Li et al. Genome Biology (2019) 20:162
https://doi.org/10.1186/s13059-019-1761-9

Genome Biology

RESEARCH Open Access

Quantitative principles of *cis*-translational control by general mRNA sequence features in eukaryotes

Jingyi Jessica Li[1*], Guo-Liang Chew[2] and Mark Douglas Biggin[3*]

Abstract

Background: General translational *cis*-elements are present in the mRNAs of all genes and affect the recruitment, assembly, and progress of preinitiation complexes and the ribosome under many physiological states. These elements include mRNA folding, upstream open reading frames, specific nucleotides flanking the initiating AUG codon, protein coding sequence length, and codon usage. The quantitative contributions of these sequence features and how and why they coordinate to control translation rates are not well understood.

Results: Here, we show that these sequence features specify 42–81% of the variance in translation rates in *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, *Mus musculus*, and *Homo sapiens*. We establish that control by RNA secondary structure is chiefly mediated by highly folded 25–60 nucleotide segments within mRNA 5′ regions, that changes in tri-nucleotide frequencies between highly and poorly translated 5′ regions are correlated between all species, and that control by distinct biochemical processes is extensively correlated as is regulation by a single process acting in different parts of the same mRNA.

Conclusions: Our work shows that general features control a much larger fraction of the variance in translation rates than previously realized. We provide a more detailed and accurate understanding of the aspects of RNA structure that directs translation in diverse eukaryotes. In addition, we note that the strongly correlated regulation between and within *cis*-control features will cause more even densities of translational complexes along each mRNA and therefore more efficient use of the translation machinery by the cell.

Keywords: Translation , *Cis*-control , RNA structure , Codon usage , Collinear regulation

Background

It is a major challenge to determine from nucleotide sequence data the rates at which eukaryotic mRNAs are translated into protein. There are two classes of *cis*-acting elements in mRNAs that determine these rates: general sequence features and gene/condition-specific elements [1–11]. The general features are secondary structure in the 5′ portion of the mRNA; upstream open reading frames (uORFs), which lie 5′ of the protein coding sequence (CDS); specific nucleotides immediately flanking the initiating AUG codon (iAUG) at the 5′ of the CDS; CDS length; and codon usage. These five features are each present in all or many mRNAs and function by affecting the formation and progress of preinitiation complexes or the ribosome. As such, they act in a wide array of physiological states and tissues. Gene or condition-specific elements, in contrast, are short sequences recognized by *trans*-acting factors such as microRNAs or regulatory proteins. A given type of specific element is present in only a subset of genes and its cognate *trans*-acting factor only functional in a subset of cells or conditions.

Here, we focus on the general elements. Our major goal is to estimate the contributions of these sequence features—and their associated biochemical mechanisms—to the total variance in translation rates, singly and in combination. We argue that from this, one can, in addition, approximate from the unexplained variance the

* Correspondence: jli@stat.ucla.edu; mdbiggin@lbl.gov
[1]Department of Statistics, Department of Biomathematics, and Department of Human Genetics, University of California, Los Angeles, CA 90095, USA
[3]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94708, USA
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

contribution of specific elements, or at least set an upper limit on that contribution. By optimizing the models that predict the contribution of each general feature, we also aim to better understand the nucleic acid sequences that encode them. Further, to establish common principles, we have analyzed in parallel data from five model eukaryotes: two yeasts, a plant, and two mammals.

The general sequence features have long been known to affect the translation rates of individual genes. The quantitative contributions of these elements to the variance in genome-wide translation rates, however, have not been well characterized. The correlation of subsets of these sequence features with ribosome profiling translation rate data has been estimated in *Saccharomyces cerevisiae* and several animal models [4, 10, 12–16] (Additional file 1: Table S1). In addition, the effect of large numbers of artificial sequence mutations on translation has been determined for 10–50 nucleotide segments proximal to the iAUG codon in *S. cerevisiae* and *Homo sapiens* [12, 17–21] (Additional file 1: Table S2). The role of RNA secondary structure in 5′ untranslated regions (5′UTRs) has also been addressed in vitro for *S. cerevisae* [22] (Additional file 1: Table S2). A prior study that we conducted was the only one to assess the contributions of all five features [10]. This work established that mRNA folding, uORFs, iAUG proximal sequence elements (APEs), and CDS length explain 58% of the variance in translation rates among *S. cerevisiae* genes, presumably due to their impact on the initiation rate. When the effect of codon usage on elongation by the ribosome is also taken into account, 80% of the variance can be explained.

Here, we have developed new models that characterize in detail the mRNA secondary structures and sequence motifs that regulate translation rates. By applying our models to five well-studied eukaryotes and by examining the correlations in control between general features and between different parts of the same mRNA, we identify the key principles common to all five species as well as phylum-specific differences.

Results

Datasets

The five model eukaryotes we chose to study are: *S. cerevisiae*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, *Mus musculus*, and *H. sapiens*. Translation rates were determined from ribosome profiling data as prior work has shown that the density of ribosomes per mRNA molecule (i.e., the translational efficiency) is a useful estimate of the rate [13, 23]. For each species, we chose an example dataset to analyze in most detail [4, 13, 24–26]. In addition, key analyses were repeated on further datasets of three species for other tissues or biological replicas [4, 14, 27, 28]. Additional file 1: Figure S1 presents the distributions of translation rates, 5′UTR length, and CDS lengths for the five example datasets. Additional file 1: Figure S2 shows the variation in translation rates for all ten datasets examined. The lengths of 5′UTRs vary among the species, with the average for *S. cerevisiae* being the shortest and that for the two mammals and *S. pombe* the longest. By contrast, the ranges of CDS lengths are more consistent among the five eukaryotes. The variance in translation rates is relatively narrow for all species and conditions. The mRNA sequences, translation rates, and other primary information for each dataset are given in Additional file 2. Table 1 provides a glossary of abbreviations used.

Control by 5′ mRNA secondary structures

Biophysical measurements of short RNA oligonucleotides in vitro have allowed the melting temperatures and Gibbs free energies of stem/loop structures to be accurately calculated for any short sequence based on a set of rules for the energies of base pairing and adjacent base stacking as well as the destabilizing entropic effect of

Table 1 Glossary of abbreviations

Abbreviation	Explanation
APE	AUG proximal element: control sequences that flank the iAUG
dAPE	Downstream APE: that part of the APE located downstream (3′) of the iAUG
uAPE	Upstream APE: that part of the APE located upstream (5′) of the iAUG
5′ofAPE	The 5′UTR sequences that lie 5′ of the APE
iAUG	Initiating AUG: the AUG codon at the 5′ end of the CDS
uAUG	Upstream AUG: an AUG codon in the 5′ UTR, i.e., an AUG at the 5′ end of an uORF
BIC	Bayesian Information Criteria: a method to select an efficient set of model features
5′ cap	The 5′ most nucleotide of the mRNA
CDS	Coding sequence: the protein coding portion of the mRNA
CDS 3′ofAPE	That part of the CDS that lies 3′ of the APE
MFE	Minimum free energy: the Gibbs free energy of folding for a defined segment of RNA
uORF	Upstream open reading frame: an uAUG containing open reading frame in the 5′UTR
PWM	Position Weight Matrix: the fraction of A, U, G, and Cs at each nucleotide position
5′region	The region spanning the 5′UTR and the 5′ most ~ 30 nucleotides of the CDS
5′UTR	5′ Untranslated region: all mRNA nucleotides 5′ of the CDS
TR	Translation rate: rate of translation per mRNA molecule at steady state
high TR	A cohort of genes with the highest translation rates
low TR	A cohort of genes with the lowest translation rates

loop length [29–32]. For example, the coefficient of the determination correlation coefficient between measured and predicted Gibbs energies for 109 4–14-bp double-stranded RNA oligonucleotides from [33] is $R^2 = 0.98$ (Additional file 3). Computational predictions of mRNA secondary structure that incorporates these and other information are quite effective: the algorithm we employed having 74% recall (percent of true-positive base pairs identified) and 79% precision (percent of base pairs identified that are true positives) [32]. Previous work has shown that the free energies of RNA folding from this or similar algorithms positively correlate with the translation rates (Additional file 1: Tables S1 and S2), supporting the view that the preinitiation complex and ribosome must unfold RNA structures in order to progress down the mRNA. Our preliminary analyses, however, showed that these prior studies had greatly underestimated the contribution of mRNA folding to translation. We therefore systematically assessed which aspects of secondary structure are most important in determining translational efficiency and used this information to improve the predictive models of rates.

Figure 1 shows the mean and the 25th percentile-75th percentile interval for the free folding energies of 35 nucleotide windows for two cohorts of genes, the 10% of genes with the highest translation rates (high TR) and the 10% with the lowest (low TR). Within the 5′UTR, the low TR cohort has smaller folding energies (i.e., are more folded) than the high TR cohort, consistent with the repression of preinitiation complexes by RNA structure. For *S. cerevisiae*, *S. pombe*, and *Arabidopsis*, the differences between the two cohorts are largest from −35 to +35, −35 to +1, and −120 to +35, respectively. For mouse and human, the differences are largest towards the 5′ cap. In the CDS region 3′ of +30, the relationship of RNA folding energy and translation rate is not consistent, with either a strong negative correlation in *S. pombe*; a weak negative correlation in *S. cerevisiae*, *Arabidopsis*, and *M. musculus*; or a positive correlation in *H. sapiens*. Prior analysis in *S. cerevisiae* implied that strongly folded RNA structures in the CDS are associated with more stable mRNAs and that the negative correlation of CDS folding energy with translation is indirect, being due to the positive correlation between translation and mRNA abundance and the negative effect of RNA turnover on mRNA abundance [10]. Given this and a lack of evidence that RNA structure affects the elongating ribosome, we have limited our models to the 5′UTR and the 5′ most part of CDS, where folding energy values and translation rates correlate positively. We term the combination of the 5′UTR and the short 5′ CDS segment the 5′ region.

Along the 5′ regions of individual mRNAs, free energy values vary dramatically (Additional file 1: Figure S3), reflecting mRNA stem/loop structures at some locations and unfolded regions at others. Despite the mean tendencies shown in Fig. 1, Additional file 1: Figure S3 reveals that individual genes can have folded or unfolded regions at almost any location.

To capture and exploit these complex distributions, we devised a number of features that each score every gene using some aspect of predicted RNA folding energy. We also constructed three multivariate linear models that combine multiple features to form feature sets. The coefficient of determination correlation coefficient (R^2) was then calculated between each feature or feature set and the translation rates (Fig. 2; Additional file 4). One feature was defined as the folding energy for the contiguous sequence from the 5′ cap to +35 (whole). The other features were determined for each of 21 window lengths varying from 6 to 100 nucleotides, the 5′ ends of the windows tiling from the 5′ cap to −1 (Fig. 2). The result is a set of 421 metrics for translational control by RNA structure for each species.

Our most accurate prediction of translation derived from a multivariate linear model that was selected by Bayesian Information Criteria (BIC) and that contains 9–33 features of whichever window length(s) provided the most useful information ("RNAfold," Fig. 2; Additional file 5). These "RNAfold" feature sets explain between 11 and 33% of the variance in rates and as such have over twice the predictive power (R^2) of models used by other groups (Additional file 1: Table S1).

Most features employed windows identified based on their free energy values. A substantial proportion of control by RNA secondary structure—often greater than a half—can be explained by the single window that has the minimum free energy within each mRNA (i.e., by the window that has the most folded structure) ("min," Figs. 2 and 3a). Less folded windows each have progressively reduced explanatory power compared to the "min" window ("10%," "25%," "75%," "90%," and "max," Fig. 3a). Likewise, features that sum Gibbs energies for the most folded window with the energies of other strongly folded segments ("sum ≤ 5%," "sum ≤ 10%," or "sum ≤ 20%") have more explanatory power than features that sum the energies of less well-folded segments ("sum > 90%" and "sum > 80%") (Fig. 2). A similar relationship is also seen for features that determine the percent of windows in each mRNA that pass some threshold on the energies of all windows in the dataset ("% ≤ 20%" vs "% > 80%" and "% > 90%") (Fig. 2). These results collectively indicate that translation rates are much more strongly influenced by the differences among the more folded segment(s) within the 5′ regions than by the differences among their less folded sequences.

Additional results imply that base pairing between sequences separated by more than 60 nucleotides does not

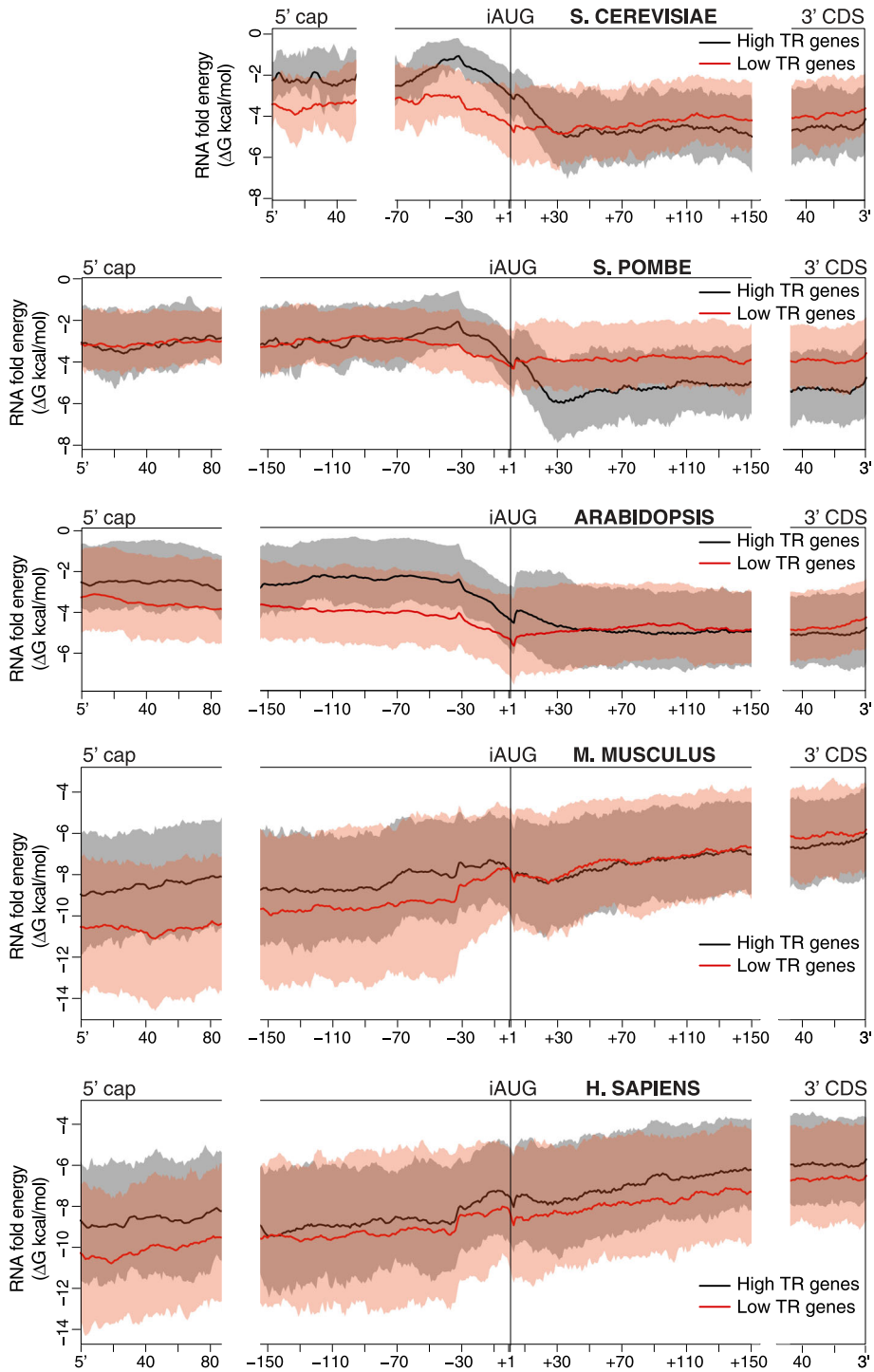


Fig. 1 Different distributions of secondary structures in the 5′ portions of mRNAs. The predicted RNA folding energy (ΔG kcal/mol) of 35 nucleotide windows (*y*-axis) are plotted for the 10% most highly translated mRNAs (high TR, black) and the 10% most poorly translated (low TR, red). The *x*-axis shows the position of the 5′ most nucleotide of each window. Windows for every one nucleotide offset were calculated. At each location, the free energies for the mean (continuous lines) and the interval between the 25th and 75th percentile (shading) are shown. mRNAs were aligned at their 5′ cap (left), at the iAUG (center), or at the 3′ of the CDS (right). In *S. cerevisiae*, *S. pombe*, and *Arabidopsis*, the differences between the two cohorts are greatest proximal to the iAUG. In the two mammals, the differences are greatest towards the 5′ cap

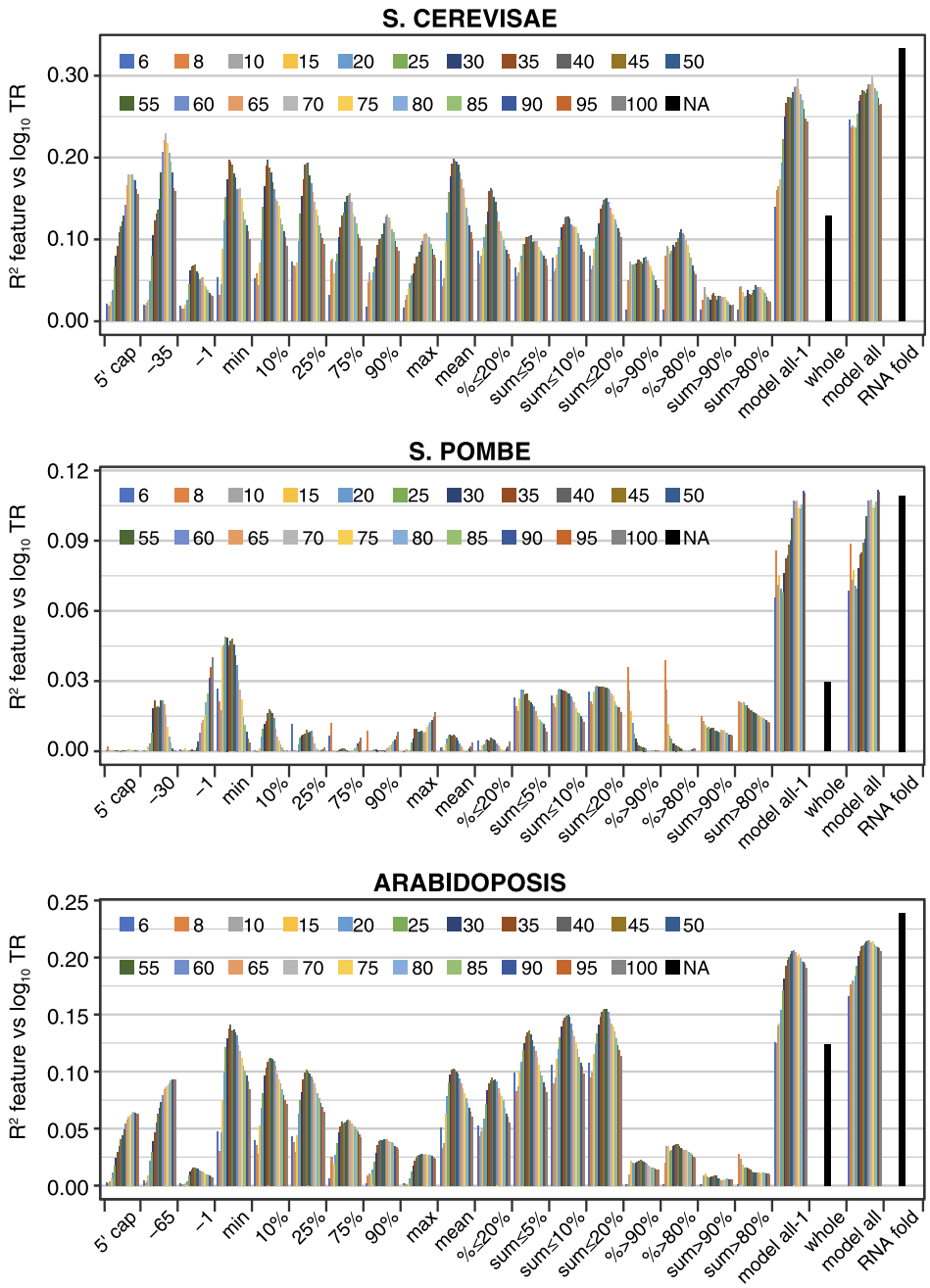


Fig. 2 Translation rates are optimally determined by a combination of RNA folding features. The R^2 coefficients of determination between $\log_{10}$ translation rates (TR) and a variety of single features and feature sets based on RNA folding energies. Nine features use the free energy value of a single window per gene (left): windows "5′ cap" through "− 1" were identified by the location of their 5′ most nucleotide, while windows "min" through "max" were identified by the rank of their free energy. Nine additional features were each calculated using multiple windows per gene (center): "mean," the mean of all windows in a gene; "% ≤ 20%," "% ≥ =90%," and "% ≥ 80%," the percent of windows in a given gene that respect a threshold on the free energy of all windows for all genes; and "sum ≤ 5%," "sum ≤ 10%," "sum ≤ 20%," "sum ≥ 80%," and "sum ≥ 90%," the sum of free energies of windows in a given gene that respect a threshold on the free energy of all windows for all genes. Twenty one window lengths from 6 to 100 nucleotides were employed for each of the above features. In addition, a final feature, "whole" (right), was calculated from the free energy of folding of the contiguous sequence from the 5′cap to + 35 for each mRNA. Three multivariate feature sets were also determined (right): "model all-1" combined all window-based features for a given window length; "model all" included in addition the "whole" feature; while "RNAfold" used Bayesian Information Criteria to select features from the complete set of features, using whichever window length variant(s) provided the most useful information. "RNAfold" was additionally constrained for *S. pombe* by removal of all windows that extend 3′ of + 30 nucleotides to avoid sequences showing a strong negative correlation of free energy values and TR. Additional file 4 provides the R^2 values for all features. Additional file 5 provides details of the features selected for the "RNAfold" model

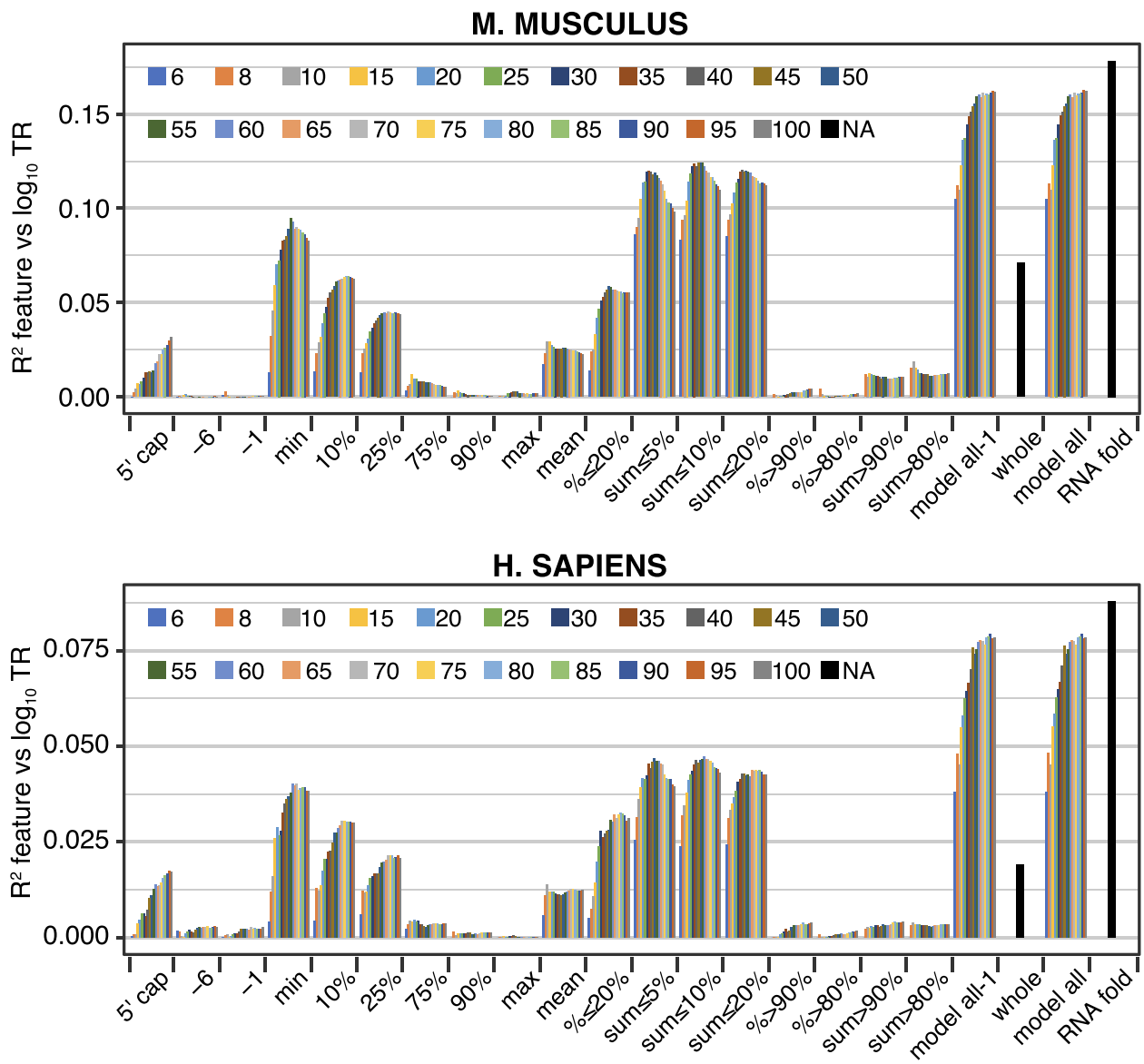


Fig. 2 (Continued)

play a dominant role in regulating translation. Because the "whole" feature includes the full 5′ region, it accounts for all long-range interactions. Models that contain all features, however, perform little better than the ones lacking "whole" (Fig. 2, compare "models all" to "models all-1").

To further define the secondary structures that control translation, we focused on the most folded window as these are strongly predictive in all species. The optimum length of "min" window ranges from 25 to 35 nucleotides in the two yeasts to 55–60 in the two mammals (Fig. 2). Windows shorter than the optimum are less predictive presumably because they exclude important secondary structures, whereas windows longer than the optimum include unfolded sequences that "dilute" the control information. The optimum length "min" windows generally contain either one or two stem/loops of varying sizes, some of which include mismatched or single nucleotide bulges within their stems (Additional file 6). For all species, the longest contiguous run of paired nucleotides within a stem contains much of the regulatory information (Fig. 3b). With the exception of *S. pombe*, however, the total number of nucleotide pairs in the window is more predictive of translation than the number of nucleotide pairs in the longest contiguous run or in the longest stem (Fig. 3b). This suggests that most nucleotide pairs in "min" windows are controlling, including those within smaller secondary

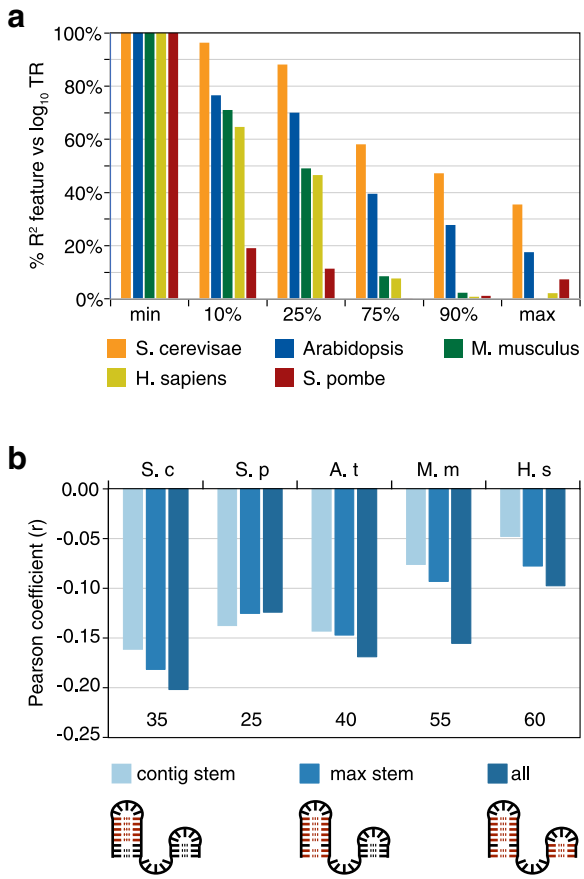


Fig. 3 The RNA structures that control translation rates. **a** The most folded window controls translation more strongly than less folded windows. The correlation between $\log_{10}$ translation rates (TR) and single windows within mRNA 5′ regions is shown. Six windows were selected based on the rank of their free energy: "min," "10%," "25%," "75%," "90%," and "max," where "min" is the most folded (smallest free energy) and "max" is the least folded. The values plotted are the R^2 coefficients of determination expressed as a percent of the R^2 coefficient for the "min" window. The window length was the optimum for each species: *S. cerevisiae*, 35; *Arabidopsis*, 40; and *M. musculus*, 55; *H. sapiens*, 60; and *S. pombe*, 25. **b** Number of nucleotide pairs describe the regulatory potential of the most folded window. The number of paired nucleotides was calculated in the most folded (min) window of each gene for the longest contiguous stem without mismatches or single nucleotide bulges (contig stem, light blue), the longest stem (max stem, mid blue), and for all pairs within the window, including those not in the longest stem (all, dark blue). The Pearson correlation coefficients between each of these three measures and $\log_{10}$ translation rates are plotted on the *y*-axis for the five example datasets. The lengths of the most folded widows are also given. For all but *S. pombe*, the correlation coefficients are more negative for measures that include more nucleotide pairs. The primary data are provided in Additional file 6

stem/loops. The different result for *S. pombe* may well be because its "min" windows have shorter optimum lengths than those of other species (25 vs 35–60 nucleotides) and tend to have only one major stem/loop (Additional file 6). Finally, It is striking that between cohorts of highly translated and lowly translated genes, while free energies and loop sizes vary significantly, the number of nucleotide pairs in contiguous runs, longest stems, or full windows is similar for all five species (Additional file 1: Figures S4 and S5).

We also defined a location-specific group of features that use the fold energy of a single window whose 5′ end is at a defined location: either at the 5′ of the mRNA (5′cap) or at a location relative to the iAUG ("-65" to "-1") (Fig. 2). These features give results consistent with the distributions of RNA folding energies shown in Fig. 1. For example, in *M. musculus* and *H. sapiens* features using windows located at the 5′ cap have more predictive power than those at or near the iAUG. While most of the features selected by BIC for the "RNAfold" feature set used windows identified by their free energies, several others belonged to the location-specific group (Additional file 5). By combining these two feature classes, we are able to more fully capture how mRNA secondary structure directs translation.

Sequence motifs in 5′ regions

One feature we used previously to explain translation rates in *S. cerevisiae* exploited sequence differences between highly and poorly translated mRNAs without regard to the biochemical mechanism(s) of control [10]. This approach proved powerful, showing that nucleotides flanking the iAUG from − 35 to + 28—a region termed the AUG proximal element (APE)—explain a third of the variance in rates [10]. To extend this strategy to the other four species, we first compared position weight matrices (PWMs) for sequences − 80 to + 35 for the most highly translated (high TR) and most poorly translated (low TR) 10% of genes (Fig. 4a; Additional file 7). The *S. pombe* and *Arabidopsis* PWMs resemble those of *S. cerevisiae*: for example, A nucleotides are enriched 5′ of the iAUG in high TR genes vs low TR genes, while Gs are depleted. *M. musculus* and *H. sapiens* mRNAs, by contrast, are GC rich, and sequence differences between high and low TR cohorts are less readily apparent. Following the strategy used earlier for *S. cerevisiae* (see the "Materials and methods" section), we assigned a score to each gene based on PWMs of varying lengths in high TR genes and defined APE boundaries by maximizing the R^2 between the PWM scores and translation rates (Fig. 4b; Additional file 1: Figure S6). *Arabidopsis* has, like *S. cerevisiae*, an extended APE, spanning nucleotides − 65 to + 33, whereas *S. pombe* and the two mammals have shorter APEs that span − 6 to + 13 or less and in which the Kozak consensus [34] plays a major role.

To increase these PWM-based models' predictive power, we used BIC to select subsets of di- and tri-nucleotides whose frequencies in each mRNA most

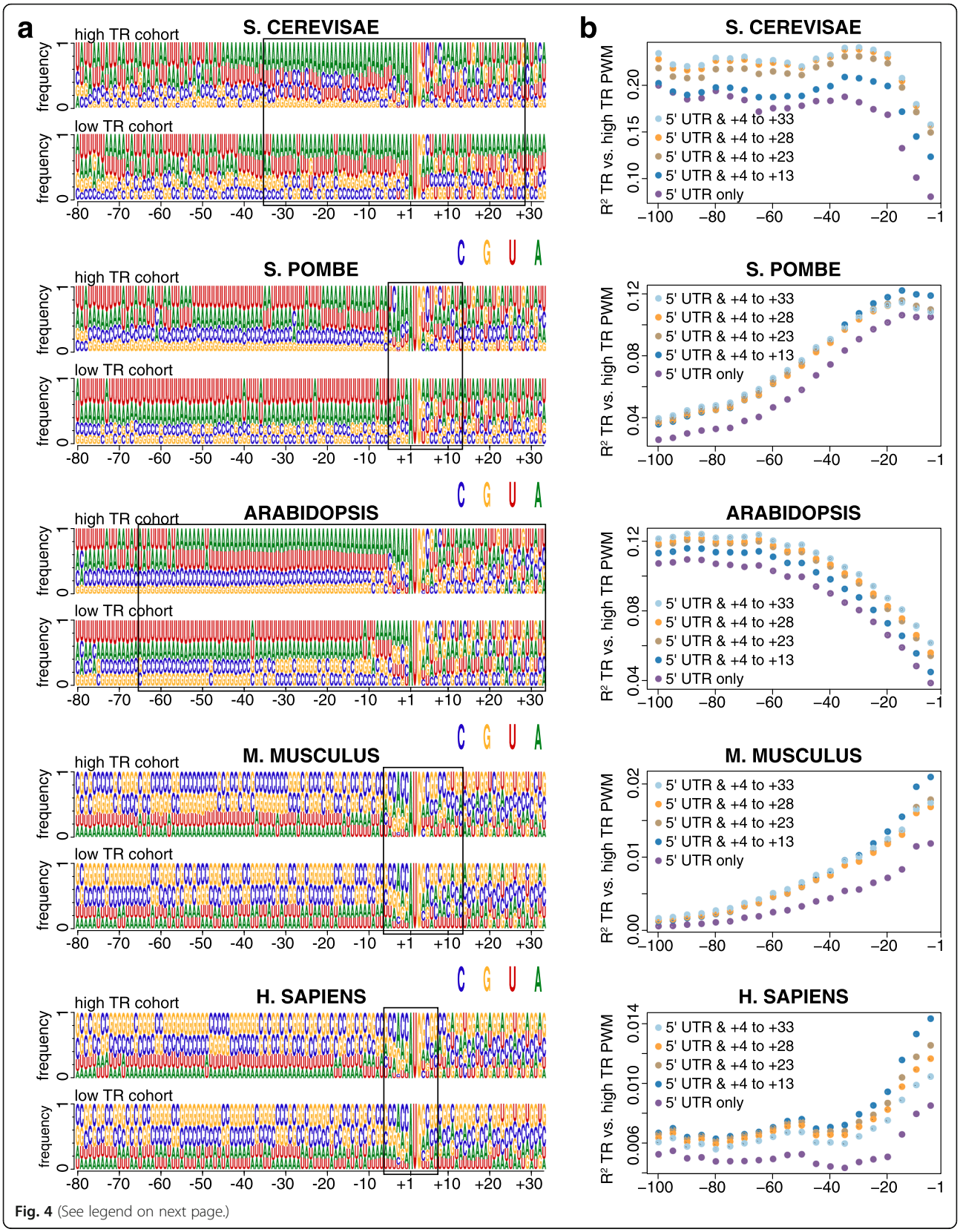


Fig. 4 (See legend on next page.)

(See figure on previous page.)

Fig. 4 The 5′ and 3′ boundaries of AUG proximal elements. **a** Position Weight Matrices (PWMs) for the 10% of mRNAs with the highest translation rate (high TR cohort) and the 10% with the lowest rate (low TR cohort). Sequence logos show the frequency of each nucleotide at each position relative to the first nucleotide of the iAUG. The location of the AUG proximal element (APE) is indicated with a black box. **b** The R^2 coefficients of determination between $\log_{10}$ translation rates (TR) and PWM scores. PWMs of varying lengths were built from the sequences of the high TR cohort, the PWMs extending 5′ from − 1 in 5 nucleotide increments and extending 3′ from + 4 in 5 or 10 nucleotide increments. Log odds scores were then calculated for all mRNAs that completely contained a given PWM. Additional file 1: Figure S6 shows more detailed mapping of the 5′ and 3′ boundaries for *M. musculus* and *H. sapiens*. The frequencies of nucleotides in the PWMs are given in Additional file 7

strongly explain translation. Di- and tri-nucleotide frequency features were selected for the portion of APE upstream of the iAUG (uAPE) and separately for the portion downstream of the iAUG (dAPE), 4–16 di- and tri-nucleotides being selected per region (Additional file 1: Figure S7, Additional file 5). In the resulting multivariate models, *S. cerevisiae*, *S. pombe*, and *Arabidopsis* APEs exert stronger control than their mammalian counterparts, the former explaining 16–33% of the variance in translation and the latter only 3–4% (Additional file 1: Figure S7).

Because the above models capture only part of the 5′ region, we also examined the motifs in the area that extends from just 5′ of the APE to the 5′ cap (5′ofAPE). We found that an effective model for this 5′ofAPE region included a five nucleotide length PWM at the 5′ cap and a subset of di- and tri-nucleotide

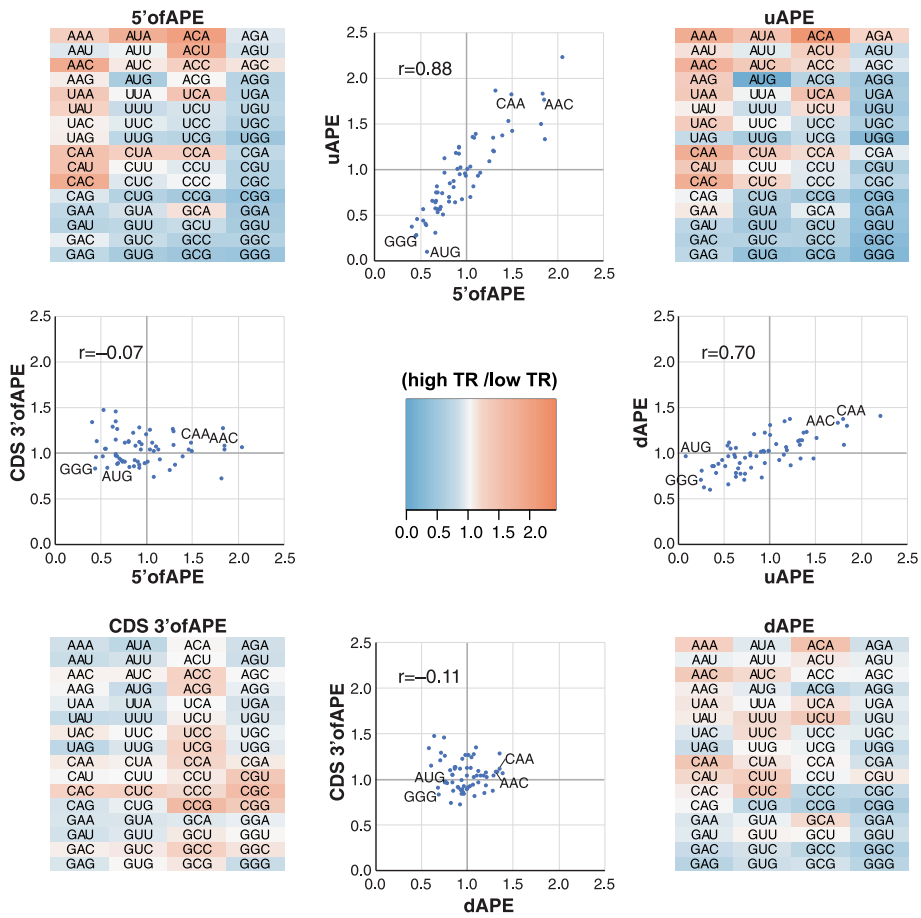


Fig. 5 Correlations between regulatory regions in *Arabidopsis* mRNAs. The four heat maps show the frequency of each tri-nucleotide in the most highly translated 10% of genes divided by its frequency in the most poorly translated 10% of genes (high TR/low TR ratio). Ratios were calculated for four separate parts of the mRNA: 5′ofAPE, uAPE, dAPE, and the CDS 3′ofAPE. Scatter plots show the correlation in high TR/low TR ratios between the selected regions. The Pearson correlation is given. Data points for four tri-nucleotides are indicated (GGG, AUG, CAA, AAC). Strong positive correlates are seen between 5′ofAPE and uAPE and between uAPE and dAPE. CDS 3′ofAPE only correlates weakly with the other regions, consistent with its different function

frequencies for the full region that was selected using BIC (Additional file 5; Additional file 1: Figure S8). In non-mammalian species, 5′ofAPE sequences are much less predictive of translation than the APE, whereas in mammals, the reverse is found (Additional file 1: Figure S9). Multivariate models that combine 5′ofAPE and APE motifs provide our most accurate prediction of translation rates using 5′ region sequence motifs (feature set "5′motifs," Additional file 1: Figure S9).

To determine if control sequences are similar within and between species, we divided each mRNA into four parts: 5′ofAPE, uAPE, dAPE, and the CDS 3′ of the APE (CDS 3′ofAPE). We then calculated the frequencies of each tri-nucleotide in each region of each gene for the high TR and low TR cohorts separately and took the ratio of the means of these values, yielding "high TR/low TR ratios" (Additional file 8). Six results stand out:

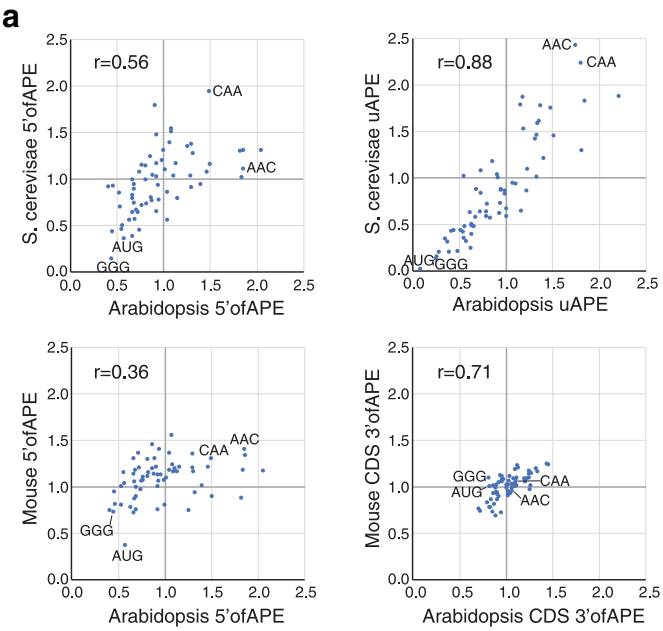
i. The (high TR/low TR) ratios in 5′ UTRs vary from 0 to 3.6, with most ratios differing markedly from one (Figs. 5 and 6; Additional file 1: Figure S10; Additional file 8). This suggests that a high proportion of all nucleotides in these regions contribute to translational regulation and that the tri-nucleotides whose ratios differ most from one make the largest contributions.

ii. In 5′UTRs, AUG has one of the smallest (high TR/low TR) ratios of any tri-nucleotide (mean (high TR/low TR ratio) = 0.31), showing that our motifs detect the inhibition of initiation at the CDS that results from translation of uORFs (Fig. 5; Additional file 1: Figure S10; Additional file 8).

iii. Within each of the two species that have extended APE sequences—*S. cerevisiae* and *Arabidopsis*—the uAPE and dAPE share strong sequence similarities, as shown by the positive correlation between their (high TR/low TR) tri-nucleotide ratios ($r = +0.41$ and $+0.70$; Additional file 1: Figure S11).

iv. As expected, the CDS 3′ofAPEs show little similarity to 5′UTRs ($r < 0.1$; Additional file 1: Figure S11).

v. Some of the strongest (high TR/low TR) ratio correlations are between the uAPEs of the three non-mammals ($r > 0.65$) and separately between the 5′ofAPEs of *M. musculus* and *H. sapiens* ($r = 0.8$) (Fig. 6). There are also, however, lower but clear correlations between the 5′ regions of mammalian and non-mammalian species (Fig. 6). Hence, while the mRNA GC contents of these two groups of eukaryotes differ greatly, the variations in the 5′ region sequences that direct translation are related across the five species.



b

	5'ofAPE	uAPE	dAPE	CDS 3'ofAPE
S.c vs S.p	0.2	0.65	0.24	0.64
S.c vs A.t	0.56	0.88	0.4	0.37
S.c vs M.m	0.33	0.32	0.03	0.38
S.c vs H.s	0.05	0.18	−0.03	−0.12
S.p vs A.t	0.39	0.68	0.1	0.66
S.p vs M.m	0.31	0.14	0.55	0.68
S.p vs H.s	0.04	0.14	0.47	−0.38
A.t vs M.m	0.36	0.15	0.05	0.71
A.t vs H.s	0.01	0.12	0.09	−0.52
M.m vs H.s	0.8	0.44	0.34	−0.32

Fig. 6 Correlations between translational *cis*-regulatory sequences in different species. **a** Scatter plots showing strong correlations between (TR high/TR low) tri-nucleotide ratios for selected genomic regions from different species. The Pearson correlations (*r*) are indicated. **b** Pearson correlation coefficients between each of the four mRNA regions among the five species. The color intensities are scaled to the correlation coefficient. Clear correlations are observed in many cases

vi. The CDS 3′ofAPE shows a positive correlation in (high TR/low TR) tri-nucleotide ratios between all species except *H. sapiens*, which shows a negative correlation with the other species (Fig. 6). This last probably reflects a previously characterized negative correlation in tRNA abundances—and as a consequence in optimum codon usage—between cancerous cell lines and other cell types [35] rather than a difference between species as the *H. sapiens* dataset was derived from cancerous HeLa cells.

A model for control by 5′ regions

mRNA 5′ regions control translation in part through their secondary structures and uORFs, both of which reduce the frequency of preinitiation complexes that reach the iAUG. Our motif models for 5′ regions likely include sequence information for these two general features as well as capturing general machinery contacts close to the iAUG and, potentially, other *cis*-elements also. It is interesting, therefore, to quantitate the overlap of regulatory information in models for RNA structure, for uORFs and for 5′ sequence motifs.

For this purpose, we represented regulation by RNA structure using the "RNAfold" feature set, repression by uORFs using the number of upstream AUGs (uAUGs) present in 5′UTRs (uAUG), and the combination of these two using a further multivariate model: "5′biochem." To calculate the overlap of the three resulting biochemical feature sets with "5′motifs," three further models were constructed that paired "5′motifs" with each biochemical feature set (Fig. 7). The regulatory overlap is given by the percent of the variance in

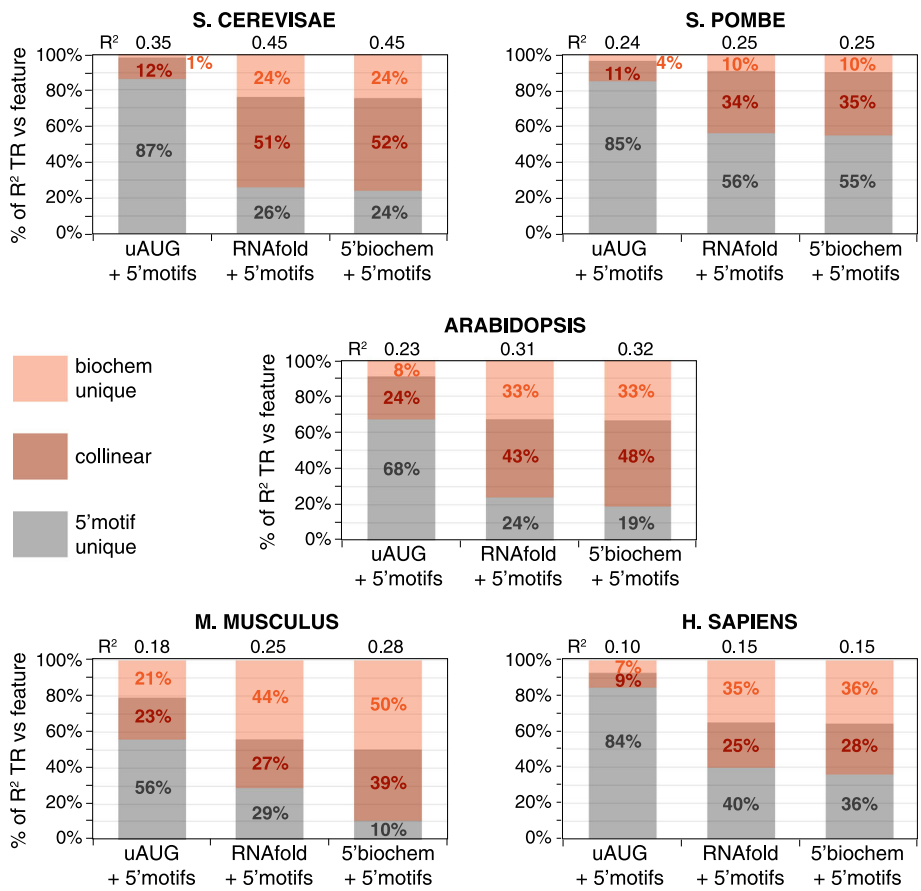


Fig. 7 Control by "RNAfold" and "uAUG" is collinear with that by "5′motifs." Two 5′ features each partially capture control by a biochemical process: "uAUG," repression due to translation of uORFs; and "RNAfold," inhibition by RNA secondary structure. A third feature, "5′motifs," captures 5′ sequences that correlate with translation, without regard to biochemical mechanism. To determine how similar information captured by "5′motifs" is to that in "uAUG" and "RNAfold," three multivariate models were constructed that combined "5′motifs" either with "uAUG," with "RNAfold," or with a third feature (5′biochem) that combines both "uAUG" and "RNAfold." The R^2 coefficients of these feature sets vs TR are shown (top). The percent of the R^2 values contributed by "5′motifs" (gray) or the biochemical process(es) (orange) is shown by length along the *y*-axis. Because some of this information is collinear, the R^2 coefficient of the combined models is less than that of the sum of the individual contributions. This collinear portion is shown by the darker orange shading, while the lighter orange and the gray shading show the unique contributions

translation that is collinear (i.e., shared) between the feature sets. Control information in "5′motifs" overlaps strongly with that in "uAUG," "RNAfold," and "5′biochem" and vice versa (Fig. 7). For example, 39–78% of the explanatory power of "5′motifs" is collinear with "5′ biochem." At the same time, "uAUG," "RNAfold," and "5′motifs" each contain unique *cis*-regulatory information not captured by the other. For instance, 48–72% of the information in the model combining "5′biochem" and "5′motifs" is uniquely obtained from the models that employ only one of these two feature sets, making the combined model our most complete for explaining the 5′ region control.

The contributions of general features to translation rates

In addition to the 5′ region general features, two others control translation rates: CDS length and codon frequency. CDS lengths are inversely correlated with translation because ribosomal subunits released from mRNAs are recycled from the stop codon to the 5′ cap (40S) or iAUG (60S) within single mRNA molecules, and this process is more efficient for a shorter CDS [9]. Codon frequencies in highly expressed mRNAs more closely match the concentrations of aminoacylated tRNAs than do those in poorly expressed mRNAs, which leads to more efficient use of the pool of translational machinery in the cell [10, 13, 35–37].

Therefore, we calculated the features sets that capture regulation by codon frequency and CDS length (features "codon" and "CDS length") and built a model that combines them with the feature sets that describe control by the 5′ region. The results were calculated for the five example datasets (*S. cerevisae*, *S. pombe* 2, *Arabidopsis* shoot 2, *M. musculus* NIH3T3 cell line, and *H. sapiens* HeLa cell line), for a second *S. pombe* dataset (*S. pombe* 1), and for the data from additional tissues or biological replicas for *Arabidopsis* (root 1 and shoot 1) and *M. musculus* (whole liver and whole kidney). The combined models explain 37–81% of the variance of translation, depending on the dataset, and at least 42% for the best predicted dataset from each species (Fig. 8b; Additional file 9; Additional file 1: Figure S12). These results are robust to variations in data as shown by the narrow 95% confidence limits from a 1000× bootstrap (Fig. 8b). Our results thus suggest a considerably larger role for the general elements than previous models implied (Additional file 1: Table S1). We have also compared the individual contributions of the general features as the percent of either the sum of their five R^2 values vs translation (Fig. 8a) or the R^2 value of a model combining all five features (Additional file 1: Figure S13). By either approach, the contributions of each feature vary more between the species than between different datasets from the same organism, suggesting that

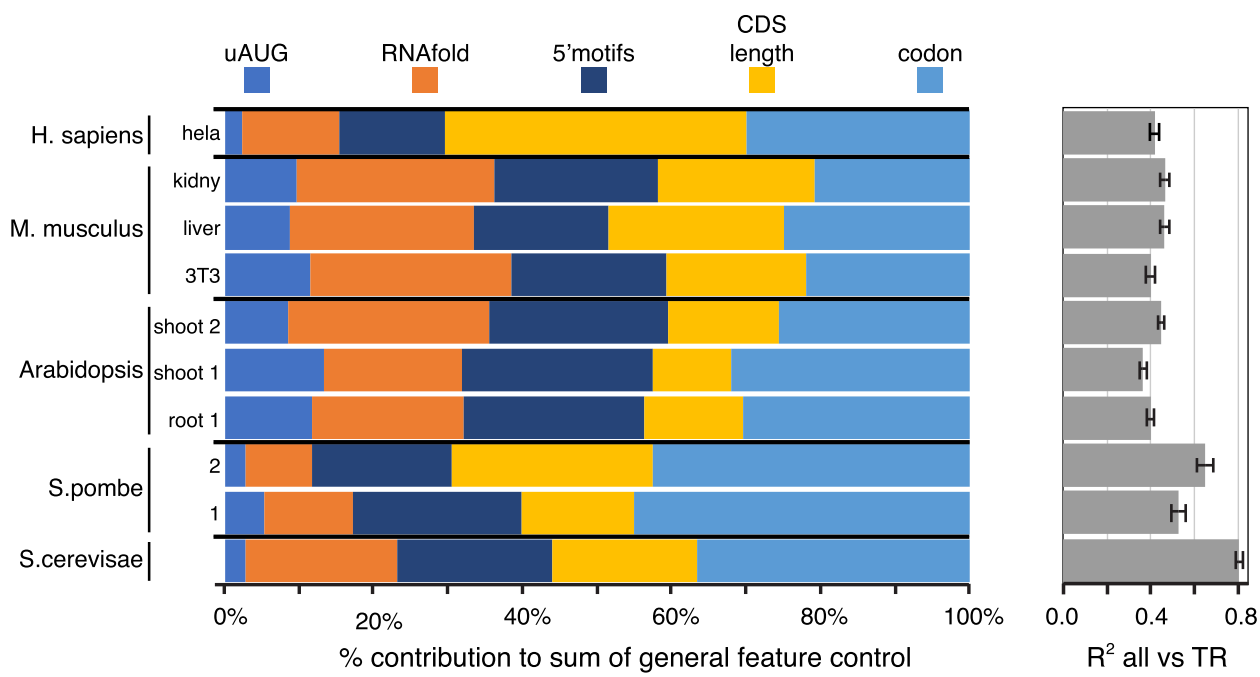


Fig. 8 The contributions of general *cis*-control features in five species and several tissues. **a** For each feature or feature set separately, its R^2 coefficient of determination vs $\log_{10}$ translation rate (TR) is given as a percent of the sum of the R^2 coefficients for linear models for each of the five features. **b** The R^2 coefficients between $\log_{10}$ TR and a multivariate linear model that combines the five features. The 95% confidence limits are shown for 1000× bootstraps with replacement. The R^2 values and the values plotted for each feature and the combined model are given in Additional file 9

aspects of translational regulation alter quantitatively between species.

Metazoan genes in particular often encode multiple mRNA isoforms. These isoforms vary in their transcription start sites, splicing patterns, and poly-A addition sites, and in some cases, isoforms for the same gene are translated at different rates [38–42]. Because our analysis employs a single representative mRNA isoform for each gene, such isoform complexity could introduce error, which would cause our models to underestimate the contribution by the general features. Adapting an approach used by others to reduce error due to multiple isoforms [16, 28, 43], we tested the ~ 30% of *Arabidopsis* and *M. musculus* genes for which only a single mRNA isoform is known (see the "Materials and methods" section). The predictive power of multivariate models for translation control by this subset of genes is, however, substantially similar to that of the models for all genes (Additional file 1: Figures S14 and S15), indicating that any variation in translation rates between mRNA isoforms is not sufficient to influence our results. Most metazoan genes that express multiple isoforms produce only one dominant form in a given tissue, and most isoforms for a given gene encode the same protein variant [40, 41]. Thus, it is not unreasonable that models employing a single representative mRNA isoform per gene should perform well.

Collinear control

The sum of the variances in translation rates predicted by each of the general features separately is 157–203% of the variance explained by the model that combines all five features (Additional file 1: Figure S13). Four of the features each represent a separate biochemical process (the exception being "5′motifs," which captures information redundant with "RNAfold" and "uAUG," see Fig. 7). The *cis*-elements that direct these four biochemical processes must, therefore, have coevolved to work in a correlated manner.

To quantitate the overlap in control, we calculated collinearity between pair-wise combinations of general features within each species. "5′motifs" was excluded, though, to avoid its redundancy with "RNAfold" and "uAUG." All feature pairs tested show collinearity (Fig. 9). This collinearity is due to the inherent correlations among the general mRNA features because if we randomly permute each feature's observed values independent of other features, the collinearity among features is expected to go down to zero. The overlap in control is largest in the two yeasts, where five out of the six pairs have a remarkable 47–100% overlap, and smallest in the two mammals, where five of the six pairs have 22–59% overlap. Some feature pairs are strongly collinear in all species: for example, "CDS length" and "codon," which show 45–62% overlap. Other pairs show dramatic differences: for example, "uAUG" is ≥ 70% collinear with "RNAfold" in *S. cerevisiae*, *S. pombe*, and *Arabidopsis* but only 7% in *M. musculus*.

The collinearity between features in the 5′ region and either "codon" or "CDS length" occurs between physically distinct parts of the mRNA. Our earlier analysis of the 5′ region motifs provides additional evidence of correlated regulation between distinct mRNA segments: for example, in the two species with long APEs, tri-

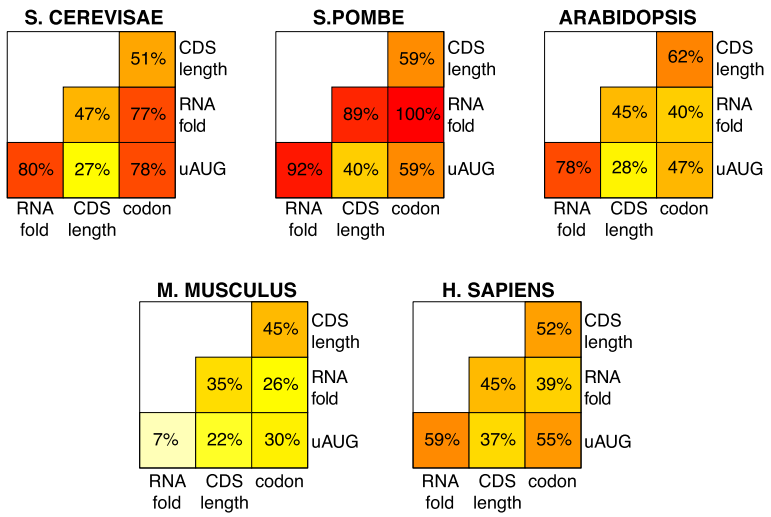


Fig. 9 Distinct biochemical processes control translation collinearly. For two features A and B, the R^2 coefficient of a model that combines both features vs $\log_{10}$ translation rate (TR) was defined as R^2AB, while the R^2 coefficients of each feature separately vs $\log_{10}$ TR were defined as R^2A and R^2B. The larger of (R^2A + R^2B-R^2AB)/R^2A or (R^2A + R^2B-R^2AB)/R^2B is given and represents the percent in overlap of regulatory control. The overlap in regulatory information varies widely between different pairs, both within and between species

nucleotide frequencies correlate strongly between the 5′ofAPE, the uAPE, and the dAPE (Fig. 5 and Additional file 1: Figure S11). To characterize the collinearity between the physically distinct parts of mRNAs further, we identified the second and third most folded windows within the 5′ regions that do not overlap with each other or with the most folded window. Sixty-three to 98% of control exerted by the second most folded window is correlated with that by the most folded window (Fig. 10a). Similarly, 67–92% of control by the third most folded window is collinear with that of the most of folded window (Fig. 10b). We also tested for collinearity within the CDS by calculating codon frequencies in N-terminal halves and separately in C-terminal halves. Control by the C-terminal half is 49–85% collinear with that by the N-terminal half (Fig. 10c).

The correlation of codon frequency with translation rates is determined both by the frequencies of amino acid in proteins and by the preference for some synonymous codons vs others [10, 44]. Synonymous codon

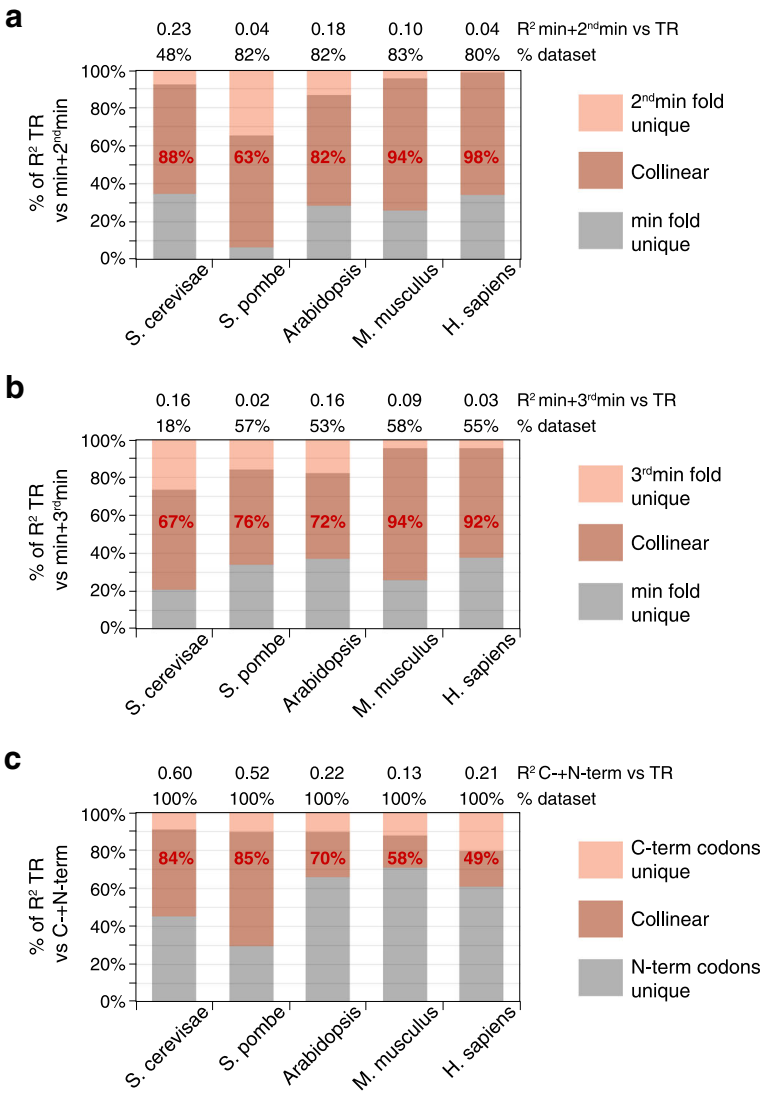


Fig. 10 Collinear control between differently located RNA segments. **a** A model was built for the combined control of translation rate (TR) by both the most highly folded window (min) and the non-overlapping second most highly folded window (2^{nd}min). The R^2 coefficients for these combined features vs $\log_{10}$ TR are given (top). The percent of the unique and collinear contributions of "min" and "2^{nd}min" to the total variance in translation explained by the model for both features is shown by length on the *y*-axis. The percent of control of TR by "2^{nd}min" that is collinear with control by "min" is shown in red text. The percent of genes in the dataset whose 5′ regions are long enough to contain the two non-overlapping windows used in the analysis are indicated (top). **b** The collinearity between the most highly folded window (min) and the non-overlapping third most highly folded window (3^{nd}min) is displayed as described in **a**. **c** The collinearity between regulation of translation by codon frequency in the N-terminal half of each protein (N-codon) and C-terminal half of each protein (C-codon) is shown as described in **a**, except that the percent of control of TR by "C-codon" that is collinear with control by "N-codon" is shown (red text), and 100% of genes in each dataset were used (top)

preferences and amino acid content are logically independent of each other, i.e., in the absence of empirical data, one cannot assume a priori that their variation between genes is related (Fig. 11a). It is possible, then, that their effect on translation could be either collinear or non-collinear. To test for any collinearity, we calculated for each gene both the frequency of each amino acid ("AA," 20 features in total) and the synonymous preference ratio for each codon ("syn.codon," 61 features in total) (Fig. 11a). Control by both "AA" and "syn.codon" correlates, as expected, with tRNA abundances (Additional file 1: Figure S16) [10, 13, 35–37]. Importantly, we find that the regulation of TR by "AA" and "syn.codon" is strongly collinear in all species (Fig. 11b).

Our various analyses thus reveal surprising, strong, and complex co-evolution of control between distinct biochemical processes and between multiple segments of mRNA.

Discussion and conclusions

Here, we have sought to better understand how and why the differences in translation rates among genes are regulated at steady state by focusing on general mRNA sequence features. We have established common quantitative principles, discussed below, by which these control elements act in diverse eukaryotes. Through optimizing methods to predict translation using mRNA structures, sequence motifs, uORFs, CDS length, and codon usage, we find that the general features together control 81% and 65% of the variance in rates in *S. cerevisiae* and *S. pombe*, respectively, and 42–46% in *Arabidopsis*, *M. musculus*, and *H. sapiens* (Fig. 8). This is significantly higher than earlier estimates (Additional file 1: Table S1). For example, prior work suggested that general features explained 39% of the variance in *S. cerevisiae* [13] and 14% in *M. musculus* [16].

The part of the variance in measured translation rates that is unexplained by our models for general features is likely due to a combination of the following:

i. Gene/condition-specific regulation by miRNAs and sequence-specific RNA-binding proteins, which our models do not detect
ii. Failure of our models to fully capture control by the general features due to, for example, uncertainty in fully predicting RNA structures
iii. Measurement error in the translation rate data

It is also possible that our models have exaggerated control by the general features due to "overfitting," which would lead to aberrant capture of some part(s) of i. and/or iii. Our conservative approach of fitting linear models to features that score known biochemical processes or simple motifs should limit this, however. The likelihood of

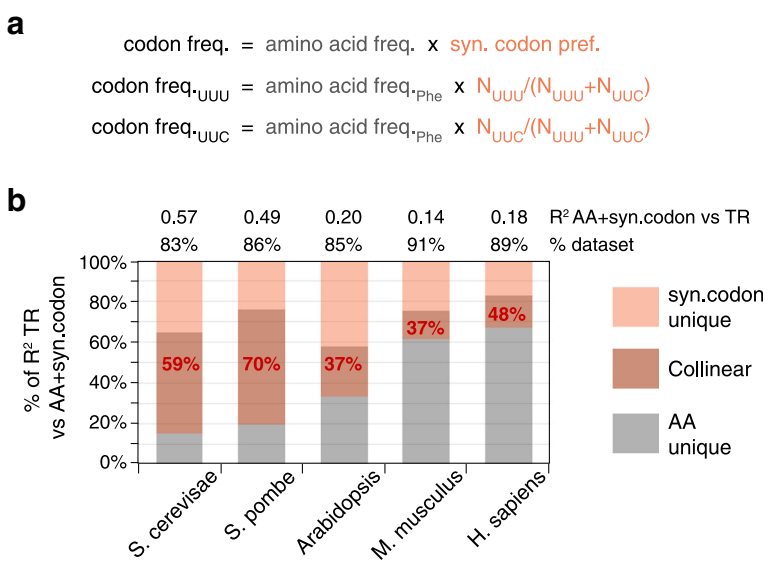


Fig. 11 Collinear control by amino acid frequency and synonymous codon preference. **a** The frequency of a codon in a gene is the product of its cognate amino acid frequency (AA) and its synonymous codon preference (syn.codon), where the latter = (number of occurrences of the codon)/(number of occurrences all codons for the cognate amino acid). Examples for the two codons for phenylalanine (Phe) are shown. A priori, there is no necessity for "AA" and "syn.codon" to be correlated. **b** A model was built for control of translation rates (TR) by the combination of "syn.codon" (61 features) and "AA" (20 features). The R^2 coefficients for this model are given (top). The percent of the unique and collinear contributions of "AA" and "syn.codon" to the total variance in $\log_{10}$ TR explained by the combined model is shown by length on the *y*-axis. The percent control of TR by "syn.codon" that is collinear with that by "AA" is shown in red text. The percent of the dataset that contains all 20 amino acids and thus could be used to calculate "syn.codon" is also shown (top). This set of genes was used for all analyses

overfitting is reduced further by our use of Bayesian Information Criterion for feature selection, which we found typically selected fewer features than other approaches, such as tenfold cross-validation (unpublished data). Our models, thus, probably have underestimated true control by the general *cis*-elements by the—unknown—percent of the variance in the translation rate data that is due to ii. and iii.

It is noteworthy that our general feature models predict translation rates more effectively in *S. cerevisiae* and *S. pombe* than in multicellular eukaryotes. miRNAs are not present in the two yeasts, whereas *Arabidopsis* encodes ~ 420 miRNAs and *M. musculus* and *H. sapiens* > 1800 each [45, 46]. Likewise, *S. cerevisiae* and *S. pombe* encode considerably fewer sequence-specific RNA-binding proteins than the multicellular eukaryotes (~ 100 vs > 350; http://cisbp-rna.ccbr.utoronto.ca [47, 48]). Thus, the general features may well play a larger role in the two yeasts than in the multicellular eukaryotes as the latter are expected to suffer greater gene and condition-specific regulation.

5′ mRNA secondary structure

Our analysis provides a more precise understanding of the mRNA secondary structures that control translation within the 5′ regions. We have found that the most folded 25–60 nucleotide segments contain far more regulatory information than the less folded parts of the 5′ regions in all five species (Figs. 2 and 3a). Previous studies, by contrast, relied on the mean of the folding energies of short windows that tile the 5′ region, on the folding of a single large window encompassing the entire region, or on windows at fixed locations (Additional file 1: Table S1). Our analysis indicates that the most folded windows alone are more predictive than these prior approaches and that models that combine highly folded windows with location-specific ones are even more effective (Figs. 2 and 3a; Additional file 5). Control by the most folded segment is dominated by the longest contiguous run of paired nucleotides within a stem (Fig. 3b). This control is enhanced by additional nucleotide pairs that lie beyond any mismatched nucleotides or single nucleotide bulges to form a longer stem and by smaller satellite stem loops that reside within the segment (Fig. 3b). Pairings between nucleotides that are separated by > 25–60 nucleotides have only a minimal effect on the translation rates (Fig. 2). The differences in structures that distinguish translation rates are surprisingly small: only 0.7–1.6 nucleotide pairs on average differentiate the most folded segments in highly translated mRNAs from those in poorly translated mRNAs across all species (Additional file 1: Figure S4).

Conserved 5′ sequence motifs

One of the approaches that we have used to define general *cis*-control elements is an unbiased method that simply categorizes the differences in the 5′ region sequence motifs between highly and poorly translated mRNAs (Figs. 4, 5, and 6). Non-mammalian species have A/T-rich 5′ regions, whereas those of mammals are G/C rich. Despite these differences, we find that when judged by the ratios of the frequencies of each tri-nucleotide in highly and in poorly translated mRNAs (high TR/low TR ratios), there are marked correlations between the *cis*-control sequences in the 5′ regions of mammalian and non-mammalian eukaryotes (Fig. 6). The fact that most tri-nucleotides' (high TR/low TR) ratios ≠ 1 and are similar across many species (Figs. 5 and 6; Additional file 1 Figure S10) strongly implies that a large percent of nucleotides in the 5′ regions are involved in regulating translation, rather than having some other function or no function. Those tri-nucleotides whose (high TR/low TR) ratios differ the most from one are presumably those most important for control.

The correlation in 5′ motifs between distant species occurs in part because our motif model captures repression by uORFs via the higher density of uAUGs in poorly translated mRNA (Fig. 6). It also results because our 5′ motifs model captures control by evolutionarily conserved nucleotides that are adjacent to the iAUG and contact the ribosome (Fig. 4) [34, 49–51]. Additionally, the correlation is probably due to the fact that di- and tri-nucleotides have differing propensities to form mRNA secondary structures because of their respective base pairing and stacking energies [29, 32, 52, 53]. The important role of RNA structure in directing translation rates will thus drive similar changes in nucleotide content between highly and poorly translated mRNAs in all species. Indeed, 38–66% of translational control explained by our 5′ motifs model is collinear with control by our model for RNA folding (Fig. 7), and those mRNA regions that show the greatest conservation of tri-nucleotide ratios between highly diverged eukaryotes also have the largest differences in RNA folding energies between highly and poorly translated mRNAs (Figs. 1, 2, and 6).

Collinear control between processes and between mRNA segments

Perhaps our most striking observation is that there is a strong correlation in control (i.e., collinearity) between general features that act through distinct biochemical processes and also between different parts of the mRNA. For example, all pair-wise combinations of the four biochemical processes show collinear regulation of translation, many by > 50% (Fig. 9). Also, the Gibbs energies of the second and third most folded mRNA segment within the 5′ regions share a remarkable 63–98% control with those of the most folded segment, and codon frequencies in the C-terminal half of CDSs share 49–85%

control with those in the N-terminal half (Fig. 10). Finally, regulation by synonymous codon preferences is 37–70% collinear with that by amino acid content, even though the two determinants of codon usage are logically independent of each other (Fig. 11). This extensive collinearity reveals a form of strongly repeated information within each mRNA that other approaches, such as testing for nucleotide sequence similarities, have failed to detect, illustrating the power of the analysis of variance (ANOVA)-based approach we have used.

Collinearity between different sections of the mRNA keeps the rate of progress of the translation machinery more constant and thus the density of translational complexes more uniform along any given mRNA (Fig. 12a). The translation rate within each mRNA is not determined by a single rate-limiting step in this case. Instead, the flux of complexes down a particular mRNA is controlled by a series of steps that are each set to allow a similar rate. The steps share regulation much in the way that the enzymes in a metabolic pathway share fractional control of metabolic flux, as defined by their flux control coefficients [54, 55]. Uncorrelated control, by contrast, would lead to a more uneven distribution of complexes. In this alternate case, the overall flux of complexes for a given mRNA would often be defined by a single rate-limiting step, with the important proviso that different steps would limit on different genes to account for the fact that all general features contribute to the overall variance in rates (Fig. 12b).

There are two known selective pressures that favor—or force—uniform densities of complexes and thus collinear control over non-collinear control.

i. Non-collinear control would increase the frequency of particle collisions when preinitiation complexes or ribosomes that are moving rapidly down an mRNA enter a region whose regulatory sequences signal a slower rate of progress (Fig. 12b, bottom). Ribosomal collisions lead to translational termination and no-go RNA decay or simply slow the mean elongation rate [56–58]. Collinear control between mRNA segments thus reduces the frequency of these deleterious events. Earlier work has shown that the rate of translation is slower near the N-terminus of the CDS then increases moderately further 3′ [13, 36]. This arrangement has been proposed to reduce collision frequency [36]. Our results extend this prior model to suggest that it is important to keep a relatively even flux of translational complexes—avoiding strong stops—from the assembly of the preinitiation complex at the 5′ cap all the way to the termination of translation at the stop codon.

ii. Collinear control requires fewer ribosomal subunits per cell than non-collinear control to achieve a given set of translation rates (compare Fig. 12a, b). This is important as the translational machinery is limiting for protein synthesis and cell growth [59–61]: for example, a two-fold reduction in the abundances of any one of a number of ribosomal proteins greatly reduces cell division rates in *Drosophila* [59], and under many conditions, ≥ 90% of ribosomal subunits are bound to mRNA and active at any instance [62, 63] and/or are present at locally high concentrations in the vicinity of each mRNA due to CDS length-dependent recycling [9]. There is thus a strong selective pressure to use the set of ribosomal

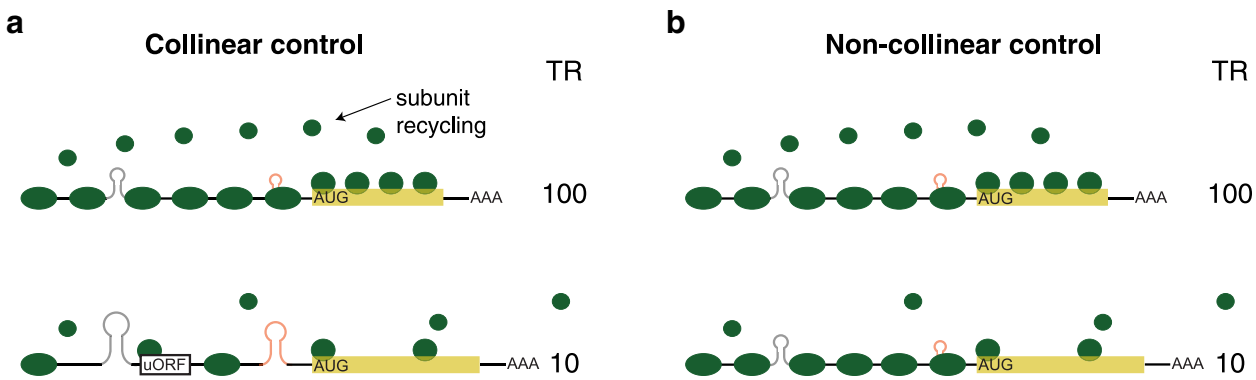


Fig. 12 Collinear control results in more uniform particle densities than non-collinear control. **a** Collinear control. The upper, high translation rate (TR) mRNA has multiple small mRNA stem/loops, no uORFs, a short CDS, and optimal codon frequency. The bottom, low TR mRNA has multiple larger mRNA stem/loops, uORFs, a long CDS, and non-optimal codon frequency. As a result, both mRNAs have relatively even densities of translational complexes along their length. The CDS length-dependent rate of ribosome subunit recycling also matches their respective translation rates. **b** Non-collinear control. Both mRNAs have multiple small mRNA stem/loops and no uORFs resulting in a high density of preinitiation complexes on the 5′UTR. The top mRNA has a short CDS and optimal codon frequency, which results in a high density of ribosomes along the CDS and efficient ribosome subunit recycling. The bottom mRNA has a suboptimal ribosomal contacts at the iAUG, which reduces the efficiency with which preinitiation complexes convert to active ribosomes. As a result, the density of ribosomes along the CDS and in the recycling step is lower than that directed by the 5′UTR. The density of translational complexes along this mRNA is thus uneven. The rates of translation of mRNAs in **a** and **b** are the same. The total number of translation machinery complexes needed in a cell, however, is greater in the non-collinear scenario than in a cell employing collinear control

subunits in the cell most efficiently. As Fig. 12 shows, this requires collinear control of the complete translation cycle, including the recycling step.

In addition to the two above explanations, a general pressure to increase or decrease protein production from a given gene could lead to multiple, independent mutations affecting different processes or the same process in different mRNA regions. This too would result in collinear control. However, we suggest instead that the discovery of strong collinear control supports other evidence that gene-specific translation rates are not determined for the most part by the need to achieve a certain protein abundance. First, translation rates only correlate poorly with protein abundance ($R^2 \sim 0.08$ in *S. cerevisiae* and *M. musculus*) [10, 64]. Second, while CDS length is an important determinant of translation rates (Fig. 8), the length of a protein is unlikely to be selected only to affect translation. Protein length is presumably set largely by protein function. There is a caveat to this, though: it is plausible that highly abundant proteins that are members of multi-subunit complexes, e.g., histones and ribosome, are selected to be short to allow efficient translation. Third, codon frequency—and thus its correlation with translation—is determined by the amino acid content of proteins as well as by the preferences for codons synonymous for the same amino acid (Fig. 11). Amino acid contents probably result from selection for protein function rather than for translational regulation. Synonymous codon frequencies in highly abundant mRNAs are selected to optimize the use of the total pool of aminoacylated tRNAs in the cell [10, 13, 35–37] (Additional file 1: Figure S16); the effect that codon usage has on ribosomal elongation rates could be a secondary consequence, one that is, besides, generally small [13, 65, 66].

Thus, both CDS length and codon frequencies are likely driven in part by forces external to the need to determine protein abundances, yet combined the two constitute a significant fraction of control by the general features (Fig. 8). The proposed need for collinearity to ensure efficient use of ribosomes by the cell will then force the other three general features to coordinate their activities with CDS length and codon frequency. This may explain why translation is not well correlated with protein abundance. Translation rates are determined by multiple selective pressures. In contrast, the selective pressure on protein abundance dominantly affects transcription, at least for the few conditions for which this has been most carefully measured by ANOVA [10, 67].

S. cerevisae and other yeasts have larger effective population sizes than *Arabidopsis*, *M. musculus*, and *H. sapiens*. As a result, the former have lower mutational loads and have been able to better optimize their genome sequences for energy-efficient cell growth than have the latter [68, 69]. These results from population genetics provide an additional explanation for our observation that the general features play a more prominent role in *S. cerevisae* and *S. pombe* than in the three multicellular eukaryotes: the general elements should be more important in organisms with larger effective population sizes, an idea that has been suggested previously for synonymous codon preference [68, 70] and which we now extend to the other general features.

High throughput mutant assays

A separate approach to identify translational *cis*-elements has tested large numbers of heterologous reporter genes bearing random sequences or mutations in 10 to 50 nucleotide segments proximal to the iAUG [12, 17–21]. In some studies, models have been developed that explain ~ 70–90% of the variance in translation resulting from these sequence changes (Additional file 1: Table S2).

It is difficult to relate the results of these reporter gene experiments to the regulation of endogenous mRNAs. The regions bearing sequence changes in the heterologous reporter assays are smaller than or lie outside of the regions where our analysis indicates that regulatory secondary structures are prevalent. Thus, the reporter assays have not captured normal control by RNA folding, let alone that by CDS length and codon usage, which are invariant in a reporter assay. In addition, the consensus sequences for uAUGs associated with repression in a heterologous assay differ dramatically from those identified by studies of uORFs in intact natural transcripts. The high-throughput reporter gene assays identified strong enrichment of A or G at nucleotide − 3 relative to uAUGs, similar to the Kozak consensus for the iAUG [20]. In endogenous mRNAs, by contrast, no specific nucleotides are strongly associated with repressive uAUGs at location − 3 [16, 25, 71]. Thus, the sequences of the *cis*-elements driving translation in the heterologous assays differ strongly from those acting in endogenous mRNAs.

One *S. cerevisiae* study did measure the protein production from natural 5′UTRs ≤ 50 nucleotides in length fused upstream of a reporter gene [21]. Although this study successfully modeled 60% of the variance in measured output, we find that this output correlates poorly with ribosome profiling data for the corresponding endogenous genes ($R^2 = 0.09$, Additional file 10). Thus, at least (60−9)/60 = 85% of the variance in reporter expression predicted by the model cannot be explained by measured endogenous translation rates.

Our analysis establishes that there is an extensive coordination (collinearity) between control by distinct parts of the mRNA and between multiple steps in the translation cycle. Heterologous reporter constructs that include only short segments of mRNAs thus fail to correctly capture *cis*-elements in part because they miss the

normal interactions between steps and between differently located *cis*-elements. It will be necessary to design alternative mutant series to understand *cis*-translational control of endogenous transcripts.

Materials and methods

Data and code

An example dataset was chosen from each of the five species to present figures for all analyses: *S. cerevisiae* [13], *S. pombe* 2 [26], *Arabidopsis* shoot 2 (dark) [25], *M. musculus* NIH3T3 cell line [4], and *H. sapiens* HeLa cell line [24]. Results from five additional datasets are shown for a subset of analyses: *S. pombe* 1 [4], *Arabidopsis* root 1 and *Arabidopsis* shoot 1 [27], and *M. musculus* liver and *M. musculus* kidney [14, 28]. The example and additional datasets were produced by a variety of protocol variants in several laboratories. Our main conclusions are true across multiple datasets from different species, tissues, and protocol variants, suggesting that they are not influenced by the specifics of particular methods. Datasets were chosen in which ribosome footprints showed the expected strong enrichment for the first nucleotide in each codon and for which there was a positive correlation between translation rates and mRNA abundances.

Where replica data was available from a study, genes without data in all replicas were removed; for each gene, its data were then averaged across all replicas; and if the data in the original publication had not been thresholded, genes with mRNA abundances < 1 RPKM were excluded. For all datasets, genes for which data on poly-A tail length were not available from Subtelny et al. were removed. The mRNA sequence isoforms used for generating sequence features were those used in the original publications for determining TR and mRNA abundance values for *S. cerevisae*, *S. pombe*, *Arabidopsis* root 1 and shoot 1, *H. sapiens*, and *M. musculus* NIH3T3. In the case of the *M. musculus* liver and kidney datasets, the mRNA isoform with the longest 5′UTR in Ensemble genome version GRCm38 was used, and for the *Arabidopsis* shoot 2 dataset, the TAIR10 representative gene model that was also employed for root and shoot was used. Genes that express a single mRNA isoform in *Arabidopsis* were identified using the AtRTD2 annotations [42] and in *M. musculus* using the Ensemble GRCm38 annotations. Additional file 2 provides all mRNA sequences, ribosome profiling data, and mRNA abundances as well as scores for each of the general features for the set of genes analyzed for each dataset.

The complete code files for reproducing the results presented are available at Zenodo: https://doi.org/10.5281/zenodo.3272133 [72]. These files include the R code, input data, processed files, and preliminary figures.

Common aspects of models

Translation rate (TR) is defined as translation efficiency (TE) values from ribosome profiling data or in the case of *S. cerevisiae* as initiation efficiency (IE) values, which are corrected TE values that take into account codon-specific elongation rates [13]. Log_{10} transformed TR values were used in all regressions.

All models employed single part, multivariate linear regressions for *S. pombe*, *M. musculus*, and *H. sapiens*; two-part, multivariate linear regressions of *Arabidopsis*; and three-part, multivariate linear regressions for *S. cerevisiae*. The multipart regressions were employed in the latter two species to allow optimal scoring of their extended APE PWMs, described below. Of necessity, multipart regression then had to be employed for all other features in these species to allow a model that combines all general features. For the multipart regressions, genes were divided into groups based on lengths of 5′ untranslated regions (UTRs), with the number of groups equal to the number of parts. For *Arabidopsis*, genes were grouped into those with 5′UTRs < 65 nucleotides and those with 5′UTRs ≥ 65 nucleotides. For *S. cerevisiae*, genes were grouped into those with 5′UTRs < 20 nucleotides; those with 5′UTRs ≥ 20 nucleotides but < 35 nucleotides; and those with 5′UTRs ≥ 35 nucleotides. Each part corresponds to a separate multivariate linear model fitted to the corresponding group of genes, with the intercept and feature coefficients allowed to differ between parts. Single-part regressions included all genes in the dataset. Models for the subset of genes that express only a single mRNA isoform (see Additional file 1: Figures S14 and S15) were produced by the same approach used for models employing all genes, though the APE boundaries were kept the same.

The coefficient of determination R^2 coefficient was based on the ordinary least squares (OLS) in all cases. The R^2 calculated in this way is equal to the square of the Pearson correlation between the observed response values and the predicted response values, and it measures the goodness of fit or the predictive power of a model.

Control by RNA folding (RNAfold)

To predict control of TR by RNA folding in the 5′ region, a series of features were developed using predicted Gibbs free energies generated by ViennaRNA RNAfold [16, 32] (Fig. 2). Most features were calculated in 21 variants based on one nucleotide offset sliding windows of lengths varying from 6 to 100 nucleotides (Fig. 2). For these window-based features, the 5′ regions used span from the 5′ cap to that part of the CDS that is covered by the window whose 5′ end maps to position − 1. The features are as follows:

Single window features

5′ cap, − 65, − 30, − 35, − 6, − 1: the free energy of the window whose 5′ nucleotide lies at the position specified.

min, 10%, 25%, 75%, 90%, max: the free energy of the window with the minimum, the 10th percentile, the 25th percentile, the 75th percentile, the 90th percentile, or the maximum energy of all windows in the 5′ region of the mRNA

Note that windows corresponding to different single-window features are allowed to overlap.

Multi window features

Mean: the mean free energy of windows of the specified length.

% ≤ 20%, % ≥ =90%, % ≥ 80%: the percent of windows in a 5′ region that have free energy ≤ 20th percentile, ≥ 90th percentile, or ≥ 80th percentile of the free energy of all windows of all genes in the dataset. Some genes will thus have scores of 0 as they lack any windows defined by these thresholds.

sum ≤ 5%, sum ≤ 10%, sum ≤ 20%, sum ≥ 80%, sum ≥ 90%: the sum of free energies defined by these thresholds out all windows of all genes in the datasets. Some genes will thus have scores of 0 as they lack any windows defined by these thresholds.

Whole 5′ region

Whole: free energy of a fold of the entire sequence from the 5′ cap to + 35

Feature sets

Feature sets fit multivariate linear models (possibly multipart, depending on the species) between $\log_{10}$ TR and multiple of the above features.

All-1: combines all features of a given window length

All: combines all features of a given window length and in addition includes the "whole" feature

RNAfold: uses forward selection with Bayesian Information Criterion (BIC) to identify an effective subset of features to predict $\log_{10}$ TR. The model is allowed to select different length windows for different features. In the case of *S. pombe*, prior to BIC selection, all windows with sequences that extend 3′ of + 30 were removed to avoid elements whose Gibbs energies have a strong negative correlation with TR. "RNAfold" is our most accurate model for control of TR by RNA structures (Fig. 2). Its BIC-selected features are listed in Additional file 5.

Control by sequence motifs (5′motifs)

Defining an optimum PWM for the iAUG proximal elements

Various length position weight matrices (PWMs) were calculated using the 10% of genes with the highest TR scores (Fig. 4a, high TR cohorts). The mRNA sequences of these sets of genes were aligned such that nucleotide + 1 corresponds to the A of the iAUG. The frequencies of A, U, C, and G at every position from − 100 to + 35 were calculated (Additional file 7). Sequences of all high TR genes were used, including those with 5′UTRs shorter than 100 nucleotides, these short 5′UTR genes contributing only to the frequencies at the 5′UTR positions they contained.

A series of PWMs were constructed that all contained position nucleotide − 1 and positions 5′ of that in five nucleotide steps to − 100, i.e., − 5 to − 1 PWM, − 10 to − 1 PWM, and − 15 to − 1 PWM. Variants of each of these PMWs were constructed that also included nucleotide + 4 and positions 3′ in 5 or 10 nucleotide steps within the protein coding sequence (CDS) to + 33, i.e., − 5 to + 13 PWM, − 5 to + 23 PWM, and − 10 to + 13 PWM. Given each PWM, say a PWM for m nucleotides from the 5′UTR and for n nucleotides of the CDS, we calculated the PWM scores of all the genes in a dataset as follows: score of gene $g = \log_{10} \prod_{\substack{i=-m \\ i\neq 0}}^{n} p\,(\text{nucleotide at position } i \text{ of gene } g)$.

We then calculated the R^2 correlation coefficient between the PWM scores based on each PWM boundary and the $\log_{10}$ TR values. The results are plotted in Fig. 4b. The locations of the approximately optimum APE PWMs are shown boxed (Fig. 4a).

An APE feature set

The APE was divided into the part 5′ of + 1 (uAPE) and the part 3′ of + 3 (dAPE). The optimum length PWMs defined above were divided into two (e.g., *S. cerevisiae* uAPE PWM − 35 to − 1 and dAPE PWM + 4 to + 28). All genes were scored with the dAPE PWM. In single regression species, all genes were also scored the corresponding uAPE. In *S. cerevisiae*, genes with 5′UTRs ≥ 35 nucleotides were scored with uAPE PWM − 35 to − 1, genes with 5′UTRs ≥ 20 nucleotides but < 35 nucleotides were scored with uAPE PWM − 20 to − 1, and genes with 5′UTRs < 20 nucleotides with uAPE PWM − 5 to − 1. In *Arabidopsis*, genes with 5′UTRs ≥ 65 nucleotides were scored with uAPE PWM − 65 to − 1, and genes with 5′UTRs < 65 nucleotides with uAPE PWM − 5 to − 1.

In addition, we calculate the 16 dinucleotide frequencies and 64 trinucleotide frequencies in the uAPE region for every gene as well as the 16 dinucleotide frequencies and 64 trinucleotide frequencies in the dAPE region for every gene. Then, we fit a multivariate linear model (possibly multipart, depending on the species) between the $\log_{10}$ TR (response) and all features for the uAPE (i.e., uAPE PWM + 16 dinucleotide frequencies + 64 trinucleotide frequencies) and separately all features for the dAPE and selected features by forward selection with

Bayesian Information Criterion. This results in a subset of features that form the uAPE feature set and the dAPE feature set (Additional file 1: Figure S7; Additional file 5). In addition, we combined the selected uAPE and dAPE feature sets to define the APE feature set (Additional file 1: Figure S7).

A 5′ofAPE feature set

A 5′cap PWM was defined using various length PWMs calculated using the 10% of genes with the highest TR scores in a procedure similar to that used to define the APE (Additional file 1: Figure S8a, high TR cohorts). All 5′cap PWMs share a common 5′ end at the 5′ cap site and differ only in length. The R^2 correlation coefficients between these PWM scores and $\log_{10}$ TR values show a local peak from 3 to 15 nucleotides from the 5′ cap (Additional file 1: Figure S8b). To simplify our model while taking into account the low R^2 values (< 1.2%), a 5′cap PWM was defined as the first 5 nucleotides of the transcript in all species.

In addition, we calculate the 16 dinucleotide frequencies and 64 trinucleotide frequencies in the region 5′ of the APE. Then, we fit a multivariate linear model (possibly multipart, depending on the species) between the $\log_{10}$ TR (response) and all features for this 5′ofAPE region (i.e., 5′cap PWM + 16 dinucleotide frequencies + 64 trinucleotide frequencies) and selected features by forward selection with Bayesian Information Criterion. The BIC-selected features for 5′ofAPE, are listed in Additional file 5, and their R^2 correlation coefficients vs TR are shown in Additional file 1: Figure S9.

A 5′motif feature set

A "5′motifs" feature set was defined by combining the 5′ofAPE, uAPE, and dAPE feature sets (Additional file 1: Figure S9).

Control by upstream open reading frames (uAUG)

For each gene, the number of AUGs upstream of the iAUG of the CDS was calculated. This feature is designated "uAUG," and its contribution to the prediction of $\log_{10}$ TR is due to the repression of CDS translation by translation of upstream open reading frames (Fig. 8; Additional file 1: Figure S13).

Control by CDS length (CDS length)

For each gene, the $\log_{10}$(number of CDS amino acids) was calculated. This feature is designated "CDS length," and its contribution to the prediction of $\log_{10}$ TR is due to the impact of CDS length on the recycling of ribosomal subunits (Fig. 8; Additional file 1: Figure S13).

Control by codon frequencies (codon)

For each gene, we calculate the frequencies of the 61 non-stop codon frequencies in its CDS nucleotide sequence. These 61 features together define the feature set "codon" (Fig. 8; Additional file 1: Figure S13).

Control by poly-A tail length

For each gene, we took the poly-A tail length defined by Subtelny et al. For all ten datasets, we find that poly-A length is weakly negatively correlated with translation rates. Prior evidence suggests that this negative correlation does not necessarily reflect a direct control of translation rates but is instead a result of two phenomena: the stabilization of mRNAs against degradation by shorter poly-A tails and the fact that translation rates are positively correlated with mRNA abundance [4, 73]. For this reason, we have not included poly-A length as a feature in our study. Note that there is a much stronger, positive correlation of poly-A tail length with translation rates in pregastrula embryos that likely reflects direct control of translation [4, 15].

Control by 5′UTR length

5′UTR length has been used as a feature to capture control by the 5′ regions in some prior studies [10, 13, 14, 16, 22, 74] (Additional file 1: Tables S1 and S2). We have found, however, that models that combine 5′UTR length with the "5′motifs," "uAUGs," and "RNAfold" features only explain an additional 0.03–0.62% of the variance in translation rates, compared with models that do not include 5′UTR length. This suggests that 5′UTR length is not a direct determinant of translation rates, but is instead largely a co-correlate of other sequence features that do directly affect translation. The fact that control by RNA folding is largely determined by most folded regions, not by less folded regions (Figs. 2 and 3a), further supports this conclusion. We, therefore, have not included 5′UTR length as a feature in our final models.

A general feature model using five features

To determine the variance in translation rates explained by "RNAfold," "5′motifs," "uAUG," "CDS length," and "codon," a multivariate model (possibly multipart, depending on the species) was used to regress $\log_{10}$ TR on all five feature sets together (Fig. 8; Additional file 1: Figure S12). The resulting R^2 correlation coefficient estimates the degree of translational control exerted by these general features.

Control by amino acid frequencies (AA) and synonymous codon preferences (syn.codon)

For each gene, we calculated the frequencies of the 20 amino acid frequencies in its CDS nucleotide sequence.

These 20 features together were defined to be the feature set "AA" (Fig. 11). Then, for each of the 61 nonstop codons, we defined a synonymous codon preference as its frequency divided by the sum of the frequencies of synonymous codons that code the same amino acid. For example, AAA, AAT, AAC, and AAG code the same amino acid, so we computed their synonymous codon preferences as AAA/(AAA + AAT + AAG + AAC), AAT/(AAA + AAT + AAG + AAC), AAC/(AAA + AAT + AAG + AAC), and AAG/(AAA + AAT + AAG + AAC), respectively. These 61 synonymous codon preferences together define the feature set "syn.codon" (Fig. 11).

Additional files

Additional file 1: Tables S1–S2. and **Figures S1–S16.** (PDF 3060 kb)

Additional file 2: Translation rate, mRNA sequence, and feature score datasets. (XLSX 29562 kb)

Additional file 3: Gibbs energies of dsRNA oligonucleotides from Xia et al. (XLSX 20 kb)

Additional file 4: RNA feature R^2 values from Fig. 3. (XLSX 81 kb)

Additional file 5: BIC-selected feature sets for "RNAfold" and "5′motifs" features. (XLSX 56 kb)

Additional file 6: RNA structures for "min" windows. (XLSX 5637 kb)

Additional file 7: APE and 5′cap PWMs. (XLSX 90 kb)

Additional file 8: Tri-nucleotide high TR/low TR ratios. (XLSX 35 kb)

Additional file 9: R^2 values for features and complete model of TR. (XLSX 65 kb)

Additional file 10: Comparison of Cupress vs Weinberg TR data. (XLSX 65 kb)

Additional file 11: Review history. (DOCX 36 kb)

Acknowledgements
We thank Soile Keranen and Xiaoyong Li for the thoughtful critiques of the drafts of this manuscript.

Review history
The review history is available as Additional file 11.

Authors' contributions
JL conducted all the statistical analyses. G-LC predicted the mRNA folding energies and secondary structures. MDB selected the datasets used. MDB, JL, and G-LC designed the analysis strategies. MDB and JL wrote the manuscript. All authors read and approved the final manuscript.

Funding
JJL's work was supported by the start-up fund of the Department of Statistics at the University of California, Los Angeles; a Hellman Fellowship from the Hellman Foundation; the PhRMA Foundation Research Starter Grant in Informatics; the Sloan Research Fellowship; the Johnson & Johnson WiSTEM2D Award; and the NIH/NIGMS grant R01GM120507. Work at Lawrence Berkeley National Laboratory was conducted under US Department of Energy Contract No. DE-AC02-05CH11231.

Availability of data and materials
The datasets supporting the conclusions of this article are included within Additional files 2, 3, and 10. Further, the complete code files for reproducing the results presented are available at Zenodo https://doi.org/10.5281/zenodo.3272133 [72]. These files include the R code, input data, processed files, and preliminary figures.

Ethics approval and consent to participate
None.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Statistics, Department of Biomathematics, and Department of Human Genetics, University of California, Los Angeles, CA 90095, USA. [2]Computational Biology Program, Public Health Sciences and Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA. [3]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94708, USA.

Received: 22 March 2019 Accepted: 11 July 2019
Published online: 09 August 2019

References
1. Hinnebusch AG. Translational regulation of GCN4 and the general amino acid control of yeast. Annu Rev Microbiol. 2005;59:407–50.
2. Gingold H, Pilpel Y. Determinants of translation efficiency and accuracy. Mol Syst Biol. 2011;7:481.
3. Svitkin YV, Yanagiya A, Karetnikov AE, Alain T, Fabian MR, Khoutorsky A, Perreault S, Topisirovic I, Sonenberg N. Control of translation and miRNA-dependent repression by a novel poly(A) binding protein, hnRNP-Q. PLoS Biol. 2013;11:e1001564.
4. Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. Poly(A)-tail lengths and a developmental switch in translational control. Nature. 2014;508:66–71.
5. Tuller T, Zur H. Multiple roles of the coding sequence 5′ end in gene expression regulation. Nucleic Acids Res. 2015;43:13–28.
6. Radhakrishnan A, Green R. Connections underlying translation and mRNA stability. J Mol Biol. 2016;428:3558–64.
7. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science*. 2016;352:1413–6.
8. Thompson MK, Gilbert WV. mRNA length-sensing in eukaryotic translation: reconsidering the "closed loop" and its implications for translational control. Curr Genet. 2016;63:613-20.
9. Fernades LD, de Moura APS, Ciandrini L. Gene length as a regulator for ribosome recruitment and protein synthesis: theoretical insights. Scientific Reports. 2017;7:17409.
10. Li JJ, Chew GL, Biggin MD. Quantitating translational control: mRNA abundance-dependent and independent contributions and the mRNA sequences that specify them. Nucleic Acids Res. 2017;45:11821–36.
11. Cottrell KA, Szczesny P, Djuranovic S. Translation efficiency is a determinant of the magnitude of miRNA-mediated repression. Sci Rep. 2017;7:14884.
12. Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. Rate-limiting steps in yeast protein translation. Cell. 2013;153:1589–601.
13. Weinberg D, Shah P, Eichhorn S, Hussmann J, Plotkin J, Bartel D. Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. Cell Reports. 2016;14:1787–99.
14. Janich P, Arpat AB, Castelo-Szekely V, Lopes M, Gatfield D. Ribosome profiling reveals the rhythmic liver translatome and circadian clock regulation by upstream open reading frames. Genome Res. 2015;25:1848–59.
15. Eichhorn SW, Subtelny AO, Kronja I, Kwasnieski JC, Orr-Weaver TL, Bartel DP. mRNA poly(A)-tail changes specified by deadenylation broadly reshape translation in Drosophila oocytes and early embryos. Elife. 2016;5:e16955.
16. Chew GL, Pauli A, Schier AF. Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. Nat Commun. 2016;7:11663.
17. Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, Segal E. Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. Proc Natl Acad Sci U S A. 2013;110:E2792–801.
18. Noderer WL, Flockhart RJ, Bhaduri A, Diaz de Arce AJ, Zhang J, Khavari PA, Wang CL. Quantitative analysis of mammalian translation initiation sites by FACS-seq. Mol Syst Biol. 2014;10:748.
19. Ben-Yehezkel T, Atar S, Zur H, Diament A, Goz E, Marx T, Cohen R, Dana A, Feldman A, Shapiro E, Tuller T. Rationally designed, heterologous *S. cerevisiae* transcripts expose novel expression determinants. RNA Biol. 2015; 12:972–84.

20. Sample PJ, Wang B, Reid DW, Presnyak V, McFadyen I, Morris DR, Seelig G. Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. BioRxiv. 2018. https://doi.org/10.1101/310375.
21. Cuperus JT, Groves B, Kuchina A, Rosenberg AB, Jojic N, Fields S, Seelig G. Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. Genome Res. 2017;27:2015–24.
22. Rojas-Duran MF, Gilbert WV. Alternative transcription start site selection leads to large differences in translation activity in yeast. RNA. 2012;18:2299–305.
23. Li GW, Burkhardt D, Gross C, Weissman JS. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell. 2014;157:624–35.
24. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature. 2010;466:835–40.
25. Liu MJ, Wu SH, Wu JF, Lin WD, Wu YC, Tsai TY, Tsai HL, Wu SH. Translational landscape of photomorphogenic Arabidopsis. Plant Cell. 2013;25:3699–710.
26. Duncan CDS, Mata J. Effects of cycloheximide on the interpretation of ribosome profiling experiments in Schizosaccharomyces pombe. Sci Rep. 2017;7:10331.
27. Hsu PY, Calviello L, Wu HL, Li FW, Rothfels CJ, Ohler U, Benfey PN. Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. Proc Natl Acad Sci U S A. 2016;113:E7126–35.
28. Castelo-Szekely V, Arpat AB, Janich P, Gatfield D. Translational contributions to tissue specificity in rhythmic and constitutive gene expression. Genome Biol. 2017;18:116.
29. Borer PN, Dengler B, Tinoco I Jr, Uhlenbeck OC. Stability of ribonucleic acid double-stranded helices. J Mol Biol. 1974;86:843–53.
30. Groebe DR, Uhlenbeck OC. Characterization of RNA hairpin loop stability. Nucleic Acids Res. 1988;16:11725–35.
31. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*. 1999;288:911–40.
32. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA package 2.0. Algorithms Mol Biol. 2011;6:26.
33. Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry. 1998;37:14719–35.
34. Kozak M. An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res. 1987;15:8125–48.
35. Gingold H, Tehler D, Christoffersen NR, Nielsen MM, Asmar F, Kooistra SM, Christophersen NS, Christensen LL, Borre M, Sorensen KD, et al. A dual program for translation regulation in cellular proliferation and differentiation. Cell. 2014;158:1281–92.
36. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell. 2010;141:344–54.
37. Novoa EM, Ribas de Pouplana L. Speeding with control: codon usage, tRNAs, and ribosomes. Trends Genet. 2012;28:574–81.
38. Blair JD, Hockemeyer D, Doudna JA, Bateup HS, Floor SN. Widespread translational remodeling during human neuronal differentiation. Cell Rep. 2017;21:2005–16.
39. Floor SN, Doudna JA. Tunable protein synthesis by transcript isoforms in human cells. Elife. 2016;5:e10921.
40. Gonzalez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. 2013;14:R70.
41. Reyes A, Huber W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. Nucleic Acids Res. 2018;46:582–92.
42. Zhang R, Calixto CPG, Marquez Y, Venhuizen P, Tzioutziou NA, Guo W, Spensley M, Entizne JC, Lewandowska D, Ten Have S, et al. A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. Nucleic Acids Res. 2017;45:5061–73.
43. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, Gilad Y. Genomic variation. Impact of regulatory variation from RNA to protein. Science. 2015;347:664–7.
44. Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR, Coller J. Codon optimality is a major determinant of mRNA stability. Cell. 2015;160:1111–24.
45. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42:D68–73.
46. Jones-Rhoades MW, Bartel DP, Bartel B. MicroRNAS and their regulatory roles in plants. Annu Rev Plant Biol. 2006;57:19–53.
47. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013;499:172–7.
48. A compendium of RNA-binding motifs for decoding gene regulation. http://cisbp-rna.ccbr.utoronto.ca. Accessed 24 Jan 19.
49. Cavener DR, Ray SC. Eukaryotic start and stop translation sites. Nucleic Acids Res. 1991;19:3185–92.
50. Nakagawa S, Niimura Y, Gojobori T, Tanaka H, Miura K. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. Nucleic Acids Res. 2008;36:861–71.
51. Pisarev AV, Kolupaeva VG, Pisareva VP, Merrick WC, Hellen CU, Pestova TV. Specific functional interactions of nucleotides at key -3 and +4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex. Genes Dev. 2006;20:624–36.
52. Seol Y, Skinner GM, Visscher K, Buhot A, Halperin A. Stretching of homopolymeric RNA reveals single-stranded helices and base-stacking. Phys Rev Lett. 2007;98:158103.
53. Sponer J, Sponer JE, Mladek A, Jurecka P, Banas P, Otyepka M. Nature and magnitude of aromatic base stacking in DNA and RNA: quantum chemistry, molecular mechanics, and experiment. Biopolymers. 2013;99:978–88.
54. Kacser H, Burns JA, Fell DA. The control of flux: 21 years on. Biochem Soc Trans. 1995;23:341–66.
55. Fell D. Understanding the control of metabolism. London: Portland Press; 1997.
56. Simms CL, Yan LL, Zaher HS. Ribosome collision is critical for quality control during no-go decay. *Mol Cell*. 2017;68:361–73 e365.
57. Shoemaker CJ, Green R. Translation drives mRNA quality control. Nat Struct Mol Biol. 2012;19:594–601.
58. Zarai Y, Margaliot M, Tuller T. On the ribosomal density that maximizes protein translation rate. *PLoS One*. 2016;11:e0166481.
59. Marygold SJ, Roote J, Reuter G, Lambertsson A, Ashburner M, Millburn GH, Harrison PM, Yu Z, Kenmochi N, Kaufman TC, et al. The ribosomal protein genes and Minute loci of *Drosophila melanogaster*. Genome Biol. 2007;8:R216.
60. Firczuk H, Kannambath S, Pahle J, Claydon A, Beynon R, Duncan J, Westerhoff H, Mendes P, McCarthy JE. An in vivo control map for the eukaryotic mRNA translation machinery. Mol Syst Biol. 2013;9:635.
61. Sinturel F, Gerber A, Mauvoisin D, Wang J, Gatfield D, Stubblefield JJ, Green CB, Gachon F, Schibler U. Diurnal oscillations in liver mass and cell size accompany ribosome assembly cycles. Cell. 2017;169:651–63 e614.
62. Princiotta MF, Finzi D, Qian SB, Gibbs J, Schuchmann S, Buttgereit F, Bennink JR, Yewdell JW. Quantitating protein synthesis, degradation, and endogenous antigen processing. Immunity. 2003;18:343–54.
63. von der Haar T. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. BMC Syst Biol. 2008;2:87.
64. Li JJ, Bickel PJ, Biggin MD. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*. 2014;2:e270.
65. Bloom-Ackermann Z, Navon S, Gingold H, Towers R, Pilpel Y, Dahan O. A comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool. *PLoS Genet*. 2014;10:e1004084.
66. Tunney R, McGlincy NJ, Graham ME, Naddaf N, Pachter L, Lareau LF. Accurate design of translational output by a neural network model of ribosome distribution. Nat Struct Mol Biol. 2018;25:577–82.
67. Li JJ, Biggin MD. Gene expression. Statistics requantitates the central dogma. Science. 2015;347:1066–7.
68. dos Reis M, Wernisch L. Estimating translational selection in eukaryotic genomes. Mol Biol Evol. 2009;26:451–61.
69. Lynch M. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A. 2007;104(Suppl 1):8597–604.
70. Dana A, Tuller T. Mean of the typical decoding rates: a new translation efficiency index based on the analysis of ribosome profiling data. G3 (Bethesda). 2014;5:73–80.
71. Johnstone TG, Bazzini AA, Giraldez AJ. Upstream ORFs are prevalent translational repressors in vertebrates. EMBO J. 2016;35:706–23.
72. Li JJ. Codes for general mRNA sequence features in five eukayotes. Zenodo. 2019; https://doi.org/10.5281/zenodo.3272133.

73. Lima SA, Chipman LB, Nicholson AL, Chen YH, Yee BA, Yeo GW, Coller J, Pasquinelli AE. Short poly(A) tails are a conserved feature of highly expressed genes. Nat Struct Mol Biol. 2017;24:1057–63.
74. Sen ND, Zhou F, Harris MS, Ingolia NT, Hinnebusch AG. eIF4B stimulates translation of long mRNAs with structured 5′ UTRs and low closed-loop potential but weak dependence on eIF4G. Proc Natl Acad Sci U S A. 2016; 113:10464–72.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC