The effect of tRNA levels on decoding times of mRNA codons
Supplementary File

Authors: Alexandra Dana[1] and Tamir Tuller[1]*.
[1]The Department of Biomedical Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel
*To whom correspondence should be addressed. TT: tamirtul@post.tau.ac.il

Supplementary Methods

Reconstructing ORFs ribosome profiles of the analyzed organisms

S. cerevisiae ribosomal profiles were reconstructed using the data published in the GEO database, accession number GSE13750 (GSM346111, GSM346114) (1). *C. elegans* read count profiles of genes expressed in the L4 larval stage were built from Illumina sequencing results (NCBI Sequence Read Archive, accession number SRR52883) (2). *E. coli* and *B. subtilis* profiles were built from the published Illumina sequencing results (GEO database, accession number GSE35641) (3).

Prior to processing, the attached linker or poly-A tail was removed from the processed fragments. To reconstruct the ribosomal profiles of the different genes, fragments were aligned to known exons and spliced junctions, using the Bowtie software (4), allowing up to two mismatches. The location of the ribosome's P site relative to the 5' end of each fragment was determined in a similar manner described in previous studies (1,5) by calculating the offset between the 5' end of the most upstream mapped fragments to the initiation site. Previous studies showed that this offset slightly differs as function of the fragments length (1,5), therefore in this study also the location of the P site relatively to the fragments 5' end was determined as function of the fragment's length. The location of the A site was therefore determined as the location of P site shifted downstream by three nucleotides.

Next, we have built a RC profile for each gene by increasing for each mapped fragment to the exon/spliced junction the counter of the nucleotides corresponding to the fragment's A site. Specifically, to overcome multiple mappings of a single fragment to an exon/spliced junction, the ribosomal RC profiles were built in two steps: in the first iteration only fragments aligning to a single location were mapped. For each successfully mapped fragment, the RC of the relevant A site (3 nucleotides) were increased by one. In the second iteration, for all fragments aligning to more than one location, the mean RC in the surrounding of the possible aligning locations was calculated (10 nt before and after the location of the A site, using only RC obtained from the first iteration); let RCM_i denote the mean RC in the surrounding candidate location i; a fragment was mapped to one of these locations, with the probability of

$$\frac{RCM_i}{\sum_i RCM_i}$$

where i depicts the index of a possible mapping location. The RC of relevant exons and spliced junctions were united to create for each gene transcript its ribosomal profile. Ribosome profiles at codon resolution were calculated by averaging the RC of each three non-overlapping consecutive nucleotides. To increase the robustness of the data, for *S. cerevisiae*, *E. coli* and *B. subtilis* ribosome profiles of genes from different replications of the experiment were averaged together.

S. cerevisiae ribosome protected fragments were polyadenylated to generate a primer site for first-strand cDNA synthesis(1). Therefore, in this analysis the processed fragments were first removed from their poly A tail and fragments shorter than 20 bases were discarded (1). To filter fragments originating from rRNA (reference genome S228c, genes transcript coordinates were taken from the Saccharomyces Genome Database - http://www.yeastgenome.org) fragments were first aligned to rRNA transcripts using the Bowtie software (4). Fragments that failed to align to rRNA genes were further processed.

The location of the most upstream mapped fragment was defined as the location of the first (upstream) **r**ead **c**ount (**RC**) peak, 10-25 nt before the beginning of the ORF. Fragment lengths that failed to show a clear single peak were discarded. According to this analysis, in case of *S. cerevisiae* the location of the A site (offset from P site + 3 nt) was determined to start of 15-16 nt downstream of the 5' end of the fragment.

C. elegans footprints were attached to a linker (AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGTGATATC) rather than polyadenilated (5). Therefore, in this study the linker was first detected and removed from the fragments in the following manner: the 5' end of the linker was estimated to be located between nucleotides 20-36 of each processed fragment. Next, the Hamming distance between the estimated linker and the published linker was calculated (in terms of number of different nucleotides); a valid linker was accepted if this distance differed by up to two nucleotides. If no valid linker was found, the fragment was rejected (6). Fragments that were successfully aligned to rRNA and tRNA transcripts were discarded (genes coordinates were downloaded from the UCSC genes data set, using the *C. elegans* WS220/ce10 genomic strain). The remaining ones were mapped to exons and spiced junctions of protein coding genes. For this organism also, the P site was determined according to the fragment's length. Fragments that failed to show a clear single peak 10-25 nt before the beginning of the ORFs were discarded. These results indicate that the A site of *C. elegans* is located 14-15 nt from the 5' of the fragments.

E. coli and ***B. subtilis*** fragments were first removed from their attached linker (CTGTAGGCACCATCAAT) as described for *C. elegans*, allowing up to one difference between the estimated linker and the real linker (due to its shorter length). The first nt of each fragment was ignored, as it frequently represents an un-templated addition during reverse transcription (7) and fragments shorter than 20 nt were discarded. Fragments were then mapped against the *E. coli* K-12 MG1655 and *B. subtilis*

str. 168 reference genome strains, accordantly. To filter fragments originating from rRNA or tRNA, those were first aligned to rRNA and tRNA transcripts (transcripts coordinates for both organisms were downloaded from the Biomart database (www.biomart.org)). Those that were successfully mapped were discarded and the remaining ones were mapped to transcripts of protein coding genes. Fragments that failed to show a clear single peak 10-25 nt before the beginning of the ORFs were discarded. The results of this analysis indicate that the A site of *E. coli* is located 16-24 nt from the 5' of the fragments while the A site of *B. subtilis* is located 19-25 nt relatively to the 5' of the fragments.

In addition, it is possible that the footprint lengths have high variability in the analyzed prokaryotes due to the usage of micrococcal nuclease instead of RNase I. To show that this fact to do not affect our results, the P site for *E. coli* and *B*. subtilis was redefined by introducing a small offset of three/six/nine nucleotides. The analyses were then repeated. The results (see Tables S9, S10) demonstrate that with such an offset the correlation between mu and tRNA copy number or tAI disappear, supporting the conjecture that the estimated P offset in this study is biologically meaningful.

Understanding the relation between the estimated parameters of the EMG model

The λ in the EMG distribution of a codon NFC models the skewness in the NFC distribution due to ribosomal jams caused by different codon translation times and translational pauses.

To demonstrate that the λ parameter is related to the skewness of the NFC distribution (mathematical definition of skewness appears in the Methods section) and to estimate the type/direction of the relation between the NFC skewness and lambda in the analyzed regime we performed the following analysis: we created an artificial NFC distribution where the μ / σ sigma parameters are set as the mean value of the estimated *B. subtilis* parameters using the EMG model (μ = 0.0075; σ = 0.0042) and changed only the λ parameter within the range of 0.6 ..4.9 (mean λ value of *B. subtilis* is 4.65). Figure S09 shows a positive correlation between λ and the skewness of a random variable with EMG distribution in this region.

Thus we expect that faster codons which have low μ values to be more skewed as a result of ribosomal jams, therefore having bigger λ values, *i.e.* we expect a negative correlation between the μ and λ parameters. For example, in the TASEP simulation we got a correlation between μ and lambda of r = -0.219, p = 0.000785. In summary, ribosomal jamming are expected to affect each codon differently; this corresponds to different skewing of the NFC distribution which is modelled by the λ parameter.

When validating this assumption, the following correlations between μ and λ were found in the four organisms: *E. coli*: r = -0.259 (p = 0.044); *B. subtilis*: r = -0.627 (p = 1.40835e-07); *C. elegans*: r = -0.353 (p = 0.0055); *S. cerevisiae*: r = -0.074 (p = 0.57). These results indicate that in three out of four organisms

the λ value is indeed inversely correlated to the μ parameter. The calculated correlations between the estimated lambda values and tAI were however not/borderline significant: *E. coli*: r = 0.18, (p = 0.17); *B. subtilis*: r= 0.16, (p = 0.21); *C. elegans*: r = -0.20, (p = 0.12); *S. cerevisiae*: r = -0.33, (p = 0.01).

Figures

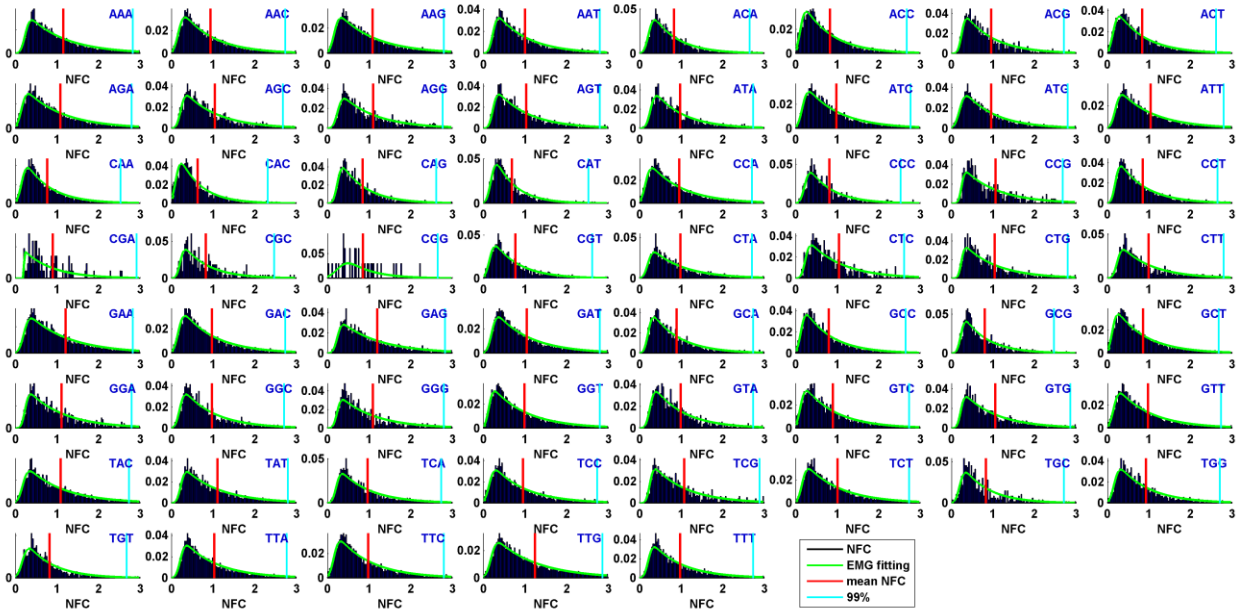

Figure S1. Codons NFC distributions of *S. cerevisiae.* Each subplot presents the NFC distribution of a codon (inline text). Black histogram: NFC values; green: EMG fitting; red: mean NFC values; cyan: 95 percentile.

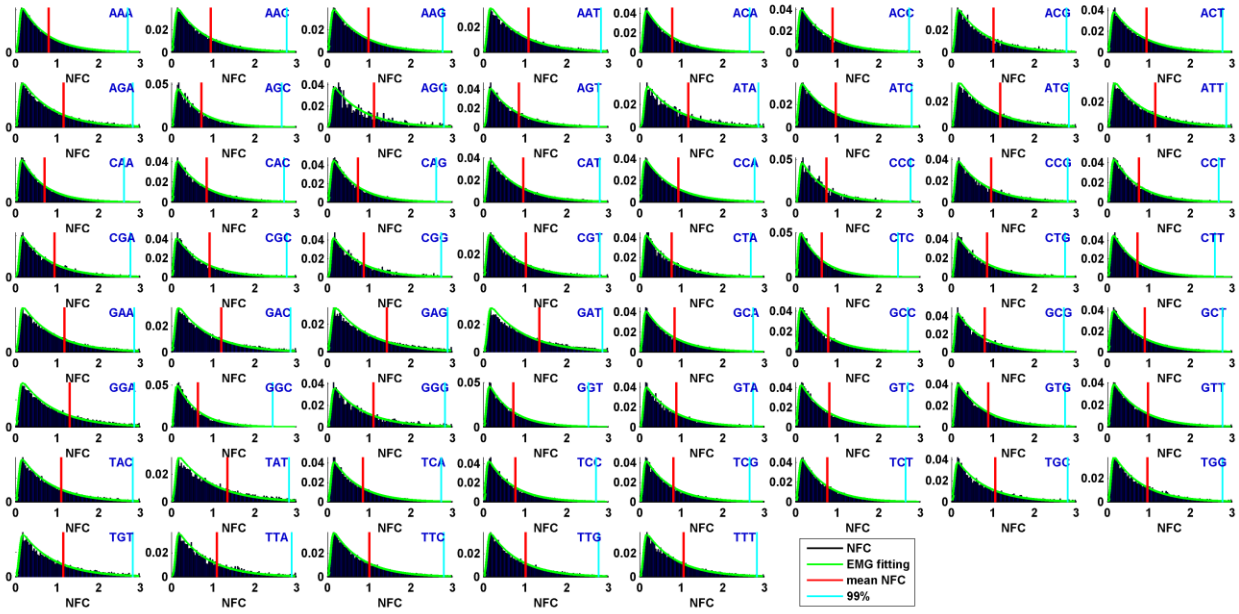

Figure S2. Codons NFC distributions of *C. elegans*. Each subplot presents the NFC distribution of a codon (inline text). Black histogram: NFC values; green: EMG fitting; red: mean NFC values; cyan: 95 percentile.

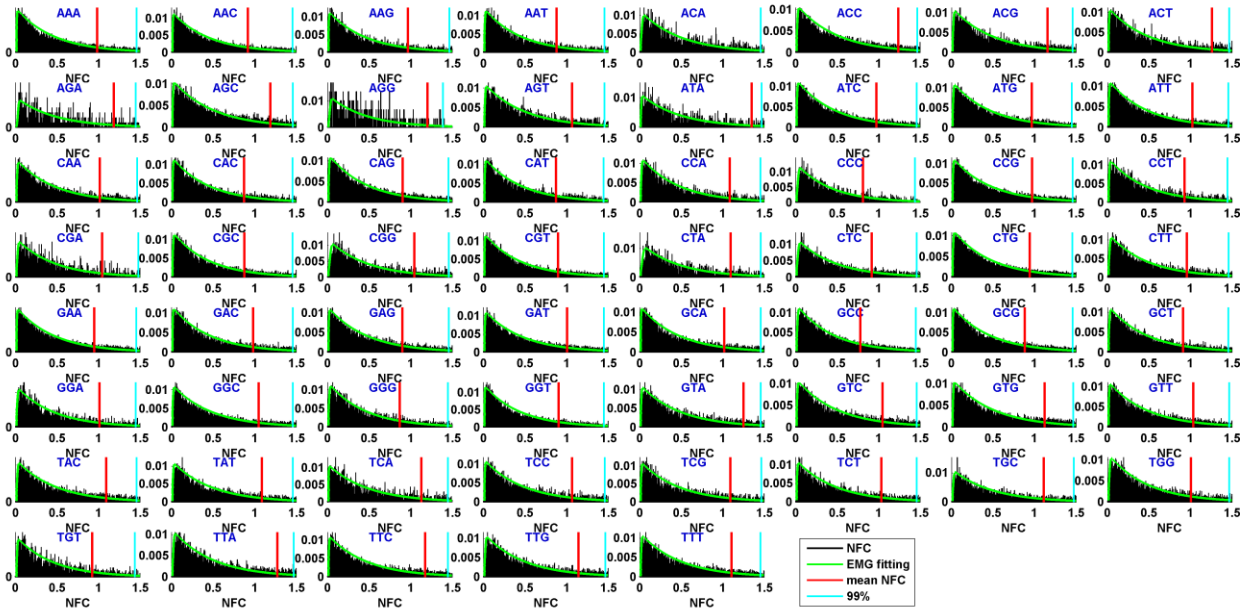

Figure S3. Codons NFC distributions of *E. coli*. Each subplot presents the NFC distribution of a codon (inline text). Black histogram: NFC values; green: EMG fitting; red: mean NFC values; cyan: 95 percentile.

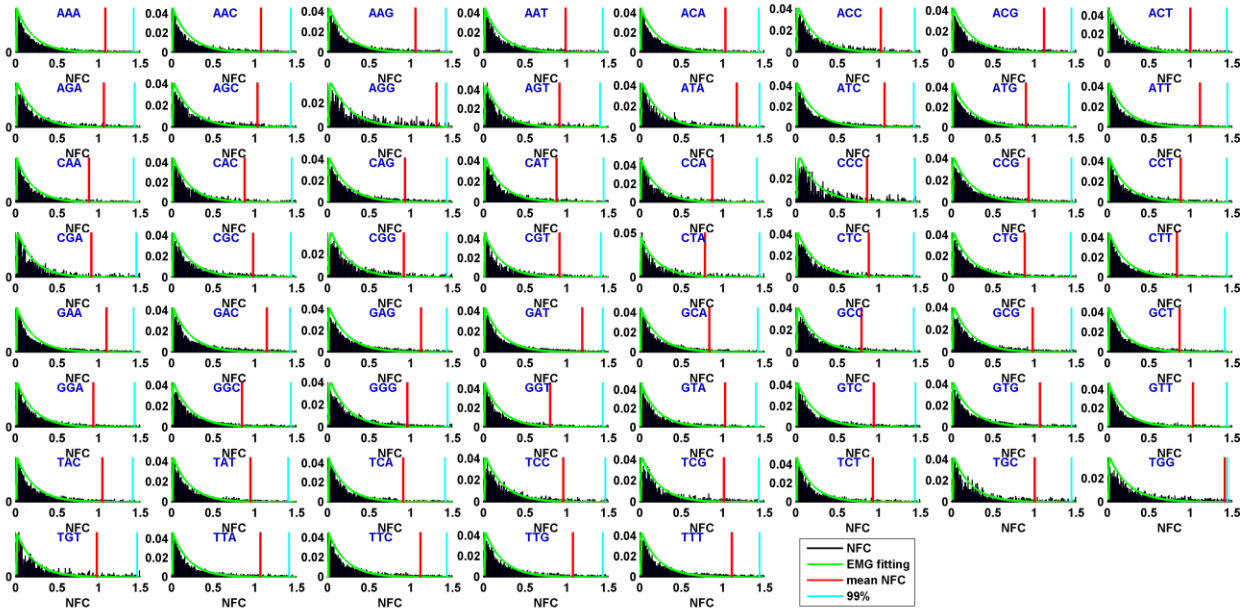

Figure S4. Codons NFC distributions of *B. subtilis*. Each subplot presents the NFC distribution of a codon (inline text). Black histogram: NFC values; green: EMG fitting; red: mean NFC values; cyan: 95 percentile.

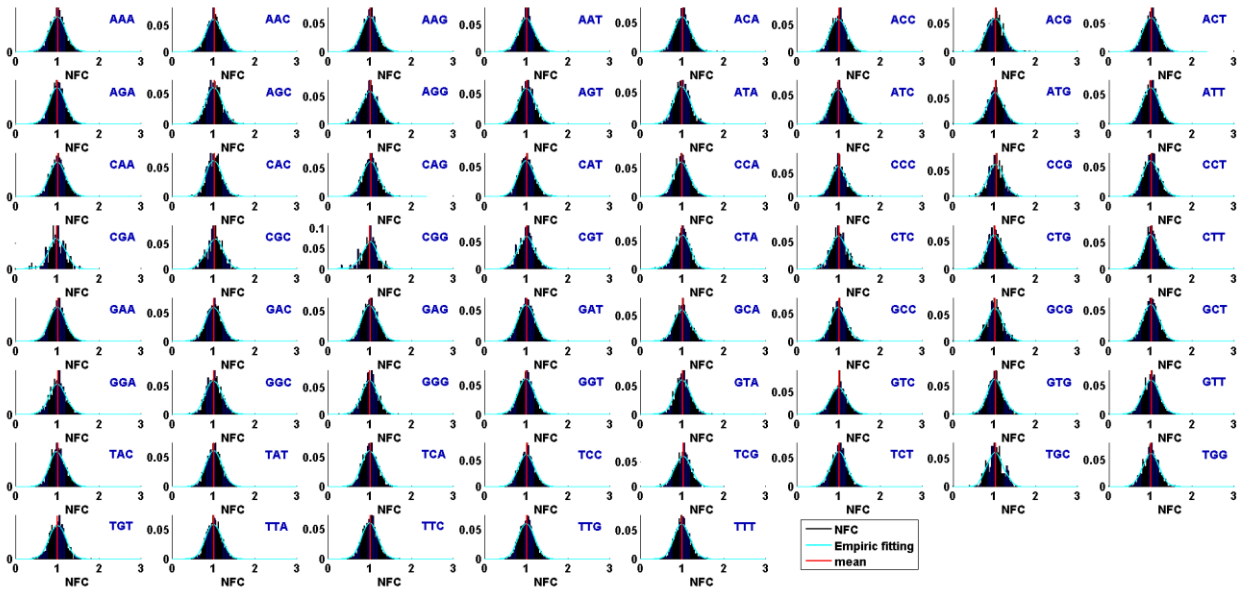

Figure S5. Demonstrating that NFC distributions are normally distributed. *S. cerevisiae* ribosome profiles were simulated using the TASEP process. All codon decoding times were set to be exponentially distributed with a rate $\psi = 1$ and initiation rate was set to 0.3. Ribosome profiles of each gene were simulated using 500 mRNA copies. The black histograms depict NFC values of each codon type, collected from all simulated genes; the cyan curve represents the empirical fitting of the NFC distributions, while the red line represents the mean NFC value for each codon type. NFC values are normally distributed (under KS-test: mean p value = 0.27)

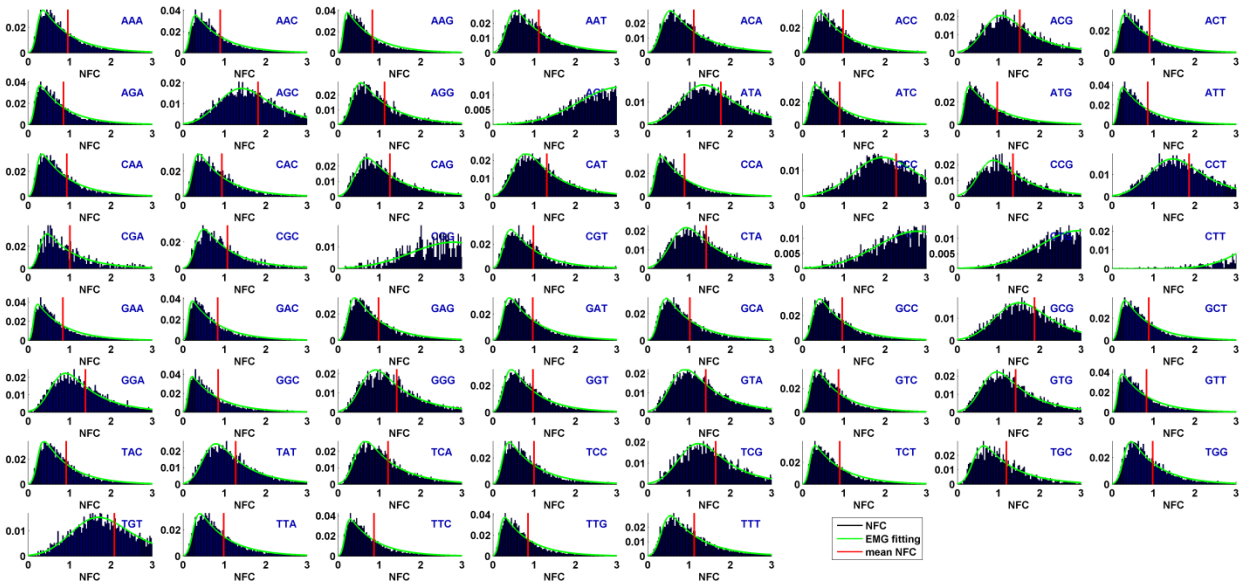

Figure S6. *S. cerevisiae* ribosomal profiles simulated using the TASEP process. The black histograms depict NFC values of each codon type, collected from all simulated genes; the green curve represents the EMG fitting of the NFC distributions, while the red line represents the mean NFC value for each codon type.

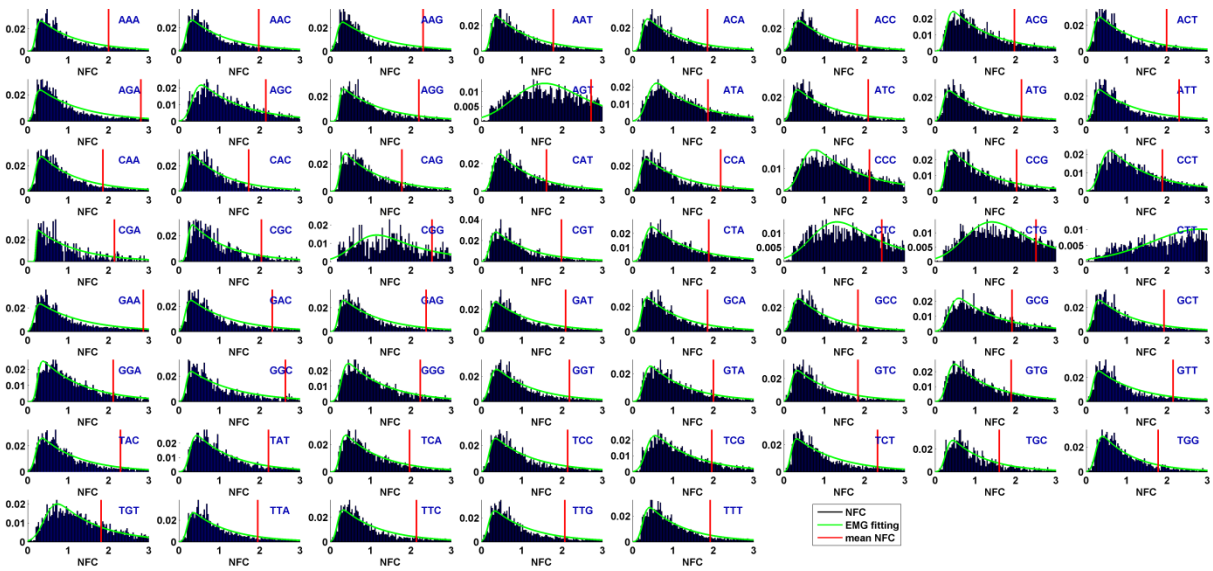

Figure S7. Simulating the influence of translational pauses on *S. cerevisiae* ribosomal profiles. Location of translational pauses were set according to the location of the translational pauses in the real ribosomal profiles. The black histograms depict NFC values of each codon type, collected from all simulated genes; the green curve represents the EMG fitting of the NFC distributions, while the red line represents the mean NFC value for each codon type.

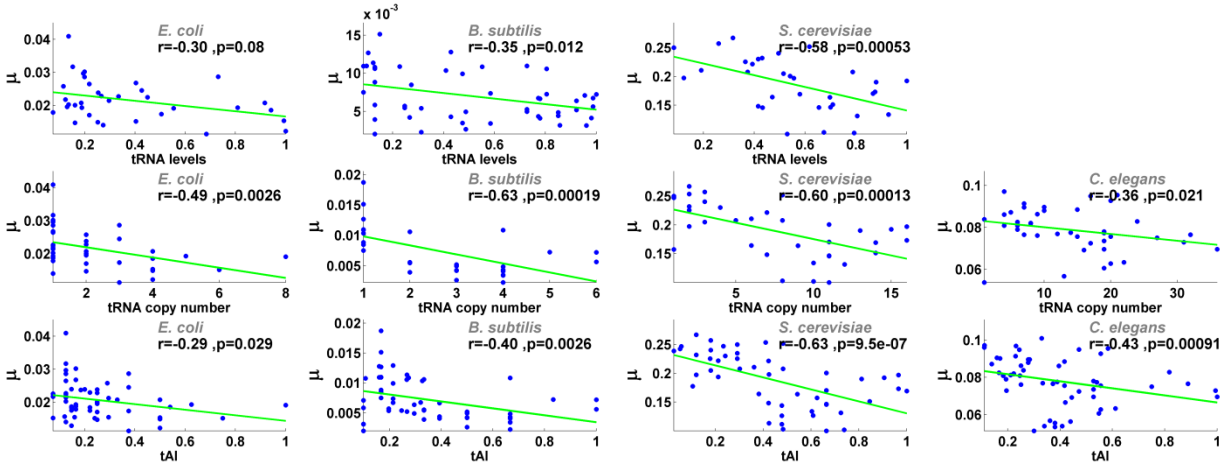

Figure S8. Spearman correlation between positive tRNA levels/copy numbers/tAI index and the μ parameter of the EMG distribution of all codons, when considering an equal amount of RC for all codon types (*E. coli:* 1500 RC, *B. subtilis*: 2000 RC, *C. elegans*: 5000 RC*, S. cerevisiae:* 1000 RC)

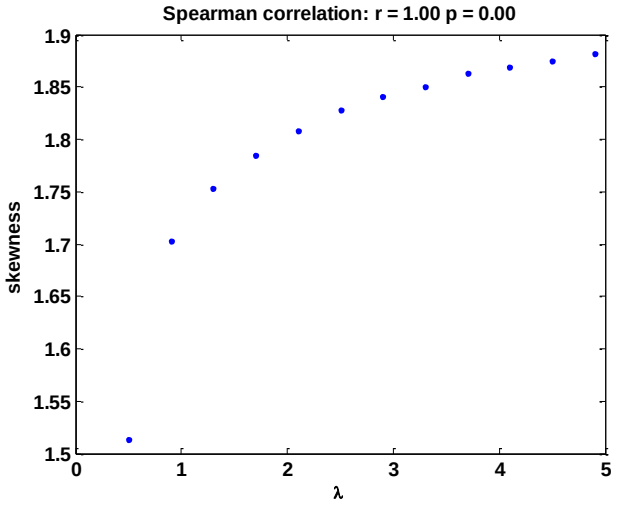

Figure S9. The relationship between skewness of the EMG distribution and the λ parameter of the EMG distribution in a regime close to the analyzed data.

Tables

Table S1. Used thresholds per organism for determining if the RC profile of a gene had sufficient RC to further participate in the analysis. Genes with a median RC lower than described in the second column were excluded. The third column depicts the number of genes participating in the analysis.

Organism	Minimal median RC per gene	Number of analyzed genes
E. coli	1	2832
B. subtilis	1	1862
C. elegans	1	3959
S. cerevisiae	1	1189

Table S2. NFC standard deviation of the analyzed organisms

Organism	Median standard deviation of NFC values	Mean standard deviation of NFC values
E. coli	1.396	1.413
B. subtilis	1.911	1.926
C. elegans	1.074	1.114
S. cerevisiae	0.969	0.943

Table S3. Simulated codons decoding times (CDT) of *S. cerevisiae*, based on the tAI index.

Codon	CDT	Codon	CDT
AAA	0.4149	GAA	0.2213
AAC	0.3017	GAC	0.2074
AAG	0.2004	GAG	0.4881
AAT	0.6873	GAT	0.4725
ACA	0.6637	GCA	0.5531
ACC	0.4191	GCC	0.4191
ACG	1.2766	GCG	1.7287
ACT	0.3017	GCT	0.3017
AGA	0.2766	GGA	1.1064
AGC	1.6595	GGC	0.2074
AGG	0.6858	GGG	1.1213
AGT	3.7803	GGT	0.4725
ATA	1.6585	GTA	1.1058
ATC	0.3204	GTC	0.3293
ATG	0.3017	GTG	1.1213
ATT	0.247	GTT	0.2371
CAA	0.3688	TAA	0
CAC	0.4149	TAC	0.4149
CAG	0.8554	TAG	0
CAT	0.9451	TAT	0.9451
CCA	0.3319	TCA	0.8295
CCC	2.3049	TCC	0.4191
CCG	1.0372	TCG	1.4557
CCT	1.6595	TCT	0.3017
CGA	0.5531	TGA	0
CGC	0.6586	TGC	0.8298
CGG	3.3191	TGG	0.5252
CGT	0.4742	TGT	1.8901
CTA	1.1064	TTA	0.4742
CTC	3.3191	TTC	0.3017
CTG	3.4574	TTG	0.2712
CTT	7.5606	TTT	0.6873

Table S4. Spearman correlation between positive tRNA levels/copy numbers/tAI index and the mean NFC value.

	corr(tRNA levels, mean NFC)	corr(tRNA copy numbers, mean NFC)	corr (tAI, mean NFC)
E. coli	r= -0.11 p = 0.5	r= -0.24 p = 0.14	r= 0.07 p = 0.62
B. subtilis	r = 0.13 p = 0.36	r = 0.01 p = 0.94	r = 0.01 p = 0.95
C. elegans	-	r = 0.23 p = 0.12	r = 0.18 p = 0.17
S. cerevisiae	r = 0.10 p = 0.55	r = -0.02 p = 0.91	r = 0.15 p = 0.26

Table S5. Spearman correlation between positive tRNA levels/copy numbers/tAI index and the mean NFC value, when considering an equal amount of RC for all codon types (*E. coli:* 1500 RC, *B. subtilis*: 2000 RC, *C. elegans*: 5000 RC*, S. cerevisiae:* 1000 RC)

	corr(tRNA levels, mean NFC)	corr(tRNA copy numbers, mean NFC)	corr (tAI, mean NFC)
E. coli	r = -0.03 p = 0.87	r= -0.18 p = 0.3	r= 0.07 p = 0.62
B. subtilis	r = -0.01 p = 0.96	r = 0.01 p = 0.95	r = 0.01 p = 0.93
C. elegans	-	r = 0.32 p = 0.04	r = 0.18 p = 0.19
S. cerevisiae	r = 0.25 p = 0.16	r =0.10 p = 0.58	r = 0.13 p = 0.36

Table S6: Akaike score when fitting the NFC data using EMG/Gaussian/exponential distribution

	AIC of EMG distribution	AIC of Gaussian distribution	AIC of exponential distribution
E. coli	223331.46	1470634.08	242875.23
B. subtilis	119060.09	1591053.56	121637.49
C. elegans	64844.67	67572.42	65576.29
S. cerevisiae	63608.54	818301.73	64382.86

Table S7: Akaike score for the EMG model with free λ parameter and constant λ parameter optimized under log-likelihood criterion for all codons

	AIC of EMG for free λ parameter	AIC of EMG for constant λ parameter
E. coli	223331.46	223667.40
B. subtilis	119060.09	119801.66
C. elegans	64844.67	65032.49
S. cerevisiae	63608.54	66239.55

Table S8. Spearman correlation between positive tRNA levels/copy numbers/tAI index and μ estimated parameter when estimating a global λ parameter for all codons based on the log-likelihood criterion.

	corr(tRNA levels, μ)	corr(tRNA copy numbers, μ)	corr (tAI, μ)
E. coli	r = -0.43, p = 0.0068	r = -0.53, p = 0.00046	r = -0.41, p = 0.001
B. subtilis	r = -0.33, p = 0.016	r = -0.75, p = 5e-07	r = -0.54, p = 7.9e-06
C. elegans	N/A	r = -0.09, p = 0.55	r =-0.26, p = 0.04,
S. cerevisiae	r = -0.61, p = 4.7e-05	r =-0.67, p = 1.1e-06	r =-0.62, p = 1e-07

Table S9. Correlation between μ and positive tRNA levels/copy numbers/tAI index when introducing a small offset of 3/6/9 nucleotides to the estimated P site of *E. coli*

	corr(tRNA levels, μ)	corr(tRNA copy numbers, μ)	corr (tAI, μ)
3 nucleotides	r = -0.30, p = 0.18	r = -0.2, p=0.36	r = -0.00, p=0.99
6 nucleotides	r = -0.04, p = 0.83	r = 0.07, p=0.71	r = 0.01, p=0.94
9 nucleotides	r= 0.13, p = 0.56	r = 0.2, p=0.37	r = 0.17, p=0.3

Table S10. Correlation between μ and positive tRNA levels/copy numbers/tAI index when introducing a small offset of 3/6/9 nucleotides to the estimated P site of *B. subtilis*

	corr(tRNA levels, μ)	corr(tRNA copy numbers, μ)	corr (tAI, μ)
3 nucleotides	r = -0.29, p = 0.039	r = -0.17, p=0.35	r = -0.07, p=0.61
6 nucleotides	r = -0.07, p = 0.6	r = -0.36, p=0.039	r = 0.22, p=0.092
9 nucleotides	r= -0.22, p = 0.13	r = -0.21, p=0.15	r = -0.05, p=0.72

Table S11. Estimated μ, σ, λ variables of all analyzed organisms.

*Table appears in a separate file

References

1. 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.
2. Stadler, M., Artiles, K., Pak, J. and Fire, A. (2012) Contributions of mRNA abundance, ribosome loading, and post- or peri-translational effects to temporal repression of C. elegans heterochronic miRNA targets. *Genome research*.
3. Li, G.W., Oh, E. and Weissman, J.S. (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, **484**, 538-541.
4. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**, R25.
5. Ingolia, N.T., Lareau, L.F. and Weissman, J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**, 789-802.
6. Dana, A. and Tuller, T. (2012) Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. *PLoS computational biology*, **8**, e1002755.
7. Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M. and Weissman, J.S. (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature protocols*, **7**, 1534-1550.