bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sequence features explain most of the mRNA stability variation across genes

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 and mechanisms regulating mRNA stability has been described, their quantitative contributions in determining mRNA half-life is unknown. Here, we built quantitative models for two eukaryotic genomes *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* that, for the first time, explain most of the half-life variation between genes based on mRNA sequence alone. The models integrate known functional cis-regulatory elements, identify novel ones, and quantify their contributions at single-nucleotide resolution. We show quantitatively that codon usage is the major determinant of mRNA stability, and that this effect depends on canonical mRNA degradation pathways. Altogether, our results integrate and quantitatively delineate mRNA stability cis-regulatory elements and provide a methodology that can serve as a scaffold to study the function of cis-regulatory elements and to discover novel ones.

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. Hence, stress genes, which must rapidly

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

respond to environmental signals, show low stability (Zeisel *et al*, 2011; Rabani *et al*, 2014). 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 (Schwanhäusser *et al*, 2011; Schwalb *et al*, 2016; Eser *et al*, 2016).

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, sequence motifs present in the 3'UTR including binding sites of RNA-binding proteins (Olivas & Parker, 2000; Vasudevan & Peltz, 2001), and, in higher eukaryotes, microRNAs (Lee, 1993). Moreover, codon usage, which affects translation elongation rate, also regulates mRNA stability (Hoekema *et al*, 1987; Presnyak *et al*, 2015; Boël *et al*, 2016; Mishima & Tomari, 2016). 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* (Muhlrad *et al*, 1995; LaGrandeur & Parker, 1999).

Since the RNA degradation machineries are well conserved among eukaryotes, the pathways have been extensively studied using *Saccharomyces cerevisiae* as a model organism (Caponigro & Parker, 1996; Parker, 2012). The general mRNA degradation pathway starts with the removal of the poly(A) tail by the Pan2/Pan3 (Boeck *et al*, 1996; Brown *et al*, 1996) and Ccr4/Not complexes (Tucker *et al*, 2001; Collart, 2003; Yamashita *et al*, 2005). Subsequently, mRNA is subjected to decapping carried out by Dcp2 and promoted by several factors including Dhh1 and Pat1 (Pilkington & Parker, 2008; She *et al*, 2008). The decapped and deadenylated mRNA can be degraded rapidly in the 3' to 5' direction by the exosome (Anderson & Parker, 1998) or in the 5' to 3' direction by Xrn1 (Muhlrad *et al*, 1994; Hsu & Stevens, 1993). 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) (Garneau *et al*, 2007; Parker, 2012).

Despite all this knowledge, prediction of mRNA half-life from a gene sequence is still not established. Most of the mechanistic studies could only be performed on

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

individual genes and it is unclear how the effects generalize genome-wide. In addition, neither the contribution of different CREs to the overall stability and to the variation between genes has been systematically studied, nor have the dependences of CREs to distinct mRNA degradation pathways.

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. We used a semi-mechanistic model which allowed us to interpret it in terms of 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. Interpretability of the model also allowed studying how the function of CREs depends on different mRNA degradation pathways. The application of the modeling approach to *S. pombe* supports the generality of these findings.

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 (Sun *et al*, 2013), which provides genome-wide half-life measurements for 4,388 expressed genes of a wild type lab strain (Fig 1). The same study also measured genome-wide mRNA half-lives for 34 strains knocked out for RNA degradation pathway genes (Table EV1), allowing for investigating how the association of candidate CREs with half-life depends on RNA degradation pathways. When applicable, associations were also investigated in the *S. pombe* genome, using half-life values for 3,614 expressed mRNAs in a wild-type lab strain from Eser and colleagues (Eser *et al*, 2016). 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. Half-life associated negatively with 5'UTR length (Spearman ρ = -0.17, $P = 3 \times 10^{-12}$ for *S. cerevisiae* and Spearman ρ = -0.27, $P = 2 \times 10^{-11}$ for *S.*

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pombe), with the coding sequence (Spearman ρ = -0.23, $P < 2 \times 10^{-16}$ for *S. cerevisiae* and Spearman ρ = -0.32, $P = 4 \times 10^{-15}$ for *S. pombe*), and with 3'UTR length (Spearman ρ = -0.06, $P = 1 \times 10^{-8}$ for *S. cerevisiae* and Spearman ρ = -0.23, $P = 6 \times 10^{-12}$ for *S. pombe*) (Fig EV1A, B, C, D, E, F). Moreover, *S .cerevisiae* half-life was found to be negatively associated with the GC-content of the 5'UTR (Spearman ρ = -0.11, $P = 3 \times 10^{-14}$) and of the 3'UTR (Spearman ρ = -0.17, $P < 2 \times 10^{-16}$), whereas the GC-content of the CDS region correlated positively with half-life (Spearman ρ = 0.27, $P < 2 \times 10^{-16}$) (Fig EV2A, C, E). In contrast, *S. pombe* showed consistent positive association between GC-content and half-life (Spearman ρ = 0.16, 0.24, 0.05 and $P < 2 \times 10^{-16}$, $P < 2 \times 10^{-16}$, P = 0.01 respectively for 5'UTR, CDS and 3'UTR, Fig EV2B, D, F). Also, secondary structure in 5'UTR (Materials and Methods) associated with RNA instability (Spearman ρ = -0.22, $P = 8 \times 10^{-12}$ for *S. cerevisiae* and Spearman ρ = -0.24, $P < 2 \times 10^{-16}$ for *S. pombe*, Fig EV1G, H).

Novel candidate cis-regulatory elements were searched for using a robust k-mer based regression, investigating all 3- to 8-mers, and followed by k-mer assembly (Materials and Methods). This recovered *de novo* the Puf3 binding motif TGTAAATA in 3'UTR (Gerber *et al*, 2004; Gupta *et al*, 2014), a well-studied CRE that confers RNA instability, as well as the Whi3 binding motif TGCAT (Colomina *et al*, 2008; Cai & Futcher, 2013). Two new motifs were found: AAACAAA in 5'UTR, and ATATTC in 3'UTR (Fig 1A). Notably, the motif TGCAT, was found in 21% of the genes (938 out of 4,388) and significantly co-occurred with the other two destabilizing motifs found in 3'UTR: Puf3 and the Whi3 binding site motifs (Fig 1B). Except for AAACAAA and TTTTTTA, all motifs associated with shorter half-lives (Fig 1A).

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

Genome-wide, occurrence of an upstream AUG (uAUG) associated with 1.37 fold shorter half-life (fold change between medians, $P < 2.2 \times 10^{-16}$), and the effect strengthened for genes with two or more AUGs (Fig 2A, B). Among the 34 knock-out

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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* (Leeds *et al*, 1992; Cui *et al*, 1995), and for the general 5'-3' exonuclease *Xrn1* (Fig 2A). The dependence on NMD suggested that the association between uAUG and shorter half-life might be due to the occurrence of a premature stop codon. Indeed, the association of uAUG with decreased half-lives was only found for genes with a premature stop codon cognate with the uAUG (Fig 2C). 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 2C). This finding likely holds for many other eukaryotes as we found the same trends in *S. pombe* (Fig 2D). This observations are consistent with single-gene studies demonstrating that translation of upstream ORFs can lead to RNA degradation by nonsense-mediated decay (Gaba *et al*, 2005; Barbosa *et al*, 2013). 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, the folding energy of the 5' initiation sequence, defined as the 5'UTR and the first 10 codons (Zur & Tuller, 2013), positively correlated with mRNA stability in *S. cerevisiae* (Spearman ρ = -0.22, $P = 8 \times 10^{-12}$ for *S. cerevisiae* and Spearman ρ = -0.24, $P < 2 \times 10^{-16}$ for *S. pombe*, Fig EV1G, H). The mRNA destabilizing function of strong secondary structure elements had been demonstrated for the gene *PGK1* (LaGrandeur & Parker, 1999; Muhlrad *et al*, 1995). Our genome-wide analysis indicates that this is a general phenomenon. Possibly, strong secondary structure in the 5' UTR interferes with ribosome scanning, thus slows down translation initiation, and therefore decreases the level of mRNA protection by the translation machinery (Roy & Jacobson, 2013; Huch & Nissan, 2014).

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Second, longer 5'UTRs associated with less stable mRNAs (Spearman ρ = -0.17, P = 3×10^{-12} for *S. cerevisiae* and Spearman ρ = -0.27, $P = 2 \times 10^{-11}$ for *S. pombe*, Fig EV1A, 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 repression elements. Hence, longer 5'UTR may confer mRNA instability by decreasing translation initiation and therefore decreasing the protection by the translation machinery.

Third, a significant association between the third nucleotide before the start codon and mRNA half-life was observed (Fig 3A). The median half-life correlated with the nucleotide frequency (Fig EV3A), leading to 1.28 fold difference between the 2,736 genes with an adenosine, the most frequent nucleotide at this position, and the 360 genes with a cytosine, the least frequent nucleotide at this position ($P = 1.7\text{x}10^{-11}$). The same correlation was also found to be significant for *S. pombe* (Fig EV3B, C). Functional effect of the start codon context on mRNA stability had been established as the long-lived *PGK1* mRNA was significantly destabilized when substituting the sequence context around its start codon with the one from the short-lived *MFA2* mRNA (LaGrandeur & Parker, 1999). Our genome-wide analysis indicates that this effect generalizes to other genes. Possibly the start codon context, which controls translation initiation efficiency (Kozak, 1987; Dvir *et al*, 2013), affects thereby mRNA stability.

Altogether, these findings indicate that 5'UTR elements, including the start codon context, may affect mRNA stability by altering translation initiation. *De novo* search for regulatory motifs identified AAACAAA to be significantly associated with longer mRNA half-lives. However, further analysis did not provide strong support for this motif, indicating it might be merely correlative (Fig EV5).

Codon usage regulates mRNA stability through common mRNA decay pathways

Codon usage has been reported to affect mRNA stability through a translation-dependent manner (Presnyak *et al*, 2015; Mishima & Tomari, 2016; Bazzini *et al*, 2016). However, it is unclear on which mRNA degradation pathway it depends.

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The species-specific tRNA adaptation index (sTAI) was used to quantify the level of codon usage bias (Sabi & Tuller, 2014). First, we confirmed on this independent half-life dataset the strong correlation between codon optimality and mRNA stability in *S. cerevisiae* (Fig 3C, Spearman $\rho = 0.55$, $P < 2.2\text{x}10^{-16}$). 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 (Fig 3D). In particular, all assayed genes of the Ccr4-Not complex, including *CCR4*, *NOT3*, *CAF40* and *POP2*, are required for wild-type level effects of codon usage on mRNA decay. Among them, *CCR4* has the largest effect. This confirmed recent studies showing that accelerated decay of non-optimal codon genes requires deadenylation activities of Ccr4-Not (Presnyak *et al*, 2015; Mishima & Tomari, 2016). Moreover, our results confirmed the dependence on Dhh1 (Radhakrishnan *et al*, 2016), but also on its interacting partner Pat1. 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 3D). Notably, we did not observe any change of effect upon knockout of *DOM34* and *HBS1* (Figure EV5), 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.

Our systematic analysis revealed two additional novel dependencies: First, on the common 5' to 3' exonuclease Xrn1, and second, on *UPF* genes, which are essential players of NMD (Fig 3D). 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 (He *et al*, 2003; Hug *et al*, 2015). 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

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Fig. 3B). This also held for *S. pombe* (Fig EV3D). 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; McCaughan *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 (Fig. 3B). 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 (Williams *et al*, 2004), exhibited significantly higher half-life (Fig EV3E, F). 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 (Fig EV5). These results strongly 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, FDR < 0.1, Materials and Methods). This analysis recovered three described motifs: the Puf3 binding motif TGTAAATA (Gerber *et al*, 2004), the Whi3 binding motif TGCAT (Colomina *et al*, 2008; Cai & Futcher, 2013), and a poly(U) motif TTTTTTA, which is likely bound by Pub1 (Duttagupta *et al*, 2005), or is part of the long poly(U) stretch that form a looping structure with poly(A) tail (Geisberg *et al*, 2014). We also identified a novel motif, ATATTC, that associated with lower mRNA half-life.

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 (Fig EV5), suggesting a potential interaction with deadenylation factors. Second, single nucleotide deviations from the consensus sequence of the

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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* (Fig EV5). We also found a significant dependence on the deadenylation (*CCR4*, *POP2*) and decapping (*DHH1*, *PAT1*) pathways (Fig EV5), consistent with previous single gene experiment showing that Puf3 binding promotes both deadenylation and decapping (Olivas & Parker, 2000; Goldstrohm *et al*, 2007). 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 (Garí *et al*, 2001). Binding of Whi3 leads to destabilization of the *CLN3* mRNA (Cai & Futcher, 2013). A subset of yeast genes are up-regulated in the Whi3 knockout strain (Cai & Futcher, 2013; Holmes *et al*, 2013). 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 = 6.9×10^{-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, suggesting possible broader and genome-wide role of Whi3 and its binding motif. However, no significant genetic dependence of the effect of the Whi3 binding motif was found (Fig EV5).

The mRNAs harboring the TTTTTTA motif tended to be more stable (Fig 5A). No positional preferences was observed for this motif (Fig 5B), so as the poly(U) stretch reported before (Geisberg *et al*, 2014). Effects of this motif depends on genes from Ccr4-Not complex and Xrn1 (Fig EV5).

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

60% between-gene half-life variance 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, Table EV2). 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 error across genes is 30%. Prediction accuracy of half-life of 30% is remarkably high 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 (Eser *et al*, 2016). The same model applied to *S. pombe* explained 42% of the total half-life variance. 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 significantly associated with half-life in the joint model indicating that they contributed independently to half-life (the complete list of features and their p-values are given in Table EV3). 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 or because 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 (55.90%) and in the joint model (44.23%) (Figure 6B). 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

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and codon usage, favorably contribute to the variation. This does not reflect, however, 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 as median fold-change across genes upon single-nucleotide variants, 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; Fig EV6). 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 (Fig EV6, 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 motif 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 (Roy & Jacobson, 2013; Huch & Nissan, 2014). 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 (Figure 6E).

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Integrative analyses of cis-regulatory elements as we used here complement mechanistic single-gene studies for important aspects. It allows assessing genome-wide the importance of CREs that have been reported previously with single-gene experiments, as Vogel and colleagues (Vogel *et al*, 2010) had shown when modeling protein abundance from mRNA levels and sequence. 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 as developed by the group of Segal (Dvir *et al*, 2013; Shalem *et al*, 2015). 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.

Causality cannot be proven through a regression analysis approach. However, we provided several complementary analyses to further assess the potential functionality of candidate CREs. These include conservation, positional preferences, and epistasis analyses to assess the dependencies on RNA degradation pathways. One of the motif we found by regression, AAACAAA in 5'UTR, was neither supported by these complementary analyses, nor it remained significant in the joint model. We think this motifs is most likely correlative. In contrast, the ATATTC motif in 3'UTR is strongly supported by these complementary analyses and is also significant in the joint model. Two of the most interesting sequence features that we identified but still need to be functionally assayed are the start codon context and 5' sequence free energy. Given their established effect on translation initiation (Kozak, 1986; Dvir *et al*, 2013; Tuller & Zur, 2014), the general coupling between translation and mRNA degradation (Huch & Nissan, 2014; Roy & Jacobson, 2013), as well as several observations directly on mRNA stability for single genes (Muhlrad *et al*, 1995; Barnes, 1998; LaGrandeur & Parker, 1999; Schwartz & Parker, 1999), they are very likely to be functional on most

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (Dvir *et al*, 2013). 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. We recovered results from recent studies about the dependencies of the effect of codon usage on RNA stability 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 et al. 2013. *S. cerevisiae* gene boundaries were taken from the boundaries of the most abundant isoform quantified by Pelechano et al. 2013. 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 et al. 2016. Reference genome version ASM294v2.26 was used to obtain sequence information.

For both half-life datasets, only mRNAs with mapped 5'UTR and 3'UTR were considered. mRNAs that have 5'UTR length shorter than 6nt were further filtered out, so that start codon context could always be defined. Half-life outliers of *S. pombe* (half-life less than 1 or larger than 250 mins) were removed.

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Codon-wise species-specific tRNA adaptation index (sTAI) of yeasts were obtained from Sabi & Tuller 2014. 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). 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: $\frac{\left(median(H_{count=1}) - median(H_{count=0})\right)}{median(H_{count=0})}$ 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: $\frac{\left(medain(H_A) - median(H_C)\right)}{median(H_A)}$ 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: $\frac{\left(median(H_{TGAG}) - median(H_{TGAC})\right)}{median(H_{TGAG})}$ 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.

Motif discovery

Motif discovery was conducted similarly to Eser et al. 2016 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 (fixed effect) while controlling the effects of the others (random effect) and for the region length as a covariate. For CDS we also used codons as further covariates. We systematically 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

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

software (Zhou *et al*, 2013). The p-values were corrected for multiple testing using Benjamini-Hochberg's FDR. Motifs were subsequently manually assembled based on overlapping significant k-mers.

Folding energy calculation

5' ensemble folding energy was calculated with RNAfold from ViennaRNA version 2.1.9 (Lorenz *et al*, 2011), with default parameters. 5' sequence was defined as 5'UTR sequence plus the first 10 codons within the coding region. Genes with 5'UTRs that are shorter than 30 bps were excluded from the analysis.

***S. cerevisiae* conservation analysis**

The phastCons (Siepel *et al*, 2005) 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 (joint model)

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*. Linear models were first fitted with all sequence features as covariates, non-significant sequence features were then removed from the final prediction models, ending up with 70 features for *S. cerevisiae* model and 73 features for *S. pombe* (each single coding codon was fitted as a single covariate). L1 or L2 regularization were not necessary, as they did not improve the out-of-fold prediction accuracy (tested with glmnet R package (Friedman *et al*, 2010)). Out-of-fold prediction was applied with 10-fold cross validation for any prediction task in this study. A complete list of model features and their p-values for both yeast species are provided in Table EV2.

Variance explained for linear model

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

Analysis of sequence feature contribution

The contribution of each sequence feature was analyzed individually as a univariate regression and also jointly in a multivariate regression model. We quantified the feature contributions in 10-fold cross-validation. The contribution of each feature *individually* was calculated as the variance explained by a univariate model. We then added the features in a descending order of their individual explained variance to a joint model, and recorded the *cumulative* variance explained as adding more features. As compensation to the issue that the additional variance explained by a single feature to a joint model depends on the order of adding it, the *drop* of variance explained upon leaving out one feature separately from the full model was also assessed.

Single-nucleotide variant effect prediction

The effects of single-nucleotide variation (except for motifs) 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 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.

Motif single-nucleotide variant analysis

Effects of motif single-nucleotide variants were predicted with linear model modified from (Eser *et al*, 2016). When assessing the effect of a given motif variation, we controlled for the effect of all other sequence features using a linear model with the other features as covariates.

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Code availability

Analysis scripts are available at: https://github.com/s6juncheng/Cheng_et_al_2016

Acknowledgements

We are thankful to 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. JC and ŽA is 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

Anderson JS & Parker RP (1998) 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.* **17:** 1497–506

Barbosa C, Peixeiro I & Romão L (2013) Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. *PLoS Genet.* **9:** e1003529

Barnes CA (1998) Upf1 and Upf2 proteins mediate normal yeast mRNA degradation when translation initiation is limited. *Nucleic Acids Res.* **26:** 2433–2441

Bazzini AA, Viso F, Moreno-mateos MA, Johnstone TG & Charles E (2016) Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *EMBO J.*: 1–17

Boeck R, Tarun S, Rieger M, Deardorff JA, Mu S & Sachs AB (1996) The Yeast Pan2 Protein Is Required for Poly (A) -binding Protein-stimulated Poly (A) -nuclease Activity *J. Biol. Chem.*. **271:** 432–438

Boël G, Letso R, Neely H, Price WN, Wong K, Su M, Luff JD, Valecha M, Everett JK, Acton TB, Xiao R, Montelione GT, Aalberts DP & Hunt JF (2016) Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature* **529:** 358–363

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:** 334–45

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

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Brown CE, Tarun SZ, Boeck R, Sachs AB, Sachs AB & Chem JB (1996) PAN3 Encodes a Subunit of the Pab1p-Dependent Poly (A) Nuclease in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **16:** 5744–5753

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

Caponigro G & Parker R (1996) Mechanisms and control of mRNA turnover in Saccharomyces cerevisiae. *Microbiol. Rev.* **60:** 233–249

Collart MA (2003) Global control of gene expression in yeast by the Ccr4-Not complex. *Gene* **313:** 1–16

Colomina N, Ferrezuelo F, Wang H, Aldea M & Garí E (2008) Whi3, a developmental regulator of budding yeast, binds a large set of mRNAs functionally related to the endoplasmic reticulum. *J. Biol. Chem.* **283:** 28670–28679

Cui Y, Hagan KW, Zhang S & Peltz SW (1995) Identification and characterization of genes that are required for the accelerated degradation of mRNAs containing a premature translational termination codon. *Genes Dev.* **9:** 423–36

Duttagupta R, Tian B, Wilusz CJ, Danny T, Soteropoulos P, Ouyang M, Joseph P, Peltz SW, Khounh DT & Dougherty JP (2005) Global Analysis of Pub1p Targets Reveals a Coordinate Control of Gene Expression through Modulation of Binding and Stability Global Analysis of Pub1p Targets Reveals a Coordinate Control of Gene Expression through Modulation of Binding and Stability *Mol. Cell. Biol.*. **25:** 5499–5513

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–E2801

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

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

Gaba A, Jacobson A & Sachs MS (2005) Ribosome occupancy of the yeast CPA1 upstream open reading frame termination codon modulates nonsense-mediated mRNA decay. *Mol. Cell* **20:** 449–460

Garí E, Volpe T, Wang H, Gallego C, Futcher B & Aldea M (2001) Whi3 binds the mRNA of the G1 cyclin CLN3 to modulate cell fate in budding yeast. *Genes Dev.* **15:** 2803–2808

Garneau NL, Wilusz J & Wilusz CJ (2007) The highways and byways of mRNA decay. *Nat. Rev. Mol. Cell Biol.* **8:** 113–126

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

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

824

Gerber AP, Herschlag D & Brown PO (2004) Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol.* **2:** e79

Goldstrohm AC, Seay DJ, Hook BA & Wickens M (2007) PUF protein-mediated deadenylation is catalyzed by Ccr4p. *J. Biol. Chem.* **282:** 109–114

Gupta I, Clauder-Münster S, Klaus B, Järvelin AI, Aiyar RS, Benes V, Wilkening S, Huber W, Pelechano V & Steinmetz LM (2014) Alternative polyadenylation diversifies post-transcriptional regulation by selective RNA-protein interactions. *Mol. Syst. Biol.* **10:** 719

He F, Li X, Spatrick P, Casillo R, Dong S & Jacobson A (2003) Genome-Wide Analysis of mRNAs Regulated by the Nonsense-Mediated and 5'to 3' mRNA Decay Pathways in Yeast. *Mol. Cell* **12:** 1439–1452

Hoekema A, Kastelein ROBA, Vasser M & Boer HADE (1987) Codon Replacement in the PGKI Gene of Saccharomyces cerevisiae : Experimental Approach To Study the Role of Biased Codon Usage in Gene Expression. **7:** 2914–2924

Holmes KJ, Klass DM, Guiney EL & Cyert MS (2013) Whi3, an S. cerevisiae RNA-binding protein, is a component of stress granules that regulates levels of its target mRNAS. *PLoS One* **8:** 10–12

Hsu CL & Stevens A (1993) 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.* **13:** 4826–4835

Huch S & Nissan T (2014) Interrelations between translation and general mRNA degradation in yeast. *Wiley Interdiscip. Rev. RNA* **5:** 747–763

Hug N, Longman D & Cáceres JF (2015) Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res.* **44:** 1483–1495

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 (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44:** 283–292

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:** 420–433

Lee RC (1993) The C . elegans Heterochronic Gene lin-4 Encodes Small RNAs with Antisense Complementarity to & II-14. *Cell* **75:** 843–854

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Leeds P, Wood JM, Lee B & Culbertson MR (1992) Gene Products That Promote mRNA Turnover in Saccharomyces cerevisiaet. *Mol. Cell. Biol.* **12:** 2165–2177

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL, Thirumalai D, Lee N, Woodson S, Klimov D, Tinoco I, Uhlenbeck O, Levine M, Tinoco I, Borer P, Dengler B, Levine N, Uhlenbeck O, Crothers D, et al (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6:** 26

McCaughan KK, Brown CM, Dalphin ME, Berry MJ & Tate WP (1995) Translational termination efficiency in mammals is influenced by the base following the stop codon. *Proc. Natl. Acad. Sci. U. S. A.* **92:** 5431–5435

Mishima Y & Tomari Y (2016) Codon Usage and 3′ UTR Length Determine Maternal mRNA Stability in Zebrafish. *Mol. Cell* **61:** 874–885

Muhlrad D, Decker CJ & Parker R (1994) Deadenylation of the unstable mRNA encoded by the yeast MFA2 gene leads to decapping followed by 5' ??? 3' digestion of the transcript. *Genes Dev.* **8:** 855–866

Muhlrad D, Decker CJ & Parker ROY (1995) Turnover Mechanisms of the Stable Yeast PGK1 mRNA. *Mol. Cell. Biol.* **15:** 2145–2156

Olivas W & Parker R (2000) The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast. *EMBO J.* **19:** 6602-6611

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

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

Pilkington GR & Parker R (2008) Pat1 contains distinct functional domains that promote P-body assembly and activation of decapping. *Mol. Cell. Biol.* **28:** 1298–1312

Presnyak V, Alhusaini N, Chen Y-H, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR & Coller J (2015) Codon Optimality Is a Major Determinant of mRNA Stability. *Cell* **160:** 1111–1124

Rabani M, Raychowdhury R, Jovanovic M, Rooney M, Stumpo DJ, Pauli A, Hacohen N, Schier AF, Blackshear PJ, Friedman N, Amit I & Regev A (2014) High-Resolution Sequencing and Modeling Identifies Distinct Dynamic RNA Regulatory Strategies. *Cell* **159:** 1698–1710

Radhakrishnan A, Chen Y-H, Martin S, Alhusaini N, Green R & Coller J (2016) The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. *Cell* **167:** 122–128.e9

Roy B & Jacobson A (2013) The intimate relationships of mRNA decay and translation. *Trends Genet.* **29:** 691–699

Sabi R & Tuller T (2014) Modelling the Efficiency of Codon-tRNA Interactions Based on Codon Usage Bias. *DNA Res.***:** 1–15

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Schwalb B, Michel M, Zacher B, Demel C, Tresch A & Gagneur J (2016) TT-seq maps the human transient transcriptome. *Science* **352:** 1225-1228

Schwanhäusser B, Busse D & Li N (2011) Global quantification of mammalian gene expression control. *Nature* **473:** 337–342

Schwartz DC & Parker R (1999) Mutations in translation initiation factors lead to increased rates of deadenylation and decapping of mRNAs in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **19:** 5247–5256

Shalem O, Sharon E, Lubliner S, Regev I, Lotan-Pompan M, Yakhini Z & Segal E (2015) Systematic Dissection of the Sequence Determinants of Gene 3' End Mediated Expression Control. *PLOS Genet.* **11:** e1005147

She M, Decker CJ, Svergun DI, Round A, Chen N, Muhlrad D, Parker R & Song H (2008) Structural Basis of Dcp2 Recognition and Activation by Dcp1. *Mol. Cell* **29:** 337–349

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W & Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050

Sun M, Schwalb B, Pirkl N, Maier KC, Schenk A, Failmezger H, Tresch A & Cramer P (2013) Global analysis of Eukaryotic mRNA degradation reveals Xrn1-dependent buffering of transcript levels. *Mol. Cell* **52:** 52–62

Tucker M, Valencia-sanchez MA, Staples RR, Chen J, Denis CL & Parker R (2001) The Transcription Factor Associated Ccr4 and Caf1 Proteins Are Components of the Major Cytoplasmic mRNA Deadenylase in Saccharomyces cerevisiae. *Cell* **104:** 377–386

Tuller T & Zur H (2014) Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Res.* **43:** 13–28

Vasudevan S & Peltz SW (2001) Regulated ARE-mediated mRNA decay in Saccharomyces cerevisiae. *Mol. Cell* **7:** 1191–1200

Vogel C, Abreu R de S, Ko D, Le S-YY, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte EM & Penalva LO (2010) Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* **6:** 400

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

Williams I, Richardson J, Starkey A & Stansfield I (2004) Genome-wide prediction of stop codon readthrough during translation in the yeast Saccharomyces cerevisiae. *Nucleic Acids Res.* **32:** 6605–6616

Yamashita A, Chang T-C, Yamashita Y, Zhu W, Zhong Z, Chen C-YA & Shyu A-B (2005) Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

turnover. *Nat. Struct. Mol. Biol.* **12:** 1054–1063

Zeisel A, Köstler WJ, Molotski N, Tsai JM, Krauthgamer R, Jacob-Hirsch J, Rechavi G, Soen Y, Jung S, Yarden Y & Domany E (2011) Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Mol. Syst. Biol.* **7:** 529

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

Zur H & Tuller T (2013) New Universal Rules of Eukaryotic Translation Initiation Fidelity. **9:** e1003136

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sequence features explain most of the mRNA stability variation across genes

Jun Cheng, Žiga Avsec, Julien Gagneur

November 2016

1 Main Figures

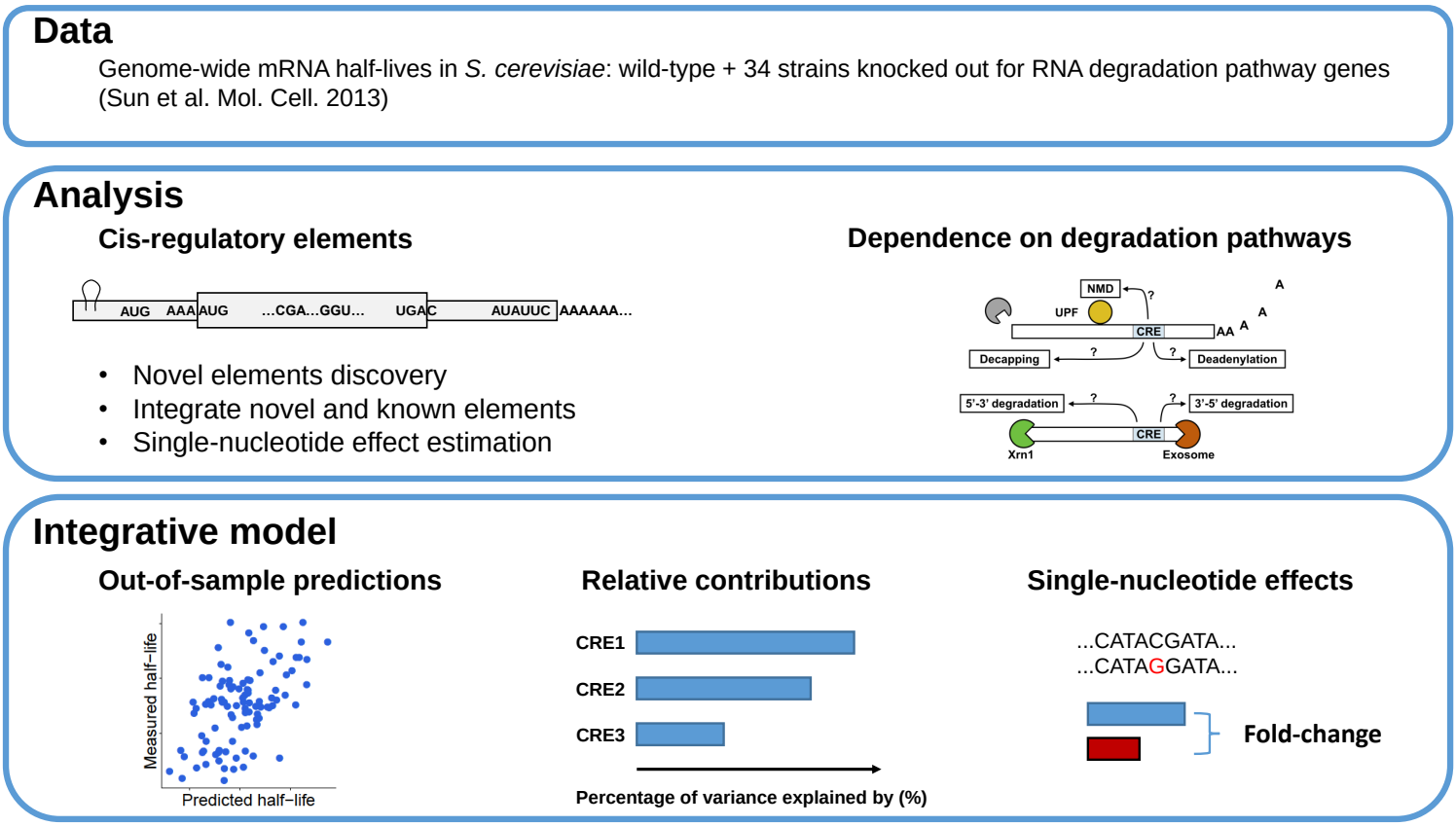


Figure 1: **Study overview.** The goal of this study is to discover and integrate cis-regulatory mRNA elements affecting mRNA stability and assess their dependence on mRNA degradation pathways. **Data)** We obtained *S. cerevisiae* genome-wide half-life data from wild-type (WT) as well as from 34 knockout 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.

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

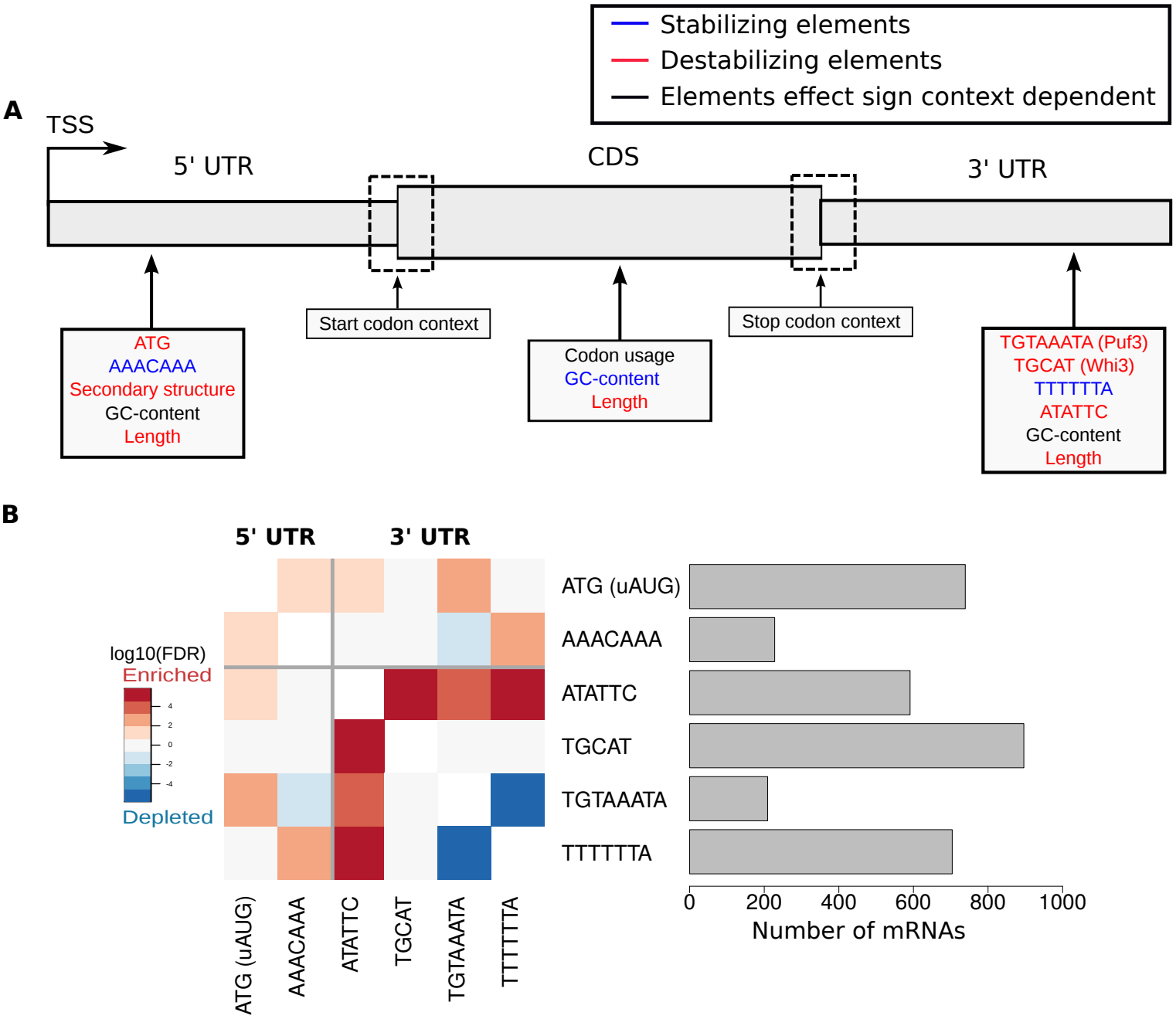


Figure 2: **Overview of identified or collected sequence features.** (**A**) Sequence features that were identified or collected from different sequence regions in this study. When applicable, stabilizing elements are shown in blue, destabilizing in red. (**B**) Co-occurrence significance (FDR, Fisher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

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

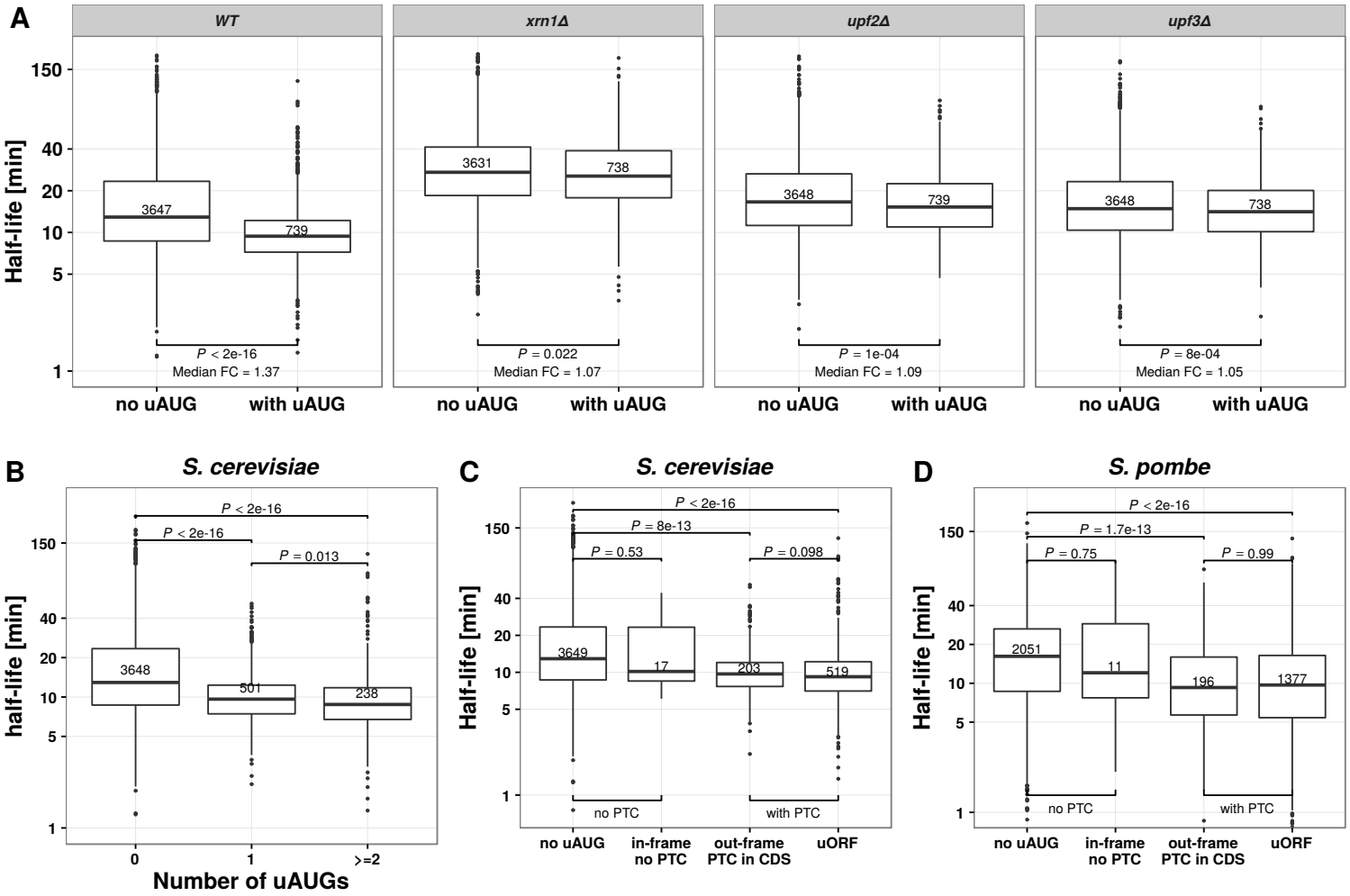


Figure 3: **Upstream AUG codon (uAUG) destabilize mRNA.** (**A**) Distribution of mRNA half-life for mRNAs without uAUG (left) and with at least one uAUG (right) in, from left to right: wild type, *XRN1*, *UPF2* and *UPF3* knockout *S. cerevisiae* strains. Median fold-change (Median FC) calculated by 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.

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

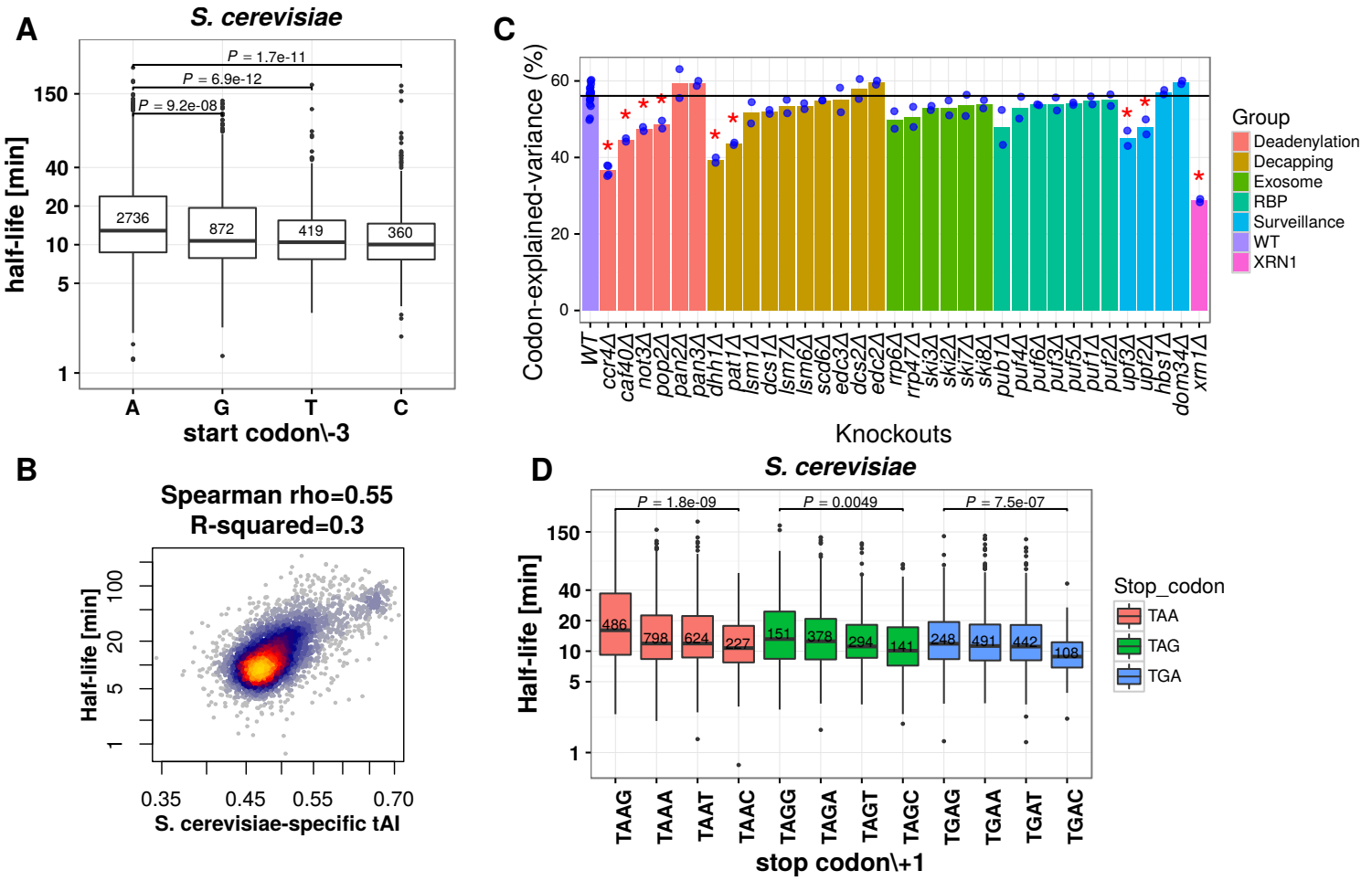


Figure 4: **Translation initiation, elongation and termination features associate with mRNA half-life.** (**A**) Distribution of half-life for mRNAs grouped by the third nucleotide before the start codon. Group sizes (numbers in boxes) show that nucleotide frequency at this position positively associates with half-life. (**B**) mRNA half-life (y-axis) versus species-specific tRNA adaptation index (sTAI) (x-axis). (**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.

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

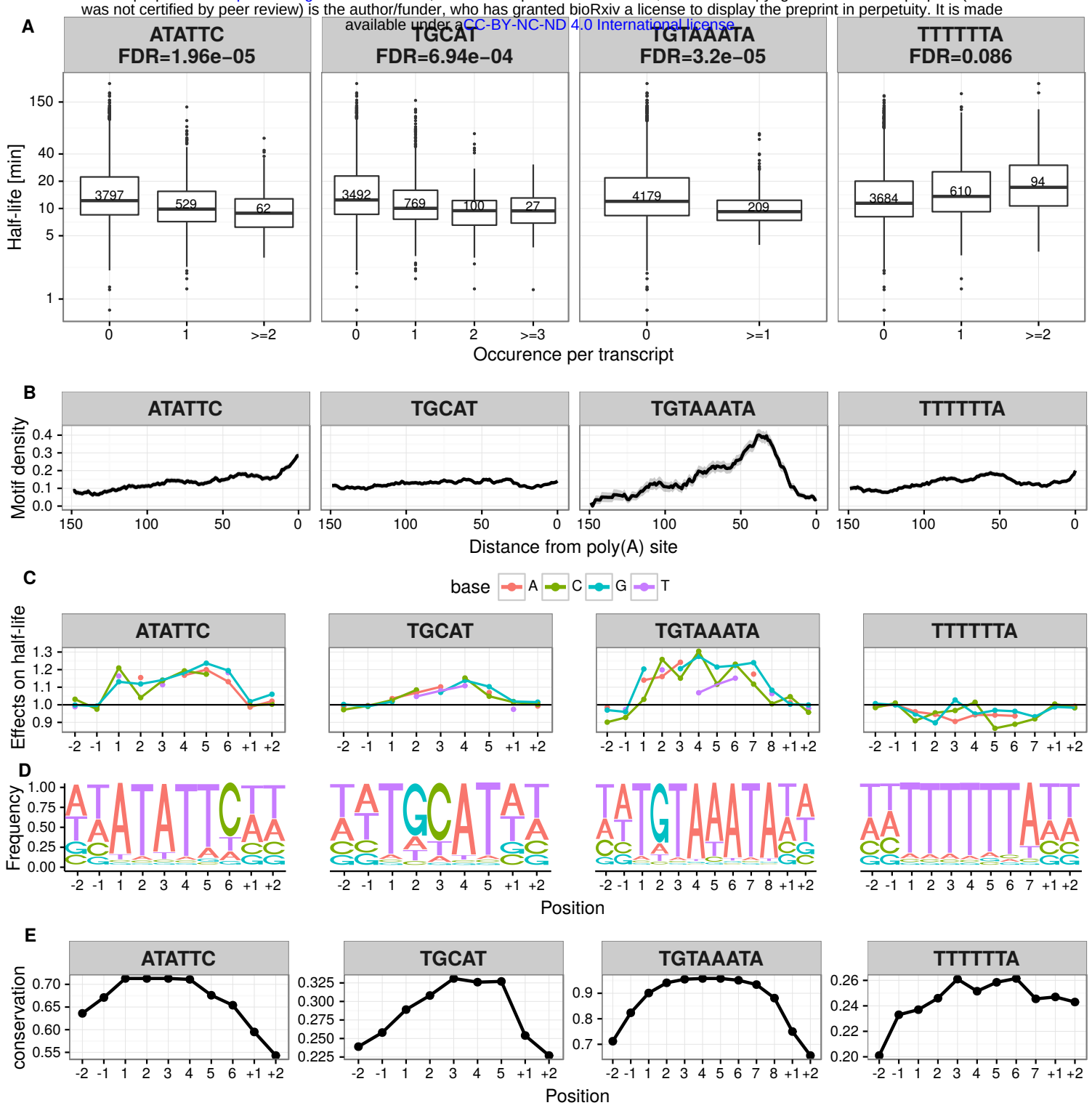
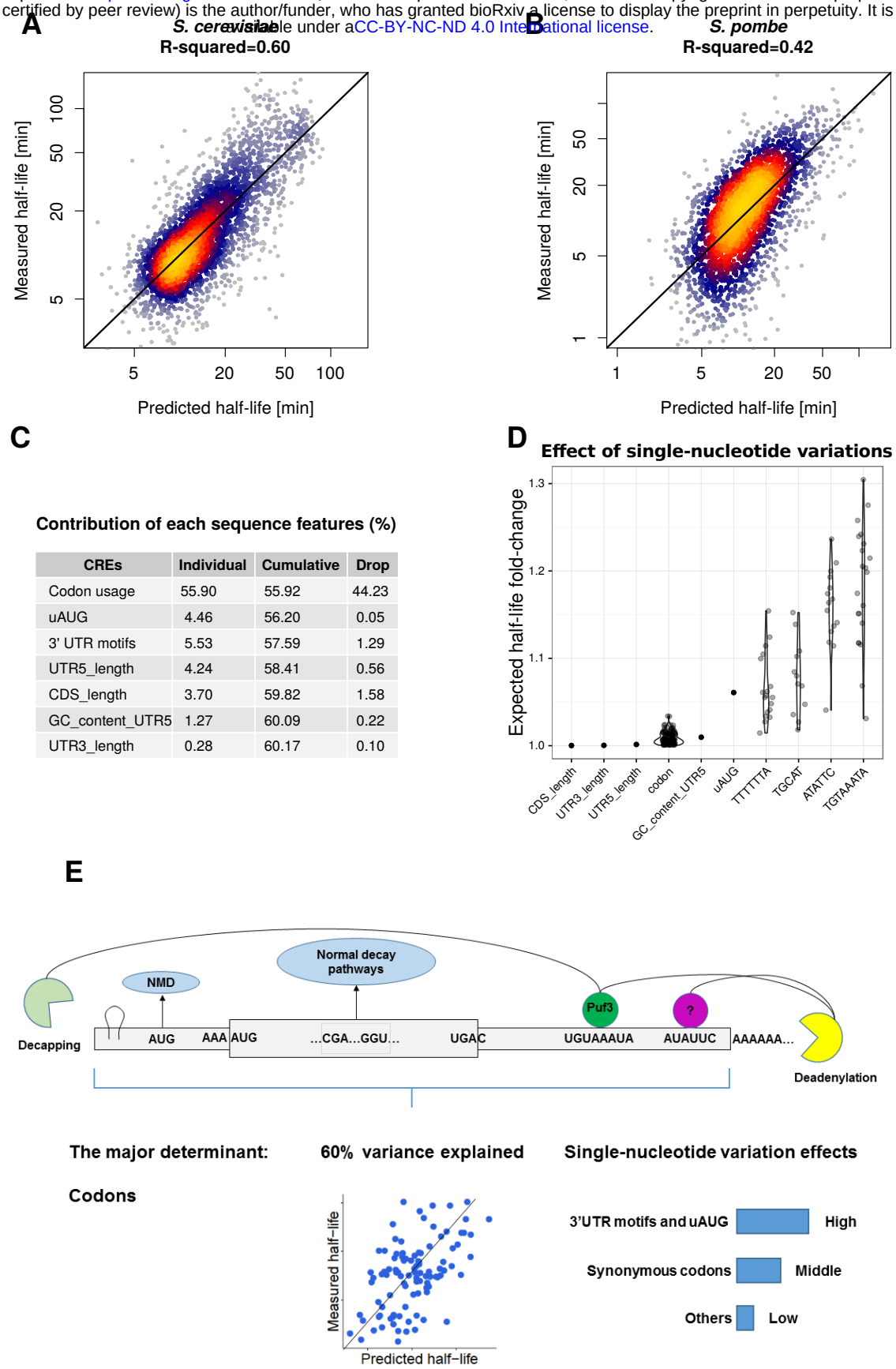


Figure 5: **3'UTR half-life determinant motifs in *S. cerevisiae*.** (**A**) Distribution of half-lives for mRNAs grouped by the number of occurrence(s) of the motif ATATTC, TGCAT (Whi3), TGTAAATA (Puf3) and TTTTTTA respectively in their 3'UTR sequence. Numbers in the boxes represent the number of members 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).

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



CREs	Individual	Cumulative	Drop
Codon usage	55.90	55.92	44.23
uAUG	4.46	56.20	0.05
3' UTR motifs	5.53	57.59	1.29
UTR5_length	4.24	58.41	0.56
CDS_length	3.70	59.82	1.58
GC_content_UTR5	1.27	60.09	0.22
UTR3_length	0.28	60.17	0.10

Figure 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 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.

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

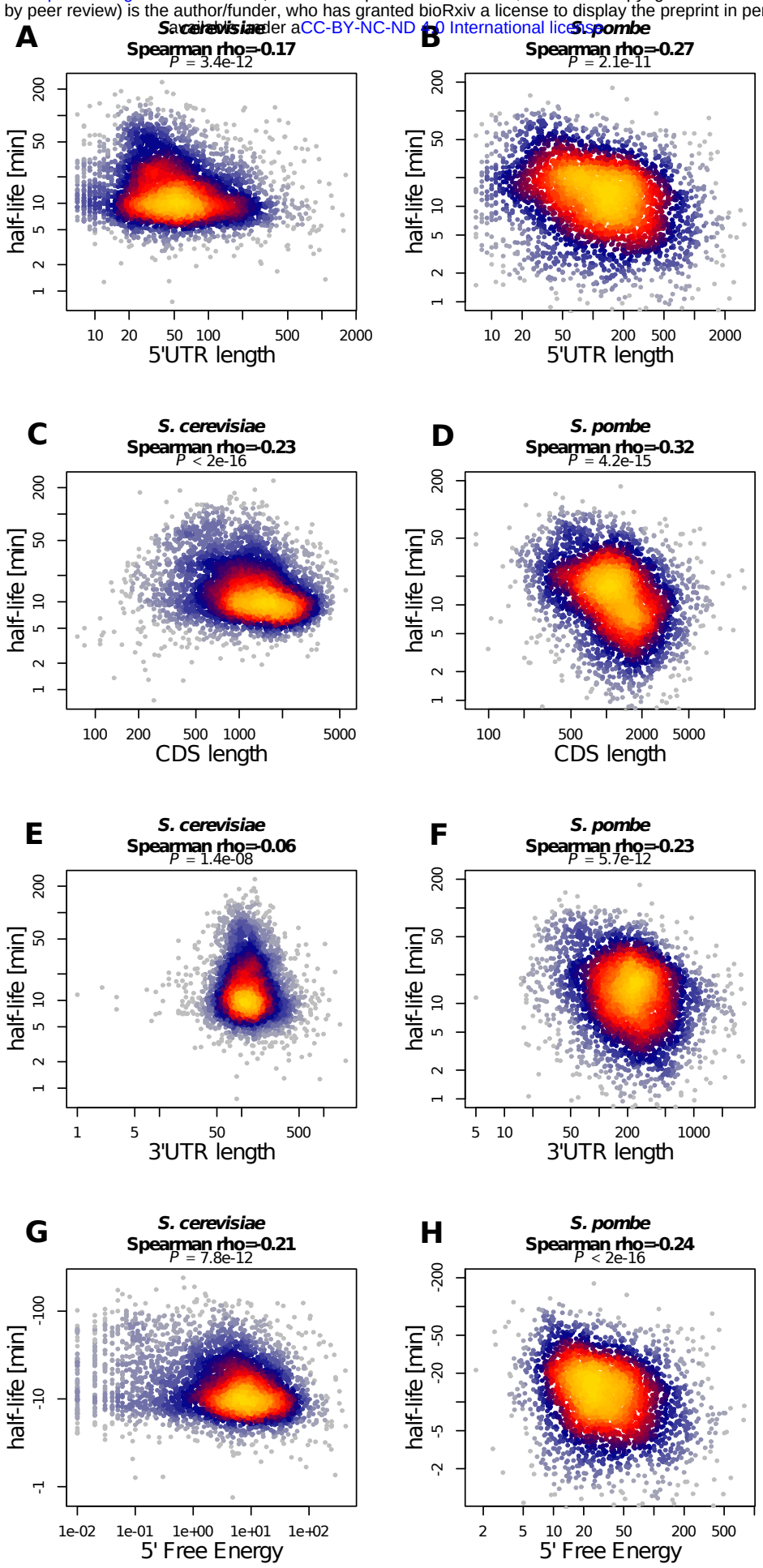


Figure EV1: **Length of 5'UTR, CDS and 3'UTR as well as 5' folding energy correlate with mRNA half-life.** (**A-B**) 5'UTR length (x-axis) versus half-life (y-axis) for *S. cerevisiae* (A) and *S. pombe* (B). (**C-D**) CDS length (x-axis) versus half-life (y-axis) for *S. cerevisiae* (C) and *S. pombe* (D). (**E-F**) 3'UTR length (x-axis) versus half-life (y-axis) for *S. cerevisiae* (E) and *S. pombe* (F). (**G-H**) 5' free energy (x-axis) versus half-life (y-axis) for *S. cerevisiae* (G) and *S. pombe* (H).

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

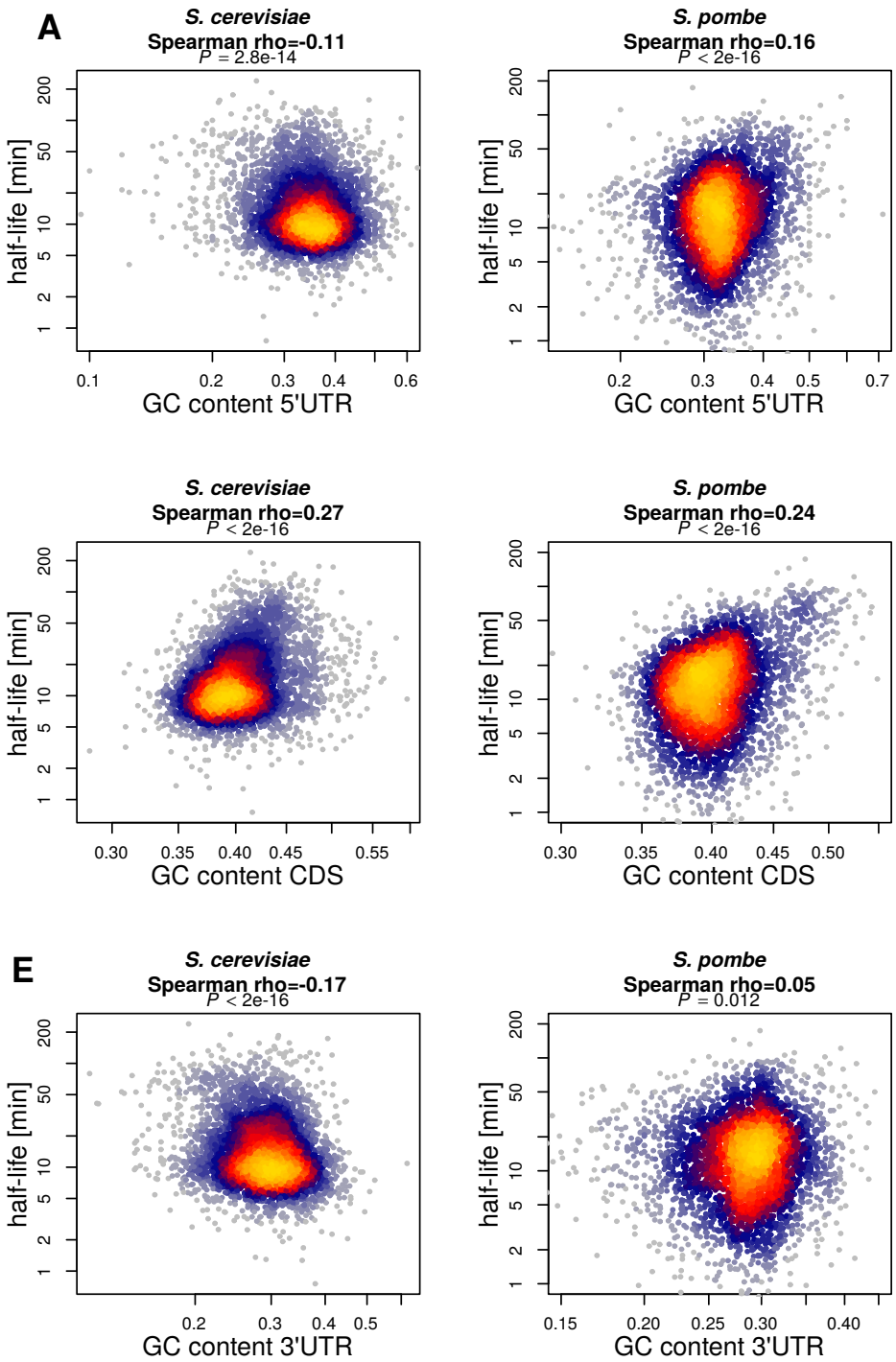


Figure EV2: **GC content of 5'UTR, CDS and 3'UTR correlate with mRNA half-life** (**A-B**) 5'UTR GC content (x-axis) versus half-life (y-axis) for *S. cerevisiae* (A) and *S. pombe*. (**C-D**) CDS GC content (x-axis) versus half-life (y-axis) for *S. cerevisiae* (C) and *S. pombe* (D). (**E-F**) 3'UTR GC content (x-axis) versus half-life (y-axis) for *S. cerevisiae* (E) and *S. pombe* (F).

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

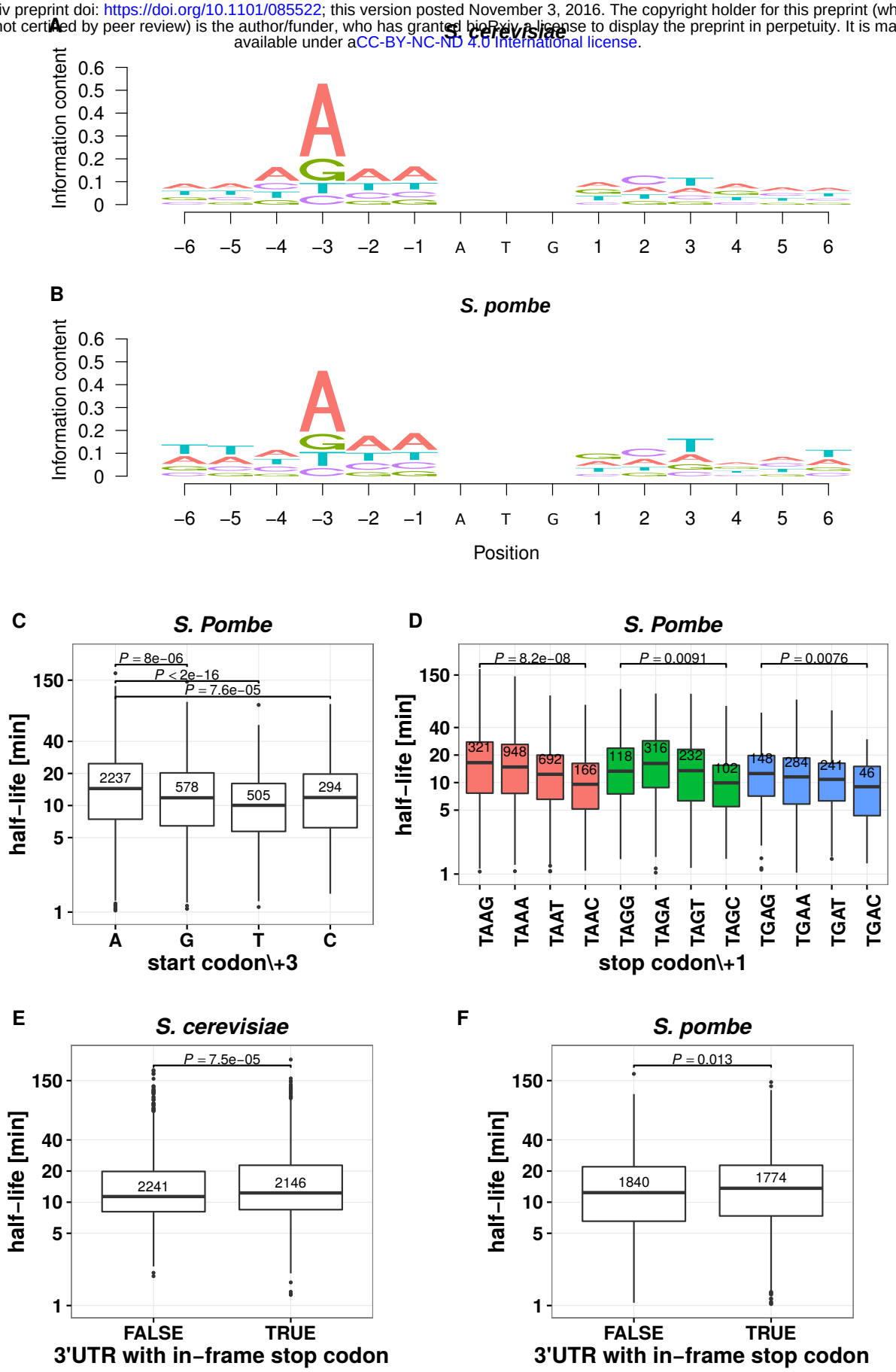


Figure EV3: **Start codon context (Kozak sequence) and *S. pombe* half-life versus start and stop codon context.** (**A-B**) Start codon context (Kozak sequence) generated from 4388 *S. cerevisiae* genes (A) and 3713 *S. pombe* genes (B). (**C**) Distribution of half-life for mRNAs grouped by the third 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. (**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.

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

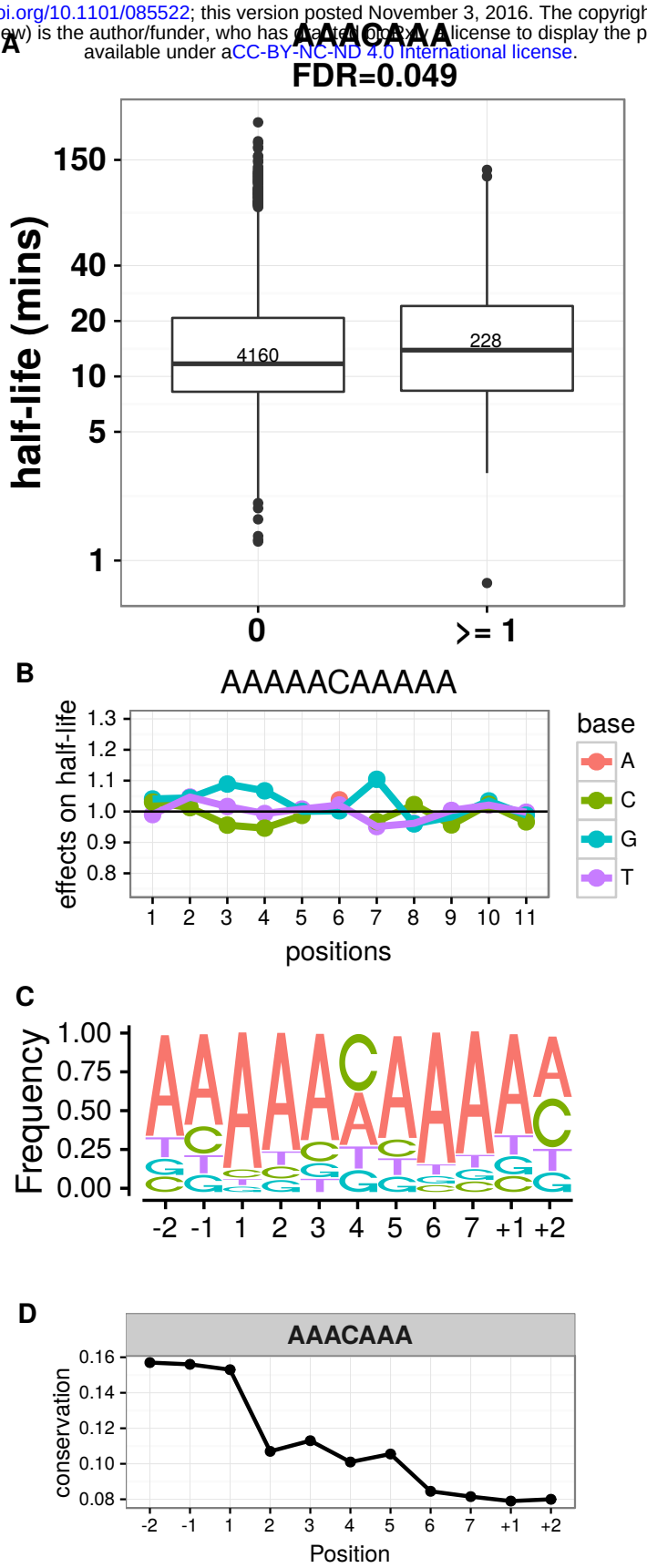


Figure EV4: ***S. cerevisiae* 5'UTR mRNA half-life associated motif.** (**A**) Distribution of half-lives for mRNAs grouped by the number of occurrence(s) of the motif AAACAAA in their 5'UTR sequence. Numbers in the boxes represent the number of members in each box. FDR were reported from the linear 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).

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

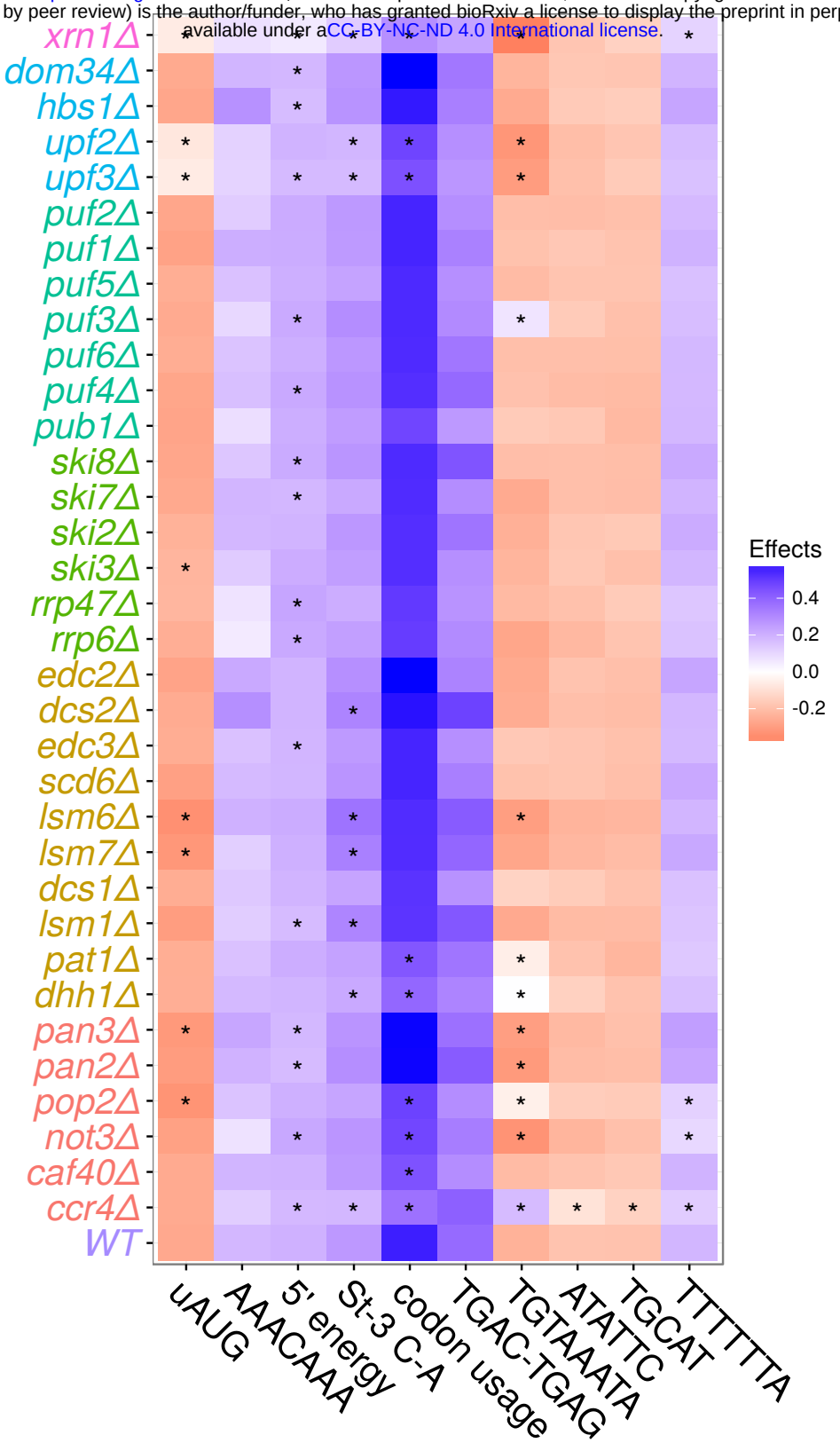


Figure EV5: **Summary of CREs effect changes across all 34 knockouts comparing with WT.** Colour represent the relative effect size (motifs, St-3 C-A, TGAG-TGAC, uAUG), correlation (5' folding energy) or explained variance (codon usage) upon knockout of different genes (y-axis) (Materials and Methods 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; TGAG-TGAC: relative median half-life difference between genes with stop codon +1 TAAG and TGAC. codon usage: codon usage explained mRNA half-lives variance. uAUG: relative median half-life difference between genes without and with upstream AUG in the 5'UTR (Materials and Methods)

bioRxiv preprint doi: https://doi.org/10.1101/085522; this version posted November 3, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

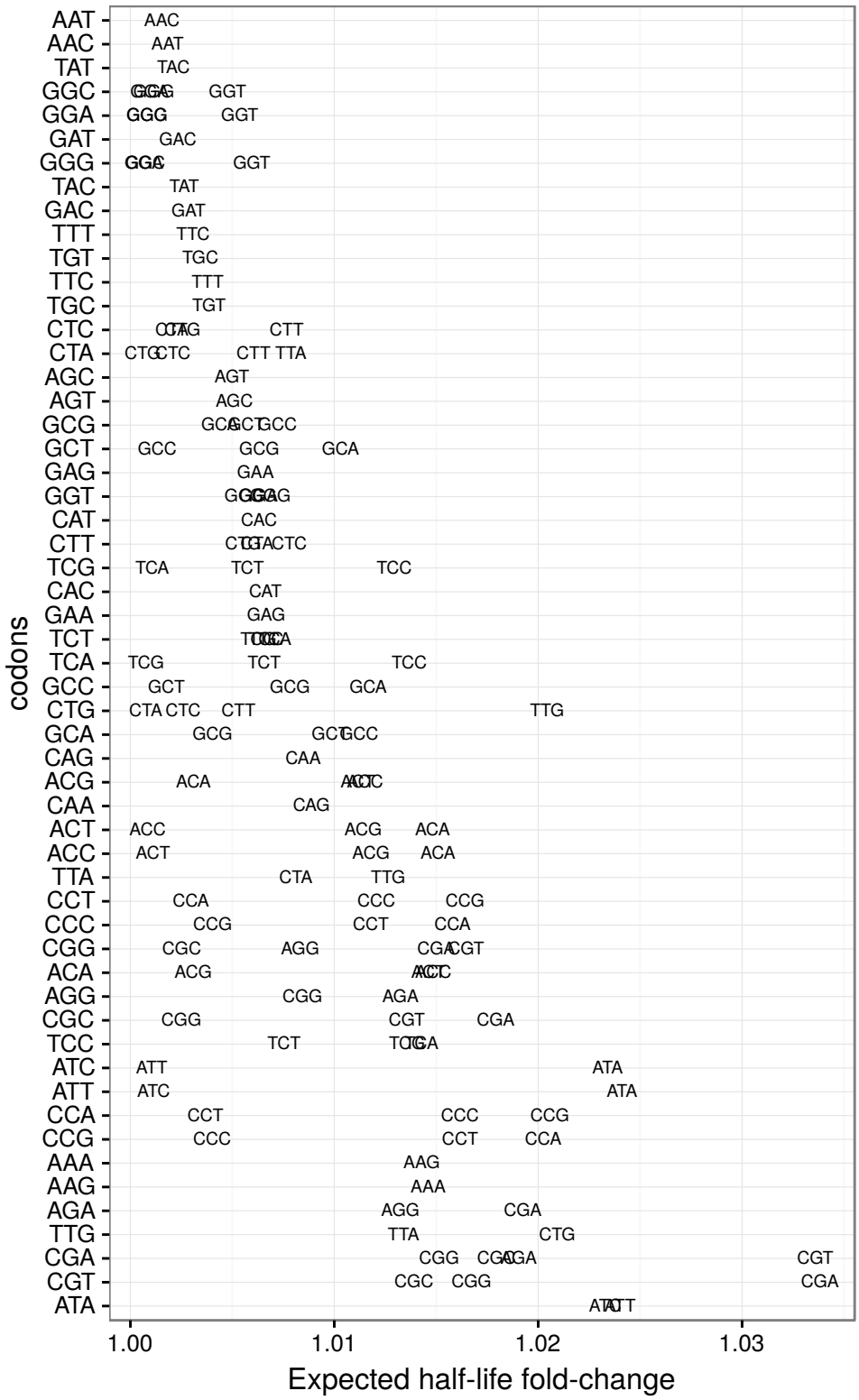


Figure EV6: **Predicted effects of synonymous codon transitions on half-life.** Expected half-life fold-change (x-axis) at each synonymous codon transitions. Each row represent transition from one codon (y-axis) to its synonymous partners. Only synonymous codons that differ by one base were considered.