bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

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

[1] Department of Statistics and Department of Human Genetics, University of California, Los Angeles, CA 90095.

[2] Computational Biology Program, Public Health Sciences and Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109.

[3] Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94708.

*** Corresponding author**. Biological Systems and Engineering Division, 1 Cyclotron Road MS 84-171, Lawrence Berkeley National Laboratory, Berkeley, CA 94708.

mdbiggin@lbl.gov (MDB)

short title: Quantitating translational control

Keywords: mRNA abundance **/** Protein abundance / Protein degradation / Translation

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Translation rate per mRNA molecule correlates positively with mRNA abundance. As a result, protein levels do not scale linearly with mRNA levels, but instead scale with the abundance of mRNA raised to the power of an "amplification exponent". Here we show that to quantitate translational control, the translation rate must be decomposed into two components. One, TR_{mD}, depends on the mRNA level and defines the amplification exponent. The other, TR_{mIND}, is independent of mRNA amount and impacts the correlation coefficient between protein and mRNA levels. We show that in *S. cerevisiae* TR_{mD} represents ~20% of the variance in translation and directs an amplification exponent of 1.20 with a 95% confidence interval [1.14, 1.26]. TR_{mIND} constitutes the remaining ~80% of the variance in translation and explains ~5% of the variance in protein expression. We also find that TR_{mD} and TR_{mIND} are preferentially determined by different mRNA sequence features: TR_{mIND} by the length of the open reading frame and TR_{mD} both by a ~60 nucleotide element that spans the initiating AUG and by codon and amino acid frequency. Our work provides more appropriate estimates of translational control and implies that TR_{mIND} is under different evolutionary selective pressures than TR_{mD}.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

The relative contributions of transcriptional and post-transcriptional control to protein expression levels in eukaryotes are the topic of ongoing debate (1-3). One view suggests that translation and protein degradation together play the dominant role because protein and mRNA abundance data correlate poorly (coefficient of determination for $\log_{10}$ transformed values $R^2_{prot–RNA}$ = 0.2–0.45) (4-9). Other work, though, has shown that the correlation is much higher when measurement error is considered ($R^2_{prot–RNA}$ = 0.66–0.83), implying that transcription dominates (10-12). In addition, the variance in translation rates affects not only the correlation coefficient between protein and mRNA, but also the slope of the relationship because translation rates increase with mRNA abundance (12). Whereas most studies assumed that protein abundances scale linearly with mRNA levels, Csardi et al. demonstrate that protein abundances scale with mRNA levels raised to the power of an "amplification exponent" ($b_{prot–RNA}$). Presumably the mRNAs of genes that are expressed at high levels, such as those for ribosomal proteins and glycolytic enzymes, contain nucleotide sequence signals that promote faster rates of translation per message than observed for less abundant mRNAs (12).

In this article, we argue that because translation affects both $R^2_{prot–RNA}$ and $b_{prot–RNA}$, the approaches used previously to quantify the contribution of translation to protein expression are improper. Prior approaches have sought to provide a single metric to estimate translational control: the impact of translation on $R^2_{prot–RNA}$. We propose that, instead, proper quantification requires that translation rates (TR) be decomposed mathematically into two components: one that is dependent on mRNA abundance (TR_{mD}) and one that is not (TR_{mIND}). For a given gene *i*

$$TR_i = TR_{mDi} \bullet TR_{mINDi},$$

where TR is the number of protein molecules translated per mRNA molecule; TR_{mIND} only affects $R^2_{prot–RNA}$ (not $b_{prot–RNA}$); and TR_{mD} only affects $b_{prot–RNA}$ (not $R^2_{prot–RNA}$).

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The traditional view of the steady-state relationship between protein and mRNA for gene *i* can be expressed as

$$prot_i = RNA_i \bullet TR_i \bullet PnD_i \qquad \text{(Materials and Methods, equation i)},$$

where prot and RNA are the number of protein molecules and mRNA molecules per cell, respectively, and PnD is the fraction of protein that is not degraded per cell cycle ($0 \leq PnD \leq 1$). Once TR is decomposed, this equation can be reformulated as

$$prot_i = a \bullet RNA_i^{b_{prot-RNA}} \bullet TR_{mIND_i} \bullet PnD_i \qquad \text{(Materials and Methods, equation viii)},$$

where a and $b_{prot-RNA}$ are positive constants for all genes. This reformulated equation has the advantage that it explicitly describes the non-linear relationship between protein and mRNA levels as well as permitting correct quantitation of translation's contribution to protein levels.

Three idealized scenarios explain the complex dependency of protein abundances on mRNA levels and the two components of translation. Plots of $\log_{10}$-transformed data are employed because the amplification exponent $b_{prot-RNA}$ is simply the linear slope of the relationship in logarithmic space (Figure 1), i.e.,

$$\log_{10}(prot_i) = \log_{10}(a) + b_{prot-RNA} \bullet \log_{10}(RNA_i) + \log_{10}(TR_{mIND_i}) + \log_{10}(PnD_i).$$

In the first scenario, translation rates are equal for all genes (i.e. TR_i = constant) as are protein degradation rates. Therefore, $R^2_{prot-RNA} = 1$ and $b_{prot-RNA} = 1$ (Figure 1A). In the second scenario, translation rates correlate perfectly with mRNA levels (i.e., $TR_i = TR_{mD_i}$), while the protein degradation rate is constant for all genes. Thus, $R^2_{prot-RNA} = 1$ and $b_{prot-RNA} > 1$. (Figure 1B). In the third scenario, translation and protein degradation rates are both uncorrelated with mRNA (i.e., $TR_i = TR_{mIND_i}$). Therefore, $R^2_{prot-RNA} < 1$ and $b_{prot-RNA} = 1$ (Figure 1C).

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The third scenario is the one most widely considered in the literature. Csardi et al. argue, though, that the truth is a hybrid of this scenario and the second scenario because translation is partially, but not fully, correlated with mRNA abundance. A Bayesian model was employed to estimate protein and mRNA abundances for 5,854 annotated protein-coding genes in *S. cerevisiae,* including 842 genes for which either protein or mRNA abundance data was lacking. From the modeled abundances, it was estimated that $b_{prot-RNA} = 1.69$ and $R^2_{prot-RNA} = 0.85$ (12).

Csardi et al.'s basic premise is important. Their Bayesian model, however, did not take into account which methods provide accurately scaled abundance data, and they did not decompose TR. Bayesian models are, in addition, inherently subjective because priors are chosen by the researcher. Therefore, we adopted a non-modeling approach that considers empirically determined abundance measurements that have been scaled using internal concentration standards, and we decomposed TR. We find that in *S. cerevisiae* TR_{mD} represents ~20% of the variance in translation and results in an amplification exponent of 1.20 with a 95% confidence interval [1.14, 1.26] and that TR_{mIND} constitutes the remaining ~80% of the variance in translation and explains ~5% of the variance in protein expression. By taking into account protein degradation data and measurement error, we also show that the expected correlation between the abundances of protein and mRNA is $R^2_{prot-RNA}$ ~0.94. This value is markedly higher than the $R^2_{prot-RNA} = 0.80$ obtained between the Bayesian model's abundance estimates for the 5,045 genes for which empirical data are available. Finally, we examined which mRNA sequence elements explain the variance in TR_{mD} and TR_{mIND} using a model that predicts 80% of the variance in TR from mRNA sequence data alone. We find that TR_{mD} is most strongly determined both by RNA secondary structure within a ~60 nucleotide element that spans the initiating AUG and by the fact that the amino acid and codon frequencies encoded in highly expressed mRNAs more closely correlate with the abundances of their cognate tRNAs than is the case for mRNAs expressed at lower levels. TRmIND, by contrast, is chiefly

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

determined by the length of the protein coding region. TR_{mIND} is thus likely under different evolutionary selective pressures than TR_{mD} and predominantly controlled by different mechanisms. Our work establishes more accurate estimates of translational control than earlier research. In addition, our analysis illustrates that decomposing translation rates allows insights into the mRNA sequence dependence of translation that would not otherwise be apparent.

Materials and Methods

Data and code

All of the data used are provided in Datasets S1–S9. The mRNA and protein abundance datasets used by Csardi et al. as input to their Bayesian model (Dataset S1) are from their file “scer-mrna-protein-raw.txt” (12). The estimates for the true abundances of mRNA and protein generated by Csardi et al.’s Bayesian model (Dataset S2) are from their file "scer-mrna-protein-absolute-estimate-sample.txt" and are for a single sample from their “SCM” values (12). The scaling-standard mRNA data are from NanoString (13,14), qPCR (15) and competitive PCR (16) studies (Dataset S3). Three scaling-standard protein datasets were measured by western blot (17), flow cytometry (18) and selected reaction monitoring mass spectrometry (19) (Dataset S3). A fourth scaling-standard protein dataset was compiled as an extension of one by von der Haar (20) to which additional data were added (21-26) (Dataset S3). The ribosome profiling data comprise median values from several studies provided by Csardi et al. (12) and, separately, the translation-initiation efficiency values from Weinberg et al. (27) (Dataset S4). Protein degradation data is from Christiano et al. (28) (Dataset S5). The mRNA sequence feature information was from Weinberg et al. and Subtelny et al. or was calculated as described in Supplementary Methods S4 (27,29) (Datasets S6–S9). The fraction of RNA not degraded was calculated from Presnyak et al. as described in Supplementary Methods S4 (30) (Dataset S6). Dataset S2 also includes our corrected versions of the Csardi et al. protein and mRNA

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

abundance data. Dataset S4 includes corrected versions of Weinberg et al. mRNA abundance and ribosome density data as well as calculated values of TR_{mD} and TR_{mIND}.

The R code used in the analyses are provided in Dataset S10. Both a word file and executable files are provided.

The relationship between the steps in protein production

For simplicity we consider the ideal case where there is no measurement error, i.e. where the true values are measured. It is assumed that the system is at steady state. We denote

RNA = the abundance of a particular mRNA (molecules per cell)

prot = the abundance of a particular protein (molecules per cell)

TR = the translation rate of a particular mRNA (the number of protein molecules translated per mRNA molecule)

TR_{mIND} = the mRNA abundance-independent component of TR

TR_{mD} = the mRNA abundance-dependent component of TR

PnD = a scaling factor that gives the fraction of a particular protein that remains undegraded per cell cycle, i.e. (1 – the fraction of the protein degraded per cell cycle); $0 \leq PnD \leq 1$.

a = a constant for all genes

b_{TR-RNA} = a constant for all genes that is the slope of the relationship between log-transformed translation rates and log-transformed mRNA levels. It thus measures the amplification of translation rates due to mRNA abundance

$b_{prot-RNA}$ = a constant for all genes that is the slope of the relationship between log-transformed protein abundance and log-transformed mRNA levels. It is thus

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

also the amplification exponent for the relationship between the unlogged abundances.

We assume that PnD is not correlated with mRNA abundance and, thus, has no impact on $b_{prot–RNA}$. This appears to be a reasonable assumption because the correlation between measured values for PnD and mRNA abundance is very low ($R^2_{PnD–RNA} < 0.005$; Supplementary Table S2).

The abundance of a chosen protein is given by

$$prot = RNA \bullet TR \bullet PnD \quad (i)$$

$$\log(prot) = \log(RNA) + \log(TR) + \log(PnD) \quad (ii)$$

In an idealized situation where log-transformed translation rates correlate perfectly with log-transformed mRNA levels (i.e. $R^2_{TR–RNA} = 1$; and $TR_{mIND} = 1$)

$$TR = TR_{mD} \bullet TR_{mIND} = TR_{mD} = a \bullet RNA^{b_{TR–RNA}}$$

$$\log(TR_{mD}) = \log(a) + b_{TR–RNA} \bullet \log(RNA) \quad (iii)$$

When log-transformed translation rates only partially correlate with log-transformed mRNA levels then

$$TR = TR_{mD} \bullet TR_{mIND} = a \bullet RNA^{b_{TR–RNA}} \bullet TR_{mIND}$$

$$\log(TR) = \log(TR_{mD}) + \log(TR_{mIND}) \quad (iv)$$

$$\log(TR) = \log(a) + b_{TR–RNA} \bullet \log(RNA) + \log(TR_{mIND}) \quad (v)$$

Combining (ii) and (v)

$$\log(prot) = \log(RNA) + \log(a) + b_{TR–RNA} \bullet \log(RNA) + \log(TR_{mIND}) + \log(PnD)$$

$$\log(prot) = \log(a) + (1 + b_{TR–RNA}) \bullet \log(RNA) + \log(TR_{mIND}) + \log(PnD) \quad (vi)$$

From (vi), the slope of the relationship between log(prot) and log(RNA) is $(1 + b_{TR–RNA})$, i.e.,

$$b_{prot–RNA} = 1 + b_{TR–RNA} \quad (vii)$$

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Combining (vi) and (vii)

$$\log(\text{prot}) = \log(a) + b_{\text{prot–RNA}} \bullet \log(\text{RNA}) + \log(\text{TR}_{\text{mIND}}) + \log(\text{PnD})$$

$$\text{prot} = a \bullet \text{RNA}^{b_{\text{prot–RNA}}} \bullet \text{TR}_{\text{mIND}} \bullet \text{PnD} \qquad \text{(viii)}$$

Estimating the slope $b_{TR–RNA}$ and the contributions of TR_{mIND} and TR_{mD} to TR.

Having defined the basic relationships between steps in protein expression, we now estimate the value for $b_{\text{TR–RNA}}$ and the contributions of TR_{mIND} and TR_{mD} to TR.

From (iv) and the fact that $\log(\text{TR}_{\text{mD}})$ and $\log(\text{TR}_{\text{mIND}})$ are uncorrelated by definition

$$\text{var}(\log(\text{TR})) = \text{var}(\log(\text{TR}_{\text{mD}})) + \text{var}(\log(\text{TR}_{\text{mIND}})) , \qquad \text{(ix)}$$

where var is the variance.

From (iii) and given that $\text{var}(\log(a)) = 0$

$$\text{var}((\log(\text{TR}_{\text{mD}})) = \text{var}(b_{\text{TR–RNA}} \bullet \log(\text{RNA})) = b_{\text{TR–RNA}}^2 \bullet \text{var}(\log(\text{RNA})) \qquad \text{(x)}$$

Combining (ix) and (x)

$$\text{var}(\log(\text{TR})) = b_{\text{TR–RNA}}^2 \bullet \text{var}(\log(\text{RNA})) + \text{var}(\log(\text{TR}_{\text{mIND}}))$$

From (x)

$$b_{\text{TR–RNA}}^2 = \text{var}((\log(\text{TR}_{\text{mD}})) / \text{var}(\log(\text{RNA}))$$

Therefore true slope

$$b_{\text{TR–RNA}} = \text{sd}(\log(\text{TR}_{\text{mD}})) / \text{sd}(\log(\text{RNA})) , \qquad \text{(xi)}$$

where sd is the standard deviation.

We considered three different regressions for estimating the value of the true slope $b_{\text{TR–RNA}}$, finding that the Ordinary Least Squares (OLS) regression described is the most appropriate (Supplementary Methods S3).

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Results

Estimates for $b_{prot–RNA}$ from protein and mRNA abundances

Csardi et al.'s estimate for $b_{prot–RNA}$ was derived using a Bayesian model to determine the true levels of mRNAs and proteins based on multiple abundance datasets from the literature and imputed values when data were lacking (12). However, the methods used to produce most of the empirical data input to this model (e.g. mRNA microarray, RNA-seq, and label-free mass spectrometry) do not employ internal concentration standards. As a result, the standard deviations of the data can be—depending on the method—either systematically compressed or systematically expanded relative to the true values (10,12,31-33). There is no guarantee that such reproducible biases can be corrected by a Bayesian model. The slope of any relationship depends on the standard deviations of the *x* and *y* values, so improperly scaled data is likely to exhibit an inaccurate slope.

We therefore re-estimated $b_{prot–RNA}$ by correcting abundances of protein and mRNA using datasets that had been derived by methods employing internal concentration standards. The internal standards are used to account for any linear or non-linear scaling bias in the raw data, and thus the final data produced by these methods should be reasonably scaled. Data for individual genes will still include some gene specific error, but the standard deviation of the whole dataset will not be much impacted by such error. We refer to these datasets as "scaling-standards". NanoString (13,14), qPCR (15) and competitive PCR (16) studies provided four independent mRNA scaling-standards (Dataset S3). Western blot (17); flow cytometry (18); selected reaction monitoring mass spectrometry (19) and a compilation of assorted methods (20-26) each provided one of four protein scaling-standards (Dataset S3). Plots of these scaling-standards against the corresponding abundance values from the Bayesian model reveal the relative scaling of each dataset: scaling-mRNA vs. Bayesian mRNA and scaling-protein vs. Bayesian protein (Figure 2 and Supplementary Figure S1). The slope of a linear regression fit to

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the log-transformed data for each of the eight pairwise comparisons was then used to correct the scaling of the Bayesian abundance estimates (Dataset S2). The Reduced Major Axis (RuMA) regression was used as it is the only one that allows the scaling of a dataset to be adjusted such that its standard deviation becomes equal to that of a scaling-standard (34) (Supplementary Methods S1).

The standard deviation of the uncorrected Bayesian protein dataset approximates those of the scaling-protein datasets (RuMA slope $\hat{b}_{\text{sprot–prot}}$ = 0.87–1.11), while the standard deviation of the uncorrected Bayesian mRNA data is less than those of the scaling-mRNA sets (RuMA slope $\hat{b}_{\text{sRNA–RNA}}$ = 1.34–1.54) (Figure 2 and Supplementary Figure S1). After correcting the scaling bias of the Bayesian data, we bootstrapped the corrected versions of the data to obtain a mean RuMA estimate of $\hat{b}_{\text{prot–RNA}}$ = 1.17 with a 95% quantile confidence interval [1.10, 1.26] (Figure 3B).

Csardi et al. used the Ranged Major Axis (RgMA) to estimate the slope $b_{\text{prot–RNA.}}$ This other type of regression yields a slope that is nearly identical to that of the RuMA regression for our corrected versions of the protein and mRNA abundance data, $\hat{b}_{\text{prot–RNA}}$ = 1.16 with a 95% quantile confidence interval [1.09, 1.25]. We also considered two additional, though more approximate, approaches to determine $\hat{b}_{\text{prot–RNA}}$. These two methods estimate that $\hat{b}_{\text{prot–RNA}}$ = 1.08 or 1.10 (see Supplementary Methods S2 and Table S1). Thus, the evidence strongly suggests that the amplification exponent $b_{\text{prot–RNA}}$ is much smaller than the previously reported value of 1.69 (Figure 3).

To investigate the basis for Csardi et al.'s higher estimate of $b_{\text{prot–RNA}}$, we compared the standard deviations of the datasets input to their Bayesian model with those of our scaling-standards and with the abundances output by the Bayesian model. While the standard

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

deviations of the 20 input protein datasets range above and below that of the protein scaling-standards, their mean scalings are similar (i.e. mean RuMA $\hat{b}_{\text{sprot–prot}}$ 0.96) and agree closely with that of the Bayesian model (Figure 4A). By contrast, while most of the 38 input mRNA datasets are scaled similarly to the mRNA scaling-standards, 33 out of the 38 input mRNA datasets have a larger standard deviation than the Bayesian model's abundance estimates (Figure 4B). The model has, in effect, given greater weight to the small minority of the input mRNA data that have the most compressed scaling. This minority is dominated by mRNA microarray data (Figure 4B), which is known to give compressed abundance estimates relative to the true values (32,33). The Bayesian model's strong weighting on biased microarray data thus appears to explain its high estimate for $b_{\text{prot–RNA}}$.

Estimates for $b_{\text{prot–RNA}}$ from ribosome profiling data

The previous study by Csardi et al. used a "toy" model to independently determine $b_{\text{prot–RNA}}$ from the slope and correlation between translation rates and mRNA abundances (12). Using averaged measurements of translation rates and mRNA abundances from several ribosome profiling studies (29,35-37), it was suggested that the toy model was consistent with $b_{\text{prot–RNA}}$ = 1.69 (12). Since our results are inconsistent with this estimate for $b_{\text{prot–RNA}}$, we have independently explored the relationship between $b_{\text{prot–RNA}}$ and ribosome profiling data. Again we adopted a non-modeling approach that defines the appropriate mathematical equations and employs the most accurate datasets available.

The correlation between measured protein degradation data and mRNA abundance data is negligible ($R^2_{\text{PnD–RNA}}$ < 0.005; Materials and Methods and Supplementary Table S2) (28). Thus, we can assume that protein degradation has no impact on $b_{\text{prot–RNA}}$. Hence, the relationship between $b_{\text{prot–RNA}}$ and $b_{\text{TR–RNA}}$ is

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$b_{prot–RNA} = 1 + b_{TR–RNA}$ (Materials and Methods, equation vii),

where $b_{TR–RNA}$ is the true slope between log-transformed translation rates versus log-transformed mRNA levels.

To estimate $b_{TR–RNA}$ we employed two available ribosome profiling datasets: one used by Csardi et al. (12), which we refer to as "Csardi–median", and another from Weinberg et al. (27) (Dataset S4). The Weinberg data eliminates a poly-A mRNA selection bias and has been corrected to reduce two additional sources of bias (27). As a result, these data show a higher correlation between translation rates and mRNA levels than previously observed (27) and appear to be more accurate than the Csardi–median data because they correlate more highly with both the mRNA and the protein scaling-standards (Supplementary Table S3). The standard deviations of the Weinberg ribosome-density and mRNA data differ modestly from that of their respective scaling-standards (mean RuMA $\hat{b}_{sprot–RD}$ = 0.98; RuMA $\hat{b}_{sRNA–RNA}$ = 1.07). We corrected this miss-scaling in the Weinberg data using the scaling-standards (Dataset S4) and then used the Ordinary Least Squares (OLS) regression to estimate $b_{TR–RNA}$ on the corrected data. The result suggests that the amplification exponent $b_{prot–RNA}$ = 1 + 0.22 = 1.22 with a 95% bootstrap quantile confidence interval [1.13, 1.29] (Figure 3C; Table S4).

Rather than correcting the Csardi–median data, we analyzed it in its original form so that we could compare analysis strategies on the same data. The result suggests that $b_{prot–RNA}$ = 1+ 0.28 = 1.28 with a 95% bootstrap quantile confidence interval [1.26, 1.31] (Table S4). Csardi et al.'s claim that ribosome profiling data were consistent with an amplification exponent of 1.69 must therefore be largely due to differences between our analysis methods and those that they employed, not the data used.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Csardi et al. estimated b_{TR-RNA} using the RgMA regression rather than OLS. For the corrected Weinberg data, RgMA $\hat{b}_{TR-RNA}$ predicts $b_{prot-RNA}$ = 1 + 0.31 = 1.31; for the Csardi–median dataset, RgMA $\hat{b}_{TR-RNA}$ predicts $b_{prot-RNA}$ = 1 + 0.55 = 1.55 (Supplementary Table S4). The RgMA slope, however, is insensitive to the correlation coefficient (Supplementary Table S5 and Methods S3). In effect, this regression assumes that the true translation rates and true mRNA levels correlate perfectly and that the poor correlations observed between the data ($R^2_{TR-RNA} \leq 0.28$; Supplementary Table S4) are due only to measurement errors that are somewhat evenly split between the TR and mRNA data. The OLS regression, by contrast, down-weights the slope as the correlation decreases (34) (Supplementary Table S5 and Methods S3). It effectively assumes that the poor correlation between translation and mRNA abundance is largely due to a genuine biological phenomenon rather than measurement error. OLS-based estimates better match current thinking that translational control includes a substantial component that is unrelated to the abundance of each mRNA. In addition, OLS $\hat{b}_{TR-RNA}$ predicts a value for the amplification exponent that is more similar to that we obtained from scaling-standard-rescaled protein and mRNA abundances (1.22 vs 1.17 respectively) than to RgMA $\hat{b}_{TR-RNA}$ (1.31 or 1.55 vs 1.17). Thus, OLS $\hat{b}_{TR-RNA}$ should give a more accurate estimate of $b_{prot-RNA}$ (see Supplementary Methods S3 for further justification). Averaging our estimate from ribosome profiling data with that from corrected protein and mRNA abundances (i.e. the estimates in Figure 3B and C) provides our most accurate estimate for $b_{prot-RNA}$ as 1.20 with a 95% confidence interval [1.14, 1.26].

Estimating mRNA abundance-dependent and independent translational control

The variance in protein levels is caused by gene-specific differences in mRNA abundances, translation rates, and protein degradation rates. Because translation rates correlate with mRNA levels, it has been suggested that the percent of the variance in true

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

protein amounts that is explained by the true individual contributions of mRNA, translation, and protein degradation sum to more than 100% (12). This argument is, however, misleading. The correlation coefficient between translation and protein abundance is not a legitimate measure of the contribution of translation to protein expression because it breaches one of the essential requirements for analysis of variance (ANOVA). ANOVA is only valid when the true explanatory variables (in this case mRNA abundance, translation and protein degradation) are fully uncorrelated with each other (i.e. when they are not collinear) and, as a result, when their marginal contributions sum to exactly 100%. Therefore, as briefly explained in the Introduction, to determine the contribution of translation rates (TR) to protein expression is it essential to decompose TR into two components: one that is dependent on mRNA abundance (TR_{mD}) and a second that is independent of mRNA abundance (TR_{mIND}), where $TR = TR_{mD} \bullet TR_{mIND}$. TR_{mIND} determines the variance in protein levels that is not explained by mRNA or protein degradation; it has no impact on $b_{prot-RNA}$. TR_{mD}, by contrast, only affects the amplification exponent $b_{prot-RNA}$; it makes no contribution to $R^2_{prot-RNA}$. The abundance of any protein i is then given by the following

$$\log_{10}(prot_i) = \log_{10}(a) + (1+b_{TR-RNA}) \bullet \log_{10}(RNA_i) + \log_{10}(TR_{mIND\ i}) + \log_{10}(PnD_i)$$

$$b_{TR-RNA} = sd(\log_{10}(TR_{mD})) / sd(\log_{10}(RNA)),$$

where sd is the standard deviation; a is positive constant for all genes; PnD is the fraction of protein not degraded; and $(1+b_{TR-RNA}) = b_{prot-RNA}$ (Materials and Methods, equations vi and xi). As one consequence of this, 100% of the variance in true protein expression is explained by the sum of the contributions of the variances of true RNA, TR_{mIND} and PnD values.

To quantitate the contribution of translation to protein expression using this new strategy, we first calculated gene-specific values of TR_{mIND} and TR_{mD} from OLS regressions of translation efficiency on mRNA abundance for both the Csardi-median and the Weinberg datasets (Figure

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5 and Dataset S4). In addition, from these same regressions we determined the percent of the variance in TR that is explained by the variances in TR_{mD} and TR_{mIND}. Assuming no measurement error, these values are 19%–21% and 79%–81% respectively (Table S4).

The contribution of TR_{mD} to protein abundance is given by the amplification exponent $b_{prot-RNA}$, which we have estimated earlier as 1.20 with a 95% confidence interval [1.14, 1.26]. The contribution of TR_{mIND} to protein abundance was derived from the linear regression of the gene specific values of protein data on TR_{mIND}. TR_{mIND} only accounted for 1%–3% of the variance in the protein abundance estimates from the Bayesian model (Supplementary Table S4). Because these percentages were surprisingly low, we recalculated the contribution of TR_{mIND} by regressing the protein scaling-standards against it to test for an unknown bias in the output of the Bayesian model. The mean contributions of TR_{mIND} to the variance in the scaling-standard protein datasets were also low: 4% (Supplementary Table S6). We also re-estimated TR_{mIND} by regressing translation efficiencies against the Bayesian mRNA abundances to avoid any potential bias in the mRNA data from the ribosome profiling studies. These re-calculated values for TR_{mIND}, though, still only explain <1% of the variance in the Bayesian protein data.

To compare our new metrics to one derived from undecomposed TR, we determined the R^2 coefficient of determination between undecomposed TR and protein abundance data. $R^2_{prot-TR}$ = 0.24–0.28 (Supplementary Table S4). This relatively high value helps expose why $R^2_{prot-TR}$ cannot be used as measure of the contribution of translation to protein abundance. TR_{mIND} represents ~80% of the variance in TR, yet $R^2_{prot-TRmIND}$ is dramatically lower than $R^2_{prot-TR}$ (0.01–0.04 vs 0.24–0.28). TR_{mD} accounts for only ~20% of the variance in TR and yet is chiefly responsible for the fact that $R^2_{prot-TR} >> R^2_{prot-TRmIND}$ (Supplementary Table S4). It is counter-intuitive that a ~20% minority of the variance in TR should have much the dominant contribution to protein expression. In effect, $R^2_{prot-TR}$ is a hybrid measure of the correlation of TR_{mIND} with

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

protein abundances combined with some part of the correlation between mRNA abundance and protein levels. Only by decomposing TR can the impact of translation be properly quantitated and provide metrics consistent with the requirement of ANOVA that explanatory variables be completely uncorrelated.

Estimating post-transcriptional control

The contribution of protein degradation to the variance of protein abundance in actively dividing yeast cells is very low because the median half-life of proteins is 3.5 times longer than the cell division rate (28). By our estimate, this contribution is ~1% (Supplementary Table S2). As explained above, the percentage contributions of the variances in the true values of mRNA, TR_{mIND}, and PnD should sum to explain exactly 100% of the variance in true protein levels (Materials and Methods, Equation v). For measured data, though, the sum of the contributions is no more than 77% (mRNA) + 4% (TR_{mIND}) + 1% (PnD) = 82% (Figure 6A and Supplementary Tables S2, S4 and S6). This discrepancy reveals another advantage of our framework. The ~18% of the variance in protein data that is unexplained (Figure 6A) should be due to measurement error. Our approach thus provides an assessment of the magnitude of error, whereas error cannot be estimated if TR is left undecomposed.

Further, if we assume that the proportion of measurement error is similar in each data class, we can estimate the contribution of the true values of each step to true protein expression. When we do this, the variance in the true values of TR_{mIND} + PnD explain ~6% of the variance in true protein levels, while TR_{mD} makes an additional contribution by increasing slope $b_{prot–RNA}$ from its ground state of 1 to more like 1.20 (Figure 6B). The expected correlation between true protein and true mRNA abundances is thus $R^2_{prot–RNA}$ ~0.94 (Figure 6B).

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The mRNA sequence determinants of TR_{mD} and TR_{mIND}

The fact that translation rates correlate with mRNA abundances suggests that highly expressed mRNAs contain features in their nucleic acid sequences that specify faster rates of translation than mRNAs present at low levels (12). Such mRNA sequence features would thus correlate with TR_{mD}. TR_{mIND}, on the other hand, is by definition fully uncorrelated with mRNA abundance and with TR_{mD}. It is plausible then that the two components of translation may be specified by different sequence elements and controlled by separate mechanisms. We therefore sought to determine if there are mRNA sequence features that specify TR_{mD} and to assess if these differ from those that define TR_{mIND}.

Detailed prior work has identified several mRNA sequence features that correlate with, and in some cases have been directly shown to affect, rates of translation (27,29,30,38-56). Extending this earlier work, we defined nine sequence features that predict between 5% – 60% of the variance in the rates of translation when tested in pairwise regressions in which only one feature is present. When all nine features are combined in a multivariate model, 80% of translation is explained (Figure 7; Supplementary Table S7 and Methods S4; Datasets S6-S9). Of note, a Translation Initiation Control Region (TICE) that flanks the AUG codon alone explains 33% of the variance in translation rates (Figures 7 and 8). The extent of the TICE was determined by testing Position Weight Matrices (PWMs) of differing lengths, which showed that the TICE is largely encoded by nucleotides -35 to +28 (Figure 8C). The -35 to -1 region is strikingly more A rich and G poor in highly translated mRNAs than in less well translated genes, while the +4 to +28 region shows more complex position specific differences with translation rate (Figure 8A and Supplementary Figure S2). Further analysis revealed that the frequencies of a subset of dinucleotides and trinucleotides within the -35/-1 and the +4/+28 regions allow more complete prediction of translation rates when combined with PWMs (Figure 8D and Dataset S8). Consistent with earlier observations (39-41,44,51), the TICE is much less likely to adopt a

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

folded RNA structure in highly translated mRNAs than it is in poorly translated mRNAs (Figure 8B), suggesting that it functions at least in part by specifying structure.

Using the nine features, the percent of the variances in TR_{mD} and TR_{mIND} that are explained by each in pairwise regressions were determined (Figure 9 and Supplementary Table S7). While TR_{mD} and TR_{mIND} both correlate with multiple features, there are significant differences in the degree to which some features explain TR_{mD} versus TR_{mIND}. CDS length has a much larger impact on TR_{mIND} than on TR_{mD} (Bonferroni corrected $p < 0.001$). On the other hand, the TICE, the frequencies of amino acids or codons encoded by the CDS, RNA folding of the CDS, and poly A tail length each explains more of TR_{mD} than TR_{mIND} (Bonferroni corrected $p < 0.001$). The remaining three features—length of the 5' untranslated region (UTR), number of AUG codons in the 5' UTR, and RNA folding in the 5' UTR—show no compelling discrimination in the degree to which they explain TR_{mIND} and TR_{mD} (Bonferroni corrected $p > 0.074$).

The mRNA sequence features that correlate with TR_{mIND} are likely to be mechanistic determinants of translation rates, see Discussion. The features that correlate with TR_{mD}, however, could in principle directly affect translation or they could instead only impact mRNA stability. Their correlation with TR_{mD} might not reflect a direct mechanistic role in translation but instead a fortuitous consequence of their impact on mRNA abundance. We therefore determined if measured mRNA degradation rate data could explain the correlation of each feature with TR_{mD} by calculating revised TR_{mD} values (TR_{mD*}) where the expected impact of RNA degradation has been removed (Supplementary Methods S4 and Table S8) (30). Only poly-A length and CDS RNA folding showed a significant reduction in their correlation with TR_{mD*} (Bonferroni corrected $p < 0.001$). The remaining features showed similar correlations with TR_{mD} and TR_{mD*} (Bonferroni corrected $p > 0.072$) (Supplementary Table S8). Thus, poly-A tail

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

length and CDS RNA folding likely act at least in part by impacting mRNA stability. The correlation of the other seven features with translation rates appears to reflect direct control of protein synthesis.

The frequencies of codons in different mRNAs correlate with the abundance of the encoded proteins (27,30,45-49,52). Codon frequencies within highly translated mRNAs more closely match the abundances of their cognate tRNAs than is the case for poorly translated messages, resulting in higher rates of translation elongation (27,30,45-49,52). The substantial correlation between amino acid or codon frequencies with TR_{mD} and with TR_{mIND} (Figure 9) therefore likely reflects control of elongation. To directly test this, we first determined the frequencies of amino acids and codons in the 10% of genes with the highest values of TR_{mD} or TR_{mIND} (top cohorts) and separately the frequencies of amino acids and codons in the 10% of genes with the lowest values (bottom cohorts) (Dataset S9). We then correlated these frequencies with tRNA abundance. All cohorts show a positive correlation (Figure 10 A and B; Supplementary Table S9). Top cohorts, however, consistently show a higher correlation than bottom cohorts, though this differences is only statistically significant for codon frequencies not for amino acid frequencies (Figure 10 A and B; Supplementary Table S9). Notably, there is a larger difference between the top and bottom TR_{mD} cohorts than seen between the top and bottom TR_{mIND} cohorts.

We also calculated the ratio of amino acid or codon frequencies between top cohorts divided by that of bottom cohorts (Dataset S9). For a given amino acid or codon, a ratio of greater than one thus indicates that it is more abundant in highly translated mRNAs than in poorly translated messages. Scatter plots comparing these ratios to tRNA abundance show positive correlations, with only the TR_{mIND} amino acid ratios not showing a significant correlation (Figure 10 C-F). The correlations are stronger for TR_{mD} ratios than for TR_{mIND} ratios. The range

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of ratios is also markedly larger for TR_{mD} than for TR_{mIND}. In particular, TR_{mD} codon ratios lie between 0.02 to 3.60 while TR_{mIND} codon ratios lie between 0.61 to 1.53, an over 50 fold difference (Figure 10 E and F; Dataset S9). We conclude that both amino acid and codon frequencies control TR_{mD} more strongly than they impact TR_{mIND} and do so by their effect on the rate of elongation by the ribosome. The differences in amino acid composition between highly abundant and less abundant proteins have a significant impact on translation rates, but the additional larger variation in the frequencies of individual codons plays a bigger role.

Discussion

We have presented a revised framework for determining the contribution of translation rates to the differences in protein expression between genes. Because translation rates partially correlate with mRNA abundance, it is not possible to provide a single metric to capture system-wide translational control. The R^2 coefficient of determination between translation rates and protein expression cannot measure translation's contribution because it mixes the contribution of translation with that of mRNA. Instead, to be consistent with the requirements of ANOVA the contributions of translation to the amplification exponent $b_{prot-RNA}$ and to $R^2_{prot-RNA}$ must be estimated separately. To achieve this, translation rates are decomposed into mRNA-abundance dependent and independent components, TR_{mD} and TR_{mIND} respectively. TR_{mD} determines $b_{prot-RNA}$, whereas TR_{mIND} and protein degradation together determine $R^2_{prot-RNA}$.

We find that in *S. cerevisiae* TR_{mD} represents ~20% of the variance in translation and results in an the amplification exponent $b_{prot-RNA}$ of 1.20 with a 95% confidence interval [1.14, 1.26], while TR_{mIND} constitutes the remaining ~80% of the variance in translation and explains ~5% of the variance in protein expression (Figure 6B). To overcome the difficulty of comparing the magnitude of contributions that are expressed by different, incommensurable metrics, we

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

suggest that the percent of the variance in translation that each explains be used. In other words, TR_{mIND} could be said to contribute 80 / 20 = 4 fold more to the control of protein levels than does TR_{mD}.

Our estimates for $b_{prot-RNA}$ are lower than that of the only previous study to assume mRNA-abundance dependent translational amplification (1.20 [1.14, 1.26] vs 1.69) (12). Because $b_{prot-RNA}$ is an amplification exponent for non-logged abundance data, this disagreement between estimates is large. $b_{prot-RNA}$ = 1.20 implies a range of mRNA abundances in the cell that is fifty fold larger than that implied by $b_{prot-RNA}$ = 1.69 (Dataset S2 and Figure 3). One of the two approaches that we used to estimate $b_{prot-RNA}$ is based on multiple protein and mRNA abundance scaling-standard datasets that were each produced using methods that employed internal concentration standards and should thus be properly scaled. Broad agreement is observed between scaling-standards from separate studies that used different methods (Figure 2). Our other estimate of $b_{prot-RNA}$ is based on the correlation between measured translation rate and mRNA abundance data. Our two independent estimates are similar (means 1.17 vs 1.22; Figure 3B and C), implying that they are reasonable. The prior estimate of $b_{prot-RNA}$ = 1.69, by contrast, used a Bayesian model to infer the scaling of true protein and true mRNA abundances from datasets that in some cases were produced by methods that yield biased scalings (Figure 4). The model had no guide for which data input was correctly scaled, and thus it had no way to determine a correct scaling. It was also previously claimed that the correlation between ribosome profiling data and mRNA abundances was consistent with $b_{prot-RNA}$ = 1.69 (12). Our analysis, however, indicates that this claim in effect assumes that true translation rates and true mRNA abundances correlate perfectly (see Results), an idea that is inconsistent with the available evidence.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Given estimates for TR_{mIND}, protein degradation and measurement error, we showed that it is possible to estimate $R^2_{prot-RNA}$ for the true abundances of proteins and mRNA. This approach suggests that $R^2_{prot-RNA}$ ~0.94 (Figure 6B). The highest previous estimate for the correlation between protein and mRNA levels was $R^2_{prot-RNA}$ = 0.86 (12). This estimate was based on modeled abundances for 5,854 protein-coding genes in *S. cerevisiae.* For 842 of these genes, however, either protein or mRNA abundance data was lacking; instead, values were imputed using a Bayesian model. When we limit the protein and mRNA abundances produced by the Bayesian model to the 5,045 genes for which empirically measured data is available, $R^2_{prot-RNA}$ = 0.80.

Our decomposition of translation rates, thus, provides an estimate for the combined contributions of translation and protein degradation that is ~3 fold lower than the smallest previous estimates based on measured protein and mRNA abundance data. Results from other approaches, though, support our estimate that $R^2_{prot-RNA}$ ~0.94. For example, ribosome profiling studies have found almost as strong a correlation between mRNA levels and the total number of protein molecules synthesized per gene (R^2 = 0.90) (27). In addition, translational regulation of specific transcripts in response to stress in *S. cerevisiae* is generally less than threefold and limited to a minority of genes (37,57). Finally, unlike animals, plants, and other fungi, *S. cerevisiae* lacks micro RNAs (58). The degree of transcript-specific translational regulation may be limited in this species, and so a particularly high correspondence between protein and mRNA abundances should be unsurprising.

These results should not be taken to suggest that translational control is unimportant, however. Translation and other steps, such as protein degradation, that do not strongly determine protein abundances, contribute to responsivity (3,11). For example, the response to environmental stimuli that change levels of specific mRNAs will be more rapid for those mRNAs

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that are inherently translated more quickly. Several metrics for control must be considered to properly appreciate the contribution of each step in regulating gene expression.

Quantifying the mechanisms that control translation

By considering which mRNA sequence features determine TR_{mD} and TR_{mIND}, we have also been able to provide insights into the mechanisms governing translation and the degree to which each exerts control. Extending detailed prior studies (27,29,30,38-56), we showed that nine sequence features can explain 80% of the variance in translation rates (Figure 7 and Supplementary Table S7). Importantly, the nine features do not all affect TR_{mD} and TR_{mIND} equally (Figure 9). TR_{mD}—and therefore the amplification exponent—is most strongly determined both by a Translation Initiation Control element (TICE) that spans nucleotides -35 to +28 and by the frequencies of amino acid and codons encoded in the open reading frame. TR_{mIND}, by contrast, is chiefly determined by the length of the protein coding region. These differences indicate that these two components of translation are under different selective pressures.

Translation initiation in eukaryotes has been proposed to be enhanced by a circularization event that brings the 5' and 3' ends of mRNAs into close proximity (55,59). The negative impact that longer ORFs have on translation rates results because this circularization appears less efficient for longer mRNAs than for shorter mRNAs (27,55,60). Given this, it can be readily understood why there might be dramatic differences in the degree to which ORF length specifies TR_{mD} versus TR_{mIND}. ORF length and mRNA abundance are under strong selective pressures that are unrelated to the control of translation rates. The relatively weak negative correlation of ORF length with TR_{mD} should thus be mostly determined by these other strong selective forces. In contrast, TR_{mIND} has no correlation with mRNA abundance, and thus

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the degree to which circularization efficiency affects translation will be fully reflected in the strong anti-correlation we observe between ORF length and TR_{mIND}.

Previous work indicates that A-rich sequences in the region -10 to -1 result in higher rates of translation initiation and that nucleotides between either +4 to +6 or +10 to +20 also play a role (38,40-42,61). Our analysis defining the TICE is consistent with this evidence, though suggests that the A-rich element is more extensive, stretching from nucleotides -35 to -1, and that all of the region from +4 to +28 is involved (Figure 8 and Supplementary Figure S2). The -35 to -1 region in highly translated mRNAs has a less folded RNA structure than in mRNAs translated at lower rates (Figure 8B) (39-41,44,51). Thus it is possible that the A-rich sequences act only by specifying unfoldedness and perhaps other aspects of structure, such as the degree of base stacking or chain flexibility. It has also been speculated, however, that A-rich sequences might stabilize the interaction of poly-A binding protein with the 5' UTR and thus enhance translation by mRNA circularization (62). The -35 to -1 portion of the TICE could thus act by two means. The TICE from +4 to +28 is not especially A-rich and instead shows a variety of location specific preferences for different bases between highly translated and poorly translated mRNAs (Figure 8A). Some of these sequence preferences may reflect evolutionary selection for protein function that are unrelated to the control of translation. Unfolded mRNA structure in the +4 to +28 region also correlates positively with translation rates, however, raising the intriguing possibility that the N-terminal nine or so amino acids could in part be selected because of the mRNA structures produced by their codons, rather than for their function within the protein (Figure 8B).

It has long been recognized that rates of translation elongation are higher for mRNAs whose frequency of codons more closely matches the relative abundance of tRNAs (27,30,45-49,52). Our analysis shows that both amino acid and codon frequencies are principally used to determine differences in translation elongation rates between differently abundant mRNAs (i.e.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

TR_{mD}) (Figures 9 and 10). These two features play a lesser role in modulating the mRNA independent variation in translation rates (i.e. TR_{mIND}) (Figures 9 and 10).

The length of poly-A tails and the degree of RNA folding in the CDS also show strong discrimination in their correlation with TR_{mD} versus with TR_{mIND} (Figure 9). The correlation of these features with TR_{mD}, however, unlike those of our other seven features, may not reflect a direct effect on translation but an impact on mRNA stability and hence mRNA abundance (Supplementary Table S8). The correlation of poly-A length and CDS RNA folding with TR may be entirely fortuitous. On the other hand, codon usage does have dramatic effects on both translation rate and mRNA stability in *S. cerevisiae*, with mRNAs that have codon frequencies optimized for rapid translation being the most stable (30,52). Our results confirm that our codon frequency feature correlates with RNA degradation ($R^2 = 0.21$, Supplementary Table S8). These results explain why codon usage is a strong determinant of TR_{mD} and why it has a less strong effect on TR_{mIND}. The control of both translation and mRNA turnover by this one sequence feature will inevitably lead to a correlation of TR and mRNA abundance and—as a consequence—an amplification exponent $b_{prot-RNA} > 1$. Any feature that impacts mRNA abundance will tend to explain TR_{mD} more so than TR_{mIND}, which is what we observe (Figure 9).

The three remaining features—5' UTR length, number of 5' UTR AUGs and 5'UTR folding—do not show significant differences in their correlation with TR_{mD} and TR_{mIND}. They contribute to both, establishing that TR_{mD} and TR_{mIND} are each specified by multiple sequence features.

Finally, because our model explains the bulk of the variance in translation, we can estimate the relative contributions of control at initiation versus control during elongation. 5' UTR length, number of 5' UTR AUGs, 5' UTR folding, and the TICE all likely affect initiation, not elongation, and collectively explain 42% of the variance in translation (Figure 11 A and B).

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Assuming that the length of the protein coding region also effects initiation rates, 58% of the variance in translation is controlled prior to elongation by the ribosome (Figure 11 C). Codon frequency controls elongation rate and determines 60% of the variance in translation (Figures 7 and 10 and Supplemental Table S7). When these six features are combined in a model, 80% of the variance in translation is explained. Initiation and elongation thus appear to share an equal role in controlling translation and to act in a substantially correlated manner (Figure 11 D). Slightly more than 60% of the control of initiation is fully correlated with elongation and vice versa: % initiation correlated with elongation = 66% = (58% + 60% - 80%) / 58% *100; % elongation correlated with initiation = 63% = (58% + 60% - 80%) / 60% *100. Initiation and elongation control features appear to act in tandem, tending to amplify the effect of each other to either both up regulate or to both down regulate rates.

Acknowledgements

We are indebted to David Weinberg, Premal Shah, Joshua Plotkin, and David Bartel for generously providing their ribosome profiling data prior to publication. Likewise, we are grateful to Craig Lawless, Robert Beynon, and Simon Hubbard for early access to their SRM MS protein abundance data. We acknowledge Kevin Weeks for pointing out that A rich RNA is likely unfolded. We thank Peter Bickel, Soile Keranen, Alisyn Nedoma, and Han Chen for thoughtful critiques of earlier drafts of this manuscript. JJL's work was supported by the start-up fund of the Department of Statistics at University of California, Los Angeles, a Hellman Fellowship from the Hellman Foundation, a PhRMA Foundation Research Starter Grant in Informatics, and NIH grant R01GM120507. Work at Lawrence Berkeley National Laboratory was conducted under U.S. Department of Energy Contract No. DE-AC02-05CH11231.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Li, J.J. and Biggin, M.D. (2015) Gene expression. Statistics requantitates the central dogma. *Science*, **347**, 1066-1067.
2. Albert, F.W. and Kruglyak, L. (2015) The role of regulatory variation in complex traits and disease. *Nat Rev Genet*, **16**, 197-212.
3. Liu, Y., Beyer, A. and Aebersold, R. (2016) On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, **165**, 535-550.
4. de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M. and Vogel, C. (2009) Global signatures of protein and mRNA expression levels. *Mol Biosyst*, **5**, 1512-1526.
5. Ghazalpour, A., Bennett, B., Petyuk, V.A., Orozco, L., Hagopian, R., Mungrue, I.N., Farber, C.R., Sinsheimer, J., Kang, H.M., Furlotte, N. *et al.* (2011) Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet.*, **7**, e1001393.
6. Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature*, **473**, 337-342.
7. Kristensen, A.R., Gsponer, J. and Foster, L.J. (2013) Protein synthesis rate is the predominant regulator of protein expression during differentiation. *Mol Syst Biol*, **9**, 689.
8. Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S. *et al.* (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature*, **513**, 382-387.
9. Ly, T., Ahmad, Y., Shlien, A., Soroka, D., Mills, A., Emanuele, M.J., Stratton, M.R. and Lamond, A.I. (2014) A proteomic chronology of gene expression through the cell cycle in human myeloid leukemia cells. *Elife*, **3**, e01630.
10. Li, J.J., Bickel, P.J. and Biggin, M.D. (2014) System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ*, **2**, e270.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

11. Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., Raychowdhury, R. *et al.* (2015) Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science*, **347**, 1259038.

12. Csardi, G., Franks, A., Choi, D.S., Airoldi, E.M. and Drummond, D.A. (2015) Accounting for experimental noise reveals that mRNA levels, amplified by post-transcriptional processes, largely determine steady-state protein levels in yeast. *PLoS Genet.*, **11**, e1005206.

13. Yassour, M., Pfiffner, J., Levin, J.Z., Adiconis, X., Gnirke, A., Nusbaum, C., Thompson, D.A., Friedman, N. and Regev, A. (2010) Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. *Genome Biol.*, **11**, R87.

14. Weiner, A., Chen, H.V., Liu, C.L., Rahat, A., Klien, A., Soares, L., Gudipati, M., Pfeffner, J., Regev, A., Buratowski, S. *et al.* (2012) Systematic dissection of roles for chromatin regulators in a yeast stress response. *PLoS Biol.*, **10**, e1001369.

15. Lipson, D., Raz, T., Kieu, A., Jones, D.R., Giladi, E., Thayer, E., Thompson, J.F., Letovsky, S., Milos, P. and Causey, M. (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol*, **27**, 652-658.

16. Miura, F., Kawaguchi, N., Yoshida, M., Uematsu, C., Kito, K., Sakaki, Y. and Ito, T. (2008) Absolute quantification of the budding yeast transcriptome by means of competitive PCR between genomic and complementary DNAs. *BMC Genomics*, **9**, 574.

17. Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K. and Weissman, J.S. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737-741.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

18. Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D.K., Noble, M., DeRisi, J.L. and Weissman, J.S. (2006) Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, **441**, 840-846.

19. Lawless, C., Holman, S.W., Brownridge, P., Lanthaler, K., Harman, V.M., Watkins, R., Hammond, D.E., Miller, R.L., Sims, P.F.G., Grant, C.M. *et al.* (2016) Direct and Absolute Quantification of over 1800 Yeast Proteins via Selected Reaction Monitoring. *Molecular and Cellular Proteomics*.

20. von der Haar, T. (2008) A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Syst Biol*, **2**, 87.

21. Picotti, P., Bodenmiller, B., Mueller, L.N., Domon, B. and Aebersold, R. (2009) Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell*, **138**, 795-806.

22. Firczuk, H., Kannambath, S., Pahle, J., Claydon, A., Beynon, R., Duncan, J., Westerhoff, H., Mendes, P. and McCarthy, J.E. (2013) An in vivo control map for the eukaryotic mRNA translation machinery. *Mol Syst Biol*, **9**, 635.

23. Carroll, K.M., Simpson, D.M., Eyers, C.E., Knight, C.G., Brownridge, P., Dunn, W.B., Winder, C.L., Lanthaler, K., Pir, P., Malys, N. *et al.* (2011) Absolute quantification of the glycolytic pathway in yeast: deployment of a complete QconCAT approach. *Mol Cell Proteomics*, **10**, M111 007633.

24. Brownridge, P., Lawless, C., Payapilly, A.B., Lanthaler, K., Holman, S.W., Harman, V.M., Grant, C.M., Beynon, R.J. and Hubbard, S.J. (2013) Quantitative analysis of chaperone network throughput in budding yeast. *Proteomics*, **13**, 1276-1291.

25. Smallbone, K., Messiha, H.L., Carroll, K.M., Winder, C.L., Malys, N., Dunn, W.B., Murabito, E., Swainston, N., Dada, J.O., Khan, F. *et al.* (2013) A model of yeast glycolysis based on a consistent kinetic characterisation of all its enzymes. *FEBS Lett*, **587**, 2832-2841.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

26. Mackenzie, R.J., Lawless, C., Holman, S.W., Lanthaler, K., Beynon, R.J., Grant, C.M., Hubbard, S.J. and Eyers, C.E. (2016) Absolute protein quantification of the yeast chaperome under conditions of heat shock. *Proteomics*, **16**, 2128-2140.

27. 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.

28. Christiano, R., Nagaraj, N., Frohlich, F. and Walther, T.C. (2014) Global proteome turnover analyses of the Yeasts S. cerevisiae and S. pombe. *Cell Rep*, **9**, 1959-1965.

29. 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.

30. 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.

31. Ahrne, E., Molzahn, L., Glatter, T. and Schmidt, A. (2013) Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics*, **13**, 2567-2578.

32. Holland, M.J. (2002) Transcript abundance in yeast varies over six orders of magnitude. *J. Biol. Chem.*, **277**, 14363-14366.

33. Dudley, A.M., Aach, J., Steffen, M.A. and Church, G.M. (2002) Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc. Natl. Acad. Sci. USA*, **99**, 7554-7559.

34. Smith, R.J. (2009) Use and misuse of the reduced major axis for line-fitting. *Am J Phys Anthropol*, **140**, 476-486.

35. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. and Weissman, J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218-223.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

36. McManus, C.J., May, G.E., Spealman, P. and Shteyman, A. (2014) Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.*, **24**, 422-430.

37. Gerashchenko, M.V., Lobanov, A.V. and Gladyshev, V.N. (2012) Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc. Natl. Acad. Sci. USA*, **109**, 17394-17399.

38. Nakagawa, S., Niimura, Y., Gojobori, T., Tanaka, H. and Miura, K. (2008) Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res.*, **36**, 861-871.

39. Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y. and Segal, E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467**, 103-107.

40. Robbins-Pianka, A., Rice, M.D. and Weir, M.P. (2010) The mRNA landscape at yeast translation initiation sites. *Bioinformatics*, **26**, 2651-2655.

41. Gingold, H. and Pilpel, Y. (2011) Determinants of translation efficiency and accuracy. *Mol Syst Biol*, **7**, 481.

42. Dvir, S., Velten, L., Sharon, E., Zeevi, D., Carey, L.B., Weinberger, A. and Segal, E. (2013) Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proc. Natl. Acad. Sci. USA*, **110**, E2792-2801.

43. 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.

44. Hinnebusch, A.G., Ivanov, I.P. and Sonenberg, N. (2016) Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science*, **352**, 1413-1416.

45. Dana, A. and Tuller, T. (2014) Mean of the typical decoding rates: a new translation efficiency index based on the analysis of ribosome profiling data. *G3 (Bethesda)*, **5**, 73-80.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

46. Sorensen, M.A. and Pedersen, S. (1991) Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J. Mol. Biol.*, **222**, 265-280.

47. Varenne, S., Buc, J., Lloubes, R. and Lazdunski, C. (1984) Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J. Mol. Biol.*, **180**, 549-576.

48. Brockmann, R., Beyer, A., Heinisch, J.J. and Wilhelm, T. (2007) Posttranscriptional expression regulation: what determines translation rates? *PLoS Comp. Biol.*, **3**, e57.

49. Guimaraes, J.C., Rocha, M. and Arkin, A.P. (2014) Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. *Nucleic Acids Res.*, **42**, 4791-4799.

50. Dacheux, E., Malys, N., Meng, X., Ramachandran, V., Mendes, P. and McCarthy, J.E. (2017) Translation initiation events on structured eukaryotic mRNAs generate gene expression noise. *Nucleic Acids Res.*

51. Hinnebusch, A.G. and Lorsch, J.R. (2012) The mechanism of eukaryotic translation initiation: new insights and challenges. *Cold Spring Harb Perspect Biol*, **4**.

52. Radhakrishnan, A. and Green, R. (2016) Connections Underlying Translation and mRNA Stability. *J. Mol. Biol.*, **428**, 3558-3564.

53. Arribere, J.A. and Gilbert, W.V. (2013) Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res.*, **23**, 977-987.

54. Rojas-Duran, M.F. and Gilbert, W.V. (2012) Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA*, **18**, 2299-2305.

55. Thompson, M.K. and Gilbert, W.V. (2016) mRNA length-sensing in eukaryotic translation: reconsidering the "closed loop" and its implications for translational control. *Curr Genet*.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

56. 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.

57. Payne, T., Hanfrey, C., Bishop, A.L., Michael, A.J., Avery, S.V. and Archer, D.B. (2008) Transcript-specific translational regulation in the unfolded protein response of Saccharomyces cerevisiae. *FEBS Lett*, **582**, 503-509.

58. Drinnenberg, I.A., Fink, G.R. and Bartel, D.P. (2011) Compatibility with killer explains the rise of RNAi-deficient fungi. *Science*, **333**, 1592.

59. Christensen, A.K., Kahn, L.E. and Bourne, C.M. (1987) Circular polysomes predominate on the rough endoplasmic reticulum of somatotropes and mammotropes in the rat anterior pituitary. *Am J Anat*, **178**, 1-10.

60. Arava, Y., Wang, Y., Storey, J.D., Liu, C.L., Brown, P.O. and Herschlag, D. (2003) Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. USA*, **100**, 3889-3894.

61. Cavener, D.R. and Ray, S.C. (1991) Eukaryotic start and stop translation sites. *Nucleic Acids Res.*, **19**, 3185-3192.

62. Gilbert, W.V., Zhou, K., Butler, T.K. and Doudna, J.A. (2007) Cap-independent translation is required for starvation-induced differentiation in yeast. *Science*, **317**, 1224-1227.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure Legends

Figure 1. Three scenarios explain the relationships between the steps in protein expression.

(A) Translation rates for all expressed genes are equal, as are protein degradation rates.

(B) Translation rates vary between genes but correlate perfectly with the amount of mRNA. Degradation rates for all proteins are constant.

(C) Translation and protein degradation rates vary but are uncorrelated with mRNA abundance.

Upper panels show the relationship between protein and mRNA levels; lower panels show the relationship between translation rates and mRNA levels. The coefficients of determination (R^2) and slopes (*b*) are indicated.

Figure 2. The slopes between Bayesian-model abundance data and scaling-standards.

(A) The four protein scaling-standards are compared to the Bayesian protein abundance data.

(B) The four mRNA scaling-standards are compared to the Bayesian mRNA data.

The colored lines are RuMA regressions that demonstrate slope $\hat{b}$. The lines have been shifted to give the same value at the origin, allowing ready comparison of the slopes. The dashed black lines show slope $b = 1$, the case where the standard deviations of the x and y values are equal and thus what would be seen if the data from the Bayesian model were scaled identically to a scaling-standard.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 3. Different predictions for slope $b_{prot-RNA}$. Red lines show the regression of protein on mRNA amount; dashed red lines show the 95% quantile confidence limits. Dashed black lines illustrate a slope of one.

(A) The RuMA regression between the uncorrected values for protein and mRNA amounts from the Bayesian model.

(B) The mean slope of the RuMA regressions for sixteen pair wise comparisons between our corrected versions of the Bayesian protein and mRNA abundance data.

(C) The true slope predicted by the relationship between the corrected Weinberg translation rate and mRNA abundance data. The predicted slope shown is 1 + the mean of the slopes of the OLS regressions for our corrected versions of the Weinberg data. The intercept in this panel was derived such that the total number of expected protein molecules per cell is the same as in panel B.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4. The RuMA slope between scaling-standards (y-axis) and the datasets input to the Bayesian model (x-axis).

(A) Protein data.

(B) mRNA data.

Each data point is the mean of the RuMA slopes between a single input dataset and each of the corresponding scaling-standards. The results are grouped by the method used to produce the input dataset, and the numbers in each group are indicated (N). The mean slope between the scaling-standards (y-axis) and the abundance estimates from the Bayesian model (x-axis) is shown by the solid black line. An RuMA slope of 1 is shown by the dashed black line, the case where the standard deviation of the dataset on the x-axis is equal to that of the mean of the scaling-standards.

Figure 5. The estimation of TR_{mD} and TR_{mIND} for a single gene. The linear regression between $\log_{10}$ translation rate data (y-axis) and $\log_{10}$ mRNA abundance data (x-axis) is shown by the red line (data from (27)). The data point for an example gene is highlighted in orange, whereas those for all remaining genes are shown in light blue circles. The gene specific values for $\log_{10}(TR_{mD})$ and $\log_{10}(TR_{mIND})$ are shown for the highlighted gene. The value for $\log_{10}(TR_{mD})$ is the intercept on the y-axis of a horizontal line that intercepts the regression at the mRNA abundance of the gene. The value for $\log_{10}(TR_{mIND}) = \log_{10}(TR) - \log_{10}(TR_{mD})$. Values are determined likewise for the remaining genes. Values for $\log_{10}(TR_{mIND})$ thus have both positive and negative values depending on if the data point lies above or below the regression. Values for $\log_{10}(TR_{mD})$ fall within the range of values for $\log_{10}(TR)$, all of which are negative.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 6. Contributions to the control protein expression.

(A) The maximum percentage contributions of estimates of mRNA abundance, protein degradation (PnD), and TR_{mIND} to the variance in measured levels of protein expression as well as the percent of the variance unexplained (Supplementary Tables S4 and S6).

(B) Left, the presumed percentage contributions of true mRNA abundance, protein degradation (PnD) and TR_{mIND} if the unexplained component in A is due to similar proportions of measurement error in each data class. Right, the mean of our estimates for the contribution of TR_{mD} to the amplification exponent $b_{prot-RNA}$. The dashed black line shows a slope of 1, the shaded area shows the increase in slope due to TR_{mD}.

Figure 7. mRNA sequences that explain translation rates. The R^2 coefficients of determination between nine mRNA sequence features and TR are shown (Supplementary Table S7 and Methods S4). A cartoon below shows to which mRNA region each feature maps. The TICE, CDS amino acid frequency and CDS codon frequency features are multi feature sets comprized of 14, 20 and 61 individual features respectively (Datasets S6 and S8). The other six are single features (Dataset S6).

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 8. The -35 to + 28 Translation Initiation Control Element (TICE)

(A) Position weight matrices (PWMs) for the 10% of mRNAs with the highest TR scores (top) and the 10% of mRNAs with the lowest TR scores (bottom). Sequence logos show the frequency of each nucleotide at each position relative the first nucleotide of the protein coding sequence (CDS) (Dataset S7).

(B) The mean predicted RNA folding energy (ΔG kcal/mol) of 35 nucleotide windows (y-axis). The x-axis shows the position of the 5' most nucleotide of each window. Windows representing every one nucleotide offset were calculated.

(C) The R^2 coefficient of determination between translation rates (TR) and PWM scores. PWMs of varying lengths were built from the sequences of the 10% of mRNAs with the highest TRs, and then log odds scores were calculated for all mRNAs that completely contained a given PWM. PWMs extending 5' from -1 in 5 nucleotide increments were tested (x-axis, right to left) and these were also extended 3' from +4 in 5 or 10 nucleotide increments (grey to black scale).

(D) The R^2 coefficients of determination between TICE mRNA sequence features and TR. PWMs corresponing to the three specified TICE mRNA regions (-31/-1, +4/+28 and -35/+28) were used to score each gene (PWM only). Alternatively, PWMs and the frequencies of a small subset of dinucleotides and/or trinucleotides were scored for each gene (PWM + di/tri nuc. freq.) (Datasets S6 and S8).

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 9. TR_{mD} and TR_{mIND} are differentially determined by mRNA sequences. The R^2 coefficients of determination between mRNA sequence features and TR_{mD} and TR_{mIND} are shown (Supplementary Table S7 and Methods S4). The Bonferroni corrected *p*-value testing if the correlations with TR_{mD} and TR_{mIND} are equal are given, with significant *p*-values shown in red. A cartoon below shows to which mRNA region each feature maps.

Figure 10. Amino acid and codon frequencies correlate with tRNA abundances. The frequencies of amino acid (AA) or codons in the CDS were determined separately for the 10% of genes with the highest scores for TR_{mD} or TR_{mIND} (top TR_{mD} or top TR_{mIND}) and for the 10% of genes with the lowest scores (bottom TR_{mD} or bottom TR_{mIND}) (Dataset S9).

(A and B) The coefficient of determination (R^2) for top and bottom amino acid or codon frequencies vs their cognate tRNA abundances (Supplementary Table S9). For amino acids, the frequencies of all cognate tRNAs for each amino acid were summed to give a combined tRNA abundance. The Bonferroni corrected *p*-value testing if the correlation of tRNA abundance with the top cohort is greater than that with the bottom cohort are given, with significant *p*-values shown in red.

(C – F) The ratios between the frequencies of amino acids or codons in the top cohort divided by those in the bottom cohort were determined (Dataset S9). Ratios > 1 thus indicate a higher frequency in the top TR_{mD} or top TR_{mIND} cohorts. Scatter plots are show between top/bottom frequency ratios and tRNA abundance along with the Pearson correlation coefficients (r). Bonferroni corrected *p*-values testing if the correlations are significant are also given, with significant *p*-values shown in red. Dashed vertical lines indicate a ratio of 1.

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 11. Progressively combining features predicts TR with increasing accuracy. Scatter plots comparing measured TR data to the results of four multi-variate models that predict TR. The R^2 coefficient of determination are also shown.

(A) A model combining features for the length of the 5' UTR, the number of open reading frames (ORFs) upstream of the initiating AUG, and the RNA folding energy of the 5' UTR.

(B) A model that combines the three features in A. with the TICE feature.

(C) A model that combines the four features in B. with the length of the CDS.

(D) A model that combines the five features in C. with the frequency of codons in the CDS.

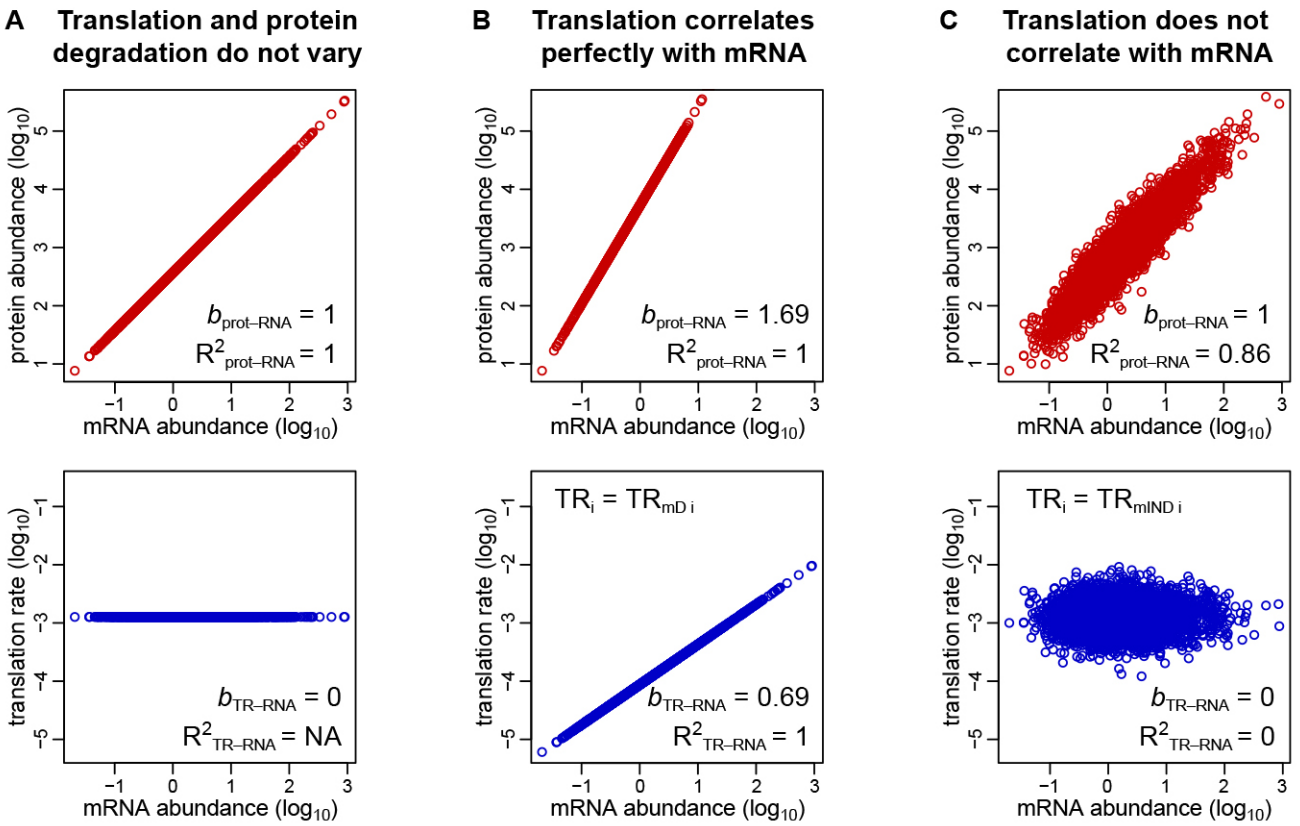
A
Translation and protein degradation do not vary
B
Translation correlates perfectly with mRNA
C
Translation does not correlate with mRNA
protein abundance (log10)
mRNA abundance (log10)
b prot–RNA = 1
R2 prot–RNA = 1
b prot–RNA = 1.69
R2 prot–RNA = 1
b prot–RNA = 1
R2 prot–RNA = 0.86
translation rate (log10)
b TR–RNA = 0
R2 TR–RNA = NA
TRi = TRmD i
b TR–RNA = 0.69
R2 TR–RNA = 1
TRi = TRmIND i
b TR–RNA = 0
R2 TR–RNA = 0

Fig. 1

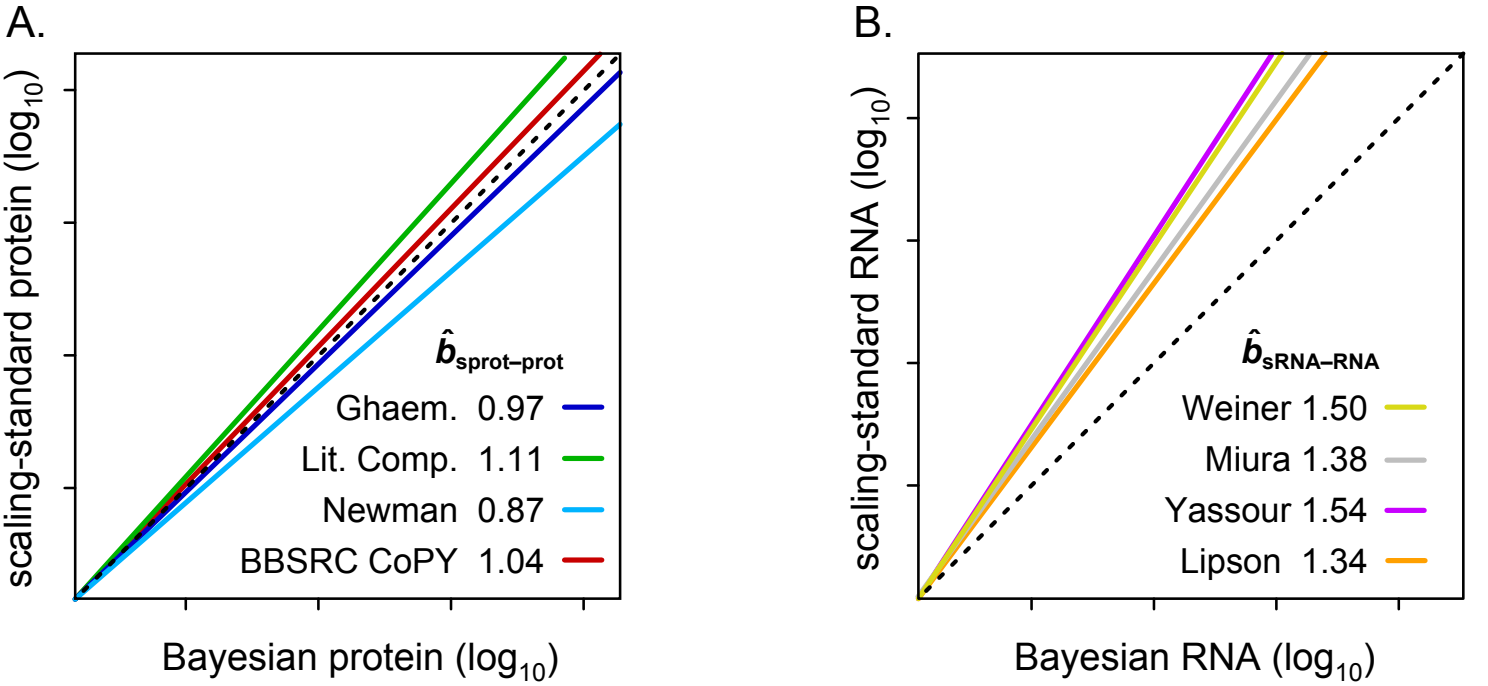
A.
scaling-standard protein (log10)
Bayesian protein (log10)
b̂sprot–prot
Ghaem. 0.97
Lit. Comp. 1.11
Newman 0.87
BBSRC CoPY 1.04
B.
scaling-standard RNA (log10)
Bayesian RNA (log10)
b̂sRNA–RNA
Weiner 1.50
Miura 1.38
Yassour 1.54
Lipson 1.34

Fig. 2

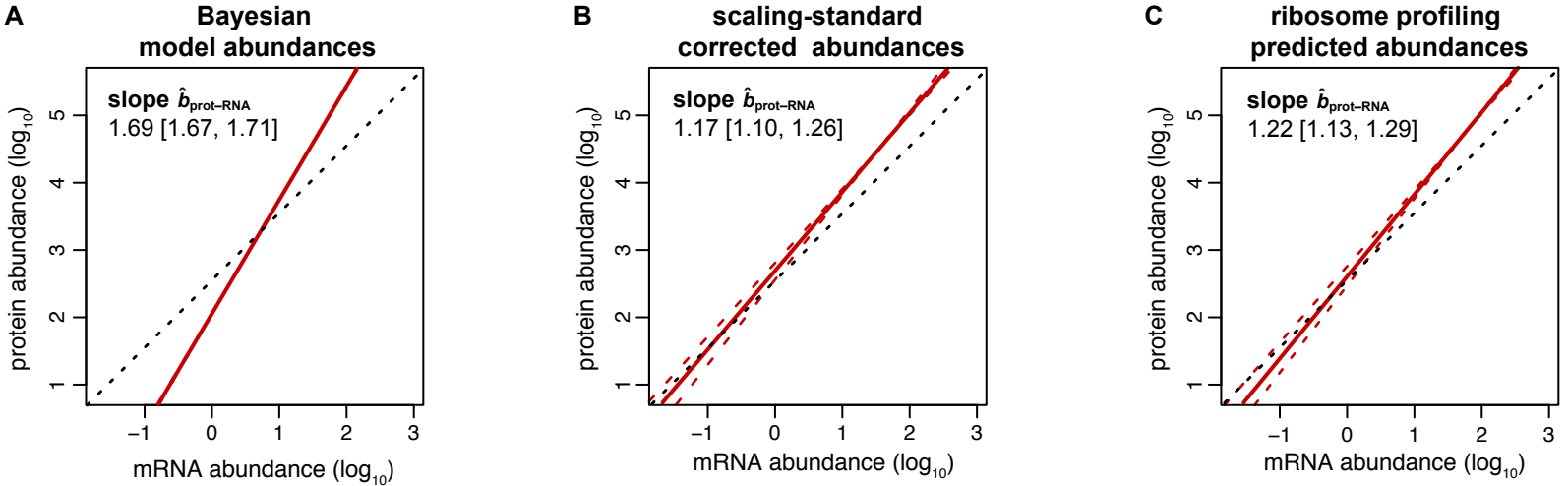
A
Bayesian
model abundances
slope $\hat{b}_{\text{prot–RNA}}$
1.69 [1.67, 1.71]
protein abundance ($\log_{10}$)
mRNA abundance ($\log_{10}$)
B
scaling-standard
corrected abundances
slope $\hat{b}_{\text{prot–RNA}}$
1.17 [1.10, 1.26]
protein abundance ($\log_{10}$)
mRNA abundance ($\log_{10}$)
C
ribosome profiling
predicted abundances
slope $\hat{b}_{\text{prot–RNA}}$
1.22 [1.13, 1.29]
protein abundance ($\log_{10}$)
mRNA abundance ($\log_{10}$)

Fig. 3

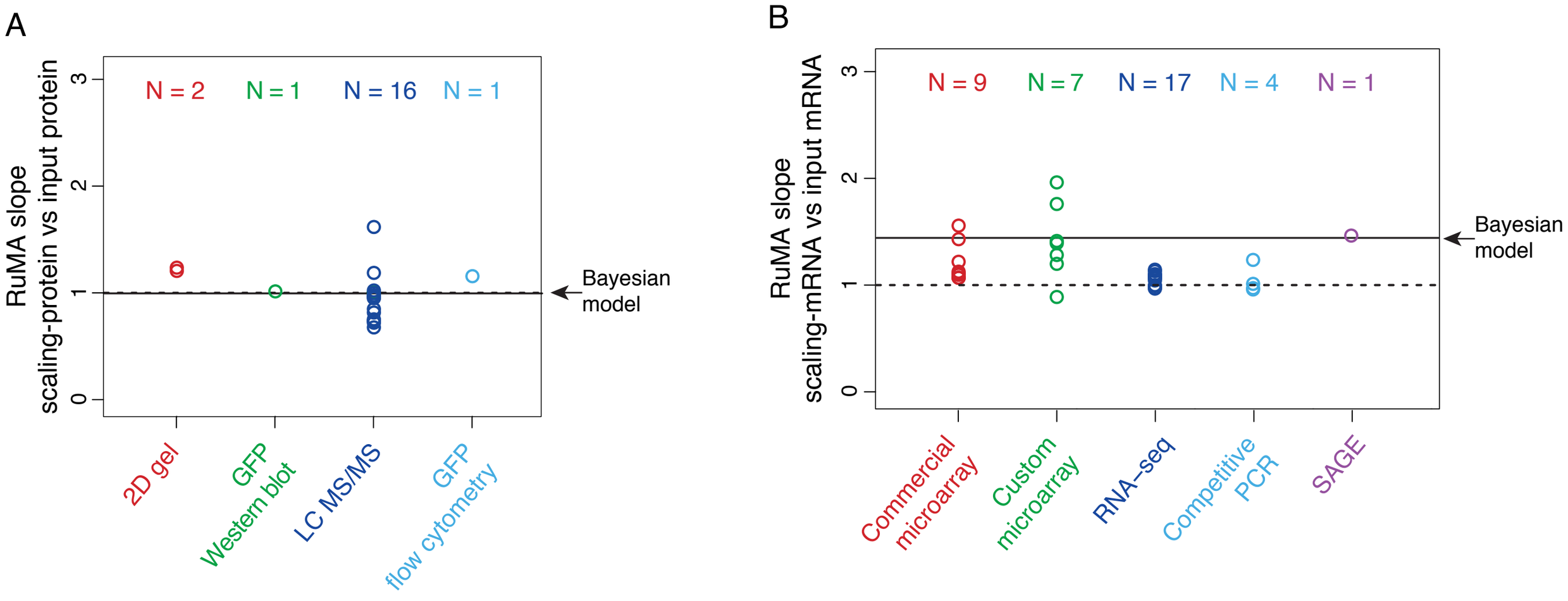
A
RuMA slope
scaling-protein vs input protein
N = 2
N = 1
N = 16
N = 1
Bayesian
model
2D gel
GFP
Western blot
LC MS/MS
GFP
flow cytometry
B
RuMA slope
scaling-mRNA vs input mRNA
N = 9
N = 7
N = 17
N = 4
N = 1
Bayesian
model
Commercial
microarray
Custom
microarray
RNA-seq
Competitive
PCR
SAGE

Fig. 4

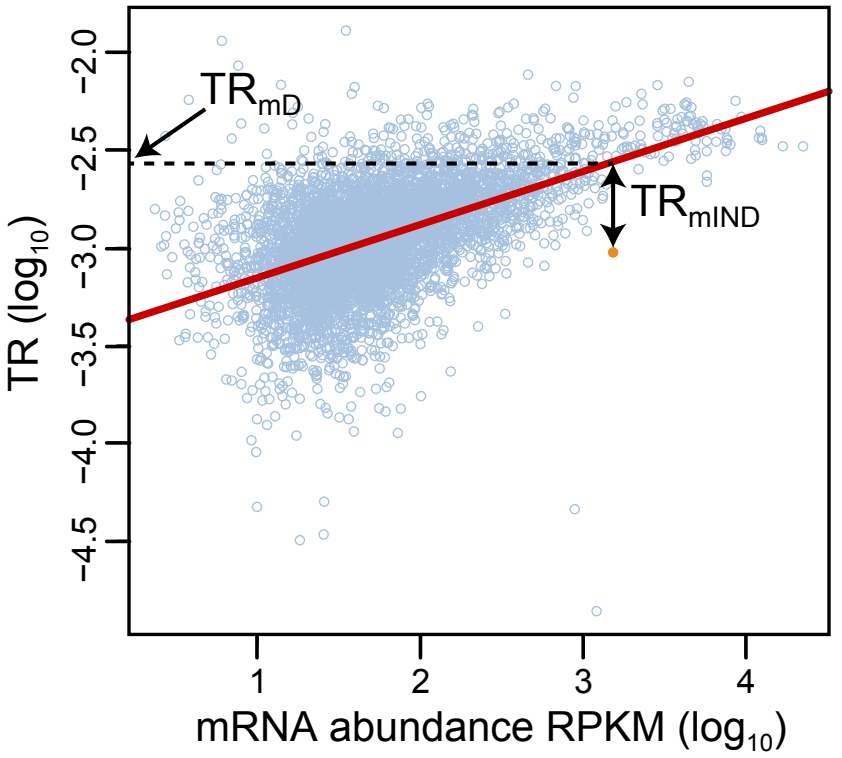
TR_mD
TR_mIND
TR (log10)
−2.0
−2.5
−3.0
−3.5
−4.0
−4.5
1
2
3
4
mRNA abundance RPKM (log10)

Fig. 5

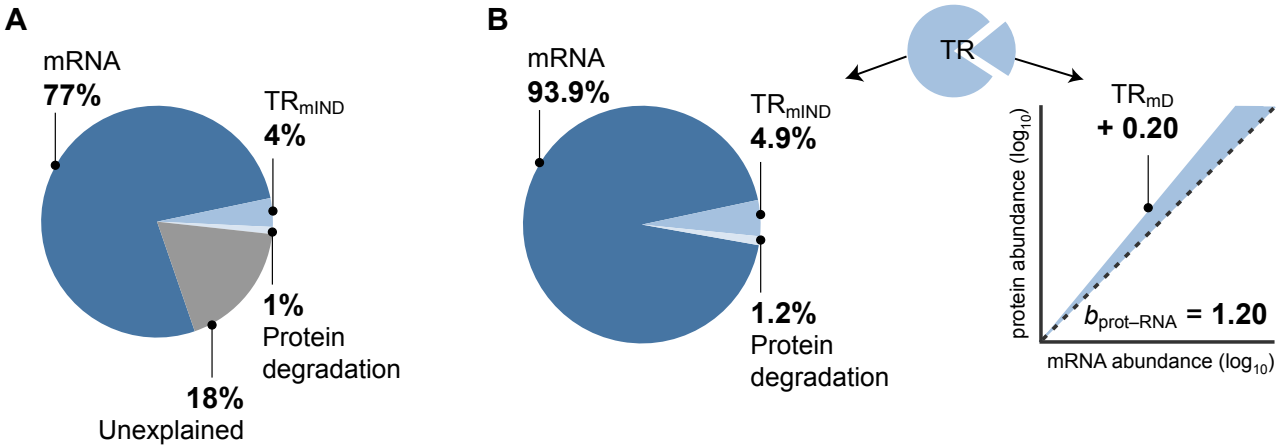
A
mRNA
77%
TRmIND
4%
1%
Protein
degradation
18%
Unexplained
B
mRNA
93.9%
TRmIND
4.9%
1.2%
Protein
degradation
TR
TRmD
+ 0.20
protein abundance (log10)
bprot–RNA = 1.20
mRNA abundance (log10)

Fig. 6

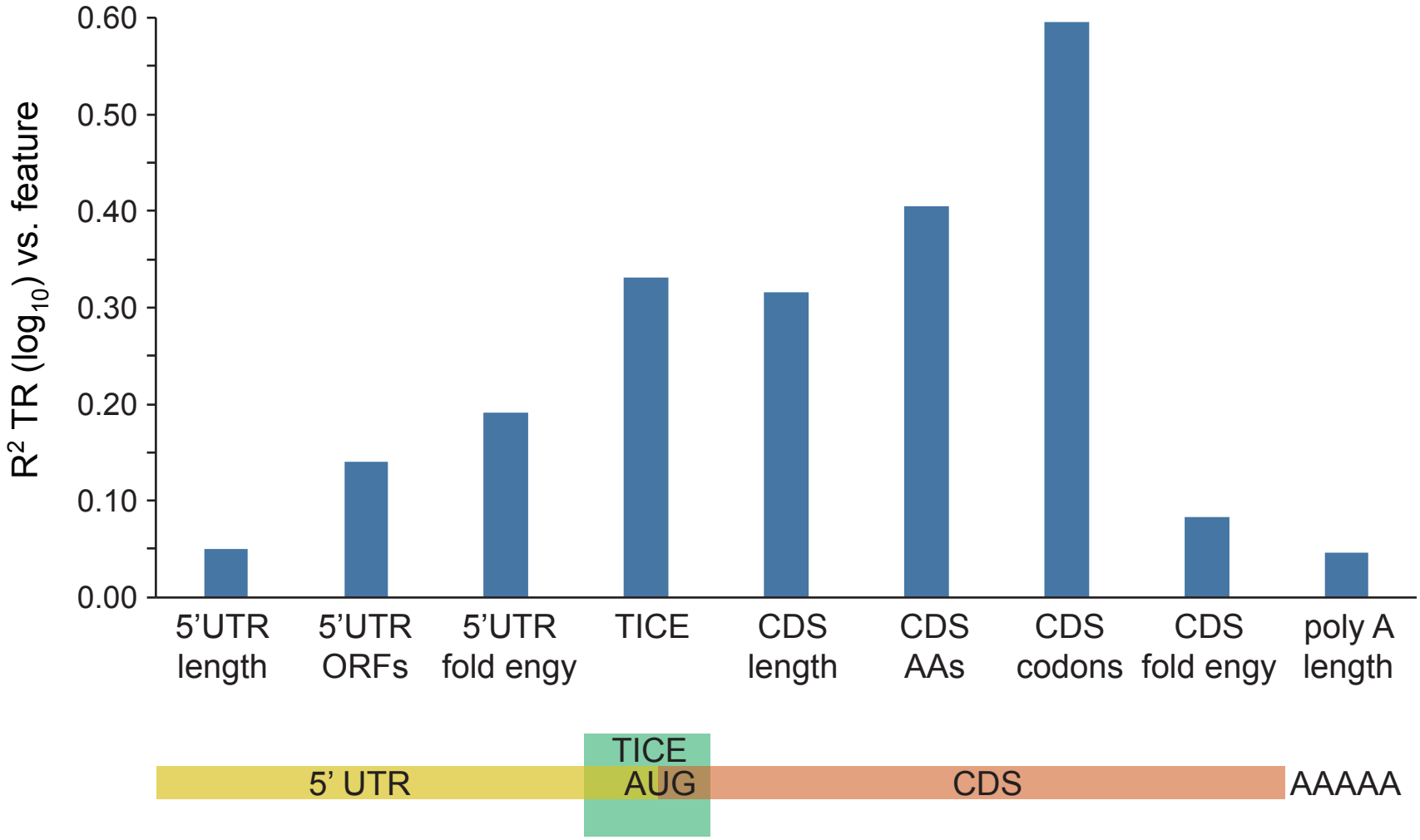
0.60
0.50
0.40
0.30
0.20
0.10
0.00
R2 TR (log10) vs. feature
5'UTR length
5'UTR ORFs
5'UTR fold engy
TICE
CDS length
CDS AAs
CDS codons
CDS fold engy
poly A length
TICE
5' UTR
AUG
CDS
AAAAA

Fig. 7

bioRxiv preprint doi: https://doi.org/10.1101/116913; this version posted July 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

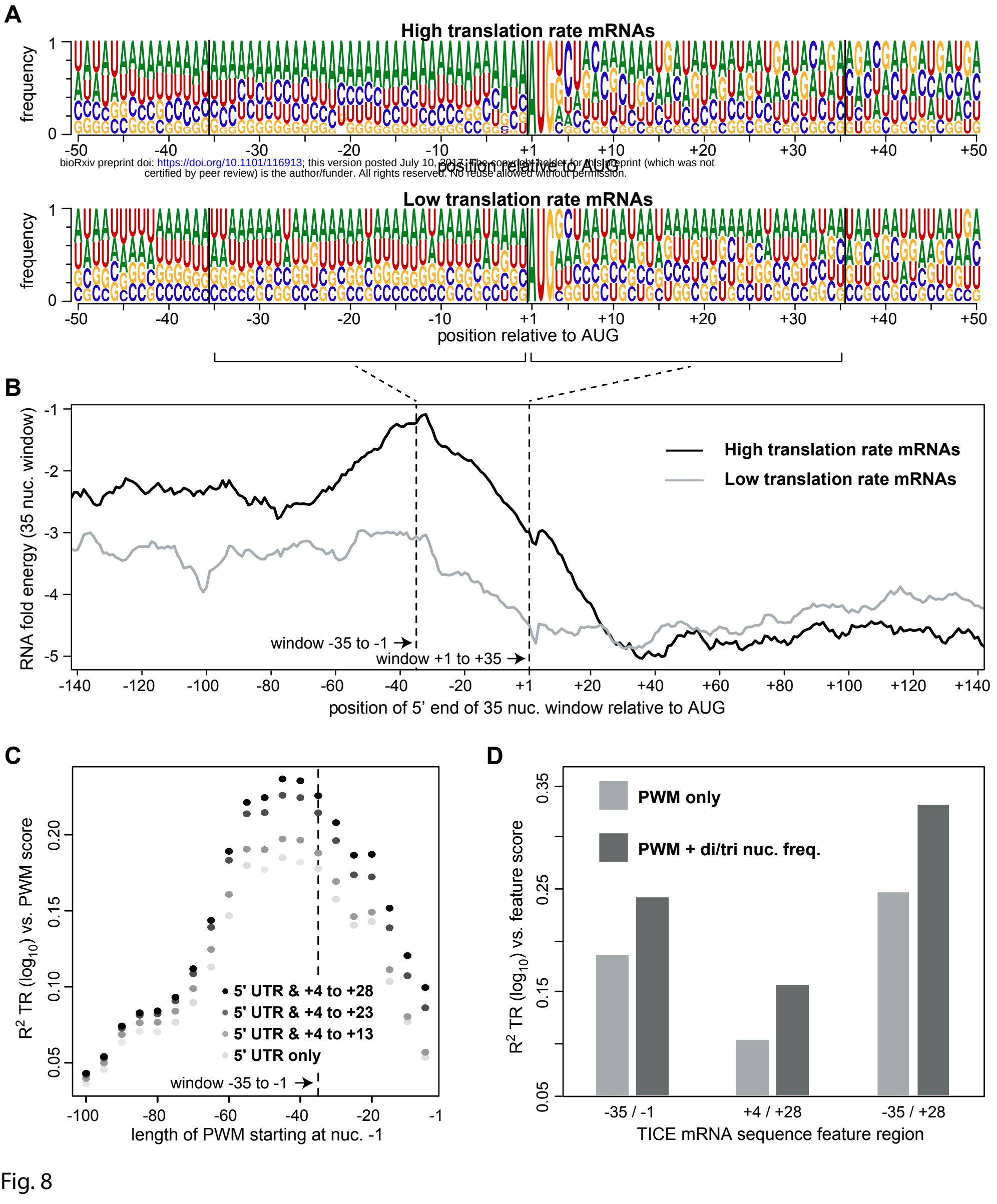

Fig. 8

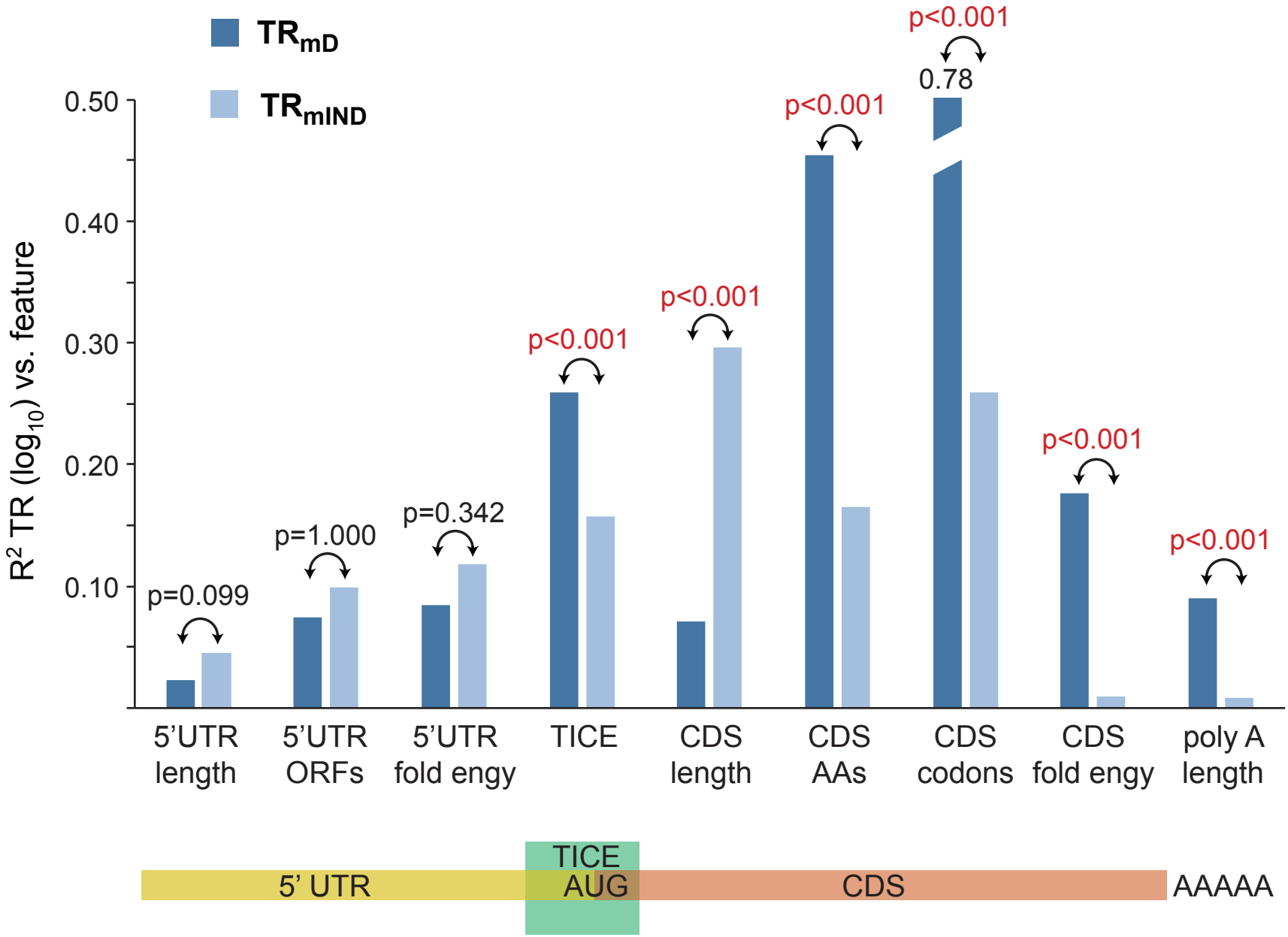
TRmD
TRmIND
R2 TR (log10) vs. feature
0.50
0.40
0.30
0.20
0.10
p=0.099
p=1.000
p=0.342
p<0.001
p<0.001
p<0.001
0.78
p<0.001
p<0.001
p<0.001
5'UTR length
5'UTR ORFs
5'UTR fold engy
TICE
CDS length
CDS AAs
CDS codons
CDS fold engy
poly A length
TICE
5' UTR
AUG
CDS
AAAAA

Fig. 9

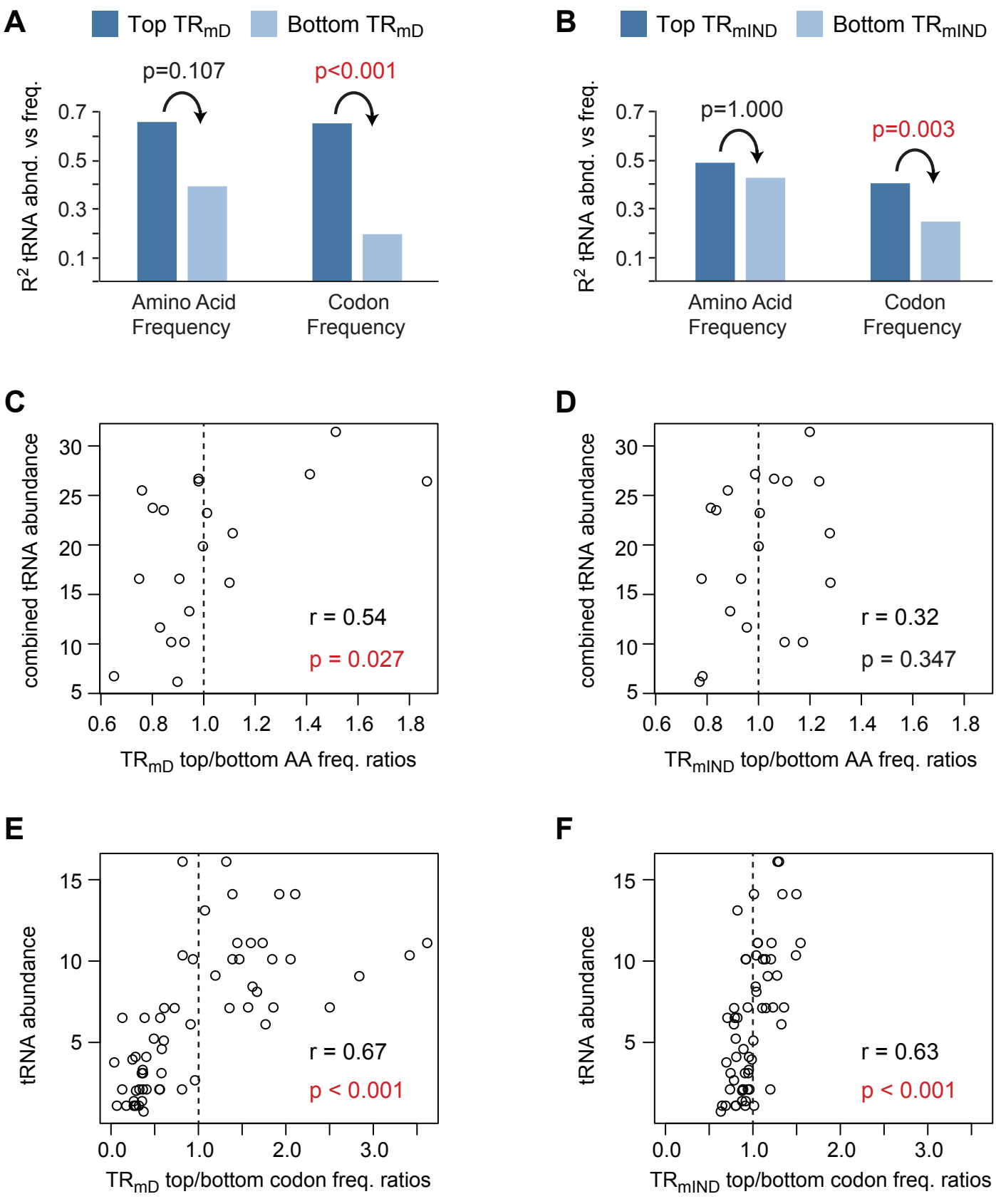
A
Top TRmD
Bottom TRmD
p=0.107
p<0.001
R2 tRNA abnd. vs freq.
0.7
0.5
0.3
0.1
Amino Acid Frequency
Codon Frequency
B
Top TRmIND
Bottom TRmIND
p=1.000
p=0.003
R2 tRNA abnd. vs freq.
0.7
0.5
0.3
0.1
Amino Acid Frequency
Codon Frequency
C
combined tRNA abundance
30
25
20
15
10
5
r = 0.54
p = 0.027
0.6
0.8
1.0
1.2
1.4
1.6
1.8
TRmD top/bottom AA freq. ratios
D
combined tRNA abundance
30
25
20
15
10
5
r = 0.32
p = 0.347
0.6
0.8
1.0
1.2
1.4
1.6
1.8
TRmIND top/bottom AA freq. ratios
E
tRNA abundance
15
10
5
r = 0.67
p < 0.001
0.0
1.0
2.0
3.0
TRmD top/bottom codon freq. ratios
F
tRNA abundance
15
10
5
r = 0.63
p < 0.001
0.0
1.0
2.0
3.0
TRmIND top/bottom codon freq. ratios

Fig. 10

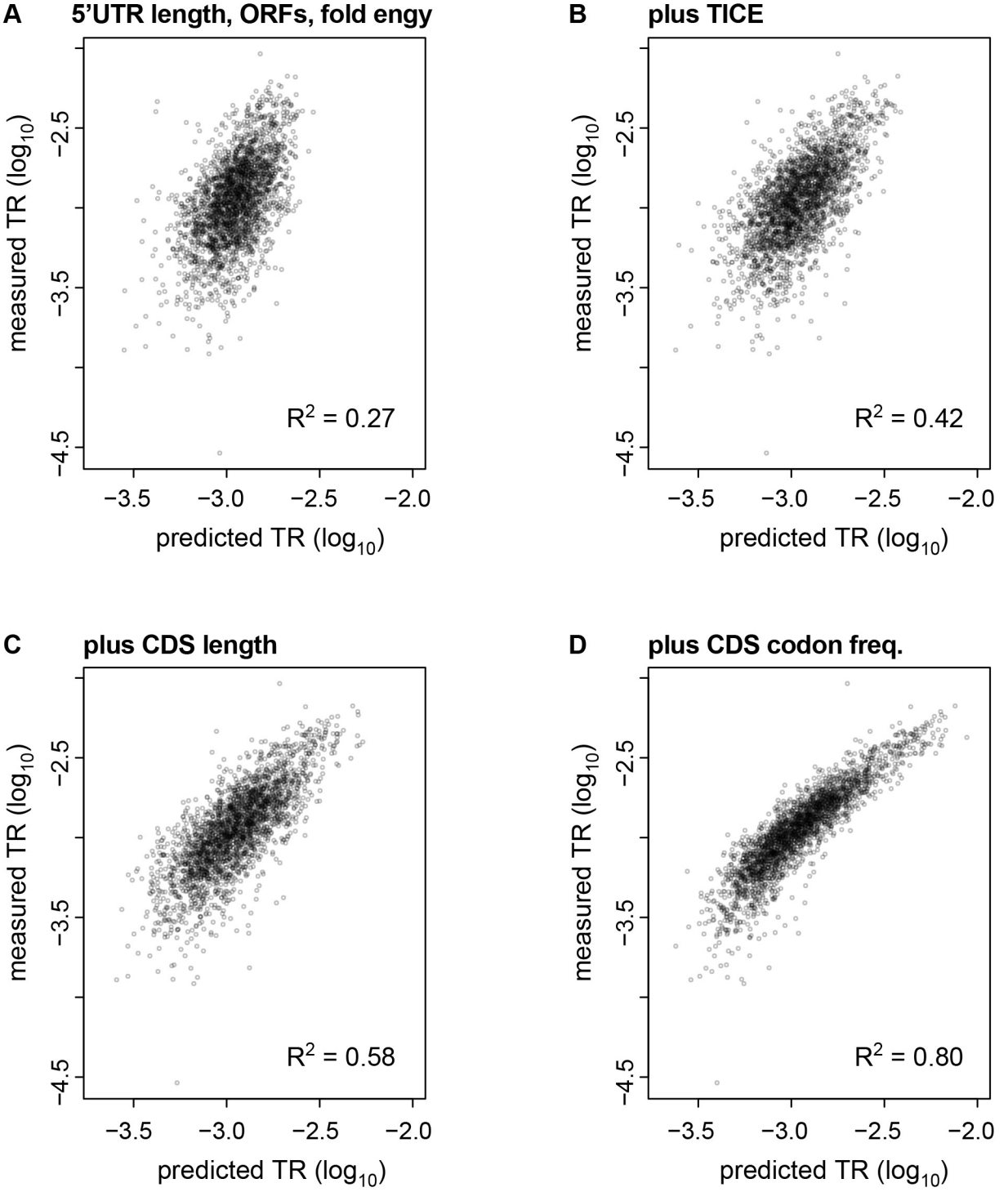
A
5'UTR length, ORFs, fold engy
B
plus TICE
C
plus CDS length
D
plus CDS codon freq.
measured TR (log10)
predicted TR (log10)
−4.5
−3.5
−2.5
−3.5
−3.0
−2.5
−2.0
R2 = 0.27
R2 = 0.42
R2 = 0.58
R2 = 0.80

Fig. 11