Quantitating Translational Control: mRNA Abundance-dependent and Independent Contributions and the mRNA Sequences That Specify Them

Jingyi Jessica Li, Guo-Liang Chew, and Mark D. Biggin

Supplementary Data

Supplementary Figures

Supplementary Tables

Supplementary Methods

Supplementary References

Supplemental Figures

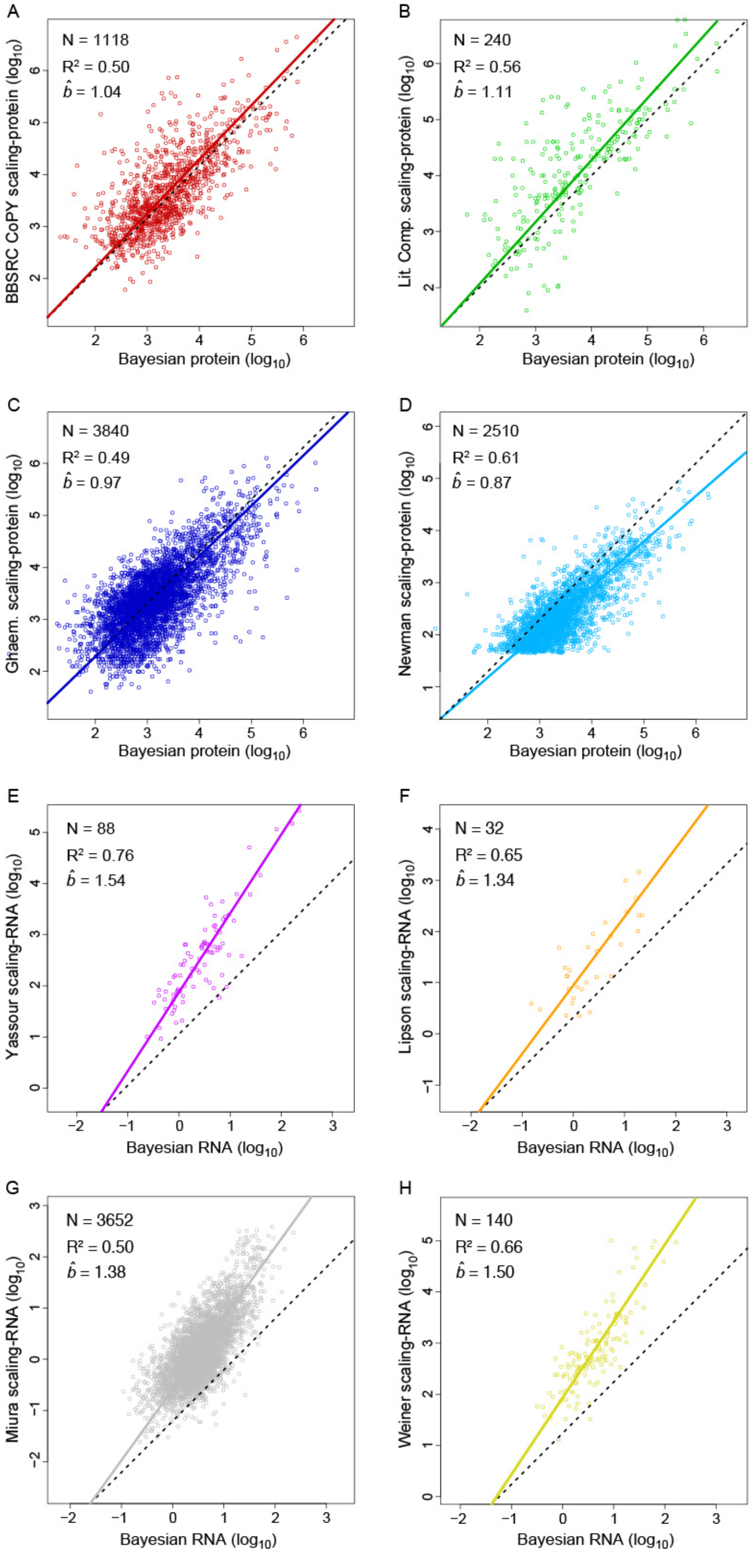

Figure S1. The relationships between the Bayesian model abundance values and scaling-standard data. (A-D) The protein abundances from Csardi et al.'s Bayesian model are compared to four scaling-standard protein datasets and (E-G) the mRNA data from Csardi et al.'s Bayesian model are compared to four scaling-standard RNA datasets. The colored lines show RuMA regressions. The dashed black lines show a slope of one, the case where the standard deviations of the x and y values are equal and thus what would be seen if the Bayesian model's abundance estimates were scaled identically to a scaling-standard. The number of genes in the intersection of the Bayesian and scaling-standard datasets (N), the coefficients of determination (R^2) and the slopes ($\hat{b}$) are indicated. The values are given in molecules per cell except for the Newman, Yassour, Lipson, Miura and Weiner data, which are in arbitrary units.

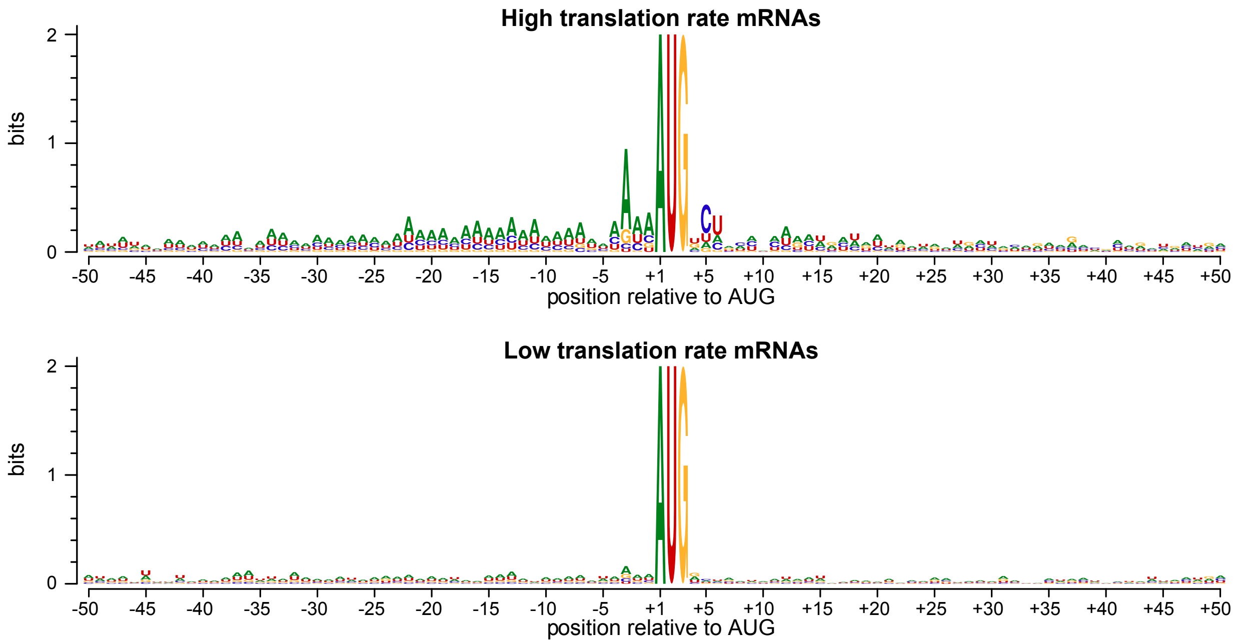

Figure S2. Sequence logos for high and low translation rate mRNAs. Logos are shown for the 10% of mRNAs with the highest TR scores (top) and the 10% of mRNAs with the lowest TR scores (bottom). The information content in bits is shown for each nucleotide at each position relative the first nucleotide of the protein coding sequence (CDS).

	Scaling-standard mRNA set (x-axis)											
Scaling-standard protein set (y-axis)	**Yassour NanoString**			**Lipson qPCR**			**Miura competitive PCR**			**Weiner NanoString**		
	N	$\mathbf{R^2_{sprot–sRNA}}$	$\hat{b}_{\mathbf{sprot–sRNA}}$	**N**	$\mathbf{R^2_{sprot–sRNA}}$	$\hat{b}_{\mathbf{sprot–sRNA}}$	**N**	$\mathbf{R^2_{sprot–sRNA}}$	$\hat{b}_{\mathbf{sprot–sRNA}}$	**N**	$\mathbf{R^2_{sprot–sRNA}}$	$\hat{b}_{\mathbf{sprot–sRNA}}$
BBSRC CoPY SRM Mass Spec.	27	0.43	0.99	6	0.78	0.62	788	0.39	1.19	41	0.38	1.26
Lit. compilation multiple methods	6	0.78	0.75	3	1.00	1.55	164	0.44	1.26	12	0.40	0.89
Ghaemmaghami GFP Western	57	0.52	1.01	14	0.40	1.05	2599	0.32	1.10	98	0.40	1.16
Newman GFP flow cytometry	34	0.61	0.90	11	0.54	0.78	1688	0.43	0.92	62	0.56	1.07

Table S1. The correlation between scaling-standard mRNA versus scaling-standard protein datasets. The number of genes in the intersection between the datasets (N); the coefficient of determination from OLS regression ($R^2_{\text{sprot–sRNA}}$), which is equal to the square of the Pearson correlation; and the RuMA estimate for the slope ($\hat{b}_{\text{sprot–sRNA}}$) are given. Log_{10} transformed data was used in every comparison. The abundance data are as described in the Materials and Methods and Dataset S3. For those cases where N ≥ 25, mean $\hat{b}_{\text{sprot–sRNA}} = 1.08$.

A.

Protein abundance dataset	**$R^2_{prot–PnD}$ (N)**
Bayesian model from multiple datasets	0.0007 (3,645)
BBSRC CoPY SRM Mass Spec.	0.0140 (1,001)
Ghaemmaghami GFP Western	0.0074 (2,970)
Newman GFP flow cytometry	0.0001 (2,121)

B.

RNA abundance dataset	**$R^2_{PnD–RNA}$ (N)**
Bayesian model from multiple datasets	0.0048 (3,645)
Weinberg RNA-seq	0.0002 (3,454)

Table S2. The correlation of protein degradation with the abundance of protein and mRNA. The fraction of each protein not degraded per cell was calculated as PnD = $e^{-k_{\mathrm{deg}} \times t}$ using protein degradation rate constants (k_{deg}) from Christiano et al. 2014 and their estimate for the cell-cycle time t = 150 min. (see Dataset S5). (A) The coefficients of determination (i.e. the squared Pearson correlation, $R^2_{prot–PnD}$) between the PnD data and four protein abundance datasets in logarithmic space. (B) The coefficients of determination ($R^2_{PnD–RNA}$) between the PnD values and two mRNA abundance datasets in logarithmic space. The number of data points in each comparison is given in brackets.

A.	**Scaling-standard mRNA dataset (y-axis)**									
Ribosome profiling study mRNA abundance data (x-axis)	**Yassour Nanostring**		**Lipson qPCR**		**Miura competitive PCR**		**Weiner Nanostring**		**Mean**	
	$\mathbf{R}^2_{\mathrm{sRNA-RNA}}$	$\hat{b}_{\mathrm{sRNA-RNA}}$	$\mathbf{R}^2_{\mathrm{sRNA-RNA}}$	$\hat{b}_{\mathrm{sRNA-RNA}}$	$\mathbf{R}^2_{\mathrm{sRNA-RNA}}$	$\hat{b}_{\mathrm{sRNA-RNA}}$	$\mathbf{R}^2_{\mathrm{sRNA-RNA}}$	$\hat{b}_{\mathrm{sRNA-RNA}}$	$\mathbf{R}^2_{\mathrm{sRNA-RNA}}$	$\hat{b}_{\mathrm{sRNA-RNA}}$
Weinberg IE	0.87	1.10	0.82	0.99	0.61	1.21	0.87	1.00	0.79	1.07
Csardi Median	0.81	1.02	0.76	0.98	0.62	1.20	0.85	1.05	0.76	1.06

B.	**Scaling-standard protein dataset (y-axis)**									
Ribosome profiling study Ribosome density data (x-axis)	**BBSRC CoPY SRM Mass Spec.**		**Lit. compilation Multiple methods**		**Ghaemmaghami GFP Western**		**Newman GFP flow cytometry**		**Mean**	
	$\mathbf{R}^2_{\mathrm{sprot-RD}}$	$\hat{b}_{\mathrm{sprot-RD}}$	$\mathbf{R}^2_{\mathrm{sprot-RD}}$	$\hat{b}_{\mathrm{sprot-RD}}$	$\mathbf{R}^2_{\mathrm{sprot-RD}}$	$\hat{b}_{\mathrm{sprot-RD}}$	$\mathbf{R}^2_{\mathrm{sprot-RD}}$	$\hat{b}_{\mathrm{sprot-RD}}$	$\mathbf{R}^2_{\mathrm{sprot-RD}}$	$\hat{b}_{\mathrm{sprot-RD}}$
Weinberg IE	0.60	1.02	0.61	1.06	0.48	0.98	0.64	0.85	0.58	0.98
Csardi Median	0.51	1.10	0.53	1.15	0.45	1.07	0.64	1.01	0.54	1.08

Table S3. The accuracy of two ribosome profiling datasets. (A) mRNA abundance data from two ribosome profiling studies are compared to four scaling-standard mRNA datasets. (B) Ribosome density data are compared to protein abundance data from four scaling-standard protein sets. The coefficient of determination (i.e. the square of Pearson correlation, R^2) and the RuMA estimate for the slope ($\hat{b}$) are given. The means of the R^2 and slope estimates are also shown. The results are for the intersection of all genes in each dataset for each pairwise comparison (Datasets S3 and S4). Log_{10} transformed data was used in every comparison.

Study	**Metrics for log$_{10}$(TR) (y-axis) vs. log$_{10}$(mRNA) (x-axis)**					**Squared correlation coefficients with protein abundance**			
	$\mathbf{R^2_{TR–RNA}}$[1]	${}^{\mathbf{OLS}}\hat{b}_{\mathbf{TR–RNA}}$[2]	${}^{\mathbf{RgMA}}\hat{b}_{\mathbf{TR–RNA}}$[3]	**%TR ≡ TR$_{\mathbf{mIND}}$**[4]	**%TR ≡ TR$_{\mathbf{mD}}$**[5]	$\mathbf{R^2_{prot–TR}}$[6]	$\mathbf{R^2_{prot–TRmIND}}$[7]	$\mathbf{R^2_{prot–TRmD}}$[8]	$\mathbf{R^2_{prot–RNA}}$[9]
Weinberg IE*	0.21	0.22	0.31	79%	21%	0.24	0.01	0.77	0.77
Csardi median	0.19	0.28	0.55	81%	19%	0.28	0.03	0.74	0.74

Table S4. Contributions of TR, TR$_{\mathbf{mD}}$ and TR$_{\mathbf{mIND}}$ to the Bayesian model's protein abundance data. The values are given for genes in the intersection of the ribosome profiling study specified (left) and the Bayesian model for protein abundance: top row based on 4,716 genes, bottom row based on 5,486 genes. The ribosome profiling data is from Dataset S4 and the protein abudance data from Dataset S2. [1-5] Metrics for the relationship between log$_{10}$(TR) (y-axis) and log$_{10}$(RNA abundance) (x-axis): [1] the coefficient of determination (i.e. the square of Pearson correlation, R^2); [2] the OLS estimate of the slope; [3] the RgMA estimate of the slope; [4] the percent of the variance in log$_{10}$(TR) that is explained by the variance in log$_{10}$(TR$_{mIND}$); [5] the percent of the variance in log$_{10}$(TR) that is explained by the variance in log$_{10}$(TR$_{mD}$). [6-9] The coefficient of determination (R^2) between the Bayesian model protein data and different datasets: [6] log$_{10}$(TR) ; [7] log$_{10}$(TR$_{mIND}$); [8] log$_{10}$(TR$_{mD}$); [9] mRNA abundance data from the specified ribosome profiling study. Note that the Weinberg and the Csardi Media ribosome profiling data strongly correlate with each other. For example, the R^2 coefficient between their estimates for translation rates is 0.71.

* The TR values in this row are Weinberg et al.'s elongation rate corrected Initiation Efficiency (IE) data. Four versions of the mRNA data were produced by correction using the four mRNA scaling-standards, and four versions of Ribosome Density data (inferred from IE data) were produced by correction using the four protein scaling-standards (Dataset S4). Sixteen corrected versions of IE, TR$_{mIND}$ and TR$_{mD}$ values were then determined using the corrected mRNA data and Ribosome Density data. The metrics presented are the means of ones derived from each of the sixteen comparisons of corrected data, where corrected IE data was always compared vs mRNA data corrected using the same mRNA scaling-standard.

True R^{2}_{y-x}	**Simulation results**			
	empirical R^{2}_{y-x}	${}^{\mathbf{RuMA}}\hat{b}_{\mathbf{y-x}}$	${}^{\mathbf{RgMA}}\hat{b}_{\mathbf{y-x}}$	${}^{\mathbf{OLS}}\hat{b}_{\mathbf{y-x}}$
1.00	1.00 [1.00 – 1.00]	1.00 [1.00 – 1.00]	1.00 [1.00 – 1.00]	1.00 [1.00 – 1.00]
0.20	0.20 [0.16 – 0.24]	1.00 [0.95 – 1.06]	1.01 [0.76 – 1.27]	0.45 [0.39 –0.50]
0.01	0.01 [0.00 – 0.03]	1.00 [0.94 –1.06]	1.31 [0.24 – 3.83]	0.10 [0.04 – 0.16]

Table S5. The impact of R^{2} on the estimates of slope *b* for different regressions. In 1,000 simulation runs, 1,000 x and y values were simulated from bivariate Gaussian distributions of (X, Y), both of which had means equal to zero and variances equal to 1. Three cohorts of 1,000 simulation runs were performed under three covariance settings where X and Y have positive Pearson correlation coefficients corresponding to R^{2} values of either 1.00, 0.20 or 0.01. The empirical values of R^{2} between x and y and the slopes defined by the RuMA, RgMA, and OLS regressions were calculated for each simulation run for each of the three R^{2} settings. For each setting (table rows), the mean values and the 95% quantile confidence intervals from these 1,000 simulations are presented in the four right-hand columns. The mean RuMA and OLS slopes behave as expected as R^{2} changes (2). The RgMA slope behaves similarly to the RuMA slope, but has broader 95% confidence intervals for lower correlation coefficients. Only the OLS slope down-weights the magnitude of the estimate of the true slope at low values of R^{2}.

A. BBSRC CoPY scaling-standard protein

Study[1]	**N**[2]	$\mathbf{R^2_{sprot–TR}}$[3]	$\mathbf{R^2_{sprot–TRmIND}}$[4]	$\mathbf{R^2_{sprot–RNA}}$[5]
Weinberg IE*	1,060	0.30	0.05	0.56
Csardi Median	1,115	0.23	0.03	0.50

B. Literature compilation scaling-standard protein

Study[1]	**N**[2]	$\mathbf{R^2_{sprot–TR}}$[3]	$\mathbf{R^2_{sprot–TRmIND}}$[4]	$\mathbf{R^2_{sprot–RNA}}$[5]
Weinberg IE*	218	0.37	0.07	0.55
Csardi Median	239	0.18	0.01	0.53

C. Ghaemmagham scaling-standard protein

Study[1]	**N**[2]	$\mathbf{R^2_{sprot–TR}}$[3]	$\mathbf{R^2_{sprot–TRmIND}}$[4]	$\mathbf{R^2_{sprot–RNA}}$[5]
Weinberg IE*	3,555	0.21	0.04	0.44
Csardi Median	3,788	0.20	0.04	0.43

D. Newman scaling-standard protein

Study[1]	**N**[2]	$\mathbf{R^2_{sprot–TR}}$[3]	$\mathbf{R^2_{sprot–TRmIND}}$[4]	$\mathbf{R^2_{sprot–RNA}}$[5]
Weinberg IE*	2,292	0.19	0.01	0.64
Csardi Median	2,476	0.22	0.05	0.60

E. Mean scaling-standard protein

Study[1]	**N**[2]	$\mathbf{R^2_{sprot–TR}}$[3]	$\mathbf{R^2_{sprot–TRmIND}}$[4]	$\mathbf{R^2_{sprot–RNA}}$[5]
Weinberg IE*	NA	0.27	0.04	0.55
Csardi Median	NA	0.21	0.03	0.51

Table S6. Contributions of TR and TR$_{\mathbf{mIND}}$ to scaling-standard protein abundance data. (A-D) Translation rates and mRNA abundance data (Dataset S4) are compared to the four scaling-standard protein datasets (Dataset S3). (E) The mean results from A-D. [1] the name and origin of the ribosome profiling study from which TR and mRNA data are taken; [2] the number of data points in the intersection between the ribosome profiling and scaling-protein data; [3] the coefficient of determination (i.e., the squared Pearson correlation, R^2) between scaling-protein and TR; [4] the coefficient of determination (R^2) between scaling-protein data and TR$_{mIND}$; [5] the coefficient of determination (R^2) between scaling-protein abundance and mRNA amounts. Log$_{10}$ transformed data was used in every comparison.

* Corrected versions of versions of IE and TR$_{mIND}$ values were determined as described in Supplementary Table S4.

Sequence feature[1]	$\mathbf{R^2}_{\textbf{feature–TR}}$[2]	$\mathbf{R^2}_{\textbf{feature–TRmD}}$[3]	$\mathbf{R\ ^2}_{\textbf{feature–TRmIND}}$[4]	***p*-value for** $\mathbf{H_0: R^2_{TRmD} = R^2_{TRmIND}}$ **vs.** $\mathbf{H_1: R^2_{TRmD} \neq R^2_{TRmIND}}$ [5]	**Bonferroni corrected *p*-value**[6]
5' UTR log length	0.050	0.023	0.045	0.011	0.099
5’ UTR ORFs	0.140	0.075	0.099	0..169	1.000
5' UTR folding engy.	0.192	0.085	0.119	0.038	0.342
-35/+28 TICE	0.331	0.260	0.158	0.000	0.000
CDS log length	0.316	0.071	0.296	0.000	0.000
CDS AA freq.	0.405	0.455	0.165	0.000	0.000
CDS codon freq.	0.596	0.779	0.260	0.000	0.000
CDS folding engy.	0.082	0.176	0.009	0.000	0.000
poly-A tail length	0.046	0.091	0.008	0.000	0.000

Table S7. Contributions of mRNA sequence features to TR, $\mathbf{TR_{mD}}$ and $\mathbf{TR_{mIND}}$. R^2 coefficients of determination were calculated for each of nine sequence features with $\log_{10}(TR)$, $\log_{10}(TR_{mD})$ or $\log_{10}(TR_{mIND})$ as the response variable for 2,450 genes using a three-part regression (Supplementary Methods S4; Datasets S4, S6 and S8). To test for a null hypothesis (H_0) that each feature has the same R^2 with $\log_{10}(TR_{mD})$ and $\log_{10}(TR_{mIND})$ with a two-sided alternative hypothesis, a permutation test with 3,000 runs was used to calculate *p*-values (Supplementary Methods S4). [1] The mRNA sequence feature; [2] the R^2 of each feature for predicting $\log_{10}(TR)$; [3] the R^2 of each feature for predicting $\log_{10}(TR_{mD})$; [4] the R^2 of each feature for predicting $\log_{10}(TR_{mIND})$; [5] the *p*-value for the null hypothesis that $R^2_{TRmD} = R^2_{TRmIND}$; [6] the Bonferroni correction of the *p*-value given in column 5, assuming nine tests. These values are plotted in Figs. 7 and 9. Details of the individual features combinded to create the -35/+28 TICE feature set are given in Dataset S8.

Sequence feature[1]	$\mathbf{R^2_{feature–RnD}}$[2]	$\mathbf{R^2_{feature–TRmD}}$[3]	$\mathbf{R^2_{feature–TRmD*}}$[4]	***p*-value** $\mathbf{H_0: R^2_{TRmD} = R^2_{TRmD*}}$ **vs.** $\mathbf{H_1: R^2_{TRmD} > R^2_{TRmD*}}$[5]	**Bonferroni corrected *p*-value**[6]
5' UTR log length	0.004	0.031	0.028	0.088	0.789
5' UTR ORFs	0.010	0.075	0.079	0.878	1.000
5' UTR folding engy.	0.006	0.084	0.078	0.048	0.435
-35/+28 TICE	0.039	0.259	0.245	0.007	0.066
CDS log length	0.006	0.093	0.106	1.000	1.000
CDS AA freq.	0.090	0.454	0.421	0.000	0.000
CDS Codon freq.	0.203	0.797	0.738	0.000	0.000
CDS folding engy.	0.035	0.178	0.149	0.000	0.000
poly-A tail length	0.016	0.087	0.073	0.000	0.000

Table S8. Correlation of mRNA sequence features with mRNA degradation rates. The fraction of RNA not degraded per cell cycle (RnD) was calculated from mRNA degradation data for 1,939 genes; and values of $\log_{10}(TR_{mD})$ were adjusted to remove the expected impact of RNA degredation by regressing it against $\log_{10}$(RnD) data to generate $\log_{10}(TR_{mD*})$ (Supplementary Methods S4; Dataset S6). R^2 coefficients of determination were calculated between each of nine sequence features and $\log_{10}$(RnD), $\log_{10}(TR_{mD})$ and $\log_{10}(TR_{mD*})$ using a three-part linear regression (Supplementary Methods S4; Datasets S4, S6–S8). To test for a null hypothesis (H_0) that each feature has the same R^2 with $\log_{10}(TR_{mD})$ and $\log_{10}(TR_{mD*})$ with a one-sided alternative hypothesis (H_1) that the R^2 with $\log_{10}(TR_{mD})$ is greater than that with $\log_{10}(TR_{mD*})$, a permutation test with 3,000 runs was used to calculate *p*-values (Supplementary Methods S4). [1] The mRNA sequence feature; [2] the R^2 of each feature for predicting $\log_{10}$(RnD); [3] the R^2 of each feature for predicting $\log_{10}(TR_{mD})$; [4] the R^2 of each feature for predicting $\log_{10}(TR_{mD*})$; [5] the *p*-value for the null hypothesis that $R^2_{TRmD} = R^2_{TRmD*}$ with a one-sided alternative hypothesis; and the Bonferroni correction of the *p*-value given in column 5, assuming nine tests. Only poly-A tail length and the folding energy of the CDS show a significantly lower correlation with $\log_{10}(TR_{mD*})$ than with $\log_{10}(TR_{mD})$, suggesting that only these two features significantly affect mRNA degradation rates.

Frequency in TR_{mD} or TR_{mIND}[1]	**R^2 top cohort freq–tRNA abnd.**[2]	**R^2 bottom cohort freq–tRNA abnd.**[3]	***p*-value** **H_0: $R^2_{Top} = R^2_{Bottom}$ vs.** **H_1: $R^2_{Top} > R^2_{Bottom}$**[4]	**Bonferroni corrected *p*-value**[5]
AA freq. TR_{mD}	0.661	0.395	0.027	0.107
codon freq. TR_{mD}	0.650	0.196	0.000	0.000
AA freq. TR_{mIND}	0.492	0.428	0.256	1.000
codon freq. TR_{mIND}	0.404	0.247	0.001	0.003

Table S9. Correlation of amino acid and codon frequencies with tRNA abundance. The frequencies of amino acids or codons were calculated for the 10% of genes with the highest TR_{mD} or TR_{mIND} values (top cohort) and separately for the 10% of genes with the lowest TR_{mD} or TR_{mIND} values (bottom cohort). The set of genes used are the 2,450 genes listed in Dataset S6. tRNA abundances were taken from Weinberg et al. 2016, Table S2 (tRNA genome copy number * wobble). A combined tRNA abundance was calculated for comparison to amino acid frequencies by summing the abundances for the codons for each amino acid. To test for a null hypothesis (H_0) that the coefficient of determination (R^2) between top cohort frequencies and tRNA abundance is the same as the R^2 between bottom cohort frequencies and tRNA abundance with a one-sided alternative hypothesis (H_1) that the former R^2 is greater than the latter R^2. The R package "cocor" was used to calculate *p*-values. [1] The amino acid or codon frequency examined in either TR_{mD} or TR_{mIND}; [2] the R^2 of each top cohort with tRNA abundance; [3] the R^2 of each bottom cohort with tRNA abundance; [4] the *p*-value for the null hypothesis that $R^2_{top} = R^2_{bottom}$ with a one-sided alternative hypothesis; and [5] the Bonferroni correction of the *p*-value given in column 4, assuming four tests.

Supplemental Methods

Note that equations numbered 12 and lower are described in the Materials and Methods in the main manuscript and equations 13 and above are presented here. All R code is provided in Dataset S10.

S1. Using scaling-standard sets to scale protein and mRNA abundances

We have scaling-standard protein and scaling-standard mRNA datasets that are presumed to be correctly scaled because the protocols that generated them used internal concentrations standards to correct any scaling bias. In addition, we have protein and mRNA abundance data from Csardi et al.'s Bayesian model whose scaling is in doubt. Adjusting the scaling of Csardi et al.'s Bayesian datasets to agree with those of the scaling-standard sets requires choice of an appropriate regression. We consider two: the Ordinary Least Squares (OLS) and the Reduced Major Axis (RuMA). As an example, we explore how to re-center and re-scale the Bayesian mRNA data to match that of scaling-standard mRNA data. Let

x = log(Bayesian mRNA abundance)
y = log(Bayesian protein abundance)
z = log(scaling-mRNA abundance)

Please note that x, y and z are vectors corresponding to measurements on the same set of genes and thus have the same dimensions. Our goal is to adjust x by a linear transformation (i.e., re-centering and re-scaling) so that the slope between y and the adjusted version of x approaches the slope between y and z. We denote the measurement errors in x, y and z as e_X, e_Y and e_Z, and the unobservable error-free measurements of x, y and z as x^{true}, y^{true} and z^{true}. Then $x = x^{true} + e_X$, $y = y^{true} + e_Y$, and $z = z^{true} + e_Z$. We assume the error terms have mean zero. We can write

$$y = b_{Y\text{–}Z} \bullet z + a + e \,, \tag{13}$$

where the slope is $b_{Y\text{–}Z}$ and e represents the translation and protein degradation effect independent of mRNA, and

$$z = b_{Z\text{–}X} \bullet x + c \,, \tag{14}$$

where the slope is $b_{Z\text{–}X}$, the intercept is c, and equation 14 is the linear function that does re-centering and re-scaling to adjust x. Therefore,

$$y = b_{Y\text{–}Z} \bullet b_{Z\text{–}X} \bullet x + (b_{Y\text{–}Z} \bullet c + a) + e \,. \tag{15}$$

And the true slope between y and x is thus

$$b_{Y\text{–}X} = b_{Y\text{–}Z} \bullet b_{Z\text{–}X} \,.$$

We considered two types of regressions for estimating the value of $b_{Y\text{–}Z} \bullet b_{Z\text{–}X}$, finding the RuMA regression described in the second option to be the most appropriate.

Option 1: Estimating $b_{Y\text{–}Z} \bullet b_{Z\text{–}X}$ by OLS regression

The OLS regression assumes that all, or at least >75% of, measurement error is in the response variable placed on the y axis (1). This regression might be appropriate for our needs if either the scaling-standard datasets or the Csardi et al. Bayesian data have considerably less error than the other. The OLS estimate of slope b for a linear model $y = b \bullet x + c$ + error is

$$^{OLS}\hat{b} = R_{y\text{-}x} \bullet sd(y) / sd(x) \,,$$

where sd is the sample standard deviation and $R_{y\text{-}x}$ is the sample Pearson correlation coefficient between measurements of x and measurements of y (1). First we assumed that the Bayesian mRNA data had the lower error than the scaling-standard datasets (i.e., $var(e_X) << var(e_Z)$), which entails determining the OLS estimates for slopes of equations 13 and 14 as

$$^{OLS}\hat{b}_{Y\text{–}Z} = R_{Y\text{–}Z} \bullet sd(y) / sd(z) \,, \tag{16}$$

$$^{OLS}\hat{b}_{Z\text{–}X} = R_{Z\text{–}X} \bullet sd(z) / sd(x) \,.$$

Then

$$^{\mathrm{OLS}}\hat{b}_{\mathrm{Y-Z}} \bullet {}^{\mathrm{OLS}}\hat{b}_{\mathrm{Z-X}} = \mathrm{R}_{\mathrm{Y-Z}} \bullet \mathrm{R}_{\mathrm{Z-X}} \bullet \mathrm{sd(y)} / \mathrm{sd(x)} .$$

However, if we use the OLS regression to estimate the slope $b_{\mathrm{Y-X}}$ in (xv), we get

$$^{\mathrm{OLS}}\hat{b}_{\mathrm{Y-X}} = \mathrm{R}_{\mathrm{Y-X}} \bullet \mathrm{sd(y)} / \mathrm{sd(x)} ,$$

which is not equal to $({}^{\mathrm{OLS}}\hat{b}_{\mathrm{Y-Z}} \bullet {}^{\mathrm{OLS}}\hat{b}_{\mathrm{Z-X}})$ because $\mathrm{R}_{\mathrm{Y-X}} \neq \mathrm{R}_{\mathrm{Y-Z}} \bullet \mathrm{R}_{\mathrm{Z-X}}$. Therefore, the OLS regression used in this way is not appropriate for correcting improperly scaled datasets.

We next tested the OLS regression for the case where the scaling-standards have much lower error than the Bayesian abundance data, i.e., $\mathrm{var}(e_{\mathrm{Z}}) << \mathrm{var}(e_{\mathrm{X}})$. For this the OLS regression of the reversed linear relationship is required, that is

$$\mathrm{x} = (1 / b_{\mathrm{Z-X}}) \bullet \mathrm{z} - (c / b_{\mathrm{Z-X}}) .$$

The OLS estimate of this slope is

$$^{\mathrm{OLS}}\left(\widehat{1/b_{\mathrm{z-x}}}\right) = \mathrm{R}_{\mathrm{X-Z}} \bullet \mathrm{sd(x)} / \mathrm{sd(z)} ,$$

$$^{\mathrm{OLS}}\left(\widehat{1/b_{\mathrm{z-x}}}\right) = \mathrm{R}_{\mathrm{X-Z}} \bullet \mathrm{sd(x)} / \mathrm{sd(z)} .$$

Then we invert the OLS estimate $\left(\widehat{1/b_{\mathrm{z-x}}}\right)$ to get an estimate of $b_{\mathrm{Z-X}}$, that is

$$^{\text{inv-OLS}}\hat{b}_{\mathrm{Z-X}} = 1 / \mathrm{R}_{\mathrm{X-Z}} \bullet \mathrm{sd(z)} / \mathrm{sd(x)} .$$

Multiplied with the estimate from (xv) of ${}^{\mathrm{OLS}}\hat{b}_{\mathrm{Y-Z}} = \mathrm{R}_{\mathrm{Y-Z}} \bullet \mathrm{sd(y)} / \mathrm{sd(z)}$, we get

$$^{\mathrm{OLS}}\hat{b}_{\mathrm{Y-Z}} \bullet {}^{\text{inv-OLS}}\hat{b}_{\mathrm{Z-X}} = \mathrm{R}_{\mathrm{Y-Z}} / \mathrm{R}_{\mathrm{X-Z}} \bullet \mathrm{sd(y)} / \mathrm{sd(x)} .$$

This, however, is not equal to the OLS estimate of $b_{\mathrm{Y-X}} = b_{\mathrm{Y-Z}} \bullet \mathrm{R}_{\mathrm{Z-X}}$ in equation 14, i.e.

$$^{\mathrm{OLS}}\hat{b}_{\mathrm{Y-X}} = \mathrm{R}_{\mathrm{Y-X}} \bullet \mathrm{sd(y)} / \mathrm{sd(x)} ,$$

because $\mathrm{R}_{\mathrm{Y-Z}} / \mathrm{R}_{\mathrm{X-Z}} \neq \mathrm{R}_{\mathrm{Y-X}}$. Therefore, the OLS regression cannot be used to correct the scaling of the Bayesian mRNA data based on the scaling-standard datasets. This is true whether we assume that the bulk of error is in the Bayesian mRNA data or in the scaling-mRNA data.

Option 2: estimating $b_{\mathrm{Y-Z}} \bullet b_{\mathrm{Z-X}}$ with the RuMA regression

The RuMA regression assumes relatively equal measurement error in the two variables (1). This is plausible in our case. The Bayesian abundance data appear to be more accurate because the protein and mRNA abundances correlate more highly with each other and with independent datasets than do the scaling-standard sets. On the other hand, the scaling sets are intrinsically more accurately scaled because they were derived using methods that employ internal controls for molecular concentration. These different classes of error could approximate to the same magnitude in the two sets. More to the point, the empirical evidence from the use of internal controls indicates that the scaling-standard sets are better scaled, which implies that we should adjust the Bayesian abundance data to have similar mean and standard deviation values as the scaling-standards.

The standard equation for the RuMA estimate of slope b for a linear model $\mathrm{y} = b \bullet \mathrm{x} + c + \mathrm{error}$ is

$$^{\mathrm{RuMA}}\hat{b} = \mathrm{sign}(\mathrm{R}_{\mathrm{y\text{-}x}}) \bullet \mathrm{sd(y)} / \mathrm{sd(x)} ,$$

where $\mathrm{sign}(\mathrm{R}_{\mathrm{y\text{-}x}})$ is 1 if x and y are positively correlated, 0 if x and y are uncorrelated, and -1 if x and y are negatively correlated (1).

Since the scaling-standards and Bayesian abundance data are positively correlated, the RuMA regression gives a direct measure of the relative scaling of data as judged by the standard deviation and seem ideal for our purposes. The RuMA estimates for the slopes of equations 13 and 14 are

$$^{\mathrm{RuMA}}\hat{b}_{\mathrm{Y-Z}} = \mathrm{sd(y)} / \mathrm{sd(z)} ,$$

$$^{\mathrm{RuMA}}\hat{b}_{Z\text{–}X} = \mathrm{sd}(z) / \mathrm{sd}(x) .$$

Then

$$^{\mathrm{RuMA}}\hat{b}_{Y\text{–}Z} \bullet {}^{\mathrm{RuMA}}\hat{b}_{Z\text{–}X} = \mathrm{sd}(y) / \mathrm{sd}(x) .$$

If we also use the RuMA regression to estimate the slope $b_{Y\text{–}X}$ in equation 15, we will get

$^{\mathrm{RuMA}}\hat{b}_{Y\text{–}X} = \mathrm{sd}(y) / \mathrm{sd}(x)$, which is equal to $^{\mathrm{RuMA}}\hat{b}_{Y\text{–}Z} \bullet {}^{\mathrm{RuMA}}\hat{b}_{Z\text{–}X}$.

So the slopes estimated by the RuMA satisfy the desired relationship for the true slopes: i.e. that

$$b_{Y\text{–}X} = b_{Y\text{–}Z} \bullet b_{Z\text{–}X} .$$

Conclusion

To require $b_{Y\text{–}X} = b_{Y\text{–}Z} \bullet b_{Z\text{–}X}$, the slopes for all three relationships in equations 13 – 15 must be estimated by the RuMA regression and the values of x must be adjusted based on the RuMA estimate for the slope $b_{Z\text{-}X}$ and c in equation 14. Hence, the adjusted values of x can be obtained from

$$x_{\mathrm{adj}} = {}^{\mathrm{RuMA}}\hat{b}_{Z\text{–}X} \bullet x + {}^{\mathrm{RuMA}}\hat{c} = \mathrm{sd}(z) / \mathrm{sd}(x) \bullet x + {}^{\mathrm{RuMA}}\hat{c} .$$

If we use the RuMA to regress y on x_{adj}, the slope is

$$\mathrm{sd}(y) / \mathrm{sd}(x_{\mathrm{adj}}) = \mathrm{sd}(y) / ({}^{\mathrm{RuMA}}\hat{b}_{Z\text{–}X} \bullet \mathrm{sd}(x)) = \mathrm{sd}(y) / (\mathrm{sd}(z) / \mathrm{sd}(x) \bullet \mathrm{sd}(x)) = \mathrm{sd}(y) / \mathrm{sd}(z) ,$$

In other words

$$^{\mathrm{RuMA}}\hat{b}_{Y-x_{\mathrm{adj}}} = {}^{\mathrm{RuMA}}\hat{b}_{Y\text{–}Z} .$$

The essential reason for this phenomenon is that we require the slope of y vs. x_{adj} to be the same as the slope between y vs. z. Since x_{adj} depends on the slope between z and x, the slope of (y vs. $b_{Z\text{–}X} \bullet x$) must be equal to $b_{Y\text{–}Z}$. That is, we need

$$b_{Y\text{–}X} = b_{Y\text{–}Z} \bullet b_{Z\text{–}X} .$$

Since the Pearson correlation coefficient does not have this transitivity, i.e. $R_{Y\text{–}X} \neq R_{Y\text{–}Z} \bullet R_{Z\text{–}X}$, the slope estimates cannot involve the Pearson correlation coefficient. Thus, approaches such as the OLS method which involves the Pearson correlation cannot be used.

Please note that this RuMA-based adjustment of x only depends on z, not y. Therefore, regardless of y being the original log protein abundance or the adjusted log protein abundance, the linear transformation of x remains the same.

We also applied this RuMA-based adjustment to correct protein abundance. By defining x = log(Bayesian protein abundance), y = log(Bayesian mRNA abundance), and z = log(scaling-protein abundance), the above arguments still apply. We only need to change the equations 13 – 15 as the following equations 13' – 15' to keep the log-transformed protein abundance as the response and the log-transformed mRNA as the explanatory variable:

$$x = b_{X\text{–}Y} \bullet y + a + e , \qquad (13')$$

where the slope is $b_{X\text{–}Y}$ and e represents the translation and protein degradation effect independent of mRNA, and

$$z = b_{Z\text{–}X} \bullet x + c , \qquad (14')$$

where the slope is $b_{Z\text{–}X}$, the intercept is c, and equation 13' is the linear function that does re-centering and re-scaling to adjust x. Therefore,

$$z = b_{X\text{–}Y} \bullet b_{Z\text{–}X} \bullet y + (b_{Y\text{–}Z} \bullet a + c) + b_{Y\text{–}Z} \bullet e . \qquad (15')$$

And the true slope between x and y is thus

$$b_{Z\text{–}Y} = b_{X\text{–}Y} \bullet b_{Z\text{–}X} .$$

By similar arguments following equation 15 about the choice of RuMA regression for adjusting mRNA abundances, we also use the RuMA regression to adjust protein abundances. Also similarly, the adjustment of the $\log_{10}$ transformed protein abundance only depends on the scaling-standard log protein abundance, not the log mRNA abundance. In our main text, we have adjusted both log(Bayesian mRNA abundance) and log(Bayesian protein abundance) based on their corresponding scaling-standards, using the steps described in the subsection “Rescaling datasets” below.

Rescaling datasets

We adjusted and rescaled the Bayesian protein and mRNA abundance data and the Weinberg ribosome density and mRNA abundance data x (on the $\log_{10}$ scale) by the corresponding scaling-standards z (on the $\log_{10}$ scale) with the following steps. Bayesian protein data was scaled using each of the protein scaling standards, Bayesian RNA data using the RNA scaling standards, Weinberg ribosome density data using the protein scaling standards, and Weinberg mRNA abundance using the RNA scaling data. Since each dataset and its scaling-standards may not have exactly the same set of genes, we found their common genes and denoted the corresponding measurements in the dataset and the scaling-standards (on the $\log_{10}$ scale) as x* and z* respectively.

1. Fit the RuMA regression on z* vs. x* to estimate the linear relationship (equation 14), resulting in estimates for the slope and intercept:
   ${}^{\mathrm{RuMA}}\hat{b}_{Z\text{–}X} = \mathrm{sd}(z^*) / \mathrm{sd}(x^*)$,
   ${}^{\mathrm{RuMA}}\hat{c}_{Z\text{–}X} = \mathrm{mean}(z^*) - ({}^{\mathrm{RuMA}}\hat{b}_{Z\text{–}X} \bullet \mathrm{mean}(x^*))$,
   where mean(x*) and mean(z*) represent the mean of the original (improperly scaled) abundance and the mean scaling-standard abundance of the common genes, respectively.
2. Adjust x (original abundance of all genes) as
   $x_{adj} = {}^{\mathrm{RuMA}}\hat{b}_{Z\text{–}X} \bullet x + {}^{\mathrm{RuMA}}\hat{c}_{Z\text{–}X}$
3. Rescale x_{adj} such that it has the same total number of molecules as the original x, i.e.
   $x_{adj\text{-}res} = \log_{10}(10^{Xadj} \bullet \mathrm{sum}(10^{X}) / \mathrm{sum}(10^{Xadj}))$,
   where $\mathrm{sum}(10^{X})$ and $\mathrm{sum}(10^{Xadj})$ refer to the total number of molecules in the original x and the adjusted x, respectively.

Note that Weinberg et al.’s ribosome density was calculated from their Translation Initiation Efficiency (IE) values as RD’ = IE • mRNA (Dataset S4) and was corrected using each of the four protein scaling standards.

Calculating $b_{prot\text{–}RNA}$ from corrected abundance data

We have four protein scaling standards and four RNA protein standards (Dataset S3). For the Bayesian model abundances, we thus derive four corrected versions of the protein abundances and four corrected versions of the mRNA abundances (Dataset S2) and 16 pair wise combinations of the corrected protein vs corrected mRNA datasets. $b_{\mathrm{prot–RNA}}$ was determined for each of the 16 corrected pairs using either the RuMA, RgMA or OLS regressions. The mean of the 16 regressions is our estimate for $b_{\mathrm{prot–RNA}}$ for a given method.

Calculating $b_{TR\text{–}RNA}$ from corrected ribosome profiling data.

We have four protein scaling standards and four RNA protein standards (Dataset S3). For Weinberg et al.’s ribosome profiling data, we thus derive four corrected versions of ribosome density and four corrected versions of mRNA abundance data (Dataset S4). Sixteen versions of corrected TR were calculated using each pair wise combination of the corrected ribosome density data and mRNA abundance data as $TR_{adj}=RD'_{adj}/mRNA_{adj}$. $b_{\mathrm{TR–RNA}}$ was determined for each of the 16 corrected versions of TR vs mRNA abundance using either the RgMA or OLS regressions. In each case, the corrected version of TR was regressed against the corrected mRNA abundance data used to calculate it, a total of 16 regressions per method. The mean of the 16 regressions is our estimate for $b_{\mathrm{TR-RNA}}$ for a given method.

Note that the Csardi median ribosome data was analyzed without correction. Instead $b_{\mathrm{TR-RNA}}$ was determined from regressions of the published TR and mRNA abundance values.

Bootstrapping to estimate the confidence limits of regression lines
We estimate the confidence limits of regressions between our rescaled protein and mRNA abundance data using the following approach.

1. Bootstrap the 5,854 genes in the Bayesian mRNA and protein abundance datasets 1,000 times. In each bootstrap run, adjust and rescale the randomly sampled (with replacement) 5,854 mRNA and protein values with the mRNA and protein-scaling standards. Since there are four mRNA scaling-standards and four protein scaling-standards, we obtain four versions of rescaled mRNA values and four versions of rescaled protein values. For each rescaled mRNA and protein pair, we estimate the slope and intercept from both RuMA and RgMA regressions. By averaging the sixteen slope estimates and the sixteen intercept estimates from the sixteen pair wise regressions, we obtain estimates for the mean slope and the mean intercept.
2. This procedure returns 1,000 (RuMA/RgMA) mean slope estimates and 1,000 mean intercept estimates. The 2.5% and 97.5% percentiles are used to construct the 95% confidence limits as follows
   Limit 1: slope = 2.5% percentile of 1,000 bootstrapped mean slope estimates;
   intercept = 97.5% percentile of 1,000 bootstrapped mean intercept estimates
   Limit 2: slope = 97.5% percentile of 1,000 bootstrapped mean slope estimates;
   intercept = 2.5% percentile of 1,000 bootstrapped mean intercept estimates

The same bootstrap approach was used to estimate the 95% confidence interval of the OLS slope between the corrected Weinberg translation rates and corrected mRNA abundance using the same set of 16 pairs of corrected TR and mRNA abundance data used to estimate $b_{\mathrm{TR-RNA}}$. Based on this, we predict the slope of protein abundance vs. mRNA abundance to be 1.22 with a 95% confidence interval [1.13, 1.29] (Figure 3C). To ensure that this predicted linear relationship satisfies the same total number of protein and mRNA molecules as in the Bayesian model abundances, we estimate the intercept and its 95% confidence interval by the following procedure, which generates the regression line and 95% confidence limits in Figure 3C.

For the regression line with our predicted slope 1.17 in Figure 3B, we estimate the intercept by solving the following equation:

$$\mathrm{sum}(10^{1.22\mathrm{Xadj\text{-}res} + \mathrm{intercept}}) = \mathrm{sum}(10^{1.17\mathrm{Xadj\text{-}res} + 2.69}),$$

where 2.69 is the estimated intercept in Figure 3B and the unknown intercept that in Figure 3C. The above equation ensures that the estimated linear relationship in Figure 3B (based on the rescaled protein and mRNA measurements) and the predicted linear relationship in Figure 3C (based on translation and mRNA measurements) will predict protein levels with the same total number of molecules given the same mRNA levels.

Calculating the Coefficient of Determination (R^2)
We calculate the R^2 based on the Ordinary Least Squares (OLS) in all cases. The R^2 calculated in this way is the square of the Pearson correlation between two variables, is a symmetric similarity measure, and measures the strength of the linear relationship between the two variables. In addition, it can be compared to other published values for R^2 between protein abundance ($\log_{10}$) and mRNA abundance ($\log_{10}$), which are all based on the OLS.

S2. Alternative approaches to estimate $\widehat{b}_{\text{prot–RNA}}$ using scaling-standards

We tested two additional approaches to estimate $\widehat{b}_{\text{prot–RNA}}$ using our protein and mRNA abundance scaling-standards.

a. The Ordinary Least Squares (OLS) estimate for $\widehat{b}_{prot–RNA}$ from scaling-standard corrected Bayesian model abundances.

In the main text we report the RuMA and RgMA slopes between the scaling-standard corrected versions of the Bayesian model mRNA and protein abundances. These two regression give similar estimates for the slope (mean $\widehat{b}_{prot–RNA}$ = 1.17 and 1.16 respectively). Figure 3B shows the RuMA slope graphically. Here we additionally consider the widely used Ordinary Least Squares (OLS) regression, again between the scaling-standard corrected versions of the Bayesian model mRNA and protein abundances. OLS $\widehat{b}_{\text{prot–RNA}}$ = 1.08 with a 95% quantile confidence interval [1.02, 1.17]. The OLS regression assumes that the fraction of the variance in protein abundance data that is not explained by mRNA data (i.e. 1- $R^2_{\text{prot–RNA}}$) is largely due to translation and protein degradation rather than measurement error in the protein and mRNA data (1). Whereas the RgMA and RuMA regressions assume that measurement errors in both protein and mRNA values are the dominant explanation (1). In the Results, we show that the majority of variance in the Bayesian protein abundance data that is not explained by Bayesian mRNA data is probably due to measurement error (Figure 6, compare A and B). For this reason, we take the RuMA and RgMA estimates to be the better approximations.

b. Determining $b_{prot–RNA}$ directly from scaling-standard data only

As an alternate approach to determine $b_{\text{prot–RNA}}$ from the scaling-standards, we directly compared scaling-standard protein to scaling-standard mRNA data. Pair wise comparisons between scaling-standard sets are much less reliable than between Bayesian data and scaling-standards because many of the scaling sets include far fewer genes than the Bayesian model (Datasets S2–3 and Supplementary Figure S1). Nonetheless, for comparisons between scaling-protein and scaling-RNA data that include >50 genes, the RuMA estimates of $\widehat{b}_{\text{sprot–gRNA}}$ = 0.92–1.26 with a mean of 1.10 (Supplementary Table S1), which supports our conclusion that $\widehat{b}_{prot–RNA} << 1.69$.

S3. Determining the appropriate regression for estimating *slope* $b_{\text{TR–RNA}}$

We considered three different regressions for estimating the value of the slope $b_{\text{TR–RNA}}$, finding that the Ordinary Least Squares (OLS) regression described in (a) is the most appropriate.

a. Estimating slope $b_{\text{TR–RNA}}$ with the Ordinary Least Squares (OLS) regression

The OLS estimate of the true value of the slope b for a linear model y = b • x + c + error is given by the equation

$$^{\text{OLS}}\widehat{b} = R_{\text{y-x}} \bullet \text{sd(y) / sd(x)} ,$$

where $R_{\text{y-x}}$ is the Pearson correlation coefficient between x and y (1). Therefore,

$$^{\text{OLS}}\widehat{b}_{\text{TR–RNA}} = R_{\text{TR–RNA}} \bullet \text{sd(log(TR)) / sd(log(RNA))} \qquad (17)$$

where $R_{\text{TR–RNA}}$ is the Pearson correlation coefficient between log-transformed TR and log-transformed RNA.

Combining (equations 11 and 17) we see that the OLS regression assumes that

$$\text{sd(log(TR}_{\text{mD}}\text{))} = R_{\text{TR–RNA}} \bullet \text{sd(log(TR))}$$

which agrees with our calculation of log(TR_{mD}) as the fitted part of regressing log(TR) on log(RNA).

b. Estimating slope $b_{\text{TR–RNA}}$ with the Reduced Major Axis (RuMA) regression

The RuMA estimate of slope b for a linear model y = b • x + c + error is given by the standard equation (1)

$$^{\text{RuMA}}\widehat{b} = \text{sd(y) / sd(x)}$$

Therefore, given that $R_{TR-RNA} > 0$,

$${}^{RuMA}\hat{b}_{TR-RNA} = sd(\ \log(TR)\ ) / sd(\ \log(RNA)\ ) \qquad (18)$$

Combining (xi) and (xviii) we see that the RuMA regression assumes that

$$sd(\ \log(TR)\ ) = sd(\ \log(TR_{mD})\ ),$$

which is only true in the idealized situation where log-transformed translation rates correlate perfectly with log-transformed mRNA levels. In practice, $R_{TR-RNA} <1$, which the OLS regression takes into account but the RuMA regression does not.

c. Estimating slope b_{TR-RNA} with the Ranged Major Axis (RgMA) regression

The RgMA estimate of slope b for a linear model y = b • x + c + error cannot be expressed by a simple equation. We used a simulation to show that, like the RuMA regression, its estimate for slope b does not scale with R_{TR-RNA} (Supplementary Table S5). In 1,000 simulation runs, 1,000 x and y values were simulated from bivariate Gaussian distributions of (X, Y), both of which had means equal to zero and variances equal to 1. Three cohorts of 1,000 simulation runs were performed under three covariance settings where X and Y have positive Pearson correlation coefficients corresponding to R^2 values of either 1.00, 0.20 or 0.01. The empirical values of R^2 between x and y and the slopes defined by the RuMA, RgMA, and OLS regressions were calculated for each simulation run for each of the three R^2 settings. The mean RuMA and OLS slopes behave as expected as R^2 changes (1). The RgMA slope behaves similarly to the RuMA slope, but has broader 95% confidence intervals at lower correlation coefficients. Only the OLS slope down-weights the magnitude of the estimate of the true slope at low values for R^2.

S4. The mRNA sequence features that explain TR, TR_{mD} and TR_{mIND}

We tested a range of published and in-house derived mRNA sequence features that predict translation rates (TR) to determine an optimum set. Nine features were selected, for which all relevant data was available for 2,450 genes (Dataset S6–9). We found that thee part linear regressions performed better than a single linear regression. In the three part regression, we divided genes into three groups based on lengths of 5' untranslated regions (UTRs): genes with 5' UTRs shorter than 20 nucleotides, genes with 5' UTRs at least 20 nucleotides but shorter than 35 nucleotides, and genes with 5' UTRs no shorter than 35 nucleotides. For genes in each group, we separately regressed their $\log_{10}(TR_{mD})$ or $\log_{10}(TR_{mIND})$ values on each feature. The fitted values of $\log_{10}(TR_{mD})$ or $\log_{10}(TR_{mIND})$ for all the 2,420 genes were used to calculate the R^2 coefficient of determination as the square of the Pearson correlation between the observed $\log_{10}(TR_{mD})$ or $\log_{10}(TR_{mIND})$ values and the fitted values (2).

Defining an optimum PWM scoring strategy

Various length position weight matrices (PWMs) were calculated using the 10% of genes with the highest TR scores (Figure 8 A, top). The mRNA sequences of this set of 245 genes were aligned such that nucleotide +1 corresponds to the A of the initiation codon, AUG. The frequencies of A, U, C, and G at every position from -100 to + 50 were calculated (Dataset S7). Sequences of all 245 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, -15 to -1 PWM etc. Variants of each of these PMWs were constructed that also included nucleotide +4 and positions 3' in 5 nucleotide steps within the protein coding sequence (CDS) to +35. i.e. -5 to +8 PWM, -5 to +13 PWM, -10 to +8 PWM etc. 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 2450 genes as follows:

$$\text{score of gene } g = \log_{10} \prod_{\substack{i=-m \\ i\neq 0}}^{n} p\ (\text{nucleotide at position } i \text{ of gene } g)$$

We then calculated the R^2 correlation coefficient between the PWM scores and the $\log_{10}$ TR values of the 2450 genes using our three part regression. The results are plotted in Figure 8C. This showed that an optimum PWM spans from nucleotide -35 to +28.

Defining a multi feature set for the -35/+28 TICE

Given the 1,447 genes with 5' UTRs > 35 nuc., we calculate a position weight matrices (PWM) of 5' UTRs -35 to -1 nuc. based on the 145 top 10% genes with the largest TR values, as well as a PWM of CDS +4 to +28 nuc. based on the 145 top 10% genes with the largest TR values. Based on each PWM, we calculate the PWM scores of all 1,447 genes. We also calculate the 16 dinucleotide frequencies and 64 trinucleotide frequencies in the -35 to -1 region for every gene, as well as the 16 dinucleotide frequencies and 64 trinucleotide frequencies in the +4 to +28 region for every gene. We refer to the 5' UTR PWM and dinucleotide & trinucleotide frequencies as the -35/-1 TICE features, while we call the CDS PWM and dinucleotide & trinucleotide frequencies as the +4/+28 TICE features. Then we fit a multivariate linear model between the log10 TR (response) and each feature set using the 1,447 genes, and select features using the forward selection with the Bayesian Information Criterion. This procedure results in six 35/-1 TICE features and eight +4/+28 TICE features (Dataset S8). After selecting these features, we recalculate the two PWM matrices based on the 245 top 10% genes with the largest TR values among all 2,450 genes. We also recalculate the dinucleotide and trinucleotide frequencies of all 2,450 genes. Finally, we define the TICE feature set based on the selected 14 features using the values calculated for all 2,450 genes.

Defining a multi feature set of number of 5' UTR ORFs

For each of the 2,450 genes, we calculate the numbers of AUGs and seven AUG variants upstream of the AUG initiating the main protein coding sequence, yielding eight separate features per gene. The seven AUG variants are UGU, AUA, UUG, CUG, AUU, AUC and ACG, all of which have been shown to promote initiation within the 5'UTRs of yeast (3). The eight features were combined in a multivariate model to give a score based on the number of putative upsteam ORFs. Dataset S8 shows the relative contributions to the model of AUG and the seven variants.

Defining multi feature sets of amino acid and codon frequencies

For each of the 2,450 genes, we calculate the 20 amino acid frequencies in its CDS amino acid sequence, as well as the 61 nonstop codon frequencies in its CDS nucleotide sequence. These values were employed in multivariate models for amino acid frequency or codon usage. Dataset S8 shows the relative contributions of each frequency to a model.

Features based on mRNA folding energies

ViennaRNA RNAfold was used to calculate free energies of folding for sliding 35 nucleotide windows as described before (4,5). Two features were selected and calculated for each gene. One, the mean score of windows from the 5' UTR, including windows that captured the first 34 nucleotides of the CDS (i.e. window -1 to +34). Two, the mean score of windows from the CDS.

Determining the percent of TR_{mD} and TR_{mIND} explained by each sequence feature

The nine sequence features selected were:

5' UTR $\log_{10}$ (length) revised from (6)
5' UTR number of ORFs
5' UTR RNA fold energy
-35 to +28 TICE
CDS $\log_{10}$ (number of amino acids) from (6)
CDS amino acid frequency
CDS codon frequency
CDS RNA fold energy
poly-A tail length from (7)

The data for each feature are given in Dataset S6. Each feature was used in our three part regression versus $\log_{10}(TR_{mD})$ and versus $\log_{10}(TR_{mIND})$ as described above.

To test for the null hypothesis that $\log_{10}(TR_{mD})$ and $\log_{10}(TR_{mIND})$ have the same R^2 coefficient of determination with each feature, we design a permutation test as follows.

1. Standardize $\log_{10}(TR_{mD})$ and $\log_{10}(TR_{mIND})$ by subtracting each by their means and subsequently dividing each by their standard deviations.
2. For B=3,000 permutation runs, in each run randomly shuffle the standardized $\log_{10}(TR_{mD})$ and $\log_{10}(TR_{mIND})$ values of every gene, resulting in two vectors of 2,450 dimensions. Then for each feature, calculate the R^2 between it and each vector and take the difference. After B=3,000 runs, 3,000 R^2 differences for each feature are obtained. These differences represent the distribution of the R^2 difference of $\log_{10}(TR_{mD})$ and $\log_{10}(TR_{mIND})$ under the null hypothesis.
3. For each feature, compare the observed R^2 difference of $\log_{10}(TR_{mD})$ and $\log_{10}(TR_{mIND})$ with the 3,000 differences generated from the permutation runs. We calculate the *p*-value as the percentage of the 3,000 differences with absolute values greater than the absolute value of the observed R^2 difference. Since we have seven features in total, we adjust the *p*-values using the Bonferroni correction, that is to multiple each *p*-value by seven, the number of tests.

A model using nine sequence features

To determine the variance in translation rates explained by all nine features, a multivariate model was used to regress all features with TR using our three part regression. Based on the resulting R^2, 80% of the variance in TR was explained by the complete model. Models using different subsets of features were also employed (Figure 11).

Effects of mRNA sequence features on mRNA degradation rates

mRNA sequence features may correlate with $\log_{10}(TR_{mD})$ either because they are direct determinants of translation or because they affect mRNA abundance via an effect on mRNA degradation rates. This question is particularly acute for poly-A tail length and CDS RNA fold energy because these two features showed a substantial correlation with $\log_{10}(TR_{mD})$, not with $\log_{10}(TR_{mIND})$ (Figure 9 and Supplementary Table S7). We therefore designed the following test to infer for each feature if its correlation with $\log_{10}(TR_{mD})$ was significantly explained by an effect on RNA stability.

$$RNA = Txn \bullet RnD$$

$$\log_{10}(RNA) = \log_{10}(Txn) + \log_{10}(RnD)$$

where RNA is the number of mRNA molecules per cell, Txn is the number of mRNA molecules synthesized per cell cycle, and RnD is the fraction of mRNA molecules not degraded per cell cycle ($0 \leq RnD \leq 1$). $\log_{10}(TR_{mD})$ correlates perfectly with $\log_{10}(RNA)$. A regression of $\log_{10}(RnD)$ against $\log_{10}(TR_{mD})$ should therefore produce modified values of $\log_{10}(TR_{mD})$ that eliminate the impact of RNA degradation and instead reflect only the number of mRNA molecules synthesized per cell cycle. We term the modified values $\log_{10}(TR_{mD*})$.

RnD values were calculated from published mRNA half-life data as

$$RnD = e^{(\ln(2)/t0.5)\times 150}$$

Where t0.5 is the mRNA half-life in minutes and 150 is the presumed length of the cell cycle in minutes. RNA half-life data was taken from (8) as it was produced using similar protocols to that used to generate the TR data from (6) that we have analyzed. Both data thus lack the poly-A selection bias found in most earlier data. The RnD values are given in Dataset S6.

To remove the collinearity between RnD and TR_{mD} (i.e. the linear effect of RnD on TR_{mD}) we regressed TR_{mD} on RnD using a single linear regression, and defined the resulting residuals as TR_{mD*} values. Then for each of the seven features in turn, we used it in a three part regression versus $\log_{10}(TR_{mD})$ and versus $\log_{10}(TR_{mD*})$ as described above.

To test for the null hypothesis that $\log_{10}(TR_{mD})$ and $\log_{10}(TR_{mD*})$ have the same R^2 coefficient of determination with each feature versus the alternative hypothesis that $\log_{10}(TR_{mD})$ has a larger R^2 than that of $\log_{10}(TR_{mD*})$, we performed a permutation test as described above.

The results of this analysis are presented in Supplementary Table S8.

The correlation of amino acid and codon frequencies with tRNA abundances (Fig. 10)

For each of the 20 amino acids and each of the 61 nonstop codons, we calculate its mean frequency among the 245 top 10% genes with the largest TR_{mD} values, its mean frequency among the 245 bottom 10% genes with the smallest TR_{mD} values, its mean frequency among the 245 top 10% genes with the largest TR_{mIND} values, and its mean frequency among the 245 bottom 10% genes with the smallest TR_{mIND} values. Then we calculate the R^2 between the TR_{mD} top mean frequencies of the 20 amino acids and combined tRNA abundance of the 20 amino acids, the R^2 between the TR_{mD} bottom mean frequencies of the 20 amino acids and combined tRNA abundance of the 20 amino acids, the R^2 between the TR_{mIND} top mean frequencies of the 20 amino acids and combined tRNA abundance of the 20 amino acids, and the R^2 between the TR_{mIND} bottom mean frequencies of the 20 amino acids and combined tRNA abundance of the 20 amino acids. We similarly calculate four R^2s for the 61 nonstop codon frequencies. We use the R package "cocor" and its function cocor.dep.groups.overlap() to compare the top R^2 and the bottom R^2, and use Bonferroni correction to adjust the resulting *p*-values. We also calculate the ratio of the TR_{mD} top mean amino acid frequencies and the TR_{mD} bottom mean amino acid frequencies, and calculate the Pearson correlation between the 20 ratios and the combined tRNA abundance of the 20 amino acids. We similarly calculate the Pearson correlation between the TR_{mIND} top / bottom ratios of the 20 amino acids and the combined tRNA abundance, the Pearson correlation between the TR_{mD} top / bottom ratios of the 61 nonstop codons and the tRNA abundance, and the Pearson correlation between the TR_{mIND} top / bottom ratios of the 61 nonstop codons and the tRNA abundance. We use R function cor.test() to test if each Pearson correlation is significantly positive and correct the resulting *p*-values using Bonferroni correction.

References

1. Smith, R.J. (2009) Use and misuse of the reduced major axis for line-fitting. *Am J Phys Anthropol*, **140**, 476-486.
2. Everitt, B.S. (2002) *The Cambridge Dictionary of Statistics*. 2nd ed. Cambridge University Press, Cambridge, UK.
3. Chang, C.P., Chen, S.J., Lin, C.H., Wang, T.L. and Wang, C.C. (2010) A single sequence context cannot satisfy all non-AUG initiator codons in yeast. *BMC Microbiol*, **10**, 188.
4. Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol*, **6**, 26.
5. Chew, G.L., Pauli, A. and Schier, A.F. (2016) Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat Commun*, **7**, 11663.
6. Weinberg, D., Shah, P., Eichhorn, S., Hussmann, J., Plotkin, J. and Bartel, D. (2016) Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell Reports*, **14**, 1787-1799.
7. Subtelny, A.O., Eichhorn, S.W., Chen, G.R., Sive, H. and Bartel, D.P. (2014) Poly(A)-tail lengths and a developmental switch in translational control. *Nature*, **508**, 66-71.
8. Presnyak, V., Alhusaini, N., Chen, Y.H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R. *et al.* (2015) Codon optimality is a major determinant of mRNA stability. *Cell*, **160**, 1111-1124.