traced in part to the fact that this sequence contains UANAUA. Hence, this Puf3 site includes the consensus efficiency element UAUAUA or its point mutants at position 3, a position in this motif where mutations allowed enhancement of expression to be retained (Fig. 4a,b). Therefore, this Puf3 binding site exemplifies a type of dual regulation based on overlapping motifs, which has been noted in 3′ UTR regulation [52, 53]; in this case, the effect on expression is likely heavily influenced by strong efficiency element activity. The Puf3 binding site might thus be competed for by stabilizing and destabilizing proteins, with the relative levels of binding by Hrp1 and Puf3 likely to depend on the surrounding RNA sequence and other regulatory factors. This result may also reflect the context dependence of the regulatory effects of Puf proteins, with Puf3 in yeast producing opposing effects—either reduced mRNA levels or increased translation—depending on metabolic state [6, 38]. Biological context-dependent stabilizing or destabilizing effects have been observed with other RNA-binding proteins as well [54]. Overall, our results for the expression consequences of Puf protein motifs in the random N50 sequence background suggest that Puf protein regulation of native mRNAs depends on additional sequence context beyond the Puf binding site. These results are similar to the sequence dependence of mRNA binding by the mouse MBNL1 and RBFOX2 proteins [55], and more broadly to other studies documenting that 3′ UTR-binding proteins associate with only a fraction of their possible binding sites in vivo [6, 56–58]. The sequence dependence of Puf binding site activity that we infer from our results—and the expression increase that we find was mediated by some Puf sites—may also explain why binding by the typically repressive Puf1 and Puf3 proteins is stabilizing for at least some native mRNAs [59]. Binding by Puf4 or Puf5 proteins may be generically stabilizing, or may increase levels of translation, in the absence of other sequence elements involved in recruiting destabilizing factors. We note, too, that we have not shown that the expression-boosting effect measured for Puf protein motifs was due necessarily to the binding of the cognate Puf proteins.

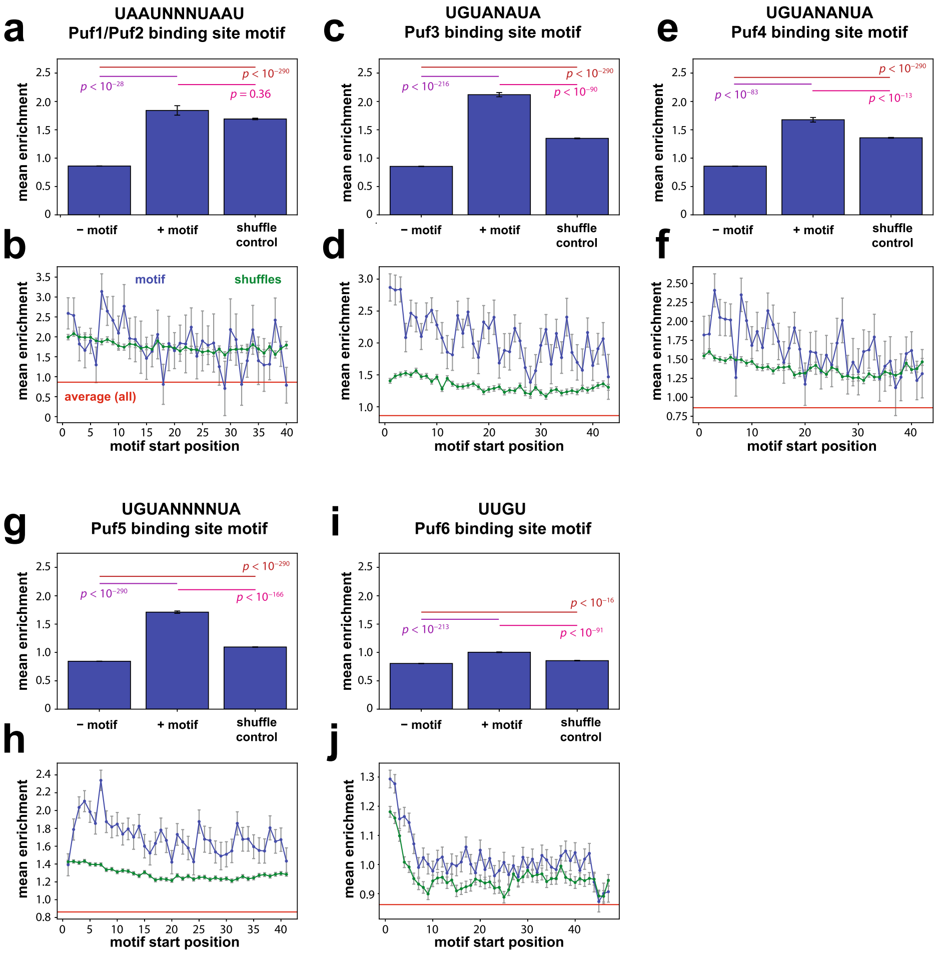

Fig. 6 Several Puf protein binding sites increase protein expression when placed in a random sequence context. **a**, **c**, **e**, **g**, **i** Average effects of each indicated Puf binding site motif on enrichment (as in Fig. 3a,c,e). Bars correspond to (left to right): mean across sequences in the N50-C library lacking the indicated motif; mean across sequences containing the indicated motif; and mean across sequences containing shuffles of the motif but not the motif itself. *p* values shown for each pairwise comparison are from Welch's two-sided *t*-tests. **b**, **d**, **f**, **h**, **j** Positional effects of the Puf protein binding sites indicated in **a**, **c**, **e**, **g**, and **i** above, and shuffled sequence controls, on average enrichment (as in Fig. 3b,d,f). Blue, average enrichment of the motif sequence; green, average enrichment of shuffles of the motif sequence; red, average enrichment across all N50 sequences. Throughout, error bars indicate s.e.m. Bars in **a**, **c**, **e**, **g**, **i** are derived from the following numbers of 3**′** UTR variants: from left to right, in **a**, *N* = 589289, *N* = 735, and *N* = 30004; in **c**, *N* = 586408, *N* = 3616, and *N* = 104188; in **e**, *N* = 587007, *N* = 3017, and *N* = 104299; in **g**, *N* = 577816, *N* = 12208, and *N* = 280954; and in **i**, *N* = 421083, *N* = 168941, and *N* = 195189. See also Supplementary Figure 6e-i for distributions of motif effects corresponding to panels **a**, **c**, **e**, **g**, and **i**

Effects of poly(U) sequences on expression

A poly(U) element near the 3′ end of yeast 3′ UTRs has been implicated in stabilizing mRNA through a proposed RNA hairpin formed with the poly(A) tail [30]. In agreement with a stabilizing effect, we observed a modest average increase in His3 expression in the N50-C library for N50 sequences containing U_8 stretches (Fig. 7a; Additional File 1: Fig. S6j). However, this boost in gene expression was weaker when the U_8 element is located in the 3′-most 25 bases of the N50 sequence (Fig. 7a), in contrast to the prior results [30], and instead was more substantial when the element was present in the 5′-most 25 bases (Fig. 7a). By calculating the average expression of

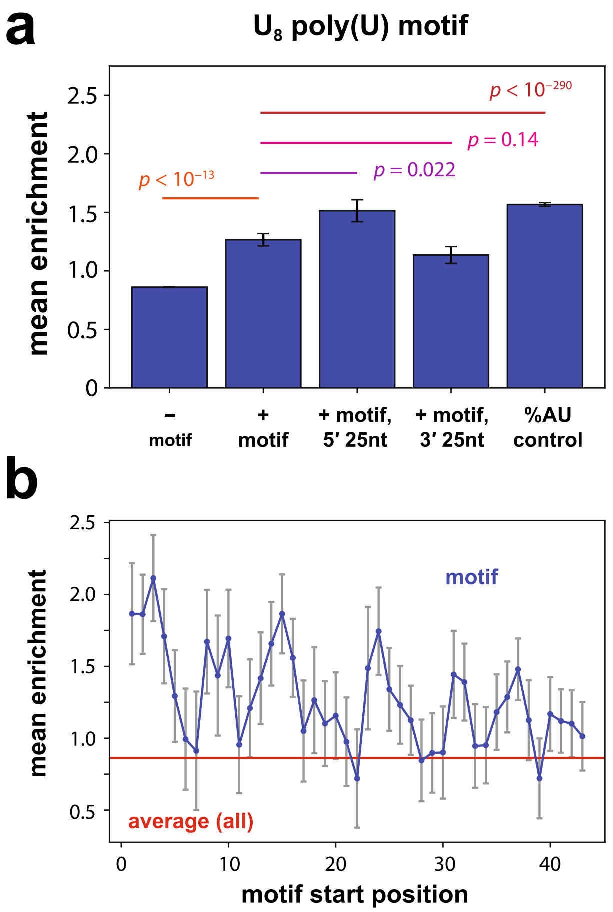

Fig. 7 Poly(U) sequence effects on gene expression across random contexts. **a** Average effects of the U_8 motif on enrichment. Plotted bars, left to right: average enrichment across sequences lacking U_8; average enrichment across sequences containing U_8; average enrichment across sequences containing U_8 in the 3′-most 25 nt of the 3′ UTR; average enrichment across sequences containing U_8 in the 5′-most 25 nt of the 3′ UTR; average enrichment of sequences containing shuffles of an 8-mer of equivalent AU content (A_4U_4) to compare with the effects of U_8. *p* values shown for each pairwise comparison are from Welch's two-sided *t*-tests. Bars in **a** are derived from the following numbers of 3′ UTR variants: from left to right, *N* = 588197, *N* = 1827, *N* = 543, *N* = 996, and *N* = 20181. See also Supplementary Figure 6j for corresponding distributions of motif effects. **b** Average enrichment across all sequences containing a U_8 motif with its 5′ end located at each N50 position (as in Fig. 3b,d,f). Blue, sequences containing the motif; red, average enrichment across all N50-C library sequences. Throughout, error bars indicate s.e.m.

sequences containing a U_8 motif at each position in the 50-mer, we found that U_8 increased expression most when present in the 5′ end of the N50, with weaker effects the closer the element is located to the 3′ end, and negligible effects at the 3′ terminus (Fig. 7b). Similar results were seen for U_6 and U_{10} stretches (Additional File 1: Fig. S8).

However, the expression enhancement associated with the U_8 sequence was smaller than the average effect of various 8-mer sequences containing 50% A and 50% U content (Fig. 7a; Additional File 1: Fig. S6j), and a U_8 sequence increased expression less than an equivalent U-rich sequence containing no more than two Us in succession (Additional File 1: Fig. S9). These results make it unclear whether the protein-level effects of poly(U) sequences are specific to the mRNA-stabilizing mechanism outlined by Geisberg et al. [30]. These findings suggest that the documented effects of 3′ UTR sequence motifs on mRNA stability may not necessarily predict expression outcomes at the protein level, as the sequence features of 3′ UTRs influence not only RNA stability but also translation.

Comparison of 3′ UTR sequence element effects to mRNA half-life measurements of native mRNAs

We sought to broadly compare the effects on protein expression of motifs in a random 50-mer background to the effects of these same motifs on mRNA stability in native sequence contexts. To make these comparisons, we leveraged published data on mRNA half-life across the yeast transcriptome [30], re-analyzing these data and calculating the average half-life of native yeast mRNAs containing various sequence features in their 3′ UTRs (see "Methods").

The average effects of each motif on N50-C library protein expression level and native mRNA half-life are plotted in Fig. 8 (see also Additional File 1: Figs. S6, S10 for the corresponding distributions of motif effects). We found that the increase in His3 protein expression associated with efficiency elements in the N50-C library matched an increase in native mRNA half-life, as expected (Fig. 8a,b), but the effect sizes were notably weaker in the native context. On the other hand, the presence of an AAWAAA positioning element sequence was associated with slightly beneficial effects on protein expression in the N50-C library, compared to slightly reduced native mRNA half-life (Fig. 8c). In a similar vein, GCGCGC was associated with reduced His3 protein expression, compared to an increase in native mRNA half-life (though with a large standard error; Fig. 8d). However, only 15 yeast genes contain a GCGCGC hexamer sequence in their 3′ UTRs, suggesting that it is evolutionarily disfavored in that context, perhaps because it typically reduces expression.

Our analysis of the average half-lives of native mRNAs containing Puf protein binding sites showed a nominally lower half-life for the Puf1/Puf2 site, although the difference was not statistically significant (Fig. 8e, green). These findings are consistent with Puf1/Puf2 sites reducing mRNA stability on average, but with this effect subject to the wide range of native mRNA half-lives and co-evolved regulatory contexts. The effect on mRNA stability was opposite to the increase in protein expression in the random N50 context, which seems to be driven by the AU content of Puf1/Puf2 sites (Fig. 8e, blue). The Puf3 binding site motif was associated with a somewhat longer mRNA half-life on average, which was similar to the effects of this element on His3 protein expression, presumably reflecting the efficiency element function of this site at both the protein and the mRNA level (Fig. 8f). However, Cheng et al. [34] found that the Puf3 binding site motif UGUAAAUA was associated with reduced half-life of native mRNAs, although this same motif became stabilizing in *puf3* and *ccr4* deletion backgrounds. These results suggest that differences in mRNA half-life results between the Cheng et al. [34] and Geisberg et al. [30] studies might relate to growth conditions. The Puf4 and Puf5 binding site motifs were both associated with reduced native mRNA half-life (Fig. 8g, h), in contrast with the increased protein expression mediated by these elements in a random N50 context. A possible explanation for this difference is that Puf4 and Puf5 sites alone increase mRNA and protein expression levels, but additional sequence features as are present in most yeast genes result in a destabilizing effect in the

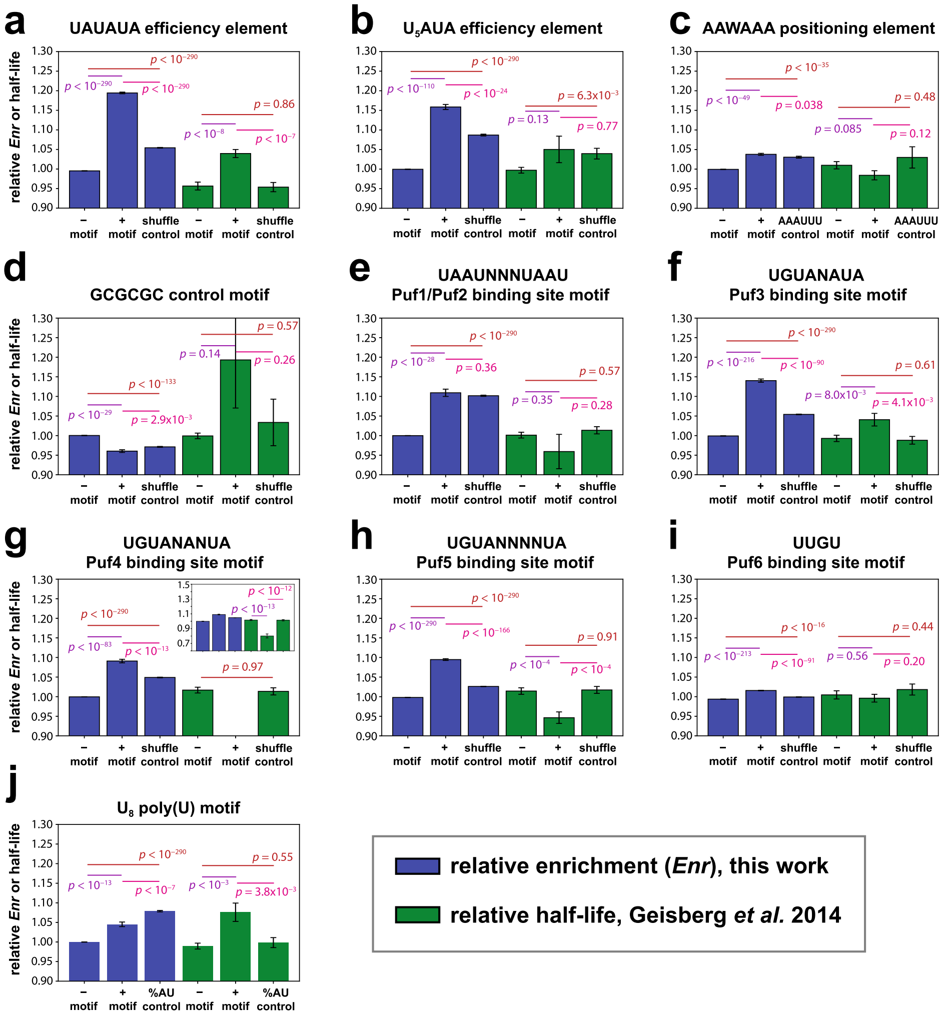

Fig. 8 (See legend on next page.)

(See figure on previous page.)

Fig. 8 Comparison of the effects of sequence motifs on expression in a random context and on mRNA stability in native genes. **a**–**j** Average effects of the noted sequence motifs on two measurements of gene expression: relative growth selection enrichment across all sequences containing the motif in the N50-C library (blue), and relative half-lives of native yeast 3′ UTRs carrying this motif [30] (green). Triplets of bars of each color in **a**–**j** correspond to (left to right): mean across sequences lacking the indicated motif; mean across sequences containing the indicated motif; and mean across sequences containing shuffles of the motif but not the motif itself, except as noted in the following. In the case of panel **c**, the third bar in each series instead represents AAAUUU, a control sequence with equivalent AU content to AAWAAA. In the case of panel **j**, the third bar in each series represents average relative enrichment or relative half-life of sequences containing shuffles of an 8-mer of equivalent AU content (A_4U_4) to compare with the effects of U_8. Throughout, error bars indicate s.e.m.; *p* values shown for each pairwise comparison are from Welch's two-sided *t*-tests. In panels **a–j**, blue bars (based on the growth selection enrichment data) are derived from the numbers of 3′ UTR variants listed for the matching panels in Figs. 3, 5, 6, and 7; green bars (based on the native mRNA half-life data) are derived from the following numbers of native yeast mRNAs: from left to right, in **a**, N = 1688, N = 1859, and N = 1273; in **b**, N = 3356, N = 191, and N = 1097; in **c**, N = 2189, N = 1358, and N = 256; in **d**, N = 3532, N = 15, and N = 67; in **e**, N = 3448, N = 99, and N = 2377; in **f**, N = 3016, N = 531, and N = 1980; in **g**, N = 3271, N = 276, and N = 2265; in **h**, N = 2799, N = 748, and N = 2531; in **i**, N = 1633, N = 1914, and N = 961; in **j**, N = 3115, N = 432, and N = 1224. See also Supplementary Figure 6 for corresponding distributions of motif effects on enrichment in the N50-C library, and Supplementary Figure 10 for distributions of motif effects on half-life in native yeast mRNAs (based on [30])

native context. Alternatively, these Puf sites might affect mRNA stability and translation differentially; such differential effects might be enabled by additional roles these motifs play besides Puf protein binding. The presence of a Puf6 binding site element was associated with only a weak nominal reduction in native mRNA half-life, which was not statistically significant (Fig. 8i, green), and a weak (but statistically significant) increase in His3 protein expression (Fig. 8i, blue). These minimal effects may reflect the nature of Puf6 regulation, with known target genes displaying multiple Puf6 binding sites in their 3′ UTRs [50, 51].

Finally, the poly(U) motif U_8 gave strikingly different results for native mRNA half-life and His3 protein expression. Among native yeast genes the presence of a U_8 sequence was associated with a longer half-life (Fig. 8j, green), consistent with Geisberg et al. [30], and with the analysis of U_6A by Cheng et al. [34]. In contrast, as noted, U_8 had no effect on protein expression beyond its AU content (Fig. 8j, blue).

Discussion

Taken together, our results indicate the importance of context in determining the expression consequences of 3′ UTR sequence features. The efficiency element emerges as a robust, context-independent regulatory sequence, with its 6-mer consensus sequence providing the largest increase in expression of any hexamer. Similarly, Puf3, Puf4, and Puf5 binding site motifs enhanced protein expression in a random context. These results suggest that an optimal efficiency element can be added to the 3′ UTR of any exogenous sequence of interest lacking this feature to increase the resultant protein expression level in yeast. Puf4 or Puf5 protein binding sites could similarly be added, although serendipitous sequence features might convert these into repressive factors; buffering the Puf motifs with surrounding random sequence might prevent this conversion. Adding AU-rich elements should also generically improve gene expression. However, the positioning element and poly(U) motifs do not display this same degree of generalizability.

As exemplified by the results for GCGCGC, the Puf binding sites and U_8 (Fig. 8), the average effects of sequence elements on native mRNA stability did not generally agree with measurements of their expression effects in a random context. This lack of concordance is presumably influenced by two important factors: first, the role of evolved sequence context in modulating 3′ UTR motif function, and second, the lack of equivalence between effects on RNA level (via mRNA stability) and protein level, consistent with literature demonstrating a lack of correlation between protein and mRNA levels [60–63]. A number of factors may contribute to this regulatory complexity: multiple proteins interacting with a motif (e.g., the Puf3 site; Fig. 8f), interactions between multiple motifs (e.g., the efficiency and positioning elements, Fig. 5c), position-dependent effects of motifs (e.g., Fig. 3b), and the effects of motifs in a random sequence context (e.g., Puf protein binding sites, Fig. 6). Furthermore, sequences such as poly(U) elements and Puf protein binding motifs may affect translation in a manner distinct from mRNA stability. A dissection of the detailed interplay between these factors at both the RNA and protein levels should be a fruitful direction for efforts to decipher the underlying regulatory grammar of the 3′ untranslated region.

Conclusions

The 3′ UTRs of mRNAs contain sequence features that regulate activities such as cleavage and polyadenylation, translation, stability, and localization. By assaying hundreds of thousands of random 3′ UTR sequences for their effects on protein expression in yeast and comparing with previous measurements of native 3′ UTR effects, we find that some of these features function similarly in any sequence context, whereas others—in particular, several Puf protein binding sites—have effects that appear to depend on other, co-evolved sequence features within natural mRNAs.

Methods

Construction of the N50-EPC and N50-C 3′ UTR libraries

We replaced the *CYC1* 3′ UTR sequence downstream of the *HIS3* stop codon on a p415-CYC1 plasmid (carrying a *LEU2* selection marker for growth on media lacking leucine) [64] with libraries of 50-bp synthetic 3′ UTR fragments. The *CYC1* terminator is relatively short (253 bp), with well-established efficiency, positioning, and cleavage sites. In the N50-EPC library, the first 102 bp were replaced with the N50 element, preserving the efficiency, positioning, and cleavage elements, while in the N50-C library, the first 151 bp were replaced with the N50 element, preserving the cleavage site. The p415-CYC1-HIS3 plasmid was linearized by inverse PCR using KAPA HiFi polymerase (Kapa Biosystems) with primers F-p415-His and R-p415-HIS (oligonucleotide sequences in Additional File 1: Table S1), which remove the first 172 bp of the native *CYC1* 3′ UTR. Template DNA was digested using DpnI, and the PCR product was isolated using a DNA Clean and Concentrate Kit (Zymo Research).

The synthetic 3′ UTR fragments were constructed from Ultramer oligonucleotides (Integrated DNA Technologies) to comprise the N50-EPC or N50-C library. The oligonucleotides encoded the N50 element and either the efficiency, positioning and cleavage elements, or the cleavage element. Each also encoded 20 bp of *CYC1* 3′ UTR sequence downstream of the cleavage site, as well as 30 bp of homology to the linearized backbone on both the 5′ and 3′ ends, for cloning by Gibson assembly [65]. The oligonucleotides were used as PCR templates and amplified by six rounds of PCR using KAPA HiFi polymerase (Kapa Biosystems) and primers F_N50_lib and R_N50_lib. We limited the cycles of PCR amplification to maintain sequence diversity in the libraries. After amplification, the PCR product was isolated using a DNA Clean and Concentrate Kit (Zymo Research).

The final libraries were assembled using Gibson assembly [65]. Briefly, four 20 μL reactions each containing 100 fmol of plasmid backbone, 200 fmol of 3′ UTR library, and 10 μL of NEB HiFi Builder 2× master mix were incubated at 50 °C for 1 h. Reactions were pooled and isolated using a DNA Clean and Concentrate Kit (Zymo Research), and samples were used to transform by electroporation 40 μL of Electromax DH10B *E. coli* (Agilent). Dilutions of 1:1000 and 1:10,000 were plated on LB agar plates supplemented with 100 μg/mL ampicillin to estimate the number of unique transformants in each library. The N50-EPC library contained approximately 4×10^6 transformants, and the N50-C library contained approximately 3.4×10^6 transformants. The remaining cells transformed with library were shaken overnight at 37 °C in LB media supplemented with 100 μg/mL ampicillin, and the plasmid library was isolated using a miniprep kit (Qiagen).

Yeast transformation

The N50-EPC and N50-C libraries were transformed into the leucine auxotrophic strain BY4741 *his3*::*KanMX* from the yeast deletion collection. The strain was struck out from a frozen glycerol stock onto YEPD plates supplemented with 200 μg/mL G418, and its auxotrophies subsequently verified by the strain's requirement for leucine and histidine for growth in SD media. Yeast were transformed using a high-efficiency yeast transformation protocol [66]. Briefly, 5 mL of culture was grown overnight at 30 °C in YEPD. The saturated culture was back-diluted into 50 mL of fresh 2× YEPD to an approximate OD_{660} of 0.1. Cultures were grown at 30 °C for approximately 6 h, until the OD_{660} reached approximately 1.0. Cells were pelleted, resuspended in 10 mL of water, split into ten separate microcentrifuge tubes, and pelleted again. Cells in each tube were resuspended in 36 μL of 1 M LiAC, 240 μL of 50% w/v PEG 3350, 50 μL of 2 mg/mL salmon sperm carrier DNA that had been denatured by boiling and 200 ng of plasmid miniprep in 36 μL of water. Tubes were transferred to a 42 °C water bath and incubated for 40 min. Cells were pelleted, resuspended in 1 mL of water, combined into a single tube, and dilutions of 1:1,000 and 1:10,000 were plated on SD-Leu agar plates and grown 48 h at 30 °C to estimate the number of unique transformants. The remaining cells were diluted into 200 mL of SD-Leu media and grown overnight with shaking at 30 °C. Aliquots of 10 mL of culture were pelleted, resuspended in 1 mL of SD-Leu, mixed with 300 μL of 50% glycerol and stored at − 80 °C.

Growth curve experiments

For each of the N50-EPC and N50-C libraries, 45 random colonies transformed with the library and three colonies transformed with a reporter plasmid with the *CYC1* terminator were used to inoculate 200 μL of SD-Leu media in a 96-well plate. The colonies were shaken overnight at 30 °C in a Biotek Synergy H1 plate reader. Two microliters of each saturated culture was used to inoculate 200 μL of SD-Leu-His media supplemented 0, 1, 3, or 5 mM 3-AT and shaken for 48 h at 30 °C in a Biotek Synergy H1 plate reader, with OD_{660} measured every 15 min. The maximum growth rate for each random library member was determined by calculating the most rapid increase in OD_{660}.

Massively parallel growth selection assay for His3 expression

Glycerol stocks of each library stored at − 80 °C were thawed and used to inoculate 100 mL of SD-Leu media. Cultures were grown overnight at 30 °C, and 5 mL of each culture was stored at 4 °C to serve as the input sample for the selection. The OD_{660} of each library was measured and approximately 2×10^8 cells were used to inoculate 100 mL of SD-Leu-His media supplemented with 1 mM 3-AT. Each culture was shaken at 30 °C until the OD_{660} measured approximately 1.0. (~ 24 h for the N50-EPC library and ~ 30 h for the N50-C library). Five milliliters of post-selection culture was stored, and plasmids from both before and after selection were isolated using the Yeast Plasmid Miniprep II Kit (Zymo Research).

A single massively parallel growth selection assay was performed for each library (N50-C and N50-EPC). Due to the very large size and completely random character of the sequence variant libraries, true biological replicates were not feasible to perform; in particular, a different set of random 50-mer sequences would be selected during any replicate transformation (or cloning step). Lack of replicates for these types of experiments has precedent in other random sequence-based high-throughput assays, e.g., [1–3, 5].

However, an alternative measure of reproducibility for these measurements can be used. Analyses of the effects of specific sequence motifs in a random sequence context across many—typically thousands—of diverse N50 sequences carrying a given element (rather than the specific enrichment of any single library sequence variant) provides an approximation of a generic random sequence background for each element of interest. Reproducibility of the effects of sequence features could thus be determined from the mean ± s.e.m. effects of each sequence element across the library, as reported throughout the figures (see also the Source Data [41]).

Preparation of sequencing libraries

Sequencing libraries were prepared as 225 bp amplicons containing the 3′ UTR libraries. Plasmids isolated before and after selection were amplified by 12–16 cycles of PCR using primers that contained Illumina adapter sequences and unique sequencing indices. PCR products were isolated using a DNA Clean and Concentrate Kit (Zymo Research) and quantified using a Qubit fluorometer. The sequencing libraries were diluted to 2 nM and denatured for sequencing following the standard Illumina protocol. DNA sequencing was performed on an Illumina Nextseq 550 instrument sequencer. To identify the set of sequences in our library, we made use of the program Bartender, which collapses similar sequences into a set of consensus sequences [67]. We ran Bartender using the following options: -t 40 -d 8 -z -1 -c 1 -l 8. This set of consensus sequences was used in all subsequent analyses, with alignments performed to these sequences using Bowtie2 [68].

Analysis of the effects of sequences in the 3′ UTR on gene expression in the N50-C and N50-EPC libraries

The growth selection results were filtered to improve confidence in the estimates of variant frequencies, as follows: only 3′ UTR variants with at least 5 reads in the input sample and at least 1 read in the output sample were subjected to further analyses (no pseudocounting was employed). This filtering yielded the analyzed library sizes of ~ 590,000 for N50-C and ~ 280,000 for N50-EPC.

For random *k*-mer sequences, a custom script was used to generate a list of all possible hexamer RNA sequences and then to determine the mean enrichment (and its standard error) for the subset of library sequences containing each hexamer in the N50 sequence. The standard error of the mean was calculated as s.e.m. = σ / $(N)^{1/2}$, where σ is the standard deviation and N is the number of variants for the given subset of sequences. The same calculations were performed for the subset of library sequences lacking each hexamer in the N50 sequence. Such calculations were performed both for the N50-C and the N50-EPC library. The resulting lists of 6-mer sequences and associated average enrichments were then sorted by enrichment of sequences containing each 6-mer to determine hexamer ranking in each library.

Analysis using kpLogo [42] was performed with sequences weighted by their enrichment scores and searching for *k*-mers of lengths 1–6. Enrichment and depletion of bases was determined using a one-sided two-tailed Student's *t* test and considered significant if the Bonferroni-corrected *p* value was < 0.05. Motif analysis was performed using the default settings on STREME [43], evaluating sequences > 2.5 standard deviations above or below the mean for motifs enriched or depleted from the libraries. Tomtom [44] was used to search these motifs against the motifs of known RNA-binding proteins.

For known 3′ UTR elements, the average effects of specific sequence elements on growth selection enrichment were calculated by using a custom script to determine mean enrichment of the subset of library sequences containing the sequence element(s) in question in the N50 sequence, making use of a string search for each element across the N50 sequences in the library. The average of sequence elements located at a specific position in the N50 were calculated as follows. FIMO [69] was used to determine the locations of all instances of a perfect match to the motif of interest in the library. Locations of each shuffled form of each motif of interest were determined in the same manner. FIMO was run with a uniform background and a *p* value threshold set at just above the expected probability of the motif in question emerging at random (e.g., $(0.25)^6$ for UAUAUA). The positionally segregated average effects of each motif were determined by using a custom script to determine mean enrichment of the subset of library sequences containing the sequence element in question with the motif 5′ end ("start" sequence output from FIMO) located at each position in the N50 sequence. These analyses were also performed with shuffled sequences derived from motifs of interest. Shuffled sequences were generated using a custom Python script. Output shuffled sequences were filtered for the criterion that they be a Hamming distance of at least half the motif length away from the starting sequence (e.g., Hamming distance of 3 for the motif UAUAUA) unless otherwise noted. The number of shuffles considered for each sequence element was as many as possible matching the above criteria, up to a maximum of 50 shuffles, unless otherwise noted.

Analysis of mRNA half-life effects of sequence elements in native yeast gene 3′ UTRs

The stability of native yeast mRNA transcripts has been described in several datasets (e.g., [27–30, 70]). We chose to compare our relative protein expression data to mRNA half-life data generated using the "anchor away" approach to stop transcription combined with a direct RNA sequencing approach [30]. We matched the mRNA sequence from the S288C reference genome with its corresponding isoform using the gene name and 3′ UTR length. Because the relative abundance of each isoform detected in that study is not reported, we used the 3′ UTR isoform with half-life nearest to the reported mean half-life of the gene as the representative 3′ UTR sequence, to hopefully avoid low abundance transcripts in our analyses. This procedure resulted in a list of 3547 representative 3′ UTR isoforms (one per gene) and their associated half-lives.

Comparison of sequence motif effects on native gene mRNA half-life vs. random library protein expression

We compared the consequences of a number of sequence elements on relative protein expression (growth selection enrichment score) in the N50-C 3′ UTR library to the consequences of these same elements on mRNA half-life across native 3′ UTRs in *S. cerevisiae*, based on the dataset [21] described in the previous section. Relative half-lives associated with each 3′ UTR (and the associated mRNAs) were calculated as $(\lambda(3'\text{ UTR}) - <\lambda>) / <\lambda>$, where λ denotes half-life and $<\lambda>$ denotes the average half-life across all genes in the data set. Similarly, relative enrichments were calculated as $(Enr_{(\mathrm{norm})} - <Enr_{(\mathrm{norm})}>) / <Enr_{(\mathrm{norm})}>$, where $Enr_{(\mathrm{norm})}$ is the normalized enrichment in the growth selection and $<Enr_{(\mathrm{norm})}>$ is the average normalized enrichment across all sequences in the N50-C library. The normalized enrichment $Enr_{(\mathrm{norm})}$ was calculated as $Enr - Enr_{(\mathrm{min})}$, where $Enr_{(\mathrm{min})}$ is the lowest enrichment among all sequences in the library. Normalized enrichment was used in the relative enrichment calculations to produce a quantity that is always positive.

Average effects of specific sequence elements on relative mRNA half-life were calculated by using a custom Python script to determine mean relative half-life of the subset of native gene 3′ UTR sequences containing the sequence element(s) in question, using a string search of the UTR sequences for each motif of interest. Similarly, the average effects of specific sequence elements on relative enrichment in the growth selection were calculated by using the same custom script to determine mean relative enrichment of the subset of N50-C or N50-EPC library 3′ UTR sequences containing the sequence element(s) in question.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02509-6.

Additional File 1:. Supplementary Figures and Tables

Additional File 2:. Peer review history

Acknowledgements

We thank members of the Fields lab for helpful discussions.

Review history

The review history is available as Additional file 2.

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

B.M.B., B.E.A., J.C., and A.S. performed experiments. A.S., B.M.B., J.C., and B.E.A. analyzed the data. A.S. and S.F. wrote the manuscript with input from B.M.B., J.C., and B.E.A. All authors read and approved the final manuscript.

Funding

This work was supported by grant 1R01 GM125809 from the National Institutes of Health.

Availability of data and materials

Source Data (including the growth selection enrichments associated with each library sequence, *k*-mer, and motif, and their associated s.e.m., as well as other pertinent information such as the counts of variants carrying each motif) are available via figshare (URL: https://figshare.com/articles/dataset/Source_Data_for_Savinov_et_al_2021_3_UTRs/16664143; doi: 10.6084/m9.figshare.16664143) [41]. The sequencing data are available via the NIH Sequence Read Archive (BioProject ID PRJNA750726) [71]. Analysis scripts are available via GitHub (https://github.com/andrewsavinov/Savinov-et-al-2021_3primeUTRs) and Zenodo (doi: 10.5281/zenodo.5149781) [72].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Genome Sciences, University of Washington, Box 355065, Seattle, WA 98195, USA. [2]Present address: Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA. [3]Department of Chemistry and Biochemistry, Creighton University, Omaha, NE 68178, USA. [4]Present address: Interdisciplinary Biological Sciences Graduate Program, Northwestern University, Evanston, IL 60208, USA. [5]Department of Medicine, University of Washington, Box 357720, Seattle, WA 98195, USA.

Received: 26 March 2021 Accepted: 30 September 2021
Published online: 18 October 2021

References

1. Cuperus JT, Groves B, Kuchina A, Rosenberg AB, Jojic N, Fields S, et al. Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. Genome Res. 2017;27(12):2015–24. https://doi.org/10.1101/gr.224964.117.
2. Sample PJ, Wang B, Reid DW, Presnyak V, McFadyen IJ, Morris DR, et al. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. Nat Biotechnol. 2019;37(7):803–9. https://doi.org/10.1038/s41587-019-0164-5.
3. de Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nat Biotechnol. 2020;38(1):56–65. https://doi.org/10.1038/s41587-019-0315-8.
4. Ireland WT, Beeler SM, Flores-Bautista E, McCarty NS, Röschinger T, Belliveau NM, et al. Deciphering the regulatory genome of Escherichia coli, one hundred promoters at a time. Elife. 2020;9:1–76. https://doi.org/10.7554/eLife.55308.
5. Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. Learning the sequence determinants of alternative splicing from millions of random sequences. Cell. 2015;163(3):698–711. https://doi.org/10.1016/j.cell.2015.09.054.
6. Mayr C. Regulation by 3′-untranslated regions. Annu Rev Genet. 2017;51(1):171–94. https://doi.org/10.1146/annurev-genet-120116-024704.
7. Shalem O, Carey L, Zeevi D, Sharon E, Keren L, Weinberger A, et al. Measurements of the impact of 3′ end sequences on gene expression reveal wide range and sequence dependent effects. PLoS Comput Biol. 2013;9(3):e1002934. https://doi.org/10.1371/journal.pcbi.1002934.

8. Shalem O, Sharon E, Lubliner S, Regev I, Lotan-Pompan M, Yakhini Z, et al. Systematic dissection of the sequence determinants of gene 3’ end mediated expression control. PLoS Genet. 2015;11(4):e1005147. https://doi.org/10.1371/journal.pgen.1005147.
9. Zhao W, Pollack JL, Blagev DP, Zaitlen N, McManus MT, Erle DJ. Massively parallel functional annotation of 3’ untranslated regions. Nat Biotechnol. 2014;32(4):387–91. https://doi.org/10.1038/nbt.2851.
10. Oikonomou P, Goodarzi H, Tavazoie S. Systematic identification of regulatory elements in conserved 3’ UTRs of human transcripts. Cell Rep. 2014;7(1):281–92. https://doi.org/10.1016/j.celrep.2014.03.001.
11. Vainberg Slutskin I, Weingarten-Gabbay S, Nir R, Weinberger A, Segal E. Unraveling the determinants of microRNA mediated regulation using a massively parallel reporter assay. Nat Commun. 2018;9. https://doi.org/10.1038/s41467-018-02980-z.
12. Litterman AJ, Kageyama R, Le Tonqueze O, Zhao W, Gagnon JD, Goodarzi H, et al. A massively parallel 3′ UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. Genome Res. 2019;29(6):896–906. https://doi.org/10.1101/gr.242552.118.
13. Siegel D, Le Tonqueze O, Biton A, Zaitlen N, Erle D. Massively parallel analysis of human 3′ UTRs reveals that AU-rich element length and registration predict mRNA destabilization. bioRxiv. 2020. https://doi.org/10.1101/2020.02.12.945063.
14. Rabani M, Pieper L, Chew GL, Schier AF. A massively parallel reporter assay of 3′ UTR sequences identifies in vivo rules for mRNA degradation. Mol Cell. 2017;68:1083–1094.e5 https://doi.org/10.1016/j.molcel.2017.11.014.
15. Slutskin IV, Weinberger A, Segal E. Sequence determinants of polyadenylation-mediated regulation. Genome Res. 2019;29(10):1635–47. https://doi.org/10.1101/gr.247312.118.
16. Bogard N, Linder J, Rosenberg AB, Seelig G. A deep neural network for predicting and engineering alternative polyadenylation. Cell. 2019;178:91–106.e23. https://doi.org/10.1016/j.cell.2019.04.046.
17. Bennetzen JL, Hall BD. The primary structure of the Saccharomyces cerevisiae gene for alcohol dehydrogenase. J Biol Chem. 1982;257(6):3018–25. https://doi.org/10.1016/S0021-9258(19)81067-0.
18. Abe A, Hiraoka Y, Fukasawa T. Signal sequence for generation of mRNA 3’ end in the Saccharomyces cerevisiae GAL7 gene. EMBO J. 1990;9(11):3691–7. https://doi.org/10.1002/j.1460-2075.1990.tb07581.x.
19. Heidmann S, Obermaier B, Vogel K, Domdey H. Identification of pre-mRNA polyadenylation sites in Saccharomyces cerevisiae. Mol Cell Biol. 1992;12(9):4215–29. https://doi.org/10.1128/mcb.12.9.4215-4229.1992.
20. Russo P, Li WZ, Guo Z, Sherman F. Signals that produce 3’ termini in CYC1 mRNA of the yeast Saccharomyces cerevisiae. Mol Cell Biol. 1993;13(12):7836–49. https://doi.org/10.1128/MCB.13.12.7836.
21. Guo Z, Sherman F. 3’-end-forming signals of yeast mRNA. Mol Cell Biol. 1995;15(11):5983–90. https://doi.org/10.1128/MCB.15.11.5983.
22. Guo Z, Sherman F. Signals sufficient for 3’-end formation of yeast mRNA. Mol Cell Biol. 1996;16(6):2772–6. https://doi.org/10.1128/MCB.16.6.2772.
23. Kessler MM, Henry MF, Shen E, Zhao J, Gross S, Silver PA, et al. Hrp1, a sequence-specific RNA-binding protein that shuttles between the nucleus and the cytoplasm, is required for mRNA 3’-end formation in yeast. Genes Dev. 1997;11(19):2545–56. https://doi.org/10.1101/gad.11.19.2545.
24. Pérez-Cãadillas JM. Grabbing the message: structural basis of mRNA 3′UTR recognition by Hrp1. EMBO J. 2006;25(13):3167–78. https://doi.org/10.1038/sj.emboj.7601190.
25. Gross S, Moore C. Five subunits are required for reconstitution of the cleavage and polyadenylation activities of Saccharomyces cerevisiae cleavage factor I. Proc Natl Acad Sci U S A. 2001;98(11):6080–5. https://doi.org/10.1073/pnas.101046598.
26. Gross S, Moore CL. Rna15 interaction with the A-Rich yeast polyadenylation signal is an essential step in mRNA 3′-end formation. Mol Cell Biol. 2001;21(23):8045–55. https://doi.org/10.1128/MCB.21.23.8045-8055.2001.
27. Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO. Precision and functional specificity in mRNA decay. Proc Natl Acad Sci U S A. 2002;99(9):5860–5. https://doi.org/10.1073/pnas.092538799.
28. Grigull J, Mnaimneh S, Pootoolal J, Robinson MD, Hughes TR. Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. Mol Cell Biol. 2004;24(12):5534–47. https://doi.org/10.1128/MCB.24.12.5534-5547.2004.
29. Miller C, Schwalb B, Maier K, Schulz D, Dümcke S, Zacher B, et al. Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. Mol Syst Biol. 2011;7(1):458. https://doi.org/10.1038/msb.2010.112.
30. Geisberg JV, Moqtaderi Z, Fan X, Ozsolak F, Struhl K. Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. Cell. 2014;156(4):812–24. https://doi.org/10.1016/j.cell.2013.12.026.
31. Gerber AP, Herschlag D, Brown PO. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. PLoS Biol. 2004;2(3):0342–54. https://doi.org/10.1371/journal.pbio.0020079.
32. 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 Biol. 2008;6(10):2297–313. https://doi.org/10.1371/journal.pbio.0060255.
33. Hasan A, Cotobal C, Duncan CDS, Mata J. Systematic analysis of the role of RNA-binding proteins in the regulation of RNA stability. PLoS Genet. 2014;10(11):e1004684. https://doi.org/10.1371/journal.pgen.1004684.
34. Cheng J, Maier KC, Avsec Ž, Petra RUS, Gagneur J. Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast. RNA. 2017;23(11):1648–59. https://doi.org/10.1261/rna.062224.117.
35. Quenault T, Lithgow T, Traven A. PUF proteins: Repression, activation and mRNA localization. Trends Cell Biol. 2011;21(2):104–12. https://doi.org/10.1016/j.tcb.2010.09.013.
36. Wang M, Ogé L, Perez-Garcia MD, Hamama L, Sakr S. The PUF protein family: Overview on PUF RNA targets, biological functions, and post transcriptional regulation. Int J Mol Sci. 2018;19(2):410. https://doi.org/10.3390/ijms19020410.
37. Olivas W, Parker R. The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast. EMBO J. 2000;19(23):6602–11. https://doi.org/10.1093/emboj/19.23.6602.
38. Der Lee C, Tu BP. Glucose-regulated phosphorylation of the PUF protein Puf3 regulates the translational fate of its bound mRNAs and association with RNA granules. Cell Rep. 2015;11(10):1638–50. https://doi.org/10.1016/j.celrep.2015.05.014.

39. Guy MP, Young DL, Payea MJ, Zhang X, Kon Y, Dean KM, Grayhack EJ, Mathews DH, Fields S, Phizicky EM Identification of the determinants of tRNA function and susceptibility to rapid tRNA decay by high-throughput in vivo analysis. Genes Dev. Cold Spring Harbor Laboratory Press; 2014;28:1721–32. https://doi.org/10.1101/gad.245936.114.
40. Gamble CE, Brule CE, Dean KM, Fields S, Grayhack EJ. Adjacent codons act in concert to modulate translation efficiency in yeast. Cell. 2016;166(3):679–90. https://doi.org/10.1016/j.cell.2016.05.070.
41. Savinov A, Brandsen BM, Angell BE, Cuperus JT, Fields S. Repository of source data. figshare. 2021. https://figshare.com/articles/dataset/Source_Data_for_Savinov_et_al_2021_3_UTRs/16664143.
42. Wu X, Bartel DP. KpLogo: Positional k -mer analysis reveals hidden specificity in biological sequences. Nucleic Acids Res. 2017;45(W1):W534–8. https://doi.org/10.1093/nar/gkx323.
43. Bailey TL. STREME: accurate and versatile sequence motif discovery. Bioinformatics. 2021;37(18):2834–40. https://doi.org/10.1093/bioinformatics/btab203.
44. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8(2): R24. https://doi.org/10.1186/gb-2007-8-2-r24.
45. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013;499:172–7.
46. Yosefzon Y, Koh YY, Chritton JJ, Lande A, Leibovich L, Barziv L, et al. Divergent RNA binding specificity of yeast Puf2p. RNA. 2011;17(8):1479–88. https://doi.org/10.1261/rna.2700311.
47. Webster MW, Stowell JA, Passmore LA. RNA-binding proteins distinguish between similar sequence motifs to promote targeted deadenylation by Ccr4-Not. Elife. 2019;8:e40670. https://doi.org/10.7554/eLife.40670.
48. Russo J, Olivas WM. Conditional regulation of Puf1p, Puf4p, and Puf5p activity alters YHB1 mRNA stability for a rapid response to toxic nitric oxide stress in yeast. Mol Biol Cell. 2015;26(6):1015–29. https://doi.org/10.1091/mbc.E14-10-1452.
49. Kalem MC, Subbiah H, Leipheimer J, Glazier VE, Panepinto JC. Puf4 mediates post-transcriptional regulation of cell wall biosynthesis and caspofungin resistance in cryptococcus neoformans. MBio. 2021;12(1):1–20. https://doi.org/10.1128/mBio.03225-20.
50. Gu W, Deng Y, Zenklusen D, Singer RH. A new yeast PUF family protein, Puf6p, represses ASH1 mRNA translation and is required for its localization. Genes Dev. 2004;18(12):1452–65. https://doi.org/10.1101/gad.1189004.
51. Jung D, Seo JS, Nam J, Kim J. Functional association of Loc1 and Puf6 with RNA helicase Dhh1 in translational regulation of Saccharomyces cerevisiae Ste12. PLoS One. 2019;14(7):e0220137. https://doi.org/10.1371/journal.pone.0220137.
52. Valley CT, Porter DF, Qiu C, Campbell ZT, Tanaka Hall TM, Wickens M. Patterns and plasticity in RNA-protein interactions enable recruitment of multiple proteins through a single site. Proc Natl Acad Sci U S A. 2012;109(16):6054–9. https://doi.org/10.1073/pnas.1200521109.
53. Crucs S, Chatterjee S, Gavis ER. Overlapping but distinct RNA elements control repression and activation of nanos translation. Mol Cell. 2000;5(3):457–67. https://doi.org/10.1016/S1097-2765(00)80440-2.
54. Winter J, Roepcke S, Krause S, Müller EC, Otto A, Vingron M, et al. Comparative 3'UTR analysis allows identification of regulatory clusters that drive Eph/ephrin expression in cancer cell lines. PLoS One. 2008;3(7):e2780. https://doi.org/10.1371/journal.pone.0002780.
55. Taliaferro JM, Lambert NJ, Sudmant PH, Dominguez D, Merkin JJ, Alexis MS, et al. RNA sequence context effects measured in vitro predict in vivo protein binding and regulation. Mol Cell. 2016;64(2):294–306. https://doi.org/10.1016/j.molcel.2016.08.035.
56. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet. 2014;15(12):829–45. https://doi.org/10.1038/nrg3813.
57. Lebedeva S, Jens M, Theil K, Schwanhäusser B, Selbach M, Landthaler M, et al. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. Mol Cell. 2011;43(3):340–52. https://doi.org/10.1016/j.molcel.2011.06.008.
58. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods. 2016;13(6):508–14. https://doi.org/10.1038/nmeth.3810.
59. Fischer AD, Olivas WM. Multiple Puf proteins regulate the stability of ribosome biogenesis transcripts. RNA Biol. 2018; 15(9):1228–43. https://doi.org/10.1080/15476286.2018.1521211.
60. Fortelny N, Overall CM, Pavlidis P, Freue GVC. Can we predict protein from mRNA levels? Nature. 2017. p. E19–20. https://doi.org/10.1038/nature22293.
61. Buccitelli C, Selbach M. mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet. 2020; 21(10):630–44. https://doi.org/10.1038/s41576-020-0258-4.
62. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell. 2016;165(3):535–50. https://doi.org/10.1016/j.cell.2016.03.014.
63. Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, Hearn J, et al. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science. 2010;329:533–8.
64. Mumberg D, Müller R, Funk M. Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. Gene. 1995;156(1):119–22. https://doi.org/10.1016/0378-1119(95)00037-7.
65. Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods. 2009;6(5):343–5. https://doi.org/10.1038/nmeth.1318.
66. Gietz RD, Schiestl RH. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat Protoc. 2007; 2(1):31–4. https://doi.org/10.1038/nprot.2007.13.
67. Zhao L, Liu Z, Levy SF, Wu S. Bartender: a fast and accurate clustering algorithm to count barcode reads. Bioinformatics. 2018;34(5):739–47. https://doi.org/10.1093/bioinformatics/btx655.
68. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923.
69. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27(7):1017–8. https://doi.org/10.1093/bioinformatics/btr064.

70. Holstege FCP, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, et al. Dissecting the regulatory circuitry of a eukaryotic genome. Cell. 1998;95(5):717–28. https://doi.org/10.1016/S0092-8674(00)81641-4.
71. Savinov A, Brandsen BM, Angell BE, Cuperus JT, Fields S. Read data. Sequence Read Archive. BioProject ID PRJNA750726. 2021.
72. Savinov A, Brandsen BM, Angell BE, Cuperus JT, Fields S. Repository of analysis code. GitHub. 2021. https://github.com/andrewsavinov/Savinov-et-al-2021_3primeUTRs (2021).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.