Sequence features explain most of the mRNA stability variation across genes in yeast

Jun cheng[1,2], Žiga Avsec[1,2], Julien Gagneur[1,2,*]

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

* Correspondence: gagneur@in.tum.de

Abstract

The stability of messenger RNA (mRNA) is one of the major determinants of gene expression. Although a wealth of sequence elements regulating mRNA stability has been described, their quantitative contributions to half-life are unknown. Here, we built a quantitative model for *Saccharomyces cerevisiae* explaining 60% of the half-life variation between genes based on mRNA sequence features alone, and predicts half-life at a median relative error of 30%. The model integrates known cis-regulatory elements, identifies novel ones, and quantifies their contributions at single-nucleotide resolution. We show quantitatively that codon usage is the major determinant of mRNA stability. Nonetheless, single-nucleotide variations have the largest effect when occurring on 3'UTR motifs or upstream AUGs. Application of the approach to *Schizosaccharomyces pombe* supports the generality of these findings. Analyzing the effect of these sequence elements on mRNA half-life data of 34 knockout strains showed that the effect of codon usage not only requires functional decapping and deadenylation, but also the 5'-to-3' exonuclease Xrn1, the non-sense mediated decay proteins Upf2 and Upf3, and does not require no-go decay. Altogether, this study quantitatively delineates the contributions of mRNA sequence features on stability in yeast, reveals their functional dependencies on degradation pathways, and allows accurate prediction of half-life from mRNA sequence.

Author Summary

The stability of mRNA plays a key role in gene regulation: It influences not only the mRNA abundance but also how quickly new steady-state levels are reached upon a transcriptional trigger. How is mRNA half-life encoded in a gene sequence? Through systematic discovery of novel half-life associated sequence elements and collecting known ones, we show that mRNA half-life can be predicted from sequence in yeast, at an accuracy close to measurement precision. Our analysis reveals new conserved motifs in 3'UTRs predictive for half-life. While codon usage appears to be the major determinant of half-life, motifs in 3'UTRs are the most sensitive elements to mutations: a single nucleotide change can affect the half-life of an mRNA by as much as 30%. Analyzing half-life data of knockout strains, we furthermore dissected the dependency of the elements with respect to various mRNA degradation pathways. This revealed the dependency of codon-mediated mRNA stability control to 5'-3' degradation and non-sense mediated decay genes. Altogether, our study is a significant step forward in predicting gene expression from a genome sequence and understanding codon-mediated mRNA stability control.

Introduction

The stability of messenger RNAs is an important aspect of gene regulation. It influences the overall cellular mRNA concentration, as mRNA steady-state levels are the ratio of synthesis and degradation rate. Moreover, low stability confers high turnover to mRNA and therefore the capacity to rapidly reach a new steady-state level in response to a transcriptional trigger (1). Hence, stress genes, which must rapidly respond to environmental signals, show low stability (2,3). In contrast, high stability provides robustness to variations in transcription. Accordingly, a wide range of mRNA-half-lives is observed in eukaryotes, with typical variations in a given genome spanning one to two orders of magnitude (4–6). Also, significant variability in mRNA half-life among human individuals could be demonstrated for about a quarters of genes in lymphoblastoid cells and estimated to account for more than a third of the gene expression variability (7).

How mRNA stability is encoded in a gene sequence has long been a subject of study. Cis-regulatory elements (CREs) affecting mRNA stability are mainly encoded in the mRNA itself. They include but are not limited to secondary structure (8,9), sequence motifs present in the 3'UTR including binding sites of RNA-binding proteins (10–12), and, in higher eukaryotes, microRNAs (13). Moreover, translation-related features are frequently associated with mRNA stability. For instance, inserting strong secondary structure elements in the 5'UTR or modifying the translation start codon context strongly destabilizes the long-lived *PGK1* mRNA in *S. cerevisiae* (14,15). Codon usage, which affects translation elongation rate, also regulates mRNA stability (16–19), Further correlations between codon usage and mRNA stability have been reported in *E. coli* and *S. pombe* (20,21).

Since the RNA degradation machineries are well conserved among eukaryotes, the pathways have been extensively studied using *S. cerevisiae* as a model organism (22,23). The general mRNA degradation pathway starts with the removal of the poly(A) tail by the Pan2/Pan3 (24) and Ccr4/Not complexes (25). Subsequently, mRNA is subjected to decapping carried out by Dcp2 and promoted by several factors including Dhh1 and Pat1 (26,27). The decapped and deadenylated mRNA can be rapidly degraded in the 3' to 5' direction by the exosome (28) or in the 5' to 3' direction by Xrn1 (29). Further mRNA degradation pathways are triggered when aberrant translational status is detected, including Nonsense-mediated decay (NMD), No-go decay (NGD) and Non-stop decay (NSD) (22,23).

Despite all this knowledge, prediction of mRNA half-life from a gene sequence is still not established. Moreover, most of the mechanistic studies so far could only be performed on individual genes or reporter genes and it is therefore unclear how the effects generalize genome-wide. A recent study showed that translation-related features can be predictive for mRNA stability (30). Although this analysis supported the general correlation between translation and stability (31,32), the model was not based purely on sequence-derived features but contained measured transcript properties such as ribosome density and normalized translation efficiencies. Hence, the question of how half-life is genetically encoded in mRNA sequence remains to be addressed. Additionally, the dependences of sequence features to distinct mRNA degradation pathways have not been systematically studied. One example of this is codon-mediated stability control. Although a causal link from codon usage to mRNA half-life has been shown in a wide range of organisms (16–19), the underlying mechanism remains poorly understood. In *S. cerevisiae*, reporter gene experiments showed that codon-mediated stability control depends on the RNA helicase Dhh1 (33). However, neither is it clear how this generalizes genome-wide nor the role of other closely related genes has been systematically assessed.

Here, we used an integrative approach where we mathematically modelled mRNA half-life as a function of its sequence and applied it to *S .cerevisiae*. For the first time, our model can explain most of the between-gene half-life variance from sequence alone. Using a semi-mechanistic model, we could interpret individual sequence features in the 5'UTR, coding region, and 3'UTR. Our approach *de novo* recovered known cis-regulatory elements and identified novel ones. Quantification of the respective contributions revealed that codon usage is the major contributor to mRNA stability. Applying the modeling approach to *S. pombe* supports the generality of these findings. We systematically assessed the dependencies of these sequence elements on mRNA degradation pathways using half-life data for 34 knockout strains, and notably delineated novel pathways through which codon usage affects half-life.

Results

Regression reveals novel mRNA sequence features associated with mRNA stability

To study cis-regulatory determinants of mRNA stability in *S. cerevisiae*, we chose the dataset by Sun and colleagues (34), which provides genome-wide half-life measurements for 4,388 expressed genes of a wild-type lab strain and 34 strains knocked out for RNA degradation pathway genes (Fig 1, S1 Table). When applicable, we also investigated half-life measurements of *S. pombe* for 3,614 expressed mRNAs in a wild-type lab strain from Eser and colleagues (6). We considered sequence features within 5 overlapping regions: the 5'UTR, the start codon context, the coding sequence, the stop codon context and the 3'UTR. The correlations between sequence lengths, GC contents and folding energies (Materials and Methods) with half-life and corresponding P-values are summarized in S2 Table and S1-S3 Figs. In general, sequence lengths correlated negatively with half-life and folding energies correlated positively with half-life in both yeast species, whereas correlations of GC content varied with species and gene regions.

Motif search (Materials and Methods) recovered *de novo* the Puf3 binding motif TGTAAATA in 3'UTR (35,36), a well-studied CRE that confers RNA instability, a polyU motif (TTTTTTA), which is likely bound by the mRNA-stabilizing protein Pub1 (12), as well as the Whi3 binding motif TGCAT (37,38). Two new motifs were found: AAACAAA in 5'UTR, and ATATTC in 3'UTR (Fig 2A). Except for AAACAAA and TTTTTTA, all motifs associated with shorter half-lives (Fig 2A). Notably, the motif ATATTC, was found in 13% of the genes (591 out of 4,388) and significantly co-occurred with the other two destabilizing motifs found in 3'UTR: Puf3 (FDR = 0.02) and Whi3 (FDR = 7× 10^{-3}) binding motifs (Fig 2B).

In the following subsections, we describe first the findings for each of the 5 gene regions and then a model that integrates all these sequence features.

Upstream AUGs destabilize mRNAs by triggering nonsense-mediated decay

Occurrence of an upstream AUG (uAUG) associated significantly with shorter half-life (median fold-change = 1.37, $P < 2 \times 10^{-16}$). This effect strengthened for genes with two or more AUGs (Fig 3A, B). Among the 34 knock-out strains, the association between uAUG and shorter half-life was almost lost only for mutants of the two essential components of the nonsense-mediated mRNA decay (NMD) *UPF2* and *UPF3* (39,40), and for the general 5'-3' exonuclease *Xrn1* (Fig 2A). The dependence on NMD suggested that the association might be due to the occurrence of a premature stop codon. Consistent with this hypothesis, the association of uAUG with decreased half-lives was only found for genes with a premature stop codon cognate with the uAUG (Fig 3C). This held not only for cognate premature stop codons within the 5'UTR, leading to a potential upstream ORF, but also for cognate premature stop codons within the ORF, which occurred almost always for uAUG out-of-frame with the main ORF (Fig 3C). This finding likely holds for many other eukaryotes as we found the same trends in *S. pombe* (Fig 3D). These observations are consistent with a single-gene study demonstrating that translation of upstream ORFs can lead to

RNA degradation by nonsense-mediated decay (41). Altogether, these results show that uAUGs are mRNA destabilizing elements as they almost surely match with a cognate premature stop codon, which, whether in frame or not with the gene, and within the UTR or in the coding region, trigger NMD.

Translation initiation predicts mRNA stability

Several sequence features in the 5'UTR associated significantly with mRNA half-life.

First, longer 5'UTRs associated with less stable mRNAs (ρ = -0.17, $P < 2 \times 10^{-16}$ for *S. cerevisiae* and ρ = -0.26, $P = < 2 \times 10^{-16}$ for *S. pombe*, S1A, B Fig). In mouse cells, mRNA isoforms with longer 5'UTR are translated with lower efficiency (42), 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 (Fig 4A). The median half-life correlated with the nucleotide frequency at this position (S4A Fig), associating with 1.28 median fold-change ($P = 1.7\text{x}10^{-11}$) between the adenosine (2,736 genes, most frequent) and cytosine (360 genes, the least frequent). The same correlation was also significant for *S. pombe* ($P = 1.2\text{x}10^{-4}$, S4A, B Fig). 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 (15). Our genome-wide analysis indicates that this effect generalizes to other genes. The start codon context, which controls translation initiation efficiency (43,44), increases ribosome density which may protect mRNA from degradation as hypothesized by Edri and Tuller (31).

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 (S5 Fig).

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

Codon usage regulates mRNA stability through common mRNA decay pathways

First, species-specific tRNA adaptation index (sTAI) (45) significantly correlated with half-life in both *S. cerevisiae* (Fig 4C, $\rho = 0.55$, $P < 2.2\text{x}10^{-16}$) and *S. pombe* (Fig S4C, $\rho = 0.41$, $P < 2.\ 2\text{x}10^{-16}$), confirming previously observed association between codon optimality and mRNA stability (17,21). Next, using the out-of-folds explained variance as a summary statistics, we assessed its variation across different gene knockouts (Materials and Methods). The effect of codon usage exclusively depended on the genes from the common deadenylation- and decapping-dependent 5' to 3' mRNA decay pathway and the NMD pathway (all FDR < 0.1, Fig 4C). In particular, all assayed genes of the Ccr4-Not complex, including *CCR4*, *NOT3*, *CAF40* and *POP2*, were required for wild-type level effects of codon usage on mRNA decay. Among them, *CCR4* has the largest effect. This confirmed a recent study in zebrafish showing that accelerated decay of non-optimal codon genes requires deadenylation activities of Ccr4-Not (18). In contrast to genes of the Ccr4-Not complex, *PAN2/3* genes which encode also deadenylation enzymes, were not found to be essential for the coupling between codon usage and mRNA decay (Fig 4C).

Furthermore, our results not only confirm the dependence on Dhh1 (33), but also on its interacting partner Pat1. Our findings of Pat1 and Ccr4 contradict the negative results for these genes reported by Radhakrishnan *et al.* (33). The difference might come from the fact that our analysis is genome-wide, whereas Radhakrishnan and colleagues used a reporter assay.

Our systematic analysis revealed two additional novel dependencies: First, on the common 5' to 3' exonuclease Xrn1, and second, on *UPF2 and UPF3* genes, which are essential players of NMD (all FDR < 0.1, Fig 4C). Previous studies have shown that NMD is more than just a RNA surveillance pathway, but rather one of the general mRNA decay mechanisms that target a wide range of mRNAs, including aberrant and normal ones (46,47). Notably, we did not observe any change of effect upon knockout of *DOM34* and *HBS1* (S6 Fig), which are essential genes for the No-Go decay pathway. This implies that the effect of codon usage is unlikely due to stalled ribosomes at non-optimal codons.

Altogether, our analysis strongly indicates that, the so-called “codon-mediated decay” is not an mRNA decay pathway itself, but a regulatory mechanism of the common mRNA decay pathways.

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 (all $P < 0.01$, Fig 4D). This also held for *S. pombe* (all $P < 0.01$, S4D Fig). A cytosine following the stop codon structurally interferes with stop codon recognition (48), thereby leading to stop codon read-through events (49). Of all combinations, TGA-C is known to be the leakiest stop codon context (50) and also associated with shortest mRNA half-life (Fig 4D). These results are consistent with non-stop decay, a mechanism that triggers exosome-dependent RNA degradation when the ribosome reaches the poly(A) tail. Consistent with this interpretation, mRNAs with additional in-frame stop codons in the 3’UTR, which are over-represented in yeast (51), exhibited significantly higher half-life ($P = 7.5\text{x}10^{-5}$ for *S. cerevisiae* and $P = 0.011$ for *S. pombe*, S4E, F Fig). 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 (S6 Fig). These results indicate that the fourth nucleotide after the stop codon is an important determinant of mRNA stability, likely because of translational read-through.

Sequence motifs in 3’UTR

Four motifs in the 3’UTR were found to be significantly associated with mRNA stability (Fig 5A, all FDR < 0.1, Materials and Methods). This analysis recovered three described motifs: the Puf3 binding motif TGTAAATA (35), the Whi3 binding motif TGCAT (37,38), and a poly(U) motif TTTTTTA, which can be bound by Pub1 (12), or is part of the long poly(U) stretch that forms a looping structure with poly(A) tail (9). We also identified a novel motif, ATATTC, which associated with lower mRNA half-life. This motif was reported to be enriched in 3'UTRs for a cluster of genes with correlated expression pattern (52), but its function remains unknown. Genes harboring this motif are significantly enriched for genes involved in oxidative phosphorylation (Bonferroni corrected $P < 0.01$, Gene Ontology analysis, Supplementary Methods and S3 Table).

Four lines of evidence supported the potential functionality of the new motif. First, it preferentially localizes in the vicinity of the poly(A) site (Fig 5B), and functionally depends on Ccr4 (S6 Fig), suggesting a potential interaction with deadenylation factors. Second, single nucleotide deviations from the consensus sequence of the motif associated with decreased effects on half-life (Fig 5C, linear regression allowing for one mismatch, Materials and Methods). Moreover, the flanking nucleotides did not show further associations indicating that the whole lengths of the motifs were recovered (Fig 5C). Third, when allowing for one mismatch, the motif still showed strong preferences (Fig 5D). Fourth, the motif instances were more conserved than their flanking bases (Fig 5E).

Consistent with the role of Puf3 in recruiting deadenylation factors, Puf3 binding motif localized preferentially close to the poly(A) site (Fig 5B). The effect of the Puf3 motifs was significantly lower in the knockout of *PUF3* (FDR < 0.1, S6 Fig). We also found a significant dependence on the deadenylation (*CCR4*, *POP2*) and decapping (*DHH1*, *PAT1*) pathways (all FDR < 0.1, S6 Fig), consistent with previous single gene experiment showing that Puf3 binding promotes both deadenylation and decapping (10,53). Strikingly, Puf3 binding motif switched to a stabilization motif in the absence of Puf3 and Ccr4, suggesting that deadenylation of Puf3 motif containing mRNAs is not only facilitated by Puf3 binding, but also depends on it.

Whi3 plays an important role in cell cycle control (54). Binding of Whi3 leads to destabilization of the *CLN3* mRNA (38). A subset of yeast genes are up-regulated in the Whi3 knockout strain (38). However, it was so far unclear whether Whi3 generally destabilizes mRNAs upon its binding. Our analysis showed that mRNAs containing the Whi3 binding motif (TGCAT) have significantly shorter half-life (FDR = $6.9\text{x}10^{-04}$). Surprisingly, this binding motif is extremely widespread, with 896 out of 4,388 (20%)

genes that we examined containing the motif on the 3'UTR region, which enriched for genes involved in several processes (S3 Table). No significant genetic dependence of the effect of the Whi3 binding motif was found (S6 Fig).

The mRNAs harboring the TTTTTTA motif tended to be more stable and enriched for translation (*P* = 1.34x10^{-03}, S3 Table, Fig 5A). No positional preferences were observed for this motif (Fig 5B). Effects of this motif depends on genes from Ccr4-Not complex and Xrn1 (S6 Fig).

60% between-gene half-life variation can be explained by sequence features

We next asked how well one could predict mRNA half-life from these mRNA sequence features, and what their respective contributions were when considered jointly. To this end, we performed a multivariate linear regression of the logarithm of the half-life against the identified sequence features. The predictive power of the model on unseen data was assessed using 10-fold cross validation (Material and Methods). Also, motif discovery performed on each of the 10 training sets retrieved the same set of motifs, showing that their identification was not due to over-fit on the complete dataset. Altogether, 60% of *S*. *cerevisiae* half-life variance in the logarithmic scale can be explained by simple linear combinations of the above sequence features (Fig 6A). The median out-of-folds relative error across genes is 30%. A median relative error of 30% for half-life is remarkably low because it is in the order of magnitude of the expression variation that is typically physiologically tolerated, and it is also about the amount of variation observed between replicate experiments (6). To make sure that our findings are not biased to a specific dataset, we fitted the same model to a dataset using RATE-seq (55), a modified version of the protocol used by Sun and colleagues (34). On this data, the model was able to explain 50% of the variance (S7 Fig). Moreover, the same procedure applied to *S. pombe* explained 47% of the total half-life variance, suggesting the generality of this approach. Because the measures also entail measurement noise, these numbers are conservative underestimations of the total biological variance explained by our model.

The uAUG, 5'UTR length, 5'UTR GC content, 61 coding codons, CDS length, all four 3'UTR motifs, and 3'UTR length remained significant in the joint model indicating that they contributed independently to half-life (complete list of p-values given in S4 Table). In contrast, start codon context, stop codon context, 5' folding energy, the 5'UTR motif AAACAAA, and 3'UTR GC content dropped below the significance when considered in the joint model (Materials and Methods). This loss of statistical significance may be due to lack of statistical power. Another possibility is that the marginal association of these sequence features with half-life is a consequence of a correlation with other sequence features. Among all sequence features, codon usage as a group is the best predictor both in a univariate model (55.23%) and in the joint model (43.84 %) (Fig 6C). This shows that, quantitatively, codon usage is the major determinant of mRNA stability in yeast.

The variance analysis quantifies the contribution of each sequence feature to the variation across genes. Features that vary a lot between genes, such as UTR length and codon usage, favorably contribute to the variation. However, this does not reflect the effect on a given gene of elementary sequence variations in these features. For instance, a single-nucleotide variant can lead to the creation of an uAUG with a strong effect on half-life, but a single nucleotide variant in the coding sequence may have little impact on overall codon usage. We used the joint model to assess the sensitivity of each feature to single-nucleotide mutations as median fold-change across genes, simulating single-nucleotide deletions for the length features and single nucleotide substitutions for the remaining ones (Materials and Methods). Single-nucleotide variations typically altered half-life by less than 10%. The largest effects were observed in the 3'UTR motifs and uAUG (Fig 6D). Notably, although codon usage was the major contributor to the variance, synonymous variation on codons typically affected half-life by less than 2% (Fig 6D; S8 Fig). For those synonymous variations that changed half-life by more than 2%, most of them were variations that involved the most non-optimized codons CGA or ATA (S8 Fig, Presnyak et al. 2015).

Altogether, our results show that most of yeast mRNA half-life variation can be predicted from mRNA sequence alone, with codon usage being the major contributor.

However, single-nucleotide variation at 3’UTR motifs or uAUG had the largest expected effect on mRNA stability.

Discussion

We systematically searched for mRNA sequence features associating with mRNA stability and estimated their effects at single-nucleotide resolution in a joint model. Overall, the joint model showed that 60% of the variance could be predicted from mRNA sequence alone in *S. cerevisiae*. This analysis showed that translation-related features, in particular codon usage, contributed most to the explained variance. This findings strengthens further the importance of the coupling between translation and mRNA degradation (56–58). Moreover, we assessed the RNA degradation pathway dependencies of each sequence feature. Remarkably, we identified that codon-mediated decay is a regulatory mechanism of the canonical decay pathways, including deadenylation- and decapping-dependent 5’ to 3’ decay and NMD (Fig 6E).

Integrative analyses of cis-regulatory elements on various aspects of gene expression (59,60) as we used here complement mechanistic single-gene studies for important aspects. They allow assessing genome-wide the importance of CREs that have been reported previously with single-gene experiments. Also, single-nucleotide effect prediction can more precisely supports the interpretation of genetic variants, including mutations in non-coding region as well as synonymous transitions in coding region. Furthermore, such integrative analyses can be combined with a search for novel sequence features, as we did here with k-mers, allowing the identification of novel candidate cis-regulatory elements. An alternative approach to the modeling of endogenous sequence is to use large-scale perturbation screens (1,44,61). Although very powerful to dissect known cis-regulatory elements or to investigate small variations around select genes, the sequence space is so large that these large-scale perturbation screens cannot uncover all regulatory motifs. It would be interesting to combine both approaches and design large-scale validation experiments guided by insights coming from modeling of endogenous sequences as we developed here.

Recently, Neymotin and colleagues (30) showed that several translation-related transcript properties associated with half-life. This study derived a model explaining 50% of the total variance using many transcript properties including some not based on sequence (ribosome profiling, expression levels, etc.). Although non-sequence based predictors can facilitate prediction, they may do so because they are consequences rather than causes of half-life. For instance increased half-life causes higher expression level. Also, increased cytoplasmic half-life, provides a higher ratio of cytoplasmic over nuclear RNA, and thus more RNAs available to ribosomes. Hence both expression level and ribosome density may help making good predictions of half-life, but not necessarily because they causally increase half-life. In contrast, we aimed here to understand how mRNA half-life is encoded in mRNA sequence. Our model was therefore solely based on mRNA sequence. This avoided using transcript properties which could be consequences of mRNA stability. Hence, our present analysis confirms the quantitative importance of translation in determining mRNA stability that Neymotin and colleagues quantified, and anchors it into pure sequence elements.

Causality cannot be proven through a regression analysis approach. Genes under selection pressure for high expression levels could evolve to have both CREs for high mRNA stability and CREs for high translation rate. When possible, we referred to single gene studies that had proven causal effects on half-life. For novel motifs, we provided several complementary analyses to further assess their potential functionality. These include conservation, positional preferences, and epistasis analyses to assess the dependencies on RNA degradation pathways. The novel half-life associated motif ATATTC in 3'UTR is strongly supported by these complementary analyses and is also significant in the joint model ($P = 5.8x10^{-14}$). One of the most interesting sequence features that we identified but still need to be functionally assayed is the start codon context. Given its established effect on translation initiation (44,62), the general coupling between translation and mRNA degradation (56–58), as well as several observations directly on mRNA stability for single genes (15,63), they are very likely to be functional on most genes. Consistent with this hypothesis, large scale experiments that perturb 5' sequence secondary structure and start codon context indeed showed a wide range of mRNA level changes in the direction that we would predict (44). Altogether, such integrative approaches allow the identification of candidate regulatory elements that could be functionally tested later on.

We are not aware of previous studies that systematically assessed the effects of cis-regulatory elements in the context of knockout backgrounds, as we did here. This part of our analysis turned out to be very insightful. By assessing the dependencies of codon usage mediated mRNA stability control systematically and comprehensively, we generalized results from recent studies on the Ccr4-Not complex and Dhh1, but also identified important novel ones including NMD factors, Pat1 and Xrn1. With the growing availability of knockout or mutant background in model organisms and human cell lines, we anticipate this approach to become a fruitful methodology to unravel regulatory mechanisms.

Materials and Methods

Data and Genomes

Wild-type and knockout genome-wide *S. cerevisiae* half-life data were obtained from Sun and colleagues (34), whereby all strains are histidine, leucine, methionine and uracil auxotrophs. *S. cerevisiae* gene boundaries were taken from the boundaries of the most abundant isoform quantified by Pelechano and colleagues (64). Reference genome fasta file and genome annotation were obtained from the Ensembl database (release 79). UTR regions were defined by subtracting out gene body (exon and introns from the Ensembl annotation) from the gene boundaries.

Genome-wide half-life data of *S. pombe* as well as refined transcription unit annotation were obtained from Eser and colleagues (6). Reference genome version ASM294v2.26 was used to obtain sequence information. Half-life outliers of *S. pombe* (half-life less than 1 or larger than 250 mins) were removed.

For both half-life datasets, only mRNAs with mapped 5'UTR and 3'UTR were considered. mRNAs with 5'UTR length shorter than 6nt were further filtered out.

Codon-wise species-specific tRNA adaptation index (sTAI) of yeasts were obtained from Sabi and Tuller (45). Gene-wise sTAIs were calculated as the geometric mean of sTAIs of all its codons (stop codon excluded).

Analysis of knockout strains

The effect level of an individual sequence feature was compared against the wild-type with Wilcoxon rank-sum test followed by multiple hypothesis testing p-value correction (FDR < 0.1). For details see Supplementary methods.

Motif discovery

Motif discovery was conducted for the 5'UTR, the CDS and the 3'UTR regions. A linear mixed effect model was used to assess the effect of each individual k-mer while controlling the effects of the others and for the region length as a covariate as described previously (Eser et al. 2016). For CDS we also used codons as further covariates. In contrast to Eser and colleagues, we tested the effects of all possible k-mers with length from 3 to 8. The linear mixed model for motif discovery was fitted with GEMMA software (65). P-values were corrected for multiple testing using Benjamini-Hochberg's FDR. Motifs were subsequently manually assembled based on overlapping significant (FDR < 0.1) k-mers.

Folding energy calculation

RNA sequence folding energy was calculated with RNAfold from ViennaRNA version 2.1.9 (66), with default parameters.

S. cerevisiae conservation analysis

The phastCons (67) conservation track for *S. cerevisiae* was downloaded from the UCSC Genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/sacCer3/phastCons7way/). Motif single-nucleotide level conservation scores were computed as the mean conservation score of each nucleotide (including 2 extended nucleotide at each side of the motif) across all motif instances genome-wide (removing NA values).

Linear model for genome-wide half-life prediction

Multivariate linear regression models were used to predict genome-wide mRNA half-life on the logarithmic scale from sequence features. Only mRNAs that contain all features were used to fit the models, resulting with 3,862 mRNAs for *S. cerevisiae* and 3,130 mRNAs for *S. pombe*. Out-of-fold predictions were applied with 10-fold cross validation for any prediction task in this study. For each fold, a linear model was first fitted to the training data with all sequence features as covariates, then a stepwise model selection procedure was applied to select the best model with Bayesian Information Criterion as criteria (*step* function in R, with k = log(n)). L1 or L2 regularization were not necessary, as they did not improve the out-of-fold prediction accuracy (tested with glmnet R package (68)). Motif discovery was performed again at each fold. The same set of motifs were identified within each training set only. A complete list of model features and their p-values in a joint model for both yeast species are provided in S4 Table. For details see Supplementary methods.

Analysis of sequence feature contribution

Linear models were first fitted on the complete data with all sequence features as covariates, non-significant sequence features were then removed from the final models, ending up with 70 features for *S. cerevisiae* model and 75 features for *S. pombe* (each single coding codon was fitted as a single covariate). A complete list of selected significant features and their p-values in a joint model were provided in S4 Table. The contribution of each sequence feature was analyzed individually as a univariate regression and also jointly in a multivariate regression model. The contribution of each feature *individually* was calculated as the variance explained by a univariate model. Features were then added in a descending order of their individual explained variance to a joint model, *cumulative* variance explained were then calculated. The *drop* quantify the drop of variance explained as leaving out one feature separately from the full model. All contributions statistics were quantified by taking the average of 100 times of 10-fold cross-validation.

Single-nucleotide variant effect predictions

The same model that used in sequence feature contribution analysis was used for single-nucleotide variant effect prediction. For **motifs,** effects of single-nucleotide variants were predicted with linear model modified from (6). When assessing the effect of a given motif variation, instead of estimating the marginal effect size, we controlled for the effect of all other sequence features using a linear model with the other features as covariates. For details see Supplementary methods. For **other sequence features,** effects of single-nucleotide variants were predicted by introducing a single nucleotide perturbation into the full prediction model for each gene, and summarizing the effect with the median half-life change across all genes. For details see Supplementary methods.

Code availability

Analysis scripts are available at: https://i12g-gagneurweb.in.tum.de/gitlab/Cheng/mRNA_half_life_public.

Acknowledgements

We are thankful to Kerstin Maier (Max Planck Institute for Biophysical Chemistry) and Fabien Bonneau (Max Planck Institute of Biochemistry) for helpful discussions on motifs and RNA degradation pathways, as well as useful feedback on the manuscript. We thank Björn Schwalb for communication on analyzing the knockout data. We thank Vicente Yépez for useful feedback on the manuscript.

Funding

JC and ŽA are supported by a DFG fellowship through QBM. JG was supported by the Bundesministerium für Bildung und Forschung, Juniorverbund in der Systemmedizin “mitOmics” (grant FKZ 01ZX1405A).

References

1. Shalem O, Dahan O, Levo M, Martinez MR, Furman I, Segal E, et al. Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. Mol Syst Biol. 2008;4(223):223.
2. Zeisel A, Köstler WJ, Molotski N, Tsai JM, Krauthgamer R, Jacob-Hirsch J, et al. Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. Mol Syst Biol. 2011;7(529):529.
3. Rabani M, Raychowdhury R, Jovanovic M, Rooney M, Stumpo DJ, Pauli A, et al. High-Resolution Sequencing and Modeling Identifies Distinct Dynamic RNA Regulatory Strategies. Cell. Elsevier Inc.; 2014;159(7):1698–710.
4. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature. 2011;473(7347):337–42.

5. Schwalb B, Michel M, Zacher B, Demel C, Tresch A, Gagneur J, et al. TT-seq maps the human transient transcriptome. 2016;352(6290).

6. Eser P, Wachutka L, Maier KC, Demel C, Boroni M, Iyer S, et al. Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. Mol Syst Biol. EMBO Press; 2016 Jan 1;12(2):857.

7. Duan J, Shi J, Ge X, Dölken L, Moy W, He D, et al. Genome-wide survey of interindividual differences of RNA stability in human lymphoblastoid cell lines. Sci Rep. 2013;3:1318.

8. Rabani M, Kertesz M, Segal E. Computational prediction of RNA structural motifs involved in post-transcriptional regulatory processes. Proc Natl Acad Sci. 2008;105(39):14885–90.

9. Geisberg J V., Moqtaderi Z, Fan X, Ozsolak F, Struhl K. Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. Cell. Elsevier Inc.; 2014;156(4):812–24.

10. Olivas W, Parker R. The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast. 2000;19(23).

11. Shalgi R, Lapidot M, Shamir R, Pilpel Y. A catalog of stability-associated sequence elements in 3' UTRs of yeast mRNAs. Genome Biol. 2005;6(10):1.

12. Duttagupta R, Tian B, Wilusz CJ, Danny T, Soteropoulos P, Ouyang M, et al. Global Analysis of Pub1p Targets Reveals a Coordinate Control of Gene Expression through Modulation of Binding and Stability. 2005;25(13):5499–513.

13. Lee RC, Feinbaum RL, Ambros V. The C . elegans Heterochronic Gene lin-4 Encodes Small RNAs with Antisense Complementarity to & II-14. 1993;75:843–54.

14. Muhlrad D, Decker CJ, Parker R. Turnover Mechanisms of the Stable Yeast PGK1 mRNA. 1995;15(4):2145–56.

15. LaGrandeur T, Parker R. 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. 1999;5(3):420–33.

16. Hoekema A, Kastelein RA, Vasser M, de Boer HA. Codon replacement in the PGK1 gene of Saccharomyces cerevisiae: experimental approach to study the role of biased codon usage in gene expression. Mol Cell Biol. 1987 Aug 1;7(8):2914–24.

17. Presnyak V, Alhusaini N, Chen Y-H, Martin S, Morris N, Kline N, et al. Codon Optimality Is a Major Determinant of mRNA Stability. Cell. 2015;160(6):1111–24.

18. Mishima Y, Tomari Y. Codon Usage and 3′ UTR Length Determine Maternal mRNA Stability in Zebrafish. Mol Cell. Elsevier Inc.; 2016;61(6):874–85.

19. Bazzini AA, Viso F, Moreno-mateos MA, Johnstone TG, Charles E. Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. EMBO J. 2016;

20. Boël G, Letso R, Neely H, Price WN, Wong K, Su M, et al. Codon influence on protein expression in E. coli correlates with mRNA levels. Nature. Nature Publishing Group; 2016;529(7586):358–63.

21. Harigaya Y, Parker R. Analysis of the association between codon optimality and mRNA stability in Schizosaccharomyces pombe. BMC Genomics. BMC Genomics; 2016;1–16.

22. Garneau NL, Wilusz J, Wilusz CJ. The highways and byways of mRNA decay. Nat Rev Mol Cell Biol. 2007;8(2):113–26.

23. Parker R. RNA degradation in Saccharomyces cerevisae. Genetics. 2012;191(3):671–702.

24. Brown CE, Tarun SZ, Boeck R, Sachs AB. PAN3 Encodes a Subunit of the Pab1p-Dependent Poly(A) Nuclease in Saccharomyces cerevisiae. 1996;16(10):5744–53.

25. Tucker M, Valencia-sanchez MA, Staples RR, Chen J, Denis CL, Parker R. The transcription factor associated Ccr4 and Caf1 proteins are components of the major cytoplasmic mRNA deadenylase in Saccharomyces cerevisiae. Cell. 2001;104:377–86.

26. Pilkington GR, Parker R. Pat1 contains distinct functional domains that promote P-body assembly and activation of decapping. Mol Cell Biol. 2008;28(4):1298–312.

27. She M, Decker CJ, Svergun DI, Round A, Chen N, Muhlrad D, et al. Structural Basis of Dcp2 Recognition and Activation by Dcp1. Mol Cell. 2008;29(3):337–49.

28. Anderson JSJ, Parker P. The 3’ to 5’ degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the SKI2 DEVH box protein and 3’ to 5’ exonucleases of the exosome complex. EMBO J. 1998 Mar 2;17(5):1497–

506.

29. Hsu CL, Stevens A. Yeast cells lacking 5'-3' exoribonuclease 1 contain mRNA species that are poly(A) deficient and partially lack the 5' cap structure. Mol Cell Biol. 1993;13(8):4826–35.

30. Neymotin B, Ettore V, Gresham D. Multiple Transcript Properties Related to Translation Affect mRNA Degradation Rates in Saccharomyces cerevisiae. G3 Genes| Genomes| Genet. 2016;6(November):3475–83.

31. Edri S, Tuller T. Quantifying the effect of ribosomal density on mRNA stability. PLoS One. 2014;9(7):e102308.

32. Lackner DH, Beilharz TH, Marguerat S, Mata J, Watt S, Schubert F, et al. A Network of Multiple Regulatory Layers Shapes Gene Expression in Fission Yeast. Mol Cell. 2007;26(1):145–55.

33. Radhakrishnan A, Chen Y-H, Martin S, Alhusaini N, Green R, Coller J. The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. Cell. Elsevier Inc.; 2016;167(1):122–32.

34. Sun M, Schwalb B, Pirkl N, Maier KC, Schenk A, Failmezger H, et al. Global analysis of Eukaryotic mRNA degradation reveals Xrn1-dependent buffering of transcript levels. Mol Cell. 2013;52(1):52–62.

35. 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).

36. Gupta I, Clauder-Münster S, Klaus B, Järvelin AI, Aiyar RS, Benes V, et al. Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions. Mol Syst Biol. 2014;10:719.

37. Colomina N, Ferrezuelo F, Wang H, Aldea M, Garí E. Whi3, a developmental regulator of budding yeast, binds a large set of mRNAs functionally related to the endoplasmic reticulum. J Biol Chem. 2008;283(42):28670–9.

38. Cai Y, Futcher B. Effects of the Yeast RNA-Binding Protein Whi3 on the Half-Life and Abundance of CLN3 mRNA and Other Targets. PLoS One. 2013;8(12):e84630.

39. Leeds P, Wood JM, Lee B, Culbertson MR. Gene Products That Promote mRNA Turnover in Saccharomyces cerevisiaet. Mol Cell Biol. 1992;12(5):2165–77.

40. Cui Y, Hagan KW, Zhang S, Peltz SW. Identification and characterization of

genes that are required for the accelerated degradation of mRNAs containing a premature translational termination codon. Genes Dev. Cold Spring Harbor Laboratory Press; 1995 Feb 15;9(4):423–36.

41. Gaba A, Jacobson A, Sachs MS. Ribosome occupancy of the yeast CPA1 upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. Mol Cell. 2005;20(3):449–60.

42. Wang X, Hou J, Quedenau C, Chen W, Angelini C, Canditiis D De, et al. Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. Mol Syst Biol. EMBO Press; 2016 Jul 18;12(7):875.

43. Kozak M. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J Mol Biol. 1987;196(4):947–50.

44. Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, et al. Deciphering the rules by which 5 ’-UTR sequences affect protein expression in yeast. Proc Natl Acad Sci. 2013;110(30):E2792–801.

45. Sabi R, Tuller T. Modelling the Efficiency of Codon-tRNA Interactions Based on Codon Usage Bias. DNA Res. 2014;21(6):511–26.

46. He F, Li X, Spatrick P, Casillo R, Dong S, Jacobson A. Genome-Wide Analysis of mRNAs Regulated by the Nonsense-Mediated and 5’ to 3’ mRNA Decay Pathways in Yeast. Mol Cell. 2003;12(6):1439–52.

47. Hug N, Longman D, Cáceres JF. Mechanism and regulation of the nonsense-mediated decay pathway. Nucleic Acids Res. 2015;44(4):1483–95.

48. Brown A, Shao S, Murray J, Hegde RS, Ramakrishnan V. Structural basis for stop codon recognition in eukaryotes. Nature. 2015;524(7566):493–6.

49. Bonetti B, Fu L, Moon J, Bedwell DM. The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in Saccharomyces cerevisiae. J Mol Biol. 1995 Aug 18;251(3):334–45.

50. Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, et al. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. Genome Res. 2011;21(12):2096–113.

51. Williams I, Richardson J, Starkey A, Stansfield I. Genome-wide prediction of stop codon readthrough during translation in the yeast Saccharomyces cerevisiae. Nucleic Acids Res. 2004;32(22):6605–16.

52. Elemento O, Slonim N, Tavazoie S. A Universal Framework for Regulatory

Element Discovery across All Genomes and Data Types. Mol Cell. 2007;28(2):337–50.

53. Goldstrohm AC, Seay DJ, Hook BA, Wickens M. PUF protein-mediated deadenylation is catalyzed by Ccr4p. J Biol Chem. 2007;282(1):109–14.

54. Garí E, Volpe T, Wang H, Gallego C, Futcher B, Aldea M. Whi3 binds the mRNA of the G1 cyclin CLN3 to modulate cell fate in budding yeast. Genes Dev. 2001;15(21):2803–8.

55. Neymotin B, Athanasiadou R, Gresham D. Determination of in vivo RNA kinetics using RATE-seq. RNA. 2014;20(10):1645–52.

56. Roy B, Jacobson A. The intimate relationships of mRNA decay and translation. Trends Genet. Elsevier Ltd; 2013;29(12):691–9.

57. Huch S, Nissan T. Interrelations between translation and general mRNA degradation in yeast. Wiley Interdiscip Rev RNA. 2014;5(6):747–63.

58. Radhakrishnan A, Green R. Connections underlying translation and mRNA stability. J Mol Biol. Elsevier B.V.; 2016;

59. Vogel C, Abreu R de S, Ko D, Le S-YY, Shapiro BA, Burns SC, et al. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol. Nature Publishing Group; 2010;6(1):400.

60. Zur H, Tuller T. Transcript features alone enable accurate prediction and understanding of gene expression in S. cerevisiae. BMC Bioinformatics. 2013;14(15):1.

61. Wissink EM, Fogarty EA, Grimson A. High-throughput discovery of post-transcriptional cis-regulatory elements. BMC Genomics. BMC Genomics; 2016;17(1):177.

62. Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell. Cell Press; 1986;44(2):283–92.

63. Schwartz DC, Parker R. Mutations in translation initiation factors lead to increased rates of deadenylation and decapping of mRNAs in Saccharomyces cerevisiae. Mol Cell Biol. 1999;19(8):5247–56.

64. Pelechano V, Wei W, Steinmetz LM. Extensive transcriptional heterogeneity revealed by isoform profiling. Nature. Nature Publishing Group; 2013;497(7447):127–31.

65. Zhou X, Carbonetto P, Stephens M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genet. 2013;9(2):e1003264.

66. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. BioMed Central; 2011;6(1):26.

67. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034–50.

68. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1–22.

Figure captions

Fig 1. Study overview. The goal of this study is to discover and integrate cis-regulatory mRNA elements a_ecting mRNA stability and assess their dependence on mRNA degradation pathways. **Data)** We obtained *S. cerevisiae* genome-wide half-life data from wild-type (WT) as well as from 34 knockout strains from Sun et al. 2013. Each of the knockout strains has one gene closely related to mRNA degradation pathways knocked out. **Analysis)** We systematically searched for novel sequence features associating with half-life from 5'UTR, start codon context, CDS, stop codon context, and 3'UTR. Effects of previously reported cis-regulatory elements were also assessed. Moreover, we assessed the dependencies of different sequence features on degradation pathways by analyzing their effects in the knockout strains. **Integrative model)** We build a statistical model to predict genome-wide half-life solely from mRNA sequence. This allowed the quantification of the relative contributions of the sequence features to the overall variation across genes and assessing the sensitivity of mRNA stability with respect to single-nucleotide variants.

Fig 2. Overview of identified or collected sequence features. (A) Sequence features that were identified or collected from different sequence regions in this study. When applicable, stabilizing elements are shown in blue, destabilizing in red. **(B)** Co-occurrence significance (FDR, Fisher test p-value corrected with Benjamini-Hochberg) between different motifs (left). Number of occurrences among the 4,388 mRNAs (right).

Fig 3. Upstream AUG codon (uAUG) destabilize mRNA. (A) Distribution of mRNA half-life for mRNAs without uAUG (left) and with at least one uAUG (right) in, from left to right: wild type, *XRN1*, *UPF2* and *UPF3* knockout *S. cerevisiae* strains. Median fold-change (Median FC) calculated by dividing the median of the group without uAUG with the group with uAUG. **(B)** Distribution of mRNA half-lives for mRNAs with zero (left), one (middle), or more (right) uAUGs in *S. cerevisiae*. **(C)** Distribution of mRNA half-lives for *S. cerevisiae* mRNAs with, from left to right: no uAUG, with one in-frame uAUG but no cognate premature termination codon, with one out-of-frame uAUG and one cognate premature termination codon in the CDS, and with one uAUG and one cognate stop codon in the 5'UTR (uORF). **(D)** Same as in (C) for *S. pombe* mRNAs. All p-values were calculated with Wilcoxon rank-sum test. Numbers in the boxes indicate number of members in the corresponding group. Boxes represent quartiles, whiskers extend to the highest or lowest value within 1.5 times the interquartile range and horizontal bars in the boxes represent medians. Data points falling further than 1.5-fold the interquartile distance are considered outliers and are shown as dots.

Fig 4. Translation initiation, elongation and termination features associate with mRNA half-life. (A) Distribution of half-life for mRNAs grouped by the third nucleotide before the start codon. Group sizes (numbers in boxes) show that nucleotide frequency at this position positively associates with half-life. **(B)** mRNA half-life (y-axis) versus species-specific tRNA adaptation index (sTAI) (x-axis) for *S. cerevisiae*. **(C)** mRNA half-life explained variance (y-axis, Materials and Methods) in wild-type (WT) and across all 34 knockout strains (grouped according to their functions). Each blue dot represents one replicate, bar heights indicate means across replicates. Bars with a red star are significantly different from the wild type level (FDR < 0.1, Wilcoxon rank-sum test, followed by Benjamini-Hochberg correction). **(D)** Distribution of half-life for mRNAs grouped by the stop codon and the following nucleotide. Colors represent three different stop codons (TAA, TAG and TGA), within each stop codon group, boxes are shown in G, A, T, C order of their following base. Only the P-values for the most drastic pairwise comparisons (A versus C within each stop codon group) are shown. All p-values in boxplots were calculated with Wilcoxon rank-sum test. Boxplots computed as in Fig 3.

Fig 5. 3'UTR half-life determinant motifs in *S. cerevisiae*. (A) Distribution of half-lives for mRNAs grouped by the number of occurrence(s) of the motif ATATTC, TGCAT (Whi3), TGTAAATA (Puf3) and TTTTTTA respectively in their 3'UTR sequence. Numbers in the boxes represent the number of members in each box. FDR were reported from the linear mixed effect model (Materials and Methods). **(B)** Fraction of transcripts containing the motif (y-axis) within a 20-bp window centered at a position (x-axis) with respect to poly(A) site for different motifs (facet titles). Positional bias was not observed when aligning 3'UTR motifs with respect to the stop codon. **(C)** Prediction of the relative effect on half-life (y-axis) for single-nucleotide substitution in the motif with respect to the consensus motif (y=1, horizontal line). The motifs were extended 2 bases at each flanking site (positions +1, +2, -1, -2). **(D)** Nucleotide frequency within motif instances, when allowing for one mismatch compared to the consensus motif. **(E)** Mean conservation score (phastCons, Materials and Methods) of each base in the consensus motif with 2 flanking nucleotides (y-axis).

Fig 6. Genome-wide prediction of mRNA half-lives from sequence features and analysis of the contributions. (A-B) mRNA half-lives predicted (x-axis) versus measured (y-axis) for *S. cerevisiae* (A) and *S. pombe* (B) respectively. **(C)** Contribution of each sequence feature individually (Individual), cumulatively when sequentially added into a combined model (Cumulative) and explained variance drop when each single feature is removed from the full model separately (Drop). Values reported are the mean of 100 times of cross-validated evaluation (Materials and Methods). **(D)** Expected half-life fold-change of single-nucleotide variations on sequence features. For length and GC, dot represent median half-life fold change of one nucleotide shorter or one G/C to A/T transition respectively. For codon usage, each dot represents median half-life fold-change of one type of synonymous mutation, all kinds of synonymous mutations are considered. For uAUG, each dot represents median half-life fold-change of mutating out one uAUG. For motifs, each dot represents median half-life fold-change of one type of nucleotide transition at one position on the motif (Materials and Methods). Medians are calculated across all mRNAs. **(E)** Overview of conclusions.

Supporting Information

S1 Fig. Length of 5'UTR, CDS and 3'UTR correlate with mRNA half-life. (A-B) 5'UTR length (x-axis) versus half-life (y-axis) for *S. cerevisiae* (A) and *S. pombe* (B). **(C-D)** CDS length (x-axis) versus half-life (y-axis) for *S. cerevisiae* (C) and *S. pombe* (D). **(E-F)** 3'UTR length (x-axis) versus half-life (y-axis) for *S. cerevisiae* (E) and *S. pombe* (F).

S2 Fig. GC content of 5'UTR, CDS and 3'UTR correlate with mRNA half-life. (A-B) 5'UTR GC content (x-axis) versus half-life (y-axis) for *S. cerevisiae* (A) and *S. pombe*. **(C-D)** CDS GC content (x-axis) versus half-life (y-axis) for *S. cerevisiae* (C) and *S. pombe* (D). **(E-F)** 3'UTR GC content (x-axis) versus half-life (y-axis) for *S. cerevisiae* (E) and *S. pombe* (F).

S3 Fig. Folding energy of 5'UTR, CDS and 3'UTR correlate with mRNA half-life. (A-B) 5' free energy (x-axis) versus half-life (y-axis) for *S. cerevisiae* (A) and *S. pombe* (B). **(C-D)** CDS free energy (x-axis) versus half-life (y-axis) for *S. cerevisiae* (C) and *S. pombe* (D). **(E-F)** 3' free energy (x-axis) versus half-life (y-axis) for *S. cerevisiae* (E) and *S. pombe* (F).

S4 Fig. Translation initiation, elongation and termination features associate with mRNA half-life. (A) Start codon context (Kozak sequence) generated from 4388 *S. cerevisiae* genes and 3713 *S. pombe* genes. **(B)** Distribution of half-life for mRNAs grouped by the third nucleotide before the start codon for *S. pombe*. Group sizes (numbers in boxes) show that nucleotide frequency at this position positively associates with half-life. **(C)** mRNA half-life (y-axis) versus species-specific tRNA adaptation index (sTAI) (x-axis) for *S. pombe*. **(D)** Distribution of half-life for mRNAs grouped by the stop codon and the following nucleotide for *S. pombe*. Colors represent three different stop codons (TAA, TAG and TGA), within each stop codon group, boxes are shown in G, A, T, C order of their following base. Only the P-values for the most drastic pairwise comparisons (A versus C within each stop codon group) are shown. **(E)** Distribution of half-life for mRNAs grouped by with or without additional 3'UTR in-frame stop codon for *S. cerevisiae*. 30 bases window after the main stop codon was considered. **(F)** Same as (E) for *S. pombe*. All p-values in boxplot were calculated with Wilcoxon rank-sum test. Boxplots computed as in Fig 3.

S5 Fig. *S. cerevisiae* 5'UTR mRNA half-life associated motif. (A) Distribution of half-lives for mRNAs grouped by the number of occurrence(s) of the motif AAACAAA in their 5'UTR sequence. Numbers in the boxes represent the number of members in each box. FDR were reported from the linear mixed effect model (Materials and Methods). **(B)** Prediction of the relative effect on half-life (y-axis) for single-nucleotide substitution in the motif with respect to the consensus motif (y=1, horizontal line). The motifs were extended 2 bases at each flanking site (positions +1, +2, -1, -2). **(C)** Nucleotide frequency within motif instances, when allowing for one mismatch compared to the consensus motif. **(D)** Mean conservation score (phastCons, Materials and Methods) of each base in the consensus motif with 2 flanking nucleotides (y-axis).

S6 Fig. Summary of CREs effect changes across all 34 knockouts comparing with WT. Colour represent the relative effect size (motifs, St-3 C-A, TGAG-TGAC, uAUG), correlation (5' folding energy) or explained variance (codon usage) upon knockout of different genes (y-axis) (Materials and Methods for detailed description). Wild-type label is shown in the bottom (WT) P-values calculated with Wilcoxon rank-sum test by comparing each mutant to wild-type level, multiple testing p-values corrected with Bonferroni & Hochberg (FDR). Stars indicating significance of statistical testing (FDR < 0.1). 5' energy: correlation of 5'end (5'UTR plus first 10 codons) folding energy with mRNA half-lives; St-3 C-A: relative median half-life difference between genes with cytosine and adenine at start codon -3 position; TGAC-TGAG: relative median half-life difference between genes with stop codon +1 TGAC and TGAG. Codon usage: codon usage explained mRNA half-life variance. uAUG: relative median half-life difference between genes without and with upstream AUG in the 5'UTR (Materials and Methods)

S7 Fig. Genome-wide prediction of mRNA half-lives from sequence features with RATE-seq data. mRNA half-lives predicted (x-axis) versus measured (y-axis) with RATE-seq data for 3,539 genes that have complete profiles of all features.

S8 Fig. Predicted effects of synonymous codon transitions on half-life. Expected half-life fold-change (x-axis) at each synonymous codon transitions. Each row represent transition from one codon (y-axis) to its synonymous partners. Only synonymous codons that differ by one base were considered.

S1 Table. List of 34 knockout strains analyzed in this study.

S2 Table. List of correlation and p-value between sequence length, GC content and folding energy with mRNA half-life for *S. cerevisiae* and *S. pombe*.

S3 Table. GO enrichment results for 3'UTR motifs.

S4 Table. Regression coefficients in the joint model for *S. cerevisiae* (Sun and Neymotin data) and *S. pombe*.

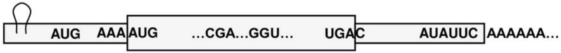
S5 Table. Out-of-fold mRNA half-life prediction results for *S. cerevisiae* (Sun and Neymotin data) and *S. pombe*.

Data

Genome-wide mRNA half-lives in *S. cerevisiae*: wild-type + 34 strains knocked out for RNA degradation pathway genes (Sun et al. Mol. Cell. 2013)

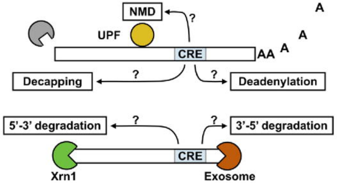
Analysis

Cis-regulatory elements



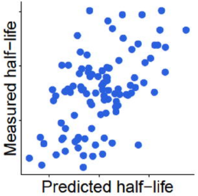
- Novel elements discovery
- Integrate novel and known elements
- Single-nucleotide effect estimation

Dependence on degradation pathways

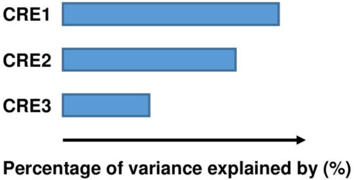


Integrative model

Out-of-sample predictions

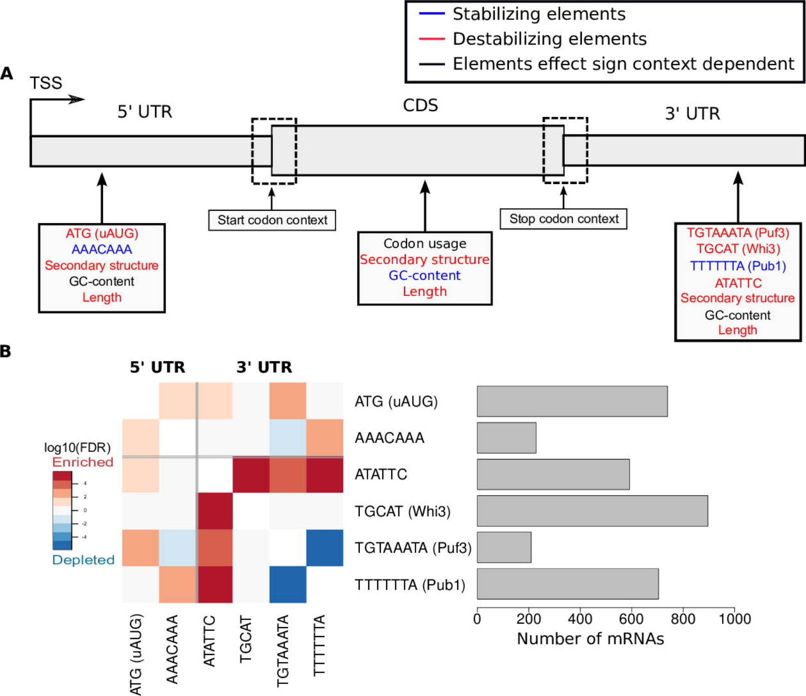


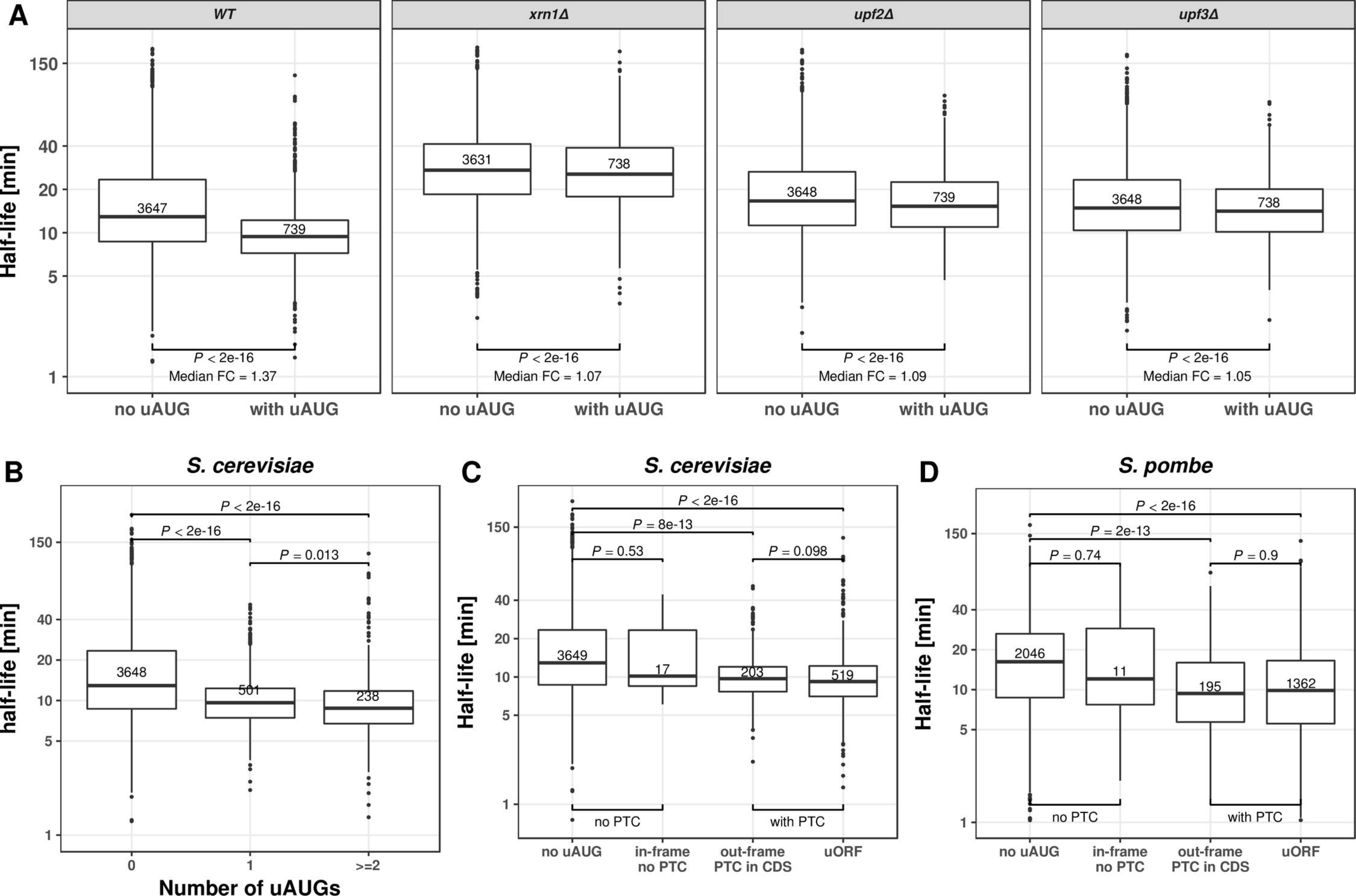
Relative contributions

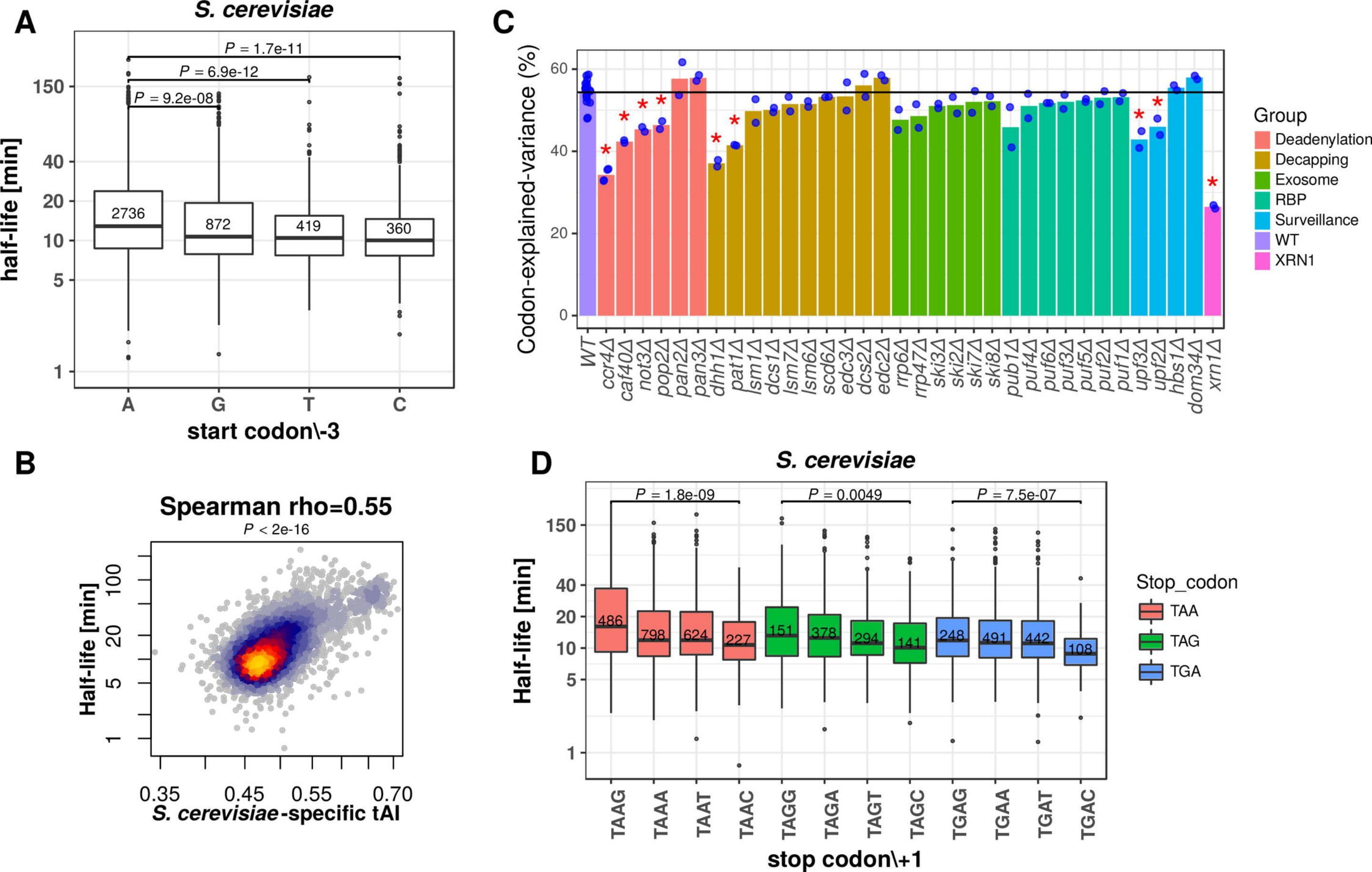


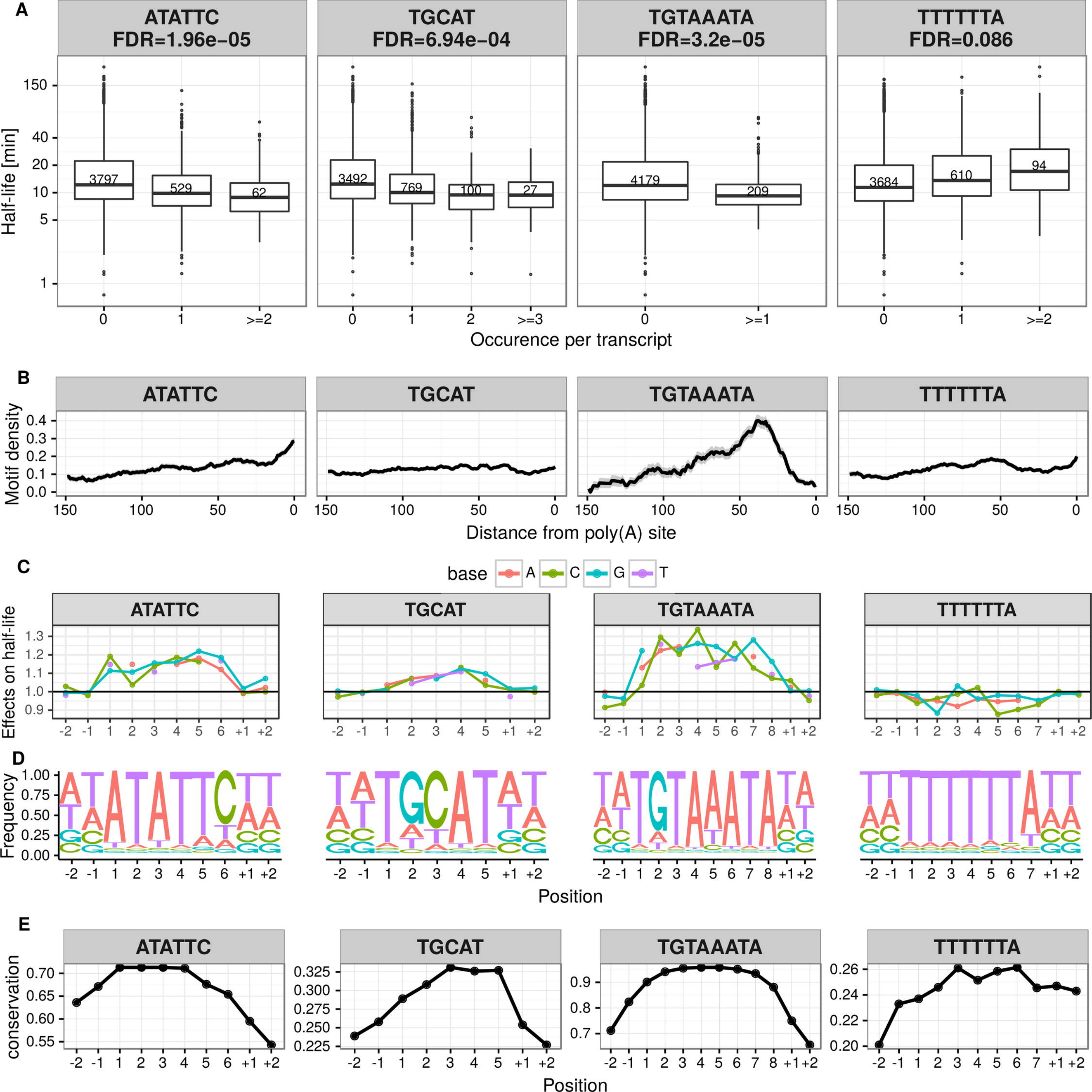
Single-nucleotide effects

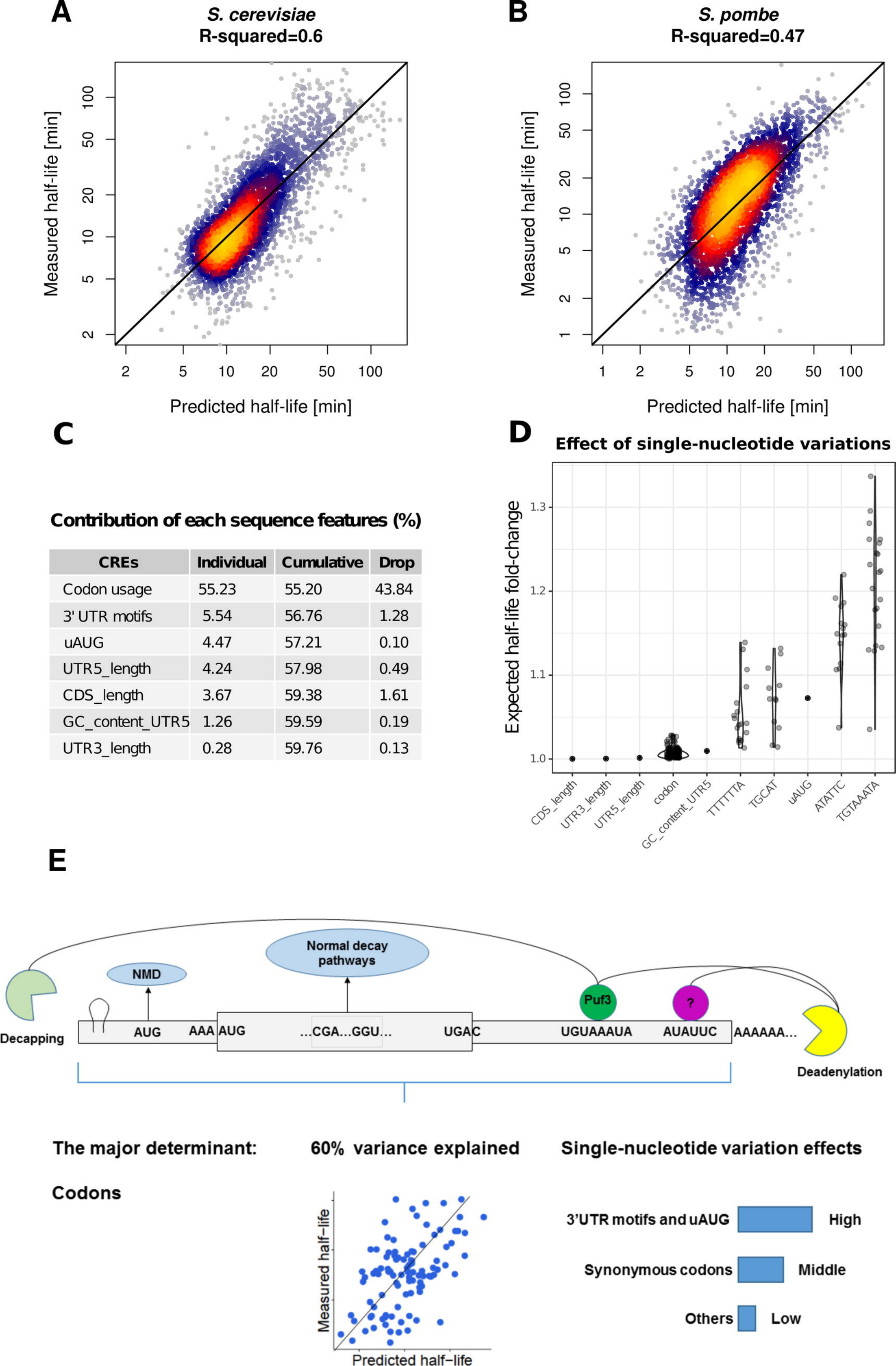
...CATACGATA...
...CATAGGATA...


Stabilizing elements
Destabilizing elements
Elements effect sign context dependent
A
TSS
5' UTR
CDS
3' UTR
Start codon context
Stop codon context
ATG (uAUG)
AAACAAA
Secondary structure
GC-content
Length
Codon usage
Secondary structure
GC-content
Length
TGTAAATA (Puf3)
TGCAT (Whi3)
TTTTTTA (Pub1)
ATATTC
Secondary structure
GC-content
Length
B
5' UTR
3' UTR
log10(FDR)
Enriched
4
2
0
-2
-4
Depleted
ATG (uAUG)
AAACAAA
ATATTC
TGCAT (Whi3)
TGTAAATA (Puf3)
TTTTTTA (Pub1)
ATG (uAUG)
AAACAAA
ATATTC
TGCAT
TGTAAATA
TTTTTTA
0
200
400
600
800
1000
Number of mRNAs


A
WT
xrn1Δ
upf2Δ
upf3Δ
Half-life [min]
150
40
20
10
5
1
3647
739
P < 2e-16
Median FC = 1.37
3631
738
P < 2e-16
Median FC = 1.07
3648
739
P < 2e-16
Median FC = 1.09
3648
738
P < 2e-16
Median FC = 1.05
no uAUG
with uAUG
B
S. cerevisiae
half-life [min]
P < 2e-16
P < 2e-16
P = 0.013
3648
501
238
0
1
>=2
Number of uAUGs
C
S. cerevisiae
Half-life [min]
P < 2e-16
P = 8e-13
P = 0.53
P = 0.098
3649
17
203
519
no PTC
with PTC
no uAUG
in-frame no PTC
out-frame PTC in CDS
uORF
D
S. pombe
Half-life [min]
P < 2e-16
P = 2e-13
P = 0.74
P = 0.9
2046
11
195
1362
no PTC
with PTC
no uAUG
in-frame no PTC
out-frame PTC in CDS
uORF


A
S. cerevisiae
P = 1.7e-11
P = 6.9e-12
P = 9.2e-08
half-life [min]
150
40
20
10
5
1
2736
872
419
360
A
G
T
C
start codon\-3
B
Spearman rho=0.55
P < 2e-16
Half-life [min]
100
20
5
1
0.35
0.45
0.55
0.70
S. cerevisiae-specific tAI
C
Codon-explained-variance (%)
60
40
20
0
WT
ccr4Δ
caf40Δ
not3Δ
pop2Δ
pan2Δ
pan3Δ
dhh1Δ
pat1Δ
lsm1Δ
dcs1Δ
lsm7Δ
lsm6Δ
scd6Δ
edc3Δ
dcs2Δ
edc2Δ
rrp6Δ
rrp47Δ
ski3Δ
ski2Δ
ski7Δ
ski8Δ
pub1Δ
puf4Δ
puf6Δ
puf3Δ
puf5Δ
puf2Δ
puf1Δ
upf3Δ
upf2Δ
hbs1Δ
dom34Δ
xrn1Δ
Group
Deadenylation
Decapping
Exosome
RBP
Surveillance
WT
XRN1
D
S. cerevisiae
P = 1.8e-09
P = 0.0049
P = 7.5e-07
Half-life [min]
150
40
20
10
5
1
486
798
624
227
151
378
294
141
248
491
442
108
TAAG
TAAA
TAAT
TAAC
TAGG
TAGA
TAGT
TAGC
TGAG
TGAA
TGAT
TGAC
stop codon\+1
Stop_codon
TAA
TAG
TGA


A
ATATTC
FDR=1.96e−05
TGCAT
FDR=6.94e−04
TGTAAATA
FDR=3.2e−05
TTTTTTA
FDR=0.086
Half-life [min]
150
40
20
10
5
1
3797
529
62
3492
769
100
27
4179
209
3684
610
94
0
1
>=2
0
1
2
>=3
0
>=1
0
1
>=2
Occurence per transcript
B
ATATTC
TGCAT
TGTAAATA
TTTTTTA
Motif density
0.4
0.3
0.2
0.1
0.0
150
100
50
0
Distance from poly(A) site
C
base
A
C
G
T
ATATTC
TGCAT
TGTAAATA
TTTTTTA
Effects on half-life
1.3
1.2
1.1
1.0
0.9
-2
-1
1
2
3
4
5
6
7
8
+1
+2
D
Frequency
1.00
0.75
0.50
0.25
0.00
-2
-1
1
2
3
4
5
6
7
8
+1
+2
Position
E
ATATTC
TGCAT
TGTAAATA
TTTTTTA
conservation
0.70
0.65
0.60
0.55
0.325
0.300
0.275
0.250
0.225
0.9
0.8
0.7
0.26
0.24
0.22
0.20
-2
-1
1
2
3
4
5
6
7
8
+1
+2
Position



Contribution of each sequence features (%)

CREs	Individual	Cumulative	Drop
Codon usage	55.23	55.20	43.84
3' UTR motifs	5.54	56.76	1.28
uAUG	4.47	57.21	0.10
UTR5_length	4.24	57.98	0.49
CDS_length	3.67	59.38	1.61
GC_content_UTR5	1.26	59.59	0.19
UTR3_length	0.28	59.76	0.13