Nucleic Acids Research, 2025, **53**, gkaf165
https://doi.org/10.1093/nar/gkaf165
RNA and RNA-protein complexes

Deciphering the landscape of *cis*-acting sequences in natural yeast transcript leaders

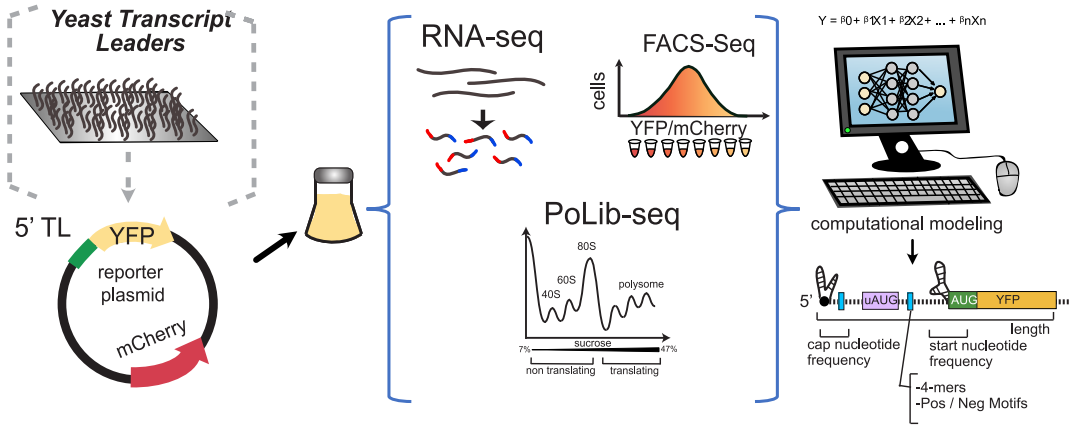
Christina Akirtava[1,2], Gemma E. May[1], C.Joel McManus[1,3,*]

[1]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, United States
[2]RNA Bioscience Initiative, University of Colorado – Anschutz, Aurora, CO 80045, United States
[3]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, United States
[*]To whom correspondence should be addressed. Email: mcmanus@andrew.cmu.edu

Abstract

Protein synthesis is a vital process that is highly regulated at the initiation step of translation. Eukaryotic 5′ transcript leaders (TLs) contain a variety of *cis*-acting features that influence translation and messenger RNA stability. However, the relative influences of these features in natural TLs are poorly characterized. To address this, we used massively parallel reporter assays (MPRAs) to quantify RNA levels, ribosome loading, and protein levels from 11,027 natural yeast TLs *in vivo* and systematically compared the relative impacts of their sequence features on gene expression. We found that yeast TLs influence gene expression over two orders of magnitude. While a leaky scanning model using Kozak contexts (−4 to +1 around the AUG start) and upstream AUGs (uAUGs) explained half of the variance in expression across TLs, the addition of other features explained ∼80% of gene expression variation. Our analyses detected key *cis*-acting sequence features, quantified their effects *in vivo*, and compared their roles to motifs reported from an *in vitro* study of ribosome recruitment. In addition, our work quantitated the effects of alternative transcription start site usage on gene expression in yeast. Thus, our study provides new quantitative insights into the roles of TL *cis*-acting sequences in regulating gene expression.

Graphical abstract



Introduction

Protein synthesis via messenger RNA (mRNA) translation is an essential process in gene expression. mRNA translation is divided into initiation, elongation, termination, and ribosome recycling steps. In rapidly growing cells, translation is largely limited by initiation, which occurs primarily through cap-dependent directional scanning. Specifically, the pre-initiation complex (PIC), which consists of the 40S ribosomal subunit and multiple initiation factors, must scan the transcript leader (TL) to locate the main coding sequence (CDS) [1]. As such, sequence elements within TLs are important contributors to initiation efficiency and gene expression. Previous work demonstrated that features such as upstream open reading frames (uORFs), mRNA structures, and mRNA-binding protein motifs influence the identification of the main start codon by the PIC [1–6]. These features have been studied largely independently, such that their relative importance in regulating gene expression from native TLs has not been systematically evaluated *in vivo*.

The Kozak context, a consensus sequence found from –6 to +1 relative to the start codon, was first identified in early studies as a key factor in translation initiation [7]. When a PIC

Received: July 12, 2024. **Revised:** February 16, 2025. **Editorial Decision:** February 17, 2025. **Accepted:** February 20, 2025

encounters a start codon in optimal Kozak context, initiation factors reorganize into a "closed" formation to initiate translation [8, 9]. In contrast, PICs presented with start codons in a sub-optimal Kozak context are more likely to remain in an "open" scanning conformation and skip initiation. The process by which the PIC bypasses an upstream start codon to initiate translation downstream is described as the leaky scanning model (LSM). Previously, a massively parallel reporter assay (MPRA) measuring translation of 2041 TL variants of the yeast RPL8A gene found that variation in the Kozak context resulted in a ∼7-fold difference in protein expression, with a preference for A-rich sequences [10]. Additional experiments confirmed this and highlighted the positive influence of adenosine at the −3 position [11–14]. Other work has shown translation can also initiate at near-AUG codons [15–22], though this appears to be generally inefficient in yeast [14, 23–27]. Though the yeast Kozak motif itself is well defined, the relative importance of Kozak sequences in the context of other elements in native yeast TLs *in vivo* remains unstudied. For example, other studies have found mRNA secondary structure can prevent start codon recognition [28, 29]. Thus, the extent to which leaky scanning and the Kozak context alone control translation from natural TLs remains unknown.

Alternative transcription start sites (aTSSs) add another layer of complexity to the function of mRNA TLs. Several studies have shown that many genes initiate transcription at multiple sites [5, 30, 31]. *In vitro* measurements of 96 native yeast TLs using luciferase assays, revealed that alternative transcript initiation sites that differed by 50–200 nt in 5′ ends could influence translation efficiency 100-fold [32], and it is conceivable that some alternative TLs could have even larger effects. Large changes in expression of alternative TLs suggest that sequence features play a crucial role in this regulation. For example, aTSSs could alter the number and identify of uORFs, RNA structures, and RNA-binding protein sites in TLs. However, the effect of such differences in natural TLs on protein expression has not been systematically evaluated.

While several foundational studies analyzed gene expression from TLs, they did not evaluate the relative impacts of sequence features in full-length native TLs *in vivo*. Studies using libraries containing a fixed sequence at the 5′ end may not capture the effects of structure or sequence preference near the cap. The fixed length, synthetic, and randomized TLs previously tested *in vivo* [10, 11, 13] were relatively short, and do not explain larger length differences or naturally occurring motifs present in native TLs. A recent study compared the recruitment of yeast ribosomes to native TLs *in vitro*, and reported several motifs associated with high and low ribosome recruitment [12]. However, although useful, translation extracts do not fully reproduce the cellular environment. Furthermore, this approach required the removal of all upstream AUGs (uORFs and N-terminal extensions), which are large contributors to gene regulation by native TLs. A systematic study *of in vivo* regulation by natural TLs is needed to more fully evaluate TL impacts on gene expression.

Here, we assayed protein expression, mRNA levels, and ribosome loading from a comprehensive library of endogenous yeast TLs and determined the relative impacts of TL features. We first compared expression levels from hundreds of genes containing alternative TLs and found that even slight changes in TL length greatly impacted expression. Next, we evaluated the extent to which Kozak sequence strengths explain gene expression variance from *Saccharomyces cerevisiae* TLs. Using start codon-Kozak pair strengths, we built an LSM which more accurately explained ribosome scanning and expression. Finally, we defined and calculated strengths of additional *cis*-acting sequence elements shown to influence expression. By combining all features in a computational model, we explain ∼80% of the variance in expression across native TLs. By determining the relative impacts of TL features on gene expression, our findings shed new light on the "rules" of directional scanning.

Materials and methods

FACS-uORF and PoLib-Seq data

All raw sequencing data were from NCBI SRA accession number PRJNA721222. The data analyzed in this study included wildtype UTRs containing uORFs (reported previously in [14] and additional wildtype UTRs that do not host uORFs. Data were processed as previously described [14]. Briefly, read pairs were merged and error corrected using FLASH2 [33] using parameters '-z -O -t 1 M 150′. The resulting merged reads were trimmed to remove the ENO2 promoter sequence (plasmid libraries, AGTTTCTTTCATAACACCAAGC) and the complementary DNA adapter sequence (RNA libraries, AGTTTCTTTCATAACACCAAGCNNNN) using cutadapt with parameters '–trimmed-only -e -0.04. RNA-seq libraries were further processed to remove extra nucleotides incorporated at the 5′ cap by reverse transcriptase (NNG [14]). Trimmed reads were counted for perfect matches to designed library constructs using custom perl scripts (DNA-seqcount.pl and RNA-seqcount.pl;) [14]. Relative YFP levels were calculated for each wildtype yeast UTR by comparison to YFP/mCherry TECAN luminometer readings taken from each of the eight FACS-sorted bins after growing the sorted yeast cultures overnight in YPD at 30°C. Each biological replicate was normalized to a 0–1 scale of YFP expression, and read counts were scaled to the proportion of cells sorted into each bin. The average YFP value for each TL was calculated as follows: YFP/mCherry = (SUM(YFPbin) * (reads/bin), where YFPbin represents the YFP/mCherry ratio measured by the TECAN, normalized to a 0–1 scale. The YFP/mCherry levels were compared across the three replicates to remove noisy TLs with inconsistent measurements (standard deviation > 0.05, <50 normalized reads).

PoLib-Seq estimates of ribosome loading were also calculated as previously reported [14]. To summarize, reads were pooled into "translating" (disome and larger) and "nontranslating" (40S, 60S, and monosome) fractions separately for each replicate. While monosomes do include actively translating ribosomes, they are predominantly found on uORFs [34]. As such, we included them in the "nontranslating" fraction, as TLs in the monosome fraction will primarily represent translation of their uORFs. Replicate 1 was downsampled by a factor of 0.826 to ensure similar proportions of translating and nontranslating reads in both replicates. A 5000 total read cutoff (combining the two replicates) was used to calculate the % translating metric for PoLib-Seq library measurements.

Leaky scanning model

Kozak scores ranging from 0 to 1 for each AUG codon were taken from our recent work [13]. These scores represent the relative amount of YFP produced by varying the sequence from −4 to +1 around the YFP start codon. We used these

as scores as approximations of the probability of initiation at start codons in each Kozak context. The LSM is used to calculate an adjusted Kozak score for each TL. The model includes all AUGs present in the TL. uAUGs detract ribosomes from the main start codon and decrease the overall Kozak score. To calculate adjusted Kozak scores that contribute to YFP, we first calculate the probability that a ribosome has skipped all start codons upstream of a productive CDS start codon:

$$P_{skip} = (1 - P_{init1}) - ((1 - P_{init1}) * P_{init2}) \ldots ((1 - P_{init1}) * P_{init2}) \ldots * P_{initn})$$

where P_{init1} representing the Kozak score of the first uORF AUG the PIC reaches in the TL (nearest AUG from 5′ end). P_{skip} represents the amount of "available/remaining" PICs that bypass uAUGs and continue scanning the TL. The adjusted Kozak score for initiation at an AUG that produce YFP is calculated as:

$$P_{CDS} = P_{AUG} * P_{skip}$$

A few TLs contained unannotated N-terminal extensions due to uAUGs that are in frame, without corresponding stop codons. Because such start codons would produce functional YFP, the adjusted Kozak score for such TLs was calculated by summing P_{CDS} scores for each productive AUG. After all AUGs have been accounted for, in-frame AUG Kozak scores are added to the adjusted score, while out-of-frame AUG Kozak scores are subtracted from the score. Thus, the LSM captures the leaky scanning events that ribosomes encounter when nearing AUGs.

TL feature compilation

Kozak scores were taken from our previous reporter studies for AUG start codons (−4 to +1) representing NNNNATGN [14]. Nucleotide frequencies of (A/T/C/G) were calculated and represented as a fraction of the whole TL. There were two separate frequency calculations. Cap-proximal frequencies were calculated using the first ≤20 nt from the 5′ end of the TL, while cap-distal frequencies were calculated using the last ≤30 nt near the start codon. For TLs shorter than these lengths, the entire sequence was used for frequency calculations. All 4-mer motifs (NNNN) that do not contain 'ATG' were included, resulting in a total of 248 unique 4-mers. Each 4-mer count was adjusted using a pseudocount of +1 and then normalized by dividing by the total number of 4-mers in the UTR plus the number of unique 4-mers (+1 for each 4-mer). Mean RNA levels were determined by averaging data over three replicates for each TL. MaxAStretch denotes the longest stretch of consecutive A's in the TL. G quartets (Num4Gs) was the number of times "GGGG" is found in the TL. For each continuous string of G nucleotides, the number of G quartets was defined as the length of consecutive Gs divided by 4 (e.g. GGGGGGG is one quartet, while GGGGGGGG is two quartets). The ViennaRNA [35] package was used for computationally predicting structure around the 5′ cap and the start codon. ΔΔG of the start codon was used for predicted structure around the mainORF. This included 30 nt around the main AUG. The energies of unwinding the RNA around uORF start codons (ΔΔG) were estimated based on previous work [28] using programs from ViennaRNA [35]. For ΔΔG predictions, we used RNAsubopt to predict 100 suboptimal folded structures for each TL, including some of the YFP (50 nt). Start codon unfolded structures were set by unpairing the -15 to +15 region around each start codon. The ΔG of each TL folded and unfolded structure was then predicted via RNAeval. The un-structured estimates were then subtracted from the structured predictions to calculate one hundred estimates of the ΔΔG for each TL. The mean values from these calculations were used as the final modeling features. The ΔG of the 5′ cap was predicted by averaging 100 trials of structure prediction via RNAfold (default parameters). For modeling purposes, the absolute value of ΔG was used so the model coefficient would be more intuitive. G-quadruplexes were estimated using the QGRS package using default parameters (command: qgrs -i input.fa -o output.txt -csv).

Motif predictions and analyses

Motif discovery was performed using the MEME suite, version 5.4.1. TLs from the top and bottom pentiles of RNA levels and relative ribosome loads (RRLs) were compared as primary and control sequences using STREME (parameters –rna –minw 6 –maxw 8 –thresh 0.05 –align left –nmotifs 3). DREME (parameters -rna -norc) was used to identify significant motifs in TLs with over- and under-predicted YFP levels, as STREME found no significant motifs. The FIMO program (parameters –norc –thresh 0.005) was used to identify locations of matches to motifs identified by STREME in the TLs from the top and bottom pentiles. To evaluate positional enrichment of identified motif sites, null distributions were generated for each set of TLs that matched each motif by randomly sampling 50 locations from each TL in each set. The cumulative distribution functions for each motif were compared to their corresponding null distribution using Kolmogorov–Smirnov tests and plotted using R (v 4.2.0). To evaluate cumulative effects of multiple motif occurrences on translation efficiency, the frequency of each motif in each TL was counted. Boxplots with 95% confidence intervals were plotted using R to compare the median YFP expression levels for TLs containing zero or more matches to each motif. [36].

TL histogram-based gradient boosting tree modeling

For modeling, we used a Histogram-based Gradient Boosting Regression (HGBR) Tree. This nonlinear model uses a gradient boosting algorithm to train fast decision trees that can learn complex feature relationships. To build the model, we first tuned the HGBR using Optuna [37], an optimization framework that automatically searches for the best performing hyperparameters given the data. Optuna performed 30 trials using 5-fold cross-validation and optimized parameters such as learning rate, tree depth, and regularization to maximize the R^2 score (final parameters: learning rate = 0.097, max iteration = 473, max depth = 9, l2 = 0.001, max bins = 27, min samples leaf = 13; final non-uAUG parameters: learning rate = 0.037, max iteration = 435, max depth = 9, l2 = 0.015, max bins = 55, min samples leaf = 13). The best parameters were then used to train the final HGBR model on a standardized dataset using sklearn's package HistGradientBoostingRegressor [38]. The data were split as 70% training and 30% testing. The model with best parameters was then run 100 times (on different shuffles of 70%–30% data) to report average R^2 score (0.796) and the best R^2 score (0.816). Permutation importance scores were calculated for every feature included in the model. This score evaluates the results of randomly shuffling each feature's values and observing how

much model performance drops. The best model and scaler were saved, visualized through scatter plots, and feature importance rankings were recorded.

Results

Using multiple MPRAs to evaluate the effects of endogenous yeast TLs on gene expression

We developed a version of FACS-seq, an MPRA system to accurately evaluate protein levels, for endogenous yeast TLs. Previously, we identified significant transcription start sites (TSSs) genome wide and re-annotated TLs in five yeast species [39]. We constructed a library of 11,027 TLs from *S. cerevisiae* and *Saccharomyces paradoxus* upstream of YFP, on plasmids that also express mCherry as an internal control (Fig. 1A). Many genes have multiple TSSs that produce alternative TLs. We assayed TLs that represented at least 10% of the total expression of their host genes and were up to 180-nt long. This includes 86% of all yeast TLs, enabling us to capture a wide range of gene expression and covering numerous sequence features. Yeast transformed with the TL library were FACS-sorted into expression bins based on the YFP/mCherry ratio (Fig. 1B). By sequencing the constructs from each bin, we calculated the mean YFP expression level for individual TLs. We performed FACS-seq in triplicate, and replicates were highly correlated (Supplementary Fig. S1; R^2∼0.98–0.99; Supplementary Table S1). Overall, the range of measured YFP expression among transcripts in our TL library varied over a 100-fold (Fig. 1C). To our knowledge these results define, for the first time, the range of gene expression driven by natural yeast TLs *in vivo*.

TLs can also impact mRNA levels which could contribute to variation in YFP expression. For example, yeast uORFs induce Nonsense Mediated Decay to various extents [14]. Other TL sequences may affect gene expression by recruiting RNA-binding proteins [40, 41] or even causing premature transcription termination [42]. To investigate this, we next performed targeted RNA-seq to estimate reporter RNA levels, normalized to plasmid DNA. Notably, this confirmed that the ENO2 promoter driving YFP used the designed TSSs 97% of the time on average [14]. RNA-seq showed mean RNA levels varied ∼100-fold (Fig. 1D), and positively correlated with YFP expression ($R = 0.392$; Fig. 1E). This is attributable in part to uORFs, which can induce nonsense-mediated decay of mRNA (Fig. 1D). Indeed, constructs with low mRNA levels were significantly enriched with uAUGs in strong Kozak contexts and U-rich motifs, with a slight positional bias nearer to the 5′ cap than expected by chance (Fig. 1G and Supplementary Fig. S2). A/C-rich motifs were overrepresented in constructs with high mRNA levels, concentrated at the 5′ cap, which may promote transcription or RNA stability. (Fig. 1F and Supplementary Fig. S2). Even after excluding TLs with uAUGs, we found AC-rich and U-rich motifs were still significantly enriched in high and low mRNA levels, respectively (Fig. 1F and G, and Supplementary Fig. S2). Thus, natural yeast TLs harbor *cis*-acting sequence features that correlate with variation in mRNA expression levels, potentially due to effects on RNA transcription or stability.

Next, we assayed ribosome loading for TLs in our library using PoLib-Seq [14, 43]. We fractionated polysomes on a sucrose gradient and calculated RRL [%translated/total; see the 'Materials and methods' section) for each TL (Fig. 2, Supplementary Fig. S3, and Supplementary Table S2). RRL varied ∼3-fold for natural yeast TLs (Fig. 2B). There was a strong positive correlation between the FACS-seq and PoLib-Seq assays (Fig. 2C; $R^2 = 0.736$). This indicates that our FACS-seq YFP measurements were largely consistent with ribosome loading. In addition, we observed a complex relationship between ribosome load and mRNA levels. TLs with the lowest ribosome load also had the lowest RNA levels, mid-range RRL values (40%–60%) had the highest RNA levels, and a decline in RNA abundance was observed for TLs with higher ribosome loads (Fig. 2D). This was independent of the presence of uORFs. This observation suggests an optimal level of ribosome loading, wherein translation appears to protect RNA from degradation up to a certain threshold, above which excessive ribosome loading may lead to RNA decay, perhaps due to ribosome collisions [44–46]. Finally, we identified sequence motifs enriched in TLs with high and low-ribosome loading (Fig. 2E and F). In general, we found A-rich sequences, including "AUA", were associated with high ribosome loading, and were enriched downstream of the 5′ cap (Supplementary Fig. S4). Notably, "AUA" was previously reported to be an enhancer element from an *in vitro* ribosome binding assay using yeast extracts [12]. Sequences associated with low ribosome loading included uAUG, concentrated at the 5′ end, and G/C rich sequences that tended to be slightly downstream of the 5′ cap (Supplementary Fig. S5).

Evaluating the impact of TL RNA structures on YFP expression

Prior work found mRNA structures can hinder PIC scanning efficiency and limit initiation [2–4, 47–49]. We next assessed how native yeast structural features affect YFP reporter expression *in vivo*. First, we used RNAfold to calculate the predicted structural stabilities of the TL cap (first 40 nt) and start codon (±15 nt) regions (the 'Materials and methods' section) [35] (Fig. 3A and B). These estimated cap structural stabilities varied over more than a 100-fold range, with the most structured caps reaching −14 kcal/mol. Structures around the start codon were estimated as the change in free energy (ΔΔG) to represent the unfolding energy required for the PIC to access the start codon (maximum = ∼34 kcal/mol). Consistent with current models, we found both structured cap and start codon regions correlated with decreased YFP levels in our library (Fig. 3B and C). Indeed, the structural stability of the TL cap region was much more repressive than stability around the start codon, suggesting structures that inhibit initial PIC loading have a larger effect on gene expression than structures that affect PIC scanning. While RNA structures downstream of the start codon can increase initiation [50], our reporters all have the same sequence downstream of the YFP start codon. As such, we did not compare structural stability downstream of the start codon with YFP expression.

RNA also folds into higher-order structures such as g-quartets which can subsequently fold into stable g-quadruplexes, hindering translation (Fig. 3A) [11, 51]. We examined TLs with the potential to form these higher-order structures to quantify their impacts on YFP expression. TL sequences containing the GGGG motif were considered to have g-quartets. Strikingly, the addition of a single g-quartet significantly decreased YFP expression in natural yeast TLs (Fig. 3D). Using QGRS Mapper [36], we predicted the formation of g-quadruplexes amongst the TLs. While g-quadruplexes were

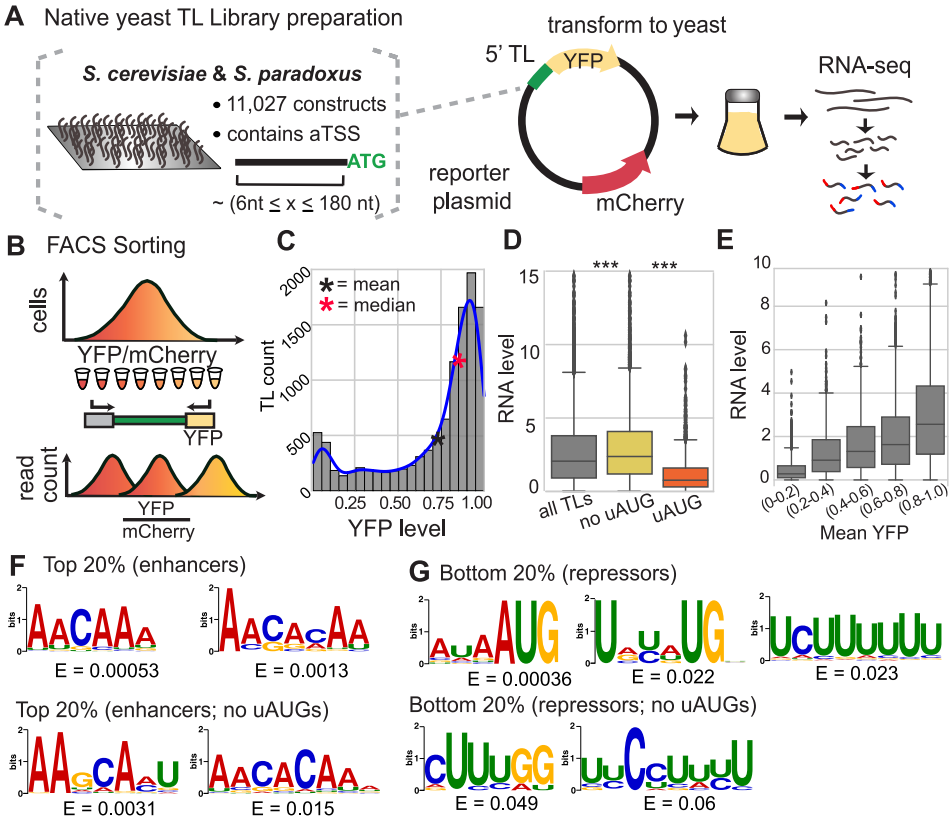


Figure 1. MPRA analysis of TL influence on protein and RNA levels. (**A**) FACS-Seq—a library of thousands of native yeast 5′ TLs, including aTSSs was cloned upstream of a single-copy YFP dual fluorescence reporter plasmid containing mCherry as an internal control. The reporter plasmids were transformed into *S. cerevisiae*. Targeted RNA-seq was used to assay RNA levels, relative to plasmid levels (RNA_{rpkm}/DNA_{rpkm}) (**B**) Cells were sorted and binned via FACS. Plasmids were extracted and sequenced to assay YFP expression levels for each reporter. (**C**) The YFP distribution for the 5′ TL library (mean = 0.713, median = 0.842). (**D**) Average RNA level of transcripts for different 5′ TL groups (all, no uAUG, uAUG). $n = 9382, 7923, 1459$; $***P <$ .001. (**E**) RNA levels for different YFP level groups. Outlier RNA-levels above 10 not shown. $n = 1255, 745, 765, 1769, 6477$. (**F**) and (**G**) STREME motifs identified in the top and bottom 20% of RNA levels with and without uAUGs.

rare (<2% of TLs) in our library, their presence was associated with significantly reduced YFP expression (Fig. 3D). These results show that RNA structures in natural yeast TLs generally reduce protein expression, and show a wide range of effects associated with predicted structural stability of TL cap regions, start codon stability, g-quartets and g-quadruplexes.

Predicting TL YFP expression using an LSM

Our next aim was to quantitatively evaluate the role of Kozak sequences [7, 8] in regulating PIC scanning and initiation in natural yeast TLs. We recently defined yeast Kozak strengths from the −4 to +1 position for AUG and near-AUG start codons using FACS-Seq [14]. Using these data, the Kozak strength of the main start codon alone [main codon model (MCM); Fig. 4A] explained roughly 13% and 22% of expression for the endogenous yeast TLs with and without uAUGs, respectively (Fig. 4B and Supplementary Fig. S6). However, this did not include the impact of uAUGs and their Kozak sequences during PIC scanning. In fact, the presence of even a single uAUG out-of-frame with the main ORF, significantly decreased expression (Fig. 4C). Although rare in our library, TLs containing strong in-frame AUGs led to increased YFP expression as seen by several outliers, consistent with the production of N-terminally extended YFP. To incorporate uAUGs in Kozak predictions, we generated an LSM (Fig. 4A). The LSM calculated probabilities of PICs initiating at uAUGs versus the main AUG based on Kozak strengths. Thus, the LSM captured the propensity of strong uAUG Kozak sequences to deter PICs from the main start codon. Meanwhile, weak Kozak sequences would permit PICs to bypass uAUGs and initiate at main AUGs with higher probability. Compared to the MCM, the LSM more accurately explained reporter expression with an $R^2 = 0.49$ (Fig. 4D). These results show an LSM using Kozak context measurements is a better predictor of gene expression. However, the discrepancy between the predicted and measured expression suggests that additional sequence features contribute to TL regulation of gene expression.

TLs features and modeling YFP

To quantitatively compare the impacts of TL sequence features on gene expression, we constructed a predictive machine learning model. Given the variation in sequence length and our library size, we employed HGBR model to evaluate the role of individual features on gene expression. We evaluated over 200 TL sequence features, including Kozak context, uAUGs, G-quartets, other structural and nucleotide composition features, and 4mers (Fig. 5C and Supplementary Table S3). The HGBR learned l sequence features that contributed to a nonlinear model of gene expression, together explaining 82% of the variance in gene expression (Fig. 5A and B). The most influential predictors of expression for our constructs

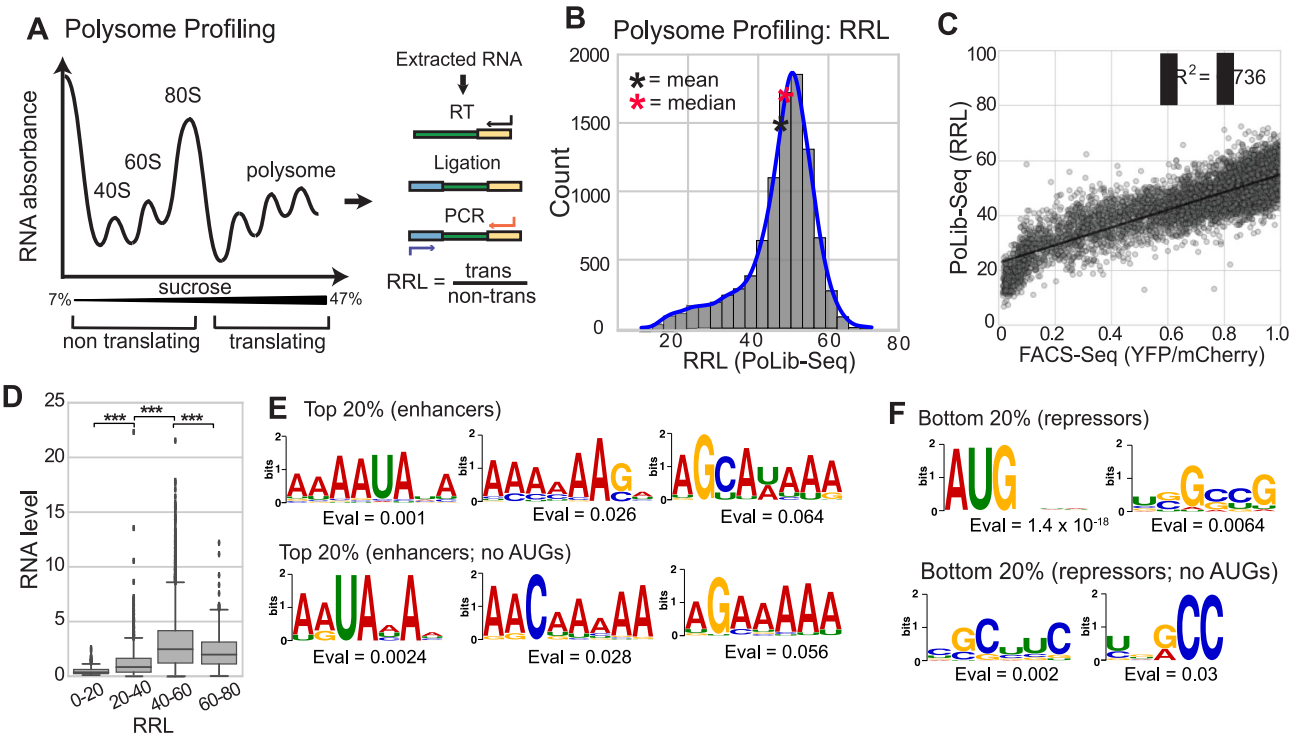


Figure 2. PoLib-Seq MPRA determines relative ribosome loading driven by natural yeast TLs. (**A**) The schematic depicts the PoLib-Seq assay used to measure the ribosome loading on 5′ TLs. Polysome extracts from wildtype *S. cerevisiae* were fractionated on a 7%–47% sucrose gradient via ultracentrifugation. UV absorbance graph represents polysome fractions from nontranslating (40S, 60S, 80S) to translating (2 polysomes +). RNA extracted from Polysome fractions was prepared for sequencing (the 'Materials and methods' section). RRL was calculated as translated/total for each TL. (**B**) Distribution of PoLib-Seq measurements of RRL for 5′ TLs. (**C**) FACS-Seq versus PoLib-Seq: Comparison of YFP/mCherry (x-axis) from FACS-Seq and RRL from PoLib-Seq (y-axis) results for the 5′ TL library. The measurements are highly correlated with R^2 of 0.736. (**D**) Mean RNA levels for different RRL level groups for all TLs. $n = 234, 1396, 7329, 423$. $^{***}P < .001$. (**E, F**) STREME motifs identified in the top (**E**) and bottom (**F**) 20% of RRL levels with and without uAUGs.

were the LSM, presence of uAUGs, and RNA levels (Fig. 5B, Supplementary Fig. S8, and Supplementary Table S4). The model also quantified the effects of mRNA folding around the 5′ cap and other TL on gene expression. For instance, structured 5′ caps are likely to hinder the binding of initiation factors, resulting in lowered PIC loading rates and decreased expression (Fig. 5C) [1, 4, 49, 52]. Additionally, strong structures around start codons could obstruct readability of the main-ORF by impeding PIC scanning and efficient translation initiation [2, 4, 28]. Moreover, the preference for increased adenine frequency around the start codon may aid in preventing the formation of strong structures. Several new 4-mer motifs were identified *in vivo*, although their exact function is unknown. From the HGBR analysis of our TL library, gene expression is significantly impacted by TL length, as expected due to increased overall TL structure and the presence of additional regulatory elements in longer TLs.

Recently, Niederer *et al.* investigated mechanisms driving ribosome recruitment to TLs lacking uAUGs *in vitro* [12] By comparing the efficiency of ribosome recruitment to 5′ UTRs in translation extracts, they uncovered several enhancer and repressive motifs. The AUA motif previously reported as an enhancer by Niederer *et al.*, was enriched in highly translating TLs (Fig. 2E and F), and significantly contributed to the HGBR model of YFP expression, indicating a positive influence albeit with a somewhat modest effect size (Supplementary Table S4). Our results were also consistent with other motifs from the previous *in vitro* study [12], although their relative impacts on gene expression *in vivo* were weaker. To more directly compare our results to Niederer *et al.*, we reevaluated the HGBR model using only TLs lacking uAUGs (Supplementary Fig. S5, $R^2 = 0.612$; Supplementary Table S5). The results reinforced the small positive impact that the AUA motif plays on expression, although the precise mechanism is unknown. Recently, it was reported that yeast eIF3 preferentially binds to the AMAYAA motif [53], which matches both the previously reported AUA sequence and the motifs we find enriched in highly translated reporters (e.g. AAAUANA). Thus it's possible the stimulatory effect results from increased eIF3 binding. Interestingly, we found this motif is enriched 30–50 nt downstream of the 5′ cap. This position is roughly adjacent to the likely PIC binding site predicted by a human cryo-EM structure [54]. These results show that motifs previously found to influence ribosome loading *in vitro* are associated with similar effects on gene expression *in vivo*, although they account for a modest amount of the variance among natural yeast TLs.

We next examined sequence motifs in the HGBR outlier predictions (over-predicted or under-predicted) to improve our model. Outliers were defined as having a change of +-0.25 in measured versus predicted YFP expression (Fig. 5A, dashed lines, the 'Materials and methods' section). Our analysis via DREME [55] identified two motifs that were enriched in the overpredicted outliers (Supplementary Fig. S7), including a "CAUUUCC" motif similar to the reported binding site

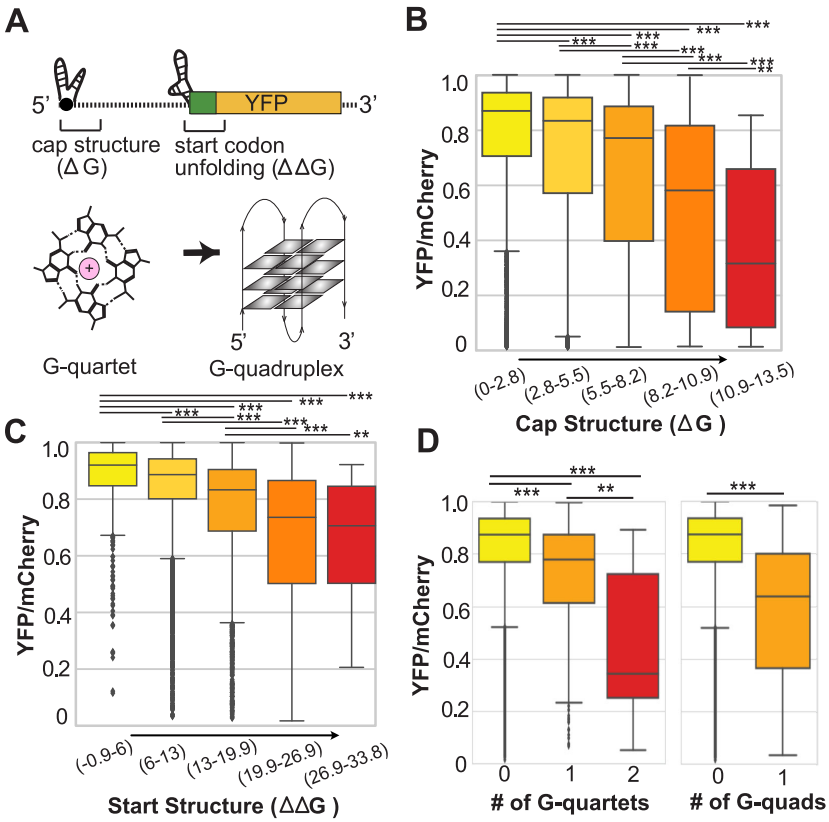


Figure 3. Impacts of natural yeast TL RNA structure on gene expression *in vivo* (**A**) Schematic displaying different mRNA structures: cap structure, start codon structure, g-quartets, g-quadruplexes. (**B**) Boxplots showing the relationship between cap structure (ΔG) and YFP levels. $n = 4279$, 3091, 1338, 248, 32 (no uAUG TLs included) (**C**) Boxplots representing the association between YFP levels and the unfolding energy of structures surrounding the main start codon (ΔΔG). $n = 498$, 5832, 2368, 266, 24 (no uAUG TLs included) (**D**) (left) G-quartet structures form when four guanines (in a row) are hydrogen bonded to each other. From the TL dataset, we saw G-quartets lead to decreased initiation. $n = 8753$, 292, 13 (no uAUG TLs included) (right) In the presence of metal ions, these G-quarters can stack on top of each other to form higher-order structures known as g-quadruplexes. The boxplots show a decrease in YFP expression in the presence of g-quadruplexes $n = 8950$, 107 (no uAUG TLs included). Only one gene (YBR196C-A) was predicted to contain two g-quadruplexes (not shown). (**A–C**) *P*-values: $***P < .001$, $**P < .01$, $*P < .05$

of translational repressor SSB1. Underpredicted outliers, in which the HGBR model predicted lower YFP levels than were actually observed, were enriched for several motifs, including in-frame AUGs that likely increase translation by creating N-terminal extension proteoforms. These findings suggest that there may be unidentified sequence elements in TLs that impact expression levels. Overall, our HGBR models helped us quantify the relative impacts of *cis*-acting sequences and structures on expression from native yeast TLs *in vivo*.

Impacts of aTSSs on protein levels

In eukaryotes, aTSSs can change gene expression levels by introducing additional sequence features. To investigate the impacts of yeast alternative TLs on gene expression, we compared YFP levels from aTSSs in the TL library. We calculated the differences in YFP expression for all pairwise aTSSs. We found that alternative TL isoforms influenced expression up to 78-fold (Fig. 6A). Indeed, changes of as little as 15 nt altered protein expression as much as ~16 fold. In most comparisons, longer TL isoforms (~2000 TLs) were less efficient at translation. In many cases, the longer TLs contain additional uAUGs or more stable RNA structures. For example, a 48 nt longer isoform of the TL from *ARG8*, an aminotransferase encoding gene, introduced a new uAUG with a strong −3A at the 5′ end (Fig. 6B). Similarly, a 105-nt difference in *AST2* TLs introduced four new uAUGs (Fig. 6B). Three uAUGs were out-of-frame with the main start codon and thus competed for PIC recognition and initiation, while the other uAUG created an N-terminal extension. The longer *AST2* transcript is also predicted to form a G-quartet/G-quadruplex structure [36] which may further contribute to its repressive nature. Although rarer in our library, longer TLs sometimes drove increased YFP expression (26%: ~600 TLs). The ubiquitin-specific protease, YER151C (UBP3) has two aTSSs that differ by 16 nt. While the sequence features are similar between the two transcripts, the longer isoform produces a less-structured cap which likely increases PIC 5′ end binding and scanning [49, 56]. One outlier was YLR265C (*NEJ1*), where a 20 nt longer TL increased expression 32-fold. The TL features did not change significantly; however, the distance from the cap to the first uORF increased, suggesting that uAUG location affects initiation as shown in May *et al.* [14]. To further evaluate the accuracy of our HGBR model for TLs, we compared the predicted and measured differences in expression from aTSSs. Notably, we found good agreement ($R^2 = 0.69$; Fig. 6C). Together, these results determine the breadth of expression differences caused by altering only the 5′ TSS in native yeast TLs *in vivo* and highlight TL features that could account for this variation.

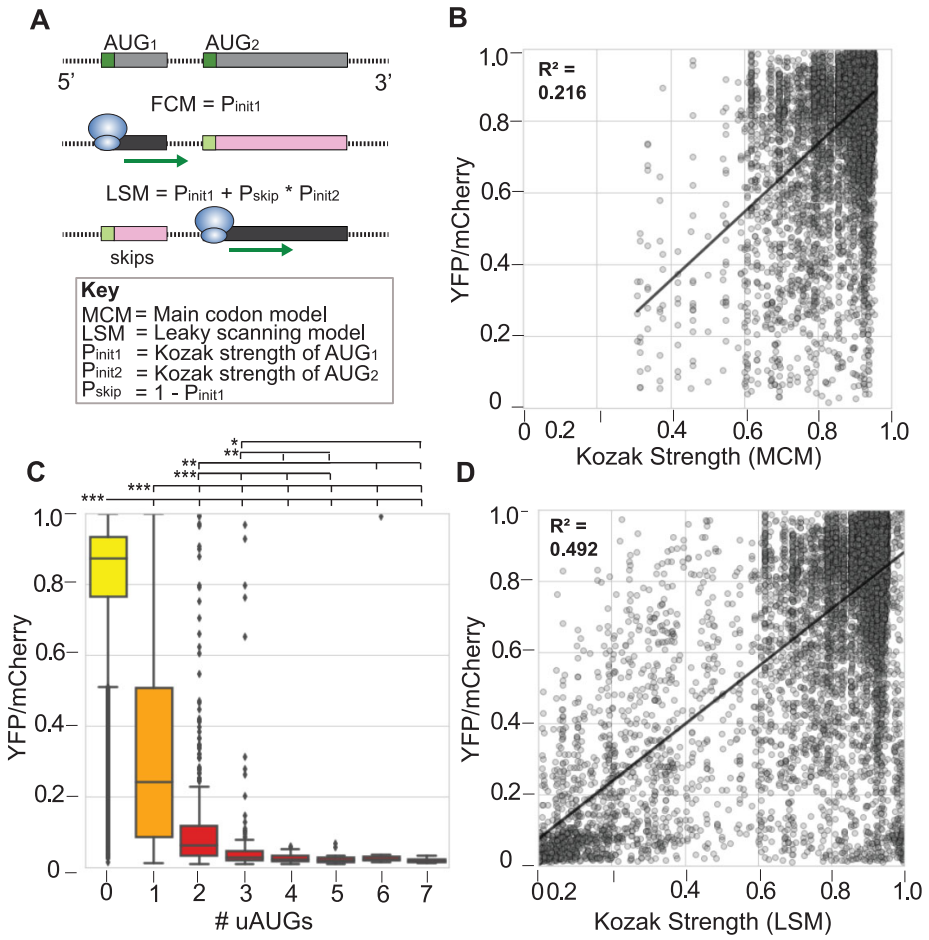


Figure 4. A simple LSM explains half of the variance of gene expression from natural yeast TLs. (**A**) Schematic describing a simple LSM. (Top) mRNA containing two AUG start codons. (Middle) The MCM predicts YFP expression using the Kozak strength for the main CDS start codon. (Bottom) The LSM predicts YFP expression using the Kozak strengths of all AUGs. This example shows the ribosome skipping the first AUG and initiating at the second ORF. The probability of initiating at YFP is given by the fraction of ribosomes that reach the CDS start codon (Pskip) times the Kozak strength at the YFP start codon (Pinit2). (**B**) The MCM explains 21.6% of the variance in YFP levels for TLs without any uAUGs ($R^2 = 0.216$; see Fig. S6 for uORF TLs). (**C**) Boxplots representing the distribution of measured YFP expression (y-axis) for native 5′ TLs binned by the number of uORFs (x-axis). Additional uAUGs further repress YFP expression. *P*-values: ***$P < .001$, **$P < .01$, *$P < .05$. uAUGs: $n = 0{:}9058$, 1:1448, 2:309, 3:126, 4:50, 5:22, 6:10, 7:4 (**D**) Linear regression model of measured YFP (y-axis) versus the LSM predicted Kozak strength (x-axis) for each 5′ TL.

Discussion

Due to their substantial influence on mRNA translation and decay, 5′ TLs play a key role in regulating gene expression. Early studies mainly examined the effects of individual TL sequence elements, such as Kozak context and RNA structure, on gene expression [2, 49, 50, 56]. More recently, MPRAs have allowed researchers to simultaneously test thousands of designer and randomized sequences *in vivo* [10, 13, 14, 57–60] and *in vitro* [12]. Although informative, these studies often involved technological compromises that could affect their interpretation. For instance, several studies appended long fixed sequences to the 5′ terminus of reporter libraries. While this facilitates PCR amplification, it also removes all variation in sequence and structure at the 5′ cap. Several previous studies also used fixed length, randomized TL sequences [10, 13, 57, 60]. Consequently, such studies cannot capture the effects of TL length and have limited structural variation. Finally, previous large-scale studies of TL effects were limited to either protein expression or ribosome loading alone. Here, we combined three separate MPRA analyses to investigate the impact of 5′ UTR sequences on mRNA levels, ribosome recruitment, and protein expression. By assaying thousands of natural yeast TLs, we quantified the respective roles of known and novel 5′ UTR *cis*-acting elements in gene expression.

The Kozak sequence context surrounding mRNA start codons has long been recognized as a major determinant of initiation efficiency [7]. The strength of Kozak contexts influences translation initiation via leaky scanning, in which PICs bypass inefficient start codons in a condition-specific manner [47, 61]. We evaluated a simple LSM incorporating all uAUGs and their Kozak contexts in our library of reporter elements. Our LSM explains ~49% of variance in expression. This indicates that the Kozak sequence, though undoubtedly important, cannot fully explain the variation in translation initiation found in yeast 5′ UTRs. Notably, it has been suggested that the Kozak sequence is less impactful in *S. cerevisiae* than in other eukaryotes [62], due to loss of protein interaction domains in yeast initiation factors. Thus, leaky scanning may play a more prominent role in other species. Similarly, previous work found that near-AUG start codons can drive significant translation initiation in mammalian tissue culture cells [63], but not in yeast [14]. Future work is needed to determine whether

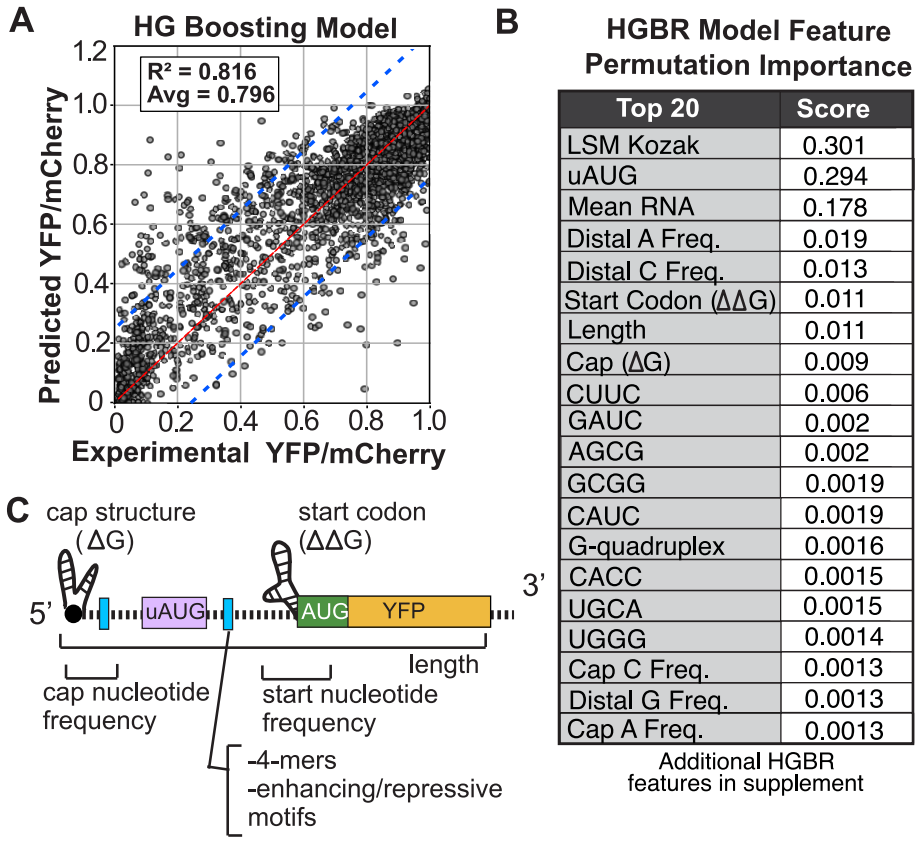


Figure 5. Nonlinear regression modeling determines the relative influence of TL features on gene expression *in vivo* (**A**) HGBR for predicting YFP expression from TL features. The scatter plot shows the measured YFP expression (y-axis) for all 5′ TLs versus the HGBR model predictions of YFP (x-axis) in WT yeast. The resulting model explains ~80% of variance in experimental YFP. (**B**) The table shows model importance scores for the significant features extracted from the HGBR model after $n = 100$ iterations (additional features shown in supplemental data). (**C**) Schematic of 5′ TL features predicted to influence YFP expression based on the EN model.

Kozak sequences, near-AUG start codons, and leaky scanning explain more of the variance in expression across natural human TLs than we observed with yeast TLs.

Although our LSM was a better predictor of yeast TL activity than start codon Kozak strength alone, TLs contain many *cis*-acting elements that could affect scanning PICs [61, 64]. For example, including constraints on uORF start codon structure was found to greatly improve translation efficiency predictions for human SERPINA1 mRNA isoform reporters [28]. Incorporation of isoform-specific structure probing data might increase the predictive power of our LSM. Indeed, we found several TL structural features influenced protein expression, including structure around the main ORF start codon and G-quadruplexes. In addition, many other features of uORFs can affect their relative influence on leaky-scanning, including uORF length, position, and the charge of encoded uORF peptides [14]. By increasing the time ribosomes occupy TLs, these features may inhibit the loading and scanning by additional PICs. For example, PIC collisions or pausing may lead to slowed main ORF initiation, a shift of translation to upstream start codons, or increased mRNA decay [65–68]. Such a reduction would give uORFs a greater influence on initiation than expected from LSMs.

The limitations of LSM models may be addressable using Totally Asymmetric Simple Exclusion Process (TASEP) models [69, 70]. For example, Andreev *et al.* described a TASEP model variant (ICIER) in which ribosomes translating uORFs displace downstream scanning PICs, while upstream scanning PICs queue behind translating ribosomes. Notably, this model can create automatic derepression of translation under stress conditions, when PICs load less frequently at mRNA 5′ ends. However, the rapid nature of translation initiation [71] may make the presence of multiple PICs on individual TLs rare. Yet, TASEP-inspired models hold promise for future prediction of translation initiation, especially with the incorporation of additional parameters, including mRNA structures.

Unlike mechanistic models such as LSM and TASEP [70], computational and machine learning methods can identify more complex TL features and interactions [10, 11, 13, 58]. Although previous studies investigated TL control and identified regulatory elements [10–12], they did not directly explain what occurs in native TLs *in vivo*. To better explain natural gene expression, we built a comprehensive TL model by combining FACS-Seq with HGBR. The resulting model explained 82% of variation in YFP for all native yeast TLs up to 180 nt *in vivo*. Our results confirmed known *cis*-acting motifs, including AUG-Kozak pairs, sequence composition, and mRNA structures, while also extending our understanding by defining their relative roles in natural TLs. Our results are consistent with motifs previously reported to affect ribosome loading *in vitro* [12], as *in vitro* "AUA" enhancer element was enriched in mRNAs with high ribosome loading *in vivo*. Several 4-mers were identified which may be indicative of structure or include unidentified RNA binding motifs. While it is unclear which

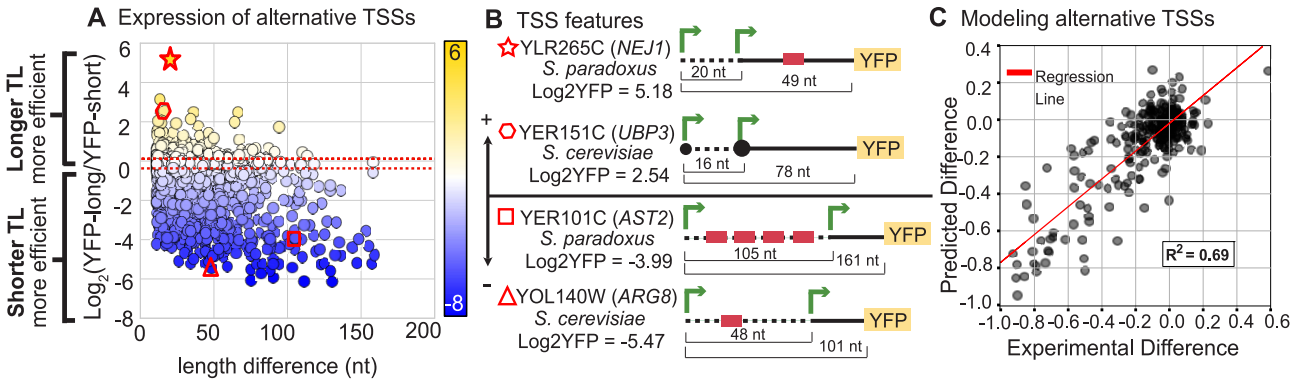


Figure 6. Effects of natural yeast aTSSs on protein expression *in vivo*. (**A**) Scatter plot compares the log fold change [$\log_2$(long/short)] of TLs for genes with aTSSs ($R^2 = 0.153$). Typically, longer TLs displayed greater changes in YFP levels. The negative values indicate the longer TL expresses less YFP than its shorter counterpart. A positive value indicates the shorter TL had lower YFP levels. The dashed lines represent the median positive (0.139) and negative (−0.411) values. Symbols identify genes represented in Fig. 2B. (**B**) Examples of alternative TLs that significantly changed YFP expression. (Top) YLR265C (*S. paradoxus*) and YER151C (*S. cerevisiae*) both had longer transcripts that increased YFP levels. (Bottom) YOL140W (*S. cerevisiae*) and YER101C (*S. paradoxus*) are examples of longer TLs which repressed YFP expression with additions of uAUGs. (**C**) HGBR model predictions of differences in YFP levels for aTSSs ($R^2 = 0.69$).

trans-acting factors recognize the AUA containing motif, notably the motif is similar to the position specific element used in mRNA cleavage and 3′ end formation [72]. We also identified C/U rich motifs associated with lower expression and ribosome loading. This motif is found in sequences bound by the translational repressor *SSD1* [73], making it a prime candidate for the corresponding *trans*-acting factor. Our model underscores the importance of the nucleotide composition at the 5′ cap and around the start codon which could influence structure and PIC binding and initiation. Thus, our HGBR model demonstrates the wide *cis*-acting potential of TLs and their direct impacts on gene expression.

Although our HGBR model accurately predicts expression, it has some limitations. Our structural predictions are constrained as they rely on predictive ΔG measurements that fail to capture various mRNA structural states. Furthermore, by omitting sequences downstream of each TL start codon, our approach does not account for RNA structures downstream that can increase initiation [50] To overcome this limitation in future studies, it would be valuable to verify mRNA structures and cap sequences via structure probing experiments. In a complementary paper, we discuss how uORF features, such as location, codon makeup, and length, impact expression [14]. These features may account for some of the unexplained variance in our current LSM and HGBR models. Furthermore, the hidden interactions between structures and uORFs are not directly examined. There may be a seesaw effect on the level of repression between these two features. Upstream structures can mask the repressive nature of uORFs by reducing PIC loading, occluding uORF starts, and minimizing PIC collisions or queuing on uORFs [29, 74]. Indeed, we see some of the overpredicted TLs containing strong 5′ cap structures along with uORFs downstream. Conversely, strong uORFs without any inhibitory structures upstream may dominate PIC usage, resulting in lowered mainORF expression. Some studies hint at such models by examining ribosome loading on uORFs versus CDSs [75, 76]. Nevertheless, PIC scanning and pausing are still not well understood. By studying PIC interactions and disomes, we may increase our understanding of scanning efficiency. Additionally, mRNAs interact with *trans*-acting factors such as 5′ RNA-binding proteins, represented by consensus binding motifs in our model. Incorporating RBP interactions from CLIP datasets could boost the performance of our TL models. Finally, because our study focused on natural TLs, our model is limited to sequences and motifs found in most yeast TLs. For example, i-motifs are not present in our library. Thus, additional unknown features or feature interactions could explain the remaining $\frac{1}{5}$ of the variance in expression that is unaccounted in our model.

Notably, we found a complex relationship between mRNA levels, translation, and protein expression. In general, TLs containing uORFs were associated with lower mRNA abundance and correspondingly lower protein levels. Consistent with this, we found TLs that had very low ribosome loading also had correspondingly low mRNA levels. However, we also observed low mRNA levels for TLs with the highest ribosome load. This suggests a model in which overloading of ribosomes leads to mRNA destabilization, perhaps via ribosome collisions and the ribosome quality control decay pathway [77, 78]. In this case, TLs may be somewhat optimized for specific protein coding genes such that collision prone transcripts have reduced ribosome loading while rapidly translating ORFs can load ribosomes more quickly [79]. If so, this would have important implications for the design of mRNA therapeutics, as the optimal rate of ribosome loading may depend on the probability of ribosome collisions in a given ORF.

Many of the aTSSs we tested are differentially regulated in yeast. Previous work suggests such transcripts may have regulatory roles. For example, shifts in TSS usage have been reported in different environmental conditions [30] and throughout meiosis [16, 80]. Studies in mammals suggest that aTSSs drive mRNA isoform diversity and have specific functions [81–83]. Our analyses showed that yeast alternative TLs can alter protein expression up to ~80-fold, revealing the vast regulatory potential and diversity of yeast TLs. Although upstream TSSs generally produced less efficient TLs, there were cases where we observed increases in expression. Further, the changes in aTSS expression were frequently attributed to the introduction or removal of regulatory features. Future work is needed to evaluate the production and significance of aTSSs. Such studies are expected to reveal regulatory

links between transcription regulation, environmental conditions, mRNA translation and turnover.

In summary, we evaluated the landscape of TL *cis*-acting features in yeast grown to log phase under unstressed conditions. However, environmental stimuli can cause dramatic changes in mRNA translation, including suppressing canonical cap translation and altering the effects of sequence features [84–89]. While current experiments of single genes under stress or varied conditions provide valuable insights, large-scale studies are needed to detect critical TL features and their role in genetic reprogramming. New methods, such as incorporation of rapidly turned over proteins or degrons [90–92], may be useful in stress response studies. It would also be interesting to investigate the direct impact of stress and the interplay of TL features on PIC scanning. Such future studies of translational response to stress may provide key insights into reestablishing homeostasis, channeling resources towards stress-related genes, and altering PIC trajectories.

Acknowledgements

The authors would like to thank the members of the McManus and Woolford laboratories at CMU for helpful discussions and suggestions. This study was supported by NIGMS grant R35GM145317 to CJM.

Author contributions: Christina Akirtava (Data curation, Software, Analysis, Investigation, Visualization, Methodology, Writing), Gemma E. May (Investigation, Methodology, Project administration, Funding acquisition, Writing), C. Joel McManus (Conceptualization, Funding acquisition, Supervision, Visualization, Analysis, Writing)

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This study was supported by NIGMS [R35GM145317 to CJM.], and CMU. Funding to pay the Open Access publication charges for this article was provided by NIH [R35GM145317 to C.J.M.].

Data availability

The raw sequencing data for FACS-uORF, PoLib-Seq, and mRNA levels were downloaded from NCBI under the SRA accession number PRJNA721222 [14].

References

1. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science* 2016;**352**:1413–6. https://doi.org/10.1126/science.aad9868
2. Kozak M. Leader length and secondary structure modulate mRNA function under conditions of stress. *Mol Cell Biol* 1988;**8**:2737–44.
3. Cigan AM, Pabich EK, Donahue TF. Mutational analysis of the HIS4 translational initiator region in *Saccharomyces cerevisiae*. *Mol Cell Biol* 1988;**8**:2964–75.
4. Kozak M. The scanning model for translation: an update. *J Cell Biol* 1989;**108**:229–41. https://doi.org/10.1083/jcb.108.2.229
5. Arribere JA, Gilbert WV. Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res* 2013;**23**:977–87. https://doi.org/10.1101/gr.150342.112
6. Wethmar K. The regulatory potential of upstream open reading frames in eukaryotic gene expression. *Wiley Interdiscip Rev RNA* 2014;**5**:765–8. https://doi.org/10.1002/wrna.1245
7. Kozak M. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res* 1984;**12**:857–72. https://doi.org/10.1093/nar/12.2.857
8. Hinnebusch AG. Structural insights into the mechanism of scanning and start codon recognition in eukaryotic translation initiation. *Trends Biochem Sci* 2017;**42**:589–611. https://doi.org/10.1016/j.tibs.2017.03.004
9. Hashem Y, Frank J. The jigsaw puzzle of mRNA translation initiation in eukaryotes: a decade of structures unraveling the mechanics of the process. *Annu Rev Biophys* 2018;**47**:125–51. https://doi.org/10.1146/annurev-biophys-070816-034034
10. Dvir S, Velten L, Sharon E *et al.* Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. *Proc Natl Acad Sci USA* 2013;**110**:E2792–801. https://doi.org/10.1073/pnas.1222534110
11. Cuperus JT, Groves B, Kuchina A *et al.* Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res* 2017;**27**:2015–24. https://doi.org/10.1101/gr.224964.117
12. Niederer RO, Rojas-Duran MF, Zinshteyn B *et al.* Direct analysis of ribosome targeting illuminates thousand-fold regulation of translation initiation. *Cell Syst* 2022;**13**:256–64. https://doi.org/10.1016/j.cels.2021.12.002
13. Sample PJ, Wang B, Reid DW *et al.* Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat Biotechnol* 2019;**37**:803–9. https://doi.org/10.1038/s41587-019-0164-5
14. May GE, Akirtava C, Agar-Johnson M *et al.* Unraveling the influences of sequence and position on yeast uORF activity using massively parallel reporter systems and machine learning. *eLife* 2023;**12**:e69611. https://doi.org/10.7554/eLife.69611
15. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 2011;**147**:789–802. https://doi.org/10.1016/j.cell.2011.10.002
16. Brar GA, Yassour M, Friedman N *et al.* High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* 2012;**335**:552–7. https://doi.org/10.1126/science.1215110
17. Ivanov IP, Firth AE, Michel AM *et al.* Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. *Nucleic Acids Res* 2011;**39**:4220–34. https://doi.org/10.1093/nar/gkr007
18. Ivanov Michel AM, Wei J, Caster SZ *et al.* Translation initiation from conserved non-AUG codons provides additional layers of regulation and coding capacity. *mBio* 2017;**8**:e00844-17. https://doi.org/10.1128/mBio.00844-17
19. Lee S, Liu B, Lee S *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci USA* 2012;**109**:E2424–32. https://doi.org/10.1073/pnas.1207846109
20. Monteuuis G, Miścicka A, Świrski M *et al.* Non-canonical translation initiation in yeast generates a cryptic pool of mitochondrial proteins. *Nucleic Acids Res* 2019;**47**:5777–91. https://doi.org/10.1093/nar/gkz301
21. Kearse MG, Wilusz JE. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev* 2017;**31**:1717–31. https://doi.org/10.1101/gad.305250.117
22. Eisenberg AR, Higdon AL, Hollerer I *et al.* Article translation initiation site profiling reveals widespread synthesis of non-AUG-initiated protein isoforms in yeast translation initiation

site profiling reveals widespread synthesis of Non-AUG-initiated protein isoforms in yeast. *Cell Syst* 2020;**11**:145–60. https://doi.org/10.1016/j.cels.2020.06.011
23. Zitomer RS, Walthall DA, Rymond BC *et al.* *Saccharomyces cerevisiae* ribosomes recognize non-AUG initiation codons. *Mol Cell Biol* 1984;**4**:1191–7.
24. Clements JM, Laz TM, Sherman F. Efficiency of translation initiation by non-AUG codons in *Saccharomyces cerevisiae*. *Mol Cell Biol*. 1988;**8**:4533–6.
25. Takacs JE, Neary TB, Ingolia NT *et al.* Identification of compounds that decrease the fidelity of start codon recognition by the eukaryotic translational machinery. *RNA* 2011;**17**:439–52. https://doi.org/10.1261/rna.2475211
26. Kolitz SE, Takacs JE, Lorsch JR. Kinetic and thermodynamic analysis of the role of start codon/anticodon base pairing during eukaryotic translation initiation. *RNA* 2009;**15**:138–52. https://doi.org/10.1261/rna.1318509
27. Saini AK, Nanda JS, Lorsch JR *et al.* Regulatory elements in eIF1A control the fidelity of start codon selection by modulating tRNA(i)(Met) binding to the ribosome. *Genes Dev* 2010;**24**:97–110. https://doi.org/10.1101/gad.1871910
28. Corley M, Solem A, Phillips G *et al.* An RNA structure-mediated, posttranscriptional model of human α-1-antitrypsin expression. *Proc Natl Acad Sci USA* 2017;**114**:E10244–53. https://doi.org/10.1073/pnas.1706539114
29. Mustoe AM, Corley M, Laederach A *et al.* Messenger RNA structure regulates translation initiation: a mechanism exploited from bacteria to humans. *Biochemistry* 2018;**57**:3537–9. https://doi.org/10.1021/acs.biochem.8b00395
30. Waern K, Snyder M. Extensive transcript diversity and novel upstream open reading frame regulation in yeast. *G3 (Bethesda)* 2013;**3**:343–52. https://doi.org/10.1534/g3.112.003640
31. Wang X, Hou J, Quedenau C *et al.* Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. *Mol Syst Biol* 2016;**12**:875. https://doi.org/10.15252/msb.20166941
32. Rojas-duran MF, Gilbert WV. Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA* 2012;**18**:2299–305. https://doi.org/10.1261/rna.035865.112
33. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 2011;**27**:2957–63. https://doi.org/10.1093/bioinformatics/btr507
34. Heyer EE, Moore MJ. Redefining the translational status of 80S monosomes. *Cell* 2016;**164**:757–69. https://doi.org/10.1016/j.cell.2016.01.003
35. Gruber AR, Lorenz R, Bernhart SH *et al.* The Vienna RNA websuite. *Nucleic Acids Res* 2008;**36**:W70–4. https://doi.org/10.1093/nar/gkn188
36. Kikin O, D'antonio L, Bagga PS. QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res* 2006;**34**:W676–82. https://doi.org/10.1093/nar/gkl253
37. Akiba T, Sano S, Yanase T *et al.* Optuna: a next-generation hyperparameter optimization framework. arXiv, https://arxiv.org/abs/1907.10902, 25 July 2019, preprint: not peer reviewed.
38. Mayer M, Bourassa SC, Hoesli M *et al.* Machine learning applications to land and structure valuation. *J Risk Fin Manag* 2022;**15**:193.
39. Spealman P, Naik AW, May GE *et al.* Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. *Genome Res* 2018;**28**:214–22. https://doi.org/10.1101/gr.221507.117
40. Hogan DJ, Riordan DP, Gerber AP *et al.* Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol* 2008;**6**:e255. https://doi.org/10.1371/journal.pbio.0060255
41. Baejen C, Torkler P, Gressel S *et al.* Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition. *Mol Cell* 2014;**55**:745–57. https://doi.org/10.1016/j.molcel.2014.08.005
42. Kuehner JN, Brow DA. Regulation of a eukaryotic gene by GTP-dependent start site selection and transcription attenuation. *Mol Cell* 2008;**31**:201–11. https://doi.org/10.1016/j.molcel.2008.05.018
43. May GE, McManus CJ. Multiplexed analysis of Human uORF regulatory functions during the ISR using PoLib-Seq. *Methods Mol Biol* 2022;**2428**:41–62. https://doi.org/10.1007/978-1-0716-1975-9_3
44. Radhakrishnan A, Green R. Connections underlying translation and mRNA stability. *J Mol Biol* 2016;**428**:3558–64. https://doi.org/10.1016/j.jmb.2016.05.025
45. Simms CL, Yan LL, Zaher HS. Ribosome collision is critical for quality control during No-go decay. *Mol Cell* 2017;**68**:361–73. https://doi.org/10.1016/j.molcel.2017.08.019
46. Bicknell AA, Reid DW, Licata MC *et al.* Attenuating ribosome load improves protein output from mRNA by limiting translation-dependent mRNA decay. *Cell Rep* 2024;**43**:114098. https://doi.org/10.1016/j.celrep.2024.114098
47. Kochetov AV. Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *Bioessays* 2008;**30**:683–91. https://doi.org/10.1002/bies.20771
48. Robbins-Pianka A, Rice MD, Weir MP. The mRNA landscape at yeast translation initiation sites. *Bioinformatics* 2010;**26**:2651–5. https://doi.org/10.1093/bioinformatics/btq509
49. Babendure JR, Babendure JL, Ding J-H *et al.* Control of mammalian translation by mRNA structure near caps. *RNA* 2006;**12**:851–61. https://doi.org/10.1261/rna.2309906
50. Kozak M. Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J Biol Chem* 1991;**266**:19867–70. https://doi.org/10.1016/S0021-9258(18)54860-2
51. Fay MM, Lyons SM, Ivanov P. RNA G-quadruplexes in biology: principles and molecular mechanisms. *J Mol Biol* 2017;**429**:2127–47. https://doi.org/10.1016/j.jmb.2017.05.017
52. Akirtava C, McManus CJ. Control of translation by eukaryotic mRNA transcript leaders-insights from high-throughput assays and computational modeling. *Wiley Interdiscip Rev RNA* 2021;**12**:e1623. https://doi.org/10.1002/wrna.1623
53. Koubek J, Kaur J, Bhandarkar S *et al.* Cellular translational enhancer elements that recruit eukaryotic initiation factor 3. *RNA* 2025;**31**:193–207. https://doi.org/10.1261/rna.080310.124
54. Brito Querido J, Sokabe M, Kraatz S *et al.* Structure of a human 48S translational initiation complex. *Science* 2020;**369**:1220–7. https://doi.org/10.1126/science.aba4904
55. Bailey TL. STREME: accurate and versatile sequence motif discovery. *Bioinformatics* 2021;**37**:2834–40. https://doi.org/10.1093/bioinformatics/btab203
56. Kozak M. Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. *Proc Natl Acad Sci USA* 1986;**83**:2850–4. https://doi.org/10.1073/pnas.83.9.2850
57. Cuperus JT, Groves B, Kuchina A *et al.* Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome Res* 2017;**27**:2015–24. https://doi.org/10.1101/gr.224964.117
58. Noderer WL, Flockhart RJ, Bhaduri A *et al.* Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol Syst Biol* 2014;**10**:748. https://doi.org/10.15252/msb.20145136
59. Ferreira JP, Overton KW, Wang CL. Tuning gene expression with synthetic upstream open reading frames. *Proc Natl Acad Sci USA* 2013;**110**:11284–9. https://doi.org/10.1073/pnas.1305590110
60. Lin Y, May GE, Kready H *et al.* Impacts of uORF codon identity and position on translation regulation. *Nucleic Acids Res* 2019;**47**:9358–67. https://doi.org/10.1093/nar/gkz681
61. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* 2016;**352**:1413–6. https://doi.org/10.1126/science.aad9868

62. Wallace EWJ, Maufrais C, Sales-Lee J *et al.* Quantitative global studies reveal differential translational control by start codon context across the fungal kingdom. *Nucleic Acids Res* 2020;**48**:2312–31. https://doi.org/10.1093/nar/gkaa060
63. Diaz de Arce AJ, Noderer WL, Wang CL. Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. *Nucleic Acids Res* 2018;**46**:985–94. https://doi.org/10.1093/nar/gkx1114
64. Akirtava C, McManus CJ. Control of translation by eukaryotic mRNA transcript leaders—Insights from high-throughput assays and computational modeling. *Wiley Interdiscip Rev RNA* 2021;**12**:e1623. https://doi.org/10.1002/wrna.1623
65. Kearse MG, Goldman DH, Choi J *et al.* Ribosome queuing enables non-AUG translation to be resistant to multiple protein synthesis inhibitors. *Genes Dev* 2019;**33**:871–85. https://doi.org/10.1101/gad.324715.119
66. Han P, Shichino Y, Schneider-Poetsch T *et al.* Genome-wide survey of ribosome collision. *Cell Rep* 2020;**31**:107610. https://doi.org/10.1016/j.celrep.2020.107610
67. Hinnebusch AG. Translational regulation of GCN4 and the general amino Acid control of yeast ∗. *Annu Rev Microbiol* 2005;**59**:407–50. https://doi.org/10.1146/annurev.micro.59.031805.133833
68. Collart MA, Weiss B. Ribosome pausing, a dangerous necessity for co-translational events. *Nucleic Acids Res* 2020;**48**:1043–55. https://doi.org/10.1093/nar/gkz763
69. Zia RKP, Dong JJ, Schmittmann B. Modeling translation in protein synthesis with TASEP: a tutorial and recent developments. *J Stat Phys* 2011;**144**:405–28. https://doi.org/10.1007/s10955-011-0183-1
70. Andreev DE, Arnold M, Kiniry SJ *et al.* TASEP modelling provides a parsimonious explanation for the ability of a single uORF to derepress translation during the integrated stress response. *eLife* 2018;**7**:e32563. https://doi.org/10.7554/eLife.32563
71. Wang J, Shin B-S, Alvarado C *et al.* Rapid 40S scanning and its regulation by mRNA structure during eukaryotic translation initiation. *Cell* 2022;**185**:4474–87. https://doi.org/10.1016/j.cell.2022.10.005
72. Zhao J, Hyman L, Moore C. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* 1999;**63**:405–45. https://doi.org/10.1128/MMBR.63.2.405-445.1999
73. Bayne RA, Jayachandran U, Kasprowicz A *et al.* Yeast Ssd1 is a non-enzymatic member of the RNase II family with an alternative RNA recognition site. *Nucleic Acids Res* 2022;**50**:2923–37. https://doi.org/10.1093/nar/gkab615
74. Ivanov IP, Shin B-S, Loughran G *et al.* Polyamine control of translation elongation regulates start site selection on antizyme inhibitor mRNA via ribosome queuing. *Mol Cell* 2018;**70**:254–64. https://doi.org/10.1016/j.molcel.2018.03.015
75. Chew G-L, Pauli A, Schier AF. Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat Commun* 2016;**7**:11663. https://doi.org/10.1038/ncomms11663
76. Johnstone TG, Bazzini AA, Giraldez AJ. Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J* 2016;**35**:706–23. https://doi.org/10.15252/embj.201592759
77. D'Orazio KN, Green R. Ribosome states signal RNA quality control. *Mol Cell* 2021;**81**:1372–83. https://doi.org/10.1016/j.molcel.2021.02.022
78. Simms CL, Thomas EN, Zaher HS. Ribosome-based quality control of mRNA and nascent peptides. *Wiley Interdiscip Rev RNA* 2017;**8**:10.1002. https://doi.org/10.1002/wrna.1366
79. Park H, Subramaniam AR. Inverted translational control of eukaryotic gene expression by ribosome collisions. *PLoS Biol* 2019;**17**:e3000396. https://doi.org/10.1371/journal.pbio.3000396
80. Cheng Z, Otto GM, Powers EN *et al.* Pervasive, coordinated protein-level changes driven by transcript isoform switching during meiosis. *Cell* 2018;**172**:910–23. https://doi.org/10.1016/j.cell.2018.01.035
81. Pal S, Gupta R, Kim H *et al.* Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res* 2011;**21**:1260–72. https://doi.org/10.1101/gr.120535.111
82. Resch AM, Ogurtsov AY, Rogozin IB *et al.* Evolution of alternative and constitutive regions of mammalian 5'UTRs. *BMC Genomics* 2009;**10**:162.
83. Wang ET, Sandberg R, Luo S *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008;**456**:470–6. https://doi.org/10.1038/nature07509
84. Wu C-C, Peterson A, Zinshteyn B *et al.* Ribosome collisions trigger general stress responses to regulate cell fate. *Cell* 2020;**182**:404–16. https://doi.org/10.1016/j.cell.2020.06.006
85. Starck SR, Tsai JC, Chen K *et al.* Translation from the 5' untranslated region shapes the integrated stress response. *Science* 2016;**351**:465. https://doi.org/10.1126/science.aad3867
86. Andreev DE, O'connor PB, Fahey C *et al.* Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *eLife* 2015;**4**:e03971. https://doi.org/10.7554/eLife.03971
87. Wek RC. Role of eIF2α kinases in translational control and adaptation to cellular stress. *Cold Spring Harb Perspect Biol* 2018;**10**:a032870. https://doi.org/10.1101/cshperspect.a032870
88. Morano KA, Grant CM, Moye-Rowley WS. The response to heat shock and oxidative stress in *Saccharomyces cerevisiae*. *Genetics* 2012;**190**:1157–95. https://doi.org/10.1534/genetics.111.128033
89. Qian S-B, Liu B. Translational reprogramming in stress response. *Wiley Interdiscip Rev RNA* 2014;**5**:301–5.
90. Houser JR, Ford E, Chatterjea SM *et al.* An improved short-lived fluorescent protein transcriptional reporter for *Saccharomyces cerevisiae*. *Yeast* 2012;**29**:519–30. https://doi.org/10.1002/yea.2932
91. Morawska M, Ulrich HD. An expanded tool kit for the auxin-inducible degron system in budding yeast. *Yeast* 2013;**30**:341–51. https://doi.org/10.1002/yea.2967
92. Jivotovskaya AV, Valášek L, Hinnebusch AG *et al.* Eukaryotic translation initiation factor 3 (eIF3) and eIF2 can promote mRNA binding to 40S subunits independently of eIF4G in yeast. *Mol Cell Biol* 2006;**26**:1355–72. https://doi.org/10.1128/MCB.26.4.1355-1372.2006

Received: July 12, 2024. **Revised:** February 16, 2025. **Editorial Decision:** February 17, 2025. **Accepted:** February 20, 2025