**Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast**

**-**

**Supplemental Information**

Jun Cheng1,2, Kerstin C. Maier3, Žiga Avsec1,2, Petra Rus3, Julien Gagneur1,2,\*

1. Department of Informatics, Technical University of Munich, Boltzmannstraße 3, 85748 Garching, Germany

2. Graduate School of Quantitative Biosciences (QBM), Ludwig-Maximilians-Universität München, Germany

3. Department of Molecular Biology, Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, 37077 Göttingen, Germany

\* Correspondence: gagneur@in.tum.de

**Supplemental Results**

**Translation initiation sequence features associate with mRNA stability**

First, longer 5’UTRs associated with less stable mRNAs (ρ = -0.17, *P* < 2 × 10-16 for *S. cerevisiae* and ρ = -0.26, *P* = < 2 × 10-16 for *S. pombe*, Supplemental Fig. S1A, B). In mouse cells, mRNA isoforms with longer 5’UTR are translated with lower efficiency (Wang et al. 2016), possibly because longer 5’UTR generally harbor more translation-repressive elements. Hence, longer 5’UTR may confer mRNA instability by decreasing translation initiation and therefore decreasing the protection by the translation machinery.

Second, a significant association between the third nucleotide 5’ of the start codon and mRNA half-life was observed (Figure 4A, B). The median half-life correlated with the nucleotide frequency at this position (Supplemental Fig. S4B), associating with 1.28 median fold-change (*P* = 1.7x10-11) between the adenosine (2,736 genes, most frequent) and cytosine (360 genes, the least frequent). The same correlation was also significant for *S. pombe* (*P* = 1.2x10-4, Supplemental Fig. S4C, D). Functional effect of the start codon context on mRNA stability has been established as the long-lived *PGK1* mRNA was strongly destabilized when substituting the sequence context around its start codon with the one from the short-lived *MFA2* mRNA (LaGrandeur and Parker 1999). Our genome-wide analysis indicates that this effect generalizes to other genes. The start codon context, which controls translation initiation efficiency (Kozak 1987; Dvir et al. 2013), increases ribosome density which may protect mRNA from degradation as hypothesized by Edri and Tuller (Edri and Tuller 2014).

Finally, de novo search for regulatory motifs identified AAACAAA motif to be significantly (FDR < 0.1) associated with longer half-lives. However, this association might be merely correlative as the motif failed for further support (Supplemental Fig, S5).

Altogether, these findings indicate that 5’UTR elements, including the start codon context, may affect mRNA stability by altering translation initiation.

**Stop codon context associates with mRNA stability**

Linear regression against the 6 bases 5’ and 3’ of the stop codon revealed the first nucleotide 3’ of the stop codon to most strongly associate with mRNA stability. This association was observed for each of the three possible stop codons, and for each codon a cytosine significantly associated with lower half-life (Supplemental Fig. S4F, also for P-values and fold changes). This also held for *S. pombe* (Supplemental Fig. S4G, also for P-values and fold changes). A cytosine following the stop codon structurally interferes with stop codon recognition (Brown et al. 2015), thereby leading to stop codon read-through events (Bonetti et al. 1995). Of all combinations, TGA-C is known to be the leakiest stop codon context (Jungreis et al. 2011) and also associated with shortest mRNA half-life (Supplemental Fig. S4F). These results are consistent with non-stop decay, a mechanism that triggers exosome-dependent RNA degradation when the ribosome translates the poly(A) tail (Guydosh and Green 2017). However, the association between the stop codon context and half-life was not weakened in mutants of the Ski complex, which is required for the cytoplasmic functions of the exosome (Supplemental Fig. S6). These results indicate that the fourth nucleotide after the stop codon is an important determinant of mRNA stability, possibly because of translational read-through. However, the effect was not significantly reduced in mutants of non-stop decay pathway (Ski genes, Supplemental Fig. S6). Further analysis and data would be necessary to assess the role of stop codon context in mRNA stability.

**Additive versus multiplicative models**

The measurement noise for half-life is multiplicative in the natural scale. Indeed, half-lives in Sun et al. were estimated by the ratios of microarray expression of two samples (labeled vs total) and noise in microarrays is essentially multiplicative. Hence, the log transformation of the response variable leads to approximately homoscedastic, additive noise. Note that the same would be true for sequencing-based data (see e.g. (Eser et al. 2016)).

Nonetheless, it is unclear whether CREs act additively or multiplicatively on decay rate in the natural scale. Two simple theories can be conceived:

i) Multiplicative model: Assuming the degradation of an RNA is the outcome of the combination of independent events, the probability of degradation in an interval of time is the product of the probability of each independent event. In this case, taking the logarithm of the rate would lead to a linear model.

ii) Additive model: Assuming in contrast that the degradation of an RNA is the outcome of exclusive events (or independent events unlikely to occur at the same time), the overall degradation rate is the sum of the individual degradation rates.

To test both models, we have fitted the additive model and compared it to the (multiplicative) model. For both we worked with the log half-life as response variable because of the multiplicative noise as explained above. We performed this comparison for the 3’UTR motifs only where we can expect that the number of motifs contribute additively to half-life assuming independent poisson process of the different motifs (so, in favor of the alternative, additive model). Both models have the same number of parameters. The mean squared error was slightly less for the multiplicative model (MSE 0.5028 for the multiplicative model compared to 0.5057 for additive model). Hence, this analysis does not support the additive model.

Because effects are small, working on logarithmic or non-logarithmic transformation probably makes little difference. Moreover, the reality is likely more complex and involve combinations of both models. Future biophysical studies would be interesting to unravel the relationships between the different cis-regulatory elements and their concerted action on RNA stability.

**Supplemental methods**

**Linear model for genome-wide half-life prediction**

1. The joint model for *S. cerevisiae* half-life prediction:

Let $y\_{i}$ be the half-life of transcript $i$:

|  |  |  |
| --- | --- | --- |
|  | $$log⁡(y\_{i})=β\_{0}+\sum\_{f\in features}^{}β\_{f}x\_{f}+ε\_{i}$$ | (1) |

where the covariates $x\_{f}$ for the following features were represented as:

$$x\_{UTR3\\_length}=log⁡(3^{'}UTR length)$$

$$x\_{UTR5\\_length}=log⁡(5^{'}UTR length)$$

$$x\_{CDS\\_deltaG}=free energy of CDS$$

$$x\_{codon\_{k}}=\frac{\#codon\_{k}}{CDS length}$$

$x\_{motif}=\# motif$ (number of perfect match of the consensus motif in 3’UTR region)

$$x\_{uAUG}=\left\{\begin{array}{c}1, \&number of uAUG>0\\0, \&number of uAUG=0\end{array}\right.$$

$$x\_{UTR5\_{GC}}=fraction of GC in 5^{'}UTR$$

2. The joint model for *S. pombe* half-life prediction:

Let $y\_{i}$ be the half-life of transcript $i$:

|  |  |  |
| --- | --- | --- |
|  | $$log⁡(y\_{i})=β\_{0}+\sum\_{f\in features}^{}β\_{f}x\_{f}+ε\_{i}$$ | (2) |

where the covariates $x\_{f}$ for the following features were represented as:

$$x\_{UTR3\\_length}=log⁡(3^{'}UTR length)$$

$$x\_{UTR5\\_length}=log⁡(5^{'}UTR length)$$

$$x\_{CDS\\_length}=log⁡(CDS length)$$

$$x\_{codon\_{k}}=\frac{\#codon\_{k}}{CDS length}$$

$$x\_{stopPlus1}=categorical variable of stop codon+1$$

$x\_{motif}=\# motif$ (number of perfect match of the consensus motif in 3’UTR region)

$$x\_{uAUG}=\left\{\begin{array}{c}1, \&number of uAUG>0\\0, \&number of uAUG=0\end{array}\right.$$

$$x\_{UTR5\_{∆G}}=free energy of 5'UTR$$

$$x\_{CDS\_{∆G}}=free energy of CDS$$

$$x\_{UTR3\_{∆G}}=free energy of 3'UTR$$

$$x\_{UTR5\_{GC}}=fraction of GC in 5^{'}UTR$$

$$x\_{UTR3\_{GC}}=fraction of GC in 3'UTR$$

**Variance explained for linear model**

The percentage of explained variance for a linear model was calculated by$ 100 \*(1 - var\_{residual} / var\_{total})$, where $var\_{residual}$ and $var\_{total}$ are the variance of regression residual and the total half-life variance respectively.

**Single-nucleotide variant effect predictions**

For **Motifs:** Briefly, consensus sequence of the motif is defined as sequence of the most frequent nucleotide at each motif position. Motif feature was created by counting the number of consensus motif matches in the corresponding sequence region allowing for one mismatch (sites). To estimate motif single-nucleotide effect, for each motif, the following linear model is fitted by maximum likelihood:

|  |  |  |
| --- | --- | --- |
|  | $$log⁡(y\_{i})=β\_{0}+\sum\_{f\in features}^{}β\_{f}x\_{f}+\sum\_{j=1}^{n\_{i}}β\_{cons}+\sum\_{k=1}^{L}β\_{k},w\_{i,j,k}+ε\_{i}$$ | (3) |

$y\_{i}$**:** the half-life of transcript $i$:

$x\_{f}:$ all covariates that remains significant in the joint model (codons were consider as a group that was significant) except the motif of evaluate

$n\_{i}:$ the number of sites for the i-th transcript

$β\_{cons}:$ effect of one consensus site

$w\_{i,j,k}:$ k-th nucleotide of the j-th site of the i-th transcripts

$L$: length of a site

For **Other sequence features:** In contrast to the motifs, effectsof single-nucleotide of the other features were done by *in silico* perturbation. For example, the effect of a single CGT to CGA transition was assessed by decreasing the count of CGT for each gene by one while increasing the count of CGA for each gene (only perturbing genes that had at least one CGT). In the case of length and GC content, we decreased the length or number of GC count by one for each gene. Note that the side effects of varying the length, such as disrupting reading-frame, were not considered. Only synonymous transitions were considered for codons.

**Analysis of knockout strains**

The sequence feature effect levels were defined as follows for different classes of sequence features:

uAUG: $\frac{\left(median\left(H\_{with\_{uAUG}}\right)- median\left(H\_{without\_{uAUG}}\right)\right)}{median\left(H\_{without\_{uAUG}}\right)}$ where $H$ stands for half-life.

Motifs: $\frac{\left(median\left(H\_{count=1 }\right)- median\left(H\_{count=0}\right)\right)}{median\left(H\_{count=0}\right)}$ where $H\_{count=0 }$ and $H\_{count=1 }$ stands for the half-life of mRNAs that has zero and one instance of the motif respectively.

Codon usage: For each knockout or wild-type, a linear model was fitted with all coding codons as covariates. The effect size of codon usage (joint effect of all codons) on half-life was defined by the explained variance of out-of-sample predictions.

Start codon -3 position: $\frac{\left(medain\left(H\_{A}\right)-median\left(H\_{c}\right)\right)}{median\left(H\_{A}\right)}$ where $H\_{A}$ and $H\_{C}$ is the half-life of mRNAs with base adenine and cytosine at start codon -3 position respectively.

Stop codon +1 position: $\frac{\left(median\left(H\_{TGAG}\right)-median\left(H\_{TGAC}\right)\right)}{median(H\_{TGAG})}$ where $H\_{TGAG}$ and $H\_{TGAC}$ is the half-life of mRNAs that has stop codon TGA followed by guanine and cytosine respectively.

5’ folding energy: genome-wide Spearman rank correlation between 5’ folding energy and half-life.

**GO enrichment analysis**

Gene ontology enrichment analysis (Biological Processes, BP) for the motif containing genes was performed with PANTHER web interface (http://www.pantherdb.org/).

**Analysis of qPCR data to validate motif**

Let$ x\_{ij}$ be the initial amount of mRNA for gene $i$ in sample $j$, $A\_{j} $be the amount of mRNAs extracted from sample $j$, $c\_{ij}$ be the concentration of gene $i$ in sample $j$.

|  |  |  |
| --- | --- | --- |
|  | $$x\_{ij}=A\_{j}\*c\_{ij}$$ | (4) |

For the PCR reaction:

|  |  |  |
| --- | --- | --- |
|  | $$x\_{ij}(1+η\_{ij})^{Ct\_{ij}}=const$$ | (5) |

where $Ct\_{ij}$ is the measured Ct value. $η\_{ij}$ is the primer efficiency. Take $log2$ at both side of equation (2):

|  |  |  |
| --- | --- | --- |
|  | $log\_{2}\left(x\_{ij}\right)+Ct\_{ij}log\_{2}(1+η\_{ij})= const$  | (6) |
|  |  |  |

$-Ct\_{ij}log\_{2}\left(1+η\_{ij}\right)$ can be seen as the normalized $Ct\_{ij}$ by primer efficiency, we refer to as $y\_{ij}$. Plug in equation (4) to equation (6):

|  |  |  |
| --- | --- | --- |
|  | $$y\_{ij}=const+log\_{2}\left(A\_{j}\right)+log\_{2}(c\_{ij})$$ | (7) |

this can be modeled as a linear regression model. Where $log\_{2}\left(A\_{j}\right)$ is the sample size, $c\_{ij}$ is the normalized gene expression.

We assume that $c\_{ij}$ is a constant across samples for the control genes *ACT1* and *TUB2*, while $c\_{ij}$ for the test genes *SFG1* and *NYV1* varies across samples (with motifs embedded, with scrambled motifs embedded).

**Reporter assay**

To increase the chance for the motifs to function in the assay, we had designed the constructs with the aim of preserving the termination region as much as possible and of avoiding interference between the reporter and the resistance locus:

* The resistance gene and the reporter gene are in convergent configuration. In contrast to tandem configuration, this avoids interference from bidirectional transcription (see validation assay in (Xu et al. 2011))
* There is a 84-bp buffer region between the termination sequence of the resistance gene and the reporter gene.
* After our the annotated transcript ends (Pelechano et al. 2013), we have 10 additional bases to ensure that transcriptional termination of the reporter gene is unaltered.
* We further checked from the raw transcript isoform data (Pelechano et al. 2013) that, the vast majority transcripts of our reporter genes used exactly this transcript 3’end.



**Scheme 1:** Illustration of validation reporter gene constructs.

**References**

Bonetti B, Fu L, Moon J, Bedwell DM. 1995. The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in Saccharomyces cerevisiae. *J Mol Biol* **251**: 334–345.

Brown A, Shao S, Murray J, Hegde RS, Ramakrishnan V. 2015. Structural basis for stop codon recognition in eukaryotes. *Nature* **524**: 493–6.

Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, Segal E. 2013. Deciphering the rules by which 5’-UTR sequences affect protein expression in yeast. *Proc Natl Acad Sci U S A* **110**: E2792-801.

Edri S, Tuller T. 2014. Quantifying the effect of ribosomal density on mRNA stability. *PLoS One* **9**: e102308.

Eser P, Wachutka L, Maier KC, Demel C, Boroni M, Iyer S, Cramer P, Gagneur J. 2016. Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. *Mol Syst Biol* **12**: 857.

Guydosh NR, Green R. 2017. Translation of poly(A) tails leads to precise mRNA cleavage. *RNA* **23**: 749–761.

Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, Kellis M. 2011. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. *Genome Res* **21**: 2096–2113.

Kozak M. 1987. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J Mol Biol* **196**: 947–950.

LaGrandeur T, Parker R. 1999. The cis acting sequences responsible for the differential decay of the unstable MFA2 and stable PGK1 transcripts in yeast include the context of the translational start codon. *RNA* **5**: 420–33.

Pelechano V, Wei W, Steinmetz LM. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**: 127–31.

Wang X, Hou J, Quedenau C, Chen W. 2016. Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. *Mol Syst Biol* **12**: 875.

Xu Z, Wei W, Gagneur J, Clauder-Münster S, Smolik M, Huber W, Steinmetz LM. 2011. Antisense expression increases gene expression variability and locus interdependency. *Mol Syst Biol* **7**: 468.