

1 Sequence features explain most of the mRNA 2 stability variation across genes in yeast

3

4 Jun cheng^{1,2}, Žiga Avsec^{1,2}, Julien Gagneur^{1,2,*}

5 1. Department of Informatics, Technical University of Munich, Boltzmannstraße 3,
6 85748 Garching, Germany

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

9 * Correspondence: gagneur@in.tum.de

10

11 Abstract

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

30

31 Author Summary

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

47

48 **Introduction**

49

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

62

63 How mRNA stability is encoded in a gene sequence has long been a subject of
64 study. Cis-regulatory elements (CREs) affecting mRNA stability are mainly encoded

65 in the mRNA itself. They include but are not limited to secondary structure (8,9),
66 sequence motifs present in the 3'UTR including binding sites of RNA-binding proteins
67 (10–12), and, in higher eukaryotes, microRNAs (13). Moreover, translation-related
68 features are frequently associated with mRNA stability. For instance, inserting strong
69 secondary structure elements in the 5'UTR or modifying the translation start codon
70 context strongly destabilizes the long-lived *PGK1* mRNA in *S. cerevisiae* (14,15).
71 Codon usage, which affects translation elongation rate, also regulates mRNA stability
72 (16–19). Further correlations between codon usage and mRNA stability have been
73 reported in *E. coli* and *S. pombe* (20,21).

74
75 Since the RNA degradation machineries are well conserved among eukaryotes, the
76 pathways have been extensively studied using *S. cerevisiae* as a model organism
77 (22,23). The general mRNA degradation pathway starts with the removal of the
78 poly(A) tail by the Pan2/Pan3 (24) and Ccr4/Not complexes (25). Subsequently,
79 mRNA is subjected to decapping carried out by Dcp2 and promoted by several
80 factors including Dhh1 and Pat1 (26,27). The decapped and deadenylated mRNA
81 can be rapidly degraded in the 3' to 5' direction by the exosome (28) or in the 5' to 3'
82 direction by Xrn1 (29). Further mRNA degradation pathways are triggered when
83 aberrant translational status is detected, including Nonsense-mediated decay (NMD),
84 No-go decay (NGD) and Non-stop decay (NSD) (22,23).

85
86 Despite all this knowledge, prediction of mRNA half-life from a gene sequence is still
87 not established. Moreover, most of the mechanistic studies so far could only be
88 performed on individual genes or reporter genes and it is therefore unclear how the
89 effects generalize genome-wide. A recent study showed that translation-related
90 features can be predictive for mRNA stability (30). Although this analysis supported
91 the general correlation between translation and stability (31,32), the model was not
92 based purely on sequence-derived features but contained measured transcript
93 properties such as ribosome density and normalized translation efficiencies. Hence,
94 the question of how half-life is genetically encoded in mRNA sequence remains to be
95 addressed. Additionally, the dependences of sequence features to distinct mRNA
96 degradation pathways have not been systematically studied. One example of this is
97 codon-mediated stability control. Although a causal link from codon usage to mRNA
98 half-life has been shown in a wide range of organisms (16–19), the underlying

99 mechanism remains poorly understood. In *S. cerevisiae*, reporter gene experiments
100 showed that codon-mediated stability control depends on the RNA helicase Dhh1
101 (33). However, neither is it clear how this generalizes genome-wide nor the role of
102 other closely related genes has been systematically assessed.

103
104 Here, we used an integrative approach where we mathematically modelled mRNA
105 half-life as a function of its sequence and applied it to *S. cerevisiae*. For the first time,
106 our model can explain most of the between-gene half-life variance from sequence
107 alone. Using a semi-mechanistic model, we could interpret individual sequence
108 features in the 5'UTR, coding region, and 3'UTR. Our approach *de novo* recovered
109 known cis-regulatory elements and identified novel ones. Quantification of the
110 respective contributions revealed that codon usage is the major contributor to mRNA
111 stability. Applying the modeling approach to *S. pombe* supports the generality of
112 these findings. We systematically assessed the dependencies of these sequence
113 elements on mRNA degradation pathways using half-life data for 34 knockout strains,
114 and notably delineated novel pathways through which codon usage affects half-life.

115

116 **Results**

117

118 **Regression reveals novel mRNA sequence features associated with** 119 **mRNA stability**

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

131 folding energies correlated positively with half-life in both yeast species, whereas
132 correlations of GC content varied with species and gene regions.

133

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

143

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

146

147 **Upstream AUGs destabilize mRNAs by triggering nonsense-** 148 **mediated decay**

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

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

168

169 **Translation initiation predicts mRNA stability**

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

171

172 First, longer 5'UTRs associated with less stable mRNAs ($\rho = -0.17$, $P < 2 \times 10^{-16}$ for
173 *S. cerevisiae* and $\rho = -0.26$, $P = < 2 \times 10^{-16}$ for *S. pombe*, S1A, B Fig). In mouse
174 cells, mRNA isoforms with longer 5'UTR are translated with lower efficiency (42),
175 possibly because longer 5'UTR generally harbor more translation-repressive
176 elements. Hence, longer 5'UTR may confer mRNA instability by decreasing
177 translation initiation and therefore decreasing the protection by the translation
178 machinery.

179

180 Second, a significant association between the third nucleotide 5' of the start codon
181 and mRNA half-life was observed (Fig 4A). The median half-life correlated with the
182 nucleotide frequency at this position (S4A Fig), associating with 1.28 median fold-
183 change ($P = 1.7 \times 10^{-11}$) between the adenosine (2,736 genes, most frequent) and
184 cytosine (360 genes, the least frequent). The same correlation was also significant
185 for *S. pombe* ($P = 1.2 \times 10^{-4}$, S4A, B Fig). Functional effect of the start codon context
186 on mRNA stability has been established as the long-lived *PGK1* mRNA was strongly
187 destabilized when substituting the sequence context around its start codon with the
188 one from the short-lived *MFA2* mRNA (15). Our genome-wide analysis indicates that
189 this effect generalizes to other genes. The start codon context, which controls
190 translation initiation efficiency (43,44), increases ribosome density which may protect
191 mRNA from degradation as hypothesized by Edri and Tuller (31).

192

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

196

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

199

200 **Codon usage regulates mRNA stability through common mRNA** 201 **decay pathways**

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

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

222

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

230 decay pathway. This implies that the effect of codon usage is unlikely due to stalled
231 ribosomes at non-optimal codons.

232

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

236

237 **Stop codon context associates with mRNA stability**

238 Linear regression against the 6 bases 5' and 3' of the stop codon revealed the first
239 nucleotide 3' of the stop codon to most strongly associate with mRNA stability. This
240 association was observed for each of the three possible stop codons, and for each
241 codon a cytosine significantly associated with lower half-life (all $P < 0.01$, Fig 4D).
242 This also held for *S. pombe* (all $P < 0.01$, S4D Fig). A cytosine following the stop
243 codon structurally interferes with stop codon recognition (48), thereby leading to stop
244 codon read-through events (49). Of all combinations, TGA-C is known to be the
245 leakiest stop codon context (50) and also associated with shortest mRNA half-life
246 (Fig 4D). These results are consistent with non-stop decay, a mechanism that
247 triggers exosome-dependent RNA degradation when the ribosome reaches the
248 poly(A) tail. Consistent with this interpretation, mRNAs with additional in-frame stop
249 codons in the 3'UTR, which are over-represented in yeast (51), exhibited significantly
250 higher half-life ($P = 7.5 \times 10^{-5}$ for *S. cerevisiae* and $P = 0.011$ for *S. pombe*, S4E, F
251 Fig). However, the association between the stop codon context and half-life was not
252 weakened in mutants of the Ski complex, which is required for the cytoplasmic
253 functions of the exosome (S6 Fig). These results indicate that the fourth nucleotide
254 after the stop codon is an important determinant of mRNA stability, likely because of
255 translational read-through.

256

257 **Sequence motifs in 3'UTR**

258 Four motifs in the 3'UTR were found to be significantly associated with mRNA
259 stability (Fig 5A, all FDR < 0.1 , Materials and Methods). This analysis recovered
260 three described motifs: the Puf3 binding motif TGTAATA (35), the Whi3 binding
261 motif TGCAT (37,38), and a poly(U) motif TTTTTTA, which can be bound by Pub1
262 (12), or is part of the long poly(U) stretch that forms a looping structure with poly(A)

263 tail (9). We also identified a novel motif, ATATTC, which associated with lower mRNA
264 half-life. This motif was reported to be enriched in 3'UTRs for a cluster of genes with
265 correlated expression pattern (52), but its function remains unknown. Genes
266 harboring this motif are significantly enriched for genes involved in oxidative
267 phosphorylation (Bonferroni corrected $P < 0.01$, Gene Ontology analysis,
268 Supplementary Methods and S3 Table).

269
270 Four lines of evidence supported the potential functionality of the new motif. First, it
271 preferentially localizes in the vicinity of the poly(A) site (Fig 5B), and functionally
272 depends on Ccr4 (S6 Fig), suggesting a potential interaction with deadenylation
273 factors. Second, single nucleotide deviations from the consensus sequence of the
274 motif associated with decreased effects on half-life (Fig 5C, linear regression allowing
275 for one mismatch, Materials and Methods). Moreover, the flanking nucleotides did not
276 show further associations indicating that the whole lengths of the motifs were
277 recovered (Fig 5C). Third, when allowing for one mismatch, the motif still showed
278 strong preferences (Fig 5D). Fourth, the motif instances were more conserved than
279 their flanking bases (Fig 5E).

280
281 Consistent with the role of Puf3 in recruiting deadenylation factors, Puf3 binding motif
282 localized preferentially close to the poly(A) site (Fig 5B). The effect of the Puf3 motifs
283 was significantly lower in the knockout of *PUF3* (FDR < 0.1 , S6 Fig). We also found a
284 significant dependence on the deadenylation (*CCR4*, *POP2*) and decapping (*DHH1*,
285 *PAT1*) pathways (all FDR < 0.1 , S6 Fig), consistent with previous single gene
286 experiment showing that Puf3 binding promotes both deadenylation and decapping
287 (10,53). Strikingly, Puf3 binding motif switched to a stabilization motif in the absence
288 of Puf3 and Ccr4, suggesting that deadenylation of Puf3 motif containing mRNAs is
289 not only facilitated by Puf3 binding, but also depends on it.

290
291 Whi3 plays an important role in cell cycle control (54). Binding of Whi3 leads to
292 destabilization of the *CLN3* mRNA (38). A subset of yeast genes are up-regulated in
293 the Whi3 knockout strain (38). However, it was so far unclear whether Whi3 generally
294 destabilizes mRNAs upon its binding. Our analysis showed that mRNAs containing
295 the Whi3 binding motif (TGCAT) have significantly shorter half-life (FDR = 6.9×10^{-04}).
296 Surprisingly, this binding motif is extremely widespread, with 896 out of 4,388 (20%)

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

300

301 The mRNAs harboring the TTTTTTA motif tended to be more stable and enriched for
302 translation ($P = 1.34 \times 10^{-03}$, S3 Table, Fig 5A). No positional preferences were
303 observed for this motif (Fig 5B). Effects of this motif depends on genes from Ccr4-Not
304 complex and Xrn1 (S6 Fig).

305

306 **60% between-gene half-life variation can be explained by sequence** 307 **features**

308 We next asked how well one could predict mRNA half-life from these mRNA
309 sequence features, and what their respective contributions were when considered
310 jointly. To this end, we performed a multivariate linear regression of the logarithm of
311 the half-life against the identified sequence features. The predictive power of the
312 model on unseen data was assessed using 10-fold cross validation (Material and
313 Methods). Also, motif discovery performed on each of the 10 training sets retrieved
314 the same set of motifs, showing that their identification was not due to over-fit on the
315 complete dataset. Altogether, 60% of *S. cerevisiae* half-life variance in the logarithmic
316 scale can be explained by simple linear combinations of the above sequence
317 features (Fig 6A). The median out-of-folds relative error across genes is 30%. A
318 median relative error of 30% for half-life is remarkably low because it is in the order of
319 magnitude of the expression variation that is typically physiologically tolerated, and it
320 is also about the amount of variation observed between replicate experiments (6). To
321 make sure that our findings are not biased to a specific dataset, we fitted the same
322 model to a dataset using RATE-seq (55), a modified version of the protocol used by
323 Sun and colleagues (34). On this data, the model was able to explain 50% of the
324 variance (S7 Fig). Moreover, the same procedure applied to *S. pombe* explained
325 47% of the total half-life variance, suggesting the generality of this approach.
326 Because the measures also entail measurement noise, these numbers are
327 conservative underestimations of the total biological variance explained by our
328 model.

329

330 The uAUG, 5'UTR length, 5'UTR GC content, 61 coding codons, CDS length, all four
331 3'UTR motifs, and 3'UTR length remained significant in the joint model indicating that
332 they contributed independently to half-life (complete list of p-values given in S4
333 Table). In contrast, start codon context, stop codon context, 5' folding energy, the
334 5'UTR motif AAACAAA, and 3'UTR GC content dropped below the significance when
335 considered in the joint model (Materials and Methods). This loss of statistical
336 significance may be due to lack of statistical power. Another possibility is that the
337 marginal association of these sequence features with half-life is a consequence of a
338 correlation with other sequence features. Among all sequence features, codon usage
339 as a group is the best predictor both in a univariate model (55.23%) and in the joint
340 model (43.84 %) (Fig 6C). This shows that, quantitatively, codon usage is the major
341 determinant of mRNA stability in yeast.

342

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

360

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

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

365

366

367 **Discussion**

368

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

379

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

395

396 Recently, Neymotin and colleagues (30) showed that several translation-related
397 transcript properties associated with half-life. This study derived a model explaining
398 50% of the total variance using many transcript properties including some not based
399 on sequence (ribosome profiling, expression levels, etc.). Although non-sequence
400 based predictors can facilitate prediction, they may do so because they are
401 consequences rather than causes of half-life. For instance increased half-life causes
402 higher expression level. Also, increased cytoplasmic half-life, provides a higher ratio
403 of cytoplasmic over nuclear RNA, and thus more RNAs available to ribosomes.
404 Hence both expression level and ribosome density may help making good predictions
405 of half-life, but not necessarily because they causally increase half-life. In contrast,
406 we aimed here to understand how mRNA half-life is encoded in mRNA sequence.
407 Our model was therefore solely based on mRNA sequence. This avoided using
408 transcript properties which could be consequences of mRNA stability. Hence, our
409 present analysis confirms the quantitative importance of translation in determining
410 mRNA stability that Neymotin and colleagues quantified, and anchors it into pure
411 sequence elements.

412
413 Causality cannot be proven through a regression analysis approach. Genes under
414 selection pressure for high expression levels could evolve to have both CREs for high
415 mRNA stability and CREs for high translation rate. When possible, we referred to
416 single gene studies that had proven causal effects on half-life. For novel motifs, we
417 provided several complementary analyses to further assess their potential
418 functionality. These include conservation, positional preferences, and epistasis
419 analyses to assess the dependencies on RNA degradation pathways. The novel half-
420 life associated motif ATATTC in 3'UTR is strongly supported by these complementary
421 analyses and is also significant in the joint model ($P = 5.8 \times 10^{-14}$). One of the most
422 interesting sequence features that we identified but still need to be functionally
423 assayed is the start codon context. Given its established effect on translation
424 initiation (44,62), the general coupling between translation and mRNA degradation
425 (56–58), as well as several observations directly on mRNA stability for single genes
426 (15,63), they are very likely to be functional on most genes. Consistent with this
427 hypothesis, large scale experiments that perturb 5' sequence secondary structure
428 and start codon context indeed showed a wide range of mRNA level changes in the
429 direction that we would predict (44). Altogether, such integrative approaches allow

430 the identification of candidate regulatory elements that could be functionally tested
431 later on.

432

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

442

443

444 **Materials and Methods**

445

446 **Data and Genomes**

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

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

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

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

463

464 **Analysis of knockout strains**

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

468

469 **Motif discovery**

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

479

480 **Folding energy calculation**

481 RNA sequence folding energy was calculated with RNAfold from ViennaRNA version
482 2.1.9 (66), with default parameters.

483

484 ***S. cerevisiae* conservation analysis**

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

491

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

493 Multivariate linear regression models were used to predict genome-wide mRNA half-
494 life on the logarithmic scale from sequence features. Only mRNAs that contain all
495 features were used to fit the models, resulting with 3,862 mRNAs for *S. cerevisiae*
496 and 3,130 mRNAs for *S. pombe*. Out-of-fold predictions were applied with 10-fold

497 cross validation for any prediction task in this study. For each fold, a linear model was
498 first fitted to the training data with all sequence features as covariates, then a
499 stepwise model selection procedure was applied to select the best model with
500 Bayesian Information Criterion as criteria (*step* function in R, with $k = \log(n)$). L1 or
501 L2 regularization were not necessary, as they did not improve the out-of-fold
502 prediction accuracy (tested with *glmnet* R package (68)). Motif discovery was
503 performed again at each fold. The same set of motifs were identified within each
504 training set only. A complete list of model features and their p-values in a joint model
505 for both yeast species are provided in S4 Table. For details see Supplementary
506 methods.

507

508 **Analysis of sequence feature contribution**

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

522

523 **Single-nucleotide variant effect predictions**

524 The same model that used in sequence feature contribution analysis was used for
525 single-nucleotide variant effect prediction. For **motifs**, effects of single-nucleotide
526 variants were predicted with linear model modified from (6). When assessing the
527 effect of a given motif variation, instead of estimating the marginal effect size, we
528 controlled for the effect of all other sequence features using a linear model with the
529 other features as covariates. For details see Supplementary methods. For **other**
530 **sequence features**, effects of single-nucleotide variants were predicted by

531 introducing a single nucleotide perturbation into the full prediction model for each
532 gene, and summarizing the effect with the median half-life change across all genes.
533 For details see Supplementary methods.

534

535 **Code availability**

536 Analysis scripts are available at: [https://i12g-](https://i12g-gagneurweb.in.tum.de/gitlab/Cheng/mRNA_half_life_public)
537 [gagneurweb.in.tum.de/gitlab/Cheng/mRNA_half_life_public](https://i12g-gagneurweb.in.tum.de/gitlab/Cheng/mRNA_half_life_public).

538

539 **Acknowledgements**

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

545

546 **Funding**

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

550

551 **References**

- 552 1. Shalem O, Dahan O, Levo M, Martinez MR, Furman I, Segal E, et al. Transient
553 transcriptional responses to stress are generated by opposing effects of mRNA
554 production and degradation. *Mol Syst Biol.* 2008;4(223):223.
- 555 2. Zeisel A, Köstler WJ, Molotski N, Tsai JM, Krauthgamer R, Jacob-Hirsch J, et
556 al. Coupled pre-mRNA and mRNA dynamics unveil operational strategies
557 underlying transcriptional responses to stimuli. *Mol Syst Biol.* 2011;7(529):529.
- 558 3. Rabani M, Raychowdhury R, Jovanovic M, Rooney M, Stumpo DJ, Pauli A, et
559 al. High-Resolution Sequencing and Modeling Identifies Distinct Dynamic RNA
560 Regulatory Strategies. *Cell.* Elsevier Inc.; 2014;159(7):1698–710.
- 561 4. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al.
562 Global quantification of mammalian gene expression control. *Nature.*
563 2011;473(7347):337–42.

- 564 5. Schwalb B, Michel M, Zacher B, Demel C, Tresch A, Gagneur J, et al. TT-seq
565 maps the human transient transcriptome. 2016;352(6290).
- 566 6. Eser P, Wachutka L, Maier KC, Demel C, Boroni M, Iyer S, et al. Determinants
567 of RNA metabolism in the *Schizosaccharomyces pombe* genome. *Mol Syst*
568 *Biol.* EMBO Press; 2016 Jan 1;12(2):857.
- 569 7. Duan J, Shi J, Ge X, Dölken L, Moy W, He D, et al. Genome-wide survey of
570 interindividual differences of RNA stability in human lymphoblastoid cell lines.
571 *Sci Rep.* 2013;3:1318.
- 572 8. Rabani M, Kertesz M, Segal E. Computational prediction of RNA structural
573 motifs involved in post-transcriptional regulatory processes. *Proc Natl Acad Sci.*
574 2008;105(39):14885–90.
- 575 9. Geisberg J V., Moqtaderi Z, Fan X, Ozsolak F, Struhl K. Global analysis of
576 mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast.
577 *Cell.* Elsevier Inc.; 2014;156(4):812–24.
- 578 10. Olivas W, Parker R. The Puf3 protein is a transcript-specific regulator of mRNA
579 degradation in yeast. 2000;19(23).
- 580 11. Shalgi R, Lapidot M, Shamir R, Pilpel Y. A catalog of stability-associated
581 sequence elements in 3' UTRs of yeast mRNAs. *Genome Biol.* 2005;6(10):1.
- 582 12. Duttagupta R, Tian B, Wilusz CJ, Danny T, Soteropoulos P, Ouyang M, et al.
583 Global Analysis of Pub1p Targets Reveals a Coordinate Control of Gene
584 Expression through Modulation of Binding and Stability. 2005;25(13):5499–
585 513.
- 586 13. Lee RC, Feinbaum RL, Ambros V. The *C. elegans* Heterochronic Gene *lin-4*
587 Encodes Small RNAs with Antisense Complementarity to *lin-14*. 1993;75:843–
588 54.
- 589 14. Muhlrاد D, Decker CJ, Parker R. Turnover Mechanisms of the Stable Yeast
590 PGK1 mRNA. 1995;15(4):2145–56.
- 591 15. LaGrandeur T, Parker R. The cis acting sequences responsible for the
592 differential decay of the unstable MFA2 and stable PGK1 transcripts in yeast
593 include the context of the translational start codon. *Rna.* 1999;5(3):420–33.
- 594 16. Hoekema A, Kastelein RA, Vasser M, de Boer HA. Codon replacement in the
595 PGK1 gene of *Saccharomyces cerevisiae*: experimental approach to study the
596 role of biased codon usage in gene expression. *Mol Cell Biol.* 1987 Aug
597 1;7(8):2914–24.

- 598 17. Presnyak V, Alhusaini N, Chen Y-H, Martin S, Morris N, Kline N, et al. Codon
599 Optimality Is a Major Determinant of mRNA Stability. *Cell*. 2015;160(6):1111–
600 24.
- 601 18. Mishima Y, Tomari Y. Codon Usage and 3' UTR Length Determine Maternal
602 mRNA Stability in Zebrafish. *Mol Cell*. Elsevier Inc.; 2016;61(6):874–85.
- 603 19. Bazzini AA, Viso F, Moreno-mateos MA, Johnstone TG, Charles E. Codon
604 identity regulates mRNA stability and translation efficiency during the maternal-
605 to-zygotic transition. *EMBO J*. 2016;
- 606 20. Boël G, Letso R, Neely H, Price WN, Wong K, Su M, et al. Codon influence on
607 protein expression in *E. coli* correlates with mRNA levels. *Nature*. Nature
608 Publishing Group; 2016;529(7586):358–63.
- 609 21. Harigaya Y, Parker R. Analysis of the association between codon optimality
610 and mRNA stability in *Schizosaccharomyces pombe*. *BMC Genomics*. *BMC*
611 *Genomics*; 2016;1–16.
- 612 22. Garneau NL, Wilusz J, Wilusz CJ. The highways and byways of mRNA decay.
613 *Nat Rev Mol Cell Biol*. 2007;8(2):113–26.
- 614 23. Parker R. RNA degradation in *Saccharomyces cerevisiae*. *Genetics*.
615 2012;191(3):671–702.
- 616 24. Brown CE, Tarun SZ, Boeck R, Sachs AB. PAN3 Encodes a Subunit of the
617 Pab1p-Dependent Poly(A) Nuclease in *Saccharomyces cerevisiae*.
618 1996;16(10):5744–53.
- 619 25. Tucker M, Valencia-sanchez MA, Staples RR, Chen J, Denis CL, Parker R.
620 The transcription factor associated Ccr4 and Caf1 proteins are components of
621 the major cytoplasmic mRNA deadenylase in *Saccharomyces cerevisiae*. *Cell*.
622 2001;104:377–86.
- 623 26. Pilkington GR, Parker R. Pat1 contains distinct functional domains that promote
624 P-body assembly and activation of decapping. *Mol Cell Biol*. 2008;28(4):1298–
625 312.
- 626 27. She M, Decker CJ, Svergun DI, Round A, Chen N, Muhlrads D, et al. Structural
627 Basis of Dcp2 Recognition and Activation by Dcp1. *Mol Cell*. 2008;29(3):337–
628 49.
- 629 28. Anderson JSJ, Parker P. The 3' to 5' degradation of yeast mRNAs is a general
630 mechanism for mRNA turnover that requires the SKI2 DEVH box protein and 3'
631 to 5' exonucleases of the exosome complex. *EMBO J*. 1998 Mar 2;17(5):1497–

- 632 506.
- 633 29. Hsu CL, Stevens A. Yeast cells lacking 5'-3' exoribonuclease 1 contain mRNA
634 species that are poly(A) deficient and partially lack the 5' cap structure. *Mol*
635 *Cell Biol.* 1993;13(8):4826–35.
- 636 30. Neymotin B, Ettore V, Gresham D. Multiple Transcript Properties Related to
637 Translation Affect mRNA Degradation Rates in *Saccharomyces cerevisiae*. *G3*
638 *Genes| Genomes| Genet.* 2016;6(November):3475–83.
- 639 31. Edri S, Tuller T. Quantifying the effect of ribosomal density on mRNA stability.
640 *PLoS One.* 2014;9(7):e102308.
- 641 32. Lackner DH, Beilharz TH, Marguerat S, Mata J, Watt S, Schubert F, et al. A
642 Network of Multiple Regulatory Layers Shapes Gene Expression in Fission
643 Yeast. *Mol Cell.* 2007;26(1):145–55.
- 644 33. Radhakrishnan A, Chen Y-H, Martin S, Alhusaini N, Green R, Collier J. The
645 DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by
646 Monitoring Codon Optimality. *Cell.* Elsevier Inc.; 2016;167(1):122–32.
- 647 34. Sun M, Schwalb B, Pirkl N, Maier KC, Schenk A, Failmezger H, et al. Global
648 analysis of Eukaryotic mRNA degradation reveals Xrn1-dependent buffering of
649 transcript levels. *Mol Cell.* 2013;52(1):52–62.
- 650 35. Gerber AP, Herschlag D, Brown PO. Extensive association of functionally and
651 cytotopically related mRNAs with Puf family RNA-binding proteins in yeast.
652 *PLoS Biol.* 2004;2(3).
- 653 36. Gupta I, Clauder-Münster S, Klaus B, Järvelin AI, Aiyar RS, Benes V, et al.
654 Alternative polyadenylation diversifies post-transcriptional regulation by
655 selective RNA-protein interactions. *Mol Syst Biol.* 2014;10:719.
- 656 37. Colomina N, Ferrezuelo F, Wang H, Aldea M, Garí E. Whi3, a developmental
657 regulator of budding yeast, binds a large set of mRNAs functionally related to
658 the endoplasmic reticulum. *J Biol Chem.* 2008;283(42):28670–9.
- 659 38. Cai Y, Futcher B. Effects of the Yeast RNA-Binding Protein Whi3 on the Half-
660 Life and Abundance of CLN3 mRNA and Other Targets. *PLoS One.*
661 2013;8(12):e84630.
- 662 39. Leeds P, Wood JM, Lee B, Culbertson MR. Gene Products That Promote
663 mRNA Turnover in *Saccharomyces cerevisiae*. *Mol Cell Biol.*
664 1992;12(5):2165–77.
- 665 40. Cui Y, Hagan KW, Zhang S, Peltz SW. Identification and characterization of

- 666 genes that are required for the accelerated degradation of mRNAs containing a
667 premature translational termination codon. *Genes Dev.* Cold Spring Harbor
668 Laboratory Press; 1995 Feb 15;9(4):423–36.
- 669 41. Gaba A, Jacobson A, Sachs MS. Ribosome occupancy of the yeast CPA1
670 upstream open reading frame termination codon modulates nonsense-
671 mediated mRNA decay. *Mol Cell.* 2005;20(3):449–60.
- 672 42. Wang X, Hou J, Quedenau C, Chen W, Angelini C, Canditiis D De, et al.
673 Pervasive isoform-specific translational regulation via alternative transcription
674 start sites in mammals. *Mol Syst Biol.* EMBO Press; 2016 Jul 18;12(7):875.
- 675 43. Kozak M. At least six nucleotides preceding the AUG initiator codon enhance
676 translation in mammalian cells. *J Mol Biol.* 1987;196(4):947–50.
- 677 44. Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, et al.
678 Deciphering the rules by which 5' -UTR sequences affect protein expression in
679 yeast. *Proc Natl Acad Sci.* 2013;110(30):E2792–801.
- 680 45. Sabi R, Tuller T. Modelling the Efficiency of Codon-tRNA Interactions Based on
681 Codon Usage Bias. *DNA Res.* 2014;21(6):511–26.
- 682 46. He F, Li X, Spatrack P, Casillo R, Dong S, Jacobson A. Genome-Wide Analysis
683 of mRNAs Regulated by the Nonsense-Mediated and 5' to 3' mRNA Decay
684 Pathways in Yeast. *Mol Cell.* 2003;12(6):1439–52.
- 685 47. Hug N, Longman D, Cáceres JF. Mechanism and regulation of the nonsense-
686 mediated decay pathway. *Nucleic Acids Res.* 2015;44(4):1483–95.
- 687 48. Brown A, Shao S, Murray J, Hegde RS, Ramakrishnan V. Structural basis for
688 stop codon recognition in eukaryotes. *Nature.* 2015;524(7566):493–6.
- 689 49. Bonetti B, Fu L, Moon J, Bedwell DM. The efficiency of translation termination
690 is determined by a synergistic interplay between upstream and downstream
691 sequences in *Saccharomyces cerevisiae*. *J Mol Biol.* 1995 Aug 18;251(3):334–
692 45.
- 693 50. Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, et al. Evidence
694 of abundant stop codon readthrough in *Drosophila* and other metazoa.
695 *Genome Res.* 2011;21(12):2096–113.
- 696 51. Williams I, Richardson J, Starkey A, Stansfield I. Genome-wide prediction of
697 stop codon readthrough during translation in the yeast *Saccharomyces*
698 *cerevisiae*. *Nucleic Acids Res.* 2004;32(22):6605–16.
- 699 52. Elemento O, Slonim N, Tavazoie S. A Universal Framework for Regulatory

- 700 Element Discovery across All Genomes and Data Types. *Mol Cell*.
701 2007;28(2):337–50.
- 702 53. Goldstrohm AC, Seay DJ, Hook BA, Wickens M. PUF protein-mediated
703 deadenylation is catalyzed by Ccr4p. *J Biol Chem*. 2007;282(1):109–14.
- 704 54. Garí E, Volpe T, Wang H, Gallego C, Fitcher B, Aldea M. Whi3 binds the
705 mRNA of the G1 cyclin CLN3 to modulate cell fate in budding yeast. *Genes*
706 *Dev*. 2001;15(21):2803–8.
- 707 55. Neymotin B, Athanasiadou R, Gresham D. Determination of in vivo RNA
708 kinetics using RATE-seq. *RNA*. 2014;20(10):1645–52.
- 709 56. Roy B, Jacobson A. The intimate relationships of mRNA decay and translation.
710 *Trends Genet*. Elsevier Ltd; 2013;29(12):691–9.
- 711 57. Huch S, Nissan T. Interrelations between translation and general mRNA
712 degradation in yeast. *Wiley Interdiscip Rev RNA*. 2014;5(6):747–63.
- 713 58. Radhakrishnan A, Green R. Connections underlying translation and mRNA
714 stability. *J Mol Biol*. Elsevier B.V.; 2016;
- 715 59. Vogel C, Abreu R de S, Ko D, Le S-YY, Shapiro BA, Burns SC, et al. Sequence
716 signatures and mRNA concentration can explain two-thirds of protein
717 abundance variation in a human cell line. *Mol Syst Biol*. Nature Publishing
718 Group; 2010;6(1):400.
- 719 60. Zur H, Tuller T. Transcript features alone enable accurate prediction and
720 understanding of gene expression in *S. cerevisiae*. *BMC Bioinformatics*.
721 2013;14(15):1.
- 722 61. Wissink EM, Fogarty EA, Grimson A. High-throughput discovery of post-
723 transcriptional cis-regulatory elements. *BMC Genomics*. *BMC Genomics*;
724 2016;17(1):177.
- 725 62. Kozak M. Point mutations define a sequence flanking the AUG initiator codon
726 that modulates translation by eukaryotic ribosomes. *Cell*. *Cell Press*;
727 1986;44(2):283–92.
- 728 63. Schwartz DC, Parker R. Mutations in translation initiation factors lead to
729 increased rates of deadenylation and decapping of mRNAs in *Saccharomyces*
730 *cerevisiae*. *Mol Cell Biol*. 1999;19(8):5247–56.
- 731 64. Pelechano V, Wei W, Steinmetz LM. Extensive transcriptional heterogeneity
732 revealed by isoform profiling. *Nature*. Nature Publishing Group;
733 2013;497(7447):127–31.

- 734 65. Zhou X, Carbonetto P, Stephens M. Polygenic Modeling with Bayesian Sparse
735 Linear Mixed Models. *PLoS Genet.* 2013;9(2):e1003264.
- 736 66. Lorenz R, Bernhart SH, Höner zu Siederdisen C, Tafer H, Flamm C, Stadler
737 PF, et al. ViennaRNA Package 2.0. *Algorithms Mol Biol. BioMed Central*;
738 2011;6(1):26.
- 739 67. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al.
740 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast
741 genomes. *Genome Res.* 2005;15(8):1034–50.
- 742 68. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear
743 Models via Coordinate Descent. *J Stat Softw.* 2010;33(1):1–22.

744

745 **Figure captions**

746 **Fig 1. Study overview.** The goal of this study is to discover and integrate cis-
747 regulatory mRNA elements affecting mRNA stability and assess their dependence on
748 mRNA degradation pathways. **Data)** We obtained *S. cerevisiae* genome-wide half-life
749 data from wild-type (WT) as well as from 34 knockout strains from Sun et al. 2013.
750 Each of the knockout strains has one gene closely related to mRNA degradation
751 pathways knocked out. **Analysis)** We systematically searched for novel sequence
752 features associating with half-life from 5'UTR, start codon context, CDS, stop codon
753 context, and 3'UTR. Effects of previously reported cis-regulatory elements were also
754 assessed. Moreover, we assessed the dependencies of different sequence features
755 on degradation pathways by analyzing their effects in the knockout strains.

756 **Integrative model)** We build a statistical model to predict genome-wide half-life
757 solely from mRNA sequence. This allowed the quantification of the relative
758 contributions of the sequence features to the overall variation across genes and
759 assessing the sensitivity of mRNA stability with respect to single-nucleotide variants.

760

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

767

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

784

785 **Fig 4. Translation initiation, elongation and termination features associate with**
786 **mRNA half-life. (A)** Distribution of half-life for mRNAs grouped by the third
787 nucleotide before the start codon. Group sizes (numbers in boxes) show that
788 nucleotide frequency at this position positively associates with half-life. **(B)** mRNA
789 half-life (y-axis) versus species-specific tRNA adaptation index (sTAI) (x-axis) for *S.*
790 *cerevisiae*. **(C)** mRNA half-life explained variance (y-axis, Materials and Methods) in
791 wild-type (WT) and across all 34 knockout strains (grouped according to their
792 functions). Each blue dot represents one replicate, bar heights indicate means across
793 replicates. Bars with a red star are significantly different from the wild type level (FDR
794 < 0.1, Wilcoxon rank-sum test, followed by Benjamini-Hochberg correction). **(D)**
795 Distribution of half-life for mRNAs grouped by the stop codon and the following
796 nucleotide. Colors represent three different stop codons (TAA, TAG and TGA), within
797 each stop codon group, boxes are shown in G, A, T, C order of their following base.
798 Only the P-values for the most drastic pairwise comparisons (A versus C within each
799 stop codon group) are shown. All p-values in boxplots were calculated with Wilcoxon
800 rank-sum test. Boxplots computed as in Fig 3.

801

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

817

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

834

835 **Supporting Information**

836 **S1 Fig. Length of 5'UTR, CDS and 3'UTR correlate with mRNA half-life. (A-B)**

837 5'UTR length (x-axis) versus half-life (y-axis) for *S. cerevisiae* (A) and *S. pombe* (B).

838 **(C-D)** CDS length (x-axis) versus half-life (y-axis) for *S. cerevisiae* (C) and *S. pombe*

839 (D). **(E-F)** 3'UTR length (x-axis) versus half-life (y-axis) for *S. cerevisiae* (E) and *S.*

840 *pombe* (F).

841

842 **S2 Fig. GC content of 5'UTR, CDS and 3'UTR correlate with mRNA half-life. (A-**

843 **B)** 5'UTR GC content (x-axis) versus half-life (y-axis) for *S. cerevisiae* (A) and *S.*

844 *pombe*. **(C-D)** CDS GC content (x-axis) versus half-life (y-axis) for *S. cerevisiae* (C)

845 and *S. pombe* (D). **(E-F)** 3'UTR GC content (x-axis) versus half-life (y-axis) for *S.*

846 *cerevisiae* (E) and *S. pombe* (F).

847

848 **S3 Fig. Folding energy of 5'UTR, CDS and 3'UTR correlate with mRNA half-life.**

849 **(A-B)** 5' free energy (x-axis) versus half-life (y-axis) for *S. cerevisiae* (A) and *S.*

850 *pombe* (B). **(C-D)** CDS free energy (x-axis) versus half-life (y-axis) for *S. cerevisiae*

851 (C) and *S. pombe* (D). **(E-F)** 3' free energy (x-axis) versus half-life (y-axis) for *S.*

852 *cerevisiae* (E) and *S. pombe* (F).

853

854 **S4 Fig. Translation initiation, elongation and termination features associate**

855 **with mRNA half-life. (A)** Start codon context (Kozak sequence) generated from

856 4388 *S. cerevisiae* genes and 3713 *S. pombe* genes. **(B)** Distribution of half-life for

857 mRNAs grouped by the third nucleotide before the start codon for *S. pombe*. Group

858 sizes (numbers in boxes) show that nucleotide frequency at this position positively

859 associates with half-life. **(C)** mRNA half-life (y-axis) versus species-specific tRNA

860 adaptation index (sTAI) (x-axis) for *S. pombe*. **(D)** Distribution of half-life for mRNAs

861 grouped by the stop codon and the following nucleotide for *S. pombe*. Colors

862 represent three different stop codons (TAA, TAG and TGA), within each stop codon

863 group, boxes are shown in G, A, T, C order of their following base. Only the P-values

864 for the most drastic pairwise comparisons (A versus C within each stop codon group)

865 are shown. **(E)** Distribution of half-life for mRNAs grouped by with or without

866 additional 3'UTR in-frame stop codon for *S. cerevisiae*. 30 bases window after the

867 main stop codon was considered. **(F)** Same as (E) for *S. pombe*. All p-values in
868 boxplot were calculated with Wilcoxon rank-sum test. Boxplots computed as in Fig 3.

869

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

881

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

896

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

900

901 **S8 Fig. Predicted effects of synonymous codon transitions on half-life.**

902 Expected half-life fold-change (x-axis) at each synonymous codon transitions. Each
903 row represent transition from one codon (y-axis) to its synonymous partners. Only
904 synonymous codons that differ by one base were considered.

905

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

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

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

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

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

914

915

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)

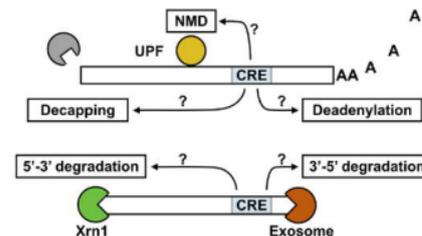
Analysis

Cis-regulatory elements



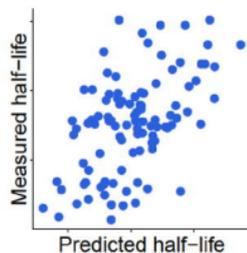
- Novel elements discovery
- Integrate novel and known elements
- Single-nucleotide effect estimation

Dependence on degradation pathways

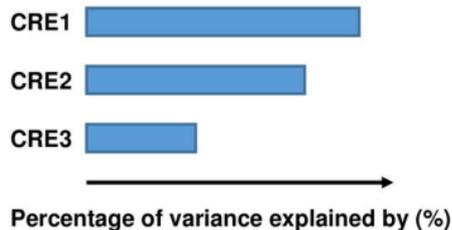


Integrative model

Out-of-sample predictions

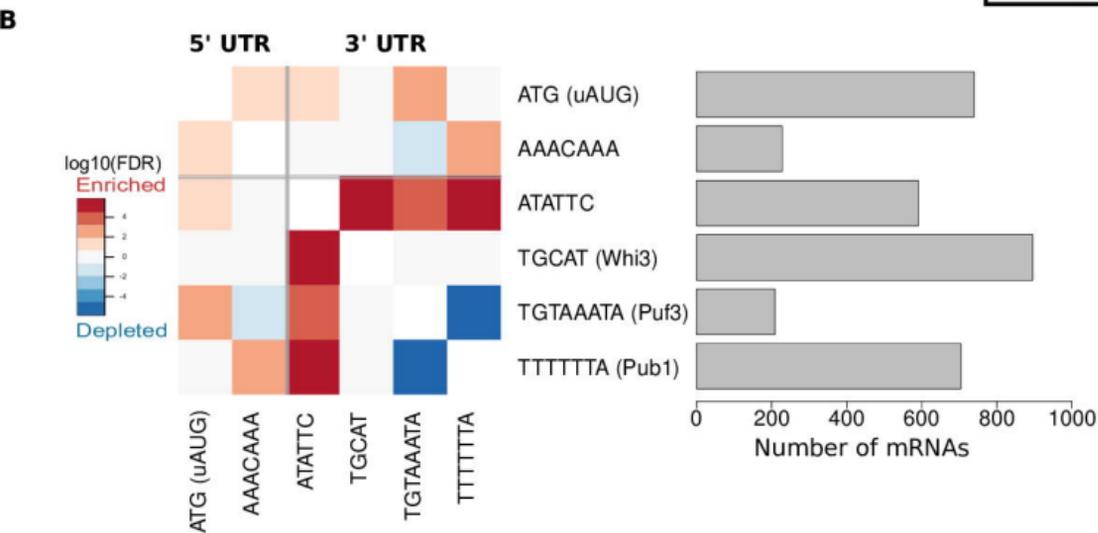
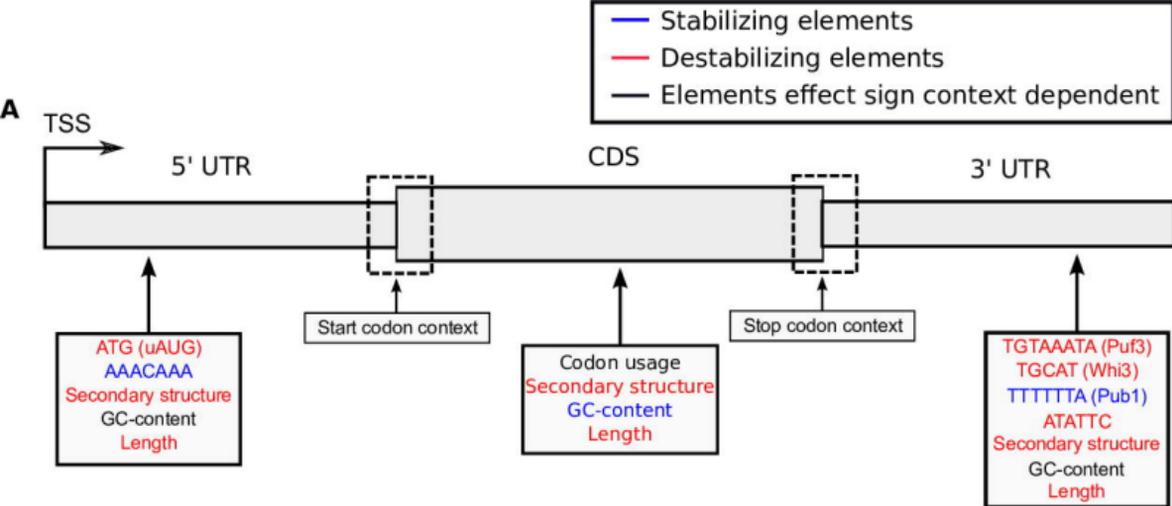


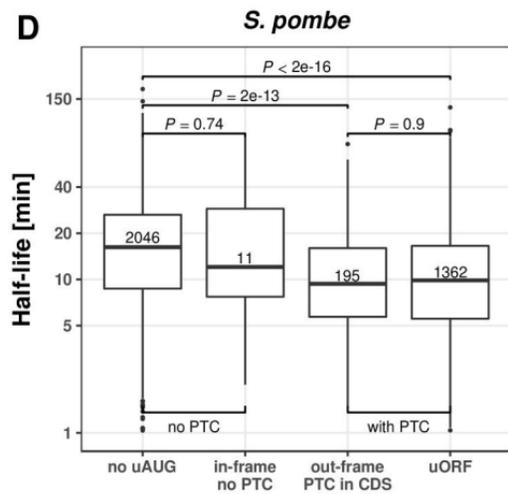
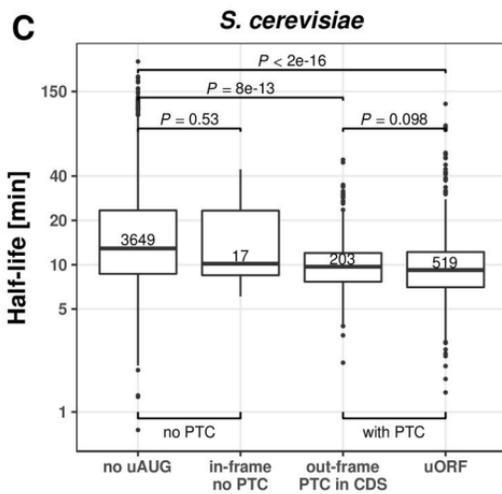
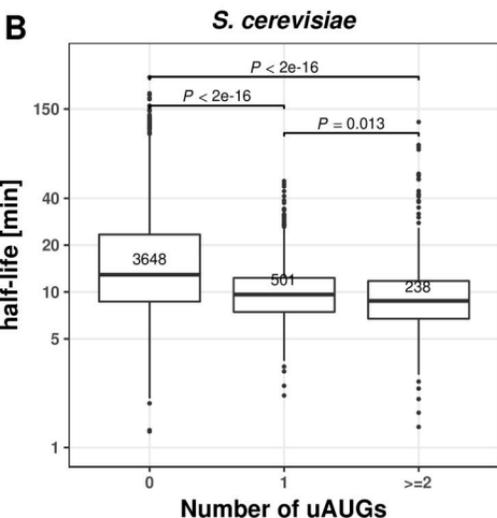
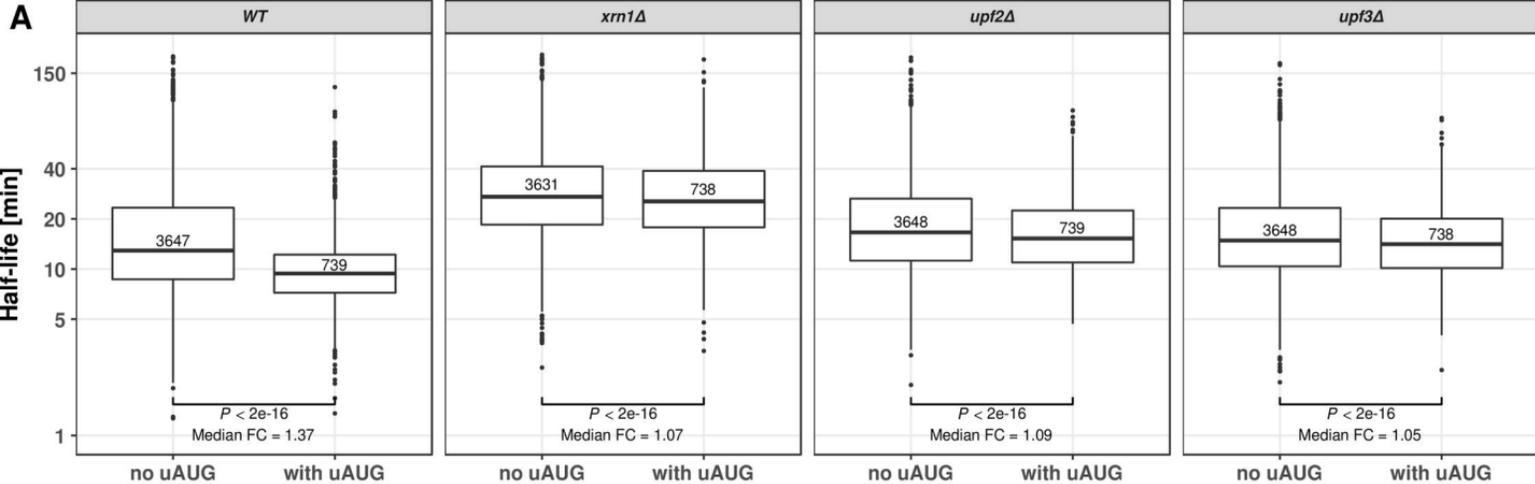
Relative contributions

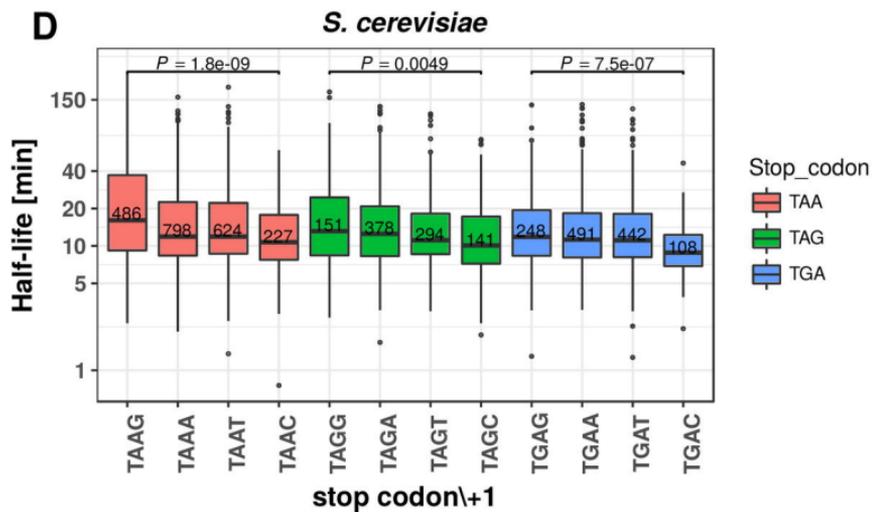
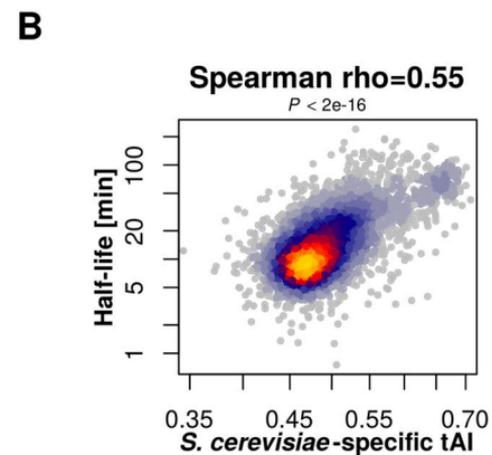
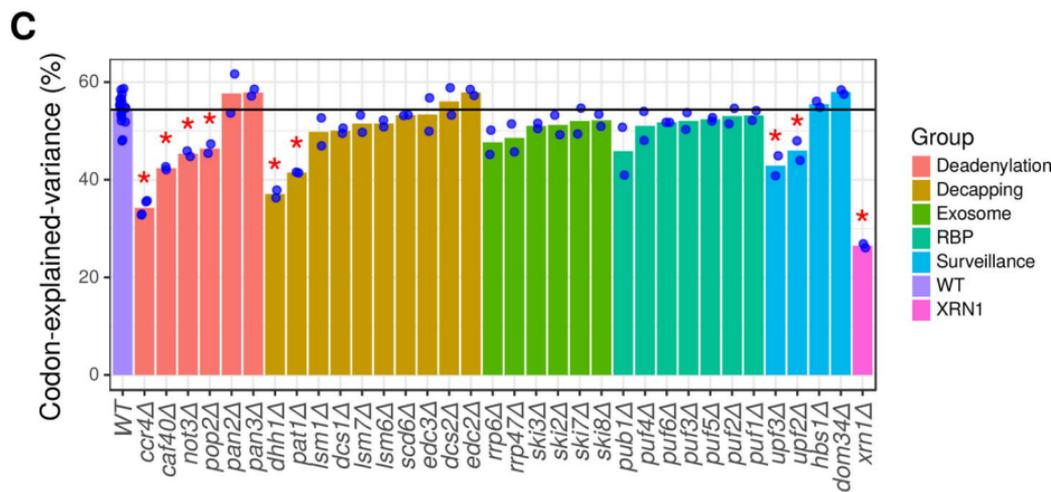
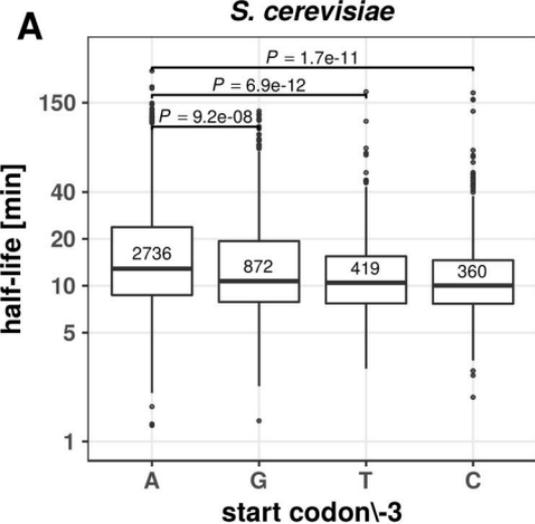


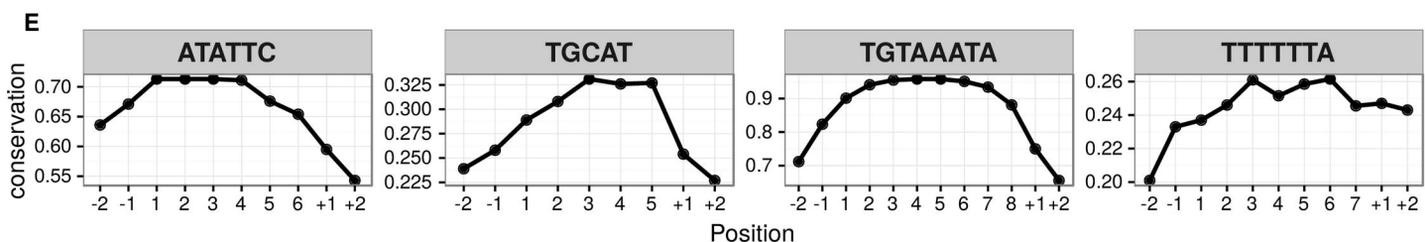
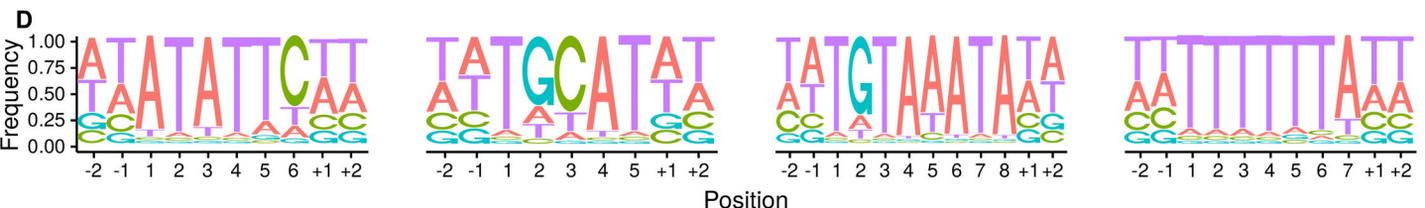
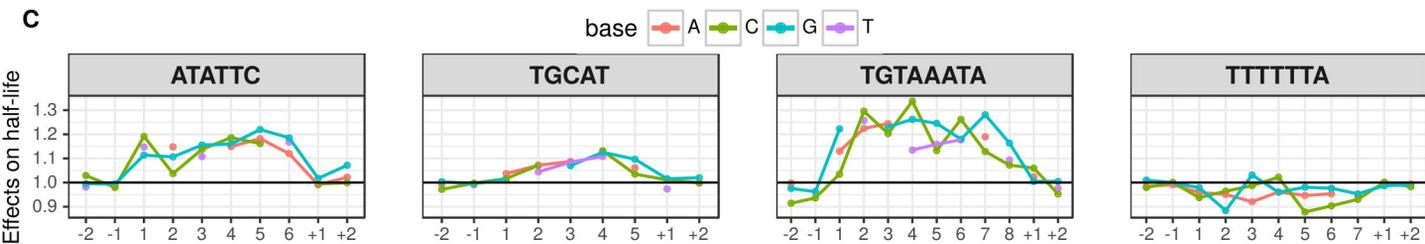
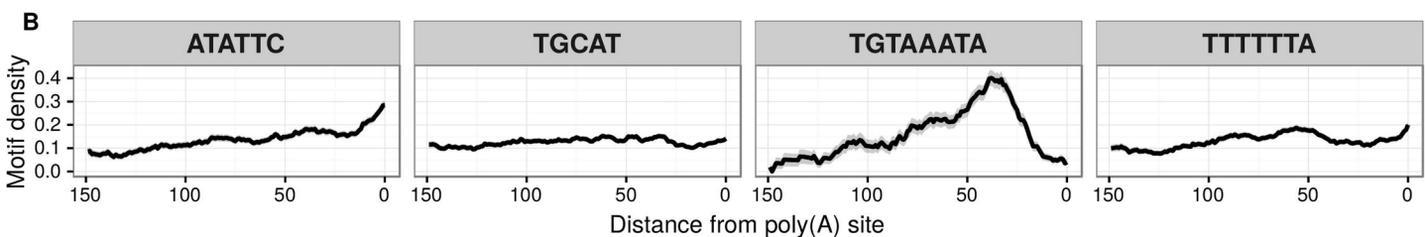
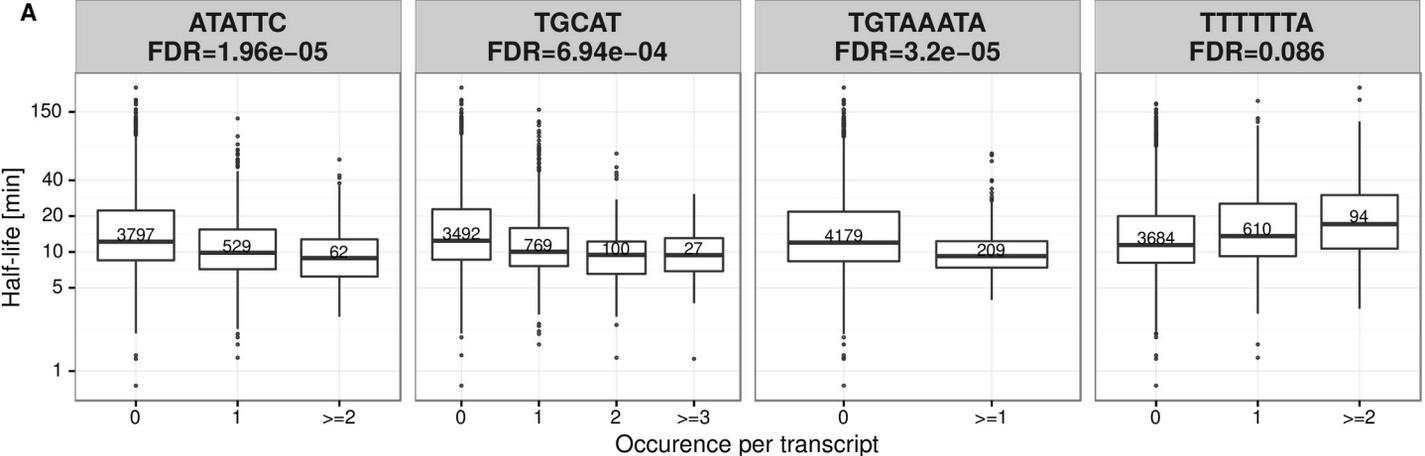
Single-nucleotide effects

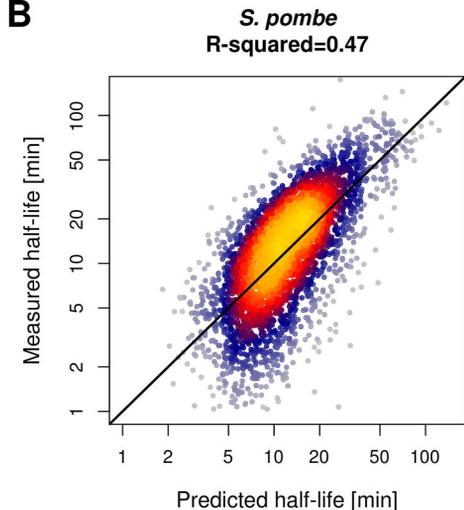
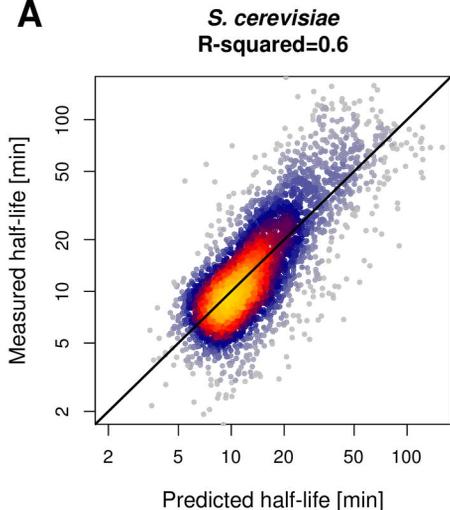










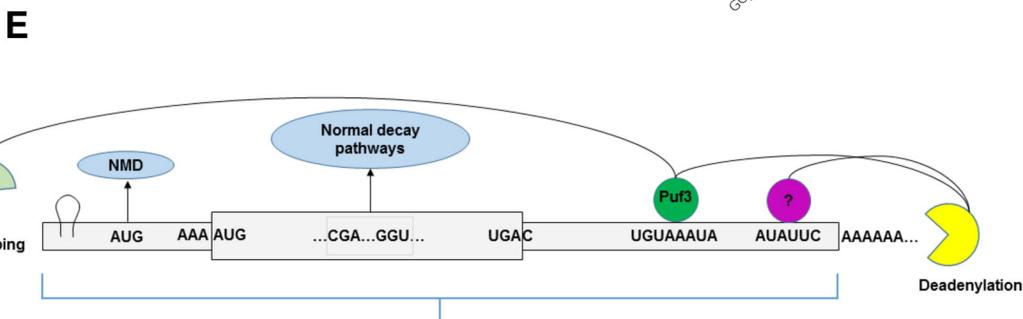
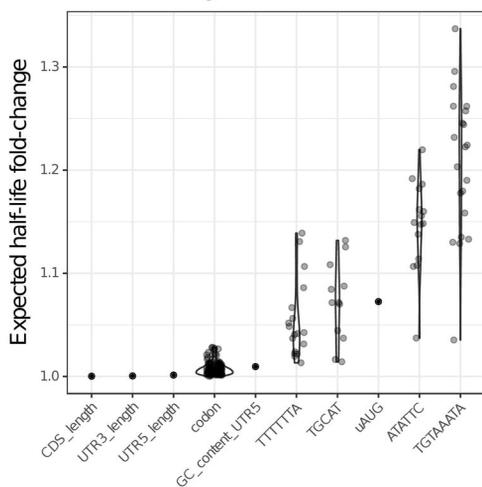


C

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

D **Effect of single-nucleotide variations**

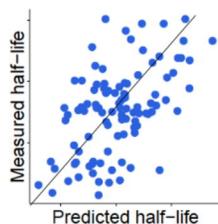


The major determinant:

60% variance explained

Single-nucleotide variation effects

Codons



3'UTR motifs and uAUG High

Synonymous codons Middle

Others Low