Supplementary Information to
Accurate measures of translation efficiency and traffic using ribosome profiling

Juraj Szavits-Nossan & Luca Ciandrini

Contents

1 Details of the model

The main definitions of the model are given in the Materials and Methods section of the main text. Here we compute the ribosomal current J and the local ribosome density ρ_i using two approximate approaches: the mean-field approximation developed in Ref. [1] and initiation-limited approximation developed in Refs. [2, 3].

The theory summarised in this section constitutes the foundations of NEAR, whose procedure is given in Section 2.

1.1 Inhomogeneous TASEP in the mean-field approximation

We define J_i to be the flux of ribosomes going from codon i to codon $i+1$, as

$$J_i = \begin{cases} \alpha P(\tau_2 = 0, \dots, \tau_{\ell+1} = 0), & i = 1 \\ k_i P(\tau_i = 1, \tau_{i+\ell} = 0), & i = 2, \dots, L - \ell \\ k_i P(\tau_i = 1), & i = L - \ell + 1, \dots, L \end{cases}$$

The conservation of density requires that

$$\frac{d\rho_i}{dt} = J_{i-1} - J_i.$$

In the stationary state $\mathrm{d}\rho_i/\mathrm{d}t = 0$ and thus $J_i = J$ is constant across the transcript. In the mean-field approximation, correlations between ribosomes are ignored, which leads to the following system of equations for J and ρ_i [1]

$$J = \alpha \left(1 - \sum_{s=1}^{\ell} \rho_{1+s} \right) \tag{1a}$$

$$J = k_i \rho_i \frac{1 - \sum_{s=1}^{\ell} \rho_{i+s}}{1 - \sum_{s=1}^{\ell} \rho_{i+s} + \rho_{i+\ell}}, \quad i = 2, \dots, L - \ell \tag{1b}$$

$$J = k_i \rho_i, \quad i = L - \ell + 1, \dots, L. \tag{1c}$$

A closed-form expression for ρ_i as a function of α and $k_2, \dots, k_L$ is not known. However, it is straightforward to express the ratio k_i/α as a function of the local densities:

$$\frac{k_i}{\alpha} = \frac{\left(1 - \sum_{s=1}^{\ell} \rho_{1+s}\right)\left(1 - \sum_{s=1}^{\ell} \rho_{i+s} + \rho_{i+\ell}\right)}{\rho_i \left(1 - \sum_{s=1}^{\ell} \rho_{i+s}\right)}, \quad i = 2, \dots, L - \ell \tag{2a}$$

$$\frac{k_i}{\alpha} = \frac{1 - \sum_{s=1}^{\ell} \rho_{1+s}}{\rho_i}, \quad i = L - \ell + 1, \dots, L. \tag{2b}$$

We will use these expressions as a starting point for the nonlinear least-squares minimisation procedure in Section 2, which is the core of NEAR.

1.2 Inhomogeneous TASEP in the initiation-limited approximation

Recently we introduced a method for computing J, ρ_i and ρ by expanding $P(C)$ in the initiation rate α [2, 3],

$$P(C) = \sum_{n=0}^{\infty} c_n(C) \alpha^n.$$

Since $P(C)$ must sum to 1 the coefficients $c_n(C)$ must obey

$$\sum_C c_n(C) = \begin{cases} 1 & n = 0 \\ 0 & n \geq 1 \end{cases}, \tag{3}$$

The biggest advantage of the power series method is that many coefficients $c_n(C)$ are in fact equal to zero. If $N(C)$ denotes the number of ribosomes in a configuration C then

$$c_n(C) = 0 \text{ if } N(C) > n.$$

This non-trivial result follows from a graph-theoretical interpretation of Markov chains applied to the TASEP. The method then delineates how to compute the coefficients $c_n(C)$ recursively starting from $n = 0$, see Refs. [2, 3] for more details.

The initiation-limited approximation amounts to replacing the series $P(C)$ with a finite sum,

$$P(C) \approx \sum_{n=0}^{K} c_n(C)\alpha^n, \tag{4}$$

In this work we compute coefficients up to and including the third order ($K = 3$), which we present in detail below. From the result in Eq. (1.2) it follows that for $K = 3$ we need to look at configurations with at most three ribosomes. These are $C = \emptyset$ (no ribosomes), $C = A_i$ (one ribosome at codon i), $C = A_iA_j$ (two ribosomes at codons i and j) and $C = A_iA_jA_k$ (three ribosomes at codons i, j and k).

1.2.1 Zero and first order

According to Eq. (1.2) $c_0(C) = 1$ if $C = \emptyset$ and is equal to 0 otherwise,

$$c_0(C) = \begin{cases} 1, & C = \emptyset \\ 0, & \text{otherwise} \end{cases}$$

This is equivalent of saying that if translation initiation is not allowed ($\alpha = 0$), then the transcript will become completely empty with probability 1.

For $n = 1$ (first order), $c_1(C) \neq 0$ only if C contains at most one particle. The corresponding coefficients $c_1(\emptyset)$ and $c_1(A_i)$ are equal to

$$c_1(\emptyset) = -\sum_{i=2}^{L} \frac{1}{k_i}, \tag{5a}$$

$$c_1(A_i) = \frac{1}{k_i},\ i = 2, \ldots, L. \tag{5b}$$

These coefficients yields the following expressions for J and ρ_i in the first-order approximation

$$J = \alpha \tag{6}$$

$$\rho_i = \frac{\alpha}{k_i}. \tag{7}$$

Since at the first order at most one particle occupies the lattice, this is equivalent to neglecting interference between ribosomes. We emphasise that the first order relation between rates and densities given in Eq. (7) is widely used in the interpretation of ribosome profiling.

1.2.2 Second order

For $n = 2$ (second order), $c_2(C) \neq 0$ only if C contains at most two particles. The equations for $c_2(C)$ are more complicated than for $c_1(C)$ and must be solved numerically. Before we write down the equations, we introduce Kronecker delta function δ_{ij} and unit step function $\theta(i)$ defined as

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad \theta[i] = \begin{cases} 1 & i \geq 0 \\ 0 & i < 0 \end{cases}. \tag{8}$$

The equations for two-particle coefficients $c_2(A_iA_j)$ then read

$$c_2(A_iA_j) = \frac{\delta_{i,2}}{e_0(A_iA_j)} c_1(A_j) + \frac{\theta[i-3]k_{i-1}}{e_0(A_iA_j)} c_2(A_{i-1}A_j) + \frac{\theta[j-1-i-\ell]k_{j-1}}{e_0(A_iA_j)} c_2(A_iA_{j-1}), \tag{9}$$

where $e_0(A_iA_j) = \theta_{j-\ell-i-1}k_i + k_j$. These equations can be solved recursively starting from $i = 2$ and iterating over j from $j = \ell + 2$ to $j = L$ with i held fixed. Then $i = 3$ is held fixed and the iteration goes over j from $j = \ell + 3$ to L and so on until $i = L - \ell$ and $j = L$. Once all $c_2(A_iA_j)$ are found, the equations for one-particle coefficients $c_2(A_i)$ are solved recursively from $i = 2$ to $i = L$,

$$c_2(A_i) = \frac{\delta_{i,2}}{k_i} c_1(\emptyset) + \frac{\theta[i-3]k_{i-1}}{k_i} c_2(A_{i-1}) + \frac{\theta[L-\ell-i]k_L}{k_i} c_2(A_iA_L). \tag{10}$$

Finally, once all $c_2(A_iA_j)$ and $c_2(A_i)$ are computed, $c_2(\emptyset)$ follows from Eq. (3),

$$c_2(\emptyset) = -\sum_{i=2}^{L-\ell} \sum_{j=i+\ell}^{L} c_2(A_iA_j) - \sum_{i=2}^{L} c_2(A_i). \tag{11}$$

Using these coefficients we can compute J and ρ_i in the second-order approximation according to

$$J = \alpha - k_L \left(\sum_{j=2}^{L-\ell} c_2(A_j A_L) + c_2(A_L) \right) \alpha^2 = \alpha - \sum_{j=2}^{\ell+1} \frac{\alpha^2}{k_i} \tag{12}$$

$$\rho_i = \frac{\alpha}{k_i} + \left(\sum_{j=2}^{i-\ell} c_2(A_j A_i) + \sum_{j=i+\ell}^{L} c_2(A_i A_j) + c_2(A_i) \right) \alpha^2. \tag{13}$$

1.2.3 Third order

For $n = 3$ (third order), $c_3(C) \neq 0$ only if C contains at most three particles. The equations for three-particle coefficients $c_2(A_i A_j A_m)$ read

$$\begin{aligned} c_3(A_i A_j A_m) = {} & \frac{\delta_{i,2}}{e_0(A_i A_j A_m)} c_2(A_j A_m) + \frac{\theta[i-3] k_{i-1}}{e_0(A_i A_j A_m)} c_3(A_{i-1} A_j A_m) \\ & + \frac{\theta[j-1-i-\ell] k_{j-1}}{e_0(A_i A_j A_m)} c_3(A_i A_{j-1} A_m) + \frac{\theta[m-1-j-\ell] k_{m-1}}{e_0(A_i A_j A_m)} c_3(A_i A_j A_{m-1}), \end{aligned} \tag{14}$$

where $e_0(A_i A_j A_m)$ is given by

$$e_0(A_i A_j A_m) = \theta[j-\ell-i-1] k_i + \theta[m-\ell-j-1] k_j + k_m. \tag{15}$$

These equations are solved recursively starting from $i = 2$ and $j = \ell + 2$ fixed and iterating over m from $2\ell + 2$ to L. Then j is increased to $\ell + 3$ and the iteration over m is repeated from $2\ell + 3$ to L. The procedure of increasing i and j and iterating over m is repeated so on until $i = L - 2\ell$, $j = L - \ell$ and $m = L$.

Once all $c_3(A_i A_j A_m)$ are found, the equations for two-particle coefficients $c_3(A_i A_j)$ are solved recursively starting from $i = 2$ and $j = \ell + 2$,

$$\begin{aligned} c_3(A_i A_j) = {} & \frac{\delta_{i,2}}{e_0(A_i A_j)} c_2(A_j) + \frac{\theta[i-3] k_{i-1}}{e_0(A_i A_j)} c_3(A_{i-1} A_j) \\ & + \frac{\theta[j-1-i-\ell] k_{j-1}}{e_0(A_i A_j)} c_3(A_i A_{j-1}) + \frac{\theta[L-\ell-j] k_L}{e_0(A_i A_j)} c_3(A_i A_j A_L). \end{aligned} \tag{16}$$

The equations for one-particle coefficients $c_3(A_i)$ are then solved recursively from $i = 2$ to $i = L$,

$$c_3(A_i) = \frac{\delta_{i,2}}{k_i} c_2(\emptyset) + \frac{\theta[i-3] k_{i-1}}{k_i} c_3(A_{i-1}) + \frac{\theta[L-\ell-i] k_L}{k_i} c_3(A_i A_L). \tag{17}$$

Finally, the coefficient $c_3(\emptyset)$ is given by

$$c_3(\emptyset) = - \sum_{i=2}^{L-2\ell} \sum_{j=i+\ell}^{L-\ell} \sum_{m=j+\ell}^{L} c_3(A_i A_j A_m) - \sum_{i=2}^{L-\ell} \sum_{j=i+\ell}^{L} c_3(A_i A_j) - \sum_{i=2}^{L} c_3(A_i). \tag{18}$$

Once all coefficients for all orders up to and including the third order are computed, J and ρ_i can be computed according to the following expressions

$$J = \alpha - \sum_{j=2}^{\ell+1} \frac{1}{k_i} \alpha^2 + k_L \left(\sum_{j=2}^{L-2\ell} \sum_{m=j+\ell}^{L-\ell} c_3(A_j A_m A_L) + \sum_{j=2}^{L-\ell} c_3(A_j A_L) + c_3(A_L) \right) \alpha^3 \tag{19}$$

$$\rho_i = f_i^{(1)} \alpha + f_i^{(2)} \alpha^2 + f_i^{(3)} \alpha^3. \tag{20}$$

The coefficients $f_i^{(1)}$, $f_i^{(2)}$ and $f_i^{(3)}$ depend only on $k_2, \ldots, k_L$ and are given by

$$f_i^{(1)} = \frac{1}{k_i} \tag{21a}$$

$$f_i^{(2)} = \sum_{j=2}^{i-\ell} c_2(A_j A_i) + \sum_{j=i+\ell}^{L} c_2(A_i A_j) + c_2(A_i) \tag{21b}$$

$$\begin{aligned} f_i^{(3)} = {} & \sum_{j=i+\ell}^{L-\ell} \sum_{m=j+\ell}^{L} c_3(A_i A_j A_m) + \sum_{j=2}^{i-\ell} \sum_{m=i+\ell}^{L} c_3(A_j A_i A_m) + \sum_{j=2}^{i-2\ell} \sum_{m=j+\ell}^{i-\ell} c_3(A_j A_m A_i) \\ & + \sum_{j=2}^{i-\ell} c_3(A_j A_i) + \sum_{j=i+\ell}^{L} c_3(A_i A_j) + c_3(A_i). \end{aligned} \tag{21c}$$

From the stationary master equation (5) in the Material and Methods section of the main text, it follows that $P(C)$ and thus ρ_i are functions of $k_2/\alpha, \ldots, k_L/\alpha$ only. Using this fact we can rewrite Eq. (20) as

$$\rho_i = g_i^{(1)} + g_i^{(2)} + g_i^{(3)}, \tag{22}$$

where

$$g_i^{(n)}\left(\frac{k_2}{\alpha}, \ldots, \frac{k_L}{\alpha}\right) = \alpha^n f_i^{(n)}(k_2 \ldots, k_L), \quad n = 1, 2, 3, \ldots. \tag{23}$$

2 The procedure of inferring elongation-to-initiation ratios $\kappa_i = k_i/\alpha$ from ribosome profiling data

2.1 Details of Monte Carlo simulations

All Monte Carlo simulations were performed using the Gillespie algorithm. In the first part of the simulation we checked the total density ρ every $100 \cdot L$ updates until the percentage error between two values of the total density ρ was less than 0.1%. After that we ran the simulation for further $M = 10^4 \cdot L$ updates during which we computed the time average of ρ_i defined as

$$\rho_i^{\text{sim}} = \frac{1}{T}\sum_{n=1}^{M} \tau_i^{(n)} dt^{(n)}, \tag{24}$$

where $\tau_i^{(n)}$ is the value of τ_i (1 if codon i is occupied by the ribosome's A-site and 0 otherwise) just before the n-th update in the simulation, $dt^{(n)}$ is the time interval between the $(n-1)$-th and the n-th iteration of the Gillespie algorithm, and $T = \sum_{n=1}^{M} dt^{(n)}$ is the total time.

2.2 Normalisation of Ribo-seq data

We used A-site ribosome profiles of 849 genes of *Saccharomyces cerevisiae* obtained by Dao Duc and Song [4] from the experimental data of Weinberg et al [5]. These genes were selected from the pool of 5887 genes to have more than 200 codons and for which the average number of A-site occurrences ("reads") per codon was greater than 10. We further reduced this list to 839 genes for which the total ribosome density (the number of ribosomes per gene length) is known from polysome profiling experiments by Mackay et al [6].

For each gene we computed the value of local experimental ribosome density r_i at codons $i = 2, \ldots, L$ according to

$$r_i = \rho \frac{N_i}{\sum_{n=2}^{L} N_n}, \quad i = 2, \ldots, L \tag{25}$$

where N_i is the number of reads at codon i obtained from ribosome profiling experiments, ρ is the total ribosome density and L is the number of codons for that particular gene. Most genes in our analysis had codons with zero reads, which implies an infinite value of the elongation rate. In order to mitigate this problem we replaced $N_i = 0$ by $N_i = 1$ and calculated r_i according to Eq. (25). The normalisation procedure failed for one gene out of 849, gene YHR094C, for which the obtained ribosome density r_i was larger than 1 at some codons.

2.3 Solving nonlinear equations for $\kappa_i = k_i/\alpha$

We looked for $\kappa_i = k_i/\alpha$ such that the theoretical ribosome density ρ_i in Eq. (20) matches the experimental density r_i,

$$\rho_i\left(\kappa_2, \ldots, \kappa_L\right) = r_i, \quad i = 2, \ldots, L. \tag{26}$$

The procedure of finding κ_i was the following. First we estimated the values of κ_i^{MF} in the mean-field approximation according to Eq. (2). Next we computed local ribosome densities $\rho_i^{\text{sim}}(\{\kappa_i^{\text{MF}}\})$ by running a stochastic simulation of the TASEP using the Gillespie algorithm in which the elongation rate k_i at codon i was set to κ_i^{MF} estimated in the mean-field approximation (i.e. setting $\alpha = 1$ in the simulation without loss of generality). Next we solved a nonlinear least squares optimisation problem using $\{\kappa_i^{\text{MF}}\}$ as a starting point, which consisted of finding $x_i = k_i/\alpha$ for which the sum of squares

$$S(\kappa_2, \ldots, \kappa_L) = \sum_{j=2}^{L} \left(\rho_j\left(\kappa_2, \ldots, \kappa_L\right) - r_j\right)^2 \tag{27}$$

was minimal.

In the initiation-limited regime we expect the local density ρ_i to increase with increasing α, which means that the first derivative of ρ_i must be non-negative,

$$\frac{\partial \rho_i}{\partial \alpha} = f_i^{(1)} + 2f_i^{(2)}\alpha + 3f_i^{(3)}\alpha^2 \geq 0. \tag{28}$$

If this criterion is not met, the approximation of ρ_i by a cubic polynomial may lead to erroneous results such as negative density. This problem can be mitigated by computing higher orders beyond the third one. However, computing higher orders takes considerably more time and memory and may not be practical for analysing many genes. Instead in the optimisation procedure we took a more pragmatic approach in which we approximated ρ_i by $g_i^{(1)}$ whenever (28) was not met

$$\rho_i = \begin{cases} g_i^{(1)} + g_i^{(2)} + g_i^{(3)}, & g_i^{(1)} + 2g_i^{(2)} + 3g_i^{(3)} > 0 \\ g_i^{(1)}, & \text{otherwise.} \end{cases} \tag{29}$$

In order to prevent unrealistic values of k_i/α and to speed up the optimisation procedure, we also restricted the search of $\kappa_2, \ldots, \kappa_L$ to

$$10^{-3} \leq \frac{k_i}{\alpha} \leq 10^{6}. \tag{30}$$

The nonlinear optimisation was preformed using NLopt library from Ref. [7]. We used a local derivative-free algorithm called BOBYQA which was developed in Ref. [8]. The optimisation search was preformed until one of the following three stopping criteria was met: (1) the fractional error $\Delta S/S$ between two iterations was less than 10^{-10}, (2) the number of evaluations $N_{\text{eval}} = 200 \cdot N$ and (3) the total run time exceeded 144 minutes. These numbers were chosen due to time constraints (maximum 1 day for the optimisation of 10 genes on a single processor) when analysing many genes; a better accuracy may be achieved for individual genes by amending the stopping criteria. After the optimisation procedure finished we recomputed local densities ρ_i^{sim} by running a stochastic simulation of the TASEP with elongation rates $k_i = \kappa_i$ and compared them to experimental values r_i.

2.4 Dealing with outliers

After optimisation of the rates explained in the previous section, comparison between local densities ρ_i^{sim} and experimental values r_i sometimes reveals codons at which the disagreement between ρ_i^{sim} and r_i is unexpectedly large.

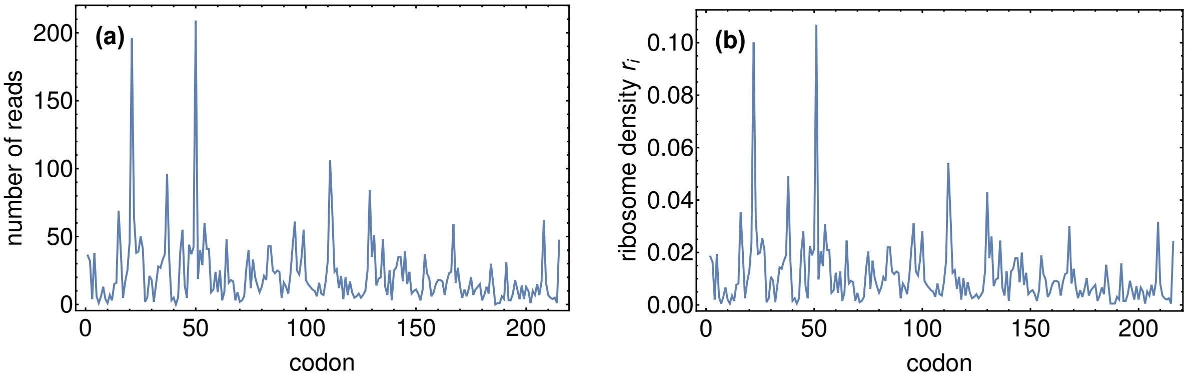

Figure S1: Ribo-seq data for gene YAL007C of *Saccharomyces cerevisiae*. (a) Number of reads of A-site positions. (b) Local ribosome density obtained by normalising ribosome profiling data according to Eq. (25).

A typical example is gene YAL007C for which the Ribo-seq and normalised Ribo-seq data are presented in Figures S1(a) and S1(b), respectively. The relative error of at least 70% between local densities $\rho_i^{\text{sim}}(\{\kappa_i\})$ predicted by the TASEP using optimised rates and experimental values r_i is found at two sets of codons: $\{7, 11, 12, 13\}$ and $\{41, 42\}$ (dashed vertical lines in Figure S2(a) and red circles in Figure S2(b)).

We first checked whether this discrepancy is due to truncation of the power series in Eq. (4). We did this by comparing ρ_i^{sim} computed from simulations using optimised rates with the theoretical prediction ρ_i^{PSM} from Eq. (20). Any discrepancy between these two profiles would indicate that the third-order approximation was not good enough and that higher-orders terms are needed. On the contrary, Figures S2(c)-(d) demonstrate that the third-order approximation ρ_i^{PSM} correctly reproduces simulated densities ρ_i^{sim}.

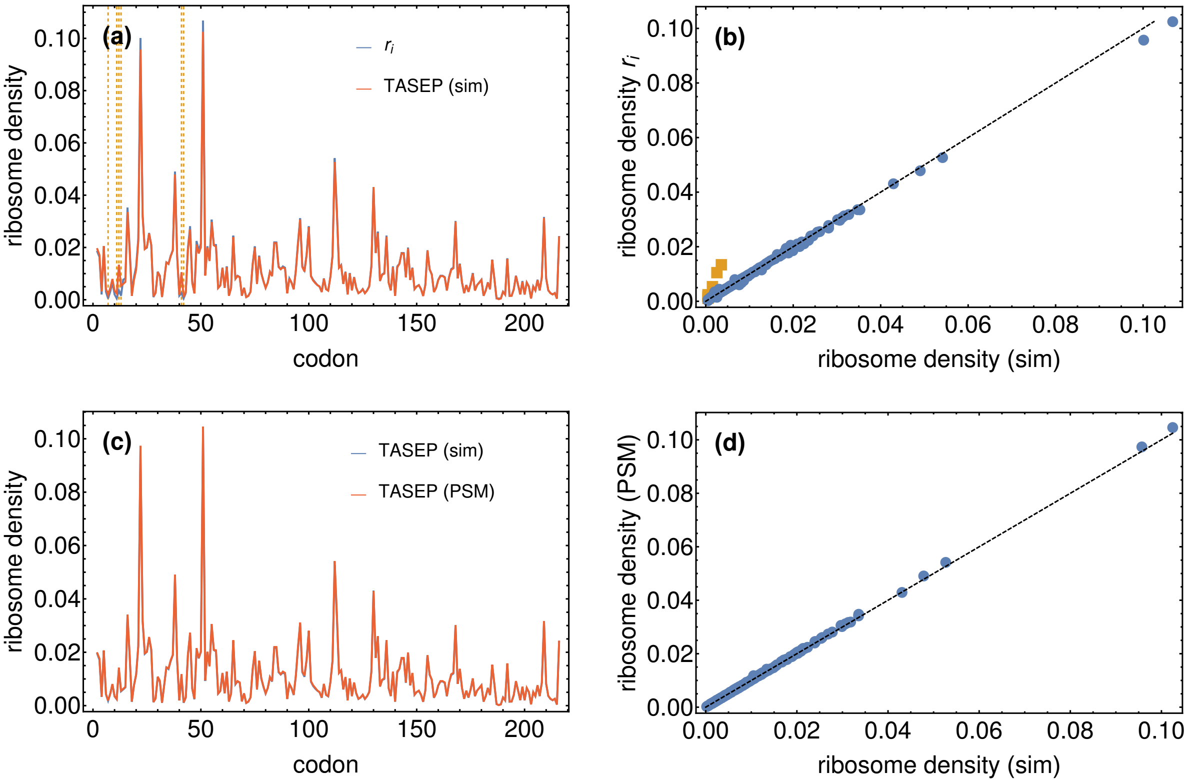

Figure S2: Results of the NEAR procedure for gene YAL007C of *Saccharomyces cerevisiae*. (a) Comparison between local ribosome density $\rho_i^{\text{sim}}(\{\kappa_i\})$ obtained using optimised rates and experimental density r_i. Vertical dashed lines denote codon positions at which the relative error between ρ_i^{sim} and r_i is larger than 70%. (b) Scatter plot of r_i vs ρ_i^{sim} showing outliers from (a) (orange squares). (c) Comparison between local ribosome density ρ_i^{PSM} obtained from Eq. (20) and simulated density ρ_i^{sim} obtained using the same optimised rates. (d) Scatter plot of ρ_i^{sim} vs ρ_i^{PSM} showing excellent agreement between the two datasets.

This leaves us with two possibilities for the explanation of the outliers. One possibility is that the optimisation procedure failed to find the best solution. This can happen because we are searching for a local minimum of the objective function S and we do not know for sure if there is a better solution somewhere in the large space of parameters.

The other possibility is that the data cannot be fully described by the TASEP model. Indeed, a closer look at the Ribo-seq data reveals a small number of reads at codons 11 and 41 (equal to 7 and 0 respectively) compared to a large number of reads 10 codons downstream (equal to 196 and 209 respectively). This scenario is unlikely to happen in the TASEP: a large density at codon i implies large density at codon $i-10$ due to excluded volume interactions between ribosomes. Perhaps the model misses an important step in mRNA translation such as premature termination due to ribosome dropping off the mRNA. It is also possible that the observed discrepancy is due to the known bias in ribosome profiling that discards clustered ribosomes. We hence develop a quality check for the estimated κ_i in Section 2.5 that excludes the estimates of codons that we rate as unreliable for the reasons explained above.

2.5 NEAR quality check of $\{\kappa_i\}$

In this section we detail the quality check that we carry out for each value of the inferred $\kappa_i = k_i/\alpha$. For each gene for which the optimisation procedure was successful (see previous section) we preformed the following five steps.

1. *Overall improvement over the initial (mean-field) prediction.* We check whether the optimisation procedure has improved the agreement with respect to the initial (mean-field) prediction. For the mean-field prediction we compute the sum of squares S_{MF}

$$S_{\text{MF}} = \sum_{i=2}^{L} \left(\rho_i^{\text{sim}}(\{\kappa_i^{\text{MF}}\}) - r_i\right)^2, \tag{31}$$

where $\{\rho_i^{\rm sim}(\{\kappa_i^{\rm MF}\})\}$ is the simulated density profile computed with the mean-field rates $\{\kappa_i^{\rm MF}\}$ and r_i is the experimental density profile. The value of $S_{\rm MF}$ is then compared to $S_{\rm opt}$ obtained from

$$S_{\rm opt} = \sum_{i=2}^{L} \left(\rho_i^{\rm sim}(\{\kappa_i\}) - r_i\right)^2 , \tag{32}$$

where $\{\rho_i^{\rm sim}(\{\kappa_i\})\}$ is the simulated density profile computed using the inferred elongation-to-initiation ratios $\{\kappa_i\}$. If $S_{\rm opt} < S_{\rm MF}$ then the TIE is taken from the simulations of the optimised system, otherwise from the MF simulations.

2. *Rate-limiting step in translation.* We verify if the optimised elongation-to-initiation ratio $\kappa_i > 1 \ \forall i$ (otherwise initiation is not the limiting step and our framework cannot be used).

3. *Applicability of the power series method.* We set a tolerance $\epsilon_{\rm PSM}$ for the power series method (3rd order approximation). For each codon we check if $|\rho_i^{\rm PSM} - \rho_i^{\rm sim}(\{\kappa_i\})|/\rho_i^{\rm sim} < \epsilon_{\rm PSM}$. If that is the case the power series approximation holds and the method is reliable. We set $\epsilon_{\rm PSM} = 0.1$.

4. *Comparison with Ribo-seq data.* If the codon passes the quality checks in points 2 and 3 then we check if the prediction is consistent with the experimental profile $\{r_i\}$ (within a tolerance $\epsilon_{\rm EXP}$) by checking if $|r_i - \rho_i^{\rm sim}(\{\kappa_i\})|/r_i < \epsilon_{\rm EXP}$, where we set $\epsilon_{\rm EXP} = 0.05$. Codons that pass this check are kept and considered reliable only if $\kappa_i < \kappa_{\rm thr}$. This last check is to discard values of κ_i that are deemed suspiciously large and are likely unphysical. Such unreasonably large values of κ_i are due to low number of reads for a particular codon, especially for codons for which we artificially increased the number reads from zero to one in order to avoid infinite elongation rates. Because such small number of reads may be due to experimental errors, we decided to exclude those codons from the final analysis. We used the threshold value $\kappa_{\rm thr} = 1000$ corresponding to an elongation rate $k \sim 120$/s (assuming $\alpha = 0.12$/s).

5. *Problematic codons.* If $|r_i - \rho_i^{\rm sim}(\{\kappa_i\})|/r_i > \epsilon_{\rm EXP}$ then the codon is excluded from the final analysis and we set $\kappa_i = -1$ (to identify the problematic codon for further analysis). Those codons fall in the most interesting class, in which the experimental data cannot be reproduced using the existing theory. We speculate that those "problematic" codons might also arise because ribosome profiling cannot detect clusters of ribosomes that will instead be predicted by our simulated profile $\{\rho_i^{\rm sim}(\{\kappa_i\})\}$. The set of codons entering in this category (in which our method can be applied but experimental data are inconsistent with the density generated by NEAR) are also gathered in the Supplementary Table 1.

6. *Falling back to the mean-field prediction.* In the case in which the codon does not pass the quality checks in points 2 and 3, the mean-field prediction $\{\kappa_i^{\rm MF}\}$ will be considered. If $|r_i - \{\rho_i^{\rm sim}(\{\kappa_i^{\rm MF}\})\}|/r_i < \epsilon_{\rm EXP}$, then the κ_i is kept and the mean-field prediction is considered reliable. Otherwise the codon is excluded from the final analysis and we set $\kappa_i = -2$.

7. *Stop codon.* We further check if the elongation-to-initiation ratio κ_L of the stop codon has been kept for the analysis. If κ_L is reliable then the ratio $\kappa_i/\kappa_L = k_i/k_L$ can be computed.

3 Computing TEE profiles

For each gene we obtain the TEE profile by computing on each codon i $\text{TEE}_i = \text{TIE}/(\kappa_i \rho_i^{\rm sim})$ (details given in the main text) on the codons where the κ_i passed the quality check. Therefore the profile is "broken" when NEAR cannot find a reliable estimate for the elongation-to-initiation ratio. We note that by definition TEE_i is a probability and as such must take values between 0 and 1. Because each of TIE, κ_i and $\rho_i^{\rm sim}$ comes with its own statistical error, TEE may occasionally be larger than 1.

References

[1] Carolyn T. MacDonald, Julian H. Gibbs, and Allen C. Pipkin. Kinetics of biopolymerization on nucleic acid templates. *Biopolymers*, 6(1):1–25, 1968.

[2] Juraj Szavits-Nossan, Luca Ciandrini, and M. Carmen Romano. Deciphering mrna sequence determinants of protein production rate. *Phys. Rev. Lett.*, 120:128101, Mar 2018.

[3] Juraj Szavits-Nossan, M. Carmen Romano, and Luca Ciandrini. Power series solution of the inhomogeneous exclusion process. *Phys. Rev. E*, 97:052139, May 2018.

[4] Khanh Dao Duc and Yun S. Song. The impact of ribosomal interference, codon usage, and exit tunnel interactions on translation elongation rate variation. *PLOS Genetics*, 14(1):1–32, Jan 2018.

[5] David E. Weinberg, Premal Shah, Stephen W. Eichhorn, Jeffrey A. Hussmann, Joshua B. Plotkin, and David P. Bartel. Improved ribosome-footprint and mrna measurements provide insights into dynamics and regulation of yeast translation. *Cell Reports*, 14(7):1787–1799, 2016.

[6] Vivian L. MacKay, Xiaohong Li, Mark R. Flory, Eileen Turcott, G. Lynn Law, Kyle A. Serikawa, X. L. Xu, Hookeun Lee, David R. Goodlett, Ruedi Aebersold, Lue Ping Zhao, and David R. Morris. Gene expression analyzed by high-resolution state array analysis and quantitative proteomics. *Molecular & Cellular Proteomics*, 3(5):478–489, 2004.

[7] Steven G. Johnson. The nlopt nonlinear-optimization package. http://github.com/stevengj/nlopt.

[8] Michael J. D. Powell. The bobyqa algorithm for bound constrained optimization without derivatives. Technical Report NA2009/06, Department of Applied Mathematics and Theoretical Physics, Cambridge England, 2009.

[9] Luca Ciandrini, Ian Stansfield, and M. Carmen Romano. Ribosome traffic on mrnas maps to gene ontology: Genome-wide quantification of translation initiation rates and polysome size regulation. *PLOS Computational Biology*, 9(1):1–10, Jan 2013.

4 Supplementary Figures

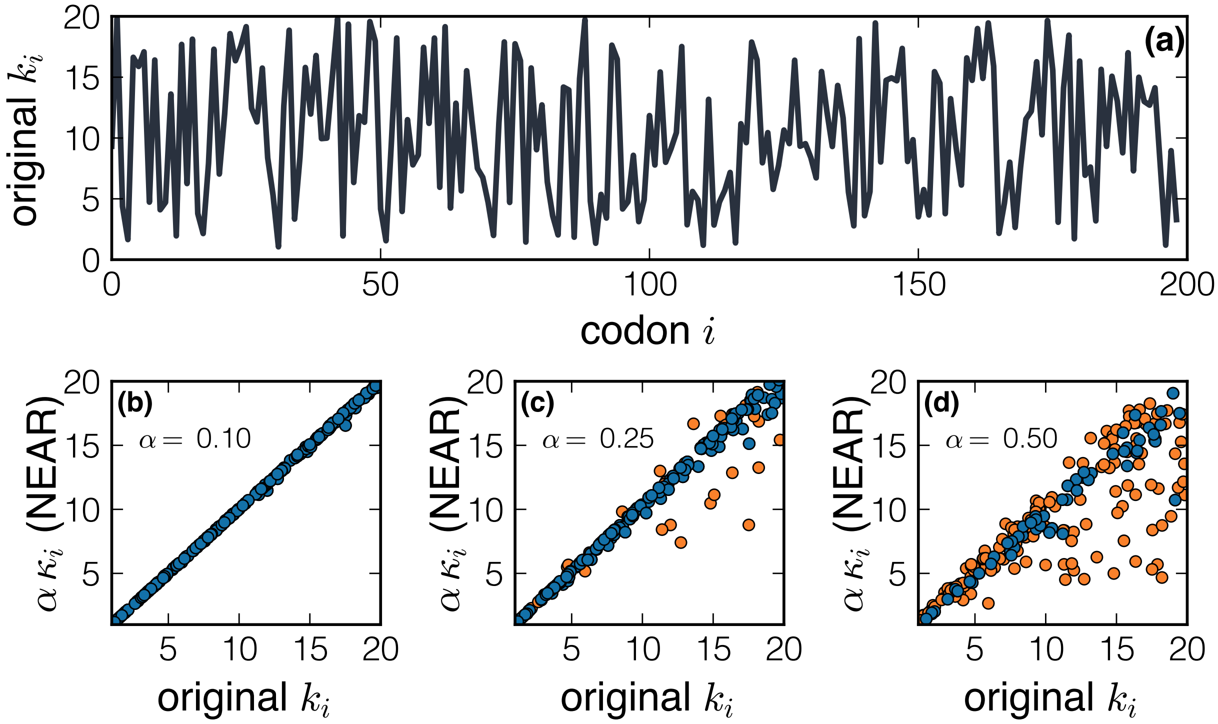

Figure S3: Testing NEAR on a random sequence of rates $\{k_i\}$. The $\{k_i\}$ profile is shown in panel (a). Panels (b)-(c)-(d) show the scatter plot between the rates $\alpha\,\kappa_i$ estimated by NEAR (y-axis) and the original rates $\{k_i\}$ (x-axis) with increasing values of the initiation rate α. The points in orange are the κ_i values that did not pass the quality check of NEAR. Since NEAR is based on a power series approximation that is reliable for small initiation rates, the method is expected to perform badly for large values of α. However, even in this regime the quality check is able to exclude codons that will not be considered in the final analysis. We remind that initiation has been estimated to be limiting and the physiological situation is the one presented in panel (b) for most of the genes [9].

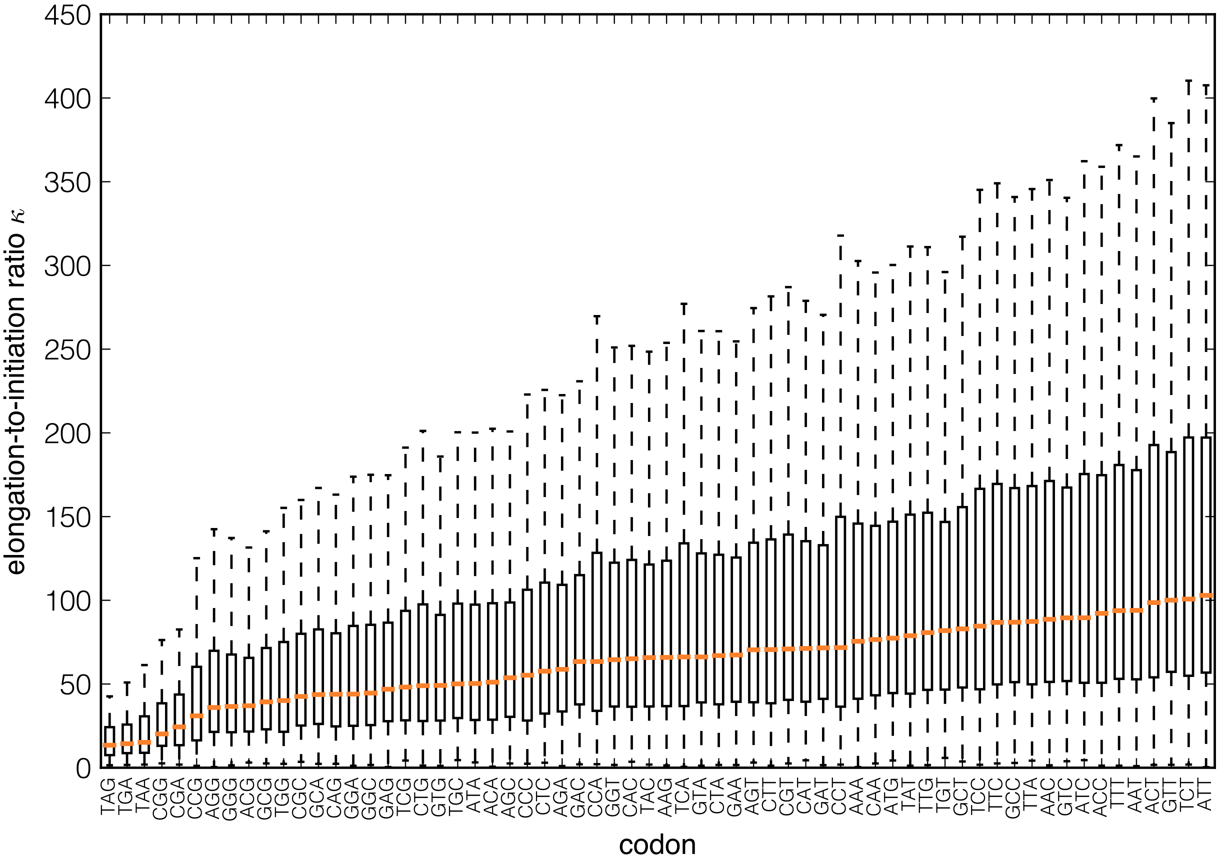

Figure S4: Distributions of elongation-to-initiation ratios κ for all codon types. We gather the ratios $\kappa_i = k_i/\alpha$ for each codon-type, and plot their distributions. In principle codons from different genes cannot be compared because they have a different initiation rate α. However, we notice a small variability in the estimates of STOP codons (first 3 codons in the plot). This is reasonable since STOP codons are supposed to be less context dependent. We also remark that, as opposed to generally believed, STOP codons are the slowest, but still ~ 10 times faster compared to initiation.

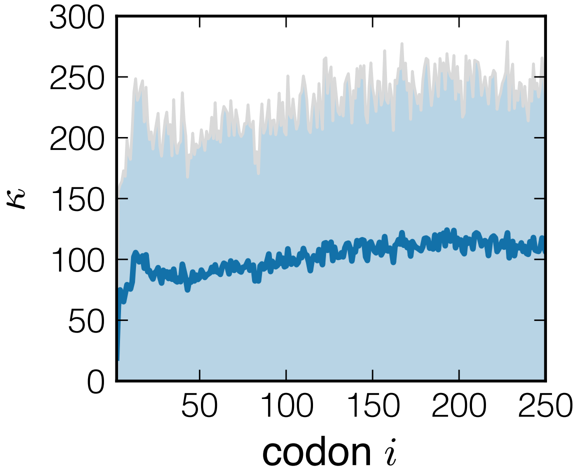

Figure S5: $\{\kappa_i\}$ profile obtained by averaging the κ_i values of each gene. The coloured area represent the standard deviation of the distribution of the elongation-to-initiation ratios at each codon position. Although high variability of elongation rates makes it difficult to conclude something on the profile of individual genes, when averaging among the entire set of genes analysed we remark that in the first part of the coding region the elongation is usually slower than more downstream in the transcript.

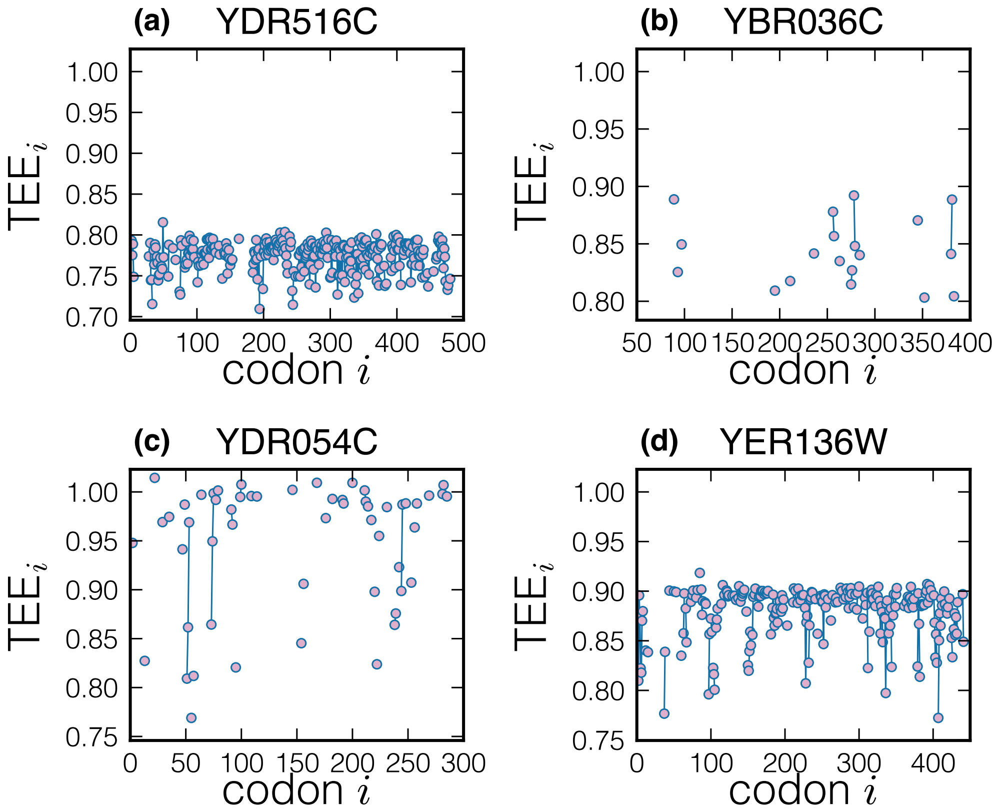

Figure S6: Profiles of TEE for four genes amongst the ones with the smallest average TEE. When the TEE is low (showing a rather uniform ribosome interference of about 15%), many κ_i do not pass the quality check, as it can be seen from the many points missing in the TEE profiles. This means that the experimental read counts are inconsistent with the model. We speculate that this is due to the bias in the ribosome profiling neglecting clusters of ribosomes. However, NEAR fills the partial information that is enclosed in the experimental profile and finds evidence of traffic.

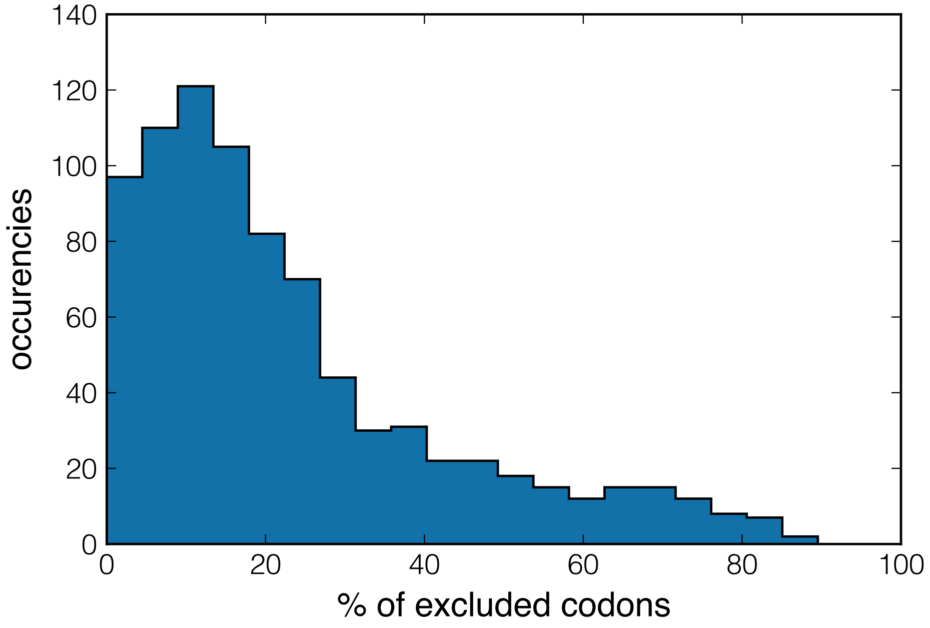

Figure S7: Percentage of codons (per mRNA transcript) that are excluded from the analysis because of the inconsistency between the underlying model and experimental densities (codons that did not pass the second part of point 4 of the quality check of Section 2.5). This is a signature of local ribosome interference that cannot be detected by ribosome profiling, and is highlighted by NEAR. The mean of the distribution is at 23%, the median at 17%.